On the Role of Logic in Information Retrieval*

Fabrizio Sebastiani
Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
Via S. Maria, 46 - 56126 Pisa (Italy)
E-mail: fabrizio@iei.pi.cnr.it

Abstract

The logical approach to information retrieval has recently been the object of active research. It is our contention that researchers have put a lot of effort in trying to address some difficult problems of IR within this framework, but little effort in checking that the resulting models satisfy those well-formedness criteria that, in the field of mathematical logic, are considered essential and conducive to effective modelling of a real-world phenomenon. The main motivation of this paper is not to propose a new logical model of IR, but to discuss some central issues in the application of logic to IR. The first issue we touch upon is the logical relationship we might want to enforce between formulae d, representing a document, and n, representing an information need; we analyse the different implications of models based on *truth*, *validity* or *logical consequentiality*. The relationship between this issue and the issue of *partiality* vs. *totality* of information is subsequently analysed, in the context of a broader discussion of the role of denotational semantics in IR modelling. Finally, the relationship between the *paradoxes of material implication* and the (in)adequacy of classical logic for IR modelling purposes is discussed.

1 Information Retrieval and modelling

In recent years, researchers in Information Retrieval (IR) have devoted an increasing amount of work to the design of *models* of IR, i.e., of theoretical descriptions of the IR task that could serve both as specifications for building running systems, and as theoretical tools for abstractly investigating the relative effectiveness of systems built along their guidelines.

Modelling is fundamentally an activity of *abstraction*. A model is a description of a system that concentrates on the most important, architectural features of the system, and leaves out details that are believed not to be fundamental to the understanding of how the system works. For instance, a model of an IR system might contain a description of how documents and information needs are represented within the system, but leave out details on the data structures used to store these representations. Of course, more or less details can be left out from the description, depending on the purpose of the model. Different levels of abstraction can then be envisaged, and different models of the same system can then be produced, each at a different level of abstraction.

*This work has been carried out in the context of the project FERMI 8134 - "Formalisation and Experimentation in the Retrieval of Multimedia Information", funded by the Commission of the European Communities under the ESPRIT Basic Research scheme.

Abstraction results in *generalisation* too, following the familiar principle according to which an abstract (or partial) description is equivalent to the class of concrete (or total) descriptions consistent with it: by leaving out non-fundamental details, the description actually becomes a model of a whole class of systems, namely those which differ from each other only by these details. A taxonomy of models can then be built that allows to classify IR systems and highlight their common premises and fundamental differences.

Traditionally, the activity of modelling is ascribed to two fundamental classes of motivations. Motivations of a *descriptive* nature reveal an intent to simply *acknowledge* "as is" the previous work of system builders, and explain the characteristics of a class of systems without being hampered by unimportant details, thus achieving clarity of exposition in the description of these systems. Motivations of a *prescriptive* (or *normative*) nature, that may be regarded as the primary driving force behind the activity of IR modelling, reveal instead an intent to *steer*, or influence, the work of system builders, dictating (or proposing) what the characteristics of a class of systems ought to be.

In the IR case (as in other subdisciplines of computer science), a third class of motivations that somehow escapes the previous, traditional classification, has recently gained prominence: motivations of a *predictive* nature reveal an intent to predict the behaviour of a real system by running "abstract experiments" on an artificial, simplified setting. The properties of the fundamentals of a system, or class of systems, can then be analytically tested: the theorist can thus be sure that the results so obtained are not influenced by the (supposedly non-fundamental) features that have been left out of the model.

Whatever the nature of the motivations for building models, however, it seems clear that, in order to achieve the breakthrough in effectiveness that nowadays applications demand, a better understanding is needed of IR and of the nature of information, and that such an understanding may be achieved only by abstracting away from particular systems and techniques, thus concentrating on the study of the core, foundational principles underlying the IR endeavour. The recent interest that researchers have shown in the logical approach to IR finds its roots in the belief that formal logic (and its companion discipline, formal semantics) is the discipline that provides the right tools for studying these foundational principles.

Surely, IR is not the first discipline within computer science in resorting to logic and methods of formal analysis. Researchers in other subfields of computer science, such as databases or programming languages, have gained much deeper insights in their discipline by analyzing it by means of logical techniques. In doing so, they have often uncovered deep-seated assumptions that had previously gone undetected, or inconsistencies that undermined the very foundations of the whole discipline, or problems previously unknown. Above all, they have gained a new perspective from which to look at their discipline, a perspective from which they have been able to take advantage of razor-sharp analytical tools whose effectiveness has been widely recognised by the computer science community as a whole.

It might even be claimed that some disciplines, such as Artificial Intelligence (AI), have been totally revolutionised by the advent of a "logic-oriented" mentality. By the late 70's, AI researchers had grown increasingly aware of the need for the methodological tools provided by mathematical logic for the construction of AI models of the domains of interest (see e.g., (Israel, 1985; McDermott, 1978)). Since then, the application of these methodologies has shaken the foundations of AI, and the theories that have withstood the test of logic have gained, as a consequence, strength and wider applicative impact, also arising interest in practitioners and theorists of other subfields of computer science (e.g., database theory,

distributed systems) or of science as a whole (e.g. economic theory). A representative example of this is offered by AI "frame-based knowledge representation languages". In the late 70's there existed a plethora of them, each based on its own custom primitives, each basically incomprehensible to anyone apart from its own designers, who nonetheless remained adamant on the fact that their languages could express what logic could not. The final insuccess of these languages prompted the very researchers that had first proposed them to recast them in terms of mathematical logic, resulting in what are now called *terminological* (or *description*) *logics* (TLs—see e.g., (Levesque & Brachman, 1987)). These have turned out to be one of the success stories of AI, to the point that AT&T relies on their inference engine to process billions of dollars' worth of customers orders (McGuinness & Alperin Resnick, 1995), and that they are now popular among computer scientists as diverse as database language designers (De Giacomo & Lenzerini, 1995) and factory-floor software engineers (Devanbu & Jones, 1994).

Although the history and underlying motivations of IR are profoundly different from those of AI, we think that IR too can benefit from the application of logical tools to the design and evaluation of its models, and our belief is indeed one of the fundamental premises that underlie our past work (Meghini et al., 1993; Sebastiani, 1994). The methodology in the use of logic that has been developed in AI and analytic philosophy is relevant also to IR applications, since logical tools are not committed to a specific domain to be modelled: documents and information needs constitute indeed one of their possible domains. Establishing logic as a common language between IR theorists and theorists of other subfields of computer science can also help IR in fostering cross-fertilisation with these other disciplines.

2 Incrementality and "modelness"

Information Retrieval, having been around almost since the invention of digital computers, has grown into a greatly complex discipline. A wealth of phenomena on which the relevance of documents to user information needs depends, have been highlighted by researchers, and a corresponding wealth of tools have been developed in the attempt to cope with them. Building realistic, worthwhile logical models of IR thus means modelling all these different phenomena, which makes for an extremely severe task.

In order to make worthwhile progress in this direction, we think that a logical model should be built *incrementally*. According to this line of thought, a relevant number of simplifying assumptions on what IR is should be made in the beginning, so as to allow the provisional design of a simple model of (an "idealised" version of) IR along methodological guidelines previously set. These simplifying assumptions should then be relaxed one by one, leading to the design of successive, more complex models that deal with a more and more complex and realistic picture of IR. Of course, care must be taken that the chosen methodology does allow the relaxation of these assumptions.

This way of proceeding places then more importance on the requirement that a first model of IR being preliminarily designed satisfy well-formedness adequacy criteria well accepted in the field of formal modelling, than on the requirement that the coverage of the model be comprehensive, i.e., that the model deals with all the wealth of phenomena which are of interest to IR. In some sense, we are saying that logical models of IR can be judged by two different yardsticks: "*modelness*" (or "*logical well-formedness*") and *coverage*.

An analysis of the logical models of IR proposed so far in the literature reveals, in our opinion, that coverage has been the primary goal to attain for most researchers who have

engaged in IR modelling through logic, and that most of the models designed so far hardly satisfy the well-formedness criteria that are standard in applied logic. Instead, we think it is much more fruitful to first design a logically well-formed, albeit simple, model of IR and incrementally make it more respondent to the complex reality of IR, than to start with a logically flawed, although comprehensive, model of IR and incrementally eliminate the (possibly inherent) flaws. This position is not only ours, but underlies practically all work in applied logic. For instance, the problem of modelling logically the diagnosis of faulty circuits has first been solved for circuits with single faults (Reiter, 1987), and only later the problem of multiple faults has been tackled (de Kleer & Williams, 1987). This and other experiences in applied logic show in our opinion that modelness is to be considered a *sine qua non* condition from the start, and coverage a goal to attain incrementally. The purpose of this paper is then to take modelness issues at heart, and discuss methodologies and criteria for the design of logical models of IR. This will be the topic of Sections 3 to 6.

3 The logical model of Information Retrieval

Although interest in logic on the part of IR researchers may be traced back at least to the early 70's (see e.g., (Cooper, 1971)), the first clear statement that IR should be understood in logical terms is, to our knowledge, due to van Rijsbergen[1]. In (van Rijsbergen, 1986) and in a number of subsequent papers (van Rijsbergen, 1989, 1992), van Rijsbergen proposed that the retrieval of document d as a result of information need n should be seen in terms of the logical formula $P(d \rightarrow n)$, where "$\rightarrow$" is the conditional connective formalised by a logic to be chosen and where $P(\alpha)$ is to be read as "the probability of α"[2]. Accordingly, the central problem of this way of looking at IR becomes that of selecting the "right" implication connective, i.e. selecting the logic whose implication connective best mirrors relevance and all the factors that influence it. The "ideal" logic should be the one in which $P(d \rightarrow n)$ equals the probability that the document represented by d is relevant to the information need represented by n. To this respect, van Rijsbergen also argued that material implication (i.e., the implication connective formalised by classical logic) is not adequate for this task. He speculated that the answer might instead lie in the brand of implication formalised by some non-classical logics, also suggesting that the **C2** logic, from the tradition of "conditional" logics (Stalnaker, 1968; Stalnaker & Thomason, 1970), might be promising in this respect.

In a "$P(d \rightarrow n)$" model, the role of probability is a key one, as logical formulae are inherently imperfect representations of documents and information needs; the relevance of a document to an information need can thus be established only up to a limited degree of certainty. In the discussion that follows, however, we will ignore this issue, and make the

[1]Darlington (1969) has been suggested by some to be the first researcher to point out that IR could be conceived in terms of logic. However, an examination of his work shows that the notion of IR that he takes into account is more akin to "data retrieval" as in database systems, and he gives no suggestion as to how the ideas presented could be extended so as to take into account information retrieval as we understand it. For this reason, we will not consider Darlington's work as relevant to our purposes; the same will happen, essentially for the same reasons, with the work of Patel-Schneider et al. (1984).

[2]In this paper we are taking a slight detour from the terminology used in van Rijsbergen's papers, as we will take *queries* to be *representations of information needs*. Information needs and documents are the "concrete entities" of our domain, while what we actually deal with in logic is their *representations*; "noise" is introduced in the translation from information needs to their representations (queries) and from documents to their representations. As a consequence, the rightmost symbol in the implication is for us not "the representation of a query", but "the representation of an information need" (or "a query").

simplifying assumption that "perfect" representations of documents and information needs can indeed be obtained, and that they are logical formulae of a simple non-probabilistic (e.g., propositional) language. This is an obviously unreasonable assumption once one attempts to build a *realistic* model of IR. It is however an assumption that will momentarily allow us to concentrate on discussing what characteristics a "good" formal model of IR should have by drawing our examples from a simple, non probabilistic logic. This is in keeping with the policy exposed in Section 2, as the very first "modelness" issues we need to clarify are those affecting very simple logical models of IR; only after these have been agreed upon may we tackle those concerning more complex ones. As a matter of fact, all the claims we make in this paper for the non-probabilistic case apply equally well once we plug in probability.

van Rijsbergen's proposal is fascinating for three quite distinct reasons. The first, more obvious reason is is that this model provides a bridge with database research, as the logical, proof-theoretic model of databases (Reiter, 1984) precisely establishes that a tuple t is to be returned as a result of query q iff $t \rightarrow q$[3].

The second reason is apparent once we consider, instead of the proof-theoretic, "symbol-crunching" level of logic, its model-theoretic, semantic level. In terms of this latter, the logical approach to IR amounts to sanctioning that relevance coincides with (set-)inclusion of information content, or semantics: only documents whose information content includes that of the information need are to be retrieved. Building effective IR models means then designing formal theories of the semantic (or information) content of documents, possibly taking into account context-dependence and other situational aspects of information (van Rijsbergen & Lalmas, 1996).

The third reason springs from the existence of the proof-theoretic aspect of logic, and is related to the predictive intent of the modelling task discussed in Section 1. The adoption of the logical model means that, once the document and the information need are given a representation, "abstract experiments" can be performed by running a theorem prover for the logic in question and establishing whether $d \rightarrow n$. The logical model of IR is then an *effective* model (in the sense of computability theory). The idea of "running abstract experiments" is especially interesting for IR, as it might produce good insights into phenomena and techniques of IR that are not yet well understood. In fact, experimental methodologies of the standard kind, while allowing to determine that certain techniques work better than others on a given experimental setting, do not always help in understanding the reasons that lie at the root of such behaviour. A logical model, instead, might prove a better "experimental" tool in this respect. We might thus envisage, for instance, coding a particular IR technique in a given logic; once the fact that this is a faithful encoding of the technique has been established (e.g., by cross-checking the results of theorem proving, or proof checking, with those obtained experimentally), the semantics of the logical language would give insight into how and why such results are obtained, and give directions as to how the technique should be implemented in real systems. If two techniques, for some yet unknown reasons, do not fit together well in real IR systems, they could both be encoded in the model, and the reasons for their interaction could be investigated by inspecting the semantics of the language.

[3]This approach may more precisely be classified along the line of the "hybrid approaches" hinted at in Footnote 11.

4 On the status of $d \rightarrow n$

van Rijsbergen's proposal has proven very influential, to the extent that a number of papers have appeared in the literature tackling, in a variety of ways, the problem of modelling IR within the above-mentioned paradigm. However, a comparative analysis of these papers reveals that it is far from clear what the logical status of the $d \rightarrow n$ formula should be, in order to indicate that the document represented by d is relevant to the information need represented by n. In particular, authors seem to take different stands on which among the following facts should indicate this:

1. $d \rightarrow n$ is *true* in some particular interpretation of the chosen logic $\mathcal{L}$;
2. d is a *logical consequence* of n in $\mathcal{L}$;
3. $d \rightarrow n$ is *valid* in $\mathcal{L}$;
4. n is *derivable* (or *provable*) from d in $\mathcal{L}$;
5. $d \rightarrow n$ is a *theorem* of $\mathcal{L}$.

These notions mean substantially different things in logic, and it should be apparent that any *logical* model of IR (or of any other real-world phenomenon, for that matter) should take a clear stand on this issue if "modelness" is to be a key concern[4]. This variety of different approaches opens up questions such as: "Is there one 'right' notion, among the ones above, on which a logical model of IR should be based?"; "Is there more than one?"; "Are there 'wrong' ones?". And a problem incumbent in the background is, of course, that if different notions are employed, *comparison* of these models may be difficult to achieve. Bruza and Huibers (1994) have quite appropriately expressed their concern on this:

> Research has not yet produced a powerful enough framework whereby information retrieval systems can be compared *inductively* instead of *experimentally*. A breakthrough in this area would mean that a theorem could be proven stating, for example, that vector space retrieval is more effective than Boolean retrieval. Such a result would not only spare us the efforts of experimentation, but more importantly, it would allow us to side step the controversies surrounding the experimental process.

Let us then try to analyse the different stands which have been taken. In (van Rijsbergen, 1989), van Rijsbergen seems to indicate that Option 4 (derivability) is the relevant notion to use:

[4] We recall that: 1) a formula α is a *logical consequence* of a set of formulae Γ when α is true in all the interpretations in which all the formulae in Γ are true; 2) a formula α is *valid* when it is true in all interpretations, i.e., when it is a logical consequence of the empty set; 3) a formula α is *derivable* (or *provable*) from a set of formulae Γ when α can be obtained by applying the rules of inference to the axioms of the logic and the formulae in Γ; 4) a formula α is a *theorem* when it can be obtained by applying the rules of inference to the axioms of the logic, i.e., when it is derivable from the empty set. While "derivability" and "theoremhood" are *syntactic* (or *proof-theoretic*) notions "truth", "validity" and "logical consequentiality" are *semantic* (or *model-theoretic*) notions. When the syntactic apparatus (i.e., the set of axioms and inference rules) of a logic is *weakly sound and complete* with respect to its semantics, theoremhood coincides with validity; when it is *strongly sound and complete*, we also obtain equivalence between derivability and logical consequentiality.

> The proposal of this paper is that all retrieval be based on a well-defined inference mechanism. This requires that objects and queries be given a formal semantics and that retrieval is expressed as a proof.

although also Option 5 (theoremhood) is no doubt consistent with this formulation. Bruza and van der Gaag (1993) explicitly side for Option 4 (derivability):

> In this approach, an information object is deemed relevant with respect to a searcher's request if this request can be proven from the characterization of the object by employing a set of rules of *inference*.

Nie (1989) has Option 1 (truth in a particular interpretation) in mind when he says:

> For document D to be a "right" answer for query Q, it must "imply" the query, i.e., $D \rightarrow Q$. (...) The symbol $\rightarrow$ does not signify the "material implication" as in classical logics. A counterexample for the material implication is that an empty document cannot imply a nonempty query (...).

as none among the other four options would result, in classical logics, in the behaviour he describes.

In their review paper on logical models of IR, Chiaramella and Chevallet (1992) agree *de facto* with Nie in embracing Option 1 (truth in a particular interpretation), as their rebuttal of classical logic is based on the truth of material implication in a single interpretation:

> If we consider the third line of this truth table, saying that when a predicate A is false and B is true, then A implies B is true has certainly no intuitive meaning: algebraic considerations somehow overrode common sense interpretations. We have left the domain of reality for the domain of mathematical abstraction.

In their model based on Terminological Logics (TLs), Meghini et al. (1993) embrace instead Option 3 (validity)[5]:

> The terminological model then sees IR as the task of retrieving, as a response to a query C, all and only those documents i such that (**sing** i) $\preceq_\Omega C$, where Ω is a TL representation of the document base [and $\preceq_\Omega$ denotes *hybrid subsumption* between terms]. In other words, IR is the task of retrieving all those documents whose membership in the class denoted by C is a direct consequence of the truth of all the assertions and axioms of Ω.

An analysis of various other works concerning logical models of IR (such as (Wong & Yao, 1991; Müller & Thiel, 1994; Fuhr, 1995; Hunter, 1995)) confirms that there is a general lack of agreement on this issue.

What stand should one take then, given that we want to take "modelness" seriously and given that these options are in general not equivalent? For ease of discussion, we will make the

[5]In reference to this work, however, it should be noted that neither of the five notions discussed above applies to TLs, as TLs deal with terms rather than formulae, while all these notions apply to *formulae* only. However, a closer look at TLs reveals that hybrid subsumption is the TL analogue of validity (on this see also Footnote 11).

simplifying assumption that our logic has a strongly sound and complete inferential apparatus. As we observed in Footnote 4, this will allow us to restrict our discussion to truth, validity and logical consequentiality; everything we will say about validity and logical consequentiality will thus also apply to theoremhood and derivability, respectively[6].

4.1 Truth

Truth is not a suitable notion on which to base a logical model of a real-world phenomenon. In the IR case, this may be seen by noting that, if we wanted to model the relevance of the document represented by d to the information need represented by n with the truth of $d \rightarrow n$, we should also specify *in which interpretation* the truth of $d \rightarrow n$ has to be evaluated. A formula cannot be true *tout court*, simply because truth is not a *property* of formulae, but a *binary relation* between formulae and interpretations. We might naively answer that we should take "the interpretation that corresponds to the real world" (i.e., the interpretation in which the sentence "Glasgow is in Scotland" evaluates to true, "Birds are mammals" evaluates to false, and so on). But what is then the truth value to which "The number of water molecules in my glass is even" should evaluate? In the "interpretation that corresponds to the real world" this sentence surely has a truth value, i.e., that number either is even or is not. The answer is that we do not have a clue to what the "interpretation that corresponds to the real world" is, because our knowledge of the real world (or, more to the point, of our domain of discourse) is *partial*, and often *fallacious* too, even in more mundane matters than those of molecular structure. And IR is no exception, as clearly argued by Lalmas and van Rijsbergen (1993):

> Partiality is an important feature of an IR system because it is common that it is unknown whether or not an item of information is contained in a document. So, assigning a truth-value to every existing formula that can be defined in the logic is meaningless.

An interpretation (i.e., "a way the world could be") is not something we can pick out and stipulate to correspond to the real world, simply because we do not have a "direct grasp" (i.e., total and infallible knowledge) of the real world. Henceforth, truth is not simply difficult or impossible to compute; it is just not an object of computation, and to speak of "computing truth" is a bit of an epistemological oxymoron. We take up again the topic of truth in Section 6.1.

4.2 Validity

This argument, we think, illustrates that the notion of truth is unsuitable for modelling substantial fragments of reality; indeed, this is the reason why, in devising a logic, a logician is usually not interested in specifying the notion of truth *per se*, but is rather interested in specifying truth as functional to the specification of the notions of validity and logical consequence[7].

[6] Quite obviously, when proposing a particular logical model for IR we should check whether this assumption holds or not in the particular logic we have chosen. In fact, among the logics that have been proposed as foundations for a model of IR, some do not enjoy this property (e.g., propositional modal logic, proposed in (Nie, 1989)).

[7] Many logicians take a logic *to be* the set of its valid formulae.

But how then could such a model be based on e.g., validity, given that the notion of validity is defined in terms of truth? The key observation is that valid formulae are true (in any interpretation) in virtue of their *form*, and not in virtue of their *content*. Because of this, we do not need to have a grasp on the real world (i.e., to know which interpretation corresponds to it) to assess the validity of a formula (in the IR case: to assess whether we should retrieve the document or not); we only need to perform a purely symbolic check of the formula itself. For instance, the formula of propositional logic

$$\texttt{John-likes-football} \vee \neg \texttt{John-likes-football} \tag{1}$$

is true in any interpretation (i.e., valid). In order to assess its validity it is not necessary to know whether in the interpretation that corresponds to the real world (or in any other particular interpretation, for that matter) John actually likes football or not; it is instead sufficient to apply the well-known syntactic rules for validity checking in propositional logic. This formula remains valid even if we substitute any other propositional formula to the two occurrences of `John-likes-football`, which indicates that content has really nothing to do with validity. The formula

$$\texttt{John-is-a-man} \rightarrow \texttt{John-likes-football} \tag{2}$$

instead, may well be true in the interpretation that corresponds to the real world, but is false in others, hence is logically not very interesting.

One might argue that (2) is at least an *informative* formula ("it says *something*"), and that valid formulae like (1), being tautologous, carry no information content. The notion of informativeness that is being hinted at here is the one by Carnap and Bar-Hillel (1953) (henceforth "CBH-informativeness"), according to which a formula is the more informative the more countermodels it has (i.e., the more interpretations falsify it)[8]; valid formulae are then minimally CBH-informative, as they have no countermodels. But in the IR case it is not the formula $d \rightarrow n$ that carries information to us; *it is the very fact that it is (or that it is not) valid*, as it informs us whether we should retrieve the document or not. For instance, if we cast the Boolean model of IR in terms of the validity of formula $d \rightarrow n$ in propositional logic, the fact that the formula $(p_1 \wedge p_2 \wedge p_3) \rightarrow (p_1 \vee p_2)$ is valid informs us that a document indexed by terms p_1, p_2 and p_3 should be retrieved as a result of the information need represented by $p_1 \vee p_2$; and the fact that the formula $(p_3 \wedge p_4 \wedge p_5) \rightarrow (p_1 \vee p_2)$ is not valid informs us that a document indexed by terms p_3, p_4 and p_5 should not be retrieved as a result of the same information need. In other words, we might say that although the valid propositional formula $(p_1 \wedge p_2 \wedge p_3) \rightarrow (p_1 \vee p_2)$ is CBH-uninformative, the meta-formula $VALID((p_1 \wedge p_2 \wedge p_3) \rightarrow (p_1 \vee p_2))$ (where $VALID$ is a meta-predicate symbol) is CBH-informative, as is also the meta-formula $\neg VALID((p_3 \wedge p_4 \wedge p_5) \rightarrow (p_1 \vee p_2))$.

This notion of "meta-level informativeness" is, of course, *task-oriented*, and the task here is IR. In fact, we are not interested in the validity or non-validity of formulae which are not of type $d \rightarrow n$; thus, a meta-formula such as $VALID(p_1 \vee \neg p_1)$ has no information content to us, as it informs us of the validity of a formula that is not meant to represent relevance of documents to information needs.

In sum, the interesting fact is that adopting a form-based notion as validity is, rather than a content-based notion as truth is, *allows us to effectively* (in the sense of computability theory) *reason about information content*.

[8]We would like to thank Gianni Amati for pointing this out to us.

4.2.1 Truth, validity and the "false document problem"

A further hint that validity, rather than truth, is a better way to go comes from the observation that, once we base a model of IR on validity, the model does not suffer from what has been called the *false document problem* (see e.g., (Chiaramella & Chevallet, 1992)). According to the supporters of Option 1 (truth in a particular interpretation), propositional logic, and classical logic in general, suffers from the problem that a so-called "false document" (i.e., a hypothetical document which is "about nothing"—we will rather call it a *totally uninformative document* (TUD)) is deemed relevant to any information need, and this is obviously unsuitable.

In fact, suppose our term language consists of the set of propositional letters $P = \{p_1, p_2, p_3, p_4\}$, so that a TUD is represented by the formula $\neg p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4$; quite obviously, in propositional logic the formula $(\neg p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4) \rightarrow \alpha$ is true for any propositional formula α *in at least one interpretation*, i.e., in the interpretation that makes all propositional letters in P true.

But this does not mean *per se* that propositional logic is unsuitable for modelling IR: it just means that a model based on truth in a particular interpretation is. In fact, it is easy to show that, if Option 3 (validity) is adopted, no "false document problem" obtains, and all and only those documents whose representation has a greater "information content" than the representation of the information need are retrieved. For those documents for which this is not the case (as in the case of a TUD), $d \rightarrow n$ will not be valid, i.e., there will be at least one assignment of truth values to the propositional letters for which $d \rightarrow n$ will be false. For these documents, it may well be that, among the assignments that make $d \rightarrow n$ true, there is some assignment that makes d false and n true, and hence makes the implication $d \rightarrow n$ *true*; but this need not worry us, as long as the implication is not *valid*. This situation will be better illustrated by means of an example.

Example 1 *Let us define a model of IR based on the validity of* $d \rightarrow n$ *in propositional logic, in the following way. Let us suppose each document is represented by a conjunction of literals (a* literal *being either a propositional letter or its negation) drawn from an alphabet* $P = \{p_1, p_2, p_3, p_4\}$*. Relying on propositional logic allows us to express the following facts:*

1. *to explicitly say that d is about* p_i*; this is obtained by including* p_i *as a conjunct in d's representation;*
2. *to explicitly say that d is not about* p_i*; this is obtained by including* $\neg p_i$ *as a conjunct in d's representation;*
3. *to take no commitment as to whether d is or is not about* p_i*; this is obtained by including neither* p_i *nor* $\neg p_i$ *as conjuncts of d's representation*[9]*.*

Note that, in this representation, a TUD is represented by the formula $\neg p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4$*.*

Let us also suppose that each information need is represented by a formula of propositional logic built out of alphabet P. The three types of facts that can be stated in representing

[9]In many approaches to Boolean retrieval, including neither p_i nor $\neg p_i$ as conjuncts of d's representation is taken to mean that d is not about p_i. This is called the *closed world assumption* (CWA); our arguments would follow similar lines even in this case, as the effect of the CWA is to say implicitly what here is specified explicitly.

	p_1	p_2	p_3	p_4	d_1	d_2	d_3	n	$d_1 \rightarrow n$	$d_2 \rightarrow n$	$d_3 \rightarrow n$
1	T	T	T	T	F	F	F	T	T	T	T
2	T	T	T	F	F	F	T	T	T	T	T
3	T	T	F	T	F	F	F	T	T	T	T
4	T	T	F	F	T	F	T	T	T	T	T
5	T	F	T	T	F	F	F	T	T	T	T
6	T	F	T	F	F	F	F	T	T	T	T
7	T	F	F	T	F	F	F	T	T	T	T
8	T	F	F	F	F	F	F	T	T	T	T
9	F	T	T	T	F	F	F	F	T	T	T
10	F	T	T	F	F	F	F	F	T	T	T
11	F	T	F	T	F	F	F	T	T	T	T
12	F	T	F	F	F	F	F	T	T	T	T
13	F	F	T	T	F	F	F	F	T	T	T
14	F	F	T	F	F	F	F	F	T	T	T
15	F	F	F	T	F	F	F	F	T	T	T
16	F	F	F	F	F	T	F	F	T	F	T

Figure 1: Validity-based relevance assessments

documents can also be stated in representing information needs: for instance, if we represent an information need by formula $p_1 \vee \neg p_2$, we mean that the system should retrieve all and only those documents that are either about p_1 or are not about p_2, regardless of whether they are also about p_3 and/or p_4. Figure 1 shows validity-based relevance assessments for the three sample documents

$$\begin{aligned} d_1 &\equiv p_1 \wedge p_2 \wedge \neg p_3 \wedge \neg p_4 \\ d_2 &\equiv \neg p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4 \\ d_3 &\equiv p_1 \wedge p_2 \wedge \neg p_4 \end{aligned}$$

with respect to information need

$$n \equiv p_1 \vee (p_2 \wedge \neg p_3)$$

Note that document d_2 is a TUD. As can be gathered from inspection of Figure 1, formulae $d_1 \rightarrow n$ and $d_3 \rightarrow n$ are valid (i.e., the corresponding columns have all T's); this means that the documents represented by d_1 and d_3 are deemed relevant to the information need represented by n, as should indeed happen. The formula $d_2 \rightarrow n$, instead, is not valid (i.e., there is an F at row 16 in the corresponding column); this means that, contrary to what might happen in truth-based approaches to relevance, the TUD represented by d_2 is correctly deemed not relevant to the information need represented by n.

It might be interesting to note that there are indeed formulae representing information needs to which a TUD would be deemed relevant: these are precisely all conjunctions of negated propositional letters that can be built by using alphabet P (and all formulae logically equivalent to them). For instance, the TUD represented by d_2 would be deemed relevant to the information need represented by $\neg p_1 \wedge \neg p_2 \wedge \neg p_3$. But this information need is at least as "pathological" as a TUD is a "pathological" document: and it is perfectly reasonable that a "document about nothing" should be deemed relevant to an "information need about nothing".

Note also that this argument, apart from accumulating further evidence against Option 1 (truth in a particular interpretation) as a suitable notion on which to base a logical model of

IR, has the practical consequence of invalidating a number of arguments (see the introduction to Section 4) that were made against the suitability of classical logic for IR modelling purposes, and that were essentially based on a truth-based view of IR modelling. We will come back to this point in Section 5.

4.3 Logical consequentiality

Let us now discuss the case of logical consequentiality. The first thing we should point out is that in Example 1 the same results we have obtained with validity would have been obtained by adopting logical consequentiality. This is due to the use of propositional logic as the basis of the simple model adopted in the example: propositional logic is in fact strongly sound and complete with respect to its standard denotational semantics, and this ensures (see Footnote 4) that, for any formulae α and β, $\alpha \rightarrow \beta$ is valid if and only if β is a logical consequence of α.

However, given that we are seeking a general (i.e., independent of the chosen logic) characterisation of what a logical model of IR should consist of, it is necessary to assess which is the notion of choice in case the two are not equivalent.

To a first approximation both are by and large suitable, as both are "form-based" notions, rather than "content-based" as truth is: while truth may be checked only with reference to a specific interpretation, validity and logical consequentiality may be established *formally* (and thus effectively), by symbolic manipulation only. These two notions are the cornerstones of logic exactly because logic is concerned with *partial and possibly fallacious* knowledge of the world, and both notions specify exactly those inferences which are compatible with this knowledge, however partial and fallacious it may be. For instance, in Example 1 we would like document d_3 to be deemed relevant to information need n: by relying on e.g., logical consequence this indeed happens, because n is compatible with the partial knowledge we have of document d_3 (partial as it does not say anything on d_3 being about p_3 or not) that is expressed by the formula $p_1 \wedge p_2 \wedge \neg p_4$.

Forgetting about truth, let us now try to assess the relative merits of validity and logical consequentiality. One aspect on which validity scores better is that logical consequentiality would seem a bit unsuitable given that what we are really interested in is extending the "$d \rightarrow n$ model" to a "$P(d \rightarrow n)$ model" (see Section 3), i.e., to a model in which we can speak of the *probability* of relevance in terms of a formula $P(d \rightarrow n) = r$ to be read "the probability that d implies n is r". Given a suitable logic (e.g., the one discussed in (Halpern, 1990)), we can express formulae such as $P(d \rightarrow n) = r$, and define appropriate notions of truth and validity for them. However, in no logic that we know of (although it might perhaps in principle be possible) we can formalise the sentence "the probability that n is a logical consequence of d is r", and define (meta-level) truth and validity conditions for it; in fact, in any logic that we know of either n is a logical consequence of d or it is not (i.e., the probability of this fact, to be expressed in the metalanguage, would always be either 1 or 0)[10], and we can hardly imagine a reasonable set of events on which such a probability distribution might be defined.

[10]Bruza (1993, page 24) recognises this difficulty when, commenting on (Cooper, 1971), he says: "Defining relevance in terms of logical consequence implies that an object is (logically) relevant with respect to an information need, or it is not. Therefore, Cooper did not view relevance as being "grey". According to Cooper, the greyness is a product of a matching algorithm which does not fully complete the inferential process.".

One aspect on which logical consequentiality gets instead a better mark is its more intuitive character. It is quite intuitive, in fact, that relevance of a document to an information need is a *consequence* of the semantic content of the document and of the information need, and possibly of other factors such as the meaning (as specified e.g., in a thesaurus) of the terms involved; and it is quite intuitive that relevance should not depend on the fact that the particular logic adopted has an implication connective in its language[11].

To sum up our argument, we may then say that at a first approximation both validity and logical consequentiality are suitable (whereas truth is not), and which is best is an issue open to debate, as each of them has advantages and disadvantages for IR modelling[12].

5 IR and the paradoxes of material implication

Since the very introduction of the logical model of IR, researchers seem to have maintained as a cornerstone of their investigations that classical logic is not adequate for IR modelling. However, in Section 4 we have seen that a number of arguments that had been used against employing classical logic for IR modelling are substantially invalid, as they are based on a misuse of logic in modelling real-world phenomena. It is our contention that also other arguments that have tried to counter classical logic are far from conclusive, and that we should perhaps come to the conclusion that the use of classical logic in IR has been dismissed too hastily.

Note that classical logic has been criticised on two accounts[13]:

1. an *extension* of classical logic is needed, because as it stands it can deal neither with (a) the *uncertainty*, or *probability*, of relevance, nor with (b) the *partiality*, or *non-binary character*, of relevance, which must be dealt with because (a) representations of documents and information needs are inherently imperfect, and because (b) (user) relevance can come in degrees;

2. a *deviation* of classical logic is needed, because as it stands (a) even if perfect representations of documents and information needs could be obtained and were formulae of classical logic, and (b) even if there existed no such a thing as partial relevance, it would not be adequate.

It is this latter criticism we do not completely agree with; one argument that has been brought to its support is the following.

A number of researchers have recalled that material implication suffers from idiosyncratic behaviour, resulting in what are known as "the paradoxes of material implication" (see e.g., (de Swart & Nederpelt, 1992) and (Haack, 1978, Chapter 3)), and have implied that

[11]Note that a seemingly "hybrid" approach is also possible: one might require that $K \models d \rightarrow q$, where K represents a body of background knowledge, possibly including characterisations of d and/or n. Although this approach involves both validity and logical consequentiality, validity is obviously the really relevant notion here.

[12]Not surprisingly, in the logical approach to AI either approach is adopted from time to time, with a tendency to adopt logical consequentiality in the case of classical logics and validity in the case of those "esoteric" logics which do not enjoy strong soundness and completeness (Levesque, 1984).

[13]We recall that, given a logic $\mathcal{L} = \langle \mathcal{A}, \mathcal{V} \rangle$ defined on alphabet $\mathcal{A}$ and with a set of valid formulae $\mathcal{V}$: (a) an *extension* $\mathcal{L}'$ of $\mathcal{L}$ is a logic $\langle \mathcal{A}', \mathcal{V}' \rangle$ such that $\mathcal{A} \subset \mathcal{A}'$ and $\mathcal{V} \subset \mathcal{V}'$; (b) a *deviation* $\mathcal{L}'$ of $\mathcal{L}$ is a logic $\langle \mathcal{A}', \mathcal{V}' \rangle$ such that $\mathcal{A} = \mathcal{A}'$ and $\mathcal{V} \neq \mathcal{V}'$. See (Haack, 1978, Chapter 9) for a more detailed discussion on this.

this renders material implication an unsuitable starting point in the attempt to model relevance of documents to information needs[14]. For instance, it is well-known that the following schemata[15], all of which have a distinctively counter-intuitive flavour once "$\rightarrow$" is interpreted as "if ... then ...", are valid in classical propositional logic:

$$\alpha \rightarrow (\beta \rightarrow \alpha) \tag{3}$$

$$\neg\alpha \rightarrow (\alpha \rightarrow \beta) \tag{4}$$

$$(\alpha \rightarrow \beta) \vee (\beta \rightarrow \alpha) \tag{5}$$

However, if relevance of documents to information needs is to be modelled by the validity of formula $d \rightarrow n$, our interest as IR theorists in an implication connective is exclusively in terms of its behaviour in the context of a formula $d \rightarrow n$ in which *neither d nor n contain occurrences of the "$\rightarrow$" symbol*. Of course, the representation of an information need may well be a formula of type $\neg p_1 \vee p_2$, which in classical logic is equivalent to $p_1 \rightarrow p_2$. But this need not concern us, as we do not understand *this occurrence of* "$\rightarrow$" as modelling relevance; in fact, the counter-intuitive character of schemata (3)÷(5) critically depends on the interpretation of *all* the occurrences of "$\rightarrow$" as "if ... then ...". For instance, if we equivalently write (4) as $\gamma \rightarrow (\gamma \vee \beta)$ (where $\gamma = \neg\alpha$), this says something which, far from being paradoxical, can be absolutely subscribed to, namely that a document about γ is to be retrieved as a consequence of an information need about γ or something else.

Hence it would seem that the only "paradoxes" that may be of interest to us are those in which "$\rightarrow$" occurs at a level of nesting equal to 1, and that we can instead disregard the others; and, to the credit of classical logic, no "paradoxes" of nesting equal to 1 have ever been pointed out (unless either the consequent or the antecedent are themselves either valid or contradictory, a case which we discuss later).

Actually, that we can disregard schemata (3) and (4) is true only in part. In fact in classical propositional logic, as in any logic in which *modus ponens* is a rule of inference[16], whenever $\alpha \rightarrow \beta$ is valid, if α is valid also β is valid. This means that schemata (3) and (4) *are* of interest to us, because they reflect the behaviour of the logic once either the information need and/or the document are *themselves* represented by valid or contradictory formulae; luckily enough, these cases turn out to be harmless, as the following discussion shows.

In IR terms, Schema (3) is pertinent to the case of information needs represented by valid formulae. In fact, from (3) and *modus ponens* we have that if n is a valid formula then $d \rightarrow n$ is also valid: in IR terms this means that *any* document will be deemed relevant to an information need represented by a valid formula. But note that such an information need is represented by formulae such as $p_1 \vee \neg p_1$ (or its logical equivalents), which corresponds to a request to retrieve "all documents that either are about p_1 or are not about p_1": it is then perfectly reasonable that all documents from the document base should be selected for retrieval, as a query such as $p_1 \vee \neg p_1$ asks exactly for this. It is also

[14] Actually, none among the papers quoted in this article mentions the "paradoxes" of material implication explicitly; it is however clear that those are the phenomena referred to. We write the word "paradox" in quotes because, in the case we are discussing, it is actually a misnomer: the "paradoxes of material implication" are not cases of *inconsistent* behaviour (as e.g., in the paradox of the Liar), but rather of *counter-intuitive* behaviour.

[15] A *schema* is an expression which stands for the class of formulae of the logic that can be obtained by uniformly substituting formulae of the logic for metavariables; for instance, the schema of propositional logic $\alpha \vee \neg\alpha$ stands for the set of formulae of propositional logic containing e.g., the formula $p \vee \neg p$ and the formula $(p \rightarrow q) \vee \neg(p \rightarrow q)$.

[16] Not all logics have this property: see the discussion on "the admissibility of γ" in (Dunn, 1986).

clear from this example that valid formulae are representations of "pathological" information needs: any normal information need will be represented by a formula that is neither valid nor unsatisfiable, but true in some proper subset of the set of interpretations of propositional logic.

Schema (4) is instead pertinent to the case of documents represented by unsatisfiable formulae. In fact, from (4) and from *modus ponens* we have that, if $\neg d$ is a valid formula (i.e., if d is unsatisfiable), then $d \rightarrow n$ is also valid: in IR terms this means that a document represented by an unsatisfiable formula will be deemed relevant to *any* information need. But note that such a document is represented by formulae such as $p_1 \wedge \neg p_1$ (or their logical equivalents), which asserts that the document "at the same time is about p_1 and is not about p_1". It is also clear from this example that unsatisfiable formulae are representations of "pathological" documents: as for information needs, any "normal" document will be represented by a formula that is neither valid nor unsatisfiable, but satisfied in some subset of the set of interpretations of propositional logic[17].

In sum, only a few among the paradoxes of material implication are pertinent to IR modelling, depending on the level of nesting of the "$\rightarrow$" connective; we have shown that the few pertinent ones do not rule out the use of classical propositional logic for IR modelling, as 1) they affect the behaviour of the model only when "pathological" documents and information needs are considered, and 2) even in these cases they arguably do not affect it in an unreasonable way.

6 The role of denotational semantics in IR modelling

In Section 4 we have argued that either validity or logical consequentiality should be the central notions in any logical model of IR. The denotational semantics of the representation language used in the models under scrutiny has played a central role in our argument[18]. In fact, it is by analysing the semantics (i.e., the truth conditions) of the formulae involved in these models that we have been able to fully understand the different impact on IR models of the five notions discussed in the previous sections. Had we relied on syntactic and proof-theoretic notions only, we would have been able to perform this analysis only with greater difficulty.

Denotational semantics is extremely important in logical analysis, as it provides a frame of reference for analysing the meaning of the logical machinery used to formalise the application domain of interest. This frame of reference is so outstanding in terms of clarity and simplicity that it has gained the status of a "standard", both within the logic community (Haack, 1978) and the programming languages community (Gordon, 1979); it thus provides a means of analysing, comparing, designing (and sometimes implementing too) logical and programming

[17] Straccia (1996) proposes a "four-valued" terminological logic for IR (further elaborated by Meghini and Straccia (1996)) which is meant to overcome the paradoxes of material implication. We think his approach is justified by the fact that, without recourse to the four-valued semantics, terminological logics would allow creating document representations that are individually satisfiable but unsatisfiable when taken together.

[18] We recall that *denotational semantics* (also known as *model-theoretic* or *Tarskian semantics*) is the standard way of formally specifying the meaning of logical languages. Such a specification is accomplished by postulating the existence of a number of "ways the world could be" (*interpretations*), and of systematically specifying in which of these interpretations the expressions of the language are true. Inference is then defined as the derivation of only those formulae that are true in all the interpretations in which the premises are also true. In the case of propositional logic, discussed in the previous sections, interpretations are usually called *truth-value assignments*.

languages that constitutes a *de facto* "interlingua" among designer, implementor, user and critic. Its importance in the analysis of logical models of IR has been recognised ever since the very first logical model of IR was proposed: van Rijsbergen (1986) makes an explicit reference to "possible worlds semantics" (PWS), a well-known type of denotational semantics, and most proposed logical models of IR appeared since then (e.g., (Crestani & van Rijsbergen, 1995; Meghini, 1995; Meghini et al., 1993; Nie, 1989; Sebastiani, 1994) rely heavily on notions from denotational semantics.

However, an analysis of this latter literature reveals that there seems to be no agreement among these researchers on the role that denotational semantics should play in IR modelling; we might go as far as saying that there seems to be no agreement as to what denotational semantics really is. Given that its importance also lies in its being a "standard", it is very important that consensus can be reached among IR theorists on how to use it.

To this respect we think that while denotational semantics is, as we said above, a frame of reference for analysing the language of a given logic (i.e., the meaning of the connectives, operators and of the other primitives of the language), it is by no means part of this language itself. The entities that populate the world of denotational semantics (e.g., possible worlds, individuals, accessibility relations, etc.) are nothing else than (immaterial) "ideas" with reference to which one may explain, or support one's intuitions about, the meaning of the (absolutely material) data structures (i.e., formulae) and inference rules of the language under consideration. For instance, the notion of "possible world" and that of "accessibility between possible worlds" are useful clues to understanding the nature of the necessity operator of modal logic or the implication connective of conditional logic. However, it is of key importance to recognise that these entities are not themselves data structures open to direct manipulation. The idea that they indeed are, and that they can thus have an active part in the IR process, lurks behind a number of papers dealing with the logical modelling of IR (e.g., (Crestani & van Rijsbergen, 1995; Nie, 1989)). In particular, the fact that the semantics of modal and conditional logics relies on "possible worlds" being grouped into graph-like structures (called *Kripke structures*) has led to the definition of models in which the theorists themselves build these networks by direct manipulation (e.g., by assigning weights to the nodes and to the "accessibility" edges of the Kripke structure). Denotational semantics stipulates instead that the characteristics of semantic structures (e.g., the weights mentioned above) are to be determined only by the logical formulae that appear in one's representation of the relevant knowledge, and are not themselves open to direct manipulation. The only way to make these structures have certain given characteristics is to *induce* these characteristics by introducing, in one's representation, formulae that constrain the semantic structures to have exactly those characteristics[19].

This argument is not meant to invalidate those models that do not comply with this criterion. It is meant instead to suggest that by direct manipulation of the semantic structures underlying a given logic one may well gain an interesting metaphor, but loses touch with inference (and, arguably, with logic), as this latter is obtained only by the indirect (rather than direct) manipulation of all (rather than one particular) semantic structures consistent with one's data.

[19]For example, in so-called "normal" modal logics (i.e., supersets of the $\mathcal{K}$ modal logic) reflexivity of Kripke structures is achieved not by adding self-loops to all nodes of these structures, but is induced by introducing in one's representation all formulae of type $L\alpha \rightarrow \alpha$.

6.1 Total knowledge, truth-based models and denotational semantics

It may be interesting to understand why models of IR that do not comply with this criterion have been so popular. We think that the reason for this is also the reason for the popularity of the truth-based models discussed in Section 4.1, and the critical issue here turns out to be, again, partiality. In fact, if one adopts a logic $\mathcal{L}$ in order to model a given domain of discourse (i.e., to encode what one knows of this domain of discourse) and this model is a finite set S of sentences of $\mathcal{L}$, the semantics of S is the set I of all the interpretations of $\mathcal{L}$ that satisfy S. This means that two interpretations in I agree on all the facts specified by S, but may disagree on the facts on which S remains uncommitted; and there usually are many such facts, as our knowledge of a given domain of discourse is, as argued in Section 4.1, usually *partial*. However, if we suppose instead that our knowledge of the domain of discourse is total, the set of interpretations contains a unique element i, as there are no facts on which S remains uncommitted and that may thus give rise to different interpretations. In logic, a set of sentences such as S is called a *complete theory*. If a formula α is a complete theory, the fact that β is a logical consequence of α reduces to the truth of β in the unique interpretation i of α, and (if *modus ponens* is a valid rule of inference) the validity of $\alpha \supset \beta$ reduces to the truth of β in i. If the language of $\mathcal{L}$ is finite, then i is also finite. The uniqueness and finiteness of i allows then its direct "representation" and manipulation, with the consequence that inference can be substituted by computations performed directly on i. As a result of total knowledge, the relationship between a set of sentences and its interpretations has no more cardinality $1 : n$ but $1 : 1$, somehow blurring the distinction between the syntactic and the semantic level (and between logic and data structure manipulation).

In terms of IR, if we assume that our knowledge of the setting we want to model is total, instead of describing this knowledge by means of a set of sentences S we may describe it by "constructing" the unique interpretation i of S; models of IR such as those of e.g., (Crestani & van Rijsbergen, 1995; Nie, 1989) may be viewed as doing exactly this. If one adopts the Boolean model of IR, one way of making this assumption is adopting the *closed world assumption*. In fact, assuming that if keyword t does not feature in the representation of document d this means that d is not about t, is equivalent to assuming that we have total knowledge of the document (for every keyword t, we know whether or not the document is about t); in this case, d is basically a truth assignment to the propositional language, and establishing the validity of $d \supset n$ can be done by checking if n is true in this truth assignment. If one adopts a more complex model such as e.g., the one by Crestani and van Rijsbergen (1995), where keywords have associated probabilities and degrees of similarity between them, for making the total knowledge assumption it is enough to assume that the probability of each keyword is known and the similarity of each pair of keywords is known[20]. In other words, the *total knowledge assumption* may be characterised by saying that, for every sentence α of the language, either α or $\neg\alpha$ is a logical consequence of what we know about the domain.

This discussion shows that a truth-based model is, in some sense, a legitimate model of a given domain of discourse, provided that there is total knowledge of this domain. In this case, though, building and directly manipulating the unique interpretation (as done in (Crestani & van Rijsbergen, 1995; Nie, 1989)) is computationally more convenient than performing inference on the corresponding set of sentences S (for a discussion of this point

[20] In a logical model in which probability is accounted for, the difference between assuming either total or partial knowledge corresponds to the difference between the *Bayesian model* and the *generalised probabilistic model* of epistemic states discussed by Gärdenfors (1988, Chapter 2).

see (Levesque, 1986)). In other words, when knowledge is total, *logic is hardly needed*, as the direct computations on the unique interpretation may completely and conveniently bypass the standard inferential processes. In some sense, the logical approach is an oversize tool for dealing with the total knowledge assumption, and its potential is fully exploited only when partiality is assumed.

7 Discussion

Although, as remarked in Section 2, "modelness" is the main topic of the present paper, no discussion on logical modelling of IR would be complete without touching on "coverage" issues too. Ultimately, it is the effectiveness achieved by systems designed along a given model to decree its success or insuccess, and the logical model will be no exception. Although a number of interesting ideas underlie much of this research, their experimentation has so far been limited: only the models by Turtle and Croft (1990) and Crestani and van Rijsbergen (1995) have undergone experimentation to IR standards (Turtle & Croft, 1991; Crestani et al., 1995)). Why is that so? We think that the problem of experimenting with logic-based models of IR has two main aspects.

The first aspect is one of *knowledge entry*. If the model has to be checked against a document collection of realistic (i.e., TREC) size, we need to produce many thousands of document representations in the form of logical formulae. The problems raised by this are twofold:

- If the model to be tested is based on some (classical or non-classical) propositional logic (this is the case of e.g., (Crestani & van Rijsbergen, 1995; Nie, 1989)), this means that algorithms that produce weighted keyword-based representations (Salton & Buckley, 1988) can be used. However, it may well be that these algorithms produce representations whose meaning is at odds with the intended semantics of the language used. Even if extensive automatic knowledge acquisition, hence experimentation, *is* in principle possible in this case, the meaning of the experimental results becomes difficult to interpret.

- If the model is based instead on structures more complex than simple sets of weighted keywords (such as the "object-oriented" structures of (Meghini et al., 1993; Lambrix & Padgham, 1995; Chevallet & Chiaramella, 1996)), the lack of algorithms that can automatically produce complex structures faithfully representing the meaning of documents is an additional problem. Although promising research exists in this field (see e.g., (Franconi, 1994)), fully automatic indexing is not yet a viable option.

For the moment being, the availability of automatic knowledge acquisition tools seems a somehow less stringent problem in logical models of non-textual document retrieval. Approaches that combine (non-logical) processing of automatically acquired representations of the *form* of such documents, with logical reasoning on manually input representations of their *content*, have been proposed (Meghini et al., 1997). The multimedia indexing case is one in which the logical approach does not have much of a gap to fill, as for multimedia there are no satisfactory automatic *content* analysis tools outside the logical camp either. In fact, while it is indeed possible to index e.g., images based on their visual features (such as colour, shape and texture), human intervention is still inevitable at present (and it appears that it is going

to be for quite some time) for indexing images conceptually: telling "red" from "yellow" can be done automatically, but telling "man" from "woman" cannot.

The second aspect is one of *inference-in-the-large*. Once proper representations of documents and information needs are created, relevance needs to be established through inference techniques of some sort. Again, the problems raised by this are twofold.

- If the model is one based on the total knowledge assumption, as discussed in Section 6.1, direct manipulations of a data structure representing the unique interpretation may conveniently bypass the standard inferential process. Although the process may still be computation-intensive (as e.g., in (Crestani et al., 1995)), it is undoubtedly much simpler due to the absence of disjunction, the operator that, practically in any logic, singlehandedly renders all attempts at large scale inference problematic.

- If the model is *not* based on the total knowledge assumption, inference of the standard (e.g., validity-based) kind must be performed. This may be extremely arduous, given the well-known negative complexity results of the validity problem for most logics. Two avenues seem of particular interest in this respect. First, inference techniques may be experimented that yield correct results only most of the times (see e.g., (Selman et al., 1992)), at the advantage of a much greater speed. The price to be paid for this would arguably be tolerable, as information retrieval results in imperfect precision and recall anyway. Second, inference techniques may be experimented with in which partial results of the computation are stored and reused for later computations. For instance, in both the techniques used in (Turtle & Croft, 1991) and (Meghini et al., 1997), one single processing of the set of document representations and of the background knowledge base serves multiple queries, and one single processing of a query representation serves the relevance assessment of multiple documents, thus consistently reducing the query-time inference operation.

All in all, this discussion points out that experimentation is still a major problem for logical models of IR. We think, however, that this research trend is starting to show results that go beyond those of a purely theoretical nature: experimentation is being increasingly tackled (see also (Fuhr, 1997)), systems are being implemented and applied to real-sized retrieval situations [21] and, especially, integration with other non-logical techniques is being attempted. We think that exactly this latter approach is where a lot of potential for the logical model resides. It is obvious to all that a lot of problems of concern in IR (especially in its multimedia incarnation) are best tackled by techniques that have not much to do with logic. Logic is suitable for reasoning about semantics, and not every aspect of concern to IR is, arguably, best conceived in terms of semantics. IR systems that integrate logical reasoning about the *content* of documents with *form* processing techniques coming from statistics, corpus linguistics and digital signal processing, may well be one of the most promising avenues for IR research.

8 Concluding remarks

In this paper we have discussed a number of issues which we deem of central importance to the logical modelling of information retrieval, arguing that progress in the field may be

[21] See e.g., the "hyperindex" system at `http://www.dstc.edu.au/cgi-bin/RDU/hib/hib`, a direct outcome of (Bruza, 1993).

accomplished only by also taking into account "well-formedness" issues of a formal nature. In logic, well-formedness criteria are not meant to satisfy aesthetic demands, but are the guarantees of the computational effectiveness of the proposed models. In particular, we have investigated the consequences of basing a model of IR on notions as diverse as truth, validity and logical consequentiality. Only these latter two, we have argued, guarantee a role for inference, as they are the semantic counterparts of theoremhood and derivability, that are computable notions. Truth instead is not computable, and is of some interest only in the case in which the total knowledge assumption is made (a very strong assumption for every modelling endeavour, and one that puts into question the real need for logic), as in this case truth and validity coincide. A related point we have made is that the semantic structures that feature in the denotational semantics of the adopted language should not be considered as data structures to be freely manipulated, lest the logical character of the enterprise is lost. We have also argued that the infamous paradoxes of material implication are not very interesting for IR modelling. All these issues, although of independent interest, have one point in common: they somehow put into question some arguments (the "deviation" arguments of Section 5) that had been made against the use of classical logic for IR. Although we do *not* claim here that classical logic is adequate for modelling IR (the "extension" arguments of Section 5 are, we think, conclusive in this respect), our arguments suggest that the rebuttal of classical logic has been too hasty. Reconsidering classical logic, maybe sieving out the real reasons of its inadequacy to the IR case, may thus be a way of better understanding the merits of non-classical approaches.

Acknowledgements

I would like to thank Gianni Amati, Iain Campbell, Fabio Crestani, Mounia Lalmas, Carlo Meghini, Jian-Yun Nie, Alessandro Saffiotti, Tefko Saracevic, Ulrich Thiel and Keith van Rijsbergen for discussing with me the topics of this paper.

References

Bruza, P. D. (1993). *Stratified information disclosure. a synthesis between hypermedia and information retrieval.* Unpublished doctoral dissertation, Katholieke Universiteit Nijmegen, Nijmegen, NL.

Bruza, P. D., & Huibers, T. W. (1994). Investigating aboutness axioms using information fields. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (pp. 112–121). Dublin, IRL.

Bruza, P. D., & van der Gaag, L. C. (1993). Efficient context-sensitive plausible inference for information disclosure. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval* (pp. 12–21). Pittsburgh, PA.

Carnap, R., & Bar-Hillel, Y. (1953). Semantic information. *British Journal for the Philosophy of Science*, *4*, 147–157.

Chevallet, J. P., & Chiaramella, Y. (1996). Extending a logic-based model with algebraic knowledge. In I. Ruthven (Ed.), *Proceedings of MIRO-95, Workshop on Multimedia Information Retrieval.* Glasgow, UK: Springer Verlag, Heidelberg, FRG.

Chiaramella, Y., & Chevallet, J. P. (1992). About retrieval models and logic. *The Computer Journal*, *35*, 233–242.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, *7*, 19–37.

Crestani, F., Ruthven, I., Sanderson, M., & van Rijsbergen, C. J. (1995). The troubles with using a logical model of IR on a large collection of documents. In *Proceedings of TREC-4, 4th Text Retrieval Conference.* Washington, DC.

Crestani, F., & van Rijsbergen, C. J. (1995). Probability kinematics in information retrieval. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieva* (pp. 291–299). Seattle, WA.

Darlington, J. (1969). Theorem proving and information retrieval. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence 4* (pp. 173–181). Edinburgh, UK: Edinburgh University Press.

De Giacomo, G., & Lenzerini, M. (1995). Making CATS out of kittens: description logics with aggregates. In *Proceedings of DL-95, 4th International Workshop on Description Logics* (pp. 143–147). Roma, Italy.

de Kleer , J., & Williams, B. (1987). Diagnosing multiple faults. *Artificial Intelligence*, *32*, 97–130.

de Swart, H., & Nederpelt, R. P. (1992). Implication. a survey of the different logical analyses of "if ... then ...". *Nieuw Archief voor Wiskunde*, *10*, 77–104.

Devanbu, P., & Jones, M. (1994). The use of description logics in KBSE systems. In *Proceedings of the 17th International Conference on Software Engineering.* Sorrento, Italy.

Dunn, J. M. (1986). Relevance logic and entailment. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of Philosophical Logic* (Vol. 3, pp. 117–224). Dordrecht, NL: Reidel.

Franconi, E. (1994). Description logics for natural language processing. In *Proceedings of the 1994 AAAI Fall Symposium on Knowledge Representation for Natural Language Processing in Implemented Systems.* New Orleans, LA.

Fuhr, N. (1995). Probabilistic Datalog: a logic for powerful retrieval methods. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (pp. 282–290). Seattle, WA.

Fuhr, N. (Ed.). (1997). *ESPRIT BRA project 8134 "FERMI". Deliverable 11: "Experimentation".* Commission of the European Communities.

Gärdenfors, P. (1988). *Knowledge in flux. Modeling the dynamics of epistemic states.* Cambridge, MA: The MIT Press.

Gordon, M. J. (1979). *The denotational description of programming languages.* Heidelberg, FRG: Springer Verlag.

Haack, S. (1978). *Philosophy of logics.* Cambridge, UK: Cambridge University Press.

Halpern, J. Y. (1990). An analysis of first-order logics of probability. *Artificial Intelligence*, *46*, 311–350.

Harper, W. L., Stalnaker, R., & Pearce, G. (Eds.). (1981). *Ifs. Conditionals, belief, decision, chance and time.* Dordrecht, NL: Reidel.

Hunter, A. (1995). Using default logic in information retrieval. In C. Froidevaux & J. Kohlas (Eds.), *Symbolic and quantitative approaches to reasoning and uncertainty* (pp. 235–242). Heidelberg, FRG: Springer.

Israel, D. J. (1985). A short companion to the naive physics manifesto. In J. R. Hobbs & R. C. Moore (Eds.), *Formal theories of the commonsense world* (pp. 427–447). Norwood, NJ: Ablex.

Lalmas, M., & van Rijsbergen, C. J. (1993). A model of an information retrieval system based on situation theory and Dempster-Shafer theory of evidence. In *Proceedings of the 1st Workshop on Incompleteness and Uncertainty in Information Systems* (pp. 62–67). Montreal, Canada.

Lambrix, P., & Padgham, L. (1995). Analysis of part-of reasoning in description logics for use in a document management application. In *Proceedings of the International Symposium on Artificial Intelligence* (pp. 102–110). Monterrey, Mexico.

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of AAAI-84, 4th Conference of the American Association for Artificial Intelligence* (pp. 198–202). Austin, TX.

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence*, *30*, 81–108. ([a] Also reprinted in (Mylopoulos & Brodie, 1989), pp. 69–82.)

Levesque, H. J., & Brachman, R. J. (1987). Expressiveness and tractability in knowledge representation and reasoning. *Computational Intelligence*, *3*, 78–93.

McDermott, D. (1978). Tarskian semantics, or: no notation without denotation! *Cognitive Science*, *2*, 272–282.

McGuinness, D. L., & Alperin Resnick, L. (1995). Description-logic based configuration for consumers. In *Proceedings of DL-95, 4th International Workshop on Description Logics* (pp. 109–111). Roma, Italy.

Meghini, C. (1995). An image retrieval model based on classical logic. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (pp. 300–308). Seattle, WA.

Meghini, C., Sebastiani, F., & Straccia, U. (1997). The terminological image model. In *Proceedings of ICIAP-97, 9th International Conference on Image Processing and Analysis.* Firenze, I. (Forthcoming)

Meghini, C., Sebastiani, F., Straccia, U., & Thanos, C. (1993). A model of information retrieval based on a terminological logic. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval* (pp. 298–307). Pittsburgh, PA. (Published by ACM Press, Baltimore, MD)

Meghini, C., & Straccia, U. (1996). A relevance description logic for information retrieval. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval* (pp. 197–205). Zürich, CH.

Müller, A., & Thiel, U. (1994). Query expansion in an abductive information retrieval system. In *Proceedings of RIAO'94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"* (pp. 461–480). New York, NY.

Mylopoulos, J., & Brodie, M. L. (Eds.). (1989). *Readings in artificial intelligence and databases.* San Mateo, CA: Morgan Kaufmann.

Nie, J.-Y. (1989). An information retrieval model based on modal logic. *Information Processing and Management*, *25*, 477–491.

Patel-Schneider, P. F., Brachman, R. J., & Levesque, H. J. (1984). ARGON: knowledge representation meets information retrieval. In *Proceedings of CAIA-84, 1st IEEE Conference on Artificial Intelligence Applications* (pp. 280–286). Denver, CO.

Reiter, R. (1984). Towards a logical reconstruction of relational database theory. In M. L. Brodie, J. Mylopoulos, & J. W. Schmidt (Eds.), *On conceptual modelling* (pp. 191–233). Heidelberg, FRG: Springer. ([a] Also reprinted in (Mylopoulos & Brodie, 1989), pp. 301–326.)

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence*, *32*, 57–95.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*, 513–523.

Sebastiani, F. (1994). A probabilistic terminological logic for modelling information retrieval. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (pp. 122–130). Dublin, IRL. (Published by Springer Verlag, Heidelberg, FRG)

Selman, B., Levesque, H., & Mitchell, D. (1992). A new method for solving hard satisfiability problems. In *Proceedings of AAAI-92, 10th Conference of the American Association for Artificial Intelligence* (pp. 440–446). San Jose, CA.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford: Basil Blackwell. ([a] Also reprinted in (Harper, Stalnaker, & Pearce, 1981), pp. 41–55.)

Stalnaker, R. C., & Thomason, R. H. (1970). A semantical analysis of conditional logic. *Theoria*, *36*, 23–42.

Straccia, U. (1996). Document retrieval by relevance terminological logics. In I. Ruthven (Ed.), *Proceedings of MIRO-95, Workshop on Multimedia Information Retrieval.* Glasgow, UK: Springer Verlag, Heidelberg, FRG.

Turtle, H. R., & Croft, W. B. (1990). Inference networks for document retrieval. In *Proceedings of SIGIR-90, 13th ACM International Conference on Research and Development in Information Retrieval* (pp. 1–24). Bruxelles, Belgium.

Turtle, H. R., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, *9*, 187–222.

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

van Rijsbergen, C. J. (1989). Towards an information logic. In *Proceedings of SIGIR-89, 12th ACM International Conference on Research and Development in Information Retrieval* (pp. 77–86). Cambridge, MA.

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

van Rijsbergen, C. J., & Lalmas, M. (1996). Information calculus for information retrieval. *Journal of the American Society for Information Science*, *47*, 385–398.

Wong, S., & Yao, Y. (1991). A probabilistic inference model for information retrieval. *Information Systems*, *16*, 301–321.