# Logic and Uncertainty in Information Retrieval

Fabio Crestani[1] and Mounia Lalmas[2]

[1] Department of Computer Science
University of Strathclyde
Glasgow G1 1XH, Scotland
fabioc@cs.strath.ac.uk
[2] Department of Computer Science
Queen Mary, University of London
London E1 4NS, England
mounia@dcs.qmw.ac.uk

**Abstract** The use of logic in Information Retrieval (IR) enables one to formulate models that are more general than other well known IR models. Indeed, some logical models are able to represent, within a uniform framework, various features of IR systems, such as hypermedia links, multimedia content, and users knowledge. Logic also provides a common approach to the integration of IR systems with logical database systems. Finally, logic makes it possible to reason about an IR model and its properties. This latter possibility is becoming increasingly important since conventional evaluation methods, although good indicators of the effectiveness of IR systems, often give results which cannot be predicted, or for that matter satisfactorily explained. However, logic by itself cannot fully model IR. In determining the relevance of a document to a query the truth value or the validity of a logical formula relating the two is not enough. It is necessary to take into account the uncertainty inherent in such a formulation. This paper gives an overview of how past and current research have combined the use of logical and uncertainty theories for the formulation of more advanced models for the representation and retrieval of information.

## 1 Introduction

*Information retrieval* (IR) is the science and technology concerned with the effective and efficient retrieval of information for the subsequent use by interested parties. The central problem in IR is the quest to find the set of relevant documents, amongst a large collection, containing the information sought thereby satisfying an information need usually expressed by a user with a query. The documents may be objects (items) in any medium, text, image, audio, or, indeed a mixture of all three. An important area of research concentrates on the modelling of objects and processes involved in the retrieval of information.

Well known models of IR are the Boolean, vector space, probabilistic, and fuzzy models; these have been studied in detail and implemented for experimentation, as well as, commercial purposes. Nevertheless, the known limitations of

these models have caused researchers to propose new models. One such model is the logical model for IR [15, 23, 24, 54, 57, 90].

In recent years there have been several attempts to define *a logic for IR* along the so-called *logical approach*, initiated by Cooper [21] and given decisive impulse by van Rijsbergen [103, 104]. Logical IR models were studied to provide a rich and uniform representation of information and its semantics, with the aim to improve retrieval effectiveness. The earliest approaches were directed to the use of classical logic, like Boolean logic. The basis of a logical model for IR is the assumption that queries and documents can be represented effectively by logical formulas. In order to retrieve a document, an IR system has to infer the formula representing the query from formulas representing the document. This logical interpretation of query and documents emphasises that *information retrieval is an inference process* that computes whether a document $d$ is relevant to a query $q$ using both information present in the document itself and external information, like for example, user knowledge. An example is given in classical logic where inference is often associated with *logical implication*: a document is relevant to a query if it implies the query, or in other words, if the query can be inferred from the document. Such an evaluation formally embodies the semantics of the information represented in the query and in the document.

This way of viewing IR is especially fascinating once we consider, instead of the proof-theoretic "symbol-crunching" level of logic, its model-theoretic, semantic level. In terms of the 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.

In addition, the use of logic to build IR models enables one to obtain models that are more general than earlier well known IR models. Indeed, some logical models are able to represent within a uniform framework various features of IR systems, such as hypermedia links [13, 97], multimedia content [19, 67], users knowledge [75], cross-lingual [73], and structured documents [58]. It also provides a common approach to the integration of IR systems with logical database systems [37]. Finally, logic makes it possible to reason about an IR model and its properties [47, 69, 89]. This latter possibility is becoming increasingly important since conventional evaluation methods, although good indicators of the effectiveness of IR systems, often give results which cannot be predicted, or satisfactorily explained.

However, logic by itself cannot fully model IR. In determining the relevance of a document to a query, the success or failure of an implication relating the two is not enough. It is necessary to take into account the *uncertainty* inherent in such an implication. The introduction of uncertainty can also be motivated from the consideration that a collection of documents cannot be considered as a consistent and a complete set of statements. In fact, documents in the collection could and often do contradict each other in any particular logic, and not all the necessary knowledge is available.

It has been shown [18, 56, 103] that classical logic, the most commonly used logic, is not adequate to represent query and documents because of the intrinsic uncertainty present in IR. To cope with uncertainty a logic for *uncertain inference* needed to be introduced. In fact, if $d \rightarrow q$ is uncertain, then we can measure its degree of uncertainty by $P(d \rightarrow q)$[1].

In 1986 Van Rijsbergen proposed the use of a non-classical conditional logic for IR [103]. This would enable the evaluation of $P(d \rightarrow q)$ using the following *logical uncertainty principle*:

> "Given any two sentences $x$ and $y$; a measure of the uncertainty of $y \rightarrow x$ related to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \rightarrow x$."

This principle was the first attempt to make an explicit connection between non-classical logics and IR uncertainty modelling. However, when proposing the above principle, Van Rijsbergen was not specific about which logic and which uncertainty theory to use. As a consequence, various logics and uncertainty theories have been proposed and investigated. The choice of the appropriate logic and uncertainty mechanisms has been a main research theme in logical IR modelling leading to a number of different approaches over the years.

In this paper we present a number of approaches to the use of logical and uncertainty models in IR. We will not address models based on classical logic, since their limitations (but also strengths) have long been recognised [56, 95]. Here we will only present models attempting to capture the uncertainty of the IR inference process, either through non-classical logics or some uncertainty theory defined on logical basis. We will call these two classes of models "logical models" and "logical-uncertainty models," respectively.

Another completely different class of models that will be presented in this paper is that of so called "meta-models." Meta-models attempts to formally study the properties and the characteristics of IR systems within a uniform logical framework. They aim at making it possible to compare IR models through formal properties of these models, instead of through their effectiveness, that can only be evaluated by means of expensive experimentation.

Thus, this paper is structured as follows. In Section 2 we introduce the concept of logical relevance, that is at the basis of the use of logic in IR. In Section 3 we present a selection of logical models of IR, while in Section 4 we present a number of logical-uncertainty models. Finally, in Section 5 we present work carried out towards developing meta-models of IR.

This paper is not intended to provide a complete survey of all attempts to using logics and uncertainty theories in IR. Such a task is outside the purpose of this paper. Also, it is not our intention to state that logical or logical-uncertainty IR models are better than other IR models, or to decide which of these models is the most appropriate. Instead, we are trying to highlight the approaches that we

[1] In most of the models presented here uncertainty is measured using probability. In this case $P(\alpha)$ stands for "the probability that $\alpha$."

believe to be most interesting to future generations of IR researchers. We hope that the lessons learned will be carried forward and new interesting development will arise in this exciting area of research.

## 2 A Logical Definition of Relevance

Relevance is one of the most important, if not "the fundamental," concept in the theory of IR. The concept arises from the consideration that if a user of an IR system has an information need, then some information stored in some documents in a document collection may be "relevant" to this need. In other words, the information to be considered relevant to a user's information need is the information that might help the user satisfy his or her information need. Any information that is not considered relevant to a user's information need, is to be considered "irrelevant" to that same information need. This is a consequence of accepting a dichotomous concept of relevance [2].

A logical definition of relevance was considered for the first time in the context of IR by Cooper in a paper written almost 30 years ago [21]. For Cooper *logical relevance* was another name for topic-appropriateness, and he addressed the problem of giving a definition of logical relevance for IR by analogy with the same problem in question-answering systems. The analogy goes only as far as having questions with a yes-no (true-false) type of answer, and while Cooper's work started by analysing question-answering systems, later he abandoned the analogy. Relevance is defined by Cooper as "logical consequence." To make this possible both queries and documents need to be represented by sets of declarative sentences. In the case of a yes-no query, the query is represented by two formal statements of the form $p$ and $\neg p$. The two statements representing the query are called "component statements." A subset of the set of stored sentences is called "premiss set" if and only if the component statement is a logical consequence of that subset. A "minimal premiss set" for a component statement is one that is as small as possible in the sense that if any of its members were deleted, the component statement would no longer be a logical consequence of the set. Logical relevance is defined as a two-place relation between stored sentences and the query represented as component statements (the representation of the information need). A first definition of logical relevance says:

> "A stored sentence is logically relevant to (a representation of) an information need if and only if it is a member of some minimal premiss set of stored sentences for some component statement of that need."

This definition of relevance is essentially just a proof-theoretic notion that has been generalised to be applicable to information needs involving more than one component statement.

[2] We will not address here the work challenging this binary view of relevance. The interested reader can look at the chapter in this book dealing with fuzzy approaches to IR or refer to other work using probabilistic approaches, like for example [2, 7].

Although logical relevance was initially defined only for sentences, it can be easily extended to apply to stored documents: a document is relevant to an information need if and only if it contains at least one sentence which is relevant to that need.

In the same paper Cooper attempted to tackle a generalisation of such a definition to natural language queries and documents. However, without a formalised language, no precise definition of the logical consequence relation is at hand, and thus we lose a precise definition of relevance. The problems of ambiguity and vagueness of natural language deny the possibility of extending the previous logical notion of relevance, despite the fact that the general idea of implication in natural language is a reasonably clear one. The definition of relevance, so far as natural language in concerned, is only a definition-in-principle — a conceptual definition — but not yet defined on a mathematical level.

Cooper also tried to tackle the problem of having "degrees of relevance," or as he wrote: "shades of grey instead of black and white" [21, pp.30]. The idea was to extend the system of deductive reasoning used to access logical relevance to a system of plausible reasoning. Cooper argued that plausible or probabilistic inference was not as well defined as deductive inference, even for formalised languages. However, he added that when such tools are formalised enough then this development would become a "sensible and indeed inescapable idea," because it would enable the ranking of documents according to an estimated probability of relevance. What he proposed was to assign a higher probability of relevance to a sentence or a document that has greater probability of belonging to a residual minimal premiss set.

Cooper went on extending the previous definition of relevance to the case of non-inferential systems and to the case of topical queries, but the extensions are not of interest in the context of this chapter. What should be retained from this discussion is that Cooper was the first to associate the topic-appropriateness sense of relevance [91] with logical implication and that he recognised the importance of evaluating the uncertainty of such implication to rank documents in relation to their estimated measure of relevance. Many other researchers followed this idea proposing the use of different logics to capture relevance.

An early common belief was that the logical implication needed to capture relevance was not the classical material implication. The reasons why the use of the classical material implication $d \supset q$ is not appropriate for IR is in the definition of material implication itself, and there are many ways of explaining why material implication is not suitable for IR (see for example [18]). For reasons of brevity, we repeat only the most important argument, that is the one most often used to dismiss the suitability of classical material implication for IR. The argument relates to the fact that the truth of the material implication $d \supset q$ is to be determined relative to a particular evaluation situation. To determine the truth of $d \supset q$ we have to compare the truth of $d$ with that of $q$. Using a truth table for $d \supset q$ one can see that when $d$ is false, no matter what the query $q$ is, $d \supset q$ will always be true. Herein lies the problem. In fact, $d$ is true only when $d$ is retrieved, but, given a retrieval situation in which $q$ is submitted, a document

$d$ is always false since it has not been retrieved yet. Therefore the real retrieval situation corresponds to the case of $d$ false and such a document is relevant to any query $q$. This obviously does not provide a suitable definition of relevance.

The idea that a non-classical form of logical implication was needed for defining relevance was first proposed by Van Rijsbergen in [103]. That initial idea was supported with stronger arguments in [102], where the logical uncertainty principle was formulated. It is now clear that it is not possible to apply the logical uncertainty principle without a combination of a (non-classical) logic formalism and uncertainty theory. Both logic and uncertainty theory alone cannot fully capture this view of relevance. Depending on where the focus lays, one can use a non-classical logic combined with a theory of uncertainty or a theory of uncertainty defined in terms on a non-classical logic. In the following we will present some examples of these two classes of approaches to the use of non-classical logics and uncertainty theories in IR.

## 3 Logical Models of Information Retrieval

Some logical models are able to capture the uncertainty inherent in the IR process. These models capture the uncertainty mainly in two ways: qualitatively by the logic itself (for example, via default rules, non-monotonicity, or background conditions), or quantitatively by adding an uncertainty theory to the logic (for example, fuzzy logic). In the following we will present some of these models, just those that we think are most representative of the work done in this area.

### 3.1 Models Based on Modal Logic and Conceptual Graphs

Modal Logic [43] adopts the notion of possible worlds [52] that correspond to the interpretations in classical logic, but which are connected to each other via an accessibility relation. The evaluation of the truth of a proposition is with respect to a possible world, and may involve the evaluation of the truth of the proposition in connected worlds (see more about possible worlds semantics in Section 4.3).

*Modal Logic* was first used to develop a logical model for IR by Nie [71, 72]. Documents are worlds, and queries are formulae. A document represented by a world $d$ is relevant to a query represented by a formula $q$ if $q$ is "true" in $d$, or if it is true in a world $d'$ accessible from $d$.

The accessibility relation captures the transformation of documents; the fact that the world $d$ is connected to the world $d'$ is interpreted as $d$ being transformed into $d'$. For example, $d'$ could contain terms that are synonymous to those contained in $d$. The accessibility relationship can have different properties. For example, transitivity, meaning that if a world $d$ is related to a world $d'$, which is itself related to a world $d''$, then the world $d$ is also related to the world $d''$. Consider the example of a hypertext system. Using Modal Logic, worlds can represent texts (nodes) and the accessibility relation can represent the links between the texts. Let $d$ be a text-world, linked to a second text-world $d'$, itself

linked to a third text-world $d''$. If $d$ does not contain information relevant to a query, but $d'$ does, we may still want to retrieve $d$. It could be that only $d''$ contains information relevant to the query. Do we still want to retrieve $d$? This decision can be formally represented by allowing the accessibility relation to be transitive or not. This example gives an indication on how it is possible to reason about the type of system needed.

The model also allows the transformation of both the query and the data set. One query can be transformed into another one using, for example, thesaural information. Query transformation is not a new approach in IR (e.g., query expansion). The novelty is that the transformation process can be formally represented, and hence reasoned upon. Transforming a data set can capture the modelling of a user's state in the retrieval process. The data set can be transformed until it reaches one that reflects the user's state.

The above approach has been implemented in a medical context [20, 69]. The results showed that such an approach was promising, but the experiment was carried out on only 36 documents, so these results are only an indication of the effectiveness of the model. Another implementation of the model uses the WordNet thesaurus to transform queries, and is applied to the CACM test collection [66]. The results showed a positive increase in effectiveness.

A variant of this model was proposed in [17]. The logic was instantiated by the formalism of conceptual graphs [96], which are graphs built out of concepts and their associated semantics. Documents and queries are represented by conceptual graphs, and the transformation process is instantiated by operations performed on the graphs. For example a graph could be transformed to another one, where one concept in the initial graph is replaced by a more general one. This formalism has been used for instance to represent medical documents and software components. The problem, however, is the automatic construction of graphs from documents or queries (see [76] for further work, and its application to image retrieval).

### 3.2 Models Based on Situation Theory

*Situation Theory* is a theory of information that provides an analysis of the concept of information and the manner in which intelligent organisms (referred to as cognitive agents) handle and respond to the information picked up from their environment [35]. The theory defines the nature of information flow and the mechanisms that give rise to such a flow. Information items are represented by types. For example, $\phi = [\dot{s}|\dot{s} \models << Swimming; Mounia; 1 >>]$ represents the information item that Mounia is swimming. Nothing is said about the truth of this type; a type is just the representation of an item of information. What makes a type true is the situation (a partially defined world) from which the information represented by that type is extracted. Situation Theory models the notion of "make true" by the support relation, denoted $\models$. If $s$ is a situation that makes the information "Mounia is swimming" true, then one can write $s \models \phi$, which should be read as "$s$ supports $\phi$."

Some IR logical model were developed based on Situation Theory [59]. A document is a situation $s$ and the query is a type $\phi$. The document is relevant to the query if there exists a flow of information from a situation $s$ to a situation $s'$ such that $s' \models \phi$. The nature of the flow depends on the so-called "constraints" which capture semantic relationships (e.g., the relationships that many people attach to white wine and Australian wine, the relationships based on synonymy, etc.). More formally, constraints are defined between types. Let $\phi$ and $\psi$ be two types that constitute the constraint $\phi \rightarrow \psi$. The application of this constraint to a situation $s$ is possible if first $s \models \phi$ and then informs of the existence of a situation $s'$ such that $s' \models \psi$: the fact $s \models \phi$ carries the information that $s' \models \psi$. A flow of information circulates between the situations $s$ and $s'$, and the nature of the flow is defined by the constraint $\phi \rightarrow \psi$.

Flows of information do not always materialise because of the unpredictable nature of situations, thus flows are often uncertain. In Situation Theory, an uncertain flow is modelled by a conditional constraint of the form $\phi \rightarrow \psi | B$, which highlights the fact that $\phi \rightarrow \psi$ holds if some background conditions captured within $B$ are met. If the background conditions are satisfied, the corresponding flow arises. The use of background conditions in an IR model acknowledges the important fact that information is seen to be dependent on a context. For example, background conditions can represent context with respect to polysemic words.

Situation theory also allows the representation of uncertainty via background conditions, although only qualitatively [53]. It can be argued that the quantitative representation of the relevance of documents to queries is there only to rank documents, the numerical values have no real significance. Therefore, a qualitative representation of uncertainty may be enough to rank documents according to their estimated relevance to a query. Based on the background conditions, the fact that a document modelled by a situation $s1$ "supports more" the information item $T$ than the document represented by the situation $s2$ would mean that the first document is more relevant to the query than is the second document. There is however much work to be done before such expressions become possible, but an elegant formalism that may be used for this purpose has already been advanced in [5].

The problem, however, with a Situation Theory based IR model is the difficult implementation of the model. The power of the theory is too complex to capture in an implementation [9]. For example, in [55], the WordNet thesaurus [66] was used to build the constraints. The implemented constraints were inappropriate (too general) for the test collection used in the experiments, hence, poor experimental results were obtained. There is however interesting work where ontologies are being developed to provide a"semantic for the Web" [3], in order to formally represent and qualitatively reason about the content of Web information objects. Situation Theory would provide an excellent framework for that purpose.

Situation Theory has also been used to develop a framework for searching on a thesaurus [8], and a meta-theory of IR [45], where the notion of aboutness was

[3] See `http://www.semanticweb.org/`

defined in terms of the flow of information. An extension of Situation Theory, called Channel Theory, was also used to develop logic-based IR models. This is discussed in the next section.

### 3.3 Models Based on Channel Theory

It is often the case that two situations are systematically related to each other, by way of a flow of information. For example: a situation where smoke is perceived is related to a situation where a fire has occurred; a situation where a person hears the door bell ringing is related to a situation where a second person is at the door pressing the bell; a situation representing a HTML document is related to a situation representing one it links to; a situation where a user views non-relevant retrieved document is related to a situation where the user adjusts his or her information need. Therefore, in addition to constraints, there are relationships that link situations. The concept of a channel is introduced in [6] to express the relationships, by way of an information flow, between two situations.

*Channel theory* defines formally channels, together with the mathematical properties that support the flow of information. For example, two operations are defined on channels: the sequential combination of channels and the parallel combination of channels. Their definitions satisfy fundamental properties of information and its flow.

One major asset of the use of channels is that the physical link between situations is conceptually defined. This allows formal representation at two levels: the link and its nature. For example, a document contains information about another document, either implicitly (e.g., the two documents are on the same topic) or explicitly, by way of citations, or links (e.g., hypermedia systems). Both of these cases can be captured with channel theory. In the first case, the nature of the flow can be defined in terms of thesaural relationships, but the link between one document and another that contains information relevant to the query is unknown. In the second case, the nature of the link is often unknown. However, the relevance of a document to a query can be calculated, since it is known that a flow of information circulates between that document and one that contains the information being sought after.

Similarly to the case for situation theory, the main problem with an IR model based on channel theory is its difficult implementation. Here in addition to the implementation of constraints, we must also provide an implementation of channels.

The use of channel theory to model IR has been investigated in [106], where the connection between IR, logic, probability and information containment was made. It was indicated that the use of channels present many potentials for theoretical IR modelling because they can apply to various IR processes present in advanced IR systems.

### 3.4 Models Based on Terminological Logic

*Terminological logics* come from the area of artificial intelligence, in particular, knowledge representation. Terminological logics derive from a large group of knowledge representation language (such as, for example, KL-ONE) based on semantic networks and inspired by the notion of frames. They provide object-oriented flavoured representations. The primary syntax starts with terms, which are either individuals or relations. Concepts are defined on top of those. For example, the concept:

$$(and\ paper\ (forall\ author\ european))$$

denotes the class of papers written by European authors only. Concepts come with a partial order $\leq$ which stands for conceptual containment. The fact that two concepts are ordered constitutes an axiom that describes thesaural knowledge. For example:

$$dog \leq (and\ animal (exactly\ 4\ leg))$$

states that dogs are four-legged animals (but not all four-legs animals are dogs).

The fact that a particular individual is an instance of a concept, is written as an assertion. For example, the following:

$$(and\ paper (forall\ author\ european))[paper1]$$

means that the individual document, named $paper1$, belongs to the concept that denotes the class of papers where all authors are European.

Given a concept $C$, defining whether an individual $i$ is an instance of this concept, that is, evaluating whether $C[i]$ holds, uses the set of given assertions describing various facts, axioms that describe thesaural knowledge, and the notion of subsumption (hierarchical domination) defined by $\leq$. The evaluation of $C[i]$ is defined similarly to classical logic. However, the semantics of the terms go beyond truth and falsity; for example, the semantics of the concept "author" will be the set of individuals that are authors.

The use of terminological logic for IR was proposed in [64]. There, documents are represented by individual constants, whereas a class of documents is represented as a concept. Queries are described as concepts. Given a query represented by a concept $Q$, the retrieval task is to find all those documents $d$ such that $Q[d]$ holds. The evaluation of $Q[d]$ uses the set of assertions describing documents, that is, we are not evaluating whether $d \rightarrow q$, but rather whether individual $d$ is an instance of the class concept $Q$.

The model has been extended to include probabilities with respect to possible-world semantics [94]. That is, all interpretations (worlds) are considered, each of them with a given probability. For example, the proposition $w(\gamma) > 0.8$ means

that the summation of the probability of those worlds (a probability distribution is defined on worlds) where $\gamma$ is true is greater than 0.8. It was shown that this approach allows the representation of subjective beliefs (e.g., the belief that a document is about a given concept is of 0.8) as well as statistical information (e.g., 80% of worlds contain some given information).

In [65], a different class of terminological logic is used, based on a variation of relevance logics and four-valued semantics proposed in [77]. The terminological logic used in [64,94] was discovered to have poor computational properties [16]. By contrast, relevance logics have language ALC as their base, which is the standard description logic on which people conduct experiments (i.e., integrations or deviations, such as probabilistic extensions, fuzzy extensions, etc.).

### 3.5 Models Based on Abductive Logic

Abduction is a way of explaining observations, expressed as formulae, by minimally extending a theory with some added hypotheses. More formally, given a theory $T$, and a formula $p$ that needs to be explained in terms of $T$, abduction leads to a set of hypotheses $H$ such that $T \cup H \models p$. The hypotheses are also referred to as abductive sentences.

A logical model based on *abductive reasoning* has been developed in [97] to build a hybrid system, where IR and hypertext facilities are combined. In this case, $p$ is a query, and $T$ is a knowledge base which captures semantic relationships (e.g., synonymy). The abductive sentences correspond to information related to documents (e.g., author, topic, etc.). The abduction process yields a structured proof, which is used to compute a solution space (a formula can be explained in different ways, so several sets of hypotheses, or solutions, can be found). Each solution generates a model that constitutes a starting point for a user to browse, either to access relevant documents, or related documents (e.g., same authors). This work is particularly attractive because it defines a logical framework that integrates IR and hypertext.

This approach was also used in an image retrieval system [67] where images were described by qualitative rules about contour, colour, texture, etc., and hence expressed as formulae.

In [87], relevance feedback is viewed as a process of explanation. In this case, a relevance feedback theory should provide an explanation of why a document is relevant to an information need. Such an explanation can be based on how information is used within documents. Abductive logic is used to provide a framework for an explanation-based account of relevance feedback.

### 3.6 Models Based on Default Logic

*Default reasoning* is concerned with the modelling of assumption of the form "birds usually fly" which are assumptions that do not always hold (e.g. penguin, ostrich). One instance of default reasoning is default theory [81], which is composed of two parts: a set of axioms, referred to as the basic theory, and a set of

default rules. The inference rules are those of classical logic plus an additional mechanism for default rules. In default logic, the following rule

$$\frac{a : b}{c}$$

indicates that if $a$ is true, and $\not b$ cannot be inferred, then infer $c$. The application of default rules to the basic theory consists of an extension. Several extensions can be obtained since default rules can lead to different conclusions, thus capturing the non-monotonicity nature of the reasoning.

The use of default logic in IR was proposed in [48,49] as a means to obtain a uniform and comprehensive framework for reasoning with keywords, for example for implementing sound query expansion. The proposed framework allows the representation of context in IR (e.g. polysemy) and the handling of exception.

### 3.7 Model Based on Belief Revision

*Belief revision* is an approach from non-monotonic reasoning that has been used in many domains [41]. The approach provides a means to formalise changes done to a knowledge base after the arrival of new information. The most interesting case arises when contradiction occurs between the old knowledge base and new knowledge (new information); the change (revision) must always lead to a consistent knowledge base. In addition, the revised knowledge base must contain the new information. Therefore, old knowledge maybe be deleted from the knowledge base, but this deletion should be minimal.

The use of belief revision in IR was attempted in [61] as a way to compute the similarity of a document to a query for retrieval purpose. The Dalal's revision operator was chosen for implementing the belief revision process, since it provides an order among proposition interpretations, where propositions model index terms. The ordering is used to formulate a similarity measure between a document and a query (expressed as formulae) based on the number of revisions necessary for the document to "reach" the query. It should be noted that both documents and queries can have, each, several interpretations, so normalisation becomes necessary.

The proposed belief revision framework has been extended in two ways. First, a syntactic characterisation of the logical formulae using DNF was advanced to overcome the problem arising from a direct implementation of logical interpretations which has an exponential complexity [62]. The characterisation allows the design of polynomial-time algorithms. Such work is crucial for the efficient implementations of logical IR models. On the other hand, in [63], retrieval situations were formally expressed in the belief revision model, which enable to formally capture, for example, user knowledge (e.g. of the collection), user information-seeking tasks (e.g. precision vs. recall search), and so on.

Finally, other work on the use of belief revision in IR can be found in [1,34].

### 3.8 Models Based on Fuzzy Logic

*Fuzzy set theory* is a formal framework well suited to model vagueness and imprecision [108]. In IR it has been successfully employed at several levels [26, 27, 51], in particular for the definition of a superstructure of the Boolean model, with the appealing consequence that existing Boolean IRS systems can be improved without redesigning them completely. Through these extensions the gradual nature of relevance of documents to user queries can be modelled.

In this paper we do not address this area of research since it is presented extensively in the paper by Bordogna and Pasi in this same book.

## 4 Logical-Uncertainty Models of Information Retrieval

Logical-uncertainty models (sometimes referred to as uncertain inference models) are based on an uncertainty theory (for instance, probability theory, semantic theory, imaging) that is defined on a logical basis. They enable a more complex definitions of relevance than other IR models (than probabilistic relevance model, for instance, which are based mainly upon statistical estimations of the probability of relevance). With logical-uncertainty models, information not present in the query formulation may be included in the evaluation of the relevance of a document. Such information might be domain knowledge, knowledge about the user, user's relevance feedback, and so on.

Another characteristic of logical-uncertainty models is that they are not as strongly collection-dependent as most other IR models. In most IR models, parameters (e.g., normalisation and weight combination parameters) are only valid for the current collection, while logical-uncertainty models can use knowledge of the user or the application domain that can be useful with many other collections.

In this section we will present some logical-uncertainty models, with particular attention to models based on an uncertainty theory defined using a non-classical logic.

### 4.1 Models Based on Probability Theory

In IR, probabilistic modelling refers to the use of a model that ranks documents in decreasing order of their evaluated probability of relevance to a user's information need [83]. Past and present research has made use of formal theories of probability and of statistics in order to evaluate, or at least estimate, those probabilities of relevance. These attempts are to be distinguished from looser ones like, for example, the "vector space model" [88] in which documents are ranked according to a measure of similarity with the query. A measure of similarity cannot be directly interpretable as a probability. In addition, similarity based models generally lack the theoretical soundness of probabilistic models.

A treatment of models based on Probability Theory is beyond the scope of this section. Good surveys of probabilistic modelling in IR are [25, 36] and we

refer the interested reader to them. The models presented in this section are based on the idea that IR is a process of uncertain inference. This research area is promising in that it is attempting to move away from the traditional approaches, and may provide the breakthrough that appears necessary to overcome the limitations of current IR systems.

There are two main types of probabilistic uncertain inference models. The first is based on non-classical logic, to which probabilities are mapped, and the second is based on Bayesian inferences. Models based of the first class can be found in Section 3, here we will only address models of the second class.

A probabilistic formalism for describing inference relations with uncertainty is provided by *Bayesian inference networks*, which have been described extensively in [68, 78]. Turtle and Croft [98, 99] applied such networks to IR. Figure 1 depicts an example of such a network. Nodes represent IR entities such as documents, index terms, concepts, queries, and information needs. We can choose the number and kind of nodes we wish to use according to how complex we want the representation of the document collection or the information needs to be. Arcs represent probabilistic dependencies between entities. They represent conditional probabilities, that is, the probability of an entity being true given the probabilities of its parents being true.

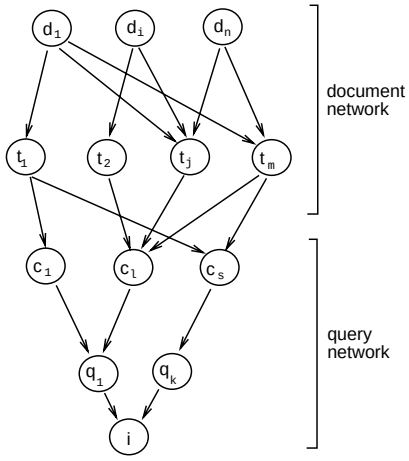


**Figure 1.** An inference network for IR.

The inference network is usually made up of two component networks: a document network and a query network. The document network represents the document collection. It is built once for a given collection and its structure does not change. A query network is built for each information need and can be modified and extended during each session by the user in a interactive and dynamic way. The query network is attached to the static document network in order to process a query.

In a Bayesian inference network, the truth value of a node depends only upon the truth values of its parents. To evaluate the strength of an inference chain going from one document to the query we set the document node $d_i$ to "true" and evaluate $P(q_k = true \mid d_i = true)$. This gives us an estimate of $P(d_i \rightarrow q_k)$. It is possible to implement various traditional IR models on this network by introducing nodes representing Boolean operators or by setting appropriate conditional probability evaluation functions within nodes [100].

One particular characteristic of this model that warrants exploration is that multiple document and query representations can be used within the context of a particular document collection (e.g., a Boolean expression or a vector). Moreover, given a single information need, it is possible to combine results from multiple queries and from multiple search strategies.

The strength of this model comes from the fact that most classical retrieval models can be expressed in terms of a Bayesian inference network by estimating in different ways the weights in the inference network [100]. Nevertheless, the characteristics of the Bayesian inference process itself, given that nodes (evidence) can only be binary (the evidence is either present or not) limits its use to where "certain evidence" [68] is available. The approach proposed by van Rijsbergen in [105], which makes use of "uncertain evidence" by using Jeffrey's conditioning [50], therefore appears more attractive.

Other approaches to the use of Bayesian inference networks in IR are presented in [38, 82, 92].

### 4.2 Models Based on Probabilistic Datalog

This area of research is based on the fact that logical IR retrieval could be viewed as a generalisation of logical database retrieval; uncertainty is introduced in the former. Datalog is a predicate logic that has been developed in the database field, and makes the link between the relational model and rule-based systems. *Probabilistic Datalog* is the probabilistic extension of Datalog [37]. For example:

$$0.7\ term(d_1, IR)$$
$$0.8\ term(d_1, DB)$$

represent two facts $term(d_1, IR)$ and $term(d_1, DB)$ they indicate that the document $d1$ is indexed by "IR" and "DB"; the weights represent the probability that the two facts are true. Retrieving documents that deal with both of these topics can be expressed by the following inference rule:

$$q_1(D) : -term(D, IR) \wedge term(D, DB)$$

If term dependence is assumed (i.e., $P(a \wedge b) = P(a) \cdot P(b)$), then the document $d1$ is retrieved with probability 0.56. If we want to retrieve every document about "IR" or "DB," the following query can be used:

$$q_2(D) : -term(D, IR)$$
$$q_2(D) : -term(D, DB)$$

The above document $d_1$ is retrieved with probability 0.94. This comes from that $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$ and assuming term independence.

The rule-based approach allows for easy formulation of various retrieval models and advanced IR systems. Consider the example of a hypermedia system. We may want to express that a document is about a term if that term indexes the document, or if the document is linked to one that is indexed by the term. This can be expressed by the following rule:

$$about(D, T) : -term(D, T)$$
$$about(D, T) : -link(D, D1) \wedge term(D1, T)$$

Consider the following query:

$$q_3 : -about(D, IR)$$

The query is looking for document directly or indirectly about "IR." The use of Probabilistic Datalog to model retrieval in hypermedia has been presented in [85]

Probabilistic Datalog is not itself a logical IR framework, but more a platform in which logical probabilistic IR models, as well as other IR models can be expressed (see for example [28]).

One of the major assets of Probabilistic Datalog is that, since it is a generalisation of the Datalog model, it can be used as standard query language for both database and IR. It can then deal with both structured data (as in database) and unstructured data (as in IR) within the same system. It also allows the uniform representation, retrieval and querying of content, fact and structural knowledge.

The work has been further extended via the development, implementation and evaluation of the POOL (Probabilistic Object-Oriented Logic) model, which allows the representation of inconsistency (using then a four-valued logic), the modelling of the document and query representation using the object-oriented paradigm [84, 86]. The result is a running IR platform, called HySpirit, that is currently being further developed and commercialised for large volumes of data [4].

### 4.3 Models Based on Logical Imaging

Several models have been developed based on the frameworks of imaging [60] and general imaging [39]. *Logical imaging* is an approach that defines the probability of conditional $P(d \rightarrow q)$ based on the notion of possible-worlds. In this

[4] See `http://www.hyspirit.de/`

approach the possible worlds (e.g., retrieval situation, document representation) are spanned by an accessibility relation defined in terms of similarity. The truth value of the implication $p \rightarrow q$ in a world $w$ depends on two cases. If $p$ is true in $w$, then $p \rightarrow q$ is true (false) in that world if $q$ is also true (false) in that world. In the other hand, if $p$ is not true in $w$, then the implication is evaluated in the worlds that differ minimally from $w$ and in which $p$ is true. The worlds in which $p$ is true are referred to as $p$-worlds.

The set of worlds comes with a probability distribution $P$, reflecting the probability of each world. The probability of a proposition $p$ is the summation of the probability of those worlds in which $p$ is true. The computation of the probability of $p \rightarrow q$ involves a shift of probability (the imaging process) from non-$p-worlds$ to their closest $p$-worlds. It can be proved that [60]:

$$P(p \rightarrow q) = P_p(q)$$

where $P_p$ is a new probability distribution, a "posterior probability," derived from $P$ by imaging on $p$.

Conditionalisation by imaging causes a revision of the prior probability on the possible worlds $w$ in such a way that the posterior probability is obtained by shifting the original probabilities from non-$p$-worlds to $p$-worlds. Each non-$p$-world moves its probability to its closest $p$-world (or set of $p$-worlds in the case of general imaging). Probability is neither created or destroyed, but just moved according to the accessibility relation (and to the opinionated probability function, in the case of general imaging). Bayesian conditionalisation, on the other hand, is obtained by cutting off all non-$p$-worlds and then proportionally magnifying the probabilities of the $p$-worlds so that the posterior probabilities still add up to 1, as required by Probability Theory. The magnification is done in the same way for every $p$-world, thus keeping constant the ratios between the probabilities assigned to these worlds. It is therefore clear that imaging and Bayesian conditionalisation yield, in general, different results [40].

However,since the transfer of probabilities is directed towards the closest $p$-worlds, this technique is just what it is needed to implement Van Rijsbergen's logical uncertainty principle. If $d$ represents the document and $q$ the query, then the relevance of the document to the query can be evaluated as:

$$P(d \rightarrow q) = P_d(q) = \sum_{t \in q} P_d(t)$$

by imaging on the document. This formula is compatible with Van Rijsbergen's logical uncertainty principle since the probability revision is minimal with regard to the accessibility relation. Alternatively, it is possible to evaluate:

$$P(q \rightarrow d) = P_q(d) = \sum_{t \in d} P_q(t)$$

by imaging on the query. Nie showed in [70] that the two conditionals $d \rightarrow q$ and $q \rightarrow d$ have a very interesting interpretation in the context of IR. The conditional $d \rightarrow q$ expresses the "exhaustivity" of the document to a query, i.e. how much of a document content is specified by the query content. In fact, $d \rightarrow q$ is intuitively equivalent to $d \subseteq q$. The conditional $q \rightarrow d$, on the other hand, expresses the "specificity" of a document to a query, i.e. how much of a query content is specified in the document content. In fact, $q \rightarrow d$ is intuitively equivalent to $q \subseteq d$. The choice of either $d \rightarrow q$ or $q \rightarrow d$ depends on the particualr requirements of the application, that is, if the application requires high recall or high precision.

The application of imaging to IR requires that:

- in any given possible world, a sentence is either true or false,
- possible worlds are more or less similar to each other (imaging uses this similarity), and
- the set of possible worlds is finite.

From an IR perspective, the second and third assumptions are acceptable. The first one, however, is problematic. For example, consider that a possible world and a sentence represent a document and an index term, respectively. The uncertainties inherent in the indexing process make a true-false assignment to index terms (sentences) within documents (worlds) an error prone task.

To build an IR system based on imaging, the similarity relation between worlds must be available. If documents correspond to worlds, then a similarity measure between documents must be computed. In practice, this measure is computed using not the documents, but their representations. As a consequence, the integrity of the similarity relation is compromised as, in practice, a document representation is an incomplete reflection of the actual document content. This is due to limitations in automatic indexing algorithms. For example, in a primitive system, the representations of a document on "Gone with the Wind" and a document on "Meteorology and Wind" may be considered fairly similar because they both contain the keyword "wind," whereas in reality the two documents are different. As a consequence, there are errors inherent in the calculation of $P(d \rightarrow q)$ when using document representation as input to the similarity calculation. Nevertheless, this approach gives a conceptually neat and concise realization of the logical uncertainty principle. Moreover, it provides a direct route for calculating $P(d \rightarrow q)$, which can be used to rank documents on likelihood of their relevance to the user. Imaging sidesteps the troublesome question of the semantics of the implication via its probabilistic basis.

General imaging [39] relaxes two earlier assumptions by asserting that: (i) truth assignments of sentences in worlds need not be true or false, and (ii) the most similar world need not be unique.

Two logical-probabilistic IR models have been developed on the concept of imaging. In the first one [31, 32], worlds model terms, and propositions model documents and queries. A term $t$ "makes a document true" if that term belongs to that document. Imaging with respect to $d$ gives the closest term to $t$ that

is contained in $d$. Of course, this is $t$ itself if $t$ is contained in the document. Imaging consists then of shifting the probabilities from term not contained in $d$ to the terms contained in $d$ (i.e., the terms that make $d$ true). The evaluation of the relevance takes into account the semantics between terms by shifting probabilities to those (semantically) closer terms contained in the document [30]. Models based on imaging and general imaging has been successfully implemented and tested on some standard test collections [32], but results with larger scale testing were inconclusive [29].

A second model based on imaging includes user's knowledge in the evaluation of the relevance of a document to a query [75]. In this model both documents and queries are propositions. Possible worlds represent different states of the data set, for example possible states of knowledge that can be held by users. A document $d$ is true in a world $w$ if the document is "consistent" (the term is used here in a broad sense) with the state of knowledge associated with that world. Worlds differ because they represent different states of knowledge and, given a metric on the world space, we can identify the closest world to $w$ for which $d$ is true. This model has the advantage that user modelling is formally included, while the model presented in [32] and detailed above only takes into account a system evaluated relevance. However, no implementation and evaluation of this model has been produced so far.

### 4.4 Models Based on Semantic Information Theory

*Semantic Information Theory* is concerned with studies in Logic and Philosophy on the use of the term information, "in the sense in which it is used of whatever it is that meaningful sentences and other comparable combinations of symbols convey to one who understands them" [42]. Notwithstanding the large scope of this description, Semantic Information Theory has primarily to do with the question of how to weigh sentences according to their informative content. The main difference with conventional information theory is that information is not conveyed by an ordered sequence of binary symbols, but by means of a formal language in which logical statements are defined and explained by a semantics. The work on Semantic Information Theory in IR concerns two research directions: the axiomatisation of the logical principles for assigning probabilities or similar weighting functions to logical sentences and the relationship between information content of a sentence and its probability. Both directions were investigate in [3, 4].

In [3], it is argued that the notion of amount of information content in IR is determined by some entropy measures, especially by one axiomatised first by Hilpinen. By using different utility functions which combine entropy and probability, several old and new models of IR are derived. In addition, the principles of a "duality theory" are presented. Retrieval based on probability theory requires the definition of an event space. When one deals with probabilities, one measures a Boolean, or a sigma algebra, of events. According to the duality theory, used also in [32] to give a representation semantics to logical imaging in the context of IR, one can either consider the set of terms $T$ of the term space as the set of the

elementary events, or consider the set $D$ of documents as the set of elementary events. In the first case one can work on a term probability space, in the second on a document probability space. Working on the term probability space leads on to the standard application of probability theory to IR. On the other hand, working on the document probability space it is possible to show how to tightly link the vector space model to the standard probabilistic model.

In [4], the concepts of simplicity, regularity, randomness, and shortest description length are studied with the purpose of formalising the informative content of documents. It is shown that the Zipf's law and the inverse document frequency weight are derived from first principles involving these concepts.

### 4.5 Models Based on Probabilistic Argumentation Systems

One basic assumption of probabilistic models of IR is that terms and documents are independent. Although this assumption has been reshaped in recent times [22], a number of investigations have demonstrated the potential usefulness of incorporating dependencies or relationships between terms (for example [80, 101]) and between documents (for example [33, 93]). When attempting to incorporate these features in the matching process, the limits of the traditional term matching approach appear more clearly, and one is often left to the use of ad-hoc schemes. For this reason and others, it has been argued by a number of researchers that the best way to model the IR process is by the use of an appropriate logic that captures these features. In fact, the expressiveness of logic makes it a very attractive framework for modelling relationships between terms and/or documents, but the complexity of its implementation makes it difficult to have large-scale applications. A possible way out of this "impasse" could be found in integrating logic in large-scale classic IR systems as a tool for solving specific problems which cannot be formalised within more conventional approaches. The different steps of the retrieval process could be done as usual, and logic could be used as a formal tool for modifying the output of certain components of the retrieval system. In this way the "logical components" could be integrated to any retrieval system working on soft term matching, e.g. the vector-space or probabilistic models, even in large-scale applications.

The work described in [79] starts from this premise and from the consideration that IR can be seen both as an inference process under uncertainty involving complex relationships between information items, and as a task of proper assessment of uncertainty. The originality of this work is in the choice of the framework used: *probabilistic argumentation systems*. Probabilistic argumentation systems provide techniques for reasoning under uncertainty which emphasise both the inference process and the assessment of uncertainty, by clearly distinguishing the qualitative and quantitative aspects of uncertainty. Probabilistic argumentation systems represent uncertainty in a clear and easily understandable way: the qualitative part is handled with propositional logic and the quantitative part is treated with probability theory. They offer a natural way to model relationships between terms and between documents, and allows complex inferences. In [79]

two applications of probabilistic argumentation systems to IR are presented, aimed at:

(1) taking into account existing hypertext links in order to improve an initial ranking of documents
(2) considering statistical similarities between query terms to improve query weighting.

These two applications can be easily integrated in a classic IR system based on term matching, even for large-scale applications [79]. It is worth noticing that even though the symbolic part of uncertain knowledge is naturally modelled with probabilistic argumentation systems, the numerical assessment of probabilities is often a very difficult problem. It is clear that if the uncertainty is incorrectly assessed, combined or propagated, the inference carried out by the logic will very probably be unable to improve retrieval effectiveness. The transformation of easily estimated statistical similarity information into a measure of uncertainty (e.g. probabilities) is one major difficulties and is still an open problem. Nevertheless, the work presented in [79] highlights that in IR the numerical and symbolic aspects of uncertainty are profoundly interlaced, and a purely symbolic or numerical approach would not bring the same insight in these problems. The theoretical foundations of probabilistic argumentation systems, which rely on the theory of evidence, make them a reliable technique for approaching problems in which the quantitative and qualitative aspects of uncertainty are of equal importance.

## 5 Meta-Models

*Meta-models* are a completely different class of models from those presented earlier. Meta-models attempts to formally study the properties and the characteristics of IR systems within a uniform logical framework [89, 107]. The advantage is that it will then be possible to compare IR systems not only with respect to their effectiveness, but also with respect to formal properties of the underlying models. For example, an application may require a system that retrieves all relevant documents (recall-oriented system), or that retrieves only relevant documents (precision-oriented system).

The use of logic to formally conduct proofs for IR purposes originated in [69], where it was showed that a logical model is a general form of many other IR models. The idea was later thoroughly investigated in [10, 47], where a framework was proposed in which different models of IR could be theoretically expressed, formally studied and compared. The framework was developed within a logic, thus allowing formal proofs to be conducted. Here meta-models are based on non-monotonic reasoning approaches.

The framework defines the aboutness relationship, denoted $\models$, which aims at capturing the notion of information containment primary to IR. Given two objects $a$ and $b$, $a \models b$ means that object $a$ is about object $b$. Axioms are defined that represent possible properties of IR systems. Examples of axioms include:

- Reflexivity: $a \models a$
- Symmetry: if $a \models b$ then $b \models a$
- Transitivity: if $a \models b$ and $b \models c$ then $a \models c$

Most IR models seem to satisfy reflexivity because if objects are documents, then if a document is submitted as a query, then this document should be retrieved. However, if only new documents should be retrieved in response to a query, then another aboutness relationship should be defined, one which is not reflexive. The IR models satisfying symmetry are the vector space model, and those based on overlap measures such as Jaccard's, or Dice's. The models based on overlap measures were shown to not satisfy transitivity.

A rule often used by meta-models and borrowed from non-monotonic reasoning is [14]:

- Right weakening: if $a \models b$ and $b \Rightarrow c$ then $a \models c$[5]

An aboutness relationship that satisfies the above rule can be precision degrading. For example, a document about "microbes" should probably not be retrieved in response to a query about "animal" although $microbe \Rightarrow animal$. This is because most preferred documents about animals would presumably deal with "birds," "dogs," and similar.

Monotonicity is another rule being investigated. Consider for example the left monotonicity rule:

- Left Monotonicity: if $a \models b$ then $a \oplus c \models a$, where $a \oplus c$ is an item yielded from the combination of $a$ and $c$

Where the combination is not restricted to the logical "and". Clearly the above rule produces inferences that are not sound (e.g. take $a = surfing$, $b = wave$ and $c = internet$). It is however not a solution to drop the rule, for example for expressing the query expansion process. Cautious monotonicity rules have been investigated for that purpose.

This research has led to the theoretical comparison of IR models [10, 44, 46]. This proceeds as follows. Two IR models are mapped down into a logic-based framework. The aboutness properties that a given model supports is then filtered out. The two models can be compared by using the particular aboutness properties they each embody. For example, if retrieval system $A$ supports a monotonic notion of aboutness and retrieval system $B$ does not, then this may suggest that system $B$ will offer more precise retrieval than $A$. Boolean retrieval and (strict) coordinate retrieval have been compared in this fashion. Boolean retrieval represents a document $d$ and query $q$ as propositional formulae. Document $d$ is deemed to be about query $q$ iff $d \models q$. Strict coordinate retrieval represents $d$ and $q$ as sets of index terms. Document $d$ is deemed to be about $q$ iff $d \subseteq q$.

The relationship between particular aboutness properties and IR effectiveness is currently an open problem. Also, which underlying logic-based framework

[5] Here $b \Rightarrow c$ means that $c$ follows from $b$ or that $c$ is informationally contained in $b$.

should be used? More investigation and experience with these frameworks may lead to an integrated, underlying theory of IR. Nevertheless, through this research, IR is gaining a clearer understanding of what aboutness is, and what properties are desirable (and not desirable) for this notion [11, 12].

## 6 Conclusions and Future Directions

This paper describes some of the work done in developing logical and logical-uncertainty models for IR. The aim in using logic to model IR is to provide expressive and uniform IR models not only to improve effectiveness, but also to have a framework where the semantics of the retrieval process can be formally described and investigated. Uncertainty theories enable to take into account the uncertainty inherent in such a formulation.

We have presented in this paper a number of approaches that address different issues. The variation in approaches reflects different vehicles deemed suitable for the modelling of IR. So far, no consensus has been reached regarding what the best vehicle is. Investigations into various logic-based frameworks will hopefully lead to a unified information-based model theory for expressing the semantics of information retrieval. Such a theory will allow us to predict the behaviour of IR systems, compare them and prove properties about them. This, we believe, is one of the major strength of logical models.

Another major asset of logical and logical-uncertainty models is that they are able to represent within a uniform framework various features of advanced IR systems, such as hypermedia objects, structured multimedia documents, users knowledge and agents. One main advantage in having a general model in IR is that it becomes possible to reason about various IR features of the model. This possibility is becoming increasingly important because conventional evaluation methods such as experimentally measuring precision and recall are sometimes insufficient.

Finally, the use of logic and uncertainty theory in IR is still in its early stages. Substantial theoretical progress has been made, but further investigation and development are required before the effectiveness of these models can be established. For instance, the implementation of logical models can be complex, and when possible, often only small document collections can be handled. Despite some implementations have provided positive results [32, 74], more experimental work is necessary to demonstrate the effectiveness of logical and logical-uncertainty models.

## Note

This paper is heavily based on some introductory and survey papers on the use of logic and uncertainty theory in IR that we already published, in particular on [27, 56,57]. In addition, large part of the work on logic and uncertainty in IR reviewed in this paper can be found in *Information Retrieval: Uncertainty and Logics*

edited by Crestani, Lalmas, and van Rijsbergen [24] and in the proceedings of the "Workshop on Logical and Uncertainty Models for Information Systems" [15,23].

## References


1. G. Amati and P.D. Bruza. A logical approach to query reformulation motivated from belief change. In *Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems*, pages 36–45, London, UK, July 1999.
2. G. Amati and F. Crestani. Probabilistic learning by uncertain samplying with non-binary relevance. In F. Crestani and G. Pasi, editors, *Soft Computing in Information Retrieval: techniques and application*, pages 292–314. Physica-Verlag, Heidelberg, Germany, 2000.
3. G. Amati and C.J. van Rijsbergen. Semantic Information Retrieval. In F. Crestani, M. Lalmas, and C.J. van Rijsbergen, editors, *Information Retrieval: Uncertainty and Logics*, pages 189–220. Kluwer Academic Publishers, Norwell, MA, USA, 1998.
4. G. Amati and C.J. van Rijsbergen. Simplicity and Information Retrieval. In F. Crestani, M. Lalmas, and C.J. van Rijsbergen, editors, *Information Retrieval: Uncertainty and Logics*, pages 281–293. Kluwer Academic Publishers, Norwell, MA, USA, 1998.
5. Z. An, A. Bell, and J.G. Hughes. Res - a logic for relative evidential support. *International Journal of Approximate Reasoning*, 8:205–230, 1993.
6. J. Barwise. *Handbook of Mathematical Logic*. Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 8th edition, 1993.
7. R.K. Belew. Rave reviews: acquiring relevance assessments from multiple users. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, CA, USA, March 1996.
8. F. C. Berger and T. W. C Huibers. A framework based on situation theory for searching on a thesaurus. In J. Rowley, editor, *The New Review of Document and text Management*, volume 1, pages 253–276, Crewe, England, 1995.
9. A. W. Black. *A Situation Theoretic Approach to Computational Semantics*. PhD thesis, University of Edinburgh, 1992.
10. P. D. Bruza and T. W. C. Huibers. Investigating aboutness axioms using information fields. In *Proceedings of ACM SIGIR*, pages 112–121, Dublin, Ireland, 1994.
11. P. D. Bruza and T. W. C Huibers. How monotonic is aboutness? Technical report, Utrecht University, The Netherlands, 1995. Technical Report UU-CS-1995-09.
12. P. D. Bruza and T. W. C Huibers. A study of aboutness in information retrieval. *Artificial Intelligence Review*, 10:1–27, 1996.
13. P.D. Bruza. *Stratified Information Disclosure: a synthesis between Hypermedia and Information Retrieval*. PhD Thesis, Katholieke Universiteit Nijmegen, The Netherlands, 1993.
14. P.D. Bruza. Intelligent filtering using nonmonotonic inference. In *Proceedings of the Australian Document Computing Symposium*, pages 1–7, Royal Melbourne Institute of Technology, Melbourne, Australia, 1996.
15. P.D. Bruza, F. Crestani, and M. Lalmas. Second Workshop on Logical and Uncertainty Models for Information Systems (DEXA-LUMIS 2000). In *Proceedings of DEXA 2000*. IEEE Press, Greenwich, London, UK, 2000.

16. P. Buongarzoni, C. Meghini, R. Salis, F. Sebastiani, and U. Straccia. Logical and computational properties of the description logic MIRTL. In *Proceedings of DL 95*, pages 80–84, Rome, Italy, 1995.
17. J. P. Chevallet. *Un modèle logique de recherche d'information appliqué au formalisme des graphes conceptuels. Le prototype ELEN et son expérimentation sur un corpus de composants logiciels.* PhD thesis, Université Joseph Fourier, Grenoble I, 1992.
18. Y. Chiaramella and J.P. Chevallet. About retrieval models and logic. *The Computer Journal*, 35(3):233–242, 1992.
19. Y. Chiaramella, P. Mulhem, and F. Fourel. A model for multimedia information retrieval. Technical report, ESPRIT Basic Research Action, Project Number 8134 - FERMI, Department of Computing Science, Glasgow University, Glasgow, UK, 1996.
20. Y. Chiaramella and J. Nie. A retrieval model based on an extended Modal Logic and its application to the RIME experiment approach. In *Proceedings of ACM SIGIR*, pages 25–43, Brussels, Belgium, 1990.
21. W.S. Cooper. A definition of relevance for Information Retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.
22. W.S. Cooper. Some inconsistencies and misnomers in probabilistic Information Retrieval. *ACM Transactions on Information Systems*, 13(1):100–111, 1995.
23. F. Crestani and M. Lalmas, editors. *Proceedings of the First Workshop on Logical and Uncertainty Models for Information Systems (LUMIS 99)*, London, UK, July 1999. Available online at: `http://www.dcs.gla.ac.uk/lumis99/`.
24. F. Crestani, M. Lalmas, and C.J. van Rijsbergen, editors. *Information Retrieval: Uncertainty and Logics.* Kluwer Academic Publisher, Norwell, MA, USA, 1998.
25. F. Crestani, M. Lalmas, C.J. van Rijsbergen, and I. Campbell. Is this document relevant? ... probably. A survey of probabilistic models in Information Retrieval. *ACM Computing Surveys*, 30(4):528–552, 1998.
26. F. Crestani and G. Pasi. Soft Information Retrieval: applications of fuzzy sets theory and neural networks. In N. Kasabov and R. Kozma, editors, *Neuro-fuzzy Techniques for Intelligent Information Systems*, pages 287–315. Physica Verlag, Heidelberg, Germany, 1999.
27. F. Crestani and G. Pasi, editors. *Soft Computing in Information Retrieval: techniques and applications.* Physica-Verlag, Heidelberg, Germany, 2000.
28. F. Crestani and T. Rölleke. Issues on the implementation of imaging on top of probabilistic datalog. In *Proceedings of the First Workshop in IR, Uncertainty and Logic.* Glasgow, Scotland, UK, September 1995.
29. F. Crestani, I. Ruthven, M. Sanderson, and C.J. van Rijsbergen. The troubles with using a logical model of IR on a large collection of documents. Experimenting retrieval by logical imaging on TREC. In *Proceedings of the TREC Conference*, pages 509–525, Washington D.C., USA, November 1995.
30. F. Crestani, M. Sanderson, and C.J. van Rijsbergen. Sense resolution properties of logical imaging. *The New Review of Document and Text Management*, 1:277–298, 1996.
31. F. Crestani and C.J. van Rijsbergen. Information Retrieval by Logical Imaging. *Journal of Documentation*, 51(1):1–15, 1995.
32. F. Crestani and C.J. van Rijsbergen. A study of probability kinematics in rnformation retrieval. *ACM Transactions on Information Systems*, 16(3):225–255, 1998.

33. W.B. Croft and R.H. Thompson. $I^3R$: a new approach to the design of Document Retrieval Systems. *Journal of the American Society for Information Science*, 38(6):389–404, 1987.
34. W.T. da Silva and R.L. Milidiú. Belief function model for Information Retrieval. *Journal of the American Society for Information Science*, 44(1):10–18, 1993.
35. K. Devlin. *Logic and Information*. Cambridge University Press, Cambridge, UK, 1991.
36. N. Fuhr. Probabilistic models in Information Retrieval. *The Computer Journal*, 35(3):243–254, 1992.
37. N. Fuhr. Probabilistic Datalog - a logic for powerful retrieval methods. In *Proceedings of ACM SIGIR*, pages 282–290, Seattle, WA, USA, 1995.
38. R. Fung and B. Del Favero. Applying bayesian networks to Information Retrieval. *Communications of the ACM*, 38(3):42–48, 1995.
39. P. Gärdenfors. Imaging and conditionalization. *Journal of Philosophy*, 79:747–760, 1982.
40. P. Gärdenfors. *Knowledge in flux: modelling the dynamics of epistemic states*. The MIT Press, Cambridge, Massachusetts, USA, 1988.
41. P. Gärdenfors, editor. *Belief Revision*. Cambridge University Press, Cambridge, UK, 1992.
42. J. Hintikka. On semantic information. In *Information and inference*. Synthese Library, Reidel, Dordrecht, The Netherlands, 1970.
43. G.E. Hughes and M.K. Cresswell. *An Introduction to Modal Logic*. Muthuen and Co. Ltd, London, UK, 1968.
44. T. Huibers, I. Ounis, and J. P. Chevallet. Axiomatization of a conceptual graph formalism for information retrieval in a situated framework. Technical Report RAP95-004, Group MRIM of the Laboratoire de Génie Informatique, Grenoble, France, 1995.
45. T. W. C. Huibers and P. D. Bruza. Situations, a general framework for studying information retrieval. In *Proceedings of the 16th British Computer Society Colloquium in Information Retrieval*, Drymen, Scotland, UK, March 1994.
46. T. W. C Huibers and N. Denos. A qualitative ranking method for logical information retrieval models. Technical Report RAP95-005, Groupe MRIM of the Laboratoire de Génie Informatique, Grenoble, France, 1995.
47. T.W.C Huibers. *An Axiomatic Theory for Information Retrieval*. PhD thesis, Utrecht University, The Netherlands, 1996.
48. A. Hunter. Intelligent text handling using default logic. In *Proceedings of IEEE Conference on Tools with Artificial Intelligence*, 1996. (to appear).
49. A. Hunter. Using default logic for lexical knowledge. In *Qualitative and Quantitative Practical Reasoning (ECSQARU'97/FAPR'97)*. Springer-Verlag, Heidelberg, Germany, 1997.
50. R.C. Jeffrey. *The logic of decision*. McGraw-Hill, New York, USA, 1965.
51. N. Kasabov and R. Kozma, editors. *Neuro-fuzzy techniques for intelligent information systems*. Physica Verlag, Heidelberg, Germany, 1998.
52. S.A. Kripke. Semantical considerations on modal logic. In L. Linsky, editor, *Reference and modality*, chapter 5, pages 63–73. Oxford University Press, Oxford, UK, 1971.
53. M. Lalmas. From a qualitative towards a quantitative representation of uncertainty on a situation theory based model of an information retrieval system. Technical report, Department of Computing Science, Technical Report TR-1995-18, University of Glasgow, Scotland, 1995.

54. M. Lalmas, editor. *Proceedings of the First International Workshop on Information Retrieval, Uncertainty and Logics*, Glasgow, Scotland, UK, July 1995.
55. M. Lalmas. Modelling Information Retrieval with Dempster-Shafer's theory of evidence: a study. In *Proceedings of the ECAI Workshop on Uncertainty in Information Systems: questions of viability*, Budapest (Hungary), September 1996.
56. M. Lalmas. Logical models in Information Retrieval: introduction and overview. *Information Processing & Management*, 34(1):19–33, 1998.
57. M. Lalmas and P.D. Bruza. The use of logic in information retrieval modelling. *Knowledge Engineering Review*, 13(2):19–33, 1998.
58. M. Lalmas and I. Ruthven. Representing and retrieving structured documents with Dempster-Shafer's theory of evidence: Modelling and evaluation. *Journal of Documentation*, 54(5):529–565, 1998.
59. M. Lalmas and C.J. van Rijsbergen. 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*, pages 62–67, Montreal, Canada, 1993.
60. D. Lewis. *Conterfactuals*. Basil Blackwell, Oxford, UK, 2nd edition, 1986.
61. D.E. Losada and A. Barreiro. Using a belief revision operator for document ranking in extended boolean model. In *Proceedings of ACM SIGIR*, pages 66–73, Berkeley, CA, USA, 1999.
62. D.E. Losada and A. Barreiro. Efficient algorithms for ranking documents. In *Proceedings of SIGIR Workshop on Formal/Mathematical Methods for Information Retrieval*, pages 16–24, Athens, Greece, 2000.
63. D.E. Losada and A. Barreiro. Retrieval situations and belief changes. In *Proceedings of DEXA-LUMIS 2000*, Greenwich, London, UK, 2000.
64. C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of Information Retrieval based on a Terminological Logic. In *Proceedings of ACM SIGIR*, pages 298–307, Pittsburgh, PA, USA, June 1993.
65. C. Meghini and U. Straccia. A relevance terminological logic for information retrieval. In *Proceedings of ACM SIGIR*, Zurich, CH, August 1996.
66. G. A. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1990.
67. A. Müller. A flexible framework for multimedia Information Retrieval. In F. Crestani, M. Lalmas, and C.J. van Rijsbergen, editors, *Information Retrieval: Uncertainty and Logics*, pages 97–128. Kluwer Academic Publishers, Norwell, MA, USA, 1998.
68. R.E. Neapolitan. *Probabilistic reasoning in expert systems*. John Wiley and Son Inc., New York, USA, 1990.
69. J. Y. Nie. *Un Modèle de Logique Générale pour les Systemes de Recherche d'Informations. Application au Prototype RIME*. PhD Thesis, Université Joseph Fourrier, Grenoble, France, 1990.
70. J.Y. Nie. An outline of a general model for Information Retrieval. In *Proceedings of ACM SIGIR*, pages 495–506, Grenoble, France, June 1988.
71. J.Y. Nie. An Information Retrieval model based on Modal Logic. *Information Processing & Management*, 25(5):477–491, 1989.
72. J.Y. Nie. Towards a probabilistic modal logic for semantic based Information Retrieval. In *Proceedings of ACM SIGIR*, pages 140–151, Copenhagen, Denmark, June 1992.
73. J.Y. Nie. CLIR and query expansion as logical inference. In *Proceedings of SIGIR Workshop on Formal/Mathematical Methods for Information Retrieval*, pages 8–15, Athens, Greece, 2000.

74. J.Y. Nie and M. Brisebois. An inferential approach to information retrieval and its implementation using a manual thesausrus. *Artificial Intelligence Review*, 10:409–439, 1996.
75. J.Y. Nie, F. Lepage, and M. Brisebois. Information Retrieval as counterfactuals. *The Computer Journal*, 38(8):643–657, 1995.
76. I. Ounis. *Un modele d'indexation relationnel pour les graphes conceptuels fonde sur une interpretation logique*. PhD Thesis, Université Joseph Fourrier, Grenoble I, 1998.
77. P. F. Patel-Schneider. A four-valued semantics for frame-based description languages. In *AAAI-86, 5th Conference of the American Association for Artificial Intelligence*, pages 344–348, Philadelphia, 1986.
78. J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo, California, 1988.
79. J. Picard. Logic as a tool in a term matching information retrieval system. In *Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems*, pages 77–90, London, UK, July 1999.
80. Y. Qiu and H.P. Frei. Concept based query expansion. In *Proceedings of ACM SIGIR*, pages 160–171, Pittsburgh, PA, USA, June 1993.
81. R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1):81–132, 1980.
82. B. Ribeiro-Neto, I. Silvia, and R. Muntz. Bayesian network models for Information Retrieval. In F. Crestani and G. Pasi, editors, *Soft Computing in Information Retrieval: techniques and application*, pages 259–291. Physica-Verlag, Heidelberg, Germany, 2000.
83. S.E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, December 1977.
84. T. Rölleke. *POOL: Probabilistic Object-Oriented Logical Representation and Retrieval of Complex Objects - A Model for Hypermedia Retrieva*. PhD Thesis, Department of Computer Science, University of Dortmund, Germany, 1999.
85. T. Rölleke and M. Blömer. Probabilistic logical Information Retrieval for content, hypertext and database querying. In *Proccedings of HIM Conference*, Dortmund, Germany, September 1997.
86. T. Rölleke and N. Fuhr. Retrieval of complex objects using a four-valued logic. In *Proceedings of ACM SIGIR*, pages 206–214, Zurich, Switzerland, 1996.
87. I. Ruthven, M. Lalmas, and C.J. van Rijsbergen. Retrieval through explanation: Inference approach to relevance feedback. In *Proceedings of 10th Annual Irish Conference on Artificial Intelligence & Cognitive Science (AICS)*, Cork, Ireland, 1999.
88. G. Salton. *Automatic information organization and retrieval*. McGraw Hill, New York, 1968.
89. S. Dominich. Formal foundation of classical information retrieval. In *Proceedings of SIGIR Workshop on Formal/Mathematical Methods for Information Retrieval*, pages 69–75, Athens, Greece, 2000.
90. S. Dominich, M. Lalmas, and C.J. van Rijsbergen. SIGIR Workshop on Formal/Mathematical Methods for Information Retrieval. *Technology Letters*, 4(1), 2000.
91. T. Saracevic. The concept of "relevance" in information science: a historical review. In T. Seracevic, editor, *Introduction to Information Science*, chapter 14. R.R. Bower Company, New York, USA, 1970.
92. J. Savoy. Bayesian inference networks and spreading activation in hypertext systems. *Information Processing & Management*, 28(3):389–406, 1992.

93. J. Savoy. A learning scheme for Information Retrieval in hypertext. *Information Processing & Management*, 30(4):515–533, 1994.
94. F. Sebastiani. A probabilistic terminological logic for modelling Information Retrieval. In *Proceedings of ACM SIGIR*, pages 122–131, Dublin, Ireland, 1994.
95. F. Sebastiani. On the role of logics in Information Retrieval. In *Proceedings of the MIRO Workshop*, Glasgow, September 1995.
96. J.F. Sowa. *Conceptual structures: information processing in mind and machine.* Addison-Wesley Publishing Company, Reading, MA, USA, 1984.
97. U. Thiel and A. Müller. Why was this item retrieved?: new ways to explore retrieval results. In M. Agosti and A.F. Smeaton, editors, *Information Retrieval and Hypertext*, chapter 8, pages 181–201. Kluwer Academic Publishers, Dordrecht, NL, 1996.
98. H.R. Turtle and W.B. Croft. Inference networks for document Retrieval. In *Proceedings of ACM SIGIR*, Brussels, Belgium, September 1990.
99. H.R. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.
100. H.R. Turtle and W.B. Croft. A comparison of text retrieval models. *The Computer Journal*, 35(3):279–290, 1992.
101. C.J. van Rijsbergen. A theoretical basis for the use of co-occurence data in Information Retrieval. *Journal of Documentation*, 33(2):106–119, June 1977.
102. C.J. van Rijsbergen. A new theoretical framework for Information Retrieval. In *Proceedings of ACM SIGIR*, pages 194–200, Pisa, Italy, 1986.
103. C.J. van Rijsbergen. A non-classical logic for Information Retrieval. *The Computer Journal*, 29(6):481–485, 1986.
104. C.J. van Rijsbergen. Toward a new information logic. In *Proceedings of ACM SIGIR*, pages 77–86, Cambridge, USA, June 1989.
105. C.J. van Rijsbergen. Probabilistic retrieval revisited. *The Computer Journal*, 35(3):291–298, 1992.
106. C.J. van Rijsbergen and M. Lalmas. An information calculus for information retrieval. *Journal of the American Society of Information Science*, 47(5):385–398, 1996.
107. S.K.M. Wong and Y.Y. Yao. On modelling Information Retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, 1995.
108. L. A. Zadeh. *Fuzzy sets and Applications: Selected Papers.* Wiley, New York, 1987.