Modeling Information Retrieval by Formal Logic: A Survey

KARAM ABDULAHHAD, GESIS - Leibniz institute for the Social Sciences, Germany and University Grenoble Alpes - CNRS - LIG, France
CATHERINE BERRUT and JEAN-PIERRE CHEVALLET, University Grenoble Alpes - CNRS - LIG, France
GABRIELLA PASI, Università degli Studi di Milano Bicocca, Italy

Several mathematical frameworks have been used to model the information retrieval (IR) process, among them, formal logics. Logic-based IR models upgrade the IR process from document-query comparison to an inference process, in which both documents and queries are expressed as sentences of the selected formal logic. The underlying formal logic also permits one to represent and integrate knowledge in the IR process. One of the main obstacles that has prevented the adoption and large-scale diffusion of logic-based IR systems is their complexity. However, several logic-based IR models have been recently proposed that are applicable to large-scale data collections. In this survey, we present an overview of the most prominent logical IR models that have been proposed in the literature. The considered logical models are categorized under different axes, which include the considered logics and the way in which uncertainty has been modeled, for example, degrees of belief or degrees of truth. Accordingly, the main contribution of the article is to categorize the state-of-the-art logical models on a fine-grained basis, and for the considered models the related implementation aspects are described. Consequently, the proposed survey is finalized to better understand and compare the different logical IR models. Last, but not least, this article aims at reconsidering the potentials of logical approaches to IR by outlining the advances of logic-based approaches in close research areas.

CCS Concepts: • **Information systems → Retrieval models and ranking**;

Additional Key Words and Phrases: Formal logics, information retrieval models, logical models, survey, uncertainty

ACM Reference format:
Karam Abdulahhad, Catherine Berrut, Jean-Pierre Chevallet, and Gabriella Pasi. 2019. Modeling Information Retrieval by Formal Logic: A Survey. *ACM Comput. Surv.* 52, 1, Article 15 (February 2019), 37 pages.
https://doi.org/10.1145/3291043

1 INTRODUCTION

In their broader sense, information retrieval systems (IRSs) aim to fulfill users' information needs expressed in a keyword-based query. More precisely, information retrieval (IR) refers to the process of selection, from a document repository, of the documents *likely* to be relevant to a particular

Authors' addresses: K. Abdulahhad, Knowledge Technologies for the Social Sciences (WTS) Department, Leibniz institute for the Social Sciences (GESIS), Unter Sachsenhausen 6-8, 50667 Köln, Germany; C. Berrut, Laboratoire LIG, University Grenoble-Alpes, CS 40700, 38058 Grenoble Cedex 9, France; J.-P. Chevallet, Laboratoire LIG, University Grenoble-Alpes, CS 40700, 38058 Grenoble Cedex 9, France; G. Pasi, Department of Informatics, Systems and Communication (DISCo), Università degli Studi di Milano Bicocca, Viale Sarca 336, 20126 Milano, Italy.

0360-0300/2019/02-ART15 $15.00
https://doi.org/10.1145/3291043

information need, formulated by a query. Based on this definition, an IRS has to deal with a collection of documents, with users' information needs, and with the notion of relevance.

In IR, *documents* are carriers of information; in their original forms, they are human-understandable objects (e.g., Web pages, articles, books, and images), which an IRS must transform into machine-understandable objects. This process is called *indexing*, and its outcome is the association of a set of features (terms in textual documents) with documents. These features constitute the basic elements employed to formally represent a document. A user's *information needs* are motivated by a user's information gap; a *query* is a representation of these needs. Once formal representations have been provided for both documents and queries, the system compares them to assess the *relevance* of each document to the considered query.

Relevance is a complex notion composed of several dimensions, such as topicality, popularity, and novelty, as has been well pointed out in the literature [33, 65, 95, 96]. An IRS can only estimate relevance, and generally topicality is the core relevance dimension. The assessment of topical relevance[1] relies on the definition of a model, the IR model, which provides a formal means to represent and compare both documents and queries. In this survey we use the term *retrieval status value* (RSV) to indicate the numerical value produced by the estimate of topical relevance.

Different mathematical theories have been employed to define IR models, which include set theory, linear algebra, probability theory, formal logics, and fuzzy set theory [8, 27]. One of the most popular IR models, the vector space model, proposes a geometric interpretation of the IR process [83], in which both documents and queries are represented as vectors in the multidimensional space of indexing terms and the relevance assessment is interpreted as a measure of the proximity of the vectors representing a document and a query via the cosine of the angle between them. Another popular point of view about relevance is the probabilistic one [5, 75, 78, 79], in which the decision relies on a probabilistic measure of the relevance of a document d to a query q.

Based on Cooper's work [19], which aims to consider the semantics of both document and query in IR, van Rijsbergen [94] claimed that "*the use of semantics comes with an appropriate logic.*" Logic modeling for IR relies on a first hypothesis that can be expressed as *a document d is relevant to a query q if we can exhibit a deduction path from the document that leads to the query, denoted* $d \rightarrow q$.

Hence, formal logics constitute a possible way to deal with the semantics of both documents and queries. The logical point of view offers appealing properties. First, logical IR models capture part of the semantics of documents and queries by means of a logical language. Second, logical models provide a distinction between the retrieval decision and its uncertainty. Van Rijsbergen proposes a second hypothesis of formulating the uncertainty of the implication $d \rightarrow q$ through the following *Logical Uncertainty Principle (LUP)*:

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

An interesting and important aspect that can be explicitly modeled when adopting a logic-based model is the possibility to account for the uncertainty of the retrieval process by means of an uncertainty function U. Defining an uncertain implication $U(d \rightarrow q)$ for representing the retrieval decision means that, in logic-based IR models, in addition to the document d and the query q, the logical implication $d \rightarrow q$ and the uncertainty function U need to be defined. In general, these definitions are based on the selection of a formal logic as an underlying mathematical framework.

[1] Topical relevance normally refers to content-based relevance [71], in which the assessment process is based only on the content of documents and queries and is independent of any contextual information, such as user background knowledge, search problem, and document popularity.

In IR, a variety of formal logics have been employed, for example, modal propositional logic [23, 69], first-order logic [17], and description logic [63, 85].

The aim of this survey is to review and comparatively analyze the various logical models of IR in a structured way by keeping a clear distinction between the formal definition of the logical model and the uncertainty that may affect it. More precisely, for each considered logical model, a categorization of the representation of documents, queries, and the adopted implication is offered. As a result, and in contrast to the previous reviews that have appeared in the literature [22, 52], a multiple-level categorization of the existing logic-based models and their uncertainty is provided in Section 2. Moreover, we review new proposals, such as [3, 59], which are models based on propositional logics and also lead to effective implementations.

The motivation that brought us to write this survey is related to a resurgence of interest in the logical IR models due to their main following potentialities. (1) Logical IR models represent a means to understand relevance, in which the inference mechanisms and their peculiarity to represent and process knowledge [7, 9, 10] offer an informed way to assess relevance. (2) Logical IR models can be seen as general IR models that are capable of representing the non-logical models [1, 67]. (3) Formal logics offer a means to naturally include in the IR process some domain knowledge (see Section 1.2). Semantic search engines [92] exploit data from the semantic web and use logic to infer, from the original query made of concepts (e.g., "Brazilian football team"), the instanced queries (e.g., "Naymar, . . . ") submitted to a classical IR system. Knowledge graphs introduced by Google [11] are semantic information used to enhance search engines through reasoning. (4) It has been recently proven that an efficient implementation of logical models can be obtained [1, 3], thus paving the way to a reconsideration of logical models in IR.

Before detailing the structure of this survey, we first briefly introduce the logical formalism of IR in Section 1.1; then, we discuss the rationale behind logic in IR (Section 1.2). For a brief introduction about formal logics, refer to Appendix A.

1.1 Logical Formalization of IR

Logical IR models proposed in the literature differ in two main aspects: (1) the chosen logic $\mathcal{L}$ and (2) the logical value of the formula $d \rightarrow q$ (or of the formula $q \rightarrow d$, as will be explained later in this section) used to determine whether the document d is relevant to the query q within this logic.

As to the second aspect, the mechanism used to determine the relevance of a document d to a query q depends on the meaning of the formula $d \rightarrow q$. Some possible meanings are [86]:

(1) for a given logic $\mathcal{L}$, $d \rightarrow q$ is *true* for a particular *interpretation* δ, denoted $\{\delta\} \models d \rightarrow q$;
(2) q is a *logical consequence* of d in $\mathcal{L}$, denoted $M(d) \models q$;
(3) the formula $d \rightarrow q$ is *valid* (a tautology) in $\mathcal{L}$, denoted $\models d \rightarrow q$;
(4) $d \rightarrow q$ can be *derived* from a set of formulas Γ in $\mathcal{L}$, denoted $\Gamma \vdash d \rightarrow q$;
(5) q can be *derived* from d in $\mathcal{L}$, denoted $d \vdash q$;
(6) $d \rightarrow q$ is a *theorem* of $\mathcal{L}$, that is, *derived* from the axioms of $\mathcal{L}$, denoted $\vdash d \rightarrow q$.

Even if logical consequence (2), validity (3), derivability (5) and theorem derivation (6) are equivalent in sound and complete logics, these notions may encompass rather different meanings depending on the considered logic $\mathcal{L}$. Any IR model adopts one of the above six meanings, and it formalizes the way to compute the relevance of a document d to a query q accordingly. As will be seen in the rest of the article, the notion of validity is the most widely used in the literature. Note that for (2) and (5), → is expressed as a meta symbol that does not belong to the logic. This point will be discussed further.

A second difference between the approaches proposed in the literature is the logic selected to model the IR process. For instance, there are IR models founded on propositional logic, first-order logic, modal logic, logical imaging, description logic, and so forth. A first issue to be addressed when formalizing the process of topical relevance assessment in a selected logic is how to express both query and document representations. Within a logical framework both a document d and a query q are formally represented as expressions (formulae) of the selected logical language. The topical connection between them is modeled by allowing the inference of a query from a document or a document from a query. The two above possibilities to set the inferential process provide, of course, different interpretations of the retrieval process; in the literature, both points of view have been explored. More formally, the retrieval process can be modeled by either $d \vdash_{\mathcal{L}} q$ or $q \vdash_{\mathcal{L}} d$. In particular, $d \vdash_{\mathcal{L}} q$ means that by applying the inference rules of $\mathcal{L}$ to d and the set of axioms of $\mathcal{L}$, we can infer q.

1.2 Rationale Behind Using Formal Logics in IR

Formal logics offer a means to include some domain knowledge in the IR process, which is useful for improving the inferential process that assesses document relevance to a query. This has not been sufficiently explored yet, but it deserves further investigation.

Let us suppose that the domain knowledge Γ is expressed by means of the formal language of $\mathcal{L}$ and that one models $d \rightarrow q$ by using derivation. In this case, the retrieval decision could be expressed as $\Gamma \cup \{d\} \vdash_{\mathcal{L}} q$, which means that by applying the inference rules of $\mathcal{L}$ to d, the knowledge Γ, and the set of axioms of $\mathcal{L}$, we can infer q. In the former representation $d \vdash_{\mathcal{L}} q$, the inference of q is based only on d and $\mathcal{L}$, whereas in the latter $\Gamma \cup \{d\} \vdash_{\mathcal{L}} q$, the inference is also based on Γ (thus offering an extension to the point (5) in Section 1.1). As previously outlined, this allows one to formally express and incorporate in a unique formal system the representation of knowledge related to a particular domain of interest.

Semantic approaches to IR actually rely on ad-hoc domain-dependent knowledge resources such as ontologies, knowledge bases, or thesauri. The formal representation and integration of domain knowledge into the considered logic constitutes an effective way to improve the quality of the retrieval process. Therefore, formal logics are a potentially useful mathematical tool to build *more accurate*[2] IR models.

In addition to the abovementioned advantages in supporting an explicit knowledge representation and integration into a logical language, formal logics also offer rich formal languages to represent the content of both documents and queries (especially for documents)[3]. With a logical language, we can easily represent a situation in which the content of d is about t_1 or t_2 and not t_3 by means of the logical sentence $d = (t_1 \vee t_2) \wedge (\neg t_3)$.

To summarize, formal logics are powerful tools for knowledge representation, knowledge integration into the IR process, and to represent the inferential nature of the retrieval decision. The most interesting aspects of using formal logics in IR are as follows.

- Formal logics are powerful tools for simulating and modeling the inferential nature of the (topical) relevance assessment process.
- Formal logics are well suited to knowledge representation [7, 9, 10] and then for building IR models being capable of formally integrating domain knowledge into the retrieval process [63, 70].
- Formal logics offer a rich logical language to represent the content of documents.

[2] A more accurate IR model means a model deciding relevance in a closer way to the domain-expert relevance judgment.
[3] It is well known that the classical bag-of-words approach to text representation fails to represent relations and connectives between words.

1.3 Structure of the Survey

In this survey, the main logic-based IR models are presented. In Section 2, we categorize the logical IR models according to their different components: considered logic, document d and query q representations, the retrieval decision, and how uncertainty is modeled and its relation to the implication. These different axes allow a multiple comparison of the different models. Section 2 concludes with a general overview of the main logical models proposed in the literature. After this transversal presentation of the state of the art, we describe several logical IR models based on various logics: propositional logic (Section 3), propositional modal logic (Section 4), and the other logics (Section 5). In Section 6, we present information systems, such as semantic web and knowledge graph, which use logical models to access knowledge or information. We present our conclusions in Section 7.

2 CATEGORIZATION OF LOGICAL IR MODELS

There are several ways to categorize[4] the logical IR models proposed in the literature. In general, they are compared either according to the expressive power of their underlying formal logic (e.g., propositional logic, description logic, and first-order logic) or according to the mathematical theories that are used to define the uncertainty function U, for example, possibility and probability. As outlined in Section 1, we categorize logical IR models according to their different components: associated logic, document d and query q representations, the retrieval decision, and uncertainty modeling and its relation to the implication. These different axes allow a multiple and deep comparison of the different models.

A coarse-grained categorization has already been used in several survey papers. For example, Crestani and Lalmas [22] organize logical models into two categories: models that are essentially based on formal logics and models that are based on logical uncertainty theories (uncertainty with logical roots). Actually, the two axes of categorization in [22] are the underlying formal logic and the uncertainty theory. The same categorization is also used in Lalmas and Bruza [52]; in addition to this way of categorization, Lalmas [51] presents the main characteristics that should be present in any logic to be suitable for IR modeling.

According to Sebastiani [87], logical IR models can be put into two main categories according to the approach in which the logic is used for modeling IR: proof theoretic and model theoretic, in other words, the group of models that uses the formal syntax and the group that uses the formal semantics of the underlying logic, respectively. Sebastiani draws also a link between proof-theoretic and model-theoretic views of logic on the one hand and open-world (partial knowledge) and closed-world (total knowledge) assumptions on the other hand. Sebastiani argues that the model-theoretic approach to IR modeling is implicitly related to the strong assumption of the total-knowledge or closed-world assumption, whereas the proof-theoretic approach is related to the partial knowledge or open-world assumption.

In this survey, we take a step forward in the categorization of logical models. Based on the fact that any logical model gives a precise definition of d, q, $\rightarrow$, and U, we propose a categorization for each component: categorizing models according to their ways to represent d and q, and categorizing models according to their interpretation of both the implication $\rightarrow$ and the uncertainty function.

Therefore, this section proposes an overview of logical IR models that will be more accurately described in subsequent sections of this survey. The overview is based on the analysis of each IR component: document and query representation (Section 2.2), retrieval implication (Section 2.3),

[4] In this paper, we use the terms *categorization* and *axis* in their general meaning, without any reference to any particular meaning in computer science.

Table 1. Logic-Based IR Models

Associated logic		IR models
Propositional Logic $\mathcal{PL}$		Losada and Barreiro [59] Picard and Savoy [74] Abdulahhad [1]
Propositional Modal Logic $\mathcal{PML}$		Nie [67–69] Crestani and van Rijsbergen[23] Nie and Brisebois[70]
Others	**Description Logic $\mathcal{DL}$**	Meghini et al. [63] Sebastiani [85]
	First-Order Logic $\mathcal{FL}$	Chevallet and Chiaramella[16]
	Probabilistic Datalog	Fuhr [34]
	Situation Theory $\mathcal{ST}$	Lalmas and Rijsbergen [53]
	Fuzzy Logic $\mathcal{FZL}$	Pasi [72] Bosc et al. [12] Ughetto et al. [91]
	Default Logic	Hunter [46]

and uncertainty (Section 2.4). Section 2.1 briefly presents the IR models and their associated logics; each will be more extensively described later in this survey.

2.1 Logic-Based IR Models and their Associated Logic

The IR models that are presented and analyzed in this survey are listed in Table 1 with their associated logic. Propositional logic and propositional modal logic are used in several IR models, whereas other logics—such as description logic, first-order logic, and so on—are associated to very few models.

2.2 Document and Query Representation

2.2.1 Language Extension for Uncertainty or Not? An aspect that distinguishes logical IR models at the level of document and query representation is whether the logical sentence that is used to represent d or q is certain or not. For example, let us assume that s_d is the logical sentence that represents the content of d and a is the logical proposition that represents the notion of "*IR*." Now, if we say that d is about "*IR*" then is this knowledge certain or not? To express the uncertainty about the content of documents and queries, researchers generally extended the formal language of the underlying logic with an uncertainty component, for example, αs in the probabilistic datalog means that α is the probability that the logical sentence s is true [34]. Accordingly, logical IR models differ by the way in which they provide both document and query representations, as well as by the existence (or not) of a formal language extension that allows expression of both document and query uncertainty. This leads to distinguishing IR models that integrate a *language extension for expressing document and query uncertainty* and those *without such an extension*.

2.2.2 Formal Syntax versus Formal Semantics? Most logical IR models represent documents and queries as logical sentences, which are syntax-related notions. However, some other models represent documents and queries using semantics-related notions such as possible worlds [87]. If the underlying logic is sound and complete, the syntax- and semantics-related notions are equivalent; thus, there is no need to say that the IR model uses the notions of syntax or semantics to represent

Table 2. Document and Query Representation Categorization

	Language extension for uncertainty	Without extension
Formal syntax	Fuhr [34] Picard et al. [74] Sebastiani [85]	Abdulahhad [1] Chevallet and Chiaramella[16] Crestani and van Rijsbergen [23] Hunter [46] Lalmas and van Rijsbergen [53] Losada and Barreiro [59] Meghini et al. [63]
Formal semantics		Bosc et al. [12] Nie [67–69] Nie et al. [70] Pasi [72] Ughetto et al. [91]

the content of documents or queries. However, in the cases in which the logic is not sound or not complete (which is the case of some families of $\mathcal{PML}$), then the two ways of representation have to be distinguished. Accordingly, two families of IR models could be distinguished: *formal syntax*–based models and *formal semantics*–based models.

Table 2 shows the IR models according to these axes of categorization.

2.3 Implication-Based Categorization

2.3.1 Discussion beyond Material Implication and Conditional. Before advancing in this axis of categorization, it is fundamental to present in detail the two notions: the conditional $d \rightarrow q$ and the material implication $d \supset q$, and to illustrate the implicit relation between them. It is also useful to explain their adequacy for IR.

In IR, there exists a debate about the convenience of using the material implication $d \supset q$ to model the retrieval decision $d \rightarrow q$ [71, 94]. The material implication leads to what is conventionally called in IR the *false document* problem [94], in which the material implication $d \supset q$ is true even if d is false. Consequently, the relevance decision could be no longer related to the user's query q.

To deal with this problem, some researchers proposed using the conditional *IF d THEN q*, denoted $d \rightarrow q$, to model the IR implication instead of the material implication [71, 94]. The conditional $d \rightarrow q$ means that it is not possible to decide about the relevance if the antecedent d is not true. Accordingly, the relevance decision is always a query-dependent decision, which is not always the case in the material implication $d \supset q$.

A more in-depth look at material implication and conditional could help for better understanding and revealing the potential relationship between the two notions. Logicians categorize material implication $s_1 \supset s_2$ as the *truth-functional* conditional to distinguish it from the *non-truth-functional* conditional $s_1 \rightarrow s_2$ [30]. In any logical sentence s, if the truth values of the constituents subsentences of s are sufficient to determine the truth value of s, then we have a truth-functional sentence; otherwise, we have a non-truth-functional sentence. For example, the sentence $s_1 \wedge s_2$ is truth functional because if the truth values of s_1 and s_2 are known, the truth value of $s_1 \wedge s_2$ can be determined.

The material implication $s_1 \supset s_2$ is a truth-functional conditional, where the truth value of the implication $s_1 \supset s_2$ can be assessed based on the truth values of s_1 and s_2. When s_1 is true and

Table 3. Material Implication and Conditional Truth Table

s_1	s_2	$s_1 \supset s_2$	$s_1 \rightarrow s_2$
T	T	T	T
T	F	F	F
F	T	T	T/F
F	F	T	T/F

s_2 is false, then $s_1 \supset s_2$ is false; otherwise, it is true (see Table 3). This is not the case in general, where the truth value of a conditional such as $s_1 \rightarrow s_2$ is indeterminable based on the truth values of s_1 and s_2. When s_1 is true and s_2 is true, then $s_1 \rightarrow s_2$ is true[5]. When s_1 is true and s_2 is false, then $s_1 \rightarrow s_2$ is false. However, in the other cases when s_1 is false, the truth value of $s_1 \rightarrow s_2$ is indeterminable and it could be true or false (see Table 3).

If we use the notation $M(s)$ to refer to the set of formal models of a logical sentence s (see Section 3.1), then $M(s_1 \supset s_2) = [M(s_1) \cap M(s_2)] \cup M(\neg s_1)$. However, we have two possible cases for the conditional $s_1 \rightarrow s_2$:

- If we suppose that the conditional $s_1 \rightarrow s_2$ is true when s_1 is false, then $M(s_1 \rightarrow s_2) = [M(s_1) \cap M(s_2)] \cup M(\neg s_1)$, exactly like the material implication $s_1 \supset s_2$.
- If we suppose that the conditional $s_1 \rightarrow s_2$ is false when s_1 is false, then $M(s_1 \rightarrow s_2) = M(s_1) \cap M(s_2)$, exactly like the logical conjunction $s_1 \wedge s_2$.

First, the material implication $s_1 \supset s_2$ is a particular kind of conditional $s_1 \rightarrow s_2$ or, in other words, the material implication $s_1 \supset s_2$ represents a stronger decision than the conditional $s_1 \rightarrow s_2$. Second, the conditional $s_1 \rightarrow s_2$ is neither a material implication $s_1 \supset s_2$ nor a conjunction $s_1 \wedge s_2$; it is something in-between.

Since conditionals are non-truth functional, they are more complicated than material implication. In fact, in the case that the antecedent s_1 of a conditional $s_1 \rightarrow s_2$ is false, the decision about the truth of $s_1 \rightarrow s_2$ is not immediate, as it is in the case of the material implication; thus, there is a process to follow to decide about the truth of $s_1 \rightarrow s_2$ [71, 94]. Accordingly, and for efficiency reasons, it would be more convenient to use the material implication $s_1 \supset s_2$, but the truth of the material implication is not a suitable choice for IR, as illustrated in the literature. To address this issue, another view of material implication—the validity of material implication, denoted $\models s_1 \supset s_2$—has been studied [2]. A logical sentence s is valid, denoted $\models s$, iff, for any formal interpretation $\delta \in \Delta$, the sentence s is true. In IR, the material implication $d \supset q$ is valid iff

- d is unsatisfiable, denoted $\not\models d$ (always false document). A document d is always false if it is empty or if it contains a contradictory information of the form $d = \cdots \wedge a \wedge \neg a \wedge \ldots$.
- q is valid, denoted $\models q$ (always true query). This means that any document would be evaluated as relevant to a valid query (i.e., a tautology).
- every model of d is also a model of q, that is, $M(d) \subseteq M(q)$. That means, to check the validity of $d \supset q$, we need to check the truth of q in each model of d or, in other words, what makes the decision possible is the situations in which d is true; this is exactly the idea underlying using conditionals in IR.

[5] Of course, this holds if → is a logical connector and not a meta-symbol.

Table 4. Retrieval Implication Categorization

	Logical connective	Logical meta-connective
Proof-theoretic	Hunter [46] Meghini et al. [63] Picard and Savoy [74]	Fuhr [34]
Model-theoretic	Abdulahhad [1] Bosc et al. [12] Chevallet and Chiaramella[16] Pasi [72] Sebastiani [85] Ughetto et al. [91]	Lalmas and van Rijsbergen [53] Losada and Barreiro[59] Nie [67–69] Crestani and van Rijsbergen [23] Nie et al. [70]

To summarize, on the one hand, the validity of material implication does not suffer from the false document problem and, on the other hand, it has a very similar behavior to the one of a conditional, that is, the situations in which d is true control the final decision. In other words, the validity of material implication has a similar behavior as a conditional but it does not have their complications.

Parallel to this analysis of material implication and conditional, there exists a cross analysis of logical IR models considering, on one side, the decision of relevance (inside or outside the logics) and, on the other side, the meta-language of the logics. Therefore, we now present these two axes.

2.3.2 *Logical Connective versus Logical Meta-connective.*

In order to categorize IR models with respect to the retrieval decision, we adopt a slightly different approach with respect to the one adopted in [87]. We start by the following question: *Does the connective $\rightarrow$ in $d \rightarrow q$ belong to the underlying logical language or not?* If it belongs, then it is a logical connective between two sentences, for example, the material implication $s_1 \supset s_2$. However, if $\rightarrow$ refers to the process of inferring a sentence s_2 from one or more sentences $\{s_1\}$ by applying the inference rules of the underlying logic, then $\rightarrow$ is a meta-connective. For example, given the set of formulae $\{s_1, s_1 \supset s_2\}$, it is possible to infer s_2 by applying Modus-Ponens, denoted $\{s_1, s_1 \supset s_2\} \vdash s_2$.

In the former case, when $\rightarrow$ is a logical connective, the retrieval decision is to show that the well-formed logical sentence $d \rightarrow q$ is valid or it is a theorem, depending on considering the semantic or the syntactic side of the logic, respectively.

In the latter case, when $\rightarrow$ is a meta-connective, the retrieval decision is to infer q from d or to show that q is a logical consequence of d depending on considering the syntactic or the semantic side of the logic, respectively.

The main difference between the two cases is that, in the case of meta-connective, we implicitly suppose that we have background knowledge, which is the document itself d.

2.3.3 *Proof-Theoretic vs. Model-Theoretic.*

Whatever $d \rightarrow q$ is a logical connective or meta-connective, there are two points of view to implement $d \rightarrow q$: the proof-theoretic and model-theoretic, according to taking the syntax or the semantics side of the logic, respectively [86, 87]. If $d \rightarrow q$ is a meta-connective, then there are two ways to express it: as proof-theoretic notion $d \vdash q$ or as model-theoretic notion $d \models q$. If the logic is sound and complete, then the two notations are equivalent. Furthermore, if we are talking about classical logics, such as propositional logic, both ways of viewing $d \rightarrow q$, either as a logical connective or as a meta-connective, are tightly related, where $\models d \supset q$ is equivalent to $d \models q$.

Table 4 shows the IR models according to these axes.

2.4 Categorization Based on Uncertainty versus Multiple Truth Values

As well outlined in [29], various calculi for uncertainty modeling exist, including probability theory, possibility theory, and belief functions. It is also possible to differentiate between *qualitative* and *quantitative* definitions of uncertainty [42]. In relation to logic, the management of uncertainty has been often modeled by means of weights attached to propositions, thus introducing an extension of the employed formal language. For example, if considering propositional logic, the association of a weight with a proposition is a meta-level association. In fact, the truth of propositions remains binary, while the associated weight can express the more or less strong inability to know if a proposition is true or false.

However, several works that have dealt with this kind of extension of the logic language have introduced several misunderstandings related to the role, the meaning, and the properties of those weights [29]. In addition, multi-valued logics have often been wrongly interpreted as logics of uncertainty, while they are basically constructed as truth-functional calculi, and they are compositional (the degree of truth of a formula can be computed based on the degrees of truth of its constituent subformulas).

Thus, while in multi-valued logics the truth value of a sentence $a \wedge b$, for example, is a function of the truth values of its components a and b (compositionality holds), the uncertainty of $a \wedge b$, denoted $U(a \wedge b)$, cannot be determined by using only the uncertainties of its components, that is, $U(a)$ and $U(b)$ (compositionality does not hold) [41].

Multi-valued logics allow definition of multi-valued propositions (e.g., fuzzy propositions when truth values take values in the interval [0,1]).

The reason to model uncertainty in relation to logic-based approaches to IR has been motivated by the need of models able to provide a ranking mechanism of the documents retrieved in response to a query. However, in a logical framework, this purpose can be achieved in two ways: (a) by incorporating an uncertainty modeling in a logical calculus (i.e., as previously outlined, by defining the function U); and (b) by adopting a multi-valued logic, such as fuzzy logic, where the propositions are interpreted as fuzzy propositions and their truth value under a given interpretation is a value in the interval [0,1]. For this reason, two possible axes that could be used to categorize the logical models are if they explicitly model uncertainty or if they rely on multi-valued logics.

In the former case, a *degree of belief* about the truth or validity of $d \rightarrow q$ must be expressed. In this case, what must be expressed is the certainty that $d \rightarrow q$ is true. In the latter case, any belief or uncertainty is expressed within the logic: for example, in fuzzy logic, each proposition has a truth value expressed as a numeric value in [0,1], which allows the expression of partial truth. This numeric degree does not express any uncertainty: as well outlined in [29], "degrees of uncertainty are clearly a higher level notion, higher than degrees of truth."

Table 5 shows the different IR models considering either the degree of belief or multi-valued logics.

2.5 Synthesis

We presented in Section 2 our fine-grained analysis and categorization of logical models. To offer the readers a self-contained survey, we present in Sections 3, 4, and 5 a synthesis of the most prominent logical IR models and their underlying mathematical logic. However, before detailing logical IR models, Table 6 presents an overview of these models. It indicates for each model its definition of each part of the IR process: term, document, query, matching, and ranking. This table offers a summary of the next sections.

The next sections present each logical IR model within its logic in the order presented in Table 6. Section 3 describes models based on propositional logic, Section 4 describes those based on propositional modal logic, and Section 5 gathers the models based on all other logics.

Table 5. Ranking Through Uncertainty Modelling or Multi-valued Logics

	IR models
Degree of belief	Abdulahhad [1]
	Chevallet and Chiaramella [16]
	Crestani and van Rijsbergen [23]
	Fuhr [34]
	Hunter [46]
	Lalmas and van Rijsbergen [53]
	Losada and Barreiro [59]
	Nie [67–69]
	Nie et al. [70]
	Picard and Savoy [74]
	Sebastiani [85]
Multi-valued logic	Pasi [72]
	Bosc et al. [12]
	Ughetto et al. [91]

3 PROPOSITIONAL LOGIC ($\mathcal{PL}$)–BASED MODELS

This section is dedicated to the $\mathcal{PL}$-based models. We start the section by a brief introduction to $\mathcal{PL}$; we then present the IR models according to the mathematical theory that is used to describe uncertainty.

3.1 Propositional Logic ($\mathcal{PL}$): Preliminary

$\mathcal{PL}$, or zero-order logic, is defined on a finite set of atomic propositions or alphabet $\Omega = \{a_1, \ldots, a_n\}$ and a finite set of connectives Υ. One of standard sets of connectives is $\Upsilon = \{\neg, \wedge, \vee\}$. Depending on the alphabet Ω and the connectives Υ, a set of well-formed logical sentences Σ is defined as follows.

- Any atomic proposition is a well-formed logical sentence: $\forall a \in \Omega, a \in \Sigma$.
- The negation of a logical sentence is also a logical sentence: $\forall s \in \Sigma, \neg s \in \Sigma$.
- The conjunction of any two logical sentences is also a logical sentence: $\forall s_1, s_2 \in \Sigma, s_1 \wedge s_2 \in \Sigma$.
- The disjunction of any two logical sentences is also a logical sentence: $\forall s_1, s_2 \in \Sigma, s_1 \vee s_2 \in \Sigma$.
- Σ does not contain any other sentences.

Material implication $\supset$ is implicitly included in Υ where, for any two logical sentences s_1 and s_2, the material implication $s_1 \supset s_2$ is equivalent to $\neg s_1 \vee s_2$.

In $\mathcal{PL}$, a formal semantics or interpretation is given to a logical sentence s through assigning a truth value (T or F) to each atomic proposition in s. Each interpretation corresponds to a mapping between all atomic propositions and the set of truth values $\{T, F\}$. The set of atomic propositions Ω thus corresponds to $2^{|\Omega|}$ possible interpretations because each atomic proposition $a \in \Omega$ is mapped to one of two possible values (T or F).

First, we define the set of all possible mappings Π_Ω between the set of atomic propositions Ω and the truth values $\{T, F\}$, where

$$\Pi_\Omega = \{\pi : \Omega \to \{T, F\} | \pi \text{ is a mapping}\},$$

Table 6. Logic-based IR Models

Model	t	d	q	$d \rightarrow q$	$U(d \rightarrow q)$
Propositional Logic $\mathcal{PL}$–based models					
Losada and Barreiro [59]	proposition	sentence	sentence	$d \models q$	Dalal's BR ($\circ_D$)
Picard and Savoy [74]	proposition	fact	sentence	$d \supset q$	cond. prob.
Abdulahhad [1]	proposition	sentence	sentence	$\models d \supset q$	Lattice+prob.
Propositional Modal Logic $\mathcal{PML}$–based models					
Nie [67, 68]	proposition	poss. world	sentence	path $d_0 \ldots d_n$	path's cost
Nie [69]	proposition	poss. world	sentence	-	imaging $P_d(q)$
Crestani and van Rijsbergen [23]	poss. world	sentence	sentence	-	imaging $P_d(q)$
Nie et al. [70]	proposition	poss. world	sentence	path $d_0 \ldots d_n$	fuzzy dist. $d(\Diamond^n q)$
Description Logic $\mathcal{DL}$–based models					
Meghini et al. [63]	-	individual	concept	assertion $d : q$	-
Sebastiani [85]	-	individual	concept	assertion $d : q$	probability
First-Order Logic $\mathcal{FL}$–based models					
Chevallet and Chiaramella [16]	-	conc. graph	conc. graph	proj. $d \leq q$	graph op. cost
Probabilistic Datalog–based models					
Fuhr [34]	-	ground facts	Bool. expr.	inference rule	probability
Fuzzy Logic $\mathcal{FZL}$–based models					
Pasi [72]	-	fuzzy set	Bool. exp.	$q \supset d$	fuzzy implication
Bosc et al. [12]	-	fuzzy set	Bool. exp.	$q \supset d$	fuzzy implication
Ughetto et al. [91]	-	fuzzy set	Bool. exp.	$q \supset d$	fuzzy implication
Situation Theory $\mathcal{ST}$–based models					
Lalmas and van Rijsbergen [53]	-	situation	infon	d supports q	cond. constraint
Default Logic based models					
Hunter [46]	proposition	clause	sentence	$d \supset q$	positioning

where $|\Pi_\Omega| = 2^{|\Omega|}$. Second, each interpretation $\delta \in \Delta_\Omega$ is a subset of the atomic propositions Ω. It corresponds to a mapping $\pi \in \Pi_\Omega$ between Ω and the truth values $\{T, F\}$. The set of interpretations Δ_Ω, which are based on a set of atomic propositions Ω, is

$$\Delta_\Omega = \{\delta = \{a \in \Omega | \pi(a) = T\} | \pi \text{ is the mapping of } \delta\},$$

where $|\Delta_\Omega| = 2^{|\Omega|}$. The notation Δ_Ω is a simplification of Π_Ω, where since atomic propositions can be mapped to only one of two possible truth values, the mapping π can be simplified to the set of atomic propositions that are mapped to T, where the other atomic propositions are implicitly mapped to F. Informally, for any alphabet Ω, the set of interpretations actually correspond to the different rows of the truth table that is built in terms of Ω. For example, suppose that Ω contains three propositions $\{a, b, c\}$, the truth table, the set of possible mappings Π_Ω, and the set of interpretations Δ_Ω as depicted in Table 7, where $\Pi_\Omega = \{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8\}$ and $\Delta_\Omega = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8\}$.

Any logical sentence $s \in \Sigma$ can be true or false in a specific interpretation $\delta \in \Delta_\Omega$. If s is true in δ, then it is denoted $\{\delta\} \models s$ and is read as δ satisfies or models s; otherwise, it is denoted $\{\delta\} \not\models s$. The truth value of an arbitrary logical sentence in Σ with respect to an interpretation is determined as follows:

Table 7. The Set of Interpretations Based on $\Omega = \{a, b, c\}$

a	b	c	π	δ
F	F	F	$\pi_1 = \{(a, F), (b, F), (c, F)\}$	$\delta_1 = \{\}$
F	F	T	$\pi_2 = \{(a, F), (b, F), (c, T)\}$	$\delta_2 = \{c\}$
F	T	F	$\pi_3 = \{(a, F), (b, T), (c, F)\}$	$\delta_3 = \{b\}$
F	T	T	$\pi_4 = \{(a, F), (b, T), (c, T)\}$	$\delta_4 = \{b, c\}$
T	F	F	$\pi_5 = \{(a, T), (b, F), (c, F)\}$	$\delta_5 = \{a\}$
T	F	T	$\pi_6 = \{(a, T), (b, F), (c, T)\}$	$\delta_6 = \{a, c\}$
T	T	F	$\pi_7 = \{(a, T), (b, T), (c, F)\}$	$\delta_7 = \{a, b\}$
T	T	T	$\pi_8 = \{(a, T), (b, T), (c, T)\}$	$\delta_8 = \{a, b, c\}$

Table 8. The Truth Table of Material Implication $\supset$

a	b	δ	$a \supset b$	
F	F	$\delta_1 = \{\}$	T	$\{\delta_1\} \models a \supset b$
F	T	$\delta_2 = \{b\}$	T	$\{\delta_2\} \models a \supset b$
T	F	$\delta_3 = \{a\}$	F	$\{\delta_3\} \not\models a \supset b$
T	T	$\delta_4 = \{a, b\}$	T	$\{\delta_4\} \models a \supset b$

- $\forall \delta \in \Delta_\Omega, \forall a \in \Omega, \{\delta\} \models a$ iff $a \in \delta$.
- $\forall \delta \in \Delta_\Omega, \forall s \in \Sigma, \{\delta\} \models \neg s$ iff $\{\delta\} \not\models s$.
- $\forall \delta \in \Delta_\Omega, \forall s_1, s_2 \in \Sigma, \{\delta\} \models s_1 \wedge s_2$ iff $\{\delta\} \models s_1$ and $\{\delta\} \models s_2$.
- $\forall \delta \in \Delta_\Omega, \forall s_1, s_2 \in \Sigma, \{\delta\} \models s_1 \vee s_2$ iff $\{\delta\} \models s_1$ or $\{\delta\} \models s_2$.

For any logical sentence s, the subset $M(s) \subseteq \Delta_\Omega$ that satisfies s is called the set of *models* of s, denoted $M(s) \models s$, where, for any interpretation $\delta \in M(s)$, we have that $\{\delta\} \models s$. That is, if we substitute each atomic proposition in s by its truth value in δ, then the truth value of s will be true. The notation $\models s$ means that s is a tautology or , or, in other words, s is true under any interpretation (i.e., $M(s) = \Delta_\Omega$). The notation $\not\models s$ means that s is false under all interpretations or **unsatisfiable** (i.e., $M(s) = \emptyset$). For example, assume that $\Omega = \{a, b\}$ and assume that the logical sentence is the material implication $a \supset b$. The set of models $M(a \supset b)$ of $a \supset b$ is depicted in Table 8, where $M(a \supset b) = \{\delta_1, \delta_2, \delta_4\}$. The set of models $M(s)$ is the set of models of a set of sentences logically equivalent to s. For example, $M(a \supset b) = M(\neg a \vee b)$ where it is well known that $a \supset b$ and $\neg a \vee b$ are equivalent. It is also known that, for any two logical sentences s_1 and s_2, we have that

$$[\models s_1 \supset s_2] \Leftrightarrow [M(s_1) \subseteq M(s_2)]$$

There are two main assumptions concerning the atomic propositions that do not occur in a sentence:

- Closed-World Assumption (CWA): For a sentence s, if an atomic proposition does not occur in s, then it is implicitly false. For example, assume that $\Omega = \{a, b, c\}$; then, under CWA, $M(a \wedge b) = \{\delta_7\}$ (Table 7), where c must be false.
- Open-World Assumption (OWA): For a sentence s, if an atomic proposition does not occur in s, then it can be either true or false. For example, assume that $\Omega = \{a, b, c\}$. Then, under the OWA, $M(a \wedge b) = \{\delta_7, \delta_8\}$ (Table 7), where c can be true or false.

$\mathcal{PL}$ is a special kind of formal logic, where it is *complete* and *sound*. Completeness and soundness govern the relation between provability $\vdash$ and satisfiability $\models$. Completeness means that if $M(s_1) \models s_2$, then $s_1 \vdash s_2$. Soundness means that if $s_1 \vdash s_2$, then $M(s_1) \models s_2$.

3.2 Belief Revision-Based Uncertainty

Losada and Barreiro [59] use $\mathcal{PL}$ to build an IR model. Each index term is an atomic proposition that could be either true or false under the interpretation offered by a particular document or query. A document d is a logical sentence formed using the index terms. A query q is also a logical sentence. The retrieval decision is a logical consequence or entailment, where d is relevant to q *iff* $d \models q$. Since the model is proposed in the $\mathcal{PL}$ framework, $d \models q$ is equivalent to $\models d \supset q$. In the formal semantics of $\mathcal{PL}$, $d \models q$ means that every model of d is also a model of q or, in other words, for every interpretation δ in which d is true, q is also true.

To estimate the uncertainty of $d \models q$, denoted $U(d \models q)$, Losada and Barreiro use the framework of Belief Revision (BR), which is a technique to formally express the notion of proximity between logical sentences. BR deals with updating an existing knowledge K with a new piece of information s, denoted $K \circ s$, where, if there is no contradiction between K and s, then the updated knowledge becomes $K \circ s = K \wedge s$; otherwise, BR deals with the *minimal change* that should be made to K in order to build an updated knowledge K' that does not contradict s, $K \circ s = K' \wedge s$. This last notion of *minimal change* is central to IR.

There are several BR techniques that deal with the syntax of the logical language, that is, formula-based approaches [66], and other techniques that deal with the formal semantics of the logic, that is, model-based approaches [73]. Losada and Barreiro use Dalal's BR operator [26], denoted $\circ_D$, which is one of the model-based approaches of BR. According to this operator, giving two interpretations δ_i and δ_j, the distance between them, denoted $dist(\delta_i, \delta_j)$, is the number of atomic propositions in which the two interpretations differ. For example, let us assume that our alphabet Ω contains three atomic propositions $\Omega = \{a, b, c\}$. Assume that we have two interpretations: $\delta_i = \{a, b\}$, which means that a and b are true whereas c is implicitly false, $\delta_j = \{a, c\}$, which means that a and c are true whereas b is implicitly false. The distance between δ_i and δ_j is

$$dist(\delta_i, \delta_j) = |(\delta_i \cup \delta_j) \setminus (\delta_i \cap \delta_j)|.$$

Hence, the distance between a logical sentence s and an interpretation δ is calculated as follows:

$$Dist(M(s), \delta) = \min_{m \in M(s)} dist(m, \delta).$$

$M(s)$ is the set of models of s or, equivalently, the set of interpretations in which s is true.

To estimate the uncertainty of the retrieval decision $U(d \models q)$ by using Dalal's BR operator, Losada and Barreiro distinguish between two modes of revision:

- Revising q by d, denoted $q \circ_D d$. Here, there are two cases:
  —q has several models $M(q)$ whereas d has only a unique model m_d (CWA).
  $$distance(d, q) = Dist(M(q), m_d)$$
  —q has several models $M(q)$ and d also has several models $M(d)$ (OWA).
  $$distance(d, q) = \frac{\sum_{m \in M(d)} Dist(M(q), m)}{|M(d)|}$$
  Now, the similarity between d and q is
  $$BRsim(d, q) = 1 - \frac{distance(d, q)}{k},$$
  where k is the number of atomic propositions appearing in q.

- Revising d by q, denoted $d \circ_D q$.

$$distance(q, d) = \frac{\sum_{m \in M(q)} Dist(M(d), m)}{|M(q)|}.$$

Now, the similarity between d and q is

$$BRsp(d, q) = 1 - \frac{distance(q, d)}{l},$$

where l is the number of atomic propositions appearing in d.

According to Losada and Barreiro, the final retrieval score is a *linear combination* of the two similarities:

$$U(d \models q) = \alpha \times BRsim(d, q) + (1 - \alpha) \times BRsp(d, q).$$

The main problem in the previous equation is that computing $BRsim(d, q)$ and $BRsp(d, q)$ is time-consuming, because the number of models $M(d)$ and $M(q)$ is exponential in the number of atomic propositions (or, equivalently, index terms).

Losada and Barreiro propose rewriting the logical sentences in Disjunctive Normal Form (DNF) and, instead of computing the distance between d and q based on interpretations, they compute it based on clauses. A sentence s in DNF is a disjunction of clauses $c_1 \vee \cdots \vee c_n$ and each clause c_i is a conjunction of literals $l_1 \wedge \ldots l_m$ and each literal l_j is either an atomic proposition a_j or its negation $\neg a_j$. The distance between two clauses is the number of atomic propositions that are positive in one clause and negative in the other, computed as follows:

$$CDist(c_i, c_j) = |\{l \in c_i | \neg l \in c_j\}|.$$

This new approach is efficient; thus, applying the model to large-scale document collections is feasible. However, this approach is applicable only in the case in which both d and q are composed of only one clause. In the general case in which d and q are DNF sentences, Losada and Barreiro illustrate that the computation of query models is still needed and the algorithm is still exponential. To overcome this problem, they propose a simplification; in this case, however, the technique of computing the uncertainty U is no longer based on BR.

To sum up, since atomic propositions correspond to the index terms, the distance measure calculates how many different terms exist between d and q. Losada and Barreiro claim that their model is equivalent to the vector space model (VSM) [83] with a more powerful language to represent the content of documents and queries but without the ability to represent term weights.

3.3 Probability-Based Uncertainty

Probabilistic argumentation systems (PASs) [40] extends propositional logic ($\mathcal{PL}$) by a mechanism that allows expression of uncertainty by using probability theory. The uncertainty is captured by means of special propositions called *assumptions*. Uncertain rules in PASs are generally defined as follows:

$$a \wedge s_1 \rightarrow s_2,$$

where → is the material implication, s_1 and s_2 are logical sentences formed by a set of propositions (the alphabet of propositional logic of PAS), and a is a logical sentence formed by another set of propositions (the assumptions). The two sets of propositions (alphabet and assumptions) are disjoint. The above rule means that s_1 implies s_2 under some circumstances represented by a. It is also possible to associate a probability with the assumptions, where $P(a)$ means $P(s_2|s_1)$ and not $P(s_1 \rightarrow s_2)$.

A knowledge base K is a set of uncertain rules of the previous form. It is possible to check whether a knowledge base K *supports* or *discounts* a particular hypothesis h, where h is any logical sentence. The support of a hypothesis h in a knowledge base K, denoted $sp(h, K)$, is the disjunction of the assumptions a_i, where for any assumption a_i we have that $(a_i \wedge K) \models h$. The *degree of support* of a hypothesis h in a knowledge base K, denoted $dsp(h, K)$, refers to the degree to which K supports h knowing that h does not contradict K. The degree of support $dsp(h, K)$ is the quantitative or numerical expression of the support $sp(h, K)$, and $dsp(h, K)$ is a type of conditional probability.

Picard and Savoy [74] use PAS to build an IR model. According to Picard and Savoy, documents, queries, and the indexing terms are atomic propositions. The content of documents and queries is represented by a set of rules of the following form:

$$a_{ij} \wedge d_i \rightarrow t_j,$$

which means that the document d_i is about or indexed by the term t_j with a degree of uncertainty $P(a_{ij}) = P(t_j|d_i)$. Semantic relations between terms are represented using the following rule:

$$b_{ij} \wedge t_i \rightarrow t_j,$$

which means that there is a relation between t_i and t_j; $P(b_{ij})$ is the strength of this relation. The strength of a relation is related to the type of this relation.

The set of previous rules forms the IR knowledge base K. The retrieval decision $d \rightarrow q$ is the material implication $d \supset q$. The uncertainty $U(d \supset q)$ is estimated in two possible ways:

- either symbolically: $U(d \supset q) = sp(q, K \wedge d)$
- or numerically: $U(d \supset q) = dsp(q, K \wedge d)$.

In both ways, the document d is assumed as a fact or it is observed. The knowledge base K is updated according to this observation to become $K \wedge d$. Actually, the model captures the degree to which the new knowledge base $K \wedge d$ supports the query q.

The formalism presented by Picard and Savoy is capable of representing the inter-terms relationships through rules of the form: $b_{ij} \wedge t_i \rightarrow t_j$. However, Picard and Savoy emphasize the ability of their formalism to represent the inter-document relationships, for example, hyperlinks, citations, and the like. They employ rules of the form: $l_{ij} \wedge d_i \rightarrow d_j$ to represent this type of relation, where d_i is related to d_j by a link directed from d_i to d_j, and the probability $P(l_{ij})$ reflects the type of relation and its strength. The main disadvantage is that estimating $P(a_{ij})$, $P(b_{ij})$, $P(l_{ij})$ in the previous rules requires the availability of relevance information. The probability $P(a_{ij})$ in the rule $a_{ij} \wedge d_i \rightarrow t_j$ is estimated by using a variant of the classical *tf.idf* measure. In general, the main difficulty in this study is the probability estimation.

3.4 Lattices and Probability-Based Uncertainty

There is a mathematical link between $\mathcal{PL}$ and probability via lattices [20, 47, 48]. There are several ways to express the link between $\mathcal{PL}$ and lattices [3, 48]. The classical way is through the formal semantics of $\mathcal{PL}$. Assume the set of atomic propositions Ω. On the one hand, the following structure $L = (2^{2^\Omega}, \cap, \cup)$ is a distributive lattice, where 2^{2^Ω} is the powerset of the powerset of Ω, $\cap$ is the meet operation, $\cup$ is the join operation, and the partial order relation defined on the lattice L is the set inclusion $\forall x, y \in 2^{2^\Omega}, x \leq y \Leftrightarrow x \subseteq y$. On the other hand, each logical sentence s in $\mathcal{PL}$ corresponds to a set of models $M(s)$, where each model $m \in M(s)$ is the set of atomic propositions that must be set to true in this model; hence, $m \in 2^\Omega$ and $M(s) \in 2^{2^\Omega}$. Furthermore, for any two sentences s_1, s_2, we have that $\models s_1 \supset s_2$ is equivalent to $M(s_1) \subseteq M(s_2)$, where $s_1 \supset s_2$ is the material implication. Accordingly, the potential link between $\mathcal{PL}$ and lattices can be formed as follows:

any logical sentence s corresponds to a node $M(s)$ on the lattice L, and the partial order relation defined on L is equivalent to the validity of material implication, where $\models s_1 \supset s_2$ is equivalent to $M(s_1) \leq M(s_2)$.

On the other side, the link between uncertainty (probability as a special case) and lattices is defined as follows: partial order relations are binary, that is, either x is smaller than y or not. However, partial order relations can be quantified where, instead of saying that x is smaller than y or not, we estimate the degree to which x includes y, denoted $Z(x, y)$:

$$\forall x, y \in L, Z(x,y) = \begin{cases} 1 & \text{if} \quad x \geq y \\ 0 & \text{if} \quad x \cap y = \bot \\ z & \text{otherwise, where} \quad 0 < z < 1. \end{cases}$$

The Z function can express the uncertainty notion. If we want that Z becomes consistent with all structural properties of distributive lattices, then Z will be equivalent to a conditional probability, where $Z(x, y) = P(x|y)$, although there are several possible implementations for Z not forcibly probability.

Accordingly, lattices can form the mathematical link between formal logics and uncertainty. Abdulahhad et al. [3] exploit this fact in order to build their IR model. In their work, the document d is a logical sentence written in DNF[6]. The query q is a logical sentence written in DNF. The retrieval decision $d \rightarrow q$ is the validity of material implication $\models d \supset q$ [2]. Before defining the uncertain implication $U(d \rightarrow q)$, the mathematical link between $\mathcal{PL}$ and lattices is redefined, where instead of sets of elements as nodes, they build a new lattice L' with nodes as flat sets of elements. In this new lattice L', each clause (not each sentence as in L) corresponds to a node and the partial-order relation is equivalent to the validity of material implication between clauses. Accordingly, d corresponds to one or several nodes in L', because d is a logical sentence written in DNF, and then it consists of one or several clauses; q also corresponds to one or several nodes in L'. The retrieval decision corresponds to the partial-order relation of L'.

The uncertainty $U(d \rightarrow q)$ is estimated by exploiting the mathematical link between lattices and uncertainty (the previously defined Z function). Assume that d and q are clauses, then

$$U(d \rightarrow q) = Z(d^*, q^*),$$

where d^*, q^* are the corresponding L' nodes of d, q, respectively. The matching score $RSV(d, q)$ is computed through the general case, where d and q consist of several clauses:

$$RSV(d,q) = U(d \rightarrow q) = \underset{c_i \in C_d}{G} \left(\underset{c_j \in C_q}{G'} \left(Z(c_i^*, c_j^*) \right) \right),$$

where C_d, C_q are the set of clauses of d and q, respectively, c_i^* is the corresponding node of the clause c_i, G is a triangular norm function, and G' is a triangular conorm.

This work shows that the vectorial inner product is a possible implementation of Z [1]. According to this observation, several operational variants of their theoretical model are built.

4 PROPOSITIONAL MODAL LOGIC ($\mathcal{PML}$)–BASED MODELS

Modal logics introduce the two notions of *necessary* and *possible* that could be added to any logic. There are thus Propositional Modal Logic ($\mathcal{PML}$), First-Order Modal Logic, and so on. We here focus on $\mathcal{PML}$. Modal logics use the *Possible Worlds* (PW) semantics to give a meaning to the previous two modalities [50]. Worlds are connected through accessibility relations. Before presenting

[6] A logical sentence in DNF is a disjunction of clauses, where a clause is a conjunction of literals and a literal is an atomic proposition or its negation.

IR models that are built on $\mathcal{PML}$ and its PW semantics, let us first give a brief introduction to the PW semantics and to the related probabilistic technique, *Imaging* [37, 56].

4.1 Possible Worlds and Imaging

In Section 3.1, we present the formal interpretation and semantics of $\mathcal{PL}$. Kripke's semantics [50], or PW semantics, is another way to give a formal meaning to logical sentences; it is recommended to give a formal semantics to $\mathcal{PML}$. For more detailed information about modal logics, refer to [44].

PW semantics has the structure $\langle W, R\rangle$, where W is a non-empty set of what is conventionally called *Possible Worlds*, and R is a binary relation $R \subseteq W \times W$, conventionally called *accessibility relations*. The structure $\langle W, R, \Vdash\rangle$ defines Kripke's system, where $\Vdash$ determines if any sentence s is true in a world w (denoted as $w \Vdash s$) or if it is not true, denoted as $w \nVdash s$. The semantics of classical $\mathcal{PL}$ connectives (conjunction $\wedge$, disjunction $\vee$, negation $\neg$) in Kripke's system is defined as follows:

- Negation: $w \Vdash \neg s$ *iff* $w \nVdash s$.
- Conjunction: $w \Vdash s_1 \wedge s_2$ *iff* $w \Vdash s_1$ and $w \Vdash s_2$.
- Disjunction: $w \Vdash s_1 \vee s_2$ *iff* $w \Vdash s_1$ or $w \Vdash s_2$.

As previously said, $\mathcal{PML}$ defines two special unary operators, *Necessarily* ($\Box$) and *Possibly* ($\Diamond$). The formal semantics of these operators in Kripke's system is as follows:

- Necessarily: $w \Vdash \Box s$ *iff* for any world w' satisfying $(w, w') \in R$, $w' \Vdash s$. Conventionally, we could say that s is necessarily true in w iff s will be true in any future moment.
- Possibly: $w \Vdash \Diamond s$ *iff* there exists a world w' satisfying $(w, w') \in R$, $w' \Vdash s$. Here, also, we could say that s is possibly true in w iff there exists at least one future moment in which s will be true.

The notation $(w, w') \in R$ means that the world w' is accessible from w through the accessibility relation R. The properties of the accessibility relation R or, equivalently, the way of interaction between the added operators ($\Box$, $\Diamond$) and the classical operators ($\wedge$, $\vee$, $\neg$), determines different families of $\mathcal{PML}$, having different expressive powers. Some families of $\mathcal{PML}$ are *sound* and *complete*.

Imaging is a process developed in the framework of *The Logic of Conditionals*. Lewis [55] demonstrated that the probability of conditionals cannot be the standard conditional probability. However, we need a revised probability (via *imaging*), where the probability of conditionals is the corresponding revised conditional probability [4, 6]. Imaging also enables the evaluation of a conditional sentence $a \rightarrow b$ without explicitly defining the operator $\rightarrow$. Technically, to define imaging, let us first assume that a probability distribution P is defined on the set of possible worlds W as follows:

$$\sum_{w \in W} P(w) = 1.$$

Imaging transfers probabilities from some worlds to other worlds and builds a new probability distribution. In imaging, any logical sentence s can only be true or false in a particular world $w \in W$. Therefore, we define $w(s) = 1$ if s is true in w, that is, $w \Vdash s$, or $w(s) = 0$ otherwise. We also define w_s to be the most similar world to w where s is true, $w_s(s) = 1$. "The most similar," "the closest," and so on, are notions related to the definition of the accessibility relation R. *Generalized Imaging* [37] relaxes the previous two assumptions, where

- The truth of s in a world w is no more binary.
$$w(s) = \begin{cases} 0 & \text{if} \quad w \not\Vdash s \\ 0 < \alpha \leq 1 & \text{otherwise.} \end{cases}$$
Furthermore, for any logical sentence s, the following condition holds: $\sum_{w \in W} w(s) = 1$.
- There could be more than one most similar world. Therefore, w_s is no more one distinct world, it is now a set of worlds, where w_s are the worlds the most similar to w where s is true, that is, $\forall w' \in w_s, w'(s) > 0$.

After defining a probability distribution on worlds, the probability can also be defined on logical sentences. For any logical sentence s,

$$P(s) = \sum_{w \in W} P(w) \times w(s).$$

Now, the imaging on a logical sentence s is the process of moving probabilities from the worlds where s is false to the most similar worlds where s is true. Imaging on s creates a new probability distribution P_s, which is defined as follows:

$$P_s(w') = \sum_{w \in W} P(w) \times I(w, w'),$$

where

$$I(w, w') = \begin{cases} 1 & \text{if} \quad w' = w_s \\ 0 & \text{otherwise.} \end{cases}$$

In generalized imaging,

$$P_s(w') = \sum_{w \in W} P(w) \times P^w(w') \times I(w, w'),$$

where

$$I(w, w') = \begin{cases} 1 & \text{if} \quad w' \in w_s \\ 0 & \text{otherwise,} \end{cases}$$

and $P^w(w')$ is the weight of the accessibility relation R between w and w' or it is the portion of the probability $P(w)$ that must be transferred to w' because, in generalized imaging, w_s is a set of worlds and $P(w)$ must be distributed on all worlds in w_s.

The probability of a condition $s_1 \rightarrow s_2$ is defined as follows [4, 21, 23]:

$$P(s_1 \rightarrow s_2) = P_{s_1}(s_2) = \sum_{w \in W} P_{s_1}(w) \times w(s_2). \tag{1}$$

Some researchers represent the retrieval decision $d \rightarrow q$ as a condition. Therefore, the previous equation is an important technique to estimate the uncertainty $U(d \rightarrow q)$ in a probabilistic way.

4.2 Path's Cost-Based Uncertainty

Nie [67, 68] uses $\mathcal{PML}$, or Kripke's formal semantics [50], to refine the LUP that is proposed by van Rijsbergen [94]. Nie defines his logic-based IR model within the formal semantics layer. The inference mechanism in Nie's model has no direct correspondence in the formal language of $\mathcal{PML}$.

Nie assumes that a document d is a set of logical sentences or, equivalently, it corresponds to a possible world. A query q is a set of propositions or a logical sentence. Any proposition a is true in d if it appears in d. To evaluate $U(d \rightarrow q)$, the model starts from the initial world d (or d_0); if q is not satisfied in d_0, then, using accessibility relations, the model moves from d_0 to d_1. If q is still not satisfied in d_1, the model moves from d_1 to d_2, and so on, until the model arrives to d_n that satisfies q. Actually, there are many paths from d_0 to d_n. Therefore, to calculate the certainty of the implication $U(d \rightarrow q)$, the path of the minimal distance is chosen. A general measure to calculate the distance from d_0 to d_n, denoted $dis(d_0, d_n)$, is defined as a function of the elementary distances $dis(d_i, d_{i+1})$.

In a symmetric approach, instead of considering documents as possible worlds, it is possible to consider queries as possible worlds. In this case, the model must find the path from q_0 to q_n that satisfies d. In both cases, the accessibility relations between possible worlds could be linguistic relations. For example, the model could move from q_i to q_{i+1} by adding the synonymous terms of the indexing terms of q_i to q_{i+1} (query expansion). Nie reformulated the LUP as follows:

> *Given any two information sets x and y; a measurement of the uncertainty of $y \rightarrow x$ relative to a given knowledge set K, is determined by the minimal extent E to which we have to add information to y, to establish the truth of $(y + E) \rightarrow x$.*

The implication $d \rightarrow q$ is thus equivalent to finding a path from the initial possible world d to another possible world d_n that satisfies q, whereas, the uncertainty $U(d \rightarrow q)$ is equivalent to the cost or the total distance of this path. The matching score $RSV(d, q)$ according to Nie is a function F of the two uncertain implications:

$$RSV(d, q) = F[U(d \rightarrow q), U(q \rightarrow d)].$$

Nie's model had not been applied to standard test collections. However, if we suppose that accessibility relations refer to inter-term semantic relations, then finding the path from d_0 to d_n or from q_0 to q_n is equivalent to document or query expansion, respectively.

4.3 Imaging (Probability)–Based Uncertainty

In the studies of Nie [67, 68], the distance between worlds or the cost of accessibility relations is informally defined. However, Nie [69] defines two sources of uncertainty in a formal way[7]:

- The truth of a proposition a in a world w is not binary, $w(a) \in [0, 1]$. The function $w()$ is recursively defined on any logical sentence, and Nie proved that it is a probability.
- The strength of the accessibility relation between two worlds w and w' depends on the type of semantic relation for example, synonymy, generalization, and specialization—that causes the transformation of w to w', where $\sum_{w' \in W} P^w(w') = 1$.

Nie [69] uses probability (imaging) to estimate the logical uncertainty $U(d \rightarrow q)$. According to Nie, a document d is a possible world and a query q is a logical sentence. First, the model of Nie makes an imaging on d to transform probabilities from the worlds that have an accessibility relation with d, to d. To compute the logical uncertainty $U(d \rightarrow q)$, Nie uses Equation (1), where

$$U(d \rightarrow q) = \sum_{w \in W} P_d(w) \times w(q).$$

The accessibility strength $P^w(w')$ is used to compute the truth function $w()$; it is also used in the imaging process to choose the closest world to d in order to build the new probability distribution P_d. In this regard, there is a preliminary experiment conducted by Amati and Kerpedjiev [4]. After simplifying the model, they applied it to a small set of documents (103 documents). Amati and Kerpedjiev conclude that the IR model has a comparable performance to vector space and probabilistic models.

[7] For readability reasons, we use the notation proposed in Section 4.1 instead of notations used in original papers.

Unlike Nie, who represents a document d as a possible world, a query q as a logical sentence. d in that case is relevant to q *iff* there is a path from d to d_n that satisfies q [67, 68], Crestani and Van Rijsbergen [23] and Crestani [21] assume that each indexing term $t \in T$ is a possible world. A document d is true in t if t appears in d. A query q is true in t if t appears in q. Crestani and Van Rijsbergen define t_d as the closest term to the term t where d is true.

Crestani and Van Rijsbergen use the imaging technique to move probabilities from the terms that do not appear in d to the terms that appear in d. They build a new probability distribution P_d from the original distribution P by imaging on d. Crestani and Van Rijsbergen have shown that the logical uncertainty $U(d \to q)$ can be estimated as follows:

$$U(d \to q) = P_d(q) = \sum_{t \in T} P_d(t) \times t(q),$$

where $t(q) = 1$ if t appears in q or $t(q) = 0$ otherwise. For the prior probabilities of terms, Crestani and Van Rijsbergen use term-discriminative power measures, such as IDF:

$$\forall t \in T, P(t) = -IDF(t) = -\log \frac{n_t}{N},$$

where N is the total number of documents and n_t is the number of documents that are true in the possible world t. The strength of the accessibility relation between two terms t_i and t_j, which computes the similarity between them, is estimated using the Expected Mutual Information Measure (EMIM) [93].

4.4 Fuzzy Measure for Uncertainty

Nie et al. [70] define the fuzzy accessibility relations between two possible worlds and the fuzzy truth value of a proposition in a possible world in order to build an IR model. They represent a document d as a possible world and a query q as a logical sentence. They redefine the two functions $P^w(w')$ and $w(s)$ as follows:

- The function $P^w(w')$ estimates the fuzzy accessibility degree between two worlds w and w', where $P^w(w) = 1$.
- For any logical sentence s, the function $w(s)$ gives the fuzzy truth value of the sentence s in the world w. This function is built based on $C_a(w)$, which represents the fuzzy truth value of an atomic proposition a in a world w.

The retrieval decision is defined in the same way as in [67–69], whereas the logical uncertainty $U(d \to q)$ is defined as follows:

$$U(d \to q) = d(\Diamond^n q),$$

which estimates how much it is possible to find a world in the future where q is true and

$$d(\Diamond^n q) = \sup_{w \in W} \Delta\left[P^d(w), w(\Diamond^{n-1} q)\right],$$

Δ is a triangular norm function, and

$$d(\Diamond q) = \sup_{w \in W} \Delta\left[P^d(w), w(q)\right].$$

To establish the fuzzy degree of accessibility between two possible worlds $P^w(w')$, Nie et al. propose an automatic way to learn the strength of inter-terms relationships, where $P^w(w')$ equals the strength of the relation that causes the transition from w to w'. The main problem is that estimating these weights requires a non-negligible amount of user feedback information.

4.5 General Remarks on Using $\mathcal{PML}$ in IR

Before making some general remarks about using $\mathcal{PML}$ in IR, it is worth mentioning a work by Röelleke and Fuhr [80], who propose a four-valued extension of Kripke possible worlds semantics.

Even though using Kripke's semantics and its related imaging technique seem to be an interesting theoretical choice, IR models based on possible worlds and imaging have some disadvantages. These models are defined in the formal semantics side, and some operations have no direct correspondence in the formal language of the logic (syntax). For example, in a condition of the form $d \to q$, the connective $\to$ has no correspondence in the language of $\mathcal{PML}$. In addition, most models directly define the logical uncertainty $U(d \to q)$ without defining what the connective $\to$ refers to. This point could also be assumed as an advantage because it allows us to skip the task of defining $\to$.

It is also not easy to define the prior probability distribution on worlds $P(w)$ and what it refers to. Furthermore, defining accessibility relations and their related cost or distance measure is also a heavy task that needs a lot of study. Finally, experiments on large document collections show poor retrieval performance of these models [99].

5 OTHER LOGICS-BASED MODELS

5.1 Description Logic ($\mathcal{DL}$)–Based Models

Description Logic ($\mathcal{DL}$) is a family of languages to represent knowledge. It is more expressive than $\mathcal{PL}$ but it has more efficient reasoning than First-Order Logic ($\mathcal{FL}$). While the reasoning in $\mathcal{FL}$ is undecidable [84], there are some families of $\mathcal{DL}$ having polynomial time complexity [49], deterministic polynomial time [61], PSpace-complete [76], or even $O(n \log n)$ in the ALN family [88]. Before presenting the IR models that use $\mathcal{DL}$ as an underlying logical framework, we briefly introduce the formal language and semantics of $\mathcal{DL}$.

5.1.1 Description Logic (Preliminary). A logical sentence in $\mathcal{DL}$ is a formulation of building blocks and connectives (operators) according to predefined rules. Different families of $\mathcal{DL}$ can be defined based on the allowed operators. Of course, in any family of $\mathcal{DL}$ there is a trade-off between the expressive power and the efficiency of related algorithms. $\mathcal{DL}$ contains three building blocks:

- *Individuals*, which are concrete objects, for example, Alex, Bob, and so on.
- *Concepts*, which define classes of objects, for example, Dogs, Cats, White, and so on.
- *Roles*, which define the role of objects or classes in relations, for example, Husband, Author, and so on.

Building blocks are linked through a set of allowed operators. Concerning concepts, there are many operators, for example, *Intersection* ($\sqcap$) and *Union* ($\sqcup$). For instance, $Dogs \sqcap White$ defines the white dogs concept. The two quantifiers *Universal* ($\forall$) and *Existential* ($\exists$) are also used to link roles with concepts, for example, $\forall Author.Human$ means that the authors of any object are humans. Two types of reasoning are defined in $\mathcal{DL}$:

- *Subsumption* between two concepts or roles ($\sqsubseteq$), for example, $Dogs \sqsubseteq FourLegsAnimals$ means that all dogs are four-legged animals but not the inverse.
- Role or concept *assertion* (:), which links concepts and roles to their individuals, for example, $Alex : Dogs$ means that Alex is a dog.

The formal semantics or interpretation gives meaning to logical sentences in $\mathcal{DL}$. We define a set of elements Δ^I that will represent the domain of interpretation; then, we define the interpretation function $.^I$, which maps

- every individual a to an element in the domain Δ^I, where $a^I \in \Delta^I$.
- every concept C to a subset of elements $C^I \subseteq \Delta^I$. Two special concepts exist: the concept that contains all individuals ($\top$), where $\top^I = \Delta^I$, and the empty concept ($\bot$), where $\bot^I = \emptyset$.
- every role R to a subset of domain Cartesian product $R^I \subseteq \Delta^I \times \Delta^I$.

Concerning operators, their formal semantics is defined as follows:

- $(C \sqcup D)^I = C^I \cup D^I$
- $(C \sqcap D)^I = C^I \cap D^I$
- $(\neg C)^I = \Delta^I \setminus C^I$
- $(\forall R.C)^I = \{x \in \Delta^I | \forall y \in \Delta^I, (x,y) \in R^I, y \in C^I\}$
- $(\exists R.C)^I = \{x \in \Delta^I | \exists y \in \Delta^I, (x,y) \in R^I, y \in C^I\}$

We say that an interpretation I is a model of a logical sentence s *iff* s is true in I, denoted $I \models s$. The truth of a logical sentence in an interpretation I is determined according to the following rules:

- $I \models a : C$ *iff* $a^I \in C^I$.
- $I \models (a,b) : R$ *iff* $(a^I, b^I) \in R^I$.
- $I \models C \sqsubseteq D$ *iff* $C^I \subseteq D^I$.

A Knowledge Base (KB) is the pair (T, A), where T is the terminological box that contains the definitions of concepts, roles, and the relations between them, for example, $C \sqsubseteq D$, and A is the assertion box that contains the relations between concepts and roles on one hand and individuals on the other hand, for example, $a : C$, $(a,b) : R$.

5.1.2 IR Models. In IR, Meghini et al. [63] use a special kind of $\mathcal{DL}$, called MIRTL, to formulate the logical implication $d \rightarrow q$, thus building an IR model. According to them, a document d is represented as an individual; in other words, a document d is the only instance of a concept D that is the intersection of all concepts where d is asserted to be an instance of, formally, $D \doteq \sqcap_{d:X_i} X_i$. A query q is represented as a concept. The relevance between a document d and a query q is mapped either to

- the individual d is an instance of the concept q, denoted $d : q$;
- or the concept D that only contains the document d is subsumed by the concept q, denoted as $D \sqsubseteq q$.

The two decisions $d : q$ or $D \sqsubseteq q$ are binary decisions, which means that the knowledge base KB either satisfies them or not. Therefore, $\mathcal{DL}$ alone does not support the partial or uncertain decision, which is more adequate to IR. To calculate the logical uncertainty $U(d \rightarrow q)$, $\mathcal{DL}$ is extended by probability [61, 85]. Sebastiani [85] extends $\mathcal{DL}$ by adding two types of probabilities:

- The degree of belief in an assertion γ, denoted $w[\gamma]$ *relop* t, where *relop* could be $=$, $\leq$, and so on (*subjective measure*). For example, $w[a : C] = 0.5$ means that the degree of our belief that the individual a is an instance of the concept C is 0.5. Conditional degree of belief could also be defined. For example, $w[a : C | a : D] = 0.6$ means that the degree of our belief that the individual a is an instance of the concept C knowing that a is an instance of the concept D is 0.6. The same interpretation holds for subsuming relations, for example, $w[C \sqsubseteq D] = 0.1$, $w[C \sqsubseteq D | C \sqsubseteq E] = 0.2$.
- Statistical information, denoted $w_{\langle x \rangle}[C]$, where C is a concept (*objective measure*). For example, $w_{\langle x \rangle}[C] = 0.3$ means that the probability that a randomly picked individual is an instance of C is 0.3. Conditional statistical information could be also defined. For example,

$w_{\langle x \rangle}[C|D]$ means that the probability that a randomly picked individual is an instance of C knowing that it is an instance of D is 0.2.

MIRTL logic [63], in addition to the above two probabilistic extensions, defines a new logic P-MIRTL [85]. In order to give a formal semantics to the new logic P-MIRTL, Sebastiani uses the notion of possible worlds in addition to the formal semantics used in $\mathcal{DL}$. To define the formal semantics, Sebastiani defines the tuple $M = \{\Delta, I, P_{dom}, P_{int}\}$, where

- Δ is a set of individuals (domain of interpretation).
- I is a set of interpretations based on Δ.
- P_{dom} is a probability distribution defined on the elements of Δ.
- P_{int} is a probability distribution defined on the elements of I.

The system's degree of belief in a probabilistic statement t in an interpretation $i \in I$, denoted $[t]_{(M,i)}$, is defined as follows:

- Statistical information for concepts: $[w_{\langle x \rangle}C]_{(M,i)} = \sum_{a \in C^i} P_{dom}(a)$ is the sum of all probabilities of those individuals that are instances of C.
- Statistical information for roles: $[w_{\langle x_1, x_2 \rangle}R]_{(M,i)} = \sum_{(a_1,a_2) \in R^i} P_{dom}(a_1) \times P_{dom}(a_2)$ is the sum of all joint probabilities of those tuples that are instances of R.
- The degree of belief in an assertion: $[w(\gamma)]_{(M,i)} = \sum_{i \in I, (M,i) \models \gamma} P_{int}(i)$ is the sum of probabilities of those interpretations that satisfy the assertion γ.

The non-probabilistic statements of P-MIRTL have the same semantics as in MIRTL. In P-MIRTL, the uncertain IR implication $U(d \rightarrow q)$ is mapped to compute the degree of belief in $d : q$ or $D \sqsubseteq q$ based on a specific domain of interpretation, a set of interpretations on that domain, a probability distribution defined on the elements of the domain, and a probability distribution defined on the interpretations.

$$U(d \rightarrow q) = [w(d : q)]_{(M,i)} \quad \text{OR} \quad U(d \rightarrow q) = [w(D \sqsubseteq q)]_{(M,i)}$$

To our knowledge, the MIRTL- or P-MIRTL-based IR models have never been tested in true retrieval situations or on standard test collections.

Later, Meghini and Straccia [64] proposed a description logic with four-valued semantics to model IR; in this case, both *lack of information* and *contradiction/inconsistency* are modeled.

There are several appealing reasons to use $\mathcal{DL}$ in IR. On the one hand, $\mathcal{DL}$ has more expressive power than $\mathcal{PL}$. $\mathcal{DL}$ enables IR models to represent documents and queries as objects (concepts or individuals) that may have, in addition to a set of indexing terms, some other properties, for example, a list of authors and publishing date. On the other hand, $\mathcal{DL}$ is originally a knowledge representation language, which means that there are many knowledge bases represented using $\mathcal{DL}$. Therefore, using $\mathcal{DL}$ to represent documents and queries enables us to easily integrate any KB described by $\mathcal{DL}$ into the IR model. However, building IR models based on $\mathcal{DL}$ has some disadvantages. It is not easy to automatically transform documents and queries from their original textual or multimedia form to concepts or individuals in $\mathcal{DL}$. Recently, there have been many studies to annotate raw text, for example, with named entities extracted from particular websites [38], with DBpedia named entities and RDF triples [31, 43], and with UMLS concepts or semantic types [62]. In addition, many families of $\mathcal{DL}$ have unpractical reasoning algorithms[8]. Furthermore, the inference in $\mathcal{DL}$ is originally a binary decision; therefore, $\mathcal{DL}$ in its original form is not suitable for IR. At the same time, extending $\mathcal{DL}$ by some notions of probability or possibility

[8] www.cs.man.ac.uk/ezolin/dl/.

make the extended logic hard to understand and very complex to reason. For example, the two probability distributions P_{dom} and P_{int} in [85] are hard to define.

5.2 Models Based on Conceptual Graphs (First-Order Logic)

Chevallet and Chiaramella [16] build an IR model based on Conceptual Graphs (CGs). According to them, a document d is a CG or, equivalently, a logical sentence in $\mathcal{FL}$, and a query q is also a CG. The retrieval decision is as follows: d is relevant to q *iff* there exists a projection $\pi(q)$ of q on d or *iff* d contains a sub-graph $\pi(q)$ that is a possible specialization of q. From the definition of the projection operation, the retrieval decision is equivalent to checking whether $d \leq q$ or $\models [\Phi(d) \supset \Phi(q)]$, where $\Phi(x)$ corresponds to the logical sentence of the conceptual graph x.

The uncertainty of the retrieval decision $U(d \rightarrow q)$ is estimated using Kripke's semantics [50] or possible worlds semantics, where a document is a possible world and the accessibility relation between worlds is one of the four main operations on CGs (copy, restriction, simplification, or join). The cost of an accessibility relation between two worlds w and w', denoted $P^w(w')$, is related to the operation that causes this transformation from w to w'. Costs are arbitrarily assigned to each operation. The uncertainty $U(d \rightarrow q)$ is estimated as in [67], where $U(d \rightarrow q)$ is the cost of the path from d to d_n in which $d_n \leq q$. To our knowledge, the model has not been tested on standard collections.

CG is a powerful formalism to express the content of documents and queries. However, it is hard to transform the content of documents and queries into CGs, and the projection operation is NP-complete [15].

5.3 Probabilistic Datalog-Based Models

Datalog is a predicate logic developed in database field. Probabilistic Datalog is an extension of deterministic Datalog using probability [34, 35]. The main difference between deterministic and probabilistic Datalog is that probabilistic Datalog defines *probabilistic ground facts* in addition to *deterministic ground facts*, which are classical predicates. Probabilistic ground facts have the form αg, where g is a deterministic ground fact (classical predicate) and $0 < \alpha \leq 1$ is the probability that the predicate g is true. Deterministic ground facts are a special case of probabilistic ground facts, where α is always 1.

Ground facts are supposed to be probabilistically independent events. Assume that $\alpha_1 g_1$ and $\alpha_2 g_2$ are two probabilistic ground facts, where $g_1 \neq g_2$. Then, the following probabilistic ground fact is correct:

$$(\alpha_1 \times \alpha_2) \quad (g_1 \wedge g_2).$$

However, it is also possible to consider some ground facts (events) to be mutually disjoint [35]. In this case, the following is correct:

$$(\alpha_1 + \alpha_2) \quad (g_1 \vee g_2).$$

Probabilistic Datalog is used in IR [34, 52], where a document d is a set of probabilistic ground facts of the form α *term*(d, t), which means that the document d is indexed by the term t and α is the probability that the predicate *term*(d, t) is true or, equivalently, α is the probability that the document d is about the term t. Fuhr [34] extends the previous definition of d, where the terms that index d can be inferred using the predicate *about* and the following inference rules:

$$\begin{array}{l} about(D,T) : -term(D,T) \\ about(D,T) : -link(D,D') \wedge about(D',T), \end{array}$$

where D, D', T are variables and the predicate $link(D, D')$ refers to the fact that D and D' are explicitly or implicitly related, for example, a hyperlink. The index of a particular document d is $about(d, T)$. The query q is a Boolean query.

The retrieval decision is defined as an inference rule. However, there are two main forms of the retrieval inference rule based on the connectives between query terms:

- Conjunction $q = t_1 \wedge t_2$:

$$q(D) : -about(D, t_1) \wedge about(D, t_2),$$

  where D is a variable.
- Disjunction $q = t_1 \vee t_2$:

$$q(D) : -term(D, t_1)$$
$$q(D) : -term(D, t_2).$$

In [36], Fuhr and Röelleke propose a four-valued extension of probabilistic Datalog.

The main advantage of using probabilistic Datalog is the clear connection between IR and database, where probabilistic Datalog is an extension of deterministic Datalog that is used in the database. However, one must first assign probabilities to ground facts, which is not a simple task. For example, Röelleke et al. [81] proposed a modular way to estimate these initial probabilities. They introduced the relational Bayes operator to probabilistic relational algebra.

5.4 Fuzzy Logic–Based Models

In this section, we focus on IR models that make use of fuzzy logic as an underlying logical framework. After presenting a few basic definitions related to fuzzy logic and fuzzy implications, we introduce the main application of fuzzy logic to IR.

5.4.1 Fuzzy Set Theory and Fuzzy Logic. The core notion of fuzzy sets is partial membership. The definition of fuzzy set provided in 1965 by L. A. Zadeh in his seminal paper [97] is the following. A fuzzy subset F of a reference domain U is characterized by a membership function associating a real number in the interval $[0, 1]$ with each element in U. The value of the membership function for a given element u represents, then, its degree of membership to F. The fuzzy set F is thus defined as a mapping $\mu_F : U \rightarrow [0, 1]$, and it offers a generalization of the characteristic function of a classical set $A : U \rightarrow \{0, 1\}$.

Fuzzy logic is a logic of vagueness that has been formalized to the aim of dealing with vague knowledge and to support approximate reasoning. The concept of vague predicate is central in fuzzy logic. In classical logic, a unary predicate identifies a subset of the universe of discourse; analogously in fuzzy logic,. a vague predicate identifies a fuzzy subset of the universe of discourse.

For example, the unary predicate *young* identifies the fuzzy subset representing the concept *young* over the possible numeric values of the age of a human being. In a given interpretation that assigns a numeric value to the variable x, the truth value of $young(x)$ is the degree of membership $\mu_{young}(x)$ of x to the fuzzy set *young*. Then, for the case in which the information in a considered interpretation is precise, the truth value of a vague predicate is a number in $[0, 1]$. In the case of vague information, the truth value is a fuzzy truth value that may be approximated by means of two values: a necessity and a possibility degree [29].

A fuzzy implication operator is a binary operator defined on $[0, 1] \times [0, 1]$ taking values on $[0, 1]$. Several definitions of the fuzzy implication operator have been provided in the literature, among which are the following:

- $a \rightarrow_{RG} b = 1$ if $a \leq b$, 0 otherwise
- $a \rightarrow_{Gd} b = 1$ if $a \leq b$, b otherwise

- $a \rightarrow_{Gg} b = 1$ if $a \leq b$, b/a otherwise
- $a \rightarrow_{L} b = 1$ if $a \leq b$, $1 - a + b$ otherwise
- $a \rightarrow_{D} b = \max(1 - a, b)$

in which RG, Gd, Gg, L, and D stand for Rescher-Gaines, Godel, Goguen, Lukasiewicz, and Dienes, respectively.

For some definitions of the implication operator, an interpretation of the interaction between the two arguments (antecedent and consequent) has been outlined. By a first interpretation, the value a is seen as a threshold that has to be reached by the value b: if the threshold is reached, the implication operator produces the value 1. If the threshold is not reached, a penalty is applied. The implication operators having this behavior are named R-implications (residuated implications); some examples are offered by the Godel, Goguen, and Lukasiewicz implications. These implications, in the case that the threshold is not reached, produce the values 0, b, b/a, and $(1 - a + b)$, respectively. A second interpretation is connected to the Dienes implication: the behavior of this implication is that the lower the a value, the higher the value of the implication (and the higher the a value the closer to b is the value of the implication). If we consider a formula with implications connected by the AND operator, the a value can be considered as an importance weight for the b value. In fact, as 1 is the neutral element for the min operator, the implications with a low a value do not have a great influence on the result of the aggregation. The implication operators having this behavior are named S-implications.

As we will see in Section 5.4.2, the distinct behavior of the considered implication operators makes their choice very important in the logical formulation of the weighted Boolean models. This choice is, in fact, related to the modeled semantics of query weights.

As outlined in [29], "*There is some confusion in the fuzzy literature on the potential of fuzzy sets for handling uncertainty ... this confusion has been around for a long time and has hampered the sound development of many-valued logics and their use in knowledge representation.*" In fact, fuzzy sets are employed to model graduality in properties; for this reason, membership degrees may represent degrees of truth of fuzzy propositions, and not degrees of uncertainty of propositions. As outlined in [28]: "*In the scope of knowledge representation, choosing a truth set is a matter of convention like choosing the range of a variable, while uncertainty is a meta-notion which reflects incomplete or contradictory knowledge.*" Again, in [29] it is outlined: "*In classical logic the convention is that truth is binary. Fuzzy set theory (and before, multiple-valued logics) has modified this convention. This shift in convention does not entitle degrees of truth to be systematically interpreted as degrees of uncertainty.*" This distinction has been materialized by two different extensions of classical logic that rely on the notion of a fuzzy set:

- Multi-valued logics, which handle vague propositions. "*The underlying algebraic structure is weaker than a Boolean algebra, and can be consistent with truth-values that lie in the unit interval and remain truth-functional*" [29];
- Possibilistic logic, where a crisp (non-vague) proposition p has a degree of certainty $N(p)$ associated with it ($N(p) = 1$ if and only if p is surely true, while $N(p) = 0$ expresses the complete lack of certainty that p is true). Moreover $N(\neg p) = 1 - PI(p)$ where $PI(p)$ is the degree of possibility of proposition p.

5.4.2 A Logic Interpretation of the Boolean Model and of Extended Boolean Models. A first work in which fuzzy logic has been employed as a useful formal framework to formalize the IR process is [72]. The same approach was independently proposed a few years later in [12, 91], but relied on the notion of fuzzy relation. In [72], the evaluation process of Boolean queries has been first re-expressed within first-order logic. Due to the set-theoretic assumptions at the basis of the Boolean model, the interpretation of a Boolean query relies on the set inclusion operator: a document d is estimated relevant to a query q composed of only a query term t if t is included in the representation of document d (a set of terms), while a document d is estimated relevant to a query $q = \neg t$ if term t is not included in the representation of document d. In a logical framework, it is straightforward to interpret the set inclusion operator by means of the material implication connective: the evaluation of a non-negated query term can be in this context equivalently represented as $t_q \rightarrow t_d$, that is, the presence of term t in query q (denoted by t_q) implies the presence of term t in the representation of document d (denoted by t_d). Similarly, the evaluation of a negated query term is represented as $not(t_q \rightarrow t_d)$. In the set-theoretic interpretation of the Boolean IR model, the binary degree of relevance of term t with respect to a document d is assessed as the membership of t in the representation of document d. Similarly, in the logical framework, the binary degree of relevance of query term t to d is assessed as the degree of truth of the implication $t_q \rightarrow t_d$ under the interpretation corresponding to the document and the considered query.

It is clear tha, by this logical interpretation of the Boolean model, what has to be assessed is not $d \rightarrow t$, as logical models seen so far do; in fact, this interpretation works at query term granularity. At this point, it is straightforward to interpret the evaluation of queries with weighted terms as a generalization of the material implication by means of fuzzy implications. In this context, the evaluation of any generic Boolean query can be easily reinterpreted in the logical framework by considering that the AND, OR, and NOT operators correspond to the logical connectives $\wedge, \vee, \neg$ respectively. While the Boolean model computes the inclusion of the query terms in the document (in the case of conjunctive queries), its fuzzy extensions compute a degree of inclusion that makes it possible to rank the retrieved documents in decreasing order of estimated relevance to the considered query. In Fuzzy Extended Boolean models, documents are represented as fuzzy subsets of the term dictionary T. Moreover, query terms are weighted and different semantics associated with query weights originates distinct fuzzy models. The degree of membership of a term to the fuzzy subset representing a document d are the usual index term weights, with the constraint to belong to the unit interval $[0, 1]$.

In the fuzzy generalization of the logic interpretation of the Boolean model, the evaluation of each (weighted) query term t is expressed as $\mu_t(q) \rightarrow \mu_d(t)$, where → is a fuzzy implication. The connectives of conjunction and disjunction are interpreted as t-norms and t-conorms, respectively, usually the min and the max operators.

The distinct behavior of the considered implication operators makes their choice very important in the logical formulation of the weighted Boolean models. This choice is related to the modeled semantics of query weights.

5.5 Situation Theory ($\mathcal{ST}$)–Based Models

Situation Theory ($\mathcal{ST}$) is a formal framework to model or represent information [9, 10]. Instead of studying whether a piece of information is true or false, $\mathcal{ST}$ studies what makes this piece of information true. According to $\mathcal{ST}$, information tells us that relations hold or not between objects. Therefore, the atomic information carriers are called *infons*, and an infon is a structure $\langle\langle R, a_1, \ldots, a_n; i\rangle\rangle$ that represents the information that the relation R holds between the objects $a_1, \ldots, a_n$ if $i = 1$, or it does not hold if $i = 0$. A *situation* is a partial description of the world, or it can be defined as a set of infons [45]. A situation s *supports* an infon f, denoted $s \models f$, if f is made true by s or, equivalently, if f can be deduced from the set of infons of s.

$\mathcal{ST}$ generalizes a set of situations having common characteristics into a *type of situation*. The notation $s : A$ refers to the fact that the situation s is of the type A. A *constraint* is a relation between two types of situation, A and A', denoted $A \Rightarrow A'$. A constraint such as $A \Rightarrow A'$ means that the occurrence of a situation $s : A$ implies the existence of a situation $s' : A'$. Uncertainty is represented as a *conditional constraint*, denoted $A \Rightarrow A'|B$, which means that the constraint $A \Rightarrow A'$ is fulfilled under some conditions B, where B itself can be a type of situation.

Lalmas and Rijsbergen [53] build an IR model–based on $\mathcal{ST}$. According to them, a document d is a situation and a query q is an infon or a set of infons. The document d is relevant to a query q *iff* d supports q, denoted $d \models q$. The uncertainty of the previous retrieval decision is estimated based on the conditional constraints

$$U(d \models q) = \begin{cases} 1 & \text{if} \quad d \models q \\ \max_{\{D' | (D \Rightarrow D' | B), \exists d' : D', d' \models q\}} \delta(D, D') & \text{otherwise,} \end{cases}$$

where D is the type of d, D' is another type related to D under some conditions B, and $\delta(D, D')$ is defined as follows:

$$\delta(D, D') = \begin{cases} 1 & \text{if} \quad D \Rightarrow D' \\ 0 < \alpha < 1 & \text{if} \quad D \Rightarrow D' | B \\ 0 & \text{otherwise.} \end{cases}$$

The function δ is based on the conditions under which it is possible to find another document d' of the type D', where d implies d'. This definition of uncertainty is very close to the definition of Nie [67].

Huibers and Bruza [45] present two examples, a Boolean model and coordination-level matching model, on how to build the situations that represent documents and queries, and what the retrieval decision $d \models q$ concretely means. Huibers and Bruza also use $\mathcal{ST}$ to define the *aboutness* relation between a document and a query in order to build a meta IR model capable of formally comparing IR models.

The main disadvantage of $\mathcal{ST}$-based IR models is the difficulty of automatically building meaningful infons for representing the content of documents and queries. Moreover, uncertainty needs to be defined in a less abstract way in order to build an implementable version of these models.

5.6 Default Logic–Based Models

Default logic is used to represent semantic relations between objects, for example, synonymy and polysemy. It is used in IR to represent some background knowledge or thesauri knowledge. Default logic defines a special inference mechanism called *default rules*, denoted $\frac{\varphi : \beta}{\psi}$, which means that, in a particular context, if φ is true and $\neg\beta$ cannot be inferred, then infer ψ.

Positioning is the process of changing a logical sentence using default rules. For example, assume a logical sentence s is $a \wedge b$, where a and b are propositions. Assume the following default rule $\frac{a : \neg c}{e}$, where c and e are also propositions. The positioned sentence s' of s, according to the previous default rule, is $a \wedge b \wedge e$.

Hunter [46] uses default logic to build an IR model. A document d is a clause or, equivalently, a conjunction of literals, where each literal is either a proposition or its negation. Each indexing term corresponds to a proposition. A query q is a logical sentence. The retrieval decision is the material implication where d is relevant to q *iff* $d \supset q$.

Semantic relations between terms are represented through default rules. For example, assume the following default rule:

$$\frac{car \wedge transport : \neg rail}{automobile}.$$

This default rule means that if the initial representation of a document d satisfies the two propositions *car* and *transport*, and the term *rail* cannot be inferred from d, then d can be expanded by the term *automobile*.

The uncertainty of $d \supset q$ is estimated through finding the positioned version d' of d using the set of default rules, where $d' \supset q$. Default logic–based IR models have the ability to build *context-dependent* models and to *qualitatively* define the logical uncertainty $U(d \rightarrow q)$. However, these models suffer from some disadvantages. The automatic learning of default rules is not an easy task and is error prone. In addition, there is no quantitative measure of uncertainty.

6 LOGIC IN IR AND KNOWLEDGE MANAGEMENT: RECENT TRENDS

In this survey, we have presented the role that logic plays in IR. From this picture, it emerges that a great interest in logic and IR has motivated several proposals in the 1990s. The reason why logical approaches for IR were less studied for a while is that the first models were not that effective. However, recently, there has been a resurgence of interest in employing logic in IR; this has been motivated by its effectiveness as measured on large datasets and by the important role that logic plays in close tasks related to information and knowledge management.

6.1 Recent Logical IR Systems

During the last decade, several approaches have tackled the issue of employing logic in information retrieval. One of the interests in using logic for IR is the fact that the results are predictable and explanatory. For example, Description Logic is explicitly used to formalize knowledge in medical IR systems [13, 77]. Belief revision [54] is being investigated to cope with vague queries. Logic is also used inside the IR system for specific tasks, for example, ranking [82], query representation [58], estimation of query difficulty [57], and meta-fusion of classical IR [39]. Zerarga and Djouadi [98] present the generalization that the logical approach provides of various IR models. Zuccon et al. [100] present logical imaging as a quantum theory interpretation, using the K kinematic operator.

6.2 Effectiveness of Logical IR Systems

Concerning the evaluation of logical IR models, there have been many recent attempts to apply logical IR models to large-scale document collections. It is important to recall that the first IR models were not able to manage such collections and were evaluated on smaller datasets. For example, Picard and Savoy [74] (Section 3.3) applied their model to the corpus CACM[9], which is a very small collection of documents (only 3,204). With respect to the $\mathcal{PML}$-based IR models (Section 4), Amati and Kerpedjiev [4] have conducted preliminary experiments on Nie's model [67, 68] (Section 4.2) by applying it to a small test collection (103 documents); they conclude that the model has a retrieval performance comparable to vector-space and probabilistic models. The imaging-based model of Crestani and Van Rijsbergen [23] (Section 4.3) has been tested with large [24] and small [25] document collections. In [24], Crestani et al. review challenges in front of applying the imaging-based model to a large document collection (a part of TREC4[10], about 165,000 documents), where they especially mention the problem of probability transfer or obtaining P_d based on P. In general, the results are not promising, maybe because of the simplifications that they are obliged to do in order to obtain an operational version of their imaging-based model. Nie et al. [70] (Section 4.4) applied their model to the corpus CACM, which is a very small collection of documents (only 3,204). Thus, it is not possible to draw clear conclusions from their study. Since reasoning is complex in $\mathcal{FL}$, there are very few attempts to implement $\mathcal{FL}$-based IR models. For example, the model of [16] (Section 5.2) is tested only on a small set of documents, where their conceptual graphs are manually obtained.

Losada and Barreiro [60] were among the first to apply their logical model (Section 3.2) to a large TREC collection (about 170,000 documents). They express queries (topics) as DNF sentences

[9] http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/.

[10] Text REtrieval Conference (http://trec.nist.gov/).

as follows: each part of the query (i.e., title, description, narrative) corresponds to one clause, and q is $title \vee description \vee narrative$. Also, documents are expressed by means of DNF logical sentences. The authors show that the retrieval effectiveness of their model is better than the one of the classical *tf.idf*-based vector space model. However, they do not compare their model to other IR models, and they do not experiment with other document collections. They also claim that most of the gain in performance comes from building a rather complex document and query representation. Zuccon et al. [99] apply the logical imaging–based model of Crestani and Van Rijsbergen [23] (Section 4.3) also on TREC collections and show that imaging-based IR models, as presented in [23], have a retrieval performance lower than the performance of some classical IR models. Recently, Abdulahhad et al. [3] applied their model (Section 3.4) to large-scale and standard corpora, for example, ImageCLEF[11] and TREC. The authors showed that their IR model performs better than classical IR models, for example, the vector space model [89], probabilistic model [79], language models [75], and information model [18]. One important conclusion of these recent works is that it is possible for logical models to have operational and effective implementations.

6.3 The Role of Logic in Information and Knowledge Management

In last 20 years, logic has been extensively employed as a formal language to define models for applications related to information and knowledge management.

In 1998, Tim Berners Lee claimed that the semantic web could be designed based on logic[12]. Since then, efforts have been spent to implement this idea and the definition of ontologies and of reasoning mechanisms on them relies on formal languages based on various kinds of logic: for example, OWL-DL[13] (based on description logic) and SWRL[14]. In 2010, John F. Sowa published a paper that analyses "*the role of logic and ontology in language and reasoning*" [90]. Sowa asserts that "*projects in artificial intelligence developed large systems based on complex versions of logic, yet those systems are fragile and limited in comparison to the robust and immensely expressive natural languages.*" He proposes an analysis that can lead "*to a more dynamic, flexible, and extensible basis for ontology and its use in formal and informal reasoning.*"

More recently, the knowledge graph has been defined as a means to represent knowledge and to reason on it. This term was introduced by Google with reference to the knowledge base that the company defined to enhance its search engine with semantic information. Since then, many companies have engaged in defining their own knowledge graphs. As pointed out in Bellomarini et al. [11], "*The reasoning core of a KGMS needs to provide a language for knowledge representation and reasoning (KRR)*" that should "*achieve a careful balance between expressive power and complexity.*" To this aim, the authors describe the VADALOG language, which they have defined.

Logic also plays an important role in ontology-based IR systems. These are hybrid systems in which the query is first represented in the logical format of the knowledge base, typically description logic [14, 32]. The output of the logical inference is then used to feed a simple bag-of-words IR system. These proposals show that at least logic can be used outside of typical IR models. We think there may be a new research path that should study the fusion of ontology logic deduction with the IR model instead of just gluing two different systems: deductive on one side and statistical on the other.

Hence, we believe that formal logics are useful tools to formally represent and integrate knowledge in IR models and also useful to reproduce the inferential nature of the retrieval process.

[11] http://www.imageclef.org/.
[12] www.w3.org/DesignIssues/Logic.html.
[13] www.w3.org/TR/owl-guide/.
[14] www.w3.org/Submission/SWRL/.

These are the main motivations beyond the choice to study logic-based IR models and to build new models. The above synthesis shows the current interest of logic and IR based on the more recent findings and research developments in different and strongly related applications.

7 CONCLUSION

In this survey, we have presented a broad picture of logic-based IR models. Different families of logics have been considered to model IR, including classical and non-classical logics. For each logic-based IR model discussed in this survey, we have reported the definitions in terms of the five main components of any IR model: term, document, query, retrieval decision, and uncertainty.

Logical IR models can be considered as theoretical guides for effective IR systems. However, they address a few main issues that are related here: binary truth values, the definition of the implication, and the generation of document and query.

Most logics—except, for example, Fuzzy Logic—rely on binary truth values: sentences are represented as classical (True/False) propositions and uncertainty is modeled by means of an external component, with the consequence of producing a hybrid framework. Moreover, to deal with IR notions such as term-weighting, there is a need to extend the formal language of the underlying logic.

Some models do not give a precise definition of the implication $\rightarrow$; instead, they directly deal with uncertainty $U(d \rightarrow q)$, as in the case of modal logic–based models. Also, only some logic-based IR models logically represent their IR components: term t, document d, query q, retrieval decision $d \rightarrow q$, and ranking $U(d \rightarrow q)$.

Finally, it is difficult in some cases to automatically generate document and query representations. This is the case for conceptual graph–based models or situation theory–based models.

These findings aside, logical IR models provide a framework for understanding and describing the main IR components (document, query, matching) and for the theoretical comparison of IR models, for example, through meta-models. In addition, defining documents and queries as logical sentences and the retrieval process as an inference makes logic-based IR models general and flexible. Furthermore, most logical IR models are capable of explicitly or implicitly integrating external knowledge into an IR model, for example, default logic. Logical IR models now provide implementation that enables their evaluation on large-scale datasets. The advent of semantic web, knowledge graph, and ontology emphasizes the place that logics and IR take in the new developments of data on the web.

APPENDIX

A FORMAL LOGICS: PRELIMINARY

Formal logics are formal systems consisting of a set of axioms and a set of inference rules (e.g., Modus-Ponens). A formal logic $\mathcal{L}$ is defined by a formal language and possibly a formal semantics.

The formal language of $\mathcal{L}$ determines the set of all well-formed sentences that can be formed based on a set of atomic elements, that is, an alphabet Ω and a set of connectives Υ, for example, conjunction $\wedge$, disjunction $\vee$, and so on. Upon this formal language, an inference mechanism $\vdash_{\mathcal{L}}$ related to $\mathcal{L}$ is defined. We say that a logical sentence s is *provable* based on a set of logical sentences Γ, denoted $\Gamma \vdash_{\mathcal{L}} s$, *iff* s can be obtained by applying the inference rules of $\mathcal{L}$ to the axioms of $\mathcal{L}$ and to the set of sentences Γ. Furthermore, $\vdash_{\mathcal{L}} s$ means that s can be obtained by applying the inference rules of $\mathcal{L}$ only to the axioms of $\mathcal{L}$. For example, let us assume that $\Gamma = \{s_1, s_1 \supset s_2\}$, where $\supset$ is the material implication. Then, in classical logics, s_2 is provable based on Γ by applying the inference rule of Modus Ponens, denoted $\{s_1, s_1 \supset s_2\} \vdash s_2$ (Modus-Ponens).

The formal semantics of $\mathcal{L}$, if available, gives meaning to the components and to the logical sentences of the formal language of $\mathcal{L}$ under a given interpretation. Let us assume that $\mathcal{L}$ is a classical logic, the formal semantics of which is defined based on a set of formal interpretations Δ. For any sentence s and any interpretation $\delta \in \Delta$, s is either true in δ, denoted $\{\delta\} \models_{\mathcal{L}} s$, or not true, denoted $\{\delta\} \not\models_{\mathcal{L}} s$[15]. Determining whether a sentence is true or not under a given interpretation depends on the formal semantics that is given to the alphabet Ω and the connectives Υ. The subset of interpretations $M(s) \subseteq \Delta$ that makes s true is called the set of models of s, denoted $M(s) \models_{\mathcal{L}} s$. The formula $M(s_1) \models_{\mathcal{L}} s_2$ or, simply, $s_1 \models_{\mathcal{L}} s_2$, means that each model of s_1 is also a model of s_2 or, equivalently, in any interpretation if s_1 is true, then s_2 is also true.

The two symbols $\vdash$ and $\models$ are related to two different sides of the logic $\mathcal{L}$. While $\vdash$ is a syntax-related or proof-theoretic notion, the $\models$ is a semantics-related or model-theoretic notion. However, if the logic $\mathcal{L}$ is *sound* and *complete*, then

- $\mathcal{L}$ is sound: If $s_1 \vdash s_2$, then $s_1 \models s_2$
- $\mathcal{L}$ is complete: If $s_1 \models s_2$, then $s_1 \vdash s_2$

Soundness means that what holds on the syntax side also holds on the semantics side. Completeness means what holds on the semantics side also holds on the syntax side.

REFERENCES

[1] Karam Abdulahhad. 2014. *Information Retrieval Modeling by Logic and Lattice. Application to Conceptual Information Retrieval.* Ph.D. Dissertation. Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique, Grenoble, France.

[2] Karam Abdulahhad, Jean-Pierre Chevallet, and Catherine Berrut. 2013. Is uncertain logical-matching equivalent to conditional probability? In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Dublin, Ireland, 825–828. DOI: https://doi.org/10.1145/2484028.2484152

[3] Karam Abdulahhad, Jean-Pierre Chevallet, and Catherine Berrut. 2017. Logics, lattices and probability: The missing links to information retrieval. *Comput. J.* 60, 7 (2017), 995–1018. DOI: https://doi.org/10.1093/comjnl/bxw034 arXiv:/oup/backfile/content_public/journal/comjnl/60/7/10.1093_comjnl_bxw034/1/bxw034.pdf

[4] Gianni Amati and Stephan S. Kerpedjiev. 1992. An information retrieval logic model: Implementation and experiments. *IEEE Transactions on Reliability* 48 (1992).

[5] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 357–389. DOI: https://doi.org/10.1145/582415.582416

[6] Horacio Arlo-Costa and Paul Egré. 2016. The logic of conditionals. In *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[7] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider (Eds.). 2003. *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press, New York, NY.

[8] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

[9] J. Barwise. 1989. *The Situation in Logic.* Center for the Study of Language and Information, Stanford, CA. http://books.google.fr/books?id=aX7RKgvpJw8C.

[10] J. Barwise and J. Perry. 1983. *Situations and Attitudes.* MIT Press, Cambridge, MA. http://books.google.fr/books?id=DCTXAAAAMAAJ.

[11] Luigi Bellomarini, Georg Gottlob, Andreas Pieris, and Emanuel Sallinger. 2017. Swift logic for big data and knowledge graphs. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17).* AAAI Press, 2–10. DOI: https://doi.org/10.24963/ijcai.2017/1

[12] Patrick Bosc, Vincent Claveau, Olivier Pivert, and Laurent Ughetto. 2009. Graded-Inclusion-Based Information Retrieval Systems. In *Proceedings of Advances in Information Retrieval: 31st European Conference on IR Research (ECIR'09), Toulouse, France, April 6-9, 2009.*Springer, Berlin, 252–263. DOI: https://doi.org/10.1007/978-3-642-00958-7_24

[13] K. Boukhari and M. N. Omri. 2017. Information retrieval based on description logic: Application to biomedical documents. In *2017 International Conference on High Performance Computing Simulation (HPCS'17).* 846–853. DOI: https://doi.org/10.1109/HPCS.2017.128

[15] Note that $\vdash_{\mathcal{L}}$ and $\models_{\mathcal{L}}$ do not belong to the formal language of $\mathcal{L}$, namely $\vdash_{\mathcal{L}} \notin \Upsilon$ and $\models_{\mathcal{L}} \notin \Upsilon$.

[14] Pablo Castells, Miriam Fernandez, and David Vallet. 2007. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. on Knowl. Data Eng.* 19, 2 (Feb. 2007), 261–272. DOI: https://doi.org/10.1109/TKDE.2007.22

[15] Michel Chein and Marie-laure Mugnier. 1992. Conceptual graphs: Fundamental notions. *Revue d'Intelligence Artificielle* 6 (1992), 365–406. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.848.

[16] Jean-Pierre Chevallet and Yves Chiaramella. 1995. Extending a logic-based retrieval model with algebraic knowledge. In *MIRO Multimedia Information Retrieval, Final Workshop.* 9 pages. https://hal.inria.fr/hal-00953974.

[17] Jean-Pierre Chevallet and Yves Chiaramella. 1998. Experiences in information retrieval modelling using structured formalisms and modal logic. In *Information Retrieval: Uncertainty and Logics*, Fabio Crestani, Mounia Lalmas, and CornelisJoost Rijsbergen (Eds.). Springer US, New York, 39–72. DOI: https://doi.org/10.1007/978-1-4615-5617-6_3

[18] Stéphane Clinchant and Eric Gaussier. 2010. Information-based models for ad hoc IR. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Geneva, Switzerland, 234–241. DOI: https://doi.org/10.1145/1835449.1835490

[19] W. S. Cooper. 1978. *Foundations of Logico-Linguistics: A Unified Theory of Information, Language, and Logic.* Springer, the Netherlands. https://books.google.it/books?id=QZ0tLAwn0yMC.

[20] R. T. Cox. 1946. Probability, frequency and reasonable expectation. *American Journal of Physics* 14, 1 (1946), 1–13. DOI: https://doi.org/10.1119/1.1990764

[21] Fabio Crestani. 1998. Logical imaging and probabilistic information retrieval. In *Information Retrieval: Uncertainty and Logics*, Fabio Crestani, Mounia Lalmas, and CornelisJoost Rijsbergen (Eds.). The Kluwer International Series on Information Retrieval, Vol. 4. Springer US, 247–279. DOI: https://doi.org/10.1007/978-1-4615-5617-6_10

[22] Fabio Crestani and Mounia Lalmas. 2001. Logic and uncertainty in information retrieval. In *Lectures on Information Retrieval*, Maristella Agosti, Fabio Crestani, and Gabriella Pasi (Eds.). Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/3-540-45368-7_9

[23] Fabio Crestani and C. J. Van Rijsbergen. 1995. Information retrieval by logical imaging. *Journal of Documentation* 51 (1995), 3–17.

[24] Fabio Crestani, Ian Ruthven, Marc Sanderson, and C. J. van Rijsbergen. 1995. The troubles with using a logical model of IR on a large collection of documents. In *Proceedings of the Fourth Text Retrieval Conference (TREC-4'95)*, D. K. Harman (Eds.). 509–526.

[25] Fabio Crestani and C. J. van Rijsbergen. 1995. Probability kinematics in information retrieval. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95).* ACM, New York, NY, 291–299. DOI: https://doi.org/10.1145/215206.215373

[26] Mukesh Dalal. 1988. Investigations into a theory of knowledge base revision: Preliminary report. In *Proceedings of the 7th National Conference on Artificial Intelligence*, Paul Rosenbloom and Peter Szolovits (Eds.), Vol. 2. AAAI Press, Menlo Park, CA, 475–479.

[27] Sandor Dominich. 2008. *The Modern Algebra of Information Retrieval.* Springer, Berlin. http://books.google.fr/books?id=uEedNKV3nlUC.

[28] Didier Dubois, Walenty Ostasiewicz, and Henri Prade. 2000. *Fuzzy Sets: History and Basic Notions.* Springer US, Boston, MA, 21–124. DOI: https://doi.org/10.1007/978-1-4615-4429-6_2

[29] Didier Dubois and Henri Prade. 2001. Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of Mathematics and Artificial Intelligence* 32, 1–4 (Aug. 2001), 35–66. DOI: https://doi.org/10.1023/A:1016740830286

[30] Dorothy Edgington. 2014. Indicative conditionals. In *The Stanford Encyclopedia of Philosophy* (Winter 2014 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2014/entries/conditionals/.

[31] Peter Exner and Pierre Nugues. 2012. Entity extraction: From unstructured text to DBpedia RDF triples. In *Proceedings of the Web of Linked Entities Workshop in Conjunction with the 11th International Semantic Web Conference (ISWC'12).* CEUR, 58–69.

[32] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. 2011. Semantically enhanced information retrieval: An ontology-based approach. *Web Semant.* 9, 4 (Dec. 2011), 434–452. DOI: https://doi.org/10.1016/j.websem.2010.11.003

[33] Thomas J. Froehlich. 1994. Relevance reconsidered—Towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *J. Am. Soc. Inf. Sci.* 45, 3 (April 1994), 124–134. DOI: https://doi.org/10.1002/(SICI)1097-4571(199404)45:3⟨124::AID-ASI2⟩3.0.CO;2-8

[34] Norbert Fuhr. 1995. Probabilistic datalog - A logic for powerful retrieval methods. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Seattle, WA,, 282–290. DOI: https://doi.org/10.1145/215206.215372

[35] Norbert Fuhr. 2000. Probabilistic datalog: Implementing logical information retrieval for advanced applications. *JASIS* 51, 2 (2000), 95–110. DOI: https://doi.org/10.1002/(SICI)1097-4571(2000)51:2⟨95::AID-ASI2⟩3.0.CO;2-H

[36] Norbert Fuhr and Thomas Rölleke. 1998. HySpirit — A probabilistic inference engine for hypermedia retrieval in large databases. In *Advances in database technology — EDBT'98*, Hans-Jörg Schek, Gustavo Alonso, Felix Saltor, and Isidro Ramos (Eds.). Springer, Berlin, 24–38.

[37] Peter Gardenfors. 1982. Imaging and conditionalization. *The Journal of Philosophy* 79, 12 (1982), 747–760. DOI: https://doi.org/10.2307/2026039

[38] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. ACM, New York, NY, 267–274. DOI: https://doi.org/10.1145/1571941.1571989

[39] Yogesh Gupta, Ashish Saini, and Ak Saxena. 2014. Fuzzy logic-based approach to develop hybrid similarity measure for efficient information retrieval. *J. Inf. Sci.* 40, 6 (Dec. 2014), 846–857. DOI: https://doi.org/10.1177/0165551514548989

[40] R. Haenni, J. Kohlas, and N. Lehmann. 2000. *Probabilistic Argumentation Systems*. Springer Netherlands, Dordrecht, 221–288. DOI: https://doi.org/10.1007/978-94-017-1737-3_6

[41] Theodore Hailperin. 1984. Probability logic. *Notre Dame J. Formal Logic* 25, 3 (07 1984), 198–212. DOI: https://doi.org/10.1305/ndjfl/1093870625

[42] Alan Hájek. 2001. Probability, logic, and probability logic. *The Blackwell Guide to Philosophical Logic*, Blackwell (2001), 362–384.

[43] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. DBpedia-entity V2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. ACM, New York, NY, 1265–1268. DOI: https://doi.org/10.1145/3077136.3080751

[44] G. E. Hughes and J. Cresswell. 1996. *A New Introduction to Modal Logic*. Routledge. https://books.google.fr/books?id=Dsn1xWNB4MEC.

[45] T. W. C. Huibers and P. D. Bruza. 1994. *Situations: A General Framework for Studying Information Retrieval*. Utrecht University, Department of Computer Science. http://books.google.fr/books?id=YtcuGwAACAAJ.

[46] Anthony Hunter. 1995. Using default logic in information retrieval. In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Christine Froidevaux and Jürg Kohlas (Eds.). Springer Berlin Heidelberg, Springer, Berlin, 235–242. DOI: https://doi.org/10.1007/3-540-60112-0_27

[47] Kevin H. Knuth. 2004. Deriving laws from ordering relations. *AIP Conference Proceedings* 707, 1 (2004), 204–235. DOI: https://doi.org/10.1063/1.1751368

[48] Kevin H. Knuth. 2005. Lattice duality: The origin of probability and entropy. *Neurocomput.* 67 (Aug. 2005), 245–274. DOI: https://doi.org/10.1016/j.neucom.2004.11.039

[49] Daphne Koller, Alon Levy, and Avi Pfeffer. 1997. P-CLASSIC: A tractable probabilistic description logic. In *Proceedings of AAAI'97*. AAAI Press, Menlo Park, CA, Providence, RI, 390–397.

[50] S. A. Kripke. 1963. Semantic analysis of modal logic I: Normal modal and propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 9 (1963), 67–96.

[51] Mounia Lalmas. 1998. Logical models in information retrieval: Introduction and overview. In *Information Processing & Management*. 34, 1 (1998), 19–33. DOI: https://doi.org/10.1016/S0306-4573(97)00041-1

[52] Mounia Lalmas and Peter D. Bruza. 1998. The use of logic in information retrieval modelling. *The Knowledge Engineering Review* 13, 3 (10 1998), 263–295. DOI: https://doi.org/10.1017/S0269888998002124

[53] Mounia Lalmas and Keith Rijsbergen. 1993. A logical model of information retrieval based on situation theory. In *14th Information Retrieval Colloquium*, Tony McEnery and Chris Paice (Eds.). Springer London, 1–13. DOI: https://doi.org/10.1007/978-1-4471-3211-0_1

[54] Raymond Y. K. Lau, Peter D. Bruza, and Dawei Song. 2008. Towards a belief-revision-based adaptive and context-sensitive information retrieval system. *ACM Trans. Inf. Syst.* 26, 2, Article 8 (April 2008), 38 pages. DOI: https://doi.org/10.1145/1344411.1344414

[55] David Lewis. 1976. Probabilities of conditionals and conditional probabilities. *The Philosophical Review* 85, 3 (1976), 297–315. http://www.jstor.org/stable/2184045.

[56] David K. Lewis. 1973. *Counterfactuals*. Harvard University Press, Cambridge, MA. http://www.worldcat.org/oclc/795075.

[57] Christina Lioma, Roi Blanco, Raquel Mochales Palau, and Marie-Francine Moens. 2009. A belief model of query difficulty that uses subjective logic. In *Advances in Information Retrieval Theory*, Leif Azzopardi, Gabriella Kazai, Stephen Robertson, Stefan Rüger, Milad Shokouhi, Dawei Song, and Emine Yilmaz (Eds.). Springer, Berlin, 92–103.

[58] Christina Lioma, Birger Larsen, Hinrich Schuetze, and Peter Ingwersen. 2010. A subjective logic formalisation of the principle of polyrepresentation for information needs. In *Proceedings of the 3rd Symposium on Information Interaction in Context (IIiX'10)*. ACM, New York, NY, 125–134. DOI: https://doi.org/10.1145/1840784.1840804

[59] David E. Losada and Alvaro Barreiro. 2001. A logical model for information retrieval based on propositional logic and belief revision. *Comput. J.* 44, 5 (2001), 410–424.

[60] David E. Losada and Alvaro Barreiro. 2003. Propositional logic representations for documents and queries: A large-scale evaluation. In *Proceedings of the 25th European Conference on IR Research (ECIR'03)*. Springer, Berlin, 219–234. http://dl.acm.org/citation.cfm?id=1757788.1757810.

[61] Thomas Lukasiewicz. 2008. Expressive probabilistic description logics. *Artif. Intell.* 172, 6-7 (April 2008), 852–883. DOI: https://doi.org/10.1016/j.artint.2007.10.017

[62] Loic Maisonnasse, Eric Gaussier, and Jean-Pierre Chevallet. 2009. Model fusion in conceptual language modeling. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval (ECIR'09)*. Springer, Berlin, 240–251. DOI: https://doi.org/10.1007/978-3-642-00958-7_23

[63] Carlo Meghini, Fabrizio Sebastiani, Umberto Straccia, and Costantino Thanos. 1993. A model of information retrieval based on a terminological logic. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Pittsburgh, PA, 298–307. DOI: https://doi.org/10.1145/160688.160753

[64] Carlo Meghini and Umberto Straccia. 1996. A relevance terminological logic for information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96).* ACM, New York, NY, 197–205. DOI: https://doi.org/10.1145/243199.243267

[65] Stefano Mizzaro. 1997. Relevance: The whole history. *Journal of the American Society for Information Science* 48, 9 (1997), 810–832. DOI: https://doi.org/10.1002/(SICI)1097-4571(199709)48:9⟨810::AID-ASI6⟩3.0.CO;2-U

[66] B. Nebel. 1992. Syntax-based approaches to belief revision. In *Belief Revision.* Cambridge University Press, New York, NY, 52–88.

[67] J. Nie. 1988. An outline of a general model for information retrieval systems. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Grenoble, France, 495–506. DOI: https://doi.org/10.1145/62437.62493

[68] Jianyun Nie. 1989. An information retrieval model based on modal logic. *Information Processing & Management* 25, 5 (1989), 477–491. DOI: https://doi.org/10.1016/0306-4573(89)90019-8

[69] Jian-Yun Nie. 1992. Towards a probabilistic modal logic for semantic-based information retrieval. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Copenhagen, Denmark, 140–151. DOI: https://doi.org/10.1145/133160.133188

[70] Jian-Yun Nie and Martin Brisebois. 1996. An inferential approach to information retrieval and its implementation using a manual thesaurus. *Artif. Intell. Rev.* 10, 5–6 (Oct. 1996), 409–439. DOI: https://doi.org/10.1007/BF00130693

[71] Jian-Yun Nie and Francois Lepage. 1998. Toward a broader logical model for information retrieval. In *Information Retrieval: Uncertainty and Logics*, Fabio Crestani, Mounia Lalmas, and CornelisJoost van Rijsbergen (Eds.). The Kluwer International Series on Information Retrieval, Vol. 4. Springer US, 17–38. DOI: https://doi.org/10.1007/978-1-4615-5617-6_2

[72] Gabriella Pasi. 1999. A logical formulation of the Boolean model and of weighted Boolean models. In *Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems, LUMIS, at ECSQARU'99.* Éditions Universitaires d'Avignon, 1–11. http://dblp.uni-trier.de/db/conf/coria/coria2011.html#AbdulahhadCB11.

[73] Paolo Penna. 2000. Succinct representations of model based belief revision. In *STACS 2000*, Horst Reichel and Sophie Tison (Eds.). *Lecture Notes in Computer Science*, Vol. 1770. Springer, Berlin, 205–216. DOI: https://doi.org/10.1007/3-540-46541-3_17

[74] Justin Picard and Jacques Savoy. 2000. A logical information retrieval model based on a combination of propositional logic and probability theory. In *Soft Computing in Information Retrieval*, Fabio Crestani and Gabriella Pasi (Eds.). Physica-Verlag HD, Heidelberg. DOI: https://doi.org/10.1007/978-3-7908-1849-9_10

[75] Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Melbourne, Australia, 275–281. DOI: https://doi.org/10.1145/290941.291008

[76] Guilin Qi and Jeff Z. Pan. 2008. A tableau algorithm for possibilistic description logic $\mathcal{ALC}$. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web (ASWC'08).* Springer, Berlin, 61–75. DOI: https://doi.org/10.1007/978-3-540-89704-0_5

[77] Saïd Radhouani, Gilles Falquet, and Jean-Pierre Chevalletinst. 2008. Description logic to model a domain specific information retrieval system. In *Database and Expert Systems Applications*, Sourav S. Bhowmick, Josef Küng, and Roland Wagner (Eds.). Springer, Berlin, 142–149.

[78] S. E. Robertson and K. S. Jones. 1976. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* 27, 3 (1976), 129–146. DOI: https://doi.org/10.1002/asi.4630270302

[79] S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Springer, New York, NY, Dublin, Ireland, 232–241. http://dl.acm.org/citation.cfm?id=188490.188561

[80] Thomas Rölleke and Norbert Fuhr. 1996. Retrieval of complex objects using a four-valued logic. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Zurich, Switzerland, 206–214. DOI: https://doi.org/10.1145/243199.243268

[81] Thomas Rölleke, Hengzhi Wu, Jun Wang, and Hany Azzam. 2008. Modelling retrieval models in a probabilistic relational algebra with a new operator: The relational Bayes. *VLDB J.* 17, 1 (2008), 5–37. DOI: https://doi.org/10.1007/s00778-007-0073-y

[82] Neil Rubens. 2006. The application of fuzzy logic to the construction of the ranking function of information retrieval systems. *CoRR* abs/cs/0610039 (2006). arxiv:cs/0610039 http://arxiv.org/abs/cs/0610039.

[83] G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (Nov. 1975), 613–620. DOI: https://doi.org/10.1145/361219.361220

[84] Christoph Schwering. 2017. A reasoning system for a first-order logic of limited belief. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17).* AAAI Press, 1247–1253. DOI: https://doi.org/10.24963/ijcai.2017/173

[85] Fabrizio Sebastiani. 1994. A probabilistic terminological logic for modelling information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Springer, New York, NY, Dublin, Ireland, 122–130. http://portal.acm.org/citation.cfm?id=188490.188544.

[86] Fabrizio Sebastiani. 1998. On the role of logic in information retrieval. *Information Processing & Management* 34, 1 (1998), 1–18.

[87] Fabrizio Sebastiani. 1999. Towards a logical reconstruction of information retrieval theory. *Cybernet. Syst* 30 (1999), 411–428.

[88] Fabrizio Sebastiani and Umberto Straccia. 1991. A computationally tractable terminological logic. In *SCAI.* 307–315.

[89] Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, Zurich, Switzerland, 21–29. DOI: https://doi.org/10.1145/243199.243206

[90] John F. Sowa. 2010. *The Role of Logic and Ontology in Language and Reasoning.* Springer Netherlands, Dordrecht, 231–263. DOI: https://doi.org/10.1007/978-90-481-8845-1_11

[91] Laurent Ughetto, Gabriella Pasi, Vincent Claveau, Olivier Pivert, and Patrick Bosc. 2010. Implication in information retrieval systems. In *Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO'10).* Le Centre de Hautes Etudes Internationales d'Informatique Documentaire, Paris, France, 61–64. http://dl.acm.org/citation.cfm?id=1937055.1937068.

[92] David Vallet, Miriam Fernández, and Pablo Castells. 2005. An ontology-based information retrieval model. In *The Semantic Web: Research and Applications*, Asunción Gómez-Párez and Jéróme Euzenat (Eds.). Lecture Notes in Computer Science, Vol. 3532. Springer, Berlin, 455–470. DOI: https://doi.org/10.1007/11431053_31

[93] C. J. van Rijsbergen. 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation* 33 (June 1977), 106–119. http://libra.msra.cn/paperdetail.aspx?id=1292204.

[94] C. J. van Rijsbergen. 1986. A non-classical logic for information retrieval. *Comput. J.* 29, 6 (1986), 481–485. http://dblp.uni-trier.de/db/journals/cj/cj29.html#Rijsbergen86.

[95] Yunjie Xu and Hainan Yin. 2008. Novelty and topicality in interactive information retrieval. *J. Am. Soc. Inf. Sci. Technol.* 59, 2 (Jan. 2008), 201–215. DOI: https://doi.org/10.1002/asi.v59:2

[96] Yunjie (Calvin) Xu and Zhiwei Chen. 2006. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.* 57, 7 (May 2006), 961–973. DOI: https://doi.org/10.1002/asi.v57:7

[97] L. A. Zadeh. 1965. Fuzzy sets. *Information and Control* 8, 3 (1965), 338–353. DOI: https://doi.org/10.1016/S0019-9958(65)90241-X

[98] Loutfi Zerarga and Yassine Djouadi. 2018. A many-sorted theory proposal for information retrieval: Axiomatization and semantics. *Knowledge and Information Systems* 55, 1 (01 Apr 2018), 113–139. DOI: https://doi.org/10.1007/s10115-017-1074-9

[99] Guido Zuccon, Leif Azzopardi, and Cornelis J. van Rijsbergen. 2009. Revisiting logical imaging for information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09).* ACM, New York, NY, 766–767. DOI: https://doi.org/10.1145/1571941.1572118

[100] G. Zuccon, L. A. Azzopardi, and C. J. van Rijsbergen. 2008. A formalization of logical imaging for information retrieval using quantum theory. In *Proceedings of the19th International Conference on Database and Expert Systems Application (DEXA'08).* IEEE Computer Society, Washington, DC, 3–8. DOI: https://doi.org/10.1109/DEXA.2008.69

Received September 2017; revised August 2018; accepted November 2018