Handling Topical Metadata Regarding the Validity and Completenes of Multiple-Source Information: A possibilistic approach

Célia Da Costa Pereira, Didier Dubois, Henri Prade, Andrea G. B. Tettamanzi

Handling Topical Metadata Regarding the Validity and Completenes of Multiple-Source Information: A possibilistic approach

Célia da Costa Pereira[1], Didier Dubois[2], Henri Prade[2], and Andrea G. B. Tettamanzi[1]

[1] Université Côte d'Azur, CNRS, I3S, France
{celia.pereira,andrea.tettamanzi}@unice.fr
[2] IRIT – CNRS, 118, route de Narbonne, Toulouse, France
dubois@irit.fr, prade@irit.fr

Abstract. We study the problem of aggregating metadata about the validity and/or completeness, with respect to given topics, of information provided by multiple sources. For a given topic, the validity level reflects the certainty that the information stored is true. The completeness level of a source on a given topic reflects the certainty that a piece of information that is not stored is false. We propose a modeling based on possibility theory which allows the fusion of such multi-source information into a graded belief base.

1 Introduction and Related Work

The relation between *beliefs* and *knowledge* plays a central role in epistemology. Much of epistemology revolves around questions about when and how our beliefs are justified or qualify as knowledge [19]. Without taking a position in this debate, in this paper, we will use the term *knowledge* when referring to information provided by an information source, but we will use the term *beliefs* to refer to a (possibly partial, incomplete, or uncertain) representation of reality obtained by combining information provided by one or more sources with metadata about its validity and completeness.

The problem of representing validity and completeness of information stored in databases has started drawing attention many years ago. For example, we can consider the model of database integrity proposed by Motro [17] and the work by Demolombe [8] who used modal logic for representing information stored in relational databases. Our aim is to consider validity and completeness in more general knowledge bases (KBs) in which the closed world assumption is not made. Therefore, a mechanism for representing uncertainty in the beliefs induced by KBs fed by sources which can provide invalid and/or incomplete pieces of information is needed.

Cholvy [6] uses the theory of evidence for proposing an interesting way to compute the extent to which an agent should believe a new piece of information provided by an imperfect information source. A difference with respect to our

work is that we explicitly associate these metadata concerning validity and completeness to topics and this allows us to describe these metadata for a source at a finer grain.

Bacchus *et al.* [3] proposed the "random worlds" method, an approach for inducing degrees of belief from KBs fed with different types of information like statistical correlations, physical laws, default rules, etc. They apply the principle of indifference and, therefore, all the possible worlds derived from the agent's KB are equally probable. The uncertainty about information is directly represented in the KB (as statistical information, defeasible information and so on), not as metadata.

We consider a possibilistic representation of beliefs to take uncertainty into account. We assume the beliefs of an agent come from various information sources, which may be more or less reliable (this has to do with information *validity*) and more or less exhaustive (this has to do with *completeness*). The validity level reflects the certainty that the information an agent stores on a given topic is true, while the completeness level reflects the certainty that, on a given topic, a missing piece of information is false.

The goal of our model is to support inferences, thus to answer queries, by providing a weighted summary of the different (and possibly conflicting) opinions of the available sources. An important point in our framework is that we provide the user (requestor) with the different answers that can be obtained from the information system in case of conflict. The user must then be aware of which sources give which answer to his/her query, and with which certainty degree.

We adapt and extend the formalism by Dubois and Prade [12] for completeness and validity of databases, to reason about the beliefs (opinions) of a source. We give a possibilistic reasoning algorithm for those beliefs, whose complexity is in the same class as reasoning on a crisp KB and less expensive than reasoning on a general possibilistic belief base. Furthermore, we combine this solution with a multi-source generalization of possibilistic logic [9] to summarize and reason about the different (and possibly conflicting) beliefs of the sources.

The paper is organized as follows: the next section states the problem we study; Section 3 gives then some background about the formal tools we use. Section 4 explains how a gradual set of beliefs can be constructed from validity and completeness metadata. Section 5 exploits multi-source possibilistic logic to merge beliefs from multiple sources. Section 7 concludes the paper.

2 Problem Statement

The problem we study can be schematically depicted as in Figure 1. We are given n KBs $K_1, \ldots, K_n$, fed by n imperfect, independent information sources $s_1, \ldots, s_n$, about which two kinds of metadata are known: for each topic, on the one hand, we know to what degree a source s_i provides *valid* information. On the other hand, we know to what degree information provided by source s_i is *complete*. Here, we use the term *knowledge base* to mean a (possibly noisy and

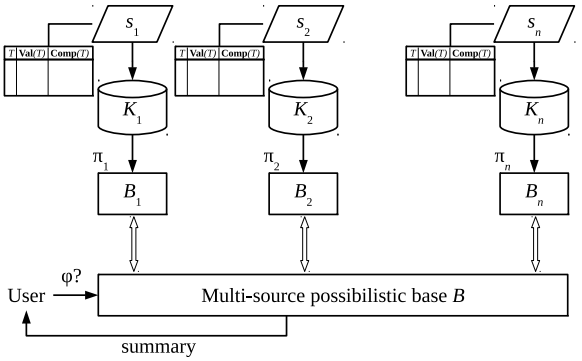

Fig. 1. A schematic illustration of an abstract information system, consisting of n knowledge bases K_i fed by n independent information sources s_i, with metadata about their validity and completeness, whence a possibility distribution π_i and a corresponding possibilistic belief base B_i are constructed and used to answer queries.

incomplete) collection of *facts*, for which the open world assumption (OWA) holds. We answer two important questions:

1. How can the facts contained in each KB K_i be combined with metadata about the validity and completeness of its source s_i to construct a gradual belief base B_i taking the uncertainty of K_i (due to its imperfection) into account?
2. How can the n belief bases be used to answer queries while merging (possibly conflicting) information coming from the n sources?

The former is a problem of metadata aggregation, whereas the latter is a problem of information (or, more properly, belief) fusion. We argue that possibility theory provides suitable tools to solve both problems.

To illustrate our proposal, we will use simple examples inspired, much like in [8], by an air travel planning application. One might suppose that some of the KBs in the system would be fed with flight information directly by an airline and some by an airport, each source being more authoritative about information which falls directly under its control and less for other information.

3 Background

3.1 Knowledge Representation Language

For the sake of simplicity, we base our treatment on a decidable fragment of pure (i.e., without function symbols and identity) first-order predicate logic, namely the Schönfinkel-Bernays class of first-order formulas [5].

Definition 1 (Language $\mathcal{L}$). *Let $\mathcal{L}_{QF}$ be the set of quantifier-free formulas inductively defined as follows:*

- *a term is either a variable or a constant (including literals denoting numbers, times, character strings, etc.);*
- *if P is an n-ary predicate symbol and $t_1, \ldots, t_n$ are terms, then $P(t_1, \ldots, t_n)$ is an (atomic) formula and $P(t_1, \ldots, t_n) \in \mathcal{L}_{QF}$;*
- $\bot, \top \in \mathcal{L}_{QF}$
- *if $\phi \in \mathcal{L}_{QF}$, then $\neg\phi \in \mathcal{L}_{QF}$;*
- *if $\phi, \psi \in \mathcal{L}_{QF}$ then $\phi \wedge \psi \in \mathcal{L}_{QF}$ and $\phi \vee \psi \in \mathcal{L}_{QF}$.*

$\mathcal{L}$ is the smallest language such that $\mathcal{L}_{QF} \subseteq \mathcal{L}$ and, if $\phi \in \mathcal{L}_{QF}$ and $x_1, \ldots, x_m$, $y_1, \ldots, y_n$ are variables, then

$$\exists x_1 \ldots \exists x_m \forall y_1 \ldots \forall y_n \phi \in \mathcal{L}.$$

A variable x is free *in formula ϕ if it is not quantified; otherwise it is* bound. *A formula without free variables is* closed. *A formula with free variables is* open. *A formula not containing variables is* ground.

The semantics of $\mathcal{L}$ can be defined as follows:

Definition 2. *The Herbrand base of $\mathcal{L}$ is the set $H_{\mathcal{L}}$ of all ground atoms in $\mathcal{L}$. An interpretation (or model) is a function $\mathcal{I} : H_{\mathcal{L}} \to \{0, 1\}$, which can also be viewed as a subset of the Herbrand base, $\mathcal{I} \subseteq H_{\mathcal{L}}$ (the set of all atoms ϕ such that $\phi^{\mathcal{I}} = 1$). We denote $\Omega = 2^{H_{\mathcal{L}}}$ the set of all interpretations.*

We observe that $H_{\mathcal{L}}$ is finite, because there are no function symbols in $\mathcal{L}$.

Definition 3 (Satisfaction). *Let P be an n-ary predicate, $\phi, \psi \in \mathcal{L}$ closed formulas and $\mathcal{I}$ an interpretation of $H_{\mathcal{L}}$:*

- $\models_{\mathcal{I}} \top$ *and* $\not\models_{\mathcal{I}} \bot$;
- $\models_{\mathcal{I}} P(t_1, \ldots, t_n)$ *if and only if* $P(t_1, \ldots, t_n) \in \mathcal{I}$;
- $\models_{\mathcal{I}} \neg\phi$ *if and only if* $\not\models_{\mathcal{I}} \phi$;
- $\models_{\mathcal{I}} \phi \wedge \psi$ *if and only if* $\models_{\mathcal{I}} \phi$ *and* $\models_{\mathcal{I}} \psi$;
- $\models_{\mathcal{I}} \phi \vee \psi$ *if and only if* $\models_{\mathcal{I}} \phi$ *or* $\models_{\mathcal{I}} \psi$;
- $\models_{\mathcal{I}} \forall x \phi(x)$ *if and only if* $\models_{\mathcal{I}} \phi(c)$ *for all constant c;*
- $\models_{\mathcal{I}} \exists x \phi(x)$ *if and only if* $\models_{\mathcal{I}} \phi(c)$ *for some constant c.*

An open formula $\phi(x_1, \ldots, x_n)$ is satisfied by $\mathcal{I}$ iff $\models_{\mathcal{I}} \forall x_1 \ldots \forall x_n \phi(x_1, \ldots, x_n)$.

It is a well-known result that the satisfiability of the formulas of $\mathcal{L}$, besides being decidable, is in the NEXPTIME-complete complexity class [15].

We impose the restriction that only ground formulas of $\mathcal{L}_{QF}$ without negation and disjunction (which we shall call *facts*) can be stored in a KB (one does not usually say, when stating facts, things like "Tom is not from NY" or "Tom is from NY or LA"). We denote such restricted language $\mathcal{L}_{\text{fact}}$.

Notice that, by Definition 2, the three languages $\mathcal{L}_{\text{fact}} \subset \mathcal{L}_{QF} \subset \mathcal{L}$ share the same identical Herbrand base $H_{\mathcal{L}}$.

Definition 4. *Let $\phi, \psi \in \mathcal{L}$: $\phi \models \psi$ if and only if, for all $\mathcal{I} \subseteq H_{\mathcal{L}}$, if $\models_{\mathcal{I}} \phi$, then also $\models_{\mathcal{I}} \psi$.*

$\mathcal{L}$ may be viewed as an abstraction of popular ways to encode information used in state-of-the-art technologies, such as relational databases, datalog, and RDF + OWL.

Example 1. The set of facts $S \subset \mathcal{L}_{\text{fact}}$,

$$S = \{ \quad \text{Flight}(\mathsf{AF1680}), \text{Origin}(\mathsf{AF1680}, \mathsf{CDG}), \text{Dest}(\mathsf{AF1680}, \mathsf{LHR}), \\ \text{Depart}(\mathsf{AF1680}, \mathsf{07{:}25}), \text{Arrival}(\mathsf{AF1680}, \mathsf{07{:}45}), \text{Airline}(\mathsf{AF1680}, \mathsf{AF}) \ \}$$

describes a morning flight connecting Paris Charles de Gaulle to London Heathrow. Formula $\phi = \exists x(\text{Flight}(x) \wedge \text{Origin}(x, \mathsf{CDG}) \wedge \text{Dest}(x, \mathsf{LHR}))$ states that there is a flight connecting those two airports.

3.2 Possibility Theory and Possibilistic Logic

Possibility theory [11] is a mathematical theory of uncertainty that relies upon fuzzy set theory [20], in that the (fuzzy) set of possible values for a variable of interest is used to describe the uncertainty as to its precise value. At the semantic level, the membership function of such set, π, is called a *possibility distribution* and its range is $[0, 1]$. By convention, $\pi(\mathcal{I}) = 1$ means that it is totally possible for $\mathcal{I}$ to be the real world, $0 < \pi(\mathcal{I}) < 1$ means that $\mathcal{I}$ is only somehow possible, while $\pi(\mathcal{I}) = 0$ means that $\mathcal{I}$ is ruled out. A possibility distribution π is said to be normalized if there exists at least one interpretation $\mathcal{I}_0$ such that $\pi(\mathcal{I}_0) = 1$.

Definition 5. *(Possibility and Necessity Measures) A possibility distribution π induces a* possibility measure *and its dual* necessity measure*, denoted by Π and N respectively. Both measures apply to a classical set of interpretation $S \subseteq \Omega$ and are defined as follows:*

$$\Pi(S) = \max_{\mathcal{I} \in S} \pi(\mathcal{I}); \tag{1}$$

$$N(S) = 1 - \Pi(\bar{S}) = \min_{\mathcal{I} \in \bar{S}} \{1 - \pi(\mathcal{I})\}. \tag{2}$$

In words, $\Pi(S)$ expresses to what extent S is consistent with the available knowledge. Conversely, $N(S)$ expresses to what extent S is entailed by the available knowledge. Among the properties of Π and N induced by a normalized possibility distribution on a finite universe of discourse Ω, we can mention, for all subsets $S \subseteq \Omega$:

1. $\Pi(S) = 1 - N(\bar{S})$ (duality);
2. $N(S) > 0 \Rightarrow \Pi(S) = 1; \quad \Pi(S) < 1 \Rightarrow N(S) = 0.$

Possibilistic logic [10] has been originally motivated by the need to manipulate syntactic expressions of the form (ϕ, α), where ϕ is a classical logic formula, and α is a certainty level, with the intended semantics that $N(\phi) \geq \alpha$, where N is a necessity measure.

A possibilistic belief base B is a set $\{(\phi_i, \alpha_i)\}_{i=1,\ldots,m}$ of possibilistic logic formulas. Clearly, B can be layered into a set of nested classical bases $B_\alpha = \{\phi_i \mid$

$(\phi_i, \alpha_i) \in B$ and $\alpha_i \geq \alpha\}$ such that $B_\alpha \subseteq B_\beta$ if $\alpha \geq \beta$. Proving syntactically $B \vdash (\phi, \alpha)$ amounts to proceeding by refutation and proving $B \cup \{(\neg\phi, 1)\} \vdash (\bot, \alpha)$ by repeated application of the resolution rule $(\neg\phi \vee \psi, \alpha), (\phi \vee \nu, \beta) \vdash (\psi \vee \nu, \min(\alpha, \beta))$. Moreover, $B \vdash (\phi, \alpha)$ if and only if $B_\alpha \vdash \phi$ and $\alpha > inc(B)$, where $inc(B)$ is the inconsistency level of B defined as $inc(B) = \max\{\alpha \mid B \vdash (\bot, \alpha)\}$. It can be shown that $inc(B) = 0$ iff B_0 is consistent, with $B_0 = \{\phi_i \mid (\phi_i, \alpha_i) \in B\}$. Thus reasoning from a possibilistic base just amounts to reasoning classically with subparts of the base whose formulas are strictly above the certainty level.

A possibilistic belief base $B = \{(\phi_i, \alpha_i)\}_{i=1,\ldots,m}$ encodes the constraints $N(\phi_i) \geq \alpha_i$. B is thus semantically associated with a possibility distribution [10]

$$\pi_B(\mathcal{I}) = \min_{i=1,\ldots,m} \max(\phi_i^{\mathcal{I}}, 1 - \alpha_i),$$

where $\phi_i^{\mathcal{I}} = 1$ if $\mathcal{I}$ is a model of ϕ_i, and $\phi_i^{\mathcal{I}} = 0$ otherwise; π_B is the largest possibility distribution, i.e., the least specific distribution assigning the largest possibility levels in agreement with the constraints $N(\phi_i) \geq \alpha_i$ for $i = 1, \ldots, m$. The distribution π_B rank-orders the interpretations $\mathcal{I}$ of the language induced by the ϕ_i's according to their plausibility on the basis of the strength of the pieces of information in B. If the set of formulas B_0 is consistent, then the distribution π_B is normalized. The semantic entailment is defined by $B \models (\phi, \alpha)$ iff $\forall \mathcal{I}, \pi_B(\mathcal{I}) \leq \pi_{\{(\phi,\alpha)\}}(\mathcal{I})$. Reasoning by refutation in propositional possibilistic logic is sound and complete, applying the syntactic resolution rule. Namely, it can be shown that $B \models (\phi, \alpha)$ iff $B \vdash (\phi, \alpha)$ and $inc(B) = 1 - \max_{\mathcal{I}} \ \pi_B(\mathcal{I})$.

Algorithms for reasoning in possibilistic logic and an analysis of their complexity, which is similar to the one of classical logic, multiplied by the logarithm of the number of levels used in the necessity scale, can be found in [14].

4 Representing and Reasoning with Validity and Completeness

When dealing with relational databases, only the statements explicitly present in the database are considered as true (valid). The others are considered as false—the closed world assumption (CWA). When dealing with more general *knowledge* bases, i.e., sets of logical formulas, from which other formulas can be deduced, the true statements are those explicitly represented in the KB, plus those which can be inferred thanks to a reasoner. However, due to the OWA, we cannot suppose that statements that cannot be inferred are false—the truth status of some statements may be unknown in case of incomplete knowledge. In fact, insofar as for any formula ϕ we have a tool to decide if ϕ can be inferred and if $\neg\phi$ can be inferred, CWA makes no sense since when neither ϕ nor $\neg\phi$ can be inferred, CWA would lead to a contradiction, unless we put syntatic restrictions on ϕ, e.g., $\neg\phi$ cannot be expressed in the language.

In this section, we recall the notions of validity and completeness for dealing with relational databases [8, 12] and adapt them to the more general case of a

KB. We treat validity and completeness of information at the fine grain of a *topic*, defined as follows.

Definition 6. *(Topic) Given a formula $\phi \in \mathcal{L}_{QF}$ without negation, the* topic *$T(\phi)$ is the set of all the ground formulas that can be obtained by substituting all the free variables in ϕ with all possible constants.*

Example 2. The topic of "all flights departing from Heathrow" may be described by the open formula $\text{Origin}(x, \mathsf{LHR})$.

Let $\mathcal{T}$ be the set of topics and let K be a KB of formulas in $\mathcal{L}_{\text{fact}}$. In practice, K is a conjunction of ground atoms in $H_{\mathcal{L}}$.

Unlike for databases, in the case of a general KB, the OWA holds and logical inferences can be performed. Therefore, we must think in terms of logical entailment of formulas.

Example 3. Assume the following KB is given:
$K = \{\text{Flight}(\mathsf{AF1680}), \text{Origin}(\mathsf{AF1680}, \mathsf{CDG}), \text{Dest}(\mathsf{AF1680}, \mathsf{LHR}), \text{Airline}(\mathsf{AF1680}, \mathsf{AF})\}$ then $K \models \exists x(\text{Flight}(x) \wedge \text{Airline}(x, \mathsf{AF}))$ (there is a flight operated by AF), but $K \not\models \exists x \forall y(\neg\text{Flight}(y) \vee \text{Airline}(y, x))$ (all flights are operated by one airline), because one cannot logically rule out other facts not contained in K (K is not complete), such as, for instance, $\text{Flight}(\mathsf{BA303})$ and $\text{Airline}(\mathsf{BA303}, \mathsf{BA})$.

In absolute terms, the notions of *validity* and *completeness* of a KB K with respect to a topic may be defined as follows:

- K is *valid* with respect to a topic iff, for every formula ϕ in that topic, $K \models \phi$ implies that ϕ is indeed true;
- K is *complete* with respect to a topic iff, for every formula ψ in that topic, $K \not\models \psi$ implies that ψ is false.

A formula may be believed to different degrees. We suppose that these degrees depend on both the degree of completeness of the set of facts contained in K and on their validity, which depends on the reliability (or trustworthiness) or even safety [7] of their information source. For example, information related to an Air France flight should be complete if the source is the Air France carrier itself. However, the completeness could be lesser if the source is a private travel agency with a partial coverage about the current flights from the different companies including those of Air France. Similarly, the degree of trust to be associated with information fed by a clerk should be less than the one to be associated with information fed by a supervisor. Still, we would like to emphasize that the way in which such degrees are obtained is out of the scope of this paper. A good source in the literature about trust can be, for example, [16], for a computational view of trust.

We assume that metadata about validity and completeness of information stored in K is given in the form of two functions, **Val** and **Comp**, which associate a degree of validity and completenes, respectively, to each topic.

Definition 7. *Let* $\mathbf{Val} : \mathcal{T} \to [0,1]$ *be such that, for all* $T \in \mathcal{T}$, $\mathbf{Val}(T)$ *is the degree to which* K *contains valid information about topic* T*, which means, for all formulas* ϕ *such that* $K \models \phi$ *and* $\phi \in T$, $N(\phi) \geq \mathbf{Val}(T)$.

Definition 8. *Let* $\mathbf{Comp} : \mathcal{T} \to [0,1]$ *be such that, for all* $T \in \mathcal{T}$, $\mathbf{Comp}(T)$ *is the degree to which* K *contains complete information about topic* T*, which means, for all formulas* ψ *such that* $K \not\models \psi$ *and* $\psi \in T$, $\Pi(\psi) \leq 1 - \mathbf{Comp}(T)$.

In practice, the **Val** and **Comp** functions may be implemented efficiently by a hash table having the formulas representing the topics as keys; a missing key would imply a degree of zero. Now, K plus the metadata provided by **Val** and **Comp** allow us to compute the degree of possibility and necessity for any arbitrary formulas ϕ and ψ as follows:

$$\Pi^-(\phi) = \min_{T:\phi\in T} 1 - \mathbf{Comp}(T), \qquad \text{if } K \not\models \phi; \tag{3}$$

$$N^+(\psi) = \max_{T:\psi\in T} \mathbf{Val}(T), \qquad \text{if } K \models \psi. \tag{4}$$

Notice that Π^- and N^+ are associated to two distinct possibility distributions π^+ (the least specific distribution induced by the necessity measure of Equation 4) and π^- (the least specific distribution induced by the possibility measure of Equation 3). We now show that if K is consistent, intersecting π^+ and π^- yields a normalized possibility distribution π, for all models $\mathcal{I}$, of the form $\pi(\mathcal{I}) = \min\{\pi^+(\mathcal{I}), \pi^-(\mathcal{I})\}$, such that there is a single model $\mathcal{I}^*$ with $\pi(\mathcal{I}^*) = 1$. We recall that normalization is the equivalent, within possibilistic logic, of consistency in crisp logic.

Let B a hypothetical possibilistic belief base corresponding to it. We now prove that such a possibility distribution exists and is normalized.

Let $H_{\mathcal{L}}$ be the Herbrand base constructed over $\mathcal{L}$ and $\Omega = 2^{H_{\mathcal{L}}}$ be the set of all interpretations. A possibilistic data base K^+ will be a collection of pairs (g_i, ν_i) made of ground atoms $g_i \in H_K \subset H_{\mathcal{L}}$, and necessity levels obtained from validity degrees as per Equation 4). The uncertain completeness assumptions comes down to the assumption of another (virtual) data base K^- containing a collection of pairs $(\neg g_j, \kappa_j)$ made of all ground atoms $g_j \in H$ that do not appear in K^+, and necessity levels obtained from completeness degrees as per Equation 3).

Theorem 1. *There exists a normalized possibility distribution* $\pi : \Omega \to [0,1]$ *of the form* $\pi(\mathcal{I}) = \min\{\pi^+(\mathcal{I}), \pi^-(\mathcal{I})\}$*, such that there is a single model* $\mathcal{I}^*$ *with* $\pi(\mathcal{I}^*) = 1$*, inducing the possibility and necessity measures of Equations 3 and 4.*

Proof. As K^+ contains only positive ground atoms $g_i \in H_K$, it is consistent. So the possibility distribution π^+ induced by K^+ is normalised. Let K^+_α be a cut of K^+. Its set of models is rectangular in the sense that it is of the form $\wedge_{\nu_i \geq \alpha} g_i$. The set of models of possibility 1 corresponds to the largest conjunction. Likewise we can consider K^- that contains only negative ground atoms $\neg g_j, g_j \notin H_K$, and is thus consistent as well. Let K^-_α be a cut of K^-. Its set of models is rectangular

in the sense that it is of the form $\wedge_{\kappa_j \geq \alpha} \neg g_j$. It is clear that everything behaves as if the actual base were $K^+ \cup K^-$. As it contains all literals in the negative or positive form only once, there is a model with positive necessity, namely, $\wedge_{g_i \in K^+} g_i \bigwedge \wedge_{g_j \in K^-} \neg g_j$ with necessity at least $\min(\min_{g_i \in K^+} \nu_i \ \min_{g_j \in K^-} \neg \kappa_j$. Hence the possibility of this model is 1, and is unique since there can be at most one model with positive necessity. The least specific possibility distribution induced by $K^+ \cup K^-$ obviously enforces the original necessity degrees as all formulas in $K^+ \cup K^-$ are logically independent from one another.

Given that such π exists, it is not important to know it or to represent one of its corresponding possibilistic bases B explicitly, since K, its associated metadata **Val** and **Comp**, together with a classical reasoner are enough to compute any possibilistic inference, as shown by the following algorithm:

Algorithm 1 (Inference from B).
Input: $K \subset \mathcal{L}_{\text{fact}}$: a KB; $\phi \in \mathcal{L}$: a formula;
Output: $N(\phi)$.
1: $\alpha \leftarrow 0$
2: **if** $K \models \phi$ **then**
3: **for** $T \in \mathcal{T}$ **do**
4: **if** $\phi \in T$ **and** $\alpha < \mathbf{Val}(T)$ **then**
5: $\alpha \leftarrow \mathbf{Val}(T)$
6: **else if** $K \not\models \neg\phi$ **then**
7: **for** $T \in \mathcal{T}$ **do**
8: **if** $\neg\phi \in T$ **and** $\alpha < \mathbf{Comp}(T)$ **then**
9: $\alpha \leftarrow \mathbf{Comp}(T)$
10: **return** α.

Property 1 *Algorithm 1 is correct (i.e., it computes $N(\phi)$).*

Proof. If $K \models \phi$, Equation 4 is applied; otherwise, Equation 3 together with duality: $N(\phi) = 1 - \Pi(\neg\phi)$.

Property 2 *The cost of Algorithm 1 is two classical inferences.*

Proof. Algorithm 1 needs to execute at most two classical inferences: the one in Line 2 and, in case $K \not\models \phi$, the one in Line 6. Checking whether a formula belongs in a topic can be done in a purely syntactic fashion (linear in the length of ϕ) and its cost is thus negligible.

Example 4. Let K be the same as in the previous example, with the following metadata:

T	$\mathbf{Val}(T)$	$\mathbf{Comp}(T)$
$\text{Origin}(x, y)$	α	β
$\text{Airline}(x, \mathsf{AF})$	γ	δ

There are four constants in K (AF, $\mathsf{AF1680}$, CDG, and LHR) and four predicates: Flight$(\cdot)$, Airline$(\cdot,\cdot)$, Dest$(\cdot,\cdot)$, and Origin$(\cdot,\cdot)$. Since there is no typing of the constants in $\mathcal{L}$, we thus construct the Herbrand base

$$\begin{aligned} H_K = \{ \quad & \text{Flight}(\mathsf{AF}), \quad \ldots, \quad \text{Flight}(\mathsf{LHR}), \\ & \text{Airline}(\mathsf{AF},\mathsf{AF}), \ldots, \text{Airline}(\mathsf{LHR},\mathsf{LHR}), \\ & \text{Dest}(\mathsf{AF},\mathsf{AF}), \quad \ldots, \quad \text{Dest}(\mathsf{LHR},\mathsf{LHR}), \\ & \text{Origin}(\mathsf{AF},\mathsf{AF}), \ldots, \text{Origin}(\mathsf{LHR},\mathsf{LHR}) \ \}, \end{aligned}$$

with $\|H_K\| = 52$, which gives $\|\Omega\| = \|2^{H_K}\| = 2^{52} \approx 4.5 \cdot 10^{15}$ interpretations. However, we do not need to explicitly construct π over such an impossibly huge domain. By applying Algorithm 1, we can easily compute, for instance:

$$\begin{aligned} N(\text{Origin}(\mathsf{AF1680},\mathsf{CDG})) &= \alpha, \\ N(\neg\text{Origin}(\mathsf{AF1680},\mathsf{CDG})) &= 0, \\ N(\text{Airline}(\mathsf{AF1680},\mathsf{AF})) &= \gamma, \\ N(\neg\text{Airline}(\mathsf{AF1680},\mathsf{AF})) &= 0, \\ N(\exists x(\text{Flight}(x) \wedge \text{Origin}(x,\mathsf{LHR}))) &= 0, \\ N(\forall x(\neg\text{Flight}(x) \vee \neg\text{Origin}(x,\mathsf{LHR}))) &= \beta. \end{aligned}$$

5 Merging Beliefs from Multiple Sources

Information is provided by different sources. So we need not only to keep track of the certainty levels of the pieces of information, but also of their sources [9]. Keeping track of sources is especially important, in case of conflicting information, to be able to report which sources support what opinions and thus give the user the elements required for a choice. This is why we need a multi-source generalization of possibilistic logic, like the one proposed in [9] and further developed in [4], to combine and reason about the belief bases obtained, as explained in the previous section, by taking the validity and completeness metadata of the source into account.

We shall denote the set of all the sources in the system by $\mathcal{S}$.

A multi-source possibilistic logic formula is a pair (ϕ, F), where ϕ is a logical formula, and $F \subseteq \mathcal{S}$ is a *fuzzy* subset of the set of the sources in the system, i.e., F belongs to the complete distributive lattice $L = [0,1]^{\mathcal{S}}$, equipped with the max-based union $\cup$, min-based intersection $\cap$, and, if we consider another fuzzy set $G \subseteq \mathcal{S}$, the inclusion $F \subseteq G \Leftrightarrow \forall a \in \mathcal{S}, F(a) \leq G(a)$.

The intended meaning of a formula (ϕ, F) is that formula ϕ is believed by a source a at least to degree $F(a)$. Each source believing ϕ somehow belongs to the fuzzy set F. The certainty of ϕ, say $C(\phi)$, is then given by the maximal degree of belief in ϕ associated to the sources in F, which believe ϕ to some extent, and, for any source $a \in \mathcal{S}$, we have that $C(\phi) \geq F(a)$ (a believes that ϕ is true at least at degree $F(a)$). Formulas of the form $(\phi, \emptyset)$ are not written (the system only considers the formulas which are somehow believed by at least one source).

Example 5. Assume there are three sources, $\mathcal{S} = \{a, b, c\}$, where a is Air France, b is British Airways, and c is the Charles de Gaulle airport. Now, let their belief bases be:

ϕ	$F(a)$	$F(b)$	$F(c)$
Dest(AF1680, LHR)	1	0	0.8
Depart(AF1680, 06:25)	0	0.5	0
Depart(AF1680, 07:25)	1	0	1
Arrival(AF1680, 07:45)	1	0.5	0

Let us consider the particular fuzzy sets of sources of the form $F = \alpha/A$, defined as

$$(\alpha/A)(a) = \begin{cases} \alpha \in (0,1], & \text{if } a \in A; \\ 0, & \text{if } a \in \overline{A}. \end{cases}$$

They correspond to a subset A of sources having the same lower bound α on the certainty level of some considered formula. The following equivalence holds between possibilistic logic bases:

$$\{(\phi, \alpha/A), (\phi, \beta/B)\} \equiv \{(\phi, (\alpha/A) \cup (\beta/B))\}. \tag{5}$$

Example 6. We will thus have:

$(\text{Dest}(\mathsf{AF1680}, \mathsf{LHR}), (1/\{a\}) \cup (0.8/\{c\}))$, $\text{Depart}(\mathsf{AF1680}, \mathsf{06{:}25}), 0.5/\{b\})$, $\text{Depart}(\mathsf{AF1680}, \mathsf{07{:}25}), 1/\{a, c\})$, $\text{Arrival}(\mathsf{AF1680}, \mathsf{07{:}45}), (1/\{a\}) \cup (0.5/\{b\}))$, $\text{Arrival}(\mathsf{AF1680}, \mathsf{08{:}45}), 0.8/\{c\})$

A multi-source possibilistic base (which, in the context of this paper, represents a summary of the opinions of multiple sources) is defined as a finite set (i.e., a conjunction) of multi-source possibilistic formulas.

Inference in multi-source possibilistic logic proceeds by refutation, as in standard possibilistic logic: given a base $B = \{(\phi_i, \alpha_i/A_i)\}_{i=1,\ldots,m}$, proving $B \vdash (\phi, F)$ amounts to proving $B \cup \{(\neg\phi, \mathcal{S})\} \vdash (\bot, F)$ by repeated application of the equivalence of Equation 5 and of the resolution rule

$$\frac{(\neg P \vee Q, \alpha/A), (P \vee R, \beta/B)}{(Q \vee R, \min(\alpha, \beta)/(A \cap B))}. \tag{6}$$

The semantics of the multi-source possibilistic logic may be given in terms of a generalization of possibility theory based on a fuzzy-set-valued possibility distribution $\pi : \Omega \to [0,1]^{\mathcal{S}}$. In the context of this work, $\Omega = 2^{H_{\mathcal{L}}}$. The fuzzy-set-valued possibility distribution π associates to every interpretation $\mathcal{I}$ a fuzzy set of sources for which $\mathcal{I}$ is possible; $(\pi(\mathcal{I}))(a)$ is the degree to which source a deems $\mathcal{I}$ possible. Distribution π is normalized if $\exists \mathcal{I}_0 \in \Omega : \pi(\mathcal{I}_0) = \mathcal{S}$. This means that the sources are *collectively* consistent since there exists at least one interpretation that all sources find fully possible. There exists another, weaker form of normalization for such a distribution, which only expresses that the sources are *individually* consistent, namely: $\bigcup_{\mathcal{I} \in \Omega} \pi(\mathcal{I}) = \mathcal{S}$. For instance, the multi-source possibilistic base $B = \{(\phi, 1/A), (\neg\phi, 1/\overline{A})\}$, where $\overline{A} = \mathcal{S} \setminus A$, is clearly not collectively consistent, but it is individually consistent. Indeed here there is partition of the sources into two subsets, those in A that support ϕ and those in $\overline{A}$ that support $\neg\phi$.

The relevant possibility and necessity measures may be defined as follows: for all formulas ϕ,

$$\Pi(\phi) = \bigcup_{\substack{\mathcal{I}\in\Omega \\ \mathcal{I}\models\phi}} \pi(\mathcal{I}), \quad N(\phi) = \bigcap_{\substack{\mathcal{I}\in\Omega \\ \mathcal{I}\not\models\phi}} \overline{\pi(\mathcal{I})}. \tag{7}$$

The distribution associated with base $B = \{(\phi_i, \alpha_i/A_i)\}_{i=1,\ldots,m}$ is

$$\pi_B(\mathcal{I}) = \begin{cases} \mathcal{S}, & \text{if } \mathcal{I} \models \phi_1 \wedge \ldots \wedge \phi_m; \\ \bigcap\limits_{i:\mathcal{I}\not\models\phi_i} (1-\alpha_i)/A_i \ \cup \ \overline{A_i}, & \text{otherwise.} \end{cases}$$

This reflects the fact that if a source in A_i believes with certainty α_i that ϕ_i is true, such a source can find possible an interpretation that violates ϕ_i only at a level that is upper bounded by $1-\alpha_i$. Multiple source possibilistic logic is sound and complete for refutation, with respect to the above semantics [9].

We have now all the formal tools needed to solve the belief fusion problem of providing a coherent answer to queries in presence of possibly conflicting beliefs. The model we propose can process queries which take the form of a formula $\phi \in \mathcal{L}$. If ϕ is closed, then the expected answer is just the fuzzy set of sources according to which ϕ holds. If ϕ is open, the expected answer is a list of substitutions of its free variables, annotated with the fuzzy set of the sources that support it.

To answer a query, the answers provided by the n belief bases are aggregated in a multi-source possibilistic base $B = \{(\phi_i, \alpha_i/A_i)\}_{i=1,\ldots,m}$, which is then used to compile the answer.

Example 7. Continuing the previous example, the result of query $\text{Dest}(x, \mathsf{LHR}) \wedge \text{Depart}(x, y) \wedge \text{Arrival}(x, z)$ requesting all flights with destination London Heathrow, together with their departure and arrival times, would be

x	y	z	$F(a)$	$F(b)$	$F(c)$
AF1680	7:25	7:45	1	0	0
AF1680	7:25	8:45	0	0	0.8

or, in a more synthetic form,

x	y	z	F
AF1680	7:25	7:45	$1/\{a\}$
AF1680	7:25	8:45	$0.8/\{c\}$

The result of query $\exists x \text{Dest}(x, \mathsf{LHR}) \wedge \text{Arrival}(x, 8{:}45)$ asking whether a flight exists with destination London Heathrow arriving at 8:45, would be, in synthetic form, $0.8/\{c\}$.

6 Related Work

Our proposal fills a gap at the intersection of two fields of in- vestigation, namely distributed information systems and possibilistic logic.

The problem of reasoning about validity and complete- ness in relational databases was first addressed by Demolombe [8] in the setting of modal logic.

Recent work on collaborative access control in distributed datalog [1] shares some common intuitions and concerns with our model. However, this approach, which is based on provenance calculus [13], does not handle uncertainty (although probabibistic c-tables are also encompassed by provenance calculus). The approach proposed in the present paper is anyway more in the spirit of possibilistic c-tables, which have been recently introduced in [18].

Finally, the idea of associating subsets of sources as supporting arguments to answers has been suggested in [2] in the context of numerical information fusion.

7 Conclusion

We have presented a solution to construct a possibilistic belief base from a crisp knowledge base using topical validity and completeness metadata. The main result is that possibilistic inferences from such belief base can be performed at the cost of two classical inferences, which is less than the cost of inference on a general possibilistic belief base. Furthermore, our solution can be straightforwardly adapted to KB representation standards like datalog and RDF + OWL.

We have also shown how to exploit the expressive power of multi-source possibilistic logic to provide the user with a comprehensive logical summary of the different opinions held by the sources. Nevertheless, it is likely that a user might be happier with receiving less detailed information in response to her queries. We see basically two directions that might be followed to alleviate the cognitive load for the end user:

- give the user the option of specifying a maximum number k of answers, to be used to select only the k most certain answers according to their supporting sources, so that each answer be simply annotated with a crisp set of sources that support it;
- if a taxonomy of sources is available (e.g., based on their sector, geographical location, etc.), the sets of sources supporting an answer could be "linguistically synthesized" (in the sense of Zadeh's [21]) by categorical labels, like "all the airlines based in the UK" or "most airport operators", which are certainly easier to understand and process than extensive lists of sources.

References

1. Abiteboul, S., Bourhis, P., Vianu, V.: A formal study of collaborative access control in distributed datalog. In: ICDT. LIPIcs, vol. 48, pp. 10:1–10:17. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2016)
2. Assaghir, Z., Napoli, A., Kaytoue, M., Dubois, D., Prade, H.: Numerical information fusion: Lattice of answers with supporting arguments. In: ICTAI. pp. 621–628. IEEE Computer Society (2011)
3. Bacchus, F., Grove, A.J., Halpern, J.Y., Koller, D.: From statistical knowledge bases to degrees of belief. Artif. Intell. 87(1-2), 75–143 (1996)

4. Belhadi, A., Dubois, D., Khellaf-Haned, F., Prade, H.: Multiple agent possibilistic logic. Journal of Applied Non-Classical Logics 23(4), 299–320 (2013)
5. Bernays, P., Schönfinkel, M.: Zum Entscheidungsproblem der mathematischen Logik. Math. Ann. 99, 342–372 (1928)
6. Cholvy, L.: Collecting information reported by imperfect information sources. In: Advances in Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2012, Catania, Italy, July 9-13, 2012, Proceedings, Part III. pp. 501–510 (2012)
7. Cholvy, L., Demolombe, R., Jones, A.J.I.: Reasoning about the safety of information: From logical formalization to operational definition. In: DAISD. pp. 345–373 (1994)
8. Demolombe, R.: Answering queries about validity and completeness of data: From modal logic to relational algebra. In: FQAS. Datalogiske Skrifter (Writings on Computer Science), vol. 62, pp. 265–276. Roskilde University (1996)
9. Dubois, D., Lang, J., Prade, H.: Dealing with multi-source information in possibilistic logic. In: ECAI. pp. 38–42 (1992)
10. Dubois, D., Lang, J., Prade, H.: Possibilistic logic. In: Handbook of logic in artificial intelligence and logic programming (vol. 3): nonmonotonic reasoning and uncertain reasoning, pp. 439–513. Oxford University Press, New York, NY, USA (1994)
11. Dubois, D., Prade, H.: Possibility Theory—An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
12. Dubois, D., Prade, H.: Valid or complete information in databases - A possibility theory-based analysis. In: DEXA. Lecture Notes in Computer Science, vol. 1308, pp. 603–612. Springer (1997)
13. Green, T.J., Karvounarakis, G., Tannen, V.: Provenance semirings. In: Libkin, L. (ed.) Proc. 26th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems, Beijing, June 11-13. pp. 31–40. ACM (2007)
14. Lang, J.: Possibilistic logic: complexity and algorithms. In: Kohlas, J., Moral, S. (eds.) Algorithms for Uncertainty and Defeasible Reasoning, pp. 179–220. Vol. 5 of Handbook of Defeasible Reasoning and Uncertainty Management Systems (Gabbay, D. M. and Smets, Ph., eds.), Kluwer Acad. Publ., Dordrecht (2001)
15. Lewis, H.R.: Complexity results for classes of quantificational formulas. Journal of Computer and System Sciences 21, 317–353 (1980)
16. Marsh, S.P.: Formalising Trust as a Computational Concept. Ph.D. thesis, Department of Computing Science and Mathematics University of Stirling (1994)
17. Motro, A.: Integrity = validity + completeness. ACM Trans. Database Syst. 14(4), 480–502 (Dec 1989)
18. Pivert, O., Prade, H.: Possibilistic conditional tables. In: Gyssens, M., Simari, G.R. (eds.) Proc. 9th Int. Symp. on Foundations of Information and Knowledge Systems (FoIKS'16), Linz, March 7-11. LNCS, vol. 9616, pp. 42–61. Springer (2016)
19. Schwitzgebel, E.: Belief. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Stanford University (Fall 2008), `http://plato.stanford.edu/archives/fall2008/entries/belief/`
20. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
21. Zadeh, L.A.: A theory of approximate reasoning. In: Hayes, J.E., Mitchie, D., Mikulich, L.I. (eds.) Machine intelligence, Vol. 9, pp. 149–194. Halstead Press, New York (1979)