Using Cognitive Entropy to Manage Uncertain Concepts in Formal Ontologies

Joaquín Borrego-Díaz and Antonia M. Chávez-González

Departamento de Ciencias de la Computación e Inteligencia Artificial
E.T.S. Ingeniería Informática-Universidad de Sevilla
Avda. Reina Mercedes s.n. 41012-Sevilla
{jborrego,tchavez}@us.es

Abstract. A logical formalism to support the insertion of uncertain concepts in formal ontologies is presented. It is based on the search of extensions by means of two automated reasoning systems (ARS), and it is driven by what we call *cognitive entropy*.

1 Introduction

The challenge of data management with logical trust arose from the statement of the Semantic Web (SW). An important problem is the need for extending or revising ontologies. Such task is, from the point of view of companies, dangerous and expensive: since every change in ontology would affect the overall knowledge of the organization. It is also hard to be automated, because some criteria for revision cannot be fully formalized. Despite its importance, the tools designed to facilitate the syntactic extension or ontological mapping do not analyze, in general, their effect on the (automated) reasoning.

Our aim is to design tools for extending ontologies in a semi-automated way, that is one of the problems present in several methods for cleaning data in the SW, when it implies ontological revision (see e.g. [1] [3]). The method is based on the preservation by extensions of the notion of *ontology robustness*, see [8]. *lattice categoricity*, (described in sect. 3), is going to be applied in a special case: the change is induced by the user, who has detected the (cognitive) necessity of adding a *notion*. That is, a vague concept which comprises a set of elements with features roughly shaped by the existing concepts. In Ontological Engineering, careful consideration should be paid to the accurate classification of objects: the notion becomes a *concept* when its behavior is constrained by new axioms that relate it to the initial concepts. This scenario emphasizes the current need for an explanation of the reasoning behind cleaning programs. That is, a *formalized explanation* of the decisions made by systems. Note that such explanations are necessary for the desirable design of logical algorithms to be used by general-purpose cleaning agents [4]. It is evident that the task will need not only specific

⋆ This work is partially supported by the project TIN2004-03884 of Spanish Ministry of Education and Science, cofinanced by FEDER founds.

automated reasoning systems (ARSs) for SW, but also those for general purpose. The reason is that some tasks are not directly related to reasoning services for the SW [2] [17] [8]. Thus, we use ARSs for first order logic theories, in favor of one reaches major generality. Among the challenges the problem raises in a dynamic setting as the SW, there are three of them which are specially interesting from the point of view of automated reasoning. They seem to obstruct the design of a fully formalised methodology [4] from classical database field:

- We can not suppose the database to be stable (because new facts could be added in the future).
- Usually, the specification of an ontology is syntactically complex, so it is very likely that classical axiomatization of database theory becomes inconsistent, even if ontology itself is consistent.
- It is possible that the database does not contain facts about the whole relations of the language.

However, some limitations can be solved by weakening the requirements imposed in both database and ontological reasoning [8] [2].

The method proposed is based on the assistance of two ARS, McCune's OTTER and MACE4 (`http://www-unix-mcs.anl.gov`). The first one, OTTER, is an automated theorem prover (ATP) based on resolution and *support set* strategy. The program allows great autonomy: its `auto2` mode suffices to find almost every automated proof that have been required. The second one, MACE4, is an automatic model finder sharing formula syntax with OTTER. It is based on Davis-Putnam-Loveland-Longemann's procedure to decide satisfiability. It has been useful for analyzing the models of the involved theories.

Finally, it would be good to add some information about MACE4. Despite it has not been formally verified to work correctly, once the result by MACE4 is determined, it is not difficult to certify that the models it gives are correct. It is necessary to use OTTER to prove that the list of models is exhaustive. Thus, MACE4 has been used as an automatic assistant to induce new results and investigate the effect of diverse axiomatizations, which must be certified later.

2 Logic-Based Ontological Extensions

Once the need for revision is accepted, the task can be seen, up to some extent -and specially when one designs her/his own logical theory-, from two points of view. The first one considers it like a task similar to belief revision, analyzing it by classic methods of AI. Nevertheless, the effort can be expensive, because it must study once again the impact of revision on the foundational features of the source ontology. The second one has a foundational character. The evolution of ontology should obey ground principles which are accepted on this matter. For example, preserving some sort of backward compatibility, if it is possible (extracted from [15]):

- *The ontology should be able to extend other ontologies with new terms and definitions.*

– *The revision of an ontology should not change the well-formedness of resources that commit themselves to an earlier version of the ontology.*

However, such principles are more adequate if the source ontology is *robust*, in the following sense [4]: *An ontology is robust if its core is clear, stable (except for extensions); if every model of its core exhibits similar properties w.r.t. the core language, and if it is capable of admitting (minor) changes made out of the core without committing core consistency.* By *core* we understand a portion of ontology that we consider as a sound theory with well known properties, and which is accepted as the best for the concepts involved. We can consider two kinds of extensions:

– *Extension by definition.* It produces conservative extensions. If definitions are not provided for the new elements, conservation can fail.
– *Ontological insertion*: Essentially new (nondefinable) concepts/relations are inserted. The task is to design good axioms to specify the new ones from core theory.

An interesting case occurs in the task of ATP-aided cleaning of logic databases. The *bottom-up change generation* in ontologies -due to the analysis of track interaction among the Knowledge Base, the ATP and the user- induces ontological revision. It can simulate new elements in ontology to be inserted (such as Skolem noise [2]). We analyze here a slightly different problem, which appears when the user is the person who decides to insert a new concept by collecting a set of data.

The *extension by definition* is the basis of *definitional methodologies* for building formal ontologies. It is based on the following principles [7]:

1. *Ontologies should be based upon a small number of primitive concepts.*
2. *These primitives should be given definite model theoretic semantics.*
3. *Axioms should only be given for the primitive concepts.*
4. *Categorical axiom sets should be sought.*
5. *The remaining vocabulary of the ontology (which may be very large), should be introduced purely by means of definitions.*

In this paper, the first three principles are assumed. The fourth one will be replaced by *lattice categoricity.* Categoricity is a strong requirement that can be hard to achieve and to preserve. Even when it is achieved, the resultant theory may be unmanageable (even undecidable) or unintuitive. This phenomenon might suggest that we restrict the analysis of completeness to coherent parts of the theory. However, it is not a *local* notion: since minor changes commit the categoricity and it is expensive to repeat the logical analysis.

With respect to the last principle, starting with a basic theory, it seems hard to define a new concept/relationship. It is better to consider it only as the starting point to build an ontology, thinking thus that we are in early steps of the process, where ontological insertions are necessary.

Finally (although it is not the topic of this paper), we would like to add that an ontological insertion should be supported by a good theory about its

relationship with the original ontology. It should as well be supported by a nice way of expanding a representative class of models of the source theory to the new one. This class of models must contain the *intended* models (those that the ontology designer wants to represent). It can be required an interpretation of the new elements which should be formalised, and a re-interpretation of the older ones, which must be compatible with basic original principles.

3 Lattice Categorical Theories

In order to solve in practice the several logical problems that ontological insertion raises we will analyze the categoricity of the structure of the concepts of the ontology. We are going to take into account compatibility which has been previously mentioned, and we will try to obtain definitions of the concepts inserted in the new ontology. We will also analyze categoricity of structure of concepts of ontology. For the sake of clarity, we suppose that the set of concepts has a lattice structure. Actually, this is not a constraint: there are methods to extract ontologies from data which produce such structure (such as the Formal Concepts Analysis [14]) and, in general, the ontology is easy to be extended by definition, verifying lattice structure. Although we think about Description Logics [5] as ontological language (the logical basis for ontology languages as OWL, see `http://www.w3.org/TR/owl-features/`), the definitions are useful for full first order logic (FOL), so we give the definitions in FOL language.

On the one hand, a *lattice categorical* theory is the one that proves the lattice structure of its basic relationships. This notion is weaker than categoricity or completeness. On the other hand, lattice categoricity is a reasonable requirement: the theory must certify the basic relationships among the primitive concepts. In [8] we argued that completeness can be replaced by *lattice categoricity* to facilitate the design of feasible methods for extending ontologies. Let us summarize these ideas.

Given a fixed FOL language, let $\mathcal{C} = \{C_1, \ldots, C_n\}$ be a (finite) set of concept symbols, let T be a theory (in the general case, definable concepts in T can be considered). Given M a model of T, $M \models T$, we consider the structure $L(M, \mathcal{C})$, in the language $L_{\mathcal{C}} = \{\top, \bot, \leq\} \cup \{c_1, \ldots, c_n\}$, whose universe are the interpretations in M of the concepts (interpreting c_i as C_i^M), $\top$ is M, $\bot$ is $\emptyset$ and $\leq$ is the subset relation. We assume from now on that $L(M, \mathcal{C})$ is requested to have a lattice structure for every theory we consider. This requirement simplifies the examples.

The relationship between $L(M, \mathcal{C})$ and the model M itself is based in two facts. The first one, the lattice L can be characterized by a finite set of equations E_L, plus a set of formulas $\Theta_{\mathcal{C}}$ categorizing the lattice under completion, that is, $\Theta_{\mathcal{C}}$ includes the domain closure axiom, the unique names axioms and, additionally, the axioms of lattice theory. Thus, every model M of $E \cup \Theta_{\mathcal{C}}$ is finite. The second one, there exists a natural translation Π of these $L_{\mathcal{C}}$-equations into formulas in the FOL language so that if E is a set of equations characterizing $L(M, \mathcal{C})$ (so $L(M, \mathcal{C}) \models E$), then $M \models \Pi(E)$.

Definition 1. *Let E be a $L_{\mathcal{C}}$-theory. We say that E is a* **lattice skeleton** *(l.s.) for a theory T if E verifies that*

- *There is $M \models T$ such that $L(M, \mathcal{C}) \models E \cup \Theta_{\mathcal{C}}$, and*
- *$E \cup \Theta_{\mathcal{C}}$ has an unique model (modulo isomorphism).*

Every consistent theory has a lattice skeleton [8]. Roughly speaking, the existence of essentially different lattice skeletons makes difficult to reason with the ontology while the existence of only one would make it easy.

Definition 2. *T is called a* **lattice categorical (l.c.) theory** *if whatever pair of lattice skeletons for T are equivalent modulo $\Theta_{\mathcal{C}}$.*

Note that if T is l.c. and E is a l.s. of T, then $T \vdash \Pi(E)$. Note also that every consistent theory T has an extension T' which is lattice categorical: it suffices to consider a model $M \models T$, and then to find a set E of equations such that $\Theta_{\mathcal{C}} \cup E$ has $L(M, \mathcal{C})$ as only model. The theory $T \cup \Pi(E)$ (and any consistent extension of it) is l.c.

Finally, we can give a formalization of *robust ontological extension*, based in the categorical extension of the ontology:

Definition 3. *Given two pairs $(T_1, E_1), (T_2, E_2)$ we will say that (T_2, E_2) is a* **lattice categorical extension** *of (T_1, E_1) with respect to the sets of concepts $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively, if $\mathcal{C}_1 \subseteq \mathcal{C}_2$ and $L(T_2, \mathcal{C}_2)$ is an E_1-conservative extension of $L(T_1, \mathcal{C}_1)$.*

For reasoning with the lattice of concepts it suffices to work with a lattice skeleton, so, to simplify, we suppose throughout that T is the self l.s.

3.1 Cognitive Support

Once formalized the notion of *lattice categorical extension*, we need to design several functions to advise how to select the best l.c. extension.

Assume that T is a theory, and L is the lattice defined by $\mathcal{C}$ in some $M \models T$. From the point of view of ontology designer, such a model M is the *intended* model that the ontology attempts to represent. Suppose that $\Delta = \{h_1, \ldots h_n\}$ is the set of facts on $\mathcal{C}$, and the user wants to classify some elements that occur in Δ by means of a new concept. We can suppose, to simplify the notation, that every fact explicit in T belongs to Δ. Let $U(\Delta)$ be the universe determined by Δ; that is, $\{\boldsymbol{a} \ : \ \exists\, C \in \mathcal{C}\ [C(\boldsymbol{a}) \in \Delta]\}$.

Given $C \in \mathcal{C}$ in Δ, we consider

$$|C|^{\Delta} := |\{\boldsymbol{a} \ : \ C(\boldsymbol{a}) \in \Delta\}| \text{ and } |C|^{\Delta}_{T} := |\{\boldsymbol{a} \in U(\Delta) \ : \ T \cup \Delta \models C(\boldsymbol{a})\}|.$$

Definition 4. *The* **cognitive support** *of C with respect to Δ, T and L, is*

$$sup^{L}_{T,\Delta}(C) := \frac{|\{\boldsymbol{a} \in U(\Delta) \ : \ \exists i[C_i \leq^{L} C \wedge T \cup \Delta \models C_i(\boldsymbol{a})]\}|}{|U(\Delta)|}$$

This support estimates the number of facts on the concept C entailed by T, normalized by the size of the universe $U(\Delta)$. Because of the computational complexity of logical reasoning, it can be hard in general to compute it: we need to seek, by logical entailment, the cone of concepts defined by C. However, this computation is trivial for lattice categorical theories:

Proposition 1. *If T is lattice categorical, then* $sup^L_{T,\Delta}(C) = \dfrac{|C|^{\Delta}_T}{|U(\Delta)|}$

The proposition holds because if $C_i \leq^L C$, then $T \models C_i \sqsubseteq C$. Thus, if $T \cup \Delta \models C_i(\boldsymbol{a})$, then $T \cup \Delta \models C(\boldsymbol{a})$.

From now on, we suppose that Δ is compounded by facts on atoms of the lattice of concepts (that is, about the most specific concepts). Note, also, that if T is l.c., then L is unique, and we will thus omit the superscript L in that case.

Corollary 1. *If $\mathcal{J} = \{C_1, \ldots, C_n\}$ is a Jointly Exhaustive and Pairwise Disjoint (JEPD) set of concepts in L, then $sup_{T,\Delta}(.)$ is a probability measure.*

Proof. It is easily seen that $\sum_{C \in \mathcal{J}} sup^L_{T,\Delta}(C) = 1$.

The **cognitive entropy** of $\mathcal{J}$ is $CH(\mathcal{J}) = -\sum_{C \in \mathcal{J}} sup_{T,\Delta}(C) \log sup_{T,\Delta}(C)$.

3.2 Entropy of Ontological Extensions

Suppose that the user decides that a set $\{a_1, \ldots, a_k\} \subseteq U(\Delta)$ induces a new concept D (provisionally, a notion). Such a notion might not be fully represented by those elements. Also, it is possible that some of them do not belong to the new concept, because of noise in the data. It might also be the case that the concept is constrained by a set Σ of axioms introduced by the user. Furthermore it is also possible that $T \cup \Sigma$ is not l. c., that is, this theory does not prove the intended lattice induced by $\mathcal{C} \cup \{D\}$. MACE4 provides the collection $\{L_1, \ldots, L_m\}$ of the lattices induced by the models of $T \cup \Sigma$. Let T_i be a lattice skeleton for L_i $(i = 1, \ldots, m)$.

Now, we focus our attention on a concrete *level* of the Ontology, where we intend to insert the new concept. The level will be a JEPD $\mathcal{J} = \{C_1, \ldots, C_k\}$ of the lattice L verifying that if the new concept D contains some of them,

$$\mathcal{J}^{L_i}_{\restriction D} = \{C_i \in \mathcal{J} \ : \ C_i \leq^{L_i} D\} \neq \emptyset$$

then $\mathcal{J}_i = (\mathcal{J} \setminus \mathcal{J}^{L_i}_{\restriction D}) \cup \{D\}$ is a JEPD in L_i. Since T_i is a l.c. extension of T, the support of D is easy to achieve:

Theorem 1. *In above conditions,* $sup_{T_i,\Delta}(D) = \sum_{C \in \mathcal{J}^{L_i}_{\restriction D}} sup^L_{T,\Delta}(C_i)$

To estimate the conditional entropy of the new extension, we consider a natural definition of *conditional support*:

$$sup_{T_i,T,\Delta}(C'|C) := \frac{|\{\boldsymbol{a} \in U(\Delta) \ : \ T \cup \Delta \models C(\boldsymbol{a}) \wedge T_i \cup \Delta \models C'(\boldsymbol{a})\}|}{|C|^{\Delta}_T}$$

This support allows to estimate the amount of new information produced by the extension by standard methods; through the *conditional entropy* associated to the two probability measures. The **conditional cognitive entropy** is :

$$CH(\mathcal{J}||\mathcal{J}_i) = -\sum_{\substack{C' \in \mathcal{J} \\ C \in \mathcal{J}_i}} sup_{T_i,\Delta}(C'|C) \log sup_{T_i,\Delta}(C'|C)$$

This sum can be simplified (assuming $0 \log 0 = 0$): if $C = C'$ or $C, C' \in \mathcal{J}$, then

$$sup_{T_i,T,\Delta}(C'|C) \log sup_{T_i,T,\Delta}(C'|C) = 0$$

and the following property holds:

Proposition 2. *In above conditions,* $sup_{T_i,T,\Delta}(C'|C) = \dfrac{|C'|^{\Delta}_{T}}{|C|^{\Delta}_{T_i}}$

This entropy is similar to Kullback-Leibler distance or relative entropy (see [16]), but using the entailment to classify the elements. It is known that it is minor than the initial entropy. In [13] similar entropies are used, but based on probabilistic assignation. Finally, in order to estimate what is the best extension for our purposes, it is necessary to compute the The **Shannon's diversity index** for each L_i. This index normalizes the amount of information produced by the extension, and is defined as

$$IH(\mathcal{J}_i) = \frac{CH(\mathcal{J}||\mathcal{J}_i)}{\log |\mathcal{J}_i|}$$

The interpretation of the index is as follows: if we select L_i with minimum $IH(\mathcal{J}_i)$, the new information produced by the new concept is minor. This option is the *cautious* one: the reparation of the source ontology is *light* and we do not expect big changes in the representation of the intended model. If we select L_i with an upper $IH(J)$, the change of the information is more relevant; we select such an extension if we regard as robust the specification of the concept given by Σ together with the facts. In general, we have to chose the l.c. extension with minor index. Intuitively, in this way we do not change too much the information of the initial ontology.

4 An Example

We would like to show a short example in the field of Qualitative Spatial Reasoning (QSR). *Region Connection Calculus* (RCC) [12] is a well-known mereotopological approach to QSR, that we can consider to be a robust ontology. For RCC, the *spatial entities* are non-empty regular sets. The primary relation between them is *connection*, $C(x, y)$, with intended meaning: *"the topological closures of* `x` *and* `y` *intersect"*. The basic axioms of RCC are $A_1 := \forall x[C(x, x)]$ and $A_2 := \forall x, y[C(x, y) \rightarrow C(y, x)]$ jointly with a set of definitions on the main spatial relations (fig. 1), and other axioms not used here (see [12]).

$DC(x,y) \leftrightarrow \neg C(x,y)$	(x is disconnected from y)
$P(x,y) \leftrightarrow \forall z[C(z,x) \rightarrow C(z,y)]$	(x is part of y)
$PP(x,y) \leftrightarrow P(x,y) \wedge \neg P(y,x)$	(x is proper part of y)
$EQ(x,y) \leftrightarrow P(x,y) \wedge P(y,x)$	(x is identical to y)
$O(x,y) \leftrightarrow \exists z[P(z,x) \wedge P(z,y)]$	(x overlaps y)
$DR(x,y) \leftrightarrow \neg O(x,y)$	(x is discrete from y)
$PO(x,y) \leftrightarrow O(x,y) \wedge \neg P(x,y) \wedge \neg P(y,x)$	(x partially overlaps y)
$EC(x,y) \leftrightarrow C(x,y) \wedge \neg O(x,y)$	(x is externally connected to y)
$TPP(x,y) \leftrightarrow PP(x,y) \wedge \exists z[EC(z,x) \wedge EC(z,y)]$	(x is a tangential prop. part of y)
$NTPP(x,y) \leftrightarrow PP(x,y) \wedge \neg\exists z[EC(z,x) \wedge EC(z,y)]$	(x is a non-tang. prop. part of y)

Fig. 1. Axioms of RCC

We have proved (by using MACE4 and OTTER) that the set of formulas E given in the figure 2 categorises under completion the lattice of the RCC-spatial relationships (given in fig. 3). The set of binary relations formed by the eight (JEPD) relations given in figure 3 is denoted by RCC8. If this set is thought to be a calculus, all possible unions of the basic relations are also used. Another interesting calculus is RCC5, based on $\{DR, PO, PP, PPi, EQ\}$.

$$\begin{array}{rll}
\top \equiv C \sqcup DR & PO \sqsubseteq \neg P \sqcap \neg Pi \sqcap \neg DR & DR \equiv EC \sqcup DC \\
NTPP \sqsubseteq \neg TPP \sqcap \neg Pi \sqcap \neg DR & C \equiv O \sqcup EC & TPP \sqsubseteq \neg Pi \sqcap \neg DR \\
O \equiv PO \sqcup P \sqcup Pi & EQ \sqsubseteq \neg PPi \sqcap \neg DR & Pi \equiv EQ \sqcup PPi \\
TPPi \sqsubseteq \neg NTPPi \sqcap \neg DR & P \equiv EQ \sqcup PP & NTPPi \sqsubseteq \neg DR \\
PPi \equiv TPPi \sqcup NTPPi & EC \sqsubseteq \neg DC & PP \equiv TPP \sqcup NTPP
\end{array}$$

Fig. 2. A skeleton for RCC

Suppose that if we insert a new spatial uncertain relation D expressing *"*x *and* y *have a isometric overlapping relation"*; that is, D covers *partial overlapping* PO and *extensional equality* EQ relationships. That is, proper part is not possible between isometric objects. This is suggested by the study of spatial relationships among identical objects (e.g. the 2-D spatial configuration of a set of coins). Thus, we consider that the new relation D satisfies

$$RCC \cup \{\forall x \forall y(PO(x,y) \rightarrow D(x,y)), \forall x \forall y(EQ(x,y) \rightarrow D(x,y))\}$$

or, in terms of skeleton, $E \cup \{PO \sqsubseteq D, EQ \sqsubseteq D\}$. MACE4 produces seven l.c. extensions (classified according to their lattices in fig. 4). All these extensions can be mereotopologically interpreted [11]. Suppose that the set that motivates the extension is:

$$\Delta := \left\{ \begin{array}{llll}
PO(m_1,m_2) & EQ(m_2,m_3) & EQ(m_3,m_4) & PO(m_1,m_3)\ DC(m_4,m_6) \\
DC(m_3,m_5) & PO(m_5,m_1) & NTPP(c_1,m_3) & EC(c_2,m_1)\ \ TPP(c_2,c_5) \\
DC(c_1,c_2) & TPPi(c_5,c_2) & NTPP(m_2,c_4) & DC(m_1,c_3)\ \ TPP(c_1,c_3)
\end{array} \right.$$

In this case, $|U(\Delta)| = 15$, and the basic JEPD is the set $\mathcal{J} = \{PO, PP, EQ, PPi, EC, DC\}$. In each L_i, $\mathcal{J}_i$ is a JEPD, so we can assign conditional entropy and

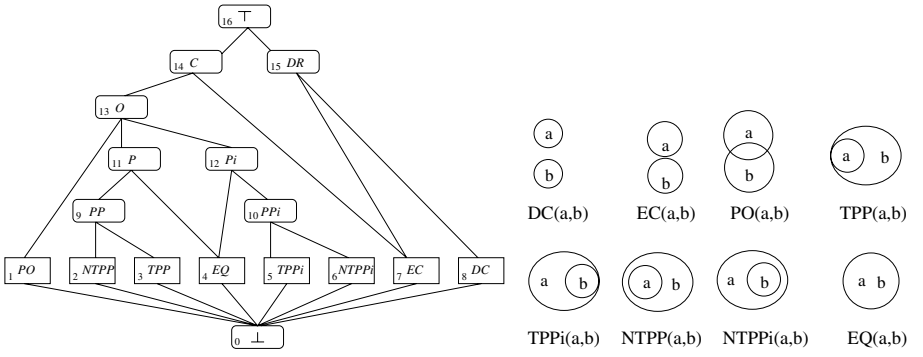


Fig. 3. The lattice of spatial relations of RCC (left) and he relations of RCC8 (right)

Shannon's diversity index to each extension.Thus, $T_2 \equiv E+\cup\{D \equiv PO \sqcup EQ\}$ is the selected l.c. extension because it has the minimum Shannon's index. On the other hand, the user's notion might be inconsistent. For instance, if the user's proposal for Σ' is $\{PO \sqsubseteq D, EQ \sqsubseteq D, P \sqsubseteq D, D \sqsubseteq O\}$, then there is not any l.c. extension, a fact that we have certified using MACE4 and OTTER.

5 Data-Driven Ontology Revision: Deficient Data Layout

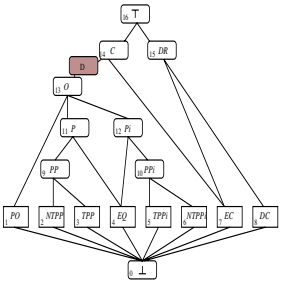
In above sections, we have formalized the insertion of a concept that will remain well defined once the appropriate extension is selected. In that case, the computation of entropies is easier than the entropies defined in this section. Now, we aim to extend the ontology in a provisional way because the deficient classification of data induces the insertion of subconcepts for refining the classification of individuals which initially were misclassified[1]. In this case the new concepts will fall in the bottom level. Therefore, we aim to extend $\mathcal{J}_L$, the JEPD set of concepts which are the atoms of the lattice $L(T, \mathcal{C})$.

The following definition formalizes the notion of *insertion of a concept with certain degree of imprecision* as subconcept of a given concept C. It has to be determined whether there is a l.c. extension of the ontology with an (atomic) subconcept μC of C. Intuitively, the meaning of $\mu C(\mathbf{a})$ is "the concept $\mathbf{a}$ falls in the concept C, but we do not know any more specific information about $\mathbf{a}$". Formally,

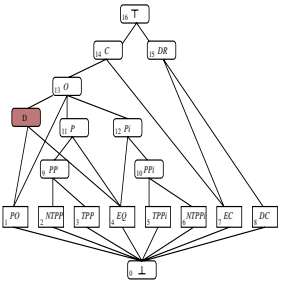
Definition 5. *Let (T, E_0) be a l.c. core and $C \in \mathcal{C}$. We say that the ontology admits an undefinition at C ($T \rightsquigarrow_w C$) if there is a l.c. extension of T, (T', E'), such that*

1. *T' is l.c. with respect to $\mathcal{C} \cup \{\mu C\}$, (where $\mu C \notin \mathcal{C}$).*
2. *$\{\mu C\}$ is an atom in the lattice $L' = L(T, \mathcal{C} \cup \{\mu C\})$.*
3. *There is not C' such that $\mu C <^{L'} C' <^{L'} C$.*

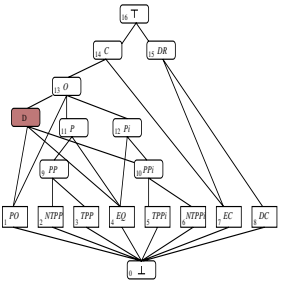
[1] This approach is inspired in the study presented at Eurocast 2007 [9].



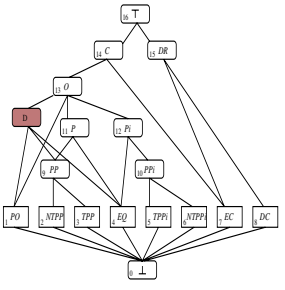
CH(L_1)=1.23
IH(L_1)=0.37



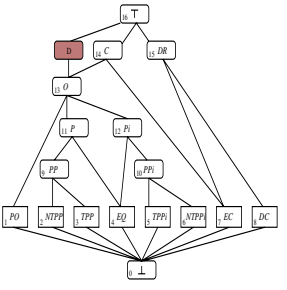
CH(L_2)=0.32
IH(L_2)=0.13



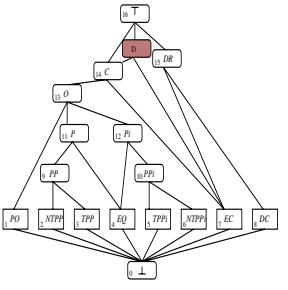
IH(L_3)=0.58
IH(L_3)=0.22



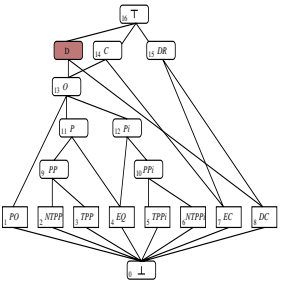
CH(L_4)=0.9
IH(L_4)=0.28



CH(L_5)=1.23
IH(L_5)=0.37



CH(L_6)=1.6
IH(L_6)=0.48



CH(L_7)=2.03
IH(L_7)=0.53

Fig. 4. The seven l.c. extensions by insertion. The grey box denotes the new relation.

Note that, in above conditions, $\mathcal{J}_L[\mu C] := \mathcal{J}_L \cup \{\mu C\}$ is a JEPD set for L' (see fig. 5, left). This requirement represents, in fact, that we have not any additional information about μC. For example, in figure 5 right, the relation $\mu C(a, b)$ means "the regions a and b are connected, but it is unknown if they overlap or they are externally connected".

The notation $T \models_\mu C(\boldsymbol{a})$ means $T \models C(\boldsymbol{a})$ and, for all $D <^L C$, $T \not\models D(a)$. In other words, $C(a)$ is the most specific knowledge on $\mathbf{a}$ entiled from T. It is easy to see that

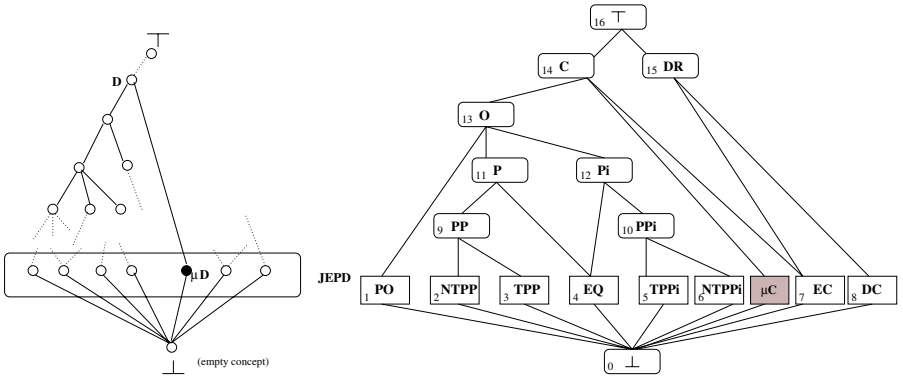


Fig. 5. The ontology (RCC, E) admits an undefinition in the concept C (*connection*) (right)

Proposition 3. *Whatever two extensions by undefinition at C of T have equivalent lattice skeletons modulo completion.*

Such a skeleton of the extension is denoted by $E[\mu C]$. We can also consider the iteration of this kind of extensions, namely $E[\mu C_1, \ldots, \mu C_k]$.

Corollary 2. *$E[\{\mu C \ : \ C \in \mathcal{C} \ \wedge \ T \rightsquigarrow_w C\}]$ is unique (modulo database completion axioms).*

5.1 Inserting Provisional Spatial Relationships in RCC

As we have already commented, the JEPD set named RCC8 is the representation of a precise classification for RCC. In order to build the adequate extension, we compute first the list of extensions obtained by inserting only one relation.

Theorem 2. *There are exactly eight extensions by undefinition of the lattice of RCC by insertion of a new relation D such that $RCC8 \cup \{D\}$ is a JEPD set.*

Such new relations can be mereotopologically interpreted [11]. A proof of this result appears at [11]. The lattices of extensions are detailed at [8]. For example, the lattice depicted in fig. 5 (right) has a skeleton $E[\mu C]$.

The next step consists in deciding which is the best l.c. extension to classify data. Suppose that $\Delta = \{h_1, \ldots h_n\}$ is the set of facts. Assume that the user believes that the set of misclassified elements is $I = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\} \subseteq U(\Delta)$ (according with user's ontology). In this case, the problem is not due to a new concept, because the user has not decided yet an insertion. Such elements are not falling on atomic concepts ($T \not\models C(\boldsymbol{a})$ for any $C \in \mathcal{J}_L$), because the user has not an specific definition of them, that is, he has got only unprecise information (as, instances of upper concepts).

It is easy to provide an extension by undefinition with complete classification of data. For each $\boldsymbol{a}_i \in I$, let $C^i \in \mathcal{C}$ such that $T \models_\mu C^i(\boldsymbol{a}_i)$. Any extension by undefinition at the set $\{C^i \ : \ i = 1, \ldots k\}$ classifies every element of $U(\Delta)$ with

a concept of the JEPD set $\mathcal{J}_{T'} := \mathcal{J} \cup \{\mu C^1, \ldots, \mu C^k\}$. Note also, that if we do not require C^i is the most specific one, the extension is not unique.

Definition 6. *Let T' be an extension by undefinition of T defined as in 5. The support of μC is defined as*

$$supp_{T',\Delta}(\mu C) = \frac{|\{\boldsymbol{a} \in U(\Delta) : \boldsymbol{a} \in I \wedge T \cup \Delta \models_\mu C(\boldsymbol{a})\}|}{|U(\Delta)|}$$

That is, the support of μC uses the number of elements for such that T proves they belong to C. In this way $supp_{T',\Delta}$ is also a probability measure on $\mathcal{J}_{T'}$. Note that this computation is equivalent to consider the support with respect to the theory $T' \cup \{\mu C(\boldsymbol{a}) \ : \ T \cup \Delta \models_\mu C(\boldsymbol{a})\}$. To simplify, we consider throughout that T' is that theory.

Theorem 3. *The extension above defined exhibits the maximum cognitive entropy among every possible extension by undefinition classifying $U(\Delta)$.*

Sketch of proof: If T'' is other extension, then some $\boldsymbol{a}_i$ of I are classified with respect to a concept which is not the most specific one. Thus the result follows by the convexity of the function $p \log p$.

A l.c. extension by undefinition with maximum entropy gives little information on new concepts. This option is a *cautious* solution to the problem, because strong requirements for the new concepts are not been imposed.

Fig. 6. Map of United States

5.2 A Motivating Example

In order to understand the problem and its solution, let us suppose that a Geographical Information System (GIS) launches agents for finding, in the SW, information about several geographical objects in United States (see fig. 6). Suppose that the set Δ found by information agents is:

`Overlap`($West$, $Mount\ Elbert$)
`PartOf`($Mount\ Elbert$, $Colorado$)
`PartOf`($Colorado$, $West$)
`ProperPartOf`($Miami$, $Florida$)
`Overlaps`($West$, $Colorado$)
`Overlaps`($Basin\ of\ Platte\ River$, $Nebraska$)
`Discrete`($Colorado$, $Basin\ of\ Missouri\ River$)
`Overlaps`($East$, $Miami$)
`PartialOverlaps`($Basin\ of\ Missouri\ River$, $West$)
`ProperPart`($East$, $Colorado$)
`PartOf`($Miami$, $Florida$)
`ProperPartInverse`($Florida$, $Miami$)
`TangentialProperPart`($Mount\ Elbert$, $Great\ Plains$)
`Discrete`($West$, $Georgia$)
`Part`($East$, $Georgia$)

Note that several facts do not provide the most specific spatial relation that it might be expressed with RCC ontology. That is the case of the fact `Overlaps`

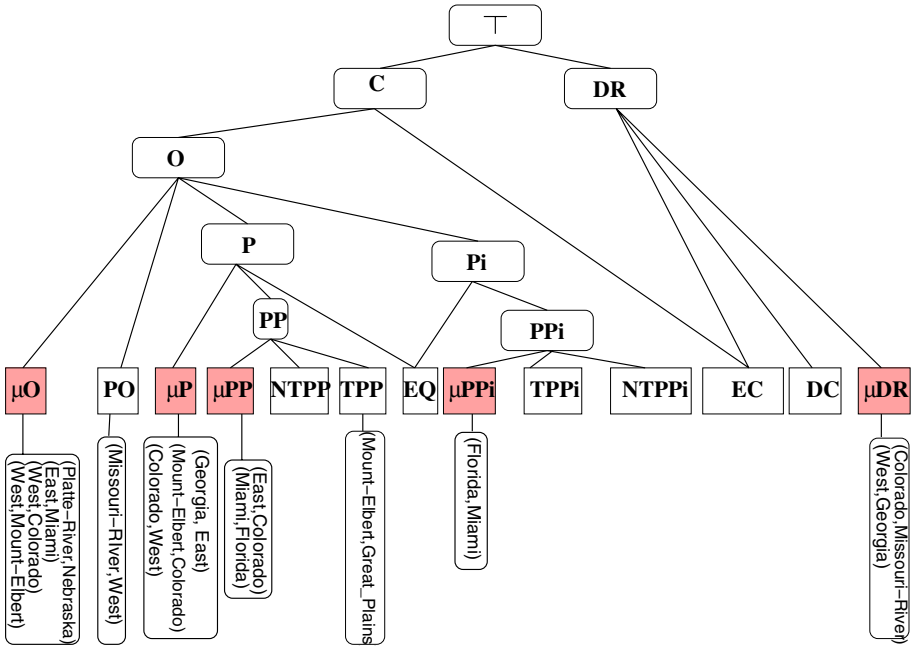


Fig. 7. Classification of data according to $E[\mu PP, \mu P, \mu PPi, \mu O, \mu DR]$

(*Basin of Platte River, Nebraska*). Both regions are overlapping, however there is no information about what level of overlapping relates these regions.

Since the GIS deals with concepts representing underspecified spatial relations such as `Overlaps, PartOf, ...`, it is hard to classify individual regions in an accurate way. They would be classify to work within a set of specific spatial-relations/concepts, a jointly exhaustive set of pairwise disjoint concepts to get the exhaustive intended classification.

The problem can be stated as follows: Given a set Δ of facts with respect to an ontology O, where most of specific information on some individuals can not entailed, to design a provisional robust extension of O to provisionally classify these concepts.

The extension of RCC for the running example will be a combination of some of the eight extensions. We are interested on finding an extension by undefinition of RCC that classifies the data and exhibits the highest entropy. According to data of the example, and th. 3, the selected extension has skeleton (see fig. 7): $E[\mu PP, \mu P, \mu PPi, \mu O, \mu DR]$. This l.c. extension reaches (by above theorem) maximum entropy, its value is 1.566. For example, $E[\mu P, \mu PPi, \mu O, \mu DR]$, shows entropy 1.326.

6 Closing Remarks

A formalization of data integration with unprecise information for the Semantic Web has been investigated. It presents a method to insert new concepts in an ontology with backward compatibility and preserving a weak form of completeness.

Although it is usual to study entropy for associating data to concepts in Ontology Learning, it is not usual to consider the *provability from ontology* like a factor, as we do. However, we think, that it will be a key issue in the SW. There are other approaches, but they deal with probabilistic objects. J. Calmet and A. Daemi also use entropy in order to revise or compare ontologies [10] [13]. This is based on the self taxonomy defined by the concepts but provability from specification is not regarded. Conditional entropy has already been considered in the similar task of Abductive Reasoning for learning qualitative relationships/concepts (usually in probabilistic terms, see e.g. [6]). The main difference between this approach and ours is that we work with probability mass distribution of *provable facts* from ontological specifications.

Finally, it should be noted that only some distributions of data will induce the user to decide an ontological insertion. Therefore, although once the distribution of data is determined, the method is fully formalized, the soundness of the extensions still depends on human decisions.

Future research lines are addressed, in the medium term, to implement the cognitive entropy into a representation of ontologies system as a tool to assist the extension of ontologies. In a long term, we will establish a theory, a formal theory in computational logic, to classify lattice categorical extensions.

References

1. Alonso-Jiménez, J.A., Borrego-Díaz, J., Chávez-González, A.M., Navarro-Marín, J.D.: A Methodology for the Computer–Aided Cleaning of Complex Knowledge Databases. In: 28th Conf. of IEEE Industrial Electronics Society IECON 2002, pp. 1806–1812 (2002)
2. Alonso-Jiménez, J., Borrego-Díaz, J., Chávez González, A.M., Gutiérrez-Naranjo, M.A., David Navarro-Marín, J.: Towards a Practical Argumentative Reasoning with Qualitative Spatial Databases. In: Chung, P.W.H., Hinde, C.J., Ali, M. (eds.) IEA/AIE 2003. LNCS (LNAI), vol. 2718, pp. 789–798. Springer, Heidelberg (2003)
3. Alonso-Jiménez, J.A., Borrego-Díaz, J., Chávez-González, A.M.: Ontology Cleaning by Mereotopological Reasoning. In: DEXA Workshop on Web Semantics WEBS 2004, pp. 132–137 (2004)
4. Alonso-Jiménez, J.A., Borrego-Díaz, J., Chávez-González, A.M., Martín-Mateos, F.J.: Foundational Challenges in Automated Data and Ontology Cleaning in the Semantic Web. IEEE Intelligent Systems 21(1), 42–52 (2006)
5. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (eds.): The Description Logic Handbook. Theory, Implementation and Applications. Cambridge University Press, Cambridge (2003)
6. Bhatnagar, R., Kanal, L.N.: Structural and Probabilistic Knowledge for Abductive Reasoning. IEEE Trans. on Pattern Analysis and Machine Intelligence 15(3), 233–245 (1993)
7. Bennett, B.: The Role of Definitions in Construction and Analysis of Formal Ontologies. In: Doherty, P., McCarthy, J., Williams, M. (eds.) Logical Formalization of Commonsense Reasoning (2003 AAAI Spring Symp.), pp. 27–35. AAAI Press, Menlo Park (2003)
8. Borrego-Díaz, J., Chávez-González, A.M.: Extension of Ontologies Assisted by Automated Reasoning Systems. In: Moreno Díaz, R., Pichler, F., Quesada Arencibia, A. (eds.) EUROCAST 2005. LNCS, vol. 3643, pp. 247–253. Springer, Heidelberg (2005)
9. Borrego-Díaz, J., Chávez-González, A.M.: A Formal Foundation for Knowledge Integration of Deficient Information in the Semantic Web. In: Moreno Díaz, R., Pichler, F., Quesada Arencibia, A. (eds.) EUROCAST 2007. LNCS, vol. 4739, pp. 305–312. Springer, Heidelberg (2007)
10. Calmet, J., Daemi, A.: From entropy to ontology. In: 4th. Int. Symp. From Agent Theory to Agent Implementation AT2AI-4 (2004)
11. Chávez-González, A.M.: Automated Mereotopological Reasoning for Ontology Debugging, Ph.D. Thesis, University of Seville (2005)
12. Cohn, A.G., Bennett, B., Gooday, J.M., Gotts, N.M.: Representing and Reasoning with Qualitative Spatial Relations about Regions. In: Stock, O. (ed.) Spatial and Temporal Reasoning, ch. 4. Kluwer, Dordrecth (1997)
13. Daemi, A., Calmet, J.: From Ontologies to Trust through Entropy. In: Proc. of the Int. Conf. on Advances in Intelligent Systems - Theory and Applications (2004)
14. Ganter, B., Wille, R.: Formal Concept Analysis, Mathematical Foundations. Springer, Berlin (1999)
15. Heflin, J.: Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment, Ph.D. Thesis, Univ. of Maryland, College Park (2001)
16. Kotz, S., Johnson, N.L. (eds.): Encyclopedia of Statistical Sciences, vol. 4, pp. 421–425. John Wiley and Sons, Chichester (1981)
17. Tsarkov, D., Riazanov, A., Bechhofer, S., Horrocks, I.: Using Vampire to Reason with OWL. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 471–485. Springer, Heidelberg (2004)