A Formal Foundation for Knowledge Integration of Defficent Information in the Semantic Web*

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, Spain
{jborrego, tchavez}@us.es

Abstract. Maintenance of logical robustness in Information Integration represents a major challenge in the envisioned Semantic Web. In this framework, it is previsible unprecise information (with respect to an ontology) is retrieved from some resources. The sound integration of such information is crucial to achieve logical soundness. We present a *data-driven* approach to classify that knowledge by means of the *cognitive entropy* of the possible robust ontology extensions and data.

1 Introduction

Knowledge Integration is a major issue in both Knowledge and Data Engineering (KDE) and Artificial Intelligence fields. Therefore, it has to be solved in one of the current projects where both fields come together, the Semantic Web (SW). In this framework, there are many situations where defficent information obstructs the use of trustworthy reasoning systems [1]. Even, it can suggest the revision of the intensional component of the Knowledge Database, namely the ontology.

In Ontological Engineering, an accurate classification of the objects is a main goal. It considers that the individuals involved in such data will remain well classified when they fall in the most specific classes of the concept taxonomy. A solution for that classification may be to introduce *provisional concepts* or *notions* for classifying individuals. Since the insertion of a notion of this kind is mainly *data-driven*, the notion is initially located in lower levels of the taxonomy of concepts. This is like that because very little is known about its definition, as well as how to subclassify their elements. In any way, we need to build a *robust extension* of the ontology to trustworthy work with the new concepts.

The subject of this paper is to present a method to insert concepts which have been induced by defficent information, into an ontology. Since data that suggests the revision is unprecise (up to certain degree), the user is not interested in to obtain a definition of the new concept. That is, one only aims to provisionally classify facts waiting for more precise information. This *data-driven* approach is investigated here. The method proposed for ontological insertion lies in extending

* Supported by project TIN2004-03884 of Spanish Ministry of Education and Science, cofinanced by FEDER founds.

the ontology to provisionally classify the individuals. The extension preserves main features of ontology source, and it can be considered as a robust extension (*lattice categoricity* [2], [3]). The main benefit of the method lies in the fact of it is fully formalized and it is semi-automated, assisted by automated reasoning systems.

There is other use case requiring this kind of ontological insertion. When the ontology engineer identifies a set of specific data (that is, facts on most specific concepts of the ontology) with the extension of a new concept, since he/she has not a formal definition of the concept, the place of the ontology in which it has to be inserted is not specified. This fact typically occurs in settings where ontology engineer detects language poorness in the ontology. This point of view gives rise to *user-driven* approaches that we have formalized in [2].

The remainder of the paper is organized as follows. In the next section we present a simple example to illustrate the problem. In section 3 the formalization of *robust ontology extension* is outlined. A kind of extension is the extension by *insertion of an undefinition* (sect. 4). The method is applied to solve the problem of the example. Finally, some final remarks about the approach are given in section 5.

2 A Motivating Example

In order to understand the problem as well as 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. Suppose that the data set Δ found by the agents is:

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

Note that several facts do not provide the most specific spatial relation that it might be expressed by the ontology. That is the case of the fact `Overlaps`($Basin\ of\ Platte\ River$, $Nebraska$). Both regions are overlapping, however there is no information about what level of overlapping relates these regions. Since the GIS deals with concepts representing underspecify spatial relations such as `Overlaps`, or `PartOf, ...`, it is hard to classify individual regions in an accurate way. They would be classified to work within a set of specific spatial-relations/concepts, a jointly exhaustive set of pairwise disjoint (JEPD) concepts to get the exhaustive intended classification.

The problem can be stated as: *Given a set Δ of facts with respect to an ontology O, where the most specific information on some individuals can not*

entailed, to design an provisional robust extension of O to provisionally classify these concepts.

The ontology for the running example is *Region Connection Calculus* (RCC), designed for (mereotopological) Qualitative Spatial Reasoning (QSR)[7]. The relations of RCC are used in both GIS and spatial databases [9]. More information on RCC can be found in [7].

The jointly exhaustive and pairwise disjoint (JEPD) set of binary relations depicted in figure 1 (right-bottom) is denoted by RCC8. The complexity of RCC8 to solve Constraints Satisfaction Problems (CSP) has been deeply studied by J.R. Renz and B. Nebel [11]. Other calculus to take into account is RCC5. It is based on the set $\{DR, PO, PP, PPi, EQ\}$. It is less precise but more manageable than RCC8. Therefore, RCC8 represents the most specific spatial relationships in RCC. The remaining relations of RCC can be regarded as *unprecise*. The special interest of authors in this ontology lies in its role as meta-ontology for visual cleaning [4].

3 Extending Ontologies with Backward Compatibility

The study of ontology revision covers a very broad spectrum of theories and techniques. It encompasses logical and engineering methods, including theories from the fields of KDE and Knowledge Representation and Reasoning. A typical case of the need of ontology revision occurs when ontology engineer detects that new data are not accurately specified/classified with respect to the current information. A first solution may be to insert some provisional concept(s) (*notion(s)*) classifying that unprecise information and to expect that new conditions will allow us to refine information for obtaining a right classification. Actually, it involves an *extension* of ontology. For instance the existence of ground literals (instances) of an abstract concept (i.e. they are non-direct instances) can be a methodological problem in ontology design. Thus, it is better to consider a new concept that provisionally represents a notion. As we have already commented, such a concept will not have subclasses; thus, it will be located at the ground level of the taxonomy of concepts.

It is necessary to point out that ontology evolution must obey basic accepted principles such as *backward compatibility*, while it is possible. In [3] a weak form of backward compatibility, useful for the aim of this paper, is introduced. In fact, it has been used for other kind of extensions in [2].

Considering an ontology as a pair (T, E) where T is the axioms set and E is a equational characterization of the intended lattice of concepts, (the *skeleton*), we say that an ontology is *lattice categorical* (l.c.) if the lattice determined by T, and denoted by $L(T, \mathcal{C})$, is unique. $\mathcal{C}$ denotes the set of concepts of T. The theory RCC is an example of l.c. theory. The only possible lattice structure exhibited by the models of RCC is that of figure 1, and a skeleton E is computed in [3].

In [3] and [2] we replaced *completeness* by *lattice categoricity* to facilitate the design of feasible methods for extending ontologies with logical soundness. The extension is defined as follows. Given $(T_1, E_1), (T_2, E_2)$, two ontologies of this

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

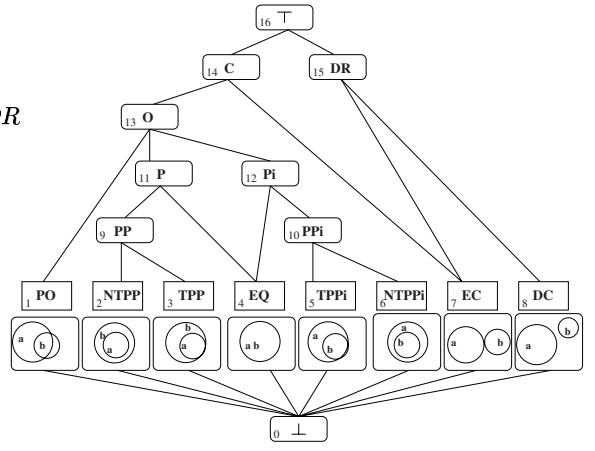

Fig. 1. The skeleton E (left) for the lattice of RCC (right)

kind with respect to the sets of concepts $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively, we say that (T_2, E_2) is a **lattice categorical extension** of (T_1, E_1) if $L(T_1, \mathcal{C}_1) \subseteq L(T_2, \mathcal{C}_2)$ and $L(T_2, \mathcal{C}_2) \models E_1$.

3.1 Cognitive Support

Once the notion of *lattice categorical extension* has been introduced, some functions for selecting the best l.c. extension have to be designed.

Suppose that $\Delta = \{h_1, \ldots h_n\}$ is a set of facts on concepts in $\mathcal{C}$. The user aims to classify some of individuals appearing in Δ by means of specific concepts. We can suppose, to simplify the notation, that every fact explicit in T belongs to Δ.

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

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

where $U(\Delta) := \{\boldsymbol{a} \ : \ \text{exists } C \in \mathcal{C} \ [C(\boldsymbol{a}) \in \Delta]\}$ is the universe determined by Δ. That is, the cognitive support estimates the number of facts on the concept C that T entails (normalized by the size of $U(\Delta)$). The computation is trivial for lattice categorical theories, [2]: $sup^L_{T,\Delta}(C) = \dfrac{|C|^{\Delta}_T}{|U(\Delta)|}$ where $|C|^{\Delta} := |\{\boldsymbol{a} \ : C(\boldsymbol{a}) \in \Delta\}|$ and $|C|^{\Delta}_T := |\{\boldsymbol{a} \in U(\Delta) \ : \ T \cup \Delta \models C(\boldsymbol{a})\}|$.

Suppose now that Δ is compounded by facts on atoms of the lattice of concepts (that is, about the most specific concepts). In this case, since $\mathcal{J} = \{C_1, \ldots, C_n\}$ is a JEPD, $sup_{T,\Delta}(.)$ is a probability measure. In general, if $\mathcal{J}$ is a JEPD set of concepts in L, and Δ is compounded by instances on concepts falling in the cone of some element of $\mathcal{J}$, then $sup_{T,\Delta}(.)$ is a probability measure on $\mathcal{J}$.

Finally, the **cognitive entropy** of $\mathcal{J}$ is

$$CH(\mathcal{J}) = -\sum_{C \in \mathcal{J}} sup_{T,\Delta}(C) \log sup_{T,\Delta}(C)$$

This entropy is the key parameter used in the *user-driven* approach [2].

4 Data-Driven Ontology Revision for Defficent Data

A defficent classification of data induces the insertion of subconcepts for refining the classification of individuals which initially were misclassified. As it is already commented, 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 unprecision* 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 more specific information about $\mathbf{a}$". Formally,

Definition 1. *Let (T, E_0) be an ontology 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 $\mu C <^{L'} C' <^{L'} C$.*

Note that, in above conditions, $\mathcal{J}_L[\mu C] := \mathcal{J}_L \cup \{\mu C\}$ is a JEPD set for L' (see fig. 2, left). This requirement represents, in fact, that we have not any additional information about μC. For example, in figure 2 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 $\mathbf{a}$ entailed by T. It is easy to see that

Proposition 1. *Whatever two extensions of T by undefinition at C have equivalent lattice skeletons modulo completion.*

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

Corollary 1. *$E[\{\mu C : C \in \mathcal{C} \wedge T \rightsquigarrow_w C\}]$ is unique (modulo database completion axioms).*

For the example, this kind of extensions for RCC have to be investigate.

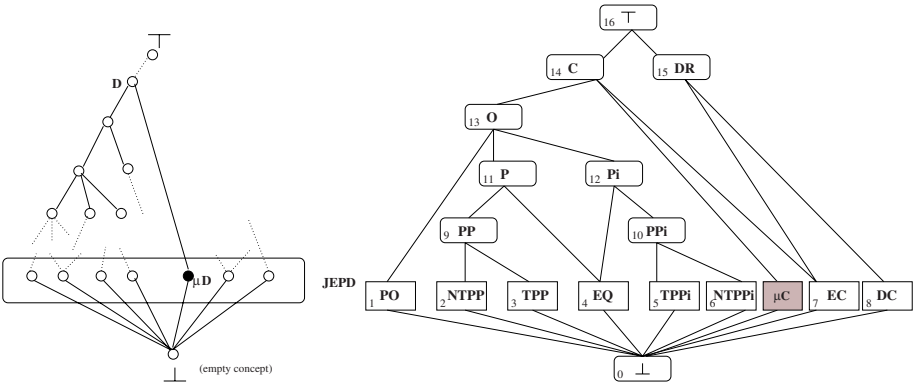

Fig. 2. The ontology admits an undefinition in the concept C (*connection*) (right)

4.1 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.

Theorem 1. *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 [6]. The lattices of the extensions are detailed at [3]. For example, the lattice depicted in fig. 1 (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 $\Delta = \{h_1, \dots h_n\}$ is the set of facts. Assume that the user believes that the set of misclassified elements is $I = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_k\} \subseteq U(\Delta)$ (according with user's ontology). In this case, the problem is not due to a new concept, because he/she 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 $\{C^i \ : \ i = 1, \dots k\}$ classifies every element of $U(\Delta)$ with a concept of the JEPD set $\mathcal{J}_{T'} := \mathcal{J} \cup \{\mu C^1, \dots, \mu C^k\}$. Note also, that if we do not require C^i is the most specific one, the extension is not unique.

Definition 2. *Let T' be an extension by undefinition of T as defined in 1. 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 $\mathcal{J}_{T'}$.

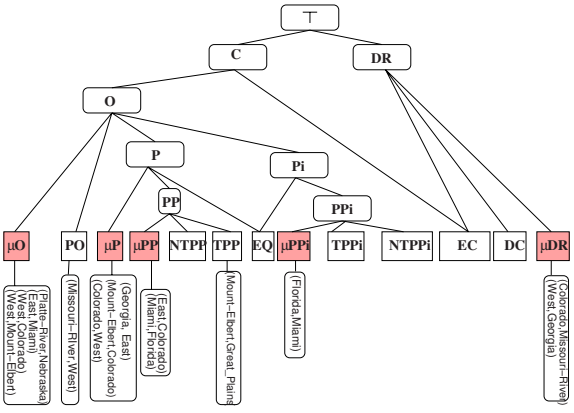

Fig. 3. Classification of data according to $E[\mu PP, \mu P, \mu PPi, \mu O, \mu DR]$

Note that this computation is equivalent to consider the support with respect to the theory $T' + \{\mu C(\boldsymbol{a}) \ : \ T \cup \Delta \models_\mu C(\boldsymbol{a})\}$. To simplify the notation, we finally consider throughout that T' is that theory.

Theorem 2. *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 w.r.t. T. 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.

The extension of RCC for the running example will be a combination of some of the eight extensions. We are interested to find an extension by undefinition of RCC that classifies the data and exhibits higher entropy. According to data and th. 2, the selected extension has skeleton (fig. 3): $E[\mu PP, \mu P, \mu PPi, \mu O, \mu DR]$. This l.c. extension has maximum entropy (by above theorem), 1.566. For example, $E[\mu P, \mu PPi, \mu O, \mu DR]$, shows entropy 1.326.

5 Closing Remarks and Related Work

A formalization of integration of unprecise data with respect to an ontology has been investigated, as well as a method to insert new concepts in an ontology with backward compatibility and preserving a weak form of completeness.

Note that reasoning services -that we need in order to build the extension with maximum entropy- can be non-decidable for first order theories. However, it is feasible for ontologies expressed in several (decidable) Description Logics, or considering the skeleton (a DL theory) as basis theory.

In [2] we formalize the insertion of a concept (possibly in a upper level) that will remain well defined once the appropriate extension is selected. In that case, the computation of the (conditional) entropies is easier than the entropies defined on this paper. The approach of this paper is different because it is not necessary user decision on new concepts. Possibly, both procedures should be combined in several situations like document enrichment tasks [10].

Entropy is usually considered for associating data and concepts of an ontology (see e.g. [5]). J. Calmet and A. Daemi also use entropy for revising or comparing ontologies [8], based on the concept taxonomy. However it is unusual to consider the *provability* as a parameter.

Finally, note that, although the method is fully formalized, the cognitive soundness of the extensions will depend of the human decision. Moreover, the iteration of the method can produce the existence of many provisional concepts without intentional component. It may be unadvisable in some cases.

References

1. Alonso-Jiménez, J.A., Borrego-Dáaz, J., Chávez-González, A.M., Martńn-Mateos, F.J.: Foundational Challenges in Automated Data and Ontology Cleaning in the Semantic Web. IEEE Intelligent Systems 21(1), 42–52 (2006)
2. Borrego-Díaz, J., Chávez-González, A.M.: Controlling Ontology Extension by Uncertain Concepts Through Cognitive Entropy. Uncertain Reasoning for the Semantic Web, URSW 2005, CEUR 173, 56–66 (2005)
3. 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. 253–257. Springer, Heidelberg (2005)
4. Borrego-Díaz, J., Chávez-González, A.M.: Visual Ontology Cleaning: Cognitive Principles and Applicability. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 317–331. Springer, Heidelberg (2006)
5. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data Driven Ontology Evaluation, Int. Conf. Lang. Resources and Evaluation (2004), `http://eprints.ecs.soton.ac.uk/archive/00009062/01/BrewsterLREC-final.pdf`
6. Chávez-González, A.: Mereotopological Automated Reasoning for Ontology Cleaning, Ph.D. Thesis, University of Seville (2005)
7. 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)
8. 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)
9. Grohe, M., Segoufin, L.: On First-Order Topological Queries. ACM Trans. Comput. Log. 3(3), 336–358 (2002)
10. Motta, E., Buckingham, S., Domingue, J.: Ontology-driven document enrichment: principles, tools and applications. Int. J. Human-Computer Studies 52(6), 1071–1109 (2000)
11. Renz, J.R., Nebel, B.: On the Complexity of Qualitative Spatial Reasoning: A Maximal Tractable Fragment of the Region Connection Calculus. Artificial Intelligence 108, 69–128 (1999)