KI - Künstliche Intelligenz (2020) 34:399–403
https://doi.org/10.1007/s13218-020-00673-8

DISSERTATION AND HABILITATION ABSTRACTS

Constructing and Extending Description Logic Ontologies using Methods of Formal Concept Analysis

A Dissertation Summary

Francesco Kriegel[1]

Received: 24 January 2020 / Accepted: 6 June 2020 / Published online: 11 July 2020

Abstract
My thesis describes how methods from Formal Concept Analysis can be used for constructing and extending description logic ontologies. In particular, it is shown how concept inclusions can be axiomatized from data in the description logics $\mathcal{EL}$, $\mathcal{M}$, Horn-$\mathcal{M}$, and Prob-$\mathcal{EL}$. All proposed methods are not only sound but also complete, i.e., the result not only consists of valid concept inclusions but also entails each valid concept inclusion. Moreover, a lattice-theoretic view on the description logic $\mathcal{EL}$ is provided. For instance, it is shown how upper and lower neighbors of $\mathcal{EL}$ concept descriptions can be computed and further it is proven that the set of $\mathcal{EL}$ concept descriptions forms a graded lattice with a non-elementary rank function.

Keywords Description logic · Formal concept analysis · Axiomatization · Concept inclusion

1 Introduction

Description Logic (abbrv. DL) [1] belongs to the field of knowledge representation and reasoning. DL researchers have developed a large family of logic-based languages, so-called *description logics* (abbrv. DLs). These logics allow their users to explicitly represent knowledge as *ontologies*, which are finite sets of (human- and machine-readable) axioms, and provide them with automated inference services to derive implicit knowledge. The landscape of decidability and computational complexity of common reasoning tasks for various description logics has been explored in large parts: there is always a trade-off between expressibility and reasoning costs. It is therefore not surprising that DLs are nowadays applied in a large variety of domains [1]: agriculture, astronomy, biology, defense, education, energy management, geography, geoscience, medicine, oceanography, and oil and gas. Furthermore, the most notable success of DLs is that these constitute the logical underpinning of the *Web Ontology Language* (abbrv. OWL) [5] in the *Semantic Web*.

Formal Concept Analysis (abbrv. FCA) [3] is subfield of lattice theory that allows to analyze data-sets that can be represented as formal contexts. Put simply, such a formal context binds a set of objects to a set of attributes by specifying which objects have which attributes. There are two major techniques that can be applied in various ways for purposes of conceptual clustering, data mining, machine learning, knowledge management, knowledge visualization, etc. On the one hand, it is possible to describe the hierarchical structure of such a data-set in form of a formal concept lattice [3]. On the other hand, the theory of implications (dependencies between attributes) valid in a given formal context can be axiomatized in a sound and complete manner by the so-called canonical base [4], which furthermore contains a minimal number of implications w.r.t. the properties of soundness and completeness.

In spite of the different notions used in FCA and in DL, there has been a very fruitful interaction between these two research areas. My thesis [6] continues this line of research and, more specifically, it describes how methods from FCA can be used to support the automatic construction and extension of DL ontologies from data.

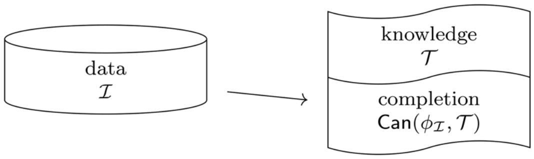
✉ Francesco Kriegel
francesco.kriegel@tu-dresden.de

[1] Institute of Theoretical Computer Science, Technische Universität Dresden, Dresden, Germany

2 Axiomatization of $\mathcal{EL}$ Concept Inclusions

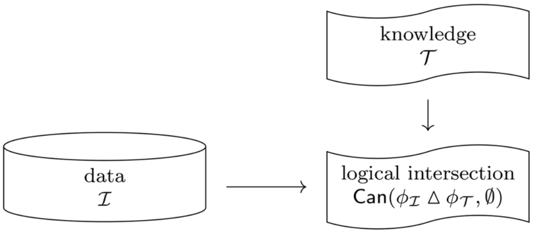
The description logic $\mathcal{EL}$ allows for tractable reasoning in polynomial time and features concept descriptions for intensionally describing collections of objects. A concept inclusion is an implication between two concept descriptions and such terminological axioms are used for describing the schema of the domain of interest. However, it might be a tedious task to formulate such axioms by hand. My thesis is concerned with the (unsupervised) axiomatization of concept inclusions under different assumptions on the input. In the following, according use cases are described.

Completions. The first use case is concerned with situations where there already exist concept inclusions describing the domain of interest and where data on the domain of interest is available. The existing concept inclusions might either have been manually formulated by a knowledge engineer or have been generated using other axiomatization techniques. Furthermore, it is required that these concept inclusions are retained, i.e., we construct a completion of these with respect to the data. More specifically, given a TBox $\mathcal{T}$ and an interpretation $\mathcal{I}$ that is a model of $\mathcal{T}$, a *completion* of $\mathcal{T}$ w.r.t. $\mathcal{I}$ (or, a *concept inclusion base* of $\mathcal{I}$ *relative* to $\mathcal{T}$) is a TBox $\mathcal{S}$ such that its union with $\mathcal{T}$ is both sound and complete for $\mathcal{I}$, i.e., $\mathcal{I}$ is a model of $\mathcal{S}$ and $\mathcal{S} \cup \mathcal{T}$ entails each concept inclusion that is valid in $\mathcal{I}$.



The problem of computing such completions can be reduced to a corresponding problem in Formal Concept Analysis, namely computing an implication base relative to an existing implication set. Algorithmic solutions for the latter exist, and my thesis describes a procedure that can compute such implication bases in a highly parallel manner where the necessary computation time is almost inverse linear proportional to the number of available CPU cores. Moreover, such completions can always be computed in exponential time and there exist interpretations for which no completion can be encoded in polynomial space.

Logical Intersections. Assume again a situation where we have an interpretation $\mathcal{I}$ and some TBox $\mathcal{T}$, but this time $\mathcal{I}$ is not a model of $\mathcal{T}$ and we do not have a preference between both. For suitably axiomatizing concept inclusions, a solution is to characterize the *logical intersection*, i.e., to find a base for the concept inclusions that are both valid in the interpretation and are entailed by the TBox.


knowledge
$\mathcal{T}$
data
$\mathcal{I}$
logical intersection
$\mathsf{Can}(\phi_{\mathcal{I}} \vartriangle \phi_{\mathcal{T}}, \emptyset)$

In order to do so, we generalize the notions of interpretations and TBoxes to a common representation: we show that both induce a so-called *closure operator*, which is a monotone, extensive, idempotent mapping on concept descriptions. The benefit is that the set of closure operators forms a lattice, i.e., it is an ordered set and for two closure operators there always exists a supremum and an infimum. These two operations on closure operators correspond to operations on the underlying data inducing the closure operator. Furthermore, we can define the notion of validity of a concept inclusion for a closure operator such that it coincides with the usual notion of validity in an interpretation and of entailment by a TBox, respectively.

Distel [2] showed that the mapping from subsets of the interpretation domain to their *model-based most specific concept descriptions* is the adjoint of the interpretation function, i.e., these form a Galois connection. It is then an immediate consequence that each interpretation $\mathcal{I}$ induces a closure operator $\phi_{\mathcal{I}}$, namely the composition of the interpretation function and the model-based most specific concept description mapping.

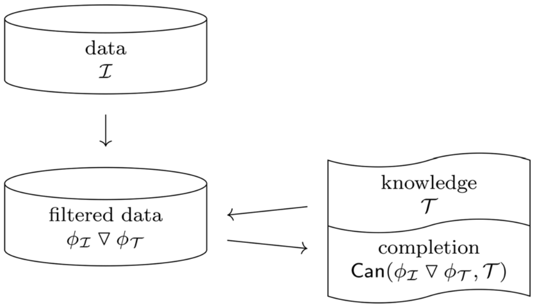
My thesis shows that each TBox $\mathcal{T}$ induces a closure operator $\phi_{\mathcal{T}}$ as well. It is obtained as the function that maps a concept description to its *most specific consequence* with respect to $\mathcal{T}$. Put simply, such most specific consequences can be computed by saturating a concept description with the concept inclusions in $\mathcal{T}$.

Now the concept inclusions valid for the infimum $\phi_{\mathcal{I}} \Delta \phi_{\mathcal{T}}$ are exactly the concept inclusions that are both valid in $\mathcal{I}$ and entailed by $\mathcal{T}$. It follows that we can characterize the logical intersection by a concept inclusion base for the closure operator $\phi_{\mathcal{I}} \Delta \phi_{\mathcal{T}}$. For computing such bases, my thesis demonstrates how the above completions can be constructed for the more general case where the data is not just an interpretation but can be described by a closure operator instead. More specifically, a so-called *canonical base* $\mathsf{Can}(\phi, \mathcal{T})$ can be computed, which is a completion of some TBox $\mathcal{T}$ w.r.t. a closure operator ϕ and has minimal cardinality among all completions of $\mathcal{T}$ w.r.t. ϕ.

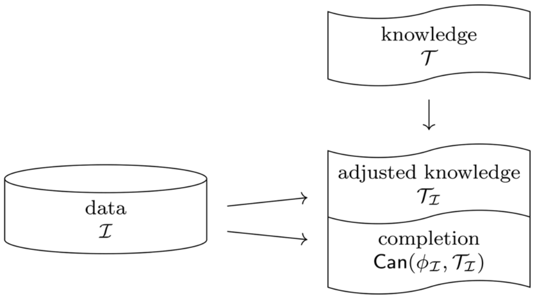
However, it might sometimes be necessary to restrict the role depth of the concept inclusions to be axiomatized. This due to the fact that logical intersections need not be finitely representable. For instance, consider the TBoxes $\mathcal{T} := \{A \sqsubseteq B_1\}$ and $\mathcal{U} := \{A \sqsubseteq B_2\}$; then for each number $n \in \mathbb{N}$, both entail the concept inclusion

$\exists r^n.(A \sqcap B_1) \sqcap \exists r^n.(A \sqcap B_2) \sqsubseteq \exists r^n.(A \sqcap B_1 \sqcap B_2)$. Obviously, there cannot exist a TBox that entails all above concept inclusions, since TBoxes must be finite.

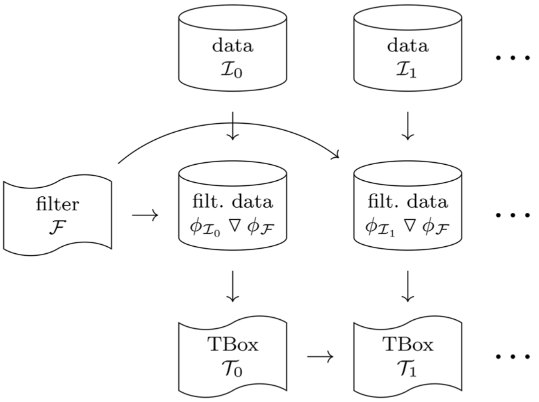
Filtering then Completing. When assuming that the TBox is more trustworthy than the interpretation, it is necessary to filter out incompatible parts of the interpretation. We utilize the supremum operation for this purpose. In particular, the supremum of $\phi_{\mathcal{I}}$ and $\phi_{\mathcal{T}}$ describes a *filtering* of the interpretation $\mathcal{I}$ with respect to $\mathcal{T}$, i.e., it only consists of the part of $\mathcal{I}$ that is a model of $\mathcal{T}$. Note that the filtering is not defined on the object level but on the extensional level instead, which means that it does not consist of objects of $\mathcal{I}$ but of sets of objects. Finally, the axiomatization of the input $\mathcal{I}$ and $\mathcal{T}$ is then obtained as the union of $\mathcal{T}$ and the completion of $\mathcal{T}$ with respect to the supremum $\phi_{\mathcal{I}} \nabla \phi_{\mathcal{T}}$.


data $\mathcal{I}$
filtered data $\phi_{\mathcal{I}} \nabla \phi_{\mathcal{T}}$
knowledge $\mathcal{T}$
completion $\mathsf{Can}(\phi_{\mathcal{I}} \nabla \phi_{\mathcal{T}}, \mathcal{T})$

Adjusting then Completing. We are now concerned with the last situation where the interpretation is preferred over the TBox. The easiest solution is, of course, to simply compute a concept inclusion base for the interpretation—however, it is then not clear how to track differences between the existing concept inclusions and the new ones in the base. Alternatively, conclusions in the existing concept inclusions can be adjusted: for each existing concept inclusion $C \sqsubseteq D$, replace D with the most specific concept description E such that $C \sqsubseteq E$ is both valid in $\mathcal{I}$ and entailed by $\mathcal{T}$, i.e., E is obtained from C by applying the infimum $\phi_{\mathcal{I}} \Delta \phi_{\mathcal{T}}$. That way, we first adjust the existing concept inclusions to the new interpretation, and afterwards we compute the completion of this adjustment w.r.t. the new interpretation.


knowledge $\mathcal{T}$
adjusted knowledge $\mathcal{T}_{\mathcal{I}}$
completion $\mathsf{Can}(\phi_{\mathcal{I}}, \mathcal{T}_{\mathcal{I}})$
data $\mathcal{I}$

Incremental Axiomatization from Streams of Data. Eventually, we want to put emphasis on the fact that all of the above techniques can be stacked and iterated. For instance, we might have a situation where new observations are available on a regular basis, i.e., a stream $(\mathcal{I}_n \mid n \in \mathbb{N})$ of interpretations is available. We further have a hand-crafted TBox $\mathcal{F}$ for filtering each of the interpretations: on the one hand, this TBox might describe a filter on *interesting* data and, on the other hand, this TBox might describe a filter on *valid* data; the concrete role of it does not matter for our purposes.


data $\mathcal{I}_0$
data $\mathcal{I}_1$
...
filter $\mathcal{F}$
filt. data $\phi_{\mathcal{I}_0} \nabla \phi_{\mathcal{F}}$
filt. data $\phi_{\mathcal{I}_1} \nabla \phi_{\mathcal{F}}$
...
TBox $\mathcal{T}_0$
TBox $\mathcal{T}_1$
...

Our goal is now as follows. For each time point n, a TBox $\mathcal{T}_n$ is to be constructed that is sound and complete for the concept inclusions that are valid in the filterings of $\mathcal{I}_0, \ldots, \mathcal{I}_n$ w.r.t. $\mathcal{F}$. Of course, we can initially construct the concept inclusion base for the filtering of the first interpretation $\mathcal{I}_0$ w.r.t. $\mathcal{F}$, yielding a TBox $\mathcal{T}_0$. For each later time point $n > 0$, the TBox $\mathcal{T}_n$ is obtained as the logical intersection of $\mathcal{T}_{n-1}$ and the filtering of $\mathcal{I}_n$ w.r.t. $\mathcal{F}$.

3 More Expressive Description Logics

The task of axiomatizing concept inclusions is also investigated for more expressive description logics. As a first target language the description logic $\mathcal{M}$ is considered. It is not a Boolean-complete logic, since it does not allow disjunctions and negations—this avoids overfitting of the resulting concept inclusions. However, reasoning complexity for $\mathcal{M}$ is much higher than for $\mathcal{EL}$: it jumps from **P**-completeness to **EXP**-completeness.[1] For this reason, the Horn fragment of $\mathcal{M}$, denoted as Horn-$\mathcal{M}$, is considered as a target language as well. Put simply, the Horn fragment is the largest fragment that can be translated into function-free logic programs (Datalog). The restriction to the Horn fragment lowers

[1] Note that this is a conjecture. In fact, it is proven only for the sublogic $\mathcal{M}^-$ without existential self-restrictions $\exists r.\mathsf{Self}$.

expressivity, but with the advantage that reasoning complexity decreases. In particular, the instance problem is **coNP**-complete for $\mathcal{M}$ and **P**-complete for Horn-$\mathcal{M}$ (both w.r.t. data complexity).[1] It hence makes sense to use Horn-$\mathcal{M}$ TBoxes in ontology-based data access applications.

As a further, more expressive description logic a probabilistic extension Prob-$\mathcal{EL}$ of $\mathcal{EL}$ is considered. It extends $\mathcal{EL}$ by the possibility to probabilistically quantify a concept description. Again, reasoning is more expensive than in $\mathcal{EL}$: the subsumption relation is **EXP**-complete. As it turns out, concept inclusion bases for probabilistic interpretations can be computed in exponential time as well, i.e., the increase in expressive power does not result in higher computational complexity of the axiomatization task.

For all above mentioned description logics that are more expressive than $\mathcal{EL}$, it is currently unclear whether most specific consequences w.r.t. a TBox exist and, if so, how these can be computed. As soon as these questions are solved with an affirmative answer, similar approaches as for $\mathcal{EL}$ can be utilized for combining knowledge from TBoxes and interpretations in a learning setting. For now, it is only possible to axiomatize the concept inclusions valid in a given interpretation.

4 A Lattice-Theoretic View on $\mathcal{EL}$

The set of $\mathcal{EL}$ concept descriptions ordered by subsumption forms a lattice in which conjunction is the infimum operation and the least common subsumer mapping is the supremum operation. In my thesis, I have investigated this lattice in more detail. It was shown that the lattice is *distributive*. Furthermore, relative pseudo-complements always exist and can be computed in polynomial time, which makes the lattice a *residuated* one.

The neighborhood relation induced by the subsumption relation contains pairs of concept descriptions where the first is strictly subsumed by the second and such that there does not exist any concept description strictly between both. A natural question is whether the transitive closure of that neighborhood relation equals the strict part of the subsumption relation, i.e., whether the subsumption relation is *neighborhood generated*. If it is, then we might walk along the neighborhood relation when searching for a concept description with specific properties without the chance to miss any interesting candidate. For empty or cycle-restricted TBoxes, the subsumption relation is indeed neighborhood generated and my thesis describes how all upper and all lower neighbors of a given concept description can be enumerated. For general TBoxes or extensions of $\mathcal{EL}$ with greatest fixed-point semantics, the subsumption relation is not neighborhood generated and suitable counterexamples are provided.

Eventually, the neighborhood relation can be utilized to define a metric (a distance function) on the set of $\mathcal{EL}$ concept descriptions. The reason is that $\mathcal{EL}$ is *of locally finite length*, i.e., all chains between two comparable concept descriptions are finite, and further that $\mathcal{EL}$ satisfies the *Jordan-Dedekind chain condition*, i.e., all maximal chains between two comparable concept descriptions have the same length. We can then simply define the distance between two comparable concept descriptions as the length of *some* maximal chain between them—clearly, such a maximal chain must be a chain of neighbors. For measuring distances between arbitrary concept descriptions, we choose the distance between the corresponding infimum and supremum. In the undirected graph where $\mathcal{EL}$ concept descriptions are the nodes and two nodes are connected if they are neighbors, that distance between two concept descriptions is the length of a shortest path between both. As it turns out, the above distance function is not an elementary function. The distance between $\top$ and $\exists r^n.(A \sqcap B \sqcap C)$ is asymptotically bounded above and below by

$$\underbrace{2^{2^{\cdot^{\cdot^{\cdot^{2^{2^{3}}}}}}}}_{n \text{ times}}.$$

5 Conclusion

My thesis describes how methods from Formal Concept Analysis can be utilized for the task of constructing and extending description logic ontologies. In particular, for the tractable description logic $\mathcal{EL}$ existing knowledge can be reused when axiomatizing concept inclusions from newly observed data. For the more expressive description logics $\mathcal{M}$, Horn-$\mathcal{M}$, and Prob-$\mathcal{EL}$ methods for mining concept inclusions from observed data are developed. Currently, existing knowledge cannot be incorporated for these logics, since it remains an open question whether most specific consequences always exist and, if so, how to compute these. All proposed methods are not only sound, but also complete.

Acknowledgements Partially supported by the DFG in the CRC 912 (HAEC). Open Access funding provided by Projekt DEAL.

References

1. Baader F, Horrocks I, Lutz C, Sattler U (2017) An introduction to description logic. Cambridge University Press, Cambridge
2. Distel F (2011) Learning description logic knowledge bases from data using methods from formal concept analysis. Doctoral thesis. Dresden, Germany: Technische Universität Dresden
3. Ganter B, Wille R (1999) Formal concept analysis: mathematical foundations, 1st edn. Springer, Berlin
4. Guigues J-L, Duquenne V (1986) Famille minimale d'implications informatives résultant d'un tableau de données binaires. Mathématiques et Sciences Humaines 95:5–18
5. Hitzler P, Krötzsch M, Rudolph S (2010) Foundations of semantic web technologies. Chapman and Hall/CRC Press, Boca Raton
6. Kriegel F (2019) Constructing and extending description logic ontologies using methods of formal concept analysis. Doctoral Thesis. Dresden, Germany: Technische Universität Dresden