Semantic Web Basics in Logical Consideration [*]

Denis Ponomaryov
Institute of Informatics Systems, Novosibirsk, Russia
ponom@iis.nsk.su

Abstract: In this paper we study the operation of incremental data extraction from declarative knowledge bases and the problem of decomposability of data in such KBs, by viewing them as first-order logical theories. The operation of incremental data extraction is closely connected with query reformulation in information retrieval and data integration, while the decomposability problem is important in the scope of modularization and distributed processing of knowledge.

1 Introduction

At present, there is a significant interest to methods and tools of declarative knowledge representation, which is in particular connected with the wide-spread notion of formal ontology and the Semantic Web paradigm. The outcome of this is development and application of new descriptive languages, as well as reasoning or deductive systems. Each of a newly appeared languages corresponds to some subset of First Order Logic (FOL). However, judging from practice, ontological engineers have realized the need of the full FOL to work with the information they encounter. In this work, we consider declarative knowledge bases as first-order (elementary) theories, i.e. sets of closed formulas of the predicate calculus. We distinguish two scenarios of their use in practical applications, namely, for search in large data repositories and for integration of heterogeneous data sources. In spite these two scenarios have much in common, they are approached differently in the field of information management. The first one is mostly considered in conjunction with the Internet search problem, however there are many other actual applications [1, 2, 8]. In this scenario, queries are formulated in terms represented by a declarative description of the subject domain of interest. Usually, there is an initial query, which is to be reformulated or strengthened/relaxed according to the relevance of search results or according to other alternative criteria. All query transformations are performed basing on the data in the given formal description of the subject domain. The second scenario is best reflected in the present research on Peer-to-Peer systems [5, 6]. The purpose of declarative descriptions in this case is to represent a conceptual schema of a data source, i.e. to describe the knowledge it provides access to. A query built in terms of one data source is reformulated in terms of another one to provide data exchange and distributed information search.

[*] This work is supported by project COMO of Russian Foundation for Basic Research (project no. 05-01-04003-NNIO_a) and Deutsche Forschungsgemeinschaft (GZ: 436 RUS 113/829/0-1).

Thus, it is necessary to find a correspondence or to build a mapping between two declarative descriptions. In most of cases, there is no need to build a correspondence between two descriptions as a whole. Instead, some part of a description containing the key query terms is needed to be mapped onto another one. How this part is chosen, greatly influences the ”precision of mapping”, which clearly, has lots of consequences.

In both scenarios, such declarative descriptions are themselves used like data sources, but the information extracted from them is mostly not sets of constants, but sets of expressions or formulas, which are treated as facts in solving a given task. Proceeding from these two scenarios, we define in Sect. 2 the operation of incremental data extraction from declarative knowledge bases. Next, we formulate the problem of decomposability by the example of this operation and propose a solution of this problem. Section 4 contains some final remarks and outlines the content of the planned talk.

2 The Operation of Incremental Data Extraction

In our work we consider declarative knowledge bases as elementary theories, i.e. consistent sets of sentences in the FOL language. By incremental data extraction from a knowledge base, we mean here a sequential selection of sentences according to some predefined strategy. We consider this to be the most general view at the use cases mentioned in the introduction. Indeed, in the first scenario, a typical algorithm starts, for example, from some set of constant symbols as an input. Then it uses relations defined on these symbols to extract new constants, then uses formulas expressing relation properties and so on. All the extracted information is used in the search. Sometimes, a choice of some set of formulas may be rejected for the reason of poor relevance of search results, and another set can be chosen instead. In most of cases, it is hard to predict an effect of usage of this or that information in a concrete search task. At least it is possible to choose between different ”types” of formulas, e.g. ground, restrictive or non-restrictive clauses. An excellent illustration of this kind of strategies can be found in papers devoted to algorithms of database schema matching [4, 3]. The operation of incremental data extraction can be based on quite different strategies, but we argue that the very basic and common strategy of this operation can be considered from a purely syntactical point of view. Further we formally define the operation of incremental data extraction.

Let $\mathcal{T}$ be an elementary theory in a signature Σ. That is, $\mathcal{T}$ consists of sentences that contain symbols only from Σ. We may assume that signatures consist only of predicate symbols, as functions of arbitrary form can be substituted by corresponding predicates via the standard representation of functions by graphs. Let us define an auxiliary function $Sig : \mathcal{T} \to 2^{\Sigma}$ that we will use throughout this paper. For any set of sentences from $\mathcal{T}$ this function gives a set of signature elements occurring in these sentences.

Definition 1 *Let $\mathcal{T}$ be an elementary theory in a signature Σ. A relation $R \subseteq \Sigma \times \Sigma$ is called a* ***syntactical relation*** *on $\mathcal{T}$, if*

$$\forall\, a, b \in \Sigma\ ((a, b) \in R \longleftrightarrow \exists\, \varphi \in \mathcal{T}\ \ (a \in Sig(\varphi)\ and\ b \in Sig(\varphi)))$$

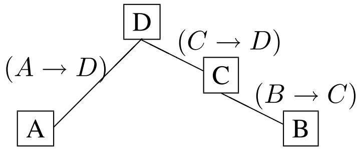

Figure 1: Representation of the thesaurus via a syntactical graph.

We will further use the symbol R to denote syntactical relations. The reader might have a doubt about this definition, as the relation R depends on the form of formulas in the theory. We will explain, how we address this problem in Sect. 3. We assume that by formally describing a subject domain, all the considered terms are mapped onto signature symbols of the constructed theory. One may consider this as determining an alphabet of a language for describing the subject domain. We also take an assumption that there is a connection between two terms, if there exists a sentence in the theory, which contains signature symbols denoting these terms. Thus, for a given theory $\mathcal{T}$ in a signature Σ we may consider a *syntactical graph* with the set of vertices equal to Σ, the set of edges equal to the set of sentences of $\mathcal{T}$ and with the incidence relation R. Let us illustrate this by a simple example with a thesaurus.

Example 1 *Let* $\Sigma = \{A, B, C, D\}$ *and* $\mathcal{T}$ $=\{\forall x(A(x) \to D(x)),\ \forall x(B(x) \to C(x)),\ \forall x(C(x) \to D(x))\}$*. The representation of this kind of a thesaurus in a form of a syntactical graph is illustrated in Fig. 1 (the quantifiers and variables are omitted for brevity).*

In the scope of the scenarios considered at the beginning of this section, we may figuratively speak about key concepts as some subset of vertices and a radius around these vertices, which represents how much of the known information about them is used in solving a concrete task (e.g., a search task).

Definition 2 *We define the operation of incremental data extraction as the following two complementary actions:*
i) extending a given subset $\sigma \subset \Sigma$ *via the relation R (i.e., for a sentence* $\varphi \in \mathcal{T}$, *we add new elements from* $Sig(\varphi)$ *to* σ, *if* $Sig(\varphi) \cap \sigma \neq \varnothing$*);*
ii) extending a given subset $S \subset \mathcal{T}$, *via the relation R (i.e., we add a sentence* $\varphi \in \mathcal{T}$, $\varphi \notin S$, *if there exists* $\psi \in S$, *such that* $Sig(\varphi) \cap Sig(\psi) \neq \varnothing$*).*

3 The Decomposability Problem

Let us generalize the definition of the syntactical relation on signature elements. To preserve mathematical correctness we will further assume that theories, we consider, are deductively closed. In this case they are uniquely defined by sets of their axioms. So, instead of speaking about a theory, we will further consider systems of axioms of this theory.

Definition 3 *Let us call two signature symbols $p,\ q \in \Sigma$ **directly connected** (by a system Φ of axioms of a theory $\mathcal{T}$), if p and q belong to one axiom $\psi \in \Phi$.*

*Correspondingly, let us call p and q **connected** (by a system Φ of axioms), if there exists a sequence of symbols $p = t_1, \ldots,\ t_q = q$, in which every pair $t_i,\ t_{i+1}$ is directly connected.*

The connectedness relation is the equivalence relation on signature symbols, so for a theory $\mathcal{T}$ in a signature Σ we may consider different connectedness (equivalence) components. It is necessary to clarify an important question concerning our idea to employ connectedness of symbols in Sect.2. Indeed, do we have unique connectedness components for different, but (logically) equivalent sets of sentences? In general, the answer is negative and the reason of this is, clearly, the syntactical nature of our approach. In practice, this means there may exist two sets of sentences that semantically mean the same, but are written differently. For example, a sentence may have *invalid* occurrences of symbols in the form of $p \vee \neg p$, which add nothing from the semantic point of view, but change the sentence syntactically. Moreover, sentences can be *glued* with each other by conjunctions in arbitrary manner. In particular, all axioms of a finitely axiomatizable theory can be glued into one axiom, from which all the sentences of the theory can be derived. This leads us to the following question: given a theory $\mathcal{T}$ in a signature Σ, is it possible to reduce $\mathcal{T}$ (axioms of $\mathcal{T}$) to a form that uniquely determines connectedness components on Σ? This question is the reformulation of the decomposability problem posed in [7] in connection with study of formal ontologies. In the following, we formulate this problem and outline our solution, which is given in more detail in [9].

Definition 4 *Let us consider a signature Σ and a theory $\mathcal{T}$ in this signature. The theory $\mathcal{T}$ is called **decomposable**, if its signature can be represented as a disjunctive union $\Sigma = \Sigma_1 \cup \Sigma_2,\ \Sigma_1 \cap \Sigma_2 = \varnothing$, such that there exists a system of axioms $S = S_1 \cup S_2$, in which sentences of $S_i,\ i = 1, 2$ contain symbols only from Σ_i. We will denote a decomposition of theory as $\mathcal{T}= S_1 \otimes S_2$, and decomposition of signature as $\Sigma = \Sigma_1 \coprod \Sigma_2$.*

A theory $\mathcal{T}$ may, obviously, have a *trivial* decomposition, in which $\Sigma_1 = \varnothing$ or $\Sigma_2 = \varnothing$, and the set of sentences S_1 (correspondingly, S_2) consists of those sentences of $\mathcal{T}$, which do not use any signature symbols. Such kind of decomposition is of no interest to us, that is why we will further assume that a theory is decomposable, iff it has a non-trivial decomposition, in which $\Sigma_1 \neq \varnothing \neq \Sigma_2$. A theory that has only trivial decompositions will be called ***non-decomposable***. For instance, if a signature Σ consists of one predicate symbol, then any theory in this signature (even defined by the empty set of axioms) is non-decomposable.

Problem 1 *Consider a theory $\mathcal{T}$ in a signature Σ, defined by some set of axioms Φ in the signature Σ. How is it possible, having the set Φ, to determine, whether $\mathcal{T}$ is decomposable?*

The question of decomposability of theories has significant importance in the field of formal knowledge representation. Since decomposability means the possibility to split a formal representation of a considered subject domain into parts, each described by a separate set of terms. For instance, when building a formal description of a subject domain, it often turns out that data obtained from an expert (or extracted automatically) is a mixture of facts that are needed to be structured in order to obtain an adequate model. In particular it may be interesting, if there exist parts of the knowledge that are independent from each other. This exactly corresponds to the question of decomposability, if one considers a formal description of a subject domain as a logical theory (say, in some subset of the language of the first-order logic). The decomposability problem is important for reasoning over large ontologies and, in particular, for checking their consistency. If ontology can be decomposed into several parts having different signatures, then they can be checked for consistency separately. This, in turn, allows for distributed execution of this operation. There are also other examples that originate from the fact, that in any field of knowledge *decomposition* always means *simplification*.

Basing on the Craig's interpolation theorem [10, 11], we have proved that for any first-order (elementary) theory $\mathcal{T}$ in a signature Σ, the disjunctive decomposition $\Sigma = \Sigma_1 \coprod \ldots \coprod \Sigma_n$ that corresponds to a decomposition of the theory $\mathcal{T}{=}\mathcal{S}_1{\otimes}\ldots{\otimes}\mathcal{S}_n$, is uniquely defined (Theorem 1 below). We have shown, how any system of axioms of a theory can be reduced to such form that uniquely determines this decomposition. Surprisingly, no matter what system of axioms one chooses, the decomposability components are always the same.

In the following, we list the main statements from the proof of this fact.

As we have mentioned above, *invalid* occurrences of symbols (like $p \vee \neg p$) in sentences of a theory influence the decomposability components. This leads us to the following remark.

Remark 1 *If a theory $\mathcal{T}$ in a signature Σ can be defined by a system of axioms, which uses only a part of signature symbols $\Sigma' \subset \Sigma$, then this theory is decomposable. It has the decomposition in a theory with the signature Σ' and theories with signatures from $\Sigma \backslash \Sigma'$, defined by sets of tautological sentences.*

Therefore, it is sufficient to consider the question of decomposability for the theory with the lesser signature. This leads to the following definition.

Definition 5 *Let us consider a signature Σ and a theory $\mathcal{T}$ in this signature. We call $\mathcal{T}$* ***reducible****, if there exists a subset $\Sigma' \subset \Sigma$ of the signature Σ and a system of axioms S for $\mathcal{T}$, which contains symbols only from Σ'. Thus, $\mathcal{T}$ is* ***reduced to the theory*** *$\mathcal{T}'$ in the lesser signature Σ'. If any system of axioms of $\mathcal{T}$ contains all signature symbols of Σ, then $\mathcal{T}$ is* ***irreducible****. Let us call* ***valid*** *all those symbols of Σ that can not be eliminated from any system of axioms of $\mathcal{T}$.*

It is possible to give the definition 6 of a reducible sentence in the same manner. Below we formulate the proposition that justifies this definition.

Proposition 1 *Consider an extension of a signature $\Sigma' \subseteq \Sigma$ and a theory $\mathcal{P}$ in the signature Σ'. Let φ be a sentence of Σ. If φ follows from $\mathcal{P}$, then there exists a sentence $\theta \in \mathcal{P}$,* *such that $\mathcal{P} \vdash \theta, \ \theta \vdash \varphi$. Besides, θ includes only those symbols from Σ' that are present in φ.*

Definition 6 *Consider a theory $\mathcal{T}$ and a sentence $\varphi \in \mathcal{T}$. φ is called* ***reducible in the theory*** *$\mathcal{T}$, if there exists a sentence $\theta \in \mathcal{T}$ that contains fewer signature symbols, than φ does, and for which $\theta \vdash \varphi$. If there are no such sentences, then we call φ* ***irreducible in the theory*** *$\mathcal{T}$.*

In order to define decomposability components, one has, first, to eliminate all invalid symbols from sentences of a theory. It is necessary to prove however, that after having this done, we will have semantically the same theory as before. In other words, everything that could be derived before, can be derived after this operation.

Proposition 2 *Let $\mathcal{T}$ be a theory in a signature Σ. Consider a set of valid symbols Σ' of the signature Σ: $\Sigma' \subset \Sigma$. Then $\mathcal{T}$ is definable by a system of axioms in the signature Σ'. Besides, such a system of axioms defines an irredicible theory.*

One of the main steps on the way to proving the uniqueness of decomposition is the following variant of the Craig's interpolation theorem.

Proposition 3 *Consider a decomposition of signature $\Sigma = \Sigma_1 \coprod \Sigma_2$ and two theories $\mathcal{P}$,$\mathcal{Q}$ in the signatures $\Sigma_1, \ \Sigma_2$ respectively. Consider a sentence φ of the signature Σ. If φ follows from the union of the theories $\mathcal{P}, \ \mathcal{Q} \ \vdash \varphi$, then there exist sentences $\theta \in \mathcal{P}$ and $\phi \in \mathcal{Q}$, such that $\mathcal{P} \vdash \theta, \ \ \mathcal{Q} \vdash \phi$ and $\theta, \phi \vdash \varphi$. Moreover, θ includes only those symbols of Σ_1 that are present in φ. Correspondingly, ϕ contains only those symbols of Σ_2 that are present in φ.*

Definition 7 *Consider a theory $\mathcal{T}$ and a sentence $\varphi \in \mathcal{T}$. Let us call φ* ***decomposable in the theory*** *$\mathcal{T}$, if there exist sentences $\theta \in \mathcal{T}, \psi \in \mathcal{T}$, such that θ, ψ contain symbols only from φ and do not have common signature symbols, neither of them is an equality formula, and $\theta, \psi \vdash \varphi$. We call θ and ψ* ***decomposition components*** *for the sentence φ. If there are no such θ and ψ, then we call φ as* ***non-decomposable in the theory*** *$\mathcal{T}$.*

Remark 2 *If a sentence φ of a theory $\mathcal{T}$ is irreducible, then its decomposition components θ and ψ are also irreducible.*

The following lemma and the corollary are the main auxiliary statements, from which the uniqueness of decomposition of a theory can be concluded. $\langle \mathcal{S}_i, \mathcal{T}^{\#} \rangle$ denotes below a theory $\mathcal{S}_i$ together with equality formulas of theory $\mathcal{T}$ (those formulas that contain variables and use no signature symbols).

Lemma 1 *For any non-trivial decomposition $\mathcal{T} = \mathcal{S}_1 \otimes \mathcal{S}_2$ in a product of theories in signatures $\Sigma = \Sigma_1 \coprod \Sigma_2$ ($\Sigma_1 \neq \varnothing \neq \Sigma_2$), every non-decomposable sentence φ of $\mathcal{T}$ that contains signature symbols follows only from $\langle \mathcal{S}_1, \mathcal{T}^{\#} \rangle$ or only from $\langle \mathcal{S}_2, \mathcal{T}^{\#} \rangle$.*

If, in addition, φ is irreducible, then it is contained either in the theory $\langle\mathcal{S}_1,\mathcal{T}^{\#}\rangle$ of the signature Σ_1, or in the theory $\langle\mathcal{S}_2,\mathcal{T}^{\#}\rangle$ of Σ_2. In particular, it contains symbols only from Σ_1 or only from Σ_2.

Corollary 1 *Consider a non-trivial decomposition of a theory $\mathcal{T} = \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n$. Then any non-decomposable and irreducible sentence $\varphi \in \mathcal{T}$ belongs to some theory $\langle\mathcal{S}_i, \mathcal{T}^{\#}\rangle$.*

These statements show that it is possible to use any system of non-decomposable and irreducible axioms of a theory to determine its decomposition components. The propositions 1 and 3 explain, how to obtain such system. Due to the paper size limitations, we skip here several steps before formulation of the main result - Theorem 1. However, those that have been mentioned here, are sufficient to derive it.

Theorem 1 *For any irreducible theory $\mathcal{T}$ in a signature Σ the disjunctive decomposition of the signature $\Sigma = \Sigma_1 \coprod \ldots \coprod \Sigma_n$, which corresponds (up to a rearrangement of components) to a decomposition $\mathcal{T}$=$\mathcal{S}_1\otimes\ldots\otimes\mathcal{S}_n$ into non-decomposable theories in signatures $\Sigma_i,\ \ i = 1, \ldots, n$, is uniquely defined. Let us fix a decomposition of the theory $\mathcal{T}$.*

Then, for an arbitrary decomposition of $\mathcal{T}$ into non-decomposable components $\mathcal{T} = \mathcal{T}_1 \otimes \ldots \otimes \mathcal{T}_m$ we have $n = m$ and after an appropriate re-enumeration of components, every $\mathcal{S}_i$ differs from $\mathcal{T}_i$ only by equality formulas from the initial theory $\mathcal{T}$. Therefore, $\langle\mathcal{S}_i,\mathcal{T}^{\#}\rangle$=$\langle\mathcal{T}_i,\mathcal{T}^{\#}\rangle$.

This theorem states that for any theory $\mathcal{T}$ its decomposition components are uniquely defined (clearly, up to equality formulas that use no signature symbols).

It is important to note that though, we have proposed a solution of the decomposability problem at theoretical level, it is necessary to come to real-world applications of the obtained results. In particular, it is necessary to develop a decomposition algorithm based on the statements above that could work for some subsets of the first-order logic. This is the subject of our research presently.

4 Summary

In this paper, we considered declarative knowledge bases represented by sets of first-order sentences. The focus of the talk will be on logical and model-theoretical frameworks for dealing with declarative descriptions and, in particular, for managing ontologies. It will turn out, that in this framework, *formal ontology*, *declarative KB* and *logical theory* are synonyms. In particular, we will raise the question for discussion, whether it is practical to assign any understanding to ontology, other than the notion of logical theory. In this scope, we will argue that most definitions of ontology employ properties that can not be checked algorithmically (at least) unless the form of ontological statements (formulas) is fixed, like in the case of thesauri. But the more general way to consider ontology is to view it just as a set of sentences in a given logical language. For instance, the language of the first-order logic, which leads us to the notion of elementary theory. In this context, we will talk about the operation of incremental data extraction and the decomposability problem. We will explain the importance of these two notions in the scope of the Semantic Web paradigm and provide both, theoretical and practical background for their consideration. By the example of the decomposability problem, we will show, how a question seeming to be just of a theoretical interest, can have valuable applications in the scope of reasoning over ontologies. In general, we will try to draw attention to the need of extended fundamental research in ontological engineering.

References

[1] G. Fu, C. Jones, A. Abdelmoty. Ontology-based spatial query expansion in information retrieval. *In Proceedings of the OTM Conferences*, 2, 2005.

[2] Hans-Michael Muller, E. Kenny, P Sternberg. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology Journal*, 2(11), 2004.

[3] S. Melnik, H. Molina-Garcia, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. *In Proceedings of the International Conference on Data Engineering (ICDE)*, 2002.

[4] E. Rahm and P. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), 2001.

[5] H. Stuckenschmidt, F. Giunchiglia, and Frank van Harmelen. Query processing in ontology-based peer-to-peer systems. *Ontologies for Agents: Theory and Experiences*, Birkhäuser, 2005.

[6] S. Castano, A. Ferrara, S. Montanelli, E. Pagani, and G. Rossi. Ontology-addressable contents in P2P networks. *In Proceedings of the 1st Workshop on Semantics in Peer-to-Peer and Grid Computing*, 2003.

[7] D. Palchunov. GABEK for Ontology Generation. *GABEK. Contributions to Knowledge Organization, Vol.2*, Wien: LIT-publishing Company, 2005.

[8] D. Ponomaryov, N. Omelianchuk, N. Kolchanov, E. Mjolsness, and E. Meyerowitz. Semantically rich ontology of anatomical structure and development for Arabidopsis thaliana (L.). *In Proceedings of the BGRS Conference*, 2006.

[9] D. Ponomaryov. On decomposability of elementary theories. Algebra and Model Theory, 5, *Collection of Papers*, Novosibirsk State Technical University, 2005.

[10] C. Chang and H. Keisler. Model theory. North-Holland Publishing Co., Amsterdam, 1973.

[11] M. Otto. An interpolation theorem. *Bulletin of Symbolic Logic*, 6, 2000.