Siberian Division of the Russian Academy of Sciences
A. P. Ershov Institute of Informatics Systems

Denis Ponomaryov

LATTICE SEMANTICS FOR INCREMENTAL DATA EXTRACTION FROM DECLARATIVE KNOWLEDGE BASES

Preprint
134

Novosibirsk 2006

In this paper we consider declarative knowledge bases as first-order (elementary) theories. We study an operation of incremental data extraction, which is closely connected with query reformulation in information retrieval and data integration. We introduce a formal definition of this operation and provide it with semantics in terms of lattices by using the main theorem of Formal Concept Analysis.

Российская академия наук
Сибирское отделение
Институт систем информатики
имени А. П. Ершова

Денис Пономарев

СЕМАНТИКА В ТЕРМИНАХ РЕШЕТОК ДЛЯ ОПЕРАЦИИ ПОСЛЕДОВАТЕЛЬНОЙ ВЫБОРКИ ДАННЫХ ИЗ ДЕКЛАРАТИВНЫХ БАЗ ЗНАНИЙ

Препринт
134

Новосибирск 2006

В данной работе мы рассматриваем декларативные базы знаний как (элементарные) теории в логике первого порядка. Мы исследуем операцию последовательной выборки данных, которая тесно свзяана с переформулированием запросов при информационном поиске и интеграции разнородных источников данных. Мы вводим формальное определение данной операции и определяем ее семанитку в терминах решеток, используя основную теорему метода анализа формальных понятий (Formal Concept Analysis).

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. The outcome of this is development and application of new descriptive languages, as well as reasoning or deductive systems. Each of 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 connection with the Internet search problem, however there are many other actual applications [1, 2, 7]. 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. 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 as 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

basic notions and the main theorem of Formal Concept Analysis in Sect. 3 as detailed as it is needed in this paper. Then we introduce the lattice semantics for the mentioned operation in Sect. 4 and consider one algorithm connected with the incremental data extraction in terms of lattices. Section 5 contains some final remarks and conclusions.

2. THE OPERATION OF INCREMENTAL DATA EXTRACTION

In our work we consider declarative knowledge bases as finitely axiomatizable elementary theories and assume that they are not deductively closed. 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 the 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 an 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 an incremental data extraction.

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

Definition 1. *Let T_Σ be an elementary theory of a signature Σ. A relation $R \subseteq \Sigma \times \Sigma$ is called a* ***syntactical relation*** *on T_Σ, if*

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

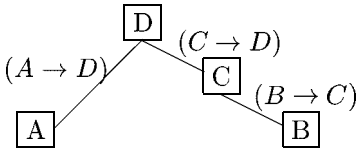

Fig. 1. Representation of the thesaurus via a syntactical graph

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 a theory. We explain how we address this problem in Sect. 4.

We assume that by formally describing a subject domain, all the considered concepts 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 concepts, if there exists a sentence in the theory, which contains signature symbols denoting these terms. Thus, for a given theory T_Σ 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 T_Σ 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* $T_\Sigma = \{\forall x(A(x) \to D(x)),\ \forall x(B(x) \to C(x)),\ \forall x(C(x) \to D(x))\}$. *The representation of this kind of a thesaurus in a form of a syntactical graph is illustrated below in Fig. 1 (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 an incremental data extraction as the following two complementary actions:*

1. *Extending a given subset* $\sigma \subset \Sigma$ *via the relation* R *(i.e., for a sentence* $\varphi \in T_\Sigma$, *we add new elements from* $Sig(\varphi)$ *to* σ, *if* $Sig(\varphi) \cap \sigma \neq \varnothing$*);*

2. *Extending a given subset* $S \subset T_\Sigma$, *via the relation* R *(i.e., we add a sentence* $\varphi \in T_\Sigma$, $\varphi \notin S$, *if there exists* $\psi \in S$ *such that* $Sig(\varphi) \cap Sig(\psi) \neq \varnothing$*).*

3. BASIC NOTIONS OF FORMAL CONCEPT ANALYSIS

Our aim is to give a formal semantics to the operation of an incremental data extraction introduced above. For this purpose we employ the main theorem of Formal Concept Analysis [9]. This method was developed by Ganter and Wille as a restructuring of the lattice theory and a formalization of the notion of *concept*. Further we define basic notions of FCA as detailed as it is needed to explain the use of the main theorem.

Definition 3. *A **formal context** is a triple (G, M, I), where G and M are sets and $I \subseteq G \times M$ is a relation between G and M. The elements of G are called **objects** and the elements of M are called **attributes** of the formal context. The relation I is called the **incidence relation** of the context.*

Definition 4. *For sets $A \subseteq G$ and $B \subseteq M$ we define an operation $'$ as follows:*
$A' = \{m \in M \mid gIm \text{ for all } g \in A\}$
and $B' = \{g \in G \mid gIm \text{ for all } m \in B\}$

Definition 5. *A **formal concept** of the formal context (G, M, I) is a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. A is called the **extent** and B is called the **intent** of the formal concept (A, B).*

Sometimes we will omit the word *formal* and call such pairs simply concepts.

Proposition 1. *Let (G, M, I) be a formal context and $A, A_1, A_2 \subseteq G$. Then the following is true:*
1. $A_1 \subseteq A_2 \Longleftrightarrow A_2' \subseteq A_1'$
2. $A \subseteq A''$.

Definition 6. *Let us define a relation $\leq$ as follows: if (A_1, B_1) and (A_2, B_2) are concepts, then $(A_1, B_1) \leq (A_2, B_2)$, if $A_1 \subseteq A_2$ (which is equivalent to $B_2 \subseteq B_1$, according to proposition 1).*
*(A_1, B_1) is called **sub-concept** of (A_2, B_2) and (A_2, B_2) is called **super-concept** of (A_1, B_1).*
*The relation $\leq$ is called the **hierarchical order** or simply the **order** of the concepts.*

Definition 7. *For a formal context $K = (G, M, I)$, the ordered set of all concepts of K is called the **concept lattice** of K.*

	1	2	3
A	$\times$		
B		$\times$	
C		$\times$	$\times$
D	$\times$		$\times$

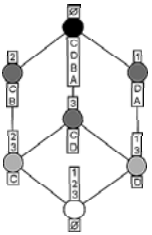

Fig. 2. A formal context for T_Σ and the resulting concept lattice

The Main Theorem of FCA *The concept lattice of a formal context is a complete lattice, in which the infimum and the supremum are given by:*

$$\bigwedge_{j\in J}(A_j,B_j) = (\bigcap_{j\in J}A_j,\ (\bigcup_{j\in J}B_j)'')$$

$$\bigvee_{j\in J}(A_j,B_j) = ((\bigcup_{j\in J}A_j)'',\ \bigcap_{j\in J}B_j)$$

4. LATTICE SEMANTICS VIA FCA

In the following, we introduce the lattice semantics for the operation of an incremental data extraction, considered in Sect. 2.

Let T_Σ be an elementary theory of a signature Σ. Consider a formal context $K = (\Sigma, T_\Sigma, I)$ with the set of objects equal to the set of signature elements, the set of attributes equal to the set of sentences of the theory T_Σ and with the incidence relation I defined as follows:

$$\forall a \in \Sigma\ \forall \varphi \in T_\Sigma\ (aI\varphi,\ if\ a \in Sig(\varphi)) \tag{1}$$

Then a formal concept in K is a pair (A,B), in which A is a set of signature elements and B is the set of all those sentences in T_Σ which contain all symbols from A. Let us illustrate this again by an example with a thesaurus.

Example 2. *Consider the theory T_Σ from example 1. Let us assign indices $\{1,2,3\}$ to the sentences in T_Σ. Then, the corresponding formal context and, more precisely, the incidence relation I can be represented by the table in Fig.2. The nodes in the resulting concept lattice are labelled with extents and intents of the corresponding formal concepts.*

Considering a declarative knowledge base as an elementary theory T_Σ of a signature Σ and having a corresponding formal context $K = (\Sigma, T_\Sigma, I)$, we define the operation of an incremental data extraction from T_Σ as a computation of the order on formal concepts in K. By this we mean a sequential computation of super-/sub-concepts for concepts from a given initial set.

The initial set of concepts for a subset $\sigma \subset \Sigma$ is computed as the antichain of minimal concepts (A_i, B_i), such that $\cup A_i \supseteq \sigma$. Conversely, the initial set of concepts for $S \subset T_\Sigma$ is computed as the antichain of maximal concepts (A_i, B_i), such that $\cup B_i \supseteq S$. This corresponds to definitions 2 and 6.

The set of concepts computed by an incremental data extraction procedure can be represented by a union of chains in the concept lattice of K. Let us call such unions of chains as *paths*.

Definition 8. *A path P is called **continuous**, if $(\varnothing, T_\Sigma) \notin P$ and $(\Sigma, \varnothing) \notin P$.*

Definition 9. *Let us consider a signature Σ and a theory T_Σ in this signature defined by some set of closed formulas (sentences).*

*The theory T_Σ is called **decomposable**, if the signature is a disjoint union of two subsets Σ_1, Σ_2, $\Sigma_1 \cap \Sigma_2 = \varnothing$, $\Sigma = \Sigma_1 \cup \Sigma_2$ and there exist theories S_1, S_2 for T_Σ, such that $T_\Sigma = S_1 \cup S_2$ and each symbol used in $S_i, i = 1, 2$ is from the signature Σ_i, respectively. We denote $T_\Sigma = S_1 \cup S_2$.*

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). For the more detailed study of decomposability of theories, please refer to the paper [8].

Proposition 2. *Let T_Σ be a decomposable theory of a signature Σ, represented by a disjoint union of two sets of sentences $T_\Sigma = S_1 \cup S_2$. Let $K = (\Sigma, T_\Sigma, I)$ be a*

formal context with the incidence relation I defined as in (1) and L be a concept lattice of K.

Then there is no continuous path in L, which contains formal concepts (A_i, B_i), such that $\cup B_i \cap S_1 \neq \varnothing$ and $\cup B_i \cap S_2 \neq \varnothing$.

Obviously, as $T_\Sigma = S_1 \cup S_2$, $S_1 \cap S_2 = \varnothing$ and $Sig(S_1) \cap Sig(S_2) = \varnothing$, the theory T_Σ can be represented by two independent formal contexts, in the sense that the sets of their objects and attributes do not intersect. This means that the theory can be represented by a join of two different concept lattices having only two common elements, which are $(\varnothing, T_\Sigma)$ and $(\Sigma, \varnothing)$.

Let us also notice that if a theory is decomposable, then a table representation of the corresponding formal context (see Example 2) has the form of a block diagonal matrix (after an appropriate rearrangement of lines and columns).

It is necessary to clarify an important question concerning our idea to represent theories in the form of lattices. Indeed, do we have a unique lattice representation for (logically) equivalent theories? In general, the answer is negative and the reason of this is the syntactical nature of our approach. However, it is possible to obtain a one-to-one correspondence, if some *normalization* of a theory is made:

1. All sentences of the theory are reduced to the conjunctive normal form and all the conjunctions are split into separate sentences;
2. Each sentence is transformed into an equivalent one, which uses the least number of signature symbols.

Regarding the second operation, it follows from the Craig's interpolation theorem [10, 11] that for any sentence φ there exists a unique sentence $\psi \sim \varphi$, such that ψ has the least number of signature symbols. In particular, all invalid occurrences of symbols (in the form of $(p \vee \neg p)$) can be eliminated.

Finally, we consider an algorithm for solving a problem, which is important in the scope of the incremental data extraction. As the input can be a subset of a signature of a theory, it is important to determine a minimal set of sentences covering this subset (further it can be extended according to a chosen strategy). We formulate this as the following problem.

Problem. *Given a theory T_Σ in a signature Σ and a subset $\sigma \subseteq \Sigma$, find a minimal set of sentences $S \subseteq T_\Sigma$, such that $Sig(S) \supseteq \sigma$.*

Clearly, there may exist several sets of sentences in T_Σ satisfying this property and the corresponding algorithm is non-deterministic. Here we briefly describe the algorithm, without mentioning implementation details. Here we assume that for

any element $a \in \Sigma$, there is at least one sentence $\varphi \in T_\Sigma$, such that $a \in \varphi$.

Algorithm.

1. Compute an antichain of maximal concepts in the lattice of the context $K = (\Sigma, T_\Sigma, I)$, where I is defined as in (1) (let us denote the obtained set of concepts by C_{max}).
2. Take a concept $(A, B) \in C_{max}$, such that the intersection $A \cap \sigma$ is maximal. Add any sentence from B to the output set U. Reduce σ by A and C_{max} by (A, B), correspondingly.
3. Repeat the previous step, until $\sigma = \varnothing$. Return U.

5. CONCLUSIONS

In this paper, we considered declarative knowledge bases represented by sets of first-order closed formulas of an arbitrary form. However it is often practical, especially regarding the normalization of theories, to restrict ourselves to specific classes of formulas. We believe, it is quite reasonable to consider Horn formulas, on which modern rule languages are based. There are almost no difficulties with the first normalization step for this sort of formulas, including extended Horn clauses, with respect to Lloyd-Topor transformations [12]. Not taking equivalences into account, the time complexity of these transformations is linear in the length of the formula. The second normalization step is proved to be of polynomial complexity [13] in this case. In general, it is important to notice that the applicability of our approach does not depend on a language, but relies much upon the normalization procedure in it.

REFERENCES

1. Fu G., Jones C., Abdelmoty A. Ontology-based spatial query expansion in information retrieval. // Proc. of the OTM Conferences. – 2005. – Vol. 2.
2. Muller H., Kenny E., Sternberg P. Textpresso: An ontology-based information retrieval and extraction system for biological literature // PLoS Biology J. – 2004. – Vol. 2(11).
3. Melnik S., Molina-Garcia H., Rahm E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching // Proc. of the Internat. Conf. on Data Engineering (ICDE). – 2002.
4. Rahm E., Bernstein P. A survey of approaches to automatic schema matching // The VLDB J. – 2001. – Vol. 10(4).
5. Stuckenschmidt H., Giunchiglia F., van Harmelen F. Query processing in ontology-based peer-to-peer systems // Ontologies for Agents: Theory and Experiences. – Birkhäuser, 2005.

6. Castano S., Ferrara A., Montanelli S., Pagani E., Rossi G. Ontology-addressable contents in P2P networks // Proc. of the 1st Workshop on Semantics in Peer-to-Peer and Grid Computing at the 12th International World Wide Web Conf. – 2003.
7. Ponomaryov D., Omelianchuk N., Kolchanov N., Mjolsness E., Meyerowitz E. Semantically rich ontology of anatomical structure and development for Arabidopsis thaliana (L.) // Proc. of the BGRS Conf. – 2006.
8. Ponomaryov D. On decomposability of elementary theories // Algebra and Model Theory. – Novosibirsk State Technical University, 2005. – Vol. 5.
9. Ganter B., Wille R. Formal Concept Analysis – Mathematical Foundations. – Springer Verlag, 1999.
10. Chang C., Keisler H. Model theory. – – Amsterdam: North-Holland Publishing Co., 1973.
11. Otto M. An interpolation theorem // Bull. of Symbolic Logic. – 2000. – Vol. 6.
12. Lloyd J., Topor R. Making Prolog more expressive // J. of Logic Programming. – 1984 – Vol. 3.
13. Dahlhaus E., Israeli A., Makowsky J.A. On the existence of polynomial time algorythms for interpolation problems in propositional logic. // *Notre Dame Journal of Formal Logic* – 1988. – Vol. 29(4).

Денис Пономарев

СЕМАНТИКА В ТЕРМИНАХ РЕШЕТОК ДЛЯ ОПЕРАЦИИ ПОСЛЕДОВАТЕЛЬНОЙ ВЫБОРКИ ДАННЫХ ИЗ ДЕКЛАРАТИВНЫХ БАЗ ЗНАНИЙ

Препринт
134

Рукопись поступила в редакцию 21.06.2006
Рецензент Ю. А. Загорулько
Редактор А. А. Шелухина

Подписано в печать 10.07.2006
Формат бумаги 60×84 1/16 Объем 0,85 уч.-изд.л., 0,94 п.л.
Тираж 60 экз.

Центр оперативной печати "Оригинал 2", г. Бердск, 49-а, оф. 7, тел./факс 8 (241) 5 38 77