Data-Driven Logical Reasoning

Claudia d'Amato[1], Volha Bryl[2], Luciano Serafini[2]

[1] Department of Computer Science - University of Bari, Italy
claudia.damato@di.uniba.it
[2] Data & Knowledge Management Unit - Fondazione Bruno Kessler, Italy
{bryl|serafini}@fbk.eu

Abstract. The co-existence of heterogeneous but complementary data sources, such as ontologies and databases describing the same domain, is the reality of the Web today. In this paper we argue that this complementarity could be exploited both for discovering the knowledge not captured in the ontology but learnable from the data, and for enhancing the process of ontological reasoning by relying on the combination of formal domain models and evidence coming from data. We build upon our previous work on knowledge discovery from heterogeneous sources of information via association rules mining, and propose a method for automated reasoning on grounded knowledge bases (i.e. knowledge bases linked to data) based on the standard Tableaux algorithm. The proposed approach combines logical reasoning and statistical inference thus making sense of heterogeneous data sources.

1 Introduction

From the introduction of the Semantic Web view [3], many domain ontologies have been developed and stored in open access repositories. However, still huge amounts of data are stored in relational databases (DBs) and managed by RDBMSs (relational database management systems). The seamless integration of these two knowledge representation paradigms is becoming a crucial research challenge. Most of the work in this area concerns what is addressed as *ontology based data access (OBDA)* [4]. In OBDA the ontology "replicates" at a higher conceptual level the physical schema of the DBMS and provides a "lens" under which the data can be viewed, and possibly adds additional semantic knowledge on the data. The connection between the ontology and the data is represented as conjunctive queries. Roughly speaking, every concept/relation of the ontology is associated to a conjunctive query which retrieves from the DB all and only the instances of such a concept/relation.

Another common situation is when existing ontologies describe domain aspects that (partially) complement data in a database. In this case the concepts of the ontologies are linkable to views of the DB. Also in this case it would be very useful to be able to combine the knowledge contained in the two information sources, for example for enriching the existing ontologies. Due to the heterogeneity of the information a crisp representation of the correspondence between the DB data and the classes and relations of the ontologies (such as the one adopted in OBDA) is not possible. A more flexible connection between the two sources of knowledge should be adopted. An option could be to exploit rules that are able to express a statistical evidence of the connection between the data in a DB and the knowledge in the ontology. For giving the intuition for this solution, let us consider the following scenario. Given an existing ontology describing people gender, family status and their interrelations with Italian urban areas[1] and a demographic DB describing Italian occupations, average salaries, etc., a possible connection between the two information sources can be described with rules as follows:

> "*clerks between 35 and 45 years old* ***living in a big city*** *are* ***male*** *and earn between 40 and 50 €*" with a *confidence* value of 0.75. (1)

where bold face terms correspond to classes and relations in the ontology, non bold face terms correspond to data in the DB. The *confidence* value can be interpreted as the probability that the specified connection between the two sources occurs. Rules of the form (1) are called *semantically enriched association rules*. They have been introduced in our previous work [5] where an inductive approach for discovering new knowledge in the form of *association rules* [1] from heterogeneous data sources is proposed.

In this paper, we revise the approach introduced in [5] by taking into account the *Open World Assumption* adopted in description logics (DLs), while in [5] association rules are extracted from the hybrid data sources by adopting an implicit *Closed Word Assumption* that is not fully compliant with the theory of ontological representation. We also make a further step towards the framework for knowledge representation and reasoning in which knowledge can be represented by a mix of logical formulas and sets of data, linked together. Specifically, we introduce a concept of a *grounded knowledge base* and the notion of *mixed model* that integrates logical knowledge expressed in terms of a description logic knowledge base, and a statistical data mining model that expresses the statistical regularities of the properties associated to a set of individuals. Finally and most importantly, we propose a method for automated reasoning on *grounded knowledge bases*, which is the result of combining logical reasoning and statistical inductive inference and learning. In particular, we propose an extension of the standard Tableaux algorithm grounded on the adoption of an heuristic to be used when random choices (i.e. the processing of a disjunction, namely when we need to decide whether an object x belongs to concept C or to concept D) have to be made during the reasoning process. The heuristic exploits the evidence coming from the data. Assume, for example, that for a given object x, which is a $Person$, a high school student, and has the property x is 15 years old, we need to decide whether x is a $Parent$ or not, and there is no statements in the knowledge base from which it is possible to infer neither x is a $Parent$ nor x is $\neg Parent$. The following association rule learned from the data (with high degree of confidence)

$$\textsc{Age} = [0, 16] \Rightarrow \neg Parent \quad 0.99$$

can be exploited to conclude that, with high probability, x is not a $Parent$.

The rest of the paper is structured as follows. In Section 2 we give basic definitions necessary to set up the framework. In Section 3 we summarize and extend the approach for learning association rules from heterogeneous sources of information pre-

[1] The concepts *"male", "parent", "big city", "medium-sized town",* and the relations *"lives_in"* are used in the ontology.

sented in [5]. Section 4 presents the data-driven Tableaux reasoning algorithm, followed by discussions and conclusions in Section 5.

2 Basic definitions

Let $\mathbf{D}$ be a non empty set of objects and $f_1, \ldots, f_n$ be n feature functions defined on every element of $\mathbf{D}$, with $f_i : \mathbf{D} \to D_i$. $\mathbf{D}$ is called the set of observed objects and $f_i(d)$ for every $d \in \mathbf{D}$ is the i-th feature observed on d. Notationally we use $\mathbf{d}$ for the elements of $\mathbf{D}$ and $\mathbf{d}_1, \ldots, \mathbf{d}_n$ to denote the values of $f_1(\mathbf{d}), \ldots, f_n(\mathbf{d})$.

Let Σ be a DL alphabet composed of three disjoint sets of symbols, Σ_C, Σ_R and Σ_I, the set of concepts symbols, the set of role symbols and the set of individual symbols. A knowledge base on Σ, is a set $\mathcal{K}$ of DL inclusion axioms and DL assertions (we assume $\mathcal{ALC}$ as DL language here). The elements of $\mathcal{K}$ are called axioms of $\mathcal{K}$. An axiom can be of the form $X \sqsubseteq Y$, where X and Y are $\mathcal{ALC}$ (complex) concepts, or $X(a)$ where X is a (complex) concept and a an individual symbol, or $R(a, b)$, and $a = b$, where R is a role symbol and a and b are individual symbols. We call $X \sqsubseteq Y$ a subsumption, and $X(a)$, $R(a, b)$, and $a = b$ assertions. $\mathcal{K}$ is formally defined as a couple $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ where $\mathcal{T}$ contains the inclusion axioms ($\mathcal{T}$ stands for Terminological part) and $\mathcal{A}$ contains the assertional axioms ($\mathcal{A}$ stands for Assertional part).

An interpretation of a DL alphabet Σ is a pair $\mathcal{I} = \left\langle \Delta_{\mathcal{I}}, \cdot^{\mathcal{I}} \right\rangle$ such that $\Delta^{\mathcal{I}}$ is a non empty set, and $\cdot^{\mathcal{I}}$ is a function that assigns to each concepts name a subset of $\Delta_{\mathcal{I}}$, to each role name a binary relation on $\Delta_{\mathcal{I}}$, and to each individual an element of $\Delta_{\mathcal{I}}$. The interpretation function can be extended to complex concepts in the usual way [2]. Satisfiability $\models$ of statements is also defined as usual [2]. $\mathcal{I} \models \mathcal{K}$ if $\mathcal{I} \models \phi$ for every axiom of $\mathcal{K}$. An interpretation $\mathcal{I}$ satisfies a knowledge base $\mathcal{K}$, (in symbols $\mathcal{I} \models \mathcal{K}$) if $\mathcal{I} \models \phi$ for every axiom of $\mathcal{K}$.

The "glue" between a dataset and a knowledge base is the so called *grounding*, which is a relation that connects the objects of the knowledge base with the data of the database. More formally: a *grounding* g of Σ on $\mathbf{D}$ is a total function $g : \mathbf{D} \to \Sigma_I$. This implies that for every $\mathbf{d} \in \mathbf{D}$ there is at least an element $a \in \Sigma_I$ with $g(\mathbf{d}) = a$. Intuitively $g(\mathbf{d}) = a$ represents the fact that the data $\mathbf{d}$ are about/correspond to object a of the knowledge base. Please note that the grounding g refers to objects that are explicitly mentioned in $\mathbf{D}$ and $\mathcal{K}$ respectively. In our framework (see Sect. 3) we assume that the grounding between $\mathbf{D}$ and $\mathcal{K}$ is already given.

3 Semantically enriched association rules

Association rules (ARs), originally introduced in [1], make it possible to represent in a rule based form some statistical regularities of the tuples in a relational database. Roughly speaking, ARs allow one to state conditional probabilities among the values of the attributes of the tuples of a database. Learning ARs is one of the fundamental tasks in data-mining.

In this section we recall how ARs can be extended to include information coming from an ontological knowledge base and how they can be used to bridge the knowledge contained in an ontology with that contained in a relational database. These rules are called *semantically enriched* ARs [5]. Hence, we revise the approach introduced in [5] by taking into account the *Open World Assumption* adopted in description logics (DL), while in [5] association rules are extracted from the hybrid data sources by adopting an implicit *Closed Word Assumption* that is not fully compliant with the theory of ontological representation. At the end of the section, the process of learning semantically enriched ARs [5] is also briefly recalled.

3.1 Association rules: an overview

Association rules [1] provide a form of rule patterns for data mining. Let $\mathbf{D}$ be a dataset made by a set of attributes $\{A_1, \ldots, A_n\}$ with domains $\mathcal{D}_i : i \in \{1, \ldots, n\}$. The basic components of an AR for the dataset $\mathbf{D}$ are itemsets. An itemset ϕ is a finite set of assignments of the form $A = a$ with $a \in \mathcal{D}(A)$. An itemset $\{A_{i_1} = a_1, \ldots, A_{i_m} = a_m\}$ can be denoted by the expression

$$A_{i_1} = a_1 \wedge \ldots \wedge A_{i_m} = a_m$$

An AR has the general form

$$\theta \Rightarrow \varphi \tag{2}$$

where θ and φ are itemsets. The *frequency* of an itemsets θ, denoted by $\mathit{freq}(\theta)$, is the number of cases in $\mathbf{D}$ that match θ, i.e.

$$\mathit{freq}(\theta) = |\{\mathbf{d} \in \mathbf{D} \mid \forall (A = a) \in \theta \ : \ f_A(\mathbf{d}) = a\}|$$

where f_A is the feature function for $\mathbf{d}$ w.r.t. the attribute A (see beginning of Sect.2).

The *support* of a rule $\theta \Rightarrow \varphi$ is equal to $\mathit{freq}(\theta \wedge \varphi)$. The *confidence* of a rule $\theta \Rightarrow \varphi$ is the fraction of items in $\mathbf{D}$ that match φ among those matching θ:

$$\mathit{conf}(\theta \Rightarrow \varphi) = \frac{\mathit{freq}(\theta \wedge \varphi)}{\mathit{freq}(\theta)}$$

A frequent itemset expresses the variables and the corresponding values that occur reasonably often together.

In terms of conditional probability, the confidence of a rule $\theta \Rightarrow \varphi$, can be seen as the maximum likelihood (frequency-based) estimate of the conditional probability that φ is true given that θ is true [8].

3.2 Semantically enriched association rules

Let $\mathcal{K}$ be a knowledge base on Σ, $\mathbf{D}$ a dataset and g a grounding of Σ on $\mathbf{D}$.

A *semantically enriched itemset* is a set containing statements of the form $f_i = a$, $C = tv$, $R = tv$ where, f_i is an attribute of $\mathbf{D}$, a a value in the range of f_i, C is a concept name of Σ_C and R is a role name of Σ_R and tv is a truth value in $\{\mathit{true}, \mathit{false}, \mathit{unknown}\}$. The elements of the itemset of the form $f_i = a$ are called *data items*, the elements of the form $C = tv$ and $R = tv$ are called *semantic items*.

A *semantically enriched* AR is an association rules made by *semantically enriched itemsets*. This means that for a certain set of individuals, both knowledge coming from the ontology and information coming from the database are available (see Sect. 3.3 for more details).

Coherently with ARs, it is possible to define the frequency of a *semantically enriched itemset* and the support of a *semantically enriched* AR. Given a grounding g of Σ on $\mathbf{D}$, the *frequency* of a *semantically enriched itemset* $\theta = \theta_d \wedge \theta_k$ (in the following also called *mixed itemset*) is the following generalization of the definition of frequency given for a standard itemset.

$$\mathit{freq}(\theta_d \wedge \theta_k) = |F|$$

where F is the following set:

$$F = \left\{ \mathbf{d} \in \mathbf{D} \left| \begin{array}{l} \forall (f_i = a) \in \theta_d,\ f_i(\mathbf{d}) = a \\ \forall (C = \mathit{true}) \in \theta_k,\ \mathcal{K} \models C(g(\mathbf{d})) \\ \forall (C = \mathit{false}) \in \theta_k,\ \mathcal{K} \models \neg C(g(\mathbf{d})) \\ \forall (C = \mathit{unknown}) \in \theta_k,\ \mathcal{K} \not\models C(g(\mathbf{d}))\ \&\ \mathcal{K} \not\models \neg C(g(\mathbf{d})) \\ \forall (R = \mathit{true}) \in \theta_k,\ \mathcal{K} \models \exists R.\top(g(\mathbf{d})) \\ \forall (R = \mathit{false}) \in \theta_k,\ \mathcal{K} \models \neg\exists R.\top(g(\mathbf{d})) \\ \forall (R = \mathit{unknown}) \in \theta_k,\ \mathcal{K} \not\models \exists R.\top(g(\mathbf{d}))\ \&\ \mathcal{K} \not\models \neg\exists R.\top(g(\mathbf{d})) \end{array} \right. \right\}$$

The support and confidence of a *semantically enriched* AR can be defined similarly.

3.3 Learning semantically enriched association rules

In [5] we proposed a framework for learning *semantically enriched* ARs from heterogeneous sources of information (namely an ontology and a relational database) grounded on the underlying *Closed World Assumption* that is not fully compliant with the theory of ontological representation. Here we extend the framework by taking into account the *Open World Assumption* usually made in DLs that is the theoretical framework underlying the OWL[2] language, namely the standard representation language in the Semantic Web [3]. With regard to this aspect it is important to note that the notions of frequency, confidence and support given in Sect. 3.2 are compliant with the *Open World Semantics*.

The approach for learning *semantically enriched* ARs is grounded on the assumption that a dataset $\mathbf{D}$ and an ontological knowledge base $\mathcal{K}$ share (a subset of) common individuals, and a grounding g of Σ on $\mathbf{D}$ is already available (see the end of Sect. 2 for more details on the grounding function). This assumption is reasonable in practice since, in the real world, there are several cases in which different information aspects concerning the same entities come from different data sources. An example is given by the public administration, where different administrative organizations have information about the same persons but concerning complementary aspects such as: personal data, income data, ownership data. Another example is given by the biological domain where research organizations have their own databases that could be complemented with existing domain ontologies.

The proposed framework is sketched in the following. To learn *semantically enriched* ARs from a dataset $\mathbf{D}$ and a knowledge base $\mathcal{K}$ grounded by g to $\mathbf{D}$, all the

[2] http://en.wikipedia.org/wiki/Web_Ontology_Language

information about the common domain of $\mathcal{K}$ and $\mathbf{D}$ are summarized (proposizionalized) in a tabular representation constructed as follows:

1. choose the primary entity of interest in $\mathbf{D}$ or $\mathcal{K}$ for extracting association rules and set this entity as the first attribute A_1 in the table $\mathbf{T}$ to be built; A_1 will be the primary key of the table
2. choose (a subset of) the attributes in $\mathbf{D}$ that are of interest for A_1 and set them as additional attributes in $\mathbf{T}$; the corresponding values are be obtained as a result of an SQL query involving the selected attributes and A_1
3. choose (a subset of) concept names $\{C_1, \ldots, C_m\}$ in $\mathcal{K}$ that are of interest for A_1 and set their names as additional attribute names in $\mathbf{T}$
4. for each $C_k \in \{C_1, \ldots, C_m\}$ and for each value a_i of A_1, if $\mathcal{K} \models C_k(a_i)$ then set to 1 the corresponding value of C_k in $\mathbf{T}$, else if $\mathcal{K} \models \neg C_k(a_i)$ then set the value to 0, otherwise set to $1/2$ the corresponding value of C_k in $\mathbf{T}$
5. choose (a subset of) role names $\{R_1, \ldots, R_t\}$ in $\mathcal{K}$ that are of interest for A_1 and set their names as additional attribute names in $\mathbf{T}$
6. for each $R_l \in \{R_1, \ldots, R_t\}$ and for each value a_i of A_1, if $\exists y \in \mathcal{K}\ s.t.\ \mathcal{K} \models R_l(a_i, y)$ then set to 1 the value of R_l in $\mathbf{T}$, else if $\forall y \in \mathcal{K}\ \mathcal{K} \models \neg R_l(a_i, y)$ then set the value of R_l in $\mathbf{T}$ to 0, otherwise set the value of R_l in $\mathbf{T}$ to $1/2$
7. choose (a subset of) the datatype property names $\{T_1, \ldots, T_v\}$ in $\mathcal{K}$ that are of interest for A_1 and set their names as additional attribute names in $\mathbf{T}$
8. for each $T_j \in \{T_1, \ldots, T_v\}$ and for each value a_i of A_1, if $\mathcal{K} \models T_j(a_i, \mathit{dataValue}_j)$ then set to *dataValue*$_j$ the corresponding value of T_j in $\mathbf{T}$, set 0 otherwise.

It is straightforward to note that for all but the datatype properties, the *Open World Assumption* is considered during the process for building the tabular representation. Numeric attributes are processed (as usual in data mining) for performing data discretization [10] namely for transforming numerical values in corresponding range of values (categorical values). An example of a unique tabular representation in the demographic domain is reported in Tab. 1 where $Person$, $Parent$, $Male$ and $Female$ are concepts of an ontological knowledge base $\mathcal{K}$, and JOB and AGE are attributes of a relational dataset $\mathbf{D}$. The numeric attribute (AGE) has been discretized.

The choice of representing the integrated source of information within tables allows for directly applying state of the art algorithms for learning association rules. Indeed, once a unique tabular representation is obtained, the well known APRIORI algorithm [1] is applied for discovering *semantically enriched* ARs from the integrated source of information[3] (see [5] for additional details and examples). Specifically, given a certain confidence threshold, ARs having a confidence value equal or greater than the fixed confidence threshold are learnt. This ensures that only significant ARs are considered while the others are discarded. As highlighted in sect. 3.1, the confidence value of the extracted *semantically enriched* ARs is interpreted as the conditional probability on the values of items in the consequence of the rule given that the left hand side of the rule is satisfied in (a model of) the available knowledge. Examples of *semantically enriched* ARs that could be learned from a table like Tab. 1 are reported in Tab. 2.

[3] Since a state of the art algorithm is adopted it is not reported in the paper. The novelty of the proposed approach consists in the way the integrated source of knowledge is built. Once this is obtained, the state of the art APRIORI algorithm is straightforwardly applied.

Table 1. Demographic example: a unique tabular representation $\mathbf{T}$

OBJECT	JOB	AGE	$Person$	$Parent$	$Male$	$Female$
x_1	Engineer	[36,45]	$true$	$true$	$true$	$false$
x_2	Policeman	[26,35]	$true$	$false$	$true$	$unknown$
x_3	Student	[16,25]	$true$	$false$	$true$	$false$
x_4	Student	[16,25]	$true$	$false$	$false$	$true$
x_5	Housewife	[26,35]	$true$	$true$	$false$	$true$
x_6	Clerk	[26,35]	$true$	$false$	$unknown$	$unknown$
...	...	...	...	...	...	...

Table 2. Demographic example: association rules

#	Rule	Confidence
1	(AGE=$[16,25]) \wedge ($JOB $= Student) \Rightarrow \neg Parent$	0.98
2	(JOB=$Policeman) \Rightarrow Male$	0.75
3	(AGE=$[16,25]) \wedge Parent \Rightarrow Female$	0.75
4	(JOB=$Primary\ school\ teacher) \Rightarrow Female$	0.78
5	(JOB=$Housewife) \wedge ($AGE $= [26,35]) \Rightarrow Parent \wedge Female$	0.85

4 Data-driven inference

We want to exploit the *semantically enriched* ARs (see Sect. 3.3) when performing deductive reasoning given DLs (namely ontological) representations. Since almost all DL inferences can be reduced to concept satisfiability [2], we focus on this inference procedure. For most expressive DL (such as $\mathcal{ALC}$) the Tableaux algorithm is employed. Its goal is to built a possible model, namely an interpretation, for the concept whose satisfiability has to be shown. If, building such a model, all clashes (namely contradictions) are found, the model does not exist and the concept is declared to be unsatisfiable.

Our goal is to set up a modified version of the Tableaux algorithm whose output, if any, is the **most plausible model**, namely the model that best fits the available data. This means to set up a data driven heuristic that should allow reducing the computational effort in finding a model for a given concept and should be also able to supply the model that is most coherent with/match the available knowledge. In this way the *variance due to intended diversity and incomplete knowledge* is reduced, namely, the number of possible models that could be built (see [7] for formal definitions). The inference problem we want to solve is formally defined as follows:

Definition 1 (Inference Problem).

Given: $\mathbf{D}$, *$\mathcal{K}$, the set R of ARs, a (possibly complex) concept E of $\mathcal{K}$, the individuals $x_1, \ldots, x_k \in \mathcal{K}$ that are instances of E, the grounding g of Σ on* $\mathbf{D}$
Determine: *the model $\mathcal{I}_r$ for E representing the* ***most plausible model*** *given the $\mathcal{K}$,* $\mathbf{D}$*, g and R.*

Intuitively, the most plausible model for E is the one on top of the ranking of the possible models $\mathcal{I}_i$ for E. The ranking of the possible models is built according to the degree up to which the models respect the ARs. The detailed procedure for building the *most plausible model* is illustrated in the following.

In order to find (or not find) a model, the standard Tableaux algorithm exploits a set of transformation rules that are applied to the considered concept. A transformation rule for each constructor of the considered language exists. In the following, the transformation rules for $\mathcal{ALC}$ logic are briefly recalled (see [2] for more details).

$\sqcap$-rule: IF the ABox $\mathcal{A}$ contains $(C_1 \sqcap C_2)(x)$, but it does not contain both $C_1(x)$ and $C_2(x)$ **THEN** $\mathcal{A} = \mathcal{A} \cup \{C_1(x), C_2(x)\}$

$\sqcup$-rule: IF $\mathcal{A}$ contains $(C_1 \sqcup C_2)(x)$, but it does not contain neither $C_1(x)$ nor $C_2(x)$ **THEN** $\mathcal{A}_1 = \mathcal{A} \cup \{C_1(x)\}$, $\mathcal{A}_2 = \mathcal{A} \cup \{C_2(x)\}$

$\exists$-rule: IF $\mathcal{A}$ contains $(\exists R.C)(x)$, but there is no individual name z s.t. $C(z)$ and $R(x, z)$ are in $\mathcal{A}$ **THEN** $\mathcal{A} = \mathcal{A} \cup \{C(y), R(x, y)\}$ where y is an individual name not occurring in $\mathcal{A}$.

$\forall$-rule: IF $\mathcal{A}$ contains $(\forall R.C)(x)$ and $R(x, y)$, but it does not contain $C(y)$ **THEN** $\mathcal{A} = \mathcal{A} \cup \{C(y)\}$

To test the satisfiability of a concept E, the algorithm starts with the ABox $\mathcal{A} = E(x_0)$ (with x_0 being a new individual) and applies to the ABox the consistency preserving transformation rules reported above until no more rules apply. The result could be all clashes, which means the concept is unsatisfiable, or an ABox containing a model for the concept E that means the concept is satisfiable.

The transformation rule for the disjunction ($\sqcup$-rule) is non-deterministic, that is, a given ABox is transformed into finitely many new ABoxes. The original ABox is consistent if and only if one of the new ABoxes is so. In order to save the computational complexity, the ideal solution (for the case of a consistent concept) should be to choose the ABox containing a model directly. Moving from this observation, in the following we propose an alternative version of the Tableaux algorithm. The main differences with respect to the standard Tableaux algorithm summarized above are:

1. the starting model for the inference process is given by the set of all attributes (and corresponding values) of $\mathbf{D}$ that are related to individuals $x_1, \ldots, x_k$ that are instances of E differently from the standard Tableaux algorithm where the initial model is simply given by the assertion concerning the concept of which the satisfiability (or unsatisfiability) has to be shown,
2. a heuristic is adopted in performing the $\sqcup$-rule, differently from the standard case where no heuristic is given,
3. the most plausible model for the concept E and individuals $x_1, \ldots, x_k$ is built with respect to the available knowledge $\mathcal{K}$, $\mathbf{D}$ and R. The obtained model is a *mixed model*, namely a model containing both information from R and $\mathcal{K}$. Differently, in the standard Tableaux algorithm the model that is built only refers to $\mathcal{K}$ and does not take into account the (assertional) available knowledge.

In the following these three characteristics are analyzed and the way for accomplish each of them is illustrated. First of all, the way in which the starting model $\mathcal{I}_r$ is built is illustrated. For each $x_i \in \{x_1, \ldots, x_k\}$, all attribute names A_i related to x_i are selected[4]

[4] As an example the following query may be performed: SELECT * FROM ⟨TABLE_NAME⟩ WHERE $A_i = x_i$. Alternatively, a subset of the attributes in $\mathbf{D}$ may be considered.

jointly with the corresponding attribute values a_i. The assertions $A_i(a_i)$ are added to $\mathcal{I}_r$. For simplicity and without loss of generality, a single individual x will be considered in the following. The generalization to multiple individuals is straightforward by simply applying the same procedure to all individuals that are (or assumed to be) instances of the considered concept.

Once the initial model $\mathcal{I}_r$ is built, all deterministic expansion rules, namely all but $\sqcup$-rule, are applied following the standard Tableaux algorithm as reported above. Instead, for the case of the $\sqcup$-rule, a heuristic is adopted. The goal of such a heuristic is twofold: a) choosing a new consistent ABox almost in one step to save computational complexity if $E(x)$ is consistent (see discussion above concerning the $\sqcup$-rule); b) driving the construction of the most plausible model given $\mathcal{K}$ and R. The approach for the assessing the heuristic is illustrated in the following.

Let $C \sqcup D$ be the disjunctive concept to be processed by $\sqcup$-rule. The choice on C rather than D (or vice versa) will be driven by the following process.

- The ARs (see Sect. 3.2) containing C (resp. D) or its negation in the *semantic items* of the right hand side of the rules are selected.
- Given the model under construction $\mathcal{I}_r$, the left hand side of each selected rule is considered and the degree of match is computed. This is done by counting the number of (both data and semantic) items in the left hand side of a rule that are contained in $\mathcal{I}_r$, and averaging this number w.r.t. the length of the left hand side of the rule. Items with uncertain (*unknown*) values are not taken into account. The degree of match for the rules whose (part of the) left hand side is contradictory w.r.t. to the model is set to 0.
- After the degree of match is computed, the rules having the degree of match equal to 0 are discarded.
- For each of the remaining rules the weighted confidence value is computed as $weightedConf = ruleConfidence * degreeOfMatch$.
- Rules that have the degree of match below a given threshold (e.g. 0.75) are discarded.
- The rule having the highest weighted confidence value is selected; in case of equal weighted confidence value of different rules, a random choice is performed.
- If the chosen rule contains $C = 1$ (resp. $D = 1$) in the right hand side, the model under construction $\mathcal{I}_r$ is enriched with $C(x)$ (resp. $D(x)$), where x is the individual under consideration.
- If the chosen rule contains $C = 0$ (resp. $D = 0$) in the right hand side, the model under construction $\mathcal{I}_r$ is enriched with $D(x)$ (resp. $C(x)$).
- In the general case the right hand side of the selected AR may contain additional items besides that involving C or D. Assertions concerning such additional items will be also added in $\mathcal{I}_r$ accordingly[5].

If there are no extracted ARs (satisfying a fixed confidence threshold) containing neither C or D in the right hand side, the following approach may be adopted.

[5] If a most conservative behavior of the heuristic has to be considered only the assertion concerning the disjunct C (resp. D) will be added in $\mathcal{I}_r$ while the additional items in the right hand side of the selected rules are not taken into account.

Given $\mathcal{I}_r$, a corresponding item set is created by transforming each assertion $A_i(a_i)$ referring to an attribute in $\mathbf{D}$ as a data item $A_i = a_i$, each concept and role assertion to a knowledge item. Specifically, each positive (not negated) assertion is transformed in $concept/role\ name = 1$, each negative assertion is transformed in $concept/role\ name = 0$. Let θ be the conventional name of such a built itemset. Four rules, $\theta \Rightarrow C = 1$, $\theta \Rightarrow C = 0$, $\theta \Rightarrow D = 1$ and $\theta \Rightarrow D = 0$ are created and their confidence value is computed (see Sect. 3.2). Then, the rule having the highest confidence (satisfying a given confidence threshold) value is selected and the corresponding right hand side will be used as a guideline for expanding $\mathcal{I}_r$.

The presented approach for the case in which no rules are available could result to be computationally expensive. As an alternative, the following criterion, grounded in the exploitation of the prior probability of C (resp. D) could be used. Specifically, the prior probability is computed, by adopting a frequency-based approach, as: $P(C) = |ext(C)|/|\mathcal{A}|$ where $ext(C)$ is the extension of C, namely the number of individuals that are instances (asserted or derived) of C and $|\cdot|$ returns the cardinality of the set extension. Similarly $P(D)$ can be defined for D. The concept to be chosen for extending $\mathcal{I}_r$ will be the one having the highest prior probability.

In the cases discussed above, the disjunctive expression is assumed to be made by atomic concept names. However, in $\mathcal{ALC}$, more complex expressions may occur as part of a disjunctive expression as: existential concept restrictions (i.e. $\exists R.A \sqcup \exists R.B$), universal concept restrictions (i.e. $\forall R.A \sqcup \forall S.B$), nested concept expression (i.e. $\exists R.\exists S.A$ or $\exists R.(A \sqcap B)$). To cope with these cases a straightforward solution is envisioned: new concept names are created for naming the cases listed above. In this way, a disjunction of atomic concept names is finally obtained. These new artificial concept names have to be added in the table representing the heterogeneous source of information (see Sect. 3.2) and the process for discovering ARs has to be run (see Sect. 3.2). This is because potentially useful ARs for treating the disjuncts may be found. It is important to note that the artificial concept names are not used for the process of discovering new knowledge in itself (as illustrated in Sect. 3.2) but only for the reasoning purpose presented in this section.

Now let us consider the following example concerning the demographic domain where the starting point for the inference is given in Tab. 3. Note that it is assumed that $Parent$ is $true$ for x_2. In the following the expansion of $(Male \sqcup Female)(x)$ for each of the objects in Tab. 3 is illustrated:

Table 3. Demographic example: data given at the inference stage

Object	Job	Age	$Parent$	$Male$	$Female$
x_1	Primary school teacher	47	$unknown$	$unknown$	$unknown$
x_2	Policeman	25	$true$	$unknown$	$unknown$
x_3	Student	20	$unknown$	$unknown$	$unknown$

– x_1:

$degreeOfMatch(rule_2)$ = 0,

$degreeOfMatch(rule_3)$ = 0,
$degreeOfMatch(rule_4)$ = 1,
$degreeOfMatch(rule_5)$ = 0,
thus the $Female$ decision is taken with $weightedConf = 0.78$.

- x_2:
  $degreeOfMatch(rule_2)$ = 1,
  $degreeOfMatch(rule_3)$ = 1,
  $degreeOfMatch(rule_4)$ = 0,
  $degreeOfMatch(rule_5)$ = 0,
  for both $rule_2$ and $rule_3$ $weightedConf = 0.75$ so we have a conflict here, and a random decision is taken.
- x_3:
  $degreeOfMatch(rule_2)$ = 0,
  $degreeOfMatch(rule_3)$ = 0,
  $degreeOfMatch(rule_4)$ = 0.5,
  $degreeOfMatch(rule_5)$ = 0,
  thus for $rule_3$ $weightedConf = 0.75 * 0.5 = 0.375$, which is below a given threshold (let it be 0.75), and so a random decision is taken.

As processing a disjunct expansion we always add assertions coming from the evidence of the available knowledge, the proposed approach should ensure that the model built is the one mostly compliant with the statistical regularities learned from data.

5 Discussion and concluding remarks

To summarize, in this paper we make a step towards the framework for knowledge representation and reasoning in which knowledge is specified by a mix of logical formulas and data linked together. We revise the preliminary results we presented in [5] by explicitly taking into account the *Open World Assumption* made in DLs. Differently from [9], where federated DBs are considered with the goal of removing structural conflicts automatically while maintaining unchanged the views of the different DBs, we focus on building a new knowledge base that is able to collect the complementary knowledge that is contained in heterogeneous sources of information. Eventually, we propose a method for data-driven logical reasoning, which is the result of combining logical reasoning and data mining methods embedded in a Tableaux algorithm. Differently from [6], where an integrated system (*$\mathcal{AL}$-log*) for knowledge representation based on DL and the deductive database language Datalog is presented, here purely relational databases are considered. Additionally, while in [6] a method for performing query answering based on constrained resolution is proposed, where the usual deduction procedure defined for Datalog is integrated with a method for reasoning on the structural knowledge, here a more expressive DL is considered and *semantically enriched* ARs are also exploited.

Our proposed mixed inference imitates in a way the cognitive reasoning process performed by humans. Indeed a human usually performs a logic reasoning process when he/she has knowledge that is assumed to be complete for a certain domain, for example, the medical domain. This step is represented in our case by the standard deductive approach. If some degrees of uncertainty occur, for instance there are strange symptoms that do not allow for a straightforward diagnosis, then *existing cases* are analyzed to support a given diagnosis or an alternative one. The existing cases would represent our external source of information (DBs and/or ARs). The process for determining a diagnosis is now driven by the integration of the logic reasoning process and inductive reasoning process that takes into account the additional cases and tries to produce a reasonable diagnosis given the additional available evidence.

The *most plausible model* that we build may be enriched with additional knowledge coming from the external source of information. Specifically, given the selected rule for resolving a disjunction (see Sect. 3.3), the information on the left hand side concerning only the external source of information could be added as part of the model under construction thus applying a sort of abductive inference. Alternatively, this additional knowledge may be exploited during the matching process for preferring a rule rather than another one (besides of the condition concerning the confidence of the rule). Particularly, as an additional criterion the level of match between the rule and the model under construction may be considered. The same would be done for additional information on the right hand side of a rule even if this case appears to be a bit more problematic.

For the future we aim at implementing the extended Tableaux algorithm for experimental purpose.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data*, pages 207–216. ACM Press, 1993.
2. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 2001.
4. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. The mastro system for ontology-based data access. *Semantic Web Journal*, 2(1):43–53, 2011.
5. C. d'Amato, V. Bryl, and L. Serafini. Semantic knowledge discovery from heterogeneous data sources. In H. Stuckenschmidt et al., editor, *Proc. of 18th Int. Conf. on Knowledge Engineering and Knowledge Management, EKAW 2012*, page To appear, 2012.
6. F. M. Donini, M. Lenzerini, D. Nardi, and A. Schaerf. $\mathcal{AL}$-log: Integrating datalog and description logics. *Journal of Intelligent Information Systems*, 10:227–252, 1998.
7. S. Grimm, B. Motik, and C. Preist. Variance in e-business service discovery. In *Proceedings of the ISWC Workshop on Semantic Web Services*, 2004.
8. David J. Hand, Padhraic Smyth, and Heikki Mannila. *Principles of data mining*. MIT Press, Cambridge, MA, USA, 2001.
9. S. Spaccapietra and C. Parent. View integration: A step forward in solving structural conflicts. *IEEE TRANS. On Knowledge and Data Engineering*, 6(2):258–274, 1994.
10. Ian Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition*. Morgan Kaufman, 2011.