Querying with Negation in Data Integration Systems

Zoran Majkić

Dept. of Computer Science,UMIACS, University of Maryland, College Park, MD 20742
zoran@cs.umd.edu
http://www.cs.umd.edu/ ∼ zoran/

Abstract. Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data. It is characterized by an architecture based on a global schema, with the set of integrity constraints, and a set of sources. In this paper we investigate the way in which Closed World Assumption on a source data base can be coherently propagated to the global schema. The problem to resolve is directly connected by the fact that a global schema has a number (possibly infinite) of minimal models, caused by the incompleteness of source databases w.r.t. the integrity constraints over global schema. The aim of this preliminary work is to open the perspective for query language with negation in Data integration framework.

1 Introduction

It is well known that a relational database theory is complete under the closed-world (CWA), domain-closure (DCA), and unique-name (UNA) assumptions: any formula is either true or false. But it is desirable to broaden the relational database concept to allow the modelling of incomplete information. In fact, it is shown that, when the global schema contains integrity constraints, even of simple forms, the semantics of the data integration system is best described in terms of a set of databases, rather than a single one, and this implies that, even in the global-as-view approach [1], query processing is intimately connected to the notion of querying *incomplete databases*: a solution to an instance of the global schema may contain values that are not among the values of the source instance.

The *main goal* of this work is to extend the formalization of a data integration system based on the relational model with integrity constraints and the query-answering for *conjunctive queries* only, given in [1], to the query-answering *with negation* in query bodies also. To do that we need some technical preliminaries, and we will use some notation from [1].

It is known that a reasoning is non-monotonic when it is built from pieces of incomplete but time-sensitive information: a closed-world assumption is a *minimization convention* which circumscribe and restrict the positive *implicit* information that a set of incomplete knowledge may involve; it induces principles and techniques that are used for querying the knowledge base in a non-monotonic manner with respect to the evolution of its contents. In fact, by introducing the CWA in data integration system, we will see that the semantic consequence relation of the ordinary data integration system (defined as a first-order logic) will be weakened: these new weakened consequence relations characterize logic systems that allow the inference of statements that are satisfied in *some*

specific models of the premises (not in all models as required by the ordinary first-order logic).

The *source schema* in a data integration system is expressed in the relational model without integrity constraints. Consequently it has a unique minimal Herbrand model which is equal to the set of all ground atoms of its predicates (relations), and corresponds to the set of all tuples of its relations in a particular source database instance. In the closed world assumption each Yes/No query over source data base corresponds to the true/false value for such ground query formula.

The main issue of this paper is to investigate the way in which such assumption on a source data base can be *coherently propagated to the global schema* (with integrity constraints) which is a real object for formalization of user queries. Let us consider, for example, the peer-to-peer systems [2, 3] where each peer encapsulates a database integration system based on sources in Web: such peer agent (Abstract Object Type) presents to users only the alphabet of a global schema in order to be able to formalize the well defined (*conjunctive*) queries; all other components of this encapsulated data integration system are hidden. The problem is that the Yes/No user queries formalized by user may be naturally interpreted as true/false facts retrieved in the source databases. This rational reasoning is enforced by the fact that user Yes/No queries posed on a global schema are equivalently translated by a peer agent , by its rewriting algorithm methods [4], into Yes/No queries over source databases.

In what follows, the example of a CWA extension of the ordinary first-order logic theory for a data integration systems [5] will be considered for a particular kind of data integration systems: in GAV (global-as view) source-to-global database mapping and with global schema with key and foreign key constraints, which is the most useful practical example and is used in some important EU projects [4].

The existential quantifiers in existentially quantified conjunctive formulas can be removed by Skolemization: each existential variable is replaced by a constant, noted ω_i (with $\Omega = \{\omega_1, \omega_2, ...\}$), distinct from all other constants initially present in the database. Those new constants are commonly called *marked (labelled) null values* [6] and denote "unknown" values that are incompletely identified in databases.

We follow the following basic principle for a data integration systems, in the framework of 3-valued Kleene logic (with true, false and *unknown* logic values), with Coherent Closed-world assumption (CCWA):

1. The repairing of the incompleteness of source databases by means of marked null values in Ω, have to be hidden to users, i.e., true/false facts contains only real source database constants in Γ. Other facts are considered as *unknown*. This assumption we denote as $\Gamma - restriction$, and the negation introduced by CCWA as "*negation as* $\Gamma - restriction$ *failure*".

In this paper we *do not* consider the mutually inconsistent information in data sources: in fact, as we will see in technical preliminaries, we assume that key constraints are not violated by data integration system. The consideration of inconsistencies, for which we can't have a unique 3-valued canonical model for a global schema, fundamentally introduces the *4-valued logic* [7] and enriched well-founded models for data integration systems.

The plan of this work is the following: after the short introduction to the Data Integra-

tion Systems with the key and foreign key integrity constraints and the GAV (Global as view) mappings, in Section 2 we present one example of an infinite canonical database with Skolem functions for it, $can(\mathcal{I}, \mathcal{D})$, and the alternative version with Skolem constants, $can_M(\mathcal{I}, \mathcal{D})$: from the last one we can take a finite subset, $\mathcal{C}_M(\mathcal{I}, \mathcal{D})$, which is sufficient for the query answering. Finally, in Section 3 is given the definition of the Coherent Closed World Assumption for Data Integration Systems based on query rewriting over the source databases which satisfy standard CWA. We present also the alternative definition based on the canonical model of the global schema which may be used also for the case when we have no query-rewriting algorithms and have to materialize a global schema.
We show that these two approaches are equivalent and that are analog to the General Closed World Assumption (GCWA) for a logic programming.

1.1 Technical preliminaries for Data Integration

In the relational model, predicate symbols are used to denote the relations in the database, whereas constant symbols denote the values stored in relations. We assume to have a fixed (infinite) alphabet Γ of constants, and, if not specified otherwise, we will consider only databases over such an alphabet. In such a setting, the UNA *unique name assumption* (that is, to assume that different constants denote different objects) is implicit.
A *relational schema* (or simply *schema*) is constituted by:
1. An *alphabet* $\mathcal{A}$ of predicate (or relation) symbols, each one with the associated arity. i.e., the number of arguments of the predicate (or, attributes of the relation).
2. A set $\Sigma_{\mathcal{G}}$ of *integrity constraints*, i.e., assertions on the symbols of the alphabet $\mathcal{A}$ that express conditions that are intended to be satisfied in every database coherent with the schema.In this framework are considered two kinds of constraints:

2.1. *Key constraints* : we assume that in the global schema there is exactly one key constraint for each relation,

2.2. *Foreign key constraints*: a foreign key constraint is a statement of the form $r_1[\mathbf{A}] \subseteq r_2[\mathbf{B}]$, where r_1, r_2 are relations, $\mathbf{A}$ is a sequence of distinct attributes of r_1, and $\mathbf{B}$ is $key(r_2)$. Such a constraint is satisfied in a database $\mathcal{DB}$ if for each tuple t_1 in $r_1^{\mathcal{DB}}$ there exists a tuple t_2 in $r_2^{\mathcal{DB}}$ such that $t_1[\mathbf{A}] = t_2[\mathbf{B}]$, where $t_1[\mathbf{A}]$ is the projection of the tuple t_1 over $\mathbf{A}$.
A *relational database* (or simply, *database*) $\mathcal{DB}$ for a schema $\mathcal{C}$ is a set of relations with constants as atomic values, and with one relation $r^{\mathcal{DB}}$ of arity n for each predicate symbol r of arity n in the alphabet $\mathcal{A}$: the relation $r^{\mathcal{DB}}$ is the interpretation in $\mathcal{DB}$ of the predicate symbol r, in the sense that it contains the set of tuples that satisfy the predicate r in $\mathcal{DB}$.
A *relational query* is a formula that specifies a set of tuples to be retrieved from a database. We consider the class of *safe conjunctive queries*. The answer to a query q of arity n over a database $\mathcal{DB}$ for $\mathcal{G}$, denoted $q^{\mathcal{DB}}$, is the set of n-tuples of constants $(c_1, \ldots, c_n)$, such that, when substituting each x_i with c_i, the formula $\exists(y_1, \ldots, y_m).conj(x_1, \ldots, x_n, y_1, \ldots, y_m)$ evaluates to true in $\mathcal{DB}$.
Yes/No queries are particular case when $n = 1$ and all variables $y_i,\ 1 \leq i \leq m$ are

replaced by constants in Γ.
A **data integration system** $\mathcal{I}$ is a triple $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, where:

- The *global schema* $\mathcal{G}$ is expressed in the relational model with constraints $\Sigma_{\mathcal{G}}$.
- The *source schema* $\mathcal{S}$ is expressed without integrity constraints.
- The *mapping* $\mathcal{M}$ is defined following the global-as-view approach: to each relation r of the global schema $\mathcal{G}$ we associate a query $\rho(r)$ over the source schema $\mathcal{S}$.

We call *global database* for $\mathcal{I}$, or simply *database* for $\mathcal{I}$, any database for $\mathcal{G}$.
Let $\mathcal{D}$ be a finite *source database instance* for $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, which is constituted by one relation $r^{\mathcal{D}}$ for each source r in $\mathcal{S}$. A database $\mathcal{B}$ for $\mathcal{I}$ is said to be *legal* with respect to $\mathcal{D}$ if:

- $\mathcal{B}$ satisfies the integrity constraints $\Sigma_{\mathcal{G}}$ of $\mathcal{G}$;
- $\mathcal{B}$ satisfies $\mathcal{M}$ with respect to $\mathcal{D}$, i.e., for each relation r in $\mathcal{G}$, the set of tuples $r^{\mathcal{B}}$ that $\mathcal{B}$ assigns to r is a superset of the set of tuples $\rho(r)^{\mathcal{D}}$ computed by the associated query $\rho(r)$ over $\mathcal{D}$, i.e., $\rho(r)^{\mathcal{D}} \subseteq r^{\mathcal{B}}$.
- we denote by $sem^{\mathcal{D}}(\mathcal{I})$ the set of databases for $\mathcal{I}$ that are legal w.r.t. $\mathcal{D}$, i.e., that satisfy both the constraints of $\mathcal{G}$, and the mapping $\mathcal{M}$ with respect to $\mathcal{D}$. If $sem^{\mathcal{D}}(\mathcal{I}) \neq \emptyset$, then $\mathcal{I}$ is said to be *consistent* w.r.t. $\mathcal{D}$.

Note that the above definition amounts to consider any view $\rho(r)$ as *sound* [5, 8], which means that the data provided by the sources are only a (incomplete) subset (possibly proper) of the data that would satisfy the relations of the global schema.
By the above definition, it is clear that the semantics of a data integration system is formulated in terms of a *set* of databases, rather than a single one.
Retrieved global database $ret(\mathcal{I}, \mathcal{D})$, for a given a finite source database instance $\mathcal{D}$, is defined in the following way:
For each relation r of the global schema, we compute the relation $r^{\mathcal{D}}$ by evaluating the query $\rho(r)$ over the source database $\mathcal{D}$. We assume that for each relation r of the global schema, the query $\rho(r)$ over the source schema $\mathcal{S}$ that the mapping $\mathcal{M}$ associates to r preserves the key constraint of r (this may require that $\rho(r)$ implements a suitable duplicate record elimination strategy), so that the retrieved global database satisfies all key constraints in $\mathcal{G}$.
Query-answering in a Data Integration System:
Let q be a conjunctive query to a data integration system $\mathcal{I}$, that is, atoms in q have symbols in $\mathcal{G}$ as predicates.
The set of **certain answers** $q^{\mathcal{I},\mathcal{D}}$ to q w.r.t. $\mathcal{I}$ and $\mathcal{D}$ is the set of tuples t of constants of the same arity as q, and $t \in q^{\mathcal{B}}$, for each $\mathcal{B} \in sem^{\mathcal{D}}(\mathcal{I})$.
Let $\Delta_{\mathcal{I}}$ be a set of formulas belonging to a language $\mathcal{L}$ used to define data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ and a formula A belonging to $\mathcal{L}$. We denote by $\models$ a *model-theoretic consequence* relation (logical entailment), defined by the following way: $\Delta_{\mathcal{I}} \models A$ when all models of the formulas contained in $\Delta_{\mathcal{I}}$ are models of A; with $\models_{\mathcal{B}} A$ we denote that a formula A is true in a particular model $\mathcal{B}$.
Then, the certain answer to $q(\mathbf{x})$ is given in the following way:
$q^{\mathcal{I},\mathcal{D}} = \{\, t \mid t \in \Gamma^{arity(q)}$ and $\models_{\mathcal{B}} q(t)$ for each $\mathcal{B} \in sem^{\mathcal{D}}(\mathcal{I})\}$.
So, in this framework, with key and foreign key constraints, we can see a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, as a logical theory $\mathcal{P}_{\mathcal{G}}$ (a definite logic program in

[1]) composed by retrieved global database $ret(\mathcal{I}, \mathcal{D})$ as extensional part of a database theory and by the integrity constraints for a global schema as its intensional part: the existential quantifiers in foreign key constraints are eliminated by introducing appropriate Skolem functions:

$HT(\mathcal{G}) = \{f_{r,i} \mid r \in \mathcal{G}$ and $i \leq arity(r)$ and $i \notin key(r)\}$

Each $f_{r,i}$ is a function symbol with the same arity as the number of attributes of $key(r)$, i.e., $arity(f_{r,i}) = arity(key(r))$. Intuitively, the role of the term $f_{r,i}(\alpha_1, \ldots, \alpha_k)$ is to denote the value in the i-th column of the tuple of r having $\alpha_1, \ldots, \alpha_k$ in the key columns. The domain of such functions is the alphabet Γ.

In [1] is presented the construction of the canonical *model* for a global schema of this logical theory $\mathcal{P}_\mathcal{G}$. This model is universal (canonical) one, that is, *initial* minimal Herbrand model such that, for every other legal database model $\mathcal{B}$ of the global schema, there is a unique homomorphism $\psi : can(\mathcal{I}, \mathcal{D}) \longrightarrow \mathcal{B}$,

and are defined the following sound and complete query rewriting algorithms:

- $exp_\mathcal{G}(\_)$, which expands the original conjunctive query $q(\mathbf{x})$ over a global schema into $exp_\mathcal{G}(q(\mathbf{x}))$ query over $ret(\mathcal{I}, \mathcal{D})$.
- $unf_\mathcal{M}(\_)$ algorithm which unfold the resulting query over $ret(\mathcal{I}, \mathcal{D})$ and returns with the query $unf_\mathcal{M}(exp_\mathcal{G}(q(\mathbf{x})))$ over a source data base $\mathcal{D}$, such that for each tuple $t \in \Gamma^{arity(q)}$, $\quad t \in q(\mathbf{x})^{\mathcal{I},\mathcal{D}}$ iff $t \in [unf_\mathcal{M}(exp_\mathcal{G}(q(\mathbf{x})))]^\mathcal{D}$.

So that the certain answer to a conjunctive query $q(\mathbf{x})$ is equal to

$q(\mathbf{x})^{\mathcal{I},\mathcal{D}} = \{\, t \mid t \in \Gamma^{arity(q)}$ *and* $\models_{M_\mathcal{D}} unf_\mathcal{M}(exp_\mathcal{G}(q(t)))\,\}$,

where $M_\mathcal{D}$ is the unique minimal Herbrand model of a source database $\mathcal{D}$.

2 Finite Canonical Database

In this paragraph we introduce a new approach to canonical model, more close to the data exchange approach [9]. It is not restricted to existence of query-rewriting algorithms, thus can be used to define Coherent Closed World Assumption for data integration systems also in absence of query-rewriting algorithms. The construction of the canonical *model* for a global schema of the logical theory $\mathcal{P}_\mathcal{G}$ for data integration system is similar to the construction of the canonical *database* $can(\mathcal{I}, \mathcal{D})$ described in [1]. The *difference* lies in the fact that, in the construction of this revisited canonical model, denoted by $can_M(\mathcal{I}, \mathcal{D})$, for a global schema, fresh *marked null values* (set Ω of Skolem constants) are used instead of terms involving Skolem functions following the idea of construction of the restricted chase of a database described in [10]. Thus, we enlarge a set of constants of our language by $\Gamma_U = \Gamma \bigcup \Omega$.

R.W.Topor and E.A.Sonenberg informally proposed the term *canonical model* to describe a model that is selected (often from many incomparable minimal Herbrand models) to represent the 'meaning' of logical programs: one now has a standard for correctness of a proper interpreter on *all* goals- it must conform to the canonical model, and succeed or fail appropriately.

Another motivation for concentrating on canonical models is the view [11] that many logic programs are appropriately thought of as having two components, an *intensional* database (IDB) that represents the reasoning component, and the *extensional* database

(EDB) that represents a collection of facts. Over the course of time, we can "apply" the same IDB to many quite different EDBs. In this context it make sense to think of the IDB as implicitly defining a transformation from an EDB to a set of derived facts: we would like the set of derived facts to be the canonical model.
Now we construct inductively the revisited canonical database model $can_M(\mathcal{I}, \mathcal{D})$ over the domain Γ_U by starting from $ret(\mathcal{I}, \mathcal{D})$ and repeatedly applying the following rule:

> if $(x_1, \ldots, x_h) \in r_1^{can_M(\mathcal{I},\mathcal{D})}[\mathbf{A}]$, $(x_1, \ldots, x_h) \notin r_2^{can_M(\mathcal{I},\mathcal{D})}[\mathbf{B}]$, and the foreign key constraint $r_1[\mathbf{A}] \subseteq r_2[\mathbf{B}]$ is in $\mathcal{G}$,
> then insert in $r_2^{can_M(\mathcal{I},\mathcal{D})}$ the tuple t such that
> - $t[\mathbf{B}] = (x_1, \ldots, x_h)$, and
> - for each i such that $1 \leq i \leq arity(r_2)$, and i not in $\mathbf{B}$, $t[i] = \omega_k$, where ω_k is a fresh marked null value.

Note that the above rule does enforce the satisfaction of the foreign key constraint $r_1[\mathbf{A}] \subseteq r_2[\mathbf{B}]$, by adding a suitable tuple in r_2: the key of the new tuple is determined by the values in $r_1[\mathbf{A}]$, and the values of the non-key attributes are formed by means of the fresh marked values ω_k during the application of the rule above.
The rule above defines the "immediate consequence" monotonic operator T_B defined by:

$$T_B(I) = I \bigcup \{\ A \mid \ \ A \in B_{\mathcal{G}}, A \leftarrow A_1 \wedge .. \wedge A_n \text{ is a ground instance of a rule in } \Sigma_{\mathcal{G}} \text{ and } \{A_1, .., A_n\} \in I\ \}$$

where, at beginning $I = ret(\mathcal{I}, \mathcal{D})$, and $B_{\mathcal{G}}$ is a Herbrand base for a global schema. Thus, $can_M(\mathcal{I}, \mathcal{D})$ is a least fixpoint of this immediate consequence operator.

Example 1: Suppose that we have two relations r and s in $\mathcal{G}$, both of arity 2 and having as key the first attribute, and that the following dependencies hold on $\mathcal{G}$:
$r[2] \subseteq s[1]$, $\quad s[1] \subseteq r[1]$.
Suppose that the retrieved global database stores a single tuple (a, b) in r. Then, by applying the above rule, we insert the tuple (b, ω_1) in s; successively we add (b, ω_2) in r, then (ω_2, ω_3) in s and so on. Observe that the two dependencies are cyclic, and in this case the construction of the canonical database requires an infinite sequence of applications of the rules. The following table represents the correspondence between old and revisited canonical database:

$r^{can(\mathcal{I},\mathcal{D})}$	$s^{can(\mathcal{I},\mathcal{D})}$	$r^{can_M(\mathcal{I},\mathcal{D})}$	$s^{can_M(\mathcal{I},\mathcal{D})}$
a, b	$b, f_{s2}(b)$	a, b	b, ω_1
$b, f_{r2}(b)$	$f_{r2}(b), f_{s2}f_{r2}(b)$	b, ω_2	ω_2, ω_3
$f_{r2}(b), f_{r2}^2(b)$	$f_{r2}^2(b), f_{s2}f_{r2}^2(b)$	ω_2, ω_4	ω_4, ω_5
$f_{r2}^2(b), f_{r2}^3(b)$	$f_{r2}^3(b), f_{s2}f_{r2}^3(b)$	ω_4, ω_6	ω_6, ω_7
..	..	..	..

Thus, the canonical model $can_M(\mathcal{I}, \mathcal{D})$ is a legal database model for the global schema. Let us introduce an unary predicate $Val(x)$, such that for any constant $c \in \Gamma_U$, $Val(c)$ is true if $c \in \Gamma$, false otherwise. Each certain answer of the original user query $q(\mathbf{x})$,

$\mathbf{x} = \{x_1, .., x_k\}$ over a global schema is equal to the answer $q_L(\mathbf{x})^{can_M(\mathcal{I},\mathcal{D})}$ of the *lifted* query $q_L(\mathbf{x}) \equiv q(\mathbf{x}) \wedge Val(x_1) \wedge ... \wedge Val(x_k)$ over this canonical model. Thus, if would be possible to materialize this canonical model, the certain answers could be obtained over such an database. Usually it is not possible because (as in the example above) this canonical model is *infinite*: In that case, we can use the revisited fixpoint semantics described in [12], based on the fact that, after some point, the new tuples added into a canonical model insert only new Skolem constants which are not useful in order to obtain *certain* answers. In fact, Skolem constants are not part of any certain answer to conjunctive query. Consequently, we are able to obtain a *finite subset* of a canonical *database*, which is big enough to obtain certain answers.

Example 2: For Example 1, the finite database
$r = \{(a, b), (b, \omega_2), (\omega_2, \omega_4)\}$, $s = \{(b, \omega_1), (\omega_2, \omega_3)\}$, is such finite least fixpoint which can be used in order to obtain certain answers to lifted queries. □

In fact, we introduced marked null values (instead of Skolem functions) in order to define and materialize such *finite* database: it *is not* a model of the data integration system (which is infinite) but has all necessary query-answering properties: it is able to give all certain answers to conjunctive queries over global schema. Thus it can be materialized and used for query answering, instead of query-rewriting algorithms.

Let denote such finite database by $\mathcal{C}_M(\mathcal{I}, \mathcal{D})$. So, we can prove the following property:

Proposition 1 *Each Yes/No query* $q(\boldsymbol{c})$, $\boldsymbol{c} = \{c_1, ..c_n\} \in \Gamma^n, n = arity(q)$ *over a canonical database* $\mathcal{C}_M(\mathcal{I}, \mathcal{D}) \subseteq can_M(\mathcal{I}, \mathcal{D})$ *returns the same logical true/false value, as the rewritten query* $unf_{\mathcal{M}}(exp_{\mathcal{G}}(q(\boldsymbol{c})))$ *over a source database* $\mathcal{D}$.
There is a unique homomorphism $\varphi : can_M(\mathcal{I}, \mathcal{D}) \longrightarrow can(\mathcal{I}, \mathcal{D})$.

Proof. The first part of this proposition is a direct consequence of the query rewriting algorithms and the fact that each Yes/No query over a source database $\mathcal{D}$ and over a canonical database $can_M(\mathcal{I}, \mathcal{D})$ returns with the true/false value. Let us prove the second part: from a definition of certain answers we have that $q^{\mathcal{I},\mathcal{D}} = \{ t \mid t \in \Gamma^{arity(q)}$ and $\models_{\mathcal{B}} q(t)$ for each $\mathcal{B} \in sem^{\mathcal{D}}(\mathcal{I})\}$. Thus, $t \in q^{\mathcal{I},\mathcal{D}}$ (or, equivalently, $\models_{can_M(\mathcal{I},\mathcal{D})} q(t)$) iff $\models_{\mathcal{B}} q(t)$ for each $\mathcal{B} \in sem^{\mathcal{D}}(\mathcal{I})$, but, $\models_{can_M(\mathcal{I},\mathcal{D})} q(t)$ iff $\models_{M_{\mathcal{D}}} unf_{\mathcal{M}}(exp_{\mathcal{G}}(q(t)))$.

Example 3: For Example 1, we define φ homomorphism as follows: for any constant in Γ, it is an identity function (this is an intrinsic homomorphism property).
For Skolem constants we have that $\varphi(\omega_1) = f_{s2}(b)$, $\varphi(\omega_{2i}) = f_{r,2}^{2i-1}(b)$ and $\varphi(\omega_{2i+1}) = f_{s,2}(\omega_{2i}) = f_{s,2}f_{r,2}^{2i-1}(b)$, for $i = 1, 2, ..$ □

3 Closed-world assumption

Non-monotonicity is used dynamically for 'jumping to conclusions' when the available information is incomplete. Logic programs with non-monotonic negation constitute a small yet quite expressive class of non-monotonic logic, which is of particular interest because *they are implementable*.

In this section we will consider a non monotonic version for a data integration system with a weakened Generalized Closed-world Assumption (GCWA) [13] generally

adopted for a logical database theories with more than one minimal Herbrand model (the usual CWA in the case when the logical theory has more than one minimal Herbrand model does not preserve the consistency).
In order to formalize the logical theory for data integration systems which involves the incompleteness of source databases, we introduce also a set of Skolem functions $\{f_1, f_2, ..\}$, with arity grater than zero, in our logical language. So, in this enlarged context of the set of constants of the database, with Skolem functions and language constants in $\Gamma_U = \Gamma \bigcup \Omega$, we are able also to assume DCA (domain-closure assumption) as the following axiom:
$\forall x((x = c_1) \vee .... \vee (x = c_n) \vee (\exists x_1, .., x_k(x = f_1(x_1, .., x_k)) \vee .... \vee (\exists y_1, .., y_j(x = f_m(y_1, .., y_j)))$ where $c_1, ..., c_n$ are all the individual constants in Γ_U and $f_1, f_2, ..$ are introduced Skolem functions.
The valid *interpretations* of these function symbols, in different models for a global schema, are functions $f_{r,i} : \Gamma_U^{arity(r)} \longrightarrow \Gamma_U$.
Notice that they are different from the set $HT(\mathcal{G})$ of Skolem functions of the paragraph 1.1, where the domain is over Γ.
From semantic point of view, the DCA implies considering Herbrand models only. The role of UNA and DCA is to allow us to consider models the interpretation domain of which stands in bijection with the set of individual constants of Γ_U. (In order to axiomatize the reasoning that is involved we must introduce the axioms that represent the properties of the equality predicate, i.e., reflexivity, symmetry and transitivity property, and the principle of substitution of equal terms).

3.1 Source database approach

Let $' \sim '$ be the 'negation as $\Gamma - restriction$ failure' operator for a CWA. For any ground formula $\psi(t),\ t \in \Gamma^{arity(\psi)}$ we denote by $\models_{M_\mathcal{D}} \sim \psi(t)$ the fact that $'\psi(t)$ is false for a source database', w.r.t. the Closed World Assumption for *source* databases. Since only positive information can be a logical consequence of a data integration system, special rules are needed to deduce negative information. We define a Coherent Closed-world assumption rule (CCWA) for our logical theory $\mathcal{P}_\mathcal{G}$, which coherently propagates CWA of a *source* database into a Data Integration System:

Definition 1. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a Data integration system, with key and foreign key integrity constraints, and with the (unique) retrieved database which satisfies all key constraints in $\mathcal{G}$.*

- *Under the Coherent Closed-world assumption the following negative atoms can be infered*
  $CCWA(\mathcal{G}) =_{def} \{ \sim r(\boldsymbol{c}) \mid r \in \mathcal{G}, \boldsymbol{c} \in \Gamma^{arity(r)}$ *and* $\models_{M_\mathcal{D}} \sim unf_\mathcal{M}(exp_\mathcal{G}(r(\boldsymbol{c}))) \}$
- *we introduce the database logic theory*
  $\mathcal{I}_{CWA} = \mathcal{P}_\mathcal{G} \bigcup CCWA(\mathcal{G}) \bigcup UNA \bigcup DCA$

In the rest of this paper we will prove that this assumption rule corresponds to the weak form of Generalized Closed-world Assumption (GCWA) [13]; in fact CCWA is weaker than GCWA and, like in Prolog, infer less negative atoms then GCWA - in GCWA can

exist the false atoms with also marked null values, while in CCWA false atoms can have only real constants (defined in source database) in Γ. Thus CCWA can be also called $\Gamma - restriction$-GCWA.
Let $q(\mathbf{c}), \mathbf{c} \in \Gamma^{arity(q)}$ be any conjunctive *ground* query over a *global* schema;
the denotation $\models\sim q(\mathbf{c})$ means that $q(\mathbf{c})$ is false in all minimal models of a global schema. Thus, the $\Gamma - restriction$-GCWA assumption (or, equivalently, CCWA), in our case, may be expressed by the following definition:
$\models\sim q(\mathbf{c})$ iff $\models_{M_\mathcal{D}}\sim unf_\mathcal{M}(exp_\mathcal{G}(q(\mathbf{c})))$.

Proposition 2 *Let $q(\boldsymbol{x})$ be any conjunctive query over a global schema. The certain answer to the query $\sim q(\boldsymbol{x})$ is given by:*
$(\sim q(\boldsymbol{x}))^{\mathcal{I},\mathcal{D}} =_{def} \{\boldsymbol{c} \mid \boldsymbol{c} \in \Gamma^{arity(q)} \; and \; \models_{M_\mathcal{D}}\sim unf_\mathcal{M}(exp_\mathcal{G}(q(\boldsymbol{c})))\}$
$= \Gamma^{arity(q)} - q(\boldsymbol{x})^{\mathcal{I},\mathcal{D}}$.

Proof. From the definition of the certain answer in a data integration systems, we have that $(\sim q(\mathbf{x}))^{\mathcal{I},\mathcal{D}} =_{def} \{\mathbf{c} \mid \mathbf{c} \in \Gamma^{arity(q)} \; and \; \models_\mathcal{B}\sim q(\mathbf{c})$ for each $\mathcal{B} \in sem^\mathcal{D}(\mathcal{I})\}$ $= \{\mathbf{c} \mid \mathbf{c} \in \Gamma^{arity(q)} \; and \; \models\sim q(\mathbf{c})\}$. From GCWA holds $\models\sim q(\mathbf{c})$ iff $\models_{M_\mathcal{D}}\sim unf_\mathcal{M}(exp_\mathcal{G}(q(\mathbf{c})))$, thus we obtain the result of the proposition.

From this proposition we obtain that each ground query with constants in Γ (without marked null values) is true or false *in all* (preferred) models of the logical theory $\mathcal{I}_{CWA}$.

3.2 Canonical (finite) database approach

A non-monotonic reasoning underlies inferences in a *specific* set of models of the premises. In standard first-order logic, a conclusion can be inferred when it is true in *all* the models of the premises. In a non-monotonic logic, a formula can be inferred when it is true in *some specific* models of the premises. These specific models are called *preferred models* [14]. By definition, the preferred models determine the specific *rationality* of the agent who conducts the reasoning: consider for example a Brokering agent [4, 15] for some peer (encapsulated data integration system).
Such rationality in choosing preferred models, determined by GCWA, is concerned with minimization conventions about the implicit positive information that can be associated with the premises: each ground query (with only ordinary database constants in Γ) over global database returns with the same truth value as a query rewritten over source database instance $\mathcal{D}$ such that answer is:
- *Yes*, if it is true in all preferred models
- *No*, if it is false in all preferred models
- *Unknown*, otherwise

Note that in this particular case, no one Yes/No user-query (with ordinary constants in Γ) will return with value "unknown". Thus unknown facts are hidden to users (theoretically, if we are able to make ground queries with also marked null values, the system has to answer by value "unknown").
The problem is to find the *right restriction* for Skolem functions, in order to obtain the prefered models that determine the Coherent Closed World rationality for the agent who conducts the reasoning.

We can use the finite canonical database $\mathcal{C}_M(\mathcal{I}, \mathcal{D})$, which is obtained without consideration of Skolem functions, in order to define such restrictions. Thus, in our case, we restrict attention only to minimal Herbrand models of a global schema where interpretations of Skolem function symbols in the database theory satisfy the following SFA assumption:

Definition 2. *Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a Data integration system, with key and foreign key integrity constraints, and with the (unique) retrieved database which satisfies all key constraints in $\mathcal{G}$. We introduce*

- *Skolem Functions Assumption (SFA)* $\forall f_{r,i}(\, imf_{r,i} \subseteq \Omega \bigcup r^{\mathcal{C}_M(\mathcal{I},\mathcal{D})}[i] \,)$.
  This assumption restrict the original set of minimal models $M_{\mathcal{P}_\mathcal{G}}$, to the set
  $M_{SFA} = \{\mathcal{B} \mid \mathcal{B} \in M_{\mathcal{P}_\mathcal{G}} \; and \; \vDash_{\mathcal{B}} \forall f_{r,i}(\, imf_{r,i} \subseteq \Omega \bigcup r^{\mathcal{C}_M(\mathcal{I},\mathcal{D})}[i] \,)\} \subseteq M_{\mathcal{P}_\mathcal{G}}$,
  so that
  $SFA(\mathcal{G}) =_{def} \{\sim r(\boldsymbol{c}) \mid r \in \mathcal{G}, \boldsymbol{c} \in \Gamma^{arity(r)} \; and \; \forall(\mathcal{B} \in M_{SFA})(r(\boldsymbol{c}) \notin \mathcal{B})\}$
- *the database logic theory* $\mathcal{I}_{SFA} = \mathcal{P}_\mathcal{G} \bigcup SFA(\mathcal{G}) \bigcup UNA \bigcup DCA$

The next theorem confirms that such restriction on interpretations (models for a global schema) is necessary and sufficient condition for the CCWA:

Theorem 1 *The CCWA assumption is equivalent to the SFA assumption, and holds*
$SFA(\mathcal{G}) = \{\sim r(\boldsymbol{c}) \mid r \in \mathcal{G}, \boldsymbol{c} \in \Gamma^{arity(r)} \; and \; r(\boldsymbol{c}) \notin r^{\mathcal{C}_M(\mathcal{I},\mathcal{D})}\}$.

Proof. Consider that for any $r \in \mathcal{G}$, $\Gamma \bigcap r^{\mathcal{C}_M(\mathcal{I},\mathcal{D})}[i] = \Gamma \bigcap r^{can_M(\mathcal{I},\mathcal{D})}[i]$, $1 \leq i \leq arity(r)$, and that $can_M(\mathcal{I}, \mathcal{D}) \in M_{SFA} \subseteq M_{\mathcal{P}_\mathcal{G}}$.
Let prove that CCWA implies SFA. Suppose that there exist some model M_0 of $\mathcal{I}_{CWA}$ with the interpretation of a function $f_{r,i}$ such that $imf_{r,i} \subsetneq r^{\mathcal{C}_M(\mathcal{I},\mathcal{D})}[i] \bigcup \Omega$, that is $\exists c \in \Gamma$ such that $c \in imf_{r,i}$ and $c \notin r^{\mathcal{C}_M(\mathcal{I},\mathcal{D})}[i]$, thus $c \notin r^{can_M(\mathcal{I},\mathcal{D})}[i]$. For such model M_0 the lifted query $q(x_i) \equiv r(x_i) \wedge Val(x_i)$ will return also with the atom $q(c)$ which is not part of the certain answer, thus $q(c)$ is false for $\mathcal{I}_{CWA}$ (that is, it is false in all models of $\mathcal{I}_{CWA}$), and, as consequence, false also for M_0, which is a contradiction. Thus all models of $\mathcal{I}_{CWA}$ satisfy SFA
Let prove that SFA implies CCWA. Suppose that it does not hold, that is, there exists some $\sim r(\mathbf{c})$, $\mathbf{c} \in \Gamma^{arity(r)}$, which is true in some model M_1 of $\mathcal{I}_{SFA}$, for which holds $\vDash_{M_\mathcal{D}} unf_{\mathcal{M}}(exp_{\mathcal{G}}(r(\mathbf{c})))$. In that case holds also $\vDash_{can_M(\mathcal{I},\mathcal{D})} r(\mathbf{c})$. Let prove that $\vDash r(\mathbf{c})$, i.e. that $r(\mathbf{c})$ exists in every model of $\mathcal{I}_{SFA}$: $can_M(\mathcal{I}, \mathcal{D})$ is the canonical model of the theory $\mathcal{I}_{SFA}$, and, consequently, there exist the unique homomorphism (which is identity function for constants in Γ) $\psi\varphi : can_M(\mathcal{I}, \mathcal{D}) \rightarrow M$, i.e. $r(\mathbf{c}) \in M$, for each model $M \in M_{\mathcal{P}_\mathcal{G}}$, and, consequently, for each model M of $\mathcal{I}_{SFA}$. Thus $r(\mathbf{c}) \in M_1$ which is in contradiction with the hypothesis.
Thus all models of $\mathcal{I}_{SFA}$ satisfy CCWA.

The different approaches of these two equivalent assumptions can be explained as follows: CCWA approach is based on the *query rewriting algorithms* in Data integration systems (negative facts are derived from source database explicitly) , while SFA approach is based on the *finite canonical database* of the global schema of the Data integration systems (i.e. also for a data exchange systems where a global database have to be materialized). Thus,
$SFA(\mathcal{G}) = CCWA(\mathcal{G})$.

Theorem 2 *The following properties hold:*

- $\mathcal{I}_{SFA}$ *is a* non-monotonic *theory which satisfy the* $\Gamma - restriction$ *Generalized closed-world assumption GCWA, such that the set of its minimal Herbrand models coincide with the preferred subset of minimal legal Herbrand models of the ordinary first-order Data integration system theory* $\mathcal{P}_{\mathcal{G}}$.
- $\mathcal{I}_{SFA}$ *is a* sound *theory: each certain answer* $q(\boldsymbol{x})^{\mathcal{I},\mathcal{D}}$ *to some conjunctive query* $q(\boldsymbol{x})$*, of a Data integration system, is deducible from* $\mathcal{I}_{SFA}$ *, that is,* $q(\boldsymbol{x})^{\mathcal{I},\mathcal{D}} = \{\, \boldsymbol{c} \mid \boldsymbol{c} \in \Gamma^{arity(q)} \; and \; \mathcal{I}_{SFA} \vdash q(\boldsymbol{c})\}$.
- $\mathcal{I}_{SFA}$ *is a* Γ-complete *theory: each ground formula* $q(\boldsymbol{c})$ *with constants* $\boldsymbol{c} \in \Gamma^{arity(q)}$ *is a true or false formula, i.e.* $\mathcal{I}_{SFA} \vdash q(\boldsymbol{c}) \;\vee\; \mathcal{I}_{SFA} \vdash\sim q(\boldsymbol{c})$. *Each ground formula, with marked null values also, is an unknown answer.*

In the Closed-world data integration theory $\mathcal{I}_{SFA}$, of any data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, we can use also negation: a conjunctive query composed by positive literals and negative literals in its body.
In fact, we can extend the global schema with a new "complement" relation $\widehat{r}(\mathbf{x})$ for each original relation $r(\mathbf{x})$, with $\widehat{r}(\mathbf{x})^{\mathcal{I},\mathcal{D}} = \Gamma^{arity(r)} - r(\mathbf{x})^{\mathcal{C}_M(\mathcal{I},\mathcal{D})}$. Thus, each negative literal in the original conjunctive query, $q(\mathbf{x})$, can be replaced by the positive literal of its "complemented" atom, and the answer to such transformed query can be retrieved from this extend canonical database.
Note that, for a particular case, given any ordinary conjunctive (without negation in its *body*) query $q(\mathbf{x})$, the query $\sim q(\mathbf{x})$ will return with set of tuples that are false in all minimal Herbrand models of $\mathcal{I}_{CWA}$, that is in all preferred legal models of a global schema in Data integration framework, so

$$\begin{aligned}(\sim q(\mathbf{x}))^{\mathcal{I},\mathcal{D}} &= \{\, \mathbf{c} \mid \mathbf{c} \in \Gamma^n \; and \; \mathcal{I}_{SFA} \vdash\sim q(\mathbf{c})\} \\ &= \Gamma^{arity(q)} - unf_{\mathcal{M}}(exp_{\mathcal{G}}(q(\mathbf{x})))^{\mathcal{I},\mathcal{D}}\end{aligned}$$

thus, in this particular case, we are able to use the ordinary query-rewriting algorithms (for positive conjunctive queries) also, in order to obtain the answer.

4 Conclusion

As this paper has shown, the problem of answering to conjunctive queries, with negative literals in its body also, in Data integration framework is possible under Coherent Closed World Assumption (CCWA) . A non-monotonic reasoning underlies inferences in a specific set of preferred models of the premises: the Skolem function assumption (SFA) introduced into a logic theory of a Data Integration System (DIS) specifies such set of preferred models, and is verified that each canonical database of such logic theory, with the specific assignment of Skolem constants, is just one of such preferred models. Thus, the CCWA axiom, based on query-rewriting algorithms, is equivalent to the SFA assumption, based on the canonical model considerations.
For the Data exchange systems, where the canonical models are really developed, the conjunctive queries with negative literals can be directly computed. For a DIS with key and foreign key integrity constraints , where are used the query-rewriting algorithms with polynomial data complexity, it is possible to obtain only answers to queries $\sim q(\mathbf{x})$, where $q(\mathbf{x})$ is a conjunctive positive query.

More generally, to deal with this problem, we are investigating suitable extensions of these algorithms for conjunctive queries with negative literals also.
A non-monotonic reasoning in query-answering, caused by CGWA, that is, on a specific set of preferential models, and the *cumulative* non-monotonic reasoning [16], caused by preferred repairs of DISs, both demonstrate that a general nature of query-answering in DIS is a non-monotonic.

References

1. A.Calì, D.Calvanese, G.De Giacomo, and M.Lenzerini, "Data integration under integrity constraints," in *Proc. of the 14th Conf. on Advanced Information Systems Engineering (CAiSE 2002)*, 2002, pp. 262–279.
2. Z. Majkić, "Weakly-coupled P2P system with a network repository," *6th Workshop on Distributed Data and Structures (WDAS'04), July 5-7, Lausanne, Switzerland*, 2004.
3. Z. Majkić, "Massive parallelism for query answering in weakly integrated P2P systems," *Workshop GLOBE 04, August 30-September 3,Zaragoza, Spain*, 2004.
4. M.Lenzerini and Z. Majkić, "General framework for query reformulation," *Semantic Webs and Agents in Integrated Economies, D3.1, IST-2001-34825, February,*, 2003.
5. Maurizio Lenzerini, "Data integration: A theoretical perspective.," in *Proc. of the 21st ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2002)*, 2002, pp. 233–246.
6. Raymond Reiter, "Towards a logical reconstruction of relational database theory," in *On Conceptual Modeling: Perspectives from Artificial Intelligence Databases and Programming Languages*, M. L. Brodie, J. Mylopoulos, and J. W. Schmidt, Eds. Springer, 1984.
7. Z. Majkić, "Many-valued intuitionistic implication and inference closure in a bilattice based logic," *35th International Symposium on Multiple-Valued Logic (ISMVL 2005), May 18-21, Calgary, Canada*, 2005.
8. Alon Y. Halevy, "Answering queries using views: A survey," *Very Large Database J.*, vol. 10, no. 4, pp. 270–294, 2001.
9. R.Fagin, P.G.Kolaitis, R.J.Miller, and L.Popa, "DATA Exchange: Semantics and query answering," in *Proc. of the 9th Int. Conf. on Database Theory (ICDT 2003)*, 2003.
10. David S. Johnson and Anthony C. Klug, "Testing containment of conjunctive queries under functional and inclusion dependencies," *J. of Computer and System Sciences*, vol. 28, no. 1, pp. 167–189, 1984.
11. Raymond Reiter, "On closed world data bases," in *Logic and Databases*, Hervé Gallaire and Jack Minker, Eds., pp. 119–140. Plenum Publ. Co., New York, 1978.
12. Z. Majkić, "Fixpoint semantic for query answering in data integration systems," *AGP03 - 8.th Joint Conference on Declarative Programming, Reggio Calabria*, pp. 135–146, 2003.
13. Jack Minker, "On indefinite data bases and the closed world assumption," in *Proc. of the 6th Int. Conf. on Automated Deduction (CADE'82)*, Springer, Ed., 1982, vol. 138 of *Lecture Notes in Computer Science*.
14. P.Besnard and P.Siegel, "The preferential-models approach to non-monotonic logics," *Non-Standard Logics for Automated Reasoning, Academic Press, London*, pp. 137–161, 1988.
15. D. Beneventano, M.Lenzerini, F.Mandreoli, and Z. Majkić, "Techniques for query reformulation, query merging, and information reconciliation-part A," *Semantic Webs and Agents in Integrated Economies, D3.2.A, IST-2001-34825*, 2003.
16. Z. Majkić, "Plausible query-answering inference in data integration," *18th International Florida Artificial Intelligence Conference (FLAIRS 2005), May 15-17, Clearwater Beach, USA*, 2005.