Querying Partially Sound and Complete Data Sources

Alberto O. Mendelzon
Department of Computer Science
University of Toronto
mendel@db.toronto.edu

George A. Mihaila*
Department of Computer Science
University of Toronto
georgem@db.toronto.edu

ABSTRACT

When gathering data from multiple data sources, users need uniform, transparent access to data. Also, when extracting data from several independent, often only partially sound and complete data sources, it is useful to present users with meta-information about the confidence in the answer to a query, based on the number and quality of the sources that participated in constructing the answer. We consider the problem of querying collections of sources with incomplete and partially sound data. We provide a method for checking the consistency of a source collection, we give a tableaux-based characterization for the set of possible worlds consistent with a given source collection and we propose a probabilistic semantics for query answers.

1. INTRODUCTION

When gathering data from multiple, independent data sources, users need an integrated view of the data. On the data access level, users want uniform access to the data; they should not need to worry where the relevant data sources are, what protocols for data access they use, and how they model the information. On the semantic level, users want to make sense of the data, to have the data presented to them in a uniform way. Also, when extracting data from several independent, often overlapping and inconsistent, data sources, it is useful to present users with meta-information about the confidence in the answer to a query, based on the number and quality of the sources that participated in constructing the answer.

In this paper, we examine some issues related to the integration of data from multiple sources, in the presence of completeness and soundness information. We provide a method for checking the consistency of a source collection and we propose a probabilistic semantics for query answers. This research generalizes the work of Grahne and Mendelzon [6] by considering arbitrary completeness and soundness estimates (values in [0,1] as opposed to just 0 or 1).

*Current affiliation: IBM T.J. Watson Research Center

1.1 Motivating Example

Consider a system that integrates information from several data sources. As a concrete case, take for example the Global Historical Climatology Network [4]. This organization collects and assembles climatic data from about 6,000 temperature stations, 7,500 precipitation stations, and 2,000 pressure stations. The earliest station data is from 1697(!) and the most recent from 1990. The domain is modeled by a global relational schema, containing one relation for each type of measurement (temperature, pressure, precipitation, *etc.*) and several additional relations storing station locations and other geographic information. For example, mean monthly temperature is recorded in a relation *Temperature*(*station, year, month, value*). Conceptually, we would like the *Temperature* relation to contain all the mean temperatures for all stations and all months between 1697 and 1990. In reality, only partial information is assembled from several data sources. For example, a data source S_1 contains station data for Canada since 1900, another source, S_2, data for American stations since 1800, another, S_3, only data for station number '438432', and so on. Assume also that all station location information is maintained in a relation *Station*(*id, latitude, longitude, country*) by a single source S_0. We can model the contents of these sources by a view over the global schema (we use the conjunctive query notation from [2]):

$$S_0 : V_0(s, lat, lon, c) \leftarrow Station(s, lat, lon, c)$$
$$S_1 : V_1(s, y, m, v) \leftarrow Temperature(s, y, m, v), \\ Station(s, lat, lon, \text{“Canada”}), After(y, 1900)$$
$$S_2 : V_2(s, y, m, v) \leftarrow Temperature(s, y, m, v), \\ Station(s, lat, lon, \text{“US”}), After(y, 1800)$$
$$S_3 : V_3(438432, y, m, v) \leftarrow Temperature(438432, y, m, v)$$
$$\ldots$$

where we assume *After* is a built-in global relation.

We consider the above view definitions as describing the *intended* content of the sources, and assume the actual content to be an approximation of it. For example, S_1 might very well contain only *some* of the temperatures and also some of the values in S_1 might be incorrect. If all the tuples in S_1 are correct, we say that the source is *sound*. If source S_1 contains all Canadian station data since 1900 we say that the source is *complete*. These notions are relative to the (unknown) complete relation *Temperature* (and to the com-

plete relations *Station* and *After*). In many situations, it is the case that a source cannot claim either, but it can provide estimates of how much of the information is covered (completeness) and how much of the information is accurate (soundness). We formally introduce these notions in Section 2.

1.2 Related Work

The data integration problem has received considerable attention in the database community. Starting from the older problem of answering queries using materialized views [11], a formalism that treats individual sources as views over a global schema has been developed for data integration. The Information Manifold project [9] gives an algorithm for computing answers to queries posed over the global schema.

Motro assumes the existence of a "real world" global database and considers data sources as approximations of it [12]. Starting from this assumption, he introduces the notions of *sound* and *complete* answers to a query: an answer given by the multidatabase system to a query q is *sound* if it is included in the (hypothetical) answer to the same query computed over the real world database and *complete* if it includes the answer computed over the real world database.

Grahne and Mendelzon [6] take a different view: instead of assuming the existence of a unique "real world" global database, they consider the uncertainty introduced by multiple sources and define a set of possible global databases consistent with a collection of sources. In their work, they consider a collection of sources where some of them are sound, some are complete, and some are both sound and complete, and give upper and lower approximations to query answers (also known as the *certain* and the *possible* answers). In particular, they prove that the answer computed by the Information Manifold algorithm coincides with the certain answer. Abiteboul and Duschka [1] consider the special case when all sources are sound or all are sound and complete.

Levy considers sources that are known to be complete for some subset of their domain and shows how they can be used to compute exact answers [10]. Also relevant is Florescu et al.'s work [3] in using probabilistic knowledge in data integration: there, information about the completeness and relative overlap of data sources is used in ordering the accesses to sources in order to maximize the likelihood of obtaining answers early in the evaluation.

Grädel, Gurevich and Hirsch [5] model uncertainty by an *observed* database together with an *error probability function*. This function assigns to each atomic fact the probability that its truth value in the observed database differs from its truth value in the *actual* database. Our approach assumes completeness and soundness information is available at the source level and infers the confidence of individual tuples.

Kifer and Li [8] define a formalism for incorporating uncertainty into expert systems. They introduce a general framework for propagating confidence values in the evaluation of a logic program. We define a similar method for computing the confidence of answer tuples for relational algebra queries.

In this paper, we generalize Grahne and Mendelzon's approach [6] to collections of sources with partial completeness and soundness.

2. THE MODEL

2.1 Global Databases and View Definitions

Let $\mathbf{rel} = \{R_1, R_2, \ldots\}$ be an infinite set of *global relation names*, and $\mathbf{loc} = \{V, V_1, V_2, \ldots\}$ an infinite set of *local relation names*. Also, consider a set $\mathbf{dom} = \{a_1, a_2, \ldots\}$ of *constants* and a set $\mathbf{var} = \{x_1, x_2, \ldots\}$ of *variables*. Associated with each relation name R is an integer k called the *arity* of R. An *atom* is an expression of the form $R(e_1, \ldots, e_k)$ where R is a relation name, k is the arity of R, and $e_1, \ldots, e_k$ are either constants or variables. A *fact* is an atom without variables.

A *global schema* is a set $\mathbf{R} = \{R_1, R_2, \ldots, R_n\}$ of global relation names. A *global database* D over $\mathbf{R}$ is a finite set of facts, each fact being over some $R_i \in \mathbf{R}$. For a fixed global database D and relation name R_i, we denote by $D(R_i)$ the collection of all the facts over R_i in D, sometimes referred to as the *extension* of R_i in D.

We model the contents of a data source by a view definition φ of the form:

$$head(\varphi) \leftarrow body(\varphi),$$

where $head(\varphi)$ is an atom over a local relation name V and $body(\varphi)$ is a sequence $b_1, b_2, \ldots, b_n$ of atoms over global relation names. We assume that all queries are *safe* (all variables in the head also occur in the body). For a fixed global database D, the result of applying φ to D, denoted $\varphi(D)$ is a collection of facts over V.

A *view extension* for a view φ is a finite set of atoms over the local relation name V in the head of the view definition φ. We will denote such a view extension by v. A view extension corresponds to the current contents of a data source.

2.2 Completeness and Soundness of Data Sources

Consider a source S defined by a view φ and containing the view extension v. For a given global database D we say that the source S is *sound* with respect to D if $v \subseteq \varphi(D)$, and that it is *complete* with respect to D if $v \supseteq \varphi(D)$. If a source is both sound and complete, we say that the source is *exact* with respect to D.

The following two definitions formally introduce the completeness and soundness measures.

DEFINITION 2.1. *[Completeness] The completeness of source S with respect to a database D is the fraction of the tuples in $\varphi(D)$ that are in v:*

$$c_D(S) = \frac{|v \cap \varphi(D)|}{|\varphi(D)|}$$

DEFINITION 2.2. *[Soundness] The soundness of source v with respect to a database D is the fraction of tuples in v which are present in $\varphi(D)$:*

$$s_D(S) = \frac{|v \cap \varphi(D)|}{|v|}$$

The completeness measure can be used in query evaluation to select sources that are most likely to contain relevant information. The soundness parameter can be used for assessing the confidence we can place in the answers provided by individual data sources and to evaluate the confidence of an answer assembled from several sources.

These parameters can be hard to estimate in practice, given that a sound and complete database is usually not available. Nevertheless, in many instances, one can use various domain-specific methods to estimate lower bounds on the soundness and completeness of a data source. For example, in accounting information systems, data analysts use statistical methods for determining whether data is free of specific types of errors at a given level of confidence [7]. The methodology includes analyzing the data processing flow and possible sources of errors at various points and checking samples of sufficient size (the sample size is inferred from the desired confidence using statistical models). Also, in the case of the climatology data, one can compute the exact size of a complete database D (number of stations $\times$ total number of months), the size of each $\varphi_i(D)$, and can also use statistical methods to detect which temperatures are abnormally high (or low) for specific stations and months[1]. This situation can be generalized to any relation $R(A_1, \ldots, A_k)$ where there exists a functional dependency $A_1, \ldots, A_l \rightarrow A_{l+1}, \ldots, A_k$ and the domains of the determining attributes $A_1, \ldots, A_l$ are known (and finite). This is a very common case for data derived from measurements of physical variables.

The completeness and soundness measures are related to the *recall* and *precision* measures used in Information Retrieval [13]. An information retrieval system — operating on a (large) collection of documents — is able to produce, in response to a user query, a subset of documents "relevant" to that query. The effectiveness of such a system is typically estimated by comparing the answers computed by the system against the "correct" answers (as compiled by a team of human experts). In that context, the *recall* is the fraction of the documents from the "correct" answer that are returned by the system, and the *precision* is the fraction of the returned documents that are deemed correct. In our context, the recall corresponds to completeness and the precision to the soundness.

2.3 Source Descriptors

A data source is modeled by a *source descriptor* of the form $\langle \varphi, v, c, s, f, r \rangle$, where

- φ is a view definition;
- v is a view extension;
- $c \in [0, 1]$ is a lower bound for the completeness;
- $s \in [0, 1]$ is a lower bound for the soundness;

3. CONSISTENCY OF A SOURCE COLLECTION

Consider a source S characterized by source descriptor $\langle \varphi, v, c, s \rangle$, as defined in Section 2.

[1] These "suspect" values are available from [4].

By giving this source descriptor, a data provider sets an implicit constraint over the possible global databases D:

$$D \text{ is such that } c_D(S) \geq c \text{ and } s_D(S) \geq s.$$

In order to give meaning to answers assembled from multiple sources, we first need to characterize the set of database instances that are consistent with the completeness and soundness measures claimed by a given collection of sources.

Consider a source collection $\mathcal{S} = S_1, \ldots, S_n$, where $S_i = \langle \varphi_i, v_i, c_i, s_i \rangle$, for $i \in [1, n]$.

The source collection $\mathcal{S}$ defines a set of possible databases, denoted $poss(\mathcal{S})$, as follows:

$$poss(\mathcal{S}) = \{D \text{ over } sch(\mathcal{S}) \ : \ c_D(v_i) \geq c_i \text{ and } s_D(v_i) \geq s_i, \text{ for all } S_i \in \mathcal{S}\}$$

where $sch(\mathcal{S})$ is the schema of $\mathcal{S}$, *i.e.* the set of all global relation names occurring in the view definitions.

In the remainder of this section we consider the following problem: given a source collection $\mathcal{S}$, determine whether $poss(\mathcal{S})$ is non-empty, in other words whether $\mathcal{S}$ is *consistent*. More precisely, this problem can be stated as:

CONSISTENCY
INSTANCE: A source collection $\mathcal{S} = \{S_1, \ldots, S_n\}$, where $S_i = \langle \varphi_i, v_i, c_i, s_i \rangle$, for $i \in [1, n]$.
QUESTION: Is there a global database D that satisfies the following conditions:

$$\frac{|\varphi_i(D) \cap v_i|}{|\varphi_i(D)|} \geq c_i \text{ and } \frac{|\varphi_i(D) \cap v_i|}{|v_i|} \geq s_i$$

for all $i \in [1, n]$?

We begin with a preliminary result that limits the search space for solutions to databases whose total size is bounded by a constant.

LEMMA 3.1. *Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a source collection, where $S_i = \langle \varphi_i, v_i, c_i, s_i \rangle$, for $i \in \{1, \ldots, n\}$. Then, $poss(\mathcal{S}) \neq \emptyset$ if and only if there exists a global database $D \in poss(\mathcal{S})$ over $sch(\mathcal{S})$ such that*

$$|D| \leq max_{i=\overline{1,n}} |body(\varphi_i)| \cdot (\Sigma_{i=1}^{n} |v_i|)$$

Proof: We only need to prove the "only if" direction. Take an arbitrary global database $G \in poss(\mathcal{S})$. For each $i \in \{1, \ldots, n\}$ construct $G_i \subseteq G$ as follows: for each fact $u \in \varphi_i(G) \cap v_i$, choose a valuation θ_u such that $head(\varphi_i)\theta = u$ and all the facts in $body(\varphi_i)$ are in G (there exists at least one such valuation, according to the definition of applying a view to a database); then, let

$$G_i = \{t : t \text{ in } body(\varphi_i)\theta_u, u \in \varphi_i(G) \cap v_i\}$$

Finally take

$$D = \bigcup_{i=1}^{n} G_i$$

We have:

$$|D| \leq \Sigma_{i=1}^{n}|G_i| = \Sigma_{i=1}^{n}|body(\varphi_i)||v_i| \leq$$

$$\leq max_{i=\overline{1,n}}|body(\varphi_i)| \cdot (\Sigma_{i=1}^{n}|v_i|)$$

We now need to prove that D is in $poss(\mathcal{S})$. For every i, we have:

$$\varphi_i(D) \cap v_i \supseteq \varphi_i(G_i) \cap v_i = \varphi_i(G) \cap v_i$$

Because $D \subseteq G$, we also have that:

$$\varphi_i(D) \cap v_i \subseteq \varphi_i(G) \cap v_i$$

and therefore

$$\varphi_i(D) \cap v_i = \varphi_i(G) \cap v_i$$

This enables us to infer that:

$$s_D(v_i) = \frac{|\varphi_i(D) \cap v_i|}{|v_i|} = \frac{|\varphi_i(G) \cap v_i|}{|v_i|} = s_G(v_i) \geq s_i$$

and

$$c_D(v_i) = \frac{|\varphi_i(D) \cap v_i|}{|\varphi_i(D)|} = \frac{|\varphi_i(G) \cap v_i|}{|\varphi_i(D)|} \geq = \frac{|\varphi_i(G) \cap v_i|}{|\varphi_i(G)|} =$$

$$= c_G(v_i) \geq c_i$$

This proves that $D \in poss(\mathcal{S})$ and concludes the lemma's proof.

We are now ready to prove the following theorem:

THEOREM 3.2. **CONSISTENCY** *is NP-complete (in the size of the view extensions).*

Proof:

i) To prove that the problem is in NP let $m = max_{i=\overline{1,n}}|body(\varphi_i)|$, $k = max\{arity(R) : R \in sch(\mathcal{S})\}$ and $p = \Sigma_{i=1}^{n}|v_i|$. The previous lemma limits the search space for a possible database to global databases with at most mp atoms involving at most mpk constants. We can fix a set $\mathbf{dom}_0$ of mpk constants ahead of time (including all the constants in view extensions). It is easy to see that if there exists a possible database in the search space mentioned earlier, there exists an equivalent possible database (modulo a bijection on $\mathbf{dom}$) that has constants only from $\mathbf{dom}_0$. Therefore, we can pick in nondeterministic polynomial time a database D over $sch(\mathcal{S})$ with constants in $\mathbf{dom}_0$ and then check in polynomial time whether $D \in poss(\mathcal{S})$ (by computing, for each i, $\varphi_i(D)$ and checking the requirements on relative completeness $c_D(v_i)$ and soundness $s_D(v_i)$). This proves that the source collection consistency problem is in **NP**.

ii) To prove NP-completeness, we construct a reduction from a special case HS* of the HITTING SET (HS) problem [**?**], which we later prove to be NP-complete as well. We state both problems here:

HITTING SET (HS)
INSTANCE: Collection $\mathcal{C} = \{A_1, A_2, \ldots, A_n\}$ of subsets of a finite set S and positive integer $K \leq |S|$.
QUESTION: Is there a subset $A \subseteq S$ such that $|A| \leq K$ and A contains at least one element from each subset in $\mathcal{C}$?

HITTING SET* (HS*)
INSTANCE: Collection $\mathcal{C} = \{A_1, A_2, \ldots, A_n\}$ of subsets of a finite set S such that A_n is a singleton, positive integer $K \leq |S|$.
QUESTION: Is there a subset $A \subseteq S$ such that $|A| \leq K$ and A contains at least one element from each subset in $\mathcal{C}$?

We transform an instance of HS* to an instance of CONSISTENCY as follows:

Let R be a fixed global relation name of arity 1. For every $i \in [1, n]$, build a source $S_i = \langle \varphi_i, v_i, c_i, s_i \rangle$ where:

- $\varphi_i : V_i(x) \leftarrow R(x)$;
- $v_i = \{V_i(a) : a \in A_i\}$;
- $c_i = 1/K$;
- $s_i = 1/|A_i|$.

We claim that a solution D of CONSISTENCY can be easily transformed into a solution A of HS* with the following mapping: $A = \{a \in S : R(a) \in D\}$.

To verify that, we need to show that $A \cap A_i \neq \emptyset$ and $|A| \leq K$.

Since D is a solution to CONSISTENCY, we know that:

$$\frac{|\varphi_i(D) \cap v_i|}{|v_i|} \geq s_i$$

But since $\varphi_i(D) = \{V_i(a) : a \in A\}$ and $v_i = \{V_i(a) : a \in A_i\}$, we can rewrite the above inequality to:

$$\frac{|A \cap A_i|}{|A_i|} \geq s_i = \frac{1}{|A_i|}, \text{ for all } i \in [1, n]$$

Hence $|A \cap A_i| \geq 1$ for all $i \in [1, n]$.

Also, we know that:

$$\frac{|\varphi_n(D) \cap v_n|}{|\varphi_n(D)|)} \geq c_n$$

which can be similarly rewritten to:

$$\frac{|A \cap A_n|}{|A|} \geq c_n = \frac{1}{K},$$

But since A_n is a singleton and $|A \cap A_n| \geq 1$ we infer that $|A \cap A_n| = 1$, therefore the above inequality becomes $|A| \leq K$, which qualifies A as a solution to HS*.

Conversely, we need to show that if HS* has a solution, than so does CONSISTENCY. Consider an arbitrary solution A'

to HS*. We claim that $D = R(A') = \{R(a) : a \in A'\}$ is a solution to CONSISTENCY. To verify that, we need to show that for all $i \in [1, n]$,

$$\frac{|A' \cap A_i|}{|A'|} \geq c_i \text{ and } \frac{|A' \cap A_i|}{|A_i|} \geq s_i$$

Since $A' \cap A_i \neq \emptyset$, we get

$$\frac{|A' \cap A_i|}{|A'|} \geq \frac{1}{|A'|} \geq \frac{1}{K} = c_i$$

and

$$\frac{|A' \cap A_i|}{|A_i|} \geq \frac{1}{|A_i|} = s_i$$

Done.

To conclude the proof, we need to prove that HS* is NP-complete. Since HS* is a special case of HS, it is in NP. We prove its completeness by showing that HS reduces to it.

LEMMA 3.3. *HS reduces to HS**

Proof

Consider an instance I of HS. We construct an instance I^* for HS* by taking $S^* = S \cup \{a\}$ where $a \notin S$ is a new element, $\mathcal{C}^* = \{A_1, A_2, \ldots, A_n, A_{n+1}\}$, where $A_{n+1} = \{a\}$, and $K^* = K + 1$.

We need to show that I has a solution if and only if I^* has a solution.

Take an arbitrary solution to I^*, A^*. As a solution, A^* is guaranteed to contain at least one element from each subset in $\mathcal{C}^*$, in particular $A_{n+1} = \{a\}$, so A^* contains a. We consider $A = S \setminus \{a\}$. For all $i \in \overline{1, n}$ we have $A \cap A_i = A^* \cap A_i \neq \emptyset$. Also, $|A| = |A^*| - 1 < (K + 1) - 1 = K$.

Conversely, if A is an arbitrary solution to I, we construct $A^* = S \cup \{a\}$. For all $i \in \overline{1, n}$ we have $A^* \cap A_i = A \cap A_i \neq \emptyset$. Also, $A^* \cap A_{n+1} \neq \emptyset$ by construction. Finally, $|A^*| = |A| + 1 \leq K + 1$. QED.

By examining the proof of Theorem 3.2, we notice that the NP-hardness was shown by reducing an NP-complete problem to a special case of the consistency problem, thus enabling us to formulate the following Corollary.

COROLLARY 3.4. *The* **CONSISTENCY** *problem remains NP-complete even if all the view definitions are identities over the same global relation.*

4. POSSIBLE DATABASES

In the previous section, we considered the problem of determining whether the set of possible databases $poss(\mathcal{S})$ generated by a source collection $\mathcal{S}$ is empty or not. Once we have determined that a given source collection is consistent, the next natural step would be to characterize the set of possible databases. In this section, we provide a representation of the set of possible databases in terms of tableaux [2, 6].

We start by introducing some necessary auxiliary concepts.

A *tableau* over a global schema $\mathbf{R}$ is a finite set of atoms over the relation names in $\mathbf{R}$.

A *constraint* over $\mathbf{R}$ is a pair (U, Θ) where U is a tableau over $\mathbf{R}$ and Θ is a set of substitutions of the form $\{x_1/e_1, \ldots, x_p/e_p\}$ where all the x_i-s appear in U, and the e_i-s are either constants or variables.

A *valuation* is a partial mapping from $\mathbf{var} \cup \mathbf{dom}$ to $\mathbf{dom}$ that is the identity on $\mathbf{dom}$. A valuation σ is said to be *compatible* with a substitution $\theta = \{x_1/e_1, \ldots, x_p/e_p\}$ if $\sigma(x_i) = \sigma(e_i)$, for all $i \in [1, p]$.

A constraint (U, Θ) is said to be *satisfied* by a database instance D if every time the tableau U can be embedded in D via a valuation σ, there is a substitution θ in Θ that is compatible with σ.

A *database template* $\mathcal{T}$ over $\mathbf{R}$ is a tuple $\langle T_1, \ldots, T_m, C\rangle$ where each T_i is a tableau over $\mathbf{R}$ and C is a finite set of constraints over $\mathbf{R}$.

EXAMPLE 4.1. *Let* $\mathcal{T} = \langle T_1, T_2, C\rangle$, *where*
$T_1 = \{R(a, x), S(b, c), S(b, c')\}$,
$T_2 = \{R(a', b'), S(b, c)\}$, *and*
$C = \{(\{R(a, x)\}, \{\{x/b\}, \{x/b'\}\})\}$. *This database template contains two tableaux, and one constraint with two substitutions.*

A database template is a compact representation for the *set* of all database instances that can be obtained by replacing the variables in tableaux with constants in such a way that all the constraints are satisfied. The following definition formalizes this.

DEFINITION 4.1. *A database template* $\mathcal{T}$ *on schema* $\mathbf{R}$ represents *the following set of global databases:*

$rep(\mathcal{T}) = \{D$: *there is a valuation* ϑ *and a tableau* T_i *in* $\mathcal{T}$ *such that* $\vartheta(T_i) \subseteq D$, *and for all* $(U, \Theta) \in C$ *in* $\mathcal{T}$ *and valuations* σ *such that* $\sigma(U) \subseteq D$ *there is a* $\theta \in \Theta$ *such that* σ *and* θ *are compatible*$\}$

EXAMPLE 4.2. *Consider the database template* $\mathcal{T} = \langle T_1, T_2, C\rangle$ *from Example 4.1. This template represents the following three global databases* $\{R(a, b), S(b, c), S(b, c')\}$, $\{R(a, b'), S(b, c), S(b, c')\}$, $\{R(a', b'), S(b, c)\}$ *and any of their supersets satisfying the constraint that whenever a occurs on the first position of an R atom, then the second component has to be b or b'. For instance,* $\{R(a, b), R(a, b'), S(b, c), S(b, c')\}$ *is a database in* $rep(\mathcal{T})$, *while* $\{R(a, c), R(a, b'), S(b, c), S(b, c')\}$ *is not in* $rep(\mathcal{T})$ *(because the atom* $R(a, c)$ *violates the constraint).*

Going back to our source collection problem, consider a source collection $\mathcal{S} = \{S_1, \ldots, S_n\}$, where $S_i = \langle \varphi_i, v_i, c_i, s_i\rangle$,

for $i \in \{1, \ldots, n\}$. We would like to express the set of possible databases $poss(\mathcal{S})$ as a set of databases represented by some template $\mathcal{T}$.

Denote by $k_i = |v_i|$, $w_i = |\varphi_i(D)|$, and $t_i = |\varphi_i(D) \cap v_i|$ (the w_i-s and t_i-s are unknowns). From the definition of $poss(\mathcal{S})$ we infer the following inequations:

$$c_D(v_i) = \frac{t_i}{w_i} \geq c_i \tag{1}$$

$$s_D(v_i) = \frac{t_i}{k_i} \geq s_i \tag{2}$$

From (2) we get:

$$t_i \geq s_i k_i \tag{3}$$

This means that in order to determine all the possible databases we can consider all combinations of subsets $u_i \subseteq v_i$ such that $|u_i| \geq s_i k_i$ in turn, and take the union of all the solutions. For each i, the selected subset u_i is seen as $\varphi_i(D) \cap v_i$, that is, the set of sound atoms in the view extension v_i.

To simplify the notation, let $U = (u_1, \ldots, u_n)$ be a fixed combination of subsets. For this combination U of subsets, (1) gives an upper bound for the size of the set $\varphi_i(D)$:

$$|\varphi_i(D)| = w_i \leq \frac{t_i}{c_i} \tag{4}$$

From the above considerations we infer that any global database D for which the result of applying the view definition φ_i to D is a superset of u_i of size not greater than t_i/c_i is a possible database.

In order to find all the solutions D, we shall first construct a database template over the global schema, $\mathbf{R}$. We define a function (T^U, C^U) from source descriptions to database templates over $\mathbf{R}$, where U is a given combination of subsets. Given a source description $S_i = \langle \varphi_i, v_i, c_i, s_i \rangle$, we set

$$T^U(S_i) = \{t : t \text{ in } body(\varphi_i)\theta \text{ and } head(\varphi_i)\theta = u, \text{ for some } u \in u_i \text{ and assignment } \theta\}.$$

Denote by $m_i = \lfloor t_i/c_i \rfloor$. We can express the cardinality constraint 4 by requiring that in any enumeration of $m_i + 1$ atoms of $\varphi_i(D)$ of the form:

$$\begin{array}{l} V_i(x^i_{1,1}, \ldots, x^i_{1,l_i}) \\ V_i(x^i_{2,1}, \ldots, x^i_{2,l_i}) \\ \ldots \\ V_i(x^i_{m_i,1}, \ldots, x^i_{m_i,l_i}) \\ V_i(x^i_{m_i+1,1}, \ldots, x^i_{m_i+1,l_i}) \end{array}$$

there be at least two identical atoms. This in turn can be expressed by requiring that any valuation that embeds the above atoms in $\varphi_i(D)$ must be compatible with one substitution of the form

$$\theta_{p,r} = \{x^i_{p,1}/x^i_{r,1}, \ldots, x^i_{p,l_i}/x^i_{r,l_i}\}, \text{ where } p, r \in [1, m_i{+}1], p \neq r$$

Therefore, we can capture the cardinality constraint 4 by setting $C^U(S_i) = (V^U(S_i), \Theta^U(S_i))$ where

$$V^U(S_i) = \{t : t \text{ in } body(\varphi_i)\theta \text{ and } head(\varphi_i)\theta = V_i(x^i_{s,1}, \ldots, x^i_{s,l_i}), \text{ for some } s \in [1, m_i + 1] \text{ and assignment } \theta\}$$

and

$$\Theta^U(S_i) = \{\theta_{p,r} : p, r \in [1, m_i + 1], p \neq r\}.$$

Finally, we set

$$T^U(\mathcal{S}) = \bigcup_{S_i \in \mathcal{S}} T^U(S_i)$$

$$C^U(\mathcal{S}) = \{C^U(S_i) : S_i \in \mathcal{S}\}$$

and

$$\mathcal{T}^U(\mathcal{S}) = \left\langle T^U(\mathcal{S}), C^U(\mathcal{S}) \right\rangle$$

This collection of database templates has the following desirable property:

THEOREM 4.1.

$$poss(\mathcal{S}) = \bigcup_{U \in \mathcal{U}} rep(\mathcal{T}^U(\mathcal{S}))$$

where $\mathcal{U} = \{U = (u_1, \ldots, u_n) : u_i \subseteq v_i \text{ s.t. } |u_i| \geq s_i|v_i|, i \in [1, n]\}$ *is the set of all allowable combinations of subsets of the view extensions.*

Proof.

Take D in $\bigcup_{U \in \mathcal{U}} rep(\mathcal{T}^U(\mathcal{S}))$. This means that there is a $U \in \mathcal{U}$ such that $D \in rep(\mathcal{T}^U(\mathcal{S}))$. Then there is a valuation v such that $v(T^U(\mathcal{S})) \subseteq D$ where $T = \bigcup_{S_i \in \mathcal{S}} T^U(S_i)$ and D satisfies the constraints $\bigcup_{S_i \in \mathcal{S}} C^U(S_i)$. Now consider an arbitrary source S_i in $\mathcal{S}$, and let u be a fact in u_i. Then there is an assignment θ such that $head(\varphi_i)\theta = u$ and all atoms in $body(\varphi_i)\theta$ are in T. By applying v it follows that all facts in $v(body(\varphi_i)\theta)$ are in D. This means that $v(head(\varphi_i)\theta)) = head(\varphi_i)\theta = u$ is in $\varphi_i(D)$, and therefore $u_i \subseteq \varphi_i(D)$. Since u_i is a subset of v_i and has at least $s_i|v_i|$ elements, it follows that the soundness of v_i w.r.t. D is at least s_i. To prove that the completeness of v_i w.r.t. D is at least c_i, suppose the opposite is true, that is:

$$\frac{|v_i \cap \varphi_i(D)|}{|\varphi_i(D)|} < c_i$$

This would imply that

$$|\varphi_i(D)| > \frac{|v_i \cap \varphi_i(D)|}{c_i} \geq \frac{|u_i|}{c_i} = m_i$$

But this would mean that D violates the constraint $C^U(S_i)$. Therefore, D makes each source S_i at least c_i complete and at least s_i sound, which qualifies it as an element of $poss(\mathcal{S})$.

Conversely, take D in $poss(\mathcal{S})$. For each $i \in [1,n]$, let $u_i = \varphi_i(D) \cap v_i$. Since each source S_i is at least s_i sound w.r.t. D it follows that each u_i has at least $s_i k_i$ elements. Therefore $U = (u_1, \ldots, u_n)$ is in $\mathcal{U}$, so, if we prove that $D \in rep(\mathcal{T}^U(\mathcal{S}))$, we are done. From the construction of T^U it follows that there is a valuation v such that $v(T^U(\mathcal{S})) \subseteq D$. We only need to show that D doesn't violate any of the constraints in $C^U(\mathcal{S})$. Suppose, therefore that D violates one of these constraints, say $C^U(S_i) = (V^U(S_i), \Theta^U(S_i))$. This means that there exists a valuation v' such that $v'(V^U(S_i)) \subseteq D$, and v' is not compatible with any of the $\theta \in \Theta^U(S_i)$. This means that the set $W_i = \{t : t = head(\varphi_i)\theta,$ s.t. $body(\varphi_i)\theta \in w(V^U(S_i))\}$ has at least $m_i + 1$ elements and since $\varphi(D) \supseteq W_i$, this means that

$$|\varphi(D)| \geq m_i + 1 = \left\lfloor \frac{|u_i|}{c_i} \right\rfloor + 1 > \frac{|u_i|}{c_i} = \frac{|\varphi_i(D) \cap v_i|}{c_i}$$

and therefore

$$\frac{|\varphi_i(D) \cap v_i|}{|\varphi(D)|} < c_i$$

which is in contradiction with the hypothesis that D is in $poss(\mathcal{S})$. *QED*.

Theorem 4.1 gives a finite representation of the set of possible global databases in terms of the set of databases represented by a collection of database templates.

5. ANSWERING QUERIES

Consider the same framework as before: a source collection $\mathcal{S} = \{S_1, \ldots, S_n\}$, where $S_i = \langle \varphi_i, v_i, c_i, s_i \rangle$, for $i \in \{1, \ldots, n\}$. In the previous section we studied the problem of computing the set $poss(\mathcal{S})$ of possible databases defined by the given source collection. The next step is to study the semantics of query answering over a source collection.

Consider a conjunctive query over the relation names in $\mathbf{R}$:

$$Q : head(Q) \leftarrow body(Q)$$

where, by convention, $head(Q)$ is an atom over a fixed atom name ans, and $body(Q)$ is a sequence $b_1, b_2, \ldots, b_m$ of atoms over global relation names. As usual, we assume that all queries are *safe* (all variables in the head also occur in the body). For a fixed global database D, the result of applying Q to D, denoted $Q(D)$ is a set of facts over ans.

We have seen that, in general, a consistent source collection doesn't uniquely define a global database, but rather a set of possible databases, which we denoted $poss(\mathcal{S})$. It is therefore natural to define the result of applying a query Q to a given source collection $\mathcal{S}$ to be the set of the results obtained by applying Q to each of the possible databases:

$$Q(\mathcal{S}) = \{Q(D) : D \in poss(\mathcal{S})\}$$

From a practical point of view, presenting the result as a collection of possible results is not very useful, and in most cases not feasible (as $poss(\mathcal{S})$ is generally large). In order to avoid this, two approximations have been proposed in the literature:

$$Q_*(\mathcal{S}) = \bigcap_{D \in poss(\mathcal{S})} Q(D)$$

and

$$Q^*(\mathcal{S}) = \bigcup_{D \in poss(\mathcal{S})} Q(D)$$

The lower approximation $Q_*(\mathcal{S})$ is also known as the *certain* answer, because it contains exactly those facts that are common to all the answers for any possible database (*i.e.* the *certain* facts). The upper approximation $Q^*(\mathcal{S})$ is called the *possible* answer, because it contains the collection of all the facts that appear in the answers for all the possible databases.

One natural question arises: could we say more about the individual atoms in the possible answer? We know that the facts in the certain answer (if any) are guaranteed to belong to the result of applying the query to any of the possible databases. How about the other atoms in the possible answer, are they "equally possible"? Intuitively, because of the different soundness and completeness bounds on each of the sources, some atoms in the possible answer would have been obtained in more of the possible worlds than others. To capture this intuition, we define the *confidence* of an fact t with respect to a query Q as the probability that t is in the result of applying Q to a database instance D chosen at random from the collection $poss(\mathcal{S})$:

$$confidence_Q(t) = Pr(t \in Q(D) | D \in poss(\mathcal{S}))$$

If the domain **dom** is finite, the above conditional probability can be computed (at least in principle) by generating all the possible global databases (in exponential time). In general, computing this confidence value is NP-hard (because it includes the consistency problem as a sub-problem).

5.1 Computing the Confidence of Base Facts

In this section we consider the special case when all the view definitions are identities over the same global relation name R and the domain **dom** is finite. We describe an algorithmic method for computing the confidence value for any fact in $poss(\mathcal{S})$ (with respect to the identity query).

Let $\mathcal{S} = \{S_1, \ldots, S_n\}$, where $S_i = \langle \varphi_i, v_i, c_i, s_i \rangle$,
$\varphi_i : V_i(x_1, \ldots, x_k) = R(x_1, \ldots, x_k)$ for $i \in \{1, \ldots, n\}$. We want to determine all the global databases D in $poss(\mathcal{S})$. In this case a global database is a set of facts over R.

We can construct an enumeration of all the facts over R with constants in **dom**:

$$\begin{aligned} t_1 &= R(c_1, \ldots, c_1, c_1) \\ t_2 &= R(c_1, \ldots, c_1, c_2) \\ t_3 &= R(c_1, \ldots, c_1, c_3) \\ &\ldots \\ t_N &= R(c_r, \ldots, c_r, c_r) \end{aligned}$$

where $\{c_1, \ldots, c_r\}$ is an enumeration of **dom** and $N = r^{arity(R)}$.

To each such fact t_i, we associate a variable x_i taking values in $\{0, 1\}$ with the following interpretation: $x_i = 1$ if and

only if $t_i \in D$. A set D of facts over R is in $poss(\mathcal{S})$ if and only if, for every $i \in [1,n]$:

$$\begin{cases} \frac{|\varphi_i(D) \cap v_i|}{|\varphi_i(D)|} \geq c_i \\ \frac{|\varphi_i(D) \cap v_i|}{|v_i|} \geq s_i \end{cases}$$

that is:

$$\begin{cases} |\varphi_i(D) \cap v_i| \geq c_i \cdot |\varphi_i(D)| \\ |\varphi_i(D) \cap v_i| \geq s_i \cdot |v_i| \end{cases}$$

which can be written as:

$$\begin{cases} \sum_{\varphi_i(t_i) \in v_i} x_i \geq c_i \sum_{i=1}^{N} x_i \\ \sum_{\varphi_i(t_i) \in v_i} x_i \geq s_i |v_i| \end{cases}$$

and finally:

$$\begin{cases} \sum_{\varphi_i(t_i) \in v_i} x_i(1-c_i) - \sum_{\varphi_i(t_i) \notin v_i} c_i x_i \geq 0 \\ \sum_{\varphi_i(t_i) \in v_i} x_i \geq s_i |v_i| \end{cases}$$

We also impose the following constraints on each variable x_i: $0 \leq x_i \leq 1$, x_i integer.

By collecting the above inequalities and variable constraints for every $i \in [1,n]$, we obtain a linear system Γ with $4N$ inequalities and N variables. To compute the confidence of a particular fact t_p, it is enough to determine $N_{sol}(\Gamma)$, the number of integer solutions of S, and $N_{sol}(\Gamma[x_p/1])$ the number of solutions for the system obtained by replacing the variable x_p with the constant 1. Then, the confidence of t_p is:

$$confidence(t_p) = \frac{N_{sol}(\Gamma[x_p/1])}{N_{sol}(\Gamma)}$$

Please note the above fraction is defined for any *consistent* source collection ($N_{sol}(\Gamma) = |poss(\mathcal{S})|$ is non-zero if and only if $poss(\mathcal{S}) \neq \emptyset$).

EXAMPLE 5.1. *Consider a collection with two sources* $\mathcal{S} = \{S_1, S_2\}$, *with* $S_1 = \langle Id_R, \{R(a), R(b)\}, 0.5, 0.5\rangle$ *and* $S_2 = \langle Id_R, \{R(b), R(c)\}, 0.5, 0.5\rangle$ *(where* Id_R *is the identity on* R*). Assume a finite domain* $\mathbf{dom} = \{a, b, c, d_1, \ldots, d_m\}$. *Then, to compute the tuple confidence for a tuple* $R(\alpha)$, *one needs to compute the number of solutions for the following system* Γ *of inequalities:*

$$\begin{cases} x_a + x_b - x_c - x_{d_1} - \cdots - x_{d_m} \geq 0 \\ x_a + x_b \geq 1 \\ x_b + x_c - x_a - x_{d_1} - \cdots - x_{d_m} \geq 0 \\ x_b + x_c \geq 1 \end{cases}$$

and of the systems $\Gamma[x_\alpha/1]$.

After solving the above system, we derive the following values for the confidences:

$$\text{confidence}(R(a)) = \text{confidence}(R(c)) = \frac{m+2}{2m+3}$$

$$\text{confidence}(R(b)) = \frac{2m+2}{2m+3}$$

$$\text{confidence}(R(d_i)) = \frac{2}{2m+3}, 1 \leq i \leq m$$

By examining the behavior for large values of m ($m \to \infty$), *we observe that* $R(b)$ *has confidence almost 1,* $R(a)$ *and* $R(b)$ *have confidence about* $1/2$ *and all the other tuples* $R(d_i)$ *have confidence close to 0. This corresponds to our intuition:* $R(b)$ *has greater confidence since it is present in both sources,* $R(a)$ *and* $R(b)$ *have a smaller confidence because each of them appears in only one source, and the other possible tuples have low confidence because they are not backed up by any source.*

5.2 Computing the Confidence of Answer Tuples

Consider a query Q over the global relations in $sch(\mathcal{S})$. For every tuple in the possible answer $Q^*(\mathcal{S})$, its confidence is given by:

$$confidence_Q(t) = Pr(t \in Q(D) | D \in poss(\mathcal{S}))$$

In this section we introduce a method for deriving the confidence of any tuple in the possible answer from the confidence of the base facts in $poss(\mathcal{S})$.

Notation: If $\{p_i\}_{i\in[1,N]}$ are the probabilities of N mutually independent events $\{E_i\}_{i\in[1,N]}$, we denote by $\bigoplus_{i\in[1,N]} p_i$ the probability of the union $\bigcup_{i\in[1,N]} E_i$, that is:

$$\bigoplus_{i\in[1,N]} p_i = 1 - \Pi_{i=1}^{N}(1-p_i)$$

DEFINITION 5.1. *For every relational query* Q *and every tuple* t *in* $Q^*(\mathcal{S})$, *we define a number* $\text{conf}_Q(t) \in [0,1]$ *as follows:*

- *if* $Q = R$, *where* R *is a relation name, we let*
$$\text{conf}_Q(t) = \text{confidence}_R(t)$$
- *if* $Q = \pi_{Att} Q'$, *we let*
$$\text{conf}_Q(t) = \bigoplus_{t' \in Q'(\mathcal{S}) \text{ s.t } \pi_{Att} t' = t} \text{conf}_{Q'}(t')$$
- *if* $Q = \sigma_\phi Q'$, *we let*
$$\text{conf}_Q(t) = \text{conf}_{Q'}(t)$$
- *if* $Q = Q' \times Q''$, *we let*
$$\text{conf}_Q(t) = \text{conf}_{Q'}(t') \cdot \text{conf}_{Q''}(t'')$$
where t' *and* t'' *are such as* $t = t' \times t''$;

THEOREM 5.1. *Let* Q *be a relational query over* $sch(\mathcal{S})$. *Then, for every tuple* t *in* $Q^*(\mathcal{S})$

$$\text{confidence}_Q(t) = \text{conf}_Q(t)$$

Proof By structural induction on Q, using standard probability laws.

6. DISCUSSION

We examined some computational issues arising when processing queries over source collections with incomplete and partially sound data sources.

We first considered the *source collection consistency problem*: given a source collection, determine whether there exists a possible global database which is consistent with all the claims of soundness and completeness of individual sources. We showed this problem to be NP-complete in the size of the data in the sources (the view extensions). In our analysis, we do not consider sources that report wrong estimates of soundness and completeness (either on purpose or because of lack of information). One interesting future direction would be to explore how a notion of *consensus* can be defined and used to detect the most trustworthy sources.

Then, we gave a finite representation of the set of possible databases in terms of tableaux and constraints. A future direction would be to use this representation to compute a finite representation of the answer to any query, along the lines of [6].

Finally, we examined the semantics of query answering over source collections with completeness and soundness metadata. We adapted the well-known notions of *certain answer* and *possible answer* to our framework. In addition to these, we introduced a notion of tuple *confidence*, and we described a method to compute it in the special case when all the views are identities and the domain is finite.

As a final remark, we note that the results in the special case when the view definitions are all identities over the same relation name are not dependent on the data model; all the results can be expressed in terms of sets and can therefore be applied in other domains, for any situation dealing with multiple, incomplete and partially incorrect (obsolete), copies of a set of objects. Examples of such situations include: multiple caches of a set of objects (*e.g.* Web pages, memory locations), multiple mirror-sites of a given site, *etc.*

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council and by Communications and Information Technology Ontario.

7. REFERENCES

[1] S. Abiteboul and O. M. Duschka. Complexity of answering queries using materialized views. In *Proceedings of the 17th Symposium on Principles of Database Systems (PODS)*, pages 254–263, Seattle, Washington, June 1998.

[2] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[3] D. Florescu, D. Koller, and A. Y. Levy. Using probabilistic information in data integration. In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB)*, pages 216–225, Athens, Greece, Aug. 1997.

[4] The Global Historical Climatology Network (home page). www.ncdc.noaa.gov/ol/climate/research/ghcn/ghcn.html.

[5] E. Grädel, Y. Gurevich, and C. Hirch. The complexity of query reliability. In *Proceedings of the 17th Symposium on Principles of Database Systems*, pages 227–234. ACM Press, 1998.

[6] G. Grahne and A. O. Mendelzon. Tableau techniques for querying information sources through global schemas. *Lecture Notes in Computer Science*, 1540:332–347, 1999.

[7] D. Kaplan and R. Krishnan. Assessing data quality in accounting information systems. *Communications of the ACM*, 41(2):72–78, Feb. 1998.

[8] M. Kifer and A. Li. On the semantics of rule-based expert systems with uncertainty. In M. Gyssens, J. Paredaens, and D. V. Gucht, editors, *Proceedings of 2nd International Conference on Database Theory (ICDT'88)*, volume 326 of *Lecture Notes in Computer Science*, pages 102–117. Springer, 1988.

[9] T. Kirk, A. Y. Levy, Y. Sagiv, and D. Srivastava. The Information Manifold. In *Proc. of the AAAI Spring Symposium on Information Gathering in Distributed Heterogeneous Environments*, Stanford, CA, Mar. 1995.

[10] A. Y. Levy. Obtaining complete answers from incomplete databases. In *Proceedings of the 22nd International Conference on Very Large Data Bases (VLDB)*, pages 402–412, Bombay, India, Sept. 1996.

[11] A. Y. Levy, A. O. Mendelzon, Y. Sagiv, and D. Srivastava. Answering queries using views. In *Proceedings of the 14th Symposium on Principles of Database Systems (PODS)*, pages 95–104, San Jose, California, May 1995.

[12] A. Motro. Multiplex: A formal model for multidatabases and its implementation. Technical Report ISSE-TR-95-103, George Mason University, Fairfax, VA, Mar. 1997.

[13] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.