

Answering queries about Validity and Completeness of data: from Modal Logic to Relational Algebra

Robert Demolombe
CERT/ONERA
France
Robert.Demolombe@cert.fr

We are never guaranteed that information stored in data or knowledge bases is a correct representation of the world. Nevertheless, there are many situations where we have strong supports for the validity and/or completeness of parts of the information.

For instance, we may know that salaries of people in a company are valid if they are inserted by people from accounting department, and that data about absenteeism are complete if they are inserted by people from the department of human resources. In that example validity and completeness depend on reliability of information sources, in other examples they may only depend on the type of data. For instance, we may know that data about health of people are valid.

In these contexts, if one asks a complex query it is not easy to infer from the meta information about validity and completeness which parts of the answer to the query are valid, and which parts are complete. This paper presents some results of an ongoing work ¹ which is intended to formalise this kind of meta information and to formalise associated queries about validity and completeness of data. We start with a formalisation in modal logic of situations where a distinction is made between, on one hand, data that have been explicitly inserted by information sources who are supposed to be, or not to be, reliable and, on the other hand, data inferred from the previous ones. Then, formalisation is adapted to situations where reliability only depends on some types of information. Finally,

¹This work was initiated in collaboration with Andrew J. Jones (see [4]). Most of the ideas presented in section 1 are the result of this work, even if they are presented in a slightly different form.

we show how, in the context of relational databases, the logical formalism can be adapted to relational algebra, provided the closed world assumption (CWA) is accepted.

1 Reliable information sources

In this section we consider situations where some information sources are supposed to be reliable in regard to validity or to completeness of some data.

A distinction is made between data base content, which is viewed as a set of beliefs and a correct description of the world. Sentences of the kind Bp can be read “the data base believes p ”, and sentence p can be interpreted as “ p is true of the world”, where B is a normal modal operator that obeys axiom schemas (K) and (D), and inference rule (Nec) [2]:

$$(K) \quad B(p \rightarrow q) \rightarrow (Bp \rightarrow Bq)$$

$$(D) \quad Bp \rightarrow \neg B\neg p$$

$$(Nec) \quad \frac{\vdash p}{\vdash Bp}$$

An information source “ a ” is reliable in regard to validity of sentence p iff the fact that a has inserted p in the data base implies that p is true of the world. The fact that a has inserted p is expressed in the form: “agent a has brought about that the data base believes p ”, which is formalised, using the action operator E_a , by the sentence: $E_a(Bp)$. The operator E_a is a classical (not normal) modal operator that obeys axiom schemas ($\neg N$) and inference rule (RE):

$$(\neg N) \quad \neg E_a(\text{true})$$

$$(RE) \quad \frac{\vdash p \leftrightarrow q}{\vdash E_a p \leftrightarrow E_a q}$$

Since no information about the world can be guaranteed to be true, in an absolute sense, the fact that an information source is reliable, itself, cannot be guaranteed to be true. However, it is assumed that this fact of being a reliable information source has a different status than other database beliefs, in the sense that the data base believes that it is a true belief. For this reason another normal modal operator K is introduced, and sentences of the kind Kp can be read “the data base “knows” p ”. For operator K we accept inference rule (Nec), and axiom schemas (K), (D) and (T’):

$$(T') \quad K(Kp \rightarrow p)$$

Notice that Kp does not imply p , that is, we do not have the axiom schema (T): $Kp \rightarrow p$. Now, reliability of information source a in regard to sentence p is defined in that way:

$$RV_a(p) \stackrel{\text{def}}{=} K(E_a(Bp) \rightarrow p)$$

This definition is extended to reliability for all the sentences of the form $p(x)$:

$$RV_a(p(x)) \stackrel{\text{def}}{=} K(\forall x(E_a(Bp(x)) \rightarrow p(x)))$$

In a similar way an information source “ a ” is defined as a reliable information source in regard to completeness of sentence p iff the fact that p is true of the world implies that “ a ” has brought about that the data base believes p . The formal definition is:

$$RC_a(p(x)) \stackrel{\text{def}}{=} K(\forall x(p(x) \rightarrow E_a(Bp(x))))$$

It is assumed that the data base “knows” whether an information source has performed, or not, the insertion of some sentence p . This assumption is formalised by the axiom schemas (OBS1) and (OBS2):

$$(OBS1) \quad E_a(Bp) \rightarrow K(E_a(Bp))$$

$$(OBS2) \quad \neg E_a(Bp) \rightarrow K(\neg E_a(Bp))$$

Moreover, it is assumed that if the data base knows that some information source has brought about that the data base believes p , then the data base believes p . This is formalised by the axiom schema (B) ² :

$$(B) \quad K(E_a(Bp)) \rightarrow Bp$$

In this approach data base content db is represented by a set of sentences ³ of the form: $E_a(Bp)$ and $\neg E_a(Bp)$. As a consequence of that representation it is possible to distinguish a situation where agent a has inserted p and he has inserted r , formally represented by: $db_1 = \{E_a(Bp), E_a(Br)\}$, and another situation where agent a has inserted $p \wedge r$, which is represented by $db_2 = \{E_a(B(p \wedge r))\}$ ⁴. From an intuitive point of view, explicit content of the data base is considered as the result of performed insertions. Of course, if in the history of the data base, p has been inserted, which is formally represented by $E_a(Bp)$, and then it has been deleted, which is formally represented by $E_a(\neg Bp)$, then neither $E_a(Bp)$ nor $E_a(\neg Bp)$ should be in db .

Meta information mdb about reliability of information sources is represented by a set of sentences of the form: $RV_a(p)$ and $RC_a(p)$. For practical reasons it is impossible to represent in db all the insertions that have *not* been performed. A solution to this problem is to accept the following inference rule (COMP) whose meaning is: if from db it is not possible to infer that a has inserted p , then it is the case that a has not inserted p . That means that db is a complete description of all the performed insertions.

$$(COMP) \quad \frac{\not\vdash db \rightarrow E_a(Bp)}{\vdash db \rightarrow \neg E_a(Bp)}$$

For a given query $q(x)$ the formal definition of the **standard answer** is the set:

$$\{ a : \vdash db \rightarrow Bq(a) \}$$

The formal definition of a **valid answer** is the set:

²Notice that from (OBS1) and (B) we have: $E_a(Bp) \rightarrow Bp$, which is a restricted form of (T).

³In some contexts, db is supposed to represent the conjunction of all the formulas in this set.

⁴Notice that we do not have the axiom schema (C) $E_a p \wedge E_a q \rightarrow E_a p \wedge q$. The reason is that each insertion is supposed to have its own independent justification.

$$\{ a : \vdash db \wedge mdb \rightarrow Kq(a) \}$$

Information about completeness of an information source in regard to p is used to infer information about the validity of $\neg p$. For instance if we have: $db = \{E_a(Bp)\}$, $mdb = \{RV_a(p), RC_a(r)\}$, and the query is $q = p \wedge \neg r$, from (COMP) we have: $\vdash db \rightarrow E_a(Bp) \wedge \neg E_a(Br)$, and from the definitions of RV and RC we have: $\vdash mdb \rightarrow K(E_a(Bp) \rightarrow p) \wedge K(\neg E_a(Br) \rightarrow \neg r)$, then, from (OBS1) and (OBS2), we have: $\vdash db \wedge mdb \rightarrow K(p \wedge \neg r)$.

2 Reliable parts of the information

In this section we suppose that reliability is defined for some parts of the information, and we do not distinguish inserted data and derived data. For instance, if it assumed that data base content is reliable in regard to validity of $p \wedge r$, then, if the data base believes $p \wedge r$, we can infer that data base “knows” $p \wedge r$, whatever this data base belief is a consequence of an insertion of $p \wedge r$, or the consequence of insertion of p and of r , and whatever are the information sources who have performed the insertions.

The technical consequence of this new approach ⁵ is that reliability of parts of the data base is defined in that way:

$$RV'(p(x)) \stackrel{\text{def}}{=} K(\forall x(Bp(x) \rightarrow p(x)))$$

$$RC'(p(x)) \stackrel{\text{def}}{=} K(\forall x(p(x) \rightarrow Bp(x)))$$

and axiom schemas (OBS1) and (OBS2) are modified in:

$$(OBS1') \quad Bp \rightarrow K(Bp)$$

$$(OBS2') \quad \neg Bp \rightarrow K(\neg Bp)$$

⁵This approach corresponds to the approach adopted by Ami Motro in [5, 6].

Comment: for (OBS1') we have adopted a stronger axiom schema than: $Bp \rightarrow B(Bp)$, because we assume that the data base is correctly informed about what he believes; the same comments applies to (OBS2').

The data base content is represented here by a set of sentences db' of the form Bp , and the meta data base is represented by a set of sentences mdb' of the form $RV'_a(p)$ and $RC'_a(p)$. For the same reason as the one mentioned in the previous section, we accept the inference rule:

$$(COMP') \frac{\not\vdash db' \rightarrow Bp}{\vdash db' \rightarrow \neg Bp}$$

The definition of a standard answer to a query is defined like in the previous section. A **valid answer** to a query $q(x)$ is a sentence $p_1(x)$ such that:

$$\vdash \forall x(p_1(x) \rightarrow q(x)) \quad \text{and} \quad \vdash mdb' \rightarrow RV'(p_1(x))$$

and such that $p_1(x)$ is maximal for consequence relation, in the sense, that if there is another sentence $p'_1(x)$ that satisfies the same properties, and which is a consequence of $p_1(x)$, then $p'_1(x)$ is logically equivalent to $p_1(x)$.

If $p_1(x)$ is a valid answer to $q(x)$, then we have: $mdb' \rightarrow \forall x(Bp_1(x) \rightarrow Kq(x))$. This means that, if the data base believes $p_1(a)$ then the data base "knows" $q(a)$, or, in other terms, the standard answer to the query $p_1(x)$ gives a subset of the standard answer to $q(x)$ which is "guaranteed" to be true of the world (this is represented in figure 1 by the fact that the extension of $Bp_1(x)$ is included in the extension of $Kq(x)$).

A **complete answer** to a query $q(x)$ is a sentence $p_2(x)$ such that:

$$\vdash \forall x(q(x) \rightarrow p_2(x)) \quad \text{and} \quad \vdash mdb' \rightarrow RC'(p_2(x))$$

and such that $p_2(x)$ is minimal for consequence relation, in the sense, that if there is another sentence $p'_2(x)$ that satisfies the same properties, and which implies $p_2(x)$, then $p'_2(x)$ is logically equivalent to $p_2(x)$.

If $p_2(x)$ is a complete answer to $q(x)$, then we have: $mdb' \rightarrow \forall x(\neg Bp_2(x) \rightarrow K\neg q(x))$. This means that, if the data base does not believe $p_2(a)$ then the data base "knows" that $q(a)$ is false, or, in other terms, the standard answer

to the query $p_2(x)$ gives a superset of the standard answer to $q(x)$ such that all the element that are not in this set are “guaranteed” not to satisfy $q(x)$ (this is represented by the fact that the extension of $Kq(x)$ is included in the extension of $Bp_2(x)$).

An intuitive interpretation of valid answer $p_1(x)$ and complete answer $p_2(x)$, is that the answer to $q(x)$, if it would be evaluated on a valid and complete description of the world, would be respectively bounded down and up by the answers to $p_1(x)$ and to $p_2(x)$ (see figure 1).

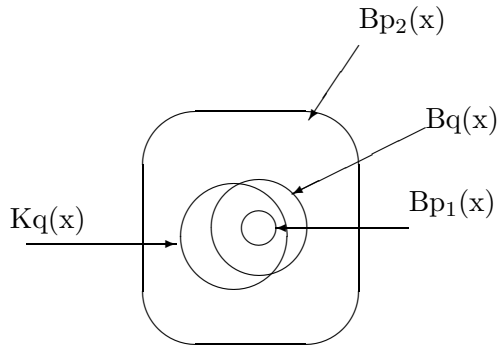


Figure 1: The standard answer $q(x)$ is bounded up and down by $p_1(x)$ and $p_2(x)$.

3 Relational data base context

According to the definitions of RV' and RC' , from $RV'(p)$ and $RV'(q)$ we can infer $RV'(p \wedge q)$, but we cannot infer $RV'(p \vee q)$ nor $RV'(\neg p)$. For similar reasons, from $RC'(p)$ and $RC'(q)$ we can infer $RC'(p \wedge q)$ and $RC'(p \vee q)$, but we cannot infer $RC'(\neg p)$.

In the context of relational data bases, since standard answers to queries are computed using relational algebra, the following axiom schemas are implicitly assumed:

$$(1) \quad B(p \vee q) \rightarrow Bp \vee Bq$$

$$(2) \quad \neg B(p) \rightarrow B(\neg p)$$

If we accept (1) and (2), from $RV'(p)$ and $RV'(q)$ we can infer $RV'(p \wedge q)$, $RV'(p \vee q)$ and $RC'(\neg p)$. In the same way, from $R'(p)$ and $RC'(q)$ we can infer $RC'(p \wedge q)$, $RC'(p \vee q)$ and $RV'(\neg p)$. These properties allow to reformulate the definition of valid answers and complete answers in the context of relational algebra. The benefit of this formulation is the possibility to use existing working relational data base systems, and also to have better performances in the evaluation of algebraic formulas than in using automated deduction techniques. Nevertheless, the determination of algebraic formulas that correspond to queries $p_1(x)$ and $p_2(x)$, in the previous sections, requires automated deduction. We briefly present now the axiomatics for the derivation of these formulas (a more detailed presentation can be found in [3]).

We shall denote by f, f', \dots formulas of the relational algebra language, and we denote by s a data base schema [7]. A data base schema extension, that is, a set of relations defined on a schema s , will be denoted se, db or w . We define an inclusion relation on algebraic formulas which is denoted by $f' \subseteq f$. We have $f' \subseteq f$ iff for every schema extension se of s we have: $f'(se) \subseteq f(se)$, where $f'(se)$ and $f(se)$ are the results of the evaluations of f' and of f on se .

The fact that $RV'(p(x))$ holds is represented in this context by a **valid view** $V(f)$, where f is a formula of relational algebra which is “equivalent” to $p(x)$ ⁶. More formally we have $V(f)$ iff for every situation if db is an extension of s that represents the data base state, and w is an extension of s that represents the world in this situation, then we have: $f(db) \subseteq f(w)$.

In a similar way the fact that $RC'(p(x))$ holds is represented by a **complete view** $C(f)$, and we have $V(f)$ iff for every situation if db is an extension of s that represents the data base state, and w is an extension of s that represents the world in this situation, then we have: $f(w) \subseteq f(db)$.

The fact that f' is a **valid subset** of f is denoted by $v_{\text{inf}}(f, f')$, and we have $v_{\text{inf}}(f, f')$ iff we have $V(f')$ and $f' \subseteq f$. The fact that f' is a **complete superset** of f is denoted by $c_{\text{sup}}(f, f')$, and we have $c_{\text{sup}}(f, f')$ iff we have $C(f')$ and $f \subseteq f'$.

The meta information mdb ” about validity and completeness of parts of the information is represented by a set of sentences of the form $V(f)$ and $C(f)$. A **valid answer** f_1 to a standard query f is an algebraic formula such that:

⁶We can easily define a translation from relational algebra to first order predicate calculus that preserves the logical meaning of sentences.

$$\vdash \text{mdb}'' \rightarrow v_{\text{inf}}(f, f_1)$$

and f_1 is maximal for the ordering relation \subseteq . A **complete answer** f_2 to a standard query f is an algebraic formula such that:

$$\vdash \text{mdb}'' \rightarrow c_{\text{sup}}(f, f_2)$$

and f_2 is minimal for the ordering relation \subseteq . The standard answer $f(w)$ evaluated on the correct representation of the world is bounded down and up by $f_1(\text{db})$ and $f_2(\text{db})$ ⁷.

The following axiom schemas allow to infer from mdb'' valid views and complete views.

⁷Notice that, if we have $v_{\text{inf}}(f, f_1)$ and $v_{\text{inf}}(f, f_2)$, then we have $v_{\text{inf}}(f, f_1 \cup f_2)$, and, if we have $c_{\text{sup}}(f, g_1)$ and $c_{\text{sup}}(f, g_2)$, then we have $c_{\text{sup}}(f, g_1 \cap g_2)$.

- (a1) $V(f) \wedge V(g) \rightarrow V(f \cup g)$
- (a2) $V(f) \wedge V(g) \rightarrow V(f \cap g)$
- (a3) $V(f) \wedge V(g) \rightarrow V(f \times g)$
- (a4) $V(f) \wedge C(g) \rightarrow V(f - g)$
- (a5) $V(f) \rightarrow V(s_c f)$
- (a6) $V(f) \rightarrow V(\pi_p f)$

- (b1) $C(f) \wedge C(g) \rightarrow C(f \cup g)$
- (b2) $C(f) \wedge C(g) \rightarrow C(f \cap g)$
- (b3) $C(f) \wedge C(g) \rightarrow C(f \times g)$
- (b4) $C(f) \wedge V(g) \rightarrow C(f - g)$
- (b5) $C(f) \rightarrow C(s_c f)$
- (b6) $C(f) \rightarrow C(\pi_p f)$

where s_c and π_p respectively denote the selection operator and the projection operator.

4 Conclusion

We have presented formal definitions of valid information sources and complete information sources in modal logic. The notions of valid answers and complete answers give information about the link between the standard answer, computed from the data base content, and the “true” answer. However, this “true” answer is in fact only a more trustworthy answer than the standard answer. It would be interesting to extend this work by generalising the idea from two levels of trustworthiness to several levels which are partially ordered.

We have shown how this formalisation can be adapted to the Motro’s approach of valid views and complete views of relational data bases. In the context of relational algebra we can compute valid answers and complete answers for any kind of formula of relational algebra, while Motro’s results are limited to queries without union operator and without difference operator.

Acknowledgements: we would like to thanks Jose Carmo for his fruitful comments.

References

- [1] L. Cholvy, R. Demolombe, and A.J. Jones. Reasoning about the safety of information: from logical formalisation to operational definition. In *Proc. of 8th International Symposium on Methodologies for Intelligent Systems*, 1994.
- [2] B. F. Chellas. *Modal Logic: An introduction*. Cambridge University Press, 1988.
- [3] R. Demolombe. Validity Queries and Completeness Queries. In *Proc. of 9th International Symposium on Methodologies for Intelligent Systems*, 1996.
- [4] R. Demolombe and A. Jones. Deriving answers to safety queries. In R. Demolombe and T. Imielinski, editors, *Nonstandard queries and answers*. Oxford University Press, 1994.
- [5] A. Motro. Completeness information and its application to query processing. In *Proc. of 12th International Conference on Very Large Data Bases*, 1986.
- [6] A. Motro. Integrity = validity + completeness. *ACM TODS*, 14(4), 1989.
- [7] J. D. Ullman. *Principles of Database Systems. Vol1 and Vol2*. Computer Science Press, 1988.