

A common logical framework to retrieve information and meta information

Robert Demolombe and Andrew Jones

ONERA/CERT, Toulouse, France demolomb@tls-cs.cert.fr

Department of Philosophy, University of Oslo, Norway jones@filosofi.uio.no

Introduction

The development of networks connecting a large number of distributed databases will shortly make available extremely large amounts of data. For instance there is currently a project whose objective is the interconnection of one hundred existing molecular-biology databases.

This kind of interconnection raises two issues: How is relevant information to be retrieved? And what is the validity of information obtained from different sources, assuming that some of them are reliable and others are not? These issues are much more important in this context than in the context of accessing a unique database because each database has been designed independently, and they are also updated independently. This abstract presents some attempts to provide solutions to these two problems in the uniform formalism of Modal Logic.

The main results, at the present time, are at the theoretical level. Nevertheless at the end of the abstract some guidelines are described which indicate how these results could be used in potential implementations.

Reasoning about topics for information retrieval

To access information stored in a large set of connected databases we have to consider two different aspects: (i) the syntactical heterogeneity of representation languages, for instance some sources have information represented in the Relational Model, and others in the Entity-Relationship Model, or in the Object Oriented Model, and (ii) the heterogeneity of database contents. In this section we only consider the heterogeneity of database contents.

Usually in this context there is a large number of types of information, and users are unable to know about all of them. So, it may happen that users ignore the existence of data that are relevant for their purposes. The proposed solution to this problem is to have “on the top” of each information source a description of its information content in terms of topics, comparable to the use of key words to describe the content of a document. The concept of key words has been

extensively used in the context of information retrieval for textual documents, but it has not been applied, as far as we know, to structured data.

If, in a first step, the problem of syntactical heterogeneity is ignored, we can assume that the overall information about all the database contents is represented by sentences of Propositional Calculus. We have defined in this context a specific Modal Logic to represent the link between a sentence p and a topic T (Cazalens, Demolombe, & Jones 1992; Cazalens & Demolombe 1992). In this logic sentences of the form $A_T(p)$ can be read: “the sentence p is about the topic T ”. The axiomatisation of this modality is not trivial, and it may be controversial. There is at least a consensus for accepting the following inference rule: (RE) from $\vdash_{CPC} p \leftrightarrow q$ and $\text{Var}(p)=\text{Var}(q)$, infer $\vdash A_T(p) \leftrightarrow A_T(q)$, where $\text{Var}(p)$ represents the propositional variables in p . The constraint on $\text{Var}(p)$ and $\text{Var}(q)$ prevents one of inferring, for example, that the sentences $a \vee \neg a$ and $b \vee \neg b$ are about the same topic whatever a and b are about. Many people are ready to accept in addition the axiom schema: (NEG) $A_T(p) \rightarrow A_T(\neg p)$. The axiom schema: (K) $A_T(p) \wedge A_T(p \rightarrow q) \rightarrow A_T(q)$ is more controversial, though it is much weaker than the inference rule: from $\vdash_{CPC} p \rightarrow q$ infer $\vdash A_T(p) \rightarrow A_T(q)$, which surely has to be rejected.

We have defined a semantics for this logic in terms of Minimal Modal Models (see (Chellas 1988)), where propositions may have three different truth values (we use Bochvar’s three valued logic, see (Gabbay & Guenther 1984)). The three valued interpretation of sentences allows the removal of the rule of substitution of equivalent sentences. The logic in its weakest form has been proved to be sound and complete.

If the content of each information source is described by a set of assumptions of the form $A_T(p)$, this logic allows inferences to be made regarding the topics of other sentences, and it can be used to retrieve pieces of information that are about a given topic. In this approach user’s requests can be expressed in terms of topics, that is in terms of concepts that are close to those used in natural language, and they can be easily

understood by all users. So they do not need to know the details of the description of each information source content.

Reasoning about safety of retrieved information

For different reasons - for instance because the frequency of data updates can be per day, per week or per year, depending on information sources, or because some information sources have valid justifications for inserting some kinds of data - some sources are quite reliable in regard to information validity, while others are not. In this situation it is not easy to determine the reliability of an answer to a complex request when this answer is obtained from several information sources.

For reasoning about reliability of agents we have formally defined the notion of reliability of an agent in Modal Logic as: $RV_a(p) \stackrel{\text{def}}{=} K_{\text{adm}}(EB_a(p) \rightarrow p)$ (Demolombe & Jones 1994). The *definiens* can be read: “the administrator adm knows that if p is an explicit belief of a database which has been inserted by agent a, then p is true of the world”, and, if this is the case, agent a is said to be reliable in regard to the validity of p (which is our reading of the *definiendum* on the left hand side of the identity sign). The rationale for this definition is that it is assumed that explicit database beliefs are the result of the performance, by some agent or agents, of acts of inserting information into the database; and some insertion actions are performed by reliable agents, while others are not.

The only inference rule for $EB_a(p)$ is the rule of equivalence: from $\vdash_{\text{CPC}} p \leftrightarrow q$ infer $\vdash EB_a(p) \leftrightarrow EB_a(q)$. In particular the situation where we have $EB_a(p) \wedge EB_a(q)$ is not equivalent to the situation where we have $EB_a(p \wedge q)$, because $EB_a(p)$, $EB_a(q)$ and $EB_a(p \wedge q)$ are the results of performing three different insertion actions that may have three independent justifications. The semantics of $EB_a(p)$ has been defined in the framework of Minimal Modal Models, and we have proved the soundness and the completeness of the axiomatics. The modality K_{adm} is defined as usual by $K_{\text{adm}}(p) \stackrel{\text{def}}{=} p \wedge B_{\text{adm}}(p)$, (this is a simplification of the knowledge concept, of course, but it is adequate for present purposes), where B_{adm} obeys the axioms of a normal modal system of type KD. Notice that the properties of $RV_a(p)$ are extremely weak; in particular, as a consequence of the above comment, $RV_a(p \wedge q)$ does not imply $RV_a(p)$, and from $RV_a(p)$ and $RV_a(q)$ we cannot infer $RV_a(p \wedge q)$.

To collate information obtained from different sources we have the axiom schema: $EB_a(p) \rightarrow Bp$, where Bp can be read: “the union of the set of information sources believes p”.

If the information about source reliability is represented by a set of assumptions of the form $RV_a(p)$, and if data explicitly stored in each source is represented by a set of assumptions of the form $EB_a(q)$, the answer

to a standard query q is “yes” if $K_{\text{adm}}(B(q))$ can be derived from this set of assumptions, and the answer to the query: “is this standard answer guaranteed to be reliable?” is “yes” if $K_{\text{adm}}(q)$ can be derived from the same set of assumptions. The second answer can be viewed as a kind of meta answer since it gives information about the status of the standard answer.

Description Logics as a framework for possible implementations

In (Cuppens & Demolombe 1988) is presented an implementation of topics to provide users with cooperative answers in the context of standard Relational databases. In this implementation, instead of a modality of the form $A_T(p)$, we have a predicate of a FOL language of the form $\text{Topic}(T,p)$ whose intended meaning is the same, and the implementation is in Prolog.

However many people think that FOL is not the best formalism when we have to cope with the problem of syntactical heterogeneity, for instance when we have to represent in the same formalism parts of the information which are represented in the Relational Model, and other parts represented in an Object Oriented Model. A better formalism for this purpose seems to be a Description Logic derived from KL1 (Blanco *et al.* Valencia 1992; Brachman 1977; Illaramendi, Blanco, & Goni 1991). Even if the expressive power has strong limitations Borgida has shown in (Borgida 1995) that a substantial fragment of FOL can be translated into a Description Logic, namely first order sentences formed with unary and binary predicates, and with no more than three distinct variables. Moreover there exist running implementations of Description Logics: see for example the systems BACK (von Luck *et al.* 1987) or MOTEL (Hustadt *et al.* 1993).

We have planned to investigate the extension of Description Logics with topics. The idea is to assign topics to roles of concepts in a way similar to the way they are assigned to sentences in Propositional Calculus. In this extension it would be possible to retrieve the value of the roles related to a given topic for all the objects in the extension of some concept.

For queries about the reliability of standard answers, we have already implemented in Prolog (Cholvy, Demolombe, & Jones 1994) a simplified version of the logic presented in (Demolombe & Jones 1994), and we plan to investigate an extension of the definition of reliability of information sources in terms of topics. For instance $RV_a(T)$ could be interpreted as “for every role r related to the topic T, the agent a is reliable in regard to the values of r”.

Another potential extension is to define a notion of relative reliability. That is, an agent a might be known to be at least as reliable as an agent b in regard to a sentence p, or in regard to a topic T. Finally another natural extension is to define a hierarchy of topics, as people do in the context of information retrieval with

thesaurus. This hierarchy should not be confused with hierarchies of concepts in Description Logics. Indeed the extension of a topic is a set of sentences, namely all the sentences that are about this topic, while the extension of a concept is a set of objects in the world.

Concluding summary: some preliminary theoretical results have been established in order to clarify the semantics of the notions of topic and of reliability, and meta data about these notions can be formally represented by sentences of the form $A_T(p)$ or $RV_a(p)$. The formalism used for this purpose employs both Normal and Classical Modal Logics, in the sense of Chellas (Chellas 1988). However, for an implementation in the context of heterogeneous existing databases it will better to adapt these results to Description Logics.

Acknowledgements: This work was partially supported by the ESPRIT Basic Research Action MED-LAR2.

References

- Blanco, J.; Illaramendi, A.; Perez, J.; and Goni, A. Valencia, 1992. Making a federated system active. In *Proc. of Int. Conference on Database and Expert Systems Applications*.
- Borgida, A. 1995. On the Relationship between Description Logic and Predicate Logic Queries. In *Proc. of 5th International Conference on Database Theory*.
- Brachman, R. 1977. *A Structural Paradigm for Representing Knowledge*. Ph.D. Dissertation, Harvard University.
- Cazalens, S., and Demolombe, R. 1992. Intelligent Access to Data and Knowledge Bases via User's Topics of Interest. In *IFIP Congress'92, Madrid*.
- Cazalens, S.; Demolombe, R.; and Jones, A. 1992. A Logic for Reasoning about Is About. Technical report, ONERA-CERT.
- Chellas, B. F. 1988. *Modal Logic: An introduction*. Cambridge University Press.
- Cholvy, L.; Demolombe, R.; and Jones, A. 1994. Reasoning about the safety of information: from logical formalization to operational definition. In *Proc. of 8th International Symposium on Methodologies for Intelligent Systems*.
- Cuppens, F., and Demolombe, R. 1988. Cooperative Answering: a methodology to provide intelligent access to Databases. In *Proc. of 2d Int. Conf. on Expert Database Systems*.
- Demolombe, R., and Jones, A. 1994. Deriving answers to safety queries. In Demolombe, R., and Imielinski, T., eds., *Nonstandard queries and nonstandard answers*. Oxford University Press.
- Gabbay, D., and Guenther, F. 1984. *Handbook of Philosophical Logic, Vol.3*. D.Reidel Publishing Company.

Hustadt, U.; Nonengart, A.; Schmidt, R.; and Timm, J. 1993. MOTEL Users Manual. Technical Report MPI-I-93-236, Max Planck Institut für Informatik.

Illaramendi, A.; Blanco, J.; and Goni, A. 1991. A uniform approach to design a federated system using BACK. In *Proc. Terminological Logic Users Workshop*.

von Luck, K.; Nebel, B.; Peltason, C.; and Schmiedel, A. 1987. The Anatomy of the BACK System. Technical Report Report 41, Technical University of Berlin.