DATA INTEGRATION AND USER MODELLING:
An approach based on Topic Maps and Description Logics

Mourad Ouziri
University Paris V René Descartes, 45 rue des Saints-Pères 75270 Paris Cedex 06

Christine Verdier[(1)], André Flory[(2)]
LIRIS-Lyon 2[(1)],LIRIS-INSA de Lyon[(2)], bât. Blaise Pascal, 7 av. J. Capelle, 69621 Villeurbanne Cedex

Keywords: Data integration, Intelligent modelling of user profile, Topic Maps, Description logics

Abstract: In the framework of intelligent information systems design, we present an intelligent data integration and user profile modelling. Our approach uses jointly Topic Maps and Description Logics. Topic Maps are used to represent semantic of distributed data. Using the formalized semantic, the distributed data is merged into one repository. Then, Description Logics are used over this repository to compute implicit semantic relations using logic reasoning such as subsumption. We present then a new user profile management which uses qualifying attributes rather than identifying attributes. Description Logics are used to formalize profiles in order to maintain consistency of right attribution to profiles.

1 INTRODUCTION

Web-based applications require access to multiple sources of data to supply relevant information. This relevance is due to the research, the choice of the user and his browsing. So, searching the good, relevant, up-to-date information in this huge volume of data is a complex task which is difficult to be done manually. Many systems are studied in the semantic web community about this subject. The main question is how to reconcile the large volume of data and the precise choice of a user. We remark that the Web offers heterogeneous sources concerning data heterogeneity, structure heterogeneity and semantic heterogeneity. Data heterogeneity is due to the use of different data types (integer, string, etc.), measurement units (meters, miles, etc.) and scales (month, day, year, etc.) associated to the values. This type of heterogeneity is solved with conversion rules defined between the data sources and the federated schema. Structural heterogeneity consists in using different data structures to represent the same object. For example the object Adress can be represented by a tuple (number, street, zip, city) or by a string. Semantic heterogeneity is crucial and is seen at conceptual level. Several cases can be noticed. An object can be represented as a attribute, or as an entity. So the merging of different data source can deal with semantic association problems. Semantic heterogenity can be observed also with synonyms, homonyms, generalized ans specialized terms: for example drug and medicine.

So we propose in this paper a web-based interface to query multiple data sources that have been merged. Data semantic is garanteed. To build this interface, we use Topic maps to represent knowledge, description logics to provide automatic reasoning on the knowledge and to help in user profiles.

We present in this first paragraph the related works, then we continue with a presentation of the Topic Maps and Description Logics approach and then we finish with the presentation of our system.

2 RELATED WORKS

2.1 Prior data integration approaches

Two main approaches concern the database community. The first approach concerns the data integration in the query. The query language is made to formulate and process multi-databases queries. The heterogeneity is resolved through the query

Ouziri M., Verdier C. and Flory A. (2005).
DATA INTEGRATION AND USER MODELLING: An approach based on Topic Maps and Description Logics.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 171-175
DOI: 10.5220/0002532701710175

directly by the user. This approach is really difficult to process when we have to query large and numerous databases (Breitbart, 1990). Federated databases consist to integrate the different views of databases into a unique and global conceptual schema (Sheth, 1990). Queries are specified on the global schema and then are parsed and shared into sub-queries. Each sub-query is sent only to one database. The major drawback of this approach is that it is often difficult to build a global schema because of data heterogeneities. The third approach is a wrapper-mediator approach (Molina, 1997) (Karp, 1995). A wrapper is designed for each datasource and its role consists in translating data into the common language of the mediator. Then, the mediator uses the data provided by the wrapper to build a global schema. Sub-queries are evaluated by the respective wrappers.

2.2 Ontology-based approach

Ontology represents a pertinent way to resolve problems related to heterogeneity. Ontology plays an important role in the knowledge representation. It allows sharing semantic interpretation of structural units. Ontology is mainly used like a global schema of datasources and a query interface. Ontology concepts are linked to datasources through a global meta-model. Theses links are used ot identify the relevant datasources and to transform queries into sub-queries with the ontology. In the literature, three main ontology approaches are proposed (Wache, 2001) (Mena, 1998): simple, multiple and hybrid ontologies. Buster (Meyer, 2001) uses hybrid ontology. In this system, the ontology is seen as a knowledge base on which semantic integration is based. The terms of a datasource are defined with a local ontology. The integration in Buster consists to define a common ontology which is used to annotate the terms resulted from the first annotation. Multiple ontologies are used in Observer (Mena, 2000). It associates an ontology to each datasource. This association is formalized by links between the concepts of the ontology and the terms of the datasource. The datasource integration is done with semantic links (synonymy, hyponymy, disjunction, etc.) between the concepts of the ontologies.

2.3 Logic-based approaches for data integration

Description logics (DL) are used to represent dependencies between concepts in different datasources in (Catarci, 1993). On these dependencies are generated reasoning. They are represented at intentional and extensional levels and make the connections between concepts belonging to different datasources. The intentional and extensional dependencies are distinctly processed because two equivalent concepts in different datasources do not imply the equivalence of the extensions. The knowledge base inter-schemas is realized with assertions which specify equivalence and subsumption relations. Therefore, a global graph is generated from the assertions and used to evaluate the queries. In (Levy, 1999), two logic-based approaches are presented. The first approach is called global as view. The global relation is a set of relations. Each relation is described with the relations of the datasources, which indicates how to obtain the instances. Queries are specified on the global relations are rewrited using their descriptions. The second approach is called local as view. The datesources are described with the global relations. This opposite approach is more interesting when the updating of the datasources is frequent. Other logic-based solutions are presented in (Goasdoué, 2000) and (Calvanese, 1998).

2.4 Web-based approach for data integration

Data integration represents an important task to access and query data in a coherent way. In the Web, data are represented with semi-structured HTML or XML models. For HTML documents, data integration consists to link HTML (or XML) documents with each other by hyperlinks. This is a static and rigid approach because semantic relationships are not considered. XML-based data integration is realized with a query language for querying multiple XML documents using one query (Cohen, 2003) or by providing a uniform view of multiple XML documents (Camillo, 2003). To integrate XML documents, a mechanism to identify multiple instances of a same real object is proposed in (De Brum, 2002). Semi-structured data models, OEM (Goldman, 1996) and XML, are used in data integration process (Papakonstantinou, 1995) (Gardarin, 2002). This type of integration is not useful for expressing semantic. XML does not give any semantics about taggs. The suitability of DL for data integration is illustrated in some projects: SIMS (Arens, 1993) and PICSEL (Goasdoué, 2000). In these systems, datasources are linked together and the knowledge is expressed manually.

3 BACKGROUND

3.1 An overview of Topic Maps

Topic Maps (Sigel, 2000) is a paradigm used ot formalize and organize human knowledge to make creation and retrieval easier in computer processing. It is also used as a mechanism for representing and optimizing resource access. As semantic networks, Topic Maps builds a structured semantic link network on these resources (Freese, 2000). A topic map is built with topics in a networked form. A topic can be anything regardless whether it exists or not. It is the formal representation of any subject, abstract or real in a computer system such as a person, John, the earth, the planet, etc. Topics are linked together by associations which express some semantics. Topic Maps applications define the nature of the associations and the role played by the topics in these associations. For example, the topics customer and doctor can be related by the association examine with the respective roles patient and doctor. So, semantics is specified by the association and especially by the roles. Associations are used ot express knowledge between topics and not between occurrences. So topics and associations represent the abstract part of a topic map. The concrete part is represented by occurrences which are resources linked to topics. In 1999, a standard defining the Topic Maps model and its syntax was edited by ISO/IEC 13250 (ISO/IEC, 1999).

3.2 An overview of Description Logics

Description Logics (DL) are logics developed to represent complex hierarchical structures and make reasoning facilities on these structures (Borgida, 1995). DL are used to build ontologies for semantic web (Horrocks, 2002). A DL is composed of two parts: abstract knowledge (Tbox) and concrete knowledge (Abox). Concrete knowledge represents a set of facts which are expressed by assertions on individuals. Abstract knowledge is expressed with concepts and roles. Concepts are unary predicates which represent an abstraction of individuals. Roles are binary predicates. They represent relations between concepts.

4 DATA INTEGRATION USING TOPIC MAPS AND DL

We present an integration process that combines Topic Maps and DL to build a semantic data integration. First, data sources are modeled with Topic Maps to represent distributed knowledge. Secondly, we use DL to represent the constraints. Constraints are useful in data integration to deduce implicit relations between concepts. The integration process is shown in the following figure:

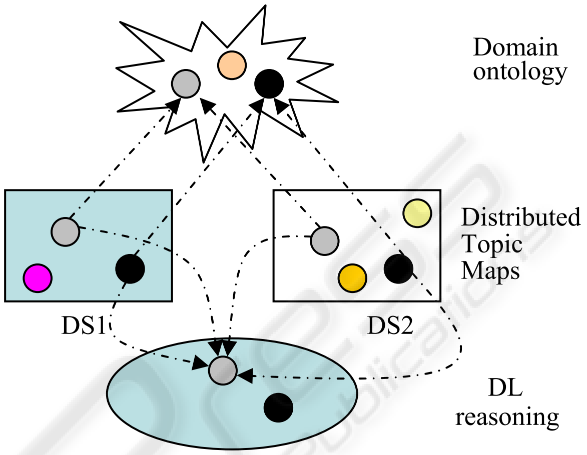

Figure 1: Data integration using TM and DL

The datasources DS1 and DS2 are represented with Topic Maps. Two datasources types are considered: relational databases and XML documents. For relational databases, topics represent tables and attributes. For XML documents, topics represent tags. Tables (or tags) are represented by topic and attributes (or sub-tags) too. These topics are connected to the ontological concepts they represent using the subjectIndicatorRef tag. Semantic integration based on Topic Maps is so natural. It consists to merge topics referencing the same ontological concept into one topic in the federated topic map.

Table 1: Representation of tables (tags) and attributes (sub-tags)

```
/*a relational table or an XML tag*/
<topic name=person
  <subjectIdentity>
    <subjectIndicatorRef xlink:href=http://www.ont.org/PSI/medicalOntology.daml# human”/>
  </subjectIdentity>
</topic>
/*a table attribute or a sub-tag*/
<topic name=adress>
  <subjectIdentity>
    <subjectIndicatorRef xlink:href=http://www.ont.org/PSI/medicalOntology.daml# adress”/>
   </subjectIdentity>
</topic>
```

Then, description logic reasoning is used to provide a consistent federated topic map because it automatically computes implicit relations between concepts. Let’s consider that the datasource DS1 and DS2 contain the two descriptions:

DS_1 : person ≐ ∀ Name.String □ ∀ Adress.String □ …

DS_2 : patient ≐ ∀ Human □ ∀ disease.String □ >=1 disease □ …

As traditional logics, DL is not able to make reasoning on distributed knowledge bases. If we do not use topic maps (especially the subjectIdentity concept), DL reasoning does not infer any relation between person and patient.

So $\text{person} \sqcap \text{patient} \sqsubseteq \varnothing$

Semantically, it is not correct.

Now let's consider that person is numan through the subjectIdentity tag of the topic map. So:

$$DS_1 : \text{human} \doteq \forall\ \text{Name.String} \sqcap \forall\ \text{Adress.String} \sqcap \ldots$$

$$DS_2 : \text{patient} \doteq \text{human} \sqcap \forall\ \text{disease.String} \sqcap >=1\ \text{disease} \sqcap \ldots$$

Therefore, DL reasoning infers that

$\text{patient} \sqsubseteq$ person and gives a semantic connection between the datasources DS1 and DS2.

5 USER PROFILES

We propose in this section to describe our system build on the previous integration data and useful for defining user profiles.The maor idea is to represent user profiles with qualifying attributes and not with traditional identity attributes, user-id and password. Therefore, user profile are represented and managed at the semantic layer. We use DL to define automatic and coherent management user profiles.

Let's consider the following DL knowledge base of user profiles:

Table 2: Example of a TBox representing user profiles

$\text{Employ} \doteq \forall\ \text{name.String} \sqcap \forall\ \text{category.String} \sqcap \forall\ \text{specialty.String}$
$\text{Doctor} \doteq \text{Employ} \sqcap \text{Category} = \text{« medical »}$
$\text{Specialist} \doteq \text{Doctor} \sqcap >=1\ \text{specialty}$
$\text{Generalist} \doteq \text{Doctor} \sqcap =0\ \text{specialty}$
$\text{OverSpecialized} \doteq \text{Doctor} \sqcap >3\ \text{specialty}$

These 5 profiles are described by the qualifying attributes *name*, *category* and *specialty*. With the DL reasoning, theses profiles are automatically organized by subsumption relationship.

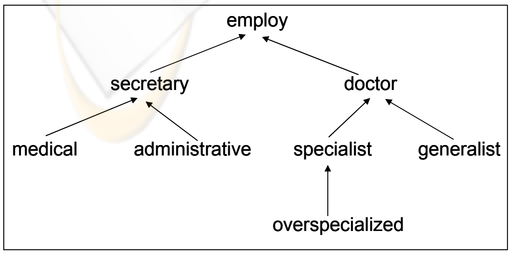

Figure 2: Automatic profiles organization

When a profile is modified, its hierarchy is automatically updated. If a new profile is defined, it is inserted in the hierarchy at the most adequate position according to its description. This hierarchy is very important for a coherent attribution of access rights to the profiles. If we grant the access right d1 to the profile Doctor, reasonably, this access right should be also granted to the profiles Specialist, Generalist and Overspecialized which are Doctors:

$$\text{specialist} \sqsubseteq \text{doctor}$$

$$\text{generalist} \sqsubseteq \text{doctor}$$

$$\text{overspecialized} \sqsubseteq \text{doctor}$$

As a general rule, the following formula should be checked for all the profiles:

$\forall$ P1,P2$\in${profiles}, if P1 subsumed_by P2 then P2.access_rights $\subseteq$ P1.access_rights

We notice that the most specific profile has the maximum access rights because it gets those of its parents. This formula allows to grant the access rights to the profiles in a coherent and automatic way. Traditionally, access rights have to be specified for each profile. In our system, the granting is automatically performed without human intervention. As the profiles and the access rights are defined, the system identifies the users and automatically assigns them to the most adequate profile. For that, users are also described with qualifying attributes. Unlike profiles, the users are described in the Abox. Let's consider the following Abox:

Employ (e1)	Employ (e2)
Name (e1, « John »)	name (e2, « Peter »)
Category (e1, « medical »)	category (e2, « medical »)
Specialty (e1, « radiology »)	

So the profiles are represented in the Tbox and the individuals are represented in the Abox. With the DL reasoning, the system classifies the users according to the Tbox and Abox with the biggest access rights. The users John and Peter are employees and respectively with a specialty "radiology" and no specialty. So they will be classified in the profiles Specialist and Generalist.

6 CONCLUSION

We have jointly used Topic Maps and DL for semantic data integration. The most advantage is to perform reasoning on distributed knowledge. Translating knowledge from Topic Maps to DL is not always possible. Topic Maps formalism includes a constraint language and constructors that is very expressive. But DL are based on reasoning algorithms defined on a set of constructors. Thus, the expressivity of DL is restricted to the reasoning algorithms. It represents a paradox between the two formalisms. In our future work, we try to improve the constraint language specification of Topic Maps. Constructors will be built to define constraints on a Topic Maps knowledge base and perform automatic reasoning.

REFERENCES

Arens, Y., Chee, C., Hsu, C., Knoblock, C., 1993. Retrieving and integrating data from multiple information sources. *In Journal of Intelligent and Cooperative Information Systems, vol.2, n°2, p. 127-158*

Borgida, A., 1995. Description Logics in data management. *IEEE Trans. On knowledge and data engineering, vol.7, n°5, p.671-682*

Breitbart, Y. and alii, 1990. Final report of the workshop on multidatabases and semantic interoperability. *University of Kentucky, Department of Computer Science, Lexington, KT 188-91, November 2-4.*

Calvanese, D., et alii, 1998. Knowledge representation approach to information integration. *In proc. Of AAAI Workshop on AI and information integration, p. 58-65*

Camillo, S.D., Heuser, C.A., Mello, R.S., 2003. *Querying heterogeneous XML sources through a conceptual scheman. Proc. of ER, p. 186-199.*

Catarci, T., Lenzerini, M., 1993. Representing and using interschema knowledge in cooperative information systems. *Journal of intelligent and cooperative information systems, vol.2, n°4, p.375-398*

Cohen, S., et alii, 2003. Xsearch: a semantic search engine for XML. *VLDB, p. 45-56*

De Brum Saccol, D., Heuser, C.A., 2002. Integration of XML data. *Proc. Of EEXTT, p. 68-80*

Freese, E., 2000. Using Topic Maps for the representaion, management and discovery of knowledge. *XML Europe 2000, Palais des congrès, Paris, 12-16 June.*

ISO/IEC 13250, 1999. *Topic Maps*, Dec. ISO/IEC FCD

Gardarin, G., Mensch, A., Tomasic, A., 2002. An introduction to the e-XML data integraiton suite. *Proc. Of EDBT, p. 297-306*

Goasdoué, F., Lattes, V., Rousser, M.C., 2000. The user of CARIN language and algorithms for information integration: the PICSEL project. *International Journal of Cooperative Information Systems (IJCIS), vol.9, n°4, p. 383-401*

Goldman, R., Chawathe, S., Crespo, A., McHugh, J., 1996. *A standard tectual interchange format for the object exchange model (OEM).* Department of computer science, Standford University, California, USA, 5 p.

Horrocks, I., Patel-Schneider, P.F., Van Harmelen, F., 2002. Reviewing the design of DAML+OIL: an ontology language for the semantic web. *Proc. Of 18th National Conf. On Artificial Intelligence, p. 792-797*

Karp, P., 1995. A strategy for database interoperation. *Journal of computational biology2, p.573-586*

Levy, A.Y., 1999. Logic-based techniques in data integration. *Workshop on logic-based artificial intelligence, Washington DC, June 14-16, 27 p.*

Mena, E., Illarramendi, A., Kashyap, V., Sheth, A., 2000. Observer: an approach for query processing in global information systems based on interoperation across pre-existing ontologies. *In the Int. Journal Distributed and Parallel Databases (DAPD), vol.8, n°2, p.223-271*

Mena, E., Kashyap, V., Illarramendi, A., Sheth, A., 1998. Domain specific ontologies for semantic information brockering on the global information infrastructure. *Int. Conf. On Formal ontologies in information systems. FOIS'98, Italy, 15 p.*

Meyer, R., et alii, 2001. Intelligent brockering of environmental information with Buster System. In Informaiton Age Economy: proc. *Of the 5th Int. Conf. 'Wirtschaftsinformatik', Physica-Verlag, 8p.*

Molina, H.G., et alii; 1997. The TSIMMIS approach to mediation: data models and languages. *In Journal of intelligent information systems, p.117-132*

Moore, G., Nishikawa, M., 2003. *The Topic Maps Constraint Language*. ISO/IEC 13250. Available at http://www.isotopicmaps.org/tmcl

Papakonstantinou, Y., Garcia-Moulina, H., Widom, J., 1995. Object exchange across heterogeneous information sources. *Proc. Of IEEE int. Conf. On Data Engineering, p.251-260*

Pepper, S. Moore, G., 2001. *XML Topic Maps (XTM).* 1.0. TopicMaps.Org Authoring Group, Aug. 2001. Available at: http://topicmaps.org/xtm/index.html

Sheth, P., Larson J.A., 1990. Federated database system for managing distributed, heterogeneous and autonomous databases. *ACM Computing surveys, ACM Press, vol.23, n°3, p.183-236*

Sigel, A., 2000. Towards knowledge organization with Topic Maps. *XML Europe 2000, Palais des congrès, Paris, 12-16 June.*

Wache, H., et alii, 2001. Ontology-based integration of information-a survey of existing approaches. *Workshop ontologies and information sharing. IJCAI 2001, 10 p.*