Accessing data in the semantic web: An intelligent data integration and navigation approaches

Mourad Ouziri
CRIP5 – René Descartes University
45, rue des Saints-Pères
75270 Paris Cedex 06, France
mourad.ouziri@univ-paris5.fr
http://www.math-info.univ-paris5.fr/crip5/Pageperso.php?numperson=369

Abstract. We present an original navigation approach to explore data in the framework of the semantic web. Before navigation, data are represented and integrated by jointly using Topic Maps and description logics. Our navigation approach improves the traditional web navigation with two specificities. First, the navigation paths are semantic instead of structural links. Second, it is a subject-centric instead nothing-centric. These two facilities increase the efficiency of information retrieval in the Web. They are implemented in an adaptive interface, which browses, gradually, data as a concept map with rest to user navigation.

1 Introduction

The Web presents a huge amount of heterogeneous collections of data. Application developments, such as e-health, e-business and digital libraries, require access to multiple sources in order to get complete and consistent information. Two many search modes can be used to search information in the Web, query language and navigation. In the first mode [1], user must be expert of the query language and know partial or complete data structure to formulate a query. Considering the number of datasources, query formulation is a fastidious task. Navigation is the most search mode used because it is natural, intuitive and does not require learning any query language syntax.

However, navigation is not efficient in huge collection of data, because it causes the well-known concept of *lost in hyperspace*. That is, user often takes progressively non relevant navigation paths. After some navigation steps, he is confronted to nondesired information space. In this situation, user does not know how to continue the navigation. For example, in e-health doctors may want to get information about a

Please use the following format when citing this chapter:
Ouziri, Mourad, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 119–128

patient. Than, they start search by selecting the given patient from a Web page giving all patients. They navigate from the given patient by following the examination link to get patient exams. In most cases, this Web page presents exaninations of all patients. After that, doctor follows navigation to get practitioners who made the examinations. Suppose that doctor would obtain result of an exam made by a known practitioner. Therefore, doctor selects the given practitioner and navigates to the Web page giving exam results. In traditional navigation, results given in this Web page should be related to all exams made by the given pratitioner for all his patients.

Moreover, when there are many navigation paths from a page, doctor may follow bad navigation paths. Addition of these mistakes brings doctor to nondesired Web pages. Unfortenatly, there is no reverse navigation to reach the nearest Web page previously navigated in order to restart more relevant navigation paths.

In this paper, we propose to integrate data about a defined-domain application and explore the integrated data space using a subject-oriented navigation. Web data are distributed and heterogeneous. To deal with distribution, we use a distributed knowledge representation, Topic Maps, to construct a conceptual graph upon the distributed data of the Web. To deal with heterogeneity, we use description logics to merge data according to their descriptions.

Then, the integrated data is used by a navigation interface. This interface adapts the integrated data to the user profile and allows exploring data in an efficient navigation mode.

2 Related Works

In order to assist user during navigation, navigation interfaces become more and more adaptive. In the context of the Web, these navigation interfaces implement the so-called adaptive hypermedia navigation. Many adaptive hypermedia conception techniques are proposed [2],[3]. These modeling techniques are based on data modeling, navigation structure modeling and abstract interface modeling.

The idea of adaptive hypermedia navigation consists to couple together the domain model and the user model using an adaptivity engine [4]. The domain model represents concepts and their associations of the application domain. The user model represents user objectives, preferences, goals and backgrounds. Adaptation is made by the adaptivity engine using rules [5]. These rules have the form: If event Then action.

Some adaptation rules are executed before presenting the document and others are executed after presenting the document. The rules of the first class are useful to select the document fragments to be shown or hidden and to define the order of these fragments and their visualization specificities. The rules of the second class are mainly used to update the user profile according to this action.

According to [5], hypermedia adaptation is performed at two levels: content adaptation and navigation adaptation. The content adaptation consists to include only pertinent fragments in the presented document, which are organized according to user preferences and goals. In AVANTI system [6], a document fragment can be described using a visual description or acoustic description. For visually handicapped persons, the fragment is presented using acoustic description. As in ELM-ART II [7], the navigation adaptation consist to reorder, annotate, colour according to pertinence of referenced document, desactivated or removed if the referenced document is not pertinent, the hyperlinks of the presented document.

In [16], adaptation is made into two steps: query preview and query refinement. In the first step, the interface presents only some attributes, which are used by the user to specify the so-called *query preview* in order to select only parts on universe of data containing pertinent information. In the second step, the interface presents more attributes, which are used to formulate a *refinement query* evaluated over the data resulted from the first step.

3 Background

3.1 An overview of Topic Maps

Topic Maps [9] is a paradigm used to 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 resources access. As semantic networks, Topic Maps builds a structured semantic link network on these resources [10].

A topic map[1] is built with topics in a networked form. A topic can be anything that is a subject 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, etc.

Topics can be linked together by associations expressing some given semantics. Topic Maps applications define the nature of the associations. Thus, semantic information or knowledge is specified by the association and especially by topic roles. In Topic Maps, associations can be defined regardless of occurrences. Associations are used to express knowledge between topics and not between occurrences. Topics and associations represent the abstract part of a topic map. The concrete part is represented by occurrences. Occurrences are resources linked to topics. A topic occurrence can be any information that is specified as relevant to a given topic. An example of a topic map represented in XTM [14] (XML for Topic Maps) is shown in Fig. 1.

In the topic map of the Figure 1, Peter and Johanna are represented by topics peter-id and johanna-id, which are associated by the association reified by the topic examination. Each topic of an association plays a role. The topic peter-id is the doctor in the association examination and johanna-id is the patient in the same association. We note that, generally associations express most important knowledge.

In Topic Maps, anything is a topic. Concepts, objects, associations and roles are represented by topics. In the previous example, individuals Peter and Johanna are represented by topics, the association is an instance of a topic, roles are topics

[1] topic map (t, m in tiny) references a knowledge base structured with respect to the Topic Maps formalism (T, M in capital letters)

referenced using PSI (Public Subject Indicator). PSI is a URI to a public topic maintained (as an ontology) apart from the topic map.

```
<topicMap  xmlns:xlink="http://www.w3.org/1999/xlink">
   <topic id="peter-id">                    /* topics of the topic map */
      <instanceOf><subjectIndicatorRef
xlink:href="http://www.xx.com/onto.daml#human"/></instanceOf>
      <topname><basename>Dr. Peter</basename></topname>
   </topic>
   <topic id="Johanna-id">
      <instanceOf><subjectIndicatorRef
xlink:href="http://www.xx.com/onto.daml#woman"/></instanceOf>
      <topname><basename>Ms. Johanna</basename></topname>
   </topic>
   <topic id="exam-id"> <topname><basename>examination</basename></topname> </topic>
   <association>                            /* topic associations of the topic map */
      <instanceOf> <topicRef="# exam -id"/></instanceOf>
      <member>
          <topicRef="#peter-id"/>
          <roleSpec><subjectIndicatorRef
xlink:href="http://www.xx.com/onto.daml#doctor"/></roleSpec>
      </member>
      <member>
          <topicRef="#Johanna-id"/>
          <roleSpec><subjectIndicatorRef
xlink:href="http://www.xx.com/onto.daml#patient"/></roleSpec>
      </member>
   </association>
</topicMap>
```

Fig. 1. An example of a topic map representing part of a knowledge base in medicine

3.2 An overview of Description Logics

DLs [12] are logics developed to represent complex hierarchical structures and make reasoning facilities over these structures. DLs are used to build ontologies for semantic Web [13]. 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 over individuals. Abstract knowledge is expressed using concepts and roles. Concepts are unary predicates, which represent an abstraction of individuals. Roles are binary predicates. They represent relations between concepts.

The abstract knowledge is expressed over concepts and roles by using constructors. Syntax and semantic of some DLs constructors are given in the table 1. Constructors semantic is given by means of an interpretation $I= (\Delta^I, .^I)$, where,

- Δ^I is a nonempty set which represents the individuals of the concrete knowledge
- $.^I$ is an interpretation function defined as:
  - $.^I (C) = C^I \subseteq \Delta^I$ for each concept C
  - $.^I (R) = R^I \subseteq \Delta^I \times \Delta^I$ for each role R

We consider that C is a concept description and R is a role description. Thus, the syntax and semantic of the DL constructors are given in [12].

DL knowledge is expressed in a ABox as set of descriptions. A description is defined using the previous constructors. The following ABox knowledge base represents the knowledge expressed using Topic Maps in Figure 1.

Table 1. An example of a TBox representing part of a medical knowledge base

doctor $\doteq$ human $\sqcap$ $\forall$ name.String $\sqcap$ $\forall$ specialty.String $\sqcap$ $\forall$ exam.patient	
patient $\doteq$ human $\sqcap$ $\forall$ name.String $\sqcap$ $\forall$ natioinality.String $\sqcap$ $\forall$ exam^{-1}.doctor $\sqcap \geq 1$ exam^{-1} $\sqcap$ $\forall$ child.Human	
patient_having_child $\doteq$ patient $\sqcap$ ≥ 1 child	
specialist $\doteq$ doctor $\sqcap$ ≥ 1 specialty	
peter : doctor	johanna : patient
(peter,'cardiologist'):specialty	(peter,johanna):exam

We have,

Δ^I = {peter, johanna},

Doctor^I = {peter}, Patient^I = {johanna}

exam^I = {(peter,johanna)}

specialty^I = {(peter,'cardiologist')}

DLs provide powerful reasoning facilities on conceptual part (TBox) and others on individuals (ABox). The significant ones are:

- *Subsumption* ($\sqsubseteq$): based on concept descriptions, this reasoning allows to explicit subsumption between two concepts. From the example of table 1, the following subsumptions are automatically computed:

  specialist $\sqsubseteq$ doctor, patient_having_child $\sqsubseteq$ patient
- *Realisation*: it consists to affect an individual of the ABox to the most appropriate concept in the TBox. Using the assertion (peter,'cardiologist'):specialty, the realization reasoning infers that peter is a specialist, that is

  peter:specialist

4 Data integration using Topic Maps and Description Logics

Data representation and integration are important and fundamental tasks to access data in homogeneous and coherent manner. On the Web, data are represented using semistructured HTML or XML models. For HTML documents, data integration consists to link HTML-documents among them using hyperlinks. This is a static and rigid approach. To explore data, Web-documents are navigated by following structural hyperlinks. This is not efficient because hyperlinks are structural and do not consider semantic relationships among documents. XML-based data integration is realized either by a query language for querying multiple XML documents using one single query [18] or by providing a uniform view of multiple XML documents [21]. In order to integrate XML documents, a mechanism to identify multiple instances of a same real object is proposed in [20].

Semistructured data models, OEM [15] and XML, are used in the data integration process [17],[19]. However, using only XML for data integration is not suitable, especially when need data semantics. That is, XML does not give any semantic about tags.

The suitability of description logics for data integration is illustrated in some projects, namely the SIMS [22] and PICSEL [11]. In these systems, datasources are linked together and with the global schema by knowledge expressed manually.

We present an integration process that combines Topic Maps and description logics to make a semantic data integration. First, data sources are represented using Topic Maps in order to track distributed knowledge. However, Topic Maps does not deal with constraints. TMCL [8] is only on specification stage. Secondly, we use description logics in order to track constraints. Constraints are useful in data integration. They are used by description logic reasoning to deduce implicit relations between concepts. The integration process is shown in the following figure:

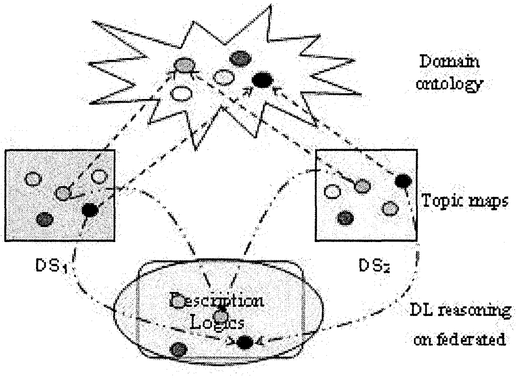

Fig. 2. . Using Topic Maps, description logic and ontology for data integration

The datasources DS_1 and DS_2 are represented using Topic Maps. Two datasource types are considered, namely relational databases and XML documents. For relational databases, topics represent tables and table attributes. For XML documents, topics represent tags. In the topic map of Fig.3, the table (or tag) person and its address attribute (sub-tag) are both represented by topics. These topics are connected to the ontological concepts that they represent using the subjectIndicatorRef tag.

So, semantic integration based on Topic Maps is natural. It consists to merge topics referencing the same ontological concept into one topic in the federated topic map.

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

Fig. 3. Representing tables/tags and attributes/sub-tags by topics

Then, the federated topic map is used by description logic reasoning in order to provide a consistent federated topic map by computing automatically implicit relations between concepts. Consider that the datasources DS_1 and DS_2 contains the two descriptions:

$\mathtt{DS_1 : person} \doteq \forall\ \mathtt{name.String}\ \Pi\ \forall\ \mathtt{address.String}\ \Pi\ \ldots$

$\mathtt{DS_2 : patient} \doteq \mathtt{human}\ \Pi\ \forall\ \mathtt{disease.String}\ \Pi\ \geq 1\ \mathtt{disease}$

As traditional logics, description logics are not able to make reasoning on distributed knowledge bases. Without consideration of Topic Maps, especially the subjectIdentity of topics, DL reasoning does not infer any relation between person and patient. That is, person Π patient $\sqsubseteq \varnothing$. Semantically, this is not correct.

Consider now that person is human through the subjectIdentity tag of the topic map. That is,

$\mathtt{DS_1 : human} \doteq \forall\ \mathtt{name.String}\ \sqcap\ \forall\ \mathtt{address.String}\ \sqcap \ldots$

$\mathtt{DS_2 : patient} \doteq \mathtt{human}\ \sqcap\ \forall\ \mathtt{disease.String}\ \sqcap \geq 1\ \mathtt{disease}\ \sqcap \ldots$

Therefore, DL reasoning infers that patient $\sqsubseteq$ person, which makes semantic connection between the datasources DS_1 and DS_2.

5 Subject-oriented navigation approach

The navigation interfaces on the Web become increasingly intelligent because they automatically adapt the visualized information for each user. Whereas fundamental element of these interfaces is the user profile [5], we present in this section a new approach called subject-oriented navigation, which adapts navigation according to user search.

Subject-oriented navigation. A subject-oriented navigation is based on the notion of subject. A subject is a user-defined object on which a search session is centered. That is, all needed information is related to the subject. Thus, the interface presents only relevant information to the user-defined subject. In the medical domain, patient-oriented navigation allows to get information about a patient, a patient record. In e-business, user would navigate to get all information about a product.

Conceptual schema exactly cardinalized. The federated topic map is visualized as concept card (see fig. 4) that represents at the same time the data and the conceptual schema. So, the concept card is a conceptual schema which represents the real data. Exact cardinalities are showed for the associations. A concept exists in the concept card only when it has relevant instances to the user. Based on this purpose, user query determines the structure and the content of the visualized concept card.

Adaptive interface. The interface that implements the subject-oriented navigation is adaptive, dynamic and progressive. The interface does not present the entire concept card at once because it is not efficient to search information in a huge volume of data. It progressively presents the concept card with the user navigation in order to adapt the content and the structure of the visualized concept card to the user needs. Indeed, as explained before, the conceptual graph visualized represents data. When the data is selected by user, the visualized concept card adapts its structure consequently. Considering the example of the Figure 4, the user selects the patient Paul in the concept card 4(left). So, the interface reconstructs a concept card centred on the patient Paul and the new cardinalities are recalculated. Paul is examined only once by a physician (Medecin in Figure 4), so the cardinality is adjusted. In the federative topic map, there are no results for this patient. Thus, the concept Result (connected to the topic Medecin) is removed from the concept card 4(right).

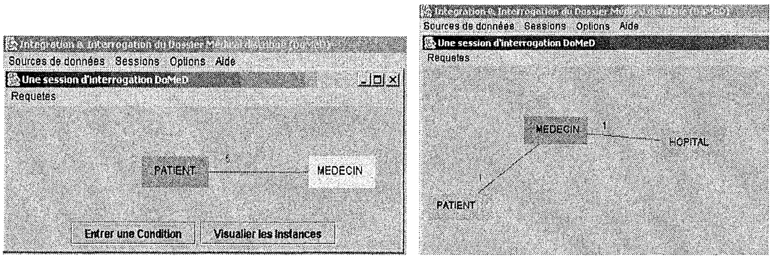
Sources de données Sessions Options Aide
Une session d'interrogation DoMeD
Requetes
PATIENT
5
MEDECIN
Entrer une Condition
Visualier les Instances
Sources de données Sessions Options Aide
Une session d'interrogation DoMeD
Requetes
MEDECIN
1
HOPITAL
1
PATIENT

Fig. 4. The left screen gives an example of a conceptual card and the right one shows a concept card obtained after user restriction and navigation to the topic Medecin in the left screen

6 Conclusion

Compared to the traditional web navigation (see Figure 5), the subject-oriented navigation is more adaptive and presents only pertinent data and navigation paths, at each navigation step. Consider that the user wants to get prescriptions of the patient John examined by the doctor Peter. He follows the path Patient, Exam, Doctor and Prescription.

At each step of the user navigation, the interface presents only pertinent data according to the user needs. It presents John's exams, doctors who examined John, and finally John's prescriptions made by Dr. Peter.

If we simulate the same navigation path using traditional approach, useless information could be presented in the html documents presented to the user at each step of the web navigation. That is, at the third step of the navigation, the doctor.html document can not present only the doctors who examined John but presents, therefore, all the doctors stored in the Web pages. At the final step, the prescriptions.html document contains all the prescriptions made by the doctor Peter for all his patients.

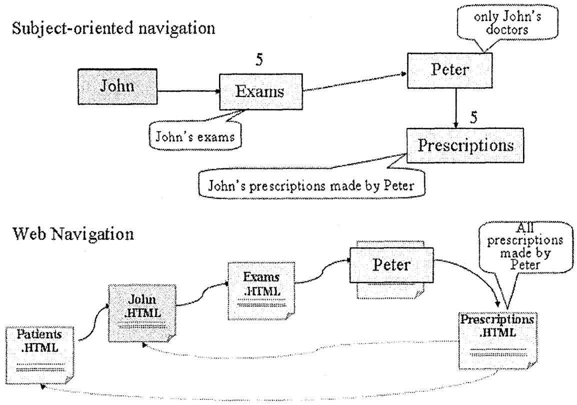

Fig. 5. Our subject navigation approach compared to the traditional navigation on the Web

References

1. Florescu, D., Levy, A., Mendelzon, A.: Database Techniques for the World-Wide Web: A Survey. *Sigmod Records*, vol. 27, n° 3, (1998) 59-74
2. Schwabe, D., Rossi, G., Barbosa, S.D.J.: Systematic Hypermedia Application Design with OOHDM. Proc. of the 7th ACM Conference on HyperText, 1996, p. 116-129
3. Garzotto, F., Paolini, P., Schwabe, D.: HDM - a model-based approach to hypermedia application design. *ACM Transactions on Information Systems* 11 (1993) 1-26
4. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. *In User Modeling and User Adapted Interaction*, vol. 6, n° 2-3, (1996) 87-129
5. De Bra, P., Brusilovsky, P., Houben, G.: Adaptive hypermedia: From Systems to Framework. *ACM Computing Surveys*, (1999) vol. 31, n° 4
6. Stephanidis, C., Paramythis, A., Sfyrakis, M., Stergiou, A., Maou, N., Leventis, A., Paparoulis, G., Karagianidis, C.: Adaptable and Adaptive User Interfaces for Disabled Users in AVANTI Project. *In 5th International Conference on Intelligence in Services and Networks (IS&N '98), Technology for Ubiquitous Telecom Services, Antwerp, Belgium,* p. 153-166
7. Weber, G., Specht, M.: User modeling and adaptive navigation support in WWW-based tutoring systems. *In proceedings of the 6th Int. Conf. on User Modeling, UM97*, Vienna, new-York, (1997) 289-300
8. Moore, G., Nishikawa, M.: The Topic Map Constraint Language. ISO/IEC 13250 (2003). Available at http://www.isotopicmaps.org/tmc
9. Sigel, A., (2000). Towards knowledge organization with Topic Maps. *XML Europe 2000, Palais des congrès Paris*, 12-16 June 2000
10. Fresse, E., (2000). Using Topic Maps for the representation, management and discovery of knowledge. *XML Europe 2000, Palais des congrès Paris*, 12-16 June 2000
11. Goasdoué, F., Lattes, V., Rousset, M.C.: The Use of CARIN Language and Algorithms for Information Integration: The PICSEL Project. International *Journal of Cooperative Information Systems (IJCIS),* vol. 9, n° 4, (2000) 383-401
12. Borgida, A.: Description Logics in data management. *IEEE Trans. on Knowledge and Data Engineering*, vol. 7, n° 5, (1995) 671-682

13.Horrocks, I., Patel-Schneider, P.F., van Harmelen, F.: Reviewing the Design of DAML+OIL: An Ontology Language for the Semantic Web. *Proc. of 18th National Conference on Artificial Intelligence, AAAI-02,* (2002) 792-797

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

15.Goldman, R. , Chawathe, S., Crespo, A., McHugh, J. A Standard Textual Interchange Format for the Object Exchange Model (OEM). *Department of Computer Science, Stanford University, California, USA, (1996) 5 p.*

16.Doan, K., Plaisant, C., Shneiderman, B. and Bruns, T.: Interface and Data Architecture for Query Preview *in Networked Information Systems. ACM Transactions on Information Systems*, 1999, vol. 17, n° 3, p. 320-341

17.Papakonstantinou, Y., Garcia-Moulina, H., Widom, J. Object Exchange Across Heterogeneous Information Sources. *Proceedings of IEEE International Conference on Data Engineering*, (1995) 251-260

18.Cohen, S., Mamou, J., Kanza, Y., Sagiv, Y.: XSEarch: A Semantic Search Engine for XML. *VLDB (2003)* 45-56

19.Gardarin, G., Mensch, A., Tomasic, A.: An Introduction to the e-XML Data Integration Suite. *Proceedings of EDBT (2002)* 297-306

20.de Brum Saccol, D., Heuser, C.A.: Integration of XML Data. *Proceedings of EEXTT (2002)* 68-80

21.Camillo, S.D., Heuser, C.A., Mello, R.S.: Querying Heterogeneous XML Sources through a Conceptual Schema. *Proceedings of ER (2003)* 186-199

22.Arens, Y., Chee, C., Hsu, C., Knoblock, C.: Retrieving and Integrating Data from Multiple Information Sources. *In Journal of Intelligent and Cooperative Information Systems,* vol. 2, n°2, (1993) 127-158