Doktorandský den ’06

Ústav informatiky
Akademie věd České republiky

Monínec, Sedlec–Prčice

20.– 22. září 2006

vydavatelství Matematicko-fyzikální fakulty
University Karlovy v Praze

Publikaci ”Doktorandský den ’06” sestavil a připravil k tisku František Hakl
Ústav Informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8

ISBN *– not yet –*

Obsah

Ontology-based Integration System

Post-Graduate Student:
ING. ZDEŇKA LINKOVÁ

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2
182 07 Prague 8

Department of mathematics
Faculty of nuclear science and physical engineering
ČVUT
Trojanova 13
120 00 Prague 2

Czech Republic

linkova@cs.cas.cz

Supervisor:
ING. JÚLIUS ŠTULLER, CSC.

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2

182 07 Praha 8

stuller@cs.cas.cz

Field of Study:
Mathematical Engineering

This work was supported by the project 1ET100300419 of the Program Information Society (of the Thematic Program II of the National Research Program of the Czech Republic) “Intelligent Models, Algorithms, Methods and Tools for the Semantic Web Realization” and by the Institutional Research Plan AV0Z10300504 “Computer Science for the Information Society: Models, Algorithms, Application”.

Abstract

Integration has been an acknowledged problem for a long time. With the aim at combining data from different sources, data integration usually provides a unified global view over these data. A crucial part of the task is the establishment of the connection between the global view and the local sources. Two basic approaches have been proposed for this purpose: Global As View (GAV) and Local As View (LAV). With the Semantic Web and its data description means, there is also another possibility - to employ ontologies for the relationship description in an integration system.

1. Introduction

Today’s world is a world of information. The expansion of World Wide Web has brought a number of information sources. However, at the same time, a number of different formats, data heterogeneity, and not yet efficient machine processing of web sources cause many problems. One of them is the reappeared problem of data integration.

Data integration is the task of combining data residing at different sources and enabling the user to process these data as one whole. Data integration has been an acknowledged data processing problem for a long time. Although there have been some projects on integration of data within particular areas, there is no universal tool for general data integration.

In general, data integration can be pursued in different layers. It is possible to consider only data, or consider also metadata (e.g. schemas). With greater data amount, the integration approach is rather non-materialized than materialized. The integration result brings virtual view over data sources that do not store any data. Therefor, the establishment of the connection to original data sources is crucial. To consider the data schemas is essential. There are some basic approaches to the design a non-materialized integration system, each with its advantages and disadvantages. The proposed approach brings an idea from the Semantic Web - a semantic extension of the current World Wide Web.

The paper is organized as follows. Section 2 describes the data integration task and basic approaches. Section 3 introduces the vision of the Semantic Web and one of its principle layers - ontologies. Section 4 presents an ontological approach to data integration. Finally, Section 5 summarizes the paper.

2. Data Integration

In data integration, the goal is to synthesize data from different data sources into one integrated data source. A user willing to process the data uses the integrated source and is freed from the knowledge where the data are and how the data are structured in the respective sources.

The integrated data source can be materialized, i. e. a new data source is created and it physically stores data, or it can be virtual, i. e. a virtual view is defined and the data remain in the sources. In materialized data integration approach, a copy of the data is made. So, with respect to actualization requirements, it is suitable for more or less stable data. Virtual data integration approach provide an interface to autonomous data sources, it can be used also for large amount of data with relatively frequently changing content. In a connection with the World Wide Web data, this approach suggests itself. It is also possible to combine both approaches. An example is an integration system that provides a virtual integrated view, but it also materializes some data in a cache. The cache is usually used for frequently accessed data.

A commonly used system architecture in virtual view approach [1] to data integration is depicted in Figure 1. A base of the system is a set of data sources to integrate. The higher layer is a set of components called wrappers. Each wrapper belongs to one local data source and it plays a role of a connector between the local source background (it means a specific model, a specific language etc. for the source) and the global one. The pure integration part of the system is presented in hierarchical layers of mediator components. A mediator can obtain information from components below it and can provide information to components above it. In general, an integration system can contain an arbitrary complex architecture.

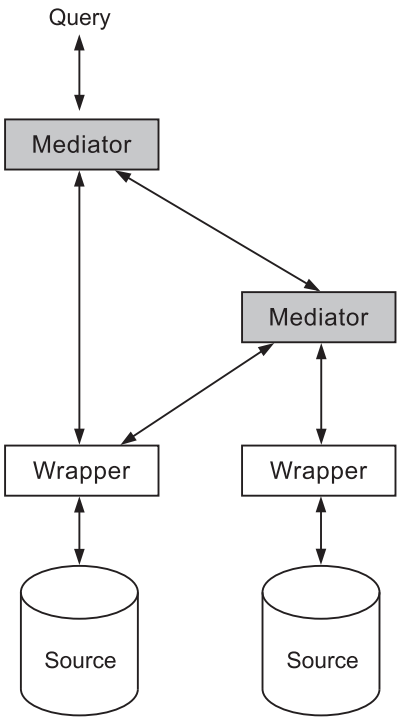

Figure 1: A mediation integration architecture

Each mediator in a hierarchy can be seen as a virtual view. These views are then used in query evaluation. A user of the integration system poses his query to a global view using a global schema. Using mediation integration, the query is reformulated and decomposed to refer to the data sources and the queries are also executed over the sources. Then obtained information is composed and the answer is given back to the user.

The main components of a data integration system are the sources with their local schemas, the global virtual view with the global schema, and the mediated system that expresses the correspondence between the global source and the local sources. So it follows that a data integration system I is a triple

$$I = (G, L, M),$$

where G is a global schema, L is a set of local schemas and M is a mediation system.

A possible way how to describe the mediation system is to use mappings. Mapping is a set of assertions that establish the connection between the element of the global schema and those of the local schema. The composition of mapping is an essential task. It plays a crucial role in query resolving - another important process of a data integration system. Two basic approaches [2, 3] have been used in order to specify the mapping. The *Global As View* (GAV) approach consists in defining the global schema as a set of views over the local schemas, while the *Local As View* (LAV) approach consists in defining the local sources as a set of views made on the global schema.

Because the GAV is based on the idea that the content of each element of the global schema should be characterized in terms of a view over the sources, this mapping tells the system how to retrieve the data. The GAV favors the system in carrying out query processing - it gives direct information on how query answering may be performed. Some GAV data integration systems do not allow integrity constraints in the global schema. Under these assumptions, query processing can be based on a simple unfolding strategy: every element of the global schema is substituted with the corresponding query over the sources. When global schema allows integrity constraints, the query processing in GAV becomes more complex - integrity constraints here can in principle be used in order to overcome incompleteness of data in the sources. In GAV query processing can look easy. However, this idea is effective when a set of sources is stable. The addition of a new source and extending the system can be difficult. The new source may have an impact on the definition of various elements of the global schema. So it can force the system designer to redesign the schema, and so to reconsider all the sources.

The LAV approach is based on the idea that the content of each source can be characterized in terms of a view over the global schema. Processing queries in LAV is a difficult task. The only knowledge we have about the data in the global schema is through the views representing the sources, and such views provide only partial information about the data. The mapping specifies the role of each source relation with respect to the global schema. It is not immediate to infer how to use the sources in order to answer queries. The LAV favors the system in the extensibility - addition of sources simply means enriching the mapping with definition of a new view over the global schema, without other changes.

To compensate the insufficiency of the LAV and GAV approaches, also their combinations have been proposed. The *Global Local As View* (GLAV) approach [4] establishes the relationships between the global schema and the sources by making both of LAV and GAV mappings and allows flexible schema definitions independent of the particular details of the sources.

3. The Semantic Web

The Semantic Web [5, 6] is intended as an extension of today's World Wide Web. It should consist of machine readable and efficiently processable data. The basis is addition of data semantics - data description will be stored together with data themselves. The full realization of the Semantic Web belongs still to the future; however, many tools, languages, theories etc. have been developed and several also implemented. The Semantic Web is based on several standards, which are defined by W3C (WWW Consortium) [7].

An important requirement for effectively machine processable data is data structuring. On the web, the main structuring method is using tags, which are parts of text containing information about the role of the text. Nowadays, the metalanguage XML (eXtensible Markup Language) [8] is used for making web document structure. It provides syntax for machine readable data. But only XML is not enough to describe data. The technique to specify the meaning of information is RDF (Resource Description Framework) [9]. It is a

basic tool of web sources metadata addition. RDF data model gives an abstract conceptual framework for metadata definition and usage. It uses XML syntax (RDF/XML) for encoding. Additionally, there is also an extension of RDF called RDF Schema [10] that is useful for class definition and class hierarchy description.

An instrument for definition of terms used either in data or in metadata are ontologies [11]. The term ontology has been used in many ways and across different communities. A popular definition of the term ontology in computer science is: an ontology is a formal, explicit specification of a conceptualization. A conceptualization refers to an abstract model of some phenomenon in the world. However, a conceptualization is never universally valid. Ontologies have been set out to overcome the problem of implicit and hidden knowledge by making the conceptualization explicit. Ontologies aim at modeling and structuring domain knowledge. It may take a variety of forms, but it will necessarily include a vocabulary of terms and some specification of their meaning. In the context of web technologies, ontology is a file or a document that contains formal definitions of terms and term relations. The Semantic Web technique for definition of ontologies is the OWL (Ontology Web Language) language [12].

4. Ontology-based mediation integration

In an ontology-based integration approach described in this paper, a conception of a virtual integration form is adopted. A global source will be also non-materialized and for the establishment of a connection to the data sources some kind of mapping will be applied. However, instead of using mapping rules as assertions for global and local schemas elements, a more complex structure covering all mapping will be employed. This approach exploits the idea that on the Semantic Web every peace of information has got defined its meaning and supposes availability of ontologies as a means for defining the concept of the data. The integration task is transformed to the building of an ontology for the integration system. This ontology from its principle should cover ontologies of all data used in the system and mappings that are in general seen as definitions of relationships between data.

Suppose, there are two data sources S_1 and S_2. Each source schema is described by an ontology: an ontology referring to the local source S_1 is O_{S_1}, an ontology of the source S_2 is O_{S_2}. The global integrated view the integration system should provide has an ontology O_G. The integration system, in Section 2 formalized as a triple $I = (G, L, M)$, has in this case representation

$$I = (O_G, \{O_{S_1}, O_{S_2}\}, O_I),$$

where O_I is an ontology of the integration system.

Ontology O_I is used to describe the mapping between elements of the global view and the local sources. O_I is also an ontology of all concepts used in the integration system I. So it follows that for ontologies of local sources is valid:

$$O_{S_1} \subseteq O_I$$
$$O_{S_2} \subseteq O_I$$

While ontologies O_{S_1} and O_{S_2} are given with the sources, O_G and O_I need to be determined. Description of O_G is relatively independent on the sources. O_G contains definition of concepts accessible directly via the global view. It is a matter of a designer who decides what will be accessible via the integration system and in what form.

Establishment of O_I is a crucial step. However, it is not an easy task. Covering O_{S_2}, O_{S_2}, O_G and their relationships, O_I is the result of task called merging ontologies. Ontology merging is studied e.g. in [13] and [14]. As in schema integration in other approaches, some conflicts [15] that have to be solved can arise. Conflicts can be of various types [16], for example terms conflicts, schema discrepancies, raw data and metadata conflicts etc. Regarding used terms, synonym and homonyms conflicts can arise. (Synonyms are different words with similar or identical meanings. Homonyms are words that are spelled the same but

with a different meaning.) In the ontology world, it is not difficult to deal with synonyms or homonyms, because there are means how to express relationship between terms. In an ontology, each term has a unique reference. Although, there can be two terms in two ontologies named in the same way, they are uniquely distinguishable, because of the context - the ontology where they are defined. This is for instance in XML syntax solved by namespaces. Within ontologies it is also possible to state that two terms are equivalent and describe the synonymic relationship, and by this to enable process it in a right way.

Example 1 There are two sources to integrate. Source 1 stores satellite images taken from a satellite Ikonos and its ontology O_1 describes only one class named `Ikonos_images` with properties `date_`,`the_geom` etc. Source 2 stores satellite images from a satellite Spot, its ontology O_2 contain class `Spot_images` with properties `date_acqui`, `the_geom` etc. Since the integration system should provide satellite images coming from different data sources, global ontology O_G contains class named `satellite_images` with properties `date` (date when the image was taken), `geom` (a geometry of the photographed region), etc. To obtain ontology of the system, O_1, O_2, O_G, and knowledge about relationships among particular concepts are merged. Ontology O_I is following:

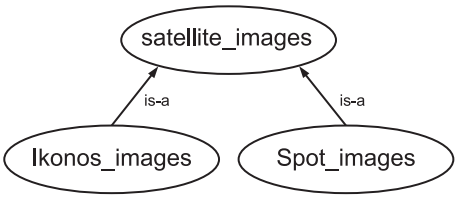

Figure 2: Ontology O_I

Ontology O_I contains three classes: `satellite_images`, `Ikonos_images`, and `Spot_images`. Images from Ikonos and images from Spot are both satellite images, so there is hierarchical class - subclass relationship between `satellite_images` and `Ikonos_images` and between `satellite_images` and `Spot_images`. `Ikonos_images` and `Spot_images` cannot be merged into one class, because it refers to different satellite images. With the knowledge of class properties semantics, there can be seen property - subproperty relationship between a global property and a relevant local property, for example `date` and `date_`. Moreover, if there were the same integrity constraints on each property from the pair, the properties can be merged and connected as equivalent. □

With a data integration system, a user poses his query on the global view in terms of the global view. In order to execute the query over the sources, where data are stored, query processing is needed. There are two approaches to query processing. The first one is query rewriting - a query is decomposed to parts referring to local sources and reformulated to be expressed in local source background. The other one is query answering - it do not pose any limitations on how a query is processed, the only goal is to exploit all possible information to compute the answer, for example find the set of data such that the knowledge logically implies that it is an answer to the query.

With mapping expressed in an ontology, for query rewriting, it is possible to adopt a rule well known in object-oriented world: a child can substitute his parent. If we are looking for all instances of class C that have property $P = x$, the query is

$$q := C(P = x).$$

Using ontology O_I, is-a hierarchy relationships give a means how to rewrite the query with respect to a specific local source. If C is not a concept of the local source schema, class C in the query is replaced with its nearest subclass C' in the is-a hierarchy. This is recursively repeated until a concept is founded in the specific local source schema, or there are no more subclasses - there is no answer. The same rule as for classes can be adopted also for properties, and the relationship property-subproperty can be employed.

In query answering approach to query processing, the is-a hierarchy is also essential. It expresses that an instance of a node is an instance of all nodes within the path from the root node to this node. Based on this rule, it can be determined if information from a local source can be an answer to the global query.

Example 2 Continuing the simple example of satellite images integration, this example shows query processing. The global view provides satellite images. The query: give all available images taken on 1st January 2001, i.e.

$$q := satellite\_images(date =' 01-01-2001'),$$

is processed as follows: `satellite_images` is not in the concept of any local source, the query is rewritten. `satellite_images` class has two child nodes `Ikonos_images` related to the source 1 and `Spot_images` related to the source 2. The reformulated query has two forms:

$$q_1' := Ikonos\_images(date =' 01-01-2001')$$

and

$$q_2' := Spot\_images(date =' 01-01-2001').$$

Because property date is not in the concept of the source 1, the query q_1 is further rewritten using property-subproperty to

$$q_1'' := Ikonos\_images(date\_ =' 01-01-2001').$$

The query q_1'' is executed over the source 1. Analogously, the query q_2' is rewritten and executed over the source 2. □

Compared with two basic approaches of mapping specification in a mediation data integration, an ontology-based approach is similar to LAV integration in a way, that the global schema is specified independently from the sources. Another similarity can be found in extending the system. When a new source is added, the ontology of the integration system O_I is enriched with a new source ontology and further possible relationships to previous version of O_I. A difference between LAV and GAV and the ontology-based integration system is in the case of a change in the layer of local or global source schemas. In case of using ontologies, the ontology of integration system is enriched with the new state. It is not needed to change any earlier part of the ontology, or even to remove some part. No other change is needed.

5. Conclusion

Data integration is a task of combining data from different data sources and enabling a user to process them as one whole. There are two classical ways of designing an integration system providing a global virtual view over the sources: GAV and LAV approaches. Both are based on definition of connection between the global view and the local sources via mappings. However, with a Semantic Web idea, there are also other possibilities that can be used. An integration system described in this paper is based on ontologies of the sources. An ontology of the integration system is defined, and it is consequently used for data query processing.

References

[1] J. D. Ullman, "Information integration using logical views", *Theoretical Computer Science* 239 (2000), pp. 189-210.

[2] M. Lenzerini, "Data Integration: A Theoretical Perspective", In Proceedings of the *21st ACM SIGMOD - SIGACT - SIGART symposium on Principles of database systems*, ACM Press, pp. 233-246, 2002.

[3] A. Cali, D. Calvanese, G. De Giacomo, and M. Lenzerini, "On the Expressive Power of Data Integration Systems", In Proceedings of the *21st Int. Conf. On Conceptual Modeling (ER 2002)*, LNCS 2503, Springer, pp. 338-350, 2002.

[4] M. Friedman, A. Levy, and T. Millstein, "Navigational plans for data integration", In Proceedings of the *16th Nat. Conf. On Artificial Intelligence (AAAI'99)*, AAAI Press, pp. 67-73, 1999.

[5] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", *Scientific American*, vol. 284, 5, pp. 35-43, 2001.

[6] M.-R. Koivunen and E. Miller, "W3C Semantic Web Activity", in the proceedings of the *Semantic Web Kick/off Seminar*, Finland, 2001.

[7] W3C (WWW Consortium), `http://www.w3.org`.

[8] Extensible Markup Language (XML), `http://www.w3.org/XML/`.

[9] Resource Description Framework (RDF), `http://www.w3.org/RDF/`.

[10] RDF Vocabulary Description Language 1.0: RDF Schema, *W3C Recommendation*, `http://www.w3.org/TR/2004/REC-rdf-schema-20040210`, February, 2004.

[11] Y. Ding, D. Fensel, M. Klein, and B. Omelayenko, "The semantic web: yet another hip?", *Data & Knowledge Engineering* 41 (2002), pp. 205-227.

[12] Web Ontology Language (OWL), `http://www.w3.org/2004/OWL`.

[13] K. Kotis, G. A. Vouros, and K. Stergiou, "Towards automatic merging of domain ontologies: The HCONE-merge approach", *Web Semantics: Science, Services and Agents on the World Wide Web* 4 (2006), pp. 60-79.

[14] H. T. Uitermark, P. J. M. van Oosterom, N. J. I. Mars, and M. Molenaar, "Ontology-based integration of topographic data sets", *International Journal of Applied Earth Observation and Geoinformation* 7 (2005), pp. 97-106.

[15] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano, "Semantic integration of heterogeneous information sources", *Data & Knowledge Engineering* 36 (2001), pp. 189-210.

[16] C.-Y. Lee and V.-W. Soo, "The conflict detection and resolution in knowledge merging for image annotation", *Information Processing and Management* 42 (2006), pp. 1030-1055.

Ústav Informatiky AV ČR
DOKTORANDSKÝ DEN '06

Vydal
MATFYZPRESS
vydavatelství
Matematicko-fyzikální fakulty
University Karlovy
Sokolovská 83, 186 75 Praha 8
jako svou *– not yet –* . publikaci

Obálku navrhl František Hakl

Z předloh připravených v systému LaTeX
vytisklo Reprostředisko MFF UK
Sokolovská 83, 186 75 Praha 8

Vydání první

Praha 2006

ISBN *– not yet –*