THE MOMIS METHODOLOGY FOR INTEGRATING HETEROGENEOUS DATA SOURCES*

Domenico Beneventano and Sonia Bergamaschi
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Modena e Reggio Emilia
Via Vignolese 905 – 41100 - Modena

Abstract: The Mediator EnvirOnment for Multiple Information Sources (MOMIS) aims at constructing synthesized, integrated descriptions of the information coming from multiple heterogeneous sources, in order to provide the user with a global virtual view of the sources independent from their location and the level of heterogeneity of their data. Such a global virtual view is a conceptualization of the underlying domain and then may be thought of as an ontology describing the involved sources. In this article we explore the framework's main elements and discuss how the output of the integration process can be exploited to create a conceptualization of the underlying domain.

Keyword: Ontologies, Heterogeneous Sources, Mediator, Global As View, WordNet

1. INTRODUCTION

Nowadays the Web is a huge collection of data and its expansion rate is very high. Web users need new ways to exploit all this available information and possibilities. A *new vision of the Web*, the Semantic Web, where resources are annotated with machine-processable metadata providing them with background knowledge and meaning, arises. A fundamental component of the Semantic Web is the ontology; this "*explicit specification of a conceptualization*"[1] allows information providers to give a understandable meaning to their documents.

MOMIS[2] is a framework for information extraction and integration of heterogeneous information sources, developed by the DBGroup (www.dbgroup.unimo.it) at the University of Modena and Reggio Emilia

* This research has been partially supported by EU IST-SEWASIE

(c) 2004 IFIP

(UNIMORE). The system implements a semi-automatic methodology for data integration that follows the *Global as View* (GAV) approach[3]. The result of the integration process is a global schema, which provides a reconciled, integrated and virtual view of the underlying sources, GVV (Global Virtual View). The GVV is composed of a set of (global) classes that represent the information contained in the sources, and it is the result of the integration process, i.e. a conceptualization of the underlying domain (domain ontology) for the integrated sources. The GVV is then semi-automatically annotated according to a lexical ontology. With reference to the Semantic Web area, where generally the annotation process consists of providing a web page with semantic markups according to an ontology, we firstly markup the local metadata descriptions and then the MOMIS system generates an annotated conceptualization of the sources. Moreover, our approach "builds" the domain ontology as the synthesis of the integration process, while the usual approach in the Semantic Web is based on "a priori" existence of ontology.

A comparison of MOMIS with others mediator systems is proposed in Table 1 (TSIMMIS[4], GARLIC[5], SIMS[6], Infomaster[7], Information Manifold[8], Observer[9]).

	MOMIS	**TSIMMIS**	**GARLIC**	**SIMS**	**Infomaster**	**IM**	**Observer**
Developer	UNIMORE University	Stanford University	IBM Almaden	ISI-USC	Stanford University	AT&T Research	Saragozza University
Sources	Structured and semistruct.	Semistruct.	Heterog.	Semistruct.	Semistruct.	Web pages	Heterog.
Data Model	ODL*I3*	OEM/MSL	GDL	Loom	KIF/KQML	Extended Relational	-
Description Logic	OLCD[10]	-	-	Loom	-	CARIN	CLASSIC
Approach	GAV	GAV	GAV	GAV/ LAV	GAV	LAV	GAV
Global View Creation	semi-automatic	manual	manual	based on wrapper description	manual	manual	automatic (DL based)
Query	Only for relational sources	Yes	Yes	Yes	Yes	Yes	Yes
Status of the project	Evolving in the Sewasie project	Completed	Evolving in other projects	Evolving in other projects	Completed	Completed	Completed

Table 1. Mediator system comparison.

(c) 2004 IFIP

2. THE MOMIS INTEGRATION METHODOLOGY

In this section, we describe the information integration process for building the GVV. The process is shown in Figure 1.

The ODL$_{I3}$ Language

As a common data model for integrating a given set of local information sources, MOMIS uses an object-oriented language called ODL$_{I3}$, which is an evolution of the OODBMS standard language ODL. ODL$_{I3}$ extends ODL with the following relationships expressing intra- and inter-schema knowledge for the source schemas: SYN (synonym of), BT (broader terms), NT (narrower terms) and RT (related terms). By means of ODL$_{I3}$, only one language is exploited to describe both the sources (the input of the synthesis process) and the GVV (the result of the process). The translation of ODL$_{I3}$ descriptions into one of the Semantic Web standards such as RDF, DAML+OIL, OWL is a straightforward process. In fact, from a general perspective an ODL$_{I3}$ concept corresponds to a *Class* of a the Semantic Web standard, and ODL$_{I3}$ relationships are translated into *properties*.

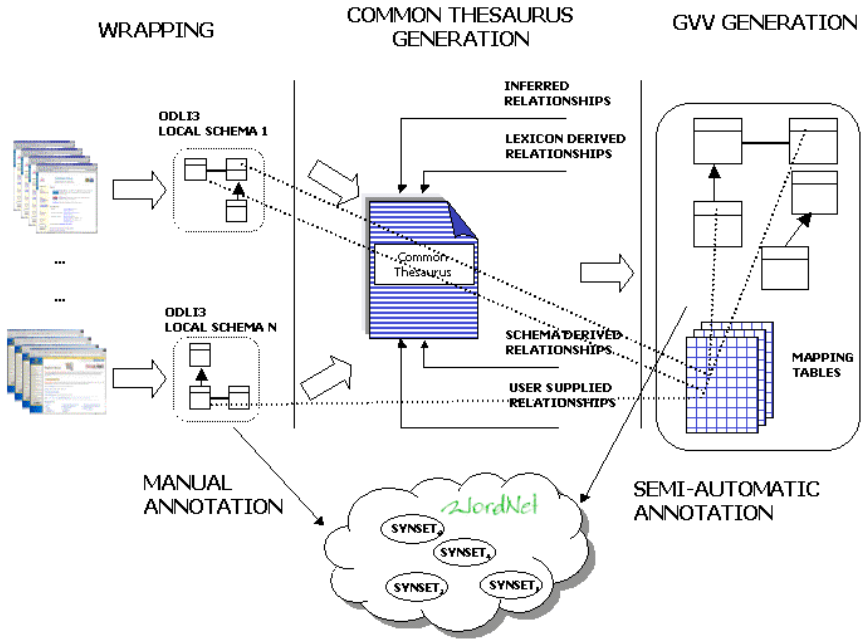

Figure 1. Overview of the ontology-generation process. The figure shows the local schemas' generation, where local schemas are annotated according to the lexical ontology WordNet, the Common Thesaurus generation, and finally the GVV global classes. In particular, these ones are connected by means of mapping tables to the local schemas and are (semi-automatically) annotated according to WordNet.

(c) 2004 IFIP

Wrapping: extracting data structure for sources

A wrapper logically converts the source data structure into the ODL_{I^3} model. The wrapper architecture and interfaces are crucial, because wrappers are the focal point for managing the diversity of data sources. For conventional structured information sources (e.g. relational databases), schema description is always available and can be directly translated. For semistructured information sources, a schema description is in general not directly available at the sources. A basic characteristic of semistructured data is that they are "self-describing" hence information associated with the schema is specified within data. Thus, a wrapper has to implement a methodology to extract and explicitly represent the conceptual schema of a semi-structured source. We developed a wrapper for XML/DTDs files. By using that wrapper, DTD elements are translated into semi-structured objects, according to different proposed methods[11].

Manual annotation of a local source with WordNet

For each element of the local schema, the integration designer has to manually choose the appropriate meaning in the WordNet[12] lexical system. The annotation phase is composed of two different steps: in the *Word Form choice* step, the WordNet morphologic processor aids the designer by suggesting a word form corresponding to the given term; in the *Meaning choice* step the designer can choose to map an element on zero, one or more senses. The annotation assigns a name (this name can be the original one or a word form chosen from the designer), and a set of meanings, to each local class and attribute of the local schema.

Common Thesaurus Generation

MOMIS constructs a Common Thesaurus describing intra and inter-schema knowledge in the form of SYN, BT, NT, and RT relationships. The Common Thesaurus is constructed through an incremental process in which the following relationships are added:

schema-derived relationships: relationships holding at intra-schema level are automatically extracted by analyzing each schema separately. For example, analyzing XML data files, BT/NT relationships are generated from couples IDs/IDREFs and RT relationships from nested elements.

lexicon-derived relationship: we exploit the annotation phase in order to translate relationships holding at the lexical level into relationships to be added to the Common Thesaurus. For example, the hypernymy lexical relation is translated into a BT relationship.

designer-supplied relationships: new relationships can be supplied directly by the designer, to capture specific domain knowledge. If a nonsense or wrong relationship is inserted, the subsequent integration process can produce a wrong global schema;

inferred relationships: Description Logics (DL) techniques of ODB-Tools[13] are exploited to infer new relationships, by means of subsump-

(c) 2004 IFIP

tion computation applied to a "virtual schema" obtained by interpreting BT/NT as subclass relationships and RT as domain attributes.

Global Virtual View (GVV) Generation

The MOMIS methodology allows us to identify similar ODL_{I^3} classes, that is, classes that describe the same or semantically related concept in different sources. To this end, *affinity coefficients* are evaluated for all possible pairs of ODL_{I^3} classes, based on the relationships in the Common Thesaurus properly strengthened. Affinity coefficients determine the degree of matching of two classes based on their names (*Name Affinity* coefficient) and their attributes (*Structural Affinity* coefficient) and are fused into the *Global Affinity* coefficient, calculated by means of the linear combination of the two coefficients[14]. Global affinity coefficients are then used by a hierarchical clustering algorithm, to classify ODL_{I^3} classes according to their degree of affinity.

For each cluster Cl, a Global Class GC, with a set of Global Attributes GA_1, …, GA_N , and a Mapping Table MT, expressing mappings between local and global attributes, are defined. The Mapping Table is a table whose columns represent the local classes, which belong to the Global Class and whose rows represent the global attributes. An element `MT[GA][LC]` is a function which represents how local attributes of LC are mapped into the global attribute GA : `MT[GA][LC]= f(LAS)` where LAS is a subset of the local attributes of LC.

Global Virtual View (GVV) Annotation

To annotate a GVV means to assign a name and a set (eventually empty) of meanings to each global element (class or attribute).

In order to semi-automatically associate an annotation to each global class, we consider the set of all its "broadest" local classes, w.r.t. the relationships included in the Common Thesaurus. On the basis of this set, the designer will annotate the global class as follows:

name choice: the designer is responsible for the choice of the name: the system only suggests a list of possible names. The designer may select a name within the proposed list or introduce a new one.

meaning choice: the union of the meanings of the "broadest" local classes are proposed to the designer as meanings of the Global Class; the designer may change this set.

A similar approach is used for Global Attributes Annotation.

3. CONCLUSION AND FUTURE WORK

MOMIS supports the semiautomatic building and annotation of domain ontologies by integrating the schemas of information sources. The MOMIS framework is currently adopted in the Semantic Web Agents in

(c) 2004 IFIP

Integrated Economies (SEWASIE) European research project (www.sewasie.org), coordinated by UNIMORE. SEWASIE aims at implementing an advanced search engine that enables intelligent access to heterogeneous data sources on the Web via semantic enrichment, providing the basis for structured secure Web-based communication. To achieve this goal, SEWASIE creates a virtual network based on Sewasie information nodes (SINodes), which consist of managed information sources, wrappers, and a metadata repository. SINodes metadata represent GVVs of the overall information sources that each manage. To maintain the GVV of a SINode, we are investigating two distinct aspects: the system overload in maintaining the built ontologies and the effects of inserting new sources that could modify existing ontologies. Future work will addresses the improving of the annotation phase by allowing the designer to face multilingual environments, that is adopting a multiligual lexical database.

REFERENCES

1. T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993.
2. S. Bergamaschi, S. Castano, D. Beneventano, M. Vincini: "Semantic Integration of Heterogeneous Information Sources", *DKE*, Vol. 36, Num. 1, Pages 215-249, 2001.
3. M. Lenzerini Data Integration: A Theoretical Perspective. *PODS '02*: 233-246, 2002.
4. S. Chawathe, Garcia Molina, H., J. Hammer, K. Ireland, Y. Papakostantinou, J.Ullman and J. Widom. The TSIMMIS project: Integration of heterogeneous information sources. *IPSJ Conference*, Tokyo, Japan, 1994.
5. M.J. Carey, L.M. Haas, P.M. Schwarz, M. Arya, W.F. Cody, R. Fagin, M. Flickner, A. Luniewski, W. Niblack, D. Petkovic, J. Thomas II, J.H. Williams, and E.L. Wimmers, Towards heterogeneous multimedia information systems: The Garlic approach. *IBM Almaden Research Center,* San Jose, CA 95120, 1996.
6. Y. Arens, C. A. Knoblock, and C. Hsu. Query processing in the sims information mediator. *Advanced Planning Technology*, 1996.
7. M. R. Genesereth, A. M. Keller, and O. M. Duschka. Infomaster: An information integration system. *SIGMOD'97*, pages 539--542, Tucson, Arizona, USA, 1997. ACM Press.
8. T. Kirk, A. Y. Levy, Y. Sagiv, and D. Srivastava. The information manifold. *AAAI Spring Symposium on Information Gathering from Heterogeneous*, 1995.
9. E. Mena, A. Illarramendi, V. Kashyap, and A. P. Sheth. Observer: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distributed and Parallel Databases*, 8(2):223--271, 2000.
10 D. Beneventano, S. Bergamaschi, S. Lodi and C. Sartori, Consistency Checking in Complex Object Database Schemata *IEEE TKDE.*, vol. 10, no. 4, 1998, pp. 576–598.
11 S. Abiteboul, P. Buneman, and D. Suciu. Data on the Web - From Relations to Semistructured Data and XML. Morgan Kaufmann, 2000.
12. A.G. Miller. A lexical database for English. *Communications of the ACM*, 38(11):39:41,1995.
13. D. Beneventano, S. Bergamaschi, C. Sartori, M. Vincini ODB-QOptimizer: a tool for semantic query optimization in OODB. *ICDE'97*, UK, April 1997.
14. S. Castano, V. De Antonellis, S. De Capitani di Vimercati. Global viewing of heterogeneous data sources. *IEEE TKDE*, 13(2), 2001.

(c) 2004 IFIP