THE HAV DATA INTEGRATION APPROACH
The Mapping in HAV

Fatima Boulçane
LIRE Laboratory
Mentouri University of Constantine

Keywords: Mediator, data integration, GAV, LAV.

Abstract: This paper provides an overview on a hybrid approach of heterogeneous data integration which we term Hybrid As View (HAV) and it focuses on the mapping between the global schema and source schemas through the partial schemas. The contribution of this approach is on two complementary axes: (i) to propose a multi-mediators architecture essentially made up of two types of components: specialized mediators and a global mediator. Each of the specialized mediators provides an integrated view of sources with the same model. The global mediator integrates the partial schemas provided by the set of the specialized mediators to provide an access on a uniform view represented by a global schema; (ii) to model the relation between the global schema and the sources through the partial schemas by combining the best of the two approaches Global As View (GAV) and Local As View (LAV).

1 INTRODUCTION

The data integration systems aim to offer a uniform view on a set of heterogeneous sources for end-users or applications. Data integration refers to the problem of combining data residing at autonomous and heterogeneous sources, and providing users with a unified global schema (Xu et al., 2004). The integration systems are often based on the three levels mediation architecture proposed in (Koffina, 2005) (Figure 1). A mediator layer represents the integration part. It interacts between the data sources and the applications or the users. To reach the data sources, the mediators call upon the wrappers, which convert the data of the sources in a data model used by the mediators.

The problems of data integration systems are closely related to the heterogeneity of the data, with their semantic difference, the differences in terms of accessibility of the sources, functions offered and availabilities of cooperation (Ullman, 1997).

We can classify the data integration systems according to the relation between the schemas of the local sources and the mediator unified global schema (Lenzerini, 2002, Cali, 2002). The definition of the global schema can be done according to the two basic approaches: the GAV approach and the LAV approach. Furthermore, hybrid approaches based on both GAV and LAV have been recently proposed (Koffina, 2005).

The GAV approach consists in defining the global schema as a view on local schemas. The principal advantage of this approach lies in the fact that the rewriting of the requests is simple. On the other hand, it is difficult to add new sources to the system (Xu et al., 2004).

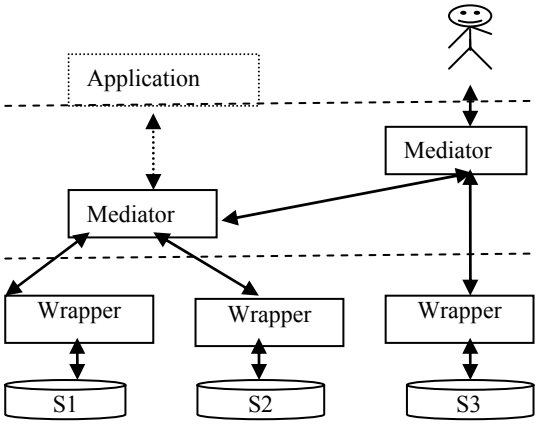

Figure 1: Architecture of mediation.

In the LAV approach, the local sources are defined like views on the global schema. With this approach, the problem of rewriting requests is more

Boulçane F. (2007).
THE HAV DATA INTEGRATION APPROACH - The Mapping in HAV.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - DISI*, pages 490-493
DOI: 10.5220/0002398004900493

complex because a request on the global schema must be reformulated according to the schemas of the sources (Li Xu, 2004).

Both GAV and LAV have some drawbacks. Thus, combining Global and Local approaches have been proposed. The first one is called GLAV (Friedman, 1999). In this approach, we are able to express a local source in terms of the global schema (LAV), a global source in terms of the local sources. Query rewriting in this approach is shown to be no harder than it is for the LAV approach (Koffina, 2005).

BAV is another data integration approach. It is a rich integration framework, which is based on the use of reversible sequences of primitive schema transformations, called transformation pathways (Boyd, 2004).

BGLaV is an alternative point of view that is neither GAV nor LAV. The approach uses source-to-target mappings based on a predefined conceptual target schema, which is specified ontologically and independently of any of the sources. The proposed data integration system is easier to maintain than both GAV and LAV, and query reformulation reduces to rule unfolding (Li Xu, 2004).

The specificity of our architecture, is that in addition to the characteristics of the current systems, seeks to promote an architecture which improves the integration of heterogeneous data by supporting the combination of two approaches GAV and LAV: HAV and to make us benefit from their advantages.

The various information integration systems containing mediator are characterized by, on the one hand, the languages used to model the global schema, the schemas of the data sources to be integrated and the requests of the users, and on the other hand, the mapping between the global schema and the schemas of the data sources to be integrated (Lenzerini 2002). This paper focuses on the mapping in HAV.

We organize the contributions in this paper as follows. Section 2 presents an overview of HAV approach. Section 3 will be devoted to the presentation of the HAV formal definition, especially the mapping. Finally, we summarize and we release some prospects in section 4.

2 OVERVIEW OF THE HAV APPROACH

In this section we present an overview of a multi-mediators architecture described in a previous work (Boulçane, 2006) essentially made up of two types of components: specialized mediators and a global mediator (Figure2). The architecture which rises from the HAV approach is an architecture where specialized mediators place themselves in the heart of the mediation architecture. They are considered as virtual sources to be requested by the global mediator via the specialized wrappers.

The specialized mediators provide each one an integrated view of sources with the same model called partial schemas, which are integrated by the global mediator into a global schema. It is essential to determine the relation between these schemas and the sources. This essentially consists in defining the correspondence between the global schema and the sources via the partial schemas. To build such a system, we propose to define the partial schemas with the LAV approach and the global schema with the GAV approach.

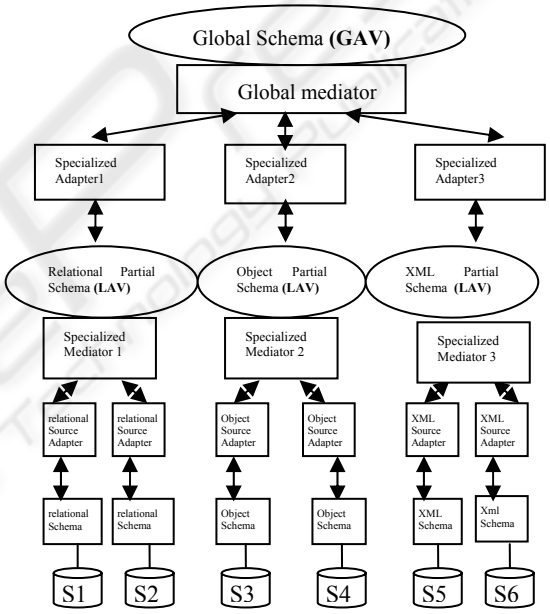

Figure 2: The multi-mediators Architecture.

3 FORMAL DEFINITION OF HAV

We use the formalism suggested by (Lenzerini, 2002) for the data integration systems based on a global schema and which we adapt to the integration system based on the HAV approach.

Definition 1 (Lenzerini, 2002): A data integration system is a triplet $(G, S, Mg_{,S})$ where:

- G is the global schema expressed in a language L_G, over the alphabet A_G. The language L_G determines the expressiveness allowed for

specifying the global schema, i.e., the set of constraints that can be defined over it.
- S is the set of the local schemas. It is modelled in the source language L_S over the alphabet A_S. The language determines the set of constraints that can be defined over it.
- $Mg_{,S}$ is the mapping between G and S.

For the needs of the HAV data integration system, we recursively use the definition 1.

Definition 2: A data integration system in HAV is a triplet $(G, I_{S,}, Mg_{,S})$ where:

- G is the global schema expressed in a language L_G, over alphabet A_G. The language L_G determines the set of constraints that can be defined over it.
- I_S, is a set of data integration systems like a triplet $(S, S_S,,M_{SS,S})$ where:
  - S is the schema of a specialized mediator, expressed in a language L_S on an alphabet A_S. The language determines the set of constraints that can be defined over it.
  - S_S is the schema of the source with the same model as S on an alphabet A_S.
  - $M_{SS,S}$ is the mapping between S and S_S, constituted by a set of assertions of the form:
    $q_{Ss} \rightarrow q_s$
    $q_s \rightarrow q_{Ss}$
- $Mg_{,S}$ is the mapping between G and S, constituted by a set of assertions of the form:
  $q_s \rightarrow q_G$
  $q_G \rightarrow q_s$

In other words, the global schema G provides an integrated view of the partial schemas S, where each one is the result schema of a specialized mediator.

Definition 3: Given the definition 2, mappings $M_{SS,S}$ and $Mg_{,S}$ in HAV approach are in the form:

$M_{SS,S}$: $S_{Si}(X) \leftarrow S_1(X_1), S_2(X_2), \ldots, S_k(X_J, Z_J)$
Where $X = U_i X_i$
S_i are relations of partial schemas
S_{Si} are local relations.

$Mg_{,S}$: $Gi(X) \leftarrow S_1(X_1), S_2(X_2), \ldots, S_k(X_k)$
Where $X = U_i X_{i,}$
Gi are global relations
S_i are relations of partial schemas

Example:

We call upon two relational sources S1 and S2, and two semi-structured sources S3 and S4.The relational sources S1 and S2 are integrated like local views on partial relational schema PS1.The sources S3 and S4 are integrated to give partial schema PS2 in XML. This corresponds to the integration with the LAV approach. The global schema is relational and is defined like a global view on partial schemas PS1 and PS2. This corresponds to the integration with the GAV approach.

The sources to be integrated contain information on films. The partial schema PS1 contains films since 1960 and their criticisms since 1990. PS2 contains films. The global schema consists of two tables. One contains information on films and the other contains articles relating to films.

The mappings $M_{SS,S}$:

We assume that the specialized shema1 (PS1) consists of two relations:

Film (Fid, title, productor, year)
Critiques (Fid, critique)

We integrate the two sources S1 and S2 on the specialized schema1. The description of these sources is:

For S1:

Film (Fid, title, year, productor) ←
Film (Fid, title, productor, year)
Film.year>1960

For S2 :

Critique (Fid, title, critique) ←
Film (Fid, title, productor, year)
Critiques (Fid, critique)
Film.year>1990,
Film.Fid =Critiques.Fid

From the description above we can conclude that the relation Film of S1contains the identifier of the film, the title of the film, the productor of the film and the year, only for films since 1960. While the relation Critique of S2 contains the identifier of the film, the title of the film and the critique of the film since 1990.

The mappings : $Mg_{,S}$:

We assume that there are two specialized schemas PS1 and PS2. We suppose that after the translation of PS2 into the relational model the relational schema of PS2 consists of one relation:

Film (Fid, title, director, kind).

We integrate PS1 and PS2 on the global schema. The description of the global schema is:

Films (Fid, title, realisator, year, kind) ←
PS1.Film(Fid, title, realisator, year, NULL)
PS2.Film(Fid, title, director, NULL, kind)
Articles (title, critique) ←
PS1.Film(title, Null)
PS1.Critiques(title, critique)

Example of requests on the global schema:
Which are criticisms of films having for title: 'Freedom'?
We send the following request to the global schema:

```
Select Films.title, Articles.critique
From Films, Article
Where Films.title = ` Freedom' And
      Films.title = Articles.title
```

The translation of the request over PS1 and PS2 (GAV) is:

```
Select PS1.Film.title, PS1.Critiques.critique,
       PS2.Film.title
From PS1.Film, PS1.Critiques, PS2.Film
Where (PS1.title='Freedom' And
        PS1.Film.Fid=PS1.Critiques.Fid Or
        PS2.title='Freedom' And
        PS2.Film.Fid=PS1.Critiques.Fid)
```

The translation of the request over the sources S1, S2, S3, and S4 (LAV) is:

```
Select S1.title, S2.Critique.critique
From S1.Film, S2.Critique, S2.Film
Where S1.title='Freedom' And
      S1.Film.Fid=S2.Critique.Fid Or
      S2.title='Freedom' And
      S1.Film.Fid=S2.Critiques.Fid
Union (the relational request on S3 and S4 is
       as follows)
Select S3.Film.title, S4.Film.critique
From S3.Film, S4.Film
Where S3.title='Freedom' And
      S3.Film.Fid=S4.Film.Fid
```

4 CONCLUSION

This paper presents HAV, a hybrid approach to data integration which combines GAV and LAV to make us benefit from their advantages, and defines the HAV mapping.

The contribution of such an approach is that HAV is more effective and practicable because the set of the partial schemas concerned by the HAV integration approach is small and stable. Thus, in HAV It will be less complex for the specialized mediators to carry out the sub-requests (because each one has a reduced number of sources to integrate, and these sources are in the same model), than if the global schema itself were built with LAV approach, so the complexity to reformulate queries is reduced.

Noting that it does not exist yet a truly definite Benchmark which makes it possible to evaluate the performances of a mediator. We intend to adapt one among those which exist with our context. In particular, the result carried out by (Dang Ngoc, 2003) seems to be promising: that it is possible to use a tree structure of mediators without harming the evaluation.

REFERENCES

Boulçane, F., 2006. An approach of mediation. In *proceedings of the IEEE ICTTA'06, International Conference on Information & Communication Technologie*. Volume 2, Page(s): 3546-3551.

Boyd, M., Lazanitis, C., Kittivoravitkula, S., Brien, P. M., and Rizopoulos, N., 2004. AutoMed: A BAV Data Integration System for Heterogeneous Data Sources. In *CAiSE'04, 16th International Conference on Advanced Information Systems Engineering*. Riga, Latvia, June 7-11, Proceedings Springer-Verlag.

Cali, A., Calvanese, D., Gracomo, G., Lenzerini, M., 2002. On the expressive power of data integration systems. *Lecture notes in computer science, volume 2503.*

Dang Ngoc, T.T., 2003. Fédération de données semi-structurées avec XML. *Thèse de Doctorat*. Université de Versailles-Saint-Quentin.

Friedman, M., Levy, A., Millstein , T., 1999. Navigational plans for data integration. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications.* Pages: 67 – 73, Orlando, Florida, United States.

Koffina, I., Serfiotis, G., Christophides, V., 2005. Foundations for Information Integration: A State of the Art. In *FORTH-ISL, Sixth Framework Programme.*

Lenzerini, M., 2002. Data Integration: A Theoretical perspective. In *Proceeding of the twenty-first ACM SIGMOD-SIGACT-SIGART, Symposium on Principles of Database Systems.*

Ullman, J.D., 1997. Information Integration using logical views. In *Proc.of the 6th Int. Conf. on Database Theory (ICDT'97).* Volume 1186 of Lecture Notes in Computer Science, pages 19-40, Springer.

Xu, L., Embley, D.W., 2004. Combining the Best of Global-as-View and Local-as-View for Data Integration. *In Information Systems Technologies and Its Applications.* ISTA. 123–136.