On the Use of Dimension Properties in Heterogeneous Data Warehouse Integration*

Marius-Octavian Olaru and Maurizio Vincini

Department of Information Engineering - DII
University of Modena, Italy
{mariusoctavian.olaru, maurizio.vincini}@unimore.it

Abstract. A new trend in Business Intelligence is the process of combining information from two or more different and heterogeneous Data Warehouses. Existing solutions rely mostly on the Extract-Transform-Load (ETL) approach, a costly and laborious process. The process of Data Warehouse integration can be greatly simplified by developing methods to semi-automatically discover semantic mappings among attributes of two or more different, heterogeneous Data Warehouse schemas, like the one proposed in this paper.

1 Introduction

The dynamic economical context that characterizes today's economical markets has increased the number of cases where companies need to integrate information coming from two or more heterogeneous, independently developed, Data Warehouses. For example, it is common for two companies to merge, or for one company to acquire one or more other companies; in both cases, the independent DWs must be integrated in order to offer management a unified view over the entire available information. A manual process of DW integration, base on ETL procedures, is laborious, time consuming and itself prone to errors. An automatic or semi-automatic method can increase the efficiency of such process. This has been proved in the data integration area where designers make use of semi-automatic tools (like [2]) as support for the mapping discovery process between independent and heterogeneous data sources. The problem of heterogeneous DW integration can be seen as a subcase of data integration, as the information to be integrated is multidimensional, that is why the authors believe that a specific solution that takes into account this particularity may yield better results than classical data integration techniques.

In this paper, we propose a semi-automatic method to discover mappings among Data Warehouses dimensions that exploits topological properties of the DW dimension levels together with semantic annotation techniques. In particular, we rely on graph theory and the normalization *Combined Wordsense Disambiguation* (CWSD) techniques developed inside the MOMIS data integration project[14].

This paper is organized as follows: Section 2 presents an overview on related work, Section 3 provides a description of the proposed method, meanwhile Section 4 draws the conclusions and presents the future work.

* This work has been published in [7].

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

2 Related Work

Until now there have been few attempts to formalize the problem of Data Warehouse Integration, and no complete solution has been yet proposed. Formalization approaches include those proposed by Kimball in [10] or the work presented in [8, 15], where the authors define the concept of *conformed dimensions*, from a theoretical point of view.

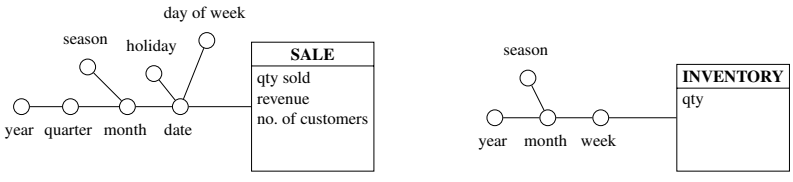
A solution to the problem is proposed in [9], where an architecture for the exchange of multidimensional information (called BIN-*Business Intelligence Network*, or *Peer-to-Peer Data Warehouse*) is proposed, and the OLAP query reformulation problem in the proposed architecture is formalized. A Peer-to-Peer Data Warehouse is a network in which the local peer has the capability of executing queries over the local Data Warehouse and to forward them to the network. The queries are then rewritten against the remote compatible multidimensional models using a set of mapping predicates between the local schema and the remote schema. The main problem of this architecture is that the mapping predicates have to be manually generated, so the approach is not scalable.

In [1] there is an attempt to automate the mapping process between multidimensional structures. The authors present a method, based on early work in data integration [4, 6], that allow designers to automatically discover schema mappings between two heterogeneous Data Warehouses. The *class similarity* (or *affinity* concept, as described in [5]) is used to find similar attributes (*facts*, *dimensions*, *aggregation levels* and *dimensional attributes*) and similarity functions for multidimensional structures based on that concept are proposed. Our method makes use of semantics, but with a different approach: [1] relies on semantic relations to discover the mappings among attributes, whereas we exploit them only as a validation technique.

3 Mapping Discovery

We propose an instance-level[12] technique for the semi-automatic discovery of mappings between heterogeneous DW dimension levels that exploits semantics and the specific graph-like structure that the partial order relationship imposes on the dimensional attributes.

Although may contain different information, dimensional hierarchies representing the same concept usually maintain the same structure. Consider, for example, that the time dimension in the first schema of Figure 1 (S_1) contains all the *days* between January



(a) SALE(S_1) (b) INVENTORY(S_2)

Fig. 1. Sample schemas

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

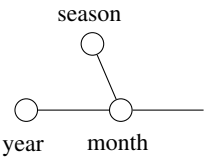


Fig. 2. A common *sub-dimension*

1^{st} 2008 and December 31^{st} 2010 (three complete years), meanwhile the time dimension from the second schema S_2 contains all the *weeks* between January 1^{st} 2010 and December 31^{st} 2011 (two complete years). A simple data intersection will reveal that the attributes values sets are only partially overlapped, due to the fact that the dimensions cover different time periods, and to the fact that the dimensions contain distinct attributes (for example, S_1 contains the dimensional level *quarter*, meanwhile S_2 contains the dimensional level *week*). However, some common topological properties are preserved among the two dimensions. For example, a *year* is an aggregation of 12 different *months* in both schemas; similarly, a *season* is composed of 3 different *months*. These common properties may be used to identify similar elements in the two dimension levels, and to generate mappings that are able to express the exact relationships among them.

The dimension in Figure 2 can thus be seen as a common *sub-dimension*[1] of the two initial dimensions that can be used to semi-automatically map levels of S_1 to levels of S_2 and *vice-versa*.

The method propose in this paper can be summarized in this steps:

1. *Candidate mappings set generation*: pairs of equivalent nodes, together with a set of 5 rules are used to generate the list of candidate mappings.
2. *Semantic validation*: the candidate mappings are then validated using a semantic approach.

3.1 Candidate Mappings Generation

As mappings predicates, we used a subset of those proposed in [9]. In particular, we considered the *equi-level, roll-up, drill-down* and *related* mapping predicates.

For the computation of the common sud-dimension, we simply consider the dimensions as directed labeled graphs, where the label of each edge is the *cardinality ratio*[2] among two different dimension levels. Using graph theory, it is possible to compute a maximum-rank common subgraph, and to identify pairs of nodes of the two initial graphs that correspond to the same node in the subgraph, that are called *equivalent*. In the example in Figure 1, the nodes $S_1.month$ and $S_2.month$ are equivalent, as are the nodes $S_1.year$ and $S_2.year$, and $S_1.season$ and $S_2.season$.

[1] A *sub-dimension* is intended as a new dimension obtained by removing one or more aggregation levels

[2] By *cardinality ratio* we simply intend the ratio among the number of different elements contained in different aggregation levels.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

We discover mappings by exploiting the following 5 rules:

Let P_x and P_y be two nodes of the first dimension such that there is a path from P_x to P_y, and P_h and P_k two nodes of the second dimension such that there is a path from P_h and P_k

1. **Rule 1**: If P_x and P_h are equivalent, add the mappings:
   * $P_x\ (equi-level)\ P_h$
2. **Rule 2**: if $P_x\ (equi-level)\ P_h$, add the mappings:
   * $P_y\ (roll-up)\ P_h$
   * $P_h\ (drill-down)\ P_y$
3. **Rule 3**: if $P_y\ (equi-level)\ P_h$, add the mappings:
   * $P_x\ (drill-down)\ P_h$
   * $P_h\ (roll-down)\ P_x$
4. **Rule 4**: if $P_y\ (equi-level)\ P_h$, add the mappings:
   * $P_x\ (drill-down)\ P_k$
   * $P_k\ (roll-up)\ P_x$
5. **Rule 5**: for every nodes P_x and P_h of the two graphs for which there has not been found any mapping rule, add the mapping:
   * $P_x\ (related)\ P_h$

Using these simple rules, it is possible to generate a complete set of mapping candidates that express the relationship among every two elements of the two schemas. For example, Rule 1 will produce a mapping $S_1.month\ (equi-level)\ S_2.month$, while Rule 3 will produce the rule $S_1.month\ (roll-up)\ S_2.week$[3].

3.2 Semantic Mapping Validation

To validate the discovered mappings, we decided to weight and prune them using a semantic approach based on Lexical Annotation. Lexical Annotation is the process of explicit assignment of one or more meanings to a term w.r.t. a thesaurus. To perform lexical annotation, we exploited the CWSD[3, 13] (*Combined Word Sense Disambiguation*) algorithm implemented in the MOMIS Data Integration System which associates to each element *label* (i.e., the name of the element) one or more meanings w.r.t. the WordNet lexical thesaurus [11].

One important issue in Data Warehousing is that often the schema and attribute labels are abbreviated, or they are Compound Nouns (CN). This makes it difficult for an automatic algorithm to assign a semantic meaning w.r.t. a thesaurus, like WordNet. The main reason for which we chose the algorithm in [14] is that it is able to increase the accuracy of the annotation process by doing both context-aware abbreviation expansion and CN disambiguation.

Starting from lexical annotations, we can discover semantic relations among attributes of the different Data Warehouses by navigating the wide semantic network of WordNet. In particular, the WordNet network includes[4]: *synonim*, *hypernim* and

[3] For the sake of simplicity, we do not present the entire mapping set.

[4] WordNet includes other semantic and lexical relations such as antonym, cause etc. which are not relevant for our approach

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

Table 1. Coefficient Assignment

P_i/P_j	equi-level	roll-up	drill-down	related
same/synonyms	1	0.7	0.7	0.7
hypernyms	0.9	0.7	1	0.8
hyponyms	0.9	1	0.7	0.8
correlated terms	0.7	0.7	0.7	1
holonyms	0.7	1	0.3	0.8
meronyms	0.7	0.3	1	0.8

meronym relationships. We added the *correlated terms* relation that can be directly derived by WordNet: two terms are correlated if they are connected by a hyponym or meronym relation to the same WordNet synset. Thus, for each identified mappings, we first annotate each label by using the algorithm in [14] and then, we discover the shortest path of semantic relations connecting the two elements into the WordNet network. The goal is to validate the mappings by computing for each of them a similarity weight on the basis of the identified WordNet paths. We computed the similarity weight by assigning to every edge (i.e., WordNet relation) of the path a coefficient using the assignment rules in Table 1. The final similarity weight is given by the product of the single coefficients (thus, long paths will usually have lower similarity weights than short or direct paths). These coefficients were defined by considering that an *equi-level* is semantically similar to a *same/synonim* relationship in WordNet, *roll-up* is similar to a *hyponim* or to a *holonym*, meanwhile a *drill-down* relationship is similar to the *hypernym* or *meronym* relationship. For the other mapping/relationship combinations we associated coefficients (lower than 1) on the basis of their relevance. For example, to the combination *drill-down/holonym*, we associated a low coefficient (0.3) as they semantically represent opposite concepts. Starting from these computed coefficients, we can prune the discovered mappings, by applying a coefficient threshold.

4 Conclusions and Future Work

In this paper, we argued that topological properties of dimensions in a Data Warehouse can be exploited to find semantic mappings between two or more different Data Warehouses. We showed how these properties can be used in conjunction with semantic techniques to efficiently generate a mapping set between the attributes of the dimensions of two independent Data Warehouses.

Our method is effective for the integration of Data Warehouses in the case of reasonable complete DW instances. If partial/scarce information is present in the DWs, then the cardinality ratio among levels might vary rendering the mapping generation step inefficient.

Another important observation we made during our research is that the mapping predicates have no exact correspondence with the WordNet semantic relations, so it is impossible to assign an exact semantic weight coefficient to a specific type of mapping. These semantic weights depend on the context of the Data Warehouse. In the future,

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

we plan to investigate on how the fine tuning of these coefficients can affect the accuracy of the mapping method. In any case, for maximum accuracy a human validation is required. This is also an issue in data integration where developers/analysts rely on semi-automatic tools to discover semantic correspondences, but unfortunately the process cannot be entirely automatic if maximum accuracy is required. As any approach proposed so far in Data Warehouse integration has flaws, we believe that a combination of approaches (like the topological/semantic approach proposed in this paper) could improve the accuracy of the mapping discovery process.

References

1. Banek, M., Vrdoljak, B., Tjoa, A.M., Skocir, Z.: Automated Integration of Heterogeneous Data Warehouse Schemas. IJDWM **4**(4) (2008) 1–21
2. Beneventano, D., Bergamaschi, S., Gelati, G., Guerra, F., Vincini, M.: MIKS : An Agent Framework Supporting Information Access and Integration. In Klusch, M., Bergamaschi, S., Edwards, P., Petta, P., eds.: AgentLink. Volume 2586 of Lecture Notes in Computer Science., Springer (2003) 22–49
3. Bergamaschi, S., Bouquet, P., Giacomuzzi, D., Guerra, F., Po, L., Vincini, M.: An Incremental Method for the Lexical Annotation of Domain Ontologies. Int. J. Semantic Web Inf. Syst. **3**(3) (2007) 57–80
4. Bergamaschi, S., Castano, S., Vincini, M.: Semantic Integration of Semistructured and Structured Data Sources. SIGMOD Record **28**(1) (March 1999) 54–59
5. Bergamaschi, S., Castano, S., Vincini, M., Beneventano, D.: Retrieving and integrating data from multiple sources: the MOMIS approach. Data Knowl. Eng. **36**(3) (2001) 215–249
6. Bergamaschi, S., Guerra, F., Vincini, M.: A Peer-to-Peer Information System for the Semantic Web. In: AP2PC. (2003) 113–122
7. Bergamaschi, S., Olaru, M.O., Sorrentino, S., Vincini, M.: Semi-automatic Discovery of Mappings Between Heterogeneous Data Warehouse Dimensions. In: International Conference on Advances in Communication and Information Technology, Amsterdam, ACEEE (2011)
8. Cabibbo, L., Torlone, R.: On the Integration of Autonomous Data Marts. In: SSDBM, IEEE Computer Society (2004) 223–
9. Golfarelli, M., Mandreoli, F., Penzo, W., Rizzi, S., Turricchia, E.: Towards OLAP query reformulation in Peer-to-Peer Data Warehousing. In Song, I.Y., Ordonez, C., eds.: DOLAP, ACM (2010) 37–44
10. Kimball, R., Ross, M.: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Volume 32. John Wiley & Sons, Inc., New York, NY, USA (2002)
11. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: WordNet: An on-line lexical database. International Journal of Lexicography **3** (1990) 235–244
12. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal **10**(4) (2001) 334–350
13. Sorrentino, S., Bergamaschi, S., Gawinecki, M.: NORMS: An automatic tool to perform schema label normalization. In: ICDE. (2011) 1344–1347
14. Sorrentino, S., Bergamaschi, S., Gawinecki, M., Po, L.: Schema label normalization for improving schema matching. Data & Knowledge Engineering **69**(12) (December 2010) 1254–1273
15. Torlone, R.: Two approaches to the integration of heterogeneous data warehouses. Distributed and Parallel Databases **23**(1) (2008) 69–97

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,