Generating Multidimensional Schemas from the Semantic Web

Oscar Romero and Alberto Abelló

Universitat Politècnica de Catalunya
Jordi Girona 1-3, E-08034 Barcelona, Spain
{oromero,aabello}@lsi.upc.edu

Abstract. In this paper, we introduce a semi-automatable method aimed to find the business multidimensional concepts from an *ontology* representing the organization domain. With these premises, our approach falls into the *Semantic Web* research area, where ontologies play a key role to provide a common vocabulary describing the meaning of relevant terms and relationships among them.

Key words: OLAP, Multidimensional Design, Ontologies, Semantic Web

1 Introduction

OLAP (*On-line Analytical Processing*) tools are intended to ease information analysis and navigation all through the business data previously integrated in a huge repository of data, the *Data Warehouse* (DW), from a *multidimensional* perspective. Despite traditional methodologies to design multidimensional DWs are typically carried out manually by DW experts, a few works automatizing the design of multidimensional databases have been presented in the last years. However, all these approaches start from a detailed analysis of the data sources to determine the multidimensional concepts in a reengineering process, as well as all of them also assume to start from a relational OLTP (*On-Line Transaction Processing*) system.

We introduce a semi-automatable method aimed to find the business multidimensional concepts from an *ontology* representing the organization or business domain. With these premises, our approach falls into the *Semantic Web* research area. This approach raises new challenges with regard to traditional modeling so that the multidimensional design process needs to be reconsidered. Mainly, we can not provide the method with end-user requirements to guide the process, since we are working over external (maybe unknown) data and, a priori, the user does not know what kind of information will be available. Moreover, we can not perform massive data mining over all existent instances due to a complexity issue, nor assume data sources are implemented over relational databases as many traditional methods do. In fact, we need to focus on the ontologies representing the knowledge contained in those sites, and narrow and guide the DW design process from knowledge captured in the ontologies, and at most, extract missing knowledge by means of samples of data from some known sites.

Section 2 discusses about the related work presented in the literature underlining those automatable approaches. Section 3 sets the foundations of our method that is presented in section 4. Finally, section 5 concludes this article.

2 Related Work

In the literature, partially automatized approaches to design multidimensional DWs ([1]) and those fully automatizing the process ([2], [3] and [4]) always start from a thorough analysis of the relational sources to determine the multidimensional concepts in a reengineering process. We would remark two main general restrictions shared by all these methods not suiting them for the multidimensional design over The Web: (1) they all work exclusively over relational sources, and (2) they work with a *table granularity*. That is, each table in the relational sources is determined to play a fact or a dimension role, overlooking their attributes; and as discussed in [5], a table, a relationship or even an attribute may be playing a fact/dimension role. Hence, in these methods, attributes within each table are considered as a whole, since they are not able to work with finer granularities. Consequently, they need a certain degree of normalization in the relational schema to work properly. Working with ontologies we will be able to get rid of these two inherent restrictions.

3 Problem Context

In this section we aim to define the context of the problem introduced and point out those criteria our method is based on; that is, those criteria allowing us to identify multidimensional concepts. Multidimensionality pays attention to two main aspects; *placement of data in a multidimensional space* and *correct summarizability of data*. Therefore, our method looks for meaningful conceptual schemas with orthogonal **Dimensions** fully functionally determining **Facts**, and free of summarizability problems.

Bearing in mind that our method input would be an ontology, we also assume the following premises: (1) the ontology is expressed in an ontology language providing basic reasoning tools such as *subsumption*, allowing us to work with taxonomies of concepts. For instance, OWL (*Web Ontology Language*), an W3C recommendation, fits properly for our purposes. (2) We have a mapping among the ontology concepts and the data sources. In the Semantic Web area this mapping is supposed to exist and, for instance, preserving the concept names in the implementation would be enough.

4 Our Method

In this section we expose an schematic view of our method composed by three well-differentiated steps (see figure 1):

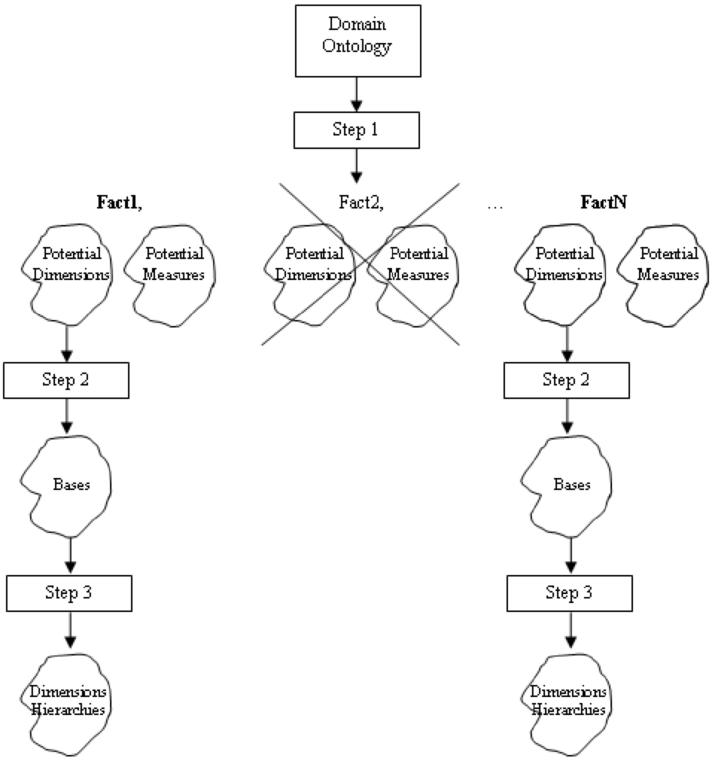

Fig. 1. An schematic view of the method proposed

First step: It looks for potential subjects of analysis (i.e. **Facts**). We consider a concept to be a potential subject of analysis if it is related to as many potential **Dimensions** (analysis perspectives) and **Measures** (factual data) as possible. So that, our aim is twofold:

- Discover potential analysis **Dimensions**: According to multidimensionality, each instance of data must be identified (i.e. placed in the multidimensional space) by a point in each of its analysis **Dimensions**. In our approach, to identify potential analysis **Dimensions** we look for concepts being functionally determined by a given concept (the potential **Fact**). To carry out this step we suggest to take advantage of the reasoning services provided by ontology languages to automatically point out **Dimensions**.
- Pointing out **Measures**: Typically, **Measures** are numeric facts allowing data aggregation. We consider any numeric *datatype* to be a **Measure** of a given **Fact** if it preserves a correct aggregation of data.

At the end of this step, we will ask the user to choose his/her subjects of interest among those concepts proposed by the method as potential **Facts** (for instance, in figure 1, the user would have disregarded the proposed *Fact2*). The rest of steps will be carried out once per each subject of analysis identified. Consequently, from each **Fact**, it will give rise to a multidimensional conceptual schema.

Second step: It points out sets of concepts likely to be used as **Base** for each **Fact** identified in previous step. We call a **Base** to those *minimal* sets of **Levels** fully functionally determining a **Fact**. **Bases** must contain *orthogonal* (i.e. functionally independent) **Dimensions**, and a set of potential **Dimensions** will be considered a *feasible* **Base** if they are able to identify all the instances of a **Fact**. In a few words, we look for concepts being able to univocally identify objects of analysis (i.e. to univocally place data in the multidimensional space).

Third step: In this step we give rise to **Dimensions** hierarchies in order to allow summarizability of data; one of the multidimensionality principles. In our approach, from every concept identified as **Dimension**, we conform their hierarchies of **Levels** from those concepts related to them by typical whole-part relationships (i.e. one-to-many relationships); or, as known in OLAP, “Roll-up” relationships.

5 Conclusions

In this paper we have introduced a semi-automated method to point out multidimensional concepts from an ontology representing our business domain. Unlike traditional approaches that work exclusively over relational sources, our approach is able to integrate information from heterogeneous data sources that describe their domain through ontologies. One of the most promising areas where to apply our method is the Semantic Web.

Acknowledgments. This work has been partly supported by the Spanish Ministerio de Educación y Ciencia under project TIN 2005-05406.

References

1. Golfarelli, M., Maio, D., Rizzi, S.: The Dimensional Fact Model: A Conceptual Model for Data Warehouses. Int. Journal of Cooperative Information Systems (IJCIS) **7**(2-3) (1998) 215–247
2. Phipps, C., Davis, K.C.: Automating Data Warehouse Conceptual Schema Design and Evaluation. In: Proc. of 4th Int. Workshop on Design and Management of Data Warehouses (DMDW’02). Volume 58 of CEUR Workshop Proceedings., CEUR-WS.org (2002) 23–32
3. Jensen, M.R., Holmgren, T., Pedersen, T.B.: Discovering Multidimensional Structure in Relational Data. In: 6th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK’04). Volume 3181 of LNCS., Springer (2004) 138–148
4. Romero, O., Abelló, A.: Multidimensional Design by Examples. In: Proc. of 8th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK 2006). Volume 4081 of LNCS., Springer (2006) 85–94
5. Cabibbo, L., Torlone, R.: A Logical Approach to Multidimensional Databases. In: Proc. of 6th Int. Conf. on Extending Database Technology (EDBT 1998). Volume 1377 of LNCS., Springer (1998) 183–197