Data-Driven Multidimensional Design for OLAP

Oscar Romero and Alberto Abelló

Universitat Politècnica de Catalunya, Barcelona, Spain
{aabello,oromero}@essi.upc.edu

Abstract. OLAP is a popular technology to query scientific and statistical databases, but their success heavily depends on a proper design of the underlying multidimensional (MD) databases (i.e., based on the *fact / dimension* paradigm). Relevantly, different approaches to automatically identify *facts* are nowadays available, but all MD design methods rely on discovering functional dependencies (FDs) to identify *dimensions*. However, an unbound FD search generates a combinatorial explosion and accordingly, these methods produce MD schemas with too many dimensions whose meaning has not been analyzed in advance. On the contrary, i) we use the available ontological knowledge to drive the FD search and avoid the combinatorial explosion and ii) only propose dimensions of interest for analysts by performing a statistical study of data.

1 Introduction

i) Our approach avoids generating too much results by mixing data mining and OLAP technologies. The purpose of this work is to demonstrate the feasibility and benefits of performing a statistical study of data to filter and prioritize the dimensional concepts found in the sources for a given fact, so that the designer can focus on these to decide and define his/her requirements for an OLAP application. ii) Furthermore, we tackle the usual assumption that a RDBMS is the most common kind of data sources we may find, by benefiting from a conceptual formalization of the domain (in our case, an OWL DL ontology) to avoid a combinatorial explosion of the statistical study.

Eventually, our approach identifies, for each fact, all the dimensional concepts and uses statistical evidences to filter out those of no relevance for data analysis.

2 Sketched Idea

Essentially, instead of blindly looking for FDs, our approach only tests combination of concepts likely to be interesting dimensional concepts for a given fact and its measures. We address the reader to [1] for further details on how to exploit the ontological knowledge available and the well-known FD theory to generate multi-concept FDs in a smart way.

Here, we extend our previous work with a statistical study to filter out those combinations of interest for the user. In [2] we can see how to perform an analysis of variance (ANOVA). This is a test designed to decide whether the difference in

J.B. Cushing, J. French, and S. Bowers (Eds.): SSDBM 2011, LNCS 6809, pp. 594–595, 2011.

the means of several samplings are due to differences in the populations or can be reasonably attributed to chance fluctuations alone. We propose to measure the importance when an attribute *partitions* a fact measure. Based on this objective evidence, we should choose the dimensional attributes based on the gain of entropy on partitioning each measure of interest. In our ANOVA tests, the hypothesis of "no difference" in the population of the different subsets is the *null hypothesis*. If this is rejected in our statistical analysis with a given confidence level, we will propose this attribute (or set of attributes) as an ***Interesting Dimension*** (**ID** from here on).

The* interesting Dimension *Function: This function is called whenever a combination of attributes is likely to be an ID (see [1]). Up to this step everything has been verified at the conceptual level. Then, we verify whether this combination of dimensional concepts is interesting to analyze a given measure by querying data. Prior to perform the statistical analysis, we first disregard candidate IDs with too many instances, since the end-user will be overwhelmed by the amount of values. Indeed, statisticians consider that useful categorical variables should have, at most, some tens of values. Relevantly, in case of querying a RDBMS, this pruning rule disregards combinations by just querying the catalog. Those combinations satisfying this rule are still candidates to be an ID, and we verify it by performing a one-way ANOVA test over data, as explained in [2], with the following query:

```
SELECT (SUM(gr.s)/(#distinct-1))/(SUM(POWER(A-grAvg,2))/(#tuples-#distinct)) AS fFisher
FROM t, (SELECT attrSet AS id, avg(A) AS grAvg, POWER(AVG(A)-(SELECT avg(A) FROM t),2) AS s
        FROM t WHERE joinConds GROUP BY attrSet) gr
WHERE attrSet=gr.id;
```

Where *attrSet* are the attributes conforming the *feasible ID* to be verified, *t* the table or tables containing those attributes (*join conditions* should be added if necessary), *#distinct* is the number of different values for *setAttr* and *#tuples* the number of tuples in the fact table. Then, the credibility of the null hypothesis is obtained by placing the result of this query in a Fisher distribution with the corresponding degrees of freedom.

However, this is not enough to decide whether this combination of attributes is an ID or not, because we could detect an ID due to the influence of another ID. Thus, once we detect an ID, we perform a two-way ANOVA test involving it and any other ID detected before (by means of a similar query) to discard the possibility that this is an ID just because another one is and there is some relationship between them (e.g., a multivalued dependency). Importantly, our approach can be used for other kind of data sources, as we would only need to adapt these SQL queries to the available data sources technology.

References

1. Abelló, A., Romero, O.: Using Ontologies to Discover Fact IDs. In: ACM 13th International Workshop on Data Warehousing and OLAP, pp. 3–10. ACM, New York (2010)
2. Wonnacott, T.H., Wonnacott, R.J.: Introductory Statistics. Wiley & Sons, Chichester (1990)