Applying OLAP Pre-Aggregation Techniques to Speed Up Query Processing in Raster-Image Databases

Angélica García Gutiérrez

School of Engineering and Science, Jacobs University Bremen, Germany
a.garciagutierrez@jacobs-university.de

Abstract. Aggregate functions are particularly useful when dealing with extremely large volumes of data. In business and statistical databases, aggregate queries have been leveraged by powerful methods such as On-Line Analytical Processing (OLAP). In contrast, current technology for raster image databases is lagging behind. A comparative study between raster image and business databases has shown similarities in their data structures and operations. This doctoral project investigates the application of OLAP pre-aggregation technology to speed up aggregate queries in raster-image databases.

1 INTRODUCTION

Remotely-sensed image data pose several challenges to existing database technology, particularly to ensure fast query performance and flexible analysis capabilities for data volumes in the order of terabytes or higher. As an example, consider the Landsat satellite program which consists approximately of 1.7 million raster-images (630 Terabytes). Typical operations on such datasets consist in extracting statistical information for a particular image channel (such as visual, infra-red or water-vapor) at a specific time and date. To speed up scaling operations, it is common to create reduced resolution copies of the original image, so called image pyramids. Further operations rely on the use of aggregate functions to reduce a multidimensional dataset into a scalar quantity.

In a different domain, mainly in business applications, similar data structures and operations are leveraged through mechanisms such as OLAP pre-aggregation, which have proved to speed up operations by factors up to 1000 times. Although the semantics of the data managed in OLAP applications are different, some characteristics fit well with the typical nature of raster-image datasets and this can be exploited.

We claim that despite the differences, there are many features that we can borrow from OLAP to build a pre-aggregation scheme for raster image databases.

Florian Probst and Carsten Keßler (Eds.): GI-Days 2007 - Young Researchers Forum.
IfGIprints 30. ISBN: 978-3-936616-48-4

The remainder of this document is organized as follows. Section 2 presents a brief overview of OLAP pre-aggregation. Section 3 presents an overview of preliminary results and current activities. Finally, a brief summary is presented in Section 4.

2 OLAP PRE-AGGREGATION

OLAP pre-aggregation refers to the process of pre-computing and storing query results that are often submitted in the database. Approaches to pre-aggregation affect both the size of the database and the response time of queries. If more values are pre-computed, a user is more likely to request a value that has already been calculated, thus the response time will be faster because the value does not need to be computed but only retrieved. However, materializing all possible combinations of aggregates is infeasible, as it typically causes a blowup in storage requirements up to 500 times the size of the base data. Thus, selecting the set of aggregates to materialize (pre-aggregate) is a trade off between speeding up query responses and minimizing the time required to keep the materializations updated.

Typical algorithms utilize cost functions in order to identify the optimal set of views to materialize. The estimation of cost functions consider the frequency in which an aggregate query is submitted to the database as well as the frequency in which the base data for answering such a query is updated (Gupta *et al.*, 1995; Shukla *et al.*, 1998). Recentlty, approaches considering user-access patterns have been proposed (Ramachandran *et al.*, 2005).

3 PRELIMINARY RESULTS AND CURRENT WORK

We have conducted a comparative study between raster-image datasets and the data structures in OLAP. The study not only showed us some striking similarities but also allowed us to identify an essential concept for pre-aggreation in OLAP: hierarchies. A hierarchy determines the level of detail in which the data can be aggregated and it is necessary to determine if a query, or part of the query, can be answered from pre-aggregated results. In contrast, aggregating data in 2-D raster-image datasets is not determined by any hierarchical structure but instead, the user can freely request an aggregate operation in the entire raster or in a specific region of it. A special case is the 3-D raster-image time-series datasets where a hierarchy can be derived for the time-dimension e.g., year-month-day. Clearly, a further formalization of the concept is necessary and this is one of the challenges in our project. A preliminary idea is to build parent-son hierarchical structures where a parent-node represents an aggregate operation and children-nodes are operations that can benefit from such an aggregate. Hence, it becomes essential to know the fundamental operations that are performed on raster-image datasets as well as a mathematical characterization for their computation. To this end, a set of fundamental operations in GIS was modeled using the algebraic framework introduced in (Baumann, 1999). Such characterization allowed us to identify operations that require data summarization (aggregation) and therefore, potential candidates to be considered for pre-aggregation (Garcia-Gutierrez *et al*., 2007).

Current activities include the adaptation of the first pre-aggregation algorithm to support 2-D datasets e.g., satellite images and digital elevation models. The prototype will be available to the GIS community in Jacobs University for purposes of evaluation and feedback.

4 Summary

We are investigating the application of OLAP pre-aggregation techniques in raster-image databases. Specifically, the emphasis is set on satellite imagery data used in GIS applications. A real-life implementation of the adapted techniques will show the impact of our results.

References

Baumann P. A database array algebra for spatio-temporal data and beyond. NGITS'99 LNCS 1649 pp.76-93, 1999.

Garcia-Gutierrez A., and Baumann P., et al. Modeling fundamental geo-raster operations with array algebra, technical report. Jacobs University Bremen, Germany, July 2007.

Gupta A., Harinarayan V., and Quass D. Aggregate-query processing in data-warehousing environments. In 21st VLDB Conference Zurich, Switzerland., 1995.

Shukla A., Naughton K., and Deshpade P. Materialized view selection for multidimensional datasets. In 24th VLDB Conf. pp.488-499, 1998.

Ramachandran K., Shah B., and Raghavan V. Dynamic pre-fetching of views based on user-access patterns in an olap system. In ACM SIGMOD, 2005.