OLAP Visualization: Models, Issues, and Techniques

Alfredo Cuzzocrea
University of Calabria, Italy

Svetlana Mansmann
University of Konstanz, Germany

INTRODUCTION

The problem of efficiently *visualizing multidimensional data sets* produced by scientific and statistical tasks/processes is becoming increasingly challenging, and is attracting the attention of a wide multidisciplinary community of researchers and practitioners. Basically, this problem consists in visualizing multidimensional data sets by capturing the *dimensionality* of data, which is the most difficult aspect to be considered. Human analysts interacting with high-dimensional data often experience disorientation and cognitive overload. Analysis of high- dimensional data is a challenge encountered in a wide set of real-life applications such as (*i*) biological databases storing massive gene and protein data sets, (*ii*) real-time monitoring systems accumulating data sets produced by multiple, multi-rate streaming sources, (*iii*) advanced *Business Intelligence* (BI) systems collecting business data for decision making purposes etc.

Traditional DBMS front-end tools, which are usually tuple-bag-oriented, are completely inadequate to fulfill the requirements posed by an interactive exploration of high-dimensional data sets due to two major reasons: (*i*) DBMS implement the OLTP paradigm, which is optimized for transaction processing and deliberately neglects the dimensionality of data; (*ii*) DBMS operators are very poor and offer nothing beyond the capability of conventional SQL statements, what makes such tools very inefficient with respect to the goal of visualizing and, above all, interacting with multidimensional data sets embedding a large number of dimensions.

Despite the above-highlighted practical relevance of the problem of visualizing multidimensional data sets, the literature in this field is rather scarce, due to the fact that, for many years, this problem has been of relevance for life science research communities only, and interaction of the latter with the computer science research community has been insufficient. Following the enormous growth of scientific disciplines like *Bio-Informatics*, this problem has then become a fundamental field in the computer science academic as well as industrial research. At the same time, a number of proposals dealing with the multidimensional data visualization problem appeared in literature, with the amenity of stimulating novel and exciting application fields such as the visualization of *Data Mining* results generated by challenging techniques like clustering and association rule discovery.

The above-mentioned issues are meant to facilitate understanding of the high relevance and attractiveness of the problem of visualizing multidimensional data sets at present and in the future, with challenging research findings accompanied by significant spin-offs in the *Information Technology* (IT) industrial field.

A possible solution to tackle this problem is represented by well-known OLAP techniques (Codd et al., 1993; Chaudhuri & Dayal, 1997; Gray et al., 1997), focused on obtaining very efficient representations of multidimensional data sets, called *data cubes*, thus leading to the research field which is known in literature under the terms *OLAP Visualization* and *Visual OLAP*, which, in the remaining part of the article, are used interchangeably.

Starting from these considerations, in this article we provide an overview of OLAP visualization techniques with a comparative analysis of their advantages and disadvantages. The outcome and the main contribution of this article are a comprehensive survey of the relevant state-of-the-art literature, and a specification of guidelines for future research in this field.

BACKGROUND

Formally, given a relational data source $\mathcal{R}$, a data cube $\mathcal{L}$ defined on top of R is a tuple $\mathcal{L} = \langle \mathcal{C},\mathcal{J},\mathcal{H},\mathcal{M} \rangle$, such that: (*i*) $\mathcal{C}$ is the data domain of $\mathcal{L}$ containing (OLAP) *data cells* storing *SQL aggregations*, such as those based on SUM, COUNT, AVG etc, computed over tuples in $\mathcal{R}$; (*ii*) $\mathcal{J}$ is the set of *functional attributes* (of $\mathcal{R}$) with respect to which $\mathcal{L}$ is defined, also called *dimensions* of $\mathcal{L}$; (*iii*) $\mathcal{H}$ is the set of *hierarchies* related to dimensions of $\mathcal{L}$; (*iv*) $\mathcal{M}$ is the set of *attributes of interest* (of $\mathcal{R}$) for the underlying OLAP analysis, also called *measures* of $\mathcal{L}$. OLAP data cubes can thus be used to effectively visualize multidimensional data sets and also support interactive exploration of such data sets using a wide set of operators (Han & Kamber, 2000), among which we recall: (*i*) *drill-down*, which descends in a dimension hierarchy of the cube by increasing the level of detail of the measure (and decreasing its level of abstraction); (*ii*) *roll-up*, which is a reverse of drill-down used to aggregate the measure to a coarser level of detail (and a finer level of abstraction); (*iii*) *pivot*, which rotates the dimensions of the cube, thus inducting data re-aggregation. Apart the visualization amenities, OLAP also offers very efficient solutions to the related problem of *representing multidimensional data sets* by means of a wide set of alternatives (Han & Kamber, 2000) according to which data cubes are stored in mass memory: (*i*) ROLAP (*Relational OLAP*), which makes use of the storage support provided by conventional RDBMS (i.e., relational tables); (*ii*) MOLAP (*Multidimensional OLAP*), which employs multidimensional arrays equipped with highly-efficient indexing data structures; (*iii*) HOLAP (*Hybrid OLAP*), which combines the two previous alternatives via storing portions of the cube on a relational support, and other portions on an array-oriented support (depending on various parameters such as the query-workload of the cube). Without further details, it is worth noticing that the efficiency of the data representation has a great impact on the effectiveness of data visualization and exploration activities.

Visual OLAP results from the convergence of BI techniques and the achievements in the scientific areas of *Information Visualization* and *Visual Analytics*. Traditional OLAP front-end tools, designed to support reporting and analysis routines primarily, use visualization merely for expressive presentation of the data. In the Visual OLAP approach, however, visualization plays the key role as the method of *interactive query-driven analysis*. A more comprehensive analysis of such a kind includes a variety of tasks such as: examining the data from multiple perspectives, extracting useful information, verifying hypotheses, recognizing trends, revealing patterns, gaining insights, and discovering new knowledge from arbitrarily large and/or complex volumes of multidimensional data. In addition to conventional operations of analytical processing, such as drill-down, roll-up, slice-and-dice, pivoting, and ranking, Visual OLAP supports further interactive data manipulation techniques, such as zooming and panning, filtering, brushing, collapsing etc.

OLAP VISUALIZATION: A SURVEY

First proposals on using visualization for exploring large data sets were not tailored towards OLAP applications, but addressed the generic problem of visual querying of large data sets stored in a database. Early experiences related to multidimensional data visualization can be found in real-life application scenarios, such as those proposed in (Gebhardt et al., 1997), where an intelligent visual interface to multidimensional databases is proposed, as well as in theoretical foundations, such as those stated in (Inselberg, 2001), which discusses and refines general guidelines on the problem of efficiently visualizing and interacting with high-dimensional data. Keim and Kriegel (1994) propose *VisDB*, a visualization system based on an innovative query paradigm. In *VisDB*, users are prompted to specify an initial query. Thereafter, guided by a visual feedback, they dynamically adjust the query, e.g. by using sliders for specifying range predicates on singleton or multiple attributes. Retrieved records are mapped to the pixels of the rectangular display area, colored according to their degree of relevance for the specified set of selection predicates, and positioned according to a grouping or ordering directive.

A traditional interface for analyzing OLAP data is a *pivot table*, or *cross-tab*, which is a multidimensional spreadsheet produced by specifying one or more measures of interest and selecting dimensions to serve as vertical (and, optionally, horizontal) axes for summarizing the measures. The power of this presentation technique comes from its ability in summarizing detailed data along various dimensions, and arranging aggregates computed at different granularity levels into a single