Chapter 2

Improving OLAP Analysis of Multidimensional Data Streams via Efficient Compression Techniques

Alfredo Cuzzocrea
ICAR-CNR, Italy and University of Calabria, Italy

Filippo Furfaro
University of Calabria, Italy

Elio Masciari
ICAR-CNR, Italy

Domenico Saccà
University of Calabria, Italy

ABSTRACT

Sensor networks represent a leading case of data stream sources coming from real-life application scenarios. Sensors are non-reactive elements which are used to monitor real-life phenomena, such as live weather conditions, network traffic etc. They are usually organized into networks where their readings are transmitted using low level protocols. A relevant problem in dealing with data streams consists in the fact that they are intrinsically multi-level and multidimensional in nature, so that they require to be analyzed by means of a multi-level and a multi-resolution (analysis) model accordingly, like OLAP, beyond traditional solutions provided by primitive SQL-based DBMS interfaces. Despite this, a significant issue in dealing with OLAP is represented by the so-called curse of dimensionality problem, which consists in the fact that, when the number of dimensions of the target data cube increases, multidimensional data cannot be accessed and queried efficiently, due to their enormous size. Starting from this practical evidence, several data cube compression techniques have been proposed during the last years, with alternate fortune. Briefly, the main idea of these techniques consists in computing compressed representations of input data cubes in order to evaluate time-consuming OLAP queries against them, thus

DOI: 10.4018/978-1-60566-328-9.ch002
obtaining approximate answers. Similarly to static data, approximate query answering techniques can be applied to streaming data, in order to improve OLAP analysis of such kind of data. Unfortunately, the data cube compression computational paradigm gets worse when OLAP aggregations are computed on top of a continuously flooding multidimensional data stream. In order to efficiently deal with the curse of dimensionality problem and achieve high efficiency in processing and querying multidimensional data streams, thus efficiently supporting OLAP analysis of such kind of data, in this chapter we propose novel compression techniques over data stream readings that are materialized for OLAP purposes. This allows us to tame the unbounded nature of streaming data, thus dealing with bounded memory issues exposed by conventional DBMS tools. Overall, in this chapter we introduce an innovative, complex technique for efficiently supporting OLAP analysis of multidimensional data streams.

INTRODUCTION

Data Stream Management Systems (DSMS) have captured the attention of large communities of both academic and industrial researchers. Data streams pose novel and previously-unrecognized research challenges due to the fact that traditional DBMS (Henzinger, Raghavan & Rajagopalan, 1998, Cortes, Fisher, Pregibon, Rogers & Smith, 2000), which are based on an exact and detailed representation of information, are not suitable in this context, as the whole information carried by streaming data cannot be stored within a bounded storage space (Babcock, Babu, Datar, Motwani & Widom, 2002). From this practical evidence, a plethora of recent research initiatives have been focused on the problem of efficiently representing, querying and mining data streams (Babu & Widom, 2001, Yao & Gehrke, 2003, Acharya, Gibbons, Poosala, & Ramaswamy, 1999, Avnur & Hellerstein, 2000).

Sensor networks (Bonnet, Gehrke & Seshadri, 2000, Bonnet, Gehrke & Seshadri, 2001) represent a leading case of data stream sources coming from real-life application scenarios. Sensors are non-reactive elements which are used to monitor real-life phenomena, such as live weather conditions, network traffic etc. They are usually organized into networks where their readings are transmitted using low level protocols (Gehrke & Madden, 2004, Madden & Franklin, 2002, Madden, Franklin, & Hellerstein, 2002, Madden, Szewczyk, Franklin & Culler, 2002). Under a broader vision, sensor networks represent a non-traditional source of information, as readings generated by sensors flow continuously, leading to an infinite, memory-unbounded stream of data.

A relevant problem in dealing with data streams consists in the fact that they are intrinsically multi-level and multidimensional in nature (Cai, Clutterx, Papex, Han, Welgex & Auvilx, 2004; Han, Chen, Dong, Pei, Wah, Wang & Cai, 2005), hence they require to be analyzed by means of a multi-level and a multi-resolution (analysis) model accordingly. Furthermore, it is a matter of fact to note that enormous data flows generated by a collection of stream sources like sensors naturally require to be processed by means of advanced analysis/mining models, beyond traditional solutions provided by primitive SQL-based DBMS interfaces. Consider, for instance, the application scenario drawn by a Supply Chain Management System (SCMS) (Gonzalez, Han, Li & Klabjan, 2006), which can be intended as a sort of sensor network distributed over a wide geographical area. Here, due to the characteristics of the particular application domain, data embedded in streams generated by supply providers (i.e., the sensors, in this case) are intrinsically multi-dimensional, and, in addition to this, correlated in nature. In more detail, multidimensionality of data is dictated by the fact that, in a typical sup-