OLAP over Uncertain and Imprecise Data Streams

Alfredo Cuzzocrea

ICAR-CNR, Italy & University of Calabria, Italy

INTRODUCTION

A critical issue in representing, querying and mining data streams consists of the fact that they are intrinsically multi-level and multidimensional in nature (Cai et al., 2004; Han et al., 2005), hence they require to be analyzed by means of multi-level and multi-resolution (analysis) models accordingly. Furthermore, it is a matter of fact to note that enormous data flows generated by a collection of stream sources *naturally* require to be processed by means of advanced analysis/ mining models, beyond traditional solutions provided by primitive SQL-based DBMS interfaces, and very often high-performance computational infrastructures, like Data Grids, are advocated to provide the necessary support to this end (e.g., (Cuzzocrea et al., 2004a; Cuzzocrea et al., 2004b; Cuzzocrea et al., 2005)), also exploiting fortunate data compression paradigms (e.g., (Cuzzocrea, 2005; Cuzzocrea, 2006a; Cuzzocrea, 2006b; Cuzzocrea and Wang, 2007; Cuzzocrea et al., 2007; Cuzzocrea et al., 2009b; Cuzzocrea & Serafino, 2009)) or data fragmentation paradigms (e.g., (Bonifati & Cuzzocrea, 2007)). Conventional analysis/mining tools (e.g., DBMS-inspired) cannot carefully take into consideration these kinds of multidimensionality and correlation of real-life data streams, as stated in (Cai et al., 2004; Han et al., 2005). From this, it follows that, if one tries to process multidimensional and correlated data streams by means of such tools, rough errors are obtained in practice, thus seriously affecting the quality of decision making processes that found on analytical results mined from streaming data.

Modern data stream applications and systems are also more and more characterized by the presence of *uncertainty* and *imprecision* that make the problem of dealing with *uncertain and imprecise data streams* a leading research challenge. This issue has recently attracted a great deal of attention from both the academic and industrial research community, as confirmed by several research efforts done in this context (Cormode & Garofalakis, 2007; Jayram et al., 2007; Aggarwal & Yu, 2008; Cormode et al., 2008; Jin et al., 2008; Zhang et al., 2008; Etuk et al., 2013).

Uncertain and imprecise data streams arise in a plethora of actual application scenarios ranging from *environmental sensor networks* to *logistic* networks and telecommunication systems, and so forth. Consider, for instance, the simplest case of a sensor network monitoring the temperature Tof a given geographic area W. Here, being Tmonitoring a natural, real-life measure, it is likely to retrieve an *estimate* of T, denoted by \tilde{T} , with a given confidence interval, denoted by [$\tilde{T}_{_{min}}$, $\tilde{T}_{_{max}}$], such that $\tilde{T}_{_{min}}$ < $\tilde{T}_{_{max}}$, having a certain probability p_r , such that $0 \le p_r \le 1$, rather than to obtain the *exact value* of T, denoted by \tilde{T} . The semantics of this confidence-intervalbased model states that the (estimated) value of T, \tilde{T} , ranges between $\tilde{T}_{_{min}}$ and $\tilde{T}_{_{max}}$ with probability p_{τ} . Also, a law describing the *probability* distribution according to which possible values of T vary over the interval [\tilde{T}_{min} , \tilde{T}_{max}] is assumed. Without loss of generality, the uniform distribution is very often taken as reference. The uniform distribution states that (possible) values in [T_{min} ,

 \tilde{T}_{max}], have *all* the same probability to be the exact value of *T*, \tilde{T} , effectively. Despite the popularity of the normal distribution, the confidence-interval-based model above is prone to incorporate any other kind of probability distribution (Papoulis, 1994).

Contrary to conventional tools, multidimensional analysis provided by On-Line Analytical Processing (OLAP) technology (Gray et al., 1997; Chaudhuri & Dayal, 1997), which has already reached an high-level of maturity, allows us to efficiently exploit and take advantages from multidimensionality and correlation of data streams, with the final aim of improving the quality of both analysis/mining tasks and decision making in streaming environments (e.g., (Cuzzocrea, 2009; Cuzzocrea & Chakravarthy, 2010)). OLAP allows us to aggregate data according to (1) a fixed logical schema that can be a star or a snowflake schema (Han & Kamber, 2000), and (2) a given SQL aggregate operator, such as SUM, COUNT, AVG etc. The resulting data structures, called data cubes (Gray et al., 1997), which are usually materialized within multidimensional arrays, allow us to meaningfully take advantages from the amenity of querying and mining data according to a multidimensional and a multi-resolution vision of the target domain, and from the rich availability of a wide set of OLAP operators (Han & Kamber, 2000), such as roll-up, drill-down, slice-&-dice, *pivot* etc, and OLAP queries, such as *range*-, *top-k*, and *iceberg* queries.

Therefore, the idea of analyzing massive exact and uncertain multidimensional data streams by means of OLAP technology makes sense perfectly, and puts the foundations for novel models and computational paradigms that can be used to efficiently extract summarized, OLAP-like knowledge from data streams, thus overcoming limitations of conventional DBMS-inherited analysis/mining tools.

By meaningfully designing the underlying OLAP (logical) model in dependence on the specific application domain and analysis goals, multidimensional models can efficiently provide support to intelligent tools for a wide set of real-life data stream application scenarios such as weather monitoring systems, environment monitoring systems, systems for controlling telecommunication networks, network traffic monitoring systems, alerting/alarming systems in time-critical applications, sensor network data analysis tools etc, according to similar insights achieved in (Cai et al., 2004; Han et al., 2005). Indeed, multidimensionality of data streams puts the basis for an extremely variegated collection of stream DW and KD tools with powerful capabilities, even beyond those of conventional ML and DM tools running on transactional data, such as multidimensional analysis (e.g., (Han et al., 2005)), correlation analysis (e.g., (Gehrke et al., 2001)), regression analysis (e.g., (Chen et al., 2002)), and so forth. The resulting representation/ analysis model constitutes what we call as OLAP stream model, which can be reasonably intended as a novel model for processing multidimensional data streams, and supporting multidimensional and multi-resolution analysis and mining tasks over exact and uncertain data streams.

Following these considerations, in this chapter we present a state-of-the-art technique for supporting OLAP over uncertain and imprecise multidimensional data streams, and provide research perspectives for future efforts in this scientific field. The remaining part of the chapter is organized as follows. First, we present a state-of-the-art technique for OLAPing uncertain and imprecise multidimensional data streams (Cuzzocrea, 2011a; Cuzzocrea, 2011b; Cuzzocrea, 2013). Then, we provide and discuss future research perspectives in the investigated topics. Finally, we summarize the contributions of this chapter and proposes future work in this scientific field.

BACKGROUND

As highlighted throughout the chapter, while there is considerable amount of research work on the problem of effectively and efficiently executing 0

8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/olap-over-uncertain-and-imprecise-datastreams/107357

Related Content

Business Analytics in Sport Talent Acquisition: Methods, Experiences, and Open Research Opportunities

Rocio de la Torre, Laura O. Calvet, David Lopez-Lopez, Angel A. Juanand Sara Hatami (2022). *International Journal of Business Analytics (pp. 1-20).*

www.irma-international.org/article/business-analytics-in-sport-talent-acquisition/290406

Organizational Capability Readiness Towards Business Intelligence Implementation

Md Shaheb Aliand Shahadat Khan (2019). International Journal of Business Intelligence Research (pp. 42-58).

www.irma-international.org/article/organizational-capability-readiness-towards-business-intelligenceimplementation/219342

Closing the Gap between Data Mining and Business Users of Business Intelligence Systems: A Design Science Approach

Ana Azevedoand Manuel Filipe Santos (2012). International Journal of Business Intelligence Research (pp. 14-53).

www.irma-international.org/article/closing-gap-between-data-mining/74733

Simulation Optimization for Finite Parameter Space

Banu Y. Ekren, Gerald W. Evans, William E. Bilesand Sunderesh S. Heragu (2014). *Encyclopedia of Business Analytics and Optimization (pp. 2171-2177).* www.irma-international.org/chapter/simulation-optimization-for-finite-parameter-space/107403

ROC Analysis in Business Decision-makings

Nan Huand Xiao-Hua Zhou (2014). *Encyclopedia of Business Analytics and Optimization (pp. 2104-2114).* www.irma-international.org/chapter/roc-analysis-in-business-decision-makings/107397