OLAP Over Probabilistic Data

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

INTRODUCTION

Probabilistic data (e.g., (Barbarà et al., 1992; Cheng et al., 2003; Dalvi & Suciu, 2004; Dalvi & Suciu, 2007; Ré & Suciu, 2008; Benjelloun et al., 2009; Agrawal et al., 2006; Sarma et al., 2008)) are becoming one of the most attracting kinds of data for database researchers, due to the fact such a format/formalism perfectly captures two novel, interesting classes of datasets that very often occur in modern database application scenarios, namely *uncertain* and *imprecise data* (e.g., (Ge et al., 2013)). Uncertain and imprecise data are indeed very popular, as uncertainty and imprecision affect the same processes devoted to collect data from input data sources and make use of these data in order to populate the target database (e.g., (Balcan et al., 2013)). Consider, for instance, the simplest case represented by a *sensory database* (Bonnet et al., 2001) populated by a sensor network monitoring the temperature *T* of a given geographic area *S*. Here, being *T* monitoring a natural, real-life measure, it is likely to retrieve an uncertain and imprecise *estimate* of *T*, denoted by $\tilde{T}$, with a given *confidence interval* (Papoulis, 1994), denoted by $\left[\tilde{T}_{min}, \tilde{T}_{max}\right]$, such that $\tilde{T}_{min} < \tilde{T}_{max}$, having a certain probability p_T, such that $0 \leq p_T \leq 1$, rather than to obtain the *exact value* of *T*, denoted by $\hat{T}$. The semantics of this confidence-interval-based model states that the (estimated) value of *T*, $\tilde{T}$, ranges between $\tilde{T}_{min}$ and $\tilde{T}_{max}$ with probability p_T. In popular *probabilistic database models* (e.g., (Benjelloun et al., 2009; Agrawal et al., 2006; Sarma et al., 2008)), confidence intervals and related probabilities are directly embedded into the *probabilistic tables* directly, thus originating *probabilistic attributes* storing *probabilistic* (*attribute*) *values*, which compose *probabilistic tuples*. Physical reasons of uncertain and imprecision of data are many-fold, and they can be found in inherent randomness and incompleteness of data, sampling errors, human errors, instrument errors, data unavailability, delayed data updates, and so forth.

Multidimensional OLAP data cubes (Gray et al., 1997) are powerful tools allowing us to support rich and multi-perspective analysis over large amounts of data sets, based on a multi-dimensional and multi-resolution vision of data. Efficiently computing OLAP data cubes over the input dataset (e.g., relational databases) is a well-known research challenge that has been deeply investigated during last decades (e.g., (Harinarayan et al., 1996; Agarwal et al., 1996)), with alternate fortune. Since probabilistic datasets are becoming very popular, it is natural and reasonable to define and introduce the problem of *efficiently computing OLAP data cubes over probabilistic data*, which has been firstly proposed in (Burdick et al., 2005). While there is a wide and rich literature on the issue of *efficiently processing and querying probabilistic databases*, despite the relevance of OLAP applications for next-generation *Data Warehousing* (DW) *and Business Intelligence* (BI) *systems* very few papers address at now the yet-interesting problem of computing and querying OLAP data cubes over probabilistic data (e.g., (Burdick et al., 2005; Burdick et al., 2006; Burdick et al., 2007)). Contrary to this actual trend, it is natural to foresee that this problem will play more and more a leading role in the context of DW and BI systems, due to the obvious popularity of uncertain and imprecise

DOI: 10.4018/978-1-4666-5202-6.ch148

datasets (e.g., environmental sensor networks, data stream management systems, alarm and surveillance systems, RFID-based applications, supply-chain management systems), beyond to classical *performance problems* that have always been of great interest in these contexts (for instance, in challenge distributed application scenarios, e.g. (Cuzzocrea et al., 2004; Cuzzocrea et al., 2005; Bonifati & Cuzzocrea, 2007)).

Inspired by these motivations, in this chapter we provide a spectrum of research contributions focused on OLAP over uncertain and imprecise relational data, ranging from theoretical models to a critical analysis of state-of-the-art proposals and a discussion on novel research perspectives.

MODELING PROBABILISTIC DATABASES AND DATA CUBES

In this Section, we investigate the problem of *modeling probabilistic databases and data cubes*, as a fundamental issue of our research. The main result of this contribution consists in providing an innovative model of *probabilistic data cubes*.

Given a probabilistic database D_P modeled in terms of a collection of probabilistic relations R_i, i.e.

$$D_P = \left\{R_0, R_1, \ldots, R_{|D_P|-1}\right\},$$

the problem we investigate in this research consists in effectively and efficiently computing and querying a data cube over D_P, $C(D_P)$, given an input *data cube schema* (Vassiliadis & Sellis, 1999) W. According to the nature of D_P, we properly define $C(D_P)$ as a *probabilistic data cube* (we detail this novel definition next). Now, focus the attention on the class of probabilistic databases considered in our research, which is inspired from fundamental works in (Benjelloun et al., 2009; Agrawal et al., 2006; Sarma et al., 2008). Given a probabilistic relation R_i in D_P modeled in terms of a collection of attributes, i.e.

$$R_i = \left\{A_{i,0}, A_{i,1}, \ldots, A_{i,|R_i|-1}\right\},$$

such that A_{i,k_j}, with k_j in {0, 1, …, $|R_i| - 1$}, denotes an attribute in R_i, two distinct sub-set of attributes in R_i can be identified. The first one, denoted by $R_i^E \subset R_i$, such that

$$R_i^E = \left\{A_{i,k_0}^E, A_{i,k_1}^E, \ldots, A_{i,k_{|R_i^E|-1}}^E\right\},$$

with k_j in {0, 1, …, $|R_i| - 1$}, stores the sub-set of *exact attributes* in R_i, i.e. attributes in R_i whose values are exact. The second one, denoted by $R_i^P \subset R_i$, such that

$$R_i^P = \left\{A_{i,k_0}^P, A_{i,k_1}^P, \ldots, A_{i,k_{|R_i^P|-1}}^P\right\},$$

with k_j in {0, 1, …, $|R_i| - 1$} stores the sub-set of probabilistic attributes in R_i, i.e. attributes in R_i whose values are probabilistic. Obviously,

$$R_i^E \bigcap R_i^P = \varnothing .$$

An exact attribute A_{i,k_j}^E in R_i^E is defined as follows:

$$A_{i,k_j}^E = \left\{V_{i,k_j}^E \mid V_{i,k_j}^E \in \mathbb{D}_{i,k_j}^E\right\},$$

where V_{i,k_j}^E denotes an exact value of A_{i,k_j}^E and $\mathbb{D}_{i,k_j}^E$ the domain of A_{i,k_j}^E, respectively. A probabilistic attribute A_{i,k_j}^P in R_i^P is defined as follows:

$$A_{i,k_j}^P = \left\{\left[V_{i,k_{j,min}}^P, V_{i,k_{j,max}}^P\right], p_{i,k_j} \middle| V_{i,k_{j,min}}^P \in \mathbb{D}_{i,k_j}^P, V_{i,k_{j,max}}^P \in \mathbb{D}_{i,k_j}^P, V_{i,k_{j,min}}^P < V_{i,k_{j,max}}^P, 0 \le p_{i,k_j} \le 1\right\} \quad \text{(a)}$$