Privacy Preserving OLAP and OLAP Security

Alfredo Cuzzocrea
University of Calabria, Italy

Vincenzo Russo
University of Calabria, Italy

INTRODUCTION

The problem of ensuring the *privacy* and *security* of OLAP data cubes (Gray et al., 1997) arises in several fields ranging from advanced *Data Warehousing* (DW) and *Business Intelligence* (BI) systems to sophisticated *Data Mining* (DM) tools. In DW and BI systems, decision making analysts aim at avoiding that malicious users access perceptive ranges of multidimensional data in order to infer *sensitive knowledge*, or *attack* corporate data cubes via violating user rules, grants and revokes. In DM tools, domain experts aim at avoiding that malicious users infer *critical-for-the-task knowledge* from authoritative DM results such as frequent item sets, patterns and regularities, clusters, and discovered association rules. In more detail, the former application scenario (i.e., DW and BI systems) deals with both the privacy preservation and the security of data cubes, whereas the latter one (i.e., DM tools) deals with *privacy preserving OLAP issues* solely. With respect to security issues, although security aspects of information systems include a plethora of topics ranging from *cryptography* to *access control* and *secure digital signature*, in our work we particularly focus on *access control techniques* for data cubes, and remand the reader to the active literature for the other orthogonal matters.

Specifically, privacy preservation of data cubes refers to the problem of ensuring the privacy of data cube cells (and, in turn, that of queries defined over collections of data cube cells), i.e. hiding sensitive information and knowledge during data management activities, according to the general guidelines drawn by Sweeney in her seminar paper (Sweeney, 2002), whereas access control issues refer to the problem of ensuring the security of data cube cells, i.e. restricting the access of unauthorized users to specific sub-domains of the target data cube, according to well-known concepts studied and assessed in the context of DBMS security. Nonetheless, it is quite straightforward foreseeing that these two even distinct aspects should be meaningfully *integrated* in order to ensure both the privacy and security of complex data cubes, i.e. data cubes built on top of complex data/knowledge bases.

During last years, these topics have became of great interest for the Data Warehousing and Databases research communities, due to their exciting theoretical challenges as well as their relevance and practical impact in modern real-life OLAP systems and applications. On a more conceptual plane, theoretical aspects are mainly devoted to study how *probability* and *statistics schemes* as well as rule-based models can be applied in order to efficiently solve the above-introduced problems. On a more practical plane, researchers and practitioners aim at integrating convenient privacy preserving and security solutions within the core layers of commercial OLAP server platforms.

Basically, to tackle deriving privacy preservation challenges in OLAP, researchers have proposed models and algorithms that can be roughly classified within two main classes: *restriction-based techniques*, and *data perturbation techniques*. First ones propose limiting the number of query kinds that can be posed against the target OLAP server. Second ones propose perturbing data cells by means of random noise at various levels, ranging from schemas to queries. On the other hand, access control solutions in OLAP are mainly inspired by the wide literature developed in the context of controlling accesses to DBMS, and try to adapt such schemes in order to control accesses to OLAP systems.

Starting from these considerations, in this article we propose a survey of models, issues and techniques in a broad context encompassing privacy preserving and security aspects of OLAP data cubes.

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

BACKGROUND

Handling sensitive data, which falls in privacy preserving issues, is common in many real-life application scenarios. For instance, consider a government agency that collects information about client applications/users for a specific *e*-government process/task, and then makes this information available for a third-party agency willing to perform market analysis for business purposes. In this case, preserving sensitive data of client applications/users and protecting their utilization from malicious behaviors play a leading role. It should be taken into account that this scenario gets worse in OLAP systems, as the interactive nature of such systems *naturally* encourages malicious users to retrieve *sensitive knowledge* by means of *inference techniques* (Wang et al., 2004a; Wang et al., 2004b) that, thanks to the wide availability of OLAP tools and operators (Han & Kamber, 2000), can reach an high degree of effectiveness and efficiency.

Theoretical background of privacy preserving issues in OLAP relies on research experiences in the context of *statistical databases* (Shoshani, 1997), where these issues have been firstly studied. In statistical databases, this problem has been tackled by means of *Statistical Disclosure Control* (SDC) techniques (Domingo-Ferrer, 2002), which propose achieving the privacy preservation of data via *trade-offing the accuracy and privacy of data*. The main idea of such an approach is that of admitting the need for data provisioning while, at the same time, the need for privacy of data. In fact, *full data hiding* or *full data camouflaging* are both useless, as well as publishing *completely-disclosed data sets*. Therefore, balancing accuracy and privacy of data is a reasonable solution to this challenge. In this context, two meaningful measures for evaluating the accuracy and privacy preservation capabilities of an arbitrary method/technique have been introduced. The first one is referred as *Information Loss* (IL). It allows us to estimate the lost of information (i.e., the accuracy decrease) due to a given privacy preserving method/technique. The second one is the *Disclosure Risk* (DR). It allows us to estimate the risk of disclosing sensitive data due to a given privacy preserving method/technique.

Duncan *et al.* (2001) introduce two metrics for probabilistically evaluating IL and DR. Given a numerical attribute A that can assume a value w with probability P_w, such that $\mathcal{D}(w)$ is the domain of w (i.e., the set of *all* the values that A can assume), a possible metrics of IL is given by the *Data Utility* (DU), which is defined as follows:

$$DU(w) = \frac{|\mathcal{D}(w)|}{\sum_{\mathrm{w}} P_w \cdot (w-1)^2} \tag{1}$$

where $|\mathcal{D}(w)|$ denotes the cardinality of $\mathcal{D}(w)$. It should be noted that DU and IL are inversely proportional, i.e. the more is IL the less is DU, and, conversely, the less is IL the more is DU.

Different formulations exist. For instance, Sung *et al.* (2006) introduce the so-called *accuracy factor* $F_{a,Q}$ of a given query Q against a data cube D, i.e. the relative accuracy decrease of the *approximate answer* to Q, denoted by $\tilde{A}(Q)$, which is evaluated on the *synopsis data cube* $\tilde{D}$ obtained from D by means of *perturbation-based techniques* (presented next), with respect to the *exact answer* to Q, denoted by $A(Q)$, which is evaluated on the original data cube D. $F_{a,Q}$ is defined as follows:

$$F_{a,Q} = 2^{-\left|\frac{\tilde{A}(Q)-A(Q)}{A(Q)}\right|} \tag{2}$$

With regards to DR, Duncan *et al.* (2001) consider the *probability with respect to the malicious user* that A can assume the value w, denoted by P_w^U, and introduce the following metrics that models DR in terms of the reciprocal of the *information entropy*, as follows:

$$DR(w) = \frac{1}{-\sum_{\mathrm{w}} P_w^U \cdot \log(P_w^U)} \tag{3}$$

Indeed, being impossible to estimate the value of P_w^U, as one should know *all* the information/knowledge held by the malicious user, in (Duncan et al., 2001) the *conditional* version of (3) is proposed as follows:

$$DR(w) = \frac{1}{-\sum_{\mathrm{w}} p(w|u) \cdot \log_2 p(w|u)} \tag{4}$$

such that $p(w|u)$ denotes the *conditional probability* that the actual value of A is w while the value known by the malicious user is u.

Just like for IL, different formulations for measuring DR exist. Sung *et al.* (2006) introduce the so-called *privacy factor* $F_{p,D}$ of a given data cube D with respect to the corresponding perturbation-based synopsis data cube $\tilde{D}$. $F_{p,D}$ is defined as follows:

5 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/privacy-preserving-olap-olap-security/11029

Related Content

A General Model for Data Warehouses
Michel Schneider (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 913-919).*
www.irma-international.org/chapter/general-model-data-warehouses/10929

Reasoning about Frequent Patterns with Negation
Marzena Kryszkiewicz (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1667-1674).*
www.irma-international.org/chapter/reasoning-frequent-patterns-negation/11042

Reflecting Reporting Problems and Data Warehousing
Juha Kontio (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1682-1688).*
www.irma-international.org/chapter/reflecting-reporting-problems-data-warehousing/11044

Real-Time Face Detection and Classification for ICCTV
Brian C. Lovell, Shaokang Chenand Ting Shan (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1659-1666).*
www.irma-international.org/chapter/real-time-face-detection-classification/11041

Can Everyone Code?: Preparing Teachers to Teach Computer Languages as a Literacy
Laquana Cooke, Jordan Schugar, Heather Schugar, Christian Pennyand Hayley Bruning (2020). *Participatory Literacy Practices for P-12 Classrooms in the Digital Age (pp. 163-183).*
www.irma-international.org/chapter/can-everyone-code/237420