

Privacy Preservation for Data Cubes

Sam Sung, Yao Liu
Department of Computer Science
National University of Singapore
Kent Ridge, Singapore 117543
ssung@comp.nus.edu.sg

Peter Ng
Department of Computer Science
University of Texas
Pan American, Edinburg TX 78539
ngp@cs.panam.edu

1. Introduction

Data privacy refers to the issue of how to preserve the confidential information in individual data cells while still being able to provide an accurate estimation of the original summation values for range queries. There are three major goals in data privacy: 1) Security - the data must be protected from being revealed; 2) Accuracy - results of analysis are valuable in a business's decision making and 3) Accessibility - data should be as easy to be accessed as possible.

However, the open nature of data warehouses creates a security conflict. Another conflict is between security and accuracy of the query results. With the increase in the number of data warehouses and OLAP users, the misuse of data warehouse is steadily growing, which has led to the needs for proper techniques to support all three goals of data privacy.

We propose a novel and simple but effective solution to fulfill all three goals. Under our scheme it is almost impossible to estimate the original data. The accuracy of range sum queries is close to 100% particularly in relation to large queries. Our scheme has no restriction on users' access of data.

2. Distortion based data privacy preserving methods

Our method is called *zero-sum* method which is a distortion based method. First, we start with an initial distortion in each cell (also called original distortion). The initially distorted data is then "adjusted" so that all the *marginal sums* of each block are zeroes. These "adjusted" distortions are the final distortions.

There are many ways to enforce these adjustments. The following is one of them: given a 2-dimensional block $m \times n$, for each row, redistribute the negative sum of the row back to each cell in the row such that the new sum becomes zero. Suppose $S_{i,*}$ denotes the distortion sum of i th row, $\frac{-S_{i,*}}{n}$ will be added to each cell of the row to make $S_{i,*}$

equal to zero. After the row adjustments are done, we repeat the same process for column adjustments. Finally each cell value of the 2-dimensional block is different from its original one but all marginal sums are remained the same.

Theorem A block of k -dimension can be converted to *zero-sum* form with k iterations.

3. Performance measurements

We use two performance measurements to evaluate the distortion method.

Privacy factor F_c : $F_c = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - x_i|}{|x_i|}$

The value F_c is the average amount of distortions over a block which indicates how closely the original value can be estimated. y_i is the distorted value, x_i is the original value and N is the number of cells.

Accuracy factor $F_{a,Q}$: $F_{a,Q} = 2^{-\left| \frac{answer - true_sum}{true_sum} \right|}$

The difference between the sum of distorted values (*answer*) and the original values (*true_sum*) over a query Q is referred to as the *accuracy loss* of Q .

4. Experimental results

The dataset is generated by the APB Benchmark program. There are four dimensions: customer, product, channel, and time. The size of each dimension is 900 (customer), 9000 (product), 9 (channel), 17 (time) respectively. The measure attribute is dollar, with range [0, 699]. We set the overall density of the Data Cube to be 20%.

Under the same conditions, the privacy of original distortions (simply add an initial distortion to each cell) is consistently better than those of adjusted distortions (obtained by *zero-sum*). However, the accuracy of adjusted distortions is consistently better than those of original distortions. With original distortion, the accuracy cannot be controlled - and in many cases, the accuracy is not acceptable. With *zero-sum* method, accuracy can be guaranteed. For example, if given a large distortion range, *zero-sum* will achieve acceptable privacy (about 50%) and high accuracy (over 98%).