ABSTRACT

When answering queries that ask for summary statistics, the query-system of a multidimensional database should guard confidential data, that is, it should avoid revealing (directly or indirectly) individual data, which could be exactly calculated or accurately estimated from the values of answered queries. In order to prevent the disclosure of confidential data, the query-system should be provided with an auditing procedure which, each time a new query is processed, checks that its answer does not allow a (knowledgeable) user to disclose any sensitive data. A promising approach consists in keeping track of (or auditing) answered queries by means a dynamic graphical data structure, here called the answer map, whose size increases with the number of answered queries and with the number of dimensions of the database, so that the problem of the existence of an efficient auditing procedure naturally arises. This chapter reviews recent results on this problem for “additive” queries (such as COUNT and SUM queries) by listing some polynomially solvable problems as well as some hard problems, and suggests directions for future work.
INTRODUCTION

The explosive increase in access to “statistical databases” has aroused concerns on the compromise of individual privacy. A statistical database (Denning, 1982; Date, 1983; Ullman, 1983; Michalewicz, 1991) is a database which contains information about individuals (companies, organizations, etc.) and whose users are only allowed to access summary statistics, but “sensitive” ones, that is, summary statistics that could lead to the exact or approximate disclosure of confidential data of single individuals. A naïve policy for guarding the confidentiality of individual data consists in leaving unanswered queries asking for sensitive summary statistics. This policy is not satisfactory because, given a set of non-sensitive summary statistics, there is a more-or-less large set of summary statistics that are implicitly released in that they can be computed in an exact or approximate way so that, if one of such summary statistics were sensitive, then the confidentiality of some individual data would be compromised. To guarantee privacy of individual records, a mechanism of inference control must be embodied in the statistical database interface according to some security policy, which can be said to be effective only if sensitive summary statistics are neither explicitly nor implicitly released.

We can imagine how an ideal control method should work. If \( Q \) is the set of previously answered queries and \( Q \) is a new query, then the value of \( Q \) is initially “locked.” Next, \( Q \) is tested for sensitivity; if the test is positive, the answer to \( Q \) will be denied; otherwise, the sensitivity test is applied to each of the summary statistics that would be implicitly revealed if \( Q \) were answered; if none of these summary statistics comes out to be sensitive, then and only then the value of \( Q \) is “unlocked” and \( Q \) is answered. Such a security procedure raises the computational problem of finding out the summary statistics that are implicitly revealed from the values of a given set of answered queries. This problem, sometimes called the “inference problem,” proves to be hard in the general case and has been solved in an efficient way only under certain restrictive assumptions. The complexity of the inference problem essentially depends on the query type and on the sensitivity criterion in use. The query type is defined by the following three parameters: the aggregation function (count, sum, average, max, min, etc.), the data type (real, nonnegative real, integer, nonnegative integer, etc.) and the data structure (simple if the answer to the query consists of a scalar, and complex if the answer consists of a structured data such as a table). Note that a complex query can be always thought of as a set of simple queries, and that average queries can be thought of being couples of queries. As to the sensitivity criteria, several proposals exist (Adam & Wortmann, 1989; Adam, Gangopadyay & Holowczak, 1999).

In this chapter, we review some recent results on the computational complexity of the inference problem for simple count and sum queries, which will be referred to as simple “additive queries,” with data of real and nonnegative real type.

The inference problem for additive queries with data of integer and nonnegative integer type is far more complicated, and some results can be found in Chin (1986)
Related Content

Matching Attributes across Overlapping Heterogeneous Data Sources Using Mutual Information
Huimin Zhao (2010). Journal of Database Management (pp. 91-110).
www.irma-international.org/article/matching-attributes-across-overlapping-heterogeneous/47421/

Regression Testing of Database Applications
www.irma-international.org/article/regression-testing-database-applications/3278/

A Link-Based Ranking Algorithm for Semantic Web Resources: A Class-Oriented Approach Independent of Link Direction
www.irma-international.org/article/link-based-ranking-algorithm-semantic/49721/

Towards an Ontology for Information Systems Development—A Contextual Approach
www.irma-international.org/chapter/towards-ontology-information-systems-development/7019/

The Expert’s Opinion
www.irma-international.org/article/expert-opinion/51077/