Joint Query Estimation from Multiple OLAP Databases

Elaheh Pourabbas*, Arie Shoshani+

* Istituto di Analisi dei Sistemi ed Informatica “Antonio Ruberti”-a-CNR, Viale Manzoni, 30 00185 Roma, Italy, e-mail: pourabbas@iasi.rm.cnr.it
+ Lawrence Berkeley National Laboratory, Mailstop 50B-3238 I Cyclotron Road, Berkeley, CA 94720, USA, e-mail: shoshani@lbl.gov

Abstract
Given an OLAP query expressed over a collection of source OLAP databases, we study the problem of computing the result OLAP target database when the underlying detail data of the source databases are not available. The problem arises when the source databases do not have the same category attributes, and it is not possible to derive the result from a single database. The method we use is the linear indirect estimator. We consider two obvious methods for computing such a target database, called the Full-cross-product (F) and the Pre-aggregation (P) methods. As for the accuracy and computational performance of these methods, while the method F provides the most accurate estimate possible, it is more expensive computationally than P. Our contribution is in proposing a third new method, called the Shortcut method (S), which is less expensive than F, but is as accurate.

1. Joint queries estimation
In this paper, we address the problem of responding to queries based only on the summary-level data of OLAP databases. That is, given that the base data is not available and that a query cannot be derived from a single summary database, we examine the process of estimating the desired result from multiple summary databases. This is typically done by a method of interpolation, called linear indirect estimator [1]. In this model, the population is supposed to be partitioned into large domains \(d\) formed by cross classification of demographic variables such as age, sex, race. Let \(i\) denotes a small area. For each of the domain \(d\), the variable of interest \(Y\) denoted by \(Y(d) = \sum Y_{i,d}\) is calculated from the survey data. It is assumed that the variable of interest and the auxiliary information are respectively available in the form of \(Y(d)\) and \(X_{i,d}\). A synthetic estimator of \(Y\) for \(i\) is defined by \(\hat{Y}(i) = \sum X_{i,d} Y(d)\). Formally, a joint query formulated on summary databases \(M_i(C_{j1},...,C_{jm})\), \(M_j(C_{j1},...,C_{jm})\) will be indicated by \(M^T(C^T,C^C,C^N)\) where \(M^T\) is a selected target summary database that can be one of \(M_i\) or \(M_j\); \(C^T\), \(C^C\), and \(C^N\) represent a set of respectively target, common, and non-common category attributes.

Definition: Let \(M_i(C^C_{m},C^N_{n})\) and \(M_j(C^C_{p},C^N_{q})\) be summary databases, and let \(\hat{M}_i(C^T_{j})\) be the target database. It can be calculated by the following methods:
(i) Full cross-product (F) it is obtained by summarizing all common and non-common category attributes in the full cross product summary database as follows:

\[
\hat{M}_i(C^T_{j}) = \sum_{i \in I} \left( M_i(C^C_{m},C^N_{n}) \right) \frac{\sum_{j \in J} M_j(C^C_{p},C^N_{q})}{\sum_{j \in J} \left( M_j(C^C_{p},C^N_{q}) \right)}
\]

(ii) Pre-aggregation (P) it is estimated by pre-summarizing all common and non-common category attributes in the summary databases then applying the linear estimator:

\[
\hat{M}_i(C^T_{j}) = M_i(\bullet) \left( \frac{\sum_{j \in J} M_j(C^T_{j})}{\sum_{j \in J} M_j(C^T_{j})} \right)
\]

(iii) Shortcut (S) it is obtained by pre-summarizing all the non-common category attributes in the summary databases and then computing the cross product and summarizing over the common attributes as follows:

\[
\hat{M}_i(C^T_{j}) = \sum_{j \in J} \left( M_i(C^C_{m},C^N_{n}) \right) \frac{\sum_{j \in J} M_j(C^T_{j})}{\sum_{j \in J} \left( M_j(C^T_{j}) \right)}
\]

In [2], we have shown that methods F and S yield the same results while F and P yield different results. The different results stem from the presence of common category attributes in the summary databases. In order to evaluate the accuracy of these methods, their average relative errors are calculated. It is stated that F (or S) provides the most accurate estimate w.r.t. P, but S is less expensive than F.

References