Synopses reconciliation via calibration in the τ-Synopses system

Yariv Matia, Yossi Matias, Leon Portman

School of Computer Science
Tel-Aviv University
matiayar@post.tau.ac.il, matias@cs.tau.ac.il, leon.portman@nice.com

Abstract. The τ-Synopses system was designed to provide a run-time environment for multiple synopses. We focus on its utilization for synopses management in a single server. In this case, a critical function of the synopses management module is that of *synopses reconciliation*: given some limited memory space resource, determine which synopses to build and how to allocate the space among those synopses. We have developed a novel approach of *synopses calibration* for an efficient computation of synopses error estimation. Consequently we can now perform the synopses reconciliation in a matter of minutes, rather than hours.

1 Introduction

Data synopses are concise representations of data sets, which enable effective processing of approximate queries to the data sets. Recent interest in approximate query processing and in effectively dealing with massive data sets resulted with a proliferation of new synopses.

The τ-Synopses system [6] was designed to provide a run-time environment for local and remote execution of various synopses. It provides the management functionality for registered synopses, and it enables easy registration of new synopses either locally or from remote SOAP-enabled platforms. The τ-Synopses system can serve as an effective research platform for experimental evaluation and comparison of different synopses, as well as a platform for studying the effective management of multiple synopses in a federated or centralized environment.

The system was previously presented in the context of *remote-synopses*, demonstrating how synopses can be managed in a distributed fashion [5]. We now focus our attention on the utilization of the τ-Synopses system for synopses management in a single server. In this case, a critical function of the Synopses Manager module is that of synopses reconciliation: given some limited memory space resource, determine which synopses to build and how to partition the available space among those synopses.

The problem of synopses reconciliation was previously studied and several algorithms were presented (e.g., [3, 2]). A basic operation in all reconciliation algorithms is that of estimating the accuracy of synopses implementations for given data sets. The common approach in obtaining such estimation is by invoking expensive queries into the original data. We present a novel approach, in

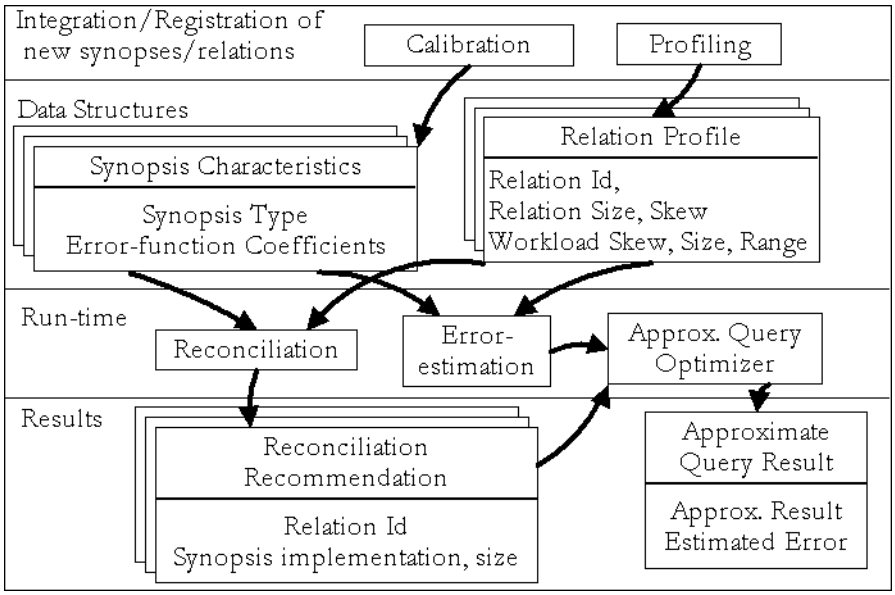

Fig. 1. Synopses Manager Architecture

which the reconciliation algorithm can consult with an error-estimation module, which can provide an error estimation without accessing the original data itself. Instead, the error-estimation module computes the synopsis approximation-error based on synopsis characteristics, which are computed by a *calibration* process in integration-time, and statistical data about the data sets, computed by a *profiling* process in registration-time. This results with an effective reconciliation process.

2 Synopsis Calibration and Synopses Reconciliation

The *Synopses Manager* module provides an automatic recommendation of which synopses to build and their sizes, based on the available synopses implementations, available memory space, and the registered relations and query workload.

As depicted in Figure 1 the Synopses Manager includes the following types of data structures: (i) *Relation Profile*, which includes statistical information about each registered relation; (ii) *Synopsis Characteristics*, which includes parameters computed at integration time for each synopsis; and (iii) *Reconciliation Recommendation*, which includes the recommendations of synopses to be built for each registered relation. It also includes the following modules: Calibration, Profiling, Error Estimation, Reconciliation, and Approximate Query Optimizer; these modules are described in more details below. A more detailed explanation of the calibration and reconciliation processes can be found in [4].

The synopsis error-estimation function. The crux of our approach is a novel calibration technique [4], which associates to every synopsis implementation T an error estimation function, EE_T.

We have found, through empirical testing, that the following function quite accurately describes the relative error of synopses implementations w.r.t. their corresponding relations and query workloads:

$$EE_T(L, Z, Q, R, S) = a_1 L^{b_1} + a_2 Z^{b_2} + a_3 Q^{b_3} + a_4 R^{b_4} + a_5 S^{b_5} + a_6 \ .$$

The arguments L, Z, Q, R and S, collectively denoted as the *Relation Profile*, are: the relation size (L), relation data distribution skew (Z), workload query skew (Q), workload query range (R), and the synopsis size (S). The a_i and b_i coefficients, collectively denoted as the *Synopsis Characteristics*, are unique to each synopsis implementation.

Calibration. This module is invoked every time a new synopsis implementation T is integrated into the system. The calibration process runs a small number of tests on the synopsis implementation, measuring its behavior under various synthetic relations and workloads. It then derives the a_i and b_i coefficients, of EE_T, using a combination of squared linear fitting and the CPLEX commercial solver [1]. The coefficients of the function are stored in the Synopsis Characteristics data structure

Error Estimation. This module is utilized by the Reconciliation and Approximate Query Optimizer modules. Given an approximate query to a relation with synopsis implementation T, the module computes the error estimation function EE_T based on the parameters available from the Relation Profile and Synopsis Characteristics data structures, resulting with the estimated approximation-error of the query.

Approximate Query Optimizer. This module has two functions: (1) triggers the building of the required synopses based on the recommendations received from the Reconciliation module; and (2) when a user submits an approximate query to the system, this module performs the query on the relevant synopsis, and also invokes the Error-Estimation module, returning both the estimated result, and the estimated approximation-error of the result.

Profiling. This module is invoked whenever a new relation or query workload are registered in the system. For relations, this module measures the cardinality of the relation (distinct count), and uses linear-squared-fitting to fit a Zipf parameter to the relation data distribution skew. For query workloads, the number and average range of the queries are calculated, and the Zipf parameter of the query distribution skew is again fitted using linear-squared-fitting. The computed statistical data is stored in the Relation Profile data structure.

Synopses reconciliation. Synopses reconciliation is basically an optimization problem – given available synopses implementations, a memory space limit, and a query workload, we would like to know the combination of synopses and their sizes that would yield the minimal error for the entire system.

The Synopses Reconciliation module can accommodate the implementations of any synopses reconciliation algorithm. The module currently has implementations of the algorithms from [3, 2], with the following modification. Whenever the error measurement of a synopsis utilization is required, it uses the Error-Estimation Module, instead of using the straight-forward measurement which involves executing costly queries into the database. The process is invoked on-demand by the administrator, and returns a recommended combination of synopses to build.

Utilizing the same reconciliation algorithms and heuristics as those in [3, 2], but replacing the action of measuring the error with a call to an error-estimation function, significantly reduces the run time of the reconciliation process while maintaining good accuracy.

3 System Demonstration

We demonstrate the calibration process for one of these synopses, showing the accuracy of the calculated EE_T function over different relations and workloads. We also demonstrate a full reconciliation process over a complex setup of relations and workloads, showing how a process that would normally take hours to complete, is completed in minutes, and compare its results to those of the optimal combination.

References

1. iLOG Inc. Ilog cplex 8.0 –user's manual, 2002.
2. H. V. Jagadish, H. Jin, B. C. Ooi, and K. L. Tan. Global optimization of histograms. In *SIGMOD '01*, pages 223–234, 2001.
3. A. C. Koenig and G. Weikum. A framework for the physical design problem for data synopses. In *Proceedings of the 8th International Conference on Extending Database Technology*, pages 627–645, 2002.
4. Y. Matia and Y. Matias. Efficient synopses reconciliation via calibration. Technical report, Tel Aviv University, 2005.
5. Y. Matias and L. Portman. τ-Synopses: a system for run-time management of remote synopses. In *International Conference on Data Engineering (ICDE), Software Demo*, pages 964–865, April 2004.
6. Y. Matias, L. Portman, and N. Drukh. The design and architecture of the τ-Synopses system. In *Proc. EDBT 06', Industrial and Application*, 2006.