τ-Synopses: A System for Run-Time Management of Remote Synopses

Yossi Matias and Leon Portman

School of Computer Science, Tel Aviv University
matias@cs.tau.ac.il, leonpo@cs.tau.ac.il

Abstract. τ-Synopses is a system designed to provide a run-time environment for remote execution of various synopses. It enables easy registration of new synopses from remote platforms, after which the system can manage these synopses, including triggering their construction, rebuild and update, and invoking them for approximate query processing. The system captures and analyzes query workloads, enabling its registered synopses to significantly boost their effectiveness (efficiency, accuracy, confidence), by exploiting workload information for synopses construction and update. The system can also serve as a research platform for experimental evaluation and comparison of different synopses.

Data synopses are concise representations of data sets, that enable effective processing of approximate queries to the data sets. Recent increased interest in approximate query processing and in effectively dealing with massive data sets resulted with a proliferation of new synopses addressing new problems as well as proposed alternatives to previously suggested synopses.

For both operational and research purposes, it would be advantageous to have a system that can accommodate *multiple synopses*, and have an easy way to integrate new synopses and manage them. The multiple synopses could be placed in remote locations for various reasons: they may be implemented on different types of platforms, they may be summarizing remote data whose transfer is undesirable or impossible due to performance or security constraints, and it would be beneficial to share the load of operating a large number of synopses using different systems for load balancing and redundancy reasons.

Motivated by the above, the τ-Synopses system was designed to provide a run-time environment for remote execution of various synopses. It enables easy registration of new synopses from remote SOAP-enabled platforms, after which the system can manage these synopses, including triggering their construction, rebuild and update, and invoking them for approximate query processing. The system captures and analyzes query workloads, enabling its registered synopses to significantly boost their effectiveness (efficiency, accuracy, confidence), by exploiting workload information for synopses construction and update. The system can serve as a research platform for experimental evaluation and comparison of different synopses.

The τ-Synopses system is independent, and can work with data sources such as existing relational or other database systems. It supports two types of users:

synopses providers who register their synopses within the system, and end-users who submit queries to the system. The system administrator defines the available data sources and provides general administration.

When a new synopsis is registered, the relevant data set and the supported queries are defined. A query submitted to the system is executed using the appropriate synopsis, based on the registration and other information. The result is returned to the user or optionally processed by other modules in the system. The system transforms updated data from its original datasource to be consistent with the format known to the synopses, so that synopses are not required to support any data transformation functionality or database connectivity logic. Any relational database or even real-time data providers can be data sources in the system.

Workload information is recorded by the system and becomes available to the registered workload-sensitive synopses.

The τ-Synopses system has the following key features:

- *multiple synopses*: The system can accommodate various types of synopses. New synopses can be added with their defined functionalities.
- *pluggable integration*: For integration purposes, a synopsis has to implement a simple interface, regardless of its internal implementation. By utilizing a light-weight host provided by the system, the synopsis can be executed on any SOAP-enabled platform.
- *remote execution*: Synopses can be transparently executed on remote machines, over TCP/IP or HTTP protocols, within local area networks or over the internet.
- *managed synopses*: The system allocates resources to synopses, triggers their construction and maintenance, selects appropriate synopses for execution, and provides all required data to the various synopses.
- *workload support*: Workload is captured, maintained and analyzed in a centralized location, and made available to the various synopses for construction and maintenance.
- *research platform*: The system provides a single, consistent source of data, training and test workload for experimental comparison and benchmarking, as well as performance measurements. It can therefore serve as an effective research platform for comparing different synopses without re-implementing them.

The core of the τ-Synopses system architecture features the following components, and depicted in Figure 1: Query Execution Engine, Synopses Manager, Updates Logger, and Workload Manager. In addition, the system includes a query-application which is used by end-users, an administration-application used by the administrator and by synopses-providers (not displayed), and a pool of registered synopses.

The Synopses Manager is used for registration and maintenance of the synopses. A new synopsis is added to the system by registering its parameters (including a list of supported queries and data sets) in the Synopses Manager Catalog.

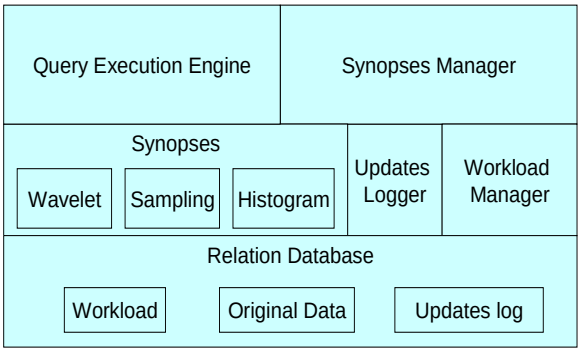

Fig. 1. Synopses Framework Architecture

The Query Execution Engine provides interface for receiving a query request from end-users and invoking the appropriate synopsis (or synopses), as determined by the Synopses Manager in order to process such query.

The Updates Logger provides all data updates to the registered synopses by intercepting data updates information in the data sources.

The Workload Manager captures, maintains and analyzes workload information for building, maintaining and testing synopses.

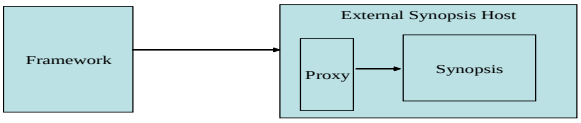

Fig. 2. Synopses Integration

Figure 2 depicts the integration process of a remote synopsis within the framework. The External Synopsis host is responsible for all communication with the system and is transparent to the synopsis. This design enables an unconstraint deployment. A remote synopsis can be integrated into the system by deploying or adapting such host into the remote system, and connecting the synopsis module locally into the host.

The system was tested by having groups of graduate and under-graduate students implement remote synopses as part of their projects, and have these synopses connect to the core system using the simple interfaces. The implemented state of the art synopses include different histograms, sketches, wavelet-synopses, etc.

We now encourage other research groups connect their synopses to the τ-Synopses system. This would allow access to a wide variety of different datasources and workloads, and a fair comparison with other synopses with little effort.