A Systematic Database Summary Generation Using the Distributed Query Discovery System

Tae W. Ryu¹ and Christoph F. Eick²

¹Department of Computer Science, California State University, Fullerton
Fullerton, CA 92834, USA
tryu@fullerton.edu

²Department of Computer Science, University of Houston, Houston
Houston, TX 77204, USA

Abstract. This paper introduces an approach to generate a database summary systematically using the distributed query discovery system, MASSON. Our approach is first to create an object-view and partition the database based on the object-view into clusters with similar properties, and then to generate the summary for each cluster. For this purpose, we propose a data set representation framework and introduce a proper similarity measure framework. The paper also describes the techniques used to generalize the generated primitive summary descriptions by MASSON and to improve the performance of the system using clustered computers and CORBA.

1 Introduction

For the current flood of data routinely generated by many organizations/Web sites, automatic tools and techniques have become necessary for analyzing the data. Recently many techniques, tools, and systems designed to discover interesting knowledge from the huge amount of data have been proposed/developed or are still under development. Some popularly used methods include classification, clustering, regression, data summarization, dependency modeling, link analysis, change and deviation detection, visualization, etc. [3].

This paper centers on the problem of database summary generation. Database summarization is useful in comprehending a database because it may provide a compact representation of the core properties of the database. It may also help us make useful inferences from a database. For example, a simple statement “Americans do not like high-fat food,” which can be a summary of a market database, allows a food company to infer advertising a new food product from the company as a low-fat product. However, generating a good summary for a database is a non-trivial problem because we need to deal with the structural data sets (e.g., typical relational or object-oriented database consists of many related relations or classes.) Several approaches to the database summary generation have been proposed. Lee and Kim [13] propose hypotheses refinement approach using fuzzy logic to summarize a database and produce a set of high-level abstract terms. An integrated system, EXPLORA [10],
searches interesting summary for a data set in the form of statement types using application-specific domain knowledge and statistical knowledge. Dhar and Tuzhilin [4] discuss user specified abstract function and aggregate principles to provide an abstract form of summary. Other approaches include generating rules for associative information between attributes [1] or for discriminating descriptions for the given data set [14].

Our approach is to partition the given database into clusters and generate the summary descriptions for each cluster. Fig. 1 shows the major steps involved in generating a database summary in this research. First, the data set to be analyzed is selected from the given database and is preprocessed. Second, the preprocessed data set is partitioned into a set of clusters with similar properties. Third, a summary is generated for each cluster. The generated summary can be further refined and generalized to produce the high-level summary descriptions for the database. For this purpose, we introduce a data set representation framework and a proper similarity measure framework for the object-based clustering. For the summary generation for each cluster, we use MASSON, a database query discovery system [18] using genetic programming (GP) [12]. In addition, we introduce the distributed MASSON developed based on clustered computers and CORBA for the performance improvement.

2 Clustering Structured Databases

The flat file is the simplest and most frequently used format in the traditional data analysis area. When using flat file format, data objects are represented through vectors in \( n \)-dimensional attribute space, each of which describes an object, and the object is characterized by \( n \) attributes, each of which has a single value. Most existing data analysis and data mining approaches assume that data sets to be analyzed are represented in a flat file format. Due to the fact that databases are more complex than flat files, database clustering faces following additional problems that do not exist when clustering flat files:

- **Support for object-view**: Databases contain objects that belong to different types; consequently, it has to support an object-view that consists of a subset of objects in the database. It is not reasonable to apply a clustering algorithm directly to the database without focusing on a view of the database. To illustrate this