Entropy Based Approximate Querying and Exploration of Datacubes

Themistoklis Palpanas *

themis@cs.toronto.edu

Dept. of Computer Science

University of Toronto

10 King's College Road, Toronto

Ontario, M5S 3G4, CANADA

Nick Koudas

koudas@research.att.com

AT&T Labs-Research

Department of Computer Science

University of Toronto

Technical Report CSRG-409

Abstract

Much research has been devoted to the efficient computation of relational aggregations and specifically the efficient execution of the datacube operation. In this paper we consider the inverse problem, that of deriving (approximately) the original data from the aggregates.

We motivate this problem in the context of two specific application areas, that of approximate query answering and data analysis. We propose a framework based on the notion of information entropy, that enables us to estimate the original values in a data set, given only aggregated information about it. We then show how approximate queries on the data from which the aggregates were derived can be performed using our framework. We also describe an alternate utility of the proposed framework, that enables us to identify values that deviate from the underlying data distribution, suitable for data mining purposes.

Finally, we present a detailed performance study of the algorithms using both real and synthetic data, highlighting the benefits of our approach as well as the efficiency of the proposed solutions.

*Contact author and address.