Using Categorical Clustering in Schema Discovery

Periklis Andritsos and **Renée J. Miller**
University of Toronto, Department of Computer Science
{periklis,miller}@cs.toronto.edu

ABSTRACT

Most techniques for managing relational schemas assume a given schema that adequately models the data [1]. However, we know that in practice, the semantics of the data may evolve over time and its schema (its table structures and constraints) is not always updated to reflect these changes [5]. Common examples include the overloading of tables to store facts of different types (for example, an order table originally designed to store service orders may be used to store various product orders as a company expands the scope of its business). Similarly, the semantics (typically represented as constraints) may evolve, perhaps because new data does not share the original semantics or because the full semantics were not captured in the original legacy design. Our long term research goal is to find techniques for discovering schemas that fit the data. In this work, we are taking an initial step toward this goal. Specifically, we are examining the benefits of using categorical clustering to discover groupings of tuples that share similar structural characteristics.

Our work is motivated by some recent studies that have used information theory to characterize good relational and nested relational (XML) schemas. Dalkilic and Robertson [4] have derived a measure for functional dependencies that quantifies the uncertainty left in a set of attributes, when another set of attributes is known. In addition, Arenas and Libkin [3] have introduced an information theoretic measure of well-designed database tables that takes into account the context in which each value appears. However, none of these methods suggest an efficient algorithm to find good decompositions or to suggest a good schema for a data set. In addition, this work considers dependencies that strictly hold in relations and so the techniques are not immediately applicable to dirty data. In real data, dependencies are often approximate, *i.e.*, there are a small number of tuples or values where given dependencies do not hold. We believe that such dependencies do exist in large legacy databases and that they can be used to help understand the semantics of a data set.

As a first step in discovering structure, we have developed an algorithm that groups similar records of relational tables containing categorical data. While clustering is a very well-studied problem, the techniques for clustering categorical data suffer from a number of limitations that make them unsuitable for use in a schema discovery application. For example, current categorical clustering algorithms do not scale to the large multi-attribute relational tables we consider. To address this and other limitations, we have introduced LIMBO, [2], a scalable hierarchical categorical clustering algorithm. LIMBO builds on the *Information Bottleneck (IB)* framework for quantifying the relevant information preserved when clustering [7]. We use it to cluster both tuples and attribute values. While the IB method has been applied before to cluster small data sets, [6], LIMBO is the first scalable hierarchical clustering algorithm to use this method. It supports a tradeoff between computation time and clustering quality. It handles large data sets efficiently and is robust over different input orders of. Finally, LIMBO produces clusterings for a range of k values (where k is the number of clusters). We take advantage of this feature to examine heuristics for selecting good clusterings within this range. This property is very important for schema discovery since we can pick a k value (among the many clusterings produced in a single application of the algorithm) for which we can achieve a good schema design.

Our objective is to produce *informative* clusters, *i.e.*, clusters that convey maximum information about their attribute values. That is, given a cluster, we wish to predict the attribute values associated with tuples of the cluster accurately. The quality measure of the clustering is then the mutual information of the clusters and the attribute values. Since a clustering is a summary of the data, some information is generally lost. Our objective will be to minimize this loss, or equivalently to minimize the increase in uncertainty as the tuples are grouped into fewer and larger clusters.

In this work, we will report on our experience using LIMBO to find clusterings with good structural characterics. That is, we plan to use LIMBO to find tuple clusters and evaluate their structural characteristics. The primary focus of our preliminary research will be to demonstrate the advantages (or disadvantages) of using the loss of information as a guide to discover relational structure.

References

[1] P. Andritsos, R. Fagin, A. Fuxman, L. M. Haas, M. Hernandez, C.-T. Ho, A. Kementsietsidis, R. J. Miller, F. Naumann, L. Popa, Y. Velegrakis, C. Vilarem, and L.-L. Yan. Schema Management. *Data Engineering Bulletin*, 25(3), Sept. 2002.

[2] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik. Limbo: A linear algorithm to cluster categorical data. Technical report, UofT, Dept of CS, CSRG-467, 2003.

[3] M. Arenas and L. Libkin. An Information-Theoretic Approach to Normal Forms for Relational and XML Data. In *PODS*, San Diego, CA, USA, 2003 (to appear).

[4] M. M. Dalkilic and E. L. Robertson. Information Dependencies. In *PODS*, Dallas, TX, USA, 2000.

[5] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining Database Structure; Or How to Build a Data Quality Browser. In *SIGMOD*, Madison, WI, USA, 2002.

[6] N. Slonim and N. Tishby. Document Clustering Using Word Clusters via the Information Bottleneck Method. In *SIGIR*, Athens, Greece, 2000.

[7] N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. In *37th Annual Allerton Conference on Communication, Control and Computing*, Urban-Champaign, IL, 1999.