Abstract

In Data mining project most of the time consuming task is to prepare a required data set for data mining analysis because in general the relational database has collection of tables and views that must be joined, aggregated and transformed in order to build the required data set. As result, most of the complex SQL queries are written multiple times independently from each other and in a disorganized manner. Therefore, the database grows with many tables and views that are not present as entities in the ER model. Similarly existing SQL aggregations having some limitations to prepare normalized data sets because they return only one column per aggregated group. To address this issue, we propose simple methods to generate SQL code to return aggregated columns in a horizontal tabular layout, where every row corresponds to an observation and every column is associated to a one variable. This new class of functions is called horizontal aggregations. Horizontal aggregations is extension of standard SQL aggregation for building data sets with a horizontal denormalized layout, which is input for most of the data mining algorithms. By providing these standard normalized data-set as an input to the Decision tree generation algorithm for generating Decision tree, similarly we can generate extended ER model.
References

- Carlos Ordonez, Sofian Maabout, David Sergio Matusevich, Wellington Cabrera, 2014 "Extending ER models to capture database transformations to build data sets for data mining", in Data and Knowledge Engineering.
- Carlos Ordonez and Zhibo Chen, 2012 "Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis", in IEEE Transaction on Knowledge and Data Engineering.
- Javier Garca-Garca, Carlos Ordonez, 2010 "Extended aggregations for databases with referential integrity issues", in Data and Knowledge Engineering.
- Carlos Ordonez, 2004 "Vertical and Horizontal Percentage Aggregations", in Proc. ACM SIGMOD Intl Conf. Management of Data (SIGMOD 04).
- Carlos Ordonez, 2004 "Horizontal Aggregations for Building Tabular Data Sets", in Proc. Ninth ACM SIGMOD Workshop Data Mining and Knowledge Discovery (DMKD 04).

Index Terms

Computer Science

Databases

Keywords

Data mining Transformation Aggregation Data preparation pivoting SQL.