Cleaning data with forbidden itemsets

Joeri Rammelaere, Floris Geerts and Bart Goethals
Universiteit Antwerpen, Belgium
{firstname.lastname}@uantwerpen.be

Methods for cleaning dirty data typically rely on additional information about the data, such as user-specified constraints that specify when a database is dirty. However, many real-world scenario's only have a dirty database available. In such a context, we propose the use of data mining methods to discover patterns that can be used to detect a certain class of potential inconsistencies.

Indeed, inconsistencies that arise from conflicting values within a single tuple can be described using itemsets. By mining infrequent, low-lift itemsets called *Forbidden itemsets*, we discover value combinations with a high chance of being conflicting. These forbidden itemsets are mined using the Eclat itemset mining algorithm with adaptations for improved efficiency. A database is then declared dirty if and only if it contains forbidden itemsets. Afterwards, we want to repair such a dirty database by modifying values. Since such modifications may introduce new forbidden itemsets, the challenge is to repair the data without introducing additional discoverable forbidden itemsets. This dynamic definition of dirtyness makes it more challenging to obtain a clean database after repairing. We address this issue by mining all itemsets which might become forbidden when using a naive repair algorithm, and exclude repairs that insert or remove occurences of these itemsets. Our repair algorithm is based on nearest neighbour imputation.

The error detection and repair algorithms are evaluated qualitatively on various datasets with categorical values. We show that Forbidden itemsets are able to discover certain errors that are not detectable with conditional functional dependencies, and that errors are typically discovered with high precision. Furthermore, the suggested repairs are of good quality and do not introduce new forbidden itemsets, as desired.