

Submission for the Dutch-Belgian Database Day (DBDBD) 2007

Name: Arne Koopman

Position/University: PhD student, Universiteit Utrecht

Supervisor: Arno Siebes

Title: Efficient Mining of Relational Data

Abstract

Relations can be found in many areas and applications. For example, we see relations in structured data such as sequences, trees and graphs when mining for patterns in web logs, chemical databases and various type of networks. More general, we see relational information in relational databases. In this case, we are interested in finding related patterns within the tables of our database. Applications are abundant, as almost every (commercially) used database is relational. Regrettably, the application of data mining techniques to these relational areas brings about intrinsic hard computational problems. Given this, we present two recent approaches to improve the efficiency within a popular area: frequent pattern mining.

Given a single table that contains structured data, it is known that frequent pattern mining results in a huge set of patterns. In practice, the user gets overwhelmed with many similar patterns; especially when handling lower support values. We show that we can find a small set of interesting patterns using the Minimum Description Length (MDL) principle. Using our algorithm named KRIMP, we attain reductions of the frequent pattern set by several orders of magnitude.

When dealing with a relational database, one could naively join all tables into one and apply the above algorithm to derive interesting patterns. Aside from the induced space complexity, we will lose structural information. Therefore, we extend frequent pattern mining from the single table case to a relational setting. As mentioned, this extension implicates a hard computational time and space complexity. At low min-sup levels, each single table leads to a large associated frequent pattern set. On top of this, to derive the candidate set of patterns for the multi-relational case, we are confronted with a combinatorial explosion. As a result, the well-known frequent pattern explosion at low min-sup settings is far worse than it is in the standard case.

We introduce an effective algorithm for the discovery of frequent, multi-relational item sets. These relational patterns show which item sets occur together. Answering questions like: 'What type of Books are bought together with what Record types?'. Hence, they provide a symmetric insight in the relation. Aside from reducing the resulting set of patterns, our KRIMP based algorithm reduces the amount of candidates dramatically. This brings down the computational complexity by orders of magnitude. In experiments we show that this approach yields a very good approximation of the naive approach, joining all tables into one huge table, while being far more efficient.