Abstract

This paper applies the data mining techniques in the area of data cleaning as effective in discovering Constant Conditional Functional Dependencies (CCFDs) from relational databases. These CCFDs are used as business rules for context dependent data validations. Conditional Functional Dependencies (CFDs) are an extension of Functional dependencies (FDs) which captures the consistency of data by supporting patterns of semantically related constants. Based on the hierarchy between FDs, CFDs, and Association Rules: Union of Association Rules are CFDs, while union of CFDs are FDs. This paper proposes the algorithms used for Association Rule discovery to be reused for CCFD Mining i.e., CFDs with constant patterns only. Three algorithms for CCFD mining namely CCFD-FPGrowth, CCFD-AprioriClose, and CCFD-ZartMNR are provided in this paper. CCFD-FPGrowth uses FP-growth algorithm to find frequent itemsets and then generates rules as constant patterns from the set of frequent itemsets using modified Agrawal Association rule Generation algorithm. CCFD-AprioriClose uses Apriori algorithm to find frequent closed itemsets and then generates rules as constant patterns from the set of frequent closed itemsets using modified Agrawal Association rule Generation algorithm. CCFD-ZartMNR uses Zart algorithm to find closed itemsets and minimal generators and then generates minimal non-redundant rules from the set of closed itemsets. Experimental results on two real-world data sets show that this approach performs well across several dimensions such as recall, runtime and scalability.
Mining Constant Conditional Functional Dependencies for Improving Data Quality

References

- Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery Journal, 8, 53-87, 2004.
Konstanz, Germany, 2000. Springer.

Index Terms

Computer Science
Information Sciences

Keywords
Data Cleaning
Constant Conditional Functional Dependency (CCFD)
Conditional Functional Dependency (CFD)
Frequent Pattern Growth (FP) tree
Frequent Itemsets
Closed Itemsets