ABSTRACT

As a huge volume of data is getting generated every day, data integration becomes important to provide a uniform view over the data collected from multiple resources. Effective data analysis methods are required to work with such massive integrated data and efficient data quality tools are also vital to ensure the correctness of the integrated data. The data store for many organizations is the databases and integrity constraints are the primary means for ensuring data integrity in databases. Functional Dependencies are the common type of integrity constraints that are specified by the database designer when the database schema is designed. In the quest for capturing more information from data in the form of constraints, functional dependencies are extended in several ways. Among various data dependencies, conditional functional dependencies, fuzzy functional dependencies and matching dependencies are widely used for data cleaning operations. In addition to the constraints specified by the designer, there are other constraints hidden in data and it becomes essential to mine dependencies from data. Several dependency discovery methods are proposed in this work to extract functional dependencies and its extensions.

The proposed work also uses information theory measure to extract various types of data dependencies from data. The use of information theory measure helps in quantifying the information content in data and serves as a good interestingness measure to identify dependencies relevant to a specific
domain. The structure of data in the database is represented using a non-negative information theory measure and this helps in reducing the dependency extraction time. Functional dependency extraction using MapReduce programming model is also proposed to extract dependencies from very large datasets.

Data integration provides the ability to manipulate data transparently across multiple distributed data sources. Schema matching and entity matching are the two main subtasks of data integration which resolves heterogeneity at schema level and data level respectively. Data dependencies also play an important role in enhancing the accuracy of schema matching and entity matching. The proposed schema matching technique uses fine grained functional dependency relationship between attributes of the database schema to find appropriate matching. An entity matching technique which uses matching dependencies that are defined using a hierarchy of functional dependency extensions is proposed which improves the accuracy of entity matching. The information theory based matching dependency extraction method helps in extracting matching rules at a faster rate. The performance of the proposed dependency extraction, schema matching and entity matching techniques is thoroughly analyzed by carrying out experiments with real-life data.