CHAPTER 7

CONCLUSION AND SCOPE FOR FUTURE WORK

In this thesis, an information theory based approach is used to extract traditional functional dependencies and its extensions like conditional, fuzzy and matching dependencies that are hidden in data. Furthermore, the discovered dependencies are applied in solving data integration issues. In summary, the proposed research work aims in extracting semantics in the form of data dependencies from data and applies the semantics to address data integration issues like schema matching and entity matching.

7.1 SUMMARY OF CONTRIBUTIONS

The following are the list of works carried out to meet the objective of the research:

- Using Information Theory measures to extract functional dependencies and its extensions from data.

- Using a parallel FD discovery algorithm based on MapReduce programming model to infer dependencies from very large data sets.

- Using the discovered dependencies to perform schema matching and entity matching.
The problem of discovering functional dependencies (FDs) from an existing relation instance has received considerable attention in the database research community. To date, even the most efficient solutions have exponential complexity in the number of attributes of the instance. An information theoretic approach is followed to extract functional dependencies from databases. Effective pruning rules are framed based on the attribute entropy, joint entropy and mutual information which does not involve computationally expensive set operations. The results show that more number of FD checks are pruned by the proposed pruning rules and only 20% of the comparisons, on an average are made by ITFD algorithm to detect the presence of FDs. For datasets with scarce number of FDs, the time complexity of the FD discovery problem in the order of \( nx2^{n-1} \) – \( n \) gets reduced to \( O(n^2) \). The proposed ITFD algorithm is an optimal solution suitable for FD mining on datasets with both poorly correlated and strongly correlated attributes.

Traditional functional dependencies can be extended in various ways and can be used for data cleaning in data integration applications. This research work considers three different variations of FDs namely CFDs, FFDs and MDs in the context of data integration. Effective pruning rules defined in terms of entropy are used for mining various FD extensions under various discovery methods. The extensions of traditional FDs capture more semantics from data in the form of rules. Algorithmic approaches are discussed, that dictates a step by step procedure to extract various dependencies from data sets. The experimental results show that the proposed approach discovers extended FDs faster, because of using effective pruning rules to reduce the dependency mining search space.

Executing functional dependency discovery algorithms on a single computer is hard and laborious with very large data sets. One of the main contributions of this study is extracting functional dependencies between
attributes from large datasets using MapReduce programming model. The grouping and aggregation programming strategy of MapReduce model is used to compute attribute entropy which is used to measure the inter attribute correlations and discover functional dependencies hidden in data. The improved algorithm can scale up with large data sets at comparatively less cost. Moreover, this distributed algorithm can accommodate the distributed nature of the input data. Furthermore, issues inherent to distributed systems, such as data transfer among nodes and node failures are taken care by Hadoop, which adds a great deal of robustness and scalability to the system.

Finally, the semantics represented by different data dependencies are exploited to improve the accuracy of schema matching and entity matching tasks in a data integration application. A schema matching technique which uses statistics from the schema instances and does not require value interpretations is proposed. In this work, functional dependency (FD) relationships between attributes of two schemas are represented in the form of a directed dependency graph. A primitive directed graph matching algorithm is used to find the matching between the two dependency graphs and find the corresponding attributes of the two schemas. Because of considering directed graphs for the graph isomorphism problem, the time complexity of matching reduces from $O(n \cdot n!)$ to $O(n^2)$. The experimental results show that the proposed approach increases the accuracy of matching as it uses fine grained functional dependency relationships between attributes to compare two schemas.

Followed by schema matching, the instance level matching is done to remove duplicate records from the integrated data. Most previous works are based on pre-defined matching rules and supervised way of detecting duplicates using trained data sets. The entity matching technique proposed in this study uses matching dependencies that are defined using a hierarchy of
data dependencies. This entity matching technique can be used to identify duplicates from any data set generated on the fly and do not require hand-coded rules to detect duplicate entities. The matching dependencies are used to identifying key matching attributes and applied as blocking keys to group similar records. The accuracy of entity matching is improved by using the matching keys discovered from data.

7.2 FUTURE EXTENSIONS

Although the dissertation has made a number of contributions to improve data quality by leveraging the power of various forms of dependencies in addressing data integration issues, it also raises several opportunities for further improvement. Some of them are reviewed in the following.

7.2.1 Extending the FD Discovery Approach

As data quality is becoming a serious issue nowadays, extracting FDs from large datasets that are dynamically changing becomes necessary for data analysis. The proposed FD discovery method can be extended in the following ways:

- The proposed FD discovery approach could be extended to work on different forms of data like XML, Probabilistic databases, spatial, temporal and spatio-temporal databases.

- Formulating information theory based extraction methods for discovering various forms of dependencies that are not considered in this thesis like sequential, differential, inclusion, metric dependencies could be a good future extension.
• FD discovery methods based on bottom up approaches could be revised by using entropy measures.

• The information theory based approach can be easily adapted for mining association rules and frequent itemsets from data.

7.2.2 Extending the Approaches for Discovering FD Extensions.

To clean data, constraints beyond CFDs are certainly needed. In future, there are several directions to carry out useful research which includes:

• There has been recent work on conditional inclusion dependencies, denoted by CINDs, which are defined along the same lines as CFDs and are demonstrated useful in data cleaning and schema matching. Efficient CIND discovery methods are required to automatically extract CINDs from data.

• It is also essential to find effective and efficient heuristic algorithms for the consistency and implication analysis of conditional constraints.

• Although preliminary work on constraint repair (Arenas et al. 2003) based on CFDs is being carried out (Cong et al. 2007), it deserves further investigation in the presence of both CFDs and CINDs.

• Methods for discovering Fuzzy FDs from fuzzy relational databases are required to extract semantic constraints from uncertain data.

• A sound and complete implication analysis and inference system for MDs is required to identify self-contradicting dependencies.
7.2.3 Extending the Parallel FD Discovery Approach

This thesis uses MapReduce model for formulating the FD discovery algorithm to work with very large datasets. There are several ways of improving the parallel approaches for dependency discovery which are described below:

- Parallel and distributed FD discovery approaches based on vertical partitioning of Data.
- Parallel and distributed approaches exploiting the MapReduce programming strategy to discover various forms of dependencies from data could be a possible extension.
- Effective pruning rules based on data distribution is required to prune the distributed FD discovery search space.

7.2.4 Extending the Schema Matching and Entity Matching Techniques

The proposed schema matching and entity matching techniques could be extended in several ways described below.

- Amalgamation of data quality management with data integration systems.
- A prototype tool for semantics extraction and data quality assessment.
- Representing semantic constraints as Ontology(s) using appropriate Description languages to make the semantics machine understandable.
It is planned to apply the research findings for dynamic heterogeneous systems where data keeps changing in an incremental way. Materialized data warehouse views need to be maintained when the data sources change. Materializing data warehouse by considering both schema level and data level is much important to handle view maintenance problem of data warehouse.

Data integration in peer to peer network is another possible area of research, where problems inherent to peer to peer systems like distributed coordination, security threats and data corruption, are to be addressed. Finally, semantic extraction from Bigdata in the context of data quality is of much important in today’s information age and it seems to be an important future work.