Semantic Integrity Constraint Rule Discovery and Outlier Detection in Relational Data as a Data Quality Mining Technique

R. Vasanth Kumar Mehta
Assistant Professor, CSE Dept.,
SCSVMV University
Enathur, Kanchipuram

S. Rajalakshmi
Assistant Professor, CSE Dept., SCSVMV University
Enathur, Kanchipuram

ABSTRACT
Data Quality is critical to the quality of patterns and analysis obtained from data. One of the important factors plaguing data is violation of Semantic Integrity, leading to inconsistency, in turn resulting in generation of bad patterns or reports when data mining or warehousing techniques are applied on such data. In this paper, a data quality mining technique is proposed to automatically generate Semantic Integrity Constraint Rules from the data. Further, this process leads to identification of Outliers, which are then to be classified as either violations or genuine cases of exception. The results of applying the proposed technique on a real-life data set are discussed. Some other data quality-related observations made in the process are listed.

General Terms
Data Quality Mining, Semantic Integrity Constraints, Outlier Detection

Keywords
Data Quality, Semantic Integrity, Outliers

1. INTRODUCTION
Data Quality can be defined as a measure of 'fitness of use' or meeting end-user expectations. It is expressed as a set of dimensions including accuracy, completeness, consistency, actuality and relevance. Data Quality is relevant because the consequences of poor data quality are experienced not only in everyday situations but also in a more severe way by organizations and businesses. Further, when data is used well beyond its regular use in transactional systems to the analytical systems, the quality of data will directly affect the outcomes of the analysis, and poor quality data will result in equal or worse lack of quality in the knowledge obtained from the data by applying knowledge discovery techniques.

The major steps in Data Mining are as follows:

Step 1. Identifying Data Sources

Step 2. Extracting required data from the sources followed by cleaning and Transformation

Step 3. Integrating the data from Different Data Sources into a unified repository

Step 4. Applying Data Mining Techniques to perform Knowledge Discovery

In this process, we observe that the first attempt to clean data is performed only beyond the point of extracting data from the underlying data sources. This implies that the Data Cleaning is done well beyond the stage when data is captured. However, mistakes at the point of entry of data into the system are best solved at the entry-level itself, rather than at any further stage in the process. Data quality problems are present in single data collections, due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.[1]

In [2], the authors have defined a taxonomy of Data Quality Problems, clearly indicating the granularity at which they occur – attribute/ tuple level, single relation level, multiple relations and multiple data sources.

In the Taxonomy of Dirty Data in [3], Kim et al. have listed out several categories of dirty data. The focus in this paper is on the category of Dirty Data that is "Not-missing but Wrong Data due to Non-enforcement of Integrity Constraints." Integrity Constraints are of several types including Entity Integrity, Referential Integrity, Semantic Integrity etc.

The solution proposed in this paper is to generate the Semantic Integrity Constraints automatically from the already input relational data using data mining techniques, thereby alleviating the cited category of dirty data. The suggested approach is a novel one to enhance Data Quality as a pre-processing activity in the Data Mining process even at the single relation level cited above. This technique would fall under the domain of Data Quality Mining [4], which refers to the use of Data Mining Techniques to enhance the quality of data.

The Semantic Integrity Constraints are very specific to the domain, and the different constraints may arise only in limited cases, thereby making them difficult to specify during the Schema Specification stage. However, not specifying the Semantic Integrity Constraints will lead to Integrity violation, thereby reducing the Data Quality, which will in turn create repercussions upstream, during the Data Analysis and Mining processes.

Further, the discovered Semantic Integrity Constraint rules can be used to detect Outliers. Hence, this technique also provides a secondary benefit of outlier detection. Such detected outliers then need to be studied further, and marked either as violations or genuine cases of data points that are not in the regular range of data.

2. LITERATURE REVIEW

In [5], the authors have proposed a system that contributes to creating semantic data automatically from relational database through URI allocation and identification using text mining techniques. The authors in [6] have proposed a solution for Semi-Automated Schema Integration as a Tool to identify and resolve Naming Conflicts – yet another Data Quality issue. In [7], the authors have presented the results of a Survey of Functional Dependency Mining from Relational Databases. Various algorithms have been listed that make use of a variety of techniques including Partitions, Difference Set, Embedded FD, Rough Sets, Bayes Net and Heuristics. In [8], the authors propose Conditional functional dependencies (CFDs) as an extension of functional dependencies (FDs) with semantics of data values and use them for generating quality assessment rules. An information theoretic approach (ITCFD) is used to detect CFDs.

From the above literature review, we can conclude that Functional Dependencies are a limited form of Integrity Constraint, and can be viewed as a basis on which Semantic Integrity Constraints can be inferred from the available data.

3. PROPOSED SOLUTION

3.1 Algorithm Outline

The proposed solution involves discovery of the Semantic Integrity Constraints from the input data using association rule mining. The input dataset is a normal relational table, which consists of N cases described by k distinct attributes.

1. The proposed algorithm is as follows:
2. Transform all attributes into binomial form.
3. Generate Frequent Item Sets using FP Growth Algorithm.
4. Generate Association Rules using appropriate Confidence Parameter.
5. Combine Rules with common Anteceent.
6. Evaluate the Rules based on the required Confidence level and identify the Integrity Constraints.
7. Testing
8. Apply the constraints and evaluate the outliers.
9. Stop.

3.2 Categorical Attribute Conversion

It is very common for attributes to have many possible values. Such categorical attributes are not suitable for association rule mining. All the polynomial or categorical attributes are transformed into a binomial form, by a mapping process consisting of generating a set of attributes , one each for each category. Similarly, continuous attributes are discretized into intervals, and the intervals are also mapped to a set of attributes, each representing a category. Now, each data item will be represented by (attribute, integer-value) pair. For example, the attribute Degree having multiple categories like B.E., M.B.A., M.C.A, will be replaced by three attributes- one each corresponding to B.E., M.B.A, M.C.A. and if a student has the Degree as B.E., the value vector would be (1,0,0).

3.3 Rule Generation

The Frequent Item Sets are generated using the FP Growth Algorithm, which uses the FP-tree data structure[9]. Frequent itemsets are groups of items that often appear together in the data. The frequent-itemsets problem is that of finding sets of items that appear together in at least a threshold ratio of transactions. This threshold is defined by the 'minimum support' criteria. Many other frequent itemset mining algorithms also exist e.g. the Apriori algorithm. A major advantage of FP-Growth compared to Apriori is that it uses only 2 data scans and is therefore often applicable even on large data sets. After the frequent item sets are discovered, a set of association rules are generated from them, based on the parameters like confidence, gain, lift etc. In this paper, the Confidence parameter is used. The confidence of a rule is defined conf(X implies Y) = supp(X ∪Y)/supp(X)[10]. Each of the generated rules represents a semantic constraint. The number of rules generated can be governed by the data quality manager by suitably altering the support and confidence parameters, depending on the stringency of the required data quality. Higher the quality requirement, more the value of support and confidence. The generated rules having the same antecedent are combined to form the Integrity constraint with the consequents being combined using the Or operator. For example, {Deg_BE=>DoB=1993} and {Deg_BE=>DoB=1994} are combined to generate the Integrity Constraint that {Deg_BE=>DoB=1993 Or 1994}. The rules can be tested by the usual method of dividing the dataset into training test and test sets, or by using k-fold cross validation technique as used in other algorithms.

Once the rules are generated, they are verified by the Domain expert or analyst. Further, the instances that are not in support of the rule are individually analyzed, and they are either categorized as instances of poor data quality or earmarked as outliers caused due to genuine exceptional conditions. The rules thus generated and verified can be used as Semantic Integrity Constraints, and can be applied to any further instances being added to the data set to ensure the correctness and validity of data.

4. EXPERIMENTAL RESULTS

The input data set of the Student Placement Activity at SCSVMV University of the 2014 batch consisted of records of 548 students, spanning 40 attributes, including the academic and personal details of the students. The complete schema is hosted at [11]. The data consists of students belonging to about 8 streams of Engineering at the Under Graduate Level and students pursuing MCA Degree at the Masters Level. The above algorithm was applied on the data set using the RapidMiner data mining tool. First, the numeric attributes were discretized by using the Discretize by Frequency operator. The polynomial attributes were converted to binomial attributes by using the Nominal to Binomial Operator. The dataset now consisting of only binomial attributes was used to generate Frequent Item Sets using the FP Growth operator. The resulting frequent item sets, when operated upon by the Create Association Rules parameter resulted in a set of Association Rules with very high confidence including the following:

{Deg_BE=>PG_NA}

This rule can be interpreted as follows - if the Degree is type B.E., the P.G. percentage is likely to be marked as Not Applicable. The rule has a very high confidence of 90%. When the items not subscribing to the rule were verified, it was observed that those were cases of data entry errors, and in fact the rule has a confidence of 100%.

Likewise, another rule that was observed with very high confidence was:

{Deg_BE=>Diploma%_NA}

This rule can be interpreted as follows - if the Degree is type B.E., the Diploma% is likely to be marked as Not Applicable.

The rule has a high confidence of 74%. When the items not subscribing to the rule were verified, it was observed that those were cases of several students entering the co-diploma % in the column meant for diploma %, which was not applicable for them. After due rectification of such errors, the rule was observed to have a confidence of 100%.

The above two examples are cited to underscore a significant point that it would be wise not to prune/ ignore a rule merely because of its low confidence level because, a large number of violations of a genuine rule would also be reflected by a low confidence value.

Further, those rules with common Antecedent were combined to form the Integrity constraints. For example, {Deg_BE=>DoB=1993} having support of 45% and {Deg_BE=>DoB=1994} with support of 50% were combined to generate the Integrity Constraint that {Deg_BE=>DoB=1993 Or 1994}. The rule thus generated had a very high confidence of 95%.

Table 1. Comparison of existing and proposed methods

DATASET	SCSVMV Placement	Kinship Data Set
Number of Acceptable Rules Discovered using proposed method	8	6
Number of Acceptable Rules Discovered using existing Functional Dependency Method	4	2
Percentage increase	**50%**	**77%**

Hence, the experimental observations support the premise that the rules with high confidence could be considered as semantic integrity constraints. Such rules are usually applicable in all cases. However, due to data quality issues, as well as some genuine outlier conditions, the rules may be violated.

4.1 Some other Quality Issues

4.1.1 Phonetic-matching words

During the course of carrying out the above experiment, an important data quality issue observed was the problem of names of the City or States being spelt differently. For example, Andhra Pradesh was spelt as Andhra Pradesh, Andhrapradesh, Andharpradesh, Andhara Pradesh, etc. (and several variations in the upper-case/lower-case usage). While seemingly innocuous, non-standardization of the above variations may lead to interpretation as separate categories/ entities. One solution could be to use the Soundex function available in the various Databases and programming languages to check the similarity of the words based on the phonetic dimension[12].

The above issue would be applicable to only those attributes which appear categorical in nature. This again can be auto discovered by observing the number of values and their frequency for every attribute. Those attributes that have fewer values, but with higher frequency can be assumed to be categorical in nature.

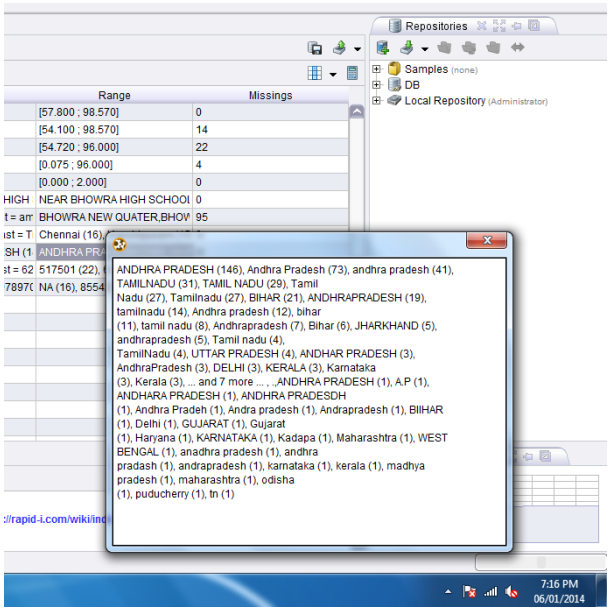

Fig 1: Need for phonetic matching.

4.1.2 Identifying Unique Attributes

Attributes that have the maximum value of statistic mode=1 are unique, and usually represent the Unique Keys. These may be marked as Unique, and are candidates for being the Primary Key. For example, in the given data set, the Roll_No attribute is unique and the maximum mode is 1.

4.1.3 Attributes with Low Support

Many of the attribute values may have very low support. In such cases, a potential solution could be to perform aggregation of the low-support attribute values.

5. CONCLUSION

In this paper, the need for ensuring data quality is discussed, followed by some of the reasons affecting data quality. As an approach to tackle consistency issues, a specific data quality mining technique is proposed that results in discovery of semantic integrity constraints from the data itself using the FP Growth algorithm and rule generation. The proposed algorithm is applied on a real-life dataset and the rules thus generated are verified to prove the correctness of the suggested technique. Further, such generated rules are applied resulting in discovery of outliers, which are then studied for violations. A few other data quality issues observed during the process are listed, and solutions discussed in brief.

6. ACKNOWLEDGMENTS

Our thanks to SCSVMV University for providing the facilities and inspiration to perform this research work.

7. REFERENCES

[1] Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." IEEE Data Eng. Bull. 23.4 (2000): 3-13.

[2] Oliveira, Paulo ; Rodrigues, Fátima ; Henriques, Pedro Rangel ; Naumann, Felix (Bearb.) ; Gertz, Michael (Bearb.) ; Madnick, Stuart E. (Bearb.): A Formal Definition of Data Quality Problems.. In: IQ :MIT, 2005

[3] Kim, Won, et al. "A taxonomy of dirty data." Data Mining and Knowledge Discovery 7.1 (2003): 81-99.

[4] Hipp, Jochen, Ulrich Güntzer, and Udo Grimmer. "Data Quality Mining-Making a Virute of Necessity." DMKD. 2001.

[5] Jeong, Chang-Hoo; Choi, Sung-Pil; Shin, Sung-Ho; Lee, Seungwoo; Jung, Hanmin; Kim, Soon-Young; Kim, Pyung, "Creating Semantic Data from Relational Database," Social Computing (SocialCom), 2013 International Conference on , vol., no., pp.1081,1086, 8-14 Sept. 2013

[6] Tahat, Said, and Kamsuriah Ahmad. "Semi-Automated Schema Integration (Icase): A Tool To Identify And Resolve Naming Conflicts." Australian Journal of Basic & Applied Sciences 7.7 (2013).

[7] Chavan, Anupama A., and Vijay Kumar Verma. "Functional Dependency Mining form Relational Database: A Survey." International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958.

[8] Extracting Data Quality Rules Using Information Theoretic Measures, International Review on Computers & Software . Jun2013, Vol. 8 Issue 6, p1321-1327. 7p.Author(s): Amshakal a, K.; Nedunchezhian, R.

[9] Yun, Unil, Gangin Lee, and Sung-Jin Kim. "Analyzing Efficient Algorithms of Frequent Pattern Mining." IT Convergence and Security 2012. Springer Netherlands, 2013. 937-945.

[10] Han, Jiawei, and Micheline Kamber. "Data mining: Concepts and techniques."China Machine Press 8 (2001): 3-6.

[11] www.kanchiuniv.ac.in/dm/dataset11

[12] Peled, Olga, et al. "Entity Matching in Online Social Networks." Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.