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.

References

- Oliveira, Paulo; Rodrigues, Fátima; Henriques, Pedro Rangel; Naumann, Felix (Bearb. ); Gertz, Michael (Bearb. ); Madnick, Stuart E. (Bearb. ): A Formal Definition of Data Quality
Semantic Integrity Constraint Rule Discovery and Outlier Detection in Relational Data as a Data Quality Mining Technique

Problems. In: IQ:MIT, 2005
- 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
- www.kanchiuniv.ac.in/dm/dataset11

Index Terms

Computer Science

Artificial Intelligence

Keywords

Data Quality Mining      Semantic Integrity Constraints      Outlier Detection