Analysis of data cleaning techniques

Marko Sumina

Interdisciplinary postgraduate programme in Statistics, University of Ljubljana
School of Economics and Business

Data cleansing involves the detection and elimination of errors and inconsistencies in data in order to improve data quality. Data collection and acquisition often introduce errors in data, e.g., missing values, typos, mixed formats, replicated entries for the same real-world entity, and violations of data integrity rules. A survey about the state of data science reveals that dirty data is the most common barrier faced by workers dealing with data. With the availability of large amounts of data, it has become increasingly evident that data curation, unification, preparation, and cleaning are key enablers in unleashing the value of data. Not surprisingly, developing effective and efficient data cleaning solutions is challenging and is rife with deep theoretical and practical problems. Error detection techniques can be either quantitative or qualitative. Specifically, quantitative error detection techniques often involve statistical methods to identify abnormal behaviors and hence have been mostly studied in the context of outlier detection. On the other hand, qualitative error detection techniques rely on descriptive approaches to specify patterns or constraints of a consistent data instance, and for that reason these techniques identify those data that violate such patterns or constraints as errors. In this presentation, I will briefly discuss which methods for data cleansing exist, which methods I used and how I used them.

Bibliography:

Chu, X., & Ilyas, I. (2013). Holistic data cleaning: Putting violations into. 29th International Conference on Data Engineering, 458–469.

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, 2.

Salem, R., & Saad, A. (2017). Fixing rules for data cleaning based on conditional function . Future computing and informatics journal.

Sidi, F., Shariat Panahy, P., & Affendey, L. (2012). Data quality: A survey of data quality dimensions. International Conference on Information Retrieval & Knowledge Management, (str. 300-304).

Ananthakrishna, R. C., & Ganti, V. (2002). Chapter 51 - Eliminating Fuzzy Duplicates in Data Warehouses. Proceedings of the 28th International Conference on Very Large Data, (str. 586–597).

Aggarwal, C. C. (2013). Outlier Analysis. Springer.