Improving Data Cleaning with Unstructured Data

Rihem Nasfi [1], Antoon Bronselaer [1], and Guy De Tré [1]

[1] Department of Telecommunications and Information Processing, Ghent University, Belgium
{rihem.nasfi, antoon.bronselaer, guy.detre}@ugent.be

In data analysis, a significant amount of erroneous or incomplete data can hinder informed organizational decisions prompting the need for automated data cleaning. Leveraging successful artificial intelligence techniques across various domains, several initiatives have introduced machine learning (ML) models to tackle these data-related issues [1, 2, 3]. In this approach, we propose a strategy to enhance data cleaning using ML models by incorporating unstructured data (e.g. texts). It also involves regulating machine learning model outputs by validating dataset constraints (e.g. set of edit rules). Preliminary results indicate promising accuracy in predicting correct values for erroneous labels, especially when the target value is embedded within unstructured data. Furthermore, a hybrid approach that combines machine learning with Parker algorithm [4] enhances the effectiveness of our proposed method.

References

[1] Rekatsinas, T., Chu, X., Ilyas, I. F. and Ré, C. (2017): HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proceedings of the VLDB Endowment*, 10(11), 1190–1201.

[2] Ilyas, Ihab F., and Theodoros Rekatsinas. (2022).Machine Learning and Data Cleaning: Which Serves the Other?. *ACM Journal of Data and Information Quality (JDIQ)*, 14(3), 1–11.

[3] Wu, R., Zhang, A., Ilyas, I., & Rekatsinas, T. (2020). Attention-based learning for missing data imputation in HoloClean. *Proceedings of Machine Learning and Systems*, 2, 307-325.

[4] Bronselaer, A., & Acosta, M. (2023). Parker: Data fusion through consistent repairs using edit rules under partial keys. *Information Fusion*, 100, 101942.