Parker: a data cleaning approach towards data fusion

Antoon Bronselaer [1] and Maribel Acosta [2]

[1] Department of Telecommunications and Information Processing, Ghent University, Belgium
antoon.bronselaer@ugent.be
[2] Department of Computer Science, Technical University of Munich, Germany
maribel.acosta@tum.de

The problem of data integration is commonly subdivided in three sub problems: schema mapping, duplicate detection, and data fusion [1]. In this setting, data fusion boils down to resolution of inconsistencies in the data. As such, we cast the data fusion problem into the setting of data cleaning. We hereby consider two types of consistency: within source consistency (modelled as a set of edit rules) and between source consistency (modelled as a partial key constraint). The crux of our approach is that although both edit rules and partial keys are subsumed by Denial Constraints (DCs) and Conditional Functional Dependencies (CFDs), finding minimal cost repairs for those more expressive constraints is computationally much more intensive [3, 4]. More precisely, the restriction to simple constraints allows for an efficient minimal-cost repair algorithm. At the same time, the reduction of expressiveness still allows to capture many inconsistencies from real-life data fusion tasks.

Experiments were done on three real-life data sets with various sizes and error rates in order to compare the effectiveness (F_1-score) and efficiency (repair time in seconds) of Parker with state-of-the-art approaches. These experiments show that, in terms of effectiveness, Parker outperforms Holistic [2] and HoloClean [3] and is comparable to Baran [5], although the latter method has clear scalability limitations. Moreover, an ablation study shows that combining partial keys with edit rules introduces a strong boost in effectiveness on all data sets. In terms of efficiency, Parker is faster than all other approaches and speed-up varies from two to five orders of magnitude.

References

[1] Bleiholder, J and Naumann, F. (2008). Data Fusion. *ACM Computing Surveys*, 41(1), 1–41.

[2] Chu, X., Ilyas, I. and Papotti P. (2013). Holistic data cleaning: Putting violations into context *Proceedings of the International Conference on Data Engineering*, 458–469.

[3] Rekatsinas, T, Chu, X., Ilyas, I. and Ré, C. (2017). Holoclean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 1190–1201.

[4] Geerts, F, Mecca, G., Papotti, P. and Santoro, D. (2019). Cleaning data with Llunatic. *The VLDB Journal*, 29, 867–892.

[5] Mahdavi, M. and Abedjan, Z. (2020). Baran: Effective Error Correction via a Unified Context Representation and Transfer Learning. *Proceedings of the VLDB Endowment*, 13 (12), 1948–1961.