SiMa: Effective and Efficient Matching Across Data Silos Using Graph Neural Networks

Christos Koutras Rihan Hai Kyriakos Psarakis Marios Fragkoulis* Asterios Katsifodimos
Delft University of Technology *Delivery Hero SE

Abstract

Given a large set of datasets split across different data silos, as well as example column relationships within those silos, how can we detect pairs of dataset columns, that are joinable or unionable across silos? Can we do this *efficiently* and *effectively*?

Organizations nowadays accumulate large numbers of heterogeneous datasets in data lakes, with the goal of gaining insights by combining those datasets. The structure (e.g., departments, teams, locations) of organizations, but also the sheer scale of their data lakes, force organizations to establish barriers for their data assets, leading to the phenomenon of *data silos*: disjoint and isolated collections of datasets, belonging to different stakeholders. Interestingly, data silos may even exist within the same organization, as individual teams enforce their own conventions and formats, as well as encapsulate knowledge about their data assets. Silo-ing data impedes collaboration and information sharing among different groups of interest.

In the data management research, the problem of finding relationships among datasets has been investigated in three different contexts: *i*) *schema matching*, with a multitude of automated methods [5]; ii) *related-dataset search* [2, 4, 1], and iii) *column-type detection* [3]. In short, traditional schema matching methods are *a)* prohibitively expensive; *b)* they cannot always be employed in the setting of data silos as they require co-locating all datasets to calculate similarities; *c)* they do not leverage existing knowledge within silos. Related-dataset search methods are not applicable to the matching problem as their goal is to search top-k related datasets to a given dataset, sacrificing recall for precision. Finally, column-type detection requires knowing the types of all columns in advance, alongside massive training data.

In this work we propose SiMa, a novel approach for discovering relationships between tabular columns across data silos. SiMa is based on the observation that *within* silos we can find existing matches among columns and train a ML model that learns to predict column relationships *across* silos: *i)* equi-joinable, *ii)* fuzzily-joinable, *iii)* unionable columns of the same domain. Towards this direction, SiMa leverages the representational power of *Graph Neural Networks* (GNNs). However, employing GNNs for the purposes of matching across data silos is far from straightforward, as we need to: *i*) transform tabular data to corresponding information-preserving graphs, *ii*) initialize nodes with suitable features, *iii*) introduce sophisticated negative sampling techniques and training schemes to optimize the learning process. Therefore, SiMa provides with effective and efficient solutions to each of these problems.

References

[1] A. Bogatu, A. A. Fernandes, N. W. Paton, and N. Konstantinou. Dataset discovery in data lakes. In *IEEE ICDE*, 2020.

[2] R. C. Fernandez, Z. Abedjan, et al. Aurum: A data discovery system. In *IEEE ICDE*, 2018.

[3] M. Hulsebos, K. Hu, M. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, C. Demiralp, and C. Hidalgo. Sherlock: A deep learning approach to semantic data type detection. In *SIGKDD*, pages 1500–1508, 2019.

[4] F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller. Table union search on open data. In *VLDB*, 2018.

[5] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDBJ*, 10(4):334–350, 2001.