Scalable Discovery of Complementary Datasets in Data Lakes

Andra Ionescu, Marios Fragkoulis, and Asterios Katsifodimos

Delft University of Technology
{a.ionescu-3, m.fragkoulis, a.katsifodimos}@tudelft.nl

Abstract

Due to the high volume of data produced by various systems and the emergence of data lakes, the necessity of discovering information that can satisfy certain data needs is increasing. Data augmentation is a technique to increase the amount of available data by either adding attributes or rows in a particular dataset. One particular type of data augmentation is to find and combine what we call *complementary datasets*. Previous work on data discovery defines two types of complementarity: entity complement and schema complement [2]. An entity complement denotes the union between two tables that come from the same source table, while a schema complement denotes the join operation. The purpose of entity and schema complement is to augment a given table with more information to increase its *value*.

In this line of work we aim at defining how complementarity can be quantified, with the goal of discovering and combining datasets in a data lake in creative ways. More specifically, one can think of a source and a target dataset are in a complementary relation if, combined, they increase their data value, given an information need. For instance, the need could be training an accurate model using the combination of those datasets [1]. At this (early) stage, we focus on the complementarity in the case of joinability and unionability, and we are modelling and quantifying complementarity using information entropy [3]. We aim at using information entropy as a function to rank the results according to the users' needs, to support the data discovery process in large data lakes.

References

1. Chepurko, N., Marcus, R., Zgraggen, E., Fernandez, R.C., Kraska, T., Karger, D.: Arda: Automatic relational data augmentation for machine learning. arXiv preprint arXiv:2003.09758 (2020)
2. Das Sarma, A., Fang, L., Gupta, N., Halevy, A., Lee, H., Wu, F., Xin, R., Yu, C.: Finding related tables. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. pp. 817–828 (2012)
3. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review **5**(1), 3–55 (2001)