A Novel Schema Matching Approach Based on Relational Embeddings

Christos Koutras, Asterios Katsifodimos, and Christoph Lofi

Delft University of Technology
{c.koutras, a.katsifodimos, c.lofi}@tudelft.nl

Abstract

Modern companies struggle with the integration of the plethora of datasets they have in their possession. Such data is typically stored across multiple storage systems using a variety of diverse schemata. Traditionally, data integration has been a mostly manual task with only limited tool support. However, due the size of current data collections, automation is the key to ensure future scalability. Early approaches towards automated data integrated focused on *Schema Matching*, i.e. the process of capturing potential relationships between different data sources and their schemata. However, most of the existing matching methods [2] rely on purely syntactic information, i.e. the symbolic representation of data as found in a database without considering their semantics or context; this limits the quality of the discovered matches.

Thus, we present a semantic matching technique, named REMA, relying on *relational embeddings* which leverage contextual information extracted from the datasets to facilitate discovery of diverse and high quality matches. In more detail, REMA is based on [1] and contributes: i) a method for transforming relational data into a heterogeneous non-directed graph. This graph serves as in input for a walk-based node embedding which encodes contextual similarity of tuples, attributes, and values, ii) a method for using these embeddings to discover high-quality schema matches. Evaluation on a series of datasets shows the ability of REMA to capture accurate relationships between columns of different relational tables, proving that it can indeed improve the accuracy of schema matching tasks. In addition, we discuss and elaborate the challenges raised by such embedding-based schema matching approaches, and the potential focus points of future work.

References

1. Koutras, C.: Data as a language: A novel approach to data integration. In: Proceedings of the VLDB 2019 PhD Workshop (2019)
2. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDBJ **10**(4), 334–350 (2001)