The Health Problems of Data Integration

Avigdor Gal

Technion - Israel Institute of Technology

Haifa

32000 Israel

`avigal@ie.technion.ac.il`

Data integration is the process of combining data residing at different data sources to generate a unified data view of these data. Schema matching generates correspondences between concepts describing the meaning of data in various heterogeneous, distributed data sources. Therefore, schema matching is recognized to be one of the basic operations required by the process of data integration and thus has a great impact on its outcome and on numerous modern applications.

We discuss the poor health of data integration and trace it back to inherent uncertainty in the schema matching process. Then, we investigate new ways to reach a common data model for schema matching. We discuss the use of matrix theory as a theoretical generic foundation to schema matching and propose to adopt the similarity matrix abstraction as a basic data model for schema matching. Such a common data model can abstract away the differences between schema matchers and focus on their similarities instead. This approach is useful in designing generic tools with wide applicability rather than solving isolated problems. Therefore, basic matching operations will be captured as matrix operations, regardless of whether the matcher itself uses a linguistic heuristic, a machine learning heuristic, *etc.* We discuss the type of abstraction that is needed for setting such a foundation and what we envision as its greatest benefits. We argue that such a framework would allow an efficient assessment of schema matcher quality, yielding a better mechanism for designing and applying new schema matchers.

Using this data model we propose a complementary approach to the design of new schema matchers. We separate schema matchers into first line and second line matchers. First line schema matchers were designed by-and-large as application of existing works in other areas (*e.g.*, machine learning and information retrieval) to schemata. Second line schema matchers operate on the outcome of other schema matchers to improve their original outcome. We aim to show the benefit in this classification, together with the use of similarity matrix as a data model, to the design of new schema matchers. Therefore, we generalize the notion of second line matchers and discuss their properties.

Given the importance of schema matching in data integration and the limitations of the state-of-the-art foundations, the significance of developing a common theoretical generic foundation for schema matching design is apparent. By having a new perspective of existing representations, we advance the state-of-the-art in providing new mechanisms to deal with the ever daunting problem of schema matching.