Summary

Entity Reconciliation using Object Similarity

Case 'Matching person entities without the existence a common identifier'

For research purposes a unified and integral view on entities in the field of police and justice is of crucial importance. For several reasons, e.g. enforcement of privacy, linking databases on primary and foreign keys is not always possible or desired. We have developed an approach that focuses on reconciliation in these situations. Our approach is based on exploiting a set of overlapping and related attributes. An attribute in this set does not uniquely identify an entity, but discriminates entities to a certain extent (i.e., the selectivity factor is not too large).

To perform reconciliations, we have combined schema information and the content of databases with available domain knowledge of experts. Schema information of different databases is used to determine what parts of a schema pertain to the same real-world entity. The content of the databases and available domain knowledge are used to define similarity functions. These functions are used to decide whether tuples in different databases refer to the same real-world entity or not.

We have implemented our approach, resulting in a prototype called EROS. We have applied EROS on two databases in the field of police and justice. It appears that our approach can be marked as quite effective, since more than 93% of the tuples have been correctly reconciliated.

For the time-being EROS is able to process only two databases at a time. Extending EROS as such that it is able to process more than two databases at a time in an efficient way is a topic for further research. Extending the knowledge system of EROS with more rules is another topic for further research. To what extent our approach can be generalized is also a topic that needs attention.