Process-based Schema Matching: From Manual Design to Adaptive Process Construction

Dissertation, 2013
Eric Peukert

Mappings between complex metadata structures are needed in a number of domains such as data integration, ontology alignment or web service composition. A mapping describes how elements of one metadata structure correspond to elements of another metadata structure. Defining such mappings is a complex and time-consuming process. It is often done manually, with the help of point and click interfaces. In the last 10 to 15 years a strong effort was made in research to partly automate the mapping process. Many schema- and ontology matching systems were developed to semi-automatically compute mapping suggestions for a user. Current systems are often not robust enough to be able to cope with different mapping problems and often face performance issues when mapping large structures. This thesis investigates how to support the user in the task of configuring schema matching systems and how to improve run-time performance.

Initially, fundamental concepts of schema matching are introduced and an overview to the existing body of work is given together with a pre-evaluation of existing approaches.

The second part of the thesis then introduces a new matching process model that supports adaptivity. It defines a set of operators for matching and filtering that can be used to create domain-specific matching processes. A condition construct is introduced that helps to adapt the match processing to the specifics of the problem at hand based on features of the input schemas or intermediate matching results. A new tool for graphically constructing and tuning matching processes is introduced. It eases development of matching processes by using a drag and drop metaphor. Furthermore, it provides visualizations for tuning and debugging matching processes. From the experience that was made by modeling matching processes a number of reappearing matching process design patterns are identified.

In a third part, the thesis introduces a novel rule-based technique for automating the construction and configuration of matching processes. The configuration of run-time performance aspects is automated by rewriting matching processes. Based on a simple cost-model, parallel combinations of matchers can be rewritten to sequential matching processes containing filter operators. By sequentializing parallel matching processes with filter-based rewrite rules significant run-time performance improvements (up to a factor of 9) could be achieved. In particular together with a so-called dynamic filter strategy improvements were achieved without changing the quality of a schema matching process. The rewrite-based approach is adopted for automatically constructing matching processes that are tailored to a given mapping problem. Based on measured features of the input schemas and intermediate results so-called matching rules can be defined. These rewrite rules rely on analyzing the input schemas and intermediate results while executing a process and rewrite the process to better fit to the problem at hand. The evaluation shows that the approach behaves more robust than existing schema matching approaches without involving the user in complex configuration tasks.