On semi-automated matching and integration of database schemas

Ünal Karakaş, Ö.

Publication date
2010

Summary

On semi-automated matching and integration of database schemas

Today, increasingly more organizations understand the need to collaborate in order to better achieve their common goals. As a result of this tendency towards increased collaboration, a line of research and development has focused on addressing data sharing and interoperation among organizations, and developing ICT tools and systems to support them. But many open challenges still remain in this area. Focusing on sharing and integration of information among independent nodes within collaborative networks, and to provision transparent access to the information stored in their databases, form the base for their effective co-working. But clearly, independent nodes model their information heterogeneously in their database schemas. Thus, before any information sharing can occur among these databases, the main challenges to be addressed include: (i) *identification and establishment of correspondences among concepts* defined in independent database schemas, (ii) *resolution of existing heterogeneities among database schemas* of the involved nodes, and (iii) *integration of these database schemas.* Typically these three tasks are complicated, even to handle manually, and this difficulty intensifies by the number of nodes within the collaborative network, as well as the size of their database schemas. A number of research approaches and prototypes have therefore aimed at automating the matching of schemas, while a few others have independently attempted at automating the integration of schemas. However, in spite of the related past efforts both in research and in commercial developments, today the matching and integration of schemas still involve a large amount of manual work and there are a large number of open research issues that remain in these areas.

This thesis proposes an *automated but supervised combined approach that addresses and merges the problems of matching and integration of relational database schemas*. Thus the main contribution of the thesis is a new combined schema matching/integration approach. A proof of concept for this approach is provided, as an implemented prototype system called SASMINT – Semi-Automatic Schema Matching and INTegration. The SASMINT system automatically identifies a large number of syntactic, semantic, and structural heterogeneities among relational database schemas. It then attempts to resolve their heterogeneity, proposing for user validation a list of potential matches among the compared schemas. The system then automatically generates an integrated schema from this list.

The Chapter 1 of this thesis presents the motivation for this work, the main research questions, the objectives, and the main contributions of this research.

Chapter 2 elaborates on the base definitions and the classification of concepts from the state of the art, in relation to architectures, approaches, and systems that enable sharing and exchange of distributed and heterogeneous information.

As heterogeneity stands at the center of the schema matching and integration processes, identification and resolution of different types of heterogeneities are crucial for the success of these processes. Chapter 3 presents different taxonomies introduced for heterogeneities, and narrows them down to the database schema heterogeneity, the main subject of the thesis.

After establishing the base for our research in the first three chapters, the rest of this dissertation addresses the design and development of our proposed approach. As a main chapter of this thesis, Chapter 4 contains a literature survey focused on (i) database integration and interoperability, (ii) database schema matching, (iii) database schema integration, and (iv) ontology matching and merging. Then our proposed solution, SASMINT, is introduced, to address a number of identified open issues. The SASMINT approach increases the accuracy of schema matching, through the weighted combination of a number of schema matching algorithms, where each algorithm resolves a different specific kind of syntactic, semantic, or structural conflicts. Furthermore, another contribution of SASMINT covered in Chapter 4 is the introduction of our so called SAMPLER technique, which semi-automatically identifies for every target domain the appropriate weights for each algorithm planned to be applied in the linguistic matching process. Chapter 4 also introduces an overview of a set of rules that enable the automatic generation of both the integrated schemas as well as the derivation constructs that represent the history of the integration process in the collaboration network. Our development of SASMINT Derivation Markup Language (SDML), which captures and supports the creation of both persisting schema match results and persisting schema integration results, is also described in this chapter**.**

Similar to any other research work, it is important to verify and validate the approach proposed by our research. For this purpose, we have implemented our approach in the SASMINT system, which is the focus of Chapter 5. The main components of the SASMINT system, as well as its operation are described in this chapter. While in our opinion, supporting user interactions through a GUI is fundamental for the semi-automated schema matching and schema integration processes, this component is typically missing from the research and development work in this area. A GUI is required to support the user with verification of the automatically identified schema conflicts and matches, as well as the proposed integrated schema. Through SASMINT's GUI, users interact with the system, set proper weights for its processes, approve/modify/disapprove its automatically generated results, and save the results of both schema matching and schema integration.

To demonstrate and evaluate the results of this research and to measure the quality of the SASMINT system, we have carried out a number of experiments. Specifically, the schema matching component of SASMINT is compared and evaluated against another state of the art schema matching system. But the approach of SASMINT is unique in that it merges the schema matching and schema integration processes. Hence, we could not find a counterpart to compare the schema integration component of SASMINT. Therefore, in our evaluation experiments we have only measured its success rate in producing accurate results. These experiments are elaborated at length in Chapter 6 of this thesis.

This thesis concludes by explaining how it has addressed the main research questions.