Combining DAML+OIL, XSLT, and Probabilistic Logics for Uncertain Schema Mappings in MIND

Henrik Nottelmann and Norbert Fuhr

Institute of Informatics and Interactive Systems, University of Duisburg-Essen, 47048 Duisburg, Germany, {nottelmann,fuhr}@uni-duisburg.de

Abstract. When distributed, heterogeneous digital libraries have to be integrated, one of the crucial tasks is to map between different schemas. As schemas may have different granularities, and as schema attributes do not always match precisely, a general-purpose schema mapping approach requires support for uncertain mappings. In this paper we present one of the very few approaches for defining and using uncertain schema mappings. We combine different technologies like DAML+OIL, probabilistic Datalog (since DAML+OIL—as similar ontology languages—lacks rules) and XSLT for actually transforming queries and documents. This declarative approach is fully implemented in the project MIND (which develops methods for retrieval in networked multimedia digital libraries). However, as DAML+OIL lacks some important features, the proposed approach is only a stepping stone for an integrated solution.

1 Introduction

Federated digital libraries (DLs) integrate a large number of legacy DLs; so they give users the impression of one coherent, homogeneous library. MIND [15] is such a system for heterogeneous and non-co-operative multimedia libraries. Heterogeneity appears in different forms: In MIND, differences in query languages, communication protocols and document models are solved by DL-specific wrappers.

Libraries also differ in the document structure (schemas). Users cannot deal efficiently with this semantic heterogeneity and therefore should only know one system-wide or personalised standard schema. In MIND, the standard schema is defined ontologically and independent from the sources. Then, queries are transformed from the standard schema into the DL schemas, and documents vice versa.

In contrast to most of the approaches available so far, MIND explicitly supports uncertain schema mappings. Schemas may have different granularity, and schema attributes do not always match precisely (e.g. authors and editors vs. the more general attribute creators). So creators cannot be mapped onto authors precisely but only with a specific probability. Systems with purely deterministic mappings fail in such settings.

In MIND, we model MIND documents and queries in DAML+OIL (the forthcoming standard ontology language) so that we can extend our approach to nested structures fairly easy. As DAML+OIL lacks rules, we specify schema mappings in probabilistic Datalog. The rules are then converted into XSLT stylesheets for transforming the DAML+OIL models. XSLT implementations are available for all major programming languages, so our approach can be used easily in other projects. DAML+OIL schema definitions,
pDatalog rules and the resulting XSLT stylesheets can be stored in textual files, so it is very easy to update the mappings when a schema is changed (without recompiling code). We are aware that there are many open questions, so this work should be considered a stepping stone.

The rest of this paper is structured as follows. The next section gives a survey of other approaches for handling heterogeneous schemas. Section 3 introduces the MIND document model. In section 4 we model MIND documents and queries in DAML+OIL. Section 5 describes how schema mappings are expressed in probabilistic logics. Section 6 summarises the major ideas presented in this paper and gives an outlook on future research.

2 Related Work

Mappings between heterogeneous schemas have been studied for quite a while, but only one of the existing approaches allows for uncertain mappings.

In the field of federated databases, two approaches are distinguished: In “local as view” (LaV), the source schemas are defined as views (mappings) over a fixed global schema. This makes it easy to add a new source, but query transformation has exponential time complexity. In contrast, the global schema is defined as a view over local schemas in the “global as view” (GaV) approach. Here, query transformation can be reduced to rule unfolding, but the global view has to be modified whenever a new source is added.

The BGLaV approach [20] combines the advantages of both worlds. The global schema is specified ontologically and independent from the sources, the source schema models the documents returned by the source, and mappings are defined by relational algebra expressions. This approach has polynomial time complexity for query transformation (like GaV) while adding new sources is fairly simple (like LaV), and is used in an extended form in MIND.

A framework for dealing with heterogeneous OSM schemas is presented in [1]. OSM models contain objects, their relationships and a predicate calculus for expressing constraints. As in BGLaV, the global schema is defined ontologically and independent from the source schemas. Interaction with an administrator is assumed (however not required) for setting up deterministic mappings between objects (and relations, respectively). These mappings can model specialisation/generalisation relationships, and string processing operators are provided.

The GaV system Demetrios [6] follows a different approach with SQL as query language for both the user and the sources. The SQL extension FRAQL is used for defining mappings from the source relations onto global relations. FRAQL allows e.g. for accessing metadata, restructuring tables and converting values.

TSIMMIS [3] is one of the early systems integrating heterogeneous digital libraries. Schema mappings are defined in a textual format with actions which are executed when a corresponding template matches a query.

MARIAN [10] shares some common aspects with our approach. Attributes can be mapped onto others with simple rules, specified with the declarative 5S language. Uncertain mappings are possible by weighting the mapping rules. In contrast to the wrapper based system MIND, MARIAN uses a harvesting approach: It periodically downloads