Challenges in Automatic Schema Matching for Data Migration

Christian Drumm

SAP Research, Karlsruhe, Germany
christian.drumm@sap.com

Introduction

Data migration is the task of transforming and integrating data originating from one or multiple legacy applications or databases into a new one. Whenever a new software application is introduced to replace existing legacy applications or whenever the application landscape is consolidated the requirement to migrate data between applications arises. During the migration process, data needs to be extracted from the source systems, transformed and loaded into the target system. This process requires solving two difficult tasks: i) *schema matching* to identify similar or semantically related elements between the source and target systems and ii) *mapping discovery* to determine mapping expressions capable of transforming instance data from the source format to the target format.

Observations & Challenges

Though analyzing existing data migration scenarios of various SAP customers we observed a number of properties common to many data migration scenarios. This properties need to be taken into consideration when developing tools for automatic schema matching in order to achieve high quality results.

Limited value of schema information: We generally observed that legacy systems are mostly not optimized towards fostering system interoperability. The used schemas typically make extensive use of technical names, abbreviations and proprietary structure. This makes it difficult to determine correspondences, not to mention mapping expressions, between schema elements. As a consequence, additional kinds of information, especially instance data, domain knowledge, and previously determined mappings, need to be considered in order to achieve reasonable quality in the schema matching and mapping discovery process.

Availability of domain knowledge: The migration process needs knowledge about both the source and the target system. Unfortunately, such knowledge is not always available at one place. The knowledge about the target system is available at its vendor, while only customers can provide detailed knowledge about their source systems. As a consequence, a migration solution developed by a vendor of a target system, e.g. SAP, should exploit the knowledge about the target system, and at the same time support effective mechanisms to incorporate the customers' knowledge about their source systems. In particular, the knowledge about the target system, e.g. field semantics, sample instances, data formats and code lists, can be specified in an ontology for automatic analysis. This manual effort is needed only once and can be quickly amortized over many migration projects.

Scope of data migration projects: Typically, the data sources involved in data migration projects are complex, resulting in large schemas to be matched. Depending on the need of the particular customer, only certain parts of the target system need to be populated with source data, which results in different target schemas for different migration projects, even for the same target system. This potential for schema reduction needs to be exploited as much as possible in order to reduce the complexity of the data migration tasks; Schema reduction can be performed with suitable involvement of the customer at the beginning of the migration process.

Accumulated mapping knowledge: In a given migration project usually many different mapping tasks need to be solved, e.g. for customer data, supplier data, purchase orders, etc. As these schemas usually contain common parts, like address data, there is a high reuse potential for the results from previous mapping tasks, especially regarding complex mapping expressions which otherwise have to be developed manually.

Ubiquitous usage of codes: In legacy as well as in current systems special codes are used ubiquitously (e.g. country code, business partner type code or gender code). These codes are either based on standardized code list like the ISO country codes or specific to a certain system. As many of these codes use similar code values to represent different information automatically creating correct matches and mapping expressions for schema elements using codes is complex.

Complex mappings: The mappings in the data migrations projects we investigated contained a significant number of complex mappings. Even 1-1 mappings between one element of a source schema and one element of a target schema often are no simply "move" operations but rather contain complex mapping expressions. In order for an automatic schema mapping approach to be applicable in industrial data migration projects, the approach must be capable of

generating at least some of these complex mappings. However, suggesting a template for a mapping expression which needs to be completed by a user might also be sufficient.

Mapping quality: The quality of automatically proposed mappings is crucial for the acceptance of an automatic approach. Regarding the mapping quality a high precision close to 1 should be the goal. The reason is that a precision close to 1 allows the user of an automatic tool to focus on the creation of the complex mappings that cannot be generated automatically instead of correcting wrong proposals.

Position Statement

Bases on these observations and challenges SAP Research developed *QuickMig*, an integrated approach for schema matching and mapping discovery in the context of data migration [1]. QuickMig is based on COMA++. Its main features are:

- **Use of sample instances:** Instead of exploiting unrelated instances between source and target system, we define and use a uniform set of standard instances.
- **Instance-based matchers:** Exploiting the availability of the same instances in the source and the target system, we developed a set of relatively simple instance-based matchers, which however are able to detect complex correspondences and mapping expressions in real-world schemas.
- **Mapping categories**: A comprehensive set of mapping categories, indicating how instance data has to be transformed from the source to the target format have been developed to capture complex mapping expressions in the data migration use case.
- **Schema reduction based on domain knowledge:** To deal with large and complex schemas, we developed a new questionnaire-based technique for schema reduction. Specified in domain nomenclature, the questionnaire allows the user to specify portions of the schemas relevant for data migration according to his understanding of the domain.

The approach was experimentally evaluated using real SAP schemas.[1] The schemas used in the evaluation are quite large (between ~500 and ~4500 elements) and range from deeply structured schemas with verbose element name to rather flat ones using crytic element names. As an example of the some of the problmes faced when mapping those schemas consider the form of address of a customer. This information is represented in one schema by the element `ANRED` using a textual representation of the form of address while it is represented by the element `FormOfAddressCode` using a coded representation in another schema. In the experiments using these schemas the QuickMig approach achieved an average precision of ~0.99, an average recall of ~0.72 and an average F-measure of ~0.84. Furthermore, QuickMig was able to identify the correct mapping categories with an average precision of ~0.97.

Due to these promising results we are currently working on a prototypical integrate of the QuickMig approach into an SAP data migration tools. This integatraion will in the future be used to test the approach in further scenarios. Besides that the current focus of our research is on the development of mapping algorithms capable of coping with the large number of code values we encountered in reald data migration scenarios.

During the workshop we would be interested in discussing schema mapping in the context of special scenarios like e.g. data migration. Especially we would be interested in discussing approached and ideas on how domain knowledge and also peculiarities of certain scenarios can beneficially be exploited when developing approaches to automatic schema mapping.

References

[1] *QuickMig - Automatic Schema Matching for Data Migration Projects.* Drumm, Christian, et al. 2007. Proc. of the Sixteenth Conference on Information and Knowledge Management (CIKM 2007).

[1] The schemas used in the evaluation of QuickMig are available for download at: http://dbs.uni-leipzig.de/de/publication/title/quickmig_evaluation_data_sets