Clio: A Semi-Automatic Tool For Schema Mapping

Mauricio A. Hernández
IBM Almaden Research Center
San Jose, CA 95120
mauricio@almaden.ibm.com

Renée J. Miller
Department of Computer Science
Univesity of Toronto
Toronto, ON, Canada
miller@cs.toronto.edu

Laura M. Haas
IBM Almaden Research Center
San Jose, CA 95120
laura@almaden.ibm.com

We consider the integration requirements of modern data intensive applications including data warehousing, global information systems and electronic commerce. At the heart of these requirements lies the *schema mapping* problem in which a source (legacy) database must be mapped into a different, but fixed, target schema. The goal of schema mapping is the discovery of a query or set of queries to map source databases into the new structure. We demonstrate *Clio*, a new semi-automated tool for creating schema mappings. Clio employs a mapping-by-example paradigm that relies on the use of *value correspondences* describing how a value of a target attribute can be created from a set of values of source attributes.

A typical session with Clio starts with the user loading a source and a target schema into the system. These schemas are read from either an underlying Object-Relational database or from an XML file with an associated XML Schema. Users can then draw value correspondences mapping source attributes into target attributes. Clio's mapping engine incrementally produces the SQL queries that realize the mappings implied by the correspondences. Clio provides schema and data browsers and other feedback to allow users to understand the mapping produced.

Entering and manipulating value correspondences can be done in two modes. In the **Schema View** mode, users see a representation of the source and target schema and create value correspondences by selecting schema objects from the source and mapping them to a target attribute. The alternative **Data View** mode offers a WYSIWYG interface for the mapping process that displays example data for both the source and target tables [3]. Users may add and delete value correspondences from this view and immediately see the changes reflected in the resulting target tuples. Also, the Data View mode helps users navigate through alternative mappings, understanding the often subtle differences between them. For example, in some cases, changing a join from an inner join to an outer join may dramatically change the resulting table. In other cases, the same change may have no effect due to constraints that hold on the source schema. Clio's Data View mode carefully selects target data examples which both illustrate a specific mapping, helping users understand what the mapping does, and which illustrate the differences from any alternative mappings, helping users differentiate mappings.

Both the Schema and Data View directly interact with the component at the heart of Clio: its Incremental Mapping Engine.Clio stores the current mapping as its internal state and, through an incremental algorithm, allows users to move to the next mapping one step at a time. modify a value correspondence. For example, when users add a new value correspondence, the mapping engine will infer and rank alternative mappings that can be formed with the new mapping. The top-ranked alternative will be suggested first and displayed on the GUI. Users can, of course, examine and choose another mapping from the set of alternatives. Details of the mapping algorithm can be found in [2].

Although the demo will concentrate on illustrating the use of the Schema View and the Data View to guide mappings, two auxiliary modules will be used as part of the demo. The first module mines source data for keys and foreign key constraints. Discovered referential constraints are then suggested as likely join-paths in mappings involving multiple source relations. The second module, the *attribute matching* module, uses a novel variation of existing classification techniques based on domain-independent feature selection to compare source and target attributes [1].

ADDITIONAL AUTHORS

Lingling Yan, C.T. Howard Ho (IBM Almaden Center), and Xuqing Tian (U.C. Berkeley)

REFERENCES

[1] C. T. H. Ho, F. Naumann, X. Tian, L. M. Haas, and N. Megiddo. Automatic Classification of Attributes Using Feature Analysis. Submitted for consideration to VLDB 2001.

[2] R. J. Miller, L. M. Haas, and M. Hernández. Schema Mapping as Query Discovery. In *Proc. of the Int'l Conf. on VLDB*, Cairo, Egypt, 2000.

[3] L. Yan, R. J. Miller, L. M. Haas, and R. Fagin. Data-Driven Understanding and Refinement of Schema Mappings. In *ACM SIGMOD Conference*, Santa Barbara, CA, May 2001.

ACM SIGMOD 2001 May 21-24, Santa Barbara, California, USA