AJAX: An Extensible Data Cleaning Tool

Helena Galhardas*
INRIA Rocquencourt, France
Helena.Galhardas@inria.fr

Daniela Florescu
INRIA Rocquencourt, France
Daniela.Florescu@inria.fr

Dennis Shasha†
Courant Institute, NYU
shasha@cs.nyu.edu

Eric Simon
INRIA Rocquencourt, France
Eric.Simon@inria.fr

Data quality concerns arise in order to correct anomalies in a single data source (e.g., duplicate elimination in a file), or to integrate data coming from multiple sources into a single new data source (e.g., data warehouse construction). In addition, the information handled may also need to undergo a formatting and normalization process so that the resulting data is structured and presented according to the application requirements.

Data quality problems usually arise when the same real object is modeled by different data representations (the "Object Identity Problem"). This may have several causes. First, data may contain *errors*, usually due to mistyping. Second, when data comes from different sources, different naming conventions may have been used. For instance, the same customer may be referred to in different tables by slightly different but correct names. Correcting the Object Identity problem and converting data formats is the job of software known as data cleaning and transformation tools. We will present such a tool: AJAX, whose main goal is to facilitate the specification and execution of data cleaning and transformation programs.

AJAX proposes a *framework* wherein the logic of a data cleaning program is modeled as a directed graph of data transformations starting from some input source data. Four types of data transformations are distinguished. The *mapping* transformation standardizes data formats when possible or simply produces records with a more suitable format. *Matching* finds pairs of records that most probably refer to the same real object. Records are compared via a matching criteria that can be arbitrarily complex. A similarity value representing the result of the matching criteria is attached to each matching pair of compared records. *Clustering* groups together matching pairs with a high similarity value by applying a given grouping criteria (e.g. by transitive closure). Finally, *merging* collapses each individual cluster into a tuple of the resulting data source. AJAX provides a "declarative" language for specifying data cleaning programs, which consists of SQL statements enriched with a set of specific primitives to express these transformations.

AJAX also provides a *graphical user interface*. It allows the user to interact with an executing data cleaning program to handle exceptional cases and to inspect intermediate results. Finally, AJAX provides a *debugger* that permits users to determine the source and processing of data for debugging purposes.

We will present the AJAX system applied to two real world problems: the consolidation of a telecommunication database, and the conversion of a dirty database of bibliographic references into a set of clean, normalized, and redundancy free relational tables maintaining the same data.

*Founded by "Instituto Superior Técnico"-Technical University of Lisbon and by INESC, in the framework of Program PRAXIS XXI (Portugal)

†Supported by National Science Foundation grant IIS-9531554

References

[1] http://caravel.inria.fr/~galharda/ajax.html.