Automated Matching of Data Mining Dataset Schemata to Background Knowledge

Stanislav Vojíř, Tomáš Kliegr, Vojtěch Svátek, and Ondřej Šváb-Zamazal

University of Economics, Prague, Dept. Information and Knowledge Engineering
{stanislav.vojir,tomas.kliegr,svatek,ondrej.zamazal}@vse.cz

1 Problem Setting

Interoperability in data mining is supported by a standard for dataset and model representation: *Predictive Model Markup Language* (PMML).[1] It allows to describe the columns (which are continuous, categorical or ordinal) of the source data table, preprocessing transformations (such as discretization of continuous values) as well as the structure of the discovered model (e.g. neural network or set of association rules).

In addition to source data, the input to the mining task typically includes expert-provided *background knowledge*. It may be related, for example, to standard ways of discretizing numerical quantities (e.g. boundaries between 'normal blood pressure' and 'hypertension', which help intuitive reading of discovered hypotheses), or may itself have the form of predictive models to which the discovered models can be compared during or after the mining process. The proposal for *Background Knowledge Exchange Format* (BKEF) [**?**], in many aspects similar to PMML, aims to support interoperability of data mining (and related) applications dealing with background knowledge.

Case studies [**?**] showed that one BKEF model typically has to be aligned with different PMML models (from different mining sessions in the same domain). The alignments are stored in the *Field Mapping Language* (FML) [**?**], expressing that a data field (column) in a PMML model semantically corresponds to an abstract 'field' in a BKEF model. However, writing FML alignments (analogous to instances of the Alignment Format [**?**] used in ontology matching) by hand is tedious and recognizing suitable correspondences may be hard; partial automation is thus desirable. Furthermore, existing tools for ontology/schema matching are not straightforwardly usable, since PMML (and even BKEF) are, compared to ontologies, more biased by data structures, but BKEF is more abstract and weakly structured than database schemata. Therefore, specific methods (inspired by existing ones) and a new tool have been devised.

2 Method, Implementation and Experiments

The matching process consists of several steps. First, the data are pre-processed into a unified format that removes most syntactic differences between PMML and BKEF. Then the *similarity* between data columns of both input resources is calculated based

[1] http://www.dmg.org/

on string measures over the *column names*, as well as based on *allowed values*. Based on the similarity matrix, *suitable alignment* (1:1 or 1:N) is finally designed.

An important feature of the system is (simple) *machine learning* from interaction with the user; as far as the positive examples are concerned, the learning component distinguishes between automatically suggested correspondences that were explicitly *confirmed* by the user (or manually entered) and those that were merely *tolerated*. The output of the learned rules is used to adjust the final similarity value.

The system has been implemented as a web application in PHP (as component to the Joomla! CMS) on server side and in JavaScript (with AJAX) on client side. Its graphical interface conveniently displays the columns of both to-be-matched resources and allows to 1) *align* one to another, 2) let one be *ignored* (in subsequent automatic matching), 3) *confirm* an automatic alignment, or 4) *revoke* any of the previous operations. Similar operations can be applied on the *values* for a pair of columns.

The system has been tested within the SEWEBAR project[2] on data from medicine and finance. Furthermore, a conventional schema matching benchmark dataset was borrowed from the *Illinois Semantic Integration Archive*, which describes universities and their courses.[3] The evaluation achieved the precision of about 70% and recall about 77% on unknown columns, while when matching the data previously aligned by the user (using the machine learning facility), the recall was improved to 90-100%.

3 Relevance for Semantic Web (and Ontology Matching) Research

Although BKEF models structurally differ from ontologies, they are close to them in covering a certain *domain* in contrast to PMML models that only cover a certain *dataset*. Experience with such asymmetric matching could cross-fertilize with the task of matching the ad-hoc mixes of vocabulary entities underlying many *Linked Data sets* to carefully engineered *domain ontologies*. A promising direction could also be that of linking BKEF entities explicitly to domain ontologies; BKEF could represent an *intermediate layer* between concrete datasets (that are subject to business-analytics processes) and ontologies as sophisticated resources that are often too abstract for industrial users.

Stanislav Vojíř and Tomáš Kliegr are supported by the IGA VSE under 26/2011. Vojtěch Svátek and Ondřej Šváb-Zamazal are supported by the CSF under P202/10/0761.

References

1. David J., Euzenat J., Scharffe F., Trojahn dos Santos C.: The Alignment API 4.0. *Semantic Web*, 2(1):3-10, 2011.
2. Kliegr T., Svátek V, Ralbovský M., Šimůnek M.: SEWEBAR-CMS: Semantic Analytical Report Authoring for Data Mining Results. *Journal of Intelligent Information Systems*, Springer, 2010 (Online First).
3. Kliegr, T., Vojíř, S., Rauch, J.: Background Knowledge and PMML – First Considerations. In: PMML Workshop at KDD'11,August 21, 2011, San Diego, CA, USA.

[2] `http://sewebar.vse.cz`

[3] `http://pages.cs.wisc.edu/~anhai/wisc-si-archive/domains/courses.html`