A Semantic Web-ready framework for processing of data mining reports

Tomáš Kliegr, Martin Ralbovský, Jan Nemrava, Jan Zemánek
Department Of Information and Knowledge Engineering
Faculty of informatics and statistics
University of Economics, Prague

This research focuses on post-processing of results of data-mining algorithms as expressed in analytical reports. Analytical report is a free-text document describing the data, preprocessing steps, DM task settings and the results. In addition, it contains information provided by the analyst - particularly background knowledge on the data fields, apriori knowledge, explanation of preprocessing steps and interpretation of the results.

Creating analytical reports manually is time-consuming and the output document is not machine-readable, which hinders the possibilities for post-processing - indexing, querying, merging and filtering.

We present a novel framework for semi-automatic generation and automatic processing of the analytical reports that falls into the SEWEBAR (Semantic Web and Analytical Reports) initiative. The framework is based on existing standards: XML technologies, PMML and Topic Maps. It is tightly connected with the workflow of the analytical reporting. In the first step, user executes a data mining task and exports the results into PMML. As part of the project, PMML support was added into Ferda and LISp-Miner data mining software. In the second step, user deploys the PMML report to a centralized online repository - Content Management System Joomla (CMS). We extended the CMS so that it presents the PMML as HTML pages, but internally the reports are preserved in PMML/XML and hence are machine-readable. The functionality that would also allow the analysts to add background knowledge in a machine-readable way is under development. PMML can be used for simple querying such as finding conflicting rules within an analytical report or finding of related rules in multiple analytical reports. Further, we have conducted first experiments with transforming the reports enriched with background information into Topic Maps, a semantic web technology, in order to support more elaborate querying and inferencing.

We successfully used TOLOG, a Topic Map query language based on Prolog and SQL, to find association rules from two different datasets whose schemas were mapped to each other using background knowledge. We are carrying out further experiments aimed at more extensive involvement of background knowledge.

The framework is being implemented in order to support both academic and scientific goals and is currently undergoing stress-test since it is going to be used by about 150 students.