Analytic Reports from KDD: Integration into Semantic Web

Václav Lín and **Jan Rauch** and **Vojtěch Svátek**
Department of Information and Knowledge Engineering, and EuroMISE Center – Cardio
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
{xlinv05,rauch,svatek}@vse.cz

The Semantic Web is expected to form a huge, heterogenous collection of both manually created and (semi-) automatically generated documents. We would like to draw attention to a specific class of documents, which mix the flavors of both types, namely to *analytic reports from KDD* (Knowledge Discovery in Databases). An analytic report from KDD (further only AR) is a textual document created by human author, presenting the results of a KDD process in a condensed form. The core of an AR are the results of data mining procedures, their interpretation and generalization performed by the author. ARs are documents well readable by humans. However, due to their regular nature, they can be easily endowed with metadata lending the embedded knowledge machine-processability. This makes ARs natural candidates for the Semantic Web. The most straightforward application of the metadata is efficient *content-based retrieval* of ARs.

The raw output of datamining procedure is a set of *statistical association rules* – formal expressions stating the existence of a quantitative relationship between two classes of objects in the analyzed database [Rauch and Šimůnek, 2000]. They are of the form $X \approx Y$, where X and Y are properties automatically generated from the input data – a simple example of property is the conjunction

$$Syst\_Blood\_Press(\geq 150) \wedge Cholesterol(too\_high)$$

Symbol $\approx$ corresponds to quantitative relation between X and Y, for instance to a statistical test (such as χ^2 or F-test). The knowledge abstracted from the 'raw' association rules can be represented by formal formulae. These formulae closely resemble association rules – the only (semantically) significant difference is the use of different operators ($\approx$), which is due to abstraction performed on the raw datamining output. Roughly speaking, these operators correspond to classes of (quantitative) relations rather than to individual relations. The semantics of ARs being precisely represented, they can be retrieved with precision and recall greater than if traditional keywords were used. The techniques of analytic report construction, indexing and (content-based) retrieval are currently tested in the medical domain; some have previously been applied to other domains such as technical diagnostics.

The ARs seem to be *conceptually* ready for knowledge sharing within the Semantic Web. A hypothetical virtual network of data miners (operating on semantically compatible data), report indexers and query engines could enable synergistic discovery and sharing of interesting relationships empirically valid in the given domain. A necessary prerequisite of the reuse of analytic reports in a non-closed environment is, however, syntactic and semantic interoperability.

The base level of *syntactic* representation is obviously XML: the first, tentative DTD defining the structure of the AR indices (which can possibly be embedded into the textual reports as metadata) has already been designed. This effort is comparable with *PMML* [PMML, 2001]; our statistical association rules are however more expressive than 'classical' association rules [Aggraval *et al.*, 1996] covered by PMML. Another rule mark-up initiative, *RuleML* [Boley, 2001], offers a suite of languages evolving from positional XML towards RDF. Our notion of statistical association rule could be added as extension to the RuleML family of rule types. Similarly to RuleML, the transformation of our statistical association rules to true (RDF–based) Semantic Web resources will require several levels of serializations.

For *semantic* interoperability, *ontologies* are the key enabler. In our approach we exploit taxonomies of (original and abstracted) database attributes, which are by themselves trivial ontologies. Richer formalisms would however be needed to facilitate e.g. the integration of ontologies used by the individual clinical sites in our medical application. We consider the use of DAML+OIL for this purpose.

Acknowledgements

The research has been partially supported by project *LN00B107* of the Ministry of Education of the Czech Rep.

References

[Aggraval *et al.*, 1996] Aggraval, R. et al: Fast Discovery of Association Rules. In: Advances in Knowledge Discovery and Data Mining. AAAI Press , 1996. 307–328

[Boley, 2001] Boley, H.: The Rule Markup Language: RDF-XML Data Model, XML Schema Hierarchy, and XSL Transformations. In: 14th Intl. Conf. Applic. of Prolog.

[PMML, 2001] PMML 2.0 – Predictive Model Markup Language. http://www.dmg.org/pmmlspecs_v2.

[Rauch and Šimůnek, 2000] Rauch, J. - Šimůnek M.: Mining for 4ft Association Rules. In: Discovery Science 2000. Springer Verlag, 2000, pp. 268 - 272