Semantic Analytical Reports: A Framework for Postprocessing Data Mining Results

Tomáš Kliegr, Martin Ralbovský, Vojtěch Svátek, Jan Nemrava

University of Economics, Prague, Dept. Information and Knowledge Engineering,
Nám. Winstona Churchilla 4, 130 67 Praha 3, Czech Republic
{tomas.kliegr|ralbovsm|svatek|nemrava}@vse.cz

Abstract. Intelligent post-processing of data mining results can provide valuable insight on the problem being solved. In this paper we present first systematic solution to post-processing that is based on semantic web technologies. The framework input is constituted by PMML and description of background knowledge. Using the Topic Maps standard, a Data Mining Association Rule Mining ontologies are introduced. Using combination of a content management system and a semantic knowledge base, the analyst can can enter new pieces information or interlink existing ones. The information is accessible either via semi-automatically authored textual analytical reports or semantic querying. Prototype implementation of the framework for association rules was implemented and is demonstrated on the PKDD'99 Financial Data set.

1 Introduction

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 data, preprocessing steps, DM task settings and results. The analyst can also include additional information such as background 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 processing of analytical reports that addresses these issues through the utilization of semantic web technologies. The framework is developed as part of the SEWEBAR (Semantic Web and Analytical Reports) initiative [19, 16, 14]. The framework is generic and should be suitable for most data-mining algorithms. However, a specific implementation of the framework needs to take into account the knowledge representation used by the selected algorithm.

In this paper, we present a prototype implementation of the framework for association rules. As part of the prototype, a data mining ontology for generalized association rules is introduced. To demonstrate the feasibility of the approach, the prototype implementation is used to post-process output of Ferda [2] association rule mining system on the PKDD'99 Financial Dataset [1].

The rest of the paper is organized as follows. In Section 2 we give an overview of the architecture of the framework and in Section 3 an overview of our prototype implementation. A case study introduced in Section 4 is used to demonstrate the benefits of the framework. Section 6 contains conclusions and a plan for future work.

2 Framework Outline

In this section, we present an outline of a new framework for post-processing results of data mining tasks. The framework is based on established standards and seamlessly integrates with existing data mining software as its input is constituted by PMML (Predictive Model Markup Language) [3], a widely adopted standard for definition and sharing of data mining and statistical models. The framework introduces BKEF – Background Knowledge Exchange Format.

PMML and BKEF specifications are stored in a Content Management System (CMS), which allows to merge information contained in one or more reports with human input. The result is a human-readable analytical report.

Further in the work-flow, PMML and BKEF specifications are transformed into a semantic representation and stored in a knowledge base. This knowledge base allows the analysts to append and interlink information in a structured way.

The knowledge base can be searched using a semantic query language. An overview of the framework is depicted on Figure 2.

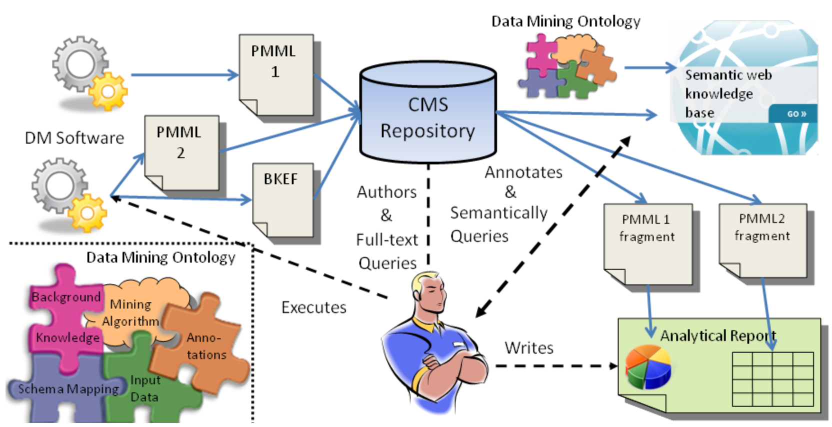

Fig. 1. Framework outline

2.1 Input Data Formats

The framework's input is constituted by the description of background knowledge and the description of the data mining models.

To the best of our knowledge there is currently no publicly available format for the exchange of background knowledge. Since the availability of such specification is vita, we are working on a new exchange format dubbed BKEF – Background Knowledge Exchange Format.

The purpose of the BKEF format is to foster exchange of information between background knowledge producers, such as special user interfaces in existing data mining systems [16], and background knowledge consumers like the framework.

The framework's input, in terms of model description, is constituted by the Predictive Model Markup Language (PMML)[3], an XML-based standard for definition and sharing of data mining and statistical models.

PMML allows two different application (the PMML producer and consumer) to share a model. Export to PMML is supported by all major data mining software implementations acting as PMML producers. There are two major types of PMML consumers – scoring and visualization applications.

PMML model specification has the following components:

- Data Dictionary: fields that are input for the model
- Mining Schema: fields used in the model
- Transformation Dictionary: derived fields created through discretization or value mapping
- Model Description: mining algorithm specific settings and output

PMML 3.2 defines the following types of models (model descriptions): association rules, cluster models, general regression, Naive Bayes, Neural Network, Regression, Ruleset, Sequences, Text models, Trees and Vector Machine.

Pieces of information contained in PMML are sufficient to fully describe the model. In the presented framework, the PMML producer is the data mining software; the framework acts as PMML consumer.

2.2 Analytical Report Authoring Support

The next element of the framework is a Content Management System (CMS), an application supporting storage and retrieval of electronic documents. The CMS is intended for storage and authoring of data mining analytical reports and related information. Analytical reports – in the CMS stage – are free-text documents whose purpose is to interpret the results of data mining tasks in the light of the background information, existing analytical and the expert's knowledge.

As such, analytical reports constitute from a large part from visualization of information contained in the PMML (such as data histograms, description of data preprocessing, model parameters,..). This allows the analyst to reuse the visualization of the needed bits of PMML in the free-text report so that existing knowledge does not have to reentered. Additionally, since this generated content preserves link to its source fragment, the corresponding part of the free-text analytical report can still be resolved to a machine-readable XML representation.

The CMS handles also handles background knowledge, an example of which is for example the allowed value range for an data field. Background knowledge can either be created by a specialized software and exported to CMS or authored directly in the CMS. In either case, the background knowledge is stored in CMS in a machine readable format (such as BKEF) and included into the free-text analytical report in the same way as PMML.

The fact that statements in analytical reports are directly backed by the source data (BKEF or PMML fragments) not only allows to search the reports as structured data but also fosters the credibility of the reports.

2.3 "Semantization" of Analytical Reports in Topic Maps

The CMS described in the previous subsection allows to query the structured content (PMML, BKEF) with XML query languages such as XQuery. However, this structured content can be further "semantized" by being stored into a knowledge base according to some ontology. This is particularly beneficial when merging heterogeneous data, such as PMML specifications of tasks executed on similar, but not same datasets, or when working with background knowledge.

As a format for interchange of semantic information, the framework uses Topic Maps, an ISO/IEC 13250 standard. Topic Maps are in principle interoperable with the semantic web formats RDF/OWL standardized by W3C. We have opted for Topic Maps for their simplicity and the availability of industry as well as open-source implementations. However, since ISO/IEC 13250 is generally interoperable with the W3C standards, the selection of a specific knowledge representation does not pose a critical issue.

Topic Maps represent information through *topics*, *associations* and *occurrences*. Topic is any entity about which a human being can lead a discourse, association represents a relationship between topics, and occurrence represents piece of information or information resource relevant to a topic. Types of topics, associations and occurrences used in a topic map constitute its ontology.

The framework defines: i) an ontology that allows to represent the structured content as instances of ontology types, ii) a transformation from structured content to the ontology.

After the analytical report is loaded into the knowledge base, it can be annotated, and interlinked with reports already present in the repository. For example, in an association rule mining task, the analyst can annotate a discovered rule with the degree of its novelty, link it with already existing rule coming either from a different report or from background knowledge or use the inference engine to search the knowledge base.

2.4 Data Mining Ontology

The Data Mining Ontology was derived from PMML and comprises of the following components: Input Data (including data transformations), Background Knowledge, Schema Mapping and Annotations. The description relating to mining algorithm (i.e. the counterpart of Mining Model from PMML) is deliberately left out, since this component is algorithm dependent and the ontology could not conceivably support every mining algorithm. The remaining areas are in general common to all data mining tasks irrespective of the specific algorithm used.

Input Data component comes out of the corresponding components of the PMML standard: Data Dictionary, Mining Schema, Transformation Dictionary. Elements prescribed by PMML Schema such as DataField were mapped to topic types, enumerations were represented as topic types with their instances representing the enumeration members. Knowledge represented as XML attributes in the PMML Schema, such as interval margins in a discretization were represented as occurrence types.

The ontology was also enriched with elements not explicitly present in the PMML XML schema. Using established naming conventions in the data-mining community, topics were organized into a taxonomy. In this way 'Category' was introduced as a superclass for both 'Discretize Bin' and 'Value Mapping', two ways for creating a Derived Field in PMML. Most importantly, association types were introduced. Instances of these association types allow to link topics and occurrences without the need of IDs or external knowledge.

Through explicitly expressing the relation between topics and occurrences through association types, additional semantics is added to the ontology compared to the XML serialization of the PMML.

Background Knowledge component allows to relate pieces of background knowledge to a specific data matrix through meta-attributes [17, 18]. Meta-attribute is a generalization of a data field. The existence of meta-attributes stem from the fact that the same property can be coded in two datasets differently. For example, there can be column 'LOAN' with possible values A - F in one dataset, and column 'STATUS' with different set of possible values such as 'bad, medium, good' in another, both columns referring to loan quality. In the ontology, the meta-attribute provides a common name for the same property (here 'loan quality'). We call the 'A-F' and 'bad, good, medium' a *realization* of a meta-attribute.

The ontology supports the pieces of background knowledge introduced in [16]: i) *basic value limits*, ii) *typical interval lengths for discretization*, *important groups of meta-attributes* and iii)*mutual influence among meta-attributes*. One meta-attribute can be assigned background knowledge for multiple realizations.

Schema Mapping allows to align a data field or derived data field used in one data mining task with its counterparts in other tasks. This is done through mapping the corresponding meta-attribute to its realizations. The current ontology version allows to express the mapping only in terms of *equivalence* of data values or categories.

Annotations are the last component of the Data Mining ontology. Annotations are pieces of knowledge that can be assigned by the analyst to some important concepts in the ontology. For example, an annotation can be assigned to an instance of Discretize Bin to explain the reason behind the discretization.

3 Prototype framework

This section describes a reference implementation of the framework described in Section 2.3 for association rule mining method GUHA and its Ferda and LISp-Miner implementations[1]. GUHA method is one of the first methods of exploratory data analysis, developed in the mid-sixties in Prague. It is a general mainframe for retrieving interesting knowledge from data. The method has firm theoretical foundations based on observational logical calculi and statistics [9]. In the following, we describe the GUHA method, the way it is reflected in the Mining Algorithm component of the Data Mining ontology and finally, we briefly describe the technologies behind the remaining parts of the framework.

3.1 GUHA Method

The GUHA method is realized by GUHA procedures, such as 4FT procedure for mining association rules. Inputs of the procedure are data and a simple definition of a set of relevant patterns (defined as formulas or a observational calculus). All the relevant patterns are automatically generated and verified against the provided data. Patterns that are true are the output of the procedure.

Throughout the history, the most used GUHA procedure was the procedure for mining association rules (initially named ASSOC), which was running about 20 years before the invention of classical *apriori* algorithm. The GUHA association rules extend the *apriori*-ones in two ways:

The first way is to enable *Boolean attributes* for antecedent and consequent. *Boolean attributes* are recursive structures that enable conjunctions, disjunctions and negations of combinations of individual items. Details can be found in [20].

The second way is to enable expressing more general kind of dependency between antecedent and consequent then *confidence* and *support*. We call these dependencies *4ft-quantifiers*. The generalized association rule can be written in form $\varphi \approx \psi$, where φ and ψ are *Boolean attributes* and $\approx$ is a *4ft-quantifier*. The quantifier is computed on the basis of *4ft-table*.

A *4ft table* is a quadruple of natural numbers $\langle a,\ b,\ c,\ d \rangle$ so that: a is the number of object from the data matrix satisfying φ and ψ (likewise for other numbers).

The *above average dependence* quantifier as example of such is defined as:

$$\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq Base$$

where p and $Base$ are user-defined parameters. It can be verbally interpreted as *Among object satisfying φ, there are at least 100p per cent more objects satisfying ψ then among all observed objects and there are at least Base observed objects satisfying φ and ψ.*

[1] `lispminer.vse.cz` and `ferda.sourceforge.net`

3.2 GUHA-based Ontology for Association Rules

It is shown e.g. in [15] that the GUHA association rules are generalization of the mainstream association rules and also of the quantitative association rules [21]. Also, majority of the quality measures for association rules such as *lift* or *f-measure* can be interpreted in terms of *4ft-table*. Utilizing these facts we have proposed an association rule mining ontology based on GUHA and the corresponding theory of observational calculi.

The main elements of this ontology are shown on Figure 2. It has been checked that this ontology allows to express the setting and discovered rules for the GUHA algorithm and the most commonly used association rule mining algorithm Apriori. Hence, it can be used in place of the Mining Algorithm component of the Data Mining ontology introduced in subsection 2.4.

3.3 Framework Implementation

The work on reference implementation described in this section gives an overview of what steps are necessary in order to implement the framework described in Section 2. Our GUHA-based prototype consists of the following parts:

Input Data The PMML Association Model was extended to incorporate GUHA features not present in the PMML 3.2. Based on our prior experience with background knowledge [23, 17, 16, 18], the work on BKEF began.

Analytical Report Authoring Support We used the open source Joomla! (`http://www.joomla.org`) CMS and extended with support for XSLT transformations[2]. XSLT transformation from GUHA specific PMML to HTML was defined. The design of transformation from BKEF to HTML is contingent on the BKEF standard being finalized.

Semantization of Analytical Reports The GUHA-based Association Rule Ontology was defined. The Ontopia Knowledge Suite[3] was used as the topic map repository and knowledge base. OKS can be either interactively browsed through using the Omnigator application or queried using *tolog*, querying language based on Prolog and SQL. The work on transformation of PMML and BKEF to this ontology is under progress.

4 Case study: Financial Dataset

The goal of this case study is to evaluate the usability of the prototype implementation in operation and to demonstrate the potential of the framework. In the experiment, we went through the whole process of designing a data mining task, performing it, generating a PMML model, uploading it to the Joomla! CMS, using it to author analytical report, semantizing it, storing it to the OKS knowledge base and finally designing and executing sample tolog queries. We used Financial Data Set, first introduced in PKDD'99 Discovery Challenge [1].

[2] an XML technology for translating between different knowledge representations.
[3] http://www.ontopia.net

4.1 Data Mining Task

The Financial Dataset consists of 8 tables describing the operations of bank customers. Among the 6181 customers, we aim to identify subgroups with high occurrence of bad loans. For the mining schema, we used columns 'duration', 'status' and 'district'; all columns coming from the Loans table.

Since there is no background knowledge specified in the PKDD'09 task setting, the following piece of background knowledge was introduced for case study purposes :

- *Background Knowledge 1* The quality of loans in Central Bohemia, the richest region of the country, is generally good. In other regions, it is in average lower. If area is expressed in terms of cities than value 'Central Bohemia' maps to value 'Prague'.
- *Background Knowledge 2* If loan quality is expressed in terms of values A - D then value *A* maps to *good*, *B* to *medium* and *C,D* to *bad*.
- *Background Knowledge 3* If loan duration is expressed in terms of months then three bins should be created $< 0; 12 >$, $< 13; 23 >$ and $< 24; inf >$.

All these pieces of background knowledge can be input in a structured way directly into the data mining systems: see [16] for BK1 and [23] for BK2 and BK3. For example, using convention introduced in [18], BK1 can be input as a following piece of structured knowledge directly to the data mining software:

$$Region(CentralBohemia) \to^{+} LoanQuality(good). \tag{1}$$

Using BK2 and BK3, the data were preprocessed in the following way: the duration column was discretized into *1 year*, *13-23 months*, *two years+* categories. A *statusAggregated* derived field was created from the status column by mapping status values *A* to *Good*, *B* to *Medium* and *C, D* to category *Bad*. Derived field *district* was created by 1 : 1 mapping from the district column, which has a municipality granularity.

In Ferda Data Miner, we formulated the following task:

$$duration(SS[1-1]) \& district(Praha) \Rightarrow statusAggregated(I[2-2])$$

Here, $duration(SS[1-1])$ means, that all subsets of the attribute of minimal and maximal length equal to 1 are created. $statusAggregated(I[2-2])$ means that intervals of status (viewed as cardinal domain) of maximal and minimal length are created. The used *4ft-quantifier* was *above average dependence*, as explained in section 3.1 with parameter $p = 0.1$.

Among the 7 rules found, the following rules were the strongest:

AR 1: $duration(13-23months) \Rightarrow statusAggregated(medium, bad)$
AR 2: $duration(1year) \Rightarrow statusAggregated(good, medium)$
AR 3: $duration(> 2y) \& district(Prague) \Rightarrow statusAggregated(good, medium)$

4.2 Representing the Knowledge with the Framework

All the knowledge relating to the task and results was then input to the CMS system. The data mining model was input automatically via PMML, the background knowledge was input manually due to ongoing work on the BKEF standard. Using tools offered by the CMS, the analyst authors the report. The analyst could be more productive through reusing the HTML representation of the structured content generated by XSLT.

In the report, the analyst stated that AR 1 is not interesting – despite its strength what concerns quantifier values. Rule AR 2 was, in turn, found surprising and useful. The analysts does not comment on Rule 3. The result of the analysts work is a human readable report, which is interlinked with its source data – primarily the PMML reports. However, since the report is written in free text, the information added by the analyst, such as information pertaining to the novelty of the association rules, is not machine readable.

This is solved by the next step in the framework. The structured content is 'semantized' through conversion to the Data Mining Ontology. The Association Rule Mining Ontology introduced in Section 3.1 was used in place of the Mining Algorithm component of the ontology. The data fields referred by background knowledge were expressed in terms of meta attributes *Area*, *Loan quality* and *Loan Duration* and their realizations *Area [Region]*, *Loan Quality [Alphabetical]* and *Loan Duration [Months]*. Utilizing the schema mapping component of the ontology, these realizations were mapped to data fields used in the mining task.

The resulting topic map was stored in the Ontopia Knowledge Suite (OKS) serving as a knowledge base. The knowledge base allows the analyst to input new pieces of knowledge - machine readable annotations. In this way, AR1 is annotated as 'Not Interesting' while AR2 is annotated as 'Surprising and Useful'.

4.3 The Added Value Of Semantization

The knowledge base allows for sophisticated search with the tolog language. The analyst can choose which 'vocabulary' he/she will use in the query. E.g. he or she can decide between querying in terms of background knowledge or a specific dataset. Two example queries are listed below.

In the first query, the analyst wants to find all discovered rules that are annotated as interesting and have a good loan quality in the consequent. Using schema mapping and meta attributes, the rule found is (as expected) AR 2: $duration(1year) \Rightarrow statusAggregated(good, medium)$.

The analyst's second query uses the tolog's inference engine to find discovered rules, whose antecedent subsumes the background knowledge rule BK 1: $Region(CentralBohemia) \rightarrow^{+} LoanQuality(good)$. The query follows:

```
select $AR from
value($BARNAME, "Region(Central Bohemia) => Loan Quality(Good)"),
topic-name($BAR, $BARNAME),
assoc:hasAntecedent($BAR : role:AssociationRule, $BARANTE : role:Cedent),
assoc:consistsOfLiteral($BARANTE : role:Cedent, $BARLIT : role:Literal),
```

```
assoc:hasCoefficient($BARLIT : role:Literal, $CAT : role:Category),
assoc:hasEquivalentCategory($CATEQ : role:CategoryEquivalency,
$CAT : role:Category),
assoc:hasEquivalentCategory($CATEQ : role:CategoryEquivalency,
$EQCAT : role:Category),
assoc:hasCoefficient($LIT : role:Literal, $EQCAT : role:Category),
assoc:consistsOfLiteral($ANTE : role:Cedent, $LIT : role:Literal),
assoc:hasAntecedent($AR : role:AssociationRule, $ANTE : role:Cedent),
instance-of($AR, topic:DiscoveredAssociationRule)?
```

The result of the query is AR3. Although the analysts has initially failed to notice that AR 3 corresponds to BK 1, the system was able to infer this automatically. The system can perform also other tasks apart from search such as merging and filtering.

5 Related Work

The current work follows up on initial attempts to data mining analytic reporting published in [14]. However, this early work did not explicitly consider ontologies or even mark-up languages such as PMML. We are not aware of any other initiative for sharing data mining results, from any domain, over the semantic web. The impact of semantic web technology on data mining has typically been perceived as shift to 'knowledge-intensive' mining, in which ontologies serve as prior knowledge [10, 5, 6] or as a means to (locally) interpret the results [22]. We also cover this aspect to some degree, although we put more stress on the data integration and inferencing (querying) aspects.

Several times before, ontologies have already been suggested as formal models of the data mining field. The applications were in data mining work flows [10, 6], grid-based data mining [5] or parameterizing a specific mining tool [7]. With the notable exception of the KDD ontology from [11] (which was developed without connection to any concrete application) and the early model [6], the ontologies are relatively poor in expressiveness and conceptual coverage. For example, the DAMON ontology described in [5] is merely a collection of taxonomies (for tasks, methods, algorithms and tools) with a few obvious relations defined among them. The connection to domain ontologies is, in turn, only elaborated in [5]. Most important, however, these KDD ontologies were only applied in the phases of the KDD process before or during the actual mining, i.e. ignoring the problem of analytical reporting.

Furthermore, domain ontologies in KDD were used for data pre-processing [12] or result post-processing [8, 22], but not for result dissemination. To our knowledge, the only attempt to bridge the conceptual gap between a concrete data schema and a pre-existing domain ontology is our own earlier work [22]; however, there the mapping was done rather ad hoc, and consequently was unreliable for subsequent reasoning (it was actually only used for suggesting syntactical 'explanation chains' from the ontology with respect to empirical associations discovered by data mining).

Finally, *PMML* has been used by various subjects in industry and research. However, search and aggregation applications over PMML documents have not been significantly reported. Use of this powerful mark-up language has typically been restricted to model exchange among mining tools. Combination of PMML with truly semantic resources has not been mentioned so far.

6 Conclusion

Intelligent post-processing of information originating and relevant to the data mining process can provide additional insight on the problem being solved compared to the common scenario when the results of data mining software are directly interpreted or deployed and no other processing occurs. The framework proposed in this paper poses, to the best of our knowledge, the first systematic solution to post-processing that is built upon semantic web technologies. This framework builds upon existing and tried standards and technologies such as PMML and content management systems. For the semantic end of the framework, we have proposed a data mining ontology derived from PMML

Using the PKDD'99 Financial Dataset and our prototype framework implementation, we have shown how the framework can ease routine tasks such as authoring an analytical report. Its main strength lies, however, in the possibility to benefit from the querying and data integration capabilities given by the use of semantic web technologies. Using Ontopia Knowledge Suite, a topic map driven knowledge base, we have executed two example queries that used background knowledge, semantic annotation and schema mapping. The knowledge base, demo, and other resources relevant to the prototype implementation are available online[4].

The future work should focus on the development of a standard for exchange of background knowledge. The high potential of background knowledge and semantic technologies for post-processing of data mining results would justify their involvement into data mining methodologies such as CRISP-DM.

References

1. Workshop notes on Discovery Challenge PKDD'99, 1999
2. Kováč M., Kuchař T., Kuzmin A., Ralbovský M.: Ferda, New Visual Environment for Data Mining. Znalosti 2006, Conference on Data Mining, Hradec Králové 2006, p. 118–129 (in Czech).
3. Raspl, S.:An Overview of PMML Version 3.0. In Proceedings of KDD-2004 Workshop on Data Mining Standards, Services and Platforms (DM-SSP 04), 2004, Seattle, WA. pp 18-22
4. Šimůnek M.: Academic KDD Project LISp-Miner. In: Advances in Soft Computing—Intelligent Systems Design and Applications, Springer Verlag 2003.
5. Cannataro, M., Comito, C.: A data mining ontology for grid programming. In: Proceedings of the 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing (SemPGrid2003), 113-134, 2003.

[4] `http://nb.vse.cz/~klit01/sewebar`

6. Žáková, M., Křemen, P., Železný, F., Lavrač, N.: Planning to Learn with a Knowledge Discovery Ontology. In: PlanLearn at ICML/COLT/UAI 2008, Helsinki, Finland, 2008.
7. Suyama, A., Yamaguchi, T.: Specifying and Learning Inductive Learning Systems using Ontologies. In: AAAI'98 Workshop on the Methodology of Applying Machine Learning: Problem Definition, Task Decomposition and Technique Selection, 29–36.
8. Domingues, M. A., Rezende S. A.: Using Taxonomies to Facilitate the Analysis of the Association Rules. In: The 2^{nd} International Workshop on Knowledge Discovery and Ontologies, held with ECML/PKDD 2005, Porto, p.59-66.
9. Hájek, P., Havránek, T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer–Verlag 1978.
10. Bernstein, A., Provost, F., Hill, S.: Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. IEEE Trans. on Knowl. and Data Eng., 17(4):503518, 2005.
11. Panov, P., Džeroski, S., Soldatova, L. N.: OntoDM: An Ontology of Data Mining. In: SADM workshop at ICDM'08, Pisa, Italy, IEEE 2008.
12. Phillips, J., Buchanan, B. G.: Ontology-guided knowledge discovery in databases. In: K-CAP'01, Victoria, Canada, October 2001.
13. Rauch J, Šimůnek M (2005) An Alternative Approach to Mining Association Rules. In: Lin T Y, Ohsuga S, Liau C J, and Tsumoto S (eds) Data Mining: Foundations, Methods, and Applications, Springer-Verlag, 2005, pp. 219–238
14. Rauch, J., Šimůnek, M.: Semantic Web Presentation of Analytical Reports from Data Mining Preliminary Considerations. In: Web Intelligence'07. Los Alamitos : IEEE Computer Society, 2007, pp. 3-7.
15. Rauch, J.: Logic of Association Rules. *Applied Intelligence*, 2005, . 22, 928. ISSN 0924-669X
16. Rauch, J., Šimůnek, M.: LAREDAM Considerations on System of Local Analytical Reports from Data Mining. Toronto 20.05.2008 23.05.2008. In: Foundations of Intelligent Systems. Berlin : Springer-Verlag, 2008, 143149. ISBN 978-3-540-68122-9. ISSN 0302-9743.
17. Dealing with Background Knowledge in the SEWEBAR Project
18. Considerations on Logical Calculi for Dealing with Knowledge In Data Mining
19. Rauch, J., Raś, Z. W., Chudán, D., Jiang, M. W., Kliegr, T., Nemrava, J., Ralbovský, M., Svátek, V., Šimůnek, M., Thompson, P., Zhang, C. X.: SEWEBAR Tinnitus – Project Presentation. In: Znalosti 2009, Annual Czecho-Slovak Conference on Knowledge Technologies, Brno, February 2009, pp. 359–363.
20. Ralbovskyý M., Kuchař T.: Using Disjunctions in Association Mining. P. Perner (Ed.), Advances in Data Mining - Theoretical Aspects and Applications, LNAI 4597, Springer Verlag, Heidelberg 2007
21. Srikant R.,Agrawal R.: Mining generalized association rules. VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, pages 407–419
22. Svátek V, Rauch J, Ralbovský.: Ontology-Enhanced Association Mining. In: Ackermann, M et al. Semantics, Web and Mining. Berlin : Springer, 2006, pp. 163–179. ISBN 978-3-540-47697-9.
23. Zeman, M., Ralbovský M., Svátek V., Rauch J.: Ontology-Driven Data Preparation for Association Mining. In: Znalosti 2009, Annual Czecho-Slovak Conference on Knowledge Technologies, Brno, February 2009, pp. 270–283.