Noname manuscript No.
(will be inserted by the editor)

SEWEBAR-CMS: Semantic Analytical Report Authoring for Data Mining Results

Tomáš Kliegr · Vojtěch Svátek · Martin Ralbovský · Milan Šimůnek

the date of receipt and acceptance should be inserted later

Abstract SEWEBAR-CMS is a set of extensions for the Joomla! Content Management System (CMS) that extends it with functionality required to serve as a communication platform between the data analyst, domain expert and the report user. SEWEBAR-CMS integrates with existing data mining software through PMML. Background knowledge is entered via a web-based elicitation interface and is preserved in documents conforming to the proposed Background Knowledge Exchange Format (BKEF) specification. SEWEBAR-CMS offers web service integration with semantic knowledge bases, into which PMML and BKEF data are stored. Combining domain knowledge and mining model visualizations with results of queries against the knowledge base, the data analyst conveys the results of the mining through a semi-automatically generated textual analytical report to the end user. The paper demonstrates the use of SEWEBAR-CMS on a real-world task from the cardiological domain and presents a user study showing that the proposed report authoring support leads to a statistically significant decrease in the time needed to author the analytical report.

Keywords data mining · association rules · background knowledge · semantic web · content management systems · topic maps

1 Introduction

Presenting results of particular data mining tasks belongs among the most significant problems in data mining research. This problem is perhaps most pressing for the association rule mining tasks, where the inability of current systems to select the interesting rules and present them to the expert in a concise form is considered as the main obstacle to using association rule mining in practical applications.

A long established line of research has been investigating interest measures that can be computed from the analyzed data to rank association rules. Newer research focuses

University of Economics, Prague, Faculty of Informatics and Statistics,
Nám. Winstona Churchilla 4, 130 67 Praha 3, Czech Republic,
Tomáš Kliegr also works at Multimedia and Vision Research Group, Queen Mary, University of London, 327 Mile End Road, London E1 4NS
E-mail: {tomas.kliegr, svatek, ralbovsm, simunek}@vse.cz

on involving domain knowledge for rule pruning or advanced search in discovered rules. The hypothesis behind the research presented here is that the solution to the association rule mining usability problem is not a single data mining algorithm, interest measure or postprocessing algorithm, but rather a flexible system that can be used 1) by a *domain expert* to provide the needed piece of information, 2) by the *data analyst* to plug-in a specific piece of software, and 3) by the *end user* to read the *analytical report* and provide feedback. SEWEBAR-CMS, a data-mining Content Management System introduced in this paper, serves as such a communication hub. The acronym SEWEBAR comes from Semantic Web – Analytical Reports.

Analytical report is a free-text document describing various elements of a data mining (DM) task: particularly the data, preprocessing steps, task setting and results. The data analyst can also include additional information such as background knowledge, explanation of preprocessing steps and interpretation of the results. Analytical reports have previously been created manually; this is however time-consuming and the output document is not machine-readable, which hinders the possibilities for postprocessing – e.g. querying, merging or filtering. SEWEBAR-CMS provides semi-automatic generation and processing of analytical reports that addresses the above issues with the help of semantic web technologies. SEWEBAR-CMS was primarily developed to support the association rule mining task, but its significant part can be reused in other mining tasks.

This paper is a significantly reworked and extended version of [25]. SEWEBAR-CMS is now demonstrated on a running example of postprocessing the results of mining from a real-world cardiological dataset (in contrast to a synthetic dataset in [25]), and a user case study is presented, which quantitatively and qualitatively evaluates the benefits of the system.

The paper is organized as follows. In Section 2 we give an overview of the architecture of the system. Section 3 describes the representation and processing of background knowledge and Section 4 the representation and processing of mining models. Section 5 describes the actual report authoring support of SEWEBAR-CMS, which leverages on previously acquired background knowledge and mining models. Section 6 explains the way in which SEWEBAR-CMS benefits from semantic processing of data mining models and background knowledge. The user case study with evaluation is presented in Section 7. An overview of related work is placed towards the end of the paper into Section 8. Section 9 contains conclusions and a plan for future work.

Throughout the paper we employ a *running example* (its individual parts numbered as Example 1 to 5), which is based on the real-world medical dataset Adamek [48] describing cardiological patients. The dataset consists of two data matrices, each covering patients in one hospital. Each matrix row contains the record of one patient. The data mining task suggested by the domain expert is to discover interesting relationships between various patient features and blood pressure, taking into account existing background knowledge. The analyst uses SEWEBAR-CMS to convey the results of the mining to the end-user (here, presumably, the domain expert herself) through the analytical report.

2 Framework Outline

Postprocessing of data mining results leading to comprehensive analytical reports encompasses integration of information from multiple sources. This involves contribution

and interaction of domain experts and data analysts. A natural environment for this integration and interaction is a Web Content Management System (CMS), an application supporting storage, retrieval and authoring of electronic documents over the web.

There are many existing commercial as well as open-source web-based content management systems. The best known systems are Alfresco (commercial, `Alfresco.com`), Drupal (open source, `Drupal.org`) and Joomla (open source, `Joomla.org`). These systems offer a wide range of functions such as user management or WYSIWYG editing for generic document authoring. To fulfill specific requirements, there are either domain-specific CMS systems or extensions of existing systems. For example, [6] describes functionality requirements on a CMS system for digital library applications.

In this paper, we demonstrate on the SEWEBAR-CMS system which functionalities should a CMS for data mining have and how it should communicate with other software agents. Together with accompanying XML formats and workflow, SEWEBAR-CMS is a part of the SEWEBAR framework [38] for analytical report authoring.

Technically, SEWEBAR-CMS is a set of domain-specific extensions for the Joomla! CMS system. An overview of the SEWEBAR-CMS system and its role in the SEWEBAR framework is given by the Figure 1.

SEWEBAR-CMS integrates with existing data mining software through the *Predictive Model Markup Language* (PMML),[1] which is a widely adopted XML-based standard for definition and sharing of data mining and statistical models. The data mining software sends PMML documents via a web service (Figure 1A).

The input from the domain expert is preserved in documents conforming to the emerging *Background Knowledge Exchange Format* (BKEF) specification. While PMML documents are produced by the DM software, BKEF documents are created directly based on human input. Background knowledge is entered via a web-based elicitation interface that is also part of SEWEBAR-CMS, and is stored in BKEF documents ((Figure 1B)).

The heart of the architecture is the CMS Repository, into which the PMML and BKEF documents are stored, and possibly linked via the so-called *Field Mapping Language* (FML). They can then be rendered to HTML through XSLT scripts, yielding *automatically generated* reports.

The semantic functionality is not directly part of the SEWEBAR-CMS. SEWEBAR-CMS only offers a web service interface for export of PMML and BKEF documents into a semantic knowledge base (Figure 1D) and an interface for querying the knowledge base (Knowledge Base Include, or shortly KBInclude). The 'semantization' essentially consists in transforming the tree-structured XML data into collections of interlinked facts with ontologically-defined semantics.

Based on an HTML visualization of PMML and BKEF documents and results of queries against the semantic knowledge base or other web-enabled structured resources , the data analyst writes an analytical report (Figure 1E). The analytical report is then the 'fine-cooked' result of the mining process.

[1] `http://www.dmg.org/pmml-v4-0.html`

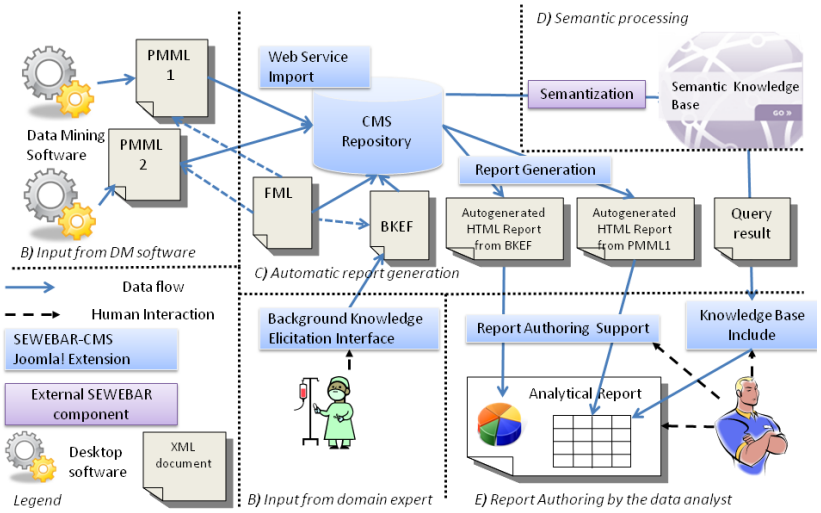

Fig. 1 Framework outline

Knowledge type	Knowledge use	Existing formalisms	BKEF
taxonomy of items	decrease granularity [42]	PMML Taxonomy	Preprocessing Hint
discretization hint	decrease granularity	E.g ontology [52]	Preprocessing Hint
interesting patterns	constrain search space	Rule Schemas [32]	Background ARs /
known patterns	result postprocessing	Rule Schemas [?]	/ Mutual Influences

Table 1 Types of background knowledge in association rule mining

3 Representation of Background Knowledge

Background (or sometimes referred to as domain) knowledge is extensively used in preprocessing of data before data mining. For example, during the setting of an association rule mining task, the data analyst needs to decide which data fields to include and how to preprocess them. Inclusion of redundant or irrelevant fields can increase the mining time and clutter results. A similar effect can have an otherwise relevant field with a high number of distinct field values, unless its granularity is decreased; this applies particularly to numerical fields. The data analyst can determine the task setting experimentally with the help of feature selection and discretization algorithms. However, this information can be also obtained from a domain expert.

Despite the potential of expert-provided background knowledge for improving the quality of data mining results, there has been so far little research effort in the area of selecting pieces of information that should be collected and little standardization efforts on devising a common specification for storing background knowledge. To the best of our knowledge, there is no established standard analogical to PMML for background knowledge in data mining.

Since the availability of a formal specification of background knowledge is vital for the framework, we have proposed the Background Knowledge Exchange Format

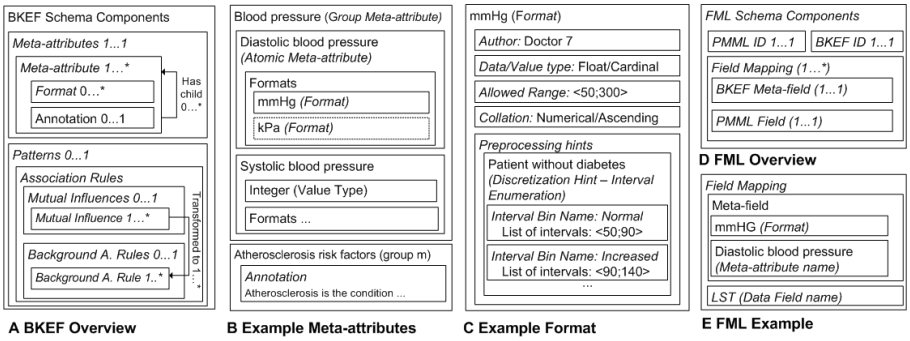

Fig. 2 BKEF and FML overview and examples

(BKEF) XML specification [26]. Table 1 presents an overview of common types of background knowledge related to association rule mining and their coverage by BKEF. The scope of a BKEF document is one *mined domain*. Figure 2A gives a brief overview of the structure of the two main components the BKEF XML Schema consists of:

- **Meta-attributes:** The basic building block of BKEF is a *Meta-attribute* [38]: an abstraction representing a property appearing in datasets from the given domain. The definition of meta-attributes can be used e.g. to automate data preprocessing as proposed in [38] or implemented in [52].
- **Patterns:** Known (confirmed), interesting (suspected to hold, but not confirmed) and rejected (proven not to hold) patterns in the mined domain. Patterns can typically be used for postprocessing the already discovered hypotheses.

A detailed description of the BKEF XML Schema is out of the scope of this paper (see [26] for more information). The following two subsections focus each on one of the main types of BKEF background knowledge, and demonstrate their various aspects on examples.

3.1 Background Knowledge: Meta-attributes

Meta-attributes can be nested to an arbitrary number of levels. A meta-attribute with no child is referred to as *Atomic Meta-attribute* and its granularity should correspond to a column in a data matrix (corresponding to a PMML's Data Field). Non-atomic Meta-attributes are *Group Meta-attributes*. A Group meta-attribute can have multiple children.

Since a property can be sometimes measured in different ways, most commonly using different units, BKEF allows each meta-attribute to have multiple *formats*. Most pieces of information relating to a meta-attribute are format-dependent. A format contains the following pieces of information:

- **Data and Value Type** specifies the format's data type (integer, float,...) and value type (cardinal, nominal or ordinal).
- **Allowed Range** specifies permissible values through an interval or enumeration.
- **Preprocessing Hints** encompass particularly expert recommendations for discretization or value mappings.

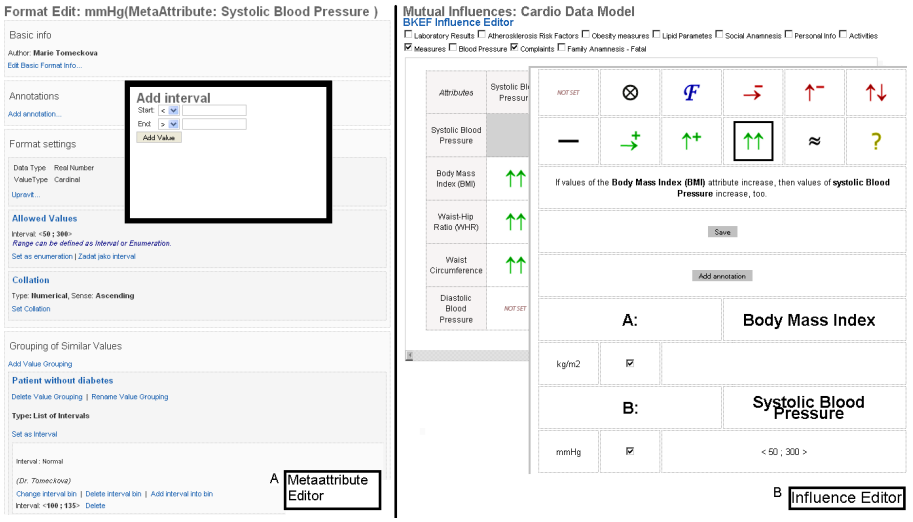

Fig. 3 Screenshots from SEWEBAR-CMS

- **Collation** specifies how to sort values, this includes enumeration-based collation suitable for ordinal formats and numerical collation suitable for cardinal formats.
- **Value Annotation** allows to annotate selected values or value ranges with a textual annotation and a machine-readable class.

It is convenient to introduce a name for an (implicit) superclass for the analogy with *Field* and *Field Value* elements in PMML (see Section 4): *Meta-field* is a format-meta-attribute pair. *Meta-field Value* is an abstraction of a possible 'value' of a meta-field – Discretize Bin, Value Mapping Bin in any of the *Preprocessing Hints* or an interval or a Value from the range specified by the Allowed Values element.

Similarly as in PMML, a concrete BKEF producer can introduce custom pieces of information (e.g. statistics such as standard deviation) using the XML extension mechanism.

In SEWEBAR-CMS, elicitation of meta-attribute knowledge contained in BKEF is done by the *Metaattribute Editor*, a Joomla! Extension. The application is wizard-based. On the first screen the user can create a new (Group/Atomic) meta-attribute or edit an existing one. The *Atomic Meta-attribute Edit screen* allows to create a new Format or edit an existing one, while the *Group Meta-attributes Edit screen* allows to (un)assign child meta-attributes. In the *Format Edit Screen* (see Fig. 3A) the user can set to a Format all the information according to the BKEF XML schema.

Example 1: BKEF Meta-attributes

To obtain background knowledge for the Adamek dataset, a medical expert was asked to use the BKEF *Metaattribute Editor* to input her knowledge. Assisted by the data analyst, she input information on 32 important clinical parameters appearing in the dataset, thus creating 32 *atomic meta-attributes*, each with one format, organized into 10 *group meta-attributes* (see Figure 1B).

Consider for example the group meta-attribute *Blood pressure* (see Figure 2B). In the dataset there are four data fields relating to blood pressure: right-hand systolic,

right-hand diastolic, left-hand systolic and left-hand diastolic. Abstracting away from the left/right hand information, the expert introduced two atomic meta-attributes for systolic and diastolic blood pressure and assigned them as child meta-attributes to the Blood pressure group meta-attributes.

Since blood pressure in Adamek is measured in millimeters of mercury (mmHg), the expert introduces the *mmHg* format for each of these meta-attributes. Within the format, additional information on diastolic/systolic blood pressure such as the preprocessing hint is conveyed. Figure 2C shows an example format.

As the *mmHg* unit is used for historical purposes[2], the blood pressure meta-attributes could be extended by the kPa Format as hinted in Figure 2B.

3.2 Background Knowledge: Patterns

Known relationships between meta-attributes are captured with BKEF Patterns. The current BKEF XML Schema allows to define two types of patterns: *Mutual Influences* and *Background Association Rules*.

The notion of *Mutual Influence* (MI) comes out of research by Rauch & Šimůnek [38], who proposed to use this construct to represent relationships between meta-attributes. Each mutual influence is assigned a *Validity* to denote its meaning. Eleven types of mutual influences and three types of validity, *unknown*, *refuted* or *proved*, were proposed. Mutual influences have graphical representation ('arrows' in Fig. 3B) that is easy to understand for domain experts. Here, we only briefly describe one of the simplest and perhaps the most common mutual influence of the type $A \uparrow\uparrow B$, which expresses that if values of meta-attribute A increase, values of meta-attribute B increase too.

Background Association Rules are GUHA-like association rules [36], but they are defined over meta-fields rather than over data fields or derived fields. A more formal definition is in [26]. A Mutual Influence can be transformed into *Background Association Rules* (BAR) of the form $A(\omega_A) \approx B(\omega_B)$, where $A(\omega_A)$ and $B(\omega_B)$ are *Basic Boolean Meta-attributes* [36]. The *coefficient* ω_X is a subset of values of meta-field X that are perceived as 'high'. If the BAR emerges from an MI then it is called Atomic Consequence of this MI [38].

In SEWEBAR-CMS, mutual influences are entered by the domain expert through the *Influence Editor* [11].

Background association rules can either be generated automatically from the mutual influence based on Format collation, which is necessary for interpreting the notion of 'high', or input manually by the domain expert through the SEWEBAR-CMS *Association Rule Builder* [**?**].

Example 2: BKEF Patterns

To obtain the second part of background knowledge for the Adamek dataset, a medical expert was asked to use the SEWEBAR-CMS Influence Editor to formalize her knowledge of patterns appearing in this domain. Some of the 75 MI patterns are listed:

1. MI1: *BMI* $\uparrow\uparrow$ *Systolic Blood Pressure (Validity: Confirmed)*

[2] Current sphygmomanometers mostly do not use mercury. Some newer devices already give readings in kilopascals (kPa), the SI measure of pressure.

2. MI2: *BMI* ↑↑ *Diastolic Blood Pressure (Validity: Unknown)*
3. ...

For example, the Mutual Influence 2 captures a relationship that, by the expert's opinion, needs further confirmation. It can be rewritten to several Atomic Consequences (setting the above-average dependence threshold to 0.2 and absolute support to 20 based on experience [38]). In the current SEWEBAR implementation, this transformation is carried out manually, while constraining the coefficients to length one.
MI2 is transformed to:

1. AC21: *BMI(Increased)* $\Rightarrow^{+}_{0.2,20}$ *DBP(Increased)*
2. AC22: *BMI(Overweight)* $\Rightarrow^{+}_{0.2,20}$ *DBP(Increased)*
3. ...

The background knowledge elicited from the expert is saved into a BKEF document and stored in the CMS. Atomic Consequences stored in the document can be used to postprocess mining results as discussed and exemplified in Section **??**.

4 Representation of Data Mining Models

4.1 Formal Representation of Data Mining Models

The largest body of input for the framework is constituted by descriptions of data mining models: settings and results of data mining algorithms running in arbitrary software environments. Concise and detailed model descriptions are crucial for postprocessing of data mining results. Obtaining these descriptions must not be, however, too costly in terms of requirements on export functionality of the data mining software.

For this reason, the framework adopts PMML, a holistic and widely-adopted XML-based standard for definition and sharing of data mining and statistical models [**?**]. PMML 4.0, the latest version of the standard at the time of writing, has the following components (second-level XML elements):

- **Header** contains task metadata such as the version of the DM application.
- **Data Dictionary** describes the input data by listing available *Data Fields* and optionally describing their content through enumerating *Values* or listing *Intervals* of permissible values.
- **Transformation Dictionary** (optional) describes the preprocessing of input data fields: typically a mapping or discretization of multiple *Data Field Values* onto one *DiscretizeBin* or *MapValues* Bin. Preprocessed values are stored in a *Derived Field*.
- **Mining Model** contains the definition of zero or more mining models (typically one). PMML 4.0 defines 13 different mining models (e.g. Association Rules, Cluster Models and Neural Networks) as reusable elements with their own unique structure. Most models begin with *Mining Schema*, which lists Mining Fields – fields used in the model, and the treatment of outlying, missing and invalid values.
- **Mining Build Tasks** (optional) contains the setting related to the training run that produced the model. Its content is not elaborated in PMML 4.0.

Since the input for the mining algorithm can be constituted either by a raw *Data Field* or a preprocessed *Derived Field*, we will used the term *Field* further in the text to denote either of the fields when referring to PMML. Similarly, the term *Field Value* will be used to denote a possible value (expressed in terms of the Data Field's Interval or Value or Derived Field's Discretize Bin or Value Mapping Bin) of a Field.

Gender	Psychical Strain	Uric Acid	BMI	Diastolic Pressure
female	small	359	45	105
female	medium	missing	31	85
female	high	320	26	90

Table 2 Adamek Data Matrix Excerpt

Example 3: Mining Models and PMML

The data matrix Adamek-Hospital1, a fragment of which is depicted in Table 2, describes 684 outpatients, each in one row.

The data analyst uses the LISp-Miner software (`lispminer.vse.cz`), which is an implementation of the GUHA ASSOC procedure [22] mining for GUHA association rules. GUHA association rules are an extension of classical Apriori-style association rules introduced in [4]. The extra features exploited in this example is the *above-average dependence interest measure* $\Rightarrow^{+}_{q,Base}$ used instead of the confidence and support and so-called *basic boolean attributes* used instead of items.

The above-average dependence interest measure can be verbally interpreted as *Among patients satisfying the antecedent, there are at least 100q per cent more objects satisfying the consequent than among all observed objects, and there are at least Base observed objects satisfying both the antecedent and consequent.* In GUHA, the term *item* in the apriori style association rule is replaced by the term Basic Boolean Attribute $b(\sigma)$, where the *Coefficient* σ is a subset of possible Values of Field b.

The task setting[3] was as follows: *above-average dependence* with thresholds $q = 0.2$, $Base = 5$; *data fields* Body Mass Index (bmi), Family status, Psychical strain, Total cholesterol (tchol), Uric Acid (uacid), Height and Gender were considered for *antecedent* and Diastolic blood pressure (dpres) for *consequent*. For mining, we only considered patients without diabetes, which then appears as fixed *condition* (syntactically, in the end of every rule). As a consequence, the *preprocessing hint/format* depicted in Fig. 2C) could be applied in a straightforward way, yielding categories such as 'Normal' or 'Increased' for Blood Pressure.

The task was finished in 1 second on a desktop PC and resulted in 450 discovered association rules.

Sample discovered rules (referred to later in the running example):

AR 1: *tchol(increased)* $\wedge$ uacid(increased) $\wedge$ *bmi(overweight..obesityII)* $\Rightarrow^{+}_{0.92,31}$ *dpres(increased) / diab(no)*
AR 445: *bmi(normal)* $\Rightarrow^{+}_{0.33,145}$ *dpres(normal) / diab(no)*
AR 450: *bmi(obesityII, obesityIII)* $\Rightarrow^{+}_{0.71,35}$ *dpres(increased) / diab(no)*

The resulting data mining model is exported into PMML and sent via a web-service to SEWEBAR-CMS.

[3] For completeness, additional setting (for explanation refer to [39]) was: a) *Coefficients*: Family status: Subset, length 1-2; BMI: Interval, length 1-3; all other: Subset, length 1-1, b) *Cedent setting*: conjunction with minimum length 0; for condition the minimum length was 1.

5 Analytical Report Authoring Support

SEWEBAR-CMS provides a single access point for information relating to the data mining task, scoping particularly the structured PMML and BKEF documents and pieces of unstructured background information.

The analytical report is written by the data analyst based on these pieces of evidence. The task of writing the analytical report is simplified by automatic visualizations of the structured content stored in the CMS in the form of human-readable reports. These *automatically generated reports* contain visualizations of information contained in PMML and BKEF using histograms, tables and automatically-generated text, but are still too verbose to be presented to the domain expert. It is important that these automatically-generated HTML reports are split into machine-readable fragments (individual tables, figures,..), which can be further reused in (custom) analytical reports.

Automatic report generation in SEWEBAR-CMS is realized using an *XSLT Transformation* Joomla! Extension and XSL transformations from PMML to HTML and from BKEF to HTML. The fragments can be nested and are marked with XML comments. Comments were chosen for practical purposes as they are not removed or garbled by off-the-shelf HTML editors.

Analytical report authoring support also comprises a Joomla! Extension called *gInclude*, which allows the analyst to include the fragments of automatically generated reports into the analytical report.

The analytical report can contain fragments of multiple automatically generated reports. Since included fragments retain information about their originating source XML document, the *gInclude Update*, a Joomla! Component extension, can be used to selectively update the analytical report if the referenced documents change.
A part of an automatically generated report is depicted on Fig. 3C. For details on report authoring support in SEWEBAR-CMS refer to [11].

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

Example 4: Report Authoring

After the data mining was completed in Example 3, the data analyst wants to convey its results through the analytical report. The BKEF document created in Examples 1 and 2 as well as the PMML document uploaded in Example 3 to the CMS can be accessed as automatically generated HTML reports. The first report contains background knowledge conveyed by the domain expert and the second all information related to the task. While writing the analytical report, the analyst tries to identify only the pieces of information important for the user. For example, s/he does not need to input details on data preprocessing, because it was done according to the domain expert's recommendation; e.g. the expert user can be generally assumed to be knowledgeable about the fact the *increased* value of diastolic blood pressure refers to the interval $<90; 140>$.

The most difficult task is the identification of discovered rules that may be potentially interesting to the report user. Referring to the excerpt of mining results shown in the example given in Example 2, the data analyst sees that the rule AR1: *tchol(Increased)* $\wedge \ldots \wedge$ *bmi(Overweight..ObesityII)* $\Rightarrow^{+}_{0.92,31}$ *dpres(increased) / diab(no)* is a confirmation of the mutual influence MI2, as every object complying to AR1 will

comply to the atomic consequence AC21: *BMI(Overweight)* $\Rightarrow^{+}_{0.2,20}$ *DBP(Increased)*. The data analyst can therefore dismiss this rule as uninteresting and exclude it from the analytical report. As we will see in the next section, identification of un/interesting rules can also be carried out automatically, through the semantic knowledge base.

6 Semantic Knowledge Base Integration

Semantic representation of mining models and background knowledge can be used to postprocess the discovered patterns in the light of the background knowledge. Particularly, the following uses are envisaged:

- **Semantic annotation:** The entire report can be for example annotated by its subject area and interestingness, and interlinked through annotation to related reports. The annotation can also be done at the granularity of individual patterns as shown in [25].
- **Semantic search:** Queries can be executed against multiple semantized PMML and BKEF documents linked by semantized FML.

The role of the CMS in the SEWEBAR framework is to support all elementary operations relating to design and retrieval of analytical reports. The technological architecture of most current CMS systems including Joomla! is adequate for this purpose – they have a relational database back-end and are written in low-level (from a semantic perspective) scripting languages.

However, some tasks, which are difficult to achieve in relational representation and imperative languages, can be naturally achieved using semantic web technologies that utilize graph data representation and declarative languages. While most CMS systems are written in PHP, almost all semantic repositories such as Sesame (`www.openrdf.org/`) or Jena (`jena.sourceforge.net`) are Java-based. Instead of trying to include semantic functionality into the CMS, which would result in low-level cross-platform technological integration, the SEWEBAR framework opts for loose web-service integration, where the processing of semantic information is outsourced into a separate system; a *Semantic Knowledge Base*.

SEWEBAR-CMS exposes the structured information (PMML, FML and BKEF documents) to the Semantic Knowledge Base through a web service. Likewise, posing queries from within the SEWEBAR-CMS is taken care for by the *Knowledge Base Include* (KBInclude) Joomla! Extension, which also communicates with outside repositories/knowledge bases via web services.

KBInclude consists of three components.

- Through the administration component the admin user defines queries, which are locally remembered and parameterized, and XSLT transformations, which are used to visualize the results of the query against the knowledge base.
- The WYSIWYG editor-plugin component allows the user to include the query into CMS documents.
- Finally, the content component is called during the rendering of the document page and ensures that queries included in the page get executed and their results embedded into the page as HTML fragments.

KBInclude was tested against the TMRAP interface of Ontopia Knowledge Suite, a SPARQL endpoint, and a custom RESTful wrapper for the Berkeley XML database.

SEWEBAR-CMS also includes an experimental GUI-based query designer for association rules.

The specific semantic representation of BKEF, FML and PMML documents is out of scope of the SEWEBAR-CMS system (see [26]). Within the first integrated prototype of the SEWEBAR framework, the ISO/IEC 13250 Topic Maps [19] standard is employed. PMML and BKEF documents are transformed into the Association Rule Mining Ontology (ARON) [24] and Background Knowledge Ontologies (BKON) [26]. In the SEWEBAR framework, we use Ontopia Knowledge Suite (OKS) as the Topic Maps knowledge base tool. We chose OKS, because it is a commercial-grade software with many deployments. The Ontopia Knowledge Suite was in 2009 open sourced. The transformation is done with an XSLT transformation (BKEF to BKON), series of update/insert *tolog* queries or via a Topic Map API.

Tolog is a Topic Map query language inspired by Datalog (a subset of Prolog) and SQL. Unlike XQuery or SQL, tolog supports inference rules, which can help deduce implicitly stated relationships between topics. A partial advantage of tolog over SPARQL as its RDF counterpart is better support of recursion, which was critically needed for the nested structure of GUHA association rules; this was actually one of the reasons for opting for Topic Maps as the first prototype's knowledge representation.

The Topic Map with semantized mining results and background knowledge can then be searched using tolog for discovered rules that are in an *interesting relation* to background knowledge patterns. The ability to traverse from background knowledge to a mining model within one query is dependent on the existence of FML mapping between metafields and datafields.

Detailed discussion of tolog queries and topic map inference in SEWEBAR is out of the scope of this paper, more details can be found in [24].

Example 5: Semantic Knowledge Base

In this example, we will use the tolog language to query for rules discovered from the Adamek-Hospital1 matrix (see example in Section 4) that confirm the background knowledge rules following from the Mutual Influence *BMI* $\uparrow\uparrow$ *Diastolic Blood Pressure*. The mapping between the cardiological background knowledge and the Adamek data matrix is described using an FML document (see Figure 2E).

First, it was necessary to semantize the PMML, BKEF and FML. In our small experiment this was done in the OKS Ontopoly editor by populating the corresponding Topic Map ontologies with instances. Next, the resulting three Topic Maps were merged.

The merged Topic Map can be queried with tolog. Consider the task of automating the search for background association rules that are specialization of a mutual influence as shown in Example 4. If an FML mapping between *bmi* datafield and *BMI* metafield and the *dpres* datafield and *Diastolic Blood Pressure* metafield is defined, the fact that AR1 confirms MI2 can be determined using a tolog query (in contrast to manual check presented in Example 4). Details of the tolog query and the underlying topic maps can be found in [24].

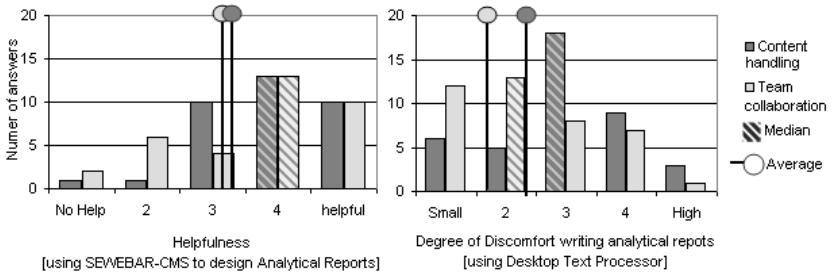

Fig. 4 Framework benefits as reflected by the questionnaire

7 Evaluation

In this section we present an evaluation of SEWEBAR-CMS from the perspective of the data analyst who authors the analytical report. The goal of the evaluation was to assess whether the Report Authoring Support toolset has an impact on the time required to author the analytical report and on its quality.

To achieve this objective, we introduced SEWEBAR-CMS into an undergraduate knowledge engineering course at UEP. One of the long established assignments in the course practicals is on data mining. Teams of 3-4 students use the LISp-Miner data mining system (or the Ferda Data Miner system, which has very similar functionality), and subsequently wrote an analytical report describing the data mining task and results. These reports are then assessed by teachers. In previous semesters, the report was written in a text processor (mostly MS Word).

In the winter semester 2009/2010 there were 44 student teams spread around 10 practicals lead by 6 teachers. The mined dataset was the Adamek dataset. We split the teams into two groups: Group W used MS Word and Group S used SEWEBAR-CMS for report authoring. All teams within one practical belonged to the same 'technology' group. The practicals were distributed among the teachers such that 4 of the 6 teachers had both a practical with group W and a practical with group S, to minimize the impact of the teaching & examination style of a particular teacher. The students were unaware of taking part in an evaluation study (which was generally designed so as to eliminate the possible impact on final assessment), additionally Group W was unaware about SEWEBAR-CMS. Group S received approximately 30 minutes of training on the usage of SEWEBAR-CMS.

After the teams handed in the assignments and these were assessed, online anonymous questionnaires were distributed to all team members, i.e. about 200 students, of which 182 completed[4] the course (either successfully or unsuccessfully) and thus can be considered as relevant respondents. 89 answers sets (i.e. from 49% of relevant respondents, under the unique provenance assumption; the students were strongly encouraged to only use the questionnaire once) were collected. The first part contained questions on the interestingness and time requirements of individual assignments in the course, and is of limited relevance for the purpose of this paper. The second part of the questionnaire, containing four questions (plus fields for free-text comments, as discussed later), was only filled in by students who indicated that they participated in

[4] The remaining students either did not attend the course at all, or left it early in the semester, and thus did not have competence to answer the questions.

the data mining task.[5] There were 33 Group S members and 41 Group W members (i.e. 74 out of 89) who filled in the second part. Additionally, two answer sets from each group were removed as outliers,[6] yielding 31+39=70 answer sets in total.

The first two questions inquired for the respondents' estimates of the total time spent by their team *mining* the data and *writing* the report. The average estimated time to *write* the analytical report in Group S was 8.49 hours and in Group W 10.12 hours. Using a one-sided test of equality of two normal population means with known variances, the null hypothesis can be rejected in favour of the alternative hypothesis that the time to author an analytical report is smaller in Group S than in Group W at the level of significance $\alpha = 0.05$.

The second two questions investigated the degree to which the respondents perceived team collaboration and content handling as a *benefit* of the SEWEBAR-CMS (Group S) or *drawback* of MS Word (Group W), respectively; the answers were graded on a discrete scale 1–5. For overview of answers refer to Fig. 4. It should be noted that while effective content handling support requires specialized tools, such as the report authoring support in SEWEBAR-CMS, the team collaboration could merely be improved by using a generic web-based collaborative text editor. The results from Group W however indicate that for the MS Word users the lack of content handling support caused greater discomfort (median 3) than the deficiencies in team collaboration facilities (median 2). On the other hand, Group S evaluated the collaborative benefits brought by the framework and the benefits from the Report Authoring Support; the median of answers on both question was 4, while the averages only differ by 0.2. This indicates that the Report Authoring Support in SEWEBAR-CMS was accepted by the users and that it probably contributes to the more efficient analytical report design.

Finally, we also noted that the nature of textual comments provided by students as 'complaints' about the technological side of the work differed between Group W and Group S. While the complaints in Group W were mainly related to inherent aspects of the manual report writing process in the text processor, many of the complaints in Group S concerned software bugs in SEWEBAR-CMS, which are an expectable feature of a brand-new software tool and can be relatively easily eliminated.

Additionally, we evaluated the quality of reports from both groups based on the number of points each of the reports received in the course assessment. Importantly, at the time of assessing, the assessors did not yet anticipate that the points would be used for comparison of the two systems afterwards. The maximum number of points was 17; the average number of points in Group S was 13.20 (21 reports) and in Group W 13.17 (24 reports). The tight correspondence of averages conforms to the educational imperative that the study should not introduce a bias to the assessment by significantly favouring one of the groups. However, by inquiring those teachers who assessed the reports by both the Group W and Group S, we saw that the reasons for negative assessment typically differed across the groups. While Group W reports frequently contained data inconsistencies, obviously introduced by manual editing (ranging from sloppy copy-pasting to mixing of data from entirely different mining sessions), the Group S reports were rather penalized for lack of added value from the students themselves (beyond the output automatically generated from PMML). The latter issue, lack of effort investment into a report that 'already looks neat', is presumably paradigmatic

[5] All tasks were assumed to be accomplished by team work, but it was not strictly enforced.

[6] As outliers we considered points located more than 1.5 interquartile ranges below the 1st or above the 3rd quartiles.

for the educational environment; it would no longer hold in a business environment, where the analyst would probably invest the time saved through automatic support into providing further personal insights in textual form.

It can be concluded that the evaluations showed that compared to the baseline scenario, the SEWEBAR-CMS framework allows to design analytical reports in smaller time, while the quality of the reports remains unaffected. The evaluation did not take account of the possibility to further process the *machine-readable information*, which is a 'by product' of the report authoring process in SEWEBAR-CMS.

The evaluation of the effectiveness of using background knowledge patterns (specifically of Rule Schemas and Implicational Rule Schemas) for pruning and filtering discovered association rules has already been presented in [32, **?**]. However, once the integration of SEWEBAR-CMS with OKS is finalized, we would like to perform the evaluation of the benefits perceived by the *domain expert* from automatic rule filtering and pruning.

8 Related Work

Our approach to knowledge-intensive web-centric data mining report authoring seems rather unique in its coverage of different types of knowledge taken into account in the whole KDD process. As its only directly comparable counterpart we identified the quite recently announced extension of the VIKAMINE system [9], which integrates data mining with a (semantic) wiki environment somewhat analogously to our CMS-based environment, leveraging on background knowledge sources. Their workflow looks even more web-centric than (the current version) of ours, as the wiki environment is used both to launch the data mining tasks and to present the results, thus closing the loop; our framework, in contrast, leaves mining task parametrization and launching to the specific tools themselves (with their heterogeneous and possibly evolving interfaces), and only takes care of elicitation of background knowledge, whose representation is presumably more stable in long term. A minor disadvantage of the approach in [9] could be reliance on proprietary knowledge formats only, while we attempt to use as much of the PMML industrial standard as possible, and also seek genericity in the newly-proposed BKEF XML Schema format. The description of the approach in [9] however does not reveal many details on the implementation and end-user functionality. We are in contact with the authors of [9] and envisage thorough comparison (and, presumably, cross-fertilisation) of both approaches in the nearest future.

As regards the use of background knowledge in KDD, its importance has been recognised from the beginning. While large-scale industrial projects merely focused on methodologies of leveraging the human expertise in different phases of the KDD cycle [27], academic research has, as early as at the beginning of 90s, occasionally attempted to formalise and subsequently automatically exploit such knowledge. First attempts to exploit prior conceptual[7] knowledge in propositional[8] machine learning (as research

[7] In this survey we omit approaches that consider background knowledge in numerical form, such as prior probability estimates or expertise-driven parameter setting for mining tools.

[8] We also omit knowledge-intensive, computationally costly approaches to learning over first-logic representation, such as Inductive Logic Programming, where prior background knowledge is an indispensable part of the learning process. These approaches have never penetrated industrial data mining except for very specific, inherently structural task settings such as those in molecular biology.

field predating present-day mainstream KDD) were often restricted to intra-attribute value (typically, taxonomical) structuring [5,8,31,43]. More sophisticated and abstract knowledge models were however sometimes also used to constrain the search and structure the learning workflow; examples are qualitative models by Clark & Matwin [14] or problem-solving methods [17,46].

This effort naturally intensified with the rise of semantic web technologies, providing standard, web-oriented languages and reasoning tools for ontological knowledge (in particular, in OWL [1] and Topic Maps [19]). The research on applying ontologies as prior knowledge in data mining is nowadays split into two rather disjoint streams.

First, domain ontologies are used to specify the semantics of individual data features as well as (known, expected, impossible etc.) relationships among them. Quite naturally, association rule mining is frequent beneficiary of such approaches, due to the inherent similarity of association hypotheses to 'relational' elements of ontological representations. For example, Antunes [7] used domain knowledge as constraints when joining itemsets; Domingues&Rezende [16] used domain taxonomies to generalize the hypotheses and to decrease their number; Tseng [49] pruned itemsets that were redundant wrt. an is-a or part-of hierarchy; Kuo [29] constrained the mining to pairs of features that were (based on their type) meaningful in the context of a medical ontology; Coulet [15] applied biomedical ontologies (namely, information on subsumption and on functionality of properties) for pre-pruning of both columns and rows in source data. In our own prior work we also attempted to directly embed (through metamodeling) background knowledge on attribute categorization and grouping into domain ontologies represented in OWL [52]. In frequent subgroup mining, as neighbour field to association mining, background knowledge on causality was used to discover causal links among subgroups [10]. In our SEWEBAR project itself, an earlier established alternative thread aims at tighter integration between background knowledge representation and the actual data mining engine (LISp-Miner) [38], while the nature of background knowledge is largely overlapping with that of BKEF; particular attention is also paid to logical consequence computation between background knowledge and discovered association rules [37]. In all these (as well as the 'rule schema' methods introduced later), possibly sophisticated knowledge processing is not coupled with standardized data mining model representations and web-based report authoring support, as in SEWEBAR-CMS.

Similar coverage to Background Association Rules used in our framework have so-called Rule Schemas, used by Olaru [32] for focusing the search for hypotheses (primarily carried out in the neighborhood of schemas). A Rule Schema has the form *rs(Condition → Conclusion [General]) [s% c%]*. Condition and Conclusion express what should appear in the rule antecedent and consequent, while the expression in the General section can appear anywhere in the rule. The optional [s% c%] values indicate minimum values of support and confidence, the two rule interest measures used in *apriori*-like association rules. The main difference between Rule Schemas and Background Association Rules is the presence of the General part in the former and the possibility to use other interest measures than confidence and support in the latter.

Besides association mining, prior knowledge was also applied e.g. in clustering [50] (as must-link and cannot-link constraints with concrete instance pairs) and decision tree mining [30] (as 'beliefs' serving for on-the-fly reweighting of attributes). In [33], 'common-sense' ontologies were used as basis for construction of new features.

Second, ontologies are used as means for automatically selecting data mining tools and constructing data mining workflows. The subject of modeling is thus the KDD do-

main itself. A few pioneering approaches appeared a decade ago [12, 18, 45]. Nowadays the field is quite vivid, as witnessed e.g. by the number of 'KDD ontology' projects presented at the ECML/PKDD 2009 workshop on 'service-oriented knowledge discovery' (SoKD, [34]). To our knowledge, however, none of these ontologies significantly covers the postprocessing phase of KDD, incl. the impact of prior domain knowledge on the exploitation of discovered hypotheses, as is the target of our Data Mining Ontology.

9 Conclusions and Future Work

The SEWEBAR framework introduces, to the best of our knowledge, the first systematic solution to supporting data mining report authoring by a web-centric system also exploiting semantic web technologies. The framework is built upon proved standards and technologies such as XML technologies, a popular open source content management system and a commercial grade Topic Map knowledge base. The principal input format is the industry standard PMML specification, which should foster adoption of the framework among data mining practitioners. The data mining ontology used by the semantic components of the framework is designed with respect to this standard as well.

On a continuous example on a medical dataset, we have shown how the SEWEBAR-CMS eases the routine task of 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 a topic-map-driven knowledge base, we have shown an example query that searched for confirmations of a background knowledge pattern in the mined results.

Since the input of the framework is constituted by PMML, the prototype SEWEBAR implementation can be easily adapted to consume results from other DM tools such as Weka or SPSS.[9] The ontology, Joomla! extensions, XML schemas, and other resources are available online at `http://sewebar.vse.cz/`.

Ongoing work focuses on integrating SEWEBAR-CMS with the OKS Knowledge Base. Once the implementation of generation of ontological instances from XML data has been finished, it will be possible to perform a thorough empirical evaluation of the benefits and computational complexity of large scale rule pruning and filtering. The future theoretical research should focus on extending the framework to algorithms mining for other representations than association rules. We are also working towards supporting RDF/OWL as an alternative knowledge representation formalism.

Acknowledgment The work described here has been supported by Grant No. ME913 of Ministry of Education, Youth and Sports, of the Czech Republic, and by Grant No. 201/08/0802 of the Czech Science Foundation, and by Grant No. IGA 21/08 of the University of Economics, Prague. We would like to thank Marie Tomečková, who gave us a valuable feedback on the expert elicitation interface, and the following colleagues who significantly contributed to SEWEBAR-CMS: Jakub Balhar, Vojtěch Jirkovský, Jan Nemrava, Stanislav Vojíř and Jan Zemánek. Last, but no least, we would like to thank teachers at the University of Economics, Prague, who devoted their time to the evaluation of the framework in the educational context.

[9] XSLT transformations need to be customized to fit the required PMML Mining Model and the possible DM tool's extensions to PMML.

References

1. OWL Web Ontology Language Overview. W3C Recommendation, 10 February 2004. http://www.w3.org/TR/owl-features/
2. W3C: XSL Transformation. Online: www.w3.org/TR/xslt. 1999
3. DMG: PMML 3.2 Specification, Online: http://www.dmg.org/pmml-v3-2.html
4. Agrawal R., Imielinski T., Swami A. N.: Mining Association Rules between Sets of Items in Large Databases. In: SIGMOD. June 1993, 22(2):207-16
5. Almuallim, H., Akiba, Y. A., Kaneda, S.: On Handling Tree-Structured Attributes in Decision Tree Learning. In: Proc. ICML 2005, Morgan Kaufmann, 1220.
6. Amato, G., Gennaro,C., Savino, P., Rabitti, F.: Functionalities of a Content Management System specialised for Digital Library Applications. AVIVDiLib05, 7th International Workshop of the EU NoE DELOS on Audio-Visual Content and Information Visualisation in Digital Libraries, Cortona, Italy, May 4-6, 2005
7. Antunes, C.: Mining Patterns in the Presence of Domain Knowledge. ICEIS (2) 2009: 188-193.
8. Aronis, J.M., Provost, F.J., Buchanan, B.G.: Exploiting Background Knowledge in Automated Discovery. In: Proc. SIGKDD-96.
9. Atzmueller M., Lemmerich F., Reutelshoefer J., Puppe J.: Wiki-Enabled Semantic Data Mining - Task Design, Evaluation and Refinement. In: DERIS2009, Design, Evaluation and Refinement of Intelligent Systems, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-545/
10. Atzmueller M., Puppe F.: A Knowledge-Intensive Approach for Semi-automatic Causal Subgroup Discovery. In: Knowledge Discovery Enhanced with Semantic and Social Information. Studies in Computational Intelligence, Volume 220/2009, Springer 2009.
11. Balhar, J., Kliegr, T., Stastny D., Vojir S.: Elicitation of Background Knowledge for Data Mining. In: Znalosti 2010, Jindrichuv Hradec, Czech Republic. Oeconomica, Prague, 2010.
12. 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), 2005, pp. 503–518.
13. 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), 2003, pp. 113–134.
14. Clark, P. Matwin, S.: Using Qualitative Models to Guide Inductive Learning. In: Proceedings of the 1993 International Conference on Machine Learning, 49-56.
15. Coulet A., Smaïl-Tabbone M., Benlian P., Napoli A., Devignes M.-D.: Ontology-guided Data Preparation for Discovering Genotype-Phenotype Relationships. In BMC Bioinformatics 9(Suppl 4): S3 (2008).
16. Domingues M. A., Rezende S. O.: Using Taxonomies to Facilitate the Analysis of the Association Rules. In: 2nd Int'l Workshop on Knowledge Discovery and Ontologies, KDO'05, at ECML/PKDD, Porto 2005.
17. van Dompseler, H. J. H., van Someren, M. W.: Using Models of Problem Solving Bias in Automated Knowledge Acquisition. In: ECAI94 - European Conference on Artificial Intelligence, Amsterdam 1994, 503507.
18. Engels, R., Lindner, G., Studer, R.: Providing User Support for Developing Knowledge Discovery Applications; A Midterm report. In: S. Wrobel (Ed.) *Themenheft der Künstliche Intelligenz*, (1) March, 1998.
19. Garshol L. M., Moore G.: Topic Maps XML Syntax. ISO/IEC JTC1/SC34, http://www.isotopicmaps.org/sam/sam-xtm/.
20. Garshol, L.M.: TMRAP - Topic Maps Remote Access Protocol. In: Maicher, L., Sigel, A., Garshol, L.M. (eds.) TMRA 2006. LNCS (LNAI), vol. 4438. Springer, Heidelberg (2007)
21. Garshol, L.M.: Towards a Methodology for Developing Topic Maps Ontologies. In: Maicher, L., Sigel, A., Garshol, L.M. (eds.) TMRA 2006. LNCS (LNAI), vol. 4438. Springer, Heidelberg (2007)
22. Hájek, P., Havránek, T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer–Verlag, 1978.
23. Kennard, J.: Mastering Joomla! 1.5: Extension Framework and Development. Packt 2007.
24. Kliegr, T., Oveăka M., Zemánek, J.: Topic Maps for Association Rule Mining. TMRA 2009. University of Leipzig 2009.
25. Kliegr M., Ralbovský M., Svátek, V, Šimůnek M., Jirkovský V., Nemrava J., Zemánek J.: Semantic Analytical Reports: A Framework for Post-Processing Data Mining Results. In: Foundations of Intelligent Systems (ISMIS'09). Springer Verlag, LNCS, 2009, 8898.

26. Kliegr M., Svátek, V, Šimůnek M., Stastný D., Hazucha A.: An XML Schema and a Topic Map Ontology for Formalization of Background Knowledge in Data Mining. In: IRMLeS-2010, 2nd ESWC Workshop on Inductive Reasoning and Machine Learning for the Semantic Web, Heraklion, Crete, Greece, May 2010. Online `http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-611/`
27. Kopanas I., Avouris N. M., Daskalaki S.: The Role of Domain Knowledge in a Large Scale Data Mining Project. In: Methods and Applications of Artificial Intelligence, LNCS Volume 2308/2002.
28. Kováč, M., Kuchař, T., Kuzmin A., Ralbovský, M.: Ferda, New Visual Environment for Data Mining. Znalosti 2006, Czech Rep., pp. 118–129 (in Czech).
29. Kuo Y.-T., Lonie A., Sonenberg L., Paizis K.: Domain ontology driven data mining: a medical case study. In: International Conference on Knowledge Discovery and Data Mining, Proceedings of the 2007 international workshop on Domain driven data mining.
30. Nazeri Z., Bloedorn E.: Exploiting Available Domain Knowledge to Improve Mining Aviation Safety and Network Security Data. In: Knowledge Discovery and Ontologies (KDO-2004), Workshop at ECML/PKDD 2004, Pisa.
31. Nunez, M.: The Use of Background Knowledge in Decision Tree Induction. Machine Learning, 6, 231250 (1991).
32. Olaru A., Marinica C., Guillet F.: Local mining of Association Rules with Rule Schemas. CIDM 2009: 118-124. http://www.claudiamarinica.com/pdf/CIDM2009.pdf
33. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. In: International Conf. Knowledge Capture, Victoria, Canada, 2001.
34. Podpečan V., Lavrač N., Kok J. N., de Bruin J. (eds.): Third Generation Data Mining: Towards service-oriented knowledge discovery (SoKD09), September 7, 2009, Slovenia.
35. Ralbovský 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, pp. 339–351
36. Rauch, J.: Logic of Association Rules. *Applied Intelligence*, 22, 2005, pp. 9–28.
37. Rauch J.: Considerations on Logical Calculi for Dealing with Knowledge in Data Mining. In: Advances in Data Management. Studies in Computational Intelligence, Volume 223/2009, Springer 2009.
38. Rauch J., Šimůnek M.: Dealing with Background Knowledge in the SEWEBAR Project. In: Knowledge Discovery Enhanced with Semantic and Social Information. Studies in Computational Intelligence, Volume 220/2009, Springer 2009.
39. Rauch J., Šimůnek M.: 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.
40. 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, pp. 143–149.
41. 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.
42. Srikant R.,Agrawal R.: Mining generalized association rules. VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, pp. 407–419.
43. Svátek, V.: Exploiting Value Hierarchies in Rule Learning. In: ECML'97, 9th European Conference on Machine Learning. Poster Papers. Prague 1997, 108–117.
44. Svátek V, Rauch J, Ralbovský.: Ontology-Enhanced Association Mining. In: Ackermann M. et al. Semantics, Web and Mining. Berlin : Springer, 2006, pp. 163–179.
45. Suyama, A., Yamaguchi, T.: Specifying and Learning Inductive Learning Systems using Ontologies. In: AAAI'98 Work. on the Methodology of Applying Mach. Learn., pp. 29–36.
46. Thomas J., Laublet, P., Ganascia, J. G.: A Machine Learning Tool Designed for a Model-Based Knowledge Acquisition Approach. In: EKAW-93, European Knowledge Acquisition Workshop, Lecture Notes in Artificial Intelligence No.723, N.Aussenac et al. (eds.), Springer-Verlag, 1993, 123138.
47. Thomas, H., Redmann T., Pressler, M., Markscheffel, B.:Towards a Graphical Notation for Topic Maps. TMRA 2008. University of Leipzig 2008.
48. Tomečková, M.: Minimal data model of the cardiological patien - the selection of data. *Cor et Vasa*, Vol. 44, No. 4 Suppl., 2004, pp. 123 (in Czech).
49. Tseng M.-C., Lin W.-Y., Jeng R.: Mining Association Rules with Ontological Information. In: Second International Conference on Innovative Comp., Inform. and Control. ICIC 2007.

50. Wagstaff K., Cardie C., Rogers S., Schroedl S.: Constrained K-means Clustering with Background Knowledge. In: ICML 2001.
51. Workshop Notes on Discovery Challenge (workshop at PKDD'99). Prague 8/1999.
52. Zeman, M., Ralbovský M., Svátek V., Rauch J.: Ontology-Driven Data Preparation for Association Mining. In: Znalosti 2009, Czech Republic, pp. 270–283.