Ontology-Driven Data Preparation for Association Mining

Martin Zeman[1], Martin Ralbovský[2], Vojtěch Svátek[2], and Jan Rauch[2]

[1] Department of Software Engineering, Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, 118 01 Praha, Czech Republic
martinzeman@email.cz
[2] Department of Information and Knowledge Engineering, University of Economics, Prague
W. Churchill Sq. 4, 130 67 Praha, Czech Republic
ralbovsm@vse.cz, svatek@vse.cz, rauch@vse.cz

Abstract. Ontologies can convey domain semantics to various phases of a KDD application through a mapping established between ontology entities and columns of the data matrix. The approach implemented in the Ferda tool focuses on providing support for the data preparation phase. Information about important data values and column groupings, once injected into a domain ontology, can be repeatedly used for creating meaninfgul categories for attributes and for defining mining tasks producing association hypotheses well-interpretable in the domain context. Tests on real data have been carried out in the domain of cardiology.

1 Introduction

Domain ontologies are potentially one of key vehicles for conveying the domain semantics to a KDD application. The traditional approach in KDD is to either treat the data relatively in isolation from the domain context, or, in the better case, to transfer the domain semantics into the individual phases of the mining process via human judgment, possibly documented using free text. This leads to effort replication, since, even in the same domain, the low-level data model to be used for analytical tasks has to be reinvented almost from scratch. Furthemore, results from disparate modelling sessions are hard to compare or integrate.

Ontologies can bring more formal rigour and better possibility of reuse to this process. It is often the case that domain experts and data mining professionals are disjoint groups of people. The latter may not have profound knowledge of the domain, but do nonetheless need some domain knowledge especially in the data preparation phase, when data are filtered, cleansed and organized. Such knowledge can in principle be either inherent part of domain ontologies from the beginning or can be injected to them by domain experts the first time the ontologies are considered for a data-intensive task.[3] I can then easily be picked from the ontologies automatically in order to support the data mining specialist. However, any attempt of ontology-enhanced KDD has to deal with structural heterogeneity issues. The structure of ontologies, even if the domain is

[3] Not necessarily tabular data mining but also e.g. information extraction from text, or ontology matching for the purpose of data integration.

the same, often strikingly differs from the structure of data tables, both formally and in their level of granularity. Therefore, the first and probably hardest step in applying ontologies in the KDD process consists in creating the *mapping* between the ontology and the source database. If the mapping is done properly, ontological knowledge can be exploited in the remaining phases of the KDD process.

In our earlier work [19] the possibilities for exploiting ontological knowledge were systematically traced over the different phases of the KDD process roughly corresponding to the CRISP-DM cycle: domain understanding, data understanding, data preparation, modeling, result interpretation and their dissemination over the semantic web. As the core approach to actual data mining, *association mining* was chosen; being a descriptive rather than predictive task, domain relevance and interpretability of results (which are to be submitted to a human expert rather than to an automated reasoning engine) are of particular importance here. More specifically, generalised association mining procedures based on the GUHA method [7] were used for simple experiments that served as proof of concept. The 'ontological engineering' part of these experiments was however mostly carried out manually.

Follow-up research, a significant part of which is presented in the current paper, extended this initial analysis both in terms of its underlying theoretical principles [13] and in terms of its support within a user-friendly KDD tool—*Ferda DataMiner* [21]. It so far focused on the *data preparation* step, which is generally considered as most time-consuming and fastening it could boost the whole KDD process.

The paper is structured as follows. Section 2 explains GUHA as underlying data mining method and outlines the roles of ontologies as prior knowledge in GUHA-based association mining in general. Section 3 familiarizes the reader with Ferda DataMiner as particular implementation of GUHA. Section 4 shows how ontologies are used to aid the data preparation process in Ferda. Section 5 illustrates the Ferda-based data preparation process with the step-by-step description of a concrete experiment on cardiological data and knowledge. Section 6 puts the current work into the context of related research. Finally, section 7 concludes the paper and shows directions for future work.

2 Ontologies as Prior Knowledge in GUHA

2.1 The GUHA Method

The GUHA method, developed in the mid-sixties, is one of the first methods of exploratory data analysis. It is a general framework for retrieving interesting knowledge from data. The method has solid theoretical foundations based on observational calculi and statistics [7]. For the purposes of this paper, let us only explain the basic principles of the method, as shown in Figure 1.

The GUHA method is realized by GUHA procedures such as the 4FT procedure (used in our work), located in the middle of the figure. Inputs of the procedure are the data and a simple definition of a possibly large set of relevant patterns. The procedure automatically generates all relevant patterns and verifies them against the source data. Patterns that are positively verified are output by the procedure.

Although GUHA is not in principle restricted to mining association rules, the most frequently used GUHA procedures mine for *generalized association rules* as defined

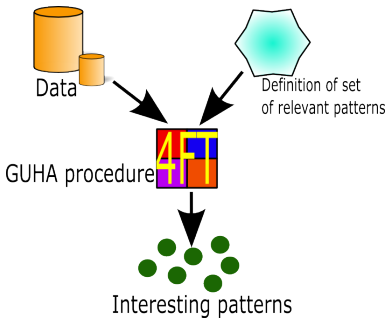

Fig. 1. The GUHA method

in [16]. The association rules are generalized in two ways: they allow a more complex structure in antecedent and consequent than just conjunction of items, and also allow to examine a wide variety of measures (relations) between the antecedent and consequent.

2.2 Using Domain Ontologies in Data Preparation for GUHA

Previous research [6, 13] proposed numerous potential ways to enhance the (GUHA-based) KDD process using ontologies. Based on these initial studies, we identified two ways of ontology usage that looked most promising. These are:

- Construction of adequate attribute categorization with the aid of ontologies
- Identification and exploitation of semantically related attributes

Construction of Attribute Categories. Data mining tools often deal with data for which some higher-level semantics could be assigned to individual values. For example, for blood pressure there are predefined values that divide the domain in a meaningful way: say, blood pressure above 140/90 mm Hg is considered as hypertension. Without proper categorization, data mining may give opaque or even misleading results.

Therefore, a way of storing the categorization information in an OWL ontology has been proposed, and tools for automatic creation of categorized attributes was implemented in the Ferda tool. Section 4 describes details.

Identification and Exploitation of Semantically Related Attributes. Examined data matrices often consist of a large number of columns representing information about real-world entities. These entities are often semantically close to each other, such relationships may however not always be transferred to the data mining phase. Ontologies have representational power to express various kinds of semantic closeness, foremost within the class taxonomy. Identification of mutually related entities can be exploited so as to meaningfully arrange the corresponding data attributes in the examined matrix (in the data preparation phase) or to construct meaningful data mining tasks (in the modelling phase).

We implemented a mechanism that identifies semantically related attributes in data and prepares them for further usage. Details are again in section 4.

3 Overview of Ferda DataMiner

Ferda (or Ferda DataMiner)[4] is a recent implementation of the GUHA method. The software evolved from a version of the older *LISp-Miner* system,[5] see [20]. Ferda started as a student project at Faculty of Mathematics and Physics, Charles University and is now under development at Department of Information and Knowledge Engineering, University of Economics (both in Prague). A complete overview of the Ferda system is presented in [9]. Since 2006, publication date of [9], the system has undergone improvements in visual appearance, performance and stability. System includes new implementations of 6 propositional GUHA procedures[15], two relational GUHA procedures [10] and an algorithm for construction of GUHA decision trees [14]. Programming abilities of the system were greatly improved, it now features a fully recursive programming language based on the lambda calculus [8].

Here, we emphasis only the features of the system important for this work. Figure 2 shows the Ferda environment. There are four boxes (modules of Ferda) displayed on the *desktop*. Boxes have properties that can be set either by the *property grid* component located in the upper right corner, or by a *setting module*, which is module designed to set complex properties (see 4.2). Important feature of the system is the contextual *box recommendation* mechanism. It advices the user on which box should be used in the next step, more precisely, it shows the user via a contextual menu all types of boxes that can be connected to the selected box. An example of usage of this functionality is in section 4.5.

4 Introduction of Ontologies into Ferda

In this section we present the different functionalities and aspects of ontology management and exploitation as they have been implemented in Ferda.

4.1 Ontology Representation Language

From the various possibilities we chose *OWL* (Web Ontology Language, version 1.1) to represent ontologies, especially because it is a standard of W3C and it is widely supported by developers. As interface for manipulating ontologies we used the Java OWL API parser.[6] The ontology was made accessible by a Ice middleware module in Ferda, which loads a local or remote OWL ontology and makes the content accessible for other modules of the Ferda system.

4.2 Mapping Ontologies to Database

When binding a particular data source to a particular domain ontology, a mapping is to be established that connects individual data columns to entities from the ontology.

[4] `http://ferda.sourceforge.net`
[5] `http://lisminer.vse.cz`
[6] Downloadable from `http://owlapi.sourceforge.net/`.

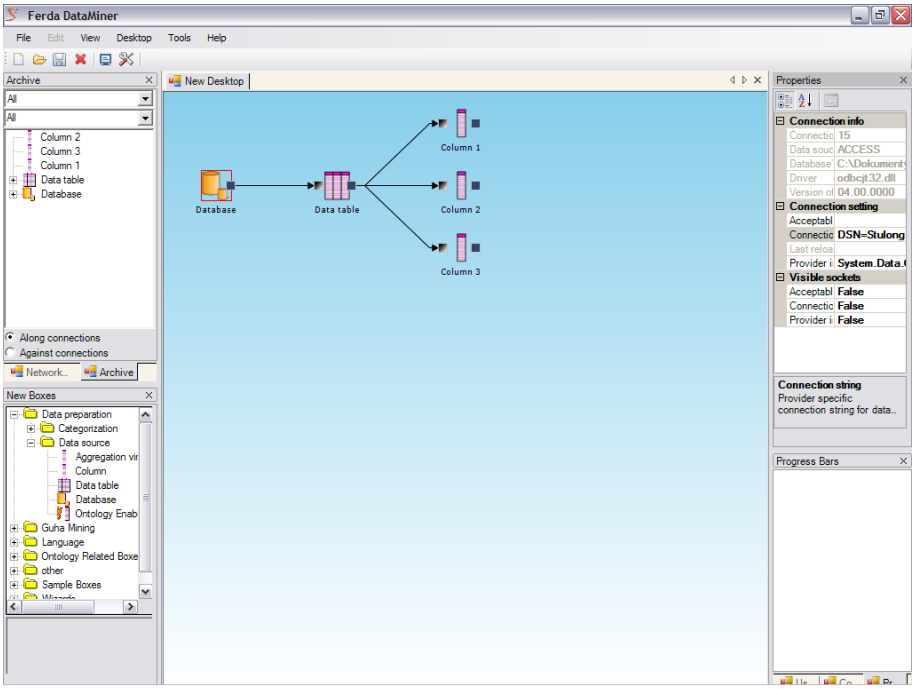

Fig. 2. The Ferda environment

In Ferda we implemented a *setting module* for manual mapping from data columns to ontology concepts. This mapping can be loaded and stored in an XML format, and thus can be created once and then repeatedly reused. Figure 3 shows the setting module that handles the mapping.

Depending on the structure of the (OWL) ontology, the most adequate entity on which a data column should be mapped could be a concept, an instance, an object property, or, possibly most adequately from the formal point of view, a datatype property. In the current version of Ferda, the user can map a data column to an ontology class or instance only: this is motivated by the experience that a *class-centric* view is much more convenient to the user (when specifying the mappings) than a *property-centric* view. We also do not allow mapping of one column to multiple classes or instances. It is however possible to map multiple columns to one class or instance of the ontology. This situation occurs when the granularity of the data source is higher than the granularity of the ontology.

Each user can create his/her own mapping, but it is desirable to share a common mapping created by the domain expert. When the database is connected to the ontology via a mapping, Ferda automatically recognizes names of concepts from the ontology and uses them besides/instead of names from the database.

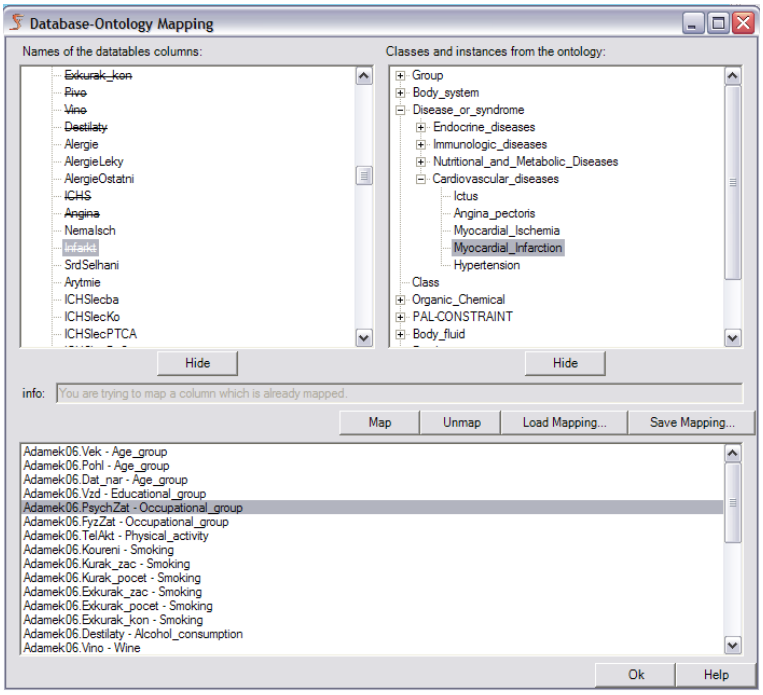

Fig. 3. Mapping database columns to ontology entities

4.3 Storing Additional Information to Ontology

As we mentioned in section 2, we need to store additional information in ontologies to enhance their usability for KDD. In order to stay within the formalism of OWL, we use a *meta-modelling* approach: for each type of additional information to be stored (such as attribute cardinality or important values dividing the domain) we created a special datatype property and set its domain to an OWL metaclass; currently we only use *owl:Class* for simplicity. Thereafter, all classes can be equipped with this additional information type.

Information useful for data mining could alternatively be assigned to (especially, datatype) properties rather than to classes. For this, it would suffice to connect the respective property to *owl:DatatypeProperty* rather than to *owl:Class*. We are observing the development of the new version of the OWL standard, *OWL 2* [1], where some meta-modelling issues are handled in a novel way.

It is important to note that including such additional information into the ontology is not merely a matter of shifting the categorisation task from the data preparation phase to some previous phase. Namely, the additional information items mentioned

- are an inherent part of the *domain*, and thus also deserve to be part of a domain ontology
- thanks to the persistence and potential web-based accessibility of the ontology, they can be repeatedly *reused* by different people at different places

– their use is *not restricted to data mining*, but they can also be exploited in other data-intensive ontology-driven tasks, such as in information extraction from text[7] or in data integration via ontology matching.

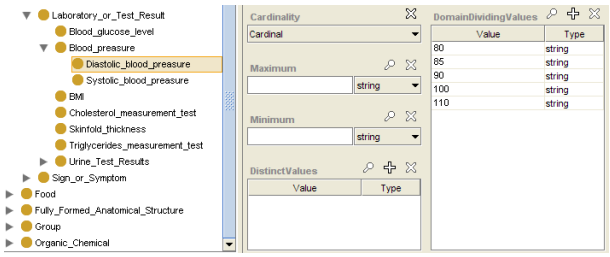

Fig. 4. Including additional information in the ontology

4.4 Categorization of Attribute

One of the main contributions of the ontology-based approach to the data preparation step is the ability to construct the attribute categorization automatically. In Ferda, *attribute* boxes provide attribute categorization. Before the ontology support was implemented, the user could only use categorization algorithms to create equidistant or equifrequency intervals, or to create the categorization manually. This task was time consuming, had to be done over and over again, and very often the resulting categorization did not reflect the semantics of data and resulted in misleading KDD results. With ontological support, the user can create an *ontology-derived attribute* box. This box loads information from the ontology and performs categorization based on additional information stored in the ontology.

4.5 Identification and Exploitation of Semantically Related Attributes

As was mentioned in section 2.2, identification and usage of mutually related attributes helps the user in both the data preparation and task design stages of the data mining cycle. We used the *box recommendation* mechanism, and offered the user semantic support (mainly at the level of naming and taxonomy) based on the ontology, as can be seen in Figure 5. When the user clicks on one item in the context menu, all data attributes corresponding to the entity from the ontology are created and added to the matrix for data mining. For example, for the concept of *Cardiovascular diseases*, attributes for individual findings such as *Angina pectoris* or *Myocardial ischemia* are added.

[7] The so-called extraction ontologies [11] share a lot with our approach mentioned.

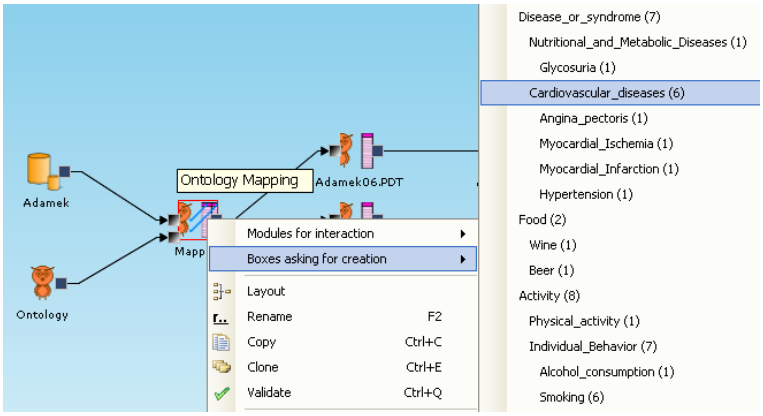

Fig. 5. Semantics of attributes from the database

5 Demonstrative Experiment

In order to demonstrate the usefulness of the presented approach, we show a step-by-step experiment that simulates the solving of a typified data mining task. We cannot compare our method to any other method, because to our best knowledge such a method does not exist. The task concerns finding strong associations between two groups of attributes in a medical database. Section 5.1 describes the medical data, section 5.2 describes the mining method and analytical question solved, section 5.3 describes the task setup, and finally, section 5.4 concludes the experiment.

5.1 ADAMEK data

The *ADAMEK* medical data set [18] was used for our experiment. The data set contains data taken from the ambulance of preventive cardiology and contains 180 attributes for each of 1122 examined patients. Related attributes are clustered to 25 thematic groups.

5.2 Mining Method and Analytical Questions

We chose GUHA-based association mining and in particular its 4FT procedure as our mining method. Although the method provides great possibilities in constructing association rules, we simplified the task so that both the antecedent and consequent only consist of a conjunction of attribute categories and the relation between them is expressed by *support* and *confidence*. In this way, mining with 4FT corresponds to classical *a-priori* mining [2], and there is no need to introduce the complexity of GUHA theory to the reader.

One can construct *analytical questions* that correspond to finding associations between two chosen groups of related attributes. We chose a medium-sized experiment that consisted in solving the following analytical questions:

What are the relations between blood pressure levels and the following diseases:

1. *Myocardial ischemia*
2. *Cardiomyopathy*
3. *Diabetes*
4. *Hypertension*
5. *Allergy*
6. *Angina Pectoris*
7. *Myocardial infarction*

5.3 Task Setup

The experiment aims to test the benefit of association mining with aid of ontologies. Therefore we presume that the data miner uses an available ontology enhanced with additional information and mapping. We used a relevant part of the *UMLS*[8] metathesaurus and semantic network as the ontology. The mapping was created manually; Table 1 shows the details.

Attribute group	**Ontology entity**	**Columns in group**	**Mapped columns**
Blood pressure	Blood pressure	4	4
ICHS	Myocardial ischemia	16	7
Diabetes	Diabetes	6	5
Hypertension	Hypertension	5	4
Allergy	Alergic anamnesis	3	2
-	Angina pectoris	1	1
-	Myocardial infarction	1	1

Table 1. Mapping details

We can see from the table (last column) that multiple database attributes were mapped on one ontology entity. This is because we used a general medical ontology but rather specific cardiologic data. The two last entities from the ontology did not have a corresponding attribute group in the data source. The number of mapped attributes is usually lower than the total number of attributes in one attribute group. This is caused by the inconsistency of the database: some attributes are erroneous and cannot be used for mining at all.

The mapping itself can be a valuable source of information. The mining tool, through it, receives information about the grouping of attributes and also the identification of unfit attributes. With the mapping available, data preparation is simplified. We take advantage of the identification of related attributes and of the automatic attribute creation. The user selects the mapping box and clicks into the *Boxes asking for creation* submenu on the chosen entity. Then s/he selects all created columns and chooses the *Ontology derived attribute* option from the *Boxes asking for creation* submenu to complete attribute creation through box recommendation. Furthermore, the information in ontology guarantees that the categorization is in a sense correct. Without ontology support the user

[8] Unified Medical Language System `http://www.nlm.nih.gov/research/UMLS/`

would have to find each column in the overall list of columns, then to choose a proper attribute box (representing categorization), and then finally adjust the box.

5.4 Results and Evaluation

In Table 2, for completeness, we list the numbers of hypotheses resulting from the above-described experiment for the seven analytical questions mentioned, with basic parameter setting *Support = 0.1*, *Confidence = 0.8*.[9]

Analytical question	1	2	3	4	5	6	7
No. of hypotheses found	191	32	33	13	0	0	0

Table 2. Numbers of hypotheses found in the illustrative experiment

The experiment demonstrates the usability of ontology support in GUHA-based association mining in two ways. Proper mapping and terms from the ontology help the user better understand the examined *data* and to properly create *attributes*. The implementation in Ferda also allows to speed up the process of constructing the data mining *task*. Without ontology usage, the user needs to perform several mouse clicks and item selections for each attribute from an attribute group. With ontology usage, all attributes from one group are created just by 3 clicks and two item selections no matter the size of the group.

6 Related Research

Although domain ontologies are nowadays a popular instrument in many diverse applications incl. e.g. text mining, they only scarcely appeared in 'tabular' KDD until very recently. One notable exception was the work by Philips & Buchanan [12], where 'common-sense' ontologies of time and processes were exploited to derive constraints on attributes, which were in turn used to construct new attributes. Our approach implemented in the visual environment of Ferda is more suitable for addressing specific domains, as it allows the user to conveniently specify domain knowledge. Our ongoing work is also aimed at detection of missing attributes, which is somewhat analogous (though not identical) to the approach of [12]. Another relevant project is that by Bogorny et al. [3], which aims to prune trivial frequent association patterns in the geospacial domain. The main purpose is however there to cope with computational complexity for the machine, and the issue of frequent trivial patterns is to some degree specific to this domain; our approach, in contrast, is well portable to different domains, and aims to primarily support the human user.

A specific stream of ontology-aware knowledge discovery is represented by bioinformatics applications that exploit (usually, shallow) ontologies in mining gene data,

[9] We consider these values as default for first iteration of association rules mining

see e.g. [5, 4]; their portability to different domains is not obvious. Another promising direction, though inherently different from our 'tabular mining' approach, is that attempting to reconciliate the notion of background knowledge in Inductive Logic Programming with that of ontology [22].

We should also mention the research done in parallel in our own group within the *SEWEBAR* project [17]. There the data mining methods used are also based on GUHA but rely on a different implemented platform called *LISp-Miner*.[10] The main focus of SEWEBAR is on the mining result exploitation phase. Prior knowledge is also used for guiding the data preparation and task definition phases to some degree, it is however currently expressed using a proprietary format rather than using a widely-used semantic web language. We are working towards harmonising both research threads.

7 Conclusions and Future Work

The approach to ontology support to association mining implemented in the Ferda tool focuses on providing support for the data preparation phase. Information about important data values and data column groupings, once injected into a domain ontology, can be repeatedly used for creating meaninfgul categories for attributes and for defining mining tasks producing association hypotheses well-interpretable in the domain context. Tests on real data have been carried out in the domain of cardiology.

In the future, we plan to extend the support to further *phases* of the association mining process, as already envisaged in [13] and [19]. We will also enhance the method of *introducing* data-related knowledge to ontologies, among other reflecting the evolution of the OWL language. It will be necessary to balance the accuracy of data-to-ontology mapping (where mapping to datatype properties seems to be most relevant) with the ergonomy for the end user (who might prefer a concept-centric view of ontology). Finally, it is likely that not only the most classical approach to GUHA-based mining (relying on four-fold table quantifiers) but also other *mining methods* implemented in Ferda could benefit from exploiting ontological knowledge.

Acknowledgment

This work was supported by the CSF project no.201/08/0802, "Application of Knowledge Engineering Methods in Knowledge Discovery from Databases".

References

1. *OWL 2 Web Ontology Language: Profiles*. W3C Working Draft 11 April 2008, online `http://www.w3.org/TR/2008/WD-owl2-profiles-20080411/`.
2. Agrawal R., Imielinski T., Swami A.: Mining association rules between sets of items in large databases. In: Proc. of the ACM SIGMOD Conference on Management of Data, p.207–216.

[10] A major difference between Ferda and LISp-Miner is that the latter has a dialogue-oriented rather than visual programming interface.

3. Bogorny, V., Engel, P., Alvares, L.O.: Enhancing the Process of Knowledge Discovery in Geographic Databases using Geo-Ontologies. In: Nigro, H. O., Cisaro, S.G., Xodo, D. (Ed.). Data Mining with Ontologies: Implementations, Findings, and Frameworks. Idea Group Inc. (2007). pp.160-181.
4. Brisson, L., Collard, M., Le Brigant, K., Barbry, P.: KTA: A Framework for Integrating Expert Knowledge and Experiment Memory in Transcriptome Analysis. In: International Workshop on Knowledge Discovery and Ontologies, held with ECML/PKDD 2004, Pisa, p.85-90.
5. Cannataro, M., Guzzi, P. H., Mazza, T., Tradigo, G., Veltri, P.: Using Ontologies in PROTEUS for Modeling Proteomics Data Mining Applications. In: From Grid to Healthgrid: Proceedings of Healthgrid 2005, IOS Press, 17-26.
6. Češpivová H. Rauch J., Svátek V., Kejkula M., Tomečková M.: Roles of Medical Ontologies in Association Mining CRISP-DM Cycle, ECML/PKDD Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa 2004.
7. Hájek P., Havránek, T.: *Mechanising Hypothesis Formation – Mathematical Foundations for a General Theory.* Springer-Verlag, 1978.
8. Kováč M.:User oriented language for solving KDD tasks. Master Thesis, Faculty of Mathematics and Physics, Charles University, Prague (to appear).
9. 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).
10. Kuzmin A.: Relational GUHA procedures. Master Thesis, Faculty of Mathematics and Physics, Charles University, Prague 2007 (in Czech)
11. Labský, M., Nekvasil, M., Svátek, V., Rak, D.: The Ex Project: Web Information Extraction using Extraction Ontologies. In: Proc. PriCKL'07, ECML/PKDD Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery. Warsaw 2007.
12. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. In: International Conf. Knowledge Capture (K-CAP, 2001), Victoria, Canada, 2001.
13. Ralbovský M.: Usage of Domain Knowledge for Applications of GUHA Procedures. Master thesis, Faculty of Mathematics and Physics, Charles University, Prague, 2006.
14. Ralbovský M., Berka P.: Implementation of GUHA Decision Trees. In: MIS 2008, Computer science conference, Josefův Důl (to appear).
15. Ralbovský M., Kuchař T.: Using Disjunctions in Association Mining. In: P. Perner (Ed.), Advances in Data Mining - Theoretical Aspects and Applications, LNAI 4597, Springer Verlag, Heidelberg 2007.
16. Rauch J.: Logic of Association Rules. *Applied Intelligence*, Vol. 22, Issue 1, p. 9 – 28.
17. Rauch, J., Šimůnek, M.: Semantic Web Presentation of Analytical Reports from Data Mining—Preliminary Considerations. In: Web Intelligence 2007, Los Alamitos: IEEE Computer Society, 2007, pp. 3 7.
18. Rauch J., Tomečková M.: System Of Analytical Questions And Reports on On Mining In Health Data A Case Study. MCCSIS 2007 [CD-ROM]. Lisabon : IADIS, 2007. pp. 176–181.
19. Svátek V., Rauch J., Ralbovský M.: Ontology-Enhanced Association Mining. In: Ackermann et al. (eds.). Semantics, Web and Mining, Springer-Verlag, 2006.
20. Šimůnek M.: Academic KDD Project LISp-Miner. In: Advances in Soft Computing—Intelligent Systems Design and Applications, Springer Verlag 2003.
21. Zeman M.: Usage of Ontologies for GUHA Procedures. Master thesis, Faculty of Mathematics and Physics, Charles University, Prague 2008.
22. Žáková, M., Železný, F.: Exploiting Term, Predicate, and Feature Taxonomies in Propositionalization and Propositional Rule Learning. ECML 2007: the 18th European Conference on Machine Learning, Springer 2007.