Ontology-Enhanced Association Mining

Vojtěch Svátek, Jan Rauch

Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
e-mail: svatek@vse.cz, rauch@vse.cz

Abstract. The roles of ontologies in KDD are potentially manifold. We track them through different phases of the KDD process, from data understanding through task setting to mining result interpretation and sharing over the semantic web. The underlying KDD paradigm is association mining tailored to our *4ft-Miner* tool. Experience from two different application domains—medicine and sociology—is presented throughout the paper.

1 Introduction

Domain ontologies, being hot topic in today's knowledge engineering research, are promising candidates for background knowledge to be used in the KDD process. They express the main concepts and relationships in a domain in a way that is consensual and comprehensible to the given professional community. The research in applied ontology and in KDD are, to some extent, two sides of the same coin. Ontologies describe the 'state-of-affairs' in a certain domain at an abstract level, and thus enable to verify the correctness of existing (concrete) facts as well as to infer new facts. On the other hand, KDD typically proceeds in the opposite direction: from concrete, instance-level patterns to more abstract ones. Semantic web mining [5] represents the junction of ontology and KDD research in their 'concrete' (instance-centric) corners. On the other hand, in this paper, we rather focus on the junction of 'abstract' corners, namely, of abstract ontologies themselves and general hypotheses produced by KDD.

The role to be played by ontologies in KDD (and even their mere usability) depends on the given mining *task* and *method*, on the *stage of the KDD process*, and also on some characteristics of the *domain* and *dataset*. The experiment described in this paper is connected with *task* of *association mining*, namely, with the *4ft-Miner* tool [17] (component of *LISp-Miner*, see `http://lispminer.vse.cz`), which is inspired by the GUHA *method* [10]. We identified four *stages* of (*4ft-Miner*-based) KDD that are likely to benefit from ontology application: data understanding, task design, result interpretation and result dissemination over the semantic web[1]. Finally, we conducted our research in two different *domains* with specific *datasets* (and available ontological resources): the domain of cardiovascular risk and that of social climate.

[1] In a pre-cursor paper [7], we explicitly used *CRISP-DM* (`http://www.crisp-dm.org`) for process decomposition; however, the phases are rather generic.

The paper is structured as follows. Section 2 describes both domain-specific applications. Section 3 recalls the basic principles of *4ft-Miner*. Sections 4, 5, 6 and 7 are devoted each to one phase of the KDD process as outlined above. Finally, section 9 reviews some related work, and section 10 shows directions for future research.

2 Overview of Applications

2.1 Cardiovascular Risk: Data and Ontologies

The *STULONG dataset* concerns a twenty-years-lasting longitudinal study of risk factors for atherosclerosis in the population of middle-aged men (see `http://euromise.vse.cz/stulong-en/`). It consists of four data matrices:

Entrance. Each of 1 417 men has been subject to entrance examination. Values of 244 attributes have been surveyed with each patient. These attributes are divided into 11 groups e. g. *social characteristics*, *physical activity* etc.

Control. Risk factors and clinical demonstration of atherosclerosis have been followed during the control examination for the duration of 20 years. Values of 66 attributes have been recorded for each one. There are 6 groups of attributes, e.g. *physical examination*, *biochemical examination* etc.

Letter. Additional information about health status of 403 men was collected by postal questionnaire. There are 62 attributes divided into 8 groups such as *diet* or *smoking*.

Death. There are 5 attributes concerning the death of 389 patients.

As ontology we used *UMLS* (Unified Medical Language System) [2], namely its high-level *semantic network* and the *meta-thesaurus* mapping the concepts picked from third-party resources onto each other. Although the central construct of UMLS is the concept-subconcept relation, the semantic network also features lots of other binary relations such as 'location of' or 'produces'. However, since the network only covers 134 high-level 'semantic types' (such as 'Body Part' or 'Disease'), the relations are only 'potentially holding' (it is by far not true that every Body Part can be 'location of' every Disease...). The meta-thesaurus, in turn, covers (a large number of) more specific concepts but relations are only scarcely instantiated, and nearly all relation instances belong to the 'location of' relation.

As additional resource, we used the knowledge accumulated in the Czech medical community with respect to risk factors of cardio-vascular diseases, in connection with the STULONG project itself. The knowledge base consists of 36 *qualitative rules*, most of which can be characterised as medical background knowledge or common-sense knowledge, e.g. "increase of cholesterol level leads to increase of triglycerides level", "increase of age leads to increase of coffee consumption", "increase of education leads to increase of responsibility in the job" or the like. Given the mentioned lack of concrete inter-concept relationships in UMLS, we adopted them, for experimental purposes, as if they were part of this ontology.

2.2 Social Climate: Data and Ontologies

In the second application, both the *ontology* and the *dataset* used for association discovery had the same seed material: the *questionnaire* posed to respondents during the *opinion poll* mapping the 'social climate' of the city of Prague in Spring 2004. The questionnaire contained 51 questions related to e.g. economic situation of families, dwelling, or attitude towards important local events, political parties or media. Some questions consisted of aggregated sub-questions each corresponding to a different 'sign', e.g. "How important is X for you?", where X stands for family, politics, religion etc.; other questions corresponded each to a single 'sign'. The questions were divided into 11 thematic groups.

While the *dataset* was straightforwardly derived from the individual 'signs', each becoming a database column[2], the *ontology* first had the form of *glossary* of candidate terms (manually) picked from the text of the questions; duplicities were removed. In conformance with most ontology engineering methodologies [9], the terms were then divided into candidates for *classes*, *relations* and *instances*, respectively. Then a *taxonomy* and a structure of *non-taxonomic relations* was (again, manually) built, while filling additional entities when needed for better connectivity of the model or just declared as important by domain expert. The instances either correspond to enumerated values of properties, e.g. `GOOD_JOB_AVAILABILITY`, or to outstanding individuals such as `PRAGUE` or `CHRISTIAN_DEMOCRATIC_PARTY`.

The current version of the ontology, eventually formalised in OWL[3], consists of approx. 100 classes, 40 relations and 50 individuals[4]. A Protégé[5] window showing parts of the class hierarchy plus the properties of class `Person` is at Fig. 1. Note that the ambition of our ontology is not to become a widely-usable formal model of social reality; it rather serves for 'simulation' of the possible role of such ontology in the context of KDD.

3 Association Mining with *4ft-Miner*

4ft-Miner mines for association rules of the form $\varphi \approx \psi$, where φ and ψ are called *antecedent* and *succedent*, respectively. Antecedent and succedent are conjunctions of *literals*. Literal is a Boolean variable $A(\alpha)$ or its negation $\neg A(\alpha)$, where A is an *attribute* (corresponding to a column in the data table) and α (a set of values called *categories*) is *coefficient* of the literal $A(\alpha)$. The literal $A(\alpha)$ is true for a particular object o in data if the value of A for o is some v such that $v \in \alpha$.

The association rule $\varphi \approx \psi$ means that φ and ψ are associated in the way defined by the symbol $\approx$. The symbol $\approx$, called *4ft-quantifier*, corresponds to a

[2] And, subsequently, an attribute for the *4ft-Miner* tool, see the next section.

[3] `http://www.w3.org/2004/OWL`

[4] By naming convention we adopted, individuals are in capitals, classes start with capital letter (underscore replaces inter-word space for both individuals and classes), and properties start with small letter and the beginning of other than first word is indicated by a capital letter.

[5] `http://protege.stanford.edu`

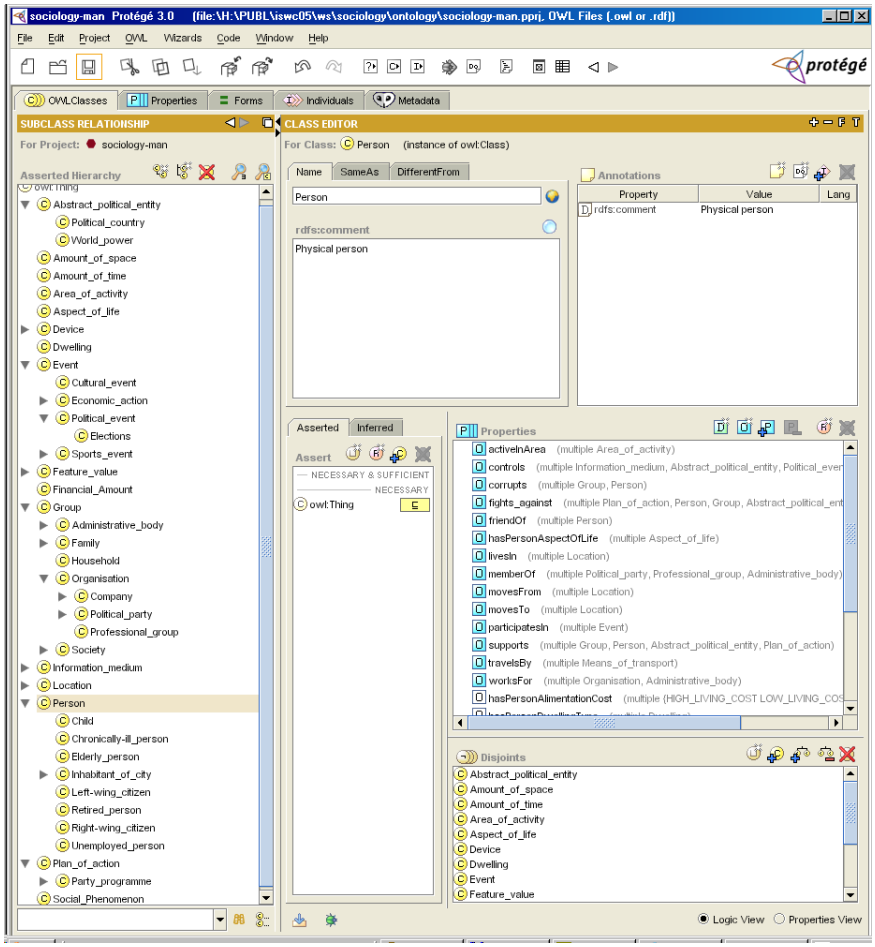

Fig. 1. Incomplete view of the ontology in Protégé

condition over the four-fold contingency table of φ and ψ. The four-fold contingency table of φ and ψ in data matrix $\mathcal{M}$ is a quadruple $\langle a, b, c, d\rangle$ of natural numbers such that a is the number of data objects from $\mathcal{M}$ satisfying both φ and ψ, b is the number of data objects from $\mathcal{M}$ satisfying φ and not satisfying ψ, c is the number of data objects from $\mathcal{M}$ not satisfying φ and satisfying ψ, and d is the number of from $\mathcal{M}$ from $\mathcal{M}$ satisfying neither φ nor ψ.

There are 16 4ft-quantifiers in *4ft-Miner*. An example of 4ft-quantifier is *above-average dependence*,
$\sim^{+}_{p,Base}$, which is defined for $0 < p$ and $Base > 0$ by the condition

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

The association rule $\varphi \sim^{+}_{p,Base} \psi$ means that among the objects satisfying φ is at least $100p$ per cent more objects satisfying ψ than among all observed objects and that there are at least $Base$ observed objects satisfying both φ and ψ.

As an example of association rule, let us present the expression

$$A(a_1, a_7) \wedge B(b_2, b_5, b_9) \sim^{+}_{p,Base} C(c_4) \wedge \neg D(d_3)$$

Here, $A(a_1, a_7)$, $B(b_2, b_5, b_9)$, $C(c_4)$ and $\neg D(d_3)$ are literals, a_1 and a_7 are categories of A, and $\{a_1, a_7\}$ is the coefficient of $A(a_1, a_7)$[6], and analogously for the remaining literals.

In order to determine the set of relevant questions more easily, we can define *cedents* (i.e. antecedent and/or succedent) φ as a conjunction

$$\varphi = \varphi_1 \wedge \varphi_2 \wedge \ldots \wedge \varphi_k$$

where $\varphi_1, \varphi_2, \ldots, \varphi_k$ are *partial cedents*. Each φ_i (itself a conjunction of literals) is chosen from the *set of relevant partial cedents*. The set of partial cedents is given in the following manner:

- the minimum and maximum *length* (i.e. the number of literals in conjunction) of the partial cedent is defined
- a set of *attributes* from which literals will be generated is given
- some attributes can be marked as *basic*, each partial cedent then must contain at least one basic attribute
- a simple definition of the set of all *literals* to be generated is given for each attribute
- *classes of equivalence* can be defined, such that each attribute belongs to at most one class of equivalence; no partial cedent then can contain two or more attributes from the same class of equivalence.

The set of all literals to be generated for a particular attribute is given by:

- the type of coefficient; there are six types of coefficients: subsets, intervals, left cuts, right cuts, cuts, one particular category
- the minimum and the maximum length of the literal
- positive/negative literal option: only positive, only negative, both.

4 Data Understanding

Within the phase of data understanding, the activity relevant for ontology exploitation is that of *data-to-ontology mapping*, the outcomes of which will be used in later phases.

In the *cardiovascular risk* application we succeeded in mapping 53 of STULONG attributes (from the Entrance dataset) on 19 UMLS semantic types and 25 metathesaurus concepts. Six attributes for which a concept could not be

[6] For convenience, we can write $A(a_1, a_7)$ instead of $A(\{a_1, a_7\})$.

found were only assigned semantic type, for example, 'responsibility in job' was assigned to semantic type Occupational Group. For subsequent processing, we only kept a light-weight fragment of UMLS containing, for each data attribute, the most adequate metathesaurus concept and the least-general semantic type subsuming this concept. We obtained a structure with five taxonomy roots: *Finding*, *Activity*, *Group*, *Food*, and *Disease or Syndrome*.

The side effect of mapping to ontology, as peculiar form of 'data understanding', was occasional identification of *redundant attributes*, which (though necessary for data management purposes) were not useful as input to data mining. For example, since the dataset contained the attribute 'age on entrance to STULONG study', the attributes 'birth year' and 'year of entrance to STULONG study' (all mapped to the Age Group semantic type) were of little use.

The mapping between STULONG data and the *qualitative rules* was straightforward, since the data were collected (more-or-less) by the same community of physicians who also formulated the knowledge base, within the same project.

For the same reason, the mapping task was relatively easy in the *social climate* application. Since the core of the ontology had been manually designed based on the text of the questions, it sufficed to track down the links created while building the ontology and maintained during the concept-merging phase. An example of mapping between a question and (fragments of) the ontology is in Fig. 2. Emphasised fragments of the text map to the concepts `Job_availability`, `Metropoly` and `Family` and to the individuals `GOOD_JOB_AVAILABILITY`, `PRAGUE`, `CENTRAL_EUROPE` and `EU`, plus several properties not shown in the diagram. Note that question no. 3 is a 'single-sign' question, i.e. it is directly transformed to one data attribute used for mining. In addition to questions, ontology mapping was also determined for *values* allowed as answers, especially for questions requiring to select concrete objects (city districts, political parties etc.).

5 Task Design

The mining process in narrow sense—*running* an individual mining session—is probably not amenable to ontologies in the case of *4ft-Miner*. The analysis of large data tables relies on optimised database-oriented algorithms, which could hardly accommodate the heterogeneity of ontological information. There is however room for ontologies in the process of *designing* the sessions, due to the sophisticated language for *4ft-Miner* task design (cf. section 3).

In the *cardiovascular risk* application, we used the mapping on ontology concepts from the previous phase so as to identify attributes that should be semantically grouped into partial cedents. We created *partial cedents* covering the attributes mapped on the five upmost classes.Although we carried out this part of the task manually, it could easily be automated.

At a higher level of abstraction, we can also operate on different *task settings*. A very general mining task setting can be decomposed into more specific tasks, which can be run faster, their results will be conceptually more homogeneous,

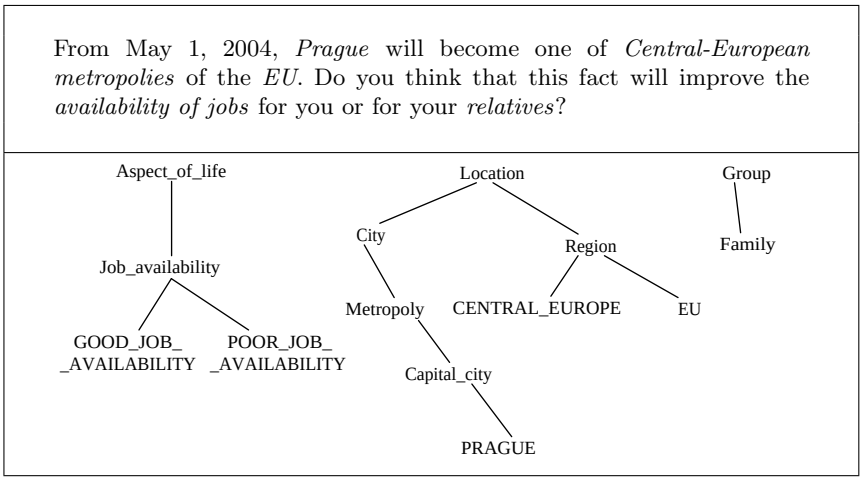

Fig. 2. Question no.3 and fragments of ontology used for its mapping

and thus can be interpreted more easily (see below). An example of task decomposition (for associations between patient activities and diseases/syndromes) is at Fig. 3. The base task (left branch) might lead to a high number of hypotheses that would be hard to interpret. We can thus e.g. separately refine the antecedent (middle branch) or succedent (right branch) of the base task to obtain more concise and homogeneous results per session.

In the *social climate* application, the ontology was not used in the task design phase. The reason was that the experiments were not guided by the interest of domain experts as in the cardiovascular risk application. So as to allow for the widest possible scope of candidate hypothesis, we thus kept the task definition maximally general: any of 96 attributes (corresponding to 'signs' from the questionnaire) was allowed in antecedent as well as in succedent[7]. As we wanted to start with (structurally) simplest hypotheses, we set the length of antecedent as well as of succedent to 1, and the cardinality of coefficient also to 1 (i.e., choice of single category). As quantifier we used the *above-average dependence* explained in section 3. The run-times were typically lower than a second.

6 Result Interpretation

Given the data-to-ontology mapping, concrete associations discovered with the help of *4ft-Miner* can be matched to corresponding semantic relations or their

[7] Actually, we would have benefited from the possibility of restricting the attributes in antecedent and succedent to be based on questions from *different* groups; such functionality however goes beyond the current task design language

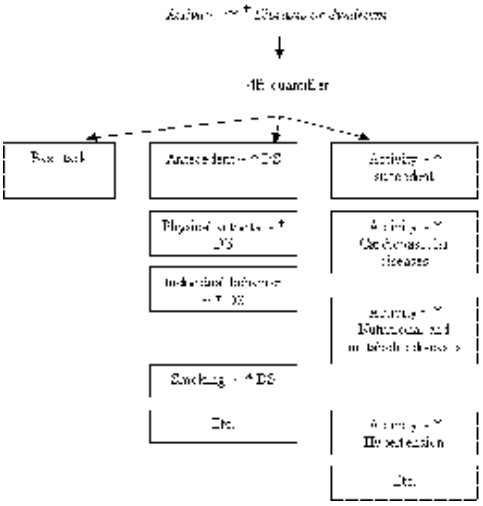

Fig. 3. Decomposition of 4ft tasks with respect to ontology

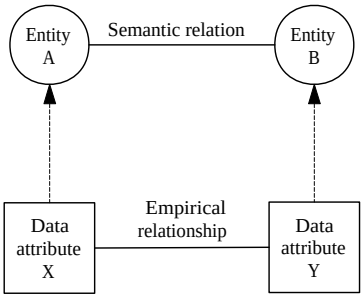

Fig. 4. Semantic relation as context to empirical relationship between attributes

more complex chains from the ontology, see Fig. 4. The semantic relation represents a potential context (e.g. explanation) for the discovered association.

In the *cardiovascular risk* application, each mining task already corresponded to a meaningful 'framing' question, such as "Searching for relationships between Activity of the patient and Diseases/Syndromes". Concrete associations discovered with the help of *4ft-Miner* could then be compared with the instance layer of the ontology; in our case, with the qualitative rules. The relationship of an association with prior knowledge is typically one of the following:

- Confirmation of prior knowledge, without additional information
- New knowledge compatible with prior knowledge
- Exception to or conflict with prior knowledge.

Let us show two examples, with their four-fold tables:

Table 1. Four-fold table for a 'confirmation' association

	Succedent	NOT Succedent	
Antecedent	10	22	32
NOT Antecedent	66	291	357
	76	313	389

Table 2. Four-fold table for a 'conflicting' association

	Succedent	NOT Succedent	
Antecedent	216	14	230
NOT Antecedent	145	14	159
	361	28	389

- The discovered association "Patients who are not physically active within the job nor after the job (Antecedent) will more often have higher blood pressure (Succedent)" was derivable from the background knowledge rule "Patients who are physically active after the job will more often have lower blood pressure" (Table 1).
- The discovered association "94% of patients smoking 5 or more cigarettes a day for more than 21 years (Antecedent) have neither myocardial infarction nor ictus nor diabetes (Succedent)" was in conflict with prior knowledge "Increase of smoking leads to increase of cardio-vascular diseases" (Table 2).

The examples are merely illustrative. In order to draw medically valid conclusions from them, we would at least need to examine the statistical validity of the hypotheses in question. As the STULONG dataset is relatively small, few such hypotheses actually pass conventional statistical tests.

In the *social climate* application, in contrast, we did not have a knowledge base with concrete rules at our disposal. We thus used the ontology itself—not to directly compare it with the hypotheses but to retrieve entity (concept-relation) chains that could serve as templates for candidate *explanations* of the hypotheses. Again, we did not have an appropriate software support for extracting entity chains (i.e. explanation templates) from the ontology, and only examined it via manual browsing. As a side-effect of chain extraction, we also identified *missing* (though obvious) links among the classes, which could be added to the ontology, and also some modelling *errors*, especially, domain/range constraints at an inappropriate level of generality.

We divided the strong hypotheses resulting from *4ft-Miner* runs into four groups, with respect to their amenability to ontology-based explanation:

1. *Strict dependencies*, e.g. the association between answers to the questions "Do you use a public means of transport?" and "Which public means of transport do you use?". Such results are of no interest in KDD and could of course be eliminated with more careful task design.

2. Relationships amounting to *obvious causalities*, for example, the association between "Are you satisfied with the location where you live?" and "Do you intend to move?" Such relationships (in particular, their strengh) might be of some interest for KDD in general; however, there is no room for ontology-based explanation, since both the antecedent and succedent are mapped on the same or directly connected ontology concepts (`Location`, `livesIn`, `movesFrom` etc.).
3. Relationships between signs that have the character of respondent's agreement with relatively *vague propositions*, for example "Our society changes too fast for a man to follow." and "Nobody knows what direction the society is taking." We could think of some complex ontology relationships, however, by Occam's razor, it is natural just to assume that the explanation link between the antecedent and succedent goes through the categorisation of the respondent as conservative/progressist or the like.
4. Relationships between signs corresponding to concrete and relatively *semantically distant* questions (in fact, those appearing in different thematic groups in the questionnaire). This might be e.g. the question "Do you expect that the standard of living of most people in the country will grow?", with answer 'certainly not', and the question "Which among the parties represented in the city council has a programme that is most beneficial for Prague?" with 'KSČM' (the Czech Communist Party) as answer. Such *cross-group* hypotheses are often amenable to ontology-based explanation.

The last hypothesis mentioned, formally written as `Z05(4)` $\sim^{+}_{0.22,64}$ `Z18(3)`, can be visualised in *4ft-Miner* by means of *four-fold contingency table*, as shown at Fig. 5, and also graphically (see [20]). The contingency table (followed with a long list of computed characteristics) shows that:

- 64 people disagree that the standard of living would grow AND prefer KSČM
- 224 people disagree that the standard of living would grow AND DO NOT prefer KSČM
- 171 people DO NOT disagree[8] that the standard of living would grow AND prefer KSČM
- 2213 people DO NOT disagree that the standard of living would grow AND DO NOT prefer KSČM.

We can see that among the people who disagree that the standard of living would grow, there is a 'substantially' higher number of people who also prefer KSČM than in the whole data sample, and vice versa[9]. The whole effort of formulating hypotheses about the reason for this association is however on the shoulders of the human expert.

In order to identify potential *explanation templates*, we took advantage of the *mapping* created prior to the knowledge discovery phase, see section 4. The

[8] More precisely, their answer to the question above was not 'certainly not'; it was one of 'certainly yes', 'probably yes', 'probably no'.

[9] This is the principle of the *above-average* quantifier, which is symmetrical.

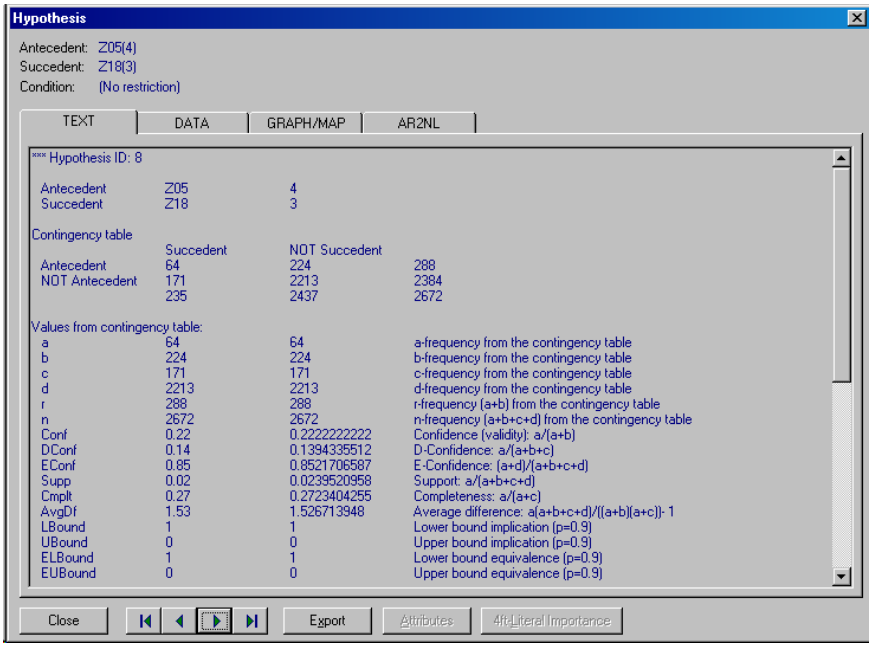

Fig. 5. Textual view of a *4ft-Miner* hypothesis

negative answer to the question about standard of living was mapped to the individual `BAD_LIVING_STANDARD` (instance of `Social_phenomenon`), and the respective answer to the question about political parties was mapped to the class `Political_party`, to its instance `KSCM`, to the class `Party_programme` and to the class `City_council`. Table 3 lists some among the possible templates,first ordered by the decreasing number of involved entities on which the hypothesis is *mapped* and then by the decreasing number of *all* involved entities. The templates do not contain intermediate classes from the hierarchy (which are not even counted for the ordering). Relations are only considered as linked to the class for which they are directly defined as domain/range, i.e. not to the class that inherits them. The symbols $\sqsubseteq$, $\sqsupseteq$ stand for subclass/superclass relationship and $\in$, $\ni$ for instance-to-class membership[10].

We can see that the 'most preferable' template suggests that the KSČM party may have some programme that may have as objective to reach the phenomenon of `BAD_LIVING_STANDARD`. The second looks a bit more adequate: the KSČM party is represented in the city council that can carry out an economic action that may have some impact on the phenomenon of `BAD_LIVING_STANDARD`. The third is almost identical to the first one. The fourth (and simplest) might

[10] Note that the description-logic-like notation is only used for brevity; a more user-oriented (e.g. graphical) representation would probably be needed to provide support for a domain expert not familiar with knowledge representation conventions.

Template	Mapped	All
KSCM ∈ **Political_party** hasPartyProgramme **Party_programme** ⊑ Plan_of_action hasObjective Social_phenomenon ∋ **BAD_LIVING_STANDARD**	4	6
KSCM ∈ **Political_party** isRepresentedIn Administrative_body ⊒ **City_council** carriesOutAction Economic_action hasImpactOn Social_phenomenon ∋ **BAD_LIVING_STANDARD**	4	7
KSCM ∈ **Political_party** hasPartyProgramme **Party_programme** ⊑ Plan_of_action envisagesAction Action ⊒ Economic_action hasImpactOn Social_phenomenon ∋ **BAD_LIVING_STANDARD**	4	8
KSCM ∈ Group informsAbout Social_phenomenon ∋ **BAD_LIVING_STANDARD**	2	3
KSCM ∈ Group carriesOut Action ⊒ Economic_action hasImpactOn Social_phenomenon ∋ **BAD_LIVING_STANDARD**	2	6
KSCM ∈ Group participatesIn Event ⊒ Economic_action hasImpactOn Social_phenomenon ∋ **BAD_LIVING_STANDARD**	2	6
KSCM ∈ Group supports Action ⊒ Economic_action hasImpactOn Social_phenomenon ∋ **BAD_LIVING_STANDARD**	2	6
KSCM ∈ Group fightsAgainst Group carriesOutAction Action ⊒ Economic_action hasImpactOn Social_phenomenon ∋ **BAD_LIVING_STANDARD**	2	7

Table 3. Explanation templates for 'standard of living' vs. 'KSČM' association

actually be most plausible: the KSČM party informs about the phenomenon of BAD_LIVING_STANDARD. Let us finally mention the fifth template, which builds on an incorrect 'inference' (caused by imprecise modelling): the party is assumed to carry out an economic action, which it (directly) can't. The relation was defined with Group and Action as subsets of its domain and range, respectively. However, the combination of Political_party (subclass of Group) and Economic_action (subclass of Action) is illegal and should have been ruled out by an axiom such as Political_party ⊒ (ALL carriesOutAction (NOT Economic_action)).

7 Result Deployment over Semantic Web

The role of ontology in the deployment phase is most crucial if the mining results are to be supplied to a wider (possibly unrestricted) range of consumer applications. A promising approach would be to incorporate the mining results into *semantic web* documents. The most straightforward way to do so is to take advantage of *analytic reports* (ARs): textual documents presenting the results of KDD process in a condensed form. ARs are produced by humans, although

the use of natural language generation was also studied [18]. Prior to entering the reports, the sets of discovered hypotheses (understood as formulae in the so-called observational calculus) can be transformed using formal *deduction rules* [15, 16] into 'canonical form' (which is, among other, free of redundancies). In the *cardiovascular risk* application, a collection of ARs has been created by junior researchers based upon results of selected *4ft-Miner* tasks on STULONG data. Similarly, in the *social climate* application, an almost exhaustive collection of (about 60) ARs have been created for different task settings (combinations of attribute groups), by undergraduate students as part of their assignment.

In order to embed the formal representation of *4ft-Miner* results themselves into the text of the reports [11], we initially used an original XML-based language conforming to early RuleML specifications [1]. A more up-to-date option would be to combine such rules with an ontology (mapped on data attributes, cf. section 4), as proposed e.g. by the Semantic Web Rule Language [**?**]. However, we also consider another option, which would go in the spirit of ontology learning [6, 12]: to use association rule mining to learn (skeletons of) *OWL ontologies* from data. The knowledge contained in the analytic reports would then be represented as ontology axioms rather than as rules, which would enable us to exploit description logic reasoners to formally compare the sets of results. The decision whether to replace rules with OWL axioms in modelling logical implications in semantic web theories would probably be based on two aspects:

- the number of variables in the rule: if there is only one variable then the expression can usually be expressed in OWL (as concept subsumption)
- the nature of the implication: if it has 'conceptual' nature then it should be modelled in OWL if possible, while if it is 'ad hoc' (say, purely empirical) then rules might be a better choice.

The first aspect makes OWL a preferable choice for our case, providing the association mining is carried out on a single data table. Then the resulting taxonomy is subordinated to a single ontology node, such as Patient (in the cardiovascular risk application) or Citizen_of_Prague (in the social climate application). Using association mining (or similar methods such as Formal Concept Analysis) to derive such taxonomy then looks quite obvious and not very novel. However, as the input data columns are not the only prior structure over the results—the initial domain ontology could also used, namely, to 'unfold' some concepts in the taxonomy into restrictions over object properties. For example, from the associations discovered in the sociological domain, we can construct taxonomy path such as Citizen_of_Prague $\sqsupseteq$ KSČM_supporter $\sqsupseteq$ Inhabitant_of_District_14 $\sqsupseteq$ Citizen_wishing_to_move_to_District_15. We can then unfold the concept of ODS_supporter to a 'to-value' restriction Inhabitant_of_District_14 $\sqsubseteq$ (supports $\in$ KSČM). In order to preserve the information content of the original hierarchy, unfolding should not be carried out for both the parent and child concept (i.e., at most every other concept along the path can be left out).

As the associations are equipped with confidence factors, we should consider some formalism for modelling impreciseness, such as fuzzy versions of OWL.

8 Envisaged Tool Support

The *LISp-Miner* tool was used for data mining in this work. It is the original tool for GUHA based procedures and includes five other mining procedures in adition to *4ft-Miner*. However since 2002 there has been a new system under development named *Ferda*[11]. Creators of *Ferda* aimed to create an open user-friendly system based on the long-term experience of *LISp-Miner*, strongly relying on principles of visual programming. 6 shows the *Ferda* environment with a setting of a task examining the validity of the quantitative rule increase of cholesterol leads to increase of triglycerides level mentioned in Section 2.1. Other main feature of the *Ferda* system is extensibility. User can easily add a new module to the system and this module can communicate with other modules via predefined interfaces. [23] describes the *Ferda* system in more details. Because of the extensibility and existing implementation of GUHA procedures in *Ferda*, it is highly preferable to implement new tools connecting ontologies and association mining in this system. Currently, it is possible in *Ferda* to construct and validate quantitative rules against hypotheses generated by data mining runs. There are also some design proposals for modules that would include ontologies into the recent task setup[12]. [24] comprises these efforts.

9 Related Work

Although domain ontologies are a popular instrument in many diverse applications, they only scarcely appeared in 'tabular' KDD until very recently. A notable exception was the work by Philips & Buchanan [14], where 'common-sense' ontologies of time and processes were exploited to derive constraints on attributes, which were in turn used to construct new attributes. Although not explicitly talking about ontologies, the work by Clark & Matwin [8] is also relevant; they used qualitative models as bias for inductive learning. Finally, Thomas et al. [21] and van Dompseler & van Someren [22] used problem-solving method descriptions (a kind of 'method ontologies') for the same purpose. There have also been several efforts to employ taxonomies over domains of individual attributes [3, 4, 13, 19] to guide inductive learning. A recent contribution that goes in similar direction with our work but is more restricted in scope is that of ??? [**?**], which uses ontologies to...

10 Conclusions and Future Work

We presented a pilot study on using ontologies to enhance the knowledge discovery process; the study was carried along most phases of the process and targeted

[11] *Ferda* can be downloaded at `http://ferda.sourceforge.net`

[12] In the future, the *Ferda* community aims to implement these proposals from automatic identification of redundant attributes, automatic categorization of attributes to automated setup of the whole task.

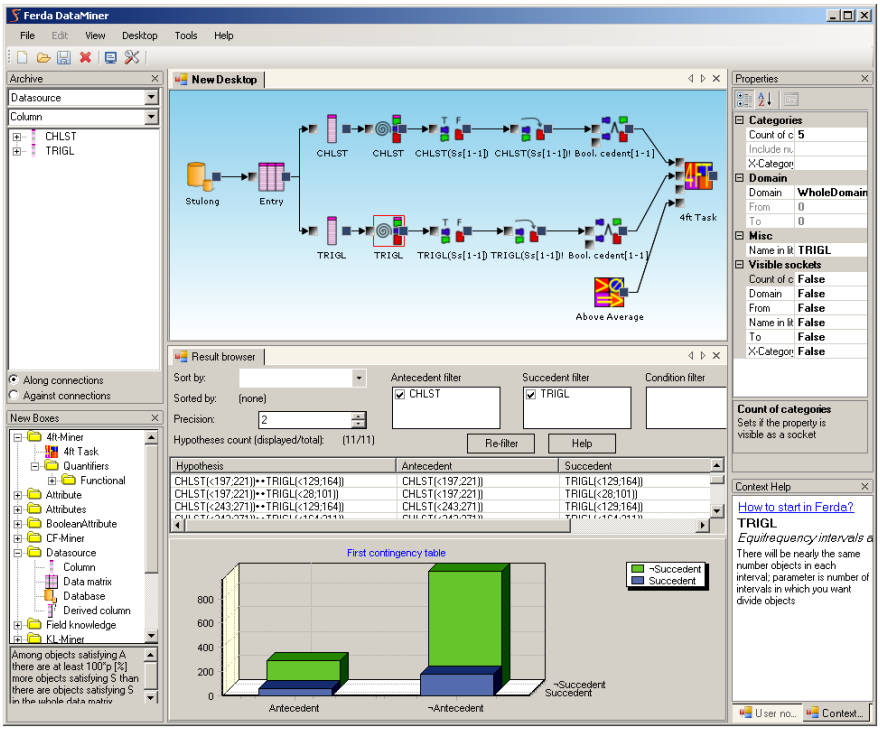

Fig. 6. The *Ferda* environment

into two applications: cardiovascular risk and social climate. We plan to use the outcome of this study to integrate ontology-handling tools to the *LISp-Miner* architecture itself. In a longer run, it would also be desirable to extend the scope of the project towards discovered hypotheses with *more complex structure*, e.g. with longer antecedents/succedents, with additional condition, or even to hypotheses discovered by means of a different procedure than *4ft-Miner*.

Acknowledgements

The research is partially supported by the grant no.201/05/0325 of the Czech Science Foundation, "New methods and tools for knowledge discovery in databases". We also acknowledge the contribution of our research colleagues and domain experts to partial results shown, namely, of Hana Češpivová, Miroslav Flek, Martin Kejkula and Marie Tomečková.

References

1. The Rule Markup Initiative, `http://www.ruleml.org/`.
2. Unified Medical Language System, `http://www.nlm.nih.gov/research/umls/`.
3. Almuallim, H., Akiba, Y. A., Kaneda, S.: On Handling Tree-Structured Attributes in Decision Tree Learning. In: Proceedings of the Twelfth International Conference on Machine Learning (ML-95). Morgan Kaufmann, 12–20.
4. Aronis, J.M., Provost, F.J., Buchanan, B.G.: Exploiting Background Knowledge in Automated Discovery. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996 (KDD-96).
5. Berendt, B., Hotho, A., Stumme, G.: 2nd Workshop on Semantic Web Mining, held at ECML/PKDD-2002, Helsinki 2002, `http://km.aifb.uni-karlsruhe.de/semwebmine2002`.
6. Buitelaar, P., Cimiano, P., Magnini, B. (eds.): Ontology Learning and Population, IOS Press, 2005.
7. Češpivová, H., Rauch, J., Svátek V., Kejkula M., Tomečková M.: Roles of Medical Ontology in Association Mining CRISP-DM Cycle. In: ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa 2004.
8. Clark, P. Matwin, S.: Using Qualitative Models to Guide Inductive Learning. In: Machine Learning - ECML'94, European Conference on Machine Learning, Catania 1994. Lecture Notes on Artificial Intelligence, Springer Verlag 1994, 360–365.
9. Gómez-Perez, A., Fernández-Lopez, M., Corcho, O.: Ontological Engineering. Springer 2004.
10. Hájek, P., Havránek, T.: *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory.* Springer-Verlag: Berlin - Heidelberg - New York, 1978.
11. Lín, V., Rauch, J., Svátek, V.: Content-based Retrieval of Analytic Reports. In: Schroeder, M., Wagner, G. (eds.). Rule Markup Languages for Business Rules on the Semantic Web, Sardinia 2002, 219–224.
12. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer, 2002.
13. Núñez, M.: The Use of Background Knowledge in Decision Tree Induction. *Machine Learning*, 6, 231–250 (1991).

14. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. In: International Conf. Knowledge Capture, Victoria, Canada, 2001.
15. Rauch, J.: Logical Calculi for Knowledge Discovery in Databases. In: Principles of Data Mining and Knowledge Discovery (PKDD-97), Springer-Verlag, 1997.
16. Rauch, J.: Logic of Association Rules. *Applied Intelligence*, 22, 9-28, 2005.
17. Rauch, J., Šimůnek, M.: An Alternative Approach to Mining Association Rules. In: Lin, T. Y., Ohsuga, S., Liau, C. J., Tsumoto, S. (eds.), Data Mining: Foundations, Methods, and Applications, Springer-Verlag, 2005, pp. 211–232
18. Strossa, P., Černý, Z., Rauch, J.:. Reporting Data Mining Results in a Natural Language. In: Lin, T. Y., Ohsuga, S., Liau, C. J., Hu, X. (ed.): Foundations of Data Mining and Knowledge Discovery. Berlin : Springer, 2005, pp. 347-362
19. Svátek, V.: Exploiting Value Hierarchies in Rule Learning. In: van Someren, M. - Widmer, G. (Eds.): ECML'97, 9th European Conference on Machine Learning. Poster Papers. Prague 1997, 108–117.
20. Svátek, V., Rauch, J., Flek, M.: Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality. In: The 2^{nd} International Workshop on Knowledge Discovery and Ontologies, held with ECML/PKDD 2005, Porto, p. 75-86.
21. 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, 123–138.
22. van Dompseler, H. J. H., van Someren, M. W.: Using Models of Problem Solving as Bias in Automated Knowledge Acquisition. In: ECAI'94 - European Conference on Artificial Intelligence, Amsterdam 1994, 503–507.
23. Kováč M., Kuchař T., Kuzmin A.: Ferda, New Visual Environment for Data Mining. In: Znalosti 2006, Conference on Data Mining, Hradec Králové 2006, 118–129
24. Ralbovský M.: Usage of Domain Knowledge for Applications of GUHA Procedures, Master thesis, Faculty of Mathematics and Physics, Charles University, Prague 2006