Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality

Vojtěch Svátek[1], Jan Rauch[1] and Miroslav Flek[2]

[1] 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
[2] University College of International and Public Relations, Prague, U Santošky 17,
150 00 Praha 5, Czech Republic
e-mail: flek@vip-vs.cz

Abstract. One of possible uses of domain ontology as prior knowledge in KDD is the generation of explanations for discovered hypotheses. We developed an ontology covering a subset of concepts and relations relevant for 'municipal social reality', and manually mapped these entities on the structure of sociological data; the data originated from a comprehensive opinion poll over citizens of the capital city of Prague. The LISp-Miner KDD tool was then applied on data, and the most conspicuous associations were matched with the ontology. Some of the extracted ontology structures seem to offer useful insight into the background of empirical associations, as high-level templates that can be instantiated to concrete explanations.

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 [3] 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.

One of the main outcomes of the first *Workshop on Knowledge Discovery and Ontologies* [5] was that the role of prior knowledge is underestimated by the KDD community, and even if this knowledge is used, it is rarely underpinned by a clear conceptual model. However, [6] demonstrated that ontologies can be beneficial in nearly all phases of the KDD (more specifically, association mining) cycle, starting from domain and data understanding, through the semantic interpretation of discovered hypotheses, and ending by exposing the hypotheses on the semantic web, e.g. in the form of annotated textual reports [9]. Here we pay attention to the middle phase, in which the ontology is to provide 'templates' for the human expert who attempts to interpret the discovered knowledge.

The paper is structured as follows. Section 2 describes the process of designing our ontology of social reality, in a bottom-up manner. Section 3 recalls the basic principles of the LISp-Miner system which was used as knowledge discovery tool. Section 4 presents the actual experiments with using the ontology as prior knowledge for (further) knowledge discovery. Finally, section 5 reviews some related work, and section 6 shows directions for future research.

2 Designing and Mapping the Ontology

2.1 State of the Art in Social Ontology Modelling

The society as such has mostly been subject of ontology research at the philosophical level. Probably the best known recent example is the work by Searle[3]. Some notions of social reality also appeared in formal ontological engineering, for example, in the 'social' fragment of the DOLCE upper-level ontology[4], which contains concepts such as 'social relationship' or 'social institution'. Similarly, Boella & van der Torre [4] recently developed an upper-level model of social reality centred around the concept of 'agent'. In a bottom-up manner, on the other hand, a tiny fragment of social reality (namely, the relationships among and the most imminent attributes of persons) has been studied by the FOAF community, see e.g. [10]. What we however needed in our project was a comprehensive formal model spanning across many heterogeneous areas; we therefore decided to create a new ontology, in a *bottom-up* manner.

2.2 Designing the Ontology

Both the *ontology* and the *dataset* used for association discovery had the same seed material: the questionnaire[5] posed to respondents during the *opinion poll* mapping the 'social climate' of the capital city of Prague in Spring 2004. The questionnaire contained 51 questions related to e.g. economic situation of families, ways of earning money and 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'. While the *dataset* was straightforwardly

[3] See [15] for a summary.
[4] `http://dolce.semanticweb.org/`
[5] The questionnaire was designed by sociology experts, entirely independent of the KDD and ontological engineering research described in this paper.

derived from the individual 'signs', each becoming a database column[6], 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 [8], 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 corresponded to enumerated values of properties (modelled according to the W3C note [14]), e.g. `GOOD_JOB_AVAILABILITY`, or to outstanding individuals such as `PRAGUE` or `CHRISTIAN_DEMOCRATIC_PARTY`, as these were often referred to in the text of the questionnaire.

The current version of the ontology, eventually formalised in OWL[7], consists of approx. 100 classes, 40 relations and 50 individuals[8]. A Protégé[9] 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. More details on the process of designing the ontology (in particular, the 'design patterns' used) can be found in [17].

2.3 Data-to-Ontology Mapping

The second and somewhat easier part of the knowledge engineering phase of our project was to *map* the attributes of the dataset to ontology concepts, relations and instances. 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 Table 1. 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.).

---

6 And, subsequently, an attribute for the LISp-Miner system, see the next section.

7 `http://www.w3.org/2004/OWL`

8 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.

9 `http://protege.stanford.edu`

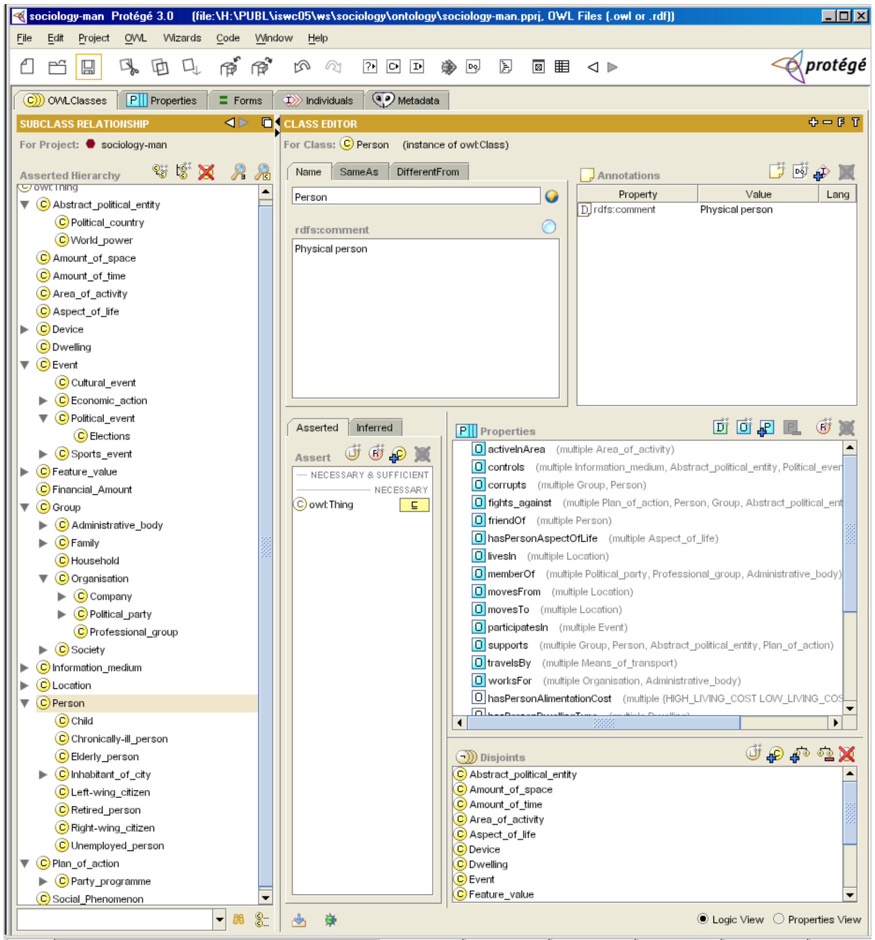

Fig. 1. Incomplete view of the ontology in Protégé

3 Association Mining with 4ft-Miner

The *4ft-Miner* procedure is the most frequently used procedure of the *LISp-Miner* data mining system [13]. It mines for association rules of the form $\varphi \approx \psi$, where φ and ψ are called *antecedent* and *succedent*, respectively[10]. 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) 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 $v \in \alpha$.

[10] *4ft-Miner* also mines for *conditional* hypotheses (i.e. with a third symbol representing a restrictive condition). We will not discuss them here, for brevity.

From May 1, 2004, *Prague* will become one of *Central-European metropolies* of the *EU*. Do you think that this fact will improve the *availability of jobs* for you or for your *relatives*?

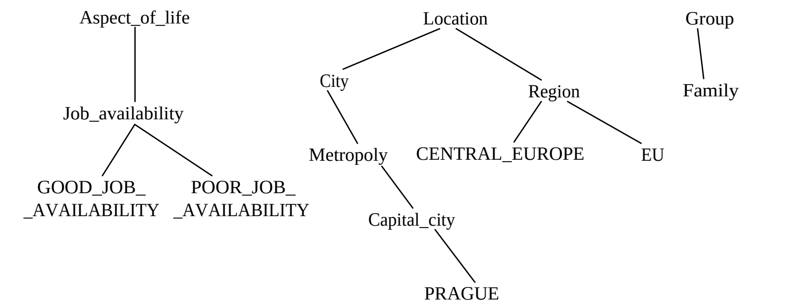

Table 1. Question no.3 and fragments of ontology used for its mapping

The association rule $\varphi \approx \psi$ means that the Boolean variables φ and ψ are associated in the way defined by the symbol $\approx$. The symbol $\approx$ is called *4ft-quantifier*. It corresponds to a 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 the 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 $A(a_1, a_7)$[11], and analogously for the remaining literals.

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

Note that the hypothesis definition language of 4ft-Miner is far richer than we described. For the sake of this paper, the description above is sufficient; for more information see e.g. the project homepage `http://lispminer.vse.cz` or [13].

4 Experiments

4.1 Overview

We experimented with various 4ft-Miner settings on the poll dataset, mostly using the *above-average dependence* quantifier explained in previous section. As we did not want to restrict the choice of antecedent and succedent of hypotheses, between which the chains of ontology entities were to be found, we 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. As we wanted to start with (structurally) simplest possible 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 value of the attribute). The run-times were typically lower than a second.

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

1. *Strict logical dependencies*, an example of which is 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 hypotheses are of no interest as KDD results in general.
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 (namely, appearing in different question 'groups' or 'clusters'). 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. We'll elaborate on this particular example in the following subsection.

Since we do not (yet) have an appropriate software support for extracting entity chains (i.e. explanation templates) from the ontology, we 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.

4.2 Example of Explanation Template Set

The hypothesis from the last example above, formally written as $\texttt{Z05(4)} \sim^{+}_{0.22,64}$ $\texttt{Z18(3)}$, could be visualised by the available means of LISp-Miner as shown at Fig. 2 and Fig. 3.

The first view presents the *four-fold contingency table*:

- 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[12] 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.

The contingency table is followed with a long list of computed characteristics.

The second view presents the same information *graphically*. 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[13]. 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 2.3. The 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`.

There are many ways of *ordering* the explanation templates; here we order them first 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

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

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

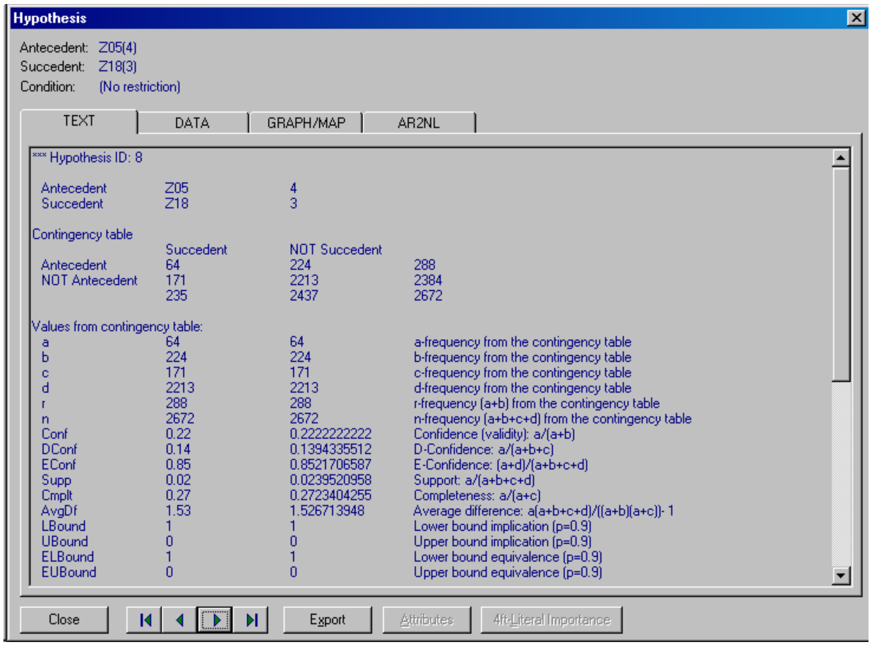

Fig. 2. Textual view of a LISp-Miner hypothesis

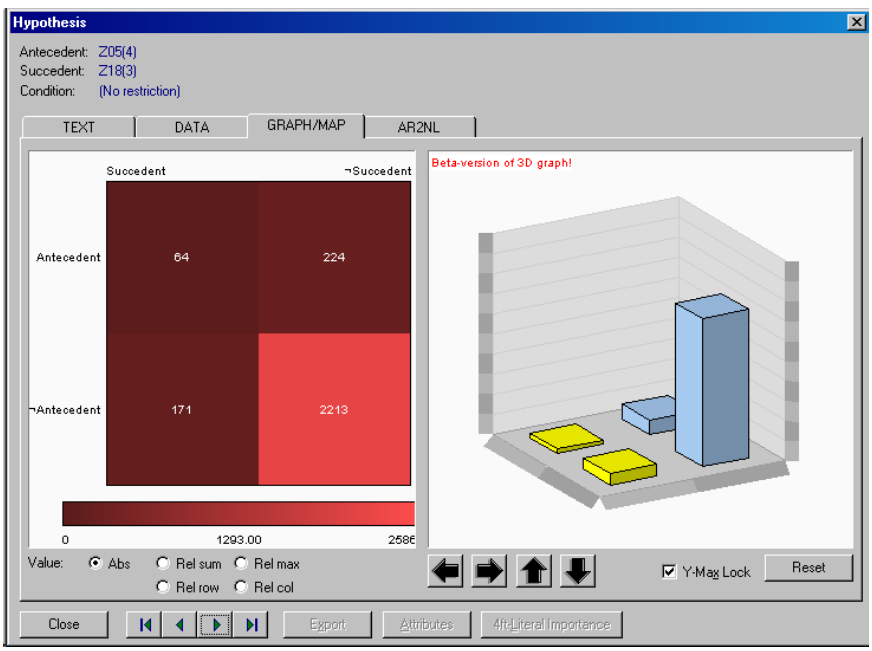

Fig. 3. Graph-based view of a LISp-Miner hypothesis

Template	Mapped	All
KSCM $\in$ **Political_party** `hasPartyProgramme` **Party_programme** $\sqsubseteq$ `Plan_of_action hasObjective Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	4	6
KSCM $\in$ **Political_party** `isRepresentedIn Administrative_body` $\sqsupseteq$ **City_council** `carriesOutAction Economic_action hasImpactOn Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	4	7
KSCM $\in$ **Political_party** `hasPartyProgramme` **Party_programme** $\sqsubseteq$ `Plan_of_action envisagesAction Action` $\sqsupseteq$ `Economic_action hasImpactOn Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	4	8
KSCM $\in$ `Group informsAbout Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	2	3
KSCM $\in$ `Group carriesOut Action` $\sqsupseteq$ `Economic_action hasImpactOn Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	2	6
KSCM $\in$ `Group participatesIn Event` $\sqsupseteq$ `Economic_action hasImpactOn Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	2	6
KSCM $\in$ `Group supports Action` $\sqsupseteq$ `Economic_action hasImpactOn Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	2	6
KSCM $\in$ `Group fightsAgainst Group carriesOutAction Action` $\sqsupseteq$ `Economic_action hasImpactOn Social_phenomenon` $\ni$ **BAD_LIVING_STANDARD**	2	7

Table 2. Explanation templates for 'standard of living' vs. 'KSČM' association

counted for the ordering). Relations, i.e. OWL properties, are only considered as linked to the class for which they are directly defined as domain/range, i.e. not to the class that just inherits them. Table 2 lists some (by far not all) possible templates, with the counts of *mapped* and *all* entities of which the template consists, respectively. The symbols $\sqsubseteq$, $\sqsupseteq$ stand for subclass/superclass relationship and $\in$, $\ni$ for instance-to-class membership[14].

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 actually be most plausible: the KSČM party informs about the phenomenon of

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

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 $\sqsupseteq$ (ALL carriesOutAction (NOT Economic_action)).

5 Related Work

Although domain ontologies are a popular instrument in many diverse applications, they only scarcely appeared in 'tabular' KDD, so far. A 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. Although not explicitly talking about ontologies, the work by Clark & Matwin [7] is also relevant; they used qualitative models as bias for inductive learning. Finally, Thomas et al. [18] and van Dompseler & van Someren [19] 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 [1, 2, 11, 16] to guide inductive learning. None of these projects however attempted to explore the role of domain ontology in *interpreting* the results of the mining process.

For a brief review of related work on *social ontology modelling* proper see section 2.1.

6 Conclusions and Future Work

We described a simple experiment in matching a social reality ontology to hypotheses discovered via data mining from poll data; abstract templates for possible explanations of the hypotheses were identified.

The work is only in its early phase, as our ontology reflects the state of affairs in our 'domain' in a very imprecise and simplified way[15]. Its further extension and refinement in close contact with the expert is envisaged; we also plan to take into account prior work in (philosophical as well as applied) social reality modelling mentioned in section 2.1. We would also like to pay more attention to expressing (mainly as relation instances) *additional heuristic knowledge* available in our domain, which could help automatically fill the templates with concrete relationships. Such a (not yet formalised) knowledge base actually arose in connection with the polls in question.

With growing body of available knowledge, *end-user tests* would become more meaningful. An important step would be to proceed from the current,

[15] An important problem, which is however not easy to overcome by state-of-the-art ontology engineering technology, is the static character of our model.

subjective, evaluation of patterns to quantitative evaluation of their efficiency in supporting the interpretation of hypotheses.

Furthermore, we would like to follow up with our earlier effort to expose KDD results on the *semantic web* [9]. Aside 'plain' empirical hypotheses, instantiated explanation templates endorsed by an expert could straightforwardly be represented.

From the point of view of *association discovery*, the experiments revealed the utility of further extensions to the task definition principles of LISp-Miner, in particular regarding the search for *cross-group* hypotheses. Such extensions would make further experiments with ontologies or similar background models more efficient.

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. An example of the last is the recently implemented procedure SD4FT; it searches for pairs of sets of objects in data such that one appears in different empirical associations than the other. Generating explanations for such hypotheses would be much more demanding but would also provide greater benefits.

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".

References

1. 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.
2. 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).
3. Berendt, B., Hotho, A., Stumme, G.: 2nd Workshop on Semantic Web Mining, held with at ECML/PKDD-2002, Helsinki 2002, `http://km.aifb.uni-karlsruhe.de/semwebmine2002`.
4. Boella, G., van der Torre, L.: An Agent-Oriented Ontology of Social Reality. In: Proc. FOIS'04, Torino 2004, Springer LNCS.
5. Buitelaar, P., Franke, J., Grobelnik, M., Paass, G., Svátek, V. (eds.): ECML/PKDD 2004 Workshop on Knowledge Discovery and Ontologies (KDO-04), Pisa 2004.
6. 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.
7. 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.

8. Gómez-Perez, A., Fernández-Lopez, M., Corcho, O.: Ontological Engineering. Springer 2004.
9. 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.
10. Matsuo, Y., Hamasaki, M., Mori, J., Takeda, H., Hasida, K.: Ontological Consideration on Human Relationship Vocabulary for FOAF. In: 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway 2004.
11. Núñez, M.: The Use of Background Knowledge in Decision Tree Induction. *Machine Learning*, 6, 231–250 (1991).
12. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. In: International Conf. Knowledge Capture, Victoria, Canada, 2001.
13. Rauch, J., Šimůnek, M.: An Alternative Approach to Mining Association Rules. In: Lin, T. Y., Ohsuga, S., Liau, C. J., and Tsumoto, S. (eds.), Data Mining: Foundations, Methods, and Applications, Springer-Verlag, 2005, pp. 219–238 (*to appear*)
14. Rector, A. (ed.): Representing Specified Values in OWL: "value partitions" and "value sets". W3C Working Group Note, 17 May 2005, online at `http://www.w3.org/TR/swbp-specified-values/`.
15. Searle, J. R.: Social Ontology: Some Basic Principles. Available online from `http://ist-socrates.berkeley.edu/~jsearle/articles.html`.
16. 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.
17. Svátek, V.: Observations from Development of Social Reality Ontology for a KDD Application. Submitted paper.
18. 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.
19. 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.