Ontologies, Questionnaires and (Mining) Tabular Data

Vojtěch Svátek
Department of Information and Knowledge Engineering, University of Economics, Prague
W. Churchill Sq. 4, 13067 Praha 3, Czech Republic
svatek@vse.cz

ABSTRACT

Questionnaires are an interesting source for ontology design, especially in connection with KDD applications. Two case studies from different domains are presented.

Keywords

Ontological engineering, knowledge discovery from databases

1. INTRODUCTION

Ontology-based knowledge discovery techniques are mostly applied on loosely structured textual or multimedial data. However, a significant portion of the information wealth of the mankind is latently present in structured databases from which it can be extracted by means of KDD (knowledge discovery from databases) techniques. In the rest of this text, we will refer to this kind of data as to *tabular*. Recently, the potential role of ontologies as prior knowledge in the KDD process has been discussed in the framework of workshops on 'Knowledge Discovery and Ontologies' [1, 2].

In this paper we also take into account a third resource, which seems to have genuine connection both to ontologies and databases, namely, *questionnaires* that are often used to collect tabular data. Typical interactions among the three types of resources are depicted inFull lines correspond to creation of resources, while dashed ones correspond to provision of additional information.

The questionnaire has twofold impact on the database: its structure is transferred into that of the data, and the textual labels clarify the semantics of the fields to humans. The texts in the questionnaire can, however, also serve as resource of ontology entities: classes, relations as well as instances (to say, values of closed questions). Similarly, the data tables (namely, values of fields corresponding to open questions) can serve as resource of instances for the ontology. Finally, the ontology can impact the analysis of tabular data in several ways: to (semi-)automatically focus the mining process, provide interpretations of discovered results, allow to expose the results on the semantic web and the like [3].

Demos and Posters of the *3rd European Semantic Web Conference (ESWC 2006)*, Budva, Montenegro, 11th - 14th June, 2006

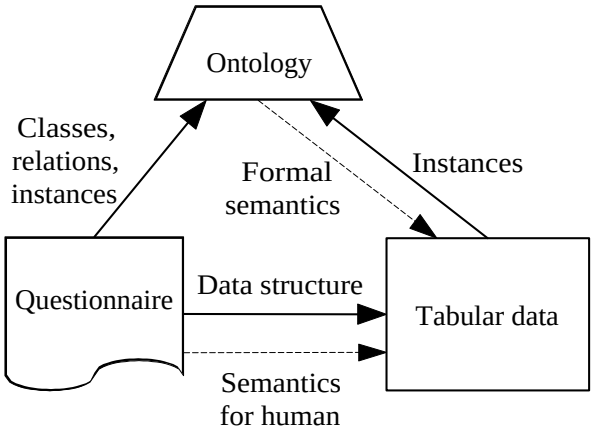

Figure 1: Interactions among the three resources

In section 2 we briefly discuss the characteristics of questionnaires as input for ontology design. In section 3 we report on case studies undertaken in two different projects. Finally, section 4 gives a summary and future plans.

2. QUESTIONNAIRES AS RESOURCES IN ONTOLOGY DESIGN

In ontology engineering methodologies, pre-existing questionnaires are hardly considered as stand-alone resources; Obviously, they are not always available and their scope is typically too narrow for a large-scale domain ontology. However, when incrementally building an ontology covering the portion of domain relevant to a given collection of tabular data, questionnaires may be quite useful:

- their *small extent* and *richness* in domain terms makes them amenable to manual processing with no or just light-weighted suppport by automated NLP
- most terms are relevant not only wrt. the domain but also wrt. the *applications* that would process the data; this alleviates the well-known 'hugeness' problem of ontology design
- the structure of questions (and answer options) may provide some cues to resulting *ontology structures*
- the *mapping* between data fields and ontology concepts is immediately available for future use.

In the following, we illustrate these advantages on two case studies.

3. CASE STUDIES

3.1 Social Reality Ontology

One of goals recently formulated in the *LISp-Miner* project[1] was to exploit domain ontologies to enhance the KDD (in particular, association mining) process and its results. For one of initial experiments, a dataset was chosen that contained information from more than 3000 respondents concerning their attitude towards various events of social and political life in the city of Prague (in 2004, the year of the country entering the EU). As there was no suitable domain ontology at hand, a new one was manually designed in a bottom-up manner, taking the comprehensive *questionnaire* (with about 50 questions) used in the poll, as starting point. Terms from the questionnaire were upgraded to concepts, relations and instances for the ontology, while keeping the mapping between the data fields and ontology entities. Only a few more entities were later added, in cooperation with a domain expert, to achieve connectivity of the whole model. The resulting (OWL) ontology contained about 100 classes, 40 relations and 50 instances. Eventually, the database was analysed using the *LISp-Miner* [4] tool, and some of the discovered associations were endowed with potential 'explanation paths' from the ontology [7].

Several explanation paths were characterised as interesting and to some degree 'plausible' by the domain expert. See for example the path "KSCM $\in$ Political_party *isRepresentedIn* Administrative_body $\sqsupseteq$ City_council *carriesOutAction* Economic_action *hasImpactOn* Social_phenomenon $\ni$ bad_living_standard". It explains the empirical association between 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, (roughly) as "KSČM party is represented in the city council, which can carry out an economic action, which may have some impact on the phenomenon of bad living standard".

3.2 Conference Organisation Ontologies

The *OntoFarm* project [6] aims at independent development of multiple ontologies of the same domain—that of *conference organisation*—thus providing a benchmark collection for ontology-processing techniques such as automated alignment, distributed reasoning or discovery of implicit design patterns. Most ontologies were designed based on human or automated analysis of either conference-support *software tools* (incl. documentation), *websites* of concrete conference series, or *insider info* on organising a conference. Eventually, we decided to include a fourth resource (obviously covering a fragment of the whole domain only), namely, the *review forms* as special kind of questionnaires. This fragment should address tasks such as identification of gaps and redundancies in the coverage of review forms or identification of potential inconsistencies in the reviews themselves.

The model was first created based on a single review form (for conference A) and then updated based on another form (for conference B), its size eventually amounted to approx. 30 classes and 20 relations. The important finding in this small study was that the structural aspects of the forms bring to light different modelling choices. For example, while form A only suggests to categorise the paper as either *theoretical* or *applicative* (which naturally leads to subclassing of class Paper), form B explicitly introduces the notion of *domain* in which the approach is applied (to be most faithfully modelled using a property such as *appliedIn*). Analogously, while form B only introduces the ordinal ranking according to *originality* (to be probably modelled as property with enumerated value set, cf. [5]), form A explicitly asks about prior papers with *similar content* (which calls for property expressing the 'similarity' relationship).

[1] `http://lispminer.vse.cz`

4. CONCLUSIONS AND FUTURE WORK

Based on two case studies, we discussed the role of questionnaires as input for ontology design, aiming at analysis of tabular data with the help of such ontology.

While the described experiments in questionnaire-based ontology design were manual, we are considering to implement a supporting tool. Such tool would definitely be interactive, would probably rely upon a POS tagger (as most ontology learning tools do), but would also include some kind of field detector (as form analysis tools do) in order to capture e.g. values for closed questions.

5. ACKNOWLEDGMENTS

The research is partially supported by the CSF grant no.201/05/0325 "New methods and tools for knowledge discovery in databases" and by the IGA VSE grant no.26/05 "Methods and tools for ontological engineering"; the OntoFarm project is also partially supported by the Knowledge Web Network of Excellence (IST FP6-507482).

6. REFERENCES

[1] M. Ackermann, B. Berendt, M. Grobelnik, V. Svátek (eds.) ECML/PKDD 2005 Workshop on Knowledge Discovery and Ontologies (KDO-05), Porto 2005.

[2] P. Buitelaar, J. Franke, M. Grobelnik, G. Paass, V. Svátek (eds.) ECML/PKDD 2004 Workshop on Knowledge Discovery and Ontologies (KDO-04), Pisa 2004.

[3] H. Češpivová, J. Rauch, V. Svátek, M. Kejkula, M. Tomečková. Roles of Medical Ontology in Association Mining CRISP-DM Cycle. In: [2].

[4] J. Rauch, M. Šimůnek. 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

[5] A. Rector (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/`.

[6] O. Šváb, V. Svátek, P. Berka, D. Rak, P. Tomášek. OntoFarm: Towards an Experimental Collection of Parallel Ontologies. In: Poster Track of ISWC 2005, Galway.

[7] V. Svátek, J. Rauch, M. Flek. Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality. In: [1].