Combining Data Mining and Ontology Engineering to enrich Ontologies and Linked Data

Mathieu d'Aquin[a], Gabriel Kronberger[b], and Mari Carmen Suárez-Figueroa[c]

[a] KMi, The Open University, Walton Hall, Milton Keynes, UK
m.daquin@open.ac.uk
[b] University of Applied Science Upper Austria, School for Informatics, Communications and Media, Softwarepark 11, 4232 Hagenberg, Austria
Gabriel.Kronberger@fh-hagenberg.at
[c] Ontology Engineering Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Boadilla del Monte, Madrid, Spain
mcsuarez@fi.upm.es

Abstract. In this position paper, we claim that the need for time consuming data preparation and result interpretation tasks in knowledge discovery, as well as for costly expert consultation and consensus building activities required for ontology building can be reduced through exploiting the interplay of data mining and ontology engineering. The aim is to obtain in a semi-automatic way new knowledge from distributed data sources that can be used for inference and reasoning, as well as to guide the extraction of further knowledge from these data sources. The proposed approach is based on the creation of a novel knowledge discovery method relying on the combination, through an iterative 'feedback-loop', of (a) data mining techniques to make emerge implicit models from data and (b) pattern-based ontology engineering to capture these models in reusable, conceptual and inferable artefacts.

Keywords: data mining, ontology engineering, linked data, ontologies

1 Introduction

Due to the rapid growth of the open data and linked data [1] movements, more and more data are being made available and directly accessible from a wide range of domains, areas and organisations. However, while representing an unprecedented, globally available resource, this Web of Data is still far from realising the promises of the Semantic Web [2], as such data are rarely associated with the formal ontologies supposed to characterise and make explicit the knowledge they contain.

Ontologies are knowledge representation artefacts that capture the concepts and relationships relevant to a specific domain. As part of the Semantic Web, they are used to provide common conceptual models over data made available online, in order to facilitate semantic interoperability and inferences. In contrast with this top-down view of capturing knowledge, data mining and knowledge discovery are traditionally concerned with the bottom-up detection of patterns and regularities in data that can be interpreted as corresponding to knowledge models in the domain of the data.

Knowledge discovery from databases [3] has for objective to make knowledge emerge from hidden patterns in large amounts of data. It generally relies on data mining techniques to identify potentially relevant hidden models. The effectiveness of data mining depends on appropriate preparation of the data and interpretation of the results, which are difficult tasks when dealing with heterogeneous, distributed data from multiple sources.

Our research hypothesis is that the interplay of, on the one hand, certain types of data mining approaches, applied at instance level and producing easily interpretable models and, on the other hand, the formalisation of conceptual knowledge in ontologies can provide a virtuous cycle where emerging knowledge is easier to interpret and integrate, and can be used to trigger the emergence of further knowledge from the data. Validating this hypothesis require advances in both the areas of data mining and of ontology engineering that we are discussing in this position paper.

The rest of the paper is organized as follows. We start with a brief overview of the fields of Knowledge Discovery and Ontology Engineering, and of their connections, concluding with the need for new methods taking advantage of the combination of these two approaches to knowledge capture. Then, our proposal for combining both data mining and ontology engineering is presented. Finally, the paper ends with a future application of the proposed approach and some general conclusions.

2 The Landscape of Knowledge Discovery and Ontology Engineering

In this section, we briefly describe the state of the art in Knowledge Discovery and Ontology Engineering as well as the specific contributions of our proposed approach.

2.1 Knowledge Discovery and Data Mining

The knowledge discovery process [3] relies on data mining for finding and extracting new and potentially useful and interesting knowledge from data. Data mining is a "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [4]. The typical process for knowledge discovery is a linear process including the following activities: selection, pre-processing, transformation, data mining, and interpretation and evaluation. Data preparation and result interpretation are often dependent on the particular domain and purpose of the application and the most time-consuming activities in the whole process.

The various algorithmic approaches to data mining range from simple techniques such as linear regression [5] which provide models which are rather easy to interpret (white-box) to more advanced approaches [6, 7] which allow the detection of highly complex patterns in the data, but also produce complex (black-box) models.

Our goal is to address this trade-off by employing a white-box approach [8] to data mining, namely Genetic Programming (GP) [9], which can be easily integrated with a-priori knowledge contained in ontologies. Due to its model representation, GP is

able to produce human interpretable results without making strong assumptions about the nature of the relationships within analysed data.

However, even GP-based data analysis suffers from the fact that the results are often complex and far from being unique. One of our objectives is therefore to extend this technique with the integration of ontological knowledge, in such a way that can be used (a) to derive adequate configurations of input variables, (b) to avoid generating trivial or inconsistent results, and (c) to support the interpretation of the resulting models through abstracting/integrating them with supporting ontological knowledge.

2.2 (Pattern-Based) Ontology Engineering

Ontologies [10], which are logical models of the concepts, entities and relationships in a domain, are nowadays one of the most common forms of knowledge representation, as they form one of the pillars of the Semantic Web. Constructing a new ontology for a specific domain is traditionally done manually, requires close cooperation between domain experts and knowledge engineers, and takes a significant amount of time. Also, maintaining an ontology with respect to the evolution of the domain (new findings and models) has to be a continuous task, which is mostly realised manually [11]. A particularly popular approach that emerged recently, following similar trends for example in software engineering, is pattern-based ontology engineering [12]. In this approach, higher-level ontological knowledge patterns are being reused and specialised in different ontologies, to avoid reproducing the same knowledge capture effort in similar modelling contexts and situations.

Considering such an approach, it seems natural to look at integrating knowledge engineering and knowledge discovery to extract potentially interesting patterns. Indeed, works have used various data and text mining algorithms to create ontologies automatically ([13, 14, 15, 16]). While these works focus on automatic ontology construction, a complete knowledge discovery method truly integrating data mining and ontology engineering is yet to be developed.

Three forms of extensions of knowledge discovery with respect to their relation to ontologies can be distinguished [17]: (1) using existing ontologies for knowledge discovery; (2) building ontologies through knowledge discovery; and (3) building and extending ontologies through knowledge discovery using existing ontologies. The last form of integration between ontologies and knowledge discovery is the one that has least been considered in the literature, and where we propose to make significant advances.

3 Combining DM and OE in the Linked Data context

The main aim of this paper is to describe a proposed new method to discover knowledge through mining large quantities of potentially distributed and heterogeneous data, using and enriching pre-existing ontological knowledge. This should represent a paradigm shift with respect to the usual ontology engineering approaches to knowledge capture, as well as significant advances of the state of the art in data mining approaches usually employed in knowledge discovery. We call this new knowledge discovery process, the ***Knowledge and Data Co-Evolution Cycle*** (see

Figure 1) as, similarly to the co-evolution process in biology [18], it creates a virtuous cycle where the creation of knowledge is led by implicit models in the data, and the enrichment of data is informed by their characterisation through explicit knowledge models.

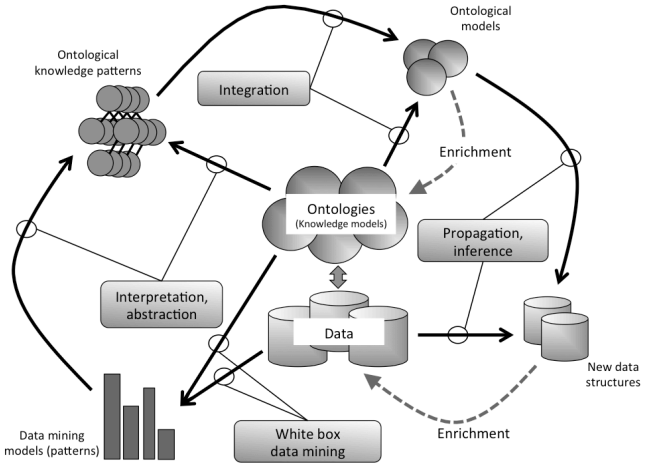

Fig. 1. The Knowledge and Data Co-Evolution cycle

The fundamental principles on which this method for knowledge discovery is based are the following ones:

1. The knowledge discovery process is bootstrapped by pre-existing data and ontologies relevant to the considered domain.
2. Both data and ontologies are evolving over time, through their interactions: ontologies are enriched with knowledge patterns abstracted from the data mining models which are extracted from the data, and data are enriched through new inferences derived from the ontologies.
3. White-box data mining techniques are used to produce interpretable patterns that can be filtered and selected on the basis of their integration with the ontologies (*Mining, Interpretation, Abstraction, Integration*).
4. Ontologies are used to select the input of the data mining techniques, based on their common relevance and on known incompleteness in the knowledge encoded in the ontologies (*Mining*).
5. New ontological models are used both for abstracting and validating (especially in terms of consistency) the identified models, as well as to infer more information, reinforcing and consolidating the data available (*Propagation, Inference, Enrichment*).
6. The process leads to multiple versions of ontologies and data, which branch over multiple iterations. Comparing these models and how they evolve over time is useful knowledge in itself.

An interesting aspect of the novel knowledge discovery method that the propose cycle represents is that it results in alternative models that can evolve independently. While this could be seen as a disadvantage in traditional knowledge discovery approaches, it makes it possible to compare different views corresponding with different

emerging ontologies and data models. This offers the possibility to select or disregard particular models based on several iterations, depending on whether or not they are converging to or diverging from other considered alternatives.

4 How the approach will be tested

The approach to knowledge discovery proposed in this paper will be experimentally tested through confrontation against distributed, heterogeneous data sources connecting the domains of tourism and economy. Particularly, we will apply our knowledge discovery cycle on real-world datasets to produce knowledge related to tourism in the Canary Islands, and to the way it is influenced by the economic state of European countries. These corresponding datasets will originate both from the institute of statistics of the Canary Islands (ISTAC)[1] and from open datasets available on the Web regarding macro-economic indicators [19] (e.g. the 'world bank' datasets[2]). Indeed, immense bases of data about global, national and regional economic conditions exist that have the potential to provide insights about the economic dependencies and future economic potential of particular regions of the world. However, these data are being largely underexploited because the large number of interconnections between these data and data from other domains, the heterogeneity of the applicable models and patterns, as well as the general ambiguity associated with these areas make it difficult to bootstrap a knowledge discovery process.

For example, a data mining model could be produced that establishes a relationship between the average domestic income in Germany and the financial results of hotels in the Canary Islands. This could lead to an extension of the ontologies indicating that a significant proportion of the income of hotels in the Canary Islands come from German nationals. Integrating this form of ontological knowledge pattern (that a type of accommodation takes income from tourist of certain nationalities) can in turn be used to guide the extraction of similar relationships (e.g., for other types of accommodations and other origin countries), or more fine-grained ones (e.g., showing different types of impact depending on the socio-economic category of the customers or cost of the hotel).

5 Conclusions

In this position paper, we have proposed a way to capture knowledge from large amounts of potentially heterogeneous and distributed data more effectively by creating a semi-automatic knowledge discovery cycle where models discovered through data mining are integrated with ontologies, to be reasoned upon and used for the extraction of further knowledge. This approach represents a paradigm shift with respect to traditional methods of both knowledge discovery and knowledge capture. Indeed, through exploiting the interplay of data mining and ontology engineering, we aim at

1 http://www2.gobiernodecanarias.org/istac/dw/indicadores/coyunturaeconomica/lstIndicadores.jsp?codAplicacion=32

2 http://data.worldbank.org/

reducing the need for time consuming data preparation and result interpretation tasks in knowledge discovery, as well as for costly expert consultation and consensus building activities required for ontology building. We will test this approach in the domains of tourism and economy.

References

1. T. Heath and C. Bizer. Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.
2. J. Domingue and D. Fensel (editors). Handbook of Semantic Web Technologies, Springer. 2011. ISBN 978-3-540-92912-3.
3. W. Frawley, G. Piatetsky-Shapiro and C. Matheus. Knowledge Discovery in Databases: An Overview. Ai Magazine, Vol. 13 (1992), pp. 57-70.
4. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining 1996: 1-34.
5. N.R. Draper and H. Smith. Applied Regression Analysis, Wiley Series in Probability and Statistics, 1998.
6. S. Haykin. Neural Networks: A Comprehensive Foundation. Prentice Hall. 1999. ISBN 0-13-273350-1.
7. I. Steinwart and A. Christmann. Support Vector Machines. Springer-Verlag, New York, 2008. ISBN 978-0-387-77241-7.
8. D.T. Larose. Data Mining Methods and Models, Wiley-Blackwell, 2006. ISBN: 978-0471666561.
9. M. Affenzeller, S.M. Winkler, S. Wagner and A. Beham. Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications. CRC Press (Taylor & Francis Group), 2009.
10. S. Staab and R. Studer (editors). Handbook on Ontologies, Springer, 2003. ISBN 978-3540408345.
11. M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta and A. Gangemi (editors). Ontology Engineering in a Networked World, Springer, 2012. ISBN 978-3-642-24793-4.
12. A. Gangemi and V. Presutti. Ontology Design Patterns. In Handbook on Ontologies (Second Edition). Springer. International Handbooks on Information Systems. 2009.
13. C. Blaschke and A. Valencia. Automatic Ontology Construction from the Literature, Genome Informatics, Vol. 13, pp 201–213, 2002.
14. P. Clerkin, P. Cunningham and C. Hayes. Ontology Discovery for the Semantic Web Using Hierarchical Clustering, Trinity College Dublin, Ireland. 2002. TCD-CS-2002-25.
15. A.E. Elsayed, S.R. El-Beltagy, M. Rafea and O. Hegazy. Applying data mining for ontology building, The Central Laboratory for Agricultural Expert Systems, Giza, Egypt. 2007.
16. O. Wuermli, A. Wrobel, S.C. Hui and J.M. Joller. Data Mining For Ontology Building: Semantic Web Overview, Diploma Thesis–Dep. of Computer Science WS2002/2003, Nanyang Technological University.
17. P. Gottgtroy. An Ontology Driven Knowledge Discovery Framework for Dynamic Domains: Methodology, Tools and a Biomedical Case. PhD Thesis, School of Computing and Mathematical Sciences, Auckland University of Technology, 2010.
18. J.N. Thompson. The coevolutionary process. University of Chicago Press, 1994. ISBN 0-226-79760-0.
19. O. Blanchard. Macroeconomics Updated (5th ed.). Englewood Cliffs: Prentice Hall. 2011. ISBN 9780132159869.