Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2007 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2007

Ontology Driven Knowledge Discovery Process: a proposal to integrate Ontology Engineering and KDD

Paulo Gottgtroy
Auckland University of Technology, Paulo.gottgtroy@aut.ac.nz

72. Ontology Driven Knowledge Discovery Process: a proposal to integrate Ontology Engineering and KDD

Paulo Gottgtroy
Auckland University of Technology
Paulo.gottgtroy@aut.ac.nz

Abstract
This paper is concerned with the integration of ontology engineering and the process of knowledge discovery in databases (KDD).It presents a hybrid life, Ontology Driven Knowledge Discovery process and methodology – ODKD, which leverages both ontology engineering and KDD taking in consideration the best industry and research practices. A brief application of the life cycle is described at the end of the paper.

Keywords: Ontology Engineering, Knowledge Engineering, Knowledge Discovery, Data Mining, knowledge discovery process

Introduction
Although a great advance has been reached independently by both the ontology engineering field and the second generation of KDD tools (the idea of a continuum process for making sense of data) in the mid 90s (Fayyad U et al. 1996), somewhat less traditional has been the investigation of the role of ontologies in incremental and cyclic approaches of knowledge discovery.
This paper is concerned with the interaction between prior knowledge (by means of ontologies) and the process of knowledge discovery. It starts describing relevant topics to the integration of ontology engineering and knowledge discovery processes. The ontology driven knowledge discovery process is then presented along with its phases and its relation to the industry life cycle. Finally, the paper is summarized and research outcomes are presented.

Related Research
There are different relevant topics to the integration of ontologies and KDD processes in the literature such as “the role of domain knowledge in KDD”, “ontology/KDD integration” and “KDD life cycle”.
Domain knowledge has been playing an important role in the knowledge engineering research since the initial development of expert systems. Most recently domain knowledge has gain importance again in the integration of ontology engineering and KDD processes:

- Domingos (Domingos P. 1999) suggests use of domain knowledge as the most promising approach for constraining knowledge discovery and for avoiding the well-known problem of data overfitting by the discovered models.
- Anand et al. (Anand et al, 1995) also identify the use of domain knowledge in KDD tasks: for description of attribute relationship rules, for hierarchical generalization trees and constraints. An example of the latter is the specification of degrees of confidence in the different sources of evidence.
- Phillips J. & Buchaman B. (Phillips J., et al 2001) propose an ontology guided methodology to gradually accumulate knowledge of databases in order gain domain knowledge in the iterative process of a KDD task.

In spite of the increase investigation in the integration of domain knowledge, by means of ontologies, and KDD, most approaches concentrate only in the data mining phase of the

knowledge discovery process while the role of ontologies in other phases of the knowledge discovery process has been relegated. There are currently three approaches being investigated in the ontology and KDD integration emergent topic research: Onto4KDD, KDD4Onto, and one approach that integrates both previous approaches, named here Onto4KDD4Onto.
Onto4KDD is defined as the application of ontologies in order to improve the KDD process. For example, domain ontologies can be used to improve the understanding of a problem and to support hypothesis-driven analysis and discovery approaches. In contrast, KDD4Onto approaches are focused on the application of mining techniques in order to automatically or semi-automatically acquire knowledge from data.

Although some researchers are addressing Onto4KDD or KDD4Onto, rare is the research that encompasses both perspectives. This work is an attempt to integrate both approaches. It bridges the gap between ontological engineering and knowledge discovery in databases in order to improve both processes.

There are also several knowledge discovery life cycles in the literature. Most of them reflect the background of their proponents, such as those originated in the database community, in the artificial intelligence community, in the decision support community, and in the information systems community (Gómez-Pérez, A. 2004).

This research takes into consideration different aspects of these life cycles in order to develop a hybrid life cycle incorporating the best practices developed in the ontology engineering field as well as the best industry practice in the knowledge discovery scenario.

To this end it adopts the Cross Industry Standard Process for Data Mining (CRISP-DM) - and implements some of the requirements for the next generation of KDD tools (CRISP-DM 2007). CRISP-DM is a comprehensive methodology and process model that defines in different levels a complete mapping for Knowledge discover in database task. Although called process for data mining, CRISP-DM is the industry standard for knowledge discovery tasks. It breaks down the life cycle of a process into six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

Ontology Driven Knowledge Discovery - ODKD

The Ontology Driven Knowledge discovery – ODKD is a methodology and process model that defines, in different levels, a mapping between ontology engineering and Knowledge discover in database (KDD) process. It defines a hybrid life cycle for knowledge discovery tasks composed of five phases. Each phase is divided in tasks related, directly or indirectly, to both ontology engineering and to CRISP-DM. The life cycle is also supported by a methodology and guidelines which support the implementation of the tasks.

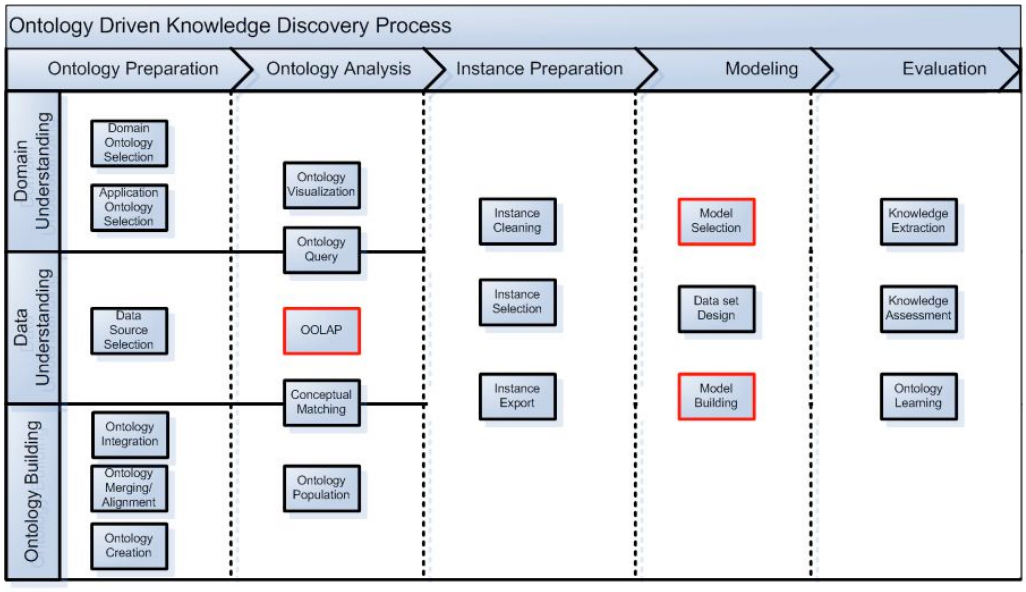

Table 1. Ontology Driven Knowledge Discovery phases and tasks.

As CRISP-DM, the ontology driven knowledge discovery process has a hierarchical breakdown into tasks – table 1: ontology preparation, ontology analysis, instance preparation, modelling, and evaluation. These tasks are responsible for the execution of specific ontology engineering and data modelling processes as well as for enabling a parallel integration of both knowledge and ontology engineering processes.

The first two phases, ontology preparation and ontology analysis are more related to the ontology engineering process of, assessment, selection and creation of an ontological model. The Instance preparation and Modelling phases are related to the data mining exercise. It is concerned with preparing data for a modelling task while selecting the most appropriated data mining technique to the problem. The last phase, Evaluation, is the most integrative phase where the results of a modelling exercise must be evaluated based on the previous knowledge and new knowledge is inserted in the ontological model.

Ontology Preparation

First and foremost, the prerequisite to a knowledge discovery exercise is data and business understanding. Without this understanding, no algorithm, regardless of sophistication, is going to provide a meaningful and useful result. Without this background a user/system will not be able to identify the problems he/she/it is trying to solve, prepare the data for mining, or correctly interpret the results.

The methodology initiates with a requirement gathering exercise which involves several stakeholders such as subject matter experts and business users, followed by the selection of problem related data and ontology sources. After reaching a consensus around the problem and the source requirements the actions are subdivided into three main pipelines, domain understanding, data understanding, and ontology building – see table 1.

The domain understanding pipeline is composed of two sub-tasks: domain ontology selection and application ontology selection. The domain ontology selection aims to select knowledge

that cover a broader problem perspective, for instance, general biomedical ontologies when dealing with a medical problem, while the application ontology selection aims in selecting specific ontological models of a problem, for instance, ontologies related to a specific disease when classifying/diagnosing patients.

The selection process is common for both domain and application ontologies. It involves several key steps, including assessing and selecting ontologies candidates (ontology investigation), knowledge quadrant analysis, and knowledge representation mapping plan.
The data understanding phase starts with an initial data collection based on the results of data sessions previously executed in the requirement gathering exercise. As in the ontology pipeline, the phase then proceeds with activities to get familiar with the data, to identify data quality problems and to map the data to the ontologies selected.

Ontology building is the last step of the ontology preparation phase. It consists of three main tasks: Ontology Integration, Ontology Merging/Alignment and Ontology Creation. Ontology integration is the incorporation of the ontologies identified in the business understanding tasks. The main objective is to reuse ontologies already developed as well as identified as available in the previous selection process in order to add a broader perspective to the problem domain.

Ontology merge/alignment is "a mapping of concepts and relations between two ontologies" (Sowa, J.F. 2001). The main goal is to incorporate and compare existent ontologies to form a body of concepts and relationships able to represent the domain and specific problem.

In spite of the current availability of large ontologies covering a very wide range of knowledge in several domains it is very likely that the ontologies acquired in the ontology merging/alignment won't be able to cover fully the necessary knowledge for a KDD task. The Ontology Creation is then concerned with the incorporation of problem specific knowledge and the creation of concepts not covered by the ontological model built in the previous task (merge/alignment).

Ontology Analysis

Ontology Analysis is related to the discovery of the first insights into the ontological model as well as to the investigation and/or checking of initial hypotheses created based on the disclosure of hidden information in the developed model.

The ODKD process defines four tasks in this phase: Ontology Visualization, Ontology Query, Conceptual Matching and Ontology Population. The first two tasks are related to the exploration of the ontological model by means of visualisation, search and analytical process. The last two tasks are related to the construction of the knowledge base by the insertion of data from the available databases and the population of instances from the selected ontologies.

Instance Preparation

The instance preparation phase covers all activities to construct the final dataset that will be fed into the modelling tool(s) from the ontological model. Instance preparation tasks are likely to be executed multiple times depending on the model being target. Tasks include concept, instances and slot (or attribute in a database taxonomy) selection, transformation and cleaning of instances as well as exportation to modelling tools. It is composed of three main

tasks in this phase: Instance Cleaning, Instance Selection and Instance Export.

Instance cleaning can be considered a special case of data cleaning, also called data cleansing or scrubbing. It deals with detecting and removing errors and inconsistencies from instances in order to improve the quality of features being selected for a modelling task.

The Instance selection task brings, probably, the most important contribution to an ontology driven knowledge discovery process. Instead of concentrating the effort in analysing the data characteristics, this task concentrates its effort in giving meaning to data and reducing the features by understanding the domain.

The instance export task is concerned with the translation of the ontological model into a format used by different data mining workbenches. The ontology driven knowledge discovery methodology then supports the reverse transformation of an ontological model into a dataset by exporting the knowledge base into a database format.

Modelling

In this phase, various modelling techniques are selected, different test sets are formed and then the models are applied. Typically, several techniques are applied for the same problem in order to validate the patterns found from different perspectives. This task is aligned with the data mining task of a KDD process.The ODKD methodology supports this phase by extracting knowledge from the ontological model and preparing a data set to the mining model selected.

Evaluation

This phase is concerned with the evaluation and acquisition of knowledge extracted by a data mining model. The knowledge acquired must be mapped and/or translated into the ontology and then evaluated before its final incorporation in the ontological model. This phase is divided in three tasks in the ODKD process: Knowledge Extraction, Knowledge Assessment and Ontology Learning.

Knowledge Extraction is responsible for the incorporation of the knowledge discovered by the data mining into the ontological model. It might also be considered as an “automatic knowledge acquisition”.

After the acquisition of the knowledge a new cycle of analysis begins to assess the knowledge incorporated. Depending on the requirements, this phase can be as simple as generating a report about the knowledge extracted or as complex as implementing a repeatable data mining process across which might involve the analysis of the current ontological model to validate the knowledge acquired or even acquire more domain knowledge to support the findings and/or go back to the ontology analysis to do extra analysis and select new test sets.

Creation of the model and assessment are generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the user can use it. In some sense, it might be compared to the deployment phase where the KDD team decides to agree with the knowledge assessed and uses its structure to update the ontological model. For example, the pattern founded might indicate that some molecular functions are responsible for triggering a disease.

This knowledge can then be used to support the decision of a medical team in the treatment of a disease.

Conclusion and Contribution

The Ontology Driven Knowledge Discovery Process is a conceptual proposal that alongside the Evolving Ontology model (Gottgtroy, P. 2006) are the basis for the development of a Framework able to enhance the process of knowledge discovery from data by adding a high level abstraction to the process which may allow a better reuse of previous knowledge as well as the creation and evaluation of new knowledge.

The main outcome of this paper is a hybrid knowledge discovery process – Table 1 - which defines a life cycle based on the principles of the most adopted KDD process in the industry. The process is divided into five phases which are composed of different tasks. The tasks have some methodological guidelines which indicate some of the best practices experimented in this research. These practices can be extended in accordance with the problem domain and KDD task. The process also presents guidelines based on both industry specific experiences and in the research developed. As a cyclic process it doesn't have a pre-defined and rigid order. To the contrary, the main objective is to create a reference life cycle which can be extended while keeping the most important tasks for the integration of ontology engineering and KDD.

The application of this life cycle, the framework and its tools are presented in (Kasabov, N. et al 2007). The process and methodology presented in this paper has been also used in different research and industry projects.

We believe that the integration of ontology engineering and KDD will play an important role in the semantic technologies adoption. This work then contributes to both research and industry applications by suggesting a hybrid methodology which describes best practices and leverages both ontology engineering and KDD processes.

References

Fayyad U et al. "From Data Mining to Knowledge Discovery: an Overview", in Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.

Domingos P., "The Role of Occam's Razor in Knowledge Discovery", Data Mining and Knowledge Discovery, an International Journal, Kluwer Academic Publishers, Vol.3, (1999), 409-425.

Anand S. S., Bell D. A., Hughes J. G. "The Role of Domain Knowledge in Data Mining", Proc. ACM CIKM '95, Baltimore MD USA, pp. 37-43.

Phillips, J. and Buchanan, B. G. "Ontology-guided knowledge discovery in databases". In Proceedings of the 1st international Conference on Knowledge Capture (Victoria, British Columbia, Canada, October 22 - 23, 2001). K-CAP '01. ACM Press, New York, NY, 123-130.

Gómez-Pérez, A., and Fernández-López, M., Corcho, O.: "Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web". Springer-Verlag, Berlin Heidelberg New York (2004)

CRISP-DM http://www.crisp-dm.org/

Sowa, J.F. (2001) Ontology, Metadata, and Semiotics. Retrievable from the internet 28/01/2006 at: http://users.bestweb.net/~sowa/peirce/ontometa.htm.

Gottgtroy, P., Kasabov, N. & MacDonell, S. "Evolving Ontologies for Intelligent Decision

Support" Publisher: Elsevier in the series "Capturing Intelligence" - Fuzzy Logic and the Semantic Web (2006).

Kasabov, N, Jain. V., Gottgtroy, P., Benevuska, L. & Joseph, F. (2007). Evolving Brain-Gene Ontology and Simulation System (BGOS): Towards Integrating Bioinformatics and Neuroinformatics Data, Information and Knowledge to Facilitate Discoveries. Special Issue of Neural Networks. In Press.