Semantic Data Preparation:
The Instance-Selection plug-in

Paulo Gottgtroy
Knowledge Engineering and Discovery Research Institute
School of Computer and Information Sciences
Auckland University of Technology
Private Bag 92006 Auckland 1020, New Zealand
E-mail: paulo.gottgtroy@aut.ac.nz

Abstract— Computer scientists have been working on data preparation techniques in order to improve the quality of the KDD results. In largely parallel research, ontologies have been widely used by the Artificial Intelligence community to represent domain knowledge and to integrate different database models. This work investigates the application of ontologies in the data preparation step of the KDD process. We present an ontology instance selection tool able to export an ontological representation into specific formats used by different data mining workbenches.

Introduction

Knowledge Discovery in Databases (KDD) is an iterative process based on the analysis of current facts or data, pre-processing to clean and transform that data, application of mining algorithms, and deployment using the mining results on new data. Computer scientists have been working on cleaning data, data transformations, selection of samples and - in case of large data sets - performing feature selection operations to find relevant variables to the problem in order to improve the quality of data processes. Their objectives include: building simpler and more comprehensible models, improving data mining performance, and helping to prepare, clean, and understand data. In largely parallel research, ontologies have been widely used by the Artificial Intelligence community to represent domain knowledge and to integrate different database models.

A common background underlying database and ontology research is well known. Although ontologists and data modelers have been working together to bridge both areas, for example, in topics like conceptual modeling, database integration and metadata representation, less work has been undertaken in relation to feature selection in large and multi-dimension spaces. Our approach explores this gap and presents an ontology feature selection tool able to translate its representation into a tabular format. This tabular format can further be exported to different data mining workbenches or data formats.

The Instance Selection plug-in is part of a set of tools which supports the Ontology Driven Knowledge Discovery (ODKD) [1]. ODKD investigates a hybrid approach bringing together the state of art of artificial intelligence methods for knowledge discovery in large databases (KDD) and ontology engineering. ODKD analyses how data mining can assist in efficient and effective large-volume data analysis in order to build a sharable and evolving knowledge repository, while at the same time leveraging the semantic content of ontologies to give intelligent support and to improve knowledge discovery in complex and dynamic domains.

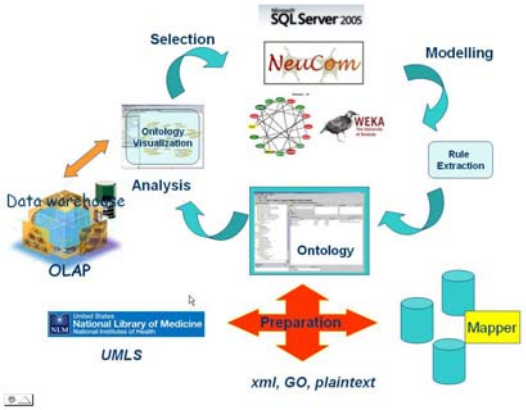

Figure 1.OKDD process.

The semantic data preparation phase of ODKD encompasses: ontology preparation, ontology analysis and instance selection - figure 1. It covers the first three steps of CRoss-Industry Standard Process for Data Mining process (CRISP-DM) [2] - business understanding, data understanding and data preparation - creating an alternative ontology driven pipeline for the data mining process.

The Instance Selection Tool supports the third step of CRISP-DM (data preparation) by converting ontology concepts into a simple tabular representation which allow further concept analysis by data mining workbenches, business intelligence tools, data visualization techniques and so on.

Instance Selection Tab

The Instance Selection Tab (figure 2) extends the instance tree tab [4] by adding a panel with a table able to store instances selected from the knowledge base through the instance tree, from results of a query and from the selection of instances in a visualization tool. We have also developed additional features to the query tab [5] and instance browser panel which allow us, for example, to select instances from the TGViz Tab [6].

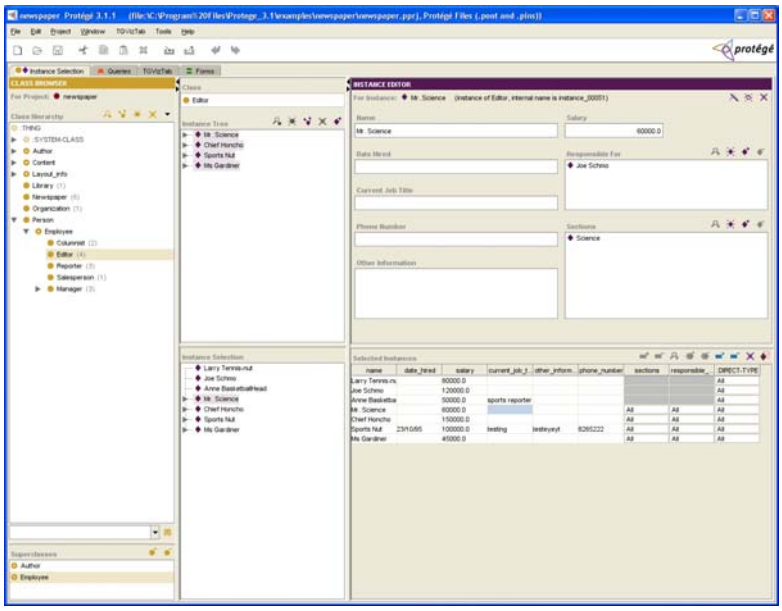

Figure 2 – Instance tree tab.

After the selection of instances, the Tab allows a full manipulation of instances and its slots defining among others: which slots should be exported, select the values of multiple value slots, select a template including the internal name (:name) and/or display slot, etc – figure 3.

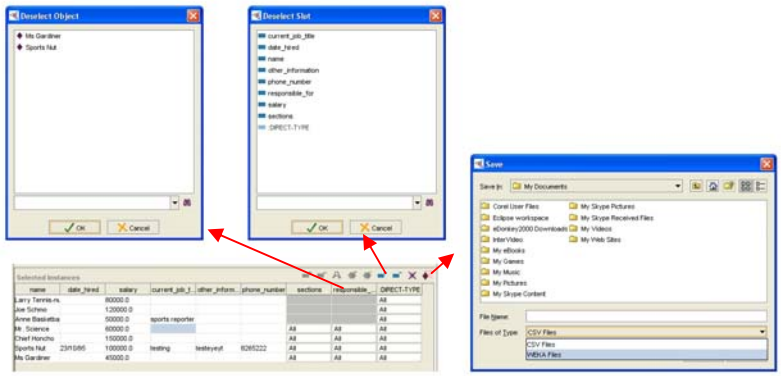

Figure 3 – Examples of instance manipulation options.

It also allows a recursive manipulation of slot type instance defining which slots of the instances should be exported for further analysis. This feature enables a full investigation of the relationships among concepts, for example, in a bioinformatics problem a user can select the genes responsible for a disease and further cluster them by the molecular function, gene expression in order to test new/different hypothesis without any extra data transformations

We believe that the Instance Selection Tab can help in any scenario where there is a need for extract/manipulate/export instances from a knowledge base without programmatically access the Protégé API.