KNOWLEDGE EXTRACTION GUIDED BY ONTOLOGIES
Database Marketing Application

Filipe Mota Pinto
Politechnic Institute of Leiria, Leiria, Portugal

Teresa Guarda
Superior Institute of Languages and Administration of Leiria, Leiria, Portugal

Keywords: Ontologies, Knowledge Discovery, Databases, Data Mining.

Abstract. The knowledge extraction in large databases has being known as a long term and interactive project. In spite of such complexity and different options for the knowledge achievement, there is a research opportunity that could be explored, throughout the ontologies support. Then this support may be used for knowledge sharing and reuse. This paper describes a research of an ontological approach for leveraging semantic content of ontologies to improve knowledge extraction in a oil company marketing databases. We attain to analyze how ontologies and knowledge discovery process may interoperate and present our efforts to propose a possible framework for a formal integration.

1 INTRODUCTION

At artificial intelligence research area, ontology is defined as a specification of a conceptualization (Gruber, 1993). Ontology specifies at a higher level, the classes of concepts that are relevant to the domain and the relations that exist between these classes. Indeed, ontology captures the intrinsic conceptual structure of a domain. For any given domain, its ontology forms the heart of the knowledge representation.

In spite of ontology-engineering tools development and maturity, ontology integration in knowledge discovery projects remains almost unrelated.

Knowledge Discovery in Databases (KDD) process is comprised of different phases, such as data selection, preparation, transformation or modeling. Each one of these phases in the life cycle might benefit from an ontology-driven approach which leverages the semantic power of ontologies in order to fully improve the entire process (Gottgtroy et al, 2004).

We dare to combine ontological engineering and KDD process in order to improve it. One of the promising interests in use of ontologies in KDD assistance is their use for guiding the process. This research objective seems to be much more realistic now that semantic web advances have given rise to common standards and technologies for expressing and sharing ontologies (Bernstein et al 2005).

This paper describes an ontological approach research for leveraging the semantic content of ontologies to effectively support the knowledge discovery in databases. We analyze how ontologies and knowledge discovery process may interoperate and present our efforts to bridge the two fields, knowledge discovery in databases and ontology learning for successful database usage projects.

2 BACKGROUND

There are different relevant topics to the KDD processes assistance also referred in literature such as “domain knowledge in KDD” (Pinto and Santos, 2009), “ontology/KDD integration” (Bernstein et al 2005), “KDD life cycle” (Zhou et al., 2009) and “KDD assisted process” (Bernstein et al 2005).

Ontologies are used to capture knowledge about some domain of interest. Ontology describes the concepts in the domain and also the relationships that hold between those concepts. Different

Mota Pinto F. and Guarda T..
KNOWLEDGE EXTRACTION GUIDED BY ONTOLOGIES - Database Marketing Application.
DOI: 10.5220/0003555103200325
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 320-325
ISBN: 978-989-8425-53-9

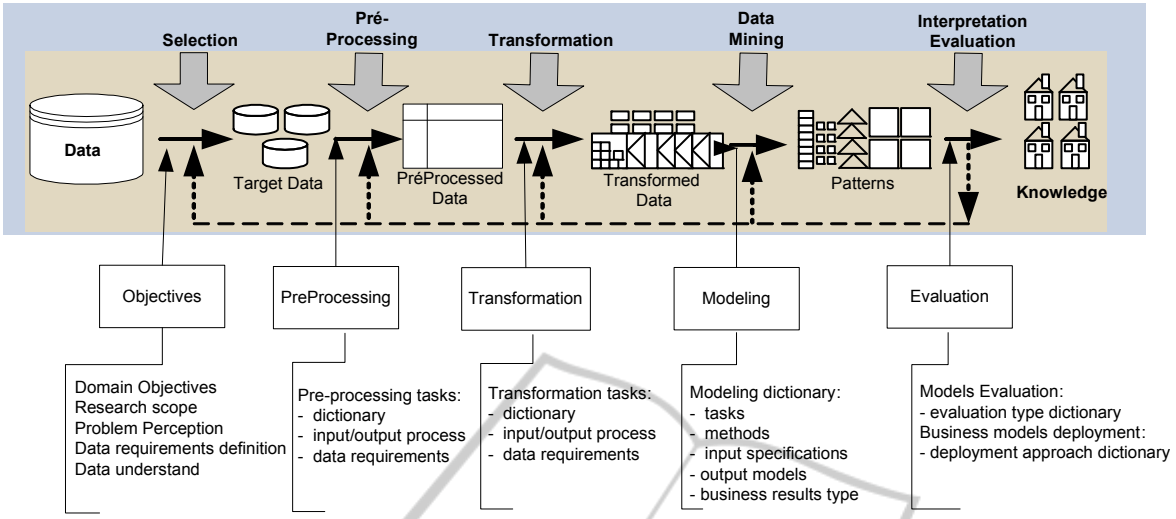

Figure 1: KDD general phase and task description workflow.

ontology languages provide different facilities. Ontology Web Language (OWL) is a standard ontology language from the World Wide Web Consortium (W3C).

2.1 Semantic Web Language Rule

To the best of our knowledge there are no standard OWL-based query languages. Several RDF -based query languages exist but they do not capture the full semantic richness of OWL. To tackle this problem, it was developed a set of built-in libraries for Semantic Web Rule Language (SWRL) that allow it to be used as a query language

The OWL is a very useful means for capturing the basic classes and properties relevant to a domain. However, these domain ontologies establish a language of discourse for eliciting more complex domain knowledge from subject specialists. Due to the nature of OWL, these more complex knowledge structures are either not easily represented in OWL or, in many cases, are not representable in OWL at all. The classic example of such a case is the relationship uncleOf(X,Y). This relation, and many others like it, requires the ability to constrain the value of a property (brotherOf) of one term (X) to be the value of a property (childOf) of the other term (Y); in other words, the siblingOf property applied to X (i.e., brotherOf(X,Z)) must produce a result Z that is also a value of the childOf property when applied to Y (i.e., childOf(Y,Z)). This "joining" of relations is outside of the representation power of OWL.

One way to represent knowledge requiring joins of this sort is through the use of the implication (→) and conjunction (AND) operators found in rule-based languages (e.g., SWRL). The rule for the uncleOf relationship appears as follows:

```
brotherOf(X,Z) AND
childOf(Y,Z)→uncleOf(X,Y)
```

2.2 Evaluation of Knowledge Reuse Effectiveness

The main objective of this research is to assist the KDD process based on ontology knowledge. Therefore, it is assumed that effectively ontology has learned from KDD domain and practice. Thus it is possible to provide users with information that are relevant to their needs at each of KDD phases.

Hence, the related task (process option) suggestion returned by the ontology will be the primary basis to determine the quality of the relevant information retrieved. For the present purpose we admit as major indices, precision and recall (Han and Kamber, 2001):

$$Precision=(Relevant \cap Selected)/(Selected\_Results)$$

Precision expresses the proportion of related results (Relevant∩Selected) among relevant results retrieved (Selected_Results). In other words, to reflect the amount of knowledge correctly identified (in the ontology) with respect to the whole knowledge available in the ontology As related results we intend the entire set of ontology elements (classes and data properties) related to the subject (e.g., to preprocessing phase: set of related classes and relationships). Also, we use selected results as the set of related results and selected at the user

question (e.g., set of suggested results for, e.g., birthDate attribute preprocessing).

$$Recall=(Relevant \cap Selected)/(Relevant\_Results)$$

Recall expresses the proportion of results retrieved (Relevant∩Selected) from related results (Relevant_Results). It is used to reflect the amount of knowledge correctly identified with respect to all the knowledge that it should identified.

During this work, precision will be used to evaluate the proportion of user interests towards the KDD phase assistance. This proportion examines how correct the ontology is suggesting tasks (options) when solicited by the user. On other hand, recall, estimates the ability that the system is able to satisfy user needs.

2.3 Database Marketing

Nowadays, organizations try to act dynamically in competitive markets so that they can find and keep their customers. In this context Database Marketing presents itself as a privileged tool marketing professionals may use to do so.

DBM refers the use of database technology for supporting marketing activities (Shepard 1998). Therefore DBM is a marketing process driven by information and managed by database technology (Grassl 2007), allowing marketing professionals to develop and to implement better marketing programs and strategies. However, DBM is usually approached using classical statistical inference, which may fail when complex, multi-dimensional, and incomplete data is available.

Through the advances in information and communication technologies, corporations can effectively obtain and store transactional and demographic data on individual customers at reasonable costs. The DBM activity framework has changed significantly over the last years. In past, database marketers applied business rules to target customers directly, based sometimes in the marketer's intuition. The current approach relies on predictive response models to target customers for offers or other marketing purposes. The challenge now is how to extract important knowledge from those vast databases. Through the use of DBM organizations get to understand the customers' preferences and behaviours through analyzing their transactional data. However, in almost cases, the data set presents several problems as the result of procedural factors, inadequate questionnaire options, refusal of response, or/and inadequate database schema's.

The set of DBM processes used in this study include the use of a KDD framework in order to create predictive models. These models may be used to, e.g., accurately estimate the probability that a customer will respond to a specific offer and can significantly increase the response rate to a product offering. The old model of design-build-sell is now being replaced by sell-build-redesign.

3 RESEARCH APPROACH

Here, we attain to the effective support the KDD process using ontologies. In order to do this we have collected a marketing database from a multinational oil company. Therefore, we used a real case study in order to effectively test and deploy to ontological work in the KDD process (Pinto et al, 2009). Our approach is defined in two distinct steps:

- Marketing database collection and preparation;
- Practical KDD ontology based development;

3.1 Marketing Database

One of the most important marketing tools used by oil companies for customer fidelization is the marketing card loyalty programs (Phillips. and Buchanan, 2001). This approach allows cardholders to obtain fuel purchase discounts, to participate in marketing campaigns or to become members of a restrict club with restrict privileges.

Since it is an open marketing system program where all oil customers may access, from the company perspective this will turn into an important information source for almost customer oriented marketing strategies definition or product offer policies.

We have collected a card loyalty program marketing database from a multinational company. This database has three main tables: card owner, station and transactions. The available data refers to the past two year's activity. The data structure is as follows:

Table 1: Data Table Card Owner.

Field	Description	Type	Domain range
IdCard	card identification	Primary key	
Idclient	Client identification	Foreigner key	
birthDate	Client birth date	Date	< today
cardClientDate	Starting card owner date	Date	<today
cardInitialDate	Starting client date	Date	<today
Postcode	Zip code	Integer	<10000
postCod3	3 Zip code	Integer	<1000
maritalStatus	Marital status	String	{cas; sol; div; viv; out}
Gender	Client sex	String	{mas,fem}
vehicleType	Vehicle type description	String	{lig, merc, pes, out}
vehicleYear	Vehicle identification year date	Number	<10000
fuelType	Fuel description type	String	{diesel, gasolina, gpl, out}

Table 2: Data Table Card Transactions.

Field	Description	Type	Domain range
IdMov	Transaction identification	Primary key	
IdCard	Card identification	Foreigner key	
Date	Transaction date	Date	< today
fuelValue	Fuel transaction amount	real	<10000
fuelLitres	Transaction liters amount	real	<3000
shopValue	Transaction value amount	real	<10000
shopUnits	Shop units transaction amount	integer	<10
stationCode	Fuel station identification	Foreigner key	

Table 3: Data Table Station.

Field	Description	Type	Domain range
stationCode	Fuel station identification code	Primary key	
stationType	Station identification type	String	{urb, rur, est}
postCode	Zip code	integer	<10000
postCod3	3 Zip code	integer	<1000

4 PRACTICAL KDD ONTOLOGY BASED DEVELOPMENT

For the KDD development we have based our work on the free open source WEKA toolkit (Witten and Frank2000). For ontological support we have used the PROTÉGÉ OWL editor and SWRL language (Knublauch et al. 2005).

Since KDD process generates output models, it was considered useful to represent them in a computable way. Such representation works as a general description of all options taken during the process. Based on PMML descriptive DM model we have introduced an OWL class in our ontology named *ResultModel* which holds instances with general form:

```
ResultModel {
domain Objective Type;
algorithm;algorithmTasks;
algorithmParameters;
workingAlgorithmDataSet;
EvaluationValue;
DeploymentValue}
```

Since the ontology contribution to the KDD process is quantitatively uncertain we have used a quality approach based on KDD team individual expert contribution.

4.1 Ontological Work

One of the promising interest of ontologies is they common understand for sharing and reuse. Hence we have explored this characteristic to effectively assist the KDD process. Indeed, this research presented the KDD assistance at two levels: Overall

process assistance based on ModelResult class and,
- KDD phase assistance. Since our ontology has a formal structure related to KDD process, is able to infer some result at each phase.

To this end, user need to invoke the system rule engine (reasoner) indicating some relevant information, e.g., at data preprocessing task: swrl:query hasDataPreprocessingTask(?dpp,"ds"), where hasDataPreProcessingTask is an OWL property which infers from ontology all assigned data type preprocessing tasks (dpp) related to each attribute type within the data set "ds". Moreover, user is also assisted in terms of ontology capability index, through the ontology index - precision, recall and PRI metrics.

Once we have a set of running KDD process registered at the knowledge base, whenever a new KDD process starts one the ontology may support the user at different KDD phases. As example to a new classification process execution the user interaction with ontology will follow the framework as described in next section. The ontology will lead user efforts towards the knowledge extraction suggesting by context. That is the ontology will act accordingly to user question, e.g., at domain objective definition (presented by user) the ontology will infer which is type of objectives does the ontology has. All inference work is dependent of previous loaded knowledge. Hence, there is an ontology limitation – only may assist in KDD process which has some similar characteristics to others already registered.

5 EXPERIMENT

To build up mining experiments we have used Weka Toolkit (Witten and Frank, 2000) which allowed not only the actual mining but also featured analysis and algorithm evaluation. These experiments did not aim to the full construction of a classification model but instead to test and analyze different approaches and further ranking.

Our system prototype operation follows general KDD framework (Figure 1) and uses the ontology to assist at each user interaction. Our experimentation was developed over a real oil company fidelity card marketing database. This database has three main tables: card owner; card transactions and fuel station.

To carry out this we have developed an initial set of SWRL rules. Since KDD is an interactive process, these rules deal at both levels: user and ontological levels. The logic captured by these rules is this section using an abstract SWRL representation, in which variables are prefaced with question marks.

```
Domain objective: customer profile
Modeling objective: description
Initial database: fuel fidelity card;
Database structure: 4 tables;
```

The most relevant rule extracted from above data algorithms use was:

if (age<27 and vehicleType="Lig" and sex="Female") then 1stUsed="p"

In this model we may say that, female card owners under 27 years of age have a "*lig*" (ligeiro) category car and use a fuel station located in range of 10 kilometers from their address.

Also, practical KDD process tasks have been done supported by SWRL ontology queries. This query tasks was manually performed by the user. Therefore, the guidance was accomplished and achieved throughout knowledge base updating with the general model:

```
INSERT record KNOWLEDGE BASE
  hasAlgorithm(J48) AND
     hasModelingObjectiveType(classif
     ication) AND
     hasAlgorithmWorkingData
     ({idCard;  age;  carClientGap;
     civilStatus;  sex;  vehicleType;
     vehicleAge;nTransactions;
     tLiters;  tAmountFuel;tQtdShop;
     1stUsed; 2stUsed; 3stUsed }) AND
      Evaluation(67,41%;  95,5%)  AND
      hasResultMoldel (J48;
     classification;"wds",PCC;0,674;0
     955)
```

The evaluation, once performed, the system automatically updates the knowledge base with a new record. The registered information will serve for future use – knowledge sharing and reuse.

From the aforementioned previous research work (Pinto and Santos, 2009), we also have used the output models and integrated them into the knowledge base.

6 CONCLUSIONS AND FURTHER RESEARCH

The KDD success still is very much user dependent. Though our system may suggest a valid set of tasks which better fits in KDD process design, it still miss the capability of automatically runs the data, develop modeling approaches and apply algorithms.

This work strived to improve KDD process supported by ontologies. To this end, we have used

general domain ontology to assist the knowledge extraction from databases with KDD process.

This research focuses the KDD development assisted by ontologies. Moreover we use ontologies to simplify and structure the development of knowledge discovery applications offering to a domain expert a reference model for the different kind of DM tasks, methodologies to solve a given problem, and helping to find the appropriate solution.

Future research work will be devoted to expand the use of KDD ontology through knowledge base population with more relevant concepts about the process. Another interesting direction to investigate is to represent the whole knowledge base in order to allow its automatic reuse.

REFERENCES

Bernstein, A., Provost, F., and Hill, S. (2005). Toward intelligent assistance for a data mining process. IEEE Transactions knowledge & data engineering, 17(4).

Domingos, P. (2003). Prospects and challenges for multi-relational data mining. *SIGKDD Explorer Newsletter, 5(1):80–83.*

Pinto, Filipe. Mota, Gago, P., and Santos, M. F. (2009). Marketing database knowledge extraction *In IEEE 13th International Conference on Intelligent Engineering Systems 2009.*

Pinto, Mota Filipe. and Santos, M. F. (2009). Database marketing supported by ontologies . *In 11th ICEIS.*

Gottgtroy, P., Kasabov, N., and MacDonell, S. (2004). *An ontology driven approach for knowledge discovery in biomedicine.*

Grassl Wolfgang (2007) "The reality of brands: Towards an ontology of marketing. *American Journal of Economics and Sociology, 58(2):313–319, April.*

Gruber, T. R. (1993). A translation approach ontology specifications. *Knowledge Acquisition, 5:199–220.*

Knublauch Holger, Horridge M., Noy N, and Wang H (2005) "The Protégé OWL Experience", *Workshop on OWL: Experiences & Directions, Galway, Ireland*

Han, J. and Kamber, M. (2001). *Data mining: concepts & Techniques. Morgan Kaufman, San Francisco, CA.*

Horrocks, I., Patel-Schneider, S., Grosof, B., and Dean, M. (2004). Swrl: A semantic web rule language - combining owl and ruleml. *Technical report, W3C.*

Phillips, J. and Buchanan, B. G. (2001). Ontology-guided knowledge discovery in databases. In ACM, editor, International Conference *On Knowledge Capture , p 123–130. I C. On Knowledge Capture.*

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. *In Proc 2nd ACM Conf e-Commerce.*

Shepard David. (1998)"Database Marketing". São Paulo: *Makron Books.*

Witten, I. H. and Frank, E. (2000). Data Mining: Practical Machine Learning Tools and Technique. *The Morgan Kaufmann Series in Data Management Systems, 2nd edition.*

Zhou Xuan, Geller James, and Perl Yehoshua and Halper Michael.(2009) "An Application Intersection Marketing Ontology", *Theoretical Computer Science, pp 143–163.LNCS. Springer Berlin.*