ISBN: 978-972-8924-63-8 © 2008 IADIS

TOWARDS A MODEL-DRIVEN ENGINEERING APPROACH OF DATA MINING*

Jesús Pardillo
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante. Spain

Jose Norberto Mazón
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante. Spain

Jose Zubcoff
Departamento de Ciencias del Mar y Biología Aplicada
Universidad de Alicante. Spain

Juan Trujillo
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante. Spain

ABSTRACT

Nowadays, data mining is based on low-level speci cations of the employed techniques typically bounded to a speci c analysis platform. Therefore, data mining lacks a modelling architecture that allows analysts to consider it as a truly software-engineering process. Here, we propose a model-driven approach based on (i) a conceptual modelling framework for data mining, and (ii) a set of model transformations to automatically generate both the data under analysis (via data-warehousing technology) and the analysis models for data mining (tailored to a speci c platform). Thus, analysts can concentrate on the analysis problem via conceptual data-mining models instead of low-level programming tasks related to the underlying-platform technical details. These tasks are now entrusted to the model-transformations scaffolding.

KEYWORDS

data mining, data warehouse, model-driven engineering, multidimensional modelling, conceptual modelling

1. INTRODUCTION

Data mining is a highly complex task which requires a great effort in preprocessing data under analysis, *e.g.*, data exploration, cleansing, and integration (Pyle 1999). Therefore, some authors suggest the suitability of data-warehousing technologies (Kimball 2002) for improving the conventional knowledge discovery in databases process (Frawley *et al.* 1991) by means of providing an integrated and cleansed collection of data over which data-mining techniques can be straight applied (Inmon 1996; Rizzi *et al.* 2006). However, the current data-mining literature has been focused on the presenting new techniques and improving the underlying algorithms (Hand *et al.* 2001), whilst the most known software platforms do not apply the data-warehousing principles during the data-mining design. To overcome this situation, several mechanisms have been proposed (Zubcoff & Trujillo 2007; Zubcoff *et al.* 2007; Zubcoff & Trujillo 2006) to model data-mining techniques in conjunction with data-warehousing technology from the early stages of design (*i.e.*, at the conceptual level). These data-mining models do not only support analysts in using and understanding the data-mining techniques in real-life scenarios, but also they allow designers to document the data-mining techniques in detail. Hence, these data-mining models are truly blueprints that can be used to manually obtain the required data-mining metadata as a basis of the implementation in a certain data-mining platform. However, this complex task is only accessible to expert analysts and requires too much effort to be successfully completed (Inmon 1996; González-Aranda *et al.* 2004).

In this work, we will go beyond the de nition of new models, here we de ne a model-driven engineering (Bézivin 2006) approach for data mining. Moreover, we propose the use of a well-known visual modeling standard, the "unified modelling language" (UML) (OMG 2008) for facilitating the design and

* This work has been supported by the ESPIA (TIN2007-67078) project (Spanish Ministry of Education), and by the QUASIMODO (PAC08-0157-0668) project (Castilla-La Mancha Ministry of Education). Jesús Pardillo and Jose-Norberto Mazón are funded by the Spanish Ministry of Education (FPU grants AP2006-00332 and AP2005-1360).

implementation tasks. In addition, our approach automatically generate a vendor-speci c data-mining implementation from a conceptual data-mining model, taking into consideration the deployment of the underneath data warehouse (*i.e.*, data under analysis).

The rest of the paper is structured as follows: Section 2 outlines the related work. Section 3 describes our model-engineering approach for data mining. Finally, Section 4 exposes conclusions, and future work.

2. RELATED WORK

Current approaches for data-mining can be classi ed on those that are a general description of data-mining process, or mathematical oriented, and propose solutions at low-abstraction level. Both approaches overlook the de nition of understandable artifacts that could be easily used by designers in a software engineering process. "Cross industry standard process for data mining" (CRISP-DM Consortium 2008) is a detailed description of the data-mining phases. Nevertheless, this standard neither proposes a concrete modeling tool nor presents a conceptual model for data mining.

The "common warehouse metamodel" (CWM) (OMG 2008) and the "predictive model markup language" (PMML) (DMG 2008) are standards for the metadata interchange proposed by vendor-independent consortiums (OMG and DMG, respectively), but they cannot be used as analysis artifacts. The "data mining extensions" (DMX) (Microsoft 2008) is a SQL-like language for coding data-mining models in the Microsoft platform, and therefore it is difficult to gain understanding of the data-mining domain. The XELOPES (Prudsys 2008) and Weka (University of Waikato 2008) provide an entire framework to carry out data mining but, once again, they are situated at very low-abstraction level, since they are code-oriented and they do not contribute to facilitate understanding of the domain problem.

All of these approaches have the same drawback, since they are focused on solving the technical scaffolding instead of providing analysts with intuitive artifacts to specify data mining. To the best of our knowledge, only the proposal described in (Zubcoff & Trujillo 2007; Zubcoff *et al.* 2007; Zubcoff & Trujillo 2006) provides a modelling framework to de ne data-mining techniques at a high-abstraction level by using UML. However, these UML-based models are mainly used as documentation.

3. MODEL-DRIVEN ARCHITECTURE FOR DATA MINING

Our model-driven engineering approach for data mining advocates de ning the underneath data warehouse (*i.e.*, data under analysis) together with the corresponding data-mining technique. The data warehouse is based on a multidimensional model which de nes the required data structures. In previous work, we have aligned this process with a model-driven approach (Pardillo & Trujillo 2008; Mazón & Trujillo 2008) in order to support designers to develop a conceptual multidimensional model and the automatic derivation of its corresponding implementation.

The novelty of our approach is twofold: (i) it is based on de ning vendor-neutral models of data-mining techniques together with the model of the underlying data warehouse, and (ii) the deployment of those data-mining techniques is done automatically. Thus, on one hand, we use a modelling approach (Zubcoff & Trujillo 2007; Zubcoff *et al.* 2007; Zubcoff & Trujillo 2006) for de ning platform-independent models for several data-mining techniques. Therefore, the model-transformation con guration has to be described in order to consider every kind of target platform. In Fig. 1, we provide an overview of the required model-transformation architecture, stressing some of the current data-mining standards and platforms in the market.

The conceptual data-mining modelling framework in data warehouses is shown at the top of this model-driven architecture. Fig. 1 also shows the transformation paths to derive several implementations through mapping data-mining models to an executable environment: CWM, XELOPES, DMX, or Weka acting as bridge. Depending on the characteristics of the analysis itself (*e.g.*, the required technique) or the data-mining solution available (*e.g.*, it can only be open-source platforms), we choose one of the transformation paths.

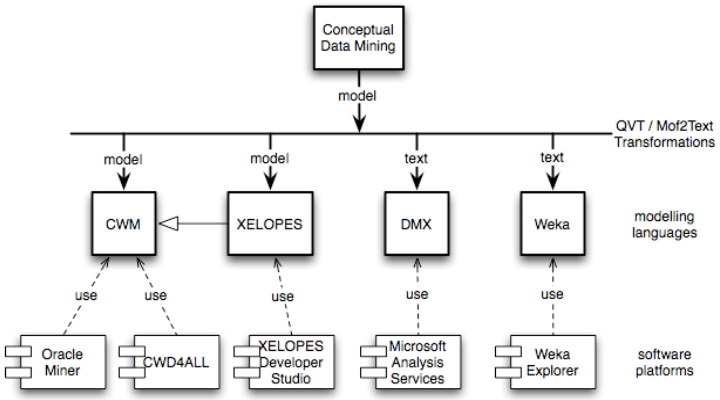

Figure 1. Model-transformation architecture for model-driven data mining

From a technical point of view, we propose the usage of the "model-driven architecture" (MDA) (OMG 2008) in order to implement these transformations between data-mining models. Within an MDA-based approach the "query/view/transformation" (QVT) language can be used as a standard mechanism for de ning formal relations between MOF-compliant models that allows the automatic derivation of a implementation. Nevertheless, there are transformations that are applied from models (*i.e.*, MOF-based) to implement code (*i.e.*, textual modelling languages). In these cases, MDA offers the "MOF models to text transformation" (Mof2Text) language that allows us to specify transformations by means of textual templates in order to automatically derive the corresponding implementation.

Due to the space constraints, we exclude an abstract speci cation of the involved mapping, an example transformation of the data under analysis that can be found in (Mazón & Trujillo 2008). Nevertheless, in Fig. 2, it is shown the implementation of this mapping (left-hand side) over the Eclipse development platform. In order to accomplish this task, we have used the MOFScript plug-in for this platform. MOFScript is a transformation-language implementation of the Mof2Text standard language that enable us to specify model-to-text transformations in the "model-driven architecture" (MDA) (OMG 2008) proposal. On the right-hand side, the resulting DMX code is shown.

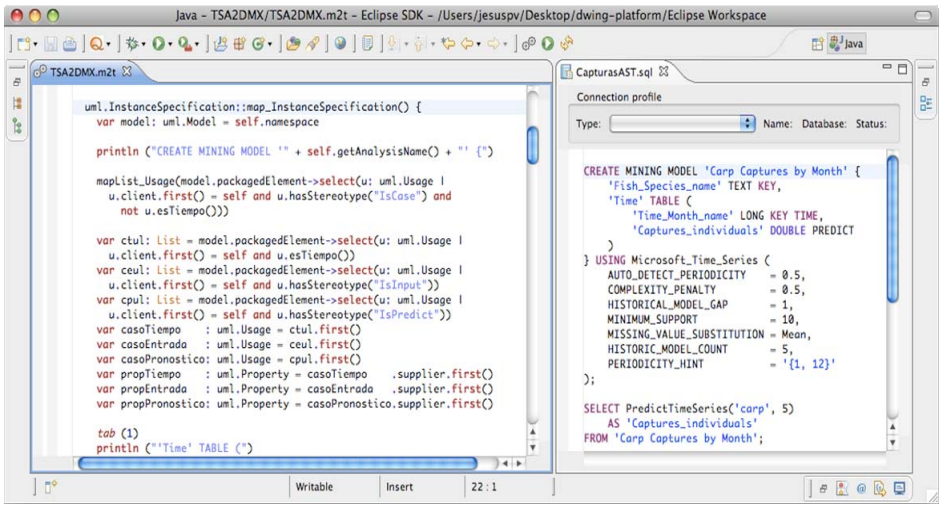

Figure 2. Example of a Mof2Text transformation and the generated code

Given Fig. 2, the mapping overview is as follows: each data mining conceptual model is mapped into a low-level model in DMX (MINING MODEL instruction). Every parameter of the analysis technique is also mapped into their DMX counterpart. The data under analysis (from the multidimensional model) is mapped into a data-mining attribute in DMX. The MOFScript engine is used to translate the conceptual model to the DMX code. Finally, within this solution, analysts can consult the data-mining results by visualising the obtained patterns and trends and extracting new knowledge from them.

4. CONCLUSION

Due to mathematical foundations of data-mining techniques, there are no formalised mechanisms to easily specify data-mining activities as a real software engineering process. In this paper, we propose a model-engineering approach for overcoming this limitation. On one hand, we provide a set of models to specify data-mining techniques in an vendor-neutral way that are close to the way of analysts thinking (*i.e.*, conceptual models). On the other hand, we provide transformations to automatically derive platform-speci c models from the conceptual ones, altogether with the deployment of data under analysis. Thus, analysts can focus on data mining at an abstract level instead of distracting by low-level details.

Therefore, the great bene t of our approach is that, once we have established the model-driven architecture for both data under analysis and analysis techniques for data mining, analysts can model their data-mining related tasks easily in a vendor-neutral way whereas the model-transformations scaffolding is entrusted to automatically implement them in a certain platform. Our ongoing work covers other high-level mechanisms to specify data-mining related tasks, and the integration of the proposed data-mining framework together with the analysis technologies traditionally employed in the data-warehouse domain.

REFERENCES

Bézivin, J., 2006. Model Driven Engineering: An Emerging Technical Space. *GTTSE*, pp. 36-64.

CRISP-DM Consortium, June 2008. CRISP-DM, version 1.0. http://www.crisp-dm.org.

DMG, June 2008. Predictive Model Markup Language (PMML). http://www.dmg.org/pmml-v3-2.html.

Hand, D.J., Mannila, H., Smyth, P., 2001. *Principles of Data Mining*. MIT Press.

Inmon, W.H., 1996. The Data Warehouse and Data Mining. *Commun. ACM*, Vol. 49, No. 4, pp. 83–88.

Kimball, R., Ross, M., 2002. *The Data Warehouse Toolkit*. Wiley.

Frawley, W. J., Piatetsky-Shapiro, G., Matheus, C. J., 1991. *KDD: An Overview*. AAAI/MIT Press.

González-Aranda, P., Menasalvas, E., Millán, S. Segovia, J., 2004. Towards a Methodology for Data Mining Project Development: The Importance of abstraction. *ICDM/FDM*, pp. 39-46.

Luján-Mora, S., Trujillo, J., Song, I.Y., 2006. A UML pro le for multidimensional modeling in data warehouses. *Data Knowl. Eng*. Vol. 59, No. 3, pp. 725.769.

Mazón, J.N., Trujillo, J., 2008. An MDA approach for the development of data warehouses. *DSS*. V. 5, No. 1, pp. 41-58.

Microsoft, June 2008. Data Mining eXtensions (DMX). http://msdn2.microsoft.com/en-us/library/ms132058(VS.90).aspx

OMG June 2008. Common Warehouse Metamodel (CWM), Uni ed Modeling Lang. (UML), Model DrivenArchitecture (MDA), Query/View/Transf. Lang. (QVT), MOF Transf. Lang. (Mof2Text). http://www.omg.org.

Pardillo, J., Trujillo, J., 2008. Integrated Model-driven Development of Goal-oriented Data Warehouses and Data Marts.*ER*. In Press.

Prudsys, June 2008. Extended Library for Prudsys Embedded Solutions (XELOPES). www.prudsys.com

Pyle, D., 1999. *Data Preparation for Data Mining*. Morgan Kaufmann.

Rizzi, S., Abelló, A., Lechtenbörger, J., Trujillo, J., 2006. Research in data warehouse modeling and design: dead or alive? *DOLAP*, pp. 3-10.

University of Waikato, June 2008. Weka. http://www.cs.waikato.ac.nz/ml/weka.

Zubcoff, J., Pardillo, J., Trujillo, J., 2007. Integrating Clustering Data Mining into the Multidimensional Modeling of Data Warehouses with UML Pro les. *DaWaK*. pp. 199-208.

Zubcoff, J., Trujillo, J., 2007. A UML 2.0 pro le to design Association Rule mining models in the multidimensional conceptual modeling of data warehouses. *Data Knowl. Eng*. Vol. 63, No. 1, pp. 44-62.

Zubcoff, J., Trujillo, J., 2006. Conceptual Modeling for Classi cation Mining in Data Warehouses. *DaWaK*. pp. 566-575.