Chapter 2.2
A Framework for Organizational Data Analysis and Organizational Data Mining

Bernd Knobloch
University of Bamberg, Germany

ABSTRACT

This chapter introduces a framework for organizational data analysis suited for data-driven and hypotheses-driven problems. It shows why knowledge discovery and hypothesis verification are complementary approaches and how they can be chained together. It presents a methodology for organizational data analysis including a comprehensive processing scheme. Employing a plug-in metaphor, data analysis process engineering is introduced as a way to set up data analysis processes based on taxonomies of tasks that have to be performed during data analysis and on the idea of re-using experience from past data analysis projects. The framework aims at increasing the benefits of data mining and other data analysis approaches, by allowing a wider range of business problems to be tackled and by providing the users with structured guidance for planning and running analyses.

INTRODUCTION

The way from raw data to business intelligence is often long and difficult to take, and, sometimes, it does not even lead to where you are heading. Since the time when data mining became popular, lots of methods and algorithms have been proposed, enabling efficient mining of large data sets. However, there are more steps to take than merely putting an algorithm to work on some data. The process embracing all these steps is known as knowledge discovery in databases (KDD). Some KDD tool vendors have covered process issues and provide features for defining sequences of processing steps tailored to meet the specific needs of individual mining cases. However, some, if not most of these approaches ignore conceptual questions, such as which business goal is to be pursued, which data are necessary for producing the desired information, which aspects of data or-

ganization have to be taken into account, etc. The need for a human-centered, process-oriented view on knowledge discovery has long been neglected. Analysts need more support in understanding how to do knowledge discovery (Brachman & Anand, 1996). As Smyth (2001) puts it, users "will almost certainly not say that they need a slightly more accurate decision tree algorithm, or a slightly faster association rule algorithm. Instead their most pressing problem is that of managing the whole process." It's well understood that methods are elementary, but they are of very limited use, if you don't know how to properly employ them to solve your business problems.

Moreover, discussion has almost always been restricted to data mining analyses. Data mining, however, is simply one of several approaches to data analysis. Focusing solely on data mining bears the risk of not applying the one analytical approach that fits the current business objective best. Why rule out the potentials of other approaches, if data mining is not the ultimate choice? Why not combine the strengths of different instruments for data analysis to obtain the best results possible? Why not think big and put the pieces of the puzzle together, constructing an overall framework for business intelligence?

This chapter introduces an organizational framework that helps mend some of the shortcomings mentioned above. That framework provides structured guidance to analysts for putting data analysis to work and covers different approaches to organizational data analysis. It may also serve as reference for documentation of project experience.

BACKGROUND

User support for data analysis is crucial in terms of efficiency. Data analysis is profitable only if the return gained from newly discovered insights is higher than the costs produced by the analysis effort. The efficiency issue has received surprisingly little attention from KDD research. However, some progress has been made.

Brachman and Anand (1996) investigate the interactions between humans and data during analysis, which are important for development of knowledge discovery support tools. Knobloch and Weidner (2000) have taken a problem-oriented approach to infer requirements for analysis tools that help reduce inefficient iterations in the analysis process.

Skalak (2001) suggests three areas for additional research in the data mining field to make data analysis processes more efficient. These areas are modeling the mining process, support for data preparation and automated model selection. Interestingly enough, these topics have already been addressed with considerable success. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a comprehensive model for analysis processes (Chapman, Clinton, Kerner, Khabaza, Reinartz, Shearer & Wirth, 2000) and may serve as recommendable reference. The European Mining Mart project deals with partially automated selection and configuration of preprocessing operations and data mining models. This is accomplished by employment of case-based reasoning to re-use best practices and by multistrategy learning methods for case adaptation (Mining Mart: Enabling End-User Data Warehouse Mining, 2002). In the MetaL project, a similar approach has been pursued, applying metaknowledge and metalevel learning to model selection and method combination (MetaL: A Meta-Learning Assistant for Providing User Support in Machine Learning and Data Mining, 2001). Both projects aim at development of user-oriented KDD support environments.

In spite of the considerable success delivered by these projects, further research is necessary. Still, emphasis lies on knowledge discovery, and other forms of data analysis are hardly addressed. The following sections intend to contribute further ideas to data analysis process research, enabling analysts to deploy data mining technology more efficiently and more effectively within organizations. To achieve that, some of the ideas outlined