Chapter VI
A Component-Based Data Management and Knowledge Discovery Framework for Aviation Studies

M. Brian Blake
Georgetown University, USA
Center for Advanced Aviation System Development, The MITRE Corporation, USA

Lisa Singh
Georgetown University, USA

Andrew B. Williams
Spelman College, USA

Wendell Norman
Center for Advanced Aviation System Development, The MITRE Corporation, USA

Amy L. Sliva
Georgetown University, USA

ABSTRACT

Organizations are beginning to apply data mining and knowledge discovery techniques to their corporate data sets, thereby enabling the identification of trends and the discovery of inductive knowledge. Since traditional transaction databases are not optimized for analytical processing, they must be transformed. This article proposes the use of modular components to decrease the overall amount of human processing and intervention necessary for the transformation process. Our approach configures components to extract data-sets using a set of "extraction hints." Our framework incorporates decentralized, generic components that are reusable across domains and

databases. Finally, we detail an implementation of our component-based framework for an aviation data set.

INTRODUCTION

Over the past decade, government and industry organizations have enhanced their operations by utilizing emerging technologies in data management. Advances in database methodology and software (i.e., warehousing of transaction data) has increased the ability of organizations to extract useful knowledge from operational data and has helped build the foundation for the field of knowledge discovery in databases (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Sarawagi, Thomas, & Agrawal, 2000; Software Suites supporting Knowledge Discovery, 2005). KDD consists of such phases as *selection*, *pre-processing*, *transformation*, *data mining*, and *interpretation/evaluation*. Selection involves identifying the data that should be used for the data mining process. Typically, the data is obtained from multiple, heterogeneous data sources. The pre-processing phase includes steps for data cleansing and the development of strategies for handling missing data and various data anomalies. Data transformation involves converting data from the different sources into a single common format. This step also includes using data reduction techniques to reduce the complexity of the selected data, thereby simplifying future steps in the KDD process. Data mining tasks apply various algorithms to the transformed data to generate and identify "hidden knowledge." Finally, the area of interpretation/evaluation focuses on creating an accurate and clear presentation of the data mining results to the user.

Excluding the data mining phase, where there are a plethora of automated algorithms and applications, the other phases are mostly human-driven. Data experts are required to complete the tasks related to the majority of steps in the KDD process as explained in the following.

- **Data formatting, loading, cleaning and Anomaly detection.** In the pre-processing phase, data experts must correct and update incorrect data values, populate missing data values, and fix data anomalies.
- **Adding important meta-data to the database.** In the data transformation phase, data must be integrated into a single model that supports analytical processing. This typically involves adding meta-data and converting data sets from text files and traditional relational schemas to star or multi-dimensional schemas.
- **User and tool-generated hints.** In the final phases (i.e., data mining and evaluation), general approaches are needed to assist users in preparing knowledge discovery routines and analyzing results. These general approaches must allow the user to manually specify potential correlation areas or "hints." In the future, the suggestion of new hints may be automated by intelligent software mechanisms.

These human-driven tasks pose problems since the initial data set, which we will refer to as the *raw data*, is large, complex and heterogeneous. Our work attempts to reduce the amount of time required for human-driven tasks in the KDD setting. General reusable components may represent a feasible solution to assist in the execution of the time-consuming processing tasks underlying KDD. In this paper, specific tasks suitable for such components are identified and characterized. In addition, a component-based framework and corresponding process are described to address these tasks.

The paper proceeds in the following section with a discussion of related work with respect to component-based KDD. The paper then