Chapter 61
Data Mining, Validation, and Collaborative Knowledge Capture

Martin Atzmueller
University of Kassel, Germany

Stephanie Beer
University Clinic of Wuerzburg, Germany

Frank Puppe
University of Wuerzburg, Germany

ABSTRACT

For large-scale data mining, utilizing data from ubiquitous and mixed-structured data sources, the extraction and integration into a comprehensive data-warehouse is usually of prime importance. Then, appropriate methods for validation and potential refinement are essential. This chapter describes an approach for integrating data mining, information extraction, and validation with collaborative knowledge management and capture in order to improve the data acquisition processes. For collaboration, a semantic wiki-enabled system for knowledge and experience management is presented. The proposed approach applies information extraction techniques together with pattern mining methods for initial data validation and is applicable for heterogeneous sources, i.e., capable of integrating structured and unstructured data. The methods are integrated into an incremental process providing for continuous validation options. The approach has been developed in a health informatics context: The results of a medical application demonstrate that pattern mining and the applied rule-based information extraction methods are well suited for discovering, extracting and validating clinically relevant knowledge, as well as the applicability of the knowledge capture approach. The chapter presents experiences using a case-study in the medical domain of sonography.

DOI: 10.4018/978-1-4666-2455-9.ch061

INTRODUCTION

Whenever data is continuously collected, for example, using intelligent documentation systems in a medical context, data mining and data analysis provide a broad range of options. The mining and analysis step is often implemented using a data-warehouse, e.g., Kimball & Ross (2002). For the data preprocessing and integration of several heterogeneous sources, there exist standardized extract-transform-load (ETL) procedures that need to incorporate suitable data schemas, and integration rules. Additionally, for unstructured or semi-structured textual data sources, the integration requires effective information extraction methods. For clinical discharge letters, for example, the structure of the letter is usually non-standardized, and thus dependent on different writing styles of different authors.

However, a prerequisite of data mining is the validation and the quality assurance of the integrated data. Especially concerning unreliable extraction and integration methods, the quality of the obtained data can vary significantly. If the data has been successfully validated, then the trust in the data mining results and their acceptance can be increased. After that, the comprehensive data set can be used for quality and knowledge management.

This chapter describes a system for integrated data mining and collaborative knowledge management, capture and refinement in a health informatics context. It presents its application using a case-study in the medical domain of sonography.

The approach applies information extraction techniques together with data mining methods for initial data validation and is applicable for heterogeneous sources integrating structured and unstructured data. We focus on an incremental level-wise approach, such that both methods can complement each other in the validation and refinement setting. Validation knowledge can also be formalized in a knowledge base, for assessing known and expected relations. After that, the consolidated data set can be applied for comprehensive quality and knowledge management.

The approach has been implemented in a clinical application context for knowledge management and data mining with data from clinical information systems, documentation systems, and clinical discharge letters. Its workflow features the integration of data mining, information extraction, data validation, and collaborative knowledge management and refinement.

The clinical application is targeted at extended quality control, profiling, and knowledge and experience management in the medical domain of sonography. This application context concerns the data integration from heterogeneous databases and the information extraction from textual documents. After that, the data can be checked with respect to deviations concerning expected relations, relations modeled in the background knowledge as well as by applying statistical validation techniques. For collaborative knowledge management tagging provides helpful techniques: It enables the flexible collection, annotation, organization, and distribution of resources and information. A combination with a wiki then provides powerful but easy to use approaches for a broad range of applications. The presented approach is thus implemented using a system for collaborative knowledge management and refinement, backed by a semantic wiki extension. For this purpose, sonographic images are collected, annotated and commented in order to serve as instructive examples for typical but also exceptional features of certain disorders. In this way, effective tutoring and discussion between the examiners can be initiated and the social collaboration can facilitate incremental knowledge capture and refinement concerning the data collected in the collaborative training system.

The rest of this chapter is structured as follows: We first briefly outline the application context given by the SONOCONSULT knowledge system. Furthermore, we present the mining and validation approach in detail. Next, we describe the social application for collaborative knowledge capture