Chapter 5
Data Mining

Martin Atzmueller
University of Kassel, Germany

ABSTRACT

Data Mining provides approaches for the identification and discovery of non-trivial patterns and models hidden in large collections of data. In the applied natural language processing domain, data mining usually requires preprocessed data that has been extracted from textual documents. Additionally, this data is often integrated with other data sources. This chapter provides an overview on data mining focusing on approaches for pattern mining, cluster analysis, and predictive model construction. For those, we discuss exemplary techniques that are especially useful in the applied natural language processing context. Additionally, we describe how the presented data mining approaches are connected to text mining, text classification, and clustering, and discuss interesting problems and future research directions.

INTRODUCTION

In the context of applied natural language processing, data mining provides for powerful approaches for the mining and discovery of regularities, or patterns, in structured data. In the text and language processing context, the input data, that is, structured data is often acquired in a preprocessing and data integration phase. The original sources are usually textual documents that are first preprocessed and integrated with other data sources. Text mining, information extraction, and data integration methods provide powerful tools for these steps.

This chapter focuses on data mining in the applied natural language processing context. Data mining is used for mining patterns or models from data. A closely related topic is text mining, which partly overlaps with data mining: While data mining considers structured data as input, text mining directly works on unstructured or semi-structured documents. In that sense, several text

DOI: 10.4018/978-1-60960-741-8.ch005

mining methods can be regarded as preprocessing steps for data mining, while data mining can be interpreted as the last step in the text mining process. The specific mining methods applied for both topics are usually quite similar, and have their origin in the data mining and knowledge discovery context.

In the following, we first describe the background of data mining and outline the general data mining process according to the CRISP-DM process model. After that, the remaining chapter is split into three parts: First, we consider *pattern mining*, including frequent pattern mining, association rule mining and subgroup mining. Next, we focus on *cluster analysis* including partitioning and hierarchical approaches. Finally, we discuss *predictive model* construction, discussing decision trees, naïve Bayes, k-nearest neighbor and rule-based (associative) classification. Since the latter two sections concerning prediction, classification and clustering are also covered in specialized chapters on text classification and text clustering, we outline the basics of the approach, put them into context and refer to the respective chapters for more details.

Considering the CRISP-DM process model, we especially concentrate on pattern mining (i.e., the knowledge discovery step in this chapter) in order to show how data mining can be applied for knowledge discovery, extraction or rapid knowledge capture using automated mining methods. Then, the extracted patterns can be applied for further analysis, knowledge engineering, and knowledge capture.

The objectives of the chapter are to provide an overview on the data mining field and to introduce prominent exemplary techniques in the context of applied natural language processing. Since data mining is closely related to text mining, we include a discussion of issues that are present in the intersection of both topics, and refer to the respective text mining chapter for more details. Furthermore, we discuss open issues, and interesting directions for future research.

BACKGROUND

Data mining, also popularly referred to as knowledge discovery in databases (KDD) is concerned with the automatic or semi-automatic extraction of patterns. These patterns represent knowledge implicitly stored in large databases, data warehouses, the Web, other massive information repositories, and data streams. Informally, data mining is used for obtaining patterns and summaries of new and nontrivial information based on the available data (*description*), alternatively for the creation of predictive models of a certain system or phenomena (*prediction*).

The literature mentions several definitions of data mining, also in relation to knowledge discovery in databases. Fayyad et al. (1996), for example, define KDD as: "the process of discovering valid, novel, interesting, and potential useful knowledge," data mining is considered as the core step of the whole KDD process, that is, the concrete knowledge discovery method. Other definitions regard data mining as the "process of discovering various models, summaries, and derived values from a given collection of data" – see Kantardzic (2002). Han and Kamber (2006) also consider data mining a step in the knowledge discovery process (i.e., as an "essential process where intelligent methods are applied in order to extract data patterns"), but choose to use the term "data mining" in favor of "knowledge discovery in databases," subsuming the older term. They therefore take a broad view of data mining functionality, and consider data mining as the general process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.

To sum up, data mining has been considered as an approach and a collection of methods, as a process, or a step in the general knowledge discovery process. In the following, we will take a broader view, similar to Han and Kamber (2008) and consider data mining as a comprehensive approach that can be mapped to a concrete process,