Chapter 7
Learning Different Concept Hierarchies and the Relations between them from Classified Data

Fernando Benites
University of Konstanz, Germany

Elena Sapozhnikova
University of Konstanz, Germany

ABSTRACT

Methods for the automatic extraction of taxonomies and concept hierarchies from data have recently emerged as essential assistance for humans in ontology construction. The objective of this chapter is to show how the extraction of concept hierarchies and finding relations between them can be effectively coupled with a multi-label classification task. The authors introduce a data mining system which performs classification and addresses both issues by means of association rule mining. The proposed system has been tested on two real-world datasets with the class labels of each dataset coming from two different class hierarchies. Several experiments on hierarchy extraction and concept relation were conducted in order to evaluate the system and three different interestingness measures were applied, to select the most important relations between concepts. One of the measures was developed by the authors. The experimental results showed that the system is able to infer quite accurate concept hierarchies and associations among the concepts. It is therefore well suited for classification-based reasoning.

DOI: 10.4018/978-1-4666-2455-9.ch007

INTRODUCTION

The complexity and amount of data on the Internet as well as in science and industry is growing rapidly. As a result of increased data complexity, it is becoming more and more difficult to find and manage relevant materials. A successful way to cope well with the enormous information flow is the use of taxonomies with hierarchically structured concepts which are usually called concept hierarchies or, more generally, ontologies, and are provided by domain experts. A common example may be a tourism ontology containing concepts such as accommodation, attractions and transport (where, for example, "hotel" and "youth hostel" are subcategories of "accommodation"). This kind of ontology formally specifies the concepts and their relationships in a domain and facilitates information search. Furthermore, ontologies are very useful for knowledge representation purposes, e.g. in the semantic Web, and therefore have become an active research field in the last few years. Unfortunately, manual creation of concept hierarchies is very cost-intensive and often impossible in face of the large scale of many domains. Additionally, it is expensive and tedious to maintain complex taxonomies manually when the data are changing frequently. For this reason, the acquisition of ontologies directly from raw text also known as ontology learning (Maedche & Staab, 2001) has recently emerged as an alternative to manual ontology building with the aim of assisting humans in ontology construction. This problem has been extensively studied in the past (Bendaoud, Hacene, Toussaint, Delecroix, & Napoli, 2007, Cimiano, Hotho, & Staab, 2005, Omelayenko, 2001).

As a subfield of text mining, these methods are closely related to the field of natural language processing. Some of them use available semantic knowledge such as the synonymous or antonymous relations between concepts. Another approach is represented by syntactical methods, which treat textual or other data as abstract structures. Besides the domain-specific text corpora, other types of data like databases may also comprise hidden or implicit taxonomic information, which can be discovered in many applications from Web mining to bioinformatics. Syntactical data-driven methods for the automatic extraction of taxonomies and concept hierarchies have been shown to be successful in assisting knowledge engineers with ontology development (Majidian & Martin, 2009). In general, any classified data containing class labels are potential sources for inferring concept hierarchies. Our previous work (Brucker, Benites, & Sapozhnikova, 2011) has shown that it is possible to extract quite accurate taxonomies by analyzing co-occurrences between labels in multi-label classification, i.e. when an instance belongs to more than one class. A movie, for example, can be classified into crime, thriller, and horror genre categories simultaneously. The presented work extends the research area to the analysis of multiple class taxonomies.

Often several ontologies may exist for the same data, each of them representing a specific point of view on a certain domain. There are a lot of studies considering similar taxonomies where the tasks of ontology merging, mapping or alignment are often dealt with in order to unify available knowledge (Choi, Song, & Han, 2006). On the other hand, multiple taxonomies may be used either for specializing in sub-domains or for providing different perspectives (Wimalasuriya & Dou, 2009). In the latter case, which is the focus of our research, a movie, for example, can be classified either by its genre into a genre taxonomy or by the producing company in a taxonomy of producers. This task of relating the taxonomies of different natures is much less investigated. In such a case, the combination of information from taxonomies providing different insights into a problem domain can lead to the discovery of new knowledge. Solving this knowledge extraction task enables, for instance, the integration of evidence from multiple data sources and thus decreases the