Chapter II
Raising, to Enhance Rule Mining in Web Marketing with the Use of an Ontology

Xuan Zhou
VPIsystems Corporation, USA

James Geller
New Jersey Institute of Technology, USA

ABSTRACT

This chapter introduces Raising as an operation that is used as a preprocessing step for data mining. In the Web Marketing Project, people's demographic and interest information has been collected from the Web. Rules have been derived using this information as input for data mining. The Raising step takes advantage of an interest ontology to advance data mining and to improve rule quality. The definition and implementation of Raising are presented in this chapter. Furthermore, the effects caused by Raising are analyzed in detail, showing an improvement of the support and confidence values of useful association rules for marketing purposes.

INTRODUCTION

Marketing has faced new challenges over the past decade. The days of the mass market are definitely over. Consumers now are exposed to numerous cable channels and satellite channels. Many people do not get their information from TV at all, but use Web sites. The population has also developed. Minorities have grown and asserted their own tastes and needs. A product that is attractive to the average white Anglo-Saxon or Italian citizen might be completely uninteresting to a first generation South American immigrant. Similarly, the market has split up by preferences. Chinese and Indian food have made major inroads and many consumers would like to cook the same

food in their homes. In short, the mass market is dead, and marketers today face the problem of advertising to many disjoint niche markets.

With the increase in available, cheap data storage, companies are keeping terabytes of information about their customers. Today, it is not outrageous to talk about one-to-one marketing. However, marketers face two problems. They may have information about previous customers, but how could they get personal information about potential customers? Secondly, if information about individuals is truly not accessible, how could they classify such individuals into small categories and then market effectively to these small categories?

To provide a solution for these two problems, the Web Marketing Project (Geller, Scherl, & Perl, 2002; Scherl & Geller, 2002) was created. This project targeted millions of publicly accessible home pages on the Web, on which people freely express their likes and dislikes. These pages are a valuable source of data for marketing purposes. One approach is to use the contact information for direct (e-mail) marketing. For example, if someone expressed his interest as music, then he might be a potential customer of music CDs. Thus the marketing can be directed towards a very narrow niche. If someone lists very detailed interests, such as the TV comedy show *The Simpsons*, the Season 8 DVD released in August, 2006, could be one of his must-buy products.

A second important use of this data is for finding interesting marketing knowledge. The data may be mined for useful correlations between interests and also between interests and demographic categories. If someone is interested in *The Simpsons*, what is the likelihood that he is interested in another comedy? What age groups are interested in particular types of TV series? The available data can be used for such investigations. The results may again be useful for marketing.

In the Web Marketing Project, we collected people's demographic and interest information from home pages and stored them in a database. There are six modules in this project, which are Web search, glossary-based extraction, database, data mining, ontology, and front end, as described in detail in Zhou (2006). In this chapter, we only focus on the ontology and data mining modules. The ontology consists of two taxonomies, one of which describes different customer classifications, while the other one contains a large hierarchy, based on Yahoo, which contains 31,534 interests. For the customer classification, an intersection ontology (Zhou, Geller, Perl, & Halper, 2006) was developed.

The data mining module uses well-known data mining algorithms to extract association rules from the given data. The WEKA (Witten & Frank, 2000) package was used at the beginning of the project. From the WEKA package, the Apriori algorithm (Agrawal & Srikant, 1994) for data mining was used. The real world data about real people tends to produce rules with unsatisfactory support values. Thus, in this research a method was developed for improving the support values of rules by using the existing ontology. This method is called "*Raising*" and will be discussed in depth later in this chapter. Moreover, due to the limitations of WEKA found during the project, the FP-Growth algorithm (Han, Pei, & Yin, 2000; Han, Pei, Yin, & Mao, 2004) was implemented and used in the second stage to correct some errors and improve the results.

The next section presents previous literature on ontologies used in rule mining. Following that, we introduce the Raising method and show how an ontology can be used to improve the support of mined rules. The effects caused by Raising on derived rules are discussed afterwards. At last, future trends and conclusions are presented at the end of this chapter.

BACKGROUND

A concept hierarchy is present in many databases either explicitly or implicitly. Some previous work