Association Rule Based Specialization in ER Models

Martine De Cock[1], Chris Cornelis[1], Ming Ren[2], Guoqing Chen[2], and Etienne E. Kerre[1]

[1] Ghent University, Fuzziness and Uncertainty Modelling Research Unit, Krijgslaan 281 (S9), 9000 Gent, Belgium
{martine.decock|chris.cornelis|etienne.kerre}@ugent.be
[2] Tsinghua University, School of Economics and Management, Beijing 100084, China
{renm|chengq}@em.tsinghua.edu.cn

Abstract. Association rules (ARs) emerged in the domain of market basket analysis and provide a convenient and effective way to identify and represent certain dependencies between attributes in a database. In this paper, we demonstrate that they also act as an appropriate aid in the construction and enrichment of entity-relationship (ER) models, structuring tools that provide high-level descriptions of data. In particular, we present different conceptual ideas for semi-automated specialization of ER models based on AR mining.

Key words: entity-relationship model, association rules, fuzzy sets, database design

1 Introduction

The entity-relationship (ER) model is a conceptual model that describes real world phenomena in terms of entities, relationships between those entities, and attributes of both of them. In an ER model for a grocery store for instance we typically encounter entity classes such as *product* and *customer*, having attributes such as *price* and *freshness date* (for the product) and *age* and *sex* (for the customer). *Purchase* is an example of a relationship class between these two entity classes, while *quantity* and *time* are examples of attributes of the purchase relationship class.

The ER model is a powerful means for business and data modelling that helps to identify essential elements of the domain of interest in a conceptual and integrated manner. Initially introduced in [10], the methodology itself has evolved considerably and has become widely accepted as a standard design tool for relational databases [11]. During the past decades, basic ER concepts have been extended in various ways, resulting in enhanced ER models. In this

Martine De Cock et al.: *Association Rule Based Specialization in ER Models*, Studies in Computational Intelligence (SCI) **5**, 203–217 (2005)
www.springerlink.com

paper we will focus on specialization: the process of defining subclasses for a given entity or relationship class. Referring to the grocery store example, for the entity class *product* we might define the subclasses[1]: *product with price in* $[0, 20[$, *product with price in* $[20, 100[$ and *product with price in* $[100, +\infty[$. *Product* is then called a superclass of these subclasses.

One of the advantages of ER models is that they are easy to understand with a minimum of training; hence they are very suitable to communicate and discuss the database design with the end user (e.g. the shop owner). Also in this respect it is more convenient to denote the subclasses of attribute values by linguistic terms in the ER model. Indeed a linguistic expression such as *cheap product* corresponds better to the shop owner's daily use of language than *product with price in* $[0, 20[$. Now regarding the implementation of the ER model, one might argue that it is against intuition to call a product of 19.90 EUR cheap and one of 20 EUR not. The transition between being cheap and not being cheap is not abrupt but gradual. Hence it makes more sense to model linguistic terms such as *cheap*, *medium* and *expensive* by fuzzy sets [30], characterized by membership functions that associate with every price a number between 0 and 1 indicating the degree to which this price can be called *cheap*, *medium* and *expensive* respectively. So-called *fuzzy* ER models have been proposed from different perspectives [4, 20, 31].

Traditionally, ER models are built upon the knowledge of business managers and database designers. However, as the real world phenomena represented by the ER model change, and our understanding of the world improves, the need arises for an extension or enrichment of the original ER model. In this paper we propose to use assocation rule mining as a tool for the semi-automatic construction and enrichment of ER models, more in particular for the specialization process. We treat the specialization of entities and relations separately (so called E-specialization and R-specialization).

Our main aim where E-specialization is concerned, is to identify subclasses of entities denoted by linguistic expressions that are common in the domain of interest. For instance typical subclasses of the entity class *product* in an interior decoration store are *bathroom accessories*, *beds and mattresses*, *storage systems*, *chairs*, *cookware*, *for the pets* etc. Each of those in turn can be divided into subclasses; for instance for *cookware* this might be *kitchen storage and containers*, *kitchen utensils and accessories*, *knives and chopping board*, *pots, pans and ovenware*. Our aim is to construct such a specialization relation between linguistic expressions (further on called "terms") automatically. To this end, we start from a collection of documents containing text that is highly related to the phenomena represented by the ER model. From this collection, we generate a document-term table. In this table we mine for association rules between terms.

If available at all, a text collection related to the domain of interest is usually already at our disposal when domain experts start constructing the

[1] $[a, b[$ denotes the real interval that contains a but not b.