An Efficient Mining of Dominant Entity Based Association Rules in Multi-databases

V.S. Ananthanarayana

Department of Information Technology
National Institute of Technology Karnataka, Surathkal - 575 025, India
anvs@nitk.ac.in

Abstract. Today, we have a large collection of data that is organized in the form of a set of relations which is partitioned into several databases. There could be implicit associations among various parts of this data. In this paper, we give a scheme for retrieving these associations using the notion of dominant entity. We propose a scheme for mining for dominant entity based association rules (DEBARs) which is *not constrained to look for co-occurrence* of values in tuples. We show the importance of such a mining activity by taking a practical example called personalized mining. We introduce a novel structure called multi-database domain link network (MDLN) which can be used to generate DEBARs between the values of attributes belonging to different databases. We show that MDLN structure is *compact* and this property of MDLN structure permit it to be used for mining vary large size databases. Experimental results reveal the efficiency of the proposed scheme.

Keywords: Multi-databases, Valueset, Personalized mining, Dominant entity, Association rules.

1 Introduction

In real-life situations, associations among data items are hidden and mining for these associations is further complicated because of possible distribution of the data across several databases [1] [2]. For example, consider *person based information system*; wherein information about the person is available as customer in SUPERMARKET database, as employee in COMPANY database, as patient in the MEDICAL CENTRE database, and as passenger in TRAVEL database. There could be an implicit association among various parts of this data. For example, there may be an association between salary, region of living, mode of traveling and disease; like *people who have salary in the range of US$ 1000 - US$ 2,000, eat frequently at CENTRAL CALCUTTA hotels and travel by air in executive class have cardiovascular diseases*, where CENTRAL CALCUTTA is marked with a high pollution rating. Such examples provide the associations of a person in different contexts. We call such a mining activity **person-based associations mining** (or **personalized mining**) and the system under which the mining activity performed is called **person warehouse**. The primary property of such a warehouse is the possibility to generate a global schema linked by

S.S. Al-Majeed et al. (Eds.): WiMoA 2011/ICCSEA 2011, CCIS 154, pp. 91–100, 2011.

one entity which we call, *dominant entity*. The associations between the values of attributes of interest using the dominant entity are called **dominant entity based associations**. In the above example, the dominant entity is *person* which is characterized by attributes *name and address*. The set of attributes which characterizes dominant entity is called *dominant entity attributes (DEA)*. The sets of attributes between which we want to find the associations are called *characteristic attributes (CA)*. In *person based information system* example quoted above, {salary, region-of-living}, {mode-of-travel}, {diseases} are characteristic attributes.

In this paper, we are addressing the problem of 'dominant entity based association mining activity in multiple databases'. For illustration purposes we use 'person warehouse' through out the paper. However, the notion can suit any domain where we have a *dominant entity* and the data warehouse.

1.1 Problem Definition

The conventional association rules are mainly based on togetherness or co-occurrence of values in *a* tuple [3] [6]. However, in the dominant entity based mining like personalized mining, we want to mine the rules of the following type: *income* $>$ *"Range X"* $\wedge$ *age* $<$ *"Range Y"* $\Longrightarrow$ *purchase* $=$ *"Costly goods"*. Note that in order to generate such rules from multiple relations/databases, support for individual values are important and associations are built using the dominant entity. Such rules are called *dominant entity based association rules, (DEBAR)*. Further, in the above rule, all "Costly goods" need not be purchased together. (The goods purchased is said to be "costly" if the cost is more than "Z Dollars"). We originate a scheme for mining for dominant entity based association rules which is *not constrained to look for co-occurrence* of values in tuples. It can look at attributes in one more databases which may be located at many places. We discuss this scheme in section 2.

In order to generate DEBARs, there is a need to link the values of attributes of interest to the values of dominant entity attributes. We propose and develop a structure called "multi-database domain link network (MDLN)" for the mining activity that involves several databases. This structure provides link between the values of characteristic attributes and the values of dominant entity attributes in an efficient manner in terms of - *space* to hold the structure; and *access time* to generate DEBARs out of it.

The rest of this paper is organized as follows. We discuss MDLN structure in section 3. We show how meaningful associations can be mined using MDLN in section 4. Experimental results are described in section 5. We conclude our study in section 6.

2 Dominant Entity Based Association Rules (DEBAR)

Let v^x_{aij} and v^y_{bkl} be the j^{th} and l^{th} values of characteristic attributes, A^x_{ai} and A^y_{bk} which belong to relations R^x_a and R^y_b of databases D_x and D_y respectively. Let $d_1, d_2, \ldots d_N$ be the N distinct values pertaining to DEA.