Constrained Cube Lattices for Multidimensional Database Mining

Lotfi Lakhal, Aix-Marseille Université, France
Alain Casali, Aix-Marseille Université, France
Rosine Cicchetti, Aix-Marseille Université, France
Sébastien Nedjar, Aix-Marseille Université, France

ABSTRACT

In multidimensional database mining, constrained multidimensional patterns differ from the well-known frequent patterns from both conceptual and logical points of view because of a common structure and the ability to support various types of constraints. Classical data mining techniques are based on the power set lattice of binary attribute values and, even adapted, are not suitable when addressing the discovery of constrained multidimensional patterns. In this paper, the authors propose a foundation for various multidimensional database mining problems by introducing a new algebraic structure called cube lattice, which characterizes the search space to be explored. This paper takes into consideration monotone and/or anti-monotone constraints enforced when mining multidimensional patterns. The authors propose condensed representations of the constrained cube lattice, which is a convex space, and present a generalized levelwise algorithm for computing them. Additionally, the authors consider the formalization of existing data cubes, and the discovery of frequent multidimensional patterns, while introducing a perfect concise representation from which any solution provided with its conjunction, disjunction and negation frequencies. Finally, emphasis on advantages of the cube lattice when compared to the power set lattice of binary attributes in multidimensional database mining are placed.

Keywords: Datacubes, Inclusion-Exclusion Identities, Lattices, Multidimensional Database Mining, Search Space

INTRODUCTION AND MOTIVATIONS

The extraction of constrained multidimensional patterns in the area of multidimensional database mining (OLAP database mining or data mining from categorical database relations) is achieved for solving various problems such as discovering multidimensional association rules (Lu, Feng, & Han, 2000), Roll-Up dependencies (Calders, Ng, & Wijsen, 2002), multidimensional constrained gradients (Dong et al., 2004), closed constrained gradients (Wang, Han, & Pei, 2006), classification rules (Liu, Hsu, & Ma,
correlation rules (Brin, Motwani, & Silverstein, 1997; Grahne, Lakshmanan, & Wang, 2000), datacube (Gray et al., 1997; Beyer & Ramakrishnan, 1999; Han, Pei, Dong, & Wang, 2001; Xin, Han, Li, & Wah, 2003), iceberg cube (Beyer & Ramakrishnan, 1999; Han et al., 2001), emerging cube (Nedjar, Casali, Cicchetti, & Lakhal, 2007; Casali, Nedjar, Cicchetti, & Lakhal, 2009b), quotient cube (Lakshmanan, Pei, & Han, 2002), and closed cube (Casali, Cicchetti, & Lakhal, 2003b; Li & Wang, 2005; Xin, Shao, Han, & Liu, 2006; Casali, Nedjar, Cicchetti, & Lakhal, 2009a). We believe that a precise semantics is required for characterizing the search space and solving these multidimensional data mining problems. Such semantics can be captured through an algebraic structure, the cube lattice, provided with a similar expression power than the power set lattice, which is used for binary database mining (transaction database mining (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996)).

Adapting to this new multidimensional context, approaches and algorithms successfully used when mining binary databases (and thus using the power set lattice as a search space) is possible but not relevant. However, such adaptations have been frequently proposed for the extraction of quantitative association rules (Srikant & Agrawal, 1996) and for classification (Liu et al., 1998; Li et al., 2001). Moreover, (Beyer & Ramakrishnan, 1999; Han et al., 2001) have extended Apriori (Agrawal et al., 1996) for computing iceberg cubes and observed that such extensions “perform terribly”. Reasons behind these failures are the following. Firstly each multidimensional attribute must be replaced by a set of binary attributes, each of which representing a single value of the multidimensional attribute (Srikant & Agrawal, 1996). If the original attribute domains are large (which is the case in data warehouses (Beyer & Ramakrishnan, 1999)), the attribute substitution results in dealing with a great number of binary attributes (called items). On the other hand, the search space, considered when mining binary databases, is the lattice representing the power set of items. Nevertheless, this large search space encompasses a great number of solutions, which are known to be semantically erroneous in a multidimensional context. Let us suppose that in a relation the attribute A has k distinct values. It is replaced by k binary attributes \( a_1, ..., a_k \). In the powerset lattice, all the couples \((a_i, a_j)\) with \(1 \leq i, j \leq k\) and \(i \neq j\) are considered and evaluating a constraint requires a costly memory space and a pass over the binary relation. But we know that the original data set does not contain any pattern \((a_i, a_j)\), simply because the initial attribute A is atomic and thus its values \(a_i\) and \(a_j\) are exclusive.

If an anti-monotone constraint w.r.t. inclusion (e.g., \(\text{Frequency()} \geq \text{threshold}\)) is enforced for mining frequent multidimensional patterns, the original complexity of levelwise algorithms is altered because the negative border (fundamental concept for the complexity analysis (Mannila & Toivonen, 1997)) used for pruning is enlarged with a possibly voluminous set of useless combinations. The problem worsens when a monotone constraint (e.g., \(\text{Frequency()} \leq \text{threshold}\)) is considered because, although erroneous, the pattern \((a_i, a_j)\) belongs to the yielded results.

In this paper, we make the following contributions.

1. We introduce and characterize the search space to be considered for multidimensional database mining problems. Such a search space only encompasses semantically valid solutions. By introducing an order relation between elements of this space, and proposing two construction operators, we define a new algebraic structure, called cube lattice. In such groundwork, the extraction of multidimensional patterns can be achieved by using conjunctions of monotone and/or anti-monotone constraints according to our order relation.

2. We formalize condensed representations of constrained cube lattices. Condensed representations based on boundary sets (or borders) avoid to enumerate all the solutions (Mannila & Toivonen, 1996, 1997).
Related Content

An Efficient Method for Discretizing Continuous Attributes
[www.irma-international.org/article/efficient-method-discretizing-continuous-attributes/42149/](http://www.irma-international.org/article/efficient-method-discretizing-continuous-attributes/42149/)

Social Network Construction in the Information Age: Views and Perspectives
[www.irma-international.org/chapter/social-network-construction-information-age/61516/](http://www.irma-international.org/chapter/social-network-construction-information-age/61516/)

Lossless Reduction of Datacubes using Partitions
[www.irma-international.org/article/lossless-reduction-datacubes-using-partitions/1821/](http://www.irma-international.org/article/lossless-reduction-datacubes-using-partitions/1821/)

Towards Harnessing Phone Messages and Telephone Conversations for Prediction of Food Crisis

SeqPAM: A Sequence Clustering Algorithm for Web Personalization
[www.irma-international.org/article/seqpam-sequence-clustering-algorithm-web/1777/](http://www.irma-international.org/article/seqpam-sequence-clustering-algorithm-web/1777/)