Chapter 7
Finding Minimal Infrequent Elements in Multi-Dimensional Data Defined over Partially Ordered Sets and its Applications

Khaled M. Elbassioni
Max-Planck-Institute für Informatik, Saarbrücken, Germany

ABSTRACT
The authors consider databases in which each attribute takes values from a partially ordered set (poset). This allows one to model a number of interesting scenarios arising in different applications, including quantitative databases, taxonomies, and databases in which each attribute is an interval representing the duration of a certain event occurring over time. A natural problem that arises in such circumstances is the following: given a database $D$ and a threshold value $t$, find all collections of “generalizations” of attributes which are “supported” by less than $t$ transactions from $D$. They call such collections infrequent elements. Due to monotonicity, they can reduce the output size by considering only minimal infrequent elements. We study the complexity of finding all minimal infrequent elements for some interesting classes of posets. The authors show how this problem can be applied to mining association rules in different types of databases, and to finding “sparse regions” or “holes” in quantitative data or in databases recording the time intervals during which a re-occurring event appears over time. Their main focus will be on these applications rather than on the correctness or analysis of the given algorithms.

INTRODUCTION
The problem of mining association rules from large databases has emerged as an important area of research since their introduction in (Agrawal et al., 1993). Typically, the different data attributes exhibit certain correlations between them, which can be summarized in terms of certain rules, provided that enough transactions or records in the database agree with these rules. For a few examples, in a database storing sets of items purchased by different customers in a supermarket, it may be interesting to
Finding Minimal Infrequent Elements

observe a rule of the form “most customers that purchase bread and butter tend also to purchase orange juice”; in a database storing personal data about individuals, it may be interesting to observe that “most individuals who are married and with age in the range 28-34 have at least 2 cars”; and in a database storing data about the time periods a given service is used by different customers, an interesting observation may take the form: “customers who make full use of the service between 2:00-3:00 on Friday tend also to use the service between 2:00-3:00 on Saturday”. Such information could be useful, for example, for placing items next to each other on supermarket shelves or providing better services for anticipated customers.

Most of the work on finding association rules divides the task into two basic steps: the first one is to identify those collections of items or attribute values that appear together frequently in the database, the so-called frequent itemsets; the second step is to generate association rules from these. While the first step has received considerable attention in the literature, with many algorithms proposed, the second step seems to be somehow overlooked. In this chapter, we will have a more careful look at this latter step, and show in fact that a lot of redundancy can be eliminated from the generated rules by solving the somewhat complementary problem of finding infrequent sets, i.e., those collections of items that rarely appear together in any transaction. This gives one important motivation for studying the problem of finding infrequent collections of values that can be assumed by the attributes of the given database. But apart from that, finding such collections is a problem of independent interest, since each infrequent collection of attribute values indicates rare associations between these values. For instance, in the database of personal data above one can observe a rule like “no individuals with age between 26 and 38 have a single car”, and in the database recording service usage, one may observe that “Fewer than 40% of the customers occupy the service on Friday between 2:00-3:00 and on Saturday between 2:00-4:00”. Another application will be given in Section 4.2, in which the objective is to discover the so-called rare association rules, which are informally rules that result from data appearing rarely in the database.

Rather than using binarization, as is common in the literature (see e.g. Srikant et al., 1995; Srikant et al. 1996), to represent the different ranges of each attribute by binary values, we shall consider more generally databases in which each attribute assumes values belonging to a partially ordered set (poset). This general framework will allow us to model a number of different scenarios in data mining applications, including the mining of association rules for databases with quantitative, categorical and hierarchical attributes, and the discovery of missing associations or “holes” in data (see (Agrawal et al., 1996; Liu et al., 1997; Mun et al., 1998)). One important feature of this framework is that it allows us to find generalized associations, which are obtained by generalizing some attribute values, for which otherwise there exist no enough support from the database transactions. As an example on the supermarket data above, it may be the case that most customers who purchase milk products tend also to purchase bread, but in the database only “cheese” and “butter” appear as items. In this case generalizing both these items to “milk products” allows us to discover the above rule.

We begin our exposition in the next section with recalling some definitions and terminologies related to partially ordered sets and give some examples of databases defined over products of posets. In Section 3, we define the main object of interest in this chapter, namely minimal infrequent elements in products of posets, describe the associated enumeration problem, and discuss how to measure its complexity. Section 4 gives some applications of such enumeration problems to finding association rules in different types of databases and to finding sparse regions on quantitative data. In Section 5, we discuss briefly the complexity of finding infrequent/minimal infrequent elements.
Related Content

A Study and Comparison of Sentiment Analysis Techniques Using Demonetization: Case Study
[www.igi-global.com/chapter/a-study-and-comparison-of-sentiment-analysis-techniques-using-demonetization/210959?camid=4v1a](www.igi-global.com/chapter/a-study-and-comparison-of-sentiment-analysis-techniques-using-demonetization/210959?camid=4v1a)

Influences of Factors of Human Resources for Innovation in Services Companies
[www.igi-global.com/chapter/influences-of-factors-of-human-resources-for-innovation-in-services-companies/117853?camid=4v1a](www.igi-global.com/chapter/influences-of-factors-of-human-resources-for-innovation-in-services-companies/117853?camid=4v1a)

Incorporating Correlations among Gene Ontology Terms into Predicting Protein Functions
Pingzho Hu, Hui Jiang and Andrew Emili (2011). *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances* (pp. 154-173).
[www.igi-global.com/chapter/incorporating-correlations-among-gene-ontology/53885?camid=4v1a](www.igi-global.com/chapter/incorporating-correlations-among-gene-ontology/53885?camid=4v1a)

Knowledge Discovery in Biomedical Data Facilitated by Domain Ontologies
[www.igi-global.com/chapter/knowledge-discovery-biomedical-data-facilitated/24907?camid=4v1a](www.igi-global.com/chapter/knowledge-discovery-biomedical-data-facilitated/24907?camid=4v1a)