Reasoning about Frequent Patterns with Negation

Marzena Kryszkiewicz
Warsaw University of Technology, Poland

INTRODUCTION

Discovering frequent patterns in large databases is an important data mining problem. The problem was introduced in (Agrawal, Imielinski, & Swami, 1993) for a sales transaction database. Frequent patterns were defined there as sets of items that are purchased together frequently. Frequent patterns are commonly used for building association rules. For example, an association rule may state that 80% of customers who buy fish also buy white wine. This rule is derivable from the fact that fish occurs in 5% of sales transactions and set {fish, white wine} occurs in 4% of transactions. Patterns and association rules can be generalized by admitting negation. A sample association rule with negation could state that 75% of customers who buy coke also buy chips and neither beer nor milk. The knowledge of this kind is important not only for sales managers, but also in medical areas (Tsumoto, 2002). Admitting negation in patterns usually results in an abundance of mined patterns, which makes analysis of the discovered knowledge infeasible. It is thus preferable to discover and store a possibly small fraction of patterns, from which one can derive all other significant patterns when required. In this chapter, we introduce first lossless representations of frequent patterns with negation.

BACKGROUND

Let us analyze sample transactional database $\mathcal{D}$ presented in *Table 1*, which we will use throughout the chapter. Each row in this database reports items that were purchased by a customer during a single visit to a supermarket.

As follows from *Table 1*, items *a* and *b* were purchased together in four transactions. The number of transactions in which set of items $\{x_1, ..., x_n\}$ occurs is called its *support* and denoted by $sup(\{x_1, ..., x_n\})$. A set of items is called a *frequent pattern* if its support exceeds a user-specified threshold (*minSup*). Otherwise, it is called an *infrequent pattern*. In the remainder of the chapter, we assume *minSup* = 1. One can discover 27 frequent patterns from D, which we list in *Figure 1*.

Table 1. Sample database $\mathcal{D}$

Id	Transaction
T_1	{*abce*}
T_2	{*abcef*}
T_3	{*abch*}
T_4	{*abe*}
T_5	{*acfh*}
T_6	{*bef*}
T_7	{*h*}
T_8	{*af*}

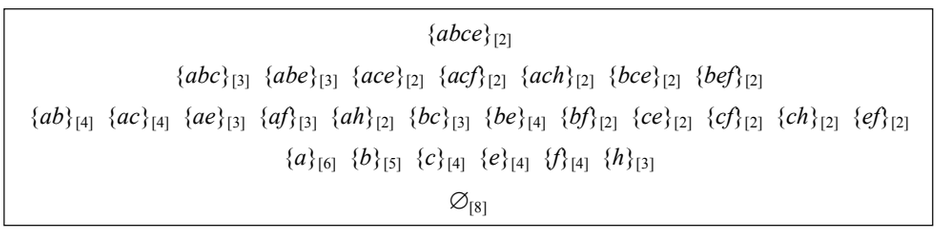
One can easily note that the support of a pattern never exceeds the supports of its subsets. Hence, subsets of a frequent pattern are also frequent, and supersets of an infrequent pattern are infrequent.

Aside from searching for only statistically significant sets of items, one may be interested in identifying frequent cases when purchase of some items (presence of some symptoms) excludes purchase of other items (presence of other symptoms). Pattern consisting of items $x_1, \ldots, x_m$ and negations of items $x_{m+1}, \ldots, x_n$ will be denoted by $\{x_1, \ldots, x_m, \text{-}x_{m+1}, \ldots, \text{-}x_n\}$. The *support of pattern* $\{x_1, \ldots, x_m, \text{-}x_{m+1}, \ldots, \text{-}x_n\}$ is defined as the number of transactions in which all items in set $\{x_1, \ldots, x_m\}$ occur and no item in set $\{x_{m+1}, \ldots, x_n\}$ occurs. In particular, $\{a(-b)\}$ is supported by two transactions in $\mathcal{D}$, while $\{a(-b)(-c)\}$ is supported by one transaction. Hence, $\{a(-b)\}$ is frequent, while $\{a(-b)(-c)\}$ is infrequent.

From now on, we will say that *X* is a *positive pattern*, if *X* does not contain any negated item. Otherwise, *X* is called a *pattern with negation*. A pattern obtained from pattern *X* by negating an arbitrary number of items in *X* is called a *variation of X*. For example, $\{ab\}$ has four distinct variations (including itself): $\{ab\}$, $\{a(-b)\}$, $\{(-a)b\}$, $\{(-a)(-b)\}$.

One can discover 109 frequent patterns in $\mathcal{D}$, 27 of which are positive, and 82 of which have negated items.

Figure 1. Frequent positive patterns discovered from database $\mathcal{D}$. Values provided in square brackets in the subscript denote supports of patterns.



In practice, the number of frequent patterns with negation is by orders of magnitude greater than the number of frequent positive patterns.

A first trial to solve the problem of large number of frequent patterns with negation was undertaken by Toivonen (1996), who proposed a method for using supports of positive patterns to derive supports of patterns with negation. The method is based on the observation that for any pattern X and any item x, the number of transactions in which X occurs is the sum of the number of transactions in which X occurs with x and the number of transactions in which X occurs without x. In other words, $sup(X) = sup(X \cup \{x\}) + sup(X \cup \{(-x)\})$, or $sup(X \cup \{(-x)\}) = sup(X) - sup(X \cup \{x\})$ (Mannila & Toivonen, 1996). Multiple usage of this property enables determination of the supports of patterns with an arbitrary number of negated items based on the supports of positive patterns. For example, the support of pattern $\{a(-b)(-c)\}$, which has two negated items, can be calculated as follows: $sup(\{a(-b)(-c)\}) = sup(\{a(-b)\}) - sup(\{a(-b)c\})$. Thus, the task of calculating the support of $\{a(-b)(-c)\}$, which has two negated items, becomes a task of calculating the supports of patterns $\{a(-b)\}$ and $\{a(-b)c\}$, each of which contains only one negated item. We note that $sup(\{a(-b)\}) = sup(\{a\}) - sup(\{ab\})$, and $sup(\{a(-b)c\}) = sup(\{ac\}) - sup(\{abc\})$. Eventually, we obtain: $sup(\{a(-b)(-c)\}) = sup(\{a\}) - sup(\{ab\}) - sup(\{ac\}) + sup(\{abc\})$. The support of $\{a(-b)(-c)\}$ is hence determinable from the supports of $\{abc\}$ and its proper subsets.

It was proved in Toivonen (1996) that for any pattern with negation its support is determinable from the supports of positive patterns. Nevertheless, the knowledge of the supports of only frequent patterns may be insufficient to derive the supports of all frequent patterns with negation (Boulicaut, Bykowski, & Jeudy, 2000), which we show beneath.

Let us try to calculate the support of pattern $\{bef(-h)\}$: $sup(\{bef(-h)\}) = sup(\{bef\}) - sup(\{befh\})$. Pattern $\{bef\}$ is frequent and its support equals 2 (see *Figure 1*). To the contrary, $\{befh\}$ is not frequent, so its support does not exceed *minSup*, which equals 1. Hence, $1 \le sup(\{bef(-h)\}) \le 2$. The obtained result is not sufficient to determine if $\{bef(-h)\}$ is frequent.

The problem of large amount of mined frequent patterns is widely recognized. Within the last five years, a number of lossless representations of frequent positive patterns have been proposed. Frequent closed itemsets were introduced in (Pasquier et al., 1999); the generators representation was introduced in (Kryszkiewicz, 2001). Other lossless representations are based on disjunction-free sets (Bykowski & Rigotti, 2001), disjunction-free generators (Kryszkiewicz, 2001), generalized disjunction-free generators (Kryszkiewicz & Gajek, 2002), generalized disjunction-free sets (Kryszkiewicz, 2003), non-derivable itemsets (Calders & Goethals, 2002), and *k*-free sets (Calders & Goethals, 2003). All these models allow distinguishing between frequent and infrequent positive patterns and enable determination of supports for all frequent positive patterns. Although the research on concise representations of frequent positive patterns is advanced, no model was offered in the literature to represent all frequent patterns with negation.

MAIN THRUST

We offer a *generalized disjunction-free literal set model* (GDFLR) as a concise lossless representation of all frequent positive patterns and all frequent patterns with negation. Without the need to access the database, GDFLR enables distinguishing between all frequent and infrequent patterns, and enables calculation of the supports for all frequent patterns.

GDFLR uses the mechanism of deriving supports of positive patterns that was proposed in Kryszkiewicz & Gajek (2002). Hence, we first recall this mechanism. Then we examine how to use it to derive the supports of patterns with negation and propose a respective naive representation of frequent patterns. Next we examine relationships