Reasoning about Frequent Patterns with Negation

Marzena Kryszkiewicz
Warsaw University of Technology, Poland

INTRODUCTION

Discovering of 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 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.

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.

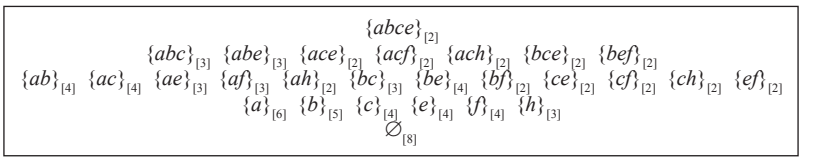
Aside from searching for only statistically significant sets of items, one may be interested in identifying frequent cases when purchase of some items (presence of some symptoms) excludes purchase of other items (presence of other symptoms). A pattern consisting of items $x_1, \ldots, x_m$ and negations of items $x_{m+1}, \ldots, x_n$ will be denoted by $\{x_1, \ldots, x_m, -x_{m+1}, \ldots, -x_n\}$. The *support of pattern* $\{x_1, \ldots, x_m, -x_{m+1}, \ldots, -x_n\}$ 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 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

Table 1. Sample database 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	$\{af\}$

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



distinct variations (including itself): $\{ab\}$, $\{a(-b)\}$, $\{(-a)b\}$, $\{(-a)(-b)\}$.

One can discover 109 frequent patterns in D, 27 of which are positive, and 82 of which have negated items. 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 and 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 positive patterns may be insufficient to derive the supports of all patterns with negation (Boulicaut, Bykowski & Jeudy, 2000), which we illustrate beneath.

Let us try to calculate the support of pattern $\{bef(-h)\}$: $sup(\{bef(-h)\}) = sup(\{bef\}) - sup(\{befh\})$. Pattern $\{bef\}$ is frequent (see Figure 1), so it is stored altogether with its support. To the contrary, $\{befh\}$ is not frequent, so the information about $\{befh\}$ and its support is not stored. Thus, we are unable to calculate the support of $\{bef(-h)\}$ based on the frequent positive patterns.

The problem of large amount of mined frequent patterns is widely recognized. Within the last decade, 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, there are few papers in the literature devoted to representing of all frequent patterns with negation.

MAIN THRUST OF THE CHAPTER

We define 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. We also define a *k-generalized disjunction-free literal set model* (*k*-GDFLR) as a modification of GDFLR for more concise lossless representing of all frequent positive patterns and all frequent patterns with at most k negated items.

Both representations may use 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 introduce a respective naive representation of frequent patterns. Next we examine relationships between specific patterns and supports of their variations. Eventually, we use the obtained results to offer GDFLR as a refined version of the naive model, and *k*-GFLDR as a generalization of GDFLR which coincides with GDFLR for $k = \infty$.

Reasoning about Positive Patterns Based on Generalized Disjunctive Patterns

Let us observe that whenever item a occurs in a transaction in database D, then item b, or f, or both also occur in the transaction. This fact related to pattern $\{abf\}$ can be expressed in the form of implication $a \Rightarrow b \vee f$. Now, without accessing the database, we can derive additional implications, such as $ac \Rightarrow b \vee f$ and $a \Rightarrow b \vee f \vee c$, which are related to supersets of $\{abf\}$. The knowledge of such implications can be used for calculating the supports of patterns they relate to. For example, $ac \Rightarrow b \vee f$ implies that the number of transactions in which $\{ac\}$ occurs equals the number of transactions in which $\{ac\}$ occurs with b plus the number of transactions in which $\{ac\}$ occurs with f minus the number of transactions in which $\{ac\}$ occurs both with b and f. In other words, $sup(\{ac\}) = sup(\{acb\}) + sup(\{acf\}) - sup(\{acbf\})$. Hence, $sup(\{abcf\}) = sup(\{abc\}) + sup(\{acf\}) - sup(\{ac\})$, which means that the support of pattern $\{abcf\}$ is determinable from the supports of its proper subsets. In general, if there is an implication related to a positive pattern, then the support of this pattern is derivable from the supports of its proper subsets (please, see (Kryszkiewicz & Gajek, 2002) for proof). If there is such an implication for a pattern, then the pattern is called a *generalized disjunctive set*. Otherwise, it is called a *generalized disjunction-free set*. We will present now a lossless *generalized disjunction-free set representation* (GDFSR) of all frequent positive patterns, which uses the discussed mechanism of deriving supports. The GDFSR representation is defined as consisting of the following components (Kryszkiewicz, 2003):

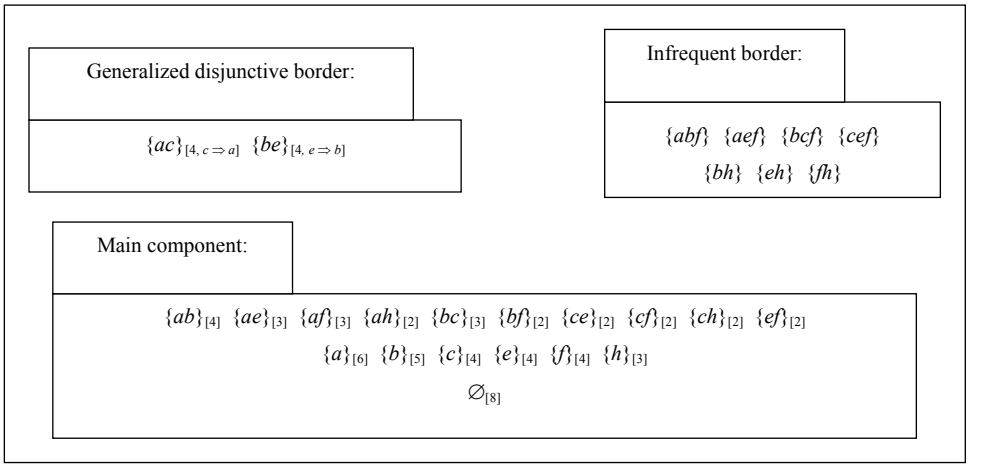
- The main component containing all frequent generalized disjunction-free positive patterns stored altogether with their supports;
- The infrequent border consisting of all infrequent positive patterns all proper subsets of which belong to the main component;
- The generalized disjunctive border consisting of all minimal frequent generalized disjunctive positive patterns stored altogether with their supports and/or respective implications.

Figure 2 depicts the GDFSR representation found in D. The main component consists of 17 elements, the infrequent border of 7 elements, and generalized disjunctive border of 2 elements.

Now, we will demonstrate how to use this representation for evaluating unknown positive patterns:

- Let us consider pattern $\{abcf\}$. We note that $\{abcf\}$ has a subset, e.g. $\{abf\}$, in the infrequent border. This means that all supersets of $\{abf\}$, in particular $\{abcf\}$, are infrequent.
- Let us consider pattern $\{abce\}$. It does not have any subset in the infrequent border, but has a subset, e.g. $\{ac\}$, in the generalized disjunctive border. Property $c \Rightarrow a$, associated with $\{ac\}$ implies property $bce \Rightarrow a$ related to $\{abce\}$. Hence, $sup(\{abce\}) = sup(\{bce\})$. Now, we need to determine the support of $\{bce\}$. We observe that $\{bce\}$ has subset $\{be\}$ in the generalized disjunctive border. Property $e \Rightarrow b$ associated with $\{be\}$ implies property $ce \Rightarrow b$ related to $\{bce\}$. Hence, $sup(\{bce\}) = sup(\{ce\})$. Pattern $\{ce\}$ belongs to the main component, so its support is known (here: equals 2). Summarizing, $sup(\{abce\}) = sup(\{bce\}) = sup(\{ce\}) = 2$.

Figure 2. The GDFSR representation found in D



Naive Approach to Reasoning About Patterns with Negation based on Generalized Disjunctive Patterns

One can easily note that implications, we were looking for positive patterns, may exist also for patterns with negation. For instance, looking at Table 1, we observe that whenever item *a* occurs in a transaction, then item *b* occurs in the transaction and/or item *e* is missing in the transaction. This fact related to pattern $\{ab(\neg e)\}$ can be expressed as implication $a \Rightarrow b \vee (\neg e)$. Hence, $sup(\{a\}) = sup(\{ab\}) + sup(\{a(\neg e)\}) - sup(\{ab(\neg e)\})$, or $sup(\{ab(\neg e)\}) = sup(\{ab\}) + sup(\{a(\neg e)\}) - sup(\{a\})$. Thus, the support of pattern $\{ab(\neg e)\}$ is determinable from the supports of its proper subsets. In general, the support of a generalized disjunctive pattern with any number of negated items is determinable from the supports of its proper subsets.

Having this in mind, we conclude that the GDFSR model can easily be adapted for representing all frequent patterns. We define a *generalized disjunction-free set representation of frequent patterns admitting negation* (GDFSRN) as holding all conditions that are held by GDFSR except for the condition restricting the representation's elements to positive patterns. GDFSRN discovered from database D consists of 113 elements. It contains both positive patterns and patterns with negation. For instance, $\{bc\}_{[3]}$, $\{b(\neg c)\}_{[2]}$, and $\{(\neg b)(\neg c)\}_{[2]}$, which are frequent generalized disjunction-free, are sample elements of the main component of this representation, whereas $\{a(\neg c)\}_{[2,\ \varnothing \Rightarrow a \vee (\neg c)]}$, which is a minimal frequent generalized disjunctive pattern, is a sample element of the generalized disjunctive border. Although conceptually straightforward, the representation is not concise, since its cardinality (113) is comparable with the number of all frequent patterns (109).

Generalized Disjunctive Patterns vs. Supports of Variations

Let us consider implication $a \Rightarrow b \vee f$, which holds in our database. The statement that whenever item *a* occurs in a transaction, then item *b* and/or item *f* also occurs in the transaction is equivalent to the statement that there is no transaction in which *a* occurs without both *b* and *f*. Therefore, we conclude that implication $a \Rightarrow b \vee f$ is equivalent to statement $sup(a(\neg b)(\neg f)) = 0$. We generalize this observation as follows:

$x_1 \ldots x_m \Rightarrow x_{m+1} \vee \ldots \vee x_n$ is equivalent to $sup(\{x_1, \ldots, x_m\} \cup \{\neg x_{m+1}, \ldots, \neg x_n\}) = 0$.

Let us recall that $x_1 \ldots x_m \Rightarrow x_{m+1} \vee \ldots \vee x_n$ implies that pattern $\{x_1, \ldots, x_n\}$ is generalized disjunctive, and $sup(\{x_1, \ldots, x_m\} \cup \{\neg x_{m+1}, \ldots, \neg x_n\}) = 0$ implies that pattern $\{x_1, \ldots, x_n\}$ has a variation different from itself that does not occur in any transaction. Hence, we infer that a positive pattern is generalized disjunctive if and only if it has a variation with negation the support of which equals 0.

Effective Approach to Reasoning About Patterns with Negation based on Generalized Disjunctive Patterns

In order to overcome the problem of possible small conciseness ratio of the GDFSRN model, we offer a new representation of frequent patterns with negation. Our intention is to store in the new representation at most one pattern for a number of patterns occurring in GDFSRN that have the same positive variation.

We define a *generalized disjunction-free literal representation* (GDFLR) as consisting of the following components:

- the main component containing each positive pattern (stored with its support) that has at least one frequent variation and all its variations have non-zero supports;
- the infrequent border containing each minimal positive pattern all variations of which are infrequent and all proper subsets of which belong to the main component;
- the generalized disjunctive border containing each minimal positive pattern (stored with its support and, eventually, implication) that has at least one frequent variation and at least one variation with zero support.

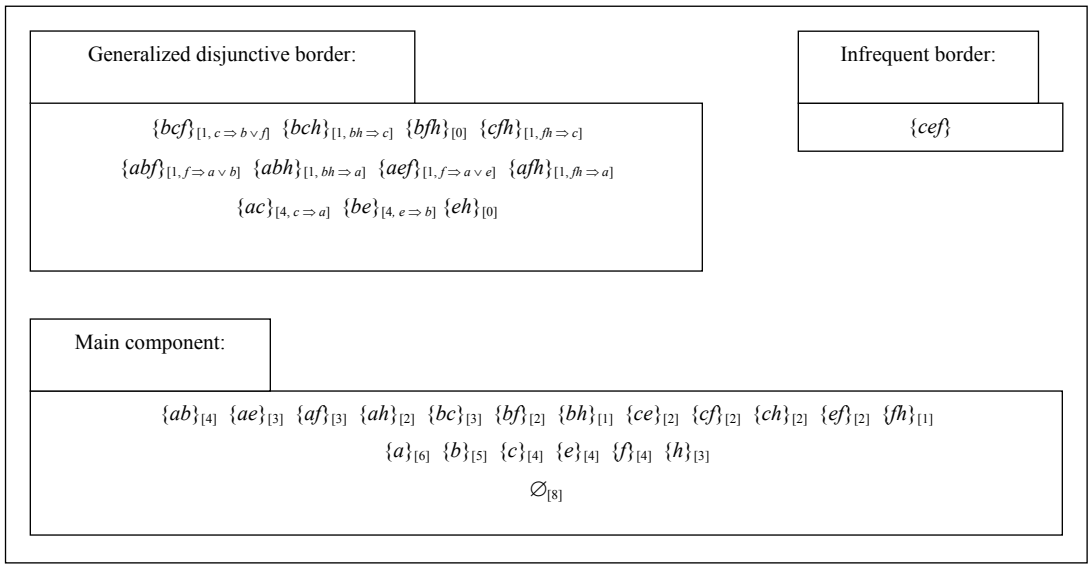
Please note that each element in the main component is generalized disjunction-free since all its variations have non-zero supports. On the other hand, each element in the generalized disjunctive border is generalized disjunctive or has support equal zero.

Figure 3 depicts GDFLR discovered in D. The main component consists of 19 elements, the infrequent border of 1 element, and generalized disjunctive border of 11 elements.

Now we will illustrate how to use this representation for evaluating unknown patterns:

- Let us consider pattern $\{a(-c)(-e)f\}$. We note that $\{acef\}$, which is a positive variation of the evaluated pattern, has subset $\{cef\}$ in the infrequent border. This means that both $\{cef\}$ and all its variations including $\{(-c)(-e)f\}$ are infrequent. Hence, $\{a(-c)(-e)f\}$, which is a superset of $\{(-c)(-e)f\}$, is also infrequent.
- Let us consider pattern $\{bef(-h)\}$. The positive variation $\{befh\}$ of $\{bef(-h)\}$ does not have any subset in the infrequent border, so $\{bef(-h)\}$ has a chance to be frequent. Since, $sup(\{bef(-h)\}) = sup(\{bef\}) - sup(\{befh\})$, we need to determine the supports of two positive patterns $\{bef\}$ and $\{befh\}$. $\{bef\}$ has subset $\{be\}$ in the generalized disjunctive border, the implication of which is $e \Rightarrow b$. Hence, $ef \Rightarrow b$ is an implication for $\{bef\}$.

Figure 3. The GDFLR representation found in D



Thus, $sup(bef) = sup(ef) = 2$ (please, see the main component for pattern $\{ef\}$). Pattern $\{befh\}$ also has a subset, e.g. $\{eh\}$, in the generalized disjunctive border. Since $sup(\{eh\}) = 0$, then $sup(\{befh\})$ equals 0 too. Summarizing, $sup(\{bef(-h)\}) = 2 - 0 = 2$, and thus $\{bef(-h)\}$ is a frequent pattern.

GDFLR is a lossless representation of all frequent patterns. It can be proved that a pessimistic estimation of the length of a longest element in GDFLR depends logarithmically on the number of records in the database. A formal presentation of this model and its properties, as well as an algorithm for its discovery and experimental results can be found in (Kryszkiewicz, 2004b). The experiments carried out on real large data sets show that GDFLR is by several orders of magnitude more concise than all frequent patterns. Further reduction of GDFLR (and GDFSRN) can be achieved by applying techniques for reducing borders (Calders & Goethals, 2003; Kryszkiewicz, 2003; Kryszkiewicz, 2004a) or a main component (Kryszkiewicz, 2004c).

In (Kryszkiewicz & Cichon, 2005), we discuss the complexity of evaluating candidate elements of the representation. We observe that the calculation of the support of a pattern with n negated items from the supports of positive patterns requires the knowledge of the support of the positive variant P of that pattern and the supports of $2^n - 1$ proper subsets of pattern P. In order to alleviate this problem, we offer a special ordering of candidate elements in (Kryszkiewicz & Cichon, 2005). The introduced ordering enables calculation of the support of a pattern with n negated items as the difference of the supports of only two patterns, possibly with negation, the supports of which were calculated earlier. The proposed algorithm using this method of calculating supports performs much faster (by up to two orders of magnitude for low support threshold values). Nevertheless, in each iteration l, it requires storing all variants of all candidates of length l and all variants of all elements of the main component of length l-1.

Reasoning about Patterns with at Most k-Negated Items

An important part of data mining is discovering patterns conforming user-specified constraints. It is natural to expect that the derivation of a part of all frequent patterns instead of all of them should take less time and should produce less number of patterns than unrestricted data mining. One can also anticipate that a representation of a part of all frequent patterns should be more concise than the representation of all frequent patterns. In this section, we define a generalized disjunction-free literal representation of frequent patterns with at most k negated items (k-GDFLR). The new representation consists of the following components (Kryszkiewicz, 2006):

- the main component containing each positive pattern (stored with its support) that has at least one frequent variation with at most k negated items and is neither generalized disjunctive nor has support equal 0;
- the infrequent border containing each positive pattern of which all variations with at most k negated items are infrequent and all proper subsets belong to the main component;
- the generalized disjunctive border containing each minimal positive pattern (stored with its support and/or implication) that has at least one frequent variation with at most k negated items and at least one variation with zero support.

Please note that for $k = 0$ (no negation allowed) k-GDFLR represents all frequent positive patterns, whereas for $k = \infty$ (unrestricted number of negations) it represents all frequent patterns: both positive ones and with negation.

It was proved in (Kryszkiewicz, 2006) that the k-GDFLR representation losslessly represents all frequent patterns with at most k negated items. The conducted experiments show that k-GDFLR is more concise and faster computable than GDFLR especially for low support values and low k.

FUTURE TRENDS

Development of different representations of frequent patterns with negation and algorithms for their discovery can be considered as a short-term trend. As a long-term trend, we envisage development of representations of patterns satisfying user imposed constraints and representations of other kinds of knowledge admitting negation, such as association rules, episodes, sequential patterns and classifiers. Such research should stimulate positively the development of inductive databases, where queries including negation are common.

CONCLUSION

The set of all positive patterns can be treated as a lossless representation of all frequent patterns, nevertheless it is not concise. On the other hand, the set of all frequent positive patterns neither guarantees derivation of all frequent patterns with negation, nor is concise in practice. The GDFSRN and GDFLR representations, we proposed, are first lossless representations of both all frequent positive patterns and patterns with negation. GDFLR consists of a subset of only positive patterns and hence is more concise than analogous GDFSRN, which admits the storage of many patterns having the same positive variation. We have proposed the *k*-GDFLR representation for even more concise, lossless representing of all frequent patterns with at most *k* negated items. In (Kryszkiewicz & Cichon, 2005), we have offered a method for fast calculation of the supports of the candidate elements of the generalized disjunction-free representations, which is based on a special ordering of the candidate patterns.

REFERENCES

Agrawal, R., Imielinski, R., & Swami, A. N. (1993, May). Mining association rules between sets of items in large databases. *ACM SIGMOD International Conference on Management of Data*, Washington, USA, 207-216.

Boulicaut, J.-F., Bykowski, A., & Jeudy, B. (2000, October). Towards the tractable discovery of association rules with negations. *International Conference on Flexible Query Answering Systems*, FQAS'00, Warsaw, Poland, 425-434.

Bykowski, A., & Rigotti, C. (2001, May). A condensed representation to find patterns. *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS'01, Santa Barbara, USA, 267–273.

Calders, T., & Goethals, B. (2002, August). Mining all non-derivable frequent itemsets. *European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD'02, Helsinki, Finland, 74-85.

Calders, T., & Goethals, B. (2003, September). Minimal k-free representations. *European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD'03, Cavtat-Dubrovnik, Croatia, 71-82.

Kryszkiewicz, M. (2001, November-December). Concise representation of frequent patterns based on disjunction–free generators. *IEEE International Conference on Data Mining*, ICDM'01, San Jose, USA, 305–312.

Kryszkiewicz, M., (2003, July). Reducing infrequent borders of downward complete representations of frequent patterns. *Symposium on Databases, Data Warehousing and Knowledge Discovery*, DDWKD'03, Baden-Baden, Germany, 29–42.

Kryszkiewicz, M. (2004a, March). Reducing borders of *k*-disjunction free representations of frequent patterns. *ACM Symposium on Applied Computing*, SAC'04, Nikosia, Cyprus, 559-563.

Kryszkiewicz, M. (2004b, May). Generalized disjunction-free representation of frequent patterns with negation. ICS Research Report 9, Warsaw University of Technology; extended version accepted to *Journal of Experimental and Theoretical Artificial Intelligence*.

Kryszkiewicz, M. (2004c, July). Reducing main components of *k*-disjunction free representations of frequent patterns. *International Conference in Information Processing and Management of Uncertainty in Knowledge-Based Systems*, IPMU'04, Perugia, Italy, 1751-1758.

Kryszkiewicz, M. (2006, April). Generalized Disjunction-Free Representation of Frequents Patterns with at Most *k* Negations. *Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference*, PAKDD'06, Singapore, 468-472.

Kryszkiewicz, M., & Cichon, K. (2005, May). Support Oriented Discovery of Generalized Disjunction-Free Representation of Frequent Patterns with Negation. *Advances in Knowledge Discovery and Data Mining*, PAKDD'05, Hanoi, Vietnam, 672-682

Kryszkiewicz, M., & Gajek, M. (2002, May). Concise representation of frequent patterns based on generalized disjunction–free generators. *Advances in Knowledge Discovery and Data Mining*, *Pacific-Asia Conference*, PAKDD'02, Taipei, Taiwan, 159–171.

Mannila, H., & Toivonen, H. (1996, August). Multiple uses of frequent sets and condensed representations. *International Conference on Knowledge Discovery and Data Mining*, KDD'96, Portland, USA, 189-194.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999, January). Discovering frequent closed itemsets for association rules. *Database Theory*, *International Conference*, ICDT'99, Jerusalem, Israel, 398–416.

Toivonen, H. (1996). Discovery of frequent patterns in large data collections. Ph.D. Thesis, Report A-1996-5, University of Helsinki.

Tsumoto, S. (2002). Discovery of positive and negative knowledge in medical databases using rough sets. In S. & A. Shinohara (eds) *Progress in Discovery Science*, Springer, Heidelberg, 543-552.

KEY TERMS

Frequent Pattern: Pattern the support of which exceeds a user-specified threshold.

Generalized Disjunction-Free Pattern: Pattern the support of which is not determinable from the supports of its proper subsets.

Generalized Disjunctive Pattern: Pattern the support of which is determinable from the supports of its proper subsets.

Item: 1) sales product; 2) feature, attribute.

Literal: An item or negated item.

Lossless Representation of Frequent Patterns: Fraction of patterns sufficient to distinguish between frequent and infrequent patterns and to determine the supports of frequent patterns.

Pattern with Negation: Pattern containing at least one negated item.

Positive Pattern: Pattern with no negated item.

Reasoning about Patterns: Deriving supports of patterns without accessing a database.

Support of a Pattern: The number of database transactions in which the pattern occurs.