

# Separating Structure from Interestingness

Taneli Mielikäinen

HIIT Basic Research Unit  
Department of Computer Science  
University of Helsinki, Finland  
`Taneli.Mielikainen@cs.Helsinki.FI`

**Abstract.** Condensed representations of pattern collections have been recognized to be important building blocks of inductive databases, a promising theoretical framework for data mining, and recently they have been studied actively. However, there has not been much research on how condensed representations should actually be represented.

In this paper we propose a general approach to build condensed representations of pattern collections. The approach is based on separating the structure of the pattern collection from the interestingness values of the patterns. We study also the concrete case of representing the frequent sets and their (approximate) frequencies following this approach: we discuss the trade-offs in representing the frequent sets by the maximal frequent sets, the minimal infrequent sets and their combinations, and investigate the problem approximating the frequencies from samples by giving new upper bounds on sample complexity based on frequent closed sets and describing how convex optimization can be used to improve and score the obtained samples.

## 1 Introduction

*Data mining* aims to find something interesting from large databases. One of the most important approaches to mine data is *pattern discovery* where the goal is to extract *interesting patterns* (possibly with some *interestingness values* associated to each of them) from data [1,2]. The most prominent example of pattern discovery is the *frequent set mining* problem:

*Problem 1 (Frequent set mining).* Given a multiset  $d = \{d_1, \dots, d_n\}$  (a *data set*) of subsets (*transactions*) of a set  $R$  of *attributes* and a threshold value  $\sigma \in [0, 1]$ , find the collection  $\mathcal{F}(\sigma, d) = \{X \subseteq R : fr(X, d) \geq \sigma\}$  where  $fr(X, d) = |cover(X, d)|/n$  and  $cover(X, d) = \{i : X \subseteq d_i, 1 \leq i \leq n\}$ .

The set collection  $\mathcal{F}(\sigma, d)$  is called the *collection of  $\sigma$ -frequent sets in  $d$* .

There exist techniques to efficiently compute frequent sets, see e.g. [3]. A major advantage of frequent sets is that they can be computed from data without much domain knowledge: The data set determines an empirical joint probability distribution over the attribute combinations and high marginal probabilities of the joint probability distribution can be considered as a reasonable way to

summarize the joint probability distribution. (Note that also a sample from the joint probability distribution is a quite good summary.) However, this generality causes a major problem of frequent sets: the frequent set collections that describe data well tend to be large. Although the computations could be done efficiently enough, it is not certain that the huge collection of frequent sets is very concise summary of the data.

This problem of too large frequent set collections have been tried to solve by computing a small irredundant subcollection of the given frequent set collection such that the subcollection determines the frequent set collection completely. Such subcollections are usually called *condensed representations* of the frequent set collection [4]. The condensed representations of the frequent sets have been recognized to have an important role in *inductive databases* which seems to be a promising theoretical framework for data mining [5,6,7,8].

The condensed representations of frequent sets have been studied actively lately and several condensed representations, such as *maximal sets* [9], *closed sets* [10], *free sets* [11], *disjunction-free sets* [12], *disjunction-free generators* [13], *non-derivable itemsets* [14], *condensed pattern bases* [15], *pattern orderings* [16] and *pattern chains* [17], have been proposed. However, not much has been done on how the condensed representations should actually be represented although it is an important question both for the computational efficiency and for the effectiveness of the data analyst.

In this paper we investigate how the patterns and their interestingness values can be represented separately. In particular, we study how to represent frequent sets and their frequencies: we discuss how the collection of frequent sets can be described concisely by combinations of its maximal frequent and minimal infrequent sets, show that already reasonably small samples determine the frequencies of the frequent sets accurately and describe how a weighting of the sample can further improve the frequency estimates.

The paper is organized as follows. In Section 2 we argue why describing patterns and their interestingness values separately makes sense. In Section 3 we study a representation of interestingness values based on random samples of data and give sample complexity bounds that are sometimes considerably better than the bounds given in [18]. In Section 4 we describe how the samples can be weighted to optimally approximate the frequencies of the set collection w.r.t. a wide variety of loss functions and show experimentally that a considerable decrease of loss can be achieved. The work is concluded in Section 5.

## 2 Separating Patterns and Interestingness

Virtually all condensed representations of pattern collections represent the collection by listing a subcollection of the interesting patterns. This approach to condensed representations has the desirable closure property that also a subcollection of (irredundant) patterns is a collection of patterns. Another advantage of many condensed representations of pattern collections consisting of a list of irredundant patterns is that the patterns in the original collection and their inter-

estingness values can be inferred conceptually very easily from the subcollection of irredundant patterns and their interestingness values.

The most well-known examples of condensed representations of pattern collections are the collections of *closed  $\sigma$ -frequent sets*:

**Definition 1 (Closed  $\sigma$ -frequent sets).** *A  $\sigma$ -frequent set  $X \in \mathcal{F}(\sigma, d)$  is closed if and only if  $fr(X, d) > fr(X \cup \{A\}, d)$  for all  $A \in R \setminus X$ .*

*The collection of closed  $\sigma$ -frequent sets is denoted by  $\mathcal{C}(\sigma, d)$ .*

The collection  $\mathcal{C}(\sigma, d)$  consists of *closures* of the sets in  $\mathcal{F}(\sigma, d)$ , i.e., the sets  $X \in \mathcal{F}(\sigma, d)$  such that  $X = cl(X, d)$  where

$$cl(X, d) = \arg \max \{fr(Y) : Y \supseteq X, Y \in \mathcal{C}(\sigma, d)\}.$$

Clearly, the frequency of the set  $X \in \mathcal{F}(\sigma, d)$  is the maximum of the frequencies of the closed frequent supersets of  $X$ , i.e., the frequency of its closure

$$fr(X, d) = fr(cl(X, d), d) = \max \{fr(Y) : Y \supseteq X, Y \in \mathcal{C}(\sigma, d)\}.$$

Although there are many positive aspects on representing the pattern collection and their interestingness values by an irredundant subset of the pattern collection, there are some benefits achievable by separating the structure of the collection from the interestingness values. For example, the patterns alone can always be represented at least as compactly as (and most of the time much more compactly than) the same patterns with their interestingness values. As a concrete example, the collection of frequent sets and the collection of closed frequent sets can be represented by their subcollection of maximal frequent sets:

**Definition 2 (Maximal  $\sigma$ -frequent sets).** *A  $\sigma$ -frequent set  $X \in \mathcal{F}(\sigma, d)$  is maximal if and only if  $Y \supseteq X, Y \in \mathcal{F}(\sigma, d) \Rightarrow X = Y$ .*

*The collection of maximal  $\sigma$ -frequent sets is denoted by  $\mathcal{M}(\sigma, d)$ .*

Clearly, the frequent sets can be determined using the maximal frequent sets: a set  $X \subseteq R$  is in the collection  $\mathcal{F}(\sigma, d)$  if and only if it is a subset of some set in  $\mathcal{M}(\sigma, d)$ .

The collection of maximal frequent sets represents the collection of (closed) frequent sets quite compactly since  $|\mathcal{M}(\sigma, d)|$  is never larger than  $|\mathcal{C}(\sigma, d)|$  but it can be exponentially smaller than the corresponding closed frequent set collection (and thus the frequent set collection, too): Let the data set  $d$  consist of the sets  $R \setminus \{A\}$  s.t.  $A \in R$ . Then the collection of (closed)  $(1/n)$ -frequent sets consists of all subsets of  $R$  except  $R$  itself but the collection of maximal frequent sets consists only of the sets  $R \setminus \{A\}, A \in R$ . Then  $|\mathcal{F}(\sigma, d)| / |\mathcal{M}(\sigma, d)| = |\mathcal{C}(\sigma, d)| / |\mathcal{M}(\sigma, d)| > 2^{|R|} / |R|$ . Also, the only case when the number of maximal frequent sets is equal to the number of frequent sets is when the collection of frequent sets consists solely of singleton sets, i.e., frequent items.

In addition to the maximal frequent sets, the frequent sets can be described also by the minimal infrequent sets:

**Definition 3 (Minimal  $\sigma$ -infrequent sets).** A  $\sigma$ -infrequent set  $X \in 2^R \setminus \mathcal{F}(\sigma, d)$  is minimal if and only if  $Y \subseteq X, Y \in 2^R \setminus \mathcal{F}(\sigma, d) \Rightarrow X = Y$ .

The collection of minimal  $\sigma$ -infrequent sets is denoted by  $\mathcal{I}(\sigma, d)$ .

Again, obtaining the frequent sets from this representation is straightforward: a set  $X \subseteq R$  is  $\sigma$ -frequent if and only if it does not contain any set  $Y \in \mathcal{I}(\sigma, d)$ .

Minimal infrequent sets for a given collection of frequent sets can be computed quite efficiently since the minimal infrequent sets are the minimal transversals in the complements of the maximal frequent sets [9].

It is not immediate which of the representations – the maximal frequent sets or the minimal infrequent sets – is smaller even in terms of the number of sets: the number  $|\mathcal{M}(\sigma, d)|$  of maximal  $\sigma$ -frequent sets is bounded by  $(|R| - \sigma n + 1) |\mathcal{I}(\sigma, d)|$  (unless the collection  $\mathcal{I}(\sigma, d)$  minimal  $\sigma$ -infrequent sets is empty) but  $|\mathcal{I}(\sigma, d)|$  cannot be bounded by a polynomial in  $|\mathcal{M}(\sigma, d)|$ , in  $|R|$  and in  $n$  (i.e., the number of transactions in  $d$ ) [19].

In practice, the representation for a given collection of frequent sets can be chosen to be the one that is smaller for that particular collection. Furthermore, instead of choosing either the maximal frequent sets or the minimal infrequent sets, it is possible to choose a subcollection determining the collection of frequent sets uniquely that contains sets from both collections:

*Problem 2 (Smallest representation of frequent sets).* Given collections  $\mathcal{M}(\sigma, d)$  and  $\mathcal{I}(\sigma, d)$ , find the smallest subset  $\mathcal{T}$  of  $\mathcal{M}(\sigma, d) \cup \mathcal{I}(\sigma, d)$  that uniquely determines  $\mathcal{F}(\sigma, d)$ .

By definition, all subsets of minimal infrequent sets are frequent and all supersets of maximal frequent sets. Thus, Problem 2 can be modeled as a minimum weight set cover problem:

*Problem 3 (Minimum weight set cover [20]).* Given a collection  $\mathcal{S}$  of subsets of a finite set  $S$  and a weight function  $w : \mathcal{S} \rightarrow \mathbb{R}$ , find a subcollection  $\mathcal{S}'$  of  $\mathcal{S}$  with the smallest weight  $w(\mathcal{S}') = \sum_{X \in \mathcal{S}'} w(X)$ .

The minimum weight set cover problem is approximable within a factor  $1 + \ln |S|$  [20].

The set cover instance corresponding to the case of representing the frequent sets by a subcollection of  $\mathcal{M}(\sigma, d) \cup \mathcal{I}(\sigma, d)$  is the following. The set  $S$  is equal to  $\mathcal{M}(\sigma, d) \cup \mathcal{I}(\sigma, d)$ . The set collection  $\mathcal{S}$  consists of the set  $S_X = \{Y \in \mathcal{M}(\sigma, d) \cup \mathcal{I}(\sigma, d) : X \subseteq Y \vee X \supseteq Y\}$  for each  $X \in \mathcal{M}(\sigma, d) \cup \mathcal{I}(\sigma, d)$ . Clearly, the solution  $\mathcal{T} \subset \mathcal{M}(\sigma, d) \cup \mathcal{I}(\sigma, d)$  determines the collection  $\mathcal{F}(\sigma, d)$  uniquely if and only if the solution  $\mathcal{S}'$  for the corresponding set cover instance covers  $S$ . Furthermore,  $w(\mathcal{S}') = |\mathcal{T}|$  when  $w(X) = 1$  for all  $X \in \mathcal{S}$ . Due to this reduction and the approximability of Problem 3 we get the following result:

**Theorem 1.** *Problem 2 is approximable within a factor  $1 + \ln |\mathcal{M}(\sigma, d) + \mathcal{I}(\sigma, d)|$ .*

A very interesting variant of Problem 2 is the case when user can interactively determine which attributes or frequent sets must be represented. This can be modeled as a minimum set cover problem, too. In the case when user only

adds attributes or frequent sets that must be represented this problem can be approximated almost as well as Problem 2 [21].

In addition to knowing which sets are frequent (or, in general, which patterns are interesting), it is usually desirable to know also how frequent each of the frequent sets is. One approach to describe the frequencies of the sets (approximately) correctly is to construct a small representative data set  $d'$  from the original data set  $d$  [22,23]. One computationally efficient and very flexible way to do this to obtain a random sample from the data set.

In the next two sections we study this approach. In Section 3 we give upper bounds on how many transactions chosen randomly from  $d$  suffice to give good approximations for the frequencies of all frequent sets simultaneously with high probability. The bounds can be computed from closed frequent sets which might be beneficial when each randomly chosen transaction is very expensive or the cost of the sample should be bounded above in advance. In Section 4 we show how the frequency estimates computed from the sample can be significantly improved by weighting the transactions using convex programming.

### 3 Sample Complexity Bounds

If the collection  $\mathcal{F}(\sigma, d)$  is known then it can be shown by a simple application of Chernoff bounds that for a sample  $d'$  of at least  $\ln(2|\mathcal{F}|/\Delta)/2\epsilon^2$  transactions, the absolute error of the frequency estimates for all sets in the collection is at most  $\epsilon$  with probability at least  $1 - \Delta$  [18].

However, the bounds given in [18] can be improved significantly since the frequent set collections have structure that can be useful when estimating the sufficient size for the sample. If the set collection  $\mathcal{F}$  is known, the data set  $d$  itself is usually a compact representation of the covers of the sets in  $\mathcal{F}$ . The goal of sampling is to choose a small set of transactions that accurately determine the sizes of the covers for each set in  $\mathcal{F}$ . This kind of sample (also for arbitrary set collections in addition to the collections of covers) is called an  $\epsilon$ -approximation [24]. The definition of  $\epsilon$ -approximation can be expressed for covers of a set collection as follows:

**Definition 4 ( $\epsilon$ -approximation).** *A finite subset  $T \subseteq [n] = \{1, \dots, n\}$  is an  $\epsilon$ -approximation for the set collection  $\mathcal{F}$  w.r.t.  $d$  if we have, for all  $X \in \mathcal{F}$*

$$\left| \frac{|T \cap \text{cover}(X, d)|}{|T|} - \frac{|\text{cover}(X, d)|}{n} \right| \leq \epsilon.$$

The sample complexity bound given in [18] is essentially optimal in the general case. However, we can obtain considerably better bounds if we look at the structure of the set collection. One of such structural properties is the VC-dimension of the collection:

**Definition 5 (VC-dimension).** *The VC-dimension  $VC(\text{cover}(\mathcal{F}, d))$  of the set collection  $\text{cover}(\mathcal{F}, d) = \{\text{cover}(X, d) : X \in \mathcal{F}\}$  is*

$$VC(\text{cover}(\mathcal{F}, d)) = \max \left\{ |T| : |\{T \cap \text{cover}(X, d) : X \in \mathcal{F}\}| = 2^{|T|} \right\}.$$

Given the  $VC$ -dimension of the collection of covers, the number of transactions that form an  $\epsilon$ -approximation for the covers can be bounded above by the following lemma adapted from the corresponding result for arbitrary set collections [24]:

**Lemma 1 (VC-dimension bound).** *Let  $\text{cover}(\mathcal{F}, d)$  be a set system of VC-dimension at most  $k$ , and let  $\epsilon \leq 1/2$ . Then there exists an  $\epsilon$ -approximation for  $\mathcal{S}$  of size at most  $\mathcal{O}(k \log(1/\epsilon)) / \epsilon^2$*

Actually, in addition to mere existence of small  $\epsilon$ -approximations, an  $\epsilon$ -approximation of size given by the  $VC$ -dimension bound can be found efficiently by random sampling [25].

One upper bound for the  $VC$ -dimension of a frequent set collection can be obtained from the number of closed frequent sets:

**Theorem 2.** *The VC-dimension of the collection  $\text{cover}(\mathcal{F}(\sigma, d), d)$  is at most  $\log |\mathcal{C}(\sigma, d)|$  where  $\mathcal{C}(\sigma, d)$  is the collection of closed frequent sets in  $\mathcal{F}(\sigma, d)$ . This bound is tight in the worst case.*

*Proof.* The  $VC$ -dimension of  $\text{cover}(\mathcal{F}(\sigma, d), d)$  is at most  $\log |\text{cover}(\mathcal{F}(\sigma, d), d)|$ . Clearly,  $|\text{cover}(\mathcal{C}(\sigma, d), d)| \leq |\mathcal{C}(\sigma, d)|$ . Thus it suffices to show that  $\text{cover}(X, d) = \text{cover}(cl(X), d)$ . However, this is immediate since, by definition,  $cl(X, d)$  is contained in each transaction of  $d$  which contains  $X$ .

The  $VC$ -dimension  $\log |\text{cover}(\mathcal{C}(\sigma, d), d)|$  is achieved by the collection of 0-frequent sets for a data set determined by  $n$  attributes  $1, 2, \dots, n$  with  $\text{cover}(i, d) = [n] \setminus \{i\}$  for each  $i \in [n]$ . The  $\mathcal{F}(\sigma, d)$  (and  $\mathcal{C}(\sigma, d)$ , too) consists of all subsets of  $[n]$ .  $\square$

The cardinality of  $\mathcal{C}(\sigma, d)$  can be estimated accurately by checking for random subset of frequent sets, how many of them are closed in  $d$ . Theorem 2 together with Lemma 1 imply the following upper bound:

**Corollary 1.** *For the set collection  $\mathcal{F}(\sigma, d)$ , there is an  $\epsilon$ -approximation of  $\mathcal{O}(\log |\mathcal{C}(\sigma, d)| \log(1/\epsilon)) / \epsilon^2$  transactions.*

These bounds are quite general upper bounds neglecting many fine details of the frequent set collections and thus usually smaller samples give approximations with required quality. It is probable that the bounds could be improved by taking into account the actual frequencies of the frequent sets. A straightforward way to improve the upper bounds is to bound the  $VC$ -dimension more tightly. In practice, the sampling can be stopped right after the largest absolute error between the frequency estimates and the correct frequencies is at most  $\epsilon$ .

## 4 Optimizing Sample Weights

Given a random subset  $d'$  of transaction in  $d$ , the frequencies in  $d$  for the sets in a given set collection  $\mathcal{F}$  can be estimated from  $d'$ . If also the correct frequencies are known, the error of the estimates can be measured.

Furthermore, the transactions in the sample can be weighted (in principle) in such a way that the frequency estimates are as good as possible. Let us denote the sample from  $d = \{d_1, \dots, d_n\}$  by  $d' = \{d'_1, \dots, d'_{n'}\}$  and let  $w_1, \dots, w_{n'}$  denote the weights of  $d'_1, \dots, d'_{n'}$ . The frequency of  $X$  in the weighted sample  $d'$  is the sum of the weights  $w_i$  of  $d'_i$ s containing  $X$ . Note that due to the weights we can assume that  $d'$  is a set although  $d$  is a multiset. Furthermore, we assume that all weights are nonnegative, i.e., no transaction in  $d'$  can be an anti-transaction.

The search for optimal weights w.r.t. a given loss function  $\ell$  can be formulated as an optimization task as follows:

$$\begin{aligned} & \text{minimize} && \ell(\mathcal{F}, d, d', w) \\ & \text{subject to} && w_i \geq 0, i = 1, \dots, n'. \end{aligned}$$

The weight  $w_i$  can be interpreted as a measure how representative the transaction  $d'_i$  is in the sample  $d'$  and thus the weights can be used also to score the transactions: On one hand the transactions with significantly larger weights than the average weight can be considered as very good representatives of the data. On the other hand the distribution of the weights tells about the skewness of the sample (and possibly also the skewness of the set collection  $\mathcal{F}$ ) w.r.t.  $d$ . If the loss function  $\ell$  is convex then the optimization task can be solved optimally in (weakly) polynomial time [26]. The most well-known examples of convex loss functions are  $L_p$  distances  $\left(\sum_{X \in \mathcal{F}} \left| fr(X, d) - \sum_{X \subseteq d'_i \in d'} w_i \right|^p\right)^{1/p}$ .

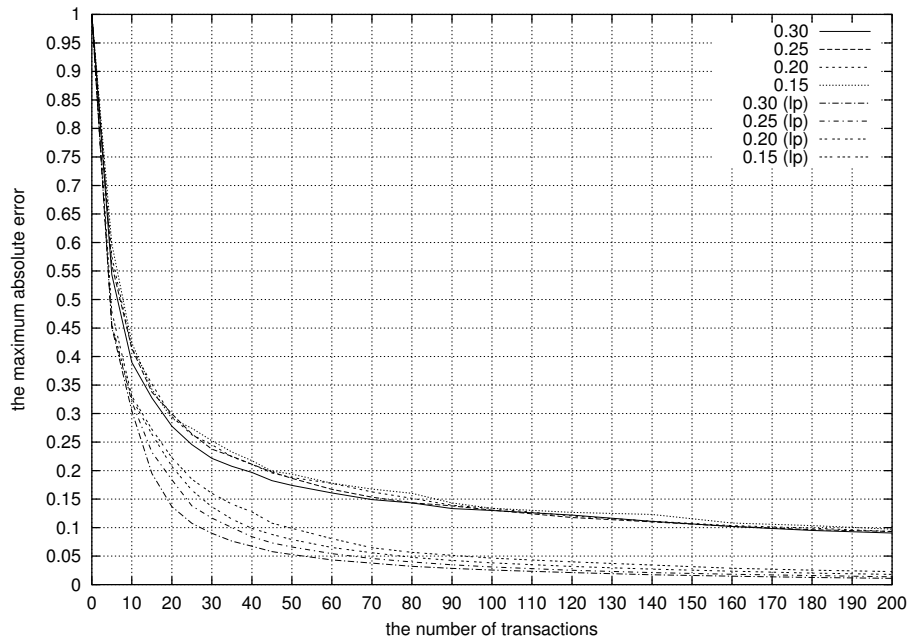
As a concrete example, let us consider the problem of minimizing the maximum absolute error of the frequency estimates. This variation of the problem can be formulated as a linear optimization task which can be solved even faster than the general (convex) optimization task:

$$\begin{aligned} & \text{minimize} && \epsilon \\ & \text{subject to} && \epsilon \geq fr(X, d) - \sum_{X \subseteq d'_i, i \in [n']} w_i \\ & && \epsilon \geq \sum_{X \subseteq d'_i, i \in [n']} w_i - fr(X, d) \\ & && w_i \geq 0, i = 1, \dots, n' \end{aligned}$$

The instance consists of  $n' + 1$  variables and  $2|\mathcal{F}(\sigma, d)| + n'$  inequalities. The number of inequalities in the instance can be further reduced to  $2|\mathcal{C}(\sigma, d)| + |S|$  by recognizing that if  $cl(X, d) = cl(Y, d)$  then the corresponding inequalities are equal, i.e., it is enough to have the inequalities for the collection  $\mathcal{C}(\sigma, d)$  of closed frequent sets. This can lead to exponential speed-ups.

To evaluate the usability of weighting random samples of transactions, we experimented with the random sampling of transactions (without replacements) and the linear programming refinement using the IPUMS Census data set from UCI KDD Repository:<sup>1</sup> IPUMS Census data set consists of 88443 transactions and 39954 attributes. The randomized experiment were repeated 80 times.

<sup>1</sup> <http://kdd.ics.uci.edu>



**Fig. 1.** IPUMS Census data

The results are shown in Figure 1. The labels of the curves correspond to the minimum frequency thresholds and the curves with (lp) are the corresponding linear programming refinements. The results show that already a very small number of transactions, when properly weighted, suffice to give good approximations for the frequencies of the frequent sets. Similar results were obtained in our preliminary experiments with other data sets.

Note that although it would be desirable to choose the transactions that determine the frequencies best instead of random sample, this might not be easy since finding the best transactions with optimal weights resembles the cardinality constrained knapsack problem that is known to be difficult [27].

## 5 Conclusions

In this paper we studied how to separate the descriptions of patterns and their interestingness values. In particular, we described how frequent sets can be described in a small space by the maximal frequent sets, the minimal infrequent sets, or their combination. Also, we studied how samples can be used to describe frequencies of the frequent sets: we gave upper bounds for the sample complexity using closed frequent sets and described a practical convex optimization approach for weighting the transactions in a given sample.

Representing pattern collections and their interestingness values separately seems to offer some benefits in terms of understandability, size and efficiency. There are several interesting open problems related to this separation of the structure and the interestingness:

- What kind of patterns and their interestingness values can be efficiently and effectively described by a sample from the data?
- How a small data set that represents the frequencies of frequent sets can be found efficiently?
- What kind of weightings are of interest to determine the costs for the maximal frequent sets and the minimal infrequent sets?
- How the representation of the patterns guides the knowledge discovery process?

## References

1. Hand, D.J.: Pattern detection and discovery. In Hand, D., Adams, N., Bolton, R., eds.: *Pattern Detection and Discovery*. Volume 2447 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag (2002) 1–12
2. Mannila, H.: Local and global methods in data mining: Basic techniques and open problems. In Widmayer, P., Triguero, F., Morales, R., Hennessy, M., Eidenbenz, S., Conejo, R., eds.: *Automata, Languages and Programming*. Volume 2380 of *Lecture Notes in Computer Science*, Springer-Verlag (2002) 57–68
3. Goethals, B., Zaki, M.J., eds.: *Proceedings of the Workshop on Frequent Itemset Mining Implementations (FIMI-03)*, Melbourne Florida, USA, November 19, 2003. Volume 90 of *CEUR Workshop Proceedings*. (2003) <http://CEUR-WS.org/Vol-90/>.
4. Mannila, H., Toivonen, H.: Multiple uses of frequent sets and condensed representations. In Simoudis, E., Han, J., Fayyad, U.M., eds.: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, AAAI Press (1996) 189–194
5. De Raedt, L.: A perspective on inductive databases. *SIGKDD Explorations* **4** (2003) 69–77
6. Imielinski, T., Mannila, H.: A database perspective on knowledge discovery. *Communications of The ACM* **39** (1996) 58–64
7. Mannila, H.: Inductive databases and condensed representations for data mining. In Maluszynski, J., ed.: *Logic Programming*, MIT Press (1997) 21–30
8. Mannila, H.: Theoretical frameworks for data mining. *SIGKDD Explorations* **1** (2000) 30–32
9. Gunopulos, D., Khardon, R., Mannila, H., Saluja, S., Toivonen, H., Sharma, R.S.: Discovering all most specific sentences. *ACM Transactions on Database Systems* **28** (2003) 140–174
10. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In Beeri, C., Buneman, P., eds.: *Database Theory - ICDT'99*. Volume 1540 of *Lecture Notes in Computer Science*, Springer-Verlag (1999) 398–416
11. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Free-sets: a condensed representation of Boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery* **7** (2003) 5–22

12. Bykowski, A., Rigotti, C.: A condensed representation to find frequent patterns. In: Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM (2001)
13. Kryszkiewicz, M.: Concise representation of frequent patterns based on disjunction-free generators. In Cercone, N., Lin, T.Y., Wu, X., eds.: Proceedings of the 2001 IEEE International Conference on Data Mining, IEEE Computer Society (2001) 305–312
14. Calders, T., Goethals, B.: Mining all non-derivable frequent itemsets. In Elomaa, T., Mannila, H., Toivonen, H., eds.: Principles of Data Mining and Knowledge Discovery. Volume 2431 of Lecture Notes in Artificial Intelligence., Springer-Verlag (2002) 74–865
15. Pei, J., Dong, G., Zou, W., Han, J.: On computing condensed pattern bases. In: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), IEEE Computer Society (2002) 378–385
16. Mielikäinen, T., Mannila, H.: The pattern ordering problem. In Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H., eds.: Knowledge Discovery in Databases: PKDD 2003. Volume 2838 of Lecture Notes in Artificial Intelligence., Springer-Verlag (2003) 327–338
17. Mielikäinen, T.: Chaining patterns. In Grieser, G., Tanaka, Y., Yamamoto, A., eds.: Discovery Science. Volume 2843 of Lecture Notes in Artificial Intelligence., Springer-Verlag (2003) 232–243
18. Toivonen, H.: Sampling large databases for association rules. In Vijayaraman, T., Buchmann, A.P., Mohan, C., Sarda, N.L., eds.: VLDB'96, Proceedings of 22nd International Conference on Very Large Data Bases, Morgan Kaufmann (1996) 134–145
19. Boros, E., Gurvich, V., Khachiyan, L., Makino, K.: On the complexity of generating maximal frequent and minimal infrequent sets. In Alt, H., Ferreira, A., eds.: STACS 2002. Volume 2285 of Lecture Notes in Computer Science., Springer-Verlag (2002) 133–141
20. Ausiello, G., Crescenzi, P., Kann, V., Marchetti-Spaccamela, A., Protasi, M.: Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties. Springer-Verlag (1999)
21. Alon, N., Awerbuch, B., Azar, Y., Buchbinder, N., Naor, J.S.: The online set cover problem. In: Proceedings of the 35th Annual ACM Symposium on Theory of Computing, ACM (2003) 100–105
22. Ioannidis, Y.: Approximations in database systems. In Calvanese, D., Lenzerini, M., Motwani, R., eds.: Database Theory - ICDT 2003. Volume 2572 of Lecture Notes in Computer Science. (2003)
23. Mielikäinen, T.: Finding all occurring sets of interest. In Boulicaut, J.F., Džeroski, S., eds.: 2nd International Workshop on Knowledge Discovery in Inductive Databases. (2003) 97–106
24. Matoušek, J.: Geometric Discrepancy: An Illustrated Guide. Volume 18 of Algorithms and Combinatorics. Springer-Verlag (1999)
25. Chazelle, B.: The Discrepancy Method: Randomness and Complexity. Paperback edn. Cambridge University Press (2001)
26. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. Volume 2 of MPS-SIAM Series on Optimization. SIAM (2001)
27. de Farias Jr., I.R., Nemhauser, G.L.: A polyhedral study of the cardinality constrained knapsack problem. *Mathematical Programming* **96** (2003) 439–467