

The Pattern Ordering Problem

Taneli Mielikäinen and Heikki Mannila

HIIT Basic Research Unit, Department of Computer Science, University of Helsinki
{Taneli.Mielikainen,Heikki.Mannila}@cs.helsinki.fi

Abstract. Many pattern discovery methods provide fast tools for finding the frequently occurring patterns in large data sets. Such pattern collections can also be used to approximate the underlying joint distribution, and they summarize the data set well. However, a large set of patterns is unintuitive and not necessarily easy to use. In this paper we consider the problem of ordering a collection of patterns so that each prefix of the ordering gives as good a summary of the data as possible. We formulate this problem for general loss functions, show that the problem has an efficient solution, and prove that its natural variant is NP-complete but the greedy approximation algorithm gives an $e/(e - 1) \approx 1.58$ approximation quality. We apply the general technique to approximation of frequencies of frequent sets, and show that the method gives good empirical results.

1 Introduction

Many pattern discovery methods provide fast tools for finding the frequently occurring patterns in large data sets. However, many methods also result in large collections of patterns which are difficult to use. There has been lots of work on techniques for pruning the pattern collections without losing too much information (see, e.g., [1,2,3,4,5]).

A collection \mathcal{S} of patterns whose frequencies are known can be used to estimate the frequencies of other patterns in several ways. For example, in frequent set mining with threshold σ , if we know the frequencies of AB , AC , and BC we can estimate the frequency of ABC by at least three methods: by $\sigma/2$, by the minimum of the frequencies of AB , AC , and BC , or by maximum entropy methods [6]. Other techniques exist, too, see e.g. [7,8,9].

In this paper we consider the following simple problem: given a collection of patterns and an estimation method for the frequencies of unknown patterns, how should we sort the known patterns in order of decreasing informativeness for the estimation? The solution of this problem gives an ordering such that each prefix is as informative as possible with respect to the following patterns.

We formulate this problem for general pattern classes, estimation methods, and loss functions. We show that the problem can be solved efficiently, and prove that its natural variant is NP-complete but the greedy method yields an $e/(e - 1)$ approximation algorithm for certain loss functions and estimation

methods, where e is the base of natural logarithms. We apply the general technique to approximation of frequencies of frequent sets, and show that the method gives good empirical results.

The rest of this paper is organized as follows. Section 2 gives background on the general framework of pattern discovery and on condensed representations. Section 3 describes the pattern discovery problem shows that an optimal solution for the problem can be used as a good approximation of the pattern collection. In Section 4 the approximation technique is illustrated with concrete estimation methods and loss functions. The technique is experimentally evaluated in Section 5. Section 6 is a short conclusion.

2 Background and Related Work

Pattern discovery, i.e., finding interesting patterns from a data set, is the central task in data mining [10,11]. The pattern discovery problem can be formulated as follows: given a pattern collection \mathcal{P} and a *quality predicate* (or an *interestingness predicate*) $q : \mathcal{P} \rightarrow \{0, 1\}$, find all interesting patterns, i.e., the patterns $p \in \mathcal{P}$ such that $q(p) = 1$. The predicate is usually defined by using a *quality measure* $\phi : \mathcal{P} \rightarrow [0, 1]$ and a threshold value $\sigma \in [0, 1]$:

$$q(p) = \begin{cases} 1 & \text{if } \phi(p) \geq \sigma, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Several measures of quality have been proposed [12]. The most prominent measure of quality is the frequency of the pattern w.r.t. the data set. Especially, frequent set mining has received considerable attention [13]. In frequent set mining, the data set d is a finite sequence $d = d_1 \dots d_n$ of subsets of a set R , the pattern collection \mathcal{P} is the collection of subsets of R , the frequency of a set $X \subseteq R$ (w.r.t. the data set d) is

$$fr(X) = fr(X, d) = \frac{|\{i : X \subseteq d_i, 1 \leq i \leq n\}|}{n}.$$

The set $X \subseteq R$ is considered interesting if and only if $fr(X) \geq \sigma$.

Finding a good measure of quality and an adequate threshold value is not easy. To avoid missing interesting patterns, very low quality threshold values might be needed. This implies that the number of patterns deemed to be interesting can be quite large. This is not necessarily a problem if the true objective of the pattern discovery was to find all the interesting patterns w.r.t. the quality predicate.

Several methods have been developed for finding *condensed representations* of the pattern collections (see e.g. [14,15,16,17,18,19,20,21,22,23,5]). The condensed representations are small descriptions of the pattern collections such that it is possible to infer the original collection of interesting patterns and the quality values (approximately) using some inference method. They depend on some structural properties of the pattern collection and the quality measure. Usually

condensed representations choose a subset of all interesting patterns and infer the quality values of the interesting patterns from that subset.

The most popular condensed representation of pattern collections is the concept of *closed patterns*. The representation depends only on the partial order of the pattern collection and the antimonotonicity of the quality measure. Let \prec be a partial order for the pattern collection \mathcal{P} . A pattern $p \in \mathcal{P}$ is *closed* if and only if its quality value is greater than any of its superpattern's quality value, i.e., if and only if

$$\forall q \in \mathcal{P} : p \prec q \Rightarrow \phi(p) > \phi(q).$$

The number of closed patterns can be much smaller than the number of all patterns.

Unfortunately even the condensed representations of the pattern collection can be very large. Thus we suggest ordering the patterns w.r.t. their informativeness. From the ordered collection of patterns, the user can interactively choose the appropriate trade-off between the number of chosen patterns and the accuracy of the approximation.

For brevity, for the rest of the paper we shall consider frequencies instead of arbitrary quality measures.

3 Pattern Ordering and Frequency Estimation

Most condensed representations of a pattern collection consist of a subset of the pattern collection. Thus a simple approach to simplify the condensed representation is to order the patterns in the condensed representation so that that the next pattern increases the knowledge about the pattern collection most.

Given a collection of patterns, we are interested in finding an ordering such that for each $i = 1, \dots, n$ the prefix p_1, \dots, p_i of the ordering p_1, \dots, p_n gives as much information about p_{i+1}, \dots, p_n as possible. To formulate this we need to define *estimation methods* and *loss functions*. An estimation method ψ takes a subcollection \mathcal{S} of all patterns \mathcal{P} with known frequencies, and provides approximations of the frequencies of all patterns. I.e., an estimation method is a function

$$\psi : \mathcal{P} \times [0, 1]^{\mathcal{S}} \rightarrow [0, 1].$$

A trivial example is the estimation method which gives the known frequencies for patterns in \mathcal{S} and 0 for everything else.

The loss function ℓ tells what penalty is to be paid for errors in estimating the frequencies. The loss function takes as inputs the true frequencies of the patterns and the estimated frequencies, and returns a score for the estimation:

$$\ell : [0, 1]^{\mathcal{P}} \times [0, 1]^{\mathcal{P}} \rightarrow \mathbb{R}.$$

A typical example of a loss function would be the L_2 metric

$$\ell(x, y) = \sqrt{\sum_{q \in \mathcal{P}} (x(q) - y(q))^2}.$$

The task of ordering the patterns can be formulated as a computational problem as follows:

Input: A pattern collection \mathcal{P} , $|\mathcal{P}| = n$, a quality measure $\phi : \mathcal{P} \rightarrow [0, 1]$, an estimation method $\psi : \mathcal{P} \times [0, 1]^{\mathcal{S}} \rightarrow [0, 1]$, $\mathcal{S} \subseteq \mathcal{P}$, and a loss function $\ell : [0, 1]^{\mathcal{P}} \times [0, 1]^{\mathcal{P}} \rightarrow \mathbb{Q}$.

Output: The pattern collection \mathcal{P} as an ordered sequence p_1, p_2, \dots, p_n such that

$$\ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|_{\{p_1, \dots, p_{i-1}, p_i\}})) \leq \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|_{\{p_1, \dots, p_{i-1}, p_j\}}))$$

for each $1 \leq i < j \leq n$, where $\phi|_{\mathcal{S}}$ is the restriction of the mapping ϕ to the set \mathcal{S} .

We call this problem the *pattern ordering problem*. The problem can be solved by a greedy algorithm as follows:

ORDER-PATTERNS($\mathcal{P}, \phi, \psi, \ell$)

```

1   $\mathcal{P}_0 = \emptyset$ 
2  for  $i \leftarrow 0$  to  $n - 1$ 
3      do  $p_{i+1} \leftarrow \arg_p \min \{ \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|_{\mathcal{P}_i \cup \{p\}})) : p \in \mathcal{P} \setminus \mathcal{P}_i \}$ 
4           $\mathcal{P}_{i+1} \leftarrow \mathcal{P}_i \cup \{p_{i+1}\}$ 
5  return  $p_1, \dots, p_n$ 

```

The running time of the algorithm depends on the efficiency of finding in each iteration i the pattern p_{i+1} that decreases the error most. The time complexity is the combined time complexity of finding the minimums. Let $M(\mathcal{P})$ be the maximum time complexity of finding a pattern p_{i+1} such that the loss for $\mathcal{P}_i \cup \{p_{i+1}\}$ is as small as possible. Then the time complexity of the algorithm is bounded by $O(nM(\mathcal{P}))$. For example, using the trivial estimation method which gives the known frequencies for the chosen patterns and 0 for the others, the minimum in each iteration can be found in logarithmic time in n using a heap [24] (assuming the loss depends only on the differences $fr(p) - \psi(p, fr|_{\mathcal{S}})$).

The ordering p_1, \dots, p_n of the pattern collection \mathcal{P} found by the algorithm ORDER-PATTERNS can be interpreted as a refining approximation of the pattern collection: each prefix $\mathcal{P}_k = \{p_1, \dots, p_k\}$ approximates the whole pattern collection \mathcal{P} . The ordering might itself shed some light to the relationships between the patterns. In addition, for several combinations of estimation methods and loss functions it can be shown that each prefix of the ordering gives a frequency approximation that is guaranteed to be at most a constant factor worse than the frequency approximation from any subset of \mathcal{P} of same size.

The greedy approach, in general, offers an efficient approach to find solutions for a wide variety of problems and several exact and approximate algorithms have been successfully derived by this approach [25,26,27,28]. Also in the case of the pattern ordering problem it is possible to show for certain estimation methods and loss functions that any prefix $\mathcal{P}_k = \{p_1, \dots, p_k\}$ of the optimal solution p_1, \dots, p_n for the pattern ordering problem is at most $e/(e-1)$ worse and has at most $(e-1)/e$ times smaller decrease in loss than any size k subset \mathcal{S} of \mathcal{P} .

(For more in-depth introduction to approximability, see e.g. [29].) On the other hand, the problem of finding the k patterns that describe the collection best can be shown to be NP-hard. Thus finding the size k optimal subset of \mathcal{P} seems to be infeasible all but very small k .

Let us define some notation. The decrease of loss w.r.t. frequency estimation without any known frequencies is denoted by

$$\Delta(\mathcal{S}) = \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi(\emptyset))) - \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi(\mathcal{S}))).$$

Let \mathcal{P}_k^* be a size k subset of \mathcal{P} with the smallest loss. The prefix of length k of the optimal solution for the pattern ordering problem is denoted by \mathcal{P}_k . The problem of finding the best k subset of the pattern collection resembles a lot the *minimum set cover problem*, i.e., given a collection \mathcal{T} of subsets of a finite set R , find the smallest subset \mathcal{S} of \mathcal{T} such that $\bigcup \mathcal{S} = R$ [30]. Thus the approximation quality of the algorithm ORDER-PATTERNS can be proven similarly to the approximability of certain variants of the minimum set cover problem [25].

First we prove Lemma 1 below which can be used to show that certain combinations of estimation methods and loss functions guarantee that $\Delta(\mathcal{P}_k) \geq \frac{e-1}{e} \Delta(\mathcal{P}_k^*)$ holds for all $1 \leq i \leq k$. I.e., the decrease of the error in the frequency estimation (w.r.t. the frequency estimation with no patterns) from the length k prefix of the pattern ordering found by the algorithm ORDER-PATTERNS is at least a fraction $(e-1)/e$ of the best decrease of the error over the size k subsets of \mathcal{P} . All one has to show is that the error decreases sufficiently in each iteration. Lemma 1 will be used in Section 4.

Lemma 1. *If*

$$\Delta(\mathcal{P}_i) - \Delta(\mathcal{P}_{i-1}) \geq \frac{\Delta(\mathcal{P}_k^*) - \Delta(\mathcal{P}_{i-1})}{k} \quad (1)$$

holds for all $1 \leq i \leq k$ then

$$\Delta(\mathcal{P}_k) \geq \frac{e-1}{e} \Delta(\mathcal{P}_k^*).$$

holds for all $1 \leq k \leq n$.

Proof. From Equation 1 we get

$$\begin{aligned} \Delta(\mathcal{P}_i) &\geq \frac{1}{k} \Delta(\mathcal{S}) + \left(1 - \frac{1}{k}\right) \Delta(\mathcal{P}_{i-1}) \geq \frac{1}{k} \Delta(\mathcal{P}_k^*) + \left(1 - \frac{1}{k}\right) \Delta(\mathcal{P}_{i-1}) \\ &\geq \frac{1}{k} \Delta(\mathcal{P}_k^*) \sum_{j=0}^{i-1} \left(1 - \frac{1}{k}\right)^j = \frac{1}{k} \Delta(\mathcal{P}_k^*) \frac{\left(1 - \frac{1}{k}\right)^i - 1}{\left(1 - \frac{1}{k}\right) - 1} \\ &= \left(1 - \left(1 - \frac{1}{k}\right)^i\right) \Delta(\mathcal{P}_k^*). \end{aligned}$$

Thus

$$\Delta(\mathcal{P}_k) \geq \left(1 - \left(1 - \frac{1}{k}\right)^k\right) \Delta(\mathcal{P}_k^*) \geq \left(1 - \frac{1}{e}\right) \Delta(\mathcal{P}_k^*)$$

as claimed. □

It is possible to show the similar result for the loss instead of the decrease of the loss:

Lemma 2. *If*

$$\begin{aligned} \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|\mathcal{P}_{i-1})) - \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|\mathcal{P}_i)) &\geq \\ \frac{1}{k} (\ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|\mathcal{P}_{i-1})) - \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|\mathcal{P}_k^*))) &\end{aligned} \quad (2)$$

holds for all $1 \leq i \leq k$ then

$$\ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|\mathcal{P}_k)) \leq \frac{e}{e-1} \ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|\mathcal{P}_k^*))$$

holds for all $1 \leq k \leq n$.

Proof. The proof is essentially identical to the proof of the Lemma 1. □

4 Case Study: Approximating by Maximums of Superpattern Frequencies

In this section we consider approximating the frequencies of the frequent patterns using the maximums of known superpattern frequencies. To define what superpattern is, we need a partial order \prec for the pattern collection \mathcal{P} . A partial order \prec for the collection \mathcal{P} is transitive ($p \prec q \wedge q \prec r \Rightarrow p \prec r$) and irreflexive ($p \prec q \Rightarrow p \neq q$) binary relation on \mathcal{P} . We denote $(p, q) \in \prec$ by $p \prec q$. We further assume that the partial order is antimonotone w.r.t. the frequencies, i.e., $p \prec q \Rightarrow fr(p) \geq fr(q)$. For example, the set inclusion relation is such a partial order. A pattern q is a superpattern of p if and only if $p \prec q$. The estimation method of maximum of superpattern frequencies is

$$\psi(p, fr|\mathcal{S}) = \max_{q \in \mathcal{S}} (\{fr(p) : p = q\} \cup \{fr(q) : p \prec q\} \cup \{0\}).$$

The smallest subset of frequent patterns that is sufficient to describe the frequencies $fr(p)$ of the frequent patterns p in \mathcal{P} correctly is called a collection of *closed frequent patterns*. More precisely, a pattern $p \in \mathcal{P}$ is closed if and only if

$$fr(p) > \max_{q \in \mathcal{P}} \{fr(q) : p \prec q\}.$$

A closure of a pattern $p \in \mathcal{P}$, denoted by $cl(p)$, is the largest superpattern $q \in \mathcal{P}$ of p such that $fr(p) = fr(q)$. The set of closures of a pattern collection \mathcal{P} is denoted by $cl(\mathcal{P})$.

Theorem 1. *The collection $cl(\mathcal{P})$ of closed frequent patterns is the smallest collection such that for all frequent patterns $p \in \mathcal{P}$ we have*

$$fr(p) = \psi(p, fr|cl(\mathcal{P})).$$

Proof. By definition, for each $p \in \mathcal{P}$ there is $q = cl(p) \in cl(\mathcal{P})$ such that $fr(p) = fr(q)$. Also, no closed pattern can be left out from the collection. \square

It follows from the definition of closed frequent patterns that they can be chosen from the collection of the frequent patterns by simply checking for each pattern whether any its superpatterns (subpatterns) have equal frequency and pruning the pattern (subpattern) if that holds. The efficiency of the method depends strongly on the pattern collection, the partial order and their representations. For example, the closed patterns from the collection \mathcal{P} of frequent sets, i.e., the closed frequent sets (or frequent closed sets), can be found in time

$$\sum_{X \in \mathcal{P}, X \neq \emptyset} \binom{|X|}{|X|-1} (|X|-1) = O(|R|^2 |\mathcal{P}|)$$

by applying the fact that $X \in cl(\mathcal{P})$ if and only if $fr(X) \neq fr(Y)$ for all $Y \in \mathcal{P}, Y \supset X, |Y| = |X| + 1$.

The problem turns out to be NP-hard if we allow errors. Let us first consider the maximum of absolute errors, i.e.,

$$\begin{aligned} \ell(fr(\mathcal{P}), \psi(\mathcal{P}, fr|\mathcal{S})) &= \max_{X \in \mathcal{P}} |fr(X) - \psi(X, fr|\mathcal{S})| \\ &= \max_{X \in \mathcal{P}} \left| fr(X) - \max_{Y \in \mathcal{S}} \{fr(Y) : X \subseteq Y\} \right| \\ &= \max_{X \in \mathcal{P}} \left(fr(X) - \max_{Y \in \mathcal{S}} \{fr(Y) : X \subseteq Y\} \right). \end{aligned}$$

Then the problem is NP-hard even for the pattern class of frequent sets:

Theorem 2. *Given a collection \mathcal{P} of frequent sets and a rational number ϵ , it is NP-hard to find a smallest subset \mathcal{S} of \mathcal{P} such that*

$$\max_{X \in \mathcal{P}} \left(fr(X) - \max_{Y \in \mathcal{S}} \{fr(Y) : X \subseteq Y\} \right) \leq \epsilon.$$

Proof. We show the NP-hardness by a reduction from the decision version of the minimum set cover problem, where the objective is, instead of finding the smallest subset $\mathcal{S} \subseteq \mathcal{T}$ such that $\bigcup \mathcal{S} = R$, to decide whether there is a subset $\mathcal{S} \subseteq \mathcal{T}$ of size at most k such that $\bigcup \mathcal{S} = R$. We can assume that each element in R occurs in some set T , the cardinality of each set in \mathcal{T} is greater than one, and no set in \mathcal{T} is contained in another set in \mathcal{T} .

Let us construct the data set d of subsets of R as follows: d consists of \mathcal{T} and appropriate number of one-element subsets of R such that $fr(\{x\}, d) = fr(\{y\}, d)$ for all $x, y \in R$. Let $\epsilon = fr(\{x\}, d) - 1/n, x \in R$.

Then for each $\mathcal{S} \subseteq \mathcal{T}, |\mathcal{S}| \leq k$, holds:

$$\bigcup \mathcal{S} = R \Leftrightarrow \max_{X \in \mathcal{P}} \left(fr(X) - \max_{Y \in \mathcal{S}} \{fr(Y) : X \subseteq Y\} \right) \leq \epsilon.$$

\square

Corollary 1. *Given a pattern collection \mathcal{P} and a rational number ϵ , it is NP-hard to find a smallest subset \mathcal{S} of \mathcal{P} such that*

$$\ell(\phi(\mathcal{P}), \psi(\mathcal{P}, \phi|\mathcal{S})) \leq \epsilon.$$

On the positive side, it can be shown for the estimation method of choosing the maximums of known superpattern frequencies that the problem of choosing size k subset of patterns such that the maximum absolute error is minimized is a special case of the minimum weight set cover, which is, given a collection \mathcal{T} of subsets of a finite set R and a weight function $w : \mathcal{T} \rightarrow [0, 1]$, to find a subset $\mathcal{S} \subseteq \mathcal{T}$ of smallest weight

$$w(\mathcal{S}) = \sum_{p \in \mathcal{S}} w(p)$$

such that $\bigcup \mathcal{S} = R$ [29].

If the loss function is, e.g., the average error instead of the maximum error, the connection to set cover is not so obvious. Also in that case the approximability guarantees can be established:

Theorem 3. *For the prefix \mathcal{P}_k of length k of an optimal solution for the pattern ordering problem and the size k subset of \mathcal{P} with the smallest loss we have*

$$\Delta(\mathcal{P}_k) \geq \frac{e-1}{e} \Delta(\mathcal{P}_k^*)$$

with respect to any loss function

$$\ell(fr(\mathcal{P}), \psi(\mathcal{P}, fr|\mathcal{S})) = \sum_{p \in \mathcal{P}} f(|fr(p) - \psi(p, fr|\mathcal{S})|)$$

where f is a convex strictly increasing function.

Proof. It suffices to show that Equation 1 holds. We have

$$\begin{aligned} & \Delta(\mathcal{P}_i) - \Delta(\mathcal{P}_{i-1}) \\ &= \sum_{p \in \mathcal{P}} f(|fr(p) - \psi(p, fr|\mathcal{P}_{i-1})|) - \sum_{p \in \mathcal{P}} f(|fr(p) - \psi(p, fr|\mathcal{P}_i)|) \\ &\geq \frac{1}{k} \left(\sum_{p \in \mathcal{P}} f(|fr(p) - \psi(p, fr|\mathcal{P}_{i-1})|) - \sum_{p \in \mathcal{P}} f(|fr(p) - \psi(p, fr|\mathcal{P}_k^*)|) \right) \\ &= \frac{\Delta(\mathcal{P}_k^*) - \Delta(\mathcal{P}_{i-1})}{k} \end{aligned}$$

because $\{p_i\} = \mathcal{P}_i \setminus \mathcal{P}_{i-1}$ is the pattern that decreases the error most and

$$\begin{aligned} & \sum_{p \in \mathcal{P}} f(|fr(p) - \psi(p, fr|\mathcal{S})|) \\ &= \sum_{p \in \mathcal{P}} \min \{ f(|fr(p) - \psi(p, fr|\mathcal{S} \setminus \mathcal{T})|), f(|fr(p) - \psi(p, fr|\mathcal{T})|) \}. \end{aligned}$$

holds for all $\mathcal{T} \subseteq \mathcal{S}$. □

The computation of the approximation can be made more efficient by observing the following fact that all but the closed patterns in \mathcal{P} can be neglected.

Theorem 4. *For all ℓ and $\mathcal{S} \subseteq \mathcal{P}$ we have*

$$\ell(fr(\mathcal{P}), \psi(\mathcal{P}, fr|\mathcal{S})) = \ell(fr(\mathcal{P}), \psi(\mathcal{P}, fr|cl(\mathcal{S}))).$$

Proof. Any pattern $p \in \mathcal{S}$ can be replaced by $cl(p)$ as $fr(p) = fr(cl(p))$, and if $\psi(p, fr|\mathcal{S}) = fr(p)$ then $\psi(p, fr|\mathcal{S}) = fr(cl(p))$. \square

5 Experiments: Approximating Frequent Sets

We implemented the ORDER-PATTERNS algorithm to evaluate the practical usefulness of the method. In the experiments we computed frequent sets with different minimum frequency thresholds for two data sets from UCI KDD Repository:¹ Internet Usage data set consisting 10104 rows and 10674 attributes, and IPUMS Census data set consisting of 88443 rows and 39954 attributes.

The estimation method was the maximum of chosen superset frequencies, i.e., $\psi(X, fr|\mathcal{S}) = \max_{Y \in \mathcal{S}} \{fr(Y) : X \subseteq Y\}$, and the loss function was the mean of absolute errors

$$\ell(fr(\mathcal{P}), \psi(\mathcal{P}, fr|\mathcal{S})) = \frac{1}{|\mathcal{P}|} \left(\sum_{X \in \mathcal{P}} fr(X) - \max_{Y \in \mathcal{S}} \{fr(Y) : X \subseteq Y\} \right).$$

The results are shown in Figure 1, and in Tables 1 and 2. The results show that relatively short prefixes can be used to obtain a good accuracy in estimating the frequencies. The inversion of the order of the error curves in Figure 1 is due to the combination of the estimation method and the loss function: As the initial frequency estimates are all zero, the average absolute error is smaller for lower minimum frequency thresholds. On the other hand the frequencies can be estimated exactly from the closed frequent sets and the number of closed frequent sets is smaller for higher minimum frequency thresholds.

6 Conclusions

We have considered the problem of ordering a pattern collection in such a way that each prefix of the ordered sequence of patterns would be as good a summary of the pattern collection as possible. A general algorithm was given, the problem complexity and the algorithm were analyzed and the approach was justified experimentally. It seems that the problem of finding good orderings of pattern collections is useful and interesting. Several open problems remain. One specially interesting one is combining pattern discovery and ordering steps: could we somehow discover patterns in an order that approximates the most informative pattern ordering. To do this exactly is impossible, but there might be some possibilities of obtaining approximate results in the style of competitive analysis.

¹ <http://kdd.ics.uci.edu>

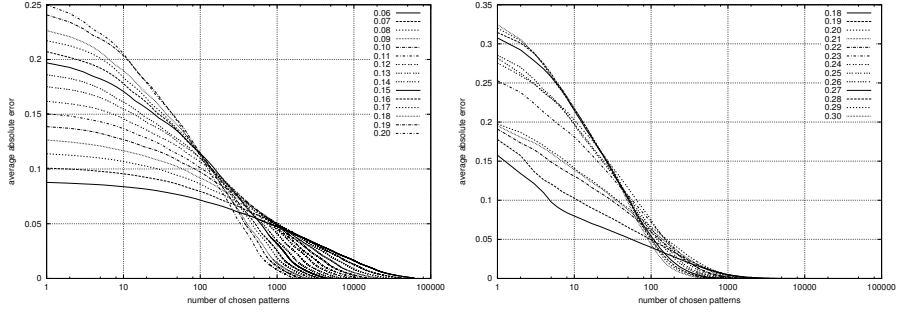


Fig. 1. Internet Usage data (left) and IPUMS Census data (right). The axes are the length of the prefix of the pattern ordering and the average absolute error of the frequency estimation from the prefix. Each curve corresponds to the minimum frequency threshold given as its label.

σ	$ \mathcal{P} $	$cl(\mathcal{P})$	$ \tau(0.001) $	$ \tau(0.005) $	$ \tau(0.01) $	$ \tau(0.02) $	$ \tau(0.04) $	$ \tau(0.08) $
0.17	3246	3246	2672	1925	1421	970	597	231
0.16	4013	4013	3254	2295	1671	1132	655	242
0.15	4983	4983	3994	2764	1995	1377	775	270
0.14	6291	6290	4955	3339	2362	1602	860	261
0.13	8000	7998	6208	4093	2881	1972	1034	281
0.12	10476	10472	7970	5118	3562	2414	1189	289
0.11	13813	13802	10267	6352	4305	2804	1284	264
0.10	18615	18594	13468	8068	5409	3395	1423	245
0.09	25729	25686	18035	10399	6920	4094	1587	203
0.08	36812	36714	24870	13681	9032	5008	1708	153
0.07	54793	54550	35441	18477	12147	6276	1803	95

Table 1. Internet Usage data. The column σ corresponds to the minimum frequency threshold. Columns $|\mathcal{P}|$ and $cl(\mathcal{P})$ correspond to the cardinalities of the frequent sets and the closed frequent sets, respectively. Each column $|\tau(x)|$ corresponds to the length of the shortest prefix found by the algorithm ORDER-PATTERNS such that the average absolute error is at most x .

σ	$ \mathcal{P} $	$cl(\mathcal{P})$	$ \tau(0.001) $	$ \tau(0.005) $	$ \tau(0.01) $	$ \tau(0.02) $	$ \tau(0.04) $	$ \tau(0.08) $
0.28	11443	1696	551	351	260	184	120	66
0.27	13843	1948	624	395	292	203	128	68
0.26	17503	2293	725	456	338	233	147	71
0.25	20023	2577	810	502	369	256	161	77
0.24	23903	3006	944	583	427	293	185	92
0.23	31791	3590	1093	661	477	328	196	85
0.22	53203	4271	1194	678	481	316	171	57
0.21	64731	5246	1454	813	573	372	189	62
0.20	86879	6689	1771	949	661	424	218	67
0.19	151909	8524	1974	953	628	363	151	27
0.18	250441	10899	2212	992	625	312	99	10

Table 2. IPUMS Census data. The columns are as in Table 1.

References

1. 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
2. Gouda, K., Zaki, M.J.: Efficiently mining maximal frequent itemsets. In Cercone, N., Lin, T.Y., Wu, X., eds.: Proceedings of the 2001 IEEE International Conference on Data Mining. IEEE Computer Society (2001) 163–170
3. 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
4. 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
5. Zaki, M.J., Hsiao, C.J.: CHARM: An efficient algorithms for closed itemset mining. In Grossman, R., Han, J., Kumar, V., Mannila, H., Motwani, R., eds.: Proceedings of the Second SIAM International Conference on Data Mining, SIAM (2002)
6. Mannila, H., Pavlov, D., Smyth, P.: Prediction with local patterns using cross-entropy. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (1999) 357–361
7. Kessler, D., Schiff, J.: Inclusion-exclusion redux. *Electronic Communications in Probability* **7** (2002) 85 – 96
8. 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
9. Pavlov, D., Mannila, H., Smyth, P.: Beyond independence: probabilistic methods for query approximation on binary transaction data. *IEEE Transactions on Data and Knowledge Engineering* (2003) To appear.
10. 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
11. 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
12. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In Hand, D., Keim, D., Ng, R., eds.: Proceedings of the Eight International Conference on Knowledge Discovery and Data Mining (KDD-2002), ACM (2002)
13. Hipp, J., Güntzer, U., Nakhaeizadeh, G.: Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations* **1** (2000) 58–64
14. Boulicaut, J.F., Bykowski, A.: Frequent closures as a concise representation for binary data mining. In Terano, T., Liu, H., Chen, A.L.P., eds.: Knowledge Discovery and Data Mining. Volume 1805 of Lecture Notes in Artificial Intelligence., Springer-Verlag (2000) 62–73
15. 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

16. 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)
17. 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
18. Kryszkiewicz, M., Gajek, M.: Concise representation of frequent patterns based on generalized disjunction-free generators. In Chen, M.S., Yu, P., Liu, B., eds.: Advances in Knowledge Discovery and Data Mining. Volume 2336 of Lecture Notes in Artificial Intelligence., Springer-Verlag (2002) 159 – 171
19. Mielikäinen, T.: Frequency-based views to pattern collections. In: IFIP/SIAM Workshop on Discrete Mathematics and Data Mining. (2003)
20. 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
21. 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), 9-12 December 2002, Maebashi City, Japan, IEEE Computer Society (2002) 378–385
22. Pei, J., Han, J., Mao, T.: CLOSET: An efficient algorithm for mining frequent closed itemsets. In Gunopulos, D., Rastogi, R., eds.: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. (2000) 21–30
23. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with TITANIC. *Data & Knowledge Engineering* **42** (2002) 189–222
24. Knuth, D.E.: Sorting and Searching. second edn. Volume 3 of The Art of Computer Programming. Addison-Wesley (1998)
25. Feige, U.: A threshold of $\ln n$ for approximating set cover. *Journal of the ACM* **45** (1998) 634 – 652
26. Guha, S., Khuller, S.: Greedy strikes back: Improved facility location algorithms. *Journal of Algorithms* **31** (1999) 228 – 248
27. Helman, P., Moret, B.M.E., Shapiro, H.D.: An exact characterization of greedy structures. *SIAM Journal on Discrete Mathematics* **6** (1993) 274 – 283
28. Kempe, D., Kleinberg, J., Éva Tardos: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2003)
29. Ausiello, G., Crescenzi, P., Kann, V., Marchetti-Spaccamela, A., Protasi, M.: Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties. Springer-Verlag (1999)
30. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Company (1979)