On the Complexity of Generating Maximal Frequent and Minimal Infrequent Sets*

E. Boros[1], V. Gurvich[1], L. Khachiyan[2], and K. Makino[3]

[1] RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway New Jersey 08854-8003; {boros,gurvich}@rutcor.rutgers.edu.

[2] Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey, 08854-8019; leonid@cs.rutgers.edu

[3] Division of Systems Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka, 560-8531, Japan; makino@sys.es.osaka-u.ac.jp

Abstract. Let A be an $m \times n$ binary matrix, $t \in \{1, \ldots, m\}$ be a threshold, and $\varepsilon > 0$ be a positive parameter. We show that given a family of $O(n^{\varepsilon})$ maximal t-frequent column sets for A, it is NP-complete to decide whether A has any further maximal t-frequent sets, or not, even when the number of such additional maximal t-frequent column sets may be exponentially large. In contrast, all minimal t-infrequent sets of columns of A can be enumerated in incremental quasi-polynomial time. The proof of the latter result follows from the inequality $\alpha \leq (m-t+1)\beta$, where α and β are respectively the numbers of all maximal t-frequent and all minimal t-infrequent sets of columns of the matrix A. We also discuss the complexity of generating all closed t-frequent column sets for a given binary matrix.

Keywords. Data mining, frequent sets, infrequent sets, independent sets, hitting sets, transversals, dualization.

1 Introduction

Let us consider an $m \times n$ binary matrix $A : \mathcal{R} \times \mathcal{C} \to \{0, 1\}$, and an integral threshold value $t \in \{1, \ldots, m\}$. To each subset $C \subseteq \mathcal{C}$ of the columns, let us associate the subset $R(C) \subseteq \mathcal{R}$ of all those rows $r \in \mathcal{R}$, for which $A(r, c) = 1$ in every column $c \in C$. The cardinality $|R(C)|$ is called the *support* of the set C. Let us call a subset $C \subseteq \mathcal{C}$ of the columns *frequent* (or more precisely, *t-frequent*) if its support is at least the given integral threshold t, i.e. if $|R(C)| \geq t$, and let us denote by $\mathcal{F}_t$ the family of all t-frequent subsets of the columns of the given binary matrix A. Let us further call a subset $C \subseteq \mathcal{C}$ *infrequent* (or *t-infrequent*)

* This research is supported in part by the National Science Foundation (Grant IIS-0118635), the Office of Naval Research (Grant N00014-92-J-1375), and Grants-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan. Visits of the second author to Rutgers University were also supported by DIMACS, the National Science Foundation's Center for Discrete Mathematics and Theoretical Computer Science.

H. Alt and A. Ferreira (Eds.): STACS 2002, LNCS 2285, pp. 133–141, 2002.

if its support does not exceed the given threshold t, i.e. if $|R(C)| < t$. Clearly, subsets of frequent sets are also frequent, and supersets of infrequent sets are also infrequent. Let us denote by $\mathcal{M}_t \subseteq \mathcal{F}_t$ the family of all maximal t-frequent sets (i.e. those which are t-frequent, but no superset of them is t-frequent), and by $\mathcal{I}_t$ the family of all minimal t-infrequent sets (i.e. those which are infrequent but all proper subsets of them are t-frequent.)

The generation of frequent sets of a given binary matrix A is an important task of knowledge discovery and data mining, e.g. it is used for mining association rules [1,2,16,21,22], correlations [8], sequential patterns [3], episodes [23], emerging patterns [10], and appears in many other applications. Most practical procedures to generate $\mathcal{F}_t$ are based on the anti-monotone *Apriori* heuristic (see [2]) and build frequent sets in a bottom-up way, running in time proportional to the number of frequent sets. It was also demonstrated recently in [9] that these methods are inadequate in practice when there are (many) frequent sets of large size (see also [4,13,19]), due the fact that $|\mathcal{F}_t|$ can be exponentially larger than $|\mathcal{M}_t|$.

These results show that it is perhaps more important to find the *boundary* of the frequent sets, i.e. the families of maximal frequent and minimal infrequent sets $\mathcal{M}_t \cup \mathcal{I}_t$ (proposed e.g. in [26]), and use those as condensed representation of the data set, as suggested in [21]. Furthermore, no algorithm using membership queries "$X \in \mathcal{F}_t$?" can generate all (maximal) frequent sets in fewer than $|\mathcal{M}_t \cup \mathcal{I}_t|$ steps (see e.g. [16]). There were several other examples presented in [21] to show the usefulness of maximal frequent sets and minimal infrequent sets, e.g. providing error bounds for the confidence of an arbitrary Boolean rule, in terms of minimal infrequent sets.

In this short paper we prove the following inequality.

Theorem 1. *If* $\mathcal{I}_t \neq \emptyset$ *then*

$$|\mathcal{M}_t| \leq (m - t + 1)|\mathcal{I}_t|. \tag{1}$$

Note that the requirement that $\mathcal{I}_t$ must be non-empty is necessary because for $|\mathcal{I}_t| = 0$ we would have $\mathcal{M}_t = \{\mathcal{C}\}$ and hence $|\mathcal{M}_t| = 1$, in contradiction with (1). The condition $\mathcal{I}_t \neq \emptyset$ thus excludes the degenerate case, when the entire column set of A is t-frequent.

Before proceeding further, let us mention some algorithmic implications of (1). It follows from the results of [5,16,17] that the incremental complexity of generating $\mathcal{M}_t \cup \mathcal{I}_t$ is equivalent with that of the transversal hypergraph problem (for definitions and related results see e.g. [11]). The latter problem is known to be solvable in incremental quasi-polynomial time [14], implying thus the same for the joint generation of maximal frequent and minimal infrequent sets. Specifically, it follows from [14] that for each $k \leq |\mathcal{M}_t \cup \mathcal{I}_t|$, we can generate k sets in $\mathcal{M}_t \cup \mathcal{I}_t$ in $poly(n, m) + k^{o(\log k)}$ time. The above inequality (1) clearly implies that if we can generate $\mathcal{M}_t \cup \mathcal{I}_t$ in time $T(|\mathcal{M}_t \cup \mathcal{I}_t|)$, then the entire set $\mathcal{I}_t$ can be generated in time $T((m - t + 2)|\mathcal{I}_t|)$. We thus conclude that the family of minimal infrequent sets $\mathcal{I}_t$ can be generated in output quasi-polynomial time,

i.e. in time bounded by a quasi-polynomial in $|\mathcal{I}_t|$. This can be further improved to show that the incremental complexity of generating $\mathcal{I}_t$ is also equivalent with that of the transversal hypergraph problem (see [6,7] for more detail). Hence

Corollary 1. *For each $k \leq |\mathcal{I}_t|$, we can compute k minimal t-infrequent sets of A in $poly(n,m) + K^{o(\log K)}$ time, where $K = \max\{k,m\}$.*

Let us note next that the matrix A can also be interpreted as the adjacency matrix of a bipartite graph $G = (\mathcal{R} \cup \mathcal{C}, E)$, i.e., in which $(r,c) \in E$ iff $A(r,c) = 1$. Then, maximal frequent sets of A correspond to maximal complete bipartite subgraphs $K_{R,C} \lhd G$, where $R \subseteq \mathcal{R}$, and $C \subseteq \mathcal{C}$. It is known (see e.g., [18]) that determining the number of maximal complete bipartite subgraphs of a bipartite graph is a very difficult #P-complete problem, and hence by the above equivalence, determining $|\cup_{t \geq 1} \mathcal{M}_t|$ is also #P-complete. (In [12] an $O(l^3 2^{2l}(m+n))$ algorithm was presented to generate all maximal complete bipartite subgraphs of a bipartite graph on $m+n$ vertices, or equivalently, to generate all maximal frequent sets of A, where l denotes the maximum of $|C||R(C)|/(|C|+|R(C)|-1)$, with the maximum taken over all maximal frequent sets C.) Strengthening these (negative) results and a statement of [20], we can show the following:

Theorem 2. *Given an $m \times n$ matrix A, a threshold t, and a subfamily $\mathcal{S} \subseteq \mathcal{M}_t$, it is NP-hard to decide if $\mathcal{S} \neq \mathcal{M}_t$, even if $|\mathcal{S}| = O(n^{\varepsilon})$ and $|\mathcal{M}_t|$ is exponentially large in n whenever $\mathcal{S} \neq \mathcal{M}_t$, where $\varepsilon > 0$ can be arbitrarily small.*

Yet, it is easy to show that determining whether or not $\mathcal{S} \neq \mathcal{M}_t$ for polylogarithmically large $|\mathcal{S}|$ can be done in polynomial time.

Finally, let us remark that the inequality (1) is best possible, as the following examples show. Let A be an $m \times (m-t+1)$ matrix, in which every entry is 1, except the diagonal entries in the first $m-t+1$ rows, which are 0. Then any $m-t$ element subset of the columns is a maximal t-frequent set, while the set $\mathcal{C}$ of all columns is the only minimal t-infrequent set. Thus we have equality in (1) for such matrices.

It is also worth mentioning that (1) stays accurate, up to a factor of $\log m$, even if $m \gg n$ and $|\mathcal{I}_t|$ is arbitrarily large. To see this, let us consider a binary matrix A with $m = 2^k$ rows and $n = 2k$ columns ($k \geq 1$, integer), such that each row contains exactly one 0 and one 1 in each pair of the adjacent columns $\{1,2\}$, $\{3,4\}, \ldots, \{2k-1, 2k\}$, and in all 2^k possible ways in the $m = 2^k$ rows. It is not difficult to see that for $t = 1$ there are 2^k maximal 1-frequent sets (every row of the matrix is the characteristic vector of a maximal 1-frequent set), and that there are only k minimal 1-infrequent sets, namely $\{2i-1, 2i\}$ for $i = 1, ..., k$. Thus, for such examples we have $|\mathcal{M}_t| = (m/\log m)|\mathcal{I}_t|$. The same examples also show that $|\mathcal{M}_t|$ cannot be bounded by a polynomial function in $|\mathcal{I}_t|$ and n, the number of columns of A. Needless to say that in general, $|\mathcal{I}_t|$ cannot be bounded by a polynomial in $|\mathcal{M}_t|$, n and m.

2 Closed Frequent Sets

Following [27], let us call a subset $C \subseteq \mathcal{C}$ of the columns *closed* if $R(C') \subsetneqq R(C)$ for all $C' \supsetneqq C$, or in other words, if $c \in C$ exactly when $A(r,c) = 1$ for all $r \in R(C)$ (see also [24,25]). Let us further denote by $\mathcal{D}_t$ the family of all closed t-frequent column sets. Clearly, we have

$$\mathcal{M}_t \subseteq \mathcal{D}_t \subseteq \mathcal{F}_t$$

for all $t = 1, ..., m$.

For the converse direction, it is also easy to see by the definitions that every closed t-frequent set is also a maximal t'-frequent set for some $t' \geq t$, implying the following claim.

Proposition 1. $\mathcal{D}_t = \cup_{t' \geq t} \mathcal{M}_{t'}$. □

Let us note next that for $C \in \mathcal{D}_t \setminus \mathcal{D}_{t+1}$ we either have a subset $C' \subset C$, $C' \neq \emptyset$ for which $C' \in \mathcal{D}_{t+1}$, or $A(r,c) = 0$ for all $c \in C$ and $r \notin R(C)$. Since the number of subsets of the latter type is limited by n and it is easy to identify those in $O(mn)$ time, all sets in $\mathcal{D}_t \setminus \mathcal{D}_{t+1}$ can be obtained in $O(nm + n|\mathcal{D}_{t+1}|)$ time by trying to increment all sets of $\mathcal{D}_{t+1}$ in all possible ways. Denoting by τ the maximum number of 1s in a column of A, we can claim that $\mathcal{D}_t = \emptyset$ for all $t > \tau$, and that $\mathcal{D}_\tau$ can easily be generated in $O(nm)$ time. Putting all these together, we can conclude that, in contrast to Theorem 2, closed frequent sets can be generated efficiently.

Proposition 2. *The family $\mathcal{D}_t$ can be generated in incremental polynomial time for any $t \in \{1, ..., m\}$.* □

Let us finally remark that in many examples we can have $|\mathcal{D}_t|$ exponentially larger than $|\mathcal{M}_t|$ and $|\mathcal{F}_t|$ exponentially larger than $|\mathcal{D}_t|$, simultaneously. To see such an infinite family of examples, let us choose k, $l > k$ and t as arbitrary positive integers, set $m = kt$, $n = kl$, and define the matrix A as follows. Let $U_i = \{(i-1)l + j \mid j = 1, ..., l\}$ for $i = 1, ..., k$, let $\mathcal{C} = \cup_{i=1}^{k} U_i$, and let $a_i \in \{0,1\}^n$ $(1 \leq i \leq k)$ be a binary vector in which $a_{ij} = 0$ if $j \in U_i$, and $a_{ij} = 1$ otherwise. Finally, let $A \in \{0,1\}^{m \times n}$ be the matrix formed by t copies, as rows, of each of the vectors a_i, $i = 1, ..., k$.

It is now easy to see that the maximal t-frequent sets in this matrix are exactly the column subsets C of the form $C = \mathcal{C} \setminus U_i$ for some $1 \leq i \leq k$. Thus, $|\mathcal{M}_t| = k$. Furthermore, the column subsets of the form $C = \mathcal{C} \setminus \cup_{i \in S} U_i$ for some nonempty subset $S \subseteq \{1, ..., k\}$ are exactly the closed t-frequent sets of A, therefore we have $|\mathcal{D}_t| = 2^k - 1$. Finally, any subset $C \subseteq \mathcal{C}$ of the columns, disjoint from at least one of the sets U_1, ..., U_k, is a t-frequent set, implying that $|\mathcal{F}_t| > 2^{(l-1)k} > |\mathcal{D}_t|^k$.

3 Proofs of Theorems 1 and 2

For the proof of Theorem 1 we shall need the following combinatorial lemma.

Lemma 1. *Given a base set V of size $|V| = m$ and a threshold $t \in \{1, \ldots, m\}$, let $\mathcal{S} = \{S_1, \ldots, S_\alpha\}$ and $\mathcal{T} = \{T_1, \ldots, T_\beta\}$ be two families of subsets of V such that*

(i) $|S| \geq t$ for all $S \in \mathcal{S}$, while $|T| < t$ for all $T \in \mathcal{T}$, and
(ii) for each of the $\alpha(\alpha-1)/2$ pairs $S', S'' \in \mathcal{S}$ there exists a $T \in \mathcal{T}$, such that $S' \cap S'' \subseteq T$.

Then $\alpha \leq (m - t + 1)\beta$, whenever $\alpha \geq 2$.

Let us remark first that if $\alpha = 1$ then the family $\mathcal{T}$ might be empty, which would violate the inequality $\alpha \leq (m-t+1)\beta$. Let us also mention that by (ii) $\beta \geq 1$ must hold whenever $\alpha \geq 2$. In addition, conditions (i) and (ii) together imply that $\mathcal{S}$ is a Sperner family, i.e. $S_i \not\subseteq S_j$ whenever $i \neq j$ (since otherwise $S_i = S_i \cap S_j \subseteq T_k$ would follow by (ii) for some $T_k \in \mathcal{T}$, contradicting condition (i).) Without loss of generality we can assume that $\mathcal{T}$ is also Sperner, for otherwise we can replace $\mathcal{T}$ by the family of all maximal sets of $\mathcal{T}$.

Proof of Lemma 1. We shall prove the Lemma by induction on t. If $t = 1$ then $\mathcal{T} = \{\emptyset\}$ by condition (i). In view of (ii), this implies that the sets of $\mathcal{S}$ are pairwise disjoint, and hence $\alpha \leq m = (m - t + 1)\beta$.

In a general step, let us define subfamilies $\mathcal{S}_v = \{S \setminus \{v\} \mid S \in \mathcal{S},\ v \in S\}$ and $\mathcal{T}_v = \{T \setminus \{v\} \mid S \in \mathcal{T},\ v \in T\}$ for each $v \in V$. Let us further introduce the notations $\alpha_v = |\mathcal{S}_v|$, and $\beta_v = |\mathcal{T}_v|$.

For vertices $v \in V$ for which $\alpha_v \geq 2$ (and thus $\beta_v \geq 1$) the families $\mathcal{S}_v$ and $\mathcal{T}_v$ satisfy all the assumptions of the Lemma with $m' = m - 1$ and $t' = t - 1$, and hence

$$\alpha_v \leq (m' - t' + 1)\beta_v = (m - t + 1)\beta_v \tag{2}$$

follows by the inductive hypothesis. Let us then consider the partition $V = V_1 \cup V_2$, where $V_1 = \{v \in V \mid \alpha_v \leq 1\}$, and $V_2 = \{v \in V \mid \alpha_v \geq 2\}$. Summing up the inequalities (2) for all $v \in V_2$, we obtain

$$\sum_{v \in V_2} \alpha_v \leq (m - t + 1) \sum_{v \in V_2} \beta_v. \tag{3}$$

On the left hand side, using condition (i) and the definition of α_v we obtain

$$\alpha t - |V_1| \leq \sum_{S \in \mathcal{S}} |S| - |V_1| \leq \sum_{S \in \mathcal{S}} (|S| - |S \cap V_1|) = \sum_{S \in \mathcal{S}} |S \cap V_2| = \sum_{v \in V_2} \alpha_v, \tag{4}$$

where the first inequality follows by $|S| \geq t$ for $S \in \mathcal{S}$, while the second one is implied by $|V_1| \geq \sum_{S \in \mathcal{S}} |S \cap V_1|$, which follows from the definition of V_1.

On the right hand side of (2) we can write

$$\sum_{v\in V_2}\beta_v=\sum_{T\in\mathcal{T}}|T\cap V_2|\le\sum_{T\in\mathcal{T}}|T|\le\beta(t-1),\tag{5}$$

where the first equality follows by the definition of β_v and $\mathcal{T}_v$, and the last inequality follows by the conditions $|T|<t$ for $T\in\mathcal{T}$.

Putting together (3),(4) and (5) we obtain $t\alpha-|V_1|\le(m-t+1)(t-1)\beta$, or equivalently that

$$\alpha\le\frac{|V_1|}{t}+\frac{t-1}{t}(m-t+1)\beta.\tag{6}$$

If $|V_1|\le m-t+1$, then

$$\frac{|V_1|}{t}+\frac{t-1}{t}(m-t+1)\beta\le(m-t+1)\beta,$$

and hence $\alpha\le(m-t+1)\beta$ by (6). On the other hand, if $|V_1|>m-t+1$, then for each set $S\in\mathcal{S}$ we have $|S\cap V_1|\ge|S|-|V_2|\ge t-|V_2|>1$. Now by the definition of the set V_1 we obtain $\alpha\le|V_1|/(t-|V_2|)=(m-|V_2|)/(t-|V_2|)\le m-t+1\le(m-t+1)\beta$. □

Proof of Theorem 1. Assume without loss of generality that $|\mathcal{M}_t|\ge 2$, for otherwise (1) readily follows from the assumption of the theorem that $|\mathcal{I}_t|\ge 1$. Let us recall that to any subset $C\subseteq\mathcal{C}$ of the columns we have associated the subset $R(C)$ of those rows $r\in\mathcal{R}$ for which $A(r,c)=1$ for every column $c\in C$. Thus, by definition we have $R(C)=\bigcap_{y\in C}R(\{y\})$, implying

$$R(C'\cup C'')=R(C')\cap R(C'')\ \text{ for all }\ C',C''\subseteq\mathcal{C}.\tag{7}$$

In its turn, (7) implies that the mapping $C\mapsto R(C)$ is anti-monotone, i.e. $R(C')\supseteq R(C'')$ whenever $C'\subseteq C''$. Furthermore, $|R(F)|\ge t$ for every maximal t-frequent set $F\in\mathcal{M}_t$, while $|R(U)|<t$ for every minimal t-infrequent set $U\in\mathcal{I}_t$. It is also easy to see that the restriction of the above mapping on $\mathcal{M}_t$ is injective, i.e. $R(F')\ne R(F'')$ for any two distinct maximal t-frequents sets of columns $F',F''\in\mathcal{M}_t$. If $F',F''\in\mathcal{M}_t$ then their union $F'\cup F''$ is not t-frequent, and hence there exists a minimal t-infrequent set $U\in\mathcal{I}_t$, for which $R(F')\cap R(F'')=R(F'\cup F'')\subseteq R(U)$. Thus, the families $\mathcal{S}=\{R(F)\mid F\in\mathcal{M}_t\}$ and $\mathcal{T}=\{R(U)\mid U\in\mathcal{I}_t\}$ satisfy the conditions of Lemma 1 with $V=\mathcal{R}$, which implies the inequality $|\mathcal{S}|\le(m-t+1)|\mathcal{T}|$. Since the mapping $C\mapsto R(C)$ is a one-to one correspondence between $\mathcal{M}_t$ and $\mathcal{S}$, we have $|\mathcal{S}|=|\mathcal{M}_t|$. Now (1) follows from the trivial inequality $|\mathcal{T}|\le|\mathcal{I}_t|$. □

Proof of Theorem 2. We reduce our problem from the following well-known NP-complete problem: Given a graph $G=(V,E)$ and an integer threshold t, determine if G contains an independent vertex set of size at least t. Let us first

substitute every vertex $v \in V$ of G by two new vertices v' and v'' connected by an edge, i.e., consider the graph $G' = (V', E')$, where $V' = \{v', v'' \mid v \in V\}$ and $E' = \{\{(v', v'') \mid v \in V\} \cup \{(v', u'), (v', u''), (v'', v'), (v'', u'') \mid (u, v) \in E\}$. Clearly, $G' = (V', E')$ has an independent set of size t if and only if G has one, moreover, if G' has one, then it has at least 2^t.

Let us now associate a matrix A to G' as follows. Let $\mathcal{C} = V'$ be the set of columns of the matrix A. To every edge $(v, w) \in E'$ we assign $t - 2$ identical rows in A containing 0 in the columns v and w, and 1 in all other columns. Furthermore, to every vertex $v \in V'$ we assign one row containing 0 in the column v and 1 in all other columns. Thus, A has $m = (t-2)|E'|+|V'| = t|V|+4(t-2)|E|$ rows, and $n = |V'| = 2|V|$ columns.

Clearly, for every edge $e = (v, w) \in E'$ the set $C_e = \mathcal{C} \setminus \{v, w\}$ is a maximal t-frequent set of A. Let $\mathcal{S} = \{C_e \mid e \in E'\}$. We claim that $\mathcal{S} \neq \mathcal{M}_t$ for this matrix, if and only if there exists an independent set I of size $|I| \geq t$ in the graph G'.

To see this claim, let assume first that $I \subseteq V'$ is an independent set of G', $|I| \geq t$. Then $R(V' \setminus I)$ contains all rows corresponding to vertices $v \in I$, and hence $|R(V' \setminus I)| \geq |I| \geq t$. Since I does not contain an edge of the graph, the set $C = V' \setminus I$ is not a subset of the member of $\mathcal{S}$, and thus it is contained by a maximal t-frequent set of the matrix A, which does not belong to $\mathcal{S}$.

For the other direction, let us assume now that $C \subseteq \mathcal{C} = V'$ is a maximal t-frequent set of A, not contained by members of $\mathcal{S}$. This latter implies that $I = V' \setminus C$ does not contain an edge of G', i.e. that I is an independent set of G'. This also implies that $R(C)$ cannot contain any of the rows corresponding to an edge of G', and hence $|R(C)| = |V' \setminus C| = |I|$. Thus, $|I| \geq t$ follows by our assumption that C is a t-frequent set, i.e. I is an independent set of size at least t.

Let us recall finally that the maximum independent set problem remains NP-hard, even if the input is restricted to cubic planar graphs (see e.g. [15]), i.e. we can assume $|E| = O(|V|)$. Therefore, we have $|\mathcal{S}| = |E'| = |V| + 4|E| = O(|V|)$, and either we have $\mathcal{S} = \mathcal{M}_t$, or $|\mathcal{M}_t| \geq |\mathcal{S}| + 2^t$. Since we can assume without loss of generality that $t = \Theta(|V|)$, we obtain the the statement of the theorem for $|S| = O(n)$, that is for $\varepsilon = 1$. For smaller values of ε it suffices to add $n^{1/\varepsilon}$ isolated vertices to G'. □

References

1. R. Agrawal, T. Imielinski and A. Swami. Mining associations between sets of items in massive databases. In: *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data*, pp. 207-216.
2. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, Fast discovery of association rules, In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy eds., *Advances in Knowledge Discovery and Data Mining*, 307-328, AAAI Press, Menlo Park, California, 1996.
3. R. Agrawal and R. Srikant. Mining sequential patterns. In: *Proceedings of the 11th International Conference on Data Engineering, 1995*, pp.3-14.

4. R.J. Bayardo, Efficiently mining long patterns from databases. In: *Proceedings of the 1998 ACM-SIGMOD International Conference on Management of Data*, pp. 85-93.
5. J. C. Bioch and T. Ibaraki, Complexity of identification and dualization of positive Boolean functions, *Information and Computation* 123 (1995) 50-63.
6. E. Boros, V. Gurvich, L. Khachiyan and K.Makino, Generating partial and multiple transversals of a hypergraph. In: *Proceedings of the 27th International Colloquium on Automata, Languages and Programming (ICALP)*, (U. Montanari, J.D.P. Rolim and E. Welzl, eds.) Lecture Notes in Computer Science **1853** pp. 588-599, (Springer Verlag, Berlin, Heidelberg, New York, 2000).
7. E. Boros, V. Gurvich, L. Khachiyan and K.Makino, Generating Weighted Transversals of a Hypergraph, DIMACS Technical Report 00-17, Rutgers University, 2000. (http://dimacs.rutgers.edu/TechnicalReports/2000.html)
8. S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generaliing association rules to correlations. In: *Proceedings of the 1997 ACM-SIGMOD Conference on Management of Data*, pp. 265-276.
9. S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In: *Proceedings of the 1997 ACM-SIGMOD Conference on Management of Data*, pp. 255-264.
10. G. Dong and J. Li. Efficient mining of emerging patterns. In: *Proceeding of the 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 43-52.
11. T. Eiter and G. Gottlob, Identifying the minimal transversals of a hypergraph and related problems, *SIAM Journal on Computing*, 24 (1995) 1278-1304.
12. D. Eppstein, Arboricity and bipartite subgraph listing algorithms, *Information Processing Letters* **51** (1994), pp. 207-211.
13. J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, In: *Proceedings of the 2000 ACM-SIGMOD Conference on Management of Data*, pp. 1-12.
14. M. L. Fredman and L. Khachiyan, On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithms*, 21 (1996) 618-628.
15. M. R. Garey and D. S. Johnson, *Computers and Intractability*, Freeman, New York, 1979.
16. D. Gunopulos, R. Khardon, H. Mannila, and H. Toivonen, Data mining, hypergraph transversals and machine learning. In: *Proceedings of the 16th ACM-SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, (1997) pp. 12-15.
17. V. Gurvich and L. Khachiyan, On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions, *Discrete Applied Mathematics*, 1996-97, issue 1-3, (1999) 363-373.
18. S. O. Kuznetsov, Interpretation on graphs and complexity characteristics of a search for specific patterns, *Nauchn. Tekh. Inf., Ser. 2 (Automatic Document. Math. Linguist.)* **23**(1), (1989) pp. 23-37.
19. D. Lin and Z.M. Kedem. Pincer-search: a new algorithm for discovering the maximum frequent set. In: *Proceedings of the Sixth European Conference on Extending Database Technology*, to appear.
20. K. Makino and T. Ibaraki, Inner-core and outer-core functions of partially defined Boolean functions, *Discrete Applied Mathematics*, 1996-97, issue 1-3 (1999), 307-326.

21. H. Mannila and H. Toivonen, Multiple uses of frequent sets and condensed representations. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, (1996) pp. 189-194.
22. H. Mannila and H. Toivonen, Levelwise search and borders of theories in knowledge discovery. Series of Publications C C-1997-8, University of Helsinki, Department of Computer Science (1997).
23. H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1 (1997), 259-289.
24. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Discovering frequent closed itemsets for association rules. *Proc. of the 7th ICDT Conference*, Jerusalem, Israel, January 10-12, 1999; *Lecture Notes in Computer Science*, **1540**, pp. 398-416, Springer Verlag, 1999.
25. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Closed Set Based Discovery of Small Covers for Association Rules, *Proc. 15emes Journees Bases de Donnees Avancees, BDA*, pp. 361-381, 1999.
26. R. H. Sloan, K. Takata, G. Turan, On frequent sets of Boolean matrices, *Annals of Mathematics and Artificial Intelligence* 24 (1998) 1-4.
27. M.J. Zaki and M. Ogihara, Theoretical foundations of association rules, *3rd SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, June 1998.