Number of Frequent Patterns in Random Databases

Loïck Lhote

To cite this version:

Loïck Lhote. Number of Frequent Patterns in Random Databases. Advances in Data Analysis, pp.33 - 45, 2009, <10.1007/978-0-8176-4799-5_4>. <hal-01082026>

HAL Id: hal-01082026
https://hal.archives-ouvertes.fr/hal-01082026

Submitted on 12 Nov 2014

Number of frequent patterns in random databases

Loïck Lhote

GREYC, CNRS UMR 6072, Université de Caen
F-14032 Caen, France
e-mail: `loick.lhote@info.unicaen.fr`

Abstract In a tabular database, patterns which occur over a frequency threshold are called frequent patterns. They are central in numerous data processes and various efficient algorithms were recently designed for mining them. Unfortunately, very few is known about the real difficulty of this mining, which is closely related to the number of frequent patterns. The worst case analysis always leads to an exponential number of frequent patterns, but experimentations show that algorithms become efficient for reasonable frequency thresholds. We perform here a probabilistic analysis of the number of frequent patterns. We first introduce a general model of random databases that encompasses all the previous classical models. In this model, the rows of the database are seen as independent words generated by the same probabilistic source [i.e. a random process that emits symbols]. Under natural conditions on the source, the average number of frequent patterns is studied for various frequency thresholds. Then, we exhibit a large class of sources, the class of dynamical sources, which is proven to satisfy our general conditions. This finally shows that our results hold in a quite general context of random databases.

Keywords: Data mining, Models of databases, Frequent patterns, Probabilistic analysis, Dynamical sources.

1 Introduction

Data mining, which applies to various fields (Astronomy, Fraud detection, Marketing, ...), aims extracting a new knowledge from large databases. We consider here tabular databases where a knowledge is represented by a collection of columns, also called a pattern.

Patterns which occur frequently at the same time in several rows are of great interest since they indicate a correlation between the columns that compose the pattern. A pattern is said frequent if it occurs over a frequency threshold, which is defined by users. Frequent patterns intervene in numerous data processes such as classification or clustering [Goethals, 2003]. They are also essential [Agrawal *et al.*, 1993] for generating the well known association rules that apply in Bioinformatic, Physics, Geography, ...

The frequent pattern mining problem was first described in [Agrawal *et al.*, 1993]; during the last decade, several algorithms have been designed to solve it [Agrawal *et al.*, 1996] [Savasere *et al.*, 1995]

[Toivonen, 1996] [Han *et al.*, 2000] [Zaki, 2000]. Their complexities are closely related to the number of frequent patterns. Whereas the worst-case analysis leads to a number of patterns which is always exponential w.r.t the number of columns, the actual behaviour appears to be quite different. The algorithms fail when the frequency threshold is too small, which suggests an exponential behaviour; however, they become efficient for reasonable frequency thresholds, which suggests a polynomial behaviour. There already exist bounds for the number of frequent patterns in [Geerts *et al.*, 2001], but they are involved and do not elucidate the influence of the frequency threshold on the number of frequent patterns.

In this article, we perform a probabilistic analysis which elucidates the real behaviour of the number of frequent patterns. There already exist such analyses for frequent pattern mining, dealing with the maximal size of the patterns [Agrawal *et al.*, 1996], or the fail rate of Apriori algorithm [Purdom *et al.*, 2004]. But the previous analyses dealt with a model based on the column independence, whereas the algorithms are precisely designed for searching correlations between columns. We introduce a general model of random databases which avoids this contradiction. Here, the rows of the database are independent words generated by the same source. A source is a probabilistic process that emits a symbol at each unit time, and the complete process builds a word. Under natural conditions on words produced by this source [Conditions 1 and 2-γ], we obtain two main results [Theorems 1 and 2] on the number of frequent patterns in two main cases: the first one is related to a fixed frequency threshold, whereas the second one deals with a linear frequency threshold [w.r.t the number of rows]. We then describe a large class of sources, called *dynamical sources*, which are proven to satisfy Conditions 1 and 2-γ [Theorem 3]. This class contains all the classical sources (memoryless sources and Markov chains), but also many other sources which may possesss a higher degree of correlations. It then follows that Theorem 1 and Theorem 2 apply to various models of databases (classical or not).

2 Model of databases

2.1 Frequent pattern mining

Frequent pattern mining is often described in the framework of *market basket analysis*, but we adopt here the more general framework of *multiple-choice questionnaire.* In this context, a set of persons (of cardinality n) answers to a number m of multiple-choice questions. The set $\mathcal{E}$ of possible answers to each question is the same, and is called the alphabet. The word of $\mathcal{E}^m$ formed by the answers of one person to all the questions is called a transaction. A natural data structure for storing all the transactions is a $n \times m$ matrix over $\mathcal{E}$.

A pattern is a set of pairs $(\texttt{question}, \texttt{answer})$ where each question appears at most once. Figure 1 gives instances of patterns in a database. A person p supports a pattern X if her transaction contains the pattern X.

	Questions						
persons	q_1	q_2	q_3	q_4	q_5	q_6	q_7
p_1	2	1	2	1	2	1	1
p_2	1	2	2	1	2	1	3
p_3	2	3	2	1	2	1	1
p_4	2	1	3	2	1	2	1

Pattern	Support	Frequency
$(q_1, 2), (q_3, 2)$	p_1, p_3	2
$(q_4, 1), (q_7, 3)$	p_2	1
$(q_5, 2)$	p_1, p_2, p_3	3

Figure1. On the left, an instance of database with 7 questions and 4 persons whose answers to the questionnaire belong to $\mathcal{E} = \{1, 2, 3\}$. On the right, instances of patterns with the associated support and frequency.

The support of a pattern X is the set of persons that support X, and the frequency of X is the size of its support. Figure 1 gives instances of patterns with their support and frequency.

A pattern is said γ-frequent in $\mathcal{B}$, with a frequency threshold $\gamma \geq 1$, if the cardinality of its support is greater than γ. In table of Figure 1, the pattern $(q_5, 2)$ is $0, 1, 2$ or 3-frequent since its frequency is 3. When the database contains at least γ copies of each possible transaction (this means that $n \geq \gamma \cdot |\mathcal{E}|^m$), all possible patterns are γ-frequent. in this case, the number of frequent patterns equals $(1 + |\mathcal{E}|)^m - 1$ [for m questions]. Now, if the matrix coefficients are all equal to v, all the patterns which contain $(\texttt{question}, v)$ are frequent [for any frequency threshold]. In this case, the number of frequent patterns equals to $2^m - 1$. In particular, it is always at least exponential (in the worst-case).

2.2 Model of random databases

Our model considers all the transactions as different words produced by the same probabilistic source defined on the alphabet $\mathcal{E}$. For instance, the word associated to the first transaction [or row or person] in Figure 1 is 2121211. Since frequent patterns aim describing correlations between questions, we always suppose that the transactions are independent, even if the persons themselves may not be independent. Finally, we are interested in asymptotics when the databases become large, with a number of persons and a number of questions which are polynomially related. The next definition summaries these three hypotheses.

Definition 1. We call *random database* a probabilistic database that satisfies the three following conditions: (i) each transaction is a word produced by the same probabilistic source over an alphabet $\mathcal{E}$, (ii) the transactions form a family of independent random variables, (iii) the number n of persons and the number m of questions are polynomially related, namely of the form $\log n = \Theta(\log m)$.

3 Main results

We study the average number of frequent patterns in random databases [Definition 1] for two types of frequency thresholds: the linear frequency threshold and the constant threshold [w.r.t n]. A general result which would hold for all existing source is certainly unexpected. This is why we introduce a natural condition on the source for each frequency threshold. In the whole section, m and n respectively denote the number of questions and persons in a *random* database $\mathcal{B}$.

3.1 Linear frequency threshold

A frequency threshold γ is said to be linear if it satisfies, $\gamma \sim r \cdot n$ (for some $r \in]0,1[$) as n tends to infinity. The probability that a person, or equivalently a word, supports the pattern X is noted p_X. The quantities p_X are essential in our different conditions. One has clearly $p_Y \leq p_X$ as soon as $X \subseteq Y$. The next condition considers sources whose pattern probability is exponentially decreasing with the size of the pattern.

Condition 1 *There exist $M > 0$ and $\theta \in]0,1[$ such that for any pattern X, the probability p_X satisfies:* $\quad p_X \leq M \cdot \theta^{|X|}$.

In practice, Condition 1 implies that questions discriminate persons. In the sequel, we will prove that various (classical) sources satisfy Condition 1.

Theorem 1. *Let $\mathcal{B}$ be a random database with parameters (n, m) generated by a probabilistic source that satisfies Condition 1 with parameters M and θ. For a linear frequency threshold $\gamma \sim r \cdot n$, the average number $F_{\gamma,m,n}$ of γ-frequent patterns is polynomial w.r.t the number m of questions,*

$$F_{\gamma,m,n} = O\left(m^{j_0}\right) \qquad \textit{with} \quad j_0 = \max\{j \geq 0 \mid M\theta^j \geq r\}$$

This polynomial behaviour explains the efficiency of the algorithms for reasonable frequency thresholds. It is also possible to obtain an estimate of $F_{\gamma,m,n}$ under the weaker condition $(1-\theta) \cdot \min(m,n) \to \infty$, but, in this case, the asymptotic behaviour is no longer polynomial w.r.t m.

3.2 Constant frequency threshold

Here, the frequency threshold γ is now constant. Given γ random transactions over m questions, the probability that γ transactions support X is p_X^γ. Hence, the average number of patterns supported by the γ transactions is

$$\Sigma_{\gamma,m} = \sum_X p_X^\gamma.$$

The sum $\Sigma_{\gamma,m}$ is proven to be greater than 1. It admits a closed form for various (classical) sources [see sections 4 and 5]. The next condition implies that, for γ constant, $\Sigma_{\gamma,m}$ is exponential w.r.t the number m of questions, :

Condition 2-γ *There exists $\theta_\gamma > 1$ such that, for large m,*

$$\Sigma_{\gamma,m} > \theta_\gamma^m \cdot \Sigma_{\gamma+1,m}$$

With Condition 2-γ, we prove our second main result.

Theorem 2. *Fix $\gamma \in \mathbb{N}^\star$ and consider a random database $\mathcal{B}$ generated by a probabilistic source that satisfies Condition 2-γ with parameter θ_γ. The mean number of γ-frequent patterns verifies*

$$F_{\gamma,m,n} = \binom{n}{\gamma} \Sigma_{\gamma,m} \cdot \left[1 + n \cdot O\left(\frac{1}{\theta_\gamma^m} \cdot\right)\right].$$

In other words, for a constant frequency threshold, the number of frequent patterns is exponential w.r.t the number m of questions, and polynomial w.r.t the number n of persons. This result explains why the algorithms fail for small frequency thresholds.

3.3 Sketch of proofs

For a given frequency threshold γ, the average number of frequent patterns is the sum over all possible patterns X and all possible supports S, of the probability that X has support S. Now, the size of the support of X follows a binomial law with parameter p_X, so that

$$F_{\gamma,m,n} = \sum_X F_{\gamma,m,n,X} \qquad \text{with} \quad F_{\gamma,m,n,X} := \sum_{k=\gamma}^{n} \binom{n}{k} p_X^k (1-p_X)^{n-k}.$$

The fundamental step transforms $F_{\gamma,m,n,X}$ into an integral. Developing $(1-p_X)^k$, doing a change of variable, inverting two signs sum and using a recurrence lead to the alternative formula

$$F_{\gamma,m,n,X} = \gamma \binom{n}{\gamma} \int_0^{p_X} t^{\gamma-1}(1-t)^{n-\gamma} dt.$$

The proofs for constant and linear thresholds separate here.
For a constant threshold, we use the bounds $1-(n-\gamma)t < (1-t)^{n-\gamma} < 1$ and get a lower bound of $F_{\gamma,m,n}$ that involves the sums $\Sigma_{m,\gamma}$ and $\Sigma_{m,\gamma+1}$, whereas the upper bound only involves $\Sigma_{m,\gamma}$. Condition 2-γ is then used to conclude.
For a linear threshold $\gamma \sim r \cdot n$, we prove that $F_{\gamma,m,n,X}$ tends to 0 if $p_X < r - \epsilon$ for some positive ϵ (with an explicit error term). Otherwise, it is bounded by 1. Hence, the sum $F_{\gamma,m,n}$ only involves patterns with probability greater than $r - \epsilon$ and Condition 1 ensures that the number of such patterns is at most polynomial.

4 Dynamical databases

The results of the previous section are valid for any database generated by any source, provided that it satisfies Conditions 1 and 2-γ. In this paper, we will prove that a large class of sources satisfy these Conditions. We now present this class, formed by a large subset of dynamical sources introduced by Brigitte Vallée in [Vallée, 2001], and further used in [Clément *et al.*, 2001] [Bourdon, 2001] [Bourdon *et al.*, 2001] The model of dynamical sources gathers classical sources as the Bernoulli sources or the Markov chains, as well as more correlated ones. It is sufficiently general and can be yet precisely analysed. This class is then a good candidate for generating general databases, that we call *Dynamical databases.* We prove the following.

Theorem 3. *A [Markovian and irreducible] dynamical source satisfies condition 1 and, for all* $\gamma \geq 1$*, condition 2-*γ*, with* $\Sigma_{\gamma,m}$ *of the form:*
$\Sigma_{m,\gamma} = \kappa_\gamma \cdot \lambda_\gamma^m (1 + O(\theta_\gamma^{-m})), \quad \kappa_\gamma > 0,\ \lambda_\gamma > 1,\ \theta_\gamma > 1$ *and* $\lambda_\gamma > \lambda_{\gamma+1}$.
In particular, Theorems 1 and 2 hold for [Markovian and irreducible] dynamical databases.

4.1 Dynamical sources

A dynamical source is defined by six elements: (i) an interval I, (ii) an alphabet $\mathcal{E}$, (iii) a topological partition $(I_\alpha)_{\alpha\in\mathcal{E}}$ of I [i.e., $\alpha \neq \beta \Rightarrow I_\alpha \cap I_\beta = \emptyset$ and $\cup_\alpha \overline{I}_\alpha = I$], (iv) a coding function $\sigma : I \to \mathcal{E}$ such that $\sigma(I_\alpha) = \alpha$, (v) a shift function T on I, of class C^2, strictly monotone on each interval I_α, and strictly expansive [namely $|T'| > \rho^{-1} > 1$ on I], (vi) an initial density f_0 on I.
Figure 2 describes some instances of dynamical sources. A dynamical source emits symbols in the following way: (i) first a random real x is chosen in I according to the initial density f_0, (ii) then, the emitted symbol at the i-th step is the symbol associated to the interval that contains the i-th iterate of x [$\alpha_i = \sigma(T^i x)$], so that the (infinite) word $\mathcal{M}(x)$ produced by the source is $\mathcal{M}(x) := \alpha_1\alpha_2\ldots$.
A dynamical source is similar to a pseudo-random generator, where a probabilistic seed is used to initialise the process, which is, after this random choice, completely deterministic.

There exist several types of dynamical sources according to the geometric or analytic properties of T. The simplest family occurs when T is affine and surjective on each interval of the partition. Such sources models the classical memoryless sources that emit symbols independently from the previous ones, but following always the same probabilistic law. When such a source is used for generating a database, the questions are not correlated. Figure 2 gives an example of Bernoulli source.

In order to introduce some correlations between questions, we first consider sources with *bounded memory*, such as Markov chains. A Markov chain emits a new symbol according to a constant probabilistic law that depends on a bounded number of previous symbols. Used to generate databases, it entails that close questions are correlated. A Markov chain is a particular dynamical source. In this case, T is piecewise affine and the image of an interval of the partition is the union of intervals of the partition. Figure 2 gives an instance of Markov chain.

In this article, we deal with more general sources, called Markovian dynamical sources. A Markovian dynamical source has the same geometry as a Markov chain [the image by T of the interval I_α is a union of such intervals], but the shift function is not necessary affine. Moreover, we suppose that the process is irreducible, i.e., the matrix $M = (m_{\alpha,\beta})$ with $m_{\alpha,\beta} = 1$ if $T(I_\beta) \cap I_\alpha \neq \emptyset$, and $m_{\alpha,\beta} = 0$ elsewhere, satisfies $M^k > 0$ for some positive integer k. Figure 2 presents a Markovian source. More general dynamical sources will be not used in this article.

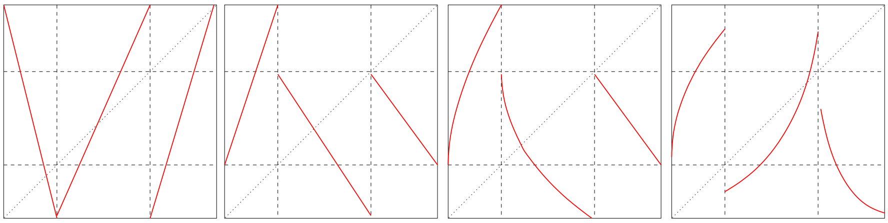

Figure2. Instances of dynamical sources (without the initial density). From left to right: a Bernoulli source, a Markov chain, a Markovian dynamical source, a general dynamical source.

4.2 Idea of proof

The main tool for analysing a dynamical source is the transfer operator. It generalizes the density transformer $\mathbf{G}$. The density transformer $\mathbf{G}$ describes the evolution of the density: one begins with some density f_0, and, after one iteration of the shift T, the new density is $f_1 = \mathbf{G}[f_0]$. We consider here the constrained operators $\mathbf{G}_\mathcal{F}$, relative to $\mathcal{F} \subset \mathcal{E}$, which are used to generate the probabilities p_X.

The set of words (or transactions) that support a pattern X is of the form $\mathcal{E}_1 \cdot \mathcal{E}_2 \cdot \ldots \cdot \mathcal{E}_m$ where $\cdot$ is the concatenation operator and $\mathcal{E}_i := \{\alpha, (q_i, \alpha) \in X\}$, and the probability p_X satisfies

$$p_X = \int_I \mathbf{G}_{\mathcal{E}_m} \circ \ldots \circ \mathbf{G}_{\mathcal{E}_1}[f_0](t)dt.$$

On a convenient functional space, the density transformer admits a unique dominant eigenvalue 1, separated from the remainder of the spectrum by a spectral gap. This spectral property, with $|T'| > \rho^{-1} > 1$ and the irreducible property entail that $p_X \leq M \cdot \rho^{m/2}$, for some positive constant M.

The sum $\Sigma_{m,\gamma}$ is the average number of patterns supported by γ random transactions. All the previous operators, defined to "describe" only one transaction, generalise for γ transactions. They give rise to operators $\mathbb{G}_{\gamma,w}$ whose m-th iterates provide an alternative expression for $\Sigma_{m,\gamma}$, namely

$$\Sigma_{m,\gamma} = \int_{I^\gamma} \mathbb{G}^m_{\gamma,2}[(f_0,\ldots,f_0)](t_1,\ldots,t_\gamma)dt_1\ldots dt_\gamma.$$

Here, $\mathbb{G}_{\gamma,w}$ is a multidimensional functional operator that admits a unique dominant eigenvalue $\lambda(\gamma,w)$, separated from the remainder of the spectrum by a spectral gap, and $\lambda(\gamma,w) > \lambda(\gamma+1,w)$. This spectral property entails a decomposition of $\mathbb{G}_{\gamma,w}$ of the form

$$\mathbb{G}^m_{\gamma,w}[F] = \lambda(\gamma,w)^m \mathbb{P}[F]\left(1+O(\theta^m_{\gamma,w})\right),$$

with $\mathbb{P}$ a projector and $\theta_{\gamma,w} < 1$. Theorem 3 follows.

5 Improved memoryless model of databases

All the existing databases are not a particuler case of Dynamical databases do not recover. Consider for instance a quite simple one, which is called the improved memoryless model. Persons and questions are independent, and each question has its own probabilistic behaviour. More precisely, the answer to the i-th question follows a Bernoulli law $B_i = (p_{i,\alpha})_{\alpha\in\mathcal{E}}$ over the alphabet $\mathcal{E}_i$, where the B_i's and the $\mathcal{E}_i$'s may be depend in the index i. In the "simple" memoryless model, used for the classical probabilistic analyses, the B_i's were the sames.

Let p denote the maximum of all the probabilities $p_{i,v}$. Since $p < 1$, the relation $p_X \le p^{|X|}$ holds and ensures Condition 1. Moreover, the sum $\Sigma_{\gamma,m}$ admits the closed formula

$$\Sigma_{\gamma,m} = \prod_{j=1}^{m}(1+\sum_{v\in\mathcal{E}} p^\gamma_{j,v})$$

and Condition 2-γ is clearly satisfied with $\theta_\gamma = (1+p/|\mathcal{E}|^\gamma)/(1+1/|\mathcal{E}|^\gamma)$.

6 Experiments

This section presents some experiments realised with classical databases of the FIMI website (Frequent Itemset Mining Implementations, `http://fimi.cs.helsinki.fi/`). In Figure 3, the plain line in the graphics represents the number of frequent patterns in function of the frequency threshold for a real database [Chess.dat] and a synthetic one [T10I4D100K.dat]. The dotted (resp. dashed) line represents the average number of frequent patterns of the simple (resp. improved) Bernoulli model naturally associated to the real database.

In the graphics, the improved model gives very good estimations whereas the simple model is quite bad. This result is not surprising for synthetic data since they have, by construction, few correlations. However, so closed results were unexpected for real life databases. The same remarks also hold for other tested databases.

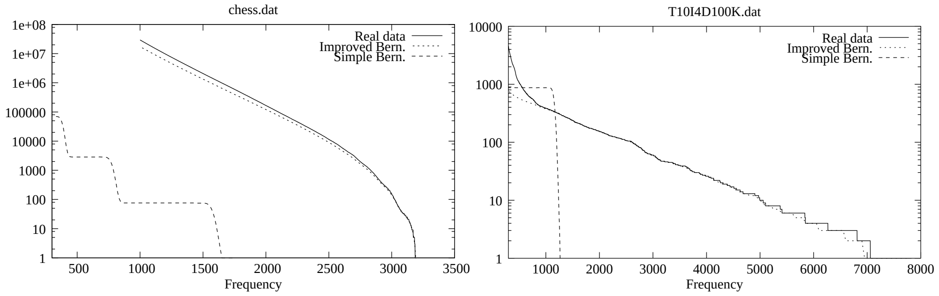

Figure3. Number of frequent patterns in function of the frequency threshold in the real database (plain line), in the associated simple Bernoulli model (dashed) and in the associated improved Bernoulli model (dotted).

References

[Agrawal *et al.*, 1993]R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, USA*, pages 207–216, 1993.

[Agrawal *et al.*, 1996]R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.

[Bourdon *et al.*, 2001]J. Bourdon, M. Nebel, and B. Vallée. On the stack-size of general tries. *Theoretical informatics ans Applications*, 35:163–185, 2001.

[Bourdon, 2001]J. Bourdon. Size and path-length of patricia tries: Dynamical sources context. *Random Structures and Algorithms*, pages 289–315, 2001.

[Clément *et al.*, 2001]J. Clément, P. Flajolet, and B. Vallée. Dynamical sources in information theory: A general analysis of trie structures. *Algorithmica*, 29(1):307–369, 2001.

[Geerts *et al.*, 2001]F. Geerts, B. Goethals, and J. Van den Bussche. A tight upper bound on the number of candidate patterns. In *IEEE International Conference on Data Mining (ICDM'01), San Jose, USA*, pages 155–162, 2001.

[Goethals, 2003]B. Goethals. Survey on frequent pattern mining, 2003.

[Han *et al.*, 2000]J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. In *ACM SIGMOD International Conference on Management of Data (SIGMOD'00), Dallas, USA*, pages 1–12, 2000.

[Purdom *et al.*, 2004]Paul W. Purdom, Dirk Van Gucht, and Dennis P. Groth. Average-case performance of the apriori algorithm. *SIAM Journal on Computing*, 33(5):1223–1260, 2004.

[Savasere *et al.*, 1995]A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *VLDB'95*, 1995.

[Toivonen, 1996]Hannu Toivonen. Sampling large databases for association rules. In *International Conference on Very Large Data Bases (VLDB'96), Mumbai, India*, pages 134–145. Morgan Kaufman, 1996.

[Vallée, 2001]B. Vallée. Dynamical sources in information theory: Fundamental intervals and word prefixes. *Algorithmica*, 29:262–306, 2001.

[Zaki, 2000]Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.*, 12(2):372–390, 2000.