Noname manuscript No.
(will be inserted by the editor)

Objectively evaluating condensed representations and interestingness measures for frequent itemset mining

Albrecht Zimmermann
albrecht.zimmermann@cs.kuleuven.be

the date of receipt and acceptance should be inserted later

Abstract Itemset mining approaches, while having been studied for more than 15 years, have been evaluated only on a handful of data sets. In particular, they have never been evaluated on data sets for which the ground truth was known. Thus, it is currently unknown whether itemset mining techniques actually recover underlying patterns. Since the weakness of the algorithmically attractive support/confidence framework became apparent early on, a number of interestingness measures have been proposed. Their utility, however, has not been evaluated, except for attempts to establish congruence with expert opinions. Using an extension of the Quest generator proposed in the original itemset mining paper, we propose to evaluate these measures objectively for the first time, showing how many non-relevant patterns slip through the cracks. **result verification; data generation; interestingness measures**

1 Introduction

Frequent itemset mining (FIM) was introduced almost twenty years ago [1] and the framework has proven to be very successful. It not only spawned related approaches to mining patterns in sequentially, tree, and graph-structured data, but due to its relative simplicity it has been extended beyond the mining of supermarket baskets towards general correlation discovery between attribute value pairs, discovery of co-expressed genes, and classification rules, etc.

The original framework used frequency of itemsets in the data (support) as a significance criterion – often occurring itemsets are assumed not to be chance occurrences – and conditional probability of the right-hand side of association rules (confidence) as a correlation criterion. This framework has clear weaknesses and other interestingness measures have been proposed in the years since the seminal paper was published [24], as well as several condensed

KU Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium

representations [15,6,10] that attempt to remove redundant information from the result set.

While each of these measures and condensed representations is well-motivated, there is as of yet no consensus about how effectively existing correlations are in fact discovered. A prime reason for this can be seen in the difficulty of evaluating the quality of data mining results. In classification or regression tasks, there is a clearly defined target value, often objectively measured or derived from expert labeling *a priori* to the mining/modeling process, that results can be compared to when assessing the goodness of fit. In clustering research, the problem is somewhat more pronounced but clusters can be evaluated w.r.t. intra-cluster similarity and inter-clusters dissimilarity, knowledge about predefined groups might be available, e.g. by equating them with underlying classes, and last but not least there exist generators for artificial data [18]. In FIM, in contrast, while the seminal paper introduced a data generator as well, that data generator has been used only for efficiency estimations and fell furthermore into some disregard after Zheng *et al.* showed that the data it generated had characteristics that were not in line with real-life data sets [28]. The current, rather small collection of benchmark sets, hosted at the FIMI repository [2], consists of data sets whose underlying patterns are unknown. As an alternative, patterns mined using different measures have been shown to human "domain experts" who were asked to assess their interestingness [8]. Given humans' tendency to see patterns where none occur, insights gained from this approach might be limited.

Interestingly enough, however, the Quest generator proposed by Agrawal and Srikant already includes everything needed to perform such assessments: it generates data by embedding source itemsets, making it possible to check mining results against a gold standard of predefined patterns. Other data generation methods proposed since [20,9,23,3] do not use clearly defined patterns and can therefore not be used for this kind of analysis. The contribution of this work is that we repurpose the Quest generator accordingly and address this open questions for the first time:

- How effective are different condensed representations and interestingness measures in recovering embedded source itemsets?

In the next section, we introduce the basics of the FIM setting, and discuss different interestingness measures. In Section 3, we describe the parameters of the Quest generator and its data generation process. Equipped with this information, we can discuss related work in Section 4, placing our contribution into context and motivating it further. Following this, we report on an experimental evaluation of pattern recovery in Section 5, before we conclude in Section 6.

2 The FIM setting

We employ the usual notations in that we assume a collection of *items* $\mathcal{I} = \{i_1, \ldots, i_N\}$, and call a set of items $I \subseteq \mathcal{I}$ an itemset, of size $|I|$. In the same manner, we refer to a transaction $t \subseteq \mathcal{I}$ of size $|t|$, and a data set $\mathcal{T} \subseteq 2^{\mathcal{I}}$ of size $|\mathcal{T}|$. An itemset I matches (or is supported by) a transaction t iff $I \subseteq t$, the support of I is $sup(I, \mathcal{T}) = |\{t \in \mathcal{T} \mid I \subseteq t\}|$, and its relative support or frequency $freq(I, \mathcal{T}) = \frac{sup(I,\mathcal{T})}{|\mathcal{T}|}$. An itemset I is *frequent* for a given threshold $\theta \in \mathbb{N}$, if $sup(I, \mathcal{T}) \geq \theta$. The *confidence* of an association rule $X \Rightarrow Y$, formed of two itemsets $X, Y \subset \mathcal{I}, X \cap Y = \emptyset$, is calculated as $conf(X \Rightarrow Y, \mathcal{T}) = \frac{sup(X \cup Y, \mathcal{T})}{sup(X, \mathcal{T})}$. When the context makes it clear which data set is referred to, we drop $\mathcal{T}$ from the notation.

Condensed representations The set of all frequent itemsets can be summarized by so-called *condensed representations*, subsets that include enough information to enumerate all frequent itemsets or even to derive their support.

An itemset I is a *closed itemset* iff I is an itemset and $\forall i \in \mathcal{I}, i \notin I : sup(I \cup i, \mathcal{T}) < sup(I, \mathcal{T})$. An itemset I is a *maximal frequent itemset* for given minimum support threshold $\theta \in \mathbb{N}$ iff $sup(I, \mathcal{T}) \geq \theta \wedge \forall I' \supset I : sup(I', \mathcal{T}) < \theta$.

2.1 Interestingness measures

The support/confidence framework has at least one major drawback in that it ignores prior probabilities. Assume, for instance, two items i_1, i_2 with $freq(i_1) = 0.6$, $freq(i_2) = 0.8$. While $freq(i_1, i_2) = 0.48$ would often denote the itemset as a high-frequency itemset, it is in fact exactly what would be expected given independence of the two items. Similarly, $conf(i_1 \Rightarrow i_2) = 0.8$, while clearly a high confidence value, would also indicate independence when compared to the prior frequency of i_2. Therefore, numerous other measures have been proposed to address these shortcoming [24].

Most of them have been proposed for assessing the quality of association rules, meaning that they relate two binary variables. Generally speaking, it is possible to use such measures more generally to assess the quality of itemsets in the following way. Given an interestingness measure $m : 2^{\mathcal{I}} \times 2^{\mathcal{I}} \mapsto \mathbb{R}$, itemset I, we can take the minimal value over all possible association rules with a single item in the right-hand side (RHS): $\min_{i \in I}\{m(I \setminus i \Rightarrow i)\}$, if the measure is to be maximized, or the maximum value in the opposite case. Several studies [21, 22, 13] have explored association rule measures w.r.t. their theoretical properties and empirical behavior and identified groups of measures that behave the same or at least very similar. Based on those groupings, we have selected the following measures:

1. Piatetsky-Shapiro (PS)

$$PS(X \Rightarrow Y) = freq(X \cup Y) - freq(X)freq(Y)$$

2. Confidence (Conf)
   See above

3. Least Contradiction (LC)

$$LC(X \Rightarrow Y) = \frac{sup(X \cup Y) - (sup(X) - sup(X \cup Y))}{sup(Y)}$$

4. Jaccard (J)

$$J(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X) + sup(Y) - sup(X \cup Y)}$$

5. J-Measure (JM)

$$JM(X \Rightarrow Y) = \max\Big\{ \frac{sup(X\cup Y)}{|\mathcal{T}|} \log_2(\frac{sup(X\cup Y)|\mathcal{T}|}{sup(X)sup(Y)}) + \frac{sup(X)-sup(X\cup Y)}{|\mathcal{T}|} \log_2(\frac{(sup(X)-sup(X\cup Y))|\mathcal{T}|}{sup(X)(|\mathcal{T}|-sup(Y))}), \frac{sup(X\cup Y)}{|\mathcal{T}|} \log_2(\frac{sup(X\cup Y)|\mathcal{T}|}{sup(X)sup(Y)}) + \frac{sup(Y)-sup(X\cup Y)}{|\mathcal{T}|} \log_2(\frac{(sup(Y)-sup(X\cup Y))|\mathcal{T}|}{sup(Y)(|\mathcal{T}|-sup(X))}) \Big\}$$

6. Goodman-Kruskal (GK)

$$GK(X \Rightarrow Y) = \frac{\sum_{i\in X} \max_{i'\in Y} freq(i\cup i') + \sum_{i\in Y} \max_{i'\in X} freq(i\cup i') - \max_{i\in X} freq(i) - \max_{i'\in Y} freq(i')}{2 - \max_{i\in X} freq(i) - \max_{i'\in Y} freq(i')}$$

Our primary aim, however, is to test the recovery of itemsets, the precursors to association rules, and we are therefore also interested in measures that have been proposed to mine interesting *itemsets*. To make it easier to discuss those more sophisticated measures, we associate each itemset with a function $I : \mathcal{T} \mapsto \{0,1\}$, with $I(t) = 1$ iff $I \subseteq t$, which allows us to define an equivalence relation based on a collection of itemsets $\{I_1, \ldots, I_k\}$:

$$\sim_{\{I_1,\ldots,I_k\}} = \{(t_1, t_2) \in \mathcal{T} \times \mathcal{T} \mid \forall I_i : I_i(t_1) = I_i(t_2)\}$$

Using this equivalence relation, a given transaction t induces a *block*: $[t] = \{t' \in \mathcal{T} \mid t' \sim_{\{I_1,\ldots,I_k\}} t\}$. The set of all blocks is referred to as the *partition* or quotient set of $\mathcal{T}$ over $\{I_1, \ldots, I_k\}$:

$$\mathcal{T}/\sim_{\{I_1,\ldots,I_k\}} = \{[t] \mid t \in \mathcal{T}\}$$

In the succeeding discussion, we label each block $b \in \mathcal{T}/\sim_{\{I_1,\ldots,I_k\}}$ with a subscript denoting what the different itemsets evaluate to. Assuming a set of itemset $\{I_1, I_2, I_3, I_4\}$, for instance, b_{1010} contains all transactions for which itemsets I_1, I_3 evaluate to 1, and itemsets I_2, I_4 evaluate to 0.

Multi-way χ^2 Brin *et al.* proposed to use the χ^2 test to evaluate itemsets directly [7]. Each item $i \in I$ is considered its own itemset and instead of a 2×2 contingency table, a multiway table with $2^{|I|}$ cells is populated by the cardinalities of the blocks derived from $\mathcal{T}/\sim_{\{i|i\in I\}}$. The χ^2-value is calculated as in the two-dimensional case. The degrees of freedom for such a table are $df(I) = 2^{|I|} - 1 - |I|$, and if the χ^2 value exceeds a given p-value for that many df, the itemset is considered significant. Brin *et al.* also propose an *interest* measure for individuals cells: $interest(b_v) = |1 - \frac{O_v}{E_v}|$, and propose to consider the combination of item presences and absences of the cell with the highest interest value the most relevant contribution of the found itemset.

Entropy Entropy effectively evaluates the "balance" of a partition induced by an itemset, i.e. the relative size or likelihood of the blocks:

$$H(\{i \mid i \in I\}) = - \sum_{b \in \mathcal{T}/\sim_{\{i \mid i \in I\}}} \frac{|b|}{|\mathcal{T}|} \log_2(\frac{|b|}{|\mathcal{T}|})$$

The entropy is highest, equal to $|I|$, if all blocks are equally likely, and 0 if there is only one block. Heikinheimo *et al.* proposed mining low-entropy itemsets [12].

Maximum Entropy evaluation Not a measure *per se*, De Bie has proposed to use maximum entropy models to sample data sets conforming to certain constraints derived from $\mathcal{T}$, e.g. row and column margins, i.e. support of individual items and sizes of transactions, in the expectation [3]. Found patterns can be reevaluated on these databases and rejected if they occur in more than a certain proportion.

3 The Quest generator

The Quest generator was introduced in the paper that jump-started the area of frequent itemset mining (FIM), and arguably the entire pattern mining field [1]. The generative process is governed by a number of parameters:

- L – the number of potentially large itemsets (source itemsets) in the data.
- N – the number of items from which source itemsets can be assembled.
- $|I|$ – the average size of source itemsets.
- $|t|$ – the average size of transactions in the data.
- $|\mathcal{T}|$ – the cardinality of the data set.
- c – the "correlation level" between successive itemsets.

The generator proceeds in two phases: it *first* generates all source itemsets, and in a *second* step assembles the transactions that make up the full data set from them. The authors, working in the shopping basket setting, aimed to model the phenomenon that certain items are typically bought together and several such groups of items would make up a transaction. This also means that the output of FIM operations can be compared to the source itemsets to get an impression of how well such mining operations recover the underlying patterns. We have reimplemented the generator and it can be downloaded at `http://www.scientific-data-mining.org`.

3.1 Source itemset generation

For each of the L source itemsets, the size is sampled from a *Poisson* distribution with mean $|I|$, which means that itemsets' sizes are not influenced by those of others. A fraction of the items used in the source itemset formed in iteration i is taken randomly from the itemset formed in iteration $i-1$. This fraction is sampled from an exponential distribution with mean c. This is the only step in which itemsets are influenced by others. The rest of the items are sampled uniformly from N. Each source itemset is assigned a weight, which will correspond to its probability of appearing in the data, sampled from an exponential distribution with unit mean, and a corruption level, i.e. a probability value that only the partial source itemset will be embedded into a transaction, sampled from a normal distribution with mean 0.5 and variance 0.1. Source itemsets' weights are normalized so that they sum to 1.0.

3.2 Transaction generation

For each of the $|\mathcal{T}|$ transactions, the size is sampled from a Poisson distribution with mean $|t|$. Source itemsets to be embedded into the transaction are chosen according to their weight, and their items embedded according to their corruption level. Importantly, this means that source itemsets are selected independently from each other. If the number of items to be embedded exceeds the remaining size of the transaction, half the time the items are embedded anyway, and the transaction made larger, in the other half the transaction is made smaller, and the items transferred for embedding into the succeeding transaction.

4 Related work

The seminal paper on FIM, which also introduced the Quest generator, was published almost twenty years ago [1]. The authors used the generator to systematically explore the effects of data characteristics on their proposed algorithm, using several different transaction and source itemset sizes, evaluating a number of values for $|\mathcal{T}|$ (9 values), N (5 values), and $|t|$ (6 values) while keeping the other parameters fixed, respectively, specifically the number of source itemsets L. It is unclear whether more than one data set was mined for each setting, an important question given the probabilistic nature of the correlation, corruption, and source itemset weight effects.

A similar kind of systematic evaluation can still be found in [26], although the authors did not evaluate different values for N (and also continue to keep L fixed throughout). The evaluation found in [11], however, already limits itself to only two Quest-generated data sets. In line with this trend, the authors of [25] used four Quest-generated data sets which they augmented by three UCI data sets [4], and PUSMB data sets that act as stand-ins for "dense" data, i.e. sets with few items coupled with large transaction sizes. The evaluation reported in [17] uses one artificial data set, one UCI data set, and the PUSMB data.

The systematic use of the Quest generator came to a virtual halt after Zheng *et al.* reported that one of the Quest-generated data sets shows different characteristics from real-life data and that algorithmic improvements reported in the literature did not transfer to real-life data [28]. Notably, the authors pointed out that CLOSET [17] scales worse than CHARM [25], a result that Zaki *et al.* verified in revisiting their work and comparing against CLOSET as well [27], and that runs contrary to the experimental evidence presented in [17] by the authors of CLOSET, probably due to the difference in used data sets.

The typical evaluation of FIM related approaches afterwards consisted of using two Quest-generated data sets, a number of UCI data sets, and the real-life data sets made available to the community, e.g. in the Frequent Itemset Mining Implementation competitions [2]. This has led to the paradoxical situation that while techniques for FIM have proliferated, the amount of data sets on which they have been evaluated has shrunk, in addition to a lack of control over these data sets' characteristics. Also, all evaluations limited themselves to evaluating efficiency questions.

In the same period, data sets begun to be characterized by the distribution of the patterns mined from them, see e.g. [20]. These analyses have given rise to techniques for "'inverse itemset mining" that, starting from FIM results, generate data leading to the same distribution of mined itemsets. While these data sets could be used for efficiency evaluations, they are dependent on the data from which patterns are mined in the first place, and the lack of clearly defined patterns prevents quality evaluations. In a similar vein falls the generator proposed in [23] which uses the MDL principle to generate data sets that will lead to similar itemsets mined, even though it serves a different purpose, namely to protect the anonymity of original data sources.

Finally, FIM research has spawned a large number of interestingness measures and literature discussing what desirable characteristics of such measures are and how similar their practical performance is [21,22,13,24]. Tan *et al.* [21] performed two types of analysis on 21 different interestingness measures: 1) they analyzed each measure in terms of the properties identified as desirable by Piatetsky-Shapiro [19] and additional properties they identified themselves, 2) they randomly generated contingency tables and compared the similarity of different measures' rankings. They identify five (non-disjunct) groups in this manner. Vaillant *et al.* [22] performed similar analyses, grouping twenty measures, only some of which had been considered by Tan *et al.*, empirically on ten UCI data sets, and comparing their results with a grouping derived from several formal properties. They identify four groups and show that the empirical and theoretical groups do not fully agree. They reprised their work in [13], slightly revising the set of measures, adding additional analyses and suggesting guidelines for selecting the appropriate interestingness measure. While this has made it possible to state whether some measures will have the same outcome, i.e. solution sets, it is at present unknown how those outcomes relate to the patterns underlying the data. The closest research has come to such evaluations are attempts to establish how well interestingness measures for association rules align with domain experts' interest [8].

5 Pattern recovery

The fact that the Quest generator assembles transactions in terms of source itemsets gives us the unique opportunity to compare the output of a frequent itemset mining operation to the original patterns. Note that this is different from the approach taken in [20,23] – in those works databases were generated that would *result* in the itemsets (or at least of the same number of itemsets of certain sizes) being mined that informed the generating process. Contrary to this, we cannot be sure that the output of the frequent itemset mining operation has any relation to the source patterns from which the data is generated, although we of course expect that that would be the case. To the best of our knowledge, this is the first time that such an objective comparison of mined to source itemsets has been performed.

5.1 Experimental setup

For reasons of computational efficiency, we use only few $(10, 100)$ source itemsets in our experiments. This allows us to mine with high support thresholds without having to expect missing (too many) source itemsets. We generate data with $N = 2000, |t| = 10, |I| = 4$, with corruption turned off. We generate 100 data sets for each setting and average the results over them. We used FPGrowth in Christian Borgelt's implementation with support threshold $100/L\%$, a generous threshold given that we can expect each transaction to consist on average of $10/4 = 2.5$ itemsets. This corresponds to a relatively easy setting since the source itemsets have high support and apart from the correlation-induced overlap, items are unlikely to appear in several itemsets. Mined itemsets can be of different type, they can correspond:

- one-to-one to *source itemsets*
- to *unions* of source itemsets
- to *intersections* of source itemsets
- to true *subsets* of source itemsets
- or be none of the above, in which case the itemset is considered *spurious*

We mine three kinds of patterns: frequent, closed, and maximal itemsets. While frequent itemsets will be guaranteed to include all source itemsets recoverable at the minimum support threshold, they will also include all of their subsets, and possibly additional combinations of items. Closed itemsets might miss some source itemsets if the probabilistic process of the Quest generator often groups two itemsets together while generating transactions, an effect that should not be very pronounced over 100 data sets, however. On the other hand can closed itemsets be expected to avoid finding subsets of frequent sets unless those are intersections of source itemsets, and to restrict supersets of source itemsets to unions of them. Maximal itemsets, finally, can be expected to consist of unions of source itemsets.

We also use this opportunity to assess the effect of data sizes on the output, e.g. whether more itemsets are mined or whether additional data helps remove

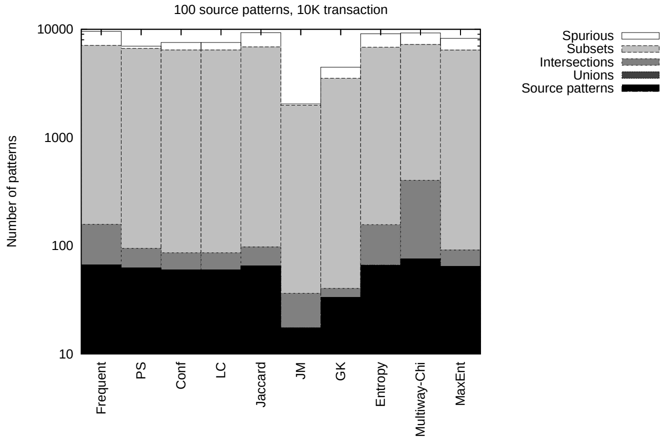

Fig. 1 Pattern types for mining all frequent itemsets for $L = 100$, no corruption

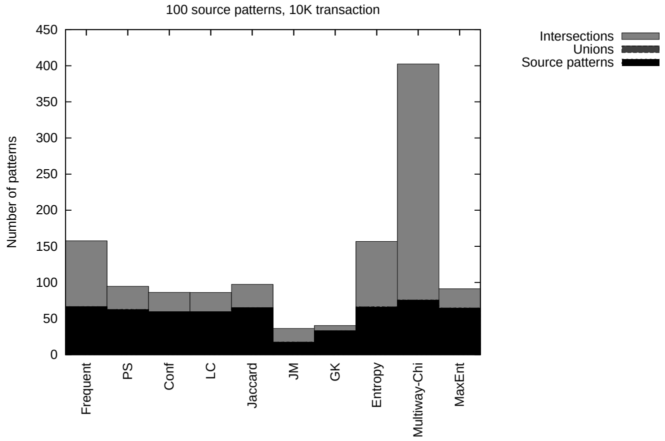

Fig. 2 Pattern types for mining frequent itemsets for $L = 100$, no corruption, without subsets and spurious itemsets

spurious patterns. To this end, we generate data with $10,000$ (10K), $100,000$ (100K), and $1,000,000$ (1M) transactions. For each pattern type, we filter using to the different measures afterwards:

a) Piatetsky-Shapiro: 0.01 since independent LHS and RHS will have a PS-score of 0.
b) Confidence: 0.5, which is a standard value used in the association rule mining literature.
c) Least Contradictions: 0.01 since the numerator will be ≤ 0 if there are at least as many counter examples as supporting examples.

d) Jaccard: it is not straight-forward how to set a threshold for the Jaccard measure since even the value for independent LHS and RHS depends on the frequency of either one. Since negatively correlating ones would get a score of 0, 0.001 would at least filter those out.
e) J-Measure: this measure measures the larger contribution to mutual entropy, which is the Shannon entropy decrease according to [5]. This is another measure for which the threshold is not straight-forward but given that independent LHS and RHS will result in no decrease at all, we set a threshold of 0.1.
f) Goodman-Kruskal: 0.01 for analogous reasons as in the case of PS.
g) Multi-way χ^2: no threshold is needed for this measure but a significance level, we choose 0.05. A high score does not always indicate that the itemset as a whole is relevant, however. To interpret selected itemsets, we use the block with the highest interest value. To give an example, if $\{i_1, i_2, i_3\}$ attains a high score but the block with highest interest is b_{101}, we interpret $\{i_1, i_3\}$ as the pattern, and i_2 as being negatively correlated with it, hence coming from a different source itemset. For this measure, we can therefore also assess how many negative correlations were identified, and how many of those correctly.
h) Entropy: there is not clear way to set a maximal threshold. We require for the entropy of itemsets to be at most half of their size.
i) Maximum entropy evaluation: we use an empirical p-value of 0.05 to reject the null hypothesis that the items in an itemset are independent from each other, i.e. an itemset must not be frequent on more than 5% of the sampled data sets. We sample 100 times from the maximum entropy model of each data set. We will therefore risk false negatives but evaluating patterns on 10000 data sets already taxes our computational resources.

These measures have different semantics, as explained in [21,22], and those works cautioned to consider the characteristics of the data mining task at hand in selecting one, or several of them. As we have explained at the beginning of Section 2.1, we apply all of them to a task that they are not necessarily well-suited for, the identification of sets of co-occurring items, and evaluate their usefulness for this task. In doing so, we also explore the view that successful prediction of an item's presence can be seen as a surrogate for the rejection of independence.

We want to understand how the different pattern types and the different interestingness measures interact with the source patterns to produce an output set. Hence we show cumulative pattern counts, i.e. the top of the bar corresponds to the total amount of itemsets mined, and different colors show the proportion of different categories of itemsets. This could be incomplete information, however: a measure could for instance have a bias for short itemsets while selecting the same *type* of itemset as another measure. We therefore also show the length distribution of itemsets in the respective output sets. Finally, the score assigned to patterns can be used to rank them and give guidance to a user in terms of which ones to inspect/consider most relevant. We therefore

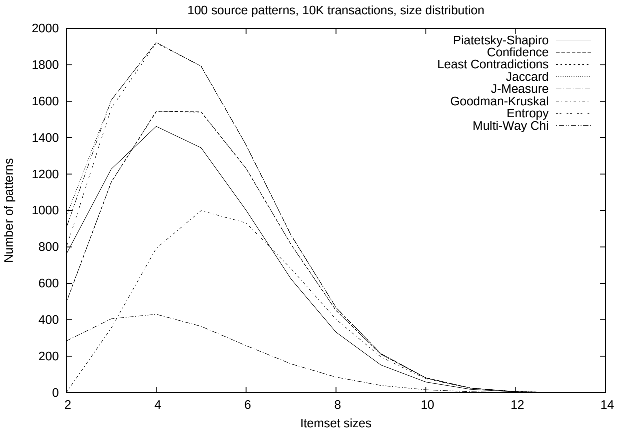

Fig. 3 Pattern size distribution for mining all frequent itemsets for $L = 100$, no corruption

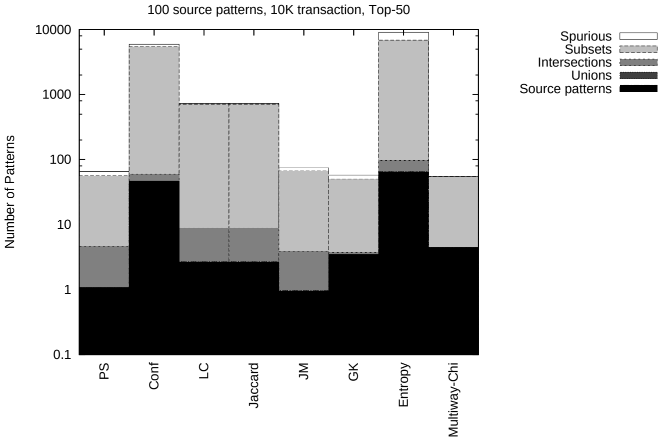

Fig. 4 Pattern types for mining the top-50 scoring frequent itemsets for $L = 100$, no corruption

select the top-50 scores for each measure, and evaluate the itemsets falling into this score interval in the same manner (types of patterns, and length distribution). If different itemsets have the same score, this set's cardinality can be larger than 50. Note that in the ideal case this should include only source itemsets for $L = 100$, while it necessarily will include other types for $L = 10$.

5.2 Data sets created without corruption of source itemsets

Due to the page limit and since the results for 10K, 100K, and 1M transactions to not differ from each other, we show only plots for 10K transactions. We also focus mainly on results for $L = 100$, since the trends for $L = 10$ remain the

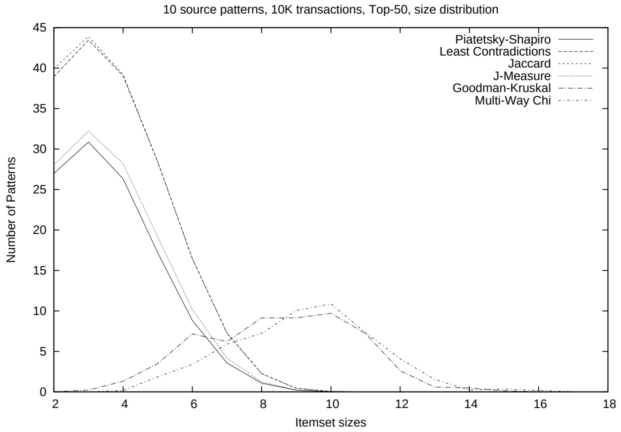

Fig. 5 Pattern size distribution for mining the top-50 scoring frequent itemsets for $L = 100$, no corruption, excluding Confidence and Entropy

same, with only some changes in the composition of the output set. The full set of plots can be downloaded at `http://www.scientific-data-mining.org`.

Frequent itemsets As Figure 1 shows, for 100 source itemsets, the vast majority of the output will be composed by subsets of those itemsets. The first column also shows quite a few spurious itemsets. While some of the interestingness measures, such as PS or JM, manage to remove most of the spurious sets, although JM does so at the expense of actual source itemsets, they cannot filter out the subsets since these consist of legitimately co-occurring items. Except for the two mentioned measures, however, all of the other measures let significant amounts of spurious itemsets pass, some filtering better (Confidence, LC, GK), some worse (Jaccard, Entropy, Multiway-Chi, MaxEnt).

Due to the large size of the output, the first figure has a log-scaled y-axis. To gain a better idea of the composition of the output, Figure 2 shows the output set with subsets and spurious sets removed.

As can be seen, the threshold is tight enough that no full unions of source itemsets appear, which unfortunately also means that quite a few source patterns are not recovered. It is interesting to see that most measures reduce the number of source itemset *intersections*, with the exception of entropy and the Multiway-Chi, which actually identifies intersections in combination with spurious items.

Figure 3 shows how itemset sizes are distributed. In particular it shows that GK has a tendency to select longer itemset, compared to the other measures.

As mentioned above (and as seen in Figure 1), using a single cut-off value still leaves too many itemsets in the output set. Figure 4 shows the effect of using the top-50 scores.

Both confidence and entropy include almost the entire set of itemsets in this score range, although confidence removes at least the spurious sets, while

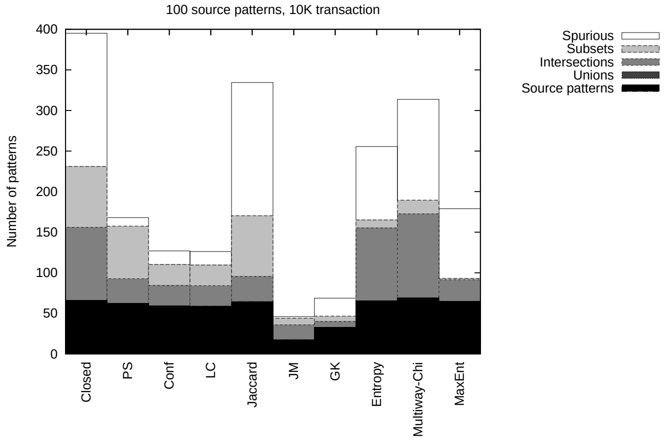

Fig. 6 Pattern types for mining all closed frequent itemsets for $L = 100$, no corruption

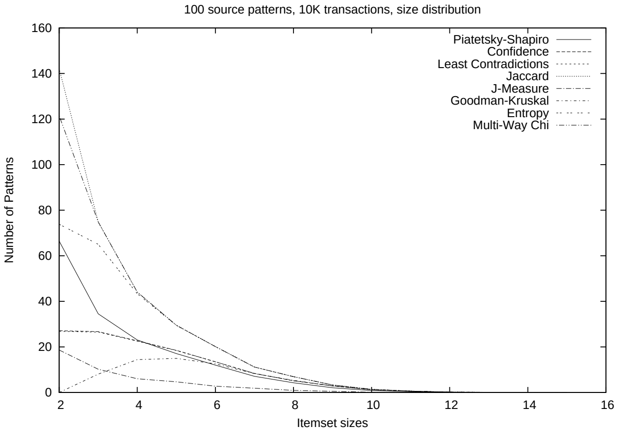

Fig. 7 Pattern size distribution for mining all closed frequent itemsets for $L = 100$, no corruption

the others return far fewer itemsets. As the figure makes clear, in most cases it is the *source itemsets* that are removed in this way, with intersections and subsets being selected. Almost no spurious patterns remain, showing that highly scoring patterns indeed consist of co-occurring items.

Figure 5 shows the size distribution of itemsets within the top-50 scores (excluding confidence and entropy since their distributions would look as in Figure 3). The majority of itemsets is relatively short, supporting the impression gained from Figure 4, with the exception of Multiway-Chi and GM that identify longer patterns as interesting. Since the Multiway-Chi score increases with the number of cells, this was to be expected.

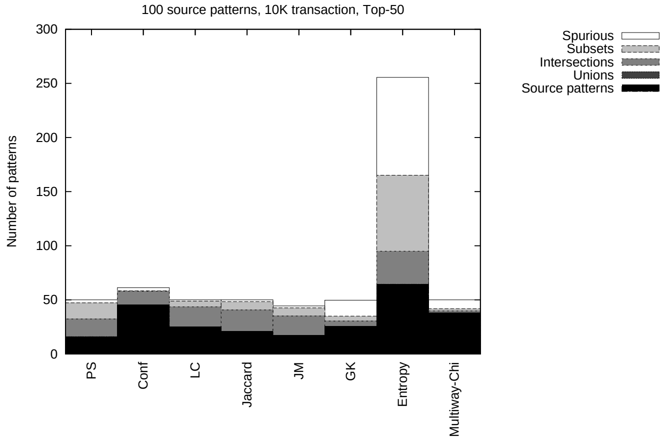

Fig. 8 Pattern types for mining the top-50 scoring closed frequent itemsets for $L = 100$, no corruption

At this point, we have to conclude tentatively that mining frequent itemsets overwhelms interestingness measures: in particular the large amount of subsets means that the result set would need quite a bit of post-processing to identify the actual patterns in the data.

Closed itemsets A straight-forward way of addressing this issue of too many subsets consists of mining closed itemsets instead, and indeed, as Figure 6 shows, the number of returned itemsets drops from almost ten thousand to a few hundred.

With many subsets removed, similar filtering trends hold as in the case of mining all frequent itemsets – in particular Jaccard and Multiway-Chi, but also Entropy and MaxEnt, are filtering spurious sets not really well. In the case of the latter three the problem seems to be that source itemsets or their subsets, when combined with spurious items, still show unexpected frequencies, appearing significant. The removal of subsets also affects the size distribution of itemsets (Figure 7).

Since there are no more subsets to overwhelm the interestingness measures, the set of itemsets having scores within the top-50 range is roughly of size 50, with the exception of entropy, as Figure 8 shows.

This plot contains the in our opinion first surprising result since the output set selected by confidence (consisting mainly of source itemsets and intersections) would actually be the most relevant. But also most of the other measures return a mix of itemsets that are related to the source itemsets, with the exception of GK, entropy, and Multiway-Chi. The latter loses its good performance, now that there are no more subsets that can be separated from spurious items with which they were combined.

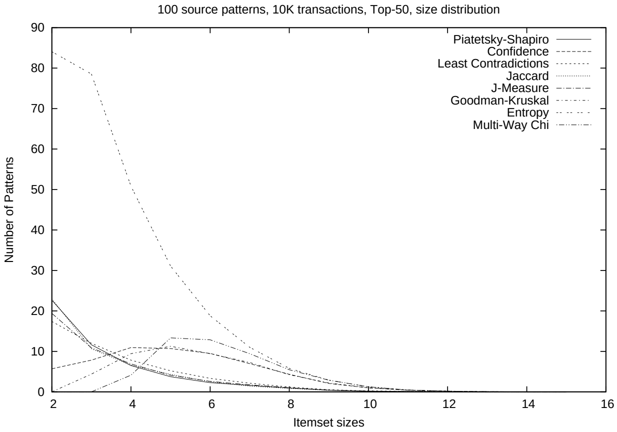

Fig. 9 Pattern size distribution for mining the top-50 scoring closed frequent itemsets for $L = 100$, no corruption

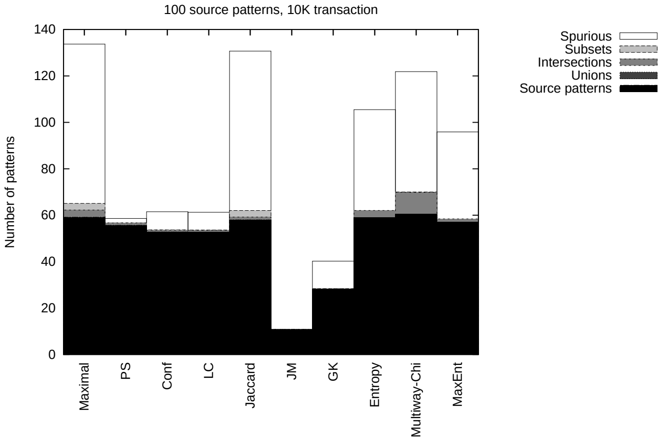

Fig. 10 Pattern types for mining all maximal frequent itemsets for $L = 100$, no corruption

Combining closed frequent itemsets with interestingness measures proves to be much more effective for recovering source itemsets and itemsets that are related to them, vindicating the theoretical considerations that led to the proposal and adoption of closed itemset mining.

Maximal itemsets An alternative to closed itemset mining consist of mining maximal itemsets. Since those can also be expected to remove intersections of source itemsets, the result set should consist mainly of source itemsets and Figure 10 seems to bear this out. The size distribution of course reflects the focus of this approach on long itemsets (Figure 11).

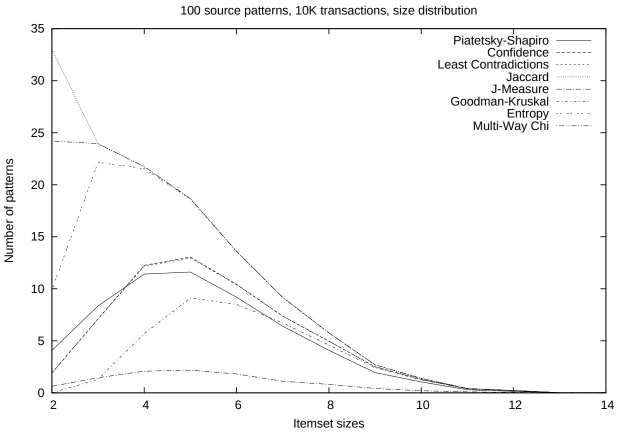

Fig. 11 Size distribution for mining all maximal frequent itemsets for $L = 100$, no corruption

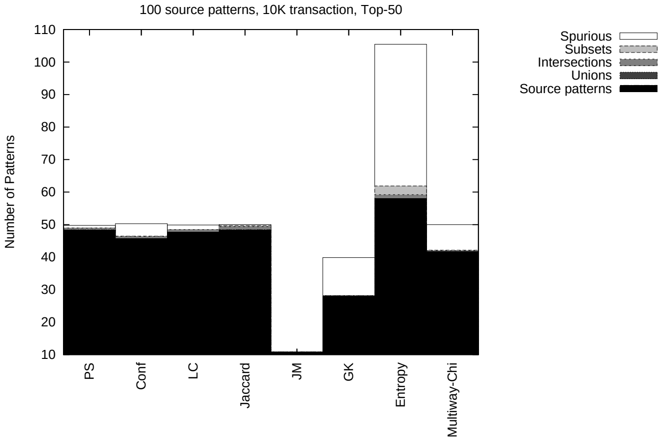

Fig. 12 Pattern types for mining the top-50 scoring maximal frequent itemsets for $L = 100$, no corruption

When selecting itemsets using the top-50 scores (Figure 12), finally, mining maximal itemsets looks like clear winner, at least if one does not use GK, Multiway-Chi, or entropy to select patterns.

The problem with this interpretation is that Figure 14, showing the result for mining maximal sets for $L = 10$, paints a different picture: while, similar to the case of $L = 100$, the full result set contains roughly the same amount of patterns as there are source itemsets, the majority of those are spurious.

The reason for this is to be found in that with 100 source patterns, $\frac{100*101}{2} =$ 5050 possible pattern combinations are possible, making it relatively unlikely that many will show up as frequent. 10 source patterns will only combine into

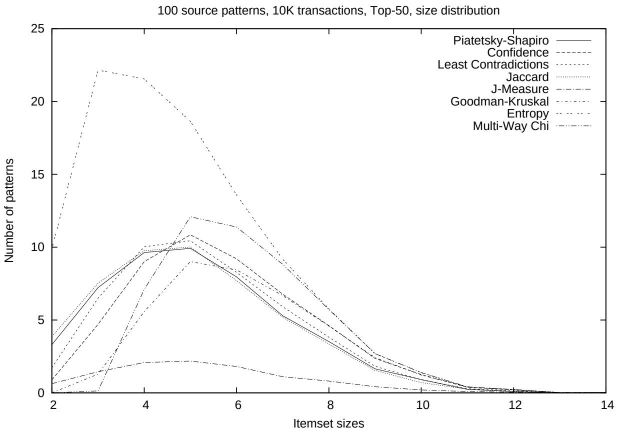

Fig. 13 Size distribution for mining the top-50 scoring maximal frequent itemsets for $L = 100$, no corruption

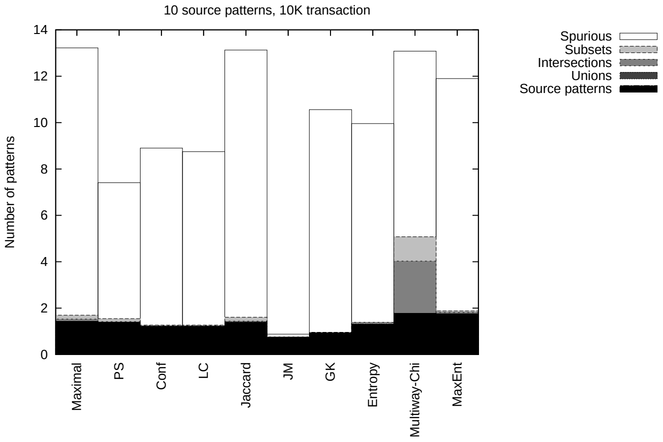

Fig. 14 Pattern types for mining all maximal frequent itemsets for $L = 10$, no corruption

55 patterns, leading to partial unions crowding out their subsets, i.e. source patterns.

Since there are fewer than 50 patterns in total, using top scores will not help in this situation, and while the set is small enough to be examined by a user, maximal pattern do not carry enough information to help with determining the relevant ones.

The number of source patterns, its interplay with the minimum support threshold, the number of highest-scoring patterns chosen, and of course thresholds for interestingness measures, will therefore have a strong effect on what kind of patterns are returned, a problem that also the use of the top-scoring closed itemsets will not fully address, as Figure 15 shows.

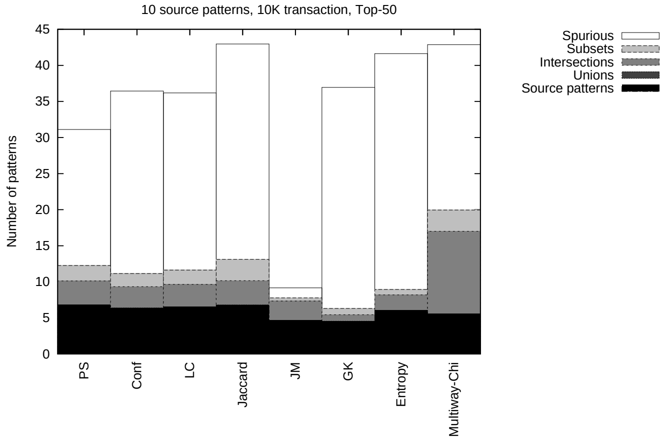

Fig. 15 Pattern types for mining the top-50 scoring closed frequent itemsets for $L = 10$, no corruption

6 Summary and Conclusions

In this work, we have for the first time evaluated condensed representations and interestingness measures for frequent itemset mining objectively. Due to a lack of data whose underlying patterns are known and whose characteristics can be easily controlled, it had been unknown whether FIM approaches recover underlying patterns.

We have revisited the Almaden Quest data generator, and exploited the fact that it constructs data from explicit patterns. By generating data sets and performing frequent itemset mining on them, we could compare the mined patterns against the source itemsets used to construct the data. We found not only that mining frequent, closed, or maximal patterns leads to result sets that include many non-relevant patterns in addition to source itemsets but also that several interestingness measures that have been proposed in the literature are only partially effective in reducing the result set to the relevant patterns.

In particular, we have seen that results may vary quite a bit depending on the number of actual source patterns in the data. Since this is one piece of information that no user will know, guidelines should be developed regarding the relationship between mined and actual patterns, depending on the generative processes. Such information could be used to relate patterns to each other in post-processing steps such as pattern set mining. Also, the Quest generator implements a relatively simple setting since all items of a source itemset are embedded together, without conditionality effects, and without any noise in the data that is independent of patterns. We therefore recommend to view our results as a kind of *upper bound* on the performance of the evaluated interestingness measures: results in more complex, let alone real-world, settings are likely to be worse and any measure that did not perform well in the setting we evaluated here is not very likely to be useful in other settings.

For future work, we therefore intend to look into more complex ways of embedding itemsets, e.g. involving noise, in particular noise that gives the appearance of regularity. Similarly, we will develop approaches for embedding partial patterns conditionally, since this would be a setting that is better suited for evaluating the performance of quality measures for predictive rules. Additionally, there have been proposals put forward for generators that generate the source itemsets themselves in a more sophisticated way, for instance in [14]. Such data can then be used to relate found patterns to generating processes, and we will use it to follow up on the results presented in this work, with the goal of getting closer to giving robust recommendations for the use of measures and pattern types in real-world settings.

Acknowledgements

We are grateful to Christian Borgelt and Tijl De Bie for their support w.r.t. the FPGrowth implementation and the MaxEnt Database Generator, respectively, and to our colleagues Matthijs van Leeuwen and Tias Guns, and the participants of Qimie 2013 for helpful discussions. Finally, we thank the anonymous reviewers for their help in improving the manuscript. The author is supported by a post-doctoral grant by the Fonds Wetenschappelijk Onderzoek Vlanderen (FWO).

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Databases. pp. 487–499. Morgan Kaufmann, Santiago de Chile, Chile (Sep 1994)
2. Bayardo Jr., R.J., Goethals, B., Zaki, M.J. (eds.): FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Brighton, UK, November 1, 2004 (2004)
3. Bie, T.D.: Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Min. Knowl. Discov. 23(3), 407–446 (2011)
4. Blake, C., Merz, C.: UCI repository of machine learning databases (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
5. Blanchard, J., Guillet, F., Gras, R., Briand, H.: Using information-theoretic measures to assess association rule interestingness. In: Han, J., Wah, B.W., Raghavan, V., Wu, X., Rastogi, R. (eds.) ICDM. pp. 66–73. IEEE, Houston, Texas, USA (Nov 2005)
6. Boulicaut, J.F., Jeudy, B.: Mining free itemsets under constraints. In: Adiba, M.E., Collet, C., Desai, B.C. (eds.) International Database Engineering & Applications Symposium, IDEAS '01. pp. 322–329 (2001)
7. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: Generalizing association rules to correlations. In: Peckham [16], pp. 265–276
8. Carvalho, D.R., Freitas, A.A., Ebecken, N.F.F.: Evaluating the correlation between objective rule interestingness measures and real human interest. In: Jorge, A., Torgo, L., Brazdil, P., Camacho, R., Gama, J. (eds.) PKDD. pp. 453–461. Springer (2005)
9. Cooper, C., Zito, M.: Realistic synthetic data for testing association rule mining algorithms for market basket databases. In: Kok, J.N., Koronacki, J., de Mántaras, R.L.,

Matwin, S., Mladenic, D., Skowron, A. (eds.) PKDD. Lecture Notes in Computer Science, vol. 4702, pp. 398–405. Springer (2007)

10. Gouda, K., Zaki, M.J.: Genmax: An efficient algorithm for mining maximal frequent itemsets. Data Min. Knowl. Discov. 11(3), 223–242 (2005)
11. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Chen, W., Naughton, J.F., Bernstein, P.A. (eds.) SIGMOD Conference. pp. 1–12. ACM (2000)
12. Heikinheimo, H., Seppänen, J.K., Hinkkanen, E., Mannila, H., Mielikäinen, T.: Finding low-entropy sets and trees from binary data. In: Berkhin, P., Caruana, R., Wu, X. (eds.) KDD. pp. 350–359. ACM (2007)
13. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research 184(2), 610–626 (2008)
14. Mampaey, M., Vreeken, J.: Summarizing categorical data by clustering attributes. Data Min. Knowl. Discov. 26(1), 130–173 (2013)
15. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: Beeri, C., Buneman, P. (eds.) ICDT. Lecture Notes in Computer Science, vol. 1540, pp. 398–416. Springer (1999)
16. Peckham, J. (ed.): SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA. ACM Press (1997)
17. Pei, J., Han, J., Mao, R.: Closet: An efficient algorithm for mining frequent closed itemsets. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 21–30 (2000)
18. Pei, Y., Zaïane, O.: A synthetic data generator for clustering and outlier analysis. Tech. rep. (2006)
19. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press (1991)
20. Ramesh, G., Zaki, M.J., Maniatty, W.: Distribution-based synthetic database generation techniques for itemset mining. In: IDEAS. pp. 307–316. IEEE Computer Society (2005)
21. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: KDD. pp. 32–41. ACM (2002)
22. Vaillant, B., Lenca, P., Lallich, S.: A clustering of interestingness measures. In: Suzuki, E., Arikawa, S. (eds.) Discovery Science. Lecture Notes in Computer Science, vol. 3245, pp. 290–297. Springer (2004)
23. Vreeken, J., van Leeuwen, M., Siebes, A.: Preserving privacy through data generation. In: Ramakrishnan, N., Zaïane, O. (eds.) ICDM. pp. 685–690. IEEE Computer Society (2007)
24. Wu, T., Chen, Y., Han, J.: Re-examination of interestingness measures in pattern mining: a unified framework. Data Min. Knowl. Discov. 21(3), 371–397 (2010)
25. Zaki, M.J., Hsiao, C.J.: ChArm: An efficient algorithm for closed association rule mining. Tech. rep., Computer Science Department, Rensselaer Polytechnic Institute (October 1999)
26. Zaki, M.J.: Scalable algorithms for association mining. IEEE Trans. Knowl. Data Eng. 12(3), 372–390 (2000)
27. Zaki, M.J., Hsiao, C.J.: Charm: An efficient algorithm for closed itemset mining. In: Grossman, R.L., Han, J., Kumar, V., Mannila, H., Motwani, R. (eds.) SDM. SIAM (2002)
28. Zheng, Z., Kohavi, R., Mason, L.: Real world performance of association rule algorithms. In: KDD. pp. 401–406 (2001)