Essential Patterns: A Perfect Cover of Frequent Patterns

Alain Casali, Rosine Cicchetti, and Lotfi Lakhal

Laboratoire d'Informatique Fondamentale de Marseille (LIF),
CNRS UMR 6166, Université de la Méditerranée
Case 901, 163 Avenue de Luminy, 13288 Marseille Cedex 9, France
lastname@lif.univ-mrs.fr

Abstract. The extraction of frequent patterns often yields extremely voluminous results which are difficult to handle. Computing a concise representation or cover of the frequent pattern set is thus an interesting alternative investigated by various approaches. The work presented in this article fits in such a trend. We introduce the concept of essential pattern and propose a new cover based on this concept. Such a cover makes it possible to decide whether a pattern is frequent or not, to compute its frequency and, in contrast with related work, to infer its disjunction and negation frequencies. A levelwise algorithm with a pruning step which uses the maximal frequent patterns for computing the essential patterns is proposed. Experiments show that when the number of frequent patterns is very high (strongly correlated data), the defined cover is significantly more reduced than the cover considered until now as minimal: the frequent closed patterns.

1 Introduction and Motivations

It is well known that frequent patterns mined from transactional databases can be extremely voluminous, and specially when data is strongly correlated. In such a context, it is difficult for the end-user to handle the extracted knowledge. Various approaches have addressed this problem and attempt to compute a concise representation, also called cover, of the whole set of frequent patterns [PBTL99, BR01, Pha02, CG02, BR03, KRG04]. Such a cover has a twofold interest: determining, at lower cost, (*i*) if an unknown pattern is frequent or not, and (*ii*) if it is the case, what is its frequency.

Unfortunately, among the covers proposed in the literature [KRG04], most of them are not proved to be concise representations and, in some cases, they can be more voluminous than the whole set of frequent patterns. In such cases, the initial objectives are not met and the difficulty to manage patterns worsens.

We call "*perfect couver*" of a set of frequent patterns a cover which is always smaller than this set. Only two have been proposed in the literature: the cover using frequent closed patterns [PBTL99, PHM00, Pha02, ZH02] and the one based on non derivable frequent patterns [CG02].

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2005, LNCS 3589, pp. 428–437, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we propose a new perfect cover based on the inclusion-exclusion identities [Nar82]. With this intention, we introduce the concept of essential pattern. We show that our representation based on frequent essential patterns is a perfect cover and has an interesting advantage when compared with the two other approaches: it is possible, not only, to retrieve the frequency of an unknown pattern but also to know the frequency of its disjunction and of its negation. Moreover, we propose a levelwise algorithm for computing the set of essential patterns. Through various experiments, we compare our approach with the one known as minimal: the cover based on closed patterns. Results are convincing since in the worse cases, when data is highly correlated, the size of our representation is significantly more reduced than the size of the closed pattern cover.

The remainder of the article is the following: in section 2 we recall the principle of the inclusion-exclusion identities. We use these identities in order to define a novel concept in section 3: the essential patterns. In the section 4, we propose the new cover based on the essential patterns. We propose a levelwise algorithm with a pruning step which uses the maximal frequent patterns for computing the frequent essential patterns in section 5. Experimental results are given in section 6. In conclusion, we resume the strengths of our contribution and the prospects for research.

2 Frequency Measures and Inclusion-Exclusion Identities

Let $\mathcal{D}$ be a transactional databse over a set of items and $X \in \mathcal{P}(\mathcal{I})$[1] a pattern, we define three weight measures, which are compatible with the weight functions defined in [STB+02], for X: (i) its frequency (denoted by $Freq(X)$), (ii) its disjunctive frequency (denoted by $Freq(\vee X)$) and (iii) its negative frequency (denoted by $Freq(\neg X)$). The disjunctive frequency of a pattern X can be seen as the probability to have at least one 1-pattern of X and the frequency of the negation stands for the probability to have no 1-pattern of X.

$$Freq(X) = \frac{|\{X' \in \mathcal{D} \mid X \subseteq X'\}|}{|\mathcal{D}|} \tag{1}$$

$$Freq(\vee X) = \frac{|\{X' \in \mathcal{D} \mid X \cap X' \neq \emptyset\}|}{|\mathcal{D}|} \tag{2}$$

$$Freq(\neg X) = \frac{|\{X' \in \mathcal{D} \mid X \cap X' = \emptyset\}|}{|\mathcal{D}|} \tag{3}$$

Example 1. - Let $\mathcal{D}$ be the following database:

We have: $Freq(AC) = 2/4$, $Freq(\vee AC) = 1$ and $Freq(\neg AC) = 0$. Since $Freq(\vee AC) = 1$, each transaction in $\mathcal{D}$ contains either the 1-pattern A, or the 1-pattern C, or both of them.

[1] $\mathcal{P}(X)$ is the powerset of X.

Table 1. Database example $\mathcal{D}$

Tid	Items
1	ABCD
2	ABD
3	CE
4	ACD

The inclusion-exclusion identities make it possible to state, for a pattern X, the relationship between the frequency, the frequency of the disjunction and the frequency of the negation, as follows:

$$Freq(X) = \sum_{\substack{X' \subseteq X \\ X' \neq \emptyset}} (-1)^{(|X'|-1)} Freq(\vee X'). \tag{4}$$

$$Freq(\vee X) = \sum_{\substack{X' \subseteq X \\ X' \neq \emptyset}} (-1)^{(|X'|-1)} Freq(X'). \tag{5}$$

$$Freq(\neg X) = 1 - Freq(\vee X) \text{ (from De Morgan Law)} \tag{6}$$

Example 2. - In our database example, we have:

1. $Freq(AC) = Freq(A) + Freq(C) - Freq(\vee AC) = 3/4 + 3/4 - 1 = 2/4$
2. $Freq(\vee AC) = Freq(A) + Freq(C) - Freq(AC) = 3/4 + 3/4 - 2/4 = 1$
3. $Freq(\neg AC) = 1 - Freq(\vee AC) = 0.$

Computing the frequency of the disjunction for a pattern can be performed along with computing its frequency and thus the execution time of levelwise algorithms is not altered. Provided with the frequency of the disjunction for the frequent patterns, a perfect cover of frequent patterns can be defined and the computation of the negation frequency is straightforward ($cf.$ De Morgan Law).

3 Essential Patterns

A pattern X is essential if and only if its disjunctive frequency is different from the disjunctive frequency of all its direct subsets. Since the disjunctive frequency function is an increasing monotone function, we do not need to examine the disjunctive frequency of each direct subset for a pattern X. Checking that the disjunctive frequency of X is different to the greatest disjunctive frequency of its direct subsets is a sufficient condition to be sure that X is an essential pattern. A more formal definition of the concept of essential pattern is given below. Then, we show that the constraint "X *is an essential pattern*" is an antimonotone constraint for the inclusion. Thus, this constraint is compatible with the frequency constraint and makes it possible to use levelwise algorithms for mining frequent essential patterns.

Definition 1. ***(Essential patterns)*** *- Let $\mathcal{D}$ be a transactional database over a set of items $\mathcal{I}$ and let $X \in \mathcal{P}(\mathcal{I})$ be a pattern. We say that $X \neq \emptyset$ is an essential pattern if and only if*

$$Freq(\vee X) \neq \max_{x \in X}(Freq(\vee X \backslash x)). \tag{7}$$

Let us denote by $\mathcal{E}$ the set of essential patterns and $\mathcal{E}(\mathbb{F})$ the set of frequent essential patterns[2]*.*

Example 3. - In our database example, the pattern AC is an essential pattern because $Freq(\vee AC) \neq Freq(\vee A)$ and $Freq(\vee AC) \neq Freq(\vee C)$.

Lemma 1. *- Let us consider the two following constraints: "X is frequent" (C_1) and "X is essential" (C_2). The conjunction of the two constraints is antimonotone for the inclusion (i.e. if X is a frequent essential pattern, then all its subsets are frequent and essential patterns).*

4 Frequency Computation Using an Improvement of the Inclusion-Exclusion Identities

The three following formulas show firstly how to compute the frequency of the disjunction from the set of essential patterns and secondly how to optimize the inclusion-exclusion identities for finding efficiently the frequency of a frequent pattern. A naive method for computing the frequency of a pattern X requires the knowledge of the disjunctive frequency of all its subsets. Formula 8 shows how we can derive the disjunctive frequency of any patterns using only essential patterns.

Lemma 2. *Let X be a set of items, then we have:*

$$Freq(\vee X) = \max_{Y \in \mathcal{E}}(\{Freq(\vee Y) \mid Y \subseteq X\}). \tag{8}$$

The formula 9 is an optimization based on the concept of essential patterns and the formula 10 is an original method for the derivation of the frequency of X.

Lemma 3. *- $\forall X \in \mathcal{P}(\mathcal{I})$, let be $Y \in Argmax(\{Freq(\vee X') \mid X' \subseteq X$ and $X' \in \mathcal{E}\})$, then we have:*

$$Freq(X) = \sum_{\substack{X' \subseteq X \\ X' \neq \emptyset}} (-1)^{|X'|-1} \begin{cases} Freq(\vee Y) \text{ if } Y \subseteq X' \\ Freq(\vee X') \text{ elsewhere} \end{cases} \tag{9}$$

Theorem 1. *- $\forall X \in \mathbb{F}$, $X \notin \mathcal{E}(\mathbb{F})$, let be $Y \in Argmax(\{Freq(\vee X') \mid X' \subseteq X$ and $X' \in \mathcal{E}(\mathbb{F})\})$, then we have:*

[2] $\mathbb{F}$ is the set of frequent patterns.

$$Freq(X) = \sum_{\substack{X' \subseteq X \\ X' \neq \emptyset \\ X' \not\subseteq Y}} (-1)^{|X'|-1} Freq(\vee X') \tag{10}$$

The set of essential patterns is not sufficient to define a perfect cover for the set of frequent patterns because we cannot decide if an unknown pattern is frequent or not. That is why we add the positive border for the frequency constraint to the set of frequent essential patterns for testing if an unknown pattern is frequent or not. If it is frequent, then theorem 1 makes it possible to to compute the frequency of its conjunction. Thus, the set of frequent essential patterns ($\mathcal{E}(\mathbb{F})$) increased with the positive border for the frequency constraint ($BD^+(\mathbb{F})$) is a perfect cover for the frequent patterns.

Definition 2. ***(Perfect cover)*** *- Let $\mathcal{D}$ a transactional database over a set of items $\mathcal{I}$ (each transaction is a subset of $\mathcal{I}$) and $\mathbb{F}$ the set of frequent patterns. We say that $\mathbb{G}$ is a cover for $\mathbb{F}$ if and only if the frequency of each element of $\mathbb{F}$ can be retrieved by using only patterns of $\mathbb{G}$ ($\forall X \in \mathbb{F}, \mathbb{G} \vDash Freq(X)$). Moreover, if $\mathbb{G} \subseteq \mathbb{F}$, the cover is called perfect.*

Theorem 2. *- Let $BD^+(\mathbb{F})$ be the positive border (i.e. the set of maximal frequent patterns) and $\mathcal{E}(\mathbb{F})$ the set of essential frequent patterns, then $BD^+(\mathbb{F}) \cup \mathcal{E}(\mathbb{F})$ is a perfect cover for the frequent patterns.*

5 The MEP Algorithm

For finding the frequent essential patterns, we propose a levelwise algorithm with a pruning step which uses the maximal frequent patterns ($BD^+(\mathbb{F})$). The algorithm MEP (Mining Essential Patterns) includes the function Max_Set_Algorithm which discovers maximal frequent patterns (*e.g.* Max-Miner [Bay98], Gen-Max [GZ01]).

Example 4. The perfect cover of our example for the threshold "$Minfreq = 2/4$" is the following: the set of frequent essential patterns is given in table 2 and the positive border $BD^+(\mathbb{F})$ is given in table 3.

We know that the pattern ABD is frequent because it belongs to the positive border. Let us compute its frequency.

Table 2. Frequent essential pattern for "$Freq(X) \geq 2/4$"

Essential pattern	Disjunctive frequency
A	3/4
B	2/4
C	3/4
D	3/4
AC	1
CD	1

Table 3. Positive border for "$Freq(X) \geq 2/4$"

Positive border
ABD
ACD

Algorithm 1 MEP Algorithm

1: $BD^+(\mathbb{F}) :=$ Max_Set_Algorithm$(\mathcal{D}, Minfreq)$
2: $L_1 = \{$frequent $1-$pattern $\}$
3: $i := 1$
4: **while** $L_i \neq \emptyset$ **do**
5: $C_{i+1} :=$ Gen_Apriori(L_i)
6: $C_{i+1} := \{X \in C_{i+1} \mid \exists Y \in BD^+(\mathbb{F}) : X \subseteq Y\}$
7: Scan the database for mining the disjunctive frequency $\forall X \in C_{i+1}$
8: $L_{i+1} := \{X \in C_{i+1} \mid \nexists x \in X : Freq(\vee X) = Freq(\vee X \backslash x)\}$
9: $i := i + 1$
10: **end while**
11: **return** $\bigcup_{j=1..i} L_j$

- We use lemma 2 to find its disjunctive frequency: $Freq(\vee ABD) = max(Freq(\vee A), Freq(\vee B), Freq(\vee D)) = Freq(\vee A) = Freq(\vee D) = 3/4$.
- We apply theorem 1 to compute its frequency:
  The patterns A and D are two frequent essential patterns included in ABD for which the disjunctive frequency is maximal. Thus we have: $Argmax(\{Freq(\vee ABD) \mid X' \subseteq X$ and $X' \in \mathcal{E}(\mathbb{F})\}) = \{A, D\}$. We need one of these two patterns to apply theorem 1, we choose $Y = A$. We obtain the following equality: $Freq(ABD) = Freq(A) + Freq(B) + Freq(D) - Freq(\vee AB) - Freq(\vee AD) - Freq(\vee BD) + Freq(\vee ABD) = Freq(B) + Freq(D) - Freq(\vee BD)$. Since the pattern BD is not an essential, we need to know its disjunctive frequency. By applying, once more, theorem 1, we obtain $Freq(B) + Freq(D) - Freq(\vee BD) = Freq(B)$. Accordingly, $Freq(ABD) = 2/4$.

We have eliminated all the patterns included between A and ABD in the inclusion-exclusion identities because the sum of their disjunctive frequencies, weighted of the good coefficient, is null.

6 Experimental Results

By providing the disjunctive and the negative frequencies, the proposed approach enriches the results obtained with the two other perfect covers proposed in the literature. Our objective is now to show, through various experiments, that the size of this new cover is often smaller than the size of the cover based on the

Table 4. Datasets

Name	Number of transactions	Average size of each transaction	Number of items
CHESS	3 196	37	75
CONNECT	65 557	43	129
PUMSB	49 046	74	2 113
PUMSB*	49 046	50,5	2 088

frequent closed patterns and this in the most critical cases: strongly correlated data. For meeting this objective, we evaluate the number of frequent essential patterns and compare it with the number of frequent closed patterns by using four datasets[3]. The characteristics of the datasets used for experiments are given in table 4. They are:

- the dataset CHESS,
- the dataset CONNECT,
- the datasets of census PUMSB and PUMSB*, extracted from « PUMS sample file ». PUMSB* is the same dataset than PUMSB from which are removed all the patterns which have a threshold greater or equal to 80%,

For all the experiments, we choose relevant minimum thresholds. In these four datasets, only encompassing strongly correlated data, the ratio between frequent patterns and the total number of patterns is high. Thus we are in the most difficult cases. For finding the positive border we use Gen-Max algorithm [GZ01]. In the dataset PUMSB*, using either the frequent closed patterns or the frequent essential patterns as a cover is advantageous: the gain compared to the set of frequent patterns for the dataset PUMSB* with the threshold $Minfreq = 20\%$ is about 45. On the other hand, for this dataset, even if the approach by essential patterns is better than the one with closed patterns, the obtained gain is near to one. In the three remaining datasets, the approach by essential is very efficient. With the dataset CHESS, many of frequent patterns are closed patterns, but the number of essential patterns is relatively small. This results in a benefit, for the

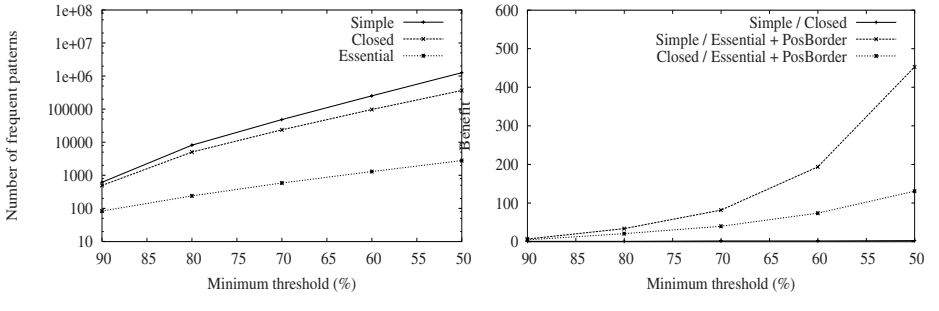

Fig. 1. Experimental results for CHESS

[3] `http://fimi.cs.helsinki.fi/`

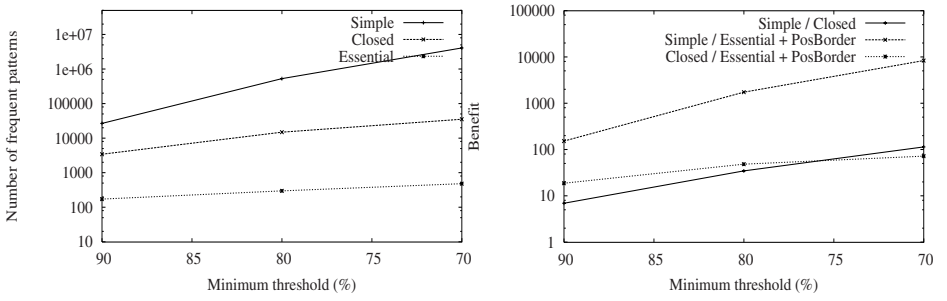

Fig. 2. Experimental results for CONNECT

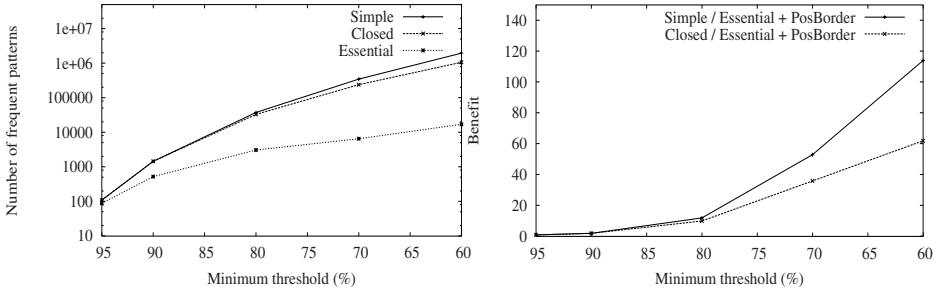

Fig. 3. Experimental results for PUMBS

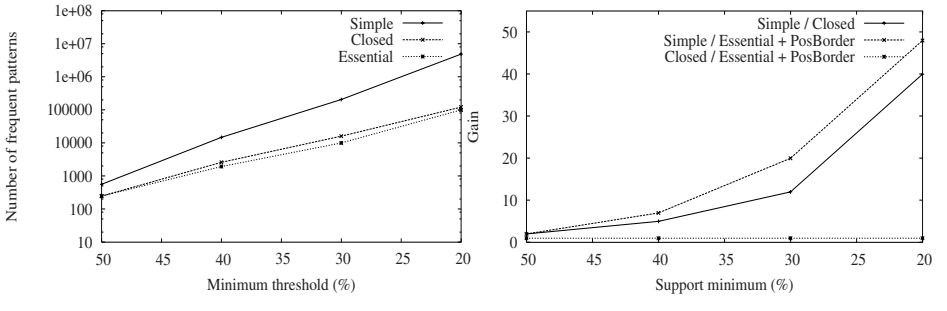

Fig. 4. Experimental results for PUMBS*

threshold $Minfreq = 50\%$, of a factor 40 compared to the original approach and of a factor 20 compared to the approach using closed patterns. With the dataset CONNECT and a threshold $Minfreq = 70\%$, the benefit compared to frequent patterns is approximately of a factor 2500 and compared to closed frequent pattern of a factor 20. We can see that with the dataset PUMSB, the benefit compared to the approach by frequent closed patterns is of a factor 20 for a threshold $Minfreq = 60\%$ and compared to the approach by frequent patterns is approximately 40.

For more readability in the figures, we have omitted "frequent patterns" in the legends. Thus "simple" means frequent patterns, "closed" stands for frequent closed pattern and "essential" symbolizes frequent essential pattern.

7 Conclusion

In this paper, we propose a novel perfect cover for the frequent patterns based on the inclusion-exclusion identities. We introduce the concept of essential pattern. The perfect cover is based (i), on one hand, on the positive border which can be used to determine if an unknown pattern is frequent or not, and (ii), on the other hand, on the frequent essential patterns which make it possible to derive the frequency of a frequent pattern by applying an optimization of the inclusion-exclusion identities. Compared with the existing perfect covers, our method makes it possible to mine at lower cost, along with the frequency of a frequent pattern, the frequency of its disjunction and negation. We have also shown, from an experimental point of view, the efficiency of our approach in the most critical cases: when the mined data is correlated, the set of frequent patterns is extremely voluminous. Having a perfect cover is specially interesting to quickly answer the ad hoc requests of decision makers.

Concerning future work, it would be interesting to define the disjunctive closure operator to reduce the number of essential patterns because this concept is similar to the concept of key. Thus, by applying this operator to each frequent essential pattern, we would also obtain a set of disjunctive closed patterns which could be a perfect cover for the frequent patterns. Since closure operators are surjective functions, the number of frequent disjunctive closed patterns will be thus lower than the number of frequent essential patterns.

Acknowledgement

We would like to thank Stéphane Lopes who has performed the experiments.

References

[Bay98] Roberto Bayardo. Efficiently mining long patterns from databases. In *Proceedings of the International Conference on Management of Data, SIGMOD*, pages 85–93, 1998.

[BR01] Artur Bykowski and Christophe Rigotti. A condensed representation to find frequent patterns. In *Proceedings of the 20th Symposium on Principles of Database Systems, PODS*, pages 267–273, 2001.

[BR03] Artur Bykowski and Christophe Rigotti. Dbc: a condensed representation of frequent patterns for efficient mining. In *Information Systems*, volume 28(8), pages 949–977, 2003.

[CG02] Toon Calders and Bart Goethals. Mining all non-derivable frequent itemsets. In *In Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD*, pages 74–85, 2002.

[GZ01] Karam Gouda and Mohammed Javeed Zaki. Efficiently Mining Maximal Frequent Itemsets. In *Proceedings of the 1st IEEE International Conference on Data Mining, ICDM*, pages 3163–170, 2001.

[KRG04] Marzena Kryszkiewicz, Henryk Rybinski, and Marcin Gajek. Dataless transitions between concise representations of frequent patterns. In *Journal of Intelligent Information System*, volume 22(1), pages 41–70, 2004.

[Nar82] Hiroshi Narushima. Principle of Inclusion-Exclusion on Partially Order Sets. In *Discrete Mathematics*, volume 42, pages 243–250, 1982.

[PBTL99] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory, ICDT*, pages 398–416, 1999.

[Pha02] Viet Phan Luong. The closed keys base of frequent itemsets. In *Proceedings of the 4th Data Warehousing and Knowledge Discovery, DAWAK*, pages 181–190, 2002.

[PHM00] Jian Pei, Jiawei Han, and Runying Mao. CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets. In *Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD*, pages 21–30, 2000.

[STB+02] Gerd Stumme, Rafik Taouil, Yves Bastide, Nicolas Pasquier, and Lotfi Lakhal. Computing Iceberg Concept Lattices with Titanic. In *Data and Knowledge Engineering*, volume 42(2), pages 189–222, 2002.

[ZH02] Mohammed Javeed Zaki and Ching-Jui Hsio. CHARM: An Efficient Algorithm for Closed Itemset Mining. In *Proceedings of the 2nd SIAM International Conference on Data mining*, 2002.