Closed Non-derivable Itemsets

Juho Muhonen and Hannu Toivonen*

Helsinki Institute for Information Technology
Basic Research Unit
Department of Computer Science
University of Helsinki
Finland

Abstract. Itemset mining typically results in large amounts of redundant itemsets. Several approaches such as closed itemsets, non-derivable itemsets and generators have been suggested for losslessly reducing the amount of itemsets. We propose a new pruning method based on combining techniques for closed and non-derivable itemsets that allows further reductions of itemsets. This reduction is done without loss of information, that is, the complete collection of frequent itemsets can still be derived from the collection of closed non-derivable itemsets. The number of closed non-derivable itemsets is bound both by the number of closed and the number of non-derivable itemsets, and never exceeds the smaller of these. Our experiments show that the reduction is significant in some datasets.

1 Introduction

Itemset mining often results in a huge amount of itemsets. Unfortunately the result usually contains a large number of redundant itemsets. Redundancy is inherent in the collection of all frequent itemsets and is the result of the fact that many itemsets are uninformative if the user is already aware of some other itemsets in that collection. In technical terms, an itemset can be considered redundant if its support can be inferred from other itemsets. Removing redundant itemsets produces a condensed representation of all frequent itemsets.

The best known condensed representations are closed itemsets [1] (or generators/free sets) and non-derivable itemsets [2]. The collection of non-derivable itemsets often is smaller than the collection of closed sets, but given an arbitrary dataset either one may contain fewer itemsets. In this paper we propose a method that combines the ideas of closed and non-derivable itemsets, and is guaranteed to be at least as efficient as the better of the two, while retaining the cabability of losslessly recovering the full collection of frequent itemsets.

2 Basic Concepts from Related Work

The frequent itemset mining problem can be described as follows [3]. We are given a set $\mathcal{I}$ of items and a dataset (multiset) $\mathcal{D}$ of subsets of $\mathcal{I}$ called transactions. A

* Currently at University of Freiburg; supported by Humboldt foundation.

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 601–608, 2006.

set X of items is called an itemset. The support $s(X)$ of X is the number of transactions that contain X. An itemset is called frequent if its support is no less than a given minimum support threshold δ. We denote the collection of all frequent itemsets by *Freq*. In the rest of the paper, we work on collections of frequent itemsets, but often drop "frequent" for lingvistical simplicity.

One condensed representation of itemsets is based on the concept of closure [1]. The closure $cl(X)$ of an itemset X is the maximal superset of the itemset X with the same support as the itemset X. The closure is always unique. An itemset X is a closed itemset if and only if its closure is the itemset X itself. We denote the collection of closed frequent itemsets by *Closed*.

Given *Freq* we can obtain *Closed* by taking the closure of each frequent itemset. Vice versa, given *Closed* and an itemset, we obtain the support of the itemset by finding its most frequent superset.

A more recent proposal [2] for pruning redundant itemsets takes advantage of the inclusion-exclusion principle. If $Y \subset X$ and $|X \setminus Y|$ is odd, then the corresponding deduction rule for an upper bound of $s(X)$ is

$$s(X) \leq \sum_{I : Y \subset I \subset X, I \neq X} (-1)^{|X \setminus I|+1} s(I). \tag{1}$$

If $|X \setminus Y|$ is even, the direction of inequality is changed and the deduction rule gives a lower bound instead of an upper bound. Given all subsets of X, and their supports, we obtain a set of upper and lower bounds for X. When the smallest upper bound equals the highest lower bound, we have obtained the exact support of X. Such an itemset is called derivable. The collection of non-derivable frequent itemsets, denoted *NDI*, is a lossless representation of *Freq*. *NDI* is downward closed [2]. In other words, all supersets of a derivable itemset are derivable, and all subsets of a non-derivable itemset are non-derivable.

Example. Consider the small transaction dataset of four items a, b, c and d and six transactions given in the left panel of Figure 1. The deduction rules for computing bounds for $\{a, c, d\}$ are given in the right panel of Figure 1.

Example Dataset

a	b	c	d
		c	d
a	b		
a	b		d
a		c	
a		c	d

Deduction rules for $s(\{a, c, d\})$

$$\begin{aligned}
s(\{a,c,d\}) &\leq s(\{a,c\}) + s(\{a,d\}) + s(\{c,d\}) \\
&\quad - s(\{a\}) - s(\{c\}) - s(\{d\}) + s(\{\}) \\
&\quad = 3 + 3 + 3 - 5 - 4 - 4 + 6 = 2 \\
s(\{a,c,d\}) &\geq s(\{a,c\}) + s(\{a,d\}) - s(\{a\}) = 3 + 3 - 5 = 1 \\
s(\{a,c,d\}) &\geq s(\{a,c\}) + s(\{c,d\}) - s(\{c\}) = 3 + 3 - 4 = 2 \\
s(\{a,c,d\}) &\geq s(\{a,d\}) + s(\{c,d\}) - s(\{d\}) = 3 + 3 - 4 = 2 \\
s(\{a,c,d\}) &\leq s(\{a,c\}) = 3 \\
s(\{a,c,d\}) &\leq s(\{a,d\}) = 3 \\
s(\{a,c,d\}) &\leq s(\{c,d\}) = 3 \\
s(\{a,c,d\}) &\geq 0
\end{aligned}$$

Fig. 1. An example dataset and deduction rules

The support of the itemset $\{a, c, d\}$ has a highest lower bound of 2, which is equal to the lowest upper bound making $\{a, c, d\}$ derivable. □

3 Closed Non-derivable Itemsets

Both the collection of frequent itemsets and the collection of non-derivable itemsets are downward closed, and the use of closed itemsets for compression utilizes this property.

Definition 1. *Let NDI be the collection of frequent non-derivable itemsets. The collection of frequent closed non-derivable itemsets is $CNDI = \{cl(X) \mid X \in NDI\}$.*

There are some subtleties. First, $CNDI \subset NDI$ does not hold in general (whereas always $Closed \subset Freq$). From this it also follows that NDI cannot be reconstructed by a straightforward downward closure of $CNDI$.

Example. Given our example dataset in Figure 1 and a support threshold of 2, we have 12 frequent, 10 closed and 10 non-derivable itemsets, given in Figure 2.

Freq	*Closed*
{}(6)	{}(6)
{a}(5) {b}(3) {c}(4) {d}(4)	{a}(5) {c}(4) {d}(4)
{a,b}(3) {a,c}(3) {a,d}(3) {b,d}(2) {c,d}(3)	{a,b}(3) {a,c}(3) {a,d}(3) {c,d}(3)
{a,b,d}(2) {a,c,d}(2)	{a,b,d}(2) {a,c,d}(2)

NDI	*CNDI*
{}(6)	{}(6)
{a}(5) {b}(3) {c}(4) {d}(4)	{a}(5) {c}(4) {d}(4)
{a,b}(3) {a,c}(3) {a,d}(3) {b,d}(2) {c,d}(3)	{a,b}(3) {a,c}(3) {a,d}(3) {c,d}(3)
	{a,b,d}(2)

Fig. 2. All, closed, non-derivable and closed non-derivable itemsets and their supports

The collection of closed itemsets has been obtained by taking the closure of each frequent itemset. With regard to NDI, we showed earlier in Figure 1 that $\{a, c, d\}$ is a derivable itemset. This is also true for $\{a, b, d\}$.

As this example shows, neither NDI nor *Closed* is a subset of the other. For instance, $\{a, b, d\}$ is closed but derivable whereas $\{b\}$ is non-derivable but not closed. On the other hand, $CNDI \subset Closed$. A comparison of these collections shows how $\{a, c, d\}$ is not in $CNDI$ since it is derivable. We emphasize that $CNDI$ cannot in general be obtained by removing all derivable itemsets from *Closed*. In this example, $|CNDI| = 9$, which is less than $|Closed| = 10$ and $|NDI| = 10$. □

Depending on the dataset, either collection of closed or collection of non-derivable itemsets may give a more condensed representation. However, the size of the collection of closed non-derivable itemsets is guaranteed to always be at most as large as the smallest of the two representations.

Theorem 1. *The size of the collection of closed non-derivable itemsets $CNDI$ is smaller than or equal to the size of the collection of non-derivable itemsets NDI: $|CNDI| \leq |NDI|$.*

Proof. By definition $CNDI = \{cl(X) \mid X \in NDI\}$. Since the closure operation gives exactly one itemset, we trivially have $|CNDI| \leq |NDI|$. □

Theorem 2. *The size of the collection of closed non-derivable itemsets $CNDI$ is smaller than or equal to the size of the collection of closed itemsets Closed: $|CNDI| \leq |Closed|$.*

Proof. By the definition of $CNDI$, given any $X \in CNDI$ there exists $Y \in NDI$ such that $X = cl(Y)$. Since $Y \in NDI \Rightarrow Y \in Freq \Rightarrow cl(Y) \in Closed \Rightarrow X \in Closed$, i.e., $CNDI \subset Closed$. □

Experiments in the following section show that the reduction by closed non-derivable itemsets may be significant in comparison to the reduction given by either of the methods alone.

The definition of $CNDI$ directly gives the simple Algorithm 1. Correctness of the algorithm is trivial and follows directly from the definition.

Algorithm 1. CloseNDI($\mathcal{D}$, δ)

```
Input: A dataset $\mathcal{D}$ and a support threshold $\delta$
Output: The collection $CNDI$ of closed non-derivable itemsets
$CNDI \leftarrow \emptyset$
Mine all non-derivable itemsets $NDI$ with support threshold $\delta$ from dataset $\mathcal{D}$
  (using methods from [2,4]).
for all $X \in NDI$ do
    $CNDI \leftarrow CNDI \cup \{cl(X)\}$
end for
return $CNDI$
```

Obtaining all frequent itemsets and their supports from the closed non-derivable itemsets is a bit more complex operation. First, given a collection of closed non-derivable itemsets we can obtain the supports of all non-derivable itemsets. Then the non-derivable itemsets in turn can be used to deduce the supports of all frequent itemsets by a levelwise search of the itemset space.

Theorem 3. *Algorithm 2 correctly recovers the frequent itemsets and their supports from the collection of closed non-derivable frequent itemsets $CNDI$.*

Algorithm 2. ExpandCNDI($CNDI$, δ)

Input: A collection $CNDI$ of frequent closed non-derivable itemsets and the corresponding support threshold δ
Output: The collection $Freq$ of frequent itemsets

```
 1: $Freq \leftarrow \emptyset$
 2: $l \leftarrow 0$
 3: while $\exists X$ such that $|X| = l$ and for all $Y \subset X$, $Y \neq X$: $Y \in Freq$ do
 4:    for all $X$ such that $|X| = l$ and for all $Y \subset X$, $Y \neq X$: $Y \in Freq$ do
 5:       if support of $X$ can be derived from supports of its proper subsets $Y$ then
 6:          $s \leftarrow$ Support of $X$ as given by deduction rules (Equation 1)
 7:          if $s \geq \delta$ then
 8:             $Freq \leftarrow Freq \cup (X, s)$
 9:          end if
10:       else if $\exists Z \in CNDI$ such that $X \subset Z$ then
11:          $s \leftarrow$ Support of the most frequent itemset $Z \in CNDI$ for which $X \subset Z$
12:          $Freq \leftarrow Freq \cup (X, s)$
13:       end if
14:    end for
15:    $l \leftarrow l + 1$
16: end while
17: return $Freq$
```

Proof. Let F be the true collection of all frequent itemsets while $Freq$ denotes the result of the algorithm. We first show that $F \subset Freq$ by induction over frequent itemsets $X \in F$ in increasing size

- For the empty set we have $s(\{\}) \geq \delta \Rightarrow \{\} \in NDI \Rightarrow cl(\{\}) \in CNDI$ by definition of $CNDI$. The algorithm can not derive the support of the empty set from its proper subsets (there are none). It thus proceeds to find the support of the closure of the empty set on line 11 of the algorithm and by properties of closed sets obtains the correct support.
- Let us now assume that the algorithm has correctly found all frequent itemsets with size less than l, and consider a frequent itemset $X \in F$ with $|X| = l$. Now $X \in F \Rightarrow X \in NDI$ or $X \in F \setminus NDI$. If $X \in F \setminus NDI$ then by definition of NDI its support can be derived from its proper subsets, which have been correctly recovered by algorithm (the inductive hypothesis). X and its support are thus correctly added to $Freq$ on line 8. If $X \in NDI$ then $cl(X) \in CNDI$ by definition of $CNDI$, and the algorithm finds the correct support of X from its closure on line 11.

To complete the proof we need to show that $Freq \subset F$. Consider any $X \in Freq$. It must have been added to $Freq$ on line 8 or 12. If it was added on line 8, $s(X) \geq \delta$ due to the test on line 7. If it was added on line 12, $s(X) \geq \delta$ since there exists $Z \in CNDI, X \subset Z$ (line 9) and by the definition of $CNDI$, Z must be frequent. □

The most time consuming phase of the Algorithm 1 is mining NDI. Time to close that collection depends on the number of itemsets in that collection. Algorithm 2 uses inclusion-exclusion to check whether an itemset is derivable. Again this is the dominant phase of the algorithm. For further analysis and efficient ways of implementing inclusion-exclusion we refer to [2,4].

4 Experiments

For an experimental evaluation of the proposed algorithms, we performed several experiments on real datasets. We implemented all algorithms in C++.

For primary comparison of methods we use four dense datasets: chess, connect, mushroom and pumsb, all obtained from the UCI Machine Learning Repository. The chess and connect datasets are derived from their respective game steps, the mushroom database contains characteristics of various species of mushrooms, and the pumsb dataset contains census data. For further perspective into the compression capabilities of the methods, we also use two sparse datasets, T10I4D100K and T40I10D100K, which contain simulated market basket data generated by the IBM Almaden Quest research group. Table 1 shows some characteristics of the used datasets.

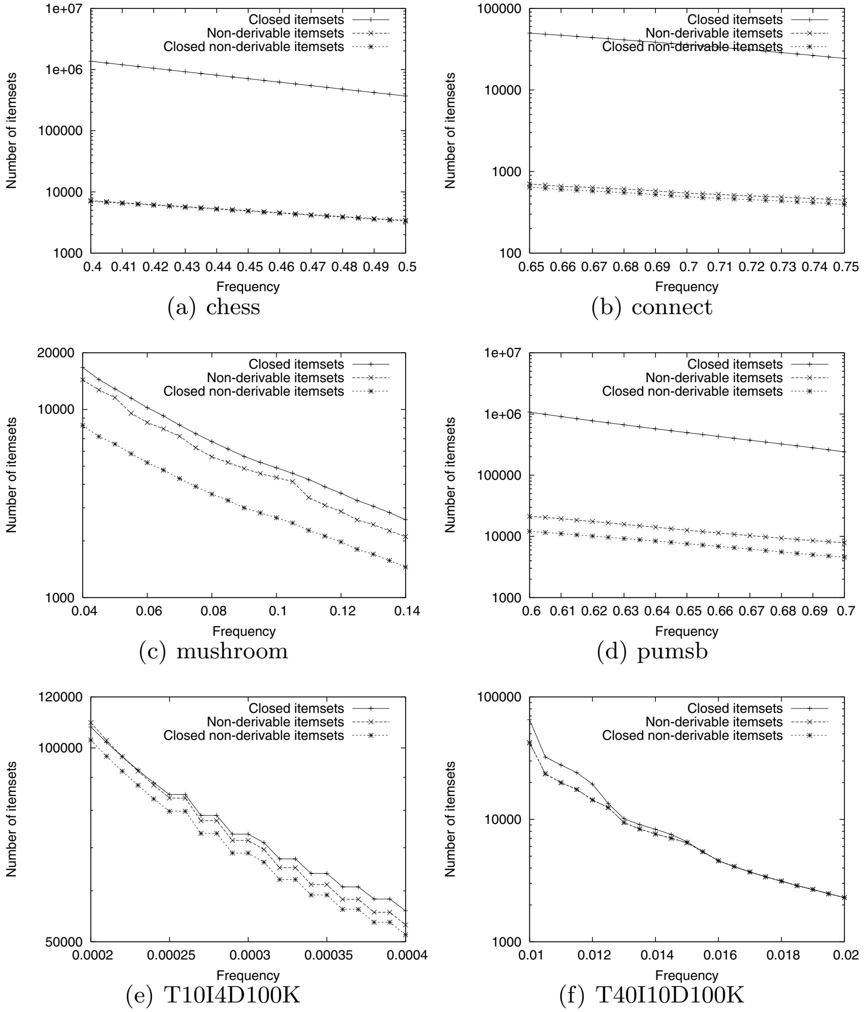
Table 1. Dataset characteristics

Dataset	Items	Avg. Trans. Size	Std. Dev. of Trans. Size	Transactions
chess	76	37	0	3 196
connect	130	43	0	67 557
mushroom	120	23	0	8 124
pumsb	7117	74	0	49 046
T10I4D100K	1000	10	3.7	100 000
T40I10D100K	1000	40	8.5	100 000

Figure 3 shows the results of the experiments: the number of closed itemsets, non-derivable itemsets and closed non-derivable itemsets, in the six different datasets as well as some of the results in numerical form.

The benefit of using closed non-derivable itemsets is biggest in mushroom and pumsb datasets, where the number of closed non-derivable itemsets is about one half of the number of non-derivable itemsets, and the reduction is even greater when compared to the number of closed itemsets (note that y-axis is logarithmic). In chess and connect datasets the compression for both non-derivable itemsets and closed non-derivable itemsets, when compared to closed itemsets, is about two orders of magnitude.

In all the dense datasets, the number of closed non-derivable itemsets is two to four magnitudes smaller than the number of all frequent itemsets. In the sparse market basket datasets, compression rates are much more modest, ranging about $15 - 35\%$. These two datasets are examples of the performance guarantee of closed non-derivable itemsets: for T10I4D100K, there are fewer closed itemsets than there are non-derivable itemsets, whereas in T40I10D100K the situation is reversed. In both cases the number of closed non-derivable itemsets is guaranteed not the exceed the smaller of these.



(a) chess

(b) connect

(c) mushroom

(d) pumsb

(e) T10I4D100K

(f) T40I10D100K

	chess	connect	mushroom	pumsb	T10	T40
Support Threshold (freq.)	1 279 (40%)	43 912 (65%)	325 (4.0%)	29 428 (60%)	20 (0.020%)	1 000 (1.0%)
$\lvert Freq \rvert$	6 472 981	9 727 092	5 131 852	19 537 366	129 876	65 237
$\lvert Closed \rvert$	1 366 834	49 707	16 733	1 075 015	107 823	65 237
$\lvert NDI \rvert$	7 185	704	14 382	21 323	109 486	42 312
$\lvert CNDI \rvert$	7 015	646	8 240	12 081	102 869	42 312

Fig. 3. The number of closed, non-derivable and closed non-derivable itemsets

5 Conclusions

We proposed a new method for lossless compression of a collection of frequent itemsets. The method takes advantage of the properties of two well-known techniques, closed itemsets and non-derivable itemsets.

We showed that the collection of closed non-derivable itemsets is a subset of the collection of closed itemsets, and that its size is also limited by the number of non-derivable itemsets, i.e., the combined method is guaranteed to yield better results than either one of the methods alone. We gave simple algorithms for producing the collection of closed non-derivable itemsets and recovering the collection of all frequent itemsets.

It is well known that closed itemsets and non-derivable itemsets give best compression rates for dense datasets, such as the UCI datasets used in our experiments, and give less benefits with sparse data, such as the IBM market basket data. Our experiments indicate that this is the case also for closed non-derivable itemsets. In our experiments with four real, dense datasets the reduction over closed itemsets was always significant $(50-99\%)$. For two of them, the reduction over non-derivable itemsets was small, but for the two others the collection of closed non-derivable itemsets was approximately 43% smaller. This shows that the collection of closed non-derivable itemsets can in practice be significantly smaller than the two other condensed representations.

It is not easy to characterize datasets that compress particularly well with closed non-derivable itemsets. An obvious factor is density: dense datasets lend themselves better for compression. The advantage over closed and non-derivable itemsets is largest when their compressions are complementary, as the use of closed non-derivable itemsets can then benefit from both, as seemed to be the case with mushroom and pumsb datasets. More research is needed to better understand the different factors. On the other hand, for practical applications such understanding is not needed: regardless of the data, closed non-derivable itemsets are the optimal choice among the three compared alternatives.

References

1. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: Proceedings of the 7th International Conference on Database Theory. Volume 1540 of Lecture Notes in Computer Science., Springer (1999) 398–416
2. Calders, T., Goethals, B.: Mining all non-derivable frequent itemsets. In: Proceedings of the sixth european conference on principles of data mining and knowledge discovery. (2002) 74–85
3. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD conference on management of data. (1993) 207–216
4. Calders, T., Goethals, B.: Quick inclusion-exclusion. In: KDID. (2005) 86–103