Closed sets based discovery of association rules

Nicolas Pasquier

Closed Set Based Discovery of Association Rules

Nicolas Pasquier

Laboratoire d'Informatique (LIMOS)
Université Blaise Pascal - Clermont-Ferrand II
Complexe Scientifique des Cézeaux
24 Avenue des Landais, 63177 Aubière Cedex France
pasquier@libd1.univ-bpclermont.fr

Plan of the Presentation

1 Association rule framework

2 Existing algorithms

3 A-Close algorithm

4 Illustration

5 Experimental results

6 Conclusion

7 Present work

1 Association Rules

- Data mining context (dataset)
  - binary relation $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$
  - $\mathcal{O}$: finite set of objets (transactions)
  - $\mathcal{I}$: finite set of items (attributes)

OID	Items			
1	A	C	D	
2	B	C	E	
3	A	B	C	E
4	B	E		
5	A	B	C	E

Figure 1: The example data mining context $\mathcal{D}$

- Itemset (set of items) support
  - proportion of objects containing the itemset

    $support(BC) = \|2,3,5\|/5 = 3/5$

- Association rules
  - implications between two itemsets

    $r : BC \rightarrow E \quad (support\%, confidence\%)$

- Association rule support
  - support of the union of antecedent and consequent of the rule

    $support(r) = support(BCE) = \|2,3,5\|/5 = 3/5$

- Association rule confidence
  - proportion of objects verifying the implication

    $confidence(r) = support(BCE)/support(BC) = 1$

- Minimum support and confidence thresholds defined by the user

2 Existing Algorithms

- Problem decomposition

  1. determination of frequent itemsets

     ($support \geq minsupport$)

  2. generation of association rules using frequent itemsets

     ($confidence \geq minconfidence$)

- The problem of extracting association rules is reduced to the problem of discovering frequent itemsets

- Pruning subset lattice $\mathcal{L}_I$ to extract frequent itemsets

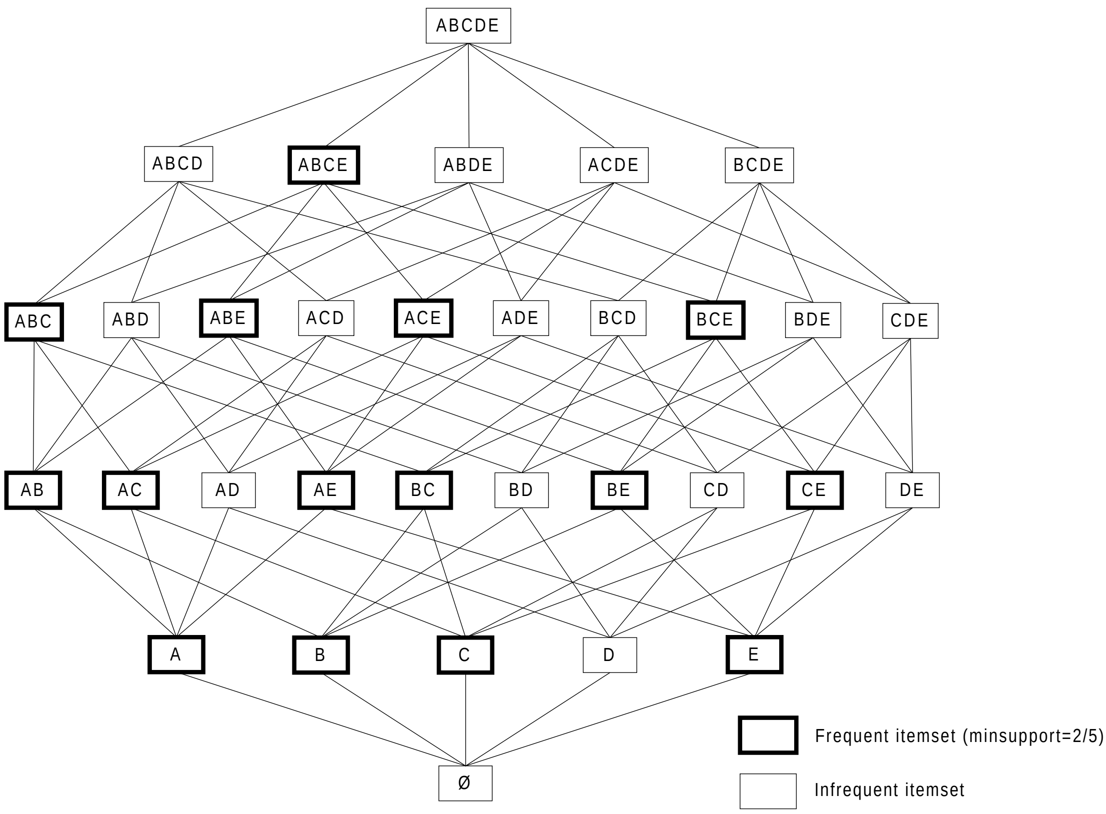


Figure 2: Subset lattice of $\mathcal{D}$

- Size is exponential $\|\mathcal{L}_I\| = 2^{\|\mathcal{I}\|}$

3 A-Close Algorithm

- Closure operator of the Galois connection of a binary relation
- Closed itemset: maximal set of items common to a set of objects

  ex: BC is not closed since $Objets(BC) = 2, 3, 5$ but $Items(2, 3, 5) = BCE$

- Problem decomposition
  1. discovering frequent closed itemsets
  2. deriving frequent itemsets from frequent closed itemsets
  3. generating association rules using frequent itemsets
- The problem of extracting association rules is reduced to the problem of discovering frequent closed itemsets
- Closed itemset properties

  i) all maximal frequent itemsets are maximal frequent closed itemsets

  ii) the support of a non-closed itemset is equal to the support of its closure

  iii) the maximal frequent closed itemsets characterise all frequent itemsets

- Pruning closed itemset lattice $\mathcal{L}_C$ to extract frequent closed itemsets

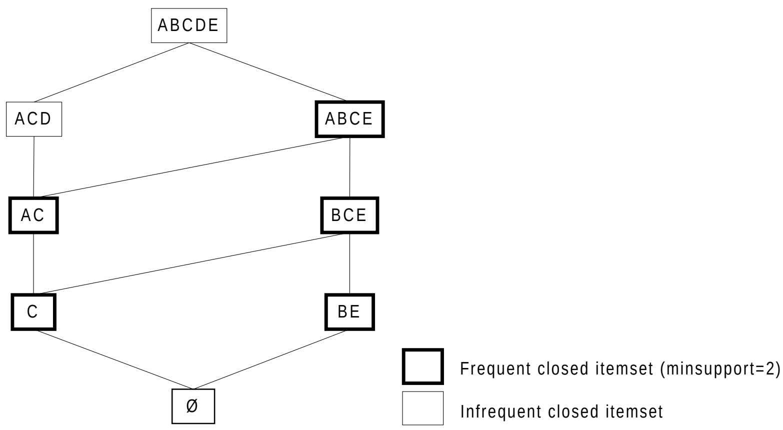


Figure 3: Closed itemset lattice of $\mathcal{D}$

- Determining minimal generator itemsets of all frequent closed itemsets
  - generators of a closed itemset: itemsets for which closure is the closed itemset
  - X is a minimal generator itemset if $\forall X' \subset X, support(X) \neq support(X')$
- Closure of an itemset is the intersection of all objects containing it

  ex: $Closure(BC) = Intersect(2, 3, 5) = BCE$

Algorithm 1 A-Close frequent closed itemset discovery

```
G_1 ← {frequent 1-itemsets};                        // scan D
for (i ← 2; G_i.generators ≠ ; i++) do
    G_i ← join generators in G_{i−1};
    Test presence of subsets(G_i) in G_{i−1};
    Determine support(G_i);                          // scan D
    Prune infrequent generators in G_i;
    Prune non-minimal generators in G_i;             // level variable ← i-1
end
Determine closures(⋃ G_i);                           // scan D
```

Support-Count ⟶

G_1

Generator	Support
{A}	3/5
{B}	4/5
{C}	4/5
{D}	1/5
{E}	4/5

Pruning infrequent generators ⟶

G_1

Generator	Support
{A}	3/5
{B}	4/5
{C}	4/5
{E}	4/5

AC-Generator ⟶

G_2

Generator	Support
{AB}	2/5
{AC}	3/5
{AE}	2/5
{BC}	3/5
{BE}	4/5
{CE}	3/5

Pruning ⟶

G_2

Generator	Support
{AB}	2/5
{AE}	2/5
{BC}	3/5
{CE}	3/5

AC-Closure ⟶

G'

Generator	Closure	Support
{A}	{AC}	3/5
{B}	{BE}	4/5
{C}	{C}	4/5
{E}	{BE}	4/5
{AB}	{ABCE}	2/5
{AE}	{ABCE}	2/5
{BC}	{BCE}	3/5
{CE}	{BCE}	3/5

Pruning ⟶

$Answer: FC$

Closure	Support
{AC}	3/5
{BE}	4/5
{C}	4/5
{ABCE}	2/5
{BCE}	3/5

Figure 4: A-Close frequent closed itemset discovery in $\mathcal{D}$ for $minsup = 2/5$ (40%)

4 Experimental Results

- Synthetic data: execution times
  - weakly correlated data: nearly all frequent itemsets are closed
  - additional time for A-Close in T20I6D100K (0.5%,0.33%): closure computations

- Census data: C20D10K
  - correlated data: few frequent itemsets are closed
  - closure mecanism allows to skip some iterations and consider less candidates

- Census data: C73D10K
  - differences between execution times can be measured in hours
  - maximal execution times: Apriori 14h, A-Close 1h15

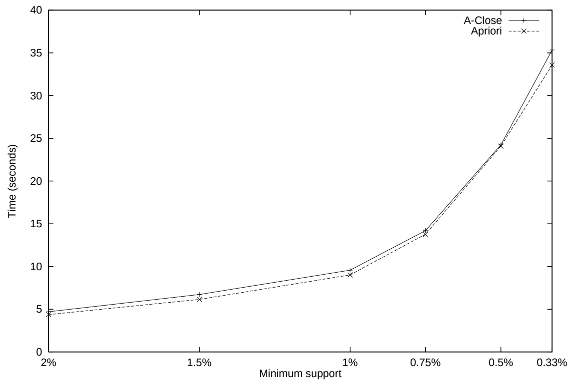
5 Conclusion

- Correlated data
  - difficult cases: long execution times
  - few frequent itemsets are closed: A-Close is particularly efficient

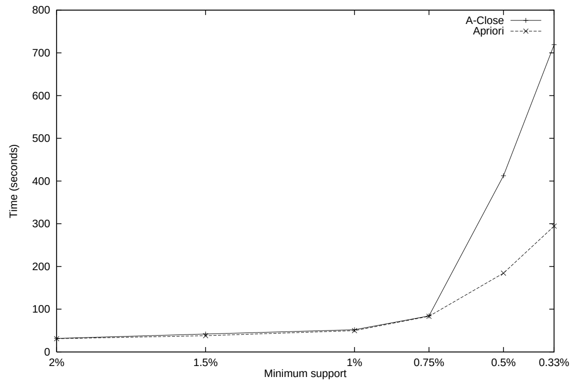
    statistical data, medical data, text data, etc.

- Weakly correlated data
  - nearly all frequent itemsets are closed
  - acceptable execution times

    synthetic data

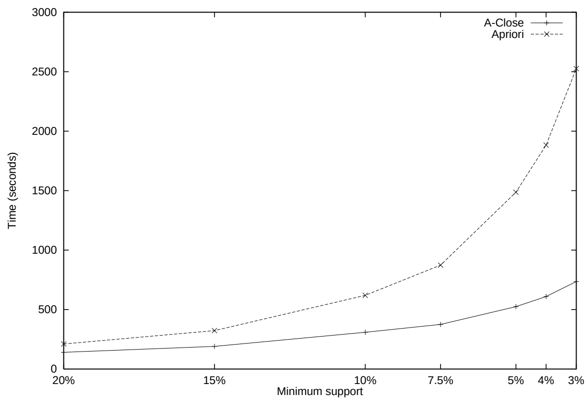


Execution times on T10I4D100K

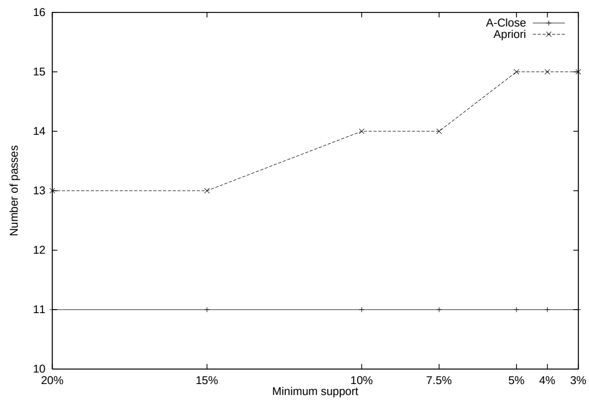


Execution times on T20I6D100K

Figure 5: Performance of Apriori and Close on synthetic data

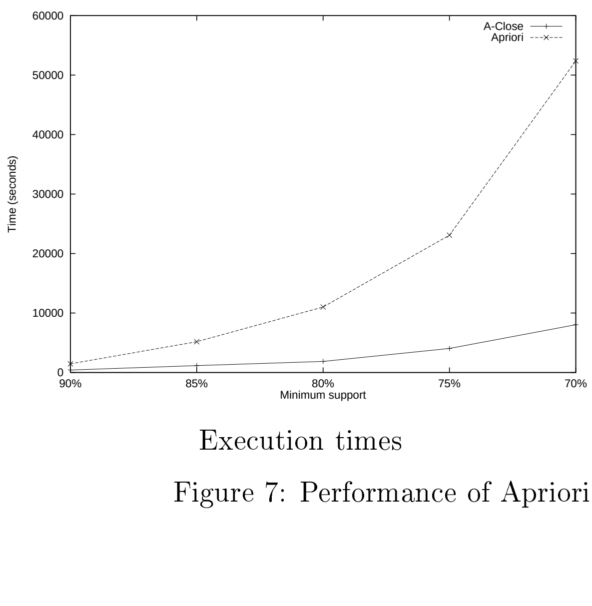


Execution times

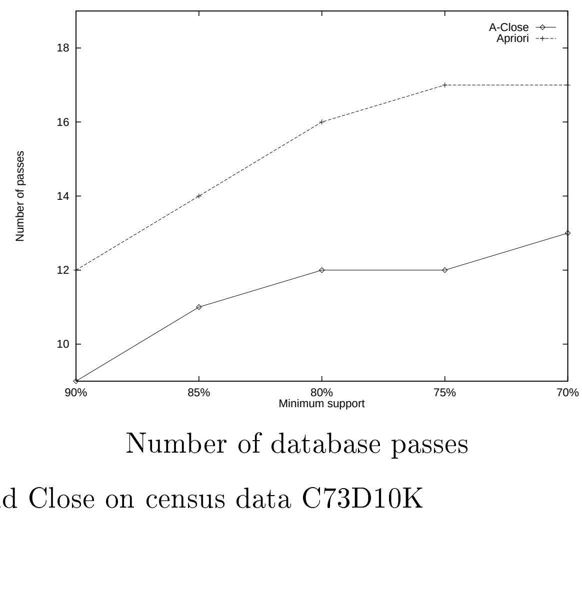


Number of database passes

Figure 6: Performance of Apriori and Close on census data C20D10K



Execution times



Number of database passes

Figure 7: Performance of Apriori and Close on census data C73D10K

6 Present Work

- Problem of the understandability and usefulness of association rules extracted

- Discovering small covers for association rules
  - small informative and structural cover for exact association rules
  - small informative cover for approximate association rules
  - small structural cover for approximate association rules

Dataset	Minimum support	Minimum confidence	Total rules	Informative cover	Structural cover
T10I4D100K	0.5%	90%	16,260	3,511	916
C73D10K	90%	90%	2,053,896	4,104	941
Mushrooms	50%	50%	1,248	87	44

Figure 8: Preliminary experimental results

References

[1] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Journal of Information Systems.* To appear.

[2] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Pruning closed itemset lattices for association rules. *Proceedings of the 14th BDA Conference on Advanced Databases*, pages 177–196, October 1998.

[3] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. *Proceedings of the 7th ICDT Int'l Conference on Database Theory*, pages 398–416, January 1999.