DATA-PEELER: Constrained-Based Closed Pattern Mining in n-ary Relations

Loïc Cerf, Jérémy Besson, Céline Robardet and Jean-François Boulicaut

INSA de Lyon - Bâtiment Blaise Pascal, 7 avenue Jean Capelle - 69621 Villeurbanne Cedex, France

Laboratoire d'InfoRmatique en Image et Systèmes d'information

Main Objectives

- Generalizing closed pattern (formal concept) extraction to n-ary relations.
- Identifying the class of constraints our proposal can efficiently handle at extraction time.

Article

In *SDM'08: Proceedings of the Eighth SIAM International Conference on Data Mining*. SIAM.

Closed n-set: a Generalization of Closed Pattern (Formal Concept) to n-ary Relations

The ternary relation, depicted beside, could represent customers (1, 2, 3 and 4) buying items (A, B and C) along three months (α, β and γ). $\langle(\alpha,\gamma),(1,2),(A,B)\rangle$ is an example of closed 3-sets:

Connection. The customers 1 and 2 buy both items A and B during the months α and γ.

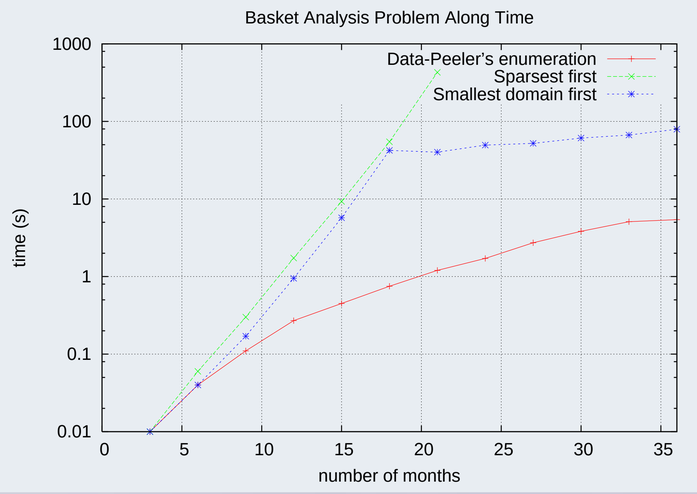
Closeness. $\langle(\alpha,\gamma),(1,2),(A,B)\rangle$ is closed w.r.t. every attribute:

- There is no other month during which these two customers buy these two items.
- No other customer buys these two items during these two months.
- No other item is simultaneously bought by these two customers during these two months.

	α			β			γ		
	A	B	C	A	B	C	A	B	C
1	1	1	1	1	1	1	1	1	
2	1	1		1			1	1	
3		1				1	1		1
4			1	1		1	1	1	1

DATA-PEELER: Extracting the Complete Collection of Closed n-sets

- D-MINER-like:
  - DATA-PEELER traverses the search space (*lattice*) in a *depth-first* way.
  - It recursively splits the search space into two *non-overlapping parts*.
- At any recursive call, *any element* (from any attribute) can be enumerated:
  - *Truly* working on n-*ary* relations.
  - A clever enumeration strategy improves the running times by orders of magnitude.
- Difficulties:
  - n-sets are *not structured by a Galois connection*.
  - Ability to exploit, at extraction time (*safe pruning*), a broad class of constraints on the n-sets.



Piecewise (Anti)-Monotonic Constraints

- DATA-PEELER can exploit, at extraction time (*safe pruning*), a broad class of constraints on the n-sets: the piecewise (anti)-monotonic constraints.
- In the expression of a piecewise (anti)-monotonic constraint, some arguments can occur several times. When such an argument grows (w.r.t. the $\subseteq$ order), some of its occurrences tend to satisfy the constraint, whereas the rest of them tend to violate it (cf. Example).

Example

Let $\epsilon \in \mathbb{R}^+$, a user-defined parameter,

$$\mathcal{C}_{\text{square}}(X^1,X^2) \equiv \frac{|X^1|}{|X^2|} - \frac{|X^2|}{|X^1|} \leq \epsilon \wedge \frac{|X^2|}{|X^1|} - \frac{|X^1|}{|X^2|} \leq \epsilon$$

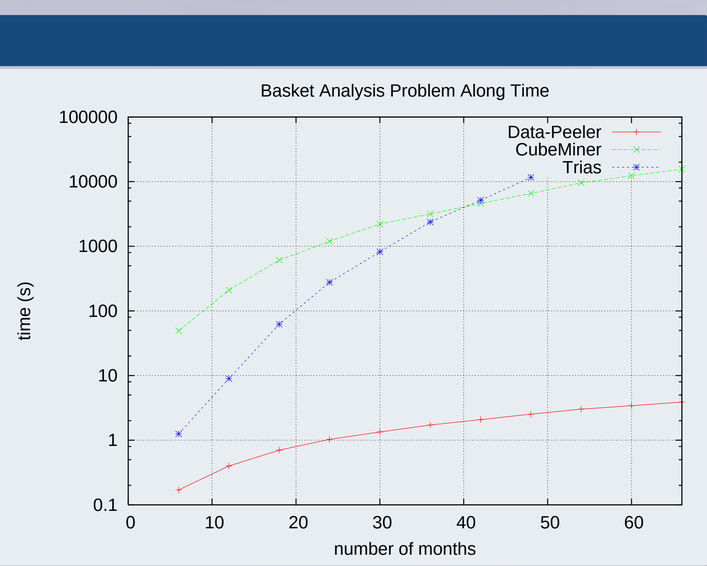
$\mathcal{C}_{\text{square}}$ is neither monotonic, nor anti-monotonic, nor succinct. It is not even convertible. However it is piecewise (anti)-monotonic. When growing (w.r.t. the $\subseteq$ order), the green occurrences of the arguments tend to satisfy $\mathcal{C}_{\text{square}}$, whereas the red ones tend to violate it.

Comparison With Competitors

General case. We are not aware of any other algorithm tackling the complete extraction of closed n-sets for $n > 3$.

Ternary case. CUBEMINER and TRIAS were *specifically* designed to extract closed 3-*sets*. DATA-PEELER *outperforms both of them* by orders of magnitude.

- CUBEMINER's performances quickly decrease with the size of the relation. CUBEMINER's division of the search space does not form a partition.
- TRIAS's performances quickly decrease with the size of the smallest dimension. TRIAS is combining extractions on binary relations.



Application to a Real-Life Co-Interest Dynamic Graph

Setting. A 4-*ary* relation was derived from the logs of `DistroWatch.com`. It indicates that, in a given country (among 40), during a given semester (among 7), two GNU/Linux distribution pages (among 350) were frequently loaded the same day by a visitor (identified by her IP address).

Results. The extracted closed 4-sets, under a weighted area constraint, were *relevant* (cf. Example).

Example

Some clearly identifiable groups of distributions:

- {Slackware, Gentoo, Ubuntu, Fedora, OpenSUSE, Debian}
- {dyne:bolic, AGNULA, MoviX, GeeXboX, ArtistX}
- {IPCop, ClarkConnect, Devil, SmoothWall, Astaro, m0n0wall}

In all the extracted 4-sets, the European countries and Australia are the most present.