Extending Set-based Dualization: Application to Pattern Mining[1]

Lhouari Nourine[1]   Jean-Marc Petit[2]

[1]Université Blaise Pascal, CNRS, LIMOS, France
[2]Université de Lyon, CNRS, INSA Lyon, LIRIS, France

ECAI 2012
Montpellier, France

[1]Funded by ANR DEFIS 2009 program, DAG project

Context: Mining interesting patterns in databases

➯ Plenty of contributions over the last 20 years

1. **Patterns**: itemsets, sequences, trees, graphs, functional dependencies, queries ...
2. **Databases**: Relational DB, Transactional DB, XML DB ... or just a collection of patterns (supposed to be large)
3. **Interestingness criteria**: frequency (and variants), satisfaction of some predicates

➯ Define a wide class of enumeration problems, some being studied for years in combinatorics, AI and databases

➯ Frequent itemset mining (**FIM**): The most studied problem in data mining

A theoretical perspective

Main theoretical framework proposed by (Mannila and Toivonen, DMKD, 1997)

- Identifying $\mathcal{RAS}$, the class of such problems reducible to FIM (i.e representable-as-sets)
  - **Isomorphism** between a poset of patterns and some set E ordered by inclusion
  - Identification of **set-based dualization** as the bottleneck for studying complexity
- $\mathcal{RAS}$ is relatively large
- However, there is an interesting open question (Gunopulos and al, ACM TODS 2003)
  How to deal with 'non representable-as-sets patterns" such as sequences, episodes or trees ?

Contributions in a nutshell

- Identifying a pattern mining problem as simple as possible and not “representable-as-sets”
  - Frequent rigid sequences with wildcard
- Studying the associated dualization problem (**SEQ**)
- Proving that SEQ is polynomially equivalent to set-based dualization
- Proposing a new theoretical framework for pattern mining problems
  - 2 new classes of problems: $\mathcal{WRAS}$ and $\mathcal{EWRAS}$

Plan

1. Preliminaries
2. Beyond $\mathcal{RAS}$
   - Weak representation as sets: The $\mathcal{WRAS}$ class
   - Efficient $\mathcal{WRAS}$: The $\mathcal{EWRAS}$ class
3. Concluding remarks

Plan

1. Preliminaries
2. Beyond $\mathcal{RAS}$
   - Weak representation as sets: The $\mathcal{WRAS}$ class
   - Efficient $\mathcal{WRAS}$: The $\mathcal{EWRAS}$ class
3. Concluding remarks

Notations (Mannila and Toivonen, DMKD, 1997)

A pattern mining problem:

- $\mathcal{L}$: set of patterns, $\preceq$ a partial order on $\mathcal{L}$.
- **d**: a database
- Q: a monotonic predicate to qualify interesting patterns X in **d**, noted Q(X,**d**).

The set of solutions is known as the theory (closed downward set) $Th(\mathcal{L}, \mathbf{d}, Q)$=$\{X \in \mathcal{L} \mid Q(X, \mathbf{d}) true\}$

Any closed downward set S can be represented by its **borders** $\mathcal{B}d^+(S)$ and $\mathcal{B}d^-(S)$.

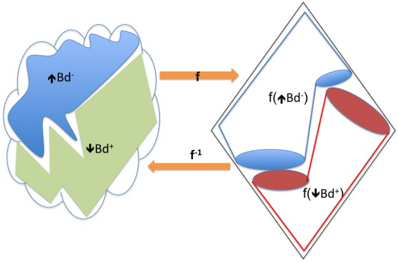
Dualization $\Leftrightarrow$ Relationship between the two borders

Isomorphism with a boolean lattice

Basic ideas:

- a bijection f between patterns $\mathcal{L}$ and some finite set R and
- structural isomorphism between $(\mathcal{L}, \preceq)$ and $(2^R, \subseteq)$

$\mathcal{RAS}$ = The class of pattern mining problems for which a representation as sets exists



$\mathcal{RAS}$, dualization and minimal transversals of hypergraphs

Dualization for $\mathcal{RAS}$?
➯ equivalent to minimal transversal of hypergraphs (TrMin)

Theorem [Mannila & Toivonen, DMKD, 1997]

Let P be pattern mining problem, $S \subseteq \mathcal{L}$ and (R, f) a representation as sets of P. Then

$$\mathcal{B}d^+(\downarrow S) = f^{-1}(\overline{TrMin(f(\mathcal{B}d^-(\downarrow S)))})$$

➯ Complexity of the decision problem is quasi-polynomial [Fredman & Khachiyan, J. Algo, 1996]

Main consequence: existence of incremental quasi-polynomial time algorithm for $\mathcal{RAS}$ [Gunopulos et al., TODS, 2003]

Plan

1. Preliminaries
2. Beyond $\mathcal{RAS}$
   - Weak representation as sets: The $\mathcal{WRAS}$ class
   - Efficient $\mathcal{WRAS}$: The $\mathcal{EWRAS}$ class
3. Concluding remarks

Limits of $\mathcal{RAS}$ (1/2)

(1) The surjectivity constraint

$\Rightarrow$the number of patterns has to be equal to 2^n, very unlikely in practice

Example with SEQ

Suppose an alphabet $\Sigma = \{a, b\}$ and an input sequence S of size 2. The set of all rigid sub-sequences of S is $\{\epsilon, a, b, aa, ab, ba, bb\}$.

Limits of $\mathcal{RAS}$ (2/2)

(2) Comparability of patterns

$\Rightarrow$The coding f guarantees the comparability of patterns, i.e. $\theta \preceq \varphi \Rightarrow f(\theta) \subseteq f(\varphi)$.

Let us consider the following coding f of sequences into sets:

- for each letter x occurring at position i in a sequence S $\Rightarrow$create a pair (i, x).

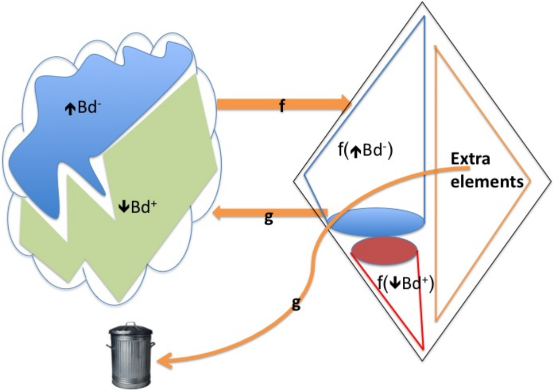
Example

Consider now two sequences *aa* and *baa*.

- $f(aa) = \{(1, a), (2, a)\}$
- $f(baa) = \{(1, b), (2, a), (3, a)\}$
- $aa \preceq baa$ but $\{(1, a), (2, a)\} \not\subseteq \{(1, b), (2, a), (3, a)\}$.

$\mathcal{WRAS}$: a new class of problems

Intuition: two functions f and g, a new bottom $\perp$ element "added" to $\mathcal{L}$, incomparability of patterns only



More formally

Definition

Let $(\mathcal{L}, \mathbf{d}, Q)$ be a pattern mining problem and $\bot$ a special pattern, $\bot \notin \mathcal{L}$. A finite set R and a pair of total functions (f, g) with $f : \mathcal{L} \rightarrow \mathcal{P}(R)$ and $g : \mathcal{P}(R) \rightarrow \mathcal{L} \cup \bot$, denoted by the triple (R, f, g), is said to be a weak representation as sets of $(\mathcal{L}, \mathbf{d}, Q)$ if

1. f and g are polynomially computable
2. for all $\theta \in \mathcal{L}$, $g(f(\theta)) = \theta$
3. for all $\theta, \varphi \in \mathcal{L}$, $f(\theta) \subseteq f(\varphi) \Rightarrow \theta \preceq \varphi$

$\mathcal{WRAS}$ = The class of such problems

➯ f is “borders-preserving” !

$\mathcal{RAS}$ vs $\mathcal{WRAS}$

Notion of extra elements

Definition

Let us denote by $\mathcal{E}$ the set of *extra elements* defined by $\mathcal{E} = \mathcal{P}(R) \setminus (\downarrow f(\mathcal{B}d^+(S)) \cup \uparrow f(\mathcal{B}d^-(S)))$.

No extra element in $\mathcal{RAS}$!
Property $(\mathcal{L}, \mathbf{d}, Q) \in \mathcal{RAS}$ implies $\mathcal{E} = \emptyset$

For $\mathcal{WRAS}$, the idea is to push those extra elements towards the positive or negative borders

$\mathcal{WRAS}$ results

Theorem

Let $(\mathcal{L}, \mathbf{d}, Q)$ be a pattern mining problem, $S \subseteq \mathcal{L}$ a downward closed set and (R, f, g) a weak representation as sets of $(\mathcal{L}, \mathbf{d}, Q)$.

(1) $\mathcal{B}d^+(S) = g(TrMin(\overline{Min_{\subseteq}(\mathcal{E} \cup f(\mathcal{B}d^-(S)))}))$

(2) $\mathcal{B}d^-(S) = g(TrMin(\overline{Max_{\subseteq}(\mathcal{E} \cup f(\mathcal{B}d^+(S)))}))$

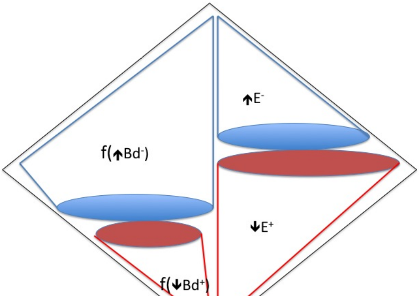
$\Rightarrow$ How to find a "condensed representation" for $\mathcal{E}$?

Notion of separating pair for extra elements

Definition

Let $\mathcal{E}^+, \mathcal{E}^- \subseteq \mathcal{E}$. $(\mathcal{E}^+, \mathcal{E}^-)$ is said to be a *separating pair* of $\mathcal{E}$ if

- $\mathcal{E}^+ \cap \mathcal{E}^- = \emptyset$,
- $\mathcal{E} \subseteq \downarrow \mathcal{E}^+ \cup \uparrow \mathcal{E}^-$,
- $f(\mathcal{B}d^+(S)) \cup \mathcal{E}^+$ and $f(\mathcal{B}d^-(S)) \cup \mathcal{E}^-$ are antichains.



The $\mathcal{EWRAS}$ class

Corollary

(1) $\mathcal{B}d^+(S) = Max_{\preceq}(g(\overline{TrMin(\mathcal{E}^- \cup f(\mathcal{B}d^-(S)))}))$

(2) $\mathcal{B}d^-(S) = Min_{\preceq}(g(\overline{TrMin(\mathcal{E}^+ \cup f(\mathcal{B}d^+(S)))}))$

Definition

$(\mathcal{E}^+, \mathcal{E}^-)$ is an efficient separating pair of $\mathcal{E}$ if $|\mathcal{E}^+|$ and $|\mathcal{E}^-|$ are bounded by a polynom in the size of the borders of $Th(\mathcal{L}, \mathbf{d}, Q)$.

$\mathcal{EWRAS}$ = $\mathcal{WRAS}$ problems having an efficient separating pair

Main theorem

The dualization problem of any $\mathcal{EWRAS}$ problem can be polynomially **reduced** to hypergraph transversal problem.

Existence of separating pairs

Do not always exist
➩ Depend of the structural properties of $(\mathcal{L}, \preceq)$

Theorem

$f(\mathcal{L})$ convex $\Rightarrow$ there exists a separating pair of $\mathcal{E}$.

No characterization of efficient separating pairs ...

For SEQ, we have shown:

1. SEQ belongs to $\mathcal{WRAS}$
2. $f(\mathcal{L}_S)$ convex
3. We have exhibited one particular efficient separating pair

➩ It follows that SEQ belong to $\mathcal{EWRAS}$

Plan

1. Preliminaries
2. Beyond $\mathcal{RAS}$
   - Weak representation as sets: The $\mathcal{WRAS}$ class
   - Efficient $\mathcal{WRAS}$: The $\mathcal{EWRAS}$ class
3. Concluding remarks

Concluding remarks

- New classes of pattern mining problems: $\mathcal{RAS} \subset \mathcal{EWRAS} \subset \mathcal{WRAS}$
- Existence of incremental quasi-polynomial time algorithms for $\mathcal{EWRAS}$
- SEQ belongs to $\mathcal{EWRAS}$

➩ very useful to clarify existing pattern mining contributions

Perspectives

- Other pattern mining problems belong to $\mathcal{EWRAS}$
- Necessary and sufficient condition for the existence of a separating pair
- Algorithmic strategies for $\mathcal{EWRAS}$

➫ Long term objective: further developing declarative approaches in data mining

`http://liris.cnrs.fr/dag`