Nouvelles classes de problèmes pour la fouille de motifs intéressants dans les bases de données [2]

Lhouari Nourine[1]

[1] Université Blaise Pascal, CNRS, LIMOS, France

SeqBio 2012
Marne la vallée, France

2. Financé par le programme DEFIS 2009 de l'ANR – Projet DAG

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

Plan

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 $Q(\mathrm{X},\mathbf{d})$.

Several questions can be asked :

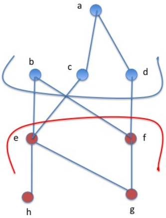
1. Frequent patterns
2. Closed frequent patterns
3. Positive and Negative borders denoted by $\mathcal{B}d^+()$ and $\mathcal{B}d^-()$.

Dualization $\Leftrightarrow$ Relationship between the two borders

Dualization

➯ Hypothesis

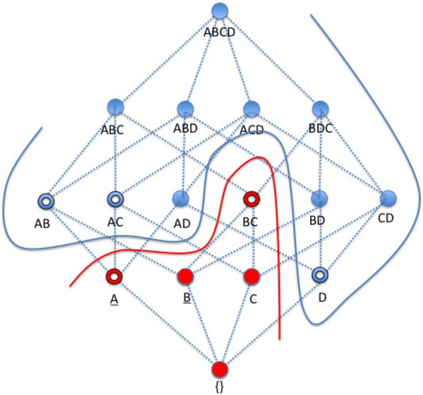
- $\mathcal{L}^*$ is structured with a partial order $\preceq$.
- Q is anti-monotonic wrt $\preceq$: for all $\theta, \varphi \in \mathcal{L}, \varphi \preceq \theta$ implies $Q(\theta) \preceq Q(\varphi)$.



➯ Maximal red elements is known as a **Positive border**

➯ Minimal blue elements is known as a **Negative border**

Dualization Problem on Boolean Lattices



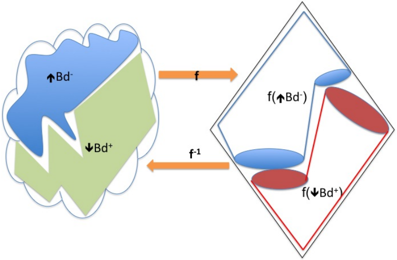
- $\mathcal{B}d^{-}(\{A, BC\}) = \{D, AB, AC\}$.

$(\mathcal{L}^*, \preceq)$ Isomorphic to a boolean lattice

Basic ideas :

- A bijection f between patterns $\mathcal{L}^*$ and the powerset of some finite set R and
- Isomorphism between $(\mathcal{L}^*, \preceq)$ and $(2^R, \subseteq)$

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



$(\mathcal{L}^*, \preceq)$ Isomorphic to a boolean lattice

Results for $\mathcal{RAS}$ problems : [Mannila &Toivonen, DMKD, 1997]

- Dualization is equivalent to minimal transversal Hypergraph,
- All algorithms for itemset mining can be used for $\mathcal{RAS}$
- Complexity depends to the computation of f^{-1}.

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

Dualization Problem

The complexity depends on the structural properties of the poset $(\mathcal{L}^*, \preceq)$.

- $(\mathcal{L}^*, \preceq)$ is isomorphic to a boolean lattice : **quasi-polynomial** (Fredman et Khachiyan 96).
- $(\mathcal{L}^*, \preceq)$ is isomorphic to a product of chains : **quasi-polynomial** (Elbassioni et al 09)
- $(\mathcal{L}^*, \preceq)$) is the set of basis of a matroid : **polynomial** (Elbassioni et al 09)
- $(\mathcal{L}^*, \preceq)$ is isomorphic to a lattice : **coNP-complete** (Babin et Kuznetsov 11).
- $(\mathcal{L}^*, \preceq)$ is a distributive lattice : **OPEN**.

Plan

Limits of $\mathcal{RAS}$

(1) The surjectivity constraint

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

(2) The embedding preserves comparability

⇒General Posets : it is not always possible preserve borders

Sequence with wildcards

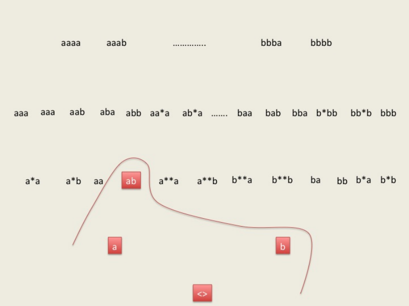
Let Σ be an alphabet and $\star \notin \Sigma$ be a wildcard. An *input sequence*, $S \in \Sigma^*$ of length $n \geq 0$.

A rigid pattern is a string $P \in (\Sigma \cup \{\star\})^*$ of length $m \leq n$ such that $P[1] \neq \star$ and $P[m] \neq \star$.

For patterns $P[1..m]$ and $Q[1..n]$, we say that P *occurs* in Q at position $p \in [1..n]$, denoted by $P \sqsubseteq_p Q$, if for every $i \in [1..m]$ $P[i] = Q[p + i - 1]$ or $P[i] = \star$.

The poset of rigid patterns

Let $\mathcal{L}_S^*$ be the set of all rigid motifs over $\Sigma \cup \star$. We have $(\mathcal{L}_S^*, \sqsubseteq)$ a partial order.



➩ $\mathcal{B}d^-(\{ab\}) = \{aa, bb, ba, a*a, a*b, b*a, b*b, a**a, a**b, b**a, b**b\}$

The poset of rigid patterns

(1) $(\mathcal{L}_S^*, \sqsubseteq)$ is not isomorphic to a boolean lattice

(1) $(\mathcal{L}_S^*, \sqsubseteq)$ cannot be embedded in a boolean lattice such that borders are preserved

$\Rightarrow$Dualisation of sequences does not belong to $\mathcal{RAS}$

➾ Mapping which does not preserve comparability ?

The encoding of Arimura 2009

Let S be a sequence of size n on Σ. The embedding is as follows :

- Let $R = \{(i, x) | i \in [1..n], x \in \Sigma\}$.
- Let $f : \mathcal{L}_S^* \to 2^R$ with

$$f(P[1..m]) = \{(i, P[i]) \mid i \in [1..m] \text{ and } P[i] \in \Sigma\}$$

For $\Sigma = \{a, b\}$ and $n = 4$, we have
$\Rightarrow R = \{(1, a), (2, a), (3, a), (4, a), (1, b), (2, b), (3, b), (4, b)\}$
$\Rightarrow f(abab) = \{(1, a), (2, b), (3, a), (4, b)\}$,
$\Rightarrow f(ab * *b) = \{(1, a), (2, b), (5, b)\}$.

The encoding of Arimura 2009

- f is not surjective :
  $\Rightarrow$Let $X = \{(1, a), (2, b)\}$ and $X' = \{((2, a), (3, b)\}$. The sequence *ab* corresponds to X while X' is not an image of f.
- f is not monotonic :
  $\Rightarrow$For the two patterns *bb* and *abb*. we have $bb \preceq abb$ whereas $f(bb) \not\subseteq f(abb)$.

Properties of the mapping

➪ Given a pattern $P \in \mathcal{L}_S^*$, then

- $f(P)$ must contain a unique symbol in each index ; and
- $f(P)$ must contains $(1, x)$ for some symbol $x \in \Sigma$.

➪ The following sets characterize elements of $\mathcal{P}(R)$ which do not have an image by the coding f.

$$\mathcal{F}^- = \{\{(i,x),(i,y)\} \text{ such that } x,y \in \Sigma, i \in [1..n]\}$$

$$\mathcal{F}^+ = \{(i,x) \text{ such that } x \in \Sigma, i \in [2..n]\}$$

Properties of the mapping

Let $\Sigma = \{a, b\}$ and $n = 4$. Then

$\mathcal{F}^+ = \{(2, a), (3, a), (4, a), (2, b), (3, b), (4, b)\}$
$\mathcal{F}^- =$
$\{\{(1, a), (1, b)\}, \{(2, a), (2, b)\}, \{(3, a), (3, b)\}, \{(4, a), (4, b)\}\}$

➯ The decoding function g is defined as follows :

$$g(X) = \begin{cases} \theta & \text{if} \quad X = f(\theta) \\ \perp & \textit{otherwise} \quad (i.e. \quad X \in \mathcal{F}^- \cup \mathcal{F}^+) \end{cases}$$

➯ Let $A \in \mathcal{P}(R)$ such that $A \notin \uparrow \mathcal{F}^-$ and $A \notin \downarrow \mathcal{F}^+$. Then $f(g(A)) = A$.

Results

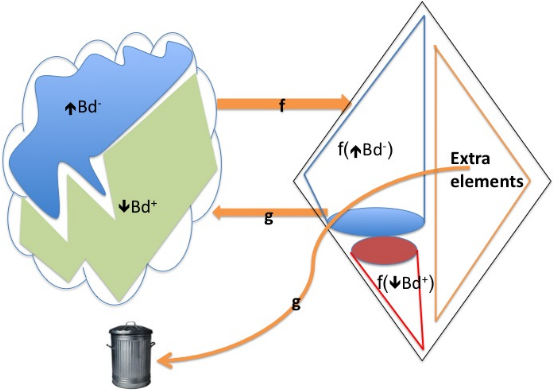
➯ There is a bijection between $\mathcal{L}_S^*$ and $\mathcal{P}(R) \setminus (\uparrow \mathcal{F}^- \cup \downarrow \mathcal{F}^+)$.

➯ $f(\mathcal{L}_S^*)$ is convex in 2^R $\Rightarrow$ Bipartition is possible

➯ f preserves the incomparability of patterns but not comparability, i.e. if $\varphi \not\preceq \theta$ then $f(\varphi) \not\subseteq f(\theta)$

- $bb \preceq abb$ but
  $f(bb) = \{(1, b), (2, b)\} \not\subseteq f(abb) = \{(1, a), (2, b), (3, b)\}$
  $\Rightarrow$ The size of the borders may increase

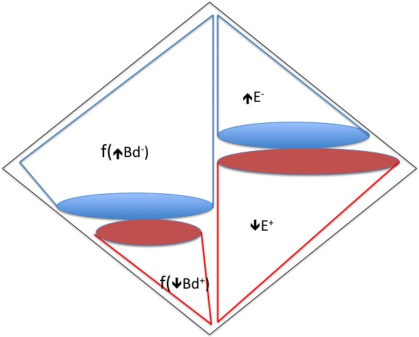
Results



Results

Theorem

The extra elements can be partionned into two parts $\mathcal{E}^+$ and $\mathcal{E}^-$



Results

Lemma

$\mathcal{E}^+$ and $\mathcal{E}^-$ are computable in polynomial time in the size of the two borders.

Main theorem

The dualization problem of sequences can be polynomially **reduced** to hypergraph transversal problem.

Plan

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