A Global Constraint for Closed Frequent Itemset Mining

N. Lazaar\textsuperscript{2} \quad Y. Lebbah\textsuperscript{1} \quad S. Loudni\textsuperscript{3} \quad \textbf{M. Maamar}\textsuperscript{1,2} \\
V. Lemière\textsuperscript{3} \quad C. Bessiere\textsuperscript{2} \quad P. Boizumault\textsuperscript{3}

\textsuperscript{1} Lab. LITIO, University of Oran 1 – Algeria \\
\textsuperscript{2} Lab. LIRMM, University of Montpellier – France \\
\textsuperscript{3} Lab. GREYC, University of Caen – France

\textbf{CP 2016}

\textit{Toulouse, France, 6 September 2016}
• It aims at finding **regularities** in a dataset

• *Find sets of products that are frequently bought together*
Set of items: \( I = \{A,B,C,D,E,F,G,H\} \)
Set of transaction: \( T = \{t1,t2,t3,t4,t5\} \)
Itemset: \( P \subseteq I \)

Cover:
Cover(AD)= \( \{t2,t3\} \)
Cover(BEFG)= \( \{t5\} \)
• The frequency of a itemset is the size of its cover.
• $\theta \in \mathbb{N}^+$ be the minimum support.

Frequent itemset mining problem:
• Extract all itemsets $P$ satisfying: $\text{freq}(P) \geq \theta$

<table>
<thead>
<tr>
<th>trans</th>
<th>items</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>B C</td>
</tr>
<tr>
<td>t2</td>
<td>A</td>
</tr>
<tr>
<td>t3</td>
<td>A C</td>
</tr>
<tr>
<td>t4</td>
<td>A D</td>
</tr>
<tr>
<td>t5</td>
<td>B</td>
</tr>
</tbody>
</table>

freq(BC) = 1
freq(EG) = 2

Frequent itemset with $\theta = 3$

Closed itemsets are closed

Can we compact these itemsets?
Closed Frequent Itemset Mining

The Need:

Extract all Closed Frequent Itemsets

How?
Many efficient algorithms for mining closed itemsets ✔

but

• dedicated to particular classes of constraints
  adding new constraints requires new implementations.

→ CP framework:
  • modeling
    - in a declarative way -> facilitates the addition of new constraints
  • solving
    - efficient solvers based on filtering
State of the art: CP approach

• **Reified model** [De Raedt et al, 2008]:

  **Variables:**
  - *item* variables: decision
  - *transaction* variables: auxiliary

  **Reified Constraints:**
  - Coverage
  - Frequency
  - Closedness

**Drawbacks**
- Additional dimension of transaction variables.
- The huge number of reified constraints -> limitation
Proposition: a **global constraint** that encodes efficiently the Closed Frequent itemsets Mining problem.

- Domain consistency with polynomial algorithm.
- No reified constraints/extra variables.
- Backtrack-free.
**CLOSEDPATTERN: Definition**

- **Encoding:** $P = \{P_1...P_n\}$: $D(P_i) = \{0,1\}$
- **Definition:**

  $\text{CLOSEDPATTERN}_{D,\theta}(P) : (\text{freq}(P) \geq \theta) \land (P \text{ is a closed pattern})$

  $\text{CLOSEDPATTERN}_{D,2}(AD)$

  \[
  \begin{cases}
  \text{freq}(AD) > 2 \\
  \text{AD is closed}
  \end{cases}
  \]
CLOSEDPATTERN: Filtering rules

3 Filtering rules

$0 \notin D(P_j)$:

**Rule 1:** full extension items

$1 \notin D(P_j)$:

**Rule 2:** infrequent items

**Rule 3:** Absent items

\[ D \text{ must be present} \]
\[ \theta = 2 \]
\[ H \text{ is infrequent} \]
\[ E \text{ assigned to } 0 \]
\[ G \Rightarrow E \]
\[ E = 0 \Rightarrow G = 0 \]
Algorithm: \( n \): items, \( m \): transactions

for each free variable

- **rule 1:** maintained \((O(n \times m))\)
- **rule 2:** maintained \((O(n \times m))\)

for each absent item

- **rule 3:** maintained \((O(n \times (n \times m)))\)

Time: \( O(n \times (n \times m)) \): Cubic
Space: \( O(n \times m) \): Quadratic
At each node:
0 $\notin D(P_j)$ : rule1
1 $\notin D(P_j)$ : rule2
1 $\notin D(P_j)$ : rule3

$\theta = 2$
Lex : variables
max_val : values

CLOSEDPATTERN vs Reified
• DC at each node on Boolean variables.

• Tree search size:
  full binary tree -> $2 \times |C| - 1$

• Backtrack-free:
  $O(|C| \times n^2 \times m)$
**CLOSEDPATTERN: Experiments**

- **Comparison with:**
  - The most efficient CP method: CP4IM (reified)
  - The most efficient ad hoc algorithm: LCM-v5.2

- **Solver:** or-tools, Intel Xeon E5-2680@ 2,5 GHz with 128 Gb

- **CLOSEDPATTERN-DC:** rules 1,2 and 3 (cubic pruning)
- **CLOSEDPATTERN-WC:** rules 1 and 2 (quadratic pruning)
**CLOSEDPATTERN: Experiments**

**CLOSEDPATTERN-DC vs CP4IM:**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>θ</th>
<th>Times(s)</th>
<th></th>
<th>#Failures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DC</td>
<td>CP4IM</td>
<td>DC</td>
</tr>
<tr>
<td>Chess</td>
<td>%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>45.92</td>
<td>136.31</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>187.89</td>
<td>467.52</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>969.40</td>
<td>1950.5</td>
<td>0</td>
</tr>
<tr>
<td>Mushroom</td>
<td>1</td>
<td>1.74</td>
<td>24.06</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>3.62</td>
<td>29.84</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>5.40</td>
<td>41.38</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>6.37</td>
<td>43.69</td>
<td>0</td>
</tr>
<tr>
<td>Pumsb</td>
<td>80</td>
<td>133.97</td>
<td>OOM</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>271</td>
<td>OOM</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>509.79</td>
<td>OOM</td>
<td>0</td>
</tr>
</tbody>
</table>

- Backtrack-free
**CLOSEDPATTERN: Experiments**

**CLOSEDPATTERN-DC vs CLOSEDPATTERN-WC:**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>θ</th>
<th>Times(s)</th>
<th>#Failures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>DC</td>
<td>WC</td>
</tr>
<tr>
<td>Chess</td>
<td>30</td>
<td>45.92</td>
<td>109.36</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>187.89</td>
<td>480.52</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>969.40</td>
<td>2288</td>
</tr>
<tr>
<td>Mushroom</td>
<td>1</td>
<td>1.74</td>
<td>3.42</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>3.62</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>5.40</td>
<td>9.96</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>6.37</td>
<td>9.94</td>
</tr>
<tr>
<td>Pumsb</td>
<td>80</td>
<td>133.97</td>
<td>640.59</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>271</td>
<td>1010.4</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>509.79</td>
<td>2150.8</td>
</tr>
</tbody>
</table>

Third rule reduces considerably the explored nodes.
### ClosedPattern: Experiments

#### ClosedPattern-DC vs CP4IM vs LCM:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>θ (%)</th>
<th>Times(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DC</td>
<td>CP4IM</td>
</tr>
<tr>
<td>Chess</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>45.92</td>
<td>136.31</td>
</tr>
<tr>
<td>20</td>
<td>187.89</td>
<td>467.52</td>
</tr>
<tr>
<td>10</td>
<td>969.40</td>
<td>1950.5</td>
</tr>
<tr>
<td>Mushroom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.74</td>
<td>24.06</td>
</tr>
<tr>
<td>0.5</td>
<td>3.62</td>
<td>29.84</td>
</tr>
<tr>
<td>0.1</td>
<td>5.40</td>
<td>41.38</td>
</tr>
<tr>
<td>0.05</td>
<td>6.37</td>
<td>43.69</td>
</tr>
<tr>
<td>Pumsb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>133.97</td>
<td>OOM</td>
</tr>
<tr>
<td>75</td>
<td>271</td>
<td>OOM</td>
</tr>
<tr>
<td>70</td>
<td>509.79</td>
<td>OOM</td>
</tr>
</tbody>
</table>
The aim: find $k$ closed itemsets:

(i) \texttt{CLOSEDPATTERN-DC}
(ii) Distinct itemsets
(iii) $lb < \text{size} < ub$

\begin{align*}
\text{chess (}\theta = 80\%, \ lb = 2, \ ub = 10) \\
\text{connect (}\theta = 90\%, \ lb = 2, \ ub = 10)
\end{align*}
The aim: find $k$ closed itemsets:

(i) \textsc{ClosedPattern-DC}

(ii) Distinct itemsets

(iii) $lb < size < ub$
**CLOSEDPATTERN: Experiments**

*k* itemsets instance

**The aim:** find *k* closed itemsets:

(i) `CLOSEDPATTERN-DC`
(ii) Distinct itemsets
(iii) $lb < size < ub$

---

**chess** ($\theta = 80\%$, $lb = 2$, $ub = 10$)

**connect** ($\theta = 90\%$, $lb = 2$, $ub = 10$)
**CLOSEDPATTERN: Experiments**

**k itemsets instance**

**The aim:** find $k$ closed itemsets:

(i) $\text{CLOSEDPATTERN-DC}$

(ii) Distinct itemsets

(iii) $lb < size < ub$

---

**chess ($\theta = 80\%, \ lb = 2, \ ub = 10$)**

**connect ($\theta = 90\%, \ lb = 2, \ ub = 10$)**
Conclusion:

- A global constraint for Closed Frequent Itemset ensuring DC
- No need for reified constraints/extra variables
- Filtering algorithm cubic in time, quadratic in space.