Symmetries in Sequential Pattern Mining

Lakhdar Saïs

CRIL - CNRS UMR 8188
Université d'Artois, France

Joint work with
Ikram Nekkache, Said Jabbour, Nadjet Kamel

http://www.cril.univ-artois.fr/decMining/

ISIP'2019, Heraklion, 9-10 May 2019

Outline

Motivation

Pattern mining

- **extract regularities** from data

Output of huge size

- Difficult to retrieve useful information.
- **Reducing the size of the output** is crucial for practical data mining
  - Search for condensed representations (e.g. closed, maximal patterns, etc.)

$\Rightarrow$ **Symmetry** ?

- for extracting regularities from data (kind of patterns?)
- for reducing the size of the output (kind of condensed representation?)

Symmetries

- Fundamental concept (structural knowledge) in Computer Science, Mathematics, Physics, Chemistry and many other domains.
  - Many human artifacts (e.g. classroom in a university, aircraft seats, circuit patterns) and entities in nature (e.g. plants, molecules, DNA sequences, atoms) exhibits symmetries.
  - $\Rightarrow$ Useful for reasoning and understanding more complex entities and systems.

Symmetries in SAT, CP and MP (ILP)

- Propositional Satisfiability (SAT)
  - Symmetry resolution proof system [Krishnamurthy 1985]
  - Tractability through symmetries in propositional calculus [Benhamou & Saïs 1992]
  - Symmetry breaking predicates [Crawford 1992]
- Constraint Programming (CP)
  - Variable and value symmetries [Puget 1993]
  - Symmetry in Constraint Programming [Gent et al. 2006]
- Mathematical Programming (MP)
  - Symmetry in Integer Linear Programming [Margot 2010]
  - Reformulations in mathematical programming: Automatic symmetry detection and exploitation [Liberti at al. 2012]

Outline

Symmetries in Itemset Mining?

1. Shin-ichi Minato: *Symmetric Item Set Mining Based on Zero-Suppressed BDDs.* DS 2006: 321-326

2. $\Rightarrow$ Said Jabbour, Lakhdar Sais, Yakoub Salhi, Karim Tabia: *Symmetries in Itemset Mining.* ECAI 2012: 432-437

How to exploit symmetries in itemset mining?

1. by dynamic integration in Apriori-like algorithms for search space pruning.

2. by rewriting the transaction databases in a preprocessing step (items elimination).
   - $\rightarrow$ new transaction database + symmetry group.
   - $\rightarrow$ condensed representation of the output.

Symmetry in Frequent Itemset Mining

Definition (Symmetry)

A symmetry of $\mathcal{D}$ is a permutation $\sigma \in \mathcal{P}(\mathcal{I})$ such that there exists a transaction renaming f over $\mathcal{T}_{id}(\mathcal{D})$ where $\sigma(\mathcal{D}) = f(\mathcal{D})$

σ = (C,E)(D,F) is a symmetry

t_i	itemset
001	A, B, E, F
002	A, B, C, D
003	C, D, E, F
004	A, C,
005	A, E,
006	C, E,
007	B, D,
008	B, F,
009	D, F,

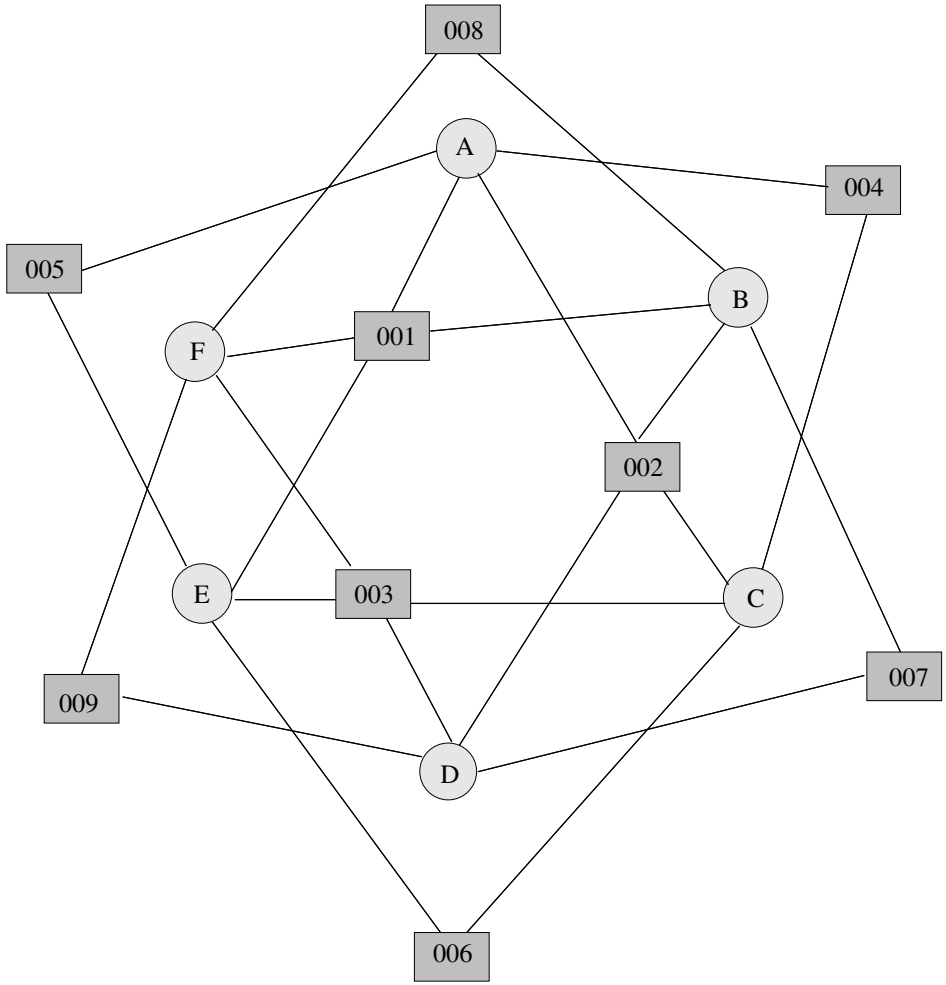
$$f(t_i) = \begin{cases} 001 & \text{if } t_i\text{=002} \\ 002 & \text{if } t_i\text{=001} \\ 003 & \text{if } t_i\text{=003} \\ 004 & \text{if } t_i\text{=005} \\ 005 & \text{if } t_i\text{=004} \\ 006 & \text{if } t_i\text{=006} \\ 007 & \text{if } t_i\text{=008} \\ 008 & \text{if } t_i\text{=007} \\ 009 & \text{if } t_i\text{=009} \end{cases}$$

Proposition

Let σ a symmetry of $\mathcal{D}$, λ a minimal support threshold and I an itemset. $I \in \mathcal{FIM}(\mathcal{D}, \lambda)$ iff $\sigma(I) \in \mathcal{FIM}(\mathcal{D}, \lambda)$.

Symmetry Detection in Transaction Databases

- Convert the original problem $\mathcal{D}$ into a colored undirected graph $\mathcal{G}$, where vertices are labeled with colors.
- Look for the automorphism group of $\mathcal{G}$.
- Symmetries of $\mathcal{D}$ are equivalent to the automorphisms of the colored undirected graph $\mathcal{G}$.
- Employ a general-purpose graph symmetry tool to uncover the symmetries [Mckay'81, Aloul'03].



t_i	itemset
001	A, B, E, F,
002	A, B, C, D
003	C, D, E, F
004	A, C
005	A, E
006	C, E
007	B, D
008	B, F
009	D, F

Symmetry Pruning

Integration in Apriori-like algorithm

$\rightarrow$ proceeds by a level-wise search of the elements of $\mathcal{FIM}(\mathcal{D}, \lambda)$.

1. Starts by computing the elements of $\mathcal{FIM}(\mathcal{D}, \lambda)$ of size 1.
2. Assuming $\mathcal{FIM}(\mathcal{D}, \lambda)$ of size n known, computes a set of candidates of size $n + 1$ so that I is a candidate if and only if all its subsets are in $\mathcal{FIM}(\mathcal{D}, \lambda)$.
3. This procedure is iterated until no more candidate is found.

Symmetry-Based Pruning in Apriori-like algos

- Let $\mathcal{D}$ be a transaction database such that $\mathcal{I}(\mathcal{D}) = \{A, B, C, D\}$ and σ is a symmetry such that σ=$(A, D)(B, C)$.
- Assume that the itemsets $\{A\}$, $\{B\}$, $\{C\}$ and $\{D\}$ are frequent. We also assume that in iteration 2, we find that the itemset $\{A, B\}$ is not frequent.

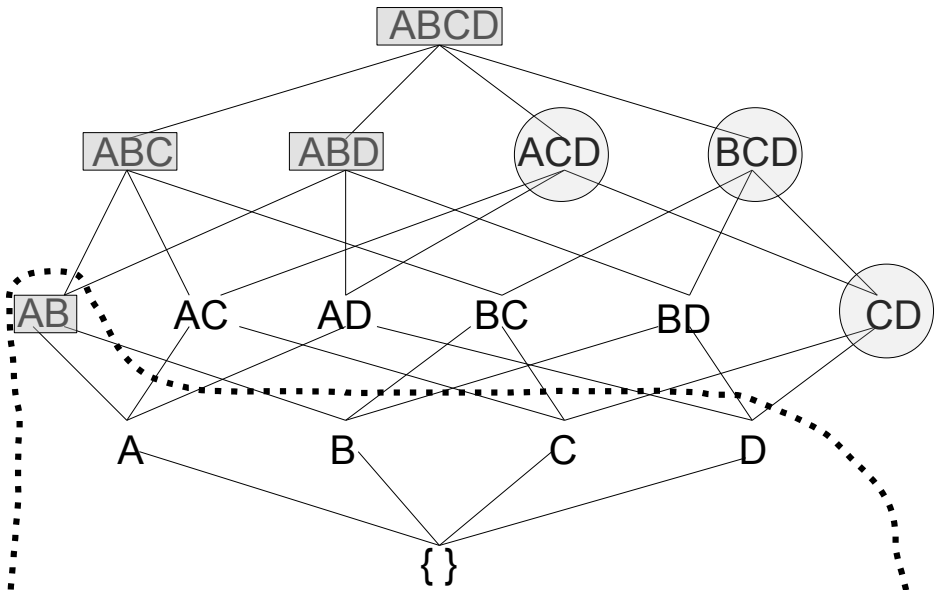


Figure: Symmetry Pruning

Symmetry Breaking

- Breaking symmetries in a preprocessing step.
  - Eliminate items from the original transaction database.
  - The frequent itemsets generated using the new transaction database together with the symmetry group can be used to retrieve the whole set of frequent itemsets of the original

Symmetry Breaking

Let $\mathcal{D}$ a transaction database and $\sigma = (a, b)(c, d)$ a symmetry

$\mathcal{FIM}(\mathcal{D}, \lambda)$=$\{a, \dots\}$, $\{b, \dots\}$, $\{c, \dots\}$, $\{d, \dots\}$, $\{a, b, \dots\}$, $\{a, c, \dots\}$, $\{a, d, \dots\}$, $\{b, c, \dots\}$, $\{b, d, \dots\}$, $\{c, d, \dots\}$

$$\begin{array}{lcl} \{a, \dots\} & \rightarrow & \{b, \dots\} \\ \{a, d, \dots\} & \rightarrow & \{b, c, \dots\} \\ \{a, c, \dots\} & \rightarrow & \{b, d, \dots\} \\ \{a, b, \dots\} & \rightarrow & \{a, b, \dots\} \end{array} \qquad \begin{array}{lcl} \{b, \dots\} & \rightarrow & \{a, \dots\} \\ \{b, c, \dots\} & \rightarrow & \{a, d, \dots\} \\ \{d, \dots\} & \rightarrow & \{c, \dots\} \\ \{b, d, \dots\} & \rightarrow & \{a, c, \dots\} \end{array}$$

Symmetry Breaking

Let $\mathcal{D}$ a transaction database and $\sigma = (a, b)(c, d)$ a symmetry

$\mathcal{FIM}(\mathcal{D}, \lambda)$=$\{a, \dots\}$, $\{b, \dots\}$, $\{c, \dots\}$, $\{d, \dots\}$, $\{a, b, \dots\}$, $\{a, c, \dots\}$, $\{a, d, \dots\}$, $\{b, c, \dots\}$, $\{b, d, \dots\}$, $\{c, d, \dots\}$

$$
\begin{array}{lcl}
\{a, \dots\} & \rightarrow & \{b, \dots\} \\
\{a, d, \dots\} & \rightarrow & \{b, c, \dots\} \\
\{a, c, \dots\} & \rightarrow & \{b, d, \dots\} \\
\{a, b, \dots\} & \rightarrow & \{a, b, \dots\}
\end{array}
\qquad
\begin{array}{lcl}
\{b, \dots\} & \rightarrow & \{a, \dots\} \\
\{b, c, \dots\} & \rightarrow & \{a, d, \dots\} \\
\{d, \dots\} & \rightarrow & \{c, \dots\} \\
\{b, d, \dots\} & \rightarrow & \{a, c, \dots\}
\end{array}
$$

Symmetry Breaking

Let $\mathcal{D}$ a transaction database and $\sigma = (a, b)(c, d)$ a symmetry

$\mathcal{FIM}(\mathcal{D}, \lambda)$=$\{a, \ldots\}$, $\{b, \ldots\}$, $\{c, \ldots\}$, $\{d, \ldots\}$, $\{a, b, \ldots\}$,
$\{a, c, \ldots\}$, $\{a, d, \ldots\}$, $\{b, c, \ldots\}$, $\{b, d, \ldots\}$, $\{c, d, \ldots\}$ + σ

$\{a, \ldots\}$	$\rightarrow$	$\{b, \ldots\}$	$\{b, \ldots\}$	$\rightarrow$	$\{a, \ldots\}$
$\{a, d, \ldots\}$	$\rightarrow$	$\{b, c, \ldots\}$	$\{b, c, \ldots\}$	$\rightarrow$	$\{a, d, \ldots\}$
$\{a, c, \ldots\}$	$\rightarrow$	$\{b, d, \ldots\}$	$\{d, \ldots\}$	$\rightarrow$	$\{c, \ldots\}$
$\{a, b, \ldots\}$	$\rightarrow$	$\{a, b, \ldots\}$	$\{b, d, \ldots\}$	$\rightarrow$	$\{a, c, \ldots\}$

- $\Rightarrow$ b can be removed from each $T \in \mathcal{D}$ if $\{a, b\} \not\subset T$
- $\Rightarrow$ d can be removed from each $T \in \mathcal{D}$ if $\{a, d\} \not\subset T$ and $\{c, d\} \not\subset T$

Symmetry Breaking

Proposition

Let $\mathcal{D}$ a transaction database and

$\sigma = (x_1, y_1)(x_2, y_2) \cdots (x_j, y_j) \cdots (x_n, y_n)$ a symmetry

$\Rightarrow y_j$ can be removed from each $T \in \mathcal{D}$ if $\{x_i, y_j\} \not\subset T, \forall i \leqslant j$

Remark

Symmetries can be broken independently

Symmetry Breaking: an example

t_i	itemset			
001	A,	B,	E,	F,
002	A,	B,	C,	D
003	C,	D,	E,	F
004	A,	C		
005	A,	E		
006	C,	E		
007	B,	D		
008	B,	F		
009	D,	F		

σ_1 = (A C)(B, D)
σ_2 = (A B)(C, D) (E F)
σ_3 = (C,E)(D,F)

t_i	itemset			
001	A,	B,	E,	F
002	A,	B,	C,	D
003	C	D	E	F
004	A,	C,		
005	A,	E,		
006	C	E		
007	B	D		
008	B	F		
009	D	F		

Table: Itempair-based Symmetry Breaking approach

Symmetry Breaking: an example

t_i	itemset
001	A, B, E, F,
002	A, B, C, D
003	C, D, E, F
004	A, C
005	A, E
006	C, E
007	B, D
008	B, F
009	D, F

σ_1 = (A C)(B, D)
σ_2 = (A B)(C, D) (E F)
σ_3 = (C,E)(D,F)

t_i	itemset
001	A, B,
002	A, B, C, D
003	
004	A, C
005	A
006	
007	
008	
009	

Table: Itempair-based Symmetry Breaking approach

Outline

Symmetry in Sequence Pattern Mining

Example

Let us consider the following sequence database over the set of symbols $\mathcal{I} = \{a, b, c, d, e, f\}$:

SID	Sequence
1	$\langle\{a, b\}, \{c\}, \{d, e\}\rangle$
2	$\langle\{b, c\}, \{a\}, \{d, e\}\rangle$
3	$\langle\{c, d\}, \{a\}, \{b, f\}\rangle$
4	$\langle\{a, d\}, \{c\}, \{b, f\}\rangle$

Table: An example of sequence database Δ

For instance, taking $\lambda = 2$, the sequence $s = \langle\{b\}, \{d, e\}\rangle$ is frequent as $support_\Delta(s) = |\{1, 2\}| = 2$.

Symmetry in Sequence Pattern Mining

SID	Sequence
1	$\langle\{a, b\}, \{c\}, \{d, e\}\rangle$
2	$\langle\{b, c\}, \{a\}, \{d, e\}\rangle$
3	$\langle\{c, d\}, \{a\}, \{b, f\}\rangle$
4	$\langle\{a, d\}, \{c\}, \{b, f\}\rangle$

$\sigma_1 = (a, c)$ and $\sigma_2 = (b, d)(e, f)$ are two symmetries of Δ.
$\sigma_1(\Delta) = f_1(\Delta)$ where f_1 is a sequence-id renaming:

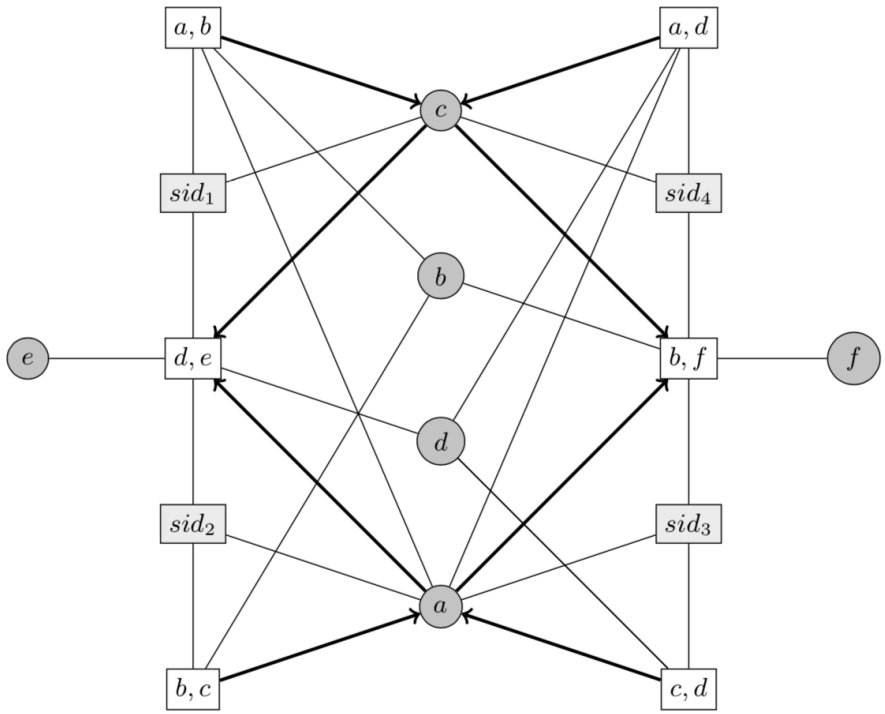
$$f_1(x) = \begin{cases} 2 & \text{if x=1} \\ 1 & \text{if x=2} \\ 4 & \text{if x=3} \\ 3 & \text{if x=4} \end{cases}$$

and $\sigma_2(\Delta) = f_2(\Delta)$ where f_2 is a sequence-id renaming:

$$f_2(x) = \begin{cases} 4 & \text{if x=1} \\ 3 & \text{if x=2} \\ 2 & \text{if x=3} \\ 1 & \text{if x=4} \end{cases}$$

Symmetry Detection Sequence Database

SID	Sequence
1	$\langle\{a, b\}, \{c\}, \{d, e\}\rangle$
2	$\langle\{b, c\}, \{a\}, \{d, e\}\rangle$
3	$\langle\{c, d\}, \{a\}, \{b, f\}\rangle$
4	$\langle\{a, d\}, \{c\}, \{b, f\}\rangle$



From Sequence Database to colored directed graph

Experiments

Symmetry detection tool:

- **bliss**: A Tool for Computing Automorphism Groups and Canonical Labelings of Graphs [1]

Dataset	$\lvert\Delta\rvert$	$\lvert\mathcal{I}\rvert$	#sym	time(s)
BIBLE	36369	13905	240	657
FIFA	20450	2990	7	235
Kosarak	25000	21144	1244	308
Leviathan	5834	9025	46	57
PubMed	17237	19931	72	310

Table: Symmetries on real sequence databases (Sequences of items)

[1] http://www.tcs.hut.fi/Software/bliss/index.html

Experiments

Dataset	$\lvert\Delta\rvert$	$\lvert\mathcal{I}\rvert$	#sym	time(s)
1	50351	37911	2984	177
2	47784	47850	2414	264
3	47555	55943	2096	352
4	47987	62413	1669	452
5	48466	67469	1332	560
6	45534	37098	2858	125
7	45451	47075	2506	196
8	46130	54683	2081	295
9	47133	60892	1801	362

Table: Symmetries on synthetic sequence databases (Sequences of itemsets)

[2]

[2] Generated with the IBM Quest Dataset Generator

Outline

Conclusion

A first step towards introducing symmetries in pattern mining

Remains to be done:

- Find real-world sequence databases where symmetry might reveal useful informations.
- Symmetry integration to sequential pattern mining algoritms

Thank you for your attention