On Preference-based (soft) Pattern Sets

Patrice Boizumault, Bruno Crémilleux
Samir Loudni and Willy Ugarte

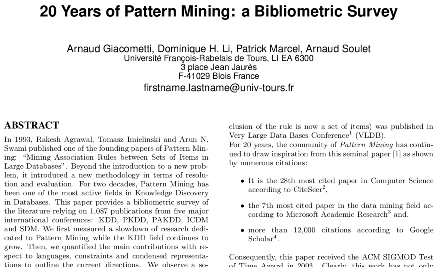
GREYC (CNRS UMR 6072)
University of Caen - France

NormaSTIC one-day workshop
Caen, May 19, 2015

On the “life story” of pattern mining

At the beginning: a runtime challenge...

“My algorithm is faster than the previous ones” (or at least some ones...)



20 Years of Pattern Mining: a Bibliometric Survey

Arnaud Giacometti, Dominique H. Li, Patrick Marcel, Arnaud Soulet
Université François-Rabelais de Tours, LI EA 6300
3 place Jean Jaurès
F-41029 Blois France
firstname.lastname@univ-tours.fr

ABSTRACT

In 1993, Rakesh Agrawal, Tomasz Imielinski and Arun N. Swami published one of the founding papers of Pattern Mining: “Mining Association Rules between Sets of Items in Large Databases”. Beyond the introduction to a new problem, it introduced a new methodology in terms of resolution and evaluation. For two decades, Pattern Mining has been one of the most active fields in Knowledge Discovery in Databases. This paper provides a bibliometric survey of the literature relying on 1,087 publications from five major international conferences: KDD, PKDD, PAKDD, ICDM and SDM. We first measured a slowdown of research dedicated to Pattern Mining while the KDD field continues to grow. Then, we quantified the main contributions with respect to languages, constraints and condensed representations to outline the current directions. We observe a so-

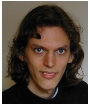
clusion of the rule is now a set of items) was published in Very Large Data Bases Conference[1] (VLDB).
For 20 years, the community of *Pattern Mining* has continued to draw inspiration from this seminal paper [1] as shown by numerous citations:

- It is the 28th most cited paper in Computer Science according to CiteSeer[2],
- the 7th most cited paper in the data mining field according to Microsoft Academic Research[3] and,
- more than 12,000 citations according to Google Scholar[4].

Consequently, this paper received the ACM SIGMOD Test of Time Award in 2003. Clearly, this work has not only

SIGKDD Explorations 2013

but, *a well-known limitation:*
too many results including many non-informative patterns, difficulty to grasp



“Is pattern mining dead or alive?”
Siegfried Nijssen, SML 2014

An up-to-date interest: from efficiency-based approaches to methods able to extract more meaningful patterns

Challenge: how to discover a manageable set of high-level and useful patterns?

- **constraint-based pattern mining**[1] (Mannila at al. DMKD'97, Ng et al. SIGMOD'98): but how to define proper constraints?
- **pattern condensed representations** (Pasquier et al. ICDT'99, Boulicaut et al. DMKD'03): designed to speed up the extraction, but closed/free patterns have many uses
- **interestingness/statistically measures/preferences** (Geng et al. ACM Computing Survey'06, Hämäläinen et al. ICDM'08)
- **a small set of patterns that compress** (Siebes et al. SDM'06)
- **pattern sets** (Knobbe et al. ECML/PKDD'06, Xin et al. KDD'06), **constraint-based pattern set mining** (De Raedt et al. SDM'07), **pairwise comparisons** (Negrevergne et al. ICDM'13, Ugarte et al. RFIA'14)
- **n-ary patterns/k-pattern sets** (Khiari et al. CP'10, Guns et al. TKDE'13)
- **global patterns** (Crémilleux et al. ICCSA'08, Giacometti et al. IDEAL'09)
- **integrating background knowledge**

In this talk: we investigate the use of user preferences based on measures

[1] Here "constraint" means: "focus on the most promising patterns"

How to get useful information in pattern mining?

What about user preferences? Examples with measures:

- *"the higher the frequency, growth rate and aromaticity are, the better the patterns"*
- *"I prefer pattern X_1 than pattern X_2 if X_1 is not dominated by X_2 according to a set of measures"*

In this talk:
skyline patterns (i.e. skypatterns) are the common theme.

Outline

- *What are the best patterns according to a set of measures?*
  a proposition: use the Pareto dominance relation

  ➥ mining skypatterns by using Constraint Programming (CP)
  (and soft-skypatterns for $\simeq$ free!)

- *What measures to keep?* Keep all the measures!

  ➥ from skypatterns to skypattern cube

- To sum up and perspectives

Skypatterns

Skypatterns: motivations

- *give the end-user an (easy) way to express his preferences according to measures*:

$$\textit{measures} \begin{cases} \textit{constraint-based data mining}: \textit{frequency}, \textit{size}, \ldots \\ \textit{background knowledge}: \textit{price}, \textit{weight}, \textit{aromaticity}, \ldots \\ \textit{statistics}: \textit{entropy}, \textit{pvalue}, \ldots \end{cases}$$

  ➥ several types of measures can be combined

- *avoid the threshold issue:*
  - what is a suitable value of the minimal frequency?
    ➥ a well-known limitation in the constraint-based pattern paradigm
  - combining several measures: how to fix several thresholds?

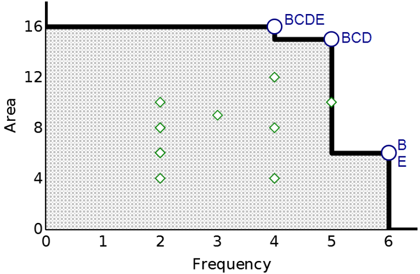
- *discovering patterns satisfying a global property*

  ➥ Pareto dominance relation

Skypatterns: an example

notion of skylines (database) in pattern mining (Soulet et al. ICDM'11)

Tid	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F



Patterns	**freq**	**area**
~~*AB*~~	2	4
~~*AEF*~~	2	6
B	6	6
BCDE	4	16
~~*CDEF*~~	2	8
E	6	6
⋮	⋮	⋮

$|\mathcal{L}_{\mathcal{I}}| = 2^6$, but only 4 skypatterns

$$Sky(\mathcal{L}_{\mathcal{I}}, \{freq, area\}) = \{BCDE, BCD, B, E\}$$

freq, *area*: *constraint-based data mining* measures

Many other measures can be addressed:

- *background knowledge:* price, aromaticity,. . .
- *statistics:* p-value,. . .

Skypatterns: more formally

M: a set of measures $\mathcal{I}$: a set of items $\mathcal{L}_{\mathcal{I}} = 2^{\mathcal{I}}$: set of patterns

Pattern (Pareto)-dominance: a pattern X_i dominates a pattern X_j w.r.t. M denoted $X_i \succ_M X_j$ iff
$\forall m \in M, m(X_i) \geq m(X_j) \wedge \exists m \in M, m(X_i) > m(X_j)$

➥ a skypattern of $\mathcal{L}_{\mathcal{I}}$ w.r.t to M is a *pattern not dominated in $\mathcal{L}_{\mathcal{I}}$ w.r.t M*

The **skypattern operator** $\mathcal{S}ky$ returns all the skypatterns w.r.t M:

$$\mathcal{S}ky(\mathcal{L}_{\mathcal{I}}, M) = \{X \in \mathcal{L}_{\mathcal{I}} | \nexists Y \in \mathcal{L}_{\mathcal{I}} : Y \succ_M X\}$$

Skylines vs skypatterns

Problem	Skylines	Skypatterns
Mining task	a set of non dominated transactions	a set of non dominated patterns
Size of the space search	$\mid \mathcal{T} \mid$	$\mid \mathcal{L}_{\mathcal{I}} \mid = \mid 2^{\mathcal{I}} \mid$
domain	a lot of works	very few works

usually: $\mid \mathcal{T} \mid << \mid \mathcal{L}_{\mathcal{I}} \mid$

$\mathcal{T}$	set of transactions
$\mathcal{I}$	set of items
$\mathcal{L}_{\mathcal{I}}$	set of patterns

Skypatterns: how to process?

A naive enumeration of all candidate patterns ($\mathcal{L}_{\mathcal{I}}$) and then comparing them is not feasible...

Two key principles:

- take benefit from the pattern condensed representation according to the condensable measures of M
  (Soulet et al. DMKD'08)

  ➥ for that purpose: skylineability to obtain M' ($M' \subseteq M$) giving a more concise pattern condensed representation

- use of Dynamic CSP to increasingly reduce the dominance area by processing pairwise comparisons between patterns

Mining skypatterns using CP (1/3)
(Ugarte et al. CPAIOR'14)

1. **Principle:**
   - starting from an initial pattern s_1 closed w.r.t. M'
   - searching a pattern s_2 not dominated by s_1
   - searching a pattern s_3 not dominated by s_1 or s_2

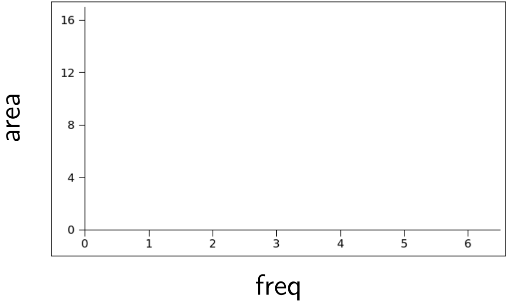
     $\vdots$
   - until there is no pattern satisfying these constraints

2. **Solving:**
   - constraints are dynamically posted during the mining step (for each candidate s_i, add the constraint $\neg(s_i \succ_M X)$)

➥ the dominance area is increasingly reduced thanks to the filtering.

Mining skypatterns using CP (2/3)

Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F



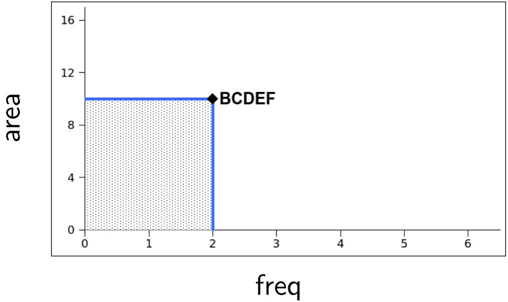
$$M = \{freq, area\}$$

$$q(X) \equiv closed_{M'}(X)$$

$$Candidates =$$

Mining skypatterns using CP (2/3)

Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F



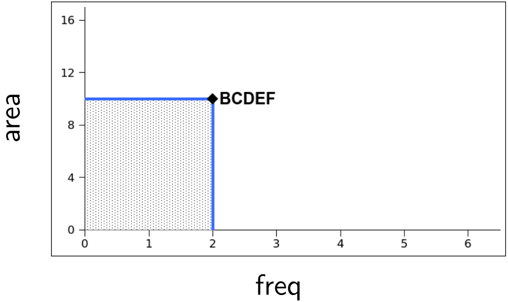
$$M = \{freq, area\}$$

$$q(X) \equiv closed_{M'}(X)$$

$$Candidates = \{\underbrace{\text{BCDEF}}_{s_1},$$

Mining skypatterns using CP (2/3)

Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F



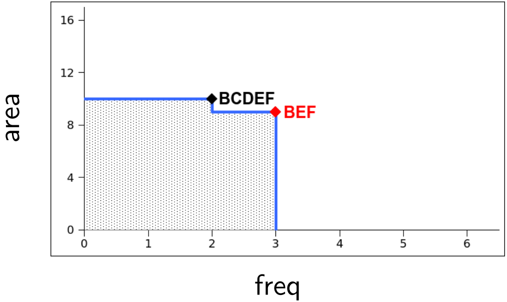
$$M = \{freq, area\}$$

$$q(X) \equiv closed_{M'}(X) \wedge \neg(s_1 \succ_M X)$$

$$Candidates = \{\underbrace{\text{BCDEF}}_{s_1},$$

Mining skypatterns using CP (2/3)

Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F



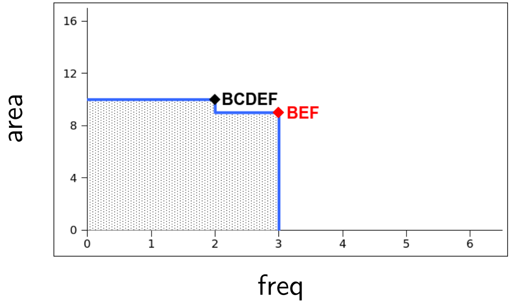
$$M = \{freq, area\}$$

$$q(X) \equiv closed_{M'}(X) \wedge \neg(s_1 \succ_M X)$$

$$Candidates = \{\underbrace{BCDEF}_{s_1}, \underbrace{BEF}_{s_2},$$

Mining skypatterns using CP (2/3)

Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F



$$M = \{freq, area\}$$

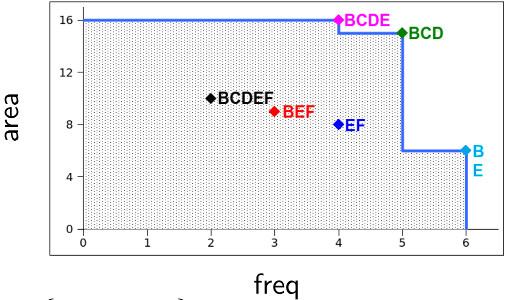
$$q(X) \equiv closed_{M'}(X) \wedge \neg(s_1 \succ_M X) \wedge \neg(s_2 \succ_M X)$$

$$Candidates = \{\underbrace{BCDEF}_{s_1}, \underbrace{BEF}_{s_2},$$

Mining skypatterns using CP (2/3)

Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F

$\mid \mathcal{L}_{\mathcal{I}} \mid = 2^6 = 64$ patterns
4 skypatterns



$$M = \{freq, area\}$$

$$q(X) \equiv closed_{M'}(X) \wedge \neg(s_1 \succ_M X) \wedge \neg(s_2 \succ_M X) \wedge \neg(s_3 \succ_M X) \wedge \neg(s_4 \succ_M X) \wedge \neg(s_5 \succ_M X) \wedge \neg(s_6 \succ_M X) \wedge \neg(s_7 \succ_M X)$$

$$Candidates = \{\underbrace{\cancel{BCDEF}}_{s_1}, \underbrace{\cancel{BEF}}_{s_2}, \underbrace{\cancel{EF}}_{s_3}, \underbrace{\underbrace{BCDE}_{s_4}, \underbrace{BCD}_{s_5}, \underbrace{B}_{s_6}, \underbrace{E}_{s_7}}_{\text{Sky}(\mathcal{L}_{\mathcal{I}}, M)}\}$$

Mining skypatterns using CP (3/3)

To sum up: mining skypatterns is achieved in a two-step approach:

1. compute the set of solutions of the query:

$$query \left\{ \begin{array}{r} q_1(X) = closed_{M'}(X) \\ q_{i+1}(X) = q_i(X) \wedge \neg(s_i \succ_M X) \text{ where } s_i\text{: solution to query } q_i(X) \end{array} \right.$$

➥ $Candidates = \{s_1, s_2, \ldots, s_n\}$

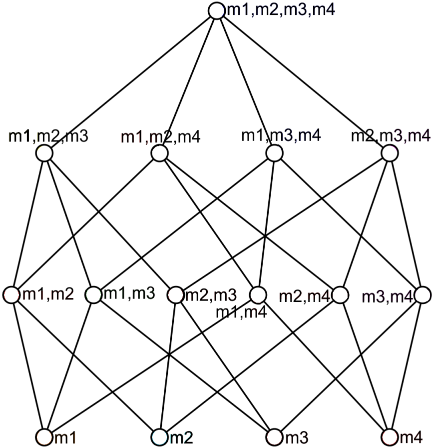
2. remove all patterns $s_i \in Candidates$ that are not skypatterns.

Experiments show that the number of candidates remains reasonably small.

From Skypatterns to Skypattern Cube

Why the skypattern cube?

- keeping all the measures is potentially useful
- what happens on a skypattern set by removing/adding measures?



$$SkypatternCube(M) = \{(M_u, \mathcal{S}ky(\mathcal{L}_{\mathcal{I}}, M_u) \mid M_u \subseteq M, M_u \neq \emptyset\}$$

Computing the skypattern cube

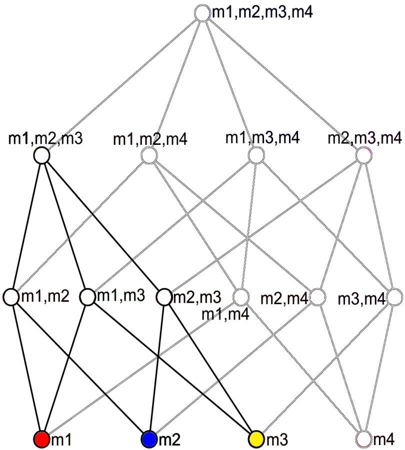
A bottom-up method in a nutshell (Ugarte et al. ECAI'14)

- mining $Sky(\mathcal{L}_{\mathcal{I}}, \{m_i\})$ for each measure $m_i \in M$
- for each parent node $M_u \subseteq M$ of the cube:
  - collect its skypatterns from the skypatterns of its child nodes (i.e., derivable skypatterns)
  - compute on the fly the non-derivable skypatterns
    ➥ use of Dynamic CSP

Bottom-up method for skypattern cube: an example

$M = \{m_1 : freq, m_2 : growth\text{-}rate, m_3 : area\}$

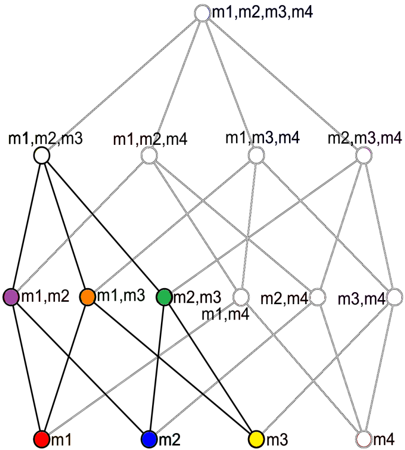
Subset of M	**Skypattern set**
$\{m_1, m_2, m_3\}$	
$\{m_1, m_2\}$	
$\{m_1, m_3\}$	
$\{m_2, m_3\}$	
$\{m_1\}$	{B, E}
$\{m_2\}$	{AEF, AF, BCDE, BCDEF, BCDF, BDE, BDEF, BDF, E, EF, F}
$\{m_3\}$	{BCDE}



Bottom-up method for skypattern cube: an example

$M = \{m_1 : freq, m_2 : growth\text{-}rate, m_3 : area\}$

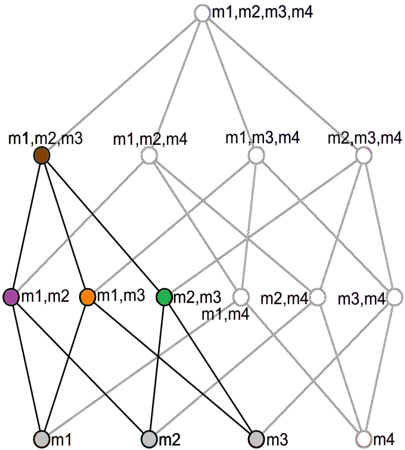
Subset of M	Skypattern set
$\{m_1, m_2, m_3\}$	
$\{m_1, m_2\}$	{E}
$\{m_1, m_3\}$	{BCD, BCDE, B, E}
$\{m_2, m_3\}$	{BCDE}
$\{m_1\}$	{B, E}
$\{m_2\}$	{AEF, AF, BCDE, BCDEF, BCDF, BDE, BDEF, BDF, E, EF, F}
$\{m_3\}$	{BCDE}



Bottom-up method for skypattern cube: an example

$M = \{m_1 : freq, m_2 : growth\text{-}rate, m_3 : area\}$

Subset of M	**Skypattern set**
$\{m_1, m_2, m_3\}$	{BCD, BCDE, E}
$\{m_1, m_2\}$	{E}
$\{m_1, m_3\}$	{BCD, BCDE, B, E}
$\{m_2, m_3\}$	{BCDE}
$\{m_1\}$	{B, E}
$\{m_2\}$	{AEF, AF, BCDE, BCDEF, BCDF, BDE, BDEF, BDF, E, EF, F}
$\{m_3\}$	{BCDE}



In practice:

- a large part of the skypatterns are collected by the derivation rules
- a sufficient condition for detecting that $Sky(\mathcal{L}_\mathcal{I}, M_u) = Derived(M_u)$

Computing the skypattern cube

An approximation-based method in a nutshell (Ugarte et al. ICTAI'14)

- **Key idea**: use a relaxation of the skypatterns (the edge-skypatterns)

  Result: the skypatterns w.r.t. any M_u are included in the set of the edge-skypatterns w.r.t. M

  The proof is based on the monotonicity of the *Edge-Sky* operator (whereas the $\mathcal{S}ky$ operator is not monotone).

- Then, the problem can be considered as computing a skyline cube in $|M|$ dimensions from the edge-skypatterns w.r.t. M

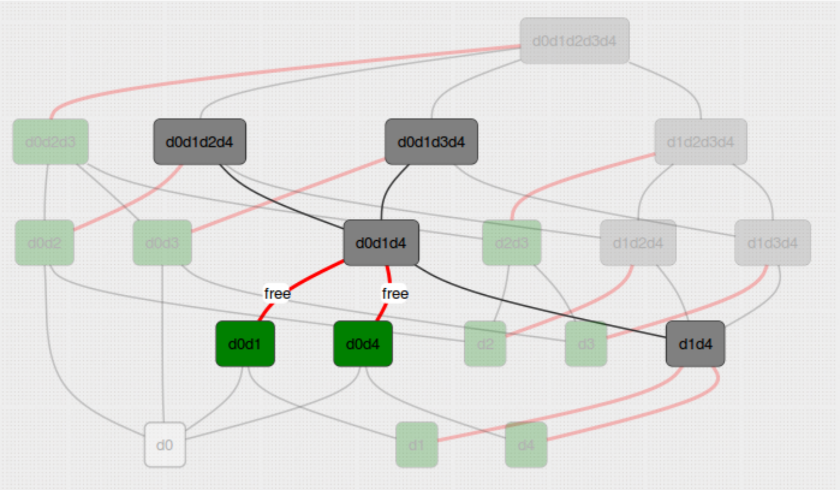
  Use of `Orion` (Raïssi et al, PVLDB 2010) to compute the closed skyline cube.

Bottom-up method versus approximation-based method

- **bottom-up method:**
  - an in-depth understanding of the different kinds of skypatterns: incomparable/indistincts ➠ Indistinct Skypattern Groups
  - elegant derivation rules
- **approximation-based method:** faster...

Skypattern cube: demo

https://sdmc.greyc.fr/skypattern/ (P. Holat)



Iris data set: d0 = freq, d1 = max(val), d2 = mean(val), d3 = area, d4 = gr 1

Concise representation of the cube:

➥ equivalence classes on measures highlight the role of measures

To sum up and perspectives

Lessons (1/2)

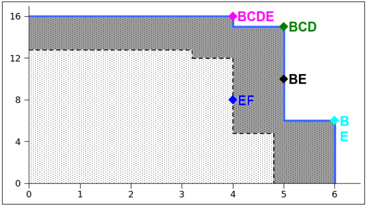
Interestingness of the CP framework

- declarative side of the CP: introducing softness is "easy":
  - changing the dominance relation ➠ soft-skypatterns
  - soft threshold constraints (Ugarte et al. DS'12)
  - top-k with soft threshold constraints (Ugarte et al. JIIS'13)
  - softness can also be useful for mining crisp patterns (cf. the cube approximation-based method)
- Dynamic CSP are a precious tool to implement:
  - pairwise comparisons (cf. the skypattern example)
  - on the fly computing (cf. the mining of non-derivable skypatterns with the cube bottom-up method)

Lessons (2/2)

Why softness? - Introducing softness is easy with CP

Stringent aspect of the classical constraint-based pattern mining framework: *what about a pattern which slightly violates the query?*



➥ introducing softness in the skypattern mining: soft-skypatterns

δ-**dominance:** a pattern X_i δ-dominates another pattern X_j w.r.t M, denoted by $X_i \succ_M^{\delta} X_j$, iff $\forall m \in M, (1 - \delta) \times m(X_i) > m(X_j)$

Same process: it is enough to update the posted constraints

Local patterns, pattern sets and more?

reminder our challenge: how to discover a manageable set of high-level and useful patterns?

➥ a general avenue: from local patterns to sets of patterns (i.e., find useful and interrelated sets of patterns)

- local patterns $(\mathcal{L}_{\mathcal{I}})$
- pattern sets (e.g., skypatterns) $(2^{\mathcal{L}_{\mathcal{I}}})$
- future?
  - interest in sets of pattern sets (e.g., skypattern cube) $(2^{2^{\mathcal{L}_{\mathcal{I}}}})$
  - vizualization methods will be helpful to go within sets of pattern sets. A novel use of the lattice structure and methods/tools such as CAMELIS (Ferré J. General Sys. 2009)?

Perspectives within NormaSTIC?

- *optimization in data mining, interactive knowledge discovery:* IVISEA project (A. Knippel and A. Pauchet)
- *sequence mining:* sports analytics: cf. Alexandre's talk
- *CP as a backbone for graph mining?*
  - chemoinformatics: L. Brun, B. Cuissart, M. Léonard, T. Lecrocq
    First approach: items to encode molecular fragments
    (cf. Willy's work)
  - graphs in bioinformatics and geomatics: cf. Géraldine's talk
  - graphs in text analysis: S. Darmoni?, A. Widlöcher?

  ➥ something to do with the graph working group?
- *sequence mining for image representation in computer vision, image clustering by combining patterns and topics* (F. Jurie?, L. Heutte?,...)
- ...

Special thanks to:

Ronan Bureau
Alban Lepailleur (CERMN)

Bertrand Cuissart
Guillaume Poezevara (GREYC)

Pierre Holat (LIPN)

Marc Plantevit (LIRIS)

Chedy Raïssi (INRIA-NGE)

Arnaud Soulet (LI)

This work is partly supported by the French Research National Agency funded projects FiCoLoFo ANR-10-BLA-0214 and Hybride ANR-11-BS002-002 and the Lower Normandy region.