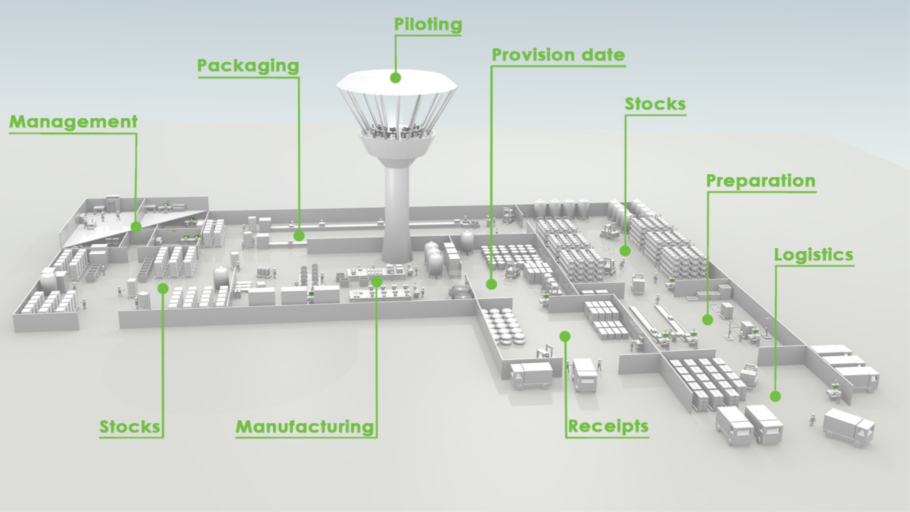
Constraint Programming for Multi-criteria Conceptual Clustering

Maxime Chabert, Christine Solnon

LIRIS, INSA Lyon, France

24 novembre 2017


Piloting
Packaging
Provision date
Stocks
Management
Preparation
Logistics
Stocks
Manufacturing
Receipts

- Simplify and reduce the cost of customization step ($\sim$ 10 000 parameters) :
  - extract relevant sets of parameter settings $\Rightarrow$ functional needs
  - reuse these sets for new client customization

Client	Price reference date	Order blocking	Order split	Stock control
Client 1	Delivery	Yes	No	Blocking
Client 2	Delivery	No	No	Alert
Client 3	Order	Yes	No	Without
Client 4	Order	Yes	Yes	Alert

- Simplify and reduce the cost of customization step ($\sim$ 10 000 parameters) :
  - extract relevant sets of parameter settings $\Rightarrow$ functional needs
  - reuse these sets for new client customization

Client	Price reference date	Order blocking	Order split	Stock control
Client 1	Delivery	Yes	No	Blocking
Client 2	Delivery	No	No	Alert
Client 3	Order	Yes	No	Without
Client 4	Order	Yes	Yes	Alert

Input

- A set of transactions $\mathcal{T}$, a set of items $\mathcal{I}$
- A relation $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$: $(t, i) \in \mathcal{R}$ iff t contains i.

Client	Price reference date	Order blocking	Order split	Stock control
Client 1	Delivery	Yes	No	Blocking
Client 2	Delivery	No	No	Alert
Client 3	Order	Yes	No	Without
Client 4	Order	Yes	Yes	Alert

Input

- A set of transactions $\mathcal{T}$, a set of items $\mathcal{I}$
- A relation $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$: $(t, i) \in \mathcal{R}$ iff t contains i.

Client	Price reference date	Order blocking	Order split	Stock control
Client 1	Delivery	Yes	No	Blocking
Client 2	Delivery	No	No	Alert
Client 3	Order	Yes	No	Without
Client 4	Order	Yes	Yes	Alert

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
t_1	1	0	1	0	0	1	1	0	0
t_2	1	0	0	1	0	1	0	1	0
t_3	0	1	1	0	0	1	0	0	1
t_4	0	1	1	0	1	0	0	1	0

Input

- A set of transactions $\mathcal{T}$, a set of items $\mathcal{I}$
- A relation $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$: $(t, i) \in \mathcal{R}$ iff t contains i.

Definitions

- Intent of $T \subseteq \mathcal{T} = \cap_{t \in T} itemset(t)$

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
t_1	1	0	1	0	0	1	1	0	0
t_2	1	0	0	1	0	1	0	1	0
t_3	0	1	1	0	0	1	0	0	1
t_4	0	1	1	0	1	0	0	1	0

$Intent(t_3, t_4) = \{i_2, i_3\}$

Input

- A set of transactions $\mathcal{T}$, a set of items $\mathcal{I}$
- A relation $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$: $(t, i) \in \mathcal{R}$ iff t contains i.

Definitions

- Intent of $T \subseteq \mathcal{T} = \cap_{t \in T} itemset(t)$
- Extent of $I \subseteq \mathcal{I} = \{t \in \mathcal{T} : I \subseteq itemset(t)\}$

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
t_1	1	0	1	0	0	1	1	0	0
t_2	1	0	0	1	0	1	0	1	0
t_3	0	1	1	0	0	1	0	0	1
t_4	0	1	1	0	1	0	0	1	0

$Extent(i_1, i_6) = \{t_1, t_2\}$

Input

- A set of transactions $\mathcal{T}$, a set of items $\mathcal{I}$
- A relation $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$: $(t, i) \in \mathcal{R}$ iff t contains i.

Definitions

- Intent of $T \subseteq \mathcal{T} = \cap_{t \in T} itemset(t)$
- Extent of $I \subseteq \mathcal{I} = \{t \in \mathcal{T} : I \subseteq itemset(t)\}$
- $FormalConcept = (T, I) \in \mathcal{T} \times \mathcal{I} s.t. T = Extent(I), I = Intent(T)$
- $freq(T, I) = \#T$, $size(T, I) = \#I$

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
t_1	1	0	1	0	0	1	1	0	0
t_2	1	0	0	1	0	1	0	1	0
t_3	0	1	1	0	0	1	0	0	1
t_4	0	1	1	0	1	0	0	1	0

$c = (\{t_1, t_2\}, \{i_1, i_6\})$

- $Extent(\{i_1, i_6\}) = \{t_1, t_2\}$
- $Intent(\{t_1, t_2\}) = \{i_1, i_6\}$
- $freq(c) = 2$, $size(c) = 2$

Input

- A set of transactions $\mathcal{T}$, a set of items $\mathcal{I}$
- A relation $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$: $(t, i) \in \mathcal{R}$ iff t contains i.

Definitions

- Intent of $T \subseteq \mathcal{T} = \cap_{t \in T} itemset(t)$
- Extent of $I \subseteq \mathcal{I} = \{t \in \mathcal{T} : I \subseteq itemset(t)\}$
- $FormalConcept = (T, I) \in \mathcal{T} \times \mathcal{I} s.t. T = Extent(I), I = Intent(T)$
- $freq(T, I) = \#T$, $size(T, I) = \#I$

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
t_1	1	0	1	0	0	1	1	0	0
t_2	1	0	0	1	0	1	0	1	0
t_3	0	1	1	0	0	1	0	0	1
t_4	0	1	1	0	1	0	0	1	0

$c = (\{t_4\}, \{i_2, i_3, i_5, i_8\})$

- $Extent(\{i_2, i_3, i_5, i_8\}) = \{t_4\}$
- $Intent(\{t_4\}) = \{i_2, i_3, i_5, i_8\}$
- $freq(c) = 1$, $size(c) = 4$

- Clustering : Partition of $\mathcal{T}$
- Conceptual clustering : Each cluster is a formal concept

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
t_1	1	0	1	0	0	1	1	0	0
t_2	1	0	0	1	0	1	0	1	0
t_3	0	1	1	0	0	1	0	0	1
t_4	0	1	1	0	1	0	0	1	0

Clustering 1 = $\{c_1, c_2\}$

- $c_1 = (\{t_1, t_2\}, \{i_1, i_6\})$
- $c_2 = (\{t_3, t_4\}, \{i_2, i_3\})$

- $minFreq(C) = \min_{c \in C} freq(c)$
- $minSize(C) = \min_{c \in C} size(c)$

- Clustering : Partition of $\mathcal{T}$
- Conceptual clustering : Each cluster is a formal concept

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
t_1	1	0	1	0	0	1	1	0	0
t_2	1	0	0	1	0	1	0	1	0
t_3	0	1	1	0	0	1	0	0	1
t_4	0	1	1	0	1	0	0	1	0

Clustering 2 = $\{c_3, c_4\}$

- $c_3 = (\{t_1, t_2, t_3\}, \{i_6\})$
- $c_4 = (\{t_4\}, \{i_2, i_3, i_5, i_8\})$

- $minFreq(C) = \min_{c \in C} freq(c)$
- $minSize(C) = \min_{c \in C} size(c)$

Declarative approaches for conceptual clustering

Constraint programming

- Binary model [Guns et al 2011] : binary variables = transaction/extents and item/intents couples
- Set model [Dao et al 2015] : set variables = extents and intents

Declarative approaches for conceptual clustering

Constraint programming

- Binary model [Guns et al 2011] : binary variables = transaction/extents and item/intents couples
- Set model [Dao et al 2015] : set variables = extents and intents

Hybrid approach [Ouali et al (2016)]

- Step 1 : Use LCM [Uno et al 2004] to extract the set $\mathcal{F}$ of all formal concepts in $\mathcal{O}(\mathcal{F})$
- Step 2 : Use ILP to select the best subset of $\mathcal{F}$ that is a partition of $\mathcal{T}$

New CP models

- Full CP model that improves the model of [Dao et al 2015] :
  - The number of clusters k is not fixed
  - Partial relaxation of the intent constraint
- Hybrid model
  - Step 1 : Use LCM [Uno et al 2004] to extract the set $\mathcal{F}$ of all formal concepts
  - Step 2 : Use CP to select the best subset of $\mathcal{F}$ that is a partition of $\mathcal{T}$

Multi-criteria conceptual clustering

- Computation of the Pareto set of non-dominated solutions

Application case experiments

- Experimentation on parameter settings of a module of the ERP Copilote

Variables

- For each t_i : G_i = cluster of t_i ($D(G_i) = \{c_1, \ldots, c_k\}$)
- For each t_i : $extent_{t_i}$ = extent of the cluster of t_i ($D(extent_{t_i}) = \mathcal{P}(\mathcal{T})$)
- For each t_i : $intent_{t_i}$ = intent of the cluster of t_i ($D(intent_{t_i}) = \mathcal{P}(\mathcal{I})$)
- Redundant variables : For each c_j : $extentCluster_{c_j}$ = extent of C_j

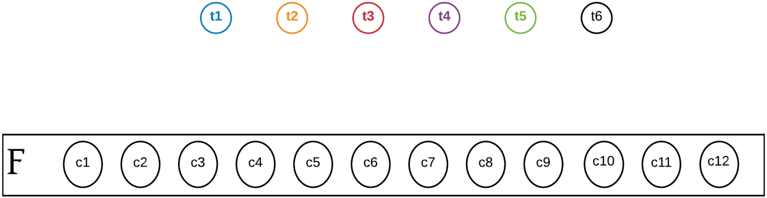
Constraints

- $\forall t \in \mathcal{T}, extent[t] = extentCluster[G_t]$
- $\forall t \in \mathcal{T}, \forall c \in [1, k_{max}], t \in extentCluster_c \Leftrightarrow G_t = c$
- $(G_{t_1} = G_{t_2}) \Leftrightarrow (Intent_{t_1} = Intent_{t_2}) \Leftrightarrow (Intent_{t_1} \subseteq itemSet(t_2))$

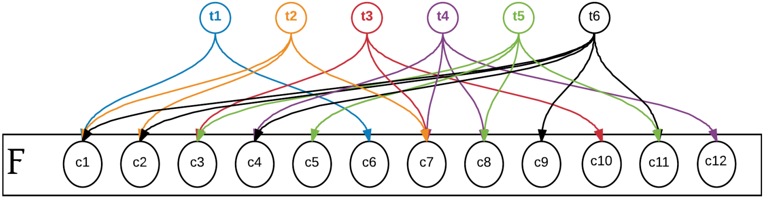
Different possible criteria to optimize

- Maximize the minimal frequency $\rightsquigarrow$ $minFreq = \min_{t \in \mathcal{T}} \#Extent_t$
- Maximize the minimal size $\rightsquigarrow$ $minSize = \min_{t \in \mathcal{T}} \#Intent_t$

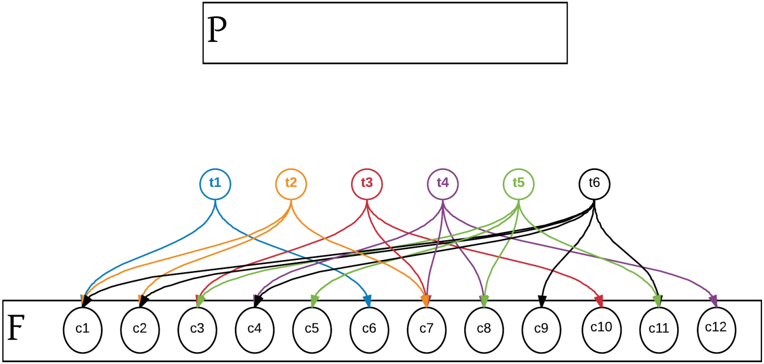

t1
t2
t3
t4
t5
t6



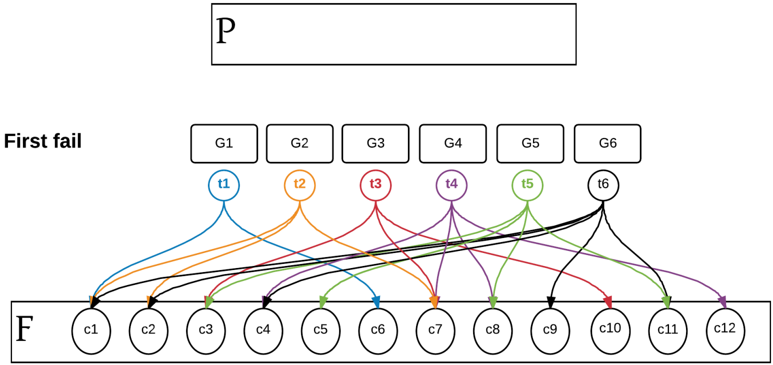
Extraction of all formal concepts of $\mathcal{T}$ with LCM [Uno et al 2004] (linear with respect to $\#\mathcal{F}$)



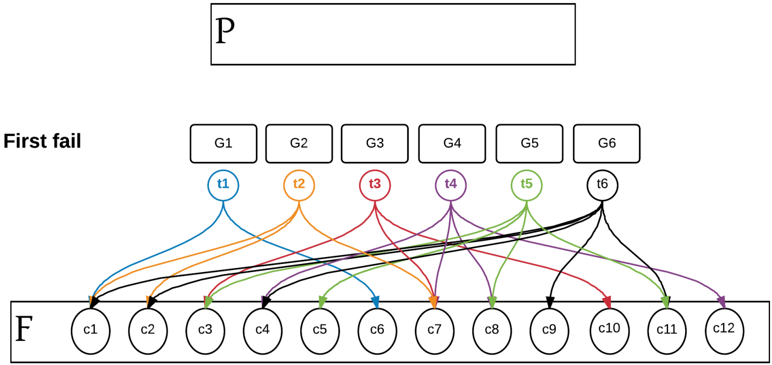
Extraction of all formal concepts of $\mathcal{T}$ with LCM [Uno et al 2004] (linear with respect to $\#\mathcal{F}$)



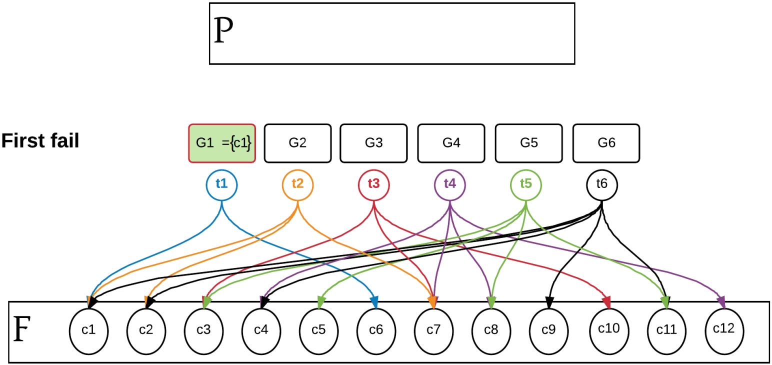
- $D(P) = \mathcal{P}(\mathcal{F})$



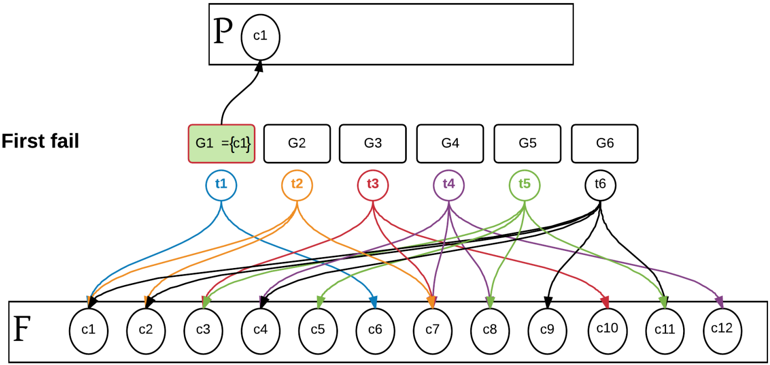
- $D(P) = \mathcal{P}(\mathcal{F})$
- For each t_i : G_i = cluster of t_i ($D(G_i) = \mathcal{P}(\{f \in \mathcal{F} : t \in extent(f)\})$)



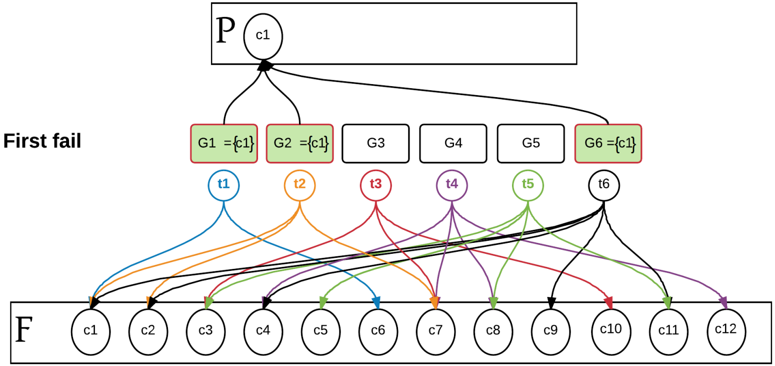
- $D(P) = \mathcal{P}(\mathcal{F})$
- For each t_i : G_i = cluster of t_i ($D(G_i) = \mathcal{P}(\{f \in \mathcal{F} : t \in extent(f)\})$)
- Integer variable k ($D(k) = [k_{min}, k_{max}]$)


P
First fail
G1 ={c1}
G2
G3
G4
G5
G6
t1
t2
t3
t4
t5
t6
F
c1
c2
c3
c4
c5
c6
c7
c8
c9
c10
c11
c12

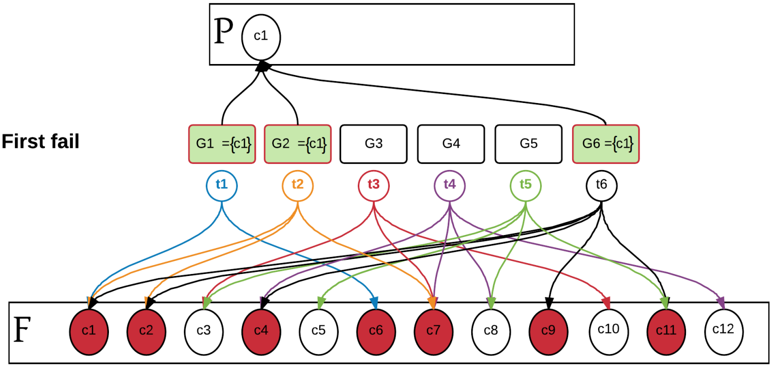
- $k = \#P$



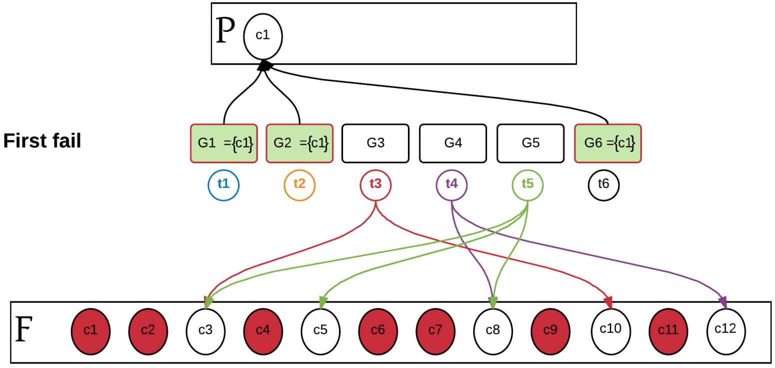
- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$



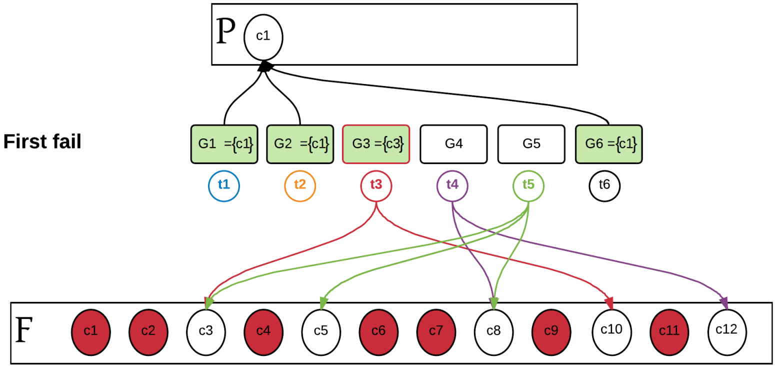
- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$



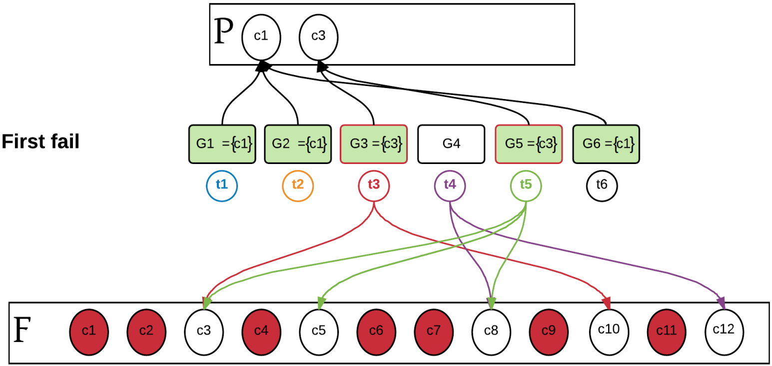
- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$



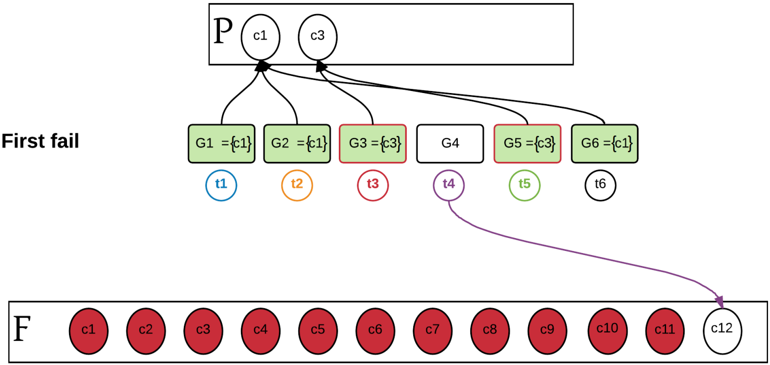
- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$



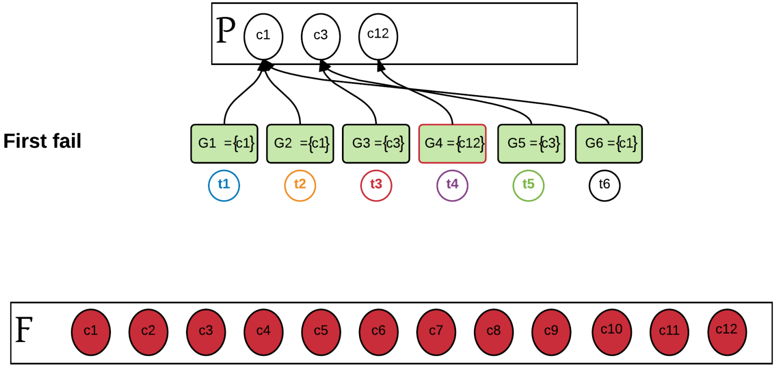
- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$



- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$

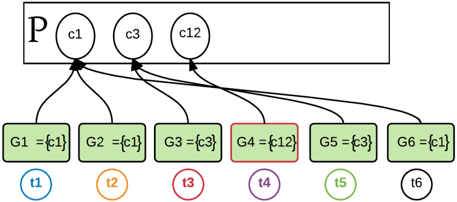


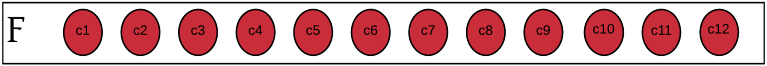
- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$



- $k = \#P$
- $\forall t \in \mathcal{T}, \{f \in \mathcal{F} : t \in extent(f)\} \cap P = G_t$
- $\forall t \in \mathcal{T}, \#(G_t) = 1$

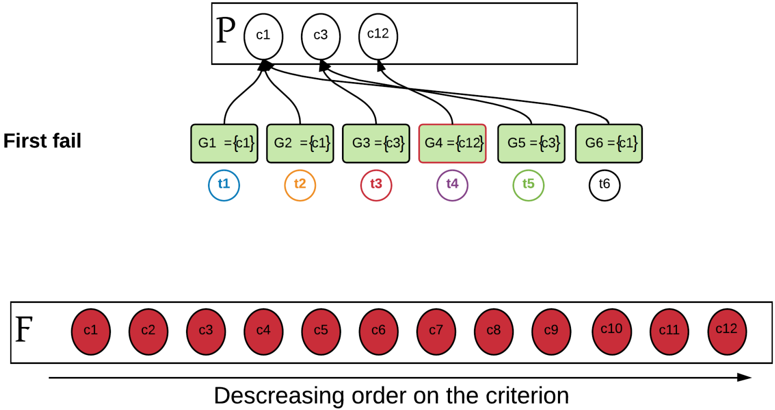
First fail





Different possible criteria to optimize :

- $minSize = \min_{c \in \mathcal{P}} \#intent(c)$
- $minFreq = \min_{c \in \mathcal{P}} \#extent(c)$



Heuristic : sort formal concepts by decreasing order on the criterion

- Classic UCI instances :

Instance	# $\mathcal{T}$	# $\mathcal{I}$	# $\mathcal{F}$	LCM Time
zoo	59	36	4 567	0.01
vote	341	48	227 031	0.54
tic-tac-toe	958	27	42 711	0.05
soybean	303	116	817 534	6.7

- Generated instances from current parameter settings of Copilote :

Instance	# $\mathcal{T}$	# $\mathcal{I}$	# $\mathcal{F}$	LCM Time
ERP1	50	27	1 580	0.01
ERP2	47	47	8 337	0.03
ERP3	75	36	10 835	0.03
ERP4	84	42	14 305	0.05
ERP5	94	53	63 633	0.28
ERP6	95	61	71 918	0.45
ERP7	160	66	728 537	5.31

- When the number of clusters k is fixed to 2

	Objective = *minFreq*				Objective = *minSize*			
	ILP	FullCP1	FullCP2	HybridCP	ILP	FullCP1	FullCP2	HybridCP
ERP1	0.2	0.0	0.2	0.2	0.2	0.0	0.3	0.2
ERP2	1.5	0.0	0.1	4.4	1.7	0.0	0.1	4.6
ERP3	1.5	0.0	0.2	9.2	1.6	0.0	0.3	9.6
ERP4	7.5	0.0	0.3	1.4	7.5	0.0	0.5	22.0
ERP5	12.5	0.0	0.5	172.2	13.1	0.0	0.6	-
ERP6	52.6	0.0	0.5	8.6	63.4	0.0	0.8	645.0
ERP7	-	0.0	2.8	-	-	0.0	4.4	-
zoo	1.0	0.0	0.2	0.5	1.1	0.0	0.2	0.7
vote	40.6	0.0	1.6	17.8	40.8	0.0	3.3	16.2
tic-tac-toe	61.3	0.2	32.5	10.9	60.7	0.4	33.2	10.9
soybean	-	0.1	1.4	63.7	-	0.0	2.5	93.4

ILP = LCM + CPLEX implementation of the hybrid model of [Ouali et al 2016]
FullCP1 = Gecode implementation of CP model of [Dao et al 2015]
FullCP2 = Choco 4 implementation of our new CP model
HybridCP = LCM + Choco 4 implementation of our new hybrid model

- When the number of clusters *k* is fixed to 3

	Objective = *minFreq*				Objective = *minSize*			
	ILP	FullCP1	FullCP2	HybridCP	ILP	FullCP1	FullCP2	HybridCP
ERP1	0.9	0.0	0.7	0.9	0.3	0.1	0.7	0.4
ERP2	2.7	0.4	0.2	1.5	1.6	0.5	0.2	17.7
ERP3	2.5	0.3	1.5	24.7	1.6	0.6	7.0	61.6
ERP4	15.0	0.3	2.8	100.6	8.3	0.8	4.6	103.4
ERP5	18.3	1.4	5.0	634.4	21.1	2.2	6.1	-
ERP6	145.8	10.3	2.7	-	93.3	14.2	7.5	-
ERP7	-	82.9	26.8	-	-	191.1	69.2	-
zoo	2.2	0.0	0.2	0.6	0.9	0.0	1.7	6.8
vote	-	2.0	19.2	150.0	243.5	3.9	12.2	69.0
tic-tac-toe	80.6	0.3	75.9	25.2	80.4	0.3	54.1	25.9
soybean	-	160.1	7.9	980.2	-	145.7	7.1	460.6

ILP = LCM + CPLEX implementation of the hybrid model of [Ouali et al 2016]
HybridCP = LCM + Choco 4 implementation of our new hybrid model
FullCP1 = Gecode implementation of CP model of [Dao et al 2015]
FullCP2 = Choco 4 implementation of our new CP model

- When the number of clusters *k* is fixed to 4

	Objective = *minFreq*				Objective = *minSize*			
	ILP	FullCP1	FullCP2	HybridCP	ILP	FullCP1	FullCP2	HybridCP
ERP1	1.0	0.0	4.3	1.4	0.3	1.0	2.5	1.6
ERP2	2.3	4.8	1.6	4.6	1.6	19.9	0.8	7.2
ERP3	3.2	20.0	1.6	2.4	1.7	252.9	7.0	61.6
ERP4	20.9	36.6	37.9	153.1	7.2	184.8	34.4	329.5
ERP5	83.7	773.6	91.9	-	40.6		58.4	-
ERP6	339.6	302.7	101.1	-	648.3	9.6	54.2	-
ERP7	-	-	742.9	-	-	-	682.5	-
zoo	3.0	0.8	4.5	1.0	1.2	1.4	7.7	9.4
vote	-	292.6	370.5	95.6	249.7	969.4	191.8	-
tic-tac-toe	-	106.0	-	-	-	105.6	-	-
soybean	-	-	166.0	-	-	-	93.3	-

ILP = LCM + CPLEX implementation of the hybrid model of [Ouali et al 2016]
HybridCP = LCM + Choco 4 implementation of our new hybrid model
FullCP1 = Gecode implementation of CP model of [Dao et al 2015]
FullCP2 = Choco 4 implementation of our new CP model

- When the number of clusters k is not fixed : $2 \leq k <$ number of transactions

	Objective = *minFreq*					Objective = *minSize*				
	k	ILP	FullCP1	FullCP2	HybridCP	k	ILP	FullCP1	FullCP2	HybridCP
ERP1	2	0.8	0.2	0.3	0.3	49	0.4	0.2	0.2	0.1
ERP2	2	1.0	0.5	0.1	0.3	42	0.8	-	0.0	0.1
ERP3	2	1.7	2.4	0.3	0.6	59	1.2	-	0.1	0.2
ERP4	2	13.6	1.2	0.4	0.8	83	18.3	2.1	0.5	0.3
ERP5	2	18.3	125.3	1.5	10.6	79	12.5	-	0.3	1.5
ERP6	2	143.3	51.7	1.1	8.0	94	-	7.2	0.5	1.9
ERP7	2	-	973.4	5.0	-	159	-	47.2	2.3	39.5
zoo	2	1.5	0.1	0.3	0.2	58	2.0	0.5	0.1	0.1
vote	2	55.2	-	33.1	20.8	317	-	-	20.4	17.2
tic-tac-toe	3	718.6	-	179.7	33.3	957	254.5	-	-	18.7
soybean	-	-	-	-	-	302	-	-	22.2	342.4

ILP = LCM + CPLEX implementation of the hybrid model of [Ouali et al 2016]
FullCP1 = Gecode implementation of CP model of [Dao et al 2015]
HybridCP = LCM + Choco 4 implementation of our new hybrid model
FullCP2 = Choco 4 implementation of our new CP model

- When the number of clusters k is not fixed : $2 \leq k <$ number of transactions

	Objective = *minFreq*					Objective = *minSize*				
	k	ILP	FullCP1	FullCP2	HybridCP	k	ILP	FullCP1	FullCP2	HybridCP
ERP1	2	0.8	0.2	0.3	0.3	49	0.4	0.2	0.2	0.1
ERP2	2	1.0	0.5	0.1	0.3	42	0.8	-	0.0	0.1
ERP3	2	1.7	2.4	0.3	0.6	59	1.2	-	0.1	0.2
ERP4	2	13.6	1.2	0.4	0.8	83	18.3	2.1	0.5	0.3
ERP5	2	18.3	125.3	1.5	10.6	79	12.5	-	0.3	1.5
ERP6	2	143.3	51.7	1.1	8.0	94	-	7.2	0.5	1.9
ERP7	2	-	973.4	5.0	-	159	-	47.2	2.3	39.5
zoo	2	1.5	0.1	0.3	0.2	58	2.0	0.5	0.1	0.1
vote	2	55.2	-	33.1	20.8	317	-	-	20.4	17.2
tic-tac-toe	3	718.6	-	179.7	33.3	957	254.5	-	-	18.7
soybean	-	-	-	-	-	302	-	-	22.2	342.4

Result interpretation

- Clusterings with high frequency : generality (low size)
- Clusterings with high size : specificity (low frequency, many clusters)

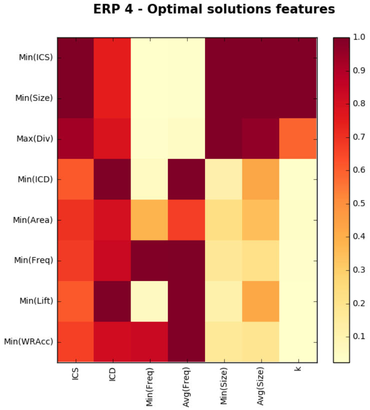
New classic criteria

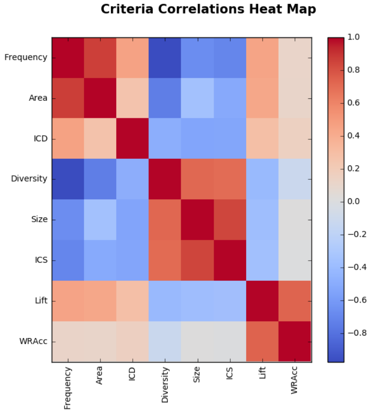
Let C be a cluster such as $C = (T, I) \in \mathcal{P}(\mathcal{T}) \times \mathcal{P}(\mathcal{I})$

- Area : $area(C) = |I| \times |T|$
- Diversity : $div(C) = \sum_{t \in T} (i \in \mathcal{I}, |(i \notin I) \wedge (i \in t)|)$ (to minimize)
- ICS : $ICS(C) = \frac{2}{|T||T-1|} \sum_{t,t' \in T} (s(t,t'))$ with $\forall t, t' \in \mathcal{T}, s(t,t') = \frac{|t \cap t'|}{|t \cup t'|}$
- ICD : $ICD(C) = \frac{1}{(|\mathcal{T}|-|T|)\times|T|} \sum_{t \in T, t' \in \mathcal{T} \wedge t' \notin T} (1 - s(t,t'))$

New advanced criteria

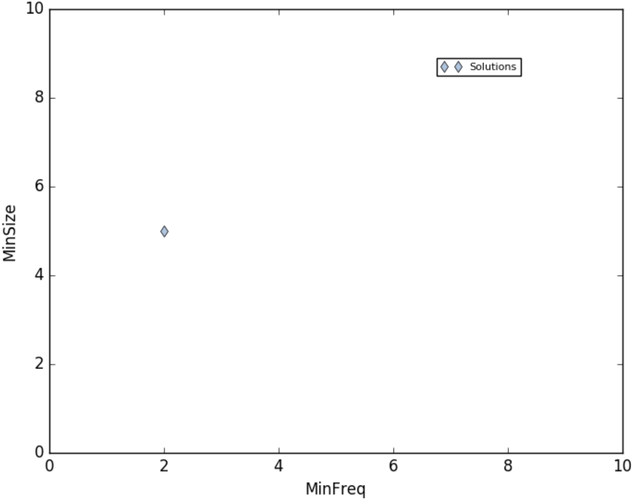
- Lift : Interest of the rule
  - Lift : $lift(X \rightarrow Y) = \frac{P(Y|X)}{P(Y)}$
  - $lift(C) = \min_{i \in I}(lift(I \setminus \{i\} \rightarrow i))$
- WRAcc (weighted related accuracy) : Weighted gain of the rule
  - $WRAcc(X \rightarrow Y) = P(X) \times (P(Y|X) - P(Y))$
  - $WRAcc(C) = \min_{i \in I}(WRAcc(I \setminus \{i\} \rightarrow i))$



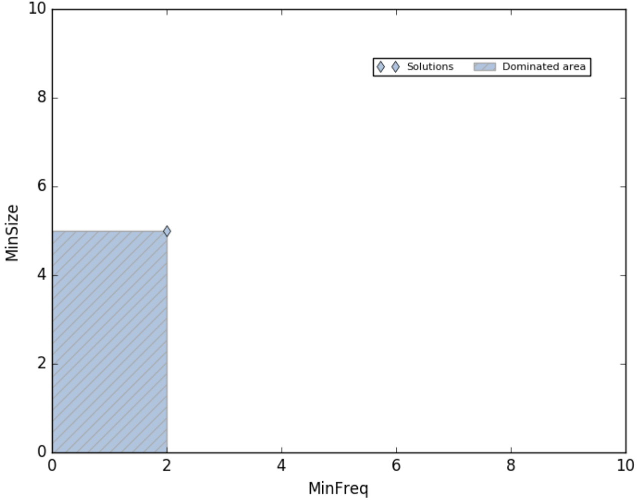


- ICD, Area and frequency : low number of clusters, high frequency
- ICS, size, diversity : high number of clusters, high size, low frequency
- WRAcc, lift : low number of clusters, high frequency

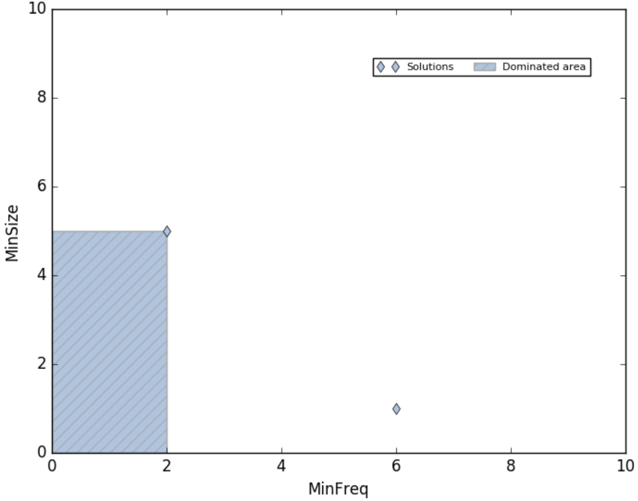
- Extract concepts with higher diversity of size and frequency and potentially higher added-value
- Use of [Gavanelli et al 2002] method that posts dynamically a new constraint for each solution found to find only non-dominated solutions



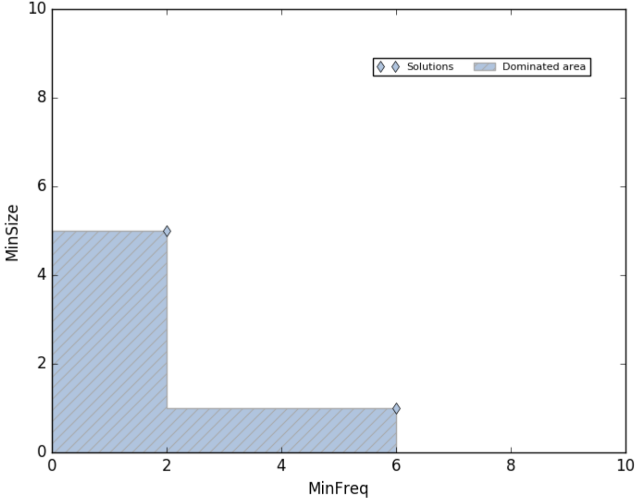
- First solution found



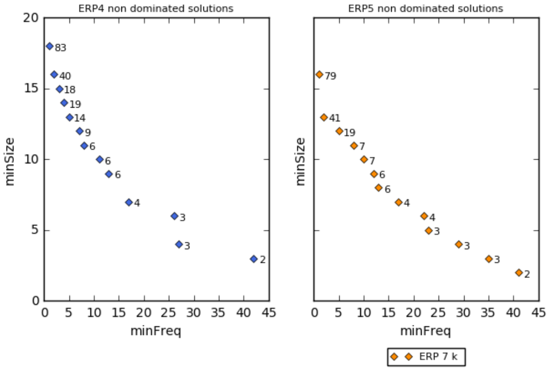
- Post of the constraint : $2 < minFreq \vee 5 < minSize$



- Next solution found



- Post of the constraint : $1 < \mathit{minFreq} \vee 6 < \mathit{minSize}$



- Adaptation of the approach : decomposition in 2 sub-problems
- 7 instances resolved in less than 2 hours
- Lack of relevancy : each point of the Pareto front correspond to hundreds of solutions

Application case

- Production planning models : planning period, data taken into account : storage, quantities, resources features etc.
- 1800 transactions and 25 items (10 parameters)

Experiments

- Experiments with classic criteria (size, freq, div, ICS, ICD, area) :
  - Lack of relevancy to the ERP expert
  - Too many clusters and redundancy (for size, ICS)
- Experiments with advanced criteria (lift and WRAcc) :
  - Extracted concepts are relevant but too short

Soft clustering

- Relax the cover constraint : at least δ transactions must be covered with $\delta < |T|$
- Better concept relevancy by ignoring some transactions
- Add the constraints :
  - $\forall t \in \mathcal{T}, \#(G_t) \leq 1$
  - Integer variable *emptyCluster* with $D(emptyCluster) = \{0, \ldots, \delta\}$
  - $nbEmpty(G, emptyCluster)$

Experiments

- Scales up on ERP instances (faster than regular clustering)
- Better solutions on WRAcc and lift criteria in our application case

Bi-clustering

- Relax the overlap constraint :
  - a transaction can belong to at most δ clusters with $\delta < |T|$
- A transaction can implement several parameter setting concepts
- Add the constraint :
  - $\forall t \in \mathcal{T}, 1 \leq \#(G_t) \leq \delta$

Experiments

- Hard to scale up

Pivot items (Expert knowledge)

- Items that have to appear in each concept
- Corresponds to parameters that :
  - can be resolved easily with a question
  - are important functionally to set up the software
- Filter the formal concepts : keep only the formal concepts that contain at least one of these items

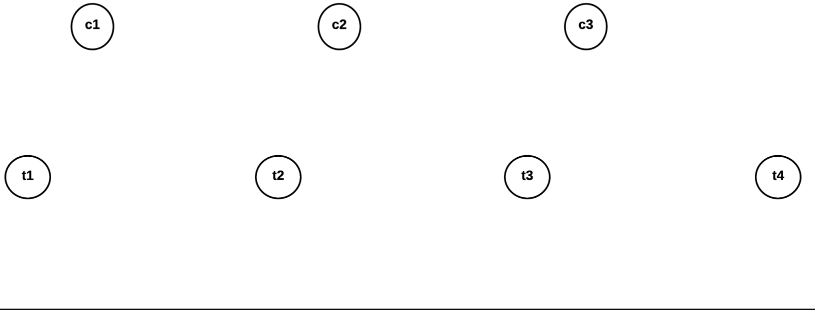
Experiments

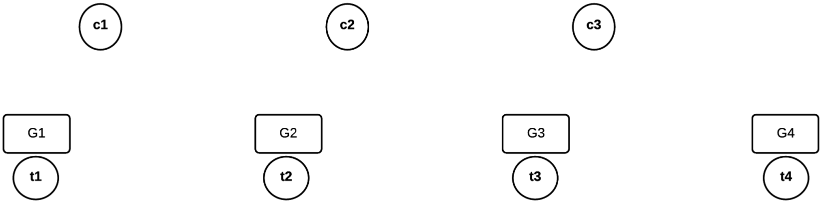
- Relevant combination with the frequency criteria (but need to do hierarchical clustering)

Further works

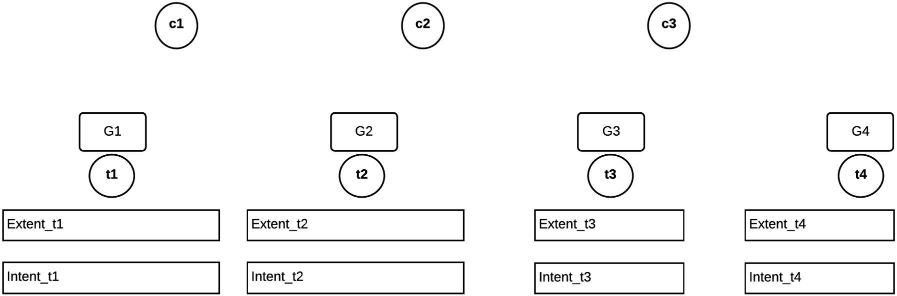
- Experiment hierarchical clustering
- Improve the model for bi-clustering
- Assess scale up properties of ILP for multi-criteria optimization and maybe combine it with CP

Thank you for your attention

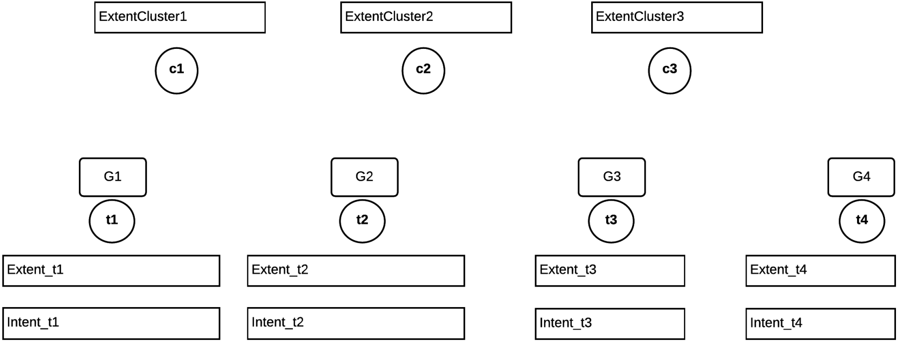

c1
c2
c3
t1
t2
t3
t4



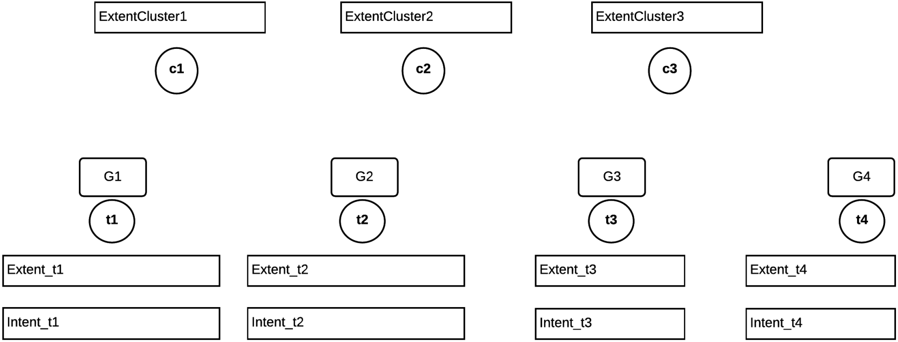
- For each t_i : G_i = cluster of t_i ($D(G_i) = \{c_1, \ldots, c_k\}$)



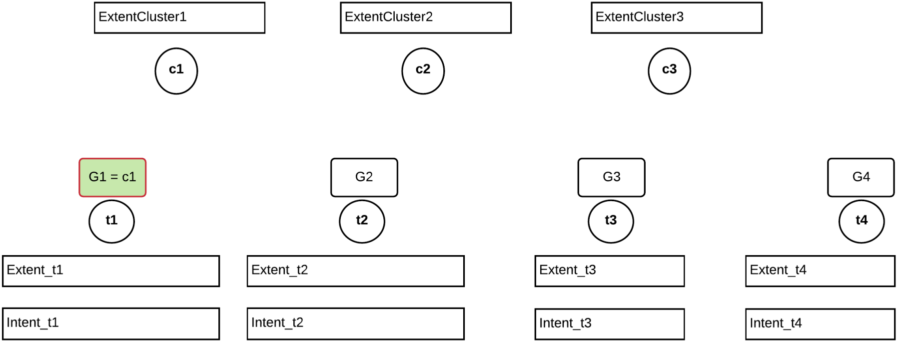
- For each t_i : G_i = cluster of t_i ($D(G_i) = \{c_1, \dots, c_k\}$)
- For each t_i : $extent_{t_i}$ = extent of the cluster of t_i ($D(extent_{t_i}) = \mathcal{P}(\mathcal{T})$)
- For each t_i : $intent_{t_i}$ = intent of the cluster of t_i ($D(intent_{t_i}) = \mathcal{P}(\mathcal{I})$)



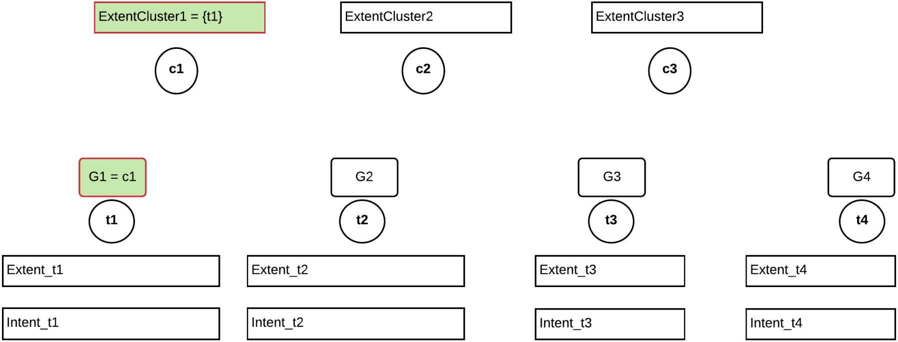
- For each t_i : G_i = cluster of t_i ($D(G_i) = \{c_1, \ldots, c_k\}$)
- For each t_i : $extent_{t_i}$ = extent of the cluster of t_i ($D(extent_{t_i}) = \mathcal{P}(\mathcal{T})$)
- For each t_i : $intent_{t_i}$ = intent of the cluster of t_i ($D(intent_{t_i}) = \mathcal{P}(\mathcal{I})$)
- Redundant variables : For each c_j : $extentCluster_{G_i} = extent_{t_i}$



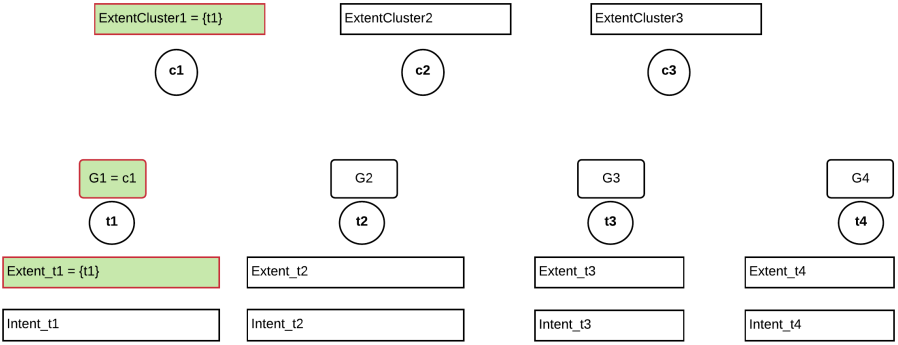
- For each t_i : G_i = cluster of t_i ($D(G_i) = \{c_1, \dots, c_k\}$)
- For each t_i : $extent_{t_i}$ = extent of the cluster of t_i ($D(extent_{t_i}) = \mathcal{P}(\mathcal{T})$)
- For each t_i : $intent_{t_i}$ = intent of the cluster of t_i ($D(intent_{t_i}) = \mathcal{P}(\mathcal{I})$)
- Redundant variables : For each c_j : $extentCluster_{G_i} = extent_{t_i}$
- Integer variable k ($D(k) = [k_{min}, k_{max}]$)



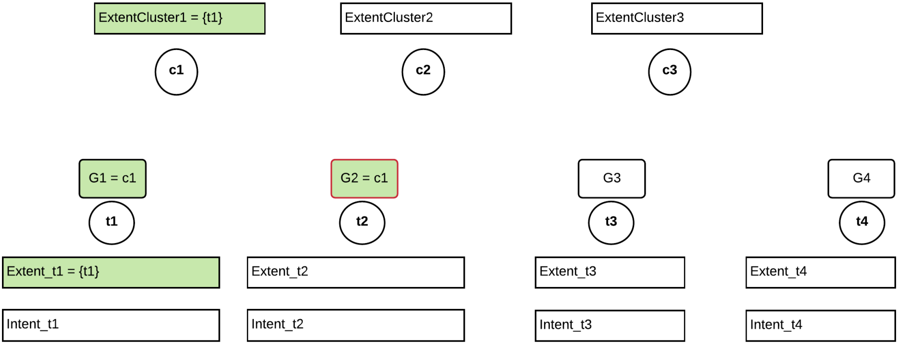
- $k = k_{max} - nbEmpty(ExtentCluster)$



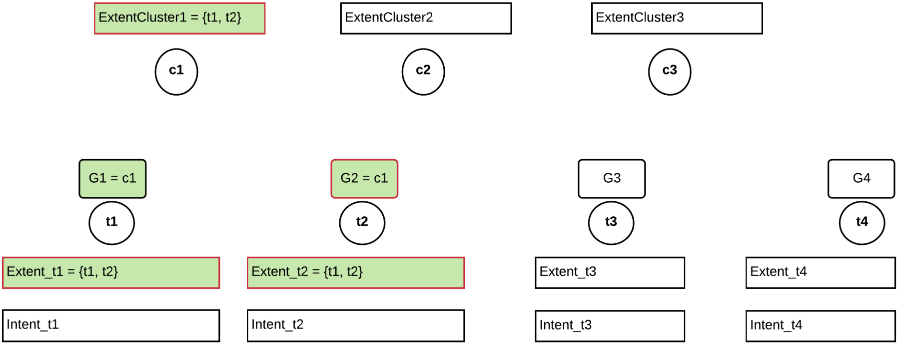
- $k = k_{max} - nbEmpty(ExtentCluster)$
- $\forall t \in \mathcal{T}, \forall c \in [1, k_{max}], t \in ExtentCluster_c \Leftrightarrow G_t = c$



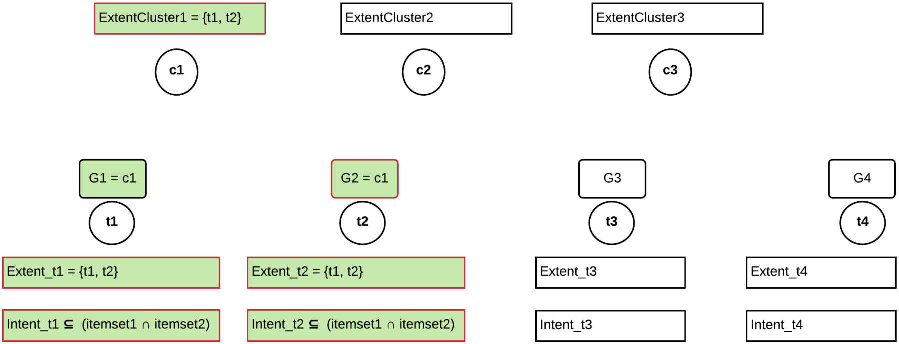
- $k = k_{max} - nbEmpty(ExtentCluster)$
- $\forall t \in \mathcal{T}, \forall c \in [1, k_{max}], t \in ExtentCluster_c \Leftrightarrow G_t = c$
- $\forall t \in \mathcal{T}, Extent[t] = ExtentCluster[G_t]$



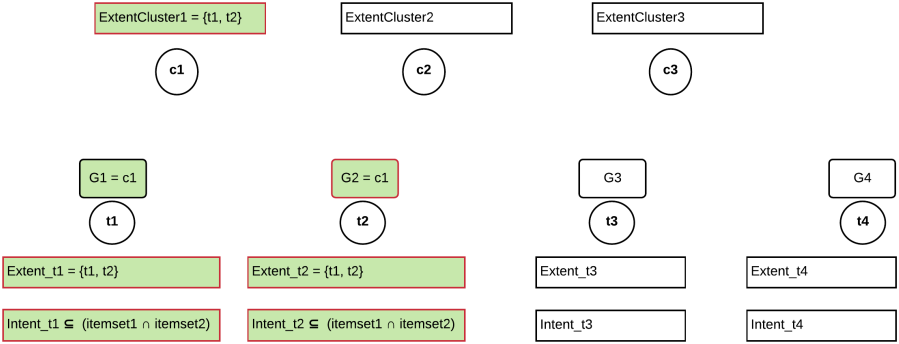
- $k = k_{max} - nbEmpty(ExtentCluster)$
- $\forall t \in \mathcal{T}, \forall c \in [1, k_{max}], t \in ExtentCluster_c \Leftrightarrow G_t = c$
- $\forall t \in \mathcal{T}, Extent[t] = ExtentCluster[G_t]$



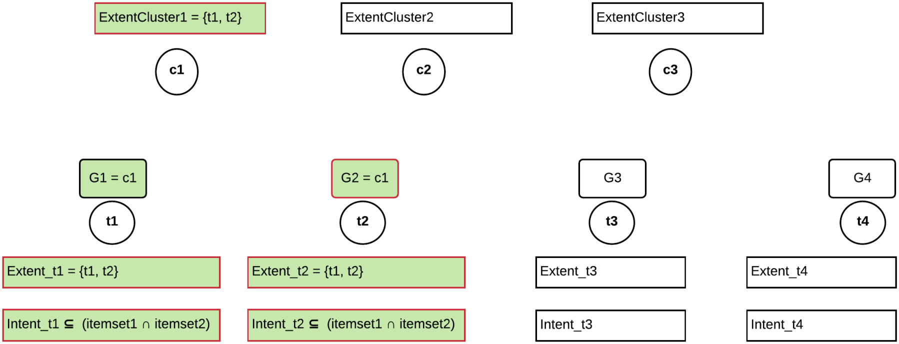
- $k = k_{max} - nbEmpty(ExtentCluster)$
- $\forall t \in \mathcal{T}, \forall c \in [1, k_{max}], t \in ExtentCluster_c \Leftrightarrow G_t = c$
- $\forall t \in \mathcal{T}, Extent[t] = ExtentCluster[G_t]$



- $k = k_{max} - nbEmpty(ExtentCluster)$
- $\forall t \in \mathcal{T}, \forall c \in [1, k_{max}], t \in ExtentCluster_c \Leftrightarrow G_t = c$
- $\forall t \in \mathcal{T}, Extent[t] = ExtentCluster[G_t]$
- $(G_{t_1} = G_{t_2}) \Leftrightarrow (Intent_{t_1} = Intent_{t_2}) \Leftrightarrow (Intent_{t_1} \subseteq itemSet(t_2))$

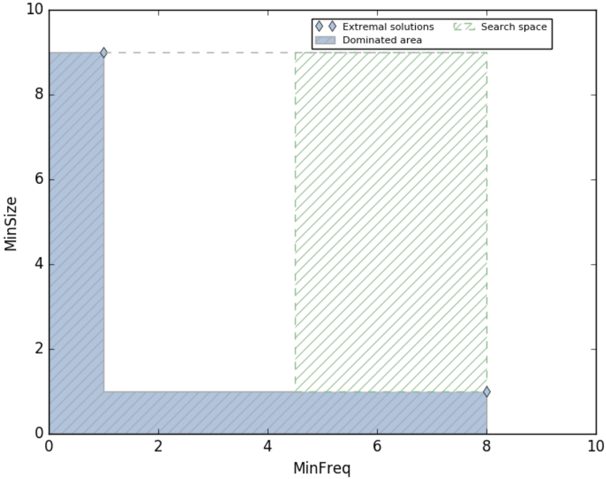


- $k = k_{max} - nbEmpty(ExtentCluster)$
- $\forall t \in \mathcal{T}, \forall c \in [1, k_{max}], t \in ExtentCluster_c \Leftrightarrow G_t = c$
- $\forall t \in \mathcal{T}, Extent[t] = ExtentCluster[G_t]$
- $(G_{t_1} = G_{t_2}) \Leftrightarrow (Intent_{t_1} = Intent_{t_2}) \Leftrightarrow (Intent_{t_1} \subseteq itemSet(t_2))$
- Symmetry breaking : $precede(G, [1, k_{max}])$

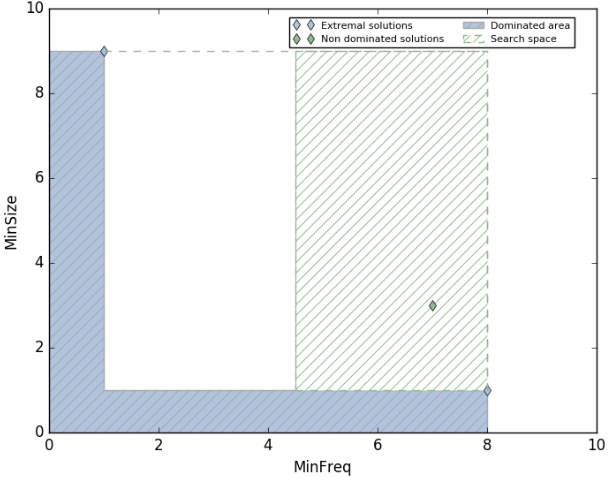


Different possible criteria to optimize :

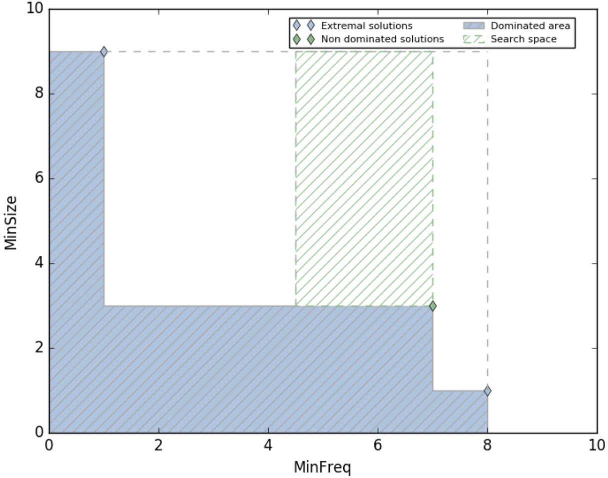
- Maximize the minimal frequency $\rightsquigarrow$ $minFreq = \min_{t \in \mathcal{T}} \# Extent_t$
- Maximize the minimal size $\rightsquigarrow$ $minSize = \min_{t \in \mathcal{T}} \# Intent_t$



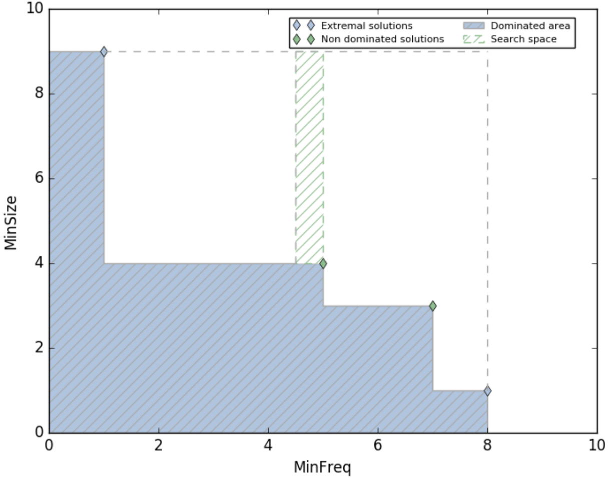
- Decomposition in 2 sub-problems : search for solutions on the high part of *minFreq* domain using the heuristic favoring frequency



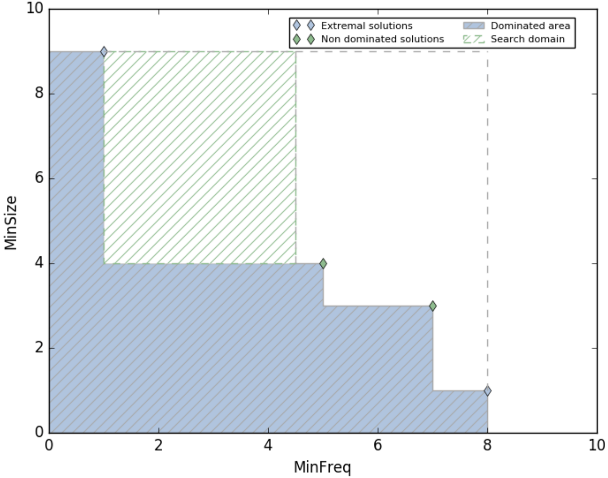
- Non-dominated solutions on high part of *minFreq* domain



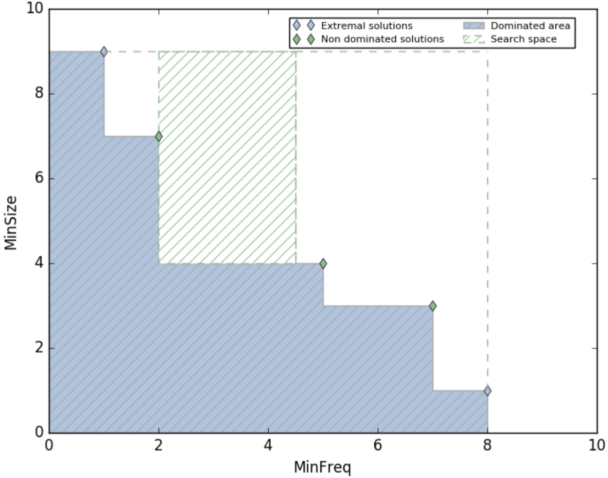
- Non-dominated solutions on high part of *minFreq* domain



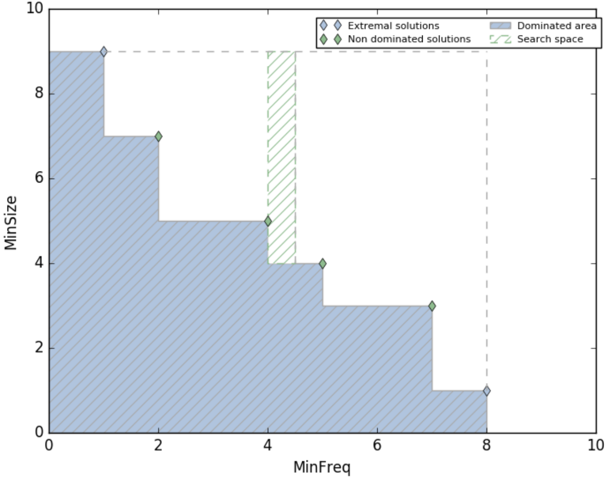
- Search for solutions on the low part of *minFreq* domain using the heuristic favoring high size solutions



- Non-dominated solutions on low part of *minFreq* domain



- Non-dominated solutions on low part of *minFreq* domain



- Non-dominated solutions on low part of *minFreq* domain