# 1 Objectif

#### Extraction des itemsets à l'aide du composant FREQUENT ITEMSETS.

La recherche des régularités dans les bases de données est l'idée principale du data mining. Ces régularités s'expriment sous différentes formes. Dans l'analyse du panier d'achats de consommateurs, l'extraction des itemsets consiste à mettre en exergue les cooccurrences entres les produits achetés c.-à-d. déterminer les produits (les items) qui sont « souvent » achetés simultanément. On parle alors d'itemsets fréquents. Par exemple, en analysant les tickets de caisse d'un supermarché, on pourrait produire des itemsets (un ensemble d'items) du type « le pain et le lait sont présents dans 10% des caddies ».

La recherche des itemsets fréquents est souvent présentée comme un préalable à l'extraction des règles d'association où l'on essaie, en sus, de mettre en évidence des relations de causalité. En reprenant notre exemple ci-dessus, une règle possible serait « ceux qui ont acheté du pain et du lait ont aussi acheté du beurre ». L'objectif est d'exploiter ce type de connaissance pour mieux agencer les rayons (mettre le beurre pas trop loin du pain et du lait) ou pour faire une offre promotionnelle ciblée (faire une promotion sur le pain et le lait dans le but d'augmenter les ventes de beurre).

En réalité, les itemsets fréquents sont en elles-mêmes porteuses d'informations. Savoir quels sont les produits achetés ensembles permet d'identifier les liens existants entre eux et, par là, de réaliser une typologie des achats ou de dégager des comportements types chez les consommateurs. Dans le cas du pain et du lait, il s'agit certainement d'achats relatifs au petit déjeuner. Si les consommateurs se mettent à acheter conjointement de la viande et du charbon, nous sommes en été, c'est la saison des barbecues...

Dans ce tutoriel, nous décrivons la mise en œuvre du composant FREQUENT ITEMSETS de Tanagra, basé sur la bibliothèque « apriori.exe » de Borgelt<sup>1</sup>. Nous utilisons un petit jeu de données pour que tout un chacun puisse reconstituer manuellement les résultats produits par le logiciel. Mais, dans un premier temps, essayons d'expliciter les différentes notions liées à l'extraction des itemsets.

## 2 Extraction des itemsets

<b>S1</b>	S2	S3	<b>S4</b>
1	0	1	0
0	1	0	0
0	0	0	1
0	1	1	1
0	1	1	0
0	1	1	0
1	1	1	1
1	0	1	0
1	1	1	0
1	1	1	0

Notre fichier comporte 10 observations (transactions) et 4 items<sup>2</sup>.

<sup>1</sup><u>http://www.borgelt.net/apriori.html</u> (Version 5.57)

<sup>2</sup> <u>http://www.dataminingarticles.com/closed-maximal-itemsets.html</u>

En ligne, nous avons des clients d'une compagnie d'assurance ; en colonne, des contrats (produits) associés à divers risques. Par exemple, le client n°1 a contracté les assurances S1 et S3, etc.

L'objectif est d'identifier les produits qui sont placés de concert.

Item. Un item correspond à un produit. Nous avons 4 items (S1, S2, S3 et S4) dans notre fichier.

**Support**. Le support d'un item est égal au nombre de transactions dans lesquelles il apparaît. Par exemple, le support de {S1} est égal à 5. Le support peut être exprimé également en termes relatifs. Dans ce cas, nous division le support (absolu) par le nombre total de transactions. Pour S1, nous avons SUP({S1]) = 5 /10 = 20%.

**Itemset**. Un itemset est un ensemble d'items (ex. {S1,S2} est un itemset de cardinal CARD({S1, S2} = 2). Le support d'un itemset comptabilise le nombre de transactions dans lesquelles les items apparaissent simultanément (ex. SUP({S1,S2}) = 3 /10 = 0.3). Un itemset peut être composé d'un singleton (ex. {S1} avec SUP({S1}) = 5/10).

**Itemset fréquent**. Un itemset est dit fréquent si son support est supérieur à un seuil défini à l'avance, paramètre de l'algorithme de recherche. Dans notre exemple, en fixant le support minimum à 2 (ou 20% en relatif), nous observons dans le schéma suivant les itemsets fréquents (toutes les combinaisons non-grisées).



Figure 1 - Extraction des différents itemsets

**Superset**. Un superset est un itemset défini par rapport à un autre itemset. Prenons un exemple pour clarifier les idées : {S1, S2, S3} est un superset de {S1, S2}. Ainsi, de manière générale, B est un superset de A, si CARD(A) < CARD(B) et que A  $\subset$  B c.-à-d. on retrouve dans B tous les items de A. Remarquons une propriété très importante du support : SUP(B)  $\leq$  SUP(A). En particulier, si A n'est

Page 2

pas fréquent, alors B ne le sera pas également. Ce résultat permet de réduire considérablement l'espace de recherche lors de l'extraction des itemsets fréquents dans une base de données.

**Itemset fermé (closed itemset)**. Un itemset fréquent est dit fermé si aucun de ses supersets n'a de support identique. Autrement dit, tous ses supersets ont un support strictement plus faible. Dans notre exemple ci-dessus, {S1, S3} est fermé car aucun de ses supersets n'a de support égal à 5/10 : SUP ({S1, S2, S3}) = 3/10, SUP({S1, S3, S4}) = 1/10.

**Itemset maximal (maximal itemset)**. Un itemset est dit maximal si aucun de ses supersets n'est fréquent. Dans notre exemple ci-dessus, {S1, S2, S3} est maximal car son superset {S1, S2, S3, S4} (il n'y en a qu'un) n'est pas fréquent avec un support de 1/10.

**Itemset générateur (generator itemset)**. Un itemset A est dit générateur s'il n'existe aucun itemset B tel que  $B \subset A$  et que SUP(B) = SUP(A). Autrement dit, l'itemset est générateur si tous ses sousitemsets ont un support strictement supérieur. Dans notre exemple, {S1, S2, S3} de support 4/10 n'est pas générateur puisqu'on trouve {S1, S2} avec un support identique. Il en est de même en ce qui concerne {S1, S3} à cause de {S1}. En revanche, {S2, S4}, de support 2/10, est générateur parce que SUP({S2}) = 7/10 et SUP({S4}) = 3/10.

## 3 Extraction des itemsets à l'aide de Tanagra

Nous chargeons le fichier «<u>itemset\_mining.xls</u> » dans le tableur Excel. Nous sélectionnons la plage de données, puis nous actionnons le menu TANAGRA / EXECUTE TANAGRA installé à l'aide de la macro complémentaire tanagra.xla<sup>3</sup>.



Tanagra est automatiquement démarré et les données chargées.

<sup>&</sup>lt;sup>3</sup> Voir <u>http://tutoriels-data-mining.blogspot.com/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html</u> pour l'installation et l'utilisation de la macro dans Excel 2007 et 2010. Elle est également fonctionnelle sous les versions antérieures d'Excel (<u>http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html</u> - Excel 97 à 2003). Enfin, une procédure analogue existe pour OpenOffice et LibreOffice (voir <u>http://tutoriels-data-mining.blogspot.com/2011/07/tanagra-addon-pour-openoffice-33.html</u>).

TANAGRA 1.4.41 - [Datase	t (tanC84D.txt)] nent Window He	In				X
Analysis Dataset (tanC84D.txt)		Dataset description 4 attribute(s) 10 example(s)				
		Attribute	Category	Informations		
		S1	Continue	-		E
		52	Continue			
		54	Continue			-
		Comp	onents			
Data visualization	Statistics	N	lonparame	tric statistics	Instance selection	
Feature construction	Feature select	tion	Regression		Factorial analysis	
PLS	Clustering		Spv learning		Meta-spv learning	
Spv learning assessment	Scoring		Asso	ciation		
Correlation scatterplot	🖉 Scatterplot		🕵 View	dataset		
Export dataset	Scatterplot w	vith label	View	multiple scatt	erplot	
L						

Nous spécifions les variables de l'analyse en insérant le composant DEFINE STATUS. Nous plaçons toutes les variables en INPUT.

TANAGRA 1.4.41 - [Dataset (tanC	84D.txt)]
Tile Diagram Component	Hindow Help
Analysis	Define attribute statuses
⊡ ⊡ Dataset (tanC84D.txt)	
Define status 1	Parameters
	Attributes : Target Input Illustration
	C S2 S2 S2 E
	<b>C</b> S4 S4
	→ I
Data visualization	
Feature construction F	
PLS	
Spv learning assessment	E, E, Clear all Clear selected
Correlation scatterplot	
Export dataset	OK Cancel Help

#### 3.1 Itemsets fréquents

Pour produire les itemsets fréquents, nous ajoutons le composant FREQUENT ITEMSETS dans le diagramme. Nous actionnons le menu PARAMETERS afin de spécifier les paramètres de l'analyse.

TANAGRA 1.4.41 - [Datas	et (tanC84D.txt)]		
Tile Diagram Compo	onent Window Help		_ 8 ×
D 📽 🔚   🎎			$\frown$
Analysis		Dataset des	requent Itemsets
□- 🛄 Dataset (tanC84D.t	kt)	(attributo(c)	
🖃 🏠 Define status 1		10 example(z)	Parameters
Frequent Ite	emsets 1 Paramete	ers egor	Min Support: 0.2
	Evecute	itinue	Max support: 1
	View	itinue	Min length : 2
		54 Continue	Max length : 4
<u></u>	][_		Itemset type
Data visualization	Statistics	Components	Frequent
Feature construction	Feature selection	n Re	Closed
PIS	Clustering	Sov	© Generators
Spy learning assessment	Scoring	٨٩٩	
A priori	A orieri DT	"Erequent Itemset	Rule base : outfreq.rul
B A priori MR	Assoc Outlier	Spy Assoc Rule	
			OK Cancel Help

Nous descendons le support minimum à 20%. Nous ne modifions pas les autres paramètres. Par défaut, l'outil extrait les itemsets fréquents de cardinal compris entre MIN LENGTH = 2 et MAX LENGTH = 4. Les itemsets composés d'un singleton ne sont donc pas générés.



Il ne nous reste plus qu'à cliquer sur le menu contextuel VIEW. Le module « apriori.exe » est lancé en sous main dixit la fenêtre de suivi. Les itemsets – au nombre de 7 – s'affichent dans la partie inférieure. Les résultats concordent avec le graphe des itemsets (les itemsets non-grisés dans Figure 1), à l'exception des itemsets de cardinal 1 que nous avons sciemment mis de côté.

#### 3.2 Itemsets fermés

Pour obtenir les itemsets fermés, nous re-paramétrons le composant en actionnant le menu contextuel PARAMETERS. Nous choisissons la seconde option (**ITEMSET TYPE = CLOSED**).

ameters					
Min Su	pport :	0.2			
Max su	pport :	1			
Min le	ength :	2			
Max le	ength :	4			2
ltem ◎ Fr ◎ Cl ◎ M ◎ G	set type requent osed aximal enerato				
Rule	base :	outfre	q.rul		
	Min Su Max su Min ka Max ke Item © Fr @ Cl @ M @ G Rule b	Min Support : Max support : Max support : Min length : Max length : Itemset type © Frequent @ Closed © Maximal © Generato Rule base :	Min Support: 0.2 Max support: 1 Min length: 2 Max length: 4 Itemset type © Frequent @ Closed @ Maximal @ Generators Rule base : outfre	Min Support : 0.2 Max support : 1 Min length : 2 Max length : 4 Itemset type © Frequent © Closed © Maximal © Generators Rule base : outfreq.rul	Min Support: 0.2 Max support: 1 Min length: 2 Max length: 4 Itemset type © Frequent © Closed © Maximal © Generators Rule base : outfreq.rul

Nous validons et nous cliquons sur VIEW. Nous obtenons 4 itemsets (en vert et bleu dans la Figure 1, à l'exception des singletons).

TANAGRA 1.4.41 - [Freque	ent Itemsets 1]					
Tile Diagram Compo	🝸 File Diagram Component Window Help 📃 🖅					
🗅 📽 🔚  🎎						
Analysis				Execution log		
Dataset (tanC84D.tx Define status 1	t) msets 1	D:\Temp\Exe\exe\apriori.exe - find frequent item sets with the apriori algorithm version 5.57 (2011.09.02) (c) 1996-2011 Christian Borgelt reading C:\Users\Maison\AppData\Local\Temp\datYB97.tmp [4 item(s). 10 transaction(s)] done [0.00s]. filtering, sorting and recoding items [4 item(s)] done [0.00s]. sorting and reducing transactions [6/10 transaction(s)] done [0.00s]. building transaction tree [8 node(s)] done [0.00s]. checking subsets of size 1 2 3 done [0.00s].				
			ITEM	SETS [#4 itemsets loaded]		
		N*         Descr           1         S4 ∧ S           2         S1 ∧ S           3         S1 ∧ S           4         S2 ∧ S	iption 52 /\ S3 52 /\ S3 53 53	11		
			Components			
Data visualization	Statistics	Statistics Nonparametric statistics Instance selection Feature construction				
Feature selection	Regression	n	Factorial analysis	PLS	Clustering	
Spv learning	Meta-spv lear	rning Spv learning assessment Scoring Association			Association	
A priori 84	A priori MR	8 A priori P	T Assoc Outlie	er 🗜 Frequent Itemse	ts 🏶 Spv Assoc Rule 🗄	
		_			a.	

#### 3.3 Itemsets maximaux

Nous réitérons le même schéma avec l'option (**ITEMSET TYPE = MAXIMAL**) dans la boîte de paramétrage. Nous obtenons à la sortie 2 itemsets, ceux qui sont en vert dans le graphe (Figure 1).

TANAGRA 1.4.41 - [Frequ	ent Itemsets 1]						
Tile Diagram Comp	onent Window Help						
🗅 📽 🖬   🎎		1					
Analysis		Execution log					
⊡-∰ Dataset (tanC84D.t ☐-∰ Define status 1 ☐-∰ Frequent It	t) D:\Temp\Exe\exe\apriori.exe - find frequent item seversion 5.57 (2011.09.02) (c) 1996-2011 Christerading C:\Users\Maison\AppData\Local\Temp\dat [0.00s]. filtering. sorting and recoding items [4 item(s)] ds sorting and reducing transactions [6/10 transactio building transaction tree [8 node(s)] done [0.00s]. checking subsets of size 1 2 3 done [0.00s].			iuent item sets with the aprio -2011 Christian Borgelt cal/Temp\dat396A.tmp [4 it [4 item(s)] done [0.00s]. /10 transaction(s)] done [0.00 done [0.00s]. .00s].	sets with the apriori algorithm ristian Borgelt at396A.tmp [4 item(s), 10 transaction(s)] done done [0.00s]. tion(s)] done [0.00s]. s].		
		N* D 1 S4 2 S1	ITEM	SETS [#2 itemsets loaded]	<b>&gt;</b>		
			Components				
Data visualization Feature selection Spv learning BA priori B BA priori MR	Statistic Regressio Meta-spv lea A priori PT Assoc Outlier	s n rning Frequ Spv As	Nonparametric statistics Factorial analysis Spv learning assessment ent Itemsets 'E Spv Assoc To ssoc Rule	Instance selection PLS Scoring ree	Feature construction Clustering Association		

#### 3.4 Itemsets générateurs

Enfin, pour obtenir les itemsets générateurs, nous sélectionnons la dernière option (**ITEMSET TYPE = GENERATORS**) dans la boîte de paramétrage. Nous obtenons 4 itemsets de cardinal égal à 2.

TANAGRA 1.4.41 - [Frequ	ent Itemsets 1]						
Tile Diagram Compo	onent Window	v Help					
D 📽 🖪   🎎							
Analysis			Execution log				
Dataset (tanC84D.txt)		D:\Tem version reading [0.00s]. filtering sorting building	D:\Temp\Exe\exe\apriori.exe - find frequent item sets with the apriori algorithm version 5.57 (2011.09.02) (c) 1996-2011 Christian Borgelt reading C:\Users\Maison\AppData\Local\Temp\dat9F8D.tmp [4 item(s), 10 transaction(s)] done [0.00s]. filtering, sorting and recoding items [4 item(s)] done [0.00s]. sorting and reducing transactions [6/10 transaction(s)] done [0.00s]. building transaction tree [8 node(s)] done [0.00s].				
			ITEN	ASETS [#4 itemsets loaded]			
		N° C	Description				
		1 S 2 S 3 S 4 S	4 /\ \$2 4 /\ \$3 1 /\ \$2 2 /\ \$3	m	, E		
Data visualization	Statist	CS.	Components Nonparametric statistics	Instance selection	Feature construction		
Feature selection	Regress	ion	an Factorial analysis PI S Clusterin		Clustering		
Spv learning	Meta-spv le	arning Spv learning assessment Scoring Association		Association			
St A priori	🕃 A priori MR 🛛 🕄 A priori PT 📰 Assoc Outlier 📲 Frequent Itemsets 💐 Spv Assoc Rule †				ts 🅄 Spv Assoc Rule 🗄 🗄		
•			III		•		
					.17		

## 4 Extraction des itemsets à l'aide de R (package « arules »)

La procédure « **apriori** » du package « **arules**<sup>4</sup> » est également basé sur la bibliothèque de Borgelt. Il est donc possible de reproduire les résultats ci-dessus en introduisant les commandes et les paramétrages adéquats dans le logiciel R. Nous décrivons succinctement les principales commandes dans ce qui suit, en leur associant les résultats affichés par l'outil.

#### 4.1 Importation des données

Nous utilisons la commande « read.xls » du package « xlsReadWrite<sup>5</sup> » pour lire le fichier de données « itemset\_mining.xls ». Voici les instructions associées.

#vider la mémoire
rm(list=ls())
#importation des données
library(xlsReadWrite) #chargement de la bibliothèque
setwd("") #modifier le répertoire de travail
#la première ligne correspond au nom des items
#les données sont situées dans la première feuille du classeur XLS
dataset <- read.xls(file="itemset_mining.xls",colNames=T,sheet=1)
#affichage des valeurs
print(dataset)

R nous fournit les informations suivantes.



#### 4.2 Extraction des itemsets fréquents

Pour extraire les itemsets fréquents, nous chargeons au préalable la bibliothèque, nous spécifions les paramètres de l'analyse, enfin nous lançons les calculs. Attention, pour que la procédure comprenne bien que les données se présentent sous la forme d'une matrice o/1 de présence-absence des items dans les transactions, nous devons effectuer un transtypage avec la commande as.matrix(...).

<sup>&</sup>lt;sup>4</sup> <u>http://cran.r-project.org/web/packages/arules/index.html</u>

<sup>&</sup>lt;sup>5</sup> <u>http://cran.r-project.org/web/packages/xlsReadWrite/index.html</u>

#chargement du package library(arules) #paramétrage de la méthode params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="frequent itemsets") #lancement des calculs result <- apriori(as.matrix(dataset), parameter = params) #affichage des itemsets inspect(result)

R affiche les itemsets, les résultats concordent en tous points à ceux de Tanagra. Ce n'est guère étonnant puisque les deux outils s'appuient sur la même bibliothèque de calcul.

```
R Console
                                                              apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)
                                 (c) 1996-2004
                                                Christian Borgelt
set item appearances ... [0 item(s)] done [0.00s].
set transactions ... [4 item(s), 10 transaction(s)] done [0.00s].
sorting and recoding items ... [4 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [7 set(s)] done [0.00s].
creating S4 object ... done [0.00s].
> inspect(result)
  items support
1 {S2,
   S4}
            0.2
2 {53,
   S4}
            0.2
3 {S1,
   S2}
            0.3
  {S1,
            0.5
   S3}
```

<u>Remarque</u> : Notons cependant que le package « **arule** » est basé sur la version **4.21 de** « **apriori.exe** » de Borgelt. L'extraction des itemsets générateurs n'est pas disponible.

#### 4.3 Extraction des autres types d'itemsets

Les résultats étant de la même teneur que ceux de Tanagra, nous nous contentons de décrire les paramètres de calculs pour chaque type d'itemset dans cette sous-section.

```
#extraction des itemsets fermés
params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="closed frequent itemsets")
result <- apriori(as.matrix(dataset), parameter = params)
inspect(result)
#extraction des itemsets maximaux
params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="maximally frequent itemsets")
result <- apriori(as.matrix(dataset), parameter = params)
inspect(result)</pre>
```

# 5 Conclusion

La sortie de la version 1.4.41 de Tanagra a été l'occasion de mettre à jour la bibliothèque externe « apriori.exe » de Borgelt (version 5.57), nettement plus performante en termes de temps de traitements<sup>6</sup>. Nous en avons profité pour introduire un nouveau composant de génération des itemsets fréquents (FREQUENT ITEMSETS), toujours basée sur la même bibliothèque, que nous décrivons dans ce didacticiel.

<sup>&</sup>lt;sup>6</sup> Voir <u>http://tutoriels-data-mining.blogspot.com/2011/09/mise-jour-de-apriori-pt.html</u>