2èmes rencontres R, Lyon, 28 juin 2013

Traiter des données de tracking avec R
Navigation sur un site web, suivi d'enquête web ou téléphonique

Anne GAYET
Directrice Datamining A,I,D,
agayet@aid,fr

CRM, Datamining, Data Quality, Big Data

Exemples de données: suites d'événements

Evénements ordonnés, souvent horodatés à la seconde

Enquête téléphonique tentatives d'appel	Enquête web écrans affichés et validés	Site web pages vues
non contact	accueil	accueil
non contact	écran Q1/Q2	identification
non contact	clic	oubli pw
non contact	clic	espace client
contact / Rdv répondant	suivant	mes contrats
interview	écran Q3	…
finalisation	clic	…
	retour	…
	suivant	…
	…	…

Les méthodes de datamining adaptées

- les règles d'association
- les motifs séquentiels
- le clustering de séquences
- le biclustering ou coclustering

Les packages R utilisés ici:

- arules: Mining Association Rules and Frequent Itemsets, Michael Hahsler, Bettina Gruen and Kurt Hornik (2013), V1,0-14,
- arulesSequences: Mining frequent sequences, Christian Buchta and Michael Hahsler, V0,2-4,
- arulesViz: Visualizing Association Rules and Frequent Itemsets, Michael Hahsler and Sudheer Chelluboina, V0,1-5,
- TraMineR: Trajectory miner: a toolbox for exploring and rendering sequence data, Alexis Gabadinho, Matthias Studer, Nicolas S, Müller, Reto Bürgin and Gilbert Ritschard, V1,8-5,

Références

[1] Nicolas S, Müller, Matthias Studer, Alexis Gabadinho, Gilbert Ritschard (2010), Analyse de séquences d'événements avec TraMineR, EGC 2010,

[2] Alexandre Pollien (FORS), Dominique Joye (ISS), Michèle Ernst Stähli (FORS), Marlène Sapin (FORS), Répondants et non-répondants dans les enquêtes, analyse des séquences de contact, 7ème colloque francophone sur les sondages 2012,

[3] Alexis Gabadinho, Gilbert Ritschard, Matthias Studer and Nicolas S, Muller (2011)
Mining sequence data in R with the TraMineR package: A user's guide

[4] Buchta, C,, Hahsler, M, (2010), "arulesSequences: Mining frequent sequences", R package, http://CRAN,R-project,org/package=arulesSequences

[5] Gabadinho, A,, Ritschard, G,, Müller, N, S,, Studer, M, (2011), Analyzing and Visualizing State Sequences in R with TraMineR, Journal of Statistical Software, 40(4), 1-37,

Ressource:
http://en,wikibooks,org/wiki/Data_Mining_Algorithms_In_R/Sequence_Mining/SPADE

Les règles d'association

Une règle simple : X → Y (Support,Confiance,Lift)

Support = proportion des séquences qui contiennent X et Y,

Ex 20% des clients ont acheté X et Y,

Confiance = P(Y/X) = proportion de séquences qui contiennent Y parmi celles qui contiennent X,

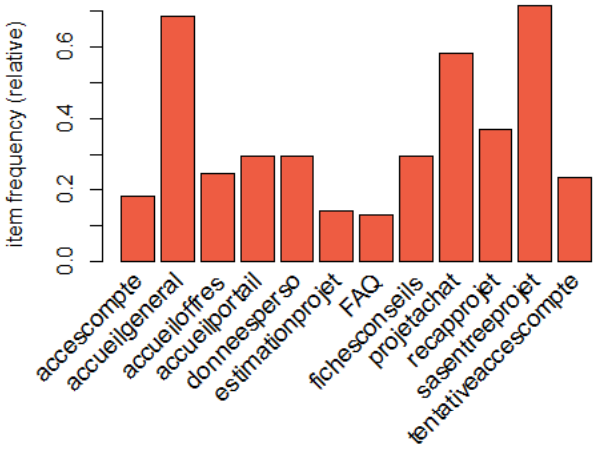
Ex: 70% des clients qui ont acheté X ont aussi acheté Y,

Règles complexes : A&B&… → Y&X&… (Support,Confiance,Lift)

Attention : l'ordre d'apparition des états n'est pas pris en compte, la flèche ne veut pas dire "ensuite"

Application à des pages vues sur un site web : un « panier » = une visite

```
itemFrequencyPlot
(PagesWeb,
support=0.2, cex =
0.7, col="tomato2")
```



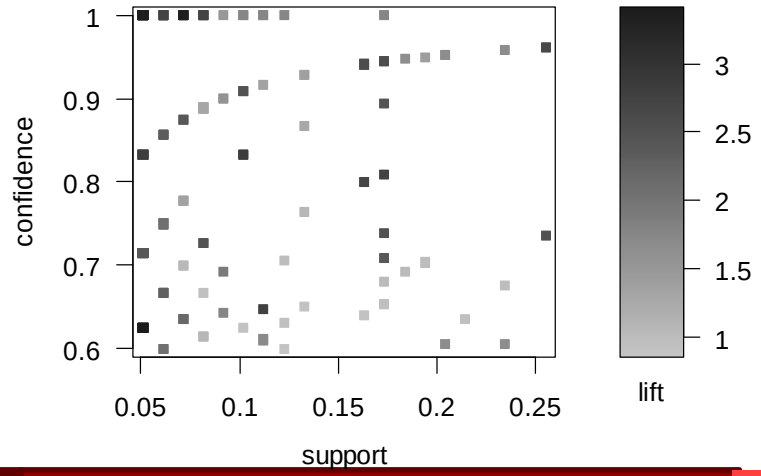
Application de l'algorithme apriori (Borgelt)

Extraction des règles

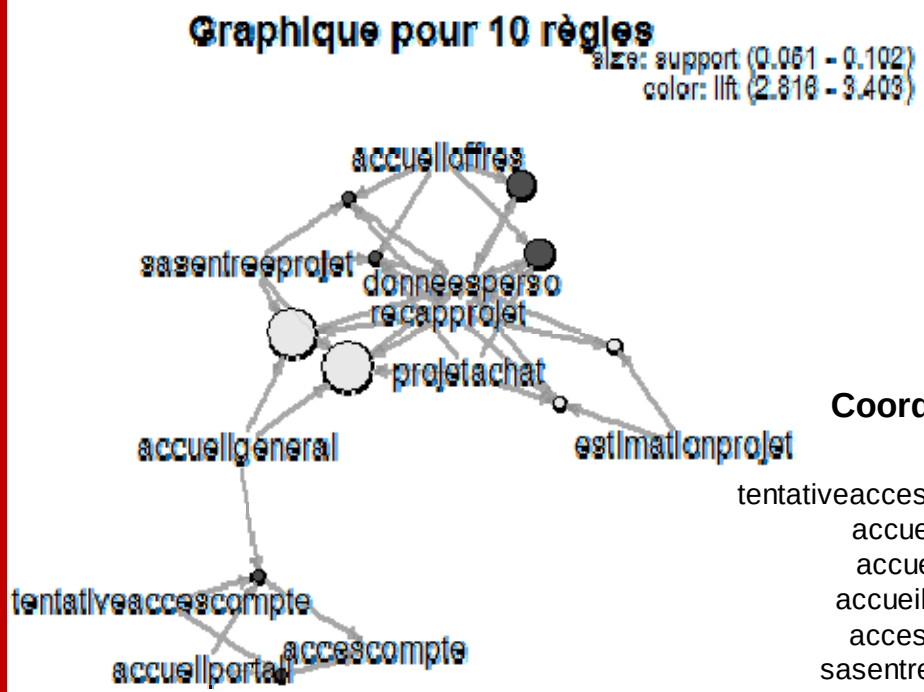
```
aid.rules<-apriori(data=PagesWeb, parameter=list(support=0.05,conf=0.6,
minlen=2, maxlen=10, target = "rules"))
```

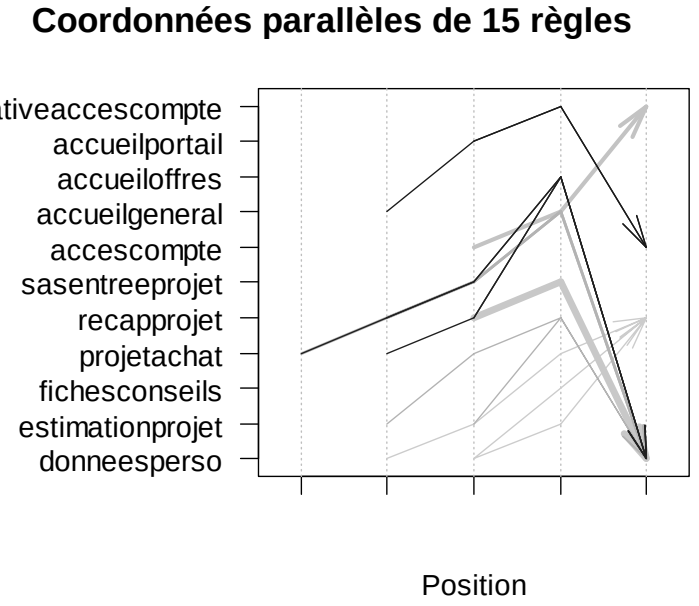
Affichage des 5 règles avec le lift le plus élevé et visualisation

lhs		rhs	support	confidence	lift
{accueilportail, tentativeaccescompte}	=>	{accescompte}	0,05	0,63	3,40
{accueilgeneral, accueilportail,tentativeaccescompte}	=>	{accescompte}	0,05	0,63	3,40
{accueiloffres,recapprojet}	=>	{donneesperso}	0,07	1,00	3,38
{accueiloffres,projetachat,recapprojet}	=>	{donneesperso}	0,07	1,00	3,38
{accueiloffres, recapprojet,sasentreeprojet}	=>	{donneesperso}	0,05	1,00	3,38



Visualisation des règles avec arulesViz





NB : avec les règles d'associations l'ordre avant/après n'est pas pris en compte si on applique un algorithme de motifs séquentiels, on pourrait utiliser les mêmes représentations graphiques et les flèches traduiront cette fois l'ordre d'apparition.

Motifs séquentiels ou séquences

Cette fois l'ordre d'apparition est pris en compte,

A **B** C D
A C **B** A D
B A A
B A D
B **A** **B** C D

A , B est de support 3
B , A est de support 4
B , D est de support 4
C , D est de support 3

ArulesSequences
Add-on de arules pour déterminer les séquences fréquentes, Il utilise l'algorithme cSPADE (Zaki),

TraMineR, Contraction de "Life Trajectory Miner for R",
gère un grand nombre d'états et la représentation de séquences d'événements ordonnées dans le temps (souvent des observations périodiques).

Extraction des motifs avec cspade


```
cspade.res<-cspade(navigation, parameter = list(support = 0.02),
control = list(verbose = TRUE))
```

Sequences	Support
<{sasentreeprojet,accueilgeneral,projetachat}>	0,02339181
<{projetachat},{recapprojet,donneesperso}>	0,02339181
<{accueilgeneral,accueilportail},{projetachat}>	0,02339181
<{recapprojet,sasentreeprojet,projetachat},{donneesperso}>	0,02339181
<{sasentreeprojet},{accueilgeneral},{accueilgeneral}>	0,02339181
<{accueilgeneral,accueilportail},{accueilgeneral}>	0,02339181
<{accueilgeneral,accueilportail},{recapprojet}>	0,02192982
<{accueilgeneral},{recapprojet,donneesperso,projetachat}>	0,02192982
<{accueilgeneral},{donneesperso,projetachat}>	0,02192982
<{sasentreeprojet},{estimationprojet}>	0,02192982

Confiance et lifts peuvent être calculés: ils concernent le dernier état vs les précédents

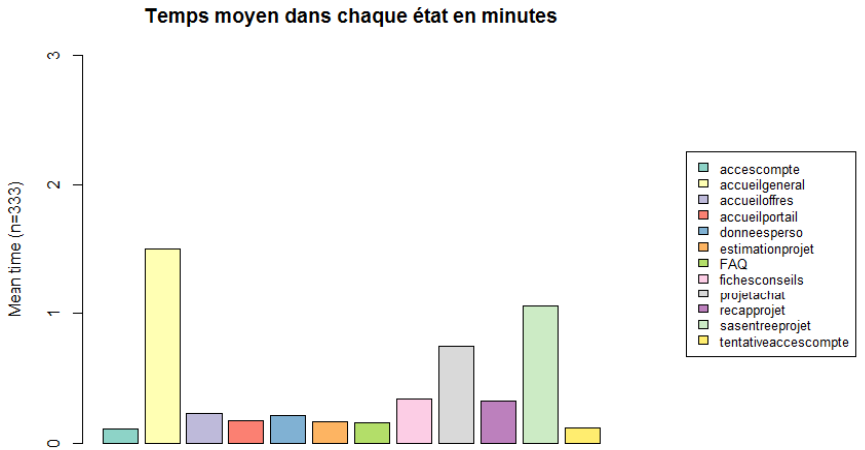
```
Regles<-ruleInduction(cspade,res, confidence=0,1,control=list(verbose=TRUE))
```

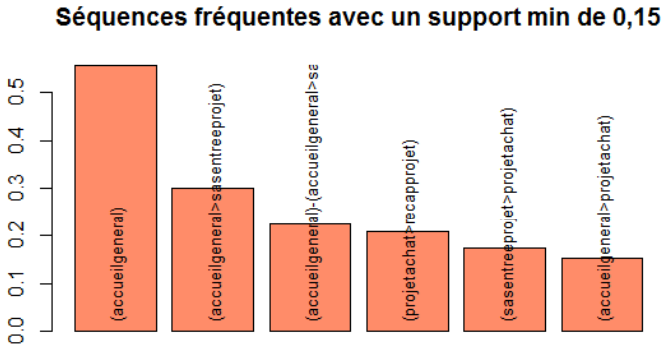
Regles	Confidence	Lift
{tentativeaccescompte,accueilgeneral}> => <{accueilgeneral}>	0,5769231	2,362966
{sasentreeprojet},{accueilgeneral}> => <{accueilgeneral}>	0,2711864	1,110728
{sasentreeprojet,accueilgeneral}> => <{accueilgeneral}>	0,3529412	1,445579
{accueilgeneral,accueilportail}> => <{accueilgeneral}>	0,3809524	1,560308
{accueilgeneral},{accueilgeneral}> => <{accueilgeneral}>	0,3170732	1,298671
{accueilgeneral}> => <{accescompte}>	0,1197605	3,033932

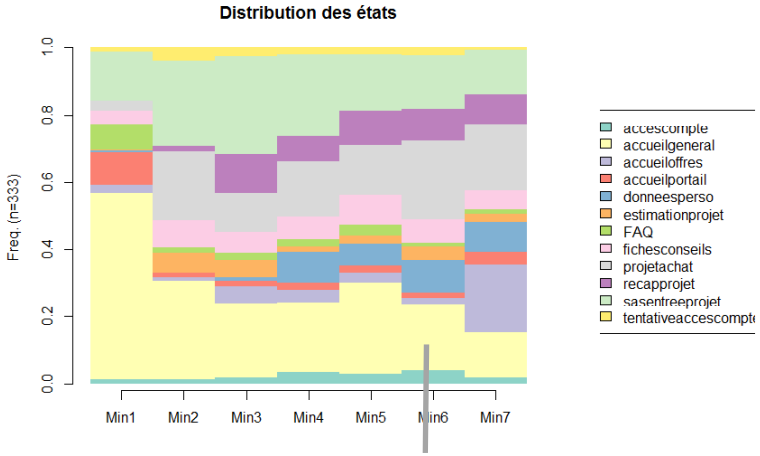
Motifs séquentiels avec TraMineR

TraMineR permet de prendre en compte l'ordre et les répétitions
→ nb de fois où on reste dans le même état

TraMiner analysant des mesures répétées: temps = nombre de répétitions

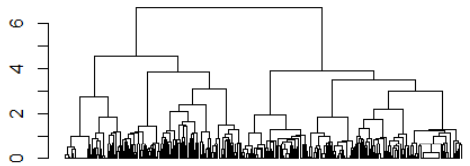






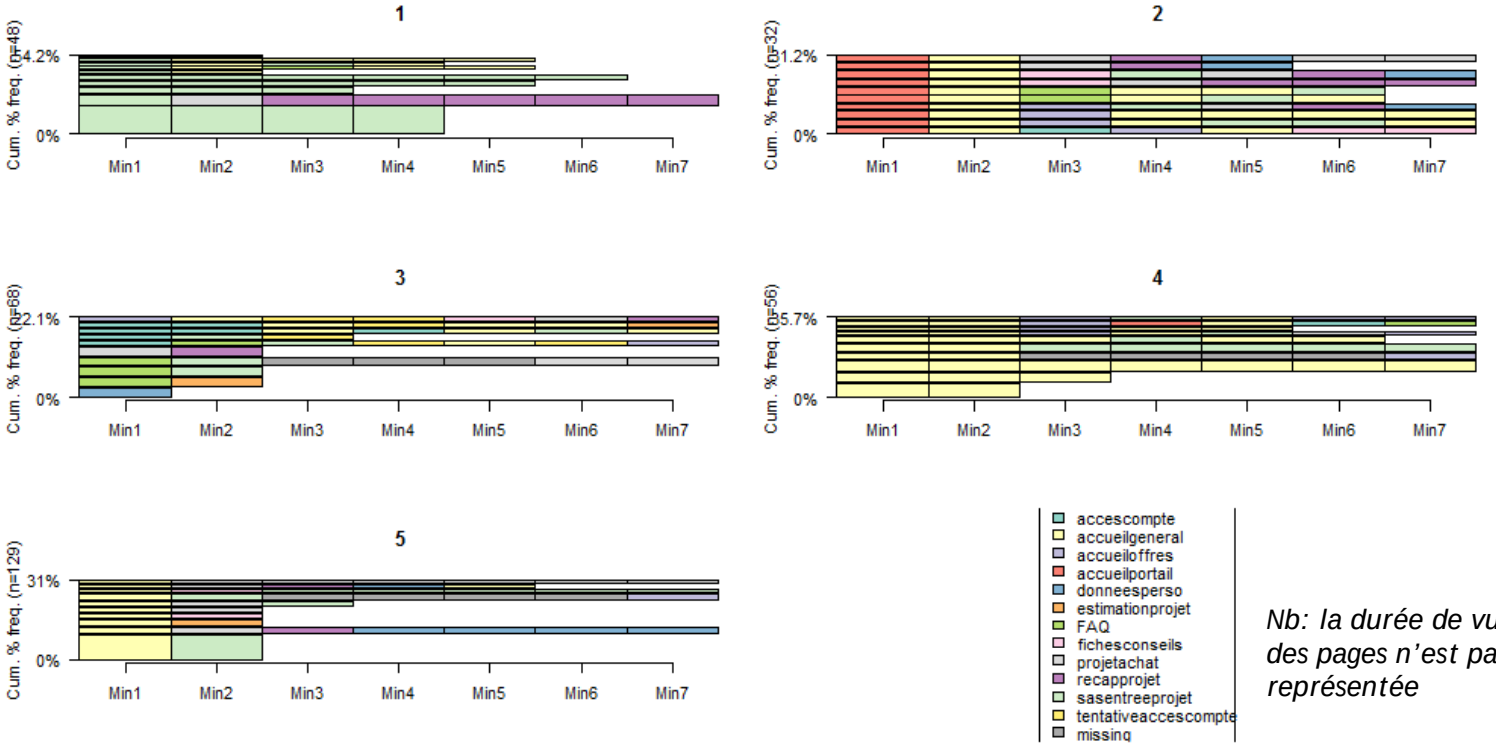
L'accueil général reste fréquent, même en position avancée dans la séquence

Matrice des transitions et Classification des séquences




```
clusterward <- agnes(seq.om, diss = TRUE, method = "ward")
seqfplot(seq.resh, group = cluster3, pbarw = T,cex.legend=1)
```

	[-> accescompte]	[-> accueilgeneral]	[-> accueiloffres]	[-> accueilportail]	[-> donneesperso]	[-> estimationprojet]	[-> FAQ]
[accescompte ->]	0.41	0.24	0.03	0.03	0.00	0.00	0.10
[accueilgeneral ->]	0.03	0.41	0.04	0.01	0.00	0.04	0.02
[accueiloffres ->]	0.00	0.31	0.33	0.05	0.00	0.00	0.05
[accueilportail ->]	0.02	0.70	0.00	0.14	0.02	0.00	0.00
[donneesperso ->]	0.00	0.12	0.07	0.05	0.45	0.00	0.00
[estimationprojet ->]	0.00	0.08	0.00	0.00	0.00	0.46	0.03
[FAQ ->]	0.00	0.26	0.02	0.02	0.00	0.04	0.19
[fichesconseils ->]	0.00	0.07	0.02	0.00	0.00	0.06	0.00



Nb: la durée de vue des pages n'est pas représentée

Algorithmes de motifs séquentiels

Algorithmes mis à disposition dans les versions « coûteuses des logiciels de datamining »
De nombreux algorithmes depuis cSPADE, de nombreuses subtilités

…

Peu d'implémentation en R

Voir (par exemple):

- PrefixSPAN et ses héritiers
- GTC/GETC
- RuleGrowth, TruleGrowth (Philippe Fournier- Viger), et l'ensemble des algorithmes implémentés dans SPMF

+ Besoin de fonctionner sur de grosses volumétries
+ En l'état beaucoup de travail de « reformatage » des données

Coclustering: faire des clusters en même temps sur les lignes et les colonnes d'un tableau

Le tableau: lignes = visites

colonnes = pages

contenu = 1/0 ou fréquences

Partitions	Pages 1	Pages 2	Pages 3	Pages 4	Pages 5
Visites 1	328	1390	140	66	381
Visites 2	48	81	14	128	**618**
Visites 3	133	86	63	**1223**	163
Visites 4	41	58	**625**	67	56
Visites 5	651	**3686**	45	118	40

De nombreuses méthodes et packages

→ communication future

Opportunités:

- Tester sur données de tracking la pertinence des algorithmes de biclustering appliqués habituellement aux microarrays
- Prétraiter la matrice des fréquences par une transformation augmentant l'importance des états peu fréquents
- Enchaîner un coclustering puis des motifs séquentiels par cluster.