1 Objectif

Description du logiciel d'extraction de règles d'association intégré dans la distribution SIPINA.

SIPINA est surtout connu pour ses algorithmes d'induction d'arbres de décision. En réalité, la distribution inclut deux autres outils peu connus du grand public : REGRESS, spécialisé dans la régression linéaire multiple, nous l'avions décrit dans un de nos anciens tutoriels[1] ; et un logiciel d'extraction de règles d'association, appelé prosaïquement « Association Rule Software » (ARS).

A l'époque (1997-1998), mon idée était de construire une série d'exécutables indépendants organisés autour de la même grille de gestion des données, un peu à la manière de la version contemporaine du logiciel STATISTICA. Des logiciels pour l'analyse factorielle et la classification automatique ont même été initiés mais non finalisés. Par la suite, je me suis rendu compte que le fonctionnement par diagramme de traitements – le standard initié par SPAD avec sa fameuse « filière » qui fait référence aujourd'hui – était plus avantageux à bien des égards, en termes de facilité d'utilisation pour l'usager, mais aussi en termes d'évolution du logiciel. Chaque nouvelle méthode implémentée engendre un composant supplémentaire simplement, et non pas un nouvel exécutable. TANAGRA a été développé dans cette optique en 2003. Initialement, chaque technique statistique devait correspondre à un plug-in (une DLL). L'architecture interne a été élaborée en ce sens d'ailleurs. Mais j'ai finalement opté pour la simplicité. Le SETUP fait à peine 2.8 Mo (version 1.4.48), la mise à jour du logiciel par le remplacement de l'exécutable principal est très simple.

Mais revenons à notre propos. Depuis TANAGRA, qui intègre plusieurs composants pour l'extraction des règles d'association, ARS est très peu utilisé. J'en parle très peu moi-même. Mais comme l'outil apparaît dans le menu intégré dans Excel via la macro complémentaire « sipina.xla »[2], plusieurs internautes m'ont demandé plus de précisions quant à son utilisation et ses spécificités.

Dans ce tutoriel, je décris la mise en œuvre du logiciel ARS lorsqu'on le lance à partir d'Excel. Comme nous le verrons par la suite, l'interaction avec le tableur introduit des opportunités très précieuses en matière d'exploration des résultats. C'est un aspect très important tant la profusion des règles peut rapidement déconcerter. Pourvoir les filtrer et les trier de différentes manières est un atout fort dans la détection des règles les plus « intéressantes » au regard des objectifs de l'utilisateur. Les outils d'Excel nous seront d'un très grand secours à cet égard.

2 Données

Le fichier « **market_basket.xlsx** »[3] décrit le contenu de n = 1361 caddies (transactions). Nous disposons d'une bibliothèque de p = 303 produits (items). En moyenne, les caddies contiennent 9.5 articles (min = 0 ; max = 303, un client aurait pris tous les produits du magasin ! bonjour le passage en

[1] « REGRESS dans la distribution SIPINA », http://tutoriels-data-mining.blogspot.fr/2011/05/regress-dans-la-distribution-sipina.html

[2] Voir http://tutoriels-data-mining.blogspot.fr/2008/03/connexion-excel-sipina.html pour l'installation de l'add-on dans Excel 2003 et versions antérieures ; on peut s'inspirer du document dédié à Tanagra http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour Excel 2007 et 2010.

[3] Récupéré sur le site http://inf.abdn.ac.uk/~hnguyen/teaching/CS5553/prac05.php (**marketbasket.csv**).

caisse) ; et les articles ont été achetés 42.7 fois en moyenne (min = 7, max = 167). Les 5 produits les plus populaires sont : « Eggs », « White bread », « 2pct milk », « Potato chips » et « 98pct fat free hamburger ». Les moins prisés sont : « Celery », « Oats and Nuts Cereals », « Chicken legs », « Nasal spray » et « Daily Newspaper ».

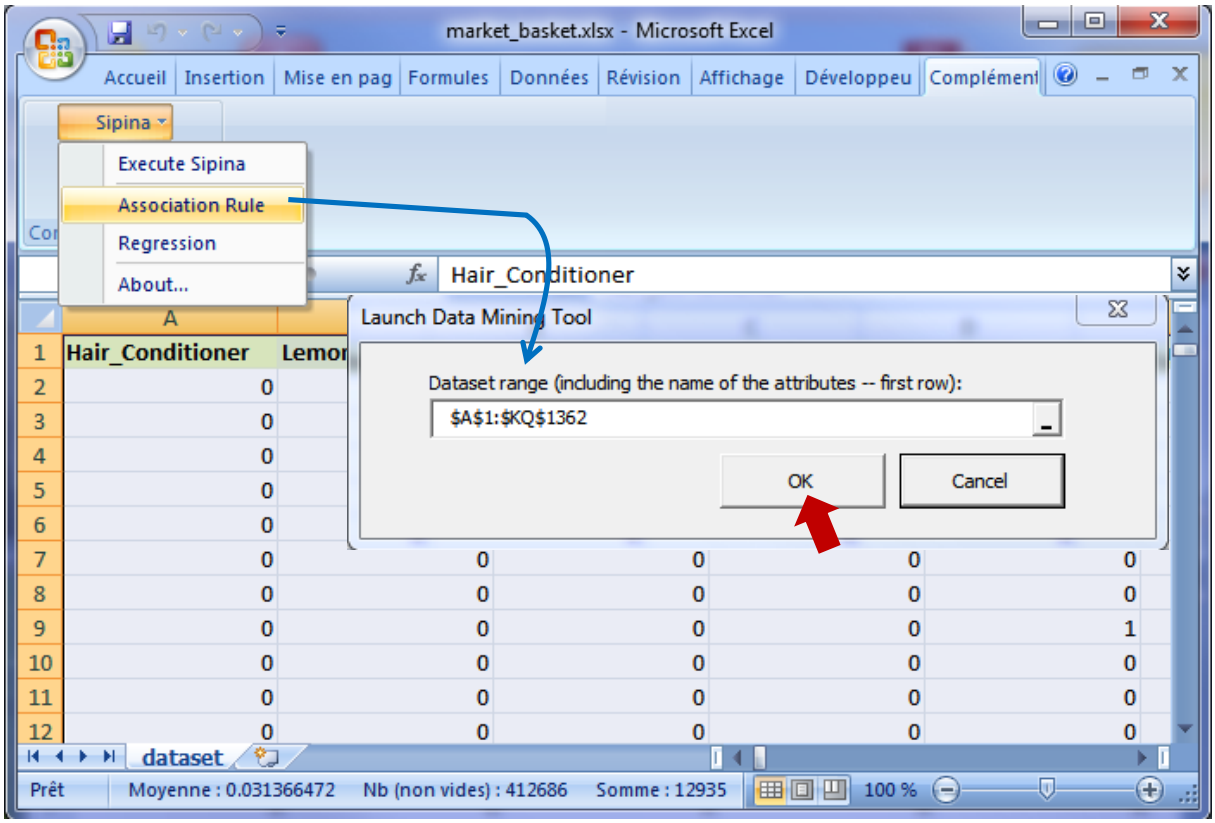
3 Construction des règles d'association avec ARS

3.1 Importation des données

Nous devons disposer d'au moins **Excel 2007** pour manipuler le fichier « market_basket.xlsx ». En effet, le nombre d'items (p = 303) excède la limitation de 256 colonnes d'Excel 2003 (et antérieures).

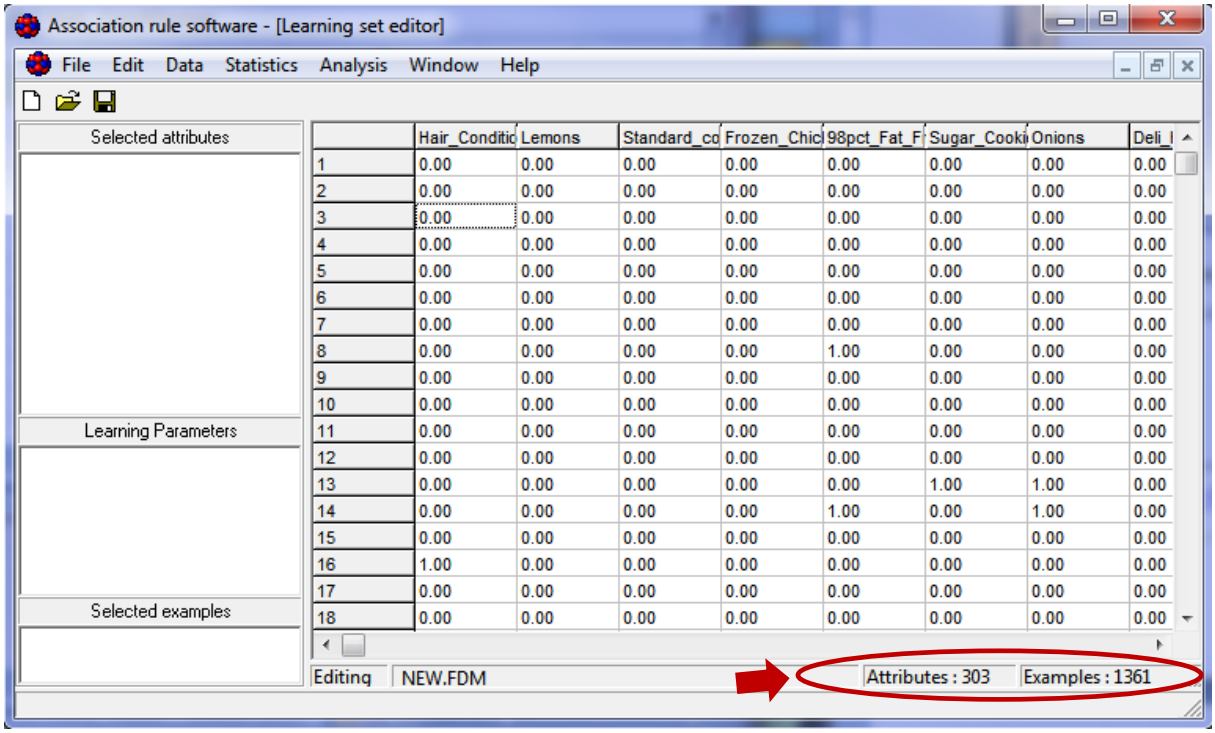
Remarque : Si vous ne disposez pas la version 2007, je vous enjoins à importer directement le fichier « market_basket.txt » au format texte avec séparateur tabulation qui accompagne ce document via le menu « FILE / OPEN / TEXT FILE FORMAT (*.txt) »[4].

Après avoir chargé le fichier dans Excel, nous sélectionnons la plage de données et nous activons le menu SIPINA / ASSOCIATION RULE. Une boîte de dialogue de confirmation apparaît. Nous vérifions les coordonnées et nous cliquons sur le bouton OK.



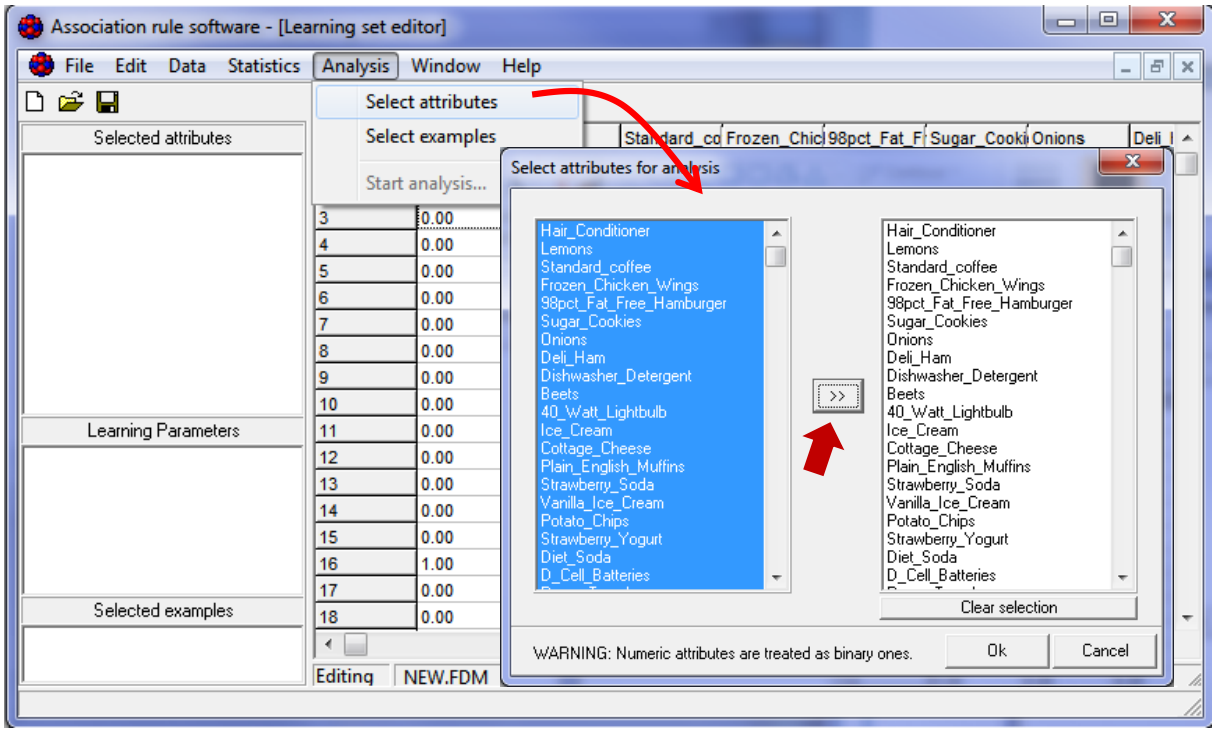
Le logiciel est automatiquement démarré et les données importées. Nous vérifions que nous avons bien n = 1361 lignes et p = 303 colonnes.

[4] Voir http://tutoriels-data-mining.blogspot.fr/2009/02/sipina-formats-de-fichiers.html



3.2 Choisir les items à traiter

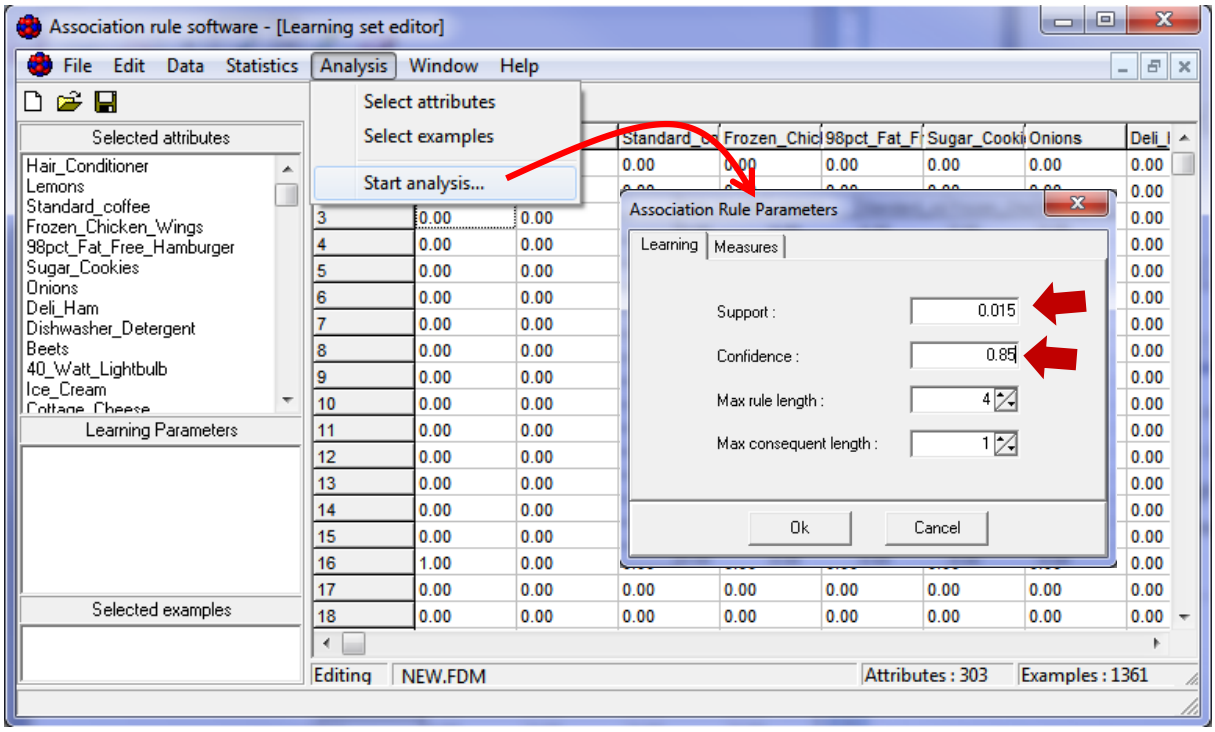
Pour spécifier les items à intégrer dans l'analyse, nous actionnons le menu ANALYSIS / SELECT ATTRIBUTES. Nous sélectionnons toutes les variables disponibles.



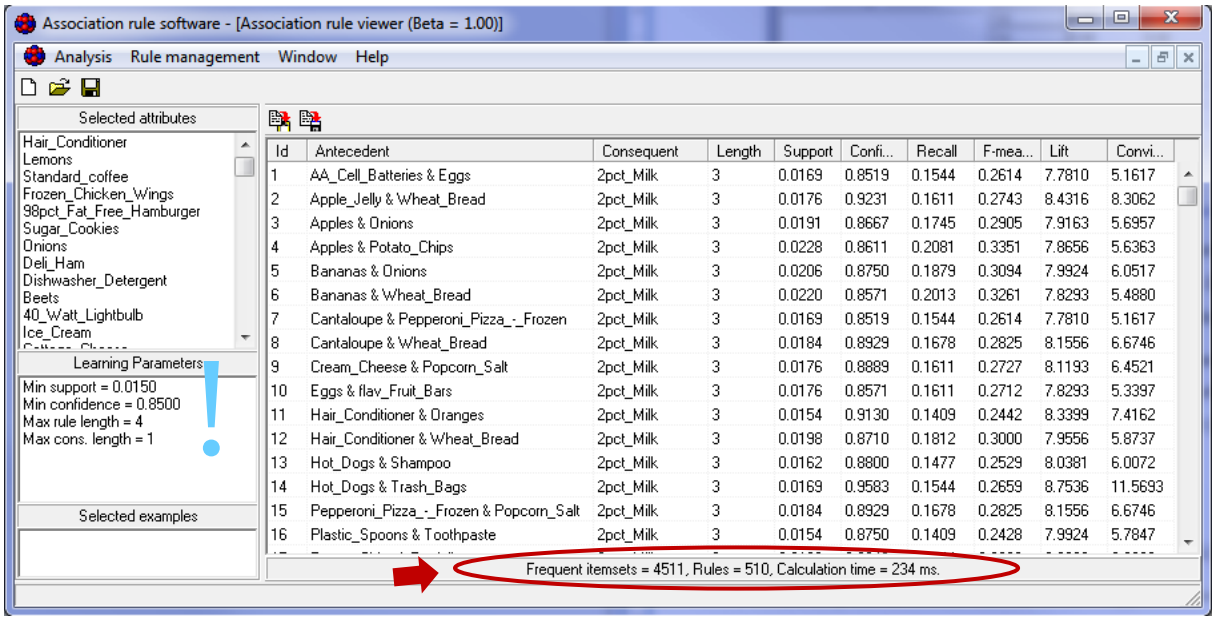
Nos variables étant numériques, ARS les prend comme des indicatrices : 0, absence de l'item dans le caddie ; 1 (ça peut être toute valeur > 0 en réalité), présence. Lorsque nous traitons des variables catégorielles, le logiciel procède automatiquement à un codage disjonctif complet en interne avant de lancer les traitements.

3.3 Paramétrage et extraction des règles

Pour lancer les calculs, nous actionnons le menu ANALYSIS / START ANALYSIS. Une boîte de paramétrage apparaît :



Nous fixons le support minimum à 0.015 (nous acceptons les règles qui couvrent au moins 1361 x 0.015 ≈ 20 individus) ; la confiance minimale à 0.85 ; la longueur maximale des règles à 4 items ; et nous acceptons 1 seul item dans le conséquent des règles. Nous validons ce choix en cliquant sur OK.



Une fenêtre de visualisation des règles apparaît. Chacune est décrite par son antécédent et son conséquent, elle est caractérisée par des indicateurs indiquant son intérêt (support, confiance, etc.)[5].

[5] Voir http://tutoriels-data-mining.blogspot.fr/2009/02/mesures-dinteret-des-regles-dans-priori.html

Nous avons **510 règles**. La première (n°1) s'écrit :

Si Achat de (AA_Cell_Batteries & Eggs) **Alors** Achat de (2pct_Milk)

Nous disposons, entres autres, de son support c.-à-d.

$$P(\text{AA\_Cell\_Batteries \& Eggs \& 2pct\_Milk}) = 0.0169$$

De sa confiance (précision) :

$$P(\text{2pct\_Milk} / \text{AA\_Cell\_Batteries \& Eggs}) = 0.8519$$

De son lift :

$$\frac{P(\text{2pct\_milk/AA\_Cell\_Batteries \& Eggs})}{P(\text{2pct\_milk})} = 7.7810$$

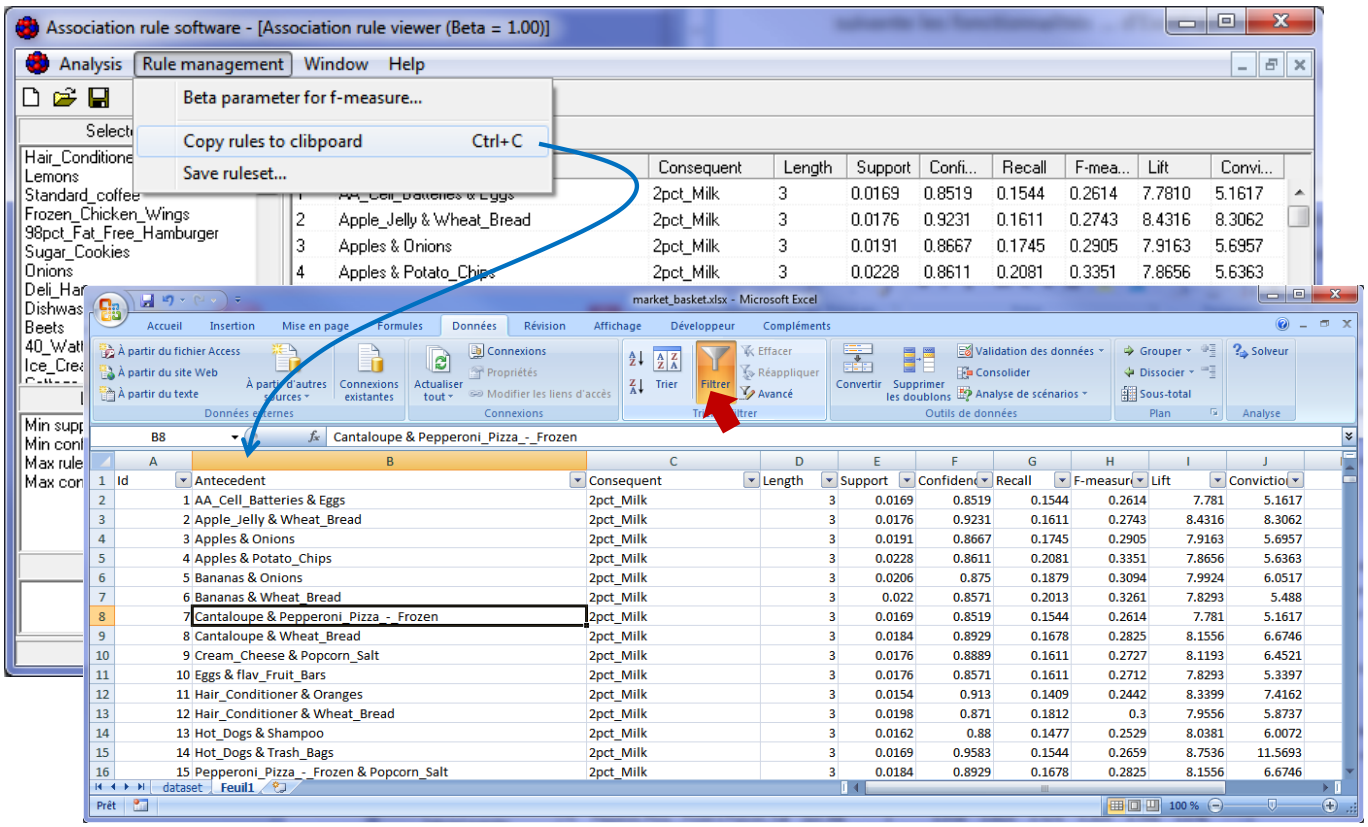
A vrai dire, rien ne distingue vraiment cet outil des composants d'extractions de règles d'association disponibles dans Tanagra. Mon idée dans ce tutoriel est surtout de mettre en avant dans la section suivante les fonctionnalités … d'Excel qui nous permettent de mieux explorer les résultats.

4 Exploration des règles sous Excel

4.1 Récupération des règles

Les algorithmes génèrent souvent un grand nombre de règles, 510 dans l'analyse qui nous concerne. Pour pouvoir les exploiter, il faut pouvoir les organiser à notre guise : les filtrer selon plusieurs critères, les trier selon les indicateurs numériques qui leur sont associés. L'outil FILTRE AUTOMATIQUE d'Excel convient parfaitement pour cela.

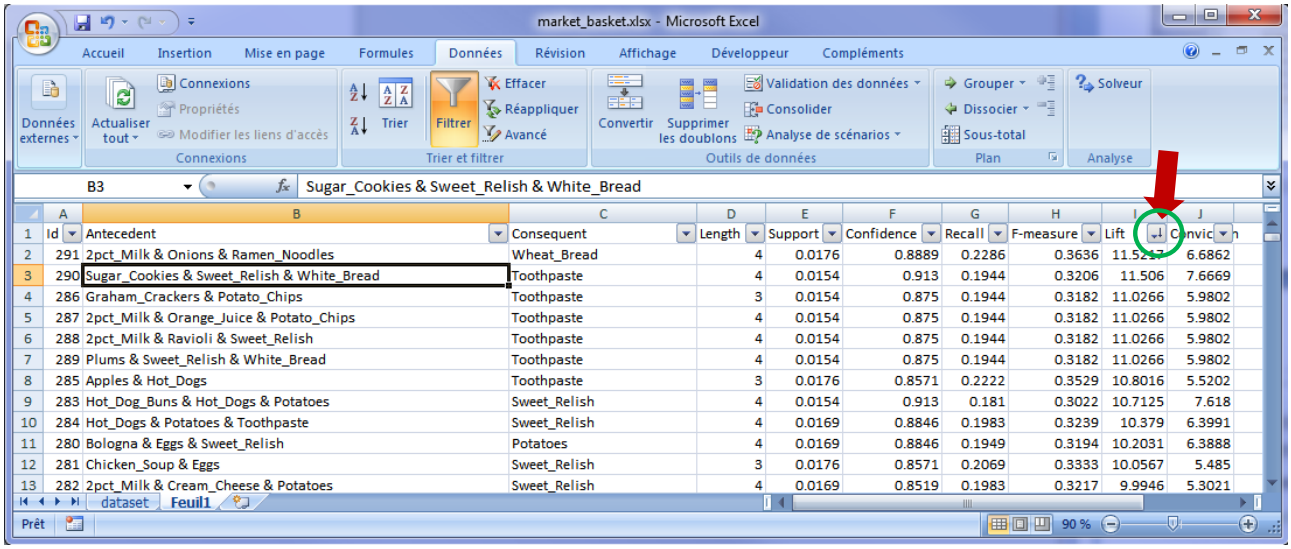
Dans ARS, nous actionnons le menu RULE MANAGEMENT / COPY RULES TO CLIPBOARD.



Nous créons une nouvelle feuille dans le classeur Excel. Nous y collons les règles et nous activons l'outil FILTRE AUTOMATIQUE (DONNEES / FILTRER dans Excel 2007). Des onglets permettant de manipuler les règles apparaissent dans chaque en-tête de colonne[6].

4.2 Trier les règles selon un critère numérique

Nous souhaitons trier les règles selon le critère LIFT. Nous cliquons sur l'onglet de la colonne correspondante, nous initions un « Tri du plus grand au plus petit ».

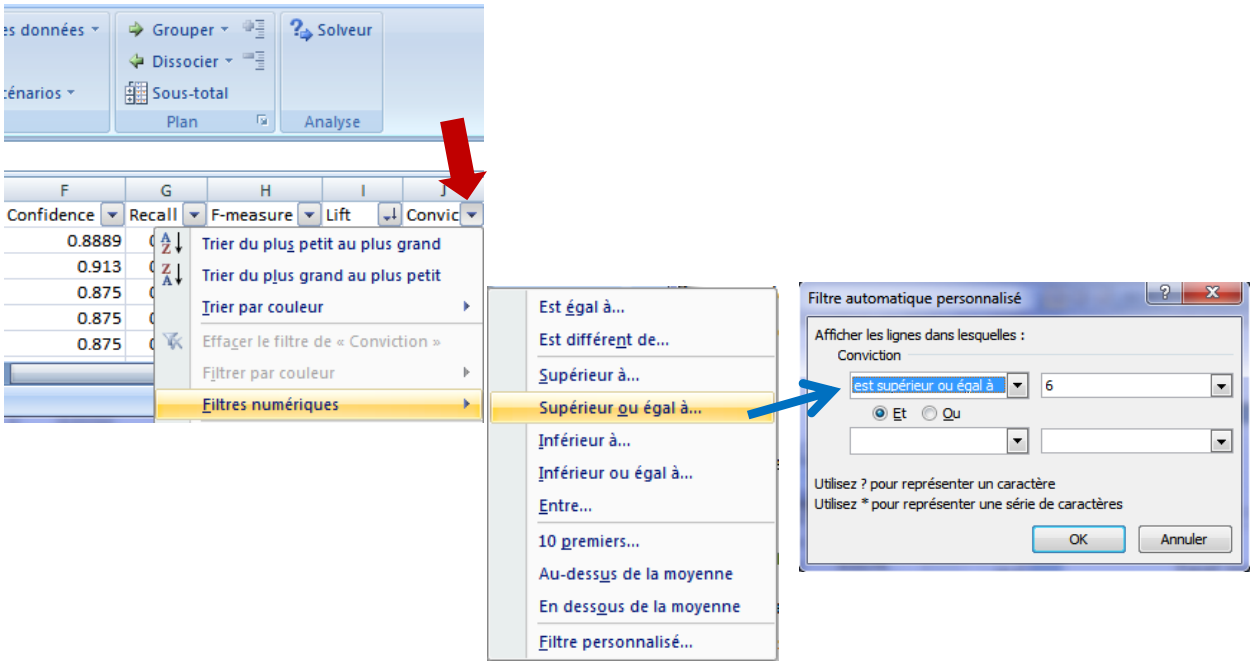


La règle avec le LIFT le plus élevé est :

Si Achat de (2pct_Milk & Onions & Ramen_Noodles) **Alors** Achat de (Wheat_Bread) [**LIFT = 11.5217**]

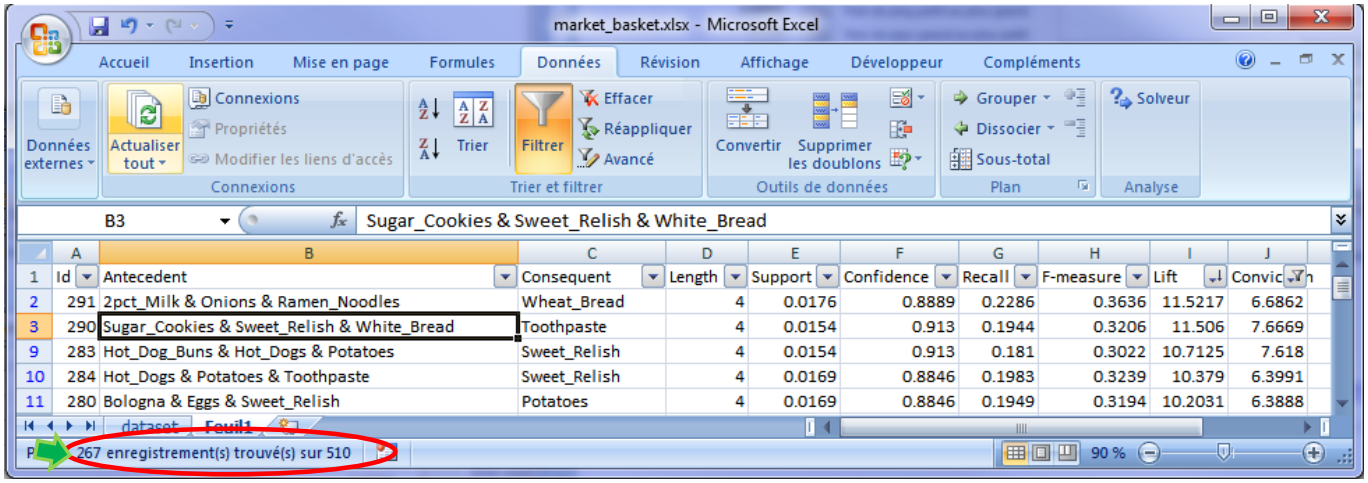
4.3 Filtrage sur un critère numérique

Nous souhaitons n'afficher que les règles avec une conviction > 6.



[6] Ces fonctionnalités sont également disponibles dans Libre Office (http://fr.libreoffice.org/).

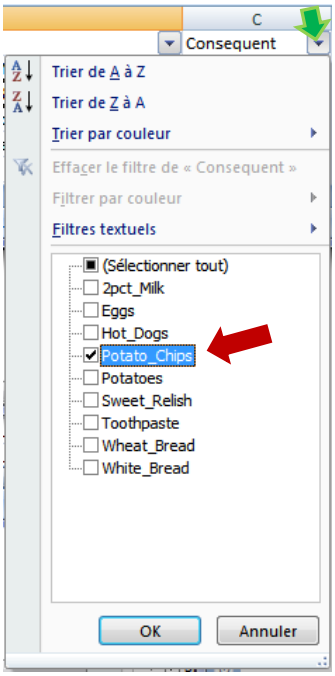
Pour ce faire, nous actionnons l'onglet correspondant et nous sélectionnons l'option « Filtre Numérique / Supérieur ou égal à… ». Nous spécifions la valeur seuil 6.



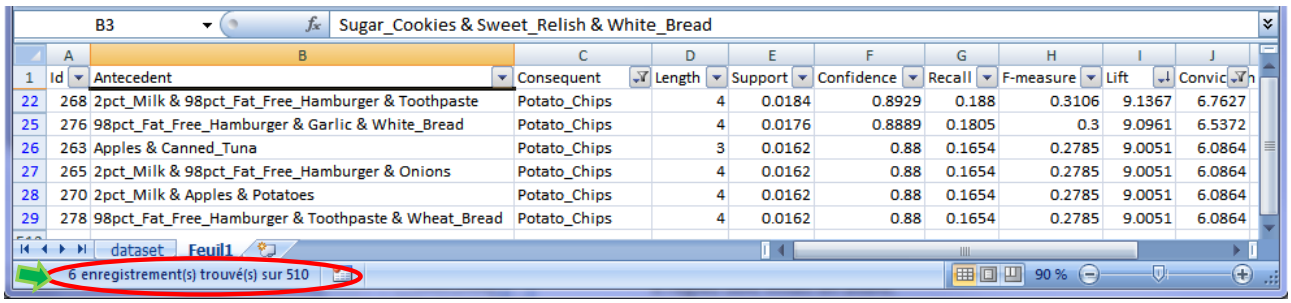
267 règles sur 510 répondent à cette spécification.

4.4 Filtrage sur le conséquent

Parmi ces règles, mettons que nous nous intéressons à celles qui mènent à la conclusion « Potato_Chips ». Nous actionnons l'onglet et nous choisissons l'item correspondant.

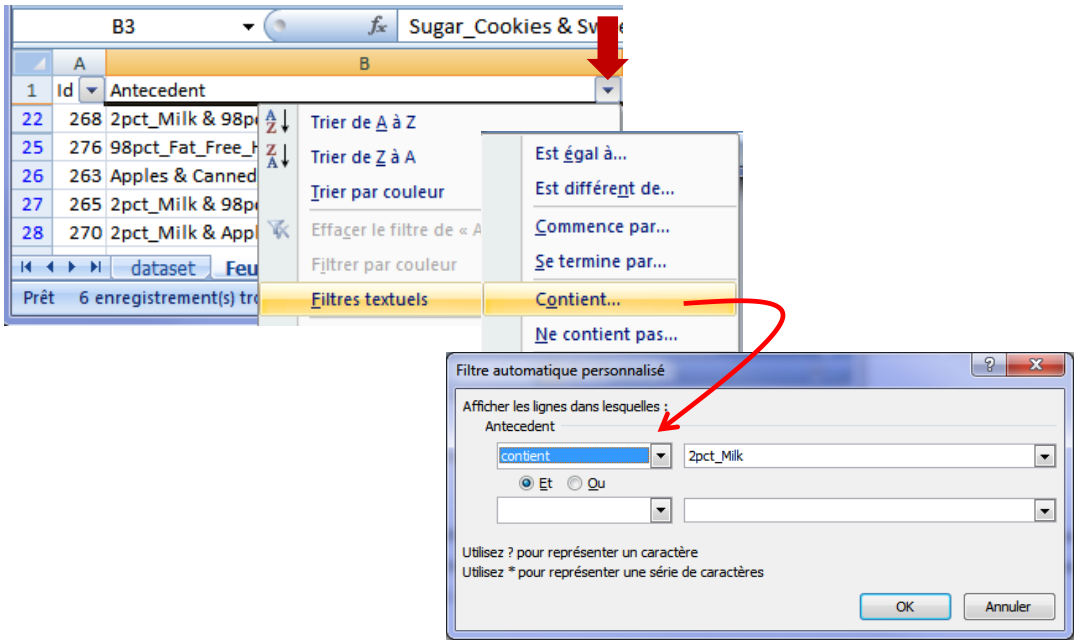


6 règles sont mises en avant.

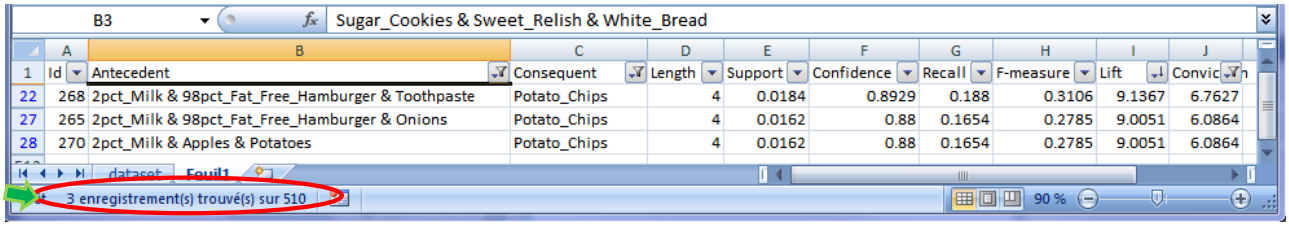


4.5 Filtrages sur l'antécédent

Parmi ces règles, nous souhaitons mettre en évidence celles dont l'antécédent comporte l'item « 2pct_Milk». Dans l'onglet de la colonne, nous activons le menu « Filtres textuels / Contient ». Nous saisissons la condition.



3 règles sont affichées maintenant.



Id	Antecedent	Consequent	Length	Support	Confidence	Recall	F-measure	Lift	Convic
268	2pct_Milk & 98pct_Fat_Free_Hamburger & Toothpaste	Potato_Chips	4	0.0184	0.8929	0.188	0.3106	9.1367	6.7627
265	2pct_Milk & 98pct_Fat_Free_Hamburger & Onions	Potato_Chips	4	0.0162	0.88	0.1654	0.2785	9.0051	6.0864
270	2pct_Milk & Apples & Potatoes	Potato_Chips	4	0.0162	0.88	0.1654	0.2785	9.0051	6.0864

Nous pouvons ainsi multiplier les combinaisons pour mettre en lumière les règles qui répondent au mieux au cahier des charges de notre étude. Certes, pour être tout à fait franc, ce dispositif n'est réellement opérationnel que si nous traitons des bases de règles de taille relativement modeste. Mais il présente le mérite de la simplicité d'utilisation. Tout le monde sait manipuler un tableur, qui est depuis longtemps, et pour longtemps encore, un des outils favoris des data miner (*Sondage KDnuggets de Mai 2012, « What Analytics, Data Mining, Big Data software you used in the past 12 months* ***for a real project*** *? » : Excel arrive en seconde position (après R). Il a toujours été dans le top du classement depuis que ce sondage annuel existe - http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html*).

5 Conclusion

ARS est avant tout d'un outil à vocation pédagogique destiné au traitement des petites bases. Nous montrons dans ce didacticiel que la possibilité d'interagir avec Excel lui ouvre des nouvelles possibilités en matière d'exploration des résultats. En effet, les outils de filtre et de tri d'Excel permettent de mieux analyser les règles générées par l'algorithme d'extraction.