Correction de l'examen final

le 23 novembre 201

1. Comment peut on traiter un aspect "profil acheteur" pour des transactions de type "achat" (panier de la ménagère) ?
   Il suffit de rajouter un ou des items liés au profils utilisateur ($Age = Jeune$, $Pet = Chien$) dans les transactions au même titre que les achats du panier.

2. Si on reste dans le domaine de la (très) grande distribution, comment pourrait on détecter des itemsets fréquents en ne disposant pas des moyens de calculs suffisants pour toute la gamme de produits ?
   On peut toujours recourir à des moyens externes de type cloud pour avoir plus de puissance de calculs et surtout de mémoire vive.
   Si on veut rester avec de moyens internes de calcul, on peut appliquer une version de Apriori Partition en découpant la base des transactions, par exemple, par sous gammes de produits (alimentaires, vêtements, produits pour bébé, etc).
   On pourrait faire aussi un découpage "vertical" : prendre des bases de transactions distinctes issues des créneau horraires significatifs.
   Dans le premier découpage on réduit le nombre de transaction et le nombre d'items, dans le second cas on réduit surtout le nombre de transaction à manipuler.

3. Proposer une méthode (originale ou adaptation des algorithmes connus) qui permettrait d'obtenir en temps raisonnable les règles d'association

à partie droite imposée.
Les solutions ne sont pas uniques.
Adaptation d'ECLAT : on calcule bien le support de tous les 1-itemsets (niveau 1 de la lattice). Lors du passage au niveau 2 on traite en premier et uniquement des branches de type S, X avec $S \in Sink$. Quant au X on prendra d'abord les $X \notin Sink$.

4. La base suivante traduit une série de transactions de type "panier".

Id	Transaction
T100	D, O, N, K, E, Y
T200	D, O, N , E, Y
T300	O, D, E
T400	K, E, Y
T500	D, O, N, E
T600	O, K, E, Y
T700	K, O, D, E

Soient les limites du support ($min\_support$) à 25% et de la confiance ($min\_confiance$) à 66%.
— Appliquant un des algorithmes vus en cours calculer tous les itemsets fréquents par rapport à $min\_support$
— A partir du résultat calculé au point précédent déterminer les associations ayant la forme $\lambda(2) \Rightarrow E, Y$ (la notation $\lambda(2)$ indique que la partie gauche a au moins deux éléments) qui sont au-delà de $min\_confiance$. Est-ce que ces règles sont intéressantes après le calcul de LIFT ?

Lors du calcul de 1-itesets on constate que support(E) = 7, alors on déroule l'exécution pour les autres items. Pas d'algo meilleur qu'un autre, toutefois parce que l'ensemble de transactions est petit j'au préféré appliquer Apriori TID (pour chaque itemset examiné on garde l'ensemble de TID qui le supporte et lors du calcul des jointure on fait l'intersection des deux ensemble de TID).
On obtient alors les itemsets fréquents suivants $\{D, N, K, Y, O, DN, Dk,$
$DY, DO, NY, NO, KY, KO, YO, DNY, DNO, DKO, DYO, NYO, KYO, DNYO\}$.
On rajoute E à tous et les itemsets fréquents sont $\{E, D, N, K, Y, O, DE, NE,$
$KE, YE, OE, DNE, DKE, DYE, DOE, NYE, NOE, KYE, KOE, YOE, DNYE,$
$DNOE, DKOE, DYOE, NYOE, KYOE, DNYOE\}$.

Les RA de la forme demandée sont : $DNO \to YE$, $DO \to YE$, $DN \to YE$, $NO \to YE$, $KO \to YE$.
On retient les RA suivantes :
$DNO \to YE$
$DN \to YE$
$NO \to YE$
$KO \to YE$
avec une confiance de $\frac{2}{3}$ *et on exclut* $DO \to YE$ *car sa confiance est de* $\frac{2}{5}$.
Pour les 4 RA retenues, la valeur $LIFT = \frac{2}{3} \times \frac{7}{4} = \frac{7}{6} > 1$

5. Si on vous donne le choix de la forme de présentation ou du codage des transactions avez-vous une préférence ? Justifiez votre réponse.
   Il peut être utile d'avoir les items avec des identifieurs munis d'un ordre total et les transactions présentées avec les items triés, tout itemset présenté lui aussi triés. D'un point de vue calcul, la réponse à la question T **supporte l'itemset** I *se donne en* $O(size(I) + size(T))$.
   D'autre part, avoir les transactions triées par leur longueur permettrait de ne plus explorer les transactions de petite longueur quand on vérifie des itemset de taille importante.