Université Sultan Moulay Slimane
Ecole National des Sciences Appliquées
Khouribga

Module : Data Mining & Texte Mining

1ère Année Master Big Data & Aide à la Décision
Semestre 2 / Année 2018/2019 / Feuille de TD N° 3
Corrigé des exercices 4-5-6-7-8

Exercice 4

Considérons les transactions d'un panier de la ménagère illustrées par le tableau suivant :

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

1. Quel est le nombre maximum de règles d'association qui peuvent être extraites à partir de ce dataset ? (y compris les règles de support nul)

   Réponse : Il y a 6 items dans le data set. En appliquant la formule de l'exercice 3, le nombre total des règles d'association possible est : $R = 3^6 - 2^{6+1} + 1 = 602$

2. Quelle est la taille maximale d'un itemset fréquent qui peut être extrait à partir de ce dataset ? (on suppose que ***min_sup*** > 0)

   Réponse : Comme la taille de la plus longue transaction est 4, la taille maximale d'un itemset fréquent est 4.

3. Quel est le nombre maximum des 3-itemsets qui peuvent être dérivés à partir de ce dataset ?

   Réponse : C'est le nombre de combinaison de 3 objets parmi 6.

$$\boxed{\binom{6}{3} = \frac{6!}{3!\times(6-3)!} = 20}$$

4. Trouver un itemset de taille ≥ 2 de support maximal.

   Réponse : {Bread, Butter}

5. Trouver une paire d'items a et b, telle que les règles $\{a\} \rightarrow \{b\}$ et $\{b\} \rightarrow \{a\}$ soient de même confiance.

   Réponse : (Bread, Butter) ou (Beer, Cookies)

Exercice 5

Dans une étape d'identification d'itemsets fréquents dans une base de données transactionnelle, nous avons trouvés que les 3-itemsets fréquents sont : $\{B, D, E\}$, $\{C, E, F\}$, $\{B, C, D\}$, $\{A, B, E\}$, $\{D, E, F\}$, $\{A, C, F\}$, $\{A, C, E\}$, $\{A, B, C\}$, $\{A, C, D\}$, $\{C, D, E\}$, $\{C, D, F\}$, $\{A, D, E\}$. Lesquels des 4-itemsets suivants peut être probablement fréquent ?

a) $\{A, B, C, D\}$: **NF** (contient $\{A, B, D\}$ qui est Non Fréquent)

b) $\{A, B, D, E\}$: **NF** (contient $\{A, B, D\}$ qui est Non Fréquent)

c) $\{A, C, E, F\}$: **NF** (contient $\{A, E, F\}$ qui est Non Fréquent)

d) $\{C, D, E, F\}$: **PF** (tous ses 3-sous- itemsets sont Fréquents)

NF : Non Fréquent

PF : Probablement Fréquent

Exercice 6

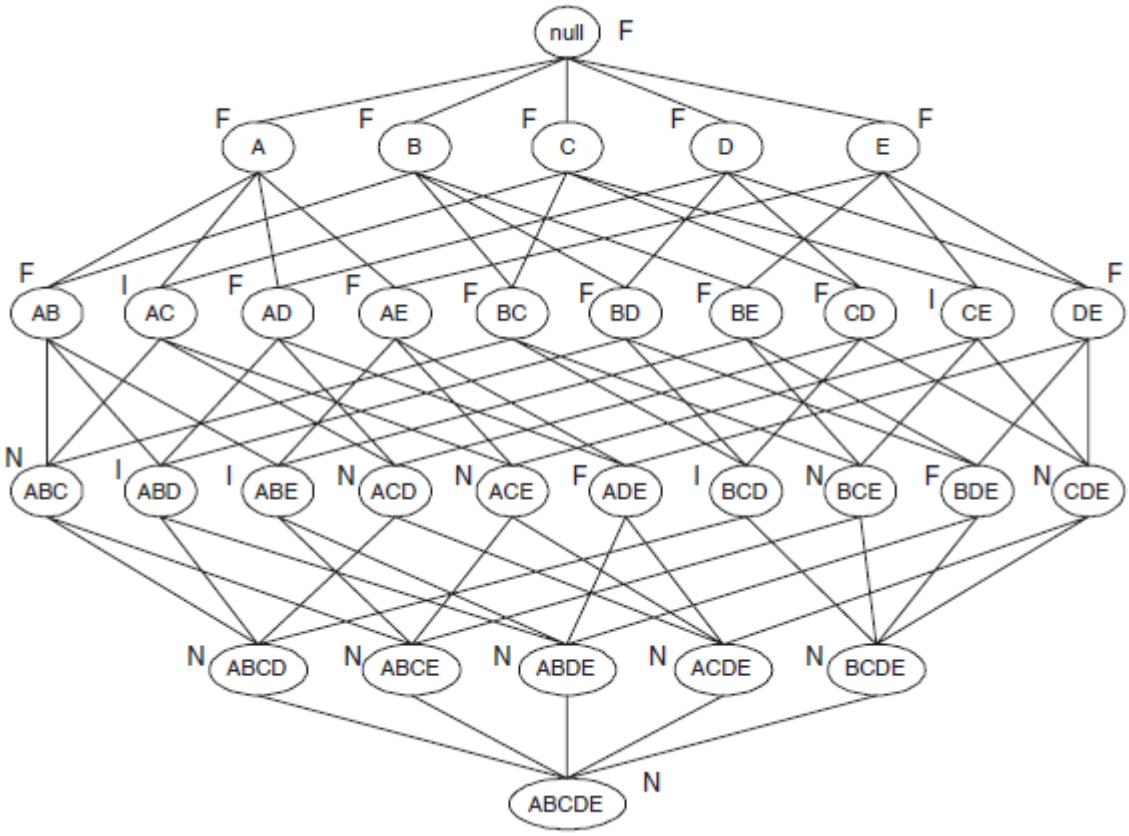
Supposons que l'algorithme Apriori est appliqué au dataset du tableau suivant :

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

On suppose que le seuil de support est ***min_sup*** = 30%.

1. Dessiner le treillis d'itemsets engendré par l'exécution de l'algorithme Apriori. Etiqueter chaque nœud du treillis avec les lettres suivantes :
   - **N** : si l'itemset est non considéré à être un itemset candidat par l'algorithme Apriori. Il y a deux raisons pour qu'un itemset soit non considéré comme un itemset candidat : (*i*) il est non généré pendant toutes les étapes de génération de candidats, ou bien (*ii*) il est généré pendant une étape de génération de candidats, mais il est supprimé pendant l'étape de l'élagage de candidats car l'un de ses sous-ensembles est non fréquent.
   - **F** : si l'itemset candidat est prouvé fréquent par l'algorithme Apriori.
   - **I** : si l'itemset candidat est prouvé non fréquent après le calcul de son support.

Réponse :



2. Quel est le pourcentage des itemsets fréquents ?

   Réponse : Le pourcentage des itemsets fréquents (incluant l'ensemble vide) est : $\frac{16}{32} = 50\%$

3. Quel est le rapport d'élagage de l'algorithme Apriori sur ce dataset ?

   Réponse : Le rapport d'élagage est le rapport du nombre des itemsets d'étiquettes N (les non considérés) par le nombre total des itemsets. C'est donc : $\frac{11}{32} = 34{,}4\%$

Exercice 7

Considérons, pour une règle d'association A → B, la mesure d'intérêt suivant :

$$M = \frac{P(B|A) - P(B)}{1 - P(B)}$$

1. Quel est l'intervalle de valeurs de la mesure M ? Préciser quand-est-ce que M atteint sa valeur maximale et sa valeur minimale.

   Réponse : L'intervalle des valeurs de la mesure M est [0, 1].

   - Pour la valeur maximale : $M = 1$, si et seulement si $P(B|A) = 1$.
   - Pour la valeur minimale : $M = 0$, si et seulement si $P(B|A) = P(B)$.

2. Qu'arrive-t-il à M lorsque :

   a. $P(A, B)$ croît alors que $P(A)$ et $P(B)$ restent inchangées ?

   Réponse :

   La mesure M peut être exprimée comme suit :

   $$M = \frac{P(A,B) - P(A)P(B)}{P(A)(1 - P(B))}$$

   $P(A, B)$ étant la probabilité conjointe de A et B. On voit alors que M augmente lorsque $P(A, B)$ croit.

   b. $P(A)$ croît alors que $P(A, B)$ et $P(B)$ restent inchangées ?

   Réponse :

   La mesure M décroit dans ce cas.

   c. $P(B)$ croît tandis que $P(A, B)$ et $P(A)$ restent inchangées ?

   Réponse :

   La mesure M diminue dans ce cas.

3. La mesure M est-elle symétrique sous permutation de variables ?

   Réponse : Non.

4. Que vaut M lorsque A et B sont statistiquement indépendants ?

   Réponse : M vaut 0, lorsque A et B sont indépendants, car dans ce cas $P(A,B) = P(A)P(B)$.

Exercice 8

Supposons que l'on dispose des données d'un panier de ménagère consistant de 100 transactions avec 20 items. Si le support d'un item a est de 25%, le support d'un item b est de 90% et le support de l'itemset $\{a, b\}$ est de 20%. Les seuils de support et de la confiance sont 10% et 60%, respectivement.

1. Calculer la confiance de la règle d'association $\{a\} \rightarrow \{b\}$. Est-elle intéressante selon la mesure de la confiance ?

   Réponse : $C(\{a\} \rightarrow \{b\}) = \frac{0,2}{0,25} = 80\%.$ ➔ La règle est donc intéressante, puisque sa confiance dépasse le seuil 60%.

2. Calculer la mesure de l'intérêt (***lift***) de la règle d'association {a} → $\{b\}$. Décrire la nature de relation entre l'item a et l'item b en termes de la mesure de l'intérêt.

   Réponse : $Lift(\{a\} \rightarrow \{b\}) = \frac{0,2}{0,25 \times 0,9} = 0.889.$ ➔ Les deux items a et b sont négativement corrélés selon la mesure de l'intérêt, puisque l'amélioration est < 1.

3. Quelle est la conclusion que l'on peut tirer à partir des résultats précédents ?

   Réponse : Une règle d'association ayant une confiance plus grande ne signifie pas qu'elle est intéressante.