USTHB,
Faculté d'Informatique
Département d'Intelligence Artificielle
Et des Sciences de données
M2, Master SII
Module: Data mining

Le 24 Octobre 2022

TD 3

Exercice 1

Considérer un ensemble de 5 documents D = {d1, d2, …, d5} et un ensemble de 11 termes T = {t1, t2, …, t11}, un document étant un ensemble de termes. La table suivante montre le contenu des documents de D :

document	termes
d1	{t1, t2, t3, t4, t5, t6}
d2	{t2, t3, t4, t5, t6, t7}
d3	{t1, t4, t5, t8}
d4	{t1, t4, t6, t9, t10}
d5	{t2, t4, t5, t10, t11}

Pour extraire les ensembles de termes fréquents avec un support minimal de 60% :

1) Appliquer l'algorithme Apriori sur D
2) Appliquer l'algorithme FP-Growth sur D
3) Appliquer l'Algorithme ECLAT sur D
4) Que peut-on en conclure ?

Exercice 2

Considérer la base suivante de cinq transactions. Supposer que *minsup* = 60% et *minconf* = 80%.

TID	**Articles achetés**
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

1) Ecrire l'algorithme FP-Growth étudié en cours. Quelle est sa complexité ?
2) Déterminer tous les itemsets fréquents en appliquant FP-Growth
3) Ecrire l'algorithme ECLAT étudié en cours. Quelle est sa complexité ?
4) Appliquer l'algorithme ECLAT sur la table des transactions ci-dessus.
5) Comparer l'efficacité des deux méthodes.

Exercice 3

Considérer un magasin avec trois points de vente géographiquement distribués sur Alger, Constantine et Oran. Chaque site a sa propre base de données, la base de données globale étant distribuée. Une transaction a le format T_j : $\{i_1, \dots, i_m\}$ où T_j est l'identifiant d'une transaction, et i_k ($1<=k<=m$) est l'identifiant d'un article acheté dans la transaction. Supposer que chaque site a la capacité d'extraire les motifs fréquents des transactions qui s'effectuent à son niveau et les envoie périodiquement à une machine centrale pour la gestion globale du magasin.

1) Proposer un algorithme efficace pour extraire les motifs fréquents globaux.
2) Proposer un algorithme efficace pour extraire les motifs fréquents spécifiques à chaque région.

Exercice 4

Considérer la base de connaissances 'vehicle' suivante :

***R*1: if** vehicleType = cycle **and** num_wheels = 2 **and** motor = no **then** vehicle = Bicycle

***R2*: if** vehicleType = cycle **and** num_wheels = 3 **and** motor=no **then** vehicle = Tricycle

***R*3: if** vehicleType = cycle **and** num_wheels =2 **and** motor = yes **then** vehicle = Motorcycle

***R*4: if** vehicleType = automobile **and** num_wheels = 2 **and** size = small **then** vehicle = SportsCar

***R*5: if** vehicleType = automobile **and** num_doors = 4 **and** size = medium **then** vehicle = Sedan

***R*6: if** vehicleType = automobile **and** num_doors = 3 **and** size = medium **then** vehicle = MiniVan

***R*7: if** vehicleType = automobile **and** num_doors = 4 **and** size = large **then** vehicle = Sports_Utility_Vehicle

***R*8: if** num_wheels = 4 **then** vehicleType = cycle

***R*9: if** num_wheels = 4 **and** motor = yes **then** vehicleType = automobile

1) Citer 3 algorithmes de détermination des motifs fréquents en évoquant leur principe et commentant leurs avantages et inconvénients.
2) Transformer la base de connaissances en une table d'entités lexicales. La première colonne indiquera le nom de la règle et la deuxième l'ensemble des entités lexicales appartenant à la règle. Ne pas considérer les mots en gras et le symbole '=' comme entités lexicales.
3) Donner la représentation verticale de cette table en inversant ses colonnes.
4) Rappeler l'algorithme Apriori et décrire chaque composant dans le détail.
5) Enumérer les motifs fréquents en appliquant l'algorithme Apriori avec un support minimal égal à 4. Des deux tables obtenues en 4) et en 5), quelle est celle qui convient le plus pour l'algorithme ? Pourquoi ?
6) Lister toutes les règles d'association (avec un support minimal égal à 44% (4/9) et une confiance minimale égale à 80%) correspondant à la métarègle suivante :

$\forall$ X = Ri (i = 1..9), *mention* (X, *item*1) ^ *mention* (X, *item*2) $\Longrightarrow$ *mention* (X, *item*3)

et interprétée comme suit:

Si item1 est mentionné dans X et si item2 est mentionné dans X alors item3 est mentionné dans X.

Exercice 5

Considérer le dataset suivant contenant 10 instances et 5 attributs nommés A, B, C, D et E. On s'intéresse à extraire des motifs fréquents pour déduire des règles d'association. Les instances font office de transactions et les valeurs des attributs d'items.

	A	B	C	D	E
I1	1	4	13	2	3
I2	1	2	12	0	7
I3	1	3	13	2	6
I4	1	4	11	2	7
I5	1	4	14	2	7
I6	0	4	15	2	7
I7	1	1	13	0	3
I8	1	4	14	0	7
I9	1	4	14	2	7
I10	1	4	12	2	7

1) Décrire avec clarté l'algorithme Apriori mais adapté à un dataset de format ci-dessus en précisant les points suivants :
    a. Les structures de données utilisées
    b. Les entrées- sorties de l'algorithme
    c. Les techniques algorithmiques
2) En déduire la complexité de l'algorithme.
3) Appliquer l'algorithme Apriori sur le dataset avec un support minimal de 40%.

Exercice 6

La base de données suivante a quatre transactions. Supposer que *minsup* = 60% et *minconf* = 80%.

client ID	*TID*	*Articles vendus* (de la forme *marque-article-catégorie*)
01	T100	{Pêcherie-Poisson, Hodna-lait, Soumam-fromage, Meilleur-Pain}
02	T200	{Meilleur-fromage, ONA-lait, Ferme-pomme, Cherchell-biscuit, Bon-pain}
01	T300	{Ouest-pomme, ONA-lait, Bon-pain, Cherchell-biscuit}
03	T400	{Bon-pain, Hodna-lait, ONA-fromage}

1) En considérant la granularité de la catégorie de l'article où un exemple d'item peut être lait, pour la règle suivante :

    $\forall X \in$ *transaction*, *achète*(X, *item*1) $\wedge$ achète(X, *item*2)) $\Longrightarrow$ achète(X, *item*3) [s, c]

    Enumérer les k-itemset fréquents pour la plus grande valeur de k, et toutes les règles d'association (avec leur support s et confiance c) contenant le k-itemset fréquent pour le plus grand k.
2) En considérant la granularité de la catégorie de la marque de l'article où un exemple d'item peut être Hodna-lait, pour la règle suivante :

    $\forall X \in$ client, *achète*(X, *item*1) $\wedge$ achète(X, *item*2)) $\Longrightarrow$ achète(X, *item*3) [s, c]

Enumérer les *k*-itemset fréquents pour la plus grande valeur de *k*.