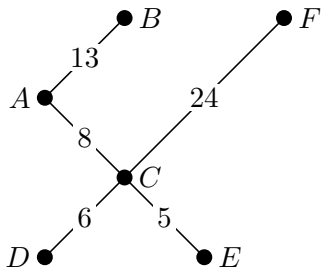
Data Mining et Data Warehousing, 3 juin 2011

Cahier fermé. Durée : 3 heures

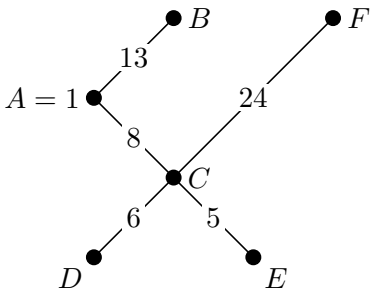
Nom et prénom
Année

Question 1 Prenons un graphe non orienté qui est acyclique et connexe, avec un entier positif sur chaque arête. La distance entre deux nœuds différents est la somme des entiers sur le chemin unique qui relie les deux nœuds. Par exemple, dans le graphe ci-dessous, la distance entre A et D est égale à $8 + 6 = 14$.



Donnez le numéro 1 au point A. Puis numérotez les autres points suivant le principe de *farthest-first traversal*.

. . . /5

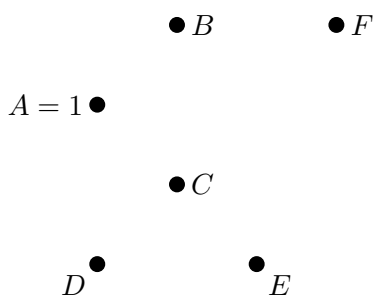


Question 2 Exécutez l'algorithme de Dasgupta et Long avec $\beta = 2$ sur les points numérotés dans la question 1. Remplissez les canevas suivants. Le coût est le **diamètre maximal**. Est-ce que l'on s'approche du "facteur 8" ?

. . . /10

R_2	=	
R_3	=	
R_4	=	
R_5	=	
R_6	=	
$lev(1)$	=	0
$lev(2)$	=	1
$lev(3)$	=	
$lev(4)$	=	
$lev(5)$	=	
$lev(6)$	=	
$\pi'(2)$	=	1
$\pi'(3)$	=	
$\pi'(4)$	=	
$\pi'(5)$	=	
$\pi'(6)$	=	

Dessinez π':



	Dasgupta et Long	Coût	**k-clustering optimal**	Coût
6-clustering	$\{A\},\{B\},\{C\},\{D\},\{E\}$	0	$\{A\},\{B\},\{C\},\{D\},\{E\}$	0
5-clustering				
4-clustering				
3-clustering			$\{A,B\},\{C,D,E\},\{F\}$	13
2-clustering				
1-clustering	$\{A,B,C,D,E,F\}$	45	$\{A,B,C,D,E,F\}$	45

Le facteur maximal observé est :

Situez chaque terme dans le cursus et expliquez de façon succincte mais précise.

Question 3 Leave-one-out.

…/5

Question 4 Fuzzy clustering.

…/5

Question 5 Ward's method.

. . . /5

Question 6 Gini index.

. . . /5

Question 7 À la page 356, les auteurs disent :

> *"We can use the closed frequent itemsets to determine the support counts of the non-closed frequent itemsets."*

Expliquez comment cela est possible. Utilisez un exemple concret et bien choisi pour illustrer votre explication.

. . . /10

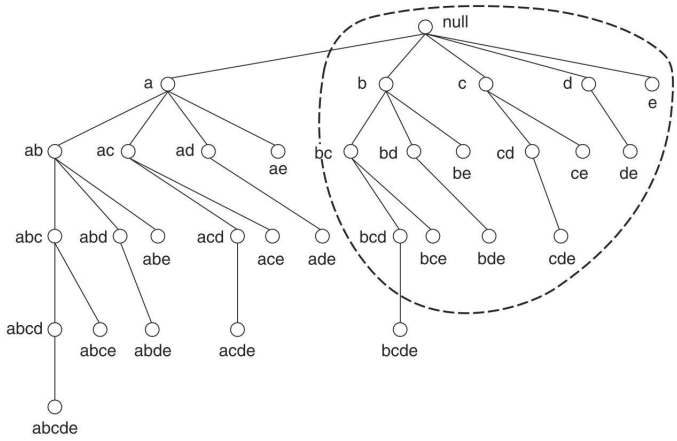


Figure 1: Generating candidate itemsets using the [X] approach.

Question 8 Voir la figure 1. Quel mot a été remplacé par [X] dans l'intitulé de cette figure ?

	.../1

Question 9 Voir la figure 1. Dans quel ordre les itemsets sont-ils traités dans l'approche [X] ? Cochez la case qui précède la phrase correcte.

	.../1

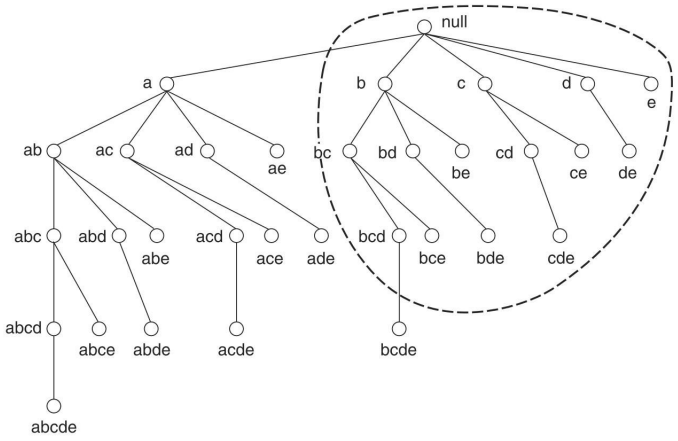
- ☐ Tous les 1-itemsets seront traités avant ab.
- ☐ bde sera traité avant c.

Question 10 Voir la figure 1. À la page 361, les auteurs disent :

> *"The* $\boxed{X}$ *approach is often used by algorithms designed to find maximal frequent itemsets. This approach allows the frequent itemset border to be detected more quickly than using a breadth-first approach. Once a maximal frequent itemset is found, substantial pruning can be performed on its subsets."*

Illustrez cette phrase de façon précise à l'aide de la figure 1.

. . . /4



Question 11 Voir la figure 1. À la page 362, les auteurs disent :

> *"The* $\boxed{X}$ *approach also allows a different kind of pruning based on the support of itemsets. For example, suppose the support for* $\{\mathsf{a}, \mathsf{b}, \mathsf{c}\}$ *is identical to the support for* $\{\mathsf{a}, \mathsf{b}\}$*. The subtrees rooted at* abd *and* abe *can be skipped because they are guaranteed not to have any maximal frequent itemsets. The proof of this is left as an exercise to the readers."*

Donnez la preuve qui est laissée aux lecteurs.

.../4

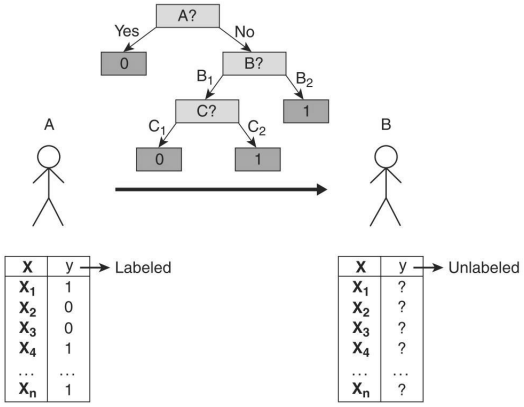


Figure 2: The [Y] principle.

Question 12 Voir la figure 2. Quel mot a été remplacé par [Y] dans l'intitulé de cette figure ?

.../1

Question 13 Expliquez la figure 2 de façon détaillée. Évitez des explications trop générales qui ne sont pas spécifiques pour la figure en question.

.../9

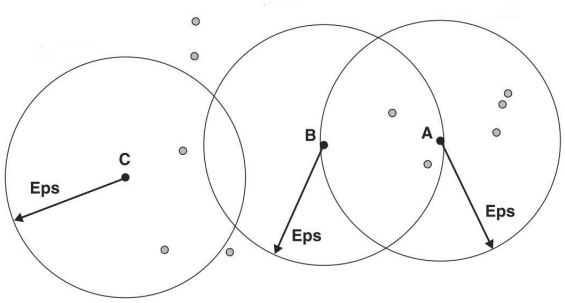


Figure 3: Z.

Question 14 Voir la figure 3. Pour quelle(s) valeur(s) de *MinPts* cette-image est-elle correcte ?

. . . /2

Question 15 Expliquez la figure 3 de façon détaillée. Évitez des explications trop générales qui ne sont pas spécifiques pour la figure en question.

. . . /8