Data Mining & Data Warehousing, 6 juin 2008

Nom et Prénom:..... Année d'études:.....

Cet examen contient 8 questions. Utiliser les 4 pages de réponse comme suit:

recto de la première page → réponse à la question 1
verso de la première page → réponse à la question 2
recto de la deuxième page → réponse à la question 3
verso de la deuxième page → réponse à la question 4
⋮
verso de la quatrième page → réponse à la question 8

1. L'algorithme *farthest-first traversal* numérote N points de 1 à N. Quelle est la complexité de cet algorithme en terme de N? Expliquez. .../3

2. À la page 380. .../2

   For asymmetric binary data, measures that do not remain invariant under the inversion operator are preferred.

   Qu'est-ce que *asymmetric binary data*? Donnez un exemple concret.

3. Exercice 2 (c) à la page 198. .../3

   Compute the Gini index for the `Gender` *attribute.*

4. La formule de Lance-Williams est utilisée dans le contexte du *hierarchical clustering* (voir la page 524): .../4

$$p(R, Q) = \alpha_A p(A, Q) + \alpha_B p(B, Q) + \beta p(A, B) + \gamma |p(A, Q) - p(B, Q)|$$

   (a) Cette formule a-t-elle une utilité pratique? Expliquez en détail.

   (b) À la page 525, on explique que les paramètres pour WPGMA sont $\alpha_A = \alpha_B = 1/2$ et $\beta = \gamma = 0$. Expliquez et discutez ces paramètres.

   (c) Dans l'article *"Performance guarantees for hierarchical clustering"*, les auteurs S. Dasgupta and Ph. M. Long proposent une nouvelle méthode de *hierarchical clustering*. Est-ce que la formule de Lance-Williams pourrait s'appliquer sur cette nouvelle méthode? Si oui, comment déterminer les paramètres $\alpha_A, \alpha_B, \beta, \gamma$?

5. Regardez la Figure 6.13 (b) *Number of frequent itemsets*, à la page 347. On pourrait penser que des *itemsets* fréquents de large taille sont plus rares. Néanmoins, la fonction pour le support = 0.1% atteint un maximum pour la taille = 10. Savez-vous expliquer ce comportement? .../3

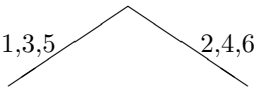
6. Exercice 12 à la page 562 présente l'algorithme du *leader*. .../4

The leader algorithm represents each cluster using a point, known as a leader*, and assigns each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold. In that case, the point becomes the leader of a new cluster.*

Comparez cet algorithme du *leader* avec l'algorithme K-means sur deux plans: (1) la qualité du clustering obtenu, et (2) la performance en temps d'exécution.

7. Exercice 10 à la page 408. Noter: . . . /4

   - Les candidats: $\{1,2,3\}, \{1,2,6\}, \{1,3,4\}, \{2,3,4\}, \{2,4,5\}, \{3,4,6\}, \{4,5,6\}$
   - La fonction de hachage:



   - Le nombre de candidats par nœud est ≤ 2.

8. Considérez la base de transactions suivante: . . . /2

tid	*items*
1	$\{a, b, c, d, e, f\}$
2	$\{a, b, c, d, e\}$
3	$\{a, d\}$
4	$\{b, d, f\}$
5	$\{a, b, c, e, f\}$

   (a) Montrez ces données dans un *FP-tree*, en utilisant un ordre dans lequel f est le dernier élément.

   (b) Dérivez le *conditional FP-tree for* f.