Professeur Michel de Rougemont

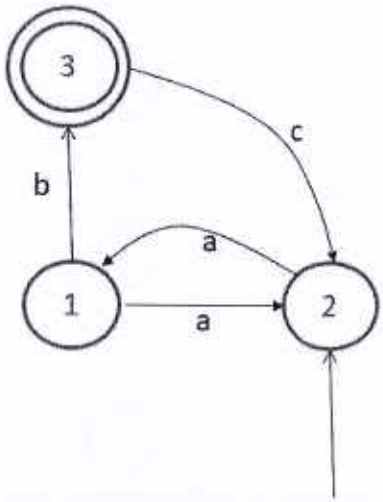
Informatique Décisionnelle - 5073
Session de Janvier 2019

Les cinq problèmes sont indépendants.

1. Problème de mots.

Soit l'automate A défini par Q={1,2,3} où 2 est l'état initial et 3 est l'état final. Les transitions sont définies par le graphe ci-dessous.

a. Trouver un mot de longueur au moins 8, accepté et un autre rejeté.
b. Décrire une expression régulière pour les mots acceptés par l'automate, c'est-à-dire partant de l'état 2 et finissant dans l'état 3.
c. Préciser si l'automate est déterministe ou non déterministe.
d. Pour la distance d'édition, quelle est la distance du mot abcabcabc au langage de l'automate. Trouver un automate distinct de A, dont le langage soit distinct de celui de A, et ε-proche de A.
e. Pour deux automates A_1 et A_2 expliquer comment classifier un mot w pour les 2 classes de mots définies par ces automates.



2. **Problème XML.**

Soit la DTD suivante :

```
<?xml version='1.0' ?>
<!ELEMENT a (a+b*)>
<!ELEMENT b (b+c)>
<!ELEMENT c (c)>
<!ELEMENT c (#PCDATA)>
```

Cette DTD peut aussi s'écrire symboliquement (Λ est le mot vide)

a : a
a : b*
b : b
b : c
c : c
c : Λ

- Trouver un arbre valide pour cette DTD avec au moins 8 nœuds et le fichier XML correspondant . Quel est le langage d'arbres défini par cette DTD.

- Pour la distance d'édition, quelle est la distance de l'arbre ci-dessous à cette DTD. Trouver un arbre valide pour cette DTD le plus proche.

- Trouver une DTD_1, qui soit ε-proche de cette DTD et dont le langage d'arbres soit différent, et une DTD_2 qui soit ε-loin de cette DTD pour la distance d'édition.

3. Problème de Schéma relationnel.

Soit le Schéma relationnel composé de 3 tables décrivant l'inscription d'étudiants dans une Université.

Students

ID	Gender	Age	City	Country

Diploms

DID	Name	Level	Area

Registration

ID	DID	Year	Mention	Grade

ID est une clé pour la table **Students,** et DID est une clé pour la table **Diploms.**
Par exemple (10, M, 21, New-York, USA) est un tuple de la table **Students.**
(100, Master Finance, M2, Economy) est un tuple de la table **Diploms.**
(10,100, 2017, Excellent, A+) est un tuple de la table **Registration.** Les notes (Grades) possibles sont A,B,C,D,E,F avec + ou – pour chaque note et les Mentions possibles sont Average, Good, Very Good et Excellent. Les niveaux (Level) sont L,M1,M2,D (Doctorat).

a. Décrire un schéma Entité-relations associé à ce Schéma relationnel.
b. Ecrire des requêtes SQL pour trouver :
    1. Combien d'étudiants étaient inscrits dans le « Master Finance » en 2016 ?
    2. Quels sont les pays d'origine des étudiants du « Master Finance » ?
    3. Les années (Year) où il y a eu des diplômés dans le « Master finance « ?
c. Quelle est la clé de la table **Registration** ?
d. Indiquer toutes les dépendances fonctionnelles de ce schéma ?

4. Problème OLAP.

On souhaite analyser la table **Registration** du schéma relationnel précédent, considérée comme un entrepôt de données pour connaitre la répartition des étudiants par genre, ville, pays, type de diplôme et par discipline.

1. Proposez un schéma étoile pour la table **Registration**. On pourra prendre ID comme mesure et la fonction d'aggrégation *count* pour compter le nombre de tuples.

2. Donnez les expressions OLAP (chemin, mesure, agrégation) associées aux trois requêtes :

   a. Analyse du nombre de diplômés par genre (Gender) en 2017,

   b. Analyse du nombre de diplômés par genre et par pays,

   c. Analyse du nombre de diplômés par discipline (Area) et par niveau (Level).

3. Donnez des expressions SQL-étendu pour la requête (c).

4. Comment analyser la note moyenne par discipline (area) ?

5. Fouille de Données

On souhaite prédire la mention d'un étudiant à partir de 1000 tuples de la table Registration, en analysant les attributs Gender, Age, Country des étudiants et Level, Area des Diplômes.

On sépare les 1000 tuples en deux sous-ensembles R_1 et R_2 de 500 tuples chacun.

a. Expliquez comment construire un arbre de décision à partir de R_1. Transformation de la table, puis sélection des attributs les plus significatifs.

b. Donner un exemple d'un arbre de décision dans ce contexte.

c. Comment est définie la qualité de la prédiction de cet arbre ? Comment la vérifier ?

d. Même question si on cherche à prédire la note (grade) avec une régression logistique.