Professeur Michel de Rougemont

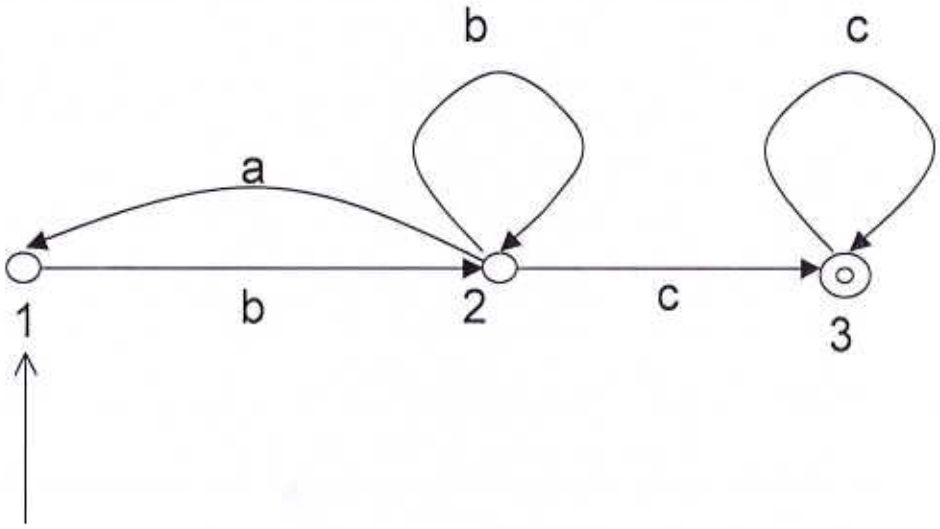
Informatique Décisionnelle
Session de Septembre 2019
4 pages

Aucun document n'est autorisé. Les cinq problèmes sont indépendants.

1. Problème de mots.

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

a. Trouver un mot de longueur au moins 8, accepté et un autre rejeté.
b. Décrire le langage défini par cet automate par une expression régulière?
c. Préciser si l'automate est déterministe ou non déterministe.
d. Pour la distance d'édition, quelle est la distance du mot ababbbbccb au langage de l'automate.
e. Trouver un automate B, dont le langage soit ε-loin de A.
f. Etant donné un mot w, expliquer comment le classifier comme un mot de A ou comme un mot de B, selon la distance d'édition.



2. Problème XML.

Soit le fichier XML suivant :

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<LIBRARY>
        <BOOK ISBN="9" LANG="fr">
                <AUTHOR>
                        <FIRSTNAME>Jean-Christophe</FIRSTNAME>
                        <LASTNAME>Bernadac</LASTNAME>
                </AUTHOR>
                <AUTHOR>
                        <FIRSTNAME>François</FIRSTNAME>
                        <LASTNAME>Knab</LASTNAME>
                </AUTHOR>
                <TITLE>Construire une application XML</TITLE>
                <PUBLISHER>
                        <NAME>Eyrolles</NAME>
                        <PLACE>Paris</PLACE>
                </PUBLISHER>
                <DATEPUB>1999</DATEPUB>
        </BOOK>
        <BOOK ISBN="7" LANG="fr">
                <AUTHOR>
                        <FIRSTNAME>Alain</FIRSTNAME>
                        <LASTNAME>Michard</LASTNAME>
                </AUTHOR>
                <TITLE>XML, Langage et Applications</TITLE>
                <PUBLISHER>
                        <NAME>Eyrolles</NAME>
                        <PLACE>Paris</PLACE>
                </PUBLISHER>
                <DATEPUB>1998</DATEPUB>
        </BOOK>
        <BOOK ISBN="5" LANG="fr">
                <AUTHOR>
                        <FIRSTNAME>William J.</FIRSTNAME>
                        <LASTNAME>Pardi</LASTNAME>
                </AUTHOR>
                <TRANSLATOR PREFIX="adapté de l'anglais par">
                        <FIRSTNAME>James</FIRSTNAME>
                        <LASTNAME>Guerin</LASTNAME>
                </TRANSLATOR>
                <TITLE>XML en Action</TITLE>
                <PUBLISHER>
                        <NAME>Microsoft Press</NAME>
                        <PLACE>Paris</PLACE>
                </PUBLISHER>
                <DATEPUB>1999</DATEPUB>
        </BOOK>
</LIBRARY>
```

a. Décrire ce fichier par un arbre XML, en abréviant le nom des tags, par leur première lettre, sans écrire les attributs.
   `<BOOK ISBN="9" LANG="fr">` sera considéré comme la balise <b>
b. Trouver une DTD pour laquelle cet arbre est valide.
c. Quel est le langage d'arbres défini par cette DTD.
d. Trouver un arbre proche, mais non valide, pour cette DTD.
e. Trouver une DTD_1, qui soit ε-loin de cette DTD pour la distance d'édition.

3. Problème de Schéma relationnel.

La société LITCH a créé une plateforme (site web + application Iphone) pour transporter des passagers en milieu urbain. Le cœur de l'application est une base de données relationnelles, qui en simplifiant se résume à 3 tables :

Les tables **VEHICULES** et **CLIENTS** ont les attributs décrits par les schémas ci-dessous.
VEHICULES(NumV, Marque, Année) décrit des véhicules avec un identifiant (NumV), une Marque et l'année du véhicule (Année).
CLIENTS(NumC, Nom, Adresse, Ville, Téléphone) décrit des clients avec un identifiant (NumC), un Nom, une Adresse, une Ville et un numéro de Téléphone.

VEHICULES

NumV	Marque	Année

CLIENTS

NumC	Nom	Adresse	Ville	Téléphone

La table **TRAFIC** décrit les différents trajets effectués par les clients avec les véhicules. Les attributs de la table **TRAFIC** décrivent un client (*NumC*), un véhicule (*NumV*) qui part d'un *Lieu de Départ*, à une certaine Date , pour arriver à un *Lieu Arrivée*, à une *Heure d'Arrivée Prévue,* avec peut-être un *Retard*, pour un *Prix* donné.

TRAFIC

NumC	NumV	Lieu Départ	Date	Lieu Arrivée	Heure Arrivée Prévue	Retard	Prix

a. Décrire un schéma Entité-relations associé à ce Schéma relationnel.
b. Ecrire des requêtes SQL pour trouver :
    1. Les Noms des Clients qui sont partis de la « rue d'Assas »,
    2. Les Marques de Véhicules qui ont desservi l'aéroport « Orly »,
    3. Le nombre de trajets vers « CDG » en « 2018 »
c. Indiquer les dépendances fonctionnelles des tables **TRAFIC et VEHICULES.**
d. Quelle est la clé de la table **TRAFIC** ?

4. Problème OLAP.

On souhaite analyser la table **TRAFIC** du schéma relationnel précédent, considérée comme un entrepôt de données pour comprendre la répartition des trajets et des ventes.

a. Proposez un schéma étoile pour la table **TRAFIC**, avec *Prix* comme mesure et Somme comme Aggrégation.

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

- Analyse des ventes par *Ville* des clients en 2018
- Analyse des ventes vers l'aéroport CDG par *Marque* des Véhicules,
- Analyse des ventes en Avril 2018, par *Ville* des Clients et par *Jour*.

c. Donnez une expression en SQL-étendu (utilisant GROUP BY) pour la 1ère requête.

5. **Fouille de Données**

L'entrepôt de données est alimenté par les données de la société LITCH, selon le schéma du problème 3. On cherche à prédire le retard des trajets, à partir de N=1000 enregistrements de cette table.

a. Dans un premier temps, on discrétise la variable *Retard* à l'aide d'une nouvelle variable R : R=0 si *Retard<5mins, R=1 si 5mins< Retard<15mins, et R=2 si Retard>15mins.*

Expliquer comment construire à partir de la table **TRAFIC** une nouvelle table **TRAFIC1,** où on détaille la date en deux attributs : Jour (Lundi, Mardi, Mercredi, Jeudi, Vendredi, Samedi, Dimanche) et Heure (0h, 1h,....23h). On remplace aussi la valeur *Retard* par *R*. Chaque tuple de **TRAFIC** génère un tuple de **TRAFIC1 :**

TRAFIC1

NumV	Lieu Départ	Jour	Heure	Lieu Arrivée	Heure Arrivée Prévue	R	Prix

b. Expliquer comment construire un arbre de décision pour la table **TRAFIC1,** en utilisant le critère du Gain d'Information. Rappeler la définition de l'erreur.

c. Donner un exemple d'arbre de décision plausible dans ce contexte. Expliquer comment on teste l'erreur du modèle. Donner un exemple de courbe de Lift dans ce cadre.

d. Expliquer comment trouver des régressions logistiques pour prédire R.

e. Comment pourrait-on prédire la variable *Retard* directement, sans la discrétiser ?