Professeur Michel de Rougemont

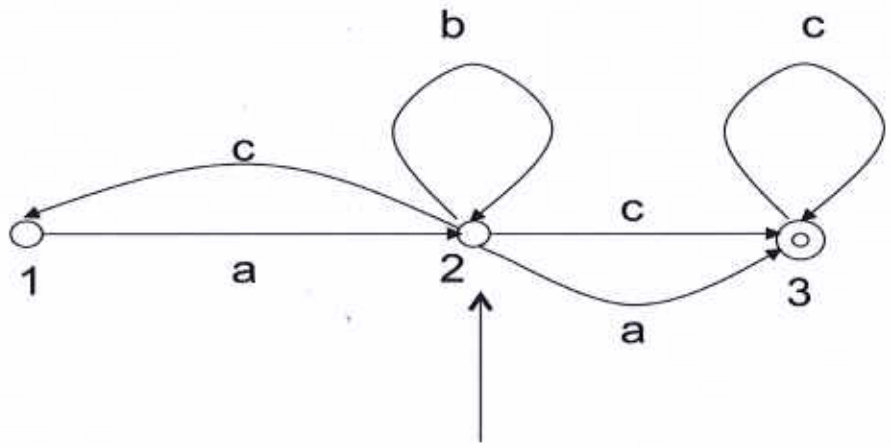
Informatique Décisionnelle
Session de Septembre 2017
4 pages

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 2 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 cabcbcbc 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 d'arbres.

Soit le fichier XML suivant :

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<LIBRARY>
        <BOOK ISBN="7" LANG="fr">
                <AUTHOR>
                        <FIRSTNAME>Alain</FIRSTNAME>
                        <LASTNAME>Michard</LASTNAME>
                </AUTHOR>
                <AUTHOR>
                        <FIRSTNAME>Al</FIRSTNAME>
                        <LASTNAME>Richard</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.

ADP (Aéroport de Paris) maintient un système d'information. Le cœur de l'application (simplifiée) est une base de données relationnelle, qui se résume à 3 tables :

Les tables **VOLS-départ, AVIONS, COMPAGNIES** ont les attributs décrits par le schéma ci-dessous.

VOLS-départ(NumV, Date, Heure, Ville, Compagnie, Type, Retard) décrit les Vols au départ de Paris, avec un identifiant (NumV), une date (jour.mois), une Ville de destination, un Identifiant pour une Compagnie, un type d'avion et un retard possible.
AVIONS(Type, Capacité) décrit la Capacité (nombre de passagers) pour chaque type d'avion.
COMPAGNIES(Id_C, Nom, Pays) décrit pour chaque identifiant Id_C le Nom de la compagnie et le Pays d'origine.

VOLS-départ

NumV	Date	Heure	Ville	Compagnie	Type	Retard

AVIONS

Type	Capacité

COMPAGNIES

Id_C	Nom	Pays

Par exemple, le tuple (AF101, 15.09, 18h, New-York, 1, A330, 1h) décrit un tuple de la table **VOLS-départ**, (A330, 120) un tuple de la table **AVIONS** et (1 , Air France, France) un tuple de la table **COMPAGNIES.**

a. Décrire un schéma Entité-relations associé à ce Schéma relationnel.
b. Ecrire des requêtes SQL pour trouver :
    1. Quels sont les Vols à destination de New-York?
    2. Quels sont les noms des Compagnies qui ont des vols le 17.06 ?
    3. Quelle est la capacité totale (en passagers) des vols à destination de New-York le 17.06 ?
c. Indiquer les dépendances fonctionnelles des tables **AVIONS et VOLS-départ.**
d. Quelle est la clé de la table **VOLS-départ**?

4. Problème OLAP.

On souhaite analyser la table **VOLS-départ** du schéma relationnel précédent, considérée comme un entrepôt de données pour comprendre les retards des vols.

a. Proposez un schéma étoile pour la table **VOLS-départ** , avec *Retard* comme mesure et Somme comme Aggrégation.

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

- Analyse des Retards par Compagnie en Juin (06),
- Analyse des Retards par Type d'avions,
- Analyse des Retards par Ville et par Mois.

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

5. **Fouille de Données**

L'entrepôt de données (la table **VOLS-départ** du problème 4) est alimenté chaque jour par les données de ADP, selon le schéma du problème 3. On cherche à prédire les retards, à partir de N=1000 enregistrements de cette table.

a. 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.* Donner un exemple d'un arbre de décision pour prédire le Retard à partir de N tuples de la table **VOLS-départ**. Expliquer comment construire un tel arbre en utilisant le critère du Gain d'Information.

b. Expliquer comment l'erreur de prédiction peut être estimée.

c. Décrire comment utiliser une régression logistique pour prédire le Retard, comme à la question a.