Professeur Michel de Rougemont

Informatique Décisionnelle (5073)

Session de Janvier 2017

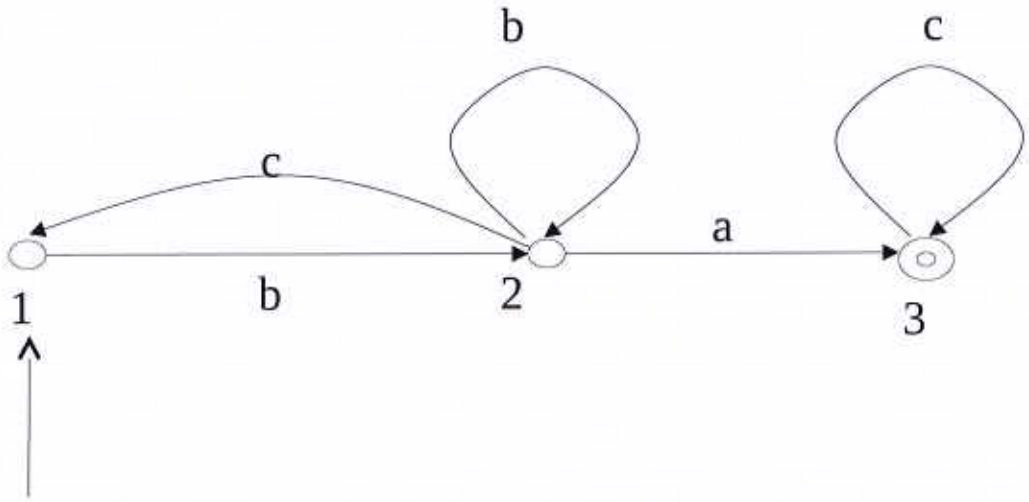
4 pages

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 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 cbcbcbcbbccb 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 avec déplacement.



2. Problème XML.

Soit le fichier XML suivant :

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<LIBRARY>
        <BOOK ISBN="9" LANG="fr">
                <AUTHOR>
                        <FIRSTNAME>Kim</FIRSTNAME>
                        <LASTNAME>Sterelny</LASTNAME>
                </AUTHOR>
                <TITLE>Dawkins vs. Gould</TITLE>
                <PUBLISHER>
                        <NAME>Icon Books</NAME>
                </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.**

Soit une base de données relationnelle, qui se résume à 4 tables :

Les tables **TRAINS, STATIONS, RESEAU** ont les attributs décrits par les schémas ci-dessous.

- **TRAINS**(NumT, Marque, Capacité, Lieu, Année) décrit les trains avec un identifiant (NumT), une Marque (le constructeur), un nombre de passagers (Capacité), un Dépôt (Lieu), et l'année de mise ne service du véhicule (Année).
- **STATIONS**(NumS, Nom, Adresse, Arrondissement) décrit une station de métro avec un identifiant (NumS), un Nom, une Adresse, un Arrondissement.
- La table **RESEAU** (Départ, Arrivée, Ligne) décrit les segments de chaque ligne de métro. Si une ligne décrit un chemin sur 20 stations, il y a 19 segments et donc 19 tuples. Par exemple (Odéon, St-Michel, 4) décrit un tuple qui décrit un segment de la ligne 4

TRAINS

NumT	Marque	Capacité	Lieu	Année

STATIONS

NumS	Nom	Adresse	Arrondissement

RESEAU

Départ	Arrivée	Ligne

La table **TRAFIC** décrit les différents trajets effectués par les trains. Les attributs de la table **TRAFIC** décrivent le trajet d'un train (*NumT*), qui part à une *Heure Départ*, d'une *Station Départ*, à une certaine *Date* , pour arriver à un *Lieu Arrivée*, à une *Heure d'Arrivée Prévue*, sur une *Ligne* du métro, avec peut-être un *Retard*, et/ou un *Incident* (0 s'il n'y a pas s'incident et 1 s'il y a un incident).

TRAFIC

NumT	Heure Départ	Station Départ	Date	Lieu Arrivée	Heure Arrivée Prévue	Ligne	Retard	Incident

a. Décrire un schéma Entité-relation associé à ce Schéma relationnel.
b. Ecrire des requêtes SQL pour trouver :
    1. Dans quel arrondissement se trouve la station « Odéon » ?
    2. Combien de lignes (distinctes) desservent la station « Odéon » ?
    3. Combien de trains ont circulé le 30 septembre 2015 sur la ligne 4 ?

c. Existe-t-il une requête SQL avec un attribut NumS qui donnerait toutes les stations accessibles depuis la station « Odéon » sur la ligne 4 ?
d. Indiquer les dépendances fonctionnelles des tables **TRAFIC et TRAINS.**
e. 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 du trafic, des retards et des incidents.

a. Proposez un schéma OLAP (Schéma étoile) pour la table **TRAFIC**, avec *Retards et Incidents* comme mesures (2 mesures) et Somme comme Agrégation.

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

- Analyse des Incidents par *Lignes* en 2006
- Analyse des Incidents par *Marque* des Trains,
- Analyse des Retards en Avril 2014, par *Ligne* 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 (la table TRAFIC du problème 4) est alimenté par les données de la RATP, selon le schéma du problème 3. On cherche à prédire les incidents, à partir de N=1000 enregistrements de cette table.

a. Expliquer comment construire un arbre de décision pour la table **TRAFIC,** en utilisant le critère du Gain d'Information. Rappeler la définition de l'erreur pour prédire s'il existe un incident (Incident=1) ou non (Incident=0). Donner un exemple d'arbre de décision dans ce contexte et son taux d'erreur.

b. Même question que pour (a), à l'aide d'une régression logistique.

c. 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 arbre de décision pour prédire le Retard.