

Informatique Décisionnelle

Session de Janvier 2018

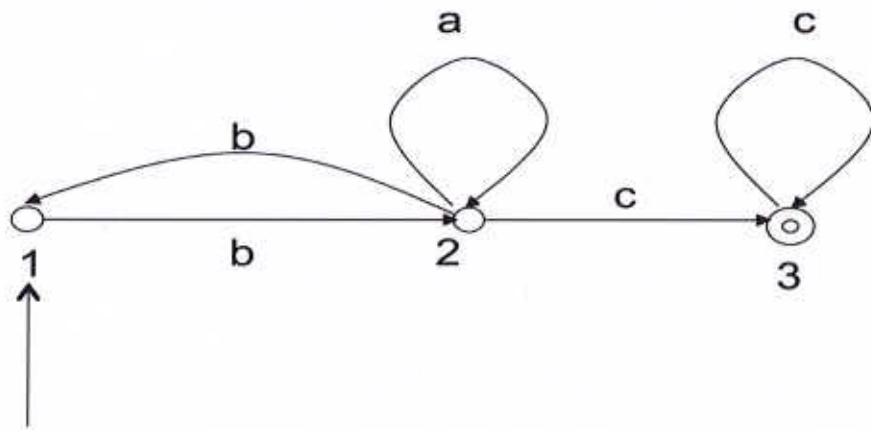
4 pages

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 de la figure.

- Trouver un mot de longueur au moins 6, accepté et un autre rejeté.
- Décrire le langage défini par cet automate par une expression régulière.
- Préciser si l'automate est déterministe ou non déterministe.
- Pour la distance d'édition, quelle est la distance du mot $bbbaaabcb$ au langage de l'automate.
- Trouver un automate B, dont le langage soit ϵ -loin de A.
- 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"?>
<MEDIA>
  <FILM LANG="fr">
    <AUTHOR>
      <FIRSTNAME>Alain</FIRSTNAME>
      <LASTNAME>Michard</LASTNAME>
    </AUTHOR>

    <TITLE>Le tigre</TITLE>
    <ACTOR>
      <FIRSTNAME>Alain</FIRSTNAME>
      <LASTNAME>Delon</LASTNAME>
    </ACTOR>
    <ACTOR>
      <FIRSTNAME>Catherine</FIRSTNAME>
      <LASTNAME>Deneuve</LASTNAME>
    </ACTOR>

    <DATEPUB>1998</DATEPUB>
  </FILM>
  <FILM LANG="an">
    <AUTHOR>
      <FIRSTNAME>William J.</FIRSTNAME>
      <LASTNAME>Pardi</LASTNAME>
    </AUTHOR>

    <TITLE>Le lion</TITLE>
    <PUBLISHER>
      <NAME>MGM</NAME>
      <PLACE>Hollywood</PLACE>
    </PUBLISHER>
    <DATEPUB>1999</DATEPUB>
  </FILM>
</MEDIA>
```

- a. Décrire ce fichier par un arbre XML, en abrégiant le nom des tags, par leur première lettre, sans écrire les attributs. Utiliser les deux premières lettres en cas d'ambiguïté.
<AUTHOR> sera considéré comme la balise <au> et <ACTOR> sera considéré comme la balise <ac>.
- 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 DTD₁, qui soit ε-loin de cette DTD pour la distance d'édition.

3. Problème de Schéma relationnel.

La ville de Paris maintient un système d'information pour gérer les prêts des livres par les bibliothèques municipales. Le système simplifié a les 4 tables suivantes :

BIBLIOTHEQUES(NumB, Nom, Arrondissement, Capacité) décrit les Bibliothèques avec un identifiant (NumB), leur Nom, L'Arrondissement, et la Capacité en m2 des bâtiments.

LIVRES(NumL, Titre, Auteur, Type, Année) décrit les livres avec un identifiant (NumL), un Titre, un Auteur (le premier auteur s'il y a plusieurs auteurs), le Type de livre (Fiction, Roman, Récit,...), l'Année de publication.

LECTEURS(NumP, Nom, Arrondissement, Genre, Age) décrit les lecteurs avec un identifiant (NumP) , leur Nom, l'Arrondissement s'ils habitent dans Paris, leur Genre (H ou F) et leur Age.

PRETS(NumB, NumL, NumP, Date, Durée) décrit le prêt par une Bibliothèque NumB, d'un livre NumL à un lecteur NumP, à une date donnée (jour/mois/année) et pour un Durée variable de 3 jours à 3 mois exprimée en nombre de jours.

PRETS

NumB	NumL	NumP	Date	Durée
------	------	------	------	-------

LECTEURS

<u>NumP</u>	Nom	Arrondissement	Genre	Age
-------------	-----	----------------	-------	-----

LIVRES

NumL	Titre	Auteur	Type	Année
------	-------	--------	------	-------

BIBLIOTHEQUES

<u>NumB</u>	Nom	Arrondissement	Capacité
-------------	-----	----------------	----------

Par exemple, le tuple (10, 20, 30, « 15/01/2017 », 90) décrit un tuple de la table **PRETS** correspondant au prêt par la Bibliothèque « 10 » du livre « 20 » au lecteur « 30 », le « 15/01/2017 », pour « 90 » jours.

- a. Décrire un schéma Entité-relations associé à ce Schéma relationnel.
- b. Ecrire des requêtes SQL pour trouver :
 1. Quels sont les titres des livres prêtés le « 15/01/2017 » ?
 2. Quels livres ont été prêtés à des hommes (« H ») en Janvier 2017 ?
 3. Quels lecteurs empruntent des livres pour plus de « 60 » jours
- c. Indiquer les dépendances fonctionnelles des tables **LECTEURS** et **PRETS**
- d. Quelle est la clé de la table **PRETS**?

4. Problème OLAP.

On souhaite analyser la table **PRETS** du schéma relationnel précédent, considérée comme un entrepôt de données pour comprendre les durées des prêts.

- a. Proposez un schéma étoile pour la table **PRETS**, avec *Durée* comme mesure et Somme comme Aggrégation.
- b. Donnez les expressions OLAP (chemin, mesure, agrégation) associées aux requêtes :
 - Analyse de la durée des prêts par Bibliothèque,
 - Analyse de la durée des prêts par type de livres et par genre des lecteurs,
 - Analyse de la durée des prêts en 2016 par arrondissement des lecteurs.
- c. Donnez une expression en SQL-étendu (utilisant GROUP BY) pour la 1^{ère} requête.

5. Fouille de Données

L'entrepôt de données (la table **PRETS** du problème 4) est alimenté chaque jour par les prêts des bibliothèques, selon le schéma du problème 3.

- a. On cherche à prédire la *Durée* des prêts, à partir de N=1000 enregistrements de cette table, en discrétisant la durée avec 3 valeurs : C, M, L pour Court, Moyen et Long. Expliquer comment construire un arbre de décision pour cette tâche et donner un exemple.
- b. Quelle est l'erreur de prédiction estimée ?
- c. Décrire comment utiliser une régression logistique pour prédire la *Durée*, comme à la question a.
- d. On souhaite regrouper les lecteurs en k classes, pour que les lecteurs de livres de même type soient dans les mêmes classes. Décrire comment construire une nouvelle table qui regrouperait ces informations et quelle technique utiliser.