Cours du Professeur Michel de Rougemont

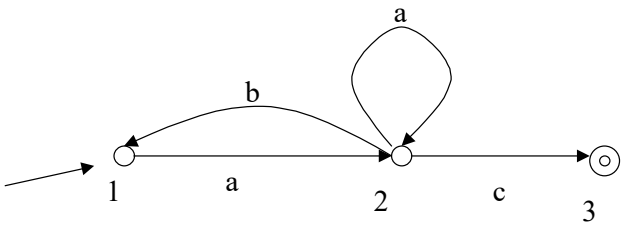
# Informatique et Techniques Numériques en Economie
## Session de Janvier 2007
## Corrigé

Tous documents autorisés. 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, 3 l'état final et les transitions sont définies par le graphe ci-dessous.

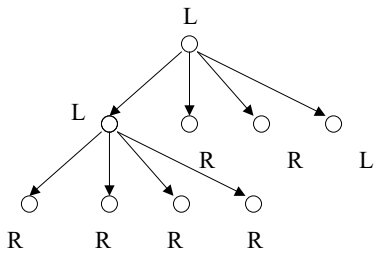
a. Le mot aaaac est accepté, le mot aaaaa est rejeté. Le langage de l'automate est L=a.((ba)*.a*)*.c
b. La distance d'Edition absolue du mot ababacba au langage de l'automate est 2, la distance relative est ¼. Il faut supprimer le c et l'insérer à la fin. La distance d'Edition avec déplacement serait de 1 et 1/8.
c. Si on supprime l'état 3 et si l'état 2 est acceptant on obtient A' dont le langage L'= a.((ba)*.a*)* est ε-proche de A. Il suffit d'enlever la dernière lettre de chaque mot.



## 2. Problème XML.

Soit la DTD suivante :
<?xml version='1.0' ?>
<!ELEMENT L (L,R*)>
<!ELEMENT L (#PCDATA)>
<!ELEMENT R (#PCDATA)>

a. Soit un arbre dont la racine est L avec un premier fils L et trois fils R. Le fichier XML est <L> <L> </L> <R> </R> <R> </R> <R> </R> </L>. L'arbre est valide. Le langage d'arbres est l'ensemble des arbres chainé à gauche par L, dont les fils sont un L et des Rs.
b. La distance d'édition absolue avec déplacement est de 1 ( 1/9 pour le distance relative). Il faut déplace le dernier L au niveau 1, comme premier L au niveau 2.




c. Ajouter <!ELEMENT L (A)> à la DTD. Il faudra enlever le A à chaque arbre pour le rendre valide pour la 1ère DTD.

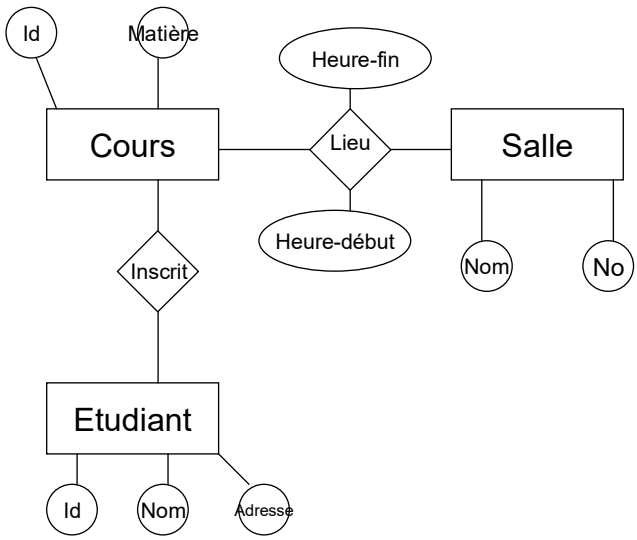
.

3. **Problème de Schéma relationnel.**

   a. Un Schéma d'une Base de Données est la définition de tables et de dépendances fonctionnelles. Une Requête est une fonction qui associe à un ensemble de tables une nouvelle table. Une requête imbriquée associe à un ensemble de tables et à une table définie par une requête, une nouvelle table. Une jointure est une opération sur deux tables qui partagent un attribut: un tuple est dans la jointure si sa projection sur chacune des deux tables existe.

   b. Un schéma entité-relation possible serait :

Schéma Entité Relation



a. Un schéma relationnel serait :

Cours(Id-C, Matière)
Salle(Nom, No)
Etudiant(Id-E, Nom, Adresse)
Inscrit(Id-E, Id-C)
Lieu(Id-C, Heure-début, Heure-fin, Nom)

b. Id-C, Nom, Id-E sont des clés pour Cours, Salle, Etudiant.

4. **Problème OLAP.**

Une université cherche à étudier les facteurs influant sur la réussite de ses étudiants aux examens. Pour cela elle décide de construire un entrepôt de données (datawarehouse).

Elle souhaite pouvoir répondre aux questions suivantes:

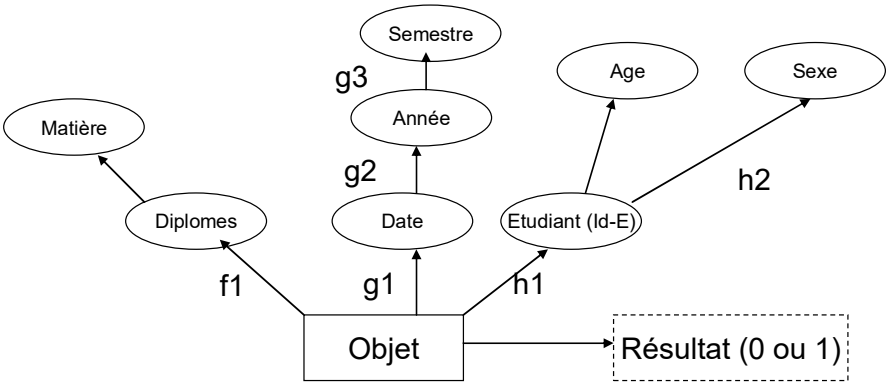
Quel est le nombre de réussites aux examens par cours, pour l'année 2003?
Quel est le nombre de réussites aux examens d'un cours obligatoire, pour l'année 2003?
Quel est le nombre de réussites aux examens par sexe (féminin, masculin), pour l'année 2003?
Combien d'étudiants ayant un âge de 22 ans ont réussi leurs examens de bases de données relationnelle?
Quel est le nombre de réussites aux examens pendant le semestre d'hiver 2002?

Pour cela elle dispose des données suivantes anonymes: Pour chaque examen passé, on connaît l'âge et le sexe de l'étudiant, le nom du cours (les cours peuvent être regroupés en cours obligatoire et cours à option), la date de l'examen, la note obtenue et si l'examen est réussi ou non.

**Réponses.**

Cours du Professeur Michel de Rougemont

Schéma OLAP



1. Le schéma étoile ci-dessous associe les hiérarchies pour les 3 dimensions principales : Diplomes, Date et Etudiant. La mesure est 0 si l'étudiant a échoué et 1 s'il a réussi son diplôme. L'aggrégation est la somme.

2. Expressions OLAP :

   - Pour le Sexe : h2°h1

   - Pour l'année : g2°g1.

   - Quel est le nombre de réussites aux examens par sexe (féminin, masculin), pour l'année 2003?

     - Filtre : année=2003, Dimension= Sexe, Mesure=Résultat, Aggrégation=Somme

     - On obtient une distribution :du type (Femme 60%, Homme 40%)

   - Combien d'étudiants ayant un âge de 22 ans ont réussi leurs examens de bases de données relationnelles?
     - Filtre : age=22 Matière= « bases de données relationnelle », Dimension= (vide), Mesure=Résultat, Aggrégation=Somme

     - Le résultat est un chiffre.

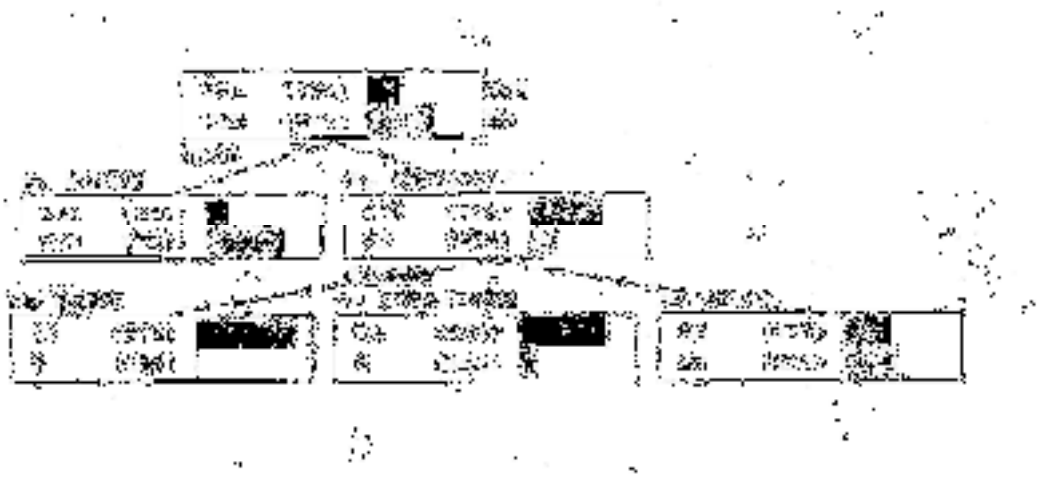
5. **Fouille de Données**

Vous avez appliqué un algorithme permettant de générer un arbre de décision sur une base de données contenant les données sur les passagers du Titanic. Cette base de données présente quatre attributs pour chacun des 2201 passagers du titanic.

- CLASS avec comme valeurs: Ist, $2^{nd}$, 3rd, crew

  - La classe dans laquelle voyageait le passager ou s'il s'agissait d'un membre d'équipage.
- SEX avec comme valeurs : female, mâle
- SURVIVED avec comme valeurs: no, yes
  - Est-ce que le passager a survécu?

La question est de trouver un lien entre la classe, le sexe et le fait d'avoir survécu ou non au naufrage du Titanic.



**Questions :**

c. L'arbre décrit la proportion de survivants (33%), décomposé par Homme/Femme puis par Classe (1,2,3) pour les femmes. Les proportions de survivantes sont 97, 86 et 47% pour les classes 1,2 et 3.

d. L'arbre de décision est partiel, car il ne donne que les proportions des femmes par classe (proportion de passagers en $1^{ère}$, $2^{ème}$ et $3^{ème}$ classe). La prédiction qu'un homme survive est de 22%, bien qu'elle soit sans-doute supérieure pour un homme de la $1^{ère}$ classe. La fonction de prédiction G, n'est pas une bonne approximation, car la probabilité d'erreur ( Pr [F(x) not= G(x)] ) ne peut pas être rendue aussi petite que souhaité.

e. Si toutes les femmes avaient survécu et qu'aucun homme n'avait survécu, l'arbre serait constitué de deux seules branches (Homme, Femme) qui prédiraient exactement la fonction F.