

1

Extraction de Connaissances et Fouille de Données

Pascal Poncelet
Centre de Recherche LGI2P
Ecole des Mines d'Alès
Pascal.Poncelet@ema.fr
<http://www.lgi2p.ema.fr/~poncelet>

Plan

- Pourquoi fouiller les données ?
- Le processus d'extraction
- Quelques domaines d'application
- Un aperçu de quelques techniques

2

Pourquoi fouiller les données ?

- De nombreuses données sont collectées et entreposées
 - Données du Web, e-commerce
 - Achats dans les supermarchés
 - Transactions de cartes bancaires
- Les ordinateurs deviennent de moins en moins chers et de plus en plus puissants
- La pression de la compétition est de plus en plus forte
 - Fournir de meilleurs services, s'adapter aux clients (e.g. dans les CRM)

3

Pourquoi fouiller les données ?

- Les données sont collectées et stockées rapidement (GB/heures)
 - Capteurs : RFID, supervision de procédé
 - Télescopes
 - Puces à ADN générant des expressions de gènes
 - Simulations générant de téraoctets de données

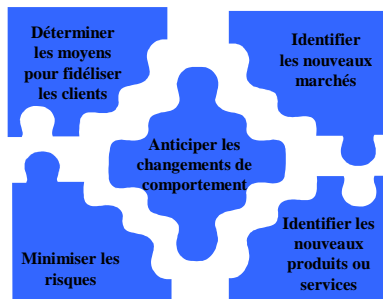
4

Pourquoi fouiller les données ?

- Les techniques traditionnelles ne sont pas adaptées
- Volume de données trop grands (trop de tuples, trop d'attributs)
 - Comment explorer des millions d'enregistrements avec des milliers d'attributs ?*
- Besoins de répondre rapidement aux opportunités
- Requêtes traditionnelles (SQL) impossibles
 - « Rechercher tous les enregistrements indiquant une fraude »*
- Croyance dans la présence de données importantes

5

Un enjeu stratégique



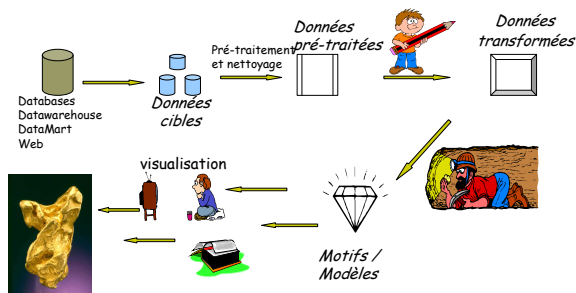
6

Qu'est ce que le Data Mining ?

- De nombreuses définitions
 - Processus **non trivial** d'extraction de connaissances d'une base de données pour obtenir de nouvelles données, valides, potentiellement utiles, compréhensibles,
 - Exploration et analyse, **par des moyens automatiques ou semi-automatiques**, de grandes quantités de données en vue d'extraire des motifs intéressants

7

Le processus de KDD



8

Données, Informations, Connaissances

Décision
•Promouvoir le produit P dans la région R durant la période N
•Réaliser un mailing sur le produit P aux familles de profil F

Connaissance (data mining)
•Une quantité Q du produit P est vendue en région R
•Les familles de profil F utilisent M% de P durant la période N

Information (requêtes)
•X habite la région R
•Y a A ans
•Z dépense son argent dans la ville V de la région R

Données
•Consommateurs
•Magasins
•Ventes
•Démographie
•Géographie

9

Data Mining ou non ?

• NON

Rechercher le salaire d'un employé

Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining

• OUI

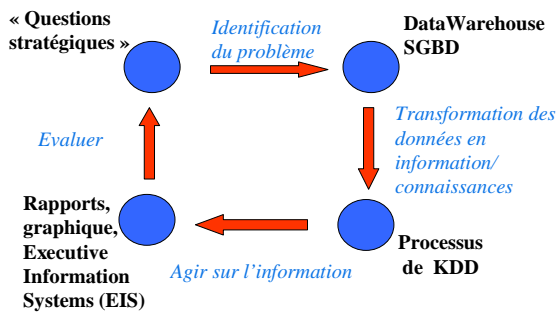
Les supporters achètent de la bière le samedi et de l'aspirine le dimanche

Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu

Contexte général

10

Cycle de vie du KDD



Contexte général

11

Applications

- ❑ Médecine : bio-médecine, drogue, Sida, séquence génétique, gestion hôpitaux, ...
- ❑ Finance, assurance : crédit, prédiction du marché, détection de fraudes, ...
- ❑ Social : données démographiques, votes, résultats des élections,
- ❑ Marketing et ventes : comportement des utilisateurs, prédiction des ventes, espionnage industriel, ...
- ❑ Militaire : fusion de données .. (secret défense)
- ❑ Astrophysique : astronomie, « contact » (;-)
- ❑ Informatique : agents, règles actives, IHM, réseau, Data-Warehouse, Data Mart, Internet (moteurs intelligent, profiling, text mining, ...)

Contexte général

12

Quid des données ?

- Grandes Bases de Données ou non ?
- Faut -il échantillonner ?
 - 100 000 enregistrements, 100 Mo par jour
 - 2 Go par jour, 100 Go par heure
 - *Déjà les petabyte (2⁵⁰) ...*
- Différents domaines
 - Bases de Données
 - Intelligence Artificielle (Machine Learning)
 - Statistiques
 - Algorithmique, ...

13

Les tâches du DM

- Data Mining : de nombreuses tâches possibles ...
 - Classification
créer une fonction qui classe une donnée élémentaire parmi plusieurs classes prédéfinies existantes
 - Régression
créer une fonction qui donne une donnée élémentaire à une variable de prévision avec des données réelles
 - Groupement (clustering)
rechercher à identifier un ensemble fini de catégories ou groupe en vues de décrire les données
 - Résumé
affiner une description compacte d'un sous-ensemble de données
 - Modélisation des dépendances
trouver un modèle qui décrit des dépendances significatives entre les variables
 - Détection de changement et déviation
découvrir les changements les plus significatifs dans les données

14

Les tâches du DM

- Non pas 1 mais n approches ... donc m techniques ...
- 3 approches principales (*R. Agrawal*) vision BD
 - Classification*
 - Règles d'association*
 - Motifs séquentiels*

15

Classification

- division de l'ensemble de données en classes disjointes en utilisant un apprentissage supervisé ou non (clustering)
- *But* : recherche d'un ensemble de prédicats caractérisant une classe d'objet et qui peut être appliqué à des objets inconnus pour prévoir leur classe d'appartenance.
- *Exemple* : une banque peut vouloir classer ses clients pour savoir si elle accorde un crédit ou non.
- *Techniques* : Arbre de décision, réseaux neuronaux, ...

16

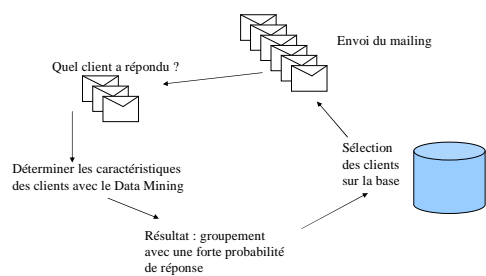
Le mailing

□ Classification... un exemple d'utilisation

- un cadeau est envoyé par mailing. Un envoi sans réponse coûte 50 € et une réponse assure 100 €.
- Pas d'envoi de mailing à un client qui aurait répondu : perte de 100 €.

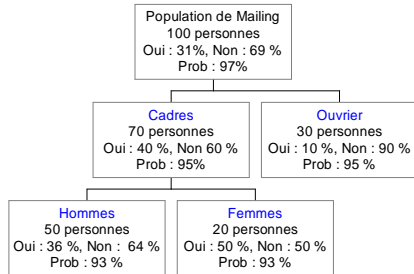
17

Le mailing



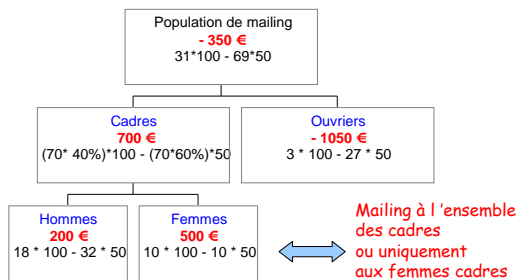
18

Résultat du mailing



19

Quantification



20

Evaluation

Prédit ↓	Matrice de coûts			TOTAL
	Payé	Retardé	Impayé	
Payé	80	15	5	100
Retardé	1	17	2	20
Impayé	5	2	23	30
TOTAL	86	34	30	150

Validité du modèle : nombre de cas exacts
 (=somme de la diagonale) divisé par le nombre total :
 $120/150 = 0.8$

21

Recherche de motifs fréquents

- Qu'est ce qu'un motif fréquent ?
 - Un motif (ensemble d'items, séquences, arbres, ...)
qui interviennent fréquemment ensemble dans une
base de données [AIS93]
- Les motifs fréquents : une forme importante de
régularité
 - Quels produits sont souvent achetés ensemble ?
 - Quelles sont les conséquences d'un ouragan ?
 - Quel est le prochain achat après un PC?

22

Recherche de motifs fréquents

- Analyse des associations
 - Panier de la ménagère, cross marketing, conception de
catalogue, analyse de textes
 - Corrélation ou analyse de causalité
- Clustering et Classification
 - Classification basée sur les associations
- Analyse de séquences
 - Web Mining, détection de tendances, analyses ADN
 - Périodicité partielle, associations temporelles/cycliques

23
