Chapitre 1 : Introduction à la fouille de données

1. Introduction

- Les données d'une entreprise doublent de volume chaque année. Elles peuvent être interrogées par des outils et langages qui ont prouvé leur efficacité.
- Par exemple, dans les SGBD relationnels, on utilise le langage SQL → cela suppose la connaissance des schémas de données, et du contenu général de la base.
- Cependant, le grand volume des données peut renfermer des connaissances que les outils classiques d'interrogation ne peuvent pas extraire
  - quel est le volume d'achat du client X, durant la période Y, quel est le meilleur client (max du volume d'achat, durée de fidélité...) → requête satisfaite par des outils classiques (SQL)
  - quels sont les caractéristiques des clients qui rompent (change d'entreprise), comment savoir si un demandeur de prêt est solvable (peut rembourser le prêt?), … → requête non (ou difficilement) satisfaite par les outils classiques.
    → besoin d'outils et techniques spécifiques pour extraire des connaissances à partir de données

- On parle du KDD (knowledge discovery from data) ou d'ECD (extraction des connaissances à partir de données)

- La fouille de données (data mining - DM) est au cœur de l'ECD et en représente le moteur.

C'est une ingénierie renfermant des outils, des techniques, des algorithmes, etc qu'elle puise dans les statistiques et analyse de données, les bases de données, l'intelligence artificielle, etc.

- Le DM permet d'aboutir à des modèles (ex : fonction mathématique, règles logiques SI condition ALORS résultats) qui doivent être validés pour devenir des connaissances utilisables par l'être humain ou par les machines.

2. Définition de la fouille de données

Les définitions du data mining ne font pas parfois la différence entre le *data mining* qui est la fouille de données et le *KDD* ou *knowledge discovery from data* qu'on peut traduire par l'extraction des connaissances à partir des données. Nous prenons les deux définitions suivantes

- Fayyad : « l'extraction de connaissances à partir des données est un processus non trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données »
- Frawley : « extraction non triviale d'informations implicite, précédemment inconnue, et potentiellement utiles à partir des données ».

Selon les deux définitions précédentes, le champ est ouvert aux techniques et applications du DM. On cite la classification, la régression, le clustering, les règles d'association, etc.

3. Historique de la fouille de données

La fouille de données est une évolution naturelle dans l'exploitation des données par les être humains en utilisant les ordinateurs. On peut résumer cette évolution dans les points suivants :

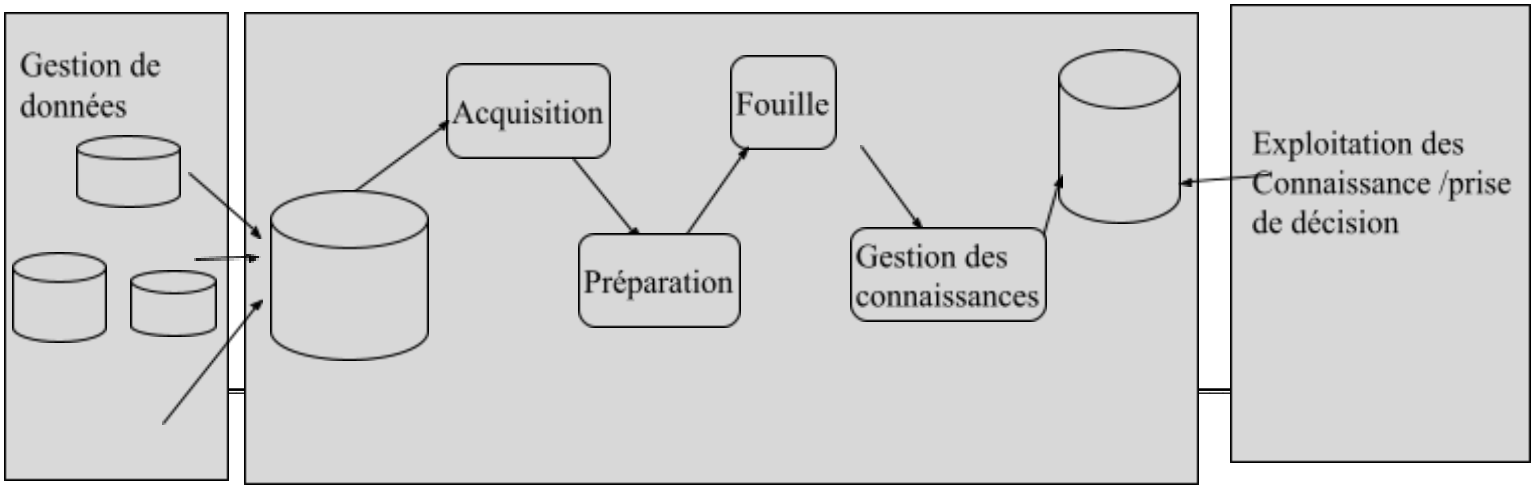
- Début de l'informatique (années 40): utilisation des ordinateurs pour les besoins de calcul
- Traitement statistique des données et analyse de données : prémices du DM
- Fin des années 80 : exploitation du contenu des bases de données pour la recherche de règles d'association : utilisation du terme database mining
- 1989 : premier atelier sur la découverte de connaissances – proposition du terme Knowledge Discovery par Gregory Piatetsky-Shapiro
- 1995 : première conférence sur le data mining.

Aussi, le DM a été influencé par

- l'explosion du volume de données produites et stockées
- la maturité des outils de reporting des données et l'évolution du besoin des utilisateurs (de la gestion de données vers la prise de décision)
- l'évolution de la relation avec les clients : vers le profiling des clients et la production orientée client.

4. Processus de la fouille de données

- Le processus de fouille de données, ou généralement d'ECD se positionne en back-end au sein de l'entreprise ou des cabinets spécialisés.
- En front-end, on trouve les activités de production des données en amont et de prise de décision en aval.
- Le schéma suivant (adapté de Zighed et Rakotomalala) résume les quatre étapes de l'ECD et le positionnement du DM au milieu comme maillon fort.

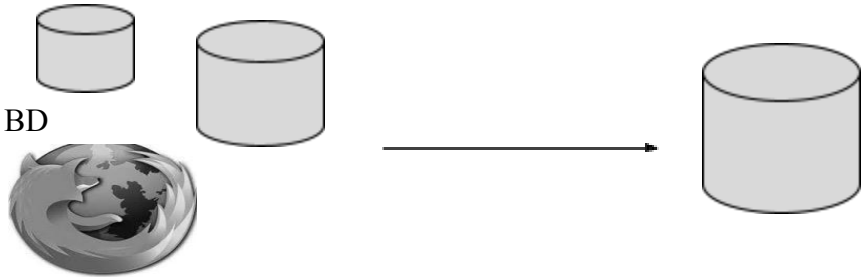


DWH

Knowledge
Base

- **Pré-étape d'ECD : Alimentation de l'entrepôt de données / production de données**

Schéma

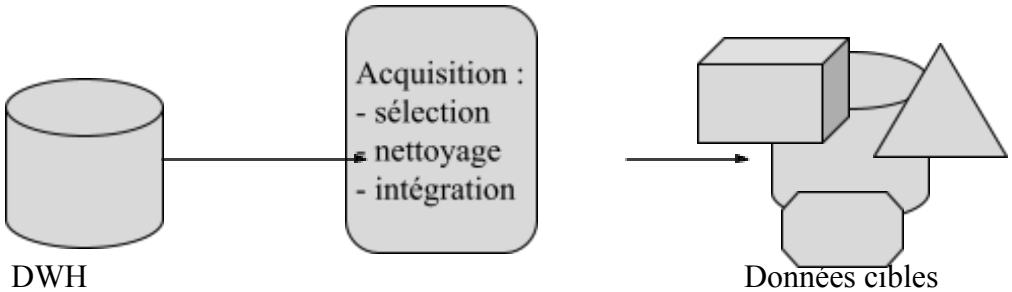


Web DWH

Description : il s'agit de la partie en amont du front office qui consiste à alimenter l'entrepôt de données (grande BdD) de l'entreprise à partir des bases de données de production, du Web ou d'autres sources (surveys, applications d'analyse, ...).

- **Etape 1 : Acquisition de données**

Schéma



DWH Données cibles

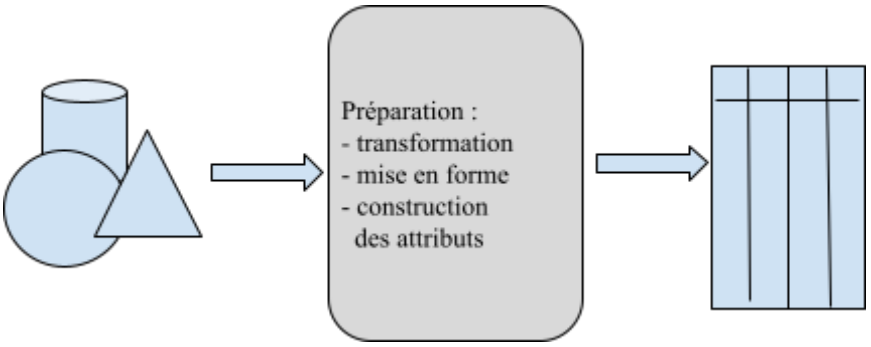
Description

- le processus d'ECD ne cible pas toutes les données de l'entreprise mais seulement celles qui serviront à résoudre le problème.
- L'acquisition permet de cibler les données utiles.
- On utilise des requêtes ad hoc (non pré définies), de l'échantillonnage (sampling), etc...
- Il n'y a pas de limite de taille des données cibles

- Les données cibles peuvent nécessiter un nettoyage (élimination des attributs mal renseignés, erronés, etc)

- **Etape 2 : Préparation de données**

Schéma



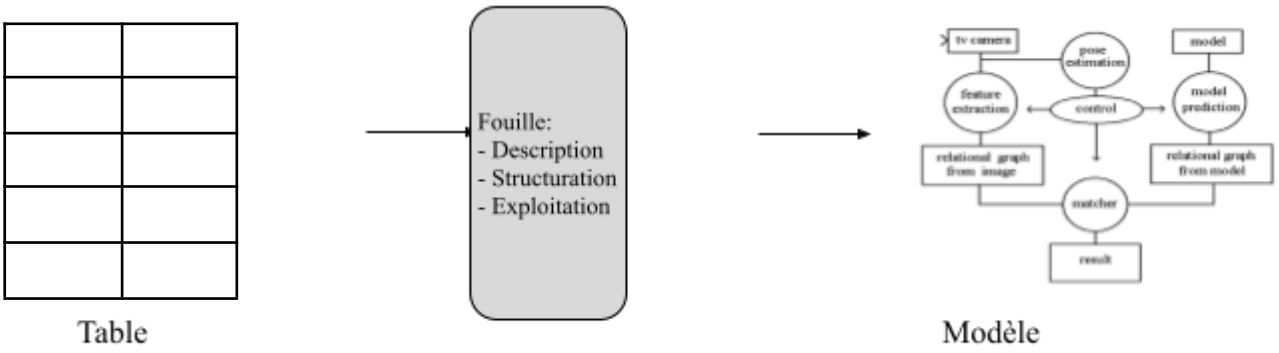
Données cibles Table

Description

- Souvent, les données exploitées par l'ECD doivent avoir une forme tabulaire (ligne/colonne).
- Si les données n'ont pas cette forme, elles sont transformées et adaptées.
- Parfois, même la forme tabulaire initiale nécessite une autre transformation (centrage, mise sous une forme binaire '0/1', etc.).
- La construction d'attributs inclut :
  - la réduction du volume de données par élimination des attributs inutiles
  - la transformation: par exemple passer d'un attribut continu (ex : température) à un attribut discret (intervalle).
  - la construction d'agrégats : il s'agit d'attributs obtenus à partir d'autres qui permettent d'effectuer des comparaisons (ex : remplacer le prix d'un appartement et la surface par le prix au mètre carré, remplacer la ville par la région, etc).
- A cette étape également, les données absentes ou erronées mais nécessaires sont traitées : on utilise différentes techniques comme : le remplacement de la valeur absente par la valeur la plus fréquente, l'estimation de la valeur absente à partir des valeurs existantes, etc.

- **Etape 3 : Fouille de données**

Schéma

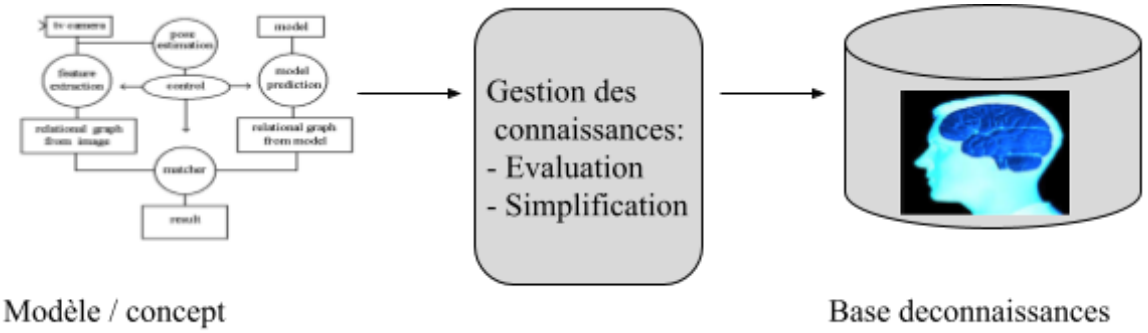


Description

- phase de fouille à proprement parler
- met en œuvre différentes techniques, selon le problème à traiter (voir section suivante)
- permet d'obtenir des modèles et des concepts à valider.

- **Etape 4 : Gestion des connaissances**

Schéma



Description

- Le modèle obtenu à partir de l'étape précédente doit être validé pour être utilisé dans des cas réel (ex : avant de pouvoir trancher sur un diagnostic d'un cas de maladie).
- On procède au calcul du taux d'erreur du modèle : si le taux d'erreur est accepté, on utilise le modèle. Sinon, le modèle n'est pas utilisé.

- **Post étape d'ECD : exploitation des connaissances / prise de décision**

Schéma

Décideur Base de connaissances

Description : une fois le modèle issu de la fouille de données validé ; le décideur l'utilise sur des données réelles pour prendre des décision selon le domaine. Ces décisions peuvent être par exemple

- d'accepter d'octroyer un prêt à un candidat,
- d'estimer le taux de vote et les résultats par candidat
- ...

5. Méthodes du data mining

- Selon le problème à traiter, il existe plusieurs méthodes de data mining.
- Ces méthodes peuvent être classés comme suit :

5.1. Visualisation et description :

- ces méthodes synthétisent et décrivent les données sous une forme visuelle qui permet une interprétation plus ou moins directe.
- elles se basent sur l'affichage d'indicateurs statistiques (moyennes, écart-type, médianes, modes, etc).
- Différentes solutions de visualisation peuvent être utilisées : les courbes ou tableaux de statistiques, les histogrammes, les nuages de points, etc.

5. 2. Classification et structuration

- l'objectif est de **classer un ensemble d'individus** afin de mieux comprendre la réalité (simplification) ou pour d'autres fins (ex : identifier des groupes de clients ayant des profils similaires afin de les cibler par des messages communs).
- Ces méthodes relèvent de l'apprentissage non supervisé car l'utilisateur ne sait pas a priori quelles classes obtenir.

- Les méthodes utilisées sont les méthodes de classification automatique ou cluster analysis.

5. 3. Explication et prédiction

- l'objectif est d'aboutir à un modèle d'explication d'un phénomène ou de prédiction. Des exemples sont (1) l'aide au diagnostic (si un patient est atteint ou pas d'une maladie), (2) la décision qu'un message est un spam, etc.
- parmi les méthodes, on cite : les arbres de décision, les réseaux de neurones, les réseaux bayésiens, les règles d'association, la régression, l'analyse discriminante

6. Quelques domaines d'application

Parmi ces domaines d'application, on cite :

- la gestion de la relation client : appelée aussi CRM. Parmi les applications :
  - profiling : connaître et regrouper les clients en profils pour mieux les cibler par des campagnes publicitaire → utilisation de la classification
  - marketing : regrouper les produits les plus fréquemment achetés ensembles dans les même stand → utilisation des règles d'association
- la médecine : aide au diagnostic des malades → utilisation des arbres de décision, des réseaux bayésiens,...
- les télécommunications
  - identification des fraudes de cartes de crédit
- la bureautique : identification des messages spam, aide contextuelle.
- la politique : prédiction des résultats des élections...