Chapitre 1 : Introduction à la fouille de données

1. Introduction

Chaque année, le volume de données d'une entreprise double. Ces données peuvent être exploitées via des outils et des langages ayant fait leurs preuves. Par exemple, dans les systèmes de gestion de bases de données relationnelles, on utilise le langage SQL. Cette approche nécessite toutefois une bonne connaissance des schémas et du contenu général de la base.

Cependant, l'important volume de données peut receler des informations que les outils d'interrogation classiques ne parviennent pas à extraire. Par exemple, il est facile de répondre à des questions comme « quel est le volume d'achat du client X sur une période donnée ? » ou « quel est le meilleur client (volume d'achat maximal, durée de fidélité la plus longue) ? » avec SQL. En revanche, des questions plus complexes telles que « quelles sont les caractéristiques des clients qui changent d'entreprise ? » ou « comment déterminer la solvabilité d'un demandeur de prêt ? » sont difficiles, voire impossibles, à traiter avec les outils classiques. D'où la nécessité de recourir à des outils et techniques spécifiques pour extraire des connaissances à partir des données.

On parle alors de KDD (Knowledge Discovery from Data) ou d'ECD (Extraction de Connaissances à partir de Données).

La fouille de données (data mining – DM) est au cœur de l'ECD et en constitue le moteur principal. Il s'agit d'une discipline d'ingénierie qui rassemble des outils, des techniques et des algorithmes issus des statistiques, de l'analyse de données, des bases de données, de l'intelligence artificielle, etc.

Le DM permet d'obtenir des modèles (par exemple, une fonction mathématique, des règles logiques du type SI condition ALORS résultat) qui doivent être validés pour devenir des connaissances exploitables par l'humain ou par des machines.

2. Définition de la fouille de données

Il arrive que les définitions du data mining ne distinguent pas clairement la fouille de données (DM) du processus global d'extraction de connaissances (KDD). Nous retiendrons ici 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 ».

Ces définitions ouvrent la voie à un large éventail de techniques et d'applications du DM, telles que la classification, la régression, le clustering, les règles d'association, etc.

3. Historique de la fouille de données

La fouille de données s'inscrit dans une évolution naturelle de l'exploitation des données par l'homme au moyen des ordinateurs. On peut résumer cette évolution comme suit :

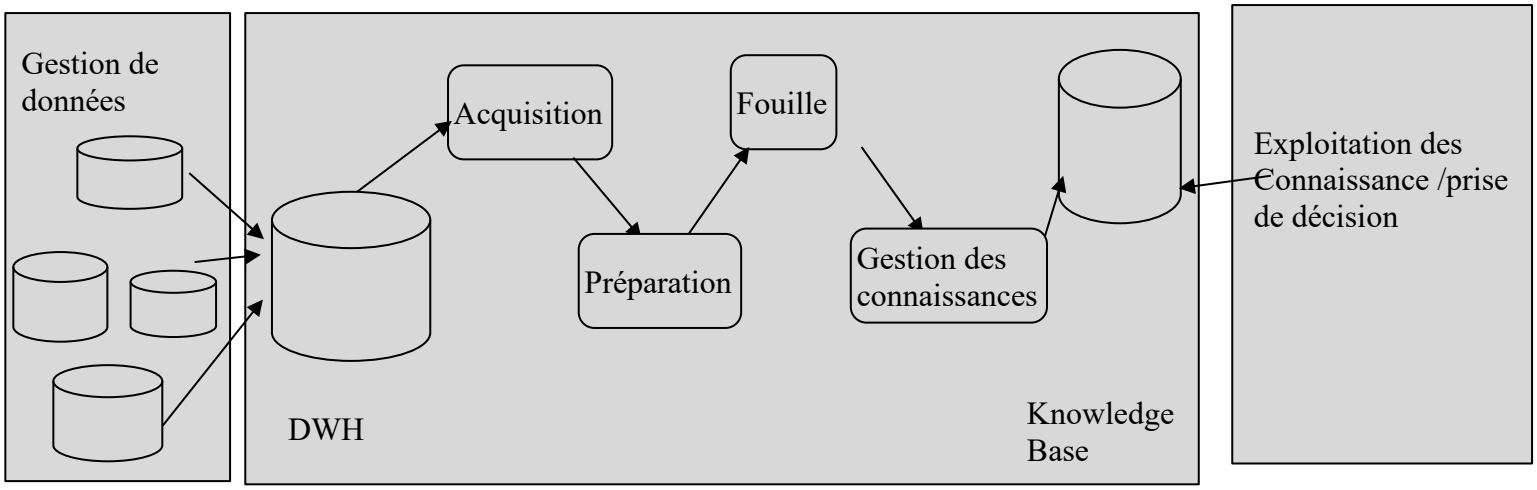
- Début de l'informatique (années 1940) : utilisation des ordinateurs pour les besoins de calcul.
- Traitement statistique et analyse des données : prémices du DM.
- Fin des années 1980 : exploitation du contenu des bases de données pour la recherche de règles d'association ; apparition du terme *database mining*.
- 1989 : premier atelier sur la découverte de connaissances ; Gregory Piatetsky-Shapiro propose le terme *Knowledge Discovery*.
- 1995 : première conférence dédiée au data mining.

Le DM a également été influencé par :

- l'explosion du volume des données produites et stockées ;
- la maturité des outils de reporting et l'évolution des besoins des utilisateurs (passage de la gestion de données à l'aide à la décision) ;
- l'évolution de la relation client, avec l'essor du profilage des clients et de la production orientée client.

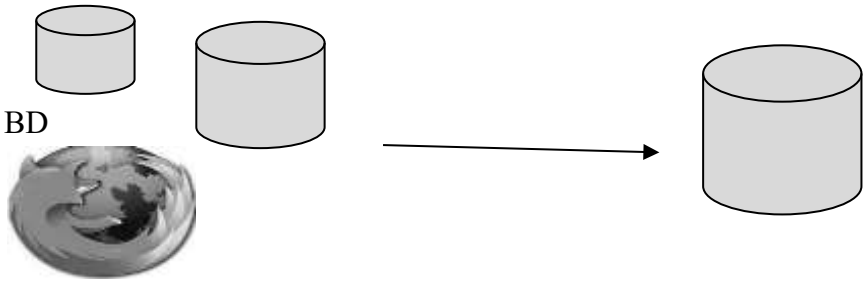
4. Processus de la fouille de données

- 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.



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

Schéma

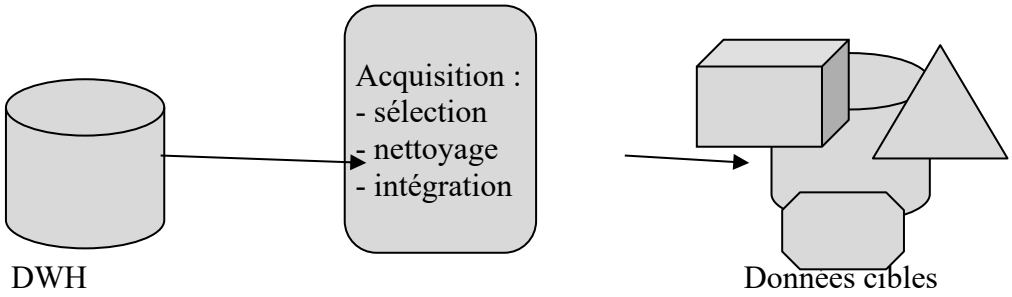


Web DWH

Description : Cette phase amont du front office consiste à alimenter l'entrepôt de données (grande base de données) de l'entreprise à partir des bases de production, du Web ou d'autres sources (enquêtes, applications d'analyse, etc.).

- **Etape 1 : Acquisition de données**

Schéma

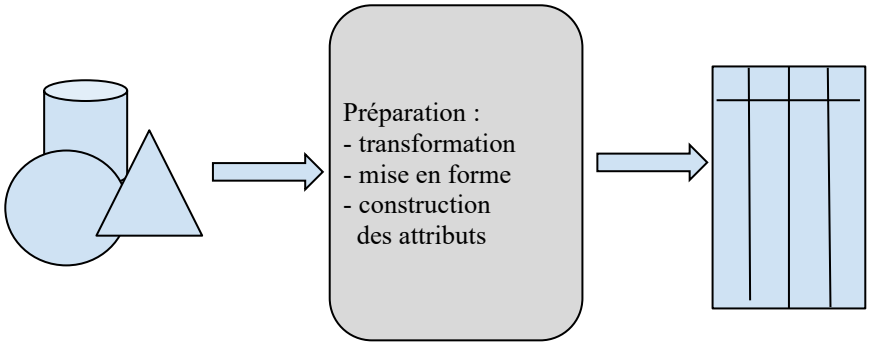


Description

- Le processus d'ECD ne cible pas la totalité des données de l'entreprise, mais uniquement celles utiles à la résolution du problème considéré.
- L'acquisition permet de sélectionner les données pertinentes.
- On utilise pour cela des requêtes ad hoc (non prédéfinies), des techniques d'échantillonnage (sampling), etc.
- La taille des données ciblées n'est pas limitée.
- Les données cibles peuvent nécessiter un nettoyage (suppression des attributs mal renseignés, erronés, etc.).

- **Etape 2 : Préparation de données**

Schéma



Données cibles Table

Description

Les données exploitées par l'ECD doivent souvent se présenter sous forme tabulaire (lignes et colonnes).
Si les données n'ont pas cette forme, elles sont transformées et adaptées.
Parfois, même une forme tabulaire initiale doit être modifiée (centrage, mise sous forme binaire 0/1, etc.).
La construction d'attributs comprend :

- la réduction du volume de données par élimination des attributs inutiles ;
- la transformation, par exemple le passage d'une variable continue (température) à une variable discrète (intervalles) ;
- la construction d'agrégats, c'est-à-dire de nouveaux attributs issus d'autres variables, facilitant les comparaisons (ex. : remplacer le prix et la surface d'un appartement par le prix au mètre carré, remplacer la ville par la région, etc.).

À cette étape, on traite également les données manquantes ou erronées mais indispensables, en utilisant diverses techniques : remplacement de la valeur absente par la valeur la plus fréquente, estimation de la valeur manquante à partir des données existantes, etc.

- **Etape 3 : Fouille de données**

Schéma

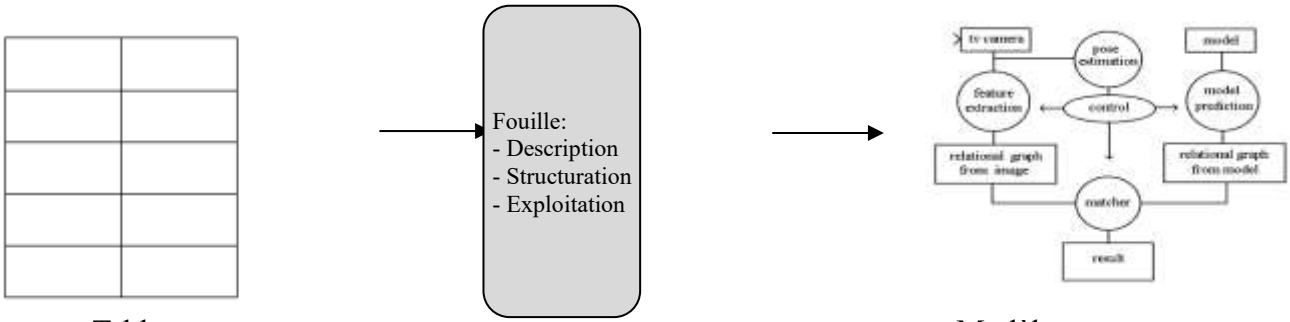


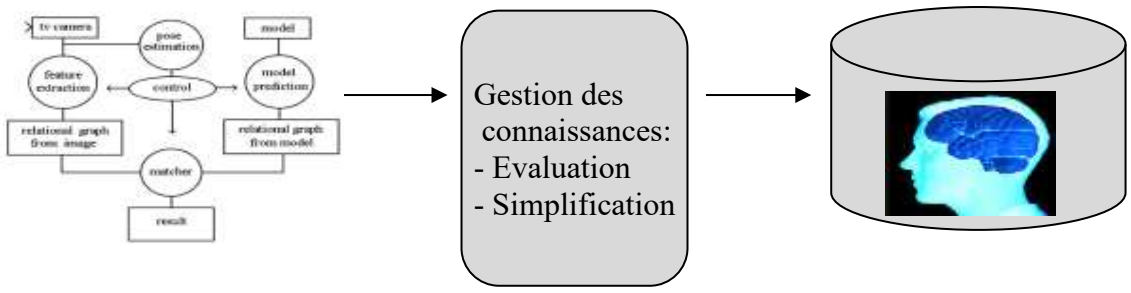
Table Modèle

Description

- Phase centrale de la fouille proprement dite.
- Met en œuvre différentes techniques, selon la nature du problème (voir section suivante).
- Permet d'obtenir des modèles et des concepts à valider.

- **Etape 4 : Gestion des connaissances**

Schéma



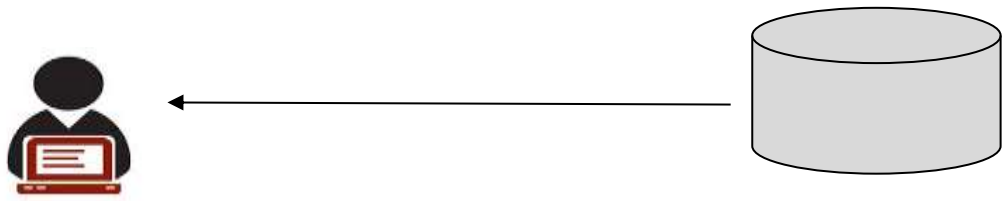
Modèle / concept Base deconnaissances

Description

- Le modèle issu de l'étape précédente doit être validé avant une utilisation réelle (par exemple, pour poser un diagnostic médical).
- On calcule le taux d'erreur du modèle : s'il est jugé acceptable, le modèle est utilisé ; sinon, il est rejeté.

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

Schéma



Décideur Base de connaissances

Description : Une fois le modèle validé, le décideur l'applique à des données réelles pour prendre des décisions propres au domaine d'application. Par exemple :

- accorder ou non un prêt à un demandeur ;
- estimer les taux de participation et les résultats d'une élection ;
- etc.

...

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 présentent les données sous une forme visuelle permettant une interprétation relativement directe.
- Elles s'appuient sur l'affichage d'indicateurs statistiques (moyennes, écarts-types, médianes, modes, etc.).
- Divers outils de visualisation peuvent être employés : courbes, tableaux de statistiques, histogrammes, nuages de points, etc.

5. 2. Classification et structuration

- L'objectif est de regrouper un ensemble d'individus afin de mieux comprendre la réalité (simplification) ou pour d'autres finalités (par exemple, identifier des groupes de clients aux profils similaires pour leur adresser des messages ciblés).

- Ces méthodes relèvent de l'apprentissage non supervisé, car l'utilisateur ne connaît pas à l'avance les classes à obtenir.

5. 3. Explication et prédiction

- L'objectif est d'élaborer un modèle capable d'expliquer un phénomène ou d'effectuer des prédictions. Exemples : (1) aide au diagnostic (déterminer si un patient est atteint d'une maladie), (2) décision de classer un message comme spam, etc.

- Parmi les méthodes utilisées, on peut citer : 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...