LE DATA MINING

La fouille de données

Dr. Nedioui Med Abdelhamid
Université Echahid Hama Lakhder Eloued - Algérie
nedioui3904@gmail.com

Chapitre 1

La fouille de données

1. Introduction

Aujourd'hui, les entreprises ont à leur disposition une masse de données importante. En effet, les faibles coûts des machines en termes de stockage et de puissance ont encouragé les sociétés à accumuler toujours plus d'informations. Cependant, alors que la quantité de données à traiter augmente énormément. Ces réservoirs de connaissance doivent être explorés afin d'en comprendre le sens et de déceler les relations entre données, des modèles expliquant leur comportement.

La croissance de la concurrence, oblige les entreprises à non plus simplement réagir au marché mais à l'anticiper. Elles doivent également cibler au mieux leur clientèle afin de répondre à ses attentes. La connaissance de son métier, de comportement de ses clients, de ses fournisseurs lui permet d'anticiper sur l'avenir.

1. Qu'est ce que la fouille de données ?

1.1 Définition :

Le Data Mining est définie par :

- Le processus de **découverte (l'extraction)** de connaissances dans les bases de données *(Knowledge Discovery in Database - **KDD**) en utilisant un* ensemble des technique et de méthodes du domaine des statistiques, des mathématiques et de l'informatique permettant **l'extraction**, à partir d'un important volume de données brutes.
- Un processus itératif et interactif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables.(Zighed et al., 2001).

1. Qu'est ce que la fouille de données ?

1.2 Domaine d'application :

Aujourd'hui le data mining a une grande importance économique du fait qu'elle permet d'optimiser la gestion des ressources (humaines et matérielles).

Elle est utilisée par exemple dans :

- **Marketing :** organisation des rayonnages dans les supermarchés en regroupant les produits qui sont généralement achetés ensemble (pour que les clients n'oublient pas bêtement 'acheter un produit parce qu'il est situé à l'autre bout du magasin).

- Par exemple, on extraira une règle du genre : "les clients qui achètent le produit X, achètent aussi le produit Y" ;

1. Qu'est ce que la fouille de données ?

1.2 Domaine d'application :

- **Scoring :** Le scoring est par exemple utilisé chez les assurances, les banques ou encore les opérateurs téléphoniques. (ex : ne pas accorder un prêt à un client qui présente un profil reconnu par le datamining comme présentant un haut risque de non remboursement.)

- Le Data Mining peut par exemple être utilisé pour déterminer quels sont les critères à prendre en compte pour considérer un client comme "réceptif en fonction du profil du demandeur de crédit, de sa demande, et des expériences passées de prêts ;

1. Qu'est ce que la fouille de données ?

1.2 Domaine d'application :

- **Médical :** les patients ayant tels et tels symptômes et demeurant dans des agglomérations de plus de centaine habitants développent couramment telle pathologie .
- **Prévention :** Plusieurs expériences ont été menées dans le domaine de prévention de crime.

- Une utilisation aux USA a par exemple été d'identifier les associations de lieu et de plages horaires auxquelles les crimes se produisaient le plus, afin de renforcer la présence policière en conséquence.

1. Qu'est ce que la fouille de données ?

1.2 Domaine d'application :

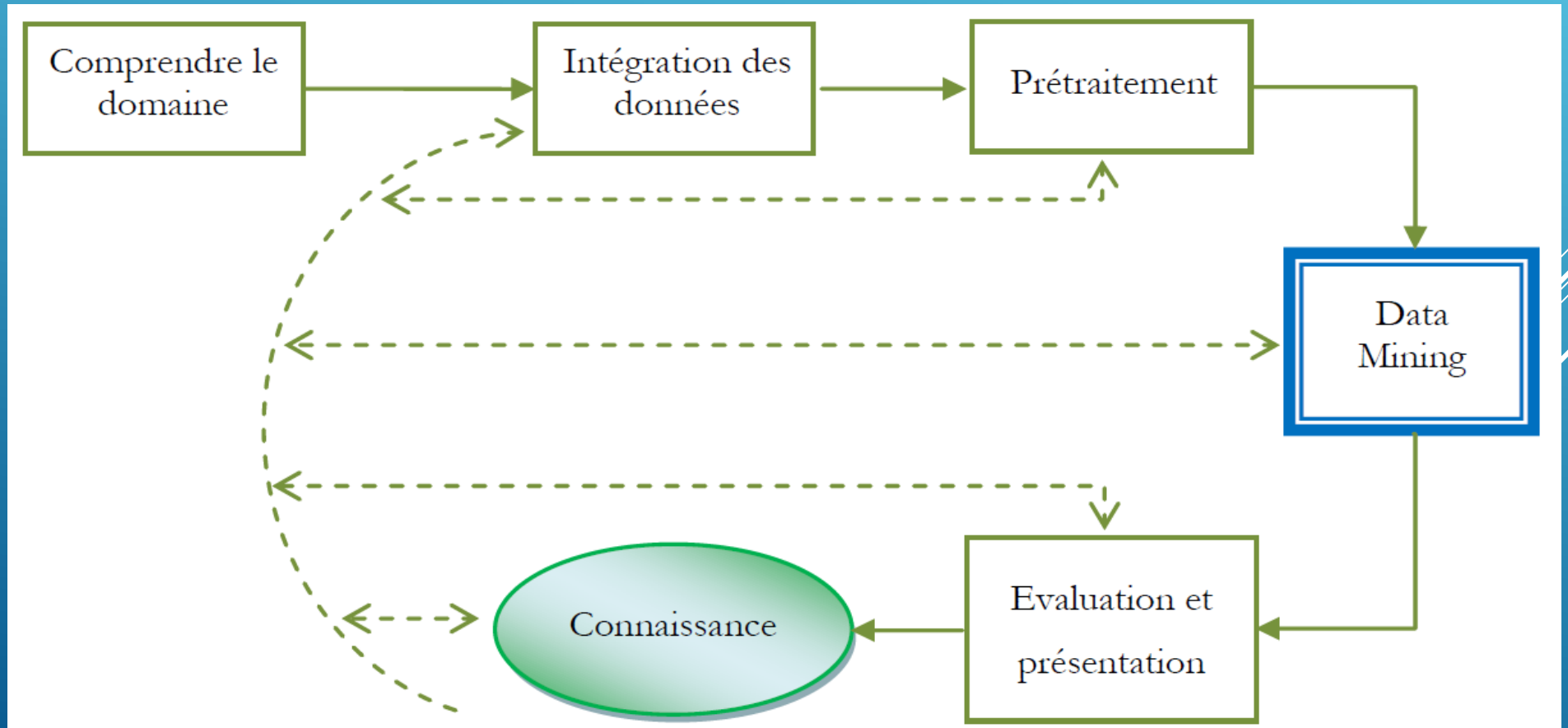
- **Web : les moteurs de recherche** utilisent des techniques de Data Mining, ce que l'on comprend aisément étant donné les volumes de données traités (rappel : 2 000 000 recherches/minute), ce que l'on appel fouille du web.

- **Ect…**

1. Qu'est ce que la fouille de données ?

1.3 Processus du data mining:

Le processus KDD, est composé de 5 étapes, n'est pas linéaire, on peut avoir besoin de revenir à des étapes précédentes pour corriger ou ajouter des données.



1. Qu'est ce que la fouille de données ?

1.3 Processus du data mining:

1. Définition et compréhension du problème :

La compréhension de problème, permet aux algorithmes de donner un résultat fiable. En effet, Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus.

Généralement, le data mining est effectué dans un domaine particulier (banques, médecine, biologie, marketing, ...etc) où la connaissance et l'expérience dans ce domaine jouent un rôle très important dans la définition du problème, l'orientation de l'exploration et l'explication des résultats obtenus.

1. Qu'est ce que la fouille de données ?

1.3 Processus du data mining:

2. Intégration des données

Cette étape permet de regrouper et de mettre en forme des données, collectées à partir de diverses origines, au sein d'une seule et même base de données. Les données peuvent provenir de différents systèmes de gestion de bases de données, de fichiers textes… Les données rassemblées, au sein d'un entrepôt de données, sont également nettoyées et codées selon un système uniforme.

Le but des opérations d'**intégration** et de **nettoyage** est de générer des entrepôts de données et/ou des magasins de données spécialisés contenant les données retravaillées pour faciliter leurs exploitations futures.

1. Qu'est ce que la fouille de données ?

1.3 Processus du data mining:

3. Prétraitement des données :

Les données collectées doivent être **préparées**. Avant tout, elles doivent être **nettoyées** puisqu'elles peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou l'éliminer complètement.

Le prétraitement comporte aussi la réduction des données qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Une méthode largement utilisée, est l'analyse en composantes principales (ACP).

1. Qu'est ce que la fouille de données ?

1.3 Processus du data mining:

4. Data Mining

Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, le clustering, ... Sont utilisées. Il est possible de définir la qualité d'un modèle en fonction de critères comme les **performances** obtenues, la **fiabilité**, la **compréhensibilité**, la **rapidité** de construction et d'utilisation et enfin l'**évolutivité**. Tout le problème du Data Mining réside dans le choix de la méthode adéquate à un problème donné. Il est possible de combiner plusieurs méthodes pour essayer d'obtenir une solution optimale.

1. Qu'est ce que la fouille de données ?

1.3 Processus du data mining:

5. Evaluation et présentation

Cette phase est constituée de l'évaluation, qui mesure l'**intérêt** des motifs extraits, et de la présentation des résultats à l'utilisateur grâce à différentes techniques de visualisation.

Cette étape est dépendante de la tâche de Data Mining employée. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des **interprétations** des modèles obtenus, bien que l'interaction avec l'expert soit importante pour que les motifs soient validés par ces experts du domaine.

2. Qu'est ce qu'une donnée ?

Le Data Mining n'est pas spécifique à un type de médias ou de données. Il est applicable à n'importe quel type d'information. Le Data Mining est utilisé et étudié pour les **Bases de Données** incluant les Bases de Données relationnelles et les Bases de Données Orientées-Objets, les data **warehouses**, les Bases de Données transactionnelles, les supports de données non structurés et semi-structurés comme le World Wide Web, les Bases de Données avancés comme les Bases de Données spatiales, les Bases de Données multimédia, les Bases de données de séries temporelles et les Bases de Données textuelles et même fichiers plats.

2. Qu'est ce qu'une donnée ?

2.1 Les fichiers plats

Les fichiers plats sont actuellement la source de données la plus commune pour les algorithmes du Data Mining et particulièrement dans le niveau de recherches. Les fichiers plats sont des fichiers de données simples dans le format texte ou binaire avec une structure connue par l'algorithme du Data Mining qui va –t– être y appliqué.

2. Qu'est ce qu'une donnée ?

2.2 Les bases de données relationnelles

Les algorithmes du Data Mining appliqués sur des Bases de Données relationnelles sont plus polyvalents que les algorithmes spécifiquement faits pour les fichiers plats puisqu'ils peuvent profiter de la structure inhérente aux bases de données relationnelles. Le Data Mining peut profiter du SQL pour la sélection, la transformation et la consolidation, il passe au-delà de ce que le SQL pourrait fournir, comme la prévision, la comparaison, la détection des déviations, etc.

2. Qu'est ce qu'une donnée ?

2.3. Les Data Warehouses (Les entrepôts de données)

Data Warehouse est un support de données assemblées de multiples sources de données (souvent hétérogènes) et est destinée à être utilisé dans l'ensemble sous le même schéma unifié. Autrement dit, les données de différents magasins peuvent être chargées, nettoyées, transformées et intégrées ensemble. Pour faciliter la prise de décisions et les vues multidimensionnelles, les Data Warehouses sont souvent modelées par une structure de données multidimensionnelle.

2. Qu'est ce qu'une donnée ?

2.4 Les bases de données transactionnelles

En général, une Base de Données transactionnelle est un fichier où chaque enregistrement représente une transaction. Une transaction contient un identifient unique de transaction (*transactionID*) et une liste d'items composant la transaction (les achats d'un client lors d'une visite). Les bases de données transactionnelles peuvent contenir d'autres informations tels que la date de la transaction, l'identifient du consommateur, l'identifient de la personne qui a vendu, et ainsi de suite.

2. Qu'est ce qu'une donnée ?

2.5 Les bases de données multimédia

Les bases de données multimédia comportent des documents sonores, des vidéos, des images et des médias en textes et audio. Elles peuvent être stockées sur des bases de données orientées objets ou objets relationnelles ou simplement sur un fichier système. Le multimédia est caractérisé par sa haute dimension ce qui rend le datamining sur ce type de données très difficile. Le data mining sur les supports des multimédias requiert exige la vision par ordinateur, l'infographie, l'interprétation des images et les méthodologies de traitement de langages naturels .

2. Qu'est ce qu'une donnée ?

2.6 Les bases de données spatiales

Ce sont des bases de données, qu'en plus de leurs données usuelles, elles contiennent des informations géographiques comme les cartes et les mondiaux ou régionaux. De telles bases de données présentent de nouveaux défis aux algorithmes de data Mining.

2. Qu'est ce qu'une donnée ?

2.7 Les bases de données de séries temporelles

Les bases de données de séries temporelles contiennent des données relatives au temps, comme les données du marché boursier ou les activités enregistrées. Ces bases de données ont couramment un flux continu de nouvelles données entrantes, qui parfois rend l'analyse en temps réel un besoin exigeant. Le data mining pour ce genre de bases de données est généralement l'étude des tendances et des corrélations entre les évolutions des différentes variables, aussi bien que la prédiction des tendances et des mouvements des variables par rapport au temps.

2. Qu'est ce qu'une donnée ?

2.8 Le World Wide Web

Le World Wide Web est le support de données le plus hétérogène et le plus dynamique disponible. Un grand nombre d'auteurs et d'éditeurs contribuent sans arrêt à son accroissement et évolution, et chaque jour un énorme nombre d'utilisateurs accède à ses ressources. Les données dans le World Wide Web sont organisées dans des documents interconnectés. Ces documents peuvent être des textes, audio, vidéos, données brutes et même des applications.

2. Qu'est ce qu'une donnée ?

2.8 Le World Wide Web

Conceptuellement, le World Wide Web est composé de trois grands composants : le contenu du Web, qui englobe les documents disponibles ; la structure du Web, qui garantie les hyperliens et les relations entre documents ; et l'usage du Web, en décrivant quand et comment les ressources seront accédées. Une quatrième dimension peut être ajoutée concernant la nature dynamique ou l'évolution des documents.

Le web mining, essaie d'aborder toutes ces questions et il est souvent divisé en contenu Web mining, la structure Web mining et l'usage Web mining.