Introduction générale au processus de Data Mining (Fouille de Données)

1.1 Historique et motivations du Processus du Data Mining :

La vision classique de la notion de traitement de l'information reposait sur l'équation :

Données + Traitements ➔ Résultats

Suite aux traitements, seuls les résultats comptaient.

Avec:

- l'explosion des quantités d'informations stockées
- le progrès important des vitesses de traitement et des supports de stockage.

est apparu le paradigme de la fouille des données (DataMining). Ce paradigme repose sur la constatation que des informations utiles se cachent dans les données. D'où:

Données+ Traitements (DM) ➔ Nouvelles informations

Plus spécifiquement, la fouille de données est un processus qui, étant donné un ensemble de données de grande taille, vise à:

- découvrir des connaissances cachées qui peuvent aider à comprendre ces données, c.a.d : Comprendre le comportement actuel des données.
- à prédire le comportement des données futures.

Historiquement, le DM peut être considéré comme une évolution naturelle de la technologie de traitement de l'information. En effet, depuis les années 60, la

technologie de l'information à évolué des simples systèmes de gestion de fichiers à des SGBD extrêmement puissants et utiles.
Avec le temps, de nombreuses compagnies et organismes possédant des filiales et agences géographiquement réparties, il est devenu pénible pour les décideurs de ces organismes de prendre des décisions sur la base de l'ensemble des données réparties sur des BDDs opérationnelles propre chacune à une filiale.
Le besoin de créer une base unifiée qui servira à l'extraction de connaissance et règles utiles pour une gestion rationnelle de toute la compagnie donna naissance au concept de DataWarHouse (entrepôt de données) et aux techniques de WareHousing (Gestion des entrepôts des Données). Brièvement, cette technologie consiste à intégrer toutes les BDDs partielles de la compagnie en une seule BDD unifiée et homogène.
Le caractère homogène de la BDD unifiée est justifié par la constatation que les BDDs partielles sont en général non homogènes en termes de version, SGBD, etc.

Avec la croissance continue des données collectées, l'exploitation des entrepôts devint difficile et fastidieuse. Le paradigme du DataMining, consistant en l'exploration des données gigantesques pour en extraire l'essentiel, vit alors le jour au début des années 90. Ce paradigme stipule que :

'Des connaissances utiles sont enfuies dans les données'.

1.2 Typologie des processus de Data mining en fonction la nature des données :
Le domaine du data mining a évolué dans le temps et s'est vu adapté aux exigences de la nature des données visées. Nous présentons ci-après des définitions des types de data mining existants :

a. Définition : Data Mining des des données Structurées :

C'est ce type de DM qui a le premier vu le jour et, de ce fait, est plus populaire et plus connu. Nous parlons ici du DM des données contenues dans les BDDs des sociétés et organismes. Autrement dit, le DM des *données structurées*. Ce type de fouille de données fut particulièrement motivé par le cas célèbre du « Panier de la ménagère » ou « Tickets de Caisse ». Ce cas particulier s'intéressa, historiquement, à extraire des connaissances utiles pour mieux maitriser la gestion des supermarchés américains.

b. Définition : Data Mining des données Non-Structurées (Multimédia):

Avec temps, le paradigme du DM se propagea à d'autres types de données, notamment:

- Les images
- Le texte
- Les signaux et time series (séries temporelles)
- Les vidéos
- Les données audio (voix, parôle, son, musique, etc.)
- Les Pages WEB et les réseaux sociaux
- Les séquences de données Biologiques (ADN)
- Etc.

C'est-à-dire, des données :

- Non caractérisées
- Plus complexes par leurs contenus pour la tâche de caractérisation
- Non structurées dans des BDDs
- Plus volumineuses

Ce type de Data mining implique aussi le Big data (voire plus loin).

Ce genre de données impose alors la difficile et complexe problématique de nécessité de caractérisation (Extraction de caractéristiques ou attributs des données sous forme de diverses nature afin de rendre le processus du data mining de ces données possible).

NB: Il faut bien comprendre ici le sens de 'Non structuré'. Il est évident qu'une image, un texte, une vidéo, etc., possèdent tous des structures, d'ailleurs très simples. De même, il existe des BDDs d'images, de vidéos, etc. (BDD MultiMedia). Si on voulait exploiter un fond d'images ou d'audio, à titre d'exemple, sous une BDD, nous pouvons les décrire par des attributs les caractérisant, sans exploiter le contenu:

- Code
- Titre
- Auteur
- Date de création
- Mots clés

Etc.

Dans ce cas, nous revenons au premier cas (DM des BDDs)

Ce qui est entendu ici par non structuré, c'est le contenu même de ces données complexes. Les caractéristiques sont alors extraites du contenu même des données. Nous parlons alors de fouille de données sur la base de leurs contenus (Content Based Data Mining) et non plus de descripteurs externes.

Ces caractéristiques sont extraites sur la base de la nature des données :

- Pour l'image, à partir des couleurs, les formes, les textures, les objets, etc.
- Pour les signaux et séries temporelles, à partir de leurs contenus en termes de valeurs des échantillons (domaine temporelle), ou coefficient de

Fourrier (contenu fréquentiel) ou les coefficients d'ondelettes (domaine temps-fréquence).

- Etc.

Pour mieux fixer les idées, voici maintenant l'évolution de la notion de Data Mining, à travers des définitions exprimées à des dates éloignées:

1.3 Définitions du Data Mining :

a. Définition1 [Fayyad et al., 1995] : Le DM est un processus itératif et interactif, par lequel on extrait des connaissances :

- Nouvelles: c.a.d, non-triviales et non connues au préalable
- Utiles: c.a.d, permettant de prendre des décisions.
- Compréhensible: c.a.d, de présentation simple.
- Valide dans le temps : c.a.d, valide actuellement et dans une période future appréciable.

Concernant la nature de ce processus :

- Itératif: Veut dire que le Data Mining nécessite généralement plusieurs passes.
- Interactif: Veut dire que l'utilisateur est dans la boucle de ce processus.

Par utilisateur, il faut comprendre aussi bien l'expert du Data mining que l'expert du domaine ou de l'activité considérée.

b. Définition2 [TUFFERY, 2014]: Le data mining est l'ensemble des méthodes scientifiques destinées à l'exploration et l'analyse de grandes bases de données informatiques en vue de détecter dans ces données

- des profils-type (pattern, motifs, régularité)
- des comportements récurrents (répétitifs, périodiques)
- des règles régissant ces données
- des tendances inconnues (non fixées a priori),

- des structures particulières restituant de façon concise l'essentiel de l'information utile pour l'aide à la décision.

Tous ces éléments constituent de la "connaissance" extraites des données.

1.4 Outils du DM:

Le datamining utilise/combine des outils de :

a. Statistiques: Exemple: Moyenne, variance, covariance, corrélation
b. Intelligence artificielle: Exemple: arbre de décision, systèmes à base de règles,
c. Reconnaissance des formes : Exemple: classification, clustering,
d. Analyse des données: Techniques de réduction des données Exemple: ACP: Analyse par composantes principales,
e. Recherche de l'Information : Exemple: Techniques d'indexation et mesures de distance et mesures de similarité
f. Et Autres domaines plus spécifiques : Exemple: Techniques de visualisation des données.

1.5 Définition ECD (Extraction des connaissances dans les données), KDD (Knowledge Discovery in Data)

Le processus du DM est au fait une étape d'un processus plus grand : L'ECD. Voici un schema générique pour ce procéssus.

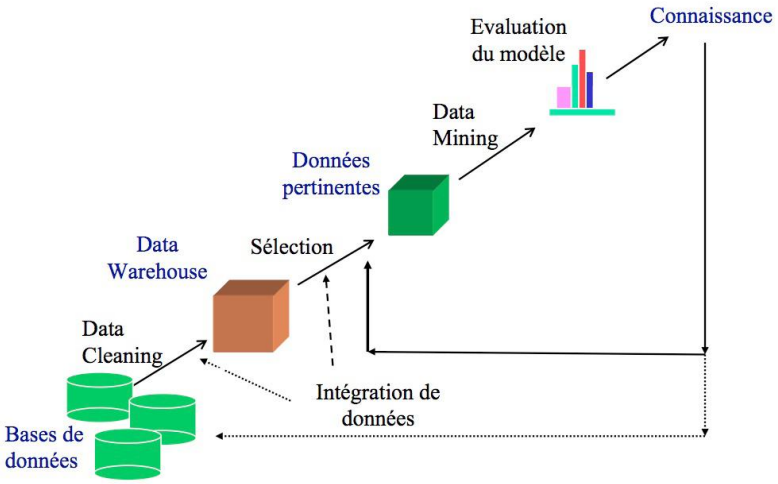

Fig.1.1 Processus du KDD - DataMining

De façon résumée, le processus du KDD comprend les étapes suivantes:

a. Nettoyage des données,
b. Integration des données,
c. Sélection des données à analyser/modéliser

Remarque : Ces étapes invoquent la notion de data warehousing (entreposage des données) qui sera présentée plus loin.

d. Application des techniques du DM, selon le besoin
e. Analyse des résultats obtenus.

Il va de soit que des

- itérations dans les différentes phases sont prévues afin de mieux régler le modèle.
- L'utilisateur (expert du data mining et expert du domaine) est impliqué dans cette phase.

1.6 Entrepôt de données (Data Warehouse) et technologie de l'entreposage (Data Warehousing) :

a. Définition : Entrepôt de données: Un entrepôt de données est une base de données regroupant plusieurs bases de données opérationnelles relatives à un organisme particulier. Cette base contiendra toutes données émanant des bases de données de l'organisme de façon homogène et unifiée afin de pouvoir effectuer les étapes du DM dessus.

b. Technologie de Gestion des Entrepôts de données: DataWarhousing Technology :

Cette partie du KDD inclut:

- Nettoyage des données - data cleaning : Elimination des données aberrantes, contradictoires, etc.
- Integration des données - data integration: Obtention d'une BDD unifié à partir de plusieurs BDD opérationnelles (hétérogènes).
- Sélection de données – Data selection: Seule un échantillon des données approprié est choisi et utilisé pour établir le modèle du DataMining.
- Analyse et traitement en-ligne - on-line analytical processing (OLAP), techniques d'analyse des données incluant les fonctionnalités de:
  - Résumé des données – data summarization,
  - Transformation des données (standardisation, etc.) en vue de l'étape du DataMining
  - Visualisation des données: Outils permettant d'apprécier visuellement la répartition des données.

1.7 Big Data : Suite au succès du processus de fouille sur les données structurées (fouille initiale, dans ce cours), il y eu extension de ce paradigme sur

les données multimédia. La fouille du multimédia constitua alors la deuxième forme de fouille de données (dite aussi, fouille des données non-structurées ou complexes). Cependant, et, suite à l'avènement des réseaux sociaux, caractérisés par une croissance (des tailles de données) vertigineuse et non préalablement imaginée, les algorithmes de fouille développés pour les deux premiers cas sont devenus obsolètes : Un troisième type de fouille est alors né : Le Big Data. Ce type de fouille est caractérisé par la règle des 3V puis des 5V.

V1 : Volume (des données) : Le Big Data traite des données de tailles extrêmement grandes et qui croissent en taille.
V2 : Vitesse : La vitesse de croissance des données doit être elle-même assez importante. Ce qui implique aussi que la vitesse nécessaire de traitement de ces données) doit aller encore plus vite pour pouvoir appréhender la croissance des données.
V3 : Variété : Au moins deux types de média sont exigés
V4 : Véracité : C'est la crédibilité accordée dans les données et leurs sources.
V5 : Valeur : C'est la valeur ajoutée issue du processus de Big Data, en termes de connaissances extraites des données.

1.8 Exemples de connaissances extraites dans des cas pratiques:

Le processus du Data Mining, sous toutes ses formes, est applicable dans divers domaines. Nous donnons quelques exemple, ci-après :

- Organisme de crédit : Il s'agit d'accorder ou non (décision) un crédit en fonction du profil du client, de sa demande, et des expériences passées de prêts ;
- Optimisation du nombre de places dans les avions, hôtels, en fonction des périodes, régions, etc.;

– Organisation des rayonnages dans les supermarchés en regroupant les produits qui sont généralement achetés ensemble (facilitation de l'achat, mais aussi incitation à l'achat).

Dans ce cadre, une règle célèbre extraite par le processus du DM est : "Les clients qui achètent le produit X, achètent généralement aussi le produit Y" ;

– Diagnostic médical : Dans le domaine médical, une règle célèbre extraite par le processus du Data Mining est : "Les patients travaillant dans telles et telles conditions et fumeurs développent couramment telle pathologie" ;

– Classification des emails selon des classes prédéfinies ou issues par apprentissage : Exemple : Messages SPAM, Message Non-SPAM, Message incluant des termes inappropriés, etc.

– Moteur de recherche sur internet : Dans ce cadre, le Data Mining est utilisé pour fouille du web afin de : Catégoriser les sites, apprendre les profils des utilisateurs, ce qui facilite l'adaptation du contenu, des suggestions et de la publicité, etc.

– Fouille du texte, en général: Catégorisation, résumé.

– Fouille des séries temporelles (time series mining) : Clustering, classification, détection d'anomalies, etc. Comme exemple de séries temporelles, nous avons : Les signaux physiologiques (ECG, EEG, Capnogramme, etc.).

1.9 Types de tâches du processus de fouille des données

Il existe principalement deux types de tâche de data mining:

a. Les méthodes *descriptives* (ou exploratoires des données) :

Ces méthodes visent à mettre en évidence des connaissances présentes dans les données, mais cachées par le volume de ces données.

Exemple:

- *segmentations* des données

- *recherches d'associations* de type X ➔Y (si X alors Y)

Ce genre de tâches a pour rôle la Réduction/Résumé des données. Elles seront présentées plus loin dans ce cours.

NB: Dans cette catégorie de méthodes du DM, il n'y a pas de variable à expliquer.

b. Les méthodes *predictives* (ou explicatives des données):

Ces méthodes visent à déduire de nouvelles informations à partir des informations présentes. Donc, ces méthodes expliquent les données.

Exemple:

- Classification d'un individu nouveau dans une classe parmi plusieurs existantes (Exemple: ClientDépensif, ClientABudget)
- Scoring: Accorder, OUI/NON, un prêt à un client,
- Prédire les revenus sur un investissement, etc.

Fin du Chapitre

Références Utiles/Utilisées :

[D'Aubigny, 2001] Gérard D'Aubigny, Discussion et commentaires. Data mining et ,(statistique, Journal de la société française de statistique, tome 142, no 1 (2001 p. 37-52.

[Cugliari, 2015] Jairo Cugliari, Fouille de Données, Master 2 IDS-Kharkiv, S2 2014-2015, Université Lumière-Lyon2.

[Sumathi & Sivanandam,] Introduction to Data Mining and its Applications, Studies in Computational Intelligence, Volume 29, Springer-Verlag Berlin Heidelberg 2006.

[Preux, 2011] Ph. Preux, Fouille de données, Notes de cours, Université de Lille 3.

[Tuffery, 2014] Stéphane Tuffery, Cours de data mining, M2 Ingénierie économique et financière, Université Rennes 1, 2014.

[Lieber, 2007] Jean Lieber, Fouille de données : Notes de cours, 2007.

[Zaki & Meira, 2014] Mohammed J. Zaki & Wagner Meira Jr, Data mining and analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.