CHAPITRE 1

FOUILLE DE DONNEE (DATA-MINING)

1.1 INTRODUCTION :

L'exploration de données connue aussi sous l'expression de fouille de données, forage de données, prospection de données, data-Mining, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.
Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances.

L'utilisation industrielle ou opérationnelle de ce savoir dans le monde professionnel permet de résoudre des problèmes très divers, allant de la gestion de la relation client à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation de sites web. C'est aussi le mode de travail du journalisme de données [1].

L'exploration de données fait suite, dans l'escalade de l'exploitation des données de l'entreprise, à l'informatique décisionnelle. Celle-ci permet de constater un fait, tel que le chiffre d'affaires, et de l'expliquer comme le chiffre d'affaires décliné par produits, tandis que l'exploration de données permet de classer les faits et de les *prévoir* dans une certaine mesure ou encore de les éclairer en révélant par exemple les variables ou paramètres qui pourraient faire comprendre pourquoi le chiffre d'affaires de tel point de vente est supérieur à celui de tel autre. [1]

1.1.1 Qu'est ce que la fouille de données

La fouille de données consiste à rechercher et extraire de l'information (utile et inconnue) de gros volumes de données stockées dans des bases ou des entrepôts de données.

Le développement récent de la fouille de données (depuis le début des années 1990) est lié à plusieurs facteurs : une puissance de calcul importante est disponible sur les ordinateurs de bureau ou même à domicile, le volume des bases de données augmente énormément, l'accès aux réseaux de taille mondiale, ces réseaux ayant un débit sans cesse croissant, qui rendent le calcul distribué et la distribution d'information sur un réseau d'échelle mondiale viable ; la

prise de conscience de L'intérêt commercial pour l'optimisation des processus de fabrication, vente, gestion, logistique.

La fouille de données a aujourd'hui une grande importance économique du fait qu'elle permet d'optimiser la gestion des ressources (humaines et matérielles). Elle est utilisée par exemple :

- Organisme de crédit : pour décider d'accorder ou non un crédit en fonction du profile du demandeur de crédit, de sa demande, et des expériences passées de prêts;
- Optimisation du nombre de places dans les avions, hôtels, ... ⇒ sur- réservation.
- Organisation des rayonnages dans les supermarchés en regroupant les produits qui sont généralement achetés ensemble (pour que les clients n'oublient pas bêtement d'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 en fin de semaine, pendant l'été, achètent généralement également le produit Y ≫;
- Organisation de campagne de publicité, promotions, ... (ciblage des offres) diagnostic médical : ≪les patients ayant tels et tels symptômes et demeurant dans des agglomérations de plus de 104 habitants développent couramment telle pathologie≫;
- analyse du génome et bioinformatique plus généralement.
- Classification d'objets (astronomies, ...)
- commerce électronique, recommandation de produits
- analyser les pratiques et stratégies commerciales et leurs impacts sur les ventes.
- moteur de recherche sur internet : fouille du web extraction d'information depuis des textes : fouille de textes
- évolution dans le temps de données : fouille de séquences

Le processus complet de fouille de données comprend plusieurs étapes :

- Collecte des informations et organisation de ces infos dans une base de données.
- nettoyage de la base de données : attributs sans valeur, ayant une valeur invalide (bruit), normalisation ;
- Sélection des attributs utiles ;
- Extraction d'information d'une base de données (Knowledge Discovery in Databases, ou KDD);
- Visualisation des données : histogramme, camembert, arbre, visualisation 3D et plus généralement, exploration interactive de données ;
- Evaluation des résultats de l'extraction de connaissance [2].

1.2 Les tâches du DataMining :

Une multitude de problèmes d'ordre intellectuel, économique ou commercial peuvent être regroupés, dans leur formalisation, dans l'une des tâches suivantes :

- **La classification** : La classification consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini
- **L'estimation :** Contrairement à la classification, le résultat d'une estimation permet d'obtenir une variable continue. Celle-ci est obtenue par une ou plusieurs fonctions combinant les données en entrée. Le résultat d'une estimation permet de procéder aux classifications grâce à un barème.
- **La prédiction** : La prédiction ressemble à la classification et à l'estimation mais dans une échelle temporelle différente. Tout comme les tâches précédentes, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé.
- **La Segmentation** : La segmentation consiste à segmenter une population hétérogène en sous populations homogènes. Contrairement à la classification, les sous-populations ne sont pas préétablies.
- **La description** : C'est souvent l'une des premières tâches demandées à un outil de Data-Mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent une exploitation supplémentaire en vue de fournir des explications.
- **L'optimisation :** Pour résoudre de nombreux problèmes, il est courant pour chaque solution potentielle d'y associer une fonction d'évaluation. Le but de l'optimisation est de maximiser ou minimiser cette fonction. Quelques spécialistes considèrent que ce type de problème ne relève pas du Data-Mining [3].

1.3 Les techniques de DataMining :

Plusieurs techniques peuvent être inscrites dans le contexte du Data-Mining, on en cite :

- **Les arbres de décision** : ce sont des outils très puissants principalement utilisés pour la classification, la description ou l'estimation. Il s'agit d'une représentation graphique sous forme d'arbre de décision représentant un enchaînement hiérarchique de règles logiques qui permet de diviser la base d'exemples en sous-groupes.

- **Les réseaux de neurones** : inspirés de la biologie, les réseaux de neurones représentent une transposition simplifiée des neurones du cerveau humain. Ils sont utilisés dans la prédiction et la classification.
- **Les algorithmes génétiques** : ce sont des méthodes d'optimisation de fonctions basée sur les mécanismes génétiques pour élaborer des paramètres de prévision des plus optimaux. Ils permettent de résoudre des problèmes divers, notamment d'optimisation, d'affectation ou de prédiction.
- **Les règles d'association** : Ce sont les méthodes les plus répandues dans le domaine de marketing et de la distribution. Elles peuvent être appliquées dans différents secteurs d'activité pour lesquels, il est intéressant de trouver des groupements d'articles qui apparaissent le plus fréquemment ensembles, et de générer des règles d'associations. Le but principal de cette technique est descriptif ou prédictif.
- **Les plus proches voisins (CBR, Case BasedReasoning)** : C'est une méthode dédiée à la classification qui peut être étendue à des tâches d'estimation. De façon similaire, le « raisonnement basé sur la mémoire », technique du Data-Mining dirigé, permet de classer ou de prédire des données inconnues à partir d'instances connues.
- **Le Clustering** : C'est une classification non supervisée, les classes possibles et leur nombre ne sont pas connus à l'avance et les exemples disponibles sont non étiquetés.

Le but est donc de découvrir des relations intéressantes qui peuvent exister Implicitement entre les données et qui permettront de regrouper dans un même groupe ou cluster les objets considérés comme similaires. Nous nous intéresserons dans nos travaux à la technique des règles d'associations. Nous détaillerons plus loin dans ce chapitre les grandes lignes de cette technique [3].

1.4 Qu'est ce qu'une donnée ?

1.4.1 Notations :

On notera X un ensemble de données. Chaque donnée est décrite par une ensemble A d'attributs. Chaque attribut a ϵ A prend sa valeur dans un certain ensemble de valeurs V_a. Ainsi, on peut considérer l'ensemble des données x dont les coordonnées balayent toutes les valeurs possibles des attributs : c'est l'espace des données que nous noterons D. Si l'on note a_1, ..., a_P les P attributs, $D = V_{a1} x\ V_{a2} x ::: x V_{aP}$. Toute donnée appartient à cet ensemble et on a X appartient à D. Il est souvent utile d'avoir une représentation géométrique de l'espace des données ; chaque attribut correspondant à un axe de coordonnées. S'il y a P attributs, l'espace des données est un espace euclidien à P dimensions [2].

1.4.2. Les différentes natures d'attributs :

Une donnée est un enregistrement au sens des bases de données, que l'on nomme aussi << individu >> (terminologie issue des statistiques) ou<<instance>> (terminologie orientée objet en informatique) ou même <<tuple>> (terminologie base de données) et << point >> ou <<vecteur>> parce que finalement, d'un point de vue abstrait, une donnée est un point dans un espace euclidien ou un vecteur dans un espace vectoriel. Une données est caractérisée par un ensemble de <<champs>>, de <<caractères>>, ou encore d'<<attributs>> (en suivant les 3 terminologies précédemment évoquées : bases de données, statistiques et conception orientée objet).

Un attribut peut être de nature qualitative ou quantitative en fonction de l'ensemble des valeurs qu'il peut prendre. Un attribut est qualitatif si on ne peut pas en faire une moyenne ; sa valeur est d'un type défini en extension (une couleur, une marque de voiture, ...).
Sinon, l'attribut est de nature quantitative : un entier, un réel, ... ; il peut représenter un salaire, une surface, un nombre d'habitants, ... On peut donc appliquer les opérateurs arithmétiques habituels sur les attributs quantitatifs, ce qui n'est pas le cas des attributs qualitatifs.

Un attribut peut également être un enregistrement (une date par exemple),donc composé lui-même de sous-attributs (jour, mois, année dans le cas d'une date), sur lequel on peut définir les opérateurs arithmétiques habituels :donc quantitatif ne signifie pas forcément <<numérique>> et, réciproquement, numérique ne signifie pas forcément quantitatif : un code postal est numérique, mais pas quantitatif[2].

1.4.3 Les différentes natures de valeur d'attribut :

Il n'est pas inutile ici de consacrer quelques lignes à ce qu'est la valeur d'un attribut. Naturellement, cette valeur est censée représenter une certaine mesure attribut nominal : valeurs d'une quantité dans le monde. Ainsi, quand on dit qu'une couleur est <<bleue>>, non ordonnées cela signifie que nous en avons une certaine perception visuelle qui est associée à ce que, en français, on désigne par le mot <<bleu>>; elle aurait pu être verte et on l'aurait appelée verte. Il est a priori impossible de comparer bleu et vert ; ce sont deux couleurs, un point c'est tout : la couleur est un attribut nominal. Sion dit qu'aujourd'hui, la température est de 20°C et qu'hier, il faisait 18°C, on peut dire que la température est plus élevée aujourd'hui qu'hier : cette fois-ci, on peut comparer les deux valeurs d'attributs, cela a un sens. Mais, si on se rappelle ses cours de physique, on sait bien que ce 20 et ce 18 sont aussi arbitraires que les mots <<bleu>> et <<vert>> : ces valeurs dépendent, notamment, des unités de mesure : la température est un attribut ordinal. Maintenant, si on dit que le nombre d'enfants de Paul est 2 et que Jacques a 3 enfants, d'une part onpeut bien affirmer que Jacques a plus d'enfants que

Paul et, de plus, ces deux nombres 2 et 3 ne sont pas arbitraires : le nombre d'enfants est un attribut absolu.

Au-delà de la distinction qualitatif/quantitatif, on voit donc apparaître des distinctions plus subtiles entre des attributs dont les valeurs sont arbitraire incomparables (attribut nominal), des attributs dont la valeur est arbitraire mais que l'on peut comparer (attribut ordinal) et des attributs dont la valeur n'est pas arbitraire (attribut absolu).

Ces différentes natures entraînent le fait que les opérations que l'on peut faire sur ces attributs ne sont pas les mêmes : on ne peut pas soustraire deux couleurs, on peut soustraire deux températures mais on ne peut pas en faire le rapport, alors que l'on peut soustraire et faire le rapport du nombre d'enfants de deux personnes.

Quand on fait des statistiques ou de la fouille de données, on effectue de nombreuses opérations arithmétiques ; hors, selon la nature de l'attribut, ces opérations sont licites ou non, Il importe donc de ne pas faire n'importe quel calcul, d'appliquer n'importe quel algorithme sans prendre garde aux attributs sur lesquels on les effectue.

Par ailleurs, il faut observer un principe d'indépendance du résultat des calculs par rapport aux unités de mesure dans lesquelles sont exprimées les valeurs des attributs : il n'y a aucune raison que l'information extraite d'une base de données change selon qu'une longueur est exprimée en millimètres, mètres ou années-lumière, De cette observation, on pose la règle suivante : on doit toujours s'arranger pour que le résultat d'une analyse ne dépende pas de l'unité de mesure [2].

1.4.4 Le bruit :

Il importe de ne pas faire comme si toutes les données ont une valeur connue,

et encore moins une valeur valide ; il faut donc gérer des données dont certains attributs ont une valeur inconnue ou invalide ; on dit que les données sont <<bruitées>>. La simple élimination des données ayant un attribut dont la valeur est inconnue ou invalide pourrait vider complètement la base de données. On touche le problème de la collecte de données fiables qui est un problème pratique très difficile à résoudre. En fouille de données, il faut faire avec les données dont on dispose sans faire comme si on disposait des valeurs de tous les attributs de tous les individus.

1.4.5 Différentes tâches d'extraction d'information

Dans certains problèmes (problème de classification supervisée), chaque donnée est affectée d'une caractéristique, par exemple une couleur. Supposons que l'ensemble des couleurs possibles soit fini et de faible cardinalité. Le problème de classification supervisée

consiste alors à prédire la couleur d'un point quelconque étant donné un ensemble de points colorés. Géométriquement, cela revient à trouver un moyen de séparer les points les uns des autres, en fonction de leur couleur. S'il n'y a que deux couleurs, un simple (hyper) plan peut suffire à les séparer ; ceux d'une certaine couleur sont d'un côté de l'hyper plan, les autres étant de l'autre côté. Dans ce cas, les points sont linéairement séparables (séparables par un objet géométrique qui ressemble à une droite, un hyperplan pour être plus précis au niveau du vocabulaire). Généralement, des points d'une couleur donnée se trouvent du mauvais côté de l'hyperplan.

Cela peut résulter d'erreurs dans l'évaluation des attributs (on s'est trompé en mesurant certains attributs, ou en attribuant sa couleur à la donnée) ; dans ce cas, les données sont bruitées. Cela peut aussi être intrinsèque aux données quine peuvent pas être séparées linéairement. Il faut alors chercher à les séparer avec un objet non hyper planaire. On peu citer diverses approches à ce problème :

- Construction d'un modèle arborescent permettant de prédire la classe d'une donnée ou d'un modèle exprimé sous forme de règles. Dans ces deux cas, le modèle obtenu est interprétable par un humain ;
- Estimation directe de la classe d'une donnée en fonction des exemples : une approche probabiliste et une approche par cas. Dans ces deux cas, il n'y a pas d'apprentissage, pas de construction de modèle ;
- Construction d'un modèle non interprétable par un humain : les réseaux de neurones qui permettent de s'attaquer à des problèmes dans lesquels les données sont décrites par de très nombreux attributs (des images par exemple), ou les machines à vecteurs supports. Dans ces deux cas, le modèle construit n'est pas interprétable par un humain ; il est utilisé par l'algorithme pour estimer la classe de nouvelles données.
- Construction d'un modèle par sélection de variables.

Pour revenir â notre exemple des points colorées, dans d'autres cas, l'ensemble des couleurs possibles est infini non dénombrables (une valeur réelle). On est alors dans un problème de régression.

Dans d'autres problèmes (problème de segmentation), on dispose d'un ensemble de points est la tâche consiste à repérer des groupes de points qui se ressemblent. On verra alors deux ensembles de techniques : les algorithmes qui proposent une segmentation des données puis des algorithmes qui construisent une représentation géométrique interprétable par un humain (dans un plan) d'un ensemble de données. C'est alors un humain qui détecte les groupes et

effectue la segmentation : l'algorithme de projection est une aide pour l'humain ; on effectue de la fouille de données assistée par ordinateur.

Plus généralement, on peut disposer d'un ensemble de données pour lequel son veut détecter des relations entre la valeur de leurs attributs. Il s'agît alors de chercher des règles d'association.

On peut aussi disposer d'un ensemble de points et d'autres points dont tous les attributs ne sont pas valués. On doit alors déterminer la valeur de ces attributs manquants. C'est alors un problème d'estimation.

On peut aussi disposer d'un ensemble de données qui décrit l'évolution au cours du temps d'une certaine entité. On veut alors prédire le comportement futur à partir du comportement observé : c'est un problème de prédiction.

Notons que dans tous ces problèmes, la notion de <<corrélation>> est omniprésente l'extraction d'information repose sur la recherche de corrélations entre des données. Ces corrélations peuvent être linéaires : c'est le cas simple. En général, on doit chercher des corrélations non linéaires.

Enfin, l'espace de données dans lequel celles-ci nous sont fournies initialement n'est pas forcément le plus adéquat. Il s'agît alors de préparer les données pour en faciliter l'extraction d'information. Cette préparation peut consister en la diminution du nombre d'attributs, à divers traitements sur les valeurs d'attributs (lissage, ...), au changement d'espace de représentation des données (projection dans un espace où les relations sont plus faciles à détecter, projection dans un sous-espace plus petit, ou représentation dans un espace obtenu par combinaison des attributs initiaux - création de features). La représentation des données dans un espace inadéquat entraîne lui-aussi du bruit, différent de celui rencontré plus haut.

1.4.6 Les étapes d'un processus d'Extraction de Connaissances à partir des données (ECD) :

Un processus d'ECD est constitué de quatre phases qui sont : le nettoyage et l'intégration des données, le prétraitement des données, la fouille de données et enfin l'évaluation et la présentation des connaissances.

La figure 1.1 récapitule ces différentes phases ainsi que les enchaînements possibles entre ces phases. Cette séparation est théorique, en pratique, ce n'est pas toujours le cas. En effet, dans de nombreux systèmes, certaines de ces étapes sont fusionnées [5].

1.4.6.1 *Nettoyage et intégration des données :*

Le nettoyage des données consiste à retravailler des données bruitées, soit en les supprimant, soit en les modifiant de manière à tirer le meilleur profit.

L'intégration est la combinaison des données provenant de plusieurs sources (base de données, sources externes, etc.…).Le but de ces deux opérations est de générer des entrepôts de données et/ou des magasins de données spécialisés contenant les données retravaillées pour faciliter leurs exploitations futures [5].

Exemple :

Soit l'exemple suivant qui présente une base de donnée d'un éditeur qui vend 5 sortes de magazines : sport, voiture, maison, musique et BD. Il souhaite mieux étudier ses clients pour découvrir de nouveaux marchés ou vendre plus de magazines à ses clients habituels.

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bemol	Rue du moulin, Paris	7/10/96	Voiture
23134	Bemol	Rue du moulin, Paris	12/5/96	Musique
23134	Bemol	Rue du moulin, Paris	25/7/95	BD
31435	Bodinoz	Rue verte, Nancy	11/11/11	BD
43342	Airinaire	Rue de la source, Brest	30/5/95	Sport
25312	Talonion	Rue du marché, Paris	25/02/98	NULL
43241	Manvussa	NULL	14/04/96	Sport
23130	Bemolle	Rue du moulin, Paris	11/11/11	Maison

Tableau 1.1 : la base de données avant le nettoyage

Intégrité de domaine : Dans notre exemple, la date d'abonnement des client 23130,31435 (11/11/11) semble plutôt correspondre à une erreur de saisie ou encore à une valeur par défaut en remplacement d'une valeur manquante ou Informations manquantes : Dans notre exemple, nous n'avons pas le type de magazine pour le client 25312. Il sera écarté de notre ensemble. L'enregistrement du client 43241 sera conservé bien que l'adresse ne soit pas connue.

Après le nettoyage on aura la base de donnée suivante :

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bemol	Rue du moulin, Paris	7/10/96	Voiture
23134	Bemol	Rue du moulin, Paris	12/5/96	Musique
23134	Bemol	Rue du moulin, Paris	25/7/95	BD
31435	Bodinoz	Rue verte, Nancy	11/11/11	BD
43342	Airinaire	Rue de la source, Brest	30/5/95	Sport
43241	Manvussa	NULL	14/04/96	Sport
23130	Bemolle	Rue du moulin, Paris	11/11/11	Maison

Tableau 1.2 : la base de données après le nettoyage [5]

1.4.6.2 Prétraitement des données

Il peut arriver parfois que les bases de données contiennent à ce niveau un certain nombre de données incomplètes et/ou bruitées. Ces données erronées, manquantes ou inconsistantes doivent être retravaillées si cela n'a pas été fait précédemment. Dans le cas contraire, durant l'étape précédente, les données sont stockées dans un entrepôt. Cette étape permet de sélectionner et transformer des données de manière à les rendre exploitables par un outil de fouille de données.

Cette seconde étape du processus d'ECD permet d'affiner les données. Si l'entrepôt de données est bien construit, le prétraitement de données peut permettre d'améliorer les résultats lors de l'interrogation dans la phase de fouille de données [5].

Exemple : soit la base de donnée nettoyée précédemment, la table 1.3 présente le résultat de prétraitement. Les clients qui ont des informations manquantes seront supprimés de la base.

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bemol	Rue du moulin, Paris	7/10/96	Voiture
23134	Bemol	Rue du moulin, Paris	12/5/96	Musique
23134	Bemol	Rue du moulin, Paris	25/7/95	BD
43342	Airinaire	Rue de la source, Brest	30/5/95	Sport

Tableau 1.3: la base de données après le prétraitement [5]

1.4.6.3 Fouille de données (Data-Mining)

La fouille de données (Data-mining en anglais), est le cœur du processus d'ECD. Il s'agit à ce niveau de trouver des pépites de connaissances à partir des données. Tout le travail consiste à appliquer des méthodes intelligentes dans le but d'extraire cette connaissance. Il est possible de définir la qualité d'un modèle en fonction de critères comme les performances obtenus, la fiabilité, la compréhensibilité, la rapidité de construction et d'utilisation et enfin l'évolutivité. Tout le problème de la fouille de données 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 globale [5].

1.4.6.4 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 fouille de données employée. En effet, bien que l'interaction avec l'expert soit importante quelle que soit cette tâche, les techniques ne sont pas les mêmes. Ce n'est qu'à partir de la phase de présentation que l'on peut employer le

terme de connaissance à condition que ces motifs soient validés par les experts du domaine. Il y'a principalement deux techniques de validation qui sont la technique de validation statistique et la technique de validation par expertise.

La validation statistique consiste à utiliser des méthodes de base de statistique descriptive. L'objectif est d'obtenir des informations qui permettront de juger le résultat obtenu, ou d'estimer la qualité ou les biais des données d'apprentissage. Cette validation peut être obtenue par :

- le calcul des moyennes et variances des attributs,
- si possible, le calcul de la corrélation entre certains champs,
- ou la détermination de la classe majoritaire dans le cas de la classification.

La validation par expertise, est réalisée par un expert du domaine qui jugera la pertinence des résultats produits. Par exemple pour la recherche des règles d'association, c'est l'expert du domaine qui jugera la pertinence des règles. Pour certains domaines d'application (le diagnostic médical, par exemple), le modèle présenté doit être compréhensible, une première validation doit être effectuée par un expert qui juge la compréhensibilité du modèle. Cette validation peut être, éventuellement, accompagnée par une technique statistique [5].

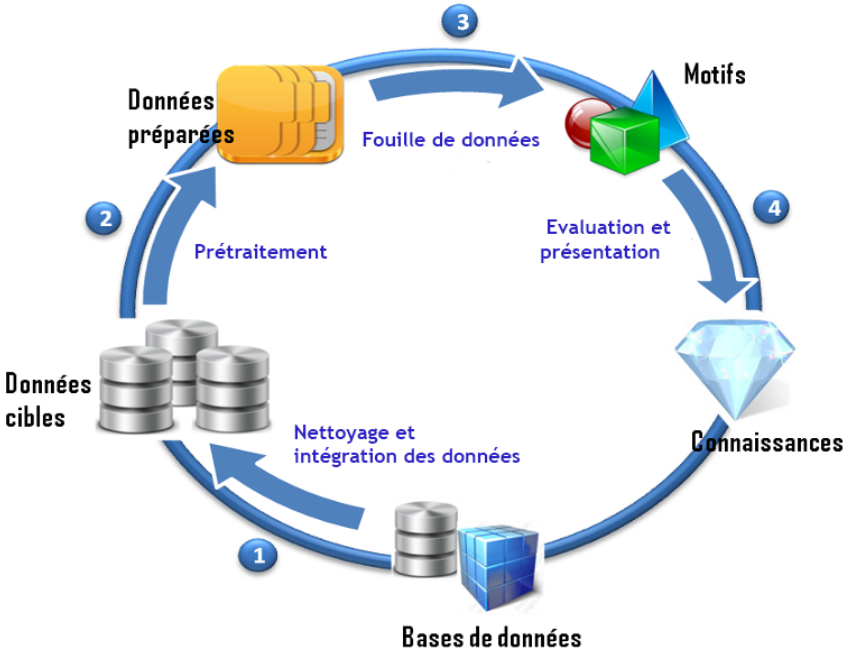


Figure 1.1 : processus d'extraction de connaissances à partir des données [5]

1.5 Conclusion :

Nous venons de présenter dans ce chapitre les concepts fondamentaux de la fouille de donnée (Data-Mining), Ainsi, le processus d'extraction des connaissances à partir des données (ECD).

Nous allons présenter dans le chapitre suivant quelques concepts concernant les règles d'association, et comment faire pour générer ces règles.