Chapitre 1

Présentation du Datamining

1 Qu'est-ce que le Datamining ?

Définition : Processus non trivial d'identification des structures inconnues, valides et potentiellement exploitables dans les bases de données.

Le datamining désigne l'ensemble des techniques et méthodes dans les domaines des statistiques, des mathématiques et de l'informatique qui permettent d'extraire d'un grand volume de données, des connaissances précises sur des éléments inconnus auparavant. Cette technique permet d'analyser et d'interpréter des données volumineuses, contenues dans une ou plusieurs bases de données afin de dégager des tendances. Le Datamining est en mesure de créer des catégories statistiques composées d'éléments similaires afin de proposer des hypothèses.

Le datamining est un procédé d'exploration et d'analyse de grands volumes de données en vue d'une part de les rendre plus compréhensibles et d'autre part de **découvrir des corrélations significatives**, c'est-à-dire des règles de classement et de prédiction dont la finalité ultime la plus courante est l'aide à la décision.

Le datamining est un procédé de production de connaissance. En terme de logique philosophique traditionnelle, le datamining consiste à **produire des jugements** (toutes les personnes sont x, la moyenne des y des personnes vaut tant, etc. : c'est l'étape de description et de compréhension des données) et des **règles de raisonnement** (si toutes les personnes sont a alors elles seront b : c'est l'étape modélisation qui permet la prédiction).

Le Datamining 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 Datamining (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, .etc.

2 Application du Datamining

Les applications de Datamining sont diverses. Il est utilisé par les entreprises pour la création des profils clients, ciblage des clients potentiels et nouveaux marchés. En finance, il est exploité pour la minimisation des risques. En bioinformatique, pour l'analyse du génome, mise au point de médicaments. En Internet, pour la détection des spams, e-

commerce, détection d'intrusion, etc.

Exemples d'applications :

1. **E-commerce**
   Dell
   Problème : 50% des clients de Dell achètent leurs machines à travers le site Web. Mais seulement 0,5% des visiteurs du site deviennent clients.
   Solution : Stocker les séquences de clicks des visiteurs, analyser les caractéristiques des acheteurs et lors de la visite d'un client potentiel, adapter le contenu du site pour maximiser la probabilité d'un achat.
   Amazon
   Opportunité : la liste des achats des clients est stockée en mémoire et par ailleurs, les utilisateurs du site notent les produits! Comment tirer profit des choix d'un utilisateur pour proposer des produits à un autre client ?
   Solution : technique dit de filtrage collaboratif permettant de regrouper des clients ayant les mêmes "goûts"
2. **Détection de fraudes pour les assurances**
   - Analyse des déclarations des assurés par un expert afin d'identifier les cas de fraudes.
   - Extraction de caractéristiques à partir de ces déclarations (type d'accident, de blessures, etc.)
   - Applications de techniques de Datamining pour identifier les caractéristiques des déclarations fortement corrélées à la fraude.

   Prêt Bancaire
   - Objectif des banques : réduire le risque des prêts bancaires.
   - Créer un modèle à partir de caractéristiques des clients pour discriminer les clients à risque des autres.
3. **Commerce**
   Organisation de rayonnage
   - **Objectifs :** Identifier les produits que les gens sont susceptibles d'acheter conjointement afin d'organiser les rayonnages.
   - **Données :** Code-Barre des produits.
   - **Méthodes :** Extractions de règles
   - **Exemples :**
     - **Résultats logiques :** le sucre et le café sont souvent proches.
     - **Résultats étranges :** dans une étude américaine, la vente de bière est plus importante si le rayon des couches n'est pas trop loin, et si sur le chemin il y a des chips, cela permet d'augmenter la vente des 3 produits.

3 Mise en œuvre d'un projet de datamining

Il est très important de comprendre que le datamining n'est pas seulement le problème de découverte de modèles dans un ensemble de donnée. Ce n'est qu'une seule étape dans tout un processus suivi par les scientifiques, les ingénieurs ou toute autre personne qui cherche à extraire les Connaissances à Partir des Données (ECD) ou *Knowledge Data Discovery (KDD)*. En 1996, un groupe d'analystes définit le datamining comme étant un processus composé de cinq étapes sous le standard CRISP-DM (*Cross-Industry Standard Process for Data Mning*) comme schématisé ci-dessous [4].

Ce processus, composé de cinq étapes, n'est pas linéaire, on peut avoir besoin de revenir à des étapes précédentes pour corriger ou ajouter des données. Par exemple, on peut

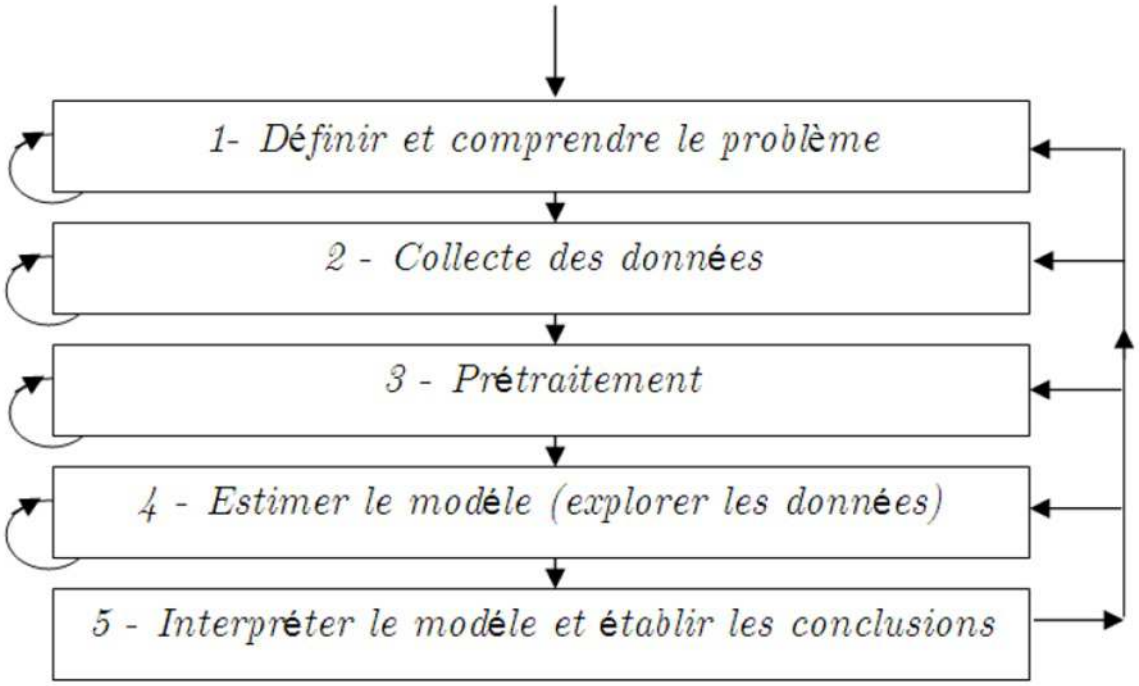

FIGURE 1.1 – Processus de datamining (CRISP-DM)

découvrir à l'étape d'exploration (5) de nouvelles données qui nécessitent d'être ajoutées aux données initiales à l'étape de collection (2).
Décrivons maintenant ces étapes :

1. **Définition et compréhension du problème :** Dans la plus part des cas, il est indispensable de comprendre la signification des données et le domaine à explorer. Sans cette compréhension, aucun algorithme ne va 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 Datamining 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. Une bonne compréhension du problème comporte une mesure des résultats de l'exploration, et éventuellement une justification de son coût. C'est-à-dire, pouvoir évaluer les résultats obtenus et convaincre l'utilisateur de leur rentabilité.
2. **Collecte des données :** dans cette étape, on s'intéresse à la manière dont les données sont générées et collectées. D'après la définition du problème et des objectifs du Datamining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la même structure. On peut avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles, etc.). Quelques traitements ne nécessitent qu'une partie des données, on doit alors sélectionner les données adéquates. Généralement les données sont subdivisées en deux parties : une utilisée pour construire un modèle et l'autre pour le tester. On

prend par exemple une partie importante (suffisante pour l'analyse) des données (70 à 80 %) à partir de laquelle on construit un modèle qui prédit les données futures. Pour valider ce modèle, on le teste sur la partie restante (20 à 30 %) dont on connaît le comportement.

3. **Prétraitement :** 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 éliminer complètement leurs enregistrements. Des données peuvent être incohérentes c.-à-d. qui sortent des intervalles permis, on doit les écarter où les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Un exemple de ces transformations est la normalisation des données qui consiste à la projection des données dans un intervalle bien précis [0,1] ou [0,100] par exemple. Un autre exemple est le lissage des données qui considère les échantillons très proches comme étant le même échantillon. 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 dans ce contexte, est l'analyse en composantes principales (ACP). Une autre méthode de réduction est celle de la sélection et la suppression des attributs dont l'importance dans la caractérisation des données est faible, en mesurant leurs variances. On peut même réduire le nombre de données utilisées par le Datamining en écartant les moins importantes. Dans la majorité des cas, le prétraitement doit préparer des informations globales sur les données pour les étapes qui suivent tel que la tendance centrale des données (moyenne, médiane, mode), le maximum et le minimum, le rang, les quartiles, la variance, ... etc. Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc., peuvent aider à la sélection et le nettoyage des données.
4. **Estimation du modèle :** 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, les réseaux bayésiens, le clustering, etc. sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat.
5. **Interprétation du modèle et établissement des conclusions :** généralement, l'objectif du datamining est d'aider à la prise de décision en fournissant des modèles compréhensibles aux utilisateurs. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des interprétations des modèles obtenus. Les expériences montrent que les modèles simples sont plus compréhensibles mais moins précis, alors que ceux complexes sont plus précis mais difficiles à interpréter.

4 Prétraitement des données

Les données extraites ne sont pas nécessairement toutes exploitables par des techniques de datamining. En effet, la plus part des techniques que nous utilisons ne traitent que des tableaux de données numériques rangées sous forme lignes/colonnes. Certaines méthodes sont plus contraignantes que d'autres. Elles peuvent par exemple exiger des données binaires, comme c'est le cas des premières techniques de recherche de règles d'association.

Les données acquises depuis l'entrepôt peuvent être de types différents. On peut y trouver des textes de longueur variables, des images, des enregistrements quantitatifs ou des séquences vidéo.

La préparation consiste à homogénéiser les données et à les disposer en tableau lignes/colonne. Formellement, chaque ligne/colonne peut être considérée comme un objet vecteur ayant un nombre fixe de composants. Ce vecteur ligne/colonne sera vu comme un objet mathématique que l'on pourra manipuler selon qu'il possède ou non certaines propriétés. Par exemple, si tous les vecteurs lignes sont des points de l'espace euclidien à p dimensions, on pourra faire appel aux techniques de datamining basées sur l'algèbre linéaire.

En général, cette transformation doit fournir un tableau ligne/colonne car il s'agit presque toujours de la structure la mieux adaptée à l'exploitation des données. Précisons que dans certaines situations, les données arrivent déjà sous une forme appropriée et qu'il n'est alors plus nécessaire de les modifier. Dans d'autres cas, elles sont dans une structure tabulaire mais exigent une transformation telle qu'un centrage par rapport à la moyenne ou une normalisation. En fait, le prétraitement est un acte de modélisation d'expert. Si l'expert ne définit pas les bonnes transformations ou les bons attributs, il ne verra alors rien dans ses données. L'expert devra par conséquent choisir un canevas pour représenter ses données et éventuellement effectuer une série de transformations pour obtenir des données adaptées aux méthodes d'exploitation.
Les principales opérations de préparation peuvent être listées comme suit :

- **Sélection de ligne/colonne :** Elle s'effectue sur des données qui sont déjà sous forme tabulaire. Il s'agit ensuite de définir un filtre qui permet de sélectionner un sous-ensemble de lignes ou de colonnes. L'objectif est soit de réduire le nombre de données soit de sélectionner les lignes ou colonnes les plus pertinentes par rapport aux préoccupations de l'utilisateur. Les techniques mises en œuvre dans ce but relèvent des méthodes statistiques d'échantillonnage, de sélection d'instances ou de sélection d'attributs. Cette sélection peut également s'effectuer selon des conditions exprimées par l'utilisateur. Par exemple, il peut ne garder que les attributs dont la moyenne est supérieure à un seuil donné ou ne conserver que les attributs qui ont un lien statistique significatif avec un attribut particulier. Ce lien sera évalué à l'aide d'une mesure d'association comme le khi-2 de Pearson ou le gain informationnel.
- **Traitement des données manquantes ou aberrantes :** Certaines données peuvent être absentes et gêner ainsi l'analyse. Il faut donc définir des règles pour gérer ou pour remplacer ces données manquantes. De nombreuses solutions sont proposées, comme le remplacement, dans le cas des données numériques continues, de toute donnée manquante par le mode de la distribution statistique (la valeur la plus fréquente) de l'attribut concerné, si ce mode existe. On peut également chercher à estimer ces valeurs manquantes par des méthodes d'induction comme la régression, les réseaux de neurones simples ou multicouches, ou les graphes d'induction. Pour le traitement des données aberrantes, il faut d'abord repérer ces dernières au moyen d'une règle préétablie. Par exemple, toutes les données numériques dont la valeur sur un attribut donné s'écarte de la valeur moyenne plus deux fois l'écart-type, pourraient être considérées comme des données possiblement aberrantes et qu'il conviendrait de traiter.
- **Transformations d'attributs :** Il s'agit de transformer un attribut A en une autre variable A' qui serait, selon les objectifs de l'étude, plus appropriée. Différentes méthodes sont pratiquées comme la discrétisation qui consiste à transformer des attributs continus en découpant le domaine de valeurs de ces attributs en intervalles afin d'obtenir des attributs qualitatifs. Il existe à cet effet pléthore de méthodes de discrétisation : supervisées ou non, à intervalles de tailles identiques, ou à intervalles à effectifs constants. On peut également centrer par rapport à la moyenne et réduire par l'écart type les valeurs des variables continues. Ce traitement leur confère certaines propriétés mathématiques intéressantes lors de la mise en œuvre de méthodes

d'analyse des données multidimensionnelles.

- **Construction d'agrégats :** Dans certaines situations particulières, il peut s'avérer que des agrégats d'attributs soient très importants pour la tâche d'analyse. Un agrégat d'attribut est un nouvel attribut obtenu selon une transformation précise. Par exemple, le prix au mètre-carré d'un appartement, défini par le rapport entre le prix de l'appartement et la surface totale de l'appartement, fournit une indication assez pertinente pour comparer les appartements ou les quartiers dans les bases de données spatiales. On peut imaginer une multitude de façons d'obtenir des agrégats. Parmi les méthodes de construction d'agrégats les plus utilisées, les méthodes factorielles telles que l'analyse en composantes principales (ACP) ou l'analyse des correspondances multiples (ACM).
- **Traitement des données complexes :** Toutes les méthodes de prétraitement citées précédemment opèrent sur des tableaux de données lignes/colonnes. Or il arrive que nous travaillions sur des données non structurées sous forme de tableaux. Par exemple, en Textmining, nous disposons d'un ensemble de textes de longueurs variées qu'il convient de ramener à une forme tabulaire. L'une des techniques les plus simples consiste à recenser l'ensemble des mots de tout le corpus et ensuite de calculer, pour chaque texte représenté par une ligne ou une colonne, la fréquence de chacun de ces mots. On obtient ainsi un tableau de comptage. Mais le codage des textes fait généralement appel à des procédures plus élaborées qui s'appuient sur la linguistique : lemmatisation, suppression des mots vides, thesaurus ou ontologies du domaine. La préparation des données images, et a fortiori vidéo, est encore plus ardue pour le non-spécialiste car il faut définir la liste des attributs qui décrivent l'image. Par exemple, on utilise les caractéristiques de l'histogramme des niveaux de gris, les attributs de texture, etc. La fouille sur des images peut nécessiter d'autres transformations en amont qui relèvent plus du traitement de l'image que de l'ECD. L'extraction de connaissances à partir de données complexes est d'ailleurs un domaine en pleine croissance.

5 Logiciels de datamining

Il existe de nombreux logiciels de Datamining qui sont faciles à installer. On trouve des logiciels libres et/ou gratuits et d'autres commerciaux, on peut citer [7] :

- **TANAGRA :** une grande partie de la panoplie des méthodes de Datamining intégrées dans une seule structure. Le mode d'utilisation du logiciel est au standard logiciels du domaine, avec notamment la définition des opérations à réaliser sur les données à l'aide d'une représentation visuelle. A voir absolument, la section didacticiels de ce site. (site : http ://eric.univ-lyon2.fr/ ricco/tanagra/)
- **SIPINA :** du même auteur que TANAGRA, il se distingue surtout par sa large palette de méthodes d'induction par arbres de décision, avec la possibilité d'interagir directement avec les modèles construits. Distribué depuis une dizaine d'années, ce logiciel est très connu dans le monde de la recherche. (site : http ://eric.univ-lyon2.fr/ ricco/sipina.html)
- **R-project :** Un logiciel gratuit que l'on associe souvent aux statisticiens mais qui en réalité convient très bien pour le Datamining. La bibliothèque des fonctions est impressionnante et s'enrichit chaque jour grâce au système des packages. Seul inconvénient, il fonctionne à l'aide d'un interpréteur de commandes. Il faut un peu de pratique pour en tirer véritablement parti. Le lien indiqué nous dirige vers mon site de cours de programmation sous R, vous y trouverez tous les liens idoines pour l'apprentissage du logiciel R. (site : http ://eric.univ-lyon2.fr/ ricco/cours/cours_programmation_R.html).

- **Weka :** L'espace de travail Weka contient une collection d'outils de visualisation et d'algorithmes pour l'analyse des données et la modélisation prédictive, allié à une interface graphique pour un accès facile de ses fonctionnalités. Les principaux points forts de Weka sont qu'il :
  - est librement disponible (en particulier gratuitement) sous la licence publique générale GNU,
  - est très portable car il est entièrement implémenté en Java et donc fonctionne sur quasiment toutes les plateformes modernes, et en particulier sur quasiment tous les systèmes d'exploitation actuels,
  - contient une collection complète de préprocesseurs de données et de techniques de modélisation, et
  - est facile à utiliser par un novice en raison de l'interface graphique qu'il contient.

  Weka supporte plusieurs outils d'exploration de données standards, et en particulier, des prétraitement de données, des agrégateurs de données (data clustering), des classificateurs statistiques, des analyseurs de régression, des outils de visualisation, et des outils d'analyse discriminante
  (site : http ://www.cs.waikato.ac.nz/ml/weka/).

6 Etude de cas

Nous présentons maintenant l'exemple qui nous servira pour illustrer chaque étape du processus. Cet exemple est issu du livre de P. Adriaans et D. Zantige [1].
Un éditeur 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. Les questions qu'il se pose sont :

1. Combien de personnes ont pris un abonnement a un magazine de sport cette année ?
2. A-t-on vendu plus d'abonnements de magazines de sport cette année que l'année dernière ?
3. Est-ce que les acheteurs de magazines de BD sont aussi amateurs de sport ?
4. Quelles sont les caractéristiques principales de mes lecteurs de magazines de voiture ?
5. Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?

Les questions qui sont proposées sont de natures différentes et mettent en jeu des processus différents. De simples requêtes SQL sont suffisantes pour répondre aux deux premières questions. Les données recherchées dans ce cas ne sont que le résultat d'un calcul simple sur un ou des groupes d'enregistrements. Ce qui distingue ces deux premières questions, c'est la notion de temps et la comparaison. Pour établir cette comparaison, les données doivent être présentes dans la base, ce qui n'est pas toujours le cas pour les bases opérationnelles.

Nous pouvons donc supposer que la requête 1 est réalisable sans technique particulière à partir des données opérationnelles sous réserve d'indexations suffisantes des tables concernées.

La requête 2 nécessite de conserver toutes les dates de souscription même pour les abonnements résiliés. Nous pouvons imaginer aussi que l'utilisateur émettra une grande variété de requêtes de ce type. Nous supposons alors que ces questions seront résolues à l'aide d'outils de création de requêtes multidimensionnelles.

La question 3 est un exemple simplifié de problème où l'on demande si les données vérifient une règle. La réponse pourrait se formuler par une valeur estimant la probabilité que la règle soit vraie. Souvent, des outils statistiques sont utilisés. Cette question peut être généralisée. On pourrait chercher des associations fréquentes entre acheteurs de magazine

pour effectuer des actions promotionnelles. On pourrait également introduire une composante temporelle pour chercher si le fait d'être lecteur d'un magazine implique d'être, plus tard, lecteur d'un autre magazine.

La question 4 est beaucoup plus ouverte. La problématique ici est de trouver une règle et non plus de la vérifier ou de l'utiliser. C'est pour ce type de question que l'on met en œuvre des outils de fouille de données.

La question 5 est également une question ouverte. Il faut pour la résoudre disposer d'indicateurs qui pourraient sur notre exemple être : les durées d'abonnement, les délais de paiement. Ce type de question a une forte composante temporelle et nécessite des données historiques. Le processus de KDD utilise des outils de création de requêtes, des outils statistiques et des outils de fouille de données. Là encore, nous nous apercevons qu'une très grande partie des problèmes de décision se traite avec des outils simples. La fouille de données quand elle est nécessaire, suit souvent une analyse des données simple ou statistique.

Phase de préparation des données Dans notre exemple, nous avons identifié quelques objectifs précis, exprimés sous forme de questions. La préparation des données consiste dans un premier temps à obtenir des données en accord avec les objectifs que l'on s'impose. Ces données proviennent le plus souvent de bases de production ou d'entrepôts. Pour illustrer cette phase, nous partons d'une liste des souscriptions d'abonnements avec cinq champs (voir Tableau 1.1).

N client	Nom	Adress	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	30/50/95	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

TABLE 1.1 – Obtention des données

L'obtention des données est souvent réalisée à l'aide d'outils de requête (*OLAP : On-Line Analytical Processing, SQL : Structured Query Language*...).

Il faut, dès à présent, noter que l'on ne peut résoudre des problèmes que si l'on dispose des données nécessaires. Il semble, par exemple, impossible de s'attaquer à la question cinq de notre exemple avec les données dont nous disposons.

On peut avoir recours à d'autres bases, achetées ou produites en un autre lieu, pour enrichir nos données. L'opération va se traduire par l'ajout de nouveaux champs en conservant souvent le même nombre d'enregistrements. Une première difficulté ici est de pouvoir relier des données qui parfois sont hétérogènes. Des problèmes de format de données apparaissent et des conversions sont souvent nécessaires. Une deuxième difficulté est l'introduction de nouvelles valeurs manquantes.

Pour notre exemple, supposons que nous ayons accès à des informations sur les clients données dans le tableau 1.2.

Nettoyage des données

Avant de pouvoir utiliser les données à travers les outils de fouille, un certain nombre de vérifications et de transformations sont nécessaires pour garantir la qualité des données.

Consolidation

Selon les choix des unités pour les dimensions, des opérations de consolidation devront

Client	Date de naissance	Revenus	Propriétaire	Voiture
Bemol	13/1/50	20 000 F	Oui	Oui
Bodinoz	21/5/70	12 000 F	Non	Oui
Airinaire	15/06/63	9 000 F	Non	Non
Manvussa	27/03/47	15 000 F	Non	Oui

TABLE 1.2 – Enrichissement des données

accompagner le chargement des données (par exemple sommer les ventes pour obtenir et enregistrer un total par jour et non pas toutes les transactions).

Uniformisation d'échelle

Pour éviter de trop grandes dispersions dans les valeurs numériques, une homogénéisation des échelles de valeurs est utile. Ne pas la réaliser peut pénaliser les outils d'analyse et de visualisation et peut-être simplement remplir inutilement les supports de stockage.

Doublons, erreurs de saisie

Les doublons peuvent se révéler gênants parce qu'ils vont donner plus d'importance aux valeurs répétées. Une erreur de saisie pourra à l'inverse occulter une répétition. Dans notre exemple, les clients numéro 23134 et 23130 sont certainement un seul et même client, malgré la légère différence d'orthographe.

Intégrité de domaine

Un contrôle sur les domaines des valeurs permet de retrouver des valeurs aberrantes. Dans notre exemple, la date de naissance du client 23130 (11/11/11) semble plutôt correspondre à une erreur de saisie ou encore à une valeur par défaut en remplacement d'une valeur manquante.

Informations manquantes

C'est le terme utilisé pour désigner le cas où des champs ne contiennent aucune donnée. Parfois, il est intéressant de conserver ces enregistrements car l'absence d'information peut être une information (ex : détection de fraudes). D'autre part, les valeurs contenues dans les autres champs peuvent être utiles. 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. *Pré traitement des données*

N client	Nom	Adress	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	NULL	BD
43342	Airinaire	Rue de la source, Brest	30/50/95	Sport
43241	Manvussa	NULL	14/04.96	Sport
23130	Bemol	Rue du moulin, Paris	NULL	Maison

TABLE 1.3 – Données après nettoyage

A ce stade du processus, les choix sont particulièrement guidés par l'algorithme de fouille utilisé et des ajustements des choix de codage sont souvent nécessaires.

Regroupements

Certains attributs prennent un très grand nombre de valeurs discrètes. C'est typiquement le cas des adresses. Lorsqu'il est important de considérer ces attributs pour la fouille de données il est obligatoire d'opérer des regroupements et ainsi obtenir un nombre de valeurs raisonnable. Dans l'exemple, nous regroupons les adresses en deux régions : Paris, province.

Attributs discrets
Les attributs discrets prennent leurs valeurs (souvent textuelles) dans un ensemble fini donné. C'est le cas de la colonne magazine dans notre exemple qui peut prendre les valeurs Sport, BD, Voiture, Maison, Musique. Deux représentations sont possibles pour ces données : une représentation verticale telle qu'elle est présentée en table 1 ou une représentation horizontale ou éclatée (voit Tableau 1.4).

N0	Sport	BD	Voiture	Maison	Musique
23134	0	1	1	0	1
31435	0	1	0	0	0
43342	1	0	0	0	0
43241	1	0	0	0	0

TABLE 1.4 – Données après nettoyage

La représentation horizontale est plus adaptée à la fouille de données et certains calculs sont simplifiés. Par exemple, la somme de la colonne sport que divise le nombre d'enregistrements calcule le pourcentage de clients ayant souscrit un abonnement à un magazine de sport.

Notons que la date d'abonnement dépend du type de magazine. De façon générale, la modification présentée en table 1.5 peut induire une perte d'information pour tous les champs en dépendance fonctionnelle avec le champ transformé. Nous choisissons de transformer le champ date d'abonnement en date du plus vieil abonnement. Il est à noter que cette transformation ne nous permet plus d'espérer générer des règles de la forme : un acheteur de BD s'abonne à un magazine de musique dans les deux ans qui suivent. Dans notre exemple, le même codage en deux valeurs 0 et 1 sera réalisé avec les champs Oui/Non issus de l'enrichissement.

Changements de type
Pour certaines manipulations, comme des calculs de distance, des calculs de moyenne, il est préférable de modifier les types de certains attributs. Dans notre exemple, la date de naissance et la date d'abonnement ne permettent pas d'effectuer simplement des comparaisons, des différences. Nous pouvons les convertir en âge ou en durée.

Uniformisation d'échelle
Certains algorithmes, comme la méthode des plus proches voisins, sont basés sur des calculs de distance entre enregistrements. Des variations d'échelle selon les attributs sont autant de perturbations possibles pour ces algorithmes. Des échelles très "étirées" vont artificiellement rendre des dimensions trop "vides".

C'est typiquement le cas pour le champ Revenus dans notre exemple. Les centaines de francs ne sont pas significatives. Nous convertissons donc les revenus en les divisant par mille. L'intervalle de valeurs est alors dans la même échelle que les dates de naissance et les durées d'abonnement.

N0	Sport	BD	Voit.	Mais.	Musique	Age	Reven.	Prop.	Voit.	Paris ?	Dur'ee abon.
23134	0	1	1	0	1	50	20	Oui	Oui	1	4
31435	0	1	0	0	0	30	12	Non	Oui	0	NULL
43342	1	0	0	0	0	37	9	Non	Non	0	5
43241	1	0	0	0	0	53	15	Non	Oui	NULL	4

TABLE 1.5 – Pré traitement des données