Chapitre 1

Introduction

l'apparition est du à la richesse du monde actuel en données et l'abondance de ces derniers mais paradoxalement pauvre en information , la croissance exponentielle en quantité de données qui sont stockées et collectées sont à l'origine de l'incapacité humaine à gérer ce flux de données et en extraire des connaissances pertinentes sans une utilisation d'outils adéquates, le besoin en exploration de données grandit et.

1 Qu'est ce que le Data mining

Le Data mining ou fouille de données est le processus d'exploration et d'extraction de connaissances ou de motifs à partir d'un volume conséquent de données en utilisant des outils , le processus comporte plusieurs étapes qui sont les suivantes :

1.1 Nettoyage de données

ceci consiste à éliminer les données bruitées

1.2 Integration de données

1.3 Sélection de données

on récupéré les données pertinentes de la base de données pour qu'elles puissent par la suite être utilisé lors de l'analyse

1.4 Transformation de données

La transformation consiste à faire des opérations d'agrégation , normalisation ou bien un résumé à fin de rendre les données sous une formes approprié pour l'exploration minière .

1.5 Data mining

c'est l'application de méthodes intelligentes pour l'extraction des modèles ou motifs de données

1.6 Évaluation des motifs

c'est une Identification des motifs pertinents représentant le mieux la connaissance

1.7 Présentation de données

c'est la visualisation et présentation des connaissances aux utilisateurs.

2 Les types de données qui peuvent être utilisés pour le data mining

Ils existent différent types de données sur les quelles le data mining peut opérer , pour cella la seule condition suffisante est que ces données doivent avoir du sens , parmi elles :

2.1 Les bases de données

c'est une collection de données interdépendantes un logiciel (système de gestion de base de données SGBD) permettant de gérer ,accéder et sécuriser ces données, les données sont sous forme de tables relationnelles suivant un certain format, chaque table comporte des colonnes qui appelées attribues et les lignes sont les tuples ou instance (un objet).

2.2 Les entrepôts de données

un entrepôt de données est un référentiel des informations collectées à partir de sources multiples, stockées sous une base de données unifiée. ce schéma est résidant généralement sur un site unique, l'entrepôt de données est construit via un processus de nettoyage des données, d'intégration, de transformation, de chargement et de rafraîchissement des données odiques et généralement modélisé par une structure de données multidimensionnelle, appelée cube de données dans lequel chaque dimension représente un ou plusieurs attributs ,ce cube de données fournit une vue multidimensionnelle des données et permet le précalcul et l'accès rapide aux données résumées.

2.3 Base données transactionnelles

une base de données transactionnelles se compose principalement de transactions mais peut aussi contenir des tables supplémentaires contenant des informations sur les transactions existantes .

2.4 Autres types de données

il existe d'autres types de données sous formes et structures polyvalentes et assez différentes sémantiquement. Ces types de données peuvent être vus dans de nombreuses applications: liées au temps données séquentielles, les flux de données , données spatiales (cartes, par exemple), données de conception technique , hypertexte et multi-données multimédia (y compris texte, image, vidéo et audio), graphique et données en réseau et le Web . Ces applications apportent de nouveaux défis, tel que la gestion des données contenant des structures spéciales et une sémantique spécifique

3 Quels sont les motifs qui peuvent être extraits ?

3.1 Classe/Concept Description: Caractérisation et Discrimination

Les entrées de données peuvent être associé à des classes ou des concepts , ces classes et concepts sont décrites selon un mécanisme de **caractérisation des données**, en résumant les données de la classe étudiée (souvent appelé le classe cible) en termes généraux, ou **discrimination des données**, par comparaison de la classe cible avec une ou plusieurs classes comparatives (souvent appelée la en contraste Des classes),ou une composition des deux techniques. le résultat de caractérisation donnent en sortie des statistiques résumé en diagramme à barres , courbes , ou bien des règles de généralisations , alors que la discrimination donne en sortie des règles de discriminations.

3.2 Motifs fréquents , associations et corrélations

Les motifs fréquents sont des modèles qui apparaissent de façon assez répétitive dans les données, il existe plusieurs types tel que les ensembles d'éléments fréquent,des sous-projets ou encore des sous-structures fréquentes séquentiels et structurés de différentes formes (graphes,arbre ...), **Association** les règles d'association sont des règles exprimés en prédicat permettant de découvrir une relation entre différente variable (attributs) dans une large base de données .
correlation est une relation liant deux ou plusieurs attributs associées dans un large volume de données.

3.3 Classification et régression pour une analyse prédictive

La Classification est le processus de recherche d'un modèle (ou fonction) qui décrit et distingue des classes de données ou des concepts. ils sont utilisé pour prédire l'étiquette de classe d'objets pour laquelle l'étiquette de classe est inconnue

Régression régression modélise des fonctions à valeurs continue , c'est une méthode statistique très utilisée pour prédire les données manquantes ou valeurs indisponibles de données numériques plutôt que des étiquettes de classe (discrètes).

3.4 clustering

Le clustering est le processus de regroupement des objets (des données) selon un degré de similarité entre eux, tel que on maximise la similarité entre les membres d'un même groupe tout en minimisant la similarité entres les membres de groupes différents (dissimilarité) . ça permet une meilleure organisation de données où chaque groupe présentent une classe ,catégorie ...

3.5 Outlier (valeurs aberrantes)

Ce sont des objets non conformes au comportement ou au modèle général des données. Ces objets de données sont valeurs aberrantes. ces objets peuvent être vu comme du bruit ou des exceptions .

3.6 Ce qui fait qu'un modèle est intéressent

Il est clair que le data mining peut générer des milliers de modèles et motifs mais l'utilité des modèles restent relatif selon le besoin , à noter que les modèles sont coûteux à produire donc il faudrait s'intéresser qu'aux modèles pertinents, pour qualifier un modèle de tel adjectif il faut :

- qu'il soit facile à assimiler et comprendre par ces utilisateurs
- qu'il procure une validation de nouvelles données ou d'essai avec un certain degré de certitude
- utile et intéressent dans le sens ou il valide ce que l'utilisateur cherche à confirmer ceci donnera naissance à DES CONNAISSANCES

.

4 Quelles genre de technologies sont appliquées et utilisées dans le data mining ?

4.1 Statistique

les statistiques étudient la collection, l'analyse, l'interprétation ou l'explication et la présentation de données.UNE modèle statistique est un ensemble de fonctions mathématiques décrivant le comportement de les objets d'une classe cible en termes de variables aléatoires et leurs distributions . on peut citer la moyenne , la médian , la variance et l'écart type.

4.2 Machine learning (apprentissage automatique)

Apprentissage machine étudie comment les ordinateurs peuvent apprendre basé sur des données. Un domaine de recherche principal concerne les programmes informatiques qui apprennent automatiquement à reconnaître des modèles complexes et prendre des décisions intelligentes basées sur des données. on distingue deux types d'apprentissage supervisé (prédire une classe)et non supervisé (clustering par exemple) ,Semi-supervisé ,active learning

4.3 Système de base de données et entrepôts de données

La recherche de systèmes de bases de données se concentre sur la création, la maintenance et l'utilisation de bases de données pour les organisations et les utilisateurs . ces systèmes sont utilisé pour les raisons suivantes : les modèles de données, les langages de requête, le traitement des requêtes,méthodes d'optimisation, stockage de données et méthodes d'indexation et d'accès qui permet une meilleure exploration de données.

on note aussi l'entrepôt qui consolide les données dans un espace multidimensionnel pour former des cubes de données facilitant ainsi leurs explorations .

4.4 Recherche d'information

recherche d'information(IR) est la science de la recherche de documents ou d'informations dans les documents.ils peuvent être du texte ou du multimédia et peuvent résider sur le Web .La recherche d'informations suppose que les données sous recherche sont non structurées et les requêtes sont formées principalement par des mots-clés.

5 Quelles sont les genres d'applications ciblé par le data mining

5.1 Business intelligence

La technologie de l'intelligence d'entreprise (BI) fournit des informations historiques, actuelles et vues prédictives des opérations commerciales. ceci en incluant les rapports, l'analyse en ligne traitement, gestion des performances de l'entreprise, veille concurrentielle, analyse comparative et analyse prédictive de comportement des clients par exemple .

5.2 Web Search engine (Moteur de recherche sur le web)

Un Moteur de recherche Web est un serveur informatique spécialisé dans la recherche des informations sur le Web , les résultats de la recherche d'une requête utilisateur sont souvent renvoyés sous forme de liste,les types des résultats peuvent consister en des pages Web, des images et d'autres types de fichiers.

6 Les problèmes qui font face au data mining

Il y a beaucoup de défis en jeux de la recherche en exploration de données:

- La méthodologie minère
- Interaction avec l'utilisateur
- L'efficacité , l'évolutivité, et la gestion de diverses types de données

Chapitre 2

Prétraitement des données

1 Qualité des données : Pourquoi faire du prétraitement sur nos données ?

On dira que les données sont de qualité si elles répondent aux exigences suivantes :

- l'exactitude
- exhaustivité
- complétude
- cohérence
- actualité (évolutif dans le temps)
- crédibilité et interprétabilité.

Mais avoir des données de qualité n'est pas chose facile , dans le monde réel beaucoup de facteurs rentrent en jeux pour compromettre la qualité des données tel que :

- valeurs manquantes (donnée incomplète) dans les tuples de données
- valeurs aberrantes (date de naissance 1700 par exemple "bruité")
- inconsistance ou encore des valeurs dupliquées .

Si alors on plonge directement dans l'exploration de nos données sans prêter attention à ces inconvenants là, la qualité des motifs ou modèles sera affectés directement ce qui causera un manque de crédibilité et de confiance dans ces derniers qui risquent être erronés.

2 Data Cleaning

L'une des principales approches de prétraitement des données , qui essaye de proposer des solutions aux différents problèmes citées plus .

2.1 Valeurs manquantes

Il existe diverse méthodes pour gérer les valeurs manquantes des données parmi elles :

2.1.1 Ignorer le tuple

cela consiste à renoncer à l'utilisation du tuple alors que ce dernier peut s'avérer intéressent par la suite, cette solution n'est efficace si la majorité de nos données souffrent d'un manque de valeurs d'un attribut ou deux .

2.1.2 Remplir les valeurs manquantes manuellement

Si la taille de nos données est conséquente cette méthode devient infaisable (elle prend énormément de temps) .

2.1.3 Utilisation d'une valeur globale pour remplir les valeurs manquantes

son inconvénient majeur est que lors de la fouille ou l'exploration ce genre de valeurs globales peuvent être prise pour des concepts intéressent .

2.1.4 Utilisation d'une mesure de la tendance centrale de l'attribut (par exemple, la moyenne ou la médiane)

2.1.5 Utilisez l'attribut moyenne ou médiane pour tous les échantillons appartenant à la même classe que le tuple donné

2.1.6 Utilisation de la valeur la plus probable

ceci peut être fait grâce à la fonction de régression , arbre de décision ou encore l'algorithme naïf de bayes,c'est généralement la technique la plus utilisée .

2.2 Les données bruitées ou erronées (Noisy data)

Les données bruitées sont des données qui manquent d'incohérence comme des valeurs aberrantes ,plusieurs techniques ont été mise au point pour faire face à ce genre de problème ,on citera:

2.2.1 Binning (nettoyer en enlevant des valeurs)

Binning se repose principalement sur le trie des valeurs des données puis il effectuera un partitionnement uniforme de façon à voir la même taille pour chaque partition ,ce qui permettra de consulter les voisins des valeurs erronées (une recherche locale) , le binning choisira par la suite de remplacer les valeurs aberrantes par la moyenne de la partition ou la médiane ou encore par le min si la valeur est plus proche du celui ci sinon par le max de l'intervalle de la partition (le plus proche voisin).

2.2.2 Régression

Afin de remplacer les valeurs erronés on peut faire appelle à une fonction de régression linéaire qui inclut deux variables (attributs) on donnera alors une valeur d'attribut pour prédire l'autre.

$$F(attribut_1) = attribut_2$$

ou bien une fonction de régression de dimension trois ou plus ou plusieurs attributs sont mise en jeux .

$$F(attribut_1, attribut_2, ..., attribut_n) = attribut_j$$

2.2.3 Analyse des outlier

en utilisant le clustering

2.3 Data Cleaning comme un processus

On peut résumer le nettoyage de données en des étapes élémentaires qui sont les suivantes:

Détection de divergence :

- Utilisez des métadonnées .
- Vérifiez la surcharge du champ.
- Vérifiez la règle d'unicité, la règle consécutive et la règle null.
- Utilisez des outils commerciaux.(data scrubbing (nettoyage) ,Data auditing(detecter des relations o corrélations)

Migration de données et intégration:

- Outils de migration de données
- Outils ETL (Extraction / Transformation / Loading)

Intégration des deux processus: Itératif et interactif

3 Intégration des données

3.1 Problème d'identification d'entité

ce problème survient lorsque on à faire à des données collecter de plusieurs sources, l'identification d'entité consiste à retrouver des attributs qui font référence à la même entité par exemple l'attribut id_consommateur ou nb_consommateur mais aussi la détection de différente structure d'un même attribut du diverse sources , généralement on utilise les méta données pour remédier à ce problème

3.2 Analyse de redondance et de corrélation

Lorsqu'un attribut peut être déduit à partir d'un autre attribut ou plus on dira qu'il est redondant, l'analyse de corrélation permet de détecter la redondance on citera :

3.2.1 X^2 corrélation pour données nominales

Pour les données nominales, une relation de corrélation entre deux attributs, A et B , peut être découvert par un X2(chi-carré) test, Hypothèse : les deux distributions sont indépendantes Les cellules (matrice r colonnes c lignes) qui contribuent le plus à la valeur X2 sont celles dont le nombre réel (observé) est très différent du nombre attendu . Plus la mesure X2 est grande, plus les variables sont susceptibles d'être liées.

$$X^2 = \sum_{i=1}^{C} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) * count(B = b_j)}{n}$$

c : valeurs distinctes que peut prendre l'attribut A
r :valeurs distinctes que peut prendre l'attribut B
o_{ij}: fréquence observé de l'événement commun $A = a_i, B = b_j$
e_{ij}: fréquence attendue (prévisible) de l'événement commun $A = a_i, B = b_j$
n: nombre de tuples totale
count(a_i): nombre de tuple ayant la valeur a_i pour A
count(b_j):nombre de tuple ayant la valeur b_j pour B

3.2.2 Coefficient de corrélation pour les données numériques

coefficient de corrélation de pearson :

$$rA, B = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i, b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

$\bar{A}$:La moyenne des valeurs de l'attribut A.
$\bar{B}$:La moyenne des valeurs de l'attribut B.
σ_A: l'écart type de A
σ_B: l'écart type de B
$\sum_{i=1}^{n}(a_i, b_i)$: produit vectoriel
r A,B est une mesure entre -1 et 1 si la valeur est positive on dira que A ,B sont positivement corrélé donc ils augmentent de façon proportionnelle, plus la valeur est proche de 1 plus ils sont corrélé , par contre si la valeur est négative on dira que A,B sont inversement proportionnelle . à noter que si r A,B = 0 , A et B sont indépendants

3.2.3 Covariance des données numériques

Dans la théorie des probabilités et les statistiques, la corrélation et la covariance sont deux mesures similaires.

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$$

La covariance est défini comme :

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

$$rA, B = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

$$Cov(A, B) = E(A.B) - \bar{A}\bar{B}$$

Comme la corrélation numérique ,A,B sont dites indépendant alors leurs covariance est nulle , positivement proportionnelle si la covariance est positive et inversement proportionnelle si covariance est négative.

3.3 Duplication de tuples

c'est lorsqu'il y a deux ou plusieurs tuples identiques pour une donnée en cas de saisie de données unique.

3.4 Détection et résolution de conflits de valeurs de données

L'intégration de données implique également la détection et résolution de conflits de valeurs de données par exemple pour une même entité du monde réel, les valeurs d'attribut provenant de sources différentes peuvent varier. Cela peut être dû à des différences de représentation, de mise à l'échelle ou d'encodage. Par exemple, un poids (attribut) peut être stocké en unités métriques (kg) dans un système et impériale britannique dans un autre (pound).