Science des données

Introduction à l'extraction de données

Themis Palpanas
Université de
Paris

Merci pour les diapositives :

- Jiawei Han
- Jeff Ullman

2

Feuille de route

- Motivation : Pourquoi l'exploration de données ?
- Qu'est-ce que l'exploration de données ?
- L'extraction de données : Sur quel type de données ?
- Fonctionnalité d'exploration de données
- Tous les motifs sont-ils intéressants ?
- Classification des systèmes d'exploration de données
- Primitives de tâches d'exploration de données
- Intégration d'un système d'extraction de données avec un système de base de données et de gestion des données.
- Principaux problèmes liés à l'extraction de données

Pourquoi l'analyse des données massives ?

- La croissance explosive des données : des téraoctets aux pétaoctets
  - Collecte et disponibilité des données
    - Outils de collecte de données automatisés, systèmes de base de données, Web, société informatisée
  - Principales sources de données abondantes
    - Affaires : Web, e-commerce, transactions, actions, …
    - Science : Télédétection, bioinformatique, simulation scientifique, …
    - Société et tout le monde : actualités, appareils photo numériques,
- Nous sommes noyés sous les données, mais affamés de connaissances !
- L'exploration de données : Analyse automatisée d'ensembles de données massives

Pourquoi l'analyse des données massives ?

- exemples de tailles de données
  - l'industrie des télécommunications (AT&T)
    - 7 Go/jour de données détaillées sur les appels
    - 15 Go/jour de données de surveillance du réseau IP
  - sites web
    - 10TB/jour de données de clics pour Yahoo !
  - détaillants
    - 20 millions de transactions commerciales/jour pour WalMart
  - projets scientifiques
    - 1,2 To/jour pour le système d'observation de la Terre (NASA)
    - 100PB/an pour l'Organisation européenne pour la recherche nucléaire (CERN)

Évolution de la technologie des bases de données

- 1960s :
  - Collecte de données, création de bases de données, SGBD IMS et réseau
- 1970s :
  - Modèle de données relationnel, mise en œuvre du SGBD relationnel
- 1980s :
  - SGBDR, modèles de données avancés (relationnel étendu, OO, déductif, etc.)
  - SGBD orientés applications (spatial, scientifique, ingénierie, etc.)
- 1990s :
  - Extraction de données, entreposage de données, bases de données multimédia et bases de données Web
- 2000s
  - Gestion et extraction de données en continu
  - L'exploration de données et ses applications
  - Technologie Web (XML, intégration de données) et systèmes d'information globaux

Qu'est-ce que l'extraction de données ?

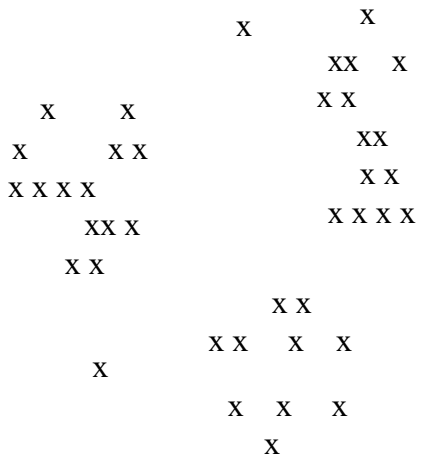
- Extraction de données (découverte de connaissances à partir de données)
  - Extraction des éléments intéressants (non triviaux, implicites, antérieurs). des modèles ou des connaissances inconnus et potentiellement utiles) à partir d'une énorme quantité de données
  - L'exploration de données : un terme mal choisi ?
- Noms alternatifs
  - Découverte de connaissances (exploration) dans les bases de données (KDD), extraction de connaissances, analyse de données/modèles, archéologie des données, dragage de données, récolte d'informations, intelligence économique, etc.
- Faites attention : Tout est-il "data mining" ?
  - Traitement simple des recherches et des requêtes
  - Systèmes experts (déductifs)

Types de motifs typiques

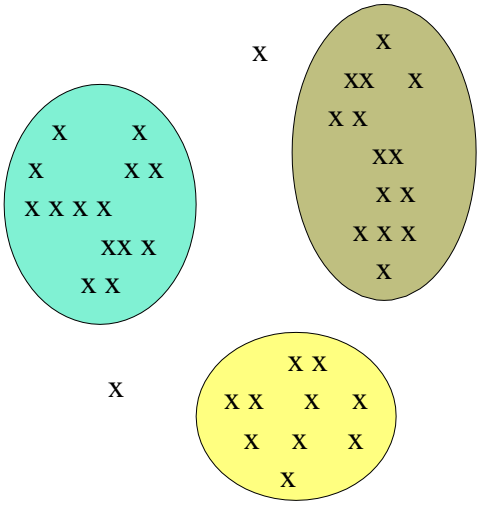
1. Les arbres de décision : des moyens succincts de classer en testant propriétés.
2. Clusters : une autre classification succincte par similarité de propriétés.
3. Modèles de Bayes, modèles de Markov cachés, ensembles d'éléments fréquents : exposent les associations importantes au sein des données.

8

Exemple : Clusters



Exemple : Clusters



Exemple : Ensembles d'éléments fréquents

- Un problème de marketing courant : examiner ce que les gens achètent ensemble pour découvrir des modèles.

1. Quelles sont les paires d'objets que l'on trouve exceptionnellement souvent ensemble à la caisse de Safeway ?
   - Réponse : les couches et la bière.

2. Quels livres sont susceptibles d'être achetés par le même client Amazon ?

Pourquoi l'extraction de données ? - Applications potentielles

- Analyse des données et aide à la décision
  - Analyse et gestion du marché
    - Marketing ciblé, gestion de la relation client (CRM), analyse du panier de la ménagère, vente croisée, segmentation du marché.
  - Analyse et gestion des risques
    - Prévision, fidélisation des clients, amélioration de la souscription, contrôle de la qualité, analyse de la concurrence
  - Détection de fraudes et détection de modèles inhabituels (valeurs aberrantes)
- Autres applications
  - Exploration de textes (groupes de discussion, courriels, documents) et exploration du Web

- Extraction de données en continu
- Bioinformatique et analyse des bio-données

Ex. 1 : Analyse et gestion du marché

- D'où proviennent les données ? - Transactions par carte de crédit, cartes de fidélité,

les coupons de réduction, les appels de réclamation des clients, ainsi que les études sur le mode de vie (public).

  - Marketing ciblé
    - Trouvez des groupes de clients "modèles" qui partagent les mêmes caractéristiques : intérêt, niveau de revenu, habitudes de dépenses, etc,
    - Déterminer les habitudes d'achat des clients au fil du temps
  - Analyse croisée des marchés : trouver les associations/co-relations entre les ventes de produits et faire des prévisions sur la base de ces associations.
  - Profilage de la clientèle - quels types de clients achètent quels produits (regroupement ou classification)
  - Analyse des besoins du client
    - Identifier les meilleurs produits pour les différents clients
    - Prévoir les facteurs qui attireront de nouveaux clients
- Fourniture d'informations sommaires
  - Rapports de synthèse multidimensionnels
    - Informations statistiques sommaires (tendance centrale et variation des données)

Ex. 2 : Analyse de l'entreprise et gestion des risques

- Planification financière et évaluation des actifs
  - analyse et prévision des flux de trésorerie
  - analyse des créances éventuelles pour évaluer les actifs
  - analyse transversale et de séries chronologiques (ratio financier, analyse des tendances, etc.)
- Planification des ressources
  - résumer et comparer les ressources et les dépenses
- Concours
  - suivre les concurrents et les orientations du marché
  - regrouper les clients en classes et appliquer une procédure de tarification par classe
  - définir une stratégie de prix dans un marché hautement concurrentiel

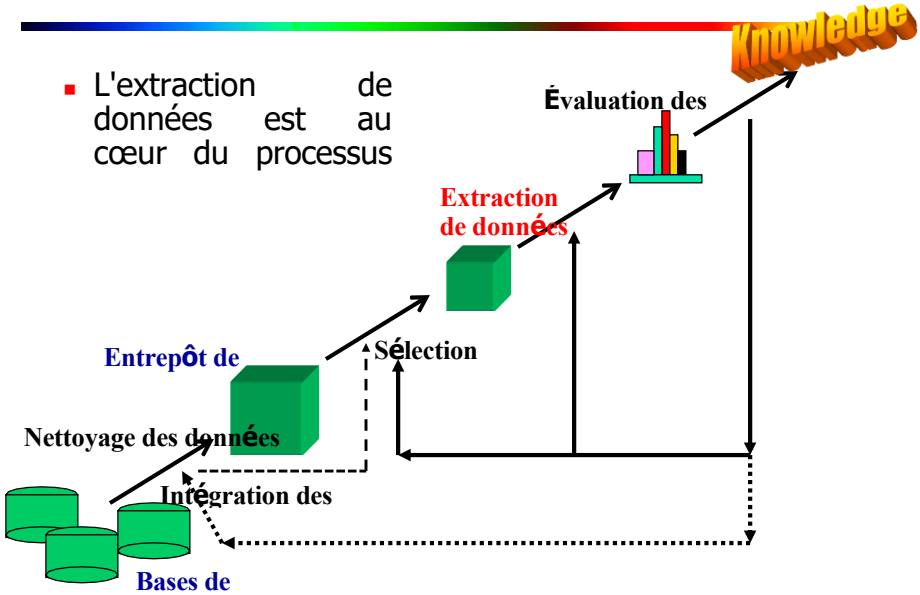
14

Ex. 3 : Détection des fraudes et exploration des modèles inhabituels

- Approches : Regroupement et construction de modèles pour les fraudes, analyse des valeurs aberrantes.
- Applications : Soins de santé, commerce de détail, services de cartes de crédit, télécommunications.
  - Assurance automobile : anneau de collisions
  - Blanchiment d'argent : transactions monétaires suspectes
  - Assurance médicale
    - Patients professionnels, cercle de médecins et cercle de références
    - Tests de dépistage inutiles ou corrélés
  - Télécommunications : fraude par appel téléphonique
    - Modèle d'appel téléphonique : destination de l'appel, durée, heure de la journée ou de la semaine. Analyser les modèles qui s'écartent d'une norme attendue.
  - Industrie du détail
    - Les analystes estiment que 38 % des pertes de magasins sont dues à des employés malhonnêtes.
  - Anti-terrorisme

Processus de découverte de connaissances (KDD)

- L'extraction de données est au cœur du processus

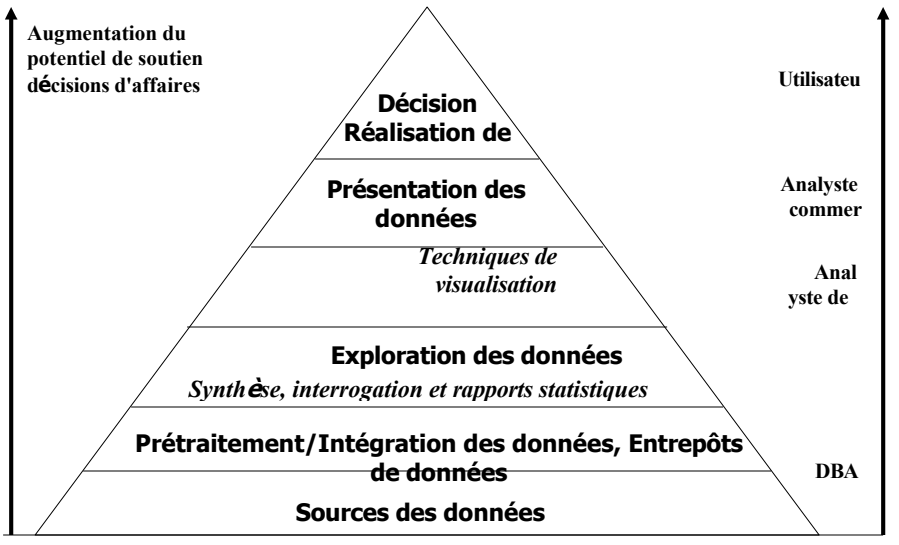


16

Le processus KDD : Plusieurs étapes clés

- Apprendre le domaine d'application
  - connaissances préalables pertinentes et objectifs de l'application
- Création d'un ensemble de données cibles : sélection des données
- Nettoyage et prétraitement des données : (peut nécessiter 60% de l'effort !)
- Réduction et transformation des données
  - Trouver des caractéristiques utiles, réduction de la dimensionnalité/variable, représentation invariante.
- Choix des fonctions de l'exploration des données
  - résumé, classification, régression, association, regroupement
- Choix de l'algorithme ou des algorithmes d'extraction
- Extraction de données : recherche de modèles d'intérêt
- Évaluation des modèles et présentation des connaissances
  - visualisation, transformation, suppression des motifs redondants, etc.
- Utilisation des connaissances découvertes

Extraction de données et intelligence économique

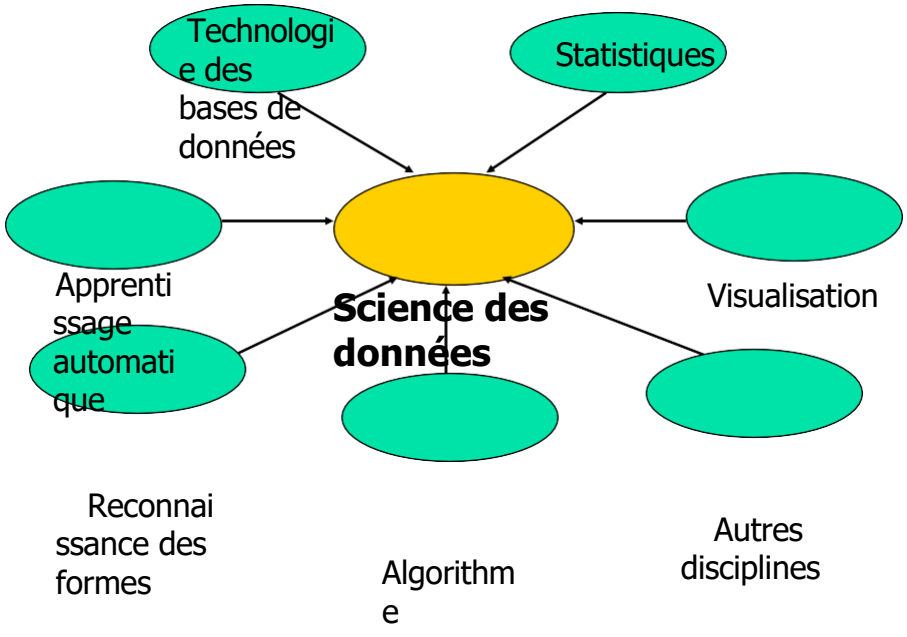


Science des données

- **La science des données** est un domaine interdisciplinaire qui...
- utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour
- extraire des connaissances et des idées de
- données structurées et non structurées

source : Wikipedia

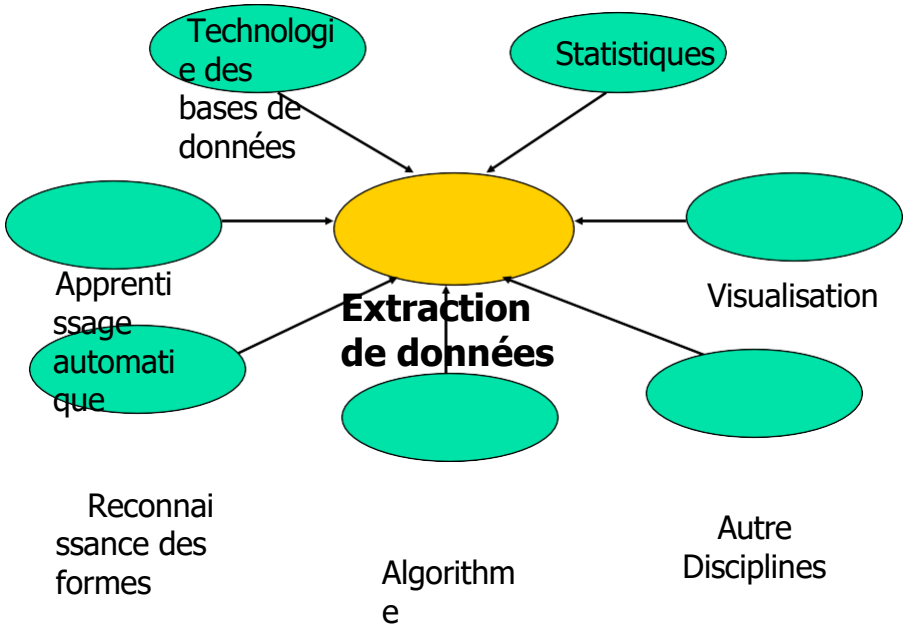
La science des données : Confluence de multiples disciplines



année :
2020

20

L'exploration de données : Confluence de multiples disciplines



année : 2000

Qu'est-ce que l'extraction de données ?

- Extraction de données (découverte de connaissances à partir de données)
  - Extraction des éléments intéressants (non triviaux, implicites, antérieurs). des modèles ou des connaissances inconnus et potentiellement utiles) à partir d'une énorme quantité de données
  - L'exploration de données : un terme mal choisi ?
- Noms alternatifs
  - Découverte de connaissances (exploration) dans les bases de données (KDD), extraction de connaissances, analyse de

données/modèles, archéologie des données, dragage de données, récolte d'informations, intelligence économique, etc.

- Faites attention : Tout est-il "data mining" ?
  - Traitement simple des recherches et des requêtes
  - Systèmes experts (déductifs)

Cultures

- Bases de données : se concentrent sur les données à grande échelle (sans mémoire principale).
- IA (apprentissage automatique) : se concentrer sur les tâches complexes.
méthodes, petites données.
- Statistiques : concentrez-vous sur les modèles.

Modèles et traitement analytique

- Pour un spécialiste des bases de données, l'exploration de données est une forme extrême de traitement analytique --- des requêtes qui examinent de grandes quantités de données.
  - Le résultat est la donnée qui répond à la requête.
- Pour un statisticien, l'exploration de données est l'inférence de modèles.
  - Le résultat est les paramètres du modèle.

24

Pourquoi pas l'analyse traditionnelle des données ?

- Une quantité énorme de données
  - Les algorithmes doivent être hautement évolutifs pour traiter des téraoctets de données.
- Haute dimensionnalité des données
  - Les microréseaux peuvent avoir des dizaines de milliers de dimensions.
- Grande complexité des données
  - Flux de données et données des capteurs
  - Données de séries chronologiques, données temporelles, données de séquences
  - Données structurées, graphes, réseaux sociaux et données multi-liées
  - Bases de données hétérogènes et bases de données patrimoniales
  - Données spatiales, spatio-temporelles, multimédia, texte et Web
  - Logiciels, simulations scientifiques
- Applications nouvelles et sophistiquées

Vue multidimensionnelle de l'extraction de données

- **Données à exploiter**
  - Relationnel, entrepôt de données, transactionnel, flux, orienté objet/relationnel, actif, spatial, séries temporelles, texte, multimédia, hétérogène, historique, WWW.
- **Des connaissances à exploiter**
  - Caractérisation, discrimination, association, classification, regroupement, analyse des tendances/déviations, analyse des valeurs aberrantes, etc.
  - Fonctions multiples/intégrées et exploitation minière à plusieurs niveaux
- **Techniques utilisées**
  - Orienté base de données, entrepôt de données (OLAP), apprentissage automatique, statistiques, visualisation, etc.
- **Applications adaptées**
  - Commerce de détail, télécommunications, banque, analyse des fraudes, extraction de données biologiques, analyse boursière, extraction de textes, extraction sur le Web, etc.

26

L'exploration de données : Schémas de classification

- Fonctionnement général
  - Exploration de données descriptives
  - Exploration prédictive des données
- Des points de vue différents conduisent à des classifications différentes
  - Vue des données : Types de données à extraire
  - Vue de la connaissance : Types de connaissances à découvrir
  - Vue de la méthode : Types de techniques utilisées
  - Vue des applications : Types d'applications adaptées

L'extraction de données : Sur quels types de données ?

- Ensembles de données et applications orientés vers les bases de données
  - Base de données relationnelle, entrepôt de données, base de données transactionnelle
- Ensembles de données et applications avancées
  - Flux de données et données des capteurs
  - Données de séries temporelles, données temporelles, données de séquences (y compris les bioséquences)
  - Données structurées, graphes, réseaux sociaux et données multi-liées
  - Bases de données objet-relationnelles
  - Bases de données hétérogènes et bases de données patrimoniales
  - Données spatiales et données spatio-temporelles
  - Base de données multimédia
  - Bases de données textuelles
  - Le World-Wide Web

28

Fonctionnalités de l'exploration de données

- Description de concepts multidimensionnels : Caractérisation et discrimination
  - Généraliser, résumer et mettre en contraste les caractéristiques des données, par exemple, les régions sèches par rapport aux régions humides.
- Modèles fréquents, association, corrélation vs. causalité
  - Couche-culotte Bière [0,5 %, 75 %] (Corrélation ou causalité ?)
- Classification et prédiction
  - Construire des modèles (fonctions) qui décrivent et distinguent des classes ou des concepts pour une prédiction future.
    - Par exemple, classer les pays en fonction de leur climat ou classer les voitures en fonction de leur consommation d'essence.
  - Prédire certaines valeurs numériques inconnues ou manquantes

Fonctionnalités de l'extraction de données (2)

- Analyse en grappes
  - L'étiquette de la classe est inconnue : Regrouper les données pour former de nouvelles classes, par exemple, regrouper des maisons pour trouver des modèles de distribution.
  - Maximiser la similarité intra-classe & minimiser la similarité inter-classe
- Analyse des valeurs aberrantes
  - Hors norme : Objet de données qui ne respecte pas le comportement général des données.
  - Bruit ou exception ? Utile pour la détection des fraudes, l'analyse des événements rares
- Analyse des tendances et de l'évolution
  - Tendance et déviation : par exemple, analyse de régression
  - Extraction de motifs séquentiels : par exemple, appareil photo numérique grande mémoire SD
  - Analyse de périodicité
  - Analyse basée sur la similarité

- Autres analyses statistiques ou axées sur le modèle

Tous les motifs "découverts" sont-ils intéressants ?

- L'extraction de données peut générer des milliers de modèles :
  Ils ne sont pas tous intéressants
  - Approche suggérée : Exploration centrée sur l'homme, basée sur les requêtes et ciblée.
- **Mesures d'intérêt**
  - Un modèle est intéressant s'il est facilement compréhensible par les humains, s'il est valable sur des données nouvelles ou des données d'essai avec un certain degré de certitude, s'il est potentiellement utile, s'il est nouveau ou s'il valide une hypothèse que l'utilisateur cherche à confirmer.
- **Mesures objectives et subjectives de l'intérêt**
  - Objectif : basé sur les statistiques et les structures des modèles, par exemple, le support,
    la confiance, etc.
  - Subjectif : basé sur la croyance de l'utilisateur dans les données,
    par exemple, le caractère inattendu, la nouveauté, la possibilité d'agir, etc.

Trouver tous et seulement les motifs intéressants ?

- Trouvez tous les motifs intéressants : Complétude
  - Un système d'exploration de données peut-il trouver tous les modèles intéressants ? Avons-nous besoin de trouver tous les modèles intéressants ?
  - Recherche heuristique ou exhaustive
  - Association, classification et regroupement
- Recherche des seuls modèles intéressants : Un problème d'optimisation
  - Un système d'exploration de données peut-il trouver uniquement les modèles intéressants ?
  - Approches

- D'abord, généraliser tous les modèles, puis filtrer ceux qui sont inintéressants.
- Générer uniquement les modèles intéressants - optimisation des requêtes par le mining

32

Signification des réponses

- Un grand risque de l'exploration de données est que vous allez "découvrir" des modèles qui n'ont aucun sens.
  - Les statisticiens l'appellent le principe de Bonferroni : (en gros) si vous cherchez des modèles intéressants à plus d'endroits que la quantité de données ne le permet, vous trouverez forcément de la merde.

Paradoxe rhénan --- (1)

- Joseph Rhine était un parapsychologue des années 1950 qui a émis l'hypothèse que certaines personnes avaient une perception extra-sensorielle.
- Il a conçu une expérience dans laquelle les sujets devaient deviner 10 cartes cachées --- rouges ou bleues.
- Il a découvert que près de 1 personne sur 1000 avait un ESP --- ils ont réussi à obtenir les 10 bonnes réponses !

34

Paradoxe du Rhin --- (2)

- Il a dit à ces personnes qu'elles avaient un ESP et les a convoquées pour un autre test du même type.
- Hélas, il a découvert que presque toutes ces personnes avaient
  ont perdu leur ESP.
- Quelle a été sa conclusion ?

Paradoxe du Rhin --- (3)

- Il a conclu que vous ne devriez pas dire aux gens qu'ils ont des ESP, elle les fait perdre.

Paradoxe du Rhin --- (4)

- Le paradoxe du Rhin : un excellent exemple de ce qu'il ne faut pas faire en matière de recherche scientifique.
- Lorsque vous recherchez un bien, assurez-vous qu'il n'y a pas tellement de possibilités que des données aléatoires produisent des faits "intéressants".

Autres questions relatives à l'extraction de motifs

- Modèles précis ou modèles approximatifs
  - Extraction d'associations et de corrélations : possibilité de trouver des ensembles de modèles précis.
    - Mais les modèles approximatifs peuvent être plus compacts et suffisants.
    - Comment trouver des modèles approximatifs de haute qualité ?
  - Extraction de séquences génétiques : les modèles approximatifs sont inhérents
    - Comment dériver un modèle approximatif efficace d'exploration de motifs
      algorithmes ? ??
- Modèles contraints et non contraints
  - Pourquoi l'exploitation minière basée sur les contraintes ?
  - Quels sont les types de contraintes possibles ?
    Comment intégrer les contraintes dans le processus

d'extraction ?

38

Pourquoi un langage de requête pour l'extraction de données ?

- Automatisé ou axé sur les requêtes ?
  - Trouver tous les motifs de manière autonome dans une base de données ? - irréaliste car les motifs pourraient être trop nombreux mais inintéressant.
- L'exploration de données doit être un processus interactif
  - L'utilisateur dirige ce qui doit être exploité
- Les utilisateurs doivent disposer d'un ensemble de primitives à utiliser pour communiquer avec le système d'extraction de données.
- Incorporation de ces primitives dans un langage de requête pour l'exploration de données
  - Une interaction plus souple avec l'utilisateur
  - Fondement de la conception de l'interface utilisateur graphique
  - Normalisation de l'industrie et des pratiques de l'extraction de données

Primitives qui définissent une tâche d'exploration de données

- Données relatives aux tâches
- Type de connaissances à exploiter
- Connaissances de base
- Mesures de l'intérêt des motifs
- Visualisation/présentation des modèles découverts

40

Primitive 1 : Données relatives à la tâche

- Nom de la base de données ou de l'entrepôt de données
- Tables de base de données ou cubes d'entrepôt de données
- Condition pour la sélection des données
- Attributs ou dimensions pertinents
- Critères de regroupement des données

Primitive 2 : Types de connaissances à exploiter

- Caractérisation
- Discrimination
- Association
- Classification/prédiction
- Regroupement
- Analyse des valeurs aberrantes
- Autres tâches d'exploration de données

Primitive 3 : Connaissance de base

- Un type typique de connaissance de base : Les hiérarchies de concepts
- Hiérarchie des schémas
  - Par exemple, rue < ville < province_ou_état < pays
- Hiérarchie de regroupement des ensembles
  - Par exemple, {20-39} = jeune, {40-59} = d'âge moyen.
- Hiérarchie dérivée des opérations
  - adresse électronique : hagonzal@cs.uiuc.edu

    nom de login < département < université < pays
- Hiérarchie basée sur des règles
  - low_profit_margin (X) <= prix (X, P_1) et coût (X, P_2) et (P_1 - P_2) < $50

Primitive 4 : Mesure de l'intérêt du motif

- Simplicité

  par exemple, la longueur des règles (d'association), la taille des arbres (de décision).
- Certitude

  par exemple, la confiance, P(A|B) = #(A et B)/ #(B), la fiabilité ou la précision de la classification, le facteur de certitude, la force de la règle, la qualité de la règle, le poids discriminant, etc.
- Utilitaire

  utilité potentielle, par exemple, soutien (association), seuil de bruit (description)
- Nouveauté

  non connu auparavant, surprenant (utilisé pour éliminer les redondances) règles)

Primitive 5 : Présentation des motifs découverts

- Des contextes/usages différents peuvent nécessiter des formes de représentation différentes.
  - Par exemple, des règles, des tableaux, des tableaux croisés, des diagrammes à secteurs ou à barres, etc.
- La hiérarchie des concepts est également importante
  - Les connaissances découvertes peuvent être plus compréhensibles lorsqu'elles sont représentées à un haut niveau d'abstraction.
  - L'exploration interactive vers le haut et vers le bas, le pivotement, le découpage en tranches et en dés offrent différentes perspectives aux données.
- Différents types de connaissances nécessitent une représentation différente : association, classification, regroupement, etc.

DMQL-A Langage de requête pour l'extraction de données

- Motivation
  - Un DMQL peut offrir la possibilité de soutenir l'exploration ad hoc et interactive des données.
  - En fournissant un langage normalisé comme SQL
    - J'espère obtenir un effet similaire à celui de SQL sur les bases de données relationnelles.
    - Fondation pour le développement et l'évolution du système
    - Faciliter l'échange d'informations, le transfert de technologies, la commercialisation et l'acceptation générale.
- Design

- DMQL est conçu avec les primitives décrites précédemment

46

Un exemple de requête en DMQL

Example 1.11 Mining classification rules. Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than $40,000, and who have bought more than $1,000 worth of items, each of which is priced at no less than $100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL[3] as follows, where each line of the query has been enumerated to aid in our discussion.

```
use database AllElectronics_db
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
mine classification as promising_customers
in relevance to C.age, C.income, I.type, I.place_made, T.branch
from customer C, item I, transaction T
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
        and C.income ≥ 40,000 and I.price ≥ 100
group by T.cust_ID
having sum(I.price) ≥ 1,000
display as rules
```


Autres langages de fouille de données & Efforts de normalisation

- Spécifications du langage des règles d'association
  - MSQL (Imielinski & Virmani '99)
  - MineRule (Meo Psaila et Ceri '96)
  - Flocons de requêtes basés sur la syntaxe Datalog (Tsur et al'98)
- OLEDB pour DM (Microsoft'2000) et DMX (Microsoft SQLServer 2005)
  - Basé sur OLE, OLE DB, OLE DB pour OLAP, C#
  - Intégration du SGBD, de l'entrepôt de données et du data mining
- DMML (Data Mining Mark-up Language) par DMG (www.dmg.org)
- Fournir une plateforme et une structure de processus pour une exploration efficace des données
  - L'accent est mis sur le déploiement de la technologie d'exploration des données pour résoudre les problèmes des entreprises.

Intégration de l'extraction de données et de l'entreposage de données

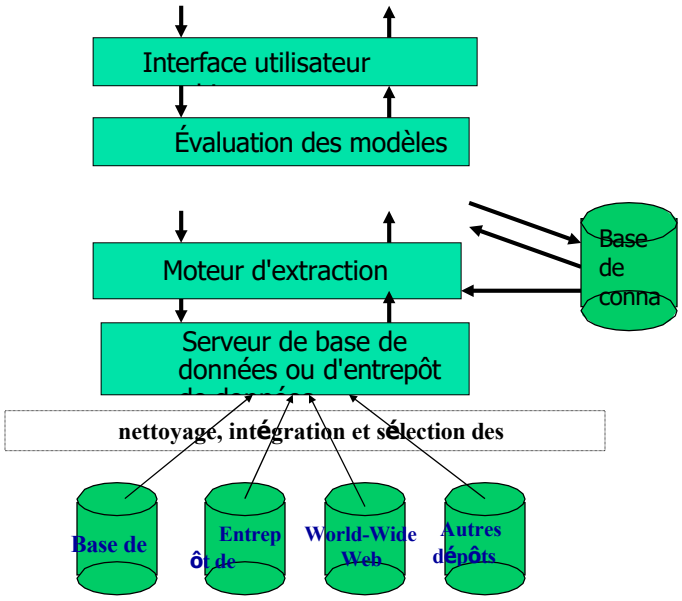
- **Couplage de systèmes d'exploration de données, SGBD, systèmes d'entrepôts de données**
  - Pas d'accouplement, accouplement lâche, accouplement semi-étanche, accouplement étanche.
- **Données d'exploitation analytique en ligne**
  - l'intégration des technologies minières et OLAP
- **Extraction interactive de connaissances à plusieurs niveaux**
  - Nécessité d'extraire des connaissances et des modèles à différents niveaux d'abstraction par forage/roulage, pivotement, découpage en tranches, etc.
- **Intégration de plusieurs fonctions minières**
  - Classification caractérisée, d'abord regroupement et ensuite association

Couplage de l'exploration de données avec les systèmes DB/DW

- Pas de traitement des fichiers à plat par couplage, non recommandé
- Accouplement libre
  - Récupération des données de la BD/DW
- Accouplement semi-étanche - performance DM améliorée
  - Fournir des implémentations efficaces de quelques primitives d'exploration de données dans un système DB/DW, par exemple, le tri, l'indexation, l'agrégation, l'analyse d'histogrammes, la jointure multivoie, le précalcul de certaines fonctions statistiques.
- Couplage étroit - Un environnement uniforme de traitement de l'information
  - DM est intégré en douceur dans un système DB/DW, la requête minière est optimisée en fonction de la requête minière, de l'indexation, des méthodes de traitement de la requête, etc.

50

Architecture : Système typique d'exploration de données



Principaux problèmes liés à l'extraction de données

- Méthodologie d'exploitation minière
  - Extraction de différents types de connaissances à partir de divers types de données, par exemple les données biologiques, les flux, le Web.
  - Performance : efficacité, efficience et évolutivité
  - Évaluation des motifs : le problème de l'intérêt
  - Incorporation des connaissances de base
  - Traitement du bruit et des données incomplètes
  - Méthodes d'extraction parallèles, distribuées et incrémentielles
  - Intégration des connaissances découvertes avec les connaissances existantes : fusion des connaissances.
- Interaction avec l'utilisateur
  - Langages de requête pour l'extraction de données et extraction ad hoc
  - Expression et visualisation des résultats de l'exploration des données
  - Extraction interactive de connaissances à plusieurs niveaux d'abstraction
- Applications et impacts sociaux
  - Extraction de données spécifiques à un domaine et extraction de données invisibles
  - Protection de la sécurité, de l'intégrité et de la confidentialité des données

52

Résumé

- L'exploration de données : Découverte de modèles intéressants à partir de grandes quantités de données
- Une évolution naturelle de la technologie des bases de données, très demandée, avec de nombreuses applications.
- Un processus KDD comprend le nettoyage des données, l'intégration des données, la sélection et la transformation des données, l'exploration des données, l'évaluation des modèles et la présentation des connaissances.
- L'extraction peut être effectuée dans divers référentiels d'information.
- Fonctionnalités de data mining : caractérisation, discrimination, association, classification, clustering, analyse des valeurs aberrantes et des tendances, etc.
- Systèmes et architectures d'exploration de données
- Principaux problèmes liés à l'extraction de données

Une brève histoire de la société d'extraction de données

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Atelier IJCAI sur la découverte de connaissances dans les bases de données)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro et W. Frawley, 1991)
- 1991-1994 Ateliers sur la découverte de connaissances dans les bases de données
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy, 1996)
- 1995-1998 Conférences internationales sur la découverte de connaissances dans les bases de données
  et l'extraction de données (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- Conférences ACM SIGKDD depuis 1998 et SIGKDD Explorations
- Autres conférences sur l'extraction de données
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD à partir de 2007

54

Conférences et revues sur l'extraction de données

- Conférences KDD
  - ACM SIGKDD Int. Conf. sur la découverte de connaissances dans les bases de données et la fouille de données (KDD)
  - Conférence de la SIAM sur l'exploration des données (SDM)
  - (IEEE) Int. Conf. sur la fouille de données (ICDM)
  - Conférence sur les principes et pratiques de la découverte de connaissances et de l'exploration de données (PKDD)
  - Conférence Asie-Pacifique sur la découverte de connaissances et l'extraction de données (PAKDD)
- Autres conférences connexes
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journaux
  - Extraction de données et découverte de connaissances (DAMI ou DMKD)
  - IEEE Trans. Sur la connaissance et l'ingénierie des données (TKDE)
  - Explorations KDD
  - ACM Trans. on KDD

Où trouver des références ? DBLP, CiteSeer, Google

- Data mining et KDD (SIGKDD : CDROM)
  - Conférences : ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal : Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Systèmes de bases de données (SIGMOD : ACM SIGMOD Anthology-CD ROM)
  - Conférences : ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journaux : IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- IA et apprentissage automatique
  - Conférences : Apprentissage automatique (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journaux : Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web et IR
  - Conférences : SIGIR, WWW, CIKM, etc.
  - Journaux : WWW : Internet et systèmes d'information Web,
- Statistiques
  - Conférences : Réunion conjointe Stat. Réunion, etc.
  - Revues : Annales de statistiques, etc.
- Visualisation
  - Comptes rendus de conférences : CHI, ACM-SIGGraph, etc.

- Journaux : IEEE Trans. visualisation et infographie, etc.

56