DIC9150 Concepts fondamentaux de l'informatique cognitive

Analyse de Données

Roger Villemaire

Département d'informatique
UQAM

28 novembre 2023

UQÀM
Université du Québec à Montréal

Plan

Données et IA

- Comme nous l'avons vu,
  - le développement de l'Intelligence Artificielle a déterminé l'importance cruciale d'avoir suffisamment d'information dans les systèmes intelligents.
- Il n'y a pas de comportement intelligent *ex nihilo*, sans relation avec le monde extérieur.

Nature des Données

- Il nous faut des concepts et des attributs fondamentaux :
  - Approche symbolique : relations, variables, etc.
  - Approche connexionniste : entrées, variables aléatoires, etc.

  et de l'information à leur sujet :
  - Approche symbolique : règles, contraintes, décisions, etc.
  - Approche connexionniste : exemples, statistiques, etc.
- De plus, tout ceci doit être sous une forme traitable par la machine.

Données et Statistiques

- On assiste à une certaine convergence entre l'analyse de données et la statistique, mais il y a toutefois un changement de paradigme :
  - en statistique classique on suppose un certain nombre d'aprioris probabilistes, de plus, les données sont rares et il faut s'en servir astucieusement pour valider des hypothèses de la façon la plus adéquate,
  - aujourd'hui on a ou on peut obtenir d'énormes quantités de données, mais les relations probabilistes ou même causales sont passablement moins claires.

Données et Entreprise

- Une source importante pour le phénomène émergent des Données Massives (Big Data) est le monde de l'entreprise :
  - Entreprise et Infrastructure informatique sont aujourd'hui indissociables.
  - Depuis au moins les années 1980, les données d'entreprise sont stockées dans des *bases de données*.

Bases de Données

- Une base de données est formée de tables (relations).
- Chaque table est formée d'attributs (variables).
  - par ex., nom, adresse, salaire, etc.
- L'ensemble des tables permet de stocker l'information
  - par ex. au sujet des clients, employés, transactions, contrats, etc. de l'entreprise.
- Les systèmes de bases de données modernes sont transactionnels, indépendants des langages de programmation, robustes, efficaces, distribués.

Entrepôts de Données

- Émergence d'*entrepôts de données* (data-warehouse) :
  - agrégation de données de plusieurs BD opérationnelles,
  - pour l'analyse (data analytics) et la prise de décision (tableau de bord).

Exemple

- Supermarché :
  - tout ce qui passe à la caisse est enregistré dans la BD,
  - et probablement associé à votre identité (carte de points ?),
  - les inventaires sont donc à jour, on peut voir l'évolution en temps-réel, prévoir les surplus (dates de péremption), les pénuries, assurer la livraison juste-à-temps (just-on-time),
  - on peut déterminer les tendances chez les clients et prévoir les besoins.

Exemple

- Dossier médical intégré :
  - on pourrait avoir accès à notre dossier médical partout (urgence),
  - incluant les prescriptions (historique, contre-indication),
  - globalement ceci pourrait permettre de mieux comprendre le fonctionnement du système de santé
    - réduction des erreurs,
    - prévision des effectifs et des ressources,
    - diminution du temps d'attente, du risque de contagion, etc.

Exemples

- Traçabilité en temps réel des bagages pour une compagnie aérienne.
- Traçabilité des rouleaux de papiers !
- RFID et gestion d'un entrepôt.

Le World Wide Web (WWW)

- Le Web est aussi une énorme source (potentielle) de données :
  - Twitter,
  - Facebook,
  - LinkedIn,
  - Wikipedia,
  - Flickr, etc.

Particularité du Web

- Manque de structure,
- sémantique pas très claire,
- pas facilement traitable par la machine.
- Néanmoins grande source de données exploitable pour :
  - compléter des dossiers de façon semi-automatique,
  - analyser des liens et des interactions,
  - déterminer un bon placement publicitaire !

Structure des données

- Données structurées
  - le sens et la position de l'information sont clairement définis
    - par ex. une base de données.
- Données semi-structurées
  - le sens et position de l'information sont en partie définis
    - document XML ou HTML.
- Données non structurés
  - le sens et position de l'information ne sont pas clairement définis ou alors très difficile à déterminer
    - texte dans une langue naturelle (TLN).

Intuition

- Nous avons, de plus en plus, accès à énormément de données, grâce à
  - la présence massive de l'informatique,
  - l'informatisation passablement complétée des entreprises,
  - le développement de l'Internet et une informatisation "sociale",
  - la décroissance phénoménale du coût du matériel,
  - la croissance exponentielle de la capacité de stockage (entreprise, individus).
- Il s'agit d'une énorme ressource qui devrait nous permettre de réaliser beaucoup de choses :
  - augmentation de la productivité/diminution des coûts,
  - développement de logiciels augmentant nos capacités,
  - réalisation des promesses de l'Intelligence Artificielle.

Science des données

- Données Massives (Big-Data) : ensembles de données énormes
  - par ex. toutes les transactions d'Amazon depuis la fondation de l'entreprise !
- Forage de Données (Data-Mining) : méthodes de l'Intelligence Artificielle et de la statistique permettant de découvrir des corrélations et des motifs (patterns) dans les données massives
  - par ex. système de recommandation : ceux qui ont acheté un robot culinaire ont aussi acheté dans les 6 mois une perceuse sans fil !
- Analyse prédictive : se servir des données massives pour prédire ou estimer un résultat ou comportement futur
  - cet étudiant gagnerait à faire le module optionnel XYZ !

Objectifs du Forage de données

- Description de motifs ou de tendances.
- Estimation d'une valeur numérique.
- Classification d'une valeur catégorielle.
- Regroupement (clustering) de valeurs en ensembles similaires.
- Association de causes et d'effets (règles si-alors avec support et confiance).

Méthodes du Forage de données

- Visualisation : représentation graphique pour développer une intuition des données.
- Prétraitement des données : champs obsolètes ou redondants, valeurs manquantes ou incohérentes, modification du format pour simplifier le traitement, intégration de sources différentes (souvent l'étape la plus difficile).
- Sélection de sous-ensembles de données ou des attributs les plus intéressants/pertinents.
- Traitement des données à l'aide d'algorithmes statistiques ou d'IA comme
  - arbres de décision, réseaux neuronaux, réseaux bayésiens, par ex., pour l'estimation ou la classification.

Doctorat en Informatique Cognitive

- Interface cognition/informatique :
  - modélisation cognitive à l'aide d'outils informatiques,
    - apport des sciences humaines, cognition (S. Robert),
    - apport de l'Intelligence Artificielle et de l'informatique,
  - recherche appliquée,
    - effectuer une synthèse des deux approches,
    - proposer une projet de recherche joignant cognition et informatique,
  - technologie cognitive,
    - les technologies cognitives de l'informatique ont atteint un bon niveau de maturité au niveau technique,
    - il reste néanmoins un potentiel d'application énorme,
    - il reste beaucoup de travail (définition du problème, formalisation, détermination des technologies les plus adéquates) à faire pour développer des applications effectives, utiles et pertinentes.
- Votre objectif maintenant : vous définir un projet !