Exploitation des données : ETL, Data Mining

Le **Data Mining** (ou exploration de données en français) est un enjeu depuis plusieurs années. Cela consiste en l'extraction et l'analyse de données afin de les transformer en informations utiles. Cette technique d'extraction permet de créer des corrélations entre les données et donc de comprendre des liens entre divers phénomènes pour pouvoir anticiper des tendances potentielles. Il repose sur l'utilisation d'algorithmes complexes dans différents domaines comme l'informatique, la santé, l'intelligence artificielle, les statistiques et bien d'autres... Le data mining utilise bon nombre d'algorithme pour l'exploitation des données (**deep learning**) et ainsi trouver des schémas précis ou différentes similitudes. Ces résultats peuvent ensuite être utilisés dans différents domaines comme la recherche, la santé, le sport ou même le tourisme : Les possibilités sont (quasi) infinies. C'est grâce à ça que aujourd'hui la plus part des applications de messagerie sont capable de terminer nos phrases ou de corriger nos fautes d'orthographe.

Application :

Les entreprises de marketing utilisent la fouille de données pour réduire le coût d'acquisition d'un nouveau client en classant les prospects selon des critères leur permettant d'augmenter les taux de réponses aux questionnaires envoyés. Les services de polices de tous les pays cherchent à caractériser les et les comportements des criminels afin de prévenir le crime, limiter les risques et les dangers pour la population. Le "scoring" des clients dans les banques est maintenant très connu, il permet de repérer les « bons » clients, sans facteur de risque à qui les organismes financiers, banques, assurances, etc. ; peuvent facilement prêter de l'argent. Les centres d'appel utilisent cette technique pour améliorer la qualité du service et permettre une réponse adaptée de l'opérateur pour la satisfaction du client. Dans la recherche du génome humain, les techniques d'exploration de données ont été utilisées pour découvrir les gènes et leur fonction.

Algorithmes :

Les méthodes descriptives permettent d'organiser, de simplifier et d'aider à comprendre l'information sous-jacente d'un ensemble important de données. Cette méthode est utilisée pour construire des normes de comportements, des groupes homogènes etc., dans un ensemble qui n'a d'apparence aucuns facteurs communs.

L’analyse factorielle est une méthode regroupant plusieurs méthodes d’analyse de tableaux de données statistiques visant à déterminer et à hiérarchiser des facteurs en corrélations.

Logiciels /Informatique:

Pour stoker et trier les données il existe de nombreux logiciels commerciaux comme libres de droit comme SPSS, RapidMiner, SAS, Excel, Oracle DM, Microsoft SQL Server… L’exploration des données découle du Machine Learning qui consiste en l’auto apprentissage des erreurs et expériences d’un ordinateur. De nombreuses techniques permettent maintenant d’affiner les recherches et besoins des entreprises pour leur offrir un maximum d’avantages. L’analyse de texte, très utilisé en plateforme Big Data avec du Deep Learning, cela permet de rechercher de manière automatique des motifs dans des textes. Cette technique est utilisée depuis plusieurs années dans les universités notamment pour repérer toute trace de plagiat.

Limite :

Les logiciels d'exploration de données donnent toujours un résultat, mais rien n'indique qu'il soit pertinent, ni ne donne une indication sur sa qualité. Cependant de plus en plus de techniques d'aide à l'évaluation sont mises en place dans les logiciels libres ou commerciaux. De plus, il peut être très difficile de restituer de manière claire soit par des graphes, des courbes ou des histogrammes, les résultats de l'analyse. Le non-technicien aura quelquefois du mal à comprendre les réponses qu'on lui apporte. Aussi, le vocabulaire est un problème, un francophone qui ne s’y connaît pas beaucoup aura du mal à comprendre le vocabulaire, parfois compliqué. Il y a aussi le problème liés à la vie privée des utilisateurs qui doit être respecté, des lois sont mises en places aujourd’hui (RGP). . De plus, des techniques de profilages sont rendues possibles par l’exploration de données pouvant poser des problèmes éthiques nouveaux.

Et plus tard ?

Avec l’apparition de toutes les plateformes d’informations, de partage, de communications, il y a une explosion du volume de données numériques. Cette explosion de données créé des enjeux colossaux en matière de sécurité. De plus, de nombreux domaine n’exploitent pas encore la fouille de données et son potentiel. Pour finir, avec l’apparition de nouvelles données et de nouveaux domaines, les techniques continuent sans cesse de se développer…

Glossaire :

Data Mining : L’exploration de données, connue aussi sous l'expression data mining 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. (Wikipédia)

ETL : Extract-transform-load est connu sous le sigle ETL. Il s'agit d'une technologie informatique intergicielle permettant d'effectuer des synchronisations massives d'information d'une source de données vers une autre. (Wikipédia)

Intergicielle : est un logiciel tiers qui crée un réseau d'échange d'informations entre différentes

applications informatiques.

Deep learning: Le deep learning est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaire.

Le Cloud computing : Le cloud computing (en français l'informatique en nuage), consiste à utiliser des serveurs informatiques distants par l'intermédiaire d'un réseau, généralement Internet, pour stocker des données ou les exploiter.

Sources :

https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es

https://www.saagie.com/fr/blog/qu-est-ce-que-le-data-mining/

https://seenthis.net/tag/data-mining