Big data, unstructured data, linked data, open data, private data, Semantics, Analytics, Business intelligence:

Mais que diable allez-vous faire dans cette galère !

Frédérique Segond

L'importance historique des données et de leur analyse

C'est une évidence : depuis le début de l'humanité, les données, sous toutes leurs formes, sont la base même de la connaissance et l'ingrédient principal de l'innovation. Les données servent donc la cause de la recherche, dont l'activité est la production des connaissances, et de l'innovation dont l'activité est la capacité à créer de la valeur en apportant quelque chose de nouveau, issu ou non de la recherche.

L'histoire foisonne d'exemples qui montrent comment les données et leur analyse ont été au centre de changements culturels, sociétaux et scientifiques. Les philosophes grecs comme Aristote ont construit, à partir du peu de données dont ils disposaient, des théories scientifiques. Peu à peu, cette approche qualitative a été complétée par une approche plus quantitative rendue possible grâce à une quantité de données disponibles de plus en plus importante.

Dans l'Antiquité, la Bibliothèque d'Alexandrie a pour but de collecter le plus de données possibles afin de que les chercheurs puissent capturer la connaissance du monde et que l'innovation puisse se faire dans différents domaines. On peut aussi citer les moines copistes qui se dédiaient, au travers de la copie, à la collecte de données afin de transmettre les connaissances à un plus grand nombre et de faire évoluer la science et la société. Au début du 17ème siècle Galileo recueille, grâce à son télescope, quantités de données. C'est à partir de l'observation de ces données qu'il développe une théorie qui est au fondement de l'astronomie moderne. Aujourd'hui encore l'astronomie continue de baser ses avancées scientifiques sur l'interprétation de grandes quantités de données.

Au 18ème siècle, les scientifiques s'éloignent, de plus en plus, des théories purement intellectuelles pour privilégier l'observation et l'expérimentation.
Le Comte de Buffon avec la publication de ses trente-six volumes de "Histoire naturelle, générale et particulière» est considéré par Darwin comme le premier auteur ayant abordé l'évolution d'une manière scientifique.
A la même époque, les encyclopédistes, sous la direction de Diderot, publient «l'Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers" dans le but de «changer la façon dont les gens pensent". Cet ouvrage est reconnu comme un vecteur intellectuel important de la révolution française qui a finalement conduit à de nouveaux modèles politiques.
Nous pouvons enfin citer Durkheim qui au 19e siècle a proposé une approche scientifique de la société basée sur des méthodes quantitatives. C'est cette approche qui a donné naissance à la sociologie moderne.

Les données à l'ère du numérique

Ce qui a radicalement changé avec l'avènement des technologies de l'Internet et de l'information, c'est que ces données qu'il était jusqu'alors très difficile de recueillir sont devenues, en un laps de temps très court, pléthores et très faciles d'accès.
En quelques années seulement, nous sommes passé du rêve, avoir accès à plus de données, au cauchemar, avoir trop de données. D'un monde de vaches maigres à un monde d'opulence. L'internet et l'utilisation généralisée des bases de données sont aujourd'hui la cause principale de la croissance exponentielle et continue des données en ligne.

De nos jours, les données ne sont plus majoritairement de type encyclopédique comme avant ; elles peuvent être des courriels, des murs de Facebook, et les échanges sur Twitter. De nos jours, les données sont recueillies non seulement à partir d'Internet, mais aussi par exemple à partir de factures, de commandes, de notes de supermarchés, de caméras, de lunettes, de téléphones portables, de voitures, de GPS, de tags RFID, et bientôt même à partir de réfrigérateurs, de poubelles ou de fours. Grâce aux logiciels embarqués la plus part des types de dispositifs électroniques que nous utilisons quotidiennement sont en mesure de fournir des données. Un autre aspect nouveau et important est la pérennité de ces données. En effet, alors qu'auparavant la plupart des données disparaissaient après avoir été utilisées dans un but précis, les données sont maintenant ***stockées, distribuées et même revendues pour être analyser et interprétées dans le meilleur des cas, à des fins d'innovation ou d'avancée scientifique***

Quelle recherche sur les données ?

La définition même de données a évolué au cours de l'histoire. Nous adoptons ici la définition générale des données comme étant des symboles tels que des mots, des chiffres, des codes ou des tables, des images. Ces symboles (données) peuvent être reliés en phrases, de paragraphes, en équation, former des concepts et des idées pour, à la fin, donner naissance à l'information. L'information peut ensuite encore être structurée et interprétée jusqu'à devenir de la connaissance.
On comprend bien qu'une fois ces données collectées il va falloir les stocker les protéger, pour certaines d'entre elles contrôler leur accès, faire en sorte de pouvoir les analyser, les croiser, les organiser, les relier, les enrichir en continu, les retrouver, y naviguer. Le but ultime étant de leur donner du sens, de les comprendre afin de mieux gérer l'information qu'elles véhiculent et de prendre les meilleures décisions, d'adopter les meilleures stratégies ou de gérer au mieux les risques, et ce, dans différents domaines et différents marchés verticaux.

Le but principal et ambitieux de l'analyse intelligente de données est d'extraire de la connaissance de différentes sources de données. Ceci implique de s'intéresser à analyse de ces données, à leur organisations, à leurs liens, à comment raisonner sur ces données. Tout cela dans le but de soutenir les entreprises dans tous les aspects de leur métier, comme par exemple, la connaissance du marché, de la compétition, la connaissance de ce que pense leur client, la connaissance de leurs données

chiffrées etc. Tout cela dans le but de leur donner tous les atouts pour prendre les meilleures décisions que ce soit au niveau marketing, financier ou technique.

Les données, nous l'avons vu, sont multiformes, nous nous concentrons prioritairement sur l'analyse des données textuelles y compris chiffrées sous leurs formes structurées et non structurées. Notre objectif dans ce domaine est d'étudier différents types d'approches permettant l'extraction et la compréhension de données non structurées et ceci à un niveau sémantique.

L'analyse intelligente des données est un domaine interdisciplinaire axé sur les méthodes permettant d'extraire des connaissances utiles à partir de données brutes qu'elles soient structurées ou non structurées.
Il ne s'agit pas de simplement collecter des mots-clés à partir de textes mais bien d'extraire de l'information c'est-à-dire des faits, des intentions et d'aider à leur donner une interprétation. Les applications de ce thème de recherche sont désormais légion au sein des entreprises, on citera par exemple, l'analyse des tendances, le traitement des demandes dans les centres d'appel, le traitement des FAQ , des courriers, des factures ou encore des opinions des clients.

Plusieurs types d'approches sont utilisés pour faire une analyse fine des données comme par exemple, les approches statistiques, l'apprentissage automatique, les approches symboliques, les approches hybrides, les différents types de logique ou la représentation des connaissances.
Nous ne nous attachons pas à un formalisme particulier notre but étant de produire des techniques d'analyses de données qui soit robustes, fines et puissent s'appliquer à un grand nombre de données et d'applications

Quel est le lien à l'innovation et au positionnement de Viseo?

De nos jours, l'innovation est généralement associée à deux ingrédients principaux: des technologies, d'une part, et des clients prêts à acheter ou à utiliser ces technologies, d'autre part. Contrairement à l'invention, l'innovation se doit d'avoir une valeur commerciale, elle est associée à l'idée de profit.
Le groupe VISEO est premier acteur multispécialiste des systèmes d'information. A ce titre, les données sont depuis longtemps au cœur des métiers du groupe. Par exemple, les ERP stockent des données structurées pour différents buts : la gestion des stocks, du personnel, des finances. La BI fournit des outils pour visualiser et comprendre ces données et au bout du compte faciliter la prise des décisions, gérer les risques, comprendre et anticiper les besoins des clients.
Aujourd'hui toute entreprise innovante traitant d'information ou de systèmes d'informations doit être en mesure d'offrir à ses clients une analyse la plus fine possible de ses données.

Viseo, grâce à ses recherches en analyse intelligente des données, va pouvoir élargir son offre et proposer à ses clients un soutien dans la compréhension de leurs données, afin d'en faire le meilleur usage dans le marketing, le développement technique, les décisions stratégiques ou la gestion des risques.

Les projets collaboratifs en cours reliés à cette thématique

GALATEAS

Le but de GALATEAS est d'offrir aux fournisseurs de contenus numériques une approche innovante leur permettant de mieux comprendre le comportement de leurs utilisateurs à travers l'analyse des informations textuelles contenues dans les journaux de transactions. Cette information permet aux utilisateurs de GALATEAS d'améliorer aussi bien la navigation à travers leur site web que la recherche multilingue de leurs contenus.

Les objectifs de GALATEAS sont les suivants:

- *Analyse des journaux de requêtes*. Analyser les logs contenant les requêtes des moteurs de recherche d'un fournisseur de contenu afin de produire des rapports sur mesure sur les utilisateurs ayant accès à cette agrégation particulière. L'analyse est basée sur des données aussi bien linguistiques que statistiques.
- *Traduction de requêtes* : Traduire des requêtes provenant d'un moteur de recherche externe en plusieurs langues cibles. Ce moteur externe utilise ces traductions pour retourner des résultats dans des langues autres que celle de la requête initiale. Les langues choisies pour GALATEAS sont: l'italien, le français, l'anglais, l'allemand, le néerlandais, l'arabe moderne et le polonais.

LEILAS

L'objectif du projet LEILAS est de proposer une fonctionnalité de géo- référenciation multi-langues sur le contenu de pages web ou de documents, permettant ainsi de leur apporter une dimension géographique.

Le projet LEILAS propose de développer la recherche appliquée sur l'identification et la normalisation des entités géographiques. Le projet développera une technologie qui permettra de reconnaitre automatiquement les références à des entités géographiques, telles que par exemple pays, région, villes, adresses, dans les documents Web et de leur associer des coordonnées spatiales.

References

Frédérique Segond, "Turning water into wine: transforming data sources to satisfy the thirst of the knowledge era", Fostering Language Resources Network (FLaReNet), Venezia, Italy, May 26-27, 2011