Des données au savoir : big data et data mining

Dossier de veille réalisé par Pierre-Alain FOUR (FRV1OO)

Juillet 2013

GRANDLYON
communauté urbaine
Direction de la Prospective et du Dialogue Public
20 rue du lac - BP 3103 - 69399 LYON CEDEX 03
www.millenaire3.com

La numérisation croissante de nos activités, la capacité sans cesse accrue à stocker des données numériques, l'accumulation d'informations en tous genres qui en découle, génère un nouveau secteur d'activité qui a pour objet l'analyse de ces grandes quantités de données. Sont alors apparues de nouvelles approches, de nouvelles méthodes, de nouveaux savoirs et in fine sans doute, de nouvelles manières de penser et de travailler. Ainsi, cette très grande quantité de données –ou big data– et son traitement –ou data mining– sous-tendent de profonds bouleversements, qui touchent à l'économie, au marketing, mais aussi à la recherche et aux savoirs. Les enjeux économiques, scientifiques et éthiques de ces données sont considérables. Le fait qu'on se situe dans un secteur en évolution constante, où les changements sont fréquents et rapides, ne rend pas l'analyse aisée… Cependant, un arrêt sur image, imparfait, nécessairement incomplet et pour partie périssable, s'avère sans doute nécessaire afin de mieux comprendre ce que sont le big data et le data mining. Pour tenter d'y voir un peu plus clair, ce dossier thématique se propose de donner un éclairage à ce phénomène.

NB : L'ensemble des documents cités ici, ainsi que les têtes de chapitres sont disponibles en version interactive sur le site www.pearltrees.com à l'adresse suivante :
http://www.pearltrees.com/ - /N-u=1_999949&N-p=83665988&N-s=1_8589707&N-f=1_8589707&N-fa=7061072

1 – Qu'est-ce que le data mining ?

On compile dans cette première section les articles généralistes, qui donnent des définitions, exposent les grands enjeux, synthétisent les questions, etc. On y trouve donc des documents qui permettent d'entrer rapidement dans le sujet du big data et du data mining.

Explorer de très grandes quantités de données

Le 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 ». On parle aussi d'exploration de données, de fouille de données, de forage de données, de prospection de données, d'Extraction de Connaissances à partir de Données (ECD) ou Knowledge Discovery from Data (KDD). Les 3 références qui suivent permettent de se faire une idée rapide sur les origines du data mining et sur ses implications

Références

1 – Exploration de données :
http://fr.wikipedia.org/wiki/Exploration_de_données
2 – Les techniques de collecte automatisée :
http://quanti.hypotheses.org/647/ - _ftn3
3 – Des données au savoir :
http://www.mysciencework.com/fr/MyScienceNews/9616/le-data-mining

Comment et pourquoi de telles quantités de nouvelles données sont-elles générées ?

Chaque jour, 118 milliards de mails sont envoyés à travers le monde et 2,45 milliards de contenus différents sont postés sur Facebook… Les nouvelles données sont produites en quantité exponentielle et d'ici moins de 10 ans plus de 10 400 milliards de gigaoctets de données transiteront tous les mois sur Internet ! Des chiffres qui donnent le vertige, surtout que les humains ne sont pas les seuls producteurs de données : les machines aussi y contribuent avec leurs cartes sim, leurs capteurs, etc.

Références

4 – Vertigineux big data :
http://www.lemonde.fr/technologies/article/2012/12/26/vertigineux-big-data_1810213_651865.html
5 – Données le vertige :
http://www.liberation.fr/economie/2012/12/03/donnees-le-vertige_864585
6 – Les données, puissance du futur :
http://www.lemonde.fr/idees/article/2013/01/07/les-donnees-puissance-du-futur_1813693_3232.html

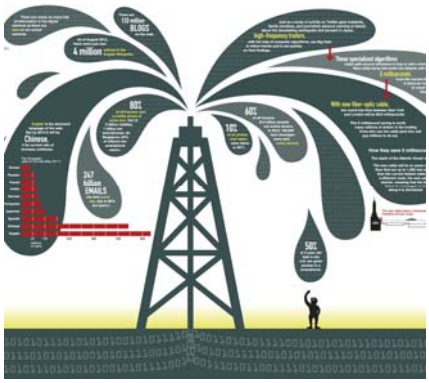
Que faire de ces données ?

Si l'on comprend bien le phénomène contemporain d'accumulation des données, il est peut-être plus difficile de percevoir en quoi ces données, mais aussi et peut être sûrement le fait que l'on soit en capacité de les traiter, soit un puissant opérateur de changement. La science, la connaissance s'appuie notablement sur la statistique, sur le comptage, etc. À partir du moment où l'on peut traiter exhaustivement un ensemble de données, où l'on peut effectuer des croisements et des tris à une échelle à peine imaginable il y a encore quelques dizaines d'années, ce sont toutes les méthodes d'analyse de notre environnement qui changent et qui se trouvent démultipliées. L'importance de ces changements est détaillée par Simon Chignard, dans un entretien paru sur le site millenaire3.com assorti d'une page de synthèse très éclairante : « Matière brute de l'information permettant la compréhension d'un phénomène, d'une réalité, la donnée est un outil d'aide à la gestion et à la décision pour les services urbains (voirie, eau, propreté), et d'évaluation des politiques publiques ».

Références

7 – A qui servent les données ?
http://www.millenaire3.com/A-qui-servent-les-donnees.1389.0.html
8 – Entretien avec Simon Chignard :
http://www.millenaire3.com/Simon-CHIGNARD-Les-grands-enjeux-de-l-opendata.122+M59e1d81e95d.0.html
9 – Les domaines du data mining :
http://datamining.blogspirit.com/4_-_les_enjeux/

2 – Des données qui ont de la valeur



Si les milieux de l'informatique sont les mieux à même de saisir les potentialités marchandes de l'accumulation et du traitement des données, il n'en va pas de même partout, où l'idée que les données ont de la valeur fait son chemin plus lentement qu'on aurait pu l'imaginer.

Quelle peut-être la valeur marchande des données ?

La constitution de données via diverses opérations informatiques constitue un potentiel de valeur dont les entreprises n'ont pas toujours conscience. Même si elles ne savent pas forcément les exploiter elles-mêmes, elles détiennent parfois des ressources qu'elles ne valorisent pas encore « Ces données et leur utilisation sont un enjeu capital pour les entreprises. Le Big Data constitue une véritable manne d'opportunités marketing ».

Références

1 – Big Data : donner de la valeur à vos données :
http://www.sfr.com/les-mondes-numeriques/services-et-innovations/07262012-1747-big-data-donner-de-la-valeur-a-vos-donnees
2 – Open Data : le marché qui valait 40 milliards :
http://www.lemag-numerique-rennais.com/2013/03/open-data-potentiel-economique-3360

Quelles politiques publiques d'incitation ?

Face à cette manne potentielle, les politiques publiques cherchent à inciter les entreprises à valoriser leurs données. Mais il y a aussi un enjeu économique pour les collectivités publiques, qui orientent leurs réflexions sur les moyens de taxer ces données. La question étant de savoir comment générer une fiscalité sur les données, tout en facilitant leur accessibilité.

Références

3 – Entretien avec Nicolas Colin :
http://www.ecrans.fr/Nicolas-Colin-Il-faut-taxer-la,16030.html
4 – Quelles politiques publiques ?
http://www.smartplanet.fr/smart-people/fabien-terraillot-redressement-productif-le-big-data-un-enjeu-de-competitivite-21868/
5 – Comment générer une fiscalité numérique ?
http://www.lesechos.fr/18/01/2013/LesEchos/21357-101-ECH_fiscalite-du-numerique---vers-une-taxation-des-donnees.htm

Des données à protéger qui sont complexes à exploiter

Cependant, cette manne de données pose de nombreux problèmes aux chercheurs spécialisés dans leur analyse. Tout d'abord, ils pointent la nécessité de mieux les protéger et d'assurer leur conservation. Par ailleurs, ils soulignent qu'elles demandent des compétences très pointues pour être traitées de manière à produire des résultats intéressants.

Références

6 – Comment protéger, traiter, valoriser les données ?
http://www.millenaire3.com/Comment-proteger-traiter-valoriser-les-donnees.1390.0.html
7 – Entretiens avec les chercheurs du Liris :
http://www.millenaire3.com/fileadmin/user_upload/interviews/LIRIS-Donnees_et_services-2012._pdf.pdf
8 – Big Data : de grandes quantités de données et des incertitudes :
http://www.smartplanet.fr/smart-technology/big-data-de-grandes-quantites-de-donnees-et-des-incertitudes-18866/

3 – Le data mining et le marketing ciblé

L'une des applications les plus marquantes du data mining réside sans doute dans le renouvellement du marketing, car le data mining permet de toucher très finement des consommateurs en établissant des profils précis et fiables de leurs goûts, de leurs modes d'achats, de leur niveau de vie, etc. De plus, plus n'est besoin de passer par un entretien, ou par du déclaratif : chacun des utilisateurs d'Internet laisse suffisamment de traces lorsqu'il surfe, tweete, publie sur Facebook, pour que son profilage soit possible, à son insu le plus souvent…

Profilage de consommateurs et de cœurs en peine…

Des entreprises, comme Criteo, se sont spécialisées par exemple dans le marketing publicitaire, capable de générer des balises personnalisées (bandeaux publicitaires) envoyées aux internautes en temps réel. Critéo revendique une « capacité d'afficher des bannières personnalisées sur les sites Internet, en les générant à la volée. Près de 300 modèles différents existent pour chaque produit. Le directeur Europe, Grégory Gazagne, explique que « deux personnes qui se rendent sur un site n'auront pas la même version d'une bannière, peut-être même qu'une d'entre elle n'aura pas de bannière Criteo ».

Références

1 – Data mining et marketing :
http://www.creg.ac-versailles.fr/spip.php?article55
2 – Ciblage comportemental :
http://fr.wikipedia.org/wiki/Ciblage_comportemental
3 – Retargeting ou reciblage publicitaire :
http://www.journaldunet.com/ebusiness/publicite/dossier/qu-est-ce-que-le-retargeting-ou-reciblage-publicitaire/qu-est-ce-que-le-reciblage-publicitaire.shtml
4 – Le reciblage publicitaire selon Criteo :
http://pro.clubic.com/webmarketing/actualite-538096-marketing-paris-criteo.html

Ciblage électoral

Avec ces technologies nouvelles, il est maintenant possible de s'adresser de manière très fine et personnalisée à l'électorat. Un atout précieux pour les candidats, qui doivent rassembler sur leur nom des électeurs de plus en plus individués, qui présentent des profils dissonants. Auparavant, l'appartenance religieuse par exemple conditionnait souvent l'orientation politique. Ce type de lien est aujourd'hui de moins en moins vrai aussi faut-il croiser de nombreuses caractéristiques pour espérer définir les orientations politiques des électeurs. Avec le data mining, il est possible de faire des tris croisés très sophistiqués. Ainsi le microciblage permet « de cibler non pas « un

marché », « le public » ou « les électeurs », mais des segments bien plus précis au sein de ces catégories ».

Références

5 – Le data mining, l'arme secrète d'Obama pour gagner :
http://www.slate.fr/monde/63859/obama-arme-secrete-gagner-data-mining-microtargeting
6 – Obama expert en stratégie média innovantes :
http://www.docnews.fr/actualites/tribune,obama-elu-meilleur-responsable-crm-monde,35,15018.html
7 – La campagne numérique d'Obama :
http://fr.slideshare.net/julie_creuilly/la-campagne-numrique-dobama

Renouveler l'analyse de l'opinion

Outre le ciblage des électeurs, le data mining permet aussi de renouveler les sondages, en améliorant très fortement les résultats et ce sans interroger un panel d'électeurs. C'est avec ce type d'algorithmes que le statisticien Nate Silver s'est acquis une très forte crédibilité, en prédisant notamment la victoire d'Obama aux dernières présidentielles américaines ainsi que les résultats qu'il obtiendrait dans les 50 États de l'Union… Nate Silver tient cependant à relativiser son travail : « Ce n'est en aucun cas révolutionnaire, comme j'ai pu l'entendre. Nous avons simplement utilisé des données qui existaient et nous les avons pondérées selon notre modèle ».

Références

8 – Les statistiques supérieures aux sondages ?
http://bigbrowser.blog.lemonde.fr/2012/11/09/gourou-des-stats-nate-silver-lautre-gagnant-de-la-presidentielle-americaine/
9 – Nate Silver et les limites du Big Data :
http://www.lemonde.fr/technologies/visuel/2013/03/11/nate-silver-et-les-limites-du-big-data_1846381_651865.html
10 – Nate Silver, saint patron des "nerds"…
http://www.lemonde.fr/technologies/article/2013/05/24/et-nate-crea-le-data_3415955_651865.html

Un nouvel espace pour la recherche en sciences sociales

Vue sous un autre angle, ces données accumulées sont une mine d'or pour les chercheurs. Certains chercheurs en comportement se sont penchés sur les attitudes des internautes utilisant des sites de rencontres. Outre qu'ils constatent que les données qu'ils utilisent sont plus fiables que celles qu'ils obtiennent en rencontrant des individus (on ment plus facilement à un enquêteur qu'à une machine…), ils peuvent faire des analyses peu politiquement correctes mais très instructives !

Références

11 – Sites de rencontres et SHS…
http://www.rslnmag.fr/post/2010/08/30/Les-donnees-des-sites-de-rencontres-en-ligne-nouveau-materiau-pour-les-recherches-en-sciences-humaines.aspx
12 – Les mathématiques de la beauté…
http://blog.okcupid.com/index.php/the-mathematics-of-beauty/

4 – Le data mining outil de prévision

Le data mining, c'est aussi un outil qui permet de démultiplier les propriétés liées au calcul de probabilité. En effet, parce qu'il permet de croiser un volume de données sans commune mesure avec celles habituellement utilisées par les probabilistes, mais surtout, parce qu'il permet d'appliquer ces calculs à de très nombreux domaines, il apparaît aujourd'hui comme capable de faire des prévisions… De la prévision à la prédiction, il n'y a que quelques lettres de différence et aujourd'hui, les scientifiques n'hésitent pas à annoncer qu'ils seront bientôt capables de prévoir l'avenir… Si les deux termes semblent s'opposer –sciences versus prédiction– on doit néanmoins convenir qu'en certains domaines, les outils aujourd'hui en usage offrent des résultats très impressionnants.

Probabilités et prédictions…

Aujourd'hui, la statistique prévisionnelle s'attaque à toutes sortes de questions : catastrophes naturelles, santé, délinquance, climat… Les outils statistiques sont nombreux et sont combinés entre eux pour améliorer les résultats, comme lorsqu'on utilise des « forets aléatoires ». Plus fascinant encore, les logiciels sont capables de s'améliorer eux mêmes et d'accumuler toujours plus de données pour booster leurs performances… En attendant, il est possible de se fier à ces analyses pour tenter d'éviter la grippe ou se faire vacciner à bon escient.

Références

1 – Prévoir en croisant des données :
http://data.blog.lemonde.fr/2013/02/13/3-millions-de-dollars-pour-un-modele-performant/
2 – Les archives du Times pour décrire l'avenir :
http://gigaom.com/2013/02/01/how-two-scientists-are-using-the-new-york-times-archives-to-predict-the-future/
3 – Prédire le futur en fouillant le web :
http://research.microsoft.com/en-us/um/people/horvitz/future_news_wsdm.pdf
4 – Analyser le web pour prévoir l'avenir :
http://www.slate.fr/lien/67943/data-mining-prediction-web
5 – Forêts Aléatoires et data mining :
http://www.decideo.fr/Les-Forets-Aleatoires-en-data-mining_a1119.html
6 – Éviter la grippe grâce au Big Data :
http://www.smartplanet.fr/smart-technology/comment-eviter-la-grippe-grace-au-big-data-21778/

Prévoir ou prévenir les crimes

Si l'idée qu'un logiciel serait capable de prévoir crimes et délits fait irrésistiblement penser au film de Spielberg « Minority report », la réalité a aujourd'hui rattrapé la fiction : le logiciel PredPol (pour predictive policing) permet d'estimer mieux qu'aucune autre technique ou analyse humaine, les lieux où risquent de se produire des délits, et conséquemment de mieux programmer les patrouilles de police et autres dispositifs préventifs.

Références

7 – La fiction rattrapée par le réel : Minority Report (2002) :
http://television.telerama.fr/tele/films/minority-report,1991902.php
8 – Un algorithme pour prévenir le crime :

http://www.rslnmag.fr/post/2012/06/18/Un-algorithme-pour-prevenirlecrime.aspx
9 – Le logiciel qui prédit les délits :
http://www.lemonde.fr/ameriques/article/2013/01/04/le-logiciel-qui-predit-les-delits_1812195_3222.html
10 – Un logiciel pour prévoir les crimes :
http://fr.euronews.com/2013/02/25/un-logiciel-pour-prevoir-les-crimes/

Se prémunir de la fraude

Autres perspectives offertes par le data mining, améliorer la lutte contre les fraudes et les « arnaques » à l'assurance. Là encore, il s'agit de mieux cibler les contrôles et apparemment, cela fonctionne : « Cette technique donne des résultats très nets (...) Dans plus de la moitié des cas, quand un contrôleur va faire un contrôle ciblé sur la base du datamining, il trouve quelque chose » affirme Hervé Drouet, directeur de la Cnaf. Les compagnies d'assurance appliquent elles aussi ce type d'analyses pour déceler les escroqueries.

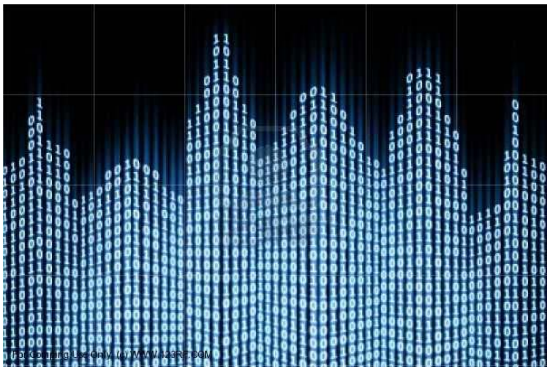
Références
11 – Lutter contre la fraude aux allocations familiales :
http://www.challenges.fr/economie/20130130.CHA5656/les-fraudes-aux-allocations-familiales-ont-depasse-les-100-millions.html
12 – Le datamining pour détecter la fraude à l'assurance :
http://www.analysepredictive.fr/gestion-des-risques/enjeux-risques/le-datamining-pour-detecter-la-fraude-a-l'assurance

Prédire l'avenir ?

Sans nullement prétendre à l'exhaustivité tant la matière est riche sur cette question des prévisions, on retiendra les résultats d'une étude qui permet de prévoir les déplacements des individus en analysant les données de géo-localisation (ou tracking) contenues dans leurs téléphones…

Référence
13 – Comment un téléphone peut prédire les déplacements :
http://www.slate.fr/lien/60475/telephones-suivi-deplacements

5 – Data mining appliqué à la ville



Le data mining est aussi mis à contribution pour améliorer la vie en ville, notamment sur les questions de déplacements. Mais cette gestion de données, compte tenu de la dimension publique, voit s'affronter des tendances contradictoires, notamment sur la question de leur ouverture (open data). Tout le problème étant finalement de définir ce qui peut être fait par des opérateurs privés et ce qui doit, notamment pour des raisons de service public, rester dans le giron public.

Optimiser la ville

Selon le chercheur Olivier Verscheure, les villes ont tout à gagner à se pencher sur leurs données pour modifier la gestion de l'urbain (déplacement, plan d'occupation des sols, gestion des fluides, optimisation des coûts énergétiques, etc) et devenir ainsi des « smarter cities » ou « villes plus intelligentes ». La question est

tellement centrale à un moment où plus de 65% de la population mondiale vivra en en ville d'ici 10 ans, qu'un institut spécialisé vient d'être créé à New-York. Outre les contenus, les données, toutes numériques quelles soient, demandent des infrastructures importantes (et sont de fortement consommatrice d'énergie). Autrement dit, les lieux de stockage aussi contribuent à façonner les villes et ces usines d'un nouveau type ont un impact sur leur environnement : Certains datas centers sont mêmes utilisés pour le chauffage urbain…

Références

1 – Big Data : des villes plus intelligentes :
http://www.smartplanet.fr/smart-people/big-data-mieux-exploiter-les-donnees-des-entreprises-20480/
2 – L'Institute for Data Science and Engineering se concentrera sur les villes intelligentes :
http://www.smartplanet.fr/smart-people/a-new-york-un-institut-sur-le-big-data-va-stimuler-la-recherche-et-leconomie-de-la-ville-19363/
3 – Comment les données reconfigurent la ville :
http://www.urbanews.fr/2013/03/25/30626-comment-notre-societe-des-donnees-va-reconfigurer-la-ville/ - .UeF5hBxVBCc
4 – Se chauffer grâce à l'énergie des serveurs informatiques :
http://www.urbanews.fr/2013/07/01/33942-se-chauffer-grace-a-lenergie-des-serveurs-informatiques/ - .UeGRrRxVBCd

Un mouvement mondial qui donne lieu à une forte compétition

L'optimisation de la gestion des villes via le big data fait l'objet d'une compétition aiguë entre les villes au niveau mondial. Une compétition qui produit de divers classements et indicateurs permettant d'établir un palmarès où sont pris en considération : la mobilité, la connectivité, les smartgrid (réseau de distribution d'électricité « intelligent » qui utilise des technologies informatiques de manière à optimiser la production, la distribution, la consommation d'énergie), les smart objects, la gestion automatisée des transports, la mise à disposition des données publiques…

Références

5 – Villes intelligentes : les 70 villes européennes les mieux notées :
http://blog.econocom.com/blog/villes-intelligentes-les-70-villes-europeennes-les-mieux-notees/
6 – 33 villes décrochent la subvention du défi « Smarter cities » :
http://www.smartplanet.fr/smart-people/33-villes-decrochent-la-subvention-du-defi-smarter-cities-13458/

Récolter les données

Même si les données commencent à s'accumuler, l'enjeu est aussi de savoir les traiter pour répondre aux différentes besoins, attentes, souhaits relatifs à la gestion d'une ville au sens large. A Toulouse une expérience retient l'attention, avant que l'ensemble de la ville ne soient couverte de capteurs, mais là c'est encore de la science fiction.

Références

7 – Toulouse analyse les sentiments des citoyens pour évaluer leurs préoccupations :
http://www.analysepredictive.fr/marketing-predictif/applications-marketing/toulouse-analyse-les-sentiments-des-citoyens
8 – Des puces sur les murs : réalités et fictions :
http://www.culturemobile.net/questions-ethique/rfid-realites-peurs-et-fantasmes/poussiere-intelligente-ville-intelligente

L’exemple de Lyon

À Lyon sont développées diverses expériences portant sur la concrétisation d’une ville rendue « intelligente » par une utilisation efficace des données qu’elle génère. Il s’agit par exemple d’optimiser la gestion des transports dans une optique de développement durable. Par ailleurs, les données sont aussi utiles pour transformer les modalités d’action des agents publics et reconfigurer les métiers. Pour Charles Népote (chef du projet « partage des données publiques » à la FING), « Le premier bénéfice de l’open data, c’est la modernisation interne des collectivités ».

Références

9 – Mobilité durable optimisée à Lyon :
http://www.optimodlyon.com/
10 – La donnée, facteur d’évolution des services :
http://www.millenaire3.com/La-donnee-facteur-d-evolution-des-services.1397.0.html
11 – Mobilité et données publiques, entretien avec Jean Coldefy :
http://www.millenaire3.com/Jean-COLDEFY-Services-de-mobilite-et-donnees-publ.122+M5637722bdee.0.html

6 – Recherche et formation sur le Pres Université de Lyon

On présente ici un rapide repérage des principaux lieux de recherche et de formation sur le PRES Lyon Saint-Étienne dédiés au data mining. On a aussi mentionné dans cette section des colloques organisés sur cette aire géographique car ils informent sur l’actualité de la recherche.

Le laboratoire ERIC

Le laboratoire ERIC pour Entrepôt, Représentation et Ingénierie des Connaissances (ou Équipe de Recherche en Ingénierie des Connaissances) travaille sur l’informatique décisionnelle. L'équipe de chercheurs du laboratoire ERIC s’est spécialisée sur la conception de nouveaux systèmes, modèles et algorithmes pour la fouille de données complexes et l'aide à la décision. Pour manipuler ces données, les chercheurs utilisent des approches principalement statistiques et inspirées de l'intelligence artificielle. Le laboratoire Eric est une unité de recherche (Équipe d’Accueil 3083) multitutelle (universités Lyon 1 et Lyon 2) membre de l’Institut des Sciences de l’Homme.

Références

1 – Laboratoire ERIC Lyon :
http://eric.univ-lyon2.fr/6-FR-prsentation
2 – Introduction au data mining / Laboratoire Eric :
http://eric.univ-lyon2.fr/~ricco/cours/slides/Introduction_au_Data_Mining.pdf

Le LIRIS

Liris : Laboratoire d'InfoRmatique en Image et Systèmes d'information. Ses activités sont regroupées dans deux départements thématiques : "Image" et

"Données, Connaissances, Services". Le Liris regroupe environ 300 personnes, dont près de 110 chercheurs et enseignants-chercheurs. Le LIRIS a 5 tutelles : le CNRS, l'INSA de Lyon, l'Université Claude Bernard Lyon 1, l'École Centrale de Lyon et l'Université Lumière Lyon 2, et des sites à La Doua, Écully et Bron.

Référence

3 – Présentation du LIRIS :
http://liris.cnrs.fr/presentation/presentation-du-laboratoire

Laboratoire Hubert Curien, département informatique, Saint-Étienne

Ce laboratoire s'intéresse à plusieurs sous domaines de l'intelligence artificielle. Il travaille notamment sur l'apprentissage automatique (Machine learning) et la fouille de données (Data mining).

Référence

4 – Laboratoire Hubert Curien, université de Saint-Étienne :
http://depinfo.univ-st-etienne.fr/recherche.php

Master Extraction des Connaissances à partir des Données (ECD) Lyon 2

Ce master repose sur une complémentarité entre une formation théorique prodiguée par des chercheurs de pointe issus des laboratoires ERIC et LINA (Laboratoire Informatique de Nantes Atlantique), et la pratique de logiciels de fouille sur des cas d'étude réels.

Référence

5 – Master extraction des connaissances Lyon 2 :
http://dea-ecd.univ-lyon2.fr/Master_2_Extraction_des_connaissances_a_partir_des_donnees_%28ECD%29.html

Master 2 Data Mining & Knowledge Management à Lyon 2

Le programme Erasmus Mundus Master en Data Mining and Knowledge Management (DMKM) est un master européen de très haut niveau en informatique, soutenu par la Commission Européenne. Il forme des spécialistes dans la Fouille de données et l'ingénierie des connaissances. Ce master, dont la formation se déroule sur 2 ans, est coordonné par Lyon 2 et est organisé par un consortium composé de six universités, réparties dans quatre pays différents.

Référence

6 – Master Data Mining & Knowledge Management (DMKM) :
http://www.univ-lyon2.fr/master-2-data-mining-knowledge-management-dmkm-erasmus-mundus-master-course-428010.kjsp

Master en Machine Learning and Data Mining à l'UJM

Un nouveau master franco-espagnol en Machine Learning and Data Mining à l'Université Jean Monnet de Saint-Étienne. Ce parcours de formation s'appuiera, pour la partie française, sur le master « Web Intelligence » cohabilité par l'Université Jean Monnet Saint-Étienne et l'École nationale supérieure des Mines de Saint-Étienne, et pour la partie espagnole, sur le master « Tecnologias de la Informatica » de l'Universidad de Alicante.

Référence

7 – Master en Machine Learning and Data Mining à l'UJM :
http://www.universite-lyon.fr/formation/un-nouveau-master-franco-espagnol-en-machine-learning-and-data-mining-a-l-ujm-197422.kjsp

Plateforme universitaire d'enquête en data SHS

Dans le cadre du lancement de DATA SHS du Larhra (LAboratoire de Recherche Historique Rhône-Alpes), une journée d'information présente la plate-forme universitaire de données d'enquêtes en sciences humaines et sociales.

Référence

8 – Plateforme universitaire d'enquête en data SHS :
http://25images.ish-lyon.cnrs.fr/player/project_index.php?id=54

Colloque défis de la gestion des grands volumes de données

Colloque : les défis de la gestion des grands volumes de données, Colloque 2013 AIM (Association Information et Management), co-organisé par l'IAE Lyon, EMLYON Business School et l'IUT Lyon 1

Référence

9 – Colloque défis de la gestion des grands volumes de données :
http://iae.univ-lyon3.fr/la-recherche-a-l-iae-lyon/centre-de-recherche-magellan/big-data-colloque-aim-2013-appel-a-communications-597315.kjsp

Enseignement et recherche

Plus largement, l'enseignement et la recherche sur la fouille de donnée est devenu un axe prioritaire et stratégique de formation. Tous les secteurs de formation s'y intéressent, depuis les écoles de commerce jusqu'aux sciences sociales. Dans entretien croisé, Jean-Michel Poggi, professeur de statistique à Paris-Descartes et Hammou Messatfa, responsable gestion du risque chez IBM, se penchent sur les besoins actuels et les enjeux à venir.

Références

10 – Le Big Data, une matière en pointe dans les écoles de commerce américaines :
http://www.smartplanet.fr/smart-business/le-big-data-une-matiere-qui-monte-dans-les-ecoles-de-commerce-americaines-23589/
11 – Big Data : les filières évoluent vers la double compétence :
http://www.letudiant.fr/educpros/entretiens/big-data-les-filieres-evoluent-vers-la-double-competence.html

7 – Les enjeux stratégiques du big data

Les enjeux de cette croissance exponentielle des données sont très vastes et touchent d'une manière ou d'une autre les individus comme les grandes entreprises et les États… Aussi afin de restreindre la focale, nous avons sélectionné des informations qui portent essentiellement sur la manière d'orienter, de travailler sur les données, comment on les conserve, comment on les partage…

Référence

1 – Les données peuvent-elles profiter à tous ?
http://lesclesdedemain.lemonde.fr/organisations/le-deluge-de-donnees-peut-il-profiter-a-tous-_a-12-1696.html

Des enjeux pour les métiers de l'informatique

Au premier chef, ce sont les métiers liés à l'informatique qui se trouvent impactés, car les compétences nécessaires pour être opérationnel sur la fouille de données sont très larges. Un entretien avec Hal Varian (doyen de l'École de gestion et des systèmes d'information et professeur d'économie à l'Université de Californie à Berkeley) donne un éclairage intéressant sur le métier de mathématicien… Par ailleurs, la recherche en data mining suppose de « repenser les outils actuels pour pouvoir profiter des nouvelles opportunités technologiques en matière de puissance de calcul ».

Références

2 – Les métiers de l'informatique en mutation :
http://lesclesdedemain.lemonde.fr/organisations/big-data-quand-le-statisticien-devient-sexy_a-12-1769.html
3 – Quelles orientations pour la recherche en data mining :
http://www.mysciencework.com/fr/MyScienceNews/9720/enjeux-du-data-mining

Travailler sur des périodes plus longues

Selon les spécialistes du data mining, l'un des écueils à éviter est de parvenir à la fois à conserver des données et à envisager des travaux qui portent sur des données accumulées sur des durées longues.

Références

4 – Travailler sur la durée :
http://www.wired.com/opinion/2013/01/forget-big-data-think-long-data/
5 – Sortir de la tyrannie du présent :
http://internetactu.blog.lemonde.fr/2013/02/15/sortir-de-la-tyrannie-du-present/
6 – Favoriser des recherches sur des périodes longues :
http://www.wired.com/wiredscience/2011/10/long-term-datasets/
7 – Comment protéger, traiter, valoriser les données ?
http://www.millenaire3.com/Comment-proteger-traiter-valoriser-les-donnees.1390.0.html

Mieux partager les données

S'agissant du partage des données, qui fait l'objet d'un Pearltrees spécifique (Open Data), on peut noter ici un point de vue original qui consiste à militer pour une mise à disposition des données privées, amassées par les entreprises, auprès d'opérateurs publics ou d'autres, à des fins non commerciales. Il s'agit en quelque

sorte d'une extension du domaine de la philanthropie : le mécénat n'est pas seulement financier, il peut passer par la mise à disposition de biens et services et donc de données…

Références

8 – Développer une philanthropie des données :
http://www.unglobalpulse.org/data-philanthropy-where-are-we-now
9 – La philanthropie des données est bonne pour les affaires :
http://www.forbes.com/sites/oreillymedia/2011/09/20/data-philanthropy-is-good-for-business/
10 – Partager les données pour améliorer leur utilisation :
http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience

8 – Enjeux épistémologiques et débats

Si le big data et le data mining suscitent un enthousiasme certain, ces nouveaux outils font l'objet de mises en garde, d'alertes méthodologiques, voire de critiques parfois virulentes… Ces critiques portent sur les méthodes, sur la croyance générée par la présence de données que l'on envisage spontanément comme fiables, sur le risque de biais méthodologiques, etc.

Ouvrir un débat sans attendre

Pour une partie des analystes, il est urgent d'ouvrir un débat sur le data mining car, comme tout domaine nouveau, il comporte des risques qu'il faut connaître et maîtriser. En effet, « l'accumulation de toutes sortes de données, open ou big data, pourrait faire croire à une source de connaissance en soi. Mais il n'en est rien : les données ne sont jamais neutres ».

Références

1 – Big Data : la nécessité d'un débat :
http://www.internetactu.net/2011/09/23/big-data-la-necessite-d'un-debat/
2 – L'explosion des données : chance ou malheur pour la connaissance ?
http://lesclesdedemain.lemonde.fr/innovation/l-explosion-des-donnees-chance-ou-malheur-pour-la-connaissance-_a-54-2141.html

Le data mining présente des risques d'erreur

De très nombreuses critiques portent sur les erreurs de méthode et les biais présents à la fois dans les données elles-mêmes et dans les outils développés pour traiter ces données. Ainsi, « certains algorithmes introduisaient des biais différents de ceux des humains, mais non moins réels ».

Références

3 – Big data et risques d'erreurs :
http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/
4 – Biais et algorithmes :
http://www.internetactu.net/2012/12/24/contourner-les-algorithmes/

Ce que l'explosion des algorithmes veut dire

Par ailleurs, plusieurs observateurs relèvent l'omniprésence des algorithmes dans nos environnements quotidiens. Ils façonnent alors nos manières de voir, et en d'autres termes, alors qu'ils sont sensés décrire le monde, ils contribuent aussi à le conditionner.

Références

5 – La pertinence des algorithmes :
http://www.internetactu.net/2012/11/29/la-pertinence-des-algorithmes/
6 – Les algorithmes façonnent-ils le monde ?
http://www.rslnmag.fr/post/2011/09/07/Kevin-Slavin-les-algorithmes-faconnent-ils-le-monde-.aspx

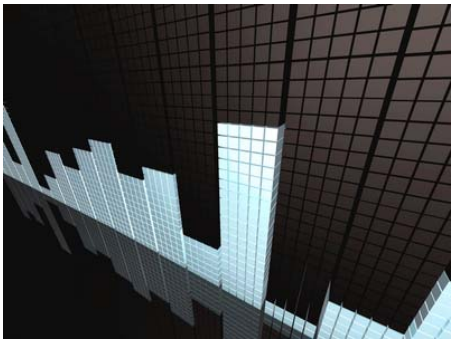
Le data mining ne serait-il qu'illusion ?

Enfin, on ne saurait omettre de mentionner des attaques en règles contre l'accumulation et la fouille de données… Pour Alan Mitchell, directeur du cabinet Ctrl-Shift « les Big Data auraient presque un côté contre-révolutionnaire : le chant du cygne d'une informatique productiviste, centralisatrice, centrée sur les grandes organisations ». Quand au professeur Steve Needel il pose l'hypothèse qu'il qualifie lui même de blasphématoire : et si le big data ne fonctionnait pas ?

Références

7 – La "grande illusion" du big data :
http://www.internetactu.net/2012/04/11/big-data-grande-illusion/
8 – Les confessions d'un blasphémateur : et si le big data ne fonctionnait pas ?
http://www.greenbookblog.org/2012/11/08/confessions-of-a-big-data-blasphemer-what-if-big-data-doesnt-work/

9 – Fouille d'image et d'autres données



Le data mining est maintenant en train de se développer non plus seulement vers les données chiffrées, mais aussi vers les images, les sons, etc.

Références

1 – Data-visualisation, machine learning…
http://www.mysciencework.com/fr/MyScienceNews/9683/data-visualisation-machine-learning

Machine learning ou apprentissage automatique

L'apprentissage automatique (machine learning en anglais), est une discipline qui travaille au développement, à l'analyse et à l'implémentation de méthodes automatisables qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques. Ainsi, des systèmes complexes peuvent être analysés, y compris pour des données associées à des valeurs symboliques (par exemple, une valeur probabilisée, c'est-à-dire un nombre assorti d'une probabilité). Ces analyses peuvent aussi concerner des données présentées sous forme de graphes ou d'arbres ou encore de courbes (d'après wikipedia).

Références

2 – Les fondements de l'apprentissage automatique :
http://mrim.imag.fr/eric.gaussier/M1ATD/FondementsAA.pdf
3 – Initiation à l'apprentissage automatique :
http://labh-curien.univ-st-etienne.fr/~fromont/ML/Cours_TSE-1-1bis_coul.pdf

Data visualisation

Qui est la manière d’exposer, de façon graphique, des données brutes.

Références

4 – Les outils de data visualisation :
http://www.aecom.org/Vous-informer/Actualites2/Les-outils-de-data-visualisation
5 – Rendre visuelles des données brutes :
http://milkcheck.fr/quel-avenir-pour-la-data-visualisation/
6 – Livre blanc sur la data visualisation, le retour de 30 entreprises :
http://pro.01net.com/editorial/598679/la-data-visualisation-dans-le-secteur-it/