ARTICLE DU DOSSIER

Recension
Les Big Data à découvert

Ouvrage collaboratif pluridisciplinaire

Recension de l'ouvrage réalisée par Jean-Christophe Thalabard, Univ. Paris-Descartes

1. Introduction

Sous ce titre, près de 220 chercheurs abordent sous la forme de 150 articles très synthétiques (2 pages), largement illustrés, les formidables potentialités offertes par les nouveaux moyens de stockage des informations et de calcul mais aussi les nouvelles contraintes qui débordent largement le seul domaine scientifique. L'ouvrage, dense (368 pages), est divisé en 9 parties, même s'il existe de larges recoupements des thèmes abordés, s'efforçant d'apporter des points de vue complémentaires, entre les parties. Il convient peut- être de noter ici que ni le titre de l'ouvrage, ni les intitulés des parties successives et encore moins les intitulés des différentes contributions ne renferme directement le terme « intelligence artificielle » (IA), que le grand public associe volontiers à « Big Data ». La porte d'entrée de l'ouvrage est donc la donnée massive, ou plutôt les différentes formes de données massives et leurs traitements et les conséquences sociétales et d'organisation que cela peut avoir. Ainsi il semblerait que pour les experts des données et de leurs traitements, l'IA ne pourrait se saisir qu'à travers les particularités de chaque domaine, en n'existant qu'à travers la somme, souvent gigantesque, de données amassées, bien loin encore de la souplesse et de l'adaptabilité de l'intelligence humaine multi-objectifs dès les premiers âges.

La première partie « **Big Data : enjeux et défis** » rappelle l'importance prise par les données massives dans des domaines aussi divers que le commerce, l'industrie, la météorologie, la géographie, la santé, sans taire les risques de dérives et les questions concernant l'individu et la société.

La partie 2 « **Données, acquisition, stockage** » indique les particularités des données de différents domaines (vivant, collections muséologique, littérature, multimédias) avant d'interroger la nature des métadonnées et leur rapport aux connaissances. Elle se poursuit par des considérations sur la qualité des données collectées suivies d'une série d'articles sur les capteurs, leurs différentes natures, leurs autonomies et leurs capacités de communication, leurs organisations en réseaux. Elle se termine par les aspects de stockage en lien avec les notions cruciales de préservation et d'archivage, de coûts énergétiques et d'impact écologique.

La troisième partie, « **Traitement des données** », décrit notamment différentes approches de traitement des données : algorithmes, interfaçages et interrogations des bases de données, intégration de données massives, hétérogènes et équidistribuées, compression, sans oublier

les aspects sécurité et d'architecture de traitement préservant la vie privée.

La quatrième partie, « **Analyse de données et apprentissage** » regroupe des thèmes variés allant de modèles de prévision à visées industrielles ou commerciales (transports, énergie), à la fouille de données (textes), aux modèles d'apprentissage (neuroscience, imagerie, reconnaissance de formes).

La cinquième partie, « **Web, réseaux sociaux et recherche d'information** » aborde l'utilisation faite des ressources accessibles via les moteurs de recherche, leurs organisations, l'accès aux données personnelles avec des conséquences sur la vie sociale, la diffusion des opinions et l'apparition de phénomènes extrémistes.

La sixième partie, « **De l'infiniment petit à l'infiniment grand** » rentre dans l'exploitation des données massives dans des domaines d'échelles variées allant de l'astronomie et l'exploration de l'espace, à la climatologie à la physique des particules en passant par le décryptage du vivant (génoscope et génomique, biologie), l'étude des écosystèmes (données Tara par exemple pour l'écosystème océan).

La septième partie, « **Santé humaine** », est plus particulièrement consacrée aux enjeux des « Big Data » dans différents champs d'activités et d'instrumentation touchant à la santé humaine : objets connectés, imagerie, cancérologie, chirurgie, environnement, sans oublier les aspects éthiques.

La huitième partie, « **Individu et Société** » reprend et développe, sous un angle sciences sociales, les impacts de nouvelles pratiques et modes de vie associés à des collections et échanges de flux de données (tels tweets, Smart Grids, Smarthome). Quels sont les enjeux sociaux ? Quelle(s) place(s) de l'humain pensant par rapport aux approches robotisées ? Quel est le futur de l'entreprise ? Quelle place pour les choix individuels ? Quel droit à la déconnexion ?

Une dernière partie de **conclusions** laisse la parole à des acteurs reconnus de la robotique, des neurosciences, de la physique théorique, du monde technologique, des sciences sociales de dresser quelques perspectives et interrogations à partir de leurs champs respectifs pour rejoindre des préoccupations plus générales.

2. Une tentative d'analyse graphique de l'ouvrage

Qui sont ces acteurs qui, sous la coordination d'un enseignant-chercheur en informatique, Mokrane Bouzeghoub et d'un chercheur du CNRS en physique théorique, Rémy Mosseri, ont apporté leur contribution à cet ouvrage ? Domaine oblige, nous nous sommes amusés à soumettre l'ouvrage à un outil d'analyse des larges bases de données, *linkage.fr*, que nos lecteurs ont pu découvrir lors d'un récent numéro de *Statistique et Société* (sous la plume de leurs auteurs, C. Bouveyron (INRIA/Univ. Nice) et P. Latouche (Univ. Paris Descartes)[1].

Il nous a fallu faire quelques choix. Nous avons choisi de garder des informations factuelles concernant les auteurs, leur appartenance, leur genre, leurs publications telles qu'apparaissant dans la base de données internationale Pubmed (nous nous sommes donc restreints aux liens pouvant exister avec les domaines biologie et/ou santé) nous donnant accès à leurs co- auteurs, leurs appartenances et les titres des journaux concernés.

1. Présidentielle 2017 : l'analyse des tweets renseigne sur les recompositions politiques - Pierre LATOUCHE, Charles BOUVEYRON, Damien MARIE, Guilhem FOUETILLOU. Statistique et Société, Vol. 5 No 3 (2017) - Ouvrir les données. Pages 39-44. http://statistique-et-societe.fr/article/view/660

Nous ne pouvons qu'insister sur le caractère très parcellaire de notre analyse en ce qui concerne le réseau collaboratif autour du seul domaine, déjà très vaste, des publications concernant les domaines de la biologie et de la santé.

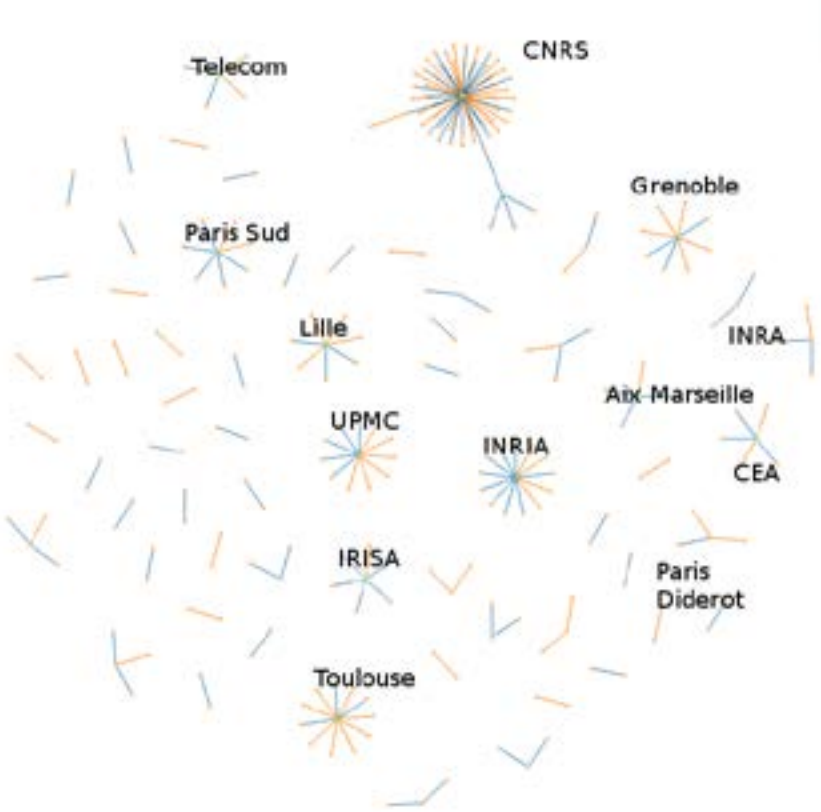
Préparer un tel fichier, qui reste plus que modeste dans le monde des « Big Data » (22°000 lignes x 8 colonnes°!) permet également de voir quelques limites de ces aspirations (via le package easyPubmed de R) d'une base de données comme Medline. En effet, les appartenances, les mots clés, les adresses sont fonctions de l'alimentation de la base qui dépend fortement des contraintes éditoriales de tel ou tel article et de sa présence dans ladite base de données. Le travail qui consiste à combler les lacunes d'un logiciel de reconnaissance automatique d'un pays ou d'une ville reste imparfait, imposant une vérification visuelle et manuelle des données incomplètes que des esprits algorithmiques s'empresseraient de vouloir imputer…

Nous nous contenterons ici de souligner les nombreux pièges qui peuvent apparaître dans toute illusion d'analyse automatique avec des outils facilement disponibles et imposent encore un certain savoir-faire vigilant.

Il convient sans doute de préciser que de tels outils sont essentiellement des moyens de représentation de données complexes, ouvrant la place à des interprétations personnelles, par nature, subjectives. La valeur d'une prédiction ne saurait se substituer à une quelconque preuve de causalité.

2.1 Répartition des auteurs par laboratoires

Figure 1 : représentation de l'ensemble des contributeurs rapportés à leur laboratoire / structure d'appartenance. Pour des questions de lisibilité, nous n'avons labellisé que les nœuds les plus denses, où se retrouvent les acteurs connus du domaine. Il convient cependant de souligner la forte inter-dépendance entre les grands organismes (CNRS, INRA, INRIA, CEA…), dont l'appréhension ici dépend fortement des appartenances indiquées par les auteurs.



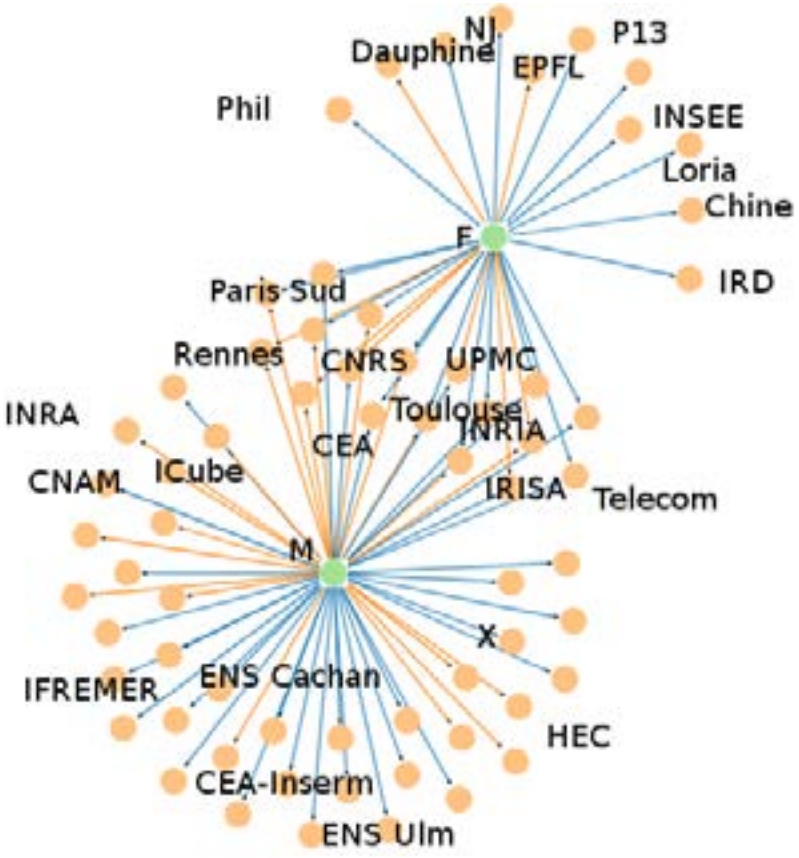
2.2 Répartition des auteurs par genre

Factuellement, parmi les 220 auteurs, 62 sont des femmes (28 %). La répartition selon les 9 parties de l'ouvrage est représentée dans le Tableau 1. Il y apparaît une nette prédominance masculine dans la plupart des parties, à commencer par l'introduction (partie 1) et la conclusion (partie 9), avec une exception pour la partie 3 « Traitement de données ». Dix-neuf de ces articles étaient écrits avec une seule auteure, vingt-cinq étaient écrits avec un ou des co- auteurs masculins. Trois étaient écrits par un binôme d'auteures et un par un trio. La répartition par appartenance est représentée sur la Figure 2.

Tableau 1 : Répartition des auteurs par genre en fonction des parties thématiques de l'ouvrage

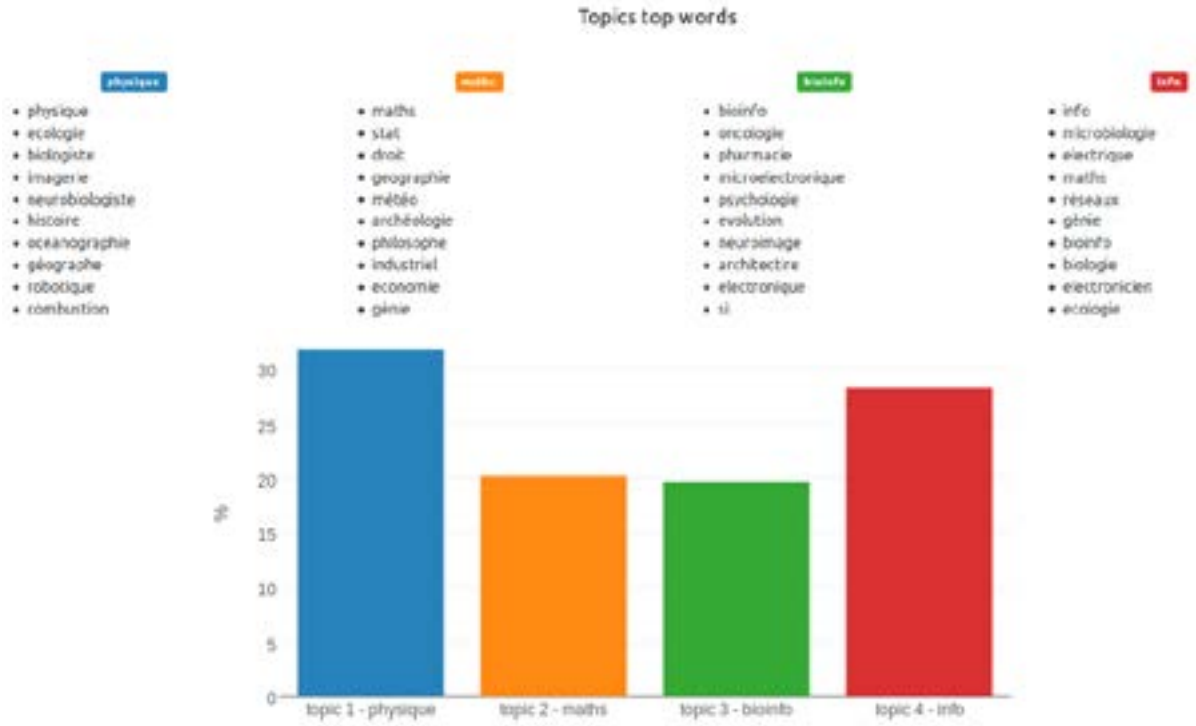
	Parties de l'ouvrage									Total
	1	2	3	4	5	6	7	8	9	
Femme	6	10	14	6	3	9	3	10	2	63
Homme	11	18	13	25	21	28	11	19	11	157
Total	17	28	27	31	24	37	14	29	13	220

Figure 2 : Distribution des contributeurs en fonction de leur genre (F : Féminin ; M : Masculin). La couleur des arêtes indique le champ disciplinaire : bleu pour informatique ; orange pour application aux domaines (biologie/ santé/ écologie...). Quelques institutions représentatives sont indiquées au niveau des nœuds. Le graphe suggère des interactions féminines plus tournées vers le monde universitaire et l'étranger, les acteurs majeurs du domaine (CNRS, INRIA, IRISA, CEA) semblant accueillir indifféremment des acteurs des deux sexes, avec une répartition similaire entre les développements informatiques et leurs applications.



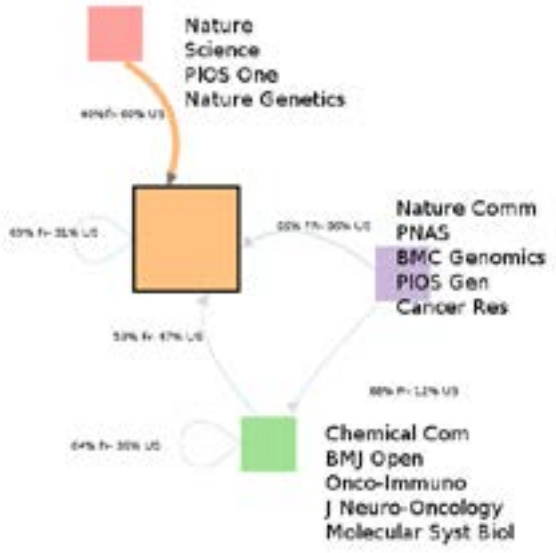
2.3 Regroupements des contributeurs (clusters) et leurs interactions (topics)

Figure 3 : Distribution des thèmes d'échanges (topics) restreints à un maximum de 4 grands thèmes entre les différents auteurs regroupés en 4 « clusters », de tailles relatives très différentes, correspondants aux poids des appartenances des contributeurs de l'ouvrage aux différents organismes et institutions. On notera la position « discrète » de la Statistique dans la colonne « maths »



2.4 Publications

Figure 4 : Étude des liens entre chercheurs à partir des publications communes. Le carré central, orange, représente l'ensemble des contributeurs et leurs co-auteurs tels que retrouvés dans PubMed. Les carrés périphériques correspondent à des grands types de journaux dont les principaux sont indiqués en regard. Les flèches indiquent les natures (en pourcentage) des collaborations franco- françaises versus franco- étrangères, dont la contribution majoritaire apparaît être nord- américaine



3. En guise de conclusion

Si la lecture de l'ouvrage peut inciter le lecteur curieux à s'essayer aux techniques d'analyses et de synthèses automatisées, rien ne peut remplacer encore une lecture plus traditionnelle au hasard des parties et des contributions, faisant découvrir la richesse des multiples points de vue au gré des inclinations du lecteur.

Il convient certainement de saluer les 2 coordinateurs de l'ouvrage, Mokrane Bouzeghoub et Rémy Mosséri, pour avoir réussi une véritable gageure de réunir une large palette d'experts pour produire cette somme d'articles courts, percutants, bien illustrés, toujours accompagnés de quelques références bibliographiques récentes, qui ne peut qu'intéresser l'honnête homme du 21ème siècle soucieux de saisir dans son élaboration l'état des connaissances portées par les multiples acteurs actuels d'un domaine aux contours par nature évolutifs qui bouge les frontières des savoirs traditionnels, tout en questionnant leurs impacts sur les sociétés, leurs économies et leurs régulations.