Journal IRIT : Noir Sur Blanc, numéro thématique "La science des Données"

Nathalie Aussenac-Gilles, Mohand Boughanem, Philippe Joly, Michelle Sibilla

ACTUALITÉ

MARS 2016

20

NUMÉRO

noir SUR blanc

IRIT
CNRS - INPT - UPS - UT1 - UT2J

De l'institut de recherche en informatique de toulouse

Fouille de données

Optimisation

LA Science des Données

Analyse de vidéo

Simulation

Analyse d'images

Directeur de la publication : Michel DAYDÉ

Secrétariat de rédaction : Véronique DEBATS
Conception et création de la maquette : Service Communication IRIT

Ont collaboré à ce numéro : Nathalie AUSSENAC, Mohand BOUGHANEM, Philippe JOLY, Michelle SIBILLA et les membres des équipes impliquées dans l'axe stratégique MDC.

Contact de la rédaction : 05 61 55 65 10 - communication@irit.fr - www.irit.fr
118 Route de Narbonne - 31062 Toulouse cedex 9

IRIT
CNRS - INPT - UPS - UT1 - UT2J

cnrs · INP TOULOUSE · UNIVERSITÉ TOULOUSE III PAUL SABATIER · UNIVERSITÉ TOULOUSE 1 CAPITOLE · UNIVERSITÉ TOULOUSE Jean Jaurès

ISSN 1620-2937

Michel DAYDÉ
Directeur de l'IRIT

Nous vivons depuis quelques années une révolution appelée « **Big Data** » conduisant à un nouveau paradigme selon lequel la science est dans les données. L'exploitation des informations contenues dans les données qu'elles soient issues des expériences scientifiques, des capteurs ou des réseaux sociaux est dorénavant une approche répandue. Cependant, pour nombre de communautés scientifiques, cette pratique est au cœur de leur recherche depuis longtemps, par exemple en physique des particules (LHC au CERN), en sciences de l'univers (astronomie, astrophysique, géosciences, …) en biologie (séquenceurs de nouvelle génération, dans le domaine médical), etc.

Le ***Big Data***, autrement dit les **données massives**, concerne le traitement d'ensemble de données qui deviennent tellement volumineuses qu'il est compliqué de les exploiter avec des outils classiques de gestion (bases de données ou autre). On le caractérise souvent par les 3 « V » : Volume, Variété, Vélocité auxquels on ajoute parfois 2 autres « V » : Véracité et Valeur.

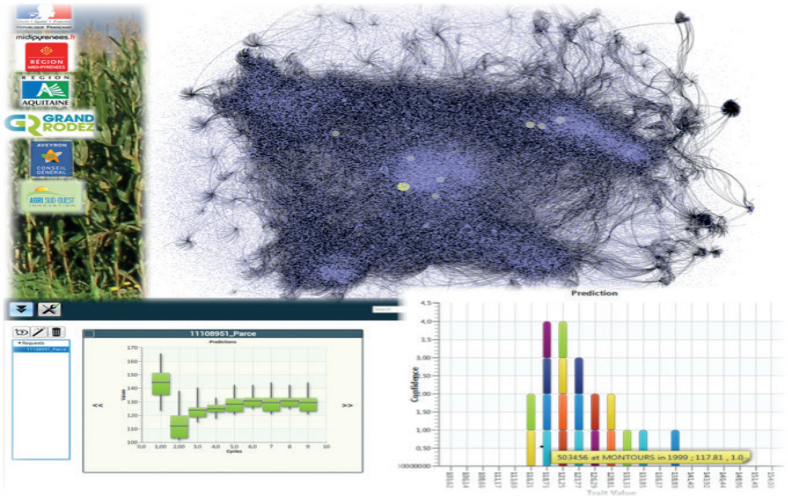
Ce succès du Big Data est tel qu'il est considéré aujourd'hui comme le quatrième pilier de la Science, les trois premiers étant la théorie, l'expérimentation et enfin la modélisation et la simulation.

L'extraction de connaissances, l'apprentissage, la fusion de données, la visualisation et la navigation dans de grands espaces de données sont donc autant d'instruments qui permettent de mieux comprendre des phénomènes ou d'élaborer de nouveaux modèles ou de nouvelles hypothèses.

La communauté informatique effectue des contributions incontournables au Big Data :

- de par les recherches qu'elle développe quasiment depuis ses origines à la fois par les outils, méthodes et nouvelles approches permettant d'aborder le Big Data à tous les niveaux, ce que nous appelons la « **Science des données** »,
- mais aussi par ses travaux sur les masses de données issues par exemple des réseaux sociaux, du Web, de l'imagerie médicale, des réseaux de capteurs, …

La **Science des données** est donc l'extraction de connaissances de données en couvrant tous les aspects liés à cette activité c'est-à-dire en allant des infrastructures matérielles et de communication, à l'aide à la décision et à la visualisation, en passant bien entendu par l'analyse de données. Elle s'appuie sur des méthodes mathématiques, statistiques, la théorie de l'information, le traitement de l'information (y compris traitement de signal), l'apprentissage (automatique ou non), l'ingénierie des données, la visualisation, l'aide à la décision, les problématiques liées au stockage de données, les infrastructures distribuées à grande échelle (grille et Cloud) et le calcul à haute performance.



Visualisation de résultats d'analyse de données génomiques - Projet GBds

La **Science des données** ne se limite évidemment pas au **Big Data** même si elle en constitue aujourd'hui un point focal.

L'IRIT, de par sa taille et sa structuration, couvre tous les aspects liés à la **Science des données**, ce qui en fait un laboratoire quasiment unique en France. L'un de ses axes stratégiques est consacré aux masses de données et calcul. Il dispose de plus d'un contexte extrêmement favorable avec :

- la plateforme en recherche d'information **OSIRIM** installée au laboratoire et largement accessible à l'extérieur,
- le **LabEx CIMI** qui lie fortement les communautés de recherche en informatique au sein de l'IRIT et en mathématiques au sein de l'IMT,
- la forte présence de **multiples communautés** en prise directe avec les problématiques **Big Data** sur le site toulousain (sciences de la vie et de l'univers, ingénierie, aérospatial, agronomie, médical, SHS, …), ce qui en fait une priorité dans plusieurs universités,
- une participation active au nouveau **GDR MaDICS** consacré aux Masses de données scientifiques.

Ce numéro du *Noir sur Blanc* est l'occasion d'expliciter notre positionnement spécifique sur la **Science des données** et de montrer en quoi il peut contribuer aux problématiques liées aux **Big Data** des autres communautés scientifiques.

Michel Daydé

La science des données

Les grandes masses de données : un nouvel enjeu

L'informatique (définie comme « science de traitement de l'information ») traite de plus en plus de données numériques produites en très grandes quantités par les entreprises, la recherche scientifique, le web, les objets connectés, les utilisateurs de logiciels ou de réseaux sociaux, … Depuis quelques années, une accélération dans le volume, la diversité des contenus et des formats de ces données conduit à parler de Big Data (masses de données). Plus que la quantité des données, ce terme évoque le fait d'atteindre les limites de capacité et de traitement des systèmes informatiques, des processeurs et bases de données, mais aussi des organisations et des modèles théoriques, pour gérer et appréhender ces données. Or celles-ci constituent un gisement à exploiter pour révéler des connaissances, imaginer de nouveaux services ou de nouveaux produits. Donc si les Big Data sont si « grandes », c'est aussi par leur valeur. L'exploitation systématique, à grande échelle, de données et de traces numériques issues de sources multiples, en fait à la fois une richesse convoitée et l'objet de méfiances car ces données concernent chacun de nous. Les traces d'usages sur le web ont fondamentalement modifié la manière de connaître les habitudes et goûts des consommateurs, révolutionné les pratiques de vente et défini de nouveaux modèles économiques, perturbant des secteurs comme la presse ou la production musicale. La recherche scientifique est également fortement modifiée dans des disciplines aussi variées que l'astronomie, la biologie ou la génétique : à côté de chercheurs menant des expériences et d'instruments collectant des données numériques de façon accélérée, comme les images d'observation du ciel produites par les télescopes, les mesures d'analyses biologiques ou médicales, les mesures de capteurs, etc., d'autres chercheurs produisent des connaissances en se consacrant à la fouille et l'analyse de ces gigantesques volumes de données.

Le défi est donc désormais de donner une réalité à ce potentiel de valorisation des données, et de dépasser les problèmes générés par leur volume, leur complexité, leur hétérogénéité, leur préservation, leur répartition spatiale sur le cloud, leur qualité, leur traitement quasiment instantané et distribué, la richesse des traitements visés, tout en respectant les utilisateurs et les personnes qu'elles concernent. Gérer, traiter automatiquement et analyser ces Big Data requiert non seulement des algorithmes et des modèles mathématiques, pour beaucoup statistiques, qui font l'objet des *Data Sciences* (ou *data analytics* ou *business intelligence*), mais aussi des architectures et infrastructures de calcul et de stockage, des choix d'implémentation, des études de complexité, des formats facilitant les traitements et l'interopérabilité … tout autant de questions pour la recherche en informatique. Traiter ces données concerne également le droit et l'éthique pour la protection des données et des personnes, et suppose des partenariats avec les disciplines productrices et utilisatrices de données.

La recherche au service des Big Data requiert des avancées tout au long du cycle de vie des données,

L'axe Masse de Données et Calcul (MDC)

Nathalie AUSSENAC-GILLES
Directrice de Recherche CNRS

Mohand BOUGHANEM
Professeur UT3 Paul Sabatier

Co-responsables de l'axe MDC

axe-mdc@irit.fr

www.irit.fr/-Masses-de-donnees-et-calcul,677-

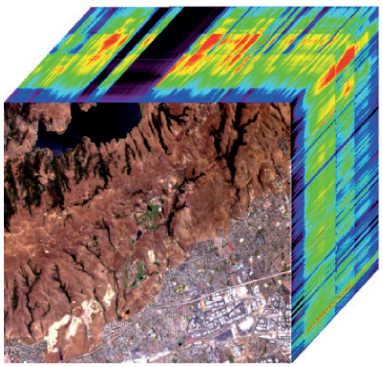
et couvre donc un spectre vaste de domaines de l'informatique et des mathématiques, allant des infrastructures de stockage aux modalités de visualisation en passant par l'étude des modèles, des algorithmes d'analyse et de leur implémentation. Nous présentons dans la suite, les paradigmes scientifiques développés à l'IRIT, de la collecte à l'exploitation des masses de données, avec une contribution plus marquée sur leur analyse et leur exploitation, ainsi que les travaux portant sur les infrastructures requises pour gérer au mieux la masse et la complexité des données et des traitements.

Collecte et extraction de données et de métadonnées

Le recueil et la collecte de données repose désormais sur des technologies de plus en plus diverses (télescopes, satellites, réseaux de capteurs, objets connectés et géolocalisés, etc.) dont la sophistication augmente la qualité et la fréquence des données produites.
L'IRIT appuie ses recherches sur des données aussi variées que des données ouvertes (données publiques des communes, de transport, etc.) et données liées sémantiques, des données scientifiques en sciences humaines (patrimoine et archéologie) ou en recherche médicale et biologique pour la simulation du vivant, des données de capteurs (bâtiment intelligent, site universitaire connecté et intelligent de neOCampus), imagerie satellite, données de réseaux sociaux (micro-blogs, blogs, forums, etc.), corpus textuels (pages web, collections d'articles, bibliothèques numériques, etc.), données techniques (mesures de vol, données de diagnostic, etc..), données d'ingénierie (mécanique des fluides, calcul de structure, géosciences, conception de nouveaux matériaux) ou flux audio-visuel (TV, vidéo-surveillance, etc.).

Dans les phases préliminaires du traitement des données, un premier défi est de filtrer les données brutes en temps réel, pour éliminer les données non pertinentes, inutiles, erronées ou redondantes, et ainsi réduire le volume à stocker.
Un deuxième défi est la capacité d'enregistrer ces données dans des infrastructures adaptées, en garantissant leur qualité. Un troisième enjeu est de générer en même temps des métadonnées adéquates en vue de permettre leur réutilisation et leur analyse en tenant compte de leur contexte de recueil et de leur provenance. Il s'agit enfin d'assurer une qualité minimale des données en éliminant celles non conformes à des contraintes connues a priori.

Par exemple, dans le cadre d'analyse d'images médicales ou satellitaires, ces premières étapes consistent en des analyses hyper-spectrales fines des images pour reconnaître automatiquement des matériaux ou des objets. Une excellente précision pour caractériser ces zones permet d'améliorer la performance des analyses multi-échelles de plus haut niveau et de mieux préserver la configuration originale et la nature continue du paysage. Dans le projet MUESLI (prix INP INNOV 'perspectives scientifiques' 2015), l'équipe SC développe cette approche au service de la surveillance des écosystèmes. Des outils statistiques extraient des informations des cubes de données hyperspectrales et LIDAR (plusieurs TéraOctets) pour identifier la nature des matériaux composant des objets d'intérêt sur les images. Dans le projet SparkinData (cf encadré p. 12), le filtrage d'images exploite à la fois l'analyse de données numériques et la caractérisation sémantique de zones ainsi identifiées.



Cube de données hyperspectrales

Un autre exemple, l'analyse et la recherche de séquences vidéo, montre l'intérêt de combiner différents types de données venant de sources complémentaires pour faciliter l'identification de séquences pertinentes et ensuite leur analyse. Ainsi, l'équipe SIG a proposé une méthode de filtrage spatial et temporel d'un flux de vidéo-surveillance, et une approche basée sur des métadonnées fournies par des victimes ou associées aux capteurs, pour faciliter l'identification de séquences relatives à un événement.

Le Centre International de Mathématiques et d'Informatique de Toulouse (CIMI) est un des laboratoires d'excellence sélectionnés dans le cadre des Programmes d'Investissement d'Avenir. Il regroupe des équipes de l'IRIT et de l'Institut de Mathématiques de Toulouse (IMT) afin de développer des synergies entre les mathématiques et l'informatique.

Les données sont porteuses de problèmes autour de la modélisation, de la représentation et de l'accès qui intéressent aussi bien les chercheurs en informatique qu'en mathématiques. Elles ont, de ce fait, conduit à l'organisation de rencontres scientifiques sur des sujets tels que le calcul haute performance ou l'apprentissage. D'autres évènements de cette nature seront programmés dans le prochains mois dans le cadre du CIMI.

http://cimi.univ-toulouse.fr

Enfin, pour mieux modéliser des connaissances, rechercher des informations ou analyser le langage naturel (analyse sémantique du lexique et du discours), des algorithmes exploitent de grandes collections de textes tirés du web ou d'applications. Au-delà des prétraitements classiques (élimination de mots vides, analyse syntaxique ou calcul de dépendances), l'IRIT développe des méthodes pour préparer des index à base de mots-clés ou de concepts (IRIS), et des approches à base d'apprentissage automatique pour extraire la structure logique d'un texte à partir de sa mise en forme matérielle (collaboration MELODI - ELIPSE), mais aussi identifier des unités de discours et des relations discursives à partir de l'analyse du langage naturel. Ces informations constituent autant de traits supplémentaires pour améliorer les analyses de plus haut niveau. Ainsi, MELODI étudie l'exploitation statistique de très grands corpus (GigaWord, Wacky Corpus) et l'analyse des distributions des mots et de leurs voisins afin de rendre compte de la sémantique lexicale des mots et de leur composition [1].

Agrégation, intégration, modélisation et stockage des données

La plupart des projets Big Data font appel à des données provenant de différentes sources dont le stockage puis le traitement à grande échelle qui nécessitent des solutions d'intégration de données hétérogènes massives, et la définition de modèles optimaux de représentation physique et logique. L'intégration et la modélisation conditionnent un accès efficace à l'information, ainsi que la qualité et la durée des traitements et analyses. Dans le cas de données hétérogènes dynamiques, les principes communément adoptés jusqu'ici, y compris de type OLAP, sont remis en cause. C'est le cas en particulier dès qu'une application intègre des données de réseaux sociaux. L'équipe SIG propose de nouveaux modèles de données assurant la réduction progressive de données et de nouveaux concepts dédiés à la modélisation multidimensionnelle (faits réflexifs, nouveaux types d'indicateurs, attributs calculables lors des analyses, attributs optionnels etc.). À partir de là, de nouveaux opérateurs d'analyses OLAP adaptés à des données réduites, calculables voire absentes sont définis [2].

Le stockage des mégadonnées est généralement distribué au sein de Clusters de machines, de Grilles ou sur le Cloud. En collaboration avec Capgemini, l'équipe SIG étudie l'utilisation des nouveaux systèmes de gestion de données, dits NoSQL. Elle a développé un benchmark pour évaluer leur adéquation à différents types et volumes de données (cf encadré, thèse Mobidick p. 13) [3].

À cette étape se pose aussi le choix de modèles assurant une représentation homogène, fidèle et riche des informations avant leur intégration. Les langages standards du web sémantique et des ressources telles que les ontologies, en associant des catégories sémantiques aux données, permettent à la fois de les enrichir et de leur conférer du sens hors de l'application qui les produit. Les données sémantisées acquièrent une valeur intrinsèque car elles peuvent être traitées globalement, intégrées à des données d'autres sources, interrogées par requêtes ou exploitées pour inférer de nouvelles connaissances.
La représentation conceptuelle s'applique à des données numériques tout comme à des documents textuels (équipes SIG, IRIS, MELODI), des images ou des vidéos. Elle permet aussi de mieux interroger des données en utilisant le langage naturel et d'évaluer des requêtes en fonction de leur structure ou de leur contenu. Dans ce cadre, MELODI développe des techniques de construction d'ontologies, d'alignement et d'interrogation de données sémantiques. L'équipe ADRIA définit des représentations et fournit des outils mathématiques pour la gestion de l'incertitude dans ces modèles.

Exploitation des données : recherche, accès et fouille de données

L'exploitation des données est traitée à l'IRIT selon deux angles complémentaires : des approches visent à faciliter la recherche et l'accès aux objets stockés en réponse à des besoins d'utilisateurs ; et en aval de ce premier angle, l'analyse des données cherche à élaborer des objets à forte valeur ajoutée, par des techniques de fouille, d'apprentissage automatique, d'agrégation et de raisonnement.

Dans la première perspective, des **modèles de recherche d'information** (RI) ont été définis à l'IRIT pour traiter de vastes collections d'informations hétérogènes et complexes (i.e., RI contextuelle), fortement dynamiques (i.e., RI sociale), ou encore nécessitant la coopération de différents acteurs pour optimiser la pertinence.
Sur ce dernier point, IRIS se démarque par l'angle théorique choisi pour formaliser des modèles de RI collaborative, alors que le vecteur central dans l'état de l'art concerne la conception de supports à l'interaction collaborative. IRIS a proposé une nouvelle approche de médiation en RI collaborative, basée sur l'apprentissage dynamique des rôles des utilisateurs et leur injection dans un modèle d'ordonnancement de documents [4]. Pour connaître et modéliser les besoins d'utilisateurs, les jeux sérieux à destination du plus grand nombre (via le crowd sourcing) constituent une des méthodes innovantes et assurent de recueillir de grandes quantités d'informations.

Ainsi le projet MyBestQuery (équipe SIG) s'appuie sur un jeu en ligne qui collecte massivement des évaluations de requêtes et des propositions de reformulations pour prédire la difficulté de requêtes selon des méthodes d'apprentissage supervisé.

Interroger des données massives et distribuées de façon optimale doit être anticipé par les algorithmes d'allocation de ressources et d'interrogation. Pour répondre à cette exigence, les recherches de l'équipe PYRAMIDE se focalisent sur la conception et le développement des nouveaux modèles d'allocation élastique de ressources pour l'optimisation dynamique de requêtes, tout en exploitant au maximum les résultats fondamentaux obtenus pour les BD parallèles et réparties, notamment les différents types de parallélisme (partitionné, dépendant (pipeline), indépendant), et la minimisation des coûts de communication [5]. Ces modèles visent à dépasser les approches classiques par l'introduction de la dimension « économique » dans la fonction objective, l'exploitation efficace des types de parallélisme, et la décentralisation du contrôle pour assurer le passage à l'échelle par l'intégration d'une politique de migration proactive à base d'agents mobiles.

Dans la deuxième perspective, les travaux menés à l'IRIT couvrent tout le cycle de la découverte de connaissances à partir de données. Les **méthodes** développées pour **l'analyse**, **la fouille** puis l'agrégation de données s'appuient sur les représentations de bas niveau obtenues après nettoyage, et exploitent différentes facettes et dimensions de ces données. Ces analyses produisent des « objets » intelligibles à forte valeur ajoutée, selon des méthodes définies en fonction des données mais aussi des interprétations à produire. Ces objets prennent la forme de connaissances formalisées, de structures (hiérarchies, graphes, etc.), de relations entre données, de résumés ou de synthèses en langage naturel. Parmi les représentations de haut niveau construites à partir de l'analyse du langage naturel, les structures de discours présentes dans les textes et dans les dialogues écrits du web font l'objet des travaux précurseurs de MELODI. Dans le projet STAC (ERC Grant), des algorithmes de ML exploitent des représentations structurées pour analyser des corpus de conversations au niveau discursif, et rendre compte des coopérations entre partenaires. L'originalité méthodologique est de coupler statistiques et modèles de la théorie des jeux afin d'extraire la structure des dialogues, de retrouver des dialogues similaires, ou de connaître l'opinion d'un des interlocuteurs. D'autres fouilles de textes visent à reconnaître automatiquement des formes d'humour et d'ironie en différentes langues [6]. L'enjeu est d'identifier les traits linguistiques les plus pertinents pour améliorer l'apprentissage à partir d'exemples annotés. Enfin, d'autres analyses de micro-blogs ont permis d'anticiper des sujets polémiques à partir de la mesure de l'évolution de la fréquence de mots clés.

Plus généralement, les données de réseaux sociaux (blogs et micro-blogs, forums, opinions, etc.) apportent un éclairage utile à l'exploitation d'autres types de données. Ainsi les équipes SIG et IRIS s'intéressent entre autres au filtrage en temps réel de flux de données de ces réseaux afin d'extraire et d'assembler (agréger) dans des objets plus riches [7], toutes les informations importantes sur un sujet ou un événement (TREC 2015). Ces approches peuvent exploiter des sources complémentaires : la contextualisation de textes courts consiste à agréger des données textuelles au sein d'un résumé permettant à un lecteur de comprendre ces textes trop succincts pour être intelligibles.

Pour répondre aux besoins de caractérisation fine et d'exploitation des données, l'IRIT dispose d'expertises reconnues en différentes **techniques d'analyse et agrégation de données** : apprentissage automatique (ML), Calcul Haute Performance (HPC) et systèmes émergents, ainsi qu'en IA (raisonnement automatique, fusion de données, modélisation de l'argumentation, etc.).

Les techniques de ML sont actuellement en plein essor. En effet, elles permettent non seulement d'identifier des régularités dans des données, de faire émerger des classes, mais aussi d'apprendre des modèles approximant le comportement de phénomènes à partir de données mesurées, et ainsi de réaliser des simulations. Les recherches menées à l'IRIT étendent les cadres habituels pour prendre en

La plateforme OSIRIM

L'**Observatoire des Systèmes d'Indexation et de Recherche d'Information Multimédia (OSIRIM)**, est une plateforme offrant une capacité de stockage de 1 Po et de calcul sur 640 cœurs. Celle-ci a vocation à héberger des projets scientifiques nécessitant la manipulation et le partage de plusieurs Teraoctets de données, telles que, par exemple, les tweets dans le but d'observer et de modéliser l'évolution d'un sujet au cours du temps. La plateforme offre des services et des outils dédiés au traitement de grands volumes de données comme le système Hadoop pour la distribution du calcul et des données.
OSIRIM est connectée à la plateforme GRID5000 spécialisée dans le calcul distribué afin d'offrir aux usagers de cette dernière de disposer d'un grand espace de stockage.

OSIRIM

✉ osirim@irit.fr http://osirim.irit.fr

compte l'incertitude et l'imprécision de la description des données et des modèles.
Elles portent également sur l'utilisation de l'apprentissage pour la visualisation de données. Plus récemment, au sein du labEx CIMI, elles visent à caractériser les propriétés théoriques de ces algorithmes et à les faire évoluer pour traiter et produire des données structurées (i.e. graphes syntaxiques ou structures de discours), ou encore apprendre à partir de données séquentielles et dynamiques (i.e. flux de données). Plusieurs équipes de l'IRIT exploitent ces algorithmes pour analyser le langage naturel oral et écrit, extraire des connaissances, rechercher des informations, analyser des images, des données numériques et des données structurées comme les graphes.

D'autres méthodes sont développées dans le cadre de recherches sur la vie artificielle (équipe VORTEX) et sur les systèmes complexes en environnement dynamique (équipe SMAC).
Des recherches visent à optimiser les performances d'algorithmes bio-inspirés tels que les réseaux de neurones, les algorithmes évolutionnaires ou les réseaux régulateurs génétiques artificiels, en intégrant des travaux de HPC. Ils sont utilisés pour produire des organismes virtuels, des créatures complexes capables de s'adapter à leur environnement en se corrigeant elles-mêmes et en s'auto-organisant, en réponse à des objectifs d'usage particuliers en environnement simulé.
Une approche alternative consiste à définir des modèles à l'aide d'un système multi-agent auto-adaptatif. L'équipe SMAC a développé un constructeur de modèles basé sur cette approche [8]. À partir d'une analyse des grandes masses de données issues de l'observation du système réel, le modèle découvre des corrélations simulant la dynamique du système. Ces travaux sont évalués en particulier dans le cadre de l'opération neOCcampus.

Enfin, l'analyse de données massives fait appel aux **techniques de HPC** et nécessite d'adapter les algorithmes à de nouvelles contraintes. Alors que le HPC optimise des traitements complexes mais identiques pour de très grands volumes de données brutes, les nouveaux jeux de données à analyser, plus hétérogènes, doivent être prétraités et évoluent dans le temps. Ils requièrent des traitements spécialisés qui induisent souvent la résolution de systèmes linéaires et non linéaires, très gourmands en temps de calcul. On est ainsi passé en dix ans de systèmes d'équations de l'ordre de quelques millions de variables à des systèmes à plusieurs milliards d'inconnues.
Lever ces verrous algorithmiques conditionne l'exploitation optimale des architectures massivement parallèles. Les recherches menées par l'équipe APO s'enrichissent de collaborations soutenues avec l'IMT via le labEx CIMI (cf encadré p. 6). Elles visent la réduction des coûts des traitements, par factorisation de matrices, par approximation de rang faible [9] ou en traitant par blocs des matrices creuses ainsi que le développement d'algorithmes pour améliorer le passage à l'échelle des méthodes de résolution sur architectures hybrides de très grande taille [10].
Des projets en commun avec le CERFACS (ADTAO, FILAOS) se focalisent sur la résolution de grands problèmes inverses (approximation de modèles à partir de données). Enfin, les équipes VORTEX et APO collaborent avec le laboratoire AMIS pour étudier de nouveaux algorithmes d'alignement, capables de traiter de grands volumes de données (20 Mo de bases chromosomiques) en réalisant des calculs complexes dans le domaine de bioinformatique pour l'anthropobiologie.

Lorsque la collecte et le filtrage de l'information à partir d'un besoin sont couplés à son analyse, on peut proposer des **méthodes d'exploration interactive et de visualisation**. Les recherches menées à l'IRIT s'intéressent entre autres à la modélisation de l'information pour l'analyse multidimensionnelle, aux interfaces de visualisation support aux analyses (cartes, graphes) et à l'adaptation des connaissances extraites aux utilisateurs. Certaines de ces approches sont intégrées à la plate-forme Tétralogie, support à l'enseignement, à l'analyse d'activités scientifiques et à l'intelligence économique.

Campagnes de test

L'IRIT participe depuis plus de dix ans à des campagnes d'évaluation de systèmes de recherche d'information (RI) telles que TREC (Text Retrieval Conference), INEX (XML Retreival) à CLEF (Cross Language Evaluation Forum), TrecVid (TREC Video Retrieval Evaluation) mais aussi OAEI (Ontology Alignment Evaluation Iinitiative).
Les tâches (tracks) proposées exploitent de très gros volumes de données (des centaines de To) de toutes natures : texte (tweets, news, forums, pages Web), vidéo, données sémantisées pour traiter de problématiques comme la sélection, le filtrage et l'agrégation en temps réel d'information autour d'un sujet, d'une entité ou d'un événement, ou encore la contextualisation de tweets ou l'alignement multilingue.
L'IRIT a obtenu des résultats significatifs : 3/18 à TREC2011 (Medical Track), 1/27 à TREC2012 (Contextual suggestion), 1/47 à CLEF2015 (Social Book Search) avec l'université de Neuchâtel et 2/11 à TREC2014 (Knowledge Base Acceleration), 1 INEX/CLEF 2013 (tweet Contextualisation).

✉ Mohand Boughanem (boughanem@irit.fr)

D'autres modalités d'interaction, plus riches, comme les techniques d'interaction spatiales (tangible, gestuelle, multi-surface) sont étudiées pour appréhender d'immenses volumes de données hétérogènes et structurées. L'équipe ELIPSE collabore avec la société Berger-Levrault pour évaluer ces techniques d'interaction dites « en entrée », afin de visualiser et manipuler les données des services publics [11].

Infrastructures et intergiciels pour la gestion des données

La valorisation de données d'entreprises s'appuie classiquement sur l'utilisation de plusieurs systèmes prenant en charge différentes parties du cycle d'analyse. Or le passage à de très grands volumes de données, la nécessité d'obtenir des réponses immédiates ou d'intégrer un grand nombre de sources pose de réels problèmes d'architecture informatique afin de réduire le coût d'accès aux données par les traitements autant que la durée des traitements.
Avec la plate-forme OSIRIM, l'IRIT dispose d'une infrastructure pour la gestion et le traitement de grandes masses de données (cf. encadré p. 8). De plus, des équipes étudient les systèmes de gestion efficace des infrastructures distribuées que sont les clouds, les centres de calcul et les centres de données. Parmi les problèmes abordés, celui de l'efficacité énergétique est l'objet du projet Green IT de l'équipe SEPIA. L'équipe développe des algorithmes de placement et d'ordonnancement de tâches et de données, de modélisation et d'estimation de la consommation d'énergie des systèmes et des applications. Elle propose l'utilisation de leviers verts au niveau du système d'exploitation et d'intergiciels pour économiser l'énergie tout en respectant les contraintes de performance des systèmes Cloud et HPC. Une autre problématique relative au stockage distribué est d'agréger des ressources et d'organiser les données de façon logique au sein de volumes virtuels en dépit de l'hétérogénéité des supports et formats physiques. L'équipe SEPIA a mis au point CloViS, un intergiciel de virtualisation du stockage qui prend en charge des bibliothèques de qualités de services (techniques de preuve d'intégrité, mécanismes de type map/reduce, etc.) enfichables à ajouter à la volée. On se rapproche ainsi de la notion de systèmes de gestion de fichiers programmable. CloViS fait actuellement l'objet d'un programme de valorisation.

En parallèle, la puissance et l'efficacité des calculateurs a augmenté rapidement grâce à l'introduction de nouvelles technologies dans les processeurs de type multi-cœurs et accélérateurs. Cela a entraîné le développement d'ordinateurs aux architectures complexes, avec une gestion très hiérarchisée des mémoires, un haut degré de parallélisme (plusieurs millions d'unités de calcul) et une très grande hétérogénéité. La recherche en calcul scientifique devient dans ce contexte le point de passage naturel, car situé à la croisée de différents champs de recherche comme l'algèbre linéaire numérique, l'optimisation, le traitement des graphes de grande taille et le calcul parallèle.

Conclusion

L'IRIT continuera d'être un acteur de la Science des données dans les années à venir en relevant plusieurs défis relatifs aux infrastructures, aux logiciels et modèles de gestion de données ainsi qu'aux algorithmes.
Les premiers concernent la définition puis l'exploitation de nouvelles représentations à partir des données : assurer que ces modèles et outils soit exploitables tout au long du cycle de vie de la donnée ; proposer des modèles fins de représentation pour mieux appréhender l'hétérogénéité des contenus ; rendre explicite et manipulable leur sémantique. Ces problématiques concernent la représentation conjointe d'informations visuelles et textuelles, la représentation hiérarchisée de documents textuels et leur apprentissage automatique. Ensuite, une de nos ambitions est d'arriver à traiter conjointement plusieurs «V» du Big Data au moment d'exploiter ces représentations, par exemple Vélocité et Volumétrie dans le cas d'analyse de microblogs en situation de crise, ou encore l'hétérogénéité et la valeur dans le cas de la maintenance préventive ou du diagnostic d'objets connectés. Une seconde ambition est de définir un cadre théorique pour prendre en compte conjointement les contextes de l'information et des utilisateurs ainsi que les interactions entre eux et avec le système. Enfin, la génération d'informations intelligibles passe par la capacité à raisonner sur des données inconsistantes et incertaines, à extraire des connaissances profondes des données, et donc de pousser plus loin l'association d'approches statistiques et symboliques. Cette génération exige aussi l'assemblage d'objets résultants de sources et analyses différentes, problème NP-difficile contraint par des exigences (comptabilité des objets, diversité, non redondance).

L'IRIT a donc fait le choix de mettre en avant ses recherches sur le passage de données à de l'information intelligible.
Pour cela, le laboratoire souhaite poursuivre une démarche d'ouverture et de collaboration avec les producteurs et utilisateurs de données au niveau régional, tant dans le milieu de la recherche qu'au niveau des entreprises. En participant également à la formation des étudiants, nous contribuons à la mise en place d'un réseau de compétences au service des différentes problématiques liées aux données. ■

Références bibliographiques

IRIT

[1] A. Bride, T. Van de Cruys, N. Asher, A Generalisation of Lexical Functions for Composition in Distributional Semantics. *ACL*, 281-291, 2015.

[2] A. Berro, I. Megdiche, O. Teste, Holistic Statistical Open Data Integration Based On Integer Linear Programming, *RCIS'15*, 468-479, IEEE 2015.

[3] M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier, Implementation of multi-dimensional databases in column-oriented NoSQL systems. *ADBIS'15*, 79-91, 2015.

[4] L. Soulier, C. Shah, L. Tamine. User-Driven System-Mediated Collaborative Information Retrieval. *ACM SIGIR 2014*, 485-494, 2014.

[5] S. Yin, A. Hameurlain, F. Morvan. Robust Query Optimization Methods with Respect to Estimation Errors: A Survey. *ACM SIGMOD Record*, Vol. 44 (3), 25-36, 2015.

[6] J. Karoui, F. Benamara, V. Moriceau, N. Aussenac-Gilles, L. Hadrich Belguith. Towards a Contextual Pragmatic Model to Detect Irony in Tweets, *ACL 2015*, 644-650, 2015.

[7] A. Kopliku, K. Pinel-Sauvagnat, M. Boughanem, Aggregated search: A new information retrieval paradigm, *ACM Computing Survey* 46(3) : 41:1-41:31, 2014.

[8] J. Boes, J. Nigon, N. Verstaevel, M.-P. Gleizes, F. Migeon. The Self-Adaptive Context Learning Pattern: Overview and Proposal. *CONTEXT 2015, Springer*, LNAI 9405, 91-104, 2015.

[9] P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, C. Weisbecker. Improving Multifrontal Methods by Means of Block Low-Rank Representations. *SIAM Journal on Scientific Computing*, 37 (3), 1451-1474, 2015.

[10] S. Gratton, M. Rincon-Camacho, E. Simon, P. Toint. Observation Thinning in Data Assimilation Computations. *EURO Journal on Computational Optimization*, 3 (1), 31-51, 2015.

[11] G. Perelman, M. Serrano, M. Raynal, C. Picard, M. Derras, E. Dubois. The Roly-Poly Mouse: Designing a Rolling Input Device Unifying 2D and 3D Interaction. *ACM-CHI '15*. 327-336, 2015.

Pour aller plus loin

S. Amer-Yahia, F. Bonchi, C. Castillo, E. Feuerstein, I Méndez-Díaz, P. Zabala, Complexity and algorithms for composite retrieval. *WWW* 79–80. 2013.

H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. 2014. Big data and its technical challenges. *Communication ACM 57 (7)*, 86-94, July 2014.

Actions fédératives : MASTODONS et MaDICS

En 2012, dans le cadre de la stratégie recherche fondamentale et interdisciplinaire sur les masses de données, le CNRS a lancé **le défi MASTODONS** dans l'objectif de produire des concepts et des solutions originaux qui n'auraient pu être obtenus sans coopération entre les différentes disciplines, et de favoriser l'émergence d'une communauté scientifique interdisciplinaire autour de la Science des données, en particulier scientifiques.

L'IRIT a participé au projet MASTODONS ARESOS sur l'analyse de grands réseaux socio-sémantiques afin de répondre à des questions du type « qui parle, de quoi, comment » grâce à la reconnaissance d'acteurs ou une analyse sociologique. Le rôle de l'IRIT ciblait la recherche d'information dans les microblogs (Identification de thématiques), la recommandation collaborative (par Crowd Indexing et tagging social) et l'étude de la production scientifique et de ses liens avec les réseaux sociaux.

Depuis 2015, **le GDR MaDICS** (Masses de Données, Informations et Connaissances en Sciences) du CNRS anime et soutient des activités de recherche autour des masses de données en sciences en fédérant une communauté interdisciplinaire. Il impulse une dynamique de rapprochement entre les producteurs et utilisateurs de masses de données et les experts en gestion et analyse de données. L'IRIT compte jouer un rôle important dans cette structure. Ainsi, lors de l'assemblée constitutive du 24-25 juin 2015, on comptait une quinzaine de membres de l'IRIT, initiateurs et/ou animateurs de 5 clusters-actions parmi les 12 proposés et discutés. Ces actions ont été présentées comme des incubateurs d'actions d'animation du GDR.

Pour en savoir plus :
http://www.cnrs.fr/mi/spip.php?article53
http://www.madics.fr/

ACOVAS (FUI, 2013-2016) implique les membres des équipes IRIS, MACAO et SIG de l'IRIT, et plusieurs industriels de l'aéronautique (Nexeya, Liebherr, Airbus, Zodiac, Gfi, Prometil). Le projet vise à développer une nouvelle version du banc d'essais de la société Nexeya.

Notre apport concerne la configuration du banc et l'exploitation des données de tests produites. Nous développons un module chargé de la configuration du banc ; cela consiste à intégrer au sein d'une base de données un ensemble de fichiers de configuration (csv). Les résultats obtenus permettent un gain de performance significatif : temps de chargement accéléré avec un facteur 80 fois plus rapide sur un nombre et des tailles de fichiers variables. Les données produites décrivent les mesures de différents capteurs. Dans le projet ACOVAS, l'exploitation de ces masses de données de tests porte sur la détection d'anomalies et d'explications : il s'agit de corréler automatiquement les alarmes (anomalies dans les mesures) avec les mesures explicatives

✉ Olivier Teste (Olivier.Teste@irit.fr)

SparkInData (FUI, 2015-2017) vise à fédérer des sources de données d'observation de la Terre pour favoriser l'émergence d'un écosystème riche de nouveaux services et usages.

Sur la base de flux continus d'images d'observation de la Terre fournis par le CNES (13 Terra Bytes/jour), différents partenaires vont construire une plateforme de service gérée par ATOS, offrant : un accès à ces données, mais également aux données de surveillance des sous-sols (de BRGM), des océans (de MERCATOR) et de cartographie (de l'IGN) ; ainsi qu'un environnement de développement (GEOMATYS) et un catalogue de services innovants pour le bénéfice des acteurs des marchés avals de l'agriculture (TERRANIS et l'IE de Purpan), de l'urbanisme (GEOSIGWEB), de la sécurité, et d'océanographie,....

Au sein de ce projet, les 4 équipes de l'IRIT (IRIS, MELODI, SC et TCI) vont proposer des solutions innovantes pour gérer l'interopérabilité des images d'observation et des méta-données associées. Au cœur de cette contribution, se trouve l'évaluation des technologies du web sémantique, des ontologies et des données liées, pour faciliter la gestion des descriptions des images, leur recherche ainsi que l'analyse de leur contenu par leurs producteurs et utilisateurs. L'IRIT contribuera ainsi à définir des briques de base et des entrepôts de données sémantiques qui favorisent un accès unifié aux données et services fournis par la plateforme SparkInData.

Nathalie Aussenac ✉ (nathalie.aussenac@irit.fr) http://sparkindata.com/

Genomic Breeding decision support (GBds) (FUI, 2012-15) a permis de créer des outils novateurs d'aide à la création de plants de maïs. Les équipes APO et SMAC de l'IRIT et UPETEC avaient pour tâche de concevoir des algorithmes de prédiction du phénotype d'un plant, capables de tenir compte des données génotypiques, agronomiques, climatiques et pédologiques fournies par les partenaires du projet (les semenciers RAGT 2n et EURALIS, et Météo France).

La volumétrie de ces données (60 000 marqueurs sur des centaines de milliers d'hybrides, milliers de parcelles, trentaine de relevés sur une dizaine d'années…), leur caractère hétérogène, le fait qu'elles soient bruitées voire souvent absentes, constituaient les obstacles à surmonter pour proposer de tels algorithmes.

Deux modèles de prédiction aux propriétés complémentaires ont été conçus :

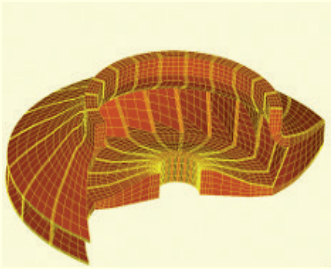
- L'équipe APO s'est concentrée sur un modèle bilinéaire pour représenter l'interaction génétique-environnement, mettant en œuvre une prédiction par une Ridge Regression, et présentant l'intérêt de nécessiter moins d'information a priori qu'avec la méthode classique.
- L'équipe SMAC et UPETEC ont défini le modèle de prédiction comme un problème d'optimisation combinatoire traité par un système multi-agent. L'originalité de cette méthode est de prendre en compte des données sans sémantique, et de limiter l'impact des données manquantes, erronées ou non significatives.

Un prototype opérationnel intégrant ces deux approches, permettant en outre de visualiser et naviguer dans la multitude des données génotypiques, a été développé par UPETEC et testé avec succès par les sélectionneurs de RAGT et EURALIS.

Carole Bernon ✉ (carole.bernon@irit.fr) www.irit.fr/GBds,1399

MUMPS (**MU**ltifrontal **M**assively **P**arallel **S**olver) est une librairie écrite en Fortran 95 et en C pour la résolution de systèmes linéaires creux de grandes tailles sur machines à grands nombres de processeurs. Elle est le fruit d'une trentaine d'années de recherches collaboratives entre l'IRIT, l'Inria, le CERFACS, le CNRS, l'Université de Bordeaux et l'ENS Lyon. La librairie MUMPS est mise à disposition de la communauté scientifique et industrielle sous licence CeCILL-C avec la possibilité de le télécharger librement sur le site internet : http://mumps-solver.org.

Les systèmes linéaires creux visés par cette librairie apparaissent dans de nombreux domaines de la simulation numérique (géophysique, mécanique des structures, économie, chimie, santé, ...).
La résolution de ces systèmes linéaires de grande taille représente une partie critique d'une simulation, très consommatrice en temps de calcul : au cours d'une simulation, chaque résolution est invoquée de très nombreuses fois, implique des Peta (1015) opérations et nécessite des Terabytes (1012) de mémoire de travail. Tout en calculant des solutions numériquement précises ou à précision contrôlée, l'enjeu est de calculer ces solutions efficacement en temps et en mémoire, en utilisant au mieux les ressources informatiques disponibles. C'est l'un des objectifs majeurs des travaux de recherche menés pour nourrir la plateforme logicielle et de recherche MUMPS.



MUMPS permet de traiter différents types de systèmes linéaires possédant plusieurs dizaines de millions d'inconnues, sur des architectures à plusieurs milliers de processeurs.
Sa reconnaissance internationale est attestée par des milliers de téléchargements par an, de la part d'industriels ou de chercheurs du monde entier, une utilisation en production dans de nombreux groupes industriels, et de nombreuses citations dans les publications scientifiques. Un Consortium d'utilisateurs (http://mumps-consortium.org), géré par l'INRIA, a été créé en 2014 pour tisser un lien privilégié entre les utilisateurs du logiciel intéressés par le soutien à son développement et les développeurs. À ce jour, six sociétés ont adhéré: EDF, Michelin, ALTAIR, LSTC, Siemens, ESI Group et Total.

THÉSES EN COURS ET SOUTENUES

Filtrage et agrégation d'informations vitales relatives à des entités

Rafik ABBES
Equ. : IRIS et MELODI (2012-2015)

Exploitation des réseaux sociaux en recherche d'information

Ismail BADACHE
Equ. : IRIS (2012-2016)

AMAS4BigData : Analyse dynamique de masses de données par système multi-agent adaptatif

Elhadi BELGHACHE
Equ. : SMAC (en cours)

Une approche multidimensionnelle pour la modélisation de la sémantique compositionnelle

Antoine BRIDE
Equ. : MELODI (en cours)

Unsupervised extraction of semantic relations using discourse information

Juliette CONRATH
Equ. : MELODI (2012-2015)

Optimization Methods for Large-scale Distributed Query Processing on Linked Data

Demirtas DAMLA
Equ. : PYRAMIDE (en cours)

Nouvelles méthodes de calcul et d'optimisation pour la comparaison génotype/phénotype: application à l'évolution des gènes de l'audition

Franklin DELEHELLE
Equ. : APO et VORTEX & labo AMIES (en cours)

MOBIDIK (nosql MOdeling of BIg Data, Information and Knowledge. Modélisation multidimensionnelle dans les bases de données NoSQL

Mohammed EL MALKI
Equ. : SIG (en cours)

Intégration holistique et entreposage automatique des données

Imen MEGDICHE-BOUSARSAR
Equ. : SIG et VORTEX (2012-2015)

Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multi-dimensionnelle en recherche d'information

Bilel MOULAHI
Equ. : IRIS (2012-2015)

Définition et évaluation de modèles de recherche d'information collaborative basés sur les compétences de domaine et les rôles des utilisateurs

Laure SOULIER
Equ. : IRIS (2011-2014)

Mokrane BOUZEGHOUB

Mokrane Bouzeghoub, spécialiste des bases de données et des systèmes d'information, est professeur à l'université de Versailles où il a dirigé successivement plusieurs équipes et projets sur le sujet.

Il est depuis 2010 directeur adjoint scientifique de l'INS2i et membre du comité de pilotage de la Mission à l'Interdisciplinarité du CNRS.
Il coordonne depuis 2012 le défi Mastodons sur les masses de données.

Noir sur Blanc : que pensez-vous des nouveaux enjeux de recherche ouverts par la « mode » Big Data ?

Mokrane Bouzeghoub : Loin d'être une mode, réservée à des entreprises branchées, le Big Data nous concerne tous, de façon durable. Derrière Big Data, on ne trouve pas seulement des techniques d'apprentissage machine mais l'ensemble des outils qui permettent non seulement de générer de la valeur à partir des données mais aussi d'en faciliter le stockage et l'accès, que les données soient de nature numérique ou symbolique.

Quant aux opportunités, elle sont nombreuses dans plusieurs domaines : meilleure connaissance des clients et de leur mode de consommation, amplification du marché par analyse et comparaison de produits, adaptation temps réel des prix, optimisation et amélioration de la mobilité urbaine, personnalisation de services, suivi d'épidémies, … Ces multiples opportunités justifient la concurrence qui fait rage entre les GAFAM[1] : mettre la main sur les plus grands corpus de données que deviendront des avantages économiques substantiels, acquérir la compétence pour exploiter ces gisements. La donnée, en particulier la donnée personnelle ou la trace numérique, est devenue une monnaie d'échange entre de multiples organismes. Le reportage récent d'Elise Lucet sur France 2 (Cash Investigation du 6 octobre 2015) illustre la bataille entre les acteurs économiques pour l'appropriation de ces données personnelles et la constitution de véritables trésors numériques.

Dans le domaine de la recherche, il faut se rappeler qu'il y a seulement quelques décennies, la constitution d'un corpus de données était quelquefois l'affaire d'une vie de chercheur (astronome, géologue, archéologue, naturaliste, …), et il fallait une seconde vie pour les analyser. Plusieurs générations de chercheurs se relayaient alors pour collecter, analyser et comprendre un phénomène. Aujourd'hui, l'acquisition de volumes de données, plusieurs milliers de fois plus importants, peut se faire en quelques heures via des satellites, des radars, des capteurs, des réseaux sociaux ou des modèles de simulation. Et leur analyse est réalisée à la même échelle de temps, contribuant ainsi à une compréhension accélérée des phénomènes, voire même des découvertes plus rapides.

NsB : y a-t-il une rupture avec les recherches faites jusqu'ici en informatique sur la gestion et l'analyse des données ?

M. B.: Je ne crois pas. C'est une évolution continue avec quelques effets d'accélération dus à des investissements spécifiques (en bioinformatique par ex.) ou à des projets emblématiques (comme LSST, LHC). Les informaticiens et les mathématiciens se sont intéressés depuis toujours au passage à l'échelle, au calcul intensif, au transport haut débit des données. D'autres thèmes de recherche se sont ajoutés comme la visualisation des grands volumes de données et l'analyse temps réel. Par contre, on peut parler de rupture de nature épistémologique dans d'autres disciplines comme les sciences humaines et sociales ou les sciences environnementales. L'accès presque instantané à des millions de documents, la possibilité de les filtrer thématiquement, la capacité d'en modéliser une partie des contenus, de les corréler pour en détecter les influences réciproques et d'en extraire les résumés des plus pertinents, offrent aux chercheurs en littérature, en histoire ou en sociologie de nouveaux horizons de prospection et d'analyse jusqu'à récemment insoupçonnés.

NsB : quelle est votre vision des équipes de pointe nationales et internationales sur ces questions ?

M. B. : Je suppose que vous n'entendez pas par là un palmarès des équipes travaillant sur le domaine? Non. Ce qui me parait important c'est que les équipes travaillant sur le Big Data ont mis en lumière un problème depuis longtemps mal compris par les décideurs: la nécessité de mettre en place des plateformes ambitieuses et de les soutenir fortement en ingénierie. Les recherches en Big Data nécessitent un effort particulier de préparation des données, de nettoyage, de configuration ou reconfiguration d'architectures de stockage ou de calcul pour ces données, d'interopérabilité entre plusieurs bases de données distribuées et hétérogènes, de réalisation de jeux de tests ambitieux, de développement et d'optimisation d'algorithmes passant l'échelle, d'offre de services de visualisation des résultats, et d'administration efficaces des centres de données. Sans infras-

[1] **G**oogle, **A**pple, **F**acebook, **A**mazon, **M**icrosoft (GAFAM)

M. B. (suite)

tructure et sans ingénieur de soutien, il n'y aura pas d'expérimentations significatives, et sans expérimentation les recherches en Big Data perdront une bonne partie de leur intérêt. C'est ce qui a motivé la campagne « ingénieurs plateformes » de l'INS2I cette année (dont a bénéficié l'IRIT) et même si elle reste modeste, elle exprime la volonté de contribuer à répondre à ce souci.

Noir sur Blanc : comment voyez-vous la place de l'IRIT dans ce panorama ?

M. B. : L'IRIT est un grand laboratoire qui couvre un spectre thématique très large et possède en son sein des compétences qui, si elles sont mobilisées autour d'un projet fédérateur, pourraient obtenir des résultats saillants dans le domaine des Big Data. Les grands thèmes du laboratoire, comme la recherche d'information, l'analyse et la synthèse d'information, l'interaction personne-machine et le calcul intensif constituent les briques fondamentales de la recherche en Big Data. Avec une originalité qui peut être mise en avant au vue des forces locales : l'intégration des Big Data et de l'IA pour apporter des solutions originales, basées sur la sémantique des données et le raisonnement, dans la décision et la recommandation. C'est une vue très extérieure au laboratoire, je vous demande pardon si elle est partielle ou erronée.

Manifestations passées

CLEF 2015

La 6e conférence CLEF (Conference and Labs of the Evaluation Forum), pour l'évaluation systématique des systèmes d'accès à l'information, par l'expérimentation sur des tâches communes, a rassemblé plus de 180 personnes en septembre 2015.

Elle était sous la présidence de Josiane Mothe, responsable de l'équipe SIG de l'IRIT, et de Jacques Savoy, professeur à l'Université de Neuchâtel.

CLEF 2015 a débuté par une conférence constituée d'articles évalués par un comité scientifique présidé par Karen Pinel-Sauvagnat (IRIT) et Jaap Kamps (Université des Pays Bas) portant sur un large éventail de problématiques dans les domaines de l'évaluation de l'accès à l'information multilingue et multimodale.

D'autre part, un ensemble de laboratoires et d'ateliers conçus pour tester les différents aspects des systèmes de recherche d'information mono et inter-langues ont permis de maintenir et d'élargir la tradition de l'évaluation CLEF et de traiter de façon détaillée ces questions d'évaluation. CLEF 2016 se déroulera début Septembre 2016 au Portugal.

www.clef-initiative.eu/

Big Data et défi industriel

Pour la 3e année, l'IMT, l'IRIT et la formation CMI SID de l'UPS, ont organisé deux journées sur les Big Data, avec le soutien des labEx CIMI et AMIES, des GDR MADICS et MascotNUM du CNRS. Des informaticiens et mathématiciens de renom ont présenté leurs recherches en assimilation de données (combinaison d'un modèle et de données observées) et en gestion de masses de données pour la modélisation biomathématique. Ils ont aussi exposé les défis systémiques liés à la « Science des données » dans des centres interdisciplinaires tels Saclay. Des ateliers ont abordé deux techniques : l'apprentissage automatique, dont l'apprentissage séquentiel facilitant l'analyse de flux de données ; le Calcul Haute Performance et les architectures multi-cœurs pour traiter des Hexabytes de données. Des exposés ont développé les défis posés à l'apprentissage automatique par l'analyse sémantique de très grands volumes de textes, la recommandation en temps réel ou l'analyse de flux d'images satellite. Les échanges entre académiques, entreprises, et étudiants ont souligné les enjeux scientifiques et économiques de ce secteur.

Manifestations à venir

Vous pouvez retrouver l'agenda complet sur :

www.irit.fr/-Agenda-

9 - 11 mars 2016

SDNR

Semaine du Document Numérique et de la Recherche d'Information

ESPE Toulouse

www.irit.fr/sdnri2016

11 avril 2016

Journée « La Science des données »

IRIT

www.irit.fr/Science_des_donnees

4 avril - 1er juillet 2016

CIPPMI

Trimestre thématique CIMI Current Issues in the Philosophy of Practice of Mathematics & Informatics

IMT et IRIT

www.cimi.univ-toulouse.fr/cippmi/

Les enseignants chercheurs de l'IRIT interviennent principalement au sein de formations en Informatique dans les universités et écoles toulousaines, tutelles de l'Institut. Ces établissements délivrent des diplômes Universitaires de Technologie (DUT, Bac+2), des diplômes de Licence (Bac+3), de Master ou/et des diplômes d'ingénieur (Bac+5).

À la rentrée 2016, une nouvelle offre de formation sera en place pour la période 2016-2021. Le terme « mention » d'un master est attribué à l'échelle du site toulousain et précise la thématique générale dans laquelle sont définis différents parcours la spécialisant.

Tous les masters auront à la fois une composante professionnelle et recherche (disparition de la distinction « pro » et « recherche »).

Les formations ci-dessous (dans leur version Septembre 2016) proposent des Unités d'Enseignement intégrant les concepts liés à la Science des données. Toutes les formations offrent la possibilité aux étudiants de faire un stage en entreprise ou en laboratoire durant le cycle de formation ; certaines peuvent être suivies en alternance, en formation continue voire à distance.

FORMATIONS

Diplôme Universitaire de Technologie

DUT « Informatique » / IUT Paul Sabatier et IUT Tlse 2 Blagnac – Université Toulouse Jean Jaurès (UT2J)

✉ Pr. M. Chevalier (Max.Chevalier@iut-tlse3.fr)
et Laurent Nonne (Laurent.Nonne@univ-tlse2.fr)

DUT « Réseaux et Télécoms » / IUT Tlse2 Blagnac -UT2J

✉ Pr. Thierry Villemur (villemur@laas.fr)

Diplômes de Licence

Licence 3 Pro en « Gestion et Traitement Informatique de Données massives » (GTID, ex AGBD) / IUT Paul Sabatier

✉ Pr. M. Boughanem (bougha@irit.fr)
et Pr. O. Marquié (olivier.marquie@iut-tlse3.fr)

Licence mention « Informatique / Univ Tlse 3 – Paul Sabatier (UT3)

✉ Pr O. Gasquet (olivier.gasquet@irit.fr)

Licence mention « Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales » (MIASHS) / Universités Tlse 1 Capitole (UT1), UT2J et UT3

✉ Pr. G. Zurfluh (gilles.zurfluh@irit.fr)
et Dr. D. Marquié (daniel.marquie@irit.fr)

Diplômes de Master ou d'Ingénieur

Master mention « Informatique » / UT3, INPT, ENAC

✉ Pr. D. Kouame et Pr. M. Paulin (kouame@irit.fr, Mathias.Paulin@irit.fr)

Master mention « Réseaux et Télécommunications » / UT3, INSA, INPT, ENAC, ISAE

✉ A. Aoun (André.Aoun@irit.fr)

Master mention « Mathématiques et Applications » / UT3, UT1, ENAC, ISAE

✉ Pr. F. Malgouyres (francois.malgouyres@math.univ-toulouse.fr)

Master mention « Bio-Informatique » / UT3

✉ Pr. G. Fichant (fichant@ibcg.biotoul.fr)
et J. Farinas (jerome.farinas@irit.fr)

Master mention « Méthodes Informatiques Appliquées à la Gestion des Entreprises » (MIAGE) / UT3 et UT1

✉ Pr. G. Zurfluh (gilles.zurfluh@irit.fr)
et Dr. D. Marquié (daniel.marquie@irit.fr)

Master Mention « « Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales » / UT2J : Master ICE - Master ISMAG

✉ Pr. B. Coulette (coulette@univ-tlse2.fr) et Pr C. Thierry (caroline.thierry@irit.fr)

Diplôme d'Ingénieur « Systèmes de Télécommunications et Réseaux Informatiques » (STRI) / UPSSI-TECH UT3

✉ Pr. A. Benzekri (benzekri@irit.fr)
et Dr. C. Galy (christine.galy@univ-tlse3.fr)

Diplôme d'Ingénieur en « Informatique et Mathématiques Appliquée » (IMA) / INP-ENSEEIHT

✉ Pr. J. Gergaud (joseph.gergaud@enseeiht.fr)

Diplôme d'Ingénieur en « Télécoms et Réseaux » (TR) / INP-ENSEEIHT

✉ Pr. M. Coulon (Martial.Coulon@enseeiht.fr)

Formation Continue

Domaine Statistiques, bioinformatique, Big Data. Stages « Intelligence économique : méthodes et outils de veille », « Programmation MapReduce sur un cluster Hadoop », « Système de stockage NoSQL (Not Only SQL) » / CNRS Formation Entreprise

✉ Pr J. Mothe (Josiane.Mothe@irit.fr)

DU « Sécurité et Sûreté des Systèmes d'information en Alternance » / UT3

✉ Pr. A. Benzekri (abdelmalek.benzekri@irit.fr)