cnrs le journal

n° 250
novembre 2010

JUSQU'OÙ IRA l'informatique ?

D'INTERNET À L'ORDINATEUR QUANTIQUE

dépasser les frontières

L'événement

Double Chooz : la traque des neutrinos est lancée

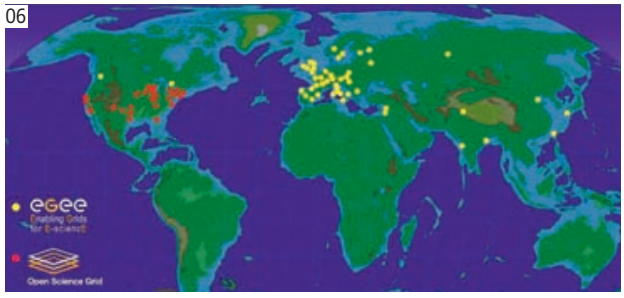
D'INTERNET À L'ORDINATEUR QUANTIQUE

Jusqu'où ira l'informatique ?

Des milliards d'informations à organiser

Un touriste à la recherche du voyage au meilleur prix. Un physicien face aux données recueillies par un accélérateur de particules. Une société d'intérim compulsant des CV afin de pourvoir une offre d'emploi. « *Tous ces exemples ont un point commun,* révèle Amedeo Napoli, du Laboratoire lorrain de recherche en informatique et ses applications[1], à Vandœuvre-lès-Nancy. *Ils renvoient à des situations où l'on fait face à un volume colossal de données parmi lesquelles on cherche à extraire une information.* » En principe, la méthode pour y parvenir est simplissime : préparer les données initiales, les confier à un algorithme de fouille et attendre que ce dernier se charge de présenter le résultat sous la forme souhaitée. Mais, dans un univers où le volume des données croît inexorablement, l'extraction de connaissances pertinentes relève de la gageure.

Illustration avec le cas de la recherche d'un séjour, comprenant vol, hôtel et location de voiture, au meilleur prix. Comme le détaille Michel Beaudouin-Lafon, du Laboratoire de recherche en informatique[2], à Orsay, « *mathématiquement, nous savons que la complexité de ce type de problème exclut qu'il puisse être résolu exactement en un temps raisonnable, dès lors que le nombre de données en entrée explose* ». Si bien qu'en pratique les programmateurs doivent ruser afin d'obtenir le résultat le moins mauvais en un temps raisonnable. Et c'est un fait, la fouille de données, à l'heure actuelle en plein essor, agrège des spécialistes de disciplines aussi différentes que l'informatique, bien sûr, mais aussi l'architecture des machines, la linguistique ou les mathématiques. Ces spécialistes empruntant aussi bien à l'intelligence artificielle, aux bases de données, aux techniques d'apprentissage et aux méthodes statistiques.



© CERN

06 Emplacements des sites impliqués dans les deux plus grandes infrastructures de grille aujourd'hui dans le monde : Egee en Europe (en jaune) et OSG aux États-Unis (en rouge).

OPTIMISER LE TRI DES DONNÉES

Une chose est certaine, plus aucun secteur n'échappe à la nécessité de développer des méthodes efficaces pour ne pas crouler sous une montagne de données inexploitables, voire impossibles à stocker. Prenons le projet ANR Midas, dont l'objectif est de réaliser un algorithme capable de résumer un important volume de données produites en temps réel, afin qu'elles puissent être stockées sur une mémoire centrale limitée pour consultation ultérieure. « *C'est typiquement le cas de figure rencontré par France Télécom, EDF ou la SNCF,* précise Pascal Poncelet, du Laboratoire d'informatique, de robotique et de microélectronique de Montpellier[3]. *Par exemple, une rame de TGV enregistre 250 informations par wagon toutes les cinq minutes afin d'anticiper des opérations de maintenance. Or il est impossible de conserver toutes ces informations. Il faut donc sélectionner les événements en fonction de leur intérêt, sachant que celui-ci évolue au cours du temps.* »

Autres gros consommateurs de techniques de fouille, les scientifiques eux-mêmes. Archétype du genre, le LHC, le collisionneur de particules géant du Cern, à Genève. Lorsqu'elle fonctionnera à plein régime, cette machine projettera des protons les uns contre les autres 40 millions de fois par seconde. Mais les physiciens estiment que seule une centaine de ces événements présenteront un intérêt et devront être enregistrés. Or ces

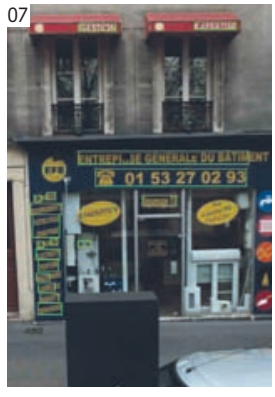
COMMENT FAIRE PARLER LES IMAGES

Désormais, nous possédons tous des milliers de photos. Les plus grosses banques d'images en recèlent des millions. Pour s'y retrouver, des outils existent. Tels ceux permettant à certains logiciels d'identifier un visage. Mais, comme le fait remarquer Matthieu Cord, du Laboratoire d'informatique de Paris-6, *« le taux de réussite est seulement compris entre 50 et 60 % ».* Typiquement, un algorithme spécialisé s'y retrouve très bien avec des informations dites de bas niveau : couleur, contraste, vecteurs de déplacement des pixels dans le cas d'une vidéo, etc. Plus délicate est leur transformation en informations de haut niveau qui rendent possible l'identification à coup sûr d'un objet ou d'un événement particulier. Ce qui n'empêche pas des applications de plus en plus performantes. Par exemple celle développée par l'équipe de Jenny Benois-Pineau, du Laboratoire bordelais de recherche en informatique[1], à Talence, en collaboration avec l'Inserm, dans le cadre du projet ANR Blanc Immed. Comme elle le précise, *« il s'agit de filmer des actions de patients atteints de la maladie d'Alzheimer chez eux et d'identifier des comportements associés à la maladie et qui sont utiles aux soignants pour suivre l'évolution des malades. »* De son côté, Matthieu Cord collabore au projet ANR iTowns, une carte numérique de Paris construite à partir de photographies, tel le service de Google Street View, à la précision du centimètre ! « *Nous développons des outils pour détecter automatiquement les personnes et les voitures afin de flouter les données personnelles,* détaille celui-ci. *Mais aussi une multitude d'objets plus ou moins enfouis dans ces images – les enseignes, les panneaux de signalisation, la végétation, les façades, etc. – pour faciliter des navigations avancées. »*

1. Unité CNRS/Université Bordeaux-I/IPB Enseirb-Matmeca Bordeaux/Université Victor-Segalen.

CONTACTS :
Jenny Benois-Pineau
› jenny.benois-pineau@labri.fr
Matthieu Cord
› matthieu.cord@lip6.fr

07



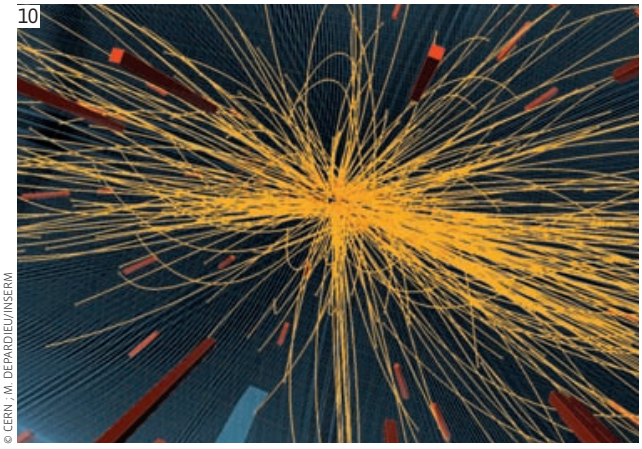
08

09

© IGN - STEREOPOLIS

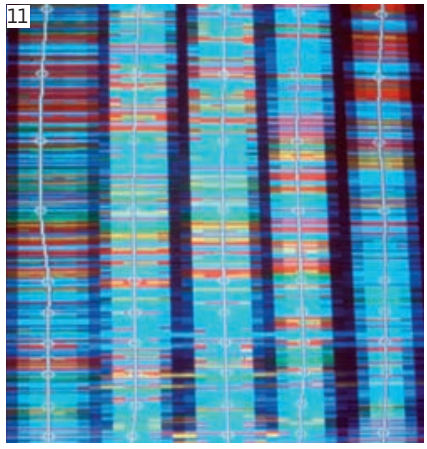
07 08 09 iTowns extrait automatiquement des informations présentes dans l'image.

10



© CERN - M. DEPARDIEU/INSERM

11



12

© C. LEBEDINSKY/INRIA

10 11 Certaines expériences, comme les collisions de particules ou le décryptage du génome, produisent d'importants volumes de données qu'il faut pouvoir trier et analyser.
12 L'étude des données scientifiques nécessite parfois de très gros moyens de calcul ainsi que la mise en réseau de machines, ici le projet Grid 5000.

DES RÉSEAUX POUR CALCULER

Les grilles informatiques sont des infrastructures virtuelles constituées d'un ensemble d'ordinateurs ou de grappes de PC géographiquement éloignés mais fonctionnant en réseau. Apparues voici quelques années sous l'impulsion de la physique des particules, elles permettent aux chercheurs et aux industriels d'accéder à moindre coût à d'importants moyens de calcul dans des domaines aussi variés que l'ingénierie, l'étude des maladies neurodégénératives ou la biochimie.
En France, l'Institut des grilles du CNRS, dirigé par Vincent Breton, fédère depuis trois ans l'activité dans ce domaine. Aux côtés de la Grid 5000, un outil spécifiquement dédié à la recherche dans le secteur des grilles, il met à la disposition des scientifiques et des industriels une grille de production rassemblant une vingtaine de milliers de processeurs disséminés dans une vingtaine de centres du CNRS, du CEA et d'universités.
Le 24 septembre dernier, ce dispositif déjà conséquent a franchi une étape supplémentaire avec la création par plusieurs organismes de recherche et universités[1] du GIS (Groupement d'intérêt scientifique) France Grilles, dont le but est de coordonner le déploiement d'une infrastructure de grille d'envergure nationale, puis de l'intégrer dans une grille européenne. Avec un objectif chiffré, annonce Vincent Breton, qui a été nommé à sa tête : *« Doubler les ressources et le nombre d'utilisateurs d'ici à 2015. »*

1. CEA, Conférence des présidents d'université (CPU), CNRS, Inra, Inria, Inserm, Renater et ministère de la Recherche.

CONTACT :
Vincent Breton
› vincent.breton@idgrilles.fr

derniers devront être sélectionnés en temps réel par des algorithmes spécialisés. « *Ce sont typiquement des algorithmes d'apprentissage, où l'ordinateur, au fur et à mesure qu'il est confronté à de nouvelles données à conserver ou à rejeter, accomplit sa tâche de mieux en mieux* », explique Michel Beaudouin-Lafon, dont l'unité collabore avec le Laboratoire de l'accélérateur linéaire[4] d'Orsay, sur la fouille de données d'accélérateurs.

UNE DÉMARCHE EMPIRIQUE

Mais les physiciens des particules ne sont pas les seuls à manipuler d'importantes quantités de données. Ainsi, l'équipe de Pascal Poncelet, en partenariat avec une équipe de l'Inserm, a développé un algorithme capable de caractériser les gènes impliqués dans différentes catégories de tumeurs du sein à partir de données de patients (informations génétiques, âge, poids, taille de la tumeur, traitement, devenir du malade…). « *Il offre aux cliniciens des informations sur les évolutions possibles d'une tumeur* », ajoute le chercheur. De même, l'équipe d'Amedeo Napoli, dans un projet en collaboration avec des astronomes, a mis au point des logiciels de fouille afin d'explorer des données sur des étoiles, dans le but de relever des caractéristiques ou des associations qui auraient pu échapper à un opérateur humain.

La fouille de données accomplit-elle pour autant des miracles ? Pas exactement. Car la discipline, qui a émergé à la fin des années 1980, est encore dans sa prime jeunesse. Conséquence, les chantiers sont légions. Pour Michel Beaudouin-Lafon, « *la plupart des démarches sont aujourd'hui empiriques. On ajuste des paramètres à la main et, lorsque cela fonctionne, on ne sait pas très bien pourquoi. Or, dans beaucoup de cas, il n'existe pas de critère quantitatif pour juger de la qualité d'informations extraites d'une base de données. Cela est laissé à l'appréciation des spécialistes du domaine* ». Et Amedeo Napoli de renchérir : « *Il y a encore beaucoup de travail à faire pour appréhender les très gros volumes. Actuellement, on peut gérer quelques milliers*

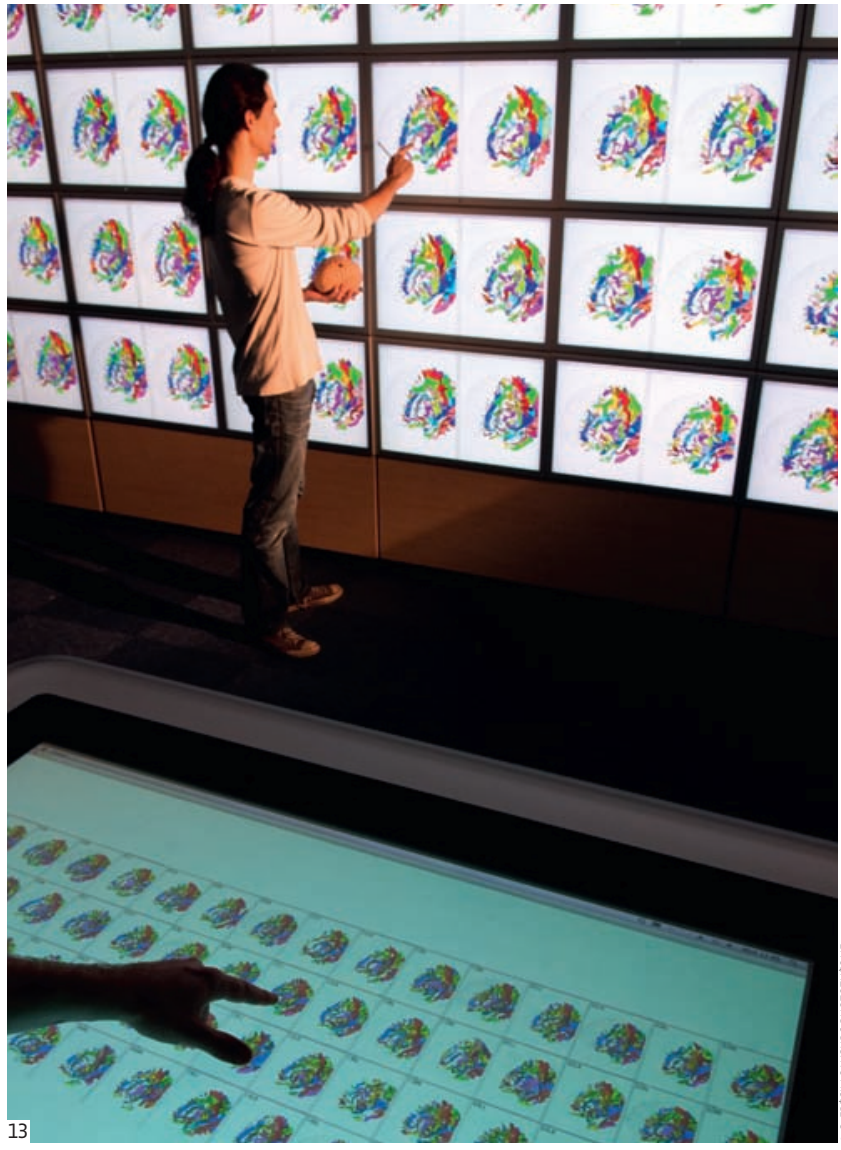
d'objets possédant quelques centaines d'attributs. Mais au-delà, on est confronté aux limites physiques des machines. »

Pour pallier cette difficulté, deux approches complémentaires sont possibles. Tout d'abord, là où une seule machine ne suffit pas, on peut faire travailler en parallèle plusieurs ordinateurs. C'est le principe de la grille *(lire l'encadré ci-contre),* poussé à l'extrême au LHC, qui dispose de 50 000 PC dispatchés dans différents centres de recherche à travers le monde, afin d'analyser l'équivalent des 3 millions de DVD de données dont les scientifiques disposeront au terme de l'expérience. Autre option, le supercalculateur, tel celui dont dispose depuis 2008 l'Institut du développement et des ressources en informatique scientifique (Idris) du CNRS, à Orsay[5]. Un monstre informatique capable de réaliser 207 milliers de milliards de calculs par seconde sur des nombres à virgule. *« Dans certains cas, typiquement la simulation d'armes nucléaires ou celle de la météo, il est difficile de morceler les données. Le superordinateur reste donc la solution »,* complète Michel Beaudouin-Lafon.

LA GESTION DU FACTEUR HUMAIN

Cependant, développer des ordinateurs ne suffit pas. De fait, à l'autre bout de la chaîne d'un processus de fouille se trouve un utilisateur humain. Se pose donc la question de la meilleure façon de lui présenter le résultat d'une recherche. Il suffit pour comprendre la problématique de penser à Google : le programme peut faire remonter plusieurs milliers d'adresses pour une requête, mais ne peut en afficher qu'une dizaine à l'écran. Comme le regrette Michel Beaudoin-Lafon, *« c'est dommage de bénéficier d'algorithmes sophistiqués pour faire remonter de l'information et de ne pas être capable de la présenter de façon correcte ».*

Pour ce faire, le Laboratoire de recherche en informatique a mis au point une plateforme d'un nouveau genre, baptisée Wild.



13 L'application Substance Grise utilisée sur la plateforme Wild sert à comparer simultanément les reconstructions 3D des cerveaux de 64 patients.

© C. FRÉSILLON/CNRS PHOTOTHÈQUE

Concrètement, un mur tapissé de 32 écrans d'ordinateurs représentant 130 millions de pixels et qui permet d'appréhender en un coup d'œil d'importantes quantités d'information. *« Nous travaillons avec huit laboratoires du plateau de Saclay sur ce projet »,* indique Michel Beaudouin-Lafon. En neurosciences, Wild permet d'afficher 64 IRM de cerveaux, *« ce qui présente un avantage indéniable lorsqu'il s'agit d'identifier une pathologie alors même que l'on observe une variabilité importante parmi les cerveaux sains »,* poursuit l'informaticien. De même, en astrophysique, certains observatoires fournissent désormais des images dont la taille excède largement celle d'un écran. Pour visualiser ces images en entier à leur résolution maximale, des outils tel que Wild font la différence. *« Je suis convaincu que ce type d'approche est amené à se développer, dans la recherche, mais aussi dans le monde industriel,* conclut Michel Beaudouin-Lafon. *Tout simplement parce que les données ne cessent d'augmenter, et les questions que l'on veut leur poser sont de plus en plus complexes et mal définies. »* Bref, il s'agit ni plus ni moins que d'éviter à la société de l'information de crouler sous son propre poids !

1. Unité CNRS/Université Henri-Poincaré/Université Nancy-II/Inria.
2. Unité CNRS/Université Paris-Sud-XI.
3. Unité CNRS/Université Montpellier-II.
4. Unité CNRS/Université Paris-Sud-XI.
5. Lire « Le CNRS s'offre un supercalculateur », *Le journal du CNRS,* n° 218, mars 2008, p. 34-35.

800 000 petaoctets,

c'est l'estimation du volume mondial de données numériques en 2009. Les experts s'attendent à une croissance de 45 % par an d'ici à 2020.

CONTACTS :
Michel Beaudouin-Lafon
› michel.beaudouin-lafon@lri.fr
Amedeo Napoli
› amedeo.napoli@loria.fr
Pascal Poncelet
› pascal.poncelet@lirmm.fr