3/4
Troisième volet
d'une série sur le thème :
« Les nouvelles frontières
de la science »
en partenariat avec le

COLLÈGE DE FRANCE 1530

…SERGE ABITEBOUL

CHERCHEUR EN INFORMATIQUE

BRUNO LÉVY POUR *ENJEUX LES ECHOS*

« Nous travaillons à des machines capables de raisonner et de décider à partir d'informations floues, voire erronées. »

PROPOS RECUEILLIS PAR
PASCALE-MARIE DESCHAMPS

Enjeux Les Echos — Votre leçon inaugurale portera sur la gestion des bases de connaissances à l'heure du Web. Pourquoi ce sujet a-t-il pris une telle importance ?

Serge Abiteboul — Une parabole recueille des signaux d'origine extraterrestre, ce sont des données ; si l'on y détecte un signal artificiel, cela devient de l'information ; et lorsqu'on parvient à décoder ce signal et à lui donner du sens, on obtient des connaissances. De tout temps, les sociétés humaines ont produit des données, des informations, des connaissances et elles les ont stockées et échangées. La nouveauté, c'est que ces données sont aujourd'hui le plus souvent numériques, stockées dans des ordinateurs et échangées via des réseaux informatiques. Surtout, ces informations – notamment dans le monde scientifique – sont désormais traitées automatiquement par des ordinateurs qui les filtrent, les transforment, les analysent pour les besoins d'expériences et de simulations. Pour donner un ordre de grandeur, la quantité d'informations stockée de manière numérique double à peu près tous les dix-huit mois. En 2012, il devrait circuler sur Internet un demi-zettaoctets. (1 zetta, c'est 10^{21} octets). Cela équivaut à environ 1 million de fois plus que la quantité d'informations correspondant à toutes les phrases qui ont été prononcées depuis l'apparition du langage.

A-t-on les outils pour gérer ces masses toujours plus grandes de données ?

S. A. — Leur gigantisme pose de nouvelles questions qui sont autant de défis scientifiques. Les progrès de la technologie des ordinateurs, des mémoires, des disques, l'augmentation des capacités, la baisse des prix, résolvent en partie les problèmes de stockage et de calculs. A titre indicatif, mon ordinateur personnel stocke environ un teraoctet (un tera c'est 10^{12} octets) et sa mémoire vive est de quelques gigaoctets (10^{9}). Il n'y a pas encore si longtemps, un dictionnaire « pesait lourd » sur un ordinateur. Aujourd'hui, on en stocke facilement plusieurs… Mais c'est quasiment inutile puisque, les machines étant en permanence connectées, toute l'information du Web est désormais à portée de doigts. La vraie difficulté n'est donc plus tant d'avoir accès à l'information que de trouver celle dont on a besoin, parfois une dizaine d'octets quelque part dans le cyberespace, peut-être juste un numéro de téléphone.

Pourtant, trouver un numéro de téléphone sur Internet ne semble pas être ce qu'il y a de plus compliqué ?

S. A. — Bien sûr. On a appris depuis le siècle dernier à gérer de gros volumes de données dans des systèmes centralisés et fermés. Ce sont, par exemple, les systèmes de gestion de bases de données d'Oracle et d'IBM, utilisés notamment pour gérer des répertoires de clients ou des catalogues de produits. Mais les volumes actuels dont on dispose sur le Web sont sans commune mesure. Et puis, à la différence des données qu'on trouve dans un système que l'on contrôle, on ne sait a priori pas grand-chose de celles que l'on trouve sur le Web. Par exemple, la BNF dispose de bases de données impressionnantes en taille, mais les archivistes savent exactement ce qu'elles contiennent. Si vous découvrez des données sur un site quelque part en Sibérie, vous ignorez tout de son contenu. Cela nous conduit à un nouveau champ de recherche baptisé le Web sémantique. L'idée est que lorsqu'on publie des données, on publie aussi des connaissances qui expliquent ces données. L'un des défis des années à venir est ainsi de rendre les machines capables de découvrir des sites riches en informations de qualité et de comprendre les données qu'ils hébergent.

Google ne sait pas déjà le faire ?

S. A. — Les outils que l'on trouve sur le Web ont déjà atteint de beaux niveaux de sophistication. Mais quand on recherche de l'information, on interroge le plus souvent du texte. L'épine dorsale du Web est en HTML, de l'hypertexte, mais du texte quand même. C'est compréhensible. Les gens aiment écrire dans leurs langues naturelles. Maintenant que fait Google, un des succès industriels les plus étonnants de l'histoire de l'humanité ? Il cherche des textes qui contiennent des mots qui vous intéressent et utilise un algorithme super-intelligent pour trouver les plus populaires et vous les proposer en tête de liste. De la recherche dans des textes ! Les résultats sont des textes ! Et quand vous posez la question : « Quelle est la capitale de la Tanzanie ? », vous obtenez comme réponse plusieurs documents qui ont été choisis parce qu'ils contenaient les mots « Tanzanie » et « capitale » et qu'il va vous falloir parcourir pour trouver votre réponse.

Pourquoi les machines ne savent-elles pas répondre plus précisément aux questions ?

S. A. — Le problème est qu'il faut passer des textes en langues naturelles à des bases de connaissances. Une base de connaissances, c'est une collection de faits élémentaires comme « Dodoma est la capitale de la Tanzanie ». Ces connaissances sont formulées dans un vocabulaire et avec une syntaxe contrôlés. Les ordinateurs sont plus à l'aise avec des connaissances qu'avec du texte en langue naturelle. Mais pour que cela fonctionne, il faut construire ces bases de connaissances. Or si les gens écrivent volontiers, ils ont plus de réticences à s'installer derrière un écran pour rentrer des connaissances au kilomètre, les vérifier. Nous travaillons donc à des outils qui iront extraire ces éléments des textes. Par exemple, Fabian Suchanek, un postdoctorant de mon équipe, quand il était en thèse, a participé au développement d'un logiciel qui extrait des connaissances de Wikipédia : 20 millions de faits élémentaires ! Et il y a bien d'autres projets en cours.

La difficulté est de passer de textes en différentes langues à des bases de connaissances du type « Dodoma est la capitale de la Tanzanie. »

Ces machines sauraient donc comprendre et parler les langues naturelles ?

S. A. — C'est la grande difficulté. On travaille depuis des dizaines d'années sur la compréhension des langues naturelles. Les progrès sont très lents. Certaines tournures de phrases sont complexes et mettent en échec les logiciels. Le même mot peut avoir de nombreux sens. Un texte peut être ambigu. Sans parler des informations dont le sens est quasiment politique, ou pour le moins sensible. A la question : « Quelle est la capitale de la France ? », nous pouvons facilement tomber d'accord sur la réponse. Mais si je demande quelle est la capitale d'Israël ou celle de la Macédoine… Enfin, les internautes publient des erreurs, volontairement ou pas. Il faut donc repérer les contradictions, trouver la réponse la plus probable.

HAL 9000, l'ordinateur de *2001, l'odyssée de l'espace* appartient encore à la science-fiction ?

S. A. — Nous n'en sommes pas là en effet. Nous essayons déjà de concevoir des machines capables de raisonner sur les connaissances qu'on leur a apportées et, à

BIOGRAPHIE

Sur son blog, Serge Abiteboul se présente comme « chercheur en informatique de profession, un peu écrivain et sculpteur amateur ». Il est aussi directeur de recherche à l'Institut national de recherche en informatique et automatique (Inria), membre de l'Académie des Sciences depuis 2008 et titulaire pour 2011-12 de la chaire Informatique et sciences numériques du Collège de France. Il y donnera sa leçon inaugurale le 8 mars 2012.

partir de ces raisonnements, de faire des propositions et de prendre des décisions. Mais les données dont la machine dispose ne sont pas forcément fiables. Or les mécanismes de raisonnement automatique partent du principe que les informations sont vraies car la logique mathématique a beaucoup de difficultés à raisonner avec des faits faux et des contradictions. Les humains, eux, savent gérer ambiguïtés, contradictions et approximations. Sur un réseau social, par exemple, si des « amis » disent que le petit copain d'Alice s'appelle Bob, d'autres que c'est Julien, ils accepteront que certains se trompent ou qu'Alice a peut-être deux petits copains. Mais une machine ? Nous cherchons donc des procédures leur permettant de raisonner à partir d'informations imprécises, voire contradictoires : ce peut être des systèmes de recommandation, d'expertise, etc. L'un des grands défis sera de pouvoir gérer l'imprécision et l'incohérence.

Pourquoi les systèmes de recommandation et d'expertise sur le Web ne sont-ils pas plus fiables ?

S. A. – Le sont-ils si peu ? Quand Wikipédia est apparu, par exemple, on a entendu de nombreuses réactions négatives : quoi, une encyclopédie faite par des gens qu'on ne connaît pas, qui produisent des informations non vérifiables, on ne peut pas s'y fier. En effet, il y a des erreurs dans Wikipédia. Elles ont été mesurées. Il n'y en a pas plus que dans une encyclopédie traditionnelle réalisée par des experts. Et en plus Wikipédia donne accès, gratuitement, à des champs de connaissances beaucoup plus larges que n'importe quelle encyclopédie classique.

Ces nouvelles bases de données auront-elles une incidence sur la manière dont on classe l'information ?

S. A. – Cela bouscule tout. Je fais partie d'un groupe de travail de l'Association for Computing Machinery (ACM) qui a pour mission de proposer une nouvelle taxonomie des articles scientifiques, autrement dit une hiérarchie de mots clés qui permettront de classer les articles scientifiques en informatique. Cet exercice est totalement futile ! Cela correspond bien à l'organisation des livres dans les rayonnages d'une bibliothèque. En témoigne la rédaction des notices papier : 3e étage, salle 145, tiroir B5, telle cote. De même, les rangements sur les « bureaux » des ordinateurs correspondent à des segmentations physiques « travail », « perso »… Cela n'a plus vraiment lieu d'être. On ne va plus chercher des livres dans les rayonnages, d'ailleurs, on ne va plus dans les bibliothèques. On tape quelques mots clefs qui permettent de retrouver des articles ou des ouvrages qu'on télécharge et feuillette à l'écran. La classification hiérarchique de domaines scientifiques est obsolète. Elle persiste par un certain mimétisme qui pousse à maintenir les habitudes jusqu'à ce que quelqu'un arrive et propose autre chose. Il en va de même de l'organisation des lettres sur les claviers. Elle remonte à l'époque où il fallait ménager les pièces mécaniques des machines à écrire. Utiliser la même organisation pour l'écran tactile d'un smartphone, c'est surréaliste. Mais il faut faire avec tous ceux qui ont appris à taper avec les anciens claviers et ne veulent pas changer.

La France et Paris ont de nombreux atouts et une excellente école informatique. Mais notre ressource rare ce sont les doctorants.

La France a-t-elle les ressources pour participer à cette prochaine aventure ?

S. A. – La France et Paris disposent de nombreux atouts, dont une brillante école de mathématiques et, bien que beaucoup plus jeune, d'une excellente école informatique. Mais notre ressource rare ce sont les doctorants. La recherche est mondialisée, la masse des doctorants est internationale, les gros volumes de neurones se trouvant en Chine et en Inde. Or ces cerveaux préfèrent aller aux Etats-Unis, ce qui explique leur domination scientifique. Sergueï Brin, le cofondateur de Google, venait de Russie. Il a atterri aux Etats-Unis, pas en France. Même si nous attirons de plus en plus d'étudiants en informatique, nous pouvons faire mieux en leur offrant davantage de bourses et en réservant un meilleur accueil à ceux qui ne parlent pas français. Les masters devraient être en anglais, sinon les étudiants ne viennent pas. Et c'est la meilleure manière de promouvoir le français car ils apprennent la langue en vivant en France, et parfois ils restent !

A quelles prochaines ruptures technologiques peut-on s'attendre ?

S. A. – On entre là dans le domaine des prédictions. Comme le disait le physicien Niels Bohr : « La prédiction est difficile, surtout lorsqu'il s'agit de l'avenir. » Mais puisque vous me le demandez, j'en ferai deux. La première : les problèmes qui demeurent sont les plus difficiles ; on va les attaquer, mais cela va prendre du temps. Par conséquent, la recherche va ralentir. Et la seconde : quelque part dans un laboratoire universitaire ou une start-up, quelqu'un a trouvé un nouveau Graal, une nouvelle application qui va donner un bon coup de pied dans la fourmilière et changer brutalement le paysage. Autrement dit, je ne peux pas vous dire à quoi il faut s'attendre. ■

BRUNO LEVY POUR *ENJEUX LES ECHOS*