Informatiques et sciences numériques

M. Serge ABITEBOUL, membre de l'Institut
(Académie des sciences), directeur de recherche
à l'Institut de recherche en informatique et en automatique,
professeur invité sur la chaire annuelle pour l'année 2011-2012

SCIENCES DES DONNÉES : DE LA LOGIQUE DU PREMIER ORDRE À LA TOILE

Serge Abiteboul est directeur de recherche à INRIA, Saclay, et membre du laboratoire LSV de l'ENS Cachan.

Leçon inaugurale

L'informatique a révolutionné nos vies. Si la vision classique des ordinateurs est celle de machines à calculer, ils servent à présent surtout à gérer des données. Le cours abordera des aspects fondamentaux de la gestion de données, y compris de ses liens profonds avec la logique mathématique et la théorie de la complexité. Le Web, que l'on peut voir comme une gigantesque base de données distribuées, sera aussi étudié avec ses facettes les plus passionnantes telles que son échelle ou les défis du calcul distribué et du Web sémantique.

L'information produite, stockée, traitée, échangée, est au cœur de l'activité des êtres vivants, des objets du monde, des associations humaines. Les systèmes informatiques nous aident à conserver cette information sous forme numérique, telle une sauvegarde quasi illimitée de notre mémoire personnelle. Ils nous aident à traiter et à échanger cette information pour communiquer entre nous. L'ordre de grandeur de l'information stockée atteint le zettaoctet : 10^{21} octets ! Le trafic d'information annuel sur Internet dépasse même cette quantité d'information accumulée. Face à ces chiffres vertigineux, deux problèmes s'imposent : où trouver la bonne information dans cette masse ? Comment déterminer ce que l'on veut conserver ?

Avec les efforts combinés d'une recherche académique dynamique, de pionniers marquants comme IBM, de jeunes géants comme Google et de *start-up* hyper créatives, les sciences des données se sont épanouies. Pourtant, le domaine tient encore de la forêt vierge quand nous atteignons la gestion de données distribuées et la Toile. Il est compliqué d'en dresser l'état de l'art ; il n'est pas simple de l'enseigner ; il n'est pas évident de prévoir quelles tendances sont là pour durer. C'est cette jungle que nous chercherons à pénétrer.

Les systèmes relationnels de gestion de bases de données sont des systèmes informatiques complexes, résultats de dizaines d'années de recherche et de développement. Ils sont parmi les plus grands succès logiciels du siècle dernier avec des produits commerciaux très répandus comme les serveurs Oracle et des systèmes gratuits très utilisés comme MySQL. Ils résultent de la combinaison de connaissances mathématiques solides (comme la logique du premier ordre), d'algorithmes très sophistiqués, et d'un *engineering* complexe.

Nous retrouvons ces trois mêmes ingrédients à la base des moteurs de recherche de la Toile. La Toile, le *World Wide Web* en anglais, s'appuie sur des documents hypermédia. Un moteur de recherche permet de fuir la navigation fastidieuse sur le graphe des pages et le monde de l'hypertexte pour plonger dans une bibliothèque numérique universelle. Si la Toile n'a sûrement pas de réponse à toutes les questions de l'internaute, la réponse à une question précise se trouve peut-être dans les masses d'informations véritablement extraordinaires disponibles. Tels des enfants, nous nous émerveillons devant les dizaines de milliards de documents de la Toile. Mais un enfant apprend, depuis son plus jeune âge, à évaluer, classer, filtrer, le volume considérable d'informations qu'il rencontre. Et nous ? Si le moteur de recherche ne nous aidait pas à nous focaliser sur un petit nombre de pages, que ferions-nous ? L'exploit technique, c'est de retrouver en un instant, grâce à un index, les pages de la Toile qui hébergent les quelques mots d'une requête. La magie, expliquée par quelques équations et des algorithmes, c'est de pouvoir retrouver, parmi les dizaines, voire centaines de millions de pages qui contiennent les mots demandés, quelques pages qui vont satisfaire l'internaute.

L'écriture nous a permis de matérialiser et d'externaliser en partie notre mémoire. L'imprimerie nous a permis de transmettre largement cette mémoire externe. On a beaucoup insisté sur le fait que la Toile diminuait considérablement les coûts de transmission de la mémoire. Nous sommes en train de découvrir que sa véritable révolution est de permettre à chacun d'apporter sa contribution personnelle au patrimoine collectif (avec des réserves comme la fracture numérique). La Toile est ainsi une juxtaposition de milliards d'individus et de tous leurs réseaux. Après les réseaux de machines, les réseaux de contenus, nous atteignons les réseaux d'utilisateurs. Des systèmes de la Toile, tels Facebook, permettent aux internautes de communiquer entre eux. Ce ne serait pas vraiment nouveau si ces nouveaux outils de communication ne conduisaient à d'autres modes de pensées, d'autres formes de relations. Surtout, phénomène véritablement passionnant, ces systèmes font émerger automatiquement, depuis les profondeurs des réseaux, des connaissances collectives. Plusieurs types d'approches permettent de construire de telles connaissances : la notation, par exemple, de produits ou d'entreprises par des internautes comme dans eBay ; l'évaluation de l'expertise des internautes comme dans « *Mechanical Turk* » ; la recommandation par exemple de produits comme dans Netflix ; la collaboration entre internautes pour réaliser collectivement une tâche qui les dépasse individuellement comme dans Wikipédia ; enfin, le *crowdsourcing* met des humains au service de systèmes informatiques, comme avec Foldit. L'émergence automatique de telles connaissances soulève toute une gamme de questions, tant philosophiques que scientifiques.

En observant les évolutions de la Toile et des sciences des données, nous pouvons imaginer ce que pourra être la Toile de demain, une Toile des connaissances, avec des millions, voire des milliards de machines interconnectées raisonnant collectivement. La fascinante Toile des documents d'aujourd'hui est fondée sur le

plaisir des gens à écrire, lire, dire, écouter du texte dans leurs langues naturelles. Les machines préfèrent échanger des connaissances plus formatées, plus rigoureuses. Avec le passage de la Toile du texte à une Toile des connaissances, elles pourront prendre plus pleinement en main la gestion de nos informations. Cela paraît une étape indispensable pour que l'humanité puisse survivre dans les flots d'information chaque jour plus cataclysmique qu'elle génère.

La Toile est multiforme et il est devenu quasi impossible de vivre sans elle. Elle est à la fois la plus belle des dentelles, trame de toutes les connaissances humaines, et terreau des plus horribles fantasmes, de toutes les violences. Il n'est pas possible, ni souhaitable, d'y renoncer, comme il n'a pas été possible de refuser l'écriture ou l'imprimerie. Et, malgré tous les écueils, nous voulons continuer à croire que la Toile contribuera à féconder un meilleur futur. Quant aux aspects plus techniques, nous nous hasarderons à affirmer que la prochaine étape en sciences des données a déjà commencé : c'est la construction de la Toile des connaissances. Des données à l'information, aux connaissances, le cheminement est logique.

Le texte intégral de la leçon inaugurale est disponible en ligne : http://lecons-cdf.revues.org/506. Il est également édité sous forme de livre imprimé dans la collection « Leçons inaugurales du Collège de France » aux Éditions Fayard.

Des versions audio et vidéo sont disponibles sur le site Internet du Collège de France : http://www.college-de-france.fr/site/serge-abiteboul/#|m=inaugural-lecture|q=/site/serge-abiteboul/inaugural-lecture-2011-2012.html|p=../serge-abiteboul/inaugural-lecture-2012-03-08-18h00.html|

Cours

Modèle relationnel

Nous discutons en premier lieu le modèle relationnel à la base des systèmes de gestion de bases de données. Ce modèle simplifie considérablement la gestion de données en servant de médiateur entre humains et machines.

Au-delà du modèle relationnel

Nous nous intéressons ensuite à des modèles qui cherchent à aller plus loin ou à faire mieux que le modèle relationnel. Il est intéressant de noter que l'informatique a révolutionné nos vies. Si la vision classique des ordinateurs est celle de machines à calculer, ils servent à présent surtout à gérer des données. Ce cours a abordé des aspects fondamentaux de la gestion de données, y compris de ses liens profonds avec la logique mathématique et la théorie de la complexité. Le Web, que l'on peut voir comme une gigantesque base de données distribuées, a aussi été étudié du point de vue de ses facettes les plus passionnantes, telles que son échelle ou les défis du calcul distribué et du Web sémantique. Les successeurs les plus célébrés du modèle relationnel se fondent sur des arbres et des graphes, comme ses prédécesseurs principaux.

Le Web sémantique

Le but du Web sémantique est de faciliter l'accès à l'information et aux connaissances. Il s'agit d'améliorer la précision des résultats de recherche et de

faciliter l'intégration de sources distinctes. L'idée est de publier des connaissances compréhensibles par des machines plutôt que du texte plus adapté à des humains.

Documents actifs et AXML

Nous nous sommes intéressés à la collaboration pour gérer des données entre des serveurs autonomes et hétérogènes. Pour cela, nous avons utilisé des arbres qui incluent des appels à des fonctions (des services du Web). Ces fonctions capturent la notion de vues (des données intentionnelles situées ailleurs) et permettent de spécifier le calcul distribué.

Moteur de recherche de la Toile

Nous avons expliqué comment, à partir d'une requête avec quelques mots clés, le moteur de recherche retrouve les pages qui semblent les plus pertinentes en utilisant principalement une indexation de la Toile et un algorithme de classement des pages par popularité.

Datalog : la renaissance

Datalog, proposé dans les années 1970, introduit la récursivité dans la partie positive des requêtes relationnelles. Nous avons décrit les traits principaux du langage et discuté de sa renaissance ces dernières années.

Gestion de données distribuées

L'utilisation de la distribution permet d'améliorer les performances des systèmes de gestion de données. La distribution se retrouve aussi dans de nombreuses applications quand les données sont naturellement distribuées entre plusieurs systèmes. Avec le Web, la gestion de données distribuées a pris une importance considérable.

Datalog distribué et Webdamlog

Nous avons parlé du langage WebdamLog et du système du même nom, dans la continuité de nos travaux récents sur un Datalog distribué.

Séminaires

Ouverture des données publiques (François Bancilhon, Data Publica)

L'ouverture des données publiques (*open data*) consiste à rendre disponible les données collectées, gérées et utilisées par la puissance publique pour accès et réutilisation par les citoyens et organisations (publiques ou privées). Dans la plupart des démocraties, de plus en plus de données sont rendues publiques par ce mouvement, lancé aux États-Unis en 2009 grâce à l'initiative « data.gov ». Ce flot de nouvelles données présente à la fois une opportunité majeure et des défis technologiques importants. L'opportunité est celle des nouvelles applications et des nouveaux usages qui peuvent être fait de ces données, et de la nouvelle compréhension

que les citoyens qui y accèdent en ont. Les défis sont multiples : ces données sont souvent pauvrement structurées et formatées, elles sont parfois de qualité médiocre, et enfin, elles sont fragmentées sous la forme de milliers ou de millions de fichiers contenant des informations complémentaires ou dupliquées. Pour utiliser ces données fragmentées, peu structurées et de qualités variables, plusieurs approches sont possibles. On peut laisser les données telles quelles et déplacer l'intelligence dans l'application qui les utilise, souvent à travers un moteur de recherche. On peut utiliser une approche de type web sémantique en convertissant les données en RDF et en établissant des liens entre des entités identifiées. Enfin, on peut les structurer sous forme de bases de données, certaines d'entre elles étant alignées sur des attributs communs (par exemple espace et temps).

Archivage du Web (Julien Masanès, Internet Memory)

Le Web représente la plus grande source d'information ouverte jamais produite dans l'histoire. Dépassant de plusieurs ordres de grandeur la sphère de l'imprimé, il offre également des caractéristiques inédites par rapport aux média qui l'ont précédé, telle l'édition collective à laquelle participe, même marginalement, une fraction importante de l'humanité, la dynamique temporelle complexe et le caractère paradoxal des traces créées, à la fois omniprésentes et fragiles. Ces caractéristiques uniques sont aussi celles qui en font une source d'information, d'analyse et d'étude majeure, dont la conservation est un enjeu important pour l'avenir. Elles obligent cependant à refonder les méthodes et les pratiques séculaires de préservation des artefacts culturels. Nous avons analysé dans ce séminaire les propriétés du Web vu depuis cet angle particulier de sa préservation et présenté quelques réflexions sur la manière dont sa mémoire peut être construite pour servir la science à l'avenir.

Raisonnement dans le Web sémantique (Marie-Christine Rousset, Université de Grenoble)

Prendre en compte la sémantique des données extraites du Web est fondamental pour construire des applications fiables intégrant ces données. Associer une sémantique formelle aux données multi-sources et multiformes du Web est un défi mais aussi une clé pour résoudre de manière robuste et générique, par des techniques de raisonnement automatique, des problèmes difficiles comme l'interopérabilité entre ressources distribuées et hétérogènes ainsi que la vérification de propriétés de sécurité ou de qualité de service spécifiées formellement. La prise en compte de la sémantique est également primordiale dans la recherche d'informations et l'évaluation de requêtes sur le Web. De nombreux travaux émanant de la communauté du Web sémantique ont été réalisés pour décrire la sémantique d'applications par la construction d'ontologies. Cependant, les problèmes de raisonnement sur les ontologies sont trop souvent déconnectés des données dont celles-ci décrivent la sémantique. Seuls quelques travaux considèrent les ontologies comme des interfaces de requêtes entre des utilisateurs ou des applications et des données. Ces travaux montrent par des arguments de complexité les limites qu'il faut imposer au pouvoir d'expression des ontologies pour pouvoir espérer obtenir des algorithmes d'évaluation de requêtes répondent dans des temps raisonnables sur de gros volumes de données. Dans ce séminaire, nous avons montré que les logiques de description sont un bon modèle pour décrire la sémantique des données du Web mais nous que

des restrictions sont nécessaires pour obtenir des algorithmes de raisonnement qui permettent, sur de gros volumes de données, de détecter des incohérences ou des corrélations logiques entre données ou sources de données, et de calculer l'ensemble des réponses à des requêtes conjonctives. Nous montrons que la famille DL-Lite a de bonnes propriétés pour combiner raisonnement et gestion de données à grande échelle, qui en font un bon candidat comme modèle de données du Web sémantique.

Les requêtes de bases de données. Logique et complexité (Moshé Y. Vardi, Rice University)

Cette présentation a donné un aperçu de la façon dont la logique mathématique fournit des bases pour l'une des technologies les plus importantes d'aujourd'hui, et a montré comment la théorie des requêtes logiques offre un éclairage précieux sur la complexité de calcul de l'évaluation de requêtes relationnelles.

Gestion des données scientifiques : ce n'est pas votre transaction quotidienne (Anastasia Ailamaki, École polytechnique fédérale de Lausanne)

Ce séminaire a discuté des défis de la gestion de données scientifiques, en l'illustrant, par exemple, avec la simulation du cerveau.

Analyse statique et vérification (Victor Vianu, University of California, San Diego)

Ce séminaire a examiné les défis et les limites intrinsèques de l'analyse statique et de la vérification. Il a identifié des situations où ce type d'approche peut être très efficace.

Le crowdsourcing de données (Tova Milo, Tel Aviv University)

Le *crowdsourcing* de données consiste à utiliser une foule d'utilisateurs du Web pour contribuer collectivement des données ou donner des opinions. Nous avons considéré la logique, les fondements algorithmiques et méthodologiques à l'œuvre pour la gestion de *crowdsourcing* de données à grande échelle.

Extraction de données à partir du Web (Georg Gottlob, Oxforf University)

Ce séminaire a traité du problème de l'extraction de données à partir du Web de façon semi-automatique et entièrement automatique.

La récolte de connaissances à partir du Web (Gerhard Weikum, Max Plank Institute, Saarbrucken)

Ce séminaire a traité des progrès récents, des possibilités de recherche et des défis dans le domaine de la récolte des connaissances sur le Web et de ses applications.

Les réseaux sociaux (Pierre Senellart, Telecom ParisTech)

Les réseaux sociaux sur le Web sont un moyen très populaire pour se connecter avec des amis, publier du contenu et partager des informations. Le séminaire a

notamment examiné les questions suivantes : 1) Comment faire pour indexer et interroger les données des réseaux sociaux ? 2) Comment expliquer la structure « petit monde » de ces réseaux ? 3) Comment utiliser des connexions sociales pour améliorer la qualité de la recherche sur le Web ou des recommandations ?

Références

Abiteboul S., *Sciences des données : de la Logique du premier ordre à la Toile*, Collège de France/Fayard, 2012.

Abiteboul S., *Sciences des données : de la Logique du premier ordre à la Toile*, Collège de France, 2012, http://lecons-cdf.revues.org/506 [doi : 10.4000/lecons-cdf.529].

Abiteboul S., Amsterdamer Y., Deutch D., Milo T. et Senellart P., « Finding optimal probabilistic generators for xml collections », *ICDT*, 2012, 127-139.

Abiteboul S., Bourhis P. et Vianu V., « Highly expressive query languages for unordered data trees », *ICDT*, 2012, 46-60.

Abiteboul S., Hubert Chan T.-H., Kharlamov E., Nutt W. et Senellart P., « Capturing continuous data and answering aggregate queries in probabilistic xml », *ACM Trans. Database Syst*., 36(4), 2011, 25.

Abiteboul S., Gottlob G. et Manna M., « Distributed xml design », *J. Comput. Syst. Sci*., 77(6), 2011, 936-964.

Abiteboul S., Manolescu I., Rigaux P., Rousset M.-C. et Senellart P., *Web Data Management*, Cambridge University Press, 2011 (sous pesse) [http://webdam.inria.fr/Jorge].

Abiteboul S., Senellart P. et Vianu V. « The erc webdam on foundations of web data management », *WWW Companion Volume*, 2012, 211-214.

Suchanek F.M., Abiteboul S. et Senellart P., « Probabilistic alignment of relations, instances and schema », *PVLDB* (Paris), 5(3), 2011, 157-168.

Suchanek F.M., Gross-Amblard D. et Abiteboul S., « Watermarking for ontologies », *International Semantic Web Conference* (1), 2011, 697-713.

Activités scientifiques principales

Membre de l'Académie des sciences et de l'Academia Europae.

Responsable du projet Webdam sur les fondements de la gestion de données distribuées de l'European Research Council (2009-2013).

Co-auteur d'un livre sur la gestion de données sur le Web.

Co-organisateur d'un workshop à Dagstuhl : *Foundations of distributed data management,* avec Alin Deutsch (University of California, San Diego), Thomas Schwentick (TU Dortmund, Allemagne), Luc Segoufin (ENS-Cachan).

Membre du bureau du Conseil académique consultatif de l'IDEX d'université Paris-Saclay.

Conférences invitées : *International Conference on Data Engineering*, 2012 ; *Annual Conference on Computer Science Logic 2012* ; *International Conference on Description Logic 2012*.