Vers une nouvelle science des risques ?

Serge Abiteboul

Vers une nouvelle science des risques ?

Serge Abiteboul
INRIA, ENS Cachan, Conseil national du numérique & Académie des sciences

Introduction

Nous assistons à une véritable explosion de la quantité de données et information disponibles. Enfouies au cœur de cette masse, des connaissances qu'on peut aller chercher. Du fait de la taille des données, cela exige le développement des sciences et des techniques autour du traitement de données. La possibilité de disposer et d'analyser d'énormes volumes de données, a évidemment d'immenses implications sur la société en général et sur l'assurance en particulier. Par exemple, l'analyse de données massives pourrait révolutionner la médecine en facilitant par exemple des traitements plus personnalisés. Pour l'assurance, cela devrait permettre une bien meilleure évaluation des risques.

Dans cet article, nous considérons l'explosion des données et l'analyse de ces données, le « Big data » pour utiliser un terme à la mode. Nous discutons des utilisations qui pourraient être faites de ces technologies notamment dans le cadre de l'évaluation des risques.

Données, information, connaissances

Les données tiennent depuis toujours une place essentielle dans le développement de l'informatique. Depuis les années 60, les logiciels de bases de données se sont imposés pour permettre le partage des données à l'intérieur d'une entreprise ou d'une organisation. Ces données qui étaient isolées dans des centres de calcul sont devenues accessibles partout dans le monde avec l'arrivée d'Internet, le réseau des *réseaux de machines*. Et puis est arrivé, le Web, le *réseau de contenus*. On a vite réalisé qu'il rendait possible le rêve de la connaissance et la culture accessibles par tous (« à des détails près » comme l'e-exclusion ou la propriété privé de certains contenus). L'étape suivante a été le développement des réseaux sociaux, des *réseaux d'individus,* basés sur le partage d'informations personnelles, la communication, la création de communauté ; l'internaute passant de simple consommateur à producteur d'information.

Nous baignons aujourd'hui dans un monde numérique. Par exemple, nous sommes entourés de milliards d'objets communicants, le web en 2008 comptait déjà plus de 1000 milliards de pages et chaque mois, les internautes réalisaient des dizaines de milliards de recherches Web. Surtout, on pense que le monde numérique double tous les 18 mois et le trafic sur Internet est déjà chaque année supérieur à tout ce que nous pourrions stocker en utilisant tous les supports, tous les disques disponibles.

Ces données et informations disponibles sur le réseau sont d'énormes gisements de connaissances à découvrir, à *valoriser*. L'analyse de données a été un domaine très actif quasiment depuis les débuts de l'informatique sous divers noms comme fouille de données ou business intelligence. Du fait de l'accroissement des capacités des disques et des mémoires, et des puissances de calcul avec des clusters jusqu'à des milliers de machines, du fait aussi de l'explosion des données disponibles, l'analyse de données pour en extraire de la valeur est redevenue à la mode, sous le nom de « Big data ».

Big data en bref

Le point de départ du Big data (en français, « Grosses données » clairement moins glamour) est de valoriser les gisements massifs de données. Quand on parle de Big data, on inclut parfois (mais pas toujours) :

(i) L'idée de croiser des données structurées, par exemple celles d'une entreprise avec des masses d'information moins structurée typiquement « sales » disponibles sur le web.

(ii) Des calculs massivement parallèles avec des technologies comme Hadoop issues des moteurs de recherche du Web. (Voir annexe.)

Comme le but est de découvrir à l'intérieur des données de nouvelles connaissances, les tâches sont les tâches classiques de l'analyse de données :

- Acquisition de données : charger les données notamment avec des outils d'ETL ;
- Intégration : combiner les données de plusieurs sources, les transformer dans un schéma unique, aligner leurs concepts ;
- Nettoyage : éliminer les réplicas, résoudre les contradictions, compléter les données… éventuellement en interagissant avec des humains (crowd-sourcing) ;
- Interrogation, surveillance, visualisation des données ;
- Analyses statistiques des données par exemple pour découvrir des corrélations ;
- Développement d'applications, de nouveaux services à partir des résultats.

Les difficultés sont nombreuses. Elles tiennent au volume des données (typiquement en téraoctets ou pétaoctets – 10^{15} ou 10^{18} octets), à leur hétérogénéité (structures différentes, multi-linguisme, etc.), à leur « vélocité » (c'est-à-dire de leur taux de changement), à leur distribution dans l'espace, aux protections éventuelles (droit d'accès, restriction sur leur usage), et aux variations dans leur qualité (erreurs, incomplétude, confiance, provenance, fraicheur, etc.). Evidemment, la nature des traitements sur ces données a une importance considérable. Un algorithme en n^3 sur un milliard d'enregistrements reste hors de portée même avec des centaines de machines.

Même s'ils s'améliorent constamment, les logiciels comme Hadoop, mentionné précédemment, sont encore relativement jeunes et compliqués à utiliser en particulier parce qu'ils demandent de faire travailler ensemble un grand nombre de machines. Quand on est confronté à une application impliquant de gros volumes de données, il faut donc s'interroger :

1. Les données sont-elles *vraiment* « Big » ? Ne suffirait-il pas d'utiliser une seule machine avec beaucoup de mémoire RAM et un stockage SSD massif.
2. Pourrait-on réduire la dimension du problème, par exemple en échantillonnant les données ?
3. Ne pourrait-on pas utiliser un algorithme plus intelligent qui éviterait une exploration systématique de l'espace des solutions ?

Quand c'est possible, on préfèrera utiliser des techniques plus classiques, notamment non parallèles, même si c'est moins passionnant que de jouer avec Hadoop et de gros clusters de machines. Evidemment, certains problèmes très parallèles sur de gros volumes de données rendent indispensable d'avoir recours à des technologies « Big data ».

Big data : mythe et réalité

On entend beaucoup le mythe : « le Big data va résoudre les problèmes de l'humanité. » En analysant de gros volumes de données, nous pourrons faire des prédications de plus en plus fines, prédire les maladies, les accidents, le climat, soigner le cancer, la pauvreté, etc. Pourquoi pas rendre le monde déterministe ? Evidemment, non. Cela reste des analyses statistiques qui calculent des probabilités et rien de plus. Reste bien sûr la part du hasard. Et puis il faut accepter les limites dues à la complexité générée par la taille des données. Il y a des calculs que l'on ne saurait pas réaliser même si on disposait de toutes les informations possibles, même si on disposait de millions de machines et de millions d'années. (En tout cas, dans le cadre de nos connaissances actuelles.)

Loin de ces beaux rêves, ce qui est surtout observable dans le Big data pour l'instant, ce sont de grosses sociétés utilisant des données privées dans des buts commerciaux, principalement pour des publicités ciblées. Plus il y a de données, plus il y a d'argent à gagner. Avec le système PRISM de la NSA et les systèmes équivalents en France, la presse s'est récemment fait l'écho d'une autre classe d'utilisation de ces technologies : L'espionnage. L'analyse des données peut servir positivement pour se protéger du terrorisme et beaucoup moins pour de l'espionnage industriel ou pour de la surveillance des citoyens. A contrario, la généralisation dans les démocraties de l'open data devrait permettre aux « data journalists » et plus généralement aux citoyens concernés, de contrôler les actions de leurs gouvernants ainsi que celles des grandes entreprises.

On pourrait aussi utiliser les technologies Big data pour résoudre d'autres problèmes. Elles sont particulièrement adaptées pour prévoir et répondre à : des crises sanitaires, des problèmes d'environnement, des catastrophes naturelles. Et généralement elles devraient permettre des avancées majeures pour résoudre les problèmes, par exemple, de santé, transport, écologie, pauvreté. Dans nombre de ces problèmes d'analyse de données, il faut combiner des analyses de gros volumes de données réalisées « off-line » et des analyses sur des données obtenues en flux en temps réel réalisées on-line ? De telles combinaisons se retrouvent ainsi, par exemple, dans le suivi personnalisé de personnes en grande difficulté, de personnes très âgées, d'élèves en échec scolaire, etc.

La santé et l'assurance, une vision de non-spécialiste

Les données en rapport avec la santé d'un individu croissent sans cesse. Evidemment le cœur en est constitué par des informations comme les examens médicaux, les diagnostics, les soins en hôpital ou non, et les prises de médicaments. Les données génomiques sont de plus simples à obtenir. (Un individu peut aujourd'hui obtenir par 23andMe pour 99$ le séquencement d'une partie importante de son génome.) A coté, des matériels comme des téléphones intelligents et de systèmes comme les réseaux sociaux produisent de plus en plus de données sur la vie quotidienne de l'individu, son alimentation, ses efforts physiques, son exposition à des pollutions particulières, etc.

L'analyse de données est de plus en plus utilisée par les chercheurs dans le domaine médical. A partir d'analyses de données, les chercheurs peuvent par exemple découvrir des corrélations entre la prise de certaines combinaisons de médicaments et des pathologies particulières. Les avantages pour les personnes sont aussi extrêmement prometteurs. Il s'agit d'abord de personnaliser les soins en adaptant les médicaments à chacun, en contrôlant les quantités prises. Et la prévention des maladies pourrait être améliorée en proposant à chacun une hygiène de vie adaptée à ses risques.

Pour les assurances, le Big data devrait permettre évidemment une meilleure évaluation des risques. Evidemment, nous sommes dans l'analyse de données, typiquement un outil scientifique qui ne se substitut pas à la science. L'analyse de gros volumes pourrait peut-être être utile en météorologie pour mieux prédire le climat mais seulement comme outil des climatologues. C'est une évidence. Pour l'évaluation des risques, la situation est similaire.

Une évaluation des risques plus sophistiquée serait un outil fantastique ne serait-ce que pour permettre une meilleure gestion des assurances ou mieux conseiller les assurés. Mais, comme souvent, la science arrive accompagnée de dérives possibles. Par exemple, ces mêmes données personnelles pourraient être utilisées pour personnaliser les polices d'assurance ; et ce type d'utilisation est déjà observé notamment aux Etats-Unis. Les personnes présentant des risques particuliers se voient proposer des polices d'assurances de prix plus élevés ou ces risques écartés de la couverture. A la limite, des scénarios pessimistes peuvent être imaginés. Une partie de la population seraient écartés des assurances médicales pour cause de risques de santé trop importants, en contradiction avec l'idée de mutualisation des risques. Tous se verraient imposer des contraintes sur leur vie quotidienne sous la menace de voir le prix de leur assurance exploser.

Liée à ce problème est celui de l'accès aux données personnelles. Un assureur doit-il avoir accès à tout ou partie d'information comme les données médicales d'un client, ses données fiscales, génomiques, les achats de ce client, sa géolocalisation, ses courriels, ses données dans les réseaux sociaux ? Ce sont les données personnelles du client. A ce titre, elles devraient lui appartenir et il devrait être seul à pouvoir décider qui y a accès et comment elles sont utilisées. Evidemment, ce n'est pas aussi simple. Par exemple, pour les progrès de la médecine, il est clair que les chercheurs doivent pouvoir faire des analyses sur les données médicales de tous. Il semble raisonnable que les résultats de ces statistiques soient publics. Mais il semble tout aussi clair que les données brutes, par exemple d'un hôpital, ne puissent pas elles devenir de l'open data même anonymisées (car il serait alors impossible de garantir la confidentialité.) Il reste beaucoup à faire pour concilier ces différents aspects.

Références

1. S. Abiteboul : *Sciences des données : de la Logique du premier ordre à la Toile*, Collège de France, 2012 : http://www.college-de-france.fr/site/serge-abiteboul/
2. S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset et P. Senellart : *Web Data Management*, Cambridge University Press, 2011 : http://webdam.inria.fr/Jorge
3. D. Agrawal et al. *Big data and cloud computing: current state and future oportunities*, EDBT/ICDT 2011
4. C. Lynch, *Big data: How do your data grow*? Nature 455, 28-29, 2008

Annexe: Le Big data et Hadoop

La technique map-reduce a été conçue par Google pour son moteur de recherche. Dans un calcul « map-reduce », on commence par découper le problème en de nombreux sous problèmes (map) que l'on confie à des machines distinctes. Ces machines résolvent les sous problèmes et envoient leurs résultats à d'autres machines qui ont pour tâche de

combiner les résultats (reduce). Le but est de pouvoir travailler sur d'énormes volumes de données en parallélisant les calculs sur des clusters de machines. Cette technique est aussi à la base des centres de données d'autres géants du Web comme Amazon, Facebook, et on la retrouve de plus en plus dans des offres de « cloud computing ».

Map-reduce ne permet de résoudre que des problèmes très parallèles. Le logiciel de map-reduce le plus populaire est Hadoop, un logiciel libre de la fondation Apache.