Le BigData et la publicité en temps réel

Nicolas GRISLAIN
Co-fondateur de la société AlephD[1]

Lorsque l'utilisateur d'un site Internet affiche une nouvelle page, sur laquelle sont réservés des espaces publicitaires, en quelques millisecondes des algorithmes déterminent quelle publicité sera affichée, et quel prix sera payé pour cela par l'annonceur. Depuis quelques années déjà, il existe des algorithmes qui choisissent pour l'annonceur le contenu précis à afficher, en fonction des caractéristiques de l'utilisateur. Désormais les espaces publicitaires sont attribués à l'issue d'une enchère, mettant en concurrence les annonceurs. D'autres algorithmes déterminent, soit pour l'annonceur soit pour l'éditeur du site, le montant à enchérir. Pour construire une règle de décision optimale, constamment évolutive, il leur faut mémoriser et analyser d'énormes quantités de données issues d'enchères précédentes. Les techniques du BigData sont indispensables. Les données traitées sont pour partie des données personnelles, mais elles ne sont utilisées que par des machines, dans le cadre d'une problématique précise qui ne permet pas d'avoir une idée globale des individus.

Sur fond d'automatisation du marketing digital, la société AlephD s'est développée en proposant aux éditeurs (sites web, applications mobiles) des outils permettant d'augmenter leurs revenus en optimisant la mise en vente de leurs espaces publicitaires. Les stratégies d'optimisation reposent sur l'exploitation de grands volumes de données : *BigData*. Ces approches puissantes permettent d'adapter son action à chaque utilisateur. En tant qu'acteur d'un secteur fortement consommateur de données personnelles, AlephD a un point de vue original sur le traitement par le marché de ces données.

Émergence des enchères publicitaires en temps réel

Le développement d'internet et la nécessité de financer la création de contenus ont, dès le milieu des années 90, contribué à la forte croissance du marché de la publicité en ligne et notamment à l'utilisation de bannières publicitaires (*display ads*).
Les bannières publicitaires, initialement intégrées de manière statique dans les pages web, l'ont rapidement été par l'intermédiaire d'*ad-servers*, permettant de programmer la diffusion de campagnes différentes selon l'utilisateur, l'heure de la journée ou tout autre critère. Les ad-servers permettent d'adapter la vente des bannières à chaque « impression », c'est-à-dire à

1. La société AlephD offre des services de traitement de données aux éditeurs de sites Internet

chaque visite d'un espace publicitaire par un utilisateur à un instant donné.
Néanmoins, la mise en relation entre le site web ou éditeur, c'est-à-dire l'entité vendant les impressions, et les annonceurs, qui achètent les espaces, reste en grande partie manuelle. En raison notamment du coût relativement élevé de l'allocation manuelle des *impressions*, une part importante de ces *impressions* reste invendue. Dans un premier temps, se sont mis en place des montages complexes de cascades entre intermédiaires, les uns passant la main aux autres s'ils ne peuvent livrer de bannière.
En 2008 pour automatiser la négociation des ventes d'*impressions* et éviter les cascades de passation de relais entre intermédiaires qui ralentissent le processus de vente, l'*ad-exchange* (bourse d'échange d'impressions publicitaires) est créée.

Fonctionnement d'un ad-exchange

Sur un *ad-exchange*, chaque impression publicitaire est mise en vente aux enchères pendant le chargement du contenu d'une page web, c'est-à-dire en quelques millisecondes : on parle de *Real-Time Bidding* (RTB). Ce processus se déroule en 3 temps :

- Un utilisateur charge la page d'un site site.com. La balise (*tag*) HTML définissant l'espace publicitaire de la page chargée émet une requête (*ad-call*) à l'*ad-exchange*.
- L'*ad-exchange* reconnaît éventuellement l'utilisateur et l'identifie par un *user id* ; il envoie un appel à enchérir (*bid-request*) à tous les acheteurs potentiels. Cette requête contient l'identifiant du placement, l'identifiant de l'utilisateur ainsi que d'autres informations associées au placement, à l'utilisateur, à l'enchère ou au protocole (adresse IP, user-agent, *referer*).
- Les acheteurs (*bidders*) répondent en fournissant leur évaluation de l'impression (*bid*). L'*ad-exchange* attribue l'espace publicitaire au plus offrant. Ce dernier paye généralement le plus haut bid en dessous du sien (second *bid*) et délivre le contenu de la publicité à l'utilisateur.

Ce processus se déroule en près de 50 ms et aboutit à l'affichage de la bannière du gagnant sur *site.com*. Les contraintes de temps réel font que les stratégies des acheteurs sont mises en œuvre par des algorithmes qui évaluent l'opportunité d'afficher une bannière à chaque impression.
L'enchère simultanée au second prix (enchère de Vickrey) est le mécanisme d'attribution dominant du marché ; il est utilisé sur la quasi-totalité des plates-formes. Dans ce cadre, le prix de réserve[2] que fixe le vendeur a un impact sur le prix de vente. Il appartient au vendeur de fixer ce *prix de réserve*, ainsi que d'autres paramètres de l'enchère, comme le niveau d'information diffusé aux acheteurs.

Analyse micro-économique de la relation acheteur vendeur sur les plates-formes d'enchères

De fait, face aux algorithmes parfois sophistiqués des acheteurs, les éditeurs réalisaient jusqu'à récemment le paramétrage de leurs enchères de manière statique. Dans ce contexte, l'avènement des *ad-exchanges* et du RTB a conduit les acheteurs à développer des stratégies d'enchère élaborées les positionnant comme *faiseurs de prix*. Au contraire, les vendeurs (éditeurs) qui n'ont pas développé les capacités technologiques pour analyser chaque enchère individuellement ni pour agir sur celles-ci se retrouvent dans la situation inconfortable de *preneurs de prix*.

2. C'est-à-dire le prix minimum demandé par l'éditeur du site

AlephD et le *big data* au service des éditeurs

Sur la base de ce constat, la société AlephD (www.alephd.com), créée fin 2012, a construit des outils d'analyse et de prise de décision en temps réel pour permettre aux éditeurs d'optimiser leurs revenus publicitaires enchère par enchère, c'est-à-dire des milliers de fois par seconde en moins de 10 ms.
Pour réaliser cette tâche, AlephD enregistre des rapports d'enchère des milliers de fois par seconde. Ces rapports contiennent a minima un identifiant d'utilisateur chargeant la publicité, un identifiant de placement publicitaire ainsi que des données de prix : premier prix et prix payé. Ces informations sont d'une part traitées en flux (*online processing*), et d'autre part stockées pour des traitements en gros (*batch processing*).
Elles représentent des volumes importants (plusieurs téraoctets de données chaque mois) ; elles arrivent avec un débit élevé (*high velocity*) et sous des formes relativement variées. Les données analysées par AlephD correspondent donc assez bien à la définition donnée par *Gartner* du big data comme coïncidence de 3Vs : *volume, velocity, variety*. En particulier, le volume de données et leur débit ne permettent pas de traiter ces données sur une seule unité de traitement, même en considérant des ordinateurs très haut de gamme.

Traitement en gros des données

Pour construire une règle de décision optimale, AlephD estime un modèle d'apprentissage automatique (*machine learning*) dont l'espace des paramètres est de très grande dimension afin de prendre en compte les relations complexes entre variables. Ce modèle de décision donne un paramétrage optimal de chaque enchère sur la base d'agrégats historiques (ensemble de grandeurs caractéristiques d'une entité) construits à la volée.
Compte tenu du volume des données, ce calcul est réparti sur différents serveurs qui réalisent chacun une part de l'estimation. Afin de gérer la complexité liée à la distribution du calcul et à l'agrégation des résultats donnés par chaque serveur, le paradigme de calcul *map-reduce* est utilisé pour exprimer le processus d'estimation.
L'approche *map-reduce* est une manière de formaliser une opération sur un grand nombre d'observations qui consiste à appliquer à chaque observation une transformation (*map*) puis à agréger les résultats des étapes *map* par une fonction d'agrégation (*reduce*). L'intérêt de cette approche est qu'un calcul exprimé dans ce formalisme est assuré de pouvoir être distribué sur un grand nombre de serveurs et donc de passer à l'échelle. Cette garantie permet à AlephD de pouvoir traiter un volume de données 10 fois plus important en multipliant le nombre de serveurs effectuant l'estimation par 10.

Traitement en ligne

Des milliers de fois par seconde, AlephD reçoit un compte-rendu d'enchère qu'elle utilise pour dresser des agrégats historiques (ensemble de grandeurs caractéristiques de l'historique d'une entité) pour chaque utilisateur.
Ce traitement est réalisé de manière distribuée par un ensemble de serveurs traitant chacun une partie du flux de données et recombinant les résultats de manière adéquate.
Ce traitement en flux permet de tenir compte en temps réel de la dernière information disponible et de pouvoir optimiser les revenus de l'éditeur de manière très réactive.

Nécessité du *BigData*

Historiquement, le réflexe du statisticien face à un jeu de données de grande taille est d'extraire un échantillon de données, c'est-à-dire de ne conserver qu'une fraction de ses observations sélectionnées aléatoirement. C'est ce que font les instituts statistiques pour calculer des

agrégats macroéconomiques ou ce que font les instituts de sondage. Cette approche est valide tant que l'on ne cherche qu'à calculer un nombre limité d'agrégats.
La problématique d'AlephD est au contraire de pouvoir quantifier un nombre très important de paramètres. En effet, l'introduction du RTB a permis aux acheteurs de pouvoir fixer un prix et acheter chaque *impression* individuellement, c'est-à-dire pour chaque utilisateur sur chaque placement. Une stratégie de vente efficace pour l'éditeur ne peut se concevoir qu'à l'échelle de l'impression et il devient nécessaire de modéliser le profil en terme d'historique d'enchère de chaque utilisateur pris individuellement.
Ce passage d'un monde où l'on calcule quelques agrégats statistiques à un monde où l'on cherche à dresser un portrait individuel de chaque utilisateur nécessite d'abandonner l'échantillonnage et de traiter les données exhaustivement.
En outre, la valeur de certaines connaissances peut décroître très rapidement dans le temps. Dans ce cas le traitement en temps réel des données peut devenir nécessaire. C'est le cas de certaines grandeurs traitées par AlephD.
Cette nécessité de prendre des décisions personnalisées pour un grand nombre d'entités sur la base de données à péremption rapide est la raison essentielle d'une approche de type *BigData*.

Traitement des données personnelles par le marché

En tant qu'acteur d'un secteur fortement consommateur de données individuelles, AlephD dispose d'un observatoire privilégié sur le traitement par le marché de ces données.
Il est clair que la numérisation croissante des activités humaines et la possibilité donnée par les outils du *BigData* mettent à mal l'anonymat en ligne et fragilisent les tentatives de protection de la vie privée. Plusieurs points sont cependant à noter :

- Dans de nombreux cas, l'exploitation commerciale de données à l'échelle individuelle se fait dans le cadre d'une problématique bien précise qui ne permet pas d'avoir une idée globale des individus ni de retrouver leur identité. Certaines entreprises fournissent, par exemple, des informations sur la probabilité qu'un utilisateur soit un robot. De nombreux robots sont conçus pour augmenter artificiellement le nombre des visites ou des clics. Identifier ces robots nécessite de modéliser ce qu'est un comportement probable d'utilisateur humain et de qualifier chaque utilisateur individuellement. Dans cet exemple, l'entreprise n'est jamais amenée à constituer un profil complet d'un individu : elle ne menace pas la vie privée des utilisateurs.
- Dans d'autre cas les statistiques constituées à l'échelle des individus ne sont utilisées que par des machines. C'est le cas d'entreprises telles que Criteo où des modèles mathématiques sont conçus pour évaluer au mieux la probabilité qu'un utilisateur clique sur une bannière. Dans ce cas toute la chaîne est automatisée et l'information non pertinente pour réaliser l'objectif fixé est naturellement délaissée. Là encore, même s'il est très facile pour Criteo de prédire votre propension à cliquer sur telle ou telle publicité sachant que vous avez cliqué sur telle page d'un site d'e-commerce, il lui est très difficile de reconstruire le profil détaillé d'un individu en particulier.
- À côté de cela, l'utilisateur prend conscience de la valeur de ses données personnelles, et il apprend à gérer la diffusion de ces données. Plus le temps passe et plus il peut décider d'échanger un accès aux informations le concernant contre un contenu gratuit qu'il juge de qualité suffisante ; il peut également décider de payer un service contre l'assurance de la protection de ses informations personnelles. Certains acteurs mettent en avant cet aspect comme un argument commercial et implémentent de réels dispositifs de protection des données (on peut citer le chiffrement des données *cloud* par Apple).

En tant que précurseur de l'exploitation en masse des données personnelles le marché de la publicité digitale préfigure sans doute les évolutions d'autres secteurs. Il illustre les risques d'exploitation abusive des données, mais aussi les forces venant limiter ce risque.