

Optimisation de détection de motifs d'activité sur des traces numériques éthiques et responsables

Le monde du marketing digital au travers du Real Time Bidding (RTB) repose sur le suivi et l'analyse du comportement des utilisateurs sur le Web. Avec la trace des utilisateurs sur la toile *via* leurs interactions sur les navigateurs Web, les smartphones, les emails ou les publicités, le RTB cherche à maximiser la diffusion d'informations sur le Web. Ainsi, en suivant le parcours digital de l'utilisateur, le RTB adapte les campagnes publicitaires à leurs profils.

Avec l'arrivée du RGPD, la confidentialité des données devient un enjeu majeur pour les sociétés travaillant dans le e-commerce. Devoir gérer les profils utilisateurs et préserver leur vie privée ajoute une complexité allant à l'encontre du principe du tracking traditionnel des utilisateurs [1,2,3]. Il devient alors nécessaire de définir des profils utilisateurs adaptés aux nouvelles normes et produire ainsi un *tracking éthique*.

Par ailleurs, le trafic généré par le processus de RTB a un énorme impact sur la consommation de ressources, aussi bien sur le réseau, les serveurs de calcul, que dans l'environnement de l'utilisateur. Récemment, la thématique du marketing responsable ou d'écoconception des supports médias (*Green Design*) émerge [4,5], mais le domaine du RTB tarde encore à évoluer sur ces conceptions modulaires du processus de tracking et d'analyse. Ainsi, la possibilité de marier le RTB avec le *Green Design* devient alors un argument de poids dans une campagne publicitaire pour donner un *tracking responsable*.

La société [Kwanko](#) cherche à relever ces deux défis en adaptant leurs processus de RTB. Fondée en 2003, Kwanko est un acteur majeur de la publicité digitale à la performance sur Web, mobile et tablettes. Elle a pour but d'accompagner les annonceurs dans le cadre de la traçabilité et de la maximisation de l'impact de leurs campagnes de publicité. Kwanko facilite la connexion des marques avec leurs audiences sur le Web.

La problématique abordée dans ce sujet de recherche est multiple. Dans le cadre d'une maximisation d'impact d'une campagne de RTB, il faut à la fois préserver la possibilité de tracer des utilisateurs pour maximiser la « transformation » (*tracking optimal*), minimiser l'impact énergétique du processus d'analyse et de tracking (*tracking responsable*) et maximiser la protection de la vie privée de l'utilisateur au vu de la réglementation (*tracking éthique*). Cette problématique combine des dimensions opposées impliquant un problème de maximisation multicontraintes.

Vers des traces numériques responsables

Nous envisageons dans une première partie de redéfinir le processus de tracking et d'analyse en microcomposants modulaires [3]. L'idée est de dissocier les données personnelles de l'analyse en produisant un modèle de données adaptatif qui servira de modèle commun pour les étapes d'analyse. La séparation en microcomposants permet de quantifier l'impact énergétique de chaque composant et ainsi de l'optimiser pour en réduire le coût.

Tout d'abord, la complexité des traitements effectués dans chaque composant associé à la quantité de données à traiter (en fonction du profil utilisateur) donne le coût de chaque étape de l'analyse. La combinaison des microcomposants à base d'opérations unitaires produit une « expression algébrique de campagne » dont les opérations sont interchangeable en vue d'une optimisation. La complexité globale de l'expression algébrique donne ainsi l'impact énergétique de l'analyse RTB.

Un des avantages de cette décomposition de la chaîne d'analyse est double. D'une part, il permet de proposer un langage de haut niveau pour exprimer les besoins d'une campagne RTB intégrant les informations ciblées de l'utilisateur et les paramètres de campagne du client. D'autre part,

l'ensemble des expressions algébriques de campagne peut être maximisé aussi bien sur leur impact, la protection de la vie privée et sur l'impact multicampagne.

Pour réduire l'impact énergétique, une heuristique initiale tentera d'allouer la tâche à l'emplacement optimal pour réduire l'impact global, soit en mutualisant les calculs multicampagnes, soit en mutualisant les profils utilisateurs. Le calcul de la pertinence d'une campagne avec le profil utilisateur pourra aussi bien se faire au niveau du navigateur que sur le serveur.

Vers des traces numériques éthiques

En seconde étape, nous reposerons sur le modèle de données commun qui servira dans le processus de tracking pour préserver le profil utilisateur. Le but étant de réduire la dépendance des modèles d'analyse classique sur les profils utilisateurs, amplifiée par la tendance au blocage de ces trackers [6]. Ainsi, nous serons à même de gérer le curseur entre la précision de l'analyse en fonction de l'adhérence des utilisateurs au profilage, tendant vers un *tracking éthique*.

À l'instar des techniques de tracking visuel [7], des stratégies de préservation de la vie privée reposent sur la définition de *motifs d'activité* pour la détection de profils spécifiques (*Activity Pattern Detection*). Il est envisageable d'orienter notre modèle de données sous forme d'Activity Pattern pour le RTB. Le profil sera ainsi analysé dans l'espace utilisateur pour générer des détections locales en fonction d'une campagne dédiée. Le résultat produit alors une recommandation pour cibler l'utilisateur avec la publicité pertinente en maximisant la protection de la vie privée.

Une autre piste envisagée est d'utiliser des techniques de définir un modèle multidimensionnel de ciblage pour les campagnes et de placer l'utilisateur dans celui-ci. Afin de garantir son anonymisation, nous nous orienterons vers des techniques de répartition aléatoire avec garantie probabiliste comme utilisée dans l'allocation sécurisée de requêtes préservant la vie privée [8]. Cette approche permettra ainsi de projeter le profil utilisateur sur des profils de campagne et de cibler l'utilisateur sans connaître celui-ci.

Vers un calcul de trace numérique optimal

Le modèle de coût basé sur l'impact énergétique reposera donc sur la complexité des composants, de leur combinaison pour l'analyse, du niveau de protection de vie privée, de la quantité de données disponibles, et du niveau de précision attendu en sortie. Une optimisation multicritère est donc nécessaire pour orienter le choix de la solution d'analyse adaptée à un ensemble de campagnes publicitaires.

L'idée pour Kwanko est de proposer un *service adaptable* à leur client en tentant de répondre à différentes dimensions de tracking difficilement compatible : éthique, responsable et optimal. Le client pourra ainsi accentuer une dimension en fonction de l'impact qu'il souhaite avoir dans sa campagne.

Profil : titulaire d'un BAC+5 en informatique, avec des connaissances solides en distribution de données, pattern mining, secure data processing, mais également une forte expérience en développement est recommandée.

Cette thèse de doctorat sera financée par un contrat CIFRE avec Kwanko, en partenariat avec le laboratoire DVRC de l'Association Léonard de Vinci (Paris La Défense) au sein du groupe digital, encadrée par Nicolas Travers (HDR) et Cédric du Mouza (HDR).

La thèse se déroulera à mi-temps chez Kwanko à Bourg-la-Reine (92340) et au DVRC à la Défense (92916).

Contact : nicolas.travers@devinci.fr

Bibliographie

- [1] T. Bujlow, V. Carela-Español, J. Solé-Pareta and P. Barlet-Ros, "A Survey on Web Tracking: Mechanisms, Implications, and Defenses," in *Proceedings of the IEEE*, vol. 105, no. 8, pp. 1476-1510, Aug. 2017.
- [2] T.-C. Li, H. Hang, M. Faloutsos, P. Efstathopoulos, "TrackAdvisor: Taking back browsing privacy from third-party trackers", *Proc. 16th Passive Active Meas. Conf.*, pp. 1-12, Mar. 2015.
- [3] Rakesh Agrawal and Ramakrishnan Srikant. 2000. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (SIGMOD '00). ACM, New York, NY, USA, 439-450.
- [4] Daniel Schien, Paul Shabajee, Stephen G. Wood, and Chris Preist. 2013. A model for green design of online news media services. In *Proceedings of the 22nd international conference on World Wide Web* (WWW '13). ACM, New York, NY, USA, 1111-1122.
- [5] Yu-Che Huang and Chia-Cheng Hsu. 2018. A study for design practice in the application of green design courses. In *Proceedings of the 10th International Conference on Education Technology and Computers* (ICETC '18). ACM, New York, NY, USA, 61-68.
- [6] Istemi Ekin Akkus, Ruichuan Chen, Michaela Hardt, Paul Francis, and Johannes Gehrke. 2012. Non-tracking web analytics. In *Proceedings of the 2012 ACM conference on Computer and communications security* (CCS '12). ACM, New York, NY, USA, 687-698.
- [7] Cassandra Carley. 2018. Balancing Privacy and Utility with Pattern Based Activity Detection: Extended Abstract. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '18). ACM, New York, NY, USA, 360-361.
- [8] Julien Loudet, Iulian Sandu Popa, Luc Bouganim: SEP2P: Secure and Efficient P2P Personal Data Processing. *EDBT 2019*: 145-156