Titre : Transparence et responsabilité des algorithmes de personnalisation

Sujet : Dans notre Société de l'Information, le profilage des internautes à des fins de recommandation et de personnalisation est devenu la norme [AEE+14], ce qui permet le développement de services ciblés sur les besoins spécifiques des individus, mais soulève aussi d'importantes questions par rapport au respect de leur vie privée. Ainsi, le consentement d'un individu par rapport à l'utilisation des traces numériques et des données personnelles collectées pour des fins de personnalisation est rarement demandé. Pour que ce consentement soit libre et éclairé, l'individu concerné devrait pouvoir inspecter les données formant son profil et aussi avoir des explications sur comment l'algorithme de personnalisation traite ce profil. Sans cette transparence sur la manière dont fonctionne le processus de profilage et de personnalisation, il devient impossible pour un individu de mettre en doute une décision automatique prise à son encontre en demandant à l'algorithme de personnalisation de lui « rendre des comptes » sur le pourquoi de sa décision [RB13]. Par exemple, on peut imaginer que sur la base des informations collectées à son sujet, un individu voit le montant de sa cotisation d'assurance médicaments doublé ou encore qu'on lui refuse l'accès à une ressource sans qu'il puisse s'opposer à cela. De plus, cette transparence est un prérequis pour pouvoir analyser les éventuels biais que pourrait avoir l'algorithme de personnalisation (par exemple en discriminant par rapport à certaines catégories sensibles de la population) et éventuellement les corriger.

À l'heure actuelle, le fonctionnement de la plupart des algorithmes de personnalisation est le plus souvent opaque et pouvoir « ouvrir la boîte noire » pour pouvoir inspecter leur fonctionnement interne est un défi complexe [Pas15]. Certains travaux préliminaires ont commencé à investiguer comment lever l'opacité de certains systèmes de profilage [Dia14,FAT14,FAT15], mais un cadre général permettant de répondre à cette problématique est toujours manquant. Pour répondre à cette problématique, l'objectif principal de cette thèse est précisément l'étude et le développement de méthodes permettant d'assurer la transparence des algorithmes de personnalisation, en particulier en reconstruisant leur fonctionnement interne à partir des entrées/sorties de ces algorithmes, dans le but d'analyser puis corriger leurs biais éventuels.

Une des premières directions de recherche explorée dans la thèse sera la caractérisation précise des données qui sont collectées sur un individu et comment elles sont transformées ensuite en profil avant d'être exploité par l'algorithme de personnalisation. Les données collectées peuvent être fournies explicitement par l'individu (par exemple dans le cas de requêtes) ou encore générées de manière implicite (par exemple son comportement de navigation). Une fois les données identifiées, l'étape suivante de la recherche est de pouvoir obtenir des exemples d'entrées\sorties pour les cas d'usage identifiés. Pour cela, il est envisagé de « sonder » directement le système de personnalisation à partir de profils générés de manière artificielle et d'observer la sortie de l'algorithme ou encore de faire appel à une approche participative où des utilisateurs pourraient directement contribuer avec leurs données à cette étude.

La seconde direction de recherche qui sera explorée est la compréhension du fonctionnement des algorithmes de personnalisation. Pour cela, nous adopterons un point de vue apprentissage-machine du problème en apprenant un modèle (typiquement un classificateur) approximant l'algorithme de profilage à partir d'un ensemble d'entraînement composé d'exemples d'entrées (ici les profils) et de sorties (par exemple l'information ou le service personnalisé). Comme il est difficile de prédire à priori quelle famille d'algorithmes d'apprentissage est le plus adaptée pour cette tâche, nous testerons différents algorithmes (arbres de décisions, réseaux de neurones, machines à vecteurs de support …) pour évaluer lequel est le plus à même d'approximer le comportement d'un algorithme de profilage.

Enfin, une fois qu'une « bonne » représentation de l'algorithme de personnalisation a été apprise, la troisième direction de recherche s'intéressera à évaluer les biais éventuels de l'algorithme et leurs impacts sur les utilisateurs. Pour quantifier ce biais, nous nous intéresserons aux métriques existantes dans la littérature en informatique [RR13], mais aussi à la conception de nouvelles métriques prenant en compte les interactions avec le droit et la sociologie. En particulier, il existe déjà des lois concernant la discrimination qui protègent particulièrement certains types des données (religion, opinion politique, origine ethnique …) pour lesquelles on peut évaluer le biais de l'algorithme. De plus, les études sociologiques existantes permettront aussi de mieux cerner la frontière pour les individus entre personnalisation légitime et discrimination injuste.

[AEE+14] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan et Claudia Diaz. *The web never forgets: Persistent tracking mechanisms in the wild*. In Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS 2014), 2014.
[Dia14] Nick Diakopoulos. *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. Tow Center. Février 2014.
[FAT14] *Fairness, Accountability, and Transparency in Machine Learning*. NIPS workshop, 2014 : http://www.fatml.org/2014/
[FAT15] *Fairness, Accountability, and Transparency in Machine Learning*. ICML workshop, 2015 : http://www.fatml.org
[Pas15] Frank Pasquale. 2015. *The Black Box Society, The Secret Algorithms that Control Money and Information*. Harvard University Press.
[RR13] Andrea Romei et Salvatore Ruggieri. *Discrimination Data Analysis: A Multi-disciplinary Bibliography*. Discrimination and Privacy in the Information Society 2013: 109-135.
[RB13] Antoinette Rouvroy et Thomas Berns. *Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individualisation par la relation ?*, *Réseaux*, n° 177, 2013.