Challenges in archiving the personalised Web

Erwan Le Merrer*[1], Camilla Penzo*[2], Gilles Tredan*†[3], and Lucas Verney*[4]

[1]INRIA Rennes – INRIA Rennes - Bretagne Atlantique, Campus de Beaulieu, Rennes, France – France
[2]Pôle d'Expertise de la Régulation Numérique (PEReN) – PEReN – Pôle d'Expertise de la Régulation Numérique 120, rue de Bercy Bâtiment Necker Télédoc 767 75572 PARIS CEDEX 12, France
[3]LAAS CNRS – LAAS CNRS TOULOUSE – France
[4]Pôle d'Expertise de la Régulation Numérique (PEReN) – PEReN – France

Abstract

Challenges in archiving the personalised Web

1 Introduction: the Web is personalised

The decision-making algorithms at the core of large online platforms tailor their content to each user by ranking, classifying, and filtering information (user personalisation). In doing so, they drive the way information reaches people and diffuses among them. People depend on search and recommendation systems, and build opinions and decisions based on what algorithms feed them with.

The Web has thus already evolved from its static initial structure, to one where each user sees an evolving, algorithmically-tailored version of the Web. To date, this algorithmic personalisation is an invisible power, as e.g., : about 62% of Facebook users are unaware of the existence of Facebook's News Feed recommendation algorithm; users tend to label content found through searches as highly trustworthy, enough to change their political views and voting behaviours; or finally, despite their critical role in selecting and ranking the most relevant information, these algorithms typically do not consider the veracity of information they present to users and might end up proposing disinformation and building filter bubbles.

In consequence, collecting the whole Internet might be of little help if one wants to understand experiences and journeys (cf the Christchurch call), as by definition personalisation makes each experience unique. Archiving user journeys among algorithmic decisions becomes vital to understand individual and group dynamics, a salient need in multiple contexts (e.g., e-Commerce, Web search, social media). Our talk focuses on challenges raised by such initiative and showcase a practical setup: the capture of YouTube recommendations during French 2021 presidential elections.

2 The problem: collecting personalisation is challenging

*Speaker
†Corresponding author: tredan@laas.fr

First, technical issues relate to the difficulty of developing tools that reliably collect personalized trajectories on the web. Second, methodological issues relate to the difficulty of observing personalized experiences.

(i) Realism and Representativity: To capture web journeys, one can emulate real users on the web, but this emulation is technically challenging. Another strategy is to capture journeys produced by actual users. Yet capturing such (private) data raises legal challenges. In both cases, the significance of the observation made as compared to the general population experience need to be questioned.

(ii) Mainstream vs Fringe: While a statistical approach might allow to capture an "average personalized experience", fringe personalization (i.e., that only impact a small user fraction) might be of high importance. Cases like Cambridge Analytica or rabbit holes argue that personalization of a minority of users can produce large impact, and hence yield high historical interest.

(iii) Platform Opacity: Platforms are opaque. Identifying what is personalized, and based on which collected information is a challenge that directly impacts personalization archiving.
(iv) Frugality (or Proportionality): the collect must be proportionate to the considered audit task. Collecting an excessive amount of data would put a high pressure on the platform infrastructure, while possibly interfering with the decisions made to future users (by the mere act of retraining the algorithms on the recent logs/users actions on the platform).

Keywords: Personalised web archiving, statistical representativity, challenges, computer science, algorithmic transparency, audit