IMPROVING PERSONALIZATION THROUGH THE RESOLUTION OF ON-SITE VISITORS

Swapneel Mehta, Deep Gandhi, Siddharth Majgaonkar, Jash Mehta, Raghav Jain
Unicode Research
swapneel.mehta@nyu.edu

ABSTRACT

We are interested in modeling user preferences to aid in personalization through an understanding of their activities and interactions outside of the news website, on social network platforms. We seek to combine on-site and off-site signals aiming to expand the team's understanding of user archetypes with the context of their off-site content preferences. We are supported by the One Fact Foundation, a 501(c)(3) nonprofit and would be excited to work with the media organization on the proposed area of personalization. This document lays out some of the details of the project that we would like to collaborate on.

1 INTRODUCTION

Inferring user interests from on-site visitor activity is a crucial problem for journalism and it could sometimes prove essential in identifying inorganic activity off-site and corresponding organic engagement for popular articles. In our proposed work, we utilise ideas similar to a previously described de-anonymization algorithm for Netflix data (Narayanan & Shmatikov, 2008). Our goal is to understand how publications like the media organization may improve personalization by connecting off-site and on-site visitors into well-defined groups. We propose to use this method to study their behaviour and find correlation in activity that is a predictor of future engagement. We focus on the identification of archetype user personas that best explains the preferences of groups of users exhibiting a particular type of on-site behavior. This prevents the identification of individual users which would violate their privacy. This approach is, in principle, aligned with Google's proposed approach[1] to replace the necessity for cookies to track user activity on websites.

For our proposed collaboration, we will conduct the analysis of a set of 1000 most recent articles from the media organization based on our proposed methodology (§2) with emphasis on scalability such that this method will scale to other articles. We will evaluate the success of this project through the identification of novel topics of interest corresponding to each user archetype, that yields meaningful signals for the personalization team to increase their on-site engagement.

2 PROPOSED METHODOLOGY

The methodology for the project can be divided into the following steps:

- Create a list of the 1000 media organization articles we plan to study based on the recency of visits by users, recency of publication, or a corresponding metric relevant to the Personalization team.
- Sample a subset of all the off-site interactions (including the equivalents of likes, shares, comments, etc.) from message-board based social media websites such as Twitter, Reddit (optionally including Meta, Instagram, 4Chan, Koo app, and Gab Social dependent on project budget) using our data mining infrastructure. These items will include account-level participation in various conversations, their interaction networks, and their friend-follower networks.
- The analysis of the mined conversations is conducted in 3 phases:

[1] https://9to5google.com/2022/01/25/google-topics-api-floc/

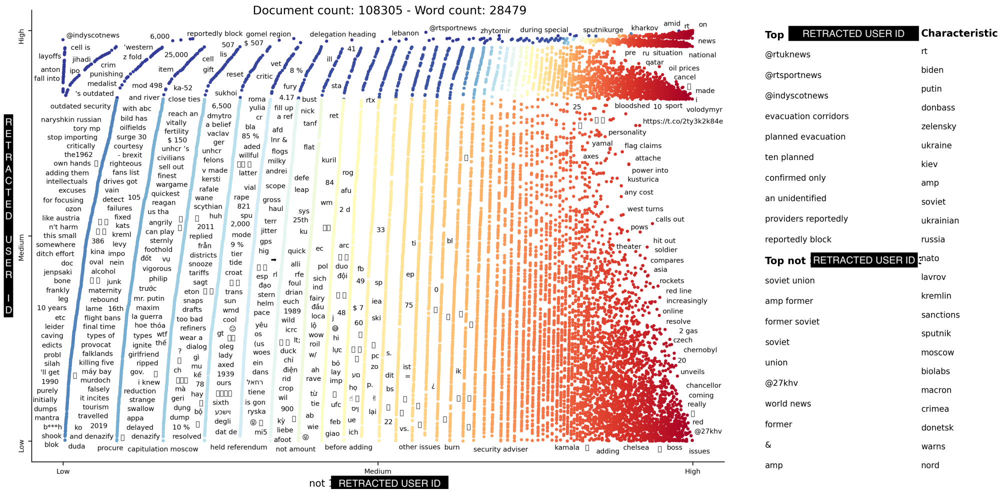


Figure 1: An example of topic modeling done for detecting anomalies about the conversations and tweets of a particular user when compared to all the others combined for a Twitter dataset.

- **Intent Detection:** We will cluster users into 198 overlapping groups based on the labels proposed in Martino et al. (2020) adapted to our use-case to track the actual intent of conversations and the proposed ”call-for-action” while discussing articles. This is useful to identify the stance of the user on the article external to their visit on-site.

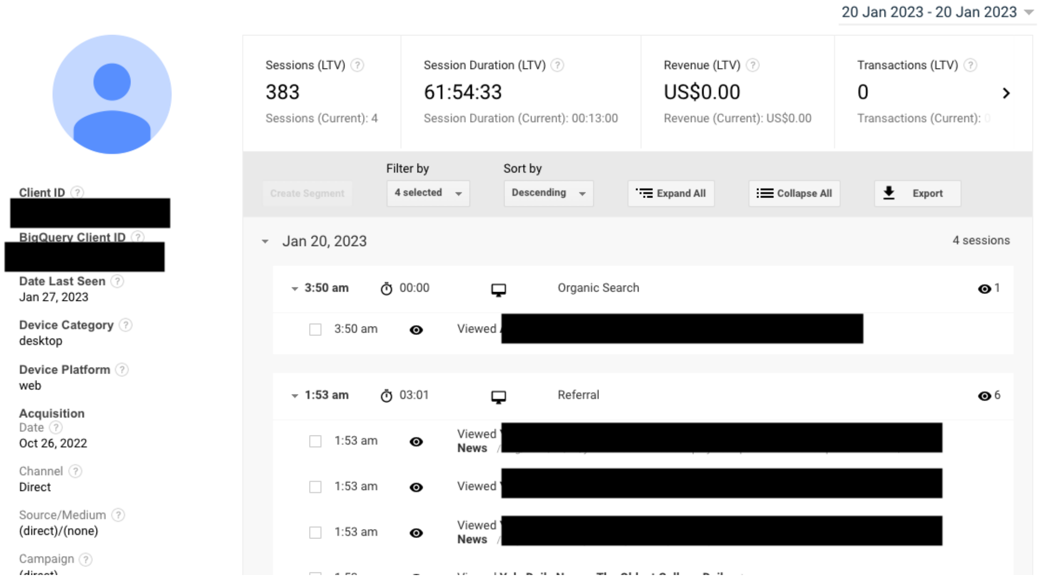
- **Topic Modeling:** We will use the ScatterText approach (Kessler, 2017) to compare the topics of conversation in each identified cluster of users to the combination of topics discussed by the other clusters. This intra and inter-cluster coherence reveals several outlier topics and helps us label a distribution of users as outliers. An example of this analysis being done for a Twitter user can be observed in fig. 1. In this case, outliers could be further studied to understand whether these might constitute gaps in coverage from the media organization or potentially signal accounts that may be engaging with media organization content in an inorganic manner lacking any consistency in topics of interest based on algorithms we have previously developed.

- **Aspect-based Sentiment Analysis:** Developing a framework to identify several *(aspect, opinion, sentiment)* pairs for every article using Gao et al. (2022). This would help identify the context (opinion) being focused on highly by different users and also give a combined score for sentiments about different aspects of the article being discussed across social media. The individual opinion focus is also stored for every cluster of a user as further metadata for the deanonymization algorithm.

- Based on the metadata obtained in the previous task (article visits by every cluster, opinion, topics discussed, etc.), we rank articles to predict which clusters on-site users belong to using the cross-correlation approach proposed in the De-Anonymization algorithm (Narayanan & Shmatikov, 2008).

We will use these correlation results to study the historical interests of a user group, validate it by reviewing their on-site activity (Figure 2), and present a report on their current and emergent interests which could be used for content creation and promotion strategies for media organization content on-site as well as to different off-site platforms including social media.

We will develop a new metric for tracking engagements using the outlier score for different users obtained from the ScatterText topic modeling, combined with their on-site activity. We expect this to improve the detection of automated site visits including bot networks and click farming.



20 Jan 2023 - 20 Jan 2023

Sessions (LTV): 383
Sessions (Current): 4

Session Duration (LTV): 61:54:33
Session Duration (Current): 00:13:00

Revenue (LTV): US$0.00
Revenue (Current): US$0.00

Transactions (LTV): 0
Transactions (Current):

Client ID
[redacted]

BigQuery Client ID
[redacted]

Date Last Seen
Jan 27, 2023

Device Category
desktop

Device Platform
web

Acquisition
Date
Oct 26, 2022

Channel
Direct

Source/Medium
(direct)/(none)

Campaign

Filter by: 4 selected
Sort by: Descending

Jan 20, 2023 — 4 sessions

3:50 am — 00:00 — Organic Search — 1
3:50 am — Viewed [redacted]

1:53 am — 03:01 — Referral — 6
1:53 am — Viewed News [redacted]
1:53 am — Viewed [redacted]
1:53 am — Viewed News [redacted]

Figure 2: An example of how Google Analytics tracks activity of a specific user on a website.

3 Our Team

Over the past 3 years, our team of social scientists, data scientists, and student researchers has worked on critical problems in the domains of artificial intelligence, machine learning, natural language processing, and social media analysis. We have published at leading conferences in these domains, and delivered invited talks at MIT, Stanford, Oxford, Twitter, Facebook, and other organizations. We have applied our AI and social science expertise in the local news space, launching pilot projects with the Vermont Digger and the Yale Daily News to study online audiences. We have organized the AI for Everyone workshop at Computation + Journalism at Columbia, and are working with The Times & the Sunday Times (UK), Deutsche Welle, and Fundamedios on various collaborative projects to develop AI tools and strategize on combating online disinformation for their newsrooms. Recently, we received a grant from the Wikimedia Foundation and Craig Newmark Philanthropies to support our work. We have previously received grants from the NYC Media Lab, Google, Amazon, and Oxford's AI4ABM Foundation.

References

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 7002–7012, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.610`.

Jason Kessler. Scattertext: a browser-based tool for visualizing how corpora differ. In *Proceedings of ACL 2017, System Demonstrations*, pp. 85–90, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL `https://aclanthology.org/P17-4015`.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*, 2020.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.