This Day in Ads: Capturing Cultural Shifts In Commercial Messages

Target: https://mc.manuscriptcentral.com/pdtc Preservation, Digital Technology and Culture

Resources

https://github.com/savingads/a-proxy

We are proposing

a tool for archivists to select a set of target websites to capture on a specific day from a specific perspective. The Python based tool will crawl the website using the techniques we learned of in the Saving Ads dataset creation. The result will be a navigable web archive of specific websites at specific time points with the central focus of preserving their advertising content from the perspective of a user that models a chosen set of attributes. The interface will also allow a more expansive capture setting that requires more resources but that can then replay captured pages from a variety of perspectives. A more expansive capture may archive the same website in different regions or languages on a target date.

Content

Introduction

The Web serves as a dynamic transcript of human interaction, capturing the ways people communicate, share knowledge, and express culture [[cite our own work]]. Web archives, such as the Internet Archive's Wayback Machine, play an important role in preserving these otherwise transient exchanges. Archives provide historians, researchers, and the public with a record of society's interactions and evolution, shaping collective memory in the form of granular microhistory [[cite]].

The preservation of personalized experiences that characterize modern Web pages is challenging. Elements such as personalized recommendations, dynamically generated content, and interactive features are often absent or simplified in archived versions, resulting in a disconnect between the original user experience and its archived representation [[cite]]. Traditional archiving methods, while effective for static content, are not well suited to capture the complexities of dynamic and user-specific web interactions [[cite]].

To address these challenges, strategies from commercial applications, such as targeted advertising and curated content, can be adapted and implemented as tools to archive personalized web experiences. For example, personalization techniques used in advertising, such as behavioral profiling and tracking, provide practical methods for capturing and

representing individualized interactions with pages that are archived but also for reproducing these interactions to simulate behavior. Through the adaptation and application of these strategies, it also becomes possible to examine how personalization shapes user behavior and influences how people interact with, interpret, and respond to personalized content based on a set of experimental criteria [[cite]].

A methodological framework based on advertising-inspired strategies, user modeling, and ontology-based systems is proposed to enable the reconstruction of personalized experiences in the replay of archived Web pages. Ontology-based approaches are presented as a way to organize and contextualize archival content, addressing the limitations of static captures when dealing with dynamic and personalized interactions and interacting characteristics that influence each other. Simulated user interactions are also considered as a means of analyzing the impact of personalization on user experience and preserving these dynamics within archived content.

The discussion first examines the theoretical foundations of this framework, including advertising-based methods, ontology modeling, and user simulation. Then the methodological components of the framework are explained, focusing on user behavior analysis, profile creation, and simulated testing. Finally, the text addresses other challenges, such as scalability, privacy, and accessibility, and explores applications and future directions for research.

The Role of Advertising Data in Personalization:

Advertisers rely on data collection to tailor content to individual users, utilizing demographic, behavioral, and contextual information to create highly personalized experiences. These data include characteristics such as age, location, and interests, along with detailed behavioral patterns such as browsing history and interaction preferences. Using this type of information as a foundation, it becomes possible to explore how similar techniques could enhance the experience of an archive user.

Advertising Data Types and Their Role in Personalization

- ### Demographic Data:

  Location Data

  Kelly et al. - 2013 - A Method for Identifying Personalized Representations in the Archives

  Kelly et al. (2013) identified issues with personalized representations in web archiving and presented a prototype that extends GeoIP and browser environment. Specifically, they provided examples that some websites display different contents including news and weather information depending on the user's GeoIP or user-provided zip code. This shows how the geolocation can impact on the contents presented to a user.

  Some sites provide local news and weather content based on the GeoIP of the requester. It provides an example of nbcnews.com displaying different content (news

and weather) based on whether the user's GeoIP can be interpreted. This demonstrates how geolocation can influence the content presented to a user

Kanoje et al. 2015

An e-tourism website was able to deliver personalized information based on the user's location and provide recommendations for nearby tourist spots

Le Merrer et al. - 2024 - Challenges in archiving the personalized web

Coarse-grained personalization involves tailoring content based on a user's inferred location, such as displaying a user interface in French to users with a French IP address

When incorporated into archival systems, geographic data can be used to generate user personas that simulate location-specific interactions. For instance, an archival replay system might reproduce a personalized web experience by simulating a user accessing content from a specific region. This could involve serving localized versions of web pages, offering regionally relevant search results, or even presenting ads and recommendations that mimic those originally shown to users in that area. By leveraging geolocation data, web archives can better preserve the cultural and regional nuances of online interactions, providing future researchers with a more accurate representation of the web as an evolving, geographically diverse ecosystem. This approach aligns with advertising techniques that dynamically adjust content based on location, demonstrating how geographic personalization can enhance both user engagement and the authenticity of archived web interactions.

Demographic data, which includes attributes such as age, location, and other user characteristics, is central to understanding online behavior. Advertisers rely on this information to create targeted campaigns that resonate with specific audience segments. Tailoring advertisements to user demographics enhances relevance and engagement, leading to increased click-through rates and overall effectiveness (Carrascosa et al., 2015; González et al., 2021). These practices allow businesses to refine their strategies by analyzing how different demographic groups respond to campaigns (Andrés et al., 2015).

The application of demographic profiling extends beyond simple categorization, enabling the development of sophisticated personalization mechanisms. Sowbhagya et al. (2022) explored how demographic information can be integrated with behavioral data to create more nuanced and accurate user profiles. Their research highlighted the importance of combining multiple data points to achieve more precise personalization, showing that demographic attributes often serve as crucial contextual markers that help interpret and predict user intentions and preferences. This multi-dimensional approach to user profiling allows for more refined targeting strategies and improved user experience customization.

- **Definition**: Demographic data includes age, gender, education level, location, and socioeconomic status.

  De Andrés, J., Pariente, B., Gonzalez-Rodriguez, M., & Fernandez Lanvin, D. (2015). Towards an automatic user profiling system for online information sites: identifying demographic determining factors. *Online Information Review*, *39*(1), 61-80.

- **Use in Advertising**: Advertisers use demographic data to target users based on general population segments, creating relevant ad content based on broad characteristics.

  Sowbhagya, M. P., Yogish, H. K., & Raju, G. T. (2022, July). User profiling for web personalization. In *2022 IEEE International Conference on Data Science and Information System (ICDSIS)* (pp. 1-5). IEEE.

- **Application to Archival Analysis**: what elements of retrieval and replay does this involve?

- ## Behavioral Data:
  - **Definition**: Behavioral data captures users' online activities, including website visits, time spent on pages, ad clicks, and past purchase history.

    Fan, X. X., Chow, K. P., & Xu, F. (2014). Web user profiling based on browsing behavior analysis. In *Advances in Digital Forensics X: 10th IFIP WG 11.9 International Conference, Vienna, Austria, January 8-10, 2014, Revised Selected Papers 10* (pp. 57-71). Springer Berlin Heidelberg.

  - **Use in Advertising**: Behavioral data allows advertisers to serve content that is highly relevant to user engagement, tracking patterns over time to predict future actions.

    Yang, Y. C. (2010). Web user behavioral profiling for user identification. *Decision Support Systems*, *49*(3), 261-271.

    Carrascosa, J. M., Mikians, J., Cuevas, R., Erramilli, V., & Laoutaris, N. (2015, December). I always feel like somebody's watching me: measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies* (pp. 1-13).

  - **Application to Archival Analysis**: Behavioral data could track user progress through archival interaction, identifying which types of content are the most directly relevant and which presentation is preferred.

Behavioral data captures the online activity of users, such as website visits, time spent on pages, ad interactions, and purchase history. This information includes explicit actions like ad clicks as well as implicit patterns such as browsing frequency or session duration, helping to create personalized digital experiences by uncovering user preferences and habits \cite{fan_web_2014}. This data provides insights into user preferences, habits, and decision-making patterns that can be leveraged for archival purposes. By analyzing behavioral patterns, researchers can identify common interaction sequences, content consumption

preferences, and navigation paths that characterize different user segments. These patterns can then be used to develop models that simulate realistic user behavior when interacting with archived content. For example, tracking how users historically navigated between related pages, engaged with interactive elements, or responded to personalized recommendations can inform the development of more authentic archive replay mechanisms. The granularity of behavioral data also allows for the identification of temporal trends and evolving user preferences, which is crucial for accurately representing how web experiences changed over time.

In advertising, businesses analyze this data to tailor content to users' observed behaviors. Identifying trends, such as frequently visited pages or recurring purchasing habits, allows for refinement of targeting strategies and the prediction of future actions \cite{yang_web_2010}. These insights form the basis for dynamic campaigns that aim to align with user engagement patterns and preferences.

The broader utility of behavioral profiling lies in its potential to capture user tendencies through models that measure factors like page view time and frequency. When combined, these techniques provide data that are used in personalization and targeted engagement \cite{fan_web_2014}\cite{yang_web_2010}.

- **Psychographic Data**: - talked to Shruti Phadke (Drexel IS)?
    - **Definition**: Psychographic data focuses on interests, values, lifestyle, and personality traits.
    - **Use in Advertising**: Advertisers use psychographic profiles to tailor ads to users based on their emotional or value-based preferences. This creates more personalized, emotionally resonating ad experiences.
    - **Application to Archival Analysis**: Psychographic data could be used to customize retrieval and replay to prioritize alignment with interests or goals.

Psychographic data focuses on understanding the attitudes, values, interests, and lifestyles of consumers. Unlike demographic data, which captures external attributes like age or location, psychographic profiling examines intrinsic characteristics that might contribute to motivation or action. This includes factors such as environmental consciousness, brand loyalty, innovation, and trust. In this way, advertisers attempt to segment audiences based on psychological traits \cite{dutta-bergman_demographic_2006}.

Predictive marketing algorithms utilize psychographic data to create personalized experiences. These algorithms analyze large and heterogeneous data sets to infer consumer preferences and anticipate future behavior, supporting the design of campaigns that are intended to resonate emotionally with users. As mentioned above, environmentally conscious consumers may respond favorably to ads that emphasize sustainability, while brand loyal users may engage

more with messaging that reinforces their affinity for specific products or services \cite{kotras_mass_2020}.

This data also supports the development of creative strategies that are designed to address the aspirations and concerns of users. By understanding psychographic traits, businesses attempt align their messaging with personal values, such as promoting health-conscious products to audiences prioritizing wellness. This integration of psychographics into advertising strategies represents a move toward mass personalization, where predictive marketing mechanisms transform how corporations understand and interact with their audiences \cite{kotras_mass_2020}.

- **Contextual Data**:

Contextual data plays a critical role in enhancing the relevance of personalized experiences during archival replay. By leveraging information about a user's environment—such as device type, location, or time of day—archival systems can adapt the presentation of content to better align with the user's current context. For example, much like advertisers dynamically adjust ads based on the user's device or browsing activity, archival systems could offer shorter, mobile-optimized lessons when detecting smartphone usage or suggest content based on a user's available time (Vassio & Mellia, 2019). These real-time adjustments ensure that archived experiences remain not only accessible but also meaningfully aligned with the user's needs during engagement, bridging the gap between static preservation and dynamic user interaction. Such applications demonstrate how lessons from advertising can inform the development of systems that preserve and reproduce personalized, contextually relevant web interactions (Vassio, Metwalley, & Giordano, 2020).

    - **Definition**: Contextual data refers to information about the user's environment when interacting with content, such as device type, location, and even time of day.
    - **Use in Advertising**: Advertisers often use contextual data to adapt ads based on the specific context, ensuring content is relevant to the user's current activity (e.g., browsing on a mobile device vs. a desktop).
    - **Application to Archival Analysis**: Archival replay systems could use contextual data to optimize the presentation of content based on the users' current environment. For instance, shorter lessons could be offered when the system detects mobile device usage or limited time availability.

1. **Real-Time Personalization Using Advertising Data Models**

- **Behavioral Adaptation Based on Advertising Techniques**:

- Explain how behavioral data can drive real-time content adjustments, much like how ads adjust dynamically based on engagement history.
- Provide examples: if a user struggles with retrieving captures relevant to a specific topic, the system may offer additional resources, much as advertising systems adjust the frequency or the type of message shown.

2. **Ethical and Privacy Considerations in the Use of Advertising Data in Education**

- **Privacy Challenges**:
  - Address the ethical concerns around the required extensive data tracking, similar to privacy debates in digital advertising. Explain how user data could be misused if not properly managed.

Vassio, L., & Mellia, M. (2019, April). Data Analysis and Modelling of Users' Behavior on the Web. In *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)* (pp. 665-670). IEEE.

→ used Tstat (installed at a PoP and monitors the network, logging information from both TCP and HTTP connections) due to privacy issue in European countries, anonymized client IP address

→ discussed the privacy issue in the following study [Vassio, L., Metwalley, H., & Giordano, D. (2020). The Exploitation of Web Navigation Data: Ethical Issues and Alternative Scenarios.]

- **Framework for Ethical Data Use**:
  - Propose a clear framework for consent and ethical data use, ensuring that the benefits of personalization in education do not come at the cost of user privacy.

3. **7. Case Study: Applying a Profile to an Engagement**

- **Using Data to Enhance Utility**:
  - Show a replay in the context of a sample persona and point out ways in which we can understand the limitations of an archive based on a profile (one might be vision impairment e.g., Victoria Van Hyning https://www.imls.gov/grants/awarded/re-252344-ols-22)

4. **8. Future Directions for Research**

- **Personalized Archival Replay Tool**

Ontology user profiling

Robal, T., & Kalja, A. (2007). Applying User Profile Ontology for Mining Web Site Adaptation Recommendations. *ADBIS Research Communications*, *325*.

Application to Archival Analysis: Elements of Retrieval and Replay

"What You See No One Saw."

Archival analysis of web content, particularly in contexts involving personalization and dynamically generated pages, requires a focus on two primary processes: retrieval and replay. Each involves several interconnected elements necessary for preserving and accurately representing the content as it appeared to users.

In the retrieval phase, the goal is to capture the web content comprehensively. This includes not only static elements like HTML and CSS but also embedded multimedia and dynamic and user-specific features rendered through JavaScript or asynchronous requests. Personalization layers, such as recommendations, location-based adaptations, or session-specific displays, require specialized techniques to variations are preserved. Effective retrieval depends on capturing contextual metadata, including user-agent strings, cookies, and geolocation data, which influence the presentation of personalized content. For instance, session emulation techniques allow can allow archivists to mimic user interactions, enabling the acquisition of elements that are conditionally displayed, such as dropdown menus or pop-ups triggered by user behavior.

Normalization is another essential component of retrieval, where content variations are consolidated to provide a clear basis for analysis. This ensures that while multiple versions of a page might exist based on demographic or session-specific data, they are indexed in a structured way that supports subsequent comparisons and analysis.

The replay phase focuses on recreating the captured content in a manner faithful to its original presentation. Accurate session reconstruction is key, as it allows for the simulation of user interactions and environmental settings, such as specific browser types or geolocations. Replay systems must maintain the functional and visual fidelity of the content, ensuring that dynamic features, such as advertisements or interactive widgets, operate as they did when first rendered. This fidelity is critical for longitudinal studies or usability analyses where researchers aim to understand how users experienced the content at a specific point in time.

Replay systems should also support temporal analysis, enabling comparisons of content across different time snapshots. This is particularly valuable for studying how web designs and personalized adaptations evolve over time. Furthermore, effective archival systems should provide search and accessibility features, allowing users to query and retrieve specific content variations and replay them with their personalized elements intact.

User Attributes Hierarchy

1. Demographic Data

- **Persona Traits:** Age, Gender, Socioeconomic Status, Education Level, Ethnicity
- **Ontology Entity:** CCO:Agent

- **Simulated Behavior:** Different user profiles (e.g., student, professional, retiree)

2. Behavioral Data

- **Persona Traits:** Browsing Patterns, Clicks, Session Length, Page View Times, Purchase History
- **Ontology Entity:** BFO:Occurrent (Processes reflecting ongoing activities)
- **Simulated Behavior:** Simulating browsing sessions or purchase behaviors

3. Psychographic Data

- **Persona Traits:** Interests, Values, Lifestyle Preferences, Attitudes
- **Ontology Entity:** BFO:SpecificallyDependentContinuant (Qualities depending on user context)
- **Simulated Behavior:** Personalizing content to match inferred preferences

4. Contextual Data

- **Persona Traits:** Device Type, Location, Time of Access, Sequence of Visited Pages, Environmental Context
- **Ontology Entity:** BFO:SpatialRegion, BFO:TemporalRegion
- **Simulated Behavior:** Simulating usage in different environments or device configurations

5. Derived Insights and Aggregates

- **Persona Traits:** Composite Scores (e.g., Engagement Index, Conversion Probability)
- **Ontology Entity:** Aggregate metrics combining attributes from the above categories
- **Simulated Behavior:** Algorithmic scenarios to predict user responses

Aspects

Accessibility

Mobile (device)

References

User Profiling_Literature

Persona_Literature

https://github.com/tencent-ailab/persona-hub

https://github.com/webrecorder/browsertrix-behaviors

https://www.selenium.dev/documentation/webdriver/, https://github.com/SeleniumHQ/selenium