Personal data visualization for exploring web activity

Sander Marx

[1] University of Bergen

1. Introduction

In recent years the talk of personal data collected from corporations has become more and more mainstream. The media often talks about the tracking and privacy concerns this personal data collection brings, in combination with GDPR (General Data Protection Regulation) this has led corporations like Google, Facebook, Twitter and TikTok to allow users to download their personal data. However, this data is often big and consists of many folders and files, making it hard for users to get meaningful info out of them without spending a lot of time scouring through their files to find relevant info. Data visualization in this context will help provide the info concisely and gets the users to absorb the info with minimal time and effort. Efforts have been made by several researchers to visualize the different types of personal data some examples include. Schrifin et al. [SRKK21] with their *TransparencyVis* tool that allows user to upload their downloaded personal data and then provide simple visualizations plotted over time to better help with exploration. Another effort by Thudt et al. [TBHC16, TBC13] focuses on positional data and how it is possible to explore old positional data to relive memories. However, none of the contributions focus on comparing visualizations from different online social networks.

The web and a user's web activity are a critical part of the daily life of most humans. However, web browsing history does not tell the whole picture, other social media like YouTube, Twitter Facebook etc. are important to consider as well. For users of Google Chrome and YouTube, Google provides personal browser data and personal YouTube watch history from their Google Takeout platform. Therefore, in the context of this paper, the focus will be primarily on browser history and YouTube watch history. We will explore how visualizing users' personal browser history and YouTube watch history can enable reflection and discovery of their behaviors. Specifically, this paper will focus on if it is possible to get a clearer picture of the user's web activity by visualizing both browser history and YouTube watch history and comparing the visualizations. We achieve this by proposing a website built with the JavaScript library, D3.JS [BOH11] as the tool for the visualizations.

2. Related Work

In this section, we will discuss the work researchers have done previously, relating to the problem of this paper. In the following section, we present the most relevant groups of related work that we have found. The section will be split into three:

- Work relating to personal data visualization
- Work relating to browser history visualization and
- Work relating to YouTube watch history visualization

2.1. Personal data visualization

In our research of related work, we have found several helpful approaches for visualizing personal data. Schrifin et al. [SRKK21] introduce their web-based tool *TransparencyVis* which allows users to interactively explore and analyze personal data exports stored by different online services like Facebook, YouTube, Google etc. the visualizations used by *TransparencyVis* is Tree-map [Shn92] of stored files and a scatter plot, plotting files stored over time and color the data points according to type. *PrivInferVis* [SSS*21] is another tool that aims to enhance the transparency of data collected from online social networks, this is similar to the motivation of *TransparencyVis*. *PrivInferVis* focuses on helping users assess and visualize their risks of attribute inference derived from personal data from multiple different social media networks. While these works primarily focus on giving more transparent ways of showing users how their privacy is being invaded by corporations, other works approach personal data visualization for the purpose of reminiscing and reflecting, this is more like the aim of the work discussed in this paper.

The work of Thudt et al. [TBHC16] focuses on positional data by proposing a web application, *Visual Mementos*. With *Visual Mementos* users can upload their personal location data retrieved from services like Google Maps, *Visual Mementos* then uses this data to visualize locations as a sequence of visited places represented with map segments, with the help of Google Street View the user is able to "relive" their memories using personal positional data. Another contribution of personal data visualization used for reminiscing and reflection is *Last History* by Baur et al. *Last History*is an interactive visualization for displaying music listening histories, along with contextual information from personal photos and calendar entries, the main goal of *Last History* is analysis and reminiscing.

2.2. Browser history visualization

There exists a handful of browser add-on extensions to help users better explore their browsing history, and a number of these extensions use visualization as a tool to achieve this. *WebHistorian* [MT16] is a browser extension designed to quickly inform users

© 2022 The Author(s)

of what their browsing data contains with visualizations built using D3.JS as well. The included visualizations are a time heat map, a network visualization between websites browsed from and to, a word cloud of search terms, and a bubble chart. *popHistory* [CKM17] is another one of these browser extensions but with a focus on interactivity and animations on the visualizations to better help users reflect on their browsing habits. *popHistory* includes an interactive sunburst visualization and an animated visualization of the user's browsing history with a bubble chart as well. Carrasco et al. [CKM17] found from their work on *popHistory* that users found it much easier to analyze and reflect on their web browsing using interactive and animated visualization. *HistoryLane* [Cht12] is a browser extension with the goal of enabling the user to gain insight into their own parallel browsing patterns over time by interactive visualization, Chtivelband [Cht12] from his research found the same results as Carrasco, namely that users found it helpful to use visualizations when analyzing browsing history.

2.3. YouTube watch history visualization

Visualization of personal YouTube watch history is sparse, and there exists few scientific works relating to this topic. Al-Hajri et al. [AHMFF14] in their paper explore the use of YouTube watch history to better navigate between watched videos using grid and sized-based visualization layouts. In Wang et al. [WY20] masters thesis an interactive visualization platform of YouTube personal data is presented that applies similar methods to the proposed work of this paper. Their visualizations include a line graph of video frequency per month and per week as well as a donut chart showing the user's most watched categories, and a word cloud of the user's popular search terms. The proposed work in our paper also shows the frequencies of watch history per month and week however our work also shows frequency per hour in the day and compares this to browser history.

3. Solution

The proposed solution is a web application presenting visualizations from users personal data developed with D3.JS.

3.1. The Data

The data used for the visualizations are taken from the Google takeout platform and consist of two JSON files one for browser history and one for YouTube watch history. When signing in to the Google takeout platform the user is allowed to choose what types of files to extract. To get the data used in the proposed solution the user must select the "Chrome" and "YouTube" module, for the YouTube module the user needs to click more formats and select JSON as the file format. The data is then sent as a zipped file that can be extracted to get the two JSON files: BrowserHistory.json and watch-history.json. These files provide temporal information of the user's history that can then be plotted over time to provide visualizations.

3.2. Temporal Charts

The proposed solution is developed with the javascript library D3.JS [BOH11] which is a flexible and efficient tool specialized for data visualizations. We split the data by the year of the entries and use D3.js to create three types of temporal charts: a Bar chart, a line chart, and a circular bar chart with the appearance of a clock. The data is plotted over different time intervals, browser and watch history by month, day of the week, and hour of the day. By visualizing over different time intervals we can learn more from the data on a more specific level than by using just one time interval. Bar graph is a good approach to this problem because it is a good tool for comparing sizes, and to show frequency distributions, this will help the users to find the specific days of the week and hours of the day that they spend the most and least time browsing the web and watching YouTube. This will give users a better view of for example excessive use of the web at times that may not be ideal. Line graph is fitting for showing trends and changes over time, this will help the users figure out how their web activity changes over the year and can help users reflect and ask them self questions about their browsing activity in specific parts of the year. The proposed solution consists of in total eight graphs:

Four bar graphs

- One for browser history by hour of the day
- One for watch history by hour of the day
- One for browser history by day of the week
- One for watch history by day of the week

Two line graphs

- One for browser history by month of the year
- One for watch history by month of the year

Two circular bar graphs

- One for browser history by hour of the day
- One for watch history by hour of the day

We provide the users with two options when visualizing browser and watch history by hour of the day, a standard bar graph and a circular bar graph meant to represent a clock. The circular bar graph meant to represent a clock has the goal of giving the user a more pleasing and familiar visualization of the data that refers to the time in a day. However, it might not be as easy to analyze the values of the bars in the circular bar graph, so a standard bar graph is also provided.

3.3. Colors

In the proposed solution we have used a color range between yellow and red on the browser and watch history by the hour while browser and watch history by month and day of the week uses a uniform blue color. The frequency distribution of hours a day is twenty-four, and because of this, we get a lot of bars with different values, by using a color range it is possible to help the user easily distinguish between high values and low values without the whole chart looking uniform, we achieve this by giving the highest value a bright red so that the users draw more attention to this bar. The frequency of days in a week is seven, meaning the graph will consist of seven bars. The values of each bar do not fluctuate as much therefore a color range on the bar graph of browser and watch history by day of the week would not make sense as this would draw the attention of the user away from the data and make them focus more on

© 2022 The Author(s)
Eurographics Proceedings © 2022 The Eurographics Association.

the colors than necessary. A line chart does not benefit much from using a color range and can often be more distracting and harder for the viewer to understand than just using a static color.

3.4. Comparison

The comparison of the two different data sets is done by placing the respective graphs beside or on top of each other in the same container. This way the user can easily analyze and compare both browsing history and watch history by glancing from one graph to the other to find possible similarities or patterns.

4. Results

The result of the proposed solution using browser and watch history from January 1st 2022 to November 30th 2022 is shown in figure 1. Here we can see six graphs, the alternative circular bar graph for watch history by hour of the day is shown in figure 2.

4.1. Hour of the day

We can start by analyzing the two first bar graphs, browser and watch history by the hour seen in figure 1. From the browser history by the hour graph, we can observe that the user's web activity is high between 13:00 to 23:00 while between 00:00 and 13:00 the user's web activity is lower. Now if we look at the watch history by the hour, we can get a different look at the user's web activity. Between 04:00 and 12:00, the activity is also low, this is the same result that we can observe from browser history by the hour. Yet, when we look at the other parts of the day, we notice a difference. The user's watch history is highest between 00:00 and 02:00 and lower between 13:00 and 23:00, this is the reverse of the values we see in the browser history graph. Now if we combine what we observe from both graphs we can conclude that the user is most active on the web between 13:00 and 02:00, this is also observable in figure 1. Had we only analyzed the graph for browser history or watch history by itself we could not have come to this solution. This shows that visualizing both browsing and watch history, in fact gives us a clearer picture of the user's web activity.

4.2. Month of year

We look at the next set of graphs that visualizes information on web activity per month of the year, the line graphs in figure 1 represent this. From the browser history by month graph, we can see that the user's web activity peak in November and is at its lowest between June and July, we can see that we have higher activity between the months of January and May, low activity between June and August, then we have the highest activity between September and November again. Now when looking at the graph visualizing watch history by month, we see that the user's web activity is at it lowest in January and then peaks to its highest value immediately after in February. The graph dips in June before quickly rising again in July after which the graph decreases until October before it rises again in November, similar to the browser history graph. The watch history graph does not have a pattern like the browser history graph where the web activity is noticeably lower in the summer months and higher the rest of the year. Yet the line graphs helps us spot spikes and trends in the data, this helps the users of the application to reflect and discover behaviours in their web activity that they previously had not known about. If we look at figure 1 as an example the user could ask them self what happened in June that caused a abrupt dip in YouTube watch history but not in browser history. Or the user could ask them self why both their browsing and YouTube activity increases in November. One can argue that this reflection and discovery would be possible with just one of the graphs, however the user gains more information that gives them better options of reflection and discovery given both the visualizations of browser history and YouTube watch history.

4.3. Day of week

Analyzing the bar graphs representing browsing and watch-history by day of the week we notice a few differences. The user's web activity when looking at browsing by day of the week is mostly higher between Monday and Thursday while being lower on the weekend between Friday and Sunday. While looking at the user's web activity from the watch history by day of the week we see that the values of the bars for the whole week are mostly uniform except for Sunday. When combining observations from both graphs we can conclude that the user's web activity is lowest on weekends especially Sundays while the rest of the week between Monday and Friday the user's web activity stays mostly the same. Had we only visualized browsing-history by day of the week, we would assume that the user's web activity is mostly focused between Tuesday and Thursday, however, when looking at both graphs we find out that the web activity is higher the entire week except Sunday, this, in turn, gives us a clearer picture of the user's web activity

5. Discussion

The presented results show that we can benefit from visualizing browser and YouTube watch history together to get a clearer picture of a user's web activity. Looking at pros and cons of the implemented solution, we find:

Pros

- The results are simple to replicate
- Visualization over different time intervals makes it simple to get a clear picture of web activity
- Simple graphs and minimal interaction make the website easy to use

Cons

- Comparing methods might be too simple to get meaningful comparisons
- No way for the user to get exact values from the graphs
- No way for the user to import data directly through the website.

The biggest issue with the current solution is that a lot of the watch history data also exists in browser history data if the user uses primarily their browser to watch YouTube. A solution to this is to remove the data that exists in both data sets, another solution is to use the YouTube API to extract watch history that was not

© 2022 The Author(s)

watched in the browser. Potential further improvements might include the addition of watch history from other platforms like Twitter, Facebook, TikTok, etc. Adding a module for importing data directly through the website and a simple guide showing the user how to retrieve their personal data. Improvement of comparing methods may include implementing methods like brushing and linking, potting both line graphs in the same view, or implementing additional visualizations like a stacked bar graph or stacked area graph.

6. Conclusion

In this paper, we have presented our design study on giving users better insight into their web activity. The study focused on creating a web application that visualizes personal data, specifically browser history and YouTube watch history. Lessons from this study show that a comparison of multiple sources of web history helps to paint a clearer picture of a user's web activity, however, it is not reasonable to assume that only visualizing browser history and YouTube watch history is enough to get the whole picture. As there exist so many different social media platforms with active users today. On top of that different nationalities use different social media provided by different actors. It will be difficult to include support for all of them in a single application for visualizing web activity. With that said we could improve the application by focusing on western audiences and providing the option to visualize personal data from the biggest and most popular actors in the west. Future work should then attempt a user study to evaluate the success of the application and help further discover opportunities for improvement.

© 2022 The Author(s)
Eurographics Proceedings © 2022 The Eurographics Association.

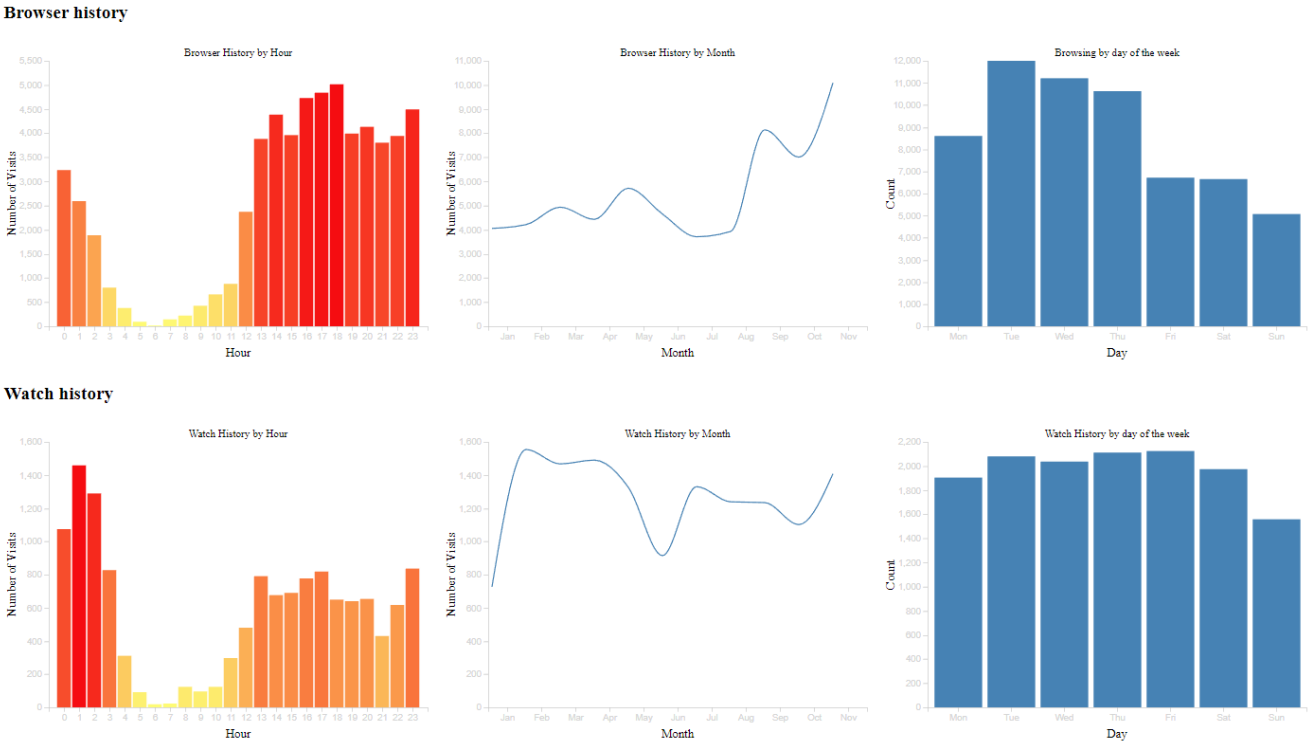

Figure 1: *Browser and Watch history visualizations. Data from January 1st 2022 to November 30th 2022*

Circular charts

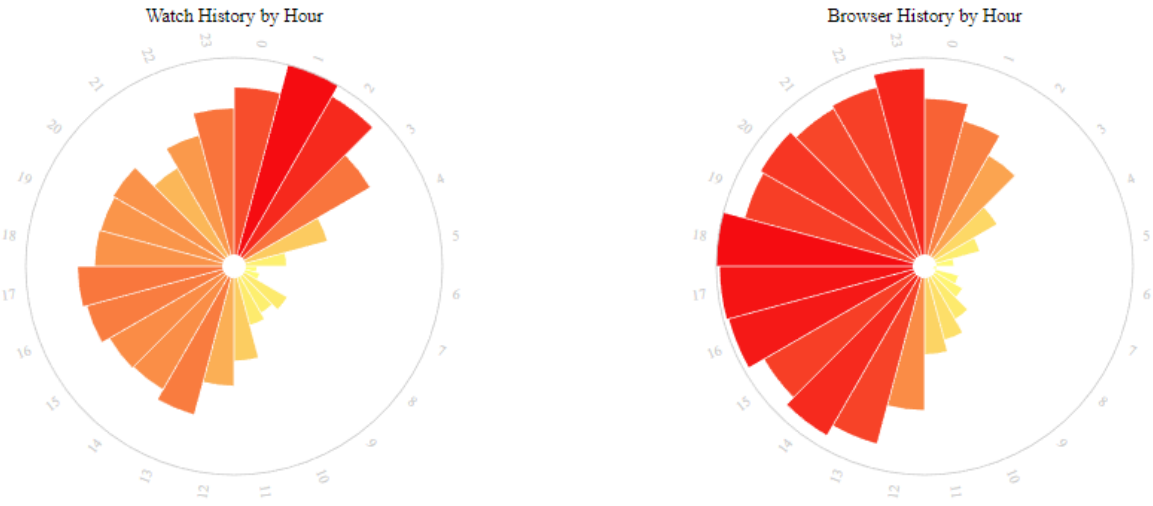

Figure 2: *Watch history and Browser history by hour in alternative circular bar graph.*

© 2022 The Author(s)

References

[AHMFF14] AL-HAJRI A., MILLER G., FONG M., FELS S. S.: Visualization of personal history for video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, Association for Computing Machinery, p. 1187–1196. doi:10.1145/2556288.2557106. 2

[BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2301–2309. doi:10.1109/TVCG.2011.185. 1, 2

[Cht12] CHTIVELBAND I.: *HistoryLane : Web Browser History Visualization Method*. Master's thesis, Blekinge Institute of Technology, 2012. URL: https://www.diva-portal.org/smash/get/diva2:833178/FULLTEXT01.pdf. 2

[CKM17] CARRASCO M., KOH E., MALIK S.: Pophistory: Animated visualization of personal web browsing history. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI EA '17, Association for Computing Machinery, p. 2429–2436. doi:10.1145/3027063.3053259. 2

[MT16] MENCHEN-TREVINO E.: Web historian: Enabling multi-method and independent research with real-world web browsing history data. In *iConference 2016 Proceedings* (2016), iSchools. doi:10.9776/16611. 1

[Shn92] SHNEIDERMAN B.: Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph. 11*, 1 (1992), 92–99. doi:10.1145/102377.115768. 1

[SRKK21] SCHUFRIN M., REYNOLDS S. L., KUIJPER A., KOHLHAMMER J.: A visualization interface to improve the transparency of collected personal data on the internet. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2021), 1840–1849. doi:10.1109/TVCG.2020.3028946. 1

[SSS*21] SIMO H., SHULMAN H., SCHUFRIN M., REYNOLDS S. L., KOHLHAMMER J.: Privinfervis: Towards enhancing transparency over attribute inference in online social networks. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (2021), pp. 1–2. doi:10.1109/INFOCOMWKSHPS51825.2021.9484595. 1

[TBC13] THUDT A., BAUR D., CARPENDALE S.: Visits: A Spatiotemporal Visualization of Location Histories. In *EuroVis - Short Papers* (2013), Hlawitschka M., Weinkauf T., (Eds.), The Eurographics Association, pp. 79–83. doi:10.2312/PE.EuroVisShort.EuroVisShort2013.079-083. 1

[TBHC16] THUDT A., BAUR D., HURON S., CARPENDALE S.: Visual mementos: Reflecting memories with personal data. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 369–378. doi:10.1109/TVCG.2015.2467831. 1

[WY20] WANG, YIXUE, YAO, SIYU: *STUDY ON INTENTION-AWARE RECOMMENDATION OF YOUTUBE VIDEOS*. Master's thesis, Cornell University, 2020. doi:10.7298/TAHM-B673. 2

© 2022 The Author(s)
Eurographics Proceedings © 2022 The Eurographics Association.