Web Scraping: A Useful Tool to Broaden and Extend Psychological Research

Speckmann, Felix

Web Scraping

A Useful Tool to Broaden and Extend Psychological Research

Felix Speckmann

Social Cognition Center Cologne, University of Cologne, Germany

Abstract: When people use the Internet, they leave traces of their activities: blog posts, comments, articles, social media posts, etc. These traces represent behavior that psychologists can analyze. A method that makes downloading those sometimes very large datasets feasible is web scraping, which involves writing a program to automatically download specific parts of a website. The obtained data can be used to exploratorily generate new hypotheses, test existing ones, or extend existing research. The present Research Spotlight explains web scraping and discusses the possibilities, limitations as well as ethical and legal challenges associated with the approach.

Keywords: web scraping, Big Data, internet research, internet research ethics

Psychologists are interested in observing and explaining human behavior. However, much psychological research happens in artificial laboratory settings, which on the one hand, leads to high internal validity and allows the testing of causal hypotheses. On the other hand, however, this approach cannot answer questions about the real-world implications of those findings. Those questions could be answered with field studies, yet psychologists conduct them less and less frequently (e.g., Sassenberg & Ditrich, 2019), a trend that will grow even stronger due to the COVID-19 pandemic. One way to ameliorate the tension between the laboratory and real-world is by collecting and using online data.

When people use the internet, they leave digital footprints of their activities (Girardin et al., 2008), sometimes actively (e.g., posts on social media, message boards, blog posts, etc.) and sometimes passively (e.g., cookies from visiting websites or entering search terms). As a result, the worldwide web contains a broad range of data concerning people's online behavior. For example, the blogging platform WordPress recorded 70.5 million new posts and 52.1 million comments per month in 2020 (Galov, 2020). Researchers can take advantage of this rich data source via a variety of methods: directly downloading complete datasets (e.g., downloading COVID-19 data from the Johns Hopkins University data repository), using application programming interfaces (APIs) to download specific data points (e.g., using the Wikipedia API to compare the number of changes between specific article versions), or by using web scraping (Singrodia et al., 2019) to access web site data directly (e.g., downloading displayed text from blogs). In the following, I will focus primarily on web scraping, as well as its capabilities and limitations, because it is an accessible and inexpensive means to access web data.

Web Scraping Examples and Challenges

Due to the multi-faceted nature of web data, research using web data is also diverse and comes from many different fields. Previous research using this method has investigated crowdfunding behavior (Agrawal et al., 2015), how to recognize depression from Twitter activity (Tsugawa et al., 2015), how to predict personal attributes and traits from Facebook likes (Kosinski et al., 2013), eating disorders (Moessner et al., 2018), political preferences (Ceron et al., 2014), research-methods blogs in psychology (Nicolas et al., 2019), and motivationally relevant key beliefs (Gordoni et al., 2019), among other topics.

Many researchers have previously suggested harnessing web data (e.g., Adjerid & Kelley, 2018; Chen & Wojcik, 2016; Heng et al., 2018; Kosinski et al., 2016), but it seems that psychology has not yet fully embraced this data source. One reason might be that for actual big data datasets, analyses often involve machine learning ("algorithms that allow computers to learn"; Segaran, 2007), which is rarely taught in psychology. Another possible reason is the heavy focus on laboratory research in many disciplines of psychology. Controlled experiments are typically required for causal claims in psychology, and web data generally does not fulfill that requirement. Moreover, web data bears the risks of HARKing (hypothesizing after results are known; Kerr, 1998) if analyzed without a specific hypothesis in mind (Landers et al., 2016) and the risk of selective reporting due to the high amount of potentially interesting variables (Adjerid & Kelley, 2018). Thus, it might appear difficult to integrate web data into one's own projects. However, web

https://doi.org/10.1027/2151-2604/a000470

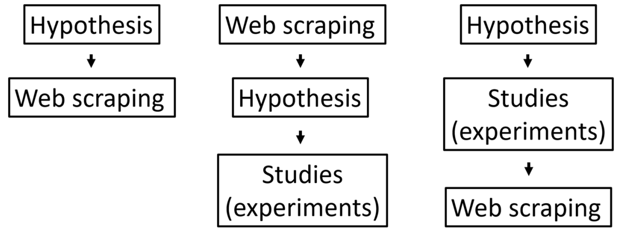

Figure 1. Three different approaches to integrating web scraping into psychological research projects.

data is not limited to enormous datasets and can potentially be analyzed using more traditional approaches such as linear regression. Keeping the aforementioned pitfalls in mind, psychological research projects can greatly profit from web scraping.

I propose categorizing web scraping approaches in psychological research into three groups, as shown in Figure 1. The first approach involves forming a hypothesis and identifying a source of web data to test the hypothesis, focusing exclusively on web data and foregoing experiments. In this case, web scraping is the only method of data acquisition. The second approach involves exploring web data with web scraping to generate hypotheses that can be (experimentally) tested in follow-up studies. The third approach starts with postulating a hypothesis, then testing the hypothesis (experimentally) using traditional data collection to subsequently corroborate the finding with web data acquired through web-scraping, extending the findings, and providing real-world context and/or applications. The web scraping procedure differs between these approaches insofar as selecting a suitable dataset depends on the sequence of steps involved. When using web scraping exploratorily (first step), any interesting dataset can be selected, whereas web scraping after hypothesis generation (second or third step) restricts the selection of suitable datasets for testing the hypothesis. Furthermore, approaches 1 and 3 benefit from preregistration. Strictly speaking, there is no proof from a reader's perspective that web-scraping data collection followed hypothesis generation, as much web data (e.g., from archives) is readily available without time stamps. However, from a writer's perspective, planning the analyses, deciding on the hypotheses, and choosing what data to use before looking at any data can facilitate a structured research workflow and help prevent HARKing. As there are good tutorials available for learning the actual coding involved in web scraping (e.g., for R: Bradley & James, 2019), I instead want to illustrate the workflow of how one might integrate web scraping into one's own projects and the concrete steps involved, following the first approach (see Figure 1) outlined above, using a straightforward example.

Predicting Politicians' Liking From Their Frequency of Occurrence in Online News (Showcase)

In this project, we examine the relationship between the frequency of mentions of politicians in online news and their popularity. Similar to the phrase "There is no such thing as bad publicity", it could be that higher media presence generally increases the liking of the people depicted, regardless of content or valence. This effect would be similar to a mere exposure effect (Zajonc, 1968) in that simply the act of repeatedly mentioning politicians increases liking towards them. Moreover, if the effect exists, we are interested in what function (linear, polynomial) best describes the relationship between exposure and liking. To test these hypotheses, we decided to web scrape online news from previous years and to download a publicly available dataset about voting behavior (see Forschungsgruppe Wahlen, Mannheim, 2020).

To obtain the relevant online news data, we used web scraping to download all articles from Spiegel Online (now only: "Spiegel", one of the most important German online news websites) from 2006 to 2019 using R (R Core Team, 2020), RStudio (RStudio Team, 2020) and the R package "rvest" (Wickham, 2020). In the next paragraph, I will explain our procedure for scraping the data from 2006. I will indicate the corresponding lines in our code for each step, which is available on the Open Science Framework (OSF; at https://osf.io/6gqkf/).

Scraping Online News From *Spiegel Online*

Generally, the best procedure is often to start small and expand from there when using web scraping. For example, rather than trying to scrape a whole year of news articles at once, it makes sense to first write code to scrape a single news article and then expand that code to include a whole day, a month, and a year of articles, which is what we did in our project. Concretely, we first wrote code to scrape a single news article. After loading all required libraries (ll. 2–6), we used the "rvest" package to download the HTML file of a single article (l. 12). This file contained all relevant text we wanted to extract, but we still had to specify which parts we wanted to scrape and how to extract them. There are several ways to target only specific parts of an HTML document, of which we used the following two. First, for the headline and subtitle, we specifically addressed the part of the document containing the headline or subtitle (using Xpath, see ll. 14–17 and ll. 19–22). Second, for the introduction and the main article, we specified certain conditions that the text had to meet to be targeted. For example, we only considered text that the website creator marked with a certain ID (l. 29) and that was also formatted in a specific style (CSS selector, see l. 33). By combining multiple conditions, it is thus possible to target very specific parts of a large amount of text and exclude other parts such as advertisements and navigation bars. After cleaning and combining all extracted text (ll. 25–84), we had successfully obtained all relevant parts of the article: headline, subtitle, introduction, and main article.

After finishing the code to download a single article, we wanted to apply the same code to download all articles published in 2006. To accomplish this, we made use of the structure of the *Spiegel Online* news archive. All articles published on a given day can be accessed via the archive page for that day. The URL of this page contains the day, month, and year (e.g., https://www.spiegel.de/nachrichtenarchiv/artikel-01.01.2006.html). By using the "rvest" package to filter all downloaded text for URLs leading to news articles, we created a list of URLs for all articles published on a single day (ll. 125–130). Applying our previous code to this list provided us with all articles of a single day. However, to obtain a whole year's worth of articles, we needed to go one step further. As previously mentioned, the URLs for archive pages all contain their respective dates. We replaced the numbers for day and month with the variables "*j*" and "*i*" [i.e., "https://www.spiegel.de/nachrichtenarchiv/artikel-j.i.2006.html" and used a loop function to access all archive pages one after another (ll. 102–113)]. Concretely, the loop function replaces the variables with numbers (e.g., $j = 3$ and $i = 2$ for March 2nd) and tries to access the resulting URL. By setting i to all numbers from 1 to 12 and j to all numbers from 1 to 31 (depending on the month), we accessed the archive pages for all days of the year 2006 and downloaded all corresponding articles.

The resulting data contained the website name, date of the article, headline, subtitle, abstract, and full article for 41,024 news articles published on *Spiegel Online* in 2006. By repeating the above process for other years and matching the resulting datasets with the voting behavior data, we formed a dataset that could provide insight into the relationship between politicians' exposure to online news and people's liking for them. For example, one could examine the correlation between an individual's popularity and the number of times they were mentioned.

Limitations, Legal and Ethical Considerations

Of course, web scraping has its limitations. Because psychological research usually focuses on the individual, but web data comes in many different shapes, it can be difficult to combine the two. When analyzing how users post on message boards, for example, it may be difficult to acquire further data (such as demographics or personality) about each user to analyze jointly with their messages. Even though it may be difficult to capture different variables from the same individual, it is still possible (e.g., see Kosinski et al., 2015; Matz, 2021, for research using Facebook). Another problem with collecting data from individuals is that this data is usually not from a representative sample. For example, Twitter's US user base is younger, more educated, and more likely Democrat than the general US population (Wojcik & Hughes, 2019). Furthermore, it is often unclear whether online content is created by human beings or computers. For example, product reviews or social media posts might be written by bots, which is something to keep in mind when collecting these types of data.

Apart from these technical challenges, web scraping also comes with additional ethical and legal challenges. Websites often explicitly forbid the automated screening of their content in terms of service (e.g., Facebook), but even if they do not, it is often debatable if certain content is ethically permissible to scrape (for a recent and comprehensive overview, see Franzke et al., 2020). In psychological research, participants usually give their explicit informed consent before any data is recorded. When people leave digital traces on the internet, researchers typically do not explicitly consent to use these data. This poses an ethical dilemma (Franzke et al., 2020) which can only fully be solved by acquiring people's consent before collecting their data. If that option is not possible, one should refrain from scraping identifying content and/or pseudonymizing all data. Furthermore, to avoid copyright and privacy issues, my recommendation is only to upload web scraping code, but never scraping data directly to data repositories. Due to national differences in laws and ethics policies, web scraping projects should also be cleared through the University's research ethics processes and other relevant research governance departments.

Conclusion

In summary, web scraping is an inexpensive and easily accessible tool that can be used in psychological research to exploratively generate hypotheses, test hypotheses, or extend existing (laboratory) findings. In this article, I suggested three ways in which web scraping may be included in psychological research and showcased the workflow of an example project. Although there are some constraints as well as ethical and legal challenges to keep in mind, in the future, hopefully, more projects will make use of this promising method.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899–917. https://doi.org/10.1037/amp0000190

Agrawal, A., Catalini, C., & Goldfarb, A. (2015). Crowdfunding: Geography, social networks, and the timing of investment decisions. *Journal of Economics & Management Strategy, 24*(2), 253–274. https://doi.org/10.1111/jems.12093

Bradley, A., & James, R. J. E. (2019). Web scraping using R. *Advances in Methods and Practices in Psychological Science, 2*(3), 264–270. https://doi.org/10.1177/2515245919859535

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16*(2), 340–358. https://doi.org/10.1177/1461444813480466

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods, 21*(4), 458–474. https://doi.org/10.1037/met0000111

Der Spiegel. (2021, November 22). *Nachrichtenarchiv. Sonntag, 1. Januar 2006*. https://www.spiegel.de/nachrichtenarchiv/artikel-01.01.2006.html

Franzke, A. S., Bechmann, A., Zimmer, M. Ess, C., & The Association of Internet Research. (2020). *Internet research: Ethical guidelines 3.0*. https://aoir.org/reports/ethics3.pdf

Forschungsgruppe Wahlen, Mannheim. (2020). *Partial cumulation of Politbarometers 1977–2019* (Version 12.0.0) [Dataset]. GESIS Data Archive. https://doi.org/10.4232/1.13631

Galov, N. (2020). *35+ WordPress statistics for the budding webmaster* [Infographic]. HostingTribunal. https://hostingtribunal.com/blog/wordpress-statistics/

Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing, 7*(4), 36–43. https://doi.org/10.1109/MPRV.2008.71

Gordoni, G., Steinmetz, H., & Schmidt, P. (2019). *Usability of web scraping of open-source discussions for identifying key beliefs*. https://doi.org/10.23668/psycharchives.2469

Heng, Y. T., Wagner, D. T., Barnes, C. M., & Guarana, C. L. (2018). Archival research: Expanding the methodological toolkit in social psychology. *Journal of Experimental Social Psychology, 78*, 14–22. https://doi.org/10.1016/j.jesp.2018.04.012

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70*(6), 543–556. https://doi.org/10.1037/a0039210

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods, 21*(4), 493–506. https://doi.org/10.1037/met0000105

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods, 21*(4), 475–492. https://doi.org/10.1037/met0000081

Matz, S. C. (2021). Personal echo chambers: Openness-to-experience is linked to higher levels of psychological interest diversity in large-scale behavioral data. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspp0000324

Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders, 51*(7), 656–667. https://doi.org/10.1002/eat.22878

Nicolas, G., Bai, X., & Fiske, S. T. (2019). Exploring research-methods blogs in psychology: Who posts what about whom, and with what effect? *Perspectives on Psychological Science, 14*(4), 691–704. https://doi.org/10.1177/1745691619835216

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2). https://www.r-project.org/

RStudio Team. (2020). *RStudio: Integrated development environment for R* (Version 1.3.1093). http://www.rstudio.com/

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science, 2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Segaran, T. (2007). *Programming collective intelligence: Building smart web 2.0 applications*. O'Reilly Media.

Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scrapping and its applications. *2019 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6. https://doi.org/10.1109/ICCCI.2019.8821809

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). Recognizing depression from Twitter activity. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, *2015-April*, 3187–3196. https://doi.org/10.1145/2702123.2702280

Wickham, H. (2020). *rvest: Easily harvest (scrape) web pages* (Version 0.3.6). https://cran.r-project.org/package=rvest

Wojcik, S., & Hughes, A. (2019). *Sizing up Twitter users*. Pew Research Center. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2, Pt. 2), 1–27. https://doi.org/10.1037/h0025848

History

Received January 29, 2021
Revision received May 26, 2021
Accepted July 28, 2021
Published online December 17, 2021

Open Data

To test our hypotheses, we decided to web scrape online news from previous years and to download a publicly available dataset about voting behavior (Forschungsgruppe Wahlen, Mannheim, 2020). This is accessible at the Open Science Framework (OSF; https://osf.io/6gqkf/).

Funding

Open access publication enabled by the University of Cologne.

Felix Speckmann
Social Cognition Center Cologne
University of Cologne
Richard-Strauss-Str. 2
50931 Cologne
Germany
felix.speckmann@uni-koeln.de