Web Scraping

An Emerging Data Collection Method for Criminal Justice Researchers

Erin J. Farley, Ph.D. & Lisa Pierotte, B.S.

Introduction

With the continual advancement of computer technology and the proliferation of the Internet, the amount of criminal justice-related information being placed on-line has dramatically increased over the last decade. As a result, public access to certain types of criminal justice data and statistical information on the Internet has rapidly expanded, presenting new and fundamentally different data access opportunities for criminal justice researchers. One method researchers are using to harness these new data access opportunities is web scraping.

Web scraping is essentially an automated tool for searching and extracting data from websites and other on-line sources. Pioneered in the fields of data science and e-commerce, web scraping provides a user with an automated way to find and collect data of interest from on-line sources that is more efficient and economical than techniques traditionally used in the past, and it arguably holds great promise for researchers working in the criminal justice community (Levy, 2017).

This brief is intended to: introduce criminal justice researchers to web scraping and explain what web scraping is and how it works; provide examples of how web scraping has been used in criminal justice research; and describe several issues one should be aware of if thinking about using this type of data collection method for criminal justice research purposes.

What is Web Scraping?

Web scraping is an automated tool for finding and extracting data from on-line sources. It utilizes computer programming

software and customized software code to mine data or other information from on-line sources in order to remove a copy of the data and store it in an external database for analysis. Typically, the data harvested through web scraping is analyzed to answer questions that could not be answered, or answered efficiently, using the data as it was originally presented on-line. Essentially, web scraping is a way to pull information from particular web pages and re-purpose it for customized analysis (Marres & Weltevrede, 2013).

Web scraping is also referred to as automated data collection, web extracting, web crawling, or web content mining. Web scraping has arguably been around since the inception of the World Wide Web, but it has primarily been utilized in the field of data science and is commonly associated with e-commerce (Marres & Weltevrede, 2013). Indeed, a form of web scraping is often used by travel-related websites readers may be familiar with, specifically those that allow consumers to compare prices for airline tickets or hotel rooms offered by different companies. In the past decade, however, the use of web scraping has emerged in several other fields including journalism, marketing, policy analysis, and psychology research (Baker & Yacef, 2009; Marres & Weltevrede, 2013; Youyou, Kosinski, & Stillwell, 2015)

How Does Web Scraping Work?

Web scraping involves the development and use of two customized software programs – a crawler and a scraper. The crawler systematically downloads data from the Internet; then the scraper systematically pulls the relevant information (unstructured, semi-structured, or structured) from the downloaded data, codes it, and relocates it in a database or file based on a pre-determined structure and format defined by the user. This new external database or file – populated with data originally presented on-line – is subsequently analyzed in ways the original on-line presentation of data did not support.

Common software programming languages like R and Python are typically used to write the software code for both the crawler and the scraper. Hence, software programming skills are essential for building and deploying a web scraper. The software code, however, is constructed based on specific search and data extraction criteria established by the researcher based on his/her understanding of the on-line data source(s) of interest and the research questions the analysis will attempt to answer. In

practice, a data source theory, developed by the researcher, guides the programmer's development of the crawler and scraper. This theory describes the researcher's and programmer's assumptions about the information source and its content, as well as their understanding of how the available data is maintained and how key measures are operationalized.

Web Scraping as a Criminal Justice Research Tool

The use of web scraping by criminal justice researchers is a relatively new phenomenon. In a search of the literature for criminal justice-related research employing web scraping as a data collection tool, only a handful of studies were found in which web scraping was utilized.

One of these studies was conducted by the Urban Institute (2017) as part of a larger exploration of how criminal background checks by employers may create barriers to employment among residents of the District of Columbia (D.C.). Background checks are utilized by potential employers, in D.C. and around the nation, to screen job applicants and to identify those with a criminal record. Having a criminal history record, however, does not necessarily mean an individual has been convicted of a crime. While a criminal history record is generated when someone is arrested, an arrest does not always result in a criminal charge; and a charge does not always result in a criminal conviction. Hence, it is possible for someone who has not been adjudicated to have engaged in criminal behavior to still have a criminal record, and this information can be, and sometimes is, used by employers to screen out job applicants, arguably unfairly limiting employment opportunities for D.C. residents with such records.

One of the key information needs in understanding the extent of this problem in DC requires determining what percentage of individuals with criminal records were and were not charged or convicted of a criminal offense. Researchers have attempted to answer this question in the past; but due to data fragmentation across law enforcement agencies and the courts, the ability to accurately answer this question for D.C. has been a challenge (Council for Court Excellence, 2011; Duane, Reimel, and Lynch, 2017).

According to the Urban Institute researchers, web scraping provided a viable way to overcome some of the existing data access and analysis issues

that resulted from this data fragmentation. Specifically, Urban Institute researchers used a web scraper to collect publicly available criminal history record data for Washington, D.C. residents over a 10-year period. These data were then used to estimate how many D.C. residents had a criminal record yet had not been convicted of a crime. The researchers determined that of the 68,000 D.C. residents who were flagged as having an arrest during the 10-year period examined, about half had not been convicted of a crime during that time span. This use of web scraping allowed Urban researchers to pull information off the web to produce more accurate estimates of the number of residents with criminal records who had not been convicted of a crime. This, in turn, better informed policy discussions regarding employment barriers for D.C. residents.

Another recent example of how web scraping has been used for criminal justice-related research involves the work being done by journalists from ProPublica Illinois, a non-profit news agency. In an article published in July 2017, David Eads describes ProPublica's efforts and ultimate failure to obtain certain information on the Cook County jail population from the Cook County Sheriff's Department through a Freedom of Information Act (FOIA) request. To overcome the data access obstacles encountered, Eads worked with computer programmers proficient in writing software code to create and deploy a web scraper for extracting publicly available data from the Cook County jail website (maintained by the Sheriff's department), including inmate names, their date of birth, and the location of the jail in which an inmate was held. The information extracted from the website using web scraping will be utilized as one part of a larger project aimed at tracking the flow of inmates through the entire criminal justice system in Illinois.

A third example comes from a National Institute of Justice-funded study currently in progress at JRSA. The study is exploring how the characteristics of various on-line advertisements for escorts, such as those posted on CraigsList and other on-line sources, can potentially be used to identify human trafficking cases. The objective of this project is to utilize the information pulled from websites (as well as from other sources like interviews) to create a profile of escort ads highly correlated with human trafficking, thereby providing law enforcement officers and prosecutors with practical guidance to more efficiently and effectively target escort ads, thereby leading to the successful prosecution of human traffickers.

As part of this project, researchers are relying upon a pre-existing, large-scale web scraping tool known as Memex. Launched by the U.S. Department of Defense in 2015, Memex searches on-line escort ads and extracts information of interest on a daily basis. Since its inception, the Memex Program[1] has pulled billions of ads off the internet to keep law enforcement informed about trends in online sex exploitation as well as to assist with anti-trafficking investigations (Sneed, 2015).

A final example involves research on bullying being conducted at Simon Fraser University. As part of a larger effort to explore environmental factors associated with bullying events, researchers utilized a web scraper to pull messages from a range of stakeholders (e.g., victims, parents, teachers, and bullies) from different countries[2] off of an international bullying website. This data collection contributed to an analysis that found bullying behavior most often took place in public settings where capable guardianship should be present, but even when potential guardians were present, their impact on bullying behavior was limited (Lam, Towle, & Cartwright, 2017).

[1] The Memex Program was created as part of the Defense Advanced Research Projects Agency.
[2] Including but not limited to: Canada, the United States, England, and Australia.

Web Scraping Issues to Consider

While the use of web scraping for criminal justice research is indeed in its infancy, the technology arguably has the potential to provide criminal justice researchers with an important new data collection tool. Given the proliferation in the amount of data being placed on-line, web scraping may serve as a viable alternative to traditional methods for accessing data through the Internet and conducting analyses that help answer important research questions.

Time Saver

Efficiency is another potential benefit of web scraping. Manually collecting data from on-line sources typically is time-consuming and labor intensive. As an automated process, web scraping can save time and reduce labor costs. Rather than reviewing websites and then manually copying and pasting relevant information from a website into another document or file for cleaning and analysis, a web scraper essentially automates these tasks, reducing the time and labor necessary to collect the information and prepare it for analysis.

However, as a novel and relatively untested approach to data collection for criminal justice research purposes, there are several issues anyone contemplating the use of web scraping should be aware of as they consider the pros, cons and feasibility of using this emerging technology.

Software Programming and Coding Skills Are Needed

While web scraping is typically used as a data collection tool to support research and analysis, developing and deploying a web scraper requires technical skills that social science researchers typically do not possess. A high level of proficiency in writing software code in computer programming languages such as Python or R is a prerequisite for developing a web scraper. Hence, a criminal justice researcher will often need to collaborate with a competent programmer or outsource the development of the web scraper, thereby incurring financial costs.

Data Quality and Interpretation

Researchers employing web scraping technology as a data collection tool also need to be concerned with the quality of the information pulled from the website and, in turn, its accurate interpretation. Web scraping was not originally created for social science research; as a result, researchers who utilize this technique may be introducing 'alien' assumptions into their research process (Marres & Weltevrede, 2013). Since web scraping does not typically involve direct communication between the researcher and those who originally collected the data and placed it on-line, data interpretation problems can easily emerge, and it can be difficult for a researcher to properly understand or verify the validity and reliability of the data. If researchers are not cognizant of these issues, they risk making inaccurate assumptions and reaching invalid conclusions.

Legal Constraints

There are also potential legal constraints for those who undertake data collection through web scraping. Indeed, an organization placing data on-line may expressly prohibit web scraping of their site, or deny access to "robots," web scrapers and other types of automatic data harvesters. These prohibitions may be, but are not necessarily, stated in the site's 'Terms of Service,' 'Terms and Conditions,' or 'Terms of Use.' Websites may also employ the use of a computer program or system such as CAPTCHA to distinguish between human and automated website users and to prevent automated data extraction (Studdenberg, 2017). Hence, researchers should

contact the organization or agency from which the website information is being drawn to ensure that web scraping is permitted. This will support transparency in the research project, and it may provide the source agency with an opportunity to provide the data of interest to the researcher through more traditional and transparent means.

Another legal issue that can emerge in the context of web scraping relates to personal privacy. The collection of personal information from a website may potentially result in a violation of personal privacy rights - even if that information is publicly available (Levy, 2017). While the standards or risks as they relate to web scraping have yet to be clearly outlined by the courts, experts recommend avoiding the collection of personal information through web scraping when possible (Hussain, 2017).

While there are some legal decisions that can be referenced for guidance (e.g., *eBay v. Bidder's Edge*, 2000; *Facebook Inc. v. Power Ventures Inc*., 2012; *Ticketmaster Corp., et al. v. Tickets.com Inc.*, 2000), in general the standards (and legal consequences) regarding what websites or types of information can and cannot be scraped by researchers have not yet been clearly established.

Website Overload

A final issue researchers should consider is the impact web scraping may have on the functionality of a website, as some web scraping attempts have inadvertently overloaded and shut down a website. This happened with an early version of a web scraper developed by Eads and colleagues for ProPublica Illinois in 2014. That web scraper continuously ran on the Cook County jail website, overwhelming and eventually crashing it. As a result, public defenders and family members of jail inmates could not access the site to find information about their clients or family members (Eads, 2016). Consequently, there are ethical and potential legal ramifications to crashing a website that anyone utilizing web scraping should be concerned with.

Summary

Considering the amount of criminal justice related data and other information available online, web scraping arguably presents researchers with a valuable new tool for collecting data and answering research questions. Properly designed and implemented, a web scraper can help researchers overcome data access barriers, collect on-line

data more efficiently, and ultimately answer research questions that were unable to be answered through traditional data collection and analysis means. We hope this fact sheet will increase awareness of this novel methodology and prompt and advance discussions on the appropriateness of its use within criminal justice research.

References

Baker, R.S.J.D. & Yacef, M. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1, pp. 3-16.

Council for Center Excellence. 2011. *Unlocking Employment Opportunity for previously Incarcerated Persons in the District of Columbia*. Washington, DC: CCE.

Duane, M., Reimal, E., & Lynch, M. (2017). *Criminal Background Checks and Access to Jobs: A Case Study of Washington, DC*. Urban Institute; Washington, DC.

Eads, D. (2017, July 24). How (and Why) We're Collecting Cook County Jail Data. *ProPublica*. Retrieved from: https://www.propublica.org/nerds/how-and-why-collecting-cook-county-jail-data

eBay, Inc. v. Bidder's Edge Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

Facebook Inc., V. Power Ventures Inc., 844 F. Supp. 2d 1025 (E.D. Cal. 2012).

hiQ Labs Inc. v. LinkedIn Corporation, No. 3:17- CV-03301 (N.D. Cal. 2017).

Hussein, Z. (2017, January 10). *Web Scraping: Pitfalls and Proactive Best Practices.* [Blog post on The Foundary Law Group Blog]. Retrieved from: http://foundrylawgroup.com/web-scraping-pitfalls-best-practices/

Lam, V.C., Towle, K., & Cartwright, B. (2017). *Bullying and Audience Behaviour: Capable Guardianship and the Environmental Backdrop of School Bullying* (PowerPoint presentation). Presented at the 2017 Annual American Society of Criminology Conference.

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), pp. 475-492.

Levy, J. (2017, August 23). *If Scraping Public Data can be Considered Criminal, Innovative Research will Suffer*. [Blog post on the Urban Wire]. Retrieved from: https://www.urban.org/urban-wire/if-scraping-public-data-can-be-considered-criminal-innovative-research-will-suffer

Marres, N. & Weltervrede, E. (2013). Scrapping the Social? Issues in Live Social Research. *Journal of Cultural Economy*, 6(3), pp. 313-335.

Ticketmaster Corp., et al. v. Tickets.com, Inc., CV 99-7654 HLH (BQRx) (C.D. Cal. 2000).

Sneed, T. (2015, January 14). How Big Data Battles Human Trafficking. *US News & World Report*. Retrieved from: https://www.usnews.com/news/articles/2015/01/14/how-big-data-is-being-used-in-the-fight-against-human-trafficking.

Review and Future Visions. *Journal of Educational Data Mining*, 1, pp. 3-16.

Studdenberg, M. (2017, July 13). Web Scraping [webinar]. In *Justice Research and Statistics Webinar Series*. Retrieved from: http://www.jrsa.org/webinars/index.html#scraping.

Youyou, W., Kosinski, M. & Stillwell, D.(2015). *Computer-Based Personality Judgments are More Accurate Than Those Made by Humans.* Proceedings at the National Academy of Sciences of the United States of America, 112: 1036-1040. Retrieved from: http://dx.doi.org/10.1073/pnas.1418680112.