Legality and Ethics of Web Scraping

Emergent Research Forum (ERF)

Vlad Krotov
Murray State University
vkrotov@murraystate.edu

Leiser Silva
University of Houston
lsivla@uh.edu

Abstract

Automatic retrieval of data from the Web (often referred to as Web Scraping) for industry and academic research projects is becoming a common practice. A variety of tools and technologies have been developed to facilitate Web Scraping. Unfortunately, the legality and ethics of using these Web Scraping tools are often overlooked. This work in progress reviews legal literature together with Information Systems literature on ethics and privacy to identify broad areas of concern together with a list of specific questions that need to be addressed by researchers employing Web Scraping for data collection. Reflecting on these questions and concerns can potentially help the researchers decrease the likelihood of ethical and legal controversies in their work. Further research is needed to refine the intricacies of legal and ethical issues surrounding Web Scraping and devise strategies and tactics that researchers can use for addressing these issues.

Keywords

Big data, web data, web scraping, web crawling, law, ethics

Introduction

In the past, social scientists were struggling to find data for their research (Munzert et al. 2015). Today, the increasing digitalization and virtualization of social processes have resulted in zettabytes (billions of gigabytes) of data available on the World Wide Web (the Web) (Cisco Systems 2016). This data provides a granular and real-time representation of numerous processes, relationships, and interactions in the social space (Krotov and Tennyson 2018). Thus, these vast volumes of Web data present academic researchers with opportunities for data collection for the purpose of answering new and old research questions with rigor, precision, and timeliness and improving organizational performance (Constantiou and Kallinikos 2015). Practitioners can leverage Web data for developing a better understanding of their customers and formulating better strategies based on these findings (Ives et al. 2016).

Unfortunately, harnessing these vast volumes of Web data presents serious technical, legal, and ethical challenges. While there has been a proliferation in tools and technologies that can be used for Web Scraping (Munzert et al. 2015), legality and ethics of data collection from the Web are still a “grey area” (Snell and Menaldo 2016). While existing legal frameworks can be applied, to some extent, to the emerging practice of Web Scraping, the ethical issues surround Web scraping have largely been ignored. This work-in-progress reviews the legal literature together with Information Systems literature on ethics and privacy to identify a preliminary set of ethical and legal considerations together with specific questions that need to be addressed when collecting data from the Web using automated tools. Compliance with these legal and ethical requirements can help industry and academic researchers decrease the likelihood of legal problems and ethical controversies in their work and, overall, foster research relying on Web data.

Literature Review

Big Web Data

The data available on the Web is comprised of structured, semi-structured, and unstructured quantitative and qualitative data distributed in the form of Web pages, HTML tables, Web databases, emails, tweets, blog posts, photos, videos, etc. (Watson 2014). Harnessing Web data requires addressing a number of technical issues related to volume, variety, velocity, and veracity of data on the Web (Goes 2014).

First, the data on the Web is often characterized by vast *volume* measured in Zettabytes (billions of gigabytes) (Cisco Systems 2016). Second, these vast data repositories available on the Web come in a *variety* of formats and rely on a variety of technological and regulatory standards (Basoglu and White 2015). Third, the data on the Web is not static; it is generated with extreme *velocity*. The final characteristic of Big Data is its *veracity* (Goes 2014). Due to the open, voluntary, and often anonymous interactions on the Web, there is an inherent uncertainty associated with availability and quality of Web data. A researcher can never be completely sure whether the needed data is or will be available on the Web and whether this data is reliable enough to be used in research (IBM 2018).

Web Scraping

Given the volume, variety, velocity, and veracity of Big Data available on the Web, collection and organization of this data can hardly be done manually by individual researchers or even large research teams (Krotov and Tennyson 2018). Because of that, researchers often resort to various technologies and tools to automate some aspects of data collection and organization. This emerging practice of using technology for collecting data from the Web is often referred to as Web Scraping (Landers et al. 2016).

Web Scraping is defined here as using technology tools for automatic extraction and organization of data from the Web for the purpose of further analysis of this data (Krotov and Tennyson 2018). Web Scraping consists of the following main, intertwined phases: website analysis, website crawling, and data organization (see Figure 1). Website analysis requires examining the underlying structure of a website or a Web repository (e.g. an online database) for the purpose of understanding how the needed data is stored. This requires a basic understanding of the World Wide Web architecture; mark-up languages (e.g. HTML, CSS, XML, XBRL, etc.); and various Web databases (e.g. MySQL). Website crawling involves developing and running a script that automatically browses the website and retrieves the needed data. These crawling applications (or scripts) are often developed using such programming languages as R and Python. This has to do with the overall popularity of these languages in Data Science and availability libraries (e.g. "rvest" package in R or Beautiful Soup library in Python) for automatic crawling and parsing of Web data. After the necessary data is parsed from the selected Web repository, it needs to be cleaned, pre-processed, and organized in a way that enables further analysis of this data. Given the volume of data involved, a programmatic approach may also be necessary to save time. Many programming languages, such as R and Python, contain Natural Language Processing (NLP) libraries and data manipulation functions that are useful for cleaning and organizing data. Oftentimes, these three phases of Web Scraping cannot be fully automated and often require at least some degree of human involvement and supervision.

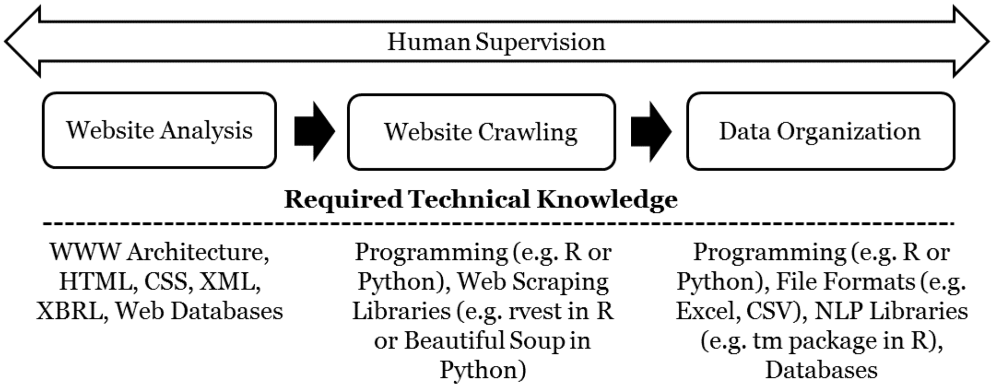
Human Supervision
Website Analysis
Website Crawling
Data Organization
Required Technical Knowledge
WWW Architecture, HTML, CSS, XML, XBRL, Web Databases
Programming (e.g. R or Python), Web Scraping Libraries (e.g. rvest in R or Beautiful Soup in Python)
Programming (e.g. R or Python), File Formats (e.g. Excel, CSV), NLP Libraries (e.g. tm package in R), Databases

Figure 1: Web Scraping (Adapted from Krotov and Tennyson 2018)

Legality of Web Scraping

While numerous tools and technologies have been available to assist researchers with Web Scraping (Munzert et al. 2015), the legality of Web Scraping is still a "grey area" in the legal field (Snell and Menaldo

2016). Here we define legality as compliance with applicable laws and legal theories. There is no legislature that addresses Web Scraping directly. As of now, Web Scraping is guided by a set of related, fundamental legal theories and laws, such as “copyright infringement”, “breach of contract”, the Computer Fraud and Abuse Act (CFAA), and “trespass to chattels” (Dreyer and Stockton 2013; Snell and Menaldo 2016). Some specific details of how these fundamental legal theories apply to Web Scraping are provided below.

Terms of Use

It has been often argued in the legal field that a website owner can effectively prevent programmatic access to a website by explicitly prohibiting this in the “terms of use” policy posted on the website. Failure to comply with these terms may lead to a “breach of contract” on the side of the website’s user (Dryer and Stockton 2013). To prosecute someone for violating the “terms of use”, the website user needs to enter an explicit agreement with the website owner to comply with the “terms of use” policy (e.g. by clicking on a checkbox). Thus, simply prohibiting Web Crawling and Web Scraping on the website may not preclude someone from crawling the website from a legal standpoint.

Copyrighted Material

Scraping and republishing data or information that is owned and explicitly copyrighted by the website owner can lead to a “copyright infringement” case (Dryer and Stockton 2013). However, a website does not necessarily own the data generated by its users. For example, a website devoted to product reviews does not necessarily own the reviews generated by the users of this website. Moreover, ideas cannot be copyrighted – only the specific form or representation of those ideas. So one can use copyrighted data to create summaries of copyrighted data. Finally, one can still use copyrighted material on a limited scale under the “fair use” principle.

Purpose of Web Scraping

Any illegal or fraudulent use of data obtained through Web Scraping is prohibited by several laws. For example, a person accessing data from the Web that is known to be confidential and protected may be prosecuted under the Computer Fraud and Abuse Act if the damage that occurred is greater than $5,000 (Dryer and Stockton 2013). On the Web, this often occurs when somebody knowingly accesses “premium content” and then resells it or continues accessing the content via an unauthorized channel after receiving a “cease and desist” letter from the owner of the website (Snell and Menaldo 2016).

Damage to the Website

If Web Scraping overloads or damages a website or a Web server, then the person responsible for the damage can be prosecuted under the “trespass to chattels” law (Dryer and Stockton 2013). However, the damage needs to be material and easy to prove in court in order for the owner of the Web server to be eligible for a financial compensation.

Ethics of Web Scraping

While existing laws and legal theories have been applied to Web Scraping in both courts and legal literature, the ethics of Web Scraping has not been addressed by prior literature. While there are many perspectives on ethics, for the purposes of this research project we view ethics as “a set of concepts and principles that guide us in determining what behavior helps or harms sentient creatures” (Paul and Elder 2006). In addition to violating existing laws, Web scraping can result in unintended harm to the “sentient creatures” that are associated with a particular website, such as the website’s owners or customers. These harmful consequences, are by definition, hard to predict (Light and McGrath 2010). Yet some possible harmful consequences of Web Scraping are discussed below.

Individual Privacy

A research project relying on data collected from a website may unintentionally compromise privacy of individuals participating in the activities afforded by the website (Mason 1986). For example, by matching the data collected from a website with other online and offline sources, a researcher can unintentionally reveal the identity of those who created the data (Ives and Krotov 2006). Even if individual privacy is not violated, the problem is that a website's customers may not have consented to any third party use of their data. Thus, using this data without a consent is a violation of the rights of research subjects (Buchanan 2017). These privacy and rights violations can lead to serious consequences for the website owner given the heightened concern with online privacy in the light of the recent privacy scandals involving such organizations and Facebook and Cambridge Analytica.

Organizational Privacy and Trade Secrets

Just like individuals have the right to privacy, organizations also have the right to maintain certain aspects of their operations confidential (Mason 1986). Automatic Web Scraping can unintentionally reveal trade secrets or simply confidential information about the organization who owns a website. For example, by automatically crawling and counting employment ads on an online recruitment website one can approximate the website's market share and revenues. It can also reveal some details and, possibly, flaws in the way the data is stored by the website (Ives and Krotov 2006). All this can damage the reputation of the company behind the website and lead to material financial losses.

Diminishing Value for the Organization

If one accesses the website omitting the Web interface made for humans, then the person will not be exposed to the advertisements that the website is using to monetize its content. Moreover, a Web Scraping project can lead to the creation of a data product (e.g. a report) that, without infringing on the copyright, makes it less likely for a customer to purchase a data product from the original owner of the data. In other words, the data product created with the help of Web Scraping, can directly or indirectly compete with the business of the website's owner (Hirschey 2014). All this may lead to financial losses to the owner of the website or, at the minimum, and unfair distribution of value from data ownership (Mason 1986).

Implications

Based on the literature presented in this paper, we generate a list of questions that need to be addressed in order to make a Web Scraping project legal (Dreyer and Stockton 2013; Hirschey 2014; Snell and Menaldo 2016) and ethical (Mason 1986; Ives and Krotov 2006; Buchanan 2017). These questions are as follows:

- Is Web Crawling or Web Scraping explicitly prohibited by the website's "terms of use" policy?
- Is the website's data explicitly copyrighted?
- Does the project involve illegal or fraudulent use of the data?
- Can crawling and scraping potentially cause material damage to the website or Web server hosting the website?
- Can the data obtained from the website compromise individual privacy?
- Can the data obtained from the website reveal confidential information about operations of the organizations providing data or the company owning the website?
- Can the project requiring the Web data potentially diminish the value of the service provided by the website?

A positive answer to any of these questions may suggest that the Web Scraping project can potentially result in lawsuits or ethical controversies. It may not be necessary to halt a research project involving a potential violation of one or more principles discussed in this paper. For example, copyrighted data can still be used in accordance with the "fair use" principle. Even if "terms of use" prohibit crawling, a permission for automatic collection of data can still be obtained from the website's owner. Still, researchers behind the projects involving a positive answer to any of these questions should reflect on how they will deal with a potential issue so that the legal and ethical requirements are still upheld.

Conclusion

The Big Data available on the Web presents researchers and practitioners with numerous opportunities. For researchers, these opportunities include leveraging this data for developing a more granular understanding of various old and new social phenomena in more timely fashion. Practitioners can leverage

Web data for developing a better understanding of their customers. But leveraging Big Data from the Web presents both researchers and practitioners with big challenges as well. Apart from the need to learn and deploy new tools and technologies capable of accommodating Big Data, researchers and practitioners intending to use Web Scraping in their research projects need to comply with a number of legal and ethical requirements. Unfortunately, due to the relative novelty of the Web Scraping phenomenon, legality and ethics of Web Scraping are still a "grey area". This work in progress is a preliminary attempt to reflect on some of the legal and ethical issues surrounding Web Scraping. A list of specific questions that need to be addressed by researchers employing Web Scraping is formulated in this paper. A negative answer to all these questions does not necessarily give a clearance to proceed with the research project. This list of questions should rather be used as a starting point for reflecting on the legality and ethics of a research project relying on Web Scraping for data acquisition. Further research aiming to refine, extend, and integrate various legal and ethical principles from which these questions are derived is needed. This research will have to adopt a multidisciplinary, socio-technical perspective. Web Scraping is truly a multidisciplinary phenomenon requiring the use of modern Big Data tools and technologies as well as knowledge of the principles of law and ethics.

REFERENCES

Basoglu, K. A., and White, Jr. C. E. 2015. "Inline XBRL versus XBRL for SEC Reporting," *Journal of Emerging Technologies in Accounting* (12:1), pp. 189-199.

Buchanan, E. 2017. "Internet Research Ethics: Twenty Years Later," in *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts,* M. Zimmer and K. Kinder-Kurlanda (eds.), Bern, Switzerland: Peter Lang International Academic Publishers, pp. xxix-xxxiii.

Dryer, A.J., and Stockton, J. 2013. "Internet 'Data Scraping': A Primer for Counseling Clients," *New York Law Journal*. Retrieved from https://www.law.com/newyorklawjournal/almID/1202610687621

Constantiou, I. D., and Kallinikos, J. 2015. "New Games, New Rules: Big Data and the Changing Context of Strategy," *Journal of Information Technology* (30:1), pp. 44-57.

Cisco Systems. 2016. "Cisco Visual Networking Index: Forecast and Methodology, 2014-2019," *White Paper*. Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

Goes, P. B. 2014. "Editor's Comments: Big Data and IS Research," *MIS Quarterly* (38:3), pp. iii-viii.

Hirschey, J. K. 2014. "Symbiotic Relationships: Pragmatic Acceptance of Data Scraping," *Berkeley Technology Law Journal* (29), pp. 897-927.

Ives, B., and Krotov, V. 2006. "Anything You Search Can Be Used Against You in a Court Of Law: Data Mining in Search Archives," *Communications of the Association for Information Systems* (18:1), pp. 593-611.

Ives, B., Palese, B., and Rodriguez, J. A. 2016. "Enhancing Customer Service through the Internet of Things and Digital Data Streams," *MIS Quarterly Executive* (15:4).

IBM. 2018. "The Four V's of Big Data," Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Krotov, V., and Tennyson, M. 2018. "Scraping Financial Data from the Web Using the R Language," *Journal of Emerging Technologies in Accounting*, Forthcoming

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., and 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.

Light, B., & McGrath, K. 2010. "Ethics and Social Networking Sites: A Disclosive Analysis of Facebook," *Information Technology & People* (23:44), pp. 290-311.

Mason, R. O. 1986. "Four Ethical Issues of the Information Age," *MIS Quarterly*, (10:1), pp. 5-12.

Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. 2015. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Chichester, UK: John Wiley & Sons, Ltd.

Paul, R., and Elder, L. 2006. The Thinker's Guide to Understanding the Foundations of Ethical Reasoning. Foundation for Critical Thinking.

Snell, J., and Menaldo, N. 2016. "Web Scraping in an Era of Big Data 2.0," *Bloomberg BNA*. Retrieved from https://www.bna.com/web-scraping-era-n57982073780/

Watson, H. J. 2014. "Tutorial: Big Data Analytics: Concepts, Technologies, and Applications," Communications of the Association for Information Systems (34:1), pp. 1247-1268.

View publication stats