2019 IEEE International Conference on Big Data (Big Data)

Web Scraping: State-of-the-Art and Areas of Application

Rabiyatou DIOUF
Université de Thies
Thies, SENEGAL
diouf.rabiyatou@ucao.edu.sn

Edouard Ngor SARR
UCAO-Saint Michel
Dakar-SENEGAL
edouard.sarr@ucao.edu.sn

Ousmane SALL
Université de Thies
Thies, SENEGAL
osall@univ-thies.sn

Babiga BIRREGAH
Université de Technologie de Troyes
Troyes, France
babiga.birregah@utt.fr

Mamadou BOUSSO
Université de Thies
Thies, SENEGAL
mbousso@univ-thies.sn

Sény Ndiaye MBAYE
Université de Thies
Thies, SENEGAL
senyr9@gmail.com

***Abstract*---Main objective of Web Scraping is to extract information from one or many websites and process it into simple structures such as spreadsheets, database or CSV file. However, in addition to be a very complicated task, Web Scraping is resource and time consuming, mainly when it is carried out manually. Previous studies have developed several automated solutions. The purpose of this article is to revisit the different existing Web Scraping approaches, categories, and tools, but also its areas of application.**

Keywords: Web-Scraping, Data Collection, Web Data Extraction.

I. INTRODUCTION

The web is a major source of information for many professionals in various sectors. It contains useful and useless, structured and non-structured information, in different formats, and from various sources. However, in addition to being a very complex activity, Web Scraping is a time- and resource-consuming task, especially when it is carried out manually. This complexity increase depending on data and collection websites. Many techniques [21] have been used to retrieve content from a web page: *Cut/Paste*, http, *Query languages for semi-structured Data* [22], *DOM* [23] or even *Web-Scraping* [25]. Many advanced techniques [21] are also used to collect data from the web. Among these, one can mention API computer languages [8, 10], robots, intelligent agents and Web Scraping. A web scraper is, therefore, a software that simulates human browsing on the web to collect detailed information data from different websites. The advantage of a scraper resides on its speed and its capacity to be automated and/or programmed. However, no matter what technique is used, the approach and the objectives remain the same: capture web data and present it in a more structured format [24]. This paper revisits existing frameworks for use, approaches, categories, and tools of Web Scraping through the identification of their strength and limits. The study has two main sections. Section 2 is an overview of Web Scraping. Section 3 presents different approaches and tools used in Web Scraping.

II. APPROACHES, CATEGORIES, AND TOOLS

A. The different approaches

a. Mimicry Approach

This category of scraper works thanks to predefined customised rules. The location of the data to be collected from a web page is preconfigured in the scraper. This mechanism is applied on DOM selectors [23] which are deduced from click-based leaning. This strategy is relatively efficient thanks to its neatness, but is less adapted when it comes to process multiple heterogeneous websites. Furthermore, if the source website modifies its graphic design, the engine should be reprogrammed to how to find the needed information. Tools such as Import.io [6] or Mozenda [14] use this approach.

b. Weight Measurement Approach

This approach is based on a generic algorithm which analyses the DOM tree [23] of a web page and measures the weight of words in each branch. Through a deduction, the algorithm chooses the node as a starting point of the main text and extracts the text from all the child nodes. The main advantage of this mechanism is that it does not require any training and can adapt itself to the graphic design changes of the source websites. However, the results are generally quite noisy.

c. Differential Approach

This approach is based on the fact that two pages from the same website will only differ in content from the body of the page. According to this logic, the menu bars, the right or left columns, and the footers are supposed to be perfectly identical between two pages of the same website. The mechanism formerly consists in applying a masking algorithm that superimposes the two pages by removing only the differences.

d. Machine Learning Approach

The general principle is to train an algorithm on a large sample of manually analyzed web pages. The machine learning [26] is based on geographical indicators of the text blocks on the page: statistical measure is done where the main text block is located compared to the other text blocks. The machine is then able to deduce by itself where the text is usually located. The larger the sampling, the more accurate the algorithm is.

A. Categories and Tools

a. Ready-made Tools

i. Browser extensions

- *Spider*: A free extension of Google Chrome. On the screen, each column represents a type of element which can be retrieved. You only need to click on the item to add it to a column. The result is available in JSON or CSV format [47]

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

- *Data Scraper* is a Google Chrome extension, which allows to extract data from a web page and to export it in CSV and/or XLS formats [44].
- *Agenty:* Chrome add-on that allows to extract data from a web page through the CSS class in a very fast and simple way [45].
- *Data Miner*: A Google Chrome extension that allows to extract data from web pages into a CSV file or Excel spreadsheet. More than 50,000 free predefined queries are available for more than 15,000 websites [46].
- *Cloump U-Scraper Plugin:* It is a Firefox extension. It converts web pages into API and connects them via Excel using a free Excel add-on. It supports the JSON schema and can process complex nested data structures [48].
- *OutWit Hub* is a powerful Firefox extension. It allows easy extraction of links, images, email addresses, data tables, etc. The extracted data can be exported to databases, CSV, HTML or Excel files, while images and documents are saved directly to your hard drive [49].
- *Dexi.io* : This Firefox extension can configure robots and recover data in real time. It supports data collection from any website [50].

ii. Software and Platforms

- *Import.io*: It is spreadsheet-based library of functions that allows the end-user to create customized formulas that can be used to enrich all data.
- *Easy Web Extract*: It is a scraper written with technology.NET with multi-format results (Excel CSV, text, XML, HTML ...). One of the limits of this tool is the time it takes to extract.
- *Web Info Extract*: A scraper that can store data in a database. Once set up, this scraper constantly monitors the web page and when new content is added to the page, depending on the change, the task assigned to the tool is updated.
- *Mozenda web*: A powerful SaaS or managed web data extraction service. It can extract data from websites and PDF.
- *Screen Scraper*: An advanced scraper that comes in three versions: Enterprise, Professional and Basic.
- *Web Data Extractor*: A web scraping tool specially designed for link extraction, Meta Tag, body text, emails, fax machines.
- *Web Content Extractor (WCE):* A simple, user-oriented scraper that is very good for putting data in different formats.
- *WebExtractor360*: An open source web scraper. It uses the standard pattern to scrape data from web pages.
- *Fminer* [16]: It is one of the best visual web scraping tools developed with Python. It has a nice schematic representation of the scraping flow and actions. It also allows to execute customized python code.
- *Weboob*: It is made up of a set of applications (QBooblyrics, QBoobMsg, QCineoob, QCookboob, QFlatBoob, QHandjoob, QHaveDate, QVideoob, QWebContentEdit, weboob-config-Qt) written in Python.
- *PySpider*: A web robot created with Python. It supports JavaScript pages and has a distributed architecture. One of the advantages of PySpider is its user-friendly interface. Data can be stored in JSON and CSV formats.

b- The libraries of programming languages

The second category of web scraping tools gathers libraries of programming languages PHP [8], Java [9] or Python [10] or NodJs. Python gathers Beautiful Soup, Newspaper [12], Lxml, etc. [12, 37, 38]. Java also offers some API web scraping such as Jsoup, Jaunt, StormCrawler, Norconex http, Collector, etc. As for NodJs, we have Cheerio and Apify [8, 9, 10, 12, 37, 38]. In addition to these basic libraries, there are many examples (Framework) of implementation allowing to quickly develop a scraper; *i.e* the remaining work of a scraper. In the two cases, these solutions are for specialists who are extracting some information or in Automatic Natural Language Processing (NLP) [11]. The first thing to note about these tools is that the majority of them are not made for the layman in that domain and they are not specialized for a specific domain. Thus, setting up these libraries for a specific domain often requires the implementation of an overlayer. Newspaper is designed for the extraction of news articles in a different language, English for instance [1].

III. AREAS OF APPLICATION

The web Scraping has a wide field of usage. It is used as a key element **in job search engines** such as Teambuilder [27]. In addition, many other **recommendation** collaborative **systems** use web scraping. As illustrations, we can note Grundy [30], GroupLens [31], Video Recommender [32], and Ringo [33] systems that first have used collaborative filtering algorithm to automate the recommendations. The recommendation system of a book from Amazon.com [34] and the PHOAKS which help internet-users find out relevant information on the Web [35] also use web scraping. In the **advertisement sector**, data collection is more based on collaborative filtering. In fact, the collaborative filtering consists of automatically establishing some forecasts on the interests of the user, by collecting similar preferences or tastes [26]. They intend to foretell the importance of elements for a particular user according to previous elements assessed by other users. Several systems of collaborative filtering based on web scraping have been developed to suggest articles and properties, notably topical news, photos and books [29]. Web scraping is also predominant in the **health sector**. Even if Web services are *de facto* considered as standards for the interoperability of bio-medical data, the SOAP protocol and REST are the two main approaches of implementation [18]. In fact, WhichGenes and PathJam, are two meta-servers that use web scraping as a means to face the analysis of the enrichment of a set of genes [17]. Tools such as Protein Information Crawler [51], DrugBank [52], ChemSpider [53], BioSpider [28], OReFil [20], and MEDPIE [19] are also based on data web scraping. There are also different examples of web scraping in the recent programs regarding the genetics' domain, molecular biology [15, 13], and general medicine [15]. In **journalism**, Web Scraping is one of the most useful methods but less understood by journalists. However, some Scrapers or tools based on web scraping such as ScraperWiki [18], bitLy [43], Blekko [42] are used to help journalists in their daily tasks. Indeed, ScraperWiki is a scraper of data developed in Python. BitLy helps grasp how users share a link for a good and widespread study. The search engine Blekko provides an incredible quantity of information thanks to internal statistics it collects while browsing through the Web. Except these five quoted sectors, it is important to know that the use of web scraping as a quick means of collecting data has become a necessity in all sectors. Indeed, in **journalism**, web scraping is becoming a vital element. Tools such as NewsPaper [12] and FactExtract [1] are used for the automatic extraction of journalistic articles. The NewsOne platform [40] extracts and gathers all news updates of many international resources. The

web scraping is also used in the search for disappeared persons [40] and accommodation [39], in the social domain [38], literature [36], marketing [37].

IV. Conclusion and Perspectives

This article present the state-of-the-art in Web Scraping. We have focused on the different approaches, categories and the web scraping tools. We have also dealt with the area of applications of web scraping. At the end of this study, we have noticed that Web scraping is more needed in one sector: journalism, though it remains the one having less specialized tools.

V. Acknowledgment

The authors express their kind regards to CEAMITIC for financing their research project Check4Decision.

References

[1]. SARR, E. N., Ousmane, SALL., & DIALLO, A. (2018, October). FactExtract: Automatic Collection and Aggregation of Articles and Journalistic Factual Claims from Online Newspaper. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 336-341). IEEE.
[2]. Rouby, A., & Tournier, T. Scraping & Crawling.
[3]. Piwowarski, B., Denoyer, L., & Gallinari, P. (2002). Un modèle pour la recherche d'information sur des documents structurés. Proceedings of the 6emes journées Internationales d'Analyse Statistique des Donnees Textuelles (JADT2002).
[4]. Haddaway, N. R. (2015). The use of web-scraping software in searching for grey literature. Grey J, 11(3), 186-90.
[5]. Webhose.io web page, January 2018, [online] Available: https://webhose.io/.
[6]. Import.io web page, January 2018, [online] Available: https://www.import.io/.
[7]. Hadi, A., & Al-Zewairi, M. (2017). Using IPython for Teaching Web Scraping. In Social Media Shaping e-Publishing and Academia (pp. 47-54). Springer, Cham.
[8]. MICHAEL SCHRENK. Webbots, Spiders, and Screen Scrapers. No Starch Press, 2007. Disponible sur http://edu.ercess.co.in/ebooks/php/Webbots,%20Spiders,%20and%20Screen%20Scrapers%2 0-%20A%20Guide%20to%20developing%20internet%20agents%20with%20PHP.pdf ISBN 978-593327-120-6, p407-416
[9]. SOUMEN CHAKRABARTI. Mining the web [online]. Bombay, India: Morgan Kaufmann Publishers, 2003. Disponible sur http://read.pudn.com/downloads75/ebook/275001/Morgan%20K aufmann%20-%20Mining% 20the%20Web%20-%20Discovering%20Knowledge%20from%20Hypertext%20Data.pdf
[10]. CHRIS HANRETTY. Scraping the web for arts and humanities [online]. Norwich, Royaume-Uni: University of East Anglia, 2013. Disponible sur http://www.essex.ac.uk/ldev/documents/going_digital/scraping_book.pdf
[11]. Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
[12]. Lucas Ou-Yang. Newspaper. https://github.com/codelucas/newspaper, 2017.
[13]. Enenbaum JD, Whetzel PL, Anderson K, et al. The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research, J Biomed Inform, 2011, vol. 44 (pg. 137-45).
[14]. Baskaran, U., & Ramanujam, K. (2018). Automated scraping of structured data records from health discussion forums using semantic analysis. Informatics in Medicine Unlocked, 10, 149-158.
[15]. Hill AW, Guralnick RP. Distributed systems and automated biodiversity informatics: genomic analysis and geographic visualization of disease evolution, Inf Knowl, 2008 (pg. 270-9).
[16]. Readability (2019). https://www.readability.com/
[17]. Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. Briefings in bioinformatics, 15(5), 788-797.
[18]. https://scraperwiki.com/ [Visited 2019-06-12]
[19]. Wall DP, Pivovarov R, Tong M, et al. Genotator: a disease-agnostic tool for genetic annotation of disease, BMC Med Genomics, 2010, vol. 3pg. 50.
[20]. Johnson S., Design & Implementation of a Pipeline for High-throughput Enzyme Function Prediction [PhD dissertation], 2006 Fairfax, Virginia George Mason University for the Degree of Master of Science Bioinformatics.
[21]. Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. Briefings in bioinformatics, 15(5), 788-797.
[22]. Boag, S., Chamberlin, D., Fernández, M. F., Florescu, D., Robie, J., Siméon, J., & Stefanescu, M. (2002). XQuery 1.0: An XML query language.
[23]. Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P. (2003, May). DOM-based content extraction of HTML documents. In Proceedings of the 12th international conference on World Wide Web (pp. 207-214). ACM.
[24]. SARR, E. N., Ousmane, S. A. L. L., & DIALLO, A. (2018, October). FactExtract: Automatic Collection and Aggregation of Articles and Journalistic Factual Claims from Online Newspaper. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 336-341). IEEE.
[25]. Sirisuriya, D. S. (2015). A comparative study on web scraping.
[26]. Vargiu, E.,& Urru, M.(2013).Exploiting web scraping in a collaborative filtering-based approach to web advertising.Artif.Intell.Research,2(1),44-54.
[27]. Karduck, A. (1994, May). TeamBuilder: a hyper-information environment for augmenting global collaborative work. In 5th IEEE COMSOC International Workshop on Multimedia Communications.
[28]. Johnson S., Design & Implementation of a Pipeline for High-throughput Enzyme Function Prediction [PhD dissertation], 2006Fairfax, Virginia George Mason University for the Degree of Master of Science Bioinformatics.
[29]. Adomavicius, G. & Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering. 2005; 17(6): 734-749. http://dx.doi.org/10.1109/TKDE.2005.99
[30]. Rich, E. User modeling via stereotypes, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.
[31]. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J. Grouplens: an open architecture for collaborative filtering of netnews. CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, New York, NY, USA. 1994; 175-186. http://dx.doi.org/10.1145/192844.192905
[32]. Hill, W., Stead, L., Rosenstein, M. & Furnas, G. Recommending and evaluating choices in a virtual community of use. CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA. 1995; 194-201.
[33]. Shardanand, U. & Maes, P. Social information filtering: algorithms for automating "word of mouth". CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1995; 210-217.
[34]. Linden, G., Smith, B. & York, J. Amazon.com recommendations. IEEE Internet Computing. 2003; 07(1): 76-80. http://dx.doi.org/10.1109/MIC.2003.1167344
[35]. Terveen, L., Hill, W., Amento, B., McDonald, D. & Creter, J. Phoaks: a system for sharing recommendations. Communication of ACM. 1997; 40(3): 59-62. http://dx.doi.org/10.1145/245108.245122
[36]. Haddaway, N. R. (2015). The use of web-scraping software in searching for grey literature. Grey J, 11(3), 186-90.
[37]. Aggrawal, N., Ahluwalia, A., Khurana, P., & Arora, A. (2017). Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media. Social Network Analysis and Mining, 7(1), 21.
[38]. Marres, N., & Weltevrede, E. (2013). Scraping the social? Issues in live social research. Journal of cultural economy, 6(3), 313-335.
[39]. Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: web scraping and analyzing craigslist rental listings. Journal of Planning Education and Research, 37(4), 457-476.
[40]. Sundaramoorthy, K., Durga, R., & Nagadarshini, S. (2017, April). NewsOne—An Aggregation System for News Using Web Scraping Method. In 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC) (pp. 136-140). IEEE.
[41]. https://searchengineland.com/library/blekko [Visited 2019-06-12]
[42]. https://searchengineland.com/library/blekko [Visited 2019-06-12]
[43]. https://bitly.com/ [Visited 2019-06-12]
[44]. https://www.data-miner.io [Visited 2019-06-12]
[45]. https://www.agenty.com/ [Visited 2019-06-12]
[46]. https://www.techopedia.com/definition/1464/data-miner [Visited 2019-06-12]
[47]. https://www.spyderproducts.com/toolpages/spyder-scraper/ [Visited 2019-06-12]
[48]. https://chrome.google.com/webstore/detail/cloump-u-scraper-plugin/kaaalbpgkcljgpbbmnnndchegnmfomjm [Visited 2019-06-12]
[49]. https://www.outwit.com/products/hub/ [Visited 2019-06-12]
[50]. https://dexi.io/product [Visited 2019-06-12]
[51]. Mayer U. Protein Information Crawler (PIC): extensive spidering of multiple protein information resources for large protein sets, Proteomics, 2008, vol. 8
[52]. Wishart DS, Knox C, GuoAC et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res, 2006, vol. 34 (pg. D668 -72).
[53]. Williams A, Using Text-Mining and Crowdsourced Curation to Build a Structure Centric Community for Chemists, 2008. http://www.slideshare.net/AntonyWilliams/using-textmining-and-crowdsourced-curation-to- build-a-structure-centric-community-for-chemists-presentation/ (22 January 2019, date last accessed.