International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at :www.ijsrcseit.com
doi : https://doi.org/10.32628/IJSRCSEIT

A Survey on the Web Scraping : In the Search of Data

Prof. K. N. Aaglave[1], Shivanjali Santosh Jadhav[2], Amaan Firoj Khatib[2], Rohini Laxman Khurangale[2]

[1]Assistant Professor, [2]Student

Department of Computer Engineering, S. B. Patil College of Engineering, Savitribai Phule Pune University

ARTICLEINFO

Article History:

Accepted: 10 Oct 2023

Published: 30 Oct2023

Publication Issue

Volume 9, Issue 10

September-October-2023

Page Number

191-195

ABSTRACT

AI-based methods and tools used for adjust themselves to scraping the data. Web scraping use machine learning and AI technologies. Data are not in structured formats and Difficulty of extracting relevant data from web pages.Need to classify web content in order to remove unwanted data.Difficulty of finding a suitable web scraping Need for a more flexible and extensible web scraping framework. Web scraping is a powerful tool for extracting data from the internet and has a wide range of applications across various industries. When done responsibly and legally, web scraping can provide valuable insights and data for businesses, researchers, and developers. It highlights Scrapy as a powerful web scraping tool, offering speed, extensibility, and efficient data extraction capabilities. Develop a more efficient and accurate way to extract and classify web content using AI and machine learning algorithms techniques.

Keywords - Machine Learning, Artificial Intelligence, Web Scraping, Data Processing, Data Extraction, Data Analysis, Ethical Concerns, Web Scraping Tools, Web Scraping Framework, User Interface.

I. INTRODUCTION

The Main domain for the project is to use Machine Learning, Which is sub category of Artificial Intelligence concepts and using python language develop a web site to scrap all data from any website automatically. Data collection methods differ depending on the subject or topic of study, the type of data sought, and the user's aims. Depending on the goals and conditions, the method's application methodology can also change without jeopardizing data integrity, correctness, or reliability[11]. There are numerous data sources on the Internet that might be employed in the design process. The technique of extracting data from websites is often know as webscraping, web extraction, web harvesting, web crawler. The purpose of Web mining is to look for models in Web data by gathering and analysing data to achieve insights. Web mining supports to increase the ability of web search engine by identifying web pages and classifying the web documents. Web mining can be divided web content mining, web structure mining and web usage mining based on information[5].

II. LITERATURE SURVEY

1. Web Scraping Techniques and Applications: A Literature Review, ChaimaaLotfi, Swetha Srinivasan, MyriamErtz, ImenLatrous, on2023, Big data analytics gives organizations a way to analyze huge data sets and gather new information. It helps answer basic questions about business operations and business performance. It also helps discover unknown patterns in vast datasets or combinations thereof. In the current data-driven world, it becomes increasingly essential that big data techniques are applied and analyzed for organizational growth[1].Access to Abundant Data, Business Performance Improvement, Research and Hypopthesis Testing, Efficient Data Collection.
2. Web Scraping for career analysis based on YouTube data APIs using web content mining, Khin Than Nyunt, NawThiriWaiKhin, on 2022, Nowadays, APIs is frequently used to refer to "web application APIs", despite the fact that there are distinct APIs for numerous different software programs. Typically, a programmer will use HTTP to request some kind of data from an API, and the API will then provide the requested data in the form of XML or JSON[2].Trending Content, Web Content Mining, Data Availability, Rapid Growth of YouTube.
3. Web Scraping Approaches and their Performance on Modern Websites, Ajay Sudhir, Naveen Ghorpade, Rohith S, S Kamalesh, Rohith R, Rohan B S on2022, When it comes to information, the internet is a gold mine. Whether you need data for your business, school, or personal use, you may uncover a wealth of information by performing an internet search. Web Scraping (WS) is a computerized method of obtaining big amounts of information from internet sites[3]. Data Extraction, Progressive Testing, Database Scraping, Ethical Considerations.
4. An industrial perspective on web scraping characteristics and open issues, Elisa Chiapponi, Marc Dacier, Olivier Thonnard, Mohamed Fangar, MattiasMattsson, Vincent Riga, on 2022.An ongoing battle has been running for more than a decade between e-commerce websites owners and web scrapers. Whenever one party finds a new technique to prevail, the other one comes up with a solution to defeat it. Based on our industrial experience, we know this problem is far from being solved[4].Qualifying Losses, E-commerce Market Significance, Adaptive Nature of Scrapers.
5. Survey Paper on Web Content Extraction & Classification, DipaliShete, Sachin Bojewar , AnkitSanghvi, on2021.Over the last few years, web data extraction has gained popularity. Product information on the Ecommerce website floods the internet with big data. Web-based business sites these days have gotten one of the most significant hotspots for getting a large amount of relevant data[5].Comparison of Techniques, Machine Learning Algorithms, Supervised Learning for Classification,Web Mining and Data Mining.
6. An Optimal Data Entry Method, Using Web Scraping and Text Recognition, Roopesh N, Akarsh M S, C. NarendraBabu, on2021.Data entry is one of the most tedious jobs which consumes huge manpower in creating structured data from the given inputs. A large amount of data entered in the system can be contrasting to the original data causing confusions, especially when the data has to be gathered from image files[6].Text Pre-processing, Comparison of OCR Tools, Versatile Output Formates, Applications in Chatbots and NLP.
7. Computer Vision-based Web Scraping for Internet Forums, Eric C. Dallmeier on 2021.With the amount of data available on websites the need to transform this data from a human-understandable format, the visual representation, to a computer-understandable format, e.g. as entries in a database, rises. The

approaches to solving web scraping that were published in the last two decades have the drawback that they all to a certain degree rely on the structure and existence of the underlying Hypertext Markup Language (HTML) or Cascading Style Sheets (CSS)[7].Web Scraping for Data Analysis, Computer Vision and Object Detection, Suitability for Discussion Forums, Data Selection and Training.

8. A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages, ERDINC¸ UZUN, on 2020. Web scraping is a process of extracting valuable and interesting text information from web pages. Most of the current studies targeting this task are mostly about automated web data extraction. In the extraction process, these studies first create a DOM tree and then access the necessary data through this tree. The construction process of this tree increases the time cost depending on the data structure of the DOM Tree[8].Improved Time Efficiency, Reduced Resource usage, Additional Information, Use of String Methods.
9. A Review on Web Scrapping and its Applications, VidhiSingrodia, AnirbanMitra on 2019.Internet grants a wide scope of facts and data source established by humans. Though, it shall consist of an enormous assortment of dissimilar and ailing organized data, challenging in collection in a physical means and problematical for its usage in mechanical processes. Since the recent past, procedures along-with various outfits have been developed to permit data gathering and alteration into organized information to be accomplished by B2C and B2B systems[9].Desktop-Based Tools, Data Variety, Automation.
10. Web Scraping: State-of-the-Art and Areas of Application, Rabiyatou DIOUF, EdouardNgor SARR, Ousmane SALL, Babiga BIRREGAH, Mamadou BOUSSO, SenyNdiaye ´ MBAYEon 2019.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[10].Efficient Data Collection, Application in Various Sectors, Wide Range of Tools.
11. Real-time applications like video conferencing and live streaming demand the reliable, effective transmission of high-quality image and video data. This model is a useful tool for these applications due to its performance in busy network environments [12].

III.LIMITATIONS OF EXISTING WORK

By the comparative study of the proposed system, we have been recognized following limitations of the system as:

- Ethical and Legal Concerns
- Rate Limiting and IP Blocking
- Resource Intensive
- Maintenance Overhead

IV.CONCLUSION

Web scraping is a valuable technique for extracting data from websites. It can be used for various purposes, such as gathering information for research, monitoring prices, or creating datasets. However, it's important to be aware of legal and ethical considerations when scraping websites, respect the website's terms of service, and avoid overloading their servers. Additionally, web scraping may require ongoing maintenance due to website changes. In conclusion, web scraping can be a powerful tool when used responsibly and ethically.

V. REFERENCES

[1]. Gaikwad, Yogesh J. "A Review on Self Learning based Methods for Real World Single Image Super Resolution." (2021).

[2]. Khetani, Y. Gandhi and R. R. Patil, "A Study on Different Sign Language Recognition Techniques," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 2021, pp. 1-4, doi: 10.1109/CCGE50943.2021.9776399.

[3]. Vaddadi, S., Arnepalli, P. R., Thatikonda, R., &Padthe, A. (2022). Effective malware detection approach based on deep learning in Cyber-Physical Systems. International Journal of Computer Science and Information Technology, 14(6), 01-12.

[4]. Thatikonda, R., Vaddadi, S.A., Arnepalli, P.R.R. et al. Securing biomedical databases based on fuzzy method through blockchain technology. Soft Comput (2023). https://doi.org/10.1007/s00500-023-08355-x

[5]. Rashmi, R. Patil, et al. "Rdpc: Secure cloud storage with deduplication technique." 2020 fourth international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC). IEEE, 2020.

[6]. Khetani, V., Gandhi, Y., Bhattacharya, S., Ajani, S. N., &Limkar, S. (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 253-262.

[7]. Khetani, V., Nicholas, J., Bongirwar, A., &Yeole, A. (2014). Securing web accounts using graphical password authentication through watermarking. International Journal of Computer Trends and Technology, 9(6), 269-274.

[8]. Kale, R., Shirkande, S. T., Pawar, R., Chitre, A., Deokate, S. T., Rajput, S. D., & Kumar, J. R. R. (2023). CR System with Efficient Spectrum Sensing and Optimized Handoff Latency to Get Best Quality of Service. International Journal of Intelligent Systems and Applications in Engineering, 11(10s), 829-839.

[9]. Nagtilak, S., Rai, S., & Kale, R. (2020). Internet of things: A survey on distributed attack detection using deep learning approach. In Proceeding of International Conference on Computational Science and Applications: ICCSA 2019 (pp. 157-165). Springer Singapore.

[10]. Mane, Deepak, and AniketHirve. "Study of various approaches in machine translation for Sanskrit language." International Journal of Advancements in Research & Technology 2.4 (2013): 383.

[11]. Shivadekar, S., Kataria, B., Limkar, S. et al. Design of an efficient multimodal engine for preemption and post-treatment recommendations for skin diseases via a deep learning-based hybrid bioinspired process. Soft Comput (2023). https://doi.org/10.1007/s00500-023-08709-5

[12]. Shivadekar, Samit, et al. "Deep Learning Based Image Classification of Lungs Radiography for Detecting COVID-19 using a Deep CNN and ResNet 50." International Journal of Intelligent Systems and Applications in Engineering 11.1s (2023): 241-250.Khin Than Nyunt,NawThiriWaiKhin “Web for career analysis based on youtube data APIs using web content mining abstract” on 2022.

[13]. Ajay Sudhir,NaveenGhorpade,Rohith S, S Kamalesh, Rohith R, Rohan B S "Web Scraping Approaches and their Performance on Modern Website"on 2022.

[14]. Chiapponi, Marc Dacier,OlivierThonnard,MohamedFangar,MattiasMattsson,VincentRigal "An industrial perspective on web scraping characteristics and open issues" on 2022.

[15]. DipaliShete,SachinBojewar,AnkitSanghvi "Survey Paper on Web Content Extraction and Classification" 0n 2021.

[16]. Roopesh N, Akarsh M S, C. NarendraBabu "An Optimal Data Entry Method, Using Web Scraping and Text Recognition" 0n 2121.

[17]. Eric C. Dallmeier "Computer Vision-based Web Scraping for Internet Forums" on 2021.

[18]. ERDINC¸ UZUN "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages" on 2019.

[19]. VidhiSingrodia, AnirbanMitra "A Review on Web Scrapping and its Applications" on 2019.

[20]. Rabiyatou DIOUF, EdouardNgor SARR, Ousmane SALL, Babiga BIRREGAH, Mamadou BOUSSO, SenyNdiaye ´ MBAYE "Web Scraping: State-of-the Art and Areas of Application" on 2019.

[21]. Gunawan, R., Rahmatulloh, A., Darmawan, I., and Firdaus, F. (2019). Comparisonof web scraping techniques: regular expression, HTML DOM and Xpath. In International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018) Comparison (Vol. 2):283-287.

[22]. Parlewar, P., Jagtap, V., Pujeri, U. , Kulkarni, M. M. S. ., Shirkande, S. T. ., &Tripathi, A. . (2023). An Efficient Low-Loss Data Transmission Model for Noisy Networks. International Journal of Intelligent Systems and Applications in Engineering, 11(9s), 267–276