JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Web-Scraping For e-Commerce Website

Daithiya Sudan K S, Perumalraja R, Kamalesh S

Department of Information Technology,
Velammal College of Engineering and Technology, Madurai, India
rpr@vcet.ac.in and skl@vcet.ac.in

***Abstract*— Web scraping is basically an interactive method for website and some other online sources to browse for and access data. To delete a replica of the information and save it in an external archive for review, it uses software engineering technology and custom software programming to extract data or any other content of on-line sources. Web scraping is often called automatic data gathering, database discovery, database crawling, or content management mining. Web scraping has possibly existed since before the start of the World Wide Web, but it has been used mainly in the context of data analytics, and is generally associated to e-commerce. Web scraping technique provides a broad collection of options and can serve various purposes: A web crawler's least necessity is to automate the normally physical work of gathering cost quotation marks and website article details. A web crawler's main requirement will be to discover formerly inaccessible sources of price data, and include a survey of all accessible price information. This scraping process is performed using different technologies which can be automatic application tools or manual methods. This paper provides the overall review of web scraping technology, how it is carried out and the effects of this technology.**

Keywords*— *Web scraping, E-commerce, Data extraction, Web crawler, automatic tools

I. INTRODUCTION

Nowadays in the modern world, data plays a very important role in satisfying the customers. In order to reach the customers' needs, every e-commerce website must have the data regarding all the products that is being sold enormously and also the products that is sold very less. Only if the data present in the store is valuable to the owners, they can adjust the products present in their website and make better sales. So to achieve that one has to get all the data there is available on the internet itself to give the customers with what they want to most at a cheaper rate. There is an enormous amount of data out on the internet which cannot be manually fetched and manipulated. One important consequence of this phenomenon for consumer researchers is that as consumers immerse themselves in the digital world, they continuously create enormous amounts of data containing information about their attitudes, revealed preferences, and behaviours. To fetch the required data in an organized manner, we use web scraping. The data that is deeply scraped using web scraping cannot be manually scraped by an ordinary human being. In simple words, web scraping helps us to decrease the amount of time required in gathering the data. If the data is gathered manually, it will take a lot of time and becomes a time consuming process. But, using the web scraping tools, one can scrap the required data in few minutes.

Purpose of Web Scraping

There are five most common and impactful use cases that businesses use to produce actionable insights and keep a watchful eye over the health and competitiveness of their industry.

- Market trend analysis
- Price monitoring
- Optimizing point of entry
- Research & Development
- Competitor monitoring

Market trend analysis -- When performing a market trend analysis, web scraped data is perfectly suited to complement and enhance the productivity and accuracy of your research methods. Acquiring insights into any particular market scenario and the larger industry environment requires a great deal of data, and web scraping provides this data with a guaranteed level of reliability and accuracy. Rigorous quality assurance standards are critical for this type of information, as subtle trends and market behaviours have been demonstrated as key indicators for market movement, and being a first-mover unlocks tremendous opportunities for any organization.

Price monitoring -- To make profitable pricing decisions, having access to a timely, reliable source of high-quality data is crucial. By scraping pricing data, market research teams are empowered to confidently consult their organizations and clients on how to best position products and services. In an online world where prices change as rapidly as new products appear, automating a healthy stream of pricing data into your market research team is essential for ensuring you have an up-to-date, reliable benchmark against which you can either compete within the market or enter it.

Optimizing point of entry -- Where you position yourself and how you price your goods is as important as your product or service itself. By using web scraping to fuel market research into your industry and location you can enter the market with confidence and a competitive upper-hand. With web scraping, a huge variety of relevant, essential information about a market can be aggregated extremely quickly, capable of fueling aggressive startup growth as well as new product launches into competitive industries.

Research and development -- Precautions are taken to ensure a healthy, thoughtful R&D cycle can dramatically reduce post-launch headaches. Whether your company produces enterprise software, video games, canned beverages, or electric cars, spending more time in R&D can be the difference between catastrophic failure and success. Web scraped data is to R&D what buttresses are to ancient structures - it supports the entire process while preventing misguidedness and disaster. Since the scope and capability of big data is truly driving an epochal change in research - of any kind - using web scraping to generate data for R&D teams unlocks tremendous insight across every aspect of the cycle.

Competitor analysis -- The traditional non-data-driven competitor monitoring processes employed by many businesses put them at real risk for disruption, creating blind spots and pain points for competitors to exploit. By integrating web scraped data into a systematic competitor monitoring process, market researchers

provide businesses with a powerful advantage and can act quickly on competitor insights to maximize revenue, market share, and growth opportunities.

Digital Footprint -- A digital footprint is data that is left behind when users have been online. There are two types of digital footprints which are passive and active. A passive footprint is made when information is collected from the user without the person knowing this is happening. An active digital footprint is where the user has deliberately shared information about themselves either by using social media sites or by using websites.

II. Literature Review

As we all know, internet and e-commerce are entirely committed towards every developed country. But we think it can be accomplished and can make a remarkable benefit to developing countries also if an ideal business purpose can be made. Ohidujja man et.al clearly discussed that E-commerce is a revolution & turning point in online business practices and can make a huge contribution to the economy and Hasan et.al also indicated that currently, e-commerce organizations have increasingly become a fundamental component of business strategy and a strong catalyst for economic development. A huge amount of research works has been done on e-Commerce which is basically on online shopping. A large group of researchers has found out and also pointed out the necessity and possibilities of Online Shopping. On the other hand, limitation of ecommerce is found and at the same time, they provided essential suggestion and came to a prediction to make Online Shopping more useful for the consumers. But the contribution of traditional marketing is also inescapable but compare to online shopping it is less effective we think. So on this basis, Mehrdad Salehi et.al found out distinguish between online marketing & traditional marketing. Though most of the people of Bangladesh especially the rural people are not enough capable of operating internet to run the online business. For that reason, they need to be dependent on traditional marketing.

E- Marketing	Traditional Marketing
Interactive advertisement. Example: website, social networking site, Google ads, banner ads, video marketing.	Contact from one side. Example: Print media (Newspaper), Broadcast Media (TV & radio ads), telemarketing.
E-marketing methods less expensive	Traditional marketing methods more expensive
Reach out maximum people	Limited audiences
Instant Comparable	Less opportunity
Save a lot of time	Need a lot of time
Less interaction	Interaction with people can make good relationship.

Fig 1. Difference between e-marketing & traditional marketing

III. System Design

System Flow — in this system, we are finding the product title, its price, total review count, ratings for feature reviews only and not for service and product review and its availability. Since on Amazon reviews there are ratings available for each review, the sentiment for the product review and service review will be equivalent to the review given by the customer so the computational task of finding the sentiments is reduced in case of service and product reviews.

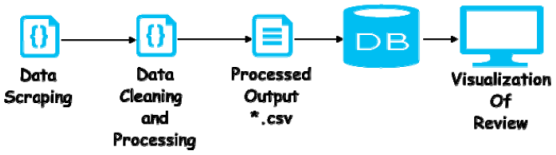

Fig 2. Basic Computations

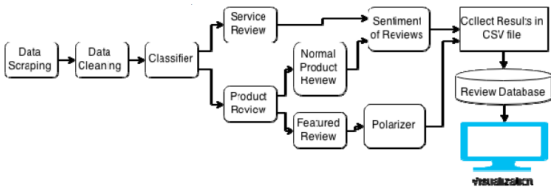

Fig 3. Detailed Block Diagram of System

The Steps are as follows:

1. **Data Scraping:**
   Crawl the amazon review URL to extract all required details from it. We need to take care of the text so as to satisfy the required format, for e.g. tags have a special meaning to the browser i.e. break read or next line, we need to explicitly convert each tag to spaces or else the crawling result will be improper. When working with online reviews there is always a question in our mind, how can I trust the reviews. This is not a problem with amazon reviews, amazon reviewers can up vote or down vote a review, this collectively is available as helpful count. We have taken a special care in extracting the data from web pages smallest necessary data is extracted for processing. The following is the list of items that we have extracted: Review of Title, Helpful Count, User Review and Date of Review. Caution: Websites uses utf-8 character set for encoding characters, but sometimes this encoding can give errors during web scraping as scraping involves matching strings and patterns. Solution to this is simply enforce the string to be coded in utf-8 format.
2. **Data Cleaning and Processing**:
   The data extracted need to be cleaned so that we get proper text review on which analysis can be performed. Cleaning of crawled data is done by removal of all special characters (such as: ":/.,'#$*^&-) in order to retrieve best results. After cleaning the crawled content copy it into a csv file. The next step is processing the cleaned data, firstly review is classified as service, feature or product review.
3. All **processed output** is stored in one csv file for further use
4. The file is then loaded into the **database** for use in visualization and summarization.
5. Finally the **summarization** of sentiments is generated as charts and displayed to the user as an attractive dashboard.

A web application may adopt one of the following measures to stop or interfere with a web scrapping tool that collects data from the given website. Those measures may identify whether an operation was conducted by a human being or a bot. Some of the major measures include the following:

i. HTML "fingerprinting" that investigates the HTML headers to identify whether a visitor is malicious or safe (Acar et al. 2013);

ii. IP reputation determination, where IP addresses with a recorded history of use in website assaults that will be treated with suspicion and are more likely to be heavily scrutinized (Sadan and Schwartz 2012);

iii. Behavior analysis for revealing abnormal behavioral patterns, such as placing a suspiciously high rate of requests and adhering to anomalous browsing patterns;

iv. Progressive challenges that filter out bots with a set of tasks, such as cookie support, JavaScript execution, and CAPTCHA.

IV. Algorithm

To extract data from amazon, there are two prerequisites. They are:

i. We have to store the URLs of all the products that we are about to extract in a text file.

ii. We need ids' of objects we are about to scrape.

The URL file will look like something that is shown as in Fig 4.

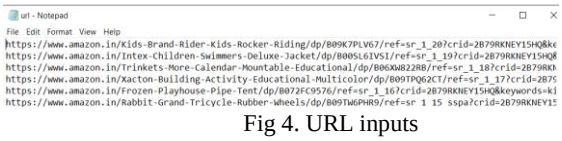

Fig 4. URL inputs

After those prerequisites are achieved, we need the below modules to be installed in the system.

i. BeautifulSoup: This will be our primary module that contains a method to access the webpages over HTTP.

ii. xhtml: This is our helper library that helps to process webpages in python programs.

iii. Requests: Makes the process of sending HTTP requests flawless. The output of the function

	A	B	C	D	E
1	Product Title	Price	Rating	Review Count	Availability
2	Kids Brand Horse Rider-Kids	₹1699.00	3.8 out of 5 stars	55 ratings	In stock.
3	Intex Kids Children Young Sw	₹365.00	4.1 out of 5 stars	59 ratings	In stock.
4	Trinkets & More Calendar Cl	₹449.00	4.2 out of 5 stars	411 ratings	In stock.
5	Xacton Bullet Blocks | Bullet	₹249.00	4.1 out of 5 stars	10 ratings	In stock.
6	Frozen Play house tent for ki	₹889.00	4.2 out of 5 stars	1767 ratings	In stock.
7	R for Rabbit Tiny Toes Grand	₹6449.00	4.3 out of 5 stars	2417 ratings	In stock.
8	R for Rabbit Tiny Toes Sport	₹6993.00	4.5 out of 5 stars	1058 ratings	In stock.
9	Wonder House Books Prince	₹169.00	3.8 out of 5 stars	51 ratings	In stock.

Fig 5. CSV file as output

Approach:

1. First, we are going to import our required libraries.
2. Then we will take the URL stored in our text file.
3. Crawl the amazon URLs to extract all required details from it.
4. Special care for required format of information must be taken, example tags have a special meaning to the browser i.e. break read or next line, we need to explicitly convert each tag to spaces or else the crawling result will be improper.
5. Cleaning the crawled data. Removal of all special characters (such as: ":/.,'#$*^&-) must be done in order to retrieve best results. This also saves our review processing time.
6. Put the crawled content into a csv file.

V. CONCLUSION

Consumers and managers create rich and diverse digital footprints that capture their behaviour. In this report, we discuss methods of transforming such data into impactful datasets for answering consumer research questions. We begin by demystifying the process of harvesting data from the internet via web scraping.

While an understanding of the basic mechanics of web scraping is a necessary condition for leveraging and realizing the full potential of data on the internet for consumer research, it is not sufficient. We also outline a structured workflow for generating credible consumer research findings via web scraping that entails four key facets: (1) design transparency, (2) analytic reproducibility, (3) analytic robustness, and (4) effect replicability and generalizability.

Web scraping can accelerate consumer research by reducing the cost and time required for data collection. In addition to collecting data by oneself using custom code, there are other ways of constructing novel web-based datasets for consumer research beyond designing custom scrapers.

Consumer researchers might also consider using tools (e.g., Mozenda, import.io, Instant Data Scraper) or outsourcing web scraping on crowdsourcing platforms (e.g., Fiverr, Upwork). While these approaches may be sufficient for answering certain research questions, they generally offer less control, flexibility, and scalability. The main goal of our structured workflow for using web scraping in consumer research is the generation of credible, replicable, and generalizable findings.

A positive side benefit of this approach for authors is that it will likely inspire other researchers to leverage the abundant data on the internet for conducting rapid, inexpensive, and high-quality replications of their work.

REFERENCES

[1]. S. A. a. A. N. S. Aljuhani, "A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones," International Journal of Advanced Computer Science and Applications, vol. 10, 2019.

[2]. L. a. L. B. Zhang, "Aspect and entity extraction for opinion mining," in Zhang, Lei and Liu, Bing, Berlin, Heidelberg, Springer, 2014, pp. 1--40.

[3]. Adjerid, Idris and Ken Kelley (2018), "Big Data in Psychology: A Framework for Research Advancement," *American Psychologist*, 73 (7), 899-917.

[4]. Antonakis, John, Samuel Bendahan, Philippe Jacquart, and Rafael Lalive (2010), "On Making Causal Claims: A Review and Recommendations," *The Leadership Quarterly*, 21 (6), 1086-120.

[6]. Balducci, Bitty and Detelina Marinova (2018), "Unstructured Data in Marketing," *Journal of the Academy of Marketing Science*, 46 (4), 557-90.

[7]. Berger, Jonah and Katherine L. Milkman (2012), "What Makes Online Content Viral?," *Journal of Marketing Research*, 49 (2), 192-205.

[8]. Edelman, Benjamin (2012), "Using Internet Data for Economic Research," *Journal of Economic Perspectives*, 26 (2), 189-206.

[9]. Datta, Hannes, George Knox, and Bart J. Bronnenberg (2018), "Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery," *Marketing Science*, 37 (1), 5-21.

[10]. Dreyer, Anthony J. and Jamie Stockton (2013), "Internet 'Data Scraping': A Primer for Counseling Clients," *New York Law Journal*.

[11]. Edelman, Benjamin (2012), "Using Internet Data for Economic Research," *Journal of Economic Perspectives*, 26 (2), 189-206.

[12]. Groves, Robert M. (2011), "Three Eras of Survey Research," *Public Opinion Quarterly*, 75 (5), 861-71.

[13]. Henkel, Alexander P., Johannes Boegershausen, Joandrea Hoegg, Karl Aquino, and Jos Lemmink (2018), "Discounting Humanity: When Consumers Are Price Conscious Employees Appear Less Human," *Journal of Consumer Psychology*, 28 (2), 272-92.

[14]. Lewis, Kevin (2015), "Three Fallacies of Digital Footprints," *Big Data & Society*, 2 (2), 1-4.

[15]. Adjerid, Idris and Ken Kelley (2018), "Big Data in Psychology: A Framework for Research Advancement," *American Psychologist*, 73 (7), 899-917.

[16]. Aguinis, Herman, Wayne F. Cascio, and Ravi S. Ramani (2017), "Science's Reproducibility and Replicability Crisis: International Business Is Not Immune," *Journal of International Business Studies*, 48 (6), 653-63.

[17]. Ali, Meiryum (2019), "The Best 25 Datasets for Natural Language Processing," https://web.archive.org/web/20190812134029/https://lionbridge.ai/datasets/the-best-25-datasets-for-natural-language-processing/.

[18]. Balducci, Bitty and Detelina Marinova (2018), "Unstructured Data in Marketing," *Journal of the Academy of Marketing Science*, 46 (4), 557-90.

[19]. Barnes, Christopher M., Carolyn T. Dang, Keith Leavitt, Cristiano L. Guarana, and Eric L.Uhlmann (2018), "Archival Data in Micro-Organizational Research: A Toolkit for Moving to a Broader Set of Topics," *Journal of Management*, 44 (4), 1453-78.

[20]. Barnes, Nick (2010), "Publish Your Computer Code: It Is Good Enough," *Nature News*, 467(7317), 753.

[21]. Berger, Jonah, Alan T. Sorensen, and Scott J. Rasmussen (2010), "Positive Effects of Negative Publicity: When Negative Reviews Increase Sales," *Marketing Science*, 29 (5), 815-27.

[21]. Bernerth, Jeremy B. and Herman Aguinis (2016), "A Critical Review and Best-Practice Recommendations for Control Variable Usage," *Personnel Psychology*, 69 (1), 229-83.

[22]. Blevins, Dane P., Eric W. K. Tsang, and Seth M. Spain (2015), "Count-Based Research in Management: Suggestions for Improvement," *Organizational Research Methods*, 18 (1), 47-69.

[23]. Cameron, A. Colin and Pravin K. Trivedi (1998), *Regression Analysis of Count Data*, New York: Cambridge University Press.

[24]. Chen, Eric Evan and Sean P. Wojcik (2016), "A Practical Guide to Big Data Research in Psychology," *Psychological Methods*, 21 (4), 458-74.

[25]. Chen, Zoey (2017), "Social Acceptance and Word of Mouth: How the Motive to Belong Leads to Divergent WOM with Strangers and Friends," *Journal of Consumer Research*, 44 (3), 613-32.
[26]. Dai, Hengchen, Cindy Chan, and Cassie Mogilner (2019), "People Rely Less on Consumer Reviews for Experiential Than Material Purchases," *Journal of Consumer Research*, forthcoming.
[27]. Datta, Hannes, George Knox, and Bart J. Bronnenberg (2018), "Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery," *Marketing Science*, 37 (1), 5-21.
[28]. Dolbec, Pierre-Yann and Eileen Fischer (2015), "Refashioning a Field? Connected Consumers and Institutional Dynamics in Markets," *Journal of Consumer Research*, 41 (6), 1447-68.