Importance of web scraping in e-commerce and e-marketing

Kasereka Henrys
MSc in Software Engineering
Bugema University
Published January 2021, HKCorporation
arisjokov@gmail.com

Abstract

Web scraping, also known as data mining, is the process of collecting large amounts of data from the web and then placing it in databases for future analysis and later use. Web scraping offers insight into price data, market dynamics, prevailing trends, practices employed by your competitors and the challenges they face. This is readily available data if you know how to get it. What many marketers don't know is how useful it can be to them. In this article, we will explain the advantage of Web scraping technique and supply a practical example that can be beneficial for e-commerce businesses and online marketers.

KEYWORDS web scraping, e-marketing, e-commerce, python.

INTRODUCTION

Websites are an ocean of limitless information which anyone and everyone can access. The new trend of technology put us to change the way we’re doing our business. The Internet now is the new place for business. Knowing how to use the Internet and the different opportunities that can bring us can be one of the keys to success in e-commerce and e-marketing.

Data crawling or web scraping or data harvesting has been into the existence for as long as the web itself. It is always associated with web content extraction, at the begin it wasn’t always served this purpose. Web scraping can be considered as a method of retrieving or extraction content from a website for Web scraping is the method of retrieving or extracting content from a website.
for the purpose of using it for purposes beyond the control of the website owner.

The first use of web scraping was to create links with test frameworks. Using tools like Selenium, companies like IP-Label have created products that allow web developers and webmasters to monitor website performance on a daily basis (xbyte, 2021). In the early times, extracting web data meant copying the text available on a web page to a local file, manually this approach was extremely inefficient and could not be used for business application. Spreadsheet software like Microsoft excel and google sheets have some basic web scraping capabilities and they were mainly used for extracting HTML tables from webpages (Slideshare, 2021).

In a world of fierce competition, every business tries to exceed itself. The question that probably worries all other entrepreneurs. In an industry where everything is customer-centric, competitor analysis is not a choice but a necessity. Having access to so much data can provide a competitive advantage in the field to which you belong (Samanta, 2021).

Since a huge amount of mixed data is constantly being generated on the web, web scraping is widely recognized as an efficient and powerful technique for collecting big data (Judit, 2001).

This concept of web scraping is not new to most data scientists, but is in fact becoming increasingly well-known due to the mass of data we can find on the internet and new startups that don't want to spend time collecting data where it can be found quickly on the internet.

OVERVIEW OF CONCEPTS

What is Ecommerce?

Ecommerce, also known as electronic commerce or internet commerce, refers to the buying and selling of goods or services using the internet, and the transfer of money and data to execute these transactions (Shopify, 20121). According to Wikipedia e-commerce Electronic is the activity of buying or selling products electronically on online services or on the Internet (Wikipedia, E-commerce, 2021).

Type of e-commerce:

- Business to business
- Business to consumer
- Consumer to consumer
- Consumer to business

What is e marketing?

E-marketing (or electronic marketing) refers to marketing carried out on the Internet. Two
synonyms of E-Marketing are Internet Marketing and Online Marketing which are frequently interchanged. Electronic marketing is the process of marketing a brand (business, product or service) using the Internet through computers and mobile devices (Conversionpipeline, 2020). Encyclopedia.com defines e-marketing as a process of planning and executing the design, distribution, promotion and pricing of products and services in a computerized and networked environment, such as the Internet and the World Wide Web, to facilitate exchanges and respond to customer requests (Encyclopedia, 2020).

**What is Web Scraping?**

Web scraping is automatically retrieving (and processing) information from websites (Bosch, 2021). Matthew define web scraping as a process of making a semi-structured document from the internet, usually in the form of web pages like HTML or XHTML, and analyze the document to take certain data from the page to be used for other purpose (Mathew, 2010).

**BENEFITS OF WEB SCRAPING**

Web scraping has emerged as an important strategy for e-commerce businesses, especially in delivering rich data-based insights.

**Practices of web scraping in e-commerce**

- Price-monitoring and Product Research
- Online price comparison
- Better Customer analysis
- Market Analysis
- Better advertisements
- Influences Marketing and Sales Strategy
- Brand monitoring
- Extract business details from business directory
- Helps in Future analysis
- Etc.

**Advantage of using web scraping**

- An online presence can be tracked. It is also an important aspect of web scraping where company profiles and reviews on websites can be removed. This can be used to see product performance, user behavior, and reaction.
- Personalized analysis and curation. This one is mainly for new websites / channels where the retrieved data can be useful for channels to know viewer behavior. This is done for the purpose of delivering targeted news to the public. So, what you watch online sets the pattern for the website to behave so that the business
knows its audience and can deliver what the audience actually likes.

- Online reputation. When the company plans its ORM strategy, the data extracted helps understand which audiences’ companies are most hoping to have an impact and which areas of responsibility may most expose the brand to reputation damage. By understanding these areas of vulnerability, the business can use them to its greatest advantage.

- Provide better targeted advertising to customers. Scrapping not only gives numbers, but also sentiment and behavioral analysis, so the business can know what audience types and choice of ads they want to see.

- To bring together public opinion. Monitor specific company pages on social media to gather updates on what people are saying about certain companies and their products. Data collection is always useful for product growth.

- Search engine results for SEO tracking. By scraping off organic search results, you can quickly find your SEO competitors for a particular search term. You can determine the title tags and the keywords they target. So you get an idea of what keywords drive traffic to a website, what categories of content get links and user engagement, what kind of resources it takes to rank your site.

- Scratch Leads: This is another important use for the sales-oriented organization in which lead generation is done. Sales teams are always hungry for data, and using web scraping technique, you can pull leads from many directories and then contact them to do an introduction to selling. The data can be extracted in any format you want and can be used for lead generation, brand development, or other purposes.

- To create vertical specific search engines. Even though this is a new thing that is popular in the market, but also requires a lot of data, so web scraping is done for as much public data as possible because this volume of data is practically impossible to collect.

**Challenges of web scraping**

- It is not always easy to know which site to scrape. Sometimes choose between scraping the data owner or an aggregator site.

- The internet is dynamic. Each web site has a particular structure, which may be subject to changes anytime.

- Data is volatile.

- Legal issues and informing web site owners.
HOW DO WEB SCRAPERS WORK?

- First step, robots.txt: One of the most important and overlooked step is to check the robots.txt file to ensure that we have the permission to access the web page without violating any terms or conditions.
- Secondly, the web scraper will be given one or more URLs to load before scraping. The scraper then loads the entire HTML code for the page in question. More advanced scrapers will render the entire website, including CSS and JavaScript elements.
- Then the scraper will either extract all the data on the page or specific data selected by the user before the project is run.
- Ideally, the user will go through the process of selecting the specific data they want from the page. For example, you might want to scrape an Amazon product page for prices and models but are not necessarily interested in product reviews.
- Lastly, the web scraper will output all the data that has been collected into a format that is more useful to the user.

Figure 1: Architecture of web scraping

Most web scrapers will output data to a CSV or Excel spreadsheet, while more advanced scrapers will support other formats such as JSON which can be used for an API.

Figure 2: Overview of web scraping system
WEB SCRAPING TECHNIQUES

There are many techniques use for scraping data on internet (wikipedia, 2021).

- Traditional copy and paste
- Text grapping and regular expression
- Hypertext Transfer Protocol (HTTP) Programming
- Hyper Text Markup Language (HTML) Parsing
- Document Object Model (DOM) Parsing
- Web Scraping Software
- Vertical aggregation platforms
- Semantic annotation recognizing
- Computer vision web-page analyzers

As you can see we have many techniques that can be used but the target is to use a technique that supplies us a correct result in a reasonable time and the best one is web scraping software.

Web scraping software

You have a choice between using an existing software or build your own script. There are many software tools available that can be used to scrape data on internet. Therefore, there are two well-known and popular methods that can be used to scrape data on the web namely the use of generic web scraping software and write code.

Use generic web scraping software

Web Scraping software falls under 2 categories. First, which can be locally installed in your computer and second, which runs in cloud browser based.

- Visual Web Ripper
- Helium Scraper
- Screen Scraper
- OutWit Hub
- Mozenda
- WebSundew
- Web Content Extractor
- import.io
- etc.

Writing code

You can hire a developer to build custom data extraction software for your specific requirement or do it by yourself. The most well-known language use in web scraping include:

- R programing (rvest, Rselenum)
- Python (BeautifulSoup, Selenium)
- Ruby
- NodeJS
- Java
Advantage and disadvantage

| Generic web scraping | Programming language |
|----------------------|----------------------|
| Closed source        | Open source          |
| Expensive            | Free                 |
| Limited              | Extremely powerful    |
| Easy to learn        | Steep learning curve |

Figure 3: Advantage and disadvantage between generic software and writing code for web scraping.

The problem with most generic web scraping software is that they are very difficult to set up and to use (webharvy, 2021). Based on the challenge of web scraping can face it will be recommended making your algorithm for web scraping.

PRACTICAL EXAMPLE

One of the advantages of web scraping in e-commerce is market analysis. In our practical example, we will show how to get data from Amazon. The section that will interest us will be to obtain the Amazon Best Sellers data in our case we will get the bestselling phone in amazon. These details can be useful to anyone who wants to start selling or manufacturing phones. Having the big picture of the market will help you improve your business and plan well.

Python

The reason we choose Python it is because it is popular, high-level programming language. Python can run on many different platforms and has a simple syntax similar to the English language, therefore, it is easy to code.

First of all, the goal of web scraping is to collect web data that would be in HTML format or Json. Python provides a library called Requests, which is a simple HTML library that allows you to integrate your Python programs with web services.

Once you find the relevant data for your project on the webpage, you can download it for valuable insight. To do this, Python provides another library called BeautifulSoup, which helps you retrieve particular content from a web page, remove HTML tags, and save the information.

The last step in web scraping is to save the collected data in a structured form. With the help of Python Pandas Library, you can store the data in the desired format.

Amazon.com

Amazon.com, Inc. is an American multinational technology company based in Seattle, Washington, which focuses on e-commerce, cloud computing, digital streaming, and artificial
intelligence. It is considered one of the Big Five companies in the U.S. In e-commerce, Amazon is leading in that domain, how we can benefit of the huge data which we find on its website which can help us for price comparing or other aspects in e-commerce (Wikipedia, wikipedia, 2021).

Implementation

```python
#!/usr/bin/env python3
# -*- coding: utf-8 -*-

## Created on Mon Jan 18 15:57:26 2021
@email: arisjokov@gmail.com
@author: henryskas

import requests
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from bs4 import BeautifulSoup
from datetime import datetime

headers = {'User-Agent':"Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:66.0) Gecko/20100101 Firefox/84.0", "Accept-Encoding":"gzip, deflate","Accept":"text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8","DNT":"1","Connection":"close","Upgrade-Insecure-Requests":"1"}
result = requests.get("https://www.amazon.com/Best-Sellers-Cell-Phones/zgbs/wireless/7072561011", headers=headers)
#result = requests.get("https://www.amazon.com/best-sellers-books-Amzon/zgbs/books", headers=headers)
print(result.status_code)
if result.status_code < 500 :  
    src = result.content
    soup = BeautifulSoup(src,'lxml')
    links = soup.find_all('span',{'class' : 'aok-inline-block zg-item'})
    df = pd.DataFrame(columns=['tittledata','pricedata','linkdata'])
    #Get all data
    for link in links :
        linkdata = link.find('a',{'class':'a-link-normal a-text-normal'})
        pricedata = link.find('span','class:p13n-sc-price')
        tittledata = link.find('div','class:p13n-sc-truncate-desktop-type2')
        # Check empty link
        if linkdata is not None and linkdata['class'] is not None:
            df2 = pd.DataFrame([{'tittledata':tittledata.text.strip(),'pricedata':pricedata.text.strip(),'linkdata':linkdata['href'].strip()}],columns=['tittledata','pricedata','linkdata'])
            df = df.append(df2,ignore_index=True)
        else :
            print("Link doesn't exist")
now = datetime.now()
current_time = now.strftime("%H%M%S")
#Split price to get a numeric number
df['pricedata'] = df['pricedata'].str.replace('$', '')
df['price'] = df['pricedata'].apply(lambda x: x.split('.')[0])
#Add a new column
df['currency']='$
#Let's find out 10 product which have the highest-priced data = df.sort_values(["price"], axis=0, ascending=False)[:10]
data['tittledata'] = data['tittledata'].str.slice(0,20)
#Plot
y_pos = np.arange(len(data['tittledata']))
plt.barh(y_pos, data['price'])
plt.yticks(y_pos, data['tittledata'])
plt.xlabel('Amount in dollars')
plt.ylabel('Labels')
plt.title('10 products with the highest prices')
plt.legend()
plt.show()
#Generate a csv
df.to_csv('result' + current_time  + '.csv')
else:
    if "To discuss automated access to Amazon data please contact" in result.text:
        print("Page was blocked by Amazon. ", result.text)
    else:
        print("Page must have been blocked by Amazon as the status code was ", result.text)
```

Figure 4: Web scraping algorithm to retrieve the bestselling products on amazon.com (Author)

Result

![Figure 5: Histogram of 10 best-selling phones at the highest prices January 18, 2021](image-url)
Table 1: 10 best-selling phones at the highest prices January 18, 2021

Table 2: Amazon Phone Best Selling List January 18, 2021

CONCLUSION

“If programming is magic, then web scraping is wizardry” (Mitchell, 2015) Said Ryan Mitchell. The presence of the Internet led to increasing source of information that can be accessed so that information seeking activities become the most common activities performed and became one of the activities that took quite a bit (Adam, 2012). The Internet will be remembered as the first place where we can collect huge amounts of data without spending a lot of energy or money. Whether in e-commerce or e-marketing, the use of the technique of web scraping will be the key to success as it will provide insight into the targeting market and help decision makers.

ACKNOWLEDGMENT

The author would like to thank Mr. Ricky Muhindo (Ph.D. Computer science candidate at Northcentral university) for his support.

REFERENCES

Adam, R. (2012). Task-Specific Information Retrieval System for Software Engineers. Journal of Computer and System Sciences, 78.

Bosch, O. t. (2021, 01 13). Uses of web scraping for official statistics. Retrieved from circabc.europa.eu:
https://circabc.europa.eu/sd/a/5e250346-44a9-471b-87f1-5b5ddb59aa77/1_Big%20Data%20Sources%20part3-Day%201-A%20Use.pdf

Conversionpipeline. (2020, 01 13). conversionpipeline. Retrieved from conversionpipeline:
https://conversionpipeline.com/what-is-emarketing/

Encyclopedia. (2020, 01 13). encyclopedia. Retrieved from encyclopedia: https://www.encyclopedia.com/finance/finance-and-accounting-magazines/emarketing

Judit, B.-I. (2001). Data collection methods on the web for. Springer.

Mathew. (2010). Guide to web scraping with PHP. Marco Tabini & Associates, Inc.

Mitchell, R. (2015). Web Scraping Using Python, First Edition. Orilley.

Samanta, J. (2021, 01 13). Data scraping: How to leverage it for your eCommerce business? Retrieved from towardsdatascience: https://towardsdatascience.com/data-scraping-how-to-leverage-it-for-your-e-commerce-business-7320e8f82358

Shopify. (20121, 01 13). what-is-e-commerce. Retrieved from shopify: https://www.shopify.com/encyclopedia/what-is-e-commerce

Slideshare. (2021, 1 10). the history of web scraping. Retrieved from slideshare: https://fr.slideshare.net/promptcloud/the-history-of-web-scraping

webharvy. (2021, 13 01). what is web scraping. Retrieved from what is web scraping: https://www.webharvy.com/articles/what-is-web-scraping.html

Wikipedia. (2021, 01 13). E-commerce. Retrieved from wikipedia: https://en.wikipedia.org/wiki/E-commerce

wikipedia. (2021, 01 17). Web scraping. Retrieved from Wikipedia Free Encycl: https://wikipedia/web_scraping

Wikipedia. (2021, 01 19). wikipedia. Retrieved from wikipedia: https://en.wikipedia.org/wiki/Amazon_(company)

xbyte. (2021, 1 10). the history of web scraping. Retrieved from xbyte: https://www.xbyte.io/the-history-of-web-scraping.php

**BIOGRAPHY OF AUTHORS**

Kasereka Henrys, he received his B.Sc. Degree in Computer Science from the University of Kinshasa. He completed his Master of Computer Science degree from Bugema University in software engineering. Got an online specialization in Data Science on coursera provided by Johns Hopkins University. He is interested in the field of data science, mobile and web programming.