Web scraping for beginners - Simple automation of online data collections

Keywords: consumer price statistics, dynamic pricing, web scraping

1. INTRODUCTION

As e-commerce is continuously gaining in importance, it needs to be considered appropriately in calculating the inflation rate. Measuring representative prices has become more difficult because the prices of certain product groups or those specified by online retailers have become increasingly volatile (dynamic pricing), and a manual measurement of volatile prices is hardly feasible. For this reason, web scraping has been increasingly used to collect data for price statistics in recent years. A generic web-based application has been developed to facilitate the use of this automated data collection technique. The application was named “ScraT” which is a compound of the words “scraping” and “tool”. The current gradual introduction of this application in price statistics aims to complement and largely replace the manual internet price collection for purposes of the consumer price index and the harmonised index of consumer prices by 2021. Additionally, other statistic divisions may soon profit by using the generic tool for their data collections. All in all, the tool supports the aim of further digitalizing the production of national statistical institutes.

2. METHODS

2.1. The incidence of dynamic pricing

Dynamic pricing is the use of automatic algorithms to change prices at short notice due to changes in market conditions or due to parameters indicating a consumer’s willingness to pay. The application of a dynamic pricing pattern is not new but has become more obvious since the growing importance of internet purchases and has additionally become popular for online retailers in order to attract customers or to increase profits. Dynamic pricing must be distinguished from the phenomenon of individualised pricing which is defined as the use of automatic algorithms to change prices due to characteristics of an individual consumer, such as the device used for purchase, the location of purchase, browser history or simply gender-related characteristics. While algorithms of dynamic pricing treat all consumers equally, individual pricing only affects consumers with pre-defined characteristics.

2.2. The application of web scraping for data collections

More and more information is freely available on the internet. Some of this information may be used for the production or validation of statistics. Web scraping has become an acknowledged technique for online data collection. Many national statistical institutes make use of web scraping for their production. One of the great advantages of using web scraping is, once a program is successfully launched, nearly infinite data collections at web shops or other web pages are possible to initiate. Web scraping therefore supports increasing the quality of statistics by simply having more data available. One of the disadvantages of web scraping is that, in the beginning, a considerable amount of effort is setting up a web scraping program which requires in-depth programming skills in appropriate languages, such as Java, Python or R.

The application of web scraping for all online data collections is therefore staff-related. Because of this, we developed a web-based application which enables staff without profound IT-skills to automate their data collections. Furthermore, with nearly infinite repetitions, potential pricing algorithms of online retailers can be observed and evaluated.

3. Results

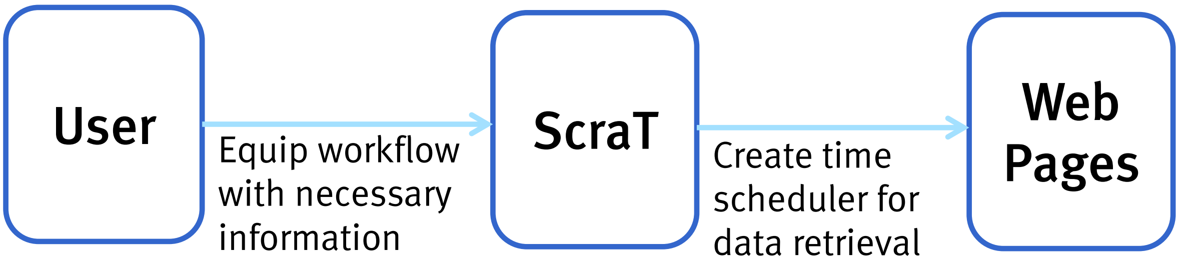
User
Equip workflow with necessary information
ScraT
Create time scheduler for data retrieval
Web Pages

Figure 1: process of data collection

The user sets up a workflow in ScraT by creating steps which resemble the steps the user makes when navigating to a page. Information can be uploaded to the workflow via an input table. Within the workflow, one can access the information and store them in variables in order to apply for the workflow. The data collection can be initiated, and nearly infinitely repeated via a time scheduler. ScraT finds the intended information with the help of XPaths. The extracted information is stored in MySQL databases and is ejected in csv-, xml-, and xls-format. Then the data may be used for analysis and index calculation with the help of SAS, Python or other programs.

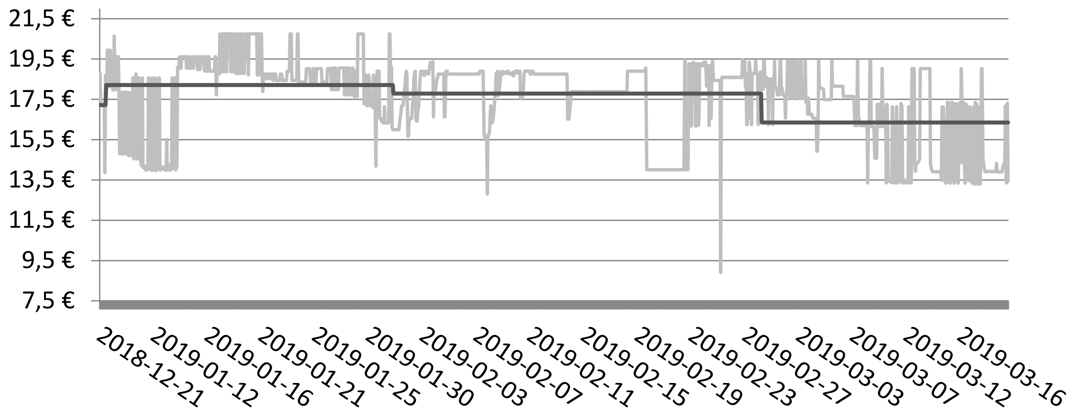

Figure 2: Prices for a men’s aftershave by a US-American e-commerce retailer

Dynamic pricing of online retailers may lead to a bias in the index calculation since the traditional way of price collection via internet is done generally at one time during the month and therefore cannot capture rapidly changing prices. Therefore, in order to display reliable price developments in indices, consumer price statistics needs to constantly monitor the pricing behaviour on the internet and apply methods to evaluate the large amount of data and integrate very volatile price developments into price indices.

4. Conclusions

Since the beginning of 2020, the application is fully operable and staff is taught on using it. By this approach, staff can easily automatize their own data collection and data collections are therefore less dependent on single staff members with profound IT-skills. The number of price quotes for production rose sharply, the sample of articles can now easily be expanded as well. Additionally, with the help of ScraT, dynamic pricing can be observed and volatile prices incorporated in indices.