Chapter 40
Big Data Techniques for Supporting Official Statistics: The Use of Web Scraping for Collecting Price Data

Antonino Virgillito
Istituto Nazionale di Statistica (ISTAT), Italy

Federico Polidoro
Istituto Nazionale di Statistica (ISTAT), Italy

ABSTRACT

Following the advent of Big Data, statistical offices have been largely exploring the use of Internet as data source for modernizing their data collection process. Particularly, prices are collected online in several statistical institutes through a technique known as web scraping. The objective of the chapter is to discuss the challenges of web scraping for setting up a continuous data collection process, exploring and classifying the more widespread techniques and presenting how they are used in practical cases. The main technical notions behind web scraping are presented and explained in order to give also to readers with no background in IT the sufficient elements to fully comprehend scraping techniques, promoting the building of mixed skills that is at the core of the spirit of modern data science. Challenges for official statistics deriving from the use of web scraping are briefly sketched. Finally, research ideas for overcoming the limitations of current techniques are presented and discussed.

INTRODUCTION

The use of Big Data in official statistics is a major topic that in recent years has been the subject of several initiatives both at the level of single National Institutes of Statistics and in the form of international collaboration projects. Among all the possible types of Big Data sources that have been tested for supporting the production of statistical indicators, the "Internet as a data source" is a highly popular one, which was subject of several experimentations in various domains, since the idea of exploiting the

DOI: 10.4018/978-1-5225-7501-6.ch040

huge amount of information available on the web is largely appealing for researchers. However, despite the apparent high accessibility of Internet data, the task of setting up a collection of web data with a level of quality that is sufficient to produce correct statistical indicators over a medium-long period of time is still an open issue.

The consumer price survey is an interesting test bed in this sense because automated collection of Internet data can replace that part of the current data collection based on repetitive centralized activities, at the same time improving the representation of modern consumption habits, more and more biased towards the use of e-commerce.

In general, various challenges have emerged for statisticians in terms of the use of Big Data for statistical purposes in this field. First of all the use of web scraping techniques as a tool to achieve big data for inflation measurement has directly to do with one of the three V's of big data, i.e., velocity. "Velocity" implies the possibility of improving timeliness in the production of statistical indicators. This is clearly true for a phenomenon, like inflation, that is characterized by temporal evolution but also for other phenomena, as unemployment, touristic flows, telecommunication, that are investigated through traditional data collection tools. Secondly web scraping techniques are widely available and Internet as data source is at disposal of all: in particular for topics for which the information are largely available on the web (and this is typically the case of consumer prices) this is a major "threat" to the monopoly of the National Statistical Institutes over data and information, currently derived by their official status. Last but not least, web scraping techniques applied to consumer prices as well as to other phenomena could offer access to a bigger amount of data compared to that accessed by the current data collection with the potential of improving the quality of the derived indicators.

In this chapter we extensively review the techniques that have been setup for collecting price data from the Internet relatively to different kinds of products. The objective is to automatically retrieve and make recognizable the information off the web page, writing it in local database/data-store/files, eliminating the use of "copy and paste" activity, currently used to collect price data. We will present and classify scraping implementation techniques that have been used in practice in different National Statistical Institutes and comment their effectiveness in the light of their application to different kinds of products, as a result of tests carried out in the Italian National Institute of Statistics.

The general problem of web scraping is to extract from semi-structured documents (i.e., web pages) some structured pieces of information that are contained in it. However, the languages that are used to implement web pages (e.g., HTML, CSS etc.) mostly constitute a syntax for describing stylistic aspects of the pages themselves, with no indication of the semantics of data they represent, which on the contrary is the specific focus of the scraping activity. This conceptual mismatch is at the basis of all the difficulties in the use of Internet as a data source, that are common to all the techniques used to implement scraping. Moreover, once a scraping process has been set up it has to be constantly monitored and maintained, in order to respond to changes in the structure of the web sites that may disrupt the correct execution of the collection process. This is an additional problem in a context such as that of an NSI, where IT resources for conducting surveys are generally scarce and survey statisticians do not have the skills to carry out such activity.

We will present the specific technical aspects of each solution and comment on the results of experimentation carried out in various National Institutes of Statistics, highlighting that although each technique provides some benefits with respect to a specific category of product, there is still no such a thing as a "one-stop-shop" solution for web scraping, that is a tool or technique that at the same time is easy to setup and maintain, preferably for non-technical users, and is not very sensitive to changes in the