

Chapter 10

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.

DOI: 10.4018/978-1-5225-2512-7.ch010

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 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

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/big-data-techniques-for-supporting-official-statistics/179969

Related Content

A Multi-Objective Ensemble Method for Class Imbalance Learning: Application in Prediction of Life Expectancy Post Thoracic Surgery

Sajad Emamipour, Rasoul Saliand Zahra Yousefi (2017). *International Journal of Big Data and Analytics in Healthcare* (pp. 16-34).

www.irma-international.org/article/a-multi-objective-ensemble-method-for-class-imbalance-learning/197439

Unstructured Healthcare Data Archiving and Retrieval Using Hadoop and Drill

Hang Yue (2018). *International Journal of Big Data and Analytics in Healthcare* (pp. 28-44).

www.irma-international.org/article/unstructured-healthcare-data-archiving-and-retrieval-using-hadoop-and-drill/223165

Proximate Breast Cancer Factors Using Data Mining Classification Techniques

Alice Constance Mensahand Isaac Ofori Asare (2019). *International Journal of Big Data and Analytics in Healthcare* (pp. 47-56).

www.irma-international.org/article/proximate-breast-cancer-factors-using-data-mining-classification-techniques/232335

A Review of Non-Linear Kalman Filtering for Target Tracking

Benjamin Ghansah, Ben-Bright Benuwa, Daniel Danso Essel, Andriana Pokuaa Sarkodieand Mathias Agbeko (2022). *International Journal of Data Analytics* (pp. 1-25).

www.irma-international.org/article/a-review-of-non-linear-kalman-filtering-for-target-tracking/294864

Brain-Computer Interface: Attack Surface Analysis

Anthony Tricheand Md Abdullah Al Momin (2022). *Security, Data Analytics, and Energy-Aware Solutions in the IoT* (pp. 161-172).

www.irma-international.org/chapter/brain-computer-interface/295907