The current issue and full text archive of this journal is available on Emerald Insight at
www.emeraldinsight.com/0007-070X.htm
BFJ 121,12
3350
Received 4 February 2019
Revised 10 June 2019 4 September 2019
Accepted 6 September 2019
British Food Journal
Vol 121 No. 12, 2019
pp. 3350-3361

0007-070X
DOI 10.1108/BFJ-02-2019-0081
Abstract
Web scraping
for food price research
Judith Hillen
Agroscope Tanikon, Ettenhausen, Switzerland
Purpose - The purpose of this paper is to discuss web scraping as a method for extracting large amounts of data from online sources. The author wants to raise awareness of the method's potential in the field of food price research, hoping to enable fellow researchers to apply this method.
Design/methodology/approach - The author explains the technical procedure of web scraping, reviews the existing literature, and identifies areas of application and limitations for food price research. Findings - The author finds that web scraping is a promising method to collect customised, high-frequency data in real time, overcoming several limitations of currently used food price data sources. With today's applications mostly focussing on (online) consumer prices, the scope of applications for web scraping broadens as more and more price data are published online.
Research limitations/implications - To better deal with the technical and legal challenges of web scraping and to exploit its scalability, joint data collection projects in the field of agricultural and food economics should be considered.
Originality/value-In agricultural and food economics, web scraping as a data collection technique has received little attention. This is one of the first articles to address this topic with particular focus on food price analysis. Keywords E-commerce, Big Data, Data collection, Food price, Digitalization
Paper type Research paper
1. Introduction
Web scraping is a relatively new method for collecting online data. The term describes the automated process of accessing websites and downloading specific information, such as prices, from each (Kienle et al., 2004). Allowing the creation of large, customised data sets at low costs, web scraping is already applied for scientific and commercial purposes in many areas, such as marketing, industrial organisations, or inflation measurement (for an overview, see Cavallo and Rigobon, 2016; Edelman, 2012).
In food price research, however, this data collection technique has received little attention. In agricultural economics and food system analysis, we mostly rely on more traditional data sources, such official price indices or retail scanner data, of consumer prices. Yet, several issues are associated with these data sources.
are associated
publish
For example, official prices and price indices for products, segments such as food, or even the whole economy are mostly published on monthly or quarterly basis, with some publication delay. The public provision by official agencies and the availability of long time series are attractive for research purposes. Yet, one must rely on correct data collection, weighting and aggregation by official sources. Because, normally, no access to the raw data is given, it is not possible to detect errors or even manipulations (Cavallo, 2013). In comparison, scanner data obtained at the point of sale at retailers are available at a higher frequency

(generally weekly) and provide more details at the product level. A main advantage is that they include transaction data, i.e. the quantities purchased of a good at a given price (Campbell and Eden, 2014; Cotterill, 1994; Silver and Heravi, 2001). However, these data need to be purchased from market research institutes such as Nielsen N.V., and can be very costly, especially if longer time series or multiple retailers and locations are required.

As an increasing number of prices is published online and as online grocery retail is slowly gaining market shares in many parts of the world (Nielsen, 2015; Rigby, 2018), web scraping may be a promising alternative to get data for food price research.

In the following, we will not give detailed instructions on how to build a web scraper, and we will omit technical details and coding issues[1]. Rather, the aim is to discuss the method's potential for agricultural and food economics research. Section 2 gives an overview of what exactly web scraping is and how it works, and weighs the pros and cons regarding food price analysis. Section 3 reviews existing applications and considers further applications for studying online and offline food prices. The paper finishes with an outlook and suggestions on how this new data collection method could best be used in the discipline of agricultural and food economics.

2. About web scraping 2.1 Definition

Throughout this paper, we use the term web scraping. However, we found several related terms and concepts, which are not always distinctively defined (for definitions, see e.g. Kienle et al., et al., 2004; Massimino, 2016; Nakash et al., 2015).

As a minimal definition, web scraping (or screen scraping, information scraping) describes the automated process of accessing web documents and downloading specific, pre-defined information, such as prices, from each, to then transform and save them into a structured format.

Web crawling, on the other hand, means accessing web content and indexing it via hyperlinks; thus, only the URL but no specific information is extracted. Instead, the full content is made available through the hyperlink but is generally not archived. Search engines, including Google Google, crawl the web, analyse the online content and compile all the links they find to match the search request. Crawlers (or spiders) are also used for price comparison tools. Shopbots are programmes that crawl websites to obtain price information from several sellers in order to find the lowest price (Hemenway and Calishain, 2004).

For a conceptual distinction between scraping and crawling, see Table I. The remainder of this paper deals only with web scraping because we are interested in collecting food price data for research from public websites.

2.2 Technical procedure

There are several ways to build a web scraper, and probably no one-size-fits-all approach exists. Although this paper does not aim to give detailed instructions on how to code a web scraper, we will briefly give an intuitive description of what a web scraper does technically and how this tool can be implemented for creating food price data sets.

Generally speaking, one needs to write a script that accesses the websites hosting the data, finds the relevant, previously defined elements and then downloads, and stores them in structured data sets. In any case, the price, product name and a timestamp recording when the content was accessed need to be stored to ensure a consistent output over time. If available and desired, additional information, such as package size, customer rating,

Process
Target
Web scraping
Web crawling
Automatically requesting web documents and Repetitively finding and fetching hyperlinks collecting information from them starting from a list of initial URLS Pre-defined data on specific websites URLs to access all kinds of information, depending on search request
information Output
Use
Downloaded data in structured format Data collection (e.g. price series)
Source: Own representation
Indexed hyperlinks, stored in database Ad hoc requests (e.g. search engines, price comparison tools)

Table I. Distinction
between web scraping and web crawling