Web scraping as a source for producing e-commerce indicators: findings from a pilot in Brazil

Thiago Meireles (NIC.br)
Marcelo Trindade Pitta (NIC.br)
Pedro Luis do Nascimento Silva (ENCE)

Abstract

The Brazilian Network Information Centre (NIC.br) is a non-profit civil entity created in 2005 to implement the decisions and projects designed by the Brazilian Internet Steering Committee. Within NIC.br, the Regional Center for Studies on the Development of the Information Society (Cetic.br) is the branch that produces Information and Communication Technologies (ICT) statistics for policymaking. As part of its portfolio, Cetic.br conducts the ICT Enterprises annual survey of companies with 10+ employees operating in Brazil. This survey aims to measure presence and use of ICT in these companies, covering topics such as infrastructure, appropriation, and use of new technologies by the private sector, as well as perceptions regarding their potential benefits to their activities.

One of the main indicators produced regularly by the ICT Enterprises survey refers to e-commerce infrastructure and practices adopted by the companies obtained by self-declared responses to the so-called E module. Considering that e-commerce implies the use of web sites, some years ago a proposal was made to investigate the possibility of collecting this information directly from the web, thus reducing the response burden for companies in the sample. If successful, this approach would help reduce survey costs and perhaps improve response rates. Furthermore, this could enable production of more timely and disaggregated statistics.

We have been experimenting with this idea since 2017 and found that it is possible to automate data collection through web-scraping to classify a company's website in one of two categories: with e-commerce and without e-commerce. However, for Brazil, web scraping failed to produce e-commerce prevalence estimates matching those obtained from traditional survey-based statistics relying on self-declared indication that the company does offer e-commerce facilities.

An alternative approach would involve combining the data collected via questionnaires from the ICT Enterprises survey (designed data) with data scrapped from the web for the sampled companies (found data). This is now being explored and is part of our ongoing pilot for replacing part of the ICT Enterprise survey's E-module.

This paper describes the findings of this pilot, discussing the differences between the web universe and the enterprise sampling frame, the challenges faced when collecting data from the web for a specified probability sample of companies, and the model developed to reliably predict a site's e-commerce availability from the web-scrapped data.