September 2016

Statistics Singapore Newsletter

ISSN 0218-6810

www.singstat.gov.sg

Experiences with the Use of Online Prices in Consumer Price Index

By
Foo Chuanyang and Lee Wen Hao, Joseph
Prices Division
Singapore Department of Statistics

Introduction

The Singapore Department of Statistics (DOS) compiles the Consumer Price Index (CPI), which is designed to measure average price changes in a fixed basket of goods and services commonly purchased by resident households over time.

Price data used in the compilation of the CPI are collected through various modes. Prices of most goods and services such as utility tariffs, petrol, school fees etc. are obtained through websites or postal and email enquiries. Prices of perishable food items sold at wet markets are collected by field interviewers.

Notably, internet purchase among households has been growing in prevalence and importance. This is due to the increasing number of traditional brick and mortar retail stores marketing their products online, coupled with greater access by households to computing and mobile equipment.

To tap into this source of price information, DOS embarked on pilot projects to integrate specific online prices into the compilation of the CPI.

This article discusses the pilot work undertaken to include the Internet as an alternative data source, specifically involving the use of web crawlers – both customized web extraction crawlers and web scraping tools – to extract online prices.

Use of Online Price Information Collected Through Web Scraping

Using price information from the Internet is an efficient way to collect data that might otherwise be costly to collect or result in response burden.

Hence, online prices for items commonly purchased via the Internet such as apparels and travel products have progressively been integrated into the CPI.

Online prices may be collected manually by combing through lists of items from websites, extracting and compiling the required information.

This approach is tedious, repetitive and labour-intensive. To collect online data more efficiently, DOS explored the use of web crawlers to capture relevant information from the Internet.

In technical terms, a web crawler is an "e-robot" programmed according to pre-defined criteria to browse through designated URLs and collect specific information from these webpages.

There are two broad groups of web crawlers, viz. customised web crawlers and the "point-and-click" types.

(a) Customised Web Crawlers

These are developed by IT programmers and specially programmed to collect data points from the exact position of each monitored website.

They are more suitable for websites with greater complexity as the required data may not be arranged in a structured manner across the pages.

DOS experimented with the use of customised web crawlers to extract information from specific websites with greater complexity. For example, collection of airfares offered by low cost carriers which accounted for a high share of online purchases in Singapore.

The web crawlers used have a simple interface with user editable selections on the destination, dates of departure and return.

The web crawlers will crawl specified websites to retrieve the required fare information based on the selections made.

As the data required are embedded within different webpages of the entire flight booking process, the web crawlers are encoded to input the necessary information in a logical sequence with appropriate time intervals between actions. Figure 1 provides a screenshot of such a customised web crawler.

Time is saved with the use of the customised web crawlers in extracting the prices information, as they automate the entry of the required parameters and transforms these data into a structured format.

(b) "Point-and-Click" Web Crawlers

This type of web crawlers does not require any programming activities and is mostly available free-of-charge.

DOS explored the use of the free services of *Import.IO* to web scrape data from websites retailing home electronics and appliances, personal effects and pharmaceutical products. These websites usually display a comprehensive list of their products with up-to-date prices.

FIGURE 1 CUSTOMISED WEB CRAWLER FOR EXTRACTION OF PRICES ON AIRFARES

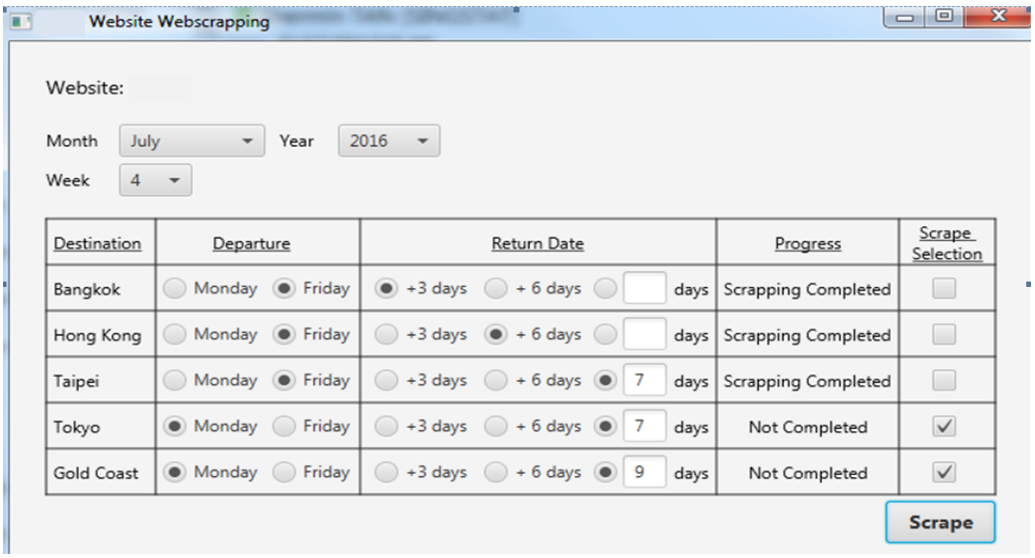
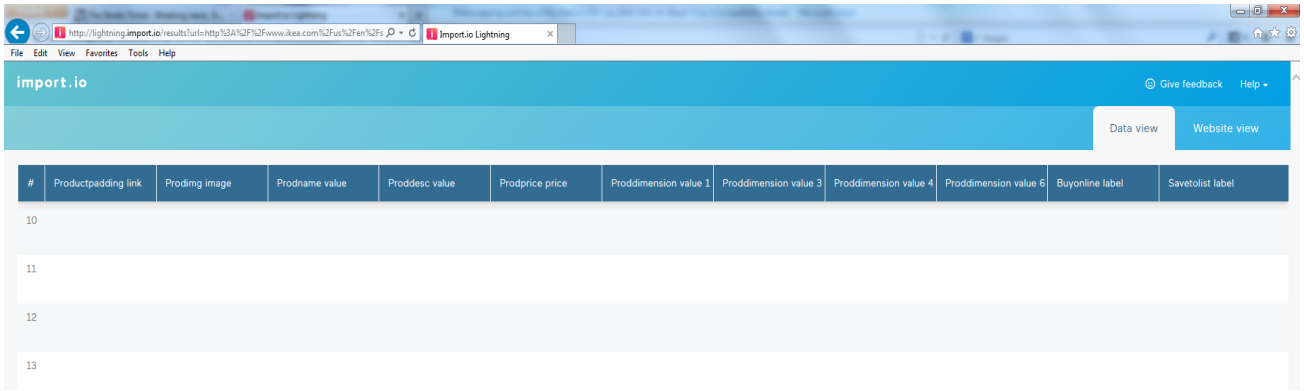


FIGURE 2 EXAMPLE OF A "POINT-AND-CLICK" WEB CRAWLER, *IMPORT.IO*



The "point-and-click" web crawler can be set up and developed in minutes for a website. For *Import.IO*, users have to "train" the tool to navigate through a designated website to locate and scrape the required data.

This involves simple "point-and-click" techniques to identify the required data, such as product description, item code and selling price, etc. and to define them neatly into respective rows and columns. Figure 2 shows a screenshot of the *Import.IO* web crawler.

Once the web crawler has been "trained" to navigate through a specific page of a website, it can be used to locate and scrape price information from other similar webpages on the same website. The time taken to extract the relevant price information from selected websites reduces substantially with the use of *Import.IO*.

Switching to Using Web Scraped Data

Before data obtained through web scraping are used in the compilation of the CPI, they are cross-checked to ensure the robustness and stability of the online prices.

For example, the data may be compared against those collected by field interviewers from the retail stores. The prices of majority of the items online and in physical retail stores are comparable. For some establishments, the prices shown online are aligned with prices in stores.

Key Learning Points

The use of web crawling technology allows a more efficient use of online prices in the compilation of CPI. The required data are retrieved off webpages and arranged in the required format automatically.

Nevertheless, there are also several issues which need to be considered.

(a) Consistency in Product Type

At the onset, price statisticians have to study the web scraped data extensively to match each product monitored in the CPI with the new data set that has been extracted online.

This requires scrutinising the entire web scraped product descriptions. When this is completed, the products are matched over time, using the mapped product descriptions. When there are revisions to the products' online descriptions, further reviews have to be made to the mapping.

(b) Expertise in Web-Programming

The use of customised web crawlers requires extensive programming knowledge and skills, as well as maintenance effort. As the page layout for each website is unique, web crawlers have to be specially developed for each website and this may be costly. Any subsequent changes to the webpages, such as the layout or design, may render the web crawler ineffective.

Compared to the customised web crawlers, the main advantage of using "point-and-click" web crawlers such as *Import.IO* is that they are easy to develop. This eliminates the need for any in-depth programming knowledge and skills.

They can hence be developed and maintained in-house by the team responsible for price collection. Such "point-and-click" web crawlers are available at very low or no costs to users.

However, the use of such web crawlers is better suited for simple websites with data arranged in a structured manner across the webpages.

Users of such tools are also highly dependent on the service continuity by the program developers. For instance, the developers may terminate the program or amend the terms and conditions of use without prior notice.

(c) Legal and Design Restrictions on Websites

Before scraping a website, it is important to review the terms and conditions of use of the websites and check against any legal restrictions imposed. Some establishments may explicitly prohibit the use of web crawlers on their websites.

For the pilot studies conducted by DOS, prior approvals were sought from the relevant establishments on the use of the web crawlers on their websites.

It is also worth noting that not all websites can be web scraped. Price information on certain websites is embedded in images and not stored as text.

This information will then become undetectable by web crawlers. In some cases, the webmasters may also set up blocking mechanisms on the websites to deter the use of web crawlers.

Other Initiatives — Use of Electronic Prices from Supermarkets

Supermarkets electronically store the prices of commodities in their database. This becomes a potential data source for DOS to tap upon.

Previously, prices were collected weekly by field interviewers via personal visits to the supermarkets. The list of items monitored from supermarkets is wide-ranging – spanning from perishable items and groceries to household appliances.

To facilitate the provision of electronic prices by the supermarkets, DOS specified the barcodes and included them in the data file sent to the supermarkets each month for their identification of the products required. These data files are encrypted with passwords to ensure data security during transmission.

The shift from traditional price collection by personal visits to electronic price data has resulted in a more efficient use of manpower. The electronic prices derived based on actual transactions are also more reflective of the monthly average price paid by consumers due to the increased number of price quotations. This improves the quality of data used for the compilation of the CPI.

Conclusion

With the prevalence of e-commerce among households, the use of online prices in the compilation of CPI becomes increasingly viable and DOS will continue to review and refine our methods.

The use of web crawling technology will also be further fine-tuned and expanded to reduce the overall workload for data collection. Evaluation of new IT developments and Smart Nation initiatives will also be carried out to ensure that DOS' data collection methods tap on the best possible approaches for efficiency and productivity.