Experiences with the Use of Online Prices in Singapore's Consumer Price Indices (CPI)

Foo Chuanyang*
Department of Statistics Singapore – foo_chuanyang@singstat.gov.sg

Lee Wen Hao, Joseph
Department of Statistics Singapore – joseph_lee@singstat.gov.sg

Abstract

Recognizing the potential use of online prices for official statistics, the Singapore Department of Statistics has embarked on a pilot project to integrate them into the compilation of Consumer Price Index (CPI). This paper starts by giving a general overview of the various modes of price collection adopted by the Department. It then details the pilot work undertaken to integrate online prices into the compilation of CPI, specifically involving the use of web scraping tools to extract prices efficiently from the Internet. The paper finally presents the experiences and learning points as well as other initiatives the Department has explored.

Keywords: crawlers; data extraction; web scraping;

1. Introduction

The price data used in the compilation of the Singapore Consumer Price Index (CPI) are obtained through a combination of data collection modes. Prices of most goods and services such as utility tariffs, petrol, school fees etc. are obtained through postal/email enquiries or websites while those of perishable food items sold at wet markets are collected by field interviewers.

As increasing number of traditional brick and mortar retail stores leverage on the online platform to market their products, coupled with greater access by households to computing and mobile equipment, households' consumption habits are observed to be shifting in recent years. More households in Singapore are making purchases over the Internet[1]. Based on the Infocomm Development Authority of Singapore's (IDA) Annual Survey[2] On Infocomm Usage In Households And By Individuals, home Internet and Broadband access rates were around 88 per cent each in 2015, about 3 – 4 percentage points higher than that of 2012. The number of online shoppers[3] in 2014 was about 1.44 million, and this has increased by a compound annual growth rate of about 14 per cent from 2012.

Given the growing prevalence and importance of internet purchase among households, as well as the availability of price data online which mirrors that of usual consumer retailers, the Singapore Department of Statistics (DOS) has embarked on pilot projects to integrate specific online prices into the compilation of the CPI. This paper discusses the pilot work undertaken to use Internet as an alternative data source for the compilation of CPI. Specifically, it details the exploratory use of customized web extraction crawlers and web scraping tools to extract prices from the Internet. The paper

[1] Based on the latest Household Expenditure Survey (HES) 2012/3, almost one-third of the households made at least one purchase over the internet.

[2] https://www.imda.gov.sg/industry-development/facts-and-figures/infocomm-usage-households-and-individuals

[3] https://www.imda.gov.sg/~/media/imda/files/industry%20development/fact%20and%20figures/infocomm%20survey%20reports/0301%202014%20hh%20public%20report%20final.pdf?la=en

finally presents the experiences and learning points from these projects, as well as other initiatives the Department has explored.

2. Use of Online Prices in the Singapore's Consumer Price Index (CPI)

The growing online retailing and availability of price information on the Internet have provided DOS an opportunity to re-examine the current price collection approach and explore the use of online prices to support the compilation of the CPI. Using price information from the Internet is an efficient way to collect data that might otherwise be costly to collect (e.g. via personal visit) or involve response burden. Hence, online prices for items commonly purchased via the Internet such as apparels, travel products (e.g. air tickets, accommodation), and cinema tickets have progressively been integrated into the Singapore's CPI.

The manual collection of online prices is, however, tedious, repetitive and labour intensive. It involves the price collectors combing through the list of items from the website to extract the required information before data entering and formatting them structurally onto spreadsheets. The collection of prices and product attributes is time consuming, especially for those websites where large amount of information is to be extracted.

In recent years, there have been extensive developments in the use of web scraping tools to capture information from the Internet. A web crawler (also termed as web scraper) is technically an "e-robot" programmed to browse through designated URLs and collect specific information from these webpages automatically according to some pre-defined criteria. It offers the opportunity to replace manual effort in price collection from websites by automating the repetitive task of extracting price information from the same website each month. National statistical offices such as the Netherlands, Austria and United Kingdom are also using such web scraping techniques for their price collection in their CPI.

Web crawlers can be generally classified into two broad groups. One group is the customised web crawlers which are developed by IT programmers and specially programmed to collect specific data points from the exact position of each monitored website. The second group is the "point-and-click" type of data crawlers which do not require any programming activities and are mostly free-of-charge to users.

In order to integrate more online prices in the Singapore CPI and make online data collection efficient, DOS has been exploring various types of web crawlers available. More details are shared in the next section.

3. Web scraping of Online Prices

(a) Customised web crawlers

Customised web crawlers are more suitable for websites with greater complexity as the data may not be arranged in a structured manner across the pages. DOS has thus embarked on a pilot project to develop customised web crawlers to extract information from specific websites e.g. airfares for low cost carriers which accounted for a high share of online purchases in Singapore, and whose websites are more complex.

Airfares for low cost carriers differ according to flight destination, departure/arrival dates, number of passengers and type of booking class, etc. Thus, the use of customised web crawlers for extracting low cost carrier fares is due to the need to input such parameters in order to simulate actual online purchasing and reproduce a specific purchasing behaviour every month. In addition, price information such as base fare, taxes, baggage add-on fee, convenience fee etc., may not be arranged in a structured manner across the airline webpages. Such customisation of user specifications renders the "point-and-click" type of web crawlers unsuitable for the extraction of airfares.

The customised web crawlers to scrape airfares for low cost carriers are written in Java computing language due to its flexibility and extensive library of solutions available. The web crawlers are designed with a simple interface with user editable selections on the destination, date of departure and return. Prior to the extraction, users need to input these parameters in order for the web crawlers to crawl the specified websites and retrieve the required fare information. Other parameters such as type of booking class and number of passengers travelling are hard-coded in the program as these are price determinants that are kept constant each month. On the interface, there is also a column to indicate the status of the web scrapping progress.

As the data required are embedded on different webpages of the entire booking process, the web crawlers are encoded to input the necessary information in a logical sequence with appropriate time intervals between actions. This caters sufficient time for the websites to load to the next page such that all the information on the subsequent pages can be captured. It also helps to minimise the load of the web crawlers on the respondents' website server.

The use of the customised web crawler resulted in savings in terms of time required to extract the prices, given that it automates the entry of the required parameters such as destination, departure/arrival dates, etc., (which will have to be repeated for every single destination) and transforms these data into structured format. Its coverage is also more comprehensive, thereby improving data accuracy with increased frequency of price collection.

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

Besides the customised web crawlers, DOS also experimented with the use of the freewares to web scrape data from the Internet. These "point-and-click" web crawlers are relatively straightforward to set up and can be developed in minutes for each website. The extraction is easy and does not require any programming skills. Once the web crawler has been developed for a particular website, the program can be saved and recycled for subsequent data extractions and downloading continuously.

For the pilot project, DOS tested on websites retailing home electronics and appliances as well as those retailing personal effects and pharmaceutical products as these websites usually display a comprehensive list of their products with up-to-date prices. The commodities are also organised in a structured manner by relevant categories such as brands/varieties, thus allowing the web crawler to identify accurately the data cells to be extracted. Unlike websites for low cost carriers, these websites also do not require further user input to simulate actual online purchasing.

To build a web scraper, the first step requires users to "train" the web crawler to navigate through a given website and how to locate and scrape the required data. It involves simple "point-and-click" techniques to identify the required data of interest, such as product description, item code and selling price, etc. and to define them neatly into respective rows and columns.

Once the crawler has been "trained" for one of the pages, it can be extended to other similar pages of the same website. The extracted data are stored on cloud servers and can be downloaded in various file formats for further processing and data analysis. With the use of these "point-and-click" web crawlers, it is observed that the time taken to extract the relevant price information from the selected websites is reduced quite substantially.

4. Switching to Web Scraped Data

Before the web scraped data are used in the compilation of the CPI, they are cross-checked with those obtained from personal visits to ensure the robustness and stability of the online prices.

Though the actual price level of the commodities may differ slightly between the two sources due to factors such as marketing strategies, the results showed that majority of the items have comparable price movements between online and physical retail stores. In fact, for some establishments, their website prices are aligned with their store prices.

5. Learning Points/Experiences

The use of web crawling technology has allowed more efficient use of online prices in the compilation of CPI. Nevertheless, there are also several issues which need to be considered.

(a) Consistency in Product Type Between Online Pricing and Field Collected Pricing

At the onset, price statisticians have to study the web scraped data extensively to match each product monitored in CPI with the new data set that was extracted. This required scrutinising all the web scraped product descriptions. When this has been 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 which requires significant amount of time and effort.

As compared to traditional price collection where field interviewers are able to confirm on the reasons for price changes with store assistants, e.g. store clearance, promotions, etc., the data obtained via web scraping do not provide such information. In order to reflect only pure price change in the CPI, officers still need to either call the store to verify or review their websites to check on possible reasons for the changes in price, e.g. anniversary sale, weekly online specials, etc.

(b) Expertise in Web-Programming to Manage Frequent Website Changes

The use of customised web crawlers also requires extensive programming knowledge and skills as well as maintenance. As the page layout for each website is unique, the web crawlers have to be specially developed for each website and this may be costly. Likewise, frequent changes in website designs require re-programming of the respective web crawlers. As the web crawler is specifically designed to scrape the data from the exact position of each web page, any amendments to site layout or design may render the web crawler ineffective. Should the website undergo a revamp e.g. change in website in terms of structure or configuration, the web crawler may need to be re-coded entirely and this process takes time. Thus, the use of these customised web crawlers is subjected to the continued non-revision / consistency of the website layout; otherwise there will be missing prices for a period of time. Given the high development and maintenance costs to develop, test and maintain these customised web scraping programs, the use of such crawlers may be more appropriate for websites which require inputting of detailed information to simulate actual online purchasing and where large amount of information has to be extracted each time.

Compared to customised web crawlers, the main advantage of using the "point-and-click" type of web crawlers is that they are easy to develop and eliminate the need for any in-depth programming knowledge and skills. These web crawlers are easily understood by non-IT personnel and hence, can be developed and maintained by price collection team themselves. No programming skills are needed to perform changes to the web crawling programs in case of website changes. Such "point-and-click" type of crawlers are also available at very low or no costs to users.

Nonetheless, it should also be noted that these readily available “point-and-click” web crawlers are better suited for simple websites with data arranged in a structured manner across the webpages and may not be appropriate for those which require customisation of user specifications. The use of such crawlers is also dependent on the service continuity by the program developers. The developers may terminate the program or amend the terms and conditions of use without prior notice. Functionality can also be changed or discontinued. During our pilot study on the use of these freewares, one of the service provider discontinued the downloading of the existing desktop application. As a result, the web crawlers had to be re-created for all monitored websites via the web version.

(c) Legal and Design Restrictions on Websites

Before proceeding to web scrape, it is also important to review the terms and conditions of use of the websites and check against any legal restrictions imposed. Some establishments may state explicitly the prohibited use of such crawlers on their websites. For our pilot studies, prior approvals were sought from various establishments on the use of the web crawlers. This also allows us to maintain a good working relationship with our stakeholders as further clarifications on the extracted data may be required from them during our data analysis stage.

Not all websites can be web scraped. For some websites, the price information is embedded in images and not stored as text. For such cases, the data are not detectable by web crawlers. In some cases, web masters also set up blocking mechanisms to deter the use of web crawlers and even falsify the data scraped when web crawlers are detected. There is also a need to explore different web scrapers to extract the required data from different websites efficiently. For e.g. some websites may display more products based on users’ continued scrolling of the page, thus, only a web scraper with “infinite scrolling” data extraction feature is suitable.

The use of online prices and web scraping tools are restricted to establishments that have an online presence. It rules out those establishments which have large market shares but do not provide online purchases. There is also a lack of information to identify the popular items e.g. total quantity sold, shelf space allocated to the items, etc.

6. Other Initiatives – Use of Electronic Prices from Supermarkets

Apart from piloting the use of web scraping techniques to capture information from the Internet, DOS has also looked into another potential data source, i.e. electronic prices from supermarkets for the compilation of CPI. As supermarkets capture the prices of commodities electronically in their database, DOS has identified this as a potential data source to tap upon.

Prior to this, prices were collected by field interviewers via personal visits to the supermarkets on a weekly basis. The list of items monitored from supermarkets is wide-ranging, spanning from perishable items to groceries to household appliances. As supermarkets tend to offer products with various packaging sizes, field interviewers had to spend much time scrutinizing the products’ descriptions and packaging to ensure that the comparable prices for the same products were collected each time.

To seek collaboration from the major supermarket chains to provide the electronic prices, DOS held meetings with them to explain how their electronic price data would be used for the compilation of the CPI. The meetings also addressed concerns or difficulties they might encounter in the provision of electronic price data, e.g. some supermarket chains were concerned about the long list of products required, the stipulated timeline for monthly submission, as well as the security of data transmission from the supermarkets to DOS.

Taking in these concerns, barcodes were specially collected by DOS and included in the data file sent to the supermarkets each month to facilitate their identification of the products required by us instead of a universal list of items available in the respondents' database. Data files are also 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, given that the reduction in manpower required for fieldwork can be channelled to other areas of work such as data checking and verification. The electronic prices which are derived based on actual transactions are also more reflective of the monthly average price paid by consumers with increased number of price quotations, thereby improving data quality for the compilation of the CPI.

7. Conclusions

The availability of reliable on-line prices which mirrors traditional retail shops and technological tools has reduced the reliance on field collection of prices. New alternatives including electronic price data and barcodes-scanner data provide a comprehensive range of data collection approaches to make CPI compilation more efficient and less labour intensive.

Looking forward, e-commerce will become more prevalent among households. The use of web crawling technology will be further fine-tuned and expanded to reduce the overall workload for data collection. DOS will continue to review the use of online prices in the compilation of CPI.

References

[1] Boettcher, Ingolf. 2015. "Automatic Data Collection on the Internet." Statistics Austria, Working Paper. http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s2p6_pap.pdf

[2] Breton, R., G. Clews, L. Metcalfe, N. Milliken, C. Payne, J. Winton, and A. Woods (2015). Research indices using web scraped data. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_UK_Research_indices_using_web_scraped_data.pdf

[3] Griffioen, R., Bosch, O. 2016. "ON THE USE OF INTERNET DATA FOR THE DUTCH CPI" Statistics Netherlands. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_Netherlands_on_the_use_of_internet_data_for_the_Dutch_CPI.pdf

[4] Import.io. https://www.import.io/

[5] Nygaard, R. (2015). The use of online prices in the Norwegian Consumer Price Index. Statistics Norway. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/d012f001b8a1cf6cca257eed008074c9/$FILE/Ragnhild%20Nygaard%20%28Statistics%20Norway-%20The%20use%20of%20online%20prices%20in%20the%20Norwegian%20Consumer%20Price%20Index.pdf

[6] Octoparse. http://www.octoparse.com/