論述 》統計 · 調查

網路耙梳對於消費者物價調查之助益與限制

隨網路科技發展，民衆透過網路消費漸增，致近年來電子商務銷售規模快速成長，其交易價格代表性亦漸提高，惟受限人力及調查經費難以增加，各國開始嘗試利用網路耙梳（Web Scraping）蒐集用於編算物價指數所需之價格。本文旨在探討目前網路耙梳技術對於物價調查所能提供的助益及限制，並期借鏡他國實務做法與經驗，增進我國後續推動效益。

許榮洲、周治良（行政院主計總處綜合統計處專員、科員）

壹、前言

隨著網路科技的快速發展與線上支付機制的日趨健全，電子商務一躍千里，根據 eMarketer 估計，2015 年以來全球電子商務市場每年均以雙位數的幅度快速成長，預期 2020 年將達 4.1 兆美元，占整體零售市場 14.6%（下頁圖 1），影響所及，電子商務交易價格之代表性，及其對物價調查品質之重要性均與日俱增，有必要增加網路價格資料之蒐集。

惟近年受查者之資料安全意識抬頭，拒查情形愈見顯著，物價調查資料之應用需求卻不斷升高，加以消費者物價指數（Consumer Price Index，以下簡稱 CPI）調查隨著商品種類的多樣化與業者因競爭致價格優惠更迭頻繁，均加重調查人力及預算之負荷；與此同時，網購市場的蓬勃發展，帶動比價網之興起，透過網路耙梳程式自動化處理，能快速找到全國主要購物網站或電商平台上，最即時完整之商品資訊，成爲各國 CPI 調查納查電子商務價格的最佳選擇，而得以在成本極小化的情況下，快速蒐集價格資料的網路耙梳程式應用，也因此成爲各國傾力研究精進的焦點。然網路耙梳爲全新技術，欲臻成熟應用必經學

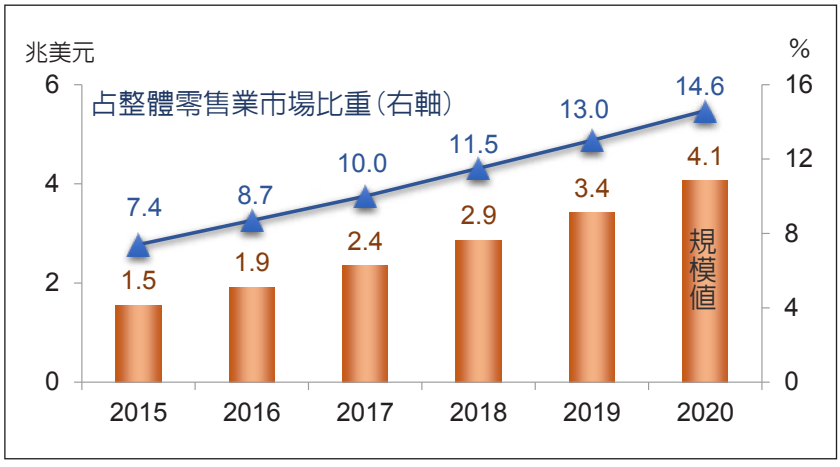
習陡坡，本文旨在探討引入網路耙梳技術對於 CPI 所能提供的助益及其限制，並期借鏡他國實務做法與經驗，爲我國後續研究發展扎根。

貳、CPI 調查流程

CPI 旨在衡量一般家庭購買消費性商品及服務價格的變化，爲如實反映整體 CPI 的價格變動，CPI 調查資料從蒐集到彙編指數發布，皆須經過縣市及行政院主計總處的嚴格審核（圖 2），以確保資料的正確性。

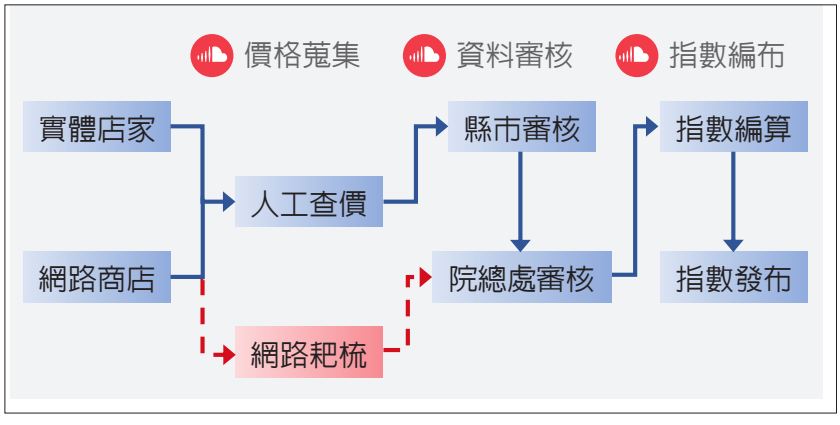
依據經濟部統計處發布之「批發、零售及餐飲業經營實況調查報告」，我國 2018 年電子商務占整體零售業營收爲 6.9%，較 2013 年 2.6%明顯提升，顯示隨行動裝置的普及與 4G 發展，電子商務已漸成爲民眾重要的消費管道，網路商店商品之價格成爲 CPI 不可或缺的資訊。我國自 2013 年起，CPI 部分調查品項之查價點已涵蓋網路商店，並逐年增加調查品項，同期間實體商店之調查品項亦持續擴增，致人力負荷不斷加重。鑑於國際發展之網路耙梳技術，可提升樣本數量，並減少人工查價負荷，行政院主計總處爰於去（2019）年開始試行，並逐步引進（如圖 2 虛線流程部分）。

圖 1　2015 年－ 2020 年全球電子商務市場概況



資料來源：eMarketer Inc.（2019）。

圖 2　編算 CPI 之資料流程圖



說　　明：實線為人工查價之資料流程，虛線為網路耙梳之資料流程。
資料來源：作者自行繪製。

參、網路耙梳作業

什麼是網路耙梳？簡單來說爲利用程式模擬人類瀏覽網頁的行爲，包含搜尋、點擊、選取、複製及貼上等動作，以達成資料蒐集、過濾、整理及儲存之目標，俾利後續應用。運用耙梳取得資料之手法略分兩種：

一、循網址耙梳

將瀏覽已知商品網址並蒐集資料的動作，改爲程式自動化耙梳，因實際應用較爲簡單，故各國在研究報告中著墨不多。

二、循目錄耙梳

在網路商店的目錄逐層掃描並耙梳所有相關商品資料（如商品名稱、重容量及價格等），程序需客製化，且資料品質不易確保，致難度較高，惟因具完整性（商店全部產品而非僅部分）及低成本（非人工，以程式自動耙梳，調查成本低）的特性，目前主要國家多已投入此領域的發展研究。

肆、主要國家發展現況

近年各國網路耙梳發展受限於資料特性及法律等因素，皆先針對特定品項進行探討及研究，以下爲主要國家發展近況：

一、英國

自 2014 年開始研究網路耙梳，實務經驗與挑戰如下：

（一）資料儲存：硬體設施須具高容量與高可靠度之儲存媒介，且須及時回應各類程式中斷（如網頁改版）訊息，以達成耙梳資料之完整性。

（二）資料分類：網頁目錄結構與 CPI 分類不同，故須建立分類機制，將儲存之資料歸至正確的 CPI 項目，如耙到「蘋果」關鍵字，若爲「蘋果 iphone」應歸手機，若是「太妃蘋果糖」應歸糖果，若是「富士蘋果」則歸生鮮蘋果，3 種分類皆不同，然由於商品推陳出新，其名稱的文字組合千變萬化，因此目前仍難確保歸類完全正確。

（三）資料清洗：若因尺寸不全（即將過期）導致價格大幅調降，則應予排除；另離群價格之品項若爲新增，也應檢視歸類是否正確。

（四）資料設算：因商品下架、沒有庫存等產生之缺漏值應予設算。

英國規劃「循目錄耙梳」資料最快在 2023 年 1 月起，逐年分批導入 CPI，初期先進行資訊產品（筆電、桌機、手機及平板等）、影音光碟（CD/DVD/ 藍光）、書籍、旅遊團費及衣服等，後續則有機票及鞋類等，其中除機票及旅遊團費爲每日耙梳外，餘爲每週。

二、美國

受限於美國法律嚴格保障消費者權益及企業資料安全，「循目錄耙梳」的研究，主要鎖定油料費，主因美國零售油品在自由競爭及州稅不同的情況下，價格走勢並不一致，需大量查價點（2017 年 12 月調查 1,332 個）方能如實反映全國城市平均價格走勢，加以其具性質單純及容易分類特性，易於進行「循目錄耙梳」所致。在取得廠商（如 GasBuddy.com）同意的前提下，美國勞工統計局（Bureau of Labor Statistics, BLS）將 2017 年 11 月至 2018 年 10 月「循目錄耙梳」資料採各查價點等權方式試算油料費指數，發現與現有 CPI 人工查價之指數結果差異不大，因此規劃 2020 年或 2021 年起，CPI 油料費價格資料將改以網路耙梳取得爲主。

三、日本

日本統計局於 2016 年 8 月至 2017 年 3 月研究耙梳家庭常在網路上購買的 94 項商品，與旅館住宿費、機票及旅遊團費等 3 項服務價格，並同步與人工蒐集資料進行比較，目前資料清洗上仍存疑慮，已投入研處，並規劃自 2020 年基期起能就前述 3 項服務全面改採網路耙梳資料計算，原因包括：

（一）網路消費者眾，且資訊完整。

（二）網路銷售價格與實體通路價格相同或走勢一致。

（三）針對旅遊專業網站耙梳，品項相對單純，

附表　各國網路耙梳價格在 CPI 之應用情形

國家	CPI 已採用 / 預計採用項目	研究中項目
英國	預計 2023 年起逐年採用：資訊產品、影音光碟、書籍、旅遊團費、衣服。	行李箱、腳踏墊、地毯、機票、珠寶、個人隨身用品、運動用品、文具及鞋類等。
美國	預計 2020 年或 2021 年採用：油料費。	機票（主要仍透過企業嫁接取得之資料為主，僅少部分採網路耙梳）。
日本	預計 2021 年 7 月（2020 年基期）採用：國外旅行團費、機票、旅館住宿費。	食品（如米）、藥品（如感冒藥）、日用品（如殺蟲劑）。
荷蘭	已採用：衣服。	機票。
挪威	已採用：機票、牙醫醫療費。	電子產品、個人照護產品。
義大利	已採用：電子產品、火車票。	機票。
比利時	已採用：跨國火車旅行、電子遊戲、鞋類。	衣服、旅館住宿費、機票、電子產品、藥妝類、書籍、影音光碟、學校住宿費等。
奧地利	已採用：機票。	火車票、旅遊團費、電子產品、衣服、旅館住宿費。
澳洲	未公開。	服裝、鞋類。
紐西蘭	未公開。	機票、書籍、音樂與電影（含下載與串流服務）、運費、通訊服務、教育、食品等。

資料來源：作者自行整理。

程式維護成本相對一般網路商店低，可持續、穩定及有效率地耙梳價格資料。

四、其他國家

近年除英國、美國與日本外，荷蘭、挪威、義大利、比利時及澳洲等亦相繼發表相關研究，主要項目如上頁附表所示。

伍、我國發展現況

行政院主計總處自 2019 年 3 月起，自行研究撰寫 Python 耙梳程式，針對國內主要大型網路商店商品進行查價，並於同年 5 月起與人工調查結果進行比對，雖「循網址耙梳」未涉及資料清洗，資料量亦較小，相較「循目錄耙梳」較易處理，仍經不斷試誤及修改程式，方取得一致性結果。自同年 10 月起，除衣著類常因尺寸不完整而特價（非屬同質產品），機票及旅遊團費亦常因無座位或關團而無價格資料等，而須「清洗」或「設算」等，致仍維持人工查價外，其餘網路商店商品皆改以網路耙梳蒐集價格。

於耙梳與人工查價並行過程中，發現前者常因網路商店修改網頁撰寫語法，導致資料可能產生誤判，因此除原有的價格審核機制（如第 69 頁圖 2）外，另納入每月抽點網頁的價格驗證機制，以確保耙梳價格之正確。

至於目前正投入研究的「循目錄耙梳」，係參考各國實務經驗，先就各界較關注之品項，採關鍵字查詢方式耙梳網路商店所有相關產品後，加以整理、分析，進而引用至 CPI 編算中，目前仍處試誤階段。以下係以關鍵字「衛生紙」耙梳後，初步遭遇的困難：

一、資料儲存

經耙梳國內主要電子商城，資料近 1 萬筆，程式執行時間約 1 小時，加上後續對價格資料的處理及納入 CPI 編算程序等，以現有的軟、硬體設備，以及程式維護之成本，僅能先行針對少數品項進行耙梳。

二、資料清洗與分類

部分耙梳結果因不符合「CPI 衛生紙、面紙及紙巾」範疇，清洗後可能會將其捨棄（因代表性不足，非 CPI 納查項目，如衛生紙盒、架）或重新歸類至適當的 CPI 品項（如廚房紙巾屬於「CPI 其他家用品」），惟因筆數過多，人工逐筆過濾作業繁重，加以商品文字組合千變萬化，需建立清洗及分類之標準化作業。

三、資料設算

當月已下架（或已無庫存）之商品，其價格之設算。

陸、結論與展望

網路耙梳可提升 CPI 網路查價的效率，也是有限預算下增加樣本的務實作法，目前規劃 3 階段目標：

一、短期目標

除衣著、機票及旅遊團費等外，網路商店查價品項皆改採「循網址耙梳」蒐集價格，惟須落實審核管控機制，以確保資料品質。

二、中期目標

汲取各國實務經驗，跨入「循目錄耙梳」領域，並培訓網路耙梳程式撰寫人力。

三、長期目標

建立各品項耙梳資料之清洗、分類機制，及優化耙梳程式，並跟隨各國發展腳步，在能力範圍內最大化「循目錄耙梳」領域。

就前述 3 階段觀察，目前甫完成短期目標，相關作業仍隨時檢討，期能兼顧 CPI 品質及作業流程優化；至於剛踏入的中期目標，則須經由不斷的研究改正，持續精進各項實務技術及作業。易言之，網路耙梳乃價格蒐集之利器，然水能載舟亦能覆舟，尤其已跨入「循目錄耙梳」領域的各國，在實務上或法律上皆遇到許多問題，因此我國未來須更謹慎持舵，以充分發揮此一技術之最大功能。

參考文獻

1. 經濟部統計處（民 108），批發、零售及餐飲業經營實況調查報告。
2. 經濟部統計處（民 103），商業經營實況調查報告。
3. 日本總務省統計局（2015-2019），第 5 次、第 8 次及第 10-14 次物價指數研究會，https://www.stat.go.jp/info/kenkyu/cpi/index.html。
4. 日本總務省統計局（2015），第 59 次、第 60 次服務業及企業統計部會，http://www.soumu.go.jp/main_sosiki/singi/toukei/kigyou/kigyou.html。
5. eMarketer Inc.（2019）, Retail Ecommerce Sales Worldwide, 2015-2020.
6. Konny（2019）etc., Big Data in the U.S. Consumer price Index: Experiences & Plans.
7. Office for National Statistics（2019）, Using alternative data sources in consumer price indices: May 2019.
8. UNECE（11-13 September 2019）, Regional Workshop on Consumer Price Indices, http://www.unece.org/index.php?id=51417.
9. UNECE（7-9 May 2018）, Meeting of the Group of Experts on Consumer Price Indices, http://www.unece.org/index.php?id=46772
10. Loon and Roels（May 2018）, Integrating big data in the Belgian CPI.
11. Office for National Statistics（2017）, Research indices using web scraped price data: August 2017 update.
12. Singapore Department of Statistics（2016）, Experiences with the Use of Online Prices in Consumer Price Index.❖