論述 》統計 · 調查

運用網路耙梳蒐集 CPI 資料之國際經驗

運用網路耙梳技術自動蒐集網路資訊，是政府統計大數據應用之可能趨勢，其資料取得方式及結構特性與傳統政府統計差異頗大，本文簡述歐洲主要國家將其運用在消費者物價指數的作業狀況。

鄭永白、龍運濤（行政院主計總處綜合統計處專員、研究員）

壹、前言

網路耙梳（Web Scraping）係利用電腦軟體技術（也稱爲網路爬蟲（Web Crawler）或網路蜘蛛（Web Spider）），以超文本傳輸協定[1]（HTTP, Hypertext Transfer Protocol）或嵌入網頁瀏覽器，取得非結構或半結構化網站內容並轉爲可資運用的結構化數據，許多知名的網路搜尋引擎皆運用此項技術建置各種服務功能，如網路比價及網站內容異動偵測等。

由於政府統計調查環境日益艱困，拒答或無回應情形漸增，在有限預算與人力下，各國莫不尋思可行替代方案，取得所需資料。根據聯合國統計委員會於 2015 年針對國際組織及各國統計機構所作的大數據調查報告（Report of the Big Data Survey 2015），已運用大數據進行政府統計專案研究中，資料採網路耙梳技術取得者達 31 件（占 17.5％），僅次於手機資料（占 23.7％），另 OECD 會員國更有高達 91％的國家，在大數據研究專案中已使用或考慮採用網路耙梳資料作爲資料來源，足見網路耙梳技術在政府統計大數據運用計畫中，頗受青睞。

政府統計運用網路耙梳工具的面向，包括企業活動概況、住宅興建許可與銷售統計、人力職缺統計、犯罪統計及物價統計等，其中又以物價統計的運用廣受各國統計機構的關注。本文彙整歐洲主要國家使用網路耙梳技術擷取網路商品價格並試行消費者物價指數（CPI）編算的應用概況，以瞭解其發展及待克服議題。

貳、於物價統計之應用發展

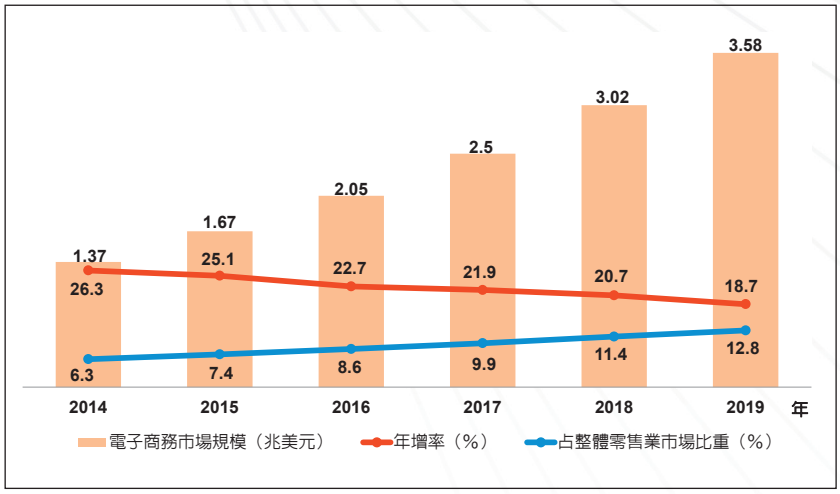
一、發展背景

隨著科技進步及數位內容蓬勃發展，商品交易方式產生巨大改變，也讓網購市場規模加速擴大，根據 eMarketer 估計，去（2015）年全球零售市場規模超過 22 兆美元，其中網路購物爲 1.7 兆美元，較 2014 年增加 25.1%，占整體零售市場 7.4%，今年起電子商務市場成長幅度雖可能趨緩，但整體占比仍持續提升，至 2019 年將達 12.8%（規模 3.6 兆美元，圖 1）。

由於網路商店經營成本與實體店舖不同，商品售價差異顯著，有效掌握網路商品價格變動之需求因而漸增，加以運用資訊技術快速取得價格資料，除節省調查成本外，亦可加速物價指數更新週期，提升編製效益。因此，近年來國外統計機構已著手研究利用網路耙梳自動蒐集以取代人工查價。

二、發展概況

圖 1　2014-2019 年全球電子商務市場概況



資料來源：eMarketer Inc.（2015）。

歐洲主要國家已先後投入相關研究（下頁附表），其中荷蘭、德國及英國等均以長期推展方式持續進行中，從其初步研究成果觀之，均肯定透過網路耙梳物價資料計算 CPI 具未來發展性，惟各國國情及運用規劃不盡相同，除相關研究內涵分歧外，指數編算方法尚未見深論，爰此，各國編算結果本文將不予探討，僅就查價項目及耙梳工具略作比較。

（一）查價項目

以英國爲例，查價對象考量市占率以三大超市網　站（Tesco, Sainsbury, Waitrose）爲主，查價項目包括食品、飲料及菸草類等 35 個 CPI 項目；而奧地利以原本利用人工方式至網路查價之商品及服務爲對象，預計取代約 5%的查價項目，主要是交通費（例如：機票、火車票及旅遊行程）、飯店住宿費、科技產品及衣著鞋類等。荷蘭在項目選取作業則有所不同，例如考量服飾產品在 CPI 編算作業上較具特殊性，向來具有向下偏誤的現象，傳統上以特徵迴歸　法（hedonic regression）來處理，利用網路耙梳商品的特徵，可以減輕人工蒐集的負擔。由以上情形觀察，多數國家雖傾向擇選網路訂購比率較高的交通票券，但

論述 》統計 · 調查

亦納入傳統指數編算上較待改善的電子產品及服飾等項目。

（二）耙梳工具

各國統計機構的資訊環境與資源多寡不一，所採用的網路耙梳查價程序及應用工具也有不同的思維，英國每天早上 5 時自動執行網路耙梳作業，約蒐集 6,500 筆商品價格資料，內容包括商品名稱、價格及折扣或贈品等相關資訊，資訊工具以 Python 開發；德國及義大利結合網路耙梳軟體 iMacros 及自行開發 Java 程式蒐集郵購公司、交通旅遊及消費性電子產品等商品價格；荷蘭爲降低人工查價成本及減少廠商受查負擔，採用 R 軟體開發網路爬蟲，蒐集特定商品（如電影票）價格。

奧地利統計局作法則截然不同，以低成本、高彈性、簡單易用、無須撰寫程式，並符合資訊安全規範爲條件，擇選視覺化網路耙梳工具 import.io 作爲自動化蒐集價格工具，甚至未來因應網站內容改變所需之系統維護，亦以不須撰寫程式爲前提，期大幅降低資訊技術門檻，打造適合物價統計部門執行網路耙梳作業之環境。據悉英國未來也將使用 import.io，以簡化開發流程及提高效率，筆者也嘗試利用此工具就 Yahoo 購物中心網頁，蒐集手機價格（下頁圖 2），彈指間即可取回價格資料加以處理，確實相當適合非資訊專業之統計人員應用。

儘管運用網路耙梳蒐集價格資訊簡捷易行，然現行技術仍有無法直接取得部分網站資訊之障礙及所獲得資訊須再加以清理分析等難題；此外，自行開發或使用既有網路耙梳工具，孰優孰劣？亦須就應用規模、資訊環境及後續維運成本等因素深入研析，方能找出最適方案。

附表　各國網路耙梳查價應用概況表（依發展期程排序）

國　別	發展期程	查價項目	耙梳工具
荷蘭	2012 年迄今	服飾、機票、電視、電影票等	R 軟體
德國	2012 年迄今	郵購公司（僅 1 家）、郵購藥局、租車、鐵路旅遊及市區觀光等	iMacros 及 Java
義大利	2013 年及 2014 年	消費性電子產品及機票等	iMacros 及 Java
英國	2014 年 6 月迄今	食品、飲料及菸草類等 35 個 CPI 項目	Python
奧地利	2015 年迄今	交通費（例如：機票、火車票及旅遊行程）、飯店住宿費、科技產品及衣著鞋類等	import.io

資料來源：作者自行整理[2]。

參、網路耙梳物價資料運用之挑戰

雖然各國普遍對運用網路耙梳資料編算 CPI 部分項目深

具信心，然仍有部分議題亟待克服，例如：

一、指數編算方法仍待探索

雖然網路耙梳可以大幅增加價格蒐集項目及頻率，讓價格資訊涵蓋更爲完整（如英國之威士忌，傳統方式每月蒐集 140 個價格，網路耙梳可增至 6,000 個），也可有效縮減人力，提升查價效率（如奧地利之機票價格查價作業所需花費的時間由 16 小時縮短至 2 小時）；惟因資料項目、時點等內涵變異甚大，有關指數編算方法及公式，並與傳統查價資料適切整併等，均有必要再深入探討。

二、資料蒐集、清理技術仍待突破

網路耙梳主要優勢爲自動蒐集價格資訊，透過軟體將網站內容之非結構性資料，自動擷取並清理後，轉入結構化之資料庫或試算表，再進行後續指數編算作業。然而，查價對象之網站內容或架構，常因商業行銷等目的不定期改版，以及網站自動阻擋耙梳程式機制，皆會衍生耙梳作業成本或障礙；此外，產品銷售週期變短、接續花色自動辨識等情形亦造成資料使用上的困擾。

三、價格代表性資訊不足

耙梳資料普遍缺乏銷售量及實際交易價格資訊，無法分辨商品是否具有市場代表性，雖有國家採用網站評論量替代銷售量作爲項目代表性考量依據，惟尚不普及；另一項發展趨勢，係採實體銷售點之交易掃描器資料（Scanner Data）或電商網站產製的網路交易資訊，所提供之完整銷售量資料，作爲查價項目代表性之判斷依據，荷蘭、奧地利等國已用來試行編算指數，我國亦針對電子發票資訊應用進行可行性分析，相關資料運用未來發展仍待觀察。

四、適法性耙梳環境待建立

網路耙梳技術係進入公開網站蒐集資料，勢必對其網站運作產生影響，所以上述已使用網路耙梳之國家，皆對自

圖 2　import.io 網路耙梳示意畫面

資料來源：作者自行繪製。

論述 》統計 · 調查

動耙梳之適法性進行探討並研訂相關規範，甚而在德國，已有法院判例禁止網路爬蟲在線上資料庫蒐集資料。但政府統計有其社會公益特性，以奧地利爲例，政府要求企業公開網站必須允許其搜尋及下載等行爲，但須遵循透明嚴謹的執行規範，避免對企業網站營運及運作效率等產生影響，包括：不得直接複製整個網站內容、網路爬蟲執行數量及頻率應以最低限度爲基礎，此外，對網站阻絕及延遲手法[3]也必須尊重。網路耙梳之運作平台係在公開之網際網路環境，且政府統計係爲產製公衆治理重要資訊，究竟應朝限制或開放發展，實爲大數據分析應用之重要課題。

肆、結語

在預算及人力成本考量下，運用網路耙梳技術蒐集網站或商品相關資料，以改善或補充現行政府統計之資料來源，國際統計機構已視爲政府統計現代化的可行方案。

然而，現階段網路耙梳運用在「適法性」、「資訊技術」、「資訊人才培育」及「維運成本」等議題仍有許多亟待完善的空間，再從各國物價指數試作情形觀察，雖運用方式或編算作業尚未有標準作法，耙梳工具亦不盡相同，惟對於取代部分傳統查價方式及提升編製效率等深具信心，爲精進我國政府統計業務，未來國際發展趨勢仍宜持續關注。

註釋

1. 超文本傳輸協定（HTTP）是網際網路上應用最爲廣泛的協議，主要用來發布及接收 HTML 此類網頁標註語言所建立網頁。
2. 聯合國歐洲經濟委員會「現代化政府統計（Modernization of Official Statistics）」計畫及歐盟「政府統計理論與研究合作（Collaboration in Research and Methodology for Official Statistics）」於 2013 年起均分別成立大數據專案，其中網路耙梳方法運用爲研究重點之一。
3. 網站管理者可利用帳號輸入、圖形驗證碼及機器人阻絕器等方式延遲或禁止自動蒐集工具在其網站執行，避免影響網站運作效率。

參考文獻

1. https://en.wikipedia.org/wiki/Web_scraping.
2. eMarketer Inc.（2015），Worldwide Retail Ecommerce Sales: eMarketer's Updated Estimates and Forcast Through 2019.
3. Ingolf Boettcher（2015）, Automatic Data Collection on The Internet（Web Scraping）, NTTS 2015, Brussels and Ottawa Group 2015, Tokyo.
4. Ingolf Boettcher（2015）, Big Data in Price Statistics - Scanner Data and Web-Scraping, Statistics Austria, Vienna.
5. Office for National Statistics（2015）, Research Indices Using Web Scrapped Data.
6. Robert Griffioen, Jan de Haan and Leon Willenborg（2014）, Collecting Clothing Data from the Internet, Statistics Netherlands, Hague.
7. Olav ten Bosch and Dick Windmeijer（2014）, On the Use of Internet Robots for Official Statistics, Meeting on the Management of Statistical Information Systems 2014, Dublin, Ireland and Manila, Philippines.❖