《漢學研究通訊》 0253-2875 第 38 卷 第 1 期 頁 43-46

以人工智慧探索人文研究趨勢——漢學研究中心「文史哲學術趨勢分析系統」介紹

Using AI to Analyze Trends in Humanities Research: Introduction to the CCS Academic Trends in Literature, History and Philosophy System

黃文德（Huang Wen-de）、廖箴（Liau Jane）*

一、前　言

隨著網際網路的發達及使用的普遍性，散佈在世界各國的漢學機構不再遙不可及，網際網路的高度易用性，促進了漢學網路資源的傳播利用，不僅能與使用者產生更直接的互動，也拓展了服務版圖。要得知全世界漢學家都做些什麼研究，將漢學機構網頁內容加以分析，探討國際漢學機構網站常出現的關鍵字，即得知哪個領域的研究是較為缺乏的，又哪些領域的研究是最熱門。藉由瞭解國際漢學研究發展趨勢，不僅可以拓展臺灣漢學界的國際視野，亦可展現國家的文化實力，使得臺灣主題的漢學研究成果，能藉此推向國際漢學界，吸引更多的國內外漢學研究學者投入。

國家圖書館漢學研究中心於 2018 年啟動「文史哲學術趨勢分析系統」（http://trends.ncl.edu.tw/）第一期建置計畫，為提供使用者對國際漢學網站進行分析與探索，引導研究者使用數位方法發掘新研究議題的平臺。系統利用網頁資料蒐集程式，自動偵測世界各國漢學機構網頁或社群網站，以人工智慧演算及文字探勘技術，分析下載之網頁最新消息及活動報導，儲存於資料庫中，利用巨量資料分析，自動過濾出欲分析研究活動的相關資訊及，並結合地理資訊系統（GIS）在世界地圖標示各地區所發生學術消息，呈現在趨勢分析系統。

第一階段系統選定 223 個國內外學術機構網頁及社群媒體、討論區等作為分析方向，包含 185 個官方網站以及 38 個社群網頁，地區分布涵蓋臺灣 50 個、中國大陸 62 個、美洲 51 個、歐洲 36 個以及亞非大洋洲 21 個。擷取的機構網頁內容主要為「最新消息」或「活動報導」，社群媒體部分則是以研究機構的 Facebook 或微博為主，討論區部分的形式較複雜，例如：「歷史學柑仔店」是以部落格形式；「故事：寫給所有人的歷史網站」是文章平臺；批踢踢（PTT）各看板則是論壇形式。網站語言之範圍以英文、繁體中文、簡體中文為主。資網站汲取範圍包含可公開接觸的網站，若為社群網站，資料結構與機構網頁所含資訊不同，但其特色是，在不侵犯隱私權的前提下，能夠挖掘較多量化資訊，除此之外，社群媒體所出現之內容，相較正式發表之論文具有更高的娛樂性、即時性，能夠更直接反映社會當下須注意之研究議題。

* 作者黃文德為漢學研究中心學術交流組編輯兼組長；廖箴為漢學研究中心學術交流組編輯兼《漢學研究通訊》主編。

本系統自 107 年 1 月 1 日至 108 年 1 月 17 日共計擷取 17,692 則消息，平均每個月擷取 1,500 則消息，本計畫為四年計畫，預計至 109 年可擷取 500 個網頁，收錄超過 10 萬則消息，更能有效協助文史哲與社會科學研究人文學者創造更多元研究面向。

二、系統設計

（一）趨勢分析首頁

整體設計以互動式分析圖表呈現活潑的首頁設計。採用響應式網頁設計（RWD），可供不同螢幕大小切換內容排版，由上而下的 6 個區塊分別為搜尋總數（預設呈現近 3 個月資料）、關鍵字熱詞分析、熱門議題分析、聲量趨勢分析、空間分布分析及熱門文章排行，圖表為互動式，點選圖表可連結進資料列表頁和原始文章連結。

1. 關鍵字熱詞分析

透過系統語意分析擷取的詞彙，預設呈現近 3 個月熱門出現的 60 個關鍵詞彙，包含熱門及躍升兩種分析模式，透過泡泡大小呈現聲量統計結果。並可依照詞彙屬性（如人物、地點、組織團體、企業品牌、關鍵字等）以不同的泡泡顏色呈現。

如圖 1 所示，自 107 年 10 月 2 日至 108 年 1 月 17 日，近 3 個月的搜尋總數為 5,802 則，包含臺灣 2,252 則、中國大陸 2,232 則、美洲 600 則、歐洲 480 則、亞洲／非洲／大洋洲 238 則。排名前 3 名的關鍵字為「民族」575 則、「北京」501 則、「核心」489 則。點擊關鍵字泡泡後，即可連結至文章列表及文章詳細內頁。

搜尋總數亦可依照時間、來源（官方網站、社群網站）、語系（繁體中文、簡體中文、英文）、作者、提及關鍵字、提及人物、提及地點、提及組織團體、提及企業品牌進行進一步的篩選。

如果點選「臺灣」並選取「社群網站」進行分析，可以看到 1,483 則搜尋結果，排名前 3 名的關鍵字為「中國」137 則、「日本」130 則、「故宮」90 則，點選「故宮」後，可以看到故宮博物院官網及 Facebook 近期發布的消息，進一步分析這 1,483 則的內容，提及人

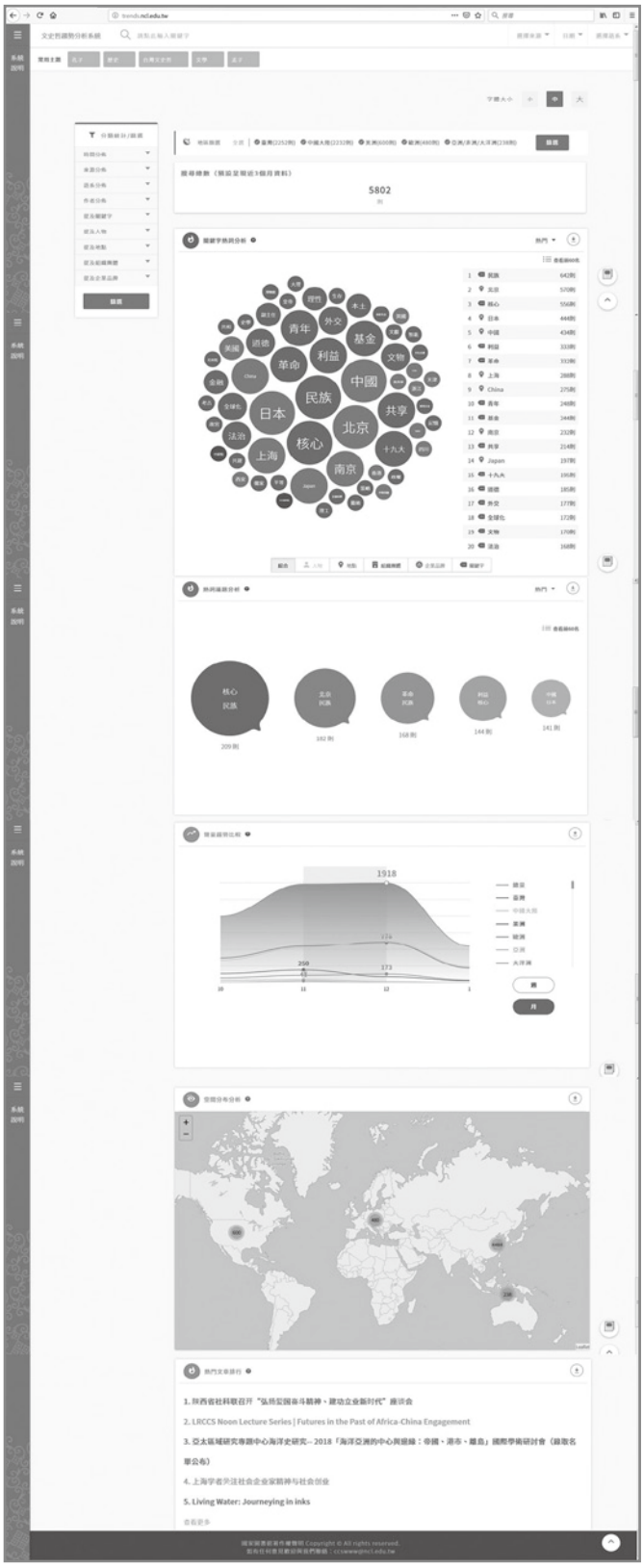

圖 1　文史哲學術趨勢分析系統首頁

物前 3 名為「孔子」、「蔣介石」、「余英時」，進一步點選「余英時」則可以看到與其相關字彙及消息，例臺灣歷史評論 Facebook 所刊登的 2018 余英時先生人文研究獎得主訊息及余英時回憶錄新書導讀講座系列活動等消息，以及 PTT 歷史版對於余英時先生的討論。

2. 熱門議題分析

熱門議題分析包含熱門及躍升兩種分析模式，預設呈現為前 5 名泡泡圖，透過系統語意分析擷取的詞彙，可快速找出相關聯的關鍵字，進行熱門議題自動對應，給予研究者提示關鍵字。所以近 3 個月的熱門議題為「核心／民族」、「北京／民族」、「革命／民族」、「利益／核心」、「中國／日本」。而從近 3 個月「躍升」分析模式來看，則會看到「中國／日本」、「核心／北京」、「日本／北京」等議題是短時間上升聲量最快的。

3. 聲量趨勢分析

聲量趨勢分析是從檢索結果中，找出網頁社群、新聞、論壇中的文章增加或減少趨勢多寡，並以線條圖的方式呈現檢索結果。點選高峰點節點，可進行進階檢索，查看該高峰點的內容。亦可分區查看各區域之文章成長趨勢，分析區域包含臺灣、中國大陸、美洲、歐洲、亞洲／大洋洲／非洲等 5 大區域。由圖 1 可以看出近 3 個月的聲量趨勢最高，也就是發表文章最多的地區是中國大陸，其次才是臺灣，亞洲及大洋洲的網頁更新的資料較少。

4. 空間分布分析

空間分布分析則是結合 GIS 技術，以地圖方式呈現資料分佈情況。搭配地圖可縮放地標功能，主階層集合為五大地區，次階層集合為國家，顯示該國家有多少筆資料。可擷取時間依照地點進行視覺統計，不同的顏色代表單一地區相同數量的座標點，用以表現該主題在某個地區研究密集程度。

5. 熱門文章排行

依照限定時間內，高聲量的熱門內容或學術研究文章，以竄升的類別和主題關鍵字分別進行排序，提供研究者參考並點選，進入詳細的文章內容。

上述圖表呈現內容，皆可點選後呈現相關文章標題、作者、發文時間、來源網站、地區，並可以選擇最新、最多關注兩種排序方式。

（二）查詢系統

除了圖表的呈現外，使用者亦可利用首頁搜尋框，進行關鍵字簡易搜尋。在輸入關鍵字時，系統自動提供搜尋紀錄、熱門搜尋關鍵字及相關推薦等關聯搜尋建議。除了搜尋建議外，使用者可利用「時間篩選」自訂時間區間來限定搜尋範圍；「來源篩選」分為官方網站及社群網站兩種，其中社群網站下又包含 Facebook 粉絲團、PTT、微博及百度貼吧，可針對特定社群網站進行篩選；「語系篩選」分為繁體中文、簡體中文及英文 3 種，文章語系的判斷標準參照來源網站之主要語言。

系統擷取的網頁以中文為主，所以分析圖表中主要呈現的語言是中文，如果以英文關鍵字進行搜尋，則可以看出英文網頁的擷取成果。例如以英文詞彙「Digital Humanities」（數位人文）作為關鍵字搜尋過去一年的資料，可以得到 546 則結果。相關的熱詞分別為「實驗 室 」、「DARIAH」、「Europe」、「Berlin」、「Germany」。「實驗室」為臺灣大學的數位人文實驗室，「DARIAH」是以歐盟為主體的「歐洲研究基礎建設策略論壇」（The European Strategy Forum on Research），結合英國、法國、德國、荷蘭、丹麥等眾多國家之力，合作推動「藝術與人文的數位研究基礎建設」（Digital Research Infrastructure for the Arts and Humanities，簡稱 DARIAH）。

網頁左側的後分類篩選可針對搜尋結果進行後分類篩選，包含側欄的分類統計／篩選列表及地區快篩。使用者可以藉由篩選工具，進一步找到有興趣的關連議題。

1. 分類統計／篩選列表：後分類篩選類別包含時間、來源、語系、作者、提及關鍵字、提及人物、提及地點與提及組織團體等，可跨類別選取條件進行篩選。
2. 地區快篩：地區篩選欄位位於圖表或搜尋結果列表上方，包含臺灣、中國大陸、美洲、歐洲及亞洲／非洲／大洋洲等 5 個地區選項，勾選後點選「送出」按鈕可進行地區篩選。地區篩選預設為選取全部地區。

（三）文章列表及詳細內頁

呈現單筆查詢文章簡要顯示資訊包含項次、標題、來源網站、作者、文章發布時間、文章來源地區、簡要顯示內文。原文 URL：點選可連結至該篇文章的原始頁面。另外也可以「分享至」將本篇文章的「原始網址」或「本頁網址」複製到外部網站。

而內文呈現部分，設定呈現原文的 50%，如欲看全文需連結至原文 URL，內文中之關鍵字以底色標示。而留言呈現，則是擷取該篇文章的留言，呈現於內文下方，留言排序依據原文留言排序呈現，留言標首顯示留言筆數。留言擷取項目包含回應者帳號、留言、留言時間等。標籤關鍵字擷取呈現本篇文章，由系統自動擷取之關鍵字及各類別標籤，包含提及關鍵字、提及人物、提及地點、提及組織團體等類別。

以下圖 2 的消息「Poets, Artists, Game Makers, and New Media」為例，可以看到發布時間是 107 年 11 月 17 日，由康乃爾大學亞洲研究系發布，除了可以連結到原網址的 URL 外，也系統自動擷取該則消息的關鍵字，例如提及人物（Brett M. de Bary、Andrew P. Campana），提及地點（Japan）、提及組織團體（Massachusetts Institute for Technology）等。

圖 2　點選文章列表後呈現之詳細內頁

系統之資料分析優化與正確性，必須透過長期性的追蹤、修正，透過使用者與學者共同參與建構與除錯，系統人工智慧文字辨識學習，以持續性、自動化方式，逐漸強化、豐富現有系統內容、檢索機制，並提高利用效能。希望能透過人文學者的參與，發掘研究議題，並將成果以普及化的方式呈現。

三、後續規劃

文史哲學術趨勢分析系統的目標是企圖匯整國際漢學研究歷程、累積研究經驗，激發人文學者發掘在人文研究上的新面向。本計畫的下一階段目標是逐步增加分析的網頁數量，藉由對於國內外漢學機構最新消息的全文分析，快速的提供「漢學研究通訊」電子報所需要的消息，擺脫現在學研究中心在尋找漢學活動消息上耗費大量時間的問題，並借用新工具之統計與分析，充實電子報內容。並預計將漢學研究中心的出版品全文加入分析內容，最終目標在於可讓研究者特過關鍵詞，掌握學術概念在網路曝光、研究發表以及出版之間的流動性。

文史哲學術趨勢分析系統

國家圖書館漢學研究中心之「文史哲學術趨勢系統」利用文字探勘與 AI 巨量分析技術，分階段觀察全球重要漢學機構網頁及社群，每天更新最新網頁資訊，藉以從龐大的資料中找出使用者所關注、需要的資訊，進而創造智慧化的資料分析。

歡迎國內外學界多加利用。

如有任何寶貴意見，歡迎來信至 ccswww@ncl.edu.tw