社群媒體巨量資料蒐集與分析——以 Facebook 與 Twitter 為例[#]

鄭宇君、陳恭、陳百齡*

一、前言

社群媒體（social media）是當代社會人們資訊分享、社交互動的核心，相較於 Web1.0 的網路媒體，Web 2.0 的社群媒體最大特色在於平臺本身不產製內容，主要由用戶自行生產各式內容，平臺供應商只是因應人們不同的社交與資訊需求，設計組建了不同形式的社群媒體平臺，鼓勵人們在平臺上創作與分享各種內容，因而個別平臺內在結構的差異影響了社群互動及資訊擴散方式，同時造就了資料結構的差別。

從巨量資料分析（big data）的角度來看，當我們要理解不同平臺的社群互動與資訊流動方式的差異，也必須理解不同平臺的內在結構及其資料特徵，才能掌握該平臺的能供性（affordance）何以能提供社群不同的互動方式。舉例而言，強調資訊快速流通的 Twitter，以 140 字元長度限縮了人們用極短的文字來表達訊息重點，使得 Twitter 在重大突發事件或災難時成了訊息即時流通的備援頻道；強調實名制的 Facebook 則以社交互動為號召，希望用戶將線上與線下的人際網絡重建於 Facebook 上，以強化用戶的黏著度，因此用戶的人際網絡成為 Facebook 最有價值的資產，許多新聞或品牌必須在 Facebook 創建帳號來發送訊息，透過其訂閱用戶的人際網絡或口碑幫助品牌訊息擴散。

本文主要介紹傳播領域與資料科學跨領域合作所成立的水火計畫研究團隊，[1] 本研究團隊共有四位主要成員，包含國立政治大學新聞學系陳百齡、玄奘

本文為科技部專題研究計畫「跨平臺社群媒體巨量資料蒐集與分析」（MOST 104-2420-H-004-043）之部分研究成果。

* 鄭宇君，玄奘大學大眾傳播學系副教授；陳恭，國立政治大學資訊科學系特聘教授兼電算中心主任；陳百齡，國立政治大學新聞學系教授兼傳播學院副院長。

[1] 水火計畫研究團隊的相關資料與研究成果，可參閱網站 https://sites.google.com/a/newliteracies.co.cc/floodfire/。

大學大眾傳播學系鄭宇君、國立政治大學資訊科學系陳恭與李蔡彥，結合傳播與資訊科學專家共同發展新的方法取徑，主要目的在建立重大事件中社交媒體資料蒐集與分析之標準作業程序，以減少巨量資料在資料處理與轉換過程中的流失與誤差，同時在人力及物力資源有限的情況下，能夠有效運用計算資源與節省研究成本。

本研究團隊以 Twitter 及 Facebook 作為資料蒐集平臺進行研究設計，發展了可蒐集這兩個平臺的工具。由於 Twitter API 可提供所蒐集關鍵字或帳號七日內貼文資料量的百分之一，而 Facebook 不開放關鍵字搜尋功能的 API，透過 Facebook Graph API 只能以粉絲頁為單位，蒐集指定粉絲頁的內容。[2] 我們面臨的研究方法議題主要有二：（1）如何在第一時間決定資料撈取方法，包括關鍵字、關鍵帳號之選擇；（2）設計有效率、具彈性的社交媒體資料蒐集、儲存與分析平臺。

因此，本文目的是以 Twitter、Facebook 為例，說明二個社群媒體資料欄位及社群互動方式的差異，根據我們先前的研究經驗，結合 Twitter 與 Facebook 官方的 API，開發出方便研究者蒐集與分析社群資料的工具，並將這二款資料蒐集工具以開放原始碼或網頁版方式釋出，開放學界相關先進使用，以達到知識與資源共享之目的。

二、不同社群媒體平臺的資料特徵

根據不同社群媒體平臺的界面設計，可以產出不同的互動與資料形式，以下以 Twitter 與 Facebook 為例，分別從用戶直觀感受的使用者界面結構進行介紹，進一步說明這些界面呈現的人際互動與資料流通，如何以數位化方式被記錄成為資料或後設資料。接下來，根據研究者經常分析的類目，我們歸納出社群媒體資料蒐集的主要面向，包括用戶檔案資料與貼文互動資料，在瞭解這些資料結構之後，以便規劃後續資料檢索、分析與詮釋之方向。

（一）Twitter 資料結構特徵

Twitter 的資料結構可分為二大部分，第一是用戶個人檔案，第二是貼文互

[2] Facebook 在使用 Graph API 時，使用規則中明確規定若要使用 Graph API 必須要先取得一組授權碼（Access Token），而此授權碼內包含著可以存取的「用戶」及「資料權限」，而取得該授權碼的方式即為透過 Facebook Login 功能，讓用戶登入並同意授權，才能取得授權資料範圍之授權碼，而且個人的資料 Graph API 無法獲得，故我們只抓取粉絲頁的資料。

動資料。在歷經多次變化之後，目前 Twitter 個人檔案的呈現方式也越加豐富，包含個人圖片、自我介紹、追蹤人數（following）、追蹤者人數（followers）、發文內容等，參見圖一。

在個人檔案中用來分析的最重要欄位，包括用戶帳號（user id，以 @user 表示，不可更改）、用戶名稱（screen name，可編輯）、用戶描述，研究者可從這些資料來判斷該用戶的身分；其餘欄位：用戶的追蹤人數（following）、追蹤者人數（followers）則可用來瞭解用戶與社群的互動關係，甚至可進一步蒐集該用戶的 following/follower 清單，以建立用戶之間的社會網絡關係，但這二個數目會隨時變化，每一刻鐘都可能新增或減少人數；另外也可透過 Twitter API 撈取特定用戶的所有貼文資料。

圖一：Twitter 個人檔案頁面

其次，在 Twitter 上，一則貼文（tweet）的結構可參見圖二，主要為貼文本文，當中內含主題標籤（hashtag）、超連結（hyperlinks）等延伸資訊。用戶與其他用戶的互動也都具體包含在一則推文之中，除了原創貼文之外，若該則貼文是轉推（分享）自其他用戶的貼文，則會顯示 RT 在貼文句首，若用戶在該則貼文內與其他用戶對話，則可直接用 mention 或 reply 表示，則在內文中會出現 @User 的標示。

圖二：一則推文Tweet的結構

（二）Facebook 資料結構特徵

在 Facebook 方面，目前多數研究者及社群數據公司所蒐集與分析的都是 Facebook Page（粉絲頁）的公開資料，以下以粉絲頁的資料結構特徵為例說明。由於 Facebook 更強調社交互動，在用戶檔案界面上的設計更加豐富多元，可以參見圖三。

圖三：Facebook page用戶檔案界面

圖三呈現了 Facebook 粉絲頁檔案中用來分析的最重要欄位，包括用戶帳號（user id，不可更改）、用戶名稱（screen name，可編輯）、粉絲頁類型、用戶描述，研究者可從 API 取得這些資料來瞭解用戶身分，並記錄粉絲頁的按讚人數。透過這些資料以判斷粉絲頁的所在地，如：臺灣、香港、美國等，或是從 Facebook API Insight 提供的資訊判斷，它是依照按讚粉絲的所在地來判斷該粉絲頁的所在地，例如：一個粉絲頁的按讚人數有 80% 來自臺灣用戶，那 Insight 會判斷為臺灣粉絲頁。

另外，Facebook 貼文互動結構比 Twitter 複雜，Twitter 將用戶互動以水平方式呈現在貼文內容裡，但 Facebook 則是垂直的階層化互動，每一則貼文具有三個數字指標：反應數（包含按讚、愛、哈、嗚、怒等五個表情符號的數目）、分享數、留言數，這三個指標可反應這則貼文引起的用戶投入程度（user engagement），而具體的貼文互動可包含三層內容:貼文（post）→評論（comment）→回應（reply），每一層內容都包含了發言者名稱與內容（文字、圖片），留言也有反應數（按讚數）為依據。Facebook 的貼文互動結構可參見圖四。

圖四：Facebook貼文互動結構

（三）比較不同社群媒體的資料特徵

社群媒體作為一個眾多使用者及內容互動之平臺，研究者蒐集到的大量社群資料中包含了眾多的資料與後設資料，為了進行資料檢索與查詢，我們根據

學術使用者所需目的，找出研究者經常使用的查詢條件，搭配社群媒體的資料特徵，建立所需要的查詢 schema。

根據本計畫團隊目前所蒐集的資料，可根據以下四個類型的資料特徵來建立查詢條件：

1. 資料平臺類型：Twitter、Facebook 或其他社群平臺。
2. 用戶類型：一般個人帳號、名人、媒體、機構或組織、機器人帳號。
3. 貼文類型：標題、貼文內容（文字、圖像、影片）、超連結。
4. 互動類型：按讚（likes）[3]、分享（share or retweet）、分享加註解、評論（comment）、回應（reply）、標籤（tag）、#hashtags。
5. 背景資料：時間資料、地理位置資料、使用載具等後設資料。

由於 Twitter、Facebook 的資料欄位差異很大，為了建立跨平臺的資料查詢機制，我們根據不同平臺的資料結構，找出共同的基本欄位，建立對應的名稱作為主要的查詢條件。因此，歸納了跨平臺資料的查詢條件及各平臺相對應的名稱，如表一：

表一：不同社群平臺的資料對應

	發言者的層級		內容的層級						
平臺	Source	user	標題	原創貼文	轉貼內容 shared post	評論	回應	Image	link
Facebook	V（page）	V	V	V（likes）	V（likes）	V（likes）	V（likes）	V	V
Twitter		V		V	V（retweets 數目）			V	V

註：V 表示該平臺有這項欄位。

從表一來看，Facebook 的資料形式最為複雜，在發言者層級即可分為粉絲頁或個別用戶發言，內容部分則為多層級的互動結構，可分為原創貼文、轉貼內容、第一層評論（comment）、第二層的回應（reply），每一種內容都有按讚數作為熱門度指標，內容同時也包含圖像或超連結；相較之下，Twitter 的資料形式較為簡單，無論是機構或個人使用者都是同樣層級的用戶帳號，貼文互動只反應在單一層內容，僅以 RT 或 @user 來區隔互動的方式，它也會帶有圖像與超連結。

[3] Facebook 後來將按讚擴充為多個表情符號，但在「按讚數」通常遠大於其他表情符號，故分析上仍以「按讚數」為主。

三、社群媒體平臺的資料蒐集與分析工具

在社交媒體資料蒐集方面，我們目前建立 Twitter 和 Facebook 粉絲頁的資料蒐集工具。

（一）Twitter 的資料蒐集工具

本研究團隊所使用的 Twitter 資料蒐集分析平臺，主要根據荷蘭阿姆斯特丹大學 Rieder 教授開發的開源軟體 DMI-TCAT 為資料蒐集工具（Borra & Rieder, 2014）進行改寫。[4] 然而，此項工具主要針對英文推文所開發，它的 Stream API 無法支援中文推文資料的蒐集，因此我們修正並增添部分功能，成為 FloodFire-TCAT-v2（水火計畫 Twitter 資料蒐集與分析工具），主要加入語言辨識功能與使用者介面，除了接受 Twitter API 原本提供的語言標識外，我們開發的工具並可區分繁體中文與簡體中文。[5]

實際進行資料蒐集時，首先，透過 Twitter Search API 蒐集帶有特定關鍵字的貼文並存於 NoSQL 資料庫，再轉到 SQL 資料庫進行資料儲存與清理，在去除重複或不完整資料後，成為乾淨的資料集，此時再計算資料集裡不重複的貼文與發文者人數。

其次，在進行資料處理的同時，我們會根據研究所需進行貼文時間的轉換（目前主要由格林威治標準時間轉換為臺灣時間）、加上語言辨識的結果（Twitter 本身提供了主要語言的標識，但繁體中文與簡體中文被 Twitter 同樣標識為 ZH，透過本研究團隊自行開發工具可將其區分為繁體中文 ZH-TW、簡體中文 ZH、廣東話等華語語系的不同書寫文字）。

最後，在獲得乾淨的事件資料集之後，研究者可根據不同研究目的進行各種類型的分析。舉例而言，社會網絡分析（social network analysis）可以得知眾多網路用戶之間的互動或群聚關係，或是找出不同語言社群中重要的連結者，如：2012 臺灣總統大選期間 Twitter 上不同語言的參與者與互動（鄭宇君、陳百齡，2014）；[6] 詞語共現網絡分析（co-concurrence analysis）可分析大量貼文中哪些詞彙之間彼此同時出現；引用來源分析（sources analysis）又稱為超連結分析（hyperlinks analysis），則用來分析大量貼文中所帶有的超連結，先進行推文內的

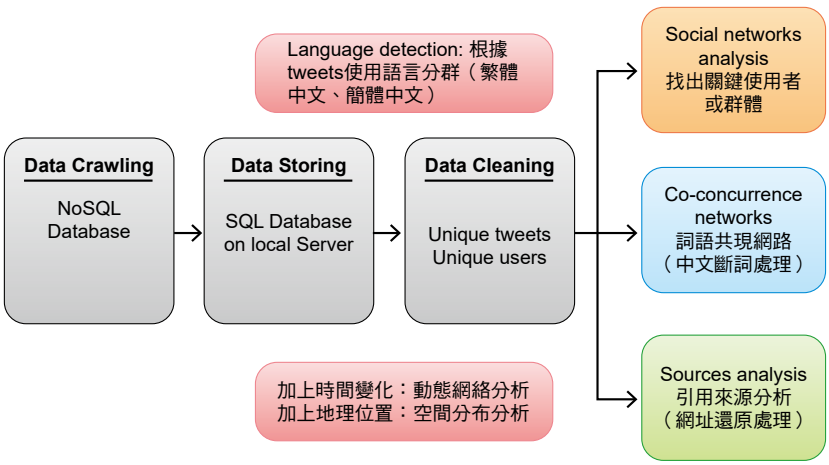
---

[4] Borra, E. & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 66(3), 262-278.

[5] 本團隊開發的 Twitter 資料蒐集工具之原始程式碼可參見 https://github.com/ninthday/ff-tcat-v2

[6] 鄭宇君、陳百齡（2014）。〈探索 2012 臺灣總統大選之社交媒體浮現社群：鉅量資料分析取徑〉，《新聞學研究》，第 120 期，頁 121-165。

短網址還原，統計哪些網域是最常被引用的超連結，如：2012 臺灣總統大選期間 Twitter 上的新聞來源引用（鄭宇君、施旭峰，2016）。[7] 藉由這些貼文資料及後設資料，研究者得以進行社交媒體巨量資料分析，從巨觀的角度來掌握事件發生的動態過程並發現洞見。社交媒體巨量資料處理流程如圖五所示：



圖五：社交媒體巨量資料處理流程

（二）Facebook Page 資料蒐集工具

本計畫團隊根據 Facebook Graph API V2.10 的規範，開發了一款 Facebook Page 的資料蒐集工具 NCCUFBTOOL，透過友善的使用者界面，方便不會寫程式的人文社會科學研究者可以透過這款工具下載 Facebook Page 公開資料。這款工具同時開放提供給學術研究與教學使用，使用者必須遵守相關之學術倫理，不得逾越特定目的之必要範圍。[8]

NCCUFBTOOL 的使用是針對粉絲頁進行貼文及互動資料撈取，登入頁面如圖六所示。使用者以自身的臉書帳號登入後，輸入所要查詢的粉絲頁 page_id，並設定所要查詢的貼文資料之開始與結束的時間區間，所撈取的資料類型，包括粉絲團公開貼文（Post）／第一層留言（Comment）／第二層留言（Level 2 Comment），以及貼文、留言之按讚數、分享數、回應數及表情符號數。使用者輸入完查詢條件後並留下 email，系統會排程執行資料蒐集，在資料蒐集完成後，會將壓縮檔案寄到使用者所指定的信箱。

[7] 鄭宇君、施旭峰（2016）。〈探索 2012 臺灣總統大選社交媒體之新聞來源引用〉，《中華傳播學刊》，第 29 期，頁 107-133。

[8] NCCUFBTOOL 的登入網址：http://140.119.163.69/PageDataCollector2/fbLogin.aspx。

圖六：NCCUFBTOOL的使用畫面

這款工具目前可蒐集指定粉絲頁在指定時間區間內的貼文、評論、回應（第二層留言），資料以 csv 檔匯出，使用者可使用 EXCEL 或相關軟體開啟，所下載的 csv 檔包含的資料欄位說明如表二所示。

表二：NCCUFBTOOL所蒐集的資料欄位

資料欄位名稱	解釋
流水號	此次抓取資料的編號，第一篇貼文為 1，第一篇貼文第一個留言為 1-1，第一篇貼文的第一個留言之第二層留言為 1-1-1，以此類推
大類別	此筆資料的類型，共有 Post, Comment, Level 2 Comment 三種
From	此筆資料發文者名稱及 id，除了粉絲頁有名稱外，個人帳號只有 id，且此 id 被臉書模糊化過，並非實際使用者 id
post_url	該篇貼文的連結
created_time	此筆資料的創建時間（已轉為臺灣時間）
post_type	該篇貼文的種類，如：影片、連結等
postID	此筆資料的 id
post_message	此筆資料的內文
status_type	此筆資料的動作，如：新增照片、分享、tag 人
link	該篇貼文中的多媒體，如：照片、影片的連結
轉貼標題（name）	該篇貼文中的多媒體之標題
留言數	此筆資料底下的留言總數
總回應數（Reactions）	此筆資料的回應總數，為六種表情符號資料加總
like	按讚數
love	愛心數
haha	哈哈數
wow	哇數
sad	難過數
angry	生氣數
分享數	該篇文章被分享數
update_time	該篇文章最後互動時間，互動包含點讚、留言、分享、編輯等

透過上述工具所蒐集的資料，不具有資訊科學背景、不會寫程式的人文社會科學研究者也能取得臉書粉絲頁之原始資料自行分析，例如：貼文及回應的時序分析、互動分析，或針對貼文與回應內容進行斷詞分析、情緒分析或人工的內容分析，同時也可藉由所蒐集資料欄位的超連結連回 Facebook 網頁，觀看該則貼文所附上照片或影音內容，進行人工編碼分析。

由於本工具目的在於提供學術研究及教學使用，設計一個使用者界面方便研究者透過 Facebook Graph API 下載資料，並不會保存使用者下載的資料。然而，本款工具所能蒐集的資料範圍會隨 Facebook 調整 Graph API 權限而改變，目前實測結果發現，若欲蒐集較長時間區間的貼文資料，建議可分數次、縮短時間區間來蒐集，以取得較完整資料，若是較大型的粉絲頁（如：貼文及回應數都相當多的新聞粉絲頁）則僅能蒐集到較近期的資料。

四、結語：社群媒體巨量資料蒐集之挑戰

透過這些社群媒體巨量資料蒐集與分析工具的應用，我們展現了傳播研究與大數據方法結合之優勢，可幫助研究者掌握巨觀的社交媒體訊息變化趨勢，特別是時間趨勢上的劇烈變化，可以用每天、每小時為單位計算變化趨勢；透過語言辨識軟體區分大量貼文裡的不同語言比例，得以比較跨語言社群的傳播模式；透過超連結分析，可以幫助我們挖掘新聞來源的引用情況。這些研究方法上的創新可讓研究者瞭解重大事件發生時在訊息流動及網路社群的動態變化。

然而，社群媒體巨量資料蒐集所面臨的最大挑戰來自 Facebook 與 Twitter 公司所釋出的 API 權限，由於越來越多的社交與商業活動在這二個平臺上進行，每天產生的資料量也越來越大，在可預見的未來，這二家公司會慢慢限縮 API 可取得的資料範圍，特別是歷史資料的取得可能更加困難。對於研究者而言，若想透過 API 取得免費的社群數據，必須在事件發生當下或儘早開始進行資料蒐集，不然資料可能隨時消失（包括粉絲頁或用戶自行刪除貼文）或是 API 無法取得一定時間之前的資料。因而，社群資料蒐集的這個特性將使得研究者越來越需即時具有問題意識，以便儘早決定社群資料的撈取。