社群媒體資料分析：特性和歷程的初探*

陳百齡、鄭宇君、陳恭**

摘要

社群媒體資料分析是藉由社群媒體所產生的鉅量資料進行觀測和分析，藉以瞭解社會科學所關注的問題。這個研究取徑，需要和資料科學以及其它學門進行跨領域團隊合作，相關論述仍然在起步階段，有待相關學門研究者投入更多關注。本文為初探性論文，試從傳播學者的角度，描述社群媒體資料分析特性、元素和歷程，透過個案分析，呈現研究人員處理社群媒體資料的歷程，以及如何針對提問，處理社群資料。

關鍵詞：社群媒體、資料分析、研究方法、鉅量資料、跨領域協力研究

*本文為科技部委託專題研究計畫〈社交媒體重大事件之傳播模式比較分析：鉅量資料數位人文取徑〉（編號 NSC 102-2420-H-004-048-MY3）之部分成果。寫作過程承蒙國立政治大學傳播學院水火計劃團隊成員參與討論，研究助理林聖翔、柯皓翔共同協助整理資料，特此誌謝。

**陳百齡為國立政治大學傳播學院新聞學系教授。聯絡方式為：pailinch@nccu.edu.tw
鄭宇君為玄奘大學大眾傳播學系副教授。聯絡方式為：colisa@gmail.com
陳恭為國立政治大學資訊科學系教授。聯絡方式為：chen.kung@gmail.com

前言

近年來，許多傳播學者都提及社群媒體對於傳播生態造成衝擊，不僅學者希望對於社群媒體內涵有更多理解，政府和企業組織，也希望透過社群媒體內容分析，瞭解社會輿情的發展趨勢，以及消費者的動向。然而社群媒體的資料分析具有鉅量資料特徵，這類資料和傳播學者慣用的傳統社會科學資料分析方法不盡相同。因此也引起學者探索的興趣。

從近年來傳播學門所發表的成果看來，雖然許多學者咸認為使用鉅量資料取徑是必然的趨勢，但目前關於社群媒體資料蒐集和分析方法的論文仍然在起步階段，這個知識領域有待學者投入更多關注。本文是一篇初探性質的論文，試著從傳播學者的角度，描述社群媒體資料分析特性、元素和歷程，透過一項個案，呈現研究人員如何進行社群媒體資料分析，以及歷程當中，研究人員如何提問和處理資料。

壹、社群媒體資料分析

廿一世紀之初崛起的社群媒體（social media），旨在「提供使用者在有界限的範圍內建立公開或是半公開的網絡服務，使用者可用以連繫彼此、或分享彼此的訊息」（boyd & Ellison, 2007: 211），包括臉書、推特、批踢踢實業工作坊等都可歸類於社群媒體[1]。於是，我們從社群平臺上觀看朋友上傳的相片和影音，也留言回應貼文，更常常轉貼和分享有用或有趣的訊息。因此，社群媒體的資訊產製機制，迥異於先前的大眾傳播媒體，它是一種「社群之間藉由對話、分散進行的內容產製、擴散和交流」的媒體平臺機制（Zeng, Chen, Lusch & Li, 2010: 13）。社群媒體的出現，不僅大幅度顛覆了人們分享資訊的方式，同時也給傳播研究帶來重大的衝擊。

[1] 誠如 van Dijck（2013）指出，將社交媒體可分為四種類型：（1）社交網絡平臺（social network sites, 簡稱 SNSs）：以社交網絡為基礎的社群網站，這也是最大宗，包括 Facebook、Twitter、Linkedin 皆屬於此類；（2）使用者產製內容平臺（user-generated content，簡稱 UGCs）：這類為使用者創作內容的分享平臺，如：YouTube, Flickr, Wiki 等；（3）交易與行銷平臺（trading and marketing sites, 簡稱 TMSs）：如：eBay, Amazon 等，以交易或行銷為主要目的之社交平臺；（4）遊戲平臺（play and game sites ,簡稱 PGS）也就是以線上遊戲為基礎的社交互動平臺。一般學者分析社群平臺研究時，主要以第一種 SNS 平臺為主。

當成千上萬閱聽人每天使用社群媒體產製、協力並擴散各種訊息，這些訊息經過在媒體平臺自動地歸類、記錄和保存，形成一個前所未有的龐大資料集。社群媒體所產出的資料，不僅包括由人所產出的內容文本（content），也包括由機器產出的後設資料（metadata）。這群資料數量龐大而多樣，Manovich（2012:2）稱之為「社群鉅量資料」（Social big data）。社群媒體承載的資料，內容涉及人們產出的大量事實、意見、想像和感受，提供了一個龐大的資料庫，可以作為蒐集和分析的對象。因此讓學者、企業主、政客和媒體工作者都躍躍欲試，企圖從這個龐大的資料集當中發掘社會、政治、文化和產業的契機。學者 Tufekei （2014: 1） 有個生動的譬喻：「社群媒體鉅量資料問世，之於人類行為研究，就有如顯微鏡之於生物學，或望遠鏡之於天文學」。上述比喻點出鉅量資料崛起為社會科學研究帶來質變：改變的不僅僅是資料分析規模而已，更及於格局和深度。

社群媒體資料分析（social media analytics）是因應社群媒體崛起而出現的一個知識領域。對於傳播學者而言，鉅量資料問世帶來的可能是一個研究方法典範的重大改變。鉅量資料最重要的意義，並不僅止於研究資料數量規模的驟增而已，但更大的挑戰，則是傳播學者「如何蒐集、整理，並交互參照這些大型資料集」（boyd & Crawford, 2012: 663）。

社群媒體資料分析是一個新興領域，學者發展出不同的定義。例如，Zeng, Chen, Lusch & Li（2010: 14）認為社群媒體資料分析是「根據特定需求，發展和評估各種資訊工具和知識結構，用以蒐集、觀測、分析、摘要，或呈現社群媒體資料」。Stieglitz, Dang-Xuan, Bruns & Neuberger（2014: 90）則認為，社群媒體資料分析是一個結合多重知識學門的研究取徑，相關學者分別從商管、經濟、社會等學門領域出發，提供學者方法論的基礎，針對大規模的社群媒體蒐集、萃取、分析資料，甚或建置資料模型，藉以解決學術或實務界所提出的問題。

社群媒體和資料分析這兩件事，可以視為當代社會資訊處理的兩種模式。諾貝爾經濟獎得主 Kahneman 指出，個人認知存在兩種思考決策模式：一方面在事件當下立即產生反應（System I），另一方面則透過推敲形成決

策（System II）；前者是「快思」，後者則是「慢想」（Kahneman, 2012）。當代社會中的社群媒體，在短時間內所產生大規模、多樣化，並且高密度的資料，類似於當代社會立即反應的「反射系統」。相對而言，社群媒體的資料分析，由研究人員進行資料蒐集、過濾、分析，經由精密卻緩慢的程序，企圖再現社會真實，則類似於「反思系統」。二者之間有如推力和拉力，藉由「快思」和「慢想」，以維持社會資訊的平衡。[2]

貳、社群鉅量資料分析的挑戰

綜合各方文獻，社群鉅量資料分析面臨多重挑戰，包括：資料龐大而複雜，以人的行為或表達為基礎所構成的資料集，不易使用自動化技術處理，資料完整性和透明度遭質疑，因此目前尚在起步階段。

一、數量規模龐大

社群資料數量規模大、資料型態多樣、並且持續不斷產出資料。社群媒體是 Web 2.0 時代的網路平臺，允許使用者產製和分享資料。這些聚集在社群平臺上的資料，不僅包括內容資料（content），也包括描述資料特性的後設資料（metadata），數量規模遠超過傳統媒體平臺。研究人員使用既有的應用程式、遵循傳統研究方法，往往難以勝任資料處理任務（Stieglitz, Dang-Xuan, Bruns & Neuberger, 2014: 91）。

二、以人為核心

社群資料是一種「以人為核心」（human-generated computing）的運算機制所產生的資料。這類資料有很大一部份來自人類的語言行動與人際／機互動。迥異於其它使用儀器觀測而產生的鉅量資料（例如，溫度雨量或空污粒子觀測資料）。

大部分的社群資料包括兩類資料，即後設資料（metadata）和內文資料（content）。後設資料是「描述資料的資料」，例如使用者帳號、發文時間和文章編號等；通常由機器建立，格式相當規律。內文資料的分析則較

[2] 這個譬喻，係得自國立政治大學陳樹衡教授的啟發。

為困難；首先，人類語言使用非常多元，文本內容中含有大量標題、關鍵詞、標記（tags）、或表情符號，以及影音串流檔案內容等，這些資料不僅格式龐雜、內容屬性高度歧異，文本意涵也非常複雜，因此需要大量人力／物力資源進行資料清洗。其次，人類語言本來就具有各種型式：意見、評價、或反諷，一詞可能多義。研究人員目前雖可透過資料科學技術如文本探勘(text-mining)或機器學習(machine learning)找尋文本態樣(patterns)，但仍難以精確捕捉文本真意（Zeng, Chen, Lusch & Li, 2010: 14）。第三，雖然許多人認為，社群媒體資料反映人們的線上互動態樣，但線上互動和社會情境密切相關，而人類互動意義又極其複雜，從社群媒體平臺有限內文資料（例如，按讚、分享、留言數量計算，或社會網絡連結的中心度）要解釋人類複雜的互動意義，或許過於簡化（Tufekei, 2014）。因此，自動化技術能夠幫助人們瞭解社群互動程度仍然有限。

三、資料完整性和透明度

社群媒體分析受到的另一個挑戰，則在於資料的完整性。社群資料是媒體平臺的重要營收來源之一，社群媒體透過販售社群資料盈利，因此並不會對公眾釋出所有資料，透過應用程式介面（API）所釋出的資料，僅為極小比例資料。以 Twitter 為例，廠商釋出給 API 的資料量，僅佔總量 1%；若要完整資料則必須耗費鉅資向其購買。正如學者 Manovich （2010）指出，儘管社群媒體擁有和保存了鉅量資料，但目前能夠即時近用完整社群媒體資料者，限於付得起鉅額資料使用費的企業和政府組織；至於學術機構和一般企業的資料分析人員，則必需設法從社群平臺釋出的資料中儘可能擷取素材，或者委託社群媒體資料蒐集廠商提供素材。當資料委託第三方蒐集時，資料完整程度又往往受限於代理廠商的技術能力。例如，社群媒體由許多平臺組構，來源非常多元，資料蒐集是否能夠及於所有資料來源，往往也值得探究。

除了資料完整性，另一值得注意的面向是資料透明度。科學研究品質判準之一，在於資料是否可複製（reproducibility），以保證相同樣程序下可以獲致相同結果。因此資料蒐集或分析程序必須具有透明度

（transparency）。在社群媒體資料蒐集場域，透明度主要表現於演算法（algorithm）機制公開與否。當代社群媒體平臺業者係以演算法機制透過程式擷取並釋出資料，資料供應廠商（外部的資料蒐集者）無論是自行爬取平臺資料或向平臺業者購買資料，皆未能獲知平臺業者之演算法內涵。社群媒體平臺往往以市場競爭為由，將演算法視為商業機密之一部分而不予公開，如此一來，研究人員蒐集資料時，便缺乏足夠透明程度，無法從演算法判斷蒐集程序是否合理，以及資料是否完整或有缺漏之虞。

基於上述種種原因，社群媒體研究蒐集資料的門檻較高，因此在這個階段以實證資料為基礎的研究，數量較為稀少。根據 Felt（2016: 4-5）所做的一項後設分析（或整合分析，meta analysis），在傳播論文資料庫 Communication and Mass Media Complete 蒐尋到的 294 篇社群媒體論文當中，83 ％ 論文依舊採用傳統方法蒐集資料，使用資訊科學技術蒐集僅佔 17 ％。而在使用傳統方法蒐集資料的論文當中，則以內容分析和問卷調查（特別是線上問卷）為大宗，各佔 21% 和 20%。另外，同一研究也發現，上述研究採集資料的社群媒體，多集中在臉書（69%）和推特（46%），跨平臺的研究相當稀少（<10%）。此外，根據江奕瑄、林翠絹（2015）的文獻分析，她們從 SSCI 資料庫檢索出 39 篇論文，但採用數據資料進行實證研究的只有 12 篇，其餘皆為理論或概念分析。國內 TSSCI 則僅有 2 篇論文，其中僅有一篇使用資料分析。上述資料也分析作者數目，鉅量資料研究大多數由多位作者合著，顯示這個領域猶在起步階段，且因跨領域而需要團隊協力研究。

四、小結

誠如上述，社群媒體產出規模龐大、多重媒材形式、以及充滿各式雜訊的訊息，傳統社會科學研究取徑相當難以處理。因此傳播學者提出的問題，必須透過資訊科學的協助，方能解決社群媒體相關的問題。因此使得社群媒體資料分析具有雙重歷程的特性。

參、社群資料分析的雙重歷程

社群資料分析具有雙重歷程（dual processes）的特徵。一方面，傳播研究者必須先發展出擬探問的問題，然後根據資料處理過程展現的形貌，探索並獲取解答。另一方面，資料處理有一定的步驟，研究者根據提問，蒐集資料，清洗資料，並將資料以視覺化方式呈現。以下先分別敘述這兩個歷程內涵，然後再說明二者之間的關係。

一、 提問／解題（problem formulation/solving）

資料分析本質上是「從問題發想到解決問題」的歷程。學術研究和實務工作者使用社群資料，以分析為手段，最終解答心中的疑惑；前者如傳播學者關注重大災情訊息如何透過社群媒體而傳佈或擴散？如何藉由社群平臺而形成社會聚合？（例如陳百齡／鄭宇君，2014）後者則例如企業或組織公關透過社群資料瞭解特定事件的輿情走向，以做為品牌形象維護或危機處理策略的參考（李瑞娟，2016）。雖然學者和實務工作者分析資料的問題意識、解題策略、精密度，以及分析結果的預期等等，或許不盡相同，但二者都包含關於問題和資料的想像，以及運用社群資料獲取答案的過程。

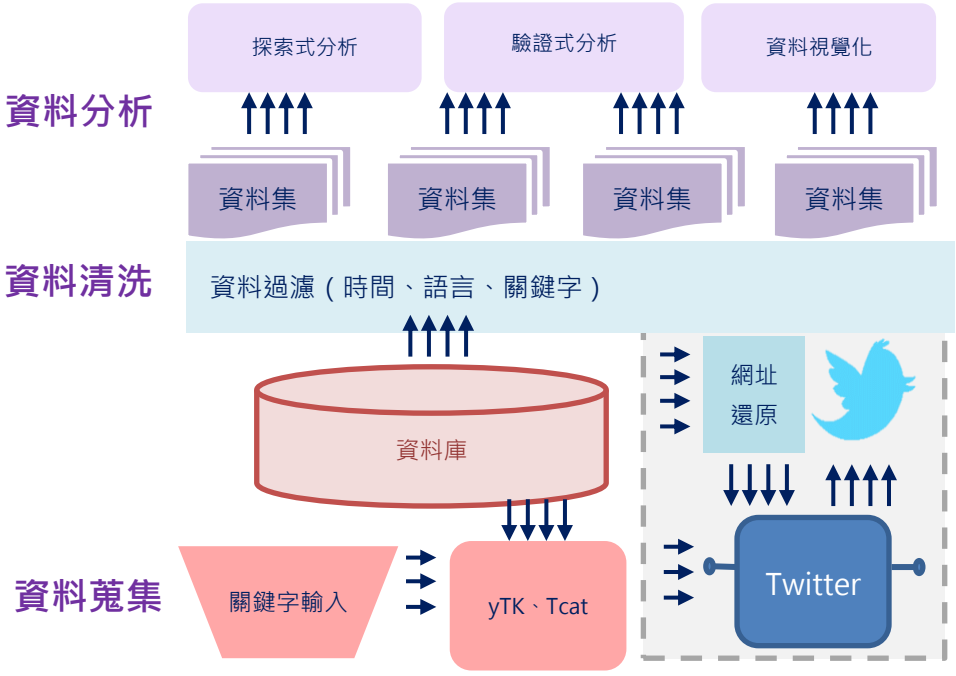
從問題發想和設定歷程看來，社群媒體資料分析和其它學術或實務的研究歷程並無大差別，但另一方面，社群媒體的資料分析既以鉅量資料為對象，和一般研究不同之處，在於充滿各種數據資料處理的細節。誠如前述，社群媒體資料主要由人產生，這類資料集不僅數量龐大、具有多種形式、充滿雜訊，而又不盡完整。因此不僅從提問到解題之間，需要藉由資料科學知識的協助，運用資料來回答心中的疑惑。因此有必要說明資料處理歷程。

二、資料處理（data processing）

另一個主要歷程是資料處理。包含五個組成元素：（1）資料／後設資料（data／metadata）：後設資料是關於資料的資料，是資料的結構信息。例如，資料生成的地理、時間資訊或暫存檔（cookies）；（2）演算法

（Algorithm）：意指平臺利用某種公式與參數來計算社交互動，演算法決定了社交平臺的競爭力，通常是商業機密；（3）通訊協定（Protocol）：用以統一不同系統的資料格式，隱蔽地引導用戶行為朝向管理者偏好方向前進；（4）介面（Interface）：可視介面意指客戶終端介面，通常是圖像化、易操作的，而不可視介面則是用來聯結軟體與硬體，API（Application Program Interface）則介於二者之間；（5）預設值（Default）：軟體中的預設值，具有引導用戶的功能（van Dijck, 2013）。社群媒體的資料處理大致可歸納為三個階段，亦即資料蒐集（data collection）、資料清洗（data clean-up），以及資料分析（data analysis）。如下圖所示

圖一：社群媒體資料處理過程的示意圖



資料來源：水火計畫團隊（2015）

（一）資料蒐集

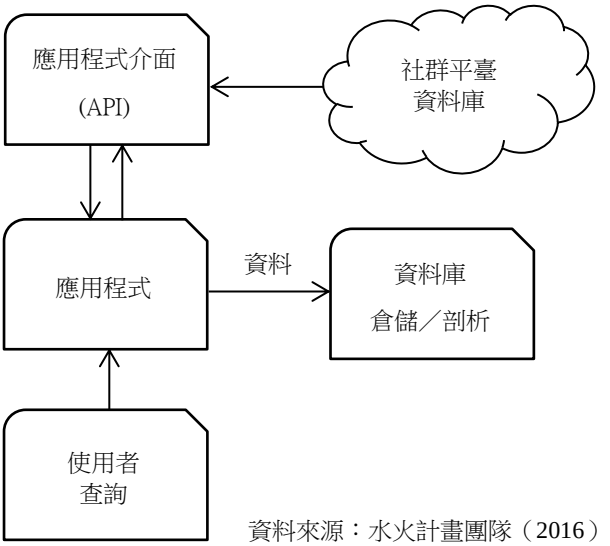
資料蒐集是社群資料處理的首發階段，也就是直接從社群媒體平臺上進行追蹤（tracking）或觀測（monitoring），擷取、清洗並整理特定平臺或閱聽人留下的數位足跡，使之成為一個可以用於分析的對象。

1. 資料來源

由於社群媒體資料數量龐大，因此資料擷取多仰賴資料科學技術，透過自動化的程式擷取資料。經過資訊技術擷取程序而獲得的特定資料的集合體，通常稱為資料集（dataset）。一般而言，在社群媒體上獲取資料集可區分為兩種方式：一種稱為「撈資料」（access by API）；另一種方式則是「扒資料」（RSS／HTMAL parsing）。

所謂「撈資料」則是由研究人員撰寫程式，根據社群媒體提供的欄位規格和權限，登入平臺業者建立的應用程式介面（application programming interface, API），程式中設定若干詞彙（例如，事件相關的關鍵詞或使用者名稱），從社群媒體伺服器擷取資料，然後下載並儲存至資料庫中，等待後續的資料清洗和分析。

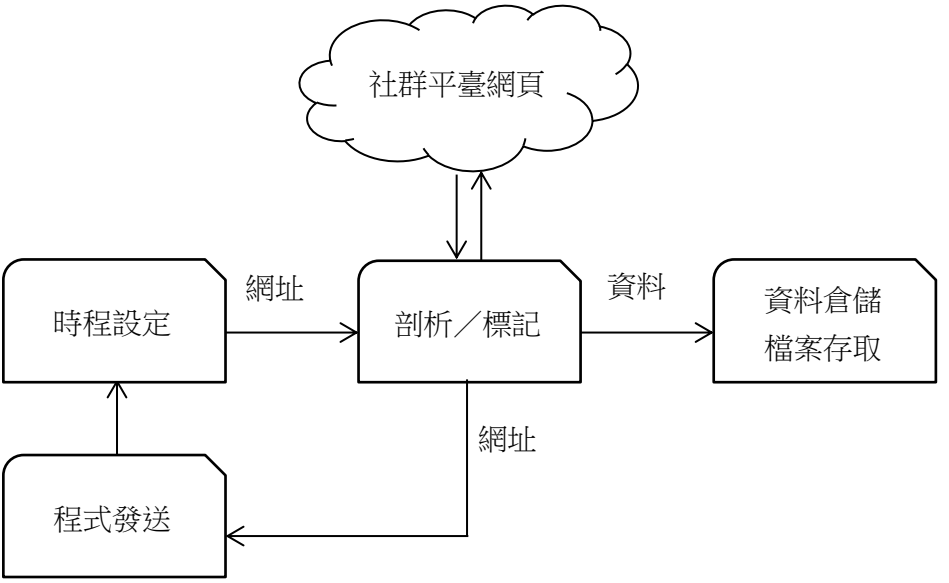
圖二：從社群平臺應用程式介面（API）撈取資料的示意圖



資料來源：水火計畫團隊（2016）

另一方面，所謂「扒資料」是指以程式模擬人類使用網頁的行為，直接從社群網站的頁面擷取資料。使用者針對社群平臺網頁，撰寫爬蟲程式（web crawler），模擬人類閱讀網頁的行為。這些爬蟲程式通常可以設定時程，向特定網頁發送訊息，以程式逐一擷取網頁中個別欄位的資料。爬蟲程式一旦擷取到資料（文本和後設資料）隨即下載到資料庫中，進行剖析和註記（parsing/marking-up），然後儲存到資料庫，等待研究者清洗和分析之用。但因扒資料必須按照網頁排版格式，於是當業者變更網頁版式，爬蟲程式也需要不斷隨之而更新。

圖三：從社群平臺網頁扒取資料的示意圖



資料來源：水火計畫團隊（2016）

無論使用那種類型的程式擷取資料，需要投資相當程度的軟硬體和人力。因此，傳播學者或實務工作者為特定目的而擷取資料，需要通過一定的技術和經濟門檻，倘若無法自行建立擷取資料的機制，則另一途徑是透過資料供應廠商（data providers）代為蒐集社群媒體資料。資料供應廠商透過上述機制固定撈取並儲存社群資料、使用者支付對價以取用廠商整理

過的資料。

2. 資料蒐集對象

資料蒐集是一種手段，通常因研究目標而異。不同的資料蒐集衍生出不同的資料類型取徑。Stieglitz & Dang-Xuan（2012: 1284-5）區分以下五種蒐集對象：

（1）以涉己事務為對象，而進行追蹤或觀測（self-involved approach）：特定機構或個人為瞭解網路社群如何看待自己，因而針對社群言論所進行的資料蒐集。包括：政黨、企業或政商人物，為瞭解自己在網路社群言論所獲的風評，所進行的資料蒐集。

（2）以特定角色為對象，而進行追蹤或觀測（actor-based approach）：研究者針對自己以外的特定組織、團體或人物所進行的資料蒐集，通常是長期的、持續的追蹤和觀測。例如，針對政治人物馬英九、蔡英文的觀測，或者針對國內兩大超商龍頭的觀測。

（3）以事件／議題（event/topics，亦稱為熱點，hot spot）為對象，進行追蹤／觀測：研究者針對網路上重大事件或討論議題所進行的資料蒐集，以瞭解網路社群對此一事件／議題的看法或評價。通常這種資料討論數量或規模具有一定的生命週期。例如，因劣質油品引發的食品安全議題，以及國際間的海域爭議，都屬於這類資料。

（4）以隨機取樣或探索方式（random/exploratory approach）進行追蹤／觀測：研究者從特定空間或時間區段，以隨機抽樣方式選取一群資料，然後再針對這個較小規模的資料集進行分析。前述幾種資料蒐集標的相對較為明確，但在不確定性或題旨尚未浮現的情境下，使用小規模資料先進行探索、比較、對照，以發現可能的問題或研究線索。

（5）以超連結為基礎（URL-based approach）的資料蒐集：社群媒體使用者經常轉述或引用新聞或社群媒體的貼文，以瞭解新聞訊息的擴散途徑和文本內涵。這些貼文通常以超連結指向特定網址，例如：根據重大空難事

件中使用者所分享或引述的新聞或社群媒體類型。

上述 Stieglitz & Dang-Xuan（2012）的分類，雖然主要針對政治傳播場域而蒐集資料，但其後文獻陸續加以沿用（例如，Stieglitz, Dang-Xuan, Bruns & Neuberger, 2014）。重大公共事件的社群媒體資料蒐集，往往亦不脫離其範疇，因此這項分類對於社會科學研究人員可能相當具有參考價值。

3. 關鍵詞／熱門詞組

社群媒體資料筆數動輒成千上萬，研究人員針對特定議題撈取資料時，有如海底撈針。現階段的資料分析人員為有效使用資訊技術資源，通常透過關鍵詞組或熱門詞彙等途徑在 API 撈取資料，或從使用者建立的文本中抽取意欲分析的內容。

社群最常撈取資料的方式是使用關鍵詞（keywords），根據一個或多個關鍵詞從社群媒體辨識或擷取目標資料。關鍵詞通常用於特定事件的資料撈取（例如，總統大選、社會運動、重大災難等）。關鍵詞設定或來自研究人員經驗、或是經閱讀資料之後所下的判斷。例如，在蒐集空難事件資料時，航空公司、航班呼號或失事地點。例如，使用馬航、Malaysia Airlines、或 MH370 等詞彙表徵 2014 年 3 月的馬航班機失蹤事件。

撈取資料所使用的關鍵詞，除了由研究者依據研究目的而設定的關鍵詞，也使用社群媒體平臺提供的熱詞標籤（hashtag）。所謂熱詞標籤是社群媒體使用者依據溝通目的而自行設定的字串，通常用來標示、歸類和串連社群議題。例如，Twitter 使用者採用 ＃Orlando 標示 2016 年 6 月間發生在美國佛州奧蘭多市的槍擊案。研究人員透過比對關鍵詞／熱詞標籤在特定期間的的數量，瞭解此一槍擊事件議題在社群媒體上的討論聲浪和範圍。

（二）資料清洗

每一個資料集內容可能是包括文字、數字、圖像或其它格式。但由於社群媒體資料含有大量非結構性資料，因此研究人員必須清洗資料。

清洗資料集之目的在於統一格式、剔除雜訊，以及將資料適度縮減至較適合處理的規模，是資料處理過程中最為費時費工的過程，如前所述，社群媒體資料主要由人產生、內容以非結構化資料為主，倘若要使用自動化的資訊技術處理資料，則須先將資料調整成為機器可辨識的格式。

資料清洗通常可歸納為以下幾項程序，亦即：格式處理（format processing）、選擇過濾（selection/filtering）、合併（integration），或者拆分（separation）。

1. 資料格式處理

研究人員將資料格式轉換或歸整為所欲處理的態樣。例如，社群媒體的發文時間，通常以 15 個數字的字串記錄下來，通常必須透過函數處理，才能轉換成為年／月／日格式，又如許多文本資料原存為 Big-5 格式，須統一轉換為 UTF-8 格式，較有利於一般軟體運算。

資料經由格式處理之後，通常還需要加以剪裁，才能使用於分析。剪裁是透過資料選擇、過濾、合併或拆分等處理程序。因此需要選擇和篩選資料。

2. 選擇資料

格式處理之後的資料集，排列有如一到多個矩陣。可以數張表單（table）儲存之，每一張表單又以欄（column）和列（row）的方式收納社群媒體資料。社群媒體資料通常可以區分為許多維度，例如使用者帳號、發文時間、留言內容等等。這些維度儲存在每一個垂直的欄位，所有欄位自左而右排列。所謂選擇資料，是指研究人員根據分析的需求，辨識和保留相關的欄位。

3. 過濾／篩選資料

每個資料集都由許多筆資料構成，每一筆資料以水平方式展現，並依照個別欄位的順序呈現其參數。每一筆資料在水平方向都依照欄位排序而存在固定位置。所謂過濾／篩選資料，是指研究人員根據分析需求，根據

個別欄位參數範圍，下達保留或刪除某筆資料的指令。例如，研究人員若要刪除重複出現使用者，是將使用者編號欄位下，凡出現兩次以上者加以刪除，使得每一筆使用者編號只出現一次。

4. 合併／拆分資料

通常研究者所欲分析的資料集當中，原本沒有可對應分析的欄位，因此研究員必須根據研究題旨，將不同欄位加以組合或拆解，因而創造出新欄位，俾供分析之用。所謂合併資料，是將數個欄位合為單一欄位；例如，按讚、留言、分享次數原本分別存在三個欄位，研究人員擬將三者統合為一個參數，於是將此三個欄位參數經過加權處理之後，存至一個新的欄位。反之，拆解資料則是將單一欄位分散為數個欄位；例如，發文時間欄位原本包含年月日三者，研究人員因目的要分析月份，將年月日區分為三個個別欄位。

上述資料經由格式處理、選擇、過濾、合併或拆分等行動，都是研究人員針對資料所進行的操弄，目的讓資料便於進一步分析。

在社群媒體資料新的過程當中，「資料維度」（dimensions）是指一群具有共同屬性資料的集合。社群資料當中，最常見的資料維度類型包括：時間（temporal）、空間（spatial）、數值（numeric）、類型（categorical）、關連（relational）。例如，發文時間，是以年／月／日等欄位呈現在資料集當中，這幾個欄位共同展現時間維度。發文地點則是空間維度，以經度和緯度等欄位共同呈現。這些維度經過清洗和歸類，便成為研究者所欲探索的目標。

然而，並非所有資料維度都自然而然對應於研究問題，且能立即適用於資料分析。通常研究人員會根據特定研究目的，組織資料維度，藉以建構出該研究的指標。

所謂「資料指標」（metrics）是指根據研究目的，使用資料維度所建立的一組函數公式，通常用來當做研究變項。例如，臉書的品牌監測人員為了要觀測或評估臉書粉絲專頁的經營成效，將臉書資料集中的按讚數、

留言數和分享數整合成為一組函數：

專頁每日互動率＝（當日按讚數＋留言數＋分享數）／該日粉絲數字 x 100%。

上述例子當中，研究人員根據研究需求「按讚數」、「留言數」和「分享數」，從資料維度中擷取、組合並加權，形成「專頁每日互動率」的指標，用來指涉粉絲專頁和閱聽人之間的「互動率」（engagement rate）。

上述「互動率」指標是由研究者建構而來。儘管資料擁有相同的數量、規模或維度，但由於指標建構方式不同，所得到的結果可能也會不同（例如，視按讚、留言、分享參與程度有別而予以加權）。因此指標是否有效，往往還需要加以評估和驗證。

就上述「互動率」指標所使用的函數而言，當研究人員將按讚數、分享數、評論數視為同值，當然可以逕行加總而呈現上述函數，但倘若研究人員認為上述活動參與程度並不相等（例如 Mayfield, 2006；Forte & Lampe, 2013），從而認為賦予量化數值時，應該有所區別（例如，認為分享數>評論數>按讚數），則在統合上述三個數值時，就可能需要考慮加權或其它資料處理方式。換言之，研究人員會根據他或她的理論，建立一套函數公式，用於計算數據。因此，儘管使用同一套數據，但研究人員從其理論出發而採納各種不同計算指標，所得到的分析結果或許也就不盡相同。

（三）資料分析

1. 探索式 VS.驗證式分析

根據資料之目的，資料分析通常可區分為兩種類型：探索式分析和驗證式分析。

「探索性資料分析」（exploratory data analysis, EDA），是指研究人員透過簡單的統計指標和視覺化工具，發現資料分佈態樣或走勢，作為進一步資料分析的基礎（Tukey, 1977；Seltman, 2015）。探索性資料分析方法經常應用在資料處理初始階段、資料意涵不明確之時。由於鉅量資料的特

徵在於樣本相當接近母體（林俊宏譯，2013），而各種資料集皆具有其獨特性，因此在分析之初，需要探索資料樣態，探索性資料分析，正如其字面所顯示：兼具探索的（explorative），以及假設的（hypothetical）歷程。它具有探索的意涵；研究人員使用簡單的統計方法和視覺化工具，讓資料態樣浮現出來。它同時也是具有假設的意涵，研究人員針對高度不確定的資料，試圖觀察資料分佈態樣和可能的異狀，以找出提問重點。例如，在重大公共事件發生之際，使用發文時間和則數建立簡單的時序模型，觀察貼文則數的集中和離散趨勢，並以圖像呈現時間關係，便能立即讓研究人員瞭解社群聲浪或走勢，以進行後續分析。

另一方面，「驗證式資料分析」則是透過分類、分群、關連以及預測等手段，針對初探結果進行進一步的假設驗證，包括使用各種統計工具建立統計模型。此外，研究人員也可能根據量化分析初探結果，縮小資料規模，進行質性分析，以發現資料內涵。

2. 資料分析類型

社群媒體資料的分析工具，迥異於傳統社會科學分析方法。本文討論以下幾種類型：時序分析、關連分析、數值分析、文本分析、以及情緒分析。這些類型並非窮舉，而僅是例示；隨著資訊科學技術和研究方法的演化，未來資料分析類型將更趨多元。以下分別說明個別分析類型內涵：

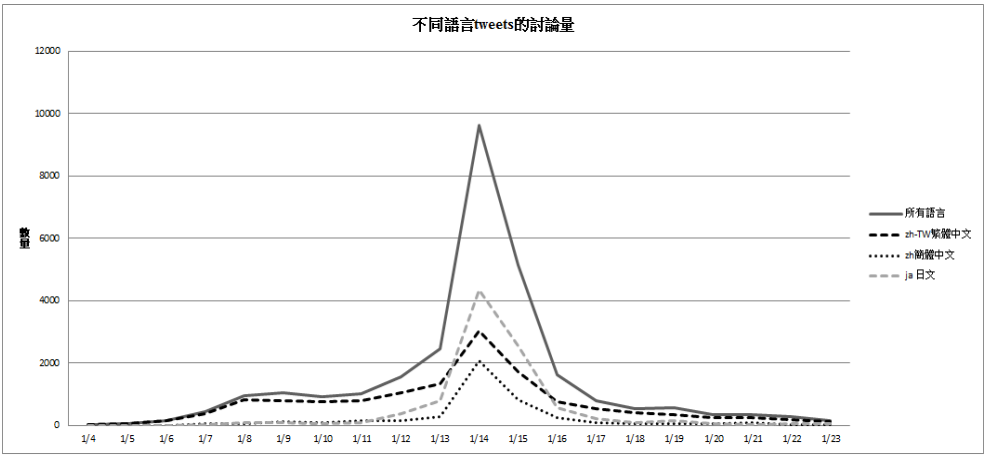
（1）時序分析（temporal/trend analysis）

通常使用時間和其它維度資料為素材，旨在觀察特定期間當中社群媒體資料變項根據時序所產生的變化。例如，研究人員透過災難事件期間的發文數量或特定文本出現的頻率變動，觀察社群媒體使用者關注程度的變化。此時以時間順序為 X 軸，而以發文數量或頻率為 Y 軸，藉以說明時序下的數量規模變動。例如，Burgess 與 Bruns（2012）針對 2010 年澳洲大選期間推文，探討選舉期間人們如何群聚在 Twitter 上討論選舉。他們搜集大選前後共 38 天熱詞標籤為 #ausvote 貼文共 41.5 萬則，比較不同時間點的發文數量，結果發現，22% 貼文發佈於選舉當天；但發文者時序觀察，

則 38 天裡曾發文的 3.7 萬使用者當中，僅 1.9 萬人（51%）是在選舉當天發文。換言之，半數參與者在選舉當天才對此一選舉事件發表意見。

時序分析也可用於預測議題趨勢或生命週期。這類分析根據資訊科學／統計建構的演算模型（例如 Hidden Markov Model），偵測並記錄社群媒體討論的生命週期，並將各種時序相關模型儲存於資料庫中，用來預測社群討論的發展趨勢（Stieglitz, Dang-Xuan, Bruns & Neuberger, 2014: 92）。下圖呈現的是 2012 年總統大選兩岸三地的 Twitter 資料分析，顯示三個語系族群（繁體中文、簡體中文和日文），在選舉前後兩週時間內的貼文數量分佈。以繁體中文為主的族群，討論聲浪崛起顯然早於其它兩個語系族群（鄭宇君、陳百齡，2014）。

圖四：2012 年總統大選推文的時序分析



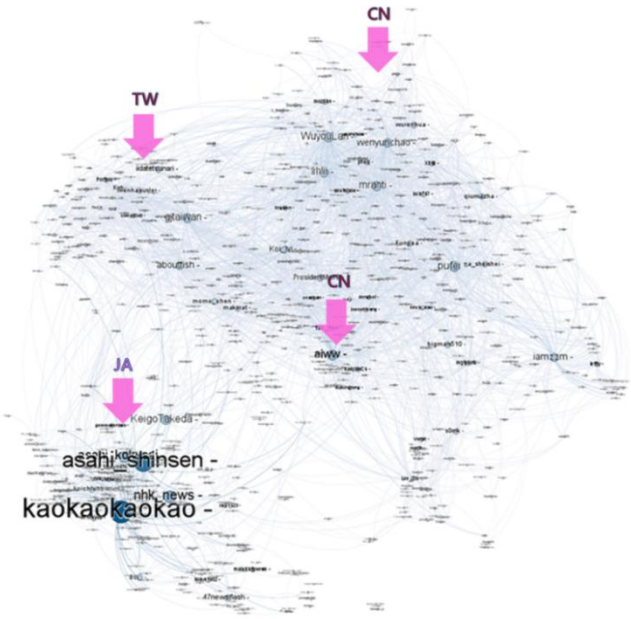
資料來源：鄭宇君、陳百齡 （2014）

（2）關連分析（relational analysis）

關連分析使用社群平臺產出的關聯資料（relational data），其主要目的在觀察使用者之間的關係脈絡。例如，當推特平臺使用者轉貼（retweet）另一使用者文章，便在平臺上留下轉貼／被轉貼者帳號，此一紀錄關連兩位使用者，可視為網絡節點（nodes）和鏈結（links），上述資料經過萃取

和編碼，可進行分析，以表徵轉貼者和被轉貼者之間的社會關係。例如，Kogan（2015）等人蒐集 2012 年美國桑迪颱風（Hurricane Sandy）期間相關 Tweeter 推文／轉推文，根據貼文上的地理位置標記分類出受災區域用戶和非受災區域用戶，再根據天災事件前、中、後期時序，再以推文／轉推文作者所構成的關連資料，以推文／轉推文作者當做節點，區分出四種社會網絡。研究指出，颱風侵襲當下災區 Twitter 使用者，較災難前後，發佈更多的信息，這些貼文透過轉貼，形成災難當下相互連結而又緊密的社會網絡。下圖資料同樣來自 2012 年總統大選兩岸三地的 Twitter 資料，但以推文者／轉推者之間的關連資料為分析對象。透過 Gephi 網絡圖像軟體的輔助，這幅網絡圖所要呈現的意涵是：雖然三個語系族群（繁體中文、簡體中文和日文），都聚焦討論 2012 年總統大選這個議題，但是三地用戶轉貼的對象各自不同，說明在這個事件當中，Twitter 用戶針對一個共同議題、三個族群各自表述的現象。

圖五：2012 年總統大選的語系族群的網絡分析



資料來源：鄭宇君、陳百齡（2014）

（3）數值／類型分析（numerical/type analysis）

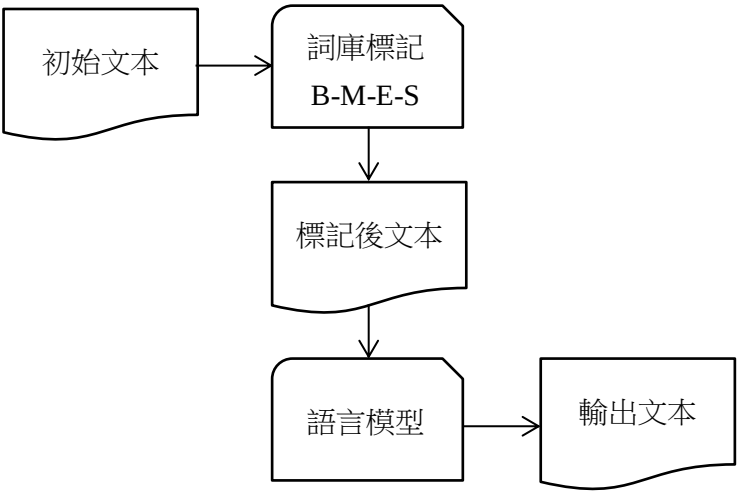
數值／類型分析關注兩個或多個資料維度之間的關係，通常把兩群資料維度當作變數，透過統計工具（如 SPSS 或 SAS），進行交叉分析。資料維度可以是數值型資料或者類別型資料。研究人員根據統計結果，說明資料維度之間的關係。例如，Bruns（2014）等人分析熱詞標籤（hashtag）和轉貼文之間的關係。例如，Burns 與 Burgess（2013）針對「阿拉伯之春」期間，在埃及和利比亞等國所發生的民眾抗爭事件，使用 Twitter 上熱詞標籤 #egypt 和 #libya 的數百萬則貼文，透過語言辨識工具整理出拉丁語（主要為英語）和非拉丁語（主要為阿拉伯語）的使用者族群，並以推文數量區分不同語系中「活躍使用者」、「高度參與者」和「比較不活躍使用者」，以探討 Twitter 上的相關發文態樣，特別是不同語言團體之間的互動。研究指出，從熱詞標籤 #egypt 和 #libya 的貼文數量觀察，埃及和利比亞兩地在推文數量和變化方面呈現差異。埃及最活躍的 1% 使用者大多使用 #egypt 這個熱詞標籤，特別當推特使用者最常使用的熱詞標籤從 #Jan25 轉到 #egypt 之後，阿拉伯語言使用者大增，甚至超過英語使用者，而主導了整個社群討論。本研究根據推文語系特徵，找出不同語言的社群而進行數量比對，從而發現不同語言社群在事件訊息傳遞的差異。

（4）文本分析（text analysis）

傳統媒體的文本分析，是由研究人員建立類目，以人為編碼員，以版面或單篇文章抽取樣本進行分析，稱為「內容分析」（content analysis）。社群媒體資料數量規模龐大，人工分析無法因應，晚近資料科學技術學門以自然語言處理（natural language processing, NLP）方法進行分析，分析稱為「文本分析」或「電腦輔助文本分析」。這種文本分析是以詞為單位，再進行詞頻統計或詞彙關連分析。中文詞彙又比英文詞彙複雜一些，主要因為中文詞彙之間沒有間距，因此必須透過斷詞（word segmentation）技術，將語句中的詞類加以切割，再除去停用詞之後，才能進一步文本分析。以下圖例描述使用監督式學習法的中文斷詞演算法：初始文本先置入於辭典中進行切割和標記，分別根據單字在詞彙的位置而註記詞首、中段詞尾

或單字詞，經過標記的文本置入語言模型中進行學習，最後輸出經過斷詞的文本，提供進一步分析之用。

圖六：中文斷詞程序（使用監督式學習法）



資料來源：中央研究院詞庫小組（2015）

大量社群媒體文本經過斷詞之後，研究人員即可運用自動化程式處理這些詞彙，包括：分析詞彙出現的頻率（word frequency）、或詞語彙在特定段落共同出現的機會（co-occurrence）。有些研究聚焦探討詞彙之間的群聚關係，例如，探討詞彙指向共同議題的機率，藉此建構出大量文本中的議題模型（topic modeling）。另外，探索大量文本中的詞彙的語意網絡，則是將文本中的詞彙視為網絡節點，尋找詞語之間的關連，試圖刻畫出詞彙網絡所構築的詞語意涵。例如，Chew 與 Eysenbach（2010）以「豬流感」（Swine flu）相關關鍵詞，蒐集了 2009 年 5 月至 12 月間 Twitter 上兩百多萬則貼文，分別進行人工編碼與情緒分析，再對所有貼文進行自動編碼，區分出幾類訊息來源（例如，新聞媒體、官方醫療組織），並進行貼文數量分析和情緒分析。結果發現， Twitter 使用者在談論豬流感時，會出現關鍵字變化（例如，使用「H1N1」術語的貼文比例從 8.8%攀升到 40.5%）；

另外，情緒也會變動（例如，隨著疫情升溫「幽默」情緒的貼文降低，而「挫敗」情緒貼文增加）。因此，研究者認為，Twitter 可用於醫療機構即時觀測疫情的重要資料源。下圖呈現的是 2009 年莫拉克風災期間網友在災情資訊網站張貼的文本，經過斷詞之後以時序呈現的分佈狀態，顯示在風災前期詞頻數量多集中在資訊獲取，但在風災中後期，詞頻逐漸集中在人員物資調度和共識的建立。

上述以計算機輔助所構成的文本分析，不僅以自動化取代傳統人工編碼。更重要的是，研究的分析單位（unit of analysis）是以詞彙為單位，比起傳統內容分析以篇章段落或版面尺寸為單位，或許更為細緻。

圖七：2009 年莫拉克風災的詞頻時序分布圖

資料來源：陳百齡、鄭宇君（2014）

（5）情緒分析（sentiment analysis）

情緒分析可視為一種特定的文本分析。研究者先從文本中挑選出目標詞彙，然後比對這群詞彙的情緒特徵，藉以判斷文本屬於正向或負向情緒（Stieglitz, Dang-Xuan, Bruns & Neuberger, 2014: 92）。情緒分析通常必須

先斷詞，然後根據事先建構的情緒辭典（事先經過情緒屬性分類的詞彙群組），或者使用機器學習方法，以人工監督或全自動方式，[3] 辨識詞彙的正／負情緒類別，經過統計和整合，藉以判斷該文本的情緒傾向。例如，美國學者 O'Connor（2010）曾經使用 Twitter 資料分析推文涉及消費者信心關鍵詞，並和民調機構所進行的調查結果相互對照。作者使用軟體分析文本中的情緒面向，結果發現，Twitter 推文的情緒用詞之詞頻和傳統民調中的消費者信心指數在某些條件下呈現正相關，在特定議題上相關係數更高達 80%。這個研究結果似乎顯示，透過文本的情緒分析，具有替代或補充傳統民意調查的潛力。因此作者認為，社群媒體資料分析可用於測量民意以及預測消費者信心。

3. 資料視覺化

資料視覺化（data visualization）是透過軟體工具輔助，將資料從數字轉換為圖像媒材，藉以展示資料的分佈型態或趨勢。一般而言，社群媒體資料分析過程中，視覺化通常使用於兩個時機：一是將視覺化作為探索工具。另一時機則是將視覺化當作敘事工具，通常發生在資料分析階段後期，主要將資料分析結果外顯化，透過圖像方式展示，和閱聽人溝通。當視覺化用於敘事，研究人員根據資料維度、分析型態，以及圖表文類（graphic genre）決定圖像如何呈現。

4. 分析類型的運用和組合

研究者在探索或解決一個問題時，研究問題（目標）和資料分析類型（手段）之間，並非一對一的對應關係；一個研究可能需要動用數種類型的分析工具。從過程觀察，研究人員在研究主題尚未明朗之前，先針對社群媒體資料進行探索式分析，例如使用較簡單的時序分析或數值分析，先描繪出事件整體時間輪廓和討論熱點，接著再根據熱點建構出不同時間點使用的關鍵詞組，以採掘文本內容或參與者，最後進行語意網絡分析或使

[3] 學者進行情緒分析時最常使用的三種技術，包括：情緒詞典法（dictionary-based methods）、監督式的機器學習法（supervised machine learning methods）以及非監督式機器學習法（non-supervised machine learning methods）。這三種資料分析技術的細節描述，參見：Boumans & Trilling (2015: 11-16)

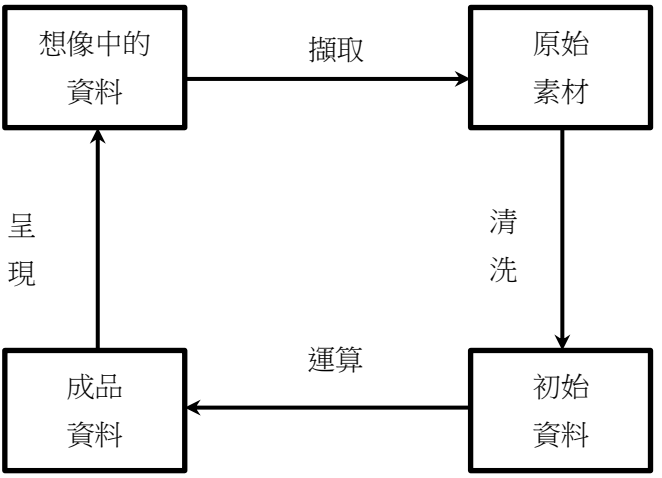
用者社會網絡分析。例如，Burgess 和 Bruns（2012）的澳洲大選的推文研究，研究人員先呈現整體資料，然後再針對貼文和語言族群，進行進一步的數值分析。鄭宇君和陳百齡（2014）分析 2012 年臺灣總統大選兩岸三地推文分析，也採取同樣策略，先以時序分析呈現整體趨勢，再根據浮現的核心議題，區分群組進行比較性的數值分析。

另一方面，要回答一個問題往往可能組合多種分析類型。以上述分析類型的個案為例，在研究過程中可以看見時序分析、文本分析和社會網絡分析等不同分析類型，因應研究人員探索資料的需求，而被組合起來。例如，前述 Chew 和 Eysenbach（2010）的豬流感研究，不僅使用數值分析，同時也使用文本分析和情緒分析。

三、資料演化：從提問答題和資料處理

如上所述，社群媒體資料分析的雙重歷程，一方面是從提問到答題的歷程，另一方面也是資料分析、清洗和分析的歷程。這兩組歷程共同涉及的是資料。在雙重歷程之間，所擷取與產製的資料不斷演化。我們且用以下圖示，說明兩個歷程如何讓資料演化。

圖八：提問解題和資料處理之間的關係



資料來源：水火計畫團隊（2016）

1. 蒐集資料

社群媒體資料分析通常以提問為起點。在資料分析最初，研究人員根據研究問題而想像資料的模樣。例如，應該從哪些平臺尋找何種資料？應該使用什麼方法擷取資料？研究團隊在想像同時，也進入實際資料處理的蒐集階段，根據想像，設定關鍵詞組，再將關鍵詞組納入透過 API 擷取程式（或設定爬蟲程式），從社群媒體平臺提取和匯集資料，而獲得原始素材。

2. 清洗資料

從社群平臺擷取的素材可能存在許多亂碼、資料缺漏，或者格式錯置。研究人員必須將這些原始素材轉換為可以直觀和取用的資料格式，例如轉換字元編碼（Big5→UTF8）或將格林威治時間轉換為臺灣時間。由於研究人員透過種種手段將資料集結構化，變成可探索的初始資料。在此同時，研究人員也進行資料的探索式分析，使用簡單的視覺化圖表（例如，透過時間和貼文或使用者數量排比，瞭解資料的生命週期），或使用描述統計（如平均值、標準差等），快速發現手頭資料的態樣，並推測這些資料是否完整？清洗是否足夠？有無雜訊？或是資料蒐集過程產生問題？甚至於資料集本身是否可用？

3. 運算／驗證資料

初始資料雖可供探索，但仍須經過研究人員根據資料維度而選擇、合併或拆解欄位，透過參數過濾資料筆數，始得成為可以執行運算指令的對象。此時，研究人員根據研究問題，選擇和組合所欲進行分析的類型，實際進行分析。例如，把所有用戶轉貼他人貼文的資料，建立一組關連資料，從而進行社會網絡分析；或者先進行一種分析之後，從分析結果再進一步進行另一種分析。換句話說，分析做為一種發現意義的手段，會在研究過程中會不斷發生。

當研究人員進行各種分析的同時，也應該根據分析結果，判斷分析結果是否回應研究問題。倘若資料無法回答問題，則再依照問題、資料和分

析結果等各環節逐一審視和反思。評估資料集和提問之間的關係，包括：資料維度建構是否構成指標？能否提供分析之用？這些指標是否足以回答研究提問？換言之，研究人員針對問題表徵（representation）和資料維度／指標進行比對，試圖發現二者之間能否對應。

最後，則是以視覺化手段呈現資料：當研究人員確認資料足以提供分析，則選擇工具將資料做整體的呈現，作為論證和詮釋之用。

在進行資料分析的過程當中，問題發現和解題歷程，以及資料處理過程，看起來似乎是兩條平行線，但實際上二者互為關連，並非各自獨立，因此研究人員必須隨時關照自己的問題和資料處理是否相呼應。這個歷程單單使用上述文字說明，或許很難理解。因此，本文將以一個資料分析個案進行說明。

肆、個案

2015 年 2 月 4 日上午 10 時 55 分，復興航空公司一架編號 G235 的 ATR 72-600 型民航班機，搭載旅客和機組人員共 58 人，從臺北松山機場起飛，預定前往金門尚義機場。班機起飛十分鐘後，即墜毀於臺北市南港區基隆河，此一空難造成 43 人死亡，屬於突發性的重大公共安全事件。在飛機墜落當下，有目擊者以手機拍攝民航機墜河過程，並隨即將手機影像上傳社群平臺推特（Twitter），經網路使用者大量轉載，引發社群媒體關注。個案中的研究團隊是某大學社群媒體研究團隊 A 團隊。A 團隊累積五年以上長期研究經驗，成員擁有跨領域知識背景，本則個案為同時期進行數個研究案之一，研究小組由三人組成，二人來自傳播，一人來自資訊科學。

本個案屬於事件型（event-based）個案。我們用這個事件說明社群媒體資料分析的歷程。一方面是因為在社群媒體資料分析領域，熱點事件分析經常出現。這類熱點事件呈現「短時間、高密度」的特性，並具有生命週期。另一方面，也因為空難事件引發國際關注，產生龐大的資料。具有較高的資料飽和度（data saturation），因此個案提供的資料較一般國內事件更多元。

以下個案目的在描述：研究人員如何透過問題和資料處理的交錯，形構資料分析的行動。本文擬以研究小組人員所建立的手稿和分析資料作為主要素材，佐以事後訪談。試圖回答以下兩個問題：（1）社群資料分析歷程中，分析人員如何和情境互動？（2）具體而言，分析人員資料進行哪些行動？

一、 發想和蒐集資料

在空難發生當天，極短時間內即吸引大量社群媒體使用者推文，這個現象引起研究團隊人員的注意。於是當天就啟動資料蒐集。研究人員必須立即決定幾件事：問題是什麼？用什麼資料？使用什麼方法蒐集資料？如何聚焦研究主題？

研究團隊長期關注重大公共事件期間人們如何使用社群媒體。空難和選舉或社會運動最大差別在於空難這種突發性的災難，無法預期，資料容易流失，由於收集資料困難，先前研究也就很有限。由於事出突然，研究人員最初提問較為廣泛：「社群媒體如何傳佈空難事件？」

其次，從什麼社群平臺擷取資料？空難事件通常也是國際事件，因此必須考量國際社會普遍使用的社群平臺。Twitter 平臺對於突發事件反應快速，在歐美國家即時訊息都是先在 Twitter 上曝光和擴散之後，然後再往其它社群媒體平臺擴散（Spasojevic, Li, Rao & Bhattacharyya, 2015）。其次，Twitter 也是最方便全球不同語言社群的用戶交換訊息的平臺（Bruns, Highfield & Burguess, 2013）。另一方面，社群媒體資料蒐集是一項技術密集的工作，須考量資料蒐集效率和系統穩定性，團隊限於預算額度無法從商業機構購買資料，而自家使用（in-house）的系統又必須使用既存的最佳技術。基於上述種種因素，最後選擇以推特為資料蒐集對象。

第三、資料蒐集方法為何？本個案中，A 團隊採用開源軟體改寫的程式，作為蒐集社群媒體資料的主要方法。這個選擇是因為撰寫和維護程式不易，而社群平臺經常改版或變更 API 格式，採用開源程式 DMI-TCAT，

由社群資料蒐集的社群共同維護，大幅降低蒐集資料成本。[4] 由於開源軟體可以在基本款之上增加新功能，因此具有擴充性。例如，原本蒐集軟體並未區分繁體和簡體中文，但因先前研究兩岸三地社群媒體討論，必須區分中文語系，因此寫入中文簡繁體的語系判讀功能。由於 DMI-TCAT 仍是使用 Twitter API 蒐集資料，因此仍受 Twitter API 本身限制所影響，像是 Twitter stream API 不支援中文關鍵字，因此無法使用收集串流推文，因此透過中文關鍵字收集推文，只能使用 Twitter search API 檢索推文功能。

最後，檢索推文需要先設定關詞組，應使用哪些關鍵詞組？社群媒體使用者採用的詞彙，未必和專業社群所用的詞彙相同。檢索貼文所用的關鍵詞組必須考量儘可能貼近使用者的用法。最後設定的關鍵詞組是以航空公司名稱、航班編號，以及機型為主的中英文詞彙（包括簡繁中文）共 9 個關鍵詞。[5] 自空難當天（2015 年 2 月 4 日）中午啟動 DMI-TCAT 從 Search API 撈資料，至 2 月 16 日中午為止，共取得 229,559 個使用者所發佈的 508,666 則貼文。這些貼文和其相關資料，即研究團隊所欲處理的原始素材。

二、 清洗資料

當原始素材齊備，下一個階段是資料的清洗。研究人員進行資料清洗之前，必須先探索以下幾件事：資料存在哪些格式？資料是什麼態樣？如何選取適當資料維度進行分析？

首先，資料結構為何？資料蒐集工具軟體 DMI-TCAT 從 Twitter API 撈取資料以後，將所撈得的原始資料欄位重新分類組合，並儲存在數個小型的關連式資料庫當中。例如，除上述舉例，原始素材還包含許多資料群組，如推文（Tweets）、提及（@mention）、網址（URL）、熱詞（hashtags）等。

---

[4] A 團隊採用的 DMI-TCAT（Twitter Capture and Analysis Toolset for Digital Methods Initiative），是由荷蘭阿姆斯特丹大學 Bernhard Rieder 團隊研發的資料蒐集軟體工具集，此工具可擷取 Twitter 串流推文（stream API）和檢索推文（search API），除資料蒐集功能外，也兼具簡單的資料分析功能（Borra, & Rieder, 2014);Felt, 2016:11-13）。該軟體置於 GitHub 上供公眾取用。參見：https://github.com/digitalmethodsinitiative/dmi-tcat/wiki/FAQ 。

[5] 這 9 個關鍵詞為：B2286, B22816, 復興航空, ATR72, GE235, TransAsia, 復航, 复兴航空, 复航。

每一個關連式資料庫都有若干欄位群組。以「熱詞」（hashtags）為例，這個群組總共包含 6 個欄位。每一欄位均有若干參數，說明熱詞和其它欄位之間的關係。例如，這個熱詞在資料庫中的編號為 1，熱詞名稱為 B22816，這個熱詞曾被一則 2015 年 2 月 4 日 12 時 24 分發出、編號 562828978835443713 的推文所註記，發文者名為 iflysims70，用戶編號為 42510577。如下表一所示：

表一：熱詞（hashtag）所構成的資料維度群組

群組名稱	欄位	參數	欄位說明
hashtags	ID	1	熱詞編號
	text	B22816	熱詞名稱
	tweet_ID	562828978835443713	含該熱詞的特定推文的編號
	created_at	2015/2/4 12:24	含該熱詞的特定推文之發文時間
	from_user_Name	iflysims70	發送含該熱詞的推文的用戶名稱
	from_user_ID	42510577	發送含該熱詞的推文的用戶編號

資料來源：DMI-CAT，整理：水火計畫團隊（2016）

上述資料群組大致上收羅了若干欄位和參數，以指向某一分析目的。上表中「欄位名稱」指向特定資料類型，一般以垂直方向分佈，參數則是以「列」的方式水平分佈。研究人員必須辨識並盤點哪些欄位為分析所需。例如，倘若研究人員想要觀察「哪些熱詞在討論中最常被發文者使用？」則至少需要熱詞編號、熱詞名稱，以及推文編號等三個欄位。但如果要觀察的是「哪些熱詞在過程中如何分佈？」，則所需要的欄位就變成熱詞編號、熱詞名稱，以及推文時間。研究人員之所以需要辨識問題，並盤點相對應的資料維度，主要原因在於，當資料維度越多，則分析越趨複雜，不利於解讀和呈現。因此研究人員必須適當減少維度，也因此在操弄資料之前，必須先決定哪些資料欄位可以回應問題，藉以降低資料維度，並簡化資料數量。

其次，如何轉換資料格式以便分析？如前所述，資料清洗重點放在選擇、過濾和合併資料，以及轉換格式。這些活動大致佔據 80%資料處理時間。因篇幅有限，將以時間格式和短網址還原為例，說明資料清洗梗概。

在 Twitter API 推文時間資料係以 14 個阿拉伯數字儲存。[6] 我們所使用的工具軟體已經貼心地為研究者先轉換過一次時間格式，包括年／月／日／時／分／秒（例如：2015／2／4　12:24:27）。然而，這個時間是英國格林威治標準時間。因空難事件發生地點在臺北，必須比對當地時間，有必要將格林威治標準時間改為臺北時間。所以研究人員進行一次轉換，將所有推文時間都加上 8 小時，變成臺北時間。

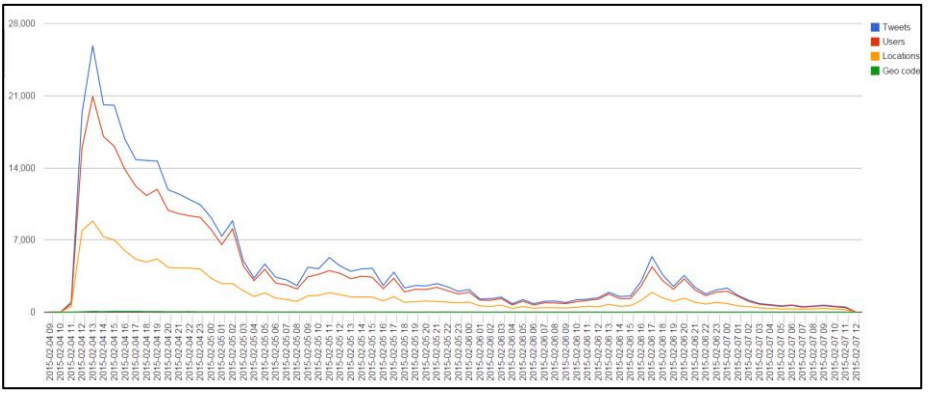
另一種轉換例子是短網址還原。Twitter 是一種微網誌（micro-blog），用戶僅有 140 字元的書寫空間，因此常以短網址（shorten URL）連結其他文本。根據本個案 URL 編號加總結果，本個案共有 81.9% 含有短網址連結。因此，用戶如何引述值得探討。但短網址無法直接辨識其網域名稱，因此必須透過短網址還原技術，將縮短的網址還原成為原本的網址格式。短網址還原基本上是一種逆向工程（reverse engineering）。其程序略為以四下個步驟：（1）先以正規表示（regular expression）逐一抽取推文中的網址；（3）再將每一則短網址以 cURL 回傳到短網址伺服器；（3）以迭代方式還原至原始真實網址（Real URL）；再以 SQL 的正規表示（regular expression）語法取出完整網域名稱（鄭宇君、施旭峰，2016）。此一網址還原結果，獲得的完整網域資料，即可用於數值分析或內容分析（鄭宇君、施旭峰，2016：）。

第三，當原始資料已經轉為初始資料，研究人員即可進行探索分析。研究人員所使用之工具軟體（如：DMI TCAT）已將資料欄位先區分為幾個關連資料，也就是預設若干類型化的資料維度，可以立即進行探索式分析。這是根據先前研究經驗，將常用的探索式分析類型收納在群組中，減少研究人員以人工匯集資料欄位的時間和精神。例如，時序分析是最常用的探索分析類型，在原始素材檔案中有一時序分析檔，收納使用者編號、

[6] 這個數字串是 1900 年以來至特定時點的秒數總和。

推文編號和發文時間等資料，可立即轉換為視覺化的時間軸圖表。如圖九所示：

圖九：探索式分析：利用 DMI TCAT 軟體界面所呈現的時序分佈



資料來源：水火計畫團隊（2016）

這圖表呈現的是一種探索式分析：時間軸的 X 軸為時間（以小時為單位），Y 軸為數量（依照顏色區分，包括推文數、用戶數、地區數，以及經緯度紀錄數等），展現數量的時間順序分佈。圖形自左向右傾斜，顯示這個事件在事發之初相當多用戶討論，經過 24 小時之後社群討論熱度即逐漸消退。研究人員觀察時間軸圖形，時間軸顯示 2 月 7 日上午 2 至 5 時之間推文量驟減，由於驟減斜度異常，可能意味著資料缺漏。因此研究人員重新啟動資料蒐集工具撈取資料，以補足原先推文資料的缺漏。

然而，並非所有類型均能滿足個別需求，此時研究人員就必須以人工匯集資料欄位。本個案屬重大空難事件，大都受到國際社群矚目，因此研究團隊希望探索「哪些語言族群參與事件討論？是否有差異？」特別是兩岸三地關注程度有何差異？欲進行此一分析，必須使用語系中的「推文編號」和「語言類型」等二欄位，並將簡體中文和繁體中文再加以區分。然而既有的資料並未建立此一資料群組，研究人員必須將此二欄位自原先的群組中抽出，重新組成一個語系分析的群組。

上述過程當中，研究人員根據資料結構，觀察資料模態，找出相對應於問題分析的資料欄位，在資料大致齊備之後，即進行初探分析，針對初探分析中的資料異常分佈，檢視資料是否缺漏，予以補強。並且將主要探討維度的資料進行匯集，形成新的資料群組。這些初步的資料整理，將有利於下一個階段的資料處理。

三、 資料分析

初步資料齊備之後，便可以進行正式的資料分析。

首先，研究團隊進行推文分析。由於先前研究指出，在選舉或社會運動中，使用者最常轉發他人推文來參與社群互動，因此轉引推文多於原創推文（original tweets）。那麼，在災難事件當中，是否依舊如此？原創推文比例會更高或更低？具體而言，在此一空難事件當中，有多少推文是原創？多少用戶轉貼加上註解？多少用戶只是純粹轉推？或是有多少用戶有多少人與某人對話？

研究人員透過數值統計來瞭解這個問題。這個問題涉及推文內容類型的數量分佈，因此會使用到「推文編號」（tweet ID）和「推文內容」（text）兩個特定欄位。其中「推文編號」內容是一串 7 位數字所組成，由推特平臺自動產生，屬於後設資料（metadata）。但「推文內容」欄位的內容，有些是由用戶自己輸入文本而產生，有的則是在用戶按轉推鍵之後，由平臺自動產生內容；屬於人產出的內容。因此，研究人員可以根據推文發送類型加以區分，再進行自動化處理。

研究人員區分發文類型為四類：（1）原創推文（original tweets）；（2）轉發推文（retweets, RT）；（3）轉發並附加個人註解（retweets with comments）；以及（4）與特定人對話（comments/reply）。根據上述四類型的操作化定義[7]，工程師撰寫程式讓電腦讀取推文內容欄位內的資料，再將判讀的結果

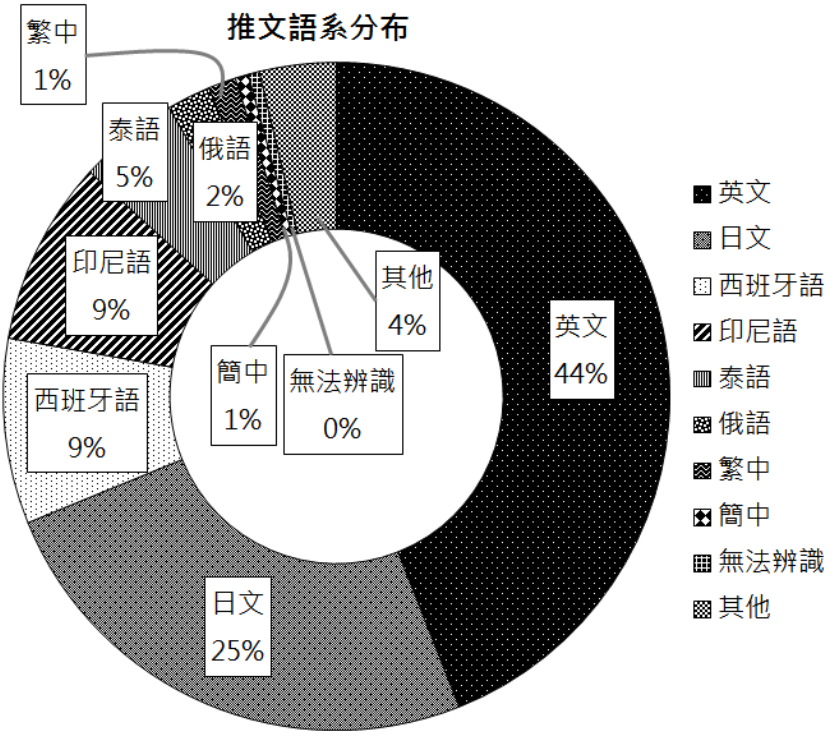
[7] 例如，原創推文的操作化定義是「用戶自己發文，不帶任何標記」。轉發推文的定義是「用戶轉發其他人推文，推文最前方為 RT」。轉發推文並加個人註解，則是「用戶轉發他人推文，並在 RT 前加註個人意見」。最後，與特定人對話，則是指「用戶發送推文與特定人對話，推文內不含 RT，且出現@user」細節參見：（鄭宇君、陳百齡，2016：132-3）。

回傳到新的欄位內，再根據各種類型的編號累積。上述分析，研究人員僅區分內容類型，後續的文本判讀，則完全委由資訊技術代勞。

其次，由於高達八成推文都附超連結網址，因此研究團隊也透過網址分析，以瞭解這個事件中的社群成員如何引述各種來源。在進行這個分析之前，先利用程式將短網址欄位內容還原為完整網域名稱。研究人員再將網域名稱加以歸為數類，包括：（1）新聞機構網站；（2）社交網絡服務；（3）影音分享平臺；以及（4）其它綜合性平臺等。根據上述分類，研究人員建立一個新的欄位「完整網域名稱」，並以「短網址編號」和「完整網域名稱」兩個欄位，計算被引述短網址類型的累加數量。此外，為了瞭解不同語系的族群引述來源是否有差異，研究團隊也將推文的語系類型加入分析，以觀察不同語系社群，在引述來源時，是否會有差異？使用「語系分類」、「短網址編號」和「完整網域名稱」等三個欄位，進行數值分析。值得注意的是，上述兩種分析雖然都屬於數值分析類型，但仍然依照內容屬性或發文屬性進行分類，並依循分類找出文本或發文類型的態樣。但以下的關連分析，則和前述兩種分析大異其趣。

最後，研究團隊希望瞭解發文內容是否可以區分為哪幾類？為找出可能的議題趨勢，研究團隊針對文本進行關連分析。研究人員選擇推文數量最多的英文推文進行分析。這個分析是以推文內容欄位為對象，研究人員先撰寫程式針對每則推文進行反向文件常見詞的統計評估，除去每則推文中的冗贅詞（如 a, the, http 等），然後以個別則推文中的單詞為節點，分析所有詞語之間的網絡關係，找出關連度最高的詞組，並呈現這些詞組之間的網絡關係，試圖呈現並詮釋詞彙之間的關係。

圖十：2015 年復航南港空難事件的語系分布圖



資料來源：水火計畫團隊（2016）

伍、討論

當代社群媒體資料分析，是研究人員根據其機構之需求，發展和評估各種資訊工具和知識結構，用以蒐集、觀測、分析、摘要，或呈現社群媒體資料所構成的一個活動。社群媒體資料分析的本質，是根據社群媒體所建構的資料，從事知識生產和再生產（knowledge production/reproduction）的歷程。本研究剖析社群媒體資料分析取徑的特性、元素和歷程。根據上述文獻和個案分析，我們獲致幾個觀察角度，可供未來研究繼續探索。以下根據本研究的陳述，說明未來研究和教學的方向。

一、 資料分析作為活動場域

這個資料分析場域，面對以人在社群媒體平臺之行動為核心的資料，不僅規模龐大，形式多元，更充滿各種雜訊。因此，社群媒體資料分析是一個活動歷程，參與的研究人員來自不同知識背景（特別是社會科學和資料科學的知識），透過彼此合作溝通完成資料分析的歷程。

誠如前述，當代社群媒體資料分析是跨域知識的活動場域，來自不同背景的研究人員，各具專長。社會科學出身的研究人員較長於將社會現實轉換為問題，以及訊息意義的詮釋。另一方面，資訊科學出身的研究人員則長於處理大量而雜亂無章的資料，使之成為可得見聞的訊息。

社群媒體資料分析這個領域正位於社會科學和資訊科學兩個學門交壤之處，對於傳播學者而言，這個領域最困難之處在於如何將問題轉換為資料處理的行動。因此，關鍵在於跨領域溝通和團隊。

二、 考量資料思維貫穿歷程的特性

社群媒體資料分析的核心是「資料思維」（thinking with data）。這個歷程是動態的（dynamic process），資料分析是從提出問題和解決問題，到尋求解答的過程，問題和資料在過程中都不斷演化，而研究人員則根據情境的變化，發展出解題策略，正如同 Scribner（1986）所描述的「一邊行動、一邊思考」（thinking in action）。這種資料分析不僅用到研究人員的心智結構，同時也體現在心智和身體的連結。因此，資料思維亦可視為心智、身體和人造物之間的連結和協力過程（connecting and collaborative process）。

三、 連結研究問題和社群資料的物質性

社群媒體研究必須掌握傳播學者所欲提問的問題，並將問題放在社群資料的情境中，研究人員尋找問題和資料之間的連結而得以發現解決問題的途徑。借用 Gibson（1979）的觀念，社群資料分析尋求的是提問（研究者主觀欲求）和情境（情境中的物質性）之間的機緣（affordances）。研

究者不僅要掌握問題，也要掌握工具軟體及資料之間的機緣特性。社群媒體資料的資料處理大都在工具軟體上進行。每一種工具軟體有其物質性（例如使用欄和列呈現資料維度、排序和數量）進而形構資料特性。因此研究人員必須善用這些物質特性的優缺點，才能針對問題找到最佳的解題策略。

對於社群資料分析的教學者而言，在情境中教學／學習是重要的。傳統研究方法教學強調方法規則和普遍性，但研究活動本質以內隱知識（tacit knowledge）為主，未必能以口語或文字說明。且資料集各有分殊性，如何在每一個個案中尋找研究問題和資料物質性之間的連結，進而在眾多解題工具中找出最佳方案，便成為訓練學生能力的重點。也因此，使用個案作為教材、讓學生「做中學習」（learning by doing），使用真實個案（real cases）引導學生思考等教學策略，有其必要性。

四、 強調探索式資料分析的重要性

兼具探索的（explorative），以及假設的（hypothetical）歷程。它具有探索的意涵；研究人員使用簡單的統計方法和視覺化工具，讓資料態樣浮現出來。它同時也是具有假設的意涵，研究人員針對高度不確定的資料，試圖觀察資料分佈態樣和可能的異狀，以找出提問重點。

正因為每一個資料集都是獨特的，卻又有某種程度的相似性，所以需要進行探索。「鷹架建構」（scaffolding）的功能，讓研究人員透過簡單的描述統計和視覺化圖表，將資料態樣再現於研究人員眼前，讓研究人員對資料開始有一定的認識，從原本生疏逐漸轉為熟稔，以利於尋找問題的解決之道。

然而，傳統的統計（無論教學或研究）通常著重在驗證和建立模型，較少聚焦於探索式分析。但是社群鉅量資料需要先行觀察資料態樣，也因此探索是社群資料分析非常重要的歷程。未來探索式分析在教育訓練中的份量，值得我們加以關注。

陸、結語

當代社會中的社群媒體，在短時間內所產生大規模、多樣化，並且高密度的資料，類似於當代社會立即反應的「反射系統」。相對而言，社群媒體的資料分析，由研究人員行資料蒐集、過濾、分析，經由精密卻緩慢的程序，企圖再現社會真實，則類似於「反思系統」。

但在現階段，社群媒體的發展迅速，而資料分析則剛剛起步，因此「快思」和「慢想」之間有如龜兔賽跑，存在重大落差。目前社群鉅量資料分析社群依舊在起步階段，社群媒體的反思系統，尚未能夠追及反射系統。

社群媒體資料分析是一個新興的知識領域。這個領域還有待學術研究社群注入活水。本文作為初探性研究，僅能描繪粗糙的輪廓。又囿於期刊篇幅限制，僅能擇要表達；內容難免掛一漏萬。但對於入門者而言，或具有參考價值；若能引發更多對話和評論，則應已達本文之初衷。

參考書目

江奕瑄、林翠絹（2015/09）。〈採用大數據探討媒體使用之學術期刊文獻分析〉，《「大數據、新媒體、使用者」研討會論文集》。臺北：風雲論壇，頁 355-368。

李瑞娟（2016）。《社交中介的政治危機傳播：探討候選人形象修復策略與媒體、公眾之互動關係》。國立政治大學新聞研究所碩士論文。

李彪（2011）。《輿情山雨欲來：網絡熱點事件傳播的空間及夠和時間結構》。北京：人民日報出版社。

林俊宏譯（2013）。《大數據：數位革命之後，資料革命登場：巨量資料掀起生活、工作和思考方式的全面革新》。臺北：天下文化。（原書 Mayer-Schönberger, V. ＆ Cukier, K. [2013]. *Big data : A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt）

林俊宏譯（2016）。《我們是誰？大數據下的人類行為觀察》。臺北：馬可孛羅 。（原書 Rudder, C. [2014]. *Dataclysm：Who we are? When we think no one is looking*. New York: Broadway Books）

許小可、胡海波、張倫、王成軍（2015）。《社交網絡上的計算傳播學》。北京：高等教育出版社。

陳百齡、鄭宇君（2014）。〈從流通到聚合：重大災難期間浮現的資訊頻道〉，《新聞學研究》，121：89-125。

楊立偉、邵功新（2016）。《社群大數據：網路口碑及輿情分析》。臺北：前程文化。

鄭宇君（2014）。〈向運算轉：新媒體研究與資科技術結合的契機與挑戰〉，《傳播研究與實踐》，7：45-61。

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》，120：121-165。

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》，29：107-133。

Ambler, T. (2011). Social media analytics. *International Journal of Advertising*, 30, (5), 918-919.

Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society*, 3, (2), 1-12.

Boumans, J., & Trilling, D. (2016). Taking stock of toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, 4, (1), 8-23.

boyd, d. m., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15, (5), 662–679.

boyd, d. m., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13, 210-230.

Bruns, A., Highfield, T., & Burgess, J. (2013). The Arab spring and social media audiences: English and Arabic twitter users and their networks. *American Behavioral Scientist*, 57, (7), 871-898.

Burgess, J.,Bruns, A. & Hjorth, L. (2013). Emerging methods for digital media research: An introduction. *Journal of Broadcasting & Electronic Media*, 57, (1), 1-3.

Burgess, J., & Bruns, A. (2012). (Not) the Twitter election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology. *Journalism Practice*, 6, (3), 384-402.

Brügger, N. & Finnemann, N. (2013). The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media*, 57, (1), 66-80.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, 5, (11), online14118.

Felt, M. (2016). Social media and the social sciences: How researchers employ big data analytics, *Big Data & Society*, 3, (1), 1-15.

Fishman, D. A. (1999). ValuJet flight 592: Crisis communication theory blended and extended. *Communication Quarterly*, 47, (4), 345-375.

Forte, A. & Lampe, C. (2013). Defining, understanding, and supporting open collaboration: Lessons from the literature. *American Behavioral Scientist*, 57, (5), 535-547.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication*, 64, 239-259.

Kahneman, D. (2012). *Thinking: Fast and slow*. New York: Penguin Books.

Kitchin, R. & Lauriault, T. P. (Forthcoming). Towards critical data studies: Charting and unpacking data assemblages and their work. *The Programmable City Working Paper 2*; pre-print version of chapter to be published in Eckert, J., Shears, A. and Thatcher, J. (eds.) *Geoweb and Big Data*. University of Nebraska Press. Retrieved September 6, 2016, from http://ssrn.com/abstract=2474112.

Kogan, M., Palen, L., & Anderson, K. M. (2015). Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. Paper presented at the Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work, Social Computing, Vancouver, BC, Canada.

Mahrt, M. & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, 57, (1), 20-33.

Manovich, L.(2011). Trending: The promises and the challenges of big social data, in Gold, M. K. (Ed.). *Debates in the Digital Humanities*. Minneapolis, MN: The University of Minnesota Press. Retrieved September 6, 2016, from http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11, 122-129, 1-2.

Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication*, *64* : 355-360.

Rutkin, A. (2014). Twitter bots grow up. *New Scientist, 223,* 2980, 20-21. Retrieved on August 18, 2016, from https://www.newscientist.com/article/mg22329804-000-twitter-bots-grow-up-and-take-on-the-world/

Seltman, H. (2015). Exploratory data analysis. *Experimental Design and Analysis*, 61-98. y, Pittsburgh, PA: Carnegie Mellon University.

Smith, A., Molinaro, M., Lee, A., & Alberto, G. (2014). Thinking with data. *The Science Teacher,* 81, (8), 58-63.

Spasojevic, Li, Rao & Bhattacharyya (2015). When-to-post on social networks, *Proceedings of the 21th International Conference on Knowledge Discovery and Data Mining*, 2127-2136, Sydney, NSW, Australia, August 10-13, 2015 in ACM Digital Library. Retrieved September 6, 2016, from http://dl.acm.org/citation.cfm?id=2788584

Stieglitz, S. & Dang-Xuan, L. (2012). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining*, 3, (4), 1277-1291.

Stieglitz, S., Deng-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, 6, (2), 89-96.

Tinati, R., Halford, S., Carr, L. & Pope, C. (2014). Big data: Methodological challenges and approaches for sociological analysis. *Sociology*, 48, (4), 663-681.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media,* Ann Arbor, Michigan, USA, June 1-4, 2014.

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. New York: Oxford University Press.

Velleman, P. & Hoaglin, D. (1981). *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis*. Boston, Mass.: Duxbury.

Zeng, D., Chen, H., Lusch, R. & Li, S. (2010). Social media and intelligence. *IEEE Intelligent Systems*, 25, (6), 13-16.

Exploring social media analytics: Characteristics and processes

Pai-lin Chen & Yu-Chung Cheng & Kung Chen

Abstract

Social media analytics (SMA) refers to the methods aiming to collecting, cleaning and analyzing data from social media platform, in order to conducting social listening. This approach of research requires inter-disciplinary team work, particular data science, statistics and other fields. Although communication scholars foresee its potential in future studies, very few articles discuss the elements and process of this research trend. This article aims to present a case study, describing the elements and the process of social media analytics, and its implications in future communication studies.

Keywords: social media, analytics, social big data, digital methods, inter-disciplinary research