災害防救科技與知識專欄 —

# 社群網路資料攀爬過濾與應用展示

劉致灝、蔣佳峰、張子瑩

國家災害防救科技中心 資訊組

## 摘要

資訊科技演進帶動新興的資訊傳遞模式，在人手一機的低頭族大量產生與社群網路的蓬勃發展之下，人們在網路世界逗留的時間也與日俱增，在臉書、LINE、IG 等社群網站上分享個人生活的行為更是常見，使用者於社群網站頻繁且大量的交換各類資訊，若在重大事件發生之時，多數人的話題開始隨之轉向，相關資訊將很快地被散佈在社群網路之中，根據事件內容的變化，討論的關鍵字詞也有所不同，例如：選舉相關文章常見的「提名」、「初選」、「候選人」等，若應用於災害發生期間，「捐款」、「救援」、「災情」等字詞的出現頻率較高。

本文將說明自國家災害防救科技中心(以下簡稱災防科技中心)建置的社群攀爬系統，介紹從社群網站蒐整災害相關文章後，透過關鍵字詞集作為基礎過濾災情相關文章，最後也以尼莎、海棠颱風作為範例，呈現社群資料分析及應用成果。

## 一、簡介

網際網路的發展加速資訊交換的效率，隨著時代的演進，人們從早期使用桌上型電腦，以靜態方式瀏覽網頁資訊或是使用者透過討論區與論壇等社群網路(Social Web)交流平台，進行互動。

由於硬體設備的縮小化，使用者從笨重的桌上型電腦，轉變成隨身帶著走的筆記型電腦，最後變成輕薄可放入口袋的智慧型手機，社群網路的演進與方便攜帶的硬體設備相輔相成之下，交換訊息的頻率越發迅速。

社群網路已經成為網路使用者最常使用的服務，根據財團法人台灣網路資訊中心調查顯示(http://technews.tw/2014/08/20/twnic-online-behavior-survay-more-mobile-sulfing/)，台灣網路使用者每天花 3.25 小時上網，其中使用社群服務的佔總時間的 64.32%，成為上網瀏覽項目的第一名，由此可知社群網路的熱門度。目前廣泛使用的社群網站如:臉書(Facebook, FB)、推特(twitter)、批踢踢(PTT)等。使用者發布資訊分享給好友或其他使用者瀏覽，一般而言，任何使用者都能對於資訊提供回饋與心得，達到資訊流通的互動性，使用者也能加入社群服務的社團，定期獲取特定主題的資訊。

從網路科技服務的發展歷程反應出使用者對於科技使用的需求，同時也改變使用者以往在資訊傳遞的方式。現今網路服務的使用者習慣將網路上瀏覽到的資訊，藉由社群網路服務平台發佈或轉貼，如旅遊資訊、美食情報或政治議題等。這些資訊會被其他社交圈或是網民瀏覽，並且不斷的被轉載或觀看，使得資訊的傳遞速度變得非常的快速。同樣地，當災害發生期間，民眾間會將最新災情的訊息傳遞於各大社群交流平台之間，防災人員若能善加利用這股能量，對於災情現況的瞭解，將有助於防災人員掌握現場的情勢。

因此，災防科技中心建立一套以社群網站為主要資料來源的攀爬系統，於災害應變期間蒐集網路上的留言與文章，透過基本、快速的過濾策略來找出與災情相關文章。

## 二、社群攀爬平台

在災防科技中心建置的社群攀爬平台中，先以資料攀爬程式(Web Hunter)擷取各大社群網站、新聞等資料來源，圖 1 為資料查詢介面，可根據時間選擇區間內發布之文章集，來源列表可選擇資料來源的種類，主題列表則是透過事先設定的正規表達式(Regular Expression, RE)初步篩選文章，大量減少非相關文章之數量，以颱風為例，當我們遇到颱風侵襲時，容易聯想到的關鍵字常為水災、淹水、豪雨等字眼，視當時災情狀況而決定適用之災害關鍵字集合，配合使用 AND(&)、OR(|)、NOT(!)等邏輯運算元組合，達到初步篩選。


圖 1 攀爬平台查詢介面

| | | |
|---|---|---|
| 地震應變-房屋 | | 地震&(((?樓\|?廈\|房屋)&(倒塌\|傾斜\|下陷))\|斷層) |
| 道路災情 | | 道路\|中斷\|坍塌\|崩塌\|落石\|封閉 |
| 活動停辦 | | 改期\|停辦\|暫停\|延期\|停止\|順延\|暫緩\|另行通知\|取消\|延後 |

圖 2 關鍵字詞類過濾範例

然後根據災害類型，設計適用的關鍵字列表，以過濾不相關文章，減少額外的分析時間、成本。以圖 2 的地震災害為例，當發生地震且有房屋受到影響時，我們將選用「地震應變-房屋」的關鍵詞組進行過濾，當文章出現「地震」且描述中具備「房屋倒塌」、「斷層」等相關字詞時，初步判定為與本次災害有相關的文章，後續再將照片、地點敘述等內容進行深入判斷。這些關鍵字的選用，主要經社群留言、報章雜誌、專家學者等提及的內容做為參考，細節將會在第三章進行說明。

除一般性的資料閱覽外，在「網路輿情觀測」功能模組中，可提供關鍵字的設定，供災防科技中心應變人員選擇指定條件後(如圖 3 所示)，以自動化方式定期更新災害相關社群輿情呈現於介面。


圖 3 網路輿情觀測平台之查詢條件

「社群資料攀爬平台」可指定來源頻道，包含 Facebook 社團/活動、Facebook Hashtag 及 PTT 八卦版等 3 個頻道內的內容作為觀測標的，以每 5 分鐘自動更新，呈現最新之討論串(討論串包含主文及其回文)，依回文最新時間或主文最新時間做為討論串排序依據。

**三、 關鍵字選取**

在一篇文章中，決定關鍵字詞的策略主要分為兩個：1. 利用統計分析計算字詞頻率的初步分析，藉此快速找出熱門關鍵字詞，此作法較為快速，但無法指定特定領域字詞，2.命名實體辨識(Named Entity Recognition, NER)，透過觀察文章特徵的機器學習，將文章結構抽象化，從中擷取特定片段，搜尋特定領域之關鍵字詞，這種作法較為費力，需標記大量訓練資料、設計合適特徵。

人類能閱讀文章並找出關鍵字句，主要為能理解文章中各單詞的意義(Meaning)，對於程式來說，文章中的字詞本身不具有任何意義性，只是由不同的字元組合的句子，這些詞義難以將其數據化，故無法被程式所「理解」。

在這前提下，簡單的做法是：無視每個字詞的意義，所謂熱門的關鍵字詞，照常理推論應是當下被熱烈討論、提及的字眼，所反映出的應該是異常高的曝光率，可以僅計算各詞類自身的出現在單一文章的詞類頻率(Term Frequency)，表示該詞類被提及的重要程

度；計算該字詞在各個文章出現的文本頻率(Document Frequency)，表示該字詞的應用廣泛程度，透過統計分析的方式，找出具代表性的關鍵詞類。

以統計分析的方式，作為直觀的權重計算方法共有三個：(1)計算特定詞類的總出現頻率(Term Frequency, $T_f$) (2)計算特定詞類出現於每一篇文章頻率(Document Frequency, $D_f$) (3)結合前述兩種計算方法，出現頻率與文章頻率倒數之乘積(TF-IDF)。

根據以上統計方法，配合報章雜誌、社群輿論留言以及專業領域會提及之災情相關內容，將這些文章進行統計分析的測試後，所得的關鍵字列表與災防領域相關的專家們討論後，依災害類型，建立災害關鍵字詞列表(如表 1 所示)，以應用於災害應變期間蒐整社群資料使用。

表 1 災害類關鍵字列表

| 災害類關鍵詞組 | 內容 |
|---|---|
| 颱風(17) | 颱風、淹水、土石流、溪水、停電、暴漲、路樹、強風、積水、暴漲、溢堤、淹水、水位、豪雨、暴雨、災情、回報 |
| 水災(10) | 淹水、溪水、暴漲、積水、暴漲、溢堤、淹水、水位、豪雨、暴雨 |
| 風災(8) | 樹倒、路樹、傾斜、招牌、掉落、倒塌、吹翻、鷹架 |

| | |
|---|---|
| 土石流(5) | 土石流、山崩、崩塌、滑落、落石 |
| 道路災情(8) | 道路、中斷、坍塌、崩塌、落石、封閉、橋梁、路基掏空、 |
| 地震(12) | 地震、樓、廈、房屋、倒塌、傾斜、下陷、斷層、土壤液化、裂、災情、回報 |

## 四、尼莎、海棠颱風實際案例分析

在本研究整理社群資料的流程可大致分為四個步驟：攀爬、過濾、定位與製圖(如圖 4 所示)，自社群網站的熱門頻道大量蒐集民眾留言、文章。這些五花八門的內容將藉由過濾機制進行刪減，通常以災情常被提及或常見的關鍵字詞作為主要篩選對象，搭配前述的正規表達式進行初步的過濾，後以判斷資料的正確性，以人工驗證、感測儀器數值比對、CCTV 的照片佐證等方式，確認資料的實際地點，最後，這些足以被稱為資訊的內容，將展示在電子地圖上呈現其時序性、區域分布。

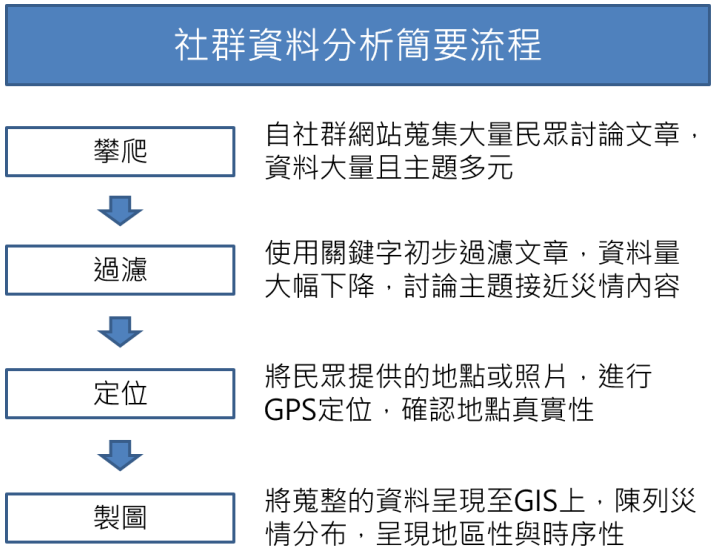


圖 4 社群資料分析的簡要流程

本研究使用的資料為攀爬平台擷取之社群攀爬資料，以尼莎、海棠颱風事件為例，詳細資料設定說明如下所示：

1. 時間：西元 2017 年 7 月 29 日至 7 月 31 日

2. 資料來源：PTT(八卦板、颱風板、各地方板等)、臉書(新聞媒體粉絲團、在地社團、熱門社團等)、噗浪搜尋(Plurk)為主要資料來源

3. 關鍵字詞組設定：

本研究使用災害常見關鍵詞組，依照該次颱風登陸情境與災害發展，依序選擇對應模式，本次計使用颱風、水災、風災、土石流、道路等災害類別關鍵詞組，詳如表 2 所示。

表 2 颱風相關關鍵字列表

| 災害類關鍵詞組 | 內容 |
|---|---|
| 颱風 | 颱風、淹水、土石流、溪水、停電、暴漲、路樹、強風、積水、暴漲、溢堤、淹水、水位、豪雨、暴雨、災情、回報 |
| 水災 | 淹水、溪水、暴漲、積水、暴漲、溢堤、淹水、水位、豪雨、暴雨 |
| 風災 | 樹倒、路樹、傾斜、招牌、掉落、倒塌、吹翻、鷹架 |
| 土石流 | 土石流、山崩、崩塌、滑落、落石 |
| 道路災情 | 道路、中斷、坍塌、崩塌、落石、封閉、橋梁、路基掏空、 |

時間取自發布尼莎颱風之陸上颱風警報開始，至解除海棠颱風之海上颱風警報為區間，共計抓取 67,352 筆文章，其中，前三名資料佔比為臉書粉絲團、PTT、Plurk 搜尋，詳細如圖 5 所示，臉書粉絲團主要透過記者拍攝災情照片後由社團的小編上傳照片、發文，

民眾瀏覽照片後再進行留言；PTT 與 Plurk 搜尋皆為一般民眾之討論內容，品質與內容則相對較差。

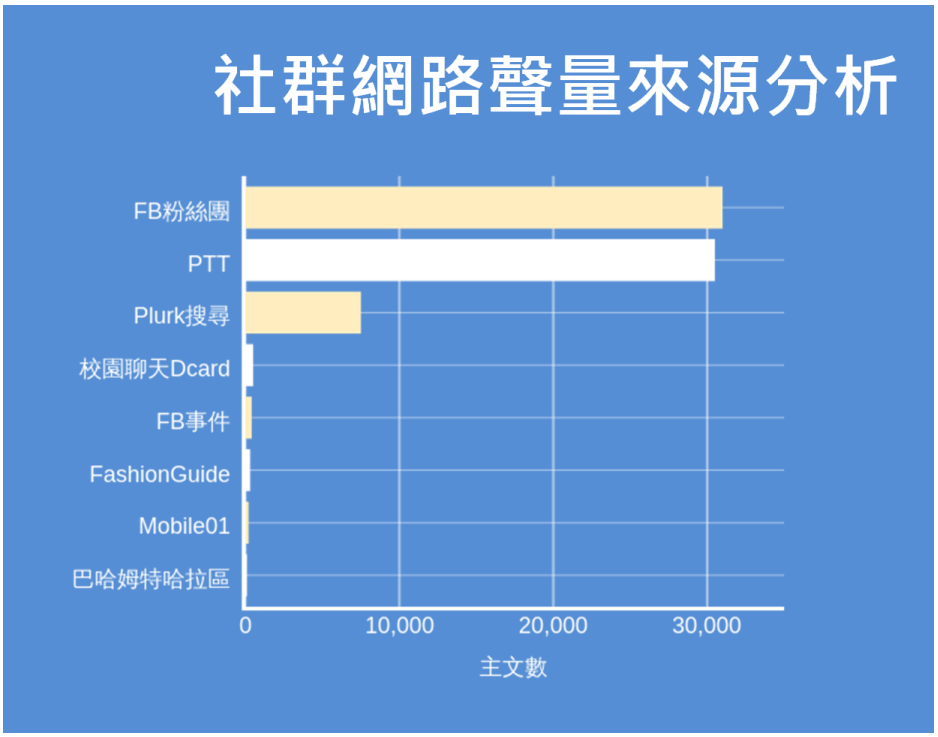


圖 5 社群網路聲量分布

聲量的趨勢追蹤部分則可以對照災防科技中心之「2017 尼莎暨海棠颱風災害報告」之時間點進行比較，民眾討論聲量分別在整起事件的開始與結束為最高討論熱度，分別為 7 月 29 日 14 時發布尼莎颱風海上颱風警報，與 7 月 31 日 8 時解除海棠颱風海上警報為兩個討論熱點，如圖 6 所示。

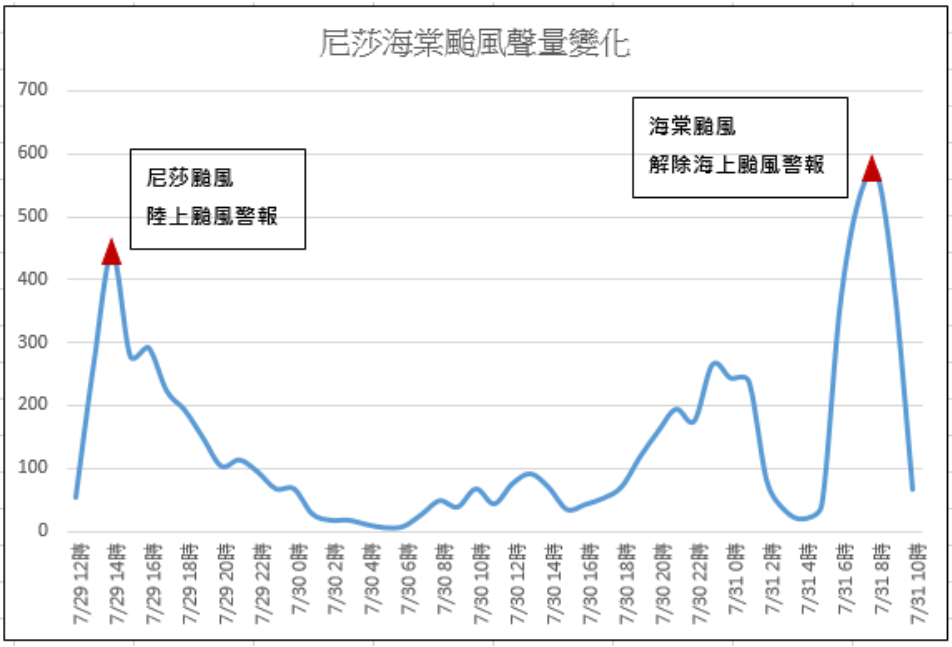


圖 6 尼莎、海棠颱風社群聲量變化趨勢

本研究將災情攀爬的災情透過關鍵字初步過濾後，經過人為判定，具備 1.詳細地址敘述、2.現場災情照片，兩者條件皆滿足的留言始成為認定之「重點災情」，實際的大量災情文章，能夠被真正的當作資訊者實屬稀少，由表 3 可知，在本次雙颱侵襲期間，尼莎颱風與海棠颱風的社群災情資訊，僅佔全文章數量的 0.1~0.15%左右，但已足夠提供一系列的災情展示追蹤。以圖 7 與圖 8 為此次屏東地區的社群災情資訊展示，我們以 GIS 呈現災情的空間分布，並能以 CCTV 輔助驗證資料的正確性，這些資料也作為當時防災人員進行分析決策調度時之參考。

表 3 尼莎、海棠颱風社群資料實際使用情形

| 颱風事件 | 災情攀爬 | 重點災情 | 情資共享 |
|---|---|---|---|
| 尼莎颱風 | 67,352 | 73 | 29 |
| 海棠颱風 | 32,976 | 51 | 33 |


圖 7 社群網路災情總整地圖


圖 8 CCTV 輔助佐證社群資料真實性

## 五、結論

社群網站在災害發生期間，能提供大量的災情資訊，然而，這些「資料」需經過先期的處理才能成為有用的「資訊」。本研究提出自社群網路應用於災害情勢分析，以關鍵字詞分析作為文章過濾機制，大量減少無關的內容，因此關鍵字詞的設定，有助節省人為驗證資料的時間。

在 2017 年尼莎、海棠颱風期間，顯示出社群網路於災情資訊能提供有效災害情資，從此案例可觀察出，社群網路可作為一個有效的資訊蒐集來源，分析文章並透過關鍵字擷取對於災害事件追蹤有一定的價值，透過社群網路來源及社群聲量分析，可觀察熱門社群網站與當前民眾討論聲量趨勢，這些都是在颱風應變期間，對於防救災單位在掌握災情現況上，提供一項新興的參考資訊。

## 六、參考文獻


1. 社群網站的使用行為：創市際調查報告，available at: https://rocket.cafe/talks/78006
2. 臺灣寬頻網路使用調查，available at: http://technews.tw/2014/08/20/twnic-online-behavior-survay-more-mobile-sulfing/
3. Andreas M. Kaplan, Michael Haenlein. (2010) Users of the World, Unite! The Challenges and Opportunities of Social Media. Volume 53, Issue 1, January-February 2010, pp. 59-68.
4. Chih-Hao Liu and Jeng-Fen Bau. (2016) Applied Social Media in Disaster Situation Response, IEEE International Conference on Applied System Innovation, May 28-June 1, Okinawa, Japan.
5. Chou, Chien-Lung, Chia-Hui Chang, and Ya-Yun Huang. "Boosted Web Named Entity Recognition via Tri-Training." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16.2 (2016): 10
6. EMC. The Digital Universe of Opportunities. Available at: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf
7. Haewoon Kwak, Changhyun Lee, Hosung Park, Sue Moon. (2010) What is Twitter, a Social Network or a News Media?. Proceeding of the 19th international conference on World Wide web, New York, USA, pp.591-600.
8. Kryvasheyeu, Yury, et al. "Rapid assessment of disaster damage using social media activity." *Science advances* 2.3 (2016): e1500779.
9. Opview社群口碑資料庫, available at: http://www.opview.com.tw/socialDB.html
10. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.* Paper presented at the Proceedings of the 19th international conference on World wide web.
11. Pantti, M and Wahl-Jorgensen, K and Cottle, S. (2012) Disasters and the Media. Peter Lang, New York, pp. 248.
12. Zheng Xiang, Ulrike Gretzel. (2010) Role of Social Media in Online Travel Information Search. Volume 31, Issue 2, April, pp.179-188.