OAC-NKUST-109-18（研究報告）

以文字探勘技術分析海洋政策在社群平台推廣成效

（成果報告）

海洋委員會補助研究

中華民國 109 年 10 月

「本研究報告僅供海洋委員會施政參考，並不代表該會政策，該會保留採用與否之權利。」

OAC-NKUST-109-18（研究報告）

以文字探勘技術分析海洋政策在社群平台推廣成效

（成果報告）

學校：國立高雄科技大學

指導教授：李建邦

學生：陳韋翔

研究期程：中華民國109年5月至109年12月

研究經費：新臺幣伍萬元

海洋委員會補助研究

中華民國 109 年 10 月

「本研究報告僅供海洋委員會施政參考，並不代表該會政策，該會保留採用與否之權利。」

目 次

表 次

圖 次

摘 要

雖然目前已經有很多文獻針對社群行銷進行研究，探討如何藉由操作社群媒體上的資源，達到有效行銷目標之策略，但卻較少有針對政府單位於社群平台的政策宣導成效進行分析。為了瞭解與探討政府單位於社群平台政策的宣導成效，本研究將以海洋委員會與海洋委員會海巡署的官方社群平台為例進行深入研析。在研究中期望透過文字探勘及語意分析的技術，針對海洋委員會與海洋委員會海巡署的官方社群平台所發表之文章內容進行分析。

由於目前 Facebook 因資安政策所以已無法直接透過 Facebook Graph API 進行資料串接獲得相關訊息，因此，本計畫改透過網路爬文的技術針對海洋委員會官方 Facebook 蒐集資料，如：貼文內容、貼文時間、回覆內容以及按讚數量等訊息。本計畫已完成爬文程式的開發，進入長期資料蒐集的階段，並依目前所蒐集到的資料進行語意分析，探討貼文內容之語意議題，並以文字雲與語意網路進行呈現。

由於目前計畫採用網路爬文技術蒐集資料，因此，後續仍需持續長期蒐集爬文內容，進而建立資料庫，最後再進行對應語意分析與貼文廣度與深度之關聯，瞭解究竟何種貼文類型或貼文時間點較易受到民眾的喜愛，並可做為未來海洋委員會與海洋委員會海巡署進行滾動式修正於社群平台的文章特性之依據。

關鍵詞：文字探勘、社群分析、語意分析、海洋委員會、海洋委員會海巡署

第一章　前言

1.1　研究緣起

隨著科技日新月異以及智慧型手機普及，加上近年來網路社群媒體的興起，為強化與民眾的互動，政府單位除了既有的網站以外，更逐漸借重各種社群平台，除了可藉由社群平台巨大的網路流量來發布訊息，增加訊息快速傳遞的機會之外，更期望能夠在第一時間點與民眾互動，因此，政府各單位的政策行銷與宣導導入社群平台成為目前極為重要的手段。

1.2　問題背景與現況分析

近年來，隨著科技進步快速，現在網路技術成熟、高科技產品問世、跨越空間限制，智慧型手機進入民眾的日常生活中，許多人在家中就能掌握古今中外的各種訊息，資訊的傳播速度越來越重要，加上資訊傳播的管道已從傳統管道，例如電視、報紙等，已經演變成網路社群媒體。因此，社群平台儼然已經成為一個重要的網路世界，每個人都能在社群平台上瀏覽來自四面八方的知識與訊息，並發表字自己的言論，與相同興趣的人們一起討論時事。目前 Facebook 臉書是世界上最大的社群網路平台之一，因此，用戶數據非常有價值，世界各國的政府單位同樣也不能錯過配合網路行銷來使政策宣傳達到更高的成效。

1.3　研究目的

地理環境四面環海的臺灣，是個標準的海島國家，而人民應要對海洋有基本的認識與親近，但是在我國既有的國民教育中較少有特別針對海洋相關議題的探討，直到西元 2018 年 4 月海洋委員會正式成立，臺灣才有統籌海洋相關議題之最高行政單位。因此，為了瞭解海洋委員會近年政策宣導的成效，本研究以海洋委員會以及海洋委員會海巡署的官方 Facebook 粉絲專頁作為蒐集資料平台。期望能透過文

字探勘與語意分析的技術來分析海洋委員會與海洋委員會海巡署的政策宣導成效。

本計畫主要研究目的為：

(1)透過本研究所定義之文章發表的廣度與深度，瞭解民眾對於政府單位在社群平台上發布政策宣導的觸及率(廣度)及意見回覆(深度)；

(2)進行文章貼文內容研析，包括貼文時間點、關鍵字選用等議題，使海洋委員會與海洋委員會海巡署的官方 Facebook 粉絲專頁能發揮更大的效益。

1.4 研究重點及預期目標

期望透過本研究成果所建立之語意分析排行榜，瞭解究竟何種貼文類型或貼文時間點較易受到民眾的喜愛，以及每則文章的留言數與按讚數量之變化，建立之廣度與深度的排行，因此，必須定期爬文獲取最新的資料並進行不同時間點資料變化之差異比較，並可做為未來海洋委員會與海洋委員會海巡署進行滾動式修正於社群平台的文章特性之依據。

第二章　研究方法與過程

本計畫原先規劃直接透過 Facebook Graph API 連接海洋委員會及海洋委員會海巡署官方粉絲專頁的資料，但因為資安的問題導致 Facebook 已關閉 Facebook Graph API 直接串接公開的粉絲專頁的資料，為了達到原先計畫的目的，本研究改採用網路爬文的技術進行資料蒐集，再將所蒐集的資料透過文字探勘的技術進行語意分析。為了說明本計畫所執行的研究方法與過程，在本章第一節與第二節將分別簡介計畫中主要應用的技術—網路爬文以及文字探勘與語意分析，而在第三節針對計畫的執行流程進行說明。

此外，因為在 2020 年 9 月時，Facebook 更新大改版導致爬文的程式碼要重新撰寫，並改用 Facebook 官方提供的純 HTML 版本 Facebook 來重新進行爬文，以獲得海洋委員會及海洋委員會海巡署官方粉絲專頁的貼文相關資料。

2.1　網路爬文

網路爬文或稱為網路爬蟲(Web Crawler)，可以理解為在網路上爬行的蜘蛛，網路就像是一張大網，而爬蟲就是在這張蜘蛛網上爬行的蜘蛛，遇到資料就會依程式撰寫的邏輯將相關的內容抓取並儲存於電腦之中。在本計畫中主要透過 Selenium 網頁測試工具，本工具可以直接以程式碼操控瀏覽器之特性，使其成為網路爬蟲必備的工具之一，在執行的過程中，透過解析原先網頁 HTML 的相關元素或標籤進而獲得 HTML 所對應的文字內容。本計畫是透過 Python 進行網路爬文的程式撰寫，並將所蒐集的資料儲存於本計畫所建置的資料庫，以進行後續文字探勘與語意分析的探討。

2.2　文字探勘與語意分析

隨著文字探勘(Text Mining)的相關技術的發展，可將既有文字進行拆解並重新

組合，以挖掘未知或潛在的知識(Delen & Crossland, 2008)，目前文字探勘主要是透過將文本中關鍵字的重要性與關聯性進行計算並作為判斷與探勘潛在知識的依據。

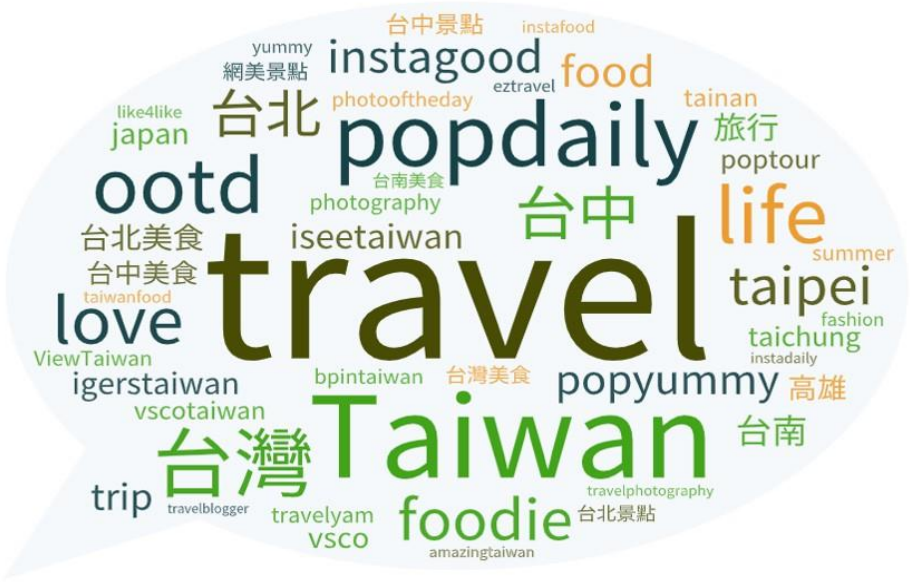

圖 1 文字雲範例：2019 Instagram 熱門#Hashta 旅遊主題分析

其中關鍵字的重要性是指關鍵字出現的頻率(Frequency)，若出現的頻率越高可反映出多數受訪者真正關心的部分，並可透過關鍵字的頻率繪製文字雲(Word Cloud)，如圖 1，透過視覺化文字雲的方式來了解在文本中重要的關鍵字。

當完成重要關鍵字的判斷後，接著可透過解析關鍵字之間的關聯性，如使用關聯規則(Association Rule)將經常共同出現的詞彙建立關聯性，進而建立語意網路以判斷其語意的真實意義。因此，進行文字探勘相關的研究，第一步驟幾乎都是先進行文字斷詞，將原本的文句斷成數個關鍵詞彙，進而探討詞會出現的頻率與關聯性，其基本概念與步驟如圖 2 所示。

在臺灣目前常見的中文斷詞系統有中央研究院所開發的「中文斷詞系統(Chinese Knowledge Information Processing, CKIP)」以及「結巴(Jieba)」兩種軟體(套件)，其中 CKIP 在西元 2019 年已經重新開放 API(Application Programming Interface)，

目前可透過 Python 語法直接進行串接，而結巴則可直接在 R 或 Python 兩種常見的資料分析軟體下載套件使用。這些中文斷詞系統可以有效提升處理中文文字分析與探索的效率。文字探勘的應用方式相當多元，其中又以「自然語言處理(Natural Language Processing, NLP)」是較接近人類自然語言的分析方式(陳怡廷, 陳麗如, & 吳姿瑩, 2016)。在 1999 年，以奇異值分解(Singular Value Decomposition, SVD)為基礎的潛在語意分析法(Latent Semantic Analysis, LSA)被發展出來，並應用於探討隱藏在關鍵字之間背後的關係(Landauer, Laham, & Foltz, 1999)。目前潛在語意分析法已廣泛被應用於各種議題，如社群網路分析(管瓊瑛 等人, 2017)，教育相關議題(Hutchison, Daigle, & George, 2018; 郭伯臣, 廖晨惠, & 張正杰, 2018)，以及查詢系統之應用(陳林志 & 陳冠瑜, 2015)。

圖 2 中文斷詞系統處理步驟

2.3 研究過程與步驟

本節將針對本研究計畫之內容與步驟進行詳細的說明，本研究步驟共分為七大步驟，並將七大步驟分於 2.3 節中的七小節進行說明，分別為第一小節：資料蒐集目標，在第一小節中主要確認在本次研究所要探討的海洋委員會與海洋委員會海巡署的官方 Facebook 之粉絲專頁的網址；第二小節：網路爬文，主要利用

Selenium 套件開啟瀏覽器，進入 Facebook 粉絲專頁後，了解 HTML 元素抓取網址內容；第三小節：資料儲存規劃，主要蒐集的資料為，貼文的內容、貼文的時間、貼文的按讚數量、貼文的回文數量，以及蒐集爬文的時間。因要了解貼文的變化量去做長期分析，所以需建置資料庫，以進行長期爬文並有效蒐集資料；第四小節：分析貼文廣度及深度，由本研究定義廣度與深度兩種衡量方式，來作為探討政府在政策宣導成效的依據；第五小節：語意分析，依爬文所蒐集之貼文內容進行彙整後，再利用斷詞系統做語意網路的分析；第六小節：對應語意分析與貼文廣度及深度之關聯，將第四小節與第五小節做結合比對，建立最佳廣度與深度的語意分析排行；第七小節：結論與建議，詳細內容如下所述。

2.3.1 資料蒐集目標

本計畫資料的蒐集目標為海洋委員會以及海洋委員會海巡署的官方 Facebook 粉絲專頁，其中海洋委員會是臺灣管理海洋相關資源的最高行政單位，於西元 2018 年 4 月 28 日正式成立。而海洋委員會官方 Facebook 粉絲專頁網址為 https://www.facebook.com/oactw/ 以及純 HTML 版本的網址為 https://mbasic.facebook.com/oactw/；海洋委員會海巡署則為負責臺灣海域及海岸巡防之最高主管機關，於西元 2000 年 1 月 28 日成立，原為國防部海岸巡防司令部、內政部警政署水上警察局等任務執行機關。在海洋委員會正式成立後，更名為海洋委員會海巡署，並於西元 2019 年 2 月 19 日開始經營 Facebook 粉絲專頁，網址為 https://www.facebook.com/CGA4U/以及純 HTML 版本的網址為 https://mbasic.facebook.com/CGA4U/。

2.3.2 網路爬文

爬文的過程中首先利用 Selenium 套件控制 WebDriver 開啟瀏覽器，接著進入官方粉絲專頁後，透過了解 HTML 的元素得知內文、貼文時間、按讚數

量以及留言數量的網頁元素為何，指定所需的網頁元素便能獲取內文、貼文時間、按讚數量以及留言數量等相關資訊。由於所爬文獲得的資訊，原先是屬於一長串的字串，為了方便後續儲存與分析之用，本計畫透過 Python 語法先將字串資料轉換成 List 型態，最後再轉換成 Data Frame 表格化後的資料，以便後續儲存於資料庫之中。

2.3.3 資料儲存規劃

由於本計畫所蒐集的資料主要為蒐集資料的時間、貼文的內容、貼文的時間、貼文的按讚數量以及貼文的回文數量，為了後續分析之方便性，如分析同一篇文章在不同時間點相關訊息之變化，因此將資料庫規劃成四大資料表，分別為來源資料表、貼文內容資料表、按讚時間表、回文資料表。各個資料表內之屬性設定與資料表之間的關聯如表 1 與圖 3 所示。

表 1 資料庫中資料表之規劃

資料表名稱	資料表內容
來源資料表	資料來源 ID、資料來源名稱 (S01：海洋委員會；S02：海洋委員會海巡署)
貼文內容資料表	貼文 ID、貼文內容、貼文時間、資料來源 ID
按讚資料表	按讚 ID、按讚數量、蒐集時間、貼文 ID
回文資料表	回文 ID、回文內容、蒐集時間、貼文 ID

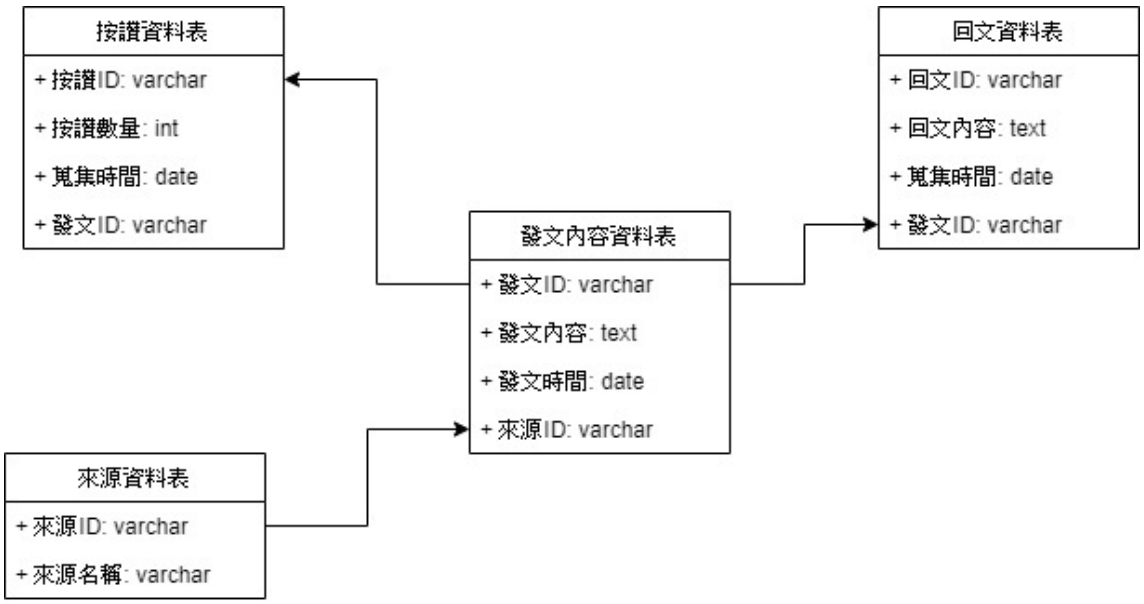

圖 3 資料表之關聯

2.3.4 分析貼文廣度及深度

本在本研究中主要定義廣度與深度兩種衡量方式，作為探討政策宣傳承效的依據，衡量方式分別介紹如下：

(a) 廣度：主要以文章的觸擊率作為衡量的依據，即依所發表的文章被按讚（Like）的數量，作為判斷文章被民眾閱讀的程度。

(b) 深度：主要以文章的回覆率作為衡量的依據，即依所發表的文章之回覆內容的數量，作為判斷文章被民眾詳細了解的程度。

針對廣度與深度之定義完成之後，則需進行每則貼文的廣度與深度，由於若直接統計每則貼文的總按讚數或總回文數時，經常會受制於貼文時間之影響，因此，在本研究中，目前暫定依表 2 與表 3 之定義將貼文進行貼文瞬間以及貼文累積的貼文廣度與深度之分類，依表之定義若 Level 越高則代表其貼文的廣度與深度越強。

表 2 貼文瞬間的廣度與深度之等級分類

時間 定義	Level 1 30分鐘內	Level 2 30分鐘	Level 3 30分鐘	Level 4 30分鐘	Level 5 30分鐘
按讚次數	100	300	600	1,000	2,000
留言數量	10	30	60	100	200

表 3 貼文累積的廣度與深度之等級分類

時間 定義	Level 1 兩天內	Level 2 兩天內	Level 3 兩天內	Level 4 兩天內	Level 5 兩天內
按讚次數	1,000	3,000	6,000	10,000	20,000
留言數量	100	300	600	1,000	2,000

根據上述之定義來說，若發表的文章被按讚數越多，代表有越多的民眾曾經接收過這則訊息，同樣地，若發表的文章被回文的數量越多，代表有越多的民眾曾經深度瞭解內容，並發表其個人看法於該文之中。

一篇貼文隨著時間推移到一定程度後，文章熱度會趨緩，按讚跟留言也會幾乎趨近於零，而這篇貼文的研究週期也將會停止，因此，本研究中亦將探討每篇文章存活的週期，並深入探討文章的存活週期是否會因為文章內容而改變。

2.3.5 語意分析

於本研究亦想瞭解貼文內容是否會影響貼文的廣度與深度，因此，將依網路爬文所蒐集之貼文內容進行分析，在研究中將採用文字探勘的技術，從非結構化的資料中，萃取出有意義的資訊，再用 Python 語法串接斷詞系統，進行語意分析。

2.3.6 對應語意分析與貼文廣度及深度之關聯

利用語意分析瞭解貼文內容後，探討出廣度與深度的關聯性，依語意分析之內容與廣度與深度的分類進行排序，建立最佳廣度與深度的語意分析排行榜。

2.3.7 結論與建議

期望透過本研究成果所建立之語意分析排行榜，瞭解究竟何種貼文類型或貼文時間點較易受到民眾的喜愛，並可做為未來海洋委員會與海洋委員會海巡署進行滾動式修正於社群平台的文章特性之依據，研究步驟圖如下圖 4。

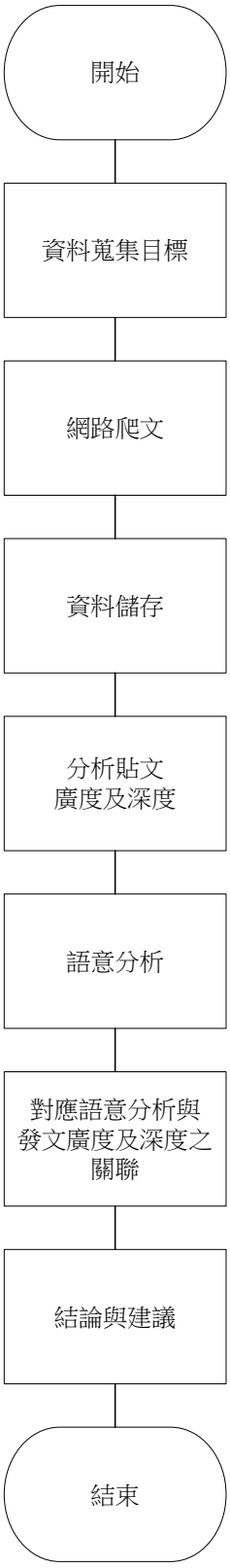
開始
資料蒐集目標
網路爬文
資料儲存
分析貼文
廣度及深度
語意分析
對應語意分析與
發文廣度及深度之
關聯
結論與建議
結束

圖 4 研究步驟圖

第三章　結果與討論

目前研究已初步透過爬文技術針對海洋委員會官方 Facebook 粉絲專頁蒐集資料，在本章中針對目前已蒐集之資料及其分析結果進行說明與報告。

因 Facebook 在 2020 年 9 月進行更新及改版，導致之前爬文的程式碼必須重新撰寫，蒐集到的資料也必須重新彙整，所以稍微改變了爬文的方式，改針對「純 HTML」版本的 Facebook 進行爬文。因此，在 3.1 節與 3.2 節將針對 Facebook 改版前所蒐集的資料與分析結果進行報告，而在 3.3 節將針對 Facebook 改版後資料介紹與說明進行資料分析解說。

3.1　Facebook 改版前資料介紹與說明

在 2020/4/20~2020/6/22 以蒐集 Facebook 改版前海洋委員會官方 Facebook 粉絲專頁所發表的文章，共有 63 則文章，資料說明如圖 7 所示。在圖 5 中，編號第 0 筆為海洋委員會官方 Facebook 粉絲專頁的置頂貼文，其貼文時間是 2019/9/30，而編號 1 是 2020/6/22 所發表的文章，編號 2~4 則都是 2020/6/20 所發表的文章，最後面編號 62 則是 2020/4/20 的貼文。

	內容	like總數
0	海洋委員會到底在幹嘛?很多人不清楚，沒關係，小編用了很久時間做了這個影片，大家看完就知道啦	5,442
1	先不要管PS5了有聽過我們海巡的「40快砲」嗎？全台都在瘋Plash Speed 5「路由器...	4,297
2	不能只有我看到全台都在瘋明天的日環蝕海巡署艦隊分署提前一天 超前部署拍到了！..... 更多	1.1萬
3	日環蝕時事問答題明天（21日），本世紀在台灣本島可看見日環蝕的最後一次機會，錯過再等195年...	3,071
4	帶著眾人祝福小虎鯨野放重返大海懷抱這次野放的小虎鯨是在4月25日晚間，迷航於高雄港口的小虎鯨...	9,748
...	...	...
58	ONLY ONLY U-U-U 唯獨你不行他就是「非洲豬瘟」，自107年以來不斷的威脅著我們...	1.1萬
59	沒有從天而降的英雄，只有挺身而出的凡人感謝全國的醫護人員，為臺灣堅守防線，全力做好工作，感謝...	9,883
60	愛自己與家人 就要懂得愛地球用行動去為自己、家人及地球盡一份心力你我能享有歲月靜好，是因為有...	3,885
61	你知道亂丟口罩的同時，不只會造成環境髒亂，還可能成為病菌孳生的溫床嗎？沒有，因為你只想到你自...	5,499
62	記得，我們的共同敵人是病毒！疫情嚴峻，人人都繃緊神經，難熬的，倒不是14天足不出戶，而是你會...	9,577

63 rows × 2 columns

圖 5 2020/4/20~2020/6/22 海洋委員會之貼文

為了驗證所爬文回來之訊息可對應且無誤，在此以編號 3 的貼文為例，其對

應在粉絲專頁真實截圖如圖 6 以及圖 7 所示，從此對應可以確認所蒐集的資料可以正確連接且資料無誤。然而從本次的驗證中亦發現，當所發表的文章其按讚數量破萬時，Facebook 會自動更動單位為萬，即無法確切得知千位數後面的詳細數據，所以未來在追蹤讚數的方面，還需要再更精準的抓取且確認資料，才能在不同時段的爬文中看出讚數的變化量，以便後續分析所發表的文章其廣度與深度之成效。

圖 6 6/20 貼文範例


All
5,994
5,149
82
39
更多
14
5
2
海洋委員會
Poppy Waffle 比利時列日鬆餅_義大利麵_咖啡_輕
店
中港香業神桌佛具
讚
新竹市警察局
讚
環寶島，環保島
讚
一木寫不完文字。五公里
讚
Porsche的逃跑日記
讚

圖 7 貼文讚數數據

3.2 Facebook 改版前資料語意分析初步結果

為了瞭解所蒐集之文章內容的語意結果，作為後續持續判斷貼文內容是否影響文章深度與廣度之關聯，因此，本計畫先初步透過結巴斷詞系統進行文句斷詞，進而建立關鍵詞彙頻率統計表，再將貼文內容的關鍵字透過視覺化建立斷詞頻率直方圖以及文字雲，最後再透過關聯規則建立語意網路，了解各辭彙之關聯，詳細內容如後續之說明。

3.2.1 斷詞結果

由於進行語意分析首先必須先將既有文句進行斷詞，而在本計畫中，目前先採用結巴斷詞套件進行分析，以蒐集資料中編號第 0 筆的置頂文為例，圖 8 為原先文章內容，透過結巴斷詞套件後，其斷詞結果的部分內容如圖 9 所示，從圖 11 中可以發現，結巴斷詞套件將原先置頂文斷成 24 個詞彙，但很明顯可以得知，有部分詞彙並沒有被正確切割，如「海洋委員會」應為一個詞彙，但卻被切成「海洋」與「委員會」，因此，若採用結巴斷詞套件時，必須另外驗證與建立個人詞彙資料庫，如圖 10 為在個人斷詞資料庫新增「海洋委員會」後的斷詞結果。而整體全部文章的斷詞詞彙頻率直方圖則如圖 11 所示。

此次斷詞分析結果可以很明顯看出頻率最高的斷詞是「更」、「多」，因為若在 Facebook 所發表的文章內文篇幅過長時，為了縮短篇幅的頁面，在文章內文尾就會顯示「……更多」，讓想看詳情的瀏覽者自行點開貼文，導致在爬文的時候，爬了一堆「……更多」，這是往後研究需要去改進。

圖 8 編號第 0 號的置頂文

V1
海洋
委員會
到底
在
幹嘛
很多
人
不
清楚
沒關係
小
編用
了
很
久
時間
做
了
影片
大家
看
完
就
啦

Showing 1 to 24 of 24 entries Previous 1 Next

圖 9 編號第 0 號的置頂文的原始斷詞結果

V1
海洋委員會
到底
在
幹嘛
很多
人
不
清楚
沒關係
小
編用
了
很
久
時間
做
了
影片
大家
看
完
就
啦

Showing 1 to 23 of 23 entries

Previous 1 Next

圖 10 新增斷詞彙後的斷詞結果

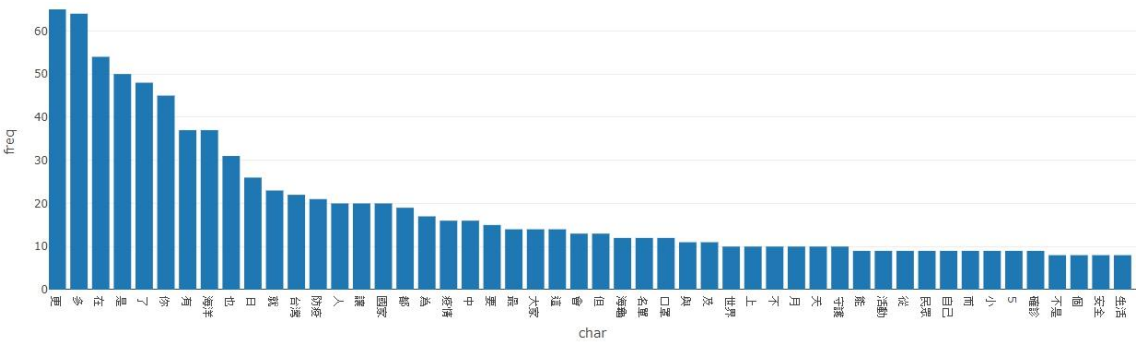

圖 11 新增斷詞彙後的文字直方圖

3.2.2 文字雲

在分別將 63 篇文章進行斷詞後，接著則以文字雲的方式呈現所蒐集資料之關鍵詞彙的文字雲，首先圖 12 為至少關鍵詞彙最小出現頻率 5 次的文字雲，由於當最小出現頻率設定為 5 時，所涵蓋的詞彙過多，無法很有效了解在該時段內，海洋委員會官方 Facebook 粉絲專頁所發表之內容的語意，因此，本計畫另外將關鍵詞彙最小出現頻率設定為 10 次，其文字雲分析結果如圖 13 所示。從圖 15 中可以很明顯發現，在該時段內剛好遇到有遊客潛水時誤碰觸到海龜以及疫情期間防疫等相關議題，所以很明顯的關鍵詞彙就是呈現相關的內容。

圖 12 文字雲分析(關鍵詞頻率大於 5)

圖 13 文字雲分析(關鍵詞頻率 10)

3.2.3 語意網路

由於文字雲僅能觀測出在特定時間內主要貼文之議題，但無法了解詞彙與詞彙之關聯，由於本計畫後續期望能夠透過建立語彙間的語意關聯來探討何種貼文形式較受民眾喜愛，因此，本計畫初步先透過關聯規則，來建立詞彙間的關係，進而建立語意網路。而在建立關聯規則時，須設定關聯規則的最小支持度（Minimum Support）與最小信賴度（Minimum Confidence），若設定的最小支持度與最小信賴度太低時，則容易產生太多規則，而造成決策上的影響。反之，設定門檻太高，則可能會因為規則太少，而面臨難以判斷的窘境。

圖 14 為先將最小支持度設定為 0.08 最小信賴度設定為 0.5 所建立的語意網路，從中可以發現其詞彙關聯較少，主要發現詞彙的關聯是在說明臺灣優秀的防疫能力，海洋委員會會持續守護邊界的安全。若放寬最小支持度與最小信賴度的標準至設定為 0.07 與 0.5 時，其結果如圖 15 所示，從圖 15 中可以發現，所產生的語意網路擴大，增加涵蓋了討論口罩防疫政策的問題。若持續將最小支持度與最小信賴度的標準放寬至 0.06 與 0.5 時，其結果如圖 16 所示，從圖 16 的結果可以知道，整體語意網路擴增為兩大網路，除了原先探討疫情相關的語意網路以外，增加了探討留言抽獎的議題。

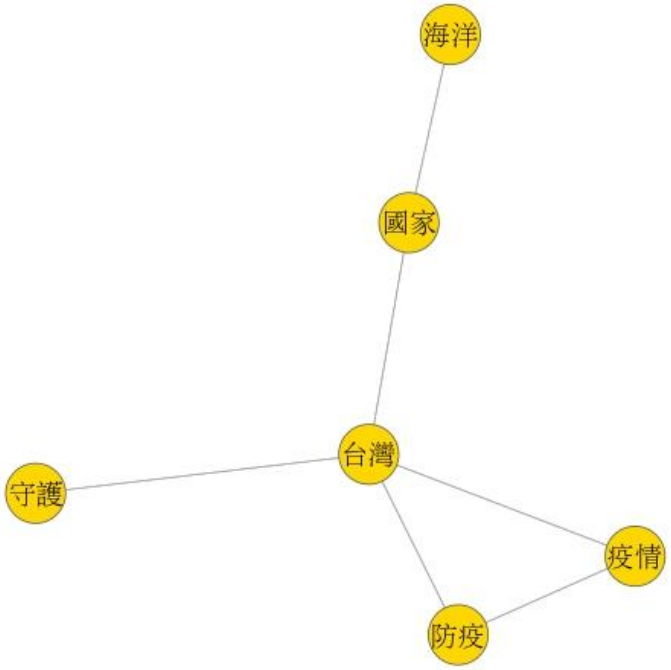

圖 14 設定最小支持度為 0.08，最小信賴度為 0.5 的語意網路

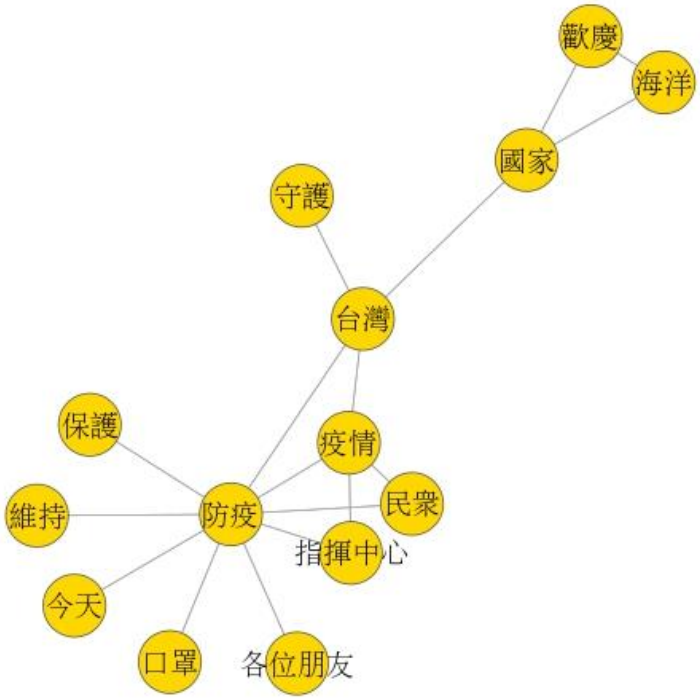

圖 15 設定最小支持度為 0.07，最小信賴度為 0.5 的語意網路

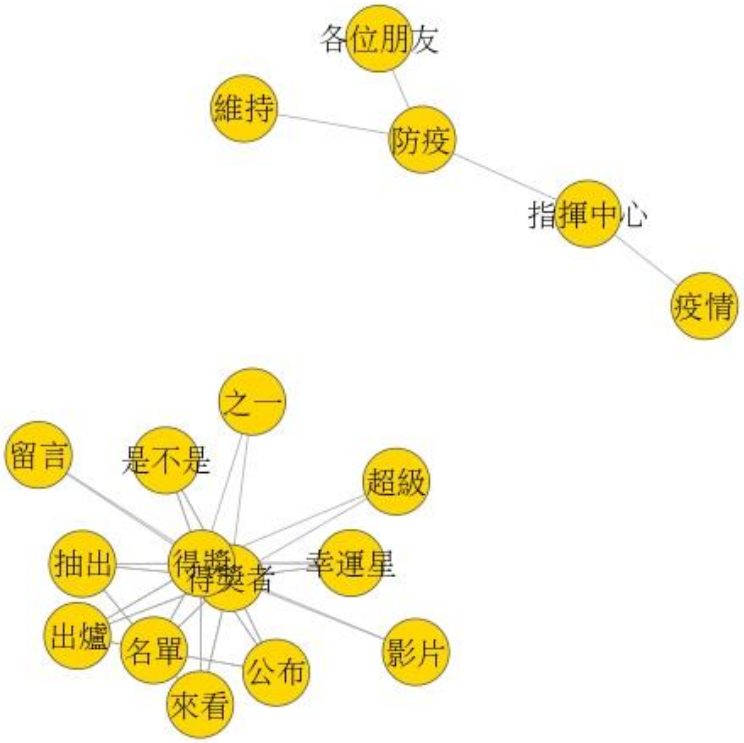

圖 16 設定最小支持度為 0.06，最小信賴度為 0.5 的語意網路

3.3 Facebook 改版後資料介紹與說明

因為 Facebook 於2020年9月更新與改版的關係，所以爬文方式進行大幅的修正，改爬取「純 HTML」版本的 Facebook 方式來獲取資料，以探討本研究之定義的廣度與深度。本計畫已完成自動化排程的爬文程式，每隔30分鐘進行一次爬文來獲取資料，而在進行貼文讚數及留言數的廣度與深度之分析時，將分析的主題分為兩種，分別為(1)貼文瞬間的等級；(2)貼文累積的等級。底下3.3.1節與3.3.2節將分別針對海洋委員會與海洋委員會海巡署 Facebook 官方粉絲專頁之資料與分析結果進行討論。

3.3.1 Facebook 抓取海洋委員會資料分析及語意分析

爬文以「純 HTML」版本的 Facebook 方式來獲取資料，每隔 30 分鐘進行一次爬文來獲取資料，此節是爬海洋委員會的 Facebook 官方粉絲專頁進行貼文讚數及留言數的廣度與深度之分析，將分析的主題分為兩種，分別為(1)貼文瞬間的等級；(2)貼文累積的等級。在目前的排程中，從 2020/9/26-202010/9，共獲取 8 則貼文，在伺服器的運作下因為學校斷電導致資料不完整，所以將分別針對其中較為完整的 4 則貼文進行分析與探討。

第一則貼文，貼文時間為 2020/10/1 11:55，而爬文的觀察時間的系統時間為 2020/10/1 12:10 ~ 2020/10/9 12:10 之間，並每間隔 30 分鐘進行一次爬文，代表本則貼文幾乎在發表後約 15 分鐘，即被本計畫所撰寫的程式擷取並儲存。圖 17 是第一則貼文的讚數及留言數之時間變化分析圖。從圖 17(a)可知，在觀察時間經過約第 80 個時間點後，則進入收斂，即貼文的觸及率已明顯降低；而從圖 17(b)的讚數的變化可知，最高變化點約在第 2-7 個時間點時，且每個時間點增加的按讚數超過 100 次，最高超過 350 次，因此，可知貼文的瞬間觸及率非常高；圖 17(c)同樣可呈現在約第 80 個時間點後，幾乎沒有新的回文

了；而從圖 17(d)可知，回文的高峰點在第 2 個時間點時，30 分鐘內增加了 7 則留言。整合本計畫所定義的貼文廣度與深度可知，此則貼文讚數貼文瞬間等級為 Level 3，留言貼文瞬間等級為 Level 1；此則貼文讚數貼文累積等級為 Level 3，留言貼文累積等級為 Level 1。

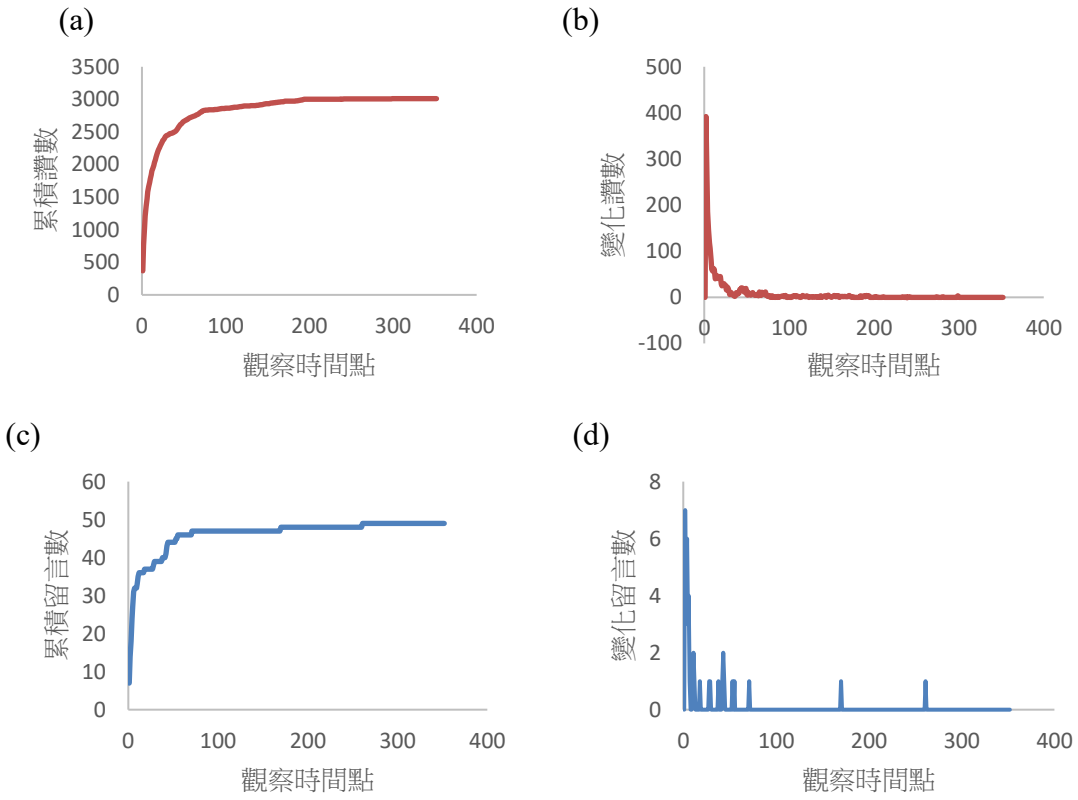

圖 17 (a)第一則文章在不同時間點的累積讚數；(b)第一則文章在不同時間點的變化讚數；(c)第一則文章在不同時間點的累積留言數；(d)第一則文章在不同時間點的變化留言數

第二則貼文，貼文時間為 2020/10/3 19:21，而爬文的觀察時間的系統時間為 2020/10/3 19:40~ 2020/10/9 12:10 之間，並每間隔 30 分鐘進行一次爬文，代表本則貼文在發表後約 30 分鐘，即被本計畫所撰寫的程式擷取並儲存。圖 18 是第一則貼文的讚數及留言數之時間變化分析圖。從圖 18(a)可知，在觀察

時間經過約第 70 個時間點後，則進入第二次收斂，即貼文的觸及率已明顯降低；而從圖 18(b)的讚數的變化可知，最高變化點約在第 1-5 個時間點時，且每個時間點增加的按讚數超過 200 次，可知貼文的瞬間觸及率非常高；圖 18(c)同樣可呈現在約第 79 個時間點後，幾乎沒有新的回文了；而從圖 18(d)可知，回文的高峰點在第 2 個時間點時，30 分鐘內增加了 37 則留言。此則貼文讚數貼文瞬間等級為 Level 4，留言貼文瞬間等級為 Level 3；此則貼文讚數貼文累積等級為 Level 3，留言貼文累積等級為 Level 3。

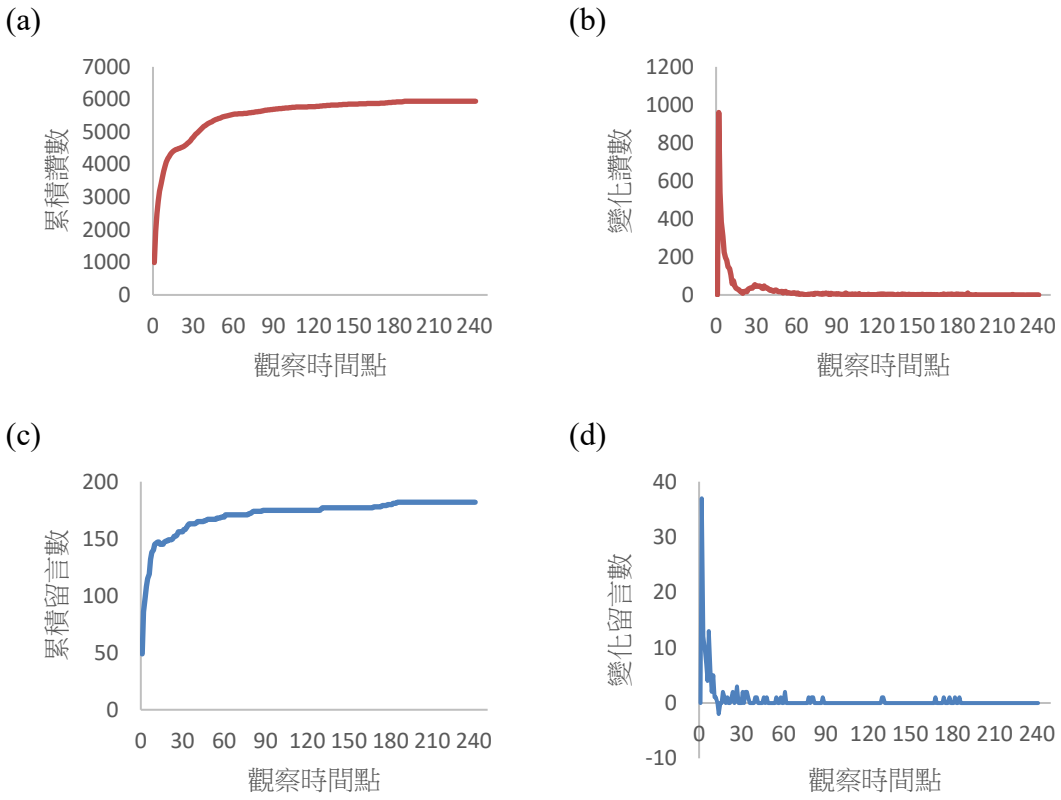

圖 18 (a)第二則文章在不同時間點的累積讚數；(b)第二則文章在不同時間點的變化讚數；(c)第二則文章在不同時間點的累積留言數；(d)第二則文章在不同時間點的變化留言數

第三則貼文，貼文時間為 2020/10/5 11:30，而爬文的觀察時間的系統時間為 2020/10/5 11:40~ 2020/10/9 12:10 之間，並每間隔 30 分鐘進行一次爬文，本

則貼文在發表後 10 分鐘，即被本計畫所撰寫的程式擷取並儲存。圖 19 是第一則貼文的讚數及留言數之時間變化分析圖。從圖 19(a)可知，在觀察時間經過約第 40 個時間點後，則進入收斂，即貼文的觸及率已明顯降低；而從圖 19(b)的讚數的變化可知，最高變化點約在第 2-12 個時間點時，且每個時間點增加的按讚數超過 100 次，最高超過 1000 次，因此，可知貼文的瞬間觸及率非常高；圖 19(c)同樣可呈現在約第 40 個時間點後，幾乎沒有新的回文了；而從圖 19(d)可知，回文的高峰點在第 2 個時間點時，30 分鐘內增加了 63 則留言。整合本計畫所定義的貼文廣度與深度可知，此則貼文讚數貼文瞬間等級為 Level 5，留言貼文瞬間等級為 Level 4；此則貼文讚數貼文累積等級為 Level 4，留言貼文累積等級為 Level 2。

第四則貼文，貼文時間為 2020/10/6 11:49，而爬文的觀察時間的系統時間為 2020/10/6 12:10~ 2020/10/9 12:10 之間，並每間隔 30 分鐘進行一次爬文，本則貼文在發表後約 21 分鐘，即被本計畫所撰寫的程式擷取並儲存。圖 20 是第一則貼文的讚數及留言數之時間變化分析圖。從圖 20(a)可知，在觀察時間經過約第 40 個時間點後，則進入收斂，即貼文的觸及率已明顯降低；而從圖 20(b)的讚數的變化可知，最高變化點約在第 2-9 個時間點時，且每個時間點增加的按讚數超過 100 次，最高接近 800 次，因此，可知貼文的瞬間觸及率非常高；圖 20(c)同樣可呈現在約第 40 個時間點後，幾乎沒有新的回文了；而從圖 20(d)可知，回文的高峰點在第 3 個時間點時，30 分鐘內增加了 12 則留言。而其中很明顯發現在第 60 個時間點有起伏一次，是因為個時間點有資料斷層，蒐集資料時間較長而導致此現象。整合本計畫所定義的貼文廣度與深度可知，此則貼文讚數貼文瞬間等級為 Level 2，留言貼文瞬間等級為 Level 1；此則貼文讚數貼文累積等級為 Level 2，留言貼文累積等級為 Level 1。

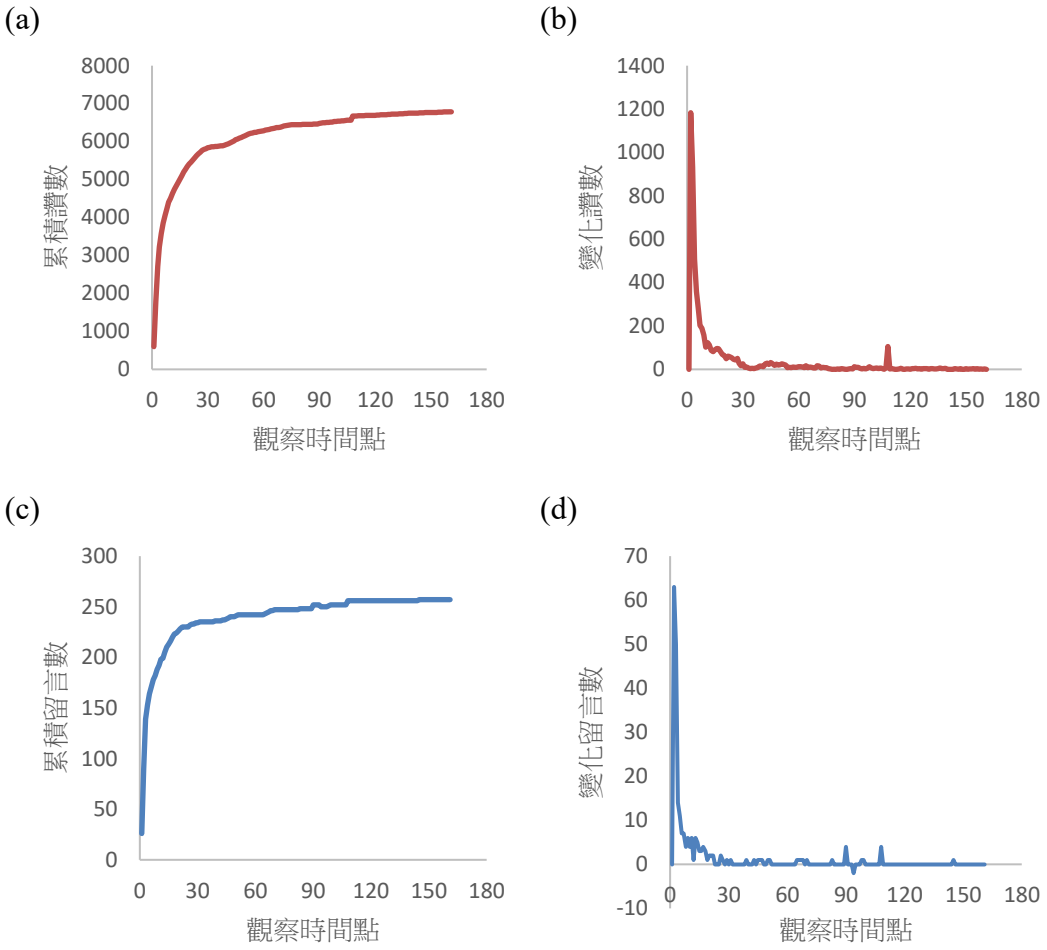

圖 19 (a)第三則文章在不同時間點的累積讚數；(b)第三則文章在不同時間點的變化讚數；(c)第三則文章在不同時間點的累積留言數；(d)第三則文章在不同時間點的變化留言數

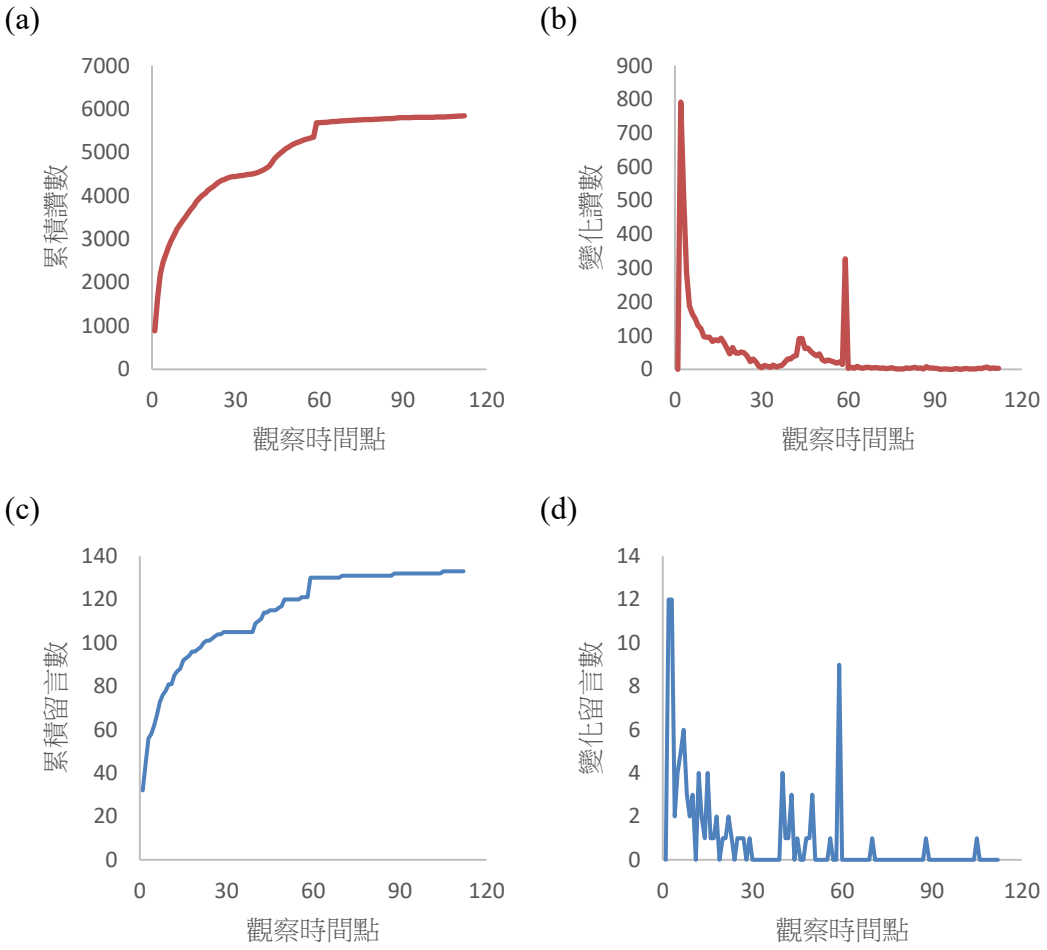

圖 20 (a)第四則文章在不同時間點的累積讚數；(b)第四則文章在不同時間點的變化讚數；(c)第四則文章在不同時間點的累積留言數；(d)第四則文章在不同時間點的變化留言數

3.3.1.1 海洋委員會 Facebook 粉絲專頁語意網路分析

爬文時間 2020/9/26-202010/9，以下圖 21-25 為蒐集期間的貼文進行語意網路分析結果。

陸棚
海域
具有
光合作用
充足
營養
鹽類
豐富
優勢
條件
不僅
培育
了
多樣化
海洋生物
更是
維繫
漁業
產業
經營
之
重要
命脈
是
絕對
禁止

Showing 1 to 50 of 1,107 entries

Previous 1 2 3 4 5 … 23 Next

圖 21 2020/9/26-202010/9 斷詞結果

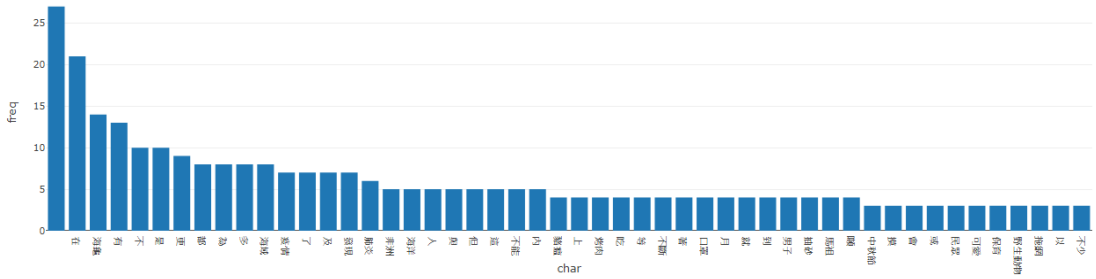

圖 22 2020/9/26-202010/9 文字直方圖

圖 23 2020/9/26-202010/9 文字雲(頻率 2)

由圖 24 以及圖 25 可明顯發現，這段期間的貼文都在討論保育海龜的議題，還有持續發燒的肺炎疫情。

圖 24 2020/9/26-202010/9 文字雲(頻率 7)

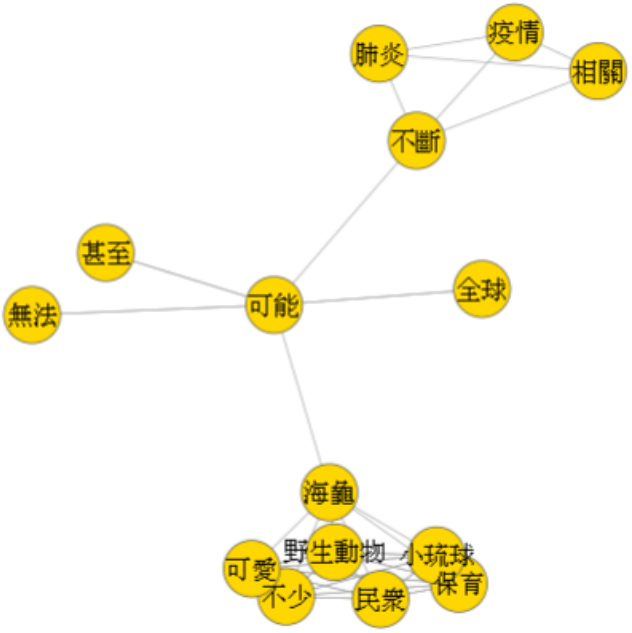

圖 25 設定最小支持度為 0.06，最小信賴度為 0.5 的語意網路

3.3.2 Facebook 抓取海洋委員會海巡署資料分析及語意分析

爬文以「純 HTML」版本的 Facebook 方式來獲取資料，每隔 30 分鐘進行一次爬文來獲取資料，此節是爬海洋委員會海巡署的 Facebook 官方粉絲專頁進行貼文讚數及留言數的廣度與深度之分析，將分析的主題分為兩種，分別為(1)貼文瞬間的等級；(2)貼文累積的等級。在目前的排程中，從 2020/9/26-202010/10，共獲取 6 則貼文，在伺服器的運作下因為學校斷電導致資料不完整，所以將分別針對其中較為完整的 2 則貼文進行分析與探討。

首先是第一則貼文，貼文時間為 2020/10/5 17:41，而爬文的觀察時間的系統時間為 2020/10/5 21:00~ 2020/10/10 23: 30 之間，並每間隔 30 分鐘進行一次爬文，代表本則貼文幾乎在發表後約 4 小時內，即被本計畫所撰寫的程式擷取並儲存。圖 26 是第一則貼文的讚數及留言數之時間變化分析圖。從圖 26 (a)可知，在觀察時間經過約第 18 個時間點後，則進入收斂，即貼文的觸及率已

明顯降低；而從圖 26(b)的讚數的變化可知，最高變化點約在第 2-7 個時間點時，在第 2-7 個時間點上增加的按讚數超過 100 次，因為在發文 4 小時候才擷取到資料，因此，貼文的瞬間觸及率應該還要更高；圖 26(c)同樣可呈現在約第 40 個時間點後，幾乎沒有新的回文了；而從圖 26(d)可知，回文的高峰點在第 31 個時間點時，30 分鐘內增加了 3 則留言。整合本計畫所定義的貼文廣度與深度可知，此篇貼文讚數貼文瞬間等級為 Level 2，留言貼文瞬間等級為 Level 1；此篇貼文讚數貼文累積等級為 Level 4，留言貼文累積等級為 Level 2。

第二則貼文，貼文時間為 2020/10/10 08:30，而爬文的觀察時間的系統時間為 2020/10/10 09:00~ 2020/10/10 23:30 之間，並每間隔 30 分鐘進行一次爬文，發文後 30 分鐘內被本計畫所撰寫的程式擷取並儲存。圖 27 是第二則貼文的讚數及留言數之時間變化分析圖。因時間不夠長的問題所以從圖 27(a)可明顯得知，曲線持續上升，很難找到收斂點；而從圖 27(b)的讚數的變化可知，最高變化點約在第 2 個時間點時，且增加的按讚數超過 4000 次，因此，可知貼文的瞬間觸及率非常高；圖 27 (c)同圖 27(a)的問題；而從圖 27(d)可知，回文的高峰點在第 2 個時間點時，30 分鐘內增加了 4000 則留言。此則貼文較為特例，讚數及留言數都非常相近。整合本計畫所定義的貼文廣度與深度可知，此篇貼文讚數貼文瞬間等級為 Level 5，留言貼文瞬間等級為 Level 5；此篇貼文讚數貼文累積等級為 Level 5，留言貼文累積等級為 Level 5。

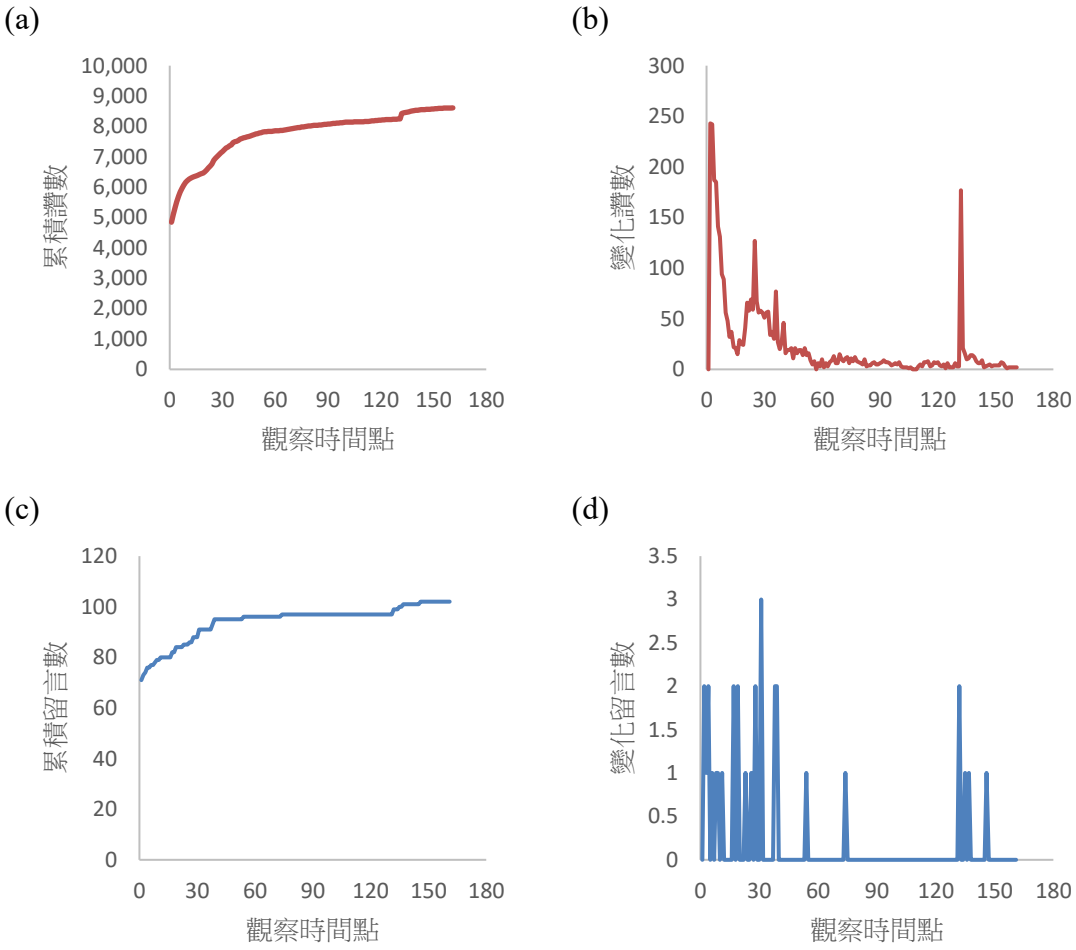

圖 26 (a)第一則文章在不同時間點的累積讚數；(b)第一則文章在不同時間點的變化讚數；(c)第一則文章在不同時間點的累積留言數；(d)第一則文章在不同時間點的變化留言數

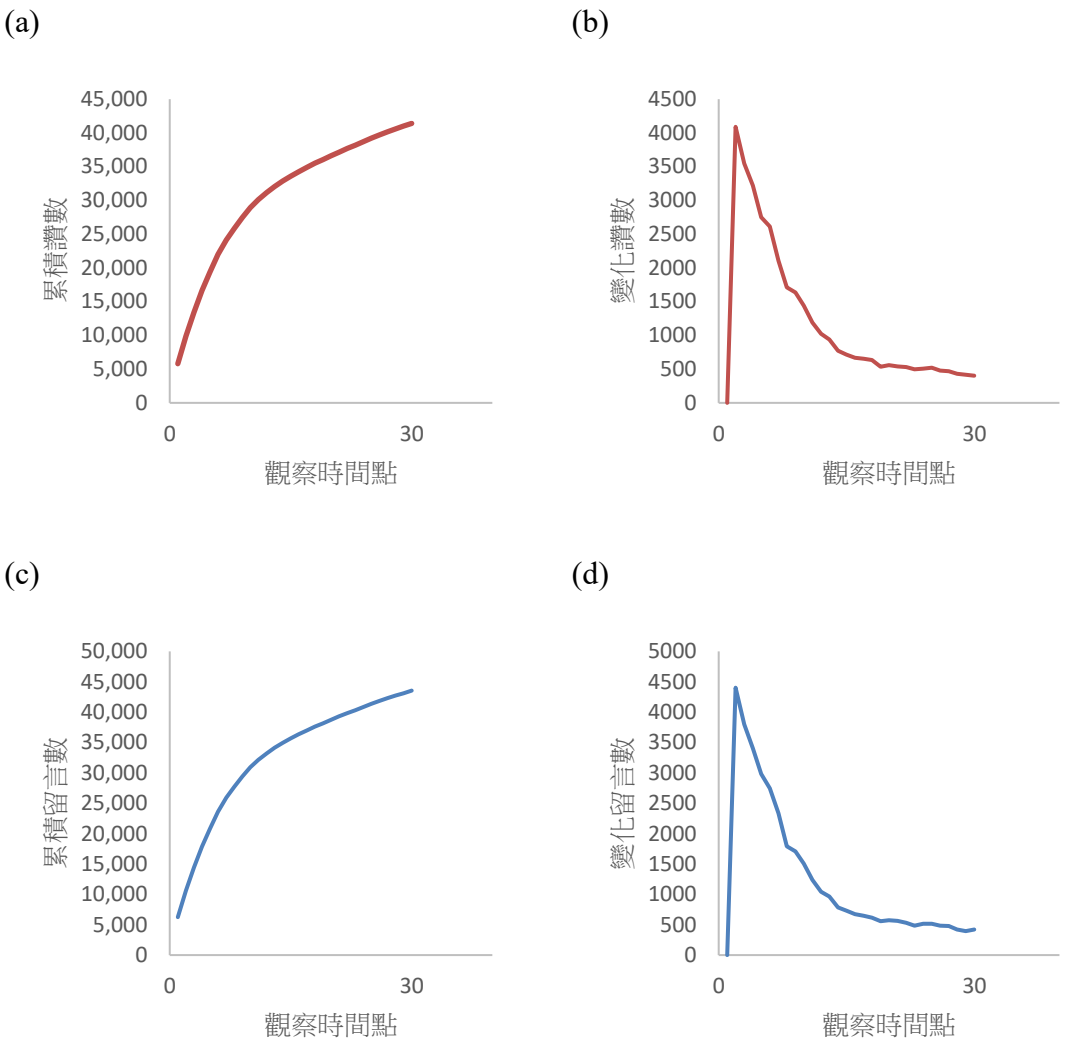

圖 27 (a)第二則文章在不同時間點的累積讚數；(b)第二則文章在不同時間點的變化讚數；(c)第二則文章在不同時間點的累積留言數；(d)第二則文章在不同時間點變化留言數

3.3.2.1 海洋委員會海巡署 Facebook 粉絲專頁語意網路分析

爬文時間 2020/9/26-202010/10，以下圖 28-32 為蒐集期間的貼文進行語意網路分析結果。

隨棚
海域
具有
光合作用
充足
營養
鹽類
豐富
優勢
條件
不僅
培育
了
多樣化
海洋生物
更是
維繫
漁業
產業
經營
之
重要
命脈
是
絕對
禁止

Showing 1 to 50 of 1,107 entries

Previous 1 2 3 4 5 … 23 Next

圖 28 2020/9/26-202010/10 斷辭結果

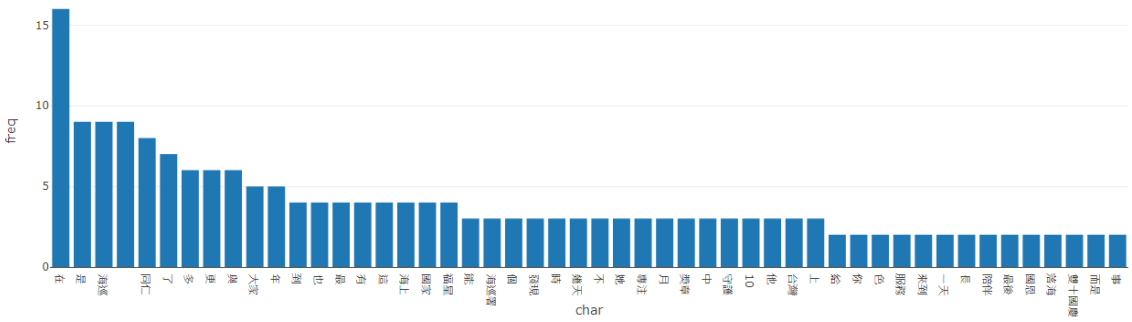

圖 29 2020/9/26-202010/10 文字直方圖

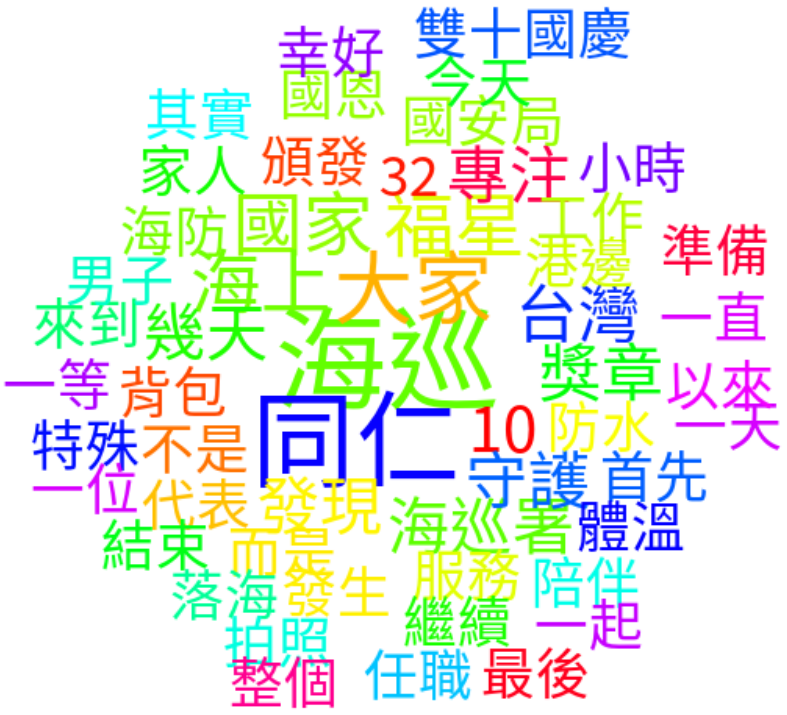

圖 30 2020/9/26-202010/10 文字雲(頻率 2)

圖 31 2020/9/26-202010/10 文字雲(頻率 7)

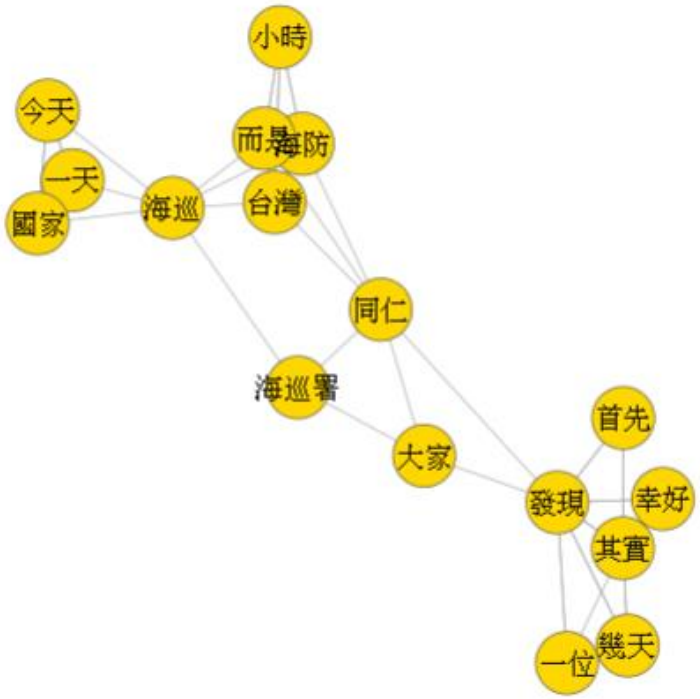

圖 32 設定最小支持度為 0.06，最小信賴度為 0.5 的語意網路

第四章　結論

目前本計畫已完成海洋委員會以及海洋委員會海巡署官方 Facebook 粉絲專頁爬文部分之內容，但未來仍須持續進行以達到本計畫所規劃之研究目的，底下將說明本計畫在執行過程中的研究結論及未來展望。

4.1　研究結論

依目前研究成果，在分析過程中須了解每則貼文在不同時間點的留言與按讚數量之變化，因此，必須定期爬文獲取最新的資料並進行不同時間點資料變化之差異比較，已經完成自動化排程的技術定期爬文，以獲得最新的資料。

在多則貼文分析下，我們發現海洋委員會的 Facebook 粉絲專頁在更新貼文上有兩個時段，1 是接近中午 12 點，2 是下午 4 點之後，依目前的貼文資料分析，由本計畫定義的廣度深度排行榜看來，在中午時段發文的效果較優於下午發文，從語意網路分析結果，近期都在討論海龜保育的議題。

4.2 研究未來發展

在結案的爬文過程中，雖然已經利用排程技術自動爬文，但是還是得進行人工作業的檢查，隨時關注爬文上要擷取之 Facebook 網頁的原始碼是否有在更新，也要確認爬文過程沒有中斷，若出了問題要馬上對程式碼的編寫進行更改，目前規劃是每日每隔 30 分鐘進行一次爬文，設為蒐集資料的頻率。

由結案報告中總共抓取到 14 則貼文，海洋委員會 8 則，海洋委員會海巡署 6 則，取 4 則海洋委員會之貼文以及 2 則海洋委員會海巡署之貼文進行分析，貼文時間距的不同，越即時抓取到的貼文比較有分析的價值性，超過兩天後的貼文基本上都已經失去熱度進行收斂，由目前分析結果顯示的讚數及留言數的排行榜，較新的貼文可以明顯分析出廣度與深度瞬間的 Level，而較舊的貼文可以明顯分析出廣

度與深度兩天內累積的 Level，其中深度(留言數)的分析 Level 不太明顯，往後希望能用 Facebook 的 API 串接，更深入去探討粉絲年齡層對於不同貼文類型的文章來進行貼文熱度的分析，以作為後續海洋委員會以及海洋委員會海巡署未來在社群平台上發文方式的修正依據。

附錄

計畫開發之程式碼

```
from selenium import webdriver
import time
profile = webdriver.ChromeOptions()
driver = webdriver.Chrome(chrome_options=profile)
driver.get("http://mbasic.facebook.com")
time.sleep(3)
driver.find_element_by_id("m_login_email").send_keys('XXX@gmail.com') #  將 USERNAME 改為你的臉書帳號
driver.find_element_by_name("pass").send_keys('XXX') #  將 PASSWORD 改為你的臉書密碼
driver.find_element_by_name("login").click() #新版臉書有更新  登入 ID
time.sleep(3)
driver.get('https://mbasic.facebook.com/CGA4U/')    #已 OK
from bs4 import BeautifulSoup #+4
import datetime
#import re
#htmltext = driver.page_source #  將網頁原始碼拿出來

soup = BeautifulSoup(htmldoc, 'html.parser')
systime=datetime.datetime.now() #抓系統時間

print(soup.prettify())
#body.prettify() #確認抓取的 html 格式

content= soup.body.find_all("div", class_="gb")
#抓取文章內容
content[0].get_text()

time = soup.body.find_all("abbr")
#抓取文章內容
time[0].get_text()

like= soup.body.find_all("a", class_="go") #.偶數純讚數
like[0].get_text()
#len(content) #一頁五篇
  #0  第一篇讚數    #1  第一篇留言數      #2  說這專頁讚
  #3  第二篇讚數    #4  第一篇留言數      #5  說這專頁讚
  #6  第三篇讚數    #7  第一篇留言數      #8  說這專頁讚
  #9  第四篇讚數    #10 第一篇留言數      #9  說這專頁讚
```

```
#12  第五篇讚數   #13 第一篇留言數   #14  說這專頁讚
message= soup.body.find_all("a", class_="go") #奇數抓留言數
message[1].get_text()
#len(content) #一頁五篇
 #0  第一篇讚數   #1  第一篇留言數     #2  說這專頁讚
 #3  第二篇讚數   #4  第一篇留言數     #5  說這專頁讚
 #6  第三篇讚數   #7  第一篇留言數     #8  說這專頁讚
 #9  第四篇讚數   #10 第一篇留言數     #9  說這專頁讚
#12  第五篇讚數   #13 第一篇留言數    #10  說這專頁讚

##合併資料
#先將資料轉為 list
content_list=[]
for i in range(0,5):
    content_list.append(content[i].get_text())
like_list=[]
for i in range(0,15,3):
    like_list.append(like[i].get_text())
time_list=[]
for i in range(0,5):
    time_list.append(time[i].get_text())
message_list=[]
for i in range(1,15,3):
    message_list.append(message[i].get_text())
systime_list=[]
for i in range(0,5):
    systime_list.append(systime.strftime('%Y-%m-%d %H:%M:%S'))

import pandas as pd
#合併成 dataframe
newlist=list(zip(content_list,like_list,time_list,message_list,systime_list))
newlist=pd.DataFrame(newlist,columns=['內容','like 總數','發文時間','留言數','系統
時間'])

newlist

#寫出匯出資料檔案類型 CSV
date1=datetime.date.today().strftime('%Y-%m-%d')
filename="d:/WAYNNE/CGA"+date1+".csv"
newlist.to_csv(filename, index=False,encoding="utf_8_sig",mode='a')

#關閉瀏覽器退出驅動程序
driver.quit()
```

參考文獻

1. 郭伯臣, 廖晨惠, & 張正杰. (2018). 應用潛在語意分析增強學生的閱讀能力. 數位學習科技期刊, 10(1), 31-55.
2. 陳林志, & 陳冠瑜. (2015). 利用語意分析模型分析谷歌部落格搜尋引擎效能. 第二十六屆國際資訊管理學術研討會, 臺北.
3. 陳怡廷, 陳麗如, & 吳姿瑩. (2016). 從部落格探索客家旅遊目的地意象之研究—自然語言處理的方法與應用. 戶外遊憩研究, 29(2), 81-111.
4. 管瓊瑛, 謝寧, 陳潔, 張桂萍, 高翊瑋, 謝邦昌, 張嘉芳, & 張耀懋. (2017). 長期照顧政策是照顧老人還是失能者？－以蔡英文臉書為例探勘民眾認知. 台灣公共衛生雜誌 36(5), 511-520.
5. 斷開中文的鎖鍊！自然語言處理 (NLP)，〈未知詞擷取作法〉(7.Nov.2018)。<https://research.sinica.edu.tw/nlp-natural-language-processing-chinese-knowledge-information/>(1.Mar.2020).
6. Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. Expert Systems with Applications, 34(3), 1707-1720.
7. Hutchison, P. D., Daigle, R. J., & George, B. (2018). Application of latent semantic analysis in AIS academic research. International Journal of Accounting Information Systems, 31, 83-96.
8. Landauer, T., Laham, D., & Foltz, P. (1999). Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report.
9. 2019 Instagram 熱門#Hashtag 分析- Social Lab 社群實驗室，< OpView 社群口碑資料庫>(6.Feb.2020)。< https://www.social-lab.cc/2020/01/social-insights/2019-instagram%E7%86%B1%E9%96%80hashtag%E5%88%86%E6%9E%90/>(1.Mar.2020).