關聯式文本探勘資訊探索實驗平台設計—以「二二八事件臺灣本地新聞史料彙編」為例

甯格致*、劉吉軒**、薛化元***、蔡銘峰****

摘要

文本探勘之資訊探索取向差異

對人文研究者來說，一個具有意義的思維或觀點的論證過程中，往往牽涉大量文本資料的爬梳，篩選出研究目標相關資訊，進而由諸多線索累積為更具輪廓的脈絡，而這些過程也往往會引發研究問題的重新定義、聚焦、深化。數位人文研究以電腦的資料處理與計算能力，協助研究者以全新的方式從資料中尋找答案，如同透過一個可移動、可調整的鏡片，以微觀、巨觀或不同視角的彈性檢視能力，分析大量人文資料，探討人文議題，解讀人文現象[1]。目前許多數位人文研究應用資訊技術於史料文本的資訊探索取向，乃先鎖定特定文字所代表之主題概念或現象，再從史料文本中搜尋比對，而以統計量化分析方法，從數量、比例等觀點，觀察主題概念或現象的顯著或差異程度，進而驗證部分假設或獲取片段式的新資訊[2][3]。這種資訊探索取向是一種被描述為「hunt and peck」的單一目標費力搜尋方式，或以「slicing」橫向切片 、連貫比較的方式，找出趨勢或異常現象[4][5]。

大量文本資料往往隱藏之意義豐富的主題資訊，其中錯綜複雜的關係與層層因果的交疊，更需要資訊技術的功能突破，以有效的關聯挖掘，具體呈現其脈絡面貌，協助人文研究者解讀與發現。單一視角的現象挖掘，雖然能彰顯特定主題資訊的意義，但通常也忽略了關聯性資訊與脈絡結構的關鍵性。有鑑於此，本研究除了採納過去社會、心理研究領域之質化分析精神，即看重每一個代表人、事、物之個體的相對存在意義，進一步考量每一個體與周遭其他個體所關聯的局部情況，將個體之間的關聯視為是一種社會定位的呈現；再進而以橫向角度觀察比較同類型個體與個體之間的異同、或以縱向角度觀察比較基於不同環境、不同時空

*國立政治大學資訊科學系研究助理，Email：floater.xkernel@gmail.com。
**國立政治大學資訊科學系副教授，Email：jsliu@cs.nccu.edu.tw。
***國立政治大學台灣史研究所教授，Email：hyh5595@gmail.com。
****國立政治大學資訊科學系助理教授，Email：mftsai.mftsai@gmail.com。

下所浮現出的脈絡樣貌。這樣的概念是利用先發掘出較具意義的關鍵個體，進一步觀察個體所擁有之關聯情況，再施以橫向合併、縱向貫穿的資訊凝聚視野，期能以一種多層次的資料維度觀點，進行較深度的資訊擷取及關聯探索，協助研究者得到更具深刻意義的發現與結果。

關鍵字：史料資訊探勘、社會網路分析、資訊檢索

Designing an Experiment Platform for Information Exploration with Relational Text Mining: A Case Study with the Taiwan 228 - event News Archive

Ke-chih Ning[*]、Jyi-shane Liu[**]、Hua-yuan Hsueh[***]、Ming-feng Tsai[****]

Abstract

This study integrates methods on computer science and social sciences and, with historian perspectives, views historical text as embedding a miniature social system. The task involves extracting relations among entities from text and performing structural analysis of the constructed entity-relationship network. One of the primary goals is to find the key-role actor and reveal its social position, which may defined by certain incidents, words, behaviors. Another further goal is to find other actors with similar social position and identify the underlying community. Finally, an abstract social role can be characterized to provide insight on the constructed social system from text. We develop an experimental platform – PARTEX, which provides text analytic tools and allows exploratory observation on relational structure among entities. Among our well-preprocessed and imported document collections, with historian inputs on key conceptual words as focal issues, the platform has been used to identify entity relations and construct the embedded social system. Discourse perspectives of position, demand, emotion, and action, are investigated with contextual parameters of boundary, association type, and relational strength. Both visual representation of the discourse-oriented social system and the quantitative measures are presented for analytic comparison. This study hopes to provide an effective text analytic tool and contribute in discovering historical implications. We intend to further improve the platform by recursive use test and validate the approach by fostering fruitful research results.

Keywords: Historical Text Mining, Social Network Analysis, Information Retrieval, Critical Discourse Analysis

[*]Research Assistant, Department of Computer Science, National Cheng-chi University. Email: floater.xkernel@gmail.com.
[**]Associate Professor, Department of Computer Science, National Cheng-chi University. Email: jsliu@cs.nccu.edu.tw.
[***]Professor, Department of History, National Cheng-chi University. Email: hyh5595@gmail.com.
[****] Assistant Professor, Department of Computer Science, National Cheng-chi University. Email: mftsai.mftsai@gmail.com.

壹、目的，工具提供，使用者導向之資訊探索平台—PARTEX

本研究目的在於將社群網絡分析方法與資訊檢索技術導入於人文研究領域中，並考量以資訊技術為工具，以人文目標為依歸，發展出一個可於大量中文文字內容，藉由關聯式網絡模型的資料呈現方法，型塑出反映自文本之重要議題的情境及脈絡，達到更有效的資訊發掘能力；並於特定實驗文本史料中，藉由概念的具體實現，呈現文本中各種關係的紋理及脈絡的檢視分析，同時反覆檢驗概念模型與觀察到的結果，確立彼此整合的可行性。目前資訊技術應用於人文研究領域的成果尚有許多發展空間，打破兩者之間隔閡的關鍵，應不在於單純提升資訊技術的介入程度，並取得數量上的回饋，而是應將現有的資訊技術方法，於使用者介面、資料處理流程、資料整合、資訊協作、資訊附加、資訊擷取、資訊視覺化等等的方式及角度，以一完整系統的觀點進行整合，最終才能成為適合人文研究者使用的分析工具，並促成人文研究者的實際使用，得到更具有價值的產出。

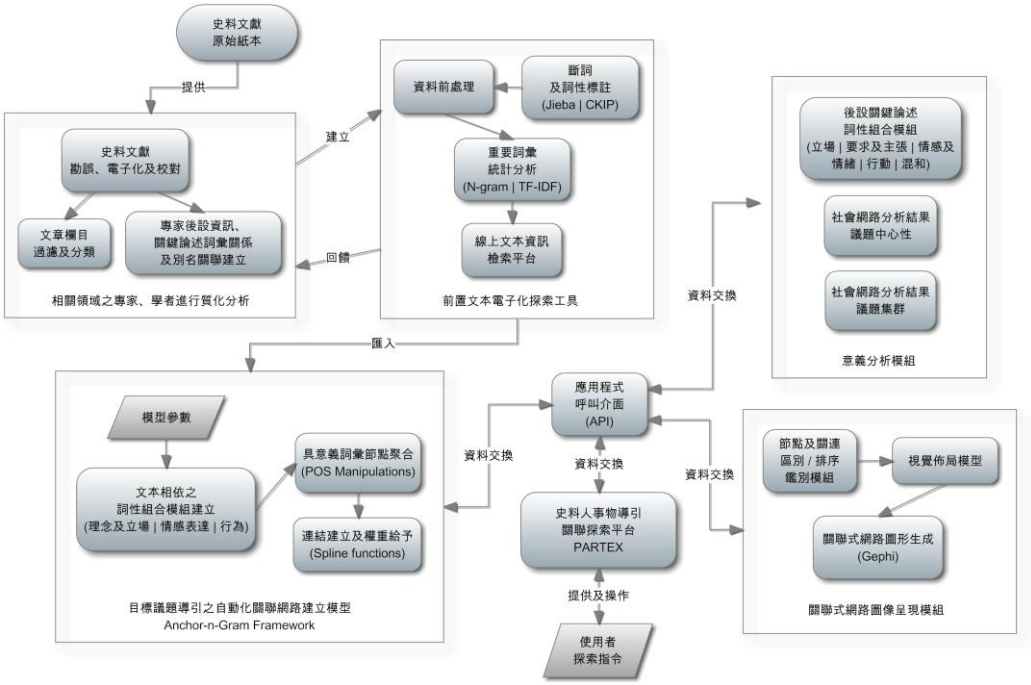

圖 1　史料文本探索平台 PARTEX

本研究以呈現人文價值為目標，以資訊科學工具為基礎，開發出一個以關鍵字檢索方式探索文本資料自身蘊含資訊的實驗觀察平台 - PARTEX，結合網路主從式伺服架構的資訊儲存、共享、協作等優勢，相較於以往紙本或是非結構化的電子文本資料檢索，能提供更快速、便利、更多層次的檢索、觀察及分析效果。PARTEX 允許藉由電子化後的文本、文本專屬的後分類資訊以及多項可變動的探索參數給予，不但可產生較符合人類觀察習慣之視覺化關聯式網路呈現成果，亦可產生數種較具有實質意涵的量化指標，讓人文研究者在各種目標議題引導之情境及資料觀察的交互比較下，探索其中人、事、

物種種複雜關係，開啟史料文本資訊中呈現出不同風貌的可能性。PARTEX 平台以模組化功能觀點來看，圖 1 的整體資料及處理程序流程圖中包括了以下主要處理程序：

1. 史料文本電子化及領域專家學者後分類資訊彙整
2. 斷詞及初步統計分析及篩選
3. 目標議題導引之關聯式網路建立模型：Anchor-n-Gram
4. 基於社會網路分析之關係意義分析模組
5. 視覺化呈現模組
6. 可調整參數及使用者互動介面

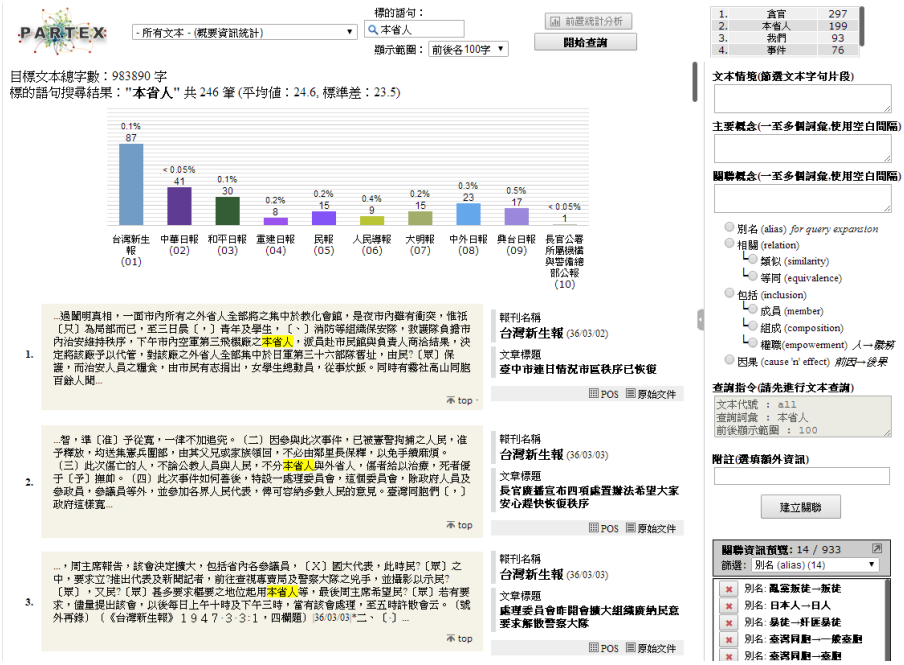

圖 2　文本史料探索平台 PARTEX 介面展示

概括而言，PARTEX 提供了史料及人文研究者一個新形態的資訊探索平台，可供儲存或分享整理完成之史料文本電子全文，並藉由領域專家以統計工具進行後分類資訊標記及初步篩選後，再從各種目標議題及探索參數出發，以半自動化方式產生具有特定意義之社會網路視覺呈現以及社會網路分析之量化結果。互動式的操作亦允許資料處理過程的階段性觀察，以利嘗試從各種角度、範圍及深度進行資訊探索，最終匯集形成一具有更上層事物具體概念意義的觀察分析結果。圖 2 即為 PARTEX 平台之概要功能及介面展示。

貳、目標議題導向之關聯網路建立—Anchor-n-Gram

本研究提出了一個目標議題導引之關聯式網路建立模型（Anchor-n-Gram），以人、事、物等概念詞彙為起點，進行資訊擷取及關聯探索，並考量斷詞的結果往往會將原本較具有意義的詞彙分解打散成為最基本的詞彙單元（Part-Of-Speech，POS），以便能夠

進行詞性標註（POS tagging）。然而這種基本的詞彙單元往往呈現的意義太具普世性（Generic），而無法充分反映該文本行文論述中試圖描繪的情境及實體之間的關連性，若依此詞彙單元建構出詞彙關聯式網路，整體的意義也會較顯得鬆散，而難以凝聚出更有意義的資訊。有鑑於此，Anchor-n-Gram 方法考量以詞性標記為基礎，建立了詞彙重整併機制，將已被分解成為單元的詞彙經由自然語言的規則性，再次整併為一較完整且指涉較為精確的詞彙。Anchor-n-Gram 處理程序的資料輸入端為已經經過中文斷詞處理並標記詞性後的資料，最後輸出的資料為詞彙關聯式網路，並可供作後續網路結構的量化分析、視覺化網路呈現、或是更上層質化意義的解釋用途。

處理程序其中的細節展開則包括：

(1) 透過領域專家學者建立出與論述語句模式相依的後分類資訊；

(2) 以統計方法彙整及篩選得到合適的詞性模組；

(3) 以迭代方式逐步整併遵循此詞性規則之語句直至收斂為止，獲得較具有意義之詞彙整併結果；

(4) 依據詞性模組的規則將符合的詞彙兩兩建立連結；

(5) 給予量化曲線函式（Spline Function），對連結之權重進行數值重新量化或是平滑作用；

(6) 最終於各論述片段及片段自身所產生的詞彙節點及關連依據相同的節點加以整併、累加。圖 3 則為 Anchor-n-Gram 處理程序之概要步驟實例。

圖 3　Anchor-n-Gram 概要步驟實例

參、以「二二八事件臺灣本地新聞史料彙編」為例之實驗成果

本研究以「二二八事件臺灣本地新聞史料彙編」史料文本為例，收錄從二二八事件引爆點翌日之 1947 年 2 月 28 日起，至 1947 年 5 月 15 日統治者清鄉結束止，蒐集主題以攸關二二八事件之新聞文本為主，兼及報上言論、官方公告及民間啟事，文本資料共分為十大報刊（其中屬於官方報刊者有三刊、屬於民間者有六刊、屬於官方偏民間者有一刊），字數共計約 88 萬餘字。我們將全文電子化並進行校訂後導入本研究開發之實驗觀察平台 PARTEX 中，在經由專家知識的輔助下，首先以「本省人」以及「外省人」做為初步觀察的關鍵詞彙，依循目標議題導引的詞彙關聯式網路模型，於「台灣新生報」以及「民報」兩報刊中分別以四種基於網路結構計算之中心性（Degree Centrality、Betweenness Centrality、Closeness Centrality、Eigenvector Centrality）、以及一種群集偵測方式（Modularity）建立出詞彙關聯網路，再依據目標關鍵論述之「立場、要求及主張」、「情感及情緒」、「行為及行動」此四種論述型式來篩選並進行綜合觀察。其中，「本省人」與「外省人」同屬事件發生的角色個體，是屬平行的橫向概念詞彙，而兩報刊的立場分屬官方及民方，是屬於縱向的詞彙，依據關聯式網路透過網路結構計算後及關鍵論述型式篩選出的關聯情況，則可視為是此多元觀點下個體所關聯的內容。藉由依循此橫向以及縱向的多維度觀察方式，我們發現了許多原本較不易經由人工閱讀及歸納得到的資訊，像是本省人與外省人在事件發生的前後，各自所關注的議題、立場及角色的差異情形、以及各種官方民間報刊於論點上的異同之處等等，都值得提供該領域不同程度視野之研究可能性。

由初步的實驗結果中我們可看出，從大量文本中以統計方式篩選出的候選關鍵詞彙往往會較偏向於在數量較大量的資訊，儘管不同文件中相同出現詞彙及次數的加權來修正此問題(例如 TF-IDF），人工審閱的工作亦需耗費大量時間人力。因此我們除了需藉由史料文本該領域專家的知識協助以及修正以外，此統計分析及篩選，連同中文斷詞的工作亦是需要謹慎處理的部分，否則資訊的良莠不齊將會在後續逐步的處理程序中有可能被過度放大，甚至導致落入錯誤的結果解釋方向。圖 4 呈現的即為在「二二八事件臺灣本地新聞史料彙編」中與本省人相關論述中使用的名詞詞彙，使用數種重要詞彙的計算方法所進行的初步篩選，其中詞彙的最小單元是以斷詞後的最小詞性單元呈現（N-gram via POS），並以詞頻以及共現網路的方式，來同步比較一般統計方法與基於網路結構之節點重要性的結果。當詞彙較短的時候，方法之間的重要性排序差異其實並不大，但是當提升到兩個詞彙組成的詞組時，由圖 4 中各種重要性評估方法所呈現的詞彙重要性排序可以看出，詞頻所呈現的結果僅有排序較前者少數詞彙與其他方法一致，其餘結果則皆較偏離其他方法。

Frequency	Degree Centrality	Eigenvector Centrality	Betweenness Centrality	Closeness Centrality
黃帝子孫 (Nb)(Na) 10	黃帝子孫 (Nb)(Na) 29	黃帝子孫 (NbNa) 1	無能政府 (VHNa) 21	外省人責任 (NaNa) 5.5
祖國懷抱 (Nc)(Na) 5	祖國懷抱 (Nc)(Na) 24	無能政府 (VHNa) 0.88	祖國懷抱 (NcNa) 19	責任責任 (NaNa) 5.5
中華民國子孫 (Na)(Nc) 4	無能政府 (VH)(Na) 22	外省人煽惑 (NaVC) 0.87	黃帝子孫 (NbNa) 12	貪官汙吏人民 (NaNa) 5.5
處長重要 (Na)(VH) 4	思想運動 (Na)(Na) 19	人心離間 (NaVC) 0.87	政治腐敗 (NaVH) 8.1	責任責備 (NaVC) 5.5
同胞愛護 (Na)(VC) 4	運動背景 (Na)(Na) 19	同胞愛護 (NaVC) 0.87	臺灣中國 (NcNc) 6.7	痛苦國家 (VHNa) 5.5
責任中國人 (Na)(Na) 4	背景運動 (Na)(Na) 19	煽惑人心 (VCNa) 0.87	治安物價 (NaNa) 5.6	幸福個人 (VHNa) 5.5
省府委員 (Nc)(Na) 4	何等思想 (VH)(Na) 19	離間情感 (VCNa) 0.87	物價狂奔 (NaVA) 5.6	人民痛苦 (NaVH) 5.5
廿三原是 (Neu)(VG) 4	含有何等 (VJ)(VH) 19	情感本省 (NaNc) 0.87	狂奔高漲 (VAVH) 5.6	國家幸福 (NaVH) 5.5
外省人民法 (Na)(Na) 4	治安物價 (Na)(Na) 18	中華民國子孫 (NaNc) 0.87	高漲教育界 (VHNc) 5.6	責備貪官汙吏 (VCNa) 5.4
要求廢止 (VF)(VC) 4	物價狂奔 (Na)(VA) 18	中華民國同胞 (NcNa) 0.87	官營事業 (ANa) 4.8	政治腐敗 (NaVH) 4.5

圖 4　全報刊：本省人 相關詞彙初步統計結果(POS-Gram=2)

Degree Centrality	Eigenvector Centrality	Betweenness Centrality	Closeness Centrality
外省人(N)14	法制委員會委員(N)0.82	省警察大隊(N)46.0	完全省自治(V)2.93
市民館(N)7	死者優於予撫卹(V)0.70	各重要幹部(V)32.7	此次傷亡(V)2.83
各重要幹部(V)7	依法嚴辦(V)0.68	法制委員會委員(N)32.0	各處長(N)2.81
人民(N)6	儘量提出(V)0.65	即刻改組各級幹部(V)28.0	應儘量採用(V)2.07
市內空軍第三飛機廠(N)6	委員互選(V)0.65	陳長官已一面應允(V)21.8	半數以上(N)2.04
空軍第三飛機廠(N)6	各重要幹部(V)0.63	外省人(N)15.6	現電力公司全(V)2.04
各處長(N)6	半數以上(N)0.57	人民(N)14.8	臺南車站(N)2.04
法制委員會委員(N)6	傷者給以治療(V)0.54	各處長(N)14.1	一切公營事業主管人(V)2.04
公教人員(N)5	應檢舉轉請處理委員會協同憲警拘拿(V)0.53	追加通過(V)14.0	各地方法院首席檢察官全(N)2.04
傷者給以治療(V)5	要互相尊重(V)0.52	本省陸海空軍(N)14.0	本省陸海空軍(N)2.04

圖 5　中華日報：本省人 詞彙關聯式網路結構中心性分析

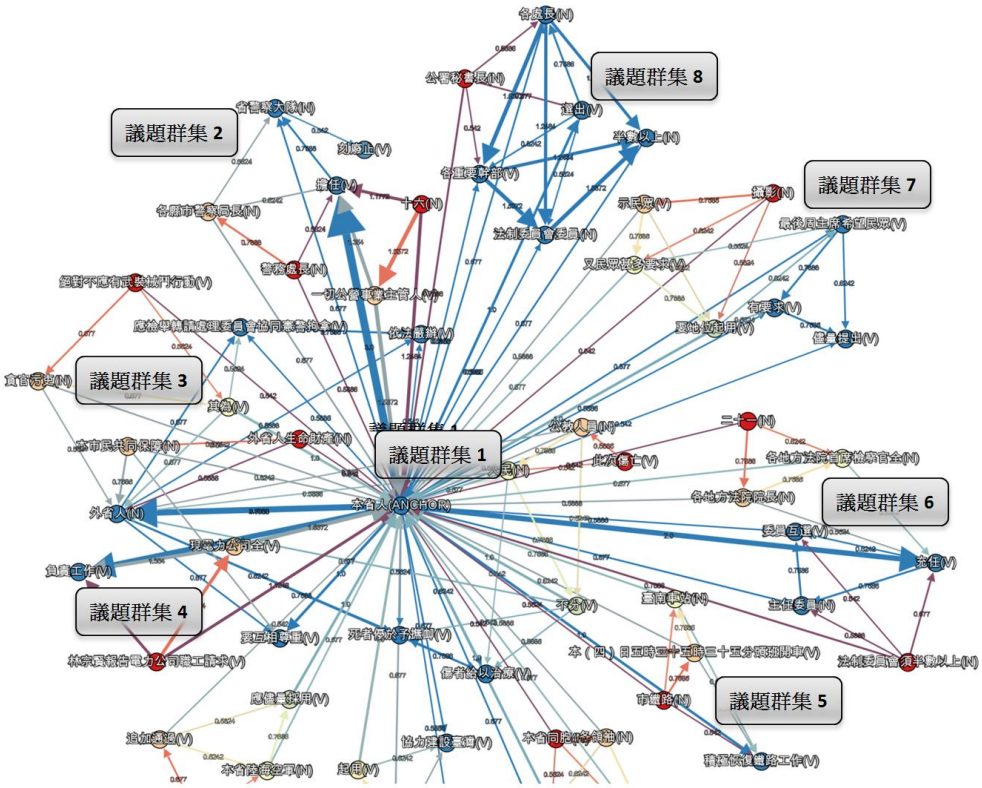

圖 6　中華日報中與本省人相關之論述情境重要詞彙關聯式網路
（中心性： Eigenvector Centrality ; 群集偵測：Modularity）

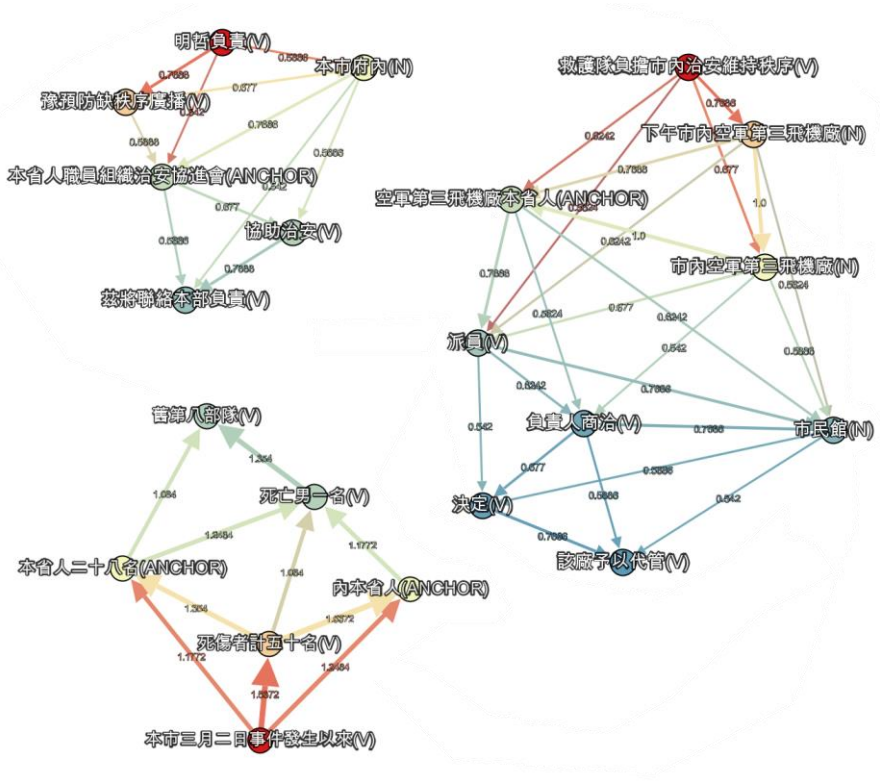

圖 7 中華日報：本省人 詞彙關聯式網路案例

而觀察以詞彙關聯式網路的實驗結果，可得知以「本省人」為關鍵詞彙時，區域性話題主要圍繞在與外省人以及其所掌控之下的相關政府機構為主，然而各分類話題之間重疊較少，密度亦較低，話題之間路徑也較遠，意味著論述的層面較為廣泛，較不聚焦在特定的話題上，且亦較多正面之呼籲及程序或法制相關的聲明，圖 5 呈現的案例則為中華日報中以「本省人」為關鍵詞彙時，相關聯的重要詞彙列表。圖 6 呈現的是在中華日報中以「本省人」為關鍵詞彙時，所相關議題之群集探勘結果，編號越小者代表話題的聚集性越高，內部所包含的子話題及關連也較為豐富。由群集所浮現的議題可看出，多半是屬於公營事業相關職位更動的程序、法治正面處理的呼籲、以及期望不要訴諸暴力的行為提倡等宣告；以本省人角度的議題群集而言，最大的集群是屬於多起武力衝突及社會事件的概述，並有部分政策上的呼籲，以及政府相關安撫的行為宣導；而次大的群集是屬於外省人、公務人員，憲兵警察於南門、臺北公園、榮町、車站本町、永樂町、太平町、等處被毆打的衝突事件相關的狀況描述；其餘亦為衝突事件相關的描述，唯圍繞的話題較少。圖 7 則亦是以中華日報為例，與本省人相關論述之重要詞彙所建立出來的詞彙關聯式網路案例。此外，我們也從多個報刊的詞彙關聯式網路實驗結果中發現，以「外省人」為關鍵詞彙時，區域性話題主要圍繞在衝突事件的發生種種描繪、武力衝突及傷亡的情況、以及相關的時間地點人員為主，並有相對少部分正向措施的呼籲，各分類話題重疊性較高，話題之間的群聚性較強也較有聚焦性。未來若以二二八事件之日期時間、報刊立場屬性、民間及官方論述型態此三種主要維度來交互產生及檢視關聯式網路的話，從其中質與量的差異及相同之處，可期望能藉由 PARTEX 平台的利基，系

統化地產生出更具有人文及歷史意義的觀察結果。

參考文獻

[1]. C. Williford & C. Henry, One Culture: Computationally Intensive Research in the Humanities and Social Sciences - A Report on the Experiences of First Respondents to the Digging Into Data Challenge. Council on Library and Information Resources.151. (2012). http://www.clir.org/pubs/reports/pub151

[2]. C. L. Liu, et al., Some chances and challenges in applying language technologies to historical studies in Chinese, International Journal of Computational Linguistics and Chinese Language Processing, 16(1-2), pp.27-46, 2011.

[3]. W. H. Cheng, et al., Ideas, events and actions: The digital humanity study of the concept formation in modern China, Proceedings of the 2014 International Conference on Digital Humanities (DH 2014), 000-000, pp.8-12, 2014.

[4]. D. Cohen, T. Hitchcock, G. Rockwell, et. al. Data mining with criminal intent. Final white paper. August 31, 2011. Available at http://criminalintent.org/ wp-content/uploads/2011/09/Data-Mining-with-Cirminal-Intent-Final1.pdf.

[5]. T. Hitchcock, R. Shoemaker, C. Emsley, S. Howard, J.McLaughlin, et al., The Old Bailey proceedings online, 1674-1913 version 7.0, 24 March 2012. Available at http://www.oldbaileyonline.org/