圖資與檔案學刊 16:1=104 (June '24) 38-68
DOI:10.6575/JILA.202406_(104).0002

基於文本分析的綜合關係網路建模
—以海外台灣左翼資料庫為例

吳怡潔* 薛化元**

【摘要】

本論文取材自左派運動領導者左雄主導的「台灣時代社」等數位典藏。該資料庫收錄台灣時代社 1970-2018 年間所出版的雜誌、內部通訊及成員間往來書信，共計文檔 498 件，內文 229 萬字。在全文數位化後，我們利用自然語言處理技術，進行命名實體識別、尋找關鍵詞語、統計重點詞彙頻率，並爬梳成員間的社會網路關係，分析結果以動態的資料視覺化呈現。如此綜合各種面向、關係的網路圖，讓研究者能探討各種關切因素，同時透過視覺化分析結果，更深入地瞭解其中的量化資訊，期盼本文成果可作為輔助社會科學領域質性與量化研究的便利工具。

關鍵詞

自然語言處理 深度學習 命名實體識別 綜合關係網路 資料視覺化

* 國立政治大學人工智慧跨域研究中心助理教授
ORCID 0000-0002-2973-5735
通訊作者 E-mail: matywu@nccu.edu.tw
** 國立政治大學台灣史研究所教授
ORCID 0000-0002-8189-8830
E-mail: hy5595@nccu.edu.tw

壹、前言

一、源起

國立政治大學圖書館自 2007 年起，致力於整理並蒐集台灣與中國近現代史各面向相關的研究資源，並於 2019 年成立特藏中心，除典藏方豪孝思堂、羅家倫、尉天驄、鄭世璠、史明、明清古籍及海外台灣人運動刊物等特色文庫外，亦積極徵集台灣研究史料資源，包括文史類的陳芳明、高明、何啟民；政治類的張君勱、民社黨、孫善豪；海外台灣人運動：徐雄彪、張維邦、台灣時代社（左雄）、台獨聯盟、艾琳達…等主題館藏，其中海外台灣人運動主題史料，涵蓋 1960-2018 年海外民主運動各種主張派別的珍貴史料。

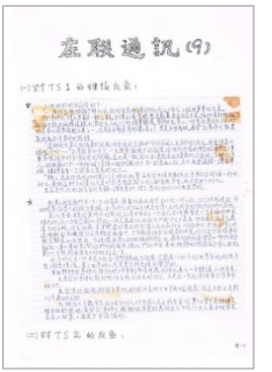
戰後海外台灣人政治運動派別林立，海外台灣左派以理論紮實及刊物深度著名，在「臺灣政治與社會發展海外史料資料庫」中，收藏了一系列包含《獨立台灣》等左翼立場出版品。其中，由左雄領導的「台灣時代社」為北美洲海外台灣左派運動的代表性團體，自《台灣人民》、《台灣革命》、乃至《台灣時代》，可說是圍繞在左雄的身上，為其運動發展過程的中心焦點（台灣時代，2020）。劉吉軒等學者（2013）曾以「臺灣政治與社會發展海外史料資料庫」收錄的 8 種海外左派刊物的論述型文章之全文文本進行分析，並以社會網路模型觀察海外左派刊物中的詞語變化，僅能以有限的公開文章內容推測當時這些左派的關注焦點及刊物間的關係。

二、史料介紹

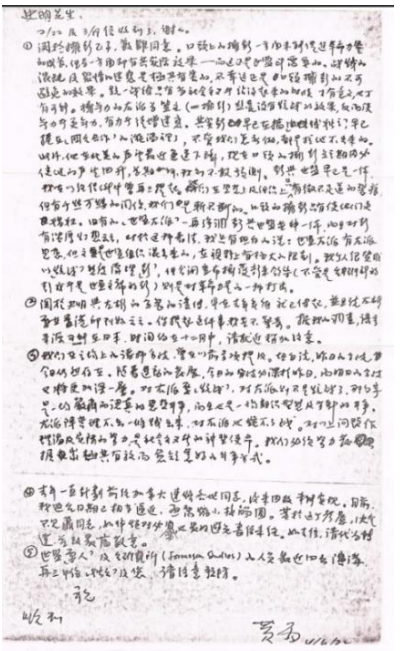
提到海外台灣左派思想者，以旅居日本的《台灣人四百年史》作者史明（本名施朝暉）較為知名。但北美洲台灣左派的代表人物左雄，自 1960 年代末期以來，陸續創辦《台灣人民》、《台灣革命》、《台灣時代》等刊物，發展左派組織。目前已知的文獻中，史明、艾琳達、張金策、鍾維達、胡民祥、林孝信等當時的運動參與者，皆提及左雄與《台灣時代》的影響力。左雄本名林重文，畢業於本校政大東方語文學系俄語組，左雄是他所有筆名中，最為人所知的（政大數位典藏，2019）。當時的社會主義者，在戒嚴時期的台灣沒有立足之地，即使旅居海外也經常遭情治單位監控。左雄

仍持續進行台灣社會主義者的串連、組織及理論宣傳。

(a)



(b)



(c)

圖 1　以左雄為中心的史料。(a)公開刊物《台灣時代》(b)內部流通文章《左聯通訊》(c)左雄以化名寫給史明的信。

資料來源：國立政治大學圖書館特藏管理組，2020。

政大圖書館自 2019 年起陸續入藏「台灣時代社史料」，包括紙質文件和電子郵件，涵蓋 1970 年到 2018 年台灣時代社在北美洲台灣人政治運動的發展歷程，文本類型囊括私人信件、內部通訊、乃至已發表的雜誌/期刊，截至 2020 年，共計完成 498 份文件，約 2,300,000 字之數位文本。透過「台灣時代社史料」的分析，可以展現當時的台灣左派團體的思想路線、統一戰線、人際網路、及具體實踐的過程。圖 1 展示這位領導者左雄的刊物《台灣時代》封面、半公開寄送的通訊文章《左聯通訊》、以及與當時旅居日本的史明的部分信件內容。

文史典藏包含從該主體延伸的相關人、事、時、地、物等訊息，就使用角度而言，理解史料的切入點也將能從點的層次，進展到線、甚至面的格局，這樣的整合與規劃，可以因應從巨觀到微觀的不同研究需求。本研究利用自然語言處理技術，進行命名實體識別、尋找關鍵詞語、統計重點詞彙頻率，並爬梳成員間的社會網路關係，分析結果以動態的資料視覺化呈現。如此綜合各種面向、關係的網路圖，讓研究者能探討各種關切因素，同時透過視覺化分析結果，更深入地瞭解其中的量化資訊，可作為輔助社會科學領域質性與量化研究的便利工具。

本論文架構如下：第二節探討自然語言處理、文本理解以及資料視覺化設計等相關文獻。第三節闡述本研究系統架構設計。第四節則就左雄文本之分析結果以視覺化方式呈現，並進行討論。最後第五節則為本文結論。

貳、相關研究

本研究旨為建立自動化文本分析流程，並以資料視覺化的形式，動態呈現分析結果。本節將從兩大面向進行文獻探討：(1)自然語言處理、文本理解，(2)視覺化設計。所得結論亦為本研究實作設計的基礎。

一、自然語言處理與理解

一般而言，在自然語言處理的流程中，首先是將連續文本拆解成單獨意義的語詞，之後方能計算前後語詞之間的關係、或是特定語詞的重要程度，故斷詞結果是後續文本理解正確與否的關鍵因素。然而，不同於歐美

語系，中文除標點外，字裡行間並不以空白分隔，故在斷詞階段，先天就存在一定的門檻。拜現今深度學習領域的轉換器（transformer）（Vaswani et al., 2017）架構之賜，基於其注意力機制，自然語言處理技術在各種文本任務中，都能夠展現更佳效果。

接下來的階段，便是開始理解一篇文本中的概念。根據 Schütze、Manning 與 Raghavan（2008）的書中指出，計算語詞頻率（Term Frequency, TF）是一個簡單、直覺的作法。由於我們目的是為了掌握核心人物的論述重點及強度，故我們就不進行額外加權（如 Inverse Document Frequency, IDF）。

此外，近年來，命名實體識別（Named Entity Recognition, NER）也是自動資訊擷取領域中，相當重要的一項技術。命名實體的概念首見於第六屆訊息理解會議（Message Understanding Conferences），其中所建立的一項子任務即是命名實體的標記與辨識（Grishman & Sundheim, 1996）。在 2010 年底，由語言資訊聯盟（Linguistic Data Consortium）釋出的語料庫 OntoNotes Release 4.0 中，更擴大命名實體定義的範圍及分類（Weischedel *et al.*, 2011）。其定義如下：

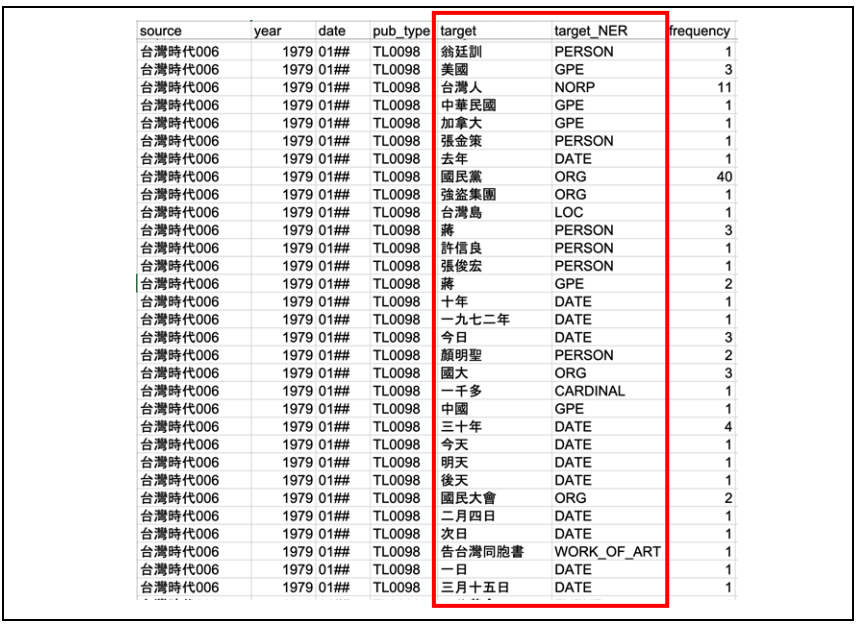
（一）命名實體類，包含：人物（person）、國籍（norp）、設施（facility）、組織（organization）、國家（gpe）、地點（location）、產品（product）、事件（event）、作品（work of art）、法律（law）、語言（language）等類別。

（二）數值類，包含：日期（date）、時間（time）、百分比（percent）、貨幣（money）、量度值（quantity）、序數（ordinal）、其他數字（cardinal）等類別。

此後，命名實體識別（NER），便成為自然語言處理領域中的一個重要課題，許多技術皆基於上述標記的語料庫發展至今。由於 NER 任務目標為解析出文本中的人、事、時、地、物等關鍵資訊，若採用精確度高的 NER 模型，將有助於自動化文本分析系統的建立。

為了同時處理繁體中文的斷詞及命名實體識別，我們採用中研院詞庫小組基於 BERT（Devlin, Chang, Lee, & Toutanova, 2019）架構訓練的模型，進行斷詞（Yang, 2021a）及 NER（Yang, 2021b）分析。圖 2 展示在《台灣時代》一篇文章完成命名實體識別分析的範例，框線部分即是辨識結果的實體字詞及其類別。我們觀察到 NER 的關鍵字詞結果，雖然結果相較詞

頻（TF）分析已精簡許多，因為包含數值分類的詞語（今天、明天、後天），在後續過程中，我們亦採用了列表過濾詞/停止詞，針對 NER 的結果進行後處理。



source	year	date	pub_type	target	target_NER	frequency
台灣時代006	1979	01##	TL0098	翁廷訓	PERSON	1
台灣時代006	1979	01##	TL0098	美國	GPE	3
台灣時代006	1979	01##	TL0098	台灣人	NORP	11
台灣時代006	1979	01##	TL0098	中華民國	GPE	1
台灣時代006	1979	01##	TL0098	加拿大	GPE	1
台灣時代006	1979	01##	TL0098	張金策	PERSON	1
台灣時代006	1979	01##	TL0098	去年	DATE	1
台灣時代006	1979	01##	TL0098	國民黨	ORG	40
台灣時代006	1979	01##	TL0098	強盜集團	ORG	1
台灣時代006	1979	01##	TL0098	台灣島	LOC	1
台灣時代006	1979	01##	TL0098	蔣	PERSON	3
台灣時代006	1979	01##	TL0098	許信良	PERSON	1
台灣時代006	1979	01##	TL0098	張俊宏	PERSON	1
台灣時代006	1979	01##	TL0098	蔣	GPE	2
台灣時代006	1979	01##	TL0098	十年	DATE	1
台灣時代006	1979	01##	TL0098	一九七二年	DATE	1
台灣時代006	1979	01##	TL0098	今日	DATE	3
台灣時代006	1979	01##	TL0098	顏明聖	PERSON	2
台灣時代006	1979	01##	TL0098	國大	ORG	3
台灣時代006	1979	01##	TL0098	一千多	CARDINAL	1
台灣時代006	1979	01##	TL0098	中國	GPE	1
台灣時代006	1979	01##	TL0098	三十年	DATE	4
台灣時代006	1979	01##	TL0098	今天	DATE	1
台灣時代006	1979	01##	TL0098	明天	DATE	1
台灣時代006	1979	01##	TL0098	後天	DATE	1
台灣時代006	1979	01##	TL0098	國民大會	ORG	2
台灣時代006	1979	01##	TL0098	二月四日	DATE	1
台灣時代006	1979	01##	TL0098	次日	DATE	1
台灣時代006	1979	01##	TL0098	告台灣同胞書	WORK_OF_ART	1
台灣時代006	1979	01##	TL0098	一日	DATE	1
台灣時代006	1979	01##	TL0098	三月十五日	DATE	1

圖 2　NER 分析用於《台灣時代》文章之範例，框線中為部分分析結果。

二、資料視覺化設計

學者 Card、Mackinlay 與 Shneiderman（1999）認為，視覺工具能幫助人類思考抽象議題；將資訊視覺化，即是藉由互動式視覺資料的呈現，讓人們能夠強化認知、深入理解，以進行決策或解釋。其中，標籤雲（TagCloud）用來突顯特定語詞的重要性或出現頻率，是最為直觀的。基於此設計，學者 Cao 與 Cui（2016）衍生出隨時間變化的動態詞雲（dynamic word clouds）設計，使用者與其互動，可觀察文本主題在不同時間點的演變。

Cao 與 Cui 同時也指出，基於網路圖形架構可以設計多面向的關係模型，所構成的資料基本元素包含實體（entities）、層面（facets）、關係（relations）、集群（clusters）以及時間趨勢（temporal trend）等，由此衍生的視覺化設計，能夠表達更複雜的資料分析的內外部關係，為操作者帶來以簡馭繁的互動體驗（Cao & Cui, 2016）。我們將考量這些原則，設計本研究的綜合關係網路互動工具。

參、系統架構設計

本研究框定的資料範圍多達上百萬字，橫跨時期將近 50 年，文本內容囊括公開發行的刊物、到私人信件，涉及不同面向的研究主題。圖 3 簡述本研究從資料整理與數位化、自動前處理與分析、到結果視覺化的流程架構。

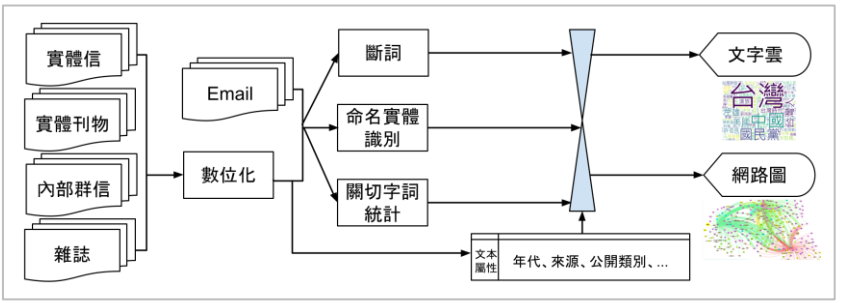


圖 3　本研究系統架構圖

為了同時提供 AI 模型、視覺化網頁介面等功能，本系統建置於配備 Nvidia GeForce RTX 3090 顯示硬體之 UBUNTU（20.04）之平台上，網站則採用 XAMPP 軟體架設（XAMPP, n.d.）。研究者可自行上傳文本後，執行指定的 AI 分析。更多套件資訊、及功能設計細節，請詳見以下各子節。

一、關鍵資訊萃取

基於前節所闡述的標籤雲設計再延伸，本研究採用 Python 中的

WordCloud 套件（Mueller, 2017）以呈現關鍵主題。圖 4 為使用文字雲分別繪製語詞頻率（TF）和命名實體識別（NER）的分析結果。我們可以觀察到，這兩者呈現出相同文本中不同的關注重點。

由於圖 4-(a)以語詞使用量為主要量化特徵，在未設定過濾詞／停止詞的前提下，文本中的代名詞（我們、他們）、助動詞（可以、必須）、或連詞（而且、如果）等詞性較容易在圖裡出現。欲瞭解作者的寫作風格，TF 或可作為一種特徵，然而作為關鍵資訊，由於個別研究存在不同需求，即便設定了過濾詞/停止詞，輔以計算 TF-IDF 等特徵，則恐怕仍然內含過多不必要的雜訊。圖 4-(b)則主要是命名實體的統計量，我們可看到圖中強調了更多人名、國名以及組織。此外，我們考量到通用性，研究者亦可設定研究重點關注字詞，我們在後續的「關切字詞表」小節中，提供較完整的方法闡述。

(a) (b)

圖 4　不同分析方法之文字雲比較。(a) TF (Yang, 2021a)；(b) NER (Yang, 2021b) 。

二、文本屬性定義

接下來，以左雄稿件資料為例，我們嘗試歸納文本中可能出現的元素，條列如下：

（一）實體信、email：寄信者、收信者、送信時間、信件內容（以下統一為文本內容）。

（二）實體刊物、雜誌：刊物/雜誌名稱、出版日期、文本內容。

（三）內部群信：寫作日期、文本內容、刊物名（部分仍為草稿狀態，我們以「左雄文稿」標示）。

對於文本內容，我們分別進行命名實體識別，以及統計關切字詞頻率。由於 NER 可取出多達十餘種實體分類，我們主要列出以下 5 種實體：(1)人物（person）；(2)組織（organization）；(3)團體（norp）；(4)地方（gpe）；(5)作品（work of art），作為繪製網路圖的基礎。

三、綜合關係網路設計

以左雄給史明的實體信為例，談論內容可能涉及人物、組織或地方。我們以這些實體元素作為節點（nodes），並將不同實體分類以不同顏色區隔。字詞出現的次數，則為談論者與談論客體（字詞）之間的關係，愈多則以愈粗的線條顯示。圖 5 展示本研究所建置的綜合關係網路圖的大致輪廓。

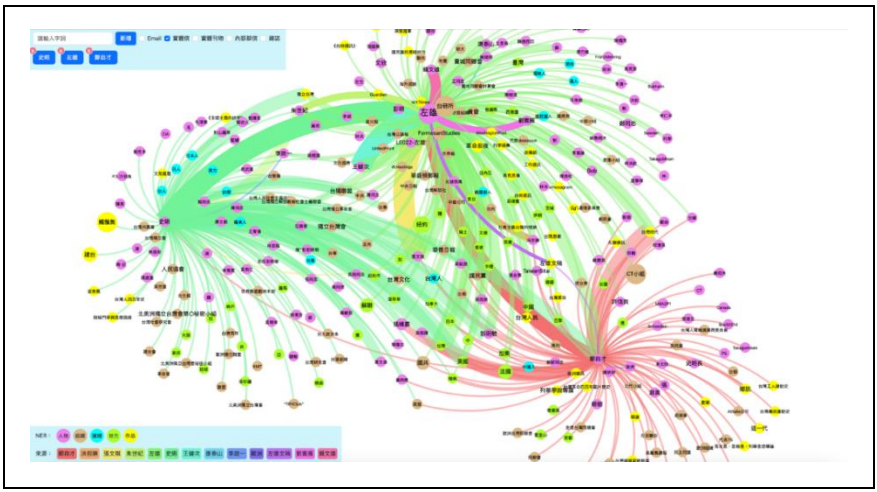


圖 5　綜合關係網路互動圖，輸入搜尋字詞為：「左雄」、「史明」、「鄭自才」。

圖 5 中左上角，呈現過濾資料的條件，除了依照文本類型調整可視資

料範圍外，另外提供輸入搜尋字串的介面，可作為集群中心繪製。以圖 5 為例，若研究者想一覽文本中左雄與僑居日本的左派領導者史明，以及以刺蔣案聞名獲釋後旅居海外的鄭自才，他們之間的互動或言談中相互提及的關係樣態，研究者可在介面中輸入這三位人物的名字，便可看到以此三個節點為中心展開的集群。

此外，若想深入瞭解意見來源（談論者、發表刊物）有多頻繁地針對特定字詞的細節，可將滑鼠指標移至邊（edge）的區域，即可呈現意見來源各在哪些文本上提及此字詞及頻率，範例於圖 6 所展示。

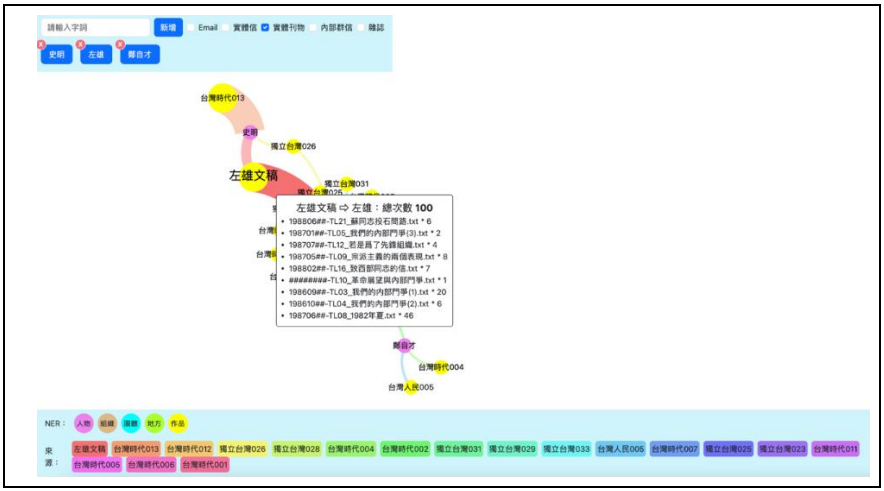


圖 6　綜合關係網路互動圖操作範例：深入瞭解「左雄」被談及在資料庫中所據文本的細節。

四、關切字詞表

由於 NER 著重於實體詞語，對於概念詞語（如：「階級」、「主義」等）並不敏感；而純粹基於斷詞結果，恐無法聚焦於目標人物文本中之左翼相關辭彙演變。故我們整理關切字詞表（關切字詞表如附錄中表 1 所示），並計算關切詞在各文本中出現的頻率。

在此，我們展示一個外來語翻譯範例：以「馬克思主義」和「馬克斯

主義」兩個字詞進行搜尋，此兩字詞則是源自「Marxism」的繁體中文翻譯，如圖 7 所示。有趣的是，我們可觀察到不同意見來源，對於特定用詞確有偏好或呈現出作者的習慣；即使我們知道兩詞指涉的是同一概念。在圖 7 中，「馬克斯主義」主要在公開刊物中較頻繁採用，如《台灣人民》、《台灣革命》、《台灣時代》等雜誌；而「馬克思主義」則使用於大多數的內部通訊、書信或 Email。即使是左雄未公開的文稿中，也呈現出兩種翻譯語偏好使用「馬克思主義」。

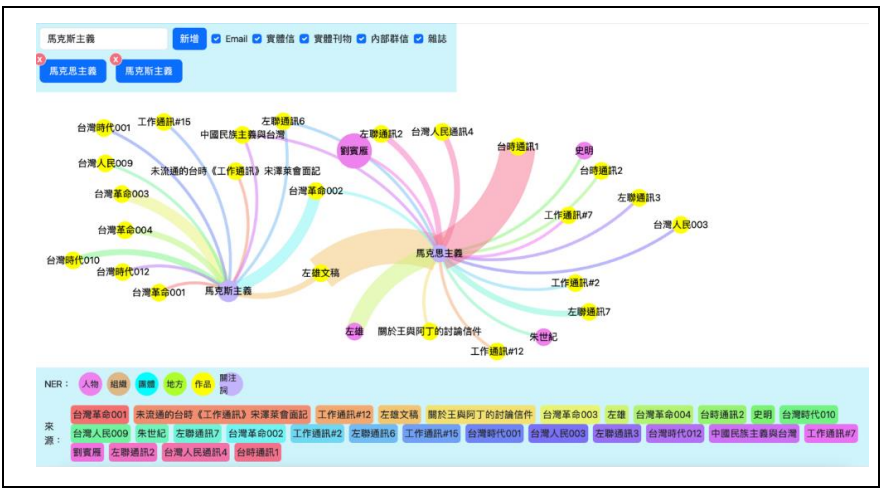


圖 7　綜合關係網路互動圖，使用關切字詞表的分析範例：「馬克思主義」和「馬克斯主義」的比較。

為了瞭解外來語的選擇及演變，我們使用 Google 趨勢（Google, n.d.），比較從 2004 年到現今，全球網路使用者在 Google 網頁上搜尋「馬克思主義」和「馬克斯主義」的數量差異，結果如圖 8 所示。在中文世界中，使用「馬克思主義」搜尋的數量無疑是占大多數的，而這亦反應在左雄私下寫作、內部流通、討論的寫作偏好。雖尚不明白為何左雄在公開刊物上選擇用「馬克斯主義」推動左派理論教育的原由，但即便到了 1970、1980 年代，無論用哪個翻譯詞彙都是國民黨當局的禁忌，也讓我們不禁遙想那個台灣人民動輒因左派刊物、讀書會等理由即可能入獄的戒嚴時期。

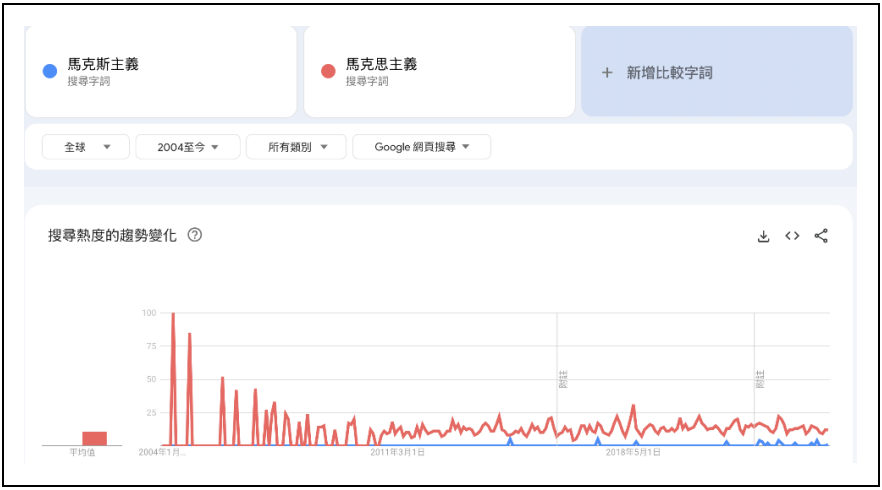


圖 8　使用 Google 趨勢，比較全球搜尋「馬克斯主義」和「馬克思主義」的頻率。（圖 8 中的上方線條代表「馬克思主義」；下方線條則代表「馬克斯主義」。）

資料來源：Google trends (Google, n.d.) 。

肆、分析與討論

在本節中，我們主要以文字雲和綜合關係網路這兩種視覺化工具，以互動方式呈現分析結果。

一、詞語重要度分析

首先我們展示 NER 分析結果，圖 4-(b)中已呈現 NER 整體樣貌。

（一）不同年代演變

圖 9 呈現不同年代的 NER 軌跡。1980 年之前，文本內容著重於個人和出版組織；在 1980 年到 1990 年之間，同時關注國內和國際議題；而在 1990 年之後，則較為關心國際形勢和歷史問題。

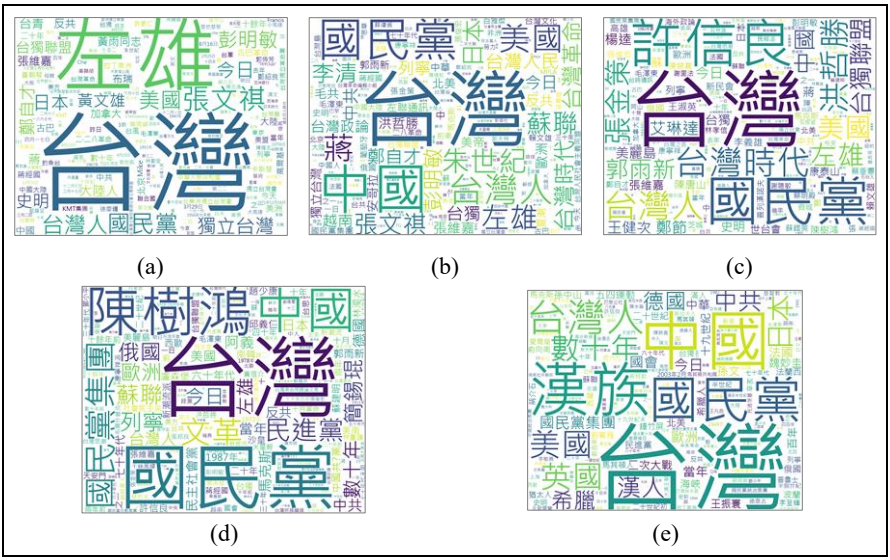


圖 9　不同年代的演變，基於 NER 分析結果。(a)1971 年之前；(b) 1972~1978 年；(c)1979~1987 年；(d)1988~1992 年；(e)1993 年之後。

（二）不同文本差異

圖 10 為區分文本類型的 NER 結果。我們可大致歸納，書信中出現最多人名，隨著文本的公開程度而提昇了國際相關詞彙的廣度和強度。在公開文章中，左雄更頻繁地論述歷史脈絡以及評論國際局勢。

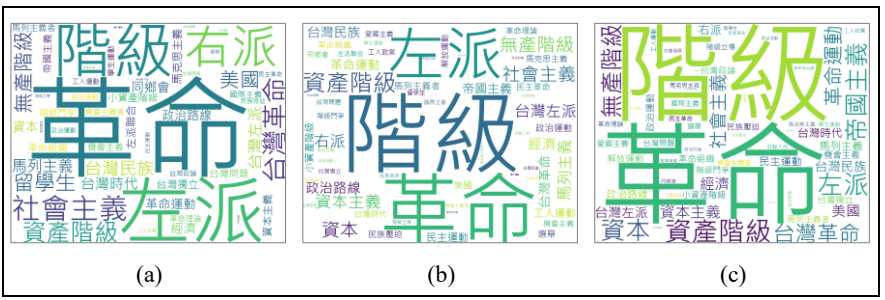


圖 10　不同文本類型，基於 NER 分析結果。(a)書信；(b)內部通訊；(c)公開文章。

二、關切詞語頻率分析

接下來是關切字詞表的詞頻分析，整體詞頻樣貌如圖 11 所示。

圖 11　基於關切字詞表之整體詞語頻率分析

（一）不同年代演變

圖 12 呈現不同年代的左翼用語的比例變化。我們可以發現圖 12-(a)(d)特別強調「階級」與「革命」。在 1971 年之前，肇因於相對少的文本量（42 件），加上國際間相繼發生的武裝革命、戰爭以及中華民國失去聯合國中國代表權等事件影響，論述「階級」敵友以實踐「革命」，是此時期當然的兩大命題。第二次的強調，出現在 1987 年台灣解除戒嚴令及蔣經國逝世的 1988 年之後，同時，國際間 1989 年也發生天安門事變、天鵝絨革命等國際大事，台灣國內隨著自由化的推動，政治路線分歧、人心思變，然而此時期的左雄文稿數量（29 件）卻比 1971 年之前更少。咸認為此為左雄重新檢討路線和批判對象的一個重要時期。

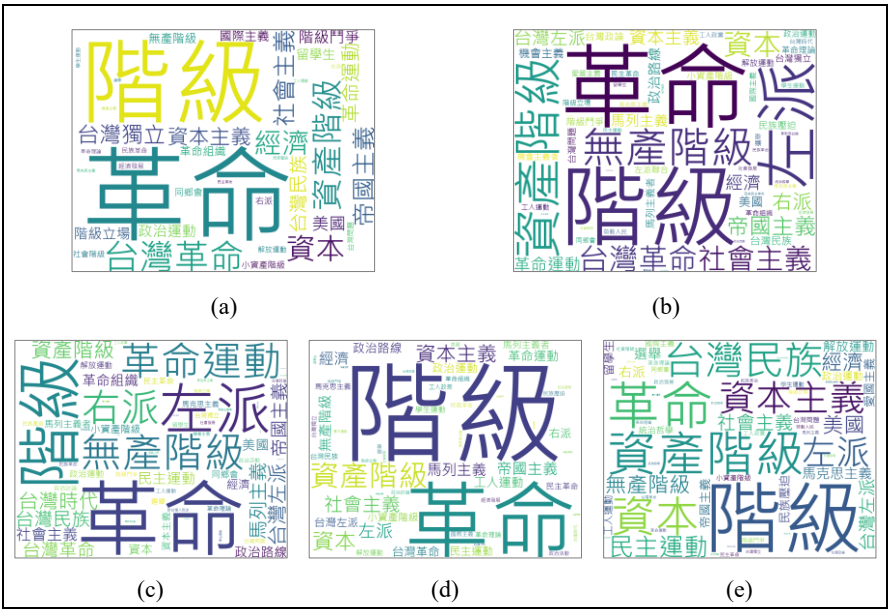


圖 12　不同年代的演變，基於關切字詞表之統計。(a)1971 年之前；(b)1972~1978 年；(c)1979~1987 年；(d)1988~1992 年；(e)1993 年之後。

（二）不同文本差異

圖 13 則是在區分不同文本類型後的分析結果。有趣的是，我們發現相較於內部通訊與公開文章，在書信文本中更頻繁地談及「右派」、「左派」。而相較於書信與內部通訊，「帝國主義」提及比例明顯較高（雖遠遠不及「階級」和「革命」）。另外，「資產階級」和「無產階級」在三類文本中皆名列前茅、且比例相似，或可將「階級」、「革命」和延伸概念「資產階級」、「無產階級」等語詞，作為左雄手稿的標幟、特徵。

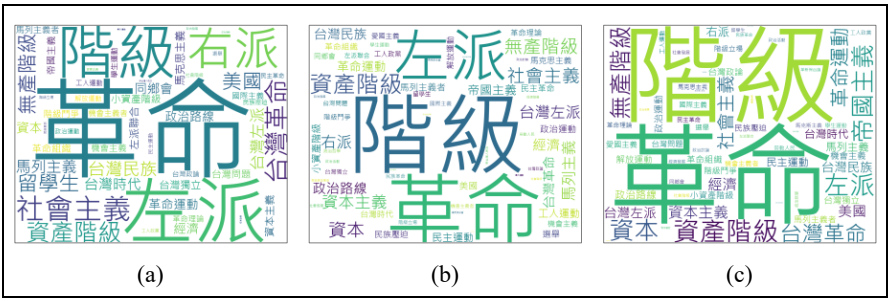


圖 13　不同文本類型，基於關切字詞表之統計。 (a)書信；(b)內部通訊；(c)公開文章。

三、綜合關係網路分析—基於 NER 結果

首先是基於 NER 分析的視覺化結果。為了更清楚的比較不同操作變因之間的變化，資料僅保留排名前 20 的詞語結果，並限制只顯示人物、組織。

（一）不同年代演變

圖 14 為不同時期的 NER 視覺化呈現。其中以圖 14-(b) 與 14-(c) 在結構上相較其他時期複雜，由於目標語詞已刪至前 20 名字詞，但(b) (c)兩圖卻有較多連結，顯示在此兩時期左雄的著述數量、或與其他人的書信互動，較其他時期為頻繁。

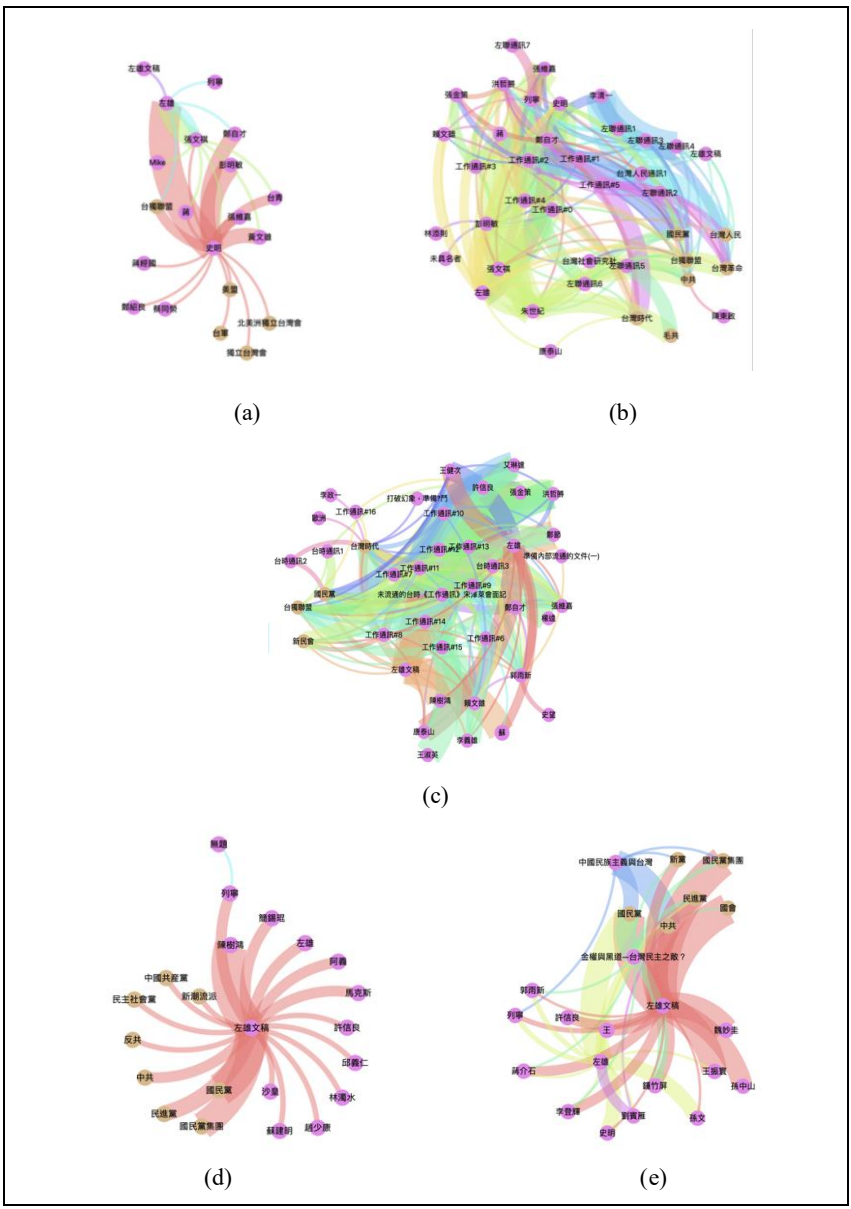
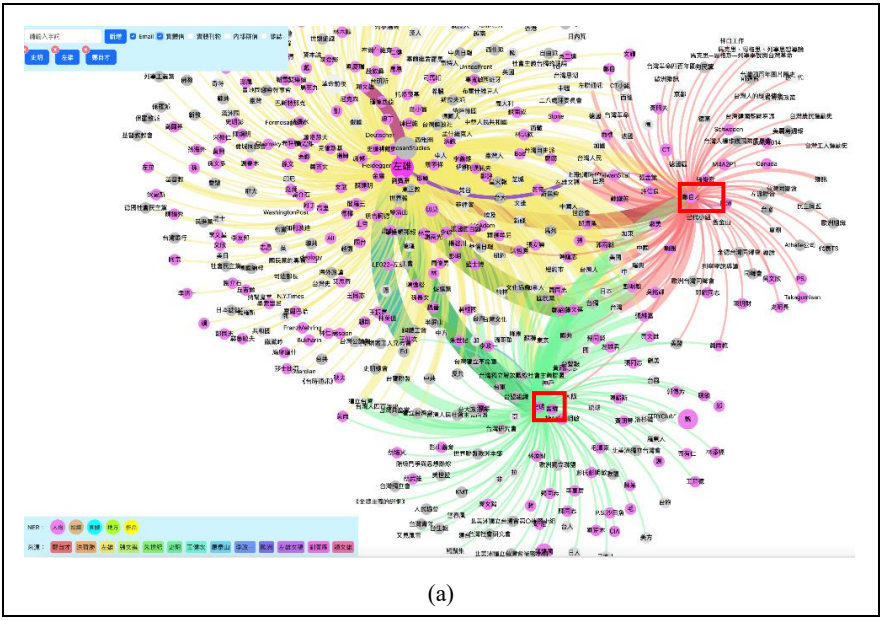


圖 14　不同年代的演變，基於 NER 分析。(a)1971 年之前；(b)1972~1978 年；(c)1979~1987 年；(d)1988~1992 年；(e)1993 年之後。

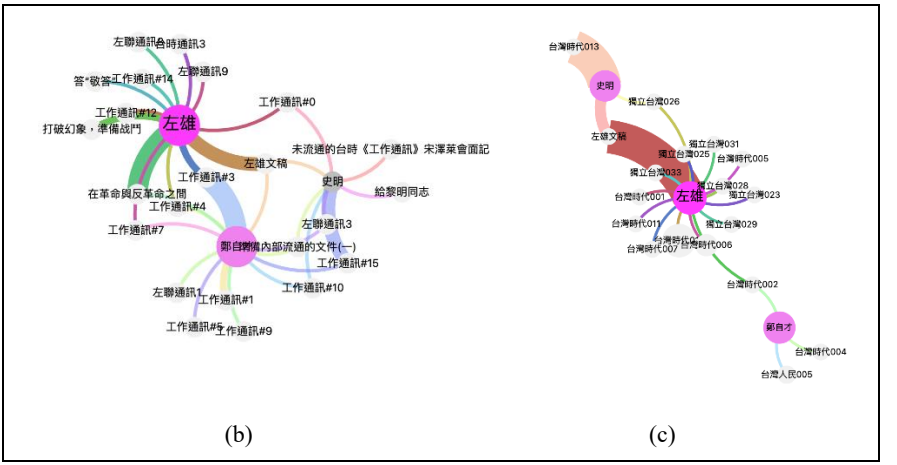
（二）不同文本差異

在這裡，我們想要探討，在固定的談論客體條件之下，不同文本類型之間會如何呈現各自的綜合網路關係樣貌？對應實際情境又如何解讀其意義？我們沿用前述圖 5 的搜尋詞：「左雄」、「史明」、「鄭自才」，並分別查看書信、內部通訊和公開文章之視覺化樣貌，如圖 15 所示。

由於左雄為大多文本的執筆者、以及部分文本的收件者，故我們更加關心「史明」和「鄭自才」與左雄的互動、或被談論的強度。圖 15-(a)即可觀察到史明和左雄的互動與鄭自才和左雄的互動相比，明顯是更加頻繁（基於邊的粗細程度）。此外，圖 15-(b)也很容易地從「工作通訊#3」以及「工作通訊#15」裡面，分別大篇幅地提及了「鄭自才」和「史明」；同樣的，圖 15-(c)則呈現了「台灣時代 013」大量的談論「史明」的情況，但只有三篇公開文章少量提及「鄭自才」。從書信、內部通訊和公開文章的綜合網路視覺圖，或可一窺左雄針對不同文本類型的定位。



(a)



(b)　(c)

圖 15　不同文本的類型，基於 NER 分析，並輸入搜尋字詞：「左雄」、「史明」、「鄭自才」。(a)書信；(b)內部通訊；(c)公開文章。

四、綜合關係網路分析—基於關切詞語

再來是關切字詞表詞頻統計的視覺化結果。我們延續上一節的分析架構，同樣探討不同時期之間、以及不同文本類型之間的綜合關係網路圖。

（一）不同年代演變

不同年代的關切詞語演變，如圖 16 所示。我們可觀察到圖 16-(a)與 16-(b)的著述量、以及每文本所提及的關切詞語量，都較其他時期更多。

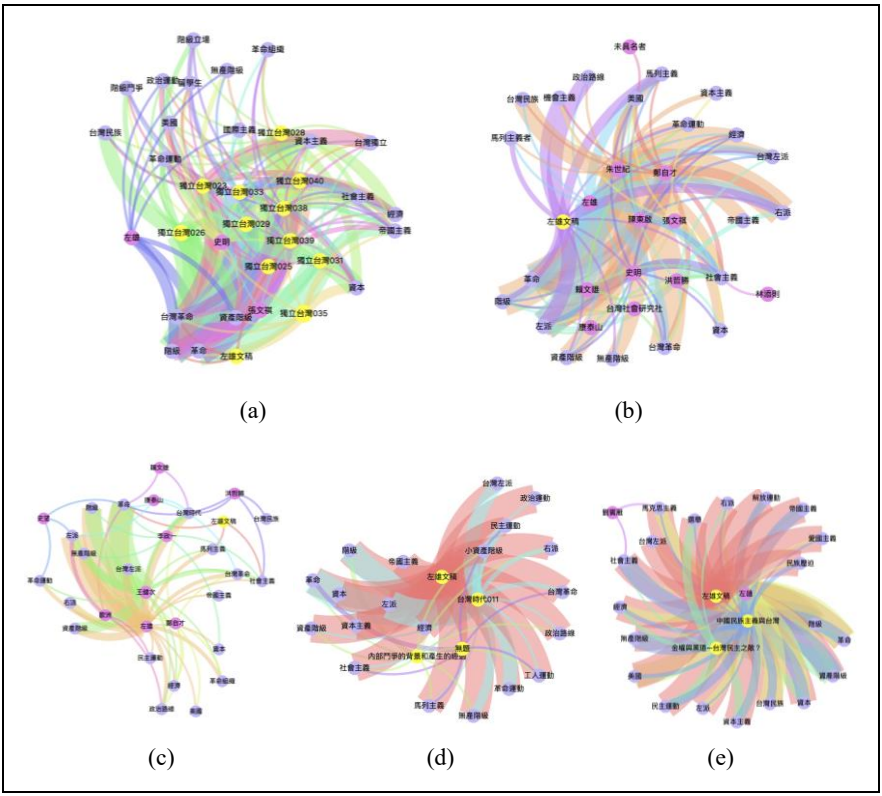


圖 16　不同時期的關係網路，基於關切字詞表之統計。(a) 1971 年之前；(b) 1972~1978 年；(c)1979~1987 年；(d)1988~1992 年；(e)1993 年之後。

綜合圖 14、圖 16 的視覺化結果，兩者結果或能相互補充。以 1971 年之前為例，文本中提及的人物、組織並不多，但該時期著述量於 5 個時期中卻相對較多；簡言之，此時期左雄的文稿主要傳達概念，與實體互動或提及的關係，則較多出現於書信中。

在圖 17-(a)的書信文本中，我們可以看到沒有一處提及「馬克斯主義」，然而在沒有寄件對象的 Email 中，卻大量使用了「馬克思主義」。在圖 17-(b)的內部通訊中，則開始出現了小量使用「馬克斯主義」的現象，然而仍能在多處內部通訊文章看到頻繁地討論「馬克思主義」，尤其在「左雄文

稿」和「台時通訊」之中。在圖 17-(c)的公開文章中，則更大量的使用了「馬克斯主義」，並主要分布在《台灣革命》、《台灣時代》兩本刊物中。

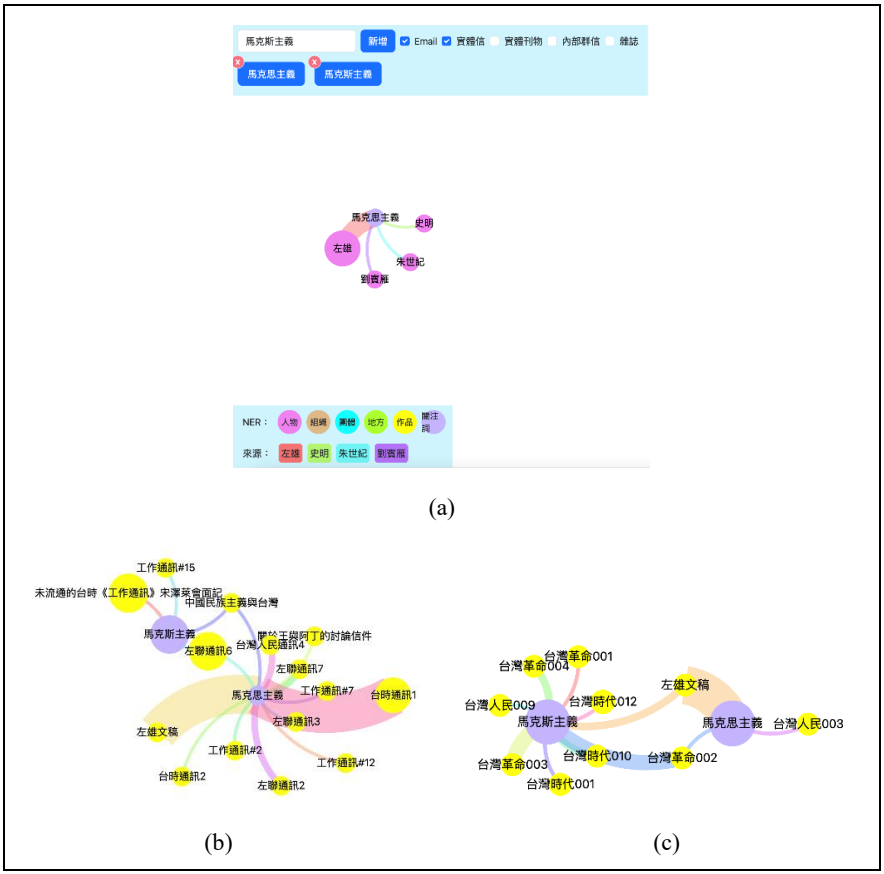


圖 17　不同文本的類型，基於關切字詞表之統計，並輸入搜尋字詞：「馬克思主義」、「馬克斯主義」。(a)書信；(b)內部通訊；(c)公開文章。

由於用哪個翻譯詞在當時都可能因為左傾思想而導致入獄，我們認為，使用不常用的，尤其是左雄幾乎很少用的「馬克斯主義」的語詞頻率，昭示了可能存在或多或少的共同作者，或不同的稿件來源。雖然，不管是內部通訊或公開文章，所有文字內容都可視為總編輯左雄的意志展現，但以

內部通訊內容研判，仍需要經過內部的文字討論，方能形成一期內容；而以實體刊物而言，則需要四至五篇文稿，方能出刊。故圖 17 的「馬克斯主義」頻率結果，或可解讀為內部通訊可能有少量複數作者的共同參與，而公開刊物，則存在較多不同寫稿作者。然而，考量到內部通訊、公開刊物在當時的編輯出版環境，仍不排除在過程中，因人工作業或設備限制，而出現此類的用字偏好落差。

伍、結論

本研究基於政治大學特藏中心所蒐集、數位化及保存的台灣民主運動之重要史料，並運用自然語言處理技術，以建立數位人文領域的研究工具。

實際案例研究則鎖定海外左翼領導者左雄（林重文）先生已數位典藏後的著述，我們主要採用命名實體識別（NER），以及搜尋所列舉之關切字詞表，作為兩大關鍵詞語來源。再根據這兩類的分析結果，分別以文字雲、綜合關係網路兩種視覺化方式呈現。除了其著述的整體輪廓以外，我們更進一步探索不同時期、不同類型的文本其中關鍵字詞組成比例，以及談論主體（公開刊物、內部信件、私人信件發信者）及談論客體（文本中之人、組織、地方等命名實體）之間的網路關係。

從分析結果研判，普遍用於文本分析的文字雲，較能夠呈現關鍵詞語頻率比重，卻無法呈現關鍵詞語的文本來源。本研究提出的綜合關係網路，除了能夠描繪語詞的文本來源，更以直覺、有效率的方式展現其重要性（以連結邊之粗細表示）。至於，命名實體詞或關切字詞，何者更為「關鍵」？我們認為兩者分別滿足了研究的不同需求。若想要抓取人事時地物的特徵，可採用 NER 分析；而若想探索特定主張用語（例如左翼文獻常用詞彙）以檢視其比重，或發表於哪些文本類型，直接列舉成表在實作上更方便直覺。

隨著大型語言模型的最新進展，我們預期分析工作將能夠更加自動化和精確，於未來研究中，將導入最新的 AI 技術，設計適用於追尋脈胳訊息的視覺化工具，以符合人文領域研究者的需求。

（接受日期：2024 年 7 月 9 日）

致謝

本研究特此感謝國立政治大學圖書館提供文本資料，廖文宏館長、張惠真編審在資料處理上的鼎力協助，政大資管所林子紘碩士生的網站技術協作，以及國家科學及技術委員會經費資助，國科會計畫【左雄與海外台灣左派運動：以《台灣時代》為中心】，編號：[MOST109-2410-H-004-150-MY2]。

參考文獻

台灣時代（2020）。*台灣時代*。檢自 https://da.lib.nccu.edu.tw/tdm_2020/content/publish.html

政大數位典藏（2019）。*左雄與台灣時代社數位史料庫*。檢自 https://cdm20070.contentdm.oclc.org/digital/collection/TL

國立政治大學圖書館特藏管理組（2020）。*臺灣政治與社會發展海外史料資料庫* [部落格文章]。檢自 https://da.lib.nccu.edu.tw/tdm_2020/index.html

劉吉軒、柯雲娥、張惠真、譚修雯、黃瑞期、甯格致（2013）。以文本分析呈現臺灣海外政治思想輪廓。在項潔編，*數位人文要義：尋找類型與軌跡*（頁 83-114）。臺北：國立臺灣大學出版中心。

Cao, N., & Cui, W. (2016). *Introduction to text visualization (Vol. 1)*. Paris: Atlantis Press. doi:10.2991/978-94-6239-186-4

Card, S. K., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: Using vision to think*. San Francisco, Calif: Morgan Kaufmann.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1,* pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423

Google. (n.d.). *Google trends*. Retrieved from https://trends.google.com/trends/

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief

history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics* (pp. 466-471). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/992628.992709

Mueller, A. (2017). *Word_cloud: A little word cloud generator in python*. Retrieved from https://github.com/amueller/word_cloud

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California. doi:10.48550/arXiv.1706.03762

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., ... & Houston, A. (2011). *Ontonotes release 4.0. LDC2011T03*. Philadelphia, Penn.: Linguistic Data Consortium, 17.

XAMPP. (n.d.). *Apache friends*. Retrieved from https://www.apachefriends.org/download.html

Yang, M. (2021a). *Ckiplab/bert-base-Chinese-ws.* Retrieved from https://huggingface.co/ckiplab/bert-base-chinese-ws

Yang, M. (2021b). *Ckiplab/bert-base-Chinese-ner*. Retrieved from https://huggingface.co/ckiplab/bert-base-chinese-ner

附錄

表 1

關切字詞表：左翼文獻常用詞彙

二～三字組	四字組			五～七字組
右派	工人政黨	社會發展	庸屬發展	小資產階級
左派	工人運動	社會階級	統治思想	政治經濟學
美國	台灣左派	封建思想	統治哲學	革新保台論
革命	台灣民族	帝國主義	勞動人民	馬列主義者
階級	台灣政論	政治活動	無產階級	馬克思主義
經濟	台灣革命	政治討論	階級立場	馬克斯主義
資本	台灣時代	政治發展	階級鬥爭	機會主義者
選舉	台灣問題	政治路線	階級理論	民族民主革命
同鄉會	台灣獨立	政治遊說	愛國主義	軍國帝國主義
留學生	左派聯合	政治運動	經濟發展	海外台灣運動
	民主革命	革命科學	解放運動	新台灣人民派
	民主運動	革命理論	資本主義	全美台灣同鄉會
	民族革命	革命組織	資產階級	
	民族壓迫	革命運動	學生運動	
	兩岸關係	貢薩維斯	機會主義	
	和平解放	馬列主義	黨外運動	
	社會主義	國際主義		

Journal of InfoLib and Archives 16:1=104 (June ’24) 38-68
DOI:10.6575/JILA.202406_(104).0002

Exploring Interpersonal Networks Based on Text Analysis: Insights from the Overseas Taiwanese Leftist Database

Yi-Chieh Wu [*] Hua-Yuan Hsueh [**]

【 Abstract 】

This study utilizes the digital archives of the “Taiwan Sitai,” led by the leftist movement leader Tso Hsiung. The database comprises 498 documents totaling 2.29 million words, including magazines, internal newsletters, and member correspondence, spanning from 1970 to 2018. Through advanced natural language processing techniques such as named entity recognition (NER) and keyword extraction (key terms), we analyzed the social network dynamics among members. The findings are illustrated with dynamic data visualizations that provide comprehensive network graphs, capturing diverse relationships and facets. Such visualizations serve as convenient tools to assist qualitative and quantitative research in social sciences.

Keywords

Natural Language Processing, Deep Learning, Named Entity Recognition,

[*] Assistant Professor, Interdisciplinary Artificial Intelligence Center, National Chengchi University
ORCID 0000-0002-2973-5735
Principal author for all correspondence E-mail: matywu@nccu.edu.tw
[**] Professor, Graduate Institute of Taiwan History, National Chengchi University
ORCID 0000-0002-8189-8830
E-mail: hy5595@nccu.edu.tw

Interpersonal Relationship Network, Data Visualization

【 Summary 】

The Special Collection Center of National Chengchi University Libraries has been dedicated to collecting various Taiwan-related data since its establishment. After the war, overseas Taiwanese abounded factions for political movement, in which overseas Taiwan leftists were famous for the solid theory and depth of publications. A series of leftist-position publications were collected in the database of the Special Collection Center of National Chengchi University Libraries. Among which, "Taiwan Sitai," led by Tso Hsiung, was a representative group of overseas Taiwan leftist movement in North America as well as the core of movement development. The publications, like *Taiwan Renmin*, *Taiwan Geming*, and *Taiwan Sitai*, were led by Tso Hsiung. The collected data covered personal letters, internal newsletters, published magazines, and journals. As of 2020, a total of 498 documents have been completed, encompassing approximately 2.3 million words of digital text.

Tso Hsiung, whose real name was Chung-Wen Lin, graduated from the Department of Oriental Languages and Culture at NCCU. "Tso Hsiung" was his most well-known pen name. Socialists could hardly maintain a foothold in Taiwan during the martial law period; they were often monitored by intelligence agencies, even when living abroad. Nevertheless, Tso Hsiung continuously advanced the linkage, organization, and theoretical propaganda among Taiwanese socialists. The analysis of the Taiwan Sitai Series Magazine could reveal the thinking processes, united fronts, interpersonal networks, and specific practice processes of leftist groups in Taiwan at that time.

Based on digital texts related to Tso Hsiung, the framed data contain several million words, span about half a century, and cover distinct research topics. To simultaneously address word segmentation and named entity recognition (NER) for traditional Chinese, a BERT-based model developed by the CKIP Lab of Academia Sinica is utilized for these analyses. Since NER focuses on entity words, such as people, events, time, places, and objects, but is not sensitive to notional words, like

"class" and "doctrine," simply relying on word segmentation results might not capture the evolution of leftist-related vocabulary in target texts. Therefore, we created a table of words of concern and calculated the frequency of each concerned word across the texts.

Furthermore, the network graph structure can be used to design the multifaceted relational data model, with the basic elements of entities, facets, relations, clusters, and temporal trends. The visualization design of these elements could express more complicated internal and external relations to provide users with a concise and highly efficient interactive experience. Visualization tools of word cloud and interpersonal relationship network are used for displaying data analysis results and comparing the differences among distinct text types and eras.

Based on the results of the named entity recognition analysis, texts before 1980 mainly focused on individuals and publishing organizations. From 1980 to 1990, the focus expanded to include both domestic and international issues. After 1990, the emphasis shifted towards international situations and historical issues. Regarding text type analysis, names were most frequently mentioned in letters. As the degree of text openness increased, both the breadth and strength of internationally relevant vocabulary also increased. Tso Hsiung even more frequently discussed historical context and commented on the international situation in public articles.

According to the analysis results of the concerned words, discussions of "class" enemies and friends, as well as the practice of "revolution," were two major propositions before 1971. The peak of the second emphasis on "class" and "revolution" occurred with the lifting of martial law in Taiwan in 1987 and the passing of Ching-Kuo Chiang in 1988. International events such as the Tiananmen Square incident and the Velvet Revolution also occurred in 1989, a year when political lines in Taiwan were shifting due to the promotion of liberalization, and people were expecting changes. This period is regarded as a crucial phase when Tso Hsiung reassessed the line and criticized objects. In terms of text type analysis, "rightists" and "leftists" were discussed more frequently in letter texts than in internal newsletters and public articles. The proportion of "imperialism" being

mentioned in public articles was significantly higher than in letters and internal newsletters.

Judging from case analyses, the proposed interpersonal relationship network could describe the text source of words, and its importance could be intuitively and efficiently displayed. This study also discusses whether named entities or concerned words are more "critical." It is concluded that each satisfies distinct research needs.

Literature archives contain information of people, events, time, places, and objects extended from the subject; comprehending the cutting point of historical data could be developed from points to lines and even surface. Such integration and planning could cope with different research needs from macro to micro. The automated framework based on natural language processing techniques is built in this study, and the analysis results are displayed with dynamic data visualization. Such a network planning comprehending various dimensions and relations allows researchers discussing various concerned factors as well as more deeply understand the quantitative information through visualization analysis results to assist in qualitative and quantitative research on social sciences and digital humanities.

Romanized & Translated References for Original Text

台灣時代（2020）。*台灣時代*。檢自 https://da.lib.nccu.edu.tw/tdm_2020/content/publish.html【Taiwan shi dai (2020). *Taiwan shi dai*. Retrieved from https://da.lib.nccu.edu.tw/tdm_2020/content/publish.html (in Chinese)】

政大數位典藏（2019）。*左雄與台灣時代社數位史料庫*。檢自 https://cdm20070.contentdm.oclc.org/digital/collection/TL【Digital Archives, National Chengchi University (2019). *Tso Hsiung yu Taiwan shi dai she shu wei shi liao ku*. Retrieved from https://cdm20070.contentdm.oclc.org/digital/collection/TL (in Chinese)】

國立政治大學圖書館特藏管理組（2020）。*臺灣政治與社會發展海外史料資料庫* [部落格文章]。檢自 https://da.lib.nccu.edu.tw/tdm_2020/index.html【Special Collection Center, National Chengchi University Libraries (2020). *Taiwan zheng*

zhi yu she hui fa zhan hai wai shi liao zi liao ku [Web log post]. Retrieved from https://da.lib.nccu.edu.tw/tdm_2020/index.html (in Chinese)】

劉吉軒、柯雲娥、張惠真、譚修雯、黃瑞期、甯格致（2013）。以文本分析呈現臺灣海外政治思想輪廓。在項潔編，*數位人文要義：尋找類型與軌跡*（頁 83-114）。臺北：國立臺灣大學出版中心。【Liu, Ji-Xuan, Ke, Yun-E, Zhang, Hui-Zhen, Tan, Xiu-Wen, Huang, Rui-Qi, & Ning, Ge-Zhi. (2013). Text analysis on overseas Taiwanese journals for political thought profiling. In Xiang, Jie (Ed.), *Essential digital humanities: Defining patterns and paths* (pp.83-114). Taipei: National Taiwan University Press. (in Chinese)】

Cao, N., & Cui, W. (2016). *Introduction to text visualization (Vol. 1)*. Paris: Atlantis Press. doi:10.2991/978-94-6239-186-4

Card, S. K., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: Using vision to think*. San Francisco, Calif: Morgan Kaufmann.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1,* pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423

Google. (n.d.). *Google trends*. Retrieved from https://trends.google.com/trends/

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics* (pp. 466-471). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/992628.992709

Mueller, A. (2017). *Word_cloud: A little word cloud generator in python*. Retrieved from https://github.com/amueller/word_cloud

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California. doi:10.48550/arXiv.1706.03762

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., ... & Houston, A. (2011). *Ontonotes release 4.0*. *LDC2011T03*. Philadelphia, Penn.: Linguistic Data Consortium, 17.

XAMPP. (n.d.). *Apache friends*. Retrieved from https://www.apachefriends.org/download.html

Yang, M. (2021a). *Ckiplab/bert-base-Chinese-ws.* Retrieved from https://huggingface.co/ckiplab/bert-base-chinese-ws

Yang, M. (2021b). *Ckiplab/bert-base-Chinese-ner.* Retrieved from https://huggingface.co/ckiplab/bert-base-chinese-ner