圖資與檔案學刊
10:1=92 (June '18) 67-87
DOI: 10.6575/JILA.201806_(92).0004

臺灣歷史人物文本檢索與探勘系統之建置

Development of a Text Retrieval and Mining System for Taiwanese Historical People

謝 順 宏*
Shun-Hong Sie

柯 皓 仁**
Hao-Ren Ke

張 素 玢***
Su-bing Chang

【摘要 Abstract】

「人物」是歷史學研究重要的實體類型之一，因此，對人物傳記的深入了解有助於歷史事件的相關研究。目前許多人物傳記資料是以數位文件的形式存在，而要以人力從大量人物傳記中爬梳、彙整資料頗為曠日廢時，宜妥為運用資訊科技協助歷史學家。此外，儘管臺灣過去已建置眾多資料庫，也有各種人物傳和可資應用的資料文獻，卻較少進行歷史人物資料庫勘考、分析工具的開發。有鑑於此，研究者乃組成研究團隊，以《新修彰化縣志・人物志》為文本來源，發展資料庫檢索、全文檢索、文本探勘與社會網絡等分析工具，協助歷史人文學進

* 國立臺灣師範大學圖書資訊學研究所博士生
Ph.D. Candidate, Graduate Institute of Library & Information Studies, National Taiwan Normal University
E-mail: mayh@ntnu.edu.tw

** 國立臺灣師範大學圖書資訊學研究所教授
Professor, Graduate Institute of Library & Information Studies, National Taiwan Normal University
E-mail: clavenke@ntnu.edu.tw

*** 國立臺灣師範大學臺灣史研究所教授
Professor, Graduate Institute of Taiwan History, National Taiwan Normal University
E-mail: 109682@ntnu.edu.tw

行研究，長期目標為建置「臺灣歷史人物資料庫（Taiwan Biographical Database, TBDB）」。本研究主旨在於描述「臺灣歷史人物資料庫」現階段所收錄之人物特性，闡述系統架構，以及說明初步成果。此外，本研究將提出一套演算法辨識《新修彰化縣志・人物志》中的命名實體（named entity），並以詩社名稱辨識為例說明。該套演算法的召回率達 96%，精確率則為 65%。最後，本研究將說明建置「臺灣歷史人物資料庫」過程中習得之經驗和未來發展方向。

Personage is an important kind of entities in the study of history. Comprehensive understanding of personage biographies is beneficial for researching into historical events. In the digital era, many personage biographies are available in digital formats; as a result, it is time-consuming and labor-intensive for researchers to explore invaluable findings from massive personage biographies. Facing this situation, researchers may be helped to utilize the information efficiently with information technologies. This article introduces the development of a text retrieval and mining system for Taiwanese historical people -- Taiwan Biographical Database (TBDB). It describes the characteristics of personages in TBDB, highlights the system architecture and preliminary achievement of TBDB, and proposes a method to recognize named entities in the personage biographies, specifically poetry societies, which achieves the recall rate of 96% and the precision rate of 65%. Finally, this article elaborates on the lessons learned through the creation of TBDB, and the future plans.

Keyword 關鍵詞

臺灣歷史人物資料庫 文本檢索 文本探勘 社會網絡分析 命名實體辨識

Taiwan Biographical Database (TBDB)；Text retrieval；Text mining；Social network analysis (SNA)；Name entity recognition

壹、緒　論

隨著資通訊技術的發展成熟，數位化學術研究（digital scholarship）逐漸成形。所謂數位化學術研究乃是指學者運用數位化的證據、探究方法、研究、出版、保存來達成學術與研究的目標，且利用數位、網路、開放的方法來展現領域的專門性（Martin, 2016）。數位化學術研究的範疇甚廣，數位人文（digital humanities）亦可容納在其中。數位人文涉及了運用數位化媒材（無論是原生數位或數位化的）和數位工具（如資料視覺化、資料與文本探勘、社會網絡分析、地理資訊系統）使人文與社會科學的學術研究發生轉變，可謂資訊科技和人文社會學科的匯流（Cambridge Digital Humanities, n.d.; Drucker, 2013; Digital humanities, 2017）。在現今人文社會學科的研究素材已大量數位化的情形下，運用電腦程式輔助人文社會學者快速整理、分析數位化研究素材，讓研究者快速完成比對和分析作業，繼而使用此結果進行相關決策，相信能使人文社會研究起事半功倍之效。

人物是歷史學研究的重要基礎，舉凡人物的個性、家庭背景、經歷、社會階層，甚至於整個社會的階層流動、婚姻與政治網絡等都是歷史研究的議題。「中國歷史人物傳記資料庫（The China Biographical Database, CBDB）」可謂目前與歷史人物研究相關最具規模的資料庫之一。CBDB 源自於一位美國的中國社會經濟史學者郝若貝（Robert M. Hartwell）的構想，自其於 1970 年代開始設計資料庫並蒐集中國歷史上的人物資料，而後由哈佛大學包弼德（Peter K. Bol）接手。目前 CBDB 的發展係由哈佛大學費正清中國研究中心、中研院歷史語言研究所、北京大學中國古代史研究中心所共同執行，截至 2017 年 8 月，CBDB 業已收集大約 417,000 位中國歷史人物的生平資料，這些人物主要分布於西元七世紀到十九世紀（Harvard University, n.d.）。CBDB 透過三種管道提供免費使用：(1)可下載、獨立運作的關聯式資料庫；(2)線上資料輸入系統；(3)線上檢索系統（Bol, Hsiang, & Fong, 2012）。CBDB 讓使用者得以運用統計、社會網絡分析、時間與空間分析等方法鑽研大量群體傳記學（Prosopography）的資料，以探索歷史問題（Bol et al., 2012）。

儘管臺灣過去已建置眾多資料庫，也有各種人物傳和可資應用的資料文獻，卻較少進行歷史人物資料庫勘考、分析工具的開發，因此研究者乃組成研究團隊，以 CBDB 為標竿，長期目標為建置「臺灣歷史人物資料庫（Taiwan Biographical Database, TBDB）」，提供臺灣歷史人物傳記資料的全文檢索、文本探勘與社會網絡分析工具和相關軟體服務。TBDB 建置初期，先以《新修彰化縣志‧人物志》為文本來源，分析 TBDB 的資料庫欄位，發展全文檢索、文本探勘與社會網絡分析工具，並運用前述工具協助歷史人文學者進行研究；俟建置有一定成果後，再擴及彰化以外的其他地區。本研究第二節介紹 TBDB 並說明其與 CBDB 不同之處；第三節從系統設計角度切入，闡述系統架構；第四節說明 TBDB 的初步成果；第五節則針對 TBDB 系統中的命名實體辨識技術發展深入探析。最後以分享

TBDB 建置經驗與說明未來規劃做結論。

貳、「臺灣歷史人物資料庫」介紹

本研究擬以中國歷史人物傳記資料庫（CBDB）為標竿，以建置臺灣歷史人物資料庫（TBDB）為目標，將歷史時間縱深延展到近、現代。CBDB 對每一位人物所收集的資訊包含以下七部分：(1)人物基本資訊、(2)社會關係、(3)入仕途徑、(4)官職種類、(5)親屬關係、(6)學術以及(7)社會身份（Fuller, 2015）。CBDB 對探索古代中國的人物來說是一個重要的傳記資料庫；然而，CBDB 所儲存的資訊不見得適用於 TBDB 的近現代臺灣人物，建置 TBDB 與 CBDB 不同的考量點如下（李宗翰、柯皓仁、張素玢、李毓嵐，2017）：

一、 時間：TBDB 收集 16 世紀之後的人物，重心在 20 世紀前後，CBDB 主要在 20 世紀以前。

二、 人物屬性：TBDB 收集在社會、政治、經濟、文化等面向的人物，包含大量非官職人員，後者則主要為官員。

三、 奠定社會地位的途徑：CBDB 以入仕為主要奠定社會地位的途徑，然而在 TBDB 中，現代的「選舉」相當程度與過去的「科舉」一樣具有重要性。

四、 境外經驗：近現代全球交通的便利，使 TBDB 的人物有更多的境外經驗，包括求學、工作、任官、移民、商業活動等。

五、 族群：TBDB 所收錄的人物不全為漢人，包括原住民、外國人（如教士、日本人、西方探險家）。

六、 社會階層：民意代表、醫生、大資本家等新階層大大突破過去的框架。

七、 職業：晚清以來，特別是日治時期臺灣地方人物所從事的職業，包括醫生、律師、商人等。

八、 社團：主要含括日治時期以後地方精英所建立的各種社團。

建置臺灣歷史人物傳記資料庫充滿挑戰性，儘管本研究團隊的長遠目標是以全臺灣人物的容納量與格局來設計，一開始卻不敢「做大」以免陷入龐大資料的泥淖。因此本研究團隊嘗試在 CBDB 打下的基礎上，步步為營，首先基於下述原因，挑選《新修彰化縣志・人物志》做為勘考、建置架構的對象（李宗翰等，2017）：

一、 彰化縣自清朝起在政治、經濟、文化、社會等方面皆有其重要地位。由於彰化縣盛產稻米，故其曾有臺灣米倉之稱；彰化縣也是臺灣第一個人口超過百萬的縣份。

二、 臺灣許多望族和社會名流皆發跡自彰化縣。

三、 研究團隊成員包含了《新修彰化縣志・人物志》的撰稿者，因此研究團隊能在符

合著作權規定下使用人物志的傳記全文（簡稱傳文）與照片；撰稿者也能以最嫻熟的研究底蘊，確實檢驗勘考工具的效度和信度，而做出正確的解讀，以發揮分析工具的價值。

四、《新修彰化縣志・人物志》具有一致的撰寫體例，有助於減輕電腦化處理的複雜度。

《新修彰化縣志・人物志》將所收錄的 887 位人物分為文化、經濟、政治、社會，字數超過百萬，人物含括彰化縣 26 鄉鎮市，時距長達 333 年。一旦相關功能發展完成，研究團隊將把 TBDB 的收錄內容擴展到彰化縣外的人物。

參、「臺灣歷史人物資料庫」系統架構

「臺灣歷史人物資料庫（TBDB）」系統的建置，除了提供傳文的檢索與瀏覽之外，更重要的是能探勘所收錄的歷史人物（簡稱傳主）間的網絡關係，輔以視覺化的呈現，協助使用者更快速地掌握資訊。為達到上述目的，初始需將傳文進行良好的剖析處理，以建構完善的底層資料；且需與歷史學者不定時接觸、訪談，瞭解其研究上的需求，從而轉化為系統功能。TBDB 的架構如圖 1 所示，分為資料層、資料處理層、服務層，分述如下。

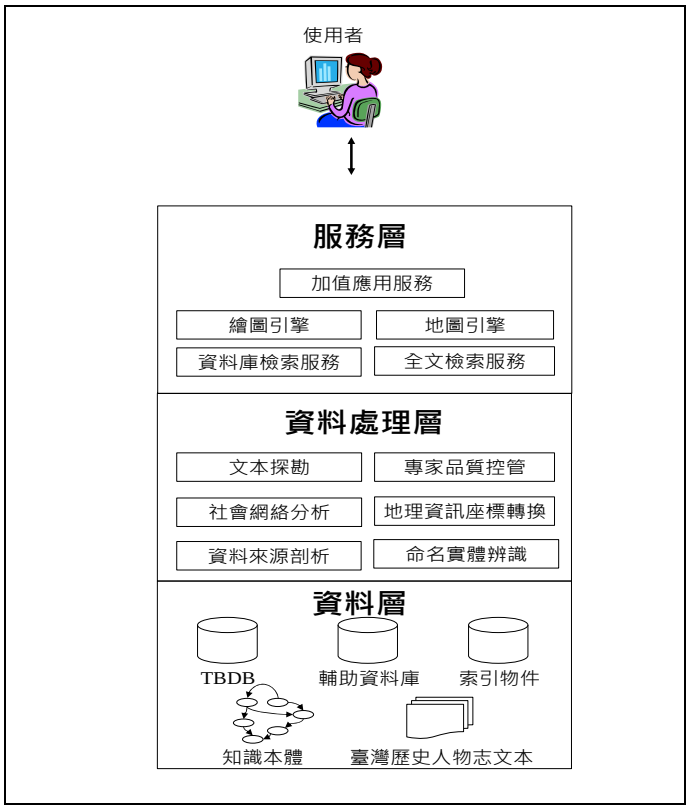


圖 1　TBDB 系統架構

資料來源：研究者自行整理

一、資料層

資料層方面包含數位物件、臺灣歷史人物資料庫 TBDB、輔助資料庫、索引物件、知識本體等部分。

首先，資料層存放了《新修彰化縣志・人物志》中人物傳記之原始文本，以及傳主相片等數位物件。

TBDB 採用物件導向資料庫的概念設計。首先釐清 TBDB 中各種重要的實體，每一種實體皆為一資料類別（class），TBDB 的實體包含：人物、地點、機構、社團、組織等；然後設計每一種實體所擁有的屬性（property），以人物而言，其屬性包含姓名、出生地、生卒年、教育程度、職業等；接著訂定實體間的關聯，例如人物與人物之間具有的親屬關係、社會關係，人物與地點間的出生地與活動地的關聯。由於 TBDB 係以物件導向程式語言 JAVA 開發，與物件導向資料庫屬於同樣的範式（paradigm），可減化、縮小程式開發與資料儲存之間的差距，也較為直覺（Brookshear & Brylow, 2015）；且 TBDB 仍處於探索、發

展階段，在人物、地點、機構、社團、組織之間存在各種可能性與不確定性，相較於關聯式資料庫需先有完善的系統分析、規劃後方能產出資料表格，並開始錄入資料，物件導向資料庫保留了最大的彈性。圖 2 為 TBDB 既存物件實體與關聯示意圖，實體物件以長方形表示，如傳主、詩社、地域，實體屬性則以橢圓形表示，如傳主姓名、生年、卒年、地點，其中姓名、生年、卒年為文數字或日期類型的屬性，地點則連結到地域的實體物件，傳主可能與一個以上的地域透過出生地、活動地等關係連結。

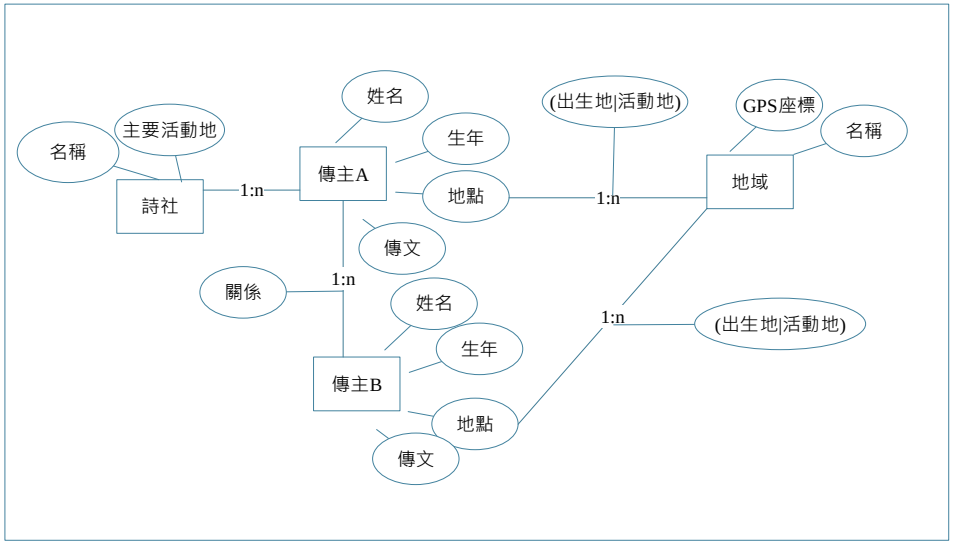


圖 2　TBDB 物件實體與關聯

資料來源：研究者自行整理

儘管如此，在建置成本與人力、維護等考量下，TBDB 的物件導向資料庫最終仍轉換成最單純並簡化結構的關聯式資料庫系統儲存，以結合物件導向資料庫易於延伸與強大的資料操作功能，及關聯式資料庫之易於管理與建置之特長（Lin, 2003）。隨著後續系統逐漸成型與研究者新增的需求，TBDB 可望再擴充既有物件屬性或其它類型之物件，並探索出更多實體之間的關聯。

TBDB 採用 Apache Lucene 檢索引擎作為基礎文本處理（包含斷詞、切字、索引建置）和全文檢索模組，因此在資料層儲存 Lucene 檢索引擎使用的索引物件。目前已建置的輔助資料庫包含地理資訊檔案和詩社名稱檔案。

未來將用知識本體概念來表徵欄位及控制詞彙，並進一步以鏈結資料（Linked Data, n.d.）格式在網路上發布。

二、資料處理層

資料處理層包含資料來源剖析、命名實體辨識、社會網絡分析、地理資訊座標轉換、文本探勘、專家品質控管等部分。

資料來源剖析的目的是將各種格式的傳文資料轉成純文字檔，以利 TBDB 後續建置 Lucene 索引，以及判斷傳文中的傳主姓名、生卒年、出生地、主要活動地等資訊後填入傳主基本資料表格。由於《新修彰化縣志・人物志》的原始傳主資料為 MS Word 格式檔案，故目前先以剖析 Word 檔案為優先，未來再視擴充之傳記資料格式發展不同的剖析功能。

在命名實體辨識方面，傳文內文所出現的人名、地名與組織名稱，為研究傳主生平的重要資訊。TBDB 嘗試設計半自動化之名詞擷取方式，進行專有名詞搜集。目前已初步完成詩社、主要活動地點、組織名稱之判讀。

社會網絡分析方面目前已完成探索《新修彰化縣志・人物志》中傳主相關詩社的情形。首先建構詩社輔助資料庫，並分析與每位傳主有關的詩社，最後以視覺化方式繪出人物、詩社、地點的關係。此外，人物志的描述中，「另有傳」（傳文描述中提到的傳主親屬、朋友亦有收錄在人物誌中）可以用來描繪收錄在人物志中之傳主的親屬或社會關係，本研究亦分析出傳主的另有傳關係，並繪出其社會網絡。

最後，為利用視覺化方式描繪出不同地點的相對位置，本研究亦進行地理資訊座標轉換的處理，將地名轉成地圖上的座標。

三、服務層

服務層包含資料庫檢索服務、全文檢索服務、繪圖引擎、地圖引擎、加值應用服務等，根據歷史學者的需求建置數位人文服務，提供給歷史學者使用。

在資料庫與全文檢索服務方面，如前所述，TBDB 採 Lucene 檢索引擎開發全文檢索或自訂多欄位檢索功能，並針對檢索結果提供層面分析與相關性之再查詢。由於 TBDB 以關聯式資料庫儲存傳主基本資料，故而可以進行關聯式資料庫的查詢。

繪圖引擎和地圖引擎方面，TBDB 採用了 D3.js（https://d3js.org/）函式庫發展網頁視覺化功能。D3 的全名為 Data-Driven Documents，2011 年由 Mike Bostock、Vadim Ogievetsky、Jeff Heer 等人聯合開發作為下一代網頁視覺化之用。TBDB 結合 D3.js 和 GeoJSON 格式資料，並搭配熱度圖（heat map），提供以多邊形方式呈現的地圖資訊格式。GeoJSON 資料格式，是一種用以處理地理資訊的 JSON 資料格式，GeoJSON 物件可以用來表示地圖上的點、線及面等幾何結構或地理特徵的集合（http://geojson.org/）。TBDB 取自政府資料開放平臺（http://data.gov.tw）所提供之「直轄市、縣市界線（TWD97 經緯度）」資料，內容包含各直轄市及縣市行政區界線圖資，再搭配鄉（鎮、市、區）行政區域界線資訊，能將系統結果，以地圖方式輔以色階即時呈現，提供操作者更加直覺的資訊。

此外，為能提供研究者詳盡且直覺之區域分佈，並期能結合現地資訊，TBDB 另引入 Leaflet（http://leafletjs.com/）元件作為進一步圖資呈現使用。Leaflet 元件容易操作與已內建大量的 API，使之容易與 TBDB 結合。TBDB 透過 MyGeoPosition.com（http://api.mygeoposition.com/geopicker/）所提供之 API，即時轉換地點為 GPS 座標，並繪製於地圖。

肆、「臺灣歷史人物資料庫」初步成果

圖 3 為 TBDB 首頁。目前系統提供了四個檢索點：人物類別（文化、經濟、政治、社會）、姓名、出生地、傳文。圖 4 為檢索結果的簡目列表，顯示傳主姓名、出生地、事蹟發生地、人物類別、生卒年等資訊，並可連結全文、檢視傳主的社會網絡關係；此外，使用者也可以選擇以 GDF 格式轉出傳主的社會網絡關係，讓使用者在 Gephi 軟體中自主操作。畫面左方則為後分類功能，提供人物類別、出生地和事蹟發生地等三層面（facet）限縮檢索結果。

圖 3　TBDB 首頁

資料來源：研究者自行整理

圖 4　檢索結果簡目顯示

資料來源：研究者自行整理

TBDB 業已實作三種社會網絡分析功能。第一種功能直接列出傳主傳文中提到的人物（無論是否為傳主），如圖 5 為辜顯榮傳文中提及人物的網絡圖。第二種功能將「另有傳關係」網絡化。所謂「另有傳」乃是在一位傳主的傳記中提及其他亦被《新修彰化縣志・人物志》所收錄的人物，透過「另有傳關係」的網絡化可以掌握重要或有名人物間關聯。圖 6 即為辜顯榮的「另有傳關係」網絡圖。前述二種網絡分析都可以選擇網路擴展的層次。第三種社會網絡分析功能展現詩社、主要活動地點、詩社成員間的關係，可藉此了解詩社成員的地理分布，例如圖 7 呈現出櫟社成員的主要活動地在鹿港、臺北、臺中、員林、彰化、大村等地。

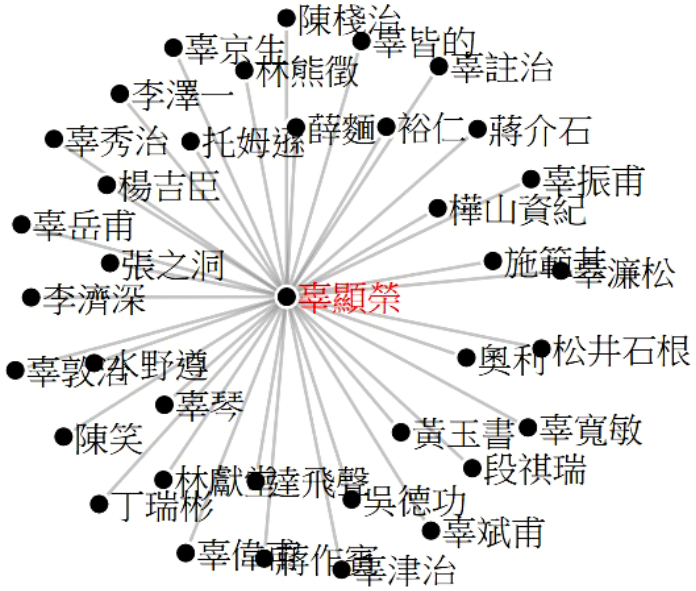


圖 5　辜顯榮傳文中提及人物的網絡圖

資料來源：研究者自行整理

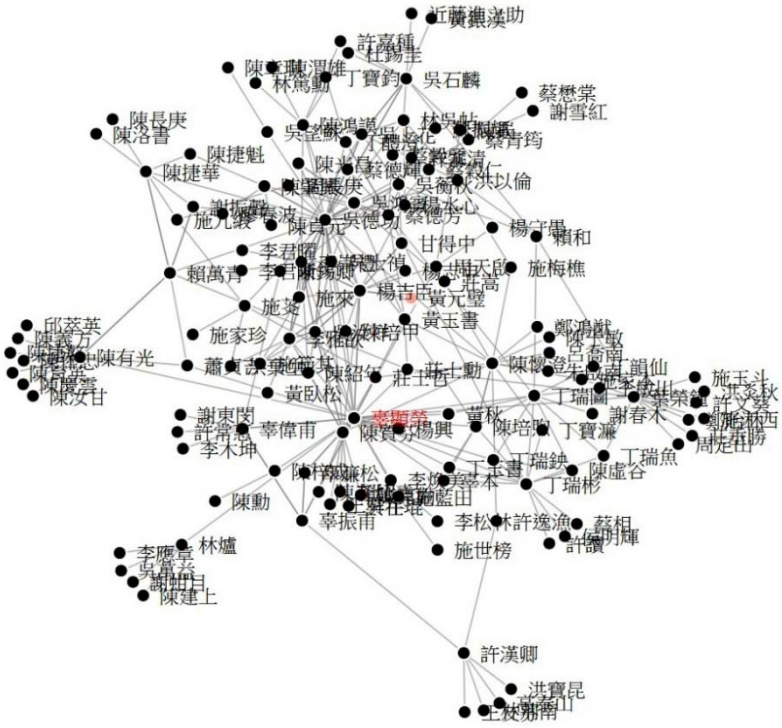


圖 6　辜顯榮的「另有傳關係」網絡圖

資料來源：研究者自行整理

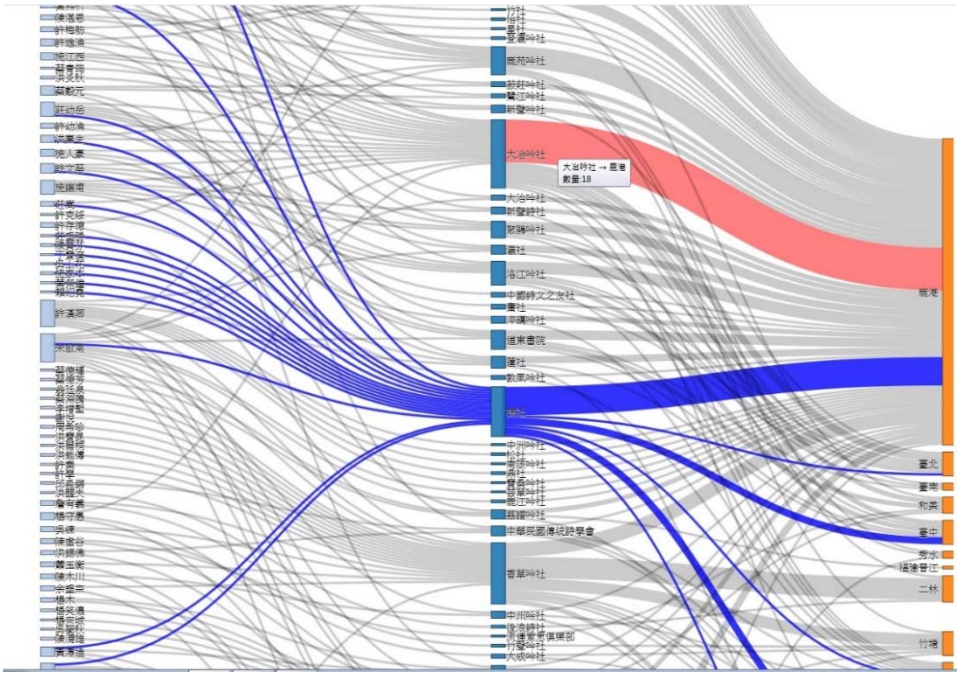


圖 7　詩社、地點、參與成員的網絡圖

資料來源：研究者自行整理

空間資訊有助於分析傳主活動。圖 8 以臺灣地圖呈現參與特定詩社成員的出生地分布情形，圖 9 則是以熱度圖將傳主活動地的分布情形視覺化。

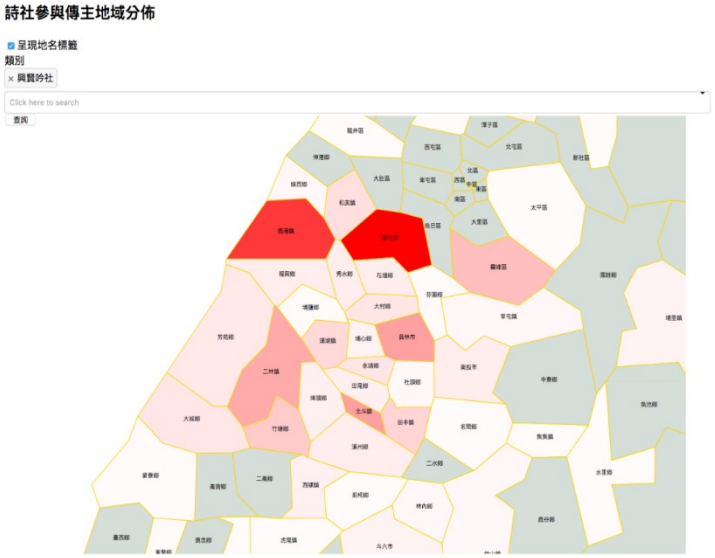


圖 8　特定詩社參與成員的出生地分布

資料來源：研究者自行整理

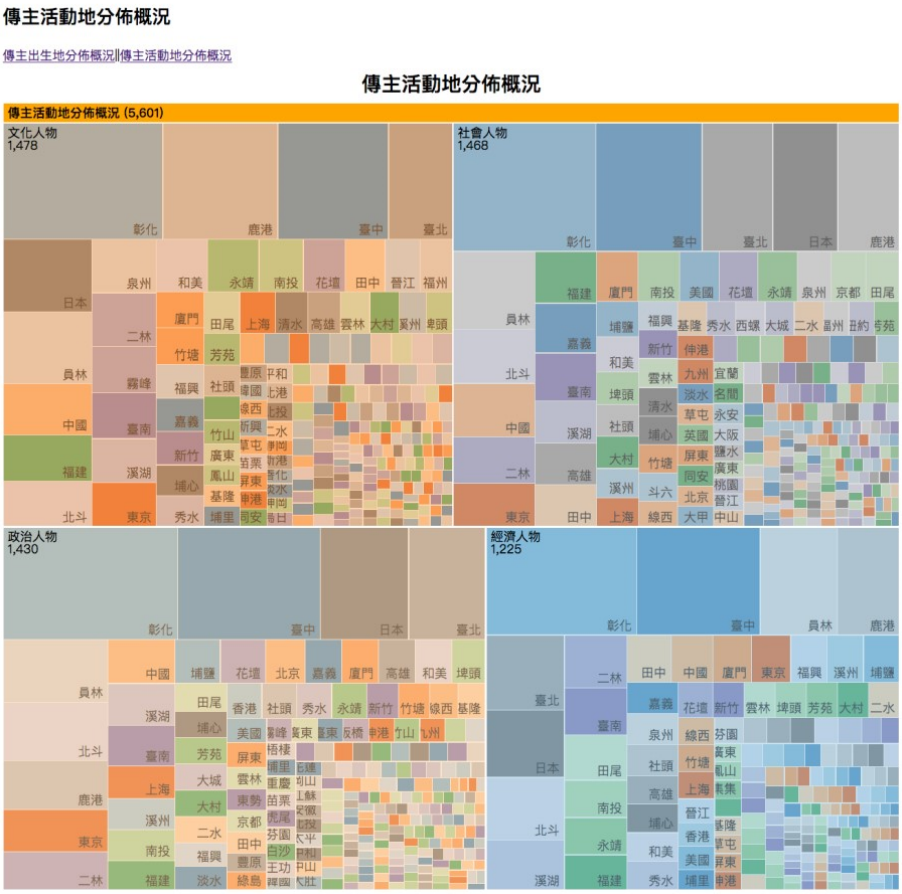


圖 9　傳主活動地概況

資料來源：研究者自行整理

根據傳主資訊，TBDB 產生若干統計圖表，除了出生地、活動地統計表之外，圖 10 為傳主卒年分布圖；圖 11 則為被提及人物排名，其中前三名依序為賴和、吳德功、施梅樵，排名第四的林獻堂因其出生於臺中霧峰而未收錄於《新修彰化縣志・人物志》，由此分析可做為 TBDB 可後續擴充之歷史人物（如：林獻堂）或地區（如：臺中）的參考。

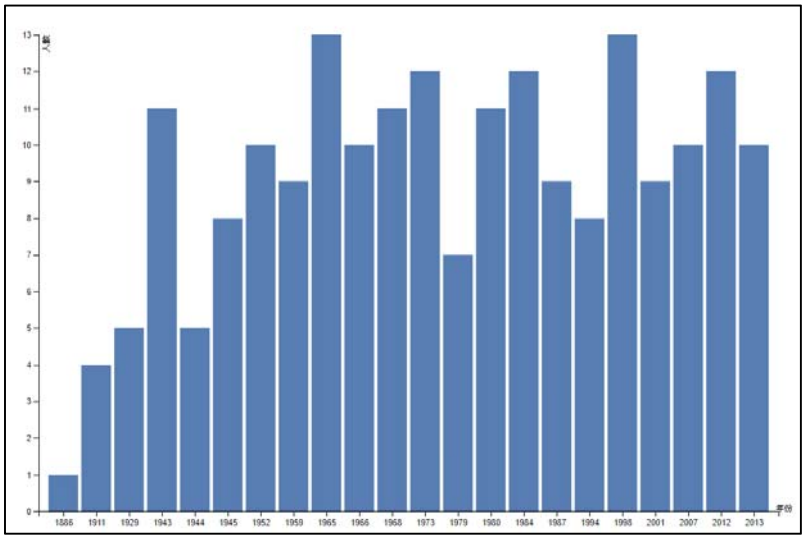


圖 10　傳主卒年分布圖

資料來源：研究者自行整理

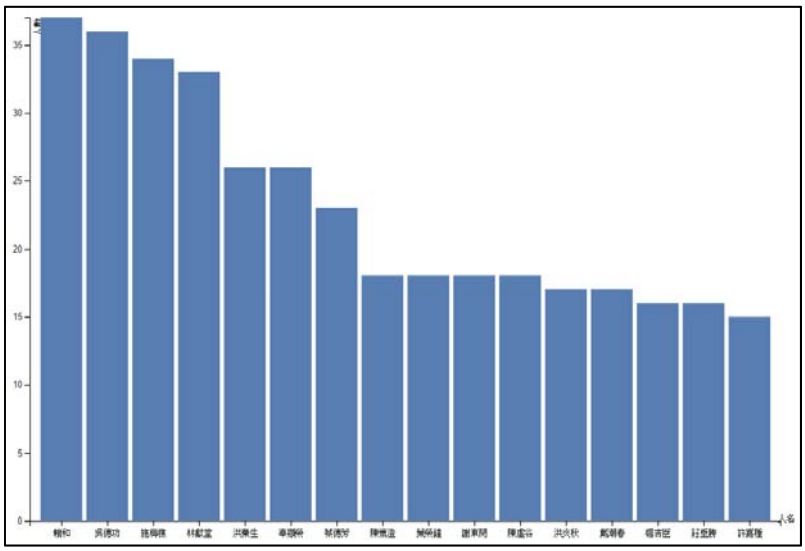


圖 11　被提及人物排名

資料來源：研究者自行整理

伍、TBDB 的命名實體辨識技術發展

與 CBDB 相似，本研究團隊企圖利用文本探勘方式自動填入 TBDB 的傳主資訊（Bol et al., 2012; Liu, Huang, Wang, & Bol, 2015）。傳文內文所出現的人名、地名與組織名稱，為描述傳主生平的重要資訊，命名實體辨識技術有助於從文本中勘考該等資料。本研究提

出一個半自動方式，擷取《新修彰化縣志・人物志》中的特定專有名詞。本節以「詩社」為例，說明技術細節。

首先，經由詩社名稱之輔助資料庫，可求得輔助資料庫內平均詩社長度 *l*，詩社最大相同字數 *n*，相似字串對 *W*。給定一傳文文本字串 *S*，若 *W* 出現在 *S* 的第 *p* 個位置，則 *substring*（*S, p-l+n, l-n*）（註 1）+*W* 和 *W*+*substring*（*S, p+n, l-n*）即可能為新的詩社名稱。以下列傳文文本為例：

> 創立「菱香吟社」，為賦風雅，是日治時期中部地區活躍的詩社之一……他被推選為菱香吟社首任社長，擅長傳統古詩……亦多次於北斗螺溪吟社、員林興賢吟社、二林香草吟社擔任課題詞宗，是促使菱香吟社茁壯有成的靈魂人物。

那麼在 n=2、l=4 的情況下，W=吟社，即可取得｛菱香吟社，螺溪吟社，興賢吟社，香草吟社｝等新詩社名稱。這個方法雖可以快速取得新詞彙；但由於中文字組合眾多，包含虛字與語助詞等會因符合資格而被取出，可能導致計算上的偏差，以下文為例，則可能產生｛各地詩社｝這個結果，造成擷取上的錯誤。

> 精於書法，詩品俊逸。昭和 5 年（1930）桃園人周石輝發行《詩報》，主要刊載各地漢語文言詩人作品與全臺各地詩社吟會、徵詩等，戰爭末期，《詩報》重心轉為漢詩與書道，開始大量刊載書法的推廣事項，他被「詩書報書道壇發行委員會」任命為參事及臨時組審查委員，可見其書法在當時頗受肯定。

為此，本研究另引入停用字詞表，過濾部份不具意義字元，以修正結果。具體而言，執行步驟如下：

一、 給定已有之輔助資料庫，求取平均字詞長度 *l*，字詞之間最大相同字數 *n*，相似字串對 *W*。

二、 給定欲擷取詞彙之文件集，進行擷取，以取得候選字詞結果。

三、 透過停用字詞表，篩選候選字詞結果，並產出最終結果集。

本方式可以用全自動的方式進行，並適用於各式擷取情況。透過專家建置的輔助資料庫自動運算並產出規則以進行擷取。但由於電腦演算法尚無法妥善處理語意關連，僅就語法、拼字組合進行擷取，而易導致偏差，故實務上，仍需適時的人工介入，並進行結果的修正。

為評估本方法之成效，本研究以《新修彰化縣志・人物志》作為實驗文本取得來源，而《新修彰化縣志・人物志》記敘著 1830 年代迄今之相關歷史軌跡，也因其涵蓋時空範圍龐大，宥以史料的駁雜，間接形塑內文部份專有名詞用字遣詞與現今有所出入，而在自

動化擷取上有一定的難度。

實驗進行步驟區分為二個部份，分為自動擷取與人工擷取的方式，自動擷取的方式透過由部份史料搜集之已存詩社名稱分析後，對全文進行擷取，並匯整產出詩社清單；而人工擷取的部份，則由三位歷史領域專家，採人工方式逐一由內文中，找出詩社清單；最後以人工作業的結果，視為最終答案，以計算其召回率與精確率。圖 12 為實際執行的過程中，系統將自動發現的詩社以特定顏色標記，以供專家判斷其正確性，若專家發現有誤，可將該詩社移除，若專家發現有遺漏的詩社，亦可將其加入。

根據專家判斷的結果作為正確答案，則前述演算法的召回率（recall）為 96%，精確率（precision）為 65%，數值呈現上看似未盡理想，但研究者認為在歷史研究中召回率（能夠找出所有相關結果的能力）較精確率為重要，因此前述演算法堪稱符合需求。另一方面，本研究也發現，系統自動判斷的「應社、旗津吟社、鹿港詩社、斐亭鐘會」等詩社卻未能被人工檢出，如圖 13 所示，這也顯示以人工擷取專有名詞，仍可能存在一定的失誤與不一致，故而以系統自動化辨識專有名詞輔助人工複檢乃是較可行的作業方式。

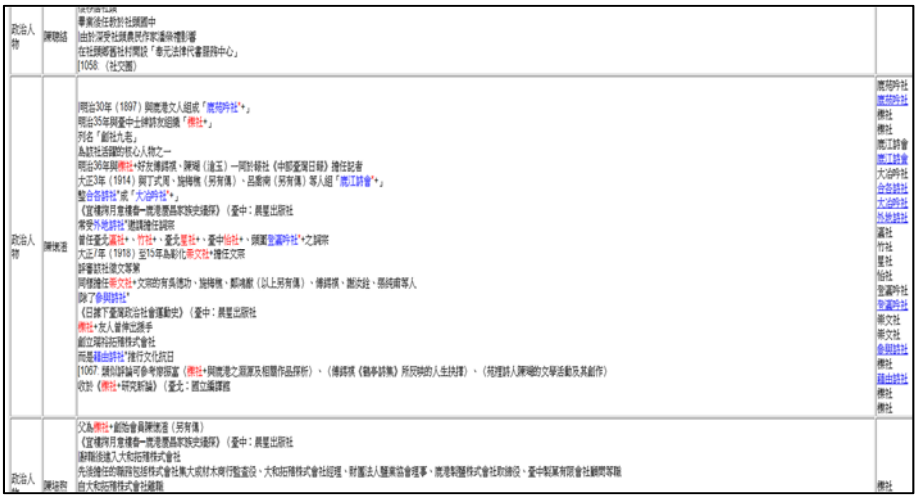
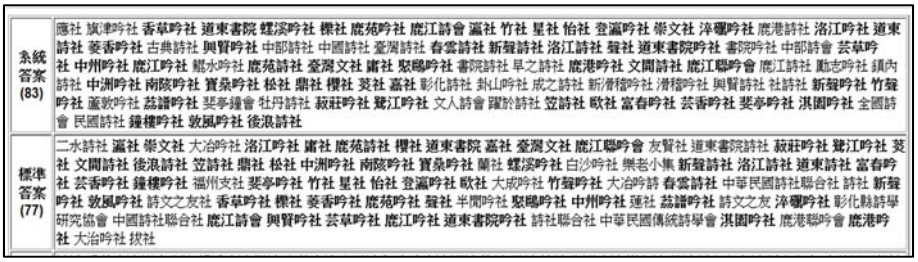


圖 12　提供專家檢核自動擷取詩社名稱正確性的功能畫面

資料來源：研究者自行整理

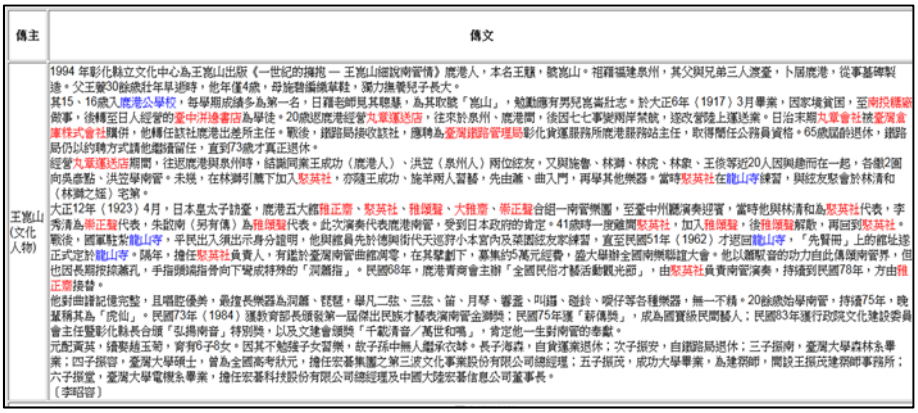


系統答案(83)	應社 旗津吟社 香草吟社 道東書院 螺溪吟社 櫟社 鹿苑吟社 鹿江詩會 瀛社 竹社 星社 怡社 登瀛吟社 崇文社 淬礪吟社 鹿港詩社 洛江吟社 道東詩社 菱香吟社 古典詩社 興賢吟社 中部詩社 中國詩社 臺灣詩社 春雲詩社 新聲詩社 洛江詩社 聲社 道東書院吟社 書院吟社 中部詩會 芸草吟社 中州吟社 鹿江吟社 鰲水吟社 鹿苑詩社 臺灣文社 庸社 聚賢吟社 書院詩社 旱之詩社 鹿港吟社 文聞詩社 鹿江聯吟會 鹿江詩社 勵志吟社 鎮內詩社 中洲吟社 南陔吟社 寶泉吟社 松社 鼎社 櫻社 莢社 嘉社 彰化詩社 卦山吟社 成之詩社 新潛稻吟社 潛稻吟社 興賢詩社 社詩社 新聲吟社 竹聲吟社 蘆敦吟社 茲譜吟社 琴亭鐘會 牡丹詩社 菽莊吟社 鷺江吟社 文人詩會 躍於詩社 笠詩社 歐社 富春吟社 芸香吟社 斐亭吟社 淇園吟社 全國詩會 民國詩社 鐘樓吟社 敦風吟社 後浪詩社
標準答案(77)	二水詩社 瀛社 崇文社 大冶吟社 洛江吟社 庸社 鹿苑詩社 櫻社 道東書院 嘉社 臺灣文社 鹿江聯吟會 友賢社 道東書院詩社 菽莊吟社 鷺江吟社 莢社 文聞詩社 後浪詩社 笠詩社 鼎社 松社 中洲吟社 南陔吟社 寶泉吟社 蘭社 螺溪吟社 白沙吟社 樂老小集 新聲詩社 洛江詩社 道東詩社 富春吟社 芸香吟社 鐘樓吟社 福州支社 斐亭吟社 竹社 星社 怡社 登瀛吟社 歐社 大成吟社 竹聲吟社 大冶吟詩 春雲詩社 中華民國詩社聯合社 詩社 新聲吟社 敦風吟社 詩文之友社 香草吟社 櫟社 菱香吟社 鹿苑吟社 聲社 半間吟社 聚賢吟社 中州吟社 蓮社 茲譜吟社 詩文之友 淬礪吟社 彰化縣詩學研究協會 中國詩社聯合社 鹿江詩會 興賢吟社 芸草吟社 鹿江吟社 道東書院吟社 詩社聯合社 中華民國傳統詩學會 淇園吟社 鹿港聯吟會 鹿港吟社 大冶吟社 拔社

圖 13　自動擷取與人工作業比對

資料來源：研究者自行整理

傳文內文所出現的人名、地名與組織名稱，為描述傳主生平與研究的重要資訊，但由於內文內容分散，內文文體因時間跨度較大、書寫方式近文言體緣故，重要名詞出現詞頻相對傳文可能未達顯著，即頻率遠低於可使用門檻值，再加上內容時間跨度較大，部份名詞、地名未能出現於現行之辭典或為出自古地名、耆老口述史料。為此，本研究嘗試引入半自動化之名詞擷取，進行人名資訊的搜集。透過現有之候選詞特徵的方式進行，以判斷候選字詞前後，是否具有特徵詞、字，再進行擷取；配合簡易中文斷詞、斷句，去除常用字詞、標點斷行後，再以詞夾子方法（張尚斌，2006）進行擷取。透過輔以人工進行簡易的判讀後，即完成特定名詞的擷取。而這些特定名詞，則可以再回饋原傳文內容，與所在之傳文進行判讀，以擷取出傳主相關的描述資訊。圖 14 為將前述方法應用於傳文中社團組織與公司的探勘，初步獲得良好的擷取效果，顯示前述演算法具有一定成效與穩定性。



傳主	傳文
王嵐山(文化人物)	1994 年彰化縣立文化中心為王嵐山出版《一世紀的擁抱 — 王嵐山細說南管情》鹿港人，本名王騏，號嵐山。祖籍福建泉州，其父與兄弟三人渡臺，卜屠鹿港，從事基礎製造。父王警30餘歲壯年早逝時，他年僅4歲，母施碧編織草鞋，獨力撫養兒子長大。 其15、16歲入鹿港公學校，每學期成績多為第一名，日籍老師見其聰慧，為其取號「嵐山」，勉勵應有男兒當壯志。於大正6年（1917）3月畢業，因家境貧困，至南投糖廠做事，後轉至日人經營的臺中洋逢書店為學徒。20歲返鹿港經營丸葉運送店，往來於泉州、鹿港間，後因七七事變兩岸禁航，遂改營陸上運送業。日治末期丸葉會社被臺灣倉庫株式會社購併，他轉任該社鹿港出差所主任。戰後，鐵路局接收該社，應聘為臺灣鐵路管理局彰化貨運服務所鹿港服務站主任，取得簡任公務員資格。65歲屆齡退休，鐵路局仍以約聘方式請他繼續留任，直到73歲才真正退休。 經營丸葉運送店期間，往返鹿港與泉州時，結識同業王成功（鹿港人）、洪笠（泉州人）兩位絃友，又與施魯、林獅、林虎、林象、王徐等近20人因興趣而在一起，各邀2圍向吳泰點、洪笠學南管。未幾，在林獅引薦下加入聚英社，亦隨王成功、施羊兩人習藝，先由簫、曲入門，再學其他樂器。當時聚英社在龍山寺練習，與絃友聚會於林清和（林獅之婿）宅第。 大正12年（1923）4月，日本皇太子訪臺，鹿港五大館雅正廚、聚英社、雅頌聲、大雅廚、崇正聲合組一南管樂團，至臺中州廳演奏迎賓，當時他與林清和為聚英社代表，李秀清為崇正聲代表，朱啟南（另有傳）為雅頌聲代表。此次演奏代表鹿港南管，受到日本政府的肯定。41歲時一度離開聚英社，加入雅頌聲，後雅頌聲解散，再回到聚英社。 戰後，國軍駐紮龍山寺，平民出入須出示身分證明，他與館員先於德興街代天巡狩小本宮內及菜園絃友家練習，直至民國51年（1962）才返回龍山寺，「先賢冊」上的館址遂正式定於龍山寺。隔年，擔任聚英社負責人，有鑑於臺灣南管曲館凋零，在其擘劃下，募集約5萬元經費，盛大舉辦全國南樂聯誼大會。他以團聚音的功力自此傳頌南管界，但也因長期按捺滿孔，手指側端指骨向下彎成特殊的「洞簫指」。民國68年，鹿港青商會主辦「全國民俗才藝活動觀光節」，由聚英社負責南管演奏，持續到民國78年，方由雅正廚接替。 他對曲譜記憶完整，且唱腔優美，最擅長樂器為洞簫、琵琶，舉凡二弦、三弦、笛、月琴、嚢盖、叫譜、雞鈴、嗩仔等各種樂器，無一不精。20餘歲始學南管，持續75年，晚輩稱其為「虎仙」。民國73年（1984）運教育部長頒發第一屆傑出民族才藝表演南管金獅獎；民國75年獲「薪傳獎」，成為國寶級民間藝人；民國83年獲行政院文化建設委員會主任暨彰化縣長合頒「弘揚南音」特別獎，以及文建會頒獎「千載清音／萬世和鳴」，肯定他一生對南管的奉獻。 元配黃英，續娶趙玉菊，育有6子8女。因其不能種子女習樂，故子孫中無人繼承衣缽。長子海森，自資運業退休；次子振安，自鐵路局退休；三子振南，臺灣大學森林系畢業；四子振容，臺灣大學碩士，曾為全國高考狀元，擔任宏碁集團之第三波文化事業股份有限公司總經理；五子振茂，成功大學畢業，為建築師，開設王振茂建築師事務所；六子振堂，臺灣大學電機系畢業，擔任宏碁科技股份有限公司總經理及中國大陸宏碁信息公司董事長。 〔李晤容〕

圖 14　社團與公司名稱的辨識

資料來源：研究者自行整理

陸、結論

人物是歷史學研究的重要基礎，也是研究歷史事件的重要基石。研究者效法哈佛大學「中國歷史人物傳記資料庫（CBDB）」，以《新修彰化縣志・人物志》為文本來源，發展臺灣歷史人物文本檢索與探勘系統，長期以建置「臺灣歷史人物資料庫（TBDB）」為目標。本研究概述 TBDB 的特色、系統架構、初步成果，並說明所發展的命名實體辨識技術。本研究最主要的貢獻在於提出一套可供建置人物文本檢索與探勘系統時的參考系統架構，並利用「臺灣歷史人物資料庫（TBDB）」的建置證明該系統架構的可行性；此外，本研究所發展的命名實體辨識技術擁有高召回率，亦能有效協助歷史學者進行人物關係的辨識。以下說明建置 TBDB 的經驗與未來規劃。

一、建置經驗

本研究團隊成立迄今將近二年，在建置 TBDB 過程中獲得以下寶貴經驗：

（一）計畫團隊由歷史學者和資訊科學家所組成，這兩個領域使用了不同語言，也導致相互了解的困難。以開放的心胸經常聚會與討論對促進雙方的理解是必要的。

（二）絢麗的工具不見得是適合的工具。在建置 TBDB 初期，資訊科學家很熱心的實作了許多工具，例如社會網絡分析、時間軸、熱度圖，這些工具也的確讓歷史學者讚嘆。但是當歷史學者進一步檢視這些工具，他們發現這些工具反而使分析更為複雜，甚至難以解釋。分析其原因，一方面是因為團隊中的歷史學者較少接觸到此類工具，在實作工具之初較難提出具體意見；另一方面則是資訊科學成員以實作工具為先，未能瞭解資料特性與意涵。故而，在兩種學科背景的成員發現問題、深入討論進而修正之後，方有目前在本研究中所呈現的資訊圖表。從這些經驗中，研究者認為一個好工具最重要的是能符合歷史學者的需求和研究流程、支持歷史學者的研究工作。

（三）電腦不是萬靈丹，舉例而言，現階段所開發的命名實體辨識技術仍無法同時達到 100%的召回率和精確率，是以仍需要有歷史背景的人員介入，判斷結果的正確性。另一方面，電腦的「中規中矩」卻也協助歷史學者辨識出許多在撰寫傳文時不慎使用的異體字（例如：黃與黃、啟與啓、昧與眛、熙與熙）。

（四）隨著數位與資通訊科技時代的來臨，許多人文社會學者急於跨入數位人文領域並喜於運用許多資料庫和工具，而可能淪於「為數位人文而數位人文」。研究者認為僅有在人文社會學者擁有問題意識的前提下，資料庫和工具才能夠發揮它們最大的效用。

二、未來規劃

TBDB 的短程規劃如下：

（一）本研究已完成雛型系統的建置，近期將對外開放使用，測試其優使性（usability），並根據使用者回饋的意見改善系統功能。

（二）為了自動化填入 TBDB 人物的欄位，未來將繼續辨識所有命名實體，並開發自動偵測親屬與社會關係的功能。於此過程中，亦將評估並逐步提升命名實體的辨識效能，如精確率。

（三）人物的親屬和社會關係是錯綜複雜的，需要好的社會網絡分析和視覺化工具。除了地理資訊外，時間資訊對歷史研究也是十分重要，未來將發展有助於探索時間資訊的工具。

（四）研究團隊未來將從 TBDB 構思新的研究議題，並展現 TBDB 能如何協助歷史研究。

TBDB 的長期規劃如下：

（一）將 TBDB 收錄之人物擴及全臺。本研究團隊未來將把 TBDB 的文本範圍從《新修彰化縣志・人物志》擴及清代臺灣方志 20 部、臺中市志、南投縣志、續修臺北市志、全台醫師名錄。未來在納入這些文本範圍時，將視其所涵蓋的傳主資訊彈性調整 TBDB 各類實體的屬性與關聯（此即採用物件導向資料庫概念設計 TBDB 的彈性），而由於傳文的用詞殊異（例如清代臺灣方志的用詞就與《新修彰化縣志・人物志》不同），因此命名實體辨識技術須加以調整。儘管如此，未來擴充 TBDB 時，仍然是以目前已發展的系統架構與各項技術為基礎。

（二）一張圖勝過千言萬語，一張團體照可能隱含了合影人的親屬或社會網絡關係。在研究團隊成員撰寫《新修彰化縣志・人物志》時即發現許多此類的團體照。未來本研究將利用群眾外包（crowd sourcing）等機制從老照片中辨識人物，從而協助建立親屬與社會關係網絡。

（三）促進 CBDB、TBDB 及類似專案計畫間的對話與系統互通（Bol et al., 2012）。

註釋

註 1：*Substring*（*S, a, b*）定義為從字串 *S* 的第 *a* 個位置擷取字元數量為 *b* 的子字串。字串 S1+字串 S2 將連結（concatenate）S1 和 S2。

致謝

本研究為科技部專題研究計畫的部分成果，計畫編號為：MOST105-2420-H-003-015、MOST105-2420-H-003-016、MOST106-2420-H-003-011。本研究為根據在第八屆數位典藏與數位人文國際研討會、太平洋鄰里協會（PNC）2017 年年會暨聯合會議（Sie, Ke, & Chang, 2017）所發表之論文增修而成。

（收稿日期：2018 年 3 月 31 日）

參考文獻

李宗翰、柯皓仁、張素玢、李毓嵐（2017 年 1 月）。從 CBDB 到 TBDB：以《新修彰化縣志・人物志》為試金石。在項潔、陳樹衡主持，*第八屆數位典藏與數位人文國際研討會（DADH 2017）*。國立政治大學數位人文團隊主辦，臺北市，中華民國。

張尚斌（2006）。*詞來子演算法在專有名詞辨識上的應用—以歷史文件為例*（未出版之碩士論文），國立臺灣大學資訊工程學研究所，臺北市。

Bol, P. K., Hsiang, J., & Fong, G. (2012). Prosopographical databases, text-mining, GIS and system interoperability for Chinese history and literature. In J. C. Meister (Ed.), *Digital Humanities 2012, Conference Abstracts* (pp. 43-51). Hamburg: Hamburg University Press.

Brookshear, J. G. & Brylow, D. (2015). *Computer science: An overview* (12th edition). Boston, N.J.: Pearson Education.

Cambridge Digital Humanities (n.d.). *Defining Digital Humanities.* Retrieved from https://www.cdh.cam.ac.uk/cdh/what-is-dh

Drucker, J. (2013). 〔Web log post〕*Intro to Digital Humanities.* U*CLA Center for Digital Humanities*: Retrieved from http://dh101.humanities.ucla.edu/?page_id=8.

Fuller, M. A. (2015). *The China Biographical Database User's Guide*, Revised Version 2.0. Retrieved from https://projects.iq.harvard.edu/files/cbdb/files/cbdb_users_guide.pdf.

Harvard University (n.d.). *China Biographical Database*. Retrieved from https://projects.iq.harvard.edu/cbdb/home.

Linked Data (n.d.) *What is the relationship between Linked Data and the Semantic Web? Frequently Asked Questions (FAQs)*. Retrieved from http://linkeddata.org/faq.

Lin, C. (2003). *Object-oriented database systems: A survey*. Retrieved from https://pdfs.semanticscholar.org/f2bf/923b8fade4ea1cfcb53683abd7aa7a1fa3a1.pdf.

Liu, C. L., Huang, C. K., Wang, H, & Bol, P. K. (2015, October). Toward algorithmic discovery of biographical information in local gazetteers of ancient china. In H. Zhao (Chair), *29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)*. Shanghai Jiao Tong University,

Shanghai, China.

Martin, L. (2016). The university library and digital scholarship: A review of the literature. In Mackenzie, A. & Martin, L. (Ed.), *Developing Digital Scholarship: Emerging Practices in Academic Libraries* (pp.3-32). London: Facet Publishing.

Sie, S. H., Ke, H. R., & Chang, S. B. (2017). Development of a text retrieval and mining system for Taiwanese historical people. In F. Lin, S. Chen, D. Wang & L. Chen (Ed.), *Proceedings of the 2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings* (pp. 56-62). doi:10.23919/PNC.2017.8203522

Digital humanities (2017). In *Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/wiki/Digital_humanities.