

# 知識天地

## 當資訊科學遇上歷史研究——兼論時空坐標系統的運用

范毅軍研究員（史語所、人社中心）、蔡宗翰教授（中央大學資工系、人社中心）、白璧玲博士後（人社中心）

### 摘要

當今資訊科學發展產生的作用，從尖端科學的探索到一般人的日常生活，幾乎無所不在。歷史學作為一門很古老的學科，也不免受到影響。然而一個仰仗大量質化的、文本史料為主的知識構建體系，如何令資訊科學有用武之地，仍然有各種觀念上以及現實技術面的障礙須待克服。我們嘗試以史語所《明實錄》全文檢索資料庫為本，運用自然語言處理的技術，企圖在兩個學科之間搭建一些可以溝通的管道。

### 一、前言

史語所前輩、民國以來的史學大家陳寅恪先生，綜其卓越的史才、史識、史德所創造的成就，垂諸今日，仍堪為我輩難以逾越的典範。姑不論陳先生學養之深厚與獨特之洞識與綜合的能力，在中共建政以後，身處嚴峻險惡的政治氛圍中，斷然拒斥歷史研究要宗奉馬列、學者要學習政治的要求，堅持「獨立之精神、自由之思想」，如此風骨，一般學者尤難望其項背！

陳先生向以過目不忘、博聞強記而知名，因之而極富傳奇色彩的是，當其晚年目擊贖足以後，主要倚仗特別助手黃萱，憑其口述竟能完成《柳如是別傳》一套三冊引經據典的巨著。此書主要藉明末清初柳如是、錢謙益之間錯綜複雜的情感生活與現世的交融，彰顯朝代遞嬗之際個人與社會間的種種糾結，同時也寓意作者個人感懷傷世的情思。

所謂「丹青難寫是精神」，為學風格與見地最難學，退而求其次，藉助現代的資訊科學，同享有陳先生個人為學之某些天賦異稟或有可能。這是史語所在上世紀90年代開始推動史籍自動化，首先完成二十五史全文檢索資料庫的初衷。希冀以電腦達成常人腦力做不到的一些事。

歷史研究首先涉及大量史料的查找，此傳統單純以查詢、檢索功能為主的全文資料庫已足堪膺其任。然而近十餘年來隨資訊科技大幅的進步，歷史研究者的需求也同步提升，不僅是史料的儲存與檢索，更在於如何對大量史料進行有效的分析、加值與應用；換言之，單純記憶之外，如何使電腦可發揮更貼近人腦思考、分析與綜合的能力，這就牽涉到資訊處理與資料庫的建置，從人工到機器學習，從非結構到結構性的過渡與發展的一些問題。歷史研究多以文本資料為主，如何進行適當有效的文本標註（mark-up）與分析，為其核心議題。

史事脈絡與內涵首須依據史料內容所涉之關鍵詞彙來進行探索。早先研究者利用漢籍全文資料庫，先據關鍵詞彙查找與研究主題相關的史料，繼以人工檢視的方式逐一進行分類、分析與綜合，從而逐步建構各種具有意義的情節（scenario）。然而，在人力限制下，此僅能聚焦於孤立的、零散的範疇。現在利用自然語言處理（natural language processing）技術，據關鍵詞彙可從巨量漢籍全文資料庫中，快速大量地擷取各種具關連性的史料，從而更有效地歸納與形成各種情節。進一步利用地理資訊科學的概念與方法，以時間與空間要素為軸，建立一個時空坐標系統；透過此一時空框架，將上述情節依點、線、面的坐標屬性一一予以定位，則可達成史事空間可視化的效果。

後者，欲透過時空坐標系統來串連與呈現史事，關鍵須具備完整的圖文整合機制。在本院數位典藏永續經營計畫之「以漢籍電子文獻資料庫之加值應用為主軸的時空坐標系統建置研究」子計畫中，我們開發了地名時空對位應用程式服務（CCTS-API），作為文本與地圖系統交互檢索之整合機制；藉由自然語言處理技術，對漢籍文本資料進行標註，並賦予基本的空間屬性；進而就事件類型、變化歷程等特性加以系統性分析，建立檢索介面。此不僅有效建立漢籍電子文獻資料庫與網路地理資訊系統的應用整合機制，還可進一步與各種外部資料庫的檢索介面進行串接，有利於資料更廣泛的加值應用與研究工作的開展。

### 二、研究實例

以具體議題為導向，探討上述模式應用於歷史研究的可行性，我們初步選定明代最龐大而重要的一手史料，圍繞明朝歷代帝王宮廷大政記事所成之《明實錄》為實驗文本，特別以其中與軍事防衛有關的衛所，作為地名識別的類型來進行測試。衛所是明代特有的軍事制度，起自太祖，終明之世各地普設都司衛所，作為地方控制、邊區防衛的機制，其興革損益涉及政治軍事形勢的變化，與明之興亡實相表裡。透過對文本所載之都司衛所名稱進行標註，並考量其時間與空間的分布狀態，綜合分析與衛所相關的各種事件與類別，預期

可通覽整個明代政軍形勢的演變及其特質。

本研究建立的衛所名稱標註系統，主要分為兩個層面，首先在於衛所候選詞的辨識。鑑於《明實錄》中的衛所構詞有一定的規則，我們採用半自動生成之構詞模板來擷取文本中的衛所候選詞；此類似規則式（Regular Expression）方法，但字串與模板匹配時允許插入、刪除與取代，故具有較大的彈性。此方法對於提升自動分詞的準確率，以及有效擷取候選詞的可行性，已見於相關研究文獻（例如Chang et al., 2015）。對於辨識出的衛所候選詞，計算其與「中華文明之時空基礎架構」（CCTS，網址http://ccts.sinica.edu.tw）衛所沿革表中，各衛所正規名稱的最小編輯距離（Minimum Edit Distance，MED），選擇MED最小的衛所正規名稱；又經由衛所名稱模板分析，得知後綴詞（衛、所、府或司）為鏈結衛所的必要條件，因而再結合後綴詞擷取結果來對應。此外，衛所設置隨時間發展多有變革，因此，須考量該衛所名稱出現在《明實錄》文本段落之紀日資料，經由中西曆轉換與時間區間計算後，再將衛所沿革表中符合條件的編碼附加至標註結果。

針對上述標註詞彙之辨識效果，我們抽選《明實錄》衛所標註結果中10%的資料量進行人工比對，評估其準確率與召回率，結果分別為84.78%與91.34%。為區別《明實錄》章節段落敘事之性質，進一步就文本之衛所標註結果，以非監督分群與分類方法來進行事件分析，初步可自動辨識並歸納出60個事件類型。

整合衛所自動標註與事件分類功能，我們建立了一個「明實錄衛所事件檢索系統」。透過網頁檢索功能適當的設計，不僅可篩選衛所事件與相關之文本段落進行排比與解讀，並可運用時間軸與統計圖表之展示功能，概覽事件類型之歷時性分布型態；同時，衛所標註所賦予的位置編碼，也可透過CCTS-API結合地圖介面，來呈現事件所述相關衛所的空間位置分布。凡此皆有利於與衛所相關的各項議題，做更深一層的探討。

舉例而言，有關衛所的功能與演變，藉由事件分群所歸納的衛所事件類型，以明初江西諸衛所的任務為例，據《明實錄》全文檢索資料庫，以系統檢索465筆段落資料結果中，占比例較高者為「派命守邊」、「置衛設員」、「陞調」、「派命」等，約佔總數的32%，同時以洪武年間事例最多。其中「置衛設員」約佔1/3（見圖一）；與漕運相關的事例總計約佔8%。另藉《明實錄》檢索系統搜尋文本段落「以袁州等衛官軍之戍贛者代領漕運」之記述，可知袁州衛又具有漕運功能，而袁州衛偏重漕運的任務，又與「袁州等衛在吉安下流便於漕運」的地理位置因素有關(引文見《明實錄》弘治元年二月條)。進一步侷限袁州衛，針對單一衛



圖一：江西諸衛所之事例類型檢索與統計

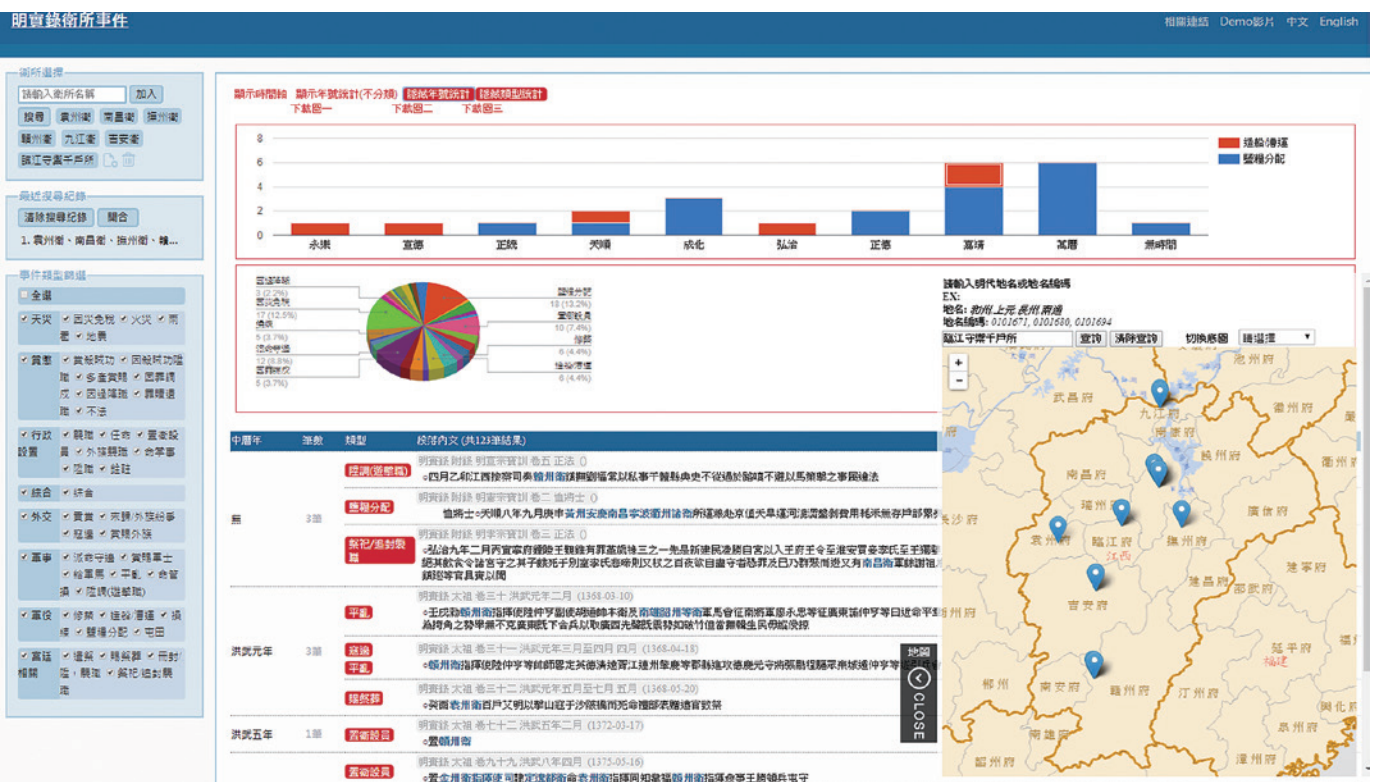
所進行分析，更多與其相關的事件類型與時間分布狀況，也可以透過事件檢索系統快速呈現。如以運糧與造船事例反映袁州衛所具漕運之功能，主要出現在弘治、嘉靖年間。據考，永樂2年(1404)以後江西大量衛軍投入軍屯，在《明實錄》中的記載，主要見於「調撥屯糧」或「受災蠲免」相關事宜；其後又明令以當地衛軍河運漕糧，使得江西大部分衛所逐漸兼有運糧或造船任務。然屯軍兼運、運軍又差役繁重的結果，導致衛所軍丁逃亡日多，至萬曆年間乃出現對漕運軍役進行改革的各種措施（于志嘉，2001）。如進一步檢索《明實錄》與漕運相關之江西其它各衛所，包括贛江沿岸的南昌、吉安、袁州，及長江沿岸的九江等，觀察其事例分布，結果也顯示嘉靖、萬曆年間與漕運相關事例居多（見圖二）。事實上，以屯軍正丁任漕運之後，轉由餘丁承擔屯田，待嘉靖年間江西一條鞭法出現，逐步規範均徭銀的徵收，限餘丁及屯田，又有依戶等高下派分正役的規定，而衛官佔役包納之弊也明文禁止（于志嘉，1997）。綜此，透過系統檢索漕運相關事例所見萬曆14年（1587）12月「丙戌更定江西南昌衛所軍餘丁差徵銀募役不許衛所官占役包納從按臣陳有年請也」條，又可一知其梗概。

### 三、結語

資訊科學技術導向輔助人文社會科學領域的研究，是一個新興的發展趨向，但也極具挑戰性。特別是以前本為主的歷史研究，希冀能有效利用資訊科學所發展出的一些功能，基本上牽涉到複雜的資訊從質化到量化的轉譯過程。我們一些歷史、資訊工程與地理資訊科學專長的同仁，有興趣於此，嘗試運用文本自動標註技術與地理資訊科學方法，做了如上一些探討與實驗性的工作。本文一方面就教於讀者，也寄望未來可擴及其他人文社科領域，會有更多的同仁加入研究團隊。

### 參考書目

- 于志嘉，1997年3月，〈明代江西衛所軍役的演變〉，《中央研究院歷史語言研究所集刊》，68（1），頁1-53。
- 于志嘉，2001年6月，〈明代江西衛所屯田與漕運的關係〉，《中央研究院歷史語言研究所集刊》，72（2），頁301-338。
- 白璧玲、賴郁婷、黃惠敏、吳承翰、蔡宗翰、范毅軍，〈文本自動標註與事件擷取技術於漢籍全文資料之時空資訊加值應用〉，第七屆數位典藏與數位人文國際研討會(DADH 2016)，2016.12.1-3 漢籍電子文獻資料庫，台北：中央研究院歷史語言研究所。
- Chang, Yung-Chun, et al. "Linguistic Template Extraction for Recognizing Reader-Emotion and Emotional Resonance Writing Assistance", ACL-IJCNLP 2015, Beijing, China.
- Tsai, Richard Tzong-Han, et al. "WeisoEvent: A Ming-Weiso Event Analytics Tool with Named Entity Markup and Spatial-Temporal Information Linking", Digital Humanities Conference 2017, Montreal, Canada.



圖二：江西漕糧衛所之事例檢索與空間可視化