

國立臺灣大學電機資訊學院資訊工程研究所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

歷史文件自動地名標註-以《清實錄》為例
Automated Annotation of Geo-information of
Historical Documents: A Case Study with the
Veritable Records of the Qing Dynasty

高欣愷

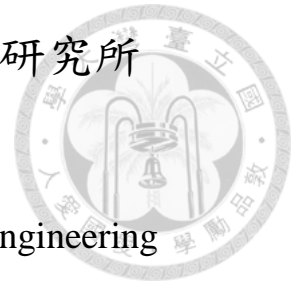
Shin-Kai Kao

指導教授：項潔 教授

Advisor: Jieh Hsiang, Professor

中華民國 102 年 7 月

July, 2013



致謝



兩年的時間一轉眼便過了，很快的自己將不再是學生，很快的自己將面對當兵、職場等新的生活。回想這兩年來，自己從對實驗室懵懵懂懂到如今已經要畢業了。這兩年來我學到許多東西，除了修課及寫程式外，學到最多的是怎麼去做好且表達一件事，在每個禮拜的 Meeting 中，不管是聽同學報告，或是自己在台上報告。老師總願意不厭其煩的教導我們如何去將一件事情說清楚，及如何將一件事不僅僅是做完而是做好。在未來，也許我不一定會走向數位人文這條路，然而，在實驗室所學到的東西，勢必能讓我在看待一件事時，能有不一樣的角度。

這篇論文的完成途中也遇過挫折與困難，幸好有許多人幫助我去克服這些困難，最後終於能夠完成。首先必須感謝項老師能夠支持這個想法，並且總是能夠提供很多有用的做法及建議。再來，也要十分感謝農堯學長，學長在 GIS 上的專業知識十分豐富，總是能夠提供我許多意見及幫助；也必須感謝杜協昌博士、宋浩、詩佩及浩洋學長提供許多他們專業的建議及協助；不管是在資訊的方面、歷史的方面以及報告的製作上都給了我很多幫助；謝謝光哲、老頭、士綱、阿姨、小黑、歐弟、柏淳、恐龍，分享了許多寶貴的經驗，不僅僅是論文的撰寫，在修課、程式能力上也十分有幫助；謝謝稷安、嘉軒、瑞安、凱勛、易徵、維謙、乃華、有為、沛強、信廷、綱政、豐成、若桓、偉儀這些實驗室的夥伴。感謝你們陪我一起討論、修課、念書、吃飯、玩樂，有你們在的實驗室，總是熱鬧而快樂，在日後我想必會不斷回味實驗室的與你們的各種趣事；也要謝謝台灣大學提供的許多資源，在我研究的過程中，發揮了很大的功用。

最後我要感謝我的家人，給了我了一個溫暖的家，讓我在一路的求學過程中，有一個舒適的環境。有了家人在背後的支持，讓我能夠健康的面對各種挑戰。還有，我也要對每一個關心我、支持我的朋友們說一聲，謝謝你們。

欣愷 民國一〇二年七月二十六日

中文摘要



歷史資料在近年來不斷進步的資訊技術下，開始能夠被數位化整理，且能結合地圖資訊，運用地理資訊系統，以空間的面向呈現歷史的脈絡，提供歷史學者一個不同的觀察角度，為人文研究帶來新的氣象。

然而，對於歷史研究者而言，地理資訊系統雖然便利，卻有一定的門檻存在，使得多數歷史研究者依然無法使用 GIS 軟體作為研究工具，使得歷史研究與地理資訊系統之間難以發揮原本預期之輔助結果。因此本研究的目標在於建構一個 GIS 能夠以直覺式的操作，被不同歷史研究者使用，因應其研究需求，自由的上傳資料，提供對應地理資訊的觀察。

本文以《清實錄》作為歷史文本的例子，結合地圖，開發一具有圖文整合能力的地理資訊系統。本文提出了使用空間資料庫及 Text Mining 技術，標註歷史文本中的地名且加入地理資訊的方法。且使用者能夠透過介面的操作，將自動化標註之結果做人工的校正。最後能將文件的地理資訊與地圖結合，呈現出文件的地理位置，使研究者能藉由視覺化的角度的觀察文件的空間資訊。期待研究者能藉由使用此系統，降低對使用 GIS 作為研究工具的抗拒，進而理解 GIS 對歷史研究的幫助。

關鍵字：HGIS、清實錄、Text Mining、地名辨識、離群點

ABSTRACT



Through the progress of information technology, a lot of historical documents have been digitized in recent years. By integrating the digital files and geographic information, a Geographic Information System (GIS) can display the context of history on maps, help historians observe phenomena which are not easily found from article alone, and provide historians a different perspective on historical research.

However, there is a technical barrier. For users without sufficient knowledge in geography, mastering GIS can be a daunting task, and the results may not necessarily meet the users' need. The goal of this thesis, then, is to develop a GIS with an intuitive user interface, through which historians can construct geographical maps that correspond to their data.

In this thesis, we use the “Veritable Records of the Qing Dynasty” as the historical text, combined with maps, and develop a WebGIS tool. We use spatial databases and text mining technologies to annotate the place name, extract the geographic vocabulary from the texts, and identify the geographic coordinates (landmarks) corresponding to the names. A user can modify the annotations or landmarks easily through the UI. The landmarks are then displayed on digital maps, thus providing the user a way to visually observe the geographical information of their data. We hope that our tool can reduce the technical difficulties that historians often encounter when using GIS, and encourage them to better utilize geographical information in their research.

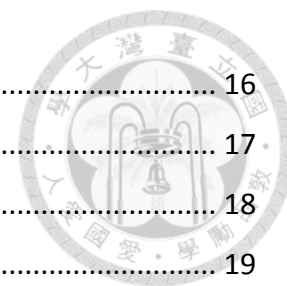
Keywords: HGIS, Veritable Records of the Qing Dynasty(清實錄), Text Mining, Location Name Recognition, Outlier points

CONTENTS



致謝	i
中文摘要	ii
ABSTRACT	iii
CONTENTS	iv
LIST OF FIGURES	vi
LIST OF TABLES	vii
Chapter 1 緒論	1
1.1 研究背景	1
1.2 研究動機	2
1.3 研究回顧	3
1.3.1 Talos-Web Annotation Tool	3
1.3.2 總督府抄錄契書地理資訊	4
1.3.3 地理編碼	5
1.3.4 詞夾子演算法於專有名詞辨識上的應用	6
1.4 論文架構	7
Chapter 2 使用資料及工具介紹	8
2.1 中華文明之時空基礎架構	8
2.2 使用史料-《清實錄》	11
2.3 系統建構工具-Timemap	12
Chapter 3 現有地理資訊系統技術回顧	13
3.1 地理資訊系統(GIS)簡介	13
3.2 常用 GIS 軟體與工具	13
3.2.1 伺服器 GIS 軟體(Server GIS)	13
3.2.2 個人電腦 GIS 軟體(Desktop GIS)	13
3.2.3 網頁地圖(WebGIS)	14
3.3 GIS 標準格式	14
3.3.1 Shapefile	14
3.3.2 KML	15
3.3.3 GeoJSON	15
Chapter 4 歷史文件地名標註系統	16

4.1 系統概述	16
4.1.1 系統架構	17
4.2 地名標註流程	18
4.2.1 詞庫式地名辨識	19
4.2.2 驗證比對地名	19
4.2.3 匯入空間資訊	23
4.3 系統功能	26
4.3.1 提交、新增、刪除文章至文本資料庫	27
4.3.2 過濾文章、時間軸呈現	29
4.3.3 校正地名標註	31
4.4 系統操作實例	31
4.4.1 工具操作流程	31
4.4.2 標註操作實例-以《清實錄》文本為例	33
Chapter 5 結論與未來工作	37
5.1 結論	37
5.2 未來工作	38
REFERENCE	39



LIST OF FIGURES



Fig. 1-1 Talos-Web Annotation Tool	4
Fig. 1-2 總督府抄錄契書地理資訊	5
Fig. 1-3 以「臺灣大學」做地理編碼後結果	6
Fig. 2-1 中華文明之時空基礎架構	8
Fig. 2-2 《中國歷史地圖集》1820 年清代地圖	9
Fig. 2-3 空間資料庫資料範例	10
Fig. 2-4 漢籍電子文獻資料庫《清實錄》數位化全文	11
Fig. 2-5 《清實錄》資料庫條目內容	12
Fig. 2-6 應用 Timemap 建立 WebGIS 之範例.....	12
Fig. 4-1 系統架構圖	18
Fig. 4-2 地名標註流程圖	19
Fig. 4-3 離群點例子	21
Fig. 4-4 同名異地處理情形	24
Fig. 4-5 利用填寫表單提交文章	28
Fig. 4-6 XML 檔案的樣式.....	29
Fig. 4-7 以系統檢索含「六百里傳諭」的文章	30
Fig. 4-8 系統運作流程	32
Fig. 4-9 新增文章	33
Fig. 4-10 經自動化標註後之結果	34
Fig. 4-11 校正標註.....	35
Fig. 4-12 校正標註後結果呈現	36

LIST OF TABLES



Table 2-1 CCTS 歷代地名圖層年代對照 10




Chapter 1 緒論

1.1 研究背景

在歷史研究中，地理學與歷史學一直是關係十分密切的兩門學科，當歷史學家面對過去時間的人、事、物時，是難以將空間的要素抽離的，若觀察歷史紀錄中的人物或事件時，不配合當代的地理背景，是難以深入史料的內裡，踏入歷史的真實脈絡之中的。因此對於多數歷史學者而言空間資訊確實是有其重要性的，然而在過去這些空間資訊的彙整、地圖的製作皆必須要以人工的方式去完成，過程費時費力，空間的變化難以察覺。

然而近年來，「數位人文」一詞已成為一新興的研究領域，隨著時間推移，大量的歷史典籍、檔案完成全文數位化，研究者可透過資訊科技從這些數位化檔案中挖掘出以往透過逐字閱讀所難以觀察到的現象。除了文字部分外愈來愈多的圖像資料也成為數位化的檔案，許多歷史地圖也被數位化為電子資料，除了單純的影像外，也有許多研究者將歷史地圖上的圖例、河川、海岸線、行政區等地形地物資訊以點、線、面標示為向量化(vector)的數位地圖。向量圖可對每筆地理資料增加描述欄位；疊合多種圖層；可匯入軟體中做更多計算及資料處理...等。由此過去所彙整的空間資訊及地圖透過資訊技術提升了許多可用性。像這樣的地理資訊系統(geographic information system，後稱 GIS)在人文研究上日益受到重視，希望透過其所擁有的多種功能，將地理要素帶入人文研究之中，使研究者能跳脫文字的侷限，配合空間與時間的整合對研究議題進行思索。

所以，近年來歷史學研究開始結合 GIS，形成所謂 HGIS(Historical GIS)的學術分類。但對多數學者來說，對於應用 GIS 當作研究工具，依舊有相當程度的抗拒。抗拒的動機一為技術的門檻，專業的 GIS 軟體功能繁多，並不容易熟悉，有些功能要求使用者必須要有一定程度的資訊技術，對於僅習慣用電腦處理文字或表格



的研究者來說，要能充分使用這些 GIS 軟體並非一件易事；其次，傳統史學研究著重在文字的思考，圖表表達的重要性並不被看重，HGIS 的呈現被視作附加的價值，而非必要的項目。[1]因此雖然 GIS 是一個十分有潛力的工具，但歷史學者使用 HGIS 進行研究的情形仍被限制住。


1.2 研究動機

中央研究院歷史語言所范毅軍研究員於〈試論地理資訊系統在歷史研究上的應用〉提及[2]：

一般研究者多著重於文字敘述，而疏於根據地圖做具體的空間分析，此事實上就等於平白忽略了一大部分史實，或是錯失了了解或分析史實的一個面向。地理資訊系統作為一個應用性的工具，正可以對這方面的缺失有所補正

在人類的歷史中，幾乎所有的活動皆與空間有密切的關係，因此，如何透過 GIS 技術結合各種人文因素，藉由視覺化的方法讓歷史研究者可以由地理的角度觀察歷史事件的脈絡，並方便說明與講解史實，變為學界與開發者所努力的目標。然而，如上節所述，HGIS 對於歷史研究者來說是有其限制在，首先技術的門檻使一般歷史研究者難以活用專業的 GIS 軟體，大部分的 GIS 軟體若要活用必須要熟悉像資料庫、SQL 等資訊技術，這類軟體對於研究者來說是不太容易上手的，也成為抗拒的原因之一；再者，有許多研究者手邊僅有文字資料，對於 GIS 使用而言是不足夠的，圖層、坐標資料等地理資料是需要花費大量人工及精神去整理的，對歷史研究者而言並不容易取得。這兩個因素，造成一般歷史研究者難以將 HGIS 作為研究工具。本研究希望能夠降低研究者使用 GIS 當研究工具的門檻，作為研究者進入 GIS 領域的一個開始。

本文第一個工作為建立一系統能夠幫助研究者透過 Text Mining 技術將文本自動化標註地名並為文本加入坐標資訊，以節省研究者需要花費大量精力整理文本的地理資訊的時間，以經整理過的地理資訊作為空間資料庫幫助研究者將自己擁有、整理的文本資料結合系統的資料，建立文本的空間脈絡。



第二個工作為透過 Google Map API 及 Timemap 時間軸工具將這些整合過的資料以視覺化方式呈現。相較於一般 GIS 軟體成本過高，像 Google Map 這類的 WebGIS 只要連上網際網路便能使用，且可透過簡單的介面操作，達到使用者所需的機能。本研究希望藉由介面的設計，盡可能降低使用者操作的門檻，透過設計的功能選單取代繁複的操作，藉此提高使用者使用 GIS 的接受程度。透過內建的資料庫做為基礎，由以往單純的使用 GIS 瀏覽模式，變為可主動提供資料進行 GIS 建立的創造模式，由被動的接受變為主動的應用。

在下文將會詳細的介紹本研究所使用的資料，以及建構系統所需要的程式語言及介面設計理念，並以實例展示本系統的功能及研究潛力。希望能達成前述的研究目標，即透過降低 GIS 使用門檻，拉近歷史研究者與 GIS 的距離，不需要具有極專業的資訊背景，亦能讓使用者嘗試使用 GIS，整合時間與空間，在歷史研究中發現一個不同的面向

1.3 研究回顧

隨著資訊技術發展，許多研究機構在 GIS 或地名標註都有相關的研究值得學習參考，以下將分別介紹這些相關研究。

1.3.1 Talos-Web Annotation Tool

網址：<http://www.talos.cti.gr/index.php/results/annotationtool>

Talos 是一個提供技術基礎，適地性服務(LBS, Location Based Services)之相關工作。Talos 開發了網頁版的標記工具，透過標記內文的動作，能連接內文與系統提供的 POI(Points of Interest)之關聯性，以此豐富了內文地理資料的詳細資訊；對於空間資料而言，此工具也加入了 geocoding 功能，讓用戶能自行新增與修正 POI 的資料，以避免遭遇 POI 資料不足的困境。

TALOS Overview

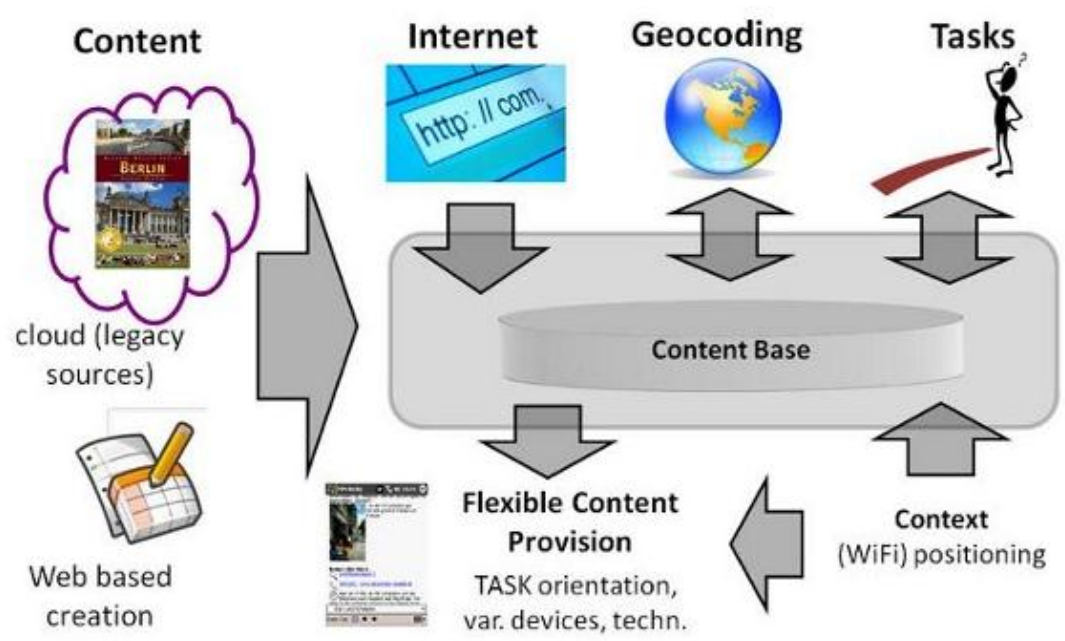


Fig. 1-1 Talos-Web Annotation Tool

1.3.2 總督府抄錄契書地理資訊

網址：<http://thdl.ntu.edu.tw/tools>

此系統分析了 THDL(Taiwan History Digital Library，台灣歷史數位圖書館)裡《古契約文書》文獻集中收錄之《台灣總督府檔案抄錄契約文書》，具「堡庄資訊」之 12,473 件契書的地域分布。透過此 GIS 使用者可以選擇使用關鍵字檢索、分類、地區及其他從契書中擷取出的詮釋資料，將契書之地理分布顯示於台灣地圖或是 1904 年台灣堡圖，此系統特點在於提供了多邊形檢索及時間軸變化，讓使用者能夠針對某一特定地區，觀察台灣古地契，空間與時間的變化結果。但該系統為針對古契書檔案進行建構，並不容易套用其他種類之歷史資料研究。



Fig. 1-2 總督府抄錄契書地理資訊

1.3.3 地理編碼

所謂地理編碼，是將地址或地標的表格紀錄與相關圖層間建立聯繫，將地理坐標分配給含對應地址的表格紀錄，並為其創建一個對應的點要素圖層。目前有相當多的 GIS 開發者使用相關的技術，以下將介紹 Google Map API 的地理編碼服務作為例子。

網址：<https://developers.google.com/maps>

此 API 服務提供使用者將地址透過 HTTP 要求存取 Geocoder，將地址轉換為地理坐標，以用來於地圖上放置標記或設定地圖位置亦能提供反向地理編碼(將坐標轉換為地址)。伺服器端接收到使用者的要求後會將地理編碼結果以 XML 或 JSON 格式輸出回用戶端，使用者可透過回傳的結果編寫程式繪製相關地圖。此服務的目標與本文相近，皆想幫助使用者將地名或地址與坐標資訊結合並繪製為地圖，差別在於此服務的目標使用者著重在網站或行動服務的開發人員，其傳回的內容需要再進程式的撰寫才能繪製為地圖，並不容易被一般使用者使用；另外，服務的資料內容為現代地址及坐標，現在網路上的地理編碼服務少有針對歷史文件去建立。

```

    "formatted_address" : "106台灣台北市大安區羅斯福路四段1號國立台灣大學",
    "geometry" : {
      "bounds" : {
        "northeast" : {
          "lat" : 25.02217180,
          "lng" : 121.54610070
        },
        "southwest" : {
          "lat" : 25.01154430,
          "lng" : 121.53327810
        }
      },
      "location" : {
        "lat" : 25.01635430,
        "lng" : 121.53676870
      },
      "location_type" : "APPROXIMATE",
      "viewport" : {
        "northeast" : {
          "lat" : 25.02217180,
          "lng" : 121.54610070
        },
        "southwest" : {
          "lat" : 25.01154430,
          "lng" : 121.53327810
        }
      }
    },
    "types" : [ "university", "establishment" ]
  },
  "status" : "OK"
}

```




Fig. 1-3 以「臺灣大學」做地理編碼後結果

1.3.4 詞夾子演算法於專有名詞辨識上的應用

中文專有名詞辨識，一直有許多人在做此方面研究，通常分為 Rule-base、Static-base、Machine learning 等三種方式，亦有使用混合式的方法，例如使用 SVM 來結合規則式與機器學習的中國地名識別方法[4]。然而這些方法有些必須要經過語言學家和各領域專家花費大量時間來建立詞庫及規則庫甚至是 POS(Part-of-speech)tagging，或是必須先花費大量時間來訓練機器。然而這些辨識方式多是針對現代文章而少有針對歷史文件去建立詞庫或詞性標記，使得這些方法使用在歷史文件上無法達到在現代報紙雜誌辨識的效果。而使用詞夾子來輔助使用在歷史文件上則是一個不錯的方式，詞夾子為利用名詞前後文特性當作規則，利用這些特性來萃取專有名詞[3]。

利用詞夾子演算法做辨識的優勢在於，其不需要先做斷詞，亦不需要大量的詞性標記或詞庫，也就是此演算法不需要像一般辨識演算法需要先花費大量人力整理資料，僅需要一份樣本詞的列表作為初始訓練詞夾子的材料。而辨識的成果卻能與採用大量語料庫的中研院斷詞系統達到差不多的精確度。



然而本文的目標並非是想在歷史文件中找出新的地名，與原本此演算法的目標不同，但根據〈《清實錄》人名擷取自動化〉[5]一文中，利用 PMI 方式針對《清實錄》斷詞，在驗證每一個斷開的詞時則利用詞夾子演算法以及人名用字來幫助計算標記人名的機率，此種作法能有效提高辨識的精確度。本研究並不像中研院具有相當大量的語料庫，難以使用統計模型來針對地名做驗證，因此在本文中進行地名標註時，為了驗證地名列表比對的結果，亦採用與前述之人名辨識類似的作法來提高精確度。

1.4 論文架構

本論文第二章「使用資料及工具介紹」將說明論文中會使用的空間資料及處理方式、實驗文本、建構系統工具。第三章「現有地理資訊系統技術回顧」，會介紹現有的 GIS 技術及被使用的情形，以及本文所採取的技術使用方式，第四章「歷史文件地名標註系統」，將詳述本論文的設計概念、整體架構及實作方式，並透過實例操作講述整個文件的標註過程，並作分析討論。第五章「結論與未來工作」將總結本論文，闡述其所解決的問題與其貢獻，再延伸討論本研究未來可發展的方向。



Chapter 2 使用資料及工具介紹

本文目標文本為歷史文本，故使用的地理資料將以歷史地名為主，而使用的地圖則為 Google Map 所提供的世界道路圖、地形圖、衛星圖做為呈現的底圖。除此之外，為了建立自動地名標註亦需要使用實際的歷史文本作為實驗文本，以下將介紹本文所使用的空間資料、實驗文本、以及建立使用者界面的 GIS 工具。

2.1 中華文明之時空基礎架構

網址：<http://ccts.sinica.edu.tw/>



Fig. 2-1 中華文明之時空基礎架構

中華文明之時空基礎架構(CCTS)由中央研究院 GIS 中心所執行，期望建構以中國為空間範圍，並以原始社會迄今的中國歷史為時間縱深，以中國文明為內涵的整合性資訊應用環境。主要對象除以學術研究與教育為主的學者、專家、與教師外，亦希望能兼顧一般性的，以時間及空間為主的資訊管理、分析、整合與呈現等應用[8]。

CCTS 的基本空間圖資以譚其驤先生主編之《中國歷史地圖集》為主要的基礎，提供上古至清代，上下逾二千年的中國歷代基本底圖，並輔之以持續整理蒐集之各類歷史地圖、遙測影像等基礎圖資。《中國歷史地圖集》共有八冊，內容主要是

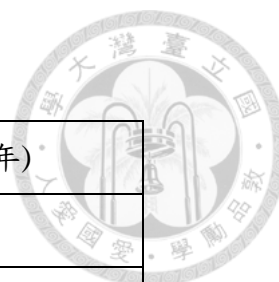
中國歷史上各時期中某個特定年份的一級及二級政區圖，而幾乎沒有對歷史事件的反映。共有圖 304 幅。以清代為例，有嘉慶二十五年(1820)及光緒三十四年(1908)兩幅地圖。



Fig. 2-2 《中國歷史地圖集》1820 年清代地圖

本文採用 CCTS 歷代地名圖層，內容為西漢至清代的基本空間圖資，共 38545 筆資料，各朝圖資採用年代如 Table 2-1 所述，本次實驗使用清代 1820 年之點坐標資料，每筆資料包括地名名稱，朝代時間帶，國家行政區，以及點坐標資料如 Fig.2-3 所示。坐標系統採用 WGS84，適用於常見的地圖系統，不需再行轉換。

本研究自中央研究院申請到此基本圖資之 Excel 檔案，透過程式將每筆資料匯入至 MySQL 資料庫做為本研究之空間資料庫，使用者將透過用戶端之 UI(User Interface)傳送要求至伺服器端程式，伺服器端再從空間資料庫中取得資料回傳至用戶端 Google Map API 呈現。



朝代	圖層時間(西元年)
西漢	西元前 7 年
東漢	140
三國	262
西晉	281
東晉	382
南北朝	497
隋	612
唐	741
五代十國	943
北宋	1111
南宋	1208
元	1330
明	1582
清	1820

Table 2-1 CCTS 歷代地名圖層年代對照

gid	NAME	begin	end	region	X	Y
38401	平政	1644	1911	清_廣東	114.907	22.8449
38402	和平	1644	1911	清_廣東	114.941	24.4385
38403	嶺岡	1644	1911	清_廣東	114.942	24.6273
38404	馴雉	1644	1911	清_廣東	114.965	23.4187
38405	官塘寨	1644	1911	清_廣東	114.967	23.5004

Fig. 2-3 空間資料庫資料範例



2.2 使用史料-《清實錄》

《清實錄》為清朝歷代皇帝統治時期之大事紀，紀錄了清代政治、經濟、文化、軍事、外交等各方面內容。

本論文針對歷史文件地名做自動化標註，採用《清實錄》做為所採用的實驗文本，其原因在於：第一，《清實錄》共 3647 萬 5317 字，數量龐大且完整，涵蓋清太祖至宣統三年，首尾完整；第二，此文本為官方紀錄，結構嚴謹，撰寫規則明確，內文編寫方式有一定結構，變異性不大。此二特性對於本研究的實驗較為適合，故採用《清實錄》作為實驗文本。

資料內容為台大數位典藏與自動推論實驗室所建構之「清實錄資料庫」，此資料庫之全文內容來源為中央研究院歷史語言所之「漢籍電子文獻資料庫」，自其中擷取內文及詮釋資料，資料庫以每日的個別事件定義為一條目，共有 325941 條。本研究取此資料庫「康熙朝」33100 條、「乾隆朝」88777 條之日期、標題、內文等三個欄位匯出成文字檔案，用於系統之實驗及訓練。

39-2
<p>巳時也。先是、 孝康章皇后、詣 慈寧宮問安。將出。衣裾若有龍繡。 太皇太后見而異之。問知有娠、顧謂近侍曰、朕 曩孕皇帝時、左右嘗見朕襖褶間、有龍盤旋、赤光燦爛。後果誕生聖子。統一寰區。今妃亦 有此祥徵。異日生子、必膺大福。至 上誕降之辰。合宮異香、經時不散。又五色光氣、充溢庭戶、與日並耀。是時、宮人以及內侍、無 不見者。咸稱奇瑞云。 上天表奇偉。神采煥發。雙瞳日懸。隆準岳立。耳 大聲洪。徇齊天縱。稍長、舉止端肅。志量恢宏。語出至誠。切中事理。讀書十行俱 下、略不遺 忘。自五齡後、好學不倦。丙夜披閱、每至漏分。凡帝王政治、聖賢心學、六經要旨、無不融會 貫通、洞徹原委。至孝性 成。繼志述事。仰承 太祖 太宗肇造鴻基、以守兼創。追念</p>
40-1
<p>世祖章皇帝耿光大烈、孺慕終身。奉事 太皇太后 皇太后、竭誠盡敬、歷久彌殷。大德好生、民物在 宥。勵精求治、日理萬幾。六十餘年、孜孜如一日。戶口繁增。風俗淳美。遠邁唐虞之 世。料敵 制勝、廟算如神。闢前古未闢之封疆。服從來 未服之方國。巡閱河工、指授方略。淮黃底定。世賴平成。且多藝多能。允文允 武。著作則上 擬典謨。吟詠則直追雅頌。精嫻細楷。妙擅擊 篲。挽弓十五鈞。用矢十三握。左右騎射、發必 中的。仁至義盡。久道化 成。如天之 鑄。如地之 載。如日月之照臨、雨露之濡潤。蓋泰運光昌。世當極治。故薦生 聖人、以贊化育。盛德大業、冠於百王。景福遐齡、超於萬禩。六齡時、嘗偕 世祖皇二子福全、皇五子常寧、問安宮中。 世祖各問其志。皇五子甫三齡未對。皇二子以 顯為賢王對。</p>

Fig. 2-4 漢籍電子文獻資料庫《清實錄》數位化全文

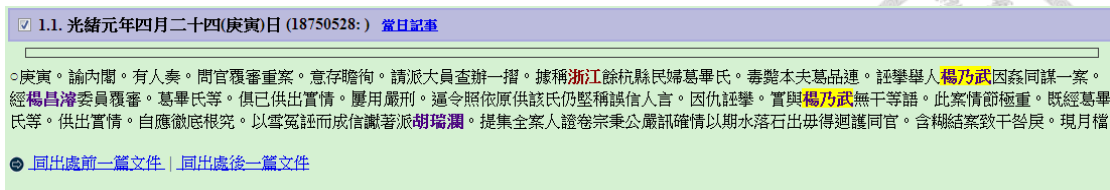


Fig. 2-5 《清實錄》資料庫條目內容

2.3 系統建構工具-Timemap

此工具為一 Javascript library，用來幫助建構使用網路地圖，整合 SIMILE Timeline 時間軸工具及 Google Map、Openlayers、Bing 等地圖服務，能讓使用者讀取 JSON、KML 或 GeoRSS 資料並同步顯示於時間軸及地圖上，將時間軸可視範圍內的條目呈現於地圖上。本研究採用此工具來建立使用者介面，原因在於其提供了許多方便的函式讀取資料與操作地圖，且具有高度的彈性使開發者能輕易加入其他功能。其整合時間軸與地圖呈現的能力，與本研究結合歷史文件與地理資訊之特性相符。本系統將 HTML、AJAX、PHP 等技術結合此工具用來建構整個用戶端的系統。

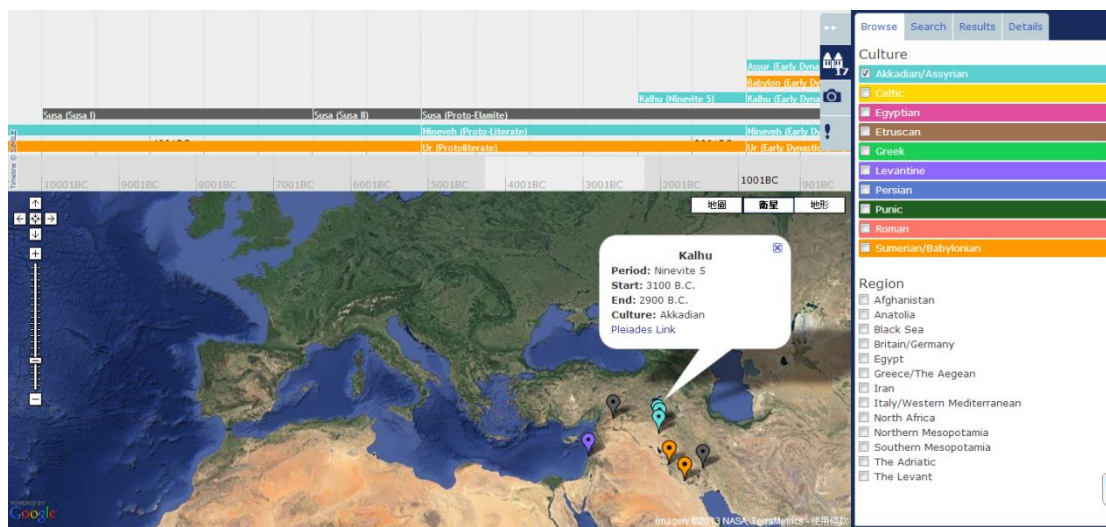


Fig. 2-6 應用 Timemap 建立 WebGIS 之範例



Chapter 3 現有地理資訊系統技術回顧

3.1 地理資訊系統(GIS)簡介

GIS 包含兩種資料型態，向量式地圖(vector)與網格式地圖(raster)，其差別在於表示方式不同，向量式資料不具連續性；而網格式資料為連續性，前者可以儲存地理的各種複雜關係，利於定位及網路分析等處理；後者則適合表達衛星影像等具地形變化的地圖資料。本研究使用向量式地圖為主，對於地理物件來說，向量式地圖包含兩種資料—空間及屬性。空間資料為利用點、縣、面於特定坐標系統進行定位，表達其地理位置；屬性資料為描述空間資料的特性，如：點的名稱、數量，面的行政區、面積等數據。

現有許多資料庫系統支援空間資訊格式，例如：PostSQL、MySQL 等，利用表格將空間資訊紀錄於其中，將空間資料與屬性資料透過使用者介面或軟體輸出給使用者運用。


3.2 常用 GIS 軟體與工具

3.2.1 伺服器 GIS 軟體(Server GIS)

Server GIS 可將大部分的 GIS 圖形計算、繪製於 Server 端完成，透過 HTTP 協定傳送至用戶端，用戶端僅需要處理使用者的輸入及地圖的顯示，能有效減少用戶端的效能負擔，此種 GIS 軟體可快速地分享 GIS，有許多網頁地圖應用採此方式開發。代表性的軟體有：ERSI(Economic and Social Research Institutut，美國環境系統研究公司)公司開發之 ArcIMS、ArcGIS Server，開放式軟體的 MapServer、GeoServer 等。

3.2.2 個人電腦 GIS 軟體(Desktop GIS)

此類型 GIS 軟體為專業地理資訊人士將地理資訊與知識加以整合、處理、使用的主要平台。代表性的軟體有：ERSI 公司的 ArcGIS 產品系列，依照應用之不



同分為許多版本，包括 ArcReader、ArcView、ArcEditor 等；另一代表性的軟體為 Quantum GIS，為自由軟體的個人電腦 GIS 軟體，又稱 QGIS，由於其開放原始碼的原因，可以被開發者修改以執行其他 GIS 任務，亦有許多開發者開發出各種擴充的功能套件擴展其功能。

3.2.3 網頁地圖(WebGIS)

WebGIS 為基於網頁所開發之 GIS 系統，使用者可透過瀏覽器直接使用，現今瀏覽器技術日漸進步，在網路上直接操作 GIS 的限制愈來愈少，相較 Desktop GIS 軟體，WebGIS 的優點在於：1.系統的泛用性，即使在不同的作業系統上，只要瀏覽器軟體支援，使用者皆可以輕鬆使用。2.平台不再被侷限於桌上型電腦，在現今智慧型手持裝置如此流行的情形下，無論是手機或平板電腦，亦能透過瀏覽器使用 WebGIS。

目前有許多大型網站提供地圖服務，如：Google Map、Bing Map 等，除了提供一般使用者瀏覽地圖、影像外，通常也提供 API 的服務，使開發者能夠用來開發地圖服務。以本研究為例，便使用了 Google Map Javascript API v3 做為開發平台。其他還有 Bing Map API、Openlayers API 等 API 也常被開發者所使用。

3.3 GIS 標準格式

OGC(Open Geospatial Consortium，開放地理空間組織)是依國際性非營利組織，制定了許多 OpenGIS 標準，例如如何表達點、線、面的空間資料，藉由這些標準讓複雜的空間資料與服務能夠供技術開發人員所使用。

3.3.1 Shapefile

由 ESRI 所制定的一種空間數據格式，Shapefile 文件用於描述空間資料，如道路、村莊等物件的地理位置。亦可用來儲存這些物件的屬性資料，如長度、名稱等。此種格式被普遍使用在地理資訊軟體界中，成為一種開放標準，絕大部分的 GIS 解能使用此種標準。



3.3.2 KML

KML(Keyhole Markup Language)為基於 XML(eXtensible Markup Language)語法標準來交換地理資訊，由 Keyhole 公司發展並維護，應用於 Google Map、Google Earth 等軟體。目前有愈來愈多 GIS 軟體支援此格式。

3.3.3 GeoJSON

GeoJSON 為一種基於 JSON(JavaScript Object Notation)的開放格式，用於表達地理資料結構，此格式可被 JavaScript 直接當作 JSON 物件使用，因此使用 JSON 物件的工具通常也能被使用在 GeoJSON 上。此格式讓地理資訊能被儲存在一個緊湊的資料格式中，支援點、線、面的空間資料，能方便的在 WebGIS 上使用。

本研究基於介面多採用 JavaScript 編寫，故傳送地理資訊之檔案格式即採用 GeoJSON。



Chapter 4 歷史文件地名標註系統

4.1 系統概述

本系統目的在於讓研究者不需熟悉 GIS 技術或資訊技術，亦能夠將古文書中的地理資訊找出並對應至地圖上。此研究假設使用者文本為《清實錄》中含地理資訊的條目，地圖則以 Google Map 提供之道路圖呈現。藉由此系統，將可呈現《清實錄》中各條目的時空資訊。

在本系統建構的過程中，首要解決的問題為如何擷取文章中的地理資訊，需要考慮的因素包括：

1. 比對的效能：

此系統以 WebGIS 呈現地圖，對使用者而言使用上與一般網站無異，因此系統的針對使用的的需求所需的反應時間將影響使用者使用此系統的意願。

2. 標註地名的準確性：

在[3]中提到地名辨識的難點主要表現在(一)地名數量大，且沒有明確規範的地名定義，不斷湧現新的地名；(二)地名長度並無限制，不似人名多為 2~4 漢字；(三)地名中內部相互成詞。而目前中文地名辨識的趨勢為採用統計模型，以及採用統計及規則相結合的方式。但在本研究中，目的在於標註出文章中在空間資料庫具有坐標資訊的地名，而非找出文章中新的地名，且為了維持效能，亦無法使用複雜度太高的比對方式，而造成反應時間下降。因此本研究採用詞庫式的標註方式，將空間資料庫中的地名清單當作詞庫，於文章中進行比對，比對出的關鍵詞即為地名資料。如此做法下可得到極好的召回率，但相對的精確度卻不高，比對出的地名可能在此文章中並不含有地名的意義在。因此，再透過使用上下文統計的詞夾子作為一地名驗證的工具，企圖取得有較好精確度的標註結果。

第二個需要解決的問題則在於如何呈現結果給使用者觀察：



1. 直覺式的使用介面：

為了讓使用者能了解數位化工具的使用，不讓使用者有太大的負擔，本研究目標讓使用者只需要透過少數操作即可得到預期目的。

2. 人工校正系統標註的結果：

因標註結果為利用資訊技術標註，而資訊技術擷取的內容無法達到 100% 準確率。所以系統提供一校正的介面，讓使用者能夠校正標註後的結果。

4.1.1 系統架構

系統分為三個部分，分別為用戶端、網路伺服器端、Google Map 伺服器端。用戶端以 AJAX(Asynchronous JavaScript and XML)為核心，負責處理資料之讀取、新增、地理資訊變更，並利用 Google Map API 操作地圖及呈現地理資訊。使用者透過 UI(User Interface)新增文本資料並提交後，文本經由 AJAX 傳送至網路伺服器端，伺服器端透過 PHP 函式處理，從空間資料庫取出地理資訊進行地名比對及標註程序，此程序會在文本中的地名詞彙加上標記，將此標記後的文本回傳至用戶端及文本資料庫，用戶端結合標記文本及 Google Map API 及 Timemap API 呈現一個可讓使用者觀察及操作地理資訊的介面。使用者亦可透過 UI 作文本的檢索，用戶端亦透過 AJAX 向伺服器的 PHP 函式要求資料，PHP 函式則依照參數向文本資料庫傳送 SQL 取得符合的標記文本，將文本及地理資訊轉為 Timemap API 可讀取的 json 檔，並傳回用戶端呈現。

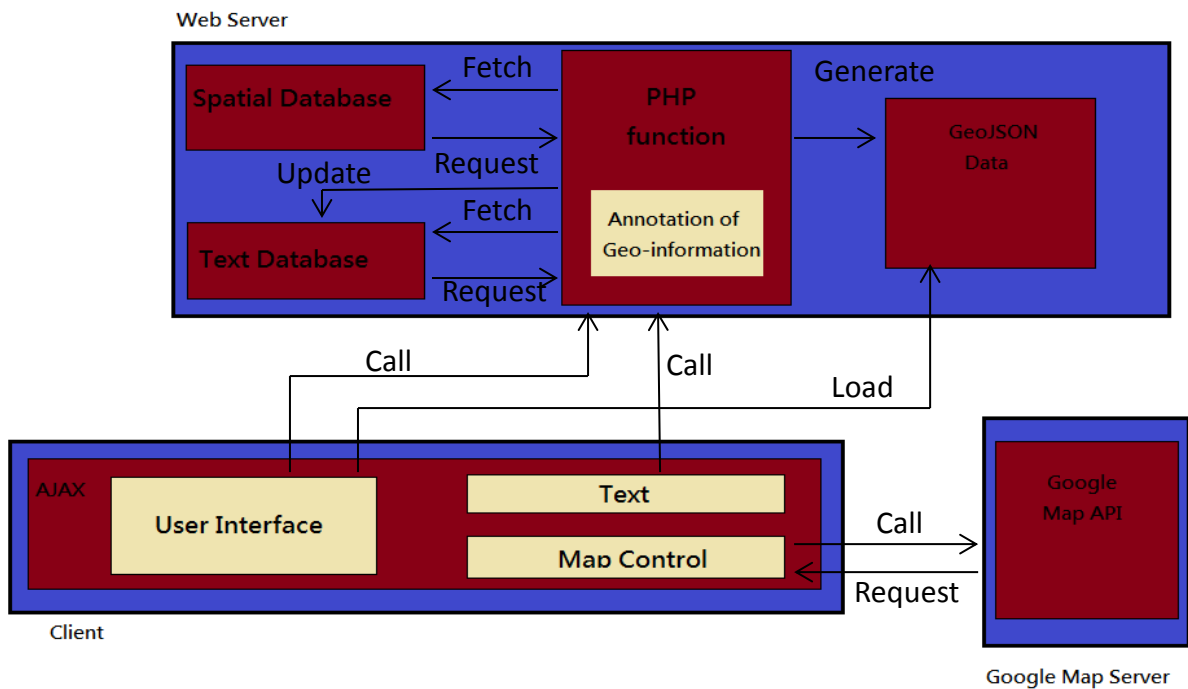


Fig. 4-1 系統架構圖

4.2 地名標註流程

本研究一重要的目的為標記出使用者文章內的地名並加入坐標，因此自動化標記地名的準確度乃為此研究一重要的面向，若自動標記準確度低落，需要花費大量人工去校正這些標記，因此標記的準確度為一重要的考量，以下將介紹此研究在使用者提交文本後如何自動化標記出地名及地理資訊的流程。

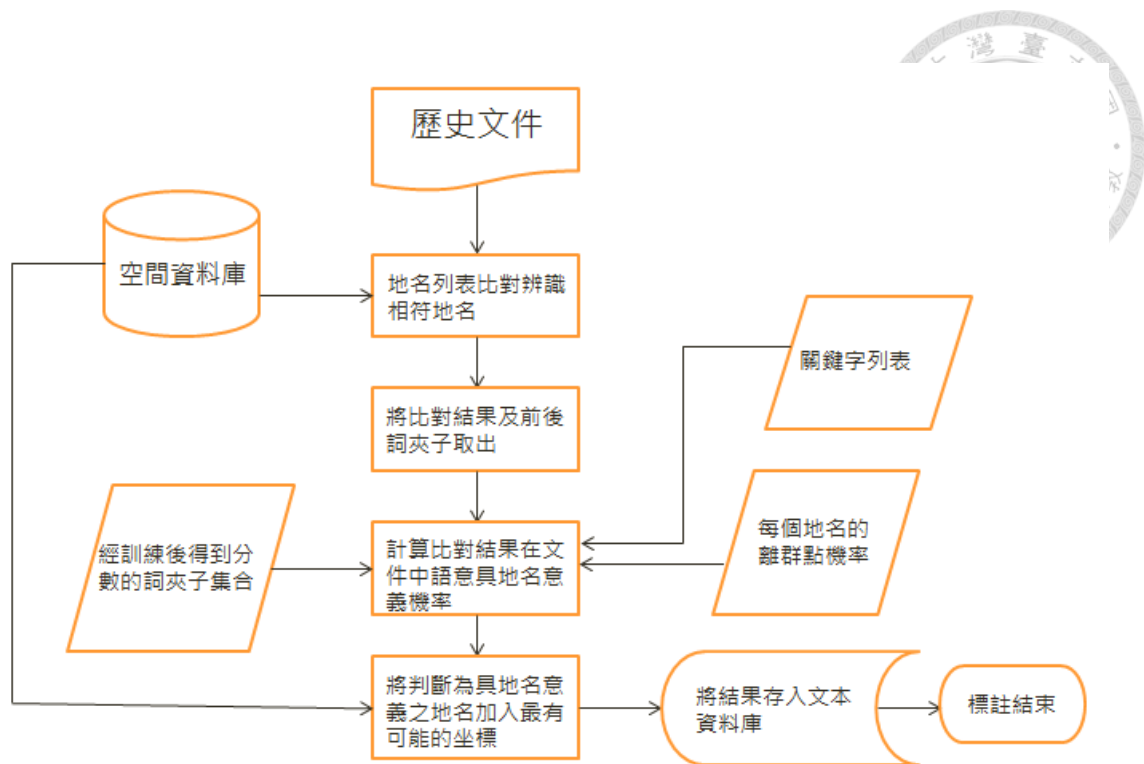


Fig. 4-2 地名標註流程圖

4.2.1 詞庫式地名辨識

本研究需要從文章的地名中找出文章的各個坐標資訊，因此地名辨識目標僅在於辨識出文章中所有出現在空間資料庫中的地名而不在意不存在於資料庫中的地名，因此採用詞庫式的辨識方法，可以擁有較快的執行速度及最高的召回率。[3]

系統辨識方式為根據字數及地點名稱，當文章被提交後，系統從空間資料庫中依照名稱字數由多到少逐一於文章中比對是否有出現相同名稱。規範須由字數多到少的原因在於若先辨識字數少的地名，若此地名為另一長地名中的一部分，則長地名便無法被辨識到。舉例而言：當辨識”庫倫伯勒齊爾”此長地名時，若先辨識到”庫倫”此短地名，則後面的”伯勒齊爾”將不被視作地名，則”庫倫伯勒齊爾”此地名將不被辨識出來。透過詞庫式的辨識方法，可用很高的召回率找出文章中的地名。

4.2.2 驗證比對地名

就前所述，辨識地名的準確度為此研究一重要面向，詞庫式地名辨識雖有快

速及高召回率的優點。但其精確度卻僅是普通，因此必須用其他方法來驗證經詞庫辨識後的地名在文章中是否含有地名的意義。

在此研究選用詞夾子作為驗證的方法。根據[3]，詞夾子為一種根據詞的前後文統計的專有名詞辨識演算法，優點在於與其他統計式的方法相比，詞夾子不需要大量的語料庫及辭典即可使用，而本研究目標為標註歷史文件，目前在歷史文件上做名詞辨識並不多，亦缺乏適合的語料庫及辭典，因此詞夾子演算法十分適合使用於本研究。

所謂詞夾子即在文本中，目標詞的前後數個字。在本研究中，以地名當作目標詞，以前後兩字作為詞夾子，舉例而言，「得旨，崇明縣知縣陳慎」一條中「崇明」為一地名，則「旨，」、「縣知」則為其詞夾子。此方法為針對比對後的地名，取出此地名的詞夾子，辨識此詞夾子是否為容易夾住地名的詞夾子。藉以計算地名在文件中語意具有地名意義的機率

詞夾子演算法雖然由於其無法辨識不曾被詞夾子夾住的專有名詞造成其召回率不高，但只要是常夾住名詞的詞夾子，當新詞被此詞夾子夾住時此新詞為專有名詞的機率很高，使演算法的精確度高，本研究便利用其精確度高的特點，當執行完詞庫式辨識後提高了召回率，再利用詞夾子驗證提高地名辨識的精確率。

為了使用詞夾子來驗證，必須先訓練出一群詞夾子並幫每個詞夾子計算分數，透過這些分數來判定哪些詞夾子為常夾到地名的詞夾子。本研究以《清實錄》為例，採用《清實錄》中字數最多的乾隆朝作為訓練文本。此文本共有 88777 條目，共 14,060,395 字。詞夾子的訓練流程如下：

Step1： 將所有樣本詞於本文中的位置找出。

Step2： 將本文中夾種樣本詞的詞夾子收集起來。計算每個詞夾子夾中樣本詞的個數。

Step3： 將上一步驟收集的詞夾子，計算每個詞夾子於本文中夾中的詞的個數。

Step4： 計算每個詞夾子夾中樣本詞的機率： $\frac{\text{詞夾子夾中樣本詞的個數}}{\text{詞夾子於本文中夾中詞的個數}}$ 作為詞夾子的



分數

此處的樣本詞即空間資料庫中的當代二至四字的地名，因文本為清實錄，故樣本詞為清代地名。透過此流程可得到一批具有分數的詞夾子，可用來驗證標註地名是否具地名意義。

而除了上下文規則外，地名本身被當作地名的機率也應該作為驗證的評估方式之一。此處導入離群點(Outlier)概念，所謂離群點在本文中定義為脫離地點群集的地點。在文本中提及的地點大部分而言應為相近的點，若有一地點明顯遠離其他被標註的地點，則此點在文章中有地名意義的機率不大。Fig.4-3 為「○工部議准、閩浙總督郝玉麟疏言、閩省灘河危險。舟楫維艱。請於建寧府浦城縣。至福州府古田縣水口。大灘七十六處。凡河心石塊。遍行鑿鑿。并修築緯道。立柱指迷。堅製官船八隻。豫備協救。以資利濟。從之。」一條的地理分布，其中「建寧府」、「浦城」、「福州府」、「古田」、「水口」皆為福建省之地名，「大灘」則為廣西之地名，藉由離群點的判斷，「大灘」在此條目中具地名意義的機率並不高。



Fig. 4-3 離群點例子

而若一地名時常被當作離群點，則此地名在各文章中具有地名意義的機率亦不大。基於此概念本文將每個地名是否為離群點的機率當作驗證地名的一個標準。首先，應先給予離群點精確的定義[7]：

給定坐標集合 D ，參數 k 為整數、 r 為半徑，若有一點 p 屬於 D ， p 以 r 為半



徑的範圍內若至多 k 個點，則 p 為該群集中的離群點。本文取 $k = \frac{|D|}{2}$ ， r 為 D 中各點的平均距離。

則演算法如下：

Neighbors(p): 與 p 距離不超過 r 之坐標集合

R: 離群點集合

Outlier(D, k, r)


```
{  R=0 //離群點集合
  for each  p ∈ D
  {
    for each  q ∈ D and q ≠ p
    {  if dist(p,q) ≤ r
      Neighbor(p)=Neighbors(p) ∪ {q}
    }
    if |Neighbor(p)| < k
      R=R ∪ {p}
  }
}
```

利用此方法，在每次完成地名標註後，做離群點的檢測，可得到每個地名被當作離群點的次數。將其除以被標記的次數，即可得到此地名在每次被標註到時，成為離取點的機率。而透過標記乾隆朝《清實錄》當作訓練，便可得到一批訓練後具有離群點機率的地名。

結合以上兩種方式，在驗證地名時，使用

$(1 + \text{詞夾子分數}) \times (\text{地名離群點機率}) + \text{特殊關鍵字分數}$

作為地名的評分標準。此特殊關鍵字分數為觀察文言文特性後，針對某些關鍵字對地名的特性作為驗證標準，即若地名之詞夾子包含營、府、鎮、廳、縣等行政



區資料，此地名擁有地名意義的機率極高。則直接將分數提高；或地名詞夾子包含「人也」、「將軍」等，此種詞夾子夾住的地名，通常與文章的地理位置關係不大，則擁有此種夾子的地名分數降低。而若地名的詞夾子不屬於被訓練過的夾子，將其分數設為 0.5，以分數超過 1 者認為是地名。

此驗證方法，以順治十八年正月九日至順治十八年九月十九日間，具有地名的條目 100 條作為實驗文本，若將精確度定義為 $\frac{\text{確實具地名意義之地名}}{\text{驗證具地名意義之地名}}$ ，則精確度由純使用地名列表比對的 68.9% 提升至 77.6%，可看出此方式有達到過濾掉不具地名意義之地名之目的。

4.2.3 匯入空間資訊

透過上一節的程序後，已辨識出使用者文本中具地名語意的地名，然僅找出地名是無法被 Google Map API 使用的。必須透過空間資料庫將每個於文章中的地名的坐標位置點出，才能夠於地圖上被點出。

於空間資料庫中，地名名稱不免有重覆的問題，也就是同一個時代內有同一個地名卻代表不同地點的情形。如清代地名”太平”便有七個坐標位置。系統遭遇此種有同名異地情形時，會將資料庫中所有此地名的坐標於使用者介面列出，讓使用者可以自行選擇正確的坐標資料，而隨著使用者更改坐標，系統呈現的地標位置也隨之不同，透過此互動介面，可使系統標註地名的彈性增加。



高宗純皇帝實錄(一)_乾隆元年十二月八日
○戶部議覆、大學士管浙江總督巡撫事嵇曾筠疏
言、1.嘉興府之用里街、柴場灣、2.台州府之關
嶺、3.青溪、4.溫州府之雙溪、5.處州府之6.青
田、7.碧湖、8.太平、均溪、9.安仁、小梅、住
溪、10.皂口、等處。皆地非市鎮。稽察難周。所有落
地稅。請概予豁免。應如所請。從之。

太平

- 太平#清_山西#POINT(111.285 35.81766)
- 太平#清_安徽#POINT(118.2022 30.32077)
- 太平#清_浙江#POINT(121.3622 28.36243)
- 太平#清_湖南#POINT(113.691 25.39669)
- 太平#清_廣東#POINT(110.0847 19.25702)
- 太平#清_廣東#POINT(111.599 21.73015)
- 太平#清_廣東#POINT(113.8402 24.58818)
- 非地名

確認

Fig. 4-4 同名異地處理情形

為了盡量減少使用者花費人工去修正此同名異地情形，預設地名的選取是有其重要性的，若能夠每次系統標記完後，系統預設選取的地點便是於文章中如使用者所預期的地點，對使用者而言可減少花費的時間，而若預設地點並非使用者所預期，使用者也能透過方便的介面做調整工作，如此便不會對使用者觀察文本有不利的影響，而增加了許多便利性。是以系統設計了一個判斷預設坐標的功能。

根據觀察《清實錄》的經驗，當文章中有多個地名出現時，通常這些地名的地點通常會是同一個省份或是鄰近區域的地點，由此判斷利用坐標之間的距離作為決定具有同名異地性質地名的預設坐標，是一個簡單且貼切的方式。在此使用經緯度來計算坐標間的距離，因此必須注意經緯度的距離必須考量地球為一橢圓球體而非平面，在計算距離時用的是球面而非平面坐標，計算兩個坐標間的最短距離是計算球面上兩點間的最短距離，即大圓距離(The Great Circle Distance)由於本研究目標在於標註古文本，在缺乏當代道路資料的情形下，並不考慮當代的實



際道路及交通情形，而是取最短的大圓距離當作兩點間距離。本研究採用 Haversine Formula 來計算兩點間的大圓距離，式子如下所示：

d:大圓距離

r:地球半徑

(ψ, λ) : 點的經度, 緯度

$$\text{havarsin}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos(\theta)}{2}$$

$$\begin{aligned} d &= 2r \arcsin\left(\sqrt{\text{havarsin}(\phi_2 - \phi_1) + \cos(\phi_1) \cos(\phi_2) \text{havarsin}(\lambda_2 - \lambda_1)}\right) \\ &= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1) \cos(\phi_2) \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \end{aligned}$$

匯入坐標實際的做法為：

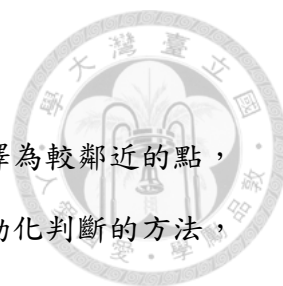
Step1:將進行地名辨識後的文章，針對各地名從資料庫中抓出對應的坐標點並存入該地名的陣列，陣列長度代表此地名的坐標數。

Step2:檢查文章中的地名數量，若僅有一地名則將此地名預設坐標設為陣列中第一筆坐標

Step3:檢查是否有陣列長度大於一的地名，若無則將所有地名之預設坐標設為該地名陣列中的坐標並結束此次空間資訊匯入。

Step4:針對陣列長度大於一的地名(稱 P)，若文章內有陣列長度為一的地名，則找出文章順序中此地名前後最接近(即相隔最少地名)且陣列長度等於一的地名(稱 P1、P2)，對 P 中每個坐標計算與 P1、P2 的大圓距離，取距離最短者為預設坐標。若無陣列長度為一的地名則跳至下一 Step。

Step5:若文章不含有陣列長度等於一的地名，則將文章開頭算起前兩個地名(稱 P3、P4)，將此兩地名陣列中的坐標彼此計算大圓距離，取具有最短距離的兩坐標設為 P3、P4 陣列內唯一坐標將其他坐標捨棄，並將 P3、P4 的陣



列長度變為一，回到 Step3。

經過此程序，在設定地名之預設坐標時，能將坐標盡量選擇為較鄰近的點，較有可能為文章中預期的地點。然就前述言，透過資訊技術自動化判斷的方法，是無法達到 100% 的精確度，勢必需要人工的校正，在本系統中，提供了讓使用者在發現系統的預設坐標並非如使用者所預期的坐標位置，可透過點擊地名，更改呈現的坐標。

透過地名比對、比對地名驗證、匯入空間資訊這三個步驟後，系統擷取出使用者提交的文章中的地理資訊，接著便可將此含地理資訊的文件存入資料庫中此使用者的表格，使用者可以透過由 Timemap API 及 Google Map API 所建立的 WebGIS 界面存取表格，將文章呈現於 WebGIS。

4.3 系統功能

系統最終目的在於讓使用者能夠透過直覺式的操作，便可以觀察到提交文本基本的地理資訊，呈現於該文本事件中具有意義的地理位置。所謂直覺式的操作，就像現代人廣泛使用智慧型手機的介面一樣，不需要閱讀冗長的使用說明，僅須要透過簡單的按鈕、選取、拖曳即可達到如使用者所預期的結果，而不會讓不熟悉使用資訊工具的使用者感到負擔，若不如此，讓使用者必須花費大量的時間熟悉使用系統功能，恐怕造成讓使用者對系統望之卻步。是以本系統功能的設計的宗旨為讓使用者能盡量減低操作的負擔，便可達到預期觀察到的結果。目前系統提供的功能如下：

1. 提交文章到文本資料庫中。
2. 對文本資料庫中的文章可做編輯及刪除。
3. 透過地圖及時間軸瀏覽文本資料庫中的文章。
4. 針對文章做全文檢索，呈現於時間軸及地圖上。
5. 當上傳不同文集，可分別瀏覽不同文集的文章，亦可一起瀏覽。

6. 對文章中被系統自動標註且加入坐標的地名，做坐標的更改。
7. 對文章中沒有被標註的地名做人工標註。

以下將介紹系統提供給使用者操作的功能及設計理念。



4.3.1 提交、新增、刪除文章至文本資料庫

系統的目的，在於幫助使用者標註出提交文章中的地名，系統目前提供兩種方式讓使用者提交文章給網路伺服器端，一是透過點選介面中的「新增文章」按鈕，讓使用者填寫出現的表單內容，如 Fig. 4-3 所示。此種方式近似於目前網路上許多人使用的微網誌的填寫方式，使用者僅需要填寫文章的詮釋資料(metadata)及全文，包括標題、起始、結束日期、分類、內文便可以將文章提交給伺服器端。當伺服器端將地名標註完並加入坐標後便會呈現在 Timemap 上。

127.0.0.1/gis/timemap.2.0.1/project/upload.html - Google Chro...

127.0.0.1/gis/timemap.2.0.1/project/upload.html

分類:

標題:

開始時間:

結束時間:

內文:

XML文件:



Fig. 4-5 利用填寫表單提交文章

這種方式讓使用者使用簡單且直覺的方式提交文章，不需要有任何資訊技術的概念。但此種方法不適用於使用者想提交大量文章的情形，若使用者想透過此方式提交多筆文章，需要一筆一筆的填寫表單，較為費時，系統提供第二種方式讓使用者能夠一次性的新增多筆文章，即透過上傳一份 XML 格式的檔案，如 Fig. 4-4 所示，讓伺服器端去分析檔案的結構，將 XML 內含的多筆文章標註地名入坐標後透過 Timemap 呈現給使用者。



```
<?xml version='1.0' encoding='utf-8'?>
<DocList>
<Doc>
<Topic>清實錄</Topic>
<Title>高宗純皇帝實錄(一)_雍正十三年九月五日</Title>
<Begin>1735-10-20</Begin>
<End>1735-10-20</End>
<Content>
○諭總理事務王大臣。前奉皇考諭旨。查郎阿回到西安之後。著史貽直來京。今因部中辦事需人。
</Content>
</Doc>
<Doc>
<Topic>清實錄</Topic>
<Title>高宗純皇帝實錄(一)_雍正十三年九月八日</Title>
<Begin>1735-10-23</Begin>
<End>1735-10-23</End>
<Content>
○又諭。甘省百姓。連年輓運軍需。荷蒙皇考聖恩。將該省應徵錢糧。連年蠲免。其本年錢糧。亦
</Content>
</Doc>
</DocList>
```

Fig. 4-6 XML 檔案的樣式

此方式讓使用者可以不需要重複的填寫表單，而能夠僅上傳一份檔案便能提交多筆數的文章。使用 XML 檔案格式，讓系統可以簡單的分析出每筆文章的詮釋資料(metadata)，然而對使用者卻有一定的技術門檻，需要有撰寫 XML 檔案的能力，但透過系統提供的範本，應可以降低這方面的門檻。

除了新增文章外，系統亦提供介面給使用者針對資料庫中的文章做編輯及刪除，方法為透過瀏覽文章時點選介面上的「編輯」、「刪除」按鈕，一樣透過直覺式的操作便可以完成。

4.3.2 過濾文章、時間軸呈現

在新增文章時，之所以要輸入詮釋資料，是系統希望能夠幫助使用者在瀏覽文章時，能更方便快速的找到想要觀察的文章，使用者在新增文章所填入的「分類」、XML 檔案中的 Topic 標籤便是用來將文章分類，透過系統的功能選單，可以選擇想呈現的分類。而系統也提供全文檢索的功能，此全文檢索的範圍為伺服器端裡使用者的文本資料庫中全部的文章的內文及標題，此功能搭配文章分類，可



在使用者觀察史料時提供過濾的功能。

而在 HGIS 中，「時間」要素是不可或缺的，系統以 Timemap 呈現，將時間軸放在最顯眼的位置，目的為讓使用者能夠一眼發現目前選取文章的時間點為何，時間軸分為三層級，分別為天、月、年，能夠清楚觀察整個分類中的文章在時間軸上的分布，同時能夠直接拖曳時間軸，改變呈現文章的時間帶，亦能夠直接輸入「西元年-月-日」來選取時間帶。以觀察選取時間帶內的文章。

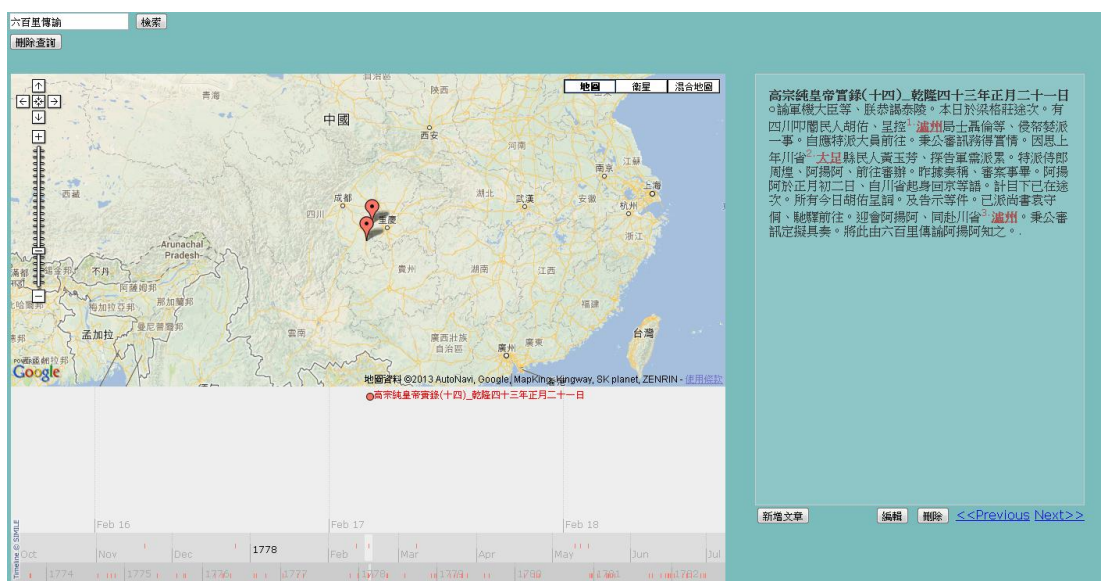


Fig. 4-7 以系統檢索含「六百里傳諭」的文章

呈現的介面如 Fig. 4-5 所示，分為三個部分，地圖、時間軸、文章。當選取時間軸上的標題或地圖上的標示時右邊的頁面會將此篇文章全文呈現，當沒有目標被選取時，地圖會呈現該時間帶內所有文章中地名的地標，當移動時間帶時地圖呈現的地標也會跟著改變。若選取目標文章後，會將其他文章之地標隱藏直到再次移動時間軸為止，使用者亦可以透過右下角的按鈕直接選取時間軸順序中前一個或後一個的文章，而不需移動時間軸，當檢索的關鍵字文章分散在較廣的時間軸時此功能會較方便於觀察。



4.3.3 校正地名標註

在 4.2 節提到當系統對文章的地名標註並非使用者所預期的將由使用者自己對這些地名標註做校正或新增。基於系統的設計理念，依舊不希望給予使用者過多負擔。透過介面的設計，目標是讓使用者能夠在閱讀文章的同時，便一邊能直觀做到校正及新增標註。

校正的流程如下：

1. 選取想更改的地名標註或反白沒被標註的地名。
2. UI 會從伺服器端空間資料庫讀取選取地名的所有坐標及行政區。
3. UI 會顯示該地名所有可能的坐標，並依使用者選取在地圖上顯示。
4. 使用者可自行選取想更改的坐標，或將該地名改為不標註。
5. 點下「送出」按鈕，UI 便會傳送更改後的標註回文本資料庫，且將改變後的結果呈現於介面上。

4.4 系統操作實例

4.4.1 工具操作流程

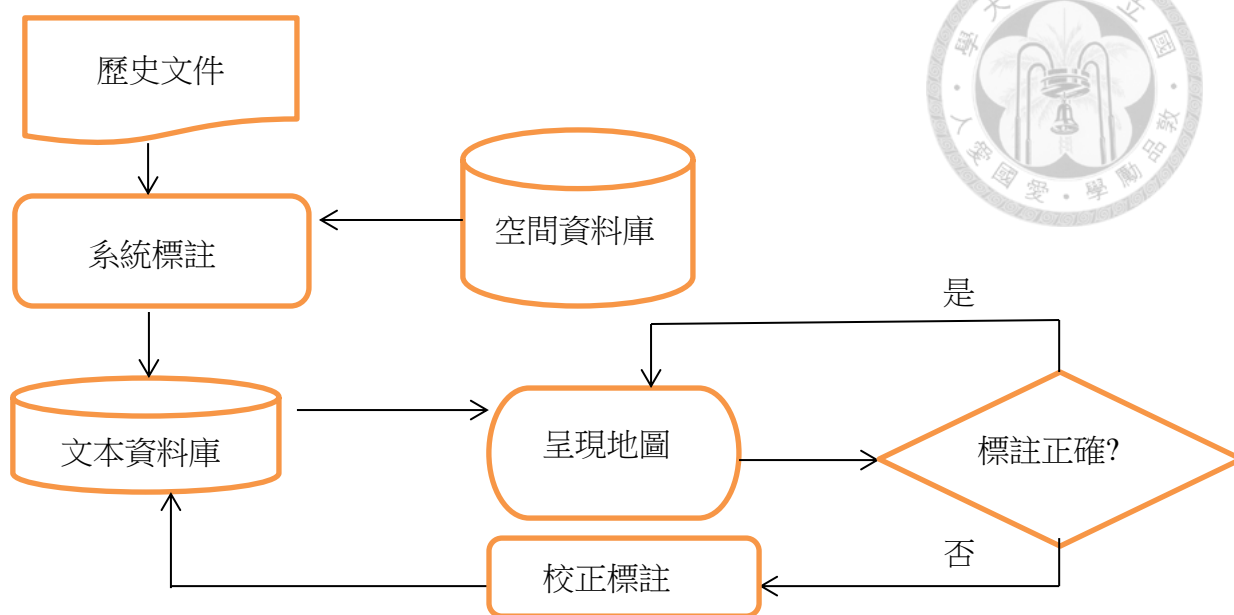


Fig. 4-8 系統運作流程

使用者進入系統後，如 Fig.4-7 所示，可選擇新增文件按鈕撰寫新文件，並按下送出，系統會透過 AJAX 將文章傳送至伺服器端進行地名標註，處理完後會將標註後的文章送回到用戶端呈現，會將標註後的地名以紅色字標記，並顯示於地圖上。若使用者欲修改標註的地名或坐標可透過反白選擇地名或點選地名以觸發修改視窗進行修改，修改完畢後會傳送至文本資料庫中進行更新，並將修改後之結果顯示於地圖上。重覆此動作，即可將上傳文本的地理位置標註清楚。

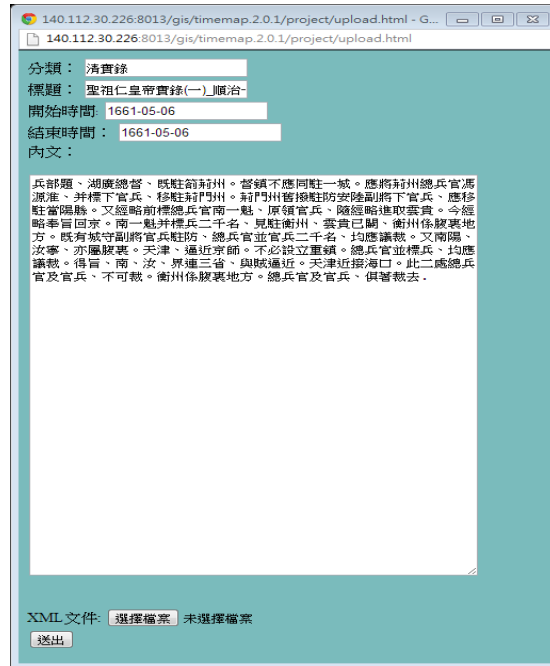


Fig. 4-9 新增文章

4.4.2 標註操作實例-以《清實錄》文本為例

在此使用《清實錄》中康熙朝實錄裡於順治十八年四月八日之事件，作為展示本系統地名標註之實例。

原文件內容：

○兵部題、湖廣總督、既駐荊州。督鎮不應同駐一城。應將荊州總兵官馮源淮、并標下官兵、移駐荊門州。荊門州舊撥駐防安陸副將下官兵、應移駐當陽縣。又經略前標總兵官南一魁、原領官兵、隨經略進取雲貴。今經略奉旨回京。南一魁并標兵二千名、見駐衡州、雲貴已闢、衡州係腹裏地方。既有城守副將官兵駐防、總兵官並官兵二千名、均應議裁。又南陽、汝寧、亦屬腹裏。天津、逼近京師。不必設立重鎮。總兵官並標兵、均應議裁。得旨、南、汝、界連三省、與賊逼近。天津近接海口。此二處總兵官及官兵、不可裁。衡州係腹裏地方。總兵官及官兵、俱著裁去。

而在經過系統自動化標註後，會得到 Fig.4-8 之地圖及以下經標記後的文章：

○兵部題、湖廣總督、既駐筓荊州。督鎮不應同駐一城。應將荊州總兵官馮源淮、并標下官兵、移駐¹荊門州。荊門州舊撥駐防²安陸副將下官兵、應移駐³當陽縣。又經略前標總兵官南一魁、原領官兵、隨經略進取雲貴。今經略奉旨回京。南一魁并標兵二千名、見駐衡州、雲貴已闢、衡州係腹裏地方。既有城守副將官兵駐防、總兵官並官兵二千名、均應議裁。又⁴南陽、汝寧、亦屬腹裏。⁵天津、逼近京師。不必設立重鎮。總兵官並標兵、均應議裁。得旨、南、汝、界連三省、與賊逼近。天津近接海口。此二處總兵官及官兵、不可裁。衡州係腹裏地方。總兵官及官兵、俱著裁去。



Fig. 4-10 經自動化標註後之結果

可發現「荊門州」、「京師」、「天津」三地並未被標註，判斷其原因為地名驗證未通過。故應再透過系統提供之校正方法，反白未被標註之地名選取正確坐標，按下確認更新文本資料庫中的此文本。如 Fig.4-9 所示。

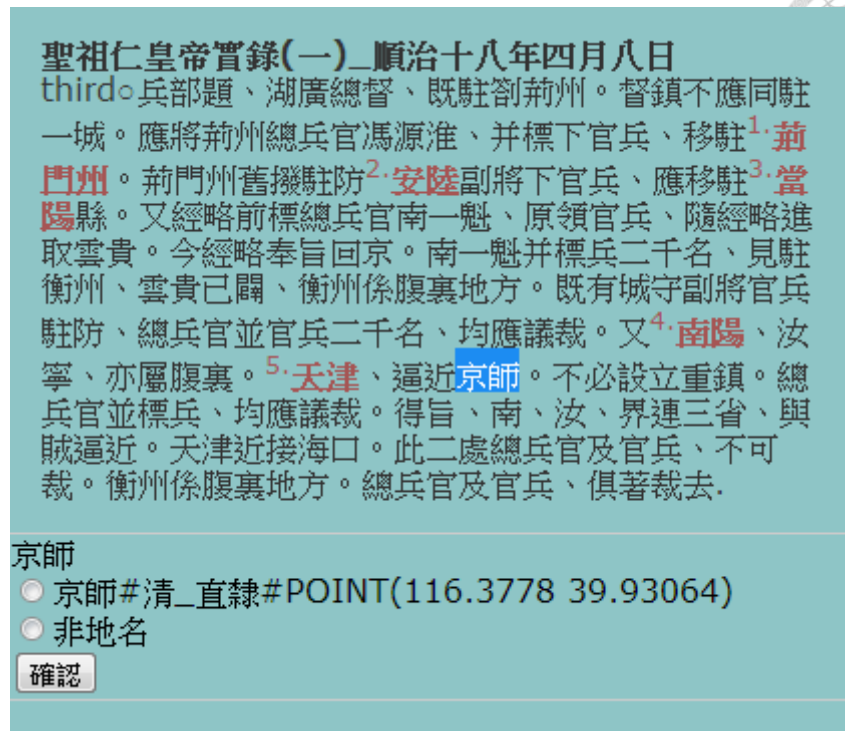


Fig. 4-11 校正標註

經過人工校正標註後地圖如 Fig.4-10，而校正後的文章如下：

○兵部題、湖廣總督、既駐劄荊州。督鎮不應同駐一城。應將荊州總兵官馮源淮、并標下官兵、移駐¹荊門州。²荊門州舊撥駐防³安陸副將下官兵、應移駐⁴當陽縣。又經略前標總兵官南一魁、原領官兵、隨經略進取雲貴。今經略奉旨回京。南一魁并標兵二千名、見駐衡州、雲貴已闕、衡州係腹裏地方。既有城守副將官兵駐防、總兵官並官兵二千名、均應議裁。又⁵南陽、汝寧、亦屬腹裏。⁶天津、逼近⁷京師。不必設立重鎮。總兵官並標兵、均應議裁。得旨、南、汝、界連三省、與賊逼近。⁸天津近接海口。此二官及官兵、不可裁。衡州係腹裏處總兵地方。總兵官及官兵、俱著裁去。



聖祖仁皇帝實錄(一) 順治十八年四月八日
 ○兵部題、湖廣總督、既駐劄荆州。營鎮不應同駐一城。應將荆州總兵官馮胤准、并標下官兵、移駐¹荊門州。²荊門州舊撥駐³安陸副將下官兵、應移駐⁴麻陽縣。又經略湖廣總兵官南一魁、原領官兵、隨經略進取雲貴。今經略奉旨回京。南一魁并總兵二千名、見駐衡州、雲貴已闕、衡州係保護地方。既有城守副將官兵駐防、總兵官並官兵二千名、均應議裁。又⁵南陽、汝寧、亦屬邊塞。⁶天津、逼近⁷京師。不必設立重鎮。總兵官並標兵、均應議裁。得旨、南、汝、界連三省、與賊逼近。⁸天津近接海口。此二官及官兵、不可裁。衡州係保護處總兵地方。總兵官及官兵、俱著裁去。

Fig. 4-12 校正標註後結果呈現




Chapter 5 結論與未來工作

5.1 結論

具有地理資訊的物件，例如古地契或官方檔案，這類文件的地理位置屬性相當有參考價值。但若不透過資訊技術，研究者若想觀察大量文件的地理分布時，必須花費大量的精神與時間，將文件逐篇的整理出地理位置，並標註於地圖上，才能夠對這些文件做地理資訊的觀察。本論文提出了針對使用者提交的各種歷史資料文件，利用 Text Mining 技術在一定準確率之上自動化擷取文件中的地名，並透過空間資料庫將這些地名加入坐標資訊。再加上使用 GIS 技術將這些文件在地圖上視覺化呈現，研究者只要再經過介面的互動介面校正因自動化地名標註產生的誤差，如此一來研究者便可以建立出一個針對此文件的簡易 HGIS。研究者可以透過這樣的視覺化工具觀察歷史文件之中的時間與空間的脈絡。

透過此系統可以幫助人文研究者跨越一般建立 GIS 的技術門檻與資料門檻。研究者可以透過系統的介面操作，將文件一筆或多筆提交給伺服器端，由伺服器端的 PHP 函式，將文件地名標註並儲存於資料庫中，研究者也可以透過介面操作將已存於資料庫的文件經過伺服器端 PHP 函式轉換為 GIS 可以讀取的 JSON 檔傳回用戶端。上述過程，研究者只需要點擊界面上的功能按鈕即可完成，而不需要有資訊背景，不需要會撰寫 PHP 程式、網頁語言、資料庫等技術。另一方面，本研究提供類似 Geocoder 的服務，系統透過空間資料庫及 Text Mining 技術幫助研究者擷取文件中的地理資訊。研究者需要提交的僅有文件本身及時間，研究者並不需要擁有大量當代的地理資訊，亦可以幫文件加入地理資訊。如此一來系統增加了研究者在使用上的彈性及便利性，降低了人文研究者使用 GIS 軟體的使用門檻，達到 GIS 直接輔助歷史研究的目的。

透過本系統可以讓使用者在使用 HGIS 時由被動性的接受資料，變為主動使用自己的資料。雖然在功能性上為了維持操作的容易性，必須將功能限制，只能做



到專業 GIS 軟體的部分功能，但已能達到讓使用者可以透過空間的方式觀察歷史資料。除了能夠呈現自身研究成果外，亦可以使用在教學上，有效節省講解 GIS 技術操作上的時間。本研究亦希望能降低一般研究者對資訊領域的排斥感，在研究者熟悉此系統後，若需要更加精確更多功能的系統，對於各種專業的 GIS 系統更容易上手。

5.2 未來工作

本研究受限於資料的取得，目前僅使用《中華文明之時空基礎架構》計畫中西漢至清代 38545 條地名的點坐標，這樣的資料數尚無法全面的提供歷史研究所需要的地理資訊協助，且地理資訊不僅僅只有點層面的使用，應該將線及面的資訊考量進去。例如過去道路、河流、海岸線、行政區的地理資訊在歷史研究上應也有相當的參考價值，未來若能取得更多的 GIS 研究使用的地圖資訊加入空間資料庫，系統勢必可以提供質量更佳的地名標註，且讓不同類型的地理資訊加入系統呈現，亦可以讓使用者能對文本做更多面向的觀察。

本系統標註地名的方式為使用詞庫比對後，再利用詞夾子作為驗證地名語意，可以說是一種地名辨識技術，而地名辨識這個課題在過去有許多研究者提出過許多方法，尤其在中國，地名辨識的相關研究一直有許多人關注，這些研究多是採用統計或規則來辨識，也有使用 Machine Learning 方式，且都具有不錯的精確度及召回率。但這些研究所採用的多是針對現代文章建立的語料庫及規則，而本研究目標為標註歷史文件，較不適用這些語料庫，因此採用不需要斷詞的詞夾子演算法。未來若能夠利用這些方法來驗證標註的地名，可以用來與詞夾子演算法做比較的實驗，對於地名標註的正確性必有正向的幫助。

目前系統為了讓使用者減輕負擔，讓使用者在提交文章時不需要填寫太多表單使得能夠描述文章的詮釋資料不多，未來應透過資訊技術或介面設計，自動化或半自動化的從文章擷取更多詮釋資料，提供更多功能選單幫助使用者過濾文章，讓使用者能透過操作介面，能清楚觀察歷史文件間的脈絡

REFERENCE



- [1] Onno Boonstra, Barriers Between Historical GIS and Historical Scholarship, *International Journal of Humanities and Arts Computing*, Volume 3, no 1-2, 2009
- [2] 范毅軍，〈論地理資訊系統在歷史研究上的應用〉，《古今論衡》2:93-96,1999
- [3] 張尚斌，〈詞夾子演算法在專有名詞辨識上的應用-以歷史文件為例〉，民國九十五年六月
- [4] 李丽双，黄德根，陈春荣，杨元生，〈SVM 与规则相结合的中文地名自动识别〉，《中文信息学报》第20卷第5期，1003—0077（2006）05—0051—07
- [5] 劉士綱，〈《清實錄》人名擷取自動化〉，中華民國 101 年 7 月
- [6] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets//Gupta Ashish, ed. *Proceedings of the 24th Conference on VLDB*. NewYork, 1998 : 392-403.

網站：



[8]中華文明之時空架構，〈<http://ccts.sinica.edu.tw>〉，檢索日期 2013 年 6 月 29 日

[9]Google Map，〈<https://maps.google.com.tw/>〉，檢索日期 2013 年 6 月 1 日

[10]總督府抄錄契書地理資訊，〈<http://thdl.ntu.edu.tw/tools>〉，檢索日期 2013 年 6 月 29 日

[11]台灣歷史文化地圖，〈<http://thcts.ascc.net/>〉，檢索日期 2013 年 6 月 29 日