計畫編號：DOH98-DC-1004

行政院衛生署疾病管制局 98 年度科技研究發展計畫

中英文媒體疫情自動分類及預警研究

研 究 報 告

執行機構：台灣大學

計畫主持人：鄭卜壬

研究人員：李佳蓉, 戴瑋彥, 高紹航, 陳健文

執行期間： 98 年 1 月 1 日 至 98 年 12 月 31 日

摘　　要

有效監測全球疾病的發展情形，預警下一波傳染疾病的可能趨勢，以利決策者作出正確且快速的反應，對牽制未來疫情的爆發是一個重要的機制。

疾病管制局已自行開發一「網路新聞地理資訊系統」，此系統目前僅能將新聞按部分疾病關鍵字做疫情搜集，尚需新技術能將新聞自動歸類，也需同時監測中英文等新聞來源的傳染病疫情、自動判斷新聞重覆性及各項疫情相關新聞重要性，以減少人力資源。

為減少「網路新聞地理資訊系統」的人力需求，加強其自動化功能，以期能應用在監測規模更大的網路疾病資料，本計畫旨在進行「中英文媒體疫情自動分類及預警」的研究。首先需建立疾病與地理資訊之中英文知識本體(ontology)以幫助疫情自動研判，為達此目的，本計畫需進行中文斷詞，從疾病新聞中擷取出疾病與地理資訊之相關中英文詞彙，並組織成階層式的分類架構。有了疾病與地理資訊的中英知識本體，中文與英文的新聞即可依疾病種類進行分類，分類為疾病相關的新聞，可再次利用地理資訊知識本體，擷取出疾病最可能的發生地區，最後，本研究擬自動群組重覆性的新聞，並進行各項疫情相關新聞重要性研判，依重要性由高至低排序出疫情相關新聞。

本計畫預期開發一「中英文媒體疫情自動分類及預警系統」，透過 Web 介面，使用者選取欲監測之一至多個疾病與一至多個中文或英文的新聞來源 (為方便調校系統整體效能，本系統初期僅針對某些特定疾病與新聞來源進行監測)，系統自動過濾出欲監測疫情的相關新聞及其可能的發生地區，在世界地圖標示哪些地區所發生哪些疫情，並以不同的標示顏色區別不同重要程度。

關鍵詞：疫情新聞分類、地理資訊系統、知識本體建置

Abstract

By effectively monitoring global diseases, one could take necessary precautions against infectious diseases according to their trends. How to help decision makers rapidly and precisely respond to the outbreaks of epidemics is, therefore, an important issue today.

The Centers of Disease Control (CDC) has developed a so-called "Web-based News Geographic Information System," which collects the epidemic-related news articles that contain a set of given keywords about some diseases. As keyword matching may bring unavoidable false alarms in document classification, many advanced techniques, such as automatic classification of news articles in Chinese and English, identification of locations that the diseases occur in, detection of duplicated news articles, and estimation of the importance of a epidemic-related article, are required to improve the performance of the CDC's system and speed up the procedure.

We expect such advanced technologies can be well applied to the application of "Automatic Disease Classification and Surveillance over Chinese and English News Articles" in a large scale, which is the goal of this project. In this project, we need to construct the ontology of disease and geographic information in Chinese and English, and apply the Chinese word segmentation technique to extract the terms about diseases and geographic information. Based on the ontology, both of the Chinese and English news articles can be classified, respectively. The geographic ontology is used to extract locations where the diseases occur. Finally, we automatically cluster duplicated news articles, judge their importance and rank them accordingly.

We plan to develop an "Automatic Disease Classification and Surveillance System over Chinese and English News Articles". Using a Web-based interface, users can select several diseases they want to monitor from one or more news sources. In order to tune our system's performance, we focus mainly on some specific diseases and news sources. Our system can automatically identify the disease-related news articles and locations the diseases possibly occur in, and mark them on a map with different colors according to difference levels of importance.

Keywords: epidemic-related news classification, geographic information system, ontology construction

前　言

(研究問題之背景與現況、研究目的)

有效監測全球疾病的發展情形，預警下一波傳染疾病的可能趨勢，以利決策者作出正確且快速的反應，對牽制未來疫情的爆發是一個重要的機制。

疾病管制局已自行開發一「網路新聞地理資訊系統」，此系統目前僅能將新聞按部分疾病關鍵字做疫情搜集，尚需新技術能將新聞自動歸類，也需同時監測中英文等新聞來源的傳染病疫情、自動判斷新聞重覆性及各項疫情相關新聞重要性，以減少人力資源。

為減少「網路新聞地理資訊系統」的人力需求，加強其自動化功能，以期能應用在監測規模更大的網路疾病資料，本計畫旨在進行「中英文媒體疫情自動分類及預警」的研究，開發一「中英文媒體疫情自動分類及預警系統」，透過 Web 介面，使用者選取欲監測之一至多個疾病與一至多個中文或英文的新聞來源，系統自動過濾出哪些新聞與監測之疫情相關，並從新聞中擷取出疫情可能發生的地區，研判各項疫情相關新聞重要性，最後利用 Google Map (http://maps.google.com)，在世界地圖標示哪些地區可能發生什麼疫情，並以不同的標示顏色區別不同的重要程度。為方便調校系統整體效能，本系統初期僅針對某些特定疾病與新聞來源進行監測。

早期的疾病監控系統主要是蒐集與分析病人就醫的書面記錄，近年來隨著資訊科技的進步，電子病歷加快醫療訊息傳遞及交換的速度，促進疾病監控系統的電腦化；然而，若要同時交換全球各地疾病的訊息，此問題變得更為複雜，各地疾病訊息的透明化程度、相互間通報管道的暢通與否，以及訊息傳遞的及時性等，都是全球疾病防疫工作所面臨的挑戰。上述訊息流通採用通報(push)的方式，亦即各地區主動通報各地發生的疾病，依照此種流通方式，若一個地區出了問題，未能及時發現或通報可能的新興疾病，極可能導致其他地區無法採取有效的疾病控制措施，「嚴重急性呼吸道症候群」(SARS)傳染病即為一例，這對國際貿易和旅行頻繁的地區，都會加速傳染病的擴散，對人口密度高的地區容易造成更大的傷害。因此，「及時的疾病相關訊息傳遞與分析」為現代化全球疾病預警系統不可或缺的能力之一。

圖 1: 報導北美腸病毒相關新聞的不同網頁

為及時獲得各地區疾病相關訊息，一個有效的解決方法是採用自動蒐集(pull)各地的訊息。隨著網際網路(World Wide Web)的盛行，目前世界各地疾病相關訊息不但廣泛散佈在網路上，並且時常更新，最顯著的例子是各地的新聞網站或入口網站，一般而言，這些網站同時報導國內與國際重要疾病新聞，以圖 1 為例，左側為北美地區最大的中文報紙「世界日報」於 2008 年 10 月 5 日報導有關一則當地新聞(http://www. worldjournal.com/wj-la-news.php?nt_seq_id=1783945)，描述南加州大學爆發腸病毒傳染病，圖一右側為香港的入口網站「香港新浪網」於 2008 年 10 月 7 日報導南加州大學發生急性腸胃道傳染病(http://news.sina.com.hk/cgi-bin/nw/show.cgi/32/1/1/896858/1. html)，當地新聞通常報導的時間與事件發生的時間非常接近，考慮新聞的即時性，這些網站有關疾病相關訊息，非常適合用來監測一個地區的衛生疾病狀況，更重要的是，這些公開的資料我們可以定期下載，或採用訂閱的方式，自動蒐集疾病相關訊息，達到「即時訊息傳遞」的目的。目前 Google News (http://news.google.com.tw)同時訂閱 350 個新聞來源，如圖 2 所示，該圖顯示以「傳染病」查詢 350 個新聞來源的結果，Google News 會將相同主題的疾病新聞合併成一個群組，其亦允許使用者以 pull 的方式訂閱新聞，當有包含某關鍵字的新聞發佈時，該網站會自動送電子郵件通知使用者，可惜 Google News 有以下 4 個限制，無法滿足一全球疾病預警系統的需求：

(1) Google News 採用「關鍵字」查詢的方式，此方式僅會找出包含該關鍵字的新聞，若一相關新聞未出現該關鍵字，則無法被選出。以圖 1 右側「香港新浪網」之新聞為例，該新聞未出現「腸病毒」，則無法以「腸病毒」為關鍵字查詢。要注意的是，疾病新聞並非皆由具醫學相關知識之記者撰寫，此限制會誤判許多相關新聞；

(2) Google News 結果僅可依新聞主題群組，然而就決策者而言，疾病發生的地區及其分佈狀況等資訊更為重要，若能從新聞內容得知疾病發生的地區，有助於決策者進行更細微的分析；

(3) 350 個新聞來源是固定的，使用者無法新增其它想要監測的來源；

(4) 無法同時處理多個語言的新聞，全球各地新聞使用不同語言，若能同時多個語言，有助於蒐集更多疾病相關新聞。

圖 2: 以「傳染病」查詢 Google News 的結果。

目前有兩個全球疾病預警系統較符合上述的需求，分別為 HealthMap (http://www. healthmap.org) 及 BioCaster (http://biocaster.nii.ac.jp)， 茲分述如下。

HealthMap(Brownstein et al.,2008; Freifeld et al., 2007)是由美國波士頓(Boston)兒童醫院以及哈佛(Harvard)醫學院的研究學者開發而成，本系統收集網際網路上的論壇與新聞，以追蹤可能爆發的傳染病，如結核病等。圖 3 顯示 HealthMap 的系統架構，本架構首先先蒐集疾病新聞，ProMED-mail (世界疫情情報網) (http://www.promedmail. org)、世界衛生組織 (http://www.who.int)、EuroSurveillance (歐洲疾病控制中心) (http://www.eurosurveillance.org)、Wildlife Disease Information Node (http/wildlifedisease. nbii.gov)、Google News、及 Moreover (http:// moreover.com)，每篇新聞分別包含標題(title)、URL (來源網址)、簡短描述(description，通常為新聞的前兩到三個句子)及新聞本文(info. text)，新聞的語言主要是以英文為主，目前系統正擴充其功能於西班牙語，分類引擎(Classification Engine)負責找出新聞內事件發生的地區及疾病名稱，由於 HealthMap 蒐集的資料幾乎都已經是疾病新聞，此分類引擎使用一包含所有地區及疾病名稱的資料庫(目前已收錄 2300 個地區名稱及 1100 個疾病名稱)，先檢查一新聞內所有出現的地區名及疾病名，再選擇一至兩個可能為事件發生的地區名及疾病名，系統架構中的 Web Backend 及 Web FrontEnd 是用來控制伺服器(server)如何與瀏覽器溝通，以增加傳送速度。圖 4 為 HealthMap 的系統介面圖，圖的左上區塊可選擇新聞來源，左下區塊可選擇欲監測的疾病，右上區塊以地圖的方式顯示發生這些疾病的地區，中間下方則是相關新聞，此網站介面提供 5 種語言(包含簡體中文)的版本。

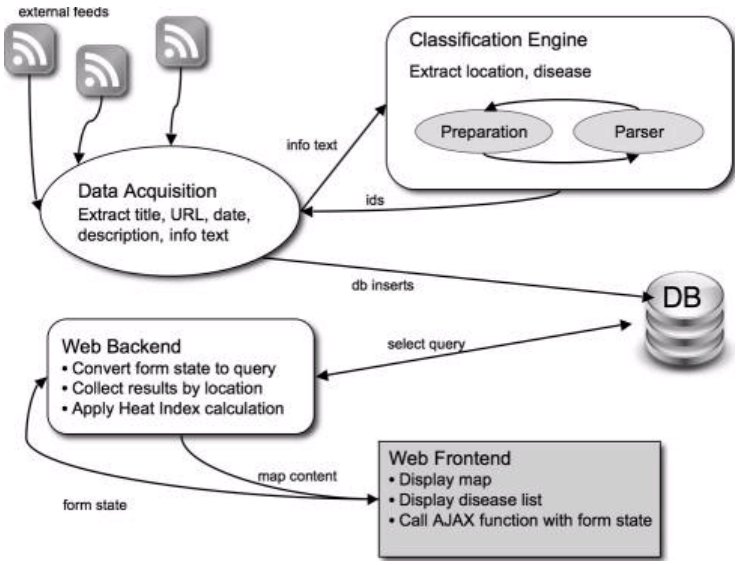


圖 3: HealthMap 的系統架構。

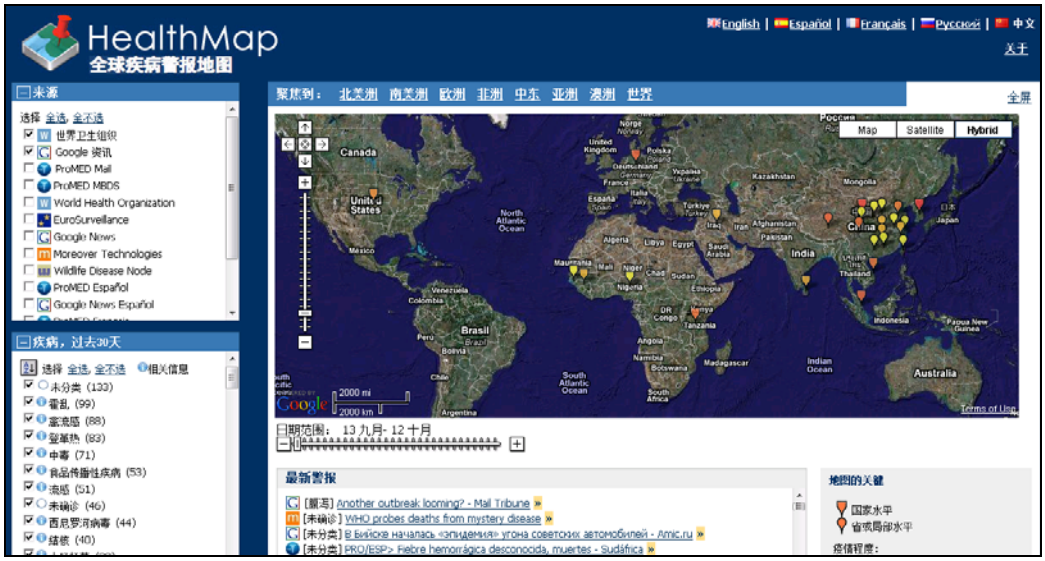


圖 4: HealthMap 的系統介面。

BioCaster(Doan et al., 2008; Kawazoe et al., 2008) 是由三個亞洲國家多所大學及研究所共同發展完成，包含日本國立情報學研究所(NII)、日本國立傳染病研究所、日本國立遺傳學研究所、日本岡山大學、越南國立大學 和泰國農業大學(Kasetsart University)，其特色是應用文件探勘(text mining)技術，嘗試找出網際網路上有關已知之突發性疾病及正在爆發的傳染病的相關新聞。圖 5 為其系統架構圖，首先系統先分類出內容有關於疾病的新聞，然後運用事先編輯好的知識本體(ontology)，擷取新聞中使用者感興趣的資訊，包含疾病發生的時間、地點、人物、疾病名稱以及相關事件，並將這些資訊自動翻譯成其它語言，最後再依疾病新聞的相關程度輸出給使用者，目前 BioCaster 過濾超過 1400 個的新聞來源，新

聞的語言以日語、泰語及越南語為主。圖 6 為其系統介面圖，此介面共有 7 種語言版本，介面的安排類似為圖 4 中 HealthMap 的系統介面，亦包含新聞來源與欲監測疾病的選擇、世界地圖的顯示方式及疾病相關新聞。

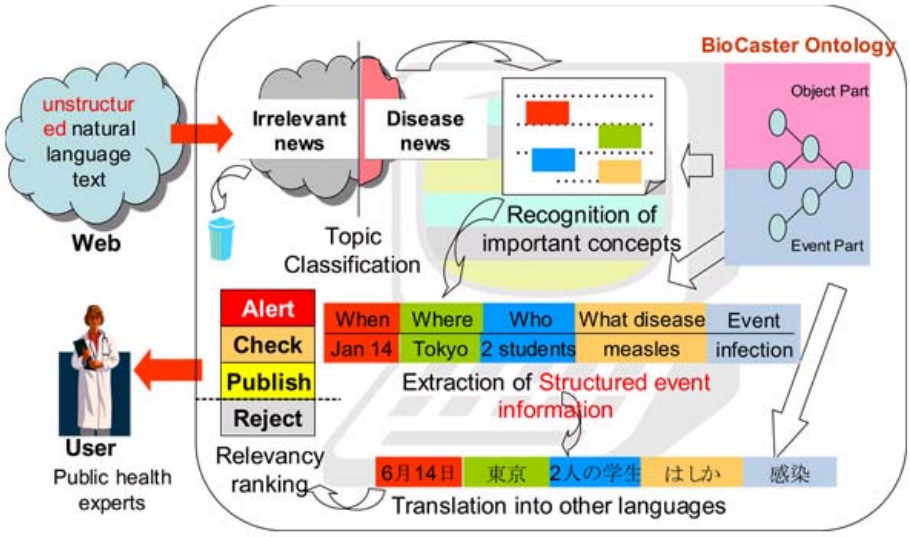


圖 5: BioCaster 的系統架構。

圖 6: BioCaster 的系統介面。

綜觀 HealthMap 與 BioCaster 兩個著名的全球疾病預警系統，前者著重於監測美加地區，後者著重於監測部份亞洲地區，其共同的優點包含

(1) 監測的新聞來源豐富。

(2) 建立龐大的疾病及地區名稱資料庫。

(3) 結合資訊、醫學、語言等領域專家共同研發。

(4) 部份技術方法發表於論文著作，並已進行系統效能評估。

(5) 以地圖的方式呈現，易於檢視疾病分佈情形。

然而，應用此二系統於台灣甚至於華人地區，仍有其侷限與不足之處，其包含

(1) 無法處理中文 (繁體中文與簡體中文) 新聞的內容，中文有斷詞問題 (word segmentation)。

(2) 缺乏中文疾病及地區名稱資料庫，此資料庫不僅收錄專業領域之專有名詞，亦需包含這些專有名詞在網路新聞的同義詞或相關詞。

疾病管制局已自行開發一「網路新聞地理資訊系統」(目前只針對禽流感)，但此系統僅能將新聞按部分疾病關鍵字做疫情搜集，尚需新技術能將新聞自動歸類且自動擷取出疾病發生的地區，此系統也需監測英文、繁體中文等新聞來源及禽流感以外的傳染病疫情，以減少人力資源，另外，此系統各項疫情相關新聞重要性目前須人工判斷，亦需有自動化的方式，提昇監測的效益。

為減少「網路新聞地理資訊系統」的人力需求，加強其自動化功能，以期能應用在規模更大的網路疾病資料，本計畫旨在進行「中英文媒體疫情自動分類及預警」的研究。首先需建立疾病與地理資訊之中英文知識本體以幫助疫情自動研判，此知識本體類似於 HealthMap 系統中的資料庫與 BioCaster 的知識本體，唯本計畫自動建置中文與英文知識本體，為達此目的，本計畫需進行中文斷詞，從疾病新聞中擷取出疾病與地理資訊之相關中英文詞彙，並組織成階層式的分類架構。有了疾病與地理資訊的中英知識本體，中文與英文的新聞即可依疾病種類進行分類，分類為疾病相關的新聞，可再次利用地理資訊知識本體，擷取出疾病最可能的發生地區，最後，本研究擬自動群組重覆性的新聞，並進行各項疫情相關新聞重要性研判，依重要性由高至低排序出疫情相關新聞。

工作項目包含

- 中英知識本體的建置
  - ✓ 知識本體包含疾病與地理資訊之階層式分類架構
  - ✓ 自動在網路上蒐集已有的疾病與地理資訊的分類架構
  - ✓ 利用文件探勘技術自動擴充現存的主題分類架構
- 中英文新聞之疫情分類
  - ✓ 為每一個欲監測之疾病建立一分類器
  - ✓ 分類器可同時處理中英文新聞
- 地理資訊擷取
  - ✓ 擷取一新聞中所有可能的地理資訊
  - ✓ 處理地理資訊之歧異性問題
  - ✓ 判別一新聞內容之疫情發生地區
- 新聞群組與重要性排序
  - ✓ 將重覆之新聞自動群組
  - ✓ 研判各項疫情相關新聞重要性
- 疾病新聞在地圖上展示

- ✓ Web 介面實作
- ✓ 在地圖上標示疾病新聞
- ✓ 依不同新聞重要程度標示不同顏色

研 究 方 法

本計畫旨在進行「中英文媒體疫情自動分類及預警」的研究，以期協助疾病管制局所開發之「網路新聞地理資訊系統」可達自動化的功能。圖 7 顯示本系統之系統架構，首先中英新聞文件經由各新聞網站所提供之 RSS 訂閱服務的方式取得，此處的 RSS 訂閱服務定義一種網路新聞頻道之數據交換規範，各新聞網站會主動傳送訂閱的內容給本系統，此內容一般為 XML 格式，經由該格式的分析，可得知新聞發佈的時間與內容等資料，收到的中英新聞經由「多語疾病分類模組」進行分類，判別哪些新聞內容與所監測的疾病有關，並且決定哪個新聞文件是關於哪個疾病，其後，「地理資訊擷取模組」負責擷取出新聞文件中疾病最可能發生的地區，此地區可能為一國家或一城市名稱，「新聞群組與相關排序」則自動群組重覆性的疾病新聞，並進行各項疫情相關新聞重要性研判，依重要性由高至低排序出疫情相關新聞，最後，「疾病地圖展示模組」以疾病發生地區的地理資訊，將疾病新聞用地圖的方式展示。以下詳述系統架構中 6 項主要工作：

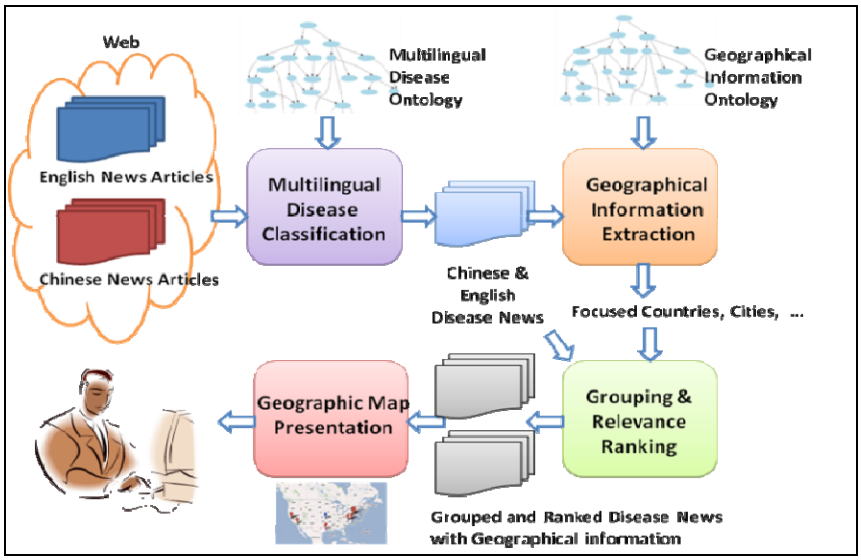


圖 7: 系統架構圖

(1) 自動蒐集網路中英文新聞

本系統透過各主要新聞網站所提供之 RSS 訂閱服務，取得該新聞網站中與健康類別相關之新聞，自動剖析其(XML 格式)新聞文件，以得知新聞發佈的時間與內容等資料。目前本系統所監測的英文新聞來源包含 Google News、Yahoo News、Baidu News、Reuter、BBC、CNN、ProMED-mail 與 WHO 主要新聞或疾病相關網站，未來易於擴充於監測其它網站。系統每隔 30 分鐘自動下載最新的文件，並過濾掉之前已被下載過的文件，為避免網頁內容在未來可能失效，目前系統會自動備份網頁內文字的部份於系統內部資料庫中。

(2) 建構中英文傳染疾病與地理資訊之知識本體(ontology)

利用資訊擷取(information extraction)的方法自動從網路抽取已整理之傳染疾病知識本體，此知識本體用以幫助疫情新聞的自動研判。利用資訊擷取的方法自動從網路抽取已整理之全球地理資訊知識本體，此知識本體用以幫助疫情發生地的自動研判。

由於網路新聞記者並非都具有專業的醫學知識背景，其撰寫之疾病新聞專業程度不一，有些新聞包含許多醫學專業術語，有些新聞則採用一般民眾易了解的詞彙編輯，導致疾病新聞分類的問題，無法僅利用疾病在醫學上之專有名詞作為判別依據，為增加網路新聞分類的準確度，中英文(包含專有名詞與非專有名詞)的同義詞與相關詞需群聚在一起，並依其語意層次組織成階層式的分類架構(Cheng et al., 2003)，進而形成一知識本體，上述的「多語疾病分類模組」與「地理資訊擷取模組」皆需要此知識本體才有辦法運作，目前大部份的知識本體需要以人工的方式建立，其成本太高且不容易維護，本計畫利用文件探勘技術嘗試自動擴充現存的中英知識本體，其內容涵蓋疾病與地理資訊。

疾病與地理資訊之中英知識本體架構極為複雜，因為知識本體中物件間的關係(relationship)可能有很多，本計畫初步僅考慮疾病與地理資訊的分類架構(taxonomy)，並整合下列兩種可能的解決方法：

(a) 自動在網路上蒐集已有的疾病與地理資訊的分類架構，

(b) 利用文件探勘技術嘗試自動擴充現存的分類架構。

在(a)「自動在網路上蒐集已有的疾病與地理資訊的主題分類架構」方面，「疾病」主題的分類架構有

- 疾病管制局的傳染病統計資料查詢系統 (http://nidss.cdc.gov.tw) ，
- Medical Subject Headings (MeSH) (http://www.nlm.nih.gov/mesh) 。

「地理資訊」的分類架構有

- Geographic Names Information System (美國地區) (http://geonames.usgs.gov) ，
- World Gazetteer (美國以外的地區) (http://www.world-gazetteer.com) ，
- United Nations Department of Economic and Social Affairs (國家) (http://unstats.un.org/unsd.) ，
- ISO 3166-1 Code List (國家名稱縮寫) (http://www.iso.org/iso/country_codes/ iso_3166_code_lists.htm)。

我們利用Web crawling技術，自動下載(crawl)上述網站的疾病與地理資訊，並剖析下載的網頁，抽取出疾病與地理資訊，然而，上述網站中的「疾病」主題與「地理資訊」的分類架構各有其限制。對「疾病」分類架構而言，網路疾病名稱可能

包含新詞及專有各種同義詞或相關詞，與多種的翻譯方式，以「禽流感」為例，其網路新聞可能的英文翻譯為“ bird flu”、“avian influenza” 或“bird virus” ，這些網路上出現的相關詞不易全部收錄在上述之分類架構。對「地理資訊」分類架構而言，地區名稱在網路上較不易出現所謂的同義詞或相關詞，但上述分類架構僅收錄洲、國家與城市，無法涵蓋所有的地理資訊。為解決第一個問題，本計畫利用文件探勘技術，嘗試自動擴充現存的分類架構(詳述如下)，為解決第二個問題，本計畫利用實體名識別技術 (named entity recognition; NER) (Gao et al., 2005)輔助判別新聞文件中可能出現的其它地名。

在(b)「利用文件探勘技術嘗試自動擴充現存的主題分類架構」方面，旨在從文件中找出其單語與雙語的同義詞(或相關詞)，以擴充主題分類架構的內容，過去同義詞(或相關詞)辭典大部份需要以人工的方式建立，其成本太高且不容易維護。因此，有些研究嘗試利用統計模型方法在各種語料庫中自動化抽取可能的雙語同義詞彙(即翻譯詞彙) (Gale et al., 1991; Rapp, 1995; Fung, 1998)，此語料庫包括並列雙語語料 (parallel bilingual text)或主題接近的雙語語料 (comparable bilingual text)，此種方法雖然避免人工抽取費時費力的缺點，但是由於並非所有語言都有足夠的語料庫，語料庫的取得成為此類方法在實務應用上的瓶頸。

隨著網際網路的快速發展，全球資訊網被視為一個龐大、跨語言且持續成長的語料庫，近年來有很多研究專注分析與利用某些包含多語言版本的網頁，集合成一個以網路為主的並列雙語語料庫，以補充特定領域並列雙語語料庫之不足(Kilgarriff, 2003)。此類研究(Nie, 1999; Resnik, 1999; Yang, 2003)假設不同語言版本的網頁可能含有相同的網頁或網站結構、類似的URL目錄及相近的網頁長度，提出的方法主要包含以下步驟：偵測並下載包含雙語或多語網頁的網站，建立特定語言的語言模型，用以預測所下載的網頁是否包含其感興趣的語言範圍，利用網頁的內容長度、結構等資訊，進一步分析該網站內所有可能成為並列雙語語料庫的網頁，對每一網頁的所有多語言版本內容，排列其可能的對應段落或句子，以組合成一個並列雙語語料庫。此方法需下載大量網頁。

本計畫嘗試以搜尋引擎的跨語言搜尋結果頁面 (search-result pages)內的豐富訊息，利用文字探勘技術，進行單語與雙語的同義詞(或相關詞)抽取之研究，此研究特別適用於各種專業領域中專有名詞的網路同義詞與相關詞(Cheng et al., 2004a; Cheng et al., 2004b)。搜尋引擎的搜尋結果頁面是搜尋引擎對某一查詢要求的查詢結果提示。基於不同的語言，文化與習慣差異, 某些語言在文件撰寫上需特別指明其翻譯的來源，以方便閱讀及了解，這些差異從翻譯抽取的觀點，正好提供豐富的語料素材。而大部份的搜尋引擎在搜尋結果頁面內，對每個指到目標網頁的網頁連結(hyperlink)旁，常輔助一些文字描述以註解其連結網頁的主題(title)及某些連結網頁內的文字片斷(snippet)，幫助查詢者判斷該網頁連結是否符合查詢需求，如有搜尋結果頁面同時包括雙語訊息，這些訊息可能可以充當雙語

語料。舉例來說，如圖 8 所示，相關於英文查詢“SARS” 的中英文網頁相當多，在搜尋結果頁面內，綜合這些不同文字描述很容易找到 SARS 在台灣繁體中文的相關翻譯為「嚴重急性呼吸道症候群」或「急性嚴重呼吸道症候群」。上述方式亦可應用於抽取單語的同義詞(或相關詞) ，如圖 9 所示，查詢「非典型肺炎」的搜尋結果頁面內，可以找到的「非典」、「未知病原體引起的肺炎」、「嚴重急性呼吸道症候群」與「院內感染」等相關詞。

圖 8: Google 對英文 “SARS” 查詢繁體中文網頁之搜尋結果頁面。

圖 9: Google 對中文 “非典型肺炎” 的中文搜尋結果頁面。

在此面臨兩個主要的問題：(i) 中文書寫上因為缺乏空白等符號區隔詞的邊界，如何從搜尋結果頁面中抽取出有效的字詞，(ii)是否可以從已抽取出有效的字詞中自動評估哪些最可能是同義詞(或相關詞)。

針對問題 (i)，即所謂的「中文斷詞」問題，本計劃首先考慮以 PAT-tree 為基礎自動抽取特定領域專有術語的關鍵詞抽取技術，這項技術已經廣泛受到引用 (Chien, 1997)，是中文檢索重要方法之一。PAT-tree 索引結構能將大量文件中任意長度的字串有效索引紀錄，加上考量字串出現頻率及分析字串前後文組合自由度以及組成文字結合強度的關鍵術語認定統計模型，可以有效過濾邊界切分 (boundary segmentation) 不完整的字串，如「嚴重急性呼吸」、「急性呼吸症候」等，這項統計模型度量可巧妙透過 PAT-tree traverse 實施，術語抽取成效極佳，對亞洲語文特別是中文極為有效。

考慮搜尋結果網頁中任意 n 個連續字或字元(character)，即 n-gram，以 PAT-tree 為基礎的斷詞法則，可採用下列公式運算某 n-gram 是否為一有效的字詞機率。

$$CD(w_1 \mathrm{K}\ w_n) = \frac{LC(w_1 \mathrm{K}\ w_n) RC(w_1 \mathrm{K}\ w_n)}{freq(w_1 \mathrm{K}\ w_n)^2},$$

其中 $LC(w_1 \ldots w_n)$為 n-gram 在搜尋結果網頁中其左邊唯一相鄰字或字元的個數，若其無任何左邊相鄰字或字元，該值表示其在搜尋結果網頁中出現的次數。同樣地，$RC(w_1 \ldots w_n)$為 n-gram 在搜尋結果網頁中其右邊唯一相鄰字或字元的個數，若其無任何右邊相鄰字或字元，該值表示其在搜尋結果網頁中出現的次數。$freq(w_1 \ldots w_n)$為 n-gram 在搜尋結果網頁中出現的次數。

由於 PAT-tree 斷詞是對搜尋結果網頁內作出長詞優先的統計結果，等於產生了可能的有效詞彙，若所搜尋結果網頁內的資訊量不夠，無法準確地評估其在統計上的重要性，因此除了 PAT-tree 外，我們考慮另一斷詞技術，此技術主要是考慮某 n-gram 中任何組成之子集合(sub n-gram)是否為一個有效字詞的機率。亦即用來判斷某 n-gram 內連續字或字元間的相關強度，在此，我們將採用下列公式運算某 n-gram 是否為一個有效的字詞機率。

$$\begin{aligned} SCP(w_1 \mathrm{K}\ w_n) &= \frac{p(w_1 \mathrm{K}\ w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} p(w_1 \mathrm{K}\ w_i)\, p(w_{i+1} \mathrm{K}\ w_n)} \\ &= \frac{freq(w_1 \mathrm{K}\ w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} freq(w_1 \mathrm{K}\ w_i)\, freq(w_{i+1} \mathrm{K}\ w_n)}, \end{aligned}$$

其中 $p(w_1 \ldots w_n)$為 n-gram 在搜尋結果網頁中出現的機率。結合上述兩個公式，我們可以得到以下的方程式。

$$SCPCD(w_1 \text{K } w_n) = SCP(w_1 \text{K } w_n) * CD(w_1 \text{K } w_n) = \frac{LC(w_1 \text{K } w_n) RC(w_1 \text{K } w_n)}{\frac{1}{n-1}\sum_{i=1}^{n-1} freq(w_1 \text{K } w_i)\, freq(w_{i+1} \text{K } w_n)}$$ 。

此方法將同時考慮任意 n-gram 中連續字或字元間的相關強度以及有效過濾邊界切分不完整的字串。

針對問題 (ii)自動評估哪些抽取出的有效字詞最可能是同義詞(或相關詞)，我們可以計算一個詞與其查詢結果頁面中的有效字詞共同出現(co-occurence)的機率，亦可計算該詞與有效字詞在文件中的前後文是否相似(similarity)，以決定兩個詞的相關程度(Cheng et al.,2004b)。

(3) 中英文疾病相關新聞之分類

「多語疾病分類模組」負責分類訂閱之中英新聞文件，以判別哪些新聞內容與所監測的疾病有關，並且決定哪個新聞文件是關於哪個疾病，此模組主要是處理一個多語文件分類的問題。長久以來，「分類」(classification)在機器學習(machine learning)、資訊擷取(information retrieval)、資料探勘(data mining)、圖形識別(pattern recognition)等領域一直是個重要的研究議題。

傳統的分類技術採用監督式學習方式，需要訓練資料(training data)訓練分類模型，以預測未看過資料的可能類別；以文件分類為例，使用者需先為每個類別蒐集相關的訓練文件，然後利用這些文件訓練各種分類模型(如 SVM、k-NN、Naïve bayes 等)，當給一個新的文件時，已訓練的分類模型可用於預測此新的文件是屬於哪一個類別(D. Lewis, 1998; Joachims, 1998)。此分類技術屬於「監督式分類」，若想要達到準確度高的分類效能，訓練文件必須適當地標記類別，且訓練文件的數量不能太少，若應用傳統的分類技術於此模組，將面臨兩個問題：

(a) 訓練文件不足的問題。

(b) 多語新聞分類的問題。

(a) 訓練文件不足的問題

對監測過去已標記(label)何種傳染病的新聞而言，傳統的分類方法問題較小，因訓練文件可從過去歷史的資料集取得，然而對於大部份的疾病，疾病管制局缺乏已標記的訓練文件。

為解決此問題，我們考慮利用已建置之疾病知識本體，由於每個疾病知識本體中的疾病，其知識本體記錄其上位語、下位語、症狀名稱、同義詞等，我們視此記錄為一特徵集合，對於缺乏訓練文件的傳染病類

別，我們檢查一給定文件是否出現此特徵集合內的詞彙，出現愈多表示愈可能是描述該疾病的新聞文件，利用規則式(rule-based)的方法，即使沒有已標記好的訓練文件，仍可辨識出其可能類別，雖然此方法屬於關鍵詞比對(keyword matching)技巧的一種，然而由於特徵集合是以網路探勘技術從網頁中擷取，其特徵覆蓋率(coverage)具一定比例。

此方法仍不適用於「新興疾病」的分類，若疾病管制局可提供少量的訓練新聞文件，例如 SARS 剛發生時，該疾病還尚未有名稱的新聞，經由少量訓練新聞文件的特徵抽取，再配合疾病專家提供的關鍵字，或可建立此類別的分類模型，當一新聞不屬於任一已知的傳染病類別時，並非直接濾掉該新聞，而是考慮該新聞屬於「新興疾病」的可能性。

(b) 多語新聞分類的問題

傳統的分類方法多討論單一語言文件的分類問題，亦即一個類別內的文件都是採用同一種語言撰寫，本模組每類別內同時包含中英文文件。我們初步考慮利用中英疾病知識本體，計算中英訓練文件中特徵的共通性與互補性，增加多語新聞分類的準確度。

對於疾管局之前以人工方式標記好的新聞為訓練資料，我們採用 SVM 的監督式學習方式，為避免不同類別之訓練資料量差異太大，每個類別 positive example 會重覆取樣至相同數量，每個類別的 negative example 則是從網路上的健康相關新聞及其他類別的文章而來。我們考慮兩種不同的分類方式，第一種是對所有疾病訓練一 SVM 模型(多重分類)，包含一特殊類別(其它類)，另一個是對每個疾病分別訓練一 SVM 模型(二元分類)。多重分類的缺點是新增一類別需重新訓練所有類別模型，二元分類的缺點是不同類別間的分類結果無法直接比較。

(4) 中英文新聞地理資訊之自動擷取

「中英地理資訊擷取模組」負責擷取出新聞文件中疾病最可能發生的地區，此地區可能為一國家或一城市名稱，利用中英地理資訊知識本體，本模組可以標示一新聞文件中是否有出現地理名稱，例如：洲名、國家名與城市名。為加速字串比對速度，適當的索引結構在實作上是必須的，如：雜湊表(hash table)。

若應用中英地理資訊知識本體於此模組，將面臨三個問題：

(a) 中英地理資訊知識本體僅收錄洲、國家與城市名，無法涵蓋所有的地理資訊。針對此問題，我們利用自然語言處理(natural language processing)技術中實體名識別方法 (NER)輔助判別新聞文件中可能出現的其它地名，實體名識別法透過學習的機制，同時計算一語言地名命名方式是否

(b) 中英地理資訊常具有語意歧異性(ambiguity)的問題。以英文為例，92% 英文地理名稱歧異性問題(Smith et al, 2001)，如圖10所示，歧異性問題可細分成兩個問題(Amitay et al., 2004)，一個問題是一地名可能同時含有地理資訊之外的語義，如：英格蘭有一個地名叫"Reading" ，該字亦有閱讀的意思，第二個問題是一地名可能同時出現在不同地理位置，如：美國有24個城市名作叫"Paris" 。為解決此問題，我們初步設計一些經驗法則，減少歧異性的發生，例如：同一則新聞內的地名可能互有關聯等，以卡姆登(Camden)為例，此地名可以是美國或澳大利亞的一個小鎮， 若該新聞出現美國的地名比澳大利亞多，卡姆登可能是美國的地名。

(c) 若一則新聞同時出現多個地名，如何決定哪個地名是疾病的發生地。解決此問題最直接的方法是利用統計法則，若某地區發生疫情，同時有許多不同語言的新聞媒體報導該消息，這些報導中，最常出現的地名很可能就是發生地。利用統計方法的缺點是當統計量不足時，無法有效計算，若碰到此問題，我們初步考慮從Google News查詢更多相關新聞，此方法的風險是須確保Google News傳回真正相關的新聞。

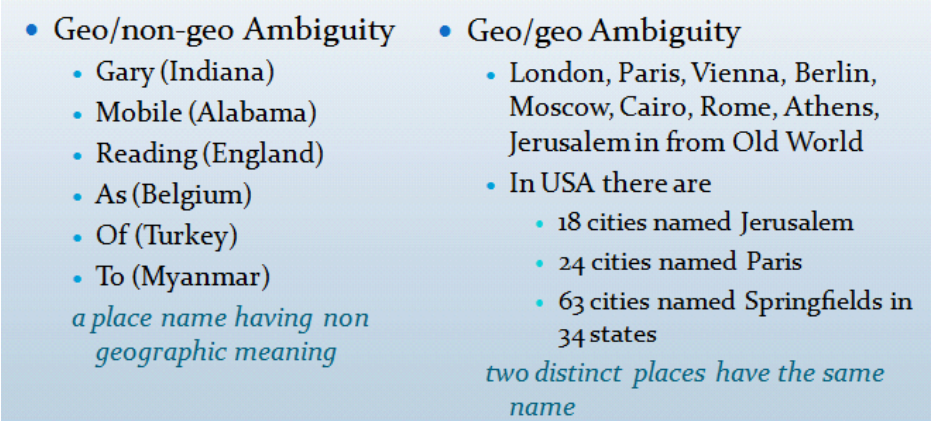


圖 10: 具語意歧異性的地理資訊。

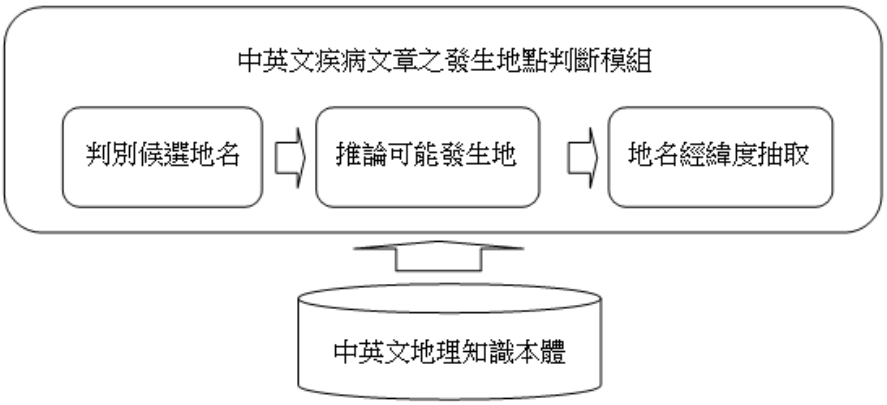


圖 11：「中英地理資訊擷取模組」的模組架構圖

為解決(b)與(c)兩個問題，本計畫嘗試許多可能的解決方法， 圖 11 為「中英地理資訊擷取模組」的模組架構，此模組將一篇文章的疾病發生地抽取出來，輸入一篇文章，由此模組的演算法計算出信心水準，根據分數輸出前幾組可能國家發生地，含城市<國家>的層級。由於疾病發生地抽取目前尚無明確且有效的方法去實現，因此我們借助實體名識別方法的技術與知識本體的推論。

在英文文件中候選地名判別方面，我們採用 Stanford NLP (Natural Language Processing) Group的NER Open Source，此Open Source的介面為輸入一篇文章，輸出一連串 Place/LOCATION的標示 (如: Taiwan/LOCATION)，我們收集後統整，最後使用中英文地理知識本體判斷是否正確。在中文文件中候選地名判別方面，我們採用中研院詞庫小組開發的實體名識別系統 (http://ckipsvr.iis.sinica.edu.tw/)，輸入一篇文章，會得到詞性標記結果，找出所有地理詞性的名詞，我們收集後統整，最後使用中英文地理知識本體判斷是否正確。

在推論可能發生地方面，

我們提出了四種演算法，分別如下：

‧方法 1「HighFeq (HF)」：

分別計算地名在文章中出現的頻率，將候選地點由頻率高到低輸出。

‧方法 2「HighFeqCountry (HFC)」：

同方法 1，只是將候選地點統一取國家層級後由高到低輸出。

‧方法 3「SurroundingText (ST)」：

取出部份文章作訓練語料，設定 Window 寬度 N，統計發生地的前後 N 單字內 bigram 詞頻。將測試語料輸入，在候選地點前後 Window 寬度 N 的 Bigram 詞依照訓練語料的詞頻加權分數。候選地點統一取國家層級後，分數高到低輸出候選名單。

‧方法 4「Combine (CB)」：

將 HFC 與 ST 的分數作 normalize，其兩者取 Linear Combination 後的分數為依據，Score function =α(HFC) + (1-α) (ST)，取國家層級後分數高到低輸出候選名單。

其中「候選地點」表示文章中出現過的地點，可能被選上為正確答案。「取國家層級 」表示候選地點透過知識本體轉換成國家層級，如果已經是國家則不變動。

舉例來說 (以 WHO 訓練語料為例)：

疾病發生地 : Indonesia

The Ministry of Health in Indonesia has confirmed an additional four cases of human infection with the H5N1 avian influenza virus. Two of these cases were fatal. There is no evidence of an epidemiological link between the cases.

The first case, a 31-year-old female from East Jakarta, Jakarta Province,

developed symptoms on 18 January, was hospitalized on 22 January and is currently in hospital. The investigation indicated that she visited a wet market where live poultry are sold three days prior to symptom onset.

The second case, a 9-year-old male from Depok Municipality, West Java, developed symptoms on 16 January, was hospitalized on 23 January and died on 27 January. Investigations into the source of his infection indicate that the case lived next door to a wet market where live poultry are sold.

The third case, a 32-year-old male from Tangerang Municipality, Banten Province, developed symptoms on 17 January, was hospitalized on 24 January and is currently in hospital. Investigations into the source of his infection are ongoing.

The fourth case, a 23-year-old female from East Jakarta, Jakarta Province, developed symptoms on 19 January, was hospitalized on 24 January and died on 27 January. Investigations into the source of her infection are ongoing.

Of the 124 cases confirmed to date in Indonesia, 100 have been fatal.

方法 1 (HF) :

候選地點有 (Jakarta, 4), (East, 2) , (Indonesia, 2), (Tangerang, 1).

方法 2 (HFC) :

透過知識本體將地點轉換成國家層級

Jakarta => Indonesia

East => India

Indonesia => Indonesia

Tangerang => Indonesia

則候選地點有 (Jakarta<Indonesia>, 4), (East<India>, 2) ,
(Indonesia<Indonesia>, 2), (Tangerang<Indonesia>, 1).

方法 3 (ST) :

由訓練語料得知前後文的詞頻如下 :

ministry of	265	has reported	93
of health	264	has confirmed	90
the ministry	147	of the	82
have been	143	has announced	67
health in	210	health of	63
to date	156	a new	61
been fatal	154	reported a	47
cases confirmed	130	confirmed the	46
date in	129	the country’s	43
confirmed to	129	cases in	41

以第一段為例，將發生地前後文有高頻字的加權

The Ministry of Health in Indonesia has confirmed an additional four cases of

human infection with the H5N1 avian influenza virus.
則候選地點有 (Jakarta<Indonesia>, 810), (Tangerang<Indonesia>, 1),
(Indonesia<Indonesia>, 0), (East<India>, 0).
方法 4 (CB) :
將 HFC 與 ST 的結果做 Normalize 之後得到:
HFC : (Jakarta<Indonesia>, 1), (East<India>, 0.5) ,
(Indonesia<Indonesia>, 0.5), (Tangerang<Indonesia>, 0.25)
ST : (Indonesia <Indonesia>, 1), (Tangerang<Indonesia>, 0.001),
(Jakarta <Indonesia>, 0), (East<India>, 0)
經過 Linear Combination 的 Score Function=α(HFC) + (1-α) (ST) ; α=0.2 之後得到候選地點 :
(Indonesia <Indonesia>, 0.9), (Jakarta <Indonesia>, 0.2),
(East<India>, 0.2), (Tangerang<Indonesia>, 0.1).

我們提出了上述幾種方法，並且進行實驗，根據實驗結果選出最適合本模組使用的方法。

(5) 相關新聞之群組與重要性判別

「新聞群組與相關排序模組」負責自動群組重覆性的疾病新聞，並進行各項疫情相關新聞重要性研判，依重要性由高至低排序出疫情相關新聞。

自動群組重覆性的問題較容易處理，因為「多語疾病分類模組」與「地理資訊擷取模組」已判別一新聞是屬於哪種疾病及可能的發生地，這些資訊都有助於決定兩則新聞是否報導同一地區的疫情，此外，亦可計算兩則新聞的相似程度，例如：共同出現詞彙所佔的比例等，皆有助於檢查重覆性的新聞。至於本模組需進行各項疫情相關新聞重要性研判，這個問題的難度較高，主要是因為新聞「重要性」的定義比較接近語意層次。

為解決重要性排序的問題，我們初步考慮一些可能突顯一網路新聞重要性的因素，例如：報導該疫情的新聞數量多寡、報導該疫情的新聞來源數量、報導該疫情新聞來源的重要程度等等。

(6) 以 Web 介面與地圖的方式

「疾病地圖展示模組」依疾病發生地區，將疾病新聞用地圖的方式展示。目前已有許多網站提供地圖展示與標示的服務。本計畫預期開發一疫情監測與預警系統，透過 Web 介面，使用者選取欲監測之一至多個疾病與一至多個中文或英

文的新聞來源，系統自動過濾出哪些新聞與監測之疫情相關，並從新聞中擷取出疫情可能發生的地區，研判各項疫情相關新聞重要性，最後利用 Google Map，在世界地圖標示哪些地區可能發生什麼疫情，並以不同的標示顏色區別不同的重要程度。

研 究 結 果 與 討 論

以下我們對 4 個主要的系統實作報告其實作或效能評估結果與討論，包括「中英文傳染疾病與地理資訊知識本體之建構」、「英文疾病相關新聞之分類」、「英文新聞中地理資訊之擷取」及「系統介面與操作模式」。

(1) 中英文傳染疾病與地理資訊知識本體之建構

知識本體為一組織架構，將同義詞與相關詞群聚一起，並依其語意層次組織成階層式的分類架構。本計畫採用資訊擷取之自動化技術，以監督式學習(supervised learning)的方法，讓使用者標示網頁中部份知識本體內容，系統自動學習擷取規則，再應用此規則抽取出大量疾病與地理資訊的相關知識，進一步組織成一同時包含中英文之知識本體，並儲存於 MySQL 資料庫中。為建構中文正體、中文簡體以及英文三個版本，本系統應用 Google Translater (http://translate. google.com) 及 LiveTrans(http://wkd.iis.sinica.edu.tw/LiveTrans/)進行翻譯。

在傳染疾病知識本體之建構方面，其類別來源是參考根據疾管局傳染病統計資料查詢系統(http://nidss.cdc.gov.tw/)的類別，目前共有 55 類傳染疾病，每個疾病都有其症狀屬性，症狀來源包含疾管局傳染病統計資料查詢系統以及日本 Biocaster 計畫(http://biocaster.nii.ac.jp)已整理好的知識本體組織而成。圖 12 以 SARS 疾病為例，顯示其關聯式表格中的部份資料。

Name	Severe Acute Respiratory Syndrome
ENTerm	SARS-CoV infection, Severe acute respiratory syndrome, SARS
SCTerm	严重急性呼吸道综合症
TCTerm	嚴重急性呼吸道綜合症
hasSymptom	Atypical pneumonia Cough Diarrhea Dyspnea Fever Headache Hyperventilation Lethargy ...

Name	Fever
ENTerm	Fever
SCTerm	发烧
TCTerm	發燒

圖 12：SARS 疾病在知識本體中的資料結構

在地理資訊知識本體之建構方面，其類別來源主要是參考根據 http://world-gazetteer. com 網站，自動擷取此網站的部份資料來並分析其內容，地理資訊知識本體主要有兩個表格，分別是 Gazetteer 及 Country 表格。

「Gazetteer 表格」: 此表單存放世界上所有地理名稱的資訊。以下是該表格的格

式與範例，目前約有 300,000 筆紀錄。

ID	En_name	Chi_name	Alt_name	Ori_name	type	Pop
470422606	New York	紐約	Nueva York		Locality	8210195
Lat	Lng	Country	P_1	P_2	P_3	
4067	-7394	United States of America	New York			

ID(該筆編號)、En_name(英文名稱)、Chi_name(英文名稱)、Alt_name(別稱)、Ori_name(該語言的原地名)、type(地點層級)、Pop (人口數)、Country(所屬國家)、P_N(往上 N 層級的地名)。

「country 表格」：此表單存放世界上國家的資料，以下是該表格的格式與範例，目前約有 240 筆紀錄

ID	State	Name	Pop	Area	Dense	Alt_name	Capital
-111	Asia	Japan	127939307	377812	338.6	Japón, Japon	Tōkyō

ID(該筆編號)、State(所屬洲名)、Name(國家名稱)、Pop(人口數)、Area(面積)、Dense(人口密度)、Alt_name(別稱)、Capital(首都)。

由於自動資訊擷取與翻譯技術可能產生錯誤，本計畫已聘請三位工讀生，以人工的方式確認資料庫中的記錄是否正確。

(2) 中英文疾病相關新聞之分類

疾病	方法	features
禽流感	X^2	Flu, influenza, bird, avian, virus, poultry, outbreak, disease, birds
	ontology	Cough, Dyspnea, Fever, Headache, Hemorrhagic, diarrhea, Lethargy, Myalgia, Sore throat, Vomit
登革熱	X^2	Dengue, fever, cases, health, mosquito, disease, mosquitoes, breeding, virus
	ontology	Arthralgia, Fever, Headache, Myalgia, Nausea, Retro-orbital pain, Vomit
愛滋病	X^2	Hiv, aids, infected, prevention, obama, virus, epidemic, health, disease
	ontology	HIV, Human Immunodeficiency Virus, Human T Cell Lymphotropic Virus type III, Acquired Immune Deficiency Syndrome Virus, HTLV-III, AIDS virus, Acquired Immunodeficiency Syndrome Virus

圖 13：三種疾病在統計及知識本體方法產生特徵之比較

傳統的分類技術採用監督式學習方式，需要訓練資料(training data)訓練分類

模型，以預測未看過資料的可能類別。以疾病新聞文件分類為例，本系統需先為每個欲監測的疾病蒐集相關的訓練文件，然後利用這些文件訓練 SVM 分類模型，當要分類一個蒐集到的新聞時，已訓練的分類模型可用於預測此新的文件是屬於哪一個類別。分類新聞文件需抽取新聞中的詞彙特徵，本計畫採用兩種特徵，一種是 unigram 和 bigram 的統計式特徵(X^2)，另一種是專家建議的特徵(即從知識本體得到的特徵屬性)，共 1000 個特徵，圖 13 比較兩種特徵的差異性。

本計畫採用疾管局之前以人工方式標記好的新聞為訓練資料，包含 2008 年 7 月 10 日到 2009 年 1 月 16 日的疾病新聞，經過分析處理後，下列左圖顯式文件數量比較多的類別，實驗僅使用禽流感、愛滋病及登革熱三個類別。為避免不同類別之訓練資料量差異太大，以下實驗每個類別 positive example 會重覆取樣至 600 篇，每個類別的 negative example 則是從網路上的健康相關新聞及其他類別的文章而來。接著測試兩種不同的分類方式，第一種是對所有疾病訓練一 SVM 模型(多重分類)，包含一特殊類別(其它類)，另一個是對每個疾病分別訓練一 SVM 模型(二元分類) ，圖 14 顯示 5-fold cross validation 的結果。

疾病	語言	數量
禽流感	中文	370
禽流感	英文	579
登革熱	中文	259
登革熱	英文	251
愛滋病	中文	359
愛滋病	英文	73
疫苗	中文	162
霍亂	中文	85
霍亂	英文	150
流感	中文	115
流感	英文	90

對所有疾病訓練一分類模型 (多種分類)

語言	準確率
英文	97.5093%
中文	97.4552%

對每個疾病都訓練一分類模型 (二元分類)

語言	準確率
英文	93.1511%
中文	94.6444%

圖 14：中英文疾病新聞分類之效能

實驗結果顯示多重分類具較高的分類準確度，可是多重分類方法在系統新增類別時，所有類別之分類模型需全部重新訓練。

由於此分類模組有助於疾管局其它計畫，以下我們特別解釋這部份的程式代碼的使用方法。此代碼在使用上有兩個主要階段，包含訓練及預測階段。在訓練階段，此代碼會對所有給定疾病之訓練文件，用 X^2 的方法抽取其統計特徵(feature selection)，接著對這些特徵計算其 tf-idf 的權重值，然後再使用 LibSVM (本系林智仁教授所研發) 訓練所需之分類模型。在預測階段，當一篇新的文章輸入系統後，先用訓練階段建立的分類模型來預測此文章的疾病類別，假使無法分到任何已訓練的類別，再使用規則式的分類方法去比對所有疾病知識本體中記錄之疾

病。規則式的分類方法採關鍵字比對(keyword matching)進行分類，每個疾病的關鍵字是從疾病知識本體中的疾病名稱和相關症狀而來，對於疾病名稱會給予較高的比對權重。

使用者可以驗證並更改所有自動分類的結果，對一個訓練資料不足的疾病，系統初期會採用規則式的分類方法，但當該疾病被驗證的文件增加到足夠的數量後，即表示其擁有足夠的訓練資料，系統就可以把跟這些疾病相關的文章當成此疾病的訓練資料，再重新訓練一個新的分類模型，之後對此疾病就可以不用再使用基本的規則式方法來分類。而對一個已訓練好的統計式分類模型，也可以因為被驗證的文章數量增加，重新訓練一個更完整的分類模型。

相關程式使用說明：

feature selection

./sh create-feature.sh [input dir] [num of feature] [output file]

[input dir] 訓練資料的資料夾

[num of feature] 回傳前幾名的特徵

[output file] 輸出的特徵檔案

Example:

./sh create-feature.sh ./data/en 1000 feature_en_1000

./sh create-feature.sh ./data/tc 1000 feature_tc_1000

create-multi-model

這支程式做的事情，主要就是讀進一個訓練資料目錄，裡面每個子目錄都是一個類別，每一個類別底下放很多此類別的訓練文件， 讀入後計算每篇文章的 tf-idef 權重值，經過處理後訓練出一個分類模型，此分類模型會出現在[output dir]/model.svm，其他相關資訊也會存在[output dir]中。

./create-multi-model

–d [output dir] 分類模型資訊輸出的地方

–tmp [tmp dir] 建分類模型過程中的暫存地方

–info [class info] 關於分類類別的檔案，格式如下

```
Class_name1
Class_name2
…
```

每一行表示一個類別(class)名稱，此名稱必須和 input dir 中每個類別的目錄名稱相同

–feature [feature file] 特徵檔，格式如下

```
flu
influenza
…
```

每一行為一個特徵名稱

–input [input dir] 要拿來訓練的目錄資料夾
此資料夾底下每一個子資料夾對應到一個類別，資料夾名稱以類別命名
–svm [svm dir] LibSVM 的位置
–cv [#cv] Cross validation 的次數
如果沒有這個參數，則會輸出一個位於[output dir]/底下的分類模型檔(model.svm)
-lang [language] 要分類的語言，en 表示英文，tc 表示繁體中文

Example:
./create-multi-model -d model_en -info class_info -feature feature_selection/feature_en_1000
-input ./data/en -svm ./libsvm-2.89 -lang en
./create-multi-model -d model_tc -info class_info -feature feature_selection/feature_tc_1000
-input ./data/tc -svm ./libsvm-2.89 -lang tc

multi-classifier

這支程式主要就是讀進一個分類模型，然後從資料庫中讀出所有未分類的文章，接著對每篇文章去做分類，再將分類的結果於資料庫中作更新

./multi-classifier
-d [input dir] 在 create 階段建造出來的分類模型資料夾
-tmp [tmp dir] 此程式在執行時中間檔暫存的空間
-feature [feature file] 特徵檔，格式如下

flu
influenza
…

每一行為一個特徵名稱
-lang [language] 目前要測試的語言為何，en 表示英文，tw 表示繁體中文
-svm [svm dir] LibSVM 的位置

Example:
./multi-classifier -d model_en -tmp tmp –feature feature_selection/feature_en_1000 -lang en -svm ./libsvm-2.89
./multi-classifier -d model_tc -tmp tmp -feature feature_selection/feature_tc_1000 -lang tc -svm ./libsvm-2.89

keyword matching (rule-based)

這支程式會對資料庫中未被成功分類的文章去做規則式的分類

./perl keyword_matching.pl [keyword_dir] [language]
[keyword_dir] 裡面存放很多檔案，每個檔案的檔名都是對應到一個疾病名稱，檔案中每行都是一個關鍵字

[language] 要對資料庫中的哪種語言做 keyword matching，en 或 tc

crawling data from DB

這支程式會去抓資料庫中已經被驗証過的文章來當作訓練資料

./perl crawl_data.pl [output dir] [language] [num_doc] [diseaseID]

[output dir] 文章要存放的資料夾
[language] 要抓的語言
[num_doc] 最多抓幾篇
[diseaseID] 要抓的疾病的 ID

(3) 中英文新聞地理資訊之擷取

以下分別討論英文與中文新聞地理資訊之擷取實驗結果。

英文新聞的地理資訊擷取實驗

資料集合：

從 WHO 官方網站上抓取 Epidemic and Pandemic Alert and Response (EPR)[http://www.who.int/csr/don/archive/year/en/]的資料，共有 1996-2008 年份，全數抓取過濾雜訊後，留下 670 篇完整的文章。

- 文章數：670
- 標準答案(Gold Standard)：皆為國家層級的地名。
- 平均文章長度：211.3 單字/文章。
- 候選地點數統計：

  未考慮取國家層級：
  - ✓ 平均：3.33 個/文章
  - ✓ 最大：16 個/文章
  - ✓ 最小：1 個/文章
  - ✓ 標準差：5.3

  考慮取國家層級：
  - ✓ 平均：2.15 個/文章
  - ✓ 最大：10 個/文章
  - ✓ 最小：0 個/文章
  - ✓ 標準差：3.42

評估方法:

我們分析了 Precision 與 Recell，並且算出 Top N inclusion rate。

Recall = (擷取出正確發生地筆數 / 所有正確發生地筆數)

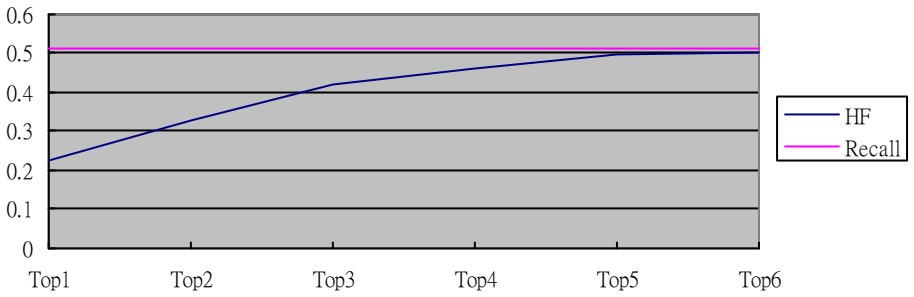
Precision = (擷取出正確發生地筆數 / 所有擷取出發生地筆數)

Top-N inclusion rate = (有多少百分比的正確率發生地會在前 N 筆傳回)

實驗結果 :

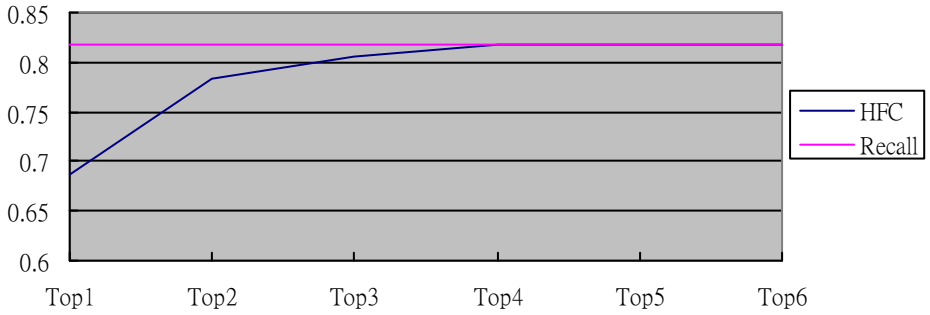
方法 1、HighFeq (HF) :

Method : HF					
Precision : 0.224		Recall : 0.513			Total : 670
Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.224	0.328	0.419	0.464	0.496	0.501



方法 2、HighFeqCountry (HFC)

Method : HFC					
Precision : 0.682		Recall : 0.818			Total : 670
Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.686	0.784	0.806	0.818	0.818	0.818



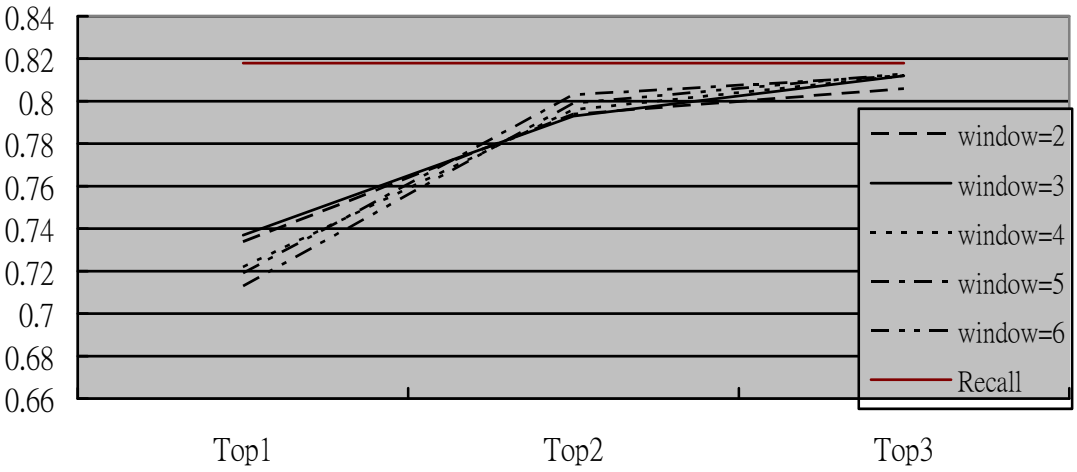
方法 3、SurroundingText (ST)

我們將部份資料集當作是訓練語料，統計前後 Window 內的 bigram 詞頻，得結果如下表：

ministry of	265	has reported	93
of health	264	has confirmed	90
the ministry	147	of the	82
have been	143	has announced	67
health in	210	health of	63
to date	156	a new	61
been fatal	154	reported a	47
cases confirmed	130	confirmed the	46
date in	129	the country’s	43
confirmed to	129	cases in	41

使用 5-fold Cross-Validation 的實驗結果

Method : ST							
Window = 2							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.734	0.818	0.734	0.794	0.806	0.812	0.818	0.818
Window = 3							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.737	0.818	0.737	0.793	0.812	0.815	0.818	0.818
Window = 4							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.722	0.818	0.722	0.796	0.812	0.815	0.818	0.818
Window = 5							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.719	0.818	0.719	0.803	0.812	0.815	0.818	0.818
Window = 6							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.713	0.818	0.713	0.799	0.813	0.815	0.816	0.818

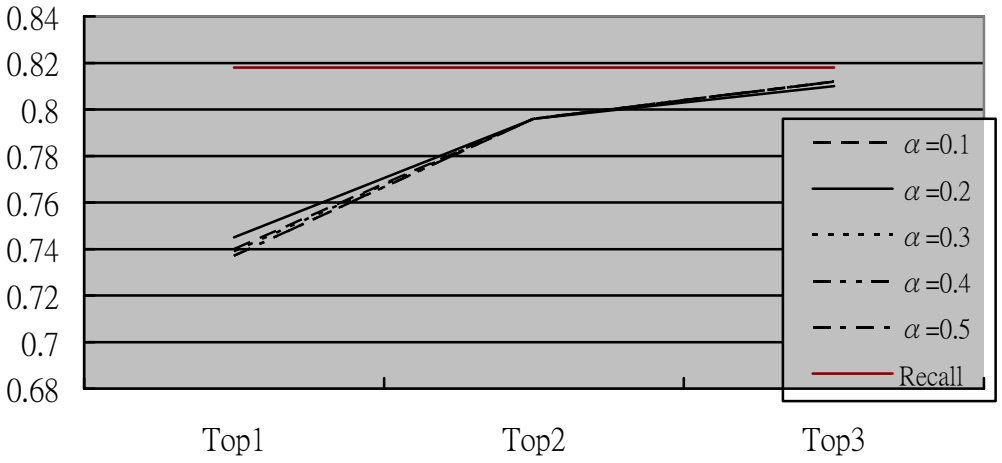


方法 4、 Combine (CB)

我們選擇 ST 方法中，結果較好的 Window = 3 進行實驗。

用 Linear Combination 合併 HFC 與 ST 的結果，給定 α 值，使用 Score function =α(HFC) + (1-α) (ST)

Method : CB							
α=0.1 ; 0.1(HFC) + 0.9(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.740	0.818	0.740	0.796	0.812	0.815	0.818	0.818
α=0.2 ; 0.2(HFC) + 0.8(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.745	0.818	0.745	0.796	0.810	0.815	0.818	0.818
α=0.3 ; 0.3(HFC) + 0.7(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.739	0.818	0.739	0.796	0.812	0.815	0.818	0.818
α=0.4 ; 0.4(HFC) + 0.6(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.737	0.818	0.737	0.796	0.812	0.815	0.818	0.818
α=0.5 ; 0.5(HFC) + 0.5(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.737	0.818	0.737	0.796	0.812	0.815	0.818	0.818



4 個方法的綜合比較

我們取出四個演算法中最佳的結果來比較，其中 ST 的 Window 取 3，CB 的 α 取 0.2。

Method	P	R	Top1	Top2	Top3	Top4	Top5
HF	0.224	0.513	0.224	0.328	0.419	0.464	0.496
HFC	0.682	**0.818**	0.686	0.784	0.806	**0.818**	**0.818**
ST	0.737	**0.818**	0.737	0.793	**0.812**	0.815	**0.818**
CB	**0.745**	**0.818**	**0.745**	**0.796**	0.810	0.815	**0.818**

從上表可以看出 CB 達到最高的 Top 1 inclusion rate，原因在於 ST 方法不足的地方，透過 Linear Combination，將 HFC 部份排名正確的拿來補足 ST 方法，因此 Top 1 inclusion rate 上升了 0.8%，Top 2 inclusion rate 上升了 0.3%。

中文新聞的地理資訊擷取實驗

資料集:

從WHO中文官方網站上抓取疾病爆發新聞-依年度分類 [http://www.who.int/csr/don/archive/year/zh/index.html] 的資料，從 2004-2009 年份，由於來源為簡體報導，先轉成繁體文章。範例如下：

禽流感–巴基斯坦的情況–最新簡報

2007 年 12 月 27 日

巴基斯坦第一起人感染 H5N1 型禽流感病例已得到證實。埃及開羅世衛組織 H5 參考實驗室和聯合王國倫敦世衛組織流感參考與研究合作中心所做的實驗室檢驗，已證實從一受影響家庭的一個病例收集的樣本中存在 A（H5N1）型禽流感病毒。該 H5N1 型陽性病例為一名來自白沙瓦地區的 25 歲男子，他於 11 月 21 日出現發熱性呼吸系統疾病，11 月 23 日住院，於 11 月 28 日死亡。目前正在進行其它實驗室分析，包括基因測序。

應巴基斯坦政府的請求，世衛組織一工作組前往巴基斯坦，參與國家當局正在對幾例人感染 H5N1 型禽流感疑似病例進行的調查。現已得出以下結論：

- 初步危險評估未見有證據表明存在持續的或社區人際傳播;
- 所有確定的密切接觸者包括受影響家庭的其他成員以及所涉保健工作者均無症狀，現已解除密切醫學觀察。

巴基斯坦衛生部已及時採取步驟，對這一事件進行調查和防控，包括隔離病人、對接觸者進行追蹤和監測、詳細的流行病學調查、提高個人保護裝備的可獲性、指定專用於治療新增疑似病例的醫院設施，以及其它感染控制措施。另外，農業當局包括糧食、農業和畜牧業部以及糧農組織，已成為積極的技術夥伴，參與了對這一範圍有限的疫情暴發的有效控制。

發生地：巴基斯坦

- 文章數：443
- 標準答案(Gold Standard)：皆為國家層級的地名。
- 平均文章長度：351.5 字/文章。
- 候選地點數統計：

  未考慮取國家層級：
  - ✓ 平均：4.23 個/文章
  - ✓ 最大：13 個/文章
  - ✓ 最小：1 個/文章
  - ✓ 標準差：5.21

  考慮取國家層級：
  - ✓ 平均：3.98 個/文章
  - ✓ 最大：14 個/文章
  - ✓ 最小：1 個/文章
  - ✓ 標準差：5.41

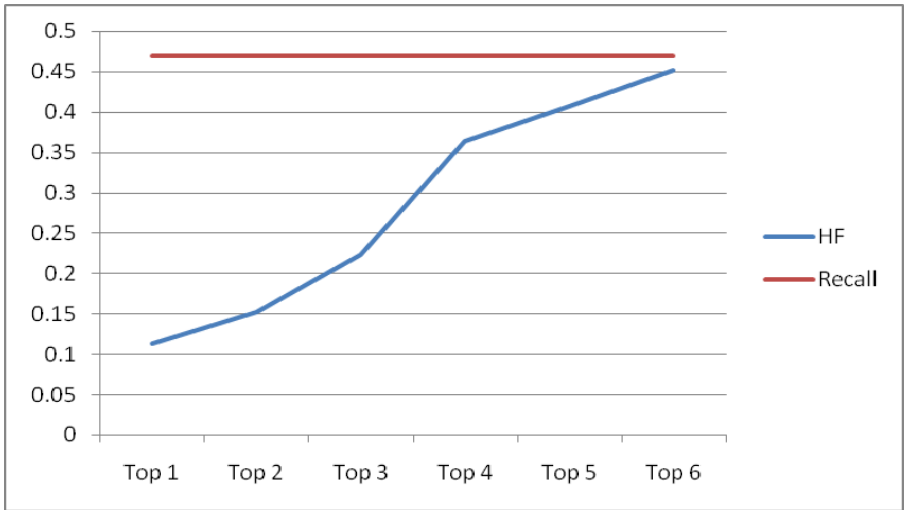
效能評估方法 :

我們分析了 Precision 與 Recell，並且算出 Top N inclusion rate。

實驗結果:

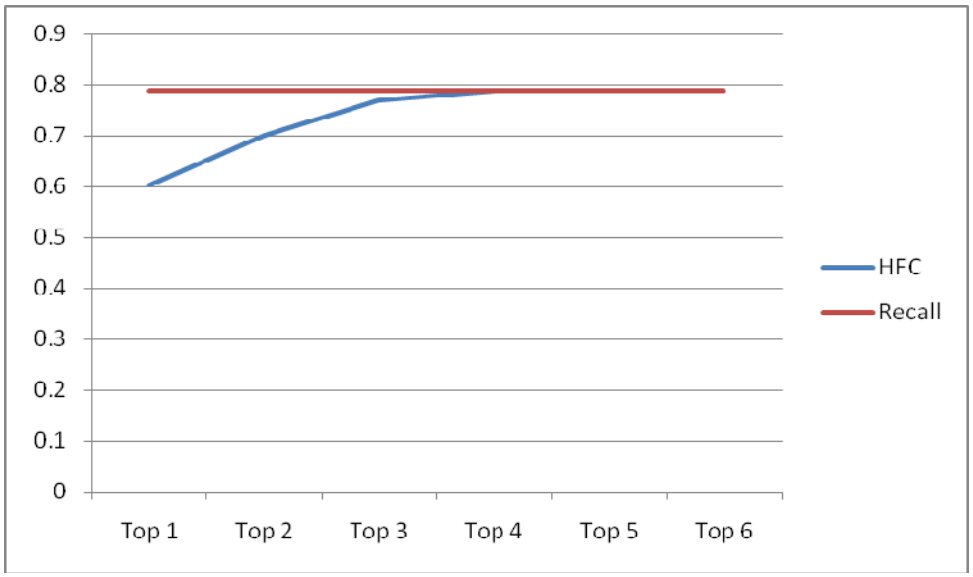
方法 1、 HighFeq (HF) :

Method : HF					
Precision : 0.113		Recall : 0.469			Total : 443
Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.113	0.152	0.223	0.364	0.406	0.451



方法 2、 : HighFeqCountry(HFC)

Method : HFC					
Precision : 0.652		Recall : 0.791			Total : 443
Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.602	0.700	0.770	0.788	0.791	0.791



方法 3、: SurroundingText (ST)

我們將 300 篇資料集當作是訓練語料，統計前後 Window size 內的 bigram 詞頻。

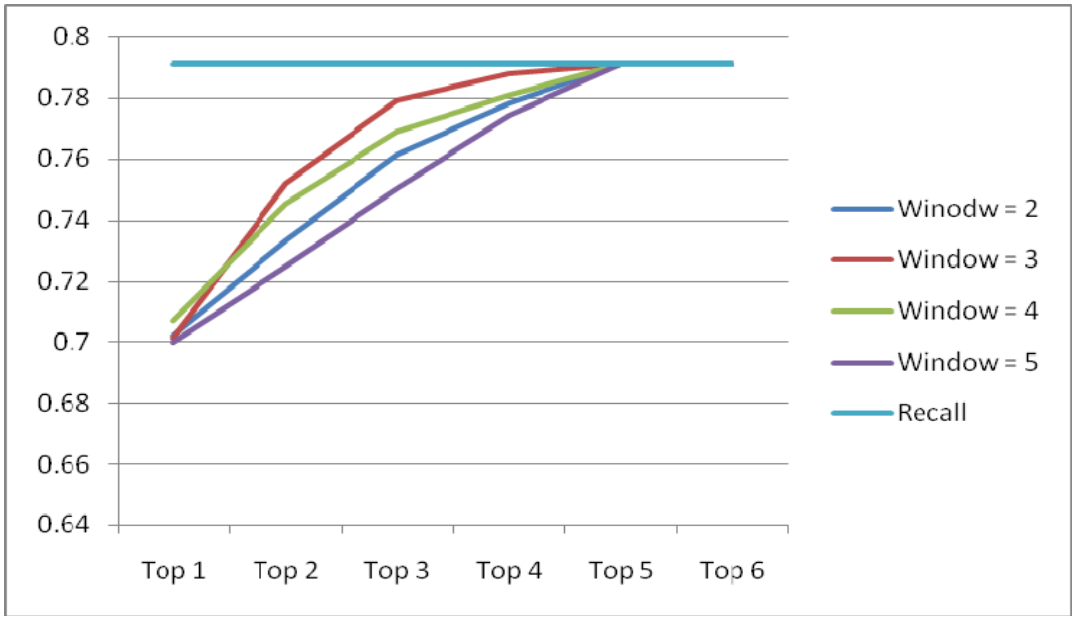
此 200 訓練語料統計數據如下：

- ✓ 平均文章長度：322.1 字/文章。
- ✓ 候選地點數統計：
  - ◆ 平均：3.34 字/文章

- 最大：11 個/文章
- 最小：1 個/文章
- 標準差：4.33

使用 5-fold Cross-Validation 的實驗結果

Method : ST							
Window = 2							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.702	0.791	0.702	0.733	0.761	0.778	0.791	0.791
Window = 3							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.701	0.791	0.701	0.752	0.779	0.788	0.791	0.791
Window = 4							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.707	0.791	0.707	0.745	0.769	0.781	0.791	0.791
Window = 5							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.700	0.791	0.700	0.725	0.75	0.774	0.791	0.791

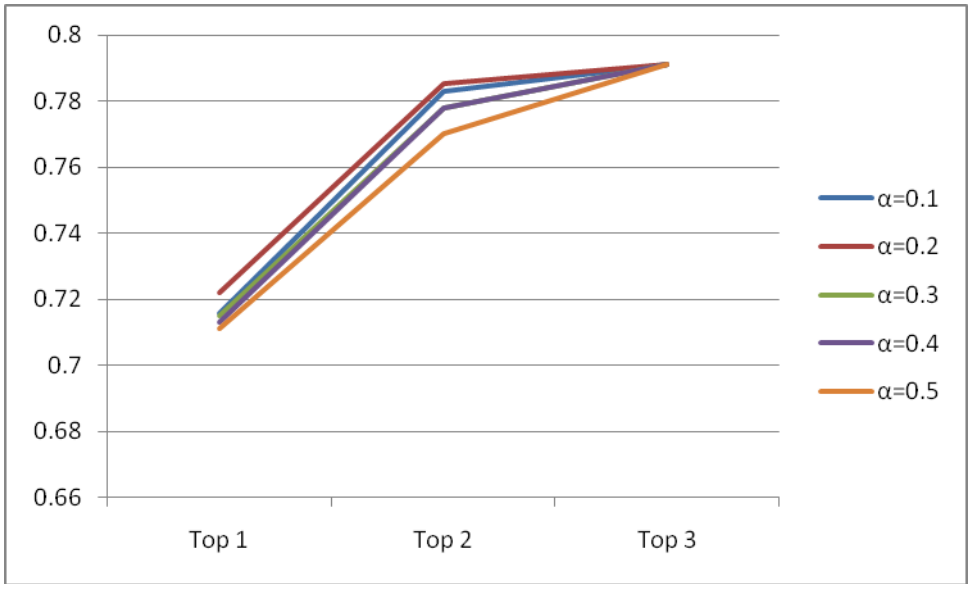


方法 4、Combine (CB)

我們選擇 ST 方法中，結果較好的 Window = 3 進行實驗。

用 Linear Combination 合併 HFC 與 ST 的結果，給定 α 值，使用 Score function =α(HFC) + (1-α) (ST)

Method : CB							
α=0.1 ; 0.1(HFC) + 0.9(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.716	0.791	0.716	0.783	0.791	0.791	0.791	0.791
α=0.2 ; 0.2(HFC) + 0.8(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.722	0.791	0.722	0.785	0.791	0.791	0.791	0.791
α=0.3 ; 0.3(HFC) + 0.7(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.715	0.791	0.715	0.778	0.791	0.791	0.791	0.791
α=0.4 ; 0.4(HFC) + 0.6(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.713	0.818	0.713	0.778	0.791	0.791	0.791	0.791
α=0.5 ; 0.5(HFC) + 0.5(ST)							
Precision	Recall	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6
0.711	0.818	0.711	0.77	0.791	0.791	0.791	0.791



4 個方法綜和比較

我們取出四個演算法中最佳的結果來比較，其中 ST 的 Window 取 3，CB 的 α 取 0.2。

Method	P	R	Top1	Top2	Top3	Top4	Top5
HF	0.113	0.469	0.113	0.152	0.223	0.364	0.406
HFC	0.602	**0.791**	0.602	0.700	0.770	0.788	**0.791**
ST	0.701	**0.791**	0.701	0.752	0.779	0.788	**0.791**
CB	**0.722**	**0.791**	**0.722**	**0.785**	**0.791**	**0.791**	**0.791**

透過 linear combination 的參數組合下，CB 在前 Top3 內能夠達到 0.791 的 Inclusion Rate，又如果只看 Top1 的情況能夠達到 0.722 的 Precision。

(4) 系統介面與操作模式

圖 15：系統主要 Web 介面

本計畫已開發一英文新聞疫情監測與預警系統，網址為 http://ir.csie.ntu.edu.tw/~irlab94/cdcWebsite/public/。如圖 15 所示，透過 Web 介面，使用者選取欲監測之一至多個疾病(圖中 Disease，目前版本支援 56 種疾病，包含具有訓練語料的禽流感、愛滋病及登革熱三種疾病)、一至多個英文的新聞來源(圖中 Source，目前版本支援 12 個新聞來源)與新聞發佈的期間(圖中 Time Period)，系統自動過濾出該期間有哪些新聞與監測之疫情相關，並從新聞中擷取出疫情可能發生的地區，

最後利用 Google Map (http://maps.google.com/)，在世界地圖標示哪些地區可能發生什麼疫情，預設以台灣為中心。同一個地點的新聞會集中在一個點上，點的顏色、大小是根據新聞的重要性做判斷，紅色為警戒色、橘色為中度警戒、藍色為普通警戒。滑鼠點到每個標示後會顯示出此處的新聞，白色視窗內顯示新聞標題以及時間，下方的棕色區塊顯示疾病種類、標題、時間以及概述。點選標題則可以回到原始的新聞連結。

本系統亦提供後端管理機制，對所有系統預測的結果，使用者可經由 Web 介面更改其內容，一方面是系統自動預測無法保證 100%的準確度，另一方面，此介面亦方便疾管局管理所有系統資料庫的資料，並決定是否發佈該監測結果於網頁上。此介面主要提供給疾管局人員使用，使用帳號密碼登入，密碼以 md5 編碼處理。功能分為「已發佈新聞管理」、「尚未確認新聞管理」、「疾病種類管理」、「新聞來源管理」、「症狀知識本體管理」，如圖 16 所示。

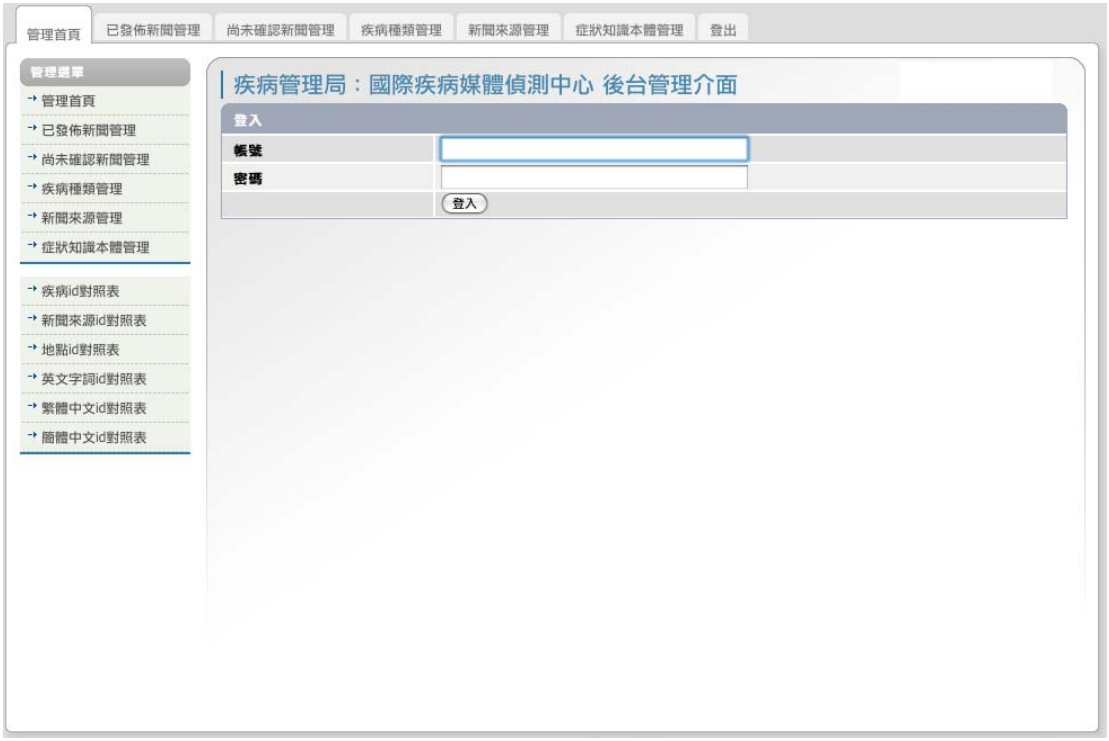


圖 16：系統管理介面

已發佈新聞管理：

此處可設定顯示在頁面上的資訊，包括疾病種類、新聞來源、標題、簡述、地點以及是否顯示在頁面上。此處的資料修改後會即時反應在已發佈的文章中。

尚未確認新聞管理：

此處是經過我們的系統判定為可能的疾病相關新聞，但尚未經過人工確認的資料清單。此處可以修改資料的所有原始資料，包括疾病種類、新聞來源、標題、簡述、地點...等，並決定是否為正確資料以及是否要發佈。是否為正確資料會影響到我們的系統，系統根據人工判定的正確結果做進一步的更新，希望藉此可以提高判別的準確率。是否顯示於頁面上則只是純粹決定是否要將資料發佈，並不會影響系統的預測機制。

疾病種類管理：

目前系統有 56 類疾病，是從 BioCaster 以及疾館局的疾病小百科而來。使用者可以在此設定疾病的英文、繁體中文以及簡體中文名稱。

新聞來源管理：

目前系統有 12 個新聞來源，此處可以設定各個新聞來源的英文、繁體中文、簡體中文名稱以及新聞媒體的官方網址。

總結本計畫期之成果報告，本計畫目前已開發出一中英文新聞疫情監測與預警系統，可線上展示，並已針對系統內各主要模組進行效能評估。相關網路探勘及檢索技術已發表於重要的國內及國際會議，包含 2 篇國內會議論文與 3 篇國際會議論文。

研 究 成 果 及 建 議

本計畫為國內有關「網路新聞地理資訊系統」自動化之先期研究，所發展之系統可有效幫助疾病管制局從事中英文媒體疫情分析的人員，加速其進行媒體疫情的分類與地點研判，進而有助其預警重要疫情在媒體報導的可能趨勢。

目前系統主要疾病分類正確率約 93%、第一候選發生地點預測正確率約 75%，因媒體報導內容的多元性，系統預測效能仍有改善空間，隨著疾病種類及新聞來源的增加，系統錯誤率可能上昇，如何加強系統與使用者的互動以維持結果的正確性變成非常重要，除了允許使用者驗證結果是否正確外，目前本系統已利用使用者的回饋資訊自動學習新的分類模型，以期當收集愈多使用者的資訊時，系統可以有愈好的效能。相關網路探勘及檢索技術已發表於重要的國內及國際會議，包含 2 篇國內會議論文與 3 篇國際會議論文。

目前本系統仍僅能協助原疾管局「網路新聞地理資訊系統」之自動化與模組化，及加入某些新的功能，如重覆與重要性判別，並不能真正地提供早期預警的機制，亦無法偵測新興的疾病，若能整合其它資源，未來可繼續進一步的研究。

參考文獻

Einat Amitay, Nadav Har'El, Ron Sivan, Aya Soffer (2004) Web-a-Where: Geotagging Web Content. In Proc. of ACM SIGIR Conference, pp. 273-280.

Brownstein, JS, Freifeld, CC, Reis, BY, Mandl, KD (2008) Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med 5(7).

Cheng, P.-J. and Chien, L.-F. (2003) Auto-Generation of Topic Hierarchies for Web Images from Users' Perspectives. In Proc. of the 12th ACM International Conference on Information and Knowledge Management (CIKM'03).

Cheng, P.-J., Pan, Y.-C., Lu, W.-H., and Chien, L.-F. (2004a). Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora. In Proc. of the Annual Meeting of the 42 th Association for Computational Linguistics (ACL'04).

Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., and Chien, L.-F. (2004b) Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In Proc. of the 27th ACM-SIGIR International Conference on Research and Development in Information Retrieval (SIGIR'04).

Chien, L.-F. (1997) PAT-tree-based keyword extraction for Chinese information retrieval, Proceedings of ACM-SIGIR'97 conference, pp. 50-59.

Doan, S., Hung-Ngo, Q., Kawazoe, A., and Collier, N. (2008), Global Health Monitor - A Web-based System for Detecting and Mapping Infectious Diseases. In Proc. of the International Joint Conference on Natural Language Processing (IJCNLP'08), pp. 951-956.

Fung, P. and Yee, L. Y. (1998) An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In Proc. of the 36th Annual Conference of the Association for Computational Linguistics, pp. 414-420.

Freifeld CC, Mandl KD, Reis BY, Brownstein JS (2007) HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. J Am Med Inform Assoc.

Gale, W. and Church, K. (1991) Identifying Word Correspondences in Parallel Texts. In Proc. of the forth DARPA Speech and Natural Language Workshop, pp. 152-157.

Gao, J., Li, M., Huang, CN, and Wu, A. (2005) Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. Computational Linguistics.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proc. of ECML.

Kawazoe, A., Chanlekha, H., Shigematsu M., and Collier, N. (2008) Structuring an Event Ontology for Disease Outbreak Detection. In BMC Bioinformatics, 9 (Suppl 3).

Kilgarriff, A. and Grefenstette, G. (2003) Introduction to the Special Issue on the Web as Corpus. Computational Linguistics 29:3, pp. 333-348.

Lewis, D. (1998) Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval. In Proc. of ECML, 1998.

Nie, J.-Y., Isabelle, P., Simard, M., and Durand, R. (1999) Cross-language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proc. of ACM-SIGIR'99 Conference, pp.74-81.

Rapp, R. (1995) Identifying Word Translations in Non-parallel Texts. In Proc. of the 35th Annual Conference of the Association for Computational Linguistics.

Resnik, P. (1999) Mining the Web for bilingual text. In Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99).

Smith, D.A., and Crane, G. (2001) Disambiguating Geographic Names in a Historical Digital Library. In Proc. of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'01), pp. 127–136,

Yang, C.C. and Li, K.W. (2003) Automatic Construction of English/Chinese Parallel Corpora. Journal of the American Society for Information Science and Technology 54(8), 730-742.

98 年度計畫重要研究成果及具體建議

（本資料須另附乙份於成果報告中）

計畫名稱：中英文媒體疫情自動分類及預警研究

主 持 人：鄭卜壬　　計畫編號：DOH97-DC-1004

1.計畫之新發現或新發明

本計畫為國內有關「網路新聞地理資訊系統」自動化之先期研究。

2.計畫對民眾具教育宣導之成果

本計畫所發展之系統主要是提供給疾病管制局的專業人員使用。經專業人員驗證並修正系統預測的結果，可發佈於公開的網站，以地圖的方式呈現某些傳染疾病發生在哪些地區，有助於一般民眾迅速找到相關的中英文媒體報導。

3.計畫對醫藥衛生政策之具體建議

本計畫所發展之系統可有效幫助疾病管制局從事中英文媒體疫情分析的人員，加速其進行媒體疫情的分類與地點研判，進而有助其預警重要疫情在媒體報導的可能趨勢。