透過 POI 的過期驗證以持續維護 POI 資料庫

張國斌
國立中央大學資工所碩士班
cheongkuokpan@gmail.com

張嘉惠
國立中央大學資工系教授
chia@csie.ncu.edu.tw

摘要

隨著智慧行動設備的普及率快速提升，查詢店家、地點等 POI(Point of Interest)資訊的服務也變成大家的日常所需，提供這種服務的背後需要有一個龐大的 POI 資料庫。在經過一段時間之後，這些資料庫的 POI 資料就不一定是最新的。如果使用者得到錯誤的資訊，將會浪費他寶貴的時間。所以如何讓 POI 資料庫保持在最新的狀態就成了一門關鍵的課題。我們希望透過持續更新資料庫，識別出已經停止營運的 POI，從而提供正確的 POI 資訊。

在本論文中，我們的系統目標在於在可行的時間內偵測資料庫內過期的 POI。方法分為兩個部分。第一部分為政府開放資料的使用，找出 POI 資料庫與開放資料共同擁有的 POI 以直接更新其狀態；第二部分則是利用網路資訊訓練 POI 過期驗證模型，偵測資料庫內已經過期的 POI。

關鍵詞：Location-based service、Crowdsourcing、Supervised learning。

1. 緒論

受到網路蓬勃發展的影響，電子地圖搜尋服務(如 Google maps、百度地圖)已經滲透到我們的日常生活中，人們透過這些服務找到感興趣的地點資料稱為 POI。POI 是由 WWW Consortium[1]所定義，基本上表示一個可用的地點資料，在地圖是有固定的位置，人們分辨一個 POI 是利用它的名稱或地址。

然而在瞬息萬變的社會上，商家的開業或是結業是一件很常見的事情.。因此，如果沒有持續的更新 POI 資料庫的話，前端的地圖搜尋服務便很有可能提供錯誤資料，導致使用者找不到原本規劃好要去的地方；反之，資料庫內的資料越正確，就越能在使用者的心中建立信心的旗幟。而直覺上讓 POI 資料庫保持新鮮的方法就是利用人工進行更新，但是一個 POI 資料庫動輒上百萬的資料量，這個方法就變得不可行。

在本論文中，我們嘗試偵測 POI 資料庫內已經過時的 POI，以防讓 POI 資料庫提供不正確的資料到使用者手上。此部分的挑戰在於缺乏有效的資料來源，店家開張有可能為了打響知名度而盡力地宣傳開張的資訊；而結業則是沒有太多的理由需要告知他人，所以一般都是默默的發生。比較可以利用的是政府開放資料—全國營業(稅籍)登記資料集和公司解散登記清冊(月份)，因為這些都是在持續更新的商家資料。透過比對需要驗證的資料，我們可以更新這些 POI 的營運狀態。不過在開放資料上也有使用上的困難之處，一、在開放資料裡所登記使用的名稱、及負責人並非營業時顧客所熟知的名稱。二、由於地址會出現可能多個表達型式。因此單靠開放資料不能夠解決所有 POI 的過期驗證。

由於網路的發達，利用網路上一些間接的資訊，例如該店家太久都沒有新的消息，那就表示有可能已經歇業了。這裡的挑戰主要有兩點，一、搜尋引擎提供商 Google 會限制單一 IP 在固定時間內可以進行搜尋的次數，這會是整個驗證過程中瓶頸；二、不容易設計出好的特徵，特徵的設計對於 POI 的過期驗證有很大的影響，不好的特徵很可能是導致驗證效能不足的原因。

本論文的研究是為了提供一個有效的 POI 資料庫更新方法，主要使用兩個方法。一、結合政府開放資料，比對需要被驗證的 POI 和政府開放資料以確定 POI 的營運狀態；二、利用網路上 POI 的相關資料，為 POI 萃取特徵以訓練分類模型，用於判斷 POI 是否已經過期。實驗結果顯示，使用我們的模組效能優於 Chuang[4]的模組。

2. 相關研究

要判斷兩個 POI 是否一樣時，有可能會面對 POI 名稱的歧義問題，一個 POI 名稱可以代表不同的 POI。Hu[6]等人利用具結構化的 DBpedia 資料來為 Wikipedia 上的地點名稱消歧義。不過一般的 POI 資料庫包含的 POI 種類很多，不一定可以為每個 POI 找到很多關於它的描述句子。另外，地址也有不同的描述方式，Lin[7]等人利用台灣郵政總局的 3+2 碼郵遞區號以完成地址標準化。但是隨著時間的推延，他們的方法不能解決老舊格式的地址。

偵測過期資訊的方法可以透過是否被新的資訊覆蓋得知，Tran[9] 等人研究從網路資訊中萃取出一些事實資訊以偵測維基百科的信息框中過期的資料。但是對於主要由地點名稱與地址組成的 POI，相同的地址或地點名稱有可能包含多個 POI，這個方法變得不可行。不過，對於一個 POI 是否過期可以視為二元分類問題。Chuang[411]等人提出基於網路弱標記資料來訓練分類模型。在訓練模型時所用到的特徵透過搜尋引擎--Google 獲得，搜尋所用的關鍵字分別是 POI 名稱、地址和 POI 名稱加地址三種，其實驗結果表示利用半監督式學習在 F 度量上得到 0.702 的分數。其後，Chuang[5]進一步改善使用的特徵，驗證的準確率達到 0.728。Chuang[5]分別是搜尋結果個數、POI 與地址之間的相關性、日期等資料。使用他們這個方法的好處是不需要實地到達每個 POI 的真實位置進行查看，但是 Chuang[5]所使用的特徵比較偏向是驗證一個 POI 的地址與 POI 的名稱的配對是否正確，忽略了一些對於 POI 過期與否有著較大代表性的特徵。我們訓練 POI 的過期驗證模型的方法與之不

[1] http://www.w3.org/2010/POI/documents/Core/latest

同的是我們設計特徵的方向主要是針對過期 POI。

Support Vector Machines(支持向量機，以下將以 SVM 表示)由 Vapnik 在 1963 年發表，很常被使用的分類模型。不過資料集需要是線性可分的條件大大的限制了這個方法，Vapnik[2]等人在 1992 提出用核技巧(kernel method，以下將以 kernel method 表示)來解決。而 SVM 一直以來都有面對二次規劃的問題，Platt[8]提出的 SMO(Sequential Minimal Optimization)解決了這個問題。

使用網路爬蟲常被網站管理者阻檔。Al-Bahadili[1]等人在提出在多核心的處理器上使用虛擬化的技術提升速度。不過他們所爬取的網頁都是來自不同的網站，我們則是針對 Google 搜尋引擎這個單一來源進行爬取，而 Google 搜尋引擎阻擋爬蟲的方法是限制單一 IP 在固定時間內可以使用的次數，所以我們每台的虛擬機器會被分配到一個實體 IP。

3. 系統架構及方法

本論文所針對驗證的 POI 資料集 POI_DB 是由 Chuang[3]在 2014 年對兩個黃頁(中華黃頁和愛評網)所爬取的資料構建而成的，裡面包含了約九十七萬筆 POI 資料。當中包含著一些舊格式的地址，例如桃園縣中壢市在 2014 年改名為桃園市中壢區。另外 POI 名稱也不一定是平常大眾稱呼該 POI 的名稱，有可能是商業類別、食物類別等、店主或擁有人的名字。

系統包含兩個驗證 POI 是否過期的部分，如圖 1 分別是左邊檢查 POI 在開放資料上的登記現狀和右邊 POI 的過期驗證。第一部分是利用 POI 的名稱和地址，在政府開放資料中查詢該 POI 的登記現狀，然後更新 POI 資料庫；第二部分則是 POI 過期驗證的主軸，透過搜尋引擎取得的資料進行過期預測，這是一個二元分類的問題，透過搜尋引擎獲得 POI 的相關資料，查詢包含「POI 名稱」、「POI 地址」和「POI 名稱 POI 地址」三種，再從搜尋結果萃取特徵，做為驗證 POI 是否過期的依據，最後依照結果更新 POI 資料庫的狀態。

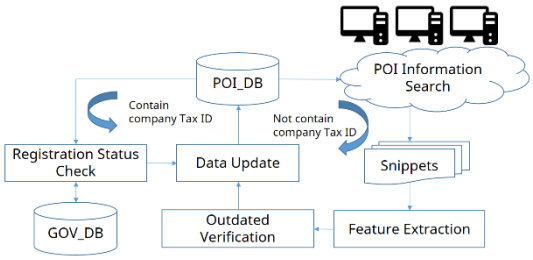

圖 1 系統架構圖

3.1 政府開放資料的利用

我們使用「全國營業(稅籍)登記資料集」和「公司解散登記清冊(月份)」，透過這兩份公開資料，我們可以直接知道我們資料庫中部分 POI 是否為過期。前者代表還沒過期的資料；後者則為已經過期的。只有當兩份資料的 POI 名稱與 POI 地址都完全一樣時，我們才會視為同一個 POI。另外，地址存在很多種不同的表達方式，我們使用「中華郵政 3+2 郵遞區應用系統[2]」的轉碼軟體來把地址進行標準化，再使用正規表達式把「里」與「鄰」刪除。

3.2 POI 驗證模型的建立

我們透過搜尋引擎收集 POI 的相關資料，利用它們擷取出數種特徵，最後訓練出模型以偵測資料庫中的過期 POI。搜尋字有三種，分別是「POI 名稱」、「POI 地址」和「POI 名稱 POI 地址」。利用這些相關資料，我們擷取出包含 Chuang[4]所使用到的特徵在內的六種不同類型的特徵，我們所使用的五種如表格 1 所示。以下是對它們的介紹。

表格 1 POI 過期驗證的特徵

類別 (個數)	描述
g_map (3)	搜尋結果中 google map 的標記狀態
date (9)	今天日期距離搜尋結果中最近日期的天數
official_site (1)	地址是否出現在官方網站中
eHowNet (22)	ehownet 上對於 POI 過期相關的詞彙
wrd_trn (70)	從訓練資料中找到對於 POI 過期相關的詞彙

Google 地圖資訊

在搜尋結果裡面有可能包含一個來自 Google Map 對 POI 描述的結果。對於這個資訊，我們使用三種狀態來表示，第一種是只有一般的網頁搜尋結果，並未出現 POI 搜尋結果、第二種是搜尋結果裡包含一個 POI 結果、第三種也是在搜尋結果中出現一個 POI 的結果，但是其包含「永久停業」的標示。對於這方面的資訊，我們會產生出 3 個特徵，分別來自三種搜尋字的搜尋結果。

與上次有消息的時間差

一個活躍或有在運作的 POI 應該隨著時間有可能出現不同的消息，如果網路上出現比較相關的訊息都比較陳舊的話，那這個 POI 有可能已經沒有在運作了；反之，POI 就有較高的機率是未過期的狀態。我們會分別產生對三種搜尋字的搜尋結果中前三名最新日期距離今天的天數作為特徵 (共 9 個)。

是否還出現在官網上

一些屬於連鎖店的POI，一般的情況下它們都會有自己的官方網站來告知民眾自己在甚麼地方有分店。官方網站一般都會有專人負責更新，盡可能不會出現錯誤的資訊。在

2 3+2 郵遞區號查詢應用系統；
http://www.post.gov.tw/post/internet/Download/index.jsp?ID=220306

這個特徵中，我們在搜尋詞「POI名稱」的搜尋結果中出現最多次的hostname當作該POI的官網。再利用官網的hostname和POI名稱組成搜尋詞「site: official_hostname POI名稱」再次對搜尋引擎進行搜尋，我們以這個搜尋結果中是否出現POI名地址產生出1個特徵。

廣義描述 POI 過期的詞彙

在人們形容一個 POI 已經過期會有一些常用關鍵字，例如倒閉、搬遷等，我們利用這些關鍵字作為種子，在詞庫中找出它們的同義詞。這些可以從廣義找到所有這類型明顯表達 POI 已經過期的關鍵字。我們選擇的詞庫為 E-HowNet (廣義知網知識本體架構)，E-HowNet 的詞庫小組從 2003 起開始打造這個詞庫，在他們的詞庫中可以找到更接近語意的同義詞彙。

描述 POI 過期的詞彙探勘

當我們形容一個 POI 已經過期可以有倒閉、結業等描述的詞彙，除了這些對於人們是明顯意思的詞彙之外，在搜尋結果裡面也應該出現某些與 POI 過期十分相關的詞彙。這些隱性的詞彙雖然對於我們不一定有太大的意思，那可能只是我們目前沒有辨法解釋。我們首先使用史丹佛的中文斷詞工具 對搜尋引擎的搜尋結果進行斷詞，再選取出與 POI 過期相關的詞彙。由於斷詞後的詞彙量太多，如果全部都使用會對後續的計算造成高額負擔，而其中也包含很多出現頻率不高的詞彙，所以我們設定最小支持度(minimum support)。其後，我們會計算各個詞彙對於 POI 過期的 information gain 和 chi-squared 分數，只有兩種分數都足夠高的詞彙才會被留下來。

取出 information gain 和 chi-squared 都是高分的詞彙的流程如下所述。對每個 POI 的相關資訊進行斷詞，只有最小支持度為 2.5% 的詞彙被保留。其後計算出它們 information gain 和 chi-squared 的分數，兩種分數的 <詞彙，分數> 序列都先經過遞增排序。之後會各自把序列的第一個詞彙，也就是分數最低的那個刪除，循環刪除直至序列總分是其原始序列總分的一半時停止。最後輸出兩種分數序列的交集，這是為了更進一步嚴謹的找出與 POI 過期相關的詞彙。最後留下來的詞彙共 70 個，每個詞彙獨立成為一個特徵，代表在三種搜尋詞所獲得的搜尋結果中有沒有出現該詞彙。

4. 實驗

在實驗的部分，我們將介紹資料集的出處以及測試資料的標記方法、觀察在不同的分類演算法下的效能表現、以各類型的特徵所訓練出的模組與基準模組的比較。

4.1 資料集

需要被驗證的 POI 資料集 POI_DB 是由 Chuang[3]在 2014 年對兩個黃頁(中華黃頁和愛評網)所爬取的資料構建而成的，裡面包含了 978,939 筆 POI 資料。營運中的政府開放資料集是使用全國營業(稅籍)登記資料集 GOV_OPENING，裡面包含 1,464,906 筆 POI 資料；沒有營運的政府開放資料集則是使用公司解散登記清冊(月份)資料集 GOV_OUTDATED，裡面包含了 75,605 個 POI 資料。經過地址標準化之後，GOV_OPENING 與 POI_DB 的交集包含 84,184 個 POI。GOV_OUTDATED 與 POI_DB 的交集包含 3,389 個 POI。

可以獲得 POI 過期資訊的來源並不多，常見的發現方法是店家的結業公告、一般民眾經過的發現、店家在以前有過一段活躍的時間，然後在某一個時間點突然開始不再出現新的消息。在 Google Maps 上，民眾可以提出 POI 已經不存在或關閉的修改意見，我們先隨機在 POI_DB 上抽取出 491,433 筆資料在 Google Map 上搜尋，其中 13,440 筆資料在回傳訊息中包含 POI 永久停業的訊息，然後再隨機抽出 2831 筆進行人工標記，標記結果包含 81 未過期資料，717 筆無法標記的資料，和 2,033 筆過期資料。被標記為過期的 POI，都是根據以下條件確保它們已經過期，以下按優先順序介紹它們的標記條件。

1. **官方的社群媒體發表停業公告**：利用社群媒體來宣傳以及增加接觸客戶的經營手法已經是越來越普遍了，如果 POI 的粉絲專頁上出現與客戶永久道別的貼文，那便代表這個店家已經沒有經營了，POI 資訊也會被標記為過期。

2. **在 POI 的官方網站上已經不存在該 POI**：官方網站主要的作用是用來宣傳和提供資訊，如果上面沒有找到該 POI 的話，那這個 POI 資訊有很高的機率已經過時了。

3. **食記或部落格上標示已停業**：網路上存在一些介紹 POI 的食記或遊記，例如愛食記、痞客邦等，他們主要是為了提供一些美食的資訊給民眾。而這些文章都是靠人手更新，如果上面有出現已歇業等關鍵詞，那有很高的機率可以相信這個 POI 資訊已經過時了。

4. **在 Google Maps 上被標示為永久停業**：在 Google Maps 上，民眾可以協助提供 POI 已經不存在或者關閉的修改意見)，一般都是出於善心，希望其他人可以獲得正確的資訊。這個動作對於回報者比較沒有利益可言，所以 Google Maps 上的過期標示也有它的可信性。但是每個人都有權利提出修改建議，難保一些因為同業競爭的業者會惡意提供錯誤的資訊傷害對手，所以為保障資料的確信性，POI 在 Google Map 上被標示為過期且一年內沒有留言的 POI 才會被人工標示為過期。

實驗所使用的測試資料中包含 4,080 筆 POI，過期的為 2,033 筆，未過期的為 2,047 筆。未過期的 POI 是隨機從全國營業(稅籍)登記資料集和我們的 POI 資料集的交集中取出。訓練資料包含 3,389 個過期 POI，3,401 個未過期 POI。未過期 POI 為全國營業(稅籍)登記資料集和我們的 POI 資料集的交集中隨機取出，且不與測試資料重複。過期 POI 則是公司解散登記清冊(月份)資料集和我們的 POI 資料集的交集。

4.2 評估

我們利用精準度 P (precision)、召回率 R (recall)、準確

率 ACC(accuracy)、F 度量 F1(F-Measure)來評估對 POI 過期驗證的效能。

我們使用所有的訓練資料，分別對應三種演算法 Naïve Bayes、Bayes Network 和 SMO 來訓練模型，使用測試資料來測試。結果如表格 2，Bayes Network 有最好的精準度；而 SMO 在召回率、F 度量和準確率表現最好。

表格 2 在 Naïve Bayes、Bayes Network 和 SMO 下過期 POI 驗證的效能比較

方法	P	R	F1	ACC
Naive Bayes	0.872	0.713	0.785	0.805
Bayes Network	**0.886**	0.751	0.813	0.828
SMO	0.809	**0.867**	**0.837**	**0.832**

我們將 Chuang 所使用的特徵與我們所使用的特徵進行比較。如圖 2，我們所使用的特徵在精準度上高出 0.315，召回率低 0.228，在 F 度量上高出 0.049，結合兩者的特徵取得最好的 F 度量 0.837。另外，為了評估我們的每個特徵的作用強度，我們將每一個特徵都獨立刪除，觀察缺少了它們之後的效能。結果如圖 3 所示，在上次有消息的時間差的這種特徵的表現最優，估計是因為在搜尋引擎沒有為 POI 找到較新的資訊時，POI 則有較大的機率處於過期的狀態；從訓練資料中找到對於 POI 過期相關的詞彙最差，我們推測是由於資料量的不足，這些詞彙過度擬合在訓練資料上了，在所有 POI 的相關資料上，它們不一定有較高的出現頻率。

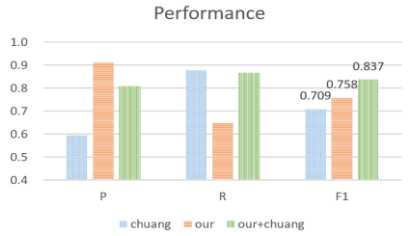

圖 2 比較 Chuang 和我們所設計的特徵的效能

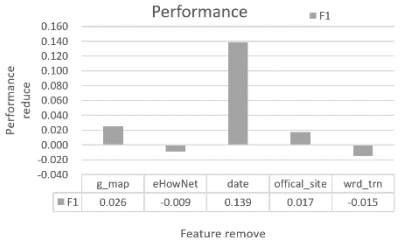

圖 3 比較各個我們所設計的特徵

在去除了從訓練資料中找到對於 POI 過期相關的詞彙這種特徵，以及加上 Chuang 的特徵後，我們對 SMO 中不同的 kernel 進行測試，結果如表格 3 所示，Pearson VII function-based universal 在各方面都表現出最好的效能，精準度為 0.908、召回率為 0.818、F 度量為 0.86 和準確率 0.869。

在一般的情況下，特徵使用的越多應該會讓效能越好，但是實際的情況不一定是這樣子，即使使用所有的特徵資料讓分類演算法訓練也不一定能夠找到最好的結果，每個特徵都是互補的關係，同時也各自包含不同程度的雜訊。我們產生出所有不同類別的特徵組合，嘗試在不同的 kernel 上進行實驗， Pearson VII function-based universal 表現出最高的 F 度量 0.91 (使用特徵分別為 Google 地圖資訊、與上次有消息的時間差和廣義描述 POI 過期的詞彙)。

表格 3 比較 SMO 中各 kernel 的效能

kernel	P	R	F1	ACC
polynomial	0.865	0.794	0.828	0.836
normalized polynomial	0.871	0.732	0.795	0.812
Pearson VII function-based universal	**0.908**	**0.818**	**0.86**	**0.869**
RBF	0.846	0.408	0.55	0.668

5. 結論

在本論文中，我們利用搜尋引擎從網路上收集 POI 的相關資料，搜尋字包含「POI 名稱」、「POI 地址」和「POI 名稱 POI 地址」三種。使用的特徵包含 google maps 的標記、最後出現的資訊時間、是否被官方網站所包含、eHowNet 上與過期 POI 有直接關係的詞彙和與過期 POI 有隱性關係的詞彙所形容的特徵。使用 SMO 算法與 polynomial kernel 可達到 F 度量 0.758，結合 Chuang[4]等人所設計可以達到 F 度量 0.837。最後，我們嘗試了不同的特徵組合下，最終以 SMO 算法與 Pearson VII function-based universal kernel 在特定的特徵組合(Google 地圖資訊、與上次有消息的時間差、廣義描述 POI 過期的詞彙等資料)上取得最高的 F 度量 0.91。

參考

[1] Al-Bahadili, H., Qtishat, H., & Naoum, R. S. (2013). Speeding up the Web Crawling process on a Multi-core processor using Virtualization. International Journal on Web Service Computing, 4(1), 19.

[2] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152). ACM.

[3] Chuang, H. M., Chang, C. H., & Kao, T. Y. (2014, September). Effective web crawling for chinese addresses and associated information. In International Conference on Electronic Commerce and Web Technologies (pp. 13-25). Springer International Publishing.

[4] Chuang, H. M., Chang, C. H. (2015, May). Verification of poi and location pairs via weakly labeled web data. In Proceedings of the 24th International Conference on World Wide Web (pp. 743-748). ACM.

[5] Chuang, H. M., Chang, C. H. (2016). POI Extraction and Relation Verification from the Web [Chuang, NCU, PhD Thesis]

[6] Hu, Y., Janowicz, K., & Prasad, S. (2014, November). Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia. In Proceedings of the 8th workshop on geographic information retrieval (p. 8). ACM.

[7] Lin Y. Y., Chang, C. H. (2014) Store Name Extraction and Name-Address Matching for Geographic Information Retrieval [Lin, National Central University, masters Thesis].

[8] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

[9] Tran, T., & Cao, T. H. (2013). Automatic Detection of Outdated Information in Wikipedia Infoboxes. Research in Computing Science, 70, 211-222.