網頁商家名稱擷取與地址配對之研究

Store Name Extraction and Name-Address Matching on the Web

林育暘　Lin Yu-Yang, 張嘉惠　Chang Chia-Hui
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
101522052@cc.ncu.edu.tw, chia@csie.ncu.edu.tw

摘要

行動裝置的普及造就大量地域性查詢的需求，其中最常見的一種查詢，就是尋找附近的餐廳或加油站。然而當使用者在電子地圖上搜尋這些地點名稱 (POI，Point of Interest)時，經常無法找到，因為電子地圖上雖有地點名稱標註，但是相關資訊不足，而這些資訊其實大多可以在網頁中找到。因此使用者大多必須開啟瀏覽器搜尋商家名稱找出地址，並把地址輸入至電子地圖查詢路線。但行動裝置螢幕小，且輸入文字不便利，如果要反覆查詢將是一件耗時耗力的工作。如果這時候有一個商家地理資訊系統能事先將網路上的商家資訊進行整合，最後提供一個 APP 直接讓使用者查詢，將可以大幅度減少使用者與裝置間的互動次數，有效的提供便利性。

為建構商家地理資料庫，Chuang 等人[9]對於包含地址網頁提出三種抓取程式，並利用 Chang 等人[2]的地址擷取程式，取得大量中文地址。Li 與 Chang[1]並定義地址相關資訊擷取問題，希望藉此豐富每個 POI 的相關資訊，提高地理資訊檢索(Geographical Information Retrieval, GIR)的召回率。然而不論是 Li 與 Chang[1]或 Chang 等人[2]或 Chuang 等人[9]的相關資訊擷取方法都僅能從多筆地址網頁擷取資訊，所得資訊有限，對於單筆地址網頁的相關資訊擷取仍尚無研究。

本研究從地址的角度出發，利用已抓取大量包含地址的網頁，先找出網頁中的地址，再藉由地址找出對應的商家名稱進行配對。換言之，給定一個已知地址，我們希望能透過網路資料擷取出該地點的名稱（如：商家名稱、政府單位…等）。舉例而言：當我們已有地址「新北市板橋區中山路二段 88 號 3F」，我們希望能知道這個地址對應的名稱「大鈞醫學美容診所」，如此即可進一步藉由地址、名稱並利用搜尋引擎收集更多額外商家資訊。這些額外資訊不僅可以有效提昇 GIS 檢索系統的召回率，也可提昇商家分類的準確率[10]。

在辨識商家名稱的部分，本篇論文使用了條件隨機域 (Conditional Random Field)[3]當作學習演算法。目前有許多關於中文組織名稱辨認的研究 [4] [5] [6] [7]，可以從新聞或一些較正式的文章中萃取出組織名稱，但是並沒有嘗試以一個 CRF-Model 直接對各種網站中的整個網頁內容進行中文組織名稱辨認。這兩者之間不同處在於新聞類文章屬於較正式的文章體裁，因此容易出現行政機關與正式的組織名稱，例如：行政院和維德食品有限公司，但是整個網路上商家組織

名稱的命名方式傾向則不同，例如：吼牛排、努哇克咖啡、阿嬤祖傳菜包肉粽仙草…等，都是商家組織名稱。另外，一個完整的網頁內容有結構與非結構化的資訊交錯呈現，雖然結構化資訊會造成自然語言文字內容的破碎，但這些結構也隱含有可利用的資訊。

為了使商家辨識能以最少人力進行自動化學習，本研究使用自動標記方式建立訓練資料，我們先針對部份的黃頁網站撰寫 Parser 取得大量商家名稱與地址的組合，並以這些已經取得的商家名稱對網頁語料進行自動標記，再利用自動標記後的語料訓練 CRF 序列標記模型。然而一個地址可能出現在多個網頁之中，僅只仰賴其中一個網頁也有失之偏頗之慮，因此我們也收集了 Google Snippets 當作訓練資料進行比較。本篇論文的第二個主題則是商家地址的配對，由於一個網頁可能包含多個商家名稱，我們對網頁以簡單的規則進行分類後，使用了啟發式（heuristic）的配對規則，利用各類型的網站所具有的表達特性，對地址與商家名稱進行配對。

本研究承續 [8] [9]之研究，經由爬取網頁上包含地址的大量網頁（包括 Yellow Page 與 Surface Web）進行商家名稱擷取。其中 Yellow Page 提供了大量商家名稱以及地址與商家的配對資料，而 Surface Web 則利用 [2]之地址擷取模型擷取出了可能含有台灣地址的網頁與地址清單。本篇論文以已知可能含有台灣地址的中文網頁、每筆網頁的地址清單、大量商家名稱清單以及已知的地址與商家名稱配對資料為基礎，提出了一個商家名稱擷取系統，方法分為四大步驟：地址網頁的前處理、商家名稱命名實體辨認、及地址－商家名稱匹配。本研究在三個模型聯合標記商家名稱的方式下，地址與商家名稱的平均配對正確率為 0.57。

關鍵詞：商家地理資訊檢索、商家名稱擷取、商家名稱與地址配對、序列標記、條件隨機域

Keywords: POI, store name extraction, name-address matching, sequence labeling, conditional random field

參考文獻

[1] S.-Y. Li, Application and Extraction of Postal Addresses and Related Information, National Central University, 2009.

[2] C.-H. Chang, C.-Y. Huang and Y.-S. Su, "Chinese Postal Address and Associated Information Extraction," The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 2012.

[3] L. D. John, M. Andrew and N. C. Fernando, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," ICML Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282-289, 2001.

[4] Z. Suxiang, Z. Suxian and W. Xiaojie, "Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields," Natural Language Processing and Knowledge Engineering, pp. 229-233, 2007.

[5] Y. Xiying, "A Method of Chinese Organization Named Entities Recognition Based On Statistical Word Frequency, Part Of Speech And Length," Broadband Network and Multimedia Technology (IC-BNMT), pp. 637-641, 2011.

[6] L. Yajuan, Y. Jing and H. Liang, "Chinese Organization Name Recognition Based on Multiple Features," Pacific Asia conference on Intelligence and Security Informatics, pp. 136-144, 2012.

[7] C.-W. Wu, R. T.-H. Tsai and W.-L. Hsu, "Semi-joint labeling for chinese named entity recognition," Proceedings of the 4th Asia information retrieval conference, pp. 107-116, 2008.

[8] Y.-S. Su, Associated Information Extraction for Enabling Entity Search on Electronic Map, National Central University, 2012.

[9] H.-M. Chuang, C.-H. Chang and T.-Y. Kao, "Effective Web Crawling for Chinese Addresses and Associated Information," in EC-Web, Munich, Germany, 2014.