

社群網路之資料探勘分析與 旅遊路徑規劃

文·圖 / 溫郁婷、彭文志

大數據資料處理與分析近年來受到產業界與學術界的重視，幾乎所有的公司，均投入大量的研發能量在相關研究。提到大數據，最常見的定義就是其4V特徵，分別為資料量規模（Volume）、資料格式的多樣化（Variety）、資料及時處理的速度（Velocity）與資料的準確性（Veracity）。舉例來說，社群網路即為典型的大數據資料，滿足了Volume（Facebook至今擁有15億持續活動的使用者）、Variety（文字描述、打卡、照片與影片等資料型態）、Velocity（平均每天有4億筆tweets產生）與Veracity（因為網路訊號不穩或是GPS誤差導致發表訊息的錯誤或缺漏）特徵。然而大部分的產業界如Google、Facebook都不願意將大量的原始資料公開，主要是透過APIs等方式，限制第三方商用。一般的產業對這些資料更是保守。因此，在從事大數據相關的研究方向，所面臨的第一個問題就是資料如何取得。與此同時隨著科技進步，現代社會非常依賴手機、行車紀錄器等行動裝置：我們在接聽電話或傳送簡訊時會透過最近基地台提供的無線通訊網路；使用地理服務如導航、鄰近景點推薦時都需要打開GPS定位；許多人會在旅遊或日常生活時將自己所在的位置打卡（check-in）或是發佈照片於Facebook, Foursquare

或Twitter等行動社群網路（Location-based Social Network），增加與社群上成員的互動。換句話來說，我們每天享受科技服務的同時都製造大量的地理位置紀錄，這些位置紀錄串聯成反映個人行為的軌跡資料。因此，在過去的研究中，我們已在社群資料的分析中，有了很多重要的研究成果，在此首先針對軌跡資料的架構所一個簡單的介紹，接下來簡介如何研發資料探勘演算法於旅遊軌跡推薦與社群影響力分析。

圖1所示為一軌跡資料探勘之架構圖，現有的軌跡資料探勘技術可依

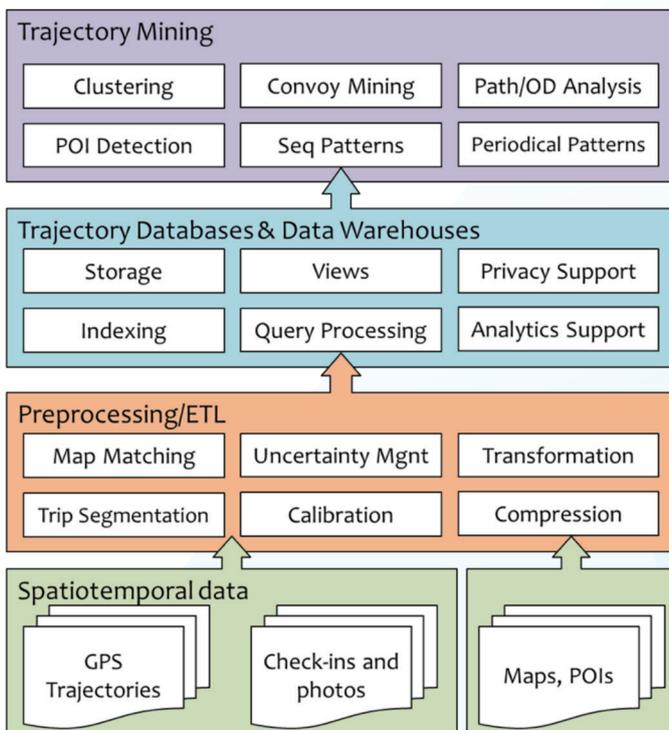


圖1：軌跡資料探勘研究層級^[4]

處理層級歸納成四類：（1）蒐集GPS軌跡或照片打卡等時空間資料（spatiotemporal data）與相關地圖、景點資訊，這部分的原始資料往往非常大量且以不同格式儲存，具有Volume與Variety特徵；（2）資料前處理階段，將軌跡資料透過對應、分段、標準化、不確定性檢測、壓縮等處理而達到Veracity特徵；（3）建立軌跡資料庫&資料倉儲，用以處理持續快速產生的資料（Velocity）；（4）運用資料探勘技術探索出實用的語意如景點、軌跡模式等。擁有這些知識後，我們可以應用在現實生活中如都市規劃、車流預測、定位廣告投放及旅遊行程規劃。

探勘社群媒體所隱含之旅遊經驗

以旅遊為例，一般人在規劃旅遊路徑時，有兩種模式，第一種是找旅行社選擇既定的行程，一切由旅行社規劃；第二種是近年來越來越盛行的自助行。第一種模式有可能遇到旅行社提供的行程時間或地點不盡理想，而第二種模式下，若是對遊玩的地點不熟悉，自行規劃旅遊路徑不僅耗時耗力，也可能造成行程規劃上出現時間計算錯誤或是想玩的景點沒玩到的遺憾。目前已經有一些平台提供旅遊路徑推薦的服務，如Yahoo! Travel根據平台上分享的旅遊路徑和使用者的偏好，搜尋符合的旅遊路線。不過現今大部分的相關服務只是列出分散的旅遊景點，或是僅依照旅遊路徑分享頁面的點擊人次作為推薦的排名，缺乏一套客觀評估的推薦機制。若能從他人的地理和軌跡資料中探勘出使用者移動模式，這些移動模式代表著使用者的經驗，藉此當作人們旅遊行程的依據，進而根據使用者的需求推薦適合的旅遊行程，即可免去遷就旅行社規劃好的行程或是花了大量時間規劃卻不滿意的旅遊路徑。

針對此問題，交大資料探勘研究團隊在2015年發表了最新研究論文^[1]，目標是能依據使用者給定的旅遊地點和時間，從使用者參與的行動社群網路中探勘出使用者的興趣偏好與人們的移動軌跡，用以自動規劃適合使用者的行程。舉例來說，考慮現今社群網路的廣泛應用，據調查

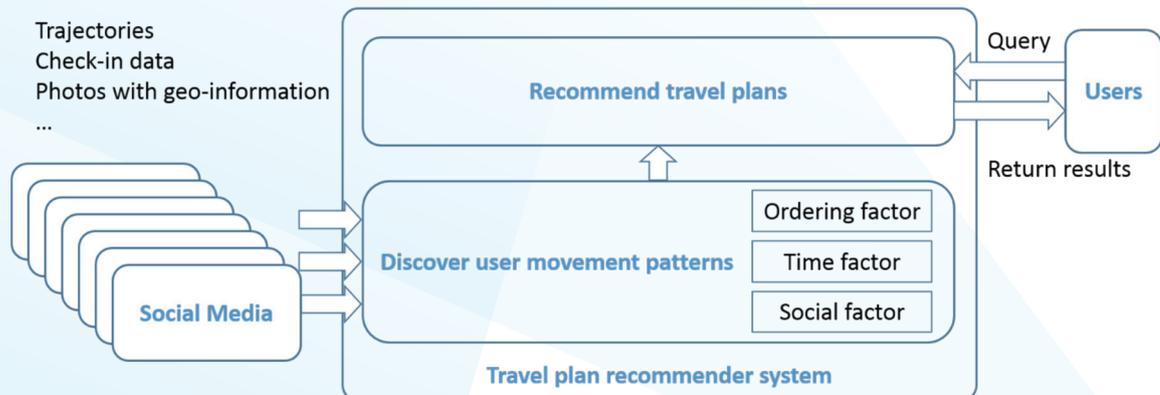


圖2：行程規劃推薦系統架構

有76%的旅行者會將旅行途中照片或是地理位置公開在社群，更有52% Facebook使用者反應會受到朋友圈的旅行照片影響其旅行相關決策，因此，蒐集了包含朋友圈關係網路及所有使用者的照片打卡紀錄的Facebook及Foursquare資料，在前處理階段先整合每人在一定時間區段內連續的地理位置紀錄（每段可視為一條旅遊軌跡）成一個龐大的軌跡資料庫，再定義影響軌跡推薦的特徵機制，最後組合出數條適合使用者的個人化推薦行程。

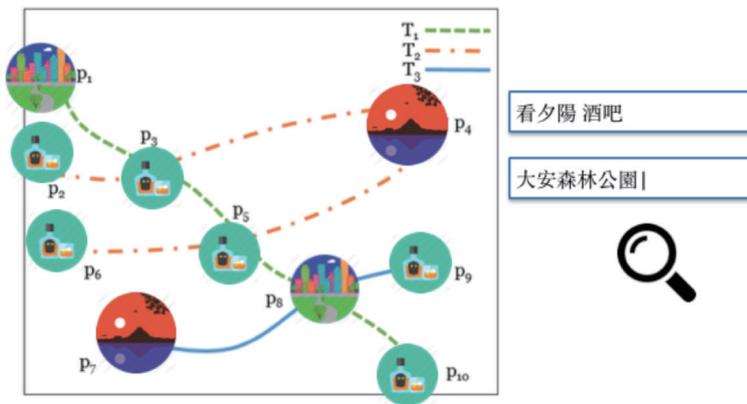


圖3：個人化旅遊推薦情景

且舉圖3的例子來說明，資料庫中有三條來自不同人的軌跡記錄（由景點 p_1, \dots, p_{10} 組成），在使用者想要旅行的地區內，同時使用者提供了一些文字來描述他的興趣。為了推薦結果的靈活與多樣性，我們不只考慮既有的軌跡而是會依照需求重組景點，再依照定義

的推薦機制得到最終的行程。那麼如何給定客觀的推薦機制呢？我們歸納出使用者規劃旅程時的三個考量：（1）拜訪符合興趣的熱門景點；（2）合適的時間安排；（3）朋友圈的行程分享或推薦也可能會影響使用者的決定，可稱為社群影響力（social influence）。

景點熱度	合適時間	社群影響力 ^[2]
<p>將軌跡拆解成兩兩景點間的移動片段，每條有向邊線上的權重代表由一端移動到另一端的移動機率。</p> <p>使用隨機漫步理論（random walk）模擬使用者在這些地點間任意移動，可透過馬可夫矩陣的運算結果找出每個景點停留機率。</p>	<p>因景點的特性不同（如夜市vs.公園），訪客的拜訪時間分布也會不同，以此為根據推測各景點的適合拜訪時段。</p> <p>我們先做一次常態分布的轉換以避免資料過度稀疏而不連續的問題，再將各時間段對應的機率值正規化，得到量化的分數。</p>	<p>中心概念為社交網路中關係越緊密，兩者相互影響力越大。我們使用「跟隨朋友拜訪同一地點」的機率做為使用者之間受影響力的程度。</p> <p>考慮影響力的傳遞性與遞減趨勢，我們使用熱擴散（heat diffusion）模型模擬影響力的傳播。</p>

社群影響力分析

圖4將行動社群網路結構化，可分為記錄使用者虛擬網路間朋友關係的social layer與景點拜訪順序的location layer，並透過使用者的拜訪行為（check-in）連接此social layer與location layer。我們認為「已經發生的事件可能影響未來事件的發生，反之，未來的事件對過去的事件的影響力相對微弱」，故必須加入時間維度記錄不同使用者拜訪同一景點的先後順序。

如圖5所示，給定任一使用者，能找出一組有影響力的使用者，能令給定使用者「跟隨」他們的腳步拜訪他們曾經去過的地點。以「跟隨機率」定義好所有邊線的權重後，我們使用熱擴散（heat diffusion）模型模擬影響力在行動社群網路中的傳播，以得到表示此網路結構的特徵值，用以做為使用者評分標準。

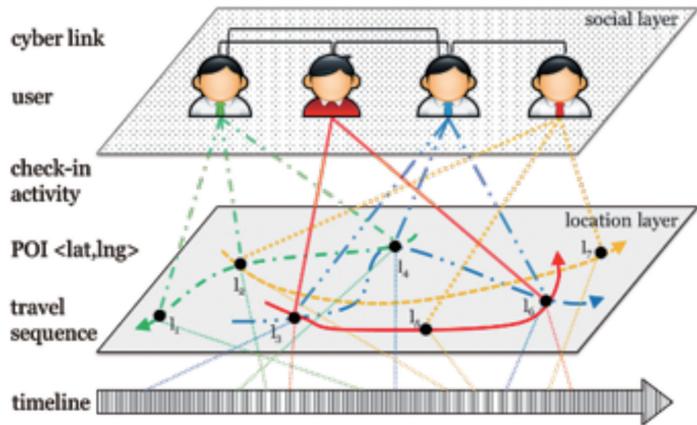


圖4：行動社群網路結構，記錄了使用者於虛擬網路間的朋友關係、使用者與地理位置間的行為記錄以及時間軸。

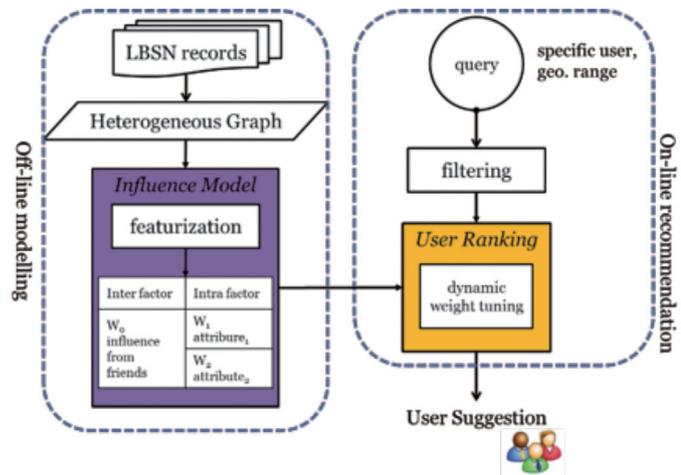


圖5：社群影響力探勘系統架構

結語

本文探討了透過軌跡探勘來找出虛擬社群與現實生活移動行為的影響力分析，著重在人與人之間的交互影響力，將社群中的朋友依信任度排序，得到可靠的使用者推薦結果；更應用此結果於自動化旅遊行程推薦，結合從他人的地理和軌跡資料中探勘出的歷史旅遊軌跡，開發出新的旅遊模式。現階段的研究還未針對社群網路中不斷產生新資料，達到有效率的更新所探勘的旅遊資訊，未來的研究方向可往資料的壓縮與更新，並深入探討隨後的隱私安全保護問題。

社群網路的研究是大數據領域一個非常重要的議題，不只限於本文章提及的旅遊相關產業，網路中的每個物件（使用者、行動裝置、交通工具等）都可以視為是一個不斷產生軌跡資料的感應器，透過這些資料可以進行都市規劃、事件預測、廣告投放、熱點偵測等各式各樣的應用。隨著相關應用的開發，人們的生活模式在近幾年以極快的頻率在變動，善用大數據背後的知識，便可創造商機並掌握社會趨勢。（本專題策畫／電資學院陳銘憲院長）

參考文獻：

- [1] Yu-Ting Wen, Kae-Jer Cho, Wen-Chih Peng, Jinyoung Yeo, Seung-won Hwang, “KSTR: Keyword-aware Skyline Travel Route Recommendation” , ICDM 2015.
- [2] Yu-Ting Wen, Po-Ruey Lei, Wen-Chih Peng, Xiaofang Zhou, “Exploring Social Influence on Location-Based Social Networks” , ICDM 2014.
- [3] Wan-Ting Hsu, Yu-Ting Wen, Ling-Yin Wei, Wen-Chih Peng, "Skyline Travel Routes: Exploring Skyline for Trip Planning", HuMoComp 2014.
- [4] Yu Zheng. Trajectory Data Mining: An Overview. ACM TIST 2015.
- [5] Hsun-Ping Hsieh and Cheng-Te Li. Mining and Planning Time-aware Routes from Check-in Data, CIKM02014.



溫郁婷小檔案

目前為交通大學網路與資訊系統博士班二年級學生，由彭文志教授與中研院陳伶志研究員共同指導。研究領域為資料探勘與機器學習，主要包含軌跡資料探勘、社群影響力分析、資訊擴散模型與推薦系統等。溫郁婷於 2011 年獲頒國科會（現為科技部）大專學生參與專題研究計畫之研究創作獎。



彭文志小檔案

臺灣大學電機工程系博士（2001），現為交通大學資訊工程系教授。研究專長為資料探勘、社群資料探勘、感測資料處理與行動資料等方面，尤其在軌跡資料探勘方面，已經發表一系列的研究成果，隨著物聯網與大數據的發展，所產生的主要資料型態包含時間與空間的資訊，這些資料可視為軌跡資料。彭教授除了有頂尖期刊與國際會議論文的發表，亦將部分成果技術授權給 hTC（宏達電），並且有多項的國內外產學合作計畫，希冀能將大數據、資料探勘與資料科學的技術，帶入更實務的資通訊系統與服務。2011 年獲頒中國電機工程師學會優秀青年電機工程師獎，並於 2012 年獲頒國科會（現為科技部）吳大猷先生紀念獎。