105年度研究報告

應用雙北開放資料之旅遊推薦創新模式設計

研究人員 / 團體名稱： 王斯霈、蔣宜衡、陳孟勛、勤業眾信

Map InterLinK

摘要

近年來隨著國民生活習慣的改變，越來越多的民眾會在週休二日或其他休假時間出遊放鬆，多則三五好友相約聯絡感情，或是做為家庭成員間的回憶紀錄，少則情侶間的浪漫點滴，或是個人的一趟清淨之旅。面對旅遊觀光上種種不同的需求、不同的意見、及不同的目的，要如何在多如繁星的旅遊景點中，選擇出最合適當前情境與心情的地點，常常會是一個難題。

單就大臺北地區的臺北市及新北市而言，就有許多可供休閒遊憩的知名景點，但每每到了假日，卻常有許多市民因此陷入選擇障礙中，想不到要到哪裡休憩才更能符合當前的需求。像這樣的困境，過往可能肇因於資訊的匱乏，但在當前的世代中，卻反而起因於資訊爆炸(Information Explosion)[1]，現代人時常需要面臨資訊超載(Information Overload)[2]的困境，所以如何協助使用者作出決策，也是當前值得處理的重要議題。

為了不浪費這些寶貴的自然與人文資源，我們想到何不將雙北市這些的旅遊景點資訊，再整合旅遊網站的評價，並透過資料整理和分析後，藉由開發一個新型態的推薦系統(Recommender Systems)[3]。跟傳統的推薦模式不一樣，本研究會先利用本國遊客及外國遊客等不同的身分來做分群，讓市民與遊客只要透過給予系統目標條件，像是出遊目的、景點類別等等，即可依照使用者需求推薦景點。

[1] 資訊爆炸(Information Explosion)：指快速增長的資訊量及大量資訊所產生的影響。資訊量的增加，使得人們的生活更加便捷，然而同時也讓管理資訊變得更困難，現代人普遍面臨資訊超載，並因而引發身體不佳、沮喪、挫折、判斷力減弱、資訊焦慮、工作滿意度降低等症狀，如何協助人們管理或降低資訊超載情形，使其不致逃避或拒絕資訊，成為資訊社會的重要課題。

[2] 資訊超載(Information Overload)：指接受太多資訊，反而影響正常的理解與決策。例如擁有大量的歷史資訊或不斷增加的新資訊，在缺乏方法去比較或是分析龐雜的資訊時，將很難去分辨哪些資訊與所需做成的決定有關，導致無法順利決策。

[3] 推薦系統(Recommender Systems)：屬於資訊過濾的一種應用，能向目標使用者推薦其可能感興趣的項目。。

另外一個更特別的地方是，本研究可以利用景點遊客數量統計的歷史資料，來推估該景點的尖峰及離峰時間，並依據遊客為本國人或外國人士等資訊進行差異推薦。有鑑於並非台灣各地區的旅遊資訊都有達到足以做有效分析的開放程度，所以本研究目前暫時鎖定在臺北市及新北市地區，做為推薦系統當前的建置目標，我們也非常期待如果未來能蒐集到全國的資料後，得以擴大系統的推薦範圍及服務人群。

目錄

圖目錄

第一章、研究主題

1.1 符合城市發展方向與願景之新方案或新制度

經由整合分散各處的多樣化資料並從中觀察出有益資訊，對於一座城市的資訊化發展與各方面的未來展望，都能起到很大的幫助，通過見微知著的觀察與綜觀全局的視野，能夠有效減少資源浪費、增進施政效能、改善市政規劃，同時更能降低人力資源的消耗，因此本研究即試圖從在這方面著手，並優先選定旅遊觀光作為首要的目標對象。

首先本研究藉由結合開放街圖(OpenStreetMap，簡稱 OSM)[4]與 Google 地圖(Google Maps)[5]的地圖資料，來完善本研究系統的圖資來源，通過運用 OSM 開放街圖豐富的地圖資料來準確定位景點及周邊設施的座標與範圍，並利用 Google 地圖提供的 API[6]來提供在路線規劃與交通時間預估上的服務，隨後也以本研究進行整合的這份地圖資訊，來完成本系統使用的地圖應用。

而就城市經濟及文化等層面的發展情況來說，也可以通過提供市民選擇旅遊去處及協助到訪遊客的觀光規劃，來達到促進消費和經濟發展，並加深市民

---

[4] 開放街圖(OpenStreetMap)：簡稱 OSM，是建構自由內容的線上地圖協作計劃，由英國人史蒂夫·克斯特(Steve Coast)創立。地圖圖資是由註冊的使用者以手持 GPS 裝置、行動裝置、航空攝影相片、衛星影像、其他自由內容，或依靠使用者對有關區域的本地知識繪製，使用者可上傳 GPS 路徑，並可編輯地圖的向量資料。地圖的向量資料以開放資料庫授權方式授權。網站由英國非營利組織 OpenStreetMap 基金會贊助維護營運。

[5] Google 地圖(Google Maps)：是 Google 公司向全球提供的電子地圖服務，地圖包含地標、線條、形狀等資訊，提供矢量地圖、衛星相片、地形圖等三種視圖，提供路線規劃、交通時間預估、地址與座標搜尋等功能，並且附有 Google 街景(Google Street View)，能提供水平方向 360°及垂直方向 180°的街道全景，讓使用者能檢視所選城市地面上街道不同位置及其兩旁的景物。

[6] 應用程式介面(Application Programming Interface)：簡稱 API，是軟體系統不同組成部分銜接的約定。是讓應用程式開發人員得以呼叫一組子程式功能，而無須考慮其底層的原始碼為何或理解其內部工作機制的細節。API 本身是抽象的，它僅定義了一個介面，而不涉及應用程式在實際實作過程中的具體操作。

對城市的文化認同(Cultural Identity)[7]；而在面對不同群眾的需求後，也會產生可以持續累積的綜合推薦成效，並期許以此在城市未來休憩地點的發展側重、旅遊觀光的推廣績效上亦能產生有效助力。

通過本研究開發的網路爬蟲(Web Crawler)[8]程式，在針對知名的旅遊網站TripAdvisor(貓途鷹)[9]擷取景點評價等網路上的公開資訊後，能夠獲取過去到訪遊客對臺北地區在景點、旅館、餐廳等地點的評價資料。經過本研究對這些資料的收集與整理，並做適當的資料探勘(Data Mining)[10]與資料分析(Data Analysis)[11]之後，通過搭配臺北市及新北市政府釋出的政府開放資料(Open Data)[12]，這份整合多方來源的資訊將能成為一份寶貴的知識來源，在本系統中能為我們連結使用者與推薦標的，而日後亦可以提供給臺北市及新北市政府的各局處業務，如觀光、交通、文化、都市發展、警消等，作為未來精進各項作為或維持優質成果的方向。

[7] 文化認同(Cultural Identity)：是對一個群體或文化的身份認同，又或者是指個人受其所屬的群體或文化影響，而對該群體或文化產生的認同感。

[8] 網路爬蟲(Web Crawler)：是一種自動化瀏覽網路的程式，可以自動採集所有能夠存取到的頁面內容，通常被廣泛運用於網際網路搜尋引擎，以取得或更新這些網站的內容和檢索方式。

[9] 貓途鷹(TripAdvisor)：是一個國際性旅遊評論網站，提供世界各地飯店、景點、餐廳等旅遊相關資訊，也包括互動性的旅遊論壇。目前網站提供 28 種語言，並擁有超過 53 萬個景點、165 萬家旅館和 270 萬家餐廳的資訊，每月可觸及超過 3 億 4 千萬名不重複的訪客。

[10] 資料探勘(Data Mining)：一門從大量資料或者資料庫中提取有用資訊的科學，目標是從資料中提取出隱含而過去未知的有價值潛在資訊。是用到人工智慧、機器學習、統計學和資料庫的交叉方法在相對較大型的資料集中發現模式的計算過程。能從資料集中提取資訊，並將其轉換成可理解的結構，以進一步使用。會對大規模資料進行自動或半自動的分析，以提取過去未知的有價值的潛在資訊。

[11] 資料分析(Data Analysis)：是將隱沒在龐雜資料中的資訊集中、萃取和提煉出來，以找出研究對象的內在規律，可以幫助人們作出判斷，以便採取適當行動。

[12] 開放資料(Open Data)：政府各機關於職權範圍內取得或做成，且依法得公開之各類電子資料，包含文字、數據、圖片、影像、聲音、詮釋資料(Metadata)等，以開放格式於網路公開，提供個人、學校、團體、企業或政府機關等使用者，依其需求連結下載及利用。

1.2　對於各機關業務提出明確問題或研提改進作為

本研究運用到北市府觀光傳播局、交通局、文化局、產業發展局、工務局、都市發展局、捷運工程局、地政局等單位資料。

在觀光傳播局的資料中，我們發現到部分資料並不是每年都有的情況，也有些資料並不能夠直接拿來運用，像是河濱自行車道.csv 資料集，即只有描述性資料，在使用上會發生不少限制及困擾。

第二章、研究方法

2.1 基本理論假設

2.1.1 協同過濾 (Collaborative Filtering)[13]

協同過濾為目前較為廣泛的推薦技術，它被應用於多個不同應用領域的環境中，如文章、網頁、電影及產品的推薦。主要是依據使用者間對項目的偏好程度來進行相似度的分群，並將每個分群中的目標使用者找出最相似偏好的鄰居，並依其偏好來予以推薦。

協同過濾有以下優點：

1. 個人化推薦：針對每個目標使用者，皆可以通過分析偏好，做出個人化的預測結果。

2. 無需內容分析：只需要運用使用者對項目的評分紀錄，通過分析使用者間或項目間的相似度，即可做出預測。

3. 發現使用者新的興趣：通過比對出的使用者或項目之同儕團體(Peer Group)[14]來做出預測，常能推薦出雖然使用者未接觸過，卻可能感到興趣的項目。

[13] 協同過濾(Collaborative Filtering)：是目前較為廣泛的推薦技術，主要是依據使用者間對項目的偏好程度來進行相似度的分群，並於每個群集中找出與目標使用者最相似偏好的使用者，依找出對象的偏好來給予目標使用者推薦。

[14] 同儕團體(Peer Group)：具同質性的人組成的群體，成員的信念及行為會受到彼此的影響。

4. 自動化(Automation)[15]程度高：能夠編寫程式來處理，通過自動化的提取歷史資料進行分析，能夠免除人工處理，自動彙整出預測結果。

協同過濾有以下缺點：

1. 冷啟動(Cold Start)[16]問題：在使用者初次進入系統時，因為尚無目標使用者的歷史資料，會導致針對該使用者在預測上將面臨困難。

2. 稀疏性(Sparsity)問題：一般使用者鮮少主動提供對於各項目的評分，這種情況將容易產生出稀疏矩陣(Sparse Matrix)[17]，導致可供預測的資料相當少，降低推薦成效。

3. 系統延伸性(Scalability)問題：當資料量過於龐大或快速累積時，計算量與運算耗時都將大幅增長，導致系統需要一定的運算過程，而難以即時性的給予結果。

[15] 自動化(Automation)：是一門綜合性技術，和資訊理論(Information Theory)、系統工程(Systems Engineering)、計算機技術(Computer Technology)、電子學(Electronics)、自動控制(Automation Control)等多門學科都有十分密切的關係，而其中又以控制理論(Control Theory)和計算機技術對自動化技術的影響最大。自動化的最大好處是可以節省人力資源，同時也可用於節約能源和材料，並改善品質(Quality)、準確度(Accuracy)和精密度(Precision)。

[16] 冷啟動(Cold Start)：是在初始階段因為缺乏資料，導致無法順利進行預測的困境，包含新使用者和新項目的問題，皆是因為缺乏對象的評分紀錄，而會導致無法對目標使用者的喜好進行準確的預測。

[17] 稀疏矩陣(Sparse Matrix)：是其元素大部分為零的矩陣。在科學與工程領域中求解線性模型時，經常出現大型的稀疏矩陣。在使用電腦儲存和操作稀疏矩陣時，經常需要修改標準演算法以利用矩陣的稀疏結構，例如以矩陣的稀疏特性，通過壓縮大幅節省稀疏矩陣的記憶體代價。

2.2 研究方法

協同過濾推薦系統(Collaborative Filtering Recommender Systems)可以分為User-Based(基於使用者)與 Item-Based(基於項目)兩種：

2.2.1 基於使用者之協同過濾 (User-based Collaborative Filtering)

分析整體使用者的資料，運用統計方法找出具有相似愛好或興趣的使用者數學模型(Mathematical Model)[18]，用這些相似的使用者所找出的數學模型來預測特定使用者面對不同產品的喜好。

一、 收集使用者資訊

收集足以代表使用者興趣的資訊。為一個二維空間的矩陣，X 座標為項目，Y 座標為使用者，並在對應的空格內填入喜好程度的評分。

二、 最近鄰搜尋 (Nearest Neighbor Search)[19]

最近鄰搜尋(Nearest Neighbor Search，簡稱 NNS)，基於使用者之協同過濾(User-based Collaborative Filtering)的出發點是與目標使用者興

[18] 數學模型(Mathematical Model)：是使用數學概念和語言來對一個系統的描述。建立數學模型的過程叫做數學建模。數學模型不只用在自然科學(如物理、生物學、地球科學、大氣科學)和工程學科(如電腦科學，人工智慧)上，也用在社會科學(如經濟學、心理學、社會學和政治科學)上；其中，物理學家、工程師、統計學家、運籌學分析家和經濟學家們最常使用數學模型。模型會幫助解釋一個系統，研究不同組成部分的影響，以及對行為做出預測。

[19] 最近鄰搜尋(Nearest Neighbor Search)：簡稱 NNS，是一個在尺度空間中尋找最近點的優化問題，多數情況下距離是由歐幾里德距離(Euclidean Distance)或曼哈頓距離(Manhattan Distance)來決定。

趣愛好相同的另一組使用者，也就是找出目標使用者的同儕團體，接著計算兩兩使用者間的相似度。

一般會根據資料的不同，進而選擇不同的演算法，目前使用上較多的相似度演算法有皮爾森相關係數(Person Correlation Coefficient)[20]、餘弦定理相似度(Cosine-based Similarity)[21]、調整性餘弦定理相似度(Adjusted Cosine Similarity)[22]、歐幾里德距離(Euclidean Distance)[23]等等。

皮爾森相關係數：$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$

餘弦定理相似度：$\mathrm{sim(X,Y)} = \frac{\vec{X} \cdot \vec{Y}}{\|X\| \cdot \|Y\|}$

調整性餘弦定理相似度：$\mathrm{sim(i,j)} = \frac{\sum_{c \epsilon I_{i,j}} (R_{i,c} - \overline{R_i})(R_{j,c} - \overline{R_j})}{\sqrt{\sum_{c \epsilon I_{i,j}} (R_{i,c} - \overline{R_i})^2} \sqrt{\sum_{c \epsilon I_{i,j}} (R_{j,c} - \overline{R_j})^2}}$

歐幾里德距離：$\mathrm{d(X,Y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

三、 產生推薦結果

藉由最近鄰集合，就能對目標使用者的興趣進行預測，以產生推薦結果。通過依據推薦目的之不同來進行不同形式的推薦，比較常見的推薦有 Top-N 推薦和關聯推薦。

---

[20] 皮爾森相關係數(Person Correlation Coefficient)：在統計學中，用於度量兩個變量 X 和 Y 之間的線性相關，其值介於-1 與 1 之間。

[21] 餘弦定理相似度(Cosine-based Similarity)：透過測量兩個向量的夾角的餘弦值來測量它們之間的相似性，查看兩向量之間的角度的餘弦值判斷是否指向相同的方向。餘弦相似度對任何維度的向量空間中都適用。

[22] 調整性餘弦定理相似度(Adjusted Cosine Similarity)：考慮到一般餘弦相似度未考慮到的尺度問題，以置中平減(Mean-centered)的方式對總體評分模式的偏差調整。例如為 1 到 5 分的評分等級，使用者 A 可能在 3 分以上就代表喜歡，但使用者 B 可能 4 分以上才表示喜歡，透過減掉使用者的平均分數來修正餘弦相似度的問題。

[23] 歐幾里德距離(Euclidean Distance)：在數學中，歐幾里德距離是歐幾里德空間(Euclidean Space)中兩點間直線距離。

Top-N 推薦是找出推薦給目標使用者的指定個數項目，為針對單一使用者產生，會對每個人產生不一樣的結果；關聯推薦則是對最近鄰使用者的記錄進行關聯規則(Association Rules)[24]探勘以產生推薦。

除了對使用者明確告知有興趣的項目之外，還可以考慮到使用者瀏覽項目的時段、次數及頻率等資料，藉此瞭解使用者的偏好，作為一種隱含回饋(Implicit Feedback)[25]，並可利用類神經網路(Artificial Neural Network，簡稱 ANN)[26]、群集分析(Cluster Analysis)[27]等方法，進行使用者資訊的分群。

[24] 關聯規則(Association Rules)：是一種在大型資料庫中發現變數之間的有趣性關係的方法，它的目的是利用一些有趣性的測量來識別資料庫中的強規則。例如沃爾瑪(Wal-Mart)的購物籃分析出的尿布與啤酒的連帶銷售關係。

[25] 隱含回饋(Implicit Feedback)：是一種對使用者行為的間接推導，因為使用者並沒有明確地提供回饋，而是通過對使用者行為意圖的推測，來作為運用的資訊。

[26] 類神經網路(Artificial Neural Network)：簡稱 ANN，在機器學習和認知科學領域，是一種模仿生物神經網絡(動物的中樞神經系統，特別是大腦)的結構和功能的數學模型或計算模型，用於對函數進行估計或近似。

[27] 群集分析(Cluster Analysis)：是對於統計數據分析的一門技術，在許多領域受到廣泛應用，包括機器學習、資料探勘、模式識別、圖像分析以及生物資訊。群集是把相似的對象通過靜態分類的方法分成不同的組別或者更多的子集(Subset)，這樣讓在同一個子集中的成員對象都有相似的一些屬性，常見的包括在坐標系中更加短的空間距離等。

2.2.2 基於項目之協同過濾 (Item-based Collaborative Filtering)

透過計算項目之間的相似度來代替使用者之間的相似度。

一、 收集使用者資訊

收集足以代表使用者興趣的資訊。為一個二維空間的矩陣，X 座標為項目，Y 座標為使用者，並在對應的空格內填入喜好程度的評分。

二、 針對項目的最近鄰搜尋

先計算已評價項目和待預測項目的相似度，並以相似度作為權重，加權各已評價項目的分數，得到待預測項目的預測值。隨後對這些組合進行相似度計算，常用的演算法同樣是有皮爾森相關係數(Person Correlation Coefficient)、餘弦定理相似度(Cosine-based Similarity)、調整性餘弦定理相似度(Adjusted Cosine Similarity)、歐幾里德距離(Euclidean Distance)等等。

三、 產生推薦結果

基於項目之協同過濾(Item-based Collaborative Filtering)在使用上因為不用考慮使用者間的差別，所以精確度上可能會比較差。但是卻有不需要使用者的歷史資料，或是進行使用者識別的好處。對於項目來講，它們之間的相似度要穩定很多，因此可以離線完成工作量最大的相似度計算步驟，從而降低了線上計算量，能有效提高推薦效率，特別是在使用者數目多於項目的情形下將尤其顯著。

2.2.3 最近鄰居法 (k-Nearest Neighbors Algorithm)

在模式識別領域中，最近鄰居法(k-Nearest Neighbors Algorithm，簡稱 k-NN，又稱 K-近鄰)[28]是一種用於分類和迴歸的無母數統計方法。在這兩種情況下，輸入包含特徵空間中的 k 個最接近的訓練樣本。

在 k-NN 分類中，輸出是一個分類族群。一個對象的分類是由其鄰居的「多數表決」確定的，k 個最近鄰居(k 為正整數，通常較小)中最常見的分類決定了賦予該對象的類別。若 k = 1，則該對象的類別直接由最近的一個節點賦予。

在 k-NN 迴歸中，輸出是該對象的屬性值。該值是其 k 個最近鄰居的值的平均值。

最近鄰居法採用向量空間模型來分類，概念為相同類別的案例，彼此的相似度高，而可以藉由計算與已知類別案例之相似度，來評估未知類別案例可能的分類。

k-NN 是一種基於實例的學習，或者是局部近似和將所有計算推遲到分類之後的惰性學習。k-近鄰算法是所有的機器學習演算法中最簡單的之一。

無論是分類還是迴歸，衡量鄰居的權重都非常有用，使較近鄰居的權重比較遠鄰居的權重大。例如，一種常見的加權方案是給每個鄰居權重賦值為 1/ d，其中 d 是到鄰居的距離。

鄰居都取自一組已經正確分類(在迴歸的情況下，指屬性值正確)的對象。雖然沒要求明確的訓練步驟，但這也可以當作是此演算法的一個訓練樣本集。

[28] 最近鄰居法(k-Nearest Neighbors Algorithm)：簡稱 k-NN，又稱 K-近鄰。是一種用於分類和迴歸的無母數統計方法。採用向量空間模型來做分類，概念為相同類別的案例，彼此之間的相似程度高，則可以藉由計算與已知類別案例之相似程度，來評估未知類別案例可能的分類。

2.2.4 協同過濾通建系統的缺點

協同過濾(Collaborative Filtering)在實務上不失為一個簡單有效的方式。但仍有以下缺點：

一、 冷啟動問題 (Cold-Start)

冷啟動問題是協同過濾推薦演算法中被廣泛關注的經典問題，因為冷啟動會顯著影響推薦系統的推薦成效，而長期影響傳統協同過濾推薦系統的發展。

例如對於電子商務推薦系統，每天都有大量未曾到訪的新使用者訪問系統，也都有相當數量的新項目被新增到系統中，假使推薦系統能夠向新顧客成功推薦青睞的商品，系統將能贏得使用者更多的信任，對商家來說因為提高顧客對系統的忠誠度而能增加客源，對使用者來說則是獲得高品質且個人化的服務；而倘若新商品能及時被推薦出去，可以有效提高產品銷售量，為商家贏得更大的經濟利益，也最終促進電子商務的發展。

目前針對冷啟動問題，學術上也提出了一些解決方法，如隨機推薦法、平均值法、眾數(Mode)法、資訊熵(Entropy)法、相似性度量(Similarity Measurement)改進法結合內容資訊的方法等，而不同的演算法也有各自的優缺點。

二、 稀疏性問題 (Sparsity)

實務上，大部分的使用者會主動給予評分的非常少，所以會造成矩陣的稀疏性，導致可以提供預測的資料相當少。也就是說，使用者-項目評分矩陣(Ratings Matrix)非常稀疏。

推薦系統通常都會利用相當龐大數量的項目來評分與推薦，但是使用者所接觸到的項目，在多數情況下只佔系統非常小的比例，因此導致資料矩陣過於稀疏，而推薦系統在篩選時，便會無法找到適合的項目推薦給使用者。

若要提高評分矩陣密度常見的主要方法，就是讓使用者對更多的產品給予評分，相對的這也會增加使用者的負擔。不過一旦讓兩者的比例差距十分懸殊時，系統就會無法做出合適的推薦。

三、 系統延伸性問題 (Scalability)

當資料量和使用者數量龐大時，相應的計算量和運算時間也會驚人的成長，因而當系統需要提供即時回饋就會產生困難。

像是最鄰近演算法的演算複雜度，是與使用者歷史交易資料或瀏覽資料的數量大小的平方成正比，當推薦系統之使用者與項目數目龐大或大量增加時，延伸性問題常會是一個相當嚴重的阻礙。

為了應對系統延伸性問題，如果事前能做好部分資料的運算，能有效協助系統於實際運作時降低整體運算量，而使用者行為等即時性的相關資料，則可以採取定期彙總的模式，如每日或每小時整合後於離線運算完，再提供新的基礎給系統使用。

2.2.5 網路爬蟲 (Web Crawler)

網路爬蟲(Web Crawler)是一種自動化瀏覽網路的程式，為擷取資料的方法之一，透過網路爬蟲可以自動將目的網頁資料擷取並儲存，將瀏覽的各方網頁資訊累積成為我方的資料，透過從多方網站進行彙整，產生豐富的資料來源，並使系統的資訊覆蓋面更加完整，而從中生成的龐大資料量

也能作為資料探勘的依據，並可以交付資料分析師將所獲得的大量資料做資料分析和評估。

另外，藉由網路爬蟲程式能有效率的協助處理重複性蒐集並整理資料的動作，使相關人員得以釋出更多的時間用於其他工作之上。

2.2.6 開放街圖 (OpenStreetMap)

OpenStreetMap(開放街圖，簡稱 OSM)是自由並開放原始碼，且覆蓋全球範圍的地圖資料庫，由英國人史蒂夫·克斯特(Steve Coast)於 2004 年發起，並採用類似維基百科(Wikipedia)[29]的協作編輯模式，以及開放的授權與格式。

OSM 開放街圖的地圖是由一般的使用者進行繪製，也因為資料來源是來自一般使用者的常態性編輯，所以內容能夠相當豐富且多元。資料的來源可能會根據手持 GPS 裝置、行動裝置、航照圖，以及其他的自由內容，甚至是單純倚仗使用者對本地的認識而得。雖然編輯者並非全是經過專業的地圖繪製訓練的人，但最終還是可以產生接近專業地理資訊水準的地圖。

因為 OSM 開放街圖的地圖圖資皆是以開放資料庫授權(Open Databse License，簡稱 ODbL)[30]方式授權，所以可以在日常生活、導航、學術、甚至商業應用。

---

[29] 維基百科(Wikipedia)：是一個建立宗旨為自由內容、公開編輯且多語言的網路百科全書協作計畫，透過 Wiki 技術使得所有人都可以簡單地使用網頁瀏覽器修改其中的內容（受保護的內容除外）。
網址：https://zh-tw.wikipedia.org。

[30] 開放資料庫授權(Open Databse License)：簡稱 ODbL，是 Open Knowledge Foundation 旗下專案 Open Data Commons 所編撰出來的授權條款，於 2009 年 6 月釋出，針對資料與資料庫進行授權規範，也提供相同方式分享的限定，以保障資料的散佈與擴散，Database Contents License 來輔助補充內容的授權規則。
網址：http://opendatacommons.org/licenses/odbl/。

2.3 研究程序[31] [32]

圖 2-1：本研究系統之研究流程圖

[31] k-平均(K-means)：k-平均演算法源於訊號處理中的一種向量量化方法，現在則更多地作為一種群集分析方法流行於資料探勘領域。k-平均群集的目的是：把 n 個點劃分到 k 個群集中，使得每個點都屬於離他最近的平均值(即群集中心)對應的群集，以之作為群集的標準。

[32] 接收者操作特徵曲線(Receiver Operating Characteristic Curve)：簡稱 ROC 曲線，是一種坐標圖式的分析工具，用於選擇最佳的信號偵測模型、捨棄次佳的模型，和在同一模型中設定最佳

2.4 資料及參考文獻蒐集

在本文中各專有名詞的定義，一部分是透過知名的線上協作網路百科全書維基百科(Wikipedia)獲得，另外也有部分的專有名詞定義是自國家教育研究院雙語詞彙、學術名詞暨辭書資訊網[33]取得。

而所使用的開放資料方面，則分別取自政府資料開放平臺[34]及臺北市政府資料開放平台[35]，獲取包括全國範圍的開放資料，還有臺北市及新北市範圍的開放資料。

閾值。在做決策時，ROC 分析能不受成本跟效益的影響，給出客觀中立的建議。

[33] 國家教育研究院雙語詞彙、學術名詞暨辭書資訊網：由國家教育研究院負責維運的詞彙查詢網站，分為三個組成部分雙語詞彙資料庫、學術名詞資訊網、辭書資訊網，原分屬行政院研考會、國立編譯館，此整合詞庫內容包括公告詞彙、機關建議詞彙、各類學術名詞譯名查詢、學術文化辭書及辭典等。
網址：http://terms.naer.edu.tw/。

[34] 政府資料開放平臺：政府為推動開放政府（Open Government）之資料公開政策，為促進民主治理，與促進創新研發，因此整合公開資料以供大眾利用。
網址：http://data.gov.tw/。

[35] 臺北市政府資料開放平台：臺北市政府為推動開放政府（Open Government）之資料公開政策，整合本府公開資料於單一入口網站以供大眾利用。
網址：http://data.taipei/。

第三章、資料運用情形

3.1 運用資料之萃取技術及分析方法

蒐集臺北地區開放資料，範圍包含旅遊景點座標、附近交通、開放時間、遊客人數、自行車租借站、公園、遊客休憩站、接駁車路線、停車場、天氣等資料，運用這些資料於本研究中。因本研究目的為推薦使用者旅遊地點，故在資料整理方面，須事先將所有景點的資料合併以待候用，如每個景點的交通資訊、地址、票價、周邊設施、附近民宿、周遭小吃等等，以方便未來查詢時，能直接帶出所有有關資訊。

在目前本研究使用的地理資訊中，除了 Google 地圖之外，另外還有使用到 OSM 開放街圖。在 Google 地圖中，特別使用到的功能為計算路線與路程，皆可以參考 Google 地圖提供的方案選擇理想的動線。另外在 OSM 開放街圖中可以看到標準的地圖，也有公車路線及腳踏車路線圖可供參考，甚至可以自行增加項目於地圖中。本研究將利用 Google 地圖來提供路線規劃，運用 OSM 開放街圖來準確定位景點及周邊設施。

有鑑於不同人有不同的感受，對於天氣大抵也可以這麼說明，有些人喜歡在晴天時出門晃晃，而有些則喜歡在陰天時遊玩，甚至有人喜歡撐著雨傘在雨中漫步，因此本研究為了符合使用者推薦的客製化，也利用到中央氣象局的天氣資料來進行天氣參考值。

因為在政府開放資料中，並不包含景點的評價資料，所以本研究將利用網路爬蟲來蒐集 TripAdvisor(貓途鷹)、玩全台灣旅遊網等包含評價資訊的網站。本研究通過匯集評價資訊，整理為外部評價，將作為推薦地點的參考。然後，利用本推薦系統之評價回饋作為內部評價。本研究將這兩項評價分開來儲存，一則可以比對滿意度是否一致，二則可以累積自有的資

料。

為了能有效運用推薦方中的 User-Based 及 Item-Based 法進行推薦，本研究將根據旅遊景點的特性進行標註，像是利用地區而非冗長的地址、景點性質如歷史背景、文化遺址、自然風景及交通方便性如大眾運輸工具轉車次數等等，能盡量考量任何面向的出遊體驗。面對推薦系統最初的冷啟動問題，本研究計劃當使用者進系統的第一步會是手機驗證，使用者可選擇是否輸入居住地區、職業、家庭成員人數等回饋，使本研究可以針對不同的族群推薦，像是根據職業來判斷此職業的人較常旅遊的地區為何。

另外，針對寒暑假及春假等放假日數較長的時段，本研究將分別給予不同學制的同學不同的推薦。對於大專以上同學會將行程時間安排較長，像是環島、溯溪、泛舟等等，對國、高中生會推薦一些可以當日來回即可的行程。鑒於本研究是利用手機驗證的方式進行，在面臨學生身分別上，本研究將利用 Facebook[36]未滿 18 歲無法註冊的方式進行配對，以確實找出真實分類。

本研究將分別探討既有資料呈現與推薦及未來數據蒐集進階推薦兩部分。

[36] Facebook：是一家線上社群網路服務網站，未有正式統一的中文譯名，台灣分公司命名為臉書。除了文字訊息之外，使用者可傳送圖片、影片、貼圖、聲音媒體訊息，和部分種類的其他檔案類型給其他使用者，以及透過整合的地圖功能分享使用者的所在位置。使用者必須註冊才能使用 Facebook，註冊後他們可以創建個人檔案、將其他使用者加為好友、傳遞訊息，並在其他使用者更新個人檔案時獲得自動通知。此外使用者也可以加入有相同興趣的群組，這些群組依據工作地點、學校或其他特性分類。使用者亦可將朋友分別加入不同的列表中管理。截至 2015 年 6 月底，每月至少瀏覽 Facebook 一次的註冊使用者達 14 億 9 千萬，約佔全球 30 億網友的一半，其中約 9680 多萬使用者每日登入，已成為世界上分布最廣的社群網站。
網址：https://zh-tw.facebook.com/。

3.1.1 既有資料推薦

首先就既有資料集而言，將運用臺北市政府之開放資料集搭配TripAdvisor(貓途鷹)、玩全台灣旅遊網等旅遊網站上的評價星等及評價內容，先做初步的景點分類及評價判斷。

接著需要對資料進行一連串的處理，首先對資料集做資料清理(Data Cleaning)[37]檢查、尋找並去除資料中的錯誤、雜訊、不一致的地方，然後進行資料整合(Data Integration)[38]以整合不同來源的資料，通過資料選擇(Data Selection)[39]從資料庫中挑選出與此次分析相關的資料，隨後進行資料變換(Data Transformation)[40]整理這些不同類型的資料，並解決開放資料集常見的資料顆粒度(Granularity)[41]大小不一致的問題。將各景點資料依據其交通資訊、觀光客統計、地理位置、適合族群、事宜時間等資料分別進行合併，以利後續推薦使用。

之後得以藉由資料探勘(Data Mining)[42]以選定的演算法來將前面步驟中整理完成的資料進行分析與處理，以從資料中獲取可能蘊含的規律，並以樣式評估(Patterns Evaluation)[43]來確定這些所得到的規律是否具備有效性，

---

[37] 資料清理(Data Cleaning)：資料探勘的第一個階段，去除資料中的錯誤、雜訊，和不一致的部分。

[38] 資料整合(Data Integration)：資料探勘的第二個階段，整合來自不同來源的資料。

[39] 資料選擇(Data Selection)：資料探勘的第三個階段，從資料庫中挑選出與此次分析相關的資料。

[40] 資料變換(Data Transformation)：資料探勘的第四個階段，將資料轉換成適當的格式彙整。

[41] 顆粒度(Granularity)：又稱資料粒度，是指資料的詳細程度。例如一個資料集中統計範圍為區鄉、鎮等級，另一個則為里或村，為合併兩個資料集，就需要改變其中一個資料集的顆粒度。

[42] 資料探勘(Data Mining)：此處指資料探勘的第五個階段，同時也是整套處理流程的核心所在，此方法由此得名；以特定演算法將彙總的資料進行分析處理，以獲取資料蘊含的規律。

[43] 樣式評估(Patterns Evaluation)：資料探勘的第六個階段，確定所得規律的有效性。

最後在知識展現(Knowledge Presentation)[44]的階段，以視覺化方式來呈現這些獲得的知識。

3.1.2 後續進階推薦

本研究於系統上蒐集使用者資訊，首先利用臺北公眾區免費無線上網(Taipei Free)[45]提供的臺北地區免費 Wi-Fi[46]服務的申請是綁定手機的特性，通過這樣的設計本研究得以運用手機來驗證使用者，累積同一使用者的個人資訊，藉此即可判斷該使用者的喜好及客製化旅遊地點。

同時也將透過選擇填寫的會員資料，調查使用者的職業及居住地區等訊息，作為後續推薦所需預先收集的資訊，得以針對不同類型使用者進行差異化的推薦。

---

[44] 知識展現(Knowledge Presentation)：資料探勘的第七個階段，也是最後一個步驟，以視覺化方式呈現知識。

[45] 臺北公眾區免費無線上網(Taipei Free)：台北市政府建置的免費無線上網服務，範圍包括台北市主要公共場所如政府機關、捷運車站、圖書館及醫院，及台北市主要幹道、住商區域及人口密集區。其室內外熱點的無線網路名稱(SSID)為 TPE-Free 以及 TPE-Free_CHT，而公車的 SSID 則是 TPE-Free Bus。並與中央行政機關室內公共區域提供免費無線上網(iTaiwan)與新北市政府無線上網(NewTaipei)建立使用者帳號雙向網路漫遊機制，可在全國 iTaiwan、新北市 NewTaipei 及 Taipei Free 的熱點，免費無線上網。

[46] Wi-Fi：是建立於 IEEE 802.11 標準的無線區域網路(Local Area Network，簡稱 LAN)技術。IEEE 802.11 定義了媒體存取控制和實體層(Physical Layer)，實體層定義在 2.4GHz 的 ISM 頻段(Industrial Scientific Medical Band)上的兩種無線調頻(Frequency Modulation，簡稱 FM)方式和一種紅外線傳輸(Infrared Data Association，簡稱 IrDA)的方式，總資料傳輸速率設計為 2Mbit/s。之後的 802.11a 定義在 5GHz ISM 頻段上的實體層，資料傳輸速率可達 54Mbit/s；而 802.11b 定義在 2.4GHz 的 ISM 頻段上，但資料傳輸速率高達 11Mbit/s 的實體層。因為 2.4GHz 的 ISM 頻段為世界上絕大多數國家通用，所以 802.11b 得到了最為廣泛的應用。

3.1.3 呈現方式

除了藉由常見的網頁及手機 APP 的方式呈現外，也試圖利用便利商店或是景區內的生活資訊站(Kiosk)，本研究也將分別探討如下：

一、 網頁介面

因應目前臺北地區的無線網路建置情況非常普及，所以設置了網頁版以提供使用者使用。網頁設計上，一開始首頁將會是作為熱門景點的介紹，當使用者註冊並登入後，才會根據使用者的歷史紀錄偏好，呈現差異化的項目。如果是新註冊的使用者，因為該使用者為第一次使用本系統，所以將面臨冷啟動問題，在此本研究的應對方式為，紀錄該使用者最初點選的分類項目至第三個點選的分類項目，以這前三個分類項目作為該使用者的優先喜好的旅遊分類。但如果當使用者於系統中累積足額的歷史紀錄，自下一次登入時，系統便能帶出針對該使用者的喜好旅遊分類，並據此推薦合適的景點、住宿、餐館等項目，提供使用者參考。

若使用者已經有想法或目標了，網頁也會提供選項，讓使用者選擇想去的地區、類別等選單式問答，或是其他條件輸入，譬如旅遊目的、出遊人數等進階問答，系統將利用這些問項給予使用者適合的地點推薦。舉例來說，若使用者想與家庭一起遊戲，那系統將針對他的目的，如推薦遊樂園或是公園等地區，讓他們可以增加家庭共同的回憶。若使用者想找一個海邊攝影，那麼系統將推薦他非中午時段的時間，以避免陽光過強導致拍攝出來的作品不佳。若使用者欲獨自一人體驗山林的寧靜，系統將會給予登山步道的推薦。若使用者不想到離家太遠的地方打發時間，系統將根據使用者提供的居住資訊，提供該地區的景點推薦。

若使用者對於推薦的地點體驗感到滿意或不愉快，可以在景點介紹後的評價中給予個人體驗評價回饋，以提供系統資訊用於下一次推薦，本研

究也可以因這些內容評價判斷是否因為季節影響了旅遊體驗，又或者使用者其實不喜歡該景點是推薦系統的誤判。

本系統將在隱私權的範圍內，本研究會簡單調查使用者的基本資料，像是職業、興趣、家庭人數、居住地區、年齡區間等，讓使用者提供一些資料，使本研究的系統更加完善。譬如本研究可以根據統計相同職業的人大多喜歡哪種旅遊景點分類，又或者本研究可以經由計算得知該職業的人最大移動距離為多少公里等其他資訊。

二、 行動應用程式 (Mobile Application)[47]

在手機應用程式 APP 的系統操作上與網頁版無異，在評價回饋部份本研究將設有雲端空間供使用者上傳照片，除了可以讓評價部分的呈現更加豐富外，也可以讓尚未決定到哪遊玩的使用者得以透過觀看實際拍攝的照片再行決定。

本研究將連結中央氣象局的天氣預測，讓使用者能夠知道接下來該地區的天氣將會是如何變化，並依此決定是否繼續行程或是更改行程。並且可以連結 Google 地圖的路線規劃的功能，讓使用者能方便的導航前往到下一個目的地。

[47] 行動應用程式(Mobile Application)：簡稱 APP，又稱手機應用程式、行動應用程式等，是指設計給智慧型手機、平板電腦和其他行動裝置上運行的應用程式。

除此之外，本系統也會與社群網站(Social Network Service，SNS)[48]相互結合，如：Facebook、Google+[49]、推特(Twitter)[50]等，提供使用者即時與好友分享當前體驗，打卡標註所在地點，或將照片上傳與朋友分享。

三、 生活資訊站 (Kiosk)[51]

在現今如此便利的社會中，台灣的便利商店分布極廣，幾乎可說是已經達到無所不在的地步，是故本研究打算利用各景區附近的便利商店，或是景區自有的 Kiosk 生活資訊站來提供服務。

這樣的情境，將更為符合現代人的需求，像是在夏天炎炎日曬下，不少人會選擇到便利商店休息、吹冷氣、購買冷飲，那何不在這相對空閒的時間中，通過本研究的服務規劃接下來的行程，尋找下一個景點目標。

---

[48] 社群網站(Social Network Service)：簡稱 SNS，是為一群擁有相同興趣與活動的人建立的線上社群。這類服務往往是基於網際網路，為使用者提供各種聯繫、交流的互動通路，如電子郵件、即時通訊服務等。多數社群網路會提供多種讓使用者互動起來的方式，可以聊天、寄信、影音、檔案分享、部落格、新聞群組等。

[49] Google+：稱作 Google Plus，簡稱 G+，是 Google 公司推出的社群網站與身分服務；除社群網站身分外，Google 也將 Google+視為其旗下眾多服務之間社交層面的補強，與傳統社群網站僅能登入單一網站的概念不同。於 2013 年 1 月超越 Twitter，成為世界上第二大的社群網站。

[50] 推特(Twitter)：是一個社群網站和微網誌服務，它可以讓使用者更新不超過 140 個字元的訊息，這些訊息也被稱作推文(Tweet)。截至 2012 年 3 月，Twitter 共有 1.4 億活躍使用者，這些使用者每天會發表約 3.4 億條推文。同時，Twitter 每天還會處理約 16 億的網路搜尋請求。非註冊使用者可以閱讀公開的推文，而註冊使用者則可以通過 Twitter 網站、簡訊或者各種各樣的應用軟體來發布訊息。Twitter 是網際網路上瀏覽量最大的十個網站之一。

[51] 生活資訊站(Kiosk)：主要是供使用者經由觸控面板的操作，自助式的使用服務內容，隨著科技的演進，變成具有共同特性而產生的新綜合體。以公用資訊站的特質，變成跨應用平台的通用設備，依業者所提供服務的不同而開發對應的軟體加以支援。

3.1.4 功能性

當系統剛開始上架，首先面臨的將會是在推薦系統中常見的冷起動問題，而且冷啟動的問題在針對每個目標使用者的初次使用時，都需要進行處理，以做出合適的推薦結果，因此面對這個重要問題本研究有如下幾個方案：

一、利用使用者點擊景點分類大項的前三個，來判斷該使用者較有興趣的景點分類為何。

二、以詢問方式進行，本研究將設計問題詢問使用者的職業及居住地區，這個方法既可以建立使用者的資訊，也可以由居住地附近的景點開始往外推薦。

三、讓使用者選擇，像是有多少人一起出遊、是否自行開車、旅遊目的等提問，利用這些資訊來判斷哪些景點適合，並進行推薦。

除了持續改進服務品質，本研究亦將增進提供的服務，目前優先規劃將增加的系統功能，如下所示：

一、 推播技術 (Push Technology)[52]

本研究通過將系統結合商業經常運用到的推播技術，例如系統可以在用餐時間快到的時候，推播景區附近合作餐廳的優惠卷，讓使用者能夠得到優惠、商家也獲得客源。除此之外，也可以結合各項官方或民間舉辦的

[52] 推播技術(Push Technology)：又稱伺服器推播(Server Push)，是指在網際網路上，由中心伺服器啟動一個通信要求的一種通信方法，相對的由顧客端開始一個通信要求的方法稱為拉播技術(Pull Technology)。

活動資訊，讓系統得以形成一個平台，使用者只要透過這個平台就能輕易得知附近的活動消息，或查詢特定地點是否有什麼活動。

這方面的應用，可以運用在像是最近流行的路跑活動，有些是只開放給特定族群報名，那這些經過特殊設計的活動，系統就可以根據使用者個人資料，將之推播給符合要求的特定族群。舉例來說，像是前陣子的九校聯合路跑，其主旨為「提升校園運動風氣，強化健康體魄，增進各校學生及教職員工之間的交流」，所以是針對九所北部大學進行活動，而當本研究的系統偵測到有這類型的活動，將會抓取「參加對象」的條件，並推播該項活動給符合的族群，在這個例子中即是推播給這九所學校的教職員生參考。而就這項活動而言，亦有提供社會組，那系統也將依據居住地區及興趣，來提供這些資訊給住在北部地區，並且興趣是路跑的族群參考。

另外像是政府單位的活動也可以藉由系統的推播功能，讓大眾瞭解政府近期舉辦的活動，一則可以當作假日去處的參考，二則可以了解政府的作為，增進民眾對相關活動的參與程度。

而針對商家推播的部分，本研究亦可以限定使用時間，譬如中午推出的午餐特餐，那使用時機就會在當天的 11 點至 13 點間，可以避免掉商家於錯誤時機發放優惠券可能造成的損失。

二、 提升旅遊品質

前陣子在各大媒體版面，皆有提到國人出遊地點出現許多外來遊客，不僅僅造成排隊時間顯著拉長，也某種程度上降低了旅遊的品質和遊客的遊興。為了解決像這樣的問題，本研究將蒐集各景點的容納人數上限、遊客人數、國內旅客人數、國外旅客人數等資訊，經過對相關資料分析後，本研究可以得知哪個季節該景點的遊客較多，並可以精細到特定月份。

而有了這樣的資訊，本研究就可以推薦怕人潮擁擠的使用者到該時段較冷門的地方，或是根據分析結果建議在哪些月份和時段相對沒那麼多人潮，供使用者進行旅程的合理規劃。

三、 回饋機制

畢竟本研究蒐集的資料是有限的，有些隱藏版的景點，及熱門餐廳往往只有當地人或熟客才會知道，所以本研究也將在系統上設置使用者回饋的功能，讓一些不被大眾知道的私房景點，得以通過熱心的使用者而公諸於世。

除了需要使用者回饋景點或餐廳的地點之外，基本的介紹、照片等等能更讓人信服的背景資料皆能上傳，其他使用者也可以針對這些資訊修改或評價。

3.2 運用資料之限制及改進建議

在開放資料獲取的過程中，有些資料並不如預期想像般能直接使用，超過半數以上的資料都需要事先處理，無論是資料合併或是資料具無意義的資訊，這些資料都會額外增加處理時間，進而影響效率及使用意願。

以下將列舉一些：

1. 2014 運動中心 csv：資料集中包含名稱、郵遞區號、地址、電話及網址，但以資料分析來說，我們更需要的資料，像是營業時間、座標，及面積等皆未包含。
2. 1050527 台北市休閒農場基本資料：資料集中包含農場名稱、地址、電話及農場主要特色簡介，因農場所包含的項目眾多，可能有餵食體驗等活動，但在資料集中無從得知相關資訊，也並無法看出這個農場是否有養殖禽畜類等資訊，這都將會影響到分析的結果。

3. 各個不同的資料集所用的欄位名稱不一致，是很嚴重的困擾，在當需要將不同資料集串接時，將沒有可以依據的欄位，增加處理上所需花費的心力。

因為有以上等等問題，導致在資料處理上，常需要花更多時間去做整理，浪費了許多時間成本。也因為許多資料集中，並不包含座標的資訊，所以本研究額外運用了 OSM 開放街圖，來完成地圖位置的搜尋與定位。

第四章、研究結論與建議

政府開放資料期許讓所有人都可自行取用原是美意，可以讓更多有想法、有能力的人都能有機會為社會盡一份心力，但倘若政府在公開資訊的追求，只是為了能在相關領域的全球排名上更好看，而為評比的各項指標要求所迷惑的話，那資料的品質和資料的充分程度都將會常有不如欲取得相關資料者預期的狀況，而在使用上更會令使用者受到極大的限制。

在英國開放知識基金會(Open Knowledge Foundation，簡稱 OKFN)[53]的開放資料 Open Data 2015 國際評比中，台灣於全球一百廿二個受調查國家中名列第一，而日本是第三十一名。

但當我們在下載政府開放資料時，時常會發現到部分台灣開放資料，在分列的細項及詮釋資料(Metadata)[54]呈現上，有時並不如日本政府在相同品項開放資料來的詳盡和清楚，我們認為這方面也許還可以參酌日本政府提供的政府開放資料[55]來進行改進。

而在部分資料的呈現上，更會有相同資料項目的前一年格式，卻與後一年的格式不一致的情況，這樣無疑是增加使用者後續利用及整合上的無端困擾，大幅增加使用者前期準備的工作量及時間。

---

[53] 英國開放知識基金會(Open Knowledge Foundation)：簡稱 OKFN，是一個於 2004 年在英國劍橋成立的非營利性組織，長期致力於在數位時代推廣各類形式的開放知識。近年來已經活躍於全球 40 多個國家和地區，並主要著重開放資料和開放政府的推廣和支持。在開放資料的推廣和教育，OKFN 和 P2PU 合作建立的資料學院(School of Data) 以及其組織社群編寫的開放資料手冊(Open Data Handbook) 和資料新聞學手冊(Data Journalism Handbook)，都得到了廣泛的關注。

[54] 詮釋資料(Metadata)：為描述資料的資料(Data about data)，主要是描述資料屬性(Property)的資訊，用來支援資料呈現，如指示儲存位置、歷史資料、資源尋找、文件記錄等功能。

[55] Data.go.jp：是日本政府為推動開放政府資料倡議，由政府機關儘量開放機械可判讀格式，允許被以營利或其他目的進行再利用之公共資料放置的平台。
網址：http://www.data.go.jp/。

在旅遊推薦系統部分，所面臨的最大問題不外乎是資料整理的過程，除了對於資料的種種前處理(Preprocessing)[56]外，當尚無特別鎖定的查詢目標，而需在龐大數目的開放資料中單純找出對使用情境有用的資訊時，就我們的經驗而言，坦白說也是蠻費時耗力的，所以在開放資料平台的瀏覽上是不是能協助使用者更輕鬆的尋找或漫無目的的閒逛，可能也會是政府在開放資料平台建置上，一個還能持續努力的重要課題。

我們希望未來能將系統的資料範圍擴及全台灣，這也能使系統在進行推薦時，不會被侷限在特定地區。就目前而言，本系統的推薦範圍暫時只涵蓋臺北市及新北市的雙北地區，而且現階段尤其對未能加入基隆市形成大臺北都會區的旅遊推薦系統感到遺憾，因為北北基地區的生活機能無疑是彼此緊密連結在一起，如果單單只對雙北市地區進行推薦難免會有所缺憾，但這方面主要還是因為我們並未取得足夠的基隆市資料，還望未來有機會能夠對基隆市提供服務。

當國人越來越會享受生活的當下，旅遊的品質特別會被大眾認真考慮與思量，而旅遊推薦的成效也將更為顯著，憑此能對更加廣大的民眾產生助益。Map InterLinK 未來還將持續努力提供更佳的服務，並盡力觸及不同地區的使用者，甚至是擴展服務到不同國家的民眾，希望能夠讓海內外的朋友都能認識到台灣的美好。

[56] 前處理(Preprocessing)：處理資料所隱藏之雜訊、不一致、遺漏、重覆之資料，另外還包括增刪欄位、整合資料等，例如刪除後續處理中不必要的資料能有效降低運算量。

第五章、參考文獻

[1] A. Jaynal, D. Kishor Kumar , Data Manipulation with R, 2/e(Paperback) (2015) , Packt Publishing

[2] Charu C. Aggarwal , Recommender System The Textbook Springer

[3] Fischetti T. , Data Analysis with R Paperback (2016) , Packt Publishing

[4] James G. , Witten D. , Hastie T. , Tibshirani R ,(2015) An Introduction to Statistical Learning with Applications in R , New York, USA: Springer

[5] Kotaro Sakamoto , Hyogo Matsui , Eisuke Matsunaga , Takahisa Jin , Hideyuki Shibuki , Tatsunori Mori , Madoka Ishioroshi , Noriko Kando , Forst: Question Answering System Using Basic Element at NTCIR-11 QA-Lab Task (2014)

[6] L. Dominique, C. Baptiste, N. Sophie, S. Patrick , English run of Synapse Développement at Entrance Exams 2014

[7] Lantz B , Machine Learning with R, 2/e (Paperback) (2015) , Packt Publishing

[8] M. Pradeepta , R Data Mining Blueprints (2016) , Packt Publishing

[9] Manning, C. D., Raghavan, P., & Schütze, H. (2012). 資訊檢索導論 [Introduction to information retrieval] (王斌 Trans.). 臺北市: 五南圖書.

[10] Min-Yuh Day, Cheng-Chia Tsai, Wei-Chun Chuang, Jin-Kun Lin, Hsiu-Yuan Chang, Tzu-Jui Sun, Yuan-Jie Tsai, Yi-Heng Chiang, Cheng-Zhi Han, Wei-Ming Chen, Yun-Da Tsai, Yi-Jing Lin, Yue-Da Lin, Yu-Ming Guo, Ching-Yuan Chien, and Cheng-Hung Lee (2016), "IMTKU Question Answering System for World History Exams at NTCIR-12 QA Lab2", The 12th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-12), Tokyo, Japan, June 7-10, 2016.

[11] P. Erik Rodriguez , Unsupervised Learning with R (2015) , Packt Publishing

[12] Ryan Mitchell. (2016). 網站擷取：使用 python [Web Scraping with Python: collecting data from the modern web] (Studio Tib. Trans.). 臺北市: 碁峰資訊.

[13] Sebastian Raschka. (2016). Python 機器學習 [Python machine learning] (劉立民, 吳建華 Trans.). 新北市: 博碩文化.

[14] Suresh K. Gorakala, Michele Usuelli , Building a Recommendation System with R (2015) , Packt Publishing

[15] Usuelli M. R Machine Learning Essentials (Paperback) (2014) , Packt Publishing

[16] W. Simon , Big Data Analytics with R (2016) , Packt Publishing

[17] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Burlington, MA: Morgan Kaufmann Publishers.

[18] 方匡南 , 朱建平 & 姜葉飛 . (2015). R 語言資料分析活用範例詳解 (H&C Trans.) , 臺北市: 碁峰資訊

[19] 李仁鐘. (2015) . 應用 R 語言於資料分析-從機器學習、資料探勘到巨量資料 . 臺北市: 松崗出版

[20] 李卓翰. (2006). 資料倉儲理論與實務 (2nd ed.). 臺北市: 學貫行銷.

[21] 林東清. (2009). 知識管理 (3rd ed.). 臺北市: 智勝文化.

[22] 陳景祥. (2010) . R 軟體 : 應用統計方法. 臺北市: 東華書局

[23] 黃文, & 王正林. (2016). 利用 R 語言打通大數據的經脈. 臺北市: 佳魁資訊

[24] 劉光, 曾敬文, & 曾慶豐. (2016). Web GIS 原理與應用開發. 北京市: 清華大學.

[25] 簡禎富, & 許嘉裕. (2014). 資料挖礦與大數據分析. 新北市: 前程文化.

作者簡介

Map InterLinK，簡稱 MILK，為王斯霈、蔣宜衡、陳孟勛，和提供技術指導的勤業眾信聯合會計師事務所所組成的團隊，關注資料分析、輿情分析、資料探勘、推薦系統等領域，目前負責建置與維護名為 MILK 的旅遊推薦系統。

王斯霈(Szupei Wang)，淡江大學商管學院資訊管理所碩士生，曾就讀同校商管學院統計學系，研究興趣包括資料探勘、機器學習、資料分析等領域，擅長運用 R 語言、SAS 等統計程式語言作統計分析。

蔣宜衡(Yi-Heng Chiang)，淡江大學商管學院資訊管理所碩士生，曾就讀同校資訊管理學系，研究興趣包括文字探勘、機器學習、推薦系統等領域，曾赴日參加日本國立情報學研究所舉辦之 NTCIR-12 研討會 QA Lab-2 課題，以 IMTKU 團隊發表。

陳孟勛(Meng-Hsun Chen)，淡江大學商管學院資訊管理所碩士生，曾就讀同校工學院資訊工程學系，研究興趣包括無線網路、行動計算、影像處理、虛擬實境等領域，擅長程式開發，涵蓋 C、Java、ASP. NET (C#)等程式語言。

勤業眾信聯合會計師事務所(Deloitte & Touche)，係指德勤有限公司（Deloitte Touche Tohmatsu Limited）之會員，其成員包括勤業眾信聯合會計師事務所、勤業眾信管理顧問股份有限公司、勤業眾信財稅顧問股份有限公司、勤業眾信風險管理諮詢股份有限公司、德勤財務顧問股份有限公司、德勤不動產顧問股份有限公司、及德

勤商務法律事務所。勤業眾信以卓越的客戶服務、優秀的人才、完善的訓練及嚴謹的查核於業界享有良好聲譽。透過德勤有限公司之資源，提供客戶全球化的服務，包括赴海外上市或籌集資金、海外企業回台掛牌、中國大陸及東協投資等。