國立台北大學資訊工程學系專題報告

社群網路資料分析及活動賓客推薦

Social Network Data Analysis and Guest Recommendation for Activities

專題組員：李力學、劉炳佑

專題編號：NTPUCSIE-103-002

執行期間：103 年 07 月　至　104 年 05 月

一、前言

近年來社群網路蓬勃發展拉近了人與人之間的距離，即使處於不同國家，不同時間區，人與人仍可互相交流自己的生活經驗，有時也使得人們可以聯繫上許多失聯已久的朋友，成為人們生活中一個重要的連結工具。而除了心情連結、拓展生活圈的功能外，基於通訊上的便利性與資料變更的即時性，社群網路也經常取代傳統的聯絡方式，被使用來籌劃和舉辦各種活動。

舉辦活動是個複雜繁瑣的工程，除了必須設定時間、地點、內容外，我們特別關注「要邀請哪些賓客」這項問題。對於一個活動而言，究竟適合什麼樣特質的賓客參加呢？是和主辦人關係密切嗎？還是個人經歷應該要跟活動主旨相呼應呢？此外我們藉由過去的經驗可以得知，同一個主辦人舉辦不同活動的時候邀請的成員也各不相同，這其中又有什麼奧秘呢？既往的的推薦研究大多是應用在處理書籍、音樂、朋友等，然而至今未曾有人考慮對於一任意活動主題，是否可以利用分析活動特性來從候補成員中推薦出適合這個活動的賓客。在本次專題中，我們想以資料探勘的角度探討這個問題，並試圖利用全新的方法解決。

二、研究方法簡介

由於我們的作法必須先蒐集資料建立模型，經過討論之後決定選擇 Facebook 作為我們的資料來源以及研究對象。

Facebook 是當今擁有超過六十億註冊人口的社群網站，資料相當豐富。我們蒐集數個在 Facebook 平台上舉辦過的活動作為訓練資料，經過前處理過後，利用我們的演算法特定出數個特別的、可代表該活動特色的關鍵字詞並將資料分群，然後依此建立群資料庫。之後若需舉辦新活動，則將其設為測試資料經過比對後加入某一群，再利用該群的群資料庫分析活動特性與候補成員作比對，最後就可輸出推薦的賓客名單。

關鍵字：資料探勘、活動舉辦、推薦賓客

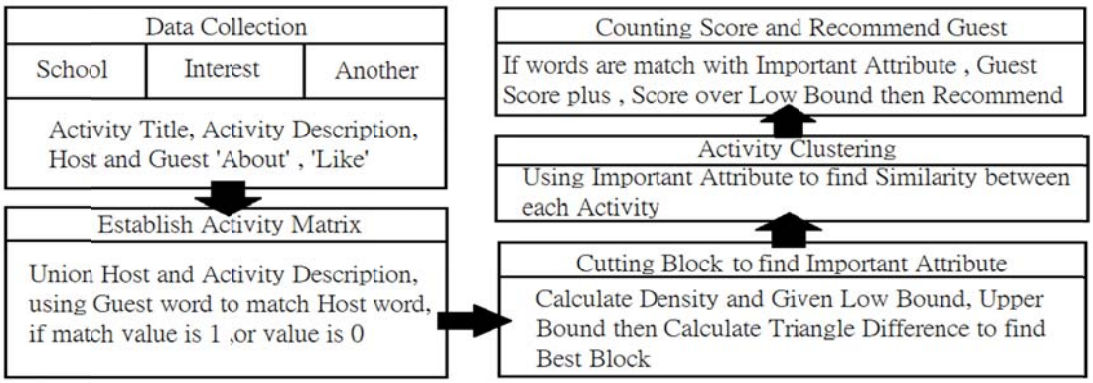

圖 1：系統運作流程

三、系統架構與演算法述論

1. 資料蒐集

根據過往經驗，我們認為所有活動大致可分為以下三種：

(i) 同儕、同事類活動

(ii) 興趣主題型活動

(iii) 其他

對各類別我們分別蒐集三個活動，一共九筆資料；各類別其中兩筆作為訓練資料建立模型，一筆則作為測試資料檢測成果。

雖然資料有著圖片、動畫、聲音等許許多多的格式，基於資料蒐集以及處理上的方便性，我們選擇蒐集文字資料並建立為文件檔作為研究主題。對於每一個活動的主旨取向，我們蒐集該活動的描述做為參考；對於該活動中每一個與會人員，包含賓客和主辦人，我們蒐集他們的「關於」和「說讚的內容」這兩個欄位。這是因為「關於」會列出既往的學歷與工作資歷，使得之後便可以藉此判斷是否有同儕同事類型的活動的傾向；「說讚的內容」則是基於人不會對不切身或不感興趣的內容表達關注的行為模式，反向推論此區的資料蒐集起來可當作個人的興趣傾向的表徵。

2. 建立活動矩陣

資料蒐集完畢後，我們先將所有文件檔送入中研院斷詞系統作詞彙切割，然後再施以前處理使之方便使用。

在建立活動矩陣的時候，我們對所有活動有兩個共通的假設：

(i) 將所有與會者和活動的關係作比對，主辦人與活動的關係程度會最大。

(ii) 將每個人和其餘所有與會者的關係作比對，主辦人與所有人之間的關係程度也會是最大。

基於這兩個假設，我們利用主辦人作為賓客和活動之間的連結，對於前處理好的資料我們先把主辦人和活動描述做比對，取出順位高的詞會當候補關鍵詞；再將候補關鍵詞和所有賓客的文件做比對，若有共通則為 1 否則為 0，建立活動的布林矩陣。

Guest \ Attribute	att1	att2	att3	att4	att5
G1	1	1	0	0	0
G2	1	1	0	0	1
G3	0	1	0	0	0
G4	0	1	0	1	1
G5	1	1	1	0	0
G6	0	0	0	1	0
G7	1	0	0	0	1

圖 2：原始布林矩陣

布林矩陣建立完成後我們對其做行列排序以將 1 集中，首先計算每個詞出現的次數做為行的重要性依據，並依此左高右低做行排序；對於每一列，我們發現可以將其視為一個二進位數值，且經排序的詞跟次方數的意義也相當近似，因此利用 Radix Sort 的方式上高下低做列排序，如此即可將 1 集中堆積在左上角，以利接下來的運算。

Guest \ Attribute	att2	att1	att5	att4	att3
G1	1	1	0	0	0
G2	1	1	1	0	0
G3	1	0	0	0	0
G4	1	0	1	1	0
G5	1	1	0	0	1
G6	0	0	0	1	0
G7	0	1	1	0	0

圖 3：行排序完成後

Guest \ Attribute	att2	att1	att5	att4	att3
G7	1	1	1	0	0
G5	1	1	0	0	1
G1	1	1	0	0	0
G4	1	0	1	1	0
G3	1	0	0	0	0
G2	0	1	1	0	0
G6	0	0	0	1	0

圖 4：列排序完成後

3. 切割矩形找出關鍵詞

矩陣排序完成後，藉由排序好的布林矩陣計算(0, 0)到(I, j)的密度，再減去其右方、下方的鄰近值後將兩者加總計算該點的三角差。藉由三角差，我們可以得知哪個點是位於資料密度的邊緣，因此將三角差最大的位置作為切割邊緣點切割出一矩形，此矩形內部之賓客即為重要賓客，關鍵候補詞即為活動關鍵詞。

Guest \ Attribute	att2	att1	att5	att4	att3
G7	1	1	1	0.75	0.6
G5	1	1	0.833	0.625	0.6
G1	1	1	0.778	0.583	0.533
G4	1	0.875	0.75	0.625	0.55
G3	1	0.8	0.667	0.55	0.48
G2	0.833	0.75	0.667	0.542	0.467
G6	0.714	0.643	0.571	0.5	0.429

圖 5：密度表

Guest \ Attribute	att2	att1	att5	att4	att3
G7	0	0	0.417	0.275	
G5	0	0.167	0.263	0.067	
G1	0	0.347	0.223	0.008	
G4	0.125	0.2	0.208	0.15	
G3	0.367	0.183	0.117	0.078	
G2	0.202	0.19	0.221	0.117	
G6					

圖 6：三角差表

但是為了確保代表性，我們經過實驗後還加入了

(i) 必須保證矩形內部含有總量一定比例以上的人數。

(ii) 必須保證矩形內部有一定比例以上的詞彙量。

(iii) 該邊緣點本身的密度須高於 0.85 的密度門檻值。

這三個條件以避免切割矩陣內部的資料的質量不足以代表該活動的特性。

4. 活動分群

訓練資料處理完畢後，我們將切割出的各活動關鍵詞做交叉比對且對其做分群，同時也將群內各活動關

鍵詞蒐集起來，建立出可代表該群的某些特色的群關鍵詞文件檔。

5. 積分計算及賓客推薦

此後每當舉行新活動，我們就可要求主辦人提供活動描述、本人同前述兩欄位的個人資料以分析活動特性，以及提供好友清單做為推薦賓客的候補名單。

主辦人提供的資料經過同樣的前處理後先依照相似度分入某一預先建立的群，利用群關鍵詞找出此新活動可能的活動關鍵詞，再比對活動關鍵詞和候補賓客文件建立布林矩陣，統計活動矩陣中各人員的 1 值作為該員之分數並上高下低排序出順位後，最終即可推薦前 N 個做為系統對該新活動推薦的賓客人選。

四、實驗結果

Testdata \ Evaluate	Accuracy	Precision	Recall
Testdata1	0.404	0.425	0.772
Testdata2	0.623	0.625	0.423
Testdata3	0.43	0.35	0.56

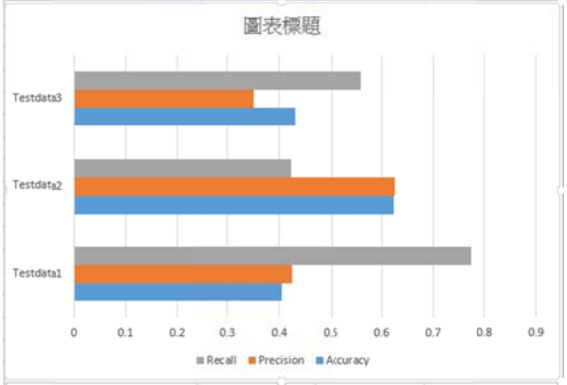

圖 7：效能分析實驗結果

為了分析系統效能，我們在各群當中加入一筆過去活動作為測試資料，並且同時額外抓取主辦人的朋友欄位，朋友區在抓取資料時有預先確保位於朋友區的成員盡可能避免和出席賓客交集以避免資料混雜；然後我們確認出席的賓客以及主辦人朋友群兩區集合起來的資料群體做賓客推薦的準確度分析，最終效能分析結果則如圖 7 所示。

抓取資料時，測試資料 1 和其中一個同樣位於同儕、同事類的訓練資料都是抓取台北大學相關，因此在群資料庫的輔助下該測試資料中所有有關台北大學的人員都被成功抓取使得 Recall 較高。而在測試資料 2 同群的訓練資料中恰巧有一和其關聯性頗高的活動，所以使得它們在 Accuracy 欄位有較好的表現。因此我們可推論當訓練資料數量擴充到某一程度，讓新進的測試資料在分群後都能夠有同性質的活動可做比對，則效能就更有可能會提升得更好。

五、未來展望

我們的研究是個新穎的想法，目前還沒有以推薦活動賓客作為主題的已發表研究。然而過程中當然也有覺得還不夠盡善盡美的地方，像是 Facebook 的自由書寫格式使得資料的統一性及可靠性都有相當程度的問題；資料蒐集後，若斷詞系統的斷點不夠精準，也會連帶影響我們的準確度；同時過濾詞彙集仍未臻完善，在未來希望能繼續改善加強這些部分。

此外即使是相同一群內，活動間主題的差異也會使得用到的詞彙不盡然相同，而在推薦的部分我們是將詞做配對，因此也會衍伸出些成功配對到的字詞不夠多的

問題。雖然此問題在資料筆數成長後因為群資料庫的成長將能獲得相當的解決，然而本次由於研究時間及人力的不足，僅能藉由實驗結果證明我們的理論可行卻無法將效能做最大化，相當可惜。希望未來能夠有人繼續接手改良我們的研究成果，將功能與效能更進一步的擴充。

六、參考文獻

無

七、資料來源

https://www.facebook.com