實作關聯規則演算法於台北市交通事故明細之分析

一、研究動機

大數據常見的運用之一，便是從顧客的消費記錄中，試圖辨別各項因素之間的關聯，例如：不同年齡層、居住於不同地區的顧客可能偏好哪些商品，或是喜好某種蛋糕的顧客必定會搭配某種咖啡等，無法直接由主觀推論而得知的 Frequent Patterns。然而，Frequent Pattern Mining 的應用並不僅限於商業用途上，在其他社會研究或是政府政策上也能發揮其功能，發掘隱藏於大量資料集中，個體與個體或事件間的因果關係。因此我們欲從臺北市政府的公開資料庫中，取得近年度 (2016) 的 A1 及 A2 類交通事故(指有人員傷亡)的事件明細，希望能從中得知哪些事件與情境組合容易造成車禍，或是造成重大傷亡，並進行預先的防範。

二、資料來源與資料集描述

<u>資料來源</u>

臺北市政府資料開放平臺 Data.Tapei (臺北市政府警察局交通警察大隊) [1]

<u>數據說明</u>

我們選以民國 105 年為分析資料集，共 50949 筆資料。原資料主要欄位如下，發生時間 (年、月、日、時、分)、地點、死傷人數、車種、性別、年齡、天候、速限、道路型態、事故位置。其中，我們調整了明細上的屬性個數與表現形式，如：

1. 將事件代碼轉以文字敘述以利觀察；
2. 將時間資訊區分成「平日/假日」、「凌晨/早上/中午/下午/晚上」；
3. 將肇事者的年齡做離散化，以 20、40、60 歲切成 4 個區間或是不明

發生年	發生月	發生日	發生時	發生分	處理別	區序	肇事地點	死亡人數	受傷人數	當事人序	車種	性別	年齡	受傷程度	天候	速限	道路型態	事故位置
106	1	3	8	58	2	01大同區	大同區大龍街	0	1	1	F01	2	70	2	8	50	14	9
106	1	3	8	58	2	01大同區	大同區大龍街	0	1	2	C03	1	38	3	8	50	14	9
106	1	4	9	14	2	01大同區	大同區環河北	0	1	1	B03	1	38	3	8	50	14	9
106	1	4	9	14	2	01大同區	大同區環河北	0	1	2	C04	2	56	2	8	50	14	9
106	1	9	15	36	2	01大同區	大同區鄭州路	0	1	1	C03	1	21	2	8	40	4	2
106	1	9	15	36	2	01大同區	大同區鄭州路	0	1	2	C03	2	21	3	8	40	4	2
106	1	14	13	21	2	01大同區	大同區南京西	0	2	1	C03	1	19	2	7	50	4	1
106	1	14	13	21	2	01大同區	大同區南京西	0	2	2	C03	1	21	2	7	50	4	1
106	1	14	17	30	2	01大同區	大同區承德路	0	1	1	C03	1	55	2	7	50	14	9
106	1	14	17	30	2	01大同區	大同區承德路	0	1	2	B03	2	50	3	7	50	14	9

▲ 原始資料集

星期	平日/假日	時間	區序	車種	性別	年齡	天候	道路型態	事故位置
星期二	平日	早上	01大同區	腳踏自行車	女	大於60歲	晴	直路	一般車道(未
星期二	平日	早上	01大同區	普通重型機	男	20-40歲	晴	直路	一般車道(未
星期三	平日	早上	01大同區	自用小客車	男	20-40歲	晴	直路	一般車道(未
星期三	平日	早上	01大同區	普通輕型機	女	40-60歲	晴	直路	一般車道(未
星期一	平日	下午	01大同區	普通重型機	男	20-40歲	晴	四岔路	交叉口附近
星期一	平日	下午	01大同區	普通重型機	女	20-40歲	晴	四岔路	交叉口附近
星期六	假日	中午	01大同區	普通重型機	男	小於20歲	陰	四岔路	交叉路口內
星期六	假日	中午	01大同區	普通重型機	男	20-40歲	陰	四岔路	交叉路口內
星期六	假日	下午	01大同區	普通重型機	男	40-60歲	陰	直路	一般車道(未
星期六	假日	下午	01大同區	自用小客車	女	40-60歲	陰	直路	一般車道(未

▲ 前處理完的資料集

三、實驗步驟

1. Relim algorithm

本次作業使用 Relim algorithm 找出頻繁出現的組合，該演算法是藉由 recursive elimination 的方式實現 [2]。與常見的演算法相比，像是 Apriori 需要不斷產生可能的組合，再搜尋資料庫判斷是否為 frequent；FP-growth 透過建立 FP tree 一種樹狀結構的方式找出所有 frequent pattern，僅需搜尋資料庫兩次，不需要產生候選組合(candidate itemset)，因此速度比 Apriori 快，但缺點是建樹時相當佔空間。而 Relim 演算法與 FP-growth 想法類似，差別在於以鏈結串列 (Linked List) 的結構儲存資料，而在找 frequent pattern 時的原則都一樣：當 A 不 frequent 時，AB 一定也不 frequent。相較於 FP-growth，Relim 結構簡單，空間利用率高且易於實現，也比 Apriori 來得有效率，因此我們採用 Relim 演算法。

2. 結果分析

我們多次比較了在設定不同 support 之下，Frequent Patterns (FP) 以及找出來的 Association Rules (AR) 的個數跟 Pattern 的 itemset 是否具有足夠合理性與有用性。藉由觀察來決定 support 和 confidence 的值。

Support	1000 (≈ 2%)	2000	3000	4000	5000 (≈ 10%)
# of FPs	3684	1257	629	364	249
# of ARs (confidence=0.5)	7523	2426	1128	629	414

經觀察後，我們認為在 Support 於 2% 至 10% 之間的 patterns 較具有可解釋性與有用性；以及 Confidence 在 50% 之下的表現較佳，因此我們進而藉由人工的方法，挑選出有用性較高的部分 frequent pattern 與 association rules，並展示於下頁表中。

Itemset of FP	Support
{'假日', '自用小客車', '40-60歲'}	1029
{'20-40歲', '女', '雨'}	1139
{'20-40歲', '凌晨'}	1200
{'交叉路口內', '早上', '20-40歲', '晴', '男'}	1404
{'交叉路口內', '普通重型機車', '平日', '四岔路', '晚上'}	1545
{'交叉路口內', '普通重型機車', '雨'}	1744
{'交叉路口內', '普通重型機車', '小於20歲'}	1745
{'40-60歲', '男', '計程車'}	1922
{'小於20歲', '晚上'}	1960
{'一般車道(未劃分快慢車道)', '普通重型機車', '平日', '20-40歲', '男'}	2115
{'晴', '自用小客車', '40-60歲', '平日'}	2141
{'交叉路口內', '普通重型機車', '平日', '20-40歲', '四岔路'}	2870
{'自用小客車', '40-60歲', '男'}	3069
{'普通重型機車', '平日', '20-40歲', '晴', '男'}	4829
{'普通重型機車', '平日', '晴', '男'}	9110

▲ 擷取部分較有意義的 Frequent Patterns

由於交通路況以及政策宣導還是因區而異，較有鑑別度，因此我們針對各區做觀察，並發現多組特別的區域特徵。

I. 車禍發生率前六名分別為：中山區(6727)、北投區(5620)、信義區(5535)、文山區(4978)、中正區(4782)、大安區(4583)。

II. 通常 1-set 或 2-set 之頻率會較高，然而 4-set 之頻率卻高達 9110，表示這是一個普遍的現象，男性普遍平日較喜歡騎機車通勤，而一般雨天會有部分民眾畏懼風雨而改搭大眾運輸工具，因此頻率才會如此之高。

{'普通重型機車', '平日', '晴', '男'}	9110

III. 20-40 歲之青年人士特別容易在中山區和信義區之交叉路口發生車禍，這是其他區沒有的現象。而其中，中山區又以四岔路之交叉路口車禍為多。

{'20-40歲', '交叉路口內', '03中山區', '四岔路'}	1032
{'20-40歲', '交叉路口內', '03中山區'}	1340
{'20-40歲', '交叉路口內', '07信義區'}	1023
{'20-40歲', '03中山區', '四岔路'}	1392
{'20-40歲', '03中山區', '平日', '四岔路'}	1041

IV. 平日於中山區發生的車禍特別集中於星期一、三、五，其他區在星期上則無顯著特徵。

{'03中山區', '平日', '星期五'}	1075
{'星期一', '03中山區', '平日'}	1036
{'星期三', '03中山區', '平日'}	1123

V. 一般車禍經常發生於交叉路口內，然而在中山與信義區（特別是平日的中山區）卻連交叉口附近都經常發生意外。

{'03中山區', '平日', '交叉口附近'}	1073
{'03中山區', '交叉口附近'}	1342
{'07信義區', '交叉口附近'}	1067

VI. 中山區和信義區之雨天車禍也較其他行政區多。中山區車禍如此頻繁之因素推測與市民大道、中山北路之路況混亂有關 [3]。

{'03中山區', '雨'}	1110
{'07信義區', '雨'}	1139

LHS	RHS	Support	Confidence
{'普通重型機車', '20-40歲', '下午', '晴'}	{'男'}	1398	0.728
{'交叉路口內', '三岔路', '普通重型機車', '平日'}	{'男'}	1835	0.708
{'普通重型機車', '平日', '下午', '晴'}	{'男'}	2159	0.718
{'平日', '凌晨'}	{'男'}	1212	0.718
{'自用小客車', '40-60歲', '男'}	{'平日'}	2255	0.735
{'平日', '乘客'}	{'女'}	1828	0.784
{'平日', '20-40歲', '下午', '男'}	{'普通重型機車'}	1582	0.628
{'平日', '晴', '男', '09北投區'}	{'普通重型機車'}	1086	0.576
{'一般車道(未劃分快慢車道)', '早上', '普通重型機車', '平日', '直路'}	{'20-40歲'}	1015	0.551
{'平日', '男', '計程車'}	{'40-60歲'}	1442	0.620

▲ 擷取部分較有意義與歸納價值的 Association Rules (LHS → RHS)

從上述的關聯規則中可以推論或觀察：

I. 一般車道早上平日之機車事故多發生於 20-40 歲人士身上，推測為早上上班時間車流量大容易發生事故。
II. 平日遭遇計程車事故之男性多介於 40-60 歲。
III. 40-60 歲駕駛自用小客車之男性多於平日發生車禍。
IV. 平日車禍事故之乘客多為女性，推測是因為女性通常自行駕駛的比例較低，通常會搭計程車、親友之車輛或大眾運輸工具。
V. 平日凌晨事故關係人多為男性，推測成年男性可能由於工作應酬之故，多於清晨才駕車返回住處，而部分青少年深夜也多在外遊蕩；另外清晨路上通常較無車輛，故駕駛人容易忽視道路速限或面對突發狀況反應不及而導致事故發生。
VI. 無論何種事故，特別是機車事故時，肇事者或受害者都以男性居多，推測和大多男性相較於女性而言，駕車風格較為急躁，以及多喜愛享受駕車樂趣有關。
VII. 20-40 歲以及北投地區的男性以騎乘機車造成的事故為多，推測可能北投地區山多路窄，機車通勤較為方便，且許多文化大學與陽明大學之教師生每日皆需上下通勤之故。

四、結論

從上一部分的實驗結果來看，我們能從交通事故資料集中萃取出頻繁發生且具有探勘有用性的事故發生組合，以及藉由地理資訊得出的 black spots。而由於原始資料集提供的明細紀錄有限，無法得知更細節的事故發生經歷，例如：是否有無關於駕駛人的其他肇事因素 (車體故障或其他外力因素，如：風速、日照等)；是否酒駕或精神不濟等更多生理因素；以及事故現場的詳細記錄 (是否逆向、超速、違規左轉等)。若能獲得上述的資料，就更能得出造成車禍事故頻繁的原因組合。再檢討這些原因，做出預防的提醒措施或是宣導工作，如：在路口處新增拍照違規的據點及告示，讓用路人能控制車速；在某些容易發生事故的時段與路段，多派駐警力掌控交通狀況，並改善該地點的基礎設施與環境，以期減少事故發生。

透過關聯規則演算法得出的結果，僅能表示在某些狀況的組合下較易有事故發生，其背後真實的因果關係與解釋還是有待更多專家知識的輔助，如：心理學、物理學等，從而協助政府調整交通政策，以適應各地區不同的交通情況，尤其是針對 Black Spot 的重點派駐；而在資料量上，我們使用的資料集會容易有資料傾斜的問題，屬性值出現的頻率不一，限制了不同事件的包含量，在 support 和 confidence 的設定上就必須多嘗試調整，或是在資料數量上做控制與平衡，才能探勘出較為完整且有用性高的 frequent pattern 及 association rules。

五、參考資料

[1] 臺北市政府警察局交通警察大隊 - 交通事故資料
http://data.taipei/opendata/datalist/datasetMeta?oid=2f238b4f-1b27-4085-93e9-d684ef0e2735
[2] Christian Borgelt (2005). Keeping Things Simple: Finding Frequent Itemsets by Recursive Elimination. *OSDM Proceedings of the 1st International Workshop on Open Source Data Mining*, pages 66-70, 2005.
[3] 十大易肇事路口，中山區最危險四路段上榜 | ETtoday 旅遊雲
https://travel.ettoday.net/article/468944.html