社群媒體分析_第四次讀書會作業

第 7 組

N104020013 吳映儒
N104020005 李姿儀
N104020011 陳冠宏
N104020012 蔡京叡
N104020015 李宇婕
N104020016 楊世華

目錄

一、探討主題

此次我們選擇的探討主題為分類運動項目，由於課堂中助教所分類的為新聞的看板類型，例如：主題為運動、兩岸的新聞，藉由訓練模型，達到自動判斷文章的類別，也因此讓我們思考是否有辦法進一步以運動新聞為主題，區分運動的種類，經過討論後，我們最終決定以五大運動種類來進行訓練，分別是：棒球、足球、籃球、羽球以及其他。

二、訓練流程

本組本次使用課程提供之文字探勘工作流程設計平台（Tarflow）進行文字探勘，並預計依據第三次讀書會分析結果再進一步分析主題，另外新增樂天棒球投手與打者之間關係進行社會網路圖的實作，第四次讀書會分析流程如圖1(GuidedLDA 主題模型)，使用之 Tarflow 名稱為 Z 。

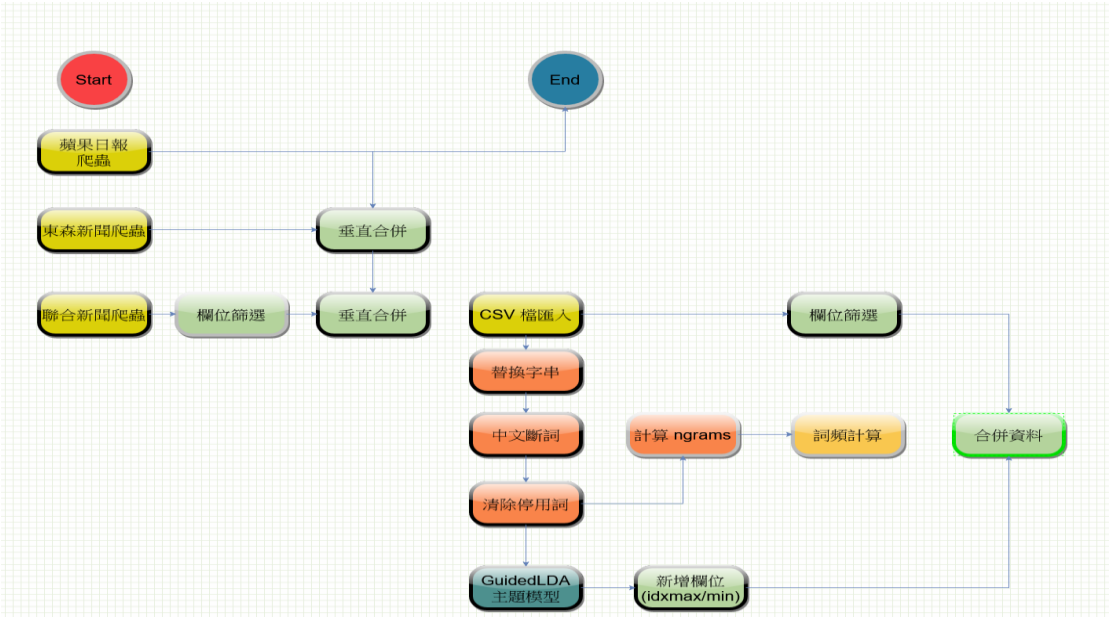

圖 1、GuidedLDA 主題模型分析流程

(一) GuidedLDA 主題模型

分析步驟如下(資料前處理部分與第三次讀書會相同)：

1. 資料爬取：我們選擇從蘋果日報、東森新聞、聯合新聞網分別爬取運動類的新聞。

   看板選擇運動版，搜尋關鍵字選擇棒球、籃球、足球、羽球，由於前陣子經典賽新聞較多，擔心結果過於導向棒球，因此排除關鍵字放入經典賽。

   替換字串：選擇替換字典 dic_sport。
2. 欄位篩選：保留了相關所需要的欄位。
3. 資料前處理：透過中文斷詞、清除停用字、ngrams 等方式，最後合併資料，相關設定方式與先前相同。
4. 人工分類：合併資料後，匯出 CSV 檔，並以人工的方式將新聞以籃球、棒球、羽球、足球、其他五大分類，進行運動項目細分，如圖 2。

Show 10 entries

	system_id	artUrl	artCatagory	artTitle	artDate	artContent	dataSource	c羅	mvp	nba	plg	sbl	一年	一度	一次	一直	一路	一軍	三分球	上場	上演
0	1	https://www.chinatimes.com/newspapers/20230101000654-260111	籃球	展望2023經典賽中華隊月底掀面紗	2023/1/1 04:10	今年台灣體壇的國際大賽征戰，3月就要從第5屆世界棒球經典賽打頭陣，期盼「最強中華隊」順利成軍，而2022杭州亞運延至今年，尋求衛冕的「舉重女神」郭婷淳、「羽球一姐」戴資穎各有挑戰，此外，在去年「魔獸」...	chinatimes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	https://www.chinatimes.com/newspapers/20230101000656-260111	其他	亞運、世大運延至今年 考驗中華健兒	2023/1/1 04:10	受新冠疫情影響，原本要在去年舉辦的成都世大運與杭州亞運都被迫延遲1年，成都世大運確定今年7月28日至8月8日舉辦，杭州亞運則於9月28日登場、10月8日開幕，中華健兒在這兩場國際賽事都將面臨歐哥考驗，...	chinatimes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

圖 2、人工分類的 CSV 檔

5. GuidedLDA 主題模型設定如圖 3，反覆測試，最後將主題數分成五群，詞頻下限設定為 50，保留主題關鍵字為 10 個，迭代次數拉高至 1000 次，詞彙頻率上線設定為 0.7。

主題種子字設定：

(1) 棒球,三振,投手,富邦悍將,中信兄弟,安打,樂天桃猿,大聯盟,大谷翔平

(2) 羽球,戴資穎,世界羽聯世界巡迴賽

(3) 足球,世界盃足球賽,梅西,C 羅,沙烏地阿拉伯聯賽,曼聯

(4) 籃球,PLG,NBA,助攻,籃板,超級籃球聯賽,林書豪,霍華德,詹姆斯

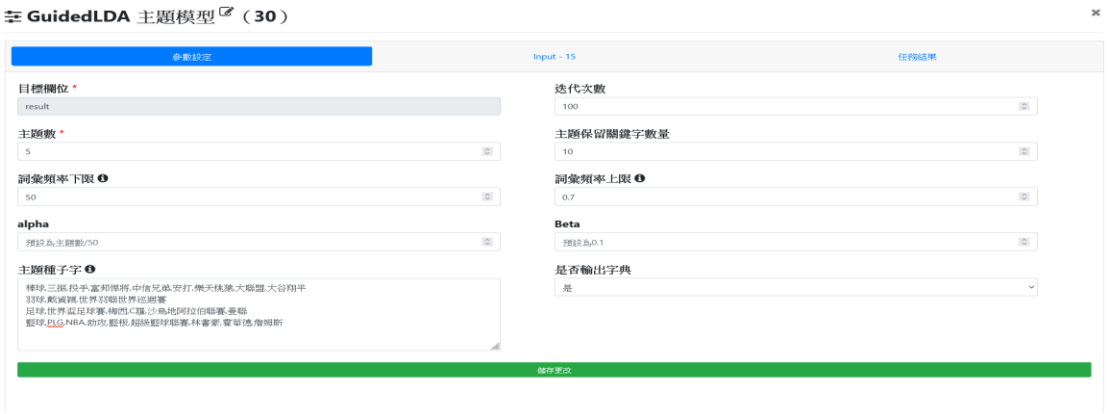

圖 3、主題模型 LDA 設定

6. 分析結果

統計資訊

50 字數	5 主題數	-1.876 主題連貫性(UMass)	-0.193 主題連貫性(PMI)
0.596 主題連貫性(Cv)	458.03 混淆度		

(1) 第一群為「籃球」

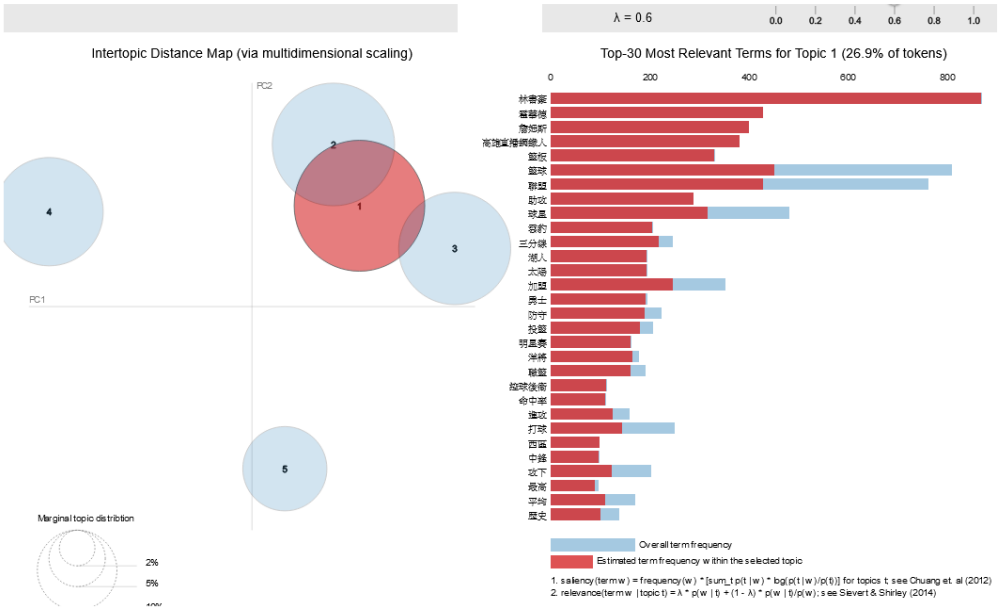

圖 4、第一類主題

(2) 第二群為「其他」

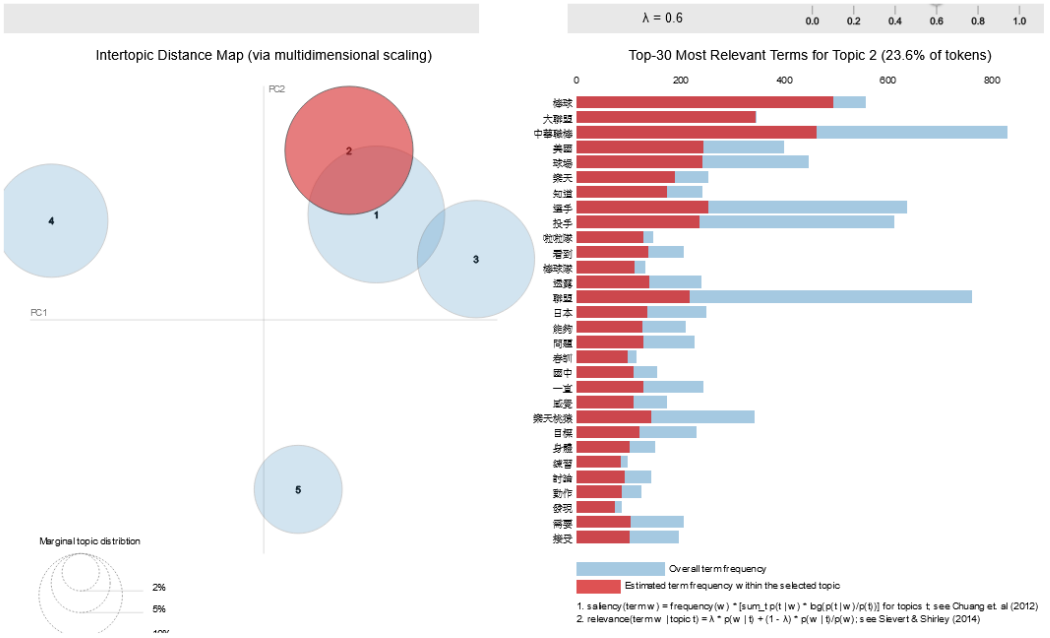

圖 5、第二類主題

(3) 第三群為「足球」

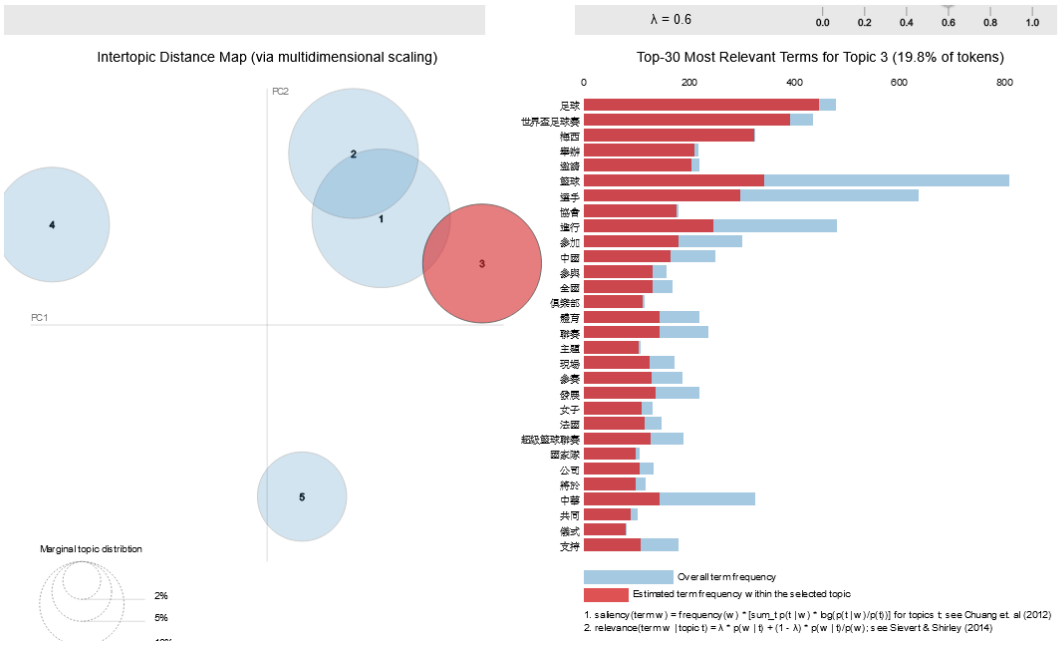

圖 6、第三類主題

(4) 第四群為「棒球」

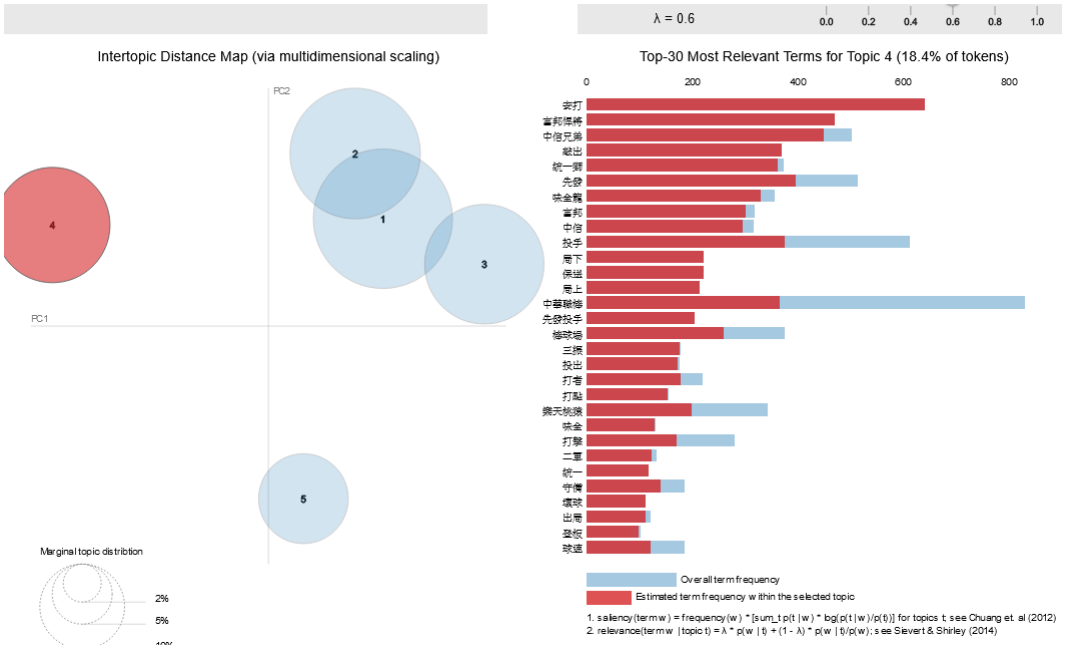

圖 7、第四類主題

(5) 第五群為「羽球」

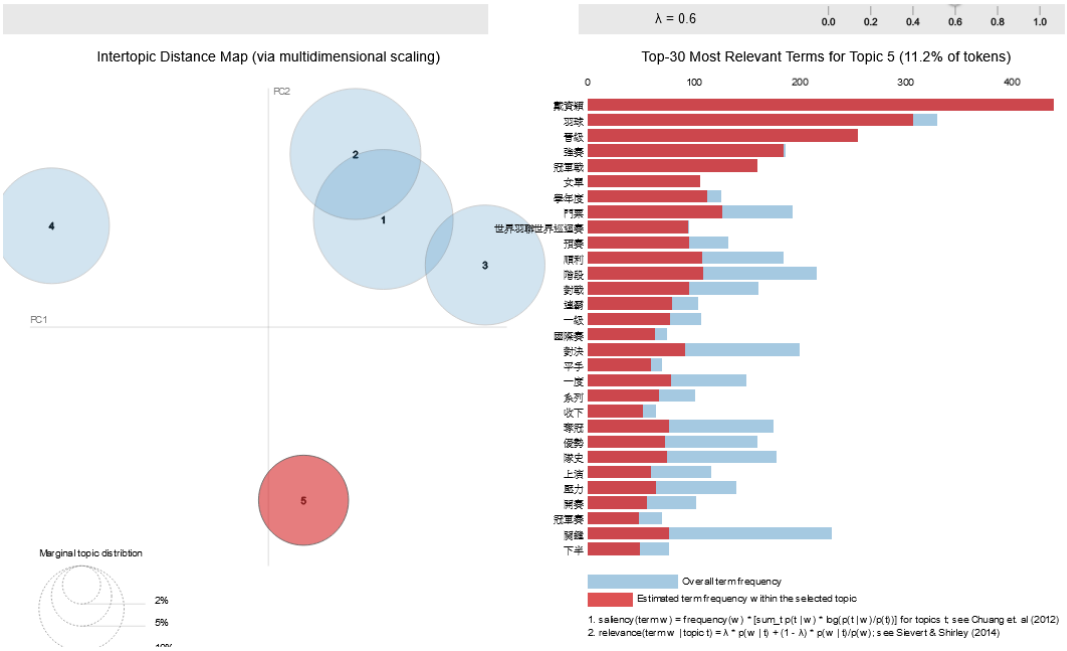

圖 8、第五類主題

主題關鍵字(λ = 1，TOP10)：

表 1、主題關鍵字結果

	1	2	3	4	5	6	7	8	9	10
Topic0	林書豪	籃球	聯盟	霍華德	詹姆斯	高雄直播鋼鐵人	籃板	球星	助攻	加盟
Topic1	棒球	中華職棒	大聯盟	選手	美國	球場	投手	聯盟	樂天	知道
Topic2	足球	世界盃足球賽	籃球	梅西	選手	進行	舉辦	邀請	參加	協會
Topic3	安打	富邦悍將	中信兄弟	先發	投手	敲出	中華職棒	統一獅	味全龍	富邦
Topic4	戴資穎	羽球	晉級	強賽	冠軍戰	門票	學年度	階段	順利	女單

(二) 社會網路圖

分析流程

1. 資料來源：2020 中華職棒上半季各投手之對戰紀錄（數據來源）
2. 資料前處理：重新設定 Gephi 要求之格式
3. 使用 Gephi：從原始資料整理好的輸入檔

三、結論

(一) GuidedLDA 主題模型

本組透過 GuidedLDA 主題模型之結果進行比對及驗證，最終得到分析結果，如下表：

表 2、比對結果

		實際					
		棒球	羽球	足球	籃球	其他	合計
預測	棒球	246	0	0	1	1	248
	羽球	17	107	13	37	10	184
	足球	40	8	110	100	67	325
	籃球	3	0	14	363	14	394
	其他	246	10	13	23	62	354
	合計	552	125	150	524	154	1505

表 3、驗證結果

	TP	**FP**	**FN**	**precision**	**recall**
棒球	246	2	306	99.19%	44.57%

	TP	FP	FN	precision	recall
羽球	107	77	18	58.15%	85.60%
足球	110	215	40	33.85%	73.33%
籃球	363	31	161	92.13%	69.27%
其他	62	292	92	17.51%	40.26%
macro	888	617	617	59.00%	59.00%

從結果來看，「羽球」這個類別的準確率較高外，有達到 80%以上，其他我們認為，「棒球」被分為「其他」之可能原因可能是因為非典型運動新聞，如採訪、記者會、球員故事、啦啦隊、合約等;「籃球」被分為「足球」的原因可能是非典型運動新聞，分類為足球主題本身的關鍵字即較不具特殊性及代表性，導致足球主題適用的關鍵字套用在籃球上也合理。

最後，本組認為非典型運動新聞特別容易被錯分，例如描述體育政策、賽前/後球員採訪、球員記者會、球員個人事件新聞、賽事宣傳、球隊人事異動等。

(二) 社會網路圖

1. 分析對戰過較多打者之投手：

卡本特、黃子鵬、霸能、尼寇力、王躍霖、王溢正，推測應該是在場上待得比較久的投手，可能是比較穩定的投手。

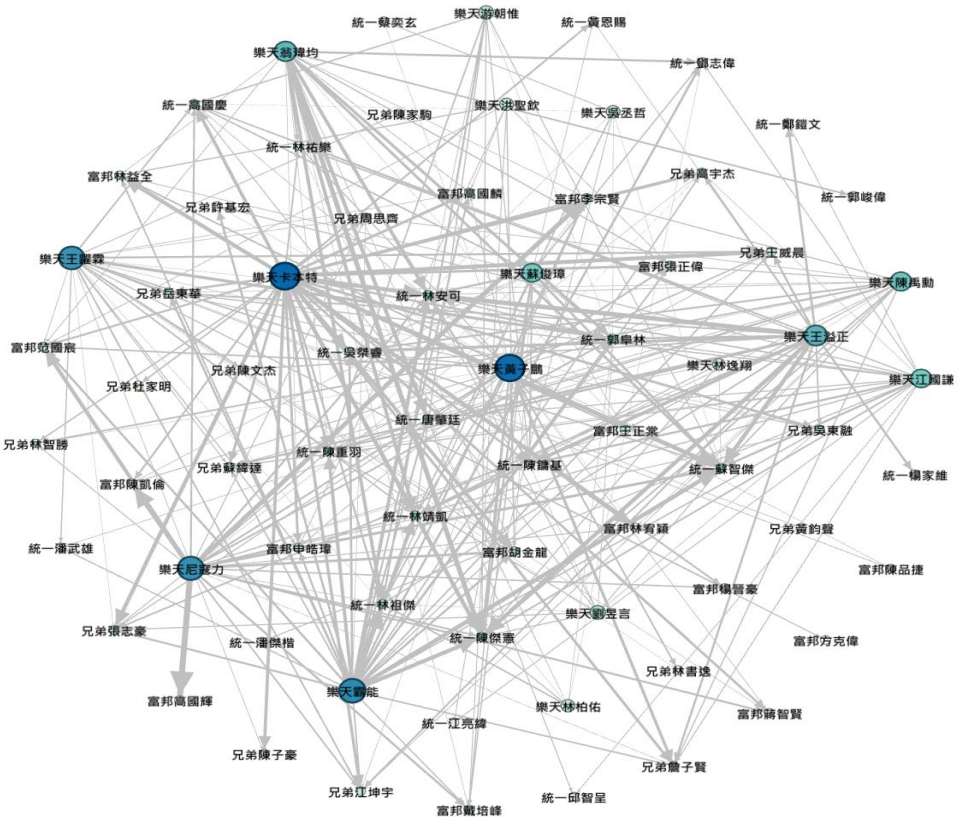

圖 9、投手分析圖 1-對戰數多

2. 分析使用伸卡球之投手：

從圖 10 可以發現，較常使用伸卡球的投手有樂天的霸能、樂天的黃子鵬，也從他們兩位投手各自的分析圖(圖 11、12)中發現較常對某幾位打者使用伸卡球，推測是因為這幾位打者對於伸卡球較不拿手。

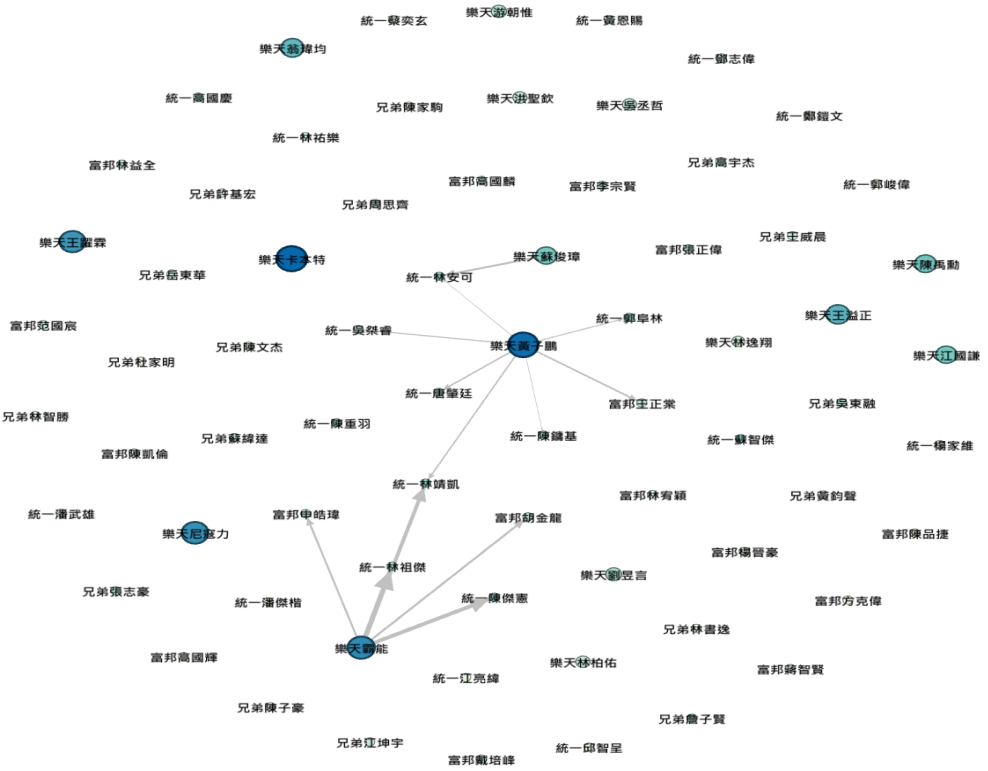

圖 10、投手伸卡球分析圖

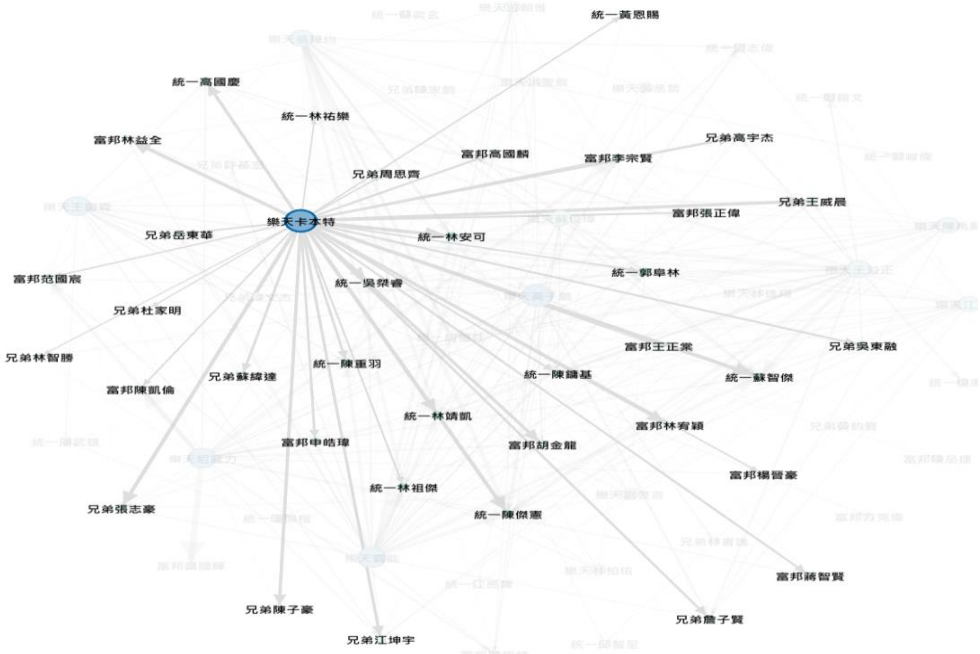

圖 11、投手卡特之伸卡球分析圖

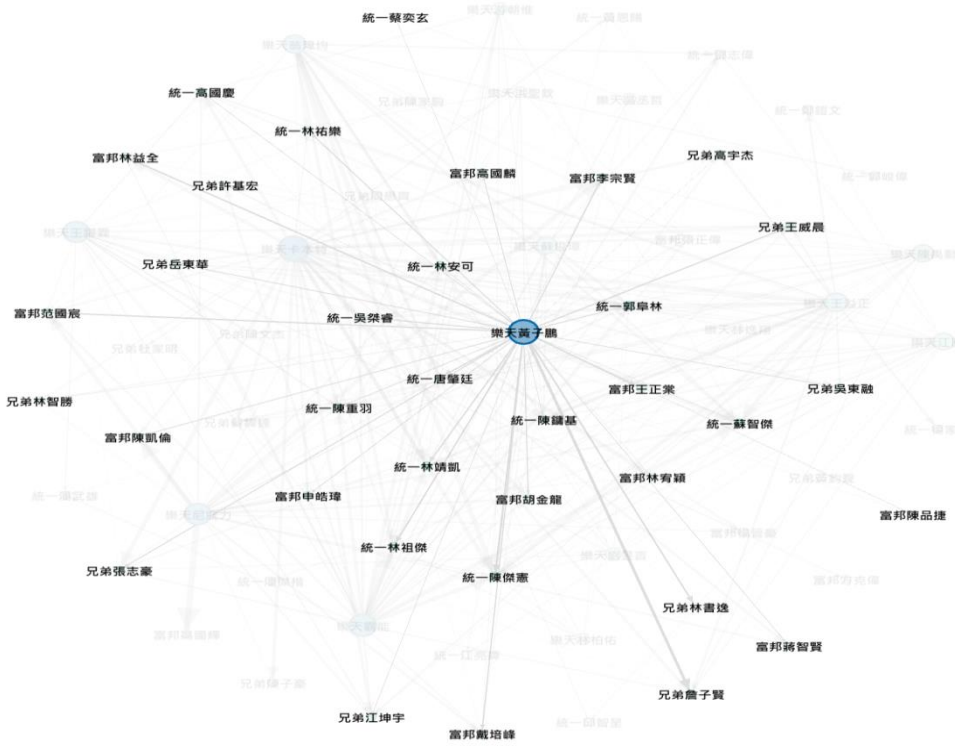

圖 12、投手黃子鵬之伸卡球分析圖

3. 特定投手針對特定打者分析：

特定投手(黃子鵬)對特定打者(王正棠)有偏愛策略(伸卡球)

有可能是隊上的策略，或是富邦的王正棠對伸卡球的表現特別不好，所以當黃子鵬對上王正棠時，會較常使用伸卡球來應對。

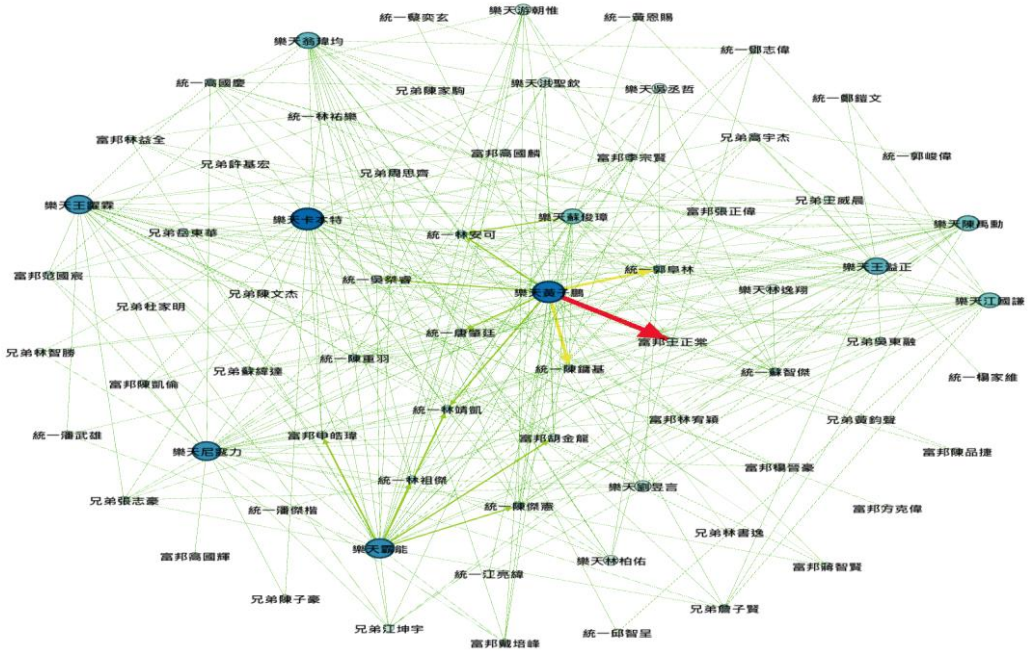

圖 13、黃子鵬針對特定打者分析圖-

4. 不同投手針對特定打者策略分析：

下圖可以看出不同投手使用相同策略對戰統一林靖凱的分析，以變化球來說，有六位投手曾經對他使用變化球，但其中可以看到樂天的霸能相較其他投手在對陣林靖凱時特別常用變化球(紅色)。

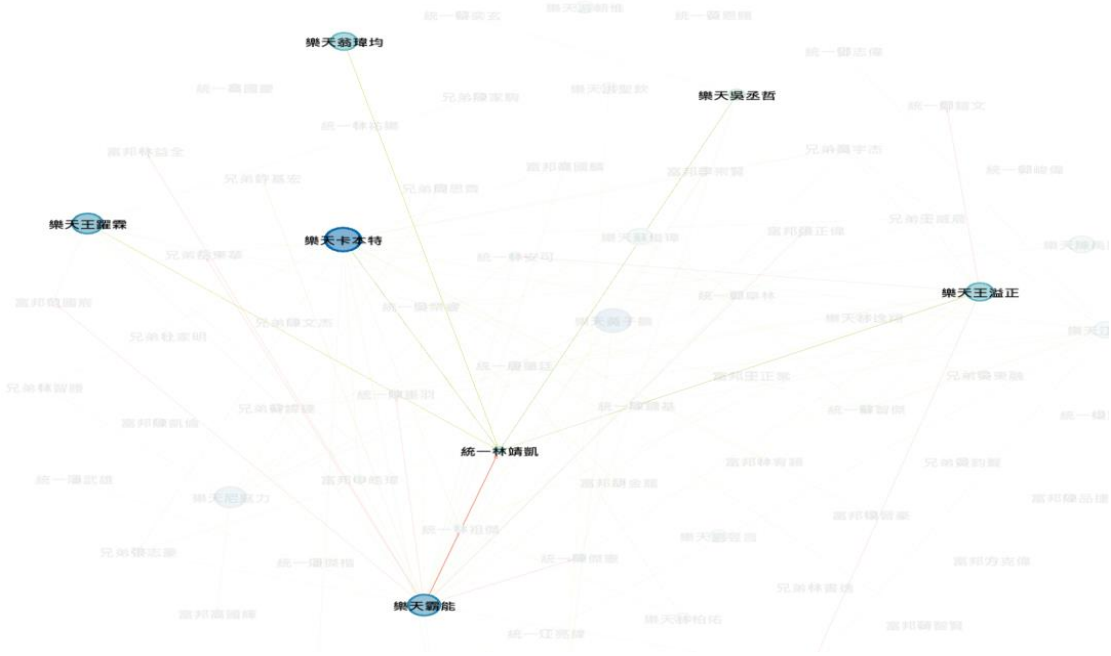

圖 14、打者林靖凱對戰分析圖-

5. 分析使用變化球之投手：

從下圖可以看到樂天的霸能在對戰時特別喜歡用變化球，再來是樂天的王益正。

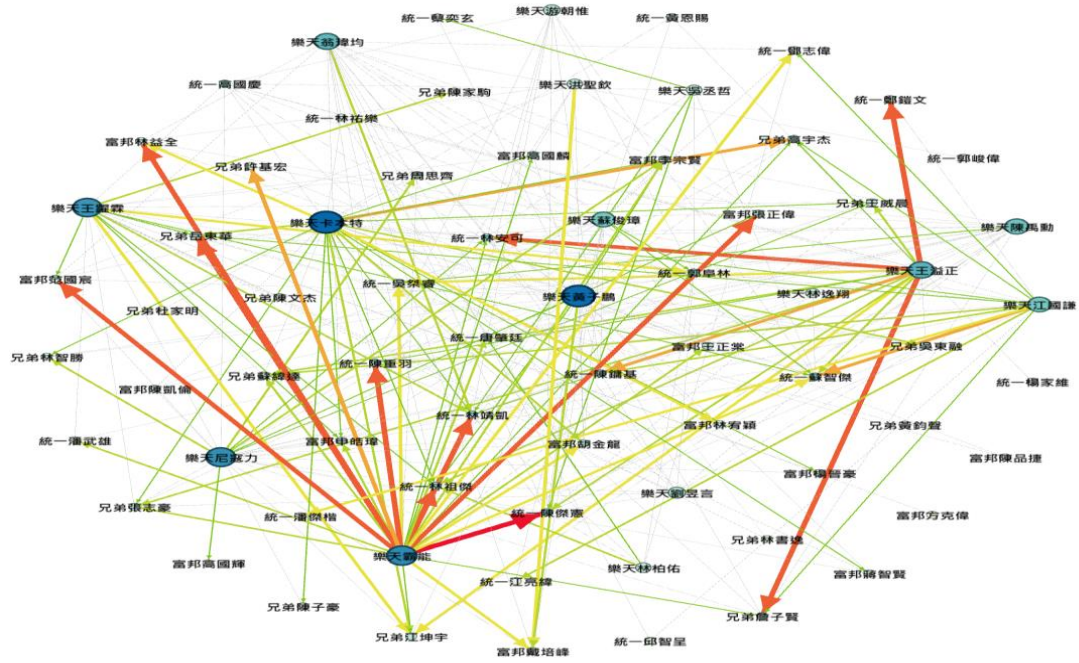

圖 15、投手選擇變化球分析圖

最常投變化球的樂天霸能，使用變化球時對陣的投手，可以看到他對某幾位打擊者時，會使用變化球來對付，猜可能是這幾位打者可能比較不會打變化球；相對的沒被投過變化球的打者可能相對比較會處理變化球，或策略上的因素。

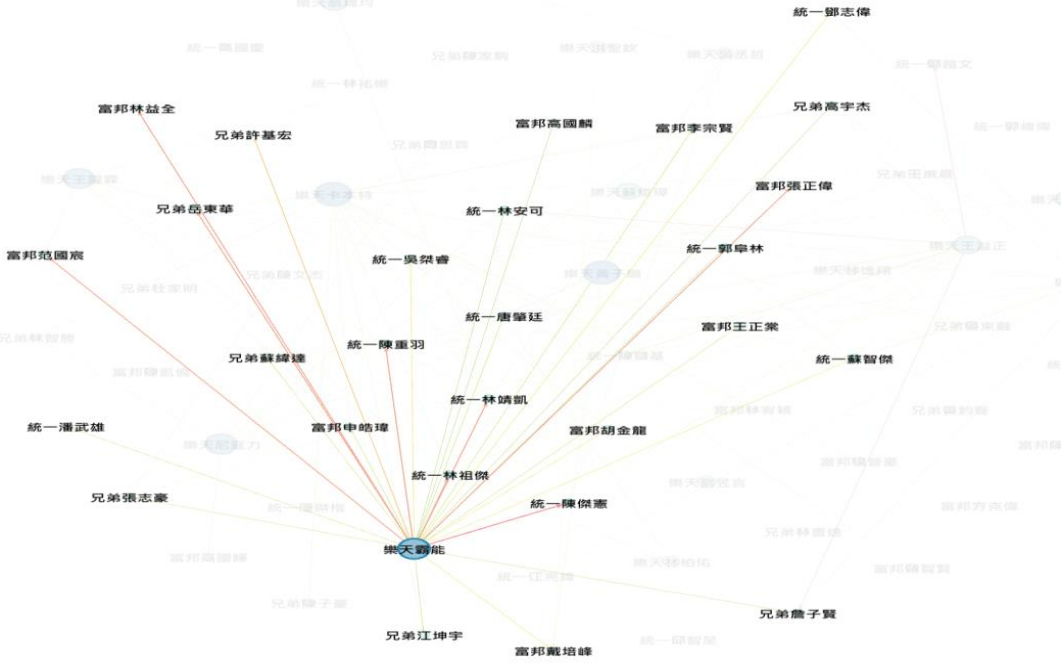

圖 16、投手霸能對戰分析圖