國立政治大學資訊管理學系

碩士學位論文

指導教授: 楊建民 博士

運用 kNN 文字探勘分析智慧型終端 App 群集之研究

The Study of Analyzing Smart Handheld Device App's Clusters by Using kNN Text Mining

研究生：曾國傑

中華民國 101 年 7 月

誌謝

在碩士班這兩年的研究生活中，首先要感謝指導教授楊建民老師的細心教導，讓我在課業的研究訓練及實務的能力培養上，都能建立扎實的基礎；也透過與老師的日常相處及籌備「商管聯盟」的相關事宜中，學習到不少生活態度與待人處世的道理；在此由衷地感謝楊老師，並致上最誠摯的祝福與謝意。而能夠順利完成本篇論文也要感謝劉文卿老師、邱光輝老師與季延平老師在論文提報與口試期間的專業指導與建議，使得論文內容能更加完整與嚴謹。

在學期間，學長姊的經驗傳承與同儕間的相互切磋都豐富了我研究所生活；感謝柏均、智民與振和等學長姐在論文上的幫助，有了你們不厭其煩的解答，讓我在論文的探索上能夠更加順利；感謝研究室的夥伴們：婉婷、鴻仁與康維的互相砥礪與教學相長，讓我在困難時不覺得孤單、快樂時也感受到加倍地喜悅；也感謝碩班同學們：雅筑、珮筠、佑純等人在論文之虞的加油打氣，讓碩士班生活除了研究之外，增添了許多歡笑與難忘的回憶，這一切都讓我更懂得珍惜，也對於成為政大資管碩士班的一份子感到非常榮幸。

也感謝一直以來總是在陪在我身旁的摯友們：依真、雨軒、雅雯、宜璇、珊珊、家瑋和榆庭，你們總是在我難過的時候給我莫大的安慰，讓我有更多的勇氣能夠去承擔挫折，也讓我更有信心去面對未來的挑戰。最後，將我最大的感謝獻給我的家人們，有了你們的支持與包容，讓我能夠無憂無慮的專心在碩士的研究生活中；並將這份喜悅與榮耀獻你們以及所有我愛的人。

摘要

隨著智慧型終端設備日益普及，使用者對 App 需求逐漸增加，各大企業也因此開創了一種新的互動性行銷方式。同時，App 下載所帶來的龐大商機也促使許多開發人員紛紛加入 App 的開發行列，造成 App 的數量呈現爆炸性成長，而讓使用者在面對種類繁多的 App 時，無法做出有效率的選擇。故本研究將透過文字探勘與 kNN 集群分析技術，分析網友發表的 App 推薦文並將 App 進行分群；再藉由參數的調整，期望能透過衡量指標的評估來獲得最佳品質之分群，以便作為使用者選擇 App 之參考依據。

為了使大量 App 進行分群以解決使用者「資訊超載」的問題，本研究以 App Store 之遊戲類 App 為分析對象，蒐集了 439 篇 App 推薦文章，並依 App 推薦對象之異同，將其合併成 357 篇 App 推薦文章；接著，透過文字探勘技術將文章轉換成可相互比較的向量空間模型，再利用 kNN 群集分析對其進行分群。同時，藉由參數組合中 k 值與文件相似度門檻值的調整來獲得最佳品質之分群；其分群品質的評估則透過平均群內相似度等指標來進行衡量；而為了提升分群品質，本研究採用「多階段分群」，以分群後各群集內的文章數量來判斷是否進行再分群或群集合併。

本研究結果顯示第一階段分群在 k 值為 10、文件相似度門檻值為 0.025 時，能獲得最佳之分群品質。而在後續階段的分群過程中，因群集內文章數減少，故將 k 值降低並逐漸提高文件相似度門檻值以獲得分群效果。第二階段結束後，可針對已達到分群停止條件之群集進行關鍵詞彙萃取，並可歸類出「棒球/射擊」與「投擲飛行」等 6 種 App 類型；其後階段依循相同分群規則可獲得「守城塔防」等 14 種 App 類型。分群結束後，共可分出 36 個群集並獲得 20 種 App 類型。分群過程中，平均群內相似度逐漸增加；平均群間相似度則逐漸下降；分群品質衡量指標由第一階段分群後的 12.65%提升到第五階段結束時的 75.81%。

由本研究可知分群之後相似度高的 App 會逐漸聚集成群，所獲得之各群集命名結果將能作為使用者選擇 App 之參考依據；App 軟體開發人員也能從各群集之關鍵詞彙中了解使用者所注重的遊戲元素，改善 App 內容以更符合使用者之需求。而以本研究結果為基礎，透過建立專業詞庫改善分群品質、利用文件摘要技術加強使用者對各群集之了解，或建立 App 推薦系統等皆可做為未來研究之方向。

關鍵字：App、kNN、群集分析、文字探勘

ABSTRACT

With the popularity of Smart Handheld Devices are increasing, the needs of "App" are spreading. Developers whom devote themselves to this opportunity are also rising, making the total number of Apps growing rapidly. Facing these kind of situation, users couldn't choose the App they need efficiently. This research uses text mining and kNN Clustering technique analyzing the recommendation reviews of App by netizen then clustering the App recommendation articles; Through the adjustments of parameters, we expect to evaluate the measurement indicators to obtain the best quality cluster to use as a basis for users to select Apps.

In order to solve the information overload for the user, we analyzed apps of the "Games" category form App store and sorted out to 357 App recommendation articles to use as our analysis target. Then we used text mining technique to process the articles and uses kNN clustering analysis to sort out the articles. Simultaneously, we fine tuning the measurement indicators to find the optimal cluster. This research uses multi-phase clustering technique to assure the quality of each cluster.

We discriminate 36 clusters and 20 categories from the clustering results. During the clustering process, the Mean of Intra-cluster Similarity increases gradually; in the contrary, the Mean of Inter-cluster Similarity reduces. The "Cluster Quality" increases from 12.65% significantly to 75.81%. In conclusion, similar Apps will gradually been clustered by its similarities, and can be used to be a reference by its cluster's name. The App developers can also understands the game elements which the users pay greater attentions and tailored their contents to match the needs of the users according to the key phrases from each cluster. In further discussion, building specialized terms database of App to improve the quality of the clustering, using summarization technique to robust user understanding of each cluster, or to build up App recommendation system is liking to be further studied via using the results by this research.

Keywords: App、kNN、Clustering、Text Mining

目錄

圖目錄

表目錄

第一章、緒論

第一節、研究背景與動機

在現今資訊科技發達的時代中，手機的普及率已達到「人手一機」的境界。隨著使用者的需求日益增加的情況下，其對手機的功能不只要求具備一般的聯絡及溝通，而是希望增加事務管理、多媒體等功能，甚至必須滿足所有食、衣、住、行的需求。因此許多 3C 品牌廠商便推出「可讓使用者自行安裝所需軟體，且軟體廠商可依此系統平台開發相容的軟體讓使用者使用」的智慧型手機。除了智慧型手機外，自從 Apple 在 2010 年推出 iPad 後，平板電腦的發展也逐漸受到大眾關注，眾多廠商開始投入平板電腦市場，就連以電子商務為主要服務發展的 Amazon 也在 2011 年 9 月推出 Kindle Fire，使得平板電腦市場的競爭愈來愈激烈，也成為了智慧型終端設備的另一個值得注目的焦點。

而在智慧型終端問世後，運行於智慧型終端上的應用程式(App)市場也逐漸擴大，根據國際研究機構 Gartner 統計：2010 年的 App 總下載量為 82 億次，2011 年倍增到 177 億次，預估到 2014 年，從蘋果的 App Store、Android Market 等各 App 平台的全球用戶之總下載量將會達到 1,850 億次，代表未來 App 在市場上潛力無窮(胡秀珠，2011)。

與此同時，研究也發現使用者在智慧型終端的使用習慣上有明顯的變化。根據美國 Flurry 公司的統計指出，使用者在 2010 年 6 月每日在 App 的平均花費時間少於網路瀏覽；但在隔年 6 月，使用者每日在 App 之平均使用時間約為 81 分鐘，已超越同時間使用者在網路瀏覽時所花費的 74 分鐘。而當使用者花愈來愈多時間在 App 功能上，其對 App 的需求不僅愈來愈高，也愈多元了(SmartMobix，2012)。

也因為使用者的大量需求,讓許多應用程式開發人員觀察到其中所隱含的商機，紛紛投入 App 的開發行列，使得 App 的數目不斷增加。龐大數量的 App，其品質參差不齊，讓使用者在選擇時不知道如何、也不容易做出決定。其主要原因是來自於過多的選擇，使得使用者面對的資訊供給量過大而造成困擾，這種情況就是一般人所熟知的資訊超載(Information Overload)或資訊超過負荷的現象。

雖然智慧型終端 App 的官方平台為了讓使用者在搜尋中能更加便利，會針對 App 進行分類，但所列出之分類往往是由官方自行訂定，沒有考慮到使用者的使用感受，而使得使用者仍無法從中選擇自己所需的 App。故為了將大量 App 分群以解決使用者「資訊超載」的問題，本研究將以 Apple 之 App Store 所提供的遊戲類(Games)App 為主，並在網路論壇中收集與其相關之推薦文章做為資料分析對象，再透過文字探勘與群集分析技術獲得 App 之最佳分群，並針對各分群結果進行群集命名。

第二節、 研究目的

依前述之背景與動機，本研究所要達成之研究目的分述如下：

1. 利用文字探勘技術對 App 推薦文進行斷詞、文件特徵選取等資料處理，並透過群集分析中的 kNN 演算法依 App 之特性將 App 分群。
2. 擬定適切的分群規則以提高 App 分群品質,進而獲得最佳之 App 分群結果。
3. 最後，萃取具代表各群的關鍵詞彙進行群集命名，以作為使用者選擇 App 之依據。

第二章、文獻探討

第一節、智慧型終端應用程式(Applications, App)

智慧型終端應用程式，亦稱為 App，現今泛指運行在智慧型終端設備上的應用程式。使用者可下載符合自身需求的 App，以擴增行動設備之功能或擷取所需資訊。其功能圍繞於日常生活中的各項需求，例如：地圖、時間、天氣、遊戲、飲食、旅遊、閱讀...等等。

而 App 隨著使用者的喜好與使用習慣，不僅造成一股下載使用的熱潮，也讓各大企業發現了一種更便利的互動行銷平台，解決了過去想要利用科技來達到廣告、行銷的目的，卻往往成本過於昂貴且不具有大眾普及性的問題；其中所隱含的龐大商機也讓許多軟體開發廠商紛紛投入 App 的開發行列，造成 App 數量呈現爆炸性的成長，面對這眾多的選擇讓使用者面臨了「資訊超載」的問題。

現今許多智慧型終端 App 平台皆提供上萬種 App 讓使用者來選購下載，以目前全球的線上軟體商店的龍頭—Apple之App Store 所提供的App數量為例(圖 2-1)，可供使用者下載使用的 App 於 2012 年 3 月就已高達約 60 萬種，其中遊戲類型的 App 為大宗，共包含約 10 萬種(148Apps.biz, 2012)。

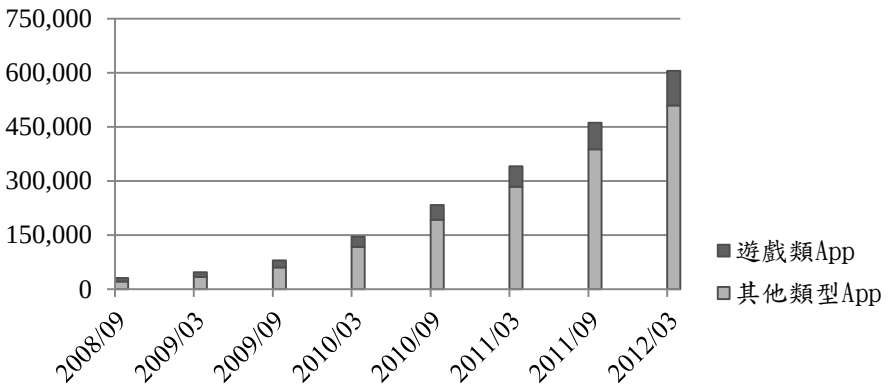

圖 2-1 App Store 提供之 App 數量統計

(資料來源：148Apps.biz, 2012)

在 App 分類部分，App Store 將 App 依功能需求分為遊戲、書籍、音樂等約 22 種類別(表 2-1)，其中因遊戲類(Games)App 的數量較多，故在該分類下又依據遊戲類型的不同，區分為動作類、冒險類等子類別(表 2-2)，以便讓使用者在選擇時能夠依照類別尋找有興趣的 App(Apple, 2012)。

表 2-1 App Store 之 App 類別

1. 工具程式	6. 社群網路	11. 書籍	16. 教育	21. 導覽
2. 天氣	7. 音樂	12. 財經	17. 新聞	22. 醫藥
3. 生活風格	8. 娛樂	13. 健康與瘦身	18. 照片和視訊	
4. 生產力工具	9. 旅遊	14. 參考	19. 運動	
5. 目錄	10. 書報攤	15. 商業	20. 遊戲	

(資料來源：Apple, 2012)

表 2-2 App Store 之遊戲 App 子類別

1. 文字	6. 家庭	11. 棋盤	16. 撲克牌
2. 角色扮演	7. 益智	12. 策略	17. 模擬
3. 兒童	8. 動作	13. 街機	18. 賭場
4. 冒險	9. 問答	14. 運動	19. 賽車
5. 音樂	10. 教育	15. 骰子	

(資料來源：Apple, 2012)

雖然 App Store 為了增加使用者在搜尋時的便利性，而將 App 進行分類；但由於 App 的類別是由官方訂定後，再由開發者於 App 申請開發時自行選擇所屬類別，故可能造成 App 的歸類過於主觀而未考慮到多數使用者真實的使用感受，抑或在 App Store 所定義的官方類別中，無法涵蓋所有開發者需要之類別，皆會使得 App 的使用者無法透過官方類別尋找到合適的 App。

第二節、 文字探勘

因為科技的發展及資訊的傳遞過程日趨便利，使得網站和網頁數量的快速成長。而人們對於許多網頁上的文章、評論或文件無法很清楚且立即了解其所要表達的潛藏資訊，也無法以簡易的方法去分割所含有的資訊內容，在此時便對文字探勘的技術產生了需求。

2.2.1. 文字探勘的定義

文字探勘又可稱為文件資料探勘(Feldman & Dagan, 1995)，是一種由資料探勘所延伸而出的技術(Fayyad, 1996;Simoudis, 1996)；依據Sullivan (2001)的定義可知，文字探勘是一種「編輯、組織及分析大量文件的方法和過程，可提供特定使用者特定的資訊，以及發現特定資訊的特徵之間的關聯」；而巫啟台(2002)認為文字探勘是近年來興起的一個文件分析的研究課題，並將其定義為「從非結構化的文字中去發現有用的或是有趣的片段、模式等規則」。

由此可知文字探勘有別於一般資料探勘的技術：無法直接應用於非結構化的文件資料上的特性，只適用於已結構化的關聯式表格資料，而是結合了資訊擷取、自然語言處理等技術，並試圖從文件資料中找出重要的項目、片語或之間的關聯程度。

雖然不同的研究學者對文字探勘的定義不盡相同，但其定義所表達的目的都在於文字探勘能夠「從大量、非結構化的文字性資料中，方便且快速地整理出有用的資訊」。且如前述，因為文件資料大多不具結構性，無法直接進行分析，故必須預先對所要進行探勘的資料進行前處理，擷取出適當的資訊後才能進行後續的分析。也因此文字探勘需整合一些資訊檢索技術，如：關鍵資訊擷取、文件自動分類、全文檢索等，以便更容易地從文件資料中取得其所需的資訊。

2.2.2. 文字探勘的架構

在學者 Tan (1999)所提出的文字探勘架構(圖 2-2)中，主要包含兩大部分：文字萃取(Text Refining)及知識淨化(Knowledge Distillation)。

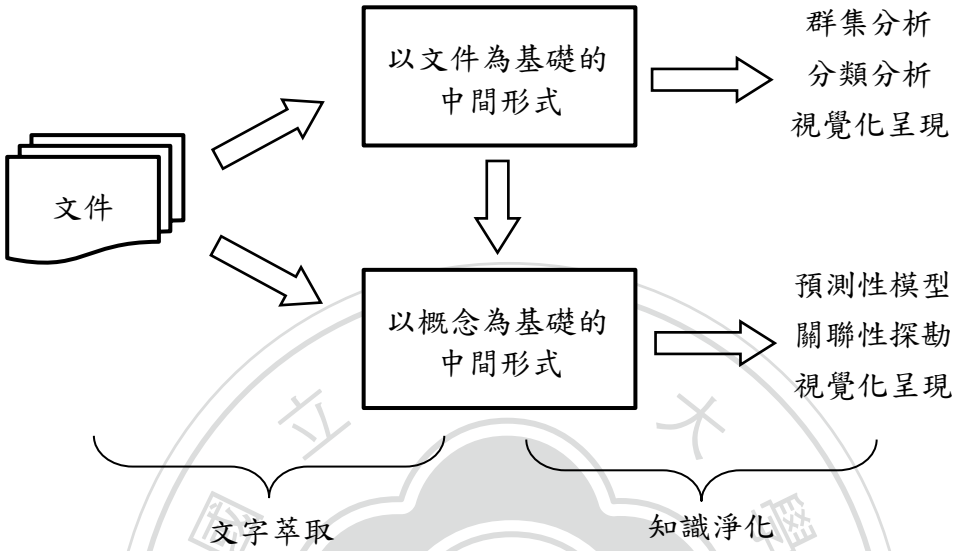

圖 2-2 文字探勘架構
(資料來源：Tan, 1999)

1. 文字萃取：將各種不規則的文件轉換成事先已決定好的中間形式；
2. 知識淨化：主要是將中間形式歸納出模式或知識。

文件所轉換的中間形式可用文件中介形式（Document-based）來表示，此時以一份文件表示一個實體，可以在文件中歸納出模式或彼此之間的關係，例如：群集分析、分類分析或視覺化呈現等；此外，也可用概念中介形式（Concept-based）來呈現，概念中介形式則是以一個物件或特定領域的概念來表示一個實體，分析結果包括預測性模型、關聯性探勘及視覺化呈現等等，亦可對文字中介形式資訊萃取，轉換成概念中介形式。

2.2.3. 文字探勘的相關技術

(一) 文章斷詞

對於任何語言來說，詞(Word)是最小有意義且可以自由使用的語言單位，所以任何語言處理的系統都必須先能分辨文本中的詞才能作進一步的處理，例如：機器翻譯、語言分析、資訊抽取等，因此自動斷詞處理便成了語言處理中不可或缺的技術。

而由於各種語言系統在結構及文法上的規則不盡相同，使得中、英文斷詞在本質與技術上存在非常大的差異。在英文的斷詞方面，是以字為單位，字與字之間以空格或是其他符號作區隔，故每個字即可以代表所含的意義及語意；但中文若以字為斷詞單位，則無法清楚得知所含的語意，故在中文文字上通常是以兩個字以上所組成的詞為單位，才具有明顯語意。也因為中文的詞與詞之間沒有一定的界線，相較於英文斷詞而言顯得複雜許多。

1. 常見的中文斷詞法

常見的中文斷詞法可分為詞庫斷詞法(Chen & Liu, 1992)、統計斷詞法(Sproat & Shih, 1990)及混和斷詞法(Nie, Brisebois, & Ren, 1996)，下列為各斷詞法的介紹：

(1) 詞庫斷詞法：

詞庫斷詞法是現今使用最廣泛的斷詞方式。需事先建立一個詞庫，再將文件中所出現的詞彙與詞庫中的詞彙互相比對，以找出有可能的分隔點。在比對的過程中通常使用「長詞優先法」來保留最完整的語意。此法具有直覺、易懂的優點，但若是新生詞不存在詞庫中或是無法掌握合適的詞庫大小時，則會降低斷詞的正確率，因此在詞庫的控制與維護上成為斷詞是否正確的關鍵。

(2) 統計斷詞法：

統計斷詞法需先經由大量文件或大型語言資料庫(Corpus)的訓練，取得足夠的統計參數(詞頻、門檻值)以作為斷詞的依據。由於不需人工定義詞彙，所以可解決複合詞、新生詞的問題，也省去了維護詞庫的負擔。但經由統計運算並無法考慮語意的正確性，因此在整個文句的表達上容易具有錯誤的可能性。且因為語言資料庫所屬的領域有所不同，故各領域的語言資料庫之間的統計參數也無法互相流通使用。

(3) 混和斷詞法：

混合斷詞法結合了詞庫斷詞法與統計斷詞法，先利用詞庫斷詞法找出許多不同組合的詞彙，再利用詞彙的統計參數來找出最佳的斷詞組合。此方法結合了上述兩個方法的優點，以增加斷詞的正確性與效率。

2. 中央研究院 CKIP 斷詞系統

CKIP 斷詞系統是由我國中央研究院詞庫小組所研發，採用的是混和斷詞法，其包含大約 10 萬多個常用中文詞彙，具有新詞辨識能力與附加詞類標記的選擇性功能。詞庫中所收錄的詞彙包含一般用詞、常用專有名詞、成語、慣用語等等。

而在此系統中，斷詞的處理分為兩部分：

(1) 斷詞：將文章根據中研院詞庫小組所維護的 10 萬多個詞彙，以一個句子為單位，把文字切割成數個獨立的詞。

(2) 標記詞性：將每一個斷詞後所產生的詞彙標記上所屬的詞性，詞性種類分別有動詞(V)、名詞(N)、連接詞(C)、語助詞(T)、副詞(D)、介詞(P)、感嘆詞(I)」、…等。該步驟有助於後續對各個詞彙的詞性之掌握與使用。

(二) 文件特徵選取

對文件進行斷詞處理後，便需篩選出具有代表性的特徵詞來表示該文件特徵。在技術上可利用統計方法去計算每個詞彙在文件中佔的權重，將權重較低的詞彙先剔除，留下的詞彙所形成的集合就能代表該文件之特徵。本研究希望能藉由此特徵集合，將資料群透過映射(Mapping)的方式將高維度的空間投影到低維度的空間，如此不僅能簡化資料分析時的計算，也能幫助我們更方便地去瞭解資料特徵間的關係。

最常見的特徵詞挑選方法為 TF-IDF(Term Frequency – Inverse Document Frequency)(Salton & Buckley, 1988)：

1. TF (詞彙頻率,Term Frequency)

在一份給定的文件中，TF 表示某一個特定詞彙在該文件中出現的次數，以代表其在文件中的重要性，而愈是重要的概念愈容易重覆出現在文件中，故在文件中出現頻率愈高的詞彙愈能代表文件所要表達的概念。

對於在某一特定文件裡的詞彙 i 來說，其重要性可表示為：

$$\mathrm{tf}_{i,j} = \frac{\mathrm{n}_{i,j}}{\sum_k \mathrm{n}_{k,j}} \quad \text{(式 2-1)}$$

其中，$\mathrm{n}_{i,j}$ 是詞彙 i 在文件 j 中的出現次數，而 $\sum_k \mathrm{n}_{k,j}$ 則是在文件 j 中所有詞彙的出現次數之總和。

2. IDF (反向文件頻率, Inverse Document Frequency)

而因為出現頻率較高的詞彙可能在每一篇文件中均會出現，則其所代表重要性便相對少於出現在較少文件內容中的詞彙，於是可透過 IDF 來修正此缺點。IDF 是一個詞彙普遍重要性的衡量標準。某一特定詞彙的 IDF，可以由總文件數除上含有該詞彙之文件數，再取對數來得出：

$$\mathrm{idf}_i = \log \frac{|N|}{df_i} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{(式 2-2)}$$

其中，$|N|$為整個文件集的文件數，而 df_i為詞彙 i 出現在整個文件集的文件數。接著，將上述兩值相乘可得：

$$\text{TF-IDF} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i = \frac{\mathrm{n}_{i,j}}{\sum_k \mathrm{n}_{k,j}} \times \log \frac{|N|}{df_i} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(式 2-3)}$$

相乘後所得的 TF-IDF 所代表的意義是詞彙在文件中的重要性是與其在文件中出現的次數成正比，但與其在所有文件集中出現的文件數成反比，原因在於若詞彙出現於其他文件的頻率愈高，則其所能代表本文件的識別力就愈低。

而為了避免因文件長度差異而影響文件集中各詞彙之權重比較，須將 TF-IDF 所算出的詞彙權重做正規化處理，方法為將權重除以文件向量中所有元素之權重平方和再開根號，公式如下：

$$\text{weight} = \frac{\mathrm{W}_t}{\sqrt{\sum_{i=1}^{\mathrm{T}} \mathrm{W}_{t_i}^2}} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(式 2-4)}$$

其中，weight 為某一特徵詞正規化後之權重， W_t為該特徵詞原始權重(即 TF-IDF)，T 為所有出現的詞彙總數。另外，文件與文件間的相似程度可使用相似度之衡量公式來加以計算。

(三) 向量空間模型

若要針對非結構化或半結構化的文件資料作處理，必須將文件資料轉化成可以用來比較、判斷的表示方式。其中最廣為運用的方法是由 Salton(1975)等人提出的向量空間模型(Vector Space Model, VSM)。

向量空間模型的主要概念是：在一個文件集中，每一個特徵詞即代表空間中的一個維度，而每個維度上的值則代表該特徵詞在文件中的重要程度，即為該維度的「權重」，其可透過文件的詞彙統計資料計算而得，而最常用的權重計算方式為前述之 TF-IDF 計算方法。由這些權重所組合而成的特徵向量(Feature Vector)

則代表在向量空間中的一篇文件。以圖 2-3 為例，在一個三維空間中，文件皆由三個不同特徵詞所組成，每個文件中特徵詞的權重皆不相同，在空間中的位置亦然不同。

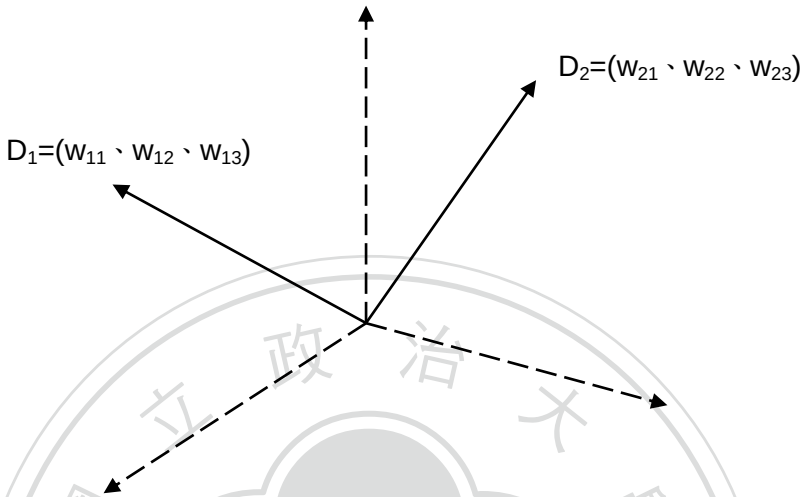

圖 2-3 向量空間模型示意圖

(資料來源：Salton et al., 1975)

若將上述此例子延伸到多維度，可以數學矩陣的方式表達及運算，如圖 2-4 所示：

$$\begin{array}{c} D_1 \\ D_2 \\ \cdots \\ \cdots \\ \cdots \\ D_i \end{array} \begin{bmatrix} w_{11} & w_{12} & \cdots & \cdots & \cdots & w_{1j} \\ w_{21} & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{i1} & \cdots & \cdots & \cdots & \cdots & w_{ij} \end{bmatrix}$$

圖 2-4 向量空間模型矩陣

(資料來源：本研究整理)

其中，D_i表示第 i 篇文件；w_{ij}表示第 j 個特徵詞在第 i 篇文件的權重值，即該矩陣是一個具有 i 篇文件與 j 個相異特徵詞的向量空間模型。

(四)　文件相似度計算

將文件以向量空間模型表達之後，文件與文件間相似的程度可透過相似度的計算以進行後續文件群集分析、分類等處理。常用的相似度計算方法為餘弦相似度(Cosine Similarity)計算法，其計算向量空間模型中兩文件所對應的向量之餘弦值，透過兩組相同基底(Base)與維度(Dimension)向量間的角度(Angle)差距來計算兩向量間的距離(You & Chen, 2006;Teng & Lee, 2007)。

其計算結果會介於 0 到 1 之間。當兩個向量間的角度差距愈小時，表示其餘弦角度愈小，餘弦值愈接近 1，即兩篇文件的相似程度愈高；反之，則相似程度愈低(陳崇正, 2009)。

如圖 2-5 所示，A、B 兩文件之向量間之餘弦相似度為 θ，而在 n 維空間的夾角公式如下：

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad \text{(式 2-5)}$$

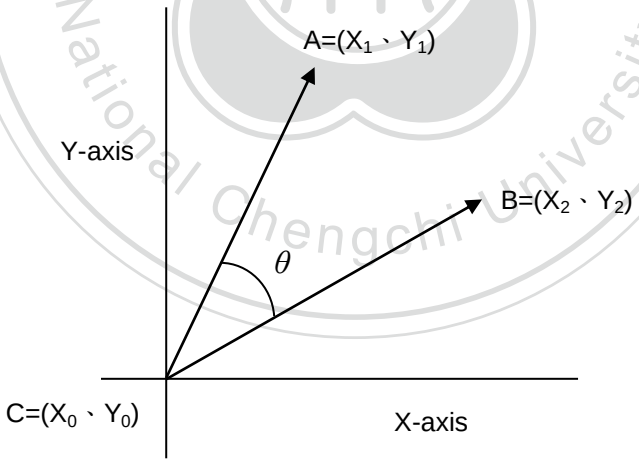

圖 2-5 二維空間中之餘弦相似度

(資料來源：陳崇正，2009)

2.2.4. 文字探勘運用於 App 推薦文章

當使用者面臨眾多選擇時，為了更瞭解其想要的產品或服務，通常會採用其他人的意見做為決策制定的參考依據，也就是所謂的口碑。而在網路的環境中，資訊的取得十分容易，網路上虛擬社群中網友的討論都是能作為自身參考的資訊來源(盧希鵬, 2005)，而這種在虛擬社群中對於產品及服務的討論正是所謂的網路口碑，亦稱為線上口碑。網路口碑的形式有很多種，像是論壇、部落格、聊天室以及電子郵件等等。

在 Engel 等人(1993)的研究中顯示：當消費者缺乏足夠的資訊而無法做出判斷，或是由於口碑來源的易得性，可節省使用者主動詢問產品的時間和心力等狀況下，消費者皆會傾向接受口碑資訊以做為參考依據；而在 Hennig-Thurau 等人(2004)的研究中指出：網路口碑會影響消費者的購買決策；加上有研究進一步發現：使用者進行遊戲的意願會受到他人的推薦或是看到他人進行遊戲而有所影響(林姿旻，2011)，故可知 App 的網路口碑能作為使用者選擇 App 時的參考依據並會影響其購買的行為。

目前以 App 為討論對象的網路口碑大多來自於網友在 App 相關網路論壇(例如：Mobile01)、電子佈告欄的討論區(例如：批踢踢的 iPhone 版)或是個人部落格上所張貼的 App 使用後之心得文章，而這些口碑文章的數量也隨著 App 的發展不斷地增加。又因為口碑文章屬於非結構化資料，無法透過常見的資料探勘方法了解資料的特性及隱含的意義，故本研究使用文字探勘技術來對口碑文章進行資料的分析與處理。

第三節、 群集分析

群集分析也稱為分群分析，在分析過程中能依照資料的相似程度，將相似性較高的文件群聚起來。該分析方法屬於非監督式(Unsupervised Learning)學習，即可將未知類別的資料從原先的集合，逐漸區分成個子集合，且能使每一個子群集內的資料具有高度的相似性，群集與群集之間則有高度的相異性。而群集分析技術已被廣泛的應用在各種領域上，像是統計、機器學習、迴歸分析、影像處理、及基因分析等。

2.3.1. 群集分析的種類

群集分析方法依性質可分為許多種類，常見的方法如：分割式群集分析(Partitioned)、階層式群集分析(Hierarchical)、密度式群集分析(Density-based Methods)等，以下將較常見的三種群集分析方法分述如下：

(一) 分割式群集分析

分割是群集分析為最早發展的分群技術，其分析前必須先決定所要分割的群集數目並挑選初始群集中心，再以重心點基礎（Centroid-based）或中心點基礎（Medoid-based）的方式進行群集分析，以使同一群集的資料能具有相似的屬性。其優點為處理過程簡單，且分群結果有一定的水準。但缺點是在進行分析前就必須決定出所要分割的群集數目並找出初始的群集中心；而在資料分布方面，也只能發掘簡易的資料點分佈形狀，無法有效率地針對高維度資料進行分析。

(二) 階層式群集分析

該群集分析方法是利用階層式的架構來產生資料分群的結果，其中又分為聚合式(Agglomerative) 以及分裂式(Divisive)兩種(Jain & Dubes, 1988)。主要的分群原理是：將相似度高的小群集合併，成為一個較大的群集，或者將較大的群集進行分離，拆解成多個彼此相似度低的小群集。所產生之階層結構，可以依據不同的使用者需求產生不同的群集數量。

(三) 密度為基礎群集分析

有別於以距離(相似度)為概念所發展出的分群方式，密度式群集分析方法是以資料的密度高低為依據，來判斷資料是否為同一群集。群集內部的資料密度應該大於外部的資料密度，亦即在同一群集內的資料密度高，而在群集外的資料密度低，其代表性演算法如：DBSCAN、DBCLASD 等。

2.3.2. 群集分析運用於文字探勘

應用群集分析方法的相關研究中，曾有學者將大量音樂文章作為資料分析對象，並以音樂文章中的和絃作為文章特徵，針對不同的特徵表示方法，提出相似度的計算方式。並藉由階層式分群等三種分群方法，將文章依據不同的音樂風格進行分群，分群的結果能降低使用者在瀏覽大量音樂資料時所需花費的時間(郭芳菲，2003)。

蕭文峰(2005)在垃圾郵件的過濾處理上，發現可透過群集分析技術將所有郵件資料集合劃分成許多群集，再找出每個群集的共同特徵來，之後便能在後續分類過濾時，讓新文件與每一群集先做比較，能夠更快地判斷出資料是否為垃圾郵件以增進分類效率。

在推薦系統的應用方面，亦有研究以唐詩詩詞為對象，希望透過文字探勘的技術來分析作者當初的創作心境及感受，並將相似程度高的詩作聚為一群。之後再根據詩作屬性建立分類模型，並依使用者所設定之心境條件，計算推薦度以推薦詩詞作品(楊智凱，2007)。

而在群集分析中使用 kNN 來達到文字探勘目的研究中，陳柏均(2011)將新聞作為資料分析對象，並以 kNN 為基礎發展出一套新的群集分析方法─RTD-based kNN，透過在向量空間中建立一個基準點，讓所有新聞文件利用與基準點的相對距離建立起遠近的關係，使得在選取前 k 個最近鄰居之前，能直接以相對關係篩選出較可能的候選文件，進而選出前 k 個最近鄰新聞文件，以減少傳統 kNN 分群所需花費的運算時間。

由相關研究可知，透過文字探勘的技術進行資料分群之實作，能依據資料的特性將大量且未知的資料集合細分為具有個別特徵的集合，進而節省使用者的搜尋成本並增加效率。以本研究而言，除了能使 App 各自成群外，還可藉由分析、觀察資料的分布情形、群集的大小及特性等，找出對使用者有用的資訊，以解決資訊超載的問題。

第四節、 k-最鄰近演算法(k-Nearest Neighbor , kNN)

k-最鄰近演算法(k-Nearest Neighbor)是在 1967 年由學者 T.M. Cover 與 P.E. Hart 所提出的，屬於基本常用之分類分析方法。而 kNN 演算法雖然經常被歸類於分類演算法中，但在實作上亦可不用事先得知訓練資料的類別，而可透過群集分析後所得到的資料特性，再對各群集進行描述，例如：Yang 等人(1999)就將 kNN 運用在「類別數未知」的新聞事件的偵測追蹤，即為一種 kNN 在群集分析上的應用。與其他群集分析方法相較之下，kNN 之建置較容易，並具有學習速度快，分群效果佳等特性。

kNN 在分群時，採用向量空間模型來進行，即在對文件進行分群前必須將文件轉換成向量空間模型，再藉由計算新進文件與其他已分群文件間的相似度，並擷取出與新進文件最相似的 k 篇已分群文件來判斷新進文件的所屬群集。判斷方式為將擷取出的 k 篇文件中，相同群集內的所有文件與新進文件的相似度加總並除以文件個數，結果數值最高的群集即為新進文件所屬的群集；而文件的相似度一般則採用 Cosine 相似度來加以計算。

以圖 2-6 為例，每個資料點的屬性即為向量空間中的維度，假設有兩屬性 A、B 與兩個已分群之群集 X 及 Y，未知樣本「S」則為新進文件。此時，kNN 會先擷取出與新進文件最相似的 k 篇文件。假設 k=3 (即擷取相似度前 3 名最相似的文件)，便從所擷取出的 3 個已分群文件中，來判定新進文件到底該歸在 X 群集或 Y 群集。假設 S 與 X 群集的相似程度為 0.5，而 S 與 Y 群集的相似程度為 0.6，那麼我們即可認定新進文件 S 所屬的群集為 Y 群集。

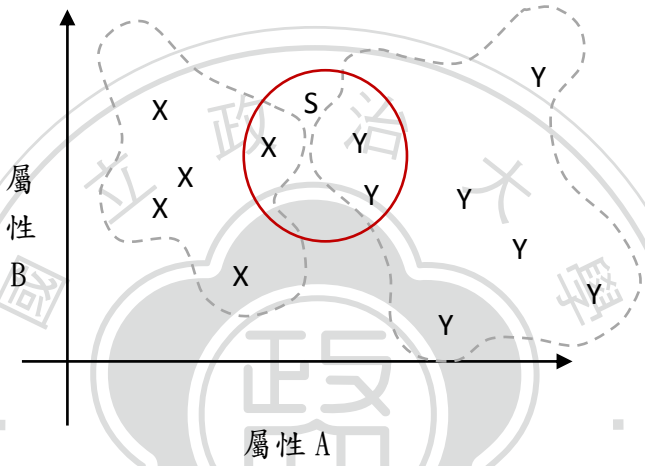

圖 2-6 kNN 群集分析示意圖

(來源：本研究整理)

本研究將利用 kNN 群集分析法，針對已轉換成向量空間模型的各篇 App 文章進行相似度的比較，將相似度較高的 App 文章聚集成群，以達成分群的目的；並透過調整 kNN 演算法中的參數：k 值與文件相似度門檻值，以及分群品質之衡量，來挑選出品質較佳的分群結果。

第三章、研究方法與設計

本研究將採用文字探勘及 kNN 群集分析技術針對網友所發表的 App 推薦文章進行 App 之分群，並透過分群品質衡量指標來找出不同參數組合之下分群效果最佳之分群結果。最後，對分群結果進行命名。

第一節、 研究架構

本研究一開始先蒐集網友於網路上發表與 App 相關之推薦文章，將推薦文內容進行文章斷詞後存入資料庫。再透過文件特徵選取中的詞彙權重計算方式，計算出文章中各詞彙之權重，並將文件轉換為向量空間模型；接著，透過向量空間模型的比較並計算出各 App 文章之相似度後，便進行 App 之分群與群集合併動作。並藉由群集分析之衡量指標對不同 k 值及文件相似度門檻值組合之下的分群結果進行評估，以獲得最佳之分群結果。最後，萃取能代表各群的關鍵詞彙來對分群結果進行群集命名。

本研究之研究架構圖如圖 3-1 所示：

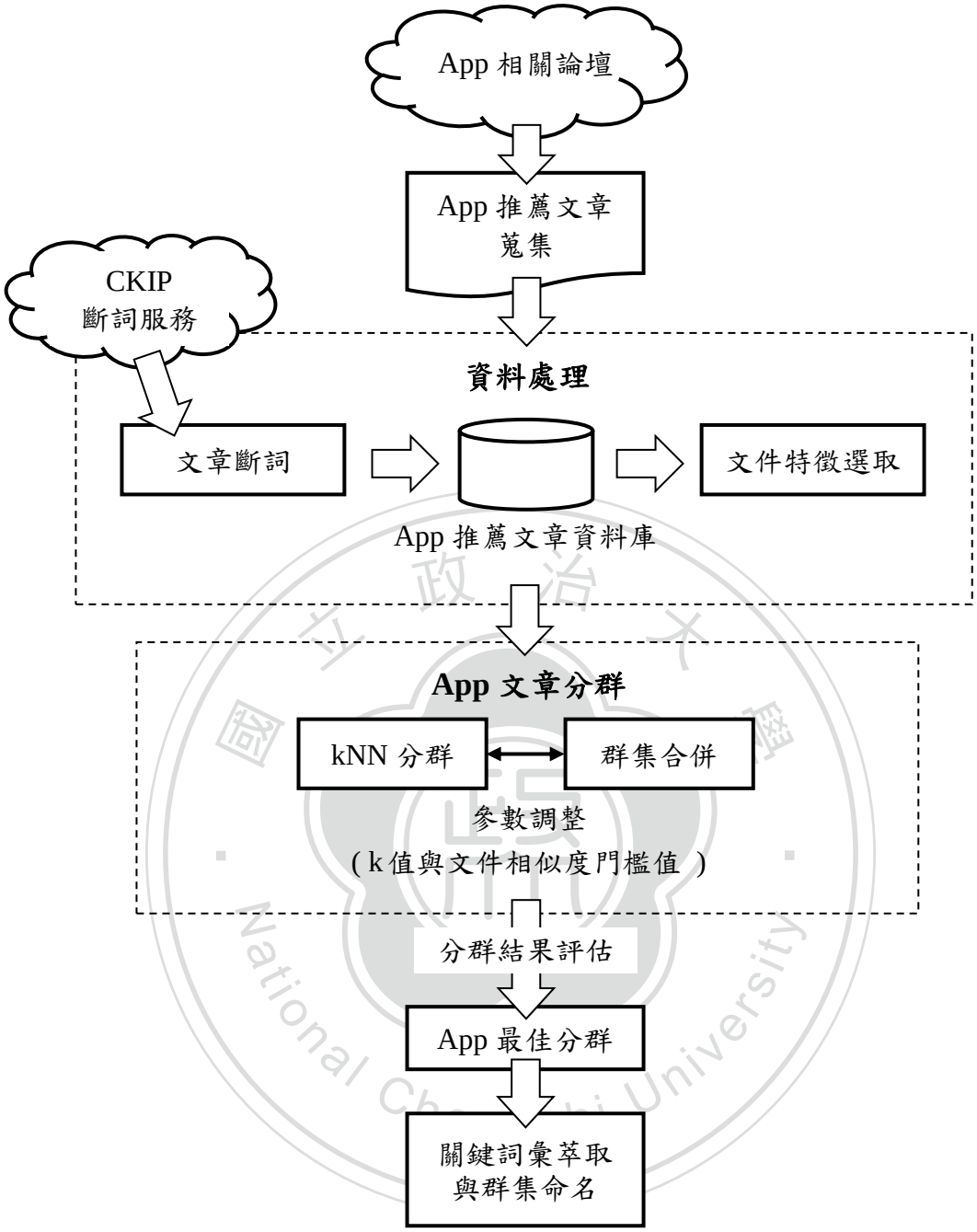
App 相關論壇
App 推薦文章
蒐集
CKIP
斷詞服務
資料處理
文章斷詞
文件特徵選取
App 推薦文章資料庫
App 文章分群
kNN 分群
群集合併
參數調整
(k 值與文件相似度門檻值)
分群結果評估
App 最佳分群
關鍵詞彙萃取
與群集命名

圖 3-1 研究架構圖

第二節、資料來源與處理

3.2.1. 資料來源

如前所述，因 Apple 的 App Store 提供使用者較大量的 App，而目前網路上也有許多針對 App 所架設的論壇及網站，提供網友在使用 App 之後能夠藉由發表文章的方式來推薦自己喜愛的 App。

故本研究便以 App Store 的 App 為主，並在 App Store 所設定的 22 項主要類別中，選擇遊戲(Games)類別的 App 為蒐集對象；而資料蒐集來源則選擇設有以使用者發表 App 相關推薦文為主討論區塊之論壇，因 App 相關論壇並無統一網頁格式，故需以人工方式蒐集文章。而論壇選擇方式為：於 Google 搜尋網頁中輸入關鍵字：「App Store、推薦、遊戲、論壇」後，於搜尋結果中選取符合本研究需求之 App 相關論壇；又仔細觀察搜尋結果排列越後面之項目，與所欲搜尋的相關論壇越無關聯，故只採用前 50 筆搜尋結果，並挑選出前三個 App 相關論壇，分別為：App01(www.app01.com.tw)、iPhone4.TW(iphone4.tw/forums/forum.php)與 Mobile01(www.mobile01.com)。接著，從中蒐集與 Apple Store 所提供之遊戲類型 App 相關之推薦文章共 439 篇；並將推薦對象相同之 App 推薦文章整合於同一份文章中。最後共整理出 357 篇文章，每一篇文章代表一種 App 遊戲。

本研究所蒐集的 App 推薦文範例如表 3-1：

表 3-1 App 推薦文範例

標題：【新發現。重要經驗分享 173 樓】CEO LIFE 私密功略分享及簡介	
發布者：cheris168	發布時間：2011-08-15 10:16
CEO LIFE 這款超有趣模擬經營的遊戲今年 2 月 11 日新鮮上架的職場 CEO 模擬軟體, 短短半年在日本與對岸吸引了數十萬的瘋狂玩家, 小弟從上星期五(8/12)也加入了這個行列, 感覺蠻有趣的, 可以邀請朋友一起加入、合夥做生意、建立子公司、培訓下屬、結交好友、開拓業務...若未來您真有自己創業的計劃, 也不妨先在這遊戲中經驗一下當 CEO 的滋味。在遊戲的當中隨時也可以與真實的朋友互通訊息, 也可以和陌生的外國朋友聊天, 互相分享心得或真實世界的生意往來。遊戲一開始會先給您一位漂亮的秘書, 當然漂亮的定義因每人觀感不同, 所以選擇也不同了... 為了讓公司強大, 首先要陪養優秀的人材, 部下的人力值分為健壯、才智、審美與外交 4 種; 然後根據部下的能力選擇職業, 建立子公司...根據子公司的職業特徵, 每隔一小時會得到相應的效益哦~ 公司愈大效益愈多, 您就有錢買鑽石或名牌包包囉~來吧~讓我們一起成為世界首富加油吧~	

(資料來源：ttp://www.mobile01.com/topicdetail.php?f=627&t=2302566&last=30656671)

3.2.2. 文章斷詞

在論壇中所蒐集到的 App 推薦文中，許多網友會透過圖片的方式來解說遊戲的玩法以及分享戰績，而這些圖片在其後是無法被進行分析的，所以在存入資料庫前，我們必須將無法透過文字探勘進行分析的圖片以及作者資訊等非主文的描述過濾掉，以便進行後續的文字探勘分析。

又在前述文獻探討中我們曾提到：因為文件是以非結構化或半結構化的方式呈現，所以為了讓推薦文章像一般資料庫中的結構化資料一樣，方便我們分析其內容資訊，就必須先將文章進行斷詞，轉換成結構化的資料。

而本研究所使用的斷詞工具是中研院詞庫小組所開發的 CKIP 中文斷詞系統，因文章透過該系統進行斷詞後，回傳的結果會針對所有詞彙加上相符的詞性且沒有每日斷詞次數之限制，故能符合本研究之需求。以下為 CKIP 系統斷詞處理前、後對照之範例：

斷詞處理前：

CEO 模擬軟體, 短短半年在日本與對岸吸引了數十萬的瘋狂玩家！

斷詞處理後：

CEO(FW) 模擬(Nv) 軟體(N) ,(COMMACATEGORY) 短短(Vi) 半年(N) 在(P) 日本(N) 與(C) 對岸(N) 吸引(Vt) 了(ASP) 數十萬(DET) 的(T) 瘋狂(Vi) 玩家(N) ！(EXCLAMATIONCATEGORY)

也因為我們可以得知文章斷詞後各詞彙的詞性，如此一來，當未來我們需要刪減文件特徵時，也能透過判斷詞彙詞性的方式，擷取出較具代表性的特徵詞彙，並刪除較不具有實質意義的詞性之詞彙；以本研究所考慮的詞性方面，較能代表使用者感受且具有實質意義的詞性為：動詞、名詞與形容詞，故本研究將過濾掉其他詞性，以減少後續運算過程，提升執行效率。

3.2.3. 文件特徵選取

在進行文件相似度計算前，需將文件轉換成向量空間模型表示，因此本研究能藉由 TF-IDF 詞彙權重計算公式，將斷詞後的各個詞彙在文章中所佔的權重值計算出來，該法考慮到詞彙在各篇文章中以及在所有文章的普遍性，並可透過正規化公式來避免各篇文章因為長度不同而造成的權重值差異問題。

由於文件中的每一個詞彙都是組成該文件的特徵，因此，選擇愈多的詞彙作為文件特徵，即愈能代表文件本身。但過多的文件特徵常會造成文件向量空間模型的維度太高，進而使得分群時的運算量過於龐大；故本研究針對每篇推薦文章計算完文章中所有詞彙之權重後，僅以詞彙權重前 80%的詞彙作為文章之關鍵詞彙，以使各文章所含之詞彙更具特徵意義並減少運算量及刪去較不重要之詞彙。

第三節、 App 文章分群

在 App 文章分群部分，首先會利用 kNN 演算法對 App 文章進行分群，接著透過群集合併來改善分群品質；最後，利用參數調整來對分群結果進行評估，以找出最佳品質之分群結果。

3.3.1. kNN 分群

在進行 App 分群時，本研究所採取的技術是 k-最鄰近演算法(kNN)演算法，該法是將文件以向量空間模型來表示，再藉由計算與已分群文件的相似度來判斷出欲分群文件可能所屬的群集。而相似度的計算是採用 cosine 相似度來加以計算。分群的步驟如下：

1. 首先，將新進的 App 推薦文章轉換為向量空間模型。
2. 接著，將新進 App 推薦文章與各個已分群之 App 推薦文章集合內之所有文章進行相似度的計算，取出前 k 份最相似的推薦文章。
3. 將這 k 份推薦文章所屬的所有群集皆列為新進推薦文的候選群集。

4. 將這 k 份推薦文章與新進推薦文章進行的所屬群集之判斷：將擷取出的 k 篇文章中，相同群集內的所有文章與新進文章的相似度加總並除以該群集所包含的文章數，計算公式如下：

$$\mathrm{P}(x, C_j) = \frac{1}{N_j} * \sum_{x_i \in KNN} Sim(x, x_i) y(x_i, C_j) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(式 3-1)}$$

其中，x為新進文章之特徵向量；$Sim(x, x_i)$為相似度計算公式；而 $y(x_i, C_j)$為類別屬性函數，即若x_i屬於群集C_j則函數值為 1，否則為 0；N_j則為第 j 群所含的文章數量。計算出新進文章與各群集之相似度後進行比較，數值最大的群集則為新進文章的所屬群集。

而在每次分群後，可能會產生部分群集所包含的文章數量過於龐大，使得群集分析品質未達到最佳，此時會針對包含文章數較多的群集再度進行分群，並透過衡量指標來評估再次分群之必要性。

3.3.2. 群集合併

進行分群時，我們所設定的 k 值及文件相似度門檻值往往會直接的影響到分群的結果，例如：文件的相似度門檻設定過高，可能會造成某一群集內只含有一份文件或是將本來應該分在同一群的文件集合拆分成兩個小群集，而使得分群品質降低。因此，我們可計算出各群集的質心，即各群集的中心點，再利用各群集質心間相似度的計算來進行群集合併，以改善 k 值及文件相似度門檻值所造成的影響。

質心計算(吳文峰，2002)公式為：

$$\overrightarrow{C_i} = \frac{1}{n_i} \sum_{d \in C_i} \vec{d} \cdots\cdots\cdots \text{(式 3-2)}$$

其中，$\vec{d}$為文件向量，$\vec{d} = (d^{(1)}, d^{(2)}, \ldots, d^{(|\vec{d}|)})$，$d^{(j)}$表示第 j 個詞彙在文件 d 中的權重，$|\vec{d}|$則為文章長度；n_i為群集C_i中的文件數。而為了使得計算後之質心的維度權重有相同的衡量標準，不受到群集文件數量的影響，故在質心計算完畢後必須進行正規化處理，以便後續可透過質心來進行群集間相似度的比較及群集合併之進行。

3.3.3. 參數調整

在 kNN 群集分析方法中，設定不同的 k 值與不同的文件相似度門檻值會得到不同的分群結果，其分群品質也不盡相同。

(一)　k 值

k 值為與欲分群文件最相似的已分群文件數量，若 k 值取得過大，這 k 個最相似的鄰居中可能會包含許多相似度並不高的已分群文件；若 k 值取得過小，那麼就有可能使得欲分群文件受到雜訊資料的影響，皆會影響到分群的品質。

(二)　文件相似度門檻值

文件相似度門檻值是指在篩選 k 個最相似的已分群文件時，相似度要超過此門檻值才能被納入候選文件集合中。門檻值設定的大小除了影響分群之後的群集內所含文章多寡，也會影響到分群的品質

本研究將會設定不同的 k 值與文件相似度門檻值組合成多種參數組合，並透過分群品質的衡量指標來選擇分群品質最佳的參數組合。

3.3.4. 分群結果評估

為了得知在不同參數組合之下，進行 kNN 分群與群集合併後分群結果之品質，本研究將利用平均群內相似度及平均群間相似度計算出分群品質的衡量指標，來判斷何種參數組合為最佳之分群。

(一) 平均群內相似度(Mean of Intra-cluster Similarity)

平均群內相似度是將每一群集內的文件，兩兩比較後將相似度加總除以比較次數以獲得各群之群內相似度。並採用加權概念，將各群計算完成之群內相似度乘上各群所含之文件數佔所有文件數的比例，即可獲得平均群內相似度。其值介於 0 到 1 之間，當平均群內相似度愈接近 1 代表群內的相似度愈高，其公式為：

$$\text{平均群內相似度} = \sum_{C_k} \frac{\sum_{d_i \in C_k} \sum_{d_j \in C_k} sim(d_i, d_j)}{N_k * (N_k - 1) * \frac{1}{2}} * \frac{N_k}{N} \quad \text{(式 3-3)}$$

其中，N為文件總數；N_k為第C_k 群之文件數量；$\frac{N_k}{N}$為第C_k群之加權值；$sim(d_i, d_j)$則是C_k群內某兩篇文件之相似度。

(二) 平均群間相似度(Mean of Inter-cluster Similarity)

而平均群間相似度則是將各群集所計算出的質心，兩兩比較後將相似度加總並除以質心比較次數而得，其公式為：

$$\text{平均群間相似度} = \frac{\sum_{C_i \in C} \sum_{C_j \in C} sim(C_i, C_j)}{C * (C-1) * \frac{1}{2}} \quad \text{(式 3-4)}$$

其中，C為群集數目；$sim(C_i, C_j)$為某兩群集質心之相似度。

(三) 分群品質(Cluster Quality)

群集分析中改善分群品質的重點在於：提升群內相似度，並降低群間相似度。故欲判斷分群品質是否良好，便可透過下列公式來進行衡量(Lai & Liu, 2009)：

$$\text{分群品質 (CQ)} = \frac{\text{平均群內相似度}}{\text{平均群間相似度}} \quad \text{(式 3-5)}$$

最後，當 App 分群結果產生後，即可同樣透過 TF-IDF 的計算萃取出代表各群集的關鍵詞彙，並對各群集進行群集命名。

3.3.5. 分群規則

(一) 多階段分群

本研究在初步分群過程中發現，若只進行一次分群會產生群集內文章數量過大或過小的現象，可能造成隱含的重要特徵過多或不足，無法進行後續分析；故為了使得分群結果隱含適切的分析資訊，並使分群品質達到最佳，本研究採用了「多階段分群」來解決此問題，意即當分群結束時，將針對文章數量較多的「大群集」進行再分群。在分群的過程中，若出現只包含了 1 篇文章的「小群集」時，因其無任何分群上的實質意義，故會透過群集間質心相似度的計算，將其合併到該分群階段中相似度最高的群集。

(二) 參數設定

本研究在進行分群時，參數的調整將以 k 值為主，每一個固定的 k 值將搭配 3 種文件相似度門檻值以組合成多種參數組合，並在多種參數組合下之分群結果中，挑選最佳品質之分群以作為各階段之最佳分群結果。

1. k 值

在第一階段分群時，將所有文章當作一個大群集，因群集內所含文章數較多，故 k 值將以 10 為起始值，調整幅度為每次向上調整 10，至多調整至 30；而在第二階段分群以後，因已經過了一次分群處理，各群集內所含文章數下降，故 k 值將以 5 為起始值，調整幅度為每次向上調整 5，至多調整 15。

2. 文件相似度門檻值

本研究針對文件相似度門檻值的設定，在第一階段分群時，將以尚未分群前之平均群內相似度作為參考標準；後續階段則依前一階段之最佳分群結果之文件相似度門檻值為基準，增加 0.005 為該階段之文件相似度門檻值；而在每一個固定的 k 值下，將設定三種不同的文件相似度門檻值，其調整幅度為 0.005，藉此觀察多種組合下之分群品質何者為最佳。

而當調整後之 k 值搭配 3 種文件相似度門檻值所獲得的 3 種分群結果，相較於調整前之 k 值的分群結果而言，無法獲得更好之分群品質，便以調整前之 k 值所獲得之 3 種分群結果中，挑選分群品質較佳的參數組合為該群集之最佳分群結果，並不再對 k 值進行向上調整動作；反之，將繼續向上調整 k 值，並觀察後續參數組合，以選出最佳品質之分群。

茲以下例說明：在第一階段分群時，k 值之起使值為 10，並搭配 3 種文件相似度門檻值可得到 3 種不同品質的分群結果(CQ_{11}、CQ_{12} 與 CQ_{13})，其中，CQ_{13} 之分群品質為三者中最佳。接著，將 k 值調整至 15，亦搭配同樣的 3 種文件相似度門檻值來得到 3 種不同品質的分群結果(CQ_{21}、CQ_{22} 與 CQ_{23})；此時，若 k 值調整後所得到的：CQ_{21}、CQ_{22} 與 CQ_{23} 之分群品質皆低於 k 值調整前之最佳分群品質：CQ_{13}，即將 CQ_{13} 視為該群集之最佳分群結果，並不繼續計算當 k 值調整至 20 時之分群結果；反之，當 CQ_{21}、CQ_{22} 與 CQ_{23} 其中之一的分群品質高於 CQ_{13}，即將 k 值再向上調整至 20，以繼續比較不同參數組合下之分群品質。

(三) 分群停止條件

本研究所設定之分群停止條件有二：

1. 首先，會透過群集內所包含的文章數量來判斷該群集是否應繼續分群，判斷標準為：將群集內含文章篇數超過 30 篇之群集判定為「大群集」，並對其繼續分群，反之，則停止分群。
2. 接著，為了避免以群集內含文章數作為繼續分群的標準，而使得已經非常相似之群集被拆分成更細的小群集，故本研究在參數調整部分亦設定了分群停止條件：即當針對某一群集繼續分群時，若在起始之 k 值與其搭配的 3 種文件相似度門檻值中，皆無法獲得分群效果，例如：文件相似度門檻值的調整對該群集無法產生分群效果，或分群後只產生許多內含 1 篇文章的小群集，合併後結果與為未分群相同時，即不再對該群集進行再分群。

第四章、研究結果

本章將透過前述之研究方法進行 App 分群及群集命名，包含各階段之分群參數調整、最佳分群品質評估、各群集關鍵詞彙萃取及定義各群集名稱等，以獲得最佳品質之 App 分群結果，作為使用者選擇 App 之參考依據。

第一節、 第一階段分群

當所有文章未分群，即將 357 篇 App 文章視為一整個大群集時，該群集編號設定為 0，其所測得之平均群內相似度約為 0.01629，以該值為第一階段分群之文件相似度門檻值的設定參考基準，故以 0.015 為起始值，並提升 0.005 及 0.01(3 種文件相似度門檻值分別為 0.015、0.02 與 0.025)；而 k 值之起使值設定為 10，因 k 值向上調整至 20 時無法獲得更好的分群品質，故不須繼續將 k 值向上調整，整理結果如表 4-1。

表 4-1 第一階段分群評估結果

群集編號		**(0)**			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
10	0.015	5	0.02166483	0.23991256	0.09030304
	0.02	5	0.02175872	0.24359091	0.08932483
	0.025	7	0.02392413	0.18915519	0.12647883
20	0.015	5	0.02142502	0.28727206	0.07458093
	0.02	5	0.02127168	0.29620999	0.07181285
	0.025	7	0.02442257	0.20933309	0.11666848

以分群品質衡量指標可知：當在 k 為 10、文件相似度門檻值為 0.025 時能獲得最佳分群，故依該分群結果為第一階段最佳分群(圖 4-1)，共分成 7 群。

由圖 4-1 可知，群集(2)、(6)與(7)因所包含的文章數符合停止條件(文章數小於或等於 30 篇)，故不需繼續往下分群；而群集(1)、(3)、(4)與(5)則須繼續進行第二階段分群。

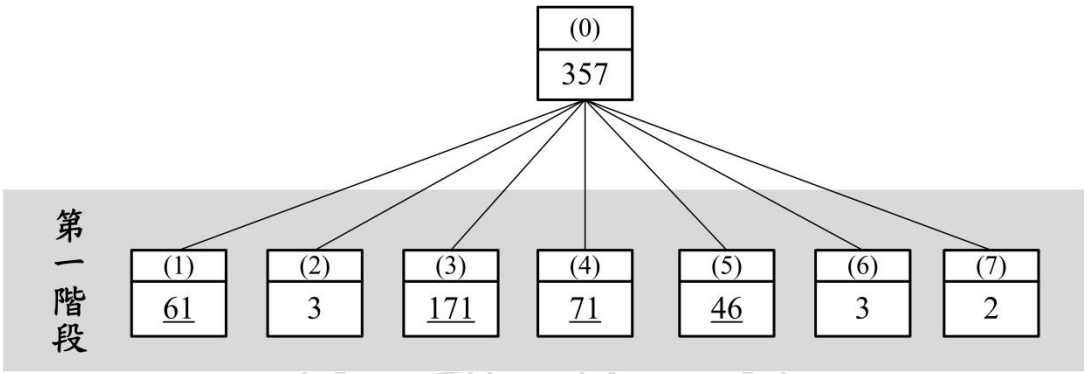

圖 4-1 第一階段分群結果

第一階段分群結束後，平均群內相似度由未分群時的 0.01629 提升至 0.02392、平均群間相似度為 0.18916；可計算其分群品質為 0.12648。

第二節、 第二階段分群

在第一階段分群後，因各群集內所包含的文章數大幅降低，若使用原始的 k 值作為起始值，在選取候選群集的 k 個最近鄰居時可能會包含過多的雜訊，故將 k 值之起使值從 10 降低為 5；而文件相似度門檻值則以第一階段所挑選之最佳分群參數組合中的文件相似度門檻值 0.025 為基準，增加 0.005、0.01 及 0.015(3 種文件相似度門檻值分別為 0.03、0.035 與 0.04)以作為第二階段之文件相似度門檻值。同樣地，將透過分群品質之衡量指標來選擇最佳分群結果，作為該階段各群集之最終分群結果；第二階段之分群評估結果整理如表 4-2。

第二階段分群評估中，群集(1)因 k 值向上調整至 10 之後，所獲得之分群品質與 k 值為 5 時相同，並無改善，故不繼續向上調整 k 值；並可知在 k 值為 5、文件相似度門檻值為 0.03 時能獲得最佳分群品質。群集(3)則在 k 值向上調整至 10 時，能夠獲得相較於 k 值為 5 時更好的分群品質，故繼續將 k 值調整為 15；但 k 值調整至 15 後所測得的 3 種分群結果之品質皆無法比 k 值為 10 時之最佳分

群品質(0.36930463)更好，故不繼續向上調整 k 值;並可知在 k 值為 10、文件相似度門檻值為 0.04 時，能獲得最佳之分群品質。同樣地，群集(4)在 k 值為 5、文件相似度門檻值為 0.04 時；群集(5)在 k 值為 5、文件相似度門檻值為 0.04 時，分別能獲得其最佳之分群品質。

表 4-2 第二階段分群評估結果

群集編號		**(1)**			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.03	2	0.030001842	0.08621828	0.347975416
	0.035	2	0.03000184	0.08621828	0.347975416
	0.04	4	0.04267441	0.13038648	0.32729165
10	0.03	2	0.030001842	0.08621828	0.347975416
	0.035	2	0.03000184	0.08621828	0.347975416
	0.04	4	0.04267441	0.13038648	0.32729165
群集編號		**(3)**			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.03	2	0.018623459	0.26150186	0.07121731
	0.035	6	0.026041642	0.12755137	0.20416591
	0.04	8	0.030164721	0.08610858	0.35031026
10	0.03	2	0.01876271	0.33230411	0.05646247
	0.035	7	0.02947889	0.12431713	0.23712651
	0.04	9	0.03338075	0.09038812	0.36930463
15	0.03	2	0.01961412	0.42646810	0.04599200
	0.035	7	0.03000169	0.13306411	0.22546793
	0.04	9	0.03372591	0.09495431	0.35518042

表 4-2 第二階段分群評估結果(續)

群集編號		(4)			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.03	無分群效果			
	0.035	2	0.02210200	0.07648028	0.28898959
	0.04	3	0.02377145	0.07502992	0.31682627
10	0.03	無分群效果			
	0.035	2	0.02210200	0.07648028	0.28898959
	0.04	3	0.02377145	0.07502992	0.31682627
群集編號		(5)			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.03	無分群效果			
	0.035	2	0.03075425	0.17153469	0.179288782
	0.04	2	0.03086379	0.17081908	0.180681181
10	0.03	無分群效果			
	0.035	2	0.03075425	0.17153469	0.179288782
	0.04	2	0.03086379	0.17081908	0.180681181

而第二階段分群結果如圖 4-2 所示；該階段分群後，群集(9)、(16)與(20)未達到分群停止條件，故須繼續分群，而已達停止條件的群集將會進行關鍵詞彙之萃取與群集命名。

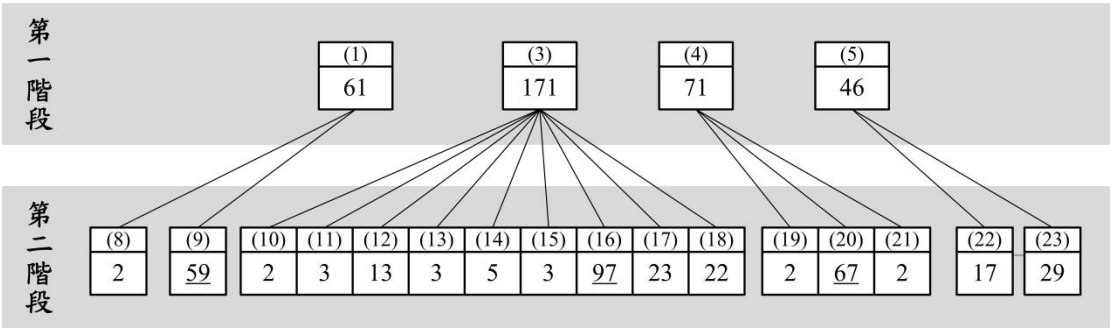

圖 4-2 第二階段分群結果

而當群集內包含的文章數量過少，可能會由少數一、兩篇文章來決定該群集的 App 類型，較無實質意義，且佔整體 App 總數量比例較低，故只對文章篇數大於 3 篇的群集進行關鍵詞彙的萃取與群集命名；以第二階段分群結果而言，群集(12)、(14)、(17)、(18)、(22)與(23)皆不需繼續分群，且所含文章篇數皆超過 3 篇，故可對其進行關鍵詞彙萃取。

關鍵詞彙的萃取方式與文件特徵選取方式類似，即將每一群集皆視為一篇大文章，藉由 TF-IDF 的方式可以計算出在各群中所出現詞彙相對於該群而言之權重。接著再透過詞彙權重的高低排列，從各群集內詞彙權重排名前 20%的詞彙集合中萃取出各群集之關鍵詞彙，並將這些關鍵詞彙轉化較適切的名稱以代表各群集。又因部分詞彙之詞性較不具特殊意義，故在此主要以詞性為名詞、動詞與形容詞的詞彙來進行關鍵詞彙之萃取。

第二階段分群結束後，群集之關鍵詞彙萃取與群集命名結果如表 4-3 所示：

表 4-3 關鍵詞彙萃取與群集命名—第二階段

群集編號	**(12)**
文章數量	13
群集命名	棒球/射擊
關鍵詞彙 (詞彙權重值)	全壘打(0.2773)、打擊(0.1183)、揮棒(0.086)、 隊友(0.0853)、投出(0.0853)、棒球(0.0785)、 贏球(0.064)、球棒(0.064)
	戰役(0.1528)、子彈(0.1493)、射擊(0.0918)、 槍械(0.0853)、武器(0.0665)、血量(0.0537)、 武裝(0.0516)、裝甲(0.0516)
所含 App 項目	Adv. Hanoi、BATTLEFIELD:BAD COMPANY 2、CrazyTouch、Deadlock:Online、Emily's Wardrobe、Flick Home Run、Homerun Battle、Modern Combat 3: Fallen Nation、OthelloCN、Robber Rabbits!、Through the Cliff hd、XiangQi by Topoc、洛克人 X
群集編號	**(14)**
文章數量	5
群集命名	投擲飛行
關鍵詞彙 (詞彙權重值)	小鳥(0.7202)、星球(0.1497)、太空(0.0679)、 引力(0.0658)、翅膀(0.0658)、拋物線(0.0658)、 終點(0.0583)、降落(0.05)、飛行(0.0449)、發射(0.0421)
所含 App 項目	Angry Birds Space、Birzzle Pandora、Cosmonauts、Early Bird、Tiny Wings

表 4-3 關鍵詞彙萃取與群集命名—第二階段(續)

群集編號	**(17)**
文章數量	23
群集命名	商家經營
關鍵詞彙 (詞彙權重值)	客人(0.3509)、餐廳(0.2271)、美髮(0.1536)、 顧客(0.1445)、經營(0.1334)、老闆(0.0976)、 上門(0.064)、服務(0.0619)、商舖(0.0512)、 隊伍(0.0471)、烹飪(0.0384)、裝潢(0.0384)、 餐點(0.0384)、美食(0.031)、髮型(0.031)
所含 App 項目	Army of Darkness Defense、Bonnie’s Brunch、Burger Queen World、Cup Puppy、DevilDark: The Fallen Kingdom、Dream Park、FIFA 12 by EA SPORTS、Fight Night Champion、Flying PuPu、Gesundheit!、iOOTP Baseball 2012 Edition、Mega Bad、Monster Pet Shop、Monsterz Revenge、My Town 2、Paco Mania、Pucca's Restaurant、Slime vs. Mushroom2、Sneak Out、SushiGoRound、Toilet War、美髮玩家 Salon Boss、開心豬仔
群集編號	**(18)**
文章數量	22
群集命名	經典桌遊
關鍵詞彙 (詞彙權重值)	大富翁(0.2718)、夾娃娃機(0.1627)、夜市(0.1312)、 卡片(0.1122)、地圖(0.1071)、股票(0.1046)、 任務(0.0543)、破產(0.0465)、獎品(0.0375)、 機台(0.0356)、道具(0.0335)
所含 App 項目	3D 可愛夾娃娃機、AHa Here、Bloons TD 4、Cuts the Buttons、Draw Slasher: Dark Ninja vs Pirate Monkey Zomb、Exitium: Saviors of Vardonia、Feed Me Oil、Gangster Rio: City of Saints、Hey Snowman、Lil' Pirates、My Horse、My Little Hero、PirateGunner、Richman 4 fun、Road Warrior Multiplayer Racing - by Top、SnuggleTruck、Subcat、The Adventures of Tintin - The Secret of the Unico、Treasure Story、夜市大亨、大家來搶錢、好鳥賣照

表 4-3 關鍵詞彙萃取與群集命名—第二階段(續)

群集編號	**(22)**
文章數量	17
群集命名	推幣積分/策略攻防
關鍵詞彙 (詞彙權重值)	推落(0.3082)、代幣(0.2201)、骰子(0.1509)、 擲骰(0.1321)、推幣機(0.0881)、推幣類(0.066)、 規則(0.0645)、累積到(0.0611)、中獎(0.0533)
	兵種(0.161)、調兵(0.0881)、裝備(0.081)、 騎士(0.0763)、兵力(0.071)、建物(0.066)、 部落(0.0533)、弓手(0.0533)、法師(0.0485)、 資源(0.0462)
所含 App 項目	Ascension: Chronicle of the Godslayer、Caligo Chaser、Caylus、coin dozer、coin factory UFO Free、coin pirates、Defense of Fortune HD、Delve :The Dice Game、Dungeon Defenders: First Wave、dungeons and coin、Elder Sign: Omens、I can read your mind!、Lock N Rock、RISK: The Official Game for iPad、Roll Through the Ages、Toy Defense、部落防衛戰
群集編號	**(23)**
文章數量	29
群集命名	音樂節奏/跳躍動作
關鍵詞彙 (詞彙權重值)	歌曲(0.3592)、節拍(0.0867)、節奏(0.0664)、 音符(0.0619)、音樂(0.0527)、好聽(0.038)、 喇叭(0.0372)、曲目(0.0372)、音感(0.0372)
	跳(0.0634)、跳舞機(0.0495)、旋轉(0.0434)、 彈跳(0.04)、跳繩(0.04)、高度(0.0372)、跳越(0.0372)、 絆倒(0.0372)、跳躍(0.0342)
所含 App 項目	7 Wonders 2 HD (Full)、Age of Barbarians、Angry Chickens Pro、Azkend、Blokus、Bumpy Road、Burnout TM CRASH!、Cytus、FANTASYxRUNNERS、Food Bar Slot、Fruit Ninja: Puss in Boots、FruitPinBall、GOGO!Biker!、HAMMER KINGDOM!、iSlash HD、JumpZ、MadMaks、Michael Jackson The Experience HD、ninja chicken、osu!stream、Scarecrow、SDBALL、Skipping NYAN-P、SPY mouse、To-Fu: The Trials of Chi、Tofu Go!、Touch Racing、勁舞甜心、謝和弦登大人音樂節奏遊戲

關鍵詞彙萃取後，可將群集(12)、(14)、(17)、(18)、(22)與(23)分別定義為「棒球/射擊」、「投擲飛行」、「商家經營」、「經典桌遊」、「推幣積分/策略攻防」與「音樂節奏/跳躍動作」之 App 類型。

而在該階段分群結束後，平均群內相似度由第一階段之 0.02392 提升至 0.03426、平均群間相似度由 0.18916 降低至 0.09060；分群品質計算為 0.37814。

第三節、 第三階段分群

該階段將針對未達分群停止條件之群集(9)、(16)與(20)進行再分群。群集(9)之文件相似度門檻值將依前一階層之最佳分群所挑選的參數組合中之文件相似度門檻值 0.03 為基準，分別以 0.035、0.04 與 0.045 作為該群集第三階段之文件相似度門檻值；而群集(16)與群集(20)因前一階段之最佳分群所挑選的參數組合中之文件相似度門檻值為 0.04，故皆使用 0.045、0.05 與 0.055 為該階段之文件相似度門檻值；三個群集之 k 值亦使用 5 為起始值，其分群之評估結果整理如表 4-4。

第三階段分群評估中，群集(9)將 k 值由 5 提升至 10 後，並未獲得更好的分群品質，故不繼續向上調整 k 值，並可知當 k 值為 5、文件相似度門檻值為 0.045 時能獲得最佳分群品質；同樣地，群集(16)在 k 值為 10、文件相似度門檻值為 0.055 時; 群集(20)在 k 值為 5、文件相似度門檻值為 0.055 時，分別能獲得最佳分群品質。

表 4-4 第三階段分群評估結果

群集編號		(9)			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.035	3	0.039400307	0.19654098	0.200468665
	0.04	4	0.04610441	0.13747620	0.335362877
	0.045	5	0.05383063	0.10666064	0.504690707
10	0.035	3	0.039400307	0.19654098	0.200468665
	0.04	4	0.04610441	0.13747620	0.335362877
	0.045	5	0.05383063	0.10666064	0.504690707

群集編號		(16)			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.045	3	0.027177983	0.08056305	0.337350489
	0.05	3	0.02880678	0.10049661	0.286644308
	0.055	3	0.02880678	0.10049661	0.286644308
10	0.045	3	0.032272933	0.1544031	0.20901739
	0.05	4	0.03600983	0.12624863	0.285229446
	0.055	4	0.03099340	0.08242656	0.376012328
15	0.045	3	0.032272933	0.1544031	0.20901739
	0.05	4	0.03600983	0.12624863	0.285229446
	0.055	4	0.03099340	0.08242656	0.376012328

表 4-4 第三階段分群評估結果(續)

群集編號		(20)			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.045	4	0.035220402	0.13246664	0.265881308
	0.05	4	0.03530220	0.13234678	0.266740174
	0.055	5	0.03833268	0.10734687	0.357091714
10	0.045	4	0.035220402	0.13246664	0.265881308
	0.05	4	0.03530220	0.13234678	0.266740174
	0.055	5	0.03833268	0.10734687	0.357091714

在第三階段分群後，所有 App 文章共分 30 群。而群集(32)內含文章數量為 86 篇，仍未達到分群停止條件，故須繼續分群，圖 4-3 為第三階段分群結果。

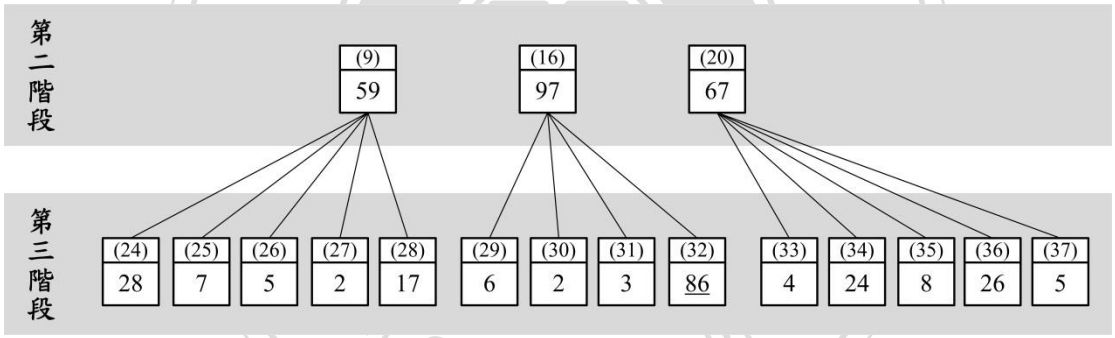

圖 4-3 第三階段分群結果

進行第四階段分群前，亦可對已分群完成之群集進行關鍵詞彙萃取與群集命名。由圖 4-3 可看出已達分群停止條件且群集內含文章篇數大於 3 篇的群集有(24)、(25)、(26)、(28)、(29)，(33)、(34)、(35)、(36)與(37)共 10 個群集，故對這些群集進行關鍵詞彙萃取(如表 4-5)。

表 4-5 關鍵詞彙萃取與群集命名—第三階段

群集編號	**(24)**
文章數量	28
群集命名	守城塔防
關鍵詞彙 (詞彙權重值)	塔防(0.2577)、敵人(0.1778)、消滅(0.1049)、 興建(0.0916)、防禦(0.0851)、房子(0.0661)、 進攻(0.0567)、佈署(0.0529)、防塔(0.0486)、 交易(0.0405)、抵禦(0.0405)、建築(0.0388)
所含 App 項目	Angry Gran、Avatar of War：The Dark Lord 、Catan、Cowboy Guns HD、Defender Chronicles HD、DOFUS: Battles、Elf Defense Eng、Fieldrunners for iPad、Fighting of Sango HD: Legend of Heroes、Fish Draw、Food Fight!、Germcraft、Gun Strike、Guns'n'Glory、Heros and Outlaws、iBomber Defense Pacific、Infinity Field HD、Kingdom Rush、Reckless Getaway、Rocket Rush - Tappi Bear、StarBunker:Guardians、The Oregon Trail、Titan HD、Tower Defense: Lost Earth HD、UltiMaze、WhatsFish、Wispin HD、決戰寶島
群集編號	**(25)**
文章數量	7
群集命名	極限運動
關鍵詞彙 (詞彙權重值)	滑雪(0.1613)、空翻(0.0968)、裝備(0.0878)、 跳躍(0.0791)、技能點(0.0671)、跑酷(0.0645)、 翻滾(0.0533)、終點(0.0461)、彈射(0.0447)
所含 App 項目	Dungeon Hunter、Fibble、Infinity Blade II、Pizza Vs. Skeletons、raging birds、Ski Safari、Wind-up knight

表 4-5 關鍵詞彙萃取與群集命名—第三階段(續)

群集編號	**(26)**
文章數量	5
群集命名	炸彈爆破
關鍵詞彙 (詞彙權重值)	炸彈(0.5652)、解除(0.2393)、集氣(0.1914)、 能量罩(0.1303)、雷射波(0.1303)、引爆(0.1065)、 爆炸(0.0988)、發射器(0.0869)、彈幕(0.0701)
所含 App 項目	Amazing Breaker HD、Bomb Ninja、deBomb、Lightning Fighter、MeWantBamboo
群集編號	**(28)**
文章數量	17
群集命名	動物生態/物理解謎
關鍵詞彙 (詞彙權重值)	動物(0.4555)、獵物(0.1685)、動物園(0.1123)、 觸手(0.1161)、棲息地(0.0696)、可愛(0.0538)、 逼真(0.0639)、豪豬(0.0464)、培育出來(0.0464)
	題(0.2554)、謎(0.1803)、繩子(0.1281)、積木(0.0644)、 思維(0.0464)、答對(0.0464)、題目(0.0464)、 記憶(0.0427)、思考(0.0424)
所含 App 項目	ARGirl、candyJack!、Carnivores:Ice Age、Catch the Candy HD、Contre Jour HD、Cut the Rope、DoubleTake!、DragonVale、happy zoo、LostWinds、Rescue Pine、Sprinkle: Water splashing firefighting fun!、The Heist、ZOOKEEPER DX Touch Edition、Zoozle、小鬼連連看、 趣怪 IQ 大考驗
群集編號	**(29)**
文章數量	6
群集命名	動物養成
關鍵詞彙 (詞彙權重值)	寵物(0.5699)、指令(0.1919)、學會(0.0959)、 叼(0.0959)、技能(0.0818)、貓咪(0.0665)、 飼養(0.0665)、手勢(0.0604)、等級(0.0557)
所含 App 項目	AlexPanda HD、Battlepath Monsters、Bean's Quest、Kinectimals、 Little Skywire、Puzzle & Dragons

表 4-5 關鍵詞彙萃取與群集命名—第三階段(續)

群集編號	**(33)**
文章數量	4
群集命名	任務解決
關鍵詞彙 (詞彙權重值)	探索(0.2036)、巫婆(0.1629)、元素(0.162)、 城市(0.1347)、小組(0.1314)、牧場(0.113)、 黑暗(0.113)、收集(0.0935)
所含 App 項目	Dark Meadow: The Pact、Iris the Captors of Elements、The final escape、Virtual City
群集編號	**(34)**
文章數量	24
群集命名	軍事戰爭
關鍵詞彙 (詞彙權重值)	軍隊(0.1003)、職業(0.0886)、騎士(0.0794)、 技能(0.0744)、陣營(0.068)、發動(0.048)、 攻擊(0.0479)、僱用(0.0454)、武器(0.0429)、 戰略(0.0409)、敵人(0.0402)
所含 App 項目	91 農場 HD、AngerOfStick2、Avengers Origins : Assemble!、BattleHeart、Braveheart、Call of Mini: Last Stand、COMMAND & CONQUER TM RED ALERT TM、Extreme Road Trip、Infect Them All、Inotia 4 PLUS: Assassin of Berkel、Knights of the Phantom Castle、Marvel KAPOW!、MARVEL VS. CAPCOM 2、Monsters Ate My Condo、N.O.V.A. - Near Orbit Vanguard Alliance HD!、Paladog、Plants vs. Zombies HD、Ranch Rush、SAMURAI BLOODSHOW、Speed Blazers、StickWars、Wolfenstein 3D Classic Lite、Zombie in My Pocket、魔導紀元
群集編號	**(35)**
文章數量	8
群集命名	水中冒險
關鍵詞彙 (詞彙權重值)	魚缸(0.5091)、青蛙(0.3322)、水族箱(0.3294)、 魚兒(0.2639)、鱷魚(0.0898)、蟾蜍(0.0599)、 游來游去(0.0599)、指引(0.033)、避開(0.033)、 經驗值(0.0316)
所含 App 項目	Liqua Pop、Pool Bar、RunBallRun、Tap The Frog 2 HD、The Lost Frog、Watee、Zuma's Revenge HD、開心水族箱

表 4-5 關鍵詞彙萃取與群集命名—第三階段(續)

群集編號	**(36)**
文章數量	26
群集命名	角色扮演
關鍵詞彙 (詞彙權重值)	飛機(0.4398)、機場(0.3357)、大亨(0.1068)、 病患(0.1068)、獵人(0.0952)、航空(0.0763)、 貨櫃(0.0763)、魚餌(0.0763)、超市(0.061)、 檢查站(0.0492)、煮菜(0.0458)
所含 App 項目	A Voice Sharpshooter、Airport Mania: First Flight HD、Amateur Surgeon 2、Anthill、AstroWings - War has begins、Carnivores: Dinosaur Hunter、Cooking Mama、Creatures & Castles、Dino Story─Pocket Pets、Downhill Xtreme、Farm Frenzy 3、Flight Control、Flight Control Rocket、Flight Tycoon、Groove Coaster、Hook'em Fishing、Hyperlight、Mr.Space!!、OvenBreak INNERSPACE、Paper Monsters、Pocket RPG、Puzzle Quest 2、SNOKIO、Temple Run、超市大亨、釣魚派對
群集編號	**(37)**
文章數量	5
群集命名	劇情犯罪
關鍵詞彙 (詞彙權重值)	黑幫(0.3362)、俠盜(0.2615)、老大(0.1808)、 殺人(0.1506)、搶劫(0.1506)、殺手(0.1494)、 臥底(0.0747)、毒品(0.0747)、犯罪(0.0747)、 警察(0.0617)、槍戰(0.0518)
所含 App 項目	Clear Vision (17+)、Grand Theft Auto 3、Mafia Rush、Max Payne Mobile、Shotgun Free 2: Duel

在第三階段之各群集關鍵詞彙萃取後，可對這些群集命名，並獲得「守城塔防」、「極限運動」、「炸彈爆破」、「動物生態/物理解謎」、「動物養成」、「任務解決」、「軍事戰爭」、「水中冒險」、「角色扮演」與「劇情犯罪」共 10 種 App 類型。第三階段分群完成後，平均群內相似度由第二階段之 0.03426 提升至 0.04430、平均群間相似度由 0.09060 降低至 0.07683；分群品質計算為 0.57655。

第四節、 第四階段分群

在第三階段分群後，群集(32)未達分群停止標準，故須進行第四階段分群。群集(32)依前一階段最佳分群結果之文件相似度門檻值 0.055 為基準，將文件相似度門檻值設為 0.06、0.065 與 0.07；k 值同樣地設為 5 為起始值，並視情況向上調整以評估不同參數組合之下之分群品質，該階段評估結果如表 4-6。

表 4-6 第四階段分群之評估結果

群集編號		**(32)**			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.06	無分群效果			
	0.065	4	0.05092507	0.10721112	0.47499801
	0.07	4	0.05170809	0.10194756	0.507202885
10	0.06	無分群效果			
	0.065	4	0.05092507	0.10721112	0.47499801
	0.07	4	0.05170809	0.10194756	0.507202885

由表 4-6 可知，群集(32)在 k 向上調整至 10 時，相較於 k 為 5 時無法獲得較佳的分群品質，故不繼續向上調整 k 值，並可知當 k 為 5、文件相似度門檻值為 0.07 時能獲得最佳分群。

而第四階段分群結果如圖 4-4 所示，並可得知群集(41)尚未達到分群停止標準，故須繼續分群；而群集(39)與(40)可進行關鍵詞彙萃取與群集命名，結果如表 4-7；並將其分別命名為「紙牌對戰」與「物件消除」類型之 App。該階段分群結束後，可使平均群內相似度由第三階段之 0.04430 提升至 0.05060、平均群間相似度則由 0.07683 降低至 0.07448；分群品質計算為 0.67942。

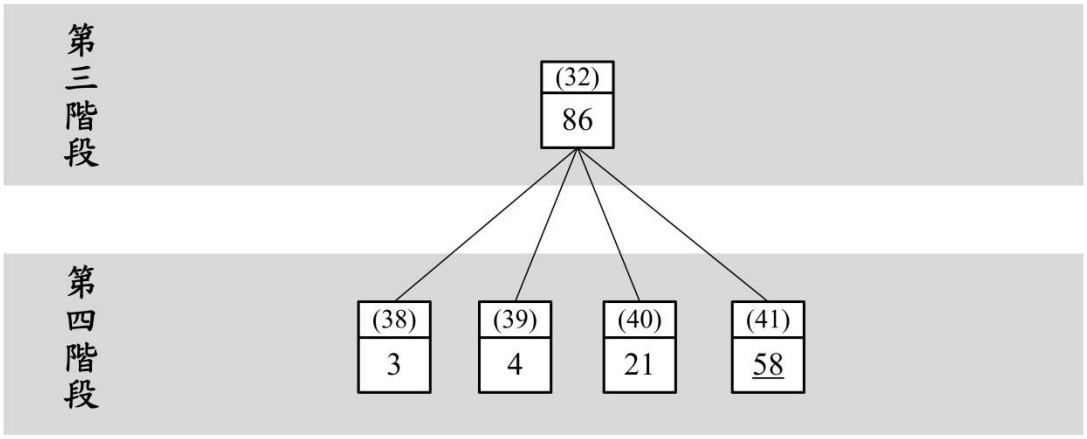

圖 4-4 第四階段分群結果

表 4-7 關鍵詞彙萃取與群集命名—第四階段

群集編號	**(39)**
文章數量	4
群集命名	紙牌對戰
關鍵詞彙 (詞彙權重值)	牌(0.4457)、牌堆(0.1084)、手牌(0.1084)、 牌面(0.1084)、敵對(0.0722)、策略(0.066)、 牌組(0.0583)、回合(0.0542)、撲克牌(0.0443)
所含 App 項目	PickRed、Reiner Knizia's Money、Tigris & Euphrates、文明復興
群集編號	**(40)**
文章數量	21
群集命名	物件消除
關鍵詞彙 (詞彙權重值)	方塊(0.5643)、寶石(0.3349)、消除(0.2522)、 消去(0.1346)、磚塊(0.1109)、鑽石(0.085)、 連消(0.0832)、同顏色(0.0693)、消掉(0.0673)、 連線(0.067)、俄羅斯(0.0535)、連成(0.0407)
所含 App 項目	A Monster Ate My Homework、Boom Boom Gems、BowQubes、Call of Atlantis、Clean Bubbles、Crow、Gu Morning、Happy Garden、Jelly Bear、Jet Ball、Jewel Frenzy、Magic Tiles、Oh! Cube、PANDA BBQ、Piggy Woogy、SotA、Space-Bean、SpinPop Lite、The Chainer、Ultimate Gem Free、電力方塊 Tesla Blocks

第五節、 第五階段分群

接著，內含文章篇數超過 30 篇的群集(41)需要進行第五階段分群，並設定文件相似度門檻值 0.075、0.08 及 0.085；而 k 值與前一階段相同，以 5 為起始值來進行分群評估，其評估結果整理如表 4-8。

當群集(41)將 k 值由 5 調整至 10 時，無法獲得更好的分群品質，故不繼續向上調整 k 值，並可知在 k 值為 5、文件相似度門檻值為 0.085 時，能獲得最佳分群品質。

表 4-8 第五階段分群之評估結果

群集編號		**(41)**			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.075	無分群效果			
	0.08	3	0.04530556	0.05978236	0.757841595
	0.085	4	0.05108477	0.06604159	0.773524245
10	0.075	無分群效果			
	0.08	3	0.04530556	0.05978236	0.757841595
	0.085	4	0.05108477	0.06604159	0.773524245

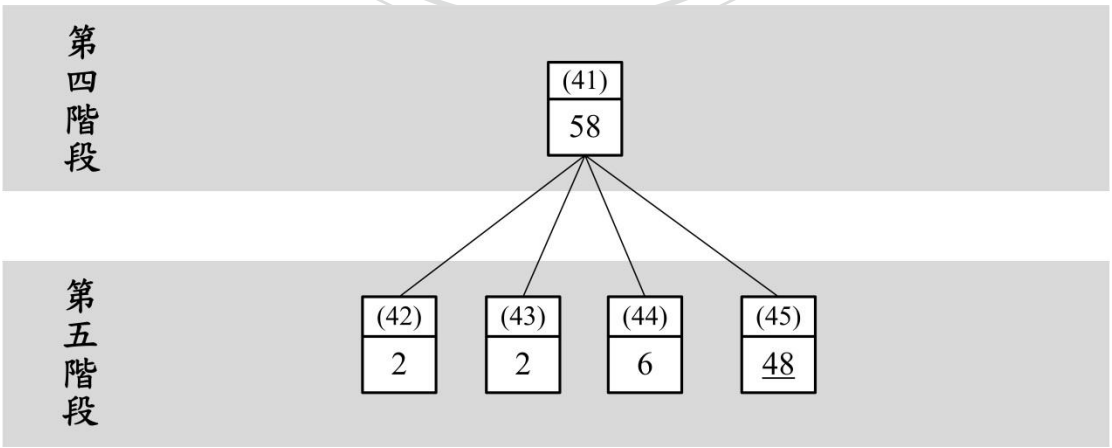

圖 4-5 第五階段分群結果

圖 4-5 為第五階段分群結果，並可對群集(44)進行關鍵詞彙萃取與群集命名。從圖中可看出群集(45)未達分群停止條件，故需繼續進入下一階段分群；但由於群集(45)在第六階段中，k 值之起始值為 5 時，所搭配的 3 種文件相似度門檻值(0.09、0.095 與 0.1)皆未能獲得分群效果(如表 4-9)，滿足分群停止條件之一，故本研究認為該群集內之 App 已達到一定之相似程度，即使該群集仍未達到內含文章數不超過 30 篇之分群停止條件，仍不對該群集繼續進行分群，並以第五階段之結果為最終分群結果。

依第五階段分群為最終分群結果需針對群集(44)與(45)進行關鍵詞彙萃取與群集命名，其結果如表 4-10，並可將其分別命名為「腦力激盪」及「博弈對決/競速賽車」類型之 App。

表 4-9 第六階段分群之評估結果

群集編號		**(45)**			
k 值	文件相似度門檻值	分群數	平均群內相似度	平均群間相似度	分群品質
5	0.09	無分群效果			
	0.095	無分群效果			
	0.1	無分群效果			

表 4-10 關鍵詞彙萃取與群集命名—第五階段

群集編號	**(44)**
文章數量	6
群集命名	腦力激盪
關鍵詞彙 (詞彙權重值)	華容道(0.3388)、轉盤(0.2177)、智能(0.1175)、 拼圖(0.1039)、布局(0.1039)、腦力(0.0622)、 下棋(0.0565)、才智(0.0565)
所含 App 項目	Another Puzzle Game、CEO LIFE、GoBang Of Chinese Style、HuaRong of Three Kingdoms、瘋狂轉盤、繽紛華容道

群集編號	**(45)**
文章數量	48
群集命名	博弈對決/競速賽車
關鍵詞彙 (詞彙權重值)	賽車(0.5162)、賽道(0.3199)、賽事(0.1052)、 跑車(0.0898)、競賽(0.0718)、競速(0.0718)、 甩尾(0.0567)、改裝(0.0522)、跑道(0.0505)
	麻將(0.2833)、打麻將(0.089)、三缺一(0.0486)、 對戰(0.0476)、連線(0.046)、大老二(0.0405)、 牌卡(0.0405)、對手(0.0345)、心臟病(0.0324)、 麻將牌(0.0324)
所含 App 項目	A+WordPuzzle、AI.Buster HD、Apex Of The Racing、Asphalt 6: Adrenaline、Backseat Driver、Baseball Superstars II Pro、Castle attack –Ultimate HD、Crash drive 3D、Death Rally、Donuts Chaser、EPOCH、Fingle、Forever Drive、Frisbee Forever、Fun Fun Sports、Funny Slap、GT Racing:Motor Academy、HeartAttack、Hook 4 Fun、Horror Racing、iQuoit、Jelly Defense、KATAMARI Amore、Lunar Racer、Mass Effect TM Infiltrator、Mini Motor Racing、Need for Speed、Real Racing HD、Reckless Racing、Shark Dash、Smash Cops、Sonic & SEGA All-Stars Racing、Sonic The Hedgehog 4TM Episode I、StoneRings-Lite、Super Laser： The Alien Fighter、Synth Racing、Tiki Kart 3D 、Tilt To Fly、Timeline for iPad、三國塔防 - 魏傳、上海麻將 3D、天下麻將、捉鬼遊戲、掌上三國、極上豪華麻將、神來也大老二、麻將 明星 3 缺 1 HD、麻將三國

最後一階段分群結束，平均群內相似度由第四階段之 0.05060 提升至 0.05396、平均群間相似度則由 0.07448 降低至 0.07118；可得該階段分群品質為 0.75809。

經過了五個階段分群後，將 357 篇 App 文章共分為 36 群(如圖 4-6)，並針對其中內含文件篇數超過 3 篇之 20 個群集進行關鍵詞彙萃取及群集命名，表 4-11 為最終分群與各 App 群集命名結果。

表 4-11 最終分群與 App 群集命名結果

群集編號	**(2)**	**(6)**	**(7)**	**(8)**	**(10)**
文章數	3	3	2	2	2
App 類型	-	-	-	-	-
群集編號	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**
文章數	3	13	3	5	3
App 類型	-	棒球/射擊	-	投擲飛行	-
群集編號	**(17)**	**(18)**	**(19)**	**(21)**	**(22)**
文章數	23	22	2	2	17
App 類型	商家經營	經典桌遊	-	-	推幣積分/策略攻防
群集編號	**(23)**	**(24)**	**(25)**	**(26)**	**(27)**
文章數	29	28	7	5	2
App 類型	音樂節奏/跳躍動作	守城塔防	極限運動	炸彈爆破	-
群集編號	**(28)**	**(29)**	**(30)**	**(31)**	**(33)**
文章數	17	6	2	3	4
App 類型	動物生態/物理解謎	動物養成	-	-	任務解決
群集編號	**(34)**	**(35)**	**(36)**	**(37)**	**(38)**
文章數	24	8	26	5	3
App 類型	軍事戰爭	水中冒險	角色扮演	劇情犯罪	-
群集編號	**(39)**	**(40)**	**(42)**	**(43)**	**(44)**
文章數	4	21	2	2	6
App 類型	紙牌對戰	物件消除	--	-	腦力激盪
群集編號	**(45)**				
文章數	48				
App 類型	博弈對決/競速賽車				

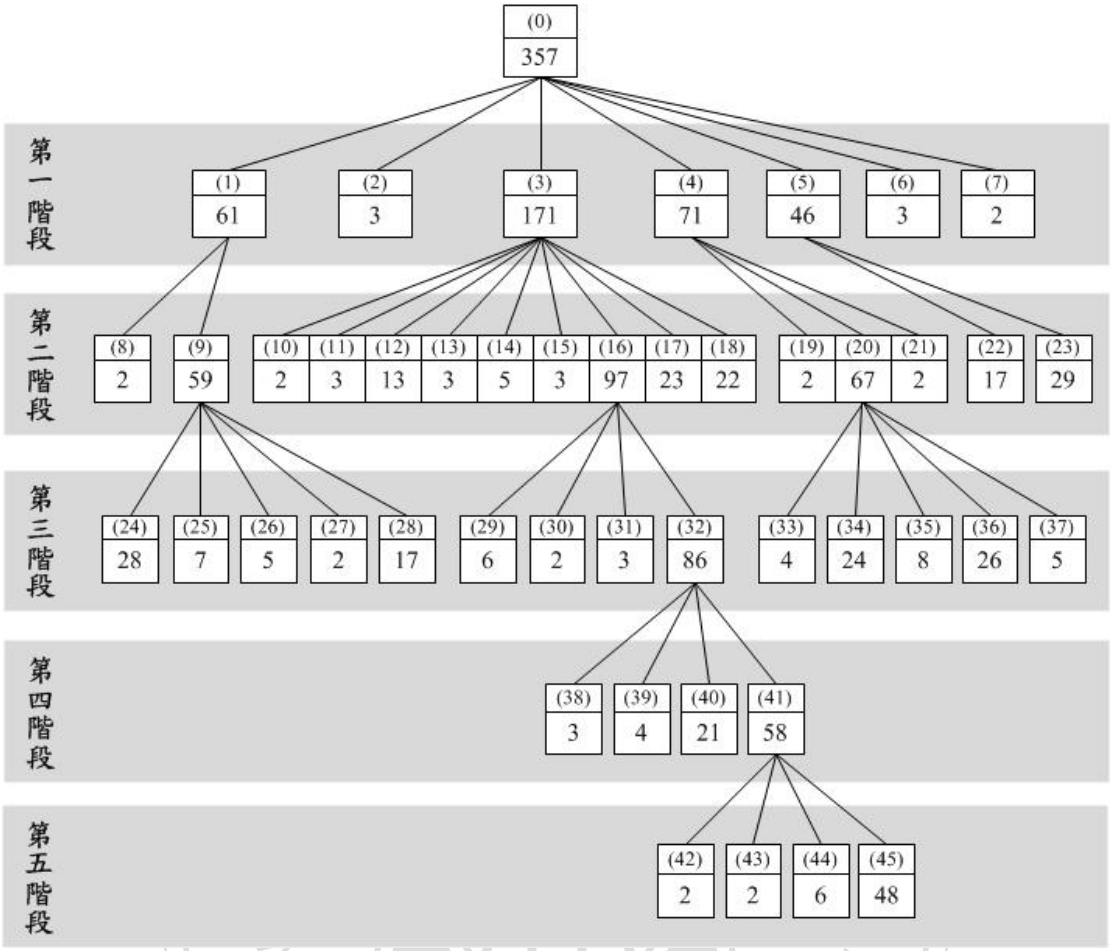

圖 4-6 App 多階段分群結果

而表 4-12 為各階段所得之分群品質，由該表可知每經過下一階段分群處理後，分群品質都會不斷提升，代表分群後之各群集內的 App 在經過各階段分群處理後亦更為相似。

表 4-12 各階段分群品質

	平均群內相似度	平均群間相似度	分群品質
第一階段	0.02392	0.18916	0.12648
第二階段	0.03426	0.09060	0.37814
第三階段	0.04430	0.07683	0.57655
第四階段	0.05060	0.07448	0.67942
第五階段	0.05396	0.07118	0.75809

第五章、結論與未來研究方向

第一節、 結論與建議

本研究於 App 推薦相關論壇中蒐集了 439 篇 App 推薦文章，將其依推薦對象 App 之異同，合併成 357 篇 App 推薦文章；並將文章透過 kNN 等文字探勘及群集分析技術進行分群，使得相似的 App 文章可聚集成群，並萃取各群集關鍵詞彙來為各群集進行群集命名。

因本研究之資料分析對象為 App 推薦文章，相較於具有某一事件為主題的新聞文章而言，具有多種不同構面(例如：遊戲玩法、主角特徵、畫面視覺等等)，使得 App 推薦文章在進行分群時，無法經由一次分群之後就獲得最佳之分群結果，故本研究提出一種新的分群概念—「多階段分群」，即將文章數量較多之大群集進行再次分群，使得分群結果能夠透過不斷分群來逐漸改善其分群品質。

由本研究結果顯示：未分群之平均群內相似度為 1.63%，在經過五階段群集分析後提升至 5.4%；而平均群間相似度由第一階段分群後所測得的 18.92%逐漸降低至 7.12%。即在各階段分群持續進行之下，平均群內相似度不斷增加，而平均群間相似度逐漸降低，分群品質之衡量指標則由第一階段分群後計算而得的 12.65%提升至第五階段分群後的 75.81%；由此可知，使用「多階段分群」後能比只進行一次分群獲得較好分群品質，而每經過一階段分群後，皆能使各資料群集的分群相似度提高，並讓相似的 App 聚集成群，最終分群結果共分成 36 個 App 群集。

在群集命名方面，第一階段分群後共將 357 篇 App 推薦文章分為 7 個群集，而第二階段分群後共可獲得 19 個群集。其中，內含文章篇數超過 3 篇且不需進行再分群之群集有：群集(12)、(14)、(17)、(18)、(22)和(23)。針對群集(12)所萃取出的關鍵詞彙有「全壘打、打擊、揮棒、隊友、投出、棒球、贏球、球棒」以及「戰役、子彈、射擊、槍械、武器、血量、武裝、裝甲」，由這些關鍵詞彙可

將群集(12)命名為「棒球/射擊」類型之 App 群集；而群集(14)可萃取出「小鳥、星球、太空、引力、翅膀、拋物線、終點、降落、飛行、發射」等關鍵詞彙，並可將該群集命名為「投擲飛行」類型之 App 群集。利用相同的關鍵詞彙萃取與群集命名方式可在該階段可獲得「商家經營」、「經典桌遊」、「推幣積分/策略攻防」以及「音樂節奏/跳躍動作」類型之 App；第三階段可獲得「守城塔防」、「極限運動」等 10 種類型之 App 群集；第四階段可獲得「紙牌對戰」與「物件消除」2 種類型之 App 群集；最後，第五階段則可獲得「腦力激盪」及「博弈對決/競速賽車」2 種類型之 App 群集。

而在第五階段分群時，群集(45)內含文章共 48 篇，並未達到「群集內文章篇數未滿 30 篇則不繼續分群」之分群停止條件，故需進行第六階段分群。但在進行第六階段分群時，當 k 值之起始值為 5，所搭配的 3 種文件相似度門檻值皆無法獲得分群效果，滿足另一分群停止條件，故毋須繼續進行分群。由群集(45)中權重前 20%之詞彙可萃取出「賽車、賽道、賽事、跑車、競賽、競速、麻將、打麻將、三缺一、對戰、連線」等關鍵詞彙，並可將該群集命名為「博弈對決/競速賽車」類型之 App 群集；而仔細觀察該 App 群集所包含之 App 可發現：被歸屬到「博弈對決/競速賽車」群集之下的 App 在玩法上皆強調連線對戰、相互對決與競賽等玩法，意即此類型的 App 對於使用者在使用感受及推薦文章的文字抒發上擁有較大的相似性。

而本研究透過使用者所撰寫之 App 推薦文所分析獲得之關鍵詞彙，以作為各群集命名的參考，與官方所訂定之分類相比，較能真實地反映出使用者的使用感受，使用者以本研究結果作為選擇 App 時的參考也更為適切； App 開發人員亦能透過各群集中權重值較高的關鍵詞彙，來做為相關 App 開發時注意的 App 關鍵元素，並能在推廣 App 時利用關鍵詞彙加強行銷文宣，以獲得使用者青睞。

第二節、 未來研究方向

本研究將文字探勘技術應用於智慧型終端 App 並分析其群集關鍵詞彙以進行群集命名。針對未來之研究方向，本節提出以下建議：

1. 建立專業領域詞庫，降低雜訊干擾。

未來可選擇具有公信力的 App 介紹網站，收集專業及大眾認可之各 App 介紹文章，進行文字探勘處理以建立 App 領域之專業詞庫。在分析使用者撰寫之推薦文章時，可透過專業詞庫的過濾，來擷取文章中曾出現於專業詞庫中的詞彙，以優化分群效果，並讓關鍵詞彙之萃取能更加精準。

2. 透過文薦摘要技術，以便讓使用者快速了解 App 類型。

利用文件摘要技術將各群集所萃取之關鍵詞彙轉換為能解釋各群集特性之摘要；透過分群結果及各篇 App 推薦文章中各個詞彙的詞性以分析每群集中的 App 推薦文章，並輸出適當的句子成為摘要；以期能讓使用者藉由閱讀摘要，快速地去了解各 App 類型之特性。

3. 利用分群結果，進行使用者選購 App 實證研究。

本研究使用平均群內相似度等衡量指標所計算之分群品質來評估分群結果之品質良窳，以挑選最佳品質之分群。未來可依本研究之分群結果為基礎，進行相關實證研究，意即分析使用者透過本研究獲得之 App 類型來選購所需的 App，並以問卷方式來調查使用者對於所選購 App 與實際使用上的滿意程度，以深入研究後續使用者對 App 選購滿意度之評估。

4. 動態擴增 App 類型，藉以觀察未來 App 趨勢。

在獲得分群結果之後，未來可批次新增 App 推薦文章，並對其進行資料處理後再利用分類技術將新文章歸類到各群集中，以更加突顯現有 App 類型之特性，亦可透過新獲得之 App 類型來觀察遊戲類 App 之發展趨勢。

5. 將分群結果作為 App 遊戲之推薦，開發相關推薦系統。

使用者可透過輸入推薦文，將輸入之推薦文進行文字探勘處理後進行興趣歸類，或是透過選擇研究中所萃取出的各群集關鍵詞彙來進行興趣歸類。歸類後可透過文章之間或文章與關鍵詞彙的相似程度，提供給使用者相似度由高到低的 App 遊戲推薦名單，以達到 App 推薦之目的。

參考文獻

英文文獻

1. 148Apps.biz. (2012). *Count of Active Applications in the App Store.* Retrieved April 20, 2012, from http://148apps.biz/app-store-metrics/?mpage=appcount
2. Apple. (2012). *iTunes Preview.* Retrieved April 20, 2012, from http://itunes.apple.com/us/genre/ios/id36
3. Chen, K. J., & Liu, S. H. (1992). Word identification for Mandarin Chinese sentences. *Proceedings of the 14th conference on Computational linguistics* , 101–107. Nantes, France.
4. Engel, J. F., Blackwell, R. D., & Miniard, P. W. (1993). *Consumer Behaviour* (7th Revised ed.). Chicago: Dryden Press.
5. Fayyad, U. M. (1996). Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Expert: Intelligent Systems and Their Applications, 11*(5), 20–25.
6. Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* , 112–117. Montreal, Canada.
7. Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing, 18*(1), 38–52.
8. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

9. Lai, C. H., & Liu, D. R. (2009). Integrating knowledge flow mining and collaborative filtering to support document recommendation. *Journal of Systems and Software*, *82*(12), 2023–2037.
10. Nie, J. Y., Brisebois, M., & Ren, X. (1996). On Chinese text retrieval. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* , 225–233. New York, USA.
11. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.
12. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, *18*(11), 613–620.
13. Simoudis, E. (1996). Reality Check for Data Mining. *IEEE Expert: Intelligent Systems and Their Applications*, *11*(5), 26–33.
14. Sproat, R. W., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, *4*(4), 336–351.
15. Sullivan, D. (2001). *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*. New York, NY, USA: John Wiley; Sons, Inc.
16. Tan, A. (1999). Text mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* , 65–70. Beijing, China.
17. Teng, W. G., & Lee, H. hsien. (2007). Collaborative Recommendation with Multi-Criteria Ratings. *Journal of Computers*, *17*(4), 69–78.
18. Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., & Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems and their Applications*, *14*(4), 32–43.

19. You, J. M., & Chen, K. J. (2006). Improving context vector models by feature clustering for automatic thesaurus construction. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing* ,1–8. Sydney, Australia.

中文文獻

1. SmartMobix. (2012). *移動裝置上使用時間何者最多？使用者：App.* 行動智庫. Retrieved January 25, 2012, from http://www.smartmobix.com.tw/flurry_20110622
2. 吳文峰. (2002). *中文郵件分類器之設計及實作.* 逢甲大學資訊工程系碩士論文.
3. 巫啟台. (2002). *文件之關聯資訊萃取及其概念圖自動建構.* 國立成功大學資訊工程學系碩士論文.
4. 林姿旻. (2011). *數位遊戲之行動載具使用者行為與開發分析—以智慧型手機為例.* 國立政治大學數位內容碩士論文.
5. 胡秀珠. (2011). *55%業者一年內推出App服務.* 創新發現誌. Retrieved March 5, 2012, from http://ideas.org.tw/magazine_article.php?f=464
6. 郭芳菲. (2003). *利用和絃特徵探勘音樂旋律曲風之研究.* 國立政治大學資訊科學學系碩士論文.
7. 陳柏均. (2011). *文件距離為基礎kNN分群技術與新聞事件偵測追蹤之研究.* 國立政治大學資訊管理學系碩士論文.
8. 陳崇正. (2009). *應用網路書籤與VSM相似度演算法於強化實踐社群的形成.* 國立中央大學資訊工程學系碩士論文.
9. 楊智凱. (2007). *唐詩推薦系統之研究.* 亞洲大學資訊科學與應用學系碩士論文.
10. 盧希鵬. (2005). *網路行銷：電子化企業經營策略.*台北市:雙葉書廊有限公司.
11. 胡國信. (2005). *具分群機制之遞增式最鄰近分類學習法 --垃圾郵件過濾之應用.* 國立屏東商業技術學院資訊管理學系碩士論文.