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

應用文字探勘分析網路團購商品群集之研究
－以美食類商品為例

The Study of Analyzing Group-buying Goods Clusters by Using Text Mining
– Exemplified by the Group-buying foods

研究生：趙婉婷

中華民國一百零一年七月

誌謝

首先感謝我的論文指導教授楊建民老師，在政大的這些年裡給予我許多幫助及指導，當我陷入思考的僵局時總是能帶給我突破性的想法，平常生活中老師的關心與鼓勵更是綿綿不絕，非常感謝楊老師的用心指導！此外感謝劉文卿老師、季延平老師以及邱光輝老師所給予我許多寶貴的指導與建議。

感謝政大資管所的同學們，每次遇到你們總是能獲得許多關心跟鼓勵，大家一起順利畢業！！特別感謝 LAB 共患難的水族箱的成員～果節、小皇及裸羅三位大大！謝謝你們與我共度許多艱困時期，給了我無數的幫助跟鼓勵，有你們的陪伴真好！還有智民學姊、柏均學長以及振和學長，謝謝你們在研究上給了我們許多幫助！LAB 的學弟們，謝謝你們的鼓勵！另外還有許多給我支持跟鼓勵的親戚與好朋友們！總是關心我的近況，鼓勵我祝福我！感謝你們！

最後，感謝的是一直支持我的家人～我最親愛的爸爸媽媽和哥哥，在我大忙論文的時期給了我無限的包容跟照顧！最特別的感謝獻給我最鍾愛的華仔兔～謝謝祢總是陪伴在姊姊身邊，當姊姊為論文忙到焦頭爛耳的時候只要看到祢摸摸祢和祢玩耍一下就是最大的休息和娛樂了！從姊姊大學進入政大開始，祢就陪伴著我，到現在姊姊碩班畢業要離開政大了，祢也離開我們變成小天使！謝謝祢帶給我們家無限的歡樂跟開心回憶，謝謝祢很體貼的選在姊姊論文完成的最後時刻才離開我...我們會永遠想念祢的！

摘要

網路團購消費模式掀起一陣風潮，隨著網路團購市場接受度提高，現今以團購方式進行購物的消費模式不斷增加，團購商品品項也日益繁多。為了使網路團購消費者更容易找到感興趣的團購商品，本研究將針對團購商品進行群集分析。

本研究以國內知名團購網站「愛合購」為例，以甜點蛋糕分類下的熱門美食團購商品為主，依商品名稱找尋該商品的顧客團購網誌文章納入資料庫中。本研究從熱門度前 1000 項的產品中找到 268 項產品擁有顧客團購網誌 586 篇，透過文字探勘技術從中擷取產品特徵相關資訊，並以「k 最近鄰居法」為基礎建置 kNN 分群器，以進行群集分析。本研究依不同的 k 值以及分群門檻值進行分群，並對大群集進行階段式分群，單項群集進行質心合併，以尋求較佳之分群結果。

研究結果顯示，268 項團購商品經過 kNN 分群器進行四個階段的群集分析後可獲得 28 個群集，群內相似度從未分群時的 0.029834 提升至 0.177428。在經過第一階段的分群後，可將商品分為 3 個主要大群集，即「麵包類」、「蛋糕類」以及「其他口感類」。在進行完四個階段的分群後，「麵包類」可分為 2 種類型的群集，即『麵包類產品』以及『擁有麵包特質的產品』，而「蛋糕類」則是可依口味區分為不同的蛋糕群集。產品重要特徵詞彙不像一般文章的關鍵字詞會重複出現於文章中，因此在特徵詞彙過濾時應避免刪減過多的產品特徵詞彙。群集特性可由詞彙權重前 20%之詞彙依人工過濾及商品出現頻率挑選出產品特徵代表詞來做描繪。研究所獲得之分群結果除了提供團購消費者選擇產品時參考外，也可幫助團購網站業者規劃更適切的行銷活動。本研究亦提出一些未來研究方向。

關鍵字：文字探勘、團購、最近鄰居法、kNN 分群

Abstract

Group-buying is prevailing, the items of merchandise diverse recently. In order to let consumer find the commodities they are interested in, the research focus on the cluster analysis about group-buying products and clusters products by the features of them.

We catch the blogs of products posted by customers, via text mining to retrieve the features of products, and then establish the kNN clustering device to cluster them. This research sets different threshold values to test, and multiply clusters big groups, and merges small groups by centroid, we expect to obtain the best quality cluster.

From the results, 268 items of group-buying foods can be divided into 28 clusters, and the mean of Intra-Similarity also can be improved. The 28 clusters can be categorized to three main clusters：Bread, Cake, and Other mouthfeel foods. We can define and name each cluster by catch the top twenty percent of the keywords in each cluster. The results of this paper could help buyers find similar commodities which they like, and also help sellers make the great marketing activity plan.

Keywords: Text Mining, Group-buying, k-Nearest Neighbors, kNN clustering

目錄

圖目錄

表目錄

第一章　緒論

第一節 研究背景與動機

近年來，網路團購的消費模式掀起一陣風潮，從前年開始便可以感受到網路團購增溫的現象，像是美國的 Groupon 買下台灣地圖日記旗下的團購網站，成立 Groupon 台灣分站，Google 與 Facebook 也分別新推出 Google offers 優惠訊息服務以及 Deals on Facebook 團購服務，Yahoo!奇摩網站也祭出大團購以及折扣$^{+}$等團購性質服務，而台灣本地則是有蕃薯藤、愛評網、愛合購等搶進團購市場。

根據資策會 MIC 的調查(蘇文彬,2011)發現,使用網路團購的網友逐年提升，顯示網路團購市場逐年擴大。隨著網路團購的市場接受度提高，許多原本由公司內部發起或親友之間呼朋引伴的實體團購行為，也轉移至網路上進行，一同團購的族群也跨及至網友及社群，大家結合共同的購物需求，集結大量的購買量向店家下單，如此一來便可獲得更優惠的折扣。除了消費者獲利之外，店家也可從團購的消費模式中獲得好處，像是在短時間內累積大量的購買人數，大量的訂單可以壓低製作成本，且網友之間轉貼、宣傳團購訊息可以使店家的曝光率增加，提升知名度及買氣，店家也可從龐大的購買人數中挖掘新顧客名單...等，雙方皆受惠，因此也更帶動團購消費市場的興盛。

隨著網路團購消費模式的崛起,現今以團購方式進行購物的消費模式不斷增加，團購商品種類也日益繁多。以國內知名團購網站「愛合購」(ihergo)為例，光美食分類中的甜點蛋糕類商品就高達近 10,000 項,在如此龐大的商品數量下，消費者不容易找到感興趣之商品，從消費者的角度來看，確實帶來資訊過載的困擾。此外，團購消費模式和一般的網購消費模式的不同處在於團購的消費者容易

受到網友們所分享的網誌內容而影響團購意願，另外也容易受到一窩蜂式的群集購買效應影響而有從眾行為的發生，在此狀態下，消費者容易因為盲從而買到不甚滿意的產品。

基於上述原因，本研究希望能以消費者的觀點找出商品的特性，再依照商品特性替商品進行分群，透過替商品分群的概念，來幫助消費者更容易找到喜歡且感興趣之商品，也協助團購網站業者規劃出更適切的行銷活動。

第二節 研究目的

本研究針對有團購經驗的消費者為研究對象，並以網路上常見的團購商品類型—『美食類』為基礎，透過文字探勘技術獲取商品特性，並結合 k 最近鄰居分群器(以下簡稱 kNN 分群器)的運用，以分出擁有不同產品特性的團購商品群。對於分群完後的結果，依據其特有的商品特徵詞去解釋並描繪此團購商品群的特質，以便日後推薦給可能有興趣的團購消費者及業者參考。

茲將此次研究之目的敘述如下：

一、運用文字探勘的技術獲取產品特徵相關資訊。

二、透過 kNN 分群器進行團購商品分群，並找出較佳之分群結果。

三、分群後所產生的團購商品群集依據其共有之產品特徵詞描繪出群集輪廓。

第二章　文獻探討

本研究以團購為主題，利用文字探勘以及群集分析的概念進行研究。本章將對團購的定義與類型，文字探勘的處理及運作，以及群集分析的 k 最近鄰居法之原理及運用進行相關文獻的探討，以作為後續研究架構建立之基礎。

第一節 團購

2.1.1 團購的定義與類型

團購(group-buying)是一種便於獲得折扣的消費模式。Anand and Aron(2003)指出其主要包含兩元素：需求聚集(demand aggregation)與數量折扣(volume discounting)，藉由匯集消費者的需求，使得價格隨著需求量的增加而下降或是獲得更多的商品數量。

團購的消費模式行之有年，以往的團購多為實體團購，其指的是傳統的生活中同一個地區的人針對共同的需求，藉由相互的溝通與協調來群體採購同一類商品，達到降低售價的目的(莊隆泰，2000)，且這樣的行為通常發生於辦公室，家庭等地方，集結眾人的需求，取得最大的購買量，以提高議價能力，使消費者達到更好的購買條件(Anand & Aron，2003) 。

隨著網路時代來臨，人與人之間的聯繫及溝通極為便利，團購消費模式也透過網路管道散播，跨越了地理與時間的限制，成為網路團購模式。楊惠琴(2006)將網路團購模式定義為一群人在網路上結集成虛擬社群，賦予網購社群交流的意味，透過互相合作，達到節省運費以及折扣互惠。

在團購的類型方面，林淑婉(2010)整理現今常見的網路團購型態，其依團購過程可分為聯合親朋好友共同購買以及主購發起號召合購網友。前者團購過程較為簡單，後者主購則需進行較繁瑣之工作，如彙整網友資料、處理款項收付、決定及通知網友合購相關資訊，以及到貨時分發給網友。

另外，廖婉如(2010)將國內目前團購平台依類型分為「電子佈告欄團購版」與「團購網站」兩大類，前者為單純提供溝通的平台，以供消費者交換訊息，並無涉及商業活動的進行；後者則扮演中間商之角色，藉由匯集各廠商資訊、提供討論空間與交易環境，來吸引消費者運用此平台。張家蓁(2010)透過觀察目前台灣團購網站的發展現況，找出較知名的營利團購網站並整理分類為合購平台、團購專門店、大型入口網站以及小型自營商入口網站，其中，「合購平台」為網路業者提供團購平台，為免費撮合店家和團購成員交易的資訊平台，由於不向店家和消費者收取任何費用，所以加入的店家數越多，也會吸引大量有團購需求的網友加入。目前台灣最大的合購網站為「愛合購(ihergo)」。

2.1.2 影響消費者團購意願

團購透過集結大家的需求，來達到獲取較低的商品價格以及節省運費的目的，如此省錢又優惠的購買活動常常帶有「呼朋引伴」的行為發生，不論是辦公室與親朋好友間的相揪成團，或是網路上集體的購買行動，皆可看出團購消費模式與「從眾行為」擁有密不可分的關係。「從眾」為社會影響的表現，其影響來源為個人受到團體中其他成員的影響(Allen，1965)。消費者為了取得群體的認同、符合群體的期望，因此會採取與群體其他成員相似的思想或行為(Wilkie，1994；Macinnis，1997)。

此外，在團購中，也可看見一窩蜂式的購買熱潮，評價良好的產品，透過參考群體的推薦及口碑效應的傳播，匯集更多的人加入購買行列，使其成為當今的熱門團購商品，可見參考群體的口碑也與團購有緊密關聯。潘倧偉(2010)指出，消費者在購買產品時，除了會參考同儕的意見外，還會參考周遭其他消費者的意見，並會根據其意見來決定自己購買的商品，因此消費者們不論是認知還是情感，意向還是行為，或多或少都會受到群體和他人的影響，這也是口碑以及從眾行為對消費者的影響力。

口碑(Word-of-Mouth)為一個不具商業意圖的口頭對話過程，傳播者與接受者主要談論的內容為特定的品牌、產品或服務的意見交換及使用經驗分享(Arndt，1967)。口碑雖然不具銷售企圖，但是它是會影響他人的產品或服務期望效用的推薦行為(Godes et al.，2005) 。呂培仕(2010)整理口碑文獻回顧(1950~2008)，其將口碑相關文獻統整後定義口碑為「買家之間根據不同的主題，以評論或推薦行為的方式，在不同的溝通管道交換訊息的溝通過程；此過程根植於傳遞者與接收者的人際網絡中，因此傳遞的訊息內容是不具商業意圖的，其可信度與影響力都被認為高於廣告和大眾媒體」。而網路時代來臨後，口碑傳播形成所謂的「網路口碑」。網路口碑(Online Word-of-Mouth)被認為是透過電子郵件、線上論壇等的網路形式進行的口碑傳播(Hanson，2000)；後來隨著網路的發展，網路口碑傳播的管道還多了聊天室、部落格、即時通訊軟體等方式(Snyder，2004)。

盧惠芬(2010)研究從眾行為影響網路團購購買意願，其結果顯示消費者容易受到參考群體推薦的影響。團購商品經網友大量分享口碑訊息、媒體大力推薦或親朋好友的推薦，常常會影響到消費者，促使其參與團購或甚而發起團購購買之。此外，潘倧偉(2010)也研究口碑與從眾行為對團購意圖的影響，其研究結果顯示從眾行為對團購意圖有正向影響，且口碑與從眾行為互相有正面影響的效果存在。

因此，從上述相關文獻可得知：團購消費者會受到參考群體的口碑及從眾行為的影響而改變其團購的消費意願。

此外，根據資策會市場情報中心(MIC)調查台灣網友上網購物行為模式(資策會，2007)發現，台灣網友上網購物行為模式以搜尋商品資訊與比價行為最普遍，且非常多數的網友會瀏覽部落格網誌的商品資訊作為購物決策的參考。

綜上述論點，本研究鎖定網友所撰寫的團購美食網誌為資料來源進行分析，團購美食網誌為網路口碑傳播的熱門管道，且網友在購物前也會瀏覽其作為購物時的決策參考，因此本研究將透過顧客團購網誌，從顧客的角度來歸納產品特性，替產品進行分群，使得偏好某特性商品的團購網友可以更容易找到感興趣的商品。

第二節 文字探勘

2.2.1 文字探勘定義

部落格網誌為非結構化的資料，其需透過文字探勘的技術來將資訊萃取出來。巫啟台(2002)提出『文件探勘』(Text Minin)是『從非結構化的文字中發掘出有用的或是有趣的片段、模型、方向、趨勢或規則』。文字探勘試圖從文件資料中找出重要的項目(Term)或片語(Phrase)、項目間的關聯強度(Association Degree)或是分類和推論規則(Classification or Prediction Rule)。文字探勘是針對非結構化(Non-structured)或半結構化(Semi-structured)的文件資料加以分析，有效率地從大量文字性資料中整理出有用的資訊，以將文件中所隱藏的珍貴知識萃取出來。

2.2.2 斷詞處理

在對文字性資料進行文字探勘前，這些資料必須先經過資料前處理的動作，而資料前處理的首要步驟，便是對文字性資料進行斷詞處理。印歐語系文件的斷詞處理與中文文件的斷詞處理有很大的不同之處，印歐語系文件在詞與詞之間以空白及其他符號隔開，因此斷詞僅需透過空格或其他符號的分隔便能將每一個單字斷開成為獨立詞彙(Nie，1996)，而中文文件是由字與標點符號以非結構化的方式所組成，單一的字元未必能成為有意義的單位，字詞與字詞間沒有明顯的邊界(喻欣凱，2008)。

中文文件的斷詞方式主要可分為三種：詞庫式斷詞法、統計式斷詞法以及混合式斷詞法，其說明如下：

(一) 詞庫式斷詞法(Chen，1992)

為目前最普遍的斷詞方式，其演算法直覺且較容易實作，主要概念為利用事先建立的詞庫與文件中的詞彙進行比對，以完成斷詞動作。由於斷詞的品質和詞庫的品質有相當大的關係，因此必須時常對詞庫的內容加以維護及更新。

(二) 統計式斷詞法(Sproat，1990)

依據大型的語料庫(corpus)上的統計資訊，以統計資訊的高低來當作斷詞的依據。優點是不受到詞庫大小詞量多寡的限制，缺點在於語料庫是屬於領域相關(Domain dependent)，因此不同語料庫間的統計資訊不適合互用(Nie，1996)。另一方面，統計式斷詞法有斷詞長度上的限制，其主要著重在二字詞的研究，因此無法完整斷出長辭彙(曾元顯，2002)。

(三) 混合式斷詞法(Nie，1996)

其整合了詞庫式斷詞法及統計式斷詞法。此方式為利用詞庫斷出不同組合的詞彙，然後以字詞的統計資訊，找出最佳的斷詞組合。此法仍需要大型的語料庫提供統計資訊。

中央研究院中文詞知識庫小組(Chinese Knowledge Information Processing Group，CKIP)所開發的中文斷詞系統是採用混合式斷詞法，其將使用者所輸入之文章或句子自動斷詞後在標示出每個詞彙的詞類標記。該系統包含一個約拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題。除了基本詞彙庫外，使用者可依需要附加領域專屬詞庫。詞類標記為選擇性功能，可附加文本中切分詞的詞類解決詞類歧義並猜測新詞之詞類。分詞系統採用之詞典俱可擴充性，使用者可依據不同領域文件，補充以領域詞典做為分詞之用(中央研究院，2012)。

斷詞服務系統的內部處理採用中央研究院中文詞知識庫小組所編列的中研院平衡語料庫詞類標記集之簡化詞類，而斷詞服務系統採用精簡詞類標記，如下表所示：

表 2-1 斷詞服務系統精簡詞類標記

精簡詞類標記	詞類說明
A	非謂形容詞
ADV	副詞、數量副詞、動詞前程度副詞、動詞後程度副詞、句副詞
ASP	時態標記
C	對等連接詞(如：和、跟)、關聯連接詞
DET	指代定詞、數量定詞、特指定詞、數詞定詞
FW	外文標記
M	量詞
N	普通名詞、專有名稱、地方詞、位置詞、時間詞、代名詞
P	介詞
POST	連接詞(如：等等)、連接詞(如：的話)、後置數量定詞、後置詞
T	字(的、之、得、地)、感嘆詞、語助詞
Vi	動作不及物動詞、動作類及物動詞、狀態不及物動詞、狀態類及物動詞
Vt	字(是、有)、動作使動動詞、動作及物動詞、動作接地方賓語動詞、雙賓動詞、動作句賓動詞、動作謂賓動詞、分類動詞、狀態使動動詞、狀態及物動詞、狀態句賓動詞、狀態謂賓動詞

(資料來源：中央研究院中文詞知識庫小組)

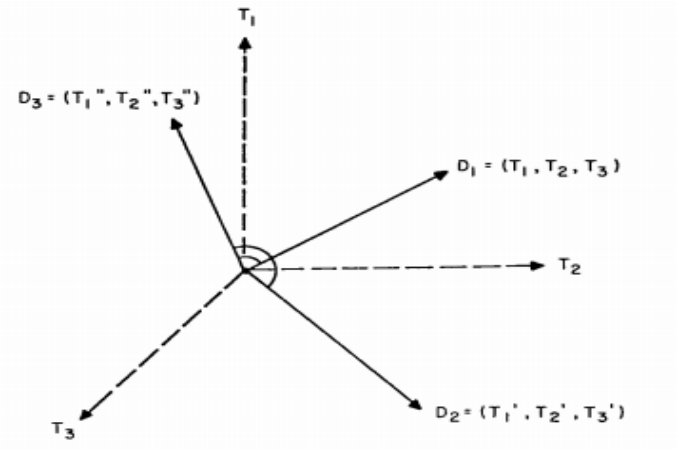
2.2.3 權重計算及特徵詞選取

文章進行斷詞後成為許多詞彙，這些詞彙對文章擁有不同的重要程度，要從這些詞彙中找出文章的重要資訊，必須先擷取出能代表文件特徵的關鍵字詞。詹益發(2009)指出要挑選出具代表性的字詞，可藉由該字詞在文章中的重要性來衡量，亦即計算該字詞在文章中的權重；一般特徵詞的篩選，先利用權重的計算方式找出候選特徵詞，再經自動化或是人工挑選方式找出進一步找出重要之特徵詞。

要從文件中擷取出代表文件的特徵詞彙，可以透過詞彙的出現頻率、出現位置或是詞彙的特性來衡量。一般較常採用的方法為 Salton(1983)所提出的 TF-IDF (Term Frequency–Inverse Document Frequency)字詞權重計算，TF(Term Frequency)為詞彙頻率，計算特徵詞彙在一篇文件中出現的頻率，數值越高代表該特徵詞彙在文件中越重要。一般來說，文件中的高頻詞彙與文件有相當高之關聯，為文件的重要特徵詞。但如果該高頻詞彙不只在該篇文件中出現頻率很高，且在所有文件中的出現次數都很高，則代表此詞彙太過普遍，不具代表性，為了避免擷取到不具代表性的詞彙，因此除了考慮 TF 值之外，還需考量逆向文件頻率(Inverse Document Frequency，IDF)。逆向文件頻率是以該詞彙出現在其他文件中的次數多寡來衡量，數值越低代表該詞彙越能將某文件與其他文件區別，因此越具代表性。 TF-IDF 為 TF 與 IDF 之平衡指標，同時考慮兩者的特性來衡量詞彙在文件中的重要程度，以挑選出具代表性之重要特徵詞彙。

2.2.4 向量空間模型

在文字探勘的領域中，向量空間模型是目前最廣為使用的資訊檢索模式(戴尚學，2003)。向量空間模型由 Gerard Salton 所提出(Salton，1975)，其目的在於將文件轉化成字詞索引的集合，每個字詞皆給予權重值(Weight)，以表達每個字在文件中的重要程度，而最常用的權重計算方式為前述 TF-IDF 計算。下圖 2-1 為向量空間模型圖，在文件集中，每篇文件以一組向量表示，維度代表關鍵字詞，而維度的數值則代表該字詞的權重。



(資料來源：Salton et al，1975)

圖 2-1 向量空間模型

為了便於文件與文件之間特徵詞彙權重值之比較，可將向量空間模型轉成以「詞彙─文件矩陣」形式來表示文件與詞彙間之關係。如下圖 2-2 所示，每一列代表一篇文章，每一欄代表一個特徵詞彙，而文章與詞彙對應到的元素(W)為權重，即該篇文章某特徵詞彙之權重值。

$$\begin{bmatrix} & Term_1 & Term_2 & \cdots & \cdots & \cdots & Term_i \\ Doc_1 & W_{11} & W_{12} & \cdots & \cdots & \cdots & W_{1i} \\ Doc_2 & W_{21} & W_{22} & \cdots & \cdots & \cdots & W_{2i} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ Doc_k & W_{k1} & W_{k2} & \cdots & \cdots & \cdots & W_{ik} \end{bmatrix}$$

(資料來源：Salton & Gill，1983)

圖 2-2 詞彙─文件矩陣

第三節 群集分析

分群是依照文件間的相似性將其分成群集，使得每一群內的文件彼此相似，亦即群內元素對某特性而言為同質，而群與群間則是互有差異，群間相似度低。分群屬於非監督式(Unsupervised Learning)學習，其不需透過已知類別的訓練資料給予訓練來做類別判斷，而是根據輸入資料的特徵將相似的歸於同一群集。

2.3.1 應用於團購領域之相關文獻探討

在團購領域的文獻中，除了透過問卷調查以統計方式分析研究結果外，也有學者結合群集分析技術加以進行。張家蓁(2010)於「資料採礦應用於消費者網路團購因素探勘之研究」中，利用問卷調查後，透過資料採礦以群集分析與關聯法則，在樣本中挖掘出潛在消費者族群與目標消費者族群，區別其網路團購的消費習性、購物行為和購買因素並加以分析。張瑜修(2011)於「消費者參與辦公室團購影響因素之研究-以宜蘭縣上班族為例」中，經過問卷調查後，採用統計軟體進行描述性統計分析，接著運用群集分析將受訪者分為三個群集，討論不同群集對各屬性的偏好及人口統計變項。

2.3.2 k 最近鄰居法的原理及運作方式

k 最近鄰居法雖然被歸類於分類演算法中，但在實作上亦可不事先設定類別及給予訓練資料，Yang et al.(1999)將其運用於「類別數未知」的新聞事件的偵測追蹤，即為 k 最近鄰居法於分群上之應用。

另外，經由戴維德(2005)研究得知，要將龐大的客戶資料加以分類與分析，進而預測顧客對於網路銀行的使用意願，利用 k 最近鄰居法的預測能力是優於決策樹以及類神經網路，且突破統計模型對資料上樣本的限制。因此，在同樣是對顧客資料加以群集分析的網路團購應用方面，也選擇透過 k 最近鄰居法原理建置分群器。

k 最近鄰居法一種最為直接簡單且具有一定精度水準的群集分析法。k 最近鄰居法(k-Nearest Neighbors, kNN)由 Cover & Hart(1967)所提出，此方法是對於一筆未知類別之資料，先找出與資料最鄰近的 k 個資料點，根據這 k 個資料點之類別，來辨別未知資料所屬類別。簡單而言，就是「物以類聚」的概念，擁有相似特徵的資料，在以其特徵形成的空間中會聚集在一起。若以向量空間中的點來表示，對於同一類別物件的這些點彼此間的距離應該會比較接近。所以對於一個未知類別的測試資料，我們只需要在訓練資料中找出和此筆資料最接近的幾個點，就可以以 k 最近鄰居法來判定此筆未知類別之測試資料的類別，其類別應與最接近的幾個點所屬類別最多的類別相同。

對於資料點與資料點間距離的計算方式，大多是採用歐幾里德距離(Euclidean distance)來計算。假設在 n 維的向量空間中有兩個點 P 跟 Q，$P = (p_1, p_2, \dots, p_n)$、$Q = (q_1, q_2, \dots, q_n)$，則歐幾里德距離的計算公式如下：

$$D_{Euclidean} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

(p_i與q_i為標準化後的特徵值)

在運作方式部分，k 最近鄰居法屬於懶散學習法，有新的測試資料時才開始做分類處理。在學習階段只是簡單的將每筆訓練資料(training data)作適當的表示後儲存起來，就完成了訓練工作。當有一筆測試資料(test data)需要分類時，再將測試資料與所有訓練資料逐一比對，找出 k 筆距離最近的訓練資料，再依據這 k 筆訓練資料所屬的類別，利用投票的方式評估此測試資料最後應歸屬的類別 (Larkey and Croft，1996)。

而 k 最近鄰居法應用在分群領域時，也就是在不事先設定類別及給予訓練資料的狀況下，也是在有分群需求時才會開始進行處理，逐一比對資料，找出前 k 個相近資料並進行所屬群集的投票來決定最後歸屬群集。和分類時的運作相比，少了透過訓練資料設定類別的訓練動作。

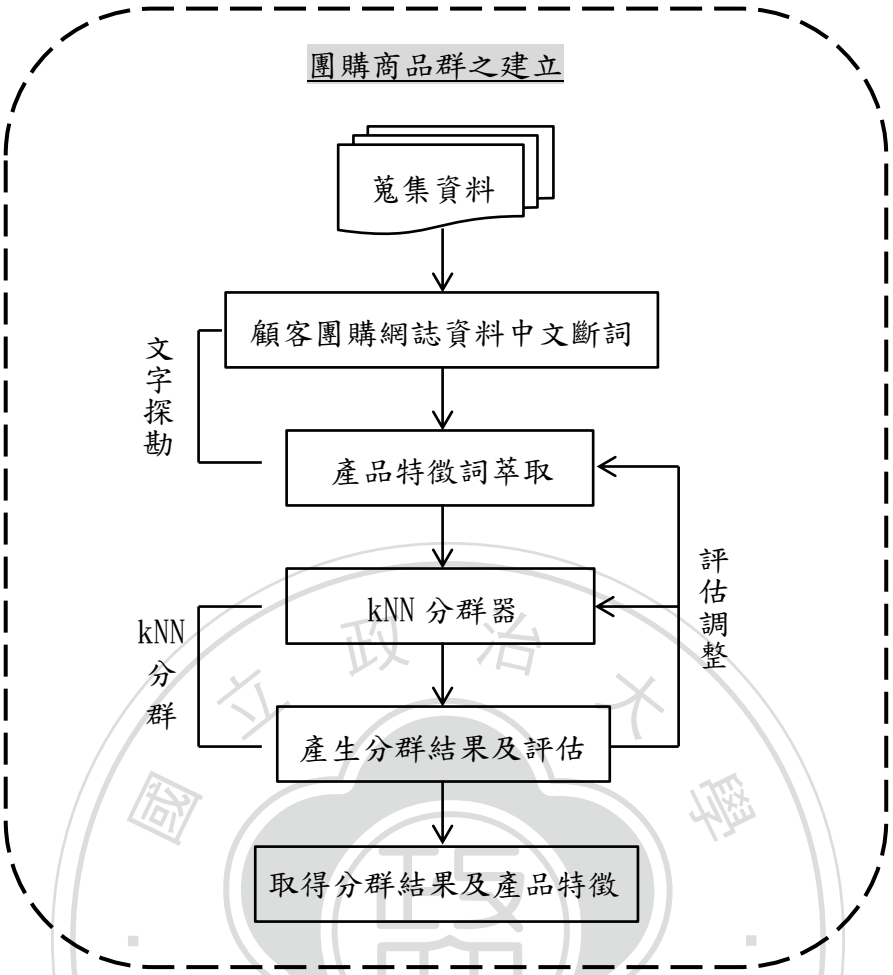
綜上述論點，本研究將以文字探勘技術對顧客團購網誌進行產品特徵擷取，再以 kNN 原理建置分群器來替團購美食商品進行群集分析。

第三章　研究方法與設計

本章節依據研究動機與目的提出研究架構。本研究主要的概念為透過替團購商品分群，以獲取團購商品隱含之商品特性。經過文獻探討之評估分析，本研究將透過文字探勘蒐集團購商品相關資料，接著利用『k 最近鄰居法』的運作原理建置 kNN 分群器來進行群集分析。kNN 分群器在替每一項商品進行分群的時候，是以該商品的商品特徵詞作為分群依據，特徵詞相似程度高的歸於同一群。分群完後的結果，每一個群集有著不同的商品特徵詞，便可依據其特有的產品特徵詞去解釋該群的商品特質。

第一節 研究架構

本研究採用中央研究院中文詞知識庫小組所研發的中文斷詞系統，作為資料斷詞之用。資料經過斷詞處理、雜訊過濾以及自訂門檻條件篩選等整理過程後所剩餘的特徵詞即為具有代表性與意義的產品關鍵字詞。運用 kNN 分群器依照產品特徵詞的相似程度替商品進行分群動作，以產生分群結果並進行評估。每一群群集即代表擁有不同產品特徵的團購商品族群。研究架構圖如圖 3-1 所示。



(資料來源：本研究整理)

圖 3-1 研究架構圖

第二節 資料處理

3.2.1 蒐集資料

資料來源部分，本研究以國內知名團購網站「愛合購」(ihergo)的美食分類下之甜點蛋糕類商品為主，依熱門買氣進行排序，鎖定前 1000 項團購美食商品，並以每一項商品為基礎，從 Google 搜尋 Bar 鍵入該團購商品名稱關鍵字，以找尋曾團購此商品的顧客網誌文章並存入資料庫。在這裡附加說明的是，之所以鎖定以團購方式進行購買的商品文章是因為團購為多數人一起進行的行動，就商品特性而言也是獲得大家認同才會一起進行購買，因此團購網誌的文章內容較能代表大眾的口味。

在鎖定的 1000 項美食團購商品中，逐項鍵入 Google 搜尋 Bar 找尋團購此商品的相關網誌後，其中有 268 項產品曾有網友撰寫過團購網誌。在網誌擷取的部份，由於一項產品可能有多位網友撰寫過團購網誌，因此本研究將依照 Google 搜尋 Bar 找尋到的順序來納入網誌，Google 搜尋引擎的排序越前面的網頁通常是熱門度與相關性較高之網頁，經過觀察數個商品的搜尋狀況後，在排序第 3 頁之後的網頁和研究所需之網誌資料較無關聯，因此決定網誌的納入範圍設定為搜尋引擎前 3 頁的網誌資料。本研究就 268 項擁有團購網誌的商品進行蒐集，共納入了 586 篇顧客團購網誌，並將顧客團購網誌以商品為基礎，相同商品的網誌會集結成為該商品的團購網誌。

3.2.2 中文斷詞

將蒐集到的顧客團購商品文章資料進行中文斷詞處理，以利研究後續特徵詞的萃取。本研究採用中央研究院中文詞知識庫小組(Chinese Knowledge Information Processing Group, CKIP)所開發的中文斷詞服務系統來進行處理。在經過中文斷詞處理後，輸出的資料皆具有 CKIP 的詞性標記。在進行特徵詞萃取前，為了避免影響分析的成效，透過中研院平衡語料庫詞類標記進行篩選，刪除斷詞後不必要的詞性，僅保留研究所需用詞之詞性。

下圖為擷取一段網友所撰寫的團購網誌內容：

美式重乳酪蛋糕

重乳酪蛋糕紮實口感，具濃郁的乳酪香氣，下層的餅乾體也很香鬆，真的很好吃，單吃也不會感到膩，會停不下來一直吃哩。是我期待中的好吃重乳酪蛋糕口感，很值得回購的一款好吃蛋糕。

(資料來源：本研究整理)

圖 3-2 顧客團購網誌(斷詞前)

此段團購網誌內容經過 CKIP 斷詞後，會在各個詞語的後面加上該詞之詞性，如下圖所示：

美式重乳酪蛋糕

重(Vi) **乳酪(N)** **蛋糕(N)** **紮實(Vi)** **口感(N)** ，(COMMACATEGORY)
具(Vt) **濃郁(Vi)** 的(T) **乳酪(N)** **香氣(N)** ，(COMMACATEGORY)
下(DET) 層(M) 的(T) **餅乾體(N)** 也(ADV) 很(ADV) **香(Vi)** **鬆(Vt)**
，(COMMACATEGORY) 真的(ADV) 很(ADV) **好吃(Vi)**
，(COMMACATEGORY) 單(ADV) **吃(Vt)** 也(ADV) 不會(ADV)
感到(Vt) **膩(Vi)** ，(COMMACATEGORY) 會(ADV) **停(Vt)** 不(ADV)
下來(Vi) 一直(ADV) **吃(Vt)** 哩(T) 。(PERIODCATEGORY) **是(Vt)**
我(N) **期待(Vt)** 中(POST) 的(T) 好(ADV) **吃重(Vi)** **乳酪(N)** **蛋糕(N)**
口感(N) ，(COMMACATEGORY) 很(ADV) **值得(Vi)** **回購(Vt)** 的(T)
一(DET) 款(M) **好吃(Vi)** **蛋糕(N)** 。(PERIODCATEGORY)

(資料來源：本研究整理)

圖 3-3 顧客團購網誌及詞性表示(斷詞後)

在本研究中，由於產品特徵詞多以名詞、動詞以及形容詞等詞性呈現，因此設立資料庫所保留的詞彙為精簡詞類標記之詞性 N、Vt、Vi 以及 A，其他非上述詞性的詞語將以過濾的方式排除。上圖以粗體字顯示之詞語即為研究中的保留字詞。

3.2.3 特徵詞萃取

為了能更精準的取出該產品的特徵詞彙，特徵詞出現次數多寡與頻率高低是一項重要的參考數據，可依字詞的重要程度過濾出常見的詞語，並自訂門檻對其進行篩選，以保留重要的特徵詞語。經篩選後所得之特徵詞即為具有代表性與意義的關鍵字詞，再以這些字詞作為分群依據。

在字詞的重要程度衡量部分，本研究採用的是最常用於計算字詞權重的 TF-IDF (Term Frequency–Inverse Document Frequency) 衡量方式，TF-IDF 傾向於過濾掉常見的詞語，以保留重要的詞語。其公式如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{2}$$

其中，

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

$$idf_i = log\left(\frac{N}{df_i}\right) \tag{4}$$

$tfidf_{i,j}$為字詞 i 在文件 j 的權重值，其值為$tf_{i,j} \times idf_i$。$tf_{i,j}$為字詞 i 在文件 j 中出現的頻率，其中$n_{i,j}$是字詞 i 在文件 j 中的出現次數，而 k 為文件 j 的總字詞數(文章長度)。idf_i為字詞 i 的逆向文件頻率（Inverse Document Frequency，IDF），其值可由總文件數目(N)除以包含字詞 i 之文件的數目(df_i)，再將得到的商取對數(log)得到。總括來說，當該詞語在某特定文件內屬於高詞語頻率，且在整個文件集合中屬於低文件頻率，便可產生出高權重的 TF-IDF。

計算完字詞的權重後，對於字詞的重要度便有了衡量的依據。一篇文章中，TF-IDF 值越高的詞彙，代表其重要性越高，極有可能為具有代表性的特徵詞，反之，TF-IDF 值越低的詞彙，可能為對文件沒有識別能力的常見字詞，抑或是容易造成分群干擾的雜訊。

本研究訂定了特徵詞選取的門檻值，來決定選取多少比例的特徵詞，以找出重要特徵詞彙。藉由篩選詞彙的動作，來提升分群品質，也降低往後進行分群時必須建置之「詞彙-文件矩陣」其維度複雜度，以增加分群執行效率。在進行特徵詞選取的時候，會先依照該文章中所有詞彙的 TF-IDF 值由高到低進行排序，接著再依據門檻值取出 TF-IDF 值前百分之 n 的詞彙，值得注意的是，由於一篇

文章中會出現許多詞彙皆擁有相同之 TF-IDF 值，因此在選取特徵值時，需先去找尋符合最低門檻值詞彙其 TF-IDF 值為多少，再將所有與其相等及大於該 TF-IDF 值的詞彙全部取出，因此取出的特徵詞彙個數占該文章詞彙總數的百分比例會大於門檻訂定之值。

此外，雖然每篇文章訂定的特徵詞選取門檻值皆是相同的，但是每篇文章真正取出的特徵詞彙數目卻因該文章通過最低門檻值的詞彙個數而有所不同，因此為了使所有文章的特徵詞都立於相同的比較基準上，必須對選取到的特徵詞彙之 TF-IDF 值進行調整。其調整方式為，特徵詞的 TF-IDF 值會依據該篇文章選中的特徵詞總數進行正規化，以獲得該特徵詞調整後的權重值。根據特徵詞總數調整權重的概念就如同根據每篇文章的長度不同而進行調整權重的概念意義相同，每篇文章所選取的特徵詞彙總數即代表該篇文章的長度。特徵詞彙的正規化權重調整公式如下：

$$W_{i,j} = \frac{tfidf_{i,j}}{\|\overrightarrow{d_j}\|} \tag{5}$$

上述公式之意涵為將該詞彙之$tfidf_{i,j}$值除以所有選中的特徵詞彙長度$\|\overrightarrow{d_j}\|$，其中$\|\overrightarrow{d_j}\|$代表該文件向量中所有權重各別平方加總再開根號(在這裡的所有權重為該文件所有被選取之特徵詞彙的 TF-IDF 值)，最後得到的$W_{i,j}$值即某一特徵詞正規化後的權重。

第三節 kNN 分群

以下將介紹本研究所採用的分群機制－k 最近鄰居法。首先將針對 kNN 分群器的運作原理介紹，接著闡述詞彙-文件矩陣的建置以及文件之間相似度的衡量方式，最後說明 kNN 分群器之相關參數設定。

3.3.1 kNN 分群器運作原理

kNN 分群器是利用 k 最近鄰居法「物以類聚」的概念，依照每一項團購商品的產品特徵詞間的相似度來替商品進行分群的動作。在 k 最近鄰居法中，計算資料點與資料點之間的距離，常用的方式是採用歐幾里德距離來計算特徵值間的差距，同樣的概念應用於文件分群的向量空間模型中，則是衡量文件與文件間的相似程度，也就是計算兩文件在 n 維空間的角度差距。在取得文件與其他文件的相似程度後，便可根據與該文件前 k 個相似的文件其所屬群集來歸納此文件應分屬至哪個群集中。

3.3.2 建置詞彙-文件矩陣

在 kNN 分群器進行分群之前，必須建置詞彙-文件矩陣，以便於計算文件間的相似度。「詞彙-文件矩陣」為文件與詞彙之權重對應矩陣表，其主要目的為將之前計算所得知每份文件的特徵詞及其權重從向量空間模型轉化為以單位向量的方式呈現。一般來說，詞彙-文件矩陣的欄即代表總詞庫中每個詞彙，列則代表文件集中的每份文件，而矩陣內容元素則為詞庫與文件對應之詞彙權重值。就本研究而言，詞彙-文件矩陣的欄並非總詞庫的所有詞彙，而是有被任一文件選取為特徵詞之詞彙才會陳列於此，透過這樣的過濾機制，可以排除沒有被任何文件選為特徵詞之詞彙，以精簡矩陣，降低運算次數。

就矩陣元素內容來看，當詞彙在此文章中被選為特徵詞時，其矩陣元素則擺放該詞彙調整後之權重值(weight)，若該詞彙並非為此文章所有之特徵詞，其矩陣元素為 0。

3.3.3 文件相似度計算

詞彙-文件矩陣建置完善後，便可依據其矩陣元素計算文件之間的相似程度。在文件分群的向量空間模型中，計算文件間的相似程度最常用的衡量方式為計算其餘弦相似度(Cosine Similarity)(Salton，1989)，其公式如下：

$$\cos(\theta) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}} \quad (6)$$

其以兩個 n 維向量間的角度差異來度量該向量間的距離，而計算所得之結果將介於 0~1 之間，當兩份文件的向量間角度越相近時，其夾角越小，所求得之計算結果越接近 1，代表兩份文件越相似；反之，則計算結果越接近 0，代表兩文件越不相似。

3.3.4 kNN 分群器

本研究依據 k 最近鄰居法的運行概念，以 PHP 及 mySQL 建置 kNN 分群器，並在 PHP Command Line Interface (CLI)環境下運作執行。在 kNN 分群器進行分群的過程中，除了需得知文件與文件之間的相似度以外，還必須決定以下兩項參數的設定，即 k 值與相似度的分群門檻值(threshold)。

「k 值」為決定文件所需參考的相近文件個數，一般 k 值通常介於 1~20 之間較為合適，且奇數比偶數好。k 值的訂定會影響分群結果的品質，如果 k 值選擇過小，則得到的參考文件數過少，對周遭文件太過敏感，也容易放大噪音數據的干擾，降低分群的準確度；倘若 k 值選擇過大，則容易將文件分屬到文件集中出現頻率很高的群集，而非將文件歸屬到特質相近的群集。

「門檻值」的意涵在於幫助分群的判斷。一個未分群的文件進入時，會依照訂定的 k 值取出取其相似度最為接近的 k 份文件，判斷這 k 份文件的所屬群集，便可得知該未分群文件與每個群集間的相似程度，並將該文件歸屬於和它相似度最相近之群集中。不過，倘若與其最相似之群集其相似程度都低於門檻值，則判斷該未分群文件不隸屬於目前存在的任何群集，應成立新的群集。另外，門檻值的高低對分群的影響很大，過低的門檻值不容易獲得好的分群效果，而過高的門檻值容易產生分群過度(Overfitting)的現象導致分群品質降低。

在未分群文件與其 k 個相近文件之相似度比較的過程中，其進行方式是把 k 份相近文件依照其所屬群集各別將相似度加總，相加結果相似度數值最高之群集即為該未分群文件之所屬群集。(若其相似值高於門檻值)

3.3.5 分群規則

本研究所制定之分群規則將依 kNN 分群器的參數設定、二次分群及合併，以及分群結束時點作說明：

1. 在 kNN 分群器的參數設定部分，每個分群階段會在不同 k 值下測試 3 種門檻值來觀察群集的分群狀況，以選取加權平均群內相似度最佳之分群結果做為該次 kNN 分群運作的參數設定。本研究設定 k 值由 5 開始，以 5 為單位遞增，若 k 值遞增後進行分群的群內相似度可較遞增前增加，則再次提高 k 值進行分群測試(本研究限制 k 值最大遞增至 15 為止)。另外門檻值變動的方式為以 0.01 為單位往上增加。
2. 二次分群為將大群集提出，單獨提高其分群門檻值來進行再次分群的動作，以避免為了將大群集進行分群而影響其他群集的聚合程度。本研究將大群集定義為群集內商品個數為 20 項產品以上時，則將該群集列為候選大群集。

3. 當出現一個群集中只有單項產品便會進行合併的動作，該項產品會合併於與其最相近之群集中，其相近程度的衡量方式是衡量群與群間的質心相似度。而可能合併的候選群集為與該群集同一階層的群集才會納入考量。
4. 分群的結束時點為：提出大群集進行下一階段分群時，若分群後的加權平均群內相似度成長幅度過小，或是分群無法產生分群效果時，則停止該大群集的分群動作。本研究訂定成長幅度過小為成長率小於 20%。

第四節 評估方法

在分群結果衡量部分，除了以人工方式檢視群集的特性是否符合群內特質相近且群間特質差異外，將透過群內相似度(Intra-Similarity)的方式來加以驗證。

群內相似度為計算群集內團購商品網誌間之相似程度，其比較方式為兩兩商品網誌進行比較，以計算彼此間之餘弦相似度，餘弦相似度之計算公式如前述章節「3.3.3 文件相似度計算」所示。當兩兩商品文件比較完之餘弦相似度累計加總後除以總共的比較次數，即為該群的群內相似度，其結果值落於 0~1 之間。計算公式如下：

$$C_k\left(\textit{群內相似度}\right) = \frac{\sum_{d_i \in C_k} \sum_{d_j \in C_k} sim(\overrightarrow{d_i}, \overrightarrow{d_j})}{N_k \times (N_k - 1) \times \frac{1}{2}} \quad (7)$$

其中，N_k為第C_k群的商品文件數量，$sim(\overrightarrow{d_i}, \overrightarrow{d_j})$為$C_k$群內兩篇商品文件之餘弦相似度。

在得知每一群的群內相似度後，將所有群集的群內相似度加總並除以總群數(C)，其所獲得之數值便為平均群內相似度(Mean of Intra-Similarity,MIS)，結果值

越接近 1 代表平均群內相似度越高，分群效果越好，其公式如下所示：

$$\text{MIS} = \frac{\sum C_k\left(\textit{群內相似度}\right)}{C} \tag{8}$$

一般情況下，小群集的群內相似度會高於大群集的群內相似度。在研究中，為了避免每一階段的分群決策受到小群集的群內相似度影響過大，因此採用加權式的平均群內相似度，且衡量範圍侷限於該階段的群集，即尚未分群時的候選大群集(父群集)以及分群後所產生的群集(子群集)。加權平均群內相似度的計算方式為：各個子群集的群內相似度乘以該群的產品個數占父群集總產品數的比例後再進行加總。

第四章　實驗結果

本研究將可蒐集到顧客團購網誌的 268 項熱門團購產品做為 kNN 分群器的輸入資料來源。進行完 CKIP 斷詞處理後，在特徵詞選取的門檻值部份，設立了 40%、60%以及 80%等 3 種不同的選取比例來進行試驗，經實驗觀察實際選取到的特徵詞彙，在選取比例為 40%時，許多產品相關的重要詞彙皆未選取到；在選取比例為 80%時，所有詞彙幾乎全部選取，因此無法產生詞彙過濾的效果;而在選取比例為 60%時，可選取出大部分的商品特徵詞彙，也可過濾掉 TF-IDF 值權重過低之詞彙。因此，本研究決定採用的特徵詞彙選取比例為 60%，選取詞彙權重前 60%的特徵詞彙做為 kNN 分群器的分群依據。在 kNN 分群器的建置部分，本研究依循其原理建置基本分群器，並實驗不同 k 值及門檻值搭配的狀況所產生的分群。以下將說明 268 項團購美食商品進行 kNN 分群器運行之分群過程及結果呈現。

第一節 各階段分群結果

以下將以表格呈現產品分群過程，並以文字說明為輔。

4.1.1　第一階段分群

下表為第一階段分群結果。未分群(只有一群)的群內相似度為 0.02983405。分群門檻值以 0.01 為單位提升，在門檻值為 0.01 時便有群集產生，擁有分群效果。首先實驗 k 值為 5，搭配三種幅度的門檻值進行分群，其結果擁有分群效果，於是再實驗 k 值為 10，搭配同樣的三種幅度門檻值進行分群，其結果也擁有分群效果，接著先比較分群結果，在上述的六種參數設定下(兩種不同 k 值個別搭配三種分群門檻值)，以 k 值為 5 以及門檻值為 0.03 時，所獲得分群結果的加權

平均群內相似度最高，由於 k 值為 10 的最佳分群結果並不優於 k 值為 5 的最佳分群結果，因此 k 值不再往上遞增至 15 來進行實驗，而本階段將採用 k 值為 5 以及門檻值為 0.03 的參數設定下所獲之分群結果(即灰色網底部分所示)。「合併次數」為單項產品群集的合併次數，而「分得群數」為該階段合併完後的群數。

表 4-1 第一階段分群

未分群：268 項產品				(群內相似度：0.029834)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.01	-	2	0.040112463	34.452%	
	0.02	-	2	0.036004996	20.684%	
	0.03	2	3	0.042315799	41.837%	
10	0.01	-	2	0.030124388	0.973%	
	0.02	-	2	0.036827896	23.442%	
	0.03	2	3	0.04037569	35.334%	

(資料來源：本研究整理)

經過第一階段分群後，可將 268 項產品分成 5 群，其中有 2 群為單項產品群集，經合併後，可獲得 3 個群集。群集 1 含有 87 項產品，群內相似度為 0.032191，群集 2 含有 157 項產品，群內相似度為 0.041533，群集 3 含有 24 項產品，群內相似度為 0.084139。此階段分群完畢後所獲得的 3 個群集皆為大群集，因此皆需進行第二階段分群。

4.1.2 第二階段分群

(一) 第二階段分群：群集 1

下表為群集 1 進行第二階段分群。由於在 k 值為 5 及 k 值為 10 的分群結果比較中，k 值為 10 的最佳分群結果與 k 值為 5 的最佳分群結果相同並無改善，因此 k 值不再往上遞增至 15 來進行實驗。

表 4-2 第二階段分群 ー 群集 1

第二階段分群 ー 群集 1：87 項產品				(群內相似度：0.032191)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.04	1	2	0.037073437	15.167%	
	0.05	1	3	0.051097793	58.733%	
	0.06	1	3	0.051097793	58.733%	
10	0.04	1	2	0.036831034	14.414%	
	0.05	1	3	0.051097793	58.733%	
	0.06	1	3	0.051097793	58.733%	

(資料來源：本研究整理)

群集 1 經過第二階段分群後，可分為 4 個子群集(群集 4～群集 7)，由於群集 5 為單項產品群集，因此需先衡量其與群集 4、群集 6 以及群集 7 的質心相似度，計算完得知群集 5 與群集 4 的質心相似度最高，因此將群集 5 的產品合併至群集 4 中。經質心合併的動作處理後，剩下 3 個子群集，其群內產品數目分別為群集 4：81，群集 6：2，群集 7：4。其中群集 4 為候選大群集，需進行第三階段分群，而群集 6 與群集 7 則不再變動，為分好群的狀態(即葉節點)。下面列出葉節點群集的狀態，並加以描述其擁有之產品品項及特徵詞彙，並替該群集命名。

在群集產品特徵詞彙的選取部分，其擷取範圍為該群出現過的詞彙中詞彙權重前 20%的字詞，這裡的詞彙權重為平均後的權重，亦即該群中某詞彙在該群各個商品網誌中的權重值加總後除以商品個數所獲得的平均值。取出權重值前 20%的詞彙後，依人工過濾並考量其出現在商品中的頻率，進而挑選出該群的產品特徵代表詞。

表 4-3 群集表：群集 6

群集 6：獨創層次感捲包牛角類製品 (2 項商品)
產品特徵詞(平均權重值)
商品主體：牛角(0.831)、金牛角(0.090) 形狀：捲包(0.064)、層次(0.043) 商家：角之館(0.176) 其他：獨創(0.058)
產品名稱
角之館三峽金牛角(焦糖瓦片)、樹林香脆牛角棒

(資料來源：本研究整理)

表 4-4 群集表：群集 7

群集 7：杏仁顆粒口感甜品 (4 項商品)
產品特徵詞(平均權重值)
商品主體：麻花(0.327) 成份：杏仁(0.490) 口感：黏牙(0.090)、顆粒(0.052)、甜滋滋(0.049)
產品名稱
今日蜜麻花、今日杏仁香片、塔吉特摩卡杏仁千層蛋糕、爆料奶酪

(資料來源：本研究整理)

(二) 第二階段分群：群集 2

下表為群集 2 進行第二階段分群。由於在 k 值為 5 及 k 值為 10 的分群結果比較中，最佳的分群結果是出現於 k 值為 5 下(搭配門檻值為 0.06 時)，因此 k 值不再往上遞增至 15 來進行實驗。

表 4-5 第二階段分群 ー 群集 2

第二階段分群 ー 群集 2：157 項產品				(群內相似度：0.041533)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.04	1	1	0.041533	0.000%	
	0.05	1	6	0.090680102	118.333%	
	0.06	1	10	0.128987803	210.567%	
10	0.04	1	1	0.041533	0.000%	
	0.05	1	6	0.093460554	125.027%	
	0.06	1	10	0.126863662	205.453%	

(資料來源：本研究整理)

群集 2 經過第二階段分群後，可獲得 10 個群集，其群內產品數目分別為群集 8：25，群集 9：13，群集 10：2，群集 11：20，群集 12：28，群集 13：43，群集 14：10，群集 16：6，群集 17：7，群集 18：3。其中群集 8、群集 12 以及群集 13 為候選大群集，需進行第三階段分群，其餘群集則為不再變動的葉節點群集。下面列出各葉節點群集的群集表。

表 4-6 群集表：群集 9

群集 9：濕潤內餡滑嫩口感布蕾類製品 (13 項商品)
產品特徵詞(平均權重值)
商品主體：布蕾(0.589)、蛋糕(0.099) 成份：巧克力(0.222)、焦糖(0.0629)、雞蛋(0.074) 口感：香醇(0.074)、滑嫩(0.059)、軟(0.067)、濕潤(0.039) 特色：內餡(0.040)、麻糬(0.071)
產品名稱
達克闇黑工場半熟蛋糕、原味蛋糕布蕾、巧克蛋糕布蕾、巧克蛋糕布蕾派、心太軟(巧克力)、心太軟(起司)、手工蛋糕布朗尼、焦糖布蕾堡、約瑟芬冰塔(鮮奶布蕾)、原味鮮奶布蕾、可可雞蛋布蕾捲、米迦原味布蕾派、鮮奶香醇布蕾派

(資料來源：本研究整理)

表 4-7 群集表：群集 10

群集 10：精美包裝手工奶油蛋糕 (2 項商品)
產品特徵詞(平均權重值)
商品主體：蛋糕(0.206) 成份：奶油(0.090)、香料(0.085) 特色：手工(0.142)、包裝(0.064)
產品名稱
拿破崙蛋糕(經典原味)、荷蘭貴族手工蛋糕

(資料來源：本研究整理)

表 4-8 群集表：群集 11

群集 11：內含新鮮水果 QQ 奶凍類製品 (20 項商品)
產品特徵詞(平均權重值)
商品主體：奶凍(0.250)、蛋糕(0.085) 成份：慕斯(0.099)、鮮奶(油)(0.048) 口味：草莓(0.653)、芒果(0.324)、巧克力(0.103)、水蜜桃(0.063) 口感：奶味(0.057)、QQ(0.036)、冰淇淋(0.033) 特色：新鮮(0.097)
產品名稱
北海道雙層草莓蛋糕、日式草莓奶凍、黑丸嫩仙草、佳樂波士頓派、塔吉特芒果奶凍千層蛋糕、芒果三明治、草莓三明治、草莓卡樂、貝里貝果、鮮果雪藏、天使水果捲、手工玫瑰黑泡芙(粉嫩草莓)、提拉奶凍、日式大福、日式巧克力奶凍、維也納紅豆牛奶麵包、芒果奶酪、草莓巧克力蛋糕、草莓慕斯、蔓越草莓慕斯

(資料來源：本研究整理)

表 4-9 群集表：群集 14

群集 14：微苦細緻提拉米蘇 (10 項商品)
產品特徵詞(平均權重值)
商品主體：提拉米蘇(0.553) 成份：咖啡(0.339)、可可粉(0.269)、慕斯(0.102) 、乳酪(0.066) 口感：苦(0.070)、細緻(0.039) 特色：餅乾層(0.075)
產品名稱
花蓮提拉米蘇、塔吉特義式提拉千層蛋糕、塔吉特英式伯爵千層蛋糕、觀音愛心家園提拉米蘇、金莎提拉米蘇、咖啡提拉米蘇、咖啡核桃瑞士捲、夏雪波士頓派(咖啡) 、皇家提拉米蘇、義式經典提拉米蘇

(資料來源：本研究整理)

表 4-10 群集表：群集 16

群集 16：乾澀鬆散桂圓製品 (6 項商品)
產品特徵詞(平均權重值)
商品主體：蛋糕(0.119)、麵包(0.109) 成份：桂圓(0.830)、核桃(0.124)、龍眼(0.117)、 口感：乾(0.089)、鬆散(0.044)、
產品名稱
桂圓蛋糕、紅酒桂圓麵包、酒釀桂圓冠軍麵包、冰沁桂圓、土雞蛋桂圓蛋糕、奕順軒桂圓蛋糕

(資料來源：本研究整理)

表 4-11 群集表：群集 17

群集 17：甜蜜濕潤爆漿口感蜂蜜蛋糕 (7 項商品)
產品特徵詞(平均權重值)
商品主體：蛋糕(0.116) 成份：蜂蜜(0.809)、雞蛋(0.037) 口感：濕潤(0.066)、甜蜜(0.057) 特色：爆漿(0.069)、液體(0.035) 製程：烘焙(0.074)
產品名稱
凹蛋糕(原味蜂蜜)、經典原味半熟蜂蜜蛋糕、蛋糕工廠蜂蜜蛋糕、凹蛋糕(蜂蜜檸檬) 、朱古力半熟蜂蜜蛋糕、牽絲太陽餅、聖淘沙蜂蜜捲(鮮奶咖椰)

(資料來源：本研究整理)

表 4-12 群集表：群集 18

群集 18：爽脆青蔥鹹蛋糕 (3 項商品)
產品特徵詞(平均權重值)
商品主體：蛋糕(0.093) 成份：蔥(0.449)、油蔥(0.234)、脆筍(0.208)、瘦肉(0.138)、美乃滋(0.079) 口感：鹹(0.281)、脆(0.056)
產品名稱
宜蘭三星蔥捲、桂夫人鹹蛋糕、豐原鹹蛋糕

(資料來源：本研究整理)

(三) 第二階段分群：群集 3

下表為群集 3 進行第二階段分群。由於在 k 值為 5 及 k 值為 10 的分群結果比較中，k 值為 10 的最佳分群結果與 k 值為 5 的最佳分群結果相同並無改善，因此 k 值不再往上遞增至 15 來進行實驗。

表 4-13 第二階段分群 ー 群集 3

第二階段分群 ー 群集 3：24 項產品				(群內相似度：0.084139)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.04	-	2	0.111224667	32.192%	
	0.05	-	2	0.111224667	32.192%	
	0.06	1	3	0.125546542	49.213%	
10	0.04	-	2	0.111224667	32.192%	
	0.05	-	2	0.111224667	32.192%	
	0.06	1	3	0.125546542	49.213%	

(資料來源：本研究整理)

群集 3 經過第二階段分群後，可獲得 3 個群集，其群內產品數目分別為群集 19：19，群集 20：2，群集 21：3，所有群集皆為葉節點群集。下面列出各群集的群集表。

表 4-14 群集表：群集 19

群集 19：香甜焦糖布丁口感類製品 (19 項商品)
產品特徵詞(平均權重值)
商品主體：布丁(0.702)、泡芙(0.176) 成份：焦糖(0.273)、香草(0.083)、鮮奶(0.083)、楓糖(0.070) 口感：香甜(0.040)、酥皮(0.046)、QQ(0.039)、綿密(0.032) 製程：烤(0.051) 保存：冷藏(0.043)
產品名稱
新美珍布丁蛋糕、李記焦糖烤布丁、和芙子脆皮泡芙、米其林葡式蛋塔、法式楓糖烤布丁、伊恩焦糖布丁、回憶香雞蛋布丁、康鼎丹比鮮奶布丁、日式芙蓉鮮奶泡芙、純手工戚風雞蛋布丁蛋糕、純粹紫米米布丁、雞蛋烤布蕾、黃金泡芙(酥皮地瓜) 、你我他之家燒烤布丁、布丁哥哥焦糖烤布丁、帕瑪森布丁捲、法式長泡芙(原味) 、熊本布蕾塔、連珍楓糖烤布丁

(資料來源：本研究整理)

表 4-15 群集表：群集 20

群集 20：紮實鳳梨纖維內餡製品 (2 項商品)
產品特徵詞(平均權重值)
商品主體：鳳梨酥(0.542) 成份：鳳梨(0.392) 口感：纖維(0.134)、紮實(0.052) 特色：內餡(0.073)
產品名稱
小潘鳳凰酥、山腳傳奇土鳳梨酥

(資料來源：本研究整理)

表 4-16 群集表：群集 21

群集 21：健康低負荷麵粉類製品 (3 項商品)
產品特徵詞(平均權重值)
成份：麵粉(0.075) 口感：綿(0.064) 特色：健康(0.175)、膽固醇(0.163)、蛋白(0.156)
產品名稱
順謚健康蛋糕(檸檬原味) 、天使蛋糕(原味) 、鹿港兔仔寮牛舌餅

(資料來源：本研究整理)

第二階段分群完畢後所獲得的群集 4、群集 8、群集 12 以及群集 13 為候選大群集，因此皆需進行第三階段分群。

4.1.3 第三階段分群

(一) 第三階段分群：群集 4

下表為群集 4 進行第三階段分群。由於在 k 值為 5 及 k 值為 10 的分群結果比較中，最佳的分群結果是出現於 k 值為 10，因此 k 值將再往上遞增至 15 來進行實驗。

表 4-17 第三階段分群 — 群集 4

第三階段分群 — 群集 4：81 項產品				(群內相似度：0.03458)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.07	2	4	0.070475741	103.805%	
	0.08	2	6	0.081352679	135.259%	
	0.09	3	7	0.094330926	172.790%	
10	0.07	2	4	0.071958222	108.092%	
	0.08	2	6	0.081352679	135.259%	
	0.09	3	7	0.094833679	174.244%	
15	0.07	2	4	0.071958222	108.092%	
	0.08	2	6	0.081352679	135.259%	
	0.09	3	7	0.094833679	174.244%	

(資料來源：本研究整理)

群集 4 經過第三階段分群後，可獲得 7 個群集，其群內產品數目分別為群集 22：40，群集 23：7，群集 24：19，群集 27：3，群集 28：3，群集 30：7，群集 31：2。其中群集 22 為候選大群集，需進行第四階段分群，其餘群集則為不再變動的葉節點群集。下面列出各葉節點群集的群集表。

表 4-18 群集表：群集 23

群集 23：營養餡料起司酥皮類製品 (7 項商品)
產品特徵詞(平均權重值)
成份：起司(0.126)、麵粉(0.063)、火腿(0.107) 口感：酥皮(0.241)、香濃(0.045) 處理方式：烤箱(0.123) 特色：營養(0.056)、餡料(0.046)
產品名稱
起酥火腿三明治、夠 PIZZA 義氏總匯(千層酥皮) 、京都起士塔、皇冠芋心肉鬆麵包、芝玫起酥蛋糕、豪華海陸披薩、起司雞肉捲

(資料來源：本研究整理)

表 4-19 群集表：群集 24

群集 24：香濃吐司類製品 (19 項商品)
產品特徵詞(平均權重值)
商品主體：吐司(0.550)、厚片(0.168) 成份：奶酥(0.228)、葡萄乾(0.151)、鮮奶(0.140) 口感：酥(0.073)、QQ(0.0513)、香(0.041)
產品名稱
甜在心手工厚片土司(香蒜)、貴客 PIZZA 和風章魚燒(奶香千層披薩) 、手工烘焙鮮奶厚片、茲蘭厚片土司、舞 Q 甜甜圈、手工紅蘿蔔厚片吐司、義珍香鮮奶吐司、全麥葡萄司康、湯種鮮奶全麥吐司、膠原鮮奶涼糕(芝麻) 、湯種鮮奶奶酥吐司、土雞蛋麵包烤布丁、奶酥大司康、沖繩黑糖吐司厚片、火花滋滋叫、無毒地瓜吐司、茶香甘藷、黃金法式雜糧麵包、黑糖麻糬土司

(資料來源：本研究整理)

表 4-20 群集表：群集 27

群集 27：膨鬆奶油戚風蛋糕 (3 項商品)
產品特徵詞(平均權重值)
商品主體：戚風(0.367) 成份：奶油(0.201)、卡士達醬(0.122) 口感：膨(0.130)、鬆(0.091) 特色：北海道(0.289)
產品名稱
北海道戚風蛋糕、北海道鮮奶戚風蛋糕、北海道 MINI 小戚風

(資料來源：本研究整理)

表 4-21 群集表：群集 28

群集 28：果泥年輪蛋糕 (3 項商品)
產品特徵詞(平均權重值)
商品主體：蛋糕(0.111) 成份：蘋果(0.424)、果泥(0.091) 特色：年輪(0.773)、味蕾(0.039)
產品名稱
年輪蛋糕、草莓蘋果年輪蛋糕、提拉米蘇蘋果年輪

(資料來源：本研究整理)

表 4-22 群集表：群集 30

群集 30：酸甜優格內餡製品 (7 項商品)
產品特徵詞(平均權重值)
成份：優格(0.549)、蔓越莓(0.412)、橙皮(0.308)、乳酪(0.069) 口感：酸(0.124)、酸甜(0.069)、 特色：天然(0.061)、餡料(0.069)
產品名稱
無油無糖全麥麵包(優格蔓越莓) 、無油無糖全麥麵包(優格香橙) 、無油無糖全麥麵包(巧克力香橙) 、高大活菌鮮奶優格、黑鑽巧克力捲、粉紅佳人(紅酒蔓越莓優格) 、蔓越莓提拉米蘇

(資料來源：本研究整理)

表 4-23 群集表：群集 31

群集 31：綿軟奶香麵皮類製品 (2 項商品)
產品特徵詞(平均權重值)
商品主體：饅頭(0.299) 成份：麵皮(0.534) 口感：綿軟(0.196)、膨鬆(0.149)、奶香(0.109)、清淡(0.074)
產品名稱
奶油銀絲卷、詠宸鮮奶饅頭

(資料來源：本研究整理)

(二) 第三階段分群：群集 8

下表為群集 8 進行第三階段分群。

表 4-24 第三階段分群 ー 群集 8

第三階段分群 ー 群集 8：25 項產品				(群內相似度：0.111189)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.07	-	1	0.111189	0.000%	
	0.08	1	2	0.13110456	17.911%	**小於 20%**
	0.09	2	2	0.13110456	17.911%	

(資料來源：本研究整理)

群集 8 經過分群後，最佳狀態的加權平均群內相似度成長率為 17.911%，小於研究訂定的標準 20%，因此不進行分群動作，k 值也不往上遞增至 10 進行實驗。下表為群集 8 的群集表。

表 4-25 群集表：群集 8

群集 8：QQ 芋頭綿密糕點類製品 (25 項商品)
產品特徵詞(平均權重值)
商品主體：涼糕(0.267)、蛋糕(0.088)、蛋糕捲(0.142) 成份：芋頭(0.657)、芋泥(0.318)、鮮奶(0.196) 口感：顆粒(0.081)、QQ(0.060)、奶香(0.040)、綿密(0.033) 特色：餡(0.055)
產品名稱
金莎巧克力卷、香帥芋泥捲、真芋泥捲、連珍芋頭球、拿破崙蛋糕(鮮奶芋頭)、提拉麻吉蛋糕捲、約瑟芬冰塔(鮮奶芋頭)、芋頭酥皮泡芙、芋頭鮮奶油波士頓、蓮藕夾心糕、飛碟蛋糕(橙桔口味)、鮮烤芋頭捲、鮮芋頭蛋糕、千巧谷鮮奶酪、波士頓捲、甜蜜水蜜桃波士頓、膠原鮮奶涼糕(芋頭)、膠原鮮奶涼糕(花生)、芋泥布蕾塔、芋頭蛋糕、芝心芋捲、雪梅娘、雪藏起司棒、鮮奶涼糕(乳酪)、鮮奶涼糕(起士)

(資料來源：本研究整理)

(三) 第三階段分群：群集 12

下表為群集 12 進行第三階段分群。

表 4-26 第三階段分群 ─ 群集 12

第三階段分群 ─ 群集 12：28 項產品				(群內相似度：0.108372)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.07	-	1	0.108372	0.000%	
	0.08	1	1	0.108372	0.000%	
	0.09	1	1	0.108372	0.000%	

(資料來源：本研究整理)

群集 12 經過分群後，無分群效果產生，因此停止此群集的分群動作。下表為群集 12 的群集表。

表 4-27 群集表：群集 12

群集 12：紮實口感乳酪類製品 (28 項商品)
產品特徵詞(平均權重值)
商品主體：乳酪(0.711)、蛋糕(0.135)、蛋皮(0.123)、蛋捲(0.073) 成份：藍莓(0.197)、櫻桃(0.163)、野莓(0.109) 口感：香氣(0.048)、酸甜(0.038)、紮實(0.032) 特色：底層(0.047)
產品名稱
日式和風輕乳酪蛋糕、美式重乳酪蛋糕、創始原味黃金乳酪球、凱特蕾烤乳酪蛋糕、崙背鮮奶乳酪蛋糕、日式櫻桃乳酪、日式魔法乳酪、日式魔法乳酪蛋皮(蛋奶素) 、櫻花覆盆子乳酪蛋糕、法式皇家野莓森林塔、福義軒手工蛋捲(芝麻) 、蔓越莓乳酪堡、藍莓乳酪塔、鄉村家常雙乳酪弗卡夏、鹹乳酪蛋糕、藍莓提拉米蘇、北海道冰鎮乳酪派、卡士達千層派(巧克力黑櫻桃) 、崙背鮮乳酪蛋糕、巨嘴鳥重乳酪蛋糕(原味) 、巨嘴鳥重乳酪蛋糕(抹茶) 、巨嘴鳥重乳酪蛋糕(檸檬) 、日本高鈣乳酪蛋糕、森田芝士乳酪、歐式黑櫻桃巧克塔、法式切達乳酪杏仁塔、法式野莓森林塔、賀米爾貝果(藍莓貝果)

(資料來源：本研究整理)

(四) 第三階段分群：群集 13

下表為群集 13 進行第三階段分群。由於在 k 值為 5 及 k 值為 10 的分群結果比較中，最佳的分群結果是出現於 k 值為 5 下(搭配門檻值為 0.09 時)，因此 k 值不再往上遞增至 15 來進行實驗。

表 4-28 第三階段分群 ─ 群集 13

第三階段分群 ─ 群集 13：43 項產品				(群內相似度：0.108329)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.07	-	2	0.113157256	4.457%	
	0.08	-	3	0.13655293	26.054%	
	0.09	2	3	0.143594209	39.921%	
10	0.07	-	2	0.113157256	4.457%	
	0.08	-	3	0.125248605	15.619%	
	0.09	2	3	0.143594209	32.554%	

(資料來源：本研究整理)

群集 13 經過第三階段分群後，可獲得 3 個群集，其群內產品數目分別為群集 33：4，群集 34：35，群集 35：4。其中群集 34 為候選大群集，需進行第四階段分群，其餘群集則為不再變動的葉節點群集。下面列出各葉節點群集的群集表。

表 4-29 群集表：群集 33

群集 33：冰淇淋口感香蕉餡蛋糕製品 (4 項商品)
產品特徵詞(平均權重值)
商品主體：蛋糕(0.083) 成份：香蕉(0.761)、巧克力(0.212)、餅乾(0.185)、慕斯(0.102) 口感：酥鬆(0.071)、冰淇淋(0.042)
產品名稱
香蕉巧克力蛋糕、拿破崙蛋糕(奧利奧提拉) 、拿破崙蛋糕(香蕉巧克力) 、香蕉慕絲

(資料來源：本研究整理)

表 4-30 群集表：群集 35

群集 35：多層次口感芝麻餡料製品 (4 項商品)
產品特徵詞(平均權重值)
成份：芝麻(0.574)、 口感：香味(0.047)、鹹鹹(0.043)、濃(0.043)、層次(0.034) 特色：養生(0.113)、餡(0.053)
產品名稱
塔吉特蕾雅起士千層蛋糕、養生黑芝麻蛋糕捲、招牌狀元餅、維也納芝麻牛奶麵包

(資料來源：本研究整理)

第三階段分群完畢後所獲得的群集 22、群集 34 為候選大群集，因此皆需進行第四階段分群。

4.1.4 第四階段分群

(一) 第四階段分群：群集 22

下表為群集 22 進行第四階段分群。由於在 k 值為 5 及 k 值為 10 的分群結果比較中，k 值為 10 的最佳分群結果與 k 值為 5 的最佳分群結果相同並無改善，因此 k 值不再往上遞增至 15 來進行實驗。

表 4-31 第四階段分群 ─ 群集 22

第四階段分群 ─ 群集 22：40 項產品				(群內相似度：0.052112)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.10	5	3	0.0830729	59.412%	
	0.11	5	4	0.08536615	63.813%	
	0.12	7	5	0.100977425	93.770%	
10	0.10	5	3	0.0830729	59.412%	
	0.11	5	4	0.08536615	63.813%	
	0.12	7	5	0.100977425	93.770%	

(資料來源：本研究整理)

群集 22 經過第四階段分群後，可獲得 5 個群集，其群內產品數目分別為群集 37：15，群集 39：6，群集 40：12，群集 41：3，群集 46：4。所有群集皆為不再變動的葉節點群集。下面列出各群集的群集表。

表 4-32 群集表：群集 37

群集 37：奶油香氣小麥麵包 (15 項商品)
產品特徵詞(平均權重值)
商品主體：麵包(0.657)、羅宋(0.202) 成份：奶油(0.144)、小麥(0.109) 口感：油膩(0.060)、乾(0.040)、香氣(0.031) 處理方式：烤箱(0.045)
產品名稱
法蘭司維也納牛奶麵包(甜奶油口味) 、奶露麵包、瑞塔納、雙喜維也納牛奶麵包、明太子法國麵包、赤穗天鹽酥烤羅宋、奶酥哈斯麵包、巴塞隆納小麥麵包、米釀荔香麵包、英式小麥麵包、帕米吉安諾起士蘿勒弗卡夏 、愛斯基冰麵包、招牌羅宋、米釀荔香、香蒜小羅宋

(資料來源：本研究整理)

表 4-33 群集表：群集 39

群集 39：鹹甜口感起司麵包 (6 項商品)
產品特徵詞(平均權重值)
商品主體：麵包(0.241) 成份：起司(0.446) 口感：鹹鹹(0.146)、甜甜(0.118)、硬(0.093)
產品名稱
雪藏拿鐵麵包、義式托斯塔尼佐黃金乳酪、法式燒布蕾、起司城堡窯烤麵包、香濃起司堡、仆街起司

(資料來源：本研究整理)

表 4-34 群集表：群集 40

群集 40：鬆軟奶油起司餡餐包 (12 項商品)
產品特徵詞(平均權重值)
商品主體：餐包(0.514)、麵包(0.088) 成份：奶油(0.197)、起司(0.403) 口感：鬆軟(0.046) 特色：爆漿(0.084)、內餡(0.041)
產品名稱
巴特里爆漿奶油餐包、低脂冰心爆漿奶油餐包、歐式小餐包、帕瑪森切達蛋糕、冰心起士派、南瓜乳酪小餐包、法國雙色起士麵包、法式小圓包、卡士達千層(原味) 、法式長起士、湯種蔓越莓乳酪餐包、起士大司康

(資料來源：本研究整理)

表 4-35 群集表：群集 41

群集 41：外酥內軟大蒜奶油麵包 (3 項商品)
產品特徵詞(平均權重值)
商品主體：麵包(0.266) 成份：奶油醬(0.106) 口感：外酥內軟(0.103) 口味：大蒜(0.551) 特色：異國(0.234)
產品名稱
大蒜法國麵包、荷蘭村蒜棒、薩爾斯堡法式麵包

(資料來源：本研究整理)

表 4-36 群集表：群集 46

群集 46：重奶香蛋糕類製品 (4 項商品)
產品特徵詞(平均權重值)
商品主體：蛋糕(0.095) 成份：牛奶(0.445)、香草(0.353) 口感：奶香(0.073)、香甜(0.072)、柔軟(0.043)
產品名稱
塔吉特貝里斯牛奶千層蛋糕、夏威夷牛奶糖(綜合) 、和風香草捲、牛奶雪露

(資料來源：本研究整理)

(二) 第四階段分群：群集 34

下表為群集 34 進行第四階段分群。

表 4-37 第四階段分群 ー 群集 34

第四階段分群 ー 群集 34：35 項產品				(群內相似度：0.136203)		
k 值	門檻值	合併次數	分得群數	加權平均群內相似度	成長率%	備註
5	0.10	1	1	0.136203	0.000%	
	0.11	1	1	0.136203	0.000%	
	0.12	1	1	0.136203	0.000%	

(資料來源：本研究整理)

群集 34 經過分群後，無分群效果產生，因此停止此群集的分群動作。下表為群集 34 的群集表。

表 4-38 群集表：群集 34

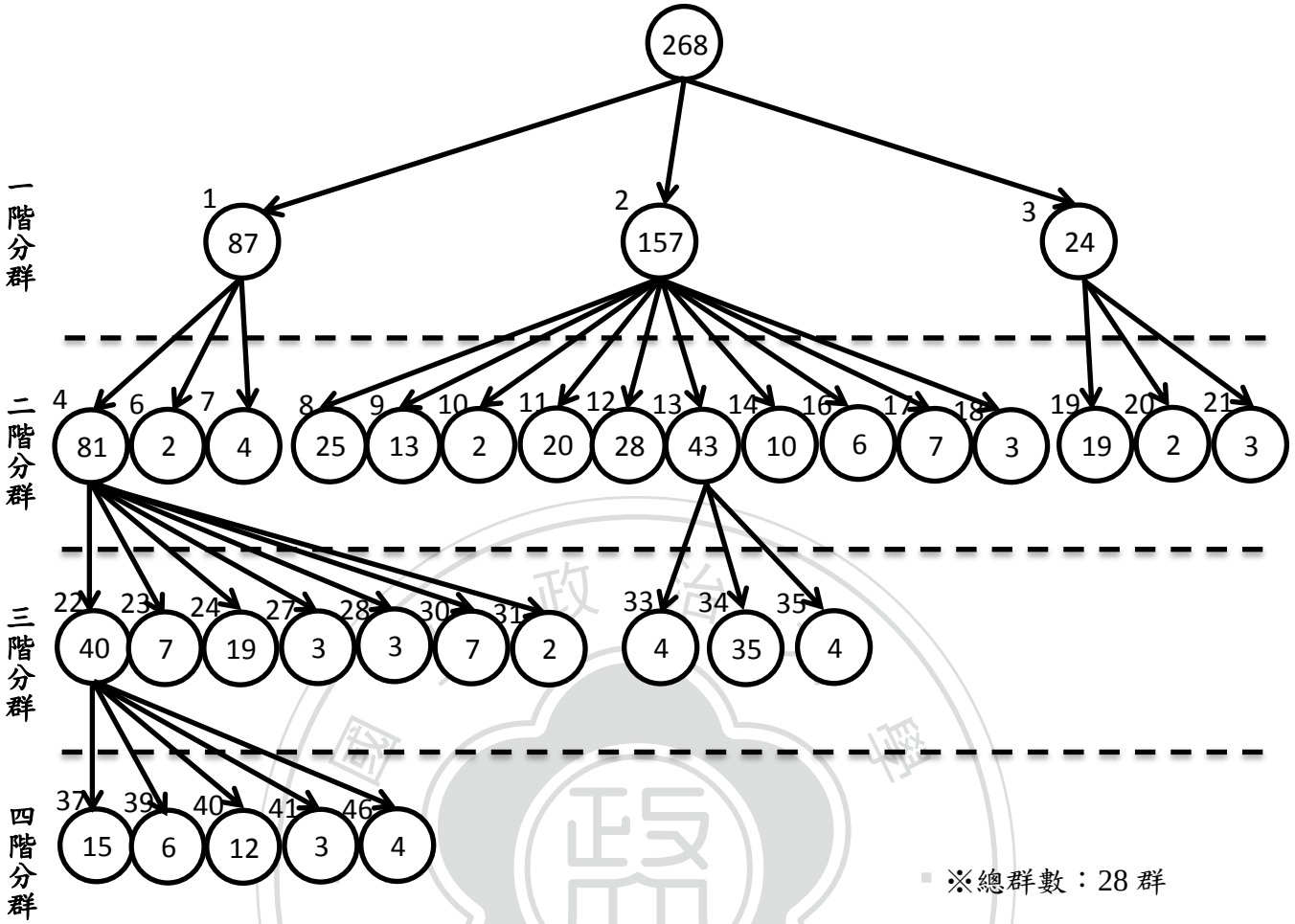
群集 34：濃郁甜膩巧克力布朗尼製品 (35 項商品)
產品特徵詞(平均權重值)
商品主體：布朗尼(0.153)、蛋糕(0.139) 成份：巧克力(0.778)、可可(0.076) 口感：苦(0.086)、濃郁(0.064)、濕潤(0.047)、甜膩(0.031)、細緻(0.021) 特色：融化(0.035)
產品名稱
黑珍珠蛋糕、塔吉特修格拉巧克力千層蛋糕、日式原味巧克力脆片蛋糕、杯子蛋糕之蛋糕杯杯、松露巧克力蛋糕捲、楓丹純巧克力蛋糕捲、法芙娜經典巧克力蛋糕、私房生巧克力蛋糕、羅密歐茱麗葉義式杯子蛋糕、阿奇諾巧克力蛋糕、香濃巧克力慕斯捲、凹蛋糕(巧克力) 、卡士達千層(經典巧克力) 、太妃甜心(焦糖巧克力加脆片) 、巧克力大司康、巧克力山藥派卡克、巧克力布蕾派(布蕾派對) 、巧克力布蕾派(米迦) 、巧克力濃情捲、巧克力爆漿布朗尼、巨嘴鳥重乳酪蛋糕(巧克力) 、帕瑪森典藏巧克力、手工玫瑰黑泡芙(黑爵巧克) 、新巧屋爆奶蛋糕、日式巧克力蛋糕、杏芳布朗尼、核桃巧克力蛋糕、法式長泡芙(瑞士巧克力) 、法朵拉巧克力捲、經典巧克力蛋糕、維也納巧克力麵包、維也納巧泥、脆皮布朗尼、超濃郁布朗尼蛋糕、花蓮提拉米蘇布朗尼蛋糕

(資料來源：本研究整理)

所有群集皆已分群完畢，結束分群。

第二節 群集結構

268 項產品經 kNN 分群器分群後可分得 28 個群集，平均群內相似度從未分群時的 0.029834 提升至 0.177428。下圖 4-1 為分群結果的群集樹狀圖，顯示各個階段的分群狀況。



※總群數：28 群

(資料來源：本研究整理)

圖 4-1 分群結果之群集樹狀圖

圓圈左上方之數字為群集代碼，圓圈內數值為該群集含有的產品個數。群集代碼按照數字從 1 開始編，缺碼則代表原本的群集為單項產品群集，藉由質心合併的動作已經將該群的產品併於其他群集中。總結葉節點的個數為 28 個，即為所分得的 28 個團購商品群集。

下表為群集總覽表，該表除了彙整 28 個群集的群集特徵外，由左邊的欄位往右看，呈現了各個群集的群集歸屬狀況。

表 4-39 群集總覽表

268 項產品	**群集 1** (麵包、吐司、奶油、起司、餐包、牛角、杏仁)	**群集 4** (麵包、吐司、奶油、起司、餐包、蛋糕、優格)	**群集 22** (麵包、餐包、奶油、起司、牛奶)	群集 37：奶油香氣小麥麵包
				群集 39：鹹甜口感起司麵包
				群集 40：鬆軟奶油起司餡餐包
				群集 41：外酥內軟大蒜奶油麵包
				群集 46：重奶香蛋糕類製品
			群集 23：營養餡料起司酥皮類製品	
			群集 24：香濃奶酥吐司類製品	
			群集 27：膨鬆奶油戚風蛋糕	
			群集 28：果泥年輪蛋糕	
			群集 30：酸甜優格內餡製品	
			群集 31：綿軟奶香麵皮類製品	
		群集 6：獨創層次感捲包牛角類製品		
		群集 7：杏仁顆粒口感甜品		
	群集 2 (乳酪、蛋糕、芋頭、草莓、捲、餡、濃郁)	**群集 13** (巧克力、蛋糕、香蕉、布朗尼、芝麻)	群集 33：冰淇淋口感香蕉餡蛋糕製品	
			群集 34：濃郁甜膩巧克力布朗尼製品	
			群集 35：多層次口感芝麻餡料製品	
		群集 8：QQ 芋頭綿密糕點類製品		
		群集 9：濕潤內餡滑嫩口感布蕾類製品		
		群集 10：精美包裝手工奶油蛋糕		
		群集 11：內含新鮮水果 QQ 奶凍類製品		
		群集 12：紮實口感乳酪類製品		
		群集 14：微苦細緻提拉米蘇		
		群集 16：乾澀鬆散桂圓製品		
		群集 17：甜蜜濕潤爆漿口感蜂蜜蛋糕		
		群集 18：爽脆青蔥鹹蛋糕		
	群集 3 (布丁、焦糖、泡芙、雞蛋、外皮)	群集 19：香甜焦糖布丁口感類製品		
		群集 20：紮實鳳梨纖維內餡製品		
		群集 21：健康低負荷麵粉類製品		

(資料來源：本研究整理)

268 項團購美食商品，共可分為 28 群。在進行第一階段的分群後，可將商品分為 3 個大群集，分別為「麵包類」(即群集 1)、「蛋糕類」(即群集 2)以及「其他口感類」(即群集 3)。在進行完四個階段的階段式分群以及單項產品群集進行質心合併後，可將「麵包類」分為 2 種類型的群集，即『麵包類產品』以及『擁有麵包特質的產品』。『麵包類產品』如：奶油香氣小麥麵包、鹹甜口感起司麵包、鬆軟奶油起司餡餐包、外酥內軟大蒜奶油麵包以及重奶香蛋糕類製品等 5 個群集；而『擁有麵包特質的產品』如：營養餡料起司酥皮類製品、香濃奶酥吐司類製品、膨鬆奶油戚風蛋糕、果泥年輪蛋糕、酸甜優格內餡製品、綿軟奶香麵皮類製品、獨創層次感捲包牛角類製品以及杏仁顆粒口感甜品等 8 個群集。「蛋糕類」大群集則依照口味區分為不同的蛋糕群集，其中由於有部分蛋糕產品為香蕉及巧克力口味搭配之組合，因此在分群的過程中有產生香蕉及巧克力組合之群集(即群集 13)，如：冰淇淋口感香蕉餡蛋糕製品、濃郁甜膩巧克力布朗尼製品以及都為濃烈口感的多層次芝麻餡料製品等 3 個群集；而其他口味的蛋糕類群集則為：QQ 芋頭綿密糕點類製品、濕潤內餡滑嫩口感布蕾類製品、精美包裝手工奶油蛋糕、內含新鮮水果 QQ 奶凍類製品、紮實口感乳酪類製品、微苦細緻提拉米蘇、乾澀鬆散桂圓製品、甜蜜濕潤爆漿口感蜂蜜蛋糕以及爽脆青蔥鹹蛋糕等 9 個群集。「其他口感類」大群集則含有：香甜焦糖布丁口感類製品、紮實鳳梨纖維內餡製品以及健康低負荷麵粉類製品等 3 個群集。

另外，在各個大群集中，本研究皆歸納出該群集的產品特徵代表詞(如表 4-39 之大群集的括號內所示)。越上層的群集所歸納出的產品特徵詞彙範圍較廣泛，而越下層群集所歸納出的產品特徵詞彙則為更細微的特徵。

第五章　結論與未來展望

第一節 結論與建議

本研究在每個分群階段皆設立不同 k 值以及 3 種門檻值，以不同的參數設定方式來進行分群。在未分群(只有一群)時，產品間的群內相似度為 0.029834，而分群門檻值從起始值 0.01，經過四個階段的分群進行後在結束時為 0.12，共可分得 28 個商品群集，而平均群內相似度則提升至 0.177428。總結整體實驗過程可分為以下結果：

1. 特徵詞萃取門檻值：

   經過觀察網友實際撰寫的團購網誌內容可得知，商品特徵詞的出現頻率並不像文章的關鍵字詞出現頻率那麼高，因此特徵詞萃取門檻值不宜訂太高。研究中觀察商品特徵詞被選取的狀況後將特徵詞萃取門檻訂為 60%。在此門檻下，不會因為特徵詞萃取比例訂太高而沒有過濾雜訊效果產生，也不會因為萃取比例訂定太低而造成商品特徵詞彙被排除掉。

2. 群集特徵代表詞擷取比例：

   在分群完畢後，依照各群集擁有的產品特徵詞彙對其進行群集特徵輪廓描繪，透過研究可知，群集特徵代表詞的範圍落於該群集質心權重前 20%的產品特徵詞彙。擷取各群集質心權重前 20%的產品特徵詞彙後再依人工過濾便可描繪出該群的群集特質，且群與群之間的群集特質皆可有所區別。

3. 在分群結果的衡量部分，除了可得知群內相似度有顯著的提升外，透過人工方式檢視群集的特性，也符合群內特質相近且群間特質差異。在分群結果中，

最大群集的產品數為35，而最小群集的產品數為2，以人工方式檢視可得知：許多商品數為 2 或 3 的小群集其商品在 268 項產品中屬於較特殊的種類，在經過分群後，這些相似種類之特殊產品皆歸屬到相同的群集中，倘若將這些少數相似之特殊產品合併於其他商品群集中，則較不易找出群集內獨特且共有的商品特徵。此外，含有產品數較多之大群集，在經過分群門檻值提升後，仍然歸屬於相同群集中，由此可見其群內產品有一定的相似特質存在。

最後所獲得之分群結果可以提供團購消費者選擇產品時參考，如消費者喜歡某種特性之商品，便可推薦其相似特性群集擁有的商品。此外，團購網站業者也可參考分群結果，針對歸為相同群集內之商品，規劃共同促銷活動，以增進成交機會。

第二節 未來研究方向

本研究透過文字探勘於顧客團購網誌中擷取商品相關資訊，再運用 kNN 分群器將網路團購商品進行群集分析，分群後之平均群內相似度有顯著提升，且透過衡量群集質心權重的方式可擷取出各個群集的群集商品特徵，並依照特徵給予命名。針對未來之研究方向，本節提出以下五點建議：

1. 本研究在商品資訊擷取部分僅採用顧客團購網誌。在現今環境下，許多網路上的心得分享已從網誌轉為以微網誌的型態發表，網誌的量越來越少，倘若能納入微網誌的資訊一併分析，應能增加產品相關資訊量，進而提高商品分群的準確率。但微網誌字數較網誌減少許多，因此對於商品相關資訊多以簡要的評論方式呈現，其處理方式還需進一步研究探討。

2. 本研究依照群集質心權重擷取前 20%的詞彙，再依人工過濾後取得產品特徵詞彙，並依面向來歸納產品特徵詞彙。倘若在獲取大量的產品特徵詞彙後能建立產品特徵詞彙庫，並依面向分門別類後，應能有更廣泛之利用。另外，也可藉由此方式達到自動化產生群集描述，應能更確切描繪出群集各個面向之輪廓。

3. 本研究之分群結果可以進一步應用於消費者滿意度的實證研究，調查本研究的分群結果是否符合消費者預期，能否增進消費者的滿意度，此外也藉以驗證版研究的分群結果。

4. 本研究經過替產品進行群集分析後已可將產品歸納分群，倘若能基於分群結果進行商品推薦，相同商品群集內的顧客優先推薦群內的團購商品，或是新進顧客先依據其有的團購商品網誌進行分群後再推薦最相近群集的團購商品，將能帶來更大的效用。

5. 本研究就目前找尋到擁有顧客團購網誌資料之團購美食商品進行分群，在未來可藉由批次擴增方式將新產品分門別類，透過 kNN 依照現有群集進行分類處理，倘若分類門檻值低於一定門檻，則為新產品創立新群集，以動態方式持續擴增團購美食群集。

參考文獻

中文文獻

1. 呂培仕，「口碑定義架構的發展：口碑文獻回顧1950~2008」，國立台灣科技大學企業管理學系，碩士論文，2010。

2. 巫啟台，「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博士班，碩士論文，2002。

3. 林淑婉，「影響網路團購再購意願因素之研究」，大同大學事業經營所，碩士論文，2010。

4. 張家蓁，「資料採礦應用於消費者網路團購因素探勘之研究」，淡江大學管理科學研究所企業經營碩士在職專班，碩士論文，2010。

5. 張瑜修，「消費者參與辦公室團購影響因素之研究-以宜蘭縣上班族為例」，佛光大學管理學系，碩士論文，2011。

6. 曾元顯，「數位文件之資訊組織與主題分析自動化之技術與應用」，台北市立圖書館館訊，2002年，第二十卷，第二期，23-35。

7. 喻欣凱，「運用支援向量機與文字探勘於股價漲跌趨勢之預測」，輔仁大學資訊管理學系，碩士論文，2008。

8. 莊隆泰,「群體採購中間商系統之研究」,國立中山大學資訊管理研究所,碩士論文,2010。

9. 詹益發,「網站部落格之顧客口碑評論分析研究-以台灣咖啡飲料市場為例」,元智大學資訊管理學系,碩士論文,2009。

10. 楊惠琴,「網路合購知覺風險與合購意向影響因素之研究」,東吳大學國際貿易學系,碩士論文,2006。

11. 廖婉如,「應用MOA理論探討團購主購者之忠誠行為－以知識交換為中介」,國立臺北大學企業管理學系,碩士論文,2010。

12. 潘倧偉,「口碑與從眾行為對團購意圖之影響—以團購美食為例」,南台科技大學行銷與流通管理系,碩士論文,2010。

13. 盧惠芬,「結合從眾行為探討影響網路團購購買意願因素」,中原大學國際貿易研究所,碩士論文,2010。

14. 戴尚學,「運用事件偵測與追蹤技術於中文多文件摘要之研究」,國立雲林科技大學資訊管理研究所,碩士論文,2003。

15. 戴維德,「使用最近鄰近分類法—以網路銀行為例」,南台科技大學國際企業系,碩士論文,2006。

英文文獻

1. Allen, V. L. (1965). Situational Factors in. Advances in Experimental and Social Psychology, Vol 2, ed. Leonard Berkowitz. New York, NY: Academic Press, 133-175.

2. Anand, K. S., & Aron, R. (2003). Group-buying on the Web: a comparison of price-discovery mechanisms. Management Science, 49(11), 1546-1562.

3. Arndt, J. (1967). Role of product-related conversations in the diffusion of a new product. Journal of Marketing research, 291–295.

4. Chen, K. J., Kiu, S. H. (1992). Word Identification for Mandarin Chinese Sentences. Fifth International Conference on Computational Linguistics, pp.101-107.

5. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13(1), 21–27.

6. Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B., Sen, S., Shi, M., Verlegh, P., et al. (2005). The Firm's Management of Social Interactions. Marketing Letters, 16(3-4) 415-428

7. Hanson, W. A. (2000). Principles of Internet Marketing. Ohio: South-Western College Publishing.

8. Larkey, L. S., & Croft, W. B. (1996). Combining classifiers in text categorization. Proceedings of the 19th annual international

9. Macinnis, H. (1997), Consumer Behavior, New York: Houghton Mifflin Company.

10. Nie, Jian-Yun, Brisebois, Martin & Ren, Xiaobo (1996). On Chinese Text Retrieval. Conference Proceedings of SIGIR, pp.225-233.

11. Salton, G., Wong, A., Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM, v.18 n.11, pp.613-620.

12. Salton, G., McGill, M. (1983). Introduction to Modern Information Retrieval,New York: McGraw-Hill.

13. Salton, G. (1989). Automatic Text Processing. Addison-Wesley, Reading, Mass.

14. Snyder, P. (2004, Jun 28). Wanted: Standards for Viral Marketing. Brandweek. 45, 21-21

15. Sproat, R, Shih , C., 1990. A Statistical Method for Finding Word Boundaries in Chinese Text. Computer Processing of Chinese and Oriental Languages, pp.336-351.

16. Wilkie, W. L. (1994), Consumer Behavior, 3rd ed., New York: John Wiley and Sons Inc.

17. Yang, Y., Carbonell, J.G., Brown, R., Pierce, T., Archibald, B. T. & Liu, X. (1999). Learning Approaches for Detecting and Tracking News Events. IEEE Intelligent Systems, v.14 n.4, pp.32-43.

網路資料

1. CKIP中文斷詞系統，中央研究院(2012)，取自http://ckipsvr.iis.sinica.edu.tw/。

2. 「資策會MIC網友上網購物行為模式調查 搜尋特定商品、自行比價、瀏覽部落格、查詢訂單 」(2007.07.19)，資策會(2012)，取自 http://mic.iii.org.tw/pop/micnews4_op_new.asp?sno=334&cred=2007/7/19，2007。

3. 蘇文彬 ，「MIC：逾2成受訪網友去年曾網路團購 」，iThome online，取自http://www.ithome.com.tw/itadm/article.php?c=65725，2011。