電子商務網站產品關聯度分析

謝怡翔 Hsieh, Yi-hsiang
Chaoyang University of Technology
s10427617@cyut.edu.tw

吳世弘 Wu ,Shih-hung
Chaoyang University of Technology
shwu(*Contact author)@cyut.edu.tw

楊媛媛 Yang, Yuanyuan
Tianjin Normal University
Echo34yy@163.com

摘要

本論文主要解決電子商務網站中過多產品分在同一類，無法區分性質上的差異問題。論文根據淘寶網等電子商務網站的產品說明為原資料，利用深度學習分析網站上的文字，找出各種產品之間階層化的關聯性。研究中將利用機器學習中不同的分群演算法，自動產生階層化的分類，並且對不同的分類結果作比較和闡述。其中關於產品的相似度將利用深度學習演算法，自動分析產品文本的相似性。研究結果將可以更清楚區分眾多產品的類型，產生有用的階層化分類。階層化分類將有助於相關業者配送，以及消費者購物時縮小選擇空間，方便購物。

關鍵詞：聚類、電子商務、商品關聯分析

Abstract

This paper mainly deals with the problem that many products in the same category in the e-commerce website, which cannot distinguish the difference of the nature. According to the product description of e-commerce websites such as Taobao, the paper uses depth learning and analysis of the text on the website to find out the relativity between different products. In the study, different clustering algorithms in machine learning will be used to automatically generate hierarchical classification, and the results of different classification will be compared and elaborated. The similarity of the product will be analyzed using the depth learning algorithm to automatically analyze the similarity of product text. The results of the study will allow for a clearer distinction between the types of products and the generation of useful hierarchical classifications. Hierarchical classification will help the distribution of relevant industry, as well as consumers to reduce the choice of shopping space, convenient shopping.

Keywords: cluster, E-commerce, Commodity Association Analysis.

1. 前言

電子商務近年來發展迅速,現有的電子商務平臺包含大型的 B2C 和/或 C2C 市場，擁有數以百萬計的產品。但是電子商務網站出售的商品數目繁多，商品分類模糊，給消費者購物帶來很多不便。電子商務網站中的產品通常被分為層次結構的分類和多層次結構的分類。產品分類在客戶導向服務和賣家推薦服務中有很大作用，可以說明消費者進行產品搜索和推薦。[3]本論文的研究目標在於觀察電子商務網站銷售的產品的分類，瞭解產品類型與分類的情況。觀察淘寶網商品分類的情況發現存在以下問題：1）商品分類太籠統，這使消費者在尋找商品的時候，難以找到需要的商品。2）商品分類階層較淺，使眾多商品分在同一類，從而使消費者難以選擇。機器學習中聚類演算法可以將相似的資訊分成一類，本實驗使用聚類演算法將產品再細分，有助於電子商務網站的產品銷售，也為消費者利用電子商務網站購物帶來方便。

2. 主要內容

我們的研究針對電子商務網站產品分類問題：過多產品分在同一類。提出的解決方法是: 搜集大量的產品資訊，使用資訊提取技術提取規律性的文字特徵，使用自然語言理解技術提取非規律性的文字特徵，產生產品特徵向量，以使用不同的分群演算法產生更細的分類樹。

我們解決方案的優勢是流程全自動化，免除人工維護。且程式保證生成的分類樹具有穩定性，生成的分類樹依然可以人工調整以滿足特殊需要。特別是流程中整合最新的深度學習演算法，提高對於非規律性文字的資訊擷取能力。

2.1 資料前處理

為方便程式處理，我們首先對原始語料進行刪去標點符號、去除停用字(如：的、嗎、

喔…等），接著執行斷詞；我們採用的是中央研究院詞庫小組 (Chinese Knowledge Information Processing Group ,CKIP)所開發的中文斷詞 API 進行斷詞。

2.1.1 利用深度學習工具程式 word2vec 將文字轉為向量

深度學習的快速發展，對現有的許多問題都提供了有效的解決方案，尤其在圖形識別，語言識別和自然語言處理方面有很大進步。

現存在許多開源的深度學習演算法工具和程式。Google 的工程師 Mikolov 在 2013 年發表的 word2vec 以及後續發展的 doc2vec 提供了對文字處理的便利功能[8,9]。Word2vec 是一個類神經網路，利用類神經網路訓練所輸入的資料，將每一個詞用多維的向量來表示，把一個詞句投射到向量空間之中，可以從中觀察到，例如相同屬性的詞可能會在向量空間之中靠得很近，甚至會觀察到一些詞性邏輯上的關係、上下文特徵。目前此項技術常使用在機器學習和資料採擷上，如果有足夠多的資料，Word2Vec 能夠基於這個詞的出現情況高度精確的預測一個詞的詞義，對於深度學習來說，一個詞的詞義只是一個簡單的信號，這個信號能用來對更大的實體分類，比如把一個文檔分類到一個類別中，從句子、段落、文章，進而衍生出更多類似的技術，例如像是 sentence2vec、paragraph2vec 和 doc2vec。

此技術在分類或尋找類似詞有相當不錯的效果，此技術用於我們研究的目標，以、應用在我們收集淘寶網和京東網站的產品資訊，建構大量的語料庫來訓練，訓練完後我們可以借此去找到相似的詞句，從中試著將商品分類或比較出差異。

2.2 分群演算法

在資料眾多的情況下，透過資料探勘技術中的聚類技術，能說明我們分析電子商務中產品的分類，藉由不同的資料登錄去比對並且觀察，從中分析相同及相異性

將物理或抽象物件的集合分組成為由類似的對象組成的多個類的過程被稱為聚類。由聚類所產生的簇是一組資料物件的集合，這些物件與同一個簇中的物件彼此相似，與其他簇中的物件相異。聚類技術[6]，是將一組資料依據相似度計算公式將其聚合或分割成若干群。聚類資料依據分群出來的目標與特性，分成許多簇，再依據同簇中的相同相似的屬性，不同簇之間的相異的屬性分析差異，尋找其中我們需要的相關性與意義。分群特性應包含：(1) 同一群的資料會有某些特性是相近的；(2) 不同群的資料會有顯著的差異，而聚類技術主要有以下類別：(1) 階層式方法，(2) 分割式方法，(3) 密集式方法，(4) 方格式方法，(5) 模型式方法[7]。

在這裡，我們用到三種主流的聚類演算法，分別是階層式分群法（Hierarchy Clustering）、K 平均演算法（K Means）、及模糊 C 平均演算法（Fuzzy C Means）。

2.2.1 階層式分群法

階層式分群法是透過圖 2-1 所示的階層架構方式，將資料層層反復的進行聚合或分割，以產生最後的樹狀結構。

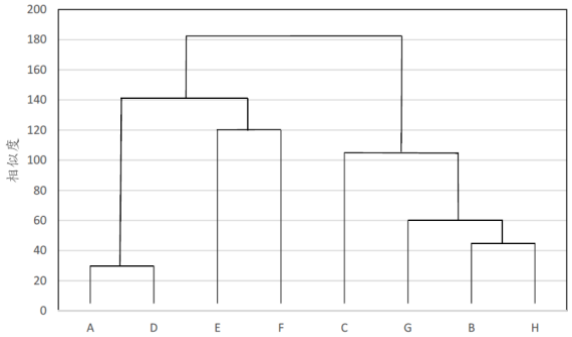

圖 2-1 階層式分群法的結構圖

階層式分群法的相似度可由距離來計算，兩分群相似度（Similarity）的計算公式：其中兩群之間的距離的 d，可使用四種不同方式進行計算：單一連接法（Single Linkage）,取群和群之間最近點的距離；平均法（Average Linkage）取群和群之間所有點的距離的平均；中心法（Centroid Linkage），取群和群之間中心點的距離；完全連接法（Complete Linkage）取群和群之間最遠點的距離。如圖 2-2 所示。

2.2.2 K 平均演算法

K 平均演算法是 MacQueen 與 1967 年所提出的分群演算法，必須事先設定分群的數量，k。k 個分群 Ci ,。計算公式如下：其中 :第 i 個分群的質心。K 平均演算法根據下列四部進行分群：1) 給定 k 值，將資料分割成 k 個非空白子集合。2）計算現在分割群組的質心，這些 質心為各個群組的中心點。3)將每個資料歸類於最接近的質心點。4）回到第二步，一直到每個群組資料沒有任何變化。

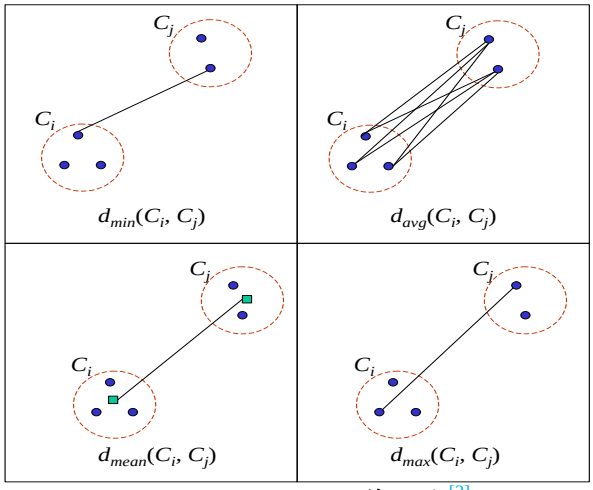

圖 2-2 四種距離計算方式[3]

2.2.3 模糊 C 平均演算法

Bezdek 於 1973 年提出的模糊 C 平均演算法，它的目標函數應用了模糊理論的概念。它使得每一組輸入向量不再屬於某一特定的分群，而以它的歸屬程度來表現它屬於某一群的程度。

資料集 X=,模糊 C 平均演算法將資料集 X 分成 c 群。令這 c 群的群心集為 V=。令資料點對群心的值為。群心計算公式如下：

模糊 C 平均演算法依照下列四步進行群：1)設定分群數 c，次方 m，誤差容忍度和起始隸屬矩陣。2)根據資料集和起始隸屬矩陣計算出起始的群心集。3）重新計算， ，修正各個群心值。4）計算誤差，若誤差在可接受範圍則停止，否則回到第三步。[3]

3. 產品相似度分析

根據我們目前收集的產品資訊，發現商品資訊分類分到第三層。第三層會有大量的相似產品，對消費者購物造成很大困擾。

針對我們要解決的問題：過多的產品屬於同一類。我們提出的結局方案是：1）收集產品資訊。2）利用資訊提取技術，提取規律性文字。3）使用自然語言理解技術提取非規律性的文字特徵 4）產生產品特徵向量 5）使用三種分群演算法產生分類樹。6）比較不同分群演算法產生的結果，比較產品分類的不同效果。實驗流程圖 3-1 如下所示：

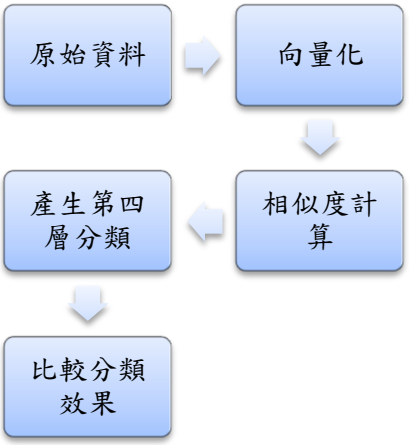

圖 3-1 實驗流程圖

3.1 資料介紹

我們收集的淘寶網產品資訊主要是三大類，分別是：百貨市場，美妝市場，食品市場。這三大類產品可以視為是第一層分類，在第一層下面有幾十個分類，視為第二層分類，在第二層之下的類別數目屬於第三層分類，類別數目有限，每一類產品數量巨大，在 1000 筆以上。

我們選擇了第三層中的五類產品做分群實驗，分別是：大豆油、運動鞋/休閒鞋、煎餅、發膜/倒膜、方便/速食品/真空即食。處理過後的資料形式如下圖 3-2（以大豆油為例）。

	A	B	C	D	E	F	G	H	I	J	K	L
1		taobaok1.id	taobaok1.	taobaok1.title	taobaok1.vec	taobaok1.vec	taobaok1.vec	taobaok1.vec3	taobaok1.vec	taobaok1.vec	taobaok1.vec	taobaok1.vec7
2	1	35120781	大豆油	自家熬制獾子油50克狗獾	0.021154	-0.029385	-0.054208	-0.016239	0.016215	-0.036228	0.031594	-0.054183
3	2	35146355	大豆油	色拉油北大荒东北九三大	0.022384	0.052325	-0.06245	0.059595	-0.015427	0.059223	0.01309	0.041106
4	3	35104540	大豆油	1.8L金龙鱼精炼一级大豆	-0.022736	0.033312	-0.026563	0.057153	0.025289	0.015627	0.053842	-0.02887
5	4	35104470	大豆油	芝焙色拉油 烘焙专用食用	-0.061724	-0.008555	-0.036213	0.017678	0.026199	-0.047653	-0.052232	0.050792
6	5	35140424	大豆油	俄罗斯原装进口大豆油 3	-0.024127	-0.043742	-0.05042	0.051998	-0.019263	0.008246	0.0059	-0.006205
7	6	35146169	大豆油	福临门一级大豆油(瓶装	0.006346	-0.057005	0.044594	-0.002801	0.044597	-0.039043	-0.052033	0.001996
8	7	35138385	大豆油	2桶包邮 5L升金龙鱼大豆	-0.03214	0.0271	-0.003242	-0.033451	-0.004915	-0.059424	-0.050359	-0.049818
9	8	35059744	大豆油	金龙鱼大豆油900ml 小瓶	-0.003693	-0.040703	0.017694	0.059556	-0.01872	0.037838	0.022159	0.002558
10	9	35060523	大豆油	芝焙色拉油 烘焙专用色拉	0.00201	-0.02072	0.023243	0.035902	-0.026911	-0.009999	0.047794	0.033001
11	10	35060871	大豆油	芝焙色拉油 食用大豆调和	-0.026389	0.046378	0.035855	0.053691	-0.060239	-0.001179	0.003858	-0.058293
12	11	35140122	大豆油	【天猫超市】福临门非转	0.009433	-0.006962	0.033397	-0.001314	-0.051598	-0.010393	-0.007471	0.048354
13	12	35157293	大豆油	满减烘焙原料芝焙色拉油	0.038202	-0.001573	-0.030622	-0.018455	-0.014933	-0.052372	0.006323	0.045992

圖 3-2 大豆油資料格式

3.2 實驗內容

根據試驗資料，我們利用三種不同的分群演算法做五個實驗。實驗工具是 R 統計軟體。R 是一種統計軟體也是一種程式設計語言。R 的前身是有 Ross Ihaka 與 Robert Gentleman（1966）所開發出來和 AT&T 貝爾實驗室 Rick Becker、John Chambers 與 Allan Wilks 等人開發的 S 語言很相似。R 有 windows、Unix、Linux 及 Apple MacOS 等不同作業系統的版本。

3.2.1 實驗一：

1) 針對第一類資料大豆油，一共有 952 筆。首先利用 K 平均演算法（kmeans）將這筆資料分成 8 群。得到的實驗結果如表 1：

表 1 大豆油（kmeans）分群結果

1	2	3	4	5	6	7	8
73	138	109	122	130	120	125	135

利用 R 語言的 ggplot2 套件得到大豆油資料用 K 平均演算法分群的圖 3-3：

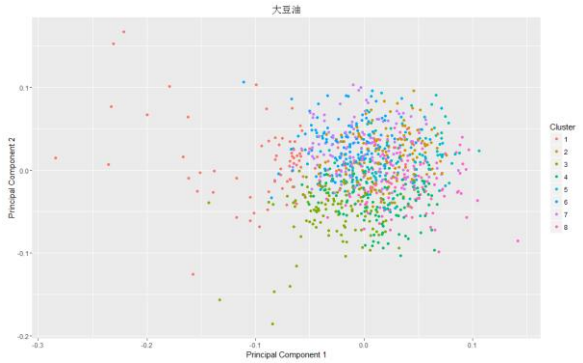

圖 3-3 大豆油 K 平均演算法分群圖示

2) 利用階層式分群演算法（hclust）將大豆油資料分成 8 群。得到的實驗結果如表 2 所示：

表 2 大豆油（hclust）分群結果

1	2	3	4	5	6	7	8
679	215	20	4	16	4	8	6

利用 R 語言畫出大豆油（hclust）分群結果圖 3-4：

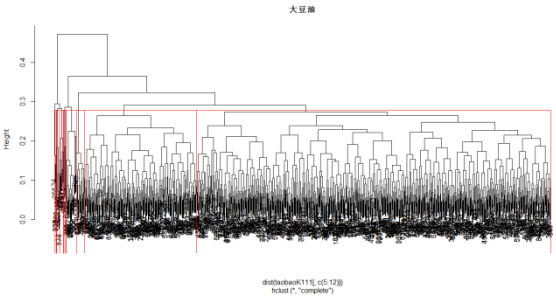

圖 3-4 hclust 分群圖示

3) 利用模糊 C 平均演算法（cmeans）將大豆油資料分成 6 群。得到的實驗結果如表 3 所示：

表 3 大豆油（cmeans）分群結果

1	2	3	4	5	6
82	342	25	116	51	336

利用 R 語言畫出大豆油（cmeans）分群結果圖 3-5：

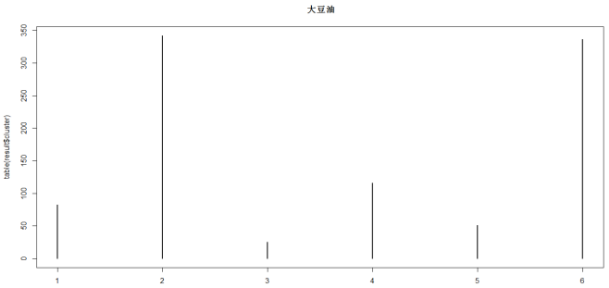

圖 3-5 大豆油（cmeans）分群結果

從大豆油的分群結果可以看到 kmeans 演算法分類結果比較均勻，每一群的類別數目相對穩定，可以很好將大類分成小類，而 hclust 演算法結果類別數目差距較大，它將大部分相似的大豆油產品分成一類，cmeans 演算法的結果介於兩者之間，大豆油產品分為數目不等的 6 類，有的類數目較多，有的較少，但類別數目還在可接受的範圍。

3.2.2 實驗二：

1) 針對第二類資料運動鞋/休閒鞋，一共有 2192 筆。首先利用 K 平均演算法（kmeans）將這筆資料分成 8 群。得到的實驗結果如表 4：

表 4 運動鞋/休閒鞋（kmeans）分群結果

1	2	3	4	5	6	7	8
356	124	313	361	43	346	344	3

利用 R 語言的 ggplot2 套件得到大豆油資料用 K 平均演算法分群的圖 3-6：

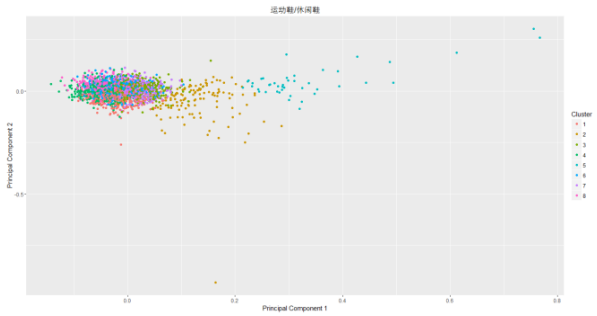

圖 3-6 運動鞋/休閒鞋 K 平均演算法分群圖示

2) 利用階層式分群演算法（hclust）將運動鞋/休閒鞋資料分成 8 群。得到的實驗結果如表 5 所示：

表 5 運動鞋/休閒鞋（hclust）分群結果

1	2	3	4	5	6	7	8
2050	19	36	1	74	7	3	2

利用 R 語言畫出運動鞋/休閒鞋（hclust）分群結果圖 3-7：

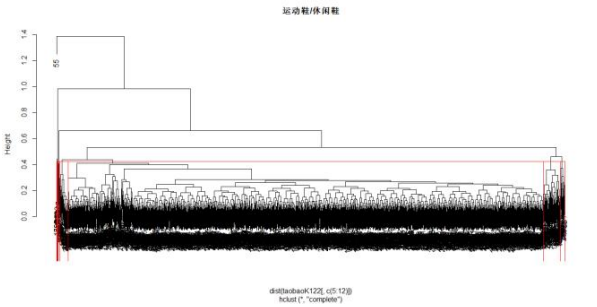

圖 3-7 運動鞋/休閒鞋（hclust）分群結果

3) 利用模糊 C 平均演算法（cmeans）將運動鞋/休閒鞋資料分成 6 群。得到的實驗結果如表 6 所示：

表 6 運動鞋/休閒鞋（cmeans）分群結果

1	2	3	4	5	6
2	1305	0	878	0	7

利用 R 語言畫出運動鞋/休閒鞋(cmeans)分群結果圖 3-8：

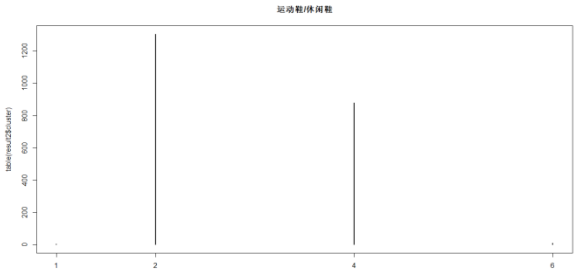

圖 3-8 運動鞋/休閒鞋（cmeans）分群結果

從運動鞋產品分類結果看，kmeans 演算法分類結果相對穩定，大部分群的數目比較接近，只有一小部分結果比較特別，需要進一步調整分群數做分類優化。而 hclust 演算法的結果依舊是將大部分相似的產品聚合在一起，類別集中分為一類。而 cmeans 演算法在遠動鞋產品的分類上和 hclust 演算法的結果有些接近，類別集中分為兩類。

3.2.3 實驗三：

1) 針對第三類資料煎餅，一共有 1451 筆。首先利用 K 平均演算法（kmeans）將這筆資料分成 8 群。得到的實驗結果如表 4：

表 7 煎餅（kmeans）分群結果

1	2	3	4	5	6	7	8
204	195	212	30	199	193	209	209

利用 R 語言的 ggplot2 套件得到煎餅資料用 K 平均演算法分群的圖 3-9：

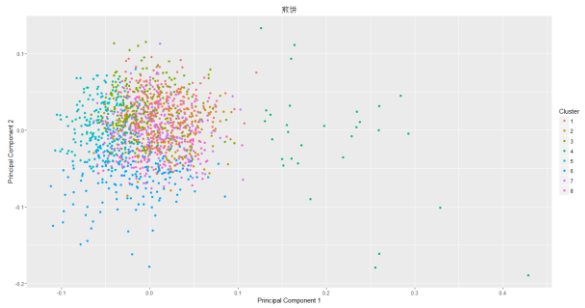

圖 3-9 煎餅 K 平均演算法分群圖示

2) 利用階層式分群演算法（hclust）將煎餅資料分成 8 群。得到的實驗結果如表 5 所示：

表 8 煎餅（hclust）分群結果

1	2	3	4	5	6	7	8
576	10	761	29	52	51	14	4

利用 R 語言畫出煎餅（hclust）分群結果圖 3-10：

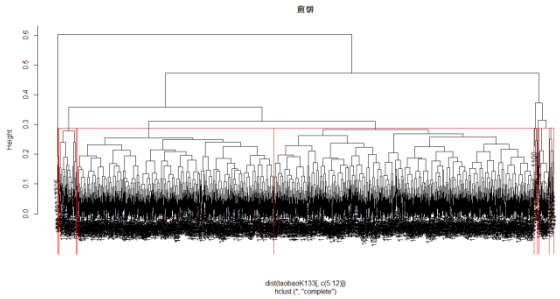

圖 3-10 煎餅（hclust）分群結果

3) 利用模糊 C 平均演算法（cmeans）將煎餅資料分成 6 群。得到的實驗結果如表 6 所示：

表 8 煎餅（cmeans）分群結果

1	2	3	4	5	6
134	80	554	94	513	76

利用 R 語言畫出運動鞋/休閒鞋(cmeans)分群結果圖 3-11：

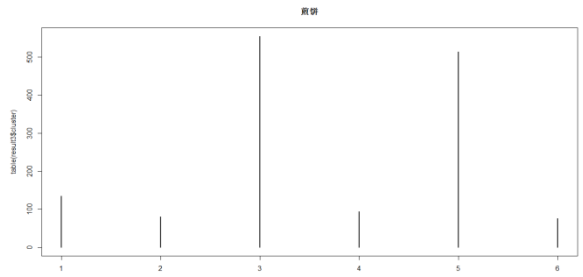

圖 3-11 煎餅（cmeans）分群結果

從煎餅的分群結果可以看出，kmeans 演算法的分群結果依舊非常均勻，偶爾出現比較特殊的類別數目可以進行分群數的調整，進行分群優化。hclust 演算法對於煎餅的分群結果也符合 hclust 分群的特點，分群結果富有層次感。大部分相似的產品分在同一類。cmeans 分群演算法的結果介於兩者之間，類別數目分佈沒有像 kmeans 很均勻，也沒有像 hclust 出現巨大差別。

3.2.4 實驗四：

1) 針對第四類資料发膜/倒膜，一共有 1104 筆。首先利用 K 平均演算法（kmeans）將這筆資料分成 8 群。得到的實驗結果如表 7：

表 9 发膜/倒膜（kmeans）分群結果

1	2	3	4	5	6	7	8
183	133	141	111	115	133	159	129

利用 R 語言的 ggplot2 套件得到发膜/倒膜資料用 K 平均演算法分群的圖 3-12：

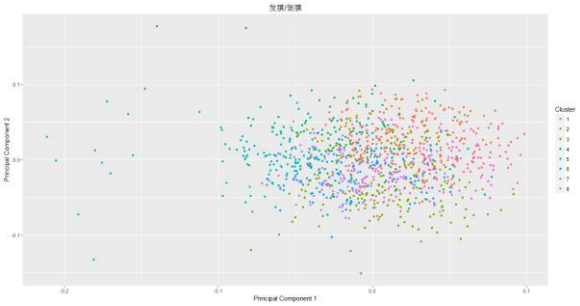

圖 3-12 发膜/倒膜 K 平均演算法分群圖示

2) 利用階層式分群演算法（hclust）將煎餅資料分成8群。得到的實驗結果如表8所示：

表 10 发膜/倒膜（hclust）分群結果

1	2	3	4	5	6	7	8
300	512	92	183	11	2	2	2

利用 R 語言畫出发膜/倒膜（hclust）分群結果圖 3-13：

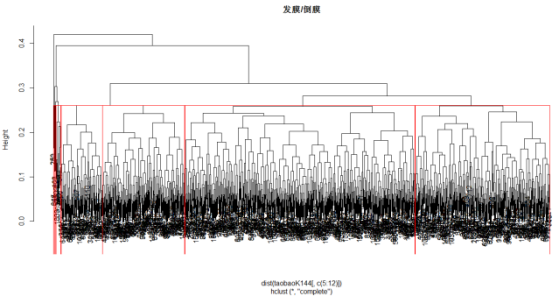

圖 3-13 发膜/倒膜（hclust）分群結果

3) 利用模糊 C 平均演算法（cmeans）將發膜/倒膜資料分成 6 群。得到的實驗結果如表 9 所示：

表 11 发膜/倒膜（cmeans）分群結果

1	2	3	4	5	6
344	326	152	17	146	119

利用 R 語言畫出发膜/倒膜（cmeans）分群結果圖 3-14：

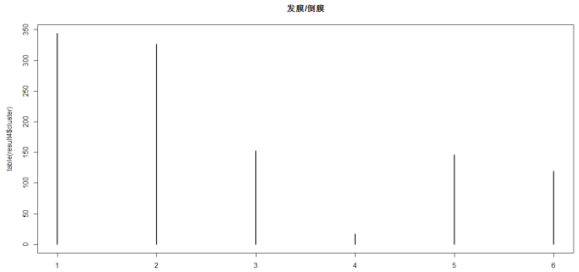

圖 3-14 发膜/倒膜（cmeans）分群結果

從发膜/倒膜的分群結果看，kmeans 演算法將資料均勻的分成 8 類，分群結果很穩定。從 hclust 演算法的分群結果看，相似的產品集中在一起，類別數目差別較大。相對 hclust 演算法的結果，cmeans 演算法的分群類別數目分佈相對均勻，但數目也有一些跳躍。

3.2.5 實驗五：

1) 針對第五類資料方便/速食品/真空即食，一共有 1104 筆。首先利用 K 平均演算法（kmeans）將這筆資料分成 8 群。得到的實驗結果如表 9：

表 12 方便/速食品/真空即食（kmeans）分群結果

1	2	3	4	5	6	7	8
183	133	141	111	115	133	159	129

利用 R 語言的 ggplot2 套件得到方便/速食品/真空即食資料用 K 平均演算法分群的圖 3-15：

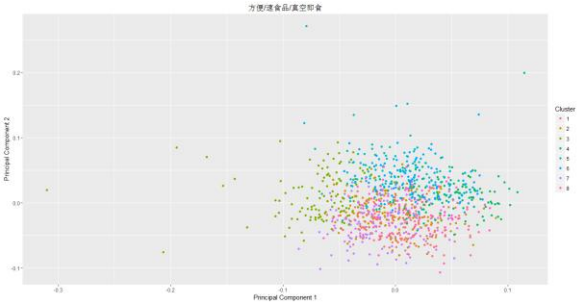

圖 3-15 方便/速食品/真空即食 K 平均演算法分群圖示

2) 利用階層式分群演算法（hclust）將方便/速食品/真空即食資料分成 8 群。得到的實驗結果如表 10 所示：

表 13 方便/速食品/真空即食（hclust）分群結果

1	2	3	4	5	6	7	8
211	681	136	6	4	2	3	1

利用 R 語言畫出方便/速食品/真空即食（hclust）分群結果圖 3-16：

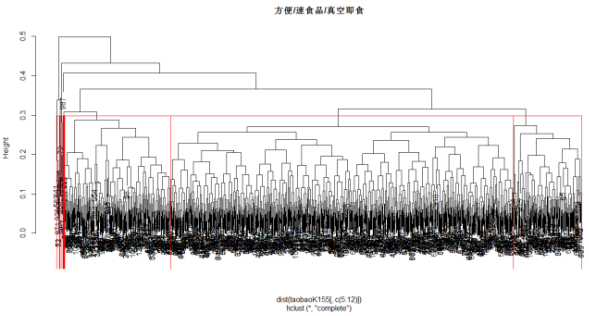

圖 3-16 方便/速食品/真空即食（hclust）分群結果

3) 利用模糊 C 平均演算法（cmeans）將方便/速食品/真空即食資料分成 6 群。得到的實驗結果如表 11 所示：

表 14 方便/速食品/真空即食（cmeans）分群結果

1	2	3	4	5	6
135	227	102	297	221	62

利用 R 語言畫出方便/速食品/真空即食（cmeans）分群結果圖 3-17：

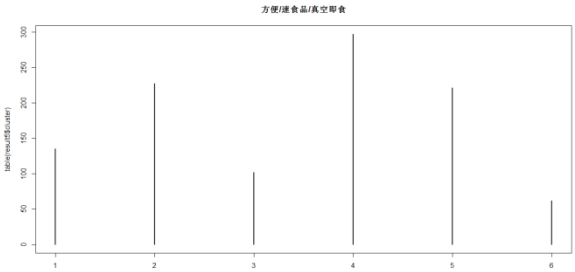

圖 3-17 方便/速食品/真空即食（cmeans）分群結果

從速食麵/速食品/真空即食產品的分類結果看，kmeans 演算法的分群結果類別分佈十分均勻，hclust 演算法的分群結果分佈十分不均勻，會將一些不相似的產品挑出來。cmeans 演算法的分群結果介於兩者之間。

3.3 實驗分析

從以上五個實驗結果可以看到 k 平均演算法分類結果較穩定，產生的類別數目分佈十分均勻，偶爾出現類別數目巨幅改變的情況，可以通過調整分群數，優化分群結果。k 平均演算法有助於淘寶經銷商將產品均勻分類，有助於消費者在更準確的類別中挑選產品，節省消費者流覽產品的時間。

階層式（hclust）分群演算法會將大量相似的產品分在前幾群，將冷門的產品（與其他產品相似度較低的產品）分在後幾群，hclust 分群有助於淘寶經銷商將冷門產品單獨分類，而剩餘的大量相似產品可以繼續利用 k 平均演算法進行分群，將兩種演算法結合在一起可以同時滿足多數消費者的需求，既可以將特殊產品分出，滿足部分消費者的需求，又可以降低第三層類別的數目。

第三種 cmeans 分群演算法的分群結果介於兩者之間，對於大部分產品的分群，既不會像 k 平均演算法一樣，使類別數目分佈很均勻，也不會像 hclust 分群演算法一樣，使類別數目差距很大，但它也會使大量相似產品分為一類，造成某一類別數目很多，沒有達到期望的情況。對於這種情況，我們可以採用再次使用 k 平均演算法繼續分群，可以產生更好的效果。

4.結論與未來展望

本論文研究商品是否可以用分群演算法將商品適當的分群。

首先商品資訊利用深度學習演算法 word2vec 計算出向量，定義向量內積為相似度計算方法，套用經典的分群演算法來分類產品。我們用淘寶網的產品說明當實驗對象。從實驗結果可以看到，我們可以控制每一個產品在第三層分類的數量，利用多種方法分出第四層後並比較各種分群演算法之間的差異。

未來我們將利用 TF-IDF 來計算各類產品的代表詞。也就是統計某類商品 title 欄位中出現的高頻率詞，並在其他群集中較少出現的詞。我們相信藉由這個方法，可以使我們比較每筆分群出的資料差異，評斷各分群演算法所產生的結果是否適合消費者。

未來可以有更多的資料來訓練實驗，訓練的句子也可以不僅僅只有 title 的欄位，商品的介紹內容、來源、廠商等等都可以當作參考作為使用，嘗試改變參數以及提高向量的維度來獲得更好的效果，在分類的過程中，嘗試繼續把分類分得更多，也可以試著將分類較細的群集適當的合併，藉由分類合併的過程觀察資料並嘗試更多方法來提高效能。

致謝

本研究依經濟部補助財團法人資訊工業策進會「106 年度虛實整合智慧商務關鍵技術與平台研發計畫計畫(4/4)」辦理。

參考文獻

[1] 廖尚文,台灣電子商務發展現況與兩岸跨境合作機會探討原文網址: http://www.cnra.org.tw/edm/20150807.pdf

[2] Jared P. Lander 著、鐘振蔚譯，R 軟體資料分析基礎與應用，旗標出版股份有限公司，2016。

[3] 李仁鐘，“應用 R 語言與資料分析”從機器學習、資料探勘到巨量資料，松崗資產管理股份有限公司，2015。

[4] Gupta V, Karnick H, Bansal A, et al. *Product Classification in E-Commerce using Distributional Semantics*[J]. arXiv preprint arXiv:1606.06083, 2016.

[5] Cevahir A, Murakami K. *Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant*[J], 2016.

[6] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome(2009)." 14.3.12 Hierarchical clustering".*The Elements of Statistical Learning* (PDF) (2nd ed.). New York: Springer. pp. 520–528.

[7] L. Chang, X. Lin, W. Zhang, J. Xu Yu, Y. Zhang, L. Qin: Optimal Enumeration: Efficient Topk .Tree Matching, In Proceedings of the VLDB Endowment, 2015.

[8] Q. Le and T. Mikolov: *Distributed representations of sentences and documents.* In Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014.

[9] 7. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. *Distributed representations of words and phrases and their compositionality.* In Advances in Neural Information Processing Systems, pages 3111--3119, 2013.

[10] L. Otero-Cerdeira, F. J. Rodriguez-Martinez, A. Gomez-Rodriguez: Ontology matching: a literature review, Expert Systems with Applications, 2015.

[11] P. Shvaiko, J. Euzenat: Ontology matching: state of the art and future challenges IEEE Transactions on Knowledge and Data Engineering, 2013.

[12] H. Tan, F. Hadzic, T. S. Dillon, E. Chang: RAZOR: mining distance-constrained embedded subtrees, In Proceedings of ICDMW, 2006.