



國立臺北科技大學

資訊工程系研究所

碩士學位論文

基於 Word2Vec 之熱門主題偵測

Popular Topic Detection based on

Vector Representation of Words



研究生：謝宗廷

指導教授：王正豪 博士

中華民國一百零五年七月

國立臺北科技大學  
資訊工程系研究所  
碩士學位論文

基於 Word2Vec 之熱門主題偵測  
Popular Topic Detection based on  
Vector Representation of Words



研究生：謝宗廷

指導教授：王正豪 博士

中華民國一百零五年七月

國立臺北科技大學  
研究所碩士學位論文口試委員會審定書

本校 資訊工程系 研究所 謝宗廷 君

所提論文 基於 Word2Vec 之熱門主題偵測

經本委員會審定通過，合於碩士資格，特此證明。

學位考試委員會

委

員：

楊凱翔

2/1 仇

王正豪

指導教授：

王正豪

所

長：

2/1 仇

中華民國 一零五 年 七 月 十三 日

# 摘要

論文名稱：基於 Word2Vec 之熱門主題偵測

頁數：39

校所別：國立臺北科技大學 資訊工程 研究所

畢業時間：一百零四學年度 第二學期

學位：碩士

研究生：謝宗廷

指導教授：王正豪 教授

關鍵詞：Topic Detection、Word2Vec、Document Clustering

社群網絡討論的主題常與社會輿論相互連動，影響力非同一般。本研究希望能夠藉由提出一個熱門主題偵測方法，讓一般大眾即使處於一個資訊爆炸的時代也能夠快速了解當今網路討論的主題趨勢。

現有的主題偵測方法，大多以文章內容擷取而來的關鍵字作為文章特徵。然而在文章相似度比較上並未考量特徵彼此的語意關係，對於在網路上用語多元且平均篇幅較短的文章，效果相當有限。本研究使用斷詞工具搭配 tf-idf 關鍵字擷取方法來擷取文章的關鍵字，並利用 Word2Vector Model 進行文章特徵的語意相似度分析及向量化，接著使用改良過後的 Hierarchical Agglomerative Clustering 演算法並使用內積作為相似度測量方法來進行主題分群，最後計算各主題熱門度藉此偵測出熱門主題。

實驗資料集使用批踢踢實業坊八卦版文章，人工標記每日五大主題並與系統偵測的熱門主題相互比較，結果顯示本研究提出的方法在 ARI、AMI 指標上及效率比傳統方法表現還要好。

# ABSTRACT

Title: Popular Topic Detection based on Vector Representation of Words

Pages: 39

School: National Taipei University of Technology

Department: Computer Science and Information Engineering

Time: July, 2016

Degree: Master

Researcher: Zong-Ting Xie

Advisor: Jenq-Haur Wang Ph.D.

Keywords: Topic Detection, Word2Vec, Document Clustering

Topic discussed in social networks are often linked to public opinion, where the influence is unusual. This Paper proposes a popular topic detection method, and hope it can let people in an information explosion age can quickly understand the trend of topic in the web.

Existing topic detection methods mostly used keywords of article content as features. However, in the similarity comparison, it does not consider semantic and syntactic relations of each other. In this paper, keywords extraction based on tf-idf method is used. We use Word2Vector Model to represent articles. Semantic and syntax are analyzed by using Word2Vector Model as vector space. Then use a clustering method modified from the Hierarchical Agglomerative Clustering and inner product similarity to cluster topics. Finally, calculate the popularity of topics to detect hot topics.

Experimental data set are collect from articles in PTT Gossiping board. We compare the topics detect by our method and the topic tagging by manual. The result shows that the proposed method has better performance than traditional method at ARI and AMI index.

## 致謝

2014 年 9 月成為網路檢索實驗室的一員，感謝老師的指導，讓我能夠進入 IR 的領域，感謝廷翰、怡靜、季霖、皓縈學長姐們的陪伴，你們影響我很多在學術上及生活上的觀念。

敬愛的父親在我碩二這一年突然離開這個世界，一時之間彷徨失措。感謝冠廷、宇辰、宏亘、學儒、庭瑋、其斌付出許多心力維持實驗室的正常運作，讓我能夠有機會騰出時間減輕家裡經濟負擔。尤其感謝冠廷平時的幫忙，時常給我有幫助的意見，與你同處一間實驗室是一件很幸運的事情。

感謝一路上支持我的家人及朋友，讓我能夠在求學道路上遇到任何問題都能夠堅持地走下去。最後我要感謝我的父母，沒有你們辛苦的栽培，就沒有今日的我。

宗廷於 2016/07/10



# 目錄

摘要 .....	i
ABSTRACT .....	ii
致謝 .....	iii
目錄 .....	iv
表目錄 .....	vi
圖目錄 .....	vii
第一章 緒論 .....	1
1.1 研究背景與動機 .....	1
1.2 研究目的 .....	1
1.3 研究貢獻 .....	2
1.4 章節概要 .....	2
第二章 相關研究 .....	3
2.1 Hot Topic Detection .....	3
2.2 Hierarchical Clustering .....	5
第三章 研究方法 .....	7
3.1 Feature Extraction .....	7
3.1.1 Segmentation .....	7
3.1.2 Keyword Extraction .....	8
3.2 Vector Representation .....	8
3.2.1 Word2Vec Model Training .....	9
3.2.2 Word Embedding .....	10
3.2.3 Vector Representation of Article .....	10
3.3 Document Clustering .....	12
3.3.1 Centroid Linkage Method .....	12

3.3.2. Dot Similarity Measure .....	13
3.4 Popularity Calculation .....	15
第四章 實驗與討論.....	17
4.1 資料來源.....	17
4.2 Word2Vec Training .....	18
4.3 分群效果評估.....	18
4.3.1. 評估指標 .....	18
4.3.2. 資料集 .....	20
4.3.3. Linkage 方法與 Similarity Measure 比較 .....	23
4.3.4. Feature Extraction 比較.....	27
4.3.5. 標題與內容大意向量比例比較 .....	28
4.3.6. 分群效果比較 .....	30
4.4 熱門主題偵測分析.....	34
第五章 結論.....	36
參考文獻.....	37





## 表目錄

表 3.1	Jieba 斷詞模式範例 .....	8
表 3.2	標題品質範例.....	11
表 4.1	主題文章列表範例.....	20
表 4.2	2016 年 6 月 1 日八卦版熱門主題.....	34



## 圖目錄

圖 2.1	linkage 方法示意圖 .....	5
圖 3.1	系統架構圖 .....	7
圖 3.2	CBOW and Skip-gram Model 示意圖 .....	9
圖 3.3	centroid linkage 方法合併示意圖 .....	13
圖 3.4	cosine 與 dot similarity measure 示意圖 .....	14
圖 3.5	群的合併示意圖 .....	15
圖 4.1	PTT 八卦版文章示意圖 .....	17
圖 4.2	contingency table .....	19
圖 4.3	資料集 A 主題規模大小分布圖 .....	21
圖 4.4	資料集 B 主題規模大小分布圖 .....	22
圖 4.5	資料集 C 主題規模大小分布圖 .....	22
圖 4.6	Linkage 方法比較 with Cosine Similarity measure (1) .....	23
圖 4.7	Linkage 方法比較 with Cosine Similarity measure (2) .....	24
圖 4.8	Linkage 方法比較 with Dot Similarity measure (1) .....	24
圖 4.9	Linkage 方法比較 with Dot Similarity measure (2) .....	25
圖 4.10	Similarity Measure 分群效果比較(1) .....	25
圖 4.11	Similarity Measure 分群效果比較(2) .....	26
圖 4.12	Similarity Measure 分群花費時間比較 .....	26
圖 4.13	Feature Extraction 比較(1) .....	27
圖 4.14	Feature Extraction 比較(2) .....	28
圖 4.15	標題與內容大意向量比例比較 (資料集 A) .....	29

圖 4.16	標題與內容大意向量比例比較 (資料集 B) .....	29
圖 4.17	標題與內容大意向量比例比較 (資料集 C) .....	30
圖 4.18	分群效果比較(資料集 A) .....	31
圖 4.19	分群效果比較(資料集 B) .....	31
圖 4.20	分群效果比較(資料集 C) .....	32
圖 4.21	分群花費時間比較.....	33
圖 4.22	熱門主題命中率.....	35



# 第一章 緒論

## 1.1 研究背景與動機

近年網路科技及社群網絡蓬勃發展連帶也影響人們獲取資訊的途徑。年輕世代的族群不再藉由傳統新聞媒體來認識世界、感知社會的溫度，他們利用網路上的新興媒體或是社群討論區來獲取新知、瞭解社會的脈動。2014 年 3 月 18 日太陽花學運過後公民意識開始崛起，社群平台關於社會議題的相關討論越來越熱絡。許多的議題透過網路在這些平台迅速地被討論，連帶影響當今傳統媒體的報導焦點，形成一股別於傳統、下對上的社會影響力。

然而網路的快速便利性除了讓資訊容易取得流通之外，也讓人邁進一個資訊爆炸的時代。許多的各種文章、討論意見在網路上充斥，使用者必須花費極大心力及時間才能從中篩選出同一議題的相關文章。本研究希望能夠透過提出一個主題偵測的方法，幫助使用者快速了解有興趣的議題相關討論。另外透過熱門度的計算，能夠幫助使用者了解當今社群網路熱門討論的議題，希望藉此能夠促進大眾對於社會議題的參與程度。

## 1.2 研究目的

2013 年 google 開源一系列 deep learning 的專案，其中 Word2Vec[1]提出了一個有別於傳統的詞向量觀念，因此受到學術界極大的關注。雖然相關研究如雨後春筍般被提出，不過以中文語料為主的應用研究還是較為缺乏。

另一方面，文件分群及主題偵測研究已經發展了一段時間，不過由於網路用語多元以及平均文章篇幅普遍較短的特性，傳統方法適用上有一定的侷限性。因此，本研究希望能夠應用時下熱門的 Word2Vec 來解決傳統熱門主題偵測上在語法與語意分析上的不足以及因向量維度過大無法短時間處理大量文件的限制，並為以中文語料為主的研究盡一份心力。

## 1.3 研究貢獻

雖然 Word2Vec 能夠將字詞轉換成詞向量，不過字詞向量化到文章向量化還是有一段距離。本研究提出一個文章向量化方法，實驗證明不對輸入文章做任何篩選，直接以此向量進行文件分群並計算熱門度能夠有效偵測熱門主題，顯示此文章向量化方法具有一定的效果。

Hierarchical Agglomerative Clustering(HAC)是常見應用於主題偵測的分群方法，很適合用來對文件分出有階層概念的主題，然而 HAC 最大的缺點是時間複雜度過高，對於需要處理大量文章的熱門主題偵測研究更突顯了效率上的劣勢。

雖然有許多群相似度以及文件相似度的測量方法可以增進速度，不過增進速度的同時相對也會犧牲一些分群品質。因此本研究提出了別於傳統 tf-idf vector space 常見的 cosine similarity 相似度測量方法。實驗證明，應用此相似度測量方法搭配以 centroid linkage 為群相似度計算方法的 HAC，除了能夠保持分群速度上的優勢之外，也能夠有效維持分群的品質。

## 1.4 章節概要

本論文共有五個章節。第一章為緒論，介紹研究動機、目的以及貢獻。第二章為相關研究，分別探討主題偵測以及分群方法的相關研究。第三章為本論文所提出的方法，包含特徵擷取、向量表示方法、文件分群方法以及熱門度的計算方法。第四章為實驗結果與討論。第五章為結論與未來展望，針對整個實驗結果做一個小結，討論本研究提出方法的優點及限制，並說明未來可以改善及發展的方向。

## 第二章 相關研究

### 2.1 Hot Topic Detection

Topic Detection and Track(TDT)相關研究已經進行很多年了。Allan 等人(1998)[2]將新文件當成 query 去查詢先前的主題，若查詢後沒有相似的主題則該文件會被當成新的主題。字詞由 tf-idf weight[3]與”surpriseness”表示，一個詞如果越不常出現則越”surprising”。Yang 等人[4]利用 tf-idf vector space model，將文件用 group-average 以及 single pass 等方法來分群，並以此來偵測新的事件主題。Schultz 等人(1999)[5]將文件中的文字去掉 stop words 並保留 idf 與 tf 高的字詞作為文件的 feature，並利用 cosine similarity measure 與 single linkage 方法進行分群來偵測主題。Makkonen 等人(2004)[6]將 vector space 切分成四個簡單的語意空間，分別為 places, names, temporal expressions and general terms，並利用四個不同的空間向量來計算文件相似度。實驗結果顯示 place 與 temporal expression 相似度辨識效果還需要改進，而且整體效率沒有比以 tf-idf vector space model 建構的文件向量還要好。Wartena 等人(2008)[7]將文件抽取出關鍵字，並以關鍵字代表文章以不同的距離測量方法進行分群，最後發現基於 Jensen-Shannon divergence of probability distributions 的距離測量方法效果表現會比 cosine similarity measure 還要好。受到此篇研究的啟發，本研究也將文件抽取出關鍵字作為文章特徵後向量化。

上述研究可以發現傳統主題偵測方法皆由 tf-idf vector space 所建構的文件向量來進行階層式分群或是 single pass 的快速分群。差異僅在使用不同的特徵擷取方法以及不同的相似度測量方法，不過以 tf-idf 為主要的 vector space model 都具有一個相同的共通點：文件的相似度與共同文字出現的次數多寡有關，這樣的相似度測量標準無法準確分辨一字多義或不同文字描述同一件事情的狀況。另外向量的維度也會與字詞數量相關，對於文件量大的計算相當不利。

為了解決傳統的 tf-idf vector space model 的問題，以 Latent Space Analysis(LSA)[8]相關研究[9]開始發展。LSA 將 tf-idf term document matrix 轉到維度較低的 latent space，以 latent space

表示的向量維度較短，而且具有隱含的語義關係。Probability Latent Space Analysis(PLSA)[10]是以 LSA 為基礎，加入了機率模型來改善 LSA，然而 PLSA 對於未曾出現過的文章很難給予初始機率，因此出現另一個改進的版本 Latent Dirichlet Allocation(LDA)[11]。LDA 將文件視為有許多個主題，文章對不同主題有各自的分布機率。有許多研究[12][13][14]基於 LDA 來進行主題偵測。然而 LDA 缺點之一是必須事先決定 Topic 的數量，這是個非常不容易的工作。而且 LDA 與 tf-idf vector space 或是 LSA 一樣還是屬於 bag of word 的延伸，依然會受 bag of words 的限制：文件量大所產生的向量維度也會變高。

另一方面，類神經網路模型相關研究開始發展，其中 word2vec 所建構的向量空間模型向量維度與神經元數量一樣是固定的，這個特性能夠解決 bag of words 向量空間隨著字詞量過大而使向量維度變高的問題。Moran 等人[15]利用 word2vec 來 expand tweets，藉此來偵測第一個主題的起源。Hashimotoa 等人[16]利用 word2vec 衍伸的方法將文章向量化，並對文章向量使用 k-means 方法來分群，藉此來偵測主題。本研究也基於 Word2Vec 提出一個文章向量化的方法藉此來進行語意分析以及解決 bag of words 的限制。

## 2.2 Hierarchical Clustering

分群方法有很多種，而常見運用在文件分群上的有 K-means, Hierarchical Clustering[17]。一般來說，若要求分群品質會使用 Hierarchical Clustering；若要求分群效率則會使用 K-means[17]。

K-means 思想最早能被追溯到 1957 的 Hugo Steinhaus[18]，但”K-means”此術語在 1967 才被 James MacQueen[19]所使用。K-means 的標準演算法則是在 1957 年被 Stuart Lloyd 作為一種脈衝碼調制的技術所提出，但直到 1982 年才被貝爾實驗室公開出版[20]。在被貝爾實驗室公開出版前，E.W.Forgy 在 1965 年也發表了相同的方法[21]。

K-means 初始會隨機挑選群組中心點，接著將每一資料點分到相似度最高的群組，接著重新計算群組中心點，反覆執行直到群組中心點不再改變為止。K-means 的優點在於能夠快速收斂分群結果，但分群效果並不太穩定，而群數 K 值在實務上也很難決定。

Hierarchical Clustering[22]有 Agglomerative 與 Divisive 兩種，前者以 bottom up 方式依序將距離最近的兩個群集合併為一群；後者以 top down 方式將群集分割成較小的兩群。前者時間複雜度為 $O(n^2 \log(n))$ 又比後者的時間複雜度 $O(2^n)$ 好上許多。Hierarchical Clustering 對於群集間距離測量有許多不同的方法[23]。

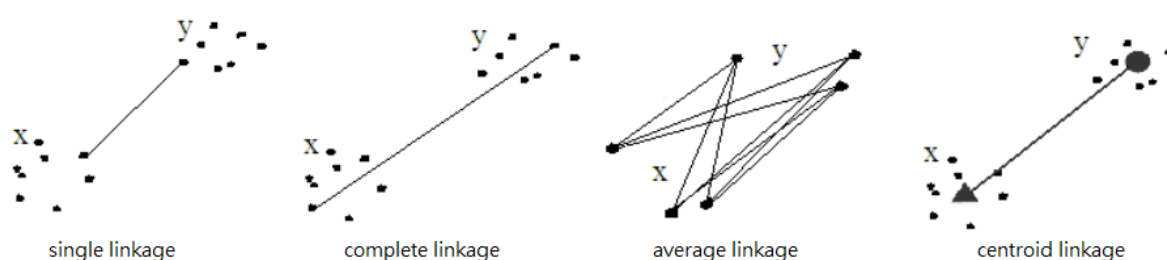


圖 2.1 linkage 方法示意圖

如圖 2.1，常見的方法有 centroid linkage(以群中心點計算距離)、average linkage(以群集各點兩兩距離取平均值)、single linkage(兩兩距離取最小值)及 complete linkage(兩兩距離取最大值)。其中，centroid linkage 方法運算速度較快，但有可能產生 Non-monotonic 的階層[24]。而



相似度也有許多測量方式，最為常見的是 cosine similarity measure。Plate 利用運算量較少的 dot similarity measure 進行粗略篩選以提升分群效率[25]。受此啟發，本研究也利用 dot similarity measure 來加快分群的速度。



## 第三章 研究方法

本研究提出的方法主要可分為四個部分：Feature Extraction、Vector Representation、Document Clustering 以及 Popularity Calculation。如圖 3.1 所示，本研究對輸入文章擷取特徵並用事先以大量資料訓練好的 Word2Vec Model 輔助向量化，接下來利用這些向量進行文章分群找出相對應的主題，最後針對各主題計算熱門度來偵測熱門主題。

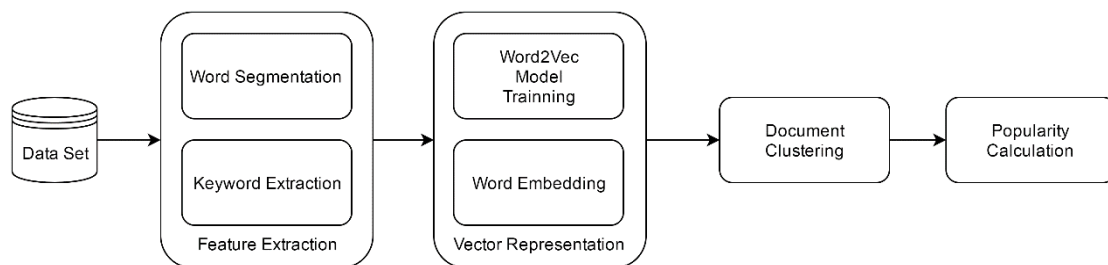


圖 3.1 系統架構圖

### 3.1 Feature Extraction

由於一整篇文章通常包含了許多較不重要的雜訊，本研究藉由關鍵字擷取方法濾除文章內容的雜訊，擷取出來的關鍵字搭配文章標題作為文章特徵，去除雜訊後的特徵對後續的文章向量化以及文章分群會有較好的效果。

#### 3.1.1. Segmentation

中文能夠表達語意的最小單位是字詞，一連串的字詞組成語句，進而形成一篇文章。由於中文的語句字詞是彼此相鄰的，不像英文字詞之間有空白分隔，因此在處理中文資料的時候，必須先將原始資料斷詞，將字詞彼此斷開以利後續處理。

本研究採用 Jieba 作為斷詞工具[26]。Jieba 斷詞共有三種模式：精確模式、全模式、搜尋引擎模式。如表 3.1 所示，精確模式會嘗試將句子精確的切開；全模式會將句子可以成詞的詞語全部都掃描出來；搜尋引擎模式則是將精確模式切出來較長的詞再次切分。本研究所採用的斷詞模式皆為搜尋引擎模式，詳細原因待 3.2.1 說明。

表 3.1 Jieba 斷詞模式範例

Jieba 斷詞模式	結果
精確模式	枝頭 / 上 / 暗褐色 / 球型 / 蒴果 / 像 / 極了 / 成串 / 鈴鐺
全模式	枝頭 / 頭上 / 暗褐 / 暗褐色 / 褐色 / 色球 / 球型 / 蒴果 / 像 / 極了 / 成串 / 串鈴 / 鈴鐺
搜尋引擎模式	枝頭 / 上 / 暗褐 / 褐色 / 暗褐色 / 球型 / 蒴果 / 像 / 極了 / 成串 / 鈴鐺

### 3.1.2. Keyword Extraction

本研究使用基於 TF-IDF 的關鍵字擷取方法，依照分數排名取前 K 個關鍵字作為文章內容的特徵。實驗結果顯示，取前 15 關鍵字會有最好的效果。另外，文章標題對於文章內容也具有相當程度的意義。所以除了內文關鍵字之外，本研究也將標題斷詞後的結果取為文章特徵。

## 3.2 Vector Representation

本研究利用預先訓練好的 Word2Vec Model 將文章特徵向量化，並將這些特徵配合不同權重合成文章向量以利後續分群。

### 3.2.1. Word2Vec Model Training

Word2Vec 是 Tomas Mikolov 等人在 2013 提出的 Word Embedding 模型，藉由類神經網路的學習，Word2Vec 能夠將字詞轉換成具有語意及語法意義的向量[27]。如圖 3.2 所示，Word2Vec 有分 Continuous Bag of Words (CBOW) 與 Skip-gram 兩種模型，前者利用字的前後文 (Context) 來預測字詞的用法，後者利用字詞來預測前後文的用法。一般來說，CBOW model 訓練速度較 Skip-gram model 快上數倍、語法資訊含量較好，但語意分析效果會比較差[27]。本研究希望能夠利用 Word2Vec 補足一般文件分類常欠缺的語意分析，因此選用語意分析效果較好的 Skip-gram 為 model。

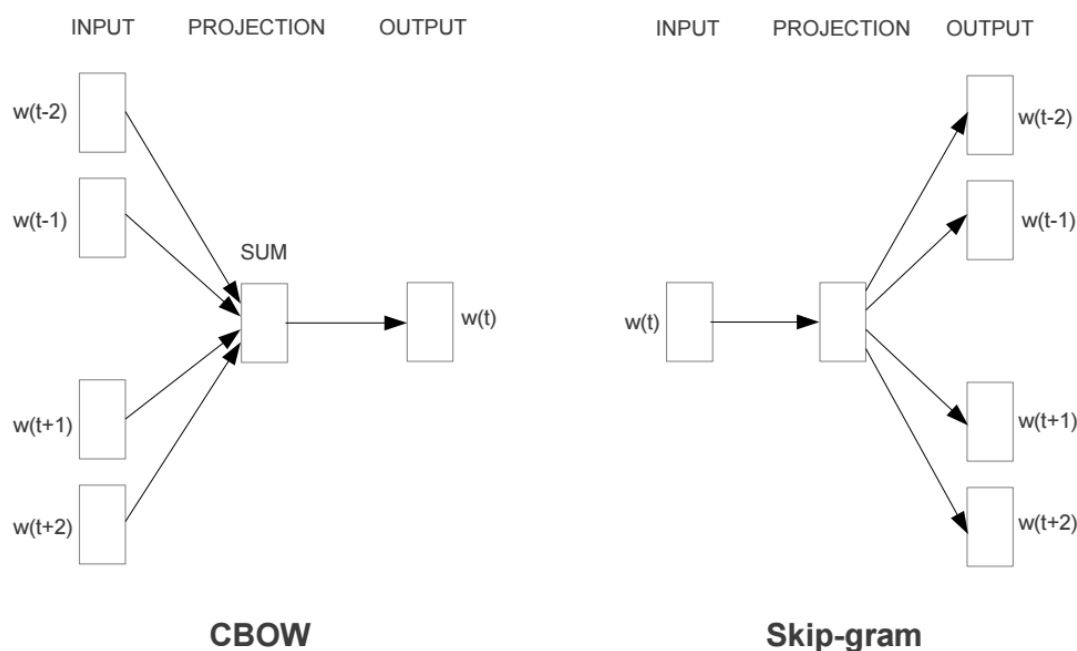


圖 3.2 CBOW and Skip-gram Model 示意圖

由於 Word2Vec 訓練方法是以字詞構成的大量句子來反覆調整神經元的權重，所以本研究必須先將訓練資料集斷句再對句子斷詞。如表 3.1 斷詞範例所示，Jieba 斷詞共有三種模式：精確模式、全模式、搜尋引擎模式。”暗褐色”在精準模式下能被精確切分出來，在全模式以及搜

尋模式下可以多切出”暗褐”以及”褐色”字詞。若是使用 CBOW model，這些多切出來的字詞會改變前後文而使預測效果變差；但若採用 Skip-gram Model，對於”暗褐色”原本的上下文來說，這些額外切出的詞語反而能夠幫助提升語意辨識效果。由於全模式會將可以成詞的所有可能列出來，有可能會比搜尋引擎模式會切出更多的詞，如”頭上”、”串鈴”。但這些詞可能跟原意毫無關係，反而會讓 Model 品質下降。因此本研究採用搜尋引擎模式作為主要的斷詞模式。

### 3.2.2. Word Embedding

經由 Word2Vec 轉出來的詞向量在語意或在語法上的會具有一定的意義，如  $\text{Vector}(\text{'big'})$  與  $\text{Vector}(\text{'bigger'})$  近似； $\text{Vector}(\text{'France'})$  與  $\text{Vector}(\text{'Italy'})$  近似[27]。而且這些向量具有某種程度的線性關係，如  $\text{Vector}(\text{'king'}) - \text{Vector}(\text{'man'}) + \text{Vector}(\text{'woman'})$  會近似於  $\text{Vector}(\text{'queen'})$ [28]。某種程度上，詞向量可以藉由簡單的相加來表達另外一個有意義的詞向量[29]，如  $\text{Vector}(\text{'大杯'}) + \text{Vector}(\text{'冰'}) + \text{Vector}(\text{'奶茶'})$  與  $\text{Vector}(\text{'大冰奶'})$  以本研究使用的測試資料集所訓練出來的 Model 計算得出的 Cosine Similarity 可以高達 0.6，可以期待 Model 訓練的效果越好，這些線性關係會更趨明顯。

### 3.2.3. Vector Representation of Article

文章特徵可以分為兩類：標題以及內容大意。基於 Word2Vec 的線性關係特性，將標題斷詞後的結果分別使用 Word2Vec Model 查詢對應的詞向量，並加總起來作為文章標題向量，如式 3.1 所示。

$$\vec{v}_{title} = \sum \vec{v}_t, t \in \text{title} \quad (3.1)$$

其中  $\vec{v}_t$  代表標題斷詞後的詞向量。文章內容大意可以由內文關鍵字組成，相同地，文章內容大意向量也可以由各關鍵字對應的向量組合成。與標題向量較為不同，內容關鍵字有重要性

之分，因此在將合成內容大意向量時，需要將關鍵字對應的權重也考慮進來，其內容大意向量合成之數學定義如下：

$$\vec{v}_{content} = \sum \vec{v}_i \cdot w_i, i \in \text{keywords of content} \quad (3.2)$$

$\vec{v}_i$ 表示第  $i$  個關鍵字向量， $w_i$ 是對應的關鍵字權重，此權重由使用 tf-idf 關鍵字擷取方法所得的分數正規化後而來。最後將標題向量及內容大意向量乘以對應的權重相加得到文章向量。

$$\vec{v}_{document} = \vec{v}_{title} \cdot w_t + \vec{v}_{content} \cdot (1 - w_t) \quad (3.3)$$

其中  $\vec{v}_{title}$  為文章標題向量， $\vec{v}_{content}$  為文章內容大意向量， $w_t$  則是標題向量比例，介於 0 到 1 之間。標題向量比例取決於資料集文章標題的”品質”，若標題品質很好則會調高標題向量的權重；反之則調降。表 3.2 **錯誤！找不到參照來源**。為標題品質的範例，當能夠直接藉由文章標題得知文章內容大意則代表標題品質良好；相對的，僅從文章標題無法得知文章內容大意則表示文章標題欠佳。如”[F B] 黑人 陳建州”，從此標題能夠得知此篇文章是討論陳建州 FB 相關的動態，但無從得知實質討論內容是什麼；標題為”正晶限時批-限時批政經”的文章大約能夠了解文章在討論該政治節目的節目內容，但無法得知當天節目內容聚焦在哪些議題。

表 3.2 標題品質範例

標題品質良好	標題品質欠佳
馬路比台灣爛的國家很多嗎	[F B] 黑人 陳建州
OECD 公布全球 38 國工時 台灣高於平均近 400	正晶限時批-限時批政經

### 3.3 Document Clustering

本研究將討論相似內容的多篇文章定義成一個主題，而一篇文章會有一個主要的討論主題。分群目標是希望能夠將屬於討論同一主要主題的輸入文章分在一起，藉此來得出主題清單以利計算各主題熱門度。

理想上的主題清單應該要有兩個特性：同一主題內的文章彼此相似度應該高度相關；而不同主題之間的相似度應該要相對較低。Hierarchical Agglomerative Clustering(HAC)是一個相當直覺且適合的方法來達成這個目標，只是本研究不以最終群的個數作為終止條件，而是以相似度門檻值為中止條件：定義一個相似度門檻值，找出最相似的兩個群，若相似度高於門檻則合併兩群，重複合併直到任兩群相似度低於門檻值。

這裡的相似度門檻值與主題的抽象程度有關，若設定的門檻值越高，主題彼此越不容易合併，但合併主題彼此之間的相似度會較高。兩相似度越高的主題合併後主題內文章彼此相關性也會較高。主題內文章相關性高代表其主題所涵蓋的範圍較為精準，換句話說，主題的抽象程度較低；若設定的門檻值越低，主題就越容易合併，主題內的文章彼此相關性會較低，主題涵蓋範圍較廣泛，抽象程度較高。與其他分群方法一樣，分群方法參數的決定相當不易。本研究藉由人工標記資料的方式來驗證分群效果並藉由分群效果來決定門檻值，較為不同的是，門檻值是具有主題抽象程度意義的。這意味著，往後以人工標記後所找出來的門檻值，對相同資料來源的文章分群，其主題的抽象程度會與人工標記主題之抽象程度相當。

#### 3.3.1. Centroid Linkage Method

如相關研究 2.2 介紹，使用 average, single 或是 complete linkage 方法的 HAC 在計算群之間的距離會先計算文件與其他群的文件兩兩之間的相似度。這樣的好處是能確保群在合併的過程中各文件的相似度依然可以保持在一定的範圍內，不過缺點就是必須要將所有文件與其他文件的相似度先計算過一遍，速度較慢。centroid linkage 方法以群的 centroid 來代表群裡面的文件集合，由於不必計算兩兩文件的相似度，因此在計算速度上較有優勢，不過因為是以重心代

表群裡所有的文件，因此在合併過程中無法保證群裡各文件相似度皆能保持在一定的範圍之內。如圖 3.3 所示，群 X 與群 Y 原先兩兩文件距離相近且各文件與其重心距離也相近，合併過後無論是兩兩文件平均距離或是各文件與重心距離也都相對變遠，這個變化隨著合併次數而趨於明顯。

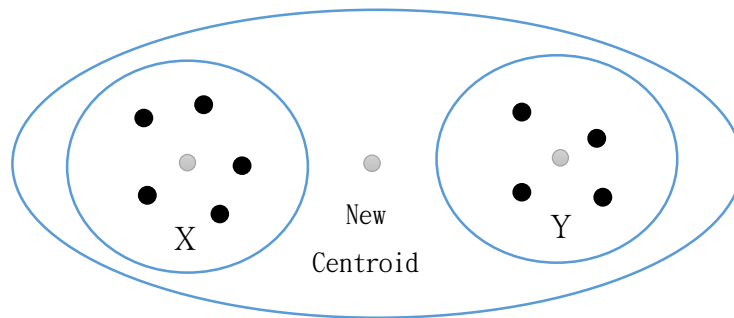


圖 3.3 centroid linkage 方法合併示意圖

HAC 雖然在分群效果表現上不錯，但是時間複雜度卻很高，雖然 centroid linkage 方法能夠加快相似度的計算，但分群效果並不若其他 linkage 方法穩定。偵測熱門主題要處理的文章量不少，分群方法的效率更顯得重要，因此本研究提出一個別於傳統的相似度測量方法，藉此保留 centroid linkage 方法在速度上的優勢並改善 centroid linkage 方法在分群結果上的隨著合併次數增加而讓平均兩兩文件相似度變低的缺點。

### 3.3.2. Dot Similarity Measure

在傳統的 tf-idf vector space model 中，普遍都是使用 cosine similarity measure 來測量文件的相似度，文件向量的夾角越小表示兩者越為相似。如圖 3.4 所示，向量 A 與向量 B, C 夾角一致，因此以 cosine similarity measure 來測量 A 與 B 文件相似度會和 A 與 C 的文件相似度相等，但通常在這個情況下 A 與 B 之間直線距離較短，直覺上 A 與 B 的相似度會較高。



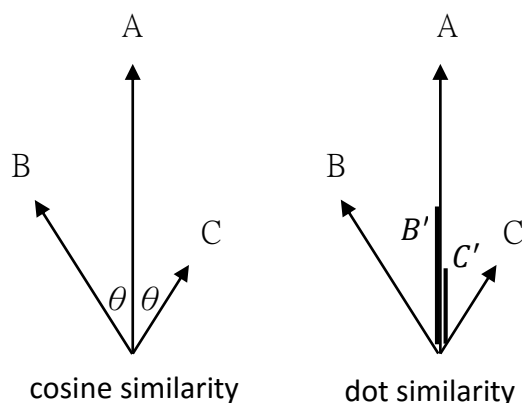


圖 3.4 cosine 與 dot similarity measure 示意圖

會有這個現象發生是因為 cosine similarity measure 只有單純考慮向量夾角而沒有考量向量本身長度。但是在傳統的 term-document 向量空間下，向量長度的意義與文件所使用的 term 種類多寡有關。在 idf 相同的情況下，由一種 term 所組成的文件向量長度會比由兩種 term 所組成的文件向量還要長。因此若不對向量長度正規化的話，相似度計算會對 term 較多的文章較為不利。另一方面，本研究所提出的向量表示方法由固定數量關鍵字之詞向量所組成，詞向量各元素由各神經元所決定，其向量長度及角度都隱含了語法及語意上的意義。因此本研究對 cosine similarity measure 做了一點修正，除了向量角度之外，將向量長度也考量進來，提出了以向量投影長度作為相似度(dot similarity measure)的概念，其計算可以簡單的向量內積而成：

$$\text{Similarity}(D_A, D_B) = \vec{v}_A \cdot \vec{v}_B = |\vec{v}_A| \cdot |\vec{v}_B| \cdot \cos \theta \quad (3.4)$$

其中  $D_A$  為文件 A； $D_B$  為文件 B； $\vec{v}_A$  為文件 A 之文件向量； $\vec{v}_B$  為文件 B 之文件向量； $\theta$  為兩文件向量之夾角。若以內積相似度測量方法來比較圖 3.4 之文件向量，則僅需比對 B 投影到 A 及 C 投影到 A 的長度即可， $B'$  長度大於  $C'$ ，因此 A 與 B 文件相似度高於 A 與 C 之文件相似度。

由於內積的值域並沒有一個固定範圍，因此需要在分群開始前將各文章向量正規化將值域限縮在-1 到 1 間，長度為一的向量內積效果等同於 cosine similarity measure。合併後的群向量

會比原個別的群向量還要短。如圖 3.5 所示，假設群 A 與群 B 大小相同，使用 centroid linkage 的分群方法其群合併後的向量取原兩群向量的平均。相對於合併前的群向量 A 與群向量 B，合併後的向量長度較短，其縮短的程度與群 A 與群 B 的相似程度有關。若兩群越為相似(群向量 A 與群向量 B 夾角越小、長度越相近)則合併後的向量長度縮短程度較小；若兩群向量越不相似，則合併後向量長度縮短的程度也就越大。

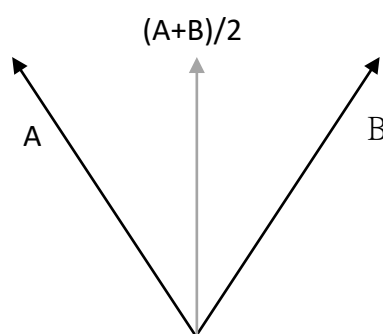


圖 3.5 群的合併示意圖

對以內積相似度測量方法搭配 centroid linkage 方法的 HAC 而言，會優先挑相似度高的兩個群合併。而隨著群合併次數的增加，群向量長度將會漸漸縮短。又內積相似度測量方法的計算明顯對長度長的向量有利，因此會優先挑選合併次數相對少的兩群合併。如 3.3.1 所提及的，centroid linkage 方法有隨著群合併次數增加而讓群內兩兩文件相似度平均變低的缺點。使用 cosine similarity measure 搭配 centroid linkage 方法因為相似度只考慮向量夾角以及合併後的向量是取其 centroid vector，所以有可能會發生某個群合併了大部分其他較小群的狀況，最大的那個群由於合併次數過多而讓群內兩兩文件相似度平均變低。使用 dot similarity measure 能夠讓每群合併的次數維持在較為平均的狀態，因此能夠獲得較穩定的分群品質

### 3.4 Popularity Calculation

針對分群後的主題清單，分別計算主題熱門度。主題熱門度以使用者的正面評價數量與負面評價數量相減而成。以本研究使用的批批踢實業訪八卦版測試資料集為例，可以藉由文章本

身的”推”以及”噓”分別代表正面評價以及負面評價。統計主題內”推”與”噓”的個數並相減來計算主題之熱門度。最後選出熱門度最高的前五大主題作為輸入文章的熱門主題。



## 第四章 實驗與討論

### 4.1 資料來源

本研究以批踢踢實業坊(PTT)八卦版之文章作為資料來源，並蒐集 2015 年 4 月 9 日至 2016 年 6 月 28 日約 92 萬篇文章作為 Word2Vec 的訓練資料集。

PTT 是以學術性質為目的，而以電子佈告欄系統 (BBS, Bulletin Board System) 為主的討論平台。目前在 PTT 註冊的人數超過一百萬人，尖峰時段容納超過十五萬名使用者同時上線。八卦版是 PTT 最熱門的看板之一，熱門時段約有數萬人同時在看板上，反黑箱服貿事件時更有十萬人同時在看板上的紀錄。看板討論內容五花八門，包含政治、社會、或者網友一些小道消息，討論主題常與社會連動，影響力可見一斑[30]。

如圖 4.1 所示，PTT 文章主要由標題、內容以及評論所構成。使用者可藉由”推”、”噓”、”箭號”表示正反及中立評論，本研究亦利用推噓評論指標作為熱門度的主要依據。

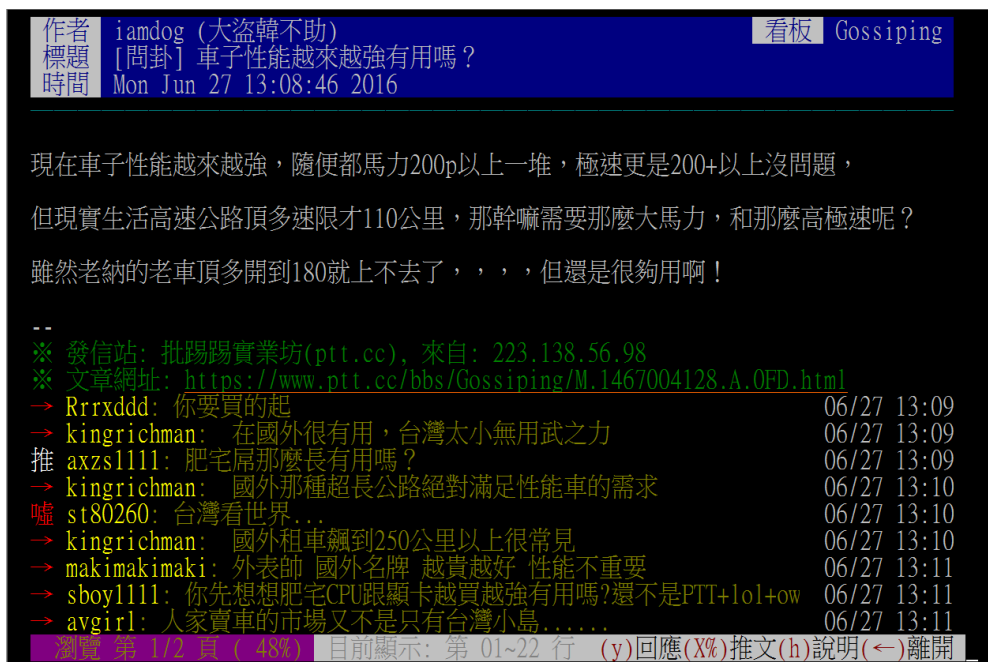


圖 4.1 PTT 八卦版文章示意圖

## 4.2 Word2Vec Training

實驗之 Word2Vec 訓練資料集來源為 PTT 八卦版文章約 92 萬篇，將文章內容濾除網址以及使用者引言後作進行訓練。本實驗使用 Gensim Word2Vec Lib[31]來訓練 word2vec model。Gensim Word2Vec model 訓練時有許多參數可以設定，本研究使用 Skip-gram 的 model algorithm(sg=1)，並將詞頻低於 5 以下的字詞忽略不計算(min\_count=5),其餘訓練參數皆使用預設值。

## 4.3 分群效果評估

本研究設計了一系列的實驗來探討不同的參數分別會對整體的分群效果造成什麼影響。分別是 linkage 方法與 similarity measure 的比較；不同的 feature extraction 方法比較；標題向量與內容大意向量比例的比較。

分群效果評估方式主要可以分成 External 與 Internal 兩種，External Validation 藉由外部所標記的正確答案直接比較分群結果；Internal Validation 藉由計算群與群之間的分散程度以及群內部的聚合程度來評估分群效果。External validation 雖然所需的標準答案需要仰賴人工標記，想要標記數目不小的答案並不太容易，但也因為能夠直接比對標準答案，所以得出來的評估值較 Internal Validation 準確。為了追求較準確的評估值，本實驗使用 External validation，人工標記了三組資料集，並以此資料集進行一系列的分群效果實驗。


### 4.3.1. 評估指標

本研究藉由人工標記資料對分群結果以 Adjust Rand Index (ARI), Adjust Mutual Information (AMI)等常見的 External validation index 指標進行評估。

Rand Index(RI)[32]藉由兩兩文件的比對來衡量分群結果與 ground truth 的相似度，其數學定義為：

$$RI = \frac{a+b}{\binom{n}{2}} \quad (4.1)$$

其中  $n$  為文件個數， $\binom{n}{2}$  為任意兩兩文件的所有可能組合， $a$  為兩文件在分群結果以及 ground truth 都屬於同一個群的個數， $b$  為兩文件在分群結果以及 ground truth 都不屬於同一個群的個數。ARI[33] 是 RI 的 chance normalization 版本，chance normalization 能夠修正分群結果與標準答案群的數量差異過大而讓分數失去參考性的問題[34]。計算 ARI 之前必須先建立 contingency table，如圖 4.2。其定義為：給定一個大小為  $n$  的集合  $S$ ，以及兩個 groups  $X$  與  $Y$  (如 ground truth 跟分群後的群集合)， $X=\{X_1, X_2, \dots, X_r\}$ ， $Y=\{Y_1, Y_2, \dots, Y_s\}$ ， $X$  與  $Y$  重疊的部分可以組成 contingency table[nij]，其中  $n_{ij}$  代表集合  $S$  同時在  $X_i$  與  $Y_j$  的元素個數。



$X \setminus Y$	$Y_1$	$Y_2$	$\dots$	$Y_s$	Sums
$X_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1s}$	$a_1$
$X_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2s}$	$a_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X_r$	$n_{r1}$	$n_{r2}$	$\dots$	$n_{rs}$	$a_r$
Sums	$b_1$	$b_2$	$\dots$	$b_s$	

圖 4.2 contingency table

而 ARI 的數學定義如下：

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} = \frac{\sum_{i,j} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}} \quad (4.2)$$

其中  $n_{ij}, a_i, b_j$  的值由 contingency table 而來。

Mutual Information(MI)[35]則是以 entropy 角度的一種分群評估指標。給定一個大小為  $N$  的集合  $S$ ，以及兩個集合  $U$  與  $V$ ， $X=\{X_1, X_2, \dots, X_R\}$ ， $Y=\{Y_1, Y_2, \dots, Y_C\}$ ，其 MI 的數學定義為：

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log \left( \frac{P(i, j)}{P(i)P'(j)} \right) \quad (4.3)$$

其中  $P(i, j) = |U_i \cap V_j| / N$ ， $P(i) = U_i / N$ ， $P'(j) = V_j / N$ 。如同 ARI，AMI[34] 是 MI 的 chance normalization 版本。其數學定義為：

$$AMI(U, V) = \frac{MI(U, V) - E[MI(U, V)]}{\max(H(U), H(V)) - E[MI(U, V)]} \quad (4.4)$$

其中  $H(U)$  以及  $H(V)$  是  $U$  以及  $V$  對應的 entropy

$$H(U) = -\sum_{i=1}^{|U|} P(i) \log(P(i)) \quad (4.5)$$

$$H(V) = -\sum_{j=1}^{|V|} P'(j) \log(P'(j)) \quad (4.6)$$

與 ARI 相同，(4.4) 的  $E[MI(U, V)]$  可以由對應的 contingency table 獲得：

$$E[MI(U, V)] = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i, b_j)} \frac{n_{ij}}{N} \log\left(\frac{N \cdot n_{ij}}{a_i b_j}\right) \cdot \frac{a_i! b_j! (N-a_i)! (N-b_j)!}{N! n_{ij}! (a_i-n_{ij})! (b_j-n_{ij})! (N-a_i-b_j+n_{ij})!} \quad (4.7)$$

其中  $(a_i + b_j - N)^+$  為  $\max(1, a_i + b_j - N)$ ， $n_{ij}, a_i, b_j$  的值由 contingency table 得到。

### 4.3.2. 資料集

實驗資料集共有三組，分別為八卦版 2016 年 6 月 24 日(資料集 A)、2016 年 6 月 15 日(資料集 B)以及 2016 年 6 月 29 日之文章各一千篇，人工標記資料集文章所屬討論主題，並以此為標準答案進行分群效果評估。

表 4.1 主題文章列表範例

文章標題	同標題文章數
他要求關版「洪素珠出現都是因批踢踢」	2
如果 PTT 被無預警強制關站	5
「與台灣民政府有驚人共通點」 藍軍要求	5
正晶限時批	1
Fw: [Live] 20160615 正晶限時批	1

中華文化復興委員會(要求關閉 PTT)	3
真天才！中華文化復興委員會要求關閉 PTT	1
中華文化復興委員會要求關 PTT	1

一個主題由許多文章組成，其文章量的多寡代表了主題規模的大小。表 4.1 為資料集 B 某一討論 PTT 關站主題之相關文章列表，可以觀察到此主題包含許多標題不同的文章，並不是單純將相同標題的文章歸類而成的。主題標記除了直接查看文章標題來判斷該文章是否屬於某一個主題之外，也會根據文章內容來判斷。如”正品限時批”該政論節目在 6 月 15 日當天討論內容與關閉 PTT 相關，所以相關討論也被歸類在此主題。

2016 年 6 月 24 日發生了兩起重大事件：桃園空服員職業工會宣布罷工以及英國脫歐公投結果出爐，因此取此一時間點作為重大事件發生的取樣資料集 A；實驗也取了無特別重大事件發生也無特別話題聚焦的 6 月 15 日作為資料集 B；第三個資料集日期取 6 月 29 日，當天討論話題聚焦在前幾天發生的事件，分別是”軍方犬隻造冊列管”以及”一例一休爭議”。

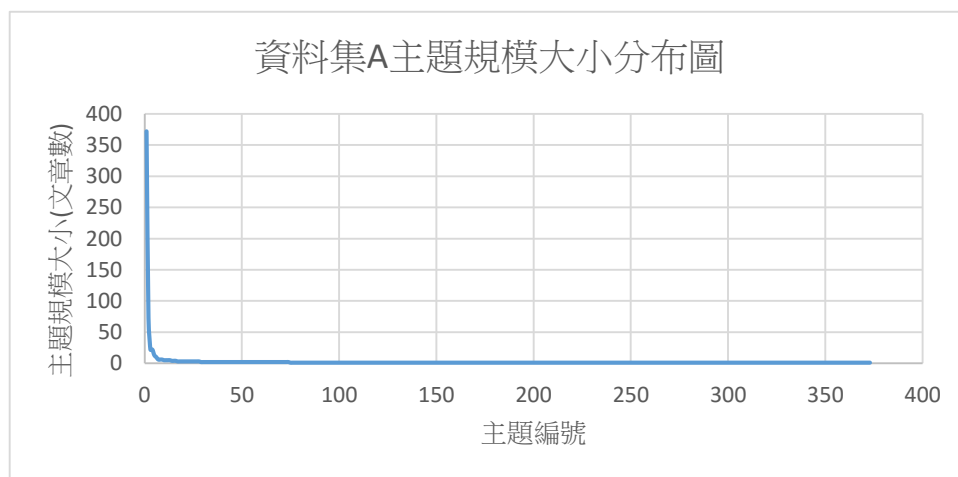


圖 4.3 資料集 A 主題規模大小分布圖



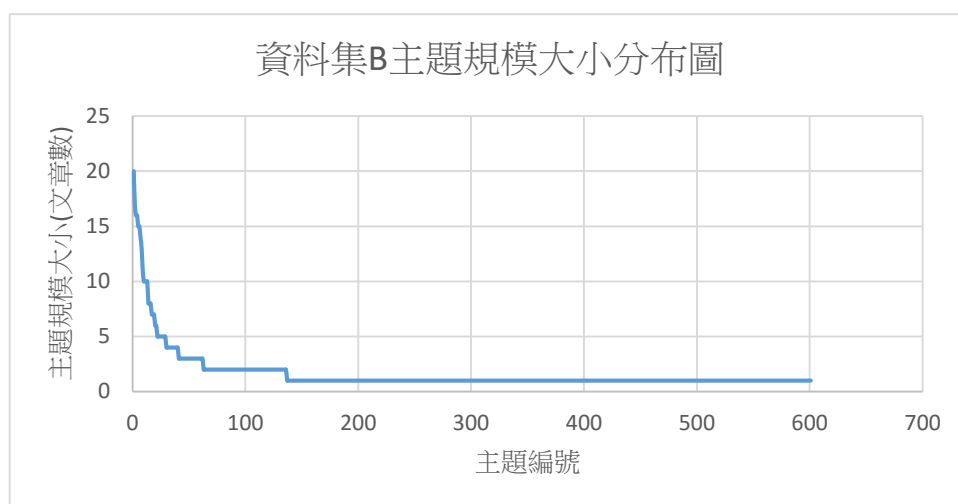


圖 4.4 資料集 B 主題規模大小分布圖

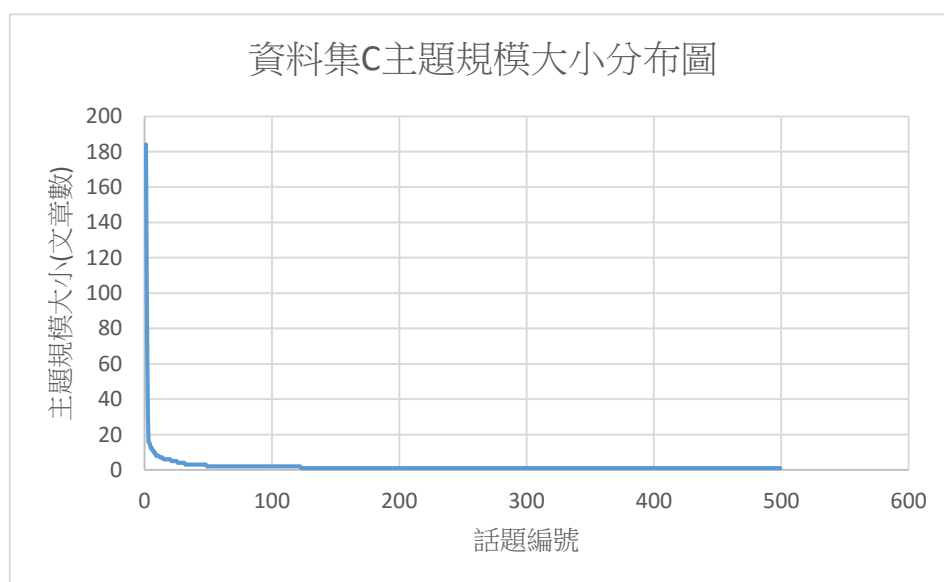


圖 4.5 資料集 C 主題規模大小分布圖

圖 4.3 顯示有重大事件發生的時候，文章討論幾乎集中在少數幾個主題，其餘討論主題的規模大小都不大。資料集 B 相較於資料集 A 較無重大事件發生，圖 4.4 顯示討論主題與資料集 A 相比較為發散，數量也較多。圖 4.5 顯示資料集 C 則介於 A 與 B 之間，雖然沒有發生特別重大的事件，但明顯有一個聚焦的討論主題。

### 4.3.3. Linkage 方法與 Similarity Measure 比較

傳統的 HAC 有許多不同 linkage 方法，本實驗挑選了常見的四種 linkage 方法運用在本論文提出的分群方法並觀察哪一種 linkage 方法對於分群效果會較好。實驗依據不同關鍵字擷取方法、擷取關鍵字的個數、關鍵字的權重有無，一共設定共十五組的不同參數組合。在這十五組參數下，來觀察使用不同的 linkage 方法或不同的 similarity measure，在不同狀況下個別最佳門檻值的 AMI 與 ARI 變化。

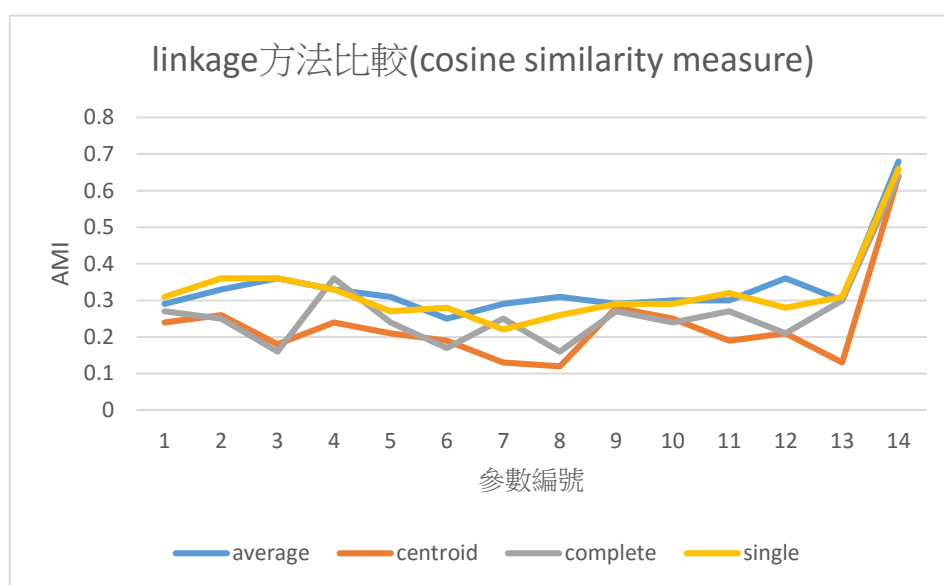


圖 4.6 Linkage 方法比較 with Cosine Similarity measure (1)

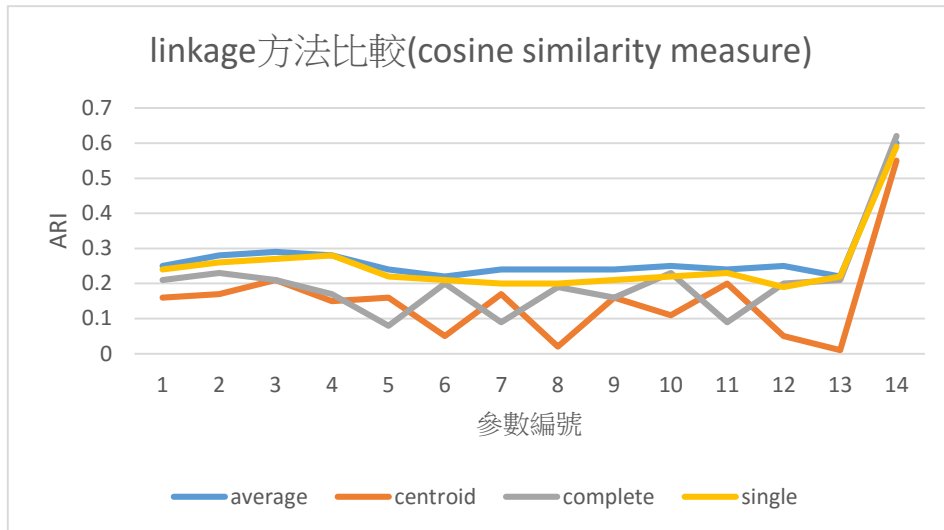


圖 4.7 Linkage 方法比較 with Cosine Similarity measure (2)

觀察圖 4.6 及圖 4.7，ARI 與 AMI 兩個指標都顯示出以 cosine similarity 為相似度測量方法的分群使用 average linkage 方法的效果在不同狀況下平均效果較好，其次是 single linkage 方法，而 centroid linkage 方法除了不穩定之外、效果也較差。

將相似度測量方法改為 dot similarity measure 再次進行觀察，如圖 4.8 及圖 4.9。無論是 ARI 或是 AMI 指標都顯示：與 cos similarity measure 相比，centroid linkage 方法搭配 dot similarity measure 之後分群品質能夠大幅提升，其分群效果大致接近於 average linkage 方法。

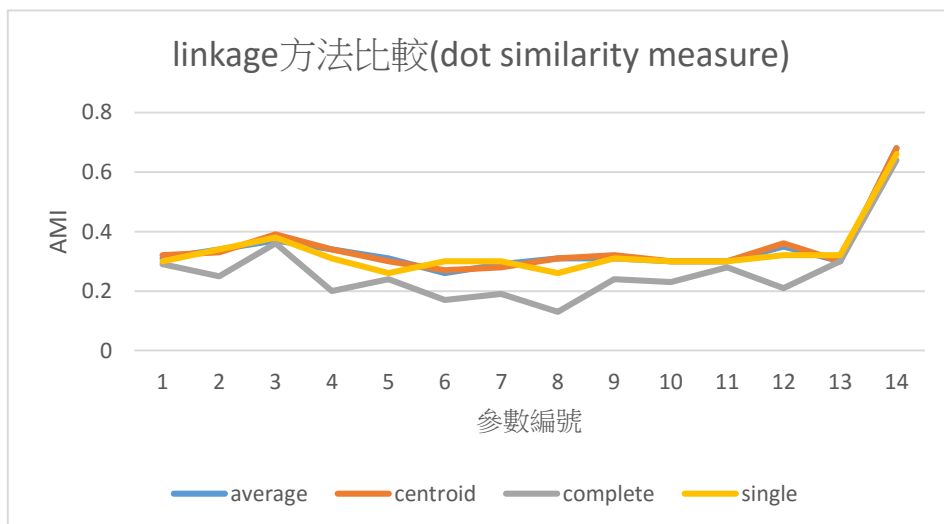


圖 4.8 Linkage 方法比較 with Dot Similarity measure (1)

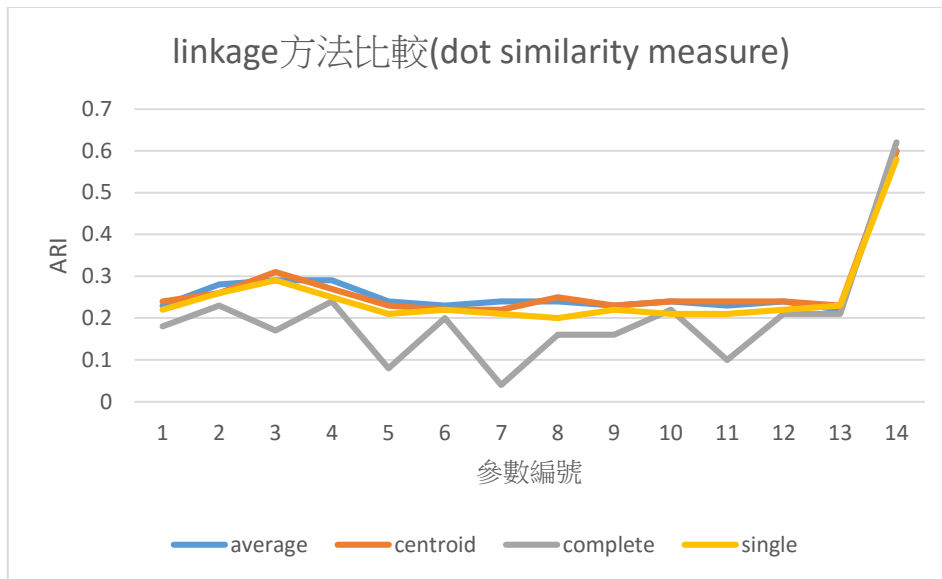


圖 4.9 Linkage 方法比較 with Dot Similarity measure (2)

將 centroid linkage 方法與 average linkage 方法特別挑出來比較有無套用 dot similarity measure 的分群效果以及分群所花費的時間，如圖 4.10、圖 4.11 以及圖 4.12。

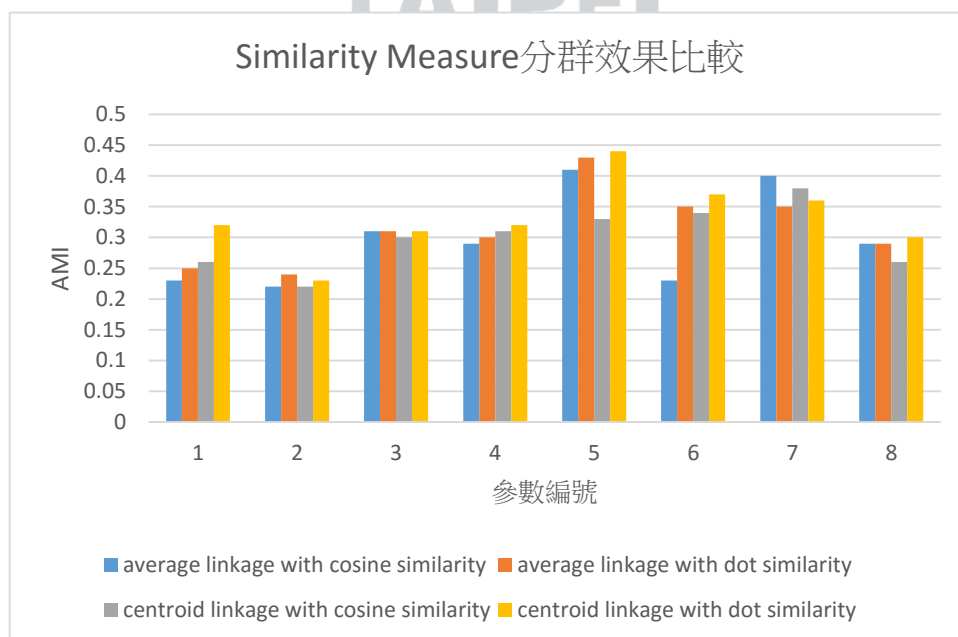


圖 4.10 Similarity Measure 分群效果比較(1)

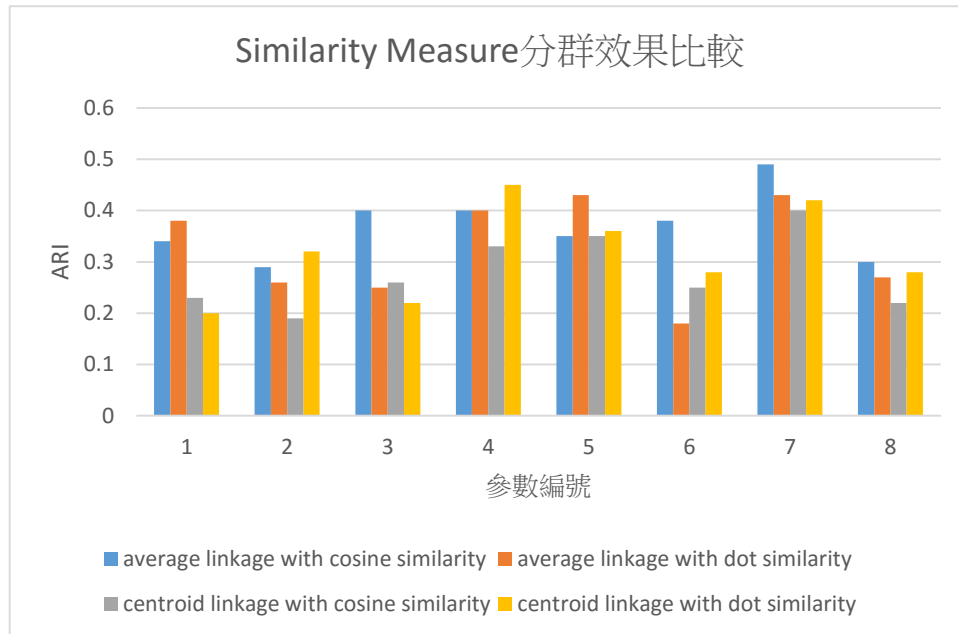


圖 4.11 Similarity Measure 分群效果比較(2)

圖 4.10 及圖 4.11 顯示 average linkage 方法改用 dot similarity measure 不一定能夠讓分群效果變好，但對 centroid linkage 方法而言，大部分的情況能有效提高分群效果，但其分群效果並不一定是最好的。

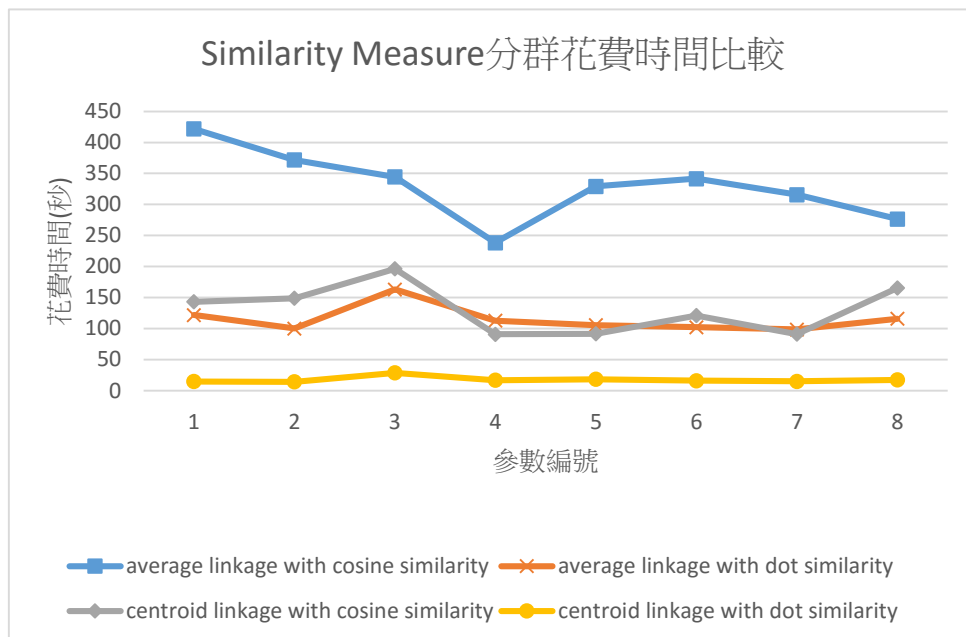


圖 4.12 Similarity Measure 分群花費時間比較

圖 4.12 顯示 average linkage 方法分群花費時間比 centroid linkage 方法多；cosine similarity measure 花費時間又比 dot similarity measure 多。與 cosine similarity measure 搭配 average linkage 方法相比，使用 dot similarity measure 搭配 centroid linkage 方法將可以減少約 20 倍的花費時間。

以上實驗結果可以證明 3.3.2 的結論：使用 dot similarity measure 搭配 centroid linkage 方法能夠維持速度上的計算優勢也能夠保持一定的分群品質。雖然這樣的組合分群效果不一定是最佳的，但是在分群速度上具有絕對的優勢。

#### 4.3.4. Feature Extraction 比較

Feature Extraction 與 Vector Representation of Article 中提到本研究使用基於 TF-IDF 的關鍵字擷取方法搭配對應的關鍵字權重組合成內容大意向量。本實驗加入以 LDA 為關鍵字擷取方法對照組，並比較有無考量特徵權重搭配不同關鍵字擷取方法在取不同關鍵字數量(K 值)下的分群效果。前小節證明了 centroid linkage 方法搭配 dot similarity measure 在速度上的優勢以及分群效果，所以本實驗繼續沿用 centroid linkage 方法與 dot similarity measure 作為分群方法。

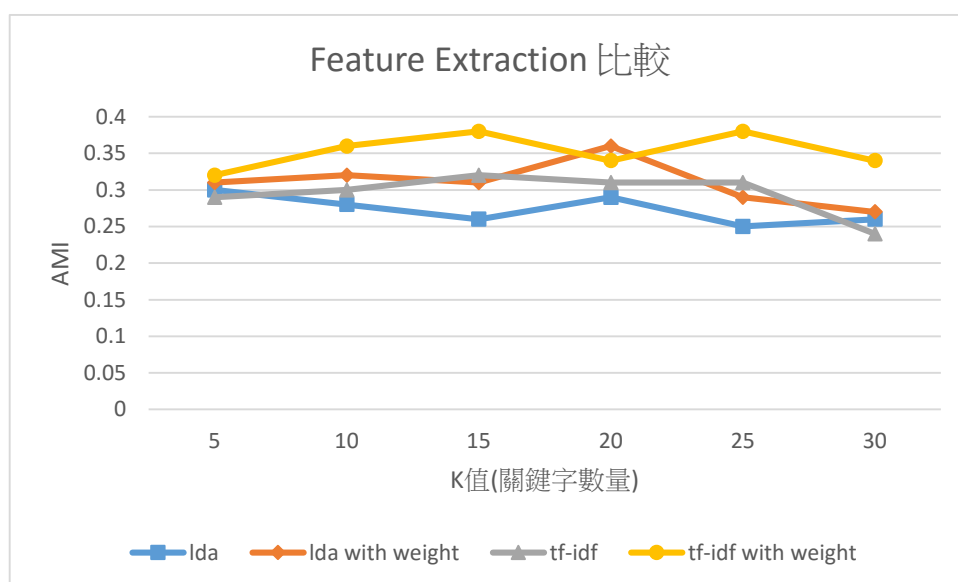


圖 4.13 Feature Extraction 比較(1)

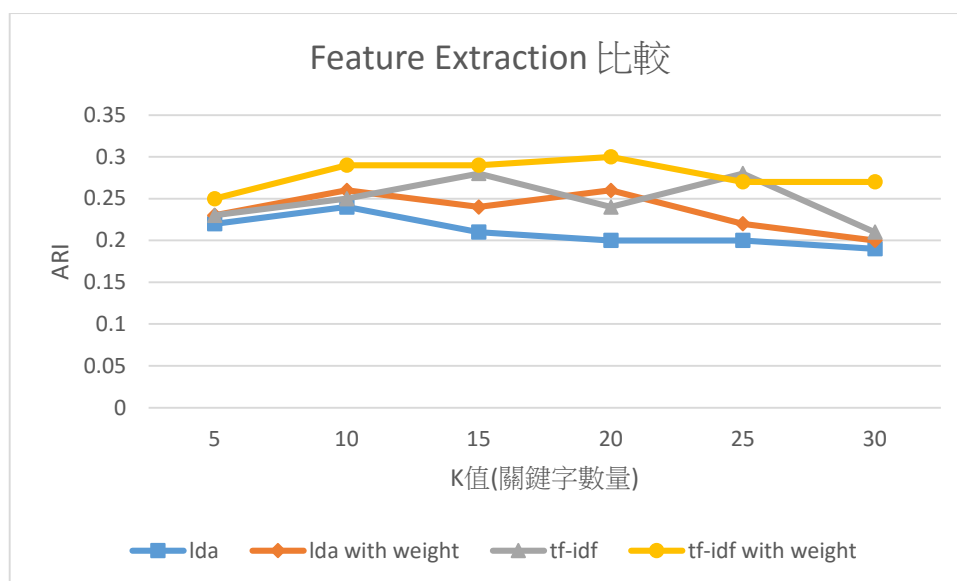


圖 4.14 Feature Extraction 比較(2)

圖 4.13 以及圖 4.14 可以觀察出在兩個評估指標下 TF-IDF 關鍵字擷取方法比 LDA 還要好；而有考量關鍵字權重的擷取方法又比沒有考量的好。整體來講，使用 TF-IDF 關鍵字擷取方法並且考量關鍵字權重的擷取方法效果會是最好的。

另外，可以觀察出關鍵字數量  $K$  等於 5 的時候有無考量關鍵字權重的差異並不太大，原因是關鍵字數量少的时候彼此之間的權重差異本身就不會太大。各方法擷取不同的關鍵字數量分群效果雖然各有差異，但關鍵字的數量多寡對分群效果彼此差異不大，關鍵字數量的影響不像關鍵字擷取方法般大。即使如此，還是可以找到一個合適的關鍵字數量  $K$  讓後續分群有個依據。觀察圖 4.13 以及圖 4.14，不同方法的極值所對應的關鍵字數量  $K$  都不太相同。上個段落提到 TF-IDF 關鍵字擷取加上考量權重的方法效果最好，可以發現  $K$  為 15, 25 時有較大的 AMI； $K$  為 20 時有最大的 ARI 值，而  $K$  為 10, 15 之 ARI 值與極值差異並不大。綜合考量 ARI 與 AMI 指標後，關鍵字數量  $K$  為 15 時會有比較好的分群效果。

#### 4.3.5. 標題與內容大意向量比例比較

在章節 Vector Representation of Article 中，提到文章向量是由標題向量與內容大意向量乘以對應的比例相加，比例取決於資料來源文章標題的品質，若標題品質很好則會調高標題向量

的比例；反之則調降。前一小節，本研究比較了不同的內容大意向量擷取方法對分群效果的影響。接下來，為了觀察標題及內容大意向量的最佳比例，本實驗以 4.3.3, 4.3.4 之實驗所得結論作為分群參數對三個資料集以不同的標題向量比例做文章分群並觀察其分群效果。

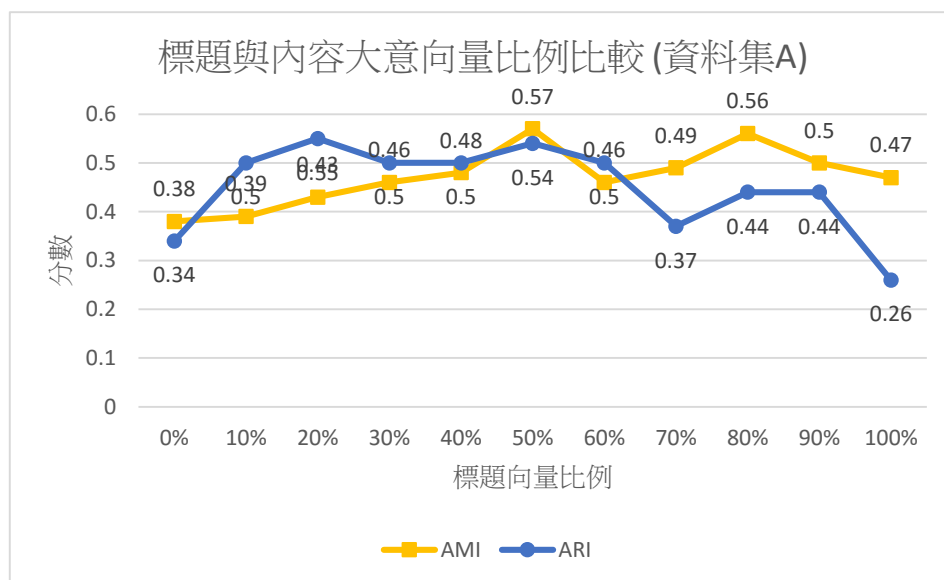


圖 4.15 標題與內容大意向量比例比較 (資料集 A)

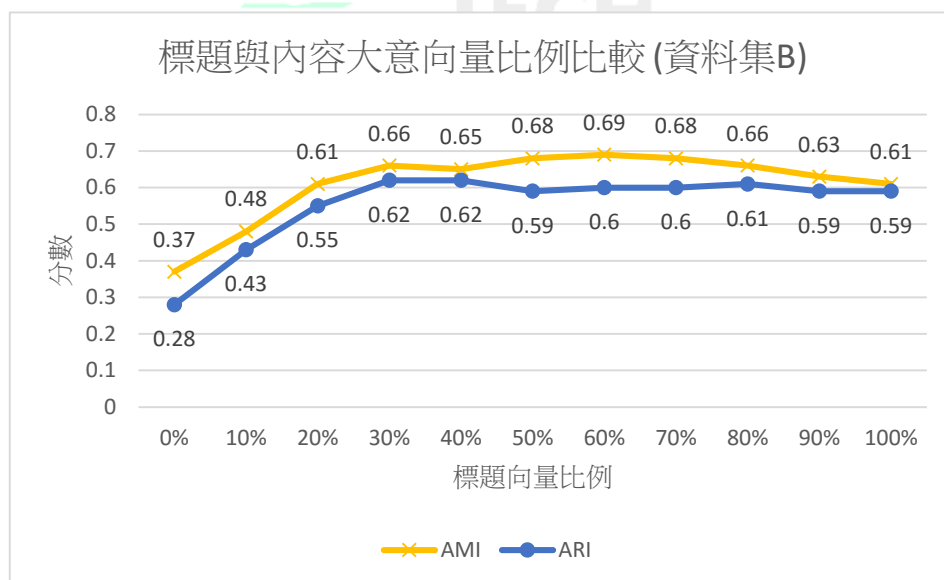


圖 4.16 標題與內容大意向量比例比較 (資料集 B)



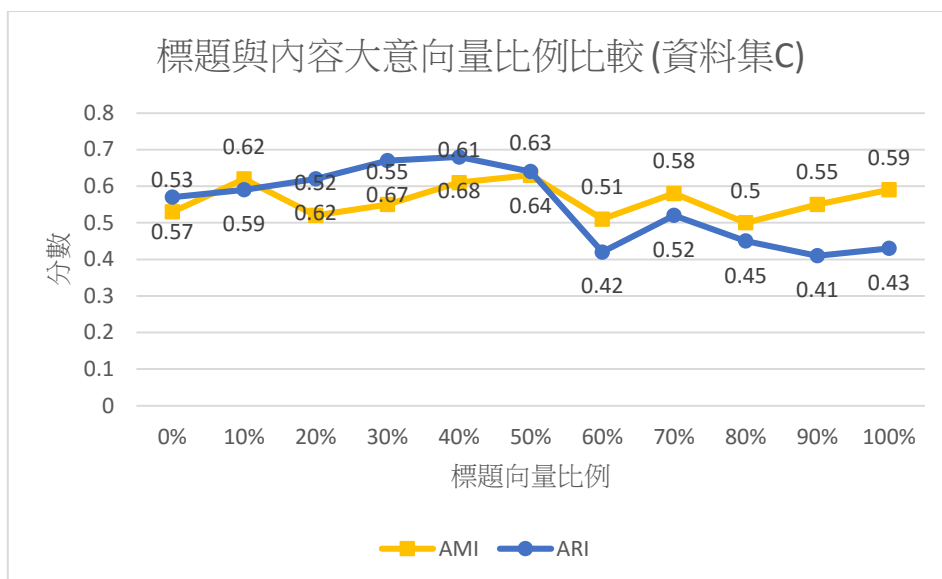


圖 4.17 標題與內容大意向量比例比較 (資料集 C)

圖 4.15、圖 4.16 以及圖 4.17 可以觀察出單純只以標題向量或內文向量合成文章向量效果都不會是最好的。資料集 A 則是在 50% ARI 與 AMI 最高，資料集 B 標題向量比例大於 30% 後的效果差異不大。資料集 C，AMI 指標並沒有一個穩定的趨勢，但 ARI 指標在 40% 左右有較好的表現。

整體來說，標題比例在 40%~60% 之間效果都不錯，比例在 50% 時效果最好。

#### 4.3.6. 分群效果比較

在此小節的實驗將進行本研究所提出的方法與傳統的 TF-IDF 文件分類方法的比較。實驗一共設立了三組對照組，並觀察這些方法在不同資料集的分群效果。對照組一(tf-idf 內文)是以將整篇文章內容計算 TF-IDF 文件向量並以 cosine similarity measure 搭配 average linkage 方法來分群；對照組二(tf-idf 標題)與對照組一相同，只是去除文章內容改以文章標題來計算文件向量，藉此來觀察是否能以文章標題為特徵就能達到非常好的效果；對照組三(word2vec)使用與前面兩對照組相同的分群方法，只是不以 TF-IDF 的空間模型向量化，而是改用本論文所提出的 Word2Vec 向量化方式。最後，是本研究所提出的方法(our method)，利用 Word2Vec 將文章特徵向量化，最後利用 dot similarity measure 與 centroid linkage 方法進行分群。

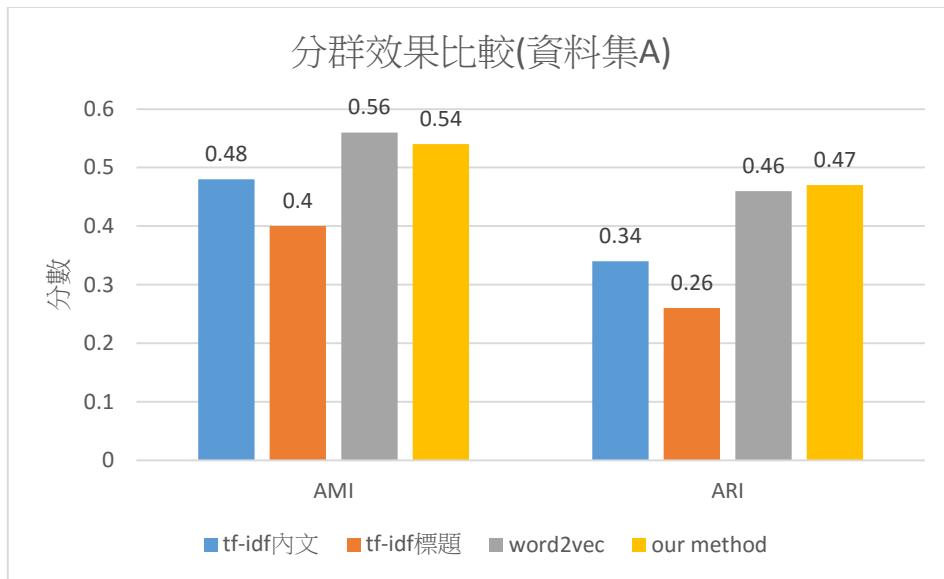


圖 4.18 分群效果比較(資料集 A)

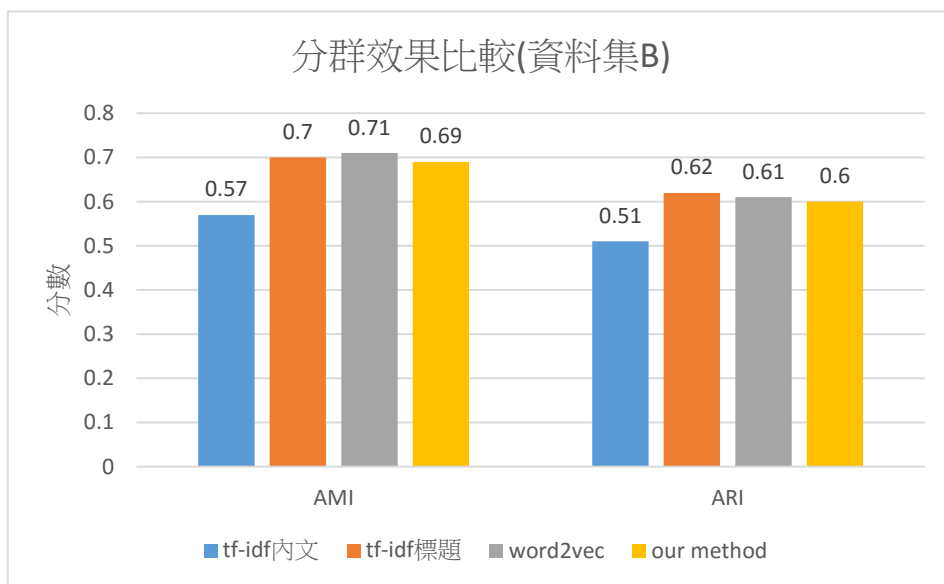


圖 4.19 分群效果比較(資料集 B)

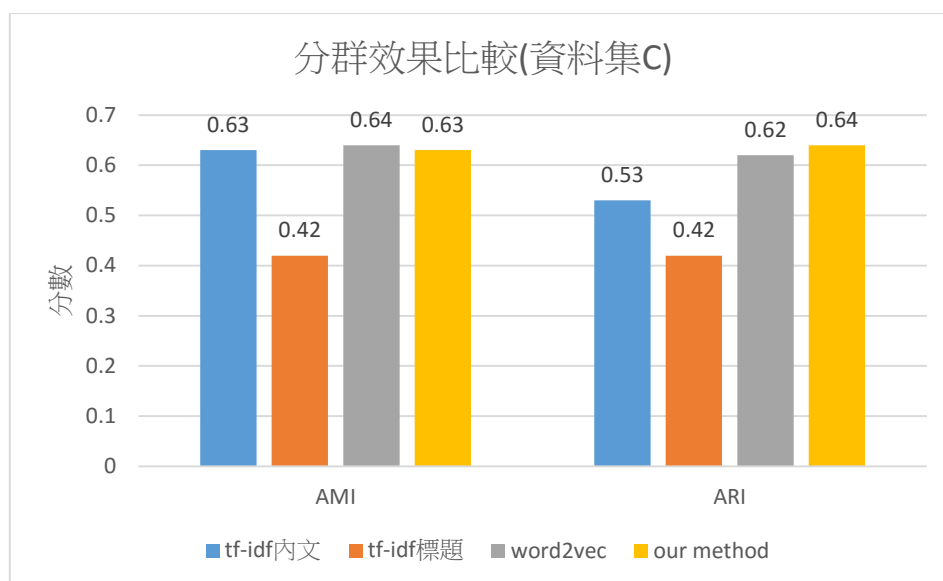


圖 4.20 分群效果比較(資料集 C)

圖 4.18、圖 4.19 以及圖 4.20 可以得知傳統方法還是具有一定的效果。如章節 3.2.3 所提及，僅使用文章標題的向量化方式很依賴資料集文章標題品質，可以看到 tf-idf 標題方法的效果在不同的資料集之間的效果並不太穩定。另外 word2vec 方法在三個資料集表現都比傳統的方法還要來的好，足以顯示運用相同的分群方法，僅改用 word2vec 來進行文章向量化就能有不錯的效果提升。our method 與 word2vec 運用了相同的文章向量化方式，差異在分群方法使用了 dot similarity measure 搭配 centroid linkage 方法。可以看出這兩種方法在不同的資料集的分群效果表現差異不多，為了瞭解各方法的執行效率，將各方法所花費的時間整理繪成圖 4.21。

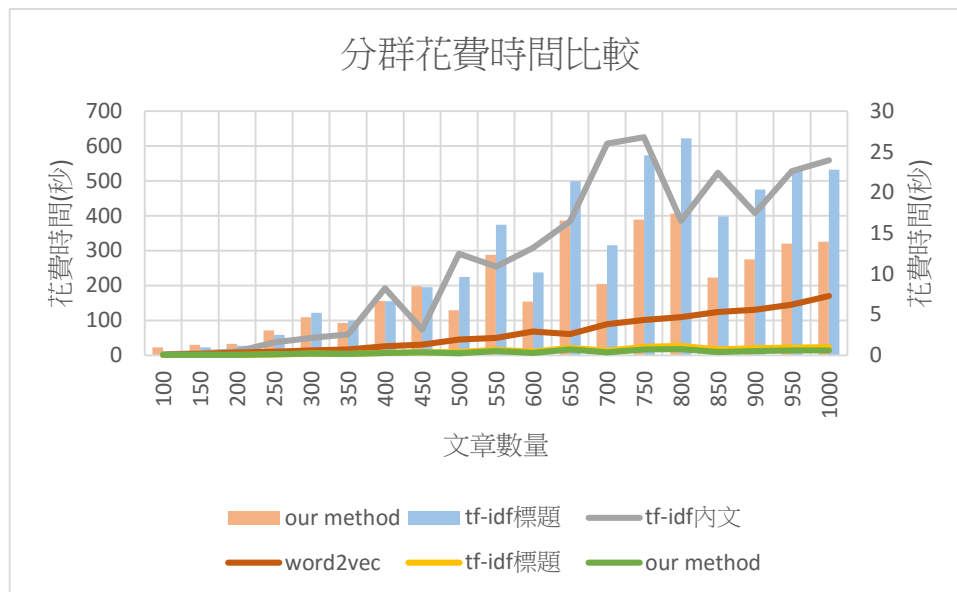


圖 4.21 分群花費時間比較

圖 4.21 折線圖部分為四個方法在不同文章量下分群所花費的時間，其中 tf-idf 標題與 our method 過於相近，所以把這兩種方法以長條圖再赴座標軸放大來比較。tf-idf 向量空間模型顯著的缺點是在向量維度會隨著字詞的量而增加，因此計算時間會隨著文章量增加而大幅上升，這也表現在 tf-idf 內文這個方法上。tf-idf 標題因為只取文章標題建向量空間，所以向量維度相對短少許多，比起 tf-idf 內文方法所花費的時間也大幅減少。word2vec 能夠將向量維度固定在一个範圍，因此花費時間比 tf-idf 內文還要來的少。our method 與 word2vec 在分群效果上差異不多，但在分群所耗費的時間差異甚大，這能夠顯示 dot similarity measure 搭配 centroid linkage 方法能夠有效的減少分群時間又能夠保持一定的分群效果。our method 與 tf-idf 標題所花費的時間相對於其他方法較少，放大後來看可以發現 our method 還是優於 tf-idf 標題，少了將近一半的時間。

傳統的方法雖然有一定的效果，但在文章量大顯得效率不高。使用 word2vec 可以改善 tf-idf 空間在文章量大時向量維度也增加的缺點，而在分群效果上也表現較好。雖然使用不同的向量空間可以減少整體運算的花費時間，但還是受限於 HAC 的效率。利用 centroid linkage 方法搭配 dot similarity measure 可以再進一步降低整體運算的時間，同時也能夠維持相當的分群品質。

## 4.4 熱門主題偵測分析

為了觀察熱門主題的偵測效果，本實驗人工挑選 2016 年 6 月 1 日至 10 日八卦版之每日五大熱門主題，並與系統偵測之五大熱門主題比較。本實驗定義主題命中率為系統標記主題與人工挑選主題相符的數量占人工挑選主題數量的比例。表 4.2 為 2016 年 6 月 1 日八卦版熱門主題範例，系統偵測之主題”仇母豬與仇女對立”與”鬧女板事件-女性真實感想”實際上同屬於人工挑選之主題”仇母豬與仇女對立(女版鬧版事件)”，而人工挑選主題”歌手隱藏神曲”沒有清楚被系統偵測，因此 6 月 1 日熱門主題命中率為 80%。圖 4.22 顯示每日主題命中率均在 60% 以上，每日平均整體命中率為 72%，顯示本研究提出的方法能夠有效偵測熱門主題。

表 4.2 2016 年 6 月 1 日八卦版熱門主題

系統偵測主題	人工挑選五大主題(沒有順序之分)
仇母豬與仇女對立	仇母豬與仇女對立(女版鬧版事件)
小時候以為很紅卻沒人看過的卡通	劉建國、李婉鈺交往風波
劉建國、李婉鈺交往風波	東海教官壓力大失控被架走
東海有教官被架走了	男童攀越圍欄掉展區，大猩猩被槍斃
鬧女板事件-女性真實感想	歌手隱藏神曲

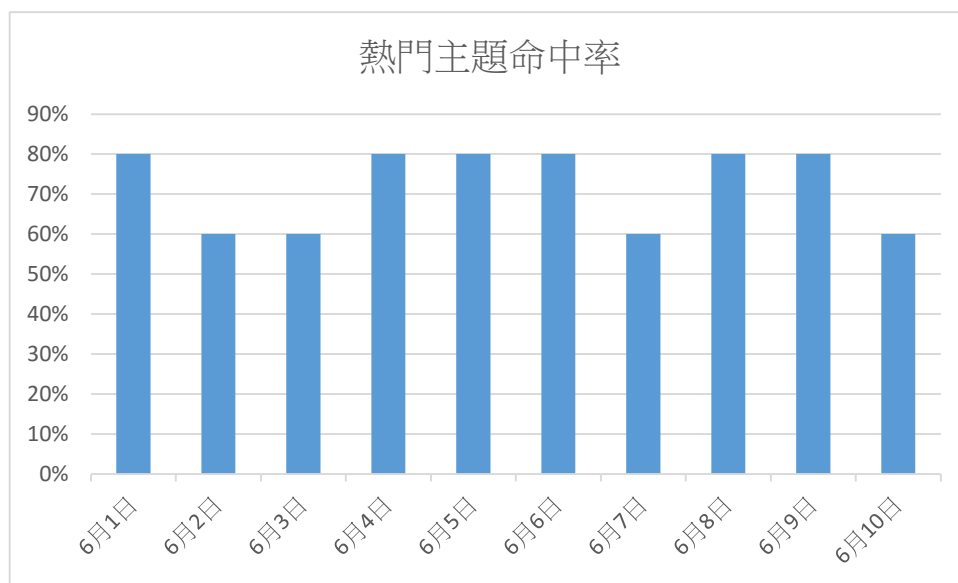


圖 4.22 熱門主題命中率



## 第五章 結論

本研究對文章內容擷取關鍵字和文章標題一同作為文章特徵，並提出一個基於 Word2Vec 的文章向量化的方法。與傳統的 tf-idf vector space model 不同，本研究提出的向量表示方法能夠將向量維度固定在一個值，而且向量隱含著語意及語法上的意義。實驗證明，以此文章向量表示法進行相似度比對具有一定的效果。

接著，本研究提出一個基於 HAC 的分群方法，將文件向量分群得出主題列表，接著對各主題計算熱門度來偵測熱門主題。HAC 分群品質相對其他分群方法穩定，但時間複雜度偏高。使用 centroid linkage 方法的 HAC 雖然能夠加快運算速度，但分群品質穩定度不甚理想，本研究利用 dot similarity measure 來提高 centroid linkage 方法的分群品質，藉此改善提高 HAC 的分群速度。實驗證明，使用 dot similarity measure 為相似度測量方法的 centroid linkage 方法與常見的 cosine similarity measure 搭配 average linkage 方法的分群品質在 ARI 與 AMI 指標下之評估效果在伯仲之間，顯示 dot similarity measure 能夠有效提高 centroid linkage 方法的分群品質。

在熱門主題偵測實驗中，系統所偵測到的熱門主題與人工標記的熱門主題相比約有七成命中率。在分群效果良好的情況下，即使本研究對於熱門度採用相對簡單的計算方法，偵測熱門主題整體而言也能有不錯的效果。

然而本研究所提的方法並非毫無限制，Word2Vec model 雖然已經經過大量資料訓練，但難免會出現沒有在訓練資料裡的新詞，而 model 無法有效的將這些新詞轉為有意義的向量。若能及時將新詞更新到原有的 model，整體效果能夠更為提升。

## 參考文獻

- [1] Google Word2vec, <https://code.google.com/archive/p/word2vec/>, Accessed: Aug. 14, 2016.
- [2] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking", Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 37-45, 1998.
- [3] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, no. 11, pp. 613–620, Nov. 1975.
- [4] Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, and X. Liu, "Learning approaches for detecting and tracking news events," IEEE Intelligent Systems, vol. 14, no. 4, pp. 32–43, Jul. 1999.
- [5] J. M. Schultz and M. Liberman,, "Topic detection and tracking using idf-weighted cosine coefficient," Proceedings of the DARPA broadcast news workshop. San Francisco: Morgan Kaufmann, pp. 189-192. 1999.
- [6] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," Information Retrieval, vol. 7, no. 3/4, pp. 347–368, Sep. 2004.
- [7] C. Wartena and R. Brussee, "Topic detection by clustering keywords." 2008 19th International Workshop on Database and Expert Systems Applications. IEEE, pp. 54-58.2008.
- [8] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse processes, vol.25, no. 2/3, pp. 259-284, 1998.
- [9] T.-C. Chou and M. C. Chen, "Using incremental PLSI for threshold-resilient online event analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 3, pp. 289–299, 2008.
- [10] T. Hofmann, "Probabilistic latent semantic indexing," Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 50-57, 1999.



- [11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993-1022, Jan. 2003.
- [12] L. Bolelli, S. Ertekin, D. Zhou, and C. L. Giles, "Finding topic trends in digital libraries," *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, ACM, pp. 69-72, 2009.
- [13] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 424-433, 2006.
- [14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Supplement 1, pp. 5228–5235, Feb. 2004.
- [15] S. Moran, R. McCreddie, C. Macdonald, and I. Ounis, "Enhancing First Story Detection using Word Embeddings," *SIGIR '16*, pp. 821-824, July. 2016.
- [16] K. Hashimoto, G. Kontonatsios, M. Miwa, and S. Ananiadou, "Topic detection using paragraph vectors to support active learning in systematic reviews," *Journal of Biomedical Informatics*, vol. 62, pp. 59–65, Aug. 2016.
- [17] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD workshop on text mining*. vol. 400. no. 1. pp. 525-526, 2000.
- [18] H. Steinhaus, "Sur la division des corp materiels en parties," *Bull. Acad. Polon. Sci* , pp. 804, 1956
- [19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. vol. 1, no. 14, pp. 281-297, 1967.
- [20] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
- [21] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics* 21, pp. 768-769, 1965.
- [22] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," *Proceedings of the 14th International Conference on Machine Learning (ML)*, Nashville, Tennessee, p.170-178, July 1997.

- [23] F. Murtagh, "A survey of recent advances in hierarchical clustering Algorithms," The Computer Journal, vol. 26, no. 4, pp. 354–359, Nov. 1983
- [24] J. Hughes, "Automatically acquiring a classification of words." PhD Thesis. The University of Leeds, p.73, 1994
- [25] T. A. Plate, "Estimating analogical similarity by dot-products of Holographic Reduced Representations," Advances in neural information processing systems (1994): 1109-1109, 1994
- [26] Jieba 斷詞系統, <https://github.com/fxsjy/jieba>, Accessed: Aug. 14, 2016.
- [27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781, 2013.
- [28] T. Mikolov, W. Yih, and G. Zweig,, "Linguistic Regularities in Continuous Space Word Representations," HLT-NAACL. vol. 13, pp. 746-751, 2013.
- [29] T. Mikolov and J. Dean, "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems, pp. 3111-3119, 2013.
- [30] 曾郁文，網路論壇中的第三人效果之研究— 以批踢踢八卦版為例，碩士論文，國立中山大學傳播管理研究所，2012
- [31] Genism Word2Vec, <https://radimrehurek.com/gensim/models/word2vec.html>, Accessed: Aug. 14, 2016.
- [32] W. M. Rand, "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical Association, vol. 66, no. 336, p. 846, Dec. 1971.
- [33] L. Hubert and P. Arabie, "Comparing partitions," Journal of Classification, vol. 2, no. 1, pp. 193–218, Dec. 1985.
- [34] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?," Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp.1073-1080, 2009.
- [35] T. M. Cover and J. A. Thomas, "Entropy, relative entropy and mutual information," Elements of Information Theory , pp. 1-55, 1991