新聞事件偵測與追蹤之分群分類演算法研究

黃純敏
國立雲林科技大學
資訊管理研究所
huangcm@yuntech.edu.tw

陳聰宜
國立雲林科技大學
資訊管理研究所
g9823738@yuntech.edu.tw

詹雅筑
國立雲林科技大學
資訊管理研究所
m10023054@yuntech.edu.tw

摘要

過去研究在進行文件群聚分析時，如以詞庫方式斷詞者，多採CKIP進行中文斷詞處理。礙於其處理傳輸量的嚴格限制，以及斷詞過於瑣碎的缺點，使得研究在處理字詞上，需多次批次上傳，斷詞結果亦需進一步過濾與合併。本研究以平行處理方式比較 CKIP 與自行開發的中文斷詞系統(Chinese Corpus Segmentation, CCS)搭配國家圖書館主題標目，做為文件分群之前置處理，研究結果證實使用專業詞庫確實可提升分群成效，事件偵測準確率高達 85%。在事件追蹤實驗中以 SVM、KNN 及 Naive Bayes 三種分類演算法做為測試評比對象，結果顯示，SVM 表現最佳，其分類準確度高達 91.33%。

關鍵字：事件偵測與追蹤、中文斷詞、分群、分類

Abstract

Numberous studies relied on CKIP to conduct term segmentation if they adopted the corpus segmentation method. Due to the transmission limitation and the need of further processing of term filtering and merging, this study suggested a professional corpus composed of subject headings along with a self-developed Chinese Corpus Segmentation (CCS). The results showed that CCS outperforms CKIP in terms of performance and term quality for the process of cluster analysis with a high precision rate of 85%. In order to provide high quality news tracking results, we compared SVM, KNN, and Naïve Bayes with regard to the accuracy of classification result. Results showed that SVM was the best among the others, with a high precision rate of 92%.

Keywords: News Event Detection and Tracking, Chinese Term Segmentation, Cluster, Classification

1. 前言

過去對於新聞文件的分群與分類處理多以主題為出發點，直到 1998 年 Allan, et al.才提出以事件(event)為分群的概念[1]，所謂事件定義為:「在某特定時間與地點所發生的事情」。如「某年某月某日受到颱風侵襲」可被視為一個新聞事件，但單獨討論「颱風」這種廣泛的議題(issue)則視為一個主題(topic)，如圖 1 所示。

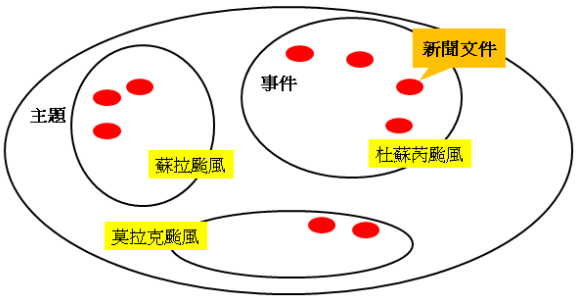

圖 1 主題、事件關係圖

新聞事件偵測與追蹤(EDT)之主要目的是

透過主動偵測事件的發生，並將同一事件的新聞群聚，爾後如有相關的報導產生則提供自動歸類到適當事件群集。其中，於事件偵測階段大多數都是以文件分群技術來實現。一般研究在進行文件群聚分析時，咸認關鍵字詞萃取的優劣影響後續分群成效。目前中文斷詞處理如以詞庫方式斷詞者，多採中研究詞庫小組(Chinese Knowledge Information Processing group, CKIP)所研發的斷詞程式，該程式的優點在於其斷詞快速兼附詞性，然而礙於每次處理傳輸量的嚴格限制，使得研究在處理字詞上，需多次批次上傳。此外，由於 CKIP 建構目的為提供一般斷詞使用，因此對於進行文件內容分析或知識萃取研究者，仍需針對斷詞結果進行過濾、合併的後續處理。是否以專業詞庫結合簡易斷詞程式，即可提升分群成效？這也是本研究希望納入實驗的議題。

此外，在新聞事件追蹤階段之呈現優劣，完全取決於事件偵測的分群結果和分類器(Classifier)的選擇。每種分類器在處理不同資料集或於不同分析環境下都有所差異，也各有優缺。然而，到目前為止，還未有研究針對新聞文件找到最適合之分類器，故本研究希透過比較 SVM、KNN 及 Naive Bayes 三種分類演算法於新聞文件分類上的表現，並挑選最佳之分類方法應用於新聞事件追蹤的呈現 。

綜上所述議題，本研究以平行處理方式比較 CKIP 與自行開發的中文斷詞系統(Chinese Corpus Segmentation, CCS)搭配國家圖書館主題標目，做為文件分群之前置處理。研究結果證實使用專業詞庫確實可提升分群成效，準確率高達 85%。在事件追蹤上以 SVM、KNN 及 Naive Bayes 三種分類演算法做為測試評比對象，結果顯示，SVM 的分類準確度最高，高達 91.33%。

本文第二節回顧與本研究相關的研究，第三節詳細描述本研究的研究架構與方法，第四節為實驗結果評估，第五節提出總結，並對往後研究發展提出未來展望。

2. 文獻探討

2.1 中文字詞處理

中文字詞處理技術的重要性在於透過分析擷取出足夠代表文件的關鍵字或特徵值。目前最常見的斷詞處理主要分為詞庫斷詞、統計斷詞與混合斷詞法。其中詞庫斷詞是以既有詞庫比對文件為取詞依據，此法擁有高品質取詞水準，但無法處理新詞;統計斷詞是以連詞詞頻為取詞辨識方式，常用來擷取新詞，但斷出之詞彙代表性與可用性備受質疑。混合斷詞法則事先利用詞庫把字詞過濾出，繼而利用統計法處理未斷出之字詞，可兼顧詞之品質與新詞之考量，此法已成為多數研究者採行的方法。

字詞分析處理中，詞庫斷詞法最常被使用，因為斷詞速度快，可提升系統之效能。近年來有關中文字詞分析相關之研究，多使用中央研究院 CKIP(Chinese Knowledge and Information Processing)詞庫小組所研發的中文斷詞系統，該系統具有未知詞辨識能力及附加詞性標記的功能，為典型的詞庫斷詞系統，包含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。然礙於每次處理傳輸量的限制，使得研究在處理字詞上，需批次上傳，而且其斷出之字詞十分瑣碎，必須再經過多重的篩選、過濾及合併，才能獲致重要關鍵字詞，因而增加後續處理的負擔。

本研究認為詞庫所蒐羅之字詞與斷詞演算法的規劃，將影響特徵值萃取的結果，進而影響後續處理結果。為驗證此一論述，本研究自行開發的中文斷詞系統(Chinese Corpus Segmentation,CCS)搭配國家圖書館主題標目，做為文件分群之前置處理。希盼專業詞庫搭配

斷詞程式可收斂關鍵詞萃取的數量，也可減輕大量資料需批次處理的時效延宕。因此實驗將以平行處理方式比較 CKIP 與 CCS，做為文件分群前置處理之差異比較。

經過斷詞處理後，仍需計算該字詞在文件中之權重，並篩選出具代表性之字詞。然而字詞權重的給定須藉由計算該字詞在單一文件的重要性(local weight)以及在整個文件集之重要性(global weight)而來。最常被採用的是TFIDF(Term Frequency Inverse Document Frequency) **[2]**，強調字詞在單一聞件出現次數愈多，且在文件集中出現次數少，則此字詞重要性愈高，公式(1)所示:

$$w_{ij} = tf_{ij} * \log \frac{N}{df_i} \qquad (1)$$

其中w_{ij}為字詞 i 在文件 j 的權重，tf_{ij}為字詞 i 在文件 j 的詞頻，df_i為字詞 i 在文件集出現的頻率，N 為整個文件集的數目。

2.2 新聞事件偵測與追蹤

由美國國防部高等研究計畫主導的「主題偵測與追蹤」(Topic Detection and Tracking，TDT)計畫，此計畫研究主題是從新聞廣播的串流中偵測及追蹤新的事件，而「新聞事件偵測與追蹤」(EDT)為其中之子項目。事件偵測可定義為:「發現包含在連續新聞串流中有關新的或先前未現的事件」[1]，為一種非監督式的學習工作，通常需分群技術來支援此階段處理。而後，事件追蹤其目的在於將後續新聞報導歸類至先前的事件群集中，為一種監督式的學習工作，也算是文件分類的一種應用。文件分群分類相關的文獻回顧將在 2.3 與 2.4 節有更深入的討論。

2.3 文件分群

分群(clustering)技術主要是將離散的資料集依某些特徵分別聚集成群[3-6]。一般而言，分群法大致可分成兩類:階層式(hierarchical)與非階層式(nonhierarchical)。階層式分群法則透過一種階層架構的方式，將資料層層反覆地進行分裂或聚合，以產生最後的樹狀(dendrogram)或巢狀(nested)架構的方式，如圖 2所示。

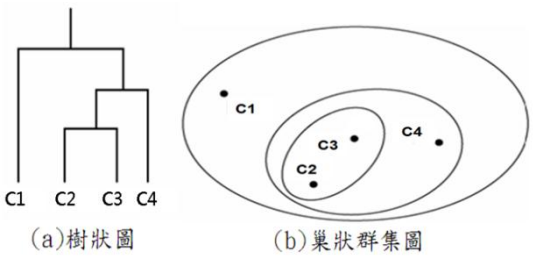

圖 2 階層式分群表示圖

以階層式分群可以清楚表達各群集間的彼此關係，更可彈性地根據不同需求，產生不同的群集數量。其優點為概念簡單，只需計算每個資料點的距離，就可建構分群結果，然而當資料量多時，處理十分費時，或者若需要某一個特定群數的分群結果，就必須從頭做起，耗時較久。其計算複雜度將以N^2的係數成長。

非階層式分群是指將資料集的 N 物件區分為 M 群。通常需先指定群數後，再用一套疊代的數學運算法，找出最佳的分群方式，將資料集分割成不等區塊以及相關的群中心，區塊間則不可重疊。其計算複雜度為$O(NM)$，遠低於階層式。

Single pass clustering 屬於非階層式分群法，由於其資料集僅需處理一次，計算效率高且複雜度低。其演算過程簡述如下：

1) 讀取文件集中第一篇文件D_1當作群集C_1。
2) 對所有文件D_i，一一與現有的所有群集作相似度的計算。
3) 依相似度計算結果，如果最大的相似度超過門檻值，指派該文件至對應的群集。

4) 重新調整新增文件的群集向量(計算質心)，否則該文件形成一個新的群集。
5) 重覆 2~4 步驟，直到沒有新進文件。

在[7] 的實驗，曾指出大量資料的即時主題偵測上，此分群法表現最佳。不少研究[8-10]也證實single-pass clustering 效能較佳。此法在執行上效率雖高，卻有分群過多的缺點。因此通常需要對所其產生的分群結果再做進一步合併。在[11]研究中，以Single pass clustering 先進行第一階段分群，再將逐一群聚相似度高的事件群集，萃取出事件與事件之間所共同描述之主題，使主題與事件形成父子階層關係，亦可達成階層效果。由於此法只允許單一文件分至相似度最高的一個群集中，故不會發生多重隸屬的分群模糊狀況，亦可稱為互斥(exclusive)或硬(hard)分群。本研究仍將沿用此種做法。由於新聞事件群集的數目會隨著時間演進或文件加入而機動調整，無法事先設定最佳群聚之數目，因此群聚數目將視實際情況再行調整。

另外，有關分群相似度的計算方法已有不少研究[12]。其中Cosine[13]函式被證實擁有較高效能表現與準確率高之計算相似度方法，Cosine如公式(2)所示:

$$\mathrm{sim}(x,c)=\frac{\sum_{j=1}^{M} w_{jx} * w_{jc}}{\sqrt{\left(\sum_{j=1}^{M} w_{jx}^{2}\right) * \left(\sum_{j=1}^{M} w_{jc}^{2}\right)}} \qquad (2)$$

其中$\mathrm{sim}(x,c)$代表新進新聞文件x對某群集c的相似度，w_{jx}為字詞j在新進新聞文件x的權重，w_{jc}為字詞j在群集c的權重，M為文件集中字詞的總數。本研究著重於分群分類之精確度，故選用此計算相似度方法。

2.4 文件分類

文件分類是將新進文件分至已事先定義好的類別中。著名的分類技術，如 SVM (Support Vector Machine)、K-NN、Naïve Bayes、Decision Tree 等，已普遍被應用於各個領域。不同研究在不同環境下得出的分類成效都有些許落差，對於哪一種分類器效果最好，都有不太一致的結論。其中，SVM 在多數研究中顯現出高精確度的分類結果[14, 15]。KNN 在資料探勘、模型辨識…等不同領域上被廣泛使用，其中在文件分類中證實擁有高準確率分類之表現[16]。Naïve Bayes 用於處理大量資料集時，能獲得準確度高且有效率的分類結果。由於本研究著重於分類之精確度，因此選用上述三種分類演算法進行本研究之事件追蹤與分類評比，希從中找出最佳之分類演算法。

2.4.1 Support Vector Machine

SVM 是一種監督式學習法，主要是針對二元分類問題，在高維度空間中尋找一個超平面作為兩類的分割，以保證分類錯誤率最低且準確度最高[17]。

SVM 將現有的資料集進行訓練，並利用這些分析出來的資料選出幾個支援向量或維度來代表整體的資料，並將少部分極端值事先剔除，然後將所挑選的支援向量或維度包裝成模型。當有測試資料需作預測時，SVM 就會將資料歸類，並利用模型將資料分成兩類，以線性可分的情況來說，假設存在訓練樣本$(x_i, y_j), \ldots, (x_j, y_j), x \in R_n, y \in \{+1, -1\}$, j 為樣本數，n 為輸入維度，假設存在一個超平面(hyperplane)能將二類樣本完全分隔，該平面描述為: $(w \cdot x)+b=0$。在二元分類中，同時希望此超平面到不同類別的距離越大越好，如此才能夠很明確的分辨這個新資料是屬於那個集合，如圖 3 所示。

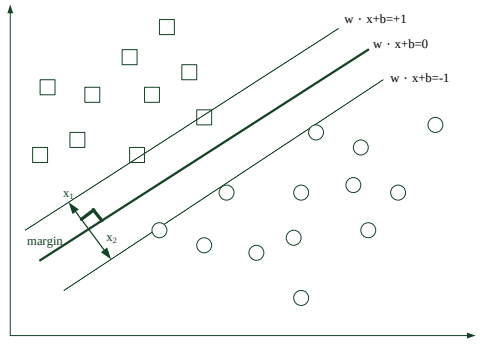

圖 3 支援向量機示意圖

2.4.2 KNN Classification

KNN 主要是利用距離的特性來計算相似度，而最常使用的則是歐氏距離[18]。KNN 其優點是簡單且容易使用[19]，能使主題追蹤發揮其獨立性且不讓非相關的主題影響作業程序。本研究採用 KNN 分類演算法於新聞事件追蹤，然而在新進文件與 k 個主題進行計算時，若要文件分類結果顯著，k 值的定義便是首要考量的因素，由於 k 值需隨不同資料集而調整變動，一般來說資料集的範圍常不固定，因此 k 值也不易確定，這將也是本研究須考慮的重點之一。

2.4.3Naïve Bayes Classifier

單純貝氏分類器主要是根據貝氏定理(Bayesian Theorem)，採用監督式的學習方式，來預測分類的結果，其運作原理是透過訓練樣本中屬性間的關聯性進行學習與記憶分類，並產生出訓練樣本的中心概念，再用學習後的中心概念，對未歸類的資料 X 進行類別 C 預測，以獲得受測試資料的目標值。貝氏定理的公式為：

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \qquad (3)$$

X 為某未歸類的資料，C 為某一類別亦即 X 屬於 C 類別的機率，如此(C 類別中出現 X 的機率)×(C 類別出現的機率)÷(X 出現的機率)。當貝氏定理應用於分類演算時，將 $X=x_1,\ldots,x_k$，其中 $x_1,\ldots,x_k$ 為未歸類資料 X 的 k 個屬性值，引進條件獨立的假設:

$$P(X=x_1, x_2,\ldots,x_k|C) = P(x_1|C) \times \ldots \times P(x_k|C) \qquad (4)$$

上述公式可以簡化為:

$$P(C|X) = P(x_1|C) \times \ldots \times P(x_k|C) \times P(C)/P(X) \qquad (5)$$

貝氏分類法應用上述公式計算出未歸類資料 X 屬於各個類別的機率，取機率值最大的類別作為 X 的類別預測。由於不同類別的機率計算中，$P(X)$的值都是相同的，因此可以忽略不計，也就是說，使 $P(x_1|C)*\ldots* P(x_k|C)*P(C)$ 值極大化的類別 C，即是未歸類資料 $X=<x_1,\ldots,x_k>$的預測類別[20]。

上述之 SVM、KNN、及 Naïve Bayes Classifier 三種分類演算法，皆各有其優點與缺點，如表 1 所示。本研究將以新聞事件偵測階段的分群結果當作實驗樣本，進而評估此三種分類演算法於新聞分類精確度及效能之表現，最後選出成效最為顯著者，作為本研究之新聞事件追蹤與呈現之分類演算法。

表 1 分類演算法比較表

	監督式/非監督式	優點	缺點
SVM	監督式	準確度高	耗時
KNN	監督式	簡單、方便使用	耗時
Naïve Bayes	監督式	較簡單	限制多

3. 研究架構與方法

本研究進行兩階段實驗，研究架構如圖 4。

第一階段為新聞事件偵測，進行『斷詞程式分群研究』；第二階段為新聞事件追蹤，進行『分類演算法研究』，詳細執行步驟敘述如下。

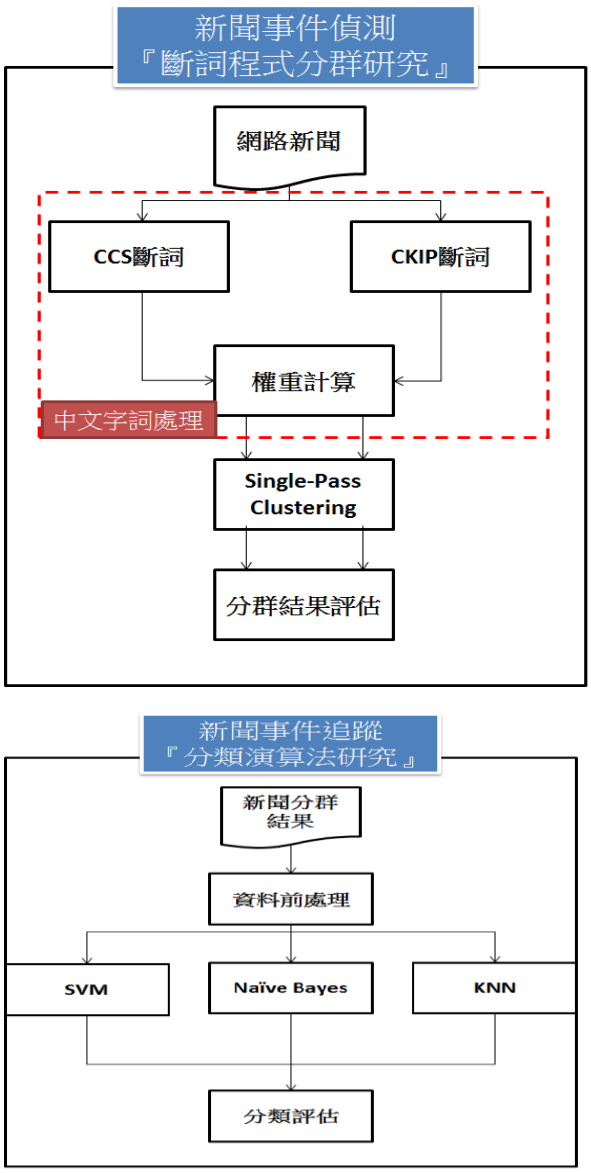

圖 4 研究架構圖

3.1 斷詞系統分群研究

本研究以 Yahoo!奇摩新聞之政治類新聞作為實驗樣本。首先將新聞內的 HTML 標籤移除，以利擷取新聞的標題、內文、以及時間等資訊。接著針對標題與內文進行字詞處理，包括「中文斷詞」以及計算「字詞權重」，最後進行事件分群及評估。

3.1.1 中文字詞處理

此步驟比較 CKIP 與 CCS 斷詞結果，以及在後續新事件偵測中，於不同門檻值設定下之分群效果之差異。以【九二共識已經發展成為兩岸關係和平發展政治基礎的重要組成部分】此一字串為例，經由 CKIP 與 CCS 處理後結果為:

CKIP	[九二][共識][已經][發展][成為][兩岸][關係][和平][發展][政治][基礎][的][重要][組成][部分]
CCS	[九二共識]已經發展成為[兩岸關係][和平]發展[政治][基礎]的重要組成部分

從斷詞結果可明顯看出，CCS 經由長詞優先處理，可保留「九二共識」和「兩岸關係」兩複合字詞，但 CKIP 則會分別將兩字詞又細分成「九二」與「共識」以及「兩岸」與「關係」。相較之下，CCS 斷出之字詞較有鑑別度，且更能具體表達此篇文章的核心概念。

完成每篇新聞文件斷詞之後，接著須經由 TFIDF 計算每個字詞於每篇文件之權重。依據本研究團隊過去研究結果顯示，若該字詞出現於標題則權重乘以十倍，若出現於第一句則權重乘以五倍，藉此強化重要字詞的重要性[11]。

3.1.2 新聞文件分群

由於權重值較低的字詞對分群結果的影響甚小，為節省系統的運算時間，本研究於每篇新聞文件取出前五個權重值最大的字詞，建立每篇新聞文件的向量空間模型(VSM)。接著運用 Single Pass Clustering，於門檻值 0.4 設定下，以 Cosine Similarity 法計算每篇文件之間的相似度後進行分群。此外，在計算新進文件

與各新聞群集的相似度時，加入時間衰退概念如公式(6)所示：

$$score(x) = \left\{\left(1 - \frac{i}{m}\right) \times sim(x, c) if c \in TW \quad (6)$$

其中 x 代表新進的文件，c 代表時間區間的其中一個群集，i 表示介於 x 以及 c 內最新的一篇文件之間所增加的文件數，m 為時間區間內所有的文件數，TW 為時間區間。在此，$score(x)$ 若大於門檻值，則標定新進文件 x 屬於舊事件；反之，若小於門檻值，則為新事件。由公式可知在時間區間內增加的文件總數越多(i 值越大，代表時間間隔越長)，新進文件與該事件群集愈疏離、越不相關。

再者，過去的研究對於事件所涵蓋的時間天數，多僅約略認為事件可能延續一週至四週，迄今仍未有過任何研究指出事件確切的天數認定，因此，在時間區間值的設定，本研究以一週為單位，輔助進行新舊事件的區隔判斷。

3.2 分類演算法研究

新聞事件追蹤主要是將新進新聞文件歸入所屬事件類別當中，故能準確的追蹤事件的發展將是本階段挑選分類演算法的考量依據。而上一階段的分群結果，將作為新聞分類的實驗樣本，從樣本中取出超過十篇以上之群集，共有 27 群，620 篇新聞。本實驗採用 weka 軟體進行 SVM、KNN 及 Naive Bayes 三種分類演算法測試。

為了轉換成 weka 軟體能處理的格式，需事先將分群結果進行轉換。首先集結每篇新聞權重最高的 5 個特徵值，以建置出特徵值與文件之對應關係，再計算每份文件出現過哪些字詞，若出現則給權重值 1，反之則為 0。如此整理出一張表格後，即可執行上述三種分類演算法。

統一匯入格式後，本研究將資料集切割成 80%為訓練資料，20%為測試資料，分別針對三種分類演算法進行實驗測試。其中，在 KNN 類演算法中，k 值得設定將會影響文件分類的精確度，因此本研究將 k 設定為 3NN、5NN、7NN、9NN 四種設定值進行分類測試。最後則比較三種演算法之分類精確度，並評選分類精確度最佳者，當作本研究之事件追蹤與分類之演算法。

4.實驗結果評估

4.1 資料集

本研究採用 Yahoo!奇摩新聞之政治類新聞中 2011 年 1 月 1 日至 2011 年 1 月 31 日，一個月份新聞量，共計 3475 筆新聞資料作為實驗樣本。

4.2 評估方法

在資訊檢索領域，準確率(Precision ratio)與召回率(Recall ratio)最被廣泛使用的評估方法；準確率是指在回傳的結果中正確的資料占了多少的比例，而召回率用來計算在所有正確的資料中，回傳了多少正確資料的比例。然而，本研究著重於分群分類的準確率，故未考慮召回率的評估方式。表 2 說明前述兩種評估方式使用之變數及演算公式

表 2 評估準則

	相關資料	不相關資料
回傳資料	A	B
未回傳資料	C	D

$$Precision = \frac{A}{A + B} \quad (7)$$

$$Recall = \frac{A}{A + C} \quad (8)$$

4.3 斷詞程式分群研究

本階段邀請三位資管背景的研究生依據 CKIP 與 CCS 分群結果逐一檢視群內每篇新聞的標題與內文，藉此判定是否討論同一議題，並採多數決求得最佳共識解答。計算方式如公式(9)。

$$Precision(c) = \frac{P_c}{N_c} \quad \textbf{(9)}$$

其中P_c為群集 c 中正確分群之新聞數，N_c 指群集 c 的新聞數。

由於考量相似度門檻值的設定，可能影響分群的結果，故本研究同時檢視在門檻值為 0.1、0.2、0.3、0.4 時，以 CKIP 與 CCS 的斷詞結果，分群精確率的表現差異。其結果如圖 5 所示。

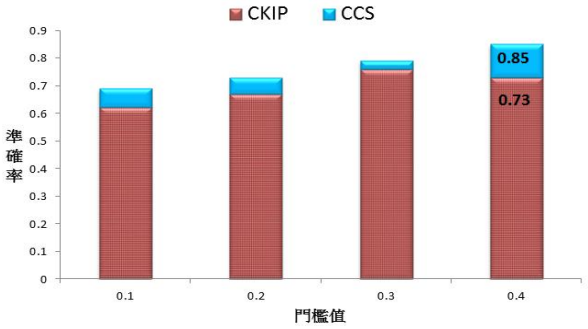

圖 5 CKIP 與 CCS 斷詞系統比較

實驗結果發現，以 CCS 斷詞結果所獲取之關鍵詞進行分群，其結果在所有門檻值設定的表現均優於 CKIP。其中 CCS 在門檻值設定為 0.4 時，分群準確度可高達 85%。CKIP 在門檻值為 0.3 時，準確率雖可達 76%，但仍略遜於 CCS。另外，在效能比較方面，以 CKIP 處理 3475 筆新聞資料斷詞、計算權重及後續分群，共花費 264.28 秒；以 CCS 則只需要花 220.07 秒，相差有 40 秒之多。實驗結果說明 CCS 在分群準確率與效能上皆優於 CKIP，也支持了本研究認為詞庫影響分群之論述。

4.4 分類演算法研究

經由上一階段得知，以 CCS 斷詞後所獲致之分群結果較 CKIP 為佳，因此本階段將以 CCS 之分群結果作為實驗樣本，並將分別透過 SVM、KNN 及 Naive Bayes 三種演算法進行測試。為謹慎起見，對於 KNN 分類演算法設定不同 k 值以觀察其間變化，發現當 k=5 時，準確率達到最高峰，為 86.65%，然而當 k 值再增大則分類準確率反而大幅降低，結果如圖 6 所示。至於三種分類演算法之比較，以 SVM 表現最佳，其準確率高達 91.33%。間接支持多數文獻所提及 SVM 優異成效之說。因此，本研究將擇選 SVM 演算法進行後續新聞事件追蹤及呈現，以期望獲得最高品質的分類成效。如圖 7:

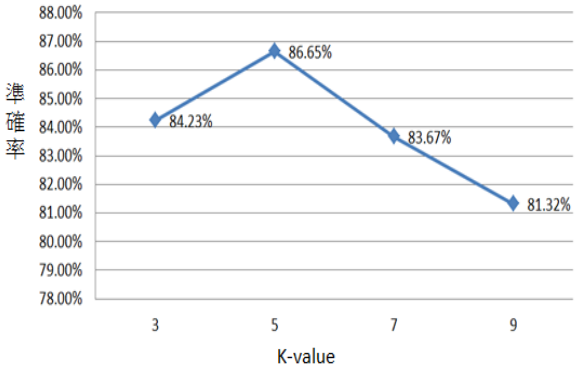

圖 6 KNN 分類結果測試

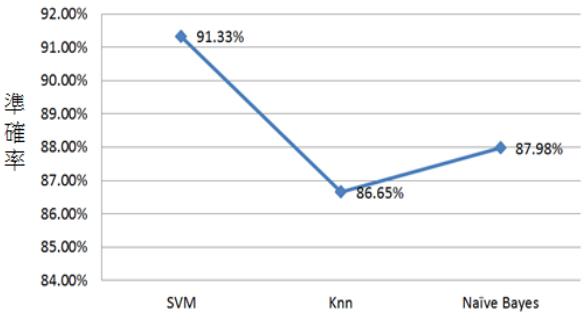

圖 7 三種分類法準確率比較

5.結論與未來展望

5.1 結論探討

本研究以新聞事件偵測與追蹤為主軸提出最佳分群與分類方法。在事件偵測與分群階段考慮到中文斷詞的品質對分群結果影響劇烈，故在中文斷詞處理中比對出最佳斷詞方法。在事件追蹤與分類階段中，比對 SVM、KNN 及 Naive Bayes 三種演算法於分類上的成效。本研究以新聞分群分類整體呈現結果達到高準確率為目標，進行一連串的實驗以及評估。實驗結果發現，透過 CCS 中文斷詞處理，其分群結果顯示有高達 85%的準確率，另外，在事件追蹤與分類階段，經過本研究測試發現，相較於 KNN 與 Naïve Bayes 兩種分類演算法，SVM 的分類準確度最高，達到 91.33%。

5.2 未來展望

本研究已在新聞事件偵測與追蹤過程中找到優異之分群與分類方法，即便如此，皆尚有精益求精之處，未來建議可朝以下幾點作更深入之研究：

(一)特徵值間關聯性辨別:

本研究在事件偵測與分群中未考慮到詞彙間的關聯，尚未解決特徵詞中同形異義與異形同義之問題，導致分群誤判情況。因此，未來也許可加入辨識關鍵詞間關聯之機制，如 LSA 技術，改善分群分類之精確度。

(二)考慮時間特徵:

本研究於實驗過程中，發現到新聞事件具有其連續性及時間限定特質，通常新聞事件都會持續一段時間，如果我們能有效偵測出事件起始與結束的時間區間模式，必然有助於分辨新舊事件之區隔，提升分群準確率。

(三)斷詞處理延伸至其他類別:

研究團隊所發展的 CCS 斷詞系統主要是針對政治類新聞所建置出來的詞庫，如要延伸至其他類別的新聞進行斷詞處理，詞彙量稍嫌不足。因此，若能借助 CKIP 豐富的詞彙量與詞性標註功能，並加入詞性合併等技術，把破碎字詞重新組織，將有效提高斷詞品質，同時未來研究將可延伸至其他類別新聞進行處理。

6. 致謝

本計畫承蒙國科會經費補助與支持，使得研究得以順利進行，謹此致謝，計畫編號：NSC 101-2221-E-224-056。

參考文獻

[1] J. Allan, *et al.*, "Topic Detection and Tracking Pilot Study: Final Report," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, 1998, pp. 194-218.

[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, pp. 513-523, 1988.

[3] N. Ye and X. Li, "A Machine Learning Algorithm Based on Supervised Clustering and Classification," presented at the Proceedings of the 6th International Computer Science Conference on Active Media Technology, 2001.

[4] Z. Hua-Jun, *et al.*, "CBC: clustering

based text classification requiring minimal labeled data," in *Third IEEE International Conference on Data Mining, 2003. ICDM 2003.*, Melbourne, FL, USA, 2003, pp. 443-450.

[5] B. Zhang and S. N. Srihari, "Fast k-Nearest Neighbor Classification Using Cluster-Based Trees," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 26, pp. 525-528, 2004.

[6] A. Kyriakopoulou and T. Kalamboukis, "Text Classification Using Clustering," in *In Proceedings of the ECML-PKDD Discovery Challenge Workshop*, Berlin, Germany, 2006.

[7] J. Allan, *et al.*, "Taking Topic Detection From Evaluation to Practice," presented at the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04, Big Island, Hawaii, USA, 2005.

[8] N. Alex, *et al.*, "Patch clustering for massive data sets," *Neurocomputing*, vol. 72, pp. 1455-1469, 2009.

[9] M. Charikar, *et al.*, "Better streaming algorithms for clustering problems," presented at the Proceedings of the thirty-fifth annual ACM symposium on Theory of computing, San Diego, CA, USA, 2003.

[10] C. Gupta and R. Grossman, "Genic: a single pass generalized incremental algorithm for clustering," in *SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, 2004.

[11] 黃純敏, *et al.*, "SOA 架構之新聞知識萃取與展現-第三年," 行政院國科會 2010.

[12] R. Naseem, *et al.*, "An Improved Similarity Measure for Binary Features in Software Clustering," in *Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on*, Bali, Indonesia, 2010, pp. 111-116.

[13] C. J. V. Rijsbergen, *Information Retrieval*, 2 ed.: Butterworth-Heinemann, 1979.

[14] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 42-49.

[15] Z. Liu, *et al.*, "Study on SVM compared with the other text classification methods," in *Education Technology and Computer Science (ETCS), 2010 Second International Workshop on*, 2010, pp. 219-222.

[16] H. Dong and S. Yin, "A improved feature weighting algorithm for Chinese text classification," in *2010 International Conference on Computer Application and System Modeling (ICCASM)*, Taiyuan Shanxi, China, 2010, pp. V6-433-V6-436.

[17] J. Sheng-Yi, "Efficient Classification Method for Large Dataset," in *2006 International Conference on Machine*

Learning and Cybernetics, 2006, pp. 1190-1194.

[18] S. Jiang, *et al.*, "A clustering-based method for unsupervised intrusion detections," *Pattern Recogn. Lett.*, vol. 27, pp. 802-810, 2006.

[19] F. Sebastiani, "Machine learning in automated text categorization,*ACM Comput. Surv.*, vol. 34, pp. 1-47,2002.

[20] 曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯, *資料探勘（Data Mining）* 台北市：旗標出版股份有限公司, 2005.