Journal of Science and Engineering Technology, Vol. 3, No. 1, pp. 9-18 (2007)

利用共生詞彙特性發展一個二階段文件群集法

李維平　吳澤民　王美淳
中原大學資訊管理研究所
中壢市中北路 200 號

摘　要

群集化（clustering）是在資料探勘領域中被廣泛應用的技術，將其概念應用於文字探勘的領域中，亦是近來的熱門研究議題。若將群集化技術應用於文件型態的資料時，常會採用向量空間模型（vector space model, VSM）來表達文件資料，然而在學術研究上卻發現有兩個缺失：一爲無法辨識文中詞彙間的關聯性，造成文件誤判。在向量空間模型中，每個關鍵詞彙所構成的維度都是獨立的，無法區別文中詞彙間的關聯性（包括一詞多義、一義多詞、以及共同發生詞彙），使得進行文件相似度的比對時可能會造成誤判的情況，降低文件群集之品質。另一缺失則爲如維度太高，易造成群集失準的問題。向量空間模型的維度是由文件集所有的關鍵詞彙之數量而定，當文件所萃取出來的關鍵字過多時，便會使得向量空間模型的維度增加，導致群集的結果也比較不準確。

爲了改善向量空間模型的兩大缺點，本文嘗試提出一個二階段的文件群集法，第一階段先將關鍵字進行群集，第二階段再利用這些關鍵字群集將文件分群；本文透過關聯規則技術的應用，來改善向量空間模型的缺失並增進文件群集的品質，此外，關鍵字群集後的結果還可以幫助文件群集作概括性的描述。本文以 Reuters-21578 文件集進行實驗評估，將本論文所提出的文件群集法與傳統的文件群集法相比較，實驗結果證實本論文所提出的方法確實能得到高品質的文件群集。

關鍵詞：文件群集，關聯規則，文件探勘，共生詞彙

A Two-Stage Document-Clustering Method Utilizing Co-Occurring Words

WEI-PING LEE, TZER-MIN WU and MEI-CHUN WANG
Department of Management Information Systems, Chung Yuan Christian University
200 Chung Pei Rd., Chung-Li, Taiwan

ABSTRACT

Clustering techniques have been developed in many application domains. When clustering text-based documents, the Vector Space Model (VSM) is often used to represent them. However, the VSM model has two major disadvantages in text-clustering research. First, the correlation between terms such as synonymy, polysemy and co-occurring words cannot be distinguished in VSM.

Second, the dimensions will increase if many keywords are retrieved from documents. These disadvantages increase the complexity when calculating similarity between document collections; moreover, the accuracy of the clustering is adversely affected.

We propose a two-stage document-clustering method to ameliorate the disadvantages of the VSM model in document clustering. In the first stage, the keywords are clustered; in the second stage, the documents are clustered from the results obtained in the first stage. The Reuters-21578 corpus was applied to test our proposed method. The results indicate that our method can improve the document-clustering quality better than other traditional clustering methods.

Key Words**:** document clustering, association rule, text mining, co-occurring words

一、緒論

（一）研究背景與動機

群集化（clustering）爲資料探勘領域中被廣泛應用的技術，將其概念應用於文字探勘的領域中，亦是近來的熱門研究議題。群集化技術能將大量的資料依某種特性自動分類，且能呈現這些資料的概觀，以有效率地描述資料 [4]。若將群集化技術應用於文件型態的資料時，常會採用向量空間模型（vector space model, VSM）來表達文件，而每篇文件都是由許多關鍵詞彙所組成，故每篇文件可視爲空間中的一個向量，而其維度則由文件集所有的關鍵詞彙之數量而定。藉由計算文件在空間中的相似度，並配合特定的群集演算法，即可完成文件分群的目的。

雖然應用向量空間模型來表達文件是個好方法，然而在學術研究上卻發現有兩大缺失：

1. 無法辨識文中詞彙間的關聯性，造成文件誤判

在向量空間模型中，每個關鍵詞彙所構成的維度都是獨立的，無法區別文中詞彙間的關聯性，包括「一詞多義」（例如：Apple computer vs. Apple pie）、「一義多詞」（例如：airline schedule vs. airplane schedule）、以及「共同發生詞彙」（例如：在介紹關聯規則的文章中，association 與 mining 兩字常常會一同在文件中出現；因爲詞彙中可能存在這三種關聯性，使得進行文件相似度的比對時，若無法有效區別其差異，將無法有效地將文件分群 [21]。

2. 若維度太高，易造成群集失準的問題

在向量空間模型中，空間的維度由文件集所有的關鍵詞彙之數量而定，當文件所萃取出來的關鍵字過多時，便會使得向量空間模型的維度增加，導致後續進行群集時的相似度運算變得較複雜，且群集的結果也比較不準確 [14, 16, 21]。後續便有學者針對改善向量空間模型爲主軸陸續進行研究 [6, 15]，另外，也有學者進行非向量空間模型的文件群集研究 [14, 23]；經過實驗比較後都證明，向量空間模型的兩項缺失會影響群集的品質。

（二）研究目的

本論文嘗試以關聯規則技術結合傳統的群集方法，提出一個二階段之文件群集法，由關聯規則技術先探勘出關鍵詞彙間的關聯性，再藉由傳統的群集法將關鍵詞彙進行分群，最後以這些關鍵詞彙群集來將文件分群。此方法不同於過去單獨以向量空間爲基礎的文件群集法，期望改善利用向量空間模型爲基礎之群集法的缺點，並完成高品質的文件群集。

（三）研究限制

因語言處理領域的研究成果，對英文語系的處理發展較爲成熟，故本論文僅採用英文文件作爲研究對象，研究結果將對英文文件的探勘較具代表性。且本論文的設計對象及實驗資料皆是針對純文字型態的文件，故多媒體文件（含有圖型、影像等多媒體元素）及網路文件（含有超鏈結的結構）等非純文字型態的資料將可能不適用於本研究，擬於後續研究再行改進。

二、文獻探討

（一）群集化

群集化是把有形或抽象的物件歸類到相似集合的過程，使得群組內的物件相似度高，群組間的相似度低。一般常見的群集演算法可分爲階層式（hierarchical）[22] 和分割式（partitioning）[8] 二種。

1. 階層式群集演算法

階層式群集演算法又可分爲「階層式聚合群集法」（hierarchical agglomerative clustering, HAC）與「階層式分裂群集法」（hierarchical divisive clustering, HDC）兩類，其主要結構都可以樹狀結構來表示，若是採用聚合演算法，則資料是由樹狀結構的底部向上方聚合；若採用分裂演算法，

則是由樹狀結構的頂端向下方層層分裂。如以階層式聚合演算法爲例，整個聚合的過程如圖 1 所示。

一開始每筆資料皆各自爲一個群集，共有 10 個群集。由彼此距離最近的二個群集開始合併；最初合併的群集是編號 7 和 10 的資料，隨後是編號 6 和 9 的資料，依此類推，到最後整個過程收斂之後，所有的資料就完全聚合爲一個群集。圖中 X 軸表示各筆資料的編號，Y 軸表示兩個群集合併時的距離。群集間彼此的距離有如下四種計算方式：

（1）單一連結法（single-linkage）：計算二群集間距離最近的二點。

（2）完整連結法（complete-linkage）：計算二群集間距離最遠的二點。

（3）平均連結法（average-linkage）：計算二群集中各點與點之間距離總和的平均。

（4）沃德法（ward's method）：計算二集群中各維度的變異數之平方和。

2. 分割式群集演算法

分割式群集演算法需要先找出 k 個隨機的中心點，再根據各文件與這些中心點的距離來決定哪些文件可合併爲同一群，隨後再於新的群集中找到新的中心點，並利用新的中心點重新計算其與各文件間的距離，反覆操作這些步驟，直到群集不再變動爲止。如：k-means [8] 即屬於此類。

（二）文件群集化（Document Clustering）

文件群集化的技術可以將大量的文件依其性質區分爲許多群集，使性質相似的文件被歸在同一群中，讓使用者能夠快速地區分文件類別，找到所需的文件，同時也能對所有的文件分佈提供一個概觀、提升資訊檢索系統的搜尋效果、有效地組織及呈現資訊、及自動建立文件的分類架構（如

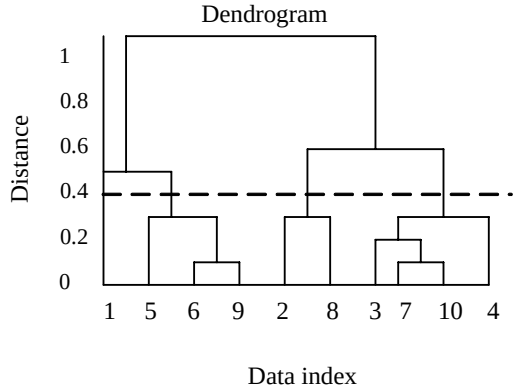

圖 1. 階層式群集化樹狀結構表示圖

yahoo 的分類目錄）[20]。

對大多數的文件群集法而言，首要步驟即是建立「向量空間模型」；在此模型中，每篇文件是由空間中的向量所組成，而空間向量的維度則由文件集所有的關鍵詞彙之數量而定。在一個有 t 維的向量空間模型中，每篇文件 D_i 可以簡單地表示成 $D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$ 的形式，其中 d_{ij} 表示第 j 個關鍵詞彙的權重 [17]。如此，任二篇文件的相似度可以很容易地利用相似度函數（如：cosine 函數）求得，並建立「文件相似度矩陣」，最後再以特定的文件群集演算法進行文件分群。

儘管應用向量空間模型表達文件是個好方法，然而在學術研究上卻發現其有兩大缺失：無法辨識文中詞彙間的關聯性，造成文件誤判；若維度太高，易造成群集失準的問題。

爲了改善向量空間模型這兩大缺失，陸續有學者針對此動機進行研究，在 1990 年 Deerwester 等學者便提出了著名的潛在語意縮減法（latent semantic indexing, LSI），主要的目的便是要解決維度過高以及詞彙關聯性之問題，而其運作主要是利用統計上的最小平方法 [21]；學者 Dhillon 在 2001 年的研究中更提出利用向量空間模型之特徵維度縮減技術－LSI，並配合 spherical k-means 群集演算法能夠大幅提升群集化的效率 [6]。

除了 LSI 的方法能夠有效縮減維度之外，2000 年 Rüger 提出新的向量空間模型之特徵維度縮減方法，主要利用一個改良的 TFIDF 字詞權重計算法，找出 ranked words 並選出排名高於某門檻值的詞彙作爲特徵維度，在對群集結果之準確率影響不大的情況下，有效縮減了特徵維度且提高分群的效率 [16]。而有些研究則只針對改善向量空間模型中，無法辨識詞彙關聯性之缺失來作爲研究動機，在 2002 年作者鍾明璇提出利用關聯規則探勘技術，改善以關鍵詞彙所建立之空間向量來進行文件群集的缺失，包括一義多詞及無法區別詞彙相關性等影響文件群集品質之重要因素，藉以提升文件群集的品質 [1]。其方法主要是利用關鍵詞彙關聯規則間之信賴度及支持度值，來重新調整原先僅以關鍵詞彙出現於文件中之次數所建立的詞彙－文件矩陣，可將文件集中之關鍵詞彙，依其關聯強度調整詞彙－文件矩陣中的權重值，使得相似的文件可獲得較高的權重值而更容易形成同一群集，經實驗證明，此法可以有效提升文件群集品質。

另外有其他學者利用非向量空間模型的方法來進行文件群集。1998 年，學者 Zamir 提出一新的群集演算法－STC

（suffix tree clustering），利用文件詞彙及片語建立一個名爲 Suffix Tree 的資料結構，再以特殊的演算法進行分群。其特點是運作效率高、同一篇文件能被重覆定義於許多群集中，且能應用於網頁的文件型態 [23]。

（三）關聯規則（Association Rule）

關聯規則，就是某些項目可能會引發其他項目出現的規則，表示爲 "$X \Rightarrow Y$"。探勘關聯規則主要是研究資料之間的關聯性，它的目的是要從銷售的交易資料庫中，發現項目（item）間的關聯。設有一個交易資料庫 D，其中的每筆交易 $T \in D$ 皆爲交易項目的集合，則 $X \Rightarrow Y$ 表示只要某筆交易 T 中出現項目 X，則也有可能會出現項目 Y [2]。例如：有 81% 的顧客在購買香煙時，同時亦會購買報紙，表示爲「*香煙⇒報紙（81%）*」；則此例中的「香煙」及「報紙」即爲彼此有特殊關聯的項目，因此藉由關聯規則技術所找出的有用規則，將對商品的行銷大有助益。

1. 關聯規則評估指標

關聯規則的評估指標，一般以最小支持度（*support*）和最小信賴度（*confidence*）爲主。最小支持度界定一個規則所必須涵蓋的最少資料數目；最小信賴度則代表這個規則的預測強度，當探勘出的規則滿足使用者訂定的最小支持度和信賴度的門檻時，這個規則才算成立。其數學表示如下：

$$support(A \Rightarrow B) = P(A \cap B) \qquad (1)$$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \Rightarrow B)}{support(A)} \qquad (2)$$

除此之外，尚有許多學者提出其它具有不同特性和用途的評估指標，如 strength、lift、interest、conviction、chi-square、Entropy、Laplace…等 [3]，但其大多仍是以 support 和 confidence 爲基礎衍生而來。

2. 關聯規則種類

關聯規則有許多種類，大體上可以將它分成以下三類 [11]：

（1）以屬性值的型態爲基礎：如果我們所關注的只是項目是否出現，則稱爲布林值的關聯規則（boolean association rule），例如「牛奶⇒麵包」即屬於這類關聯規則。如果我們也一併關注項目的購買單位數，便稱爲有重複項目的關聯規則（association rule with repeated items），例如「2 單位牛奶⇒3 單位麵包」。如果我們所要描述的規則其項目或屬性是一個數值，這種就稱爲數量關聯規則（quantitative association rule）。

（2）以規則中所涵蓋的資料維度爲基礎：如果在關聯規則中的項目或屬性僅參照單一的維度時，我們稱之爲單一維度關聯規則（single dimensional association rule），例如我們將「牛奶⇒麵包」的關聯規則寫成「購買（X, "牛奶"）⇒購買（X, "麵包"）」，則其著眼的是「購買」這個維度。反之，如果關聯規則中的項目或屬性參照兩個以上維度時，便稱爲複合維度關聯規則（multidimensional association rule）。

（3）以規則中所涵蓋的抽象層級爲基礎：如果在關聯規則中的項目或屬性可以屬於不同的概念層級，例如「年齡（X, "中年"）⇒購買（X, "味全果汁牛奶"）」（"中年" 對於年齡而言屬於較高層級概念，但 "味全果汁牛奶" 對於購買項目而言屬於較低層級概念），則稱這類規則爲跨層級關聯規則（multilevel association rule）。反之，如果沒有參照到不同層級的項目或屬性規則，則稱爲單一層級關聯規則（single-level association rule）。

3. 關聯規則演算法

關聯規則的演算法主要可以分爲兩大類：（1）利用 Apriori-like 的方法產生 candidate set，並找出符合最小支持度的大項目集合（large itemsets），再依據大項目集合產生關聯規則；（2）使用 Non Apriori-like 的方法，找出大項目集合。

第一類的方法是以 Apriori 演算法 [2] 爲基礎，爲關聯規則探勘技術中，最早被提出且運作穩健的演算法。它們共同的特點是第一次的 candidate set（以 C_1 表示）是直接掃描資料庫一次而得到，而其他的 C_k（k>1）產生方式都包含了兩個主要步驟：第一個是合併產生 candidate set，第二個則是將這些項目集合中，含有不是前一次作業的大項目集合者去除，然後針對這些留下來的 candidate set，以掃描資料庫的方式獲取其支持度，再將未滿足最小支持度要求的項目集合去除掉，即得到所謂的大項目集合。

4. 探勘文件的關聯規則

應用關聯規則於文字探勘中，可找出關鍵詞彙之間的關聯性。依循資料庫的關聯規則探勘精神，Singh [18] 將非結構化或半結構化的文章對應到結構化的資料表格中，將一篇文章對應爲一筆交易資料，使用稀疏矩陣（sparse matrix）

避免過多結構化資料的比對運算，並運用概念相關性（concept-relatives）從稀疏矩陣得到大項目集合，這些大項目集合即爲符合使用者需求並具高度相關性的知識。

Feldman [9]則提出了一種改良的文件關聯規則探勘技術—最大關聯規則（maximal association rules），以此來計算關鍵詞彙同時發生的頻率，藉以探勘文件中的重要資訊。Singh [19] 以擴充的概念階層（extended concept hierarchy, ECH）建構背景知識；擴充了概念與概念之間的兄弟關聯關係，可探勘出四種規則：（1）一般規則（general rules）、（2）父法則（parent rules）、（3）子規則（child rules）、（4）兄弟規則（sibling rules）。

三、研究架構

（一）二階段文件群集法

本文以非向量空間模型的方法爲主，採用兩階段分群法：第一階段先利用關聯規則及分群技術將關鍵字分群，主要目的是希望在此階段先辨識出字彙間的關聯性，同時將會影響分群品質的干擾資訊，亦即沒有任何關聯性的字彙予以刪除，接著將關鍵字彙分群，第二階段再利用關鍵字彙的群集，透過比對函數將文件分群。圖 2 爲本論文之方法的概念圖。

（二）演算法

以下簡述本論文所採用的演算法及其時間複雜度。

1. 關聯規則探勘（apriori）

L_1 = {Large 1 – itemsets};

for (k=2; $L_{k-1}=\phi$; k++) **do begin**

C_k = apriori - gen(L_{k-1}); //New candidates

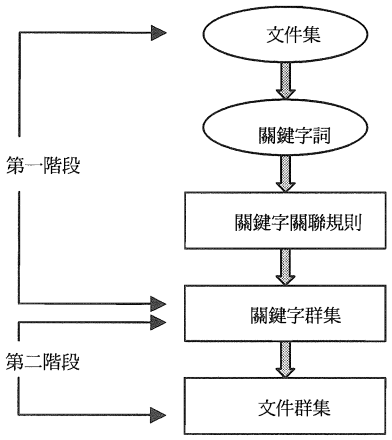

圖 2. 二階段文件群集法運作概念圖

forall transactions t ∈ D **do begin**

C_t = subset(C_k, t); //Candidates contained in t

forall candidates c ∈ C_t **do**

c.count ++;

end

L_k={c ∈ C_k | c.count ≥ minsup}

end

Answer=$U_k L_k$;

其時間複雜度爲 $O(N^L)$，其中 N 爲文件集包含的文件數量，L 關鍵字數量。

2. 關鍵字分群（hierarchical agglomerative clustering, HAC）

Given: a set D={d_1, d_2, …, d_m}of documents

a function **sim**: S(D)* S(D)→R

for i:=1 to m **do**

c_i :=d_i **end**

C={c_1, c_2, …, c_m }

j :=m+1

while |C|>1

$(c_{m1},c_{m2}) := \arg\max_{((c_u,c_v)\in C\times C)} sim(c_u,c_v)$

$c_j=c_{m1} \cup c_{m2}$

$C:=C \setminus \{c_{m1},c_{m2}\} \cup c_j$

j:=j+1

其時間複雜度爲 $O(N^2)$，其中 N 爲文件集包含的文件數量。下面將詳細地說明本研究運作方式。

（三）群集步驟

1. 關鍵字群集

（1）資料前處理：將每篇文件中所需用到的資料（如：文件編號、類別、標題及內文）解析出來，並且建立停用字詞列表（stop-words list），也就是事先定義不具文件代表性的字詞，使得下一步資訊萃取時能避免擷取這些字詞。

（2）關鍵資訊萃取：辨別文件中的每一個詞彙；去除停用字詞：將停用字詞從文件中去除，留下有意義的字詞；還原字根（stemming）：英文文件中常有許多名詞、動詞，會以不同的型態（如：單數型、複數型、現在式、過去式…等）出現，甚至還有形容詞及副詞的詞類變化，如：beauty、beauties、beautiful、beautifully…等，均屬於同一字根的不同變化，且都有相同的概念，但若未加處理，則在向量空間模型中，其將分別被視爲不同的字，結果使得空間維度爆增，以及後續運算失準。所以將這些具有相同概念但不同型態變化的字還

原成最原始的字根，有助於減少文字處理的數量及後續運算的精準；計算詞彙顯著值（權重）：本研究採用 *TF* 值作爲詞彙顯著值（權重）的計算方法。

經過上述步驟，將會萃取出文件集內的詞彙集及每個詞彙的 *TF* 值。之後便可藉由選擇 *TF* 值的大小，決定每篇文件中將以那些詞彙作爲該文件的關鍵詞彙。

（3）關聯規則探勘：從文件集中萃取出所有關鍵詞彙，將之轉換爲結構化的資料格式，存入關聯式資料庫中以進行關聯規則探勘。先將每篇文件對應爲在資料庫中的一筆交易（transaction）；而代表每篇文件的關鍵詞彙則對應爲該筆交易的資料項（item）。再利用 Apriori [2] 演算法找出長度爲二（如："A ⇒ C"）的關聯規則，藉此分析關鍵詞彙兩兩間的關聯性。並藉由調整最小支持度（minimal support）及最小信賴度（minimal confidence）的門檻值，以找出有意義的規則。

（4）建立關鍵字相似度矩陣：我們將字彙之關聯規則的平均信賴度（confidence）[14] 視爲字彙之間的相似度，例如：*confidence*（A ⇒ B）=0.66，*confidence*（B ⇒ A）=0.57，則 A、B 兩個字彙的相似度便是取其平均信賴度 0.615。最後，建立出來的關鍵字相似度矩陣便如表 1 所示。

（5）分群演算法：在許多研究中指出，利用階層式聚合群集法 HAC 產生的群集品質優於分割式群集法（如 k-means），但運算速度也相對較慢 [5, 7, 13]。本論文因著重於提升群集化的品質，而非群集化的速度，故選用階層式聚合群集法來進行文件群集。

在階層式聚合群集法的運算過程中，當二個群集合併爲一群後，該群集與其它群集間彼此的距離需重新計算，其中沃德法與完整連結法 [10, 23]，在群集化的過程中會產生較爲平衡的二元樹，使得被歸於同一群的文件能有較高的相關性；因此，本論文採用完整連結法來計算群集間的距離。

經過群集演算法將關鍵字群集完成，分析者可以從群集的結果中發覺被群集爲同一群之關鍵字之間所存在較強的關聯性，因此更可以幫助文件群集後對文件進行概括性描述；另外，由於群集數是由使用者自行訂定的，因此，使用者也可以修正所要分群的群集數以得到較好的群集結果。

表 1. 關鍵字相似度矩陣

	A	B	C	D
A	1.000	0.615	0.660	0
B	0.615	1.000	0.570	0.280
C	0.660	0.570	1.000	0
D	0	0.280	0	1.000

2. 文件群集

利用關鍵字群集結果將文件分群：藉由關鍵字的群集結果，可以進一步找出文件的群集，只要找出每篇文件該對應至哪一個關鍵字群集中，即可完成群集。

在此步驟主要是藉由比對函數（matching score）的設計，計算每筆文件在每個關鍵字的群集（C_j）中所佔的之分數；若某篇文件在某個關鍵字的群集中所得的分數最高者，則將該篇文件歸類至該群中；重覆計算所有的文件，直到所有的文件被歸類完畢，則完成文件的群集。

比對函數則是參考準確率（precision）及求全率（recall）兩參數所設計的調和參數；而準確率所代表的意義是某關鍵字群集的準確度，求全率所代表的是某關鍵字群集中，對某篇文件的關鍵字涵蓋程度。比對函數的計算公式如下所示，其中，P 代表關鍵字準確率；R 代表求全率；D_i 代表某篇文件的關鍵字；C_j 代表某關鍵字群集中的關鍵字。

$$\text{Score} = \frac{2PR}{P+R} \qquad (3)$$

$$\text{Where P} = \frac{D_i \cap C_j}{C_j} \text{ and R} = \frac{D_i \cap C_j}{D_i}$$

四、實驗結果

（一）測試文件集

本論文針對英文文件進行群集化，實驗中所採用的文件集是 Reuters-21578 Distribution 1.0，其包含了路透社（Reuters Newswire）自 1987 年 2 月 26 日到 1987 年 10 月 9 日所收集的新聞，共有 21,578 篇文件。這些文件大多已經過人工定義其類別，並將歸類的結果註明在每篇文件中的 "TOPICS" 屬性內。這些人工分類的結果常在文件群集的實驗中作爲比較群集良窳的重要依據。

（二）評估準則

本論文採用的衡量指標是在資訊檢索領域中用來評估效能的標準衡量指標－Precision、Recall 及 F-measure，來評估群集結果的準確性 [15]。

Precision 代表某個群集的準確度；Recall 代表某個群集

中，對某種類別之文件的涵蓋程度 [12, 15]；而 F-measure 是調和平均數，組合了以上二種衡量指標，用以衡量群集的效果 [13]。當 Precision 與 Recall 的值皆很高時，F-measure 的值才會很高，此亦表示群集的效果良好，即群集內的資料相似度高，群集間的資料相似度低。

圖 3 顯示了在一個文件集內，Precision 與 Recall 的關係圖，其中，N_t：被人工定爲類別 t 的文件數；N_c：群集 c 中的文件數；$N_t \cap N_c$：在群集 c 中包含類別 t 的文件數。

對於任意一個文件類別 t 和群集 c 而言，Precision（P）、Recall（R）、F-measure（F）的計算方式如下：

$$P_{c,t} = \frac{N_t \cap N_c}{N_c} \quad (4)$$

$$R_{c,t} = \frac{N_t \cap N_c}{N_t} \quad (5)$$

$$F_{c,t} = \frac{2P_{c,t}R_{c,t}}{P_{c,t} + R_{c,t}} \quad (6)$$

$$F_t = Maximal(F_{c,t}) \quad (7)$$

在求出每個文件類別在每個群集中的 F-measure 值（$F_{c,t}$）之後，取每個類別在各群集中的最大 $F_{c,t}$ 值，作爲該類別的 F-measure 值；再計算 $F_{overall}$ 以求出所有類別的平均 F-measure 值，此即代表整體群集的效果；其中 T 爲文件集內所有文件類別的集合，$|t|$ 爲文件集內某一類別 t 的文件數量：

$$F_{overall} = \frac{\sum_{t \in T} |t| \cdot F_t}{\sum_{t \in T} |t|} \quad (8)$$

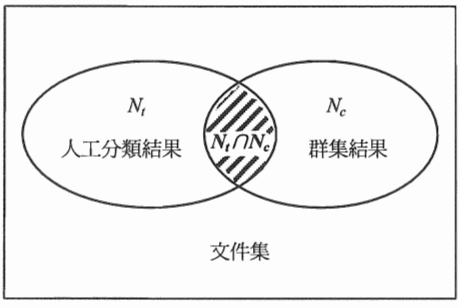

圖 3. Precision 與 Recall 的關係圖

（三）實驗設計

本論文中將調整四個可能影響群集品質的參數—文件數量、關鍵詞彙數量、關聯規則的支持度、關聯規則的信賴度，分別設計不同的實驗，並比較本論文所提出的群集法，其品質是否優於傳統群集方式。同時，本論文也將和鍾明璇 [1] 之研究進行比較（後文將簡稱爲 JMS）；在許多研究中指出，利用階層式聚合群集法 HAC 產生的群集品質優於分割式群集法（如 k-means），但運算速度也相對較慢 [13]。本論文因著重於提升群集化的品質，而非群集化的速度，故在傳統的文件群集法上將選用階層式聚合群集法來進行文件群集品質方面的比較。而在實驗參數的設定上，我們將所需擷取的關鍵字群集數設定爲實驗文件集所包含之類別數，以便評估文件分群後的群集品質。

1. 實驗一：文件數量對不同群集法的影響

文件的數量增加，文件內容的複雜度亦隨之增加，則群集的效果將有可能受到影響。此實驗採用三個不同文件數量的測試文件集，測試在不同文件數量的情況下，本論文所提出的群集法是否仍能提升群集效果。

（1）測試資料：首先在 Reuters-21578 文件集內選擇五個涵蓋文件數量最多的類別－"EARN、ACQ、MONEY-FX、CRUDE、GRAIN"，再於各類別中隨機選取出數十篇文件，並去除同時被歸屬於二個類別以上的文件，最後產生三個測試文件集：TD_a、TD_b 及 TD_c。每個文件集內含有四個以上的文件類別，平均每篇文件的關鍵詞彙數量控制在約十個左右。表 2 爲此三個測試文件集的描述。

表 2. 實驗一之測試文件集的描述

測試文件集	總文件數	TF	關鍵詞彙數	類別(文件數)
TD_a	50	2	534	EARN(24) ACQ(17) CRUDE(5) MONEY-FX(4)
TD_b	104	2	887	EARN(47) ACQ(47) MONEY-FX(5) CRUDE(4) GRAIN(1)
TD_c	325	2	2610	EARN(151) ACQ(132) MONEY-FX(17) CRUDE(24) GRAIN(1)

（2）實驗結果：本實驗我們將關聯規則之最小支持度定爲 5%，而最小信賴度定爲 25%。由表 3 中的實驗結果可知：本論文所提出的群集法，在文件數量較少時（如：測試文件集 TD_a），都比文件數量較多時的群集結果爲佳，提升度皆達 40% 以上，其中在文件數量最多的情況下（如：測試文件集 TD_c），提升度更是高達 85.27%。

過去的群集法在文件數量逐漸增多時其 F-value 會逐漸下降，推斷其原因爲當文件數量增加時，文件集整體的複雜度亦隨之增加，有可能包含更多特徵不明顯而難以歸類的文件，使得群集的整體效果降低。然而，使用本論文之方法進行群集可以發現其 F-value 是隨著文件數量的增加而漸漸提高，推斷其原因爲本論文之群集方法中的關聯規則探勘步驟不但可以發掘詞彙間的關聯性，更可以有效過濾干擾資訊，減少關鍵字彙，以有效幫助群集結果的品質。

2. 實驗二：關鍵詞彙數量對不同群集法的影響

一般而言，用以表達文件特徵的關鍵詞彙愈多，則愈能夠描述文件的特色；但相對的，關鍵詞彙愈多，也可能含有愈多的雜訊，反而影響群集的效果。此實驗採用三個不同關鍵詞彙數量的測試文件集，測試在不同關鍵詞彙數量的情況下，本論文所提出的群集法是否仍能提升群集效果。

（1）測試資料：首先在 Reuters-21578 文件集內隨機選取出五十篇文件，並去除同時被歸屬於二個類別以上的文件，之後調整文件的 TF 值，藉以控制文件中關鍵詞彙的數量，最後產生三個含有同樣文件但關鍵詞彙數量不同的測試文件集：TT_a、TT_b 及 TT_c。表 4 爲此三個測試文件集的描述。

（2）實驗結果：本實驗我們將關聯規則之最小支持度定爲 5%，而最小信賴度定爲 25%。由表 5 中的實驗結果可知：當探勘出的關鍵詞彙數量越少時，本論文之方法越能發揮其品質上的效果，尤其在關鍵詞彙最少的資料集（如：測試文件集 TT_a）其 F-value 更是提升了兩倍以上。推斷結果可能原因是當關鍵詞彙越多雖然越能代表文件的特徵，但是卻也可能納入了干擾資訊，造成群集上的資訊模糊，導致群集失準，而本論文之方法，除了可利用 TF 值在限定關鍵詞彙數量外，也利用關聯規則過濾了干擾資訊，因此，所得到的群集都有較好的品質。

3. 實驗三：關聯規則支持度（support）對不同群集法的影響

關聯規則支持度的門檻值愈高，則產生的關聯規則愈少；反之則產生的關聯規則愈多。本實驗在測試藉由調整關聯規則的最小支持度所產生出不同數量的關聯規則，會對本論文提出的群集法之效果造成什麼樣的影響。

（1）測試資料：首先在 Reuters-21578 文件集內隨機選取出五十篇文件，並去除同時被歸屬於二個類別以上的文件，最後產生三個含有同樣屬性之文件的測試文件集：TS_a、TS_b 及 TS_c。而在關聯規則探勘的過程中，將分別對不同的測試文件集調整不同的關聯規則最小支持度。表 6 爲此三個測試文件集的描述。

（2）實驗結果：本實驗我們將關聯規則之最小信賴度固定爲 25%。由表 7 中的實驗結果中得知：本論文之方法將不會因爲最小支持度的門檻值高低而影響 F-value，換句話說，最小支持度的門檻值設定所影響

表 3. 實驗一之結果

測試文件集	$F_{overall}$ (傳統群集法)	$F_{overall}$ (JMS)	$F_{overall}$ (本論文)	提升百分比
TD_a	0.354	0.392	0.497	40.40%
TD_b	0.323	0.348	0.539	66.87%
TD_c	0.292	0.310	0.541	85.27%

表 4. 實驗二之測試文件集的描述

測試文件集	總文件數	*TF* 值	關鍵詞彙數
TT_a	50	3	348
TT_b	50	2	534
TT_c	50	1	889

表 5. 實驗二之結果

測試文件集	$F_{overall}$ (傳統群集法)	$F_{overall}$ (JMS)	$F_{overall}$ (本論文)	提升百分比
TT_a	0.307	0.322	0.686	123.45%
TT_b	0.354	0.392	0.497	40.40%
TT_c	0.349	0.381	0.492	40.97%

表 6. 實驗三之測試文件集的描述

測試文件集	總文件數	TF	最小支持度	關聯規則數
TS_a	50	2	5%	705
TS_b	50	2	7%	281
TS_c	50	2	9%	120

表 7. 實驗三之結果

測試文件集	$F_{overall}$ (傳統群集法)	$F_{overall}$ (JMS)	$F_{overall}$ (本論文)	提升百分比
TS_a	0.349	0.364	0.514	47.28%
TS_b	0.349	0.381	0.492	40.97%
TS_c	0.349	0.371	0.513	47.28%

的是關聯規則數目，並不代表會對關鍵字數目有大幅影響，因此，無論最小支持度的門檻值設定高低，將不會對本文之群集方法品質有絕對的影響。

4. 實驗四：關聯規則信賴度對群集結果的影響

本實驗在測試藉由調整關聯規則的最小信賴度所產生出不同數量的關聯規則，會對本論文提出的群集法之效果造成什麼樣的影響。

（1）測試資料：首先在 Reuters-21578 文件集內隨機選取出五十篇文件，並去除同時被歸屬於二個類別以上的文件，最後產生三個含有同樣屬性之文件的測試文件集：TS_a、TS_b 及 TS_c。而在關聯規則探勘的過程中，將分別對不同的測試文件集調整不同的關聯規則最小信賴度。表 8 爲此三個測試文件集的描述。

（2）實驗結果：本實驗我們將關聯規則之最小支持度固定爲 5%。由表 9 中的實驗結果中可發現：當最小信賴度定的越高，所萃取的關鍵規則數量越少時，利用本論文之方法其結果會有較高的 F-value，相對於傳統方法更有 48% 的提升度；推斷其原因可能是因爲，當最小信賴度定的越高也代表其規則的可信度越高，更能突顯本論文方法的貢獻。

表 8. 實驗四之測試文件集的描述

測試文件集	總文件數	TF	最小信賴度	關聯規則數
TC_a	50	2	25%	281
TC_b	50	2	50%	226
TC_c	50	2	75%	166

表 9. 實驗四之結果

測試文件集	$F_{overall}$ (傳統群集法)	$F_{overall}$ (JMS)	$F_{overall}$ (本論文)	提升百分比
TC_a	0.349	0.381	0.492	40.97%
TC_b	0.349	0.388	0.492	40.97%
TC_c	0.349	0.383	0.517	48.14%

（四）實驗討論

經由上述四個實驗的結果發現，每次群集化後所求得的 F 值皆介於 0.5 至 0.7 之間，同時也可以發現文件中所萃取出關鍵字對於本論文方法的群集結果有極重要之影響，而在最重要的關鍵資訊萃取步驟上，本身即是一門專業的學域，針對不同語言、不同文件特性，都對應了不同的萃取技術，因此，若是能運用適當的關鍵字萃取技術，則對於群集品質將能更有效地提升。

五、結論與建議

（一）結論

一般而言，文件群集化的二大步驟爲：（1）萃取文件特徵，並將文件轉換成向量空間模型表示之。（2）利用特定的群集演算法進行文件群集。然而，向量空間模型本身有其先天的缺失，一爲無法區別文中詞彙間的關聯性，另一爲萃取的關鍵字過多導致向量空間中維度過高，此皆可能導致後續群集運算失準。而本論文利用關聯規則探勘技術找出文中關鍵詞彙彼此的關聯性，並過濾干擾資訊，同時本論文之方法也以關聯規則之平均信賴度取代相似度函數，大大降低運算複雜度與成本。實驗結果發現，無論在不同的文件數量或關鍵詞彙數量之條件下，本論文所提出的群集方法皆能有效提升群集的品質。

（二）後續研究方向

爲更有效地提高文件群集的品質，後續的研究可以朝更進一步的關聯規則探勘來加強：

1. 加入數量關聯規則（quantitative association rule）

本文在關聯規則探勘上並沒有考慮關鍵字出現的次數（TF 值），未來的研究可以透過數量關聯規則技術，加入關鍵字出現次數之要素，期望利用探勘出來的結果來調整相似度的計算以及比對函數的權重值。

2. 找出共同發生的片語（co-occurring text phrase）

可利用序列型樣探勘（mining sequential patterns）之演算法來找出文字間的片語（phrase）以及找出片語之間的關聯性；在這裡所謂的片語是指關鍵字的出現有順序性，而非一般文法中認定的片語。希望藉由片語間關聯的發掘來提升文件群集的品質。

參考文獻

1. 鍾明璇（民 91），應用關聯規則技術有效輔助以向量

空間模型爲基礎之文件群集法，中原大學資訊管理學系碩士論文。

2. Agrawal, R. and R. Srikant (1994) Fast algorithms for mining association rules. Proceedings of the 20th international Conference on Very Large Databases, Santiago, Chile.
3. Bayardo, R. J. Jr. and R. Agrawal (1999) Mining the most interesting rules. Conference on Knowledge Discovery in Data Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California.
4. Chen, M. S., J. Han and P. S. Yu (1996) Data mining: An overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883.
5. Cutting, D. R., D. R. Karger, J. O. Pedersen and J. W. Tukey (1992) Scatter/Gather: A cluster-based approach to browsing large document collections. 15th International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark.
6. Dhillon, I. S. and D. S. Modha (2001) Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1), 143-175.
7. Dubes, R. C. and A. K. Jain (1988) *Algorithms for Clustering Data*, Prentice Hall, Upper Saddle River, NJ.
8. Faber, V. (1994) Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22, 138-144.
9. Feldman, R., W. Klosgen, B. Y. Yaniv, G. Kedar and V. Reznikov (1997) Pattern based browsing in document collections. Proceedings of First European Symposium on Principles of Data Mining and Knowledge Discovery, London, UK.
10. Griffith, A., H. C. Luckhurst and P. Willet (1986) Using inter-document similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37, 3-11.
11. Han, J. and M. Kamber (2000) *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA.
12. Kowalski, G. (1997) *Information Retrieval Systems – Theory and Implementation*, Kluwer Academic Publishers, Norwell, MA.
13. Larsen, B. and C. Aone (1999) Fast and effective text mining using linear-time document clustering. Proceedings of the fifth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, San Diego, California.
14. Moore, J., E. H. Han, D. Boley, M. Gini, R. Gros, K. Hasting, G. Karypis, V. Kumar and B. Mobasher (1997) Web page categorization and feature selection using association rule and principal component clustering. 7th Workshop on Information Technologies and Systems (WITS'97), Atlanta, Georgia.
15. Rijbergen, C. J. Van (1979) *Information Retrieval*, 2nd Ed, 114-115. Butterworths, London, UK.
16. Rüger, S. M. and S. E. Gauch (2000) *Feature Reduction for Document Clustering and Classification: Technical Report DTR 2000/8*, Computing Department of Imperial College, London, UK.
17. Salton, G. and M. McGill (1983) *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY.
18. Singh, L., B. Chen, R. Haight and P. Scheuermann (1999) An algorithm for constrained association rule mining in semi-structured data. Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining, London, UK.
19. Singh, L., P. Scheuermann and B. Chen (1997) Generating association rules from semi-structured documents using an extended concept hierarchy. Proceedings of the sixth international conference on Information and knowledge management, Las Vegas, Nevada.
20. Steinbach, M., G. Karypis and V. Kumar (2000) *A Comparison of Document Clustering Techniques: Technical Report #00-034 (2000)*, University of Minnesota, Minneapolis, Minnesota.
21. Sullivan, D. (2001) *Document Warehousing and Text Mining*, 326. Wiley Computer Publishing, New York, NY.
22. Willet, P. (1988) Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(5), 557-597.
23. Zamir, O. and O. Etzioni (1998) Web document clustering: A feasibility demonstration. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia.

收件：94.11.29　修正：95.03.29　接受：95.06.08