改良式階層聚合演算法之研究

黃純敏
雲林科技大學 資訊管理學系
副教授
huangcm@yuntech.edu.tw

李宜樺
雲林科技大學 資訊管理學系
碩士生
g9323713@yuntech.edu.tw

陳泠謹
雲林科技大學 資訊管理學系
碩士生
g9423727@yuntech.edu.tw

摘要

以往提出之分群演算法中以階層式聚合演算法之分群效果為佳，但當其面對龐大資料量時，因其執行效能不佳而成為大家漸漸不使用此一方法之主因。本研究有鑑於此提出改良式階層聚合演算法，運用 K-way 之概念，降低樹狀結構之高度及其時間複雜度，並結合潛在語意分析提升分群之精確度。

本研究使用 Google 搜尋引擎對熱門話題之搜尋結果做為實驗資料集，實驗結果發現改良式階層聚合演算法之精確度高於 0.99、熵低於 0.003 以下。改良式階層聚合演算法在合併群集時，可有效的運用潛在語意分析之特質，將相同概念之網頁合併，進而提升分群之精確度，並具有群間密度低、群內密度高之特質，此外執行速度並不因資料量之大小而影響其效率；整體而言，本研究之改良式階層聚合演算法優於傳統階層聚合演算法。

關鍵詞：分群演算法、階層聚合演算法、潛在語意分析、資訊檢索

1. 前言

現今人們仰賴搜尋引擎為其獲取全球化之資訊，但目前搜尋引擎主要採用關鍵字比對內容進行檢索，此檢索結果往往傳回數以萬計的網站，在此龐大的資訊量下，使用者仍需費很多時間一一過濾資訊，以獲取所需，故就使用者之角度而言，此種搜尋結果呈現方式顯然不夠便利。

資料探勘領域提出以分類或是分群之方式將資料加以整理，以方便使用者閱讀並降低過濾資訊之時間。目前運用至搜尋引擎之作法，僅採用預先定義之類目(如 taxonomy 等)，將網站進行適當的分類，如 Yahoo 等搜尋引擎。但此類別之設定無法隨時間而成長或改變，與現實無法相互配合；再者實際運用網頁分類至搜尋引擎之網站少之又少，即便提供此功能，其搜尋結果均偏向商業化之資訊，因而對使用者而言，這樣的資訊範圍過於狹隘。近幾年，美國某些知名企業、機關與學校(如 AOL、NASA、Stanford HighWire Press 等) 採用最新的檢索分群技術，有效解決資訊爆炸問題並提升使用者於資訊搜尋判讀的效率；目前此一分群概念之應用範圍有限，若將其應用於搜尋引擎之改良，或可對資料龐大之搜尋結果依概念進行動態分群。

探究目前之分群方式，一般於呈現文件資料時採用向量空間模型，但此向量空間模型僅能表現詞彙-文件之關係，無法表達文件中詞彙間之關聯性，且尚未解決「同形異義」及「異形同義」之問題，因而此向量空間模型之結果，可能導致文件相似度比對時產生誤判情況而影響整個分群結果。另外，文件之分群演算法中，一般認為階層聚合演算法(Hierarchical Agglomerative Algorithm; HAC)之分群結果較佳，但若實驗資料集過於分散或包括數個主題時，受限於其演算法－合併距離差最小之網頁，卻容易導致其分群效果並未實際達到群內密度高，群間密度低之條件；再者此一演算法最常被批評之缺點為面對龐大之搜尋結果時，將造成分群執行效能不佳之結果 (Tamayo, Slonim et al. 1999)。

本研究提出改良式階層聚合演算法(Revised Hierarchical Agglomerative Clustering Method; RHAC)，並以搜尋引擎 Google 之搜尋結果為實驗樣本，希藉此改良傳統聚合演算法之缺點；本研究之研究目的列示如下：

1. 採用潛在語意分析方法，解決網頁中內含「同形異義」及「異形同義」之問題，並呈現文件之語意表達。
2. 採用潛在語意分析方法，改良階層聚合關係，以提升分群之精確度。
3. 運用 B - Tree 概念之多分支特性，降低執行之時間複雜度，且讓分群之結果更貼切於資料集所呈現之主題。

2. 文獻探討

2.1. 網頁分群

由於網頁所蘊涵之資訊量繁多且雜亂，因此過往研究常針對某些特定之網頁資訊，並配合不同之分群演算法進而達到網頁分群之目的。一般常見可被運用之網頁資訊包含如下：

1. 網頁內容：容易從中獲取特徵值。
2. 超連結：可分為網頁內頁之超連結及連結至其他外來網頁；此超連結表達網頁與網頁間之潛

在語意，且其連結必具有相關性 (Hou and Zhang 2002)。

先前研究主要均針對上述兩項資訊進行網頁分群，Vlajic & Card(1999)先將網頁依其性質區分為組織、部門等類別，再進行網頁及超連結之相似度比對並以類神經網路之分群方法-適應性超文件分群(Adaptive hypertext clustering; AHC)得到分群之結果(Vlajic and Card 1999)。Hou & Zhang(2002)主要著重於超連結資訊，其運用相似度矩陣進行一系列分割，最後形成階層式分群結果 (Hou and Zhang 2002)。

Mukhopadhyay & Sing (2004)基於將網頁事先分類之作法無法符合時代之變遷，因而將分類轉為分群，並結合共同引述(co-citation)方法，認為任二個網頁如被同一來源網頁所引用，則此二網頁應可建立關聯，進而連結相關網頁，萃取出群集(Mukhopadhyay and Sing 2004)。另外 Maggini, Rigutini, & Turchi (2004)指出搜尋引擎搜尋結果之文件相似度雖高，然礙於向量空間模型維度高且無法表達詞彙間之關係，故導致文件分群並不容易；因而採用解決此問題之常用技術－奇異值分解(Singular Value Decomposition; SVD)、概念矩陣分解(Concept Matrix Decomposition; CMD)結合分群 K-means 演算法；最後藉由專家評估此兩種方法及傳統向量空間模型之效果，以奇異值分解結合 K-Means 分群演算法勝出 (Maggini, Rigutini et al. 2004)。

基於仲介網站為能讓使用者可以迅速獲取其所需的廠商資訊，因而對仲介網站之排列資訊由傳統產品類別轉換為以廠商為基礎之分群呈現。因此本研究著重分析使用者之瀏覽行為；再者，由於向量空間模型之維度高將影響分群之效能，Xu 等人提出機率潛在語意分析(Probabilistic Latent Semantic Analysis)整合改良式 K-Means 分群演算法，此改良式 K-Means 演算法允許文件重覆出現於不同群集(Xu, Zhang et al. 2005)。但此方法以 ACM 知識管理研討會資料為測試集，但此測試集因已事先分類而無法達到網頁分群之目的，且著重於以專業領域(知識管理)為基礎之文件，因此無法測試其方法之一般性應用。

綜上所述，網頁分群主要乃為解決網頁分類無法隨時間變化而自動增減類別之缺撼，因此分群之前提，需以未經由人工或自動之分類文件為基礎；此外就搜尋引擎之結果而言，其回傳資訊通常為跨領域，因此網頁分群除考量網頁內容及超連結間之相似度外，尚需進一步衡量網頁內容間之語意關聯，以解決跨領域資訊及異形同義或同形異義之文件內容精確分群問題。

2.2. 分群演算法

分群演算法主要可分為階層式分群法(Hierarchical clustering)、分割式分群法(Partitional clustering) (Han and Kamber 2001)。分群演算法之目的在於找出語料庫中相似的幾個群集，以及各群集之代表點，使用這些代表點來表達原先大量的語料文件，以達成兩項目標：

1. 降低計算量
2. 資料壓縮

上述兩種分群法中，以階層式分群法之分群精確較佳，但因其結構為二元樹，且若語料庫蘊涵雜訊過高時，將影響分群之精確。再者，面對資料量龐大時，執行效能並不佳 (Tamayo, Slonim et al. 1999)；即便如此，仍有許多網頁分群使用階層式分群法，可能因為其可適用於任何領域之資料 (Yao, Chen et al. 2000)且在網頁分群的表現上，效果較佳(Willet 1988; Zamir, Etzioni et al. 1997)。

2.3. 潛在語意分析

潛在語意分析(Latent Semantic Analysis; LSA)藉由統計的方法自動從資料集中萃取詞彙並表示其於資料中之內涵，可視為向量空間模型之延伸，由 Deerwester 等人於 1990 年提出，主要應用於解決同形異義及異形同義問題 (Deerwester, Dumais et al. 1990)。Landauer 等人於 1998 年指出潛在語意分析能評估文件內容所隱含的知識，且可適當表現人類在知識上的推演過程 (Landauer, Foltz et al. 1998)。

潛在語意分析之基本概念以低維度的共同語意因子呈現原先文件和字詞之間的關聯。一般是以奇異值分解(Singular Value Decomposition; SVD)和維度約化(Dimension Reduction)為基礎。奇異值分解是以文件所隱含的知識，抽象轉換到語意空間中，而維度約化能去除文件知識在語意空間中的雜訊，使潛在語意分析能更精確推演出文件所隱含的知識。

3. 研究方法

3.1. 系統架構

本研究提出之系統架構圖如圖 1 所示，處理流程以近期發生之事件且為人們所關注之焦點(如達文西密碼)，做為關鍵字萃取依據。旋即利用 Google 搜尋引擎搜尋，並從搜尋結果中擷取網頁結果之超連結及網頁資訊。迄獲取足夠之網頁資訊後，則針對網頁進行詞彙處理－斷句、斷詞、詞性標注、權重計算、向量空間模型建立等，進而取得網頁之特徵值。

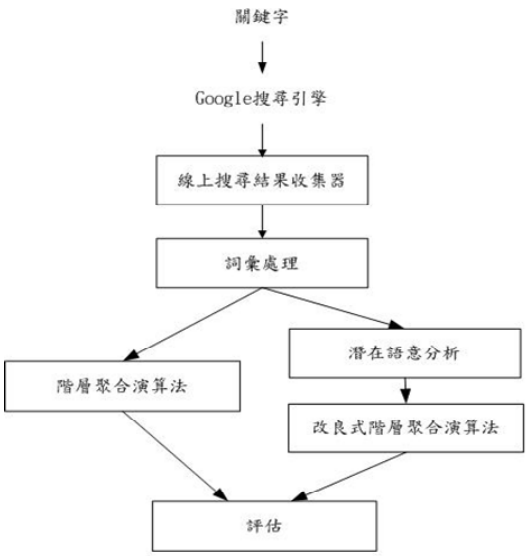

圖 1 本研究系統架構圖

本研究分二階段進行：

1. 階層聚合演算法：

   以向量空間模型來表達文件與詞彙之關係，並搭配傳統階層聚合演算法流程進行網頁之分群。

2. 改良式階層聚合演算法：

   針對向量空間模型，藉由潛在語意分析之前置處理－奇異值分解，來降低向量空間模型之維度。進一步運用潛在語意分析，萃取出網頁間共同語意並搭配改良式階層聚合演算法進行分群。改良式階層聚合演算法主要著重於有效地運用 K-way 之概念來降低階層式樹狀高度，減少時間複雜度。

最後，針對上述之兩種作法－階層聚合演算法及改良式階層聚合演算法進行分群結果精確度和執行效能之評估。

3.2. 線上搜尋結果收集器

本研究設計收集器，自動下載線上搜尋結果，以自動獲得大量且多元之搜尋結果。考量可供下載前提下且網頁分群所需之特徵需是未事先分類者，因此收集器以 Google 搜尋引擎搜尋結果為下載對象。

為獲取足夠資訊，進一步透過超連結自動下載其網頁內容；考量網站內之超連結亦隱含網頁語意資訊，因此本研究除擷取搜尋結果之網頁內容與超連結外，並深入探勘下一層的網頁內容。

考量本研究後續處理所需之網頁內容相關屬性，線上搜尋結果收集器於下載網頁的同時，亦針對網頁的純文字原始檔進行雜訊之去除，此雜訊包含 HTML 標籤、網頁中之宣傳、廣告及與使用者互動之原始碼，只保留網站名稱、網址、網站內容及網頁層級。

3.3. 詞彙處理

線上搜尋結果收集器所擷取之網頁內容在進入分群步驟前，需先對資料庫中之網頁內容逐一進行詞彙處理，並產出向量空間模型以代表文件之特徵。詞彙處理之流程如圖 2，以下將針對每一子流程做一詳細敘述。

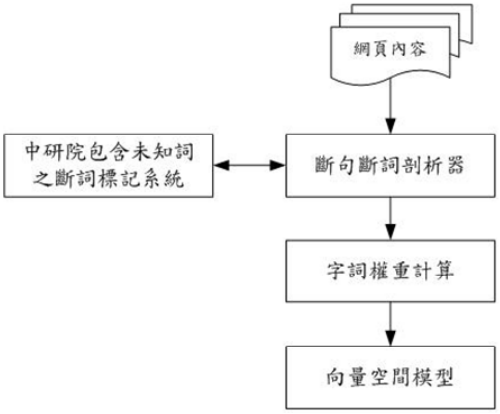

圖 2 詞彙處理流程

3.3.1. 斷句斷詞剖析器

本研究之斷句處理主要依照句號、問號、驚嘆號來當作句子分隔之依據，隨後進行詞庫斷詞法。本研究採用「中央研究院資訊科學研究所詞庫小組」所建構之「八萬詞庫」作為斷詞法之詞庫，另外為補強文件中人名、地名與機關名稱之重要性，尚納入教育部所提供之新詞等以補強中研院之詞庫。

經由詞庫斷詞法處理後之網頁內容，去除其中部份無特別意義的字詞，如我們、你們、他們等。將剩下未辨認出之語句交由中研院包含未知詞之斷詞標詞系統斷出其餘具意義之字詞。

3.3.2. 字詞權重計算

由於字詞剖析器判斷出網頁所包含之字詞，仍無法代表其為網頁之特徵。因此，需計算詞彙於網頁中之權重以衡量此詞彙於網頁之重要性。考量每篇網頁之長短不一，因而採用平滑之詞彙權重方法 - Smooth TFIDF (Croft and Harper 1979)。其 *TF* 計算公式如公式 1，其中 L 為資料庫中網頁長度之平均值，K 為常數，通常介於 1.0 至 2.0 之間，f 表示特定詞彙於單一網頁中之出現次數。

$$TF = \frac{f(K+1)}{f+KL} \qquad \text{(公式 1)}$$

IDF(Inverse document frequency)計算公式如公

式 2，其中N為資料庫之網頁數，n_t為擁有詞彙t之文件數。

$$IDF_t = \log(\frac{N - n_t + 0.5}{n_t + 0.5}) \quad \text{(公式 2)}$$

3.3.3. 向量空間模型

向量空間模型概念主要是將每份文件或段落、句子以向量來表示。本研究以網頁內容為基礎，將網頁內容中所包含之詞彙，視為向量空間之元素。因此，透過向量空間模型的概念，網頁 D 可視為 $D=(t_1,t_2,...t_n)$，其中 t 為網頁中關鍵詞彙。而利用前述 Smooth－TFIDF 來計算字詞權重，每份網頁 D 可視為 $D=(w_1,w_2,....w_n)$ (Salton and McGill 1983)。

3.4. 階層聚合演算法

階層聚合演算法由樹狀結構的底部開始層層聚合。一開始將每一網頁視為一個群集，假設現在擁有 N 筆網頁，則將這 N 筆網頁視為 N 個群集，亦即每個群集包含一筆網頁。其分群流程如下 (Jang 2005)：

1. 將每筆網頁視為一個群集C_i。
2. 找出所有群集間，距離最接近的兩個群集C_i、C_j。
3. 合併C_i、C_j成為一個新的群集。
4. 假設目前的群集數目多於預期的群集數目，則反覆重覆步驟 2 - 3，直到群集數目已將降到我們所要求的數目。

階層式聚合演算法計算群集間之距離有四種計算方式：單一連結(single-linkage)、完整連結演算法(complete-linkage)、平均連結(average-linkage, AL)及沃德法(Ward's method, WM)；此四種方法中以完整連結及平均連結之效果較佳，Chuang和Chien (2002)之研究結果也顯示平均連結之分群效果較佳(Chuang and Chien 2002)；因此本研究採用平均連結聚合演算法，其意義為：群集間距離為不同群集間之各點與各點間距離總和的平均，如公式 3。其中 n_{c_i} 表示C_i之資料個數，公式 4 為歐幾里德之距離計算公式，Tap為文件a中第p個詞彙。

$$d(C_i, C_j) = \frac{1}{n_{c_i} n_{c_j}} \sum_{a \in c_i, b \in c_j} d(a,b) \quad \text{(公式 3)}$$

$$d(a,b) = \sqrt{(\sum_{p=1}^{m} Tap - Tbp)^2} \quad \text{(公式 4)}$$

3.5. 潛在語意分析

本研究於詞彙處理後進行潛在語意分析，藉此得到新的語意空間，此語意空間能更精確的描述詞彙與網頁所代表之意義。

茲舉例本研究於實驗階段所設定之關鍵詞彙之一－達文西密碼之部份搜尋結果，其內容主要可區分為達文西(A1)、達文西密碼書籍之內容(B1、B2、B3)、電影(F1、F2)、宗教－抹大拉(M1、M2)，詳細列示於表 1。

表 1 達文西密碼之部份搜尋結果

編號	標題	網址
A1	達文西密碼非官方網站	http://www.books.com.tw/activity/DAVINCI2/art11.htm
B1	電子明周	http://www.mingpaoweekly.com/htm/movie/main_movie_davinci.asp
B2	達文西密碼不知大家看沒有分享給大家	http://pc-551.tccn.edu.tw/phpbb/viewtopic.php?p=4462&sid=bda5ca9b8fad659c7e4725dc8f7ce9e1
B3	達文西密碼－天使與魔鬼	http://www.wert.18kg.idv.tw/
F1	電影－達文西密碼	http://carrieli.why3s.net/blog/archives/000246.html
F2	雅虎香港網站分類>達文西密碼	http://hk.dir.yahoo.com/Regional/Countries_and_Regions/United_States/Entertainment/Movies_and_Films/Genres_and_Titles/Drama/The_Da_Vinci_Code/
M1	Hong Kong Anglican hurch (Episcopal)	http://www1.hkskh.org/index_ch.php?do=viewtopic&forumid=32&topicid=2870
M2	揭開達文西密碼的面紗	http://life.fhl.net/Culture/davincode/

本研究考量網頁文件中往往存在許多無意義之詞彙，且此矩陣之大小將影響後續分群之時間複雜度。因此，挑選網頁中之詞彙詞性為 Na、Nb、Nc 三種詞彙詞性，且該詞彙需於資料集中出現次數需大於 2、詞彙長度大於等於 2，做為潛在語意分析向量矩陣基準。

依上述八項搜尋結果所建一簡略矩陣 X，矩陣內數值為該詞彙於網頁之權重，且經由奇異值分解後得 U、S、V 矩陣。在此，僅列示 S 矩陣。達文西密碼之奇異值分析如圖 3 所示。

X =

	電影	宗教	羅浮宮	信仰	福音	蒙娜麗莎	藝術家	湯姆漢克
A1	0	0	1.1426	0	0	55.031	102.54	0
B1	0	2.367	4.4259	0	0	0	0	0
B2	0	2.4617	0	0	0	0	0	0
B3	1.4112	12.484	11.318	8.5845	0	0	8.5812	0
F1	11.317	0	21.059	0	0	19.768	0	36.837
F2	5.6829	5.673	0	0	0	0	0	18.6334
M1	0	3.1979	0	2.6562	91.079	0	0	0
M2	0	15.097	0	7.797	102.12	4.0831	7.6086	0

S =

138.2763	0	0	0	0	0	0	0
0	116.7893	0	0	0	0	0	0
0	0	49.99276	0	0	0	0	0
0	0	0	20.3182	0	0	0	0
0	0	0	0	12.40953	0	0	0
0	0	0	0	0	4.062085	0	0
0	0	0	0	0	0	2.130856	0
0	0	0	0	0	0	0	0.146636

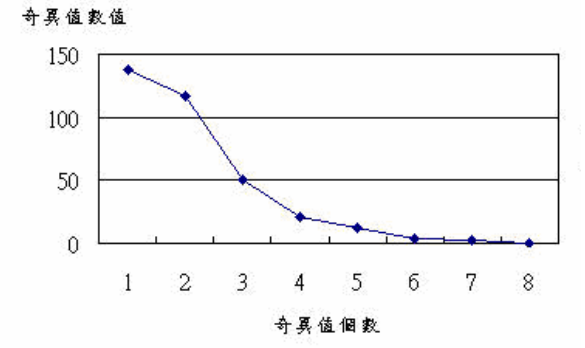

圖 3 奇異值分析圖

由圖 3 可得知其陡降點為 4，故縮減維度設定為 3，維度縮減後得 X'矩陣。潛在語意分析前後所產生之矩陣 X 及 X'，可發現在矩陣 X 中，羅浮宮、蒙娜麗莎等詞彙並未出現於 F2 中，但 F2 包含了電影和湯姆漢克二詞彙。因此，在經過奇異值分解與維度約化後，羅浮宮在 F2 的權重由矩陣 X 中的 0，提升為矩陣 X'的 7.0039，蒙娜麗莎則提升為 4.6985。此外，宗教一詞於 B3 原矩陣 X 中之值為 12.484，但經由奇異值分解與維度約化後，降其值為矩陣 X'的 0.5200；此現象顯示，宗教一詞對 B3 對於達文西密碼書籍內容而言並不重要。

X'=

	電影	宗教	羅浮宮	信仰	福音	蒙娜麗莎	藝術家	湯姆漢克
A1	-0.035	0.9280	1.8712	0.6509	-0.125	54.874	102.61	-0.376
B1	0.4987	0.1475	0.8210	0.0432	0.2230	0.5959	-0.219	1.6090
B2	0.0392	0.0342	0.0648	0.0162	0.2448	0.0542	-0.003	0.1265
B3	1.1619	0.5200	2.0398	0.2086	1.6657	5.0057	6.2486	3.7312
F1	11.830	3.0012	19.597	0.7520	-0.338	17.558	1.1817	38.154
F2	4.2626	1.1091	7.0039	0.2830	0.4411	4.6985	-2.614	13.755
M1	-0.089	8.9436	-0.160	5.0662	90.376	-0.182	0.0330	-0.312
M2	0.0870	10.283	0.2721	5.8192	102.71	4.1330	7.6425	0.2336

3.6. 改良式階層聚合演算法

本研究有鑑於階層聚合演算法採用兩兩合併且其樹狀結構為二元樹之方式，導致階層聚合演算法在面臨龐大資料量時，執行效能不佳。為此，本研究提出改良式階層聚合演算法，其流程如下：

1. 將每筆網頁視為一個群集C_i。
2. 計算所有群集間之平均連結距離，找出距離大於門檻值之群集$C_i,\ldots,C_j$，並計算$C_i,\ldots,C_j$合併後各點到合併後之群中心的距離平方和(沃德法)。若小於門檻值，則進行步驟 3，若大於門檻值，則降低合併之群集數並計算沃德法之距離，直至小於門檻值。
3. 合併$C_i,\ldots,C_j$成為一個新的群集。
4. 假設目前的群集數目多於預期的群集數目，則反覆重覆步驟 2-4，直到群集數目已將降到本研究所要求之數目。

本研究所提出之改良式階層聚合演算法為避免一次合併過多群集且一次合併之群集過大，而產生群集之精確率降低問題。因此，改良式階層聚合演算法利用平均連結聚合演算法來保持群集間之距離，並挑出候選可合併之群集後，利用沃德法維持群集內之密度。沃德法之概念為群集合併後，各點到合併後的群中心的距離平方和，u 代表 $c_i \cup \ldots \cup c_j$ 之平均值，如公式 5 所示。

$$d(C_i,\ldots,C_j) = \sum_{a \in c_i \cup \ldots \cup c_j} \|a-u\|^2 \qquad \text{(公式 5)}$$

4. 實驗設計與結果

4.1. 評估準則

本研究針對階層聚合演算法與潛在語意分析結合改良式階層聚合演算法之分群結果品質進行評估。下列提出二種分群品質評估指標 (Khaled M.Hammouda and Mohamed S.Kamel 2004)：

1. 精確度(Precision)：

精確度為群集中正確分群網頁數之比例，藉由精確度可獲知每群之分群密度是否集中於概念 i，精確度之計算如公式 6 所示。

$$P = Precision(i,j) = \frac{N_{ij}}{N_j} \qquad \text{(公式 6)}$$

N_{ij}為群集j中擁有相同概念i之網頁數，N_j指群集j中內含網頁之數。

2. 熵(Entropy)：

熵為鑑定群集、階層分群法同一層級之同質性高低之計算方法；熵愈低，則表示群集之同質性愈高。反之，熵愈高，則表示群集之同質性愈低。公式 7 為計算群集 j 之熵(Ej)，階層分群法之亂度則加總同一層級中所有群集之熵，如公式 8。

$$Ej = -\sum_i p_{ij} \log(p_{ij}) \qquad \text{(公式 7)}$$

$$E_c = \sum_{j=1}^{m} (\frac{N_j}{N} \times E_j) \qquad \text{(公式 8)}$$

4.2. 實驗資料來源

本研究之實驗資料來源為 Google 搜尋既定事件之結果。有鑑於傳統階層聚合演算法具有資訊量過多則執行效率不佳的問題，因而本研究以能擷取 Google 最大資訊量為前提。資料自動下載後，過濾相同網址、資訊量過少之資料筆數，最後之篩選結果即為本研究之實驗資料集。

表 2 實驗資料集數

熱門話題 / 資料筆數	人名			生活資訊			書籍		
	王建民	琦君	趙建銘	飛蚊症	地震	基測	藍海策略	達文西密碼	世界是平的
搜尋結果	640	766	586	647	781	700	620	677	534
實驗筆數	340	282	389	354	352	248	290	265	254

表 2 為本研究之實驗資料集，關鍵字為針對 2006/06/08 之熱門話題。有鑑於人們獲取資訊來源主要為新聞或書籍，本研究針對此二類收集所下關鍵字之資料。新聞之關鍵字擷取為透過奇摩新聞所提供之功能，包括：一週內最多人搜尋者、最多人轉寄者及最多人瀏覽者之資訊中，擷取出六項關鍵字，並分為兩類－人名與生活資訊。其中，人名為王建民、琦君、趙建銘；生活資訊類為飛蚊症、地震、基測；書籍之關鍵字擷取為摘錄自博客來網路書店之暢銷書排行榜，其關鍵字為藍海策略、達文西密碼、世界是平的。上述 9 項關鍵字再利用 Google 搜尋最大資料量，並藉由自動下載線上搜尋結果之收集器收集資料。過濾後之實驗筆數如表 2 所示。

4.3. 潛在語意分析之維度縮減門檻值分析

本研究所提出之改良式階層聚合演算法需先經由潛在語意分析，萃取文件之重要概念，形成新的語意空間。本研究運用吳忻萍 (1996) 於維度縮減分析時所採用之陡階檢定法，運用此方法潛在語意分析縮減維度門檻值之設定 (吳忻萍 1996)，各類熱門話題之奇異值分析如表 3。

表 3 關鍵詞彙之陡降點

熱門話題 / 陡降點	人名			生活資訊			書籍		
	王建民	琦君	趙建銘	飛蚊症	地震	基測	藍海策略	達文西密碼	世界是平的
陡降點	7	3	6	14	11	5	9	10	9

4.4. 改良式階層聚合演算法之門檻值實驗與結果

本研究針對熱門話題之三類別中取出陡降點較大者－王建民、飛蚊症、達文西密碼作為此階段之實驗資料集。本研究之實驗設定為：以潛在語意分析維度縮減為 13 之前提下，變動平均連結距離與沃德法距離之設定值。本研究實驗三項關鍵詞彙之精確度、熵及效率，做為 9 項關鍵詞彙實驗之門檻值設定依據。

本研究觀察各種門檻值之設定所產生之效果後，挑選三組平均連結距離與沃德法距離門檻值做為實驗之準則，此三組設定如下：

A. 平均連結距離 350，沃德法距離 70000。
B. 平均連結距離 500，沃德法距離 100000。
C. 平均連結距離 650，沃德法距離 130000。

經由上述之門檻值設定後，每提升一個階層其平均連結距離、沃德法距離會以倍數成長。實驗結果，飛蚊症、達文西密碼所形成之階層數為 3，王建民因其實驗集主題過於分散，在 A、B 門檻值下，層級 1 時並未做任何的合併動作，總層級數為 4。在衡量精確度、熵時，本研究將其層級 2 視為分群結果之層級 1，原層級 1 則略過不計。最後，精確度、熵之實驗結果將會區分為層級 1 與層級 2 來進行最佳門檻值設定之選擇分析。

4.4.1. 精確度

三項關鍵詞(王建民、飛蚊症、達文西密碼)基於上述三組設定之門檻值條件後，層級 1 所形成之實驗結果如圖 4，層級 2 之結果於圖 5。

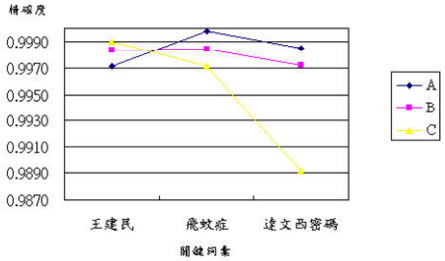

圖 4 層級 1 關鍵詞彙與門檻值設定之精確度

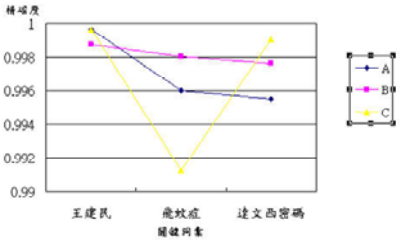

圖 5 層級 2 關鍵詞彙與門檻值設定之精確度

從圖 4 中可發現，C 門檻值在層級 1 中，其精確度為最低，且震盪幅度最大；B 門檻值則精確度較穩定，且其值介於另二種門檻值之中；此外，從圖 5 中可得知，B 門檻值於層級 2 時，有二項關鍵詞彙(飛蚊症、達文西密碼)之表現均優於其它二項門檻值設定。

4.4.2. 熵

關鍵詞彙與門檻值設定之熵實驗結果層級 1 如圖 6，層級 2 如圖 7 所示。

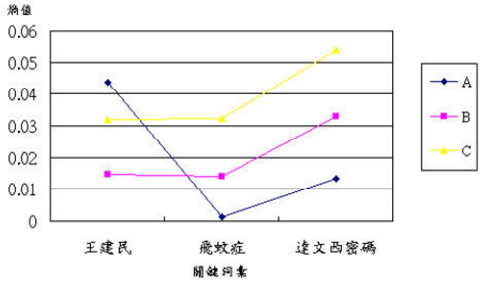

圖 6 層級 1 關鍵詞彙與門檻值設定之熵值

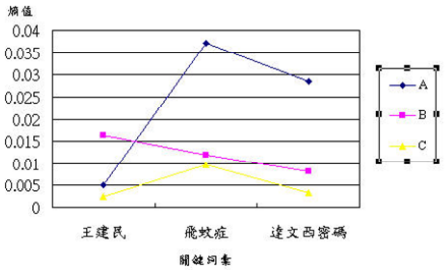

圖 7 層級 2 關鍵詞彙與門檻值設定之熵值

圖 6、7 可得知，A 門檻值會隨關鍵詞彙的不同，而有明顯之震盪； B 門檻值，熵介於其他兩者之間。圖 7 中，A 門檻值其亂度為三種門檻值中最高，乃因其於層級 1 時作亂度最小之群集合併處理，但至層級 2 沃德法之門檻值提升時，其合併後之亂度亦因主題之分散而提升。

4.4.3. 執行效能

改良式階層聚合演算法於三種關鍵詞彙與距離門檻設定之執行效能，列示如圖 8。

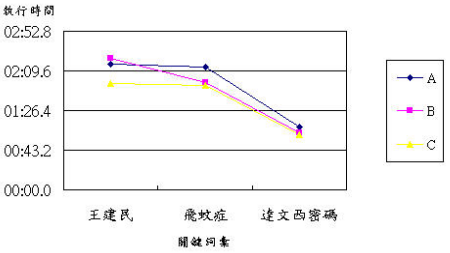

圖 8 關鍵詞彙與門檻值設定之執行效能

由圖 8 中，在 A 門檻值前提下，執行效能於二項關鍵詞彙中較差，C 門檻值之效能較佳。綜上所述，於考慮精確度、熵搭配執行效能之條件下，選擇 B 門檻值設定。

4.5. 改良式階層聚合演算法與階層聚合演算法之評估

此階段探討改良式階層聚合演算法於 B 門檻值及維度約化 13 之設定下，所形成之分群結果與階層聚合演算法之比較。以下列示在達文西密碼以 8 項搜尋結果為前提下，傳統階層聚合演算法(圖 9)與改良式階層聚合演算法之分群結果(圖 10)，其中 *Ci* 表示層級 *j* 中之群集編號。

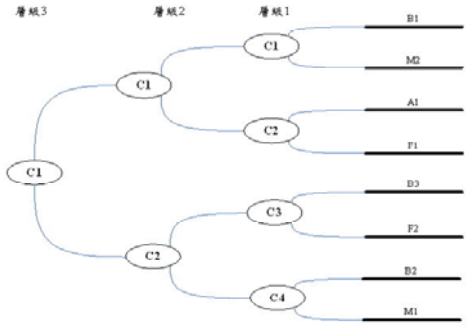

圖 9 階層聚合演算法之分群結果

從圖 9 之層級 1 中可得知 C1、C4 的合併，均包含達文西密碼一書的內容和宗教之概念來做為合併。而達文西密碼之電影則分別與書籍、達文西進行合併；層級 2 時，C1、C2 均包含達文西密碼之書、電影及宗教之概念，此種情況解釋出階層聚合演算法明顯受限於自身演算法中，與距離差最小之網頁合併，因而造成相同層級中多個群集擁有相同概念。

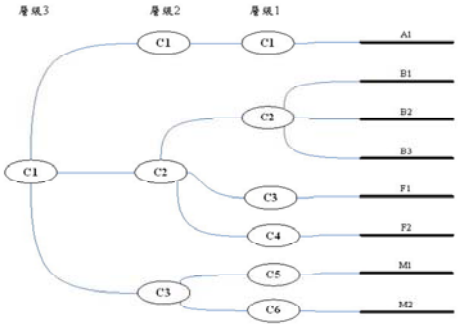

圖 10 改良式階層聚合演算法之分群結果

圖 10 為改良式階層聚合演算法之分群結果。層級 1 時，僅有 B1、B2、B3 通過平均連結距離與沃德法距離之門檻值進行合併，且該群集集中於達文西密碼書之內容;層級 2 之門檻值為層級 1 之 2 倍，在層級 2 中會合併層級 1 的三群，而概念則集中於達文西密碼之書與電影之網頁內容上。宗教之網頁 M1、M2 亦於層級 2 中合併。而達文西(A1)並未與任何之群集進行合併，乃因該網頁為探討達文西此人本身之相關議題，如畫作等，而非如達文西密碼書中之內容，或是宗教-抹大拉之類的議題。

綜合傳統階層聚合演算法與改良式階層聚合演算法之結果，可獲知改良式階層聚合演算法透過潛在語意分析能確實地萃取出網頁中所探討之概念。進而產生新的語意空間，並提升精確度；熱門話題在兩種演算法之效能比較如圖 11、12、13 所示，其中精確度、熵之評估取$\lfloor H/2 \rfloor$及其往上一層。其中，H 指階層式分群結果之樹狀高度。階層聚合演算法之高度為 9，因而取 5、6 層級之平均做為衡量之數值。改良式階層聚合演算法則高度集中於 3，若其高度為 4 且層級 1，且未進行任何合併動作，則視層級 2 為層級 1。此法以 1、2 層之精確度、熵值之平均做為衡量。

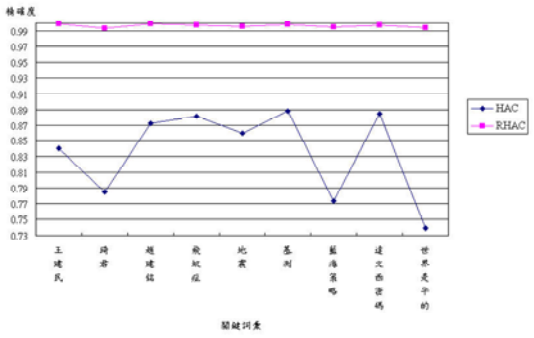

圖 11 精確度

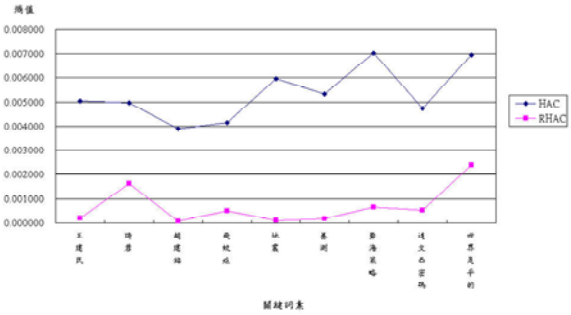

圖 12 熵

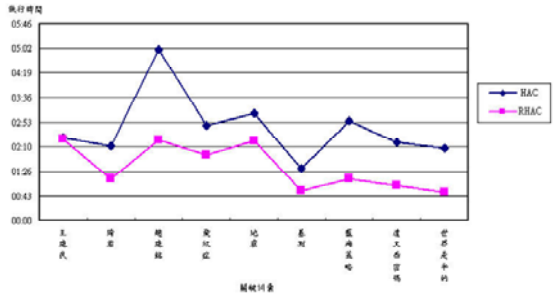

圖 13 執行時間

圖 11、12、13 呈現出本研究所提之改良式階層聚合演算法之精確度在三類熱門話題中皆高於 0.99、熵低於 0.003 之下。8 項關鍵字之執行時間皆較傳統階層聚合演算法時間來得低；其中，趙建銘一詞於傳統階層聚合演算法中執行效能最高，但本研究所提出之改良式階層聚合演算法仍可穩定在 02:53 分的執行效能下；另外，以王建民做為關鍵詞，其兩種方法之執行時間相近，乃因王建民之資訊概念較為分散。因此，改良式階層聚合演算法於層級 1 時未進行任何的合併動作，進而造成執行時間逼近於傳統階層聚合演算法。

5. 結論與未來研究方向

5.1. 結論

本研究所提出之改良式階層聚合演算法，乃建基於潛在語意分析之上。藉由潛在語意分析解決「同形異義」和「異形同義」之問題，並萃取資料集之潛在語意概念。在潛在語意分析之維度縮減係數的選擇上，可與陡階檢定相配合，以避免捨去具有重要意義之奇異值，導致分群之品質不佳。

改良式階層聚合演算法運用 K-way 概念，藉此加速相似概念之合併速度，以降低時間複雜度及樹狀結構之高度。為避免發生一次合併過大而導致精確率不佳的問題，本研究結果沃德法維持群集內之密度以及平均連結聚合演算法保持群集間之距離。實驗結果顯示，傳統階層聚合演算法除了在大量資料下執行效能較差之外，其分群之品質亦受限於需合併最相似之網頁，而容易導致同一層級之群集間會擁有相同概念，未能確切的加以區分出群集來。反觀，改良式階層聚合演算法可有效地解決上述之問題，且於實驗資料集之測試下，精確度高於 0.99、熵低於 0.003，執行效能不受限於資料筆數之大小，能有效且快速的進行分群。而影響改良式階層聚合演算法之效率，乃在於資料集之亂度會影響底層合併之效率，將來若為解決亂度高時之效率問題，建議可適度地調整距離門檻值。綜上所述，本研究所提之改良式階層聚合演算法不論於效率、精確度上都較傳統階層聚合演算法佳。

5.2. 未來研究方向

本研究乃針對搜尋引擎之搜尋結果進行分群，期許將來能實際運行於搜尋引擎中，建議後續之研究可有下列幾個方向：

1. 本研究僅針對分群演算法進行改良，未來可實際運用改良式階層聚合演算法於搜尋引擎中，並加以比較加入前／後之成效。
2. 由於奇異值維度約化係數及距離之門檻值，乃針對本實驗之資料集設計，未來可針對不同搜尋引擎所提供之資料，做一統整之評估，以決定一個適配於搜尋引擎之門檻值。
3. 本研究實驗 Google 搜尋結果發現，搜尋結果中仍有許多與使用者所下之關鍵字較不相關之結果。建議可藉由潛在語意分析等方法，萃取出相關性較高之結果給予使用者。
4. 如搜尋引擎可事先對內存之網頁進行粗略之分群，未來可針對萃取出之結果，進行即時動態分群之研究。至於如何針對新進之網頁與現有資料庫之分群結果做一快速整併之方法，仍需未來研究努力突破此一問題。

參考文獻

[1] Chuang, S.-L. and L.-F. Chien (2002). Towards automatic generation of query taxonomy: a hierarchical query clustering approach.

[2] Croft, W. B. and D. J. Harper (1979). "Using Probabilistic Models of Document Retrieval Without Relevance Information." Journal of Documentation **35**(4): 285-295.

[3] Deerwester, S., S. T. Dumais, et al. (1990). "Indexing by latent semantic analysis." Journal of the American Society for Information Science **41**(6): 391-407.

[4] Han, J. and M. Kamber (2001). Data Mining: Concepts and Techniques. San Francisco, Morgan Kaufmann.

[5] Hou, J. and Y. Zhang (2002). A Matrix Approach for Hierarchical Web Page Clustering Based on Hyperlinks. Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) - (WISEw'02). , IEEE Computer Society

[6] Jang, J.-S. R. (2005). Data Clustering and Pattern Recognition

[7] Khaled M.Hammouda, S. M., IEEE and S. M. Mohamed S.Kamel, IEEE (2004). "Efficient Phrase-Based DOcument Indexing for Web Document Clustering." IEEE Transactions on Knowledge and Data Engineering **16**: 1279-1296.

[8] Landauer, T. K., P. W. Foltz, et al. (1998). "An introduction to latent semantic analysis." Discourse Processes **25**(2): 259-284.

[9] Maggini, M., L. Rigutini, et al. (2004). Pseudo-Supervised Clustering for Text Documents.

[10] Mukhopadhyay, D. and S. R. Sing (2004). An algorithm for automatic Web-page clustering using link structures.

[11] Salton, G. and M. J. McGill (1983). Introduction to Modern Information Retrieval. New York, McGraw-Hill Co.

[12] Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." Proc Natl Acad Sci U S A **96**(6): 2907-2912.

[13] Vlajic, N. and H. C. Card (1999). An adaptive neural network approach to hypertext clustering.

[14] Willet, P. (1988). Recent Trends in Hierarchical Document Clustering: a Critical Review. Information Processing and Management.

[15] Xu, G., Y. Zhang, et al. (2005). Using Probailistic Latent Sematic Analysis for Web Page Grouping. 15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Application.

[16] Yao, Y., L. Chen, et al. (2000). Using cluster skeleton as prototype for data labeling. IEEE Transactions on Systems.

[17] Zamir, O., O. Etzioni, et al. (1997). Fast and Intuitive Clustering of Web Document Third International Conference on Knowledge Discovery.

[18] 吳忻萍 (1996). 以隱藏語意索引為基礎的中文資訊檢索. 資訊管理學系. 台北, 國立台灣大學. **碩士:** 56.