圖書資訊學研究 4：1 (December 2009)：133-164

運用文件分群與概念關聯分析技術協助網誌瀏覽：任務導向評估方法

Supporting Browsing of Blogs via Text Clustering and Concept Analysis Techniques: A Task-Oriented Evaluation Approach

吳怡瑾

I-Chin Wu

天主教輔仁大學資管理系 助理教授
Assistant Professor
Department of Information Management
Fu-Jen Catholic University

方友杉

Yu-Shang Fang

天主教輔仁大學資管理系 研究生
Graduate Student
Department of Information Management
Fu-Jen Catholic University

喻欣凱

Hsin-Kai Yu

天主教輔仁大學資管理系 研究生
Graduate Student
Department of Information Management
Fu-Jen Catholic University

投稿日期：2009.11.17；接受日期：2009.12.16
email: 吳怡瑾icwu.fju@gmail.com；方友彬azcdchar@msn.com；喻欣凱wuinfans@gmail.com

【摘要 Abstract】

由知識管理的角度觀之，網誌（blog）提供一種新型態的知識交換模式，即網誌空間提供作者與讀者積極的互動轉換關係，促進知識的交流。由於大部份的網誌內容是偏於生活中的隨筆，其形式較無法像一般的網站，在設計之初就能規劃好所欲討論之內容及整體架構。本研究主要結合階層式分群法與概念關聯分析以協助網誌作者與讀者能夠正確的重新組織文章並進而有效的瀏覽文章。研究以使用者網誌的文章為例，設計考量使用者觀點之階層式文件分群方法將網誌之文章進行分群，以達成彈性分群之目的並由使用者角度重新組織網誌文章。研究並透過正規概念分析（formal concept analysis）方法建構與顯示各群組關鍵字詞之階層關係，主要目的為協助分群之命名並更有效的協助讀者進行資料的搜尋與瀏覽。研究並嘗試透過設計問卷與模擬使用者搜尋任務（task-oriented approach）以評估方法之有效性，實證結果顯示，結合分群與概念關聯呈現均有助網誌作者與讀者對網誌內容的理解，提升瀏覽效率與節省搜尋時間。

Blogs are a good online tool that encourages information exchange and knowledge sharing. However, blog users often face two challenges. First, blog contents are often categorized vaguely or inadequately by authors. Second, many popular blogs sort content by date. Consequently, when a blog user does not possess the right keywords for information retrieval, he or she must take time to browse the contents by chronological order until the relevant content is identified. This study addressed the problems by using the hierarchical agglomerative clustering and formal concept analysis methods to re-classify blog contents. To evaluate the effectiveness of the proposed automated text clustering solution, we further conducted user task-oriented evaluation. The results showed that the technique can help authors (bloggers) define new categories and refine existing categories. In addition, the concept hierarchy applied in each category helped the blog users to quickly discover the needed information.

[關鍵字 Keywords]

網誌；概念階層；正規概念分析；階層式文件分群；任務導向評估法

Blog; Concept hierarchy; Formal concept analysis; Hierarchical agglomerative clustering; Task-oriented evaluation approach

壹、緒論

在2001年網路泡沫化後，企業開始醞釀新型態的網路經營模式，除了銷售商品及服務外，也重新正視、定位使用者，如：Facebook、Flicker、Bloglines、Del.icio.us、Socialtext等，皆是以使用者為本而提供新型態的社會媒體服務（social media service）網站。近幾年，隨著Web 2.0之知識共享平台概念的興起，網誌（blog）與維基協作（wiki）等更為網路促進知識交流與傳遞的重要媒體與平台（Hill, et al., 2006; Rosenbloom, 2004）。哈佛商學院教授McAfee（2006）由Web 2.0的技術中萃取6項元素，透過這些元件集成的技術，提出「Enterprise 2.0」的概念，這六個元件分別是：搜尋（search）、連結（links）、書寫（authoring）、標籤（tags）、延伸（extensions）與信號（signals），企圖為企業的知識工作者架構出方便的技術與工具，例如網誌、維基協作等技術，達成企業協同運作以應用群體智慧。

過去一般網路上的討論平台大多是提供群體一個共享資訊的單一環境，想要有個人的發表空間，需瞭解網頁建構等相關技術，因此，一般人較不容易建置屬於自己的網路發表空間。待以網誌為基礎之相關網站出現後，使用者不用懂太多的網頁設計語法便可以輕鬆地擁有獨具風格的個人發表空間、自由地發表資訊，使文章的發佈更即時（Rosenbloom, 2004）。近幾年，網誌除了提供使用者抒發情感、生活記事、交換想法極佳的網路空間外，更廣為應用於支援教學、學術研討、組織知識管理等正式活動。Divitini、Haugalokken與Morken（2005）提出教師可以使用網誌支援教學相關活動。Yao（2006）提出利用網誌協助研究者完成研究活動與交流。此外，網誌不只是將使用者的意見與想法發表在網路上，透過搜尋、回應、連結、標籤、訂閱等社群互動機制功能，更可以讓興趣相投的網友交換意見並深入瞭解彼此。網誌是一種以文章的形式在網路上發表內容的一種共享空間，其呈現的方式乃是按照發表順序倒敘排列，將最新的文章置於頂端之網頁，其特色為個人或群體依時間作的一種記錄，具有不斷更新的特色。而目前大部份使用者主要透過自訂分類或以標籤（tag）組織網誌文章，因此，當網誌文章逐漸累積增加

時，使用者常需重新組織分類。目前網誌分類的缺點為不夠精細，常會產生某類別之文章群數過大，或是類別所包含的內容過於繁雜，導致網誌作者不易從龐大的文章群中找尋到所需要的資訊。對網誌的讀者而言，需要花費大量的時間瀏覽網誌，才能找到特定的資料，或是對整個網誌中所談論到的概念以及範圍獲得瞭解。網誌分群相關研究仍有限，主因在於傳統對網誌概念為使用者之隨筆作品。文字選用對於分群品質之效力仍待驗證，Brooks與Montanez（2006）提出由使用者標籤進行分群，初步成果顯示透過使用者標籤有助於作者文章分群，但並無法協助讀者明確瞭解文章內容。

本研究為針對使用者網誌進行文章分群，協助使用者建立新的分類，並依照新建立的類別建構概念關聯圖呈現使用者網誌主題為研究目的。研究採用中央研究院（簡稱中研院）中文詞知識庫小組（Chinese Knowledge and Information Processing，簡稱CKIP）提供之中文斷詞服務進行文章斷詞（中文詞知識庫小組，無年代；Chen & Liu, 1992），並依照TF-IDF字頻方式計算字詞權重、選取文章關鍵詞並計算文章間之相似度（Salton & Buckley, 1988）。研究採用階層式文件分群方法（hierarchical agglomerative clustering，簡稱HAC）將網誌文章進行分群，進以作為協助使用者建立新分類之依據。研究接著透過正規概念分析法（formal concept analysis，簡稱FCA）分析各群文件字詞以視覺化呈現網誌各類別之關鍵詞階層關係，其主要目的乃是協助網誌作者建立更貼近網誌內容的新分類；並建構以關鍵詞為基礎之概念關聯以呈現網誌網誌各類別之內容。因此，研究預期解決問題為：(1)解決網誌類別的文章數量過於龐大的問題：由於網誌容易產生某些類別之文章群數過大的問題（甚至有些網誌的作者撰寫網誌時是不進行分類的），故本研究擬透過自動化文件分群及呈現各類別關鍵詞概念圖的方式，協助網誌作者建立新的類別，降低類別的文章數量過於龐大的問題；(2)協助網誌作者調整文章的類別：由於網誌作者在撰寫網誌文章的時候並無法預知將來還會再建立的文章類別，因此一開始文章所屬的類別可能在日後將需要重新調整其所屬類別。為了節省網誌作者花在重新瀏覽與更改文章類別所花費的時間，本研究將透過自動文件分群以及概念關聯圖呈現的方式給予網誌作者協助；(3)利用概念關聯表達網誌的內容：為了能讓網誌讀者更容易瞭解以及搜尋網誌中眾多的資訊內容，本研究將透過概念關聯的建構，呈現讀者該網誌中各個類別所包含的概念（關鍵詞）以及概念之間的關係，協助讀者對於該網誌的瞭解。本研究希望利用文件分群技術協助使用者訂定適合網誌內容之類別；此外，將以正規概念分析技術

建構與呈現概念階層樹，幫助網誌作者訂定類別名稱並且有效率的增加資訊搜尋的效力。預期研究結果將有助於網誌作者進行更貼近其網誌內容之新分類，協助讀者對網誌的理解，以節省過多的瀏覽與搜尋時間。

本文涵蓋以下部份：首先，本文在文獻探討部份介紹文件前處理、文件分類技術、正規概念分析法與概念圖的理論與應用；接著介紹研究所使用網誌系統之分析架構，並說明研究假設與方法、相關演算法與說明範例；後文續而詳述實驗步驟及研究採用之任務導向實驗評估模式，並討論評估結果；最後提出結論及未來研究方向。

貳、文獻探討

本研究採用文字探勘技術分析網誌文章內容，主要採用中研院提供之CKIP斷詞服務進行網誌文章斷詞以及詞性標注，並以階層式分群法進行文件分群，其結果作為協助網誌作者建立新分類之依據。研究並以正規概念法分析文章中關鍵詞之間的關係，建構網誌各類別之關鍵詞關聯，視覺化呈現類別之內容。

一、中文斷詞與關鍵詞擷取

對於中文而言，「詞」是最小且具有意義的語言單位，因此，進行語言的處理時，必須先對文章內容做斷詞的動作。基本上自動分詞大多利用詞典中所收錄的詞來和文本做比對，以找出可能包含的詞。但是透過詞典來做分詞所遇到最大的瓶頸就是「未知詞」（詞典中沒有的詞）的擷取，根據統計，一篇文章當中約有3%~5%的詞彙是未知詞，新聞類的文章更是遠高於此比例。而某些類型未知詞的詞構非常複雜，也不一定具有強烈的統計特性。未知詞的擷取問題一直是中文語言處理上一個重要且困難的研究課題，必須利用句子中的結構特性進行斷詞的分析，馬偉雲（無年代）、Chen與Bai（1998）對於未知詞的辨認與擷取是透過其下作法：(1)初步斷詞；(2)未知詞偵測；(3)中國人名擷取；(4)歐美譯名擷取；(5)複合詞擷取；(6)由下而上合併演算法(bottom-up merging algorithm)；(7)重新斷詞。

中文斷詞不存在任何單一方法可以完成所有的斷詞動作（無論是透過詞庫或是各種詞性分析方法）；當處理不同領域的文章時，該領域的專有名詞極有可能造成系統無法辨認，產生斷詞上的錯誤。斷詞方法主要分為(1)詞庫比對；(2)文法剖析；(3)統計分析。現今段詞多以連詞（n-gram）方式，或由系統提供事先建好的詞庫，輔以定期更新或系統

自動擷取新詞的方式以自動建構關鍵詞（索引詞）。文法剖析法主要透過自然語言處理技術的文法剖析程式，剖析文件中之名詞片語，再過濾不適合之詞彙。統計分析法為分析文件收集足夠的統計數據後，再從中擷取關鍵詞，最簡單的統計參數是計算詞彙發生的頻率，將詞頻落在某一範圍的詞彙取出。本研究使用中研院中文詞知識庫小組（CKIP）所建構之詞庫，該詞庫採用結合詞庫比對法與統計分析法之混合斷詞法，以確保能自動擷取重要的未知詞。中研院CKIP小組之斷詞主要步驟為(1)以中研院詞典中的九萬目詞為基礎，進行字詞分類與選取；(2)採用中央標準局提出的「中文資訊用分詞國家標準草案」的分詞原則；(3)詞類標記。

在完成文章的斷詞動作之後，開始進行關鍵詞（key term）擷取與文件分群等程序。關鍵詞的意義，即是一篇文章中最能代表該篇文章意義的詞。關鍵詞的選擇可透過所建立的詞庫以進行篩選或是透過計算字詞在文章中出現的頻率TF-IDF（term frequency-inverse document frequency）予以選擇。一般來說，資訊檢索可透過文字權重（term weighting）計算以達成關鍵字詞選擇（Salton & Buckley, 1988）。最常見到的方法則是TF-IDF，字頻（term frequency，簡稱TF）代表的是該字詞在文章中出現的次數，若TF越高，則表示這個詞在這篇文章中的重要性越高；而反轉頻率（inverse document frequency，IDF）則代表所有出現該詞的文章篇數之倒數，若IDF值越大，則表示該字詞越具有獨特性（即鑑別力）。文字權重之公式如下（Salton & Buckley, 1988）：

$$w(k_i,d)=\frac{1}{\sqrt{\sum_i\left(tf(k_i,d)\times\log(N/df(k_i))\right)^2}}tf(k_i,d)\times\log\frac{N}{df(k_i)} \quad (1)$$

公式中之第一項表示正規化因子，即根據每篇文章關鍵字數做正規化；公式中$tf(k_i, d)$為關鍵字k_i在文件d中出現的頻率（TF）；N 表示總文件數；$df(k_i)$為關鍵字k_i共出現於多少篇文件；因此，$\log\frac{N}{df(k_i)}$ 即為文件反轉頻率（IDF）。

二、文件分群

文件分群是資料分析中常見的方法之一，資料分群主要精神是將大量資料分成若干群組，使每個群組內有較高的相似度，而群組間則有較低的相似度。資料分群可以用來將一堆資料做有意義的分割，方便資料簡化與分割處理。到目前為止，許多學者對此提出許多研究結果以及新的分群方法，分群技術主要有切割式分類法（partition-based clustering）

和階層式分類法（hierarchical clustering）兩種；另外還有結合其他資訊技術而產生的分類方法，如利用模糊（fuzzy）觀念的模糊分類法（fuzzy clustering），類神經網路的類神經網路分類法（artificial neural network for clustering），使用基因演算法的基因演進分類法（evolutionary approach for clustering），混合各解決法的混合法（mixture-resolving and mode-seeking）等（Han & Kamber, 2000; Jain, Murty, & Flynn, 1999）。

階層式分群演算法（hierarchical clustering algorithms）可以再細分為聚合法（agglomerative algorithm）與分裂法（divisive algorithm）二種。聚合法是先將每一筆資料視為一個群聚（cluster），然後每次將特性最相近的兩個群聚合而為一，直到群聚數目達到事先所設定的停止條件為止。一般應用階層式分群演算法時，較常使用聚合法，而其群聚距離的評估方式目前有幾種agglomerative hierarchical clustering演算法，分別是單一鏈結（single link）、完全鏈結（complete link）和群平均相似度（group average），其差異為計算群與群之間相似度（inter-cluster similarity）所採用的方法不同。

就以上三種分群演算法結果的品質來說，採用complete link演算法計算相似度會形成比較嚴謹的結果，也就是在同一群裡面的資料，會有比較高的相似度。然而在single link以及group average的分群結果顯示，儘管資料間的相似度很小，但是結果還是被分在同一個群聚中。Single link演算法的特性，只考慮到兩群之中是否有相似度高的物件，如果有就把兩群合併，因此，將導致分群結果會有較大的群集與一些包含較少物件的小群集；甚或是物件幾乎串在一起，形成一個大群集，即為所謂的鏈結效應（chaining effect）（蔡宜龍，2002）。不過就計算時間複雜度方面來說，single link與group average的時間複雜度為O(n2)，而complete link的時間複雜度為O(n3)。所以就計算時間複雜度以及分群結果品質的諸多考量之下，group average演算法算是一個比較常被使用來做文章分群的方法（Chuang & Chien, 2004; Jain, et al., 1999; Liu, Wu, & Chen, 2006），故本研究採取group average演算法進行HAC分群。

三、由正規概念分析法產生概念關聯

正規概念分析法（以下簡稱FCA）是一種從資料集合（data sets）中發現概念結構（conceptual structures）的資料分析理論（Ganter & Wille, 1999）。FCA於1982年由Wille提出，並且快速的發展並應用到許多領域如：醫學、心理學、音樂學、語言學、資料庫、圖書館學、資訊科學、軟體工程、生態學及其它領域（Wille, 1982，轉引自Everts, Park, &

Kang, 2006）。FCA的特色是它可以從資料集合的結構中產生圖形化形象（graphical visualization），特別是在社會科學上常常無法充分的捕捉（captured）量化的分析（quantitative analyses）。FCA也擴充了正式分析方法（formal analysis methods），並且在統計及概念分析領域之中可以互補，將包含有物件（objects）以及屬性（properties）的資料來源矩陣（matrix）以圖形化的方式表達出來。本研究將利用FCA的理論以建構概念與概念之間的階層關係。由於本研究中的「概念」是採用關鍵詞向量集合產生的，所以需要對概念作比較明確的定義。一般而言概念之間的關係可區分為獨立（independent）、交錯（intersection）和繼承（inheritance）三種關係。本研究最重視的就是第三種情況，也就是繼承的關係，如果一個概念B可以被另一個概念A所完全包含的話，我們可以說概念B完全包含於概念A，概念A與概念B有繼承的關係。比較小的概念會繼承比較大的概念的特性。利用這種繼承關係，發展出以關鍵詞為主的概念階層。

將概念關聯繪製成概念圖目前已是教育上常見的一種工具，近年來有不少研究課題著重在將學習的概念加以轉換並圖表化以呈現給學者（林柏成，2002；吳育龍，2000；Jonassen, Reeves, Hong, Harvey, & Peters, 1997；McAleese, 1994, 2000），而這些關鍵詞和關聯所組成的圖表，被許多研究者和學習者用來協助診斷誤解、增進學習效率的方法和一瞥學習者的學習成效，並以學生所畫的概念圖作為評量之依據（高宜敏，2001）。陳年興與孫振凱（2002）透過網際網路上搜尋引擎的功能，針對全國博碩士資訊網進行關鍵詞的擷取，並分析關鍵詞出現在網頁上的關聯，建構網際網路上的知識分佈圖，以輔助線上教學，讓學習者探索網際網路上的資料時能夠更有效率，避免學習者發生認知迷失與資訊超載的情況。陳道輝（2003）利用論文之標題、關鍵詞與關鍵詞出現之頻率外，更利用論文中的摘要內容建立知識關鍵詞之間的關聯強度及關聯相似度，最後依據關聯強度與關聯相似度，實作出自動化建立知識關聯圖的系統。

除此之外，概念圖尚可呈現領域的知識演進歷程，陳年興、謝盛文與黃琬婷（2007）成功建構資管領域之知識概念圖與知識關聯圖，並且加入時間維度和關係程度的因素，成功地展現出資管領域知識之演進歷程。巫啟台（2002）提出一個資訊萃取及其概念圖分析架構，利用自然語言處理詞性標記技術，找出文件中專有名詞之重要詞彙，再以關聯樣版將檔案中的關聯資訊萃取出來，提供一個從文字性資料中有效發掘關聯資訊的方法，並且以「項目

－關聯－項目」之形式展現關聯資訊之概念圖。Chen、Kinshuk、Wei與Chen（2008）利用文字探勘技術，自動化地建構數位學習（e-Learning）領域之概念圖，區分為「以所有研究主題為主的概念圖」與「以單一研究主題為主的概念圖」，並且針對時間因素，實作不同時間區段之關鍵詞。提供該領域之研究者進行課程設計，以及幫助欲瞭解數位學習相關知識之學習者一個詳盡且有用的參考架構。

參、系統架構與模組

研究對象為網誌作者與網誌閱讀者，因此，研究目的在協助網誌作者進行更貼近其網誌內容之分類、協助網誌讀者對喜好網誌作者其內容的理解，節省過多的瀏覽與搜尋時間。預期研究結果為：(1)利用文件分群技術能協助網誌作者重新組織分類；(2)利用文件分群技術以及建構概念階層有助於讀者對所閱讀blog的理解；(3)透過正規概念分析技術建構之概念階層可以協助網誌作者依據各類別所產生之關鍵字詞關聯階層，替網誌分類更準確的命名。

研究首先以HAC演算法將網誌之文章進行分群，後續採用FCA技術的繼承關係判別建構各群組的主題關聯階層圖。圖1所示為系統主要模組，分別為文件前處理模組（text pre-processing module）、網誌分群模組（blogs clustering module）與關鍵字概念階層產生模組（keyword-based concept generation and visualization module）。此外，系統提供使用者介面輔助使用者可以依其偏好而進行彈性的網誌分群與概念呈現。

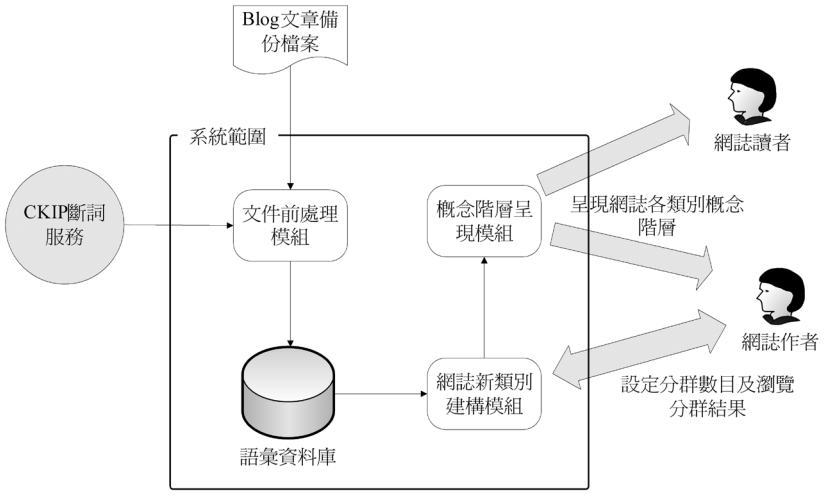

圖1 系統架構圖

- 文件前處理模組：該模組主要進行網誌備份文章之前處理，其主要包含：標籤判別與移除、使用中研院CKIP斷詞網路服務進行網誌文章斷詞、詞性定義與關鍵字詞選擇。
- 網誌分群模組：該模組目的在透過研究所採用之HAC演算法將網誌文件進行重新分群。該模組將提供使用介面，讓使用者可以瀏覽網誌分群結果並比較先前所設定之所有類別，進而讓作者建立更貼切網誌內容之文章分類群組。
- 關鍵字概念階層產生模組：該模組主要目的在透過分群結果結合FCA技術呈現網誌文件之關鍵字集階層關係。使用者將可以透過關鍵字集自訂分群群組名稱或者參考關鍵字概念階層，自訂文件標籤，滿足Web 2.0使用者自訂標籤的精神。

肆、網誌群集與概念關聯分析

㈠文件前處理模組

文件前處理階段主要分為以下步驟：

1. 刪除文章格式：第一步驟為移除所有備份網誌文章的格式、擷取文章標題及內文所組成的資料、並存為文字檔案。
2. 網誌文章斷詞與文章未知詞擷取：利用中研院CKIP所建構的詞彙庫對網誌文章進行斷詞。而未知詞的萃取，除了針對特定類型未知詞作專屬的詞構分析之外，還將所有種類的未知詞的詞構以免境式語法（context free grammar）搭配一個簡單的bottom-up merging algorithm。此方式能夠解決大部份統計特性低的未知詞擷取問題（馬偉雲，無年代）。由於網誌文章中有部份屬於個人紀錄或情感抒發，往往有不少使用者定義的新詞彙，研究主要進一步將CKIP線上斷詞服務所發現的未知詞組進行篩選以選擇出重要的新詞彙。
3. 選取文章之關鍵詞：關鍵詞的權重計算方式為透過TF與IDF計算而得（Salton & Buckley, 1988）。研究選取名詞為代表關鍵字，此外研究將計算每篇網誌文章之平均名詞個數，將每篇文章之關鍵詞依照權重由高至低排列，取前k個名詞作為代表該篇文章之關鍵詞，研究中所取名詞數量約為20。

㈡網誌分群模組

網誌分群模組將透過向量空間模式（vector space model）計算文章之間的相似度，透過HAC分群方式將所有網誌文章進行重新分群，並將其分群結果提供網誌作者瀏覽，以作為其建立新類別之依據。其步驟如下：

1. 計算文章之相似度：利用空間向量計算文章之相似度（Salton, Wong, & Yang, 1975），如公式(2)所示。

$$Sim(D_1, D_2) = \cos(\overrightarrow{D_1}, \overrightarrow{D_2}) = \frac{\sum_{k=1}^{n} w_{1k} \times \sum_{k=1}^{n} w_{1k}}{\sqrt{\sum_{k=1}^{n} w_{1k} \times \sum_{k=1}^{n} w_{1k}}} \quad (2)$$

公式中w_{1k}與w_{2k}分別為字詞k在文件D1和文件D2中的權重（即TF-IDF值）。

2. 進行階層式分群：根據第一步結果，建立一矩陣表示出每篇文章之相似度，如表1。階層式分群為在一開始將每篇文章皆視為一個群，接下來以相似度由高至低進行聚合，如圖2所示。
3. 網誌作者依據分群結果建構新的類別：利用本系統所實做之使用者介面，將網誌原先設定之類別內容以及透過HAC之分群結果呈現給網誌作者，並讓使用者進行新類別的建立。

表1 文章相似度矩陣範例

	D1	D2	D3	D4	D5
D 1	1	0.5	0.4	0.2	0.1
D 2		1	0.7	0.6	0.25
D 3			1	0.15	0.3
D 4				1	0.8
D 5					1

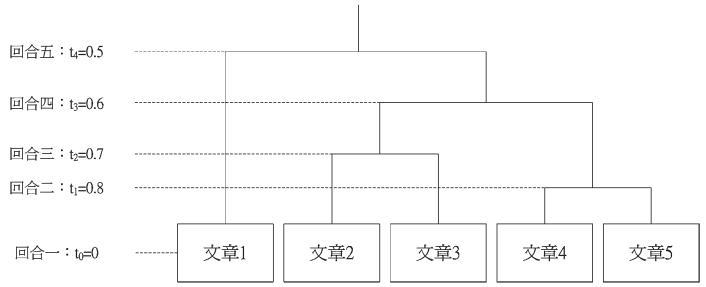

圖2 文件階層式分群之示意圖

㈢概念階層呈現模組

概念階層呈現模組主要將分群結果結合FCA技術呈現各群網誌文件的關鍵詞階層關係。

1. 產生各類別文件群之關鍵詞的關係矩陣：每份網誌文件皆會由文件前處理模組產生代表關鍵字，因此，系統將每份文件與所有文件所組成的字詞集合進行比對，若文件內有對應之詞彙，則標示為「 1 」；若無則為「 0 」，如此可以產生文件群與關鍵詞的二元關係矩陣，如表2所示。FCA的本文內容（ context ）定義為L，文件群定義為G，關鍵詞集合定義為M，文件群與關鍵詞的相互關係則為I，因此以上關係式可表示為：L : = (G , M , I)。
2. 計算概念間之階層關係：當一個關鍵詞kw1的文件集合皆可以被另一個關鍵詞kw2的文件集合所完全包含時，則稱此關鍵詞kw1為關鍵詞kw2的子屬性。
3. 產生整體的關鍵詞網狀架構關聯：同一個概念可能有許多的父概念，一個概念也可能有許多的子概念，因此本研究藉由計算不同概念間之關係以求得整體的關鍵詞間的關聯，最後建構出群集之網狀架構關聯圖。
4. 各類別概念關聯階層的瀏覽：利用本系統所提供之使用者介面，讓使用者可以自行選擇及瀏覽網誌各類別之概念關聯階層。在選擇欲瀏覽之類別後，建立其類別之概念階層以圖形化方式加以呈現，如圖3所示並說明如下：

・灰色的文字框（ 例如：生物 ）是該群組內的關鍵詞，亦為FCA中的屬性。

・白色的文字框（例如：公雞）是該群組內的文章標題，亦為FCA中的物件。

若兩個不同的灰色的文字框（關鍵詞）之間有線段連接，表示文字框之間有繼承關係，線段下方的關鍵詞會繼承線段上方關鍵詞的屬性。例如：關鍵詞 生物 繼承關鍵詞 。雙色端點（◒）代表的是關鍵詞的末端節點，在文章標題（白色文字框）下方的關鍵詞（灰色文字框）也就是表示該文章其它的關鍵詞。

表2 文件與關鍵詞之二元關係矩陣範例

	kw 1	kw 2	kw 3	Kw 4	kw 5	…	kw i
D 1	1	0	1	0	1		0
D 2	1	0	1	1	0		1
D 3	0	0	1	0	1		1
⋮	⋮	⋮	⋮	⋮	⋮		⋮
D n	1	1	1	0	0		1

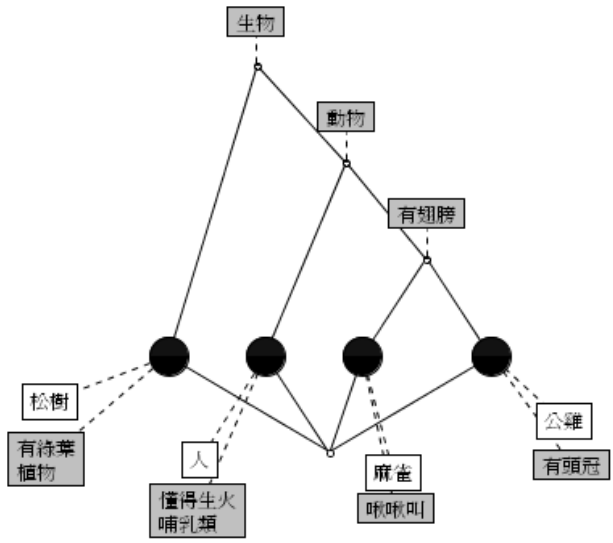

圖3 概念階層呈現範例

伍、實驗設計與討論

一、實驗流程

研究目的為協助網誌作者重新組織分類並透過瀏覽所建構的概念關係圖以瞭解所有網誌文章的脈絡。實驗採用的資料為目前廣受歡迎的無名小站網誌文章，並依照研究目的依序進行兩個部份的實驗。圖4為實驗流程說明並輔以一網誌實例說明實驗步驟。

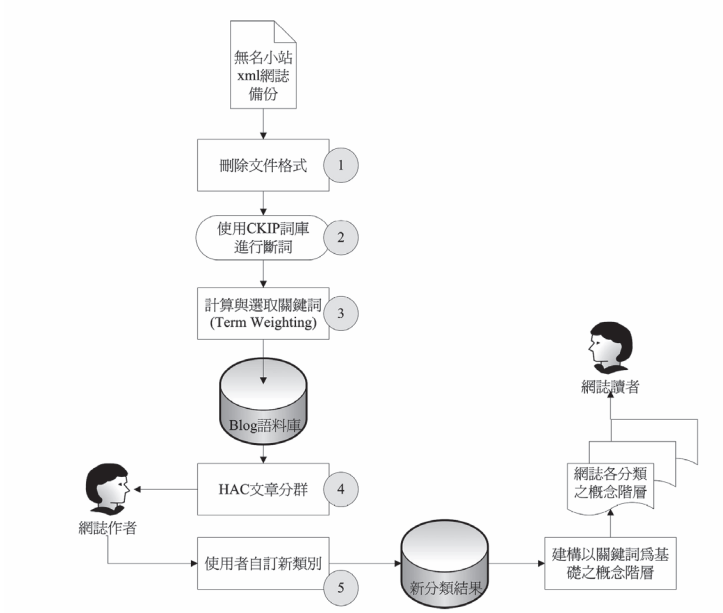

圖4 實驗流程

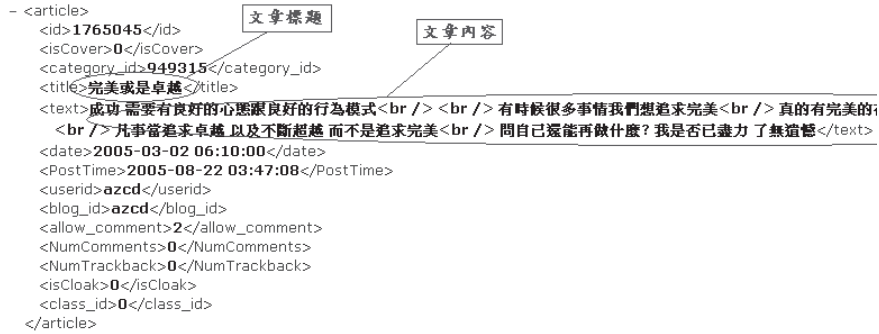

圖5 網誌備份XML檔案之文章標題及內容

㈠刪除文章格式

下載無名網站之網誌備份，由於其檔案格式為XML（如圖5所示），須移除不必要的標籤並取出文章的標題與內容，將每篇網誌文章以文字檔案儲存。

㈡使用CKIP詞庫進行斷詞

將儲存後的網誌文章修改至符合CKIP斷詞服務之格式，透過CKIP網路服務進行文章內容的斷詞，並將傳回之斷詞結果存入資料庫中（如圖6所示）。在進行關鍵詞權重計算前，系統採用停用表（stop word list）過濾不適用的關鍵詞，如：「我」、「我們」等。

art_id	art_t	term_n	term	term_f
11	完美或是卓越	1	完美	Vi
11	完美或是卓越	2	或是	C
11	完美或是卓越	3	卓越	Vi
11	完美或是卓越	4	成功	Vi
11	完美或是卓越	5	需要	Vt
11	完美或是卓越	6	有	Vt
11	完美或是卓越	7	良好	Vi
11	完美或是卓越	8	的	T
11	完美或是卓越	9	心態	N
11	完美或是卓越	10	跟	P
11	完美或是卓越	11	良好	Vi
11	完美或是卓越	12	的	T
11	完美或是卓越	13	行爲	N
11	完美或是卓越	14	模式	N
11	完美或是卓越	15	有時候	ADV
11	完美或是卓越	16	很多	DET
11	完美或是卓越	17	事情	N
11	完美或是卓越	18	我們	N

圖6 網誌文章之詞與詞性

㈢計算與選取關鍵詞

本研究以名詞（N）為研究對象，計算每篇文章中之名詞權重值，由高至低進行排列每篇文章之名詞的權重，取權重較高的前20個名詞作為代表該篇文章之關鍵詞。

㈣HAC文章分群

首先使用者決定分群群數，系統採用向量空間模式計算網誌文章之相似度，計算公式如公式(2)。研究於進行HAC演算法時根據計算時間複雜度以及分群結果品質的考量之下，採用group average方法計算群與群之間的相似度（inter-cluster similarity），主要為評估群聚距離。假設C表示群集合，n為群數目，C={*c1, c2,...,cn*}，若為未經改良之分群，*n*表示為

所有網誌群及數目。在group average方法中，兩群c_i與c_j之間的相似度定義為c_i所包含的所有文件與c_i所涵蓋所有文件，把文件與文件間的相似度作一對一的加總平均，以代表兩群的相似度，如公式(3)所示：

$$sim(c_i, c_j) = \frac{1}{|c_i||c_j|} \sum_{dk \in ci, dl \in cj} sim(d_k, d_l) \quad (3)$$

㈤使用者訂定新類別

網誌文章經由HAC技術完成文章分群後，將結果呈現給網誌作者並請網誌作者建立更適合網誌的新分類，系統介面如圖7所示。圖中左邊方框為分群結果，右邊為作者之原本分類。使用者可以根據分群結果建立新的類別群組、修正原本的分類文章或者刪除群組。此步驟在完成所有網誌類別命名後，將於下一步驟建立每個類別的概念階層。

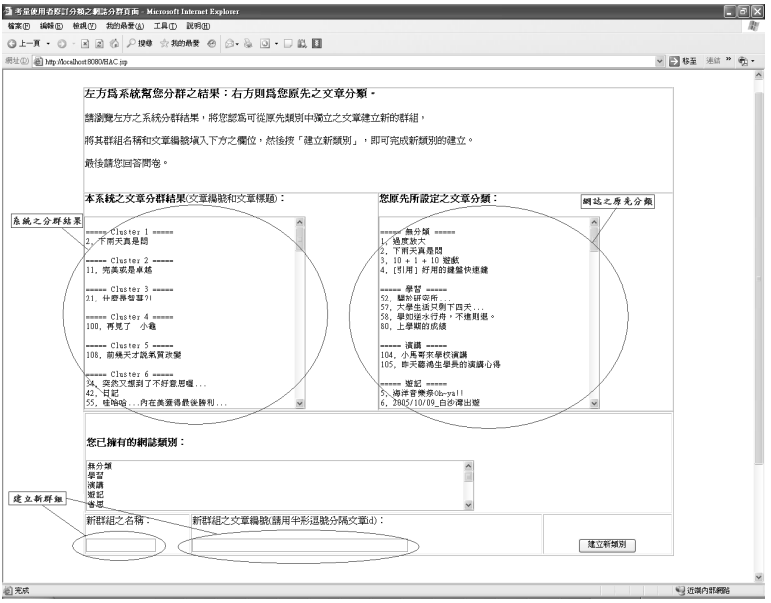

圖7 使用者設定新類別頁面

㈥建構以關鍵詞爲基礎之概念階層

此步驟將依照使用者在步驟㈤所訂定之網誌類別，分別建構每個類別之概念階層。首先要建立每個類別所包含之文章、名詞之二維矩陣，接下來利用FCA法對關鍵詞（概念）進行階層關係的分析。圖8即為概念階層建構之實例，定義如第三部份之概念階層呈現模組所描述，灰色背

景之文字框內容為關鍵字詞；白色背景之文字框內容為文章之標題。關鍵詞（概念）之間之階層關係乃由上而下依序建立，在線段下方之關鍵詞表示其繼承線段上方之關鍵詞（概念）。

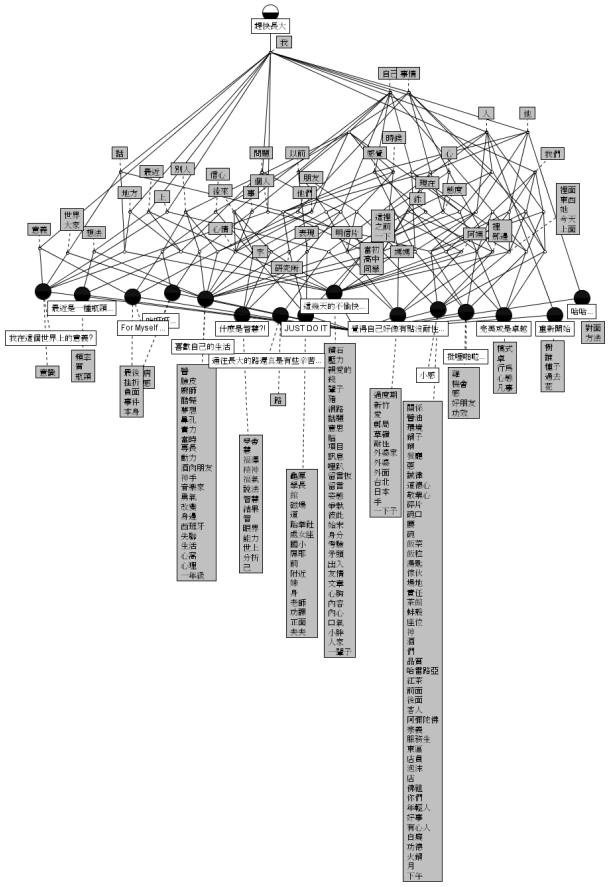

圖8 概念階層建構實例

二、實驗評估

實驗的評估對象分為兩部份，分別是「網誌作者」與「網誌讀者」。對網誌作者進行「文件分群」以及「概念關聯呈現」兩個部份，進行評估；而在網誌讀者則是透過使用者搜尋任務設計（task-oriented approach）以評估方法之有效性（Vegas, Crestani, & Fuente, 2007; White, Jose, & Ruthven, 2003）。

㈠資料樣本

我們針對7份無名網誌資料進行實驗分析，此7份網誌文章篇數平均108篇（最少的網誌資料為48篇），並且從建立開始迄今平均撰寫月數為25個月（最少的期間為5個月），並且大多仍有網誌資料在陸續撰寫，網誌內容多為生活記事或是省思與討論等等，網誌作者主要為資訊相關科系之學生或畢業生。

㈡評估方式

實驗評估方式將進行半結構化式的評估（包含問卷問項的部份以及訪問分群以及概念呈現的使用心得），並將問項分為分群前之調查（Q1以及Q2）、文件分群（Q3至Q5）、建構概念關聯圖輔助作者進行文章之新分類以及協助瀏覽（Q6至Q9）等方面進行。問卷主要針對文件分群以及概念關聯呈現的有用性、可接受性、對網誌作者進行以下問題的5等級（degree）之李科特量表（Likert scale）評等，如附錄A所示。

三、針對網誌作者之評估與觀察

本實驗針對網誌作者的評估分為兩個部份，分別是問卷量化的結果，用以衡量網誌作者對於文章分群以及概念關聯呈現的滿意度；以及對網誌作者進行非量化的訪問，瞭解作者在系統使用以及瀏覽上的心得與意見。

㈠量表結果

我們將問卷的尺度「非常不同意」、「略不同意」、……、「非常同意」化為數字「1」、「2」、……、「5」。並將7位網誌作者（樣本）之評估結果計算平均值和標準差，將其統計結果呈現如表3所示。此外，7個樣本之網誌資訊，如表4所示。

㈡分群結果之觀察

觀察一：網誌作者能夠瞭解某些分群的結果，且有程度不等的準確度

經過分群後，樣本1與樣本2之作者皆發現原本不存在之高度相關的群組，樣本1從分群結果中發現第15群是「演講」高度相關的群組，以及「省思」相關之文章群組，如圖9與圖10所示。圖12的樣本3新發現了「3C產品」高度相關的群組，該網誌之作者表示此群組的文章分類完全正確，也沒有該分至此處的文章散落在別的群組。原先並未建立任何網誌類別的樣本6，經過分群後發現了「職場」、「棒球」高度相關的群組，但是分別還是有2篇以及7篇文章散落到別的文章群組；此外並發現與「吃」相關之群組。

表3 網誌作者對分群以及概念呈現的評估統計

問項	平均值	標準差
分群進行前之調查		
Q1：您覺得目前的blog需要更妥善的分類？	4.00	1.15
Q2：您覺得靠自己將blog文章進行分類會是一件很花費時間的工作？	4.29	1.50
分群結果之評估		
Q3：您覺得此分群結果是否能為您所接受？	3.14	1.21
Q4：您覺得此分群有助於節省您進行分類的時間？	3.57	0.98
Q5：您覺得瀏覽分群的結果有助於您建立新的類別？	4.00	1.00
概念關聯階層之評估		
Q6：您覺得概念關聯階層所呈現的意義能夠被你所接受？	4.14	0.69
Q7：您覺得瀏覽概念關聯階層有助於建立新的類別？	4.00	1.00
Q8：您覺得概念關聯階層式否有助於網誌的瀏覽？	4.43	0.79
Q9：您覺得概念關聯階層式否有助於網誌新類別的命名？	4.00	0.58

觀察二：部份群組包含多個討論主題

群組的聚合是透過計算文章相似度，將具有相似內容之文章群集起來，但由於該法無法真正理解字詞對人的意義，將產生比較不易為人所理解的群組，所以只要群組間的內容相似度夠高，便會將兩個不同的群組加以合併。例如：下圖新產生的群集10由「遊記」及「影劇」兩個特性之群組所聚合而成，由於該群組只有兩個特性，故還可以從中辨識，並且給予其意義，例如圖13所示為將此文章群命名為「娛樂」群組。但倘若一個群組聚集了更多特性，有可能是涵蓋多個領域的特殊議題，須由作者自行判斷會被合併之原因以進行群組命名。

觀察三：群集的鏈結效應問題尚未完全解決

研究採用group average計算分群相似度以降低鏈結效應，但實際分群結果仍有樣本產生一個大群的群組以及落單之單篇文章所形成的群組的結果。整體而言，除了樣本4因為文章多屬於抒情散文，導致分群結果無法正確的表達作者當初的分類並產生「鏈結效應」外，其它樣本在瀏覽過HAC分群結果後皆建立出正確與新的類別，結果列於表4。

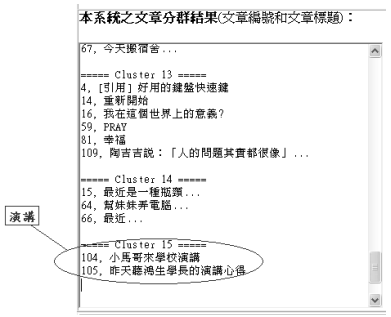

圖9 樣本1的「演講」群組

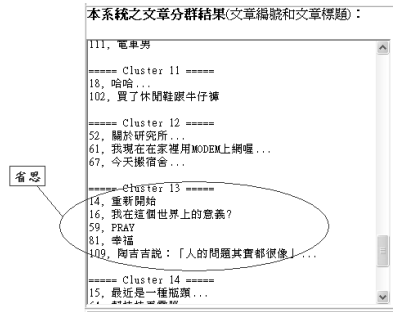

圖10 樣本1的「省思」群組

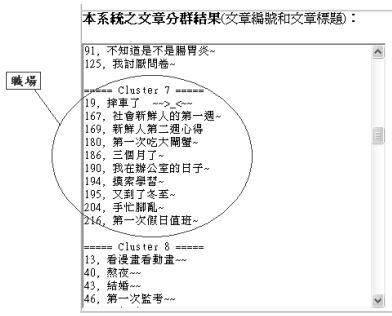

圖11 樣本6的「職場」群組

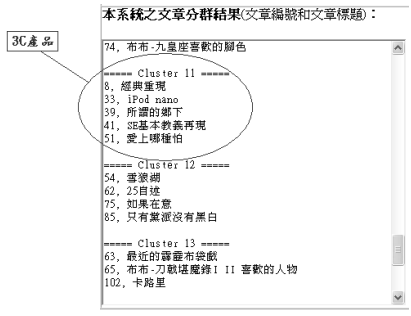

圖12 樣本3的「3C產品」群組

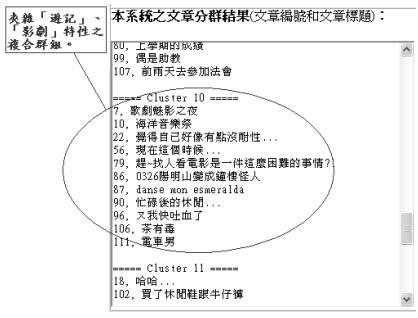

圖 13 多個主題之複合特性群組

表4 以HAC分群協助網誌作者進行分類之結果

網誌樣本	文章篇數	原先網誌分類	瀏覽HAC分群後所建立之新分類
樣本1	116	遊記、省思、叉燒飯、日記、所見所聽、雜類、冷笑話	學習、演講
樣本2	155	運動、Funny、遊記、知識、隨筆、心得、日記	歌詞、異想、電視電影感想、書籍感想
樣本3	110	過日子、記日子、鬼日子、早上八點你在做什麼、好東西、鳥東西、鬼東西、這樣過著她的生活、說話、廢話、概念、店、人、事、做	3C產品、布袋戲、出遊
樣本4	49	我的人、我的事、我的話、股票、論文、茶葉、霹靂、保齡、3C、閱讀	
樣本5	58	燒錄回憶、腦袋監視器、硬碟備份、音樂吸滴	研究所、竹友會
樣本6	217		棒球、吃的、職場、研究所學習、玩樂
樣本7	48		感情、觀影心得、流水帳日記、生活隨筆、藝術心得

㈢概念關聯之觀察

研究應用FCA建構出每個類別的概念（關鍵詞）階層圖，期望能透過概念圖的方式來呈現網誌各分類的內容，讓網誌的讀者能夠更快瞭解該網誌各分類所談論之內容，以及協助網誌作者透過該方法，將過大群的文章群體分出新的分類以解決「鏈結效應」之問題。

觀察一：網誌作者能夠瞭解概念關聯階層所呈現的意義並有助於網誌的瀏覽

在瀏覽各類別之概念階層圖後，除了樣本4以外的網誌作者認為概念階層關聯意義為「普通」，其他受試者皆認為概念圖的階層關聯能夠為網誌作者所理解。而樣本4不能夠理解概念階層之關聯的原因有三：(1)樣本4文章數目較少且多屬於心情記事的文章導致第一階段的分群結果產生了「鏈結效應」，因此，後續的概念圖亦無法適切的表達群組的意義；(2)樣本4網誌所撰寫的內容多為抒情類的文章，然而關於心情方面的名詞即使化成概念圖呈現，也較難看出有意義的內容，故選擇「普通」。整體而言，大部份的網誌作者都同意透過建構不同類別的概念圖能夠協助網誌作者對該類別內容進行瀏覽。

觀察二：網誌作者認為概念關聯階層有助於網誌新類別的命名

大部份網誌作者皆同意能夠透過瀏覽該類別的概念圖幫助使用者進行網誌新類別的命名。在樣本1中，有一包含很多文章的類別群組（72篇文章），名為「日記」。作者表示本來覺得日記這個分類的文章數目雖然遠遠超過其他幾個分類，但由於文章數目太多也不知從何分起，但在瀏覽過該類別的概念階層之後，便發現可以將該類別中抽出一些文章成立新類別，例如「研究所」、「籃球」、「我與朋友」，等新類別。樣本2的「運動」類別，在瀏覽概念階層之後又可再分為「洋基」等類別，如圖14所示。樣本7原本所建立的「藝術心得」類別，在瀏覽過概念關聯圖後便將類別名稱改為「西洋藝術史」，更貼近該類別的文章內容，如圖15所示。

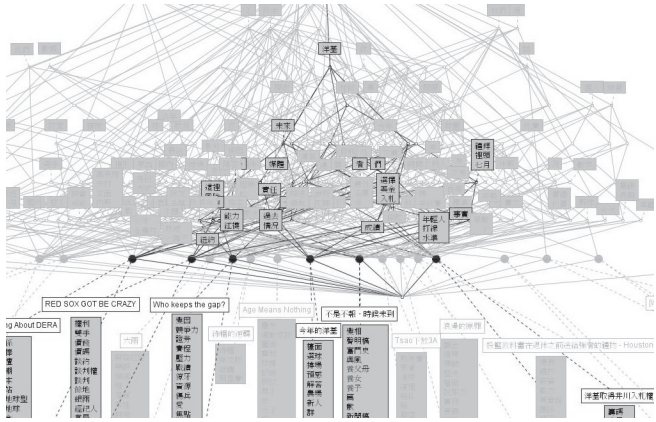

圖14 瀏覽概念階層建立新類別實例1

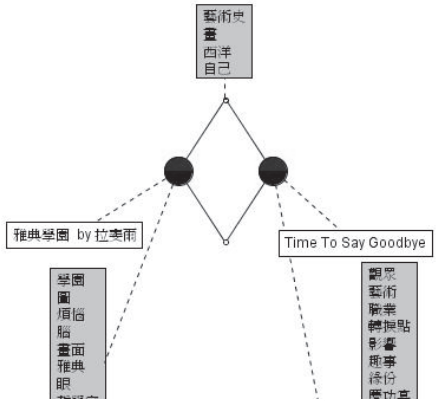

圖15 瀏覽概念階層協助類別命名實例2

四、針對網誌讀者之評估

為了評估本研究以HAC階層式分群以及利用FCA繪製概念圖呈現網誌的可行性。此部份從網誌讀者的角度進行評估，以使用者搜尋任務設計的方式設計能夠模擬真實情境中不同資訊需求之搜尋任務，主要針對內容正確性、易理解性、有用性等方面進行評估。

㈠網誌搜尋任務設計

本研究企圖模擬真實的資訊搜尋情況，進而設計四種類型的搜尋任務問題。此四類問題分別反映搜尋上的不同的資訊需求，這些問題的類型簡述如下：(1)找尋一件事實（search for a fact），例如：尋找某某人物之居住地；(2)找尋幾個項目（search for a number of items），例如：尋找作者在大學時所修過的三門課；(3)決策搜尋（decision search），例如：尋找作者最有可能偏好哪一類之書籍？(4)背景搜尋（background search），例如：尋找有關網誌中提到的某人物之相關資訊。受測前將提供使用者進行搜尋任務的情境（scenario），讓使用者充分瞭解他們需要什麼樣的資訊以及為何需要該資訊。此外為了避免讓讀者找到錯誤答案，問題的設計則以簡單、清楚、避免模稜兩可為考量。以下為搜尋任務之問題：

(1) 假設你是一位參加演唱會的愛好者，你得知1號網誌的作者也有看演唱會的經驗，本搜尋任務是請你從他的網誌中找出他曾去過哪位明星的演唱會。

(2) 假設2號網誌的作者是你最近新結交的朋友，你對他很有興趣，所以想知道他平日都和誰在來往，此搜尋任務要請你從他的網誌中搜尋出任兩位他的朋友之姓名或是暱稱。

(3) 樣本7是一個運動的愛好者，在他的網誌裡常會談論到他的興趣以及平日的休閒活動等資訊，本搜尋任務是請你從他的網誌中找出他可能最喜歡的運動。

(4) 假設你在網路上無意間瀏覽到樣本7的網誌，你發覺這個人很有趣所以想跟他作朋友，此搜尋任務要請你找出與他相關的資訊（例如找出他隸屬於哪個單位做事，職稱為何等等）。

㈡評估對象與結果

評估對象為10位主修資訊相關科系以及平日有撰寫或瀏覽網誌習慣之學生。評估進行約50分鐘，10位受測者都將使用兩種系統分別針對2項搜尋任務進行搜尋，總共進行4項搜尋任務。每項搜尋任務時間上限為10分鐘。為了降低受測者對於使用同樣系統進行搜尋任務所產生的學習效

應，故在整個評估過程的中間將兩組人員交換，利用不同的方式進行搜尋任務。整個實驗過程的搜尋任務與人員分配如表5。

表5 搜尋任務與其利用之系統一覽

	搜尋任務1~2	搜尋任務3~4
實驗A組（樣本一~樣本五）	本研究所提出之概念圖瀏覽方法進行搜尋	利用作者原先設定之類別進行搜尋
實驗B組（樣本六~樣本十）	利用作者原先設定之類別進行搜尋	本研究所提出之概念圖瀏覽方法進行搜尋

㈢任務蒐尋結果與訪談

表6為網誌讀者進行搜尋任務所費時間之統計結果，其中之數據以秒為單位。以灰色為背景的數值乃是代表該樣本進行該問項是透過本系統所提出之FCA建構各類別概念圖方式協助瀏覽的方式。其中除了樣本九沒有找到問項一之答案以外，其他樣本皆能夠正確完成搜尋任務，透過上表我們可以看出透過FCA建構各類別概念圖，提供讀者瀏覽之搜尋方式，相較於網誌作者本身所提供之分類，更能節省時間。

表6 進行搜尋任務所花費之時間統計

		問項一	問項二	個人平均	個人標準差	問項三	問項四	個人平均	個人標準差
	樣本一	33	59	46	18	95	148	122	37
	樣本二	35	139	87	74	158	92	125	47
	樣本三	60	180	120	85	148	180	164	23
A組	樣本四	139	60	100	56	102	447	275	244
	樣本五	102	105	104	2	205	125	165	57
	組平均	74	109			142	198		
	組標準差	46	52			45	143		

		問項一	問項二	個人平均	個人標準差	問項三	問項四	個人平均	個人標準差
	樣本六	270	127	199	101	40	60	50	14
	樣本七	167	91	129	54	40	117	79	54
	樣本八	320	119	220	142	60	100	80	28
B組	樣本九	600	58	329	383	85	130	108	32
	樣本十	27	171	99	102	32	51	42	13
	組平均	277	113			51	92		
	組標準差	213	42			21	35		

在以上各項搜尋任務結束後，皆再對受測者進行完成該項任務後之使用觀感評估。評估的方式採用五維度之李科特量表方式，分為「非常不同意」、「略不同意」、「普通」、「略同意」、「非常同意」五個選項，分別用數字1~5代替。而其問項為：(1)你覺得你已正確完成本搜尋任務？(2)你覺得此次的搜尋任務難度很高？(3)你覺得進行本次搜尋任務時，系統能夠協助你進行資料的搜尋？(4)你覺得在此次搜尋任務中，系統能夠提供你更多與問題相關之資訊？

以下將評估結果與使用者的心得列點如下：

(1) 透過各類別建構FCA關鍵詞概念圖的瀏覽，有助於協助讀者對於該類別重要概念的掌握。

(2) 透過各類別所繪製出的概念圖呈現方式，可以一次將整個類別所有文章的相關關鍵詞列於同一個畫面，可以節省讀者往返瀏覽、搜尋的時間。

(3) 若使用者不知道其欲查詢資訊之相關關鍵詞時，透過FCA所建構出之網誌各類別的關鍵詞概念圖，更能明顯協助讀者進行資料的搜尋。

(4) 雖然透過關鍵詞概念圖的方式可以瞭解該類別所談及之關鍵詞，但有些關鍵詞之間的關係依然無法單從概念圖上瞭解，必需閱讀整篇文章之後才能瞭解其關聯。

五、綜合討論

㈠利用HAC文件分群協助作者建立新分類

(1) 大部份網誌作者同意利用HAC文件分群方式可以協助新類別的建立，以及節省分類的時間。

(2) 由於網誌作者不易從大群的文章群中找出具有特色的子群組，故對於協助作者建立新分類的目的而言，網誌作者會偏好文章篇數少、代表性高、分群數目多的分群方式，較容易看出整個網誌的各個主題。

㈡利用FCA建構概念關聯呈現網誌

(1) 大部份網誌作者同意利用概念圖可以協助呈現該網誌各類別之內容，以及協助該類別的重新命名。

(2) 利用FCA建構概念圖呈現網誌內容，在分析專門領域的類別時（舉例如「棒球」）能夠發揮最好的效果。而在分析有關抒發心情的類別時，FCA的效果便會降低，因為抒發心情的類別含有較多重要的形容詞，沒有伴隨形容詞的名詞有時候無法看出其代表的意義，再者描寫心情相關的名詞具有比較抽象的特性，即便是以概念圖的方式將其呈現，也不易看出意義。

(3) 網誌的撰寫風格也會影響FCA的準確性，例如本例可以從「棒球」類別中發現「投手」這個關鍵詞屬性可以包含「王建民」、「郭泓志」、「曹錦輝」之類的關鍵詞，但發現網誌作者在撰寫關於這些球星之文章時，不一定會提及「投手」這個關鍵詞，便會導致FCA進行分析時，無法發現這些球星與投手之間的關聯。

(4) 網誌讀者皆同意透過繪製網誌各類別概念圖，提供其瀏覽的方式能夠協助讀者便利的掌握其類別所提及之重要內容，並可節省瀏覽時間。

(5) 有時候讀者在找尋網誌之特定資訊時，不一定知道與其資訊相關之關鍵詞，在此情況下透過FCA建構各類別關鍵詞概念圖的方式便能更有效地協助讀者進行資料的搜尋與瀏覽。

陸、結論與未來研究方向

隨著Web 2.0之知識共享平台概念的興起，網誌的撰寫與交流已成為了一種熱門的知識交換媒介，網誌作者與讀者藉由網誌進行積極的資訊

交流與互動，促進知識的分享。研究主要貢獻為使用文件分群技術以協助網誌作者重新組織分類並協助網誌讀者對所閱讀網誌的理解。另一方面，研究使用FCA技術建構各類別之概念階層以協助網誌作者更準確的替網誌分類命名。

研究評估方法分別由網誌作者與網誌讀者角度分析與探討之。在網誌作者部份，評估結果顯示透過階層式分群結果的瀏覽能夠協助使用者建立新類別，進而協助網誌作者進行新類別的建立；此外，雖然非所有作者皆滿意分群結果，但是，大部份網誌作者皆同意網誌文章的重新分群有助於文章的重新組織整理。研究更進一步發現透過FCA建構概念階層可以有效協助網誌作者對大群組進行分析，因此，可以協助網誌作者對該群組內容進行瀏覽。在網誌讀者的評估部份，本研究嘗試使用使用者搜尋任務設計的評估方式請10位網誌讀者分別進行四項搜尋任務。從量化的評估數據中我們觀察到相較於利用網誌原本的類別進行搜尋，透過FCA繪製網誌各類別關鍵詞概念圖的方式進行搜尋更能降低花費的時間以及降低搜尋難度。本實驗另一主要貢獻為嘗試模擬真實的使用者資訊搜尋情況，進而設計四種類型的搜尋任務問題：㈠找尋一件事實；㈡找尋幾個項目；㈢決策搜尋；與㈣背景搜尋。透過不同的搜尋任務可以更深入分析方法有適用性與有效性。

研究主要以名詞為重要關鍵詞的篩選對象，進而計算文章相似度以及建構概念關聯之呈現。研究結果顯示FCA在分析專門領域的類別時能夠發揮最好的效果；而不善於分析有關「抒發心情」之類別，乃因為「心情」相關的類別若沒有伴隨形容詞將無法表達其意義，再者較抽象的名詞，單看其本身可能會流於空泛，故效果較差。因此，研究所提出之方法較適用於學術性質或探討特定議題之網誌而不適用於心情隨筆之抒情文，此為研究主要之限制。整體而言，透過視覺化方式呈現網誌內容，有助網誌作者與讀者對網誌內容的理解進而達成有效率之瀏覽。但在分群操作介面部份，研究於使用者進行分群評估過程中觀察發現，方便的操作介面亦為使用者對分群結果滿意度的重要因素，因此，未來將持續進行相關測試與介面修正以期設計出更友善之輔助系統。未來研究將嘗試解決文章同義異詞問題（例如：「棒球」與「野球」即為同義異詞），以改善文件分群以及概念關聯呈現的品質。基於以上部份，研究將考量網誌作者透過自訂標籤的方式，輔助判別同義詞以衡量網誌標籤語是否能提升分群精確度。

致謝

本研究經費承國科會專題研究計畫補助（編號：NSC 95-2416-H-030-012 & 97-2410H-030-030-MY2），僅以此致謝。研究並感謝主編與審查委員寶貴的建議。

參考文獻

中文詞知識庫小組（無年代）。**中文句結構樹建立原則**。上網日期：2006年3月1日，檢自：http://godel.iis.sinica.edu.tw/CKIP/treebank.htm

巫啟台（2002）。**文件之關聯資訊萃取及其概念圖自動建構**。未出版之碩士論文，國立成功大學資訊工程研究所，台南市。

吳育龍（2000）。**於網際網路上應用概念圖輔助學習之研究**。未出版之碩士論文，私立中原大學資訊工程研究所，中壢市。

林柏成（2002）。**利用知識地圖診斷數學問題之研究**。未出版之碩士論文，私立中原大學資訊工程研究所，中壢市。

馬偉雲（無年代）。**未知詞擷取作法**。上網日期：2006年3月1日，檢自：http://ckipsvr.iis.sinica.edu.tw/uwe.htm

高宜敏（2001）。**合作競爭式的網路分享建構學習環境**。未出版之碩士論文，國立交通大學資訊科學系，新竹市。

陳年興、謝盛文、黃琬婷（2007，1月）。自動化建構具時間向度之知識結構映射圖——以資訊管理領域之知識及其演進為例。**資訊管理學報，14**(1)，1-32。

陳年興、孫振凱（2002）。透過網頁分析建構知識分佈圖以輔助線上教學。在淡江資訊管理學系編，**第十三屆國際資訊管理學術研討會論文集**（頁675-682）。臺北縣：編者。

陳道輝（2003）。**利用學位論文資訊萃取資訊相關領域之研究主題關聯性**。未出版之碩士論文，國立中山大學資訊管理研究所，高雄市。

蔡宜龍（2002）。**特殊領域文件分群之系統設計與研究——以佛學資料為例**。未出版之碩士論文，國立台灣大學資訊工程學研究所，台北市。

Brooks, C. H., & Montanez, N. (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. *WWW '06: Proceedings of the 15th International Conference on World Wide Web* (pp. 625-632). New York: ACM Press.

Chen, K. J., & Liu, S. H. (1992). Word identification for mandarin Chinese sentences. *COLING '92: Proceedings of the 14th International Conference on Computational Linguistics* (pp. 101-107). Nantes: France.

Chen, K. J., & Bai, M. H. (1998). Unknown word detection for Chinese by a corpus-based learning method. *International Journal of Computational linguistics and Chinese Language Processing, 3*(1), 27-44.

Chen, N. S., Kinshuk, Wei, C. W., & Chen, H. J. (2008). Mining e-Learning domain concept map from academic articles. *Computers & Education, 50*(3), 1009-1021.

Chuang, S. L., & Chien, L. F. (2004). A practical web-based approach to generating topic hierarchy for text segments. *CIKM '04: Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management* (pp. 127-136). New York: ACM Press.

Divitini, M., Haugalokken, O., & Morken, E. M. (2005). Blog to support learning in the field: Lessons learned from a fiasco. *ICALT '05: Proceedings of the 5th IEEE International Conference on Advanced Learning Technology* (pp. 219-221). Kaohsiung, Taiwan: IEEE Computer Society.

Everts, T. J., Park, S. S., & Kang, B. H. (2006). Using formal concept analysis with an incremental knowledge acquisition system for web document management. *ACSC '06: Proceedings of the 29th Australasian Computer Science Conference* (pp. 247-256). Hobart: ACM Press.

Ganter, B., & Wille, R. (1999). *Formal concept analysis: Mathematical foundations*. Berlin Heidelberg: Springer-Verlag.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann Publisher.

Hill, M. D., Gaudiot, J. L., Hall, M., Marks, J., Prinetto, P., & Baglio, D. (2006). A wiki for discussing and promoting best practices in research. *Communications of the ACM, 49*(9), 63-64.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264-323.

Jonassen, D. H., Reeves, T. C., Hong, N., Harvey, D., & Peters, K. (1997). Concept mapping as cognitive learning and assessment tools. *Journal of*

Interactive Learning Research, 8(3-4), 289-308.

Liu, D. L., Wu, I. C., & Chen, W. H. (2006). Mining and supporting task-stage knowledge: A hierarchical clustering technique. *Lecture Notes in Artificial Intelligence, 4333*, 178-188, Springer Berlin/Heidelberg.

McAfee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration. *MIT Sloan Management Review, 47*(3), 21-28.

McAleese, R. (1994). A theoretical view on concept mapping. *Association for Learning Technology Journal, 2*(1), 38-48.

McAleese, R. (2000). Concept mapping: A critical review. *Innovations in Education and Training International, 36*(4), 351-360.

Rosenbloom, A. (2004). The Blogosphere. *Communications of the ACM, 47*(12), 31-33.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513-523.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Vegas, J., Crestani, F., & Fuente, P. D. L. (2007). Context representation for Web search results. *Journal of Information Science, 33*(1), 77-94.

White, R. W., Jose, J. M., & Ruthven, I. G. (2003). A task-oriented study on the influencing effects of query-biased summarization in Web searching. *Information Processing and Management, 39*(5), 707-733.

Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival (Ed.), Ordered Sets (pp. 445-470). Dordrecht-Boston: D. Reidel Publishing Company.

Yao, J. (2006). Supporting research with weblogs: A study on web-based research support systems. *WI-IATW '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 161-164). Washington, D.C.: IEEE Computer Society.

附錄A

半結構化問卷之問項

請您在分群進行之前回答以下問題	**非常不同意**	**略不同意**	**普通**	**略同意**	**非常同意**
Q1：您覺得目前的blog需要更妥善的分類？					
Q2：您覺得靠自己將blog文章進行分類會是一件很花費時間的工作？					
請您瀏覽分群後之結果，並回答以下問題	**非常不同意**	**略不同意**	**普通**	**略同意**	**非常同意**
Q3：您覺得此分群結果是否能為你所接受？					
Q4：您覺得此分群有助於節省您進行分類的時間？					
Q5：您覺得瀏覽分群的結果有助於您建立新的類別？					
請您瀏覽概念關聯階層後回答以下問題	**非常不同意**	**略不同意**	**普通**	**略同意**	**非常同意**
Q6：您覺得概念關聯階層所呈現的意義能夠被你所接受？					
Q7：您覺得瀏覽概念關聯階層有助於建立新的類別？					
Q8：您覺得概念關聯階層式否有助於網誌的瀏覽？					
Q9：您覺得概念關聯階層式否有助於網誌新類別的命名？					

註：我們將問卷的尺度「非常不同意」、「略不同意」、……、「非常同意」化爲數字「1」、「2」、……、「5」。

附錄B

問卷評估結果

Q1：您覺得目前的blog需要更妥善的分類？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	3	5	4	4	2	5	5
Q2：您覺得靠自己將blog文章進行分類會是一件很花費時間的工作？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	4	5	5	5	1	5	5
Q3：您覺得此分群結果是否能為您所接受？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	3	2	4	2	2	4	5
Q4：您覺得此分群有助於節省您進行分類的時間？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	4	3	4	2	3	5	4
Q5：您覺得瀏覽分群的結果有助於您建立新的類別？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	4	4	4	2	4	5	5
Q6：您覺得概念關聯階層所呈現的意義能夠被你所接受？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	4	4	4	3	5	5	4
Q7：您覺得瀏覽概念關聯階層有助於建立新的類別？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	5	5	4	2	4	4	4
Q8：您覺得概念關聯階層是否有助於網誌的瀏覽？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	4	5	5	4	5	3	5
Q9：您覺得概念關聯階層是否有助於網誌新類別的命名？							
	樣本1	樣本2	樣本3	樣本4	樣本5	樣本6	樣本7
評比	4	5	4	3	4	4	4