國立交通大學

資訊管理研究所

碩士論文

應用社會性推薦於學術社群

Using Social Recommendation in Academic Community

研究生：粘怡祥

指導教授：柯皓仁

中華民國 九十七 年 七 月

應用社會性推薦於學術社群

Using Social Recommendation in Academic Community

研 究 生：粘 怡 祥　　Student:Yi-Hsiang Nien
指導教授：柯 皓 仁　　Advisor: Dr. Hao-Ren Ke

國 立 交 通 大 學
資 訊 管 理 研 究 所
碩 士 論 文

A Thesis
Submitted to Institute of Information Management
College of Management
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
In
Information Management

June 2008

Hsinchu, Taiwan, the Republic of China

中華民國　九十七 年 七 月

應用社會性推薦於學術社群

指導教授：柯 皓 仁　　研究生：粘 怡 祥

國立交通大學資訊管理研究所

摘要

網際網路提供一個開放的平台，利用網路取得資訊已經成為最方便的管道。面對網路上充斥的大量資訊，使用者在找尋資訊時，相當的費時也不容易聚焦。推薦系統成為改善資訊過載問題的方法之一。使用者除了本身的主觀喜好之外，其行為容易受到人際關係的影響，於是虛擬社群與社會網路，成為許多使用者獲得資訊情報的最佳來源。

本研究主要的目在於提出結合主題概念萃取與社會網路分析之資訊推薦系統，以提供符合使用者需求之推薦資訊。本研究利用關鍵字分群演算法，萃取出使用者感興趣的主題概念；並且分析使用者社會網路，進行使用者分群，以形成主題社群；經由分析社群成員的主題偏好，預測使用者的潛在興趣，建構出更符合使用者需求的資訊推薦系統，以提升資訊推薦的品質。

關鍵詞：分群演算法、社會網路分析、推薦系統、機構典藏、學術社群

Using Social Recommendation in Academic Community

Advisor: Dr. Hao-Ren Ke Student: Yi-Hsiang Nien

Institute of Information Management

National Chiao Tung University

Abstract

Internet provides an open platform, which becomes a convenient channel to obtain information. Facing huge amount of information, users will spend a lot of time and become out of focus when they search information on Internet. Recommendation systems become one solution for information overloading. Besides the subjective preference of a user, interpersonal relationship will affect his/her behavior, and the concepts of virtual community and social network become one feasible information source for deriving interpersonal relationship.

This thesis proposes a recommendation system combining topic concept extraction and social network analysis to meet users' needs. This thesis uses the keyword clustering algorithm to extract topic concepts that users are interested in, followed by the formation of topic communities by analyzing the users' social network to cluster the users. By analyzing the preferences of community members, the system can predict the potential interests and improve the quality of recommendation.

Keyword：Clustering, Social Network Analysis, Recommendation System, Institutional Repository, Academic Community

誌　　謝

隨著畢業口試的結束，兩年的研究所生涯也到達了尾聲，一路走來，心中除了感謝還是感謝。

首先要感謝指導教授柯皓仁老師諸多指導與協助，引導我一步步地將論文完成，心中的喜悅真是無法以筆墨形容。感謝口試委員謝建成老師與陳光華老師，在口試時所給予的建議，使得本論文能更加嚴謹。

同時也要感謝研究室同伴建穎、有盈、姿婷及揚書，所辦的淑惠與欣欣，博班學長姐曜輝與栩嘉，APC研究室以及網路研究室的各位，不論在研究上或生活上，大家總是一起努力打拚，共同分享喜怒哀樂，你們都是陪伴我一起成長的好夥伴，與大家相處的點點滴滴，都將是我人生中美好的回憶。

最後則是感謝家人與女友雅君的包容與愛護，讓我能無後顧之憂的朝目標前進，有你們的支持才有現在的我！

粘怡祥　謹誌

2008年7月

目　　錄

表 目 錄

圖 目 錄

第一章　緒論

1.1 研究背景與動機

隨著網際網路的普及與資訊技術的發達，網路提供了一個完全開放的資訊平台，讓資訊得以快速地傳播、大量地複製與儲存，利用網路取得資訊已經成為最方便的管道。但是面對網路上充斥的大量資訊，不但內容繁雜而且難以過濾，使用者在找尋資訊時，相當的費時也不容易聚焦，以單純瀏覽的方式，在網路上尋找資訊變得相當沒有效率，這使得人們在網路世界中寸步難行。搜尋引擎(Search Engine)與推薦系統(Recommender System)的出現，成為改善資訊過載(Information Overload)問題的兩大利器。

搜尋引擎是屬於一種資訊檢索(Information Retrieval)的方法，當使用者在資訊需求較為明確的情況下，主動地以查詢的方式，來獲取其需要或是感興趣的資訊，這是目前網路使用者最常使用的工具之一，例如 Google、Yahoo 及百度等搜尋引擎。而在使用者較不清楚本身的資訊需求時，可以透過推薦系統以被動的方式取得資訊；推薦系統主要是利用資訊過濾(Information Filtering)的方法，在使用者與網站的互動過程中，發掘使用者潛在的需求或興趣，用以比對推薦標的，過濾出符合使用者需求的資訊，並進行推薦的動作，例如 Amazon 與 eBay 等電子商務網站，便是推薦系統應用上的佼佼者。

推薦系統已廣泛地運用在各種行銷策略上，更成為許多電子商務網站的核心功能。推薦系統要得到較佳的推薦效果，首要的步驟是收集使用者資訊，不論是外顯資訊的取得，例如使用者在網站上登錄的基本資料，以及對其購買產品或使用過的服務做評比，或是隱性資訊的收集，例如將使用者的歷史交易紀錄，以及瀏覽網站之行為加以分析；在取得使用者的相關資訊後，透過各種推薦方法的運用，進而提供符合使用者需求的產品資訊，以作為使用者購買決策的參考指標。利用有效的推薦資訊，不但可以節省使用者的搜尋時間，更能為企業帶來更多的顧客，提升顧客忠誠度，並且增加企業的收益，這使得個人化的推薦機制成為重

要議題。

目前常應用在推薦系統的方法主要分為內容導向(Content-based)與協同過濾(Collaborative Filtering)二種。內容導向方法主要是針對使用者喜好的項目進行內容分析，經由分析項目的屬性特徵後，進而判斷並找出使用者可能有興趣的項目，再將其結果推薦給使用者。協同過濾方法的主要目的在於發掘與其他使用者間的關聯性，找出具有共同興趣的使用者以形成社群；經由分析社群成員共同的興趣與喜好，以此作為推薦的參考依據，並用以推論或預測目標使用者的潛在偏好。

除了上述所提到的兩種方法，Iskold [1]也以消費者的觀察角度，將推薦系統分類如下：

1. 個人化推薦(Personalized Recommendation)：依據消費者個人過去在網站中的行為進行推薦。
2. 社會性推薦(Social Recommendation)：依據過去和個人有相似行為的消費者來進行推薦。
3. 產品為導向的推薦(Item Recommendation)：依據產品本身的特性進行推薦。
4. 綜合以上三種方法之推薦模式。

根據以上的脈絡，從消費者個人的資料入手，發現不足後，便運用所有消費者資料，繼而整合之前所開發的技術，產生綜合策略；雖然各種分類角度有不同的陳述，但結果卻是殊途同歸。

使用者除了本身的主觀喜好之外，其行為容易受到人際關係的影響。網際網路提供使用者一個高度自主的環境，讓使用者可以輕易地進行資訊分享與訊息交換，使得網路上的人際互動頻繁，無形中形成虛擬社群(Virtual Community)或社會網路(Social Network)等非正式組織，例如部落格(Blog)、社交網路網站(Social Networking Site)與線上遊戲(Online Game)等；於是虛擬社群與社會網路，成為許多使用者獲得資訊情報的最佳來源，Zhong[26]更是明確地指出，虛擬社群與社

會網路提供了從事推薦活動時的基礎資訊。

Staab[32]的研究指出，在市場行銷的戰役中，口耳相傳(Word of Mouth)對消費者行為的改變有重要影響；社會網路的發掘與分析，有助於企業制定病毒行銷(Viral Marketing)策略，在進行產品或服務的推薦活動時，使消費者間口耳相傳的正面效果達到最大化，與忽略消費者互動及網路效應的傳統行銷模式相比，可以使企業取得較高的獲利。

Hotta[10]認為將使用者的社會網路整合到推薦演算法中，可以獲得較佳的推薦效果。在該研究中利用使用者基本資料與瀏覽網站的記錄檔(Log)，建構使用者設定檔(User Profile)，其中包含使用者偏好表(Preference Table)，並且以使用者點擊網頁的內容與使用者社會網路連結來更新偏好表，最後以資訊過濾的方法，產生符合使用者偏好的推薦資訊。

推薦系統有助於使用者獲得所需的資訊，並且能節省搜尋的時間成本；社會網路對使用者決策行為的影響，也成為進行資訊推薦時不可忽視的重要因素。因此本研究將深入探討如何運用社會網路提升資訊推薦的品質。

1.2 研究目的

本研究主要的目在於提出結合主題概念萃取與社會網路分析之資訊推薦系統，並將其應用於學術社群。本研究利用關鍵字分群演算法，萃取出使用者感興趣的主題概念；並且分析使用者社會網路，進行使用者分群，以形成主題社群；經由分析社群成員的興趣偏好，預測使用者的潛在興趣，建構出更符合使用者需求的資訊推薦系統，以提升資訊推薦的品質;整體研究目標如下：

1. 主題概念萃取

萃取出文件中的重要關鍵字，利用關鍵字分群，達到主題概念萃取的目的，藉以瞭解使用者所關注的興趣與議題。

2. 形成主題社群

將個別使用者轉換成以向量空間模型(Vector Space Model)來表示，並結合使用者的社會網路，將相似度高且具有相同主題興趣的使用者群聚在一起，以形成主題社群。

3. 資訊推薦

經由主題社群的產生，針對使用者個人的主題偏好，進行個人化推薦；此外，更進一步分析社群成員共同的興趣與喜好，預測使用者的潛在偏好，以建構出更符合使用者需求的推薦系統，提升資訊推薦的品質。

1.3 研究流程

本研究之研究流程如圖1-1所示，首先闡述本研究之動機與目的，並且收集分群演算法、社會網路分析與推薦系統之相關文獻資料，探討其發展趨勢及有待改善之處；而後針對先前三方面文獻的回顧，勾勒出一整合社會網路分析於推薦系統之資訊推薦方法；並發展雛型系統進行資訊推薦，藉由分析實證結果得到本研究之結論，並提供後續研究之建議方向。

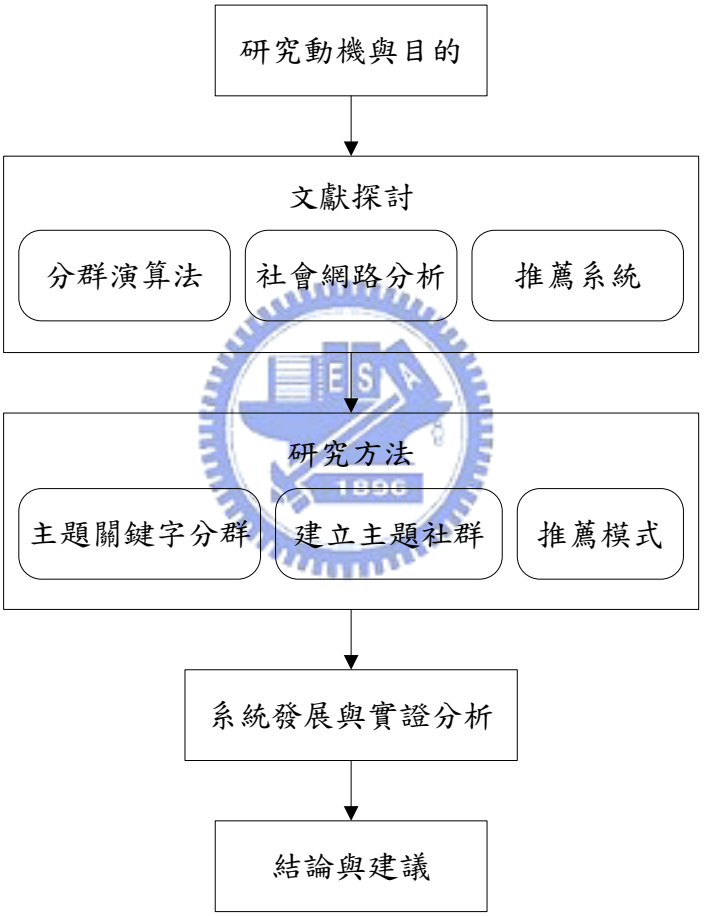

圖 1-1 研究流程圖

1.4 論文架構

本研究共分為五章，說明如下：

第一章 緒論

說明本研究之研究背景與動機、研究目的及研究流程。

第二章 文獻探討

針對分群演算法、社會網路分析及推薦系統三方面與本研究相關之文獻進行探討。

第三章 研究方法

描述本研究之資料前置處理流程、研究方法，並介紹系統之推薦模式。

第四章 系統發展與實證分析

經由雛型系統發展，進行資訊推薦，並就推薦結果進行分析與比較，以得知推薦之效能與適用性。

第五章 結論

總結本研究並提出未來可供後續研究之方向。

第二章 文獻探討

本章從相關文獻之蒐集、整理與分析，來探討分群演算法、社會網路以及推薦方法之相關理論。分群演算法主要針對常用的劃分式(Partitional)與階層式(Hierarchical)兩種方法進行討論[2]，以k-means[14]及Chameleon[8]演算法舉例說明，並且對本研究所採用之主題關鍵字分群法(Topic Keyword Cluster)[9]做進一步討論。社會網路方面乃從其定義與建構社會網路之基本單元做說明，並且列舉三項常用於分析社會網路的指標，以發掘個人在社會網路中所扮演的角色。至於推薦方法的理論探討，則是針對廣泛使用的內容導向(Content-based)與協同過濾(Collaborative Filtering)兩種方法進行闡述，並列舉相關之實際應用作為比較。

2.1 分群演算法

分群演算法在資料探勘(Data Mining)領域上是應用相當普遍的一種技術，主要的目的在於將資料集裡的資料區分成數個群集(Cluster)，使得每一個群集內資料間的相似性高，而不同的群集之間的資料相似性低。分群演算法的選擇取決於資料的類型、分群的目的與應用，根據Jain[2]的歸納，分群演算法主要區分為劃分式(Partitional)分群法與階層式(Hierarchical)分群法兩大類，如圖2-1所示。

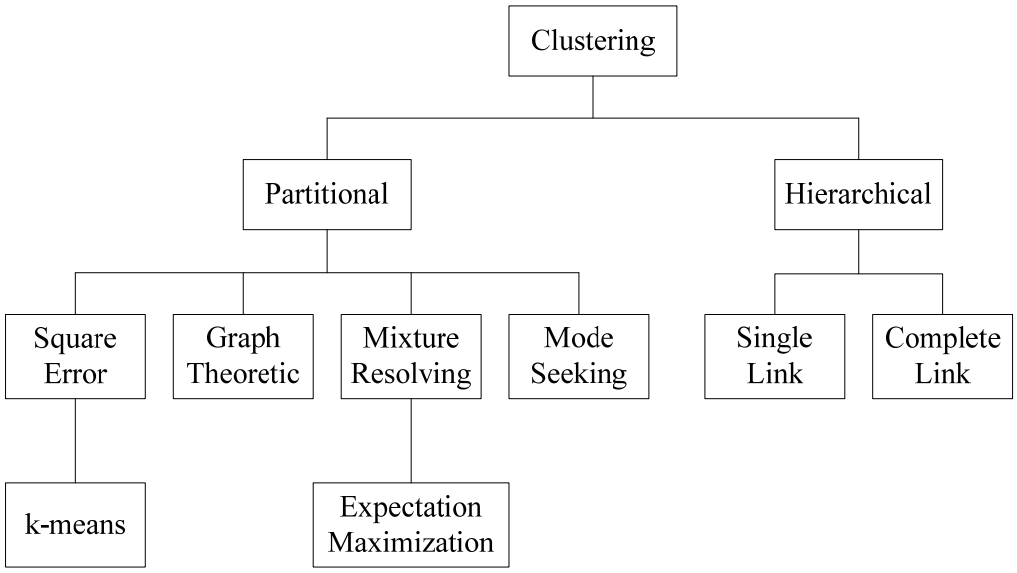

圖 2-1 分群演算法之分類[2]

2.1.1 劃分式分群法

劃分式分群法是分群演算法中最早發展的一種方法。劃分式分群法簡單地來說，是由使用者指定將資料物件分割成k個群集，每一個劃分(Partition)代表一個群集，也就是在分群前必須先定義目標分群的個數，此為劃分式分群法的特色。

k-means是劃分式分群法中最基本卻也是應用最為廣泛的方法，在1967年時由MacQueen所提出[14]，採用歐幾里得(Euclidean Distance)距離作為對資料分群的基礎，距離公式如方程式(2-1)所示。分群過程中必須同時確保兩個條件：

1. 位於相同群集內的物件，彼此間相似度高，群集中心點為所有物件的向量平均值，物件與群集中心點的距離愈小，則表示相似度愈高。
2. 位於不同群集內的物件，彼此間相似度低，即屬於不同群集的物件其距離愈大愈好。

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2\text{-}1)$$

k-Means演算法的詳細步驟如表2-1所示：

表 2-1 k-Means演算法步驟

輸入：n 個物件與分群的數目 k。
輸出：k 個群集。
(1) 任意選擇 k 個物件作為初始的群集中心。
(2) 重複以下步驟，直到群集的分佈不再改變。
(2.1) 利用距離最近者相似度最高的原則，依序計算每個物件與k個群集中心點的距離，將每個物件歸屬到最相似的群集。
(2.2) 計算群組內所有物件間距離的平均值，作為各個群集的中心點。

k-means 在結果群集是密集的、且群集與群集間有明顯區隔時，有相當不錯的成效。若是分群資料中具有雜訊(Noise)或者是離群值(Outlier)時，其分群結果

將會受雜訊的影響而失真。總括而言，k-means 的優點為演算法簡單且計算快速、效率高。缺點是使用群集的算術平均數做為群集的中心，使得群集的結果容易受到雜訊或是離群值所影響而降低其正確性；而且 k-means 只適用於數值型的資料，涉及具有分類屬性的資料則不適用。

2.1.2 階層式分群法

階層式分群法是利用樹狀結構來表示資料分群的結果。每一個群集所代表的節點可以往下再分裂成數個子群，或者同一階層的群集節點也可以往上再聚集成一個更大的群集節點；因此階層式群集演算法的好處，就是可以藉由選擇不同的階層來檢視不同詳細程度的分群結果。階層式分群法一般區分為聚合式(Agglomerative)以及分裂式(Divisive)兩大類[2]。圖2-2分別描述聚合式與分裂式兩種方法在一個包含五個物件的資料集合{a, b, c, d, e}上的處理過程。

1. 聚合式

聚合式方法採用的是由下而上(Bottom-up)的分群策略。聚合式方法開始時將每一個資料物件當作單一群集，然後尋找相似度最高的群集，當相似度大於既定的臨界值時，則往上一個階層聚合成一個更大的群集；經由反覆進行群集聚合的步驟，直到所有的資料物件聚合成同一個群集，或是符合終止條件為止。

2. 分裂式

分裂式方法所採用的策略與聚合式方法正好相反，分裂式分群法是一種由上而下(Top-down)的方式。分裂式方法開始時將全部的資料物件當作同一個群集，然後尋找相異度最高的群集，再往下一個階層分裂成較小的子群集；經由反覆進行群集分裂的步驟，直到每個子群集都只有一個物件，或是符合終止條件為止。

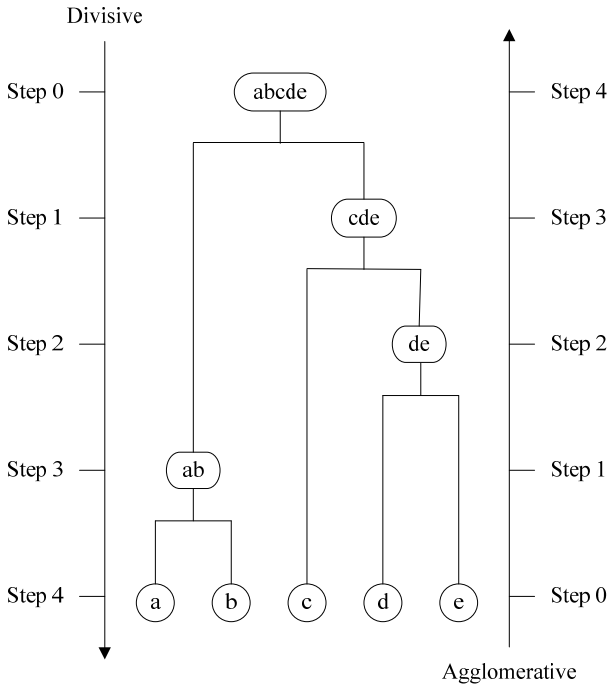

圖 2-2 階層式演算法處理流程

在聚合與分裂的過程中，測量群集間距離的方式又可分為單一連結法(Single Link)以及完全連結法(Complete Link)兩種。單一連結法使用的是最短距離法，即是以兩個群集間資料物件的距離最短者作為群組間的距離，而完全連結法則是使用最長距離法，即是以兩個群集間資料物件的距離最長者為群集間的距離。

階層式分群法的概念雖然簡單，但是經常會遇到合併或分裂點選擇的困難。資料物件一旦被合併或分裂，下一步驟的處理將依循先前的群集結果繼續進行，並且群集之間不能交換物件。因此，若選擇了不適當的合併或分裂的條件，可能會導致最終分群結果不佳。除此之外，此種分群方法的可擴展性不佳，因為合併或分裂的決定需要檢查及估算大量的物件或群集。

Chameleon演算法[8]是在階層式分群法中採用動態模型的演算法。在分群的過程中，當兩個群集間的互連性和相似度有高度相關時，則合併這兩個群集；基於動態模型的合併過程，有利於同質性群集的發現。Chameleon演算法的步驟如下，流程圖如圖2-3所示。

1. 將資料建立成一個稀疏圖形(Sparse Graph)，每個點代表一個資料物

件，資料物件間的關係即代表一個具有權重的邊。

2. 以 k-Nearest Neighbor 演算法[19]將圖形分割為多個子圖，即把相對較大的群集，劃分成較小的子群集。
3. 透過聚合式演算法依照相似度反覆進行合併子群集，直到最後的結果群集產生。以兩個子群集間的相對互連性(Relative Inter-connectivity)與相對近似性(Relative Closeness)來決定相似度。

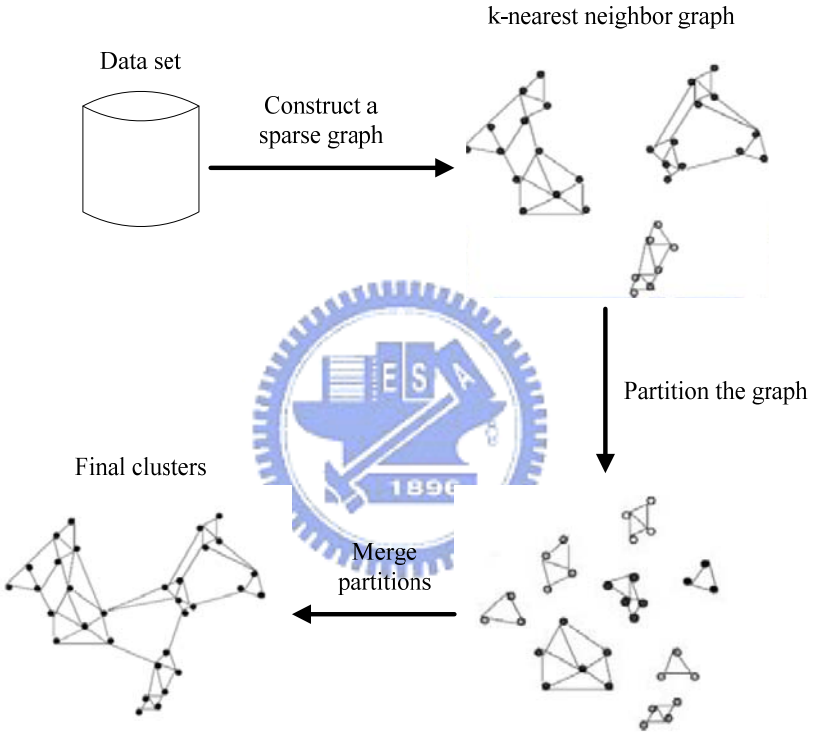

圖 2-3 Chameleon 演算法流程圖[8]

2.1.3 主題關鍵字分群法

主題關鍵字分群法(Topic Keyword Cluster)[9]是應用在文件分群上的一種方法，採用與Chameleon演算法類似的想法。在分群的步驟上，首先取出文件中的關鍵字(Keyword)，將關鍵字分群之後，再計算文件與關鍵字群集之相似度，最後將文件對應至最相似的關鍵字群集。詳細分群步驟如下：

1. 選擇關鍵字

關鍵字的選擇關係到分群結果的準確性，以及是否能夠適當地表達所代表的主題。根據Koller[6]的研究，用來表示文件最適當的關鍵字個數為10~25個，過多的關鍵字反而會降低重要關鍵字的顯著性，圖2-4為關鍵字的數量對分群準確率的影響。因此透過停用字集(Stop Word)可移除詞頻較高的關鍵詞及功能詞(Function Word)，以達到篩選關鍵字的目的

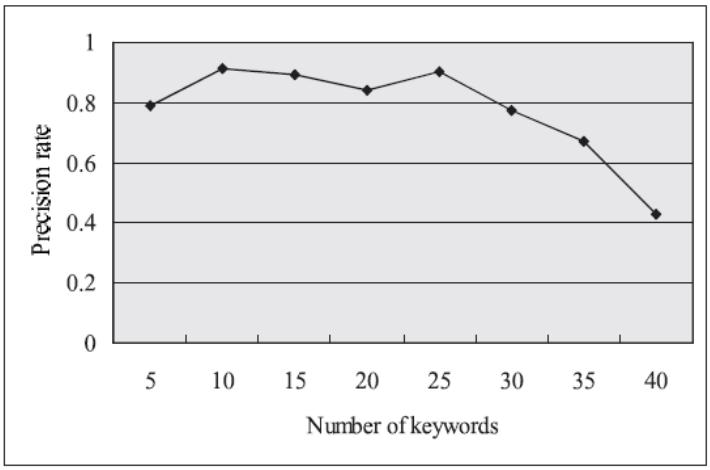

圖 2-4 關鍵字數量與分群準確率關聯圖[9]

2. 計算關鍵字間的關聯性

當關鍵字間共同出現的頻率很高，則代表這些關鍵字間具有關聯性。主題關鍵字分群法於是利用Mutual Information[27]，計算關鍵字共現的頻率作為關鍵字間的語意相關，計算方式如方程式(2-2)所示，分子為兩關鍵字共同出現的頻率，分母則取兩關鍵字在語料庫中出現次數的最大值。

$$r_{ij} = \frac{f(t_i \cap t_j)}{\max\{f(t_i), f(t_j)\}} \qquad (2\text{-}2)$$

3. 建立關鍵字網路圖

根據圖形理論(Graph Theory)建立關鍵字網路圖，以關鍵字代表網路圖中的一個點(Vertex)，關鍵字間的語意相關度為一個邊(Edge)，如圖2-5 (a)所示。接著對網路圖內的連線進行刪減，只保留大於平均語意相關度的連線，

將原本的網路圖修正為稀疏網路圖。如圖2-5(b)所示

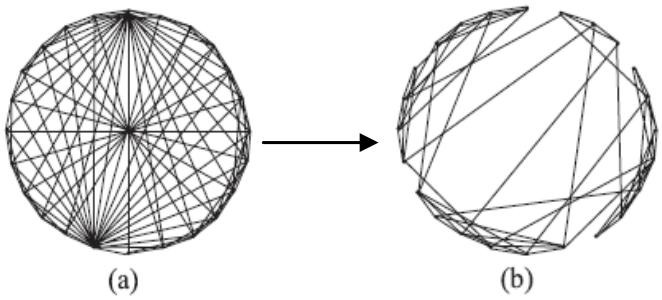

圖 2-5 關鍵字網路圖[9]

4. 進行關鍵字分群

關鍵字分群的主要步驟如圖2-6所示。首先選取出候選的關鍵字，候選關鍵字是由稀疏網路圖中選取權重大於平均的點。點權重CW_i的計算如方程式(2-3)所示，w_i表示點v_i的權重值，即關鍵字i在語料庫中TF-IDF值的加總；r_{ij}則表示所有與點v_i有連線的點v_j間之語意相關度，m為與點v_i有連線的點個數。

$$w_{ij} = tf_{ij} \times idf_i - \text{the weight of term } i \text{ in the document } d_j$$
$$w_i = \sum_{j=1}^{N} w_{ij}$$
$$CW_i = w_i + \frac{\sum_{j=1}^{m} r_{ij}}{m} \qquad (2\text{-}3)$$

接著將候選關鍵字利用k-Nearest Neighbor演算法[19]進行分群，以每個候選關鍵字組為中心，向外還原先前與候選關鍵字組內的點有直接連線關係的邊，形成候選關鍵字子群，並計算每個子群的權重。找出候選關鍵字子群中互連性(Inter-connectivity)最強的兩個群將之合併，直到子群間的互連相關度(Relative Inter-connectivity)都小於門檻值後停止，即得到關鍵字子群。

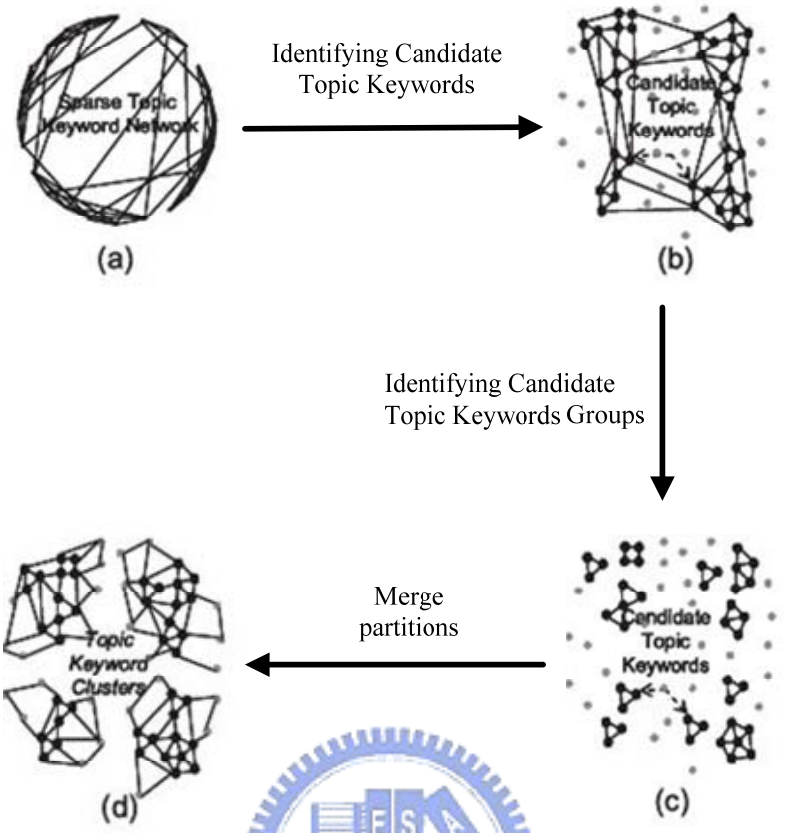

圖 2-6 Topic Keyword Cluster主要步驟[9]

5. 修正關鍵字分群結果及文件分群

在合併候選關鍵字子群的過程中，會造成每一群包含的關鍵字個數產生差距，進而影響文件分群的正確性，因此需要將每個子群內的關鍵字保持在一定的差距內，當子群內的關鍵字數目大於平均關鍵字數目時，則依序移除點權重最小的關鍵字。關鍵字子群修正完成後，將文件與每一個關鍵字群以餘弦相似度(Cosine Similarity)計算相似度，並將文件對應至相似度最高的關鍵字子群中，以達到文件分群的目的。相似度計算方式如方程式(2-4)所示

$$sim(x,\ y) = \frac{\sum_{j=1}^{n} x_{ij} \times y_{ij}}{\sqrt{\sum_{j=1}^{n} x_{ij}^{\ 2}} \times \sqrt{\sum_{j=1}^{n} y_{ij}^{\ 2}}} \qquad (2\text{-}4)$$

2.2 社會網路分析

社會網路(Social Network)是指在社會組織中，個人或是組織間相互連結的關係集合[29]。社會網路分析(Social Network Analysis)是一種研究社會結構、組織系統、人際關係、團體互動的概念與方法，是在社會計量學(Sociometry)基礎上所發展出來的分析方法。社會學是最早注意到社會網路現象並且從事研究分析的領域。心理學家Moreno[16]利用形式社會學派觀點和現象觀察的方式，將人際行為、人際關係數量化，並且以圖形方式表達，呈現人與人之間互動的方向性、接觸的距離等。

社會網路分析研究領域中，最著名的理論之一為「六度分隔」(Six Degrees of Separation)理論[40]。利用信件傳遞實驗，發現從寄件者到收件者之間，平均轉寄了六次，根據這樣的實驗，得出了六度分隔的概念。六度分隔理論指的是互不相干的兩個人，可經由六個人連結出某種關係，也就是最多透過六個人，就能夠認識世界上任何一個陌生人。

經由社會網路分析，可以描繪出原本無法察覺的各種網路關係，無論是人際關係、疾病傳播與文化時尚等，都可以運用社會網路來分析與解釋，並且利用各項量測指標，得以評估社會網路的狀況，瞭解社會網路中的角色，來幫助解決所面臨的問題，於是愈來愈多的研究致力於從不同角度發掘社會網路及其應用。

Tyler[17]利用電子郵件建構社會網路，將社會網路以圖形代表，接著使用中介中心性(Betweenness Centrality)分析，區分出網路圖中的社群結構，用以表示組織中的社群分佈，並且定義出具有領導地位的社群。

Mika[28]發展了一套名為Flink的系統，藉由分析網頁內容、電子郵件、論文著作以及FOAF(Friend of a Friend)檔案，來發掘語意網(Semantic Web)研究領域學者的社會網路，並以視覺化的方式表現社群中的社會網路關係，以及建構語意網研究領域的主題本體論(Ontology)。

Liu[34]則是經由分析過去參與ACM、IEEE及JCDL (ACM/IEEE Joint

Conference on Digital Libraries)成員所發表的論文著作，從共同作者(Co-author)的角度出發，分析在數位圖書館研究領域的社會網路關係，並且提出以此社會網路為基礎的AuthorRank，來和分析網頁超連結關係所構成的PageRank[22]作一比較。

POLYPHONET[35]是以JSAI(Japan Society of Artificial Intelligence)研討會的參與成員為依據，利用搜尋引擎找回和參與成員相關的網路文件，並且從中萃取出和個人相關的關鍵詞，來代表個人感興趣的研究主題，並計算成員間的情境相似度(Context Similarity)以找出潛在的社會網路關係。

2.2.1 社會網路分析單元

社會網路分析的目的在於檢視人們在社會、經濟，文化等框架(Framework)中所扮演的角色，並藉此分析個人與個人、個人與群體、群體與群體之間的互動關係及影響。社會網路的分析單元共有行動者(Actor)、關係(Relation)及聯繫(Tie)三種[42]。分別說明如下：

1. 行動者

網路中所定義的人、事、物，為網路的主體。社會網路分析著重在行動者之間的關係。

2. 關係

兩個行動者間由於某種關係的存在而影響彼此之互動。關係的特徵可經由內容、方向、強度及主動或被動關係來說明[31]。

(1) 內容：內容就是指兩行為者間之關係發生的原因與關係建構基礎。

(2) 方向：關係可分成有方向性(Directed)及無方向性(Undirected)。

(3) 強度：關係也有著程度不同的強度。其衡量方式可能因為不同的關係型態而有所不同。

(4) 主動關係或被動關係：關係產生時，因為行為者本身意向之主被動的不同，也是一種關係的特徵。

3. 聯繫

聯繫是指兩行動者間的關係組合[23]。當行動者間建立某種形式的關係時，必須透過某種途徑達成關係的建立，使行動者互相連結。聯繫所含的關係可能由一種或是多種的關係組合而成，聯繫又可分成弱聯繫(Weak Ties)及強聯繫(Strong Ties)，另外亦可分為直接與間接兩種聯繫模式。

Granovetter[25]提出構成聯繫強度的四項屬性，分別為時間(Amount of Time)、情感強度(Emotional Intensity)、親密(Intimacy)以及相互服務(Reciprocal Services)。他認為家人、朋友以及彼此接觸頻繁的人皆屬於強聯繫；而弱聯繫是指透過本身關係以外的人事物所產生的關聯。研究中指出弱聯繫所提供的資訊或資源較強聯繫更為有用。

2.2.2 社會網路量測指標

在社會網路分析的層面上，依照不同網路層級的特性，具有不同的量測指標。個別行動者的分析主要透過程度中心性(Degree Centrality)、中介中心性(Betweenness Centrality)及緊密中心性(Closeness Centrality)[21]。各項指標之說明如下：

1. 程度中心性

計算特定網路成員與其他成員聯繫的數量，分數愈高表示其在網路中具有重要的地位，其表示式如方程式(2-5)。

$$d(i) = \sum_{j=1}^{h} m_{ij} \tag{2-5}$$

$-m_{ij}$: the edge for vertex i to vertex j

2. 中介中心性

定義為網路關係中，任兩個成員的互動必須透過某個關鍵行動者連結的程度，亦即衡量一個成員是否占據在其他成員相互聯絡的重要捷徑上，其表示式如方程式(2-6)。

$$b(i) = \sum_{j=1,k\neq i}^{n} \frac{g_{jik}}{g_{jk}} \tag{2-6}$$

$-\ g_{jik}$: the number of geodesics between j and k that contain i

$-\ g_{jk}$: the number of geodesics between j and k

3. 緊密中心性

衡量成員和其他連結點之間的最短路徑加總，值愈小者表示和大多數成員之間的關係較為緊密，其表示式如方程式(2-7)。

$$c(i) = \sum_{j=1}^{N} \frac{1}{d_{ij}} \tag{2-7}$$

$-\ d_{ij}$: the shortest path of i to j

Faust[20]認為網路中的行動者基於結構上的相似性，可分類為不同的網路角色。以中心性作為分類的準則，區別主要的網路角色如下：

1. 網路中心(Hub)

即具有較高程度中心性之行動者。由於建立的連結多，在網路中較為活躍，通常是決策的主導者、意見的領袖。

2. 橋樑(Bridge)

即具有較高中介中心性之行動者。橋樑行動者在不同社群間扮演聯繫的角色，愈多社群倚賴特定行動者，而無其他替代溝通管道時，該行動者所具有的橋樑特質愈為重要。

2.3 推薦系統

推薦系統的目的是從大量資訊中找出使用者最可能感興趣的部份，減少使用者主動搜尋的機會成本。推薦系統已廣泛地運用在各種行銷策略上，成為許多電子商務網站的核心功能，利用有效的推薦資訊，不但可以節省使用者的搜尋時間，更能為企業帶來更多的顧客，提升顧客忠誠度，並且增加企業的收益，這使得推薦機制成為重要議題。

Schafer[13]認為在運用推薦系統的機制下，對於電子商務上可獲取的效益為以下三項：

1. 將瀏覽者變成購買者(Browsers into Buyers)

使用者在網站上往往只是快速地瀏覽網頁內容，而不會主動購買商品，利用推薦系統可以適時地推薦使用者感興趣的產品及服務，引發使用者的購買慾望，使瀏覽者也成為購買者。

2. 交叉銷售(Cross-sell)

推薦系統透過對使用者提供已購買商品以外的產品建議，來產生交叉銷售的效益，如果所推薦的產品符合顧客需求，則可以提高平均的交易量。舉例來說，推薦系統可依據使用者購物車中的產品資訊推論使用者的興趣，並依其興趣進行額外的產品推薦。

3. 提高忠誠度(Loyalty)

推薦系統透過學習的機制，瞭解使用者需要什麼，並且預測使用者的需求以進行推薦。與其他的競爭者相比，若能利用推薦系統提供使用者需要的資訊，則使用者將再度到訪能提供最符合其需求的網站，藉此可改善企業與消費者的關係，同時提高顧客的忠誠度。

推薦系統在進行推薦之前，必須先取得使用者的相關資料，瞭解使用者對於哪些資訊感興趣，才能使推薦符合個人需求，並且增加推薦的準確性。系統必須獲得三種資料才能做出推薦，包含人口統計資料、物品的特徵屬性及使用者

的偏好[41]，說明如下：

1. 人口統計資料

人口統計資料可做為辨別使用者類型的依據，例如推薦系統可分別採用性別、年齡、教育、職業與薪水等統計資訊，區分出不同類型的使用者。經由分析各種類型使用者的喜好特徵，可以依據不同的類型進行不同的推薦。例如對於年輕女性，可能會推薦她們流行服飾或是保養品，高收入的中年男性，則可能推薦高爾夫球。

2. 物品的特徵屬性

物品本身的屬性也可以做為推薦的依據，即具備類似或互補屬性的產品將優先被推薦。屬性一般又可以分為外在屬性(Extrinsic Feature)及內在屬性(Intrinsic Feature)。外在屬性是指無法用自動化方式取得的特徵，如品牌形象、顏色、物品從屬關係等；相反地，內在屬性通常能藉由分析物品內容得到，其內容或關鍵的屬性都可以被萃取出來。物品具備的屬性繁多，為了管理上的方便，通常套用階層式架構，依不同的層次加以分類，將資訊有條理地組織起來。在資訊量太大的情況下，以類別的推薦來代替物品的推薦會更具有意義。

3. 使用者的偏好

使用者偏好(User Preference)是個人化推薦的重要依據，唯有瞭解使用者本身的喜好才能達成個人化的推薦。即分析個人的興趣與喜好，建構使用者設定檔，經由與產品項目間特徵屬性的比對，進行符合個人喜好的推薦。單純用人口統計資料或物品特徵屬性來做推薦僅能作一般性的推薦，不會有個別的差異。

一般常用的推薦機制有內容導性式推薦和協同過濾式推薦，分述如下：

2.3.1 內容導向式推薦

內容導向式推薦主要依據使用者個人過去所購買或曾接觸過的商品，得知使用者的喜好，進而推薦使用者相近的商品。內容導向式推薦必須為每位使用者建立使用者設定檔，對使用者的興趣詳加描述，推薦系統將物品的特徵屬性和使用者的輪廓做比對，相似度高的物品將優先被推薦。主要的構想是從資訊檢索領域延伸而來，以物件的屬性或特徵為每個商品建立向量，以向量間的餘弦相似度值判斷兩個物件是否相關。當兩物件間向量的夾角越小時，代表相似度越高，反之則越小。

內容導向式推薦常應用於可解析內容或描述之相關資訊推薦，例如網頁、文件與新聞等，由於可將資訊之內容或關鍵屬性萃取並加以分析，因此成為內容推薦之主要依據。內容導向式推薦也因為先天上的限制，具有以下幾項缺點：

1. 音樂、電影、照片等多媒體資料，無法以文字表達，其內容維度因而較難清楚定義，導致較難對此類型的商品進行推薦。
2. 內容導向式推薦系統因為是依照使用者過去的喜好做為推薦依據，因此往往僅能推薦出使用者最喜好的商品，無法找到一些對顧客來說較特殊或不一樣的產品，可能無法發現顧客的潛在需要。

內容導向式推薦的應用相當廣泛，一般使用資訊檢索的技術來分析文件內容，並透過文件與使用者設定檔的比較來向使用者進行推薦。以下分別介紹內容導向式推薦的實際運用。

InfoFinder[3]是經由訊息資料集(Set of Messages)或是其他的線上文件，來得知使用者的資訊喜好類別。此系統的特點在於使用啟發式(Heuristic)的搜尋來取得有意義的片語，優點在於不需要很多文件樣本就可以正確取得使用者的偏好。

ANATAGONOMY[12]則是從使用者瀏覽網頁的操作行為來學習使用者的個人偏好。系統分為學習引擎(Learning Engine)以及評分引擎(Scoring Engine)兩個部份。學習引擎從使用者操作行為中分析使用者偏好；而評分引擎則負責新聞文

件的評分，並依據使用者設定檔建立個人化的電子報。研究中比較明顯性評分(Explicit Rating)與隱含性評分(Implicit Rating)兩種模式的效果。在明顯性模式中，使用者每閱讀完一篇文章，必須以分數來評估該文章與本身興趣的關連性；在隱含性模式中，系統會記錄使用者閱讀文章時的動作，推導出對使用者對該文章感興趣的程度。

SmartPad[33]為到賣場購物的使用者提供個人化的商品清單。此系統為所有的商品建立一顆分類樹，經由分析使用者過去的消費記錄，計算出使用者對特定商品類別的偏好程度。二商品間的相似度即以它們在分類樹中所在的相對位置表達。SmartPad以使用者對商品類別的偏好程度，與商品間的相似度，為使用者提供個人化的商品推薦清單。

2.3.2 協同過濾式推薦

協同過濾式推薦藉由發掘系統內使用者間的相關性，以此做為推薦的參考依據，並用以推論或預測目標使用者的潛在偏好。使用者通常喜歡參考來自於與他相同興趣之人的建議，透過統計分析的過程，找出與使用者本身最為相似的使用者，依據相似使用者過去的交易記錄或其他屬性，預測使用者對於尚未見過事物的喜好程度，進而推薦喜好程度最高的商品清單給特定使用者。

在 Sarwar[4]的研究中，將協同過濾式推薦分成以下三部份，如圖 2-7 所示。

1. 輸入資料的表示法(Representation of Input Data)

   建立出系統中使用者購買產品的行為模式。將使用者的交易紀錄建立在一個矩陣R中，以r_{ij}代表使用者i購買產品j的關聯。

2. 形成鄰近者(Neighborhood Formation)

   經由分析使用者偏好與計算使用者間的相似度，找出具有相似偏好的使用者社群，以社群成員之偏好及相似度做為進行推薦之依據。

3. 產生推薦(Recommendation Generation)

   將使用者社群內所偏好的前n項產品資訊推薦給目標使用者，提供其做

為交易的參考依據。

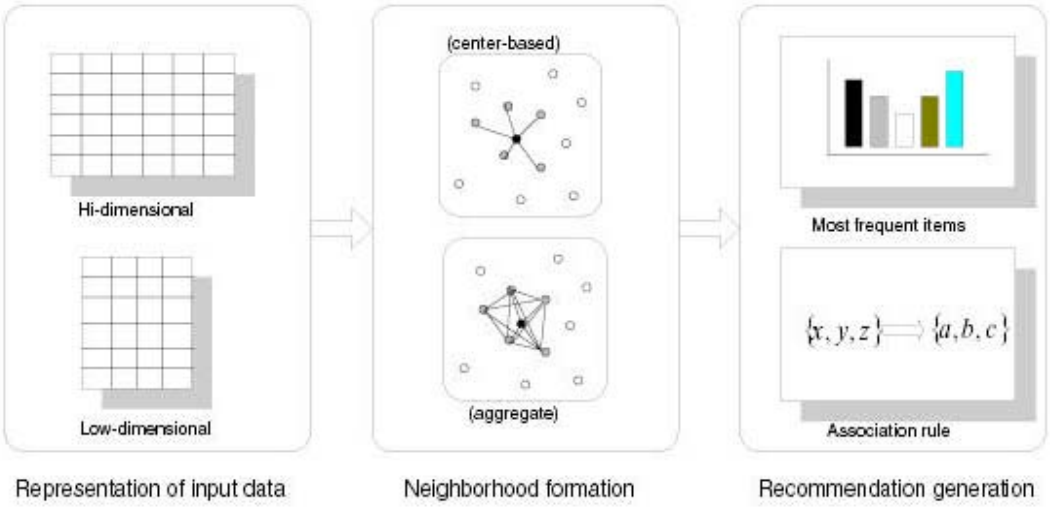

圖 2-7 推薦系統的功能組成[4]

協同過濾依據其他顧客的意見來為目標顧客推薦產品，因此所推薦之產品可能與使用者從前的喜好大不相同，但是卻能發掘出消費者的潛在需要。以下介紹協同過濾式推薦的實際運用。

Tapestry[5]的主要目的在過濾電子郵件，之後延伸到所有的電子文件。系統能夠讓使用者對電子郵件加上註解(Annotation)，由使用者來判斷文件品質的優劣，並且使用者可以透過查詢的方式，過濾瀏覽自己有興趣的電子郵件及註解紀錄。但是此系統需要使用者被動地以查詢來過濾信件，而非系統主動地對使用者進行推薦。

ReferralWeb[11]是一個結合社會網路與協同過濾的系統，主要利用使用者在網路文件中的共現(Co-occurrence)情形，發掘社會網路關係，並以此進行資訊推薦。該研究中提出的資料來源: (1)個人網頁的連結；(2) 論文著述間的共同作者(Co-author)與共同引用(Co-citation)關係；(3) 使用者在網路論壇中的交流記錄；(4)組織架構圖。接著當使用者查詢某個關鍵字時，系統便會先從與使用者較相關的人找起，以幫助使用者很快地找到該社會網路中的專家。

SiteSeer[18]是利用使用者的瀏覽器書籤(Browser Bookmark)判斷使用者的偏好，比較使用者間書籤的相似度來找出具有相同偏好的使用者，並將相同偏好者書籤裡的其他網站推薦給目標使用者。使用者會把網站加入書籤，表示使用者對此網站具有相當程度的偏好，所以將書籤視為一個可信賴的隱含性評分。此外，從使用者書籤的分類架構，也可以看出使用者的資訊分類方式，可作為呈現推薦資訊時的依據。

第三章　研究方法

本研究以交通大學機構典藏系統 (National Chiao Tung University Institutional Repository, NCTUIR) [38]所收集的期刊論文做為語料庫(Corpus)，選取標題(Title)、摘要(Abstract)、關鍵字(Keyword)及作者(Author)欄位做為資料來源。後續研究中所提及之「使用者」即為NCTUIR中記錄之「作者」。

NCTUIR是將交通大學本身的研究產出，如期刊、會議論文、研究報告、投影片與教材等，以數位的方式保存，並建立網路平台，提供檢索與使用的系統，對內是研究人員的交流平台，能夠保存記錄研究傳承與發展；對外則能幫助展現研究能量，提高學術成果的能見度與影響力，增加一個研究成果被使用的管道。

在本章中將闡述本研究提出的結合主題關鍵字分群演算法與社會網路分析的資訊推薦方法。首先說明前置處理的步驟；再詳細描述進行主題關鍵字分群、社會網路分析、形成主題社群以及資訊推薦的方法，本章會在每一節詳述每一個流程的步驟。

3.1 前置處理

前置處理的目的在於提高資料的正確性，避免雜訊的干擾。此步驟包含斷詞切字(Tokenization)、小寫化(Lowercasing)、刪除停用字(Stop Word Removing)、詞性標記(Part-Of-Speech)、片語化(Chunking)以及詞幹還原(Stemming)。

3.1.1 斷詞切字與小寫化

斷詞切字主要是用來找出文字的分界，在英文文件中是利用空格和標點符號來判斷句子與單字，以達到斷詞切字的目的。字彙的大小寫會影響到計算字詞頻率時的差異及詞性的判斷，為了避免判斷錯誤，本研究將全部的字小寫化，再進行斷詞切字的處理。

3.1.2 刪除停用字

停用字是指文章中沒有語意，一旦脫離文句語境來解釋，便沒有任何涵義存在，但可用來平順語意的字詞。停用字一般包含介系詞、指代詞、連接詞與助詞等。本研究以頻率來計算字詞的重要程度，這些停用字經常出現於文件中，反而會被誤判為具有相當程度之重要性，因此在前置處理中將其過濾。表 3-1 為部份停用字列表。

表 3-1 部份停用字列表

Stop word list					
about	after	became	because	can	could
down	during	each	everyone	few	former
get	give	hence	however	if	indeed
keep	latterly	less	many	moreover	neither
never	often	otherwise	part	perhaps	rather
re	same	several	take	thereby	until
upon	very	via	well	whatever	yet
yourself	the	whom	anywhere	an	for

3.1.3 詞性標記

英文文件中的重要概念大多是由名詞或名詞片語所組成，故詞性標記也是篩選字詞的重要步驟之一。本研究採用LingPipe NLP Toolkit[37]進行實作，LingPipe是一套處理自然語言的JAVA API，提供的功能包含有詞性標記(Part-of-Speech Tagging)、中文斷詞(Chinese Word Segmentation)、情感分析(Sentiment Analysis)、語言辨別(Language Identification）等等。其中詞性標記功能是以Brown Corpus作為語料庫，Brown Corpus是由美國Brown大學於1960年代所建立，是一個根據系統性原則採集樣本的標準語料庫，收集了約一百萬個字詞，並加上詞性標記。以圖3-1的原文範例進行詞性標記，其結果如表3-2所示。

Some combinatorial characteristics of matrix multiplication on regular two-dimensional arrays are studied. From the studies, the authors are able to design many efficient varieties of the cylindrical array and the two-layered mesh array for matrix multiplication.

圖 3-1 原文範例

表 3-2 詞性標記處理結果

#	Token	(Prob:Tag)*				
0	Some	0.991:dti	0.007:rb	0.001:ql	0.000:np	0.000:vb
1	combinatorial	0.973:jj	0.027:nn	0.000:nns	0.000:nn$	0.000:nns$
2	characteristics	0.995:nns	0.005:nn	0.000:jj	0.000:vbz	0.000:rb
3	of	1.000:in	0.000:jj	0.000:nn	0.000:rb	0.000:vbg
4	matrix	1.000:nn	0.000:nns$	0.000:jj	0.000:vb	0.000:np$
5	multiplication	1.000:nn	0.000:vb	0.000:vbd	0.000:jj	0.000:vbn
6	on	0.997:in	0.003:rp	0.000:nn	0.000:ql	0.000:cs
7	regular	1.000:jj	0.000:nn	0.000:rb	0.000:nns	0.000:ql
8	two-dimensional	1.000:jj	0.000:nn	0.000:nns	0.000:vbg	0.000:rb
9	arrays	0.992:nns	0.007:nn	0.000:np	0.000:vbz	0.000:nps
10	are	1.000:ber	0.000:nn	0.000:vb	0.000:rb	0.000:ql
11	studied	0.999:vbn	0.001:vbd	0.000:jj	0.000:rb	0.000:ql
12	.	0.797:.	0.066:np	0.059:rb	0.039:nn	0.026:nns
13	From	1.000:in	0.000:np$	0.000:np	0.000:rb	0.000:nr$
14	the	1.000:at	0.000:jj	0.000:dts	0.000:nn	0.000:np
15	studies	1.000:nns	0.000:nn	0.000:nps	0.000:vbz	0.000:nns$
16	,	1.000:,	0.000:rb	0.000:wpo	0.000:in	0.000:abl
17	the	1.000:at	0.000:jj	0.000:vb	0.000:dts	0.000:nn
18	authors	1.000:nns	0.000:nn	0.000:nn$	0.000:nns$	0.000:nps
19	are	1.000:ber	0.000:vb	0.000:ql	0.000:jj	0.000:rb
20	able	0.999:jj	0.001:nn	0.000:vb	0.000:nns	0.000:vbn
21	to	0.954:to	0.046:in	0.000:nn	0.000:jj	0.000:nn$
22	design	0.953:vb	0.047:nn	0.000:jj	0.000:vbz	0.000:vbg
23	many	0.951:ap	0.025:abn	0.023:ql	0.000:jj	0.000:rb
24	efficient	0.985:jj	0.015:nn	0.000:rb	0.000:nns	0.000:vb
25	varieties	1.000:nns	0.000:vbz	0.000:nn	0.000:nps	0.000:np
26	of	1.000:in	0.000:rb	0.000:rp	0.000:vb	0.000:vbd
27	the	1.000:at	0.000:jj	0.000:ql	0.000:dts	0.000:pp$$
28	cylindrical	1.000:jj	0.000:nn	0.000:nn$	0.000:nns	0.000:np$
29	array	1.000:nn	0.000:np	0.000:jj	0.000:vbn	0.000:vb
30	and	1.000:cc	0.000:vb	0.000:nn	0.000:rb	0.000:vbd
31	the	1.000:at	0.000:vb	0.000:jj	0.000:ql	0.000:pps
32	two-layered	0.985:jj	0.015:vbn	0.000:nns$	0.000:nn	0.000:vbg
33	mesh	0.898:nn	0.101:jj	0.001:vb	0.000:np	0.000:nns$
34	array	0.999:nn	0.001:vbd	0.000:vbn	0.000:vb	0.000:np
35	for	0.994:in	0.006:cs	0.000:nn	0.000:jj	0.000:rb
36	matrix	1.000:nn	0.000:nns$	0.000:jj	0.000:vb	0.000:np$
37	multiplication	1.000:nn	0.000:vb	0.000:jj	0.000:vbd	0.000:vbn
38	.	0.988:.	0.007:np	0.002:nn	0.001:rb	0.001:nns

3.1.4 片語化

英文句子是由有意義的語意單位所組成，以單一字彙在語意判斷上是不足夠的，後續在建立語意關聯時，必須統計字詞出現的頻率，若沒有經過片語化會把字面相同但是意義不同的單字計算在一起，這樣就無法區分出語義的歧異(Word

Sense Ambiguity)，故需將單一字彙組合成片語以表達正確的語意。片語化同樣採用LingPipe NLP Toolkit進行實作。以圖3-1的原文範例進行片語化，其結果如表3-3所示。

表 3-3 片語化處理結果

POS	Phrase
noun	combinatorial characteristics
noun	matrix multiplication
noun	regular two-dimensional arrays
verb	studied.
noun	studies
noun	author
verb	design
noun	efficient varieties
noun	cylindrical array
noun	two-layered mesh array
noun	matrix multiplication

3.1.5 詞幹還原

詞幹轉換即是去除型態學(Morphology)上的詞類型態變化，英文時常因為時態或是句型文法變化將字詞的字尾形態改變，當進行資料擷取時，形態變化會導致無法準確地計算字詞出現的頻率，進而影響字詞相關度的計算結果。詞幹轉換便是將經過變形的字尾以統一的結尾表示。本研究採用Porter演算法[39]進行詞幹轉換。

3.1.6 特徵選擇

本研究以向量空間模型(Vector Space Model)來代表個別使用者與關鍵字之關聯，每一個單一的字詞都代表向量空間的一個維度(Dimension)，沒有經過特徵選擇的過程，將造成相當高維且稀疏的向量空間，在分群過程中耗費時間在處理不具代表性或甚至是無意義的字詞上，同時也可能降低重要字詞的顯著性。故本研究依據下列幾項規則對字詞進行過濾及篩選，以達到維度縮減(Dimension

Reduction)[7]的目的：

1. 僅保留名詞與名詞片語為侯選關鍵字。

論文中的主題概念通常以名詞或名詞片語表示，故本研究中僅保留名詞與名詞片語為侯選關鍵字。

2. 移除在語料庫出現次數過多的字詞。

當一個字詞在語料庫內經常出現時，幾乎可以確定此字詞屬於過於常見且不具有代表性之字詞。本研究將出現次數定義為出現該字詞的文章篇數，且出現次數的上限為語料庫內文件篇數的10%。

3. 移除在語料庫出現次數過少的字詞。

若一個字詞出現的次數太少，則此字詞幾乎可以確定不適合用以表達文件的概念。本研究訂定出現次數的下限為3次。

3.2 主題關鍵字分群

在主題關鍵字分群過程中，依據圖形理論(Graph Theory)將關鍵字及語意關係建立語意網路圖，進行過濾及分群，以達到主題概念萃取的目的。主題概念萃取的方法為將2.1.3節中所提之主題關鍵字分群(Topic Keyword Cluster)[9]演算法加以修改，以符合本研究的需要。

3.2.1 使用者模型

在資訊檢索的領域中，最常使用TF-IDF(Term Frequency-Inverse Document Frequency)來評估詞彙在文件中的重要性。計算公式如方程式(3-1)所示。TF指的是某一字詞在一篇文件或資訊內容中出現的頻率，當TF值愈高時，代表該文件與此字詞的關聯性愈高；IDF為在全部文件中有多少文件包含此一字詞之倒數，IDF值愈高則此字詞愈能代表此文件。經由TF-IDF判斷文件與字詞的相關屬性，可以進一步瞭解文件所代表的概念，藉由相似度的計算，可以達到文件分群的效果。

$$tf_{ij} - idf_i = freq_{ij} \times \log_2 \frac{N}{n_i} \tag{3-1}$$

$-\ freq_{ij}$: frequency of term i in the document j

N : # of documents

n_i : # of documents that contain the term i

本研究的目的在於發掘使用者感興趣的主題，以及達到使用者分群的效果，為此，採用TF-IAF(Term Frequency-Inverse Author Frequency)[30]來衡量使用者與關鍵字間的關聯，並且利用相似度的計算，達到使用者分群的目的。其計算如方程式(3-2)所示。TF指某一字詞在與使用者相關文件中出現的頻率，當TF值愈高時，代表使用者與該字詞的關聯性愈高；IAF為全部使用者中，有多少使用者曾經使用過此一字詞之倒數，IAF值愈高則此字詞愈能代表使用者。在計算完TF-IAF之後，則每個使用者皆可以向量的形式來呈現，如方程式(3-3)所示。

$$
\begin{aligned}
&tf_{ij} = freq_i - \text{frequency of term } i \text{ associated with author } j \\
&iaf_i = \log_2 \frac{N}{n_i} - \begin{cases} N : \#\text{ of authors} \\ n_i : \#\text{ of author that use the term } i \end{cases} \\
&w_{ij} = tf_{ij} \times iaf_i
\end{aligned} \tag{3-2}
$$

$$
\begin{aligned}
&U_j = (w_{1j},\ w_{2j},\ .....,\ w_{mj}) \\
&- w_{ij} = \begin{cases} tf_{ij} \times iaf_i \text{ , if term } i \text{ is associated with author } j \\ 0 \quad \text{, otherwise} \end{cases} \\
&- m : \#\text{of keywords}
\end{aligned} \tag{3-3}
$$

3.2.2 計算語意相關度

在計算語意相關度時，以「共現」(Co-occurrence)原則來判斷兩個關鍵字之間是否具有語意關係，首先需要訂出一個範圍，在這個範圍內出現的關鍵字才具有語意相關性。本研究以句子為範圍，即當兩個關鍵字在同一個句子內出現才表示其具有語意相關性。

由於本研究採用論文為資料來源，根據論文的特性，標題與關鍵字通常是表達文件的主題概念，出現於標題與關鍵字欄位中之字詞往往具有較重要的語意關係，故在計算關鍵字間語意相關度時，透過增加此類關鍵字之權重來強化其代表性，如方程式(3-4)所示。計算出語意相關度之後，並對其進行篩選，門檻值為所有語意相關度的平均，只取大於門檻值的語意關係。

$$
r_{ij} = \begin{cases} 1 & \text{, if term } i \text{ \& term } j \text{ are both in title or keyword} \\ \dfrac{f(t_i \cap t_j)}{\max\{f(t_i), f(t_j)\}} & \text{, otherwise} \end{cases} \tag{3-4}
$$

3.2.3 建立語意網路圖

根據圖形理論之原理，每個關鍵字都可表示為一個點，點權重為個別關鍵字在使用者間TF-IAF值的加總，再加上該關鍵字所有語意相關度平均。如方程式(3-5)所示，關鍵字間的關係表示成一個邊，邊權重即為關鍵字的語意相關度，如前述方程式(3-4)，由這些點與邊即可組合成一個語意網路圖。

$$w_i = \sum_{j=1}^{N} w_{ij}$$

$$CW_i = w_i + \frac{\sum_{j=1}^{h} r_{ij}}{h} \qquad (3\text{-}5)$$

$-\ N$: # of authors

$-\ h$: the degree of vertex v_i

得到基本的語意網路圖後，依照下列的步驟對此網路圖進行處理：

1. 移除網路圖內不重要的邊

對網路圖內的邊進行刪減，取門檻值為網路圖內所有語意關係度的平均，刪除小於門檻值的連線，將原本的網路圖修正為稀疏網路圖。如圖3-2所示。

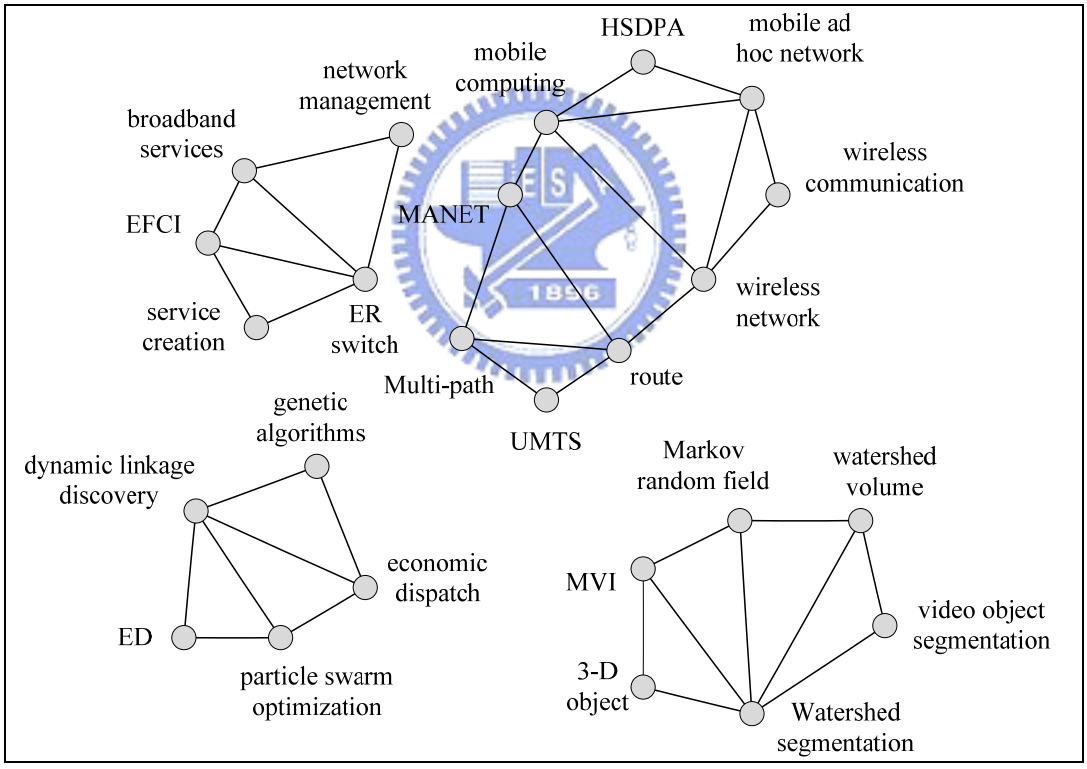

圖 3-2 稀疏網路圖

2. 移除稀疏網路圖內不重要的點

在移除不重要的邊之後，接著移除稀疏網路圖中不重要的點，將門檻值訂為所有點權重的平均，移除小於平均值的點。如圖3-3所示。

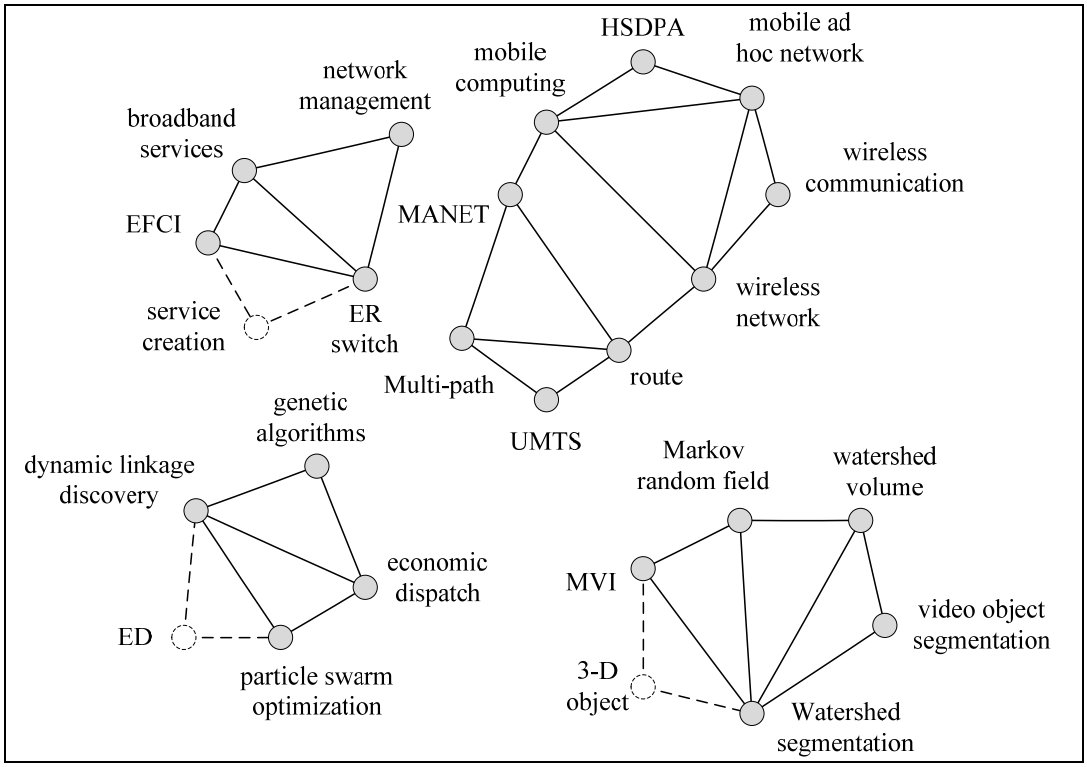

圖 3-3 選出重要關鍵字

3.2.4 關鍵字分群

在前一步驟所建立的語意網路圖中，圖內的點即代表語料庫中重要的關鍵字，將這些關鍵字利用演算法進行分群運算，所獲得的每一個群即代表萃取出的一種主題概念，此步驟在本研究中稱之為主題萃取，將重要的步驟詳述如下：

1. k-Nearest Neighbor Approach [19]

考慮圖中的每個點，取與該點最相近的k個點為一組，每組都為一個連通圖(Connected Graph)，本研究稱之為候選關鍵字組。如圖3-4所示。k值的選擇會影響分群數目的多寡，當k值愈大，群數便會愈少，而每一群包含的關鍵字也會較多，但當群內的關鍵字數量過多時，反而無法表達出清楚的主題概念，降低重要關鍵字的代表性。

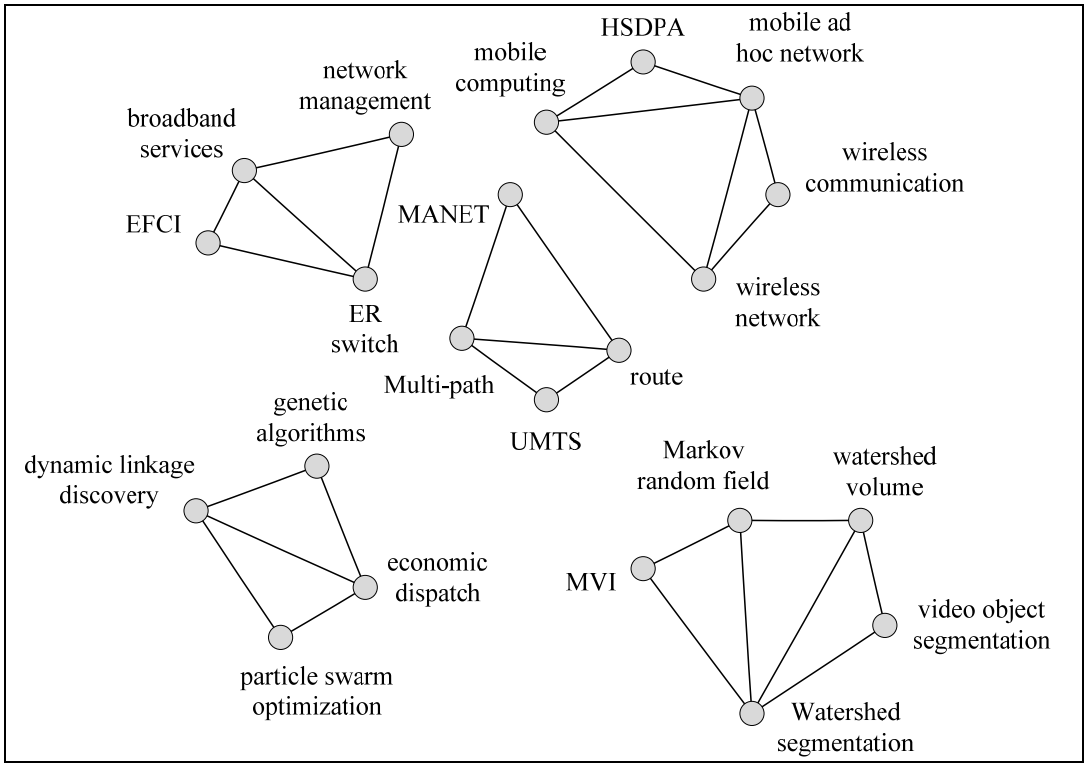

圖 3-4 k-Nearest Neighbor Graph

2. 產生候選關鍵字子群

以每個候選關鍵字組為中心，向外還原先前與候選關鍵字組內的點有直接連線關係的邊，形成候選關鍵字子群，並計算每個子群的權重，權重計算方式為該群內所有邊權重的總和，如方程式(3-6)所示，G_k表示某一候選關鍵字組k，r_{ij} 則是G_k內包含的語意關係。

$$W_{G_k} = \sum_{r_{ij} \in G_k} r_{ij} \tag{3-6}$$

3. 合併候選關鍵字子群

產生候選關鍵字子群之後，找出互連性(Inter-connectivity)最強的兩個子群將之合併，直到子群間的互連相關度(Relative Inter-connectivity)都小於門檻值後停止。互連性強度的判別是依據兩個候選關鍵字子群的互連相關度來計算，計算方式為連接兩個子群的邊之權重總和再除上兩個子群的權重總和，如方程式(3-7)所示，當兩個子群有交集的邊具有相當程度關係時，即將

之合併。如圖3-5所示。

$$RI(G_i, G_j) = \frac{|W_{E(G_i,G_j)}|}{|W_{G_i}| + |W_{G_j}|} \tag{3-7}$$

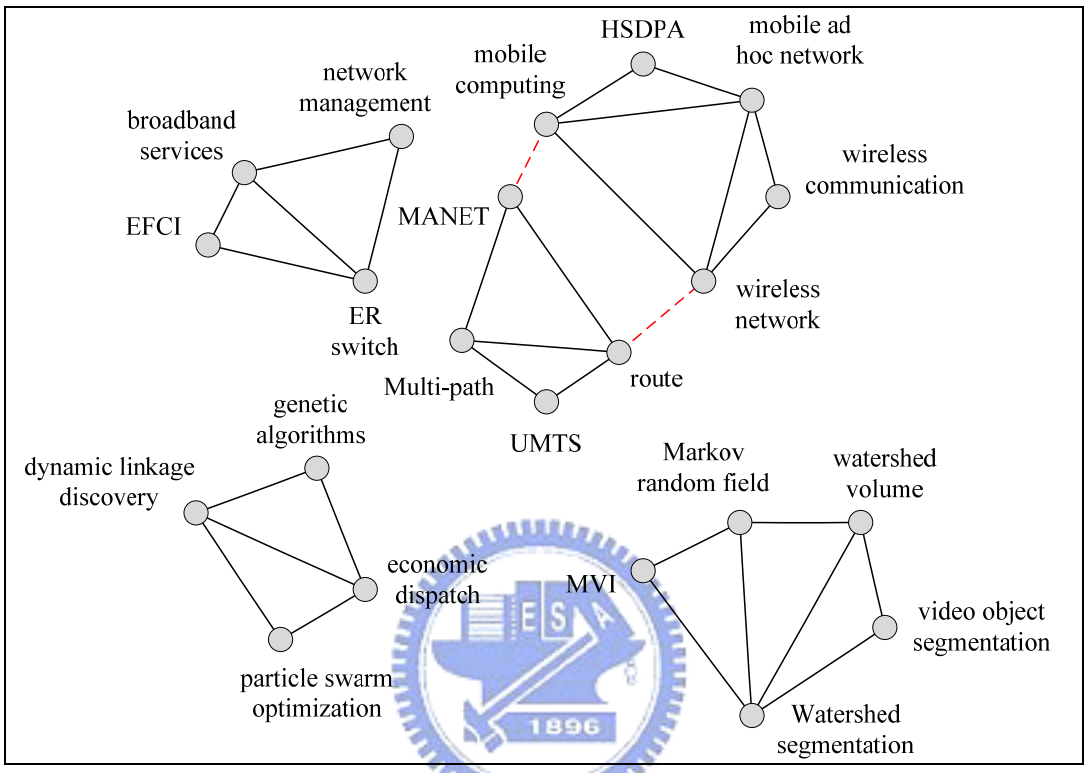

圖 3-5 合併關鍵字子群

4. 修正並產生主題關鍵字分群

合併候選關鍵字子群之後，子群內關鍵字的個數會有不同，此現象將會影響分群的正確性，因此本研究希望每個子群內的關鍵字個數保持在一定的差距內。故當子群內的關鍵字個數大於平均個數時，則移除與該群最不相關，即點權重最小的關鍵字。

若是子群內包含的關鍵字比平均個數少，但是子群權重卻大於平均權重時，表示此群內包含重要的關鍵字關係，故將該群保留；若子群經修正後仍小於平均權重，則表示該群內的沒有重要的關鍵字關係，故將該群直接刪除。子群權重的計算方法如方程式(3-8) ~(3-10)所示，利用每個子群之邊權重平均值及連線密度(Connected Density)來計算子群的權重。連線密度為該

子群內的邊個數除以該子群可能最大邊數，如方程式(3-8)所示。邊權重的計算方式如方程式(3-9)。

$$CD(G) = \frac{|E(G)|}{(|V(G)| \times |V(G)-1|)/2} \tag{3-8}$$

$$AS(G) = \frac{\sum_{r_{ij} \in G} r_{ij}}{|E(G)|} \tag{3-9}$$

$$CW = CD(G) \times AS(G) \tag{3-10}$$

以圖3-6為例，{ER switch, broadband services, EFCI, network management}包含四個關鍵字，{HSDPA, mobile ad hoc network, mobile computing, MANET, wireless network, wireless communication, route, Multi-path, UMTS}一共包含九個關鍵字，{genetic algorithms, dynamic linkage discovery, particle swarm optimization, economic dispatch}包含四個關鍵字，{Markov random field, MVI, watershed volume, Watershed segmentation, video object segmentation}則包含五個關鍵字，每群的平均個數為(4+9+4+5)/4=5.5，表示每個子群最多只能包含五個關鍵字。因此需將包含關鍵字MANET的子群進行修正，依序將不重要的關鍵字HSDPA、Multi-path、UMTS及wireless communication移除，以達到平衡子群內關鍵字個數的目的。若是子群內包含的特徵少，但是子群權重卻大於平均權重時，表示此群內包含重要的關鍵字關係，故將該群保留；反之，若子群經修正後仍小於平均權重，則表示該群內的沒有重要的關鍵字關係，故將該群直接刪除。

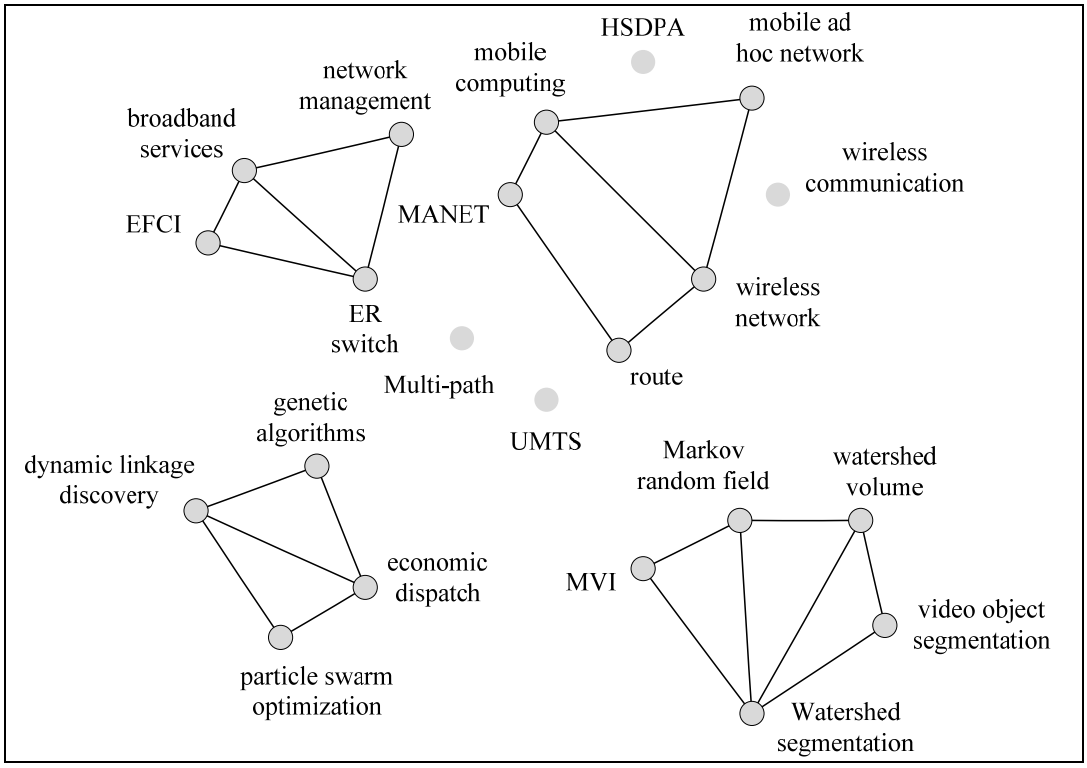

圖 3-6 修正關鍵字子群

3.2.5 關鍵字分群標記

標記之目的是為了讓使用者能夠快速且容易瞭解每一個分群所代表的主題概念，表3-4為主題分群標記之範列。由觀察得知名詞及名詞片語較能表達群的主題概念，且名詞片語的重要性又大於名詞。故本研究依照下列兩項規則來挑選主題概念：

1. 利用人力過濾出有意義的關鍵字；
2. 取權重最高的關鍵字做為最後的群標記。

表 3-4 關鍵字分群標記範例

Topic label	Keywords	Associate weight
Mobile Computing	MANET	351.1895
	mobile ad hoc network	152.5828
	route	140.446
	mobile computing	126.5852
	wireless network	95.5504
Genetic Algorithm	genetic algorithm	97.3025
	dynamic linkage discovery	66.415
	particle swarm optimization	53.332
	GA	41.3458
	economic dispatch	27.0827
Semantic Query	Metadata	53.2695
	Semantic query	40.1823
	Digital library	27.066
	Structure clustering	27.041
	Structure expression	26.966
Neural Network	neural network	72.7728
	optimization problem	49.7963
	constraint	43.1446
	energy function	38.9312
Watershed Segmentation	watershed volume	82.466
	Watershed segmentation	78.002
	multiview images	66.1728
	Markov random field	47.2012
	3-D model	33.5364
Parallel Algorithm	Matrix Multiplication	62.054
	cylindrical array	53.2935
	Parallel algorithms	49.8796
	two-layered mesh array	38.9645
	regular arrays	38.6312

3.3 建立主題社群

在前一小節萃取出主題概念後，利用餘弦相似度計算使用者向量與各主題分群的相似度，進而產生對特定主題感興趣的使用者社群，其目的在於發掘出具有相同興趣的使用者，藉由分析或預測使用者的喜好，推薦使用者感興趣的論文清單。

在建立使用者主題社群的同時，不僅僅專注於個別使用者的關鍵字分佈與主題興趣，也把社會網路的互動關係對使用者的影響考慮進來，經由分析使用者的社會網路，來衡量使用者間的相關係數，進而調整使用者與關鍵字的關聯。

3.3.1 使用者社會網路

社會網路會由不同的觀察角度，產生不同的結果。採用研究論文為資料來源時，通常經由共同作者或共同引用兩個角度來發掘社會網路。本研究是藉由分析使用者間的共同作者關係建立使用者的社會網路。在衡量使用者社會網路時，採用Shah[24]所提出之方法，利用Jaccard coefficient來計算使用者相關係數。首先將使用者與文件之間的關係以矩陣W表示，如方程式(3-11)與圖3-7所示。

$$W_{ij} = \begin{cases} 1, \text{if user } i \text{ is one of the authors in document } j \\ 0, \text{otherwise} \end{cases} \tag{3-11}$$

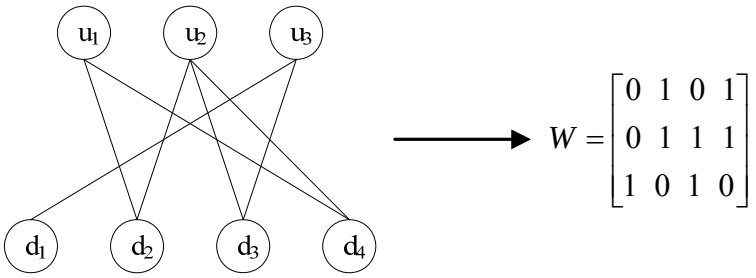

圖 3-7 使用者與文件之關係矩陣

接著以 $S = W \times W^T$ 表示使用者的共同作者關係，矩陣元素 S_{ij} 代表使用者間共同發表的論文篇數。如圖 3-8 所示。

$$S = W \times W^T = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 3 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

圖 3-8 使用者共同作者矩陣

隨後利用 Jaccard coefficient 來計算使用者間的相關程度，即將矩陣 S 中的元素 S_{ij} 除以$|W_i|+|W_j|-S_{ij}$($|W_i|$表示使用者 i 所撰寫的文章篇數)，如方程式(3-12)所示；則使用者間的社會網路關係可以圖 3-9 表示。

$$J_{ij} = \frac{S_{ij}}{|W_i| + |W_j| - S_{ij}} \tag{3-12}$$

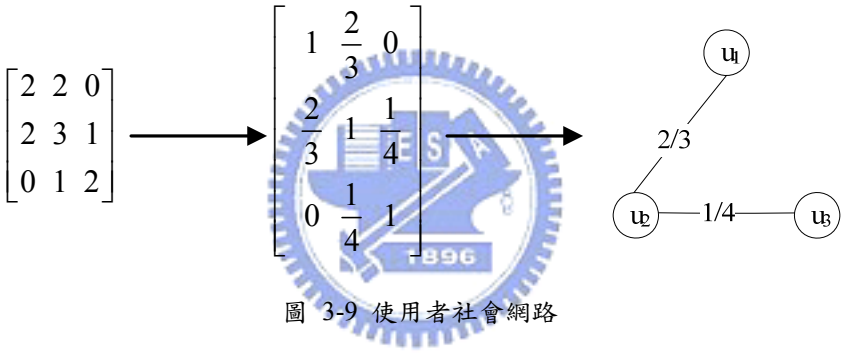

圖 3-9 使用者社會網路

為了避免透過使用者社會網路對使用者本身的影響程度過大，對矩陣 S 中的 Jaccard coefficient 作一調整，並且以矩陣 R 代表調整後的使用者相關係數，如方程式(3-13)所示。調整係數 α 為一介於 0 與 1 之間的值，當 α 愈接近 0，則代表使用者間互相影響的程度較小，反之則影響較大。

$$R_{ij} = \begin{cases} 1 & ,\text{if } i = j \\ \alpha \cdot J_{ij} & ,\text{otherwise} \end{cases} \tag{3-13}$$
$$-\alpha : 0 < \alpha < 1$$

3.3.2 使用者分群

使用者分群的依據是使用者對主題分群間的相似度來衡量。也就是說當使用者和特定主題分群關鍵字有較多的對應關係，代表使用者對此主題的偏好也會較

大，於是本研究透過相似度的計算，將使用者歸類到相似度較高的主題，以達到使用者分群的目的。

將所有使用者向量模型以N×m的矩陣U表示，N代表使用者數目，m代表所有關鍵字數目，矩陣U中每一行可表示個別使用者與關鍵字間的關聯。利用代表使用者間相關係數的矩陣R，乘上以使用者向量模型構成的矩陣U，形成一新的矩陣U'代表更新後的使用者向量模型，如圖3-10所示。

$$\mathrm{U}=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & & & \\ w_{N1} & w_{N2} & \cdots & w_{Nm} \end{bmatrix}$$

$$\mathrm{U'}=\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1N} \\ R_{21} & R_{22} & \cdots & R_{2N} \\ \vdots & & & \\ R_{N1} & R_{N2} & \cdots & R_{NN} \end{bmatrix} \times \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & & & \\ w_{N1} & w_{N2} & \cdots & w_{Nm} \end{bmatrix} = \begin{bmatrix} w'_{11} & w'_{12} & \cdots & w'_{1m} \\ w'_{21} & w'_{22} & \cdots & w'_{2m} \\ \vdots & & & \\ w'_{N1} & w'_{N2} & \cdots & w'_{Nm} \end{bmatrix}$$

圖 3-10 更新使用者向量模型

接著利用餘弦相似度，計算使用者向量模型與個別主題的相似度，如方程式(3-14)所示。當使用者與主題間的相似度大於門檻值時，則認為使用者對此主題具有相當程度的偏好，於是將其歸類到該主題。表3-5為使用者分群之範例。

$$U'_j = (w'_{1j}, w'_{2j}, \ldots, w'_{3j}) \text{ where } j = 1, 2, \ldots, N$$
$$- w'_{ij} : \text{the weight of keyword } i \text{ associated with user } j$$
$$- N : \#\text{ of users}$$
$$- m : \#\text{ of keywords}$$

$$C_k = (t_{k1}, t_{k2}, \ldots, t_{km}) \text{ where } k = 1, 2, \ldots, p$$
$$- t_{km} = \begin{cases} 1, \text{if keyword } i \in C_k \\ 0, \text{otherwise} \end{cases}$$
$$- p : \#\text{ of clusters}$$

$$SU_{jk} = \left\{ sim\left(U_j, C_1\right), sim\left(U_j, C_2\right), \ldots, sim\left(U_j, C_k\right) \right\} \quad (3\text{-}14)$$
$$where\ j = 1, 2, \ldots, N\ ;\ k = 1, 2, \ldots, p$$

表 3-5 使用者分群範例

Topic	Name	Weight
Mobile Computing	Yuh-Shyan Chen	0.7314
	Yen-Ku Liu	0.7291
	Bing-Rong Lin	0.7291
	Chi-He Chang	0.7105
	Chi-Ming Hsieh	0.6993
Genetic Algorithm	Ying-Ping Chen	0.881
	Ming-Chung Jian	0.881
	Wen-Chih Peng	0.7508
	Ying-Hong Liao	0.5302
	Ming-Da Wu	0.4556
Semantic Query	Su-Hsien Huang	0.9532
	Wei-Pang Yang	0.7021
	Hao-Ren Ke	0.6059
	Jen-Yuan Yeh	0.4986
	I-Heng Meng	0.3326
Neural Network	K. T. Sun	0.9476
	J. J. Shann	0.5939
	Jyh-Da Wei	0.4999
	Y. S. Sun	0.4945
	C. -C. Chuang	0.4945
Watershed Segmentation	Yu-Pao Tsai	0.8418
	Yi-Ping Hung	0.8418
	Zen-Chung Shih	0.8384
	Yao-Xun Chang	0.8367
	Cheng-Hung Ko	0.8268
Parallel Algorithm	Jong-Chuang Tsay	0.901
	Yeh-Chin Ho	0.9001
	Pen-Yuang Chang	0.8521
	Sy Yuan	0.7446

3.4 推薦模式

建立使用者社群之後，在個別社群中的成員，都具有相似的主題興趣，但是由於多重主題(Multiple Topics)[9]的屬性存在，使得一個使用者可能對多種主題都具有偏好。表3-6為使用者多重主題之範例。有鑑於此，本研究的推薦模式分為個人化推薦與社群推薦等兩種，茲分述如下：

1. 個人化推薦

依據內容導向方法，對使用者進行論文推薦，即計算社群內成員所撰寫的論文與個別成員的相似度，選取相似度最高的n篇論文給予推薦，同時排除使用者本身所撰寫的論文。

2. 社群推薦

由於多重主題屬性的存在，可以透過分析社群成員對其他主題的興趣分佈，統計出具有較高偏好比重的主題，推薦與該主題最相關的n篇論文給使用者。

表 3-6 使用者多重主題範例

Name	Interest Topic	Preference
Yi-Bing Lin	Mobile Computing	0.5774
	End-to-end Security	0.1713
Ying-Dar Lin	TCP	0.4739
	PIM-SM	0.46
	Network Management	0.3425
	Routing Protocol	0.2674
	Bandwidth Requests	0.2569
Hsin-Chia Fu	SPDNN	0.6769
	Neural Network	0.4515
	Divide-and-conquer Learning	0.2448
Hao-Ren Ke	Semantic Query	0.6059
	Memory Cache	0.4898
	Content Management	0.469
Wen-Chih Peng	Genetic Algorithm	0.7508
	Mobile Computing	0.1628
Yuan-Cheng Lai	TCP	0.6945
	Routing Protocol	0.3787
	Network Management	0.3073

第四章　系統發展與實證分析

本章將敘述雛型系統發展、評估方法與實驗結果，並根據實驗的結果進行討論與分析。首先說明雛型系統之架構及功能介面，接著由專家將語料庫內的作者分類，並以此為標準答案評估將語料庫內的作者根據第三章所提出之分群方法加以分群後的準確率(Precision)與回現率(Recall)[15]；在評估系統推薦資訊的好壞上，使用Kappa Statistics[36]評估專家同意度，以此作為計算推薦準確性之依據；此外也將對使用者分群的結果進行討論，並且根據Matsuo[35]所提出之方法，建構使用者間的社會網路，並對此社會網路結構進行分析。

4.1 系統發展

4.1.1 系統架構

本研究為驗證所提出推薦機制的效果，以NCTUIR之期刊論文為資料來源發展雛型系統。系統架構圖如圖4-1所示。系統經由分析NCTUIR之期刊論文內容，透過主題關鍵字分群萃取出主題概念，以作者間的共同作者關係進行社會網路分析；綜合以上兩者之結果，進行作者分群的動作，形成對特定主題感興趣的主題社群，再針對社群成員進行資訊推薦；推薦模式分為個人化推薦與社群推薦，個人化推薦以使用者本身具偏好的主題進行推薦，社群推薦則是分析社群成員的主題偏好分佈，統計出具有較高偏好比重的主題進行推薦。

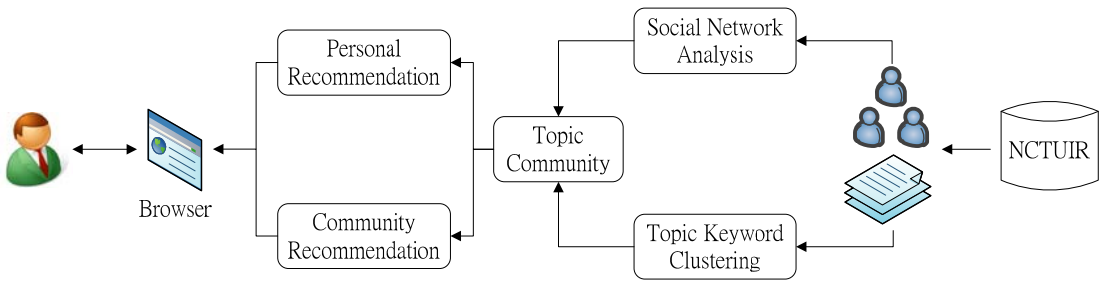

圖 4-1 系統架構圖

4.1.2 系統介面

系統主要提供兩種瀏覽模式，分別是以作者姓名排序瀏覽，以及作者所屬主題社群瀏覽兩種方式，作者瀏覽介面如圖4-2與圖4-3所示。

圖 4-2 依作者姓名排序瀏覽

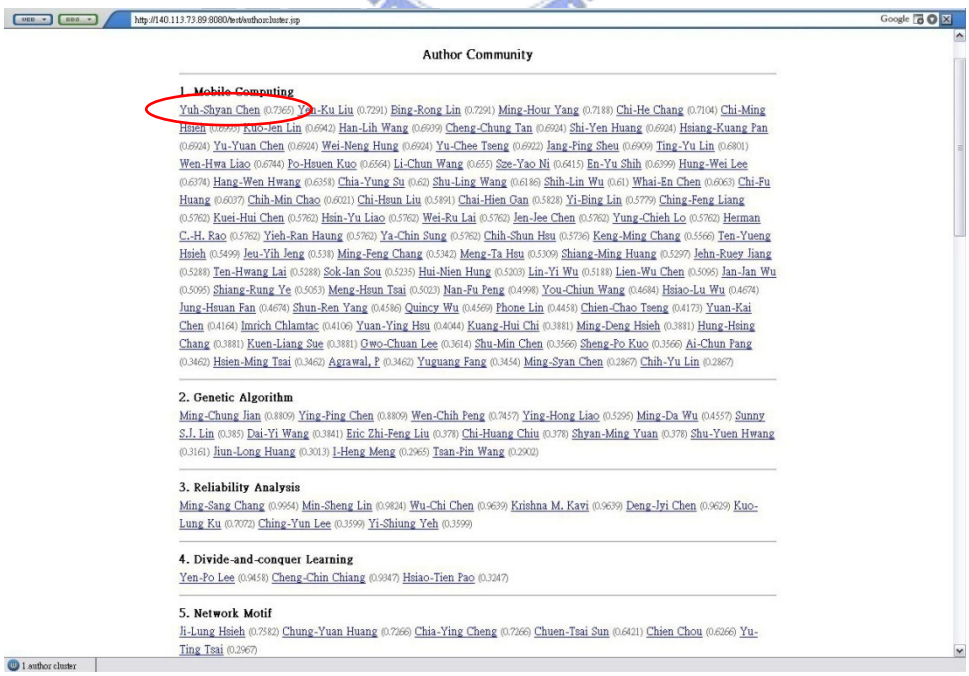

圖 4-3 依作者所屬主題社群瀏覽

點選作者姓名後，可瀏覽作者個人資料，包含作者姓名、感興趣的研究主題、主題社群成員與作者本身發表的論文，研究主題與主題社群皆由系統分析作者偏好與社會網路所產生；於作者感興趣之主題後之括號內為該作者對該主題的偏好值；有標示星號(*)的社群成員代表與該作者具共同作者關係，在社群成員姓名後之括號內為與該作者的相似度。

以圖4-4為例，作者姓名為Wei-Pang Yang；所感興趣的主題為Semantic Query、Content Management以及Content-based Image Retrieval，與個別主題的偏好度依序為0.708、0.5425以及0.3189；該作者與所屬主題社群成員Hao-Ren Ke具有共同作者關係，兩者之相似度為0.6498；該作者發表之論文分別列示共同作者、發表期刊及發表時間。

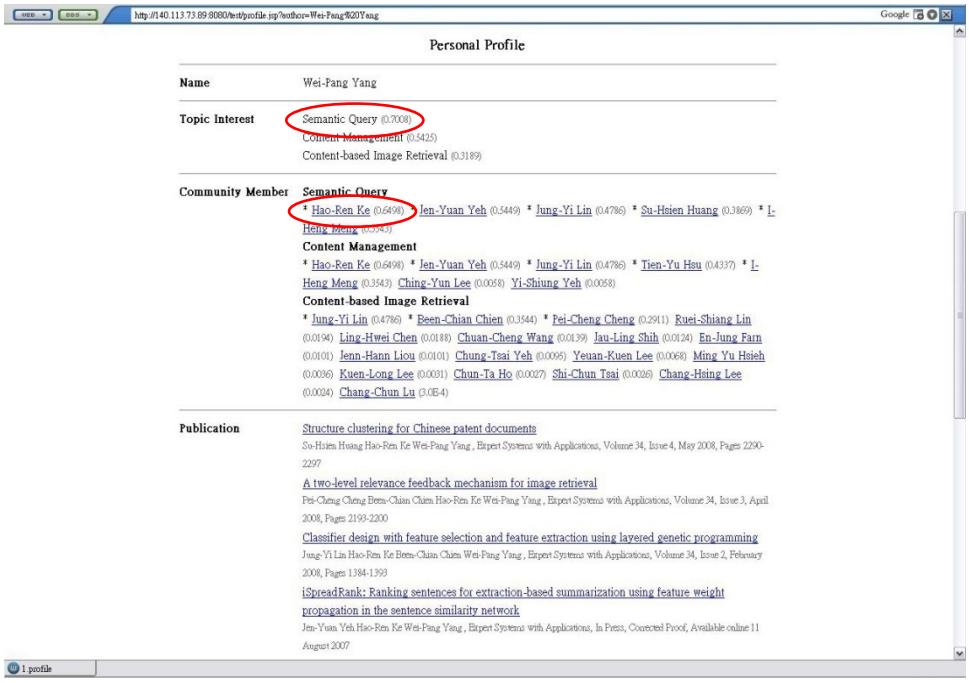

圖 4-4 個人資料介面

除上述之個人資料外，系統也針對作者個人主題偏好進行個人化推薦，以及經由分析社群成員主題偏好分佈，而產生之社群推薦，圖4-5為系統推薦之介面。

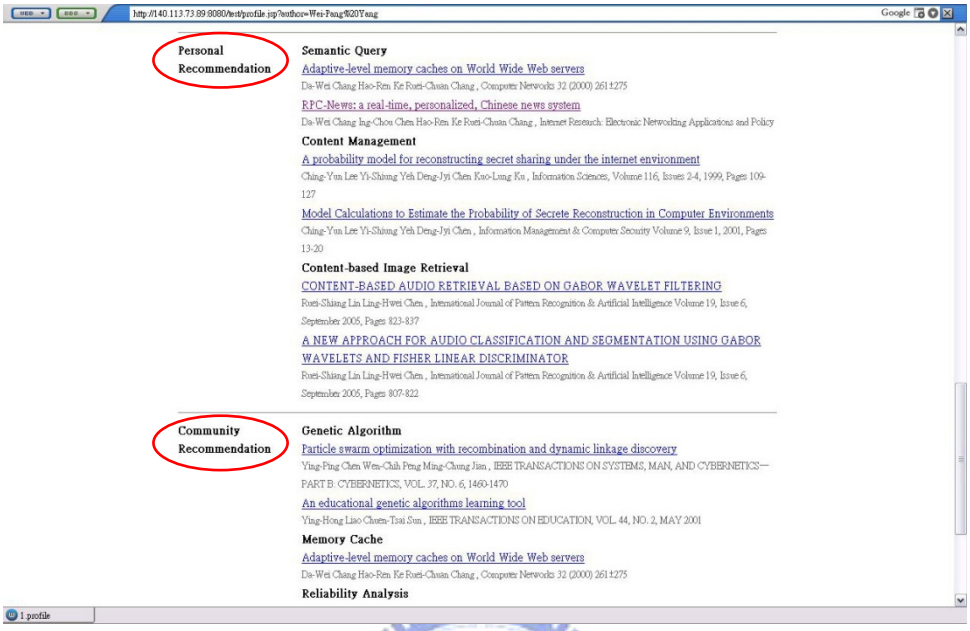

圖 4-5 系統推薦介面

圖4-6為論文文件之詳細內容，於系統中點選論文之標題，即可連結到NCTUIR顯示論文的詳細內容，包含題名、作者、摘要、關鍵字以及發表的期刊與時間等，進一步點選「檢視/開啟」則可以瀏覽論文之全文內容。

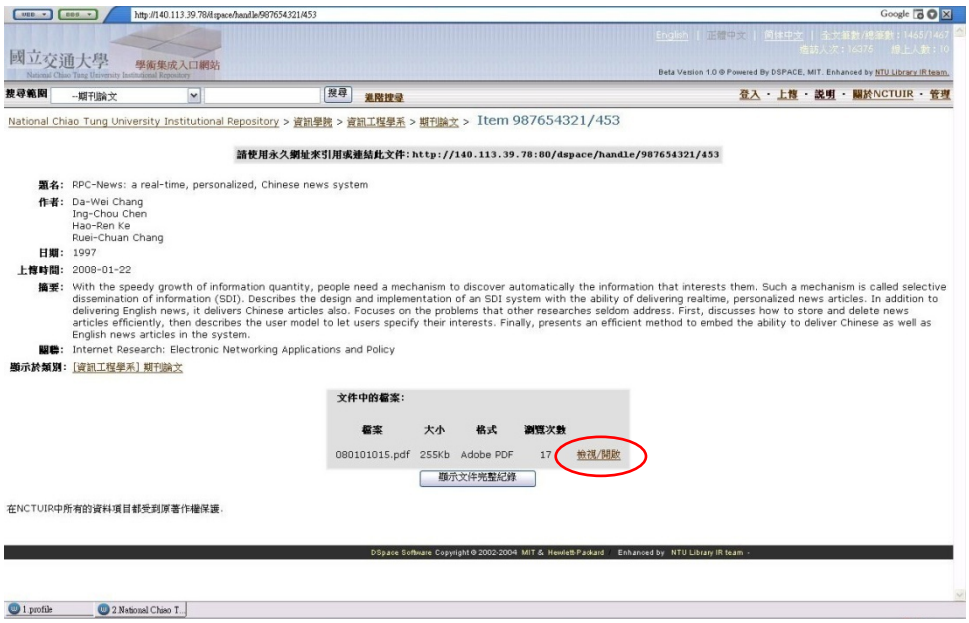

圖 4-6 文件內容介面

4.2 評估方法

4.2.1 以專家評估分群結果

在分群的過程中，使用者不一定只對單一主題感興趣，使用者往往具有多重主題的現象。本研究在使用者分群的過程中，允許多重主題的存在，但是在進行結果評估的過程中，使用者只允許具有單一主題，即與使用者相似度最大的主題分群。

用以評估本研究所提出方法的語料庫為 NCTUIR 內資訊學院的作者與其論文著述，其中包含 235 位作者與 226 篇學術論文，以及 894 對共同作者關係。由於沒有分類的架構，故先請專家將語料庫內的作者分類。首先將系統分群的結果分類，亦即將相近的群歸屬於同一類，總計分成 Network Communication、Artificial Intelligence、Computer Graphics、Information Retrieval、Computer System、Information Security、Graph Theory 及 Software Engineering 等八個類別，類別中的分群皆具有相近的概念，如表 4-1 所示。接著請專家依序對個別使用者進行分類之動作，分類之依據以上述八個類別為主，若使用者皆不屬於上述分類，則將之歸類到「其他」項目中。

表 4-1 主題分群之類別

Class label	Cluster label
Network Communication	Mobile Computing Routing Protocol PIM-SM Bandwidth Requests TCP Network Management
Artificial Intelligence	Genetic Algorithm Network Motif Brick Motif Content Analysis Neural Network SPDNN Divide-and-conquer Learning
Computer Graphics	Content-based Image Retrieval Watershed Segmentation Toboggan Approach
Information Retrieval	Semantic Query Content Management
Computer System	Memory Cache Parallel Algorithm
Information Security	End-to-end Security
Graph Theory	Interconnection Network
Software Engineering	Reliability Analysis

本研究採用資訊檢索中常用的準確率與回現率兩項指標[15]，來評估分群結果的好壞。準確率表示在所檢索出的文章中，相關文章的比例；回現率則表示所有相關的文章中，被檢索出來的比例，其計算方式如方程式(4-1)與(4-2)所示。

$$\text{Precision}=\frac{\text{Relevant} \cap \text{Retrieved}}{\text{Retrieved authors}} \tag{4-1}$$

$$\text{Recall}=\frac{\text{Relevant} \cap \text{Retrieved}}{\text{Relevant authors}} \tag{4-2}$$

4.2.2 以專家評估推薦結果

在進行人工判別實驗結果時，必須評估專家對推薦結果的同意度，通常利用可信度(Reliability)及有效性(Validity)來區別。可信度指專家在評估過程中標示的一致性，而有效性是指專家評估的樣本中可用的樣本數。本研究採用Kappa Statistics[36]來評估專家的同意度，以表4-2為例說明之。假設有40位病人，分別由兩位醫生診斷病情，Yes表示診斷結果為不健康，No則表示健康，則Kappa Statistics計算如下所示。

Kappa=(Observed agreement-Chance agreement)/(1-Chance agreement)
Observed agreement=(10+24)/40=0.85
Chance agreement=0.3×0.35+0.7×0.65=0.56
Kappa=(0.85-0.56)/(1-0.56)=0.659

表 4-2 Kappa Statistics範例

		Doctor A					
		No		Yes		Total	
Doctor B	No	10	(25.0%)	2	(5.0%)	12	(30.0%)
	Yes	4	(10.0%)	24	(60.0%)	28	(70.0%)
Total		14	(35.0%)	26	(65.0%)	40	

並計算標準差得到0.127，當信賴水準(Confidence Level)達95%時，信賴區間(Confidence Interval)為(0.411, 0.907)，將結果對照Kappa的參考表格，如表4-3所示。此範例的Kappa值介於0.61到0.80之間，同意度為Substantial，同意度值為0.85。在評估推薦結果時，則取專家具有一致性意見的樣本，計算資訊推薦的準確率。

表 4-3 Kappa statistics[36]

Kappa	Strength of agreement
0.00	Poor
0.01~0.20	Slight
0.21~0.40	Fair
0.41~0.60	Moderate
0.61~0.80	Substantial
0.81~1.00	Almost perfect

4.3 實驗結果

4.3.1 分群結果評估

本研究在進行分群結果評估時，首先由兩位具有資訊背景的碩士班研究生對分群結果進行分類，再請第三位具有相同背景的研究生，針對上述分類具有不同意見的結果進行分類。最終分類之結果所包含的主題及作者數如表 4-4 所示，235位作者共分為九個類別。

表 4-4 專家分類標示結果

Class label	# of authors
Network Communication	111
Artificial Intelligence	28
Information Retrieval	7
Computer System	6
Computer Graphics	23
Information Security	10
Graph Theory	29
Software Engineering	4
Others	17
Total	235

本研究利用共同作者的社會網路，計算Jaccard coefficient代表使用者間的相關係數，並以α值調整其大小，作為社會網路對使用者偏好之影響(見3.3.1)。實驗中將α值由0至1間做調整，0表示分群過程不考慮使用者社會網路關係；1表示分群過程直接採用Jaccard coefficient為其相關係數。依序計算其準確率與回現率，其結果如表4-5與圖4-7表示。

根據實驗結果，發現使用者社會網路對分群之準確率影響不大，呈現穩定狀態；對於回現率之提升則有較佳之效果，代表其能發掘出更多具有關聯性之使用者。在後續對推薦結果進行評估的實驗中，取α值為0.3作為分群之依據，並以此分群結果進行推薦之動作。

表 4-5 Precision與Recall

α value	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Precision	0.7071	0.6917	0.6981	0.7107	0.7172	0.7209	0.7209	0.7209	0.7209	0.7209	0.7209
Recall	0.6271	0.7606	0.7785	0.7839	0.7817	0.7828	0.7828	0.7828	0.7828	0.7828	0.7828

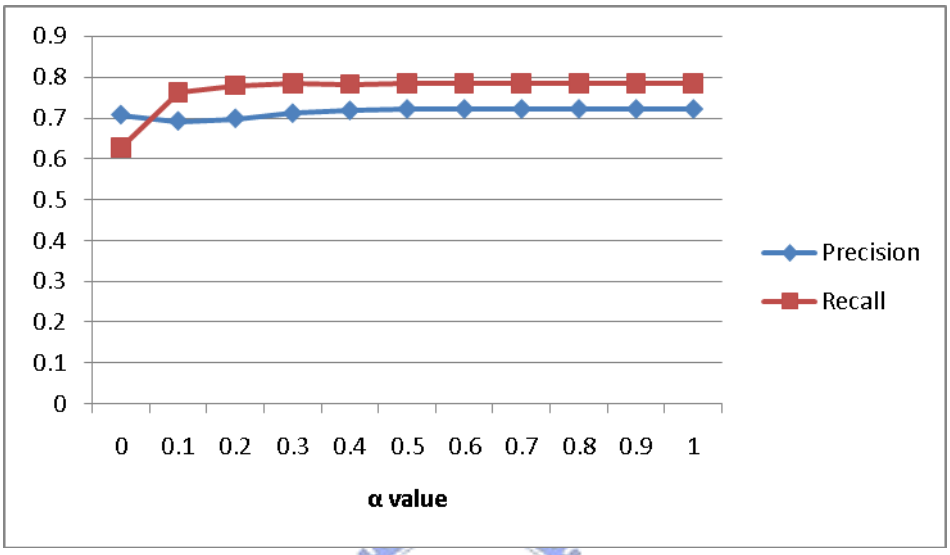

圖 4-7 Precision與Recall

4.3.2 推薦結果評估

本研究在進行推薦結果評估時，由兩位具有資訊背景的碩士班研究生對資訊推薦之結果進行評估，Yes表符合使用者需求之推薦，No表不符合需求之推薦，標示結果如表4-6所示；接著以Kappa Statistics進行同意度分析，針對實驗中有對使用者進行推薦動作的219筆資料，計算其Kappa值為0.764，對照表4-4得知，專家的同意度為Substantial，同意度值為0.95；其標準差為0.068，當信賴水準達95%時，信賴區間為(0.632, 0.897)。

表 4-6 專家評估推薦結果

		Expert A					
		No		Yes		Total	
Expert B	No	21	(9.6%)	9	(4.1%)	30	(13.7%)
	Yes	2	(0.9%)	187	(85.4%)	189	(86.3%)
Total		23	(10.5%)	196	(89.5%)	219	

Kappa=(Observed agreement-Chance agreement)/(1-Chance agreement)
Observed agreement=(21+187)/219=0.949
Chance agreement=0.105×0.137+0.895×0.863=0.786
Kappa=(0.949-0.786)/(1-0.786)=0.764

針對專家具有相同意見之推薦結果，總共有208筆，計算其推薦之準確率。評估結果認為符合使用者需求之推薦有187筆，推薦之準確率為187/208=0.899，顯見系統之推薦效果，頗能符合使用者需求。

4.4 討論與分析

4.4.1 社會網路分析

本研究採用NCTUIR內資訊學院的作者與其論文著述為資料來源，其中包含235位作者與226篇學術論文，以及894對共同作者關係。圖4-8為依據所收錄的論文篇數所作之統計，論文收錄的篇數介於1篇到41篇，只收錄1篇文章的作者有129位，佔全部作者的55%；收錄少於5篇的作者有93%。表4-7為收錄論文篇數大於5篇的作者。

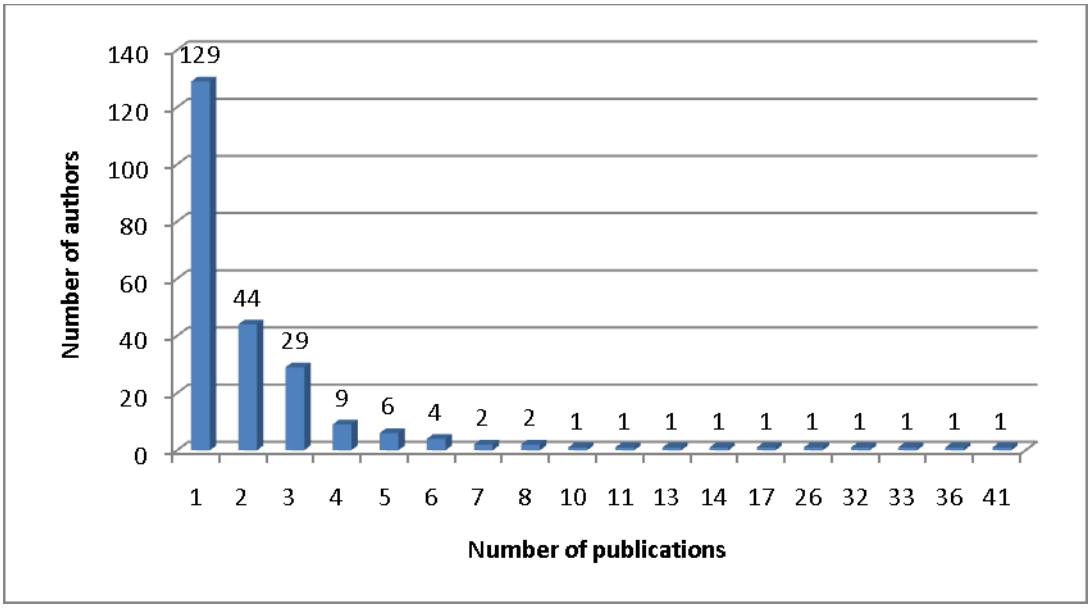

圖 4-8 收錄論文統計

表 4-7 收錄論文數大於5篇之作者

Name	Publications
Yu-Chee Tseng	41
Jimmy J. M. Tan	36
Lih-Hsing Hsu	33
Yi-Bing Lin	32
Ying-Dar Lin	26
Ling-Hwei Chen	17
Chuen-Tsai Sun	14
Jang-Ping Sheu	13
Hsin-Chia Fu	11
Hao-Ren Ke	10
Wei-Pang Yang	8
Wen-Guey Tzeng	8
Chien-Chao Tseng	7
Tseng-Kuei Li	7
Wen-Chih Peng	6
Chang-Hsiung Tsai	6
Deng-Jyi Chen	6
Yuan-Cheng Lai	6

共同作者的統計數據如圖4-9所示。共同作者數介於1到6位作者之間，由單一作者所發表的論文只有6篇，佔全部論文數的3%；由2到6位作者所共同發表的論文篇數，共有220篇，佔全部的97%。由此項統計資料顯示，由共同作者發表論文著述在NCTUIR資訊學院中是普遍存在的現象。

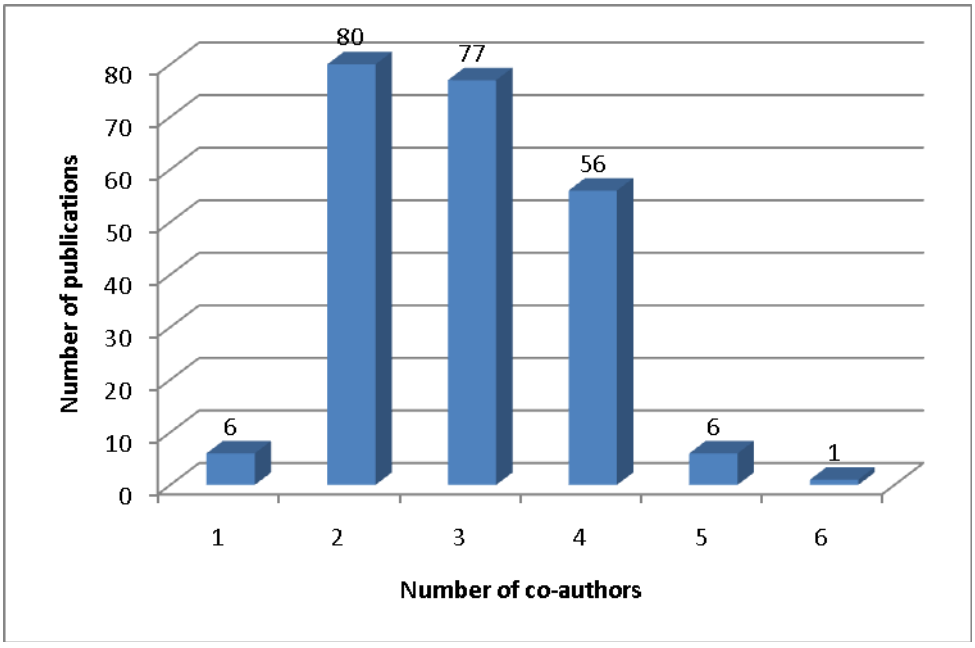

圖 4-9 共同作者統計

由共同作者關係所建構之社會網路如圖4-10所示。共同作者的社會網路並非只形成一個單一的連通圖(Connected Graph)，而是由大小不同的網路元件(Component)所組成；其中最大的一個元件包含89位作者，佔所有作者的38%，程度中心性較高者為Yu-Chee Tseng與Yi-Bing Lin，如圖4-11所示；最小的一個元件只包含4位作者，只佔所有作者的1.7%，該社會網路之網路中心為Jong-Chuang Tsay，如圖4-12所示。

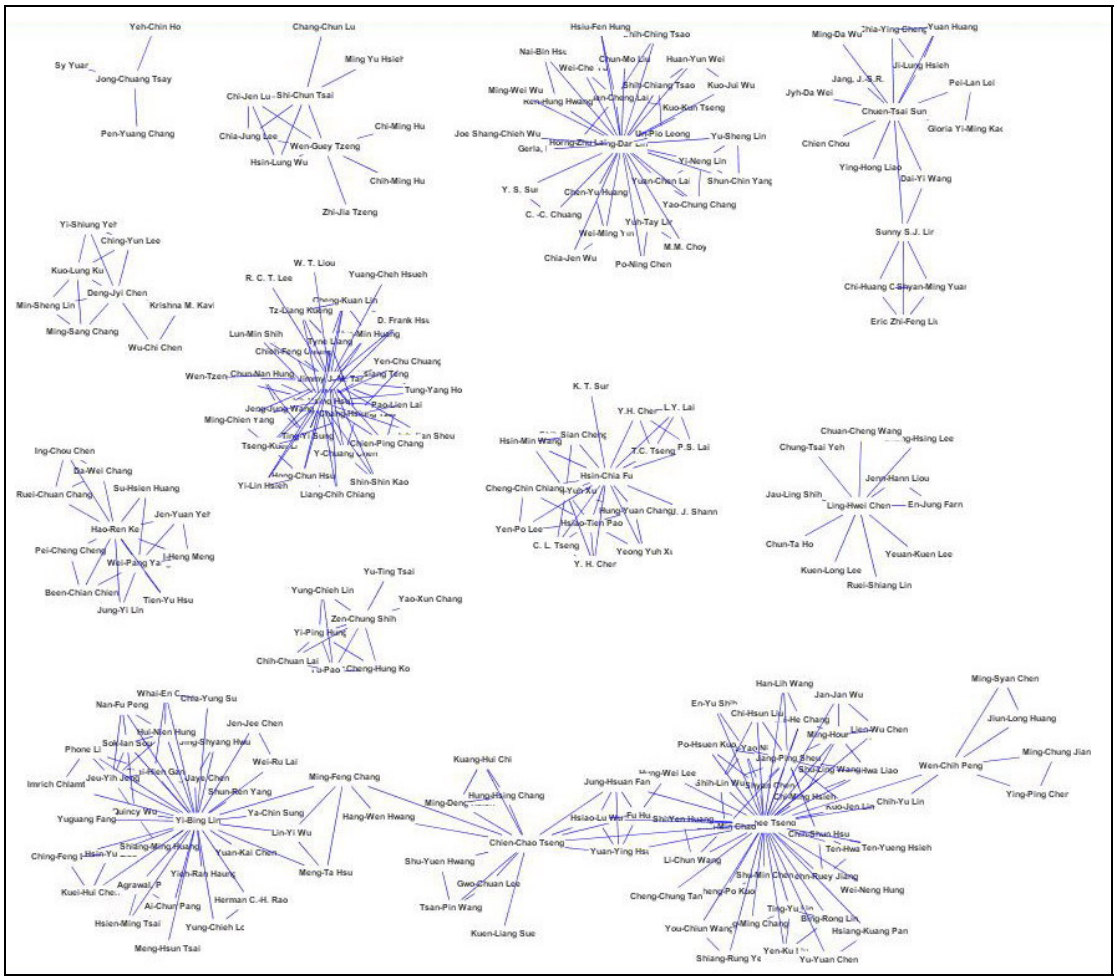

圖 4-10 共同作者社會網路

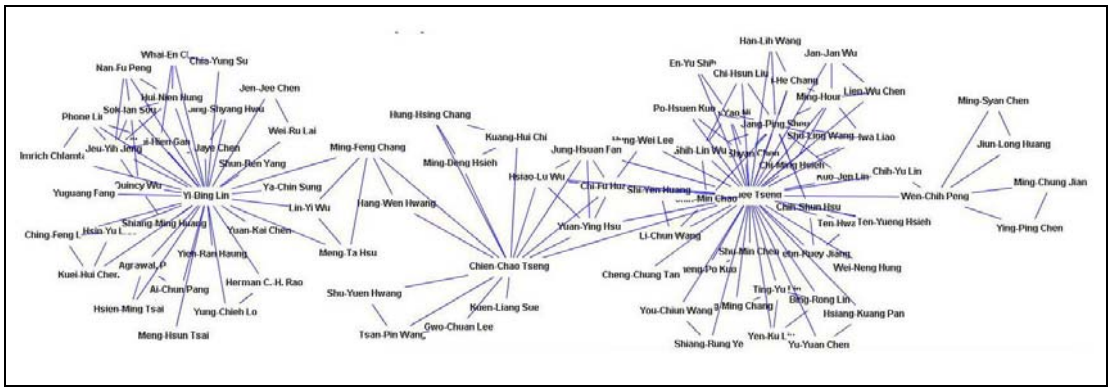

圖 4-11 最大網路元件

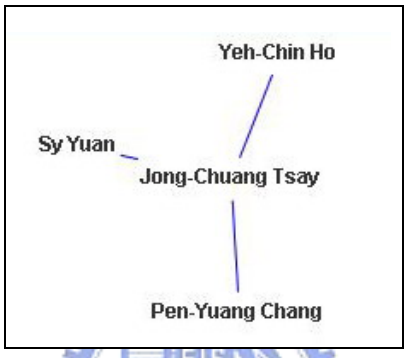

圖 4-12 最小網路元件

針對共同作者社會網路，分別計算此社會網路之程度中心性、中介中心性與緊密中心性。其結果如表4-8所示。程度中心性愈高，則代表其為網路中心，並與其他作者所建立的連結多，在網路中較為活躍；中介中心性愈高，代表其為網路中之橋樑，扮演溝通聯繫的角色；緊密中心性愈高，代表與其他作者間距離愈短，表示其能較快取得資訊。

表 4-8 中心性分析

Rank	Degree		Betweenness		Closeness	
1	Yu-Chee Tseng	43	Yu-Chee Tseng	2660.333	Yu-Chee Tseng	0.678
2	Yi-Bing Lin	32	Chien-Chao Tseng	2180.500	Chien-Chao Tseng	0.678
3	Ying-Dar Lin	29	Yi-Bing Lin	2081.333	Ming-Feng Chang	0.678
4	Jimmy J. M. Tan	29	Ming-Feng Chang	1792.000	Chi-Fu Huang	0.677
5	Lih-Hsing Hsu	26	Ying-Dar Lin	376.500	Hsiao-Lu Wu	0.677
6	Hsin-Chia Fu	16	Wen-Chih Peng	340.000	Yuan-Ying Hsu	0.677
7	Jang-Ping Sheu	15	Jimmy J. M. Tan	213.167	Jung-Hsuan Fan	0.677
8	Chien-Chao Tseng	14	Lih-Hsing Hsu	133.167	Yi-Bing Lin	0.677
9	Chuen-Tsai Sun	12	Chuen-Tsai Sun	91.000	Hang-Wen Hwang	0.677
10	Hao-Ren Ke	11	Hsin-Chia Fu	86.000	Jang-Ping Sheu	0.677
11	Ling-Hwei Chen	10	Ling-Hwei Chen	44.000	Wen-Chih Peng	0.677
12	Wei-Pang Yang	8	Jang-Ping Sheu	38.333	Meng-Ta Hsu	0.677
13	Hsiao-Tien Pao	8	Sunny S.J. Lin	36.000	Lin-Yi Wu	0.677
14	Zen-Chung Shih	7	Hao-Ren Ke	32.833	Ming-Hour Yang	0.677
15	Chang-Hsiung Tsai	7	Chi-Fu Huang	22.500	Chih-Yu Lin	0.677
16	Jeu-Yih Jeng	7	Wen-Guey Tzeng	21.333	Sze-Yao Ni	0.677
17	Yeong-Yuh Xu	7	Shi-Chun Tsai	15.333	Wen-Hwa Liao	0.677
18	Deng-Jyi Chen	7	Deng-Jyi Chen	12.000	Shih-Lin Wu	0.677
19	Wen-Guey Tzeng	7	Zen-Chung Shih	12.000	Chih-Shun Hsu	0.677
20	Ming-Hour Yang	7	Wei-Pang Yang	8.833	Chi-He Chang	0.677

Matsuo[35]於研究中認為，當使用者間具有某種程度上的情境相似度(Context Similarity)，則可視為具有潛在的社會網路關係。依據這個論點，計算使用者間的相似度，當彼此的相似度大於門檻值時，則建立其社會網路關係；同時計算程度中心性、中介中心性與緊密中心性，將各主題分群內各中心性指標最高的作者列出，其結果如表4-9所示。因此本研究認為在各主題分群中，程度中心性較高之作者，為該領域中之專家。

表 4-9 主題分群之中心性分析

Cluster	Degree		Betweenness		Closeness	
Mobile Computing	Yu-Chee Tseng	42	Chien-Chao Tseng	1092.04	Ming-Syan Chen	0.641
Interconnection Network	Jimmy J. M. Tan	29	Jimmy J. M. Tan	38.04	Yuang-Cheh Hsueh	0.485
Routing Protocol	Ying-Dar Lin	21	Ying-Dar Lin	201.50	Ming-Wei Wu	0.469
PIM-SM	Ying-Dar Lin	21	Ying-Dar Lin	201.50	Ming-Wei Wu	0.469
TCP	Ying-Dar Lin	21	Ying-Dar Lin	201.50	Ming-Wei Wu	0.469
Network Management	Ying-Dar Lin	21	Ying-Dar Lin	201.50	Yuh-Tay Lin	0.429
Neural Network	Hsin-Chia Fu	14	Hsin-Chia Fu	69.00	Jyh-Da Wei	0.427
SPDNN	Hsin-Chia Fu	14	Hsin-Chia Fu	69.00	J. J. Shann	0.456
Divide-and-conquer Learning	Hsiao-Tien Pao	10	Hsiao-Tien Pao	28.00	Yen-Po Lee	0.456
Content-based Image Retrieval	Ling-Hwei Chen	10	Ling-Hwei Chen	34.50	Ming Yu Hsieh	0.442
Semantic Query	Hao-Ren Ke	9	Hao-Ren Ke	24.20	Su-Hsien Huang	0.446
Memory Cache	Hao-Ren Ke	9	Hao-Ren Ke	24.20	Da-Wei Chang	0.446
Content Management	Hao-Ren Ke	9	Hao-Ren Ke	24.20	Ching-Yun Lee	0.439
Reliability Analysis	Min-Sheng Lin	7	Min-Sheng Lin	1.00	Ching-Yun Lee	0.439
Watershed Segmentation	Zen-Chung Shih	7	Zen-Chung Shih	6.50	Yu-Ting Tsai	0.439
Toboggan Approach	Zen-Chung Shih	7	Zen-Chung Shih	6.50	Yao-Xun Chang	0.439
End-to-end Security	Wen-Guey Tzeng	7	Wen-Guey Tzeng	20.00	Jing-Shyang Hwu	0.427
Genetic Algorithm	I-Heng Meng	6	Sunny S.J. Lin	8.00	Tsan-Pin Wang	0.427
Network Motif	Chuen-Tsai Sun	6	Chuen-Tsai Sun	8.33	Ji-Lung Hsieh	0.437
Brick Motif Content Analysis	Chuen-Tsai Sun	6	Chuen-Tsai Sun	8.33	Gloria Yi-Ming Kao	0.437
Bandwidth Requests	Shih-Chiang Tsao	3	Chen-Yu Huang	21.00	Wei-Ming Yin	0.469
Parallel Algorithm	Jong-Chuang Tsay	3	Jong-Chuang Tsay	0.00	Jong-Chuang Tsay	0.431

第五章　結論與建議

5.1 結論

本研究致力於改善資訊推薦的效能，主要的目在於提出結合主題概念萃取與社會網路分析之資訊推薦系統，以提供符合使用者需求之推薦資訊。本研究首先利用關鍵字分群，萃取出使用者感興趣的主題概念，接著分析共同作者關係，建構社會網路以形成主題社群，經由主題社群的產生，分析社群成員的興趣與喜好，以預測目標使用者的潛在偏好，建構出更符合使用者需求的推薦系統，以提升資訊推薦的品質。經由實驗與統計分析的驗證，將本研究的結果整理如下：

1. 主題概念萃取

將經過特徵選擇的關鍵字建立關係，根據共現原則建立相關度，並且產生語意網路，由語意網路圖找出重要的核心概念，採用k-Nearest Neighbor Approach演算法找出核心，同時考慮分群結果的平衡性，以最適當的數個關鍵字表現主題概念。在所有235位作者，226篇論文中，共產生22個主題概念。

2. 形成主題社群

萃取出主題概念後，利用餘弦相似度計算使用者向量與各主題分群的相似度，進而產生對特定主題感興趣的使用者社群。在建立主題社群的同時，不僅僅專注於個別使用者的關鍵字分佈與主題興趣，也把社會網路的互動關係及其影響考慮進來，經由分析使用者間共同作者關係建立社會網路，以Jaccard coefficient衡量使用者間相關係數，調整使用者與關鍵字的關聯。經由實驗發現，社會網路對使用者分群之準確率影響不大，呈現穩定狀態；對於回現率之提升則有較佳之效果，代表其能發掘出更多具有關聯性之使用者。

3. 資訊推薦

進行資訊推薦時，以使用者感興趣的主題，進行內容導向式推薦；由於多重主題屬性的存在，透過分析社群成員對其他主題的興趣分佈，統計出具

有較高偏好比重的主題，以此對社會成員進行推薦。本研究採用Kappa Statistics評估專家對推薦結果的同意度，同時考慮可信度及有效性。實驗結果推薦之準確率為0.899，顯見系統之推薦效果，頗能符合使用者需求。

經由討論與分析，在所採用之NCTUIR語料庫中針對資訊學院進行共同作者社會網路分析，發現由兩位作者以上所發表的論文篇數佔所有論文的97%，由此項統計資料顯示，以共同作者發表論文著述在NCTUIR資訊學院中是普遍存在的現象。又分別計算程度中心性、中介中心性與緊密中心性，發現在各主題分群中，程度中心性較高之作者，可視為該領域中之專家。

5.2 後續建議

推薦系統的發展隨著應用技術與領域的不同而不斷演進，本研究對於將社會網路應用於推薦系統做了深入的研究，然而由於系統實驗仍有其限制，後續仍然有許多地方值得探討，以下針對後續研究方向提出建議：

1. 建立主題知識本體(Ontology)

在進行主題萃取的過程中，利用階層式分群法以樹狀結構表示主題分群之結果，產生主題概念階層，可發掘主題概念之從屬關係。經由使用者主題偏好之關聯，建立主題概念之連結，以形成主題之知識本體，可幫助使用者瞭解本身處於何種階層層級，未來可朝那些研究方向前進。

2. 使用者評分之應用

使用者評分可分為明顯性評分與隱含性評分。明顯性評分為使用者依對目標物感興趣程度給予主觀評分，例如要求使用者在閱讀完文件後給予評分；隱含性評分的估計通常以使用者的瀏覽行為做依據，例如所花費的閱讀時間以及滑鼠的操作等。經由使用者評分可以更精確瞭解使用者偏好所在，使資訊推薦更符合使用者需求。

3. 社會網路之階層擴展

社會網路有助於社群之形成，本研究著眼於與使用者本身有直接關聯的

社會網路關係，未來可經由建立在共同社會網路中之使用者關係，進一步探討社會網路之資訊流動及影響。例如使用Floyd-Warshall演算法可找出位於同一社會網路中，兩兩使用者間的最短路徑，則可經由節點的分析，研究其對使用者的影響。

參考文獻

[1] A. Iskold, (2007) “The Art, Science and Business of Recommendation Engines.” http://www.readwriteweb.com/archives/recommendation_engines.php

[2] A. K. Jain, M. N. Murty, & P. J. Flynn, “Data clustering: A review,” *ACM Computing Surveys*, vol. 31, pp. 264-323, 1999.

[3] B. Krulwich, & C. Burkey, “The InfoFinder agent: Learning user interests through heuristic phrase extraction,” *IEEE Expert: Intelligent Systems and Their Applications*, vol. 12, pp. 22-27, 1997.

[4] B. Sarwar, G. Karypis, J. Konstan, & J. Riedl, “Analysis of recommendation algorithms for e-commerce,” *Proceedings of the 2nd ACM conference on Electronic commerce*, pp. 158-167, 2000.

[5] D. Goldberg, D. Nichols, B. M. Oki, & D. Terry, “Using Collaborative Filtering to Weave An Information Tapestry,” *Communications of the ACM*, vol. 35, pp. 61-70, 1992.

[6] D. Koller, & M. Sahami, “Hierarchically classifying documents using very few words,” *Proceedings of 14th the International Conference on Machine Learning*, pp.170–178, 1997.

[7] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, pp. 1-47, 2002.

[8] G.. Karypis, E. H. Han, & V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” *Computer*, vol. 32, pp. 68-75, 1999.

[9] H. C. Chang, & C. C. Hsu, “Using topic keyword clusters for automatic document clustering,” *Transactions on Information and Systems*, vol. 88, pp. 1852-1860, 2005.

[10] H. Hotta , “User profiling system using social networks for recommendation”, *In Proceedings of 8th International Symposium on Advanced Intelligent Systems* , 2007.

[11] H. Kautz, B. Selman, & F. Park, “Referral Web: Combining social networks and collaborative filtering,” *Communications of the ACM,* vol. 40 , pp. 63-65, 1997.

[12] H. Sakagami, & T. Kamba, “Learning Personal Preferences on Online Newspaper Articles from User Behaviors,” *Computer Networks and ISDN Systems*, vol. 29, pp. 1447-1455, 1997.

[13] J. B. Schafer, J. Konstan, & J. Riedi, “Recommender systems in e-commerce,” *Proceedings of the 1st ACM conference on Electronic commerce*, pp. 158-166, 1999.

[14] J. MacQueen, “Some methods for classification and analysis of multivariate

observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281-297, 1967.

[15] J. Makhoul, F. Kubala, R. Schwartz, & R. Weischedel, "Performance measures for information extraction," *Proceedings of DARPA Broadcast News Workshop*, pp. 249-252, 1999.

[16] J. Moreno, *Who Shall Survive?* New York: National Institute of Mental Health, 1934.

[17] J. R. Tyler, D. M. Wilkinson, & B. A. Huberman, "Email as spectroscopy: Automated discovery of community structure within organizations," *Communities and technologies*, pp. 81-96, 2003.

[18] J. Rucker, & M. J. Polanco, "Siteseer: Personalized navigation for the web," *Communications of the ACM*, vol. 40, pp. 73-76, 1997.

[19] K. C. Gowda, & G. Krishna, "Agglomerative clustering using the concept of mutual nearest neighbourhood," *Pattern Recognition*, vol. 10, pp. 105-112, 1978.

[20] K. Faust, "Comparison of methods for positional analysis:Structural and general equivalences," *Social Networks*, vol. 10, pp.313-341, 1988.

[21] L. C. Freeman, "Centrality in Social Networks: Conceptual clarification," *Social Networks*, vol. 1, pp. 215-239, 1979.

[22] L. Page, & S. Brin, "The anatomy of a large-scale hypertextual Web search engine," In Proceedings of the seventh international World-Wide Web conference, 1998.

[23] L. Garton, C. Haythornthwaite, & B. Wellman, (1997) "Studying Online Social Networks," http://jcmc.huji.ac.il/vol3/issue1/garton.html

[24] M. A. Shah, "ReferralWeb: A resource location system guided by personal relations," Master's thesis, M.I.T., 1997.

[25] M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociology Theory*, vol. 1, pp. 201-233, 1983.

[26] N. Zhong, J. Liu & Y. Yao, "In search of the wisdom web," *Computer*, vol. 35, pp. 27-31, 2002.

[27] P. Athanasios, *Probability, Random Variables and Stochastic Processes.*, Second Edition ed. New York: McGraw-Hill, 1984.

[28] P. Mika, "Flink: Semantic Web technology for the extraction and analysis of social networks," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol 3, pp. 211-223, 2005.

[29] P. Pattison, *Algebraic models for social networks.*, Cambridge University Press, 1993.

[30] S. E. Chan, R. K. Pon, & A. F. Cárdenas, "Visualization and Clustering of Author Social Networks," *International Conference on Distributed Multimedia Systems*

Workshop on Visual Languages and Computing, pp. 30-31, 2006.
[31] S. P. Borgatti, (1998) "What Is Social Network Analysis?" http://www.analytictech.com/networks/whatis.htm
[32] S. Staab, P. Domingos, P. Mike, J. Golbeck, D. Li, T. Finin, A. Joshi, A. Nowak, & R. R. Vallacher, "Social networks applied," *IEEE Intelligent Systems*, vol. 20, pp. 80-93, 2005.
[33] V. Kotlyar, M. S.Viveros, S. S. Duri, R. D. Lawrence, & G. S. Almasi, "A case study in information delivery to mass retail markets," *In Proceedings of the 10th International Conference on Database and Expert Systems*, 1999.
[34] X. Liu, J. Bollen, M. L. Nelson, & H. V. de Sompel, "Co-authorship networks in the digital library research community," *Information Processing and Management*, vol 41, pp. 1462-1480 , 2005.
[35] Y. Matsuo, J. Mori, & M. Hamasaki, "POLYPHONET: An advanced social network extraction system from the Web ", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, pp. 262-278, 2007.
[36] Kappa Statistics - http://www.dmi.columbia.edu/homepages/chuangj/kappa
[37] LingPipe NLP Toolkit - http://alias-i.com/lingpipe/
[38] NCTUIR - http://ir.lib.nctu.edu.tw/
[39] Porter Stemming Algorithm - http://tartarus.org/~martin/PorterStemmer
[40] D. J. Watts 著, 傅士哲, 謝良瑜譯, "6 個人的小世界" 大塊文化, 2004.
[41] 楊永芳, "語意擴充式文件推薦方法之研究" 中山大學資訊管理研究所碩士論文, 2001.
[42] 翁明正, 翁頌舜, "應用本體論與社會網路分析於個人化推薦" 2007 數位科技與創新管理研討會論文集, 423-434, 2007.