國立交通大學

管理學院資訊管理學程碩士班

碩士論文

企業部落格文章之自動化分類與推薦

Automatic Classification and Recommendation of Enterprise Blog Articles

研究生：蔡曉菁

指導教授：劉敦仁博士

中華民國九十九年七月

企 業 部 落 格 文 章 之 自 動 化 分 類 與 推 薦

Automatic Classification and Recommendation of Enterprise Blog Articles

研 究 生：蔡曉菁　　Student: Shiao-Jing Tsai

指導教授：劉敦仁　　Advisor: Dr. Duen-Ren Liu

國立交通大學

管理學院資訊管理學程碩士班

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

July 2010

Hsinchu, Taiwan, the Republic of China

中 華 民 國 九 十 九 年 七 月

研究生：蔡曉菁　　指導教授：劉敦仁 博士

國立交通大學管理學院資訊管理學程碩士班

摘要

企業競爭已經進入以知識為基礎之後知識經濟時代，不斷創新及持續學習成為組織決勝的關鍵。建立企業組織知識價值觀並透過良好的知識分享和知識管理的機制，才能有效提昇組織價值和能力。

近年來，文件自動化分類與推薦方法已成功應用於知識管理和知識分享，自動化分類文件可幫助文件有效管理與取用；透過分析使用者興趣，預測使用者對知識的偏好，主動推薦使用者有興趣、相關知識文件，可加速組織內知識的傳遞、分享與使用。

本研究針對企業內部部落格文章設計分類與推薦兩大機制。分類是透過文件與類別特徵檔相似度概念進行分類；推薦機制包含社群分析與文件推薦兩模組，社群分析考量使用者興趣與個人屬性檔相關性，進行使用者多社群分析。文件推薦是結合多社群與混合式過濾推薦，並使用四種推薦方法「社群最大分數法」、「社群興趣最大分數法」、「使用者權重最大分數法」、「使用者權重興趣最大分數法」。依比例混合之內容式推薦（使用者對文章興趣程度）與協同式推薦（鄰居使用者是否點閱該文章、與目標使用者相似度、使用者權重、鄰居使用者對文章興趣指數）得出文件預測興趣分數、推薦清單。

實驗結果顯示，本研究所提分類法可有效率將文件自動化分類，且社群分析時考量使用者相關性可有效提高文件推薦準確率。

關鍵字：知識管理、分類、虛擬社群、推薦

Automatic Classification and Recommendation of Enterprise Blog Articles

Student: Shiao-Jing Tsai Advisor: Dr. Duen-Ren Liu

Master Program of Institute of Information Management

College of Management

National Chiao Tung University

Abstract

In the area of knowledge economics, enterprises need innovation and learning ability to compete. Efficient knowledge sharing and management mechanism can help organizations to gain value and competitive advantages. Recently, Automatic document classification and personalized recommendation methods have been applied to knowledge management and sharing. Efficient recommendation based on knowledge preferences can speed up knowledge spread in organizations.

This research proposes document classification and recommendation mechanisms for enterprise blog articles. Blog articles are classified based on the similarity of article profiles and category profiles. The proposed recommendation mechanisms cluster users into multi-communities based on users' interests and personal attributes. Based on the multi-communities of users, four hybrid approaches which combine community-based approach, content-based filtering and collaborative filtering, are proposed to make article recommendations. Experimental evaluations are conducted to evaluate the effectiveness of the proposed approaches.

Keywords：Knowledge Management, Classification, Community, Recommendation

誌謝

首先，感謝劉敦仁教授的細心指導，在論文研究過程中，犧牲許多休假時間指導我們專班學生論文研究，給予許多建議與方向，讓我能順利口試過關，完成碩士學位。另外，感謝口試委員王朝煌教授、李瑞庭教授、羅濟群教授的寶貴意見指導，讓我的論文研究更加完整、完善。

這兩年的在職專班進修雖然辛苦，期間學到許多新的知識、認識許多不同領域同學，非常感謝實驗室同學獻祥、居逸，論文研究緊張之餘總是有你們笑話紓壓，一起努力、口試過關；還有瓊雯、秋霞、子翔的陪伴，每當遇到問題時，大家都能一起討論、互相幫忙、彼此激勵，謝謝你們豐富了我這兩年的求學生涯，給予我許多珍貴意見。

還要感謝在求學期間，職場上協助我的好同事們，沒有你們的協助，我無法在兩年的時間內就能順利取得碩士學位，還能同時兼顧工作。

最後，感謝我的家人、朋友一直默默的支持和陪伴，還有男友志昇這段時間載我桃園、新竹兩頭跑，犧牲休假時間陪我一起閉關做論文，鼓勵我、爲我打氣，讓我能夠順利完成兩年的碩班學位。

蔡曉菁　2010/07

目錄

1. 緒論 1

1.1. 研究背景 1

1.2. 研究動機 2

1.3. 研究目的 3

1.4. 論文組織架構 4

2. 文獻探討 5

2.1. 資訊檢索與資訊過濾 5

2.1.1. 以文件為基礎之資訊檢索與資訊過濾 5

2.1.2. 關鍵字權重 6

2.2. 分類 7

2.3. 社群 8

2.4. 分群方法 9

2.5. 推薦方法 10

2.6. 企業部落格 13

3. 企業部落格之自動化分類與推薦方法 15

3.1. 企業部落格之自動化分類與推薦方法之架構 15

3.2. 文件前置處理、文件特徵檔分析 17

3.3. 分類 21

3.3.1. 類別特徵檔分析 21

3.3.2. 文件、類別相似度計算 22

3.4. 社群分析 23

3.4.1. 使用者特徵檔分析 24

3.4.2. 使用者相似度計算 25

3.4.3. 知識社群分群 29

3.5. 推薦 32

3.5.1. 多社群文件推薦 32
3.5.2. 社群最大分數法（Community Max Score - CMS） 34
3.5.3. 社群興趣最大分數法（Community Interest Max Score - CIMS） 36
3.5.4. 使用者權重最大分數法（User Weight Max Score - UWMS） 38
3.5.5. 使用者權重興趣最大分數法（User Weight Interest Max Score - UWIMS） 40
4. 實驗與評估 43
4.1. 實驗資料 43
4.2. 評估標準 43
4.3. 實驗工具 44
4.4. 實驗結果與評估 45
4.4.1. 分類 45
4.4.2. 推薦 47
5. 結論與未來方向 56
5.1. 結論 56
5.2. 未來方向 57
REFERENCE 59

表目錄

表 2–1 知識部落格應用模組 14

表 3–1 文件特徵檔 20

表 3–2 文件 Di 和 類別 Cj 之特徵檔範例 23

表 3–3 使用者特徵檔轉換 26

表 3–4 使用者特徵檔範例 27

表 3–5 使用者 u1 和 u2 的特徵檔範例 28

表 3–6 知識工作者相似度矩陣 30

表 3–7 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表 36

表 3–8 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表 38

表 3–9 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表 40

表 3–10 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表 42

表 4–1 部落格有效分類表 45

表 4–2 分類實驗結果 46

表 4–3 推薦實驗資料 47

圖目錄

圖 2–1 兩文件在空間向量中的表示 6

圖 3–1 企業部落格之自動化分類與推薦方法整體架構 15

圖 3–2 企業部落格之自動化分類方法之架構 16

圖 3–3 企業部落格之自動化推薦方法之架構 17

圖 3–4 文件前置處理模組 18

圖 3–5 實作斷詞功能系統 19

圖 3–6 文件特徵檔分析流程 22

圖 3–7 社群分析流程 24

圖 3–8 使用者特徵檔分析流程 25

圖 3–9 社群特徵檔分析流程 31

圖 4–1 Precision–CMS 48

圖 4–2 Precision–CIMS 50

圖 4–3 Precision–UWMS 51

圖 4–4 Precision–UWIMS 53

圖 4–5 Precision - Total 55

圖 4–6 Precision – 4 methods comparison 55

1. 緒論

1.1. 研究背景

對於企業組織而言，員工是公司最重要的資產，組織智慧是來自於員工知識創造，故企業如何使員工主動將其內涵之知識萃取出來，形成知識創造、分享的機制，為組織智慧資產累積的重要關鍵。且知識分享的平台除了單向的貢獻知識內容外，最重要的是更希望組織其他使用者能夠透過平台的互動、討論、評價，透過這樣的互動過程，才能建立組織中知識的流通和知識分享的循環機制。故企業開始導入 Web 2.0 之概念，企圖藉由開放性、個人化之企業內部部落格平台，引發員工自發性的紀錄個人工作歷程、專業知識、個人相關喜好…等，亦或以組織為單位之組織部落格、工作業務導向之 CFT（Cross Function Team）部落格、興趣社群（社團）部落格，加深組織內部互動與員工向心力。

隨著資訊科技快速的進步，組織中資訊累積的速度不斷快速增加，企業組織中的知識內容也隨著快速巨量累積。對組織競爭力而言，便利的平台、工具，使得組織快速的累積資訊愈趨容易，組織開始面臨另一大挑戰，如何在巨量資訊中發掘有價值的資訊，並能及時將資訊快速地傳遞到對的同人手中，讓知識水平擴散、分享、應用，加快組織的反應速度與解決問題之能力，以獲得最大效益。目前與多組織之都面臨，當員工試圖從組織龐大的知識庫中，獲得所需的知識文件時，往往會耗費相當多的時間，甚至無法搜尋到適合文件，因此陸續有資訊過濾（Information Filter）和資訊擷取（Information Extraction）的概念提出，並有相關方法開始被提出應用，初期主要的概念是透過關鍵字比對的查詢搜尋，以找到所需的知識內容。然而此類方式無法解決同義詞相關問題，耗費使用者時間且無法判斷文件相關性。故文件推薦的概念開始被提

出探討，其中又以『協同過濾推薦』(collaborative filter recommendation) 推薦方法最常被應用於個人化文件推薦方法中，其主要概念是根據目標使用者 (target user) 對其閱讀過的知識文件，進行知識興趣分析，再根據其他使用者的閱讀行為，找到和目標使用者興趣相似的鄰居群 (neighborhood)，再透過該群內成員對閱讀過的知識文件的興趣程度、評分資料 (rating)，進而預測目標使用者對各知識文件的興趣程度、評分分數。

以此為基礎之個人化推薦應用領域非常的廣泛，例如 Knowledge Pump[7]，根據使用者對文章的興趣，找出相同知識領域的群體及文件加以推薦。GroupLens[10]，根據使用者閱讀的新聞內容，找出共同新聞閱讀興趣的鄰居，並推薦未看過，但目標使用者可能有興趣的其他新聞。數位圖書館近年也是大量使用個人化文件推薦，根據目標讀者對書籍或文章閱讀的分類內容，分析出具共同興趣的讀者，再推薦該群體亦感興趣的其他書籍文章給目標使用者。其他相關應用還包括網路書店、電子商務、音樂推薦等。

1.2. 研究動機

對組織而言，伴隨著資訊科技的進步及網路世界的發達，各種資訊透過網路達到無遠弗屆的力量，更累積了大量的資訊及知識。組織內部欲藉由導入 Web 2.0 之概念之企業內部部落格來強化員工知識保存，提供組織內外部知識評價機制、提高知識分享與應用效率，故需解決部落格之多站台結構下，知識有效管理與分享機制。故本研究希望可以提供文件自動化分類功能，能跨部落格定義出組織分類，有效將知識文件分門別類，有助於部落格知識管理與應用。

在資訊爆炸的情況下，推薦系統的重要性也相對的提高，本研究應用協同過濾推薦方法進行推薦研究，協同過濾推薦概念為找尋具擁有共同興趣的社群，社群是由一群擁有共同興趣的人所組成的，群內的成員可以相互分享共同領域的專業知識、工作經驗和文件。在以往的研究中，知識文件推薦系統都只針對個人的單一興趣進行分析，然而使用者的興趣是多元化，因此本研究希望可以找出各種領域的知識，滿足使用者多元興趣的知識需求。並將知識文件推薦給對其領域有相當興趣的使用者。另外，本研究考量組織內員工特徵檔屬性，希望在考量組織員工本身興趣社群關係之外，亦加入員工基本屬性分析員工屬性間相關程度，將此元素加入社群分析，因而形成不同目標知識社群，藉以分析文件推薦與員工屬性之間的關係。透過社群建立與推薦機制，可以有效滿足使用者的知識興趣需求，並提昇組織中知識分享機制的運作效能。

1.3. 研究目的

本研究主要針對組織知識需求，定義出相關的知識分類，透過監督式學習方式，分析各知識文件特徵檔，透過 Cosine 相似度計算新文件與已分類文件之相似度，建立一自動化分類機制，協助組織有效管理各部落格之知識文件，並幫助組織員工有效、快速取得需要的知識文件。另外，本論文提出四種混合式推薦方法計算文件預測「興趣分數」，分別是「社群最大分數法（Community Max Score）」、「社群興趣最大分數法（Community Interest Max Score）」、「使用者權重最大分數法（User Weight Max Score）」、「使用者權重興趣最大分數法（User Weight Interest Max Score）」，首先針對組織中使用者的知識興趣需求進行分析，根據使用者知識興趣相似度與使用者個人屬性相似度，自動分析建立各種知識興趣內容的社群，知識工作者可參與多個社群。接

著，根據每種相似度得到分群結果，四種混合式推薦方法會透過調整內容式推薦與協同式推薦比例計算出文件預測「興趣分數」，得到推薦清單。

在本研究中，特別提出使用者屬性特徵檔的概念，根據組織使用者之間會有個人屬性特性參與不同的知識領域內容為基礎，在計算使用者相似度時，加入此元素形成單一、複合兩種使用者相似度，藉以分析員工屬性檔與社群知識興趣相關性。本研究主要貢獻兩大部分，分別為（一）自動化分類機制（二）自動化多社群推薦機制。

1.4. 論文組織架構

本篇論文共有五個章節，第二章主要說明本研究參考之相關文獻，內容包括資訊過濾與資訊檢索、分類、社群、分群以及推薦方法、企業部落格介紹。第三章的部份則是詳細說明本研究提出方法之架構、內容和流程，第四章為實驗結果及評估，第五章則為研究的結論和未來研究方向說明。

2. 文獻探討

本章節中，主要是介紹跟本研究相關的參考文獻，其中包括資訊過濾與資訊檢索、分類、社群、分群以及推薦方法、企業部落格介紹。

2.1. 資訊檢索與資訊過濾

隨著網際網路和電腦科技的快速發展，資訊存在的形式也日益複雜，包括影音、圖片、文字、動畫等等形式，以致目前在資訊檢索實作上愈趨複雜。本研究主要針對文字型式的資料內容進行自動化分類與推薦機制研究，因所使用工具限制，本研究使用繁體中文文件作為研究對象，希望透過資訊檢索和資訊過濾的技術，從文件中擷取出該文章的重要關鍵特徵，形成文件的關鍵字特徵檔。再藉由文件關鍵字特徵檔概念，延伸到類別、使用者、社群等實體。

2.1.1. 以文件為基礎之資訊檢索與資訊過濾

資訊檢索（Information Retrieval）是指因應使用者之資訊需求提供查詢的方法以及查詢的過程，帶領使用者找到符合他們需求的資訊[13]，透過自動化的資訊檢索系統解決資訊超載（Information overload）的問題，幫助使用者從超載的資源中獲得能夠解決或管理問題的資訊。

目前資訊檢索[3][11]主要的技術有四種，分別為布林模式（Boolean model）、向量空間模式（vector space model）及機率檢索模式（probabilistic retrieval model）、推理網路模型（Inference Network Model）。其中向量空間模式是資訊檢索中較被廣為應用的方

法[17]，向量空間模型最早由 Gerard 提出[15]。在此模型中，一個文件被描述成由一系列關鍵詞（Term）組成的向量，如圖 2-1 所示：

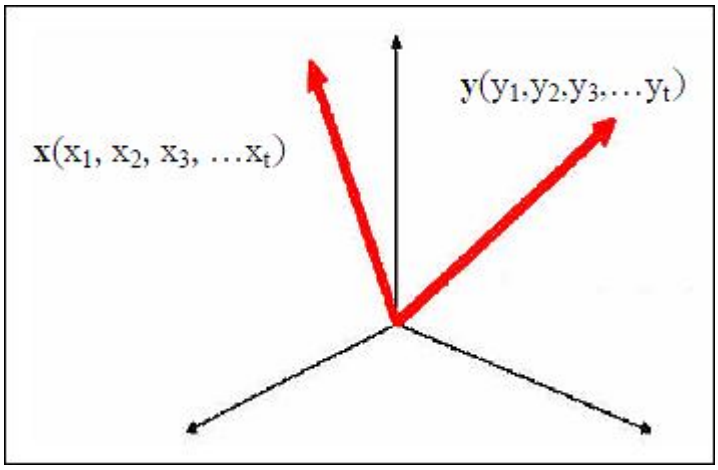

圖 2－1 兩文件在空間向量中的表示

文件透過模型計算關鍵詞權重，可將文件關鍵詞權重組合視為空間向量來表示，當使用者透過查詢語找尋資訊時，則比較文件和查詢語之間的相似程度，最後將相似度高的文件以重要性高至低的排序方式，或以設定門檻值的方式，將檢索結果回饋給資訊需求者。

資訊過濾是過濾掉不相關的資訊，強調的是主動、長期與個人化的資訊服務，如何對使用者進行長期學習而找出使用者資訊需求，進而找出符合使用者需求的資訊文件[9]。故資訊過濾著重在使用者特徵檔學習的技術和演算法。目前資訊過濾濾技術大致可分為內容式資訊過濾(Content-based Information Filtering)及協同式資訊過濾(Collaborative Information Filtering)、混合式資訊過濾(Hybrid Information Filtering)三種。

2.1.2. 關鍵字權重

如何透過一組關鍵字來表示該文件內容之重要特徵詞屬性，在資訊檢索的領域中，

以 TF-IDF (Term Frequency/Inverse Document Frequency)應用最為廣泛[16]。TF-IDF 包含兩個部份，(1) TF(Term Frequency)，字詞頻率，表示字詞在文章中出現的次數，若出現次數愈高，則表示重要性愈大。 (2) IDF (Inverse Document Frequency)，反文件頻率，指字詞在其他文章中出現的次數，若出現次數愈高，則其鑑別率會愈低，即公式 2-1 中的 IDF。例如，介係詞「的」時常出現在許多文件，則此　字詞對該篇文件的重要性即會降低。

TF-IDF 的計算方式如下：

$$W_{i,j} = TF_{i,j} \times IDF_i = TF_{i,j} \times \log \frac{N}{n_i} \quad \text{(公式 2-1)}$$

$W_{i,j}$：關鍵字權重，字詞 i 在文章 j 中的重要性。

$TF_{i,j}$：字詞頻率，字詞 i 在文章 j 出現的次數。

$IDF_{i,j}$：反文件頻率，字詞 i 出現在其他文件集中的次數。

N：文件集中全部的文件數目

n_i：文件集中包含字詞 i 的文件數目

2.2. 分類

文件分類的目的，在對文件進行分門別類的加值處理， 使得文件易於管理、利用。協助使用者可以更快速、正確的取得所需要的文件，增加文件的使用率。由於資訊技術普及運用，組織內的知識文件不斷快速累積，難以有效的管理與利用， 人工分類又會

因使用者個體而有所不同，使得分類無統一的定義，降低分類一致性，加上人工分類相當耗時。文件自動化分類的需求也就因應而生。

本研究之自動化文件分類是透過監督式學習法來建立分類模型。預先定義好文件的分類項目，蒐集訓練文件由專家進行類別分析後，將文件歸屬至對應類別，再將分類完之訓練資料提供給分類器作為訓練資料，進行分類演算法學習，建立出分類的模型，接著再將新文件透過分類模型分析歸屬類別。目前較常使用文件分類演算法[17]為 KNN（K-Nearest Neighbor）[6]與 SVM（Support Vector Machine）[5]演算法，本研究爲考量資料量與演算效能，提出類別特徵檔概念，類別特徵檔是透過各類別之訓練文件特徵檔形成，新未分類文件則透過其文件特徵檔與各類別特徵檔進行相似度計算，作為分類模型，將新文件歸類到與其相似度最高之類別。目前文件常見之相似度計算方法[18]包含 Pearson correlation coefficients、Cosine-based Similarity、Adjusted Cosine Similarity。

2.3. 社群

社群概念近年來陸續被重視，且廣泛的應用於各種領域上。關於社群定義，Rheingold（1993）提出，「虛擬社群是社會的集合體，當足夠數量的群眾，在網路上進行了足夠的討論，並付出足夠的情感，以發展人際關係的網路，則虛擬社群因而形成」。Shafer（1999）則提出，「社群為有著相同的興趣或目標，並隨著時間加深相互間瞭解的一群人」。社群關係可分為直接與間接關係，直接關係如共同工作專案團隊，間接關係則是透過第三者間接發掘的，例如使用者 u1 和 u2 雖然沒有直接關係，但分別跟 u3 為有共同興趣之直接關係，則 u1 和 u2 為非正式關係之社群[2]。在本研究中，社群泛指由一群對特定之領域、專業知識、工作內容、具有共同興趣或共同經驗的關係的人所組成，

組織希望透過社群的建置，開發共同興趣資訊的平台，提供組織內員工交換知識或資訊流通的環境，有效發展知識管理和知識分享的機制。

2.4. 分群方法

本研究目的為多社群自動化推薦文件，組織中包含了大量的知識和資訊，需要一社群分析的過程，分析組織知識中有多少種類社群知識存在，因本研究在社群分析過程無法預知有多少種類社群知識存在，故不事先設定社群種類與個數，將透過分群的演算法來達到多社群分析。

分群（clustering）[1]是一種將資料分類成群的方法，其主要的目的乃在於找出資料中較相似的幾個群聚（clusters），一般而言，分群法主要分為兩大類，分別是階層式分群法（hierarchical clustering）與分割式分群法（partitional clustering）兩類，另外還有以格子為基礎的群聚演算法。針對不同的需求和資料特性，需應用不同的分群方法來分析處理，本研究提出的是多社群推薦概念，採用的是以格子為基礎的 CLIQUE 分群方法。以下三種方法說明如下：

（一）階層式分群法（hierarchical clustering）：

群數（number of clusters）可以由大變小，或是由小變大，來進群聚的合併或分裂，最後再選取最佳的群數。階層式又分成群聚分析(Agglomerative)和分割分析(Divisive)，其中有四個基本的分群法，即 Single-link、Complete-Link、Average-Link 和 Ward’s，這些方法主要的差別在於對於群和群之間距離的計算方法。

（二）分割式分群法（partitional clustering）：

先指定群數後，再用一套疊代的數學運算法，找出最佳的分群方式以及相關的群中心，目的是希望盡量減小每個群聚中，每一點與群中心（cluster center）的距離平方差（square error）。常見的方法有 mode-seeking、Squared Error、Nearest Neighbor Clustering 和 K-means 等分群方法，其中 K-means 經常用來處理文件分群。

（三）CLIQUE：

CLIQUE 分群方法是以格子為基礎之分群法，優點能處理高維度之資料，且富彈性。格子基礎的群聚演算法主要是將資料空間量化成許多格子，以格子為單位進行群聚運算，如此可以大量的減少群聚的時間。格子基礎群聚演算法很容易受到其格子大小的影響，若是格子太大則所找出的群聚外型會比較粗糙，故使用者必須輸入適當切割間隔值，如此才能分割出適當的格子。本研究選擇 Clique 分群方法主要有下列原因，第一，此分群法不需預先設定分群數，而是可以根據內容自動判斷。第二，每個物件可以被分到多群。第三，群內任兩點皆具大於門檻的相似度，即可完全代表一主題內容。

2.5. 推薦方法

在資訊爆炸時代，系統越來越講求個人化服務，故推薦系統(recommender system)的應用與發展越來越快速、完善。使用者以往需要花費許多時間在眾多知識中尋找有用、適合的知識，瀏覽許多文件，才能找到自己有興趣的文件。有此陸續有學者開始研究自動化推薦的方法，如何藉由個人特徵、行為模式等，發展出個人化模型，進而提供

個人化的資訊與服務。企業也可透過文件推薦方法，提高組織知識使用率，讓知識快速正確的傳遞給對的人員，使知識分享的機制更加完善且有效率，有助於組織知識管理和分享。

協同過濾推薦最早是 1992 由 Goldberg 等學者所提出的 Tapestry[8]，為解決員工電子郵件過量的問題，讓員工決定自己的感興趣的郵件類型，以便有效過濾符合使用者興趣之電子郵件。1994 又有 GroupLens[10]被提出，此系統主要是應用在新聞的過濾篩選，提供讀者其感興趣的新聞，透過讀者看過內容後給一個評比的分數，以讀者對過去感興趣的新聞內容在未來也會有興趣之假設為前提，結合同一項目不同使用者之評分資料以進行新聞的推薦。協同過濾推薦方法廣泛的應用在各領域，例如，Siteseer[14]利用相鄰使用者的書籤(bookmark)進行推薦，Knowledge Pump[7]對使用者感興趣之文章進行推薦，Ringo[19]對音樂進行推薦，MovieLens 電影推薦系統、YouTube，其他包括亞馬遜網路書局之電子商務應用等。

推薦方法有三個，即以內容式基礎推薦方法（Content-based recommendation）、協同過濾推薦方法 Collaborative filtering recommendation）、混合式推薦（Hybrid recommendation）方法，簡要說明如下：

（一）內容式過濾推薦（Content-based filtering recommendation）[4][12]：

針對項目內容分析，計算該目標使用者對項目的喜好程度，進而找出目標使用者喜歡的項目。因為內容式推薦是針對項目內容去做分析，會導致以下問題：

（a）分析物品只能用特徵集合，對於聲音、圖片、藝術品、影像、文字等沒辦法被特徵化的內容無法處理。

（b）無法找出與和目標使用者過去喜好不同但目標使用者可能有興趣之項目，會失去許多的潛在推薦的可能性。

（c）對於內容品質無法透過群體力量有效的分辨。

（二）協同式過濾推薦（Collaborative filtering recommendation）：

協同過濾推薦為目前應用較廣泛的方法，主要是利用具有共同興趣、共同經驗之群體的喜好，來推薦目標使用者其可能感興趣的資訊，強調個人透過合作的機制，藉由過去的行為紀錄，進而分析使用者之間的行為偏好相似度，找出與目標使用者相似度接近的鄰居，透過相似鄰居的行為建議，給予目標使用者資訊推薦，亦可推薦使用者之前尚未發現過卻有興趣的資訊。協同過濾式推薦可分成二大步驟，第一，利用相似度計算方法分析群體之間使用者彼此的相似程度，以尋找相似鄰居，並進行分群。第二，預測目標使用者對項目興趣分數，以相似鄰居為對項目的興趣程度為基礎，利用協同過濾推薦方法，預測目標使用者對該項目之喜好分數[11]。喜好分數愈高代表愈符合使用者興趣，可推薦給使用者。協同過濾推薦解決了內容導向推薦方法的缺點：

（a）不需要分析項目內容。

（b）可找出與和目標使用者過去喜好不同但目標使用者可能有興趣之項目，添加潛在推薦的可能性。

（c）可透過群體力量（評分、點閱）分片內容品質。

但協同過濾推薦仍有許多限制如下：

（a）冷起始（cold start），指在系統上完全沒有任何使用的紀錄、使用紀錄過少的目

標使用者，或是尚未有點閱、交易、評價等紀錄的項目，會無法正確且有效率的推薦，本研究即過透過加入員工屬性檔計算員工相似度以解決新使 用者問題。

（b）稀疏性（sparsity），當使用者所接觸到的物件大部分都只佔系統非常小的比例。當產品與使用者兩者的比例差距十分懸殊時，系統就沒有辦法找到適合的推薦者來進行推薦。

（三）混合式推薦（Hybrid recommendation）：

為了改進內容式過濾推薦與協同式過濾推薦的缺點，合併雙方的優點，結合兩種方法的推薦系統就很自然而然的產生出來。這就是所謂的混合式推薦（Hybrid recommendation）。使用內容式過濾技術建立的使用者 profiles，用以計算使用者之間的相似度，然後再使用協同過濾的技術，分析相似的使用者閱讀、購買行為，將相似使用者之高興趣項目推薦給目標使用者。

2.6. 企業部落格

隨著科技不斷的進步，企業競爭已經進入以創新思考、協同合作思維，1996 年「經濟合作開發組織」（OECD, 1996）正式指出，全球經濟發展型態也轉為以知識為基礎之後知識經濟時代，不斷創新及持續學習成為組織決勝的關鍵。員工是組織最重要資產，企業如何開發員工潛力、創意、增進協同合作，並建立全新的知識價值觀並落實知識分享文化，加深員工對組織的向心力、認同感。因此，結合了知識儲存、團隊學習、社群、訂閱特色的部落格議題已逐漸被重視與應用。Fortune Magazine 認為部落格是 2005 年最新科技趨勢；Harvard Business Review 將企業部落格列為 2005 年最具突破性的發明之

一，可見部落格已成為後知識經濟的趨勢新議題。

部落格一詞最早出現於 1997 年，部落格是 weblog 的簡稱；而 weblog 一詞中的 web 是指網路，log 則指紀錄檔，亦隨著部落格軟體技術的發展愈趨多元化。

企業部落格朝知識部落格方向發展，希望透過此平台，提供組織未來在知識工作者的例行工作、專案經驗儲存、協同知識分享與互動，發掘有價知識，助於組織員工協同合作、團隊學習、知識管理與即時溝通，若能將知識價值與工作績效整合，可提升組織知識品質與平台績效[22]。目前組織知識平台可分為下列應用：

表 2－1 知識部落格應用模組

知識工作者部落格	**協同合作部落格**
（CEO 日誌）	（組織合作部落格）
（高階主管日誌）	（團隊合作部落格）
（知識專家日誌）	（專案合作部落格）
（知識工作者日誌）	（供應鏈合作部落格）
專案部落格	**創新部落格**
（專案管理日誌）	（創意點子部落格）
（專案經理日誌）	（創新部落格）
（客服日誌）	

3. 企業部落格之自動化分類與推薦方法

此章節主要介紹本研究提出企業部落格之自動化分類與推薦方法。整體架構圖如下圖3－1所示：

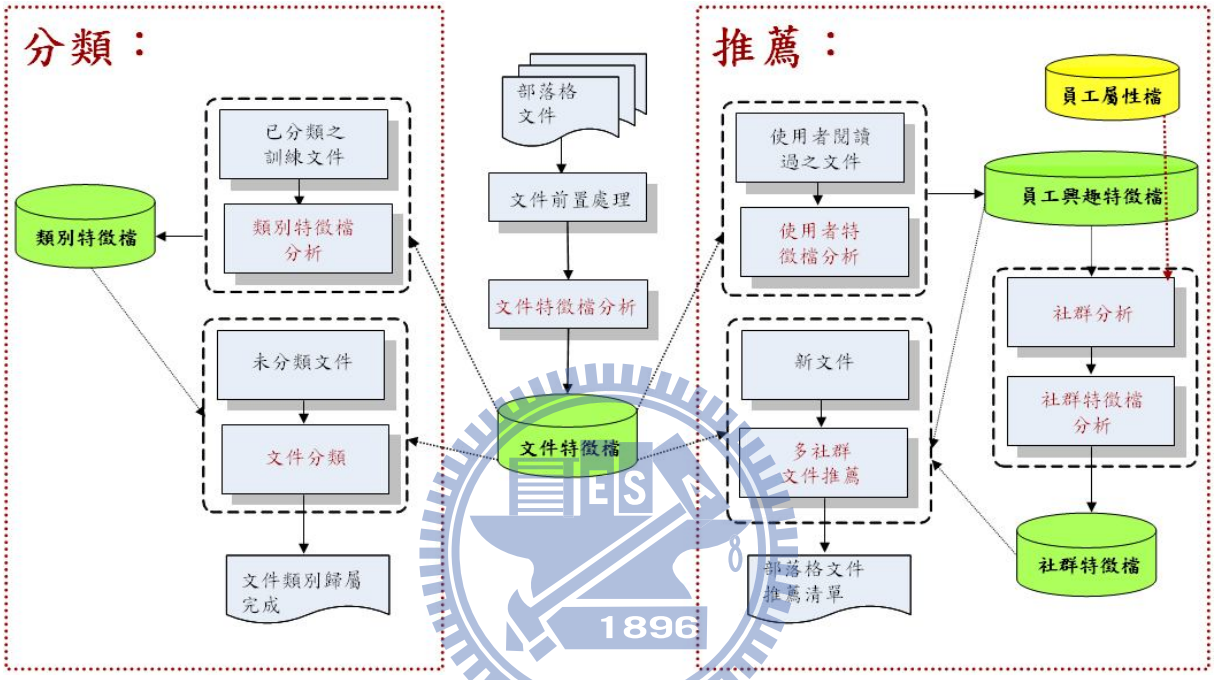

圖 3－1 企業部落格之自動化分類與推薦方法整體架構

3.1. 企業部落格之自動化分類與推薦方法之架構

本章節主要介紹提出之企業部落格之自動化分類與推薦方法之架構。企業部落格之自動化分類與推薦方法分成兩大模組，一為分類模組，另一為推薦模組。為了達到文件分類與推薦的目的，必須分析組織中的文件知識內容，針對知識工作者閱讀過的所有知識文件做前置處理，將知識文件的內容以文件特徵檔表示，將在3.2文件前置處理章節中介紹。

章節3.3分類，針對類別特徵檔形成與分類機制作介紹。分類模組主要是針對文件作單一類別之分類，透過各類別之類別特徵檔形成，計算新文件與各類別之相似度已達到文件自動化分類機制，分類方法架構如下：

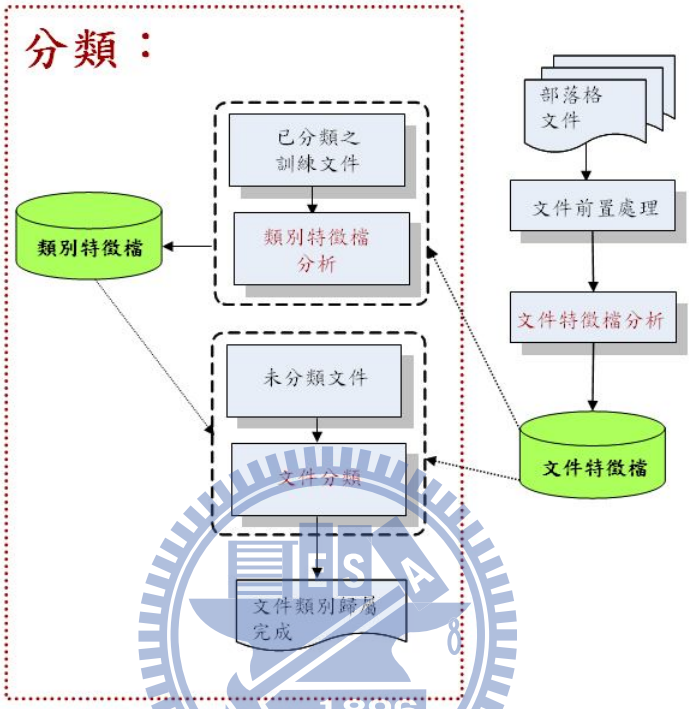

圖 3－2 企業部落格之自動化分類方法之架構

推薦模組又分為兩個部份，第一個部份是建構組織知識社群，參與社群的成員會對社群所代表的興趣知識具相當程度的興趣，但不代表該成員的全部興趣。第二個部份透過知識社群做文件推薦。多社群文件推薦方法的流程如圖3-3：

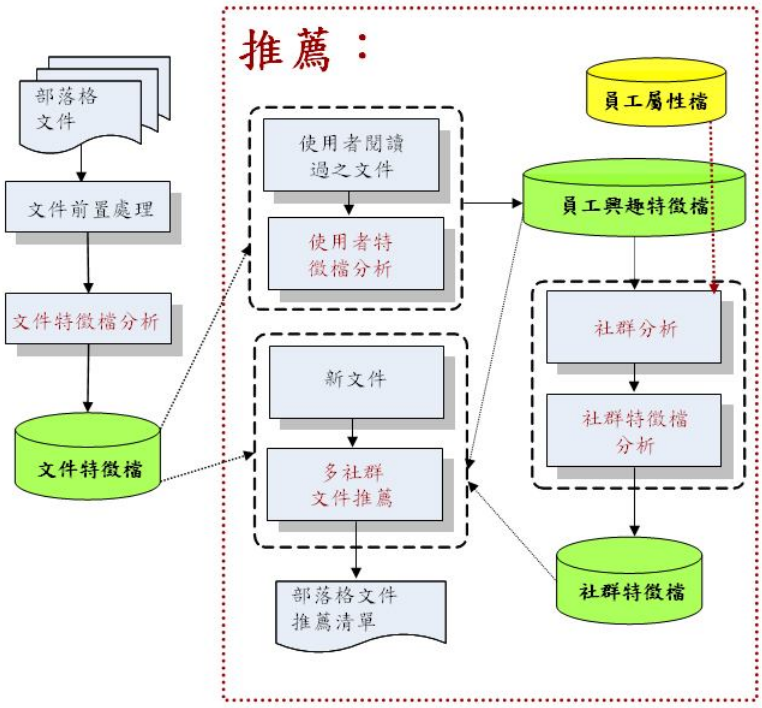

圖 3－3 企業部落格之自動化推薦方法之架構

接續章節3.4、3.5將介紹推薦機制，第一部份是產生知識社群，分成四個主要步驟；第一步驟以知識工作者閱讀過的文件內容為基礎產生知識工作者知識興趣的特徵檔；第二步驟是根據知識工作者感興趣特徵檔計算知識工作者間的相似度；第三步驟計算知識工作者之個人屬性檔相似度；第四步驟是運用兩種工作者相似度，結合知識工作者感興趣特徵檔與知識工作者個人屬性檔來計算知識工作者間的相似度(複合)，找出組織中各種不同知識領域的知識社群，每位知識工作者可參與多個知識社群。第二部份文件推薦，結合多知識社群分析與混合式濾推薦的方法，產生符合目標知識工作者知識興趣的推薦清單。

3.2. 文件前置處理、文件特徵檔分析

本研究透過空間向量概念來表示知識文件內容，經過文件前置處理的程序，將知

識文件統一轉換成由關鍵字和權重分數表示的文字特徵檔，進而運用此特徵檔於分類、推薦機制。圖3-4是文件前置處理、特徵檔分析流程，透過前置處理的步驟，刪除重複或不重要的文字，降低特徵檔中不必要資訊的出現，以減少在應用過程中的複雜度。

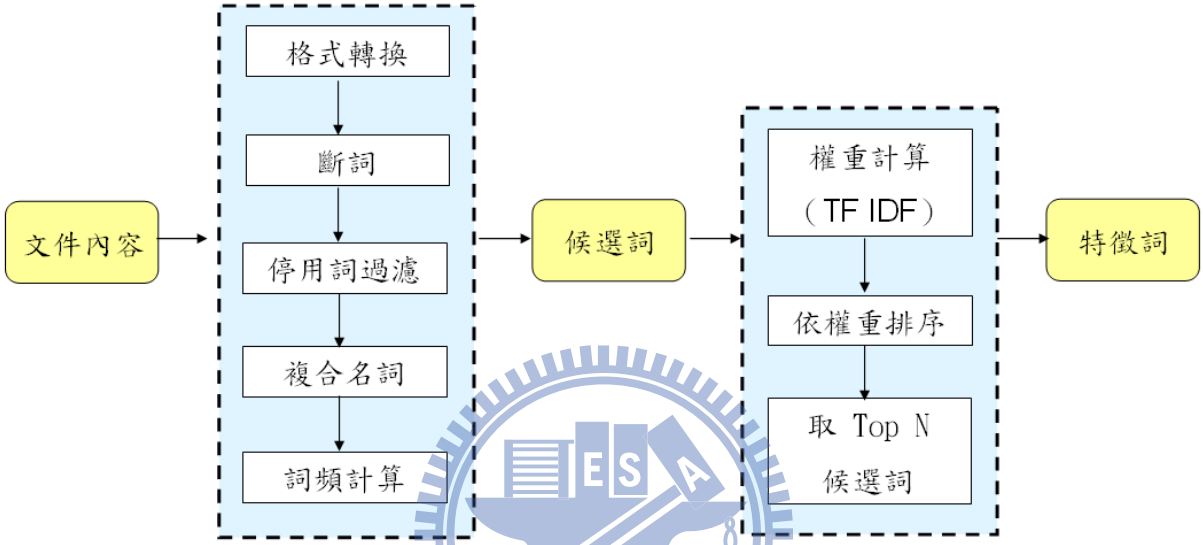

圖 3－4 文件前置處理模組

(一) 格式轉換

本研究使用部落格平台上之知識文件，這類知識文件皆為WEB HTML格式內容，故在使用知識內容之前，需要清除不必要之HTML Tag（<html></html>、<a>），且部落格文件內容隨技術演變，涵蓋內容越趨多元化，如圖檔、動畫、影片檔、外部嵌入元件（影片、Flash）等，無法涵蓋在本研究之空間向量表示範圍內，故須排除掉此類元素。將上述格式轉換處理後，可以萃取出知識文件之真正內容，以便進行後續分析處理。

(二) 斷詞

本研究針對中文內容之知識文件進行分類與推薦，以中文為例，詞為最小單位，是最小有意義且可以自由使用的語言單位。本研究使用中研院之CKIP斷詞服務進行本文內容之斷詞，如圖3-5。CKIP為一具有新詞辨識能力並附加詞類標記的選擇性功能之中文斷詞系統。有拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題，並提供該詞所對應中研院定義出一套"平衡語料庫詞類標記集"之詞類。

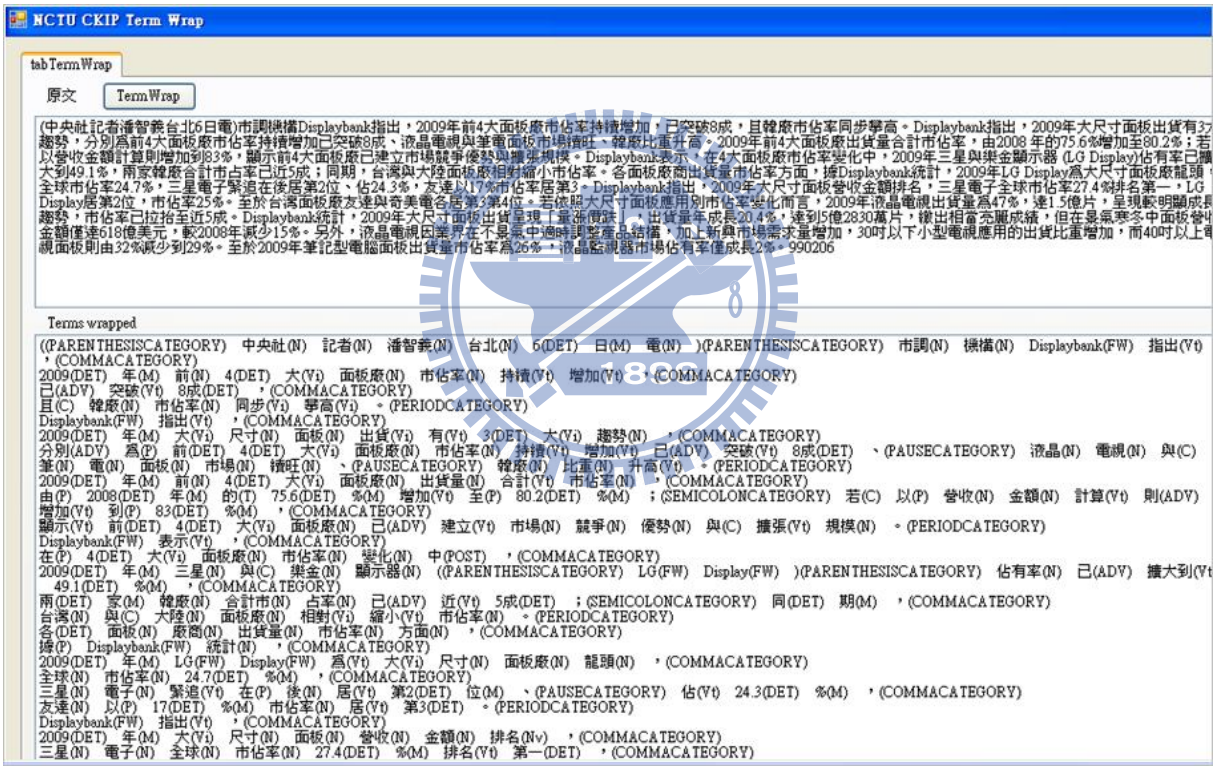

圖 3－5 實作斷詞功能系統

但CKIP限定僅能針對繁體中文之內容進行斷詞處理，故本研究僅針對企業部落格上繁體中文文件進行研究。此服務目前之斷詞正確率約為 95 %-96%[23]。

(三) 停用字過濾

在知識文件中，有些字詞經常性被的大量使用，但這些字詞無法代表文章的關鍵字，例如代名詞、be動詞或是冠詞。本步驟的目的即是移除此類字詞，根據研究學者提出的停用字列表，將文件中有出現在停用字列表的字詞移除。

(四) 複合名詞

中文名詞會有復合名詞情形發生，例如”金融”、”海嘯”，應為複合性名詞金融海嘯，此複合詞才具有文件內容代表性意義，故本研究會針對此類重要複合性名詞進行名詞合併動作。

(五) TF-IDF

TF-IDF主要是計算字詞在文件中的權重分數，TF是指字詞出現的頻率(Term Frequency)，一字詞在文章中出現愈多次，表示重要性愈高，IDF是指字詞的反文件頻率(Inverted Document Frequency)，表示字詞出現在其他文章的頻率，若某字詞在很多文件出現頻率都很高，即表示該字詞對於單篇文件的代表性並不高。

(六) 特徵詞選擇

根據TF-IDF計算各篇文件的字詞權重分數後，將字詞依權重降幕排序取Top N的方式挑選代表知識文件的特徵辭。如下表3-1所示：

表 3－1 文件特徵檔

Term	Weight
太陽能	3.17545
市場	2.74075
能源	1.6394
多晶矽	1.23640

面板	0.97437

此表用來表示某文件之文件特徵檔，各文件由其特徵詞與特徵辭權重組成二維度的矩陣，以供後續分類、推薦分析。

3.3. 分類

本研究之分類模型使用監督式學習法，分類器透過已分類之訓練資料進行學習，建構出各類別特徵檔，再將文件與各類別特徵檔以Cosine相似度計算法運算，求得該文件之歸屬類別。

3.3.1. 類別特徵檔分析

各類別特徵檔取得流程如圖3-6，計算類別內文件之特徵詞對類別的重要性，計算計算方式如公式3-1所述，計算完各特徵詞之權重之後，再依特徵詞權重降幕排序取TOP N 做為該類別之特徵檔。

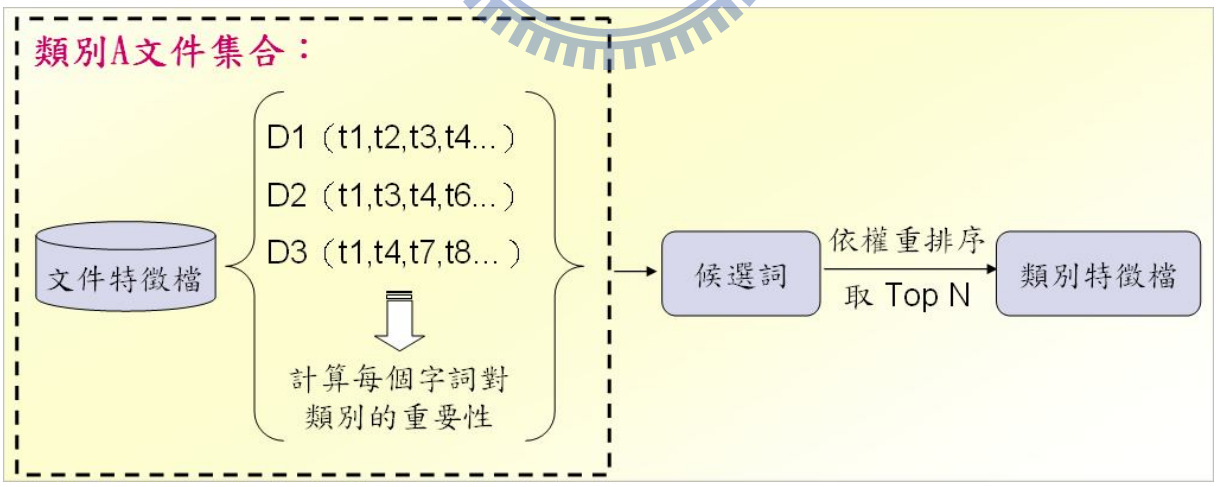

圖 3－6 文件特徵檔分析流程

$$CW_{c,i} = \frac{\sum_{d \in D_c} w_{i,d}}{|D_c|} \qquad \text{(公式 3-1)}$$

- $CW_{c,i}$：在類別c中，字詞 i 的權重分數
- D_c：類別c中所有文件集合
- $w_{i,d}$：字詞 i 在文件d中的權重分數
- $|D_c|$：類別c中所有文件總數

3.3.2. 文件、類別相似度計算

本研究是透過文件特徵檔與各類別特徵檔之間進行相似度計算，將文件歸屬至相似度分數最高之類別。本研究採用Cosine相似度進行計算。Cosine相似度計算方式為公式3-2：

$$COS_{Di,Cj} = \frac{\sum_{t=1}^{k} w_{t,Di} \times w_{t,Cj}}{\sqrt{\sum_{t=1}^{k} w_{t,Di}^2} \times \sqrt{\sum_{t=1}^{k} w_{t,Cj}^2}} \qquad \text{(公式 3-2)}$$

- $COS_{Di,Cj}$：以cosine方式計算，文件Di 與類別Cj之相似度分數
- $w_{t,Di}$：字詞 t 在文件 Di 特徵檔中的權重分數
- $w_{t,Cj}$：字詞 t 在類別 Cj 特徵檔中的權重分數

相似度計算範例如下：

表 3－2 文件 Di 和 類別 Cj 之特徵檔範例

Di		Cj	
Term	**Weight**	**Term**	**Weight**
太陽能	0.4468	太陽能	0.8529
價格	0.3135	電池	0.801
建築	0.2327	地球日	0.4373
多晶矽	0.2285	多晶矽	0.4332
薄膜	0.2203	能源	0.3325
面板	0.1824	地球	0.3013
下跌	0.1563	產能	0.2556
跌幅	0.1285	夏普	0.2389
FPD	0.0821	面板	0.237
市場	0.0627	二氧化碳	0.1893

$$\sum_{t=1}^{n} w_{t,u1}^2 = 4.2468^2 + 0.3135^2 + 0.2327^2 + 0.2285^2 + 0.2203^2 + 0.1563^2 + 0.1285^2 + 0.0821^2 + 0.0627^2 + 0.1824^2 = 0.399$$

$$\sum_{t=1}^{n} w_{t,u2}^2 = 0.8529^2 + 0.8010^2 + 0.4373^2 + 0.4332^2 + 0.3325^2 + 0.3013^2 + 0.2556^2 + 0.2389^2 + 0.237^2 + 0.1893^2 = 2.1637$$

$$COS_{u1,u2} = \frac{0.3135 \times 0.8010 + 0.2285 \times 0.3325 + 0.1285 \times 0.2389}{\sqrt{0.399} \times \sqrt{2.1637}} = 0.385$$

3.4. 社群分析

使用者特徵檔代表該使用者感興趣的知識內容，依據使用者相似程度，進行Clique分群，形成多知識社群，並依據社群內成員特徵檔分析社群特徵檔，完成社群分析。另外在使用者相似度計算，除了採用興趣相似度，加入個人屬性檔考量；將興趣相似度與使用者個人屬性相似度依權重調配計算後，形成複合相似度作為分群依據。

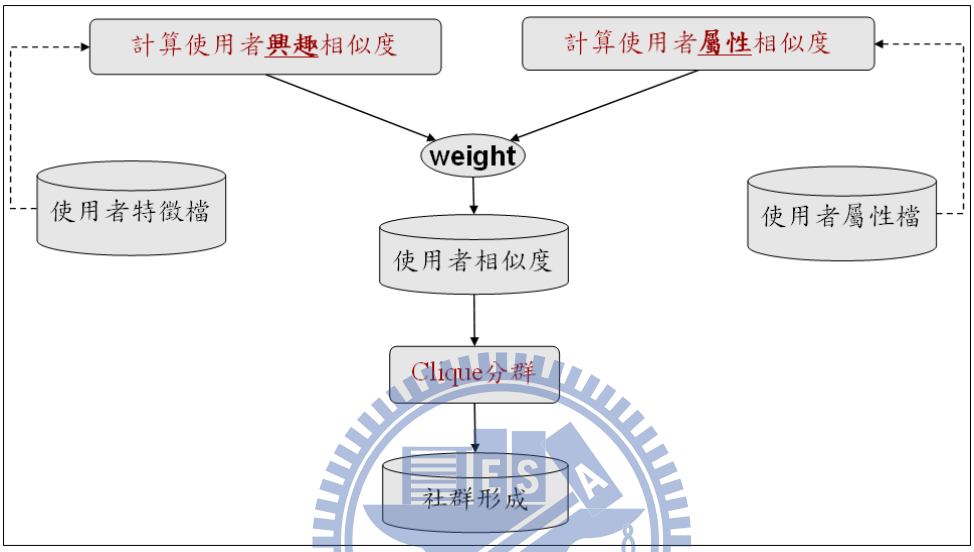

圖 3－7社群分析流程

3.4.1. 使用者特徵檔分析

使用者特徵檔代表該使用者感興趣的知識內容，此步驟如圖3-8，取出使用者閱讀過之文件的特徵檔，針對每個特徵詞權重作加總平均計算，計算方式為公式3-3。依計算後之權重排序，取出Top N 候選詞作為該使用者之特徵詞集合，形成使用者之特徵檔，詳細說明如下：

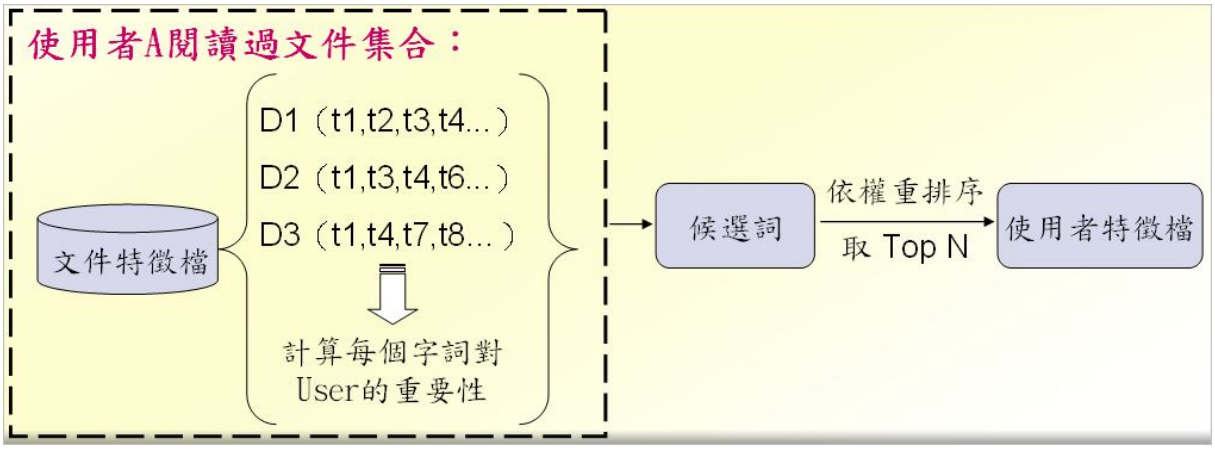

圖 3－8 使用者特徵檔分析流程

$$UW_{u,i} = \frac{\sum_{d \in D_u} w_{i,d}}{|D_u|} \qquad \text{(公式 3-3)}$$

- $UW_{u,i}$：字詞 i 對使用者 u 的權重分數
- D_u：使用者 u 曾經閱讀過的文件集合
- $w_{i,d}$：字詞 i 在文件d中的權重分數
- $|D_u|$：使用者 u 閱讀過的文件總數

3.4.2. 使用者相似度計算

在協同過濾推薦的方法中，相同興趣的使用者分析是重要的關鍵，透過知識興趣相似使用者的行為進而預測目標使用者的喜好。因本研究主要針對企業部落格文件探討，在探討使用者相似度計算，除考量使用者特徵檔(知識興趣)部份，更加入組織使用者之個人屬性，例如組織流程相關性強的群組，對特定領域知識內容應較感興趣，

或是新進同仁會對新手知識等內容較感興趣。故本研究取得下列五項個人屬性特質，並根據專家定義將各屬性的特徵值依連續、不連續做正規轉換如表3-3，左側欄位為各屬性原始值，右側欄位為轉換過後之屬性值。

表 3－3 使用者特徵檔轉換

	部門		性別s		職級t		年資y		年齡a	
u1	IT	>0	F	2	99	1	<3個月	1	<20	1
u2	IT	>1	M	1	90	2	3個月~<1年	2	20~<30	2
u3	BU	>1	F	2	30	3	1年~<3年	3	30~<40	3
u4	TU	>1	M	1	20	4	3年~<5年	4	40~<50	4
u5	OU	>1	F	2	10	5	5年~<8年	5	50~<60	5
u6	QM	>1	M	1	<10	6	>8年	6	>60	6

（一）使用者屬性檔相似度計算

將使用者各屬性轉換後之值視為空間向量表示，利用空間向量距離的觀念，計算出兩點向量距離並將距離值正規化至0～1，兩使用者之相似度即為1-distance(U1,U2)，使用者屬性檔相似度公式3-4詳細說明如下：

$$Dis_{u1,u2} = \frac{1}{5}\left(\begin{array}{c} dept(0,1) + |s\_u1 - s\_u2| + \frac{1}{5} \times |t\_u1 - t\_u2| \\ + \frac{1}{5} \times |y\_u1 - y\_u2| + \frac{1}{5} \times |a\_u1 - a\_u2| \end{array} \right) \quad \text{(公式 3-4)}$$

$$Sim_{u1,u2} = 1 - Dis_{u1,u2} \ , \ (0 \le \mathrm{Sim} \le 1)$$

- $Dis_{u1,u2}$：使用者 u1 與 u2 向量距離
- $dept(0,1)$：使用者 u1 與 u2 部門是否相同，同為0，不同為1
- s_u1、s_u2：使用者 u1 與 u2 性別屬性轉換後之值

- t_u_1、t_u_2：使用者 u1 與 u2 職級屬性轉換後之值
- y_u_1、y_u_2：使用者 u1 與 u2 年資屬性轉換後之值
- a_u_1、a_u_2：使用者 u1 與 u2 年齡屬性轉換後之值
- $Sim_{u1,u2}$：使用者 u1 與 u2 相似度

使用者屬性檔相似度計算範例如下：

$$Dis_{u1,u2} = \frac{1}{5}(0 + |2-1| + \frac{1}{5}|1-2| + \frac{1}{5}|1-2| + \frac{1}{5}|1-2|) = 0.32$$

$$Sim_{u1,u2} = 1 - 0.32 = 0.68$$

$$Dis_{u1,u5} = \frac{1}{5}(1 + |2-2| + \frac{1}{5}|1-5| + \frac{1}{5}|1-5| + \frac{1}{5}|1-5|) = 0.68$$

$$Sim_{u1,u2} = 1 - 0.68 = 0.32$$

$$Dis_{u2,u5} = \frac{1}{5}(1^2 + |1-2| + \frac{1}{5}|2-5| + \frac{1}{5}|2-5| + \frac{1}{5}|2-5|) = 0.76$$

$$Sim_{u1,u2} = 1 - 0.76 = 0.24$$

表 3－4 使用者特徵檔範例

	u1	u2	u5
部門	IT	IT	OU
	V0		V1
性別	F	M	F
	2	1	2
職級	99	90	10
	1	2	5
年資	1.5月	0.9年	5年
	1	2	5
年齡	18	25	52
	1	2	5

（二）使用者特徵檔相似度計算

將使用者特徵檔以Cosine相似度計算方式為公式3-5：

$$COS_{u1,u2} = \frac{\sum_{t=1}^{k} w_{t,u1} \times w_{t,u2}}{\sqrt{\sum_{t=1}^{k} w_{t,u1}^2} \times \sqrt{\sum_{t=1}^{k} w_{t,u2}^2}} \qquad \text{(公式 3-5)}$$

- $COS_{u1,u2}$：以cosine方式計算，使用者 u1 與使用者 u2 相似度分數
- $w_{t,u1}$：字詞 t 在使用者 u1 特徵檔中的權重分數
- $w_{t,u2}$：字詞 t 在使用者 u2 特徵檔中的權重分數

使用者特徵檔相似度計算範例如下：

表 3－5 使用者 u1 和 u2 的特徵檔範例

U1		U2	
Term	**Weight**	**Term**	**Weight**
股市	0.4468	物價	0.8529
太陽能	0.3135	太陽能	0.801
指數	0.2327	消費卷	0.4373
矽晶圓	0.2285	手機	0.4332
交易所	0.2203	矽晶圓	0.3325
股票	0.1824	台股	0.3013
下跌	0.1563	團購	0.2556
面板	0.1285	面板	0.2389
分析家	0.0821	利空	0.237
股海	0.0627	優惠	0.1893

$$\sum_{t=1}^{n} w_{t,u1}^2 = 0.4468^2 + 0.3135^2 + 0.2327^2 + 0.2285^2 + 0.2203^2$$
$$+ 0.1563^2 + 0.1285^2 + 0.0821^2 + 0.0627^2 + 0.1824^2 = 0.399$$

$$\sum_{t=1}^{n} w_{t,u2}^2 = 0.8529^2 + 0.8010^2 + 0.4373^2 + 0.4332^2 + 0.3325^2$$
$$+ 0.3013^2 + 0.2556^2 + 0.2389^2 + 0.237^2 + 0.1893^2 = 2.1637$$

$$COS_{u1,u2} = \frac{0.3135 \times 0.8010 + 0.2285 \times 0.3325 + 0.1285 \times 0.2389}{\sqrt{0.399} \times \sqrt{2.1637}} = 0.385$$

本研究針對兩項使用者相似度，按權重加總而得一複合使用者相似度，分別使用單一(興趣特徵值相似度)、複合兩種相似度進行知識社群分群、推薦，比較兩項參數對結果差異性。

3.4.3. 知識社群分群

知識社群分析，主要目的是將使用者分群，本研究透過使用者相似度當做分群依據，並以每位使用者應擁有多項知識興趣為前提，可參與一至多個知識社群，分群後之每個知識社群代表某種特定領域、組合的知識興趣內容；我們透過Clique分群技術進行知識社群分析，分群完之後，再依據群內的使用者特徵檔，分析出各社群的特徵檔。

(一) Clique分群

根據使用者的特徵檔分析知識社群。知識社群有兩大特性，第一，群內差異小：每一個社群代表某一特定的知識內容，可能為單一主題或由多個主題所組成的，且社群內的成員興趣相近，且具有共同的興趣。第二，使用者可能有多個興趣知識領域，因此一個使用者可能參與多個知識社群。本研究利用使用者相似度，設定一門檻值，透過Clique分群方法將使用者分群。以下為Clique分群範例：

範例：

為找出組織中的知識社群，先計算所有使用者彼此相似度如下表3-6，設定相似度門檻值為0.4。接著透過Clique分群，找出所有的社群組合，分群的步驟說明如下：

表 3－6 知識工作者相似度矩陣

	A	B	C	D	E
A	—	0.1315	0.6147	0.4143	0.2911
B	0.1315	—	0.0428	0.9774	0.9087
C	0.6147	0.0428	—	0.4021	0.1429
D	0.4143	0.9774	0.4021	—	0.9637
E	0.2911	0.9087	0.1429	0.9637	—

步驟一：找出相似度超過門檻值(>0.4)的所有組合。根據知識工作者為間的相似度，找出相似度大於0.4之使用群。

{A,C} {A,D} {B,D} {B,E} {C,D}{D,E}

步驟二：往上第三層、第四層…續找共同群，一直到最後一層。

{A,C} {A,D}{C,D}→{ACD}

{B,D} {B,E}{D,E}→{BDE}

步驟三：刪除重複的群

~~{A,C} {A,D} {B,D} {B,E} {C,D}{D,E}~~

{ACD} {BDE}

在此範例中最後分成兩個知識社群ACD及*BDE*，群內的使用者相似度高，群間的相似度低。且可看到使用者D會參與兩個社群，社群內的成員相互具有高度相關的興趣內容，每一個社群皆可代表著一個特定知識興趣。

(二) 社群特徵檔

社群特徵檔代表該社群所代表之興趣知識內容，此步驟如圖3-9，根據社群內成員的使用者特徵檔，將成員特徵詞權重作加總平均計算。針對每個特徵詞權重作加總平均計算，計算方式如公式3-6。依計算後之權重排序，取出Top N 候選詞作為該社群之特徵詞集合，形成社群之特徵檔，詳細流程說明如下：

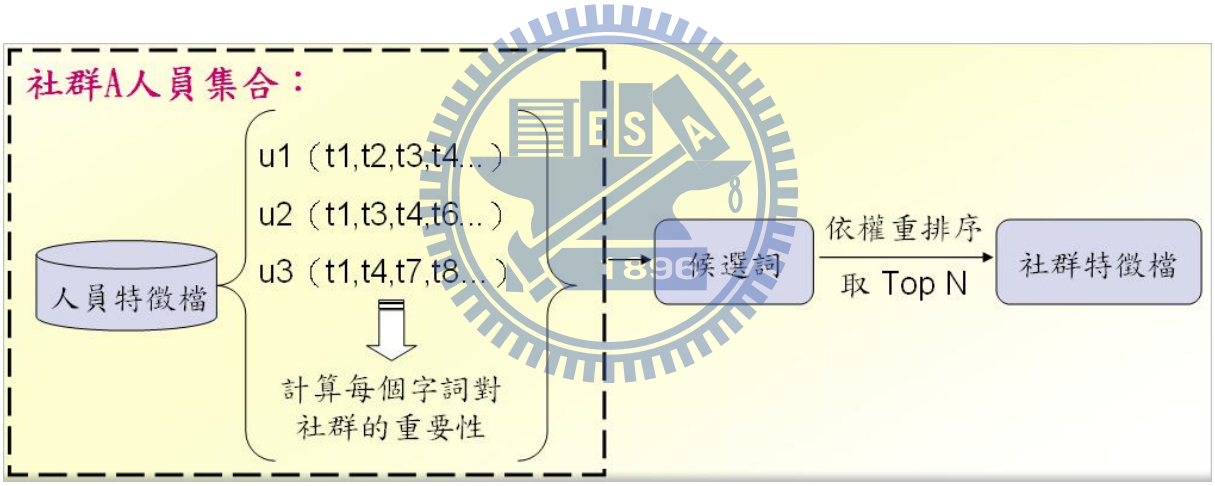

圖 3－9 社群特徵檔分析流程

$$GW_{g,i} = \frac{\sum_{u \in U_g} w_{i,u}}{|U_g|} \qquad \text{(公式 3-6)}$$

- $GW_{i,j}$：字詞 i 對社群 g 的權重分數
- U_g：社群 g 所含的所有使用者集合

- $w_{i,u}$：字詞 i 對使用者 u 的權重分數
- $|U_g|$：社群 g 所含的所有使用者人數

3.5. 推薦

本章節中，我們將會詳細介紹本研究提出以多社群為基礎之推薦步驟與方法，透過兩種相似度—知識興趣相似度與個人屬性相似度，按不同比例加總平均計算，每一種權重比例計算出的使用者相似度組合經過Clique分群法後都可獲得不同的分群結果。針對每一種分群結果，本研究使用混合式推薦法進行推薦，內容式推薦分析使用者知識興趣與文件的相關程度計算出「興趣指數」，本論文提出四種混合式推薦方法計算文件預測「興趣分數」，分別是「社群最大分數法（Community Max Score）」、「社群興趣最大分數法（Community Interest Max Score）」、「使用者權重最大分數法（User Weight Max Score）」、「使用者權重興趣最大分數法（User Weight Interest Max Score）」，四種方法皆會透過調整內容式推薦與協同式推薦比例計算出文件預測「興趣分數」，得到推薦清單。

3.5.1. 多社群文件推薦

多社群之文件推薦分成三大步驟，首先選擇工作者參與的知識社群，接著預測知識工作者對文章的喜好分數，最後根據此分數產生推薦清單。各步驟說明如下:

(一) 社群選擇

社群選擇即是從所有的知識社群中，挑選出使用者參與的所有社群。使用者有可

能參與一個以上的社群，即擁有多元的興趣領域。

(二) 「興趣分數」預測

本研究是採用混合式過濾推薦為基礎之推薦方法，內容式推薦過濾部分藉由計算使用者感興趣之知識內容與該文件特徵檔相似程度，作為使用者對該文件的預測「興趣指數」。使用者對文件之興趣指數是透過用文件與使用者相似度計算，計算方式如公式3-7所述：

$$COS_{Di,Uj} = \frac{\sum_{t=1}^{k} w_{t,Di} \times w_{t,Uj}}{\sqrt{\sum_{t=1}^{k} w_{t,Di}^2} \times \sqrt{\sum_{t=1}^{k} w_{t,Uj}^2}} \qquad \text{(公式 3-7)}$$

- $COS_{Di,Uj}$： 以cosine方式計算，文件 Di 與使用者 Uj 之興趣指數
- $w_{t,Di}$： 字詞 t 在文件 Di 特徵檔中的權重分數
- $w_{t,Uj}$： 字詞 t 在使用者 Uj 特徵檔中的權重分數

協同過濾推薦部分藉由與目標使用者相似度高的鄰近使用者，以其興趣、喜好、閱讀行為來預測目標使用者對文件的興趣程度，故預測時會考量相似鄰居和目標使用者的相似度。此步驟會根據上一步驟選出的興趣社群，分別將各社群內的成員對所有知識文件的興趣指數計算出來，並考慮社群內成員與目標使用者的相似度。最後混合內容式推薦「興趣指數」計算出預測「興趣分數」，以預測目標使用者在各知識領域中，對文件的興趣程度。

(三) 推薦清單

根據所預測之目標使用者對文件的興趣程度由高至低做排序作為推薦清單中。本研究以多社群為基礎，提出四種混合式文件推薦方法：「社群最大分數法（Community Max Score）」、「社群興趣最大分數法（Community Interest Max Score）」、「使用者權重最大分數法（User Weight Max Score）」、「使用者權重興趣最大分數法（User Weight Interest Max Score）」，以下四小節將介紹這四種推薦方法。

3.5.2. 社群最大分數法（Community Max Score - CMS）

社群最大分數法主要目的是為滿足知識工作者多元的興趣知識，針對目標使用者感興趣的所有領域，分別預測使用者對該文章落於每一領域時的「興趣分數」，進而從各社群中產生滿足工作者興趣之推薦清單。在本研究中，由於知識工作者參與的知識社群不單只有一個，在預測「興趣分數」時，會有一篇文章在多個知識社群中出現且具有不同預測「興趣分數」。

例如，預測目標使用者對一篇「太陽能未來十年分析」議題文章的分數，由於根據目標工作者參與「太陽能產業分析」和「健康醫療」兩個知識社群，因而會產生一篇文章有兩個預測「興趣分數」。最後選擇最高的「興趣分數」代表文件的「興趣分數」，意為「興趣分數」越高，表示在該知識興趣領域中，此知識文件是目標知識工作者感興趣的，且代表此文件與此知識社群相關性也較高，最後根據「興趣分數」高低進行文件推薦。

社群最大分數法之「興趣分數」計算方式如公式3-8所述，以目標使用者所屬的各知識社群為主，找出有點閱該文件之相似使用者，計算其與目標使用者相似度加總平

均，做為該社群鄰居使用者對此文件整體興趣程度，搭配特定比例之目標使用者對此文件「興趣指數」，計算出最後文件預測「興趣分數」。分別計算目標使用者在不同社群中對文件的預測「興趣分數」，最後挑選出分數最高的文件預測「興趣分數」，代表知識工作者*u*的文件*d*的預測「興趣分數」：

$$\hat{P}_{u,d} = \underset{G \in UG}{MAX}\left(\alpha \times II_{u,d} + (1-\alpha) \times \frac{\sum_{i \in G_{r,d}} SIM_{u,i}}{\left|G_{r,d}\right|} \right) \qquad \text{(公式 3-8)}$$

- $\hat{P}_{u,d}$：知識工作者 u 對文件 d 的預測「興趣分數」
- UG：使用者 u 參與的所有社群
- α：混合式推薦中內容式推薦比例（0~1）
- $II_{u,d}$：使用者 u 對文件 d 的興趣指數
- $G_{r,d}$：社群 G 中有點閱文件 d 之使用者集合
- $SIM_{u,i}$：使用者 u 和相似使用者 i 的相似度

透過上述公式3-8計算出目標使用者對文件的預測「興趣分數」，在不同興趣社群中，每一篇知識文件會得到不同的預測「興趣分數」。範例說明如下：

預測使用者U1對文件D1的「興趣分數」，首先選擇U1有參與的知識興趣社群，有G1、G2，如表3-7所示，計算各成員與目標使用者U1的相似度及分析各成員對D1點閱行為。由於不同社群分別代表著不同興趣知識內容，故同一文件D1對目標使用者在不

同知識社群會得到不同的「興趣分數」。

表 3－7 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表

	G1			G2		
	U1	**U2**	**U3**	**U1**	**U3**	**U4**
D1（興趣指數）	0.2	0.7	0.2	0.2	0.8	0.4
ifRead	-	1	0	-	1	1
SIM（與使用者 U1 相似度）	1	0.5	0.4	1	0.5	0.6
UW（使用者與社群相似度）	0.33	0.6	0.5	0.33	0.7	0.8

利用公式計算目標使用者 U1 對文件 D1 預測分數，當 α=0.3，結果如下所示：

$$\hat{P}_{u,d} = MAX\left(0.3\times 0.2 + 0.7\times\frac{0.5}{1} \quad ， \quad 0.3\times 0.2 + 0.7\times\frac{0.5+0.6}{2}\right)$$

$$= MAX(0.41，0.445) = 0.445$$

最後選擇G1和G2的最大預測「興趣分數」，即0.445為目標使用者U1對文件D1的預測「興趣分數」。

3.5.3. 社群興趣最大分數法（Community Interest Max Score - CIMS）

社群興趣最大分數法與社群最大分數法主要差異在於協同推薦部份，多加入分析相似使用者本身對該文件興趣程度(「興趣指數」)，提高內容式推薦混合比例。此方法亦可滿足知識工作者多元的興趣知識，針對目標使用者感興趣的所有領域，分別預測使用者對該文章落於每一領域時的「興趣分數」，進而從各社群中產生滿足工作者興趣之推薦清單。

社群興趣最大分數法之「興趣分數」計算方式如公式3-9所述，以目標使用者所屬的各知識社群為主，找出有點閱該文件之相似使用者，計算相似使用者對該文件「興趣指數」、其與目標使用者相似度後加總平均，做為該社群鄰居使用者對此文件整體興趣程度，搭配特定比例之目標使用者對此文件「興趣指數」，計算出最後文件預測「興趣分數」。分別計算目標使用者在不同社群中對文件的預測「興趣分數」，最後挑選出分數最高的文件預測「興趣分數」，代表知識工作者*u*的文件*d*的預測「興趣分數」：

$$\hat{P}_{u,d} = \underset{G \in UG}{MAX}\left(\alpha \times II_{u,d} + (1-\alpha) \times \frac{\sum_{i \in G_{r,d}} SIM_{u,i} \times II_{i,d}}{\left| G_{r,d} \right|} \right) \qquad \text{(公式 3-9)}$$

- $\hat{P}_{u,d}$：知識工作者 u 對文件 d 的預測「興趣分數」
- UG：使用者 u 參與的所有社群
- α：混合式推薦中內容式推薦比例（0~1）
- $II_{u,d}$：使用者 u 對文件 d 的興趣指數
- $G_{r,d}$：社群 G 中有點閱文件 d 之使用者集合
- $SIM_{u,i}$：使用者 u 和相似使用者 i 的相似度
- $II_{i,d}$：相似使用者 i 對文件 d 的興趣指數

透過上述公式3-9計算出目標使用者對文件的預測「興趣分數」，在不同興趣社群中，每一篇知識文件會得到不同的預測「興趣分數」。範例說明如下：

預測使用者U1對文件D1的「興趣分數」，首先選擇U1有參與的知識興趣社群，有G1、G2，如表3-8所示，計算各成員與目標使用者U1的相似度及分析各成員對D1點閱行為。由於不同社群分別代表著不同興趣知識內容，故同一文件D1對目標使用者在不同知識社群會得到不同的「興趣分數」。

表 3－8 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表

	G1			G2		
	U1	U2	U3	U1	U3	U4
D1（興趣指數）	0.2	0.7	0.2	0.2	0.8	0.4
ifRead	-	1	0	-	1	1
SIM（與使用者 U1 相似度）	1	0.5	0.4	1	0.5	0.6
UW（使用者與社群相似度）	0.33	0.6	0.5	0.33	0.7	0.8

利用公式計算目標使用者 U1 對文件 D1 預測分數，當 α=0.3，結果如下所示：

$$\hat{P}_{u,d} = MAX\left(0.3\times0.2+0.7\times\frac{0.5\times0.7}{1}\text{，}0.3\times0.2+0.7\times\frac{0.5\times0.8+0.6\times0.4}{2}\right)$$

$$= MAX(0.305\text{，}0.248) = 0.305$$

最後選擇G1和G2的最大預測「興趣分數」，即0.305為目標使用者U1對文件D1的預測「興趣分數」。

3.5.4. 使用者權重最大分數法（User Weight Max Score - UWMS）

使用者權重最大分數法與社群最大分數法不同點主要是使用者權重最大分數法多考量社群內成員與該社群的相關性，故多加入計算社群內成員與社群相似度，當作預測目標使用者「興趣分數」參數之一。由於目標使用者可參與一個以上的知識社群，

則一篇文章在不同社群會有不同預測分數，最後選擇最大之預測分數來當成文件的「興趣分數」。將所有文件預測「興趣分數」排序列出對目標知識工作者最後推薦清單。使用者權重最大分數法文件的分數預測計算方式如公式3-10所述：

$$\hat{P}_{u,d} = \underset{G \in UG}{MAX}\left(\alpha \times II_{u,d} + (1-\alpha) \times \frac{\sum_{i \in G_{r,d}} SIM_{u,i} \times UW_{i,G}}{\left| G_{r,d} \right|} \right) \qquad \text{(公式 3-10)}$$

- $\hat{P}_{u,d}$：知識工作者 u 對文件 d 的預測「興趣分數」
- UG：使用者 u 參與的所有社群
- α：混合式推薦中內容式推薦比例（0~1）
- $II_{u,d}$：使用者 u 對文件 d 的興趣指數
- $G_{r,d}$：社群 G 中有點閱文件 d 之使用者集合
- $SIM_{u,i}$：使用者 u 和相似使用者 i 的相似度
- $UW_{i,G}$：相似使用者 i 在社群 G 的權重 (以使用者和該社群之相似度表示)

上述公式中的G是指所有目標使用者有參與的知識社群，先針對每一社群內所有相似使用者，根據每位相似使用者與目標使用者相似度，乘上此相似使用者在此社群中的權重，再做一正規化處理得出群內相似使用者對文件推薦興趣程度。最後再與目標使用者本身對該文件的「興趣指數」以權重調整加總後計算出最後使用者對文件預測「興趣分數」。範例說明如下：

預測使用者U1對文件D1的「興趣分數」，首先選擇U1有參與的知識興趣社群，有

G1、G2，如表3-9所示，計算各成員與目標使用者U1的相似度及分析各成員對D1點閱行為。由於不同社群分別代表著不同興趣知識內容，故同一文件D1對目標使用者在不同知識社群會得到不同的「興趣分數」。

表 3－9 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表

	G1			G2		
	U1	U2	U3	U1	U3	U4
D1（興趣指數）	0.2	0.7	0.2	0.2	0.8	0.4
ifRead	-	1	0	-	1	1
SIM（與使用者 U1 相似度）	1	0.5	0.4	1	0.5	0.6
UW（使用者與社群相似度）	0.33	0.6	0.5	0.33	0.7	0.8

利用公式計算目標使用者 U1 對文件 D1 預測分數，當 α=0.3，結果如下所示：

$$\hat{P}_{u,d} = MAX\left(0.3\times0.2+0.7\times\frac{0.5\times0.6}{1}\ ,\ 0.3\times0.2+0.7\times\frac{0.5\times0.7+0.6\times0.8}{2}\right)$$

$$= MAX(0.27\ ,\ 0.3505) = 0.3505$$

最後選擇G1和G2的最大預測「興趣分數」，即0.3505為目標使用者U1對文件D1的預測「興趣分數」。

3.5.5. 使用者權重興趣最大分數法（User Weight Interest Max Score - UWIMS）

使用者權重興趣最大分數法與使用者權重最大分數法主要差異在於協同推薦部份，多加入分析相似使用者本身對該文件興趣程度（「興趣指數」），提高內容式推薦混合比例。此方法亦可滿足知識工作者多元的興趣知識，針對目標使用者感興趣的所有

領域，分別預測使用者對該文章落於每一領域時的「興趣分數」，進而從各社群中產生滿足工作者興趣之推薦清單。使用者權重興趣最大分數法文件的分數預測計算方式如公式3-11所述：

$$\hat{P}_{u,d} = \underset{G \in UG}{MAX}\left(\alpha \times II_{u,d} + (1-\alpha) \times \frac{\sum_{i \in G_{r,d}} SIM_{u,i} \times UW_{i,G} \times II_{i,d}}{\left| G_{r,d} \right|} \right) \qquad \text{(公式 3-11)}$$

- $\hat{P}_{u,d}$：知識工作者 u 對文件 d 的預測「興趣分數」
- UG：使用者 u 參與的所有社群
- α：混合式推薦中內容式推薦比例（0~1）
- $II_{u,d}$：使用者 u 對文件 d 的興趣指數
- $G_{r,d}$：社群 G 中有點閱文件 d 之使用者集合
- $SIM_{u,i}$：使用者 u 和相似使用者 i 的相似度
- $UW_{i,G}$：相似使用者 i 在社群 G 的權重 (以使用者和該社群之相似度表示)
- $II_{i,d}$：相似使用者 i 對文件 d 的興趣指數

上述公式中的G是指所有目標使用者有參與的知識社群，先針對每一社群內所有相似使用者，根據每位相似使用者與目標使用者相似度，乘上此相似使用者在此社群中的權重、該相似使用者對該文件的「興趣指數」，再做一正規化處理得出群內相似使

用者對文件推薦興趣程度。最後再與目標使用者本身對該文件的「興趣指數」以權重調整加總後計算出最後使用者對文件預測「興趣分數」。範例說明如下：

預測使用者U1對文件D1的「興趣分數」，首先選擇U1有參與的知識興趣社群，有G1、G2，如表3-9所示，計算各成員與目標使用者U1的相似度及分析各成員對D1點閱行為。由於不同社群分別代表著不同興趣知識內容，故同一文件D1對目標使用者在不同知識社群會得到不同的「興趣分數」。

表 3－10 目標使用者 U1 所屬社群 G1 及社群 G2 資訊表

	G1			G2		
	U1	**U2**	**U3**	**U1**	**U3**	**U4**
D1（興趣指數）	0.2	0.7	0.2	0.2	0.8	0.4
ifRead	-	1	0	-	1	1
SIM（與使用者 U1 相似度）	1	0.5	0.4	1	0.5	0.6
UW（使用者與社群相似度）	0.33	0.6	0.5	0.33	0.7	0.8

利用公式計算目標使用者 U1 對文件 D1 預測分數，當 α=0.3，結果如下所示：

$$\hat{P}_{u,d} = MAX\left(0.3\times0.2+0.7\times\frac{0.5\times0.6\times0.7}{1}\text{，}0.3\times0.2+0.7\times\frac{0.5\times0.7\times0.8+0.6\times0.8\times0.4}{2}\right)$$

$$= MAX(0.207\text{，}0.2252) = 0.2252$$

最後選擇G1和G2的最大預測「興趣分數」，即0.2252為目標使用者U1對文件D1的預測「興趣分數」。

4. 實驗與評估

在本章節中，我們利用企業部落格之實際的資料，以驗證本研究所提出之自動化分類與推薦方法，的確能協助企業提昇知識管理品質與知識文件使用率。以下就實驗資料、評估標準、實驗工具、實驗結果說明。

4.1. 實驗資料

本研究所使用之實驗資料，是取自某企業部落格平台，截至本實驗資料蒐集完成階段，共有350部落格，有效持續經營之部落格約有100個，另外單月平均點閱人次約有6000～7000人，資料月份涵蓋2009年全年，各部落格設有權限控管機制，本研究實驗資料僅取用繁體中文文件內容。本研究分為分類、推薦兩模組，各模組使用資料將於實驗敘述中說明。

4.2. 評估標準

本研究採用之評估標準分別為Precision。Precision是指查準率。以自動化分類模組為例，即計算由系統自動分類文件中，其符合正確分類之文件的比例。分子為系統分類正確的文件數，分母為系統分類之所有文件數。計算方式如公式4-1所述：

$$\Pr ecision = \frac{\#\,of\ correctly\_classified\_documents}{\#\,of\ total\_classified\_documents} \qquad \text{(公式 4-1)}$$

例如若系統分類100篇文章，其中70篇為分類正確，則Precision為0.7。

推薦之評估標準為計算推薦的文章中，其符合使用者興趣範圍相關文件的比例，以分析系統推薦文章的準確性。分子為系統推薦且符合使用者需求的文章數，分母為系統推薦的文章數。計算方式如公式4-2所述：

$$\Pr ecision = \frac{\#\ of\ correctly\_recommended\_documents}{\#\ of\ \ total\_recommended\_documents} \qquad \text{(公式 4-2)}$$

本研究針對推薦驗證方式，是採用使用者歷史點閱記錄進行Precision驗證，因為本研究所提出之混合式過濾法，其中協同過濾推薦部份之特性即為發掘、推薦目標使用者未發現之領域的文件，在本研究之企業部落格中，因部落格數量多，且部分文件、站台有權限卡控機制，即使系統所推薦之文件是目標使用者知識興趣相關，但在目標使用者歷史點閱紀錄中，仍有可能該目標使用者不知該篇文件的存在，抑或對該文件不具有存取權限，故即使系統推薦之文件符合目標使用者知識興趣，驗證Precision之歷史點閱紀錄中，該目標使用者仍無此文件之點閱紀錄，導致實驗分析驗證時出現Precision低估現象。針對上述現象，本研究之Precision評估法會有以下前提，系統推薦之文件，若為使用者未曾點閱過之部落格所屬文件，將此推薦文件移除不納入Precision評估，避免因推薦文件為使用者無權限、未知站台相關文件導致Precision低估。透過此評估方法，可藉由實際歷史點閱紀錄，驗證本研究所提出之自動化文件推薦方法精準度。

4.3. 實驗工具

本研究使用之工具如下：

1. Microsoft Visual Studio 2005 C# 開發程式

2. 資料庫使用Microsoft SQL Server 2005作為資料儲存工具

3. 文件本文內容之斷詞、詞性擷取、詞頻計算，皆透過中研院斷詞研究小組提供之CKIP Service。

4.4. 實驗結果與評估

4.4.1. 分類

分類實驗先由部落格管理人員，根據企業需求定義，事先定義類別，再於各類中挑選100筆文章作為實驗訓練資料，經部落格管理人員針對企業需求、部落格平台目前資料分析後，共取得有效分類16類如下表所述。在隨機於各類別文件選取40篇作為測試資料進行實驗。

表 4－1 部落格有效分類表

Type	Category	Type	Category
內部流程相關	Company-OU	**外部環境相關**	經濟動態
	Company-BU		面板產業動態
	Company-IU-HR		太陽能產業動態
	Company-TU	**企業管理面向、其他**	商業企管
	Company-IU-FIN		智權
	Company-IU-IT		Green
	Company-IU-Q		休閒育樂
	Company-IU-ESH		旅遊美食

	Company-IU-LG		消費生活
	-		健康醫療

實驗結果如下：

表 4－2 分類實驗結果

Type	Category	TrainData	Exp Data	Err	Correct
內部流程相關	Company-OU	100	40	7	33
	Company-IU-HR	100	40	8	32
	Company-TU	100	40	7	33
	Company-IU-IT	100	40	4	36
	Company-IU-Q	100	40	5	35
	Company-IU-ESH	100	40	6	34
外部環境相關	經濟動態	100	40	8	32
	面板產業動態	100	40	4	36
	太陽能產業動態	100	40	2	38
企業管理面向、其他	商業企管	100	40	11	29
	智權	100	40	2	38
	Green	100	40	6	34
	休閒育樂	100	40	6	34
	旅遊美食	100	40	8	32
	消費生活	100	40	7	33
	健康醫療	100	40	6	34
Total	**16**	**1600**	**640**	**97**	**543**

$$\Pr ecision = \frac{543}{640} = 84.84\%$$

4.4.2. 推薦

本研究以企業部落格之以2009/01月之部落格資料為推薦訓練資料，2009/02之部落格資料作為測試資料。在社群分析過程中，部份目標使用者與其他使用者間的相似度皆低於門檻值，沒有被分到任何知識社群中，故系統也不會產生推薦清單，此類使用者系統將隨機推薦文件作為系統推薦清單。

表 4－3 推薦實驗資料

項目	訓練文件	實驗文件
筆數	964	876

項目	使用者
人數	30

本研究實驗過程中需要設定幾個參數，部分為實驗中動態參數，部分參數為參考文獻最佳值、經過實際實驗測試取得下列最佳值，以進行後續研究分析：

1. 特徵檔之特徵詞個數：30。

2. β：複合使用者相似度中，使用者屬性相似度所佔之比例，為實驗中動態參數。本實驗使用四組參數，β=0、β=0.2、β=0.3、β=0.4。

3. Clique分群使用者相似度門檻：0.4。

4. α：混合式過濾推薦法中，內容式推薦法所佔之比例，為實驗中動態參數。本實驗使用十一組參數，α=0、0.1、0.2、0.3、0.4 … 0.8、0.9、1。

實驗結果如下：

本次實驗針對本文所提出之四種方法，搭配兩動態（α、β）參數，取30位使用者進行系統推薦實驗分析。以此30位使用者於2009/01點閱2009/01發布之文件紀錄作為訓練資料，此30位使用者於2009/02點閱2009/02發布之文件記錄作為測試評估依據。將系統針對30位使用者推薦清單，排除推薦清單中有使用者未曾拜訪過之部落格所屬文件、無權限文件，每位使用者取Top 20推薦清單進行Precision評估，獲得下列結果：

（一）Precision – CMS：

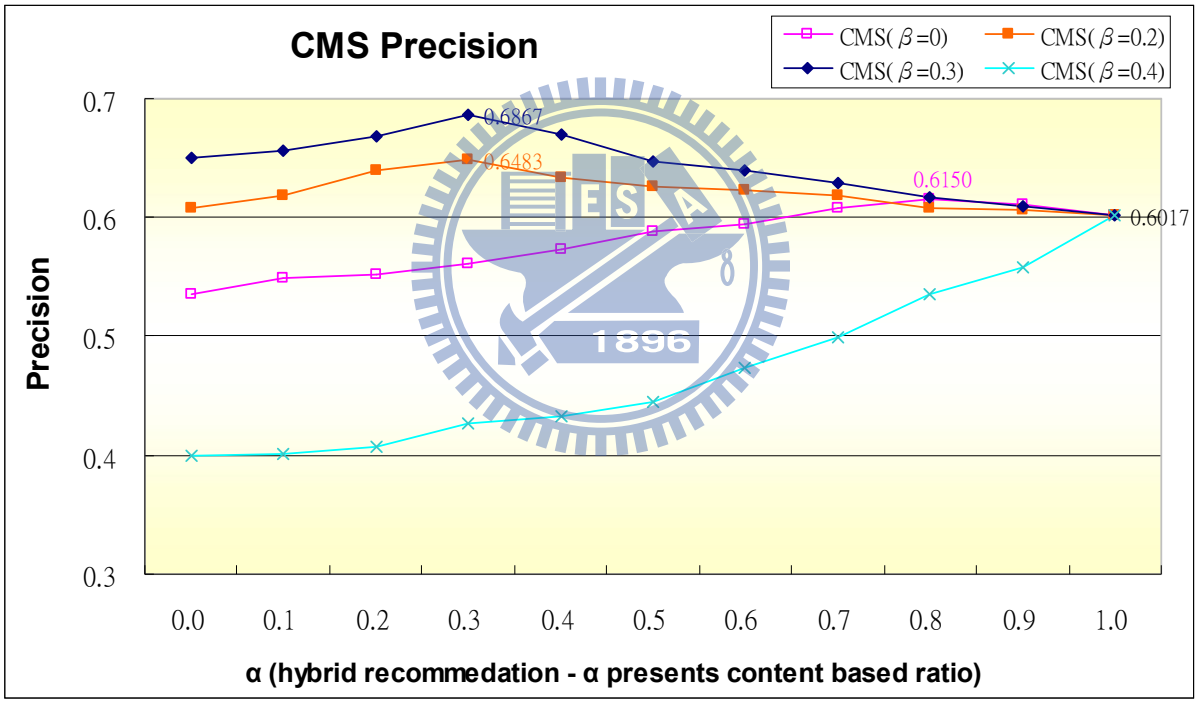

圖 4－1 Precision－CMS

圖4-1為使用CMS方法進行推薦的準確率整體分析，整體表現受使用者屬性相似度參數（β）影響最大，當**β=0.3**時（即使用者屬性相似度為0.3、興趣相似度0.7情況下），使用CMS整體表現最好，且配合**α=0.3**（即內容式推薦比例為0.3、協同式推薦比例為0.7情況下）出現Precision最佳值**0.6867**。β=0.2時搭配不同α比例整體表

現線形和β=0.3相似，都是在α=0.3表現最好，但整體表現都較β=0.3差。另外實驗發現，當β=0.4時，整體Precision隨著推薦方法中之協同比例提高（α遞減）卻持續下降至約0.4。β=0為完全採用使用者興趣相似度做推薦分群依據，整體Precision表現受協同比例提高（α遞減）稍微提高後即漸漸下降，線型較為平穩，Precision僅較β=0.4好。

由上分析可得知，CMS整體Precision表現受使用者屬性相似度參數（β）影響較大，當完全使用使用者興趣相似度分群，30人中會有8人沒有任何與其相似度高於門檻值0.4之相似使用者，故沒有參與任何知識社群，推薦清單為隨機推薦，當β提高至0.3時，未被分群人數降至3人，故降低隨機推薦人數而提高準確率。但當β提高至0.4時，分群後之知識社群知識興趣相似度大為降低，反而導致群內知識興趣散亂不集中，整體推薦Precision大幅下滑。內容式過濾推薦參數比例（α）表現受分群結果影響，整體而言混合式過濾推薦方式較單一（內容式、協同式）過濾推薦Precision高，且須搭配有加入使用者屬性相似度分群結果，提高協同推薦比例較能提升Precision。

（二）Precision – CIMS：

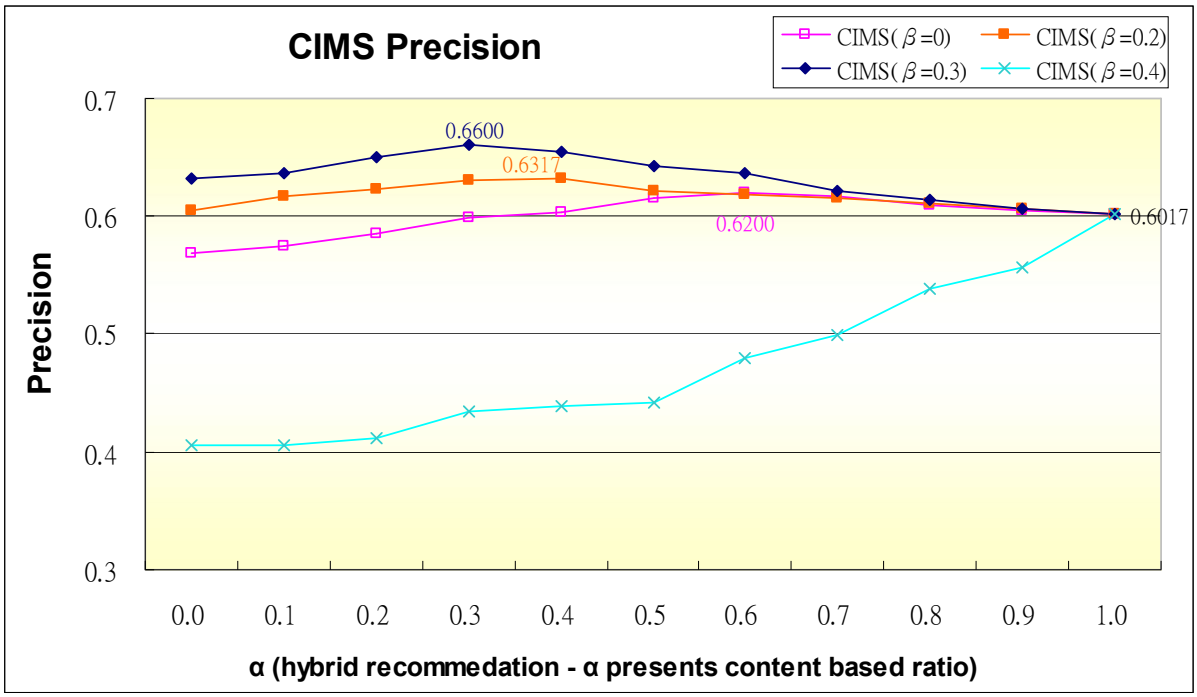

圖 4－2 Precision－CIMS

圖4-2為使用CIMS方法進行推薦的準確率整體分析，整體表現受使用者屬性相似度參數（β）影響最大，當**β=0.3**時（即使用者屬性相似度為0.3、興趣相似度0.7情況下），使用CIMS整體表現最好，且配合**α=0.3**（即內容式推薦比例為0.3、協同式推薦比例為0.7情況下）則出現Precision最佳值**0.66**。β=0.2時搭配不同α比例整體表現線形和β=0.3大致相似，當α=0.4表現最好為0.6317，整體而言表現都較β=0.3差。當β提高至0.4時，整體Precision隨著推薦方法中之協同比例提高（α遞減）卻持續下降至約0.4。β=0為完全採用使用者興趣相似度做推薦分群依據，整體Precision表現受協同比例提高(α遞減)先提高後即漸漸下降，當α≧0.5時期表現與β=0.2相近，整體Precision僅較β=0.4好。

由上分析可得知，CIMS整體Precision表現受使用者屬性相似度參數（β）影響較大，主要為增加使用者屬性相似度比例來計算分群依據之相似度，可減少未被分群

使用者數，故降低隨機推薦人數而提高準確率。但過高之使用者屬性相似度參數（β）比例（0.4），反而會使分群後之知識社群知識興趣相似度大為降低，導致群內知識興趣散亂不集中，大幅降低整體推薦Precision。內容式過濾推薦參數比例（α）表現受分群結果影響，整體而言混合式過濾推薦方式較單一（內容式、協同式）過濾推薦Precision高。

（三）Precision – UWMS：

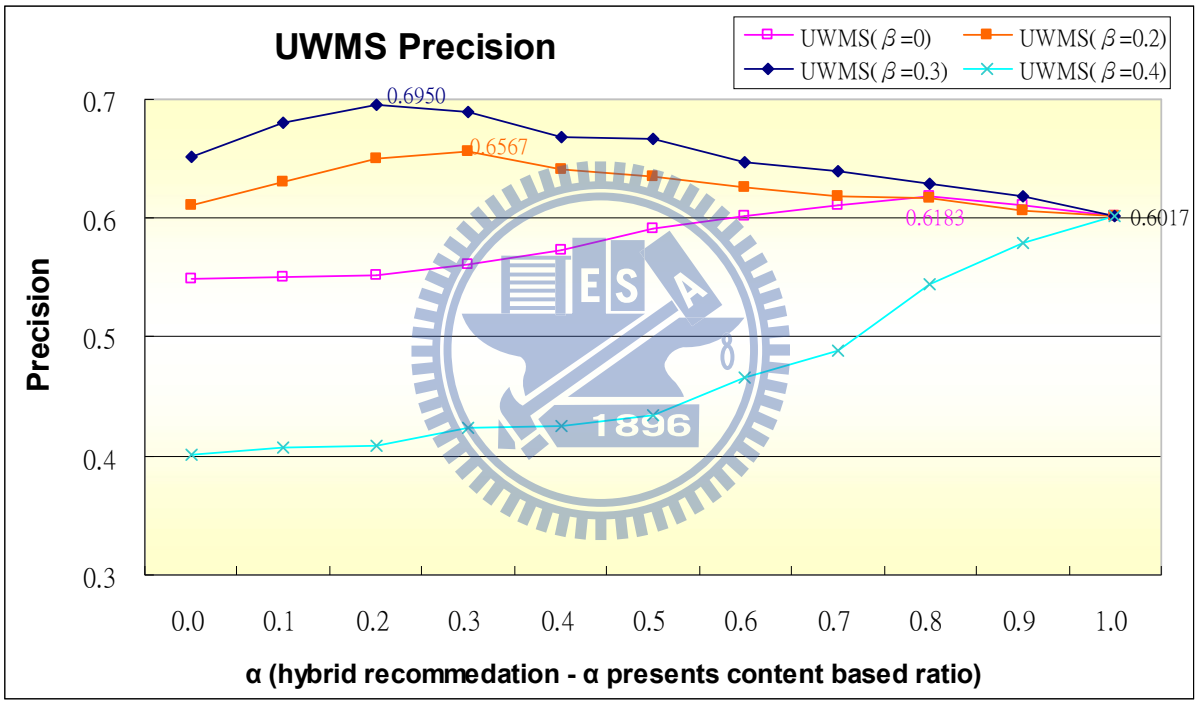

圖 4－3 Precision－UWMS

圖4-3為使用UWMS方法進行推薦的準確率整體分析，整體表現受使用者屬性相似度參數（β）影響最大，當**β=0.3**時（即使用者屬性相似度為0.3、興趣相似度0.7情況下），使用UWMS整體表現最好，且配合**α=0.2**（即內容式推薦比例為0.2、協同式推薦比例為0.8情況下）出現Precision最佳值**0.6950**。β=0.2時搭配不同α比例整體表現線形和β=0.3相似，α=0.3時表現最好，整體表現都較β=0.3差。當β提高至0.4

時，整體Precision隨著推薦方法中之協同比例提高（α遞減）卻持續下降至約0.4。β=0為完全採用使用者興趣相似度做推薦分群依據，整體Precision表現受協同比例提高（α遞減）稍微提高後即漸漸下降，線型較為平穩，Precision僅較β=0.4好。

由上分析可得知，UWMS整體Precision表現受使用者屬性相似度參數（β）影響較大，透過調整使用者屬性相似度參數（β）比例可影響分群結果與群內使用者知識興趣集合，可能降低未被分群隨機推薦使用者提高Precision，亦可能導致分群後之知識社群知識興趣相似度大為降低而降低Precision。內容式過濾推薦參數比例（α）表現受分群結果影響，整體而言混合式過濾推薦方式較單一（內容式、協同式）過濾推薦Precision高，且須搭配有加入使用者屬性相似度分群結果，提高協同推薦比例較能提升Precision。

（四）Precision –UWIMS：

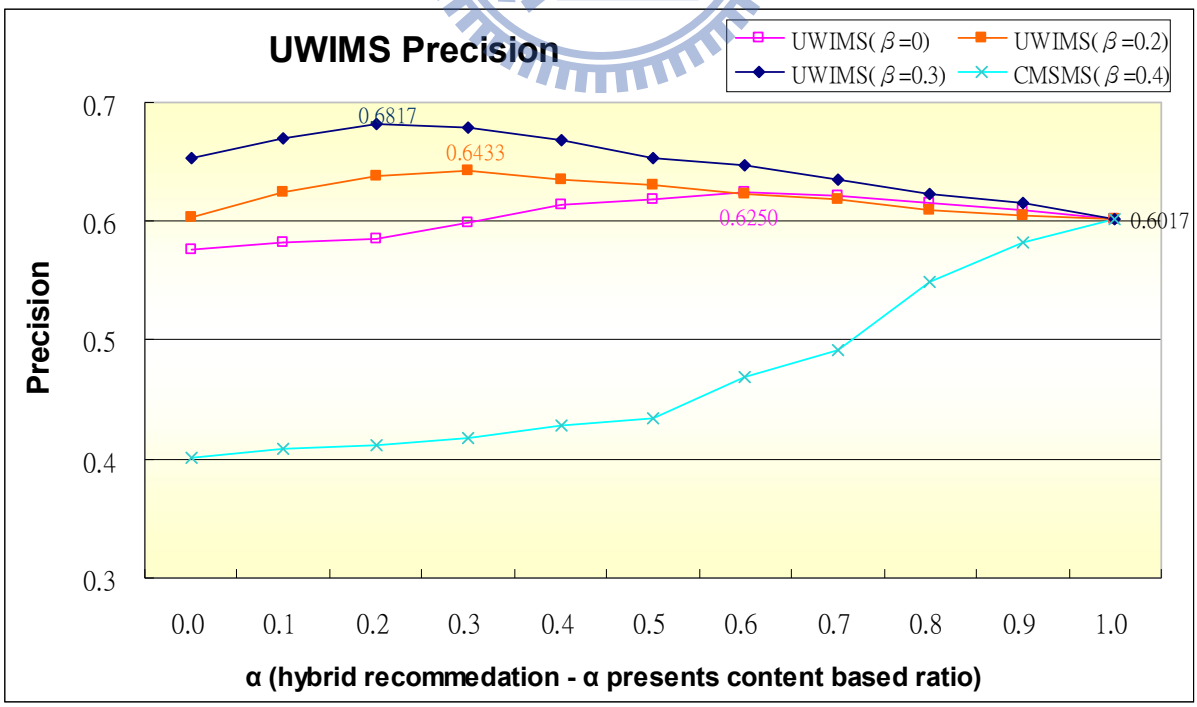

圖 4－4 Precision－UWIMS

圖4-3為使用UWIMS方法進行推薦的準確率整體分析，整體表現仍受使用者屬性相似度參數（β）影響最大，當**β=0.3**時（即使用者屬性相似度為0.3、興趣相似度0.7情況下），使用UWIMS整體表現最好，且配合**α=0.2**（即內容式推薦比例為0.2、協同式推薦比例為0.8情況下）則出現Precision最佳**0.6817**。β=0.2時搭配不同α比例整體表現線形和β=0.3相似，整體表現仍較β=0.3差。當β提高至0.4時，整體Precision隨著推薦方法中之協同比例提高（α遞減）快速下降。β=0時在α≧0.6表現與β=0.2相近，隨協同比例越提高（α遞減），Precision反而開始下降。β=0.4時，整體Precision隨著推薦方法中之協同比例提高（α遞減）卻持續快速下降。

由上分析可得知，UWIMS整體Precision表現受使用者屬性相似度參數（β）影響較大。內容式過濾推薦參數比例（α）表現受分群結果影響，整體而言混合式過濾推薦方式較單一（內容式、協同式）過濾推薦Precision高。

（三）整體方法評估：

根據上述各實驗後，我們整理整體推薦方法評估如圖4-5、圖4-6可得下列分析：

1. 加入使用者屬性相似度計算使用者相似度進行分群，影響到整體多社群分析後的分群的結果。可解決部分協同式推薦中使用者Cold Start 問題，提高推薦整體Precision。但是並非每一知識領域都和使用者屬性有絕對相關性，故本研究發現，在加入個人屬性檔相似度概念後，部分使用者出現系統所推薦清單Precision降低的現象，群內的使用者不見得完全都是擁有相同知識興趣領域，部分是因為個人屬性相關，故目標使用者運用群內相似鄰居進行「興趣分數」預測時會受到此類鄰近使

用者影響。故過高的使用者屬性相似度比例，會造成社群知識興趣不集中，降低推薦整體Precision。

2. 混合式推薦可較單一內容式推薦或是單一協同式推薦提高Precision，透過協同推薦機制，可發現目標使用者未察覺之興趣知識，擴大目標使用者知識興趣領域。

3. 多加入相似使用者於社群權重，可在進行推薦時，將與該社群知識興趣越相近之相似使用者推薦影響力提高，由CMS與UWMS比較與CIMS與UWIMS比較可看出，的確可提高整體推薦Precision。

4. 多考量相似使用者「興趣指數」，比較圖4-5中CMS與CIMS，CIMS於協同部分多考量相似使用者對該文件的「興趣指數」，使得整體內容式推薦比例增加，CIMS整體Precision較CMS低；UWMS與USIMS情況亦類似，相同 β 參數值分群結果下，多加入考量相似使用者對該文件的「興趣指數」，會降低整體Precision。

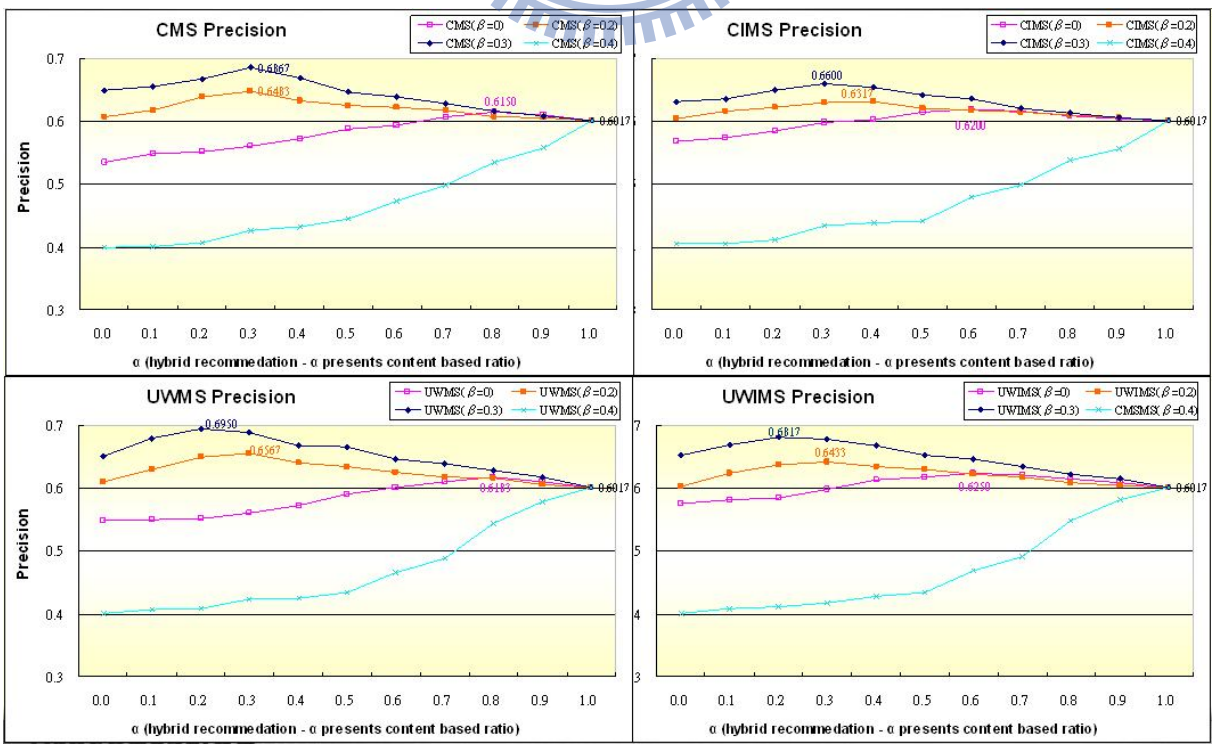

圖 4－5 Precision - Total

最後，本實驗CMS、CIMS、UWMS、UWIMS四種方法都是在β=0.3分群情況下表現最好，圖4.6整理了β=0.3時四種方法比較。四種方法在混合式推薦比例α為0.2、0.3時都可有較好的Precision，故混合式推薦的確可以提高推薦Precision。四種推薦方法Precision分別是UWMS＞CMS＞UWIMS＞CIMS，使用者權重最大分數法優於社群最大分數法，多考量相似使用者之興趣指數會降低Precision。

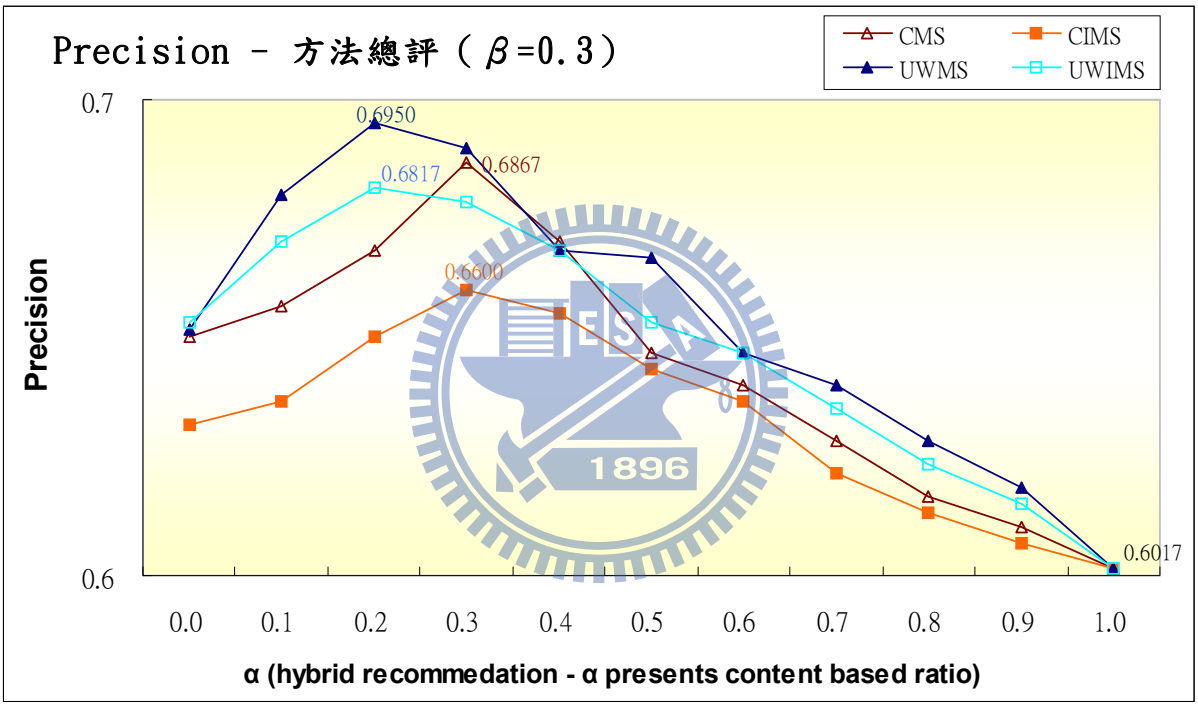

圖 4－6 Precision – 4 methods comparison

5. 結論與未來方向

5.1. 結論

現今企業競爭已經進入以創新思考、協同合作思維，以知識為基礎之後知識經濟時代，不斷創新及持續學習成為組織決勝的關鍵。隨著科技的進步，企業員工可獲得的資訊大幅增加，往往會耗費大量時間成本來尋找所需要的資訊，為了方便企業內知識的有效管理，並能將組織知識快速分享、運用，節省企業時間成本和人力，故建立一具有自動化分類機制與個人化文件推薦系統是必要的。由系統管理者定義符合企業需求之一致性分類，可幫助企業同仁更快速地取得資訊，另外也可分析企業組織知識的整體集中性、分布情形，再透過使用者的點閱行為，發揮群體力量進行點閱分析，可協助組織了解企業知識需求類別，爲知識加值。

本研究所提出之分類方法，係透過已分類文件的特徵檔歸納出該類別的特徵檔，運用此方式而不使用KNN或是SVM方式，是由於KNN演算法需要將每篇文件與所有已分類文件進行相似度比對，由於企業每天都持續有知識文件的增加，使用KNN會耗費較多時間；SVM方法則是要針對每項類別進行分類模型訓練，故也需要花費較多時間進行訓練、參數調整，故考量效能，本研究使用將各類別中所有已分類文件特徵檔分析成類別的特徵檔的方式，因此新進文件在類別分析時，只要與各類別特徵檔進行特徵檔相似度比對即可，可節省許多時間增進效能，只要在進行特徵詞分析過程中能有效過濾、擷取出正確關鍵的特徵詞，運用此方式分類方法經本研究實驗可達到80%以上。

另外本研究所使用之多知識社群分析來進行知識文件推薦，有效協助企業將部落格知識文件對員工個人化推薦，並有將近七成推薦準確率。從實驗結果可得知社群分

析過程除了考量群內使用者之知識興趣，使用者之個人屬性檔在部分知識興趣社群中也有關聯性存在，故在加入使用者特徵檔進行相似度分析後，對部分同仁推薦清單準確率的確有提高，但同時也有部分同仁推薦清單準確率降低，因為並非所有知識社群分布都與員工特徵有相關，在加入員工特徵檔後可解決部分員工Cold Start（缺乏點閱紀錄）的問題，故可提高整體的推薦準確率。

5.2. 未來方向

（一）多類別分類機制

本研究所提出的自動化分類機制，僅針對文件進行單一分類的研究，但從實驗結果可發現，許多文件所包含的主題涵蓋了許多類別，導致進行分類時會出現難以分類的狀況。舉例而言，近年來因為環保意識抬頭，企業營運、產品、策略都受到環保意識影響，許多議題都與環境保護有一定的相關，故企業也定義了「Green」類別蒐集相關環保議題，當一篇文章在探討經濟市場中碳交易時，該歸屬於「經濟動態」類別或「Green」類別難以界定，且由於文件議題涵括兩大主題，會導致與分類錯誤的機率提高。故未來方向可針對文件進行多類別分析。

（二）類別、知識社群對應

目前推薦機制所使用的多知識社群分析，社群本身代表意義即為特定的知識領域，目前知識社群結果並無分析知識社群類型，也無法確保目前知識社群分析結果品質，若可與自動化分類結果進行比對分析該知識社群所屬知識領域為何，相互調整已達最佳化分類定義、知識社群數，讓兩大機制更為完整。

（三）增加個人屬性資訊

本研究在進行知識社群分析時，除了考量使用者知識興趣相似程度外，還加入了組織內使用者個人屬性檔關係進行分群，目前僅使用「部門」、「年齡」、「性別」、「年資」、「職級」五大屬性，可再多蒐集一些使用者屬性進行分析。另外，若能再進一步分析使用者在企業流程中相關性程度並作為知識社群分群依據，可使得資訊推薦制更加有效率。

（四）知識社群、個人屬性相關性分析

並非每一知識社群都與此五大使用者相關屬性相關，本研究未針對各社群知識領域分析使用者屬性影響程度，故可針對使用者屬性與各知識社群相關程度進行研究調整權重，避免因加入使用者屬性而導致推薦準確率降低。

Reference

English：

[1] Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan,P. (1998), "Automatic subspace clustering of high dimensional data for data mining applications," ACM SIGMOD Record, 27(2): 94-105.

[2] Alani, H., Dasmahapatra, S., O'Hara, K. and Shadbolt, N. (2003), "Identifying Communities of Practice through Ontology Network Analysis," IEEE Intelligent Systems, 18(2): 18–25.

[3] Baeza-Yates, R. and Ribeiro-Neto, B. (1999), "Modern Information Retrieval," Addison-Wesley Longman Publishing Company.

[4] Balabanovic, M and Shoham, Y. (1997), "Fab: Content-Based, Collaborative Recommendation," Communications of the ACM 40(3): 66-72.

[5] Cortes C. and Vapnik V., (1995), Support-Vector Networks, Machine Learning.

[6] Cover, T. M., and P. E. Hart, (1967), Nearest Neighbor Pattern Classification. Institute of Electrical and Electronics Engineers Transactions on Information Theory, Vol. 13, No. 1, pp. 21-27.

[7] Glance, N., Arregui, D. and Dardenne, M. (1997), "Knowledge pump: Community-centered collaborative filtering," Fifth DELOS Workshop. Filtering and Collaborative Filtering. Budapest, ERCIM report, ERCIM-98-W001.

[8] Goldberg, D., Nichols, D., Oki B.M. and Terry, G. (1992), "Using Collaborative Filtering to Weave an Information Tapestry," Communications of the ACM, 35(12):

61-70.

[9] Hanani, U., Shapira, B. and Shoval, P. (2001), "Information filtering: Overview of issues, research and systems," User Modeling and User Adapted Interaction 11(3): 203-259.

[10] Konstan, J.A., Miller B.N., Maltz, D., Herlocker, J.L., Gordon L.R. and Riedl, J. (1997), "GroupLens: applying collaborative filtering to Usenet news," Communications of the ACM, 40(3): 77-87.

[11] Mitra, M. and Chaudhuri, B.B. (2000), " Information Retrieval from Documents: A Survey," Information Retrieval, 2: 141–163.

[12] Mooney, R. J. and Roy, L. (2000), "Content-based book recommending using learning for text categorization," Proceedings of the ACM International Conference on Digital Libraries, 195-204.

[13] Robertson, S.E. (1981), "The methodology of information retrieval experiment," Information Retrieval Experiment. Chapt. 1: 9-31.

[14] Rucker, J. and Polanco, M. J. (1997), "Siteseer: Personalized navigation for the web.," Communications of the ACM, 40(3): 73-75.

[15] Salton, G., Wong, A. and Yang, C.S. (1975), A vector space model for automatic indexing. Communications of the ACM, 18(11):613-620.

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

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

[18] Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. , (2001), “Item-based collaborative filtering recommendation algorithms,” Proceedings of the 10th international conference on World Wide Web, 285-295 .

[19] Shardanand, U. and Maes, P. (1995), “Social information filtering: Algorithms for automating "word of mouth",” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 210-217.

[20] Wagner, JC and Bolloju, N. (2005), Supporting Knowledge Management in Organizations with Conversational Technologies: Discussion Forums, Weblogs and Wikis, Journal of Database Management, 16, 2, April-June, pp i-viii.

Chinese：

[21] 中央研究院，中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw

[22] 朱家聲 (2006) ，“部落格在知識型企業的發展方向之研究”， 實踐大學企業創新發展研究所論文

[23] 廖佩君 (2008)，“以多社群為基礎之部落格文件推薦.”，交通大學資訊管理研究所碩士論文。