國立交通大學

資訊管理研究所

碩士論文

以標籤為基礎之個人化文件推薦

Personalized Document Recommendation Based on Tags

研究生：林子瑋

指導教授：劉敦仁 博士

中華民國九十八年六月

以標籤為基礎之個人化文件推薦

Personalized Document Recommendation

Based on Tags

研 究 生：林 子 瑋　　Student: Tzu-Wei Lin

指導教授：劉 敦 仁　　Advisor: Dr. Duen-Ren Liu

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
Information Management
June 2009
Hsinchu, Taiwan, the Republic of China

中 華 民 國 九 十 八 年 六 月

以標籤為基礎之個人化文件推薦

研究生：林子瑋　　指導教授：劉敦仁 博士

國立交通大學資訊管理研究所

摘要

資訊科技日漸成熟與網際網路逐漸普及化後，以 web2.0 為概念發展的網站成為近年趨勢。標籤標記(tagging)已變成 web2.0 網站熱門技術，讓使用者不僅藉由標籤註記與分類資訊內容，還可以有效且快速管理與組織資訊。另外，衍生標籤建議、查詢語擴展、整合標籤於推薦系統等應用服務。標籤標記透過群體合作方式增加新的詮釋資料。大多數整合標籤於推薦的研究都假定相同標籤文字具有同樣意義，而忽略標籤語意的模糊性。

本研究提出以標籤為基礎之個人化文件推薦。本研究認為可以將一個標籤視為一個興趣，相同標籤在不同使用者間具有不同涵義，因此本研究以使用者標記文章建立專屬的標籤特徵檔來代表使用者之標籤興趣。另外，本研究從標籤持續時間與標籤標記量分析標籤時間效應以區別使用者興趣喜好之程度。實驗結果顯示標籤有助於個人化文件推薦，只考慮使用者標籤興趣對於推薦效果之提升有限，還必須考量標籤時間效應才足夠。

關鍵字：標籤標記、標籤時間效應、個人化文件推薦、協同過濾

Personalized Document Recommendation Based on Tags

Student: Tzu-Wei Lin Advisor: Dr. Duen-Ren Liu

Institute of Information Management

National Chiao Tung University

Abstract

As information technology is maturing and Internet is getting more popular, web2.0 based websites are emerging to promote knowledge sharing. Tagging on web2.0 becomes a hot technique that users not only create tags to annotate and categorize content but also organize and manage information efficiently. Moreover, it launches related research on tag suggestions, query extension, and recommender considering tags. Tagging increases new meta-data through social intelligence. Most recommendation research on tags assumes that the same tag has the same meaning, which ignores the vagueness of the semantics of tags.

This study proposes personalized document recommendation based on tags. A user's tagging is regarded as the user's interest. The same term (tag) may denote different meanings to users when they apply the same term to tag documents. In this paper, a personalized tag profile representing user's interest is generated from the user's tagging of documents. Besides, we analyze the time effect of tags according to the duration of tag and the proportion of documents belonging to the tag. Novel recommendation methods incorporating tag profiles and the time effect of tags are proposed to improve the quality of documents recommendation. Experiment result shows that the proposed approach can effectively improve the quality of document recommendation.

Keywords: Tagging, Time Effect of Tag, Personalized Document Recommendation, Recommendation Based in Tag

誌謝

這是一種緣分，竹風吹拂下寫著這句話，接著，綿延更多的感謝。

當時，搖擺於三校的選擇，定與不定、決定之後又改變，心境上像流浪的小孩，真的不知道該往何方。最後，來到新竹完成兩年的碩士學業，也寫下這一頁，充滿說不盡的感謝。首先，感謝母親不離不棄撫養我長大，一同渡過艱難時刻，一起成長。同時，感謝指導教授劉敦仁博士的諄諄教誨，在論文上給予極大的幫助，才能在不斷之挫敗下順利完成論文。另外還要感謝邱登裕教授與羅濟群教授在論文口試時給了我很多寶貴的建議，讓我的論文研究內容更完整，在此致上深深的謝意。

在兩年中，非常感謝實驗室同學、學弟妹和學長姐的陪伴，感謝錦慧學姐和宇軒學長、純和學長在研究上的指導與鼓勵，感謝一起努力打拼的同窗好友邱璇、偉珍、佩芸、瓊瑤，有你們一起努力的感覺真好！也謝謝實驗室的學弟妹榮笙、其捷、雅婷和卉芳為實驗室帶來愉快的氣氛，豐富了我研究所生活的每一天。也感謝欣欣、淑惠、乃嘉在日常生活中的幫助。最後，感謝棒球隊的夥伴們陪我舒解壓力。要感謝的人太多了，謝謝一路上幫助我鼓勵我的朋友們！

2009.7 林子瑋

目錄

圖目錄

表目錄

第一章 緒論

1.1. 研究背景

近年以 web2.0 為概念發展的網站如雨後春筍般紛紛現世。過去瀏覽者扮演著被動的消費者角色，只能照單全收網站訊息；現今轉變成主動的內容提供者，與其他人分享資訊內容，使得網路空間累積資比以往更豐富完整的資訊。使用者可以在無名小站打造個人專屬部落格、上傳影音娛樂至 youtube[1]空間，或者在 plurk[2]平台即時分享生活點滴與關心朋友近況。

在 web2.0 網站開發中，標籤標記(tagging)是目前最受歡迎的概念。標籤標記讓使用者建立大眾分類(folksonomy)架構，而且提供具彈性的機制以進行組織、管理、搜尋資訊等活動。大眾分類是由使用者的標籤資料產生較鬆散的分類樹狀結構。從架構中大眾觀點瞭解某個資訊內容所包含涵義。另外，只要幾個辭彙便可分類資訊，使用者不用遵循系統定義好的分類架構。不過，不同使用者會因生活型態、文化、習慣而導致對於文字的解讀天差地遠，進而造成語意上差異，常見現象可分成一詞多義(polysemy)、多詞同義(synonymy)、或具有特別意義的詞彙。

網路科技進步與 web2.0 網站盛行使得資訊隨手可得，相對地亦讓個人面臨資訊過載問題，有效地尋找符合需求的資訊便成為重要的議題。推薦技術因應而生，其中協同過濾方法(Collaborative filtering)與內容式過慮法(Content-based filtering)是最常被使用的方法，前者概念是找出目標使用者(target user)的相似喜好者，再予預測目標使用者對推薦項目的喜好分數，來達成推薦成效；後者則是記錄所點閱項目，形成興趣特徵檔(profile)，接著尋找與特徵檔相似之文件，進而達到推薦之目的。使用者喜愛共同項目

[1] http://www.youtube.con

[2] http://www.plurk.com

的原因不見得相同，上述兩者皆忽略這個情境因素。標籤標記增加具個人化的項目詮釋資料(metadata)，讓研究者能夠從這個角度進行推薦之探討。

1.2. 研究動機

標籤標記技術快速成為網路平台的新服務，除了原本讓使用者能夠標記資訊功能外，還衍生標籤推薦與查詢語擴展等其他應用，也造成學術界逐漸投入大量精神於標籤相關領域研究。

學界研究標籤成果越來越多，應用方式也各有不同。Bischoff 學者深入標籤結構之探討，利用多個標籤資訊有助於搜尋所需之項目以及挖掘與其相關標籤[8]；Vig 藉由使用者標籤與被標記物的關係與標籤慣用偏好來解釋推薦之原因[22]。早期標籤探討多著重於標籤建議(tag suggestion)，意即協助使用者選擇標籤對物件作標記[20]。在語意方面，Rattenbury 自動化從 Flickr 的標籤擷取事件與地點之語意結構**錯誤! 找不到參照來源。**。另外，應用標籤資訊於搜尋研究，Kuo 運用標籤雲(tag cloud)將搜尋結果產生摘要[10]；Wang 與 Davison 利用 SVM 建立分類器，再進行查詢語擴展[24]。

以屬性為基礎或者使用者評分資訊為基礎的推薦研究非常常見，相對地整合標籤於推薦系統研究篇數並不多。在傳統推薦方法上，項目的屬性值對於每個使用者都是一樣的，然而標籤資訊會因為不同使用者具有不同意義。Nakamoto 由標籤角度來尋找目標使用者的相似鄰居使用者來改善協同過濾推薦方法[14]。Shepitsen 先對全部標籤分群，接著考慮使用者標籤與各群關係以及項目與各群間關係，以推薦使用者合適的項目[19]。上述皆假設相同標籤具有相同意義，未考量不同使用者對於相同標籤具有不同認知。本研究將使用者標籤記錄整合於協同過濾與內容式過濾推薦，並加以探討，根據使用者各種不同標籤來達成推薦的成果。

1.3. 研究目的

本研究將標籤資訊分別結合協同過濾與內容式過濾的推薦方法，並且分析標籤時間效應，建立適用於個人化文件推薦機制。citeulike[3]網站提供學術研究者建置個人虛擬圖書館以便收藏個人興趣之文章，同時進行分享收藏結果，並以標籤標記所收錄文章，本研究採用該網站資料當作實驗資料。

每個人在文字標籤表達上，都會含有各式各樣模糊與歧義，即使相同文字也可能有些許不同。因此本研究中有以下兩個假設：

假設 1. 一個使用者標籤特徵檔都代表一個使用者興趣特徵檔。

假設 2. 在同一個使用者下，相同標籤皆是具有固定涵義；但是，相同標籤在不同使用者間具有不同涵義。

在上述兩個假設下，建立使用者專屬的標籤特徵檔，以及進行標籤長期短期時間效應分析。從協同過濾推薦方法與內容過濾推薦方法角度，設計以標籤資訊觀點出發的推薦技術。本研究將標籤標記資訊剖析成標籤特徵檔與標籤時間效應分析，期望藉由這兩項新資訊可以有效提升個人化文件推薦品質。

1.4. 論文組織架構

以下說明本論文的研究動機與背景，相關文獻的探討，說明以標籤為基礎的個人化文件推薦的方法以及實驗結果，並且依據研究結果進行探討，提出未來可行的研究方向。本論文共分成五章，各章內容簡述如下。

第一章為緒論。說明本研究背景與動機、研究目的，與本論文架構。

第二章為文獻探討。主要包含資訊檢索、協同過濾推薦法、內容式過慮推薦法。

[3] http://www.citeulike.org

第三章為標籤為基礎的個人化文件推薦方法。本研究從標籤標記資料分析標籤特徵檔以及標籤標記時間記錄檔兩種新資訊，期盼藉由新資訊的結合獲得更好的推薦效果。本章詳述設計的概念與設計模型。

第四章實驗結果與分析。針對本研究提出的標籤為基準推薦方法進行實驗，整理實驗結果，並對數據進行分析討論。

第五章結論與未來研究方向。本章整理研究的主要發現與結論，並對可能的改進方向與未來可行的研究議題提出建議。

第二章 文獻探討

在本章節中，主要介紹本研究相關的文獻，其中包括資訊過檢索、協同過濾推薦法、內容式過濾推薦法。

2.1. 大眾分類法

大眾分類法(folksonomy)名稱為 Thomas Vander Wal 創造[21]，指群眾自行對照片、文件、音樂等媒體內容定義關鍵字。許多網站採用標籤標記服務，如：分享書籤網站「del.icio.us[4]」、相片分享網站「Flickr[5]」、音樂評論網站「GenieLab[6]」。傳統分類法(taxonomy)由專家或是系統開發者事先對資訊內容進行分類而建立階層式分類架構；而大眾分類法則是使用者為本身需求以自己熟悉的語言標記內容，形成扁平無階層關係分類結構。

大眾分類法滿足用戶視覺偏好以及讀圖心理，較搜尋引擎的輸入框更直接。定義的標籤來自於使用者共同建立，容易取得使用者認同感。不過，由於個人可以使用的詞彙未受控制，加上不同人以各自不同方式標記，因而產生語意含糊[16][23]。另外，群眾對同樣事物存在著各種不同的見解，但隨者定義標籤數量的增加，最後能夠產生數個趨近一致、並且收斂的共識標籤。大眾分類法不盡然是個完美分類方案，表 2-1 為其優缺點。

表 2-1 大眾分類法之優缺點

優點	缺點
● 具回饋性，創造溝通與分享空間。 ● 直接反應使用者需求。 ● 可以包含少數人的分類觀點。	● 個人可以任意的詞彙，又因不同人價值觀與認知不一樣，易造成語意模糊。 ● 過多詞彙組成一個標籤或者是一句話，如：美食義大利麵捷運站附近。

[4] del.icio.us http://delicious.com/

[5] flickr http://www.flickr.com/

[6] genieLabhttp://genielab.com/

● 以標籤來搜尋，可以挖掘使用者意想不到的資訊內容而增加驚喜感。	● 同義字的困擾，蘋果、麥金塔、麥金塔電腦都是指蘋果麥金塔電腦。

2.2. 資訊檢索

資訊過濾是可用來解決資訊超載問題的有效工具，其運作原理主要是藉由分析使用者行為來獲取其偏好或興趣，進而過濾或篩選出使用者所需的資訊。其概念為根據對使用者特徵檔(user profile)的長期學習模式，自動定義出使用者的資訊需求，找出符合使用者需求的資訊文件[6]。資訊過濾除了應用傳統資訊檢索的技術外，主要著重於使用者特徵檔學習的技術和演算法，以提高資訊需求的正確性及有效性。資訊擷取必須被動地等待使用者下達自行定義的查詢語(query)才能進行後續的分析運算，而資訊過濾則是經過長時間學習使用者特徵檔，以及根據使用者對文件的評分，主動協助使用者找出他有興趣或有需求的相關資訊，進而達到資訊過濾和擷取的效能，這種方法目前已廣泛地應用在知識管理和推薦的系統中。

資訊檢索在文件的應用上非常的廣泛，最終目的是根據使用者的查詢語，找到符合使用者的文件，滿足使用者的資訊需求。目前主要的技術有三種，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)。布林模式是利用三個邏輯運算子 AND、OR、NOT 來比對文件內容，將文件視為一群索引詞，以 1 和 0 代表相似度，1 為符合，0 為不符合，也就是說，文件中的索引詞要和使用者所下查詢完全符合才會成為檢索結果。該模式之運作方式上較為簡單、檢索速度快、可以用不同欄位資料來限定檢索範圍，對主題明確的檢索（如明確的作者名稱、標題名稱）非常有效，然一般使用者比較難以利用此種模式表達較為複雜的查詢。機率檢索模式是計算文件中出現的關鍵字屬於某類別的機率，若關鍵字未出現在測試文件中則代表不相關，反之則代表相關，再計算相關性，最後分類結果是屬於機率最大的類別。常見分類器有 Naïve Bayes 分類器、決策樹分類器、KNN 分類器、TFIDF 分類器等。向量空間模式是將文件內所有詞彙轉換至空間向量，計算關鍵字彼此之間的相似

程度，下一節將針對向量空間模式說明。其中向量空間模式是資訊檢索中較被廣為應用的方法[18]，將文件和查詢語(query)以多維度的向量形式表示，例如，以二維度的方式表達文件，其中包含關鍵字及權重。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排序的方式或門檻值設定的方式，回饋給資訊需求者。

將一篇文章擷取出重要的關鍵字以表示該文章的內容和屬性，在資訊檢索的領域中 TF-IDF (Term Frequency/Inverse Document Frequency)應用最為廣泛[18]。TF-IDF 概念為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，即中的 TF；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，即中的 IDF。TF-IDF 的計算方式如下：

$$W_{i,j} = TF_{i,j} \times IDF_i = TF_{i,j} \times \log \frac{N}{n_i} \qquad \text{(2.1 式)}$$

2.3. 協同過濾推薦方法

最早利用協同過濾推薦方法的是由 Goldberg 等學者提出的 Tapestry[5]，目的是過濾電子郵件，透過使用者定義的查詢語，挖掘出符合使用者興趣的電子郵件。爾後，協同過濾推薦方法相關的系統大量地被提出，其中最有名的是 GroupLens[9]，主要針對使用者感興趣的新聞進行推薦，有別於 Tapestry 中藉由使用者查詢語定義的過濾方式，GroupLens 則透過相似鄰居的計算方法，主動地找到具共同興趣的鄰居，以進行新聞的推薦。另外，Siteseer[17]利用相鄰使用者的書籤(bookmark)進行推薦，Knowledge Pump **錯誤! 找不到參照來源。**對使用者感興趣之文章進行推薦，Ringo[18]對音樂進行推薦，其他應用包括電子商務、學術論文等。

協同過濾主要是利用群體觀點來產生推薦項目給特定的個人使用者，故強調的是一種人與人之間的合作，利用過去的歷史記錄，計算各使用者間偏好行為的相似度，找出

喜好相近的鄰居者(neighbors) 並透過這些鄰居者所組成的群組之意見或建議，來產生目標使用者尚未接觸過卻可能有興趣的資訊推薦給目標使用者[5][9]。

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使用者之間的興趣相似程度，以尋找相似鄰居，可採用餘弦函式(2.2 式)皮爾森係數(2.3 式)，以分析出相似鄰居群。

$$sim(u,u') = \cos(\vec{u},\vec{u'}) = \frac{\vec{u} \times \vec{u'}}{\| \vec{u} \|_2 \times \| \vec{u'} \|_2} = \frac{\sum_{o=1}^{K} r_{u,o} r_{u',o}}{\sqrt{\sum_{o=1}^{K} r_{u,o}^{\ 2}} \sqrt{\sum_{o=1}^{K} r_{u',o}^{\ 2}}} \qquad (2.2 式)$$

$$sim(u,u') = \frac{\sum_{o \in O_{uu'}} (r_{u,o} - \overline{r_u})(r_{u',o} - \overline{r_{u'}})}{\sqrt{\sum_{o \in O_{uu'}} (r_{u,o} - \overline{r_u})^2} \sqrt{\sum_{o \in O_{uu'}} (r_{u',o} - \overline{r_{u'}})^2}} \qquad (2.3 式)$$

傳統協同過濾推薦是根據使用者相似度，利用 KNN 以及門檻值為基礎的方法找尋相似鄰居群，KNN 是找尋最相似之 K 個鄰居，門檻值為基礎的方法則是設定一相似度門檻值，超過門檻值之相似使用者則可列為鄰居。第二，預測使用者分數，以相似鄰居為基礎，透過對物品的評分分數及相似度分數，利用協同過濾推薦，預測使用者對該物品之喜好分數[18]。最後計算出來之預測分數，分數愈高代表愈符合使用者興趣，選擇分數高的項目即可推薦給使用者。以下為常見的預測分數之計算。

$$r_{u,o} = \frac{1}{N} \sum_{u' \in \bar{U}} r_{u',o} \qquad (2.4 式)$$

$$r_{u,o} = k \sum_{u' \in \bar{U}} sim(u,u') \times r_{u',o} \qquad (2.5 式)$$

$$r_{u,o} = \overline{r_c} + \frac{\sum_{u' \in \bar{U}} sim(u,u') \times (r_{u',o} - \overline{r_{c'}})}{\sum_{u' \in \bar{U}} sim(u,u')} \qquad (2.6 式)$$

2.4. 內容式過濾推薦法

內容式過濾推薦方法的觀點是個人在面對選擇之時，往往會選擇和印象中接近或是相似的物件，而這些物件會包含個人喜好的特徵。內容式過濾推薦的基礎是先對物品內容的分析而形成物件特徵檔(content profile) (2.7 式)，再依據使用者歷史行為記錄建立個人特徵檔(user profile) (2.8 式)；當進行推薦時，比對物件與個人特徵檔相似度(2.9 式)，推薦相似度較高的物件給予使用者。在計算文字為主的物件時，會採用餘弦定理(2.10 式)。

$$ContentProfile(o_j) = (w_{1j}, ...w_{ij}) \quad (2.7\text{ 式})$$

$$ContentBaedUserProfile(u) = (w_1, ...w_i) \quad (2.8\text{ 式})$$

$$sim(u,o) = \cos(\overrightarrow{w_u}, \overrightarrow{w_o}) = \frac{\overrightarrow{w_u} \times \overrightarrow{w_o}}{\| \overrightarrow{w_u} \|_2 \times \| \overrightarrow{w_o} \|_2} = \frac{\sum_{i=1}^{K} w_{i,u} w_{i,o}}{\sqrt{\sum_{i=1}^{K} w_{i,u}^{\ 2}} \sqrt{\sum_{i=1}^{K} w_{i,o}^{\ 2}}} \quad (2.9\text{ 式})$$

藉著分析顧客所喜好的產品之特徵來建立顧客的個人特徵檔，然後根據個人特徵檔來推薦產品[25]。其依據的基礎是對物品內容的分析，而不是人的評價。更簡單地來說，內容式過濾的方法就是根據顧客過去喜好的產品，推薦類似的產品給予顧客。內容導向式過濾的方法比較適合應用於文件的推薦，也已經被應用於網頁[1][15]、書籍[13]以及新聞文章[7][11]的推薦。

個人特徵檔來自於歷史記錄，一個新的使用者缺乏使用記錄，故無法進行推薦活動；另外，內容式過濾受限於物件內容分析，使得物件內容本身需要擁有足夠的資訊才能有效地進行分析。

2.5. 整合標籤於推薦系統

不少網站已經採用標籤標記服務，也衍生很多其他相關應用，將標籤整合於推薦系統是其中一例。標籤最常見用於改善使用者相似度計算[14][15]。另外也有將標籤與物件內容進行分群以進行推薦[19]。Shiwan 等學者藉由 WordNet 建立標籤語意以計算標籤語意相似度來改善尋找相似鄰居[26]。Nakamoto 等學者提出演算法的雛形未進行實驗，利用標籤改善尋找相似鄰居；同時，求出鄰居之所有標記物件與推薦候選物件相似度之最大值作為鄰居對此推薦候選物件之評分。爾後，該名學者又進行改良，以 ded.icio.us 網站為實驗資料，先蒐集部分資料當作訓練資料，利用 EM 演算法切割出 100 主題領域，計算每個網站在各個主題領域的比重，以及每名使用者使用的標籤在各個主題領域的比重，以改善推薦品質[14]。

Andriy 等學者以標籤結合分群概念進行推薦。其概念如圖 2-1，將單一使用者的標籤進行分群，然後對全部使用者所產生的標籤進行分群；接著藉由全部使用者所產生的標籤分群把推薦項目與使用者連起來[19]。上述研究都假設標籤為相同意義，沒有區分相同標籤在使用者間會具有不同意義。

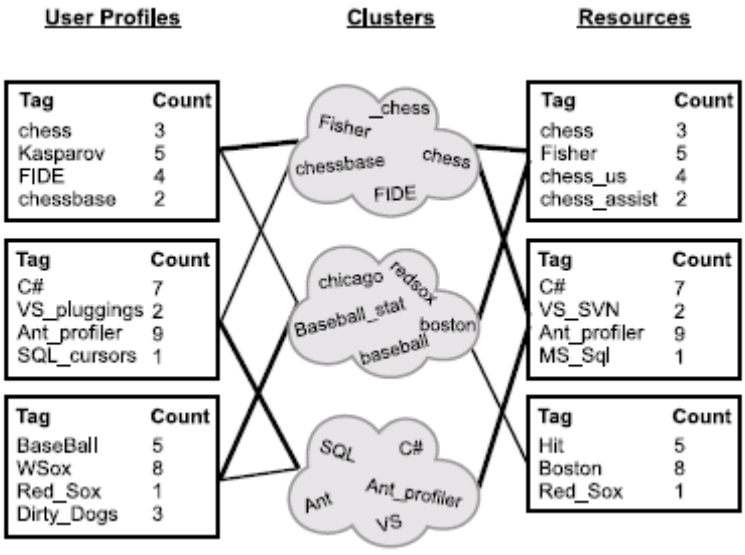


圖 2-1 應用標籤於推薦之分群技術概念

第三章 以標籤為基礎之個人化文件推薦方法

本章主要介紹本研究提出以標籤為基礎之個人化文件推薦方法，第一節概要說明標籤為基礎之推薦流程，第二節說明如何進行文件前置處理，第三節述說如何建立特徵檔，第四節說明標籤時間效應分析法則，與以標籤為基礎之個人化文件推薦設計。

3.1. 標籤為基礎之個人化文件推薦流程

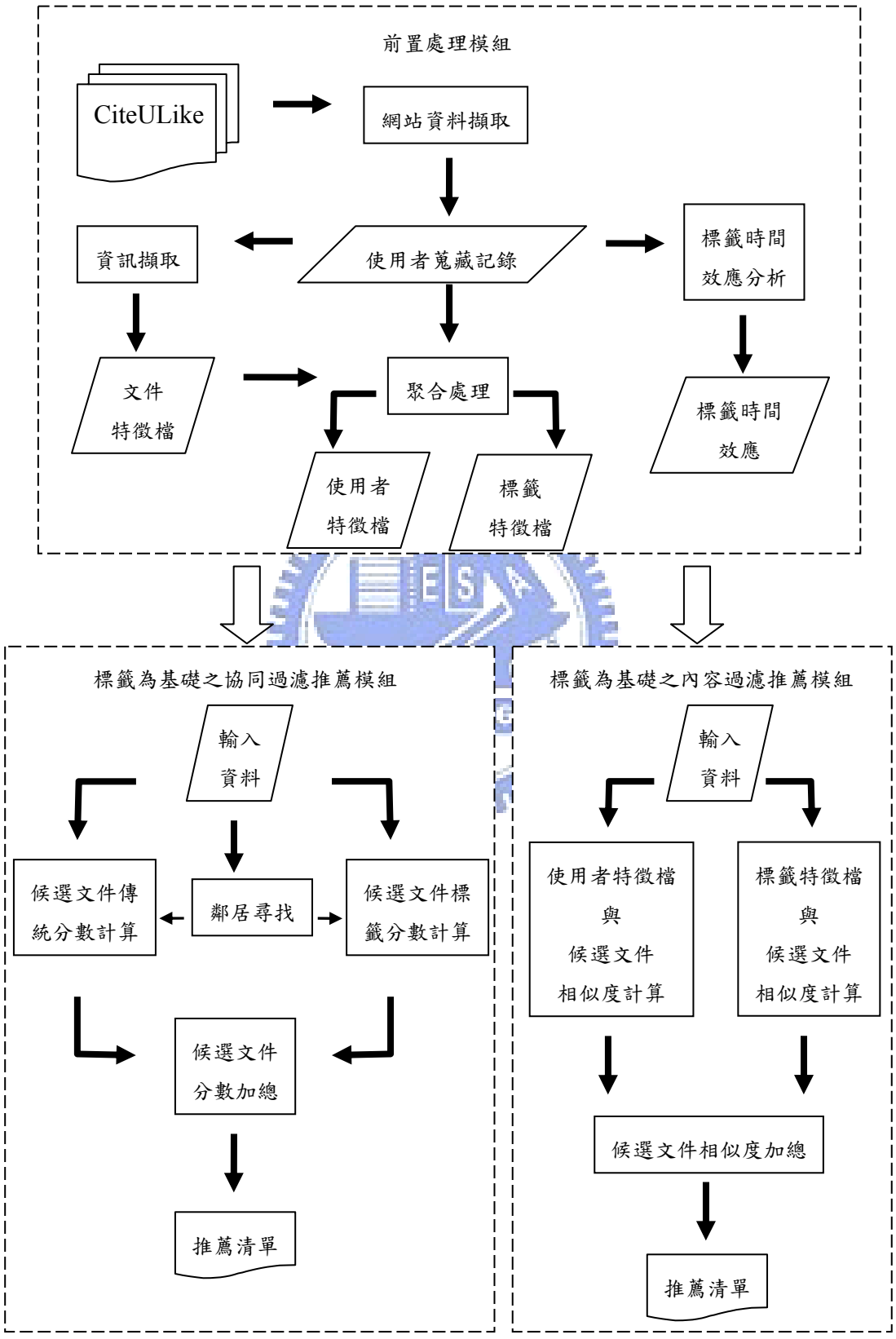


圖 3-1 以標籤為基礎之個人化推薦流程圖

論文研究架構分成兩大部分，前置處理與標籤為基礎之個人化文件推薦。前者進行將資料轉換成程式所需格式;後者又可分成標籤為基礎之協同過濾推薦與標籤為基礎之內容過濾推薦，將於本章第四節逐一說明。本研究資料擷取自於citeulike網站，該網站提供建立個人虛擬圖書館服務，即使用者可以收藏自己感到興趣之學術文章，並且准許其以自行定義標籤標記所收錄文章。

每個人在文字標籤表達上，都會含有各式各樣模糊與歧義，即使相同文字也可能有些許不同。因此本研究有兩個假設，一個使用者標籤特徵檔都代表一個使用者興趣特徵檔；在同一個使用者下，相同標籤皆是具有固定涵義；但是，相同標籤在不同使用者間具有不同涵義。研究中每個人將有個人專屬的標籤特徵檔，即便標籤文字一樣。

在圖3-1中清楚得知整個研究流程，首先自citeulike網站擷取資料，剖析成使用者收藏文件記錄、標籤標記文件記錄與文件內容。接著，由文件內容經過資訊擷取處理建立文件特徵檔，再組成使用者特徵檔與標籤特徵檔，並且從標籤標記記錄分析標籤時間效應。最後，以改良的標籤為基礎之協同過濾推薦與內容過濾式推薦產生推薦清單。

3.2. 文件前置處理

學術文章是由文字組成的，須透過文件前置處理的步驟，將文件統一轉換成由關鍵字與權重組成的文件特徵檔，才能進行後續的研究。本研究採取文件中標題與摘要進行前置處理的步驟，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，以提高關鍵字擷取的正確性，降低矩陣的雜訊。

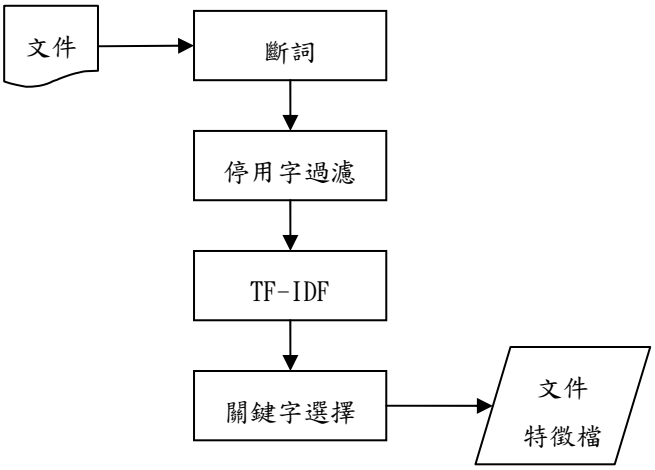


圖 3-2 文件特徵檔建置流程圖

(一)斷詞

詞是最小有意義且可以自由使用的語言單位。任何語言處理的系統都必須先能分辨文本中的詞才能進行進一步的處理，例如機器翻譯、語言分析、語言解讀、資訊抽取。

主要目的是簡化文章中的字詞，有些字詞代表相同的意義，但是因為詞性不同，而被視為不同的單字內容，例如，「cluster」、「clustering」、「clusters」為三個不同的單字，但是其本意皆為分群。在本研究中利用Porter Stemming的方法，將詞性不同但詞義相同的字詞轉換成同一單字，以簡化文章中的字詞數量，並且提高關鍵字詞的重要性。

(二)停用詞過濾

文件撰寫者編輯知識文件，經常使用某些字詞，但這些字詞不足以代表文章的關鍵字，例如連接詞、介詞、副詞等。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[3]，將文件中出現於停用字列表的字詞移除。

(三)TF-IDF

TF-IDF 主要是計算字詞在文件中權重分數，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，出現頻率愈高，則該字詞的可代表性就愈低。在本研究中採用的 TF-IDF 計算如下：

$$W_{i,j} = TF_{i,j} \times IDF_i = \frac{freq_{i,j}}{\sqrt{\sum_{i=1}^{|Dj|} (freq_{i,j})^2}} \times \log \frac{N}{n_i} \qquad (3.1\ 式)$$

其中，中的符號代表含意說明如下：

$W_{i,j}$ ： 關鍵字權重，字詞 i 在文章 j 中的權重。

$TF_{i,j}$ ： 字詞頻率，字詞 i 在文章 j 中出現的頻率。

IDF_i ： 反文件頻率，字詞 i 出現在其他文件集中的頻率。

$freq_{i,j}$ ： 字詞次數，字詞 i 在文章 j 中出現的次數。

N： 全部文章數。

$|D_j|$： 關鍵字數，文章 j 的向量大小。

n_i ： 字詞 i 出現在其他文件集的次數

(四)關鍵字選擇

根據TF-IDF計算後的字詞權重分數，利用Top N的方式挑選代表一份文件的關鍵字。表3.1所示：

表 3-1 文件特徵檔例子

Term	Weight
custom	3.85260745
market	1.379382136
segment	0.466892051
servic	0.377730741
valu	0.259009804

3.3. 特徵檔

本節將依序介紹文件特徵檔、使用者特徵檔、標籤特徵檔。文件特徵檔經過上節文件前處理後，會得到關鍵字與其權重建構而成的特徵檔，表達如下。

$$DocProfile_d = (w_{1d}, ...w_{id})$$

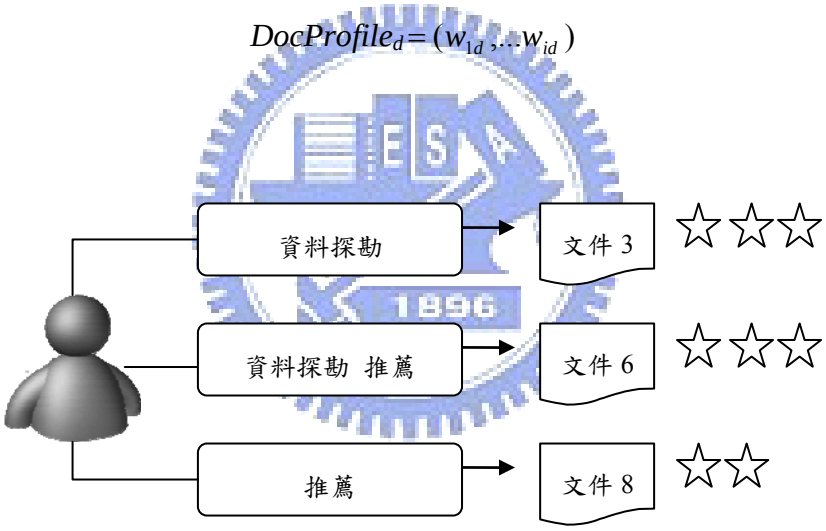


圖 3-3 標籤標記文件圖

從圖 3-3 中得知使用者收藏文件 3、文件 6、文件 8，使用者特徵檔便是由所收錄文件建構而成，意即文件 3、文件 6、文件 8 之文件特徵檔以使用者文章評分為權重以加權平均方式而組成使用者特徵檔，如 3-2 式。

$$UserProfile^u = (w_{1u}, ...w_{iu})$$

$$W_{i,u} = \frac{\sum_{d \in Doc^u} Score_d^u \times W_{i,d}}{\sum_{d \in Doc^u} Score_d^u} \quad (3.2\text{ 式})$$

$W_{i,u}$ ： 關鍵字權重，字詞 i 在使用者 u 中的權重。

$Score_d^u$ ： 文件分數，使用者 u 給該收錄文章 d 的分數。

$W_{i,d}$ ： 關鍵字權重，字詞 i 出現在文章 d 中的權重。

Doc^u ： 文件集合，使用者 u 所收錄的文件集合。

圖 3-3 顯示使用者以「資料探勘」標記文件 3 與文件 6，「推薦」標記文件 6 與文件 8，一共使用過兩個標籤，故該名使用者會有兩個專屬標籤特徵檔。「資料探勘」標籤特徵檔與「推薦」標籤特徵檔分別由文件 3 與文件 6、文件 6 與文件 8 的文件特徵檔，以使用者文章評分透過加權平均方式組成標籤特徵檔，如 3-3 式。

$$TagProfile_t^u = (w_{1t}, \ldots w_{it})$$

$$W_{i,t}^u = \frac{\sum_{d \in Doc_t^u} Score_d^{ut} \times W_{i,d}}{\sum_{d \in Doc_t^u} Score_d^{ut}} \quad (3.3\text{ 式})$$

$W_{i,t}^u$ ： 關鍵字權重，字詞 i 在標籤 t 中的權重。

$Score_d^{ut}$ ： 文件分數，使用者 u 給被標籤 t 標記的文章 d 的分數。

$W_{i,d}$ ： 關鍵字權重，字詞 i 出現在文章 d 中的權重。

Doc_t^u ： 文件集合，使用者 u 以標籤 t 標記該文件集合。

3.4. 以標籤為基礎之個人化文件推薦方法

本節介紹標籤時間效應分析法則與以標籤為基礎之推薦設計，後者分別將標籤資訊整合於傳統協同過濾推薦與內容式過濾法以加強推薦之成效。

3.4.1. 標籤時間效應分析

本研究中使用者為進行學術研究而收藏文件，不是漫無目的地閱讀文件，使用者的標籤量不會因收錄文件數量而呈線性或指數成長。本研究將標籤視為一個使用者興趣，認為標記於文章的標籤是收藏這份文件的動機之一。使用者可能大部份都使用某個標籤，或者某段特定期間都使用另一個標籤，因此，分析標籤時間效應以加強瞭解使用者興趣。

其分析概念為標籤出現的時間越長其重要性越高，標籤標記文件數量越多其重要性也越高，前者為興趣持續程度之考量，後者為興趣強度之考慮，圖 3-4 呈現該概念。計算方式為標籤出現時間長度占整個時間長度的比例以及標籤標記文件數量占全部收藏文件數量比例。本研究從長期時間與短期時間進行效應分析，分別是 3.4 式與 3.5 式。

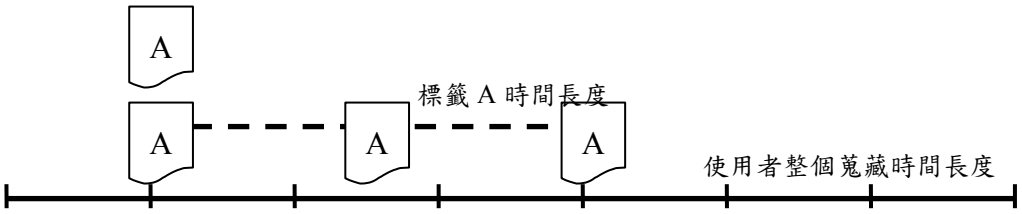


圖 3-4 標籤標記文件時間概念圖

$$LongTerm_{ta}^{Ua} = \frac{Duration^{ta}}{Duration^{Ua}} \times \frac{\#DOC^{ta}}{\#DOC^{Ua}} \qquad (3.4\ 式)$$

$Duration^{ta}$ ： 標籤時間長度，標籤 *ta* 最後標記的文章收錄時間減去標籤 *ta* 第一次標記的文章收錄時間。

$Duration^{Ua}$ ： 使用者時間長度，使使用者 *Ua* 最後標記的文章收錄時間減去使用者 *Ua* 第一篇標記的文章收錄時間。

$\#DOC^{ta}$ ： 文件數量，標籤 *ta* 標記的文章數目。

$\#DOC^{Ua}$ ： 文件數量，使用者 *Ua* 的收藏文章數目。

$$ShortTerm_{ta}^{Ua} = \frac{ShortTermDuration^{ta}}{ShortTermDuration^{Ua}} \times \frac{\#Doc^{ta}inShortTermPeriod}{\#Doc^{Ua}inShortTermPeriod} \quad (3.5\ 式)$$

$ShortTermDuration^{ta}$ ： 標籤短期時間長度。

$ShortTermDuration^{Ua}$ ： 使用者短期時間長度，在研究中為使用者時間長度之 1/3。

$\#Doc^{ta}inShortTermPeriod$ ： 文件數量，短期時間內標籤標記的文章數目。

$\#Doc^{Ua}inShortTermPeriod$ ： 文件數量，短期時間內使用者的收藏文章數目。

由於短期內出現新的標籤，與持續出現的標籤具有不同意義，持續出現的標籤比較相對新的標籤更為重要。圖 3-5 為短期時間標籤標記標籤，A 與標籤 B 同樣標記 2 篇文件，但是標籤 A 過去就出現過了，表示在短期內使用者仍就對這興趣有興趣，在相同標記量下，相對來說標籤 A 重要性大過於標籤 B。

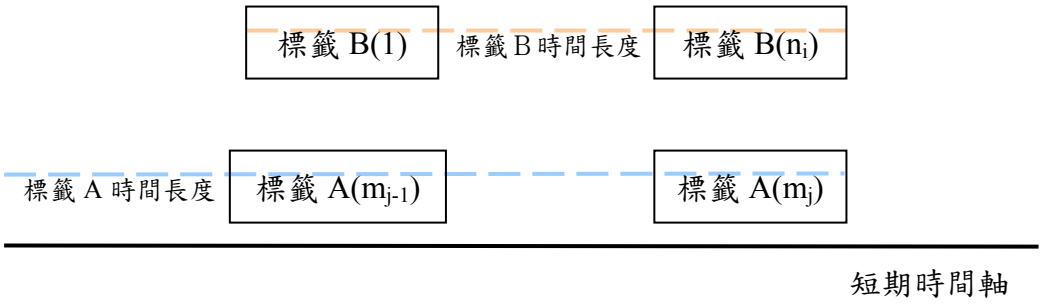


圖 3-5 短期時間標記記錄

$$ShortTermDuartion^{ta} = \begin{cases} Document_{ta}^{1} \in ShortTerm_{d}^{Ua}, Time_{ta}^{d,n} - Time_{ta}^{d,1} \\ Document_{ta}^{1} \notin ShortTerm_{d}^{Ua}, ShortTermDuration^{Ua} - (Time_{Ua}^{d,n} - Time_{ta}^{d,n}) \end{cases}$$

(3.6 式)

$Document_{ta}^{1}$ ： 文件，標籤 *ta* 標記的第一篇文件。

$ShortTerm_{d}^{Ua}$ ： 文件集合，使用者 *Ua* 短期文件集合。

$Time_{ta}^{d,n}$ ： 文件收藏時間，標籤 *ta* 標記最後一份(*n*)文件 *d* 的時間。

$Time_{ta}^{d,1}$ ： 文件收藏時間，標籤 *ta* 標記第一份文件 *d* 的時間。

$Time_{Ua}^{d,n}$ ： 文件收藏時間，使用者 *Ua* 標記最後一份(*n*)文件 d 的時間。

以圖 3-5 為例詳細說明標籤時間效應分析。圖 3-5 為某位使用者的收藏記錄時間圖，在 9 天時間之內，分別以三種標籤(A、B、C)共標記了 13 篇學術文章。另外，在研究中短期時間為最近整個收藏時間的三分之一時期，在該例短期時間便為 3 天(7 天~9 天)。

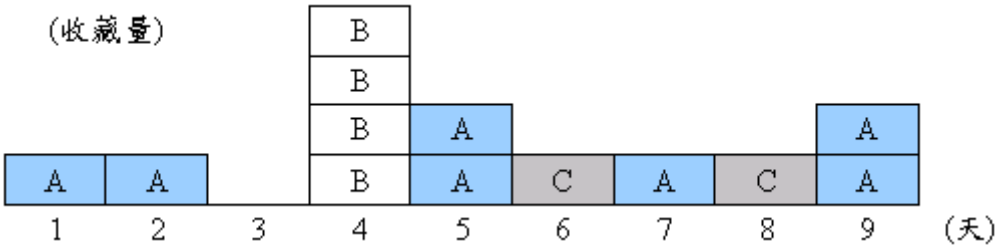


圖 3-6 使用者收藏文件記錄圖

標籤 A 標記長期時間長度為 9 天(1 天~9 天)，共有 7 篇文章；短期時間長度為 3 天(7 天~9 天)，共有 3 篇文章。標籤 B 標記長期時間長度為 1 天，共有 4 篇文章；在短期時間內沒有任何記錄。標籤 C 標記短期時間長度為 1 天，僅有一篇文章。標籤時間效應分析結果如表 3-2 所示。

表 3-2 標籤時間效應分析例子

標籤	長期強度	短期強度
標籤 A	$LongTerm_{ta}^{Ua} = \frac{9}{9} \times \frac{7}{13}$	$ShortTerm_{ta}^{Ua} = \frac{3}{3} \times \frac{3}{4}$
標籤 B	$LongTerm_{ta}^{Ua} = \frac{1}{9} \times \frac{4}{13}$	$ShortTerm_{ta}^{Ua} = \frac{0}{3} \times \frac{0}{4}$
標籤 C	$LongTerm_{ta}^{Ua} = \frac{3}{9} \times \frac{2}{13}$	$ShortTerm_{ta}^{Ua} = \frac{2}{3} \times \frac{1}{4}$

3.4.2. 以標籤為基礎之協同過濾推薦方法

本章節說明以標籤為基礎之協同過濾推薦方法如何進行。首先找出目標使用者的鄰居，$Nbr(Ua)$(Ua 為目標使用者)的使用者相似度計算方式有二種，餘弦函式與皮爾森係數，最後從鄰居標記該候選文件的標籤計算標籤預測分數。

先列出傳統協同過濾與標籤為基礎的線性組合，如 3.7 式。稍後再說明標預測分數($Ptagscore_d$)的細節。傳統方式為從鄰居評分記錄計算該候選文件的預測分數，為 3.7 式參數(1- α)結合部分。

$$Pscore_d^{Ua} = (1-\alpha)\left(\overline{Score^{U_a}} + \left(\frac{\sum_{Ui \in Nbr(Ua)} SIM(User\,Profile^{Ua}, User\,Profile^{Ui}) \times (Score_d^{Ui} - \overline{Score^{Ui}})}{\sum_{Ui \in Nbr(Ua)} \left| SIM(User\,Profile^{Ua}, User\,Profile^{Ui}) \right|}\right)\right) + \alpha \frac{\sum_{Ui \in Nbr(Ua)} Ptagscore_d^{Ua,Ui}}{\left| Nbr(Ua) \right|} \quad (3.7\text{ 式})$$

$Pscore_d^{Ua}$ ： 預測分數，目標使用者 Ua 的文章 d 之預測分數。

$\overline{Score^{U_a}}$ ： 平均分數，目標使用者 Ua 給文章的平均分數。

$\overline{Score^{Ui}}$ ： 平均分數，鄰居使用者 *Ui* 給文章的平均分數。

$Score_d^{Ui}$ ： 文章分數，鄰居使用者 *Ui* 給文章 *d* 的分數。

$Ptagscore_d^{Ua,Ui}$ ： 預測分數，從標籤角度建立起預測分數，見 3.8。

$Nbr(Ua)$ ： 尋找鄰居，尋找目標使用者 *Ua* 的鄰居。

標籤預測分數概念為從鄰居標記候選文件的標籤求出對應的目標使用者標籤，再從對應的標籤分數聚合最後結果。標籤預測分數如 3.9 式；找出對應標籤如 3.11 式；每位使用者標籤分數計算如 3.8 式。

$$Score_t^{Ua} = \frac{\sum_{d \in D_t^{Ua}} Score_d^{Ua}}{\left| D_t^{Ua} \right|} \quad (3.8\text{ 式})$$

$Score_{tai}^{Ua}$ ： 使用者 *Ua* 對標籤 t 的分數。

$Score_d^{Ua}$ ： 使用者 *Ua* 對文件 d 的分數。

$\left| D_t^U \right|$ ： 使用者 *Ua* 用標籤 t 標記的文件數量。

D_t^U ： 使用者 *Ua* 用標籤 t 標記的文件集合

先從表 3-3 與圖 3-6 為例說明標籤預測分數流程，再列出算式。

表 3-3 多名使用者之文件評分

	Doc1	Doc2	Doc3	Doc4	Doc5
User1	1		3		2
User2			2	3	
User3	2	1		1	
User4		3			
User5		2			2

目標使用者為表 3-3 中 User1，倘若經過使用者相似度計算，得知 User3 與 User4 為 User1 的鄰居。Doc2 為推薦候選文件，求其標籤構面預測分數。圖 3-7 為 User1 使用過的全部標籤「推薦」、「隱私」、「文字探勘」，以及 User3 與 User4 標記 Doc2 的標記記錄，分別是「協同過濾」、「搜尋引擎」、「資料探勘」與「分群演算法」、「搜尋引擎」、「資訊擷取」。

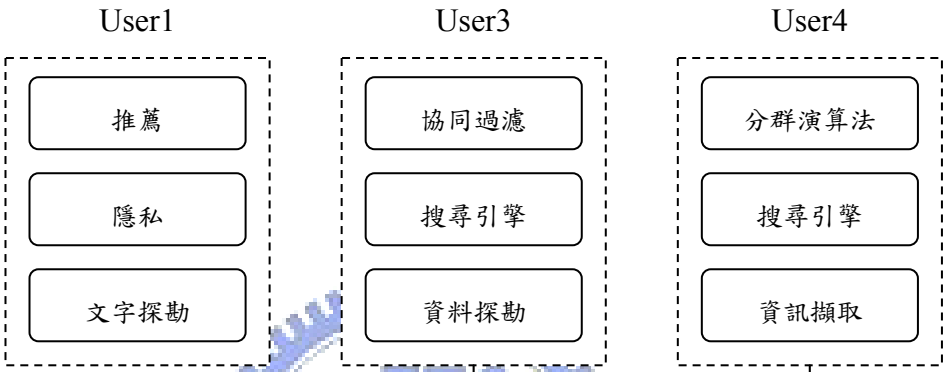


圖 3-7 鄰居標記候選文件概念圖

計算鄰居的標籤特徵檔與目標使用者的標籤特徵檔相似度，相似度最大即為對應的標籤。。

User3 標記 Doc2 的標籤「協同過濾」、「搜尋引擎」、「資料探勘」之對應標籤分別是「推薦」、「文字探勘」、「文字探勘」。User4 標記 Doc2 的標籤「分群演算法」、「搜尋引擎」、「資訊擷取」之對應標籤皆為「文字探勘」。

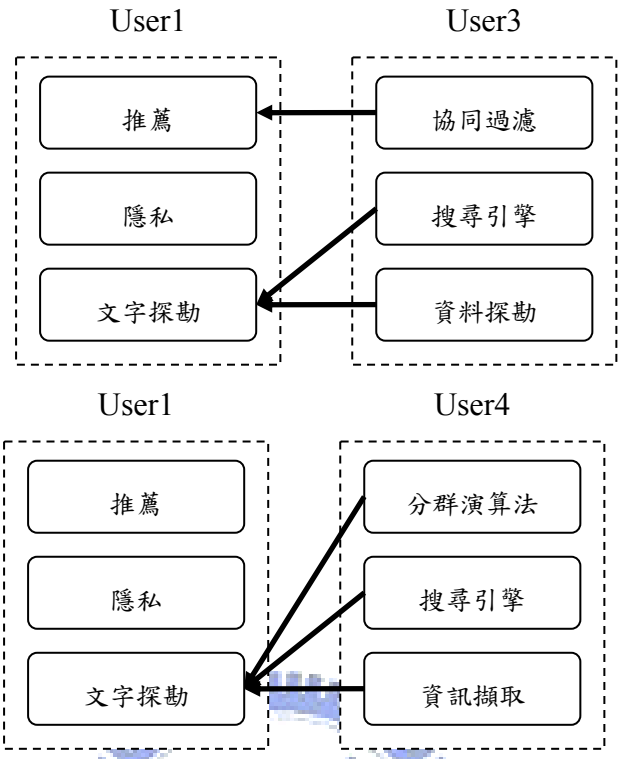


圖 3-8 鄰居標記候選文件之標籤對應目標使用者標籤概念圖

對「推薦」、「文字探勘」標籤分數以標籤相似度結合對應標籤時間效應加權平均方式計算出標籤預測分數，如 3.8 式。鄰居 User4 部分亦同，再將兩個標籤分數平均就是最後結果，標籤時間效應可以是標籤長期時間效應或者標籤短期時間效應。由鄰居使用者標記文章的標籤與目標使用者的標籤計算相似度得到對應標籤，如 3.11 式。利用皮爾森尋找鄰居時，以標籤標示記錄計算標籤間相似度，如 3-10 式。

$$Ptagscore_d^{Ua,Ui} = \frac{\sum_{t_i \in TagSet_d^{Ui};tai \in \max tagmapping(Ua,Tag_{ti}^{Ui})} Score_{tai}^{Ua} \times sim(Tag\ \Pr ofile_{tai}^{Ua}, Tag\ \Pr ofile_{ti}^{Ui}) \times TimeFactor_{tai}^{Ua}}{\sum_{t_i \in TagSet_d^{Ui};tai \in \max tagmapping(Ua,Tag_{ti}^{Ui})} sim(Tag\ \Pr ofile_{tai}^{Ua}, Tag\ \Pr ofile_{ti}^{Ui}) \times TimeFactor_{tai}^{Ua}} \quad (3.9\ 式)$$

$Tag\,\mathrm{Pr}ofile_{tai}^{Ua}$ ： 目標使用者 *Ua* 的特定標籤描繪檔 *tai*。

$Tag\,\mathrm{Pr}ofile_{ti}^{Ui}$ ： 鄰居使用者 *Ui* 的特定標籤描繪檔 *ti*。

tai ： 對應標籤，來自 *tagmapping()*，見 3.11 式。

TimeFactor ： 時間因素，依據長期與短期推薦而不同，分別為 $LongTerm_{ta}^{Ua}$ 與 $ShortTerm_{ta}^{Ua}$ 。

$Score_{tai}^{Ua}$ 標籤分數，目標使用者 *Ua* 的標籤 *tai* 的標籤分數，計算方式如 3.8。

$$sim(Tag_{tai}^{Ua}, Tag_{ti}^{Ui}) = \frac{\sum_{i=1}^{|D|} r_{tai,j} r_{ti,j}}{\sqrt{\sum_{j=1}^{|D|} r_{tai,j}^{\ 2}} \sqrt{\sum_{j=1}^{|D|} r_{ti,j}^{\ 2}}} \qquad (3.10\ 式)$$

Tag_{tai}^{Ua} ： 標籤標記文件記錄，使用者 *Ua* 使用標籤 *tai* 標記文件記錄。

Tag_{ti}^{Ui} ： 標籤標記文件記錄，使用者 *Ui* 使用標籤 *ti* 標記文件記錄。

$|D|$ ： 文件集合大小，使用者 *Ua* 文件集合與使用者 *Ui* 的文件集合聯集大小。

$r_{t,j}$ ： 標籤 *tai* 標記文件 *j* 與否；否為 0，反之為 1。

$r_{t',j}$ ： 標籤 *ti* 標記文件 *j* 與否；否為 0，反之為 1。

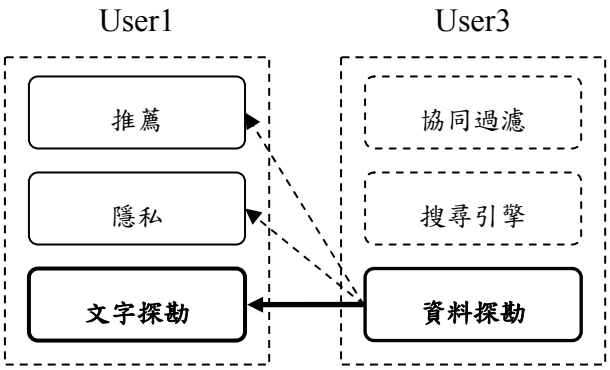


圖 3-9 標籤對應概念

圖 3-9 中 User3 有三個標記 Doc2 的標籤，其中以「資料探勘」標籤為例說明找出對應的標籤步驟。先將「資料探勘」標籤分別與 User1 全部標籤進行相似度計算，即與「推薦」、「隱私」、「文字探勘」進行餘弦函式計算，接著取相似度最大值的標籤作為對應的標籤，例如「資料探勘」標籤對應到文字探勘」標籤。研究中設定θ值為 0.1，假若缺乏對應標籤，則以 0 作為相似度。

$$tai = \max tagmapping\,(Tag_{ta}^{Ua}, Tag_{ti}^{Ui})$$

$$= \max_{taj \in TagSet^{Ua}} \left\{ sim(Tag\,\mathrm{Pr}ofile_{taj}^{Ua}, Tag\,\mathrm{Pr}ofile_{ti}^{Ui}) > \theta \right\} \qquad (3.11\ 式)$$

$Tag\,\mathrm{Pr}ofile_{tai}^{Ua}$：　目標使用者 *Ua* 的特定標籤描繪檔 *tai*。

$Tag\,\mathrm{Pr}ofile_{ti}^{Ui}$：　鄰居使用者 *Ui* 的特定標籤描繪檔 *ti*。

θ：　門檻值。

3.4.3. 以標籤為基礎之內容式過濾推薦方法

本章節說明以標籤為基礎之內容式過濾推薦方法如何進行，主要概念是計算標籤特徵檔與文件特徵檔的相似度。使用者有三個標籤分別與文件特徵檔計算餘弦相似度，再運用標籤時間效應作加權總合，其中標籤時間效應可以是標籤長期時間效應或者標籤短期時間效應，如不使用標籤時間效應，便採取平均方式。

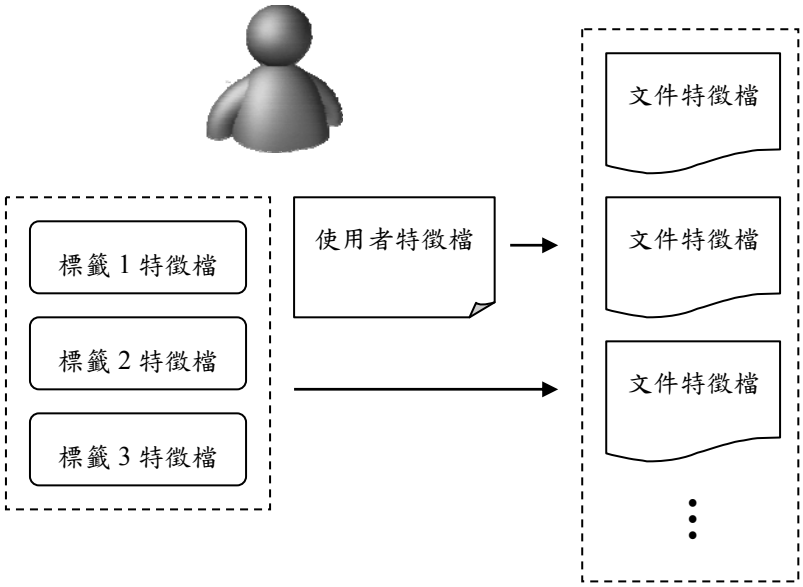


圖 3-10 以標籤為基礎之內容是推薦概念圖

傳統內容式過慮法是計算使用者特徵檔與文件特徵檔的相似度。3.12 式為兩者的線性組合。

$$Pscore_d^{Ua} = (1-\alpha)\, sim(User\,\mathrm{Pr}ofile^{Ua}, Doc\,\mathrm{Pr}ofile_d) + \alpha \frac{\sum_{ta \in TagSet^{Ua}} sim(Tag\,\mathrm{Pr}ofile_{ta}^{Ua}, Doc\,\mathrm{Pr}ofile_d) \times TimeFactor_{ta}^{Ua}}{\sum_{ta \in TagSet^{Ua}} TimeFactor_{ta}^{Ua}} \quad (3.12\text{ 式})$$

$User\,\mathrm{Pr}ofile^{Ua}$ ： 目標使用者 *Ua* 的使用者描繪檔。

$Doc\,\mathrm{Pr}ofile_d$ ： 文章 d 的文件描繪檔。

$Tag\,\mathrm{Pr}ofile_{ta}^{Ua}$ ： 目標使用者 *Ua* 的特定標籤描繪檔 *ta*。

3.13 式為混合標籤長期時間效應與標籤短期時間效應，其中 $\alpha+\beta+\gamma=1$。

$$Pscore_d^{Ua} = \alpha\ sim(User\,\mathrm{Pr}ofile^{Ua}, Doc\,\mathrm{Pr}ofile_d) + \beta \frac{\sum_{ta \in TagSet^{Ua}} sim(Tag\,\mathrm{Pr}ofile_{ta}^{ua}, Doc\,\mathrm{Pr}ofile_d) \times LongTerm_{ta}^{Ua}}{\sum_{ta \in TagSet^{Ua}} LongTerm_{ta}^{Ua}} + \gamma \frac{\sum_{ta \in TagSet^{Ua}} sim(Tag\,\mathrm{Pr}ofile_{ta}^{Ua}, Doc\,\mathrm{Pr}ofile_d) \times ShortTerm_{ta}^{Ua}}{\sum_{ta \in TagSet^{Ua}} ShortTerm_{ta}^{Ua}} \quad (3.13\ 式)$$

$User\,\mathrm{Pr}ofile^{Ua}$ ： 目標使用者 Ua 的使用者描繪檔。

$Doc\,\mathrm{Pr}ofile_d$ ： 文章 d 的文件描繪檔。

$Tag\,\mathrm{Pr}ofile_{ta}^{Ua}$ ： 目標使用者 Ua 的特定標籤描繪檔 ta。

第四章 實驗與評估

在本章節中利用實際的資料，以驗證本研究所提出以標籤為基礎之個人化文件推薦方法，能在視標籤為興趣之基礎下，確實提昇個人化文件推薦之成效。以下就實驗資料、實驗方法、實驗結果與評估詳加以說明。

4.1. 資料蒐集

實驗資料來自於 citeulike 網站，該網站提供建立個人虛擬圖書館服務，即收藏自己感到興趣之學術文章，並且准許以自行定義標籤標記所收錄文章，並以 1 至 5 分方式評分文章。實驗資料在 2 月 15 日將所有資料完成抓取，將資料彙整成使用者收藏資料表<user_id, document_id, rating>、標籤標記資料表<user_id, document_id, tag>、文件內容資料表<doc_id, title, abstarct>。透過人工篩選出使用者間交集數量大的資料，剩下 162 名使用者與 46700 篇學術文章，每名使用者平均使用 170 個標籤。

表 4-1 資料集

項目	文件	使用者
數量	46700	162

4.2. 實驗方法

研究中為模擬系統是否成功推薦，必須將資料分成訓練資料(training data)與測試資料(testing data)。使用訓練資料建立第三章所提到的特徵檔，經過推薦流程後得到推薦清單，檢驗前十篇推薦文章是否為測試資料中大於 3 分的文件以得知推薦成效。

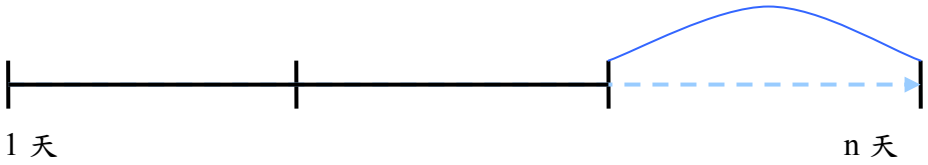


圖 4-1 測試資料切割概念圖

為求短期與長期比較的公平性，選擇最近的時間作測試。因此，取出使用者最後三分之一收藏資料的 60%來當作測試資料(等同取出 20%資料)，意即將資料分成三個時期，拿出最後時期的資料當作測試資料，如圖 4-1 淺藍色虛線；剩餘的資料便用來當作訓練資料。

4.3. 評估標準

本研究採用之評估標準分別為由Recall、Precision組成的F1-measure。以下分別介紹上述評估標準。

(一) 準確率

$$\Pr ecision = \frac{\#correctly\_recommended}{\#recommended} \qquad (4.1式)$$

Precision是指準確率，由系統推薦文章中符合使用者相關文章的比例，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，例如若系統推薦100篇文章，其中60篇為使用者評估為相關的，則Precision為0.6。

(二) 召回率

$$\mathrm{Re}call = \frac{\#correctly\_recommended}{\#relevant} \qquad (4.2式)$$

Recall是指召回率，計算使用者評分為具相關性的文章中，系統有正確推薦的比例。分子為系統推薦且符合使用者需求的文章數，分母為使用者評為相關性的文章數，例如使用者評估後有200篇文章為相關的，而系統共推薦100篇文章，其中有60篇為使用者評估為相關的，則Recall為0.3。

(三) F1-measure

$$F1 = \frac{2 \times \mathrm{Re}call \times \mathrm{Pr}ecision}{\mathrm{Recall} + \mathrm{Precision}} \qquad (4.3式)$$

Precision和Recall的值皆介於0到1之間，最好的評估值為當Precision和Recall皆為1時，故透過F1-measure結合此兩種評估方法，同時考慮Precision和Recall評估的分數。F1值介於0到1間，愈高表示系統的效能愈好，即只有在準確率和召回率的值皆大時，F1的值才會高。

4.4. 實驗結果分析

4.4.1. 實驗一：提出方法與傳統方法之比較

該實驗比較所提出方法之推薦效能，主要大類有兩種內容式過濾法與協同過濾法，每種又區分成傳統模式、結合標籤特徵檔、結合標籤長期時間效應與結合標籤短期時間效應。協同過濾法又可再分成利用餘弦函式與皮爾森系數尋找鄰居。為了讓圖表清楚呈現，替每個方法設定縮寫名，表 4-2 為縮寫名與其說明。其中協同過濾又區分餘弦與皮爾森方式尋找鄰居。

表 4-2 方法名縮寫

編號	縮寫名	說明內容
1	CBF	使用者特徵檔與文件特徵檔相似度之內容式過慮
2	cbf_tag	標籤特徵檔與文件特徵檔相似度之內容式過慮推薦
3	cbf_long	標籤特徵檔與文件特徵檔相似度考量長期時間效應之內容式過慮推薦
4	cbf_short	標籤特徵檔與文件特徵檔相似度考量短期時間效應之內容式過慮推薦
5	CF_cos	使用者餘弦相似度之協同過濾
6	cos_tag	標籤餘弦相似度之使用者餘弦相似度協同過濾
7	cos_long	標籤餘弦相似度考量長期時間效應之使用者餘弦相似度協同過濾
8	cos_short	標籤餘弦相似度考量短期時間效應之使用者餘弦相似度協同過濾
9	CF_corr	使用者皮爾森相似度之協同過濾
10	corr_tag	標籤標記相似度之使用者皮爾森相似度之協同過濾
11	corr_long	標籤標記餘弦相似度考量長期時間效應之使用者皮爾森相似度之協同過濾
12	corr_short	標籤標記餘弦相似度考量短期時間效應之使用者皮爾森相似度之協同過濾

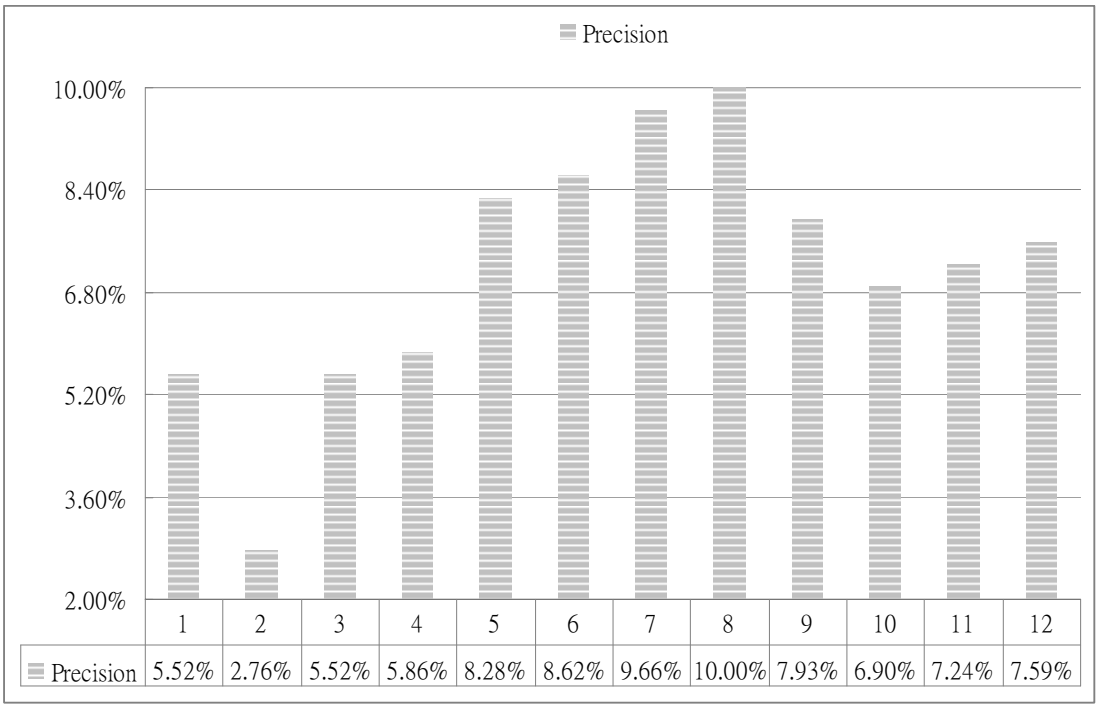
從圖 4-2、圖 4-3 與圖 4-4 中發現 CF 推薦成效較 CBF 高；而運用餘弦函式尋找鄰居又較皮爾森來得高。CF 可有效過濾出使用者感到興趣的文章，所以 CBF 不能而導致表現較差；而在本資料集合中，皮爾森無法找到品質較佳的鄰居，相對地餘弦函式可以找到品質佳的鄰居因而表現較佳。

以分組來觀察，編號 1~4 為 CBF 組、編號 5~8 為 CF(餘弦)組、編號 9~12 為 CF(皮爾森)組，僅有 CF(餘弦)結合新方法的成效會比原先方法佳。尋找品質佳的鄰居為第一步驟，意即有效率地篩選文件，再結合標籤時間效應分析可以提升推薦品質。

觀察各組內部情況，雖然不見得結合新方法的成效會比原先方法佳，但會發

現結合標籤短期時間效應優於結合標籤長期時間效應，而結合標籤長期時間效應又優於標籤特徵檔。後者說明標籤特徵檔僅瞭解使用者興趣，無法曉得使用者興趣程度，不足以提升推薦品質。

cbf_tag 成效差異甚大，其原因為使用者特徵檔由全部收錄文件組成，而標籤特徵檔只由部分收錄文件組成，無法充分表達使用者興趣，只說明該標籤興趣組成成份。因此，標籤為基礎之內容式過濾推薦成效並不是很好，需要加上標籤時間效應才能將雜訊去除以提升推薦品質。



	1	2	3	4	5	6	7	8	9	10	11	12
Precision	5.52%	2.76%	5.52%	5.86%	8.28%	8.62%	9.66%	10.00%	7.93%	6.90%	7.24%	7.59%

圖 4-2 提出方法與傳統方法比較之 Precision

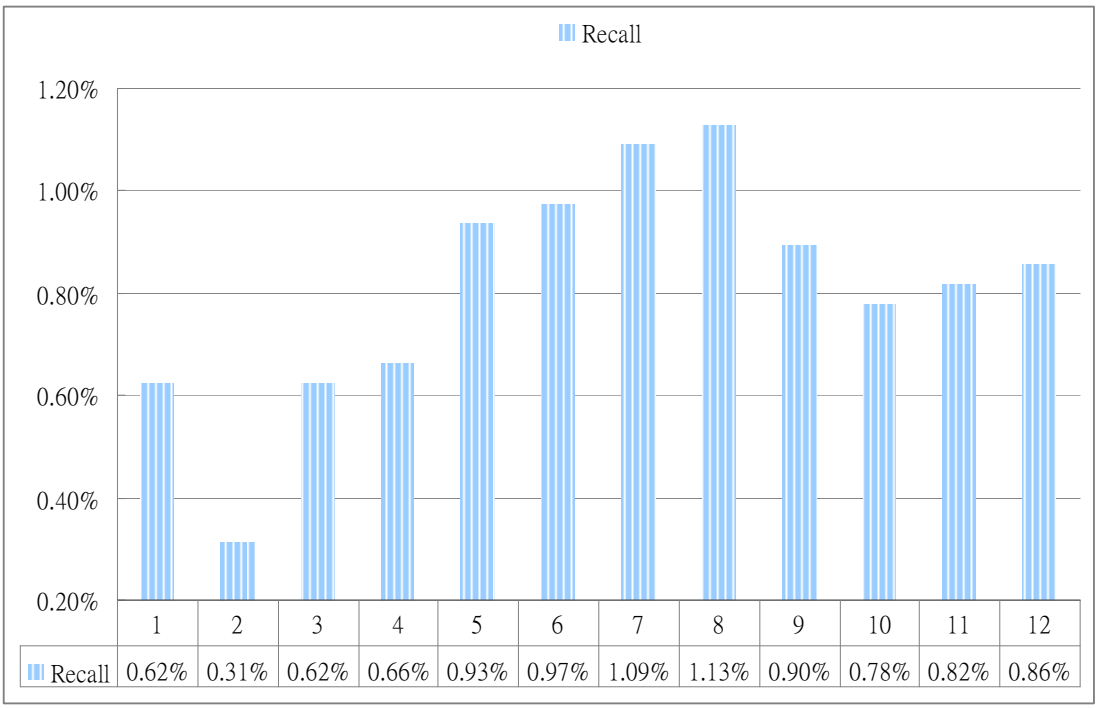


圖 4-3 提出方法與傳統方法比較之 Recall

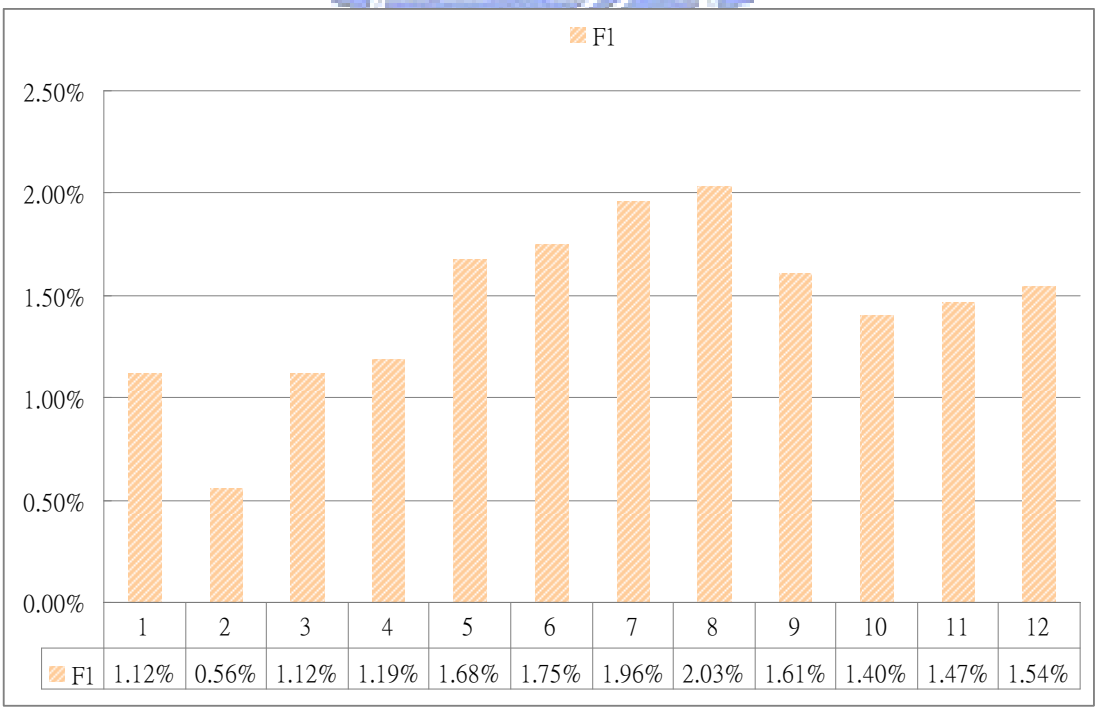


圖 4-4 提出方法與傳統方法比較之 F1

表 4-3 將圖 4-2、圖 4-3 與圖 4-4 整合成一張表。

表 4-3 所有方法之成效比較

	Precision	Recall	F1
CBF	5.52%	0.62%	1.12%
cbf_tag	2.76%	0.31%	0.56%
cbf_long	5.52%	0.62%	1.12%
cbf_short	5.86%	0.66%	1.19%
cf_cos	8.28%	0.93%	1.68%
cf_cos_tag	8.62%	0.97%	1.75%
cf_cos_long	9.66%	1.09%	1.96%
cf_cos_short	**10.00%**	**1.13%**	**2.03%**
cf_corr	7.93%	0.90%	1.61%
cf_corr_tag	6.90%	0.78%	1.40%
cf_corr_long	7.24%	0.82%	1.47%
cf_corr_short	7.59%	0.86%	1.54%

4.4.2. 實驗二：線性組合

此實驗目的為驗證傳統方式與提出方法之間是否有互補性，補足彼此不足，藉以提升推薦之成效。

表 4-4 九種線性組合

CBF_tag	(1-α) * CBF + α * tag
CBF_long	(1-α) * CBF + α * long
CBF_short	(1-α) * CBF + α * short
CF_cos_tag	(1-α) * CF_cosine + α * cos_tag
CF_cos_long	(1-α) * CF_cosine + α * cos_long
CF_cos_short	(1-α) * CF_cosine + α * cos_short
CF_corr_tag	(1-α) * CF_correlation + α * corr_tag
CF_corr_long	(1-α) * CF_ correlation + α *corr_long
CF_corr_short	(1-α) * CF_correlation + α * corr_short

傳統方式與提出方法互相結合下共有 9 種組合，內容式過濾法與協同過濾法，每種分別與結合標籤特徵檔、結合標籤長期時間效應與結合標籤短期時間效應進行線性組合。協同過濾法又可再分成利用餘弦函式與皮爾森系數尋找鄰居，如表 4-4。(1-α) 為傳統方式的權重，而α為新方法的權重。經過參數組合，可以得知何種參數下，線性組合會有較佳的效果。x 軸參數為α值，越偏向 1，新方法比重越大。

從圖 4-5、圖 4-6、圖 4-7 分別 CBF 線性組合之 Precision、Recall、F1，可從三圖中發現 CBF _tag 的標籤比較越大，成效越差。CBF_long 與 CBF_short 的推薦效果皆會逐漸提升，在α=0.6 時達到最高，而慢慢掉落。CBF 結合標籤短期時間效應大多時候比結合標籤長期時間效應與結合標籤特徵檔之線性組合高。

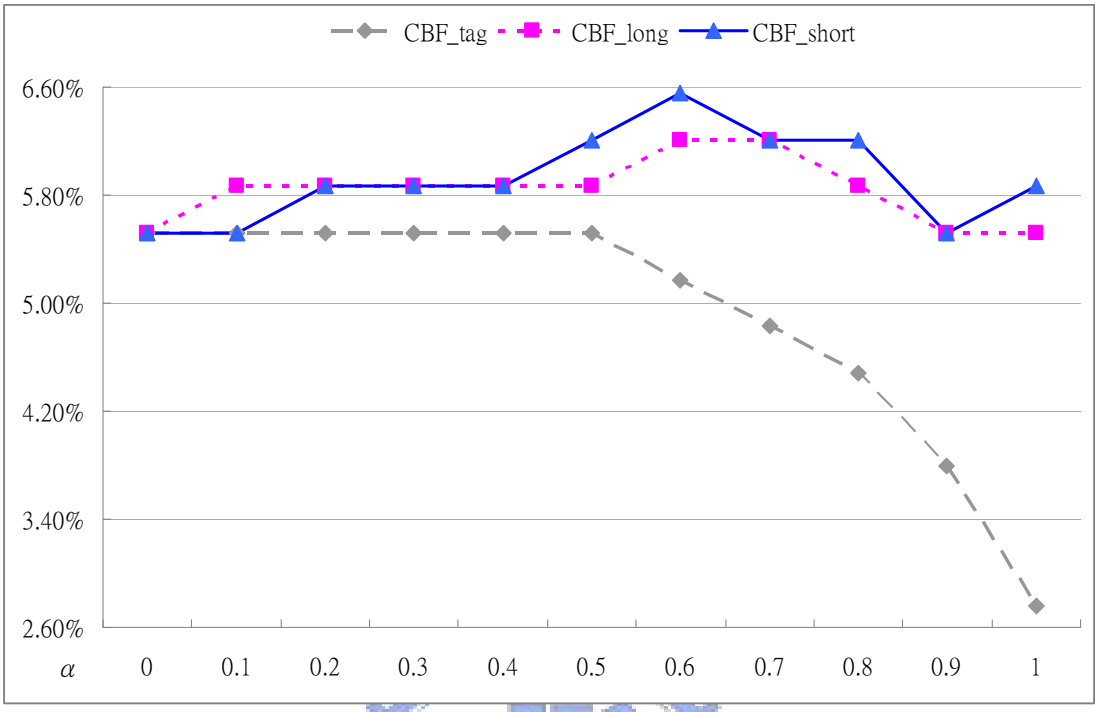


圖 4-5 CBF 線性組合之 Precision

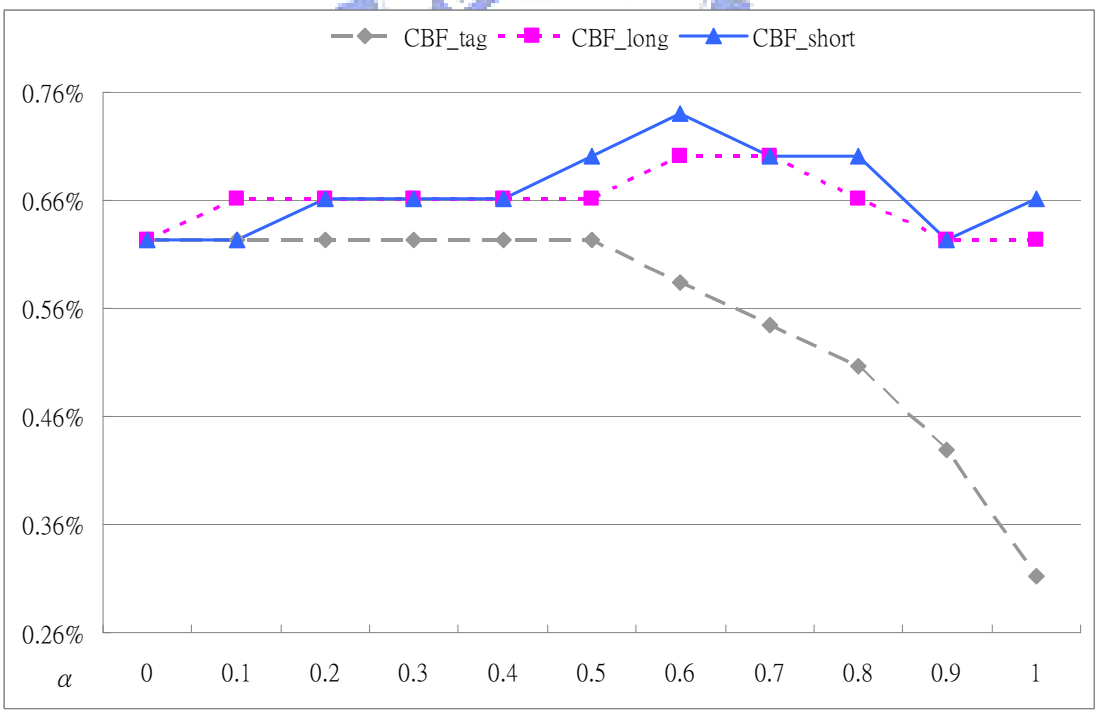


圖 4-6 CBF 線性組合之 Recall

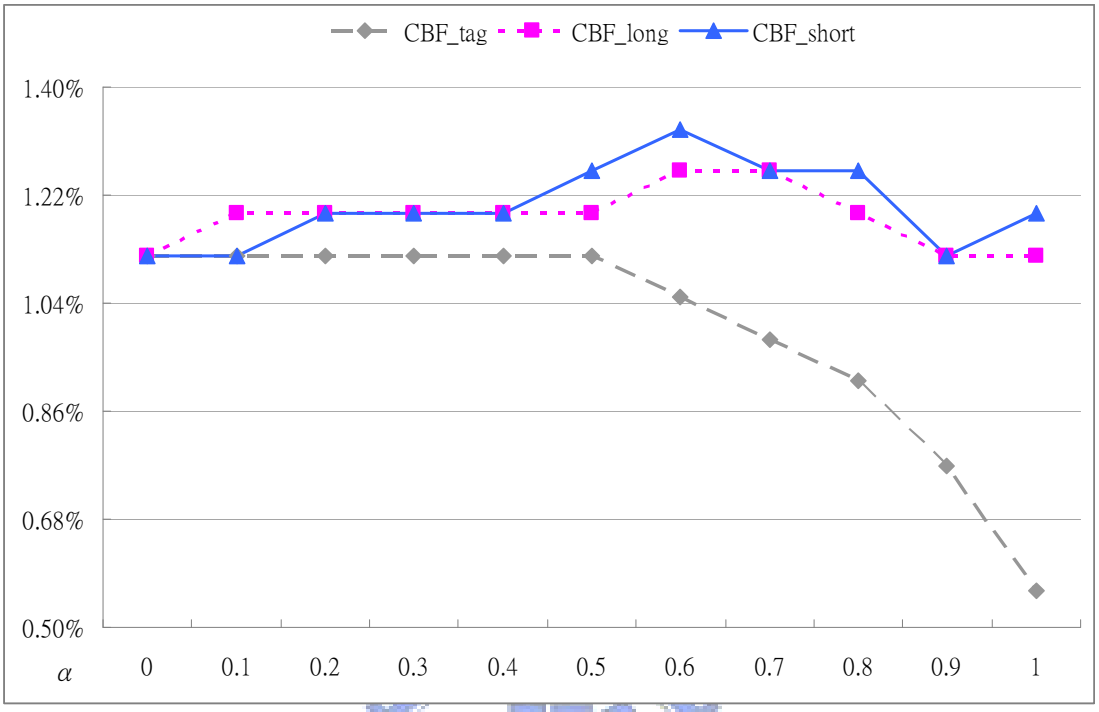


圖 4-7 CBF 線性組合之 F1

從圖 4-8、圖 4-9、圖 4-10 分別 CF(餘弦)線性組合之 Precision、Recall、F1，可從三圖中發現三種結合新方法之線性組合，會隨著 α 值上升達到最高之後而下降，其中以 CF_cos_short 效果最佳，在 α=0.7 時最高；CF_cos_long 在 α=0.8 時最高；CF_cos_tag 隨著 α 值上升，α=0.5 時最佳。另外，可以發現 α>=0.5 的推薦品質皆大於等於 α<0.5 的結果。

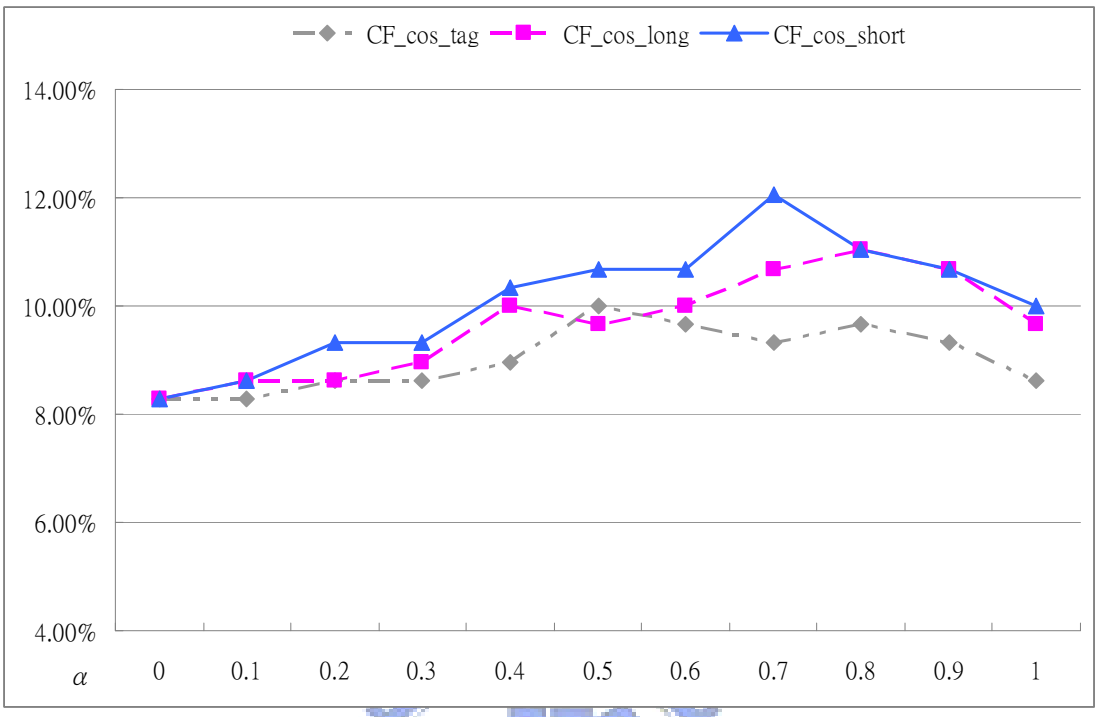


圖 4-8 CF(餘弦函式)線性組合之 Precision

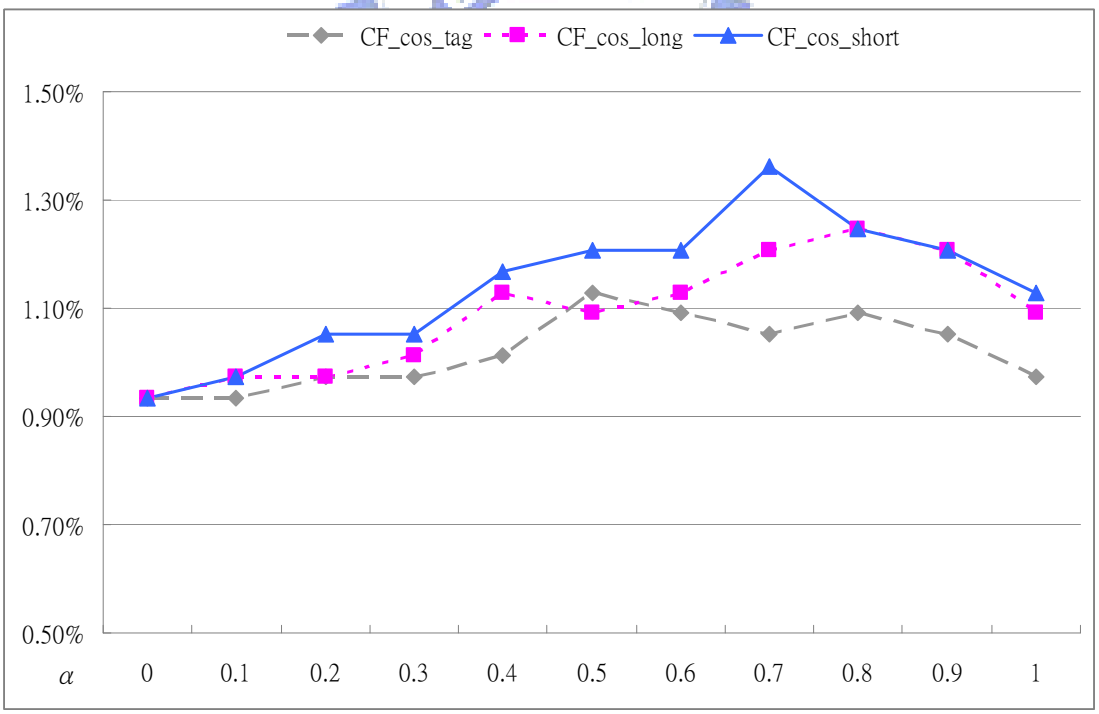


圖 4-9 CF(餘弦函式)線性組合之 Recall

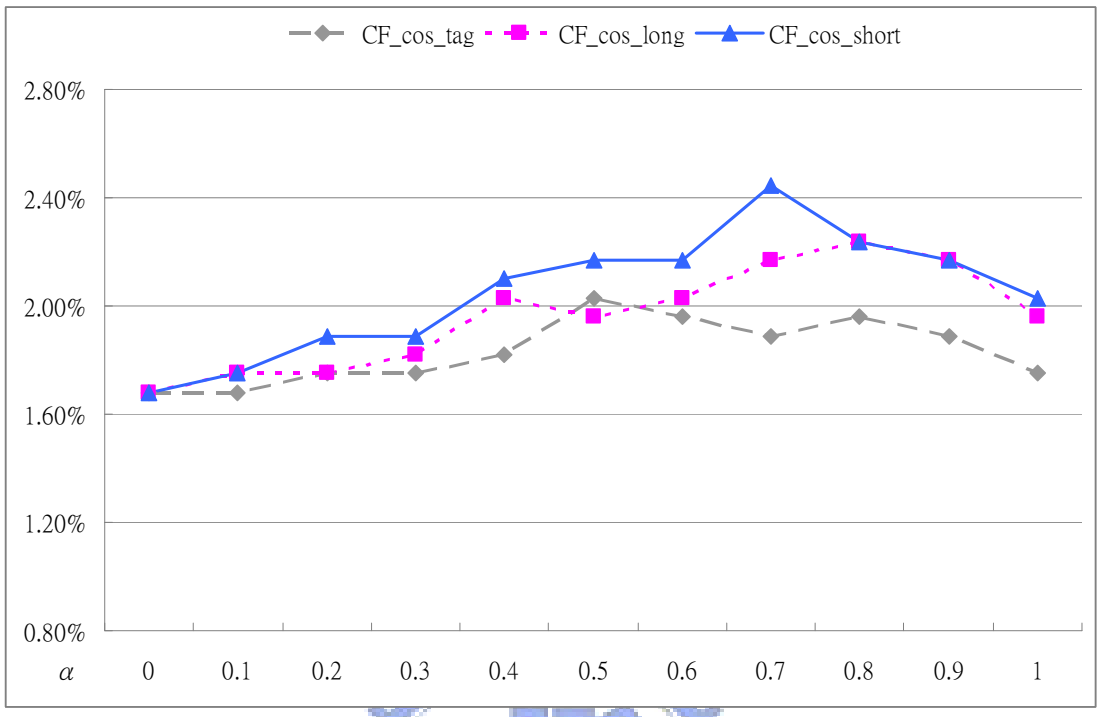


圖 4-10 CF(餘弦函式)線性組合之 F1

從圖 4-11、圖 4-12、圖 4-13 分別 CF(皮爾森)線性組合之 Precision、Recall、F1，可從三圖中發現三種結合新方法之線性組合，，會隨著 α 值上升達到最高之後而下降，其中 CF_corr_short 效果最佳，在 α=0.7 時最高；CF_corr_long 在 α=0.6 時最高；CF_corr_tag 會先下降而後緩步提升，α=0.7 時最佳，然後下降。在圖中可見三條曲線當達到最後高後，成效相對於 CF(餘弦)之線性組合較快下降。

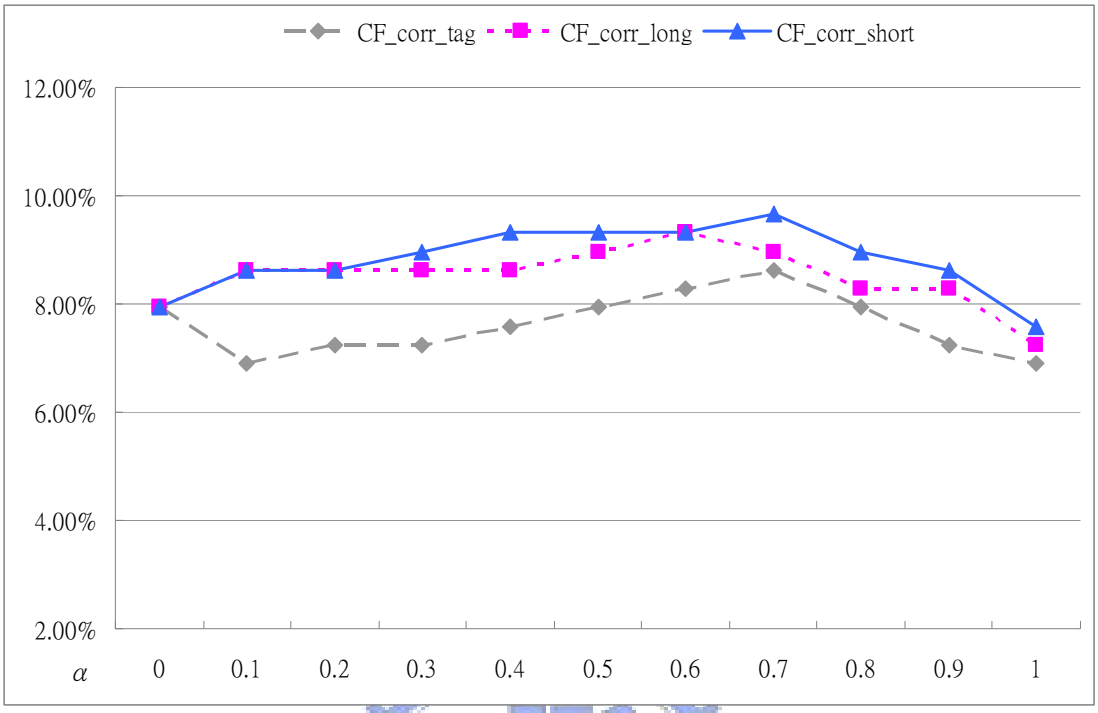


圖 4-11 CF(皮爾森)線性組合之 Precision

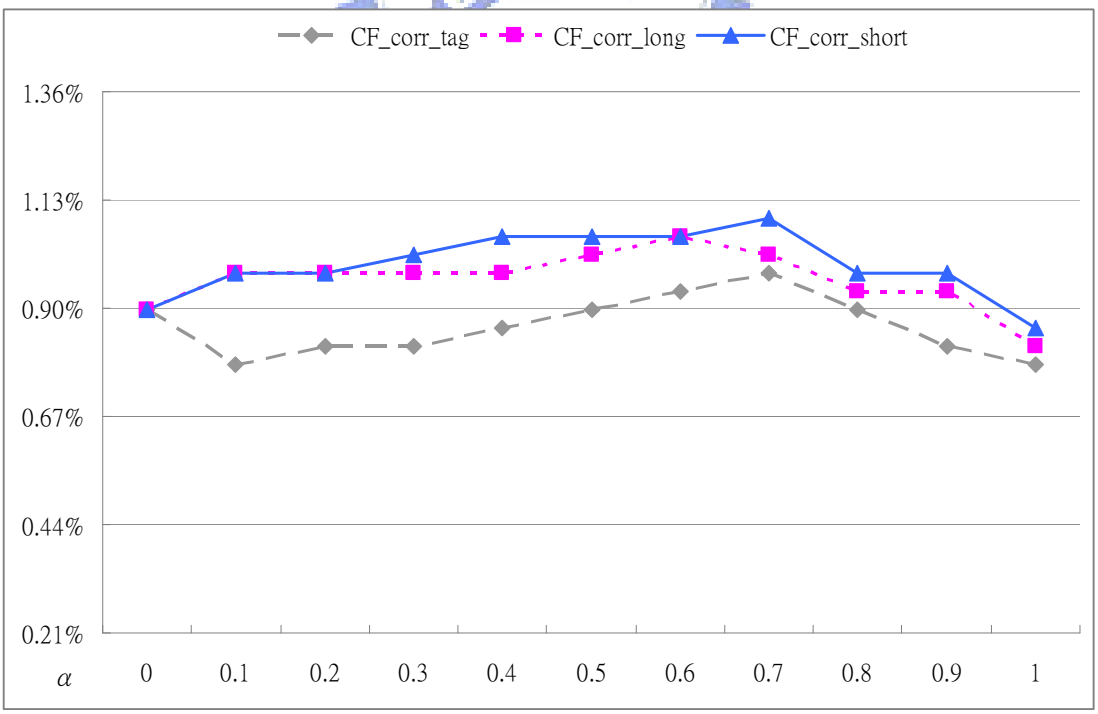


圖 4-12 CF(皮爾森)線性組合之 Recall

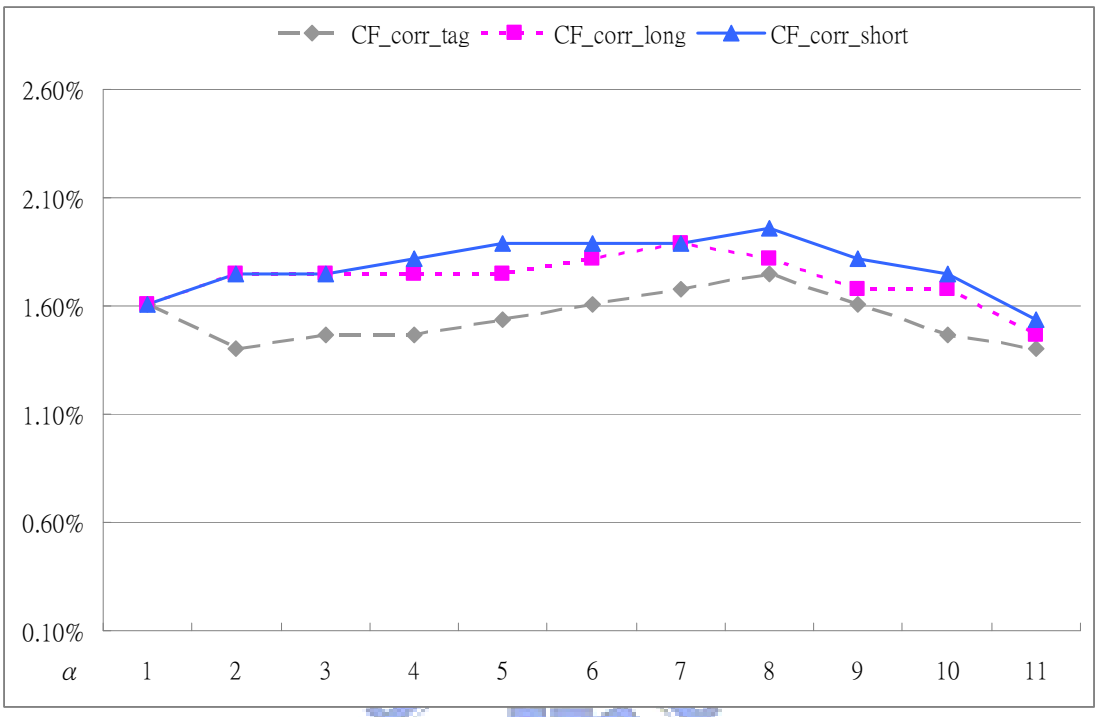
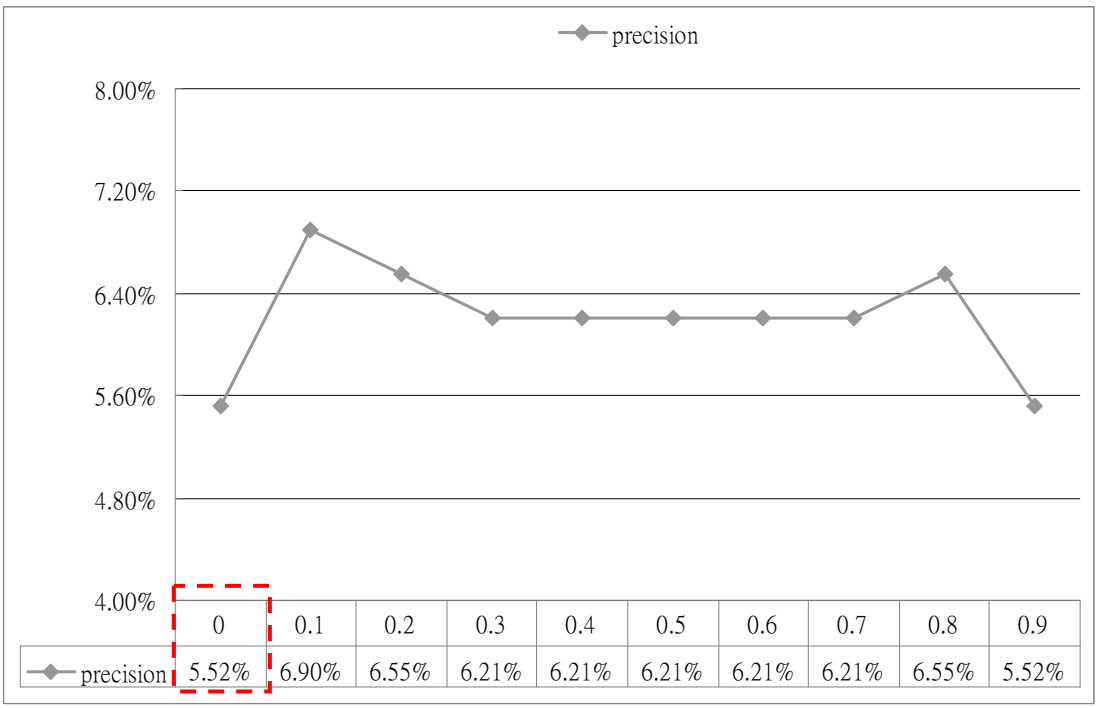


圖 4-13 CF(皮爾森)線性組合之 F1

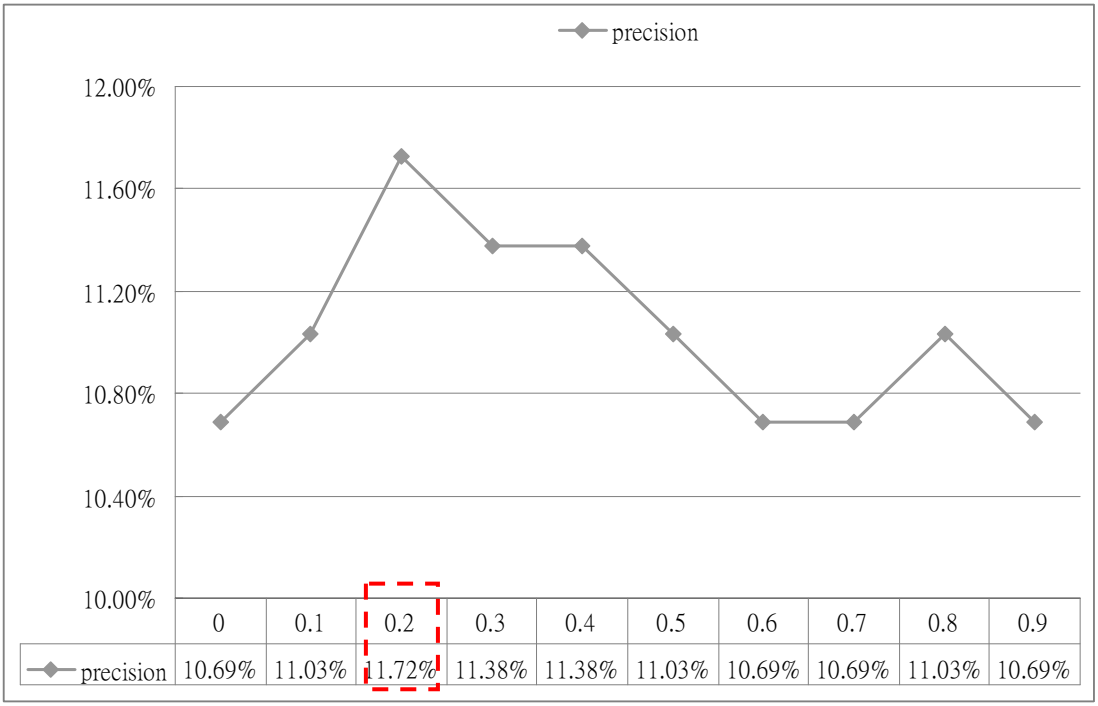
4.4.3. 實驗三：長期因素與短期因素混合

此實驗目的為驗證以標籤長期時間效應為基礎與以標籤短期時間效應為基礎之內容式過慮間是否有互補性，補足彼此不足，藉以提升推薦之成效。因為數據眾多，只列出最好部份之 Precision 圖。圖 4-14 為 CBF 之長期時間與短期時間混合，當 α=0.1、β=0.1、γ=0.8 時值最大，且優於 CBF 結合其他新方法。圖 4-15 為 CF(餘弦)之長期時間與短期時間混合，當 α=0.1、β=0.2、γ=0.7 時值最大；圖 4-16 為 CF(皮爾森)之長期時間與短期時間混合，當 α=0.1、β=0.2、γ=0.7 時值最大。僅有 CBF 之長期時間與短期時間混合後會優於 CBF 結合其他新方法。



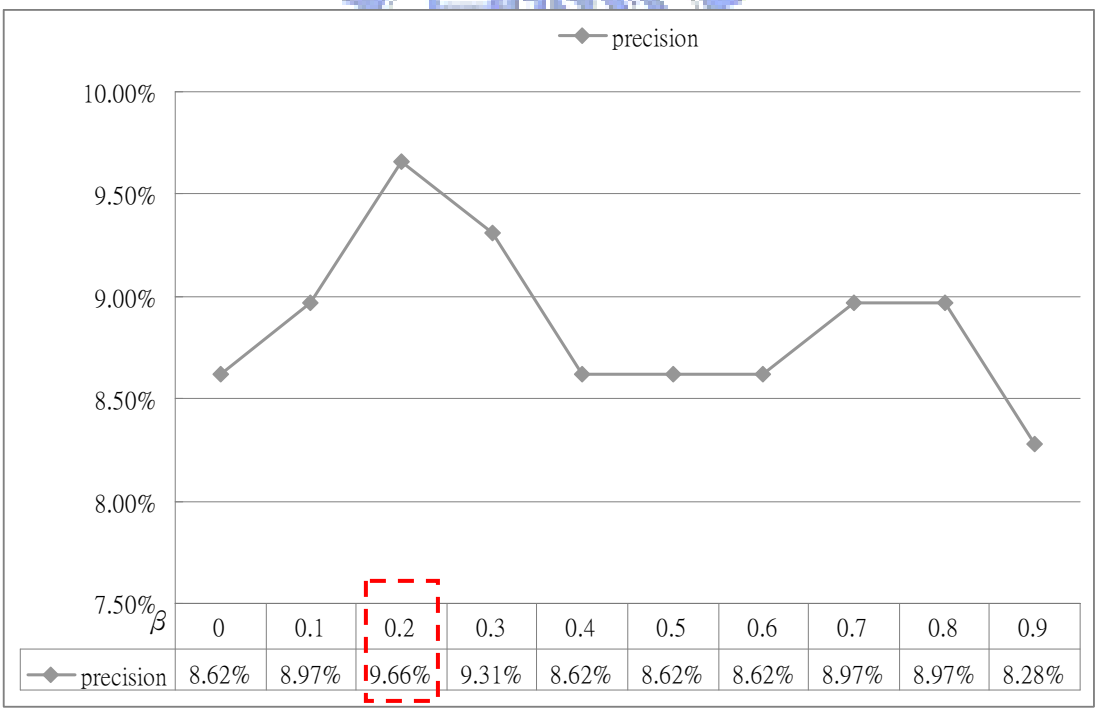
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
precision	5.52%	6.90%	6.55%	6.21%	6.21%	6.21%	6.21%	6.21%	6.55%	5.52%

圖 4-14 α=0.1, CBF 之長期時間與短期時間混合之 Precision



β	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
precision	10.69%	11.03%	11.72%	11.38%	11.38%	11.03%	10.69%	10.69%	11.03%	10.69%

圖 4-15 α=0.1, CF(餘弦)之長期時間與短期時間混合之 Precision



β	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
precision	8.62%	8.97%	9.66%	9.31%	8.62%	8.62%	8.62%	8.97%	8.97%	8.28%

圖 4-16 α=0.1, CF(皮爾森)之長期時間與短期時間混合之 Precision

第五章 結論與未來研究方向

5.1. 結論

處在資訊過載的時代，知識工作者如何有效率找到符合所需文章，以減少耗費更多的時間成本，因此打造個人化文件推薦系統是有存在的必要性，同時，亦能達到知識分享的目的。

個人會擁有很多興趣，並且興趣喜好程度不一。我們以標籤特徵檔描述個人興趣，以及利用時間效應分析表達興趣喜好程度。本研究提出以標籤為基礎之個人化文件推薦方法是以內容式過濾推薦與協同過濾推薦為雛形，再結合協同標籤為切入角度來發展新的推薦方式，分別為標籤特徵檔、標籤長期時間效應分析以及標籤短期時間效應分析。

藉由瞭解文件是否屬於個人興趣以及興趣喜好程度來進行推薦文件。從實驗可得知標籤特徵檔僅是說明使用者興趣，並無法表達興趣程度，需要加入標籤時間效應才能有所改善，意即知曉興趣程度來加強推薦品質。另外，標籤短期時間效應都優於標籤長期時間效應，長期時間涵蓋較多偶而為之的興趣，標籤短期時間效應可以過濾出持續性與強度佳的興趣，所以才有比較好的效果。最後，協同過濾推薦相較於內容式過濾推薦能透過尋找鄰居方式，有效篩選出候選推薦文件，再結合標籤時間效應可以提升推薦品質。

5.2. 未來研究方向

個人化文件推薦為知識分享的重要一環。個人化文件推薦協助使用者有效地找尋所需文件，大幅減少搜尋成本，以增加個人工作效率。在未來整合標籤標記於推薦的研究中可分成幾點作為努力方向，分述如下。

■ 使用者標籤分群

研究中將標籤視為一個興趣，許多興趣可能相似，若將標籤分群可以有效地區分使用者的興趣，不僅從標籤特徵檔瞭解興趣，也可以從一組相似的標籤文字說明興趣。

■ 文件之標籤結構

一份文件被許多人用許多標籤標記，從使用者需求分析這些記錄再轉換成標籤結構，得知文件會滿足何種需求。

■ 即時推薦

本研究中以模擬方式進行實驗，意即把資料切割成訓練資料與測試資料，若能採取真實使用者進行實驗，以即時性方式推薦文章會更具實務上效益。

■ 興趣標籤結構

未來的研究中可以對標籤資料進行統計分析，瞭解使用者的長期與短期興趣標籤穩定性。依據穩定性將資料區開來，再分別以常騎與短期進行探討。

參考文獻

[1] M. Balabanovic, Y. Shoham, "Fab: Content-Based, Collaborative Recommendation," *Communications of the ACM*, pp.66-72, 1997.

[2] K. Bischoff, C.S. Firan, W. Nejdl, R. Paiu, "Can all tags be used for search?" *Proceeding of the 17th ACM conference on Information and knowledge management*, p.p.193-202., 2008

[3] C. Fox, "*A stop list for general text*," ACM SIGIR Forum, vol.24, no. 1, p.p.19-21, 1990

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

[5] N. Glance, D. Arregui, M. Dardenne, "Knowledge pump: supporting the flow and use of knowledge," *Information Technology for Knowledge Management*, 1998.

[6] U. Hanani, B. Shapira, P. Shoval, "Information filtering: Overview of issues, research and systems," *User Modeling and User Adapted Interaction*, vol.11, no.3, p.p.203-259, 2001.

[7] T. Kamba, K. Bharat, M.C. Albers, "The Krakatoa Chronicle-an interactive personalized newspaper on the Web," *Proceedings of the Fourth International World Wide Web Conference*, p.p.11-14, 1995.

[8] H.L. Karen, B.M. Leandro, S.T. Lars, "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms," *Proceedings of 23rd Annual ACM Symposium on Applied Computing*, p.p.16–20, 2008.

[9] J.A. Konstan, B.N. Miller, D. Maltz *et al*, "GroupLens: applying collaborative filtering to Usenet news," *Communications of the ACM*, vol.40 no.3, p.p.77-87, 1997.

[10] B.Y.L. Kuo, T. Hentrich, B.M. Good *et al.*, "Tag clouds for summarizing web search results," *Proceedings of the 16th international conference on World Wide Web*, p.p.1203 – 1204, 2007.

[11] K. Lang, “Newsweeder: Learning to filter netnews,” *Proceedings of the Twelfth International Conference on Machine Learning*, p.p.331-339, 1995.

[12] A. Mathes, “Folksonomies cooperative classification and communication through shared metadata” *Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign*, 2004.

[13] R.J. Mooney, L. Roy “Content-Based Book Recommending Using Learning for Text Categorization,” *Proc. ACM SIGIR, 99 Workshop Recommender Systems: Algorithms and Evaluation*, 1999.

[14] R.Y. Nakamoto, S. Nakajima, J. Miyazaki *et al.*, “Reasonable tag-based collaborative filtering for social tagging systems,” *Proceeding of the 2nd ACM workshop on Information credibility on the web*, p.p.11-18, 2008.

[15] M. Pazzani, D. Billsus, “Learning and Revising User Profiles: The Identification of Interesting Web Sites,” *Machine Learning*, p.p.313-331, 1997.

[16] T. Rattenbury, N. Good, M. Naaman, “Towards automatic extraction of event and place semantics from flickr tags,” *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, p.p.03-110, 2008.

[17] J. Rucker, M.J. Polanco, “Siteseer: Personalized navigation for the web,” *Communications of the ACM*, vol.40, no.3, p.p.73-75, 1997.

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

[19] A. Shepitsen, J. Gemmell, B. Mobasher *et al.*, “Personalized recommendation in social tagging systems using hierarchical clustering,” *Proceedings of the 2008 ACM conference on Recommender systems*, p.p.259-266, 2008.

[20] B. Sigurbjornsson , R. Van Zwol, “Flickr tag recommendation based on collective knowledge,” *Proceeding of the 17th international conference on World Wide Web*, p.p.327-336, 2008.

[21] T. Vanderwal, “Off the Top: Folksonomy Entries.” 2005.

[22] J. Vig, S. Sen, J. Riedl, "Tagsplanations: explaining recommendations using tags," *Proceedings of the 13th international conference on Intelligent user interfaces*, p.p.47-56, 2009.

[23] J. Voss, "Tagging, Folksonomy & Co - Renaissance of Manual Indexing?" *Proceedings of the International Symposium of Information Science*, p.p.234–254, 2007.

[24] J. Wang, B.D. Davison, "Explorations in tag suggestion and query expansion," *Proceeding of the 2008 ACM workshop on Search in social media*, p.p.43-50, 2008.

[25] P.S. Yu, "Data mining and personalization technologies," *Proceedings of the sixth international conference on database system for advanced application*, p.p.6-13, 1999.

[26] S. Zhao, N. Du, A. Nauerz *et al.*, "Improved recommendation based on collaborative tagging behaviors," *Proceedings of the 13th international conference on Intelligent user interfaces*, p.p.413-416, 2008.