結合詞彙共現網路與協同過濾之維基推薦服務系統

劉雯凱
輔仁大學資訊管理研究所
498745276@mail.fju.edu.tw

董惟鳳
輔仁大學資訊管理系
076144@mail.fju.edu.tw

摘要

近年來，Web2.0 應用的快速發展，維基(Wikis)與社交網路網站(Social Network Web Site)的使用普及，也因此產生更多元的協同知識的建立與使用，推薦系統將有助於使用者知識的取得與發現(Knowledge Discovery)，本研究提出一個混合式的線上文件內容推薦技術，主要使用詞彙共現網路分析與文件協同過濾等方法的整合來建構出一個線上文件推薦服務模式與架構，提供使用者以文件關鍵字進行搜尋與知識探勘，利用詞彙關聯找到其他相關程度較高之關鍵字，最後再依照文件評分系統，推薦給使用者關聯主題下較符合使用者偏好之文件。藉由擴展根據關鍵字的相關文件的決策資訊，提升查詢的決策效率。本研究以維基百科的精華文章做為推薦模型的資料來源，建構一個自動化的線上文件推薦雛型系統並進行驗證。

關鍵詞：共現字分析、協同過濾、文件檢索、推薦機制、服務設計

1. 前言

網際網路被視為一個大型的文件資料庫，儲存著各式各樣的文件內容，Web2.0 更可以說是群體智慧的表現，其中，維基(Wikis)即是結合Web2.0精神的一種資訊服務，亦是內容管理系統(Content Management System, CMS)具代表性的一項應用，「它所強調的是一種開放式的知識分享觀念、強化使用者互動，讓網路上的使用者都可以對知識內容做自由的編輯與修改的一個知識管理協作平台，這些知識的正確性仰賴成員之間彼此的監督力量來達成」(朱伯昇，2006)。雖然維基百科與傳統網頁內容的產生方式，有著相當大的變革，但一般使用者在查詢所需要的線上文件內容時，亦可能因為接收到的訊息過多且文章品質可能參差不齊，導致想取得真正所需之內容更加困難。

推薦系統是一種可以減少使用者在過濾大量搜尋結果負擔的一種資訊過濾(Information filtering)機制，用來幫助使用者在複雜的資訊空間中做出決定，基於對使用者的認識以及偏好，針對其需求做出分析，推薦其可能感興趣的項目，不但可以減少使用者搜尋、過濾所需資料的時間成本與困難，還能降低資訊超載(Information overload)現象，本研究提出一種混合模型,雖一樣以關鍵字檢索為出發，但過程中結合了詞彙共現網路分析，將相關程度較強之詞彙串聯起來，找出涵蓋這些具有同一主題性質的線上文件後，再以推薦系統常用的協同過濾方法，找出較符合使用者偏好的文章，產生排序後的文件推薦清單，以期幫助使用者花費較少的時間就能在這一類的線上內容管理系統(Web CMS)或是文件管理系統中找到兼顧內容與品質的文件。

服務科學是一門跨領域的科學，整合科學、工程、心理學、法律與管理等範疇，並以嚴謹的科學方法，將不同領域的專業知識，整合到服務的創新與管理當中，以建立高度專業之服務系統，以及發展出適合使用者的服務類型，透過精緻與複雜的合作關係設計，更可提高服務參與者的互動。本研究以服務科學研究為目的，提出個人化資訊推薦服務系統設計，做為網路線上文件的創新服務應用。

2. 相關文獻

2.1 個人化資訊推薦服務

「服務」意味著兩人面對面的互動，其中一方提供服務，而另一方則接受它；現今，「服務」的領域以及互動關係已經相對變的複雜許多，而「服務系統」(Service Systems)則是可以在不同的設計背景下，結合所創造出來的價值，實踐服務科學之跨領域整合的系統化服務創新與自動化價值共創，服務設計背景包含了人與人的服務接觸、科技的增進、自助服務、多管道以及多平台的服務(Maglio, et al., 2006; Spohrer, et al., 2007)。例如，當使用者登入到 Amazon.com 或其他類似的網路書店時，該網站將透過歷史購物清單以及使用過的搜尋字串，把通用目錄將轉換成個人化目錄並產生推薦清單，反映出使用者的偏好，而一般傳統書店，卻只能藉由具有經驗的員工針對顧客的購買偏好提出建議，Shafer 等學者(2001)曾經表示，「Amazon.com 或其他類似的網路零售商提供了一個相當複雜的推薦服務，該服務的背後需要蒐集與分析數以百萬計的交易和查詢資料，才能夠根據使用者的瀏覽行為動態調整目錄內容與定價」。

因為各種技術的引進，改變了傳統的服務設計基礎，從人與人的服務接觸轉換到以科技輔助的自助服務上，藉由資訊技術的幫助，可自動蒐集使用者在網路上的行為，不需要每次都要使用者主動提供個人資料與相關偏好資訊；也因為技術的引進，微妙的改變了服務中的人際與訊息互動關係。

WWW 在 1990 年代中期成為主流並以相當驚人的速度成長，尤其是在服務系統中，「使用者介面整合了全球網路服務的資源，需要更多複雜的分析與測量技術來克服服務的效率與品質問題」(Edmunds, et al., 2007; Wiggins, 2007)。「只需點擊一次即可(just a click away)對於自助服務和大多數網站都是重要的議題，因為它與服務的可用性與品質有關，成功的線上服務有很大的程度取決於使用者通過網站介面的體驗」(Massey, et al., 2008)。

2.2 維基(Wikis)文章品質的評估

網頁搜尋是資訊檢索領域中一項典型的研究問題，搜尋結果的覆蓋率通常被用來評估一個搜尋引擎(Steve Lawrence and C. Lee Giles, 1998)，而像是網頁排名(Lawrence Page, et al., 1999)和點擊率(Jon M. Kleinberg, 1999)等連結分析技術則是用來衡量網頁受歡迎的程度，利用這些技術，便可以將搜尋結果進行排序，增加搜尋的效率。「網頁排名的得分可以利用點擊次數來計算，無論是從其他網頁連結過來的數目，或是從本身連結出去，具有較高的網頁排名便可認定該網頁具有較高的品質」(Brian Amento, et al., 2000; Xiaolan Zhu, et al., 2000)。

網頁品質是一項主觀的衡量，並沒有絕對的標準，除了網頁排名和點擊率外，相關文獻中也探討了許多其他的指標來衡量網頁的品質(Brian Amento, et al., 2000; Jiwoon Jeon, et al.,2006)，其中，(Xiaolan Zhu and Susan Gauch, 2000)使用了六項指標，並且發現納入這些指標後，普遍改善了搜尋的效果。但是，「這些品質的衡量指標可能並不適合用於維基百科，因為維基百科上的文章都依循著相同的設計介面並且提供相同的取用性，另外，許多數據來源還必須從維基百科外部取得」(Jiwoon Jeon, et al.,2006)。

維基百科豐富的資料內容也引起許多研究者的興趣，包含利用文章編輯歷史的資料格式來評估文章品質(Andrew Lih, 2004)、Honglei Zeng 等學者(2006)運用動態貝氏網路從修改的歷史紀錄中計算文章可被信任的程度等；如何從大量的線上文件中取得品質較高的文章亦是本研究的重點工作之一，本研究亦將以維基百科文章為分析對象，依照使用者偏好，建立相關品質衡量指標。

2.3 線上文件推薦系統

「推薦系統是一種能夠有效透過個人

化的方式，在大量的資訊中，引導使用者選擇出最有用或感興趣的資訊」(劉崇汎等，2006)，推薦系統的基礎概念是從認知科學(cognitive science)、逼近理論(approximation theory)、資訊檢索(information retrieval)、預測理論(forecasting theories)、管理科學(management science)等領域所延伸而來的，1990 年代中期，學者們開始著重在倚賴評分機制的推薦問題上，推薦系統便形成了獨立的研究領域，「從學者 Goldberg 等人在 1992 年提出協同過濾（Collaborative Filtering）後，Resnick 與 Varian 在 1997 年正式提出了推薦系統（Recommender Systems）」(楊亨利、黃仁智，2008)。

根據 Resnick 等學者(1994)的分類，推薦系統主要分成兩類：(1)內容式(Content-based)過濾：以使用者本身過去所購買的物品或喜好，推薦相似的產品。(2)協同式(Collaboration)過濾：推薦的產品是透過具有類似喜好或經驗的其他使用者的喜好。爾後，有許多學者將上述兩種方式混合使用，便逐漸形成第三種分類，(3)混合式(Hybrid)過濾：結合內容式及協同式過濾方法。

無論在學術界或業界，協同過濾式的推薦系統相當多，「Grundy system」被認定為是最早的協同過濾推薦系統，它使用型態(stereotype)建立每個使用者的模型(model)，推薦相關書籍給目標使用者參考，後來的協同過濾推薦系統還有 Tapestry system、GroupLens、Video Recommender 及 Ringo 等，另外，Amazon.com 則與 Grundy system 同樣是運用協同過濾技術在書籍的推薦上。

由於內容式推薦及協同式推薦各有其優缺點，為了讓推薦系統更有效率，過去許多學者會將兩種方法搭配使用，如 Pazzani 等(1999)及 Melville 等(2002)學者結合了內容式過濾與協同式過濾方法形成混合式的推薦方法，目的在提高推薦的正確性並避免各種過濾方法的缺點。

(Huang et al., 2000)採混合方法作為基礎，應用在數位圖書館的領域上，為中文書籍建立了圖形化的推薦系統。(Torres et al, 2004)亦混合了內容式過濾及協同式過濾兩大方法，幫助數位圖書館建立起一套推薦系統，其中，內容式過濾方法使用餘弦相似度找出內容相近的論文，而協同式過濾部分則採用最近鄰演算法，並依照文章的引用關係作為排序的建議，這些學者也發現這種混合方法比單一演算法來的更有效率。

3. 維基推薦服務設計

本研究將提出一個新的服務架構，如圖 1 所示，採用文字探勘方法中的詞彙共現網路分析加上推薦系統中常用的協同過濾方法構成混合式的關鍵字檢索及線上文件推薦服務，讓使用者更快速且正確的取得符合自身偏好的資訊。

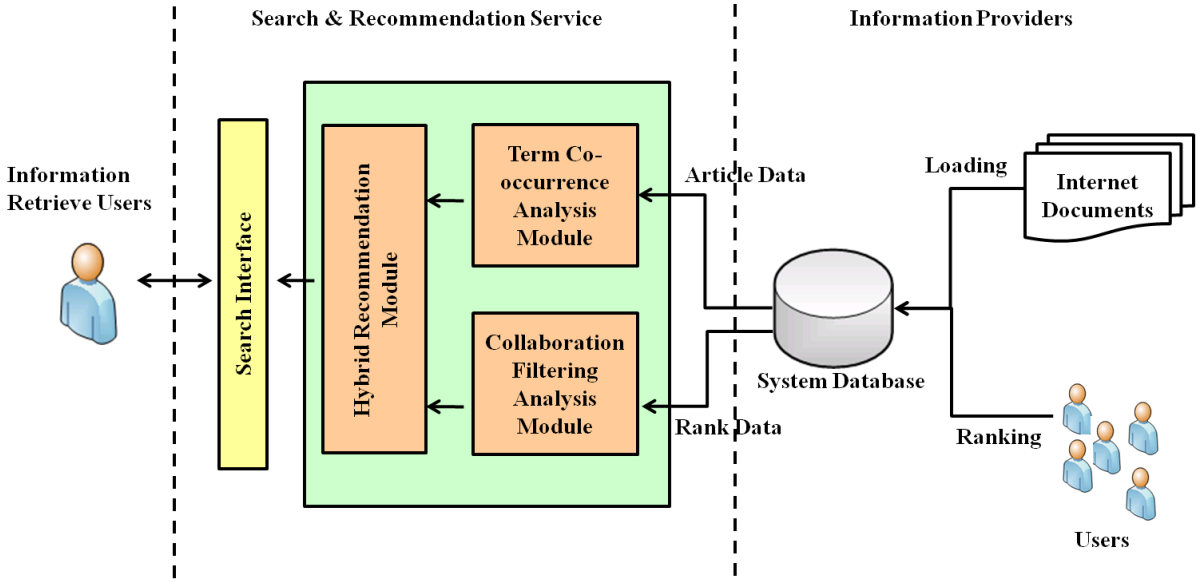

圖 1 研究架構圖

3.1 詞彙共現分析(Term Co-occurrence Analysis)模組

3.3.1 詞彙擷取與權重計算

文字探勘工具已漸趨成熟，本研究選擇使用 SQL Server 軟體 Business Intelligence Development Studio 所提供之 Term Search 及 Term Lookup 相關功能，建立相關專案，除了可自動化進行斷詞工作外，亦能方便的擷取出論文摘要之詞彙並計算各詞彙出現在每篇文章中之次數。

在評估詞彙的重要程度時，TF-IDF (term frequency – inverse document frequency)是常用的一種計算方法，Term frequency(TF)表示詞彙出現在文件中的頻率，詞彙的重要性隨著在某文件中出現的次數成正比增加，但該詞彙如果重複出現在多篇文件時，該詞彙可能就不具價值，隨著在資料群中出現的頻率成反比下降，所以利用 TF 和 IDF 相乘的結果計算出詞彙的重要性。(如公式所示)

$$W_{i,j} = TF * IDF = \frac{n_{i,j}}{\sum_k n_{k,j}} * \log\left(\frac{|D|}{|d: d \ni t_i|}\right)$$

其中詞彙頻率（term frequency, TF）為給定字詞在一文件中出現的次數，再經過正規化的數值；而逆向文件頻率（inverse document frequency, IDF）由總文件數目|D|與包含該詞語之文件的數目做計算。

3.3.2 詞彙相似度

詞彙共現關係(Mutual Information) 可表示兩詞彙間關聯程度，根據詞彙在文件中出現頻率，以及詞彙發生相鄰的機率計算而得。詞彙共現關係越高，表示兩詞彙間的關聯性越強，詞彙共現關係也已廣泛地被應用在各種機器翻譯以及資訊檢索相關研究上。

共現分析最早是應用在科學文獻的共現關係上(White & McCain, 1989)，它可以用來自動架構主題關聯性，目前在資訊擷取(IR)以及文獻計量學(Bibliometrics)的領域當中，已經成為主要的分析工具之一。而在向量空間模式中，最常被使用來設計相似度函數的工具為餘弦係數(Cosine coefficient)，係數值介於 0 到 1 之間，當餘弦值越接近 1，代表著向量夾角越小，兩詞彙之間有極高的相似度。反之，餘弦值越低，代表著兩詞彙的相似度越低。餘弦係數公式如下式所示：

$$\mathbf{Cosine}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{t} \mathbf{x_i y_i}}{\sqrt{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} \mathbf{y}_i^2}}$$

其中 x、y 代表兩個不同詞彙，$\mathbf{x_i}$代表詞彙 x 在第 i 篇論文中所出現的次數，$\mathbf{y_i}$代表詞彙 y 在第 i 篇論文中所出現的次數，而 t 表示為文章總數。

3.2 協同過濾分析(Collaborative Filtering Analysis)模組

協同式推薦藉由紀錄使用者的偏好及相關統計方法，將相同偏好的使用者區分在同一群體，利用相同群體的評價來推測同一群體中其他使用者的偏好，舉例來說，Amazon.com 及 ebay.com 皆有使用這方面的技術，可根據某顧客以往的購買行為以及具有相似購買行為的顧客群，推薦該顧客可能喜歡的商品，雖然不是百分之百的準確，但這方面的應用已經越來越廣泛。

從使用者角度出發的協同過濾推薦具有許多優點，包含：內容分析的自動化程度高、可以共用其他使用者的經驗以提高推薦的精確度、發現潛在的興趣偏好以及新資訊的推薦能力等。雖然協同過濾作為推薦的機制已經受到相當廣泛的應用，但其仍有許多問題上待解決，問題包含：稀疏性問題(sparsely problem)、冷啟動問題(cold-start problem)及擴充性問題(scalability problem)。

以使用者為基礎（User-based）的協同過濾必須找到與使用者偏好相同的一群使用者，計算兩使用者之間的相似度，目前較多使用的相似度演算法有 Person Correlation Coefficient、Cosine Similarity、Adjusted Cosine Similarity 等，本研究將採用 Pearson Correlation Coefficient 進行分析，相關係數值的範圍為-1~1，當值接近 1 時表示兩個變數之間有很強的正向關係，接近-1 時表示有很強的負向關係，接近 0 時則表示沒有線性相關，

$$\mathbf{sim_{a,u}} = \frac{\sum_{i=1}^{m}(\mathbf{r_{a,i}} - \bar{\mathbf{r}}_\mathbf{a}) * (\mathbf{r_{u,i}} - \bar{\mathbf{r}}_\mathbf{u})}{\sqrt{\sum_{i=1}^{m}(\mathbf{r_{a,i}} - \bar{\mathbf{r}}_\mathbf{a})^2 * \sum_{i=1}^{m}(\mathbf{r_{u,i}} - \bar{\mathbf{r}}_\mathbf{u})^2}}$$

其中$\mathbf{sim_{a,u}}$表示主要使用者(a, 本研究中為進行搜尋動作的使用者)與其他使用者(u)的相關係數，$\mathbf{r_{a,i}}$表示主要使用者某篇論文的評分，$\bar{\mathbf{r}}_\mathbf{a}$是主要使用者對論文的平均分數，m 則是論文的總數。

有了最相近的使用者集合，就可以針對主要使用者的偏好(評分)進行預測，產生推薦清單，本研究則採用下列公式計算加權平均後的結果作為主要使用者對論文的評分預測。

$$\mathbf{P_{ai}} = \bar{\mathbf{r}}_\mathbf{a} + \frac{\sum_{u=1}^{k}(\mathbf{r_{u,i}} - \bar{\mathbf{r}}_\mathbf{u}) * \mathbf{sim_{a,u}}}{\sum_{u=1}^{k} \mathbf{sim_{a,u}}}$$

其中$\mathbf{P_{ai}}$表示為主要使用者對文章 i 的預測評分，$\mathbf{sim_{a,u}}$表示主要使用者(a)與使用者(u)之間的相關係數，k 表示選擇鄰近使用者的數目，$\mathbf{r_{u,i}}$表示使用者 u 對文章 i 的評分，$\bar{\mathbf{r}}_\mathbf{a}$是主要使用者對文章的平均分數。

圖 3 簡易表示此小節綜合概念，當預測主要使用者對第五篇論文的分數時，首先會利用 Pearson 相似度公式找到相關性較高具有相同偏好的鄰近使用者 User2 和 User3，而鄰近使用者數目可視資料量的多寡作調整，接著透過預測評分之公式計算出結果。

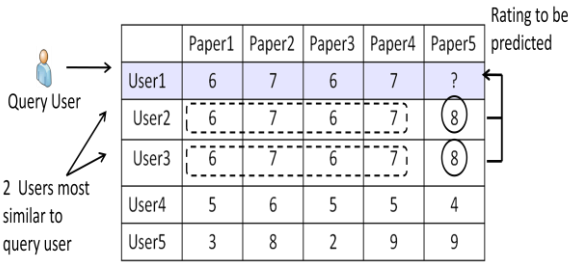

圖 2 協同過濾概念示意圖

為了運用協同過濾技術，必須紀錄使用者的相關輪廓，本研究將利用模擬方式進行此部份的研究。

3.3 結合詞彙共現及協同過濾之混合式推薦模組(Hybrid Recommendation Module)

當使用者搜尋某關鍵字時，便可透過詞彙網路的基礎找到其他關聯程度較高之詞彙。

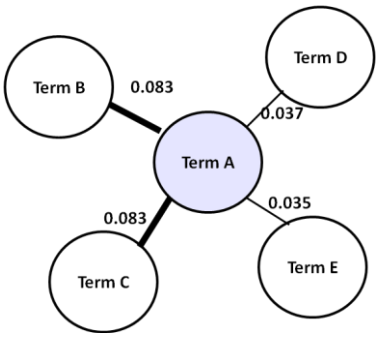

圖 3 詞彙共現網路示意圖

由詞彙共現網路分析圖形中可以看出，與詞彙A具有關聯性的詞彙分別為詞彙B、C、D、E其中又以詞彙B和詞彙C的相關性較高(cosine coefficient 數值均為 0.083)，因此，除了使用者的搜尋目標「詞彙A」外，本系統亦會推薦給使用者包含「詞彙B」及「詞彙C」的所有文章清單。

除了找出包含關聯性較強的詞彙並列出包含該關鍵字的文章清單外，其中，清單的順序是以詞彙出現在文章中的頻率及前一小節協同過濾方法所計算得來的分數作加權平均後的結果為依據，與一般文件資料庫使用論文名稱、作者、年份等排序基礎不同，以期替使用者省下瀏覽大量的論文的時間，更快得到真正感興趣或對使用者本身有幫助的文件。

表 1 產生推薦清單

Term	Document	FQ.	CF.	MN.
TERM A	Doc.27	185	5	0.50
TERM A	Doc.18	118	3	0.19
TERM A	Doc.33	6	6	0.02
TERM B	Doc.96	49	8	0.80
TERM B	Doc.05	7	5	0.07
TERM C	Doc.57	4	5	0.50
TERM C	Doc.22	2	8	0.40

其中，FQ.為該詞彙出現在某篇文章中之頻率，CF.為某使用者的評分資訊，MN.為加權平均後的數值，公式如下：

$$\mathbf{MN_{t,d}} = \frac{\mathbf{FQ_{t,d}}}{\mathbf{MAX}(\mathbf{FQ_{t,d}})} + \frac{\mathbf{CF_{t,d}}}{\mathbf{MAX}(\mathbf{CF_{t,d}})}$$

其中 t 表示某關鍵字，d 為包含關鍵字 t 之文章。

4.研究結果

本研究服務系統架構流程如下圖：

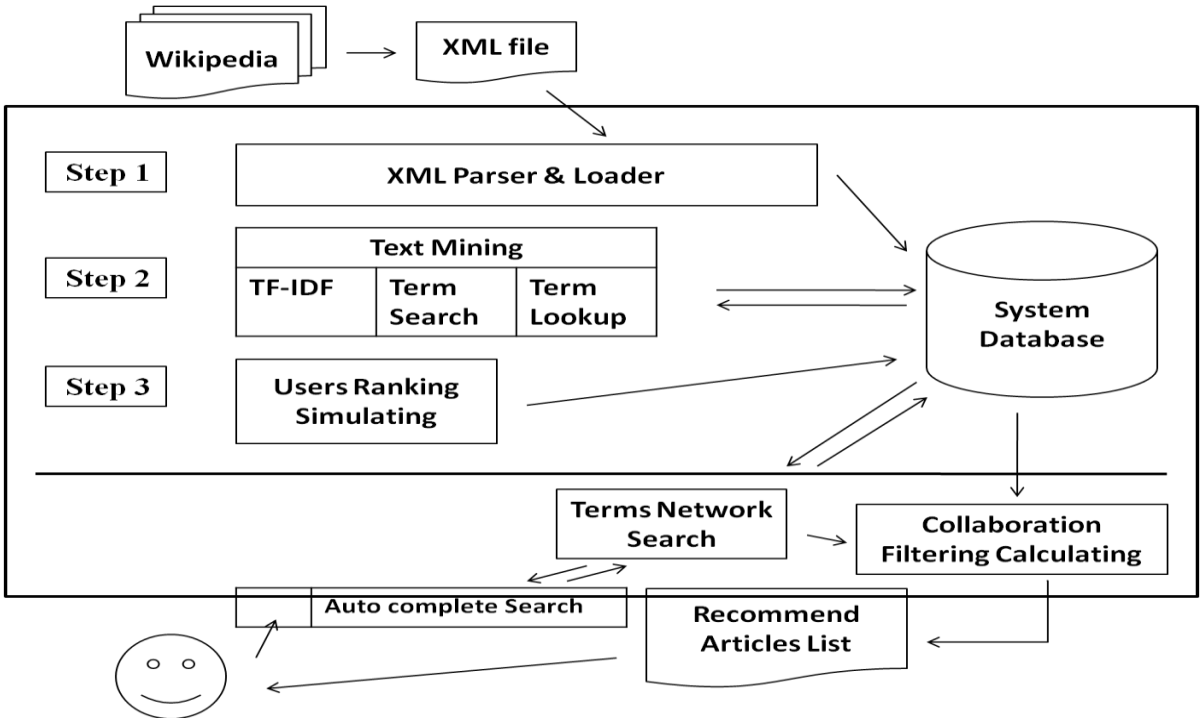

圖 4 維基推薦服務系統架構流程圖

使用者欲查詢維基百科文章(本研究簡稱文件)，輸入關鍵字進行搜尋時，本系統會利用 AJAX 自動完成技術，提供依照 TF-IDF 權重排序過的關鍵詞彙讓使用者選擇，並透過 Cosine 相似度計算出其他與搜尋詞彙共現程度較高之詞彙，接著找出包含這些詞彙之文章，最後再依照協同過濾計算的加權結果，將這些文章加以排序並推薦給使用者做為參考。

4.1 實驗資料來源

維基百科(Wikipedia)是線上且免費的，使用者可以輕易讀取它的內容，也可以貢獻相關的知識進行編輯，維基百科儼然已經成為世界上內容最豐富、協同作業的百科全書,目前已擁有超過兩百種語言版本，超過一千三百萬的使用者，英文版也有三百五十萬篇以上的文章，透過網路社群的力量，維基百科對文章內容有相當嚴格的管控，可迅速對有問題的文章內容進行處理。

目前在英文版文章中,有三千多篇被評選為精華文章，評選的準則包含：具有專業水準、全面性的、有文獻或可靠資料支撐的、無偏見的以及較不受事件發生或時間影響的，故本研究選擇此資料來源進行分析。

維基百科提供多種方式下載其文章內容，本研究選擇以該網站所提供之功能(如圖 4-1)，將精華文章匯出成 XML 檔案。本研究亦自行撰寫了相關程式將該網站提供之 XML 檔案經適當解析後儲存到系統資料庫中。

4.2 詞彙共現分析

SQL Server 軟體在詞彙擷取功能提供了 TF-IDF 計算，透過詞彙擷取轉換編輯器，便可以計算出每個關鍵詞彙的 TF-IDF 數值，產生全部詞彙共 420103 個，下表擷取了前 50 個權重較高之詞彙片語，分數介由 5938.386 到 2276.253 之間。藉由產生這樣子的清單，便可瞭解所有文件中權重較高的詞彙，可提供使用者作為查詢線上文件資料庫的參考。

表 2 權重前五十名之詞彙片語

RANK	TERM	SCORE	RANK	TERM	SCORE
01	New York	5938.386439	26	fs player	2963.675314
02	tropical storm	5847.974379	27	Winter Olympics	2860.484289
03	United States	5465.817734	28	San Francisco	2830.072209
04	New York Times	5089.560487	29	Associated Press	2804.150504
05	video game	4787.730638	30	New Jersey	2742.828498
06	United Kingdom	4579.835567	31	Star Trek	2736.280315
07	NHL season	4554.907817	32	font color	2727.411083
08	English football	4288.301689	33	span id	2718.057261
09	National Hurricane Center	4255.716502	34	prime minister	2694.613475
10	vcite journal	4215.477527	35	South Park	2665.88104
11	World War II	3964.765519	36	Cambridge University Press	2627.45656
12	New South Wales	3945.489831	37	Summer Olympics	2564.327567
13	ref name	3872.267412	38	Soviet Union	2487.330857
14	New Zealand	3480.239908	39	El Greco	2480.226229
15	Virginia Tech	3388.48316	40	Western Australia	2471.411097
16	Rolling Stone	3346.748792	41	North America	2459.621726
17	Final Fantasy	3339.78048	42	Link FA	2437.337654
18	dead link	3334.906405	43	Los Angeles	2428.862668
19	ref group	3231.435618	44	Oxford University Press	2407.363486
20	ice hockey	3222.197981	45	South Australia	2396.803768
21	Puerto Rico	3088.976164	46	BBC NEWS	2370.773713
22	World War	3081.874657	47	BBC Sport	2332.916965
23	Greater Manchester	3060.276226	48	Washington Post	2312.110498
24	Kingdom Hearts	3053.772544	49	vcite web	2311.996555
25	solar system	3051.731666	50	North Carolina	2276.253563

透過上述步驟找出存在於每篇文章的詞彙後，本研究接著使用 SQL Server 軟體所提供之「詞彙查閱」功能，計算出每個詞彙出現在每篇文章的次數，有了詞彙與文章的對應關係後，便可計算出詞彙共現 Cosine 相似度數值，為了明確顯示詞彙共現網路，圖 5 中選擇以表 1 權重最高之五十個詞彙片語，透過 UCINET 軟體繪製出網路圖形，並選擇顯示出中介中心性 (Betweenness Centrality)最高之二十五個節點。

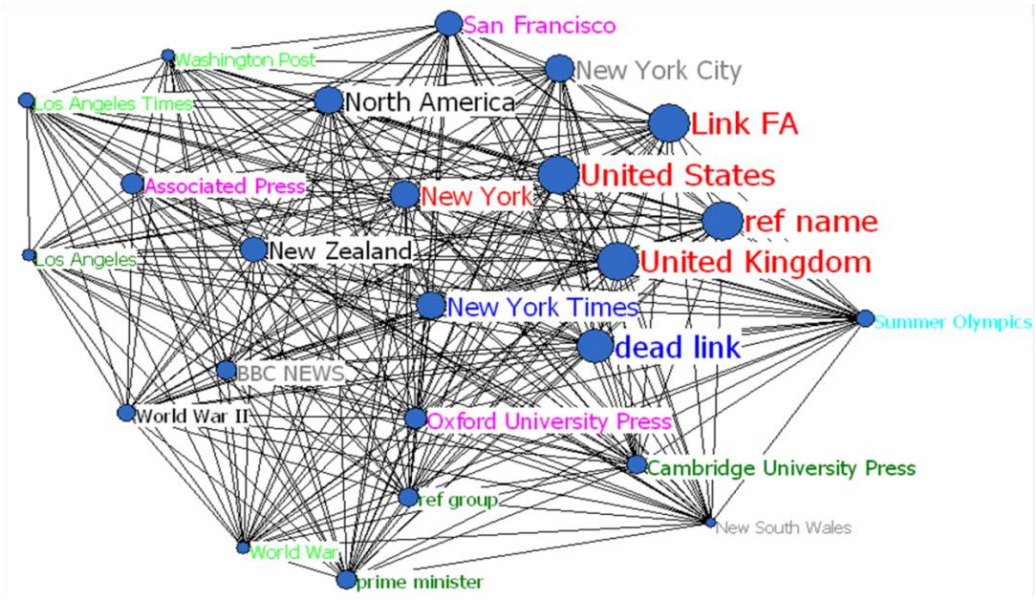

圖 5 詞彙片語網路圖

4.3 協同過濾之資料模擬結果

研究過程中，模擬了十位使用者對每篇文章進行評分的動作，分數以亂數方式產且介於 1~10 分之間，並以 0 分表示使用者尚未評分之項目，藉由使用者相似度之計算，便可得到尚未評分項目之預測分數，以 USER1 為例，計算所得最相近之使用者為 USER10 和 USER2，皮爾森相關係數分別為 0.003210915799316599 及 0.003210915799316507，經由協同過濾方法計算後的預測評分如表 3 所示。

表 3 模擬協同過濾計算結果(使用者 1)

文章名稱	USER1	USER2	USER3	USER4	USER5	USER6	USER7	USER8	USER9	USER10
0.999...	7	5	7	4	8	5	3	4	10	9
2009 …	6	0	7	5	7	5	0	4	0	9
Puerto …	8	10	0	3	0	3	5	9	5	4
Axis naval …	6	7	6	8	0	3	9	0	3	0
Battle of …	3	5	9	5	4	8	5	5	6	7
Guy …	3	0	10	9	8	3	7	0	4	3
Rachel …	3	9	8	0	6	7	4	9	0	10
Golden …	9	10	7	6	3	0	10	3	7	8
William …	9	9	6	5	5	8	8	0	6	3
Ima Hogg	6	0	4	9	8	6	7	4	0	6
…..	…	…	…	…	…	…	…	…	…	…

4.4 服務介面與結果顯示

圖 6 是使用者一(USER1)作查詢後的結果，當使用者輸入欲查詢的關鍵字詞彙「Harry Potter」後，本系統會自動運算出與「Harry Potter」詞彙共現程度較高的其他詞彙，如「J.K.Rowling」等，列出包含該詞彙的文章，並依照詞彙出現在文章中的頻率以及協同過濾的評分兩者之加權平均作為排序的依據。圖 7 則是模擬使用者二(USER2)的查詢結果，雖然查詢的詞彙依舊為「Harry Potter」，但因為個人的喜好不同，使得協同過濾的結果影響了推薦文章排列的順序。

Auto Recommendation service

USER1

Harry Potter

SUBMIT

關鍵字	文章標題	FQ	CF	MN
Harry Potter	Religious debates over the Harry Potter series	185	5	0.5
	J. K. Rowling	114	8	0.49297
	Emma Watson	118	3	0.19135
	Jonathan Strange & Mr Norrell	6	6	0.01946
	The Simpsons Movie	2	8	0.00865
	Chess	1	8	0.00432
	Hippocrates	1	7	0.00378
	Meet Kevin Johnson	1	7	0.00378
	I Know Why the Caged Bird Sings	1	7	0.00378
	Nigel Kneale	1	6	0.00324
	University of Michigan	1	5	0.0027
	The Shape of Things to Come (Lost)	1	5	0.0027
	Sirius	1	5	0.0027
	The Beginning of the End (Lost)	1	3	0.00162
J. K. Rowling	J. K. Rowling	49	8	0.8
	Religious debates over the Harry Potter series	7	5	0.07143
	Chess	1	8	0.01633
	Jonathan Strange & Mr Norrell	1	6	0.01224
	Emma Watson	1	3	0.00612
Hogwarts	Religious debates over the Harry Potter series	4	5	0.5
	J. K. Rowling	2	8	0.4
	Emma Watson	2	3	0.15
	Jonathan Strange & Mr Norrell	1	6	0.15
Azkaban	J. K. Rowling	9	8	0.6
	Emma Watson	12	3	0.3
	Religious debates over the Harry Potter series	3	5	0.125
Deathly Hallows	J. K. Rowling	9	8	0.65455
	Religious debates over the Harry Potter series	8	5	0.36364
	Emma Watson	11	3	0.3

圖 6 使用者 1 的查詢結果

Auto Recommendation service

USER2

Harry Potter

SUBMIT

關鍵字	文章標題	FQ	CF	MN
Harry Potter	J. K. Rowling	114	8	0.49297
	Emma Watson	118	7	0.44649
	Religious debates over the Harry Potter series	185	4	0.4
	Jonathan Strange & Mr Norrell	6	8	0.02595
	The Simpsons Movie	2	6	0.00649
	Meet Kevin Johnson	1	9	0.00486
	University of Michigan	1	8	0.00432
	Nigel Kneale	1	8	0.00432
	The Beginning of the End (Lost)	1	7	0.00378
	Chess	1	6	0.00324
	Sirius	1	5	0.0027
	I Know Why the Caged Bird Sings	1	5	0.0027
	Hippocrates	1	4	0.00216
	The Shape of Things to Come (Lost)	1	3	0.00162
J. K. Rowling	J. K. Rowling	49	8	0.8
	Religious debates over the Harry Potter series	7	4	0.05714
	Jonathan Strange & Mr Norrell	1	8	0.01633
	Emma Watson	1	7	0.01429
	Chess	1	6	0.01224
Hogwarts	Religious debates over the Harry Potter series	4	4	0.4
	J. K. Rowling	2	8	0.4
	Emma Watson	2	7	0.35
	Jonathan Strange & Mr Norrell	1	8	0.2
Azkaban	Emma Watson	12	7	0.7
	J. K. Rowling	9	8	0.6
	Religious debates over the Harry Potter series	3	4	0.1
Deathly Hallows	Emma Watson	11	7	0.7
	J. K. Rowling	9	8	0.65455
	Religious debates over the Harry Potter series	8	4	0.29091

圖 7 使用者 2 的查詢結果

5. 結論

目前 Web2.0 共享平台日益擴增，維基(Wikis)、部落格(Blog)、與社交網站 face book 等，大量使用者的參與知識分享，帶來了更龐大的網路分享文件，本研究探討以分析維基文件之間的關聯性產生一種有效率的查詢推薦方式，然而，「文件推薦系統」是一種可以減少使用者在過濾大量搜尋結果負擔的一種資訊過濾(Information filtering)方法，可減少使用者搜尋、過濾所需資訊的時間成本，並降低資訊超載現象(Information overload)。本研究也以使用者的偏好與檢索需求，發展具有「詞彙共現

分析」與「協同推薦」機制的整合進行文件推薦。針對使用者的偏好與需求，作更進一步的推薦基礎。本研究採用詞彙共現分析方法，讓使用者在檢索時描述不完整或對相關知識有限等情形下，能自動延伸至可能隱藏的相關主題詞彙，增加搜尋廣度；並同時運用協同過濾方法，根據個人偏好作出推薦預測，並依此排列搜尋結果的順序，提升推薦效果。

詞彙共現分析以及協同過濾方法在文字探勘和推薦系統的領域中，都是經常被使用到的技術，本研究將之結合並搭配熱門的維基百科文章進行個人化的線上文件推薦服務系統設計，大幅提升實務應用貢獻。本研究除自行撰寫的程式外，搭配軟體工具輔助，可快速建立完整的自動化分析模組,且能應用於多種不同的資料來源，如此便可提升分析效率並降低研究門檻。

研究限制則是在以使用者為基礎的協同過濾方法時，必須蒐集使用者對文件之偏好，但目前大多數線上文件管理系統，包括本研究將分析的維基(Wikis)文章，都尚未提供相關的評分機制，故本研究相關文件評分資料將使用模擬方式產生。

參考文獻

[1]朱伯昇，*基於 Wiki 技術之知識管理系統設計*，文化大學資訊管理研究所碩士論文，2006。

[2]陳言熙，*運用文字探勘技術協助建構公司治理本體知識*，國立政治大學會計研究所碩士論文，2006。

[3]曾元顯，*高階神經網路及其應用*，國立台灣大學資訊工程研究所博士論文，1992。

[4]黃群弼，*中文繁簡等義詞自動辨識之研究*,國立政治大學資訊科學所碩士論文，2009。

[5]楊亨利、黃仁智，*具整體觀點考量之推薦系統：以家庭親子為例*，中華管理評論國際學報，第十一卷三期，2008/08。

[6]溫文喆，*詞義相似度的社會網路分析研究*，國立政治大學資訊科學研究所碩士論文，2008。

[7]劉崇汎等，*智慧型個人化多媒體推薦系統之建置*，成功大學資訊工程學系，2006。

[8]Andrew Lih, Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource, *In Proc. of the 5th International Symposium on Online Journalism*, 2004.

[9]Bailey, J., Gruhl, D., Maglio, P. and Spohrer, J., Steps Toward a Science of Service Systems. *IEEE Computer*, 40(1), 2007, pp.71-77.

[10] Brian Amento, et al., Does “authority” mean quality? predicting expert quality ratings of Web documents, *In Proc. of SIGIR*, 2000, pp.296-303.

[11] Councill Lee Giles and Steve Lawrence, Searching the World Wide Web, *Science*, 280(5360), l 1998, pp.98-100.

[12] Drucker. S., Edmunds, A., Morris, D. and White, R., Instrumenting the Dynamic Web. *Journal of Web Engineering*, 6(3), 2007, pp.244-260.

[13] Honglei Zeng, et al., Computing trust from revision history. *In Proc. of the 2006 International Conference on Privacy, Security and Trust*, 2006.

[14] Huang, Z., Chung, W., Ong, T-H., Chen, H.: A graph-based recommender system for digital library. *In: Proc. Joint Conf. on Digital Libraries*, 2002, pp. 65-73.

[15] Jiwoon Jeon, et al., A framework to predict the quality of answers with non-textual features, *In Proc. of SIGIR’06*, 2006, pp.228-235.

[16] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 1999, pp.604-632.

[17] Khatri, V., Massey, A. and Montoya-Weiss, M., Online Services, Customer Characteristics and Usability Requirements, *Hawaii International Conference on System Sciences*, 41, 2008.

[18] Konstan, J. A., Riedl, J. and Shafer, J., E-Commerce Recommendation

Applications, *Data Mining and Knowledge Discovery*, 5 (1/2), 2001, pp.115-153.

[19] Kreulen, J., Maglio, P., Spohrer, J. and Srinivasan, S., Service Systems, Service Scientists, SSME, and Innovation. *Communications of the ACM*, 49(7), 2006, pp.81-85.

[20] Lawrence Page, et al.,. The PageRank citation ranking: Bringing order to the Web, 1999.

[21] Melville, P., Mooney, R.J. and Nagarajan, R., Content-Boosted Collaborative Filtering for Improved Recommendations, *Proc. 18th Nat'l Conf. Artificial Intelligence*, 2002.

[22] Michael J. Pazzani, "A Framework For Collaborative, Content-Based and Demographic Filtering," *Artificial Intelligence Review*, Vol. 13, No. 5-6 , 1999, pp.393-408.

[23] Torres R., McNee S. M., Abel M., Konstan J. A., and Riedl J.,Enhancing Digital Libraries with TechLens, *In Proc. of the 4th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)*, pp.228–236, 2004.

[24] Wiggins, A., Data-Driven Design: Using Web Analytics to Validate Heuristics, *Bulletin of the American Society for Information Science and Technology*, 33(5), 2007, pp.20-24.

[25] Xiaolan Zhu and Susan Gauch, Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web, *In Proc. of SIGIR'00*, 2000, pp.288-295.