國立臺灣大學文學院圖書資訊學研究所

碩士論文

Department of Library and Information Science

College of Liberal Arts

National Taiwan University

Master's Thesis

書目資料為中介的協力過濾式推薦於圖書館之應用

Application of Collaborative Filtering Recommendation Mediated by Bibliographic Data in Library

史修竹

Hsiu-Chu Shih

指導教授：陳光華　博士

Advisor: Kuang-Hua Chen, Ph.D.

中華民國 113 年　1 月

January, 2024

doi:10.6342/NTU202400232

BLANK PAGE

doi:10.6342/NTU202400232

國立臺灣大學碩士學位論文
口試委員會審定書

書目資料為中介的協力過濾式推薦於圖書館之應用

Application of Collaborative Filtering Recommender System Mediated by Bibliographic Data in Library

本論文係史修竹君（學號 R06126014）在國立臺灣大學圖書資訊學研究所完成之碩士學位論文，於民國一一三年一月十九日承下列考試委員審查通過及口試及格，特此證明

指導教授：　陳光華

口試委員：

唐牧群

楊東謀

蕭宗銘

系主任、所長　　（簽名）

doi:10.6342/NTU202400232

摘 要

推薦系統（Recommendation System）是圖書館協助使用者篩選資料的一種工具，但因多數資料缺乏使用紀錄，使得推薦系統面臨資料稀疏等問題，影響推薦效能。本研究以臺北市立圖書館的書目資料和 2013 至 2018 年的借閱紀錄為基礎，選擇「題名」、「作者」、「分類號」、「主題標目」、「出版社」等 5 種書目資料，提出以書目資料作為中介之混和推薦方法，檢視其預測效果。

本研究以 2013 年至 2017 年的借閱紀錄為訓練資料集、2018 年借閱紀錄為答案資料集，將訓練資料集的書籍借閱紀錄轉換為書目資料借閱紀錄，再依據書目資料的共被借閱率以 Message Passing Clustering（MPC）分群方法進行階層分群，建構 5 個單種書目資料的相似度矩陣。利用相似度矩陣將個人借閱紀錄轉為興趣特徵，得出預測結果。最後與答案資料集中同一人之借閱紀錄相比較，透過預測分數、預測結果餘弦相似度與前 N 名預測結果評估 5 種書目各自與整合的預測結果。

評估結果顯示「主題標目」與「出版社」的整體預測分數最佳，單次借閱均分大於 5.5，但是前 20 名預測結果中，「題名」、「作者」成功預測答案資料集的次數較多，分別有 2 萬 5,939 次與 6,841 次。考量真實情境下使用者的認知負荷，且「題名」、「作者」與「出版社」預測結果之排序成高度相關，建議選擇預測成功次數較多的「題名」取代「作者」，並將「出版社」的預測結果作為補充。此外，「主題標目」預測結果與其他 4 種書目資料呈低度相關，亦可提供使用者不同面向的新資訊。本研究針對書目資料用於混和推薦系統之結果可提供未來書籍推薦方法研究以及實務運用之參考。

關鍵詞：混合推薦模型、協力過濾、書目資料、書目資訊、階層分群

doi:10.6342/NTU202400232

Abstract

Recommendation system is a tool that libraries can assist users to filter data. Lack of usage records for most data leads to challenges, such as data sparsity. This study proposes a hybrid recommendation method based on the Taipei Public Library borrowing records from 2013 to 2018. Five types of bibliographic data: title, author, classification number, subject heading, and publisher, were selected as intermediaries to deal with data sparsity issues.

This study transformed borrowing records to bibliographic data borrowing records. Then, it hierarchically clustered these bibliographic data by co-borrowed rates with the Message Passing Clustering (MPC) method. The clustered results were utilized to construct 5 similarity matrices for each type of bibliographic data. Personal records were converted into interest features by these matrices and obtain predictions. Finally, the predictions of each type of bibliographic data and the integrated predictions were evaluated based on prediction score, cosine similarity, and the top N predictions.

The results show that "subject heading" and "publisher" have better prediction scores. Average scores of records are above 5.5. However, within the top 20 predictions, "title" and "author" are better, with 25,939 and 6,841 success predictions, respectively. Considering user's cognitive load and high correlation between "title," "author," and "publisher," this study suggest to use "title," which has more success predictions, rather than "author." "Publisher" can be used as a supplementary source. Additionally, since the predictions of "subject heading" are less correlated with the other four types of bibliographic data, it can provide users some different perspectives of new information. The results of this study on using bibliographic data in hybrid recommendation system can serve as a reference for future research and practical applications in book recommendation methods.

Keywords: Hybrid Recommendation Method, Collaborative Filtering, Bibliographic Data, Bibliographic Information, Hierarchical Clustering

doi:10.6342/NTU202400232

謝辭

本篇論文緣於一個對未來迷茫之人的執念，經過了長達 4 年的掙扎，終於勉強孵化而出。為什麼我會想做推薦方法的研究，直至今日也沒有一個明確的答案。

當初只是給予一心想投入公職，但又毫無準備的自己，建立避免失業之退路的藉口。「雖然我沒有一定要讀研究所，但既然要讀，我要做推薦方法的研究」，不論從知識、技能、人際資源上，選擇這個主題都沒有任何的優勢，就只是單純的一個願望，以及莫名其妙覺得可以做到的謎之自信。這樣的我，想必未來還會因此遇上不少困難、吃上不少苦頭吧。

就這樣，踏上了一條截然不同的學習之旅，直至終點。路上東彎西繞不在話下，上班之後更是幾乎難以推進，4 年工作期間投入在論文的時間恐怕不到 10%。然而，這一切還是完成了，受到許多人的幫助、鼓勵、建議與溫柔對待。最要感謝的莫過於指導教授陳光華老師，老師無比的耐心與精準的建議，是本篇論文能完成的最大助力；同時也要感謝黃建智學長不遺餘力地跟缺乏程式概念與技術的我討論修改方向，提出各種有效的建議，讓一切成為可能。

感謝計畫書與論文口試委員唐牧群老師、楊東謀老師、蕭宗銘老師提供的諸多建議與修改方向，讓本文的架構更加完善與豐富；感謝臺北市立圖書館願意提供豐富的借閱紀錄，讓我的研究可以順利執行；感謝我的同學朋友、家人們的支持與包容，敬豪、誌紘，這是一條獨自前行的道路，因為我的工作與個性影響，這條路注定更加孤獨，感謝你們對我的鼓勵與陪伴。

最後，本文能夠完成還要感謝另一個人的存在，意晴，因為你，我才能理解自己有多麼固執，理解後剩下的就是我自己的選擇。你為我展開了完全自由的選項，並在敦促我朝一個方向前進同時，包容我的偷懶、不積極、逃避與自責。這原本應該是一個惡性循環，但只有你能打斷這個連鎖反應，讓我重設自己的心靈，再次有動力開始向前。我曾經覺得「平安喜樂」是一個古老、俗套的祝福語，但在此刻與

doi:10.6342/NTU202400232

未來，我無比希望你能永遠平安、常保喜樂。未來，是一條沒有終點的路，希望我們能一起走向那些未來的未來。

將本篇論文獻給充滿執念的我，而我將獻上自己，讓我愛的人們與愛我的人們能展露笑顏。

doi:10.6342/NTU202400232

目次

doi:10.6342/NTU202400232

doi:10.6342/NTU202400232

圖次

doi:10.6342/NTU202400232

表次

doi:10.6342/NTU202400232

第一章、緒論

第一節、研究動機

圖書館典藏的資訊數量龐大，隨著人類的知識發展，不論是數位或實體館藏都持續增加，並且增加的速度越來越快。然而，人類使用圖書館館藏的成長速度並未以等比例增加，也就是說被有被使用的館藏占整體的比例持續降低；同時，圖書館的借閱紀錄經常會集中在少數館藏，而不會平均分配在所有的館藏。從圖書館服務管理的角度來看，未被使用的館藏固然可能是因為品質不佳，不值得使用者利用，但考量館藏數量與使用者寶貴的注意力與時間，更可能是因為使用者與圖書館之間互動效率低落，難以找到所有他需要的館藏。傳統上，圖書館站在被動的角度等待使用者利用搜尋工具查詢所需的館藏，因此在缺乏主動推薦工具的情況下，使用者常常無法有效、快速地得知圖書館內有哪些館藏符合他的需求（余明哲，2003；Liao, Hsu, Cheng, & Chen，2010）。

從使用者的角度來看，知識本就隨著人類社會文化與科技的發展持續分割、重組，整體知識架構越來越複雜，如果不是特定領域的專家，很難辨識、理解該領域中不同知識間的關係。此外，環境中充斥的大量資訊更讓使用者在面對不熟悉的領域時，無法有效地篩選與辨認資訊。一般人需要資訊的時候，多處於對知識架構不熟悉的情況，這就是唐牧群與吳宛青（2009）提到的，資訊使用者面臨的雙重困境。第一個困境是資訊超載，從大量資訊過濾出對使用者有用的部分是困難且使用者不願面對的任務；第二個困境則是使用者缺乏足夠的知識判斷資訊對自己的有用程度，事實上，使用者賴以判斷資訊價值的資訊，往往就是能滿足其資訊需求的資訊，如果使用者擁有判斷的能力，其資訊需求通常也就隨之消失。雖然唐牧群與吳宛青（2009）的研究僅將這種雙重困境侷限在使用文化產品的情境，但是這種困境

doi:10.6342/NTU202400232

同樣也會發生在其他情境，不論是學術研究、日常購物、生涯規劃、個人理財，都會因為資訊超載與知識不足而導致選擇、決策的困難。

使用者在缺乏領域知識與明確的篩選標準時，Simon（1955）認為使用者會試圖降低篩選與辨認資訊對自身價值的認知負荷，這種時間與認知能力有限的情況被他稱為有限理性（Bounded Rationality）。唐牧群與吳宛青（2009）同樣發現圖書館的使用者會傾向使用有較高機率觸發資訊偶遇的查詢途徑，進而避免產出檢索詞的認知負荷。Higgins（1999）、Dellarocus（2003）也認為使用者在面對過量資訊時，不會依照資訊內容去辨識其價值，而是更依靠資訊的二手資訊與書目資料來判斷資訊的有用程度，這也是真正會影響使用者的選書決策的因素。

面對使用者在尋找資訊遇到的雙重困境，從圖書館的立場不可能藉由減少提供的資訊來協助使用者，這樣不但與圖書館設立宗旨相違背，也會損害使用者取用資訊的權利，進而衍生資訊審查的問題。因此圖書館必須從其他角度解決使用者的問題，推薦系統（Recommendation System）就是其中一個解決方案。推薦系統是一種過濾資訊的系統，依照使用者過去的行為或是其他相關資訊，預測特定使用者的資訊需求，可以讓使用者免於資訊超載，並給予個人化的推薦資訊。推薦系統最基礎的機制是針對使用者對尚未接觸、使用的物件或資訊，預測使用者對該物件或資訊的評價（Adomavicius & Tuzhilin, 2005; Liao et al., 2010），其目的是希望協助使用者更有效率地篩選資訊。由於推薦系統是根據使用者行為或是資料主題進行推薦，使用者在瀏覽推薦清單時，也能從中了解該知識領域的概況，降低其發想檢索詞彙的認知負荷。

推薦系統依照其使用的資料分為三類方法等（Adomavicius & Tuzhilin, 2005; Su & Khoshgoftaar, 2009）：

1. 基於內容（Content-based）：主要利用物件的屬性進行推薦，推薦主題、屬性相似的物件。

doi:10.6342/NTU202400232

2. 協力過濾（Collaborative Filtering）：主要利用使用者的行為資料，根據使用者的行為相似性進行推薦，其目標是找出與個別使用者相似的其他使用者，並根據這些相似使用者的行為提供該使用者推薦結果。
3. 混合方法（Hybrid Method）：混合上述兩種方法，結合二者的優點，彌補兩個方法各自的缺點，如冷啟動與過度專業化。

不同推薦方法都有各自的優缺點以及適用情況，基於內容的推薦方法根據個別使用者曾使用的物件，以及物件屬性相似性給予推薦，但是物件屬性的區隔如果太詳細，可能造成推薦過度專業化，使推薦結果缺乏多樣性。此外，基於內容的推薦方法需要使用者累積一定數量的使用行為後才能產生較好的推薦結果，新使用者因為缺乏使用行為，基於內容的推薦系統也無法有效提供推薦。協力過濾的推薦方法根據使用者行為記錄進行推薦，不需要分析物件屬性，不會有推薦專業化的問題，但因為推薦完全依賴使用者的行為紀錄，同樣無法讓新使用者有效利用。此外，協力過濾推薦方法的運作也深受資料稀疏的限制，且對於未被使用的新物件缺乏有效推薦手段（余文哲，2003；Zhang，2016；Adomavicius & Tuzhilin, 2005；Mansur, Patel & Patel, 2017；Su & Khoshgoftaar, 2009）。

針對目前推薦系統面臨的限制，本研究希望利用唐牧群與吳宛青（2009）、謝宜瑾與唐牧群（2013）提出的選書決策影響因素，應用圖書館的書目資料與借閱紀錄，提出結合物件屬性與使用者行為紀錄的混合推薦方法，解決圖書館使用者面臨的資訊超載困境，降低在圖書館中篩選資訊時的認知負荷。期望藉由將書籍借閱紀錄轉化為「題名」、「作者」、「出版社」、「主題標目」、「分類號」等五種書目資料借閱紀錄，並根據共同被借閱的比例計算書目資料的相似度，最後根據個別使用者真實的借閱紀錄產出興趣特徵，提出個人化推薦清單。

doi:10.6342/NTU202400232

第二節、研究目的

本研究利用書籍的書目資料作為協力過濾推薦方法計算相似度的對象，採用此方法是因為圖書館使用的圖書分類法是一種架構嚴謹的資訊分類體系，但它同時也是明確而單一的知識架構，缺乏兼容跨領域主題、反映多元主題的彈性。每本書即使包含多個主題，仍然只能有一個分類號，使得依賴分類號進行推薦的基於內容推薦方法受到很大的限制，無法呈現書籍主題的多樣性。此外，人類的知識架構十分複雜，變化頻繁，很難找出一個穩定、具公信力的知識架構供基於內容的推薦系統參考。卜小蝶（2007）的研究也發現在使用者的借閱紀錄上，有許多非層級式的類號關聯規則出現，凸顯了分類法的知識架構與使用者的認知是有差異的。

因此本研究提出的混合推薦方法會結合基於內容推薦方法與基於物件（Item-based）的協力過濾推薦方法，使用五種影響選書決策的書目資料推薦書籍，期望藉由將書籍借閱紀錄轉換為五種書目資料借閱紀錄，一方面豐富可供計算的資料，緩解資料稀疏的問題；二方面透過借閱紀錄探索共同被借閱的書目資料間隱含的不同主題和關聯，通過由下而上的分群方法，反映真實的知識架構、呈現書籍多元主題；最後，由於不同書籍可能有相同的作者、出版社，使用書目資料計算相似度有助於縮小需計算的矩陣大小，並降低資料稀疏對推薦結果造成的影響。

此外，利用書目資料表示書籍間相似度的方法也可以讓新書籍在沒有借閱紀錄時，依據其書目資料被其他使用者借閱過情形，受到推薦系統的推薦。最後，在應用此推薦方法時，可事前建立書目資料相似度矩陣並紀錄書籍之間的相似度。提供即時個人化推薦結果時，不需重新計算整體的借閱紀錄，而是根據已計算好的書目資料相似度與目前的個人借閱紀錄找出推薦分數高的書籍，只要定期更新書目相似度，即可維持推薦結果的有效性。

綜合本節所述，本研究之研究問題有三個：

1. 利用「題名」、「作者」、「出版社」、「主題標目」、「分類號」等五種書目資料各自得出的書目資料相似度之推薦結果預測效果如何？

doi:10.6342/NTU202400232

2. 前項五種書目資料的推薦結果間相似程度如何？

3. 若將五種書目資料的推薦結果整合後，整體預測效果如何？

第三節、研究範圍與限制

本研究之範圍與限制如下：

1. 本研究使用的借閱資料為臺北市立圖書館 2013 年 1 月~2018 年 12 月之間的借閱紀錄，所計算出的書目資料相似度受限於時間範圍與特定圖書館使用者社群特徵，可能無法直接應用於其他圖書館。
2. 本研究僅使用五種既存的書目資料，但唐牧群與吳宛青（2009）、謝宜瑾與唐牧群（2013）的研究另提及的「名人推薦」、「封面設計」等選書決策影響因素。然而現有書目資料庫 F 未有上述資訊，且研究者人力與時間有限，前述兩項影響因素並無法應用於本研究。
3. 臺北市立圖書館提供的 6,394 萬 5,077 筆借閱紀錄與 64 萬 8,596 種書目資料有程度不等的闕漏與亂碼，如借閱紀錄遺失 5,386 筆資料，占所有借閱紀錄的 0.008%，主題標目缺漏 18 萬多筆，占所有書目資料種數的 28.68%；作者闕漏 11,368 筆，占 1.75%；分類號缺漏 13 萬多筆，占 20.83%。

doi:10.6342/NTU202400232

doi:10.6342/NTU202400232

第二章、文獻回顧

本章將簡單介紹推薦系統的概念，並在第二節與第三節分別回顧基於內容與協力過濾兩種主要推薦方法的應用方式與優缺點。其次，因本研究使用階層分群方法計算書目相似度，因此在第四節將進一步介紹不同類型的分群方法。第五節則回顧整合多種因素進行協力過濾推薦的相關研究。本章最後一節會回顧以圖書館為場域的推薦系統相關研究。

第一節、推薦系統

推薦系統是一種通過分析系統中的物件資訊與使用者行為，來幫助使用者處理資訊超載並提供個人化的建議、服務、內容的系統，它可以協助人們處理日常生活中那些需要依賴他人提供建議的事務（Adomavicius & Tuzhilin, 2005; Park, Kim, Choi, & Kim, 2012; Su & Khoshgoftaar, 2009）。

推薦系統依照提供建議的方式分為兩類：

- 基於內容推薦（Content-based Recommendations）：依照使用者過去的使用行為或是物件本身的訊息，推薦與之相似的物件。
- 協力過濾推薦（Collaborative Filtering Recommendations）：依照使用者過去的行為推薦和他行為相似的其他使用者。

這兩種方法各有其優缺點，因此有許多研究提出不同的混合方法（Hybrid Method），希望能結合上述兩種方法，以求截長補短之效（Adomavicius & Tuzhilin, 2005; Su & Khoshgoftaar, 2009）。

基於內容的推薦方法（Content-based Recommendations）源於資訊檢索領域對內容相似性的研究，其中最重要的是對內容關鍵字的萃取技術。基於內容的推薦方法主要是從物件內容中萃取關鍵字，或是將物件的屬性作為關鍵字，並根據使用者

doi:10.6342/NTU202400232

過去的使用紀錄，如點擊、購買、評分等行為來建構使用者的興趣特徵，最後根據使用者的興趣特徵與物件內容關鍵字的相關性來計算推薦分數。除了這種基於關鍵字相關性的方法外，也有一部分基於內容的推薦方法會建構模型，以預測未受過使用者評分的物件之推薦分數。基於內容的推薦方法需要面對的問題包含：非文字資料的關鍵字萃取不易、關鍵字的文字模糊性、推薦過度專業化與新使用者的冷啟動問題，上述問題都可能影響基於內容推薦方法的推薦準確度（Adomavicius & Tuzhilin, 2005; Park et al., 2012）。

另一方面，協力過濾推薦（Collaborative Filtering Recommendations）是根據其他使用者的使用行為來預測當前使用者對不同物件的滿足程度，藉由找出與使用者相似的另一群使用者，並根據這群相似使用者的使用行為來提供推薦結果，而這種相似可以是顯性指標，如滿意度評分；也可以是隱性指標，比如網頁點擊、購買、停留時間等操作行為（Su & Khoshgoftaar, 2009）。協力過濾推薦方法的優點是不需處理關鍵字萃取的正確性和尺度問題，僅依照使用者特徵或行為的相似性提供推薦，因此可以推薦使用者不曾使用的物件類別，增加推薦多樣性。但它仍需要面對許多問題，包含：新使用者的推薦失準、新資料缺乏推薦方法、使用者與需推薦的物件數量龐大，相對而言使用者的行為資料十分稀疏，影響推薦結果的表現。其他還有諸如字詞模糊性、使用者隱私等問題（Adomavicius & Tuzhilin, 2005）。

混合方法（Hybrid Method）結合基於內容的方法與協力過濾的方法，以提高推薦準確度並解決兩者各自的問題。混合的方法有許多種，可以是將兩種方法的推薦結果整合；也可以是將利用物件內容取代使用者的行為，如此便可以基於使用紀錄中所有物件的內容關鍵字尋找相似的使用者群；或是反過來根據使用者行為，將內容關鍵字做矩陣分解（Adomavicius & Tuzhilin, 2005）。Su and Khoshgoftaar（2009）指出有許多混合方法可以解決資料稀疏的問題，比如 Content-boosted 協力過濾方法（Melville, Mooney, & Nagarajan, 2002），將電影的屬性，如標題、演員、導演等等內容屬性，利用受試者的評分將這些屬性透過貝式文字分類器（Bayesian Text

doi:10.6342/NTU202400232

Classifier）分為六級，從而豐富可分析的資料，並且因為是針對電影的屬性作分析，解決了新電影無法被推薦的問題。Mooney and Roy（2000）則是抽取書籍的屬性與內文的頻繁關鍵字作為計算推薦分數的對象，根據使用者對書籍的評分去計算每個屬性、關鍵字對該使用者的推薦分數，最後得出個人化的推薦結果。

下兩節將依序介紹基於內容與協力過濾兩種推薦方法的主要應用方式與限制。

第二節、基於內容推薦方法

基於內容的推薦方法以物件相似性為推薦依據，其目的是找出與使用紀錄裡使用者「滿意」的物件最「相似」的其他物件。依照推薦物件類型、欲解決問題的不同，這裡的「滿意」與「相似」可以使用不同資料和不同計算方式。

目前基於內容的推薦方法多應用在文字資訊的推薦，通過從物件中抽取關鍵字作為其特徵值，基於內容的推薦方法可以藉由比較物件之間的特徵值計算兩者的相似度。常用的相似度計算方式為餘弦相似度或 Pearson 相關性，現在也有許多研究使用機率模型、線性分類器（Linear Classifiers）等方法，根據使用者的使用行為來預測他對其他未使用物件的態度，或是根據使用者對推薦結果的態度、行為，利用相關回饋的技術改善改善推薦結果。

基於內容的推薦方法可以有效的提供使用者個人化的推薦結果，即使是新物件也能即時推薦，但是此方法在應用上會遇到一些困難（Adomavicius & Tuzhilin, 2005; Park et al., 2012; Pazzani & Billsus, 2007）：

1. 非文字資料的關鍵字萃取不易：內容自動分析功能目前僅在文字資料上有比較好的表現，應用於非文字類的資料，如圖片、音樂等表現較差。
2. 關鍵字的模糊性：單純以關鍵字來代表物件的特徵，遇到關鍵字有同形異義或異形同義的狀況時，就會出現同一組關鍵字代表不同主題、內容物件的情形，影響推薦結果的準確度與品質。

doi:10.6342/NTU202400232

3. 推薦結果過度專業化：基於內容推薦方法在決定關鍵字的尺度時，若是關鍵字分割的過於粗略，會讓推薦結果中出現過多不相關的物件，但若是分割的過於詳細，也會導致推薦結果侷限在很小的主題範圍，缺乏多樣性，失去推薦的效果。
4. 冷啟動：由於基於內容的推薦方法只會利用使用者自身行為紀錄提供推薦結果，因此新使用者在缺乏使用紀錄的情況下，系統無法完整辨識使用者的興趣特徵，無法在一開始就獲得準確的推薦結果。

為了解決基於內容推薦方法的限制，Pazzani and Billsus（2007）提到在某些情況下，如推薦餐廳、電影、商店，把使用者對物件的評論內容作為物件特徵的附加資訊，豐富推薦方法可用的資料。Lops, Jannach, Musto, Bogers, & Koolen（2019）認為近年來對基於內容推薦方法的改進主要可分為資料與演算法兩個方面。在資料方面可利用 Linked Open Data 或是使用者產生的標籤、評論等資料來豐富物件的特徵，針對圖形或多媒體物件，則可以利用類神經網絡等深度學習技術來判斷這類物件的內容特徵，彌補此類物件過去只能使用文字詮釋資料推薦的限制。因為用於分析的資料類型增加，異質性資訊網路（Heterogeneous Information Network）的技術被用於呈現不同資料關係，進而讓基於內容的推薦方法能辨識並處理物件間不同的關係。在演算法方面，深度學習的應用日漸廣泛，使用基於詮釋路徑（Meta-Path）方法分析異質性資訊網路，以及利用嵌入（Embeddings）方式編碼物件的特徵關係等方法被用於基於內容的推薦方法，使得推薦系統能夠分析使用者行為的時間與順序，利用不同類型的資料改善推薦結果。

第三節、協力過濾推薦方法

協力過濾的推薦方法與基於內容推薦方法最大的差異在於，協力過濾推薦方法是根據其他使用者提供的行為資料來預測一個使用者的行為，此方法評估的是使用者相似性，而基於內容的推薦方法評估的則是物件相似性。依照協力過濾推薦

doi:10.6342/NTU202400232

方法預測的方式可分為基於記憶（Memory-based）與基於模型（Model-based）兩種方法。前者基於使用者過去的使用行為，尋找和他行為相似的其他使用者的使用行為，進而預測使用者對未接觸的其他物件之使用行為；後者則傾向運用所有使用者的使用行為建立一個通用的模式預測使用者的行為（Adomavicius & Tuzhilin, 2005）。

基於記憶（Memory-based）的協力過濾通常使用餘弦相似性與 Pearson 相似性代表兩個使用者之間的相似程度，此方法因易於應用且效果顯著而被廣泛使用（Su & Khoshgoftaar, 2009）。依照計算相似性的對象，基於記憶的方法又可以分為基於使用者（User-based）與基於物件（Item-based）兩種。基於使用者的協力過濾方法是藉由計算使用者行為、評分的相似性，來找出相似的使用者群進行推薦；基於物件的方法則是根據全體使用者的行為資料，計算物件之間的相似性，從而找出與特定物件相似的其他物件（Mansur, Patel, & Patel, 2017）。

基於模型（Model-based）的協力過濾並非是計算使用者之間的相似程度，它是從所有數據中分析整體使用者的行為模式，並以機率的方式預測個別使用者的行為，此方法的優勢是在資料稀疏時會比基於記憶的推薦方法來得可信（Su & Khoshgoftaar, 2009）。

協力過濾推薦方法也有它的限制，這些會影響推薦效果的問題包含（Adomavicius & Tuzhilin, 2005; Su & Khoshgoftaar, 2009）：

1. 新物件冷啟動問題

   協力過濾推薦方法是依據使用者的行為紀錄，因此新進入系統的物件在缺乏使用紀錄的情況下，無法被推薦。

2. 新使用者冷啟動問題

   與基於內容的推薦方法相同，在缺乏使用者行為紀錄的情況下，協力過濾推薦方法無法尋找相似的使用者。

doi:10.6342/NTU202400232

3. 資料稀疏

通常來說，推薦系統擁有的使用者行為紀錄遠少於它所需的數量，因為多數推薦系統需要推薦的物件種類很多，使用者很少會大量使用這些物件。在推薦比如電影、書籍、餐廳、服裝等物件時，多數的物件可能只有很少人使用過，這會影響協力過濾推薦方法的表現。

4. 資料規模與運算時間

如上點所述，隨著使用者與物件的數量不斷增加，推薦系統所需要處理的資料越來越大，因此需要越來越久的時間計算，最終超出使用者可以忍受的範圍。

5. 字詞模糊性

與基於內容推薦方法面臨問題相同，協力過濾推薦方法也會因為相同物件有不同名稱，或是名稱在詞綴、修飾語上出現不同，就會被判定是不同的物件，導致物件的使用紀錄無法有效整合。

6. 先令攻擊（Shilling Attacks）

先令攻擊指的是使用者或物件銷售者惡意的創造特定的行為紀錄，其目的可能是希望提高物件的能見度，或是影響競爭物件的能見度，但此類行為對協力過濾推薦方法卻可能造成很大的影響。

7. 灰羊問題（Gray sheep）

「灰羊」是部分偏好與主流使用者不同的使用者，對他們來說協力過濾推薦方法的表現不佳，因為針對某些偏好的推薦完全無效，但針對另外一些偏好的推薦卻有效。而偏好與整體使用者完全不同的使用者被稱為「黑羊」，因為推薦結果完全無用，因此這類人不會成為使用者，反而不是協力過濾推薦方法要處理的問題。

doi:10.6342/NTU202400232

上述問題有部分可以透過混合兩種推薦方法來解決。比如新物件的冷啟動問題可以利用基於內容推薦方法中的物件相似度來彌補，先令攻擊與灰羊問題也都能藉由調整基於內容推薦方法在混合方法中的比重獲得緩解。

另一方面，新使用者、資料稀疏、資料規模與運算時間、字詞模糊性等問題是兩種方法都會遇到的問題。Adomavicius and Tuzhilin（2005）提到有些研究根據物件的流行度、使用者的個人特徵來解決新使用者的冷啟動問題。

資料稀疏與資料規模的問題都肇因於推薦系統中物件與使用者的數量很多，但可供運用的行為紀錄卻不足。這不只導致推薦系統無法有效計算物件、使用者的相似度，也會耗費很多時間在計算數值極低或是不重要的相似度。Adomavicius and Tuzhilin（2005）指出面對資料稀疏問題時，可以根據使用者背景資料，比如利用人口統計、物件使用的順序或時間長短、頁面瀏覽的順序等資料來豐富使用者的行為資料；或是從更多角度去看待使用者與資料，考慮更多種的情境因素與物件特徵的分類，比如使用資料時地點、時間、購買食物的價格、風味、服務態度等等，利用不同背景、情境因素與物件特徵來增加推薦系統可以運用的資料，優化推薦結果。

另一方面，使用者的行為紀錄中可能存在相似度極高，或是影響力極低的資料，此時藉由主成分分析、奇異值分解（SVD）、矩陣分解等降維技術可以從行為紀錄中找出重要的部分，進一步合併相似的資料或挑選出具代表性的資料，減少推薦系統需要計算的資料量，也能解決資料稀疏的問題（Adomavicius & Tuzhilin, 2005）。Su and Khoshgoftaar （2009）也認為除了利用矩陣分解的降維技術縮小「使用者—物件」或「物件—物件」相關矩陣，減少矩陣中的空值外。使用關聯規則（Asoociation Rules）方法找出出現頻率較高的物件配對，或是利用分群（Clustering ）技術把使用者分群，針對群內的使用者推薦，都有助於緩解資料稀疏的影響。

本研究的協力過濾推薦方法主要使用分群技術計算書籍之間的相似度，因此下節本文將回顧分群技術的不同方法。

doi:10.6342/NTU202400232

第四節、分群方法

分群方法是基於模型的協力過濾推薦方法常用的技術之一，藉由將相似的物件或是使用者分為同一群，盡可能減少群內物件或使用者的差異，並且提高不同分群之間的差異，利用分群的結果來呈現使用者之間的相似程度。資料或分群之間的相似性，常會使用 Pearson 相關係數、歐幾里德距離（Euclidean distance）、馬哈蘭距離（Mahalanobis distance）等指標。評估分群結果則主要看分群間的距離，距離計算方式通常有四類，可依照推薦需求決定（Su & Khoshgoftaar, 2009; Han, Kamber, & Pei, 2010）：

1. 最小距離：分群之間最接近的兩資料點的距離，此類方法被稱為最近鄰居分群法。
2. 最大距離：分群之間距離最遠的兩資料點之間的距離，此類方法被稱為最遠鄰居分群法。
3. 平均值距離：兩分群各自的重心之間的距離。
4. 平均距離：兩分群間所有點之間的距離平均。

分群方法在形成分群時，共有 Partitioning、Density-based、Hierarchical、Grid-based、Model-based 等五種方法，不同方法的分群結果會有不同的特性。

Partitioning 分群方法是藉由最大化分群之間的距離達到分群的目標，這些分群之間通常是互斥，但也可以利用模糊分割技術產生非互斥的分群。Partitioning 方法的基本方法是設定一個分群數量 K，然後試圖讓 K 個分群之間的距離最大化。最常用的方法為 K-means 方法與 K-medoids 方法，這兩個方法因為相對高效率與易於應用的特性而被廣為使用。這兩者的主要差異在於分群代表點的選擇方式不同，K-means 分群方法會先隨機選擇 K 個點作為分群中心，然後將所有資料各自分配給距離最近的分群中心。根據第一次分群結果重新計算出分群的重心後，把重心視為該分群新的中心點後重新分配資料，迭代執行上述步驟直到分群的結果穩定。K-medoids 分群方法同樣是隨機選擇 K 個資料作為分群中心，然後將所有資料

doi:10.6342/NTU202400232

分配給最近的分群中心，但它更新分群中心的方式是隨機將一個非分群中心的資料點取代另一個分群中心，再次分配資料後計算整體分群的結果，如果分群結果更好則接受這次分群。此方法相較 K-means 方法更不易受離群值影響，但是當資料量與分群數都很大時，運算複雜度會相對提高很多，需要更長時間運算。

Density-based 分群方法選擇分群中心的依據是某資料周圍的資料數量，也就是特定資料點的一定距離內資料的密度。如果一個資料周圍有不少於特定數量的其他資料時，該資料就會被認為是分群中心，而其他在特定距離內的資料則屬於該分群。當某些資料既是分群中心，也是其他分群的資料時，Density-based 的方法會將這些分群合併，從而產生不規則形狀的分群結果。此方法的優點是能找出非球狀的分群，而且不需要指定分群數量，也能避免離群值對分群結果的影響，但是如何決定距離與密度的閾值是此種方法的難題。

Hierarchical 分群方法是依照特定的標準將資料整體逐步階層化的分割或聚合，將資料組織成階層結構。Hierarchical 分群方法可以分為聚合式與分裂式，前者是以由下而上的方式聚合資料；後者則是以由上而下的方法將資料分割。聚合式比分裂式更常被使用，因為分裂式分群方法為了節省時間，在分群之後不會重複考慮先前的分群決策，因此當初期的分群出現錯誤時，將會嚴重影響分群品質。事實上，所有分群方法都會切斷不同群資料之間的關係，因此都可能出現錯誤的分群決策，影響分群品質。階層分群方法可以透過整合其他分群方法來改善這些問題，比如 BIRCH（Balances Iterative Reducing and Clustering using Hierarchies）就是先根據資料產生分群特徵樹（Cluster Feature Tree）來描述資料的多層次特徵，再運用其他的分群方法針對每個葉節點分群。CURE（Clustering using Representatives）是另一種階層分群方法，它先使用 Partitioning 的分群方法，然後找出每個分群的重心作為代表點，再針對這些代表點進行分群，此方法有助於減少離群值對分群結果的影響，也能找出非球形的分群。

doi:10.6342/NTU202400232

Grid-based 分群方法的特色是利用多解析度(Multiresolution)的網格資料結構，將整體資料量化後投射到網格空間中，並切分成數量有限的小方格進行分群。此方法優點是運算時間與資料本身的數量無關，而是受研究者設定的方格數量影響較大。STING（Statistical Information Grid）分群方法是一個典型的 Gird-based 分群方法，它使用多個的統計數據做閾值，建構一個階層化的網格空間，不同階層中使用不同的統計數據作為閾值。當資料被輸入時，就依照不同的閾值，在每一層中被分入不同的方格，直到無法再被分入下一層方格中為止。此方法的準確度與研究者一開始設定閾值時的粒度息息相關，而且上一層中被分在不同群的資料之間的其他關係無法在下一層中被察覺，這使分群結果的形狀不是呈現水平就是呈現垂直，影響分群的品質與準確度。CLIQUE（Clustering in Quest）分群方法整合了 Density-based 與 Gird-based 的方法，從整體的資料空間中尋找密度較高的資料集合，並將相互接觸的高密度資料集合都視為同一個分群。此方法的好處是對於資訊輸入順序並不敏感，且不須假設資料的分布狀態。但缺點是準確度較低。

Model-based 分群方法是將嘗試將資料套入數學模型中進行預測，這類的方法常會假設資料的來自隨機分布的狀態。比如 COBWEB 這種分群方法會先針對一群為分群的資料，根據其屬性或是特色將資料分群，然後再根據機率與條件機率建構階層的分群結果。類神經網絡也是 Model-based 的分群方法，其步驟與 K-means 分群方法類似，都是將資料分配給距離最近的分群中心點，但它並不會一次分配所有資料後調整分群中心點，而是逐步的輸入資料，當分配了資料 A 給分群中心 X 之後，就會將分群中心調整到 A 與 X 之間，才分配下一個資料。

本節中介紹的五種分群方法是以不同的方式評估資料的相似程度，實際上因應使用的資料、推薦的對象、研究者欲解決之問題的不同，需要結合不同的分群方法，本節中所介紹實例也多為如此。而本研究使用 Geng, Deng and Ali（2008）提出的 Message Passing Clustering （MPC）分群方法則屬於聚合式的 Hierarchical 分群方法，將分別將五種書目資料分群，用五個因素代表書籍的相似度，所以接下來，

doi:10.6342/NTU202400232

本文會回顧一些同樣使用多個因素進行協力過濾推薦的相關研究，推薦對象包含餐廳、電影、書籍、會議規劃等等。

第五節、多因素協力過濾推薦方法

通常推薦系統僅處理使用者與推薦的物品兩個因素，但實際上使用者的需求與決策會受到很多因素的影響，比如同一專業的學生知識學習歷程有一定順序，不同的資訊需求也會依序產生，某些旅遊行程適合與不同旅伴一起參與(Adomavicius & Tuzhilin, 2001; Zhang, 2016）。因此，有不少研究試著應用不同的因素，來改善推薦的結果，比如 Melville, Mooney and Nagarajan（2002）利用電影的劇本、演員等屬性豐富協力過濾推薦系統中電影的使用資料，從而為使用者尋找更多的相似使用者。Adomavicius and Tuzhilin（2001）將電影類型（如喜劇、動作片）作為一種因素，也將使用者觀賞的時間點視為另一個因素，從使用者、電影類型、觀賞時間三個維度去判斷使用者的相似度。Adomavicius, Sankaranarayanan, Sen and Tuzhilin （2005）、Rahman（2013）進一步用使用者、電影、時間、地點、同行者五個維度來計算使用者的相似度。Mooney and Roy（2000）針對書籍推薦，利用作者、題名、主題詞、摘要、出版評論、消費者評論等因素結合使用者評分，建立使用者的偏好書籍特徵檔，用以推薦其他書籍。楊亨利與張文祥（2008）提出的推薦系統主要針對商務會議的安排需求，將商務會議的條件分為：人、事、時、地、物五個維度，分別依照使用者過去的選擇建立分類樹，再根據使用者的身分進行分群之後，根據同群其他使用者的選擇給予推薦。

第六節、圖書館的推薦系統相關研究

目前針對圖書館的推薦方法研究大多是使用協力過濾方法，但還是有使用基於內容的推薦方法，比如從使用者的查詢指令與書籍內文、書目資料來抽取關鍵字，

doi:10.6342/NTU202400232

作為基於內容推薦的書籍特徵（Lopes et al., 2008）或是計算使用者與書籍的相似度（Lai & Zeng, 2013）。

在圖書館的環境中，使用協力過濾方法的推薦系統研究最常使用關聯規則來計算書籍間的相似度，關聯規則的目的是找出使用者的行為中出現次數多且特殊的模式，作為推薦的依據，此方法找出的規則可以是「資料—資料」、「資料—使用者」或「使用者—使用者」，在尋找規則時主要的指標有（余明哲，2003；羅子文，2007：Han, Kamber, & Pei, 2012）：

- 信心水準：A 書被借閱後，B 書也被借閱的機率
- 支援度：A 書和 B 書同時被借閱的次數
- 重要度：A 書和 B 書同時被借閱的次數 / 沒有借 A 書的 B 書借閱次數

關聯規則推薦研究的典型如卜小蝶（2002）、呂家賢（2005）、戴玉旻（2002）等利用借閱紀錄尋找不同分類號之間的關聯規則，此外關聯規則也能利用時間間隔來找出借閱的順序，把無方向性的關聯規則變成有向的規則（吳安琪，2001），或是依照兩本書之間借閱的間隔加權關聯規則的強度（蔡秀滿與莊宛螢，2007）。余明哲（2003）則是把關聯規則結合基於內容推薦方法的特徵，除了借閱紀錄上共同出現的次數外，還加入不同分類號在分類法上與使用者興趣特徵檔的距離作為推薦分數的計算依據。

另一個常用到的技術是分群（Clustering），也是基於模型的推薦方法之一。此方法的好處是利用人口統計資料或是使用行為將資料或使用者分群，從而縮減矩陣的大小。在合併同樣背景的資料後，也能降低資料稀疏的影響，並減少運算所需的時間。陳弘霖（2010）、邱宏彬、湯鎰聰與陳揮明（2005）、鄭玉玲（2003）都是根據使用者的背景資料作為分群的指標。Liao 等人（2010）是依照 DDC 分類號來將書籍分群，從不同群的書目資料抽取獨特的關鍵字代表該分群，用來呈現讀者的興趣特徵檔，依照讀者的興趣特徵檔的相似度來替使用者尋找相似的其他使用者。另外的分群方法如 Rajagopal and Kwan（2012）採用 K-means 分群方法分群書

doi:10.6342/NTU202400232

籍的主題標目（Subject heading）。Sirikayon, Thusaranon and Pongtawevirat（2018）則用經過矩陣分解的借閱紀錄計算使用者的相似度，並以前 K 個最相似的使用者作為使用者的推薦依據。

從過去的文獻我們可以總結推薦系統最主要的限制在於資料，因為資料使用紀錄的稀少讓推薦系統無法觀察到足夠的使用者行為來建立相關主題的連結，同時推薦系統對使用紀錄的依賴也讓處理新的使用者與資料成為一個問題。過去的研究藉由人口資料、資料屬性與情境資料來擴展使用者間的關係，或是使用主成分分析、奇異值分解、顯著詞分析等技術合併特徵相似的資料，試圖去解決這些問題。

本研究希望應用唐牧群與吳宛青（2009）、謝宜瑾與唐牧群（2013）提出的選書決策影響因素，發展一個可應用於圖書館環境的推薦方法。以基於物件的協力過濾方法為基礎，參考 Melville, Mooney, and Nagarajan（2002）、Mooney and Roy（2000）等人的方法，將一筆書籍借閱紀錄轉換為多筆書目資料借閱紀錄，不但能豐富共被借閱關係，增加可運用的資料；也因為書籍可能有相同書目資料，書目資料數量將會小於等於書籍數量，可以縮減書籍相似度矩陣的大小，達到降低資料稀疏的影響與減少運算負荷的效果。而書目資料的共被借閱矩陣將利用 Geng, Deng and Ali（2008）提出的 Message Passing Clustering （MPC）分群方法進一步分群，建立書目資料的相似度矩陣。此方式可以不用即時運算推薦分數，只需要針對個別使用者的使用紀錄，利用預先計算好的相似度矩陣找尋最相似的書目資料即可，針對圖書館新採購的書籍也能根據書目資料進行推薦。

doi:10.6342/NTU202400232

doi:10.6342/NTU202400232

第三章、研究方法

本研究以臺北市立圖書館為研究場域，利用唐牧群與吳宛青（2009）、謝宜瑾與唐牧群（2013）提出的影響讀者選書決策之作者、題名、主題、出版者等四類因素，提出融合基於內容的推薦方法與基於物件的階層分群協力過濾推薦方法的混合推薦方法。其中「主題」因素在書目資料中可對應到主題標目與分類號兩種不同的書目資料，因此本研究共採用五種書目資料，分別為作者、分類號、主題標目、題名、出版社。

本研究首先會將借閱紀錄分為「訓練資料集」與「答案資料集」，將訓練資料集的書籍借閱紀錄轉化為書目資料的借閱紀錄來取得更多可用於計算的資料後，以書目資料的共被借閱率來進行階層分群，計算出每種書目各自的相似度矩陣。然後，找出訓練資料集與答案資料集中都有借閱紀錄的使用者，根據其訓練資料集中的借閱紀錄建構個人化興趣特徵，並與答案資料集中的真實借閱紀錄相比較，檢視訓練資料集產出的五種書目資料相似度預測效益。

本章第一節將說明所使用的資料概況，第二節說明研究流程，包含推薦方法的建構與推薦結果的評估。

第一節、研究對象

本研究使用之資料有兩種資料，分別是臺北市立圖書館館藏目錄的書目資料，以及 2013 年至 2018 年間的借閱紀錄，書目資料抽取的時間點為 2019 年 3 月 25 日，共 64 萬 8,596 種書目資料。該館使用的是 MARC 21 機讀編目格式，本研究使用的五種書目資料分別抽取自下列 MARC 21 欄位：

1. 作者：100、110、111、245
2. 題名：245

doi:10.6342/NTU202400232

3. 分類號：084

4. 主題標目：650、654

5. 出版社：260

第二個資料集為借閱紀錄，包含臺北市立圖書館 2013 年 1 月 1 日至 2018 年 12 月 31 日間共 6 年份的借閱紀錄，總計有 6,393 萬 9,691 筆。借閱紀錄會再被區分為訓練資料集與答案資料集，訓練資料集主要用於計算書目資料相似度，包含 2013 年 1 月 1 日至 2017 年 12 月 31 日間 5 年份的借閱紀錄，總共 5,389 萬 6,988 筆；答案資料集則用於驗證推薦結果，包含 2018 年 1 月 1 日至 12 月 31 日間 1 年份的借閱紀錄，共有 1,004 萬 2,703 筆。

第二節、研究流程

本研究的流程分為五個階段：

一、資料清理

二、轉換書目資料借閱紀錄

三、建構書目資料相似度矩陣

四、推薦結果與評估

五、評估整合推薦的效益與書目資料相關性

一、資料清理

資料清理階段是將書目資料與借閱紀錄預先處理，刪除無效的書目資料並減少書目資料分歧，同時刪除無效的借閱紀錄。

1. 書目資料清理

臺北市立圖書館的書目資料編號稱作 CatKey，同一種書的不同複本（Copy）、集數（Volume）可能有不同的條碼號，但 Catkey 都會相同。因此本研究以臺北市立圖書館的 Catkey 作為唯一編碼（Unique Code），分別擷取本研究所需的題名、

doi:10.6342/NTU202400232

作者、出版社、分類號與主題標目等五種書目資料，建立書目資料對應表，用以將書籍借閱紀錄轉換為書目資料借閱紀錄。CatKey 與書目資料對應表範例如表一。

表一、書目資料對應表範例

	MARC 欄位		
CatKey	100	245	……
1	彭百顯, 1949-	理想的社會 : 福利與安全 / 彭百顯著	
17	-	小蟀哥愛唱歌/ 布萊爾利（Jane Brierley）原著 ; 伍爾夫（Tony Wolf）繪圖	
96	-	班雅明作品選: 單行道.柏林童年/ 瓦爾特.班雅明（Walter Benjamin）著 ; 李士勛, 徐小青譯	

因一本書籍的書目資料中包含多位作者或多個主題標目的狀況相當常見，所以除了分類號通常每個 CatKey 只有 1 個外，題名、作者、主題標目與出版社都可能存在 1 個 CatKey 對應多個書目資料的情形，這些書目資料可能代表書籍不同面向，滿足不同使用者的需求，本研究基於豐富借閱紀錄的目的皆予以保留。本研究使用的書目資料類型分述如下：

(1) 作者：全部保留。

(2) 題名：主要保留書籍之正題名與副題名。並列題名多為翻譯作品原名或中文譯名，為方便研究者檢視，「中文正題名+外文並列題名」與「外文正題名+中文並列題名」兩種情況會優先保留中文題名；「外文正題名+外文並列題名」的情況則保留英文題名，以利後續辨識與分析。

(3) 分類號：保留中國圖書分類法的完整分類號為原則，但刪除用以表示資料類型的標記，如「R」、「AC」，前者為參考資料，後者為錄音帶。非中

doi:10.6342/NTU202400232

國圖書分類法的資料亦不予保留，避免出現不同分類法間，同一分類號代表意義卻不同的情況。

(4) 主題標目：全部保留。

(5) 出版社：保留具出版事實之出版者，刪除對內容形塑無貢獻之「印刷」、「經銷」、「代理」等團體。為豐富可分析資料，針對「未言明出版職責」者，因多無其他並列出版者，因此認定其為出版者；針對「發行者」之認定，因書目資料中屢見出版者與發行者角色之混淆，在 648,596 種書目中，有 62,841 個 CatKey 有發行者，但其中只有 52,184 個 CatKey 同時並列發行者與出版者，剩餘 10,657 個 CatKey 則僅有發行者，因此將發行者予以保留，以免損失大量可分析書目。

為了避免同一個目標有不同的著錄形式，針對書目資料的文字內容則依表二的書目資料轉換原則處理，減少因人工編目造成的著錄錯誤，或因不同時期編目原則不一致造成的歧異。經整理後的書目資料對應表範例如表三，CatKey 與經轉換的五種書目資料的數量摘要如表四。

表二、書目資料轉換原則

項次	轉換原則	轉換前	轉換後
1	個人/團體名稱保留原文	布萊爾利(Jane Brierley)著	janebrierley
2	大寫字母轉為小寫		
3	刪除空格與縮寫點		
4	刪除貢獻方式說明		

doi:10.6342/NTU202400232

表三、資料清理後書目資料對應表（以作者為例）

CatKey			……
1	彭百顯		
17	janebrierley	tonywolf	
96	walterbenjamin	李士勛	徐小青

表四、CatKey 與五種書目資料數量表

類型	CatKey	作者	題名	分類號	主題標目	出版社
種數	648,596	342,375	484,094	21,357	27,829	30,346

2. 借閱紀錄清理

本研究的借閱紀錄共有兩個資料集，訓練資料集與答案資料集，前者為 2013~2017 年的借閱紀錄，共有 5,389 萬 6,988 筆；後者為 2018 年的借閱紀錄，共有 1,004 萬 2,703 筆。為了計算書目資料共被借閱率，本研究將訓練資料集中僅出現 1 次的使用者予以刪除，因為這些人的借閱紀錄只有 1 筆，沒辦法產生共被借閱次數。訓練資料集共刪除 3 萬 9,591 筆借閱紀錄，佔訓練資料集的 0.07%；答案資料集共刪除 8,905 筆，佔答案資料集的 0.09%，佔比都相當低。原始借閱紀錄筆數與清理過後的「多次讀者」借閱紀錄筆數如表五。

表五、借閱紀錄清理情形

資料集	原始借閱紀錄數	計算共被借閱率所用紀錄數
訓練資料集	53,896,988	53,857,397
答案資料集	10,042,703	10,033,798

doi:10.6342/NTU202400232

二、轉換書目資料借閱紀錄

為了計算書目資料的共被借閱率，本研究利用步驟 1 所建置的表三（資料清理後書目資料對應表）將「使用者→CatKey」的借閱紀錄（如表六）轉換為「使用者→書目資料」（範例如表七），分別建立作者、題名、分類號、主題標目、出版社共五種「使用者→書目資料」的借閱次數矩陣，紀錄每位使用者對書目資料的借閱次數，如果一個 CatKey 對應到多個書目資料，每個書目資料將被紀錄原始的借閱次數，不會均分次數或做其他處理。

由於不會所有的 CatKey 都出現在訓練資料集中，亦有部分 CatKey 存在書目資料佚失，表八紀錄了經過資料清理與轉換後，用到後續計算書目共借閱率的書目資料種數與借閱紀錄總數。

表六、「使用者→CatKey」的借閱次數矩陣（以作者為例）

	書目資料		使用者		
CatKey			甲	乙	丙
1	彭百顯		0	1	0
17	janebrierley	tonywolf	10	5	0
96	徐小青	李士勛	0	7	5

表七、「使用者→書目資料」的借閱次數矩陣（以作者為例）

	書目資料			
使用者	janebrierley	徐小青	李士勛	彭百顯
甲	10	0	0	0
乙	5	7	7	1
丙	0	5	5	0

doi:10.6342/NTU202400232

表八、訓練資料集中的 CatKey 數量與轉換後的書目資料數量表

類型	種數	借閱紀錄總數
CatKey	531,347	53,857,397
作者	323,683	93,922,669
題名	457,109	53,454,132
分類號	20,732	43,958,528
主題標目	26,835	45,304,641
出版社	29,249	54,826,241

三、建構書目資料相似度矩陣

建構書目相似度矩陣分為三個步驟，將借閱紀錄轉換為共被借閱率，針對同一種書目資料進行階層分群，最終得出書目資料間的相似度。各步驟依序說明如下：

1. 建構共被借閱率矩陣

為了建構書目資料相似度矩陣，本研究先將使用者的書目資料借閱紀錄建構出借閱次數矩陣（表七），以公式(1)轉換為共被借閱次數矩陣（表九）並進行正規化，得出共被借閱率矩陣（表十），使冷門與熱門 CatKey 的書目資料可以互相比較，避免較冷門的書目資料，即使高機率被共同借閱，卻因為共被借閱次數低而被視為相似度低。轉換方式如公式(1)、正規化方式如公式(2)。

$$\begin{bmatrix} B_1B_1 & \cdots & B_1B_d \\ \vdots & \ddots & \vdots \\ B_dB_1 & \cdots & B_dB_d \end{bmatrix} = \begin{bmatrix} U_1B_1 & \cdots & U_nB_1 \\ \vdots & \ddots & \vdots \\ U_1B_d & \cdots & U_nB_d \end{bmatrix} \cdot \begin{bmatrix} B_1U_1 & \cdots & B_dU_1 \\ \vdots & \ddots & \vdots \\ B_1U_n & \cdots & B_dU_n \end{bmatrix} \quad (1)$$

B：書目資料、d：書目資料種數、U：使用者、n：使用者人數

$$\begin{bmatrix} \frac{B_1B_1}{\sum_{i=1}^{d} B_iB_1} & \cdots & \frac{B_1B_d}{\sum_{i=1}^{d} B_iB_d} \\ \vdots & \ddots & \vdots \\ \frac{B_dB_1}{\sum_{i=1}^{d} B_iB_1} & \cdots & \frac{B_dB_d}{\sum_{i=1}^{d} B_iB_d} \end{bmatrix} \quad (2)$$

B：書目資料、d：書目資料種數

doi:10.6342/NTU202400232

轉換後的共被借閱率矩陣如表十，矩陣中的數值代表兩個書目資料間互相的共被借閱次數，占各自所有共被借閱次數的比例。

表九、「書目資料↹書目資料」的共被借閱矩陣範例

	書目資料			
書目資料	janebrierley	徐小青	李士勛	彭百顯
janebrierley	-	70	70	5
徐小青	70	-	84	0
李士勛	70	84	-	0
彭百顯	5	0	0	-
總計	145	154	154	5

表十、「書目資料↹書目資料」的共被借閱率矩陣範例

	書目資料			
書目資料	janebrierley	徐小青	李士勛	彭百顯
janebrierley	-	0.4545	0.4545	1
徐小青	0.4828	-	0.5455	0
李士勛	0.4828	0.5455	-	0
彭百顯	0.0345	0	0	

使用共被借閱率除了可避免 CatKey 共被借閱次數規模影響相似度計算外，也能觀察兩個書目資料之間的單向相似度，比如針對書目資料「janebrierley」與「徐小青」在表九中共被借閱次數為 70，但因為「janebrierley」總共有 145 次共被借閱，而「徐小青」有 154 次，也就是說「徐小青」比「janebrierley」更熱門。在共被借閱率矩陣中，[janebrierley, 徐小青]的數值因此大於[徐小青, janebrierley]的數值，前者為 0.4828，後者為 0.4545。相較之下，兩者間的共被借閱關係，對「janebrierley」來說更重要。

doi:10.6342/NTU202400232

共被借閱率可以避免冷門書籍和熱門書籍之間，只有極少共被借閱次數卻有極高共被借閱率的情況。舉一個極端的狀況說明，假設書目資料 X 與 Y 的共被借閱次數為 1，並且 X 不曾與其他書目資料共同被借閱，在計算 X 與 Y 的共被借閱率時，不論以借閱次數、X 或 Y 的共被借閱次數作為分母都不適合，即使分別用 X 與 Y 的共被借閱率取平均，也會因為 X 的共被借閱率是 1，導致 X 與 Y 的共被借閱率最低也有 0.5，因此將兩者分別計算比較能呈現書目資料共被借閱狀況。

2. 以使用 MPC 方法進行階層分群

為了反映書目資料之間的複雜主題關係，本研究將根據表十的共被借閱率矩陣進行階層分群(Hierarchical Clustering)。雖然也可以直接根據表中單向的共被借閱率代表書目資料間的相似性，但研究者期望透過階層分群，找出書籍主題間的中介關係，亦可豐富書目資料間的關係，有助於改善資料稀疏之問題。若單純用依據具方向性共被借閱率，以關聯規則方法進行推薦的話，無法呈現書籍之間複雜的主題關係，不利於向外延伸推薦主題。

舉例來說，一個想要開飲料店的人可能會先收集調製飲料的資訊，然後再學習成本管控、定價等經營管理技術，最後才借閱登記納稅等商業法律相關的資訊。當他第一步先借閱了飲料調製的書籍，關聯規則推薦方式只能推薦下一步的經營管理書籍，而無法同時顧及經營管理與商業法律資料的高度相關，只能等使用者借閱過經營管理書籍後，才會提供下一步的資料，但若以階層分群方法就可以同時推薦不同主題的書籍，並依照相關程度排序。

本研究利用共被借閱率將書目資料做階層分群，呈現書目資料間豐富的關係。分群方法主要參考 Geng、Deng and Ali（2008）提出的 Message Passing Clustering（MPC）階層分群方法。MPC 的概念十分簡單，針對需分群資料評估是否有任兩項互相為關係最緊密，若有則將其合併，重複此步驟直到所有資料都無法再被合併。與常用的 K-means 或是傳統的階層分群方法相比，此方法的優點是可以平行計算

doi:10.6342/NTU202400232

多個資料點，同時產生多組合併對象。此外，不像 K-means 方法會受到起始點選擇的影響，不管從何資料作為起始點，或重複多少次，獲得的分群結果都是一樣。

使用此方法的原因是，分群時使用的每一筆書目資料共被借閱率，都只對與該數據相關的兩個書目資料有意義，任三個書目資料間的共被借閱率可能不滿足三角不等式，使得書目資料共被借閱率矩陣投射的空間是扭曲的，不適合使用 K-means 或是近鄰分群方法。MPC 方法的運作只考慮兩個書目資料的共被借閱率進行分群，符合本研究的資料特性和分群需求。然而，MPC 方法在分群時不會優先合併共被借閱率最高的資料，而是針對每筆資料分別評估，導致高估熱門書目資料間的的相似度。因為熱門書目資料被借閱次數高，與較多書目資料都有共被借閱次數，次數分散使得共被借閱率普遍偏低。但熱門書目資料間發生共被借閱的機會又很高，很可能與另一本主題不相關的熱門書目戶為彼此共被借閱率最高對象，在很早期就被 MPC 方法合併。可能出現共被借閱率<0.1 與>0.5 的兩組資料都在同一階層被合併，而被賦予相同的相似分數。

為了避免高共被借閱率與低共被借閱率都被賦予相同的相似分數，本研究會先設定每一階層的共被借閱率之閾值，來優先合併高共被借閱率的書目資料。實行方式包含下列三個步驟：設定閾值、使用 MPC 分群、重新計算書目資料相似度網路中各點的共被借閱率。

(1) 設定閾值

為了避免熱門書目資料對分群結果的影響，設定閾值來優先合併高共被借閱率的書籍，可以避免熱門書目資料過早被分到同一群。因為 MPC 分群關注的是「互為最相似」的兩個書目資料，因此會將未經分群的書目資料共被借閱率矩陣（表十）中，抽取每個書目資料在矩陣內的最高共被借閱率，以表十為例就是[1, 0.5455, 0.5455, 0.4828]，並將這組數據由高至低排序取其十分位數切成十份，第一階分群時就以第一十分位數為閾值，當一對書目資料互為最相似，且雙方對彼此的共被借閱率皆高於閾值時才會合併。第二階分群時

doi:10.6342/NTU202400232

閾值則改以第二十分位數的共被借閱率，計算數值最高的 20%資料，依序遞增直到所有書目資料相似度都被納入分群，每一階分群皆只針對高於該階閾值的共被借閱率做分群。

(2) 使用 MPC 分群

根據「書目資料↹書目資料」的共被借閱率矩陣進行分群，MPC 方法會依序偵測每一個書目資料與其他書目資料的共被借閱率，並找出數值最高者，最後檢查是否有任兩個書目資料互為彼此共被借閱率最高的對象，並且兩個共被借閱率都高於閾值，若有則將其分為同一群。

(3) 重新計算網絡中各點的共被借閱率

將被分為同一群的書目資料合併，同一群的共被借閱次數相加後，再回到步驟 1，尋找與合併後的書目資料共被借閱次數最高的另一個書目資料。以表十為範例說明的話，書目資料「janebrierley」與「徐小青」會被分為一群，「李士勛」與「彭百顯」則不會被分群。合併之後的共被借閱次數表格將如表十一，相似度如表十二所示。

依照分群結果重新計算共被借閱率後，重新回到步驟 2 使用 MPC 分群方法重新檢視，直到不再有書目資料能合併時，就將目前的分群結果視為第一階分群，並重新回到步驟 1 將閾值放寬，進入第二階分群。經過反覆分群後，可以得到十階分群的書目資料網絡，理論上第十階時可以將所有具共被借閱關係的書目資料合併。

表十一、一次分群後的共被借閱矩陣範例

	書目資料			總次數
書目資料	janebrierley+徐小青	李士勛	彭百顯	
janebrierley+徐小青	-	154	5	159
李士勛	154	-	0	154
彭百顯	5	0		5

doi:10.6342/NTU202400232

表十二、一次分群後的共被借閱率矩陣範例

	書目資料		
書目資料	janebrierley+徐小青	李士勛	彭百顯
janebrierley+徐小青	-	0.9686	0.0314
李士勛	1	-	0
彭百顯	1	0	-

3. 建構書目資料相似度矩陣

在完成十階分群後，本研究根據書目資料被分為同群的階層次序設定其相似度，第一階層被分為同一群的書目資料間相似度為 10，第二階層相似度則為 9，之後等差遞減至第十階層同群者的相似度為 1。

四、推薦結果與評估

利用書目資料相似度矩陣，本研究可根據個別使用者在訓練資料集中的借閱紀錄計算個人對書目資料的興趣特徵，五種書目資料各自形成一組興趣特徵。比較興趣特徵與答案資料集中的借閱紀錄，即可評估興趣特徵的推薦效益如何。由於並不是所有的使用者都在訓練資料集與答案資料集中有借閱紀錄，因此實際用於評估的使用者共 256,762 人。

建構使用者個人興趣特徵時需要先將使用者的書籍借閱紀錄，依照 5 種書目資料被借閱的情況轉換成書目資料借閱特徵。比如表十三中的使用者(abc)借閱了 4 本書，分別由 4 位不同的作者撰寫，那 abc 的作者借閱特徵值就是 4 位作者各四分之一；4 本書出自 3 家出版社，月亮出版有 2 本，因此它的特徵值就有二分之一，其它兩家各為四分之一，轉換後的書目資料借閱特徵如表十四。

將使用者的書目資料之借閱特徵與上一步驟建置的書目資料相似度矩陣取內積，便可以得出單一讀者對五種書目中每一項書目資料的推測興趣程度，作為讀者的興趣特徵，因為相似度介於 0~10 之間，因此取平均值的推測興趣程度最大值為

doi:10.6342/NTU202400232

10。最小值為 0，數值越大代表本研究推測使用者對該書目資料越有興趣，（如表十五）。

表十三、使用者借閱紀錄（範例）

使用者姓名	abc				
	書目資料				
Catkey	作者	分類號	主題標目	題名	出版社
23	小坪	835	-	風	太陽出版
59764	小名	987	畫冊	火	月亮出版
113258	曉華	104	佛教	水	月亮出版
995	小張	588	股票	土	星星出版

表十四、使用者 abc 個人借閱特徵

使用者姓名		abc							
書目資料									
作者	權重	分類號	權重	主題標目	權重	題名	權重	出版社	權重
小坪	0.25	835	0.25	畫冊	0.25	風	0.25	太陽出版	0.25
小名	0.25	987	0.25	佛教	0.25	火	0.25	月亮出版	0.5
曉華	0.25	104	0.25	股票	0.25	水	0.25	星星出版	0.25
小張	0.25	588	0.25			土	0.25		

doi:10.6342/NTU202400232

表十五、使用者 abc 之分類號興趣特徵分數表

使用者：abc	其他分類號	100	830	890	998
分類號	權重	相似度	相似度	相似度	相似度
835	0.25	8	8	5	3
987	0.25	0	8	6	9
104	0.25	7	8	2	7
588	0.25	2	0	2	1
推測興趣程度		4.25	6	3.75	5

取得使用者的五組書目資料興趣特徵後，透過以下方式評估興趣特徵對答案資料集的借閱紀錄之預測效果。

1. 預測分數

觀察使用者在答案資料集裡借閱過的書目資料在興趣特徵中的分數高低，若平均分數越高，代表興趣特徵可以反映使用者的閱讀行為。計算公式如公式(3)

$$\text{Score} = \begin{bmatrix} U_1B_1 \\ \vdots \\ U_1B_d \end{bmatrix} \cdot [U_1A_1 \quad \cdots \quad U_1A_d] \tag{3}$$

U_iB_d：使用者U_i對書目資料B_d的興趣值

U_iA_d：使用者U_i在答案資料集中借閱書目資料A_d的次數

2. 餘弦相似性

將興趣特徵在書目資料空間中的向量與同一位使用者的答案資料集借閱紀錄在書目資料空間中的向量做比較，若餘弦相似度越大則代表兩向量越相似，即興趣特徵能預測答案資料集的借閱行為。計算公式如公式(4)

$$\text{Similarity} = \cos\theta\,(U_iB, U_i\text{A}) = \frac{\sum_{j=1}^{d} U_iB_j \times U_iA_j}{\sqrt{\sum_{j=1}^{d} U_iB_j{}^2} \times \sqrt{\sum_{j=1}^{d} U_iA_j{}^2}} \tag{4}$$

doi:10.6342/NTU202400232

3. 答案資料集的預測排名

答案資料集為使用者未來的實際借閱情形，若是這些真實被借閱的 Catkey 在預測結果中排名靠前，代表預測結果對使用者具效用，使用者參考預測結果，可以提前發現他們未來有興趣的 Catkey。反之，若是排名靠後則使用者無法得知這些 Catkey，因為人的認知能力與時間都是有限，太長的預測結果無法被使用者有效利用。

由於計算所有 Catkey 的預測分數再做排名之排名需要耗費大量儲存空間，相較預測分數與餘弦相似性的計算結果都是一個整合的數值，若要計算每個 Catkey 的預測排名，每位使用者都需要紀錄 64 萬 8,596 個資料（Catkey 總數），所需儲存空間預計高達 640GB（25 萬*64 萬*uint32）。本研究在電腦效能與儲存空間的取捨下，將 25 萬 6,762 名使用者分為 121 組，每組 2,122 人（2122*121= 256,762），一組一組計算使用者的個人興趣特徵後，隨機抽選該組中 200 人計算 Catkey 的預測排名，總計隨機選出 2 萬 4,200 位使用者計算完整之預測分數，並計算這些使用者在答案資料集中實際借閱過的 Catkey 之預測排名。

五、評估整合推薦的效益與書目資料相關程度

由於書目資料具有能反映不同面向的主題，本研究嘗試以調和平均數（Harmonic Mean）來整合五種書目資料的預測結果，並與個別書目資料的預測結果相比較，觀察是否整合後的結果有何差異。整合方式如公式(5)。

$$H(U_i, C_j) = \frac{5}{\frac{1}{T_{rank}(U_i, C_j)} + \frac{1}{A_{rank}(U_i, C_j)} + \frac{1}{CN_{rank}(U_i, C_j)} + \frac{1}{P_{rank}(U_i, C_j)} + \frac{1}{SH_{rank}(U_i, C_j)}} \quad (5)$$

U_i：第i位使用者、C_j：第j個$Catkey$、T_{rank}：在題名中的預測排名、

A_{rank}：在作者中的預測排名、CN_{rank}：在分類號中的預測排名、

P_{rank}：在出版社中的預測排名、SH_{rank}：在主題標目中的預測排名

另外，若是不同書目資料的推薦結果高度相關，代表它們反應的結果相似，實務上可以選擇預測成功率較高的書目資料作為代表，減少資料運算需求。由於不同

doi:10.6342/NTU202400232

書目資料的預測分數無法直接比較，為了比較不同書目資料產出之個人書籍推薦結果間的相關程度，並評估是否有相互取代的可能性，因此本研究使用預測排名分析不同書目資料之預測結果的相關程度。因為預測排名是次數尺度(Ordinal Scale)變項，故採用 Spearman 相關係數評估書目資料間的相關程度，Spearman 相關係數的計算方式如公式(6)。

$$r_S = \rho_{R(C_a),R(C_b)} = \frac{\sum_{j=1}^{N} R(C_a)_j \times R(C_b)_j - \frac{\left[\sum R(C_a)_j\right] \times \left[\sum R(C_b)_j\right]}{N}}{\sigma R(C_a) \times \sigma R(C_b)} \quad (6)$$

$R_a(C_j)$：書目資料a的預測排名中第j個$Catkey$的$Rank$

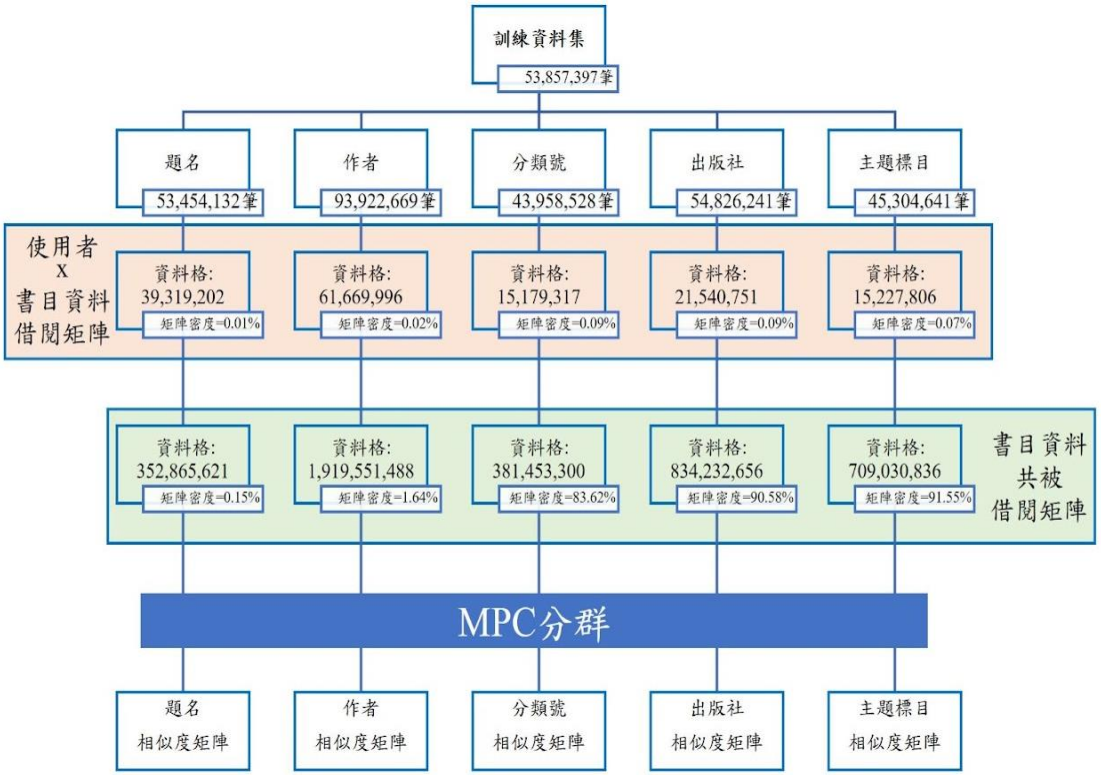
σ：同一書目資料所得的預測排名之標準差

doi:10.6342/NTU202400232

第四章、研究結果

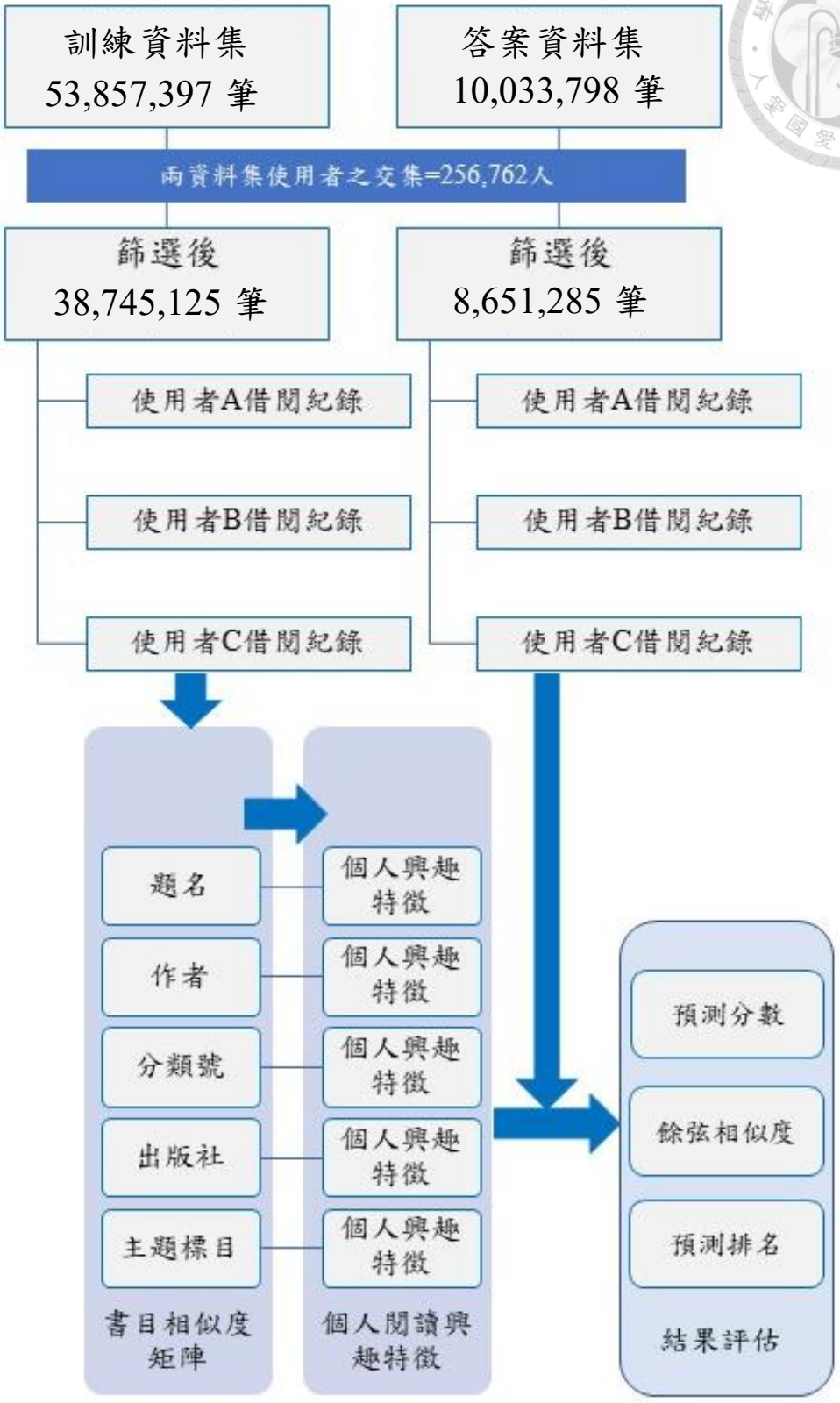
第一節　實驗流程

本節首先依照上一章研究方法之流程，說明訓練資料集與答案資料集的資料變化。在建構書目相似度矩陣時，訓練資料集的資料變化如圖一，推薦結果之產出與評估流程如圖二：



圖一、書目資料相似度矩陣計算流程圖

doi:10.6342/NTU202400232



圖二、個人興趣特徵產製與評估流程圖

doi:10.6342/NTU202400232

第二節、個人興趣特徵與答案資料集之相似性

得出使用者的個人興趣特徵後，本研究透過預測分數與餘弦相似度兩項指標，評估訓練資料集呈現的興趣特徵與讀者在答案資料集中借閱行為的相似程度。

1. 預測分數

由於部分使用者的借閱紀錄可能因書目資料闕漏而呈現空值，表十六呈現了五種書目資料各自的有效使用者人數、答案資料集中有效使用者的「書目資料借閱次數」、人均借閱次數，以及預測分數總分。為了比較有不同借閱次數的書目資料間的差異，將預測分數總分除以總借閱次數，可得出「單次借閱均分」，反映答案資料集中每一筆借閱紀錄，平均來說在預測結果中得到幾分，分數越高代表預測結果越接近答案資料集。

表十六、五種書目資料的預測分數摘要表

書目資料	使用者人數	答案資料集總借閱次數	使用者人均借閱次數	預測分數總分	單次借閱均分
題名	252,199	8,595,315	34.08148	8,379,356	0.97488
作者	249,409	19,267,124	77.25112	40,814,432	2.11835
分類號	248,901	7,149,867	28.72575	30,621,739	4.28284
出版社	251,931	8,909,936	35.36657	50,044,024	5.61665
主題標目	252,034	15,636,294	62.04042	91,458,408	5.84911

預測分數中，由於「作者」與「主題標目」有許多 1 個 Catkey 對應多個書目資料的情形，因此轉化為答案資料集後的總借閱次數明顯高於另外三種書目資料。在單次借閱紀錄的平均預測分數上，「題名」分數最低，僅有 0.97488 分，「主題標目」、「出版社」平均分數最高，分別為 5.61665 與 5.84911。若單依預測分數這一項指標，出版社與主題標目產出的興趣特徵與答案資料集中的借閱紀錄最相似。

doi:10.6342/NTU202400232

為了進一步分析預測分數對個別讀者的效益，減少極大量借閱的或極少借閱的使用者對整體結果的影響，因此本研究進一步計算每位使用者各自的單次借閱平均，找出所有使用者的平均數，反映出每一位使用者接收到的預測分數，避免部分讀者表十七分別呈現了個別讀者得分平均，個別讀者單次得分平均，並擷取借閱次數介於四分位距（IQR）間的使用者，分析其平均得分和單次得分平均。

表十七、五種書目特徵的人均預測得分摘要表

書目資料	單次借閱均分	人均得分平均	人均單次得分平均	四分位距（IQR）間		
				單次借閱均分	人均得分平均	人均單次得分平均
題名	0.97488	33.22518	0.93774	0.93260	18.29125	0.92992
作者	2.11835	163.64459	1.77858	1.80260	76.27002	1.75497
分類號	4.28284	123.02779	4.18575	4.25113	69.96873	4.22734
出版社	5.61665	198.64179	5.46677	5.45603	110.59598	5.43297
主題標目	5.84911	362.88123	5.81610	5.90975	205.24708	5.89305

比較整體數值與 IQR 之間的數據，可以觀察以下情形：

(1) 不論整體數據或是 IQR 間的數據，「主題標目」、「出版社」的平均分數都是最高，人均得分與人均單次得分都較高，顯示這兩種書目資料對答案資料集中的借閱紀錄預測分數最高。

(2) 依照借閱次數排序時，五種書目資料人均借閱次數都呈現平均數>中位數的右偏樣態，顯示受到少數高借閱次數族群的影響較大。（題名 M=34>MD=17；作者 M=77>MD=38；分類號 M=28>MD=15；出版社 M=35>MD=18；主題標目 M=62>MD=30）

doi:10.6342/NTU202400232

(3) 「分類號」、「主題標目」這兩種書目資料在 IQR 間數據的人均單次得分較整體數據提升，凸顯這兩種書目資料的右偏型態，並不是因為有高分的大量借閱使用者存在，而是有較高比例的使用者獲得高分。

(4) 「題名」、「作者」、「出版社」在 IQR 間數據的人均單次得分較整體數據減少，加上整體資料分布呈右偏型態，顯示高借閱次數族群對整體資料的影響較大，因此在排除借閱次數會多與最少 25%的使用者後，人均單次得分下降。

2. 餘弦相似度

餘弦相似度代表書目資料的向量空間中，讀者個人興趣特徵與借閱紀錄的向量夾角餘弦值，數值越大代表兩個向量越相似，五種書目資料分別計算的個人興趣特徵與兩個資料集中借閱紀錄的餘弦相似度如表十八。

表十八、使用者個人興趣特徵與借閱行為的餘弦相似度摘要表

書目資料	整體借閱行為	訓練資料集借閱行為	答案資料集借閱行為	不重複答案資料集借閱行為（扣除兩資料集重複目標）
題名	0.01333	0.01227	0.00564	0.00550
作者	0.03467	0.03245	0.01535	0.01429
分類號	0.10018	0.09302	0.06522	0.05873
出版社	0.13753	0.12829	0.08564	0.07318
主題標目	0.05618	0.05331	0.03803	0.04083

利用餘弦相似度可以評估不同書目資料產出的個人興趣特徵與使用者針對該種書目資料的借閱行為有多相似，由上表可見「出版社」的使用者個人興趣特徵與資料集中的借閱行為相似度最高，其次為「分類號」，顯示這兩種書目資料所產出的興趣特徵與讀者的實際行為最接近。

doi:10.6342/NTU202400232

「主題標目」的餘弦相似度名列第三，但在不重複答案資料集的餘弦相似度這項指標上表現與其他書目資料不同。一般來說，不重複資料集因為扣除了答案資料集與訓練資料集中重複的書目資料，餘弦相似度應該會降低，被扣除的那些書目資料就是用於計算興趣特徵的原始資料。「主題標目」的興趣特徵與答案資料集的餘弦相似度為 0.03803，而與不重複答案資料集的餘弦相似度反而提升至 0.04083，數值的提升代表以「主題標目」推算出的興趣特徵，相較答案資料集整體，反而與未曾借閱的書目資料有更高的相似度。

第三節、預測排名

根據使用者對書目資料的興趣特徵可以計算使用者對所有 Catkey 的興趣分數，本研究針對隨機抽選的 24,200 位使用者計算興趣特徵並轉換為 Catkey 興趣分數，這些使用者共借閱 817,606 次，彙整後的「使用者→Catkey」的借閱矩陣中有 646,049 格資料。

由於實務上使用者無法瀏覽全部推薦結果，因此排名變得十分重要，本研究針對預測排名在前 5 名、前 10 名、前 20 名與前 649 名（所有 Catkey 數量的 0.1%）的表現，比較五種書目資料加上利用調和平均數轉換的綜合預測排名之效益，數據紀錄如表十九。

不論是哪一種排名，題名的預測成功次數都高於其他種書目資料，前 5 名的預測成功次數就有 22,786 次，前 20 名的則增加至 25,939 次。而且題名預測成功的書籍排名都很前面，前 5 名與前 649 名的預測成功次數只增加了 23.40%，代表大多數預測成功的結果排名都是前 5 名。

前 5 名預測成功次數排名第二的書目資料是分類號，共有 5,036 次，但隨著前 N 名的標準放寬，前 649 名的預測成功次數僅增加 7 次，增加 0.14%，顯示該書目資料僅在前 5 名的表現較佳。相較之下，作者前 5 名預測成功次數雖然只有 277 次，但若放寬至前 10 名，就增加為 5,245 次，超過分類號的預測成功次數，放寬

doi:10.6342/NTU202400232

到前 649 名的預設成功次數增加至 7,178 次，較前 5 名的預測成功次數增加 2369.68%。

表十九、書目資料預測排名前 5、10、20、649 名的預測成功次數比較表

	預測成功次數			
書目資料	前 5 名	前 10 名	前 20 名	前 649 名
綜合預測	285	741	828	5,154
題名	22,786	25,358	25,939	28,117
作者	277	5,245	6,841	7,178
分類號	5,036	5,040	5,040	5,043
出版社	21	24	54	1,944
主題標目	0	4	5	108

綜合預測的結果並未收到較好效果，前 20 名預測成功次數僅有 828 次，排行第四。雖然前 649 名預測成功次數達 5,154 次，排名第三，但並未與第四名的分類號拉開明顯差距。在預測分數與餘弦相似度分數較高的出版社與主題標目，在前 N 名的預測成功次數都很少，顯示其推薦結果，對於認知負荷有限的一般使用者來說沒有效益，單就預測排名這項指標無法與其他種書目資料相比。

第四節、書目資料預測排名相關性

五種書目資料反映書籍不同的主題或特徵，各自產出不同的推薦清單，但如果不同種書目資料間的相似性較高，在實際運用時可以考慮選擇預測準確度較高的書目資料，減少需運算資料量並節省時間。反之，若是兩個書目資料的推薦清單相關程度很低，可能代表它們反映了書籍的不同主題，值得進一步評估該如何整合兩份推薦清單。

doi:10.6342/NTU202400232

本研究分別針對 256,762 位使用者各自的五種書目資料推薦清單計算預測排名相關係數，五種書目資料的推薦排名間之相關係數平均值如表二十。

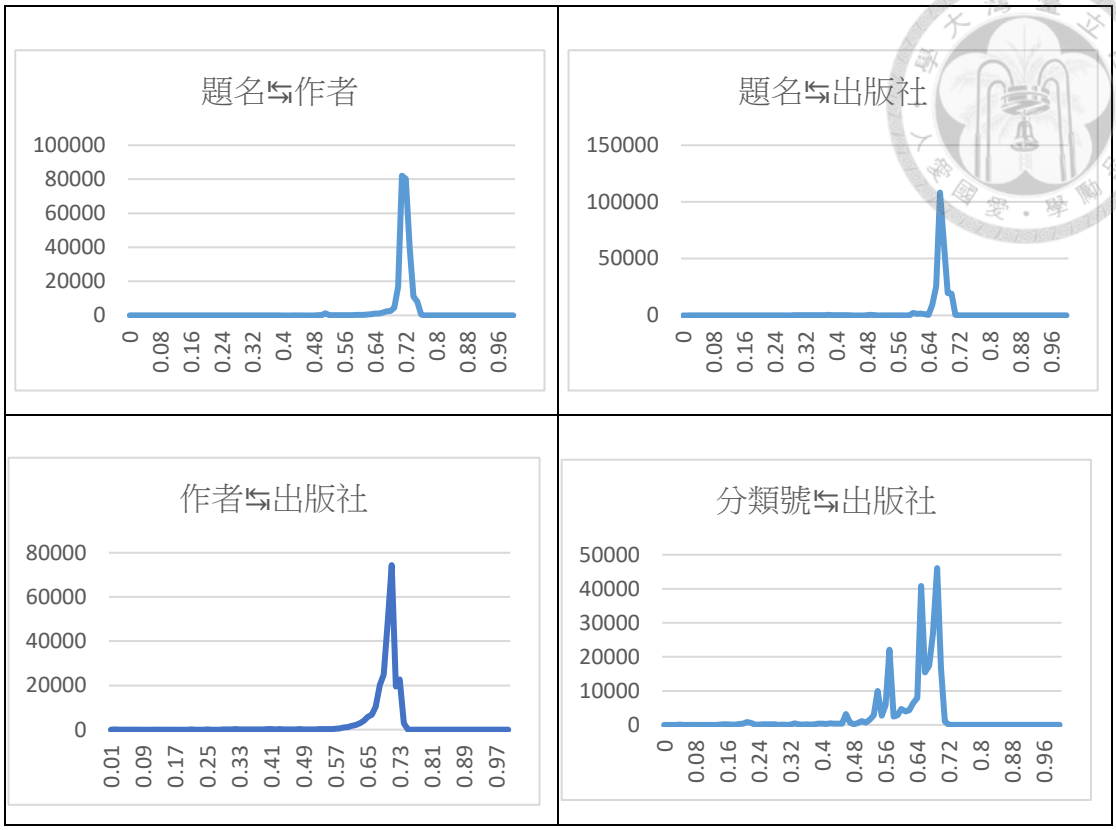
表二十、書目資料推薦清單相關係數平均數比較表

書目資料	題名	作者	分類號	出版社	主題標目
題名	1	0.70876	0.48564	0.66515	0.38244
作者	0.70876	1	0.51693	0.68668	0.36350
分類號	0.48564	0.51693	1	0.62614	0.19357
出版社	0.66515	0.68668	0.62614	1	0.37317
主題標目	0.38244	0.36350	0.19357	0.37317	1

一般來說，相關係數介於 0.6~0.99 視為高度相關，0.4~0.59 視為中度相關，小於 0.39 被視為低度相關。從上表中可以發現，高度相關的包含「題名↹作者」、「題名↹出版社」、「作者↹出版社」、「分類號↹出版社」。進一步檢視這 4 種相關係數的資料分布，資料分布圖如圖三。

依據資料分布圖可發現，「題名↹作者」、「題名↹出版社」、「作者↹出版社」等三種書目相關係數之資料分布尚屬正常，平均值正負 1 倍標準差以內使用者的佔全體 88%以上，相關係數低於 0.6 的使用者人數分別僅有 1.115%、2.744%、1.152%，因此可推論對多數讀者來說，前述書目資料的預測結果之間相關程度較高。「分類號↹出版社」雖然也是高度相關，但其資料分布呈現多個峰值，同時「分類號」與「作者」、「題名」之間也無呈現高度相關，故無法評估是否可取代或被其他書目資料取代。

doi:10.6342/NTU202400232



圖三、高度相關書目資料之相關係數分布圖

本章第一節至第三節的各項分析顯示，高度相關的「題名」、「作者」與「出版社」等三種書目資料間，「題名」與「作者」在預測排名前 20 名預測成功的次數表現較佳，「出版社」的預測分數、興趣特徵與答案資料集餘弦相似度則表現較好。「作者↹出版社」之間出版社也是在預測分數與餘弦相似度上表現較好，前 20 名的預測次數則是作者佔優。比較資料如表二十一。

表二十一、題名、作者與出版社之預測數據比較表

書目資料	人均單次預測得分平均	不重複答案資料集借閱行為餘弦相似度	前 5 筆預測成功次數	前 10 筆預測成功次數
題名	0.93774	0.00550	22,786	25,939
作者	1.77858	0.01429	277	6,841
出版社	5.46677	0.07318	21	54

doi:10.6342/NTU202400232

上述數據顯示，雖然「題名」、「作者」與「出版社」的預測結果之間有高度相關，但是這些預測結果的特徵有所不同，出版社在平均預測分數與餘弦相似度較高，代表它的預測結果「整體」來說表現較好，但若考量到使用者有限的認知能力，題名與作者在前 20 筆的預測情況明顯更佳。實務上運用時或許可以參考此一現象，針對初次使用或是非重度閱讀者，可以優先依據前 20 名預測成功次數最多的題名，以及和「題名」、「作者」非高度相關，前五名預測成功次數也很高的分類號進行推薦，面對重度使用者或對預測結果不滿的使用者，再提供依據出版社所提出的預測結果做為參考。

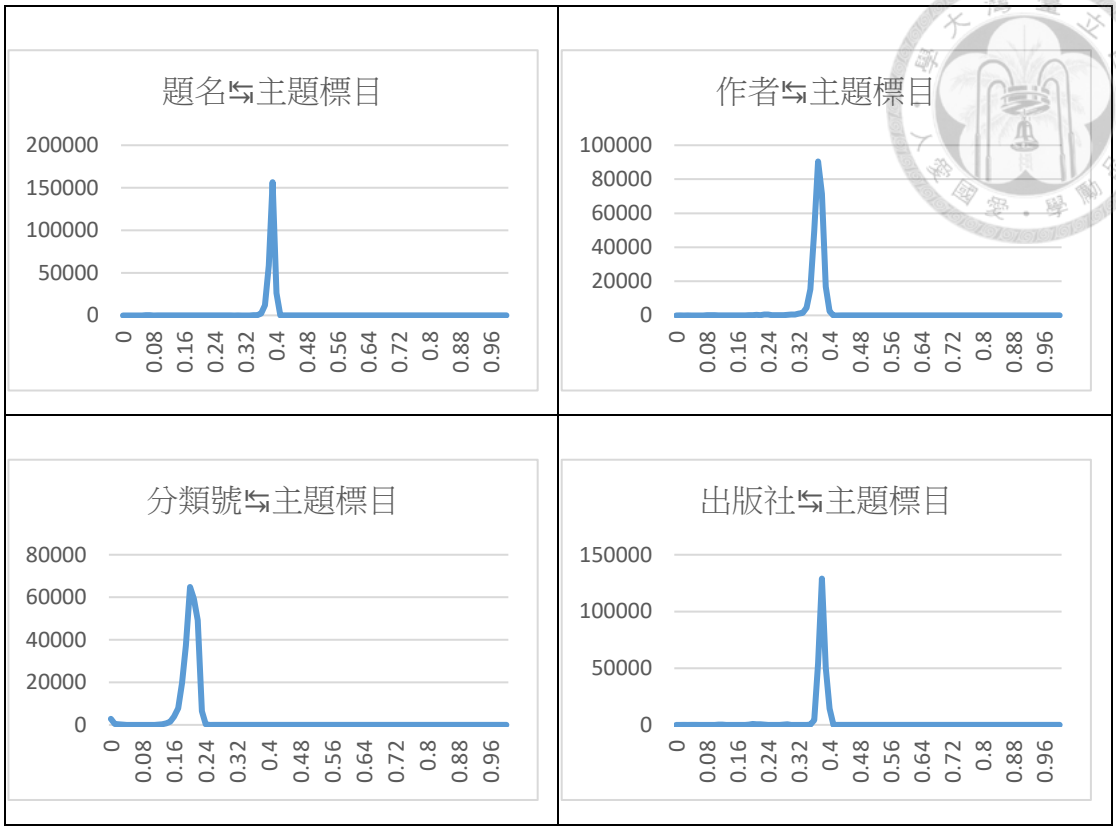
另一方面，「主題標目」的預測結果與其他每一種書目資料都呈現低度相關，預測結果的人均預測分數在五種書目中排名第一，表示「主題標目」和「出版社」的情況類似，排名高的預測結果效果不佳，但是整體預測結果表現不錯，主題標目與其他書目資料間相關係數分布如圖四，主題標目預測數據摘要如表二十二。

在比較五種書目資料之預測結果間的相關係數後，由於「題名」、「作者」、「出版社」間的高度相關，未來應用時或可考慮不納入「作者」，選擇在前 20 名預測成功次數更多的「題名」，以及整體預測分數較高的「出版社」作為補充。

表二十二、主題標目之預測數據摘要表

書目資料	人均單次預測得分平均	不重複答案資料集借閱行為餘弦相似度	前 5 筆預測成功次數	前 20 筆預測成功次數
主題標目	5.81610	0.04083	0	5

doi:10.6342/NTU202400232



圖四、主題標目與其他書目資料之相關係數分布圖

另一方面，「主題標目」的預測結果雖然在前 20 名的預測效果非常差，但是它的預測分數卻是最高，可見其預測結果仍有參考價值。在相關係數上，它與其他四種書目資料都是低度相關，反映出它呈現與其他種書目資料不同的主題面向，考量使用者在真實環境中不可能接觸到每一件有興趣閱讀的資料，提供多元化的結果有其意義。基於「主題標目」的預測結果可視為另一種次要的推薦清單，而不與主要的推薦清單合併呈現。

doi:10.6342/NTU202400232

doi:10.6342/NTU202400232

第五章、結論

第一節、研究結論

圖書館中收藏著龐大的資訊，遠超過使用者可以接觸與判斷的數量，多數情況下並非使用者不願意尋找，而是尋找資訊、判斷資訊價值的認知負荷超過其預期資訊可帶來的利益。推薦系統是一個可以協助使用者減輕認知負荷的工具，透過基於內容的方式聚合相同主題資訊，或是基於其他使用者經驗的協力過濾方法，找出具有相同興趣與目標的其他使用者曾經接觸過的資訊，都可以有效的協助使用者限縮須尋找、篩選、判讀的資訊範圍。

然而，因為不同推薦方法有各自的優點與限制，本研究希望應用書籍的書目資料為中介，仿照基於內容的推薦方法，將單筆的書籍借閱紀錄轉換為多筆針對不同書目資料的借閱紀錄，豐富借閱紀錄，改善協力過濾方法常見的資料稀疏問題。同時借重協力過濾方法分析實際使用者行為的特性，反映人類知識中複雜而多元的架構與學習途徑，提供多元的主題關聯性給使用者參考。最後，因為不同書籍可能有相同的書目資料，以書目資料為推薦依據可以減少需計算的資料種數，降低運算負荷，並且對未被借閱的書籍，只要其書目資料與其他被外借的書籍相同，就能被納入推薦結果中，改善推薦方法的冷啟動問題。

本研究將書籍借閱紀錄轉換為五種書目資料「題名」、「作者」、「分類號」、「出版社」、「主題標目」的借閱紀錄，並以 Message Passing Cluster 方法進行分群。建構由下而上的分類樹，依照其共被借閱率分為 1 至 10 階，以建構書目資料間的相似度矩陣。根據個別使用者在訓練資料集的借閱紀錄計算其個人興趣特徵，並與其在答案資料集中的借閱紀錄比較，以預測分數、餘弦相似度與預測排名等三種指標評估其個人興趣特徵的推薦效益。

doi:10.6342/NTU202400232

預測分數顯示，「主題標目」與「出版社」兩種書目資料產出的興趣特徵在與答案資料集借閱紀錄相乘後得出的單次預測均分最高，平均為 5.84911 與 5.61665 分。若考量個體差異與極端值影響，這兩種書目資料在人均單次借閱均分、四分位距（IQR）間的人均單次預測均分都還是最高。單以此指標來看「主題標目」與「出版社」的預測效果最好。

在餘弦相似度上，則是「出版社」與「分類號」的表現最好，答案資料集中的不重複借閱紀錄之餘弦相似度分別為 0.07318 與 0.05879，相較之下，「題名」與「作者」的預測分數與餘弦相似度都明顯地低於其他 3 種書目資料。

雖然預測分數與餘弦相似度都可以評估預測結果的準確度，但實務上使用者只能接收有限的推薦清單，推薦數量過多造成另一種資訊超載，因此本研究分別檢視了 5 種書目資料的前 5、前 10、前 20 與前 649 名（總書籍種數的 0.1%）的預測結果，比較五種書目資料預測結果在答案資料集中預測成功的次數，以評估預測結果之效益。

預測排名顯示，以調和平均數整合五種書目的綜合預測結果在前 5 名預測成功 285 次，前 10 名為 741 次，前 20 名為 828 次，前 649 名則增加至 5,154 次。不過使用單一書目資料「題名」或「作者」的預測成功次數都高於綜合推薦結果，在前 20 名分別有 25,939 次與 6,841 次。「分類號」的前 20 名預測成功次數為 5,040 次，略少於綜合預測結果。在預測分數與餘弦相似度分數較高的「出版社」與「主題標目」，分別只有 54 次與 5 次，預測成功次數極低，顯示其推薦結果對於認知負荷有限的一般使用者來說缺乏效益，應優先使用「題名」或「作者」作為推薦依據。

本研究進一步比較使用者以不同種書目資料計算的完整預測排名間之相關係數，觀察不同種書目資料的預測排名是否有高度相關，可以相互取代的情況。分析結果顯示「題名」、「作者」、「出版社」三種書目資料間的人均相關係數都達到 0.6 以上，屬於高度相關，進一步比較三者在預測分數、餘弦相似度與預測排名的

doi:10.6342/NTU202400232

表現，建議選擇前 20 名預測成功次數最多的「題名」取代「作者」的預測結果。並視使用者需求，納入整體預測分數與餘弦相似度表現較好的「出版社」的預測結果，提供給願意接受更多推薦結果的重度使用者，或是對現有推薦結果不滿意的使用者，滿足其需求。

同時，「主題標目」的預測排名與其他書目資料的預測排名都呈現低度相關，但在預測分數與餘弦相似度的表現都不錯，顯示其預測結果仍有效益，與眾不同的推薦排序反而可能反映書籍主題的不同面向。因此本研究建議將「主題標目」的預測結果用於產製另一份書單，提供使用者不同面向的新資訊，增加資訊偶遇的機會。

總結來說，在本研究中依據「分類號」、「出版社」、「主題標目」產出的興趣特徵效果較好，預測分數與餘弦相似度都高於「題名」與「作者」。但若考量使用者有限的認知能力，應優先考慮依據「題名」、「作者」產出的預測結果，這兩者的前 20 筆預測結果較能成功預測答案資料集中真實借閱紀錄，對多數一般使用者來說效益較好。至於整合五種書目資料的推薦清單，在前 20 名只有預測成功 828 次，使用效益不如單一書目。實際應用上，由於「題名」與「作者」的預測結果高度相關，建議以預測成功次數較高的「題名」作為推薦依據，整體預測分數較高、興趣特徵與答案資料集餘弦相似度高的「出版社」、「主題標目」的預測結果則可用於發現新事物，與其他書目資料皆呈現低度相關的「主題標目」更有機會讓使用者到接觸不同主題或是滿足其潛在資訊需求。

第二節、未來研究之建議

本研究僅針對「題名」、「作者」、「分類號」、「出版社」、「主題標目」等五種可能影響選書決策之書目資料進行推薦，未來可進一步納入其他文獻中提及的「名人推薦」、「封面設計」等選書決策影響因子，或是同樣可能與書籍內容主題高度相關的書目資料「叢書名」，提供更豐富的分析資料。

doi:10.6342/NTU202400232

另一方面，由於本研究並未篩選「大量借書」的異常使用者（如 1 年借閱上千本書籍的使用者），或是只有借閱極少書籍的「罕用使用者」，前者產生的原因包含館員策辦主題書展時以公務用或個人借閱證預約、借閱圖書；部分讀者為了各館提供的閱讀獎勵，大量借還圖書但並未閱讀，這些使用者會產出大量無關聯的共被借閱關係，影響書目資料之間的關聯強度。「罕見使用者」則因為借閱紀錄少，無法建構足夠完整的個人興趣特徵，難以產出有效的推薦結果。未來可研究如何制定合適的篩選標準，或是比較訓練資料集中不同借閱次數的使用者，其預測分數與預測排名之表現如何。

doi:10.6342/NTU202400232

參考文獻

卜小蝶（2002）。使用者導向之圖書分類關聯分析研究。**圖書資訊學刊，17**，81-94。

吳安琪（2001）。**利用資料探勘的技術及統計的方法增強圖書館的經營與服務**（碩士論文）。取自 https://hdl.handle.net/11296/3hmq5s

余明哲（2003）。**圖書館個人化館藏推薦系統**（碩士論文）。取自 https://hdl.handle.net/11296/s8vph5

呂家賢（2005）。**運用資料探勘技術於大學圖書館圖書資源推廣利用之研究**（碩士論文）。取自 https://hdl.handle.net/11296/w7k84a

邱宏彬、湯鎰聰、陳揮明（2005）。數位圖書館個人化檢索與推薦服務之設計與實作。**資訊管理研究，5**，1-23。

唐牧群、吳宛青（2009）。由透鏡理論看大學圖書館讀者選書決策過程。**圖書資訊學刊，7**（1/2），37-52。

蔡秀滿與莊宛螢（2007）。使用加權移動視窗模式之圖書資料探勘。**電腦學刊，17**（4），79-96。

戴玉旻（2002）。**圖書館借閱記錄探勘系統**（碩士論文）。取自 https://hdl.handle.net/11296/hgj6h8

謝宜瑾、唐牧群（2013）。從透鏡模式探討影響讀者尋書滿意度之因素─以 aNobii 為例。**圖書資訊學研究，8**（1），69-119。

羅子文（2007）。**Web2.0 概念的圖書館個人化推薦系統**（碩士論文）。取自 https://hdl.handle.net/11296/4uam4c

Adomavicius, G., & Tuzhilin, A. (2001). *Extending recommender systems: A multidimensional approach.* Proceedings of the International Joint Conference on

doi:10.6342/NTU202400232

Artificial Intelligence (IJCAI-01), Workshop on Intelligent Techniques for Web Personalization (ITWP2001).

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS), 23*(1), 103–145. doi: 10.1145/1055709.1055714

Adomavicius, G., & Tuzhilin, A. (2005.) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734-749.

Geng, H., Deng, X., & Ali, H. H. (2008). Message Passing Clustering (MPC): a knowledge-based framework for clustering under biological constraints. *International Journal of Data Mining and Bioinformatics, 2*(2), 95-120.

Han, J., Kamber, M., Pei, J. (Eds.). (2012). *Data Mining: Concepts and Techniques.* (3rd Edition). doi: 10.1016/B978-0-12-381479-1.00010-1.

Lai, Y., & Zeng, J. (2013). A cross-language personalized recommendation model in digital libraries. *The Electronic Library*, *31*(3), 264–277. doi:10.1108/EL-08-2011-0126

Liao, I E., Hsu, W. C., Cheng, M. S., & Chen, L. P. (2010). A library recommender system based on a personal ontology model and collaborative filtering technique for English collections. *The Electronic Library, 28*(3), 386-400.

Lopes, G. R., Souto, M. A.M., Wives, L. K., & de Oliveira, J. P. M. (2008). *A personalized recommender system for digital libraries.* Proceedings of the 14th Brazilian Symposium on Multimedia and the Web - WebMedia 2008 (pp. 59-66). doi:10.1145/1666091.1666103

doi:10.6342/NTU202400232

Lops, P., Jannach, D., Musto, C., Bogers, T., & Koolen, M. (2019). Trends in content-based recommendation. *User Modeling and User-Adapted Interaction, 29*(2), 239-249. doi: 10.1007/s11257-019-09231-w

Mansur, F., Patel, V., & Patel, M. (2017). *A review on recommender systems.* 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 1-6. doi: 10.1109/ICIIECS.2017.8276182

Melville, P., Mooney, R. J., & Nagarajan, R. (2002.) *Content-boosted collaborative filtering for improved recommendations*. Proceedings of the 18th National Conference on Artificial Intelligence (AAAI '02), 187–192. Retrieved from https://www.aaai.org/Papers/AAAI/2002/AAAI02-029.pdf

Mooney, R. J., & Roy, L. (2000). *Content-based book recommending using learning for text categorization*. Proceedings of the Fifth ACM Conference on Digital Libraries, 195–204. doi: 10.1145/336597.336662

Park, D. H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications, 39*(11), 10059-10072. doi: 10.1016/j.eswa.2012.02.038

Pazzani, M.J., & Billsus, D. (2007). Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) Lecture Notes in Computer Science: The Adaptive Web. (pp. 325-341). doi: 10.1007/978-3-540-72079-9_10

Pham, M. C., Cao, Y., Klamma, R., & Jarke, M. (2011). A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis. *Journal of Universal Computer Science (j-jucs), 17*(4). 583-604.

Rahman, M. M. (2013). Contextual Recommender Systems Using a Multidimensional Approach. *International Journal of Intelligent Information Systems, 2*(4), 55-63. doi: 10.11648/j.ijiis.20130204.11

doi:10.6342/NTU202400232

Rajagopal, S., & Kwan, A. (2012). *Book Recommendation System using Data Mining for the University of Hong Kong Libraries.* In B. Fox (Chair), CITE Research Symposium 2012 (CITERS 2012). Symposium conducted at the meeting of CITE (Centre for Information Technology in Education) and Faculty of Education, HKU Hong Kong. Retrieved from http://hub.hku.hk/handle/10722/164694

Sirikayon, C., Thusaranon, P., & Pongtawevirat, P. (2018). *A collaborative filtering based library book recommendation system.* Proceedings of 2018 5th International Conference on Business and Industrial Research (ICBIR), 106-109. doi: 10.1109/ICBIR.2018.8391175

Su, X., & Khoshgoftaar, T. M. (2009). A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence, 2009,* 1–19. doi: 10.1155/2009/421425

Zhang, F. (2016). A Personalized Time-Sequence-Based Book Recommendation Algorithm for Digital Libraries. *IEEE Access, 4,* 2714–2720. doi: 10.1109/ACCESS.2016.2564997

doi:10.6342/NTU202400232