# LDA 和使用紀錄為基礎的線上電子書主題趨勢發掘方法

洪崇洋
國立中山大學資訊管理學系

李建祥
國立中山大學資訊管理學系

黃三益
國立中山大學資訊管理學系

## 摘要

數位內容產業日漸蓬勃，使用者可透過網際網路隨時隨地在電子書檢索平台或圖書館自動化系統檢索、閱讀圖書，因此圖書館購買電子書做為館藏的比例亦逐年增加。有鑑於圖書種類繁多，當圖書館進行電子書採購時通常需參考使用量統計報表做成訂購的決策。然而，由書的點閱量，很難精確預測圖書受喜愛的類型。因此本研究使用 Latent Dirichlet Allocation(LDA)方法，基於圖書內容建置主題模型，結合電子書平台的 COUNTER 統計報表，加權並發掘主題的變化，由不同的角度來觀察使用統計資訊。另外實驗中亦採用卡方獨立性檢定來比較 LDA 主題與美國國會圖書館分類法及圖書主題標目的相關性。

關鍵字: 電子書、使用記錄、主題模型、主題、LDA、LCC、LCSH

# An Approach to eBook Topics Trend Discovery Based on LDA and Usage Log

## Abstract

With the growth of digital content industry, publishers start to provide online services for ebook search, reading and downloading. Nowadays more and more libraries have purchased ebooks as an important part of the library collection. To access the online resources users can link directly to publisher's ebook portal or via the OPAC system. In general, librarians make the decision of purchasing ebooks by referencing ebook usage report. However, the number of requests for ebooks are not sufficient for identifying popular topics. In this study, we combine LDA topic model with the usage report to discover popular topics, while using chi-square independence test to assess whether topic model is independent of classification method and subject heading method in the bibliographic.

Keywords: Ebook, Usage Log, Topic Model, Topic, LDA, LCC, LCSH

# 壹、研究動機與目地

近年來政府積極推動文創及數位內容產業發展，圖書館採購電子書做為館藏的比重亦逐年增加。使用電子書可減少庫存管理上的壓力，減化書籍流通過程，增加資源利用率，且透過網際網路可讓多位使用者同時間閱讀電子書內容。另外隨著閱讀習慣的改變，電子書已成為市場上的主流，出版社及系統服務業者無不積極投入這一個市場。圖書館在眾多的出版社當中，如何選擇最適合讀者的內容，需經過仔細的評估，圖書館採購電子書時除了依據內容本身的價值之外通常也會參考使用統計或讀者的推薦，在有限的預算下做最好的選擇。

對於圖書館來說，統計報表是一個很重要的參考資訊，由於各電子書檢索平台功能設計方式不同，統計的標準亦可能產生差異，為了提供標準電子書統計報表、大部份圖書館會要求系統服務商或出版社依循 COUNTER (COUNTER - Counting Online Usage of Networked Electronic Resources)的規範來提供報表數據。COUNTER 報表規範了統計數據的呈現方式，但相對的也限制了顯示的資訊，其顯示的是個別書籍的使用狀況，所以我們無法得知讀者所感興趣的主題，更無法了解閱讀的趨勢變化。

近年來 LDA (Latent Dirichlet Allocation) 被廣泛運用於各領域，它透過統計學的方式以機率生成主題模型(Topic Model)，提供了一個有效的方法來協助資訊的探索並發掘資料的特徵(Anthes, 2010)。在 LDA 模型裏面每一份文件是由主題的機率分佈所構成，主題中包含的是文件中隨機抽取出來的字彙，透過分析可以讓我們了解文件所描述的主題。由於一般統計報表所產生的統計數據能夠提供的資訊有限，所以本研究探討如何運用圖書文字內容來發掘其中有用的資訊以利於電子書採購的決策。換言之，透過 LDA 的方法來產生主題模型，結合圖書使用記錄進行主題內容加權以呈現主題趨勢的變化，由於加權後主題的變化反應了讀者的圖書類型喜好，所以其提供相對於傳統點擊統計之外更貼近於圖書內容包含的資訊給圖書採購人員作為採購決策之參考。

在圖書的書目記錄中通常包含有圖書的分類方法及透過各領域專家給于圖書的主題標目資訊，由於 LDA 主題模型是由圖書的文字內容所產生，所以書目資訊描述的內容可能和主題模型間相關，所以本研究亦透過 Information Entropy 的計算及 Chi-Square 的方法來驗證其相關性，實驗結果顯示 LDA 產生的主題模型與圖書的書目記錄相關性極低，因此主題模型具有相當高參考的價值，另外我們亦發現出版社所提供的電子書當中，有些圖書並未包含圖書分類及主題標目，所以主題模型亦可做為識別圖書內容的參考。

# 貳、文獻探討

## 一、LDA 主題模型

LDA 是一個以統計為基礎的主題模型，在這一個模型當中假設文件是由一堆的主題按某種機率分佈隨機混合所產生，每一個主題是一個多項式分佈的組合，主題被所有的文件所共享，每一份文件包含各主題的分佈。如圖 1 所示，一個主題當中包含有許多的字彙，同時一個文件是由主題的分佈所組合，範例中分別以不同的顏色來做區別。

| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

圖 1 LDA 主題及包含的字彙以顏色區分、資料來源 (Blei, Ng, & Jordan, 2003)

依圖 2 所示，主題模型的機率分佈透過 hyper-parameters α、β 對主題模型進行控制。其中 α 與 β 為 Dirichlet Prior，α 主要控制主題於文字件上的分佈、而 β 主要控制主題當中文字的分佈、$\theta_i$ 是主題在文件 i 中的機率分佈、$\varphi_k$ 是文字在主題 k 中的機率分佈、$Z_{ij}$ 表示主題中的 j 個字彙於文件 i 的分佈，$W_{ij}$ 代表文件 i 中的字彙、N 為文件中的字彙總數，而 M 為所有文件的數量(Blei, Ng, & Jordan, 2003)。

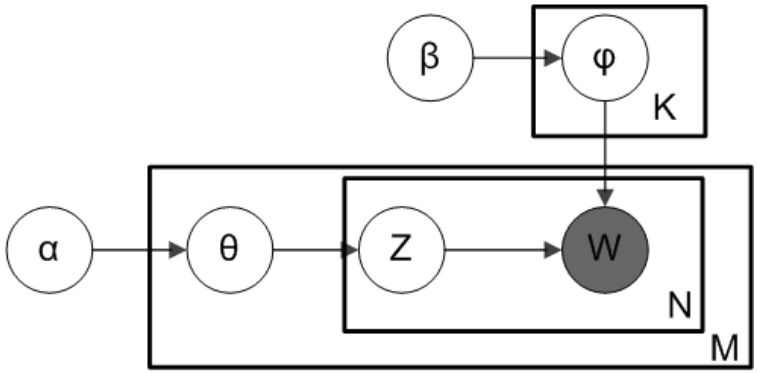


圖 2 Graphical Model of the Smoothed LDA Model、資料來源 ("Wikipedia - Latent Dirichlet Allocation,")

二、COUNTER 統計報表

統計報表的目地就是希望圖書館能夠更好的理解購買的線上服務是如何被使用。要達成這一個目的需要建立一套標準協議讓使記錄可以被良好的記錄與管理，並呈現一致化的格式(Shepherd)。在 2002 年三月 COUNTER (Counting Online Usage of Networked Electronic Resources) 正式釋出。它提供了圖書館、出版社及代理商一個使用統計的參考標準，我們可以用它建立一個具備開放性、一致性、可被信任及相容與各平台的統計報表。COUNTER 目前已被許多出版社廣泛採用，並且在台灣各項電子資源的採購中亦經常被要求要具備這項資訊。

依據 2011 年十月份 Draft Release 4 的 COUNTER Code of Practice 所描述，目前 COUNTER 支援的內容範圍包含 Journal、Database、Book 及 Multimedia 內容。統計數據的呈現主要依報表的型態來做區分，依其登入的帳號或是 IP 位置來做識別及分析。出版社或系統服務商亦可依不同的身份需求來提供統計報表的資訊，例如依個人、組織/機構、聯盟及聯盟成員等不同種類的統計報表。報表格式的輸出必需為 CSV、Microsoft Excel 或其他方便匯入 Microsoft Excel 表格的資料格式。另外亦可提供 XML 格式的報表及報表對應的 XML DTD 檔案(COUNTER - Counting Online Usage of Networked Electronic Resources  Home)。

三、美國國會圖書館分類法與圖書館標題表

美國國會圖書館分類法（Library of Congress Classication，簡稱 LCC）是一個圖書分類的方法，它在 19 世紀末、20 世紀初由美國國會圖書館所發展。這一個分類法在美國大多數的圖書館中被採用，它同時也是全世界最被廣泛使用的圖書分類方法(Library of Congress Classification)。它由 21 個主要的類別所組成，每一個類別由一個英文字母來表示。主類別往下可以再細分為次類別。次類別編碼的方式包含第一碼的分類號，由前兩個或三個英文字母所組成，例如主類別 N (Art) 其下包含有次類別 NA (Architecture)、NB (Sculpture)、ND (Painting) 及其他的分類。

美國國會圖書館標題表（Library of Congress Subject Headings，簡稱 LCSH）是編目人員編輯標題時的必備工具，和 LCC 分類表一樣，標題表是為了館藏編目的用途而建立，進行主題標目編目的人員通常具備有一定的專業能力，才能由圖書內容中來歸納出圖書的主題。LCSH 讓書目記錄包含圖書的主題資訊，協助館藏資料的分類與檢索(國家圖書館編目園地全球資訊網)。在(Khosh-khui, 1987)的研究中發現 LCC 分類法與 LCSH 主題標目具有相關性，相關的程度主要是依 LCC 類別而有所差異，比如 LCC 類別 T(Technology)與 LCSH 的相關性最高，而 LCC 類別 C(General Works)與 LCSH 的相關性最低。

## 參、主題模型建立的方法

一、系統架構

索平台上的使用記錄，在進行主題建立及主題加權之前，必需透過系統化的方式來進行資料的處理、轉換並儲存分析結果，如圖 3 所示。

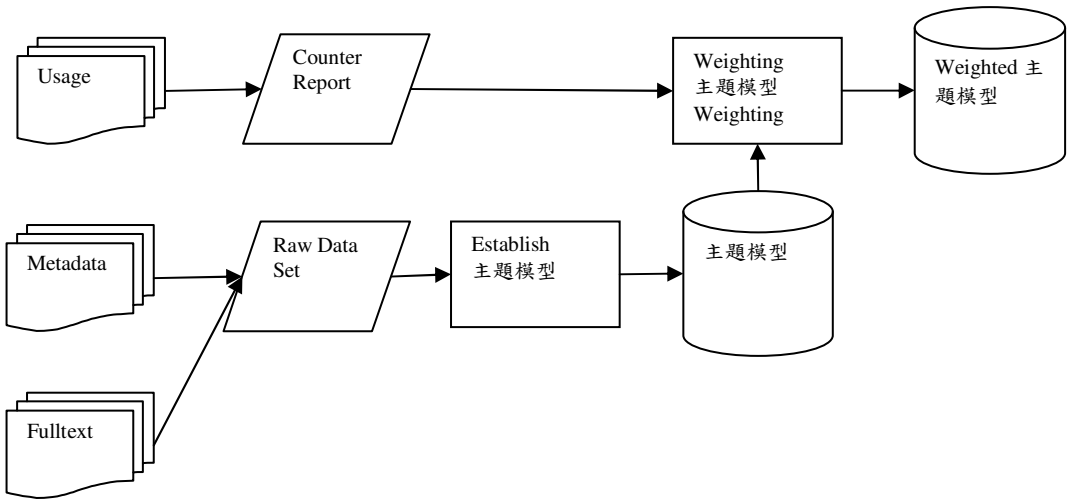


圖 3 LDA 主題模型建立及加權系統結構

文字資料經過一連串的處理過程產生未加權及加權後的主題模型，輸出的結果相關資料儲存在關聯式資料庫當中，透過 SQL 語法就可以進行簡易的資料分析。這一個架構主要包含四個主要步驟，以下簡單描述各步驟的目的：

(1) 圖書文字資料及使用記錄處理

在這一個步驟中會準備文字匯整資料(Bag of Words)及用來加權主題的統計資訊，用來建立主題模型，文字資料的來源包含書目資訊及圖書全文，透過程式進行資料的清理、篩選後產生符合 LDA 工具可以接受的資料格式。使用記錄的部份則是依據 COUNTER REPORT 的使用記錄處理規範來進行過濾並依使用單位的訂購內容、範圍及時間來彙總每本書籍的開啟次數。詳細的步驟請參閱本節之二、三小節。

(2) 主題模型的建置

這一個步驟透過 LDA 轉換工具，使用第一階段產生的文字資料來建置主題模型，在建置前需先決定要產生的主題數量、主題包含的字彙數量、Hyper-parameter $\alpha$ 及 $\beta$ 值。輸出的結果儲存為文字檔案型態，其中包含了字彙與主題的對應、字彙與各主題分佈機率、主題與各圖書分佈機率、相關參數設定等。其中主題模型的建置所需的時間依原始文字資料的大小，主題設定數量多寡所影響。

(3) 使用記錄加權主題

在這一個步驟中使用第一階段產生的統計資料為依據，計算每月、每本圖書在該月的使用率比重，然後透過這一個數值來加權第二階段所產生的主題模型中所有的主題。加權過程中僅處理該月份有被讀者開啟過的圖書其包含的主題，若圖書當月的統計數據為 0 則其包含的主題則不進行加權。

(4) 匯入關聯資料庫

為了更有效數據分析、圖書推薦及資料管理，將前述步驟分析產生的結果及主題模型透過關聯式資料庫來進行儲存。

## 二、文字資料前置處理

文字資料使用 Columbia University Press 出版社(CUP) 1,324 本西文圖書，每一本書均包含書目記錄、摘要、目次及全文資料，文字資料檔案大小約為 1.8GB。全文資料的部份是萃取自出版社提供的電子書原始 PDF 檔案，至於書目資訊則是出版社另外提供的 Meta-data。在所有 1,324 本的圖書中共有 1,128 本書含有主題標目的訊息，其中共找出 1,789 種主題及 3,086 條欄位記錄，由於一筆書目當中可能包含數筆主題標目的資訊，在本實驗中是將每一個主題標目視為獨立的項目來計算數量，所以有可能同一個分類在單筆書目中被記錄多次。另外主題標目亦可以階層的方式來做顯示，不過在本實驗中僅使第一層的主題標目資料來進行分析。我們透過分類號進一步觀察各分類所佔的比重來了解圖書的內容分佈，所有圖書 1,324 本中共有 1,319 本書標註有 LCC 的分類資訊，其中約 50% 的圖書被歸類為 Language and Literature 及 Social Sciences 的分類。

在(Magdy & Darwish, 2008)的研究中將圖書內容分為幾個部份 1. BC (Book Content,全文)、2. BH (Book Heading,每頁第一列內容)、3. TOC (Table of Content,目次與關鍵字索引頁，若沒有目次則取圖書全文前 3,000 個字元)及 4. BT (Book Title,圖書主題)，透過結合不同區塊的資料來評估不同組合的檢索效率。同時其在研究中指出使用 BC 建立的索引相較於使用 BH+BT 組合建立的索引，其檢索效率差異程度在 20%以下，但是在索引檔案大小的差異上確超過 95%。同時在其實驗結果中顯示單純使用 BC 建立的索引其檢索效率相近於使用 BH+TOC+BT 建立索引的組合，另外若單純使用 TOC 來建立索引則無法產生良好的檢索效率，其可能是由於 TOC 中所隱含的字數頻率過少的影響。

在本研究中首先嚐試使用 BT+TOC+BH 的方式來建立文字資料，透過專家的觀察與評估其產生的主題並沒有明顯較佳，同時花費更多的時間在主題模型的建立上，所以後來的實驗調整成使用 BT+TOC 並配合圖書的摘要來建立文字匯總的資料，使用摘要最主要的因素是考量其當中含有較多圖書內容描述性的文字，同時亦可補強部份圖書當中 TOC 包含資訊過少的缺點。在本研究中文字的處理分為三個步驟:

1. 選擇圖書文字內容: 依(Magdy & Darwish, 2008)的方式採用 BT+TOC 兩個部份的文字資料，另外本研究中亦加上圖書的摘要，這一個部份亦可避免部份圖書文字過少而無法產生具代表性的主題。
2. 過濾文字內容: 將文字中包含的 Stopwords、數字、字元數小於 3、標點符號、出現頻率較高及已知的無效詞彙移除。這一個過程需要在主題模型建立完成，檢視輸出的結果，重覆進行調整與實驗。
3. 文字的轉換: 將第二步驟產生的文字資料，整理成可被 LDA 工具所接受的格式。在處理的過程中亦需記錄文字資料與書目之間的對應，做為後續資料匯入及分析的參考。

在(Newman, Hagedorn, Chemudugunta, & Smyth, 2007)的研究中指出若文字資料沒有經過適當的清理步驟，LDA 所產生的主題可解讀性會降低。可以利用反覆檢視主題模型輸出結果，由文字資料中移除無效的字彙來增加主題的可用性，經過適當的文字內容清理後，由 1,324 本書當中共產生 3,406,224 個字彙，檔案大小約為 25MB，做為 LDA 主題模型建立的文字資料來源。

三、使用記錄前置處理

本研究中採用的使用記錄取自 IG Publishing 公司(IGP) 所提供的電子書平台，相較於其他的 ebook aggregator 例如: eBrary, MyiLibrary 或 EBL 所開發的電子書檢索系統，在 IGP 的電子書平台架構中將每一個出版社視為是獨立產品與服務，所以使用記錄是由 IGP 的 Columbia University 電子書平台上直接取得。本研究採用國立中山大學 2010 整年度的使用記錄做為實驗數據，並使用 COUNTER REPORT BR1 所規範的方法來處理使用記錄，也就是依每一本書被要求並開啟的次數彙總，並以月為彙總單位。而主題的加權即是使用每本書的彙總開啟次數來做計算。

四、LDA 參數選擇

建立 LDA 主題模型需要設定 Hyper-Parameter 的 $\alpha$ 及 $\beta$ 值，同時需要決定主題的數量，由於主題的好壞需要透過產生結果的觀察來做決定，所以在本實驗當中使用 LDA 工具的預設值，即是設定 $\alpha$=50/K、$\beta$=0.1 來執行。另外主題數量的部份則是採用(Griffiths & Steyvers, 2004)建議的方法來計算主題模型的 perplexity 值，當 Perplexity 的值越低時，則該主題數量就是最適當的一個值，由圖 5 可以觀察到，當主題數量約等於 100 時則 perplexity 的值開始趨於平緩。在(Newman, et al., 2007)的研究中亦指出選擇主題數量是一個 Trade-off 的過程，設定過多或過少都會讓主題無法良好表達內容。本研究中我們設定主題的數量為 100 除了基於內容的表達能力之外同時考量到實驗數據的大小、硬體設備及程式效能的限制而決定。

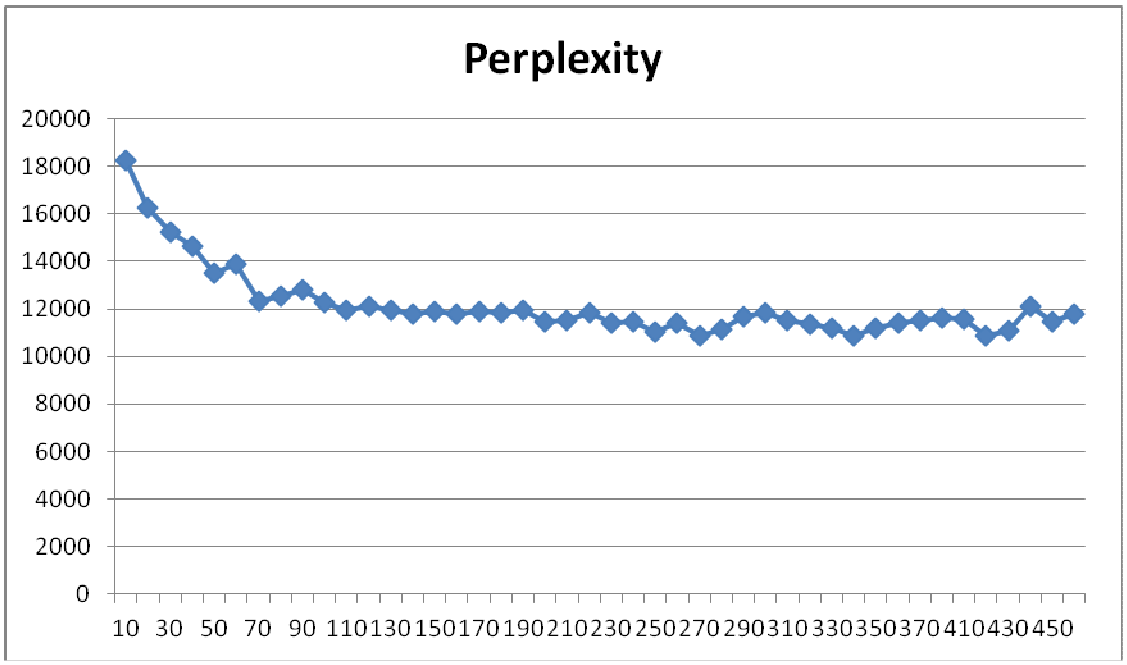


圖 5 主題模型 Perplexity 的計算結果

五、主題模型建置

在本研究中使用的工具是 JGibbLDA，它採用 Gibbs Sampling 的方式來進行主題模型的建置，使用的程式語言是 Java，使用 JGibbsLDA 工具的最主要考量是其開放

原始碼，所以可以透過程式碼來理解其運作的流程及資料的處理。同時程式使用 JAVA 語言來撰寫，除了跨平台的特性之外，未來亦方便未來整合到正式的系統上來使用。由於主題模型的建立時間較長，所以亦可依 Iteration 的執行次數設定階段儲存。在 Griffiths 與 Steyvers (Griffiths & Steyvers, 2004)的研究中指出 Iteration 執行的次數並非越多越好，我們可利用工具所提供的階段儲存功能觀察 主題模型 的變化，若已呈現收斂狀態就可以停止工具，採最後一次的執行儲存結果來使用。

表 1 為主題模型建置完成，依機率分佈總合排名前十的主題，由這一個表中可以觀察到主題內容亦比較偏向 Language and Literature 及 Social Sciences 的分類，這兩個分類也是圖書內容中所佔比例最高的項目。另外透過主題所產生的字彙例如 Social Culture、Political、encomic、literature、History 等，我們也可以發現 Columbia University 圖書的內容偏重於人文科學及社會科學方面的主題。嘗試由主題的字彙來識別其所代表的意義，雖然並非所有的主題都能夠給予適當的標籤，不過我們還是可以由字彙當中來猜測其可能代表的意義。

表 1 依機率分佈排行前十名的主題

| 主題編號 | 主題字彙 |
|---|---|
| 73 | see women life history literature language fiction culture literary social modern cultural work writing between nature book world self death |
| 82 | john see william american james new robert george charles thomas richard david henry paul joseph edward war michael world york |
| 22 | see political social national women party labor state rights new education politics movement economic government society public class reform war |
| 91 | social see work care family services health practice case community child welfare development service children abuse assessment research programs management |
| 64 | see theory philosophy critique history hegel marx self language heidegger political german kant culture nietzsche jean adorno world paul power |
| 4 | see theory science human language natural philosophy psychology nature scientific social analysis anesthesia behavior mind systems problem view physics definition |
| 25 | see trade economic international market bank development capital investment business policy financial industry world tax foreign production oil system markets |
| 17 | gay see sexual women lesbian sex social family men children marriage relationships gender lesbians identity families class violence male parents |
| 81 | policy war act see american committee national foreign economic plan reagan elections administration conference campaign john public kennedy nixon bill |
| 92 | see china soviet chinese war relations policy taiwan foreign russian military revolution communist ccp union east korea russia treaty sino |

六、LDA 主題加權

主題加權的方式是依 COUNTER REPORT BR1 的統計資料為基礎，分別計算每一個單位、每個月份其訂購的圖書當中，被開啟過的圖書的次數統計，針對統計數據

的加總進行正規化(normalize)之後得到各別圖書的使用率比重。主題模型加權透過各別圖書的使用率比重相乘於各圖書包含的主題之機率來調整其重要性。

主題模型加權演算法執行方式如下:
M: 主題數目
N: 訂閱的書本數目
W: 一維陣列儲存每一本書的使用機率
D: N×M 陣列儲存每一本書屬於每一個主題的機率
B: 一維陣列儲存使用加權後的每一主題機率
T: 訂購的時間範圍

```
SET M to the number of topics  // 100 in our experiment
SET N to the number of titles subscribed by a given library
SET W to an array storing usage ratios of a title in a given period
SET D to a N×M matrics storing the topic probability of a title
SET B to an array storing the usage-weighted probability of a topic
Input M, N, W, D
Output B
Begin
        For j=0→M do

    B[j] = 0
                For i=0→N do

                        B[j] = W[i]×D[i, j]+B[j]

                End
        End
        Return B
End
```

主題加權演算法的運作模式是基於使用單位，於特定的時間範圍內，所定購的所有圖書，透過迴圈取得每一本圖書 b 在該時段的使用率比重，然後再依 b 所包含的所有主題來進行加權的操作。加權後產生的結果將寫回關聯式資料庫當中做保存。表 2 為加權後主題的變化，使用 2010 整年度的使用記錄來進行加權產生的結果。

表 2 加權後排行前十名的主題 (依 2010 年使用記錄加權)

| 主題編號 | 主題字彙 |
|---|---|
| 10 | see literature china taiwan chinese japanese literary wang japan new culture zhang cultural qigong chen journal women poetry taiwanese modern |
| 73 | see women life history literature language fiction culture literary social modern cultural work writing between nature book world self death |
| 64 | see theory philosophy critique history hegel marx self language heidegger political german kant culture nietzsche jean adorno world paul power |
| 25 | see trade economic international market bank development capital |

| | |
|---|---|
| | investment business policy financial industry world tax foreign production oil system markets |
| 24 | see shih wang chi ching chu ming chih ang ing tzu chang yuan yang hsi shu liu chou ien eng |
| 82 | john see william american james new robert george charles thomas richard david henry paul joseph edward war michael world york |
| 22 | see political social national women party labor state rights new education politics movement economic government society public class reform war |
| 4 | see theory science human language natural philosophy psychology nature scientific social analysis anesthesia behavior mind systems problem view physics definition |
| 92 | see china soviet chinese war relations policy taiwan foreign russian military revolution communist ccp union east korea russia treaty sino |
| 3 | jewish see jews judaism american israel family hebrew torah education jacob orthodox rabbi life synagogue school ben joseph abraham new |

當某個主題出現在所有圖書的機率越高時，代表越多的圖書與該主題有關。因此，高機率的主題可以代表這群圖書主要的內容描述，其重要性也較高，我們依照主題的機率值排序並比較加權前、後的排序差異。表 3 顯示，排序的結果中未加權的主題其編號 73 排名第在 1 位，但是在加權之後主題編號 10 卻變成了第 1 名。同時在結果當中除了主題編號 10 是由第 28 名晉升為第一名之外，亦可發現主題編號 24 及 3 由後面的名次提升到前十名。

表 3 比較加權前與加權後的主題 (2010 整年度)

| 加權前主題 | | 加權後主題 | |
|---|---|---|---|
| **原始排名** | **加權前主題** | **原始排名** | **加權後主題** |
| 1 | 73 | 28 | **10** |
| 2 | 82 | 1 | 73 |
| 3 | 22 | 5 | 64 |
| 4 | 91 | 7 | 25 |
| 5 | 64 | 25 | 24 |
| 6 | 4 | 2 | 82 |
| 7 | 25 | 3 | 22 |
| 8 | 17 | 6 | 4 |
| 9 | 81 | 10 | 92 |
| 10 | 92 | 39 | **3** |

## 肆、統計分析與實驗結果

本節觀察並分析加權後主題的變化，依每月的使用記錄產生主題變化趨勢圖。另外也將比較 LCC 分類法、LCSH 主題標目與 LDA 主題模型之間的關聯。其結果分述於後。

### 一、主題加權結果觀察

依電子書檢索平台所產生的使用記錄，透過加權的方式來改變主題模型的機率分佈，其結果影響了主題重要性，透過機率值來進行排序後我們得到表 4 的內容。我

們依排名前十名的主題當中所包含的字彙，由專家給于適當的標籤做為主題項目的識別。

表 4 加權後主題對應標籤表

| 加權後主題 | | |
|---|---|---|
| **排名** | **加權主題** | **主題標籤** |
| 1 | 10 | Chinese Literature |
| 2 | 73 | Political Biography |
| 3 | 64 | Philosophy |
| 4 | 25 | Economic Financial |
| 5 | 24 | Chinese Novel |
| 6 | 82 | Bible Story |
| 7 | 22 | Political Economic |
| 8 | 4 | Neuroscience and Philosophy |
| 9 | 92 | Chinese Political |
| 10 | 3 | Jewish History |

由圖 6 的全年度主題變化趨勢可以觀察到只有在特定的月份產生較高的變化，其主要是受到使用率加權主題的影響，例如主題編號 25 (Economic Financial) 只有在六月份有較高的機率值，主要是受這一個月份多為學生準備期末考試或繳交報告的時間，但是相對於其他月份來說則幾乎沒有任何顯著的改變。另外我們再觀察主題編號 10 (Chinese Literature) 可以發現這一類的書在 9 月份之後的機率持續保持較高的比例，可能的理由為開學之後圖書資源的使用率較高，而 Chinese Literature 比較顯著的因素除了中山大學設有中文系所之外，亦和 Columbia University Press 所收錄的內容範圍有相關，在所有圖書當中 P(LANGUAGE AND LITERATURE)在所有圖書中佔有最高的比例，同樣的狀況也可在圖七的主題編號 92 (Chinese Political) 上面發現。

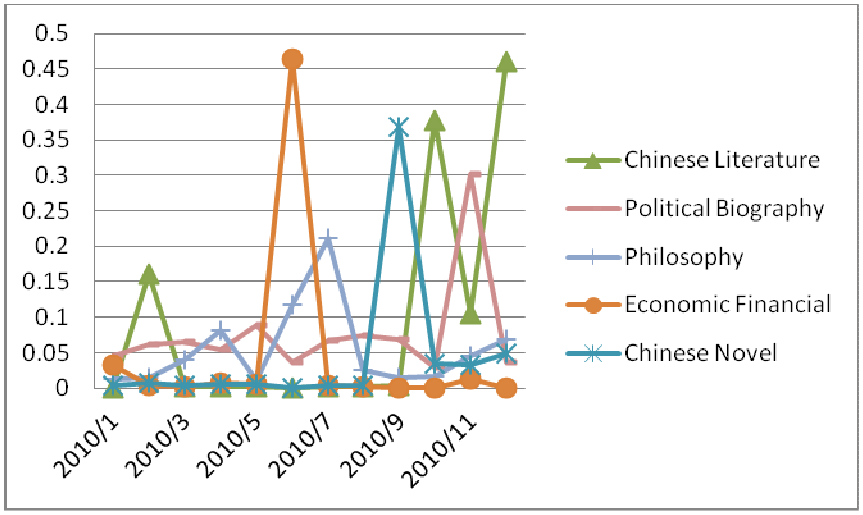


圖 6 主題編號 10, 73, 65, 25, 24 全年度變化趨勢圖

我們再觀察圖 7 呈現的趨勢，它並沒有像圖 6 產生如此明顯的變化，除了主題編號 92 之外，其他的主題趨勢相對來說較為平緩，主題編號 92 的變化除受圖書內容所影響之外，同時也與讀者的閱讀喜好有相關，我們另外可以發現主題編號 82、22、4 及 3 在我們分析的內容中佔有一定的比例值，但是並沒有特別的顯著。

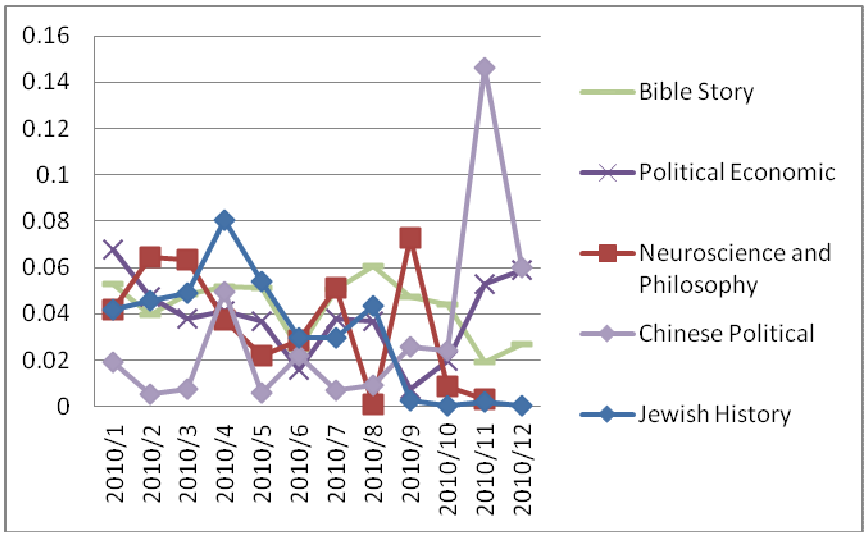


圖 7 主題編號 82, 22, 4, 92, 3 全年度變化趨勢圖

進一步再比較未加權與加權主題之間機率的變化，透過機率累計的方式來做比較。累計的計算方式是先針對未加權主題的機率進行排序，由大至小，依序由 1,324 本書中取出主題進行加總，然後，再取其平均值而得到每一個主題的平均機率做為比較的基準線。而計算加權後主題機率累計的方式，則是透過加總使用單位訂購的時間範圍中，曾經被使用過的圖書，加權後的機率值，同樣的由大至小、依序由使用過的圖書中取出主題進行加總，由於在使用記錄加權處理時已做過正規化，所以其機率值不需要再取平均值。

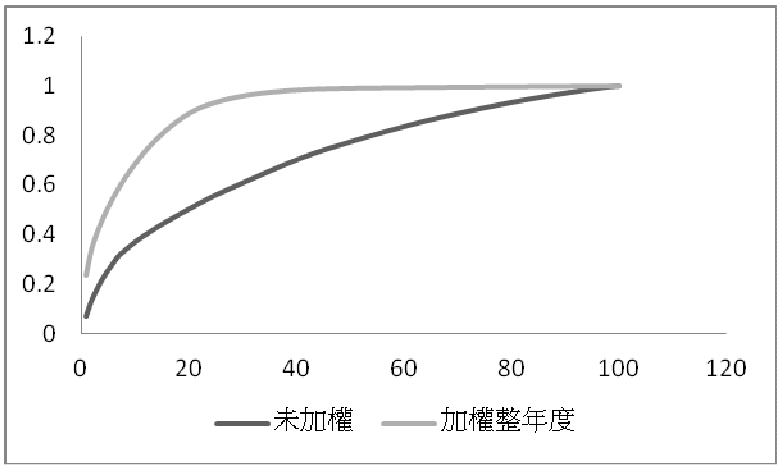


圖 8 未加權與加權主題機率累計比較 (整年度)

本實驗中我們使用 100 個主題，透過圖 8 可以觀察到加權後的機率值產生很明顯的變化，當累計的主題越多時則變化的幅度越小，我們可以發現上圖在 25 個主題之後曲線就變的平緩，所以前 25 個主題可以在趨勢上應該也會有較明顯反應，另一方面我們再比較每一個月份的主體機率累計相對於基準線亦可發現累似的狀況產生。由此可以看出，透過使用加權後，各主題的機率差異性明顯提升。因此較少的主題即可涵蓋較多的累計機率。

二、LCC 與主題模型關聯性

書目資料中包含了 LCC 分類號，這一個欄位是由出版社或圖書館員依據圖書的內容所賦予，所以我們可以透過分類號來識別圖書的主要分類。在本研究比較圖書內容建立的主題是否與 LCC 分類號之間存在關係，同時也進一步的觀察是否能夠使用主題來決定圖書的分類。我們透過 LCC 相對於圖書主題資訊熵(Information Entropy)的計算來觀察兩者之間是否有明顯的關係存在。下列的方程式是用來進行某一個分類號的資訊熵計算，其中 $p_i$ 為該分類號在第 i 個主題的分佈機率，n 為 Topic 的總數。

資訊熵的計算 $H(X)=-\sum_{i=1}^{n} p_i \log_2(\frac{1}{p_i})$

計算的結果如表五所示， LCC 分類與主題之間沒有固定的規則，每一個類別的資訊熵都不同，如果主題數量設定越多則亂度越高，資訊熵越大。同時圖書的類別也會影響到這一個數值，我們可以發現 S(AGRICULTURE) 類別的資訊熵較小，代表它跟 LDA 所分的某些主題較接近，而 G(GEOGRAPHY. ANTHROPOLOGY. RECREATION)及 P(LANGUAGE AND LITERATURE)則無法用 LDA 的主題來代表。另外 S 類別的亂度隨著主題增加而變大，所以可以推論過多的主題分散了每一個主題的特徵值，所以在類別的識別上的適用性就會降低。

表 5 依 LCC 分類第一層計算主題模型的資訊熵

| 項次 | 分類號 | 數量 | 主題數量/資訊熵 (Entropy) | | | |
|---|---|---|---|---|---|---|
| | | | 100 | 50 | 25 | 10 |
| 1 | A | 2 | 3.294328 | 2.911226 | 2.991528 | 1.889249 |
| 2 | B | 48 | 5.010965 | 4.132428 | 3.34721 | 2.390229 |
| 3 | C | 10 | 4.284662 | 4.238595 | 3.573282 | 2.635353 |
| 4 | D | 131 | 5.303272 | 4.576113 | 3.895108 | 2.330363 |
| 5 | E | 53 | 4.245757 | 3.782793 | 3.221278 | 2.130033 |
| 6 | F | 17 | 4.033627 | 3.804693 | 3.616406 | 2.375371 |
| 7 | G | 53 | 5.215972 | 4.759288 | 3.996787 | 2.888559 |
| 8 | H | 264 | 4.924049 | 4.420709 | 3.643541 | 2.433152 |
| 9 | I | 1 | 2.232357 | 1.66878 | 1.583531 | 0.950275 |
| 10 | J | 58 | 4.351098 | 3.933277 | 3.070711 | 2.133933 |
| 11 | K | 23 | 3.82121 | 3.510087 | 3.053169 | 2.10732 |
| 12 | L | 7 | 3.207077 | 2.117627 | 2.05282 | 1.103106 |
| 13 | M | 14 | 3.165828 | 3.053393 | 1.958364 | 1.847804 |
| 14 | N | 15 | 4.644991 | 4.141896 | 3.512223 | 2.683227 |
| 15 | P | 287 | 5.507219 | 4.665999 | 3.953866 | 2.710985 |
| 16 | Q | 97 | 4.518889 | 3.723138 | 2.646837 | 1.829386 |
| 17 | R | 35 | 3.530006 | 3.335437 | 2.256981 | 1.627877 |
| 18 | S | 10 | 2.214694 | 1.470081 | 1.047311 | 0.887751 |
| 19 | T | 21 | 4.26008 | 3.558285 | 3.064709 | 2.761337 |
| 20 | U | 19 | 3.682151 | 2.745754 | 2.347146 | 1.616622 |
| 21 | Z | 5 | 3.08241 | 3.118979 | 2.998983 | 2.496377 |

LCC 與主題模型的相關度部份使用了 Chi-Square 獨立性檢測，使用主題數 100、50、25 及 10 個進行檢定，在 alpha=0.1 或 0.05 的狀況下均拒絕虛無假設，LCC

與主題機率間並不相關。所以透過主題可以發掘圖書分類法未包含到，但的確和圖書內容非常相關的資訊。

三、LCSH 與主題模型關聯性

書目資料中的主題標目是由圖書館專業人員，依據圖書內容，由控制的詞彙當中給于書目記錄主題標籤，在(Noh, Hagedorn, & Newman, 2011)的研究中指出 LCSH 的主題與 LDA 主題模型是具有相關性的。本實驗使用資訊熵的計算方式，取發生頻率最高的前 20 個 LCSH 主題標目，依不同的主題數目來計算主題標目的亂度，我們的實驗結果顯示主題數量越多則產生的亂度越高，越不利於分類的進行。另外在 LCSH 主題的部份 Psychoanalysis 及 Social Service 及的亂度值最低，而 Jews 及 Women 主題的亂度值最高，分類效果可能最差。

LCSH 與主題模型相關性的部份，同樣使用 Chi-Square 獨立性檢測，在本實驗中使用出現頻率最高的前 20 個 LCSH 主題項目來做計算，由於過多的主題數量得到的卡方值 $\chi^2$ 與自由度 DF 過大無法進行比較，所以將主題的範圍設定在 10、25 及 50 個。實驗結果顯示當 $\alpha$=0.05 時、不論主題數量是 10、25 或者是 50 都無達到顯著的水準，所以 LCSH 和主題模型之間並不相關，與 (Noh, Hagedorn, & Newman, 2011) 的研究產生差異的最主要因素是評估的方式不同，本研究中僅使用 Chi-Square 來進行檢測但是並未透過使用者來做評分。

LDA 方法所產生的主題排除了人為主觀因素，依文字內容為基礎，所以能反應圖書的內容，適合當做輔助參考資訊。在(Noh, et al., 2011)的研究中也指出 LDA 主題模型可以做為另一種描述圖書主題內容的方式，同時當圖書內容中含有多種的主題，書目中的 LCSH 主題內容範圍較廣時，則使用 LDA 方法產生的主題更能反應圖書的資訊，更具有參考的價值。

## 伍、結論與未來研究建議

在本研究中以使用記錄來加權主題模型，同時使用卡方分配獨立性檢定比較美國國會圖書館分類法及國會圖書館主題標目與主題模型之間的相關性。我們發現依使用記錄加權後的主題，明顯改變了主題的機率，進而影響了主題的趨勢變化。主題趨勢讓圖書館透過另一個角度來觀察使用者的行為，產生有別於一般統計報表的資訊，同時主題相較於書目資料中的圖書館分類法與主題標目，LDA 產生的主題能獨立於書目之外，提供一個有價值的參考資訊。

LDA 主題模型亦可以用來豐富書目資料，主題所歸納出來的標籤亦可於資料的檢索，或協助建構圖書分類，增加資源的利用率(Newman, et al., 2007)。在本實驗中嘗試設定不同的主題數目及內容區塊來進行分析，發現分析的過程需耗費很長的時間，所以在有限的硬體資源及環境下只針對 Columbia University 1,324 本書來建立主題模型。未來若要整合主題模型至實際運作的電子書檢索平台上需考量到圖書資料的更新及轉換程式的效率，所以無法像實驗環境一樣花費數天的時間來完成單一出版社的主題模型，所以主題數量及參數的選擇影響到主題的產生品質、執行效率及系統的實用性。

在主題模型建置前需事先設定 α、β 參數與主題數量，當來源文字資料越龐大、主題數目設定越多則程式需耗費的時間越長。在許多研究中的顯示 α、β 值與主題的數量影響了主題產生的品質，雖然透過 Perplexity 的計算可以協助選擇統計上最佳的主題數量，

但是對於主題內容的解讀反而可能造成反效果(Chang, et al., 2010)，所以主題數量的選擇要考量到實際運用面的需求，針對不同的內容做調整，經過多次的實驗，由主題結果中來觀察並決定模型的好壞，或者是透過領域專家的協助來決定主題的好壞。

在主體加權的方面，本研究尚未透過讀者的經驗來判斷加權主題相對於於使用統計的相關性，僅能透過比較的方式來顯示加權前後主題的差異，所以在未來的研究中可以嘗試以問卷來搜集使用者對於主題重要性的看法，進一步來驗證主題加權方法的實用性。在與 LCC 和 LCSH 比較方面，本研究的卡方檢定顯示加權主題模型與 LCC 和 LCSH 有明顯的差異，但對於特殊的應用，比如推薦方面，是否能有幫助，仍需做進一步的探討。

## 參考文獻


1. Anthes, G. (2010). Topic models vs. unstructured data. *Commun. ACM, 53*(12), 16-18. doi: 10.1145/1859204.1859210
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res., 3*, 993-1022. doi: 10.1162/jmlr.2003.3.4-5.993
3. Chang, J., Boyd-graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2010). Reading Tea Leaves: How Humans Interpret Topic Models %U http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.992.
4. COUNTER - Counting Online Usage of Networked Electronic Resources. from http://www.projectcounter.org/
5. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5228-5235. doi: 10.1073/pnas.0307752101
6. Khosh-khui, S. A. (1987). *Relationship Between LCSH and LCC Notationsin Different Classes of LCC.* Staff Publications-Library, Texas State University.
7. Library of Congress Classification. from http://www.loc.gov/catdir/cpso/lcc.html
8. Magdy, W., & Darwish, K. (2008). *Book search: indexing the valuable parts*. Paper presented at the Proceeding of the 2008 ACM workshop on Research advances in large digital book repositories, Napa Valley, California, USA. http://dl.acm.org/citation.cfm?doid=1458412.1458429
9. Newman, D., Hagedorn, K., Chemudugunta, C., & Smyth, P. (2007). *Subject metadata enrichment using statistical topic models*. Paper presented at the Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, Vancouver, BC, Canada.
10. Noh, Y., Hagedorn, K., & Newman, D. (2011). *Are learned topics more useful than subject headings*. Paper presented at the Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries, Ottawa, Ontario, Canada.
11. Shepherd, P. T. COUNTER: towards reliable vendor usage statistics. [Conceptual Paper]. *VINE, 34*(4). doi: 10.1108/03055720410570975