國立中山大學資訊管理學系

碩士論文

Department of Information Management

National Sun Yat-sen University

Master Thesis

以 LDA 和使用紀錄為基礎的線上電子書主題趨勢發掘方法

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

研究生：洪崇洋 撰

Chung-Yang Hung

指導教授：黃三益 博士

Dr.San-Yih Hwang

中華民國 101 年 1 月

Jan 2012

國立中山大學研究生學位論文審定書

本校資訊管理學系碩士在職專班

研究生洪崇洋（學號：N984020017）所提論文

以LDA和使用紀錄為基礎的線上電子書主題趨勢發掘方法
An Approach to eBook Topics Trend Discovery Based on LDA and Usage Log

於中華民國 101 年 1 月 12 日經本委員會審查並舉行口試，符合碩士學位論文標準。

學位考試委員簽章：

召集人 陳嘉玫 陳嘉玫　　委　員 黃三益 黃三益

委　員 張德民 張德民　　委　員 _____

委　員 _____　　委　員 _____

指導教授(黃三益) 黃三益 (簽名)

致謝

進入職場十餘年後能夠重新回到學校是一件非常值得欣慰的事情，能夠進入中山資管就讀更是我生命中的一個驚喜。在這兩年半的求學時光中我得到了新的知識、想法與概念、認識了中山的好老師、結交了許多的朋友、我的第一個孩子誕生、第二個孩子也即將來到世上...雖然在職進修的過程當中發生了許多的事情，家庭、工作與學業無法很好的兼顧，但是我很高興能夠做了這一個選擇，同時我相信這是一個正確的決定。

畢業是另一個階段的開始，我對家人有更多的責任，我對工作需要付出更多的努力，我對自己的未來有更多的期望。最後要感謝我的太太在我求學過程當中能體諒並支持我的決定、感謝黃三益老師細心指導，在我遇到挫折時不斷鼓勵，感謝我的老闆Dr. Aung同意讓我回學校唸書、同時也要感謝所有中山資管的老師、課程當中一起努力過的同學及學弟的熱心協助，讓我在這段求學過程的最後劃下一個好的句點。

洪崇洋 謹識

中山資管所

論文提要

學年度: 100

學期: 1

校院名稱: 國立中山大學

系所名稱: 資訊管理研究所

論文名稱(中): 以 LDA 和使用紀錄為基礎的線上電子書主題趨勢發掘方法

英文名稱(英): An Approach to eBook Topics Trend Discovery Based on LDA and Usage Log

學位類別: 碩士

語文別: 中文

學號: N984020017

提要開放使用: 是

頁數 : 48

研究生(中)姓: 洪

研究生(中)名: 崇洋

研究生(英)姓: Hung

研究生(英)名: Chung-Yang

指導教授(中)姓名: 黃三益

指導教授(英)姓名: San-Yih Hwang

關鍵字(中): 電子書、使用記錄、主題模型、主題、LDA、LCC、LCSH

關鍵字(英): Ebook, Usage Log, Topic Model, Topic, LDA, LCC, LCSH

摘要

網際網路的發展及科技的進步讓數位內容產業日漸蓬勃，出版業者紛紛開始提供線上電子書檢索、閱讀及下載服務，使用者不受地域或時間的限制，隨時隨地都能使用電腦來閱讀數位內容，另外一方面圖書館購買電子書做為館藏的比例亦逐年增加。使用電子資源的方式，可透過連線到電子書檢索平台或透過圖書館自動化系統檢索，由館藏目錄中直接鏈結至電子書平台進行使用。這一個方式相較於實體館藏來說沒有流通數量上的限制，同時提昇了圖書資源的利用率。

提供電子書檢索服務的出版社或系統整合業者眾多，圖書內容包羅萬象，考量到有限的預算條件下，圖書館採購電子書除了參考讀者的推薦之外亦需要評估電子資源的使用率，做最有效率的投資。目前最普遍的方式是使用統計報表，其通常由出版社所提供。

本研究使用 Latent Dirichlet Allocation 簡稱 LDA 的方法，基於圖書的內容來建置主題模型，然後結合電子書檢索平台的使用統計報表，運用主題模型的加權來發掘電子書讀者閱讀主題的變化，進而提供一個具參考價值的訊息。我們在實驗中並比較了其他兩種方式：美國國會分類法和主題標目法。實驗結果證實透過主題加權方法產生的主題模型與其他兩種方法顯著不同，可以提供另一方面的有用資訊。

關鍵字: 電子書、使用記錄、主題模型、主題、LDA、LCC、LCSH

Abstract

With the growth of digital content industry, publishers start to provide online services for ebook search, reading and downloading. Users can access to online resources from anywhere, any place with laptop or mobile devices at any time. 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. Compared to the library circulation process, ebooks are more convenient to patrons and improve the utilization of library online resources.

There are various kinds of ebooks available in the market, so libraries have to focus their investment on the most valuable online resources. Usage statistics report plays an important role in providing valuable information to libraries. It is usually based on the standard of COUNTER to generate the statistic reports, although it provides when and where users access to specific ebooks, it fails show the general topics and how they change.

In this study, we introduce a post process method to weighting the LDA topic model via the usage statistic report to emphasize the changes of topic and compare it to the classification method and subject heading method in the bibliographic, namely LCC and LCSH respectively. The result show that weighted topic model significantly affect the ranking of topics, and the topic model are independent from the classification method and the subject heading method in the bibliographic record.

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

目錄

圖目錄

表目錄

第一章　諸論

1.1 研究背景

近年來政府積極推動文創產業及數位內容的發展，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，減少了庫存管理上的壓力、減化了書籍流通過程，增加了資源的利用率，透過網際網路同時間可以讓多位使用者閱讀電子書內容。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。

依台灣學術電子書聯盟為例,截至2011年為止已有 87個學術單位成為會員，依採購的電子書量統計(2008-2010)，如表一-1所示，每年採購數量都保持在一萬本以上。

表 一-1 TAEBC 2008-2010 採購圖書統計

購案型式	2008 冊數	2009 冊數	2010 冊數
Collection	10,043	10,667	12,411
Titles	6,762	4,152	3,007

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。

1.2 研究動機

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例如

圖一-1 顯示的格式。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。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	BR1 : Number of Successful Title Requests by Month and Title																	
2	Title	Publisher	Platform	ISBN	Type	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	Total
3	A calculus of suffer	columbia unive	ebridge	023105	request	9	10	7	15	4	4	5	12	0	0	0	0	66
4	A chinese village: ta	columbia unive	ebridge		request	9	4	3	6	4	2	3	5	0	0	0	0	36
5	A communion of su	columbia unive	ebridge	978023	request	3	7	8	6	3	5	5	12	0	0	0	0	49
6	A cultural history of	columbia unive	ebridge	023106	request	9	7	6	6	4	1	7	10	1	0	0	0	51
7	A derrida reader: be	columbia unive	ebridge	978023	request	8	6	10	8	2	2	8	11	0	0	0	0	55
8	A field of honor: writ	columbia unive	ebridge	978023	request	6	7	5	8	6	4	4	11	0	0	0	0	51
9	A framework for imr	columbia unive	ebridge	023112	request	9	5	5	7	4	0	3	10	0	0	1	0	44
10	A genetic and cultu	columbia unive	ebridge	978023	request	7	3	6	7	4	7	7	11	0	0	0	0	52
11	A guide to oriental c	columbia unive	ebridge	023106	request	8	12	5	8	3	1	4	7	0	0	0	0	48
12	A history of greek li	columbia unive	ebridge	023101	request	4	7	9	3	6	2	5	7	0	0	0	0	43
13	A history of housing	columbia unive	ebridge	023106	request	7	5	6	3	3	4	7	11	0	0	0	0	46
14	A history of latin lite	columbia unive	ebridge	023101	request	8	7	7	11	4	2	3	10	0	0	0	0	52
15	A limited partnershi	columbia unive	ebridge	023112	request	9	7	2	7	3	2	7	12	0	0	0	0	49
16	A modern heretic ar	columbia unive	ebridge	023110	request	10	8	7	14	4	5	4	9	0	0	0	0	61
17	A natural history of	columbia unive	ebridge	023112	request	11	7	5	5	5	5	3	10	0	0	0	0	51
18	A partisan century:	columbia unive	ebridge	978023	request	5	4	5	10	1	3	5	11	0	0	0	0	44
19	A popular guide to c	columbia unive	ebridge	023104	request	4	9	5	9	5	4	6	5	0	0	0	0	47
20	A possible peace b	columbia unive	ebridge	978023	request	10	4	9	9	1	2	6	8	0	0	0	0	49
21	A private life	columbia unive	ebridge	978023	request	7	9	7	5	7	1	7	14	0	0	0	0	57
22	A revolution in eatin	columbia unive	ebridge	978023	request	2	0	0	0	0	0	0	0	0	0	0	0	2
23	Aesthetic nervousn	columbia unive	ebridge	978023	request	0	0	0	0	0	0	0	0	0	11	0	1	12
24	Agents of bioterroris	columbia unive	ebridge	978023	request	0	0	0	0	0	0	1	0	0	0	0	0	1

圖 一-1 COUNTER REPORT 圖書統計報表範例

對於出版社來說若能夠了解讀者的喜好，可以協助其制定未來收錄內容的規劃，產生差異化的服務，提供客戶更具有策略價值的報表而不是一般的統計數據，與其他競爭者在市場上做出區隔。

1.3 研究目地

近年來 LDA (Latent Dirichlet Allocation) 被廣泛運用在各領域，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，所以本研究討論如何來處理

圖書的文字內容，進而發掘文字中所包含的訊息，透過 LDA 的方法來產生主題模型，並結合圖書使用記錄來進行主題內容加權，產生主題變化趨勢。

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一個輔助的資訊提供給讀者來參考。

1.4 論文架構

本論文的架構分為五個章節，依研究的順序組成。第一章描述目前電子書產業的發展現況，研究的動機，傳統報表所遭遇的問題並提出使用 LDA 的方式來建立主題模型及產生主題趨勢; 第二章說明研究中參考的相關文獻，LDA 模型的構成，主題模型參數的選擇、目前出版社經常採用的電子資源統計標準及圖書的分類方法及主題標目的探討; 第三章說明研究方法、實驗系統的架構，如何結構化處理圖書文字內容、如何依使用記錄產生符合 COUNTER 標準的統計報表、另外並描述 LDA 主題模型的建立及如何利用報表匯整資訊來進行主題的加權; 第四章分析加權前後主題的變化、圖書分類法及主題標目相對於 LDA 主題模型的差異及相關性; 第五章描述本研究的結論，實驗中所遇到的一些困難及將來可以嚐試的研究方向。

第二章　文獻探討

2.1 LDA 主題模型

Latent Dirichlet Allocation 簡稱 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。如圖二-1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 二-1 LDA 主題及包含的字彙以顏色區分、資料來源 (Blei, Ng, & Jordan, 2003)

依圖二-2 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、θ_i 是主題在文件 i 中的機率分佈、φ_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)。

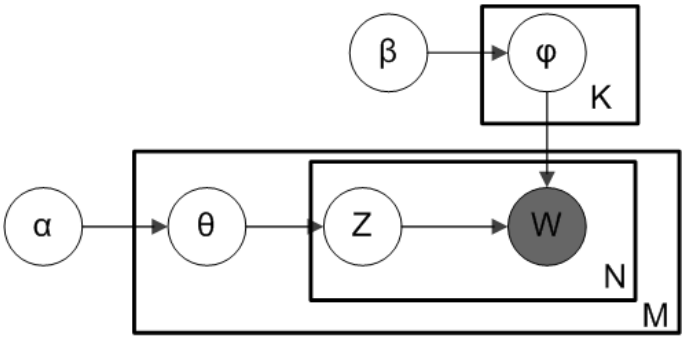


圖 二-2 Graphical Model of the Smoothed LDA Model、資料來源 ("Wikipedia - Latent Dirichlet Allocation,")

依時間資訊來產生 LDA 主題模型的方法、主要採用下列三種方法:

1. 時間演化主題 (Topic Over Time, TOT)，這一個模型把連續的時間資訊引入生成模型，產生話題變化趨勢，但無法對於新的文件進行擴展，必須重新建模(Wang & McCallum, 2006);
2. 動態主題模型（Dynamic 主題模型, DTM）(Blei & Lafferty, 2006) 與線上主題模型 (Online LDA Model, OLDA)(AlSumait, Barbara, & Domeniconi, 2008)，這一個模型是將時間視為離散的狀態，將文件透過時間段進行區分來建模;
3. 第三個方法在整個文件中建模，然後再依每一個文件的時間區段來建立子集，產生出該子集的主題(Hall, Jurafsky, & Manning, 2008)。

本研究中所採用的主題模型加權方法，透過使用記錄來加權主題模型是一種後處理的方法。由於圖書的內容不會經常性的更新，所以依圖書內容所產生的主

題模型可以透過加權的方式來突顯不同時間點主題所產生的變化，而不需重新建置或分析來建置新的主題模型。相對於上述的三種方法來說本研究提出的主題加權方法降低了資料處理的複雜度及資料分析的困難度，同時主題加權所產生的趨勢更適合搭配統計報表做為輔助性參考的資訊。

2.2 LDA 參數的選擇

LDA 的建置需設定 Hyperparameters 的α、β及主題數量。其中α及β的大小對於主題在文件的分佈及主題與字彙的分佈有重要的影響，較大的β值通常會導致較粗糙的主題，較大的α值會導致較粗糙的文件主題分佈，相對來說使用較小的α與β值來提高較特殊主題被發掘的機率。在主題數量的選擇上則可以透過建立完成的主題模型計算其Perplexity值，較小的Perplexity值則建立出來的主題模型最能代表文字的內容(Griffiths & Steyvers, 2004)。而選擇主題數量較多時，Perplexity值較小，但使用者常不易理解。在(Maskeri, Sarkar, & Heafield, 2008)的研究中指出，適當的α與β值的設定應該透過實驗結果來觀察，調整至最合理的數目。

主題數量的選擇透過Perplexity的計算來選擇，要耗費較長的時間來進行測試，同時亦不能保證主題中所包含的字彙都可以被良好的解釋，仍然需要透過專家的評估才能決定最適當的數量。在 (Chang, Boyd-graber, Gerrish, Wang, & Blei, 2010) 的研究，嚐試評估不同主題模型在最佳 Perplexity 數值時的解讀性，這個研究比較了 pLSI、LDA 及 CTM 三個種類的主題模型，透過 Word Intrusion 及 Topic Intrusion 的方式，讓受測者針對主題內容中的字彙或者是文件所包含主題的相關性進行評分，其實驗結果顯示較佳(較小)的 Perplexity 並不代表主題的字彙容易被解讀，甚至會降低字彙可被理解的程度，所以主題數量的選擇應該依真實世界的需求來決定。

在(Sun, Han, Gao, & Yu, 2009)的研究中亦提出利用 Q function 的方式來決定主題的數量。Q function 最主要被應用在評估 Network Clustering結果的好壞，透過資料所產生的 Network Nodes 之間的關聯路徑來計算 Q 值，在LDA主題數的選擇上，若資料是以 Document Network 的方式來組合，則可以使用Q function 的方式來決定主題數量。

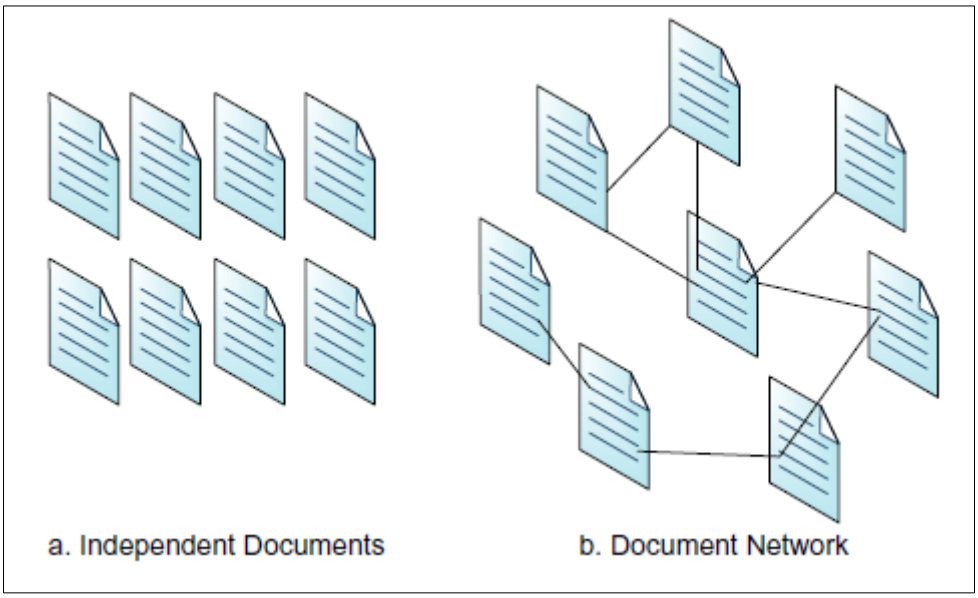


圖 二-3 比較 a.獨立型文件結構與 b.網路型文件結構(Sun, Han, Gao, & Yu, 2009)

2.3 Collapsed Gibbs Sampler

建構 LDA 主題模型的方式通常會透過 Variational Methods 或 Gibbs Sampling 的方式，由於 Gibbs Sampling 的實現方式較容易，所以在許多的研究當中被使用。Gibbs Sampling 是 Markov Chain (MCMC) 的一種實現，也就是在母體機率分佈未知而各別樣本機率已知的狀況下，透過大量的抽樣及演算法的迭代計算，其樣本分佈會逐漸收斂並趨近於母體的機率分佈(Gibbs sampling)。

我們感興趣的是文件內所隱含的主題機率分佈θ_i、文字在主題 k 中的機率分佈φ_k及每一個字被指定為主題的機率Z_{ij}。其中透過Z_{ij}可用來計算出θ_i及φ_k，因為

Z為兩個機率分佈的充份統計量,因此我們可以簡化原 Gibbs Sampling 的演算法，單純由 Z_i 中來進行取樣。這一個方式稱呼為 Collapsed Gibbs Sampler(Darling, 2011)。

實現 LDA collapsed Gibbs Sampler 的步驟包含了變數的設定、隨機初始化、重覆取樣及迭代的計算得知 θ_i 及 φ_k 的值。

Collapsed Gibbs Sampler 演算法表示如下，令:

$n_{d,k}$: 在文件 d 中 word 被指定為 topic k 的次數

$n_{w,k}$: word 被指定為 topic k 的次數

n_k: 任一個 word 被指定為 topic k 的次數

d:單一文件

w: d 其中的一個 word

z: w 所被分配的 topic

N: d 所包含的 word 總數

Input: words **w** ∈ documents **d**

Output : topic assignments **z** and counts $n_{d,k}$, $n_{w,k}$, and n_k

Begin

Randomly initialize **z** and increment counters

Foreach iteration **do**

For i=0 → N -1 **do**

Word ← w[i]

Ntopic ← z[i]

$n_{d,topic}$-=1; $n_{word,topic}$-=1; n_{topic}-=1

For k=0 → K – 1 **do**

$$P(z=k|)=(n_{d,k}+\alpha_k)\frac{n_{w,k}+\beta_w}{n_k+\beta*W}$$

End

topic ← sample from (p|)

z[i] ← topic

$n_{d,topic}$+=1; $n_{word,topic}$+=1; n_{topic}+=1

End

End

Return z, $n_{d,k}$, $n_{w,k}$, n_k

End

2.4 COUNTER 統計報表

線上資源使用快速成長，內容供應商及圖書館皆同意資源使用必需透過一致性的方式來做評估。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。要達成這一個目地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。在電子書方面 COUNTER Book Report 的部份共包含了六種標準的報表格式: (COUNTER - Counting Online Usage of Networked Electronic Resources Home)。

1. Book Report 1: 每月報表，依圖書被開啟成功的次數統計

2. Book Report 2: 每月報表，依圖書章節被成功請求的次數統計

3. Book Report 3: 每月報表，依圖書被開啟失敗，權限不足次數統計

4. Book Report 4: 每月報表，依系統服務類型存取失敗，權限不足次數統計

5. Book Report 5: 每月報表，依圖書執行檢索及存取次數的統計

6. Book Report 6: 每月報表，依服務類型執行檢索及存取次數的統計。

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）。

2.5 美國國會圖書館分類法

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。

分類的結構以階層方式組成，再往下細分可以使用 1-4 碼的單一數字或數字的範圍來設定為更特定的分類，通常是以特殊地點、時間範圍、書目格式等資訊來加以區分。不過在主題與主題之間的數字並沒有直接的關系，這一個分類法採

用較鬆散的結構，所以需透過主題上一層的類別來加以鏈結或識別其中的關聯性(Library of Congress Classification)。

表 二-1 美國國會圖書館分類法、第一層類別

Class	Description	Class	Description
A	GENERAL WORKS	B	PHILOSOPHY. PSYCHOLOGY. RELIGION
C	AUXILIARY SCIENCES OF HISTORY	D	WORLD HISTORY AND HISTORY OF EUROPE, ASIA, AFRICA, AUSTRALIA, NEW ZEALAND, ETC.
E	HISTORY OF THE AMERICAS	F	HISTORY OF THE AMERICAS
G	GEOGRAPHY. ANTHROPOLOGY. RECREATION	H	SOCIAL SCIENCES
I	未使用	J	POLITICAL SCIENCE
K	LAW	L	EDUCATION
M	MUSIC AND BOOKS ON MUSIC	N	FINE ARTS
O	未使用	P	LANGUAGE AND LITERATURE
Q	SCIENCE	R	MEDICINE
S	AGRICULTURE	T	TECHNOLOGY
U	MILITARY SCIENCE	V	NAVAL SCIENCE
W	未使用	X	未使用
Y	未使用	Z	BIBLIOGRAPHY. LIBRARY SCIENCE. INFORMATION RESOURCES (GENERAL)

2.6 美國國會圖書館標題表

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。LCC 分類與 LCSH 的關聯性如表二-2 所示，p 為 LCC 與 LCSH 出現頻率的符合程度，r 為主題標目間的相關性、s 為信心水準值。

表 二-2 Comparison Of Probability, Correlation between LC Subject Headings and LCC Notations in LCC Classes

LCC Main Classes	LCSH f	LCC f	p	r	s
A (General works)	127	53	0.42	0.93	.001
B (Philosophy/Religion)	2006	1481	0.74	0.61	.001
C (...Science of History)	214	137	0.64	0.60	.001
D (History: General)	1349	983	0.73	0.33	.001
E (History: America)	367	285	0.78	0.59	.001
F (History: United States)	395	339	0.86	0.41	.001
G (Geography...)	806	644	0.80	0.46	.001
H (Social Sciences)	3816	2832	0.74	0.49	.001
J (Political Sciences)	746	484	0.65	0.46	.001
K (Law)	1080	614	0.57	0.25	.001
L (Education)	678	543	0.80	0.65	.001
M (Music)	372	271	0.73	0.33	.001
N (Fine Arts)	579	370	0.64	0.83	.001
P (Literature)	2320	1482	0.64	0.29	.001
Q (Science)	2080	1645	0.79	0.59	.001
R (Medicine)	1782	1502	0.84	0.38	.001
S (Agriculture)	293	239	0.82	0.40	.001
T (Technology)	2109	1833	0.87	0.66	.001
U (Military Science)	120	91	0.76	0.82	.001
V (Naval Science)	64	44	0.69	0.80	.001
2 (Bibliography)	418	342	0.28	0.83	.001
ALL Classes	6142	5010	0.82	0.70	.001

第三章　主題模型建立的方法

本章分為六個小節來說明使用記錄的處理、文字資料的轉換、主題模型的建置、儲存及主題的加權方法，各小節內容大綱如下:

1. 描述本實驗的主要架構，其中所包含的四個主要處理步驟概括說明。
2. 說明使用記錄檔案的來源以及格式，同時應該透過什麼樣的方式來做資料的清理並產生符合 COUNTER REPORT 的統計數據。
3. 說明文字資料的來源以及格式，文字內容的過濾條件及處理方法，書目資料及全文內容區塊的選擇。
4. 說明 LDA 主題模型參數的選擇,如何透過 Perplexity 來評估主題的數量，同時說明其他的研究中是如何決定最適合的主題數量。
5. 介紹 jGibbsLDA 工具的使用，輸入檔案格式及輸入檔案格式的處理，及如何使用關聯式資料庫來儲存 LDA 主題模型。
6. 說明主題模型加權的方式及加權演算法，同時比較加權前後主題模型產生的變化。

3.1 系統架構

我們使用 LDA 的方法，由圖書的文字內容中來產生主題模型，利用電子書檢索平臺的使用記錄來進行主題加權。

資料的來源包含了圖書的書目資料,全文資料及電子檢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，如圖三-1 所示。

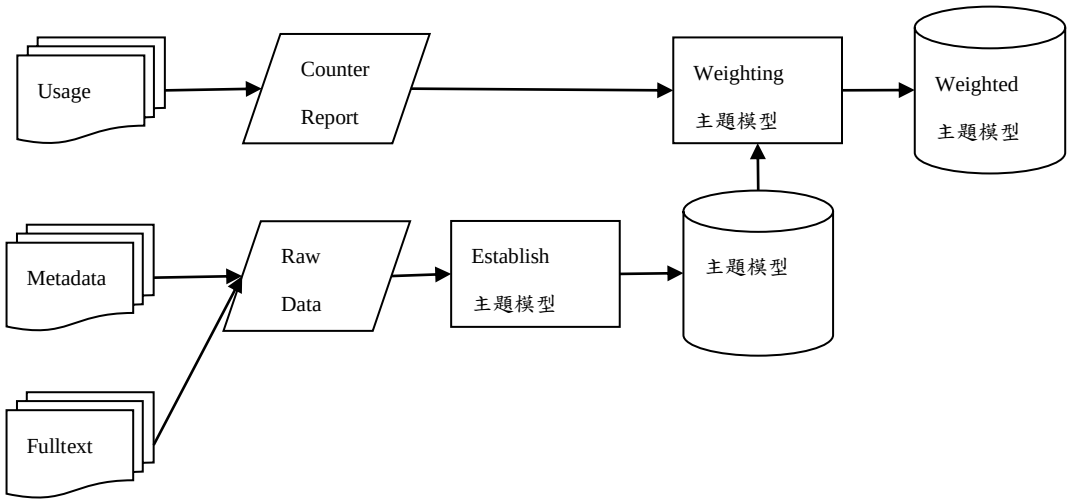


圖 三-1 LDA 主題模型建立及加權系統結構

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，輸出的結果及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，關於詳細的說明請參閱第三章其他小節的內容:

(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、刪選後產生符合 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、範圍及時間來產生統計報表。

(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 α 及 β 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

為了更有效的進行數據分析、圖書推薦及資料管理，未加權及加權後的主題模型皆使用關聯式資料庫來儲存。未加權的主題模型儲存指的是在第二個步驟透過 LDA 工具所產生的主題分佈、主題字彙等相關資料，匯入資料庫中保存; 加權後的主題模型儲存則是在第三個步驟，依使用單位訂購時間及內容依統計資料加權圖書主題的結果，匯入資料庫中保存。

3.2 文字資料前置處理

3.2.1 資料來源

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。Columbia University Press 出版社是在 1893 年所成立，它是美國歷史最悠久的大學出版社之一，每年發行 160 種以上的圖書(Columbia University Press) 出版社透過授權的方式將內容提供給電子書平台的服務業者來進行銷售，通常亦是由平台的服務業者來進行資料的處理，轉換並發佈到線上平台來提供服務，通常圖書館購買電子書時多是透過這類的電子書平台服務業者。

目前 Columbia University 收錄的內容如表三-1 所示:

表 三-1 Columbia University Press 電子書收錄內容

Asian Studies and Literature	Biological Sciences
Business	Culinary History
Current Affairs	Economic
Enviromental Sciences	Film and Media studies
Finance	History
International Affairs	Literary Studies
Middle Eastern Studies	New York City History
Philosophy	Neuroscience
Paleontology	Politcal Theory
Religion	Social Work

全文資料的部份是由出版社提供的電子書原始 PDF 檔案，透過輔助的程式來進行 PDF 文字的抽取，不過在 PDF 檔案當中並未包含有完整的書目資訊，所以書目資訊的部份出版社會另外提供，另外書目資訊的部份經過適當的轉換處理之後會直接保存在關聯式資料庫當中。

在所有圖書 1,324 本中共有 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，表三-2 列出數量排名前二十個主題標目。

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。

表 三-2 數量排名前二十的主題標目資訊

主題標目	**數量**	**主題標目**	**數量**
Jews	30	criticism	12
Women	30	democracy	12

social service	28	world politics	12
american literature	15	english literature	12
popular culture	14	women and literature	12
Globalization	14	liberalism	11
Psychoanalysis	14	african americans	11
english fiction	14	international relations	10
Terrorism	13	environmental policy	10
motion pictures	13	science	10

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，依圖三-2 所示。(各分類的意義請參考第二章關於 LCC 的說明)

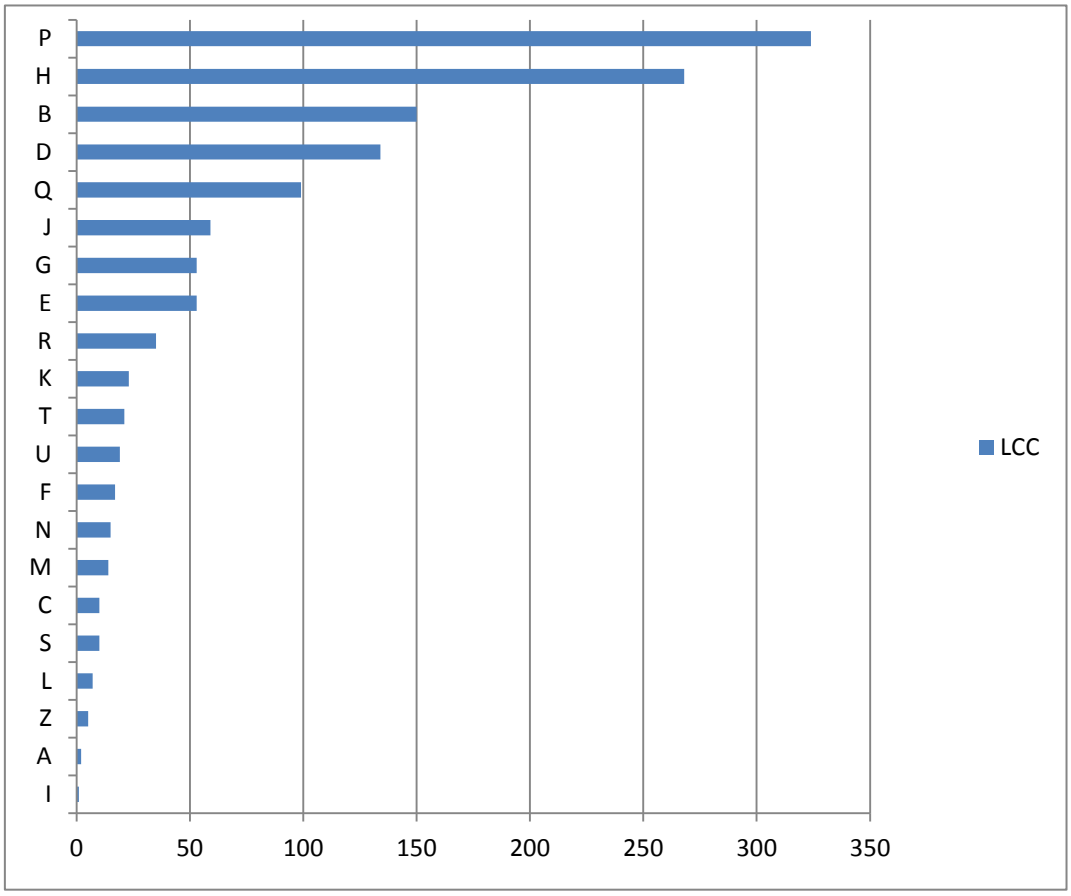


圖 三-2 LCC 分類第一層、圖書分佈

3.2.2 資料處理方式

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。

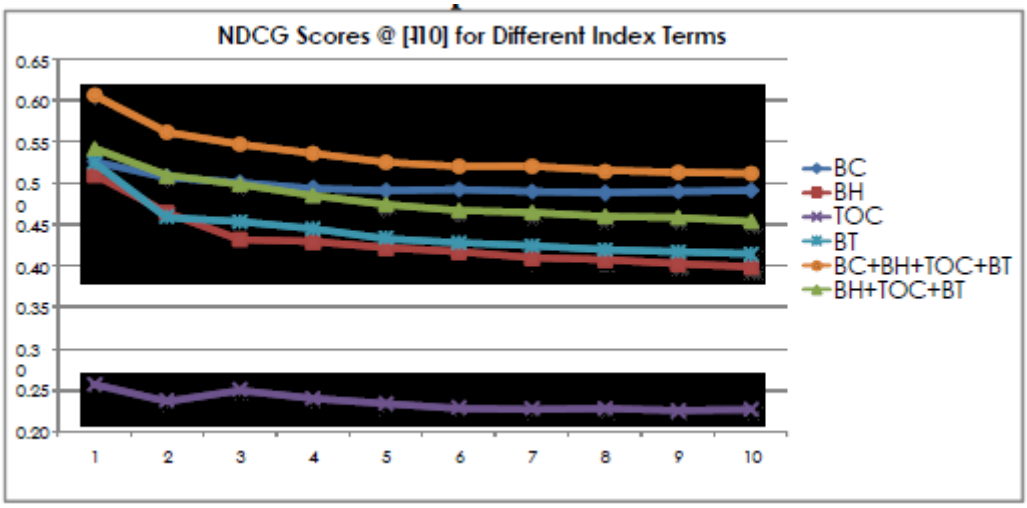


圖 三-3 使用 NDCG 方式評估不同文字組合，觀察 top-n 記錄中所含的資訊量

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 Abstract 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。

3.3 使用記錄前置處理

3.3.1 資料來源

本研究中採用的使用記錄是由 IG Publishing 公司(IGP) 所開發的電子書平台所提供，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。

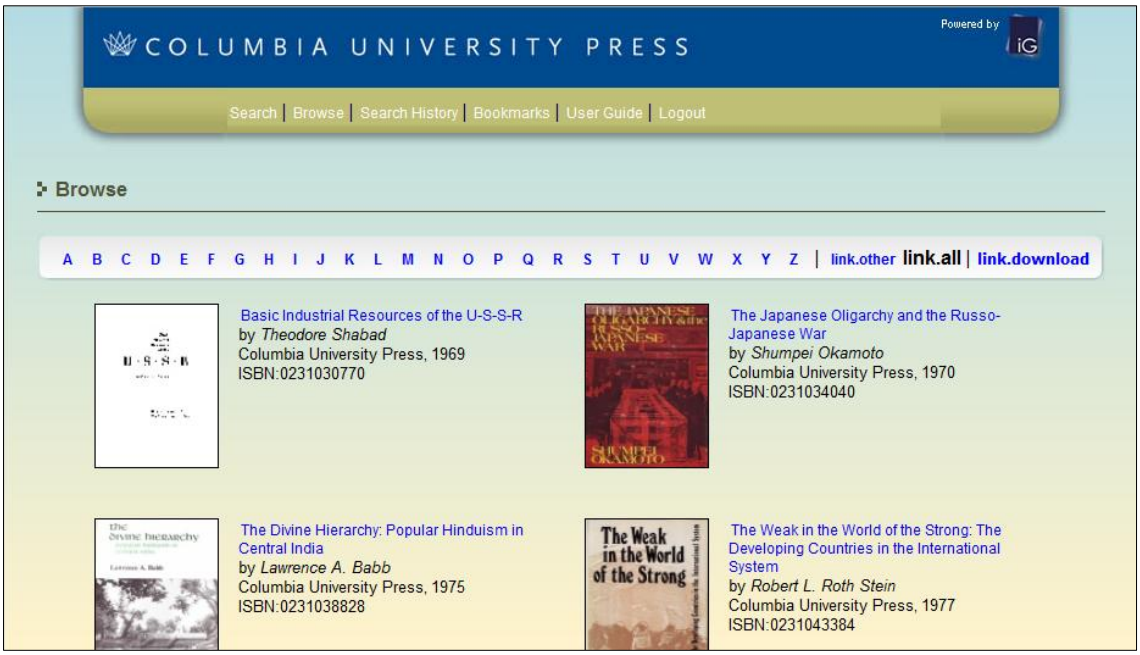


圖 三-4 IGP 電子書檢索平台

IGP 電子書檢索平台運行在 Windows Server 2003 上面，使用 IIS 做為 Web Server 來提供服務，雖然 IIS 有提供 Http Server Log，不過在使用記錄的部份這個平台是依自定的格式來產生，採取這一個方式最主要的目地是考量到 IIS 記錄的訊息過多，同時無法方便加入客製的一些資訊，所以 IGP 平臺使用自定義的圖書使用記錄格式，方便未來統計目地的需求。

我們為了將來相關研究的進行，本研究中採用國立中山大學 2010 整年度的使用記錄做為實驗數據。在使用記錄檔案格式上主要為文字型態，其記錄的內容包含訂購的客戶資訊、系統操作及圖書開啟次數等相關資訊，使用記錄欄位請參考表三-3 的說明。

表 三-3 IGP 電子書平台自定使用記錄欄位說明

項次	欄位內容	欄位說明
1	110101	記錄建立日期
2	09:38:06	記錄建立時間

3	PBDEGHJKUIRCJRDJTHNWuIJwLQIzNS.igroupTH	Session ID
4	110101	記錄建立日期
5	09:38:10	記錄建立時間
6	US	HS 使用 IP 登入 US 使用帳號登入
7	Userid	登入帳號
8	Publishers	使用者身份
9	Demo	使用單位名稱
10	58.10.147.44	使用者 IP 位址
11	58.10.147.44 / ADDR:58.10.147.44*	使用者 IP 位址
12	Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.0.19) Gecko/2010031422 Firefox/3.0.19	瀏覽器資訊
13	DEMO\\COL\\B0000626	圖書代碼
14	http://www.igpublish.com/ama%2Debooks/main.nsp	頁面參照
15	/columbia-ebooks/Cover-thumb.nsp	連線 URL
16	PEFLQFgKLNTEK.publishersTH	系統碼

3.3.2 資料處理方式

使用記錄無法直接用來做主題的加權，必需依使用單位訂購內容、訂購時間範圍及操作記錄進行分析，產生匯整的數據，才能利用這一個資料來加權主題模型。本研究中使用 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統計，匯整的統計資料以月為單位。

在使用記錄處理的部份 CONTER 亦有明確的規範，透過它所建議的方式可以降低各出版社自行分析可能產生的錯誤及偏差值，其中包含了下列幾個要求:

- 記錄中成功及有效的 Requests 必需要被統計，依 NCSA Return Codes 的規範 Web Server Log 必需包含有 HTTP200 或 HTTP304 的註記。
- 記錄中 Page Request 若包含圖片資訊、網頁格式、及其他未使用到的內容必需被忽略。

- 記錄中每一個鏈結的操作若被識別為 Double-Clicks 僅記數為一次。識別 Double-Clicks 的方式為判斷同一個網頁的鏈結記錄，其間隔時間應不低於十秒。在判別 Double-Clicks 的部份，亦可透過下列的方式 ：

  1. 使用 IP Address 的資訊相較於使用記錄進行 Double-Clicks 識別
  2. 使用 Session-Cookie 追蹤使用者登入，進行 Double-Clicks 識別
  3. 當瀏覽器的 Cookie 功能啟用時，應該利用 Cookie 機制進行 Double-Clicks 識別
  4. 當使用者使用帳號/密碼登入系統時，應該利用帳號進行 Double-Clicks 識別

  上述的四種方式，在 Double-Clicks 的識別上，依序由上到下可以提供不同層級的可信度（由低至高）。

- 由於 PDF 檔案的下載時間比 HTML 頁面產生的時間較長，因此在 Double-Clicks 的識別上，若 30 秒之內存取至同一 PDF 檔案的一個或數個 Requests 應該被識為單一個 Request。

使用記錄檔案清理完成之後，我們進一步再依 COUNTER REPORT BR1 的格式產生匯整的統計數據。統計資料依使用單位訂購的圖書內容、時間範圍來做區分。由於本實驗採用的使用記錄格式是由電子書檢索平台自定，所以無法透過使用記錄來取得 HTTP 狀態碼，不過依然可以透過其他的欄位來識別使用者是否合法登入，並產生符合標準的統計報表。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Customer Name:**	National Sun Yat-sen University														
2	**Publisher :**	Columbia University Press eBooks														
3	**Access Period :**	JAN/2010 to DEC/2100														
4																
5	**TITLE**	**ISBN**	**TYPE**	**JAN/2010**	**FEB/2010**	**MAR/2010**	**APR/2010**	**MAY/2010**	**JUN/2010**	**JUL/2010**	**AUG/2010**	**SEP/2010**	**OCT/2010**	**NOV/2010**	**DEC/2010**	**TOTAL**
6																
7	"rail, steam, and s	0231134746	request													0
8	A calculus of suffe	0231051867	request	9	10	4	14	1	1	3	12					54
9	A case-study of li	0-231-04982	request													0
10	A chinese village:		request	9	4	3	6	2	1	2	5					32
11	A communion of s	9780231136	request	3	7	5	5	1	1	1	12					35
12	A cultural history	0231062958	request	9	7	4	3	1	1	2	10	1				38
13	A cultural history	0-231-0629	request													0
14	A derrida reader: I	9780231066	request	8	6	7	7	1	1	4	11					45
15	A field of honor: w	9780231124	request	6	7	2	7	1	1	2	11					37
16	A framework for in	0231120826	request													0
17	A framework for in	0231120826	request	9	5	3	6	1		1	10			1		36
18	A genetic and cult	9780231133	request	7	3	6	7	3	1	1	11					39
19	A guide to orienta	0-231-06674	request													0
20	A guide to orienta	0231066740	request	8	12	4	5	1	1	3	7					41
21	A history of greek	0231017677	request	4	7	5	2	2	1	2	7					30
22	A history of housi	0-231-0629	request													0
23	A history of housi	0231062966	request	7	5	4	3	1	1	3	11					35
24	A history of latin li	0231018487	request	8	7	5	10	1	1	2	10					44
25	A limited partners	0231120842	request	9	7	2	5	1	1	2	12					39
26	A modern heretic	0231106262	request	10	8	6	12	1	1	1	9					48
27	A natural history c	0231129947	request	11	7	5	5	1	1	1	10					41
28	A partisan century	9780231103	request	5	4	5	9	1	1	2	11					38
29	A popular guide to	0-231-04019	request													0

圖 三-5 2010 年 1-12 月 BR1 統計數據範例

依圖三-4 所示，使用單位的顯示依每一本圖書，每月的成功開啟次數來做匯整，而主題的加權即是使用匯整的數據來做計算。

3.4 LDA 參數選擇

建立 LDA 主題模型需要設定 Hyper-Parameter 的α及β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α=50/K、β=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，由圖三-5 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。

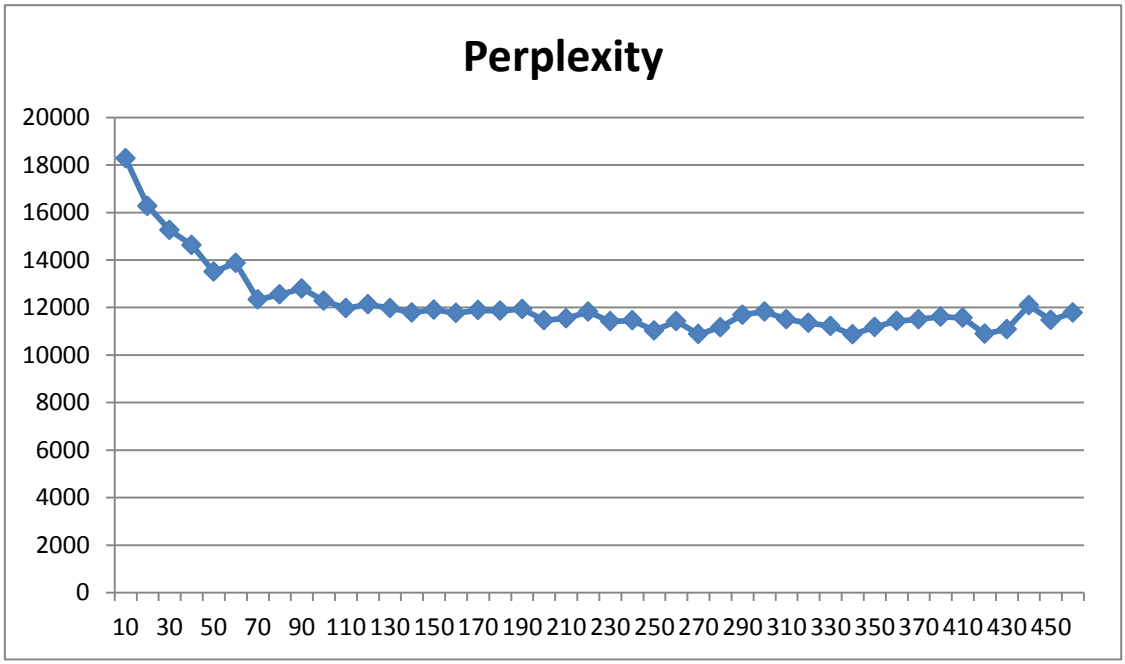


圖 三-6 主題模型 Perplexity 的計算結果

3.5 主題模型建置

3.5.1 工具的選擇

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，它適合於下列幾種應用情境:

- 資訊檢索方面（分析語義/隱含主題/大量文字的概念結構）
- 文件的分類、分群及 Text/Web 的資料探勘等目地
- 協同過濾
- 依內容為基礎的圖像分群、物件識別及其他電腦圖像上的應用
- 其他應用範圍如生物科技領域等

在 JGibbsLDA 工具中可以設定主題模型建置時的各項參數，表三-4 為各項參數的說明:

表 三-4 JGibbsLDA 參數說明

項次	參數	資料格式	參數說明
1	-est		建立基礎主題模型
2	-alpha	double	設定 LDA hyper-parameter Alpha 值. 預設值為 50/K（K = 設定的主題數量）
3	-beta	double	設定 LDA hyper-parameter Beta 值，預設值為 0.1
4	-ntopics	int	設定 Topics 的數量，預設值為 100
5	-niters	int	設定 Gibbs Sampling 迭代執行的次數，預設值為 2,000 次
6	-savestep	int	設定主題模型階段儲存間隔，預設值每 200 次迭代，執行階段儲存
7	-twords	Int	設定每一個 Topic 要顯示的 Words 數量，預設值為 0。假如設定數量大於 0 則依設定值，該 Topic 可能性最高的 Words 會被顯示並儲存
8	-dir	string	設定訓練資料的目錄位置
9	-dfile	string	設定訓練資料的檔案名稱

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。

3.5.2 輸入資料格式

來源資料為純文字檔案格式，在匯入檔案的第一列記載的是全部文件的數量，而接下來的內容，每一列均代表一本圖書所有文字的集合。

範例格式如下:

[M]

[$document_1$]

[$document_2$]

...

[$document_M$]

在每一份文件當中，每一個字彙之間需使用空白做為分隔

[$document_i$] = [$word_{i1}$] [$word_{i2}$] ... [$word_{iNi}$]

在實驗當中重新組合了書目資料的題名、目次、關鍵字索引頁及圖書的摘要，匯整成單一個檔案，並且記錄每一列文字資料相對應資料庫當中的那一筆圖書記錄，透過這一個資訊才能讓程式將 LDA 所產生的各種主題機率值正確的回寫到資料庫當中。

3.5.3 輸出資料格式

這一個工具所輸出的資料為純文字檔案格式，記錄了主題在各文件上的分佈及每一個字彙在主題上的分佈等資訊。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 & 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。關於 JGibbsLDA 工具輸出的檔案格式如表三-5 所示:

表 三-5 JGibbsLDA 輸出檔案型態說明

項次	儲存檔案名稱	檔案說明
1	<model_name>.others	這一個檔案記錄 主題模型各階段建立時，設定

		的初始參數及訓練檔案資訊，其中包含 alpha 值、beta 值、ntopics(主題數量)、ndocs(文件數量)、nwords(vocabular 大小)及 liter(迭代次數)
2	<model_name>.phi	這一個檔案儲存每一個 Word 被指定 Topic 的機率分佈。例如 $p(word_w \mid topic_t)$，每一列為一個主題，每一欄為 vocabulary 中的 Word
4	<model_name>.theta	這一個檔案儲存每一份文件所包含 Topics 的機率分佈。例如 $p(topic_t \mid document_m)$，每一列為一份文件，每一欄為一個主題
5	<model_name>.tassign	這一個檔案儲存訓練資料中每一個 Word 被指定為特定 Topic 的對應。每一列為一份文件，格式為 $<word_{ij}>:<topic\ of\ word_{ij}>$
6	<model_file>.twords	這一個檔案儲存每一個 Topic 所包含的 Words，顯示的數量可透過命令列來設定
7	Wordmap.txt	儲存每一個 Word 與系統自動編號的對應

3.5.4 主題模型的建置

表三-6 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。

再嚐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 三-6 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

3.5.5 LDA 資料庫的設計

由於 LDA 建置工具輸出的資料為文字檔案格式，為了方便進行管理及後續分析的進行，本研究使用關聯式資料庫來儲存主題模型，資料庫 ER 設計圖如 3-5 所示。資料庫中包含了使用單位、書目資料、出版社資料、使用統計資料、圖書使用的比重、未加權的主題模型資料、已加權的主題模型等相關資訊。

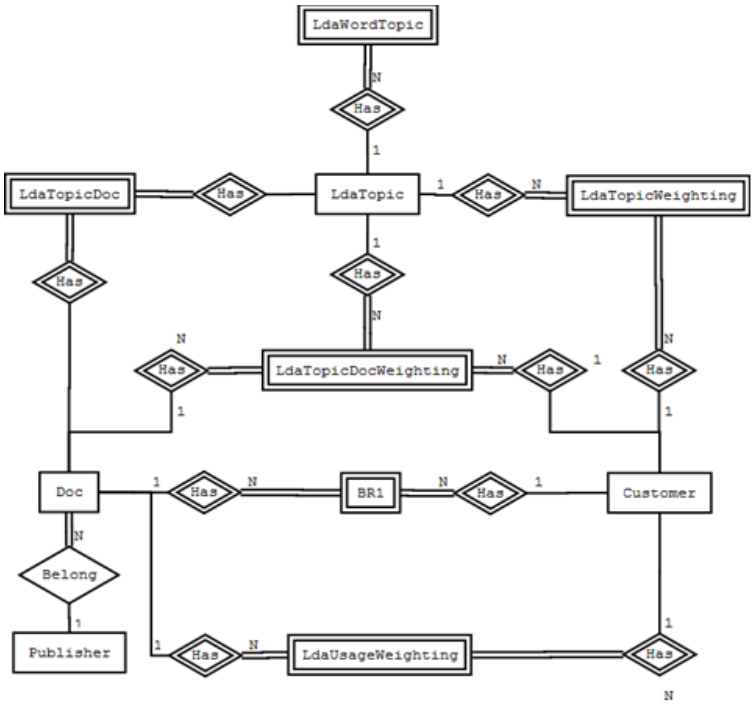


圖 三-7 LDA 資料庫 ER 設計圖

其中已加權的主題模型，依使用單位、各月份統計資料、訂購的圖書為基礎，然後再依統計資料中各圖書的使用比重來進行加權的計算，如表三-7 的說明。

表 三-7 資料庫表格說明

項次	表格	功能說明
1	Publisher	出版社基本資料
2	Doc	圖書基本資料
3	Customer	使用單位基本資料
4	Br1	依 COUNTER 圖書報表格式，分析使用記錄檔案，儲存使用單位每月開啟電子書的數量
5	LdaTopic	依 LdaTopicDoc 為基礎，記錄各出版社包含之圖書各 Topics 機率分佈總合
6	LdaTopicDoc	每本圖書於主題模型每一個 Topic 機率分佈
7	LdaWordTopic	記錄每個 Topic 中所包含的 Words 及其機率分佈
8	LdaUsageWeighting	記錄使用單位每月圖書開啟書量加權值，依 BR1 的統計數據進行計算的結果

9	LdaTopicWeighting	依 LdaTopicDocWeighting 為基礎，記錄使用單位每月，加權後各圖書 Topics 機率分佈總合
10	LdaTopicDocWeighting	使用單位每月使用圖書，依 LdaTopicWeighting 為基礎加權後於 主題模型 中每一個 Topic 的機率分佈

3.6 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一般化的處理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。

表三-8 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。

表三-8 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

當主題的機率較高時代表其在所有圖書當中出現的比例亦比較高，具有較高的重要性，可以代表這群圖書主要的內容描述，我們透過主題機率分佈的排序來找出影響性較高的主題做為加權前、後的比較。

表三-9 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。

表三-9 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**

由於使用記錄產生的統計資料是針對各別的使用單位，各單位的使用狀況都不一樣，訂購的圖書內容或者是使用的期限也都不同，所以在加權主題結果的儲存上，需為每一個使用單位做個別的分析及資料的儲存，同時搭配訂購的時間範圍來做劃分。

在本實驗中是採每月為單位來做分析，例如 2010 年的部份，針對中山大學圖書館的使用記錄，在加權主題的結果上，由 1 月份開始至 12 月為止，共會儲存 12 份的加權主題模型資料，僅針對該月有使用過的圖書，對於該月份沒有使用過的圖書則不進行處理，其在該月的主題產生機率則視為零。

第四章　實驗結果

4.1 前言

本章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。最後再以文件主題機率分佈為基礎利用 Jensen Shannon 方法找出內容相近的書，進一步來觀察主題模型與內容之間的相關性。

4.2 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四-1 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 四-1 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

由圖四-1 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，可以合理解釋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四-2 的主題編號 92 (Chinese Political) 上面發現。

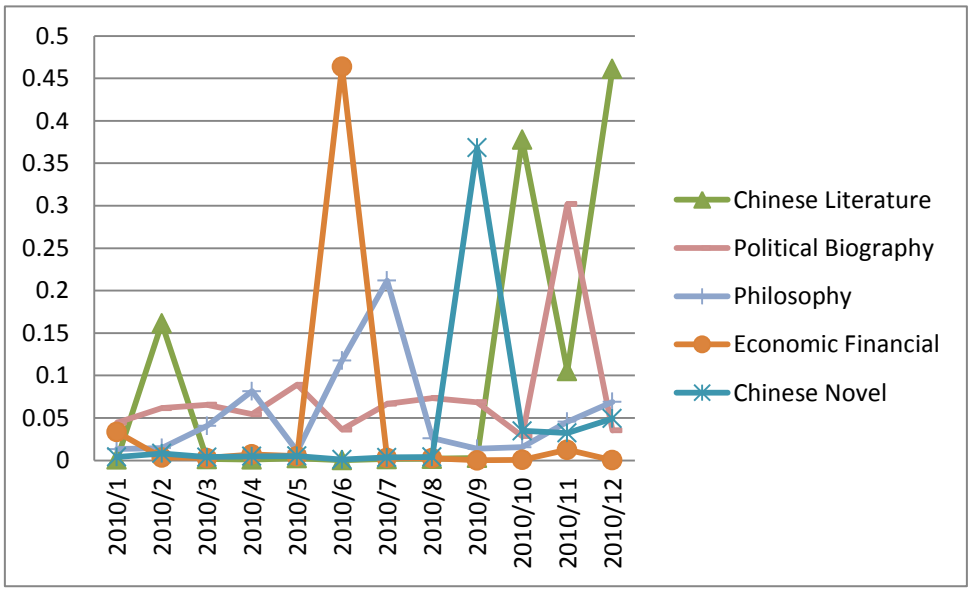


圖 四-1 主題編號 10, 73, 65, 25, 24 全年度變化趨勢圖

我們再觀察圖四-2 呈現的趨勢，它並沒有像圖四-1 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。

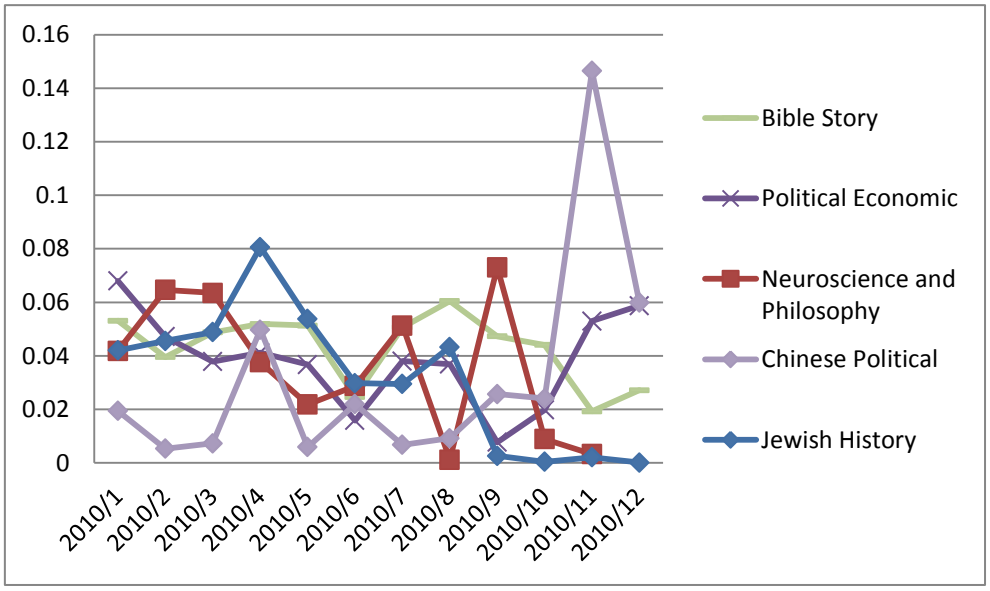


圖 四-2 主題編號 82, 22, 4, 92, 3 全年度變化趨勢圖

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一般化，所以其機率值不需要再取平均值。

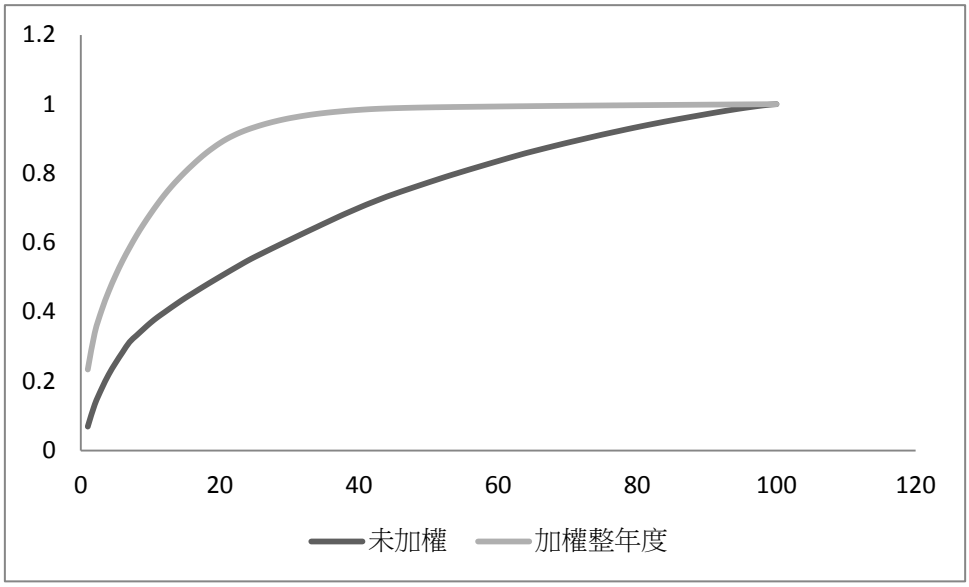
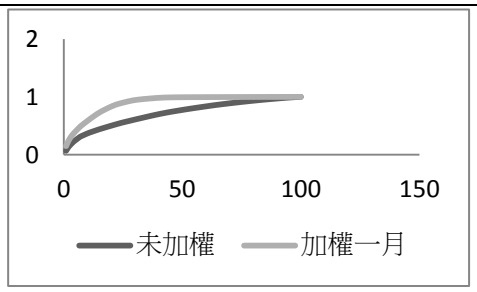
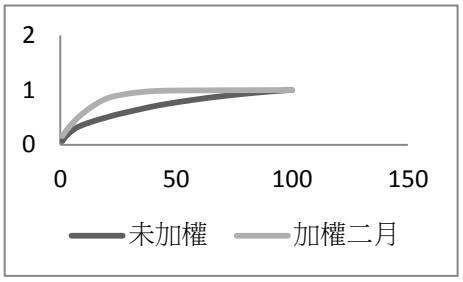


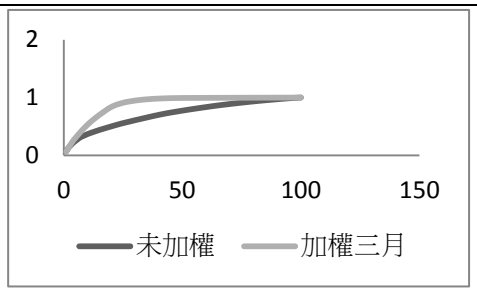
圖 四-3 未加權與加權主題機率累計比較 (整年度)

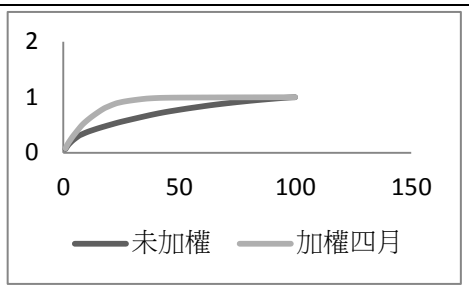
本實驗中我們使用 100 個主題，透過圖四-3 可以觀察到加權後的機率值產生很明顯的變化，當累計的主題越多時則變化的幅度越小，我們可以發現上圖在 25 個主題之後曲線就變的平緩，所以前 25 個主題可以在趨勢上應該也會有較明顯反應，另一方面我們再比較每一個月份的主體機率累計相對於基準線亦可發現累似的狀況產生。由此可以看出，透過使用加權後，各主題的機率差異性明顯提升。因此較少的主題即可涵蓋較多的累計機率。

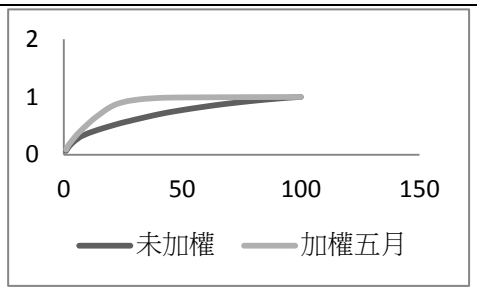
表 四-2 未加權與加權主題機率累計比較 (2010 一至十二月)

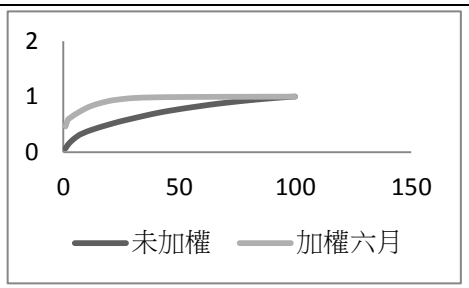


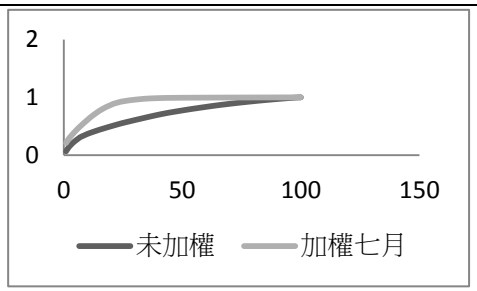


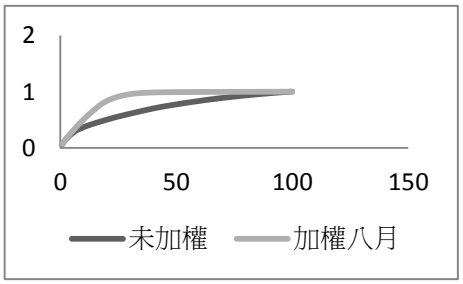


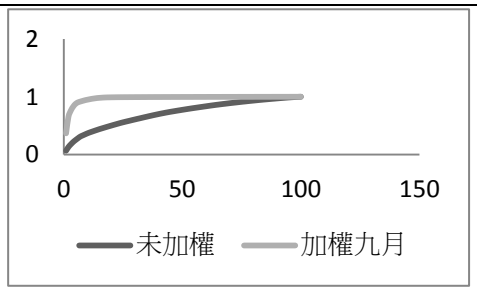


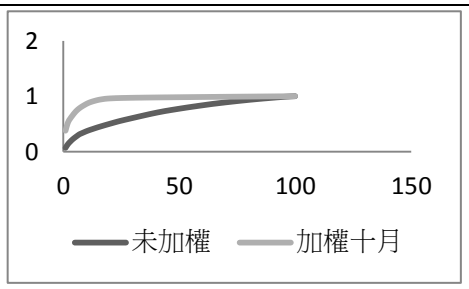


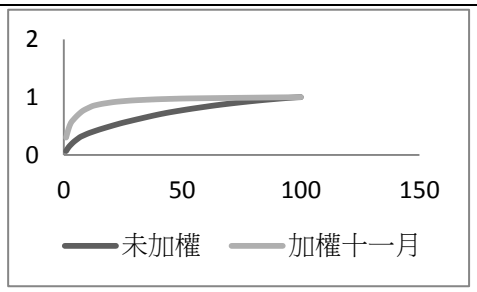


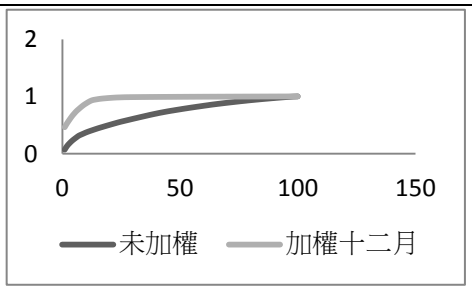












4.3 LCC 與主題模型關聯性

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 $H(X)=-\sum_{i=1}^{n} p_i \log_2(\frac{1}{p_i})$

計算的結果如表四-3 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 四-3 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，如表四-4 所示，在實驗中使用主題數 100、50、25 及 10 個，我們發現計算的結果均無法透過檢測，所以可以說明 LCC 與主題機率間並不相關，所以亦不可能透過主題來推測 LCC 的分類。

表 四-4 Chi-Square 獨立檢測，檢定 LCC 與主題模型的獨立性

項次	主題數量	自由度	X^2	檢定結果
1	100	130482	73655.6	Reject
2	50	64582	37999.65	Reject
3	25	31632	18737.13	Reject
4	10	11862	7290.171	Reject

4.4 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，依表四-5 的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。

表 四-5 依 LCSH 主題標目計算主題模型的資訊熵

項次	主題	主題數量/資訊熵 (Entropy)			
		100	**50**	**25**	**10**
1	jews	3.777132	3.172172	3.314006	2.457882
2	women	4.383585	3.836623	3.72439	2.653474
3	social service	2.265817	1.607074	1.295703	0.689896
4	united states	5.120702	4.576993	3.745369	2.447077
5	american literature	3.81016	3.020742	2.401108	1.097221
6	popular culture	3.884846	3.289381	2.802201	1.573595
7	globalization	3.174381	2.845019	2.048639	1.974482
8	psychoanalysis	1.99542	1.395683	1.016353	0.41275
9	english fiction	2.564103	1.92602	2.16446	1.277201
10	motion pictures	2.601964	2.926272	2.177495	1.95395
11	terrorism	2.546489	2.926758	2.810019	1.583136
12	world politics	3.605818	2.905595	2.013274	1.117567
13	criticism	3.283179	2.908275	1.962204	1.249676
14	women and literature	3.865296	3.008708	2.663648	1.830734
15	democracy	3.505789	3.361623	2.682261	2.151832
16	english literature	3.763294	2.319638	1.968547	1.676471
17	african americans	2.925981	2.6247	2.429885	1.749949
18	liberalism	3.533795	3.069488	2.669205	1.768781

19	environmental policy	3.048519	2.384029	2.707478	2.231373
20	literature and society	3.632643	2.716928	2.233888	1.923977
21	american fiction	3.231662	1.949243	2.765716	1.359774

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 X^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。如表四-8 的實驗結果顯示當α=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。

表 四-6 Chi-Square 獨立檢測，比較 LCSH 與主題模型的相關性

項次	主題	主題數量/DF, X^2					
		DF	**50**	**DF**	**25**	**DF**	**10**
1	Jews	833	572.749	408	223.8073	153	112.6885
2	Women	980	420.1871	480	205.5788	180	76.09051
3	social service	833	372.0508	408	188.797	153	86.12867
4	american literature	539	260.0363	264	107.2141	99	48.58697
5	popular culture	539	217.796	264	123.974	99	45.09498
6	globalization	343	109.2554	168	78.54041	63	15.88493
7	psychoanalysis	441	147.8533	216	70.73955	81	43.07893
8	english fiction	392	175.4719	192	98.68598	72	31.19404
9	motion pictures	392	197.3859	192	147.5268	72	35.78724
10	Terrorism	392	161.1483	192	79.60843	72	30.85327
11	world politics	539	197.8555	264	111.8378	99	52.47941
12	Criticism	441	153.2612	216	88.79001	81	43.68991
13	women and literature	490	272.6199	240	102.8229	90	41.23342
14	Democracy	441	132.2767	216	89.27941	81	32.49891
15	english literature	441	204.5763	216	114.7373	81	40.27503
16	african americans	441	207.9803	216	103.71	81	48.0982

17	Liberalism	343	167.3522	168	86.74271	63	37.75114
18	environmental policy	245	227.1158	120	73.36189	45	24.93145
19	literature and society	294	139.9799	144	77.68051	54	28.82024
20	american fiction	294	175.0097	144	58.98812	54	26.70507

4.5 LCC、LCSH 與主題相關性觀察

主題在文件當中的機率值越高，表示其在文件中出現的次數也越多，每一份文件在各主題所包含的機率都不相同，我們可以透過比較主題機率值來找出相近的圖書。由 4.3 及 4.4 的實驗中証明無法透過 LDA 產生的主題來決定書目中的圖書分類,同時 LDA 主題與主題標目的相關性亦不明顯,本研究利用 Jensen Shannon 方法依主題分佈的機率為基礎，來找出相近的圖書，進一步來觀察 LCC 及 LCSH 與主題內容的差異。

首先由 LCC 分類號來觀察，以圖書編號 COLB0000024、COLB0000159、COLB0000589 三本圖書為基礎各別找出最相近的五本書，我們觀察表四-7、四-8 及四-9 可以發現主題機率分佈相近的圖書，在書目記錄中的分類存在差異，皆無法被歸納為單一類別的圖書。

表 四-7 依圖書編號 COLB0000024 找出前五筆相近的圖書，比較 LCC 分類號

圖書編號	LCC 分類號	書名
COLB0000024	JS100.E43	The Dynamics Of Computing
COLB0000646	T176	Limited by Design: R&D Laboratories in the U.S. National Innovation System
COLB0000427	JS344.E4	Computers and Politics: High Technology in American Local Governments
COLB0000530	RB152.5	Toxic Exposures: Contested Illnesses and the Environmental Health Movement
COLB0000111	JS344.E4	The Management of Information Systems
COLB0000908	KF3467	Fetal Protection in the Workplace: Women's Rights, Business Interests, and the Unborn

表 四-8 依圖書編號 COLB0000159 找出前五筆相近的圖書，比較 LCC 分類號

圖書編號	LCC 分類號	書名
COLB0000159	GN345	Human Nature and History: A Response to Sociobiology
COLB0001037	GN315	Theories of Man and Culture
COLB0000503	PN81	Theory's Empire: An Anthology of Dissent
COLB0000097	HM24.A4	Structure and Meaning: Relinking Classical Sociology
COLB0000512	BD450	Being Human: Historical Knowledge and the Creation of Human Nature
COLB0000625	PN81	The Range of Interpretation

表 四-9 依圖書編號 COLB0000589 找出前五筆相近的圖書，比較 LCC 分類號

圖書編號	LCC 分類號	書名
COLB0000589	E185.97.W4	The Education of Booker T. Washington: American Democracy and the Idea of Race Relations
COLB0001159	PN4877	The Dream of a New Social Order: Popular Magazines in America
COLB0000781	LD1250	Changing the Subject: How the Women of Columbia Shaped the Way We Think About Sex and Politics
COLB0000887	HN57	The Refuge of Affections: Family and American Reform Politics, 1900-1920
COLB0001283	E178.6	Contested Democracy: Freedom, Race, and Power in American History
COLB0000778	E185.615	Black Leadership

另外我們再觀察這三本書的 LCSH 的內容，依表四-10、四-11 及四-12 所顯示的資訊可以顯示主題標目的內容在每一本圖書的書目當中都不太相同，在主題的數量及內容上都有差異。

表 四-10 依圖書編號 COLB0000024 找出前五筆相近的圖書，比較 LCSH 主題標目

圖書編號	LCSH 主題標目

COLB0000024	Municipal government, Data processing.
COLB0000646	Research, Industrial, United States, Laboratories.,Technology and state, United States.
COLB0000427	Local government, United States, Data processing.
COLB0000530	Environmentally induced diseases.,Asthma, Etiology.,Breast, Cancer, Etiology.,Persian Gulf syndrome, Etiology.,Environmental Exposure, adverse effects.,Asthma, etiology.,Breast Neoplasms, etiology.,Environmental Health, trends.,Persian Gulf Syndrome, etiology.,Public Policy.
COLB0000111	Municipal government, United States, Data processing.
COLB0000908	Pregnant women, Employment, Law and legislation, United States.,Fetus, Legal status, laws, etc., United States.,Fetus, Abnormalities.,Pregnancy, Complications.

表 四-11 依圖書編號 COLB0000159 找出前五筆相近的圖書，比較 LCSH 主題標目

圖書編號	LCSH 主題標目
COLB0000159	Ethnology, Philosophy.,Culture.,Sociobiology.,Social history.
COLB0001037	Ethnology.,Anthropologists.
COLB0000503	Criticism.,Literature, History and criticism, Theory, etc.
COLB0000097	Sociology.,Civilization, Modern.
COLB0000512	Philosophical anthropology.
COLB0000625	Literature, History and criticism, Theory, etc.,Translating and interpreting.,Interpretation (Philosophy),Canon (Literature)

表 四-12 依圖書編號 COLB0000589 找出前五筆相近的圖書，比較 LCSH 主題標目

圖書編號	LCSH 主題標目
COLB0000589	Racism, United States.,Racism, Political aspects, United States.,African Americans, Civil rights, History.,Civil rights movements, United States, History.,Democracy, United States.
COLB0001159	American periodicals, History.,Journalism, Social aspects, United States.,Popular culture, United States.
COLB0000781	Feminism and higher education, New York (State), New York, History, 20th century.,Women in higher education, New York (State), New York, History, 20th century.,Coeducation, New York (State), New York,

	History, 20th century.
COLB0000887	Social reformers, United States, Biography.,Families, United States, History.,Progressivism (United States politics)
COLB0001283	Democracy, United States, History.,Power (Social sciences), United States, History.,Radicalism, United States, History.,Social movements, United States, History.
COLB0000778	African American leadership.

由於圖書文字當中包含的資訊太過廣泛，主題僅依靠機率的方式來產生，並未經過專家的解讀，相較於書目記錄是經由圖書館員或領域專家所提供，所以 LDA 方法所產生的主題模型與書目記錄當中存在有差異，無法取代圖書館的分類及主題標目的內容。

但是 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，所以更能反應圖書的內容，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。

第五章　結論與未來研究建議

5.1 結論

在本研究中以使用記錄來加權主題模型、同時使用 Jensen Shannon 方法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)。在本實驗中嚐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。

5.2 未來研究建議

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。

未來在參數的選擇上若能針對不同的學科領域，訂出一個通用參數值，讓 α、β 與主題數量的設定有一個參考的標地，則在實務上就不需要為了尋求最佳的結果而花費許多時間來調整參數。

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，本研究證明了加權主題模型與 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。

第六章　參考文獻

AlSumait, L., Barbara, D., & Domeniconi, C. (2008, 15-19 Dec. 2008). *On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking.* Paper presented at the Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on.

Anthes, G. (2010). Topic models vs. unstructured data. *Commun. ACM, 53*(12), 16-18. doi: 10.1145/1859204.1859210

Blei, D. M., & Lafferty, J. D. (2006). *Dynamic topic models*. Paper presented at the Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania.

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

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.

. Columbia University Press. from http://cup.columbia.edu/

. COUNTER - Counting Online Usage of Networked Electronic Resources. from http://www.projectcounter.org/

. COUNTER - Counting Online Usage of Networked Electronic Resources　Home. from http://www.projectcounter.org/

Darling, W. M. (2011). *A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling*.

. Gibbs sampling. from http://en.wikipedia.org/wiki/Gibbs_sampling

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

Hall, D., Jurafsky, D., & Manning, C. D. (2008). *Studying the history of ideas using topic models*. Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii.

Khosh-khui, S. A. (1987). *Relationship Between LCSH and LCC Notationsin Different Classes of LCC.*　Staff Publications-Library, Texas State University.

. Library of Congress Classification. from http://www.loc.gov/catdir/cpso/lcc.html

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

Maskeri, G., Sarkar, S., & Heafield, K. (2008). *Mining business topics in source code using latent dirichlet allocation*. Paper presented at the Proceedings of the 1st India software engineering conference, Hyderabad, India.

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.

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.

Shepherd, P. T. COUNTER: towards reliable vendor usage statistics. [Conceptual Paper]. *VINE, 34*(4). doi: 10.1108/03055720410570975

Sun, Y., Han, J., Gao, J., & Yu, Y. (2009). *itopicmodel: Information network-integrated topic modeling*.

Wang, X., & McCallum, A. (2006). *Topics over time: a non-Markov continuous-time model of topical trends*. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA.

. 國家圖書館編目園地全球資訊網. from http://catweb.ncl.edu.tw/