主題複分效用分析：以NBINET為例

吳政叡 (Cheng-Juei Wu)

輔仁大學圖書資訊系專任教授

E-mail: lins1022@mails.fju.edu.tw

中文摘要

圖書館透過主題複分，一方面將資料加以更詳細的細分，來協助使用者更準確的找到其所想要的資料；另一方面也使資料能分到各細目下，以降低資料的集中度。為了探究主題複分的實際功效，作者自「中文圖書標題表」隨機抽取80個標題，然後利用國家圖書館NBINet來進行主題檢索分析。結果發現雖然每個複分標題平均記錄數目並不高，例如高筆數組（每個標題資料總數100以上）的總平均值為35.1筆。但這祇是表面的假象，進一步的分析發現，主題複分並未發揮其預期的功能，因為大多數的資料仍然集中在極少數的複分標題中。以高筆數組的標題來說，第1高複分的記錄數目大多數超出平均值甚多，其記錄數目從55筆到3049筆，其中半數標題的第1高複分記錄數目超過200筆，祇有2個標題的第1高複分記錄數目在100筆以下；所佔資料總筆數的百分比則從28.8%到98.5%。綜括來說，不但是絕大部份的資料，集中在極少數的標題上；而且在這些少數的超高使用頻率的標題（即高筆數組）中，大多數的資料又集中在極少數的複分標題中，高筆數組第1高複分的平均記錄數目高達541筆。由此看來，主題複分並未完全發揮實際功效，而且此部分的使用者資訊超載問題是非常嚴重。

==========

Effectiveness analysis of Subject Subdivision: Using NBINet

In this study, we randomly select 80 subjects from the Chinese Subject Headings as the searching terms on the NBINet, which is the largest

bibliographic database in Taiwan, to measure the effectiveness of subject subdivision commonly used in the subject analysis. At first, it seems that subject subdivision works very well since the average number of records of subheadings is small. For example, the average number of records of subheadings is 35.1 for the group of high-usage subjects, which mean the number of records including each of them is more than 100. However, careful analysis exposes that it is only an illusion because most of the records are in few subdivisions actually. For instance, the number of records of the highest subdivision, ranging from 55 to 3049, is well over the average, ranging from 4.1 to206.4, for every heading in the group of high-usage subjects. In addition, on average, these subdivisions occupy 81.1% of records, ranging from 28.8% to 98.5%. These experimental results clearly show that the effectiveness of subject subdivision is pretty bad in reality. This indicates the potential problem of information overload can not be solved by the current practice of subject subdivision.

關鍵字：中文圖書標題表，主題分析，標題，資訊超載，Chinese Subject Headings，Subject Analysis，Subject Heading，Information Overload，NBINet。

一、前言

主題（subject）或標題（subject heading）的定義為：「一個詞或一組字用以表示資料討論的主題。」[註 1]，除了基本的書目資料外，主題可以說是目前著錄人員用來揭露資料內容最重要的工具。另外一個用來揭露資料內容的工具—分類號，由於肩負排架功能，祇能有一個 [註 2]，這使得分類號不若主題來的靈活。

根據研究顯示，主題檢索是圖書館使用者在線上檢索時最常用的方法 [註3-4]。然而以往圖書館界對於主題或標題的探討，大都集中在控制詞彙的應用（即實踐主題標目的統一原則 [註5]）與詞彙的選用、主題標目的結構 [註6]、主題標目的複分方式 [註7]、主題分析的作法 [註8-9]、標題表缺失 [註10-11]等，近年來雖然也有許多探討主題與檢索效益的關係 [註12]，不過仍然是在傳統的主題標目結構下來探討。

目前臺灣地區圖書館編目人員在給與圖書標題時的主要依據為「中文圖書標

題表」，於1993年出版，是目前臺灣地區最完整的中文標題表。作者在「中文圖書標題表的使用情況分析：以NBINet為例」一文中 [註13]，藉由隨機抽樣方式，發現「中文圖書標題表」的另外一項缺失。因為隨機抽取的標題在國家圖書館NBINet來進行主題檢索時，分析實驗結果後發現約有26%的標題從未被使用過。如此高的未被使用比例，似乎也意謂「中文圖書標題表」中若干標題有重新檢討的必要。

由於本文的重點在關注主題複分與資訊超載的關係，和主題在書目資料庫中的資料分佈狀況，作者從「中文圖書標題表」中隨機抽取80個標題 [註14]，以全國最大的「全國圖書資訊網路系統」（National Bibliographic Information Network，簡稱NBINet）為實驗對象 [註15]，在西元2002年4月5-7日三天進行調查。NBINet具有國家書目資料庫的特質，目前參與的合作館已擴及各類型圖書館共70所。另外根據國家圖書館NBINet簡介網頁上公布的統計資料顯示 [註16]，截止至2002年（民國91年）3月底止，書目記錄筆數幾近373萬筆。以記錄類型來看，圖書資料佔絕大多數，有3,557,611 筆；記錄語文來分，絕大多數為中文資料，有2,938,320 筆。由上述的統計數據來看，NBINet可以說是國內最主要和具代表性的中文圖書書目資料庫，因此也是測試「中文圖書標題表」的最佳實驗系統。

二、主題複分效用的表面假象

為了觀察主題在書目資料庫中的資料分佈狀況，作者在「中文圖書標題表的使用情況分析：以NBINet為例」一文中，製作了抽樣標題之記錄數量的次數分配表。發現就標題的整體使用情況來看，低使用頻率的標題（使用次數不超過30者）有60個，佔全體80個標題的75%（即3/4），其合計的被使用機率祇約為4.6%。相反的，會造成使用者資訊超載之超高使用頻率的標題（資料筆數超過100以上者）有10個，約佔全體80個標題的13%（即1/8）。但其總資料筆數為7752，約佔全體資料筆數8804的88%。換言之，極少數1/8的標題，卻佔全體使用機率的88%。

由前面的分析可知，由於絕大部份的資料，集中在極少數的標題（主題）上，因此潛在的資訊超載問題是很嚴重的。為此，圖書館界引進了各種形式的複分，將資料加以更詳細的細分，來協助使用者更準確的找到其所想要的資料，另一方面也使資料能分到各細目下，以降低資料的集中度。

然而正如作者在「模糊邏輯在主題分析的應用：標題權值的計算方式」一文中指出的 [註17]，主題複分似乎並無法達成預期的功效和解決資訊超載的問題。[註18] 然而，此項觀察祇是根據作者隨意使用的四個通用詞彙--天主教、宗

教、心理學、電腦，因此其立論的基礎並穩固。

在本文中作者用嚴謹（自「中文圖書標題表」隨機取樣）和較具規模方式（80個標題）來重新檢視主題複分的效用。實驗過程如下：先自「中文圖書標題表」隨機取樣標題；其次，將取樣得到的標題依序來利用NBINet查詢，並記錄結果；最後依據每個標題的資料總筆數來整理、分組、和分析。

資料整理和分組的方式如下：由於當資料筆數很少時，既無主題複分的需要，也無資訊超載的憂慮，因此作者將資料總筆數不超過10者省略。

（1）資料總筆數不超過30者省略：由於當資料筆數很少時，既無主題複分的需要，也無資訊超載的憂慮，因此作者將此類資料省略不用。

（2）資料分組方式：依據資料總筆數分成二組，分別是高筆數組（資料總數100以上）和中筆數組（資料總數31-100）。

（3）每組標題數目固定為10：為了便於比較，以上二組的標題數目劃一為10，因此是按標題取樣順序和查詢結果來分組，每組祇取前10個適宜歸入該組的標題。

作者根據上述的三個資料整理和分組的原則，整理出表1和2二個表格如下。每個表格都含有標題名稱、標題存在形式（主題複分）數目、資料總筆數、和平均值（資料總筆數/主題複分數目）等四項資訊。

表1. 高筆數（資料總數100以上）之主題抽樣結果。[註19]

標題名稱	標題存在形式（主題複分）數目	資料總筆數	平均值（資料總筆數/主題複分數目）
產科	24	445	18.5
共產主義	44	476	10.8
民俗音樂	47	191	4.1
中等教育	155	1984	12.8
時間	19	689	36.3
食物治療	15	3096	206.4
初等教育	35	207	5.9

眼	25	156	6.2
景觀工程	16	284	17.8
夫妻	7	224	32.0

表2. 中筆數（資料總數31-100）之主題抽樣結果。[註19]

標題名稱	標題存在形式（主題複分）數目	資料總筆數	平均值（資料總筆數/主題複分數目）
文字學	17	92	5.4
投影幾何	1	32	32.0
電腦犯罪	8	52	6.5
膽固醇	1	63	63.0
生活教育	9	73	8.1
兵家	8	58	7.3
器械體操	1	37	37.0
泛函分析	2	63	31.5
貨幣政策	15	82	5.5
奧林匹克運動會	10	93	9.3

表3. 二個分組的總和與平均值的統計表格。[註19]

組別	標題存在形式（主題複分）數目總和	資料筆數總和	平均值（資料筆數總和/主題複分數目總和）
高筆數組	387	7752	20.0
中高筆數組	72	645	9.0

如果單從表3的統計數字來看，似乎主題複分有發揮其效用。即便是再仔細查看表1和2的統計資料，可以發現祇有高筆數組的標題「食物治療」平均值206.4筆記錄，和中筆數組的標題「膽固醇」平均值63.0筆記錄，這二個標題平均值較高和有資訊超載問題外，其餘標題的平均記錄值看來都還可以。

然而，正如作者已經在「模糊邏輯在主題分析的應用：標題權值的計算方式」一文中點出的 [註20]，這祇是表面的假象。一如下面的表4，表面的統計平均值看來沒問題，但是細部分析後，會發現大部份的資料仍然集中在極少數的（複分）細目中，因此資料過度集中的現象仍然存在，所以資訊超載的問題依然無法解決。有興趣的讀者，請自行參閱「模糊邏輯在主題分析的應用：標題權值的計算方式」一文。

表4. 四個隨意標題的結果結果表格。[註21]

標題名稱	標題存在形式（主題複分）數目	資料總筆數	平均值（資料總筆數/主題複分數目）
天主教	67	720	10.8
宗教	284	1900	6.7
心理學	64	1083	16.9
電腦	512	10010	19.6

三、主題複分效用的細部分析

作者在「中文圖書標題表的使用情況分析：以NBINet為例」一文中，除了發現約有26%（即1/4強）的標題從未被使用過，也觀察到約有23%（即1/4弱）的標題，祇有1種標題存在形式，換言之，這些標題並未有使用到任何複分的功能。此外也發現到每個標題的複分數目平均為9.36（以61個標題計算，扣除21個存在形式數目為0的標題）。然而，大部份標題的複分數目在1-5（含），有40個標題，約佔61個標題（扣除21個存在形式數目為0的標題）的66%。

為了分析主題複分效用的表面假象，和更細緻的了解每個標題的複分情況，作者根據每個標題個別複分的記錄筆數，重新整理出表5和6二個表格如下，並且抽出每個標題前三多複分的記錄筆數，再計算其所佔的百分比（該複分的記錄筆

數 / 該標題的總記錄筆數)，最後增列平均值以便比較。每個表格都含有標題名稱、第1高複分的記錄數目、第2高複分的記錄數目、第3高複分的記錄數目、和平均值（資料總筆數/主題複分數目）等五項資訊。

表5. 高筆數組標題個別複分前3高的記錄筆數和百分比結果。[註19]

標題名稱	第1高複分的記錄數目和百分比	第2高複分的記錄數目和百分比	第3高複分的記錄數目和百分比	前3高複分的總記錄數目和百分比	平均值（資料總筆數/主題複分數目）
產科	317（71.2%）	48（10.8%）	27（6.1%）	392（88.1%）	18.5
共產主義	185（38.9%）	127（26.7%）	41（8.6%）	353（74.2%）	10.8
民俗音樂	55（28.8%）	22（11.5%）	19（9.9%）	96（50.3%）	4.1
中等教育	689（34.7%）	329（16.6%）	113（5.7%）	1131（57.0%）	12.8
時間	551（80.0%）	96（13.9%）	9（1.3%）	656（95.2%）	36.3
食物治療	3049（98.5%）	16（0.5%）	14（0.5%）	3079（99.5%）	206.4
初等教育	80（38.6%）	51（24.6%）	9（4.3%）	140（67.6%）	5.9
眼	123（78.8%）	16（10.3%）	7（4.5%）	146（93.6%）	6.2
景觀工程	149（52.5%）	88（31.0%）	11（3.9%）	248（87.3%）	17.8
夫妻	212（94.6%）	4（1.8%）	4（1.8%）	220（98.2%）	32.0

表6. 中筆數組標題個別複分前3高的記錄筆數和百分比結果。[註19]

標題名稱	第1高複分的記錄數目和百分比	第2高複分的記錄數目和百分比	第3高複分的記錄數目和百分比	前3高複分的總記錄數目和百分比	平均值（資料總筆數/主題複分數目）
文字學	51（55.4%）	13（14.1%）	6（6.5%）	70（76.1%）	5.4
投影幾何	32（100%）	0（0%）	0（0%）	32（100%）	32.0
電腦犯罪	44（84.6%）	2（3.8%）	1（1.9%）	47（90.4%）	6.5
膽固醇	63（100%）	0（0%）	0（0%）	63（100%）	63.0
生活教育	54（74.0%）	6（8.2%）	4（5.5%）	64（87.7%）	8.1
兵家	28（48.3%）	19（32.8%）	4（6.9%）	51（87.9%）	7.3

器械體操	37（100%）	0（0%）	0（0%）	37（100%）	37.0
泛函分析	62（98.4%）	1（1.6%）	0（0%）	63（100%）	31.5
貨幣政策	27（32.9%）	15（18.3%）	13（15.9%）	55（67.1%）	5.5
奧林匹克運動會	57（61.3%）	12（12.9%）	8（8.6%）	77（82.8%）	9.3

從表5可以看出，以高筆數組的標題來說，第1高複分的記錄數目，其記錄數目從55筆到3049筆，其中半數標題的第1高複分記錄數目超過200筆，祇有2個標題的第1高複分記錄數目在100筆以下。因此不但是全部超出其個別平均值甚多；除一個標題是少幅超越外，其餘也是超出組的總平均值（35.1筆，參考下面表7）甚多。第1高複分標題，所佔資料總筆數的百分比，則從28.8%到98.5%。

表6亦反映出相同的結果，以中筆數組的標題來說，除了三個無複分的標題外，其餘的第1高複分記錄數目，也超出其個別平均值甚多。第1高複分標題，所佔資料總筆數的百分比，則從32.9%到98.4%（扣除三個無複分的標題）。

如果合計前3高複分標題的記錄數目來看，以高筆數組而言，所佔資料總筆數的百分比則從50.3%到99.5%（參考表5），平均值為81.1%。換言之，4/5強的資料集中在前3高的複分標題中。此種資料分佈極不平均的情況，亦存在於中筆數組，其前3高複分標題所佔資料總筆數的百分比，從67.1%到100%（參考表6），平均值為89.2%。

以下作者將各組標題之總平均值和最高複分標題的各種比較，整理成表7如下。從下表可以清楚看出，第1高複分的平均百分比在60%以上，高筆數組第1高複分的平均記錄數目更達541筆，而總平均值35.1筆，相差幾近506筆記錄，由此可見表面的總平均值假象嚴重失真，掩蓋了實際上嚴重的使用者資訊超載問題。此結果證實了作者在「模糊邏輯在主題分析的應用：標題權值的計算方式」一文中 [註22]，根據隨意使用的四個通用詞彙所得到的結果吻合，因此主題複分並無法達成預期的功效和解決資訊超載的問題。

表7. 各組標題總平均值和最高複分標題的各種比較。[註19]

	高筆數組	中筆數組
總平均值（資料總筆數/主題複分數目）	35.1	20.6
第1高複分的平均記錄數目	541	40.5

第1高複分的平均百分比	61.7%	75.5%
第1高複分中最高記錄數目	3049	63
第1高複分中最低記錄數目	55	27

四、結語

作者在「模糊邏輯在主題分析的應用：標題權值的計算方式」一文中，曾利用國家圖書館的NBINet來進行簡易的實驗，以探討目前主題分析作業的效益和主題檢索的便利性，結果發現資訊超載和資料分佈不平均的情況嚴重，同時目前圖書館主題分析作業中所採行的主題複分，並不足以解決資訊超載的問題。[註23]

後來作者在「光碟資料庫中主題分佈頻率的初步調查」一文中 [註24]，以圖書資訊界經常使用的library Literature和Lisa兩種光碟資料庫為對象，進行同樣的調查，結果顯示光碟資料庫上主題分佈不平均現象亦存在，並足以造成使用者的資訊負荷超載。

接著為了解標題在編目上的使用情況，作者在「中文圖書標題表的使用情況分析：以NBINet為例」一文中，藉由隨機抽取的標題在國家圖書館NBINet來進行主題檢索，結果後發現約有26%的標題從未被使用過。同時少數1/8的標題，卻佔全體使用機率的88%，顯示絕大部份的資料，集中在極少數的標題上。[註25]

本文是上述研究的延伸，因為圖書館一向透過主題複分方式，一方面將資料加以更詳細的細分，來協助使用者更準確的找到其所想要的資料；另一方面也使資料能分到各細目下，以降低資料的集中度。所以本文的重點在關注主題複分與資訊超載的關係，和主題在書目資料庫中的資料分佈狀況。

為了探究主題複分的實際功效，作者自「中文圖書標題表」隨機抽取80個標題，然後利用國家圖書館NBINet來進行主題檢索分析，實驗進行時間在西元2002年4月5-7日三天。

實驗結果顯示，雖然每個複分標題平均記錄數目並不高，例如高筆數組（每個標題資料總數100以上）的平均值為35.1筆（參考表7），組內除了一個標題的平均筆數特別高（206.4筆），絕大多數都在20以下（參考表1）。但這祇是表面的假象，進一步的分析發現，主題複分並未發揮其預期的功能，因為大多數的資料仍然集中在極少數的複分標題中。

以高筆數組的標題來說，第1高複分的記錄數目，其記錄數目從55筆到3049筆，其中半數標題的第1高複分記錄數目超過200筆，祇有2個標題的第1高複分記錄數目在100筆以下。因此不但是全部超出其個別平均值甚多（參考表5）；除一個標題是少幅超越外，其餘也是超出組的總平均值（35.1筆）甚多。第1高複分標題，所佔資料總筆數的百分比則從28.8%到98.5%。

如果合計前3高複分標題的記錄數目來看，以高筆數組而言，所佔資料總筆數的百分比則從50.3%到99.5%（參考表5），平均值為81.1%。換言之，4/5強的資料集中在前3高的複分標題中。此種資料分佈極不平均的情況，亦存在於中筆數組（每個標題資料總數31-100），其前3高複分標題所佔資料總筆數的百分比，從67.1%到100%（參考表6），平均值為89.2%。

綜括來說，不但是絕大部份的資料，集中在極少數的標題上 [註26]；而且在這些少數的超高使用頻率的標題（即高筆數組）中，大多數的資料又集中在極少數的複分標題中，高筆數組第1高複分的平均記錄數目高達541筆。由此看來，主題複分並未完全發揮實際功效，而且此部分的使用者資訊超載問題是非常嚴重。

註釋

註 1：黃淵泉，中文圖書分類編目學（台北市：學生書局，民 85 年 4 月），頁11。

註 2：薛理桂，「分類與編目之發展趨勢」，國立成功大學圖書館館刊第1 期（民87 年 4 月），頁36-48。

註 3：盧秀菊，「中文主題標目與標題表」，中國圖書館學會會報59 期（民 86 年12 月），頁 26。

註 4：曾繁絹，「中文標題檢索效益之研究--以國立臺灣大學TULIPS系統為例」，大學圖書館第2卷第1 期（民 87 年 1 月），頁101。

註 5：陳麥麟屏和林國強，美國國會圖書館主題標目（台北市：三民書局，民 78年 12 月），頁27-29。

註 6：同註5，頁41-48。

註 7：同註5，頁49-73。

註 8：陳昭珍，「主題檢索理論之探討—主題分析（上）」，書農9 期（民 81 年 12 月），頁 11-27。

註 9：陳佳君，「從知識結構探討主題分析」，書府16 期（民 84 年 6 月），頁 38-47。

註 10：同註3，頁 31-38。

註11：侯漢清，「評《中文圖書標題表》—兼談標題表的敘詞化改進」，圖書與資訊學刊31 期（民 88 年 11 月），頁 17-23。

註12：同註4，頁100-123。

註13：吳政叡，「中文圖書標題表的使用情況分析：以NBINet為例」，審查中。

註14：實驗所使用的「中文圖書標題表」，乃是直接取自國家圖書館編目組網站中編目規範標準之「中文圖書標題表」，網址是http://192.83.186.1/catweb/2-1-4.htm。

註15：文獻上發現NBINet的中文名稱有二個—「全國圖書資訊網路系統」和「全國圖書書目資訊網」，在國家圖書館有關NBINet簡介的網頁中（http://NBINet.ncl.edu.tw/screens/libinfo_chia.html），這二種名稱都有使用，不過在正式的「全國圖書資訊網路系統合作編目要點」中，是使用「全國圖書資訊網路系統」一詞，因此本文也使用此詞。

註16：國家圖書館「NBINet簡介」的網頁網址為http://NBINet.ncl.edu.tw/screens/libinfo_chia.html。

註17：吳政叡，「模糊邏輯在主題分析的應用：標題權值的計算方式」，圖書與資訊學刊 40 期（民 91 年 2 月），頁 10-17。

註18：同註17，頁12-13。

註19：請注意表1中「標題存在形式數目」的用法，標題存在形式數目0表示標題在書目資料庫中不存在，標題存在形式數目1表示書目資料庫中祇有標題本身，未有任何其他的複分形式存在。此外，NBINet的書目紀錄有重複情形，換言之，並未將各合作圖書館相同的書目紀錄合併，因此這裡的紀錄數量有偏高趨勢。不過對一個如此龐大的書目資料庫而言（將近373萬筆記錄），重複記錄畢竟是極少數，同時也是稀釋散佈在全部資料庫中，因此對最後分析結果的影響，應該是可以忽略的。至於NBINet書目資料庫的其他缺失，在「NBINet合作編目資料庫內容發展之探討」一文中已有詳盡

的描述和探討。林淑芬，「NBINet合作編目資料庫內容發展之探討」，國家圖書館館刊88年2 期（民 88 年 12 月），頁 4-5。

註20：同註18。

註21：此四個標題的資料來自「模糊邏輯在主題分析的應用：標題權值的計算方式」一文，請參見註17。

註22：同註18。

註23：在此項簡易的實驗中，作者祇是隨意使用四個通用的詞彙--天主教、宗教、心理學、電腦，因此其立論的基礎並不穩固，尚需使用更嚴謹的方式來進一步查核。吳政叡，「模糊邏輯在主題分析的應用：標題權值的計算方式」，圖書與資訊學刊 40 期（民 91 年 2 月），頁 10-17。

註24：吳政叡，「光碟資料庫中主題分佈頻率的初步調查」，審查中。

註25：同註13。

註26：同註13。