模糊邏輯在主題分析的應用：標題權值的計算方式

吳政叡 (Cheng-Juei Wu)

輔仁大學圖書資訊系專任副教授

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

中文摘要

作者首先藉由在國家圖書館NBINET上進行的幾個主題檢索個案分析，展現目前圖書館透過主題複分來克服資訊超載問題的方式仍然有很大的侷限性。接著作者提出導入模糊邏輯概念，來解決目前主題分析在設定標題時，因為採用古典的二元邏輯所面臨的困境和不足之處。最後，對於模糊邏輯在主題分析的實際應用方式上，本文提出一套標題權值的給定標準和計算公式，並詳細闡釋此公式的特色與優點。

==========

The Application of Fuzzy Logic on Subject Analysis: A Formula for Subject Weights

Besides classification, for the moment, subject analysis is the main tool for cataloguers to expose the contents of materials such as books; however, this tool is inflexible because of the use of classical binary logic. In this study, we wanted to conquer this shortage by using Fuzzy Logic. Firstly, we investigated the usability of subject analysis via some keyword searches on NBINET, the biggest bibliographic database in Taiwan. The case study of four subject searches showed that information overload was quite severe in all four cases. This experimental result indicated that the current practice of subject subdivision alone was not sufficient to avoid information overload. In order to conquer this problem, Fuzzy Logic was introduced to the current practice of subject analysis and a formula for subject weights

was proposed in this work. The features and advantages of proposed formula were analysed as well.

關鍵字：主題分析，標題，模糊邏輯，資訊超載，Subject Analysis，Fuzzy Logic，Information Overload，NBINET。

一、前言

作者在去年（西元2000）底曾經受人所託代為收集購買一批天主教相關書籍，準備捐贈給國外某天主教大學的圖書館。當時的第一直覺就是利用國家圖書館的全國圖書書目資訊網（簡稱NBINET）來收集書目資訊，可是發現利用主題「天主教」來檢索時，有數百筆回覆的書目資料，當下即望著螢幕興嘆，雖然開始時耐著性子勉強一筆筆過濾，但是在察看數十筆後，最後仍然不得不放棄。這個親身經歷使作者體認到人類認知與處理能力的有限，也體會到「資訊超載」這個現代人普遍共有的夢魘。

人類認知與處理能力的有限，和資訊超載現象，是自遠古以來即存在的事實。當資料累積超過一定數量時，單憑人類的記憶能力無法應付時，如何對資料加以適當描述，成為有效利用和整理資料的必然作法，因此自古以來圖書著錄一直是圖書館員的首要核心工作。有關圖書著錄一詞的解釋，根據黃淵泉在《中文圖書分類編目學》一書中的定義為--「是將書籍的內容和形式特徵，按既定的編目規則記錄起來，以方便讀者來利用或是圖書館員來管理。」 [註 1] 現在另外一個常使用的同義詞是資訊組織。 [註 2]

除了基本的書目資料，如書名、作者、出版社、與出版日期等，分類號和主題（subject）可以說是圖書著錄人員用來揭露資料內容的兩個主要工具，其中分類號因為肩負排架功能，祇能有一個，因此如薛理桂在『分類與編目之發展趨勢』一文中提及 [註 3]，傳統的分類法，是採用「單一分類 -- 單一位置」（single classification--single placement）概念。無論是國會圖書館分類法或是杜威十進分類法，均無法克服此缺失。由此可知，分類號不若主題來的靈活，所以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。

主題（subject）或標題（subject heading）在黃淵泉《中文圖書分類編目學》一書定義為：「一個詞或一組字用以表示資料討論的主題。」 [註 4] 標題的重要性由以前圖書館在卡片目錄盛行時代，一般均會有書名目錄、分類目錄、作者目錄、與標題目錄等四種目錄可以窺知[註5]。

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6]）與詞彙的選用、主題標目的結構[註7]、主題標目的複分方式[註8]、主題分析的作法[註9-10]、標題表缺失 [註 11-12]等，近年來雖然也有許多探討主題與檢索效益的關係，例如『中文標題檢索效益之研究--以國立臺灣大學TULIPS系統為例』[註13]，不過仍然是在傳統的主題標目結構下來探討。

至於本文中的另外一個涉及領域--模糊邏輯，它是由Zadeh首先在1965年的『Fuzzy Sets』一文中提出的概念 [註 14]，最初並未在美國受到重視，反而是在歐洲和日本陸續有一些研究和實驗計畫在不斷的推動[註 15]。直至1980年代開始，日本人首先成功將模糊邏輯應用在家電產品上，並且推出很多模糊邏輯的家電商品後，模糊邏輯才在美國本土重新受到重視，並且在全世界的學術圈興起研究熱潮。時至今日，模糊邏輯已成為機器學習（或是機器智慧）的一個重要分支領域。雖然模糊邏輯最成功的應用是在控制方面[註 16]，但是在檢索（或是模糊檢索系統）方面的研究也為數不少[註 17-22]，祇是似乎未能形成一股氣候，也未能影響或是落實到圖書館的主題分析作業上。

雖然模糊邏輯正如其名稱所暗示，與哲學中的「邏輯」有密切的關係，不過一般模糊邏輯的教科書都是從數學中的「集合」角度來切入和介紹，事實上目前模糊邏輯的應用，不管是在控制方面或是檢索部分，利用數學中的「集合」遠比使用哲學中的「邏輯」來的便利。

以「集合」的術語來說，古典的二元邏輯可以對映到古典（或是一般）的集合，而模糊邏輯則對映到「模糊集合」（Fuzzy Set）。兩者的不同在於，一般集合中，元素與集合的關係祇能是屬於（1）或者是不屬於（0）；但是在模糊集合中，元素與集合的關係可以是屬於到某種程度，以成員值（Membership Value）或權值（Weight）來表示，其值的範圍在0和1之間（[0, 1]）。

同樣的，在檢索操作中常常使用的布林邏輯，也可以對映到集合上。布林邏輯中的「and」可以對映到集合的「交集」（Intersection），「or」對映到集合的「聯集」（Union），「not」對映到集合的「補集」（Complement）。由於在一般集合中，交集、聯集、補集都已有明確的定義和操作計算方式，而模糊集合也已經將此三種操作（事實上是幾乎所有古典集合中的操作）做了適當的展延，因此任何一般集合能夠處理的，模糊集合也能加以處理，事實上古典集合可以說成是模糊集合的一個特例。

有關模糊邏輯的基本概念，以及一般集合和模糊集合的基本觀念和操作，在Ross的《Fuzzy Logic with Engineering Applications》一書中有詳盡清楚的闡釋[註 23]，有興趣的讀者請自行參閱。至於（模糊）檢索系統的基本模式，請參考

『Analysis of Fuzzy Operators for High Quality Information Retrieval』一文。[註 24]

由於根據研究顯示，主題檢索是圖書館使用者在線上檢索時最常用的方法[註25-26]，因此本文探討如何應用模糊邏輯的概念於主題分析，來改善主題檢索時資訊超載的情況。

二、主題檢索個案分析

雖然書籍的數目和成長速度遠不及網頁，但是現存的書目記錄，以OCLO而言，已有數千萬筆書目記錄；即便是臺灣地區為主的全國書目網（NBINET），截至西元2000年8月止，也累積有170萬筆以上的書目記錄。[註 27] 在此種數量下，正如作者在前言中所提及的親身經歷，即便是應用現有主題分析中的技巧，如各種複分方式，仍不足以解決資訊超載的問題。

以「天主教」為主題來利用國家圖書館的NBINET檢索時 [註 28]，總共有720筆資料分佈在67個以「天主教」起頭的各式各樣複分標題 [註 29]，例如：天主教、天主教--中國、天主教--中國--建築等。表面看來每一標題平均有10.75筆資料，似乎利用複分可以有效來解決資料數量過多的問題。然而進一步細究卻發現每一標題的資料分佈情況是極為懸殊和不平均，以此個案為例，有最多資料筆數的前四個標題分別為天主教--信仰、天主教、天主教--傳道、天主教--祈禱，分別有167、119、95、92筆資料，因此前四多資料的標題總共有473筆，佔全體資料筆數的65.7% ，同時超過一半的標題祇有一筆書目資料。從以上的個案分析，可以清楚顯示光是依靠主題複分並不足以解決資訊超載的問題，因為前四多資料的標題都有近百筆書目資料，很顯然是超過使用者的負荷能力。

此種資訊超載的情況，將隨著主題詞彙的常用性增加而日益惡化，例如使用較「天主教」更常用的主題「宗教」來檢索時，總共有1900筆資料分佈在284個複分標題，然而其資料分佈情況依然是極為不平均的。如宗教有212筆，宗教--中國有92筆等。

此種資訊超載和資料分佈不平均的情況，在其他領域同樣存在，作者以「心理學」來檢索時，總共有1083筆資料分佈在64個複分標題，其中心理學一詞即佔670筆（為資料總數的62.9% ），資料分佈不平均的情形更為嚴重。同樣的情形亦發生在以「電腦」檢索時，總共有10010筆資料分佈在512個複分標題，其中電腦一詞佔1958筆（為資料總數的19.6% ），資料分佈不平均的情形同樣嚴重。

由上述的個案分析可知，越是普遍和常用的主題詞彙，資訊超載和資料分佈不平均的情況也越嚴重。尤其當單一主題詞彙（如心理學和電腦）的書目資料

達數百筆，甚而達數千筆時，使用者常常祇能望而興嘆，偏偏這些常用的主題詞彙，卻是使用者在主題檢索時較常使用的詞彙，使得資訊超載的問題更是不容忽視。

三、古典和二元邏輯現行主題分析作法的缺失

從模糊邏輯的角度來說，現存的主題分析作法有如古典的二元邏輯，當編目人員針對某書給與一些標題時，這些出現的標題可以視為其權值為1，其餘標題表中的標題權值則都為0。

這種非1即0的二元分法，很明顯有一些缺陷存在。以厚達數百頁且資料含蓋面較廣的書來說，即有以下可能的潛在問題：

（1）如果編目人員給了3個標題，是否代表這3個標題都同等重要，或者是都跟該書完全相關。

（2）由於有非1即0的限制，極有可能使編目人員祇考慮完成相關的標題，使得標題數目偏低，因此一本書的內容無法完全顯示出來，例如作者在利用NBINET來進行「天主教」的主題檢索時，即發現很多書目資料祇有1或2個標題。

（3）由於有非1即0的限制，使編目人員在主題複分時可能非常猶豫，而傾向以較籠統的總稱（例如電腦、心理學等）為限，造成主題複分在減少資訊負載，和防止使用者資訊超載的功能不彰[註 30]，正如前面一節中，針對四個標題（即天主教、宗教、心理學、電腦）所做的主題檢索結果顯示，主題複分並未能發揮功用，且籠統總稱標題的資料筆數都很多，較嚴重的如心理學，甚至佔到資料總數的62.9% 。

（4）使用者的檢索需求和情境各有不同，有人可能希望收集任何有一些關連的資料，另外也有祇要完全相關的，面對這些林林總總不同的檢索需求，二元邏輯非1即0的傳統主題分析作法，是無法面面俱的。

作者認為上述的問題，可以藉由打破較僵硬的二元邏輯概念，改採較具彈性的模糊邏輯觀念而得以解決。因為基本上模糊邏輯是以0-1的權值區間（即[0,1]）來取代古典二元邏輯的非0即1，此時編目人員可以根據某些標準來設定標題的權值，如0、0.1、0.5、0.8、1等，來區分不同標題在該書的重要性或是相關性。另一方面，使用者也可以根據自身的檢索需求來設立資料過濾的門檻（threshold），例如祇有標題權值達到0.8（含以上）的資料才要。

四、標題權值的計算方式

如何導入模糊邏輯於圖書館實際的主題分析作業上，作者以為關鍵在製作一套圖書館編目人員和使用者皆可輕易了解和運用的標題權值給定標準和方法。一方面，圖書館界習於透過合作編目和書目資料共享來減輕編目的負荷與服務讀者，自然而然希望標題權值的給定，能有一套大家可共同遵循的標準和程序，才不致造成編目人員在進行主題分析時無所適從，和增加書目資料共享時的困擾。另一方面，若無一套標題權值的給定方法，使用者在進行主題檢索時也無法知道該用什麼權值來查詢，以及該權值所代表的意義。

至於這套標題權值的給定標準，作者有如下的初步建議：首先，權值的級數須要恰當，級數太少，將無法達到透過模糊邏輯來減少使用者在主題檢索的資訊負載，以及避免資訊超載的功能，因此也失去導入模糊邏輯於主題分析作業的意義；級數太多，將使權值變得過於瑣碎，鄰近權值間的差異太小，一方面造成編目人員作業時的不便與困擾，一方面對於使用者在實際檢索的幫助也不大。

作者以為恰當的權值級數為10（即0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9、1）或是20（即0.05、0.1、0.15、…、0.9、0.95、1），前者10個級數適用較小規模的書目資料庫，後者20個級數適用於大規模或者是全國性（如國家圖書館的NBINET）的書目資料庫。

其次作者建議如下的標題權值計算方式：

（1）計算該標題實際所佔篇幅比例。如果是約佔15% 的篇幅，則換算成0.15。

（2）編目人員評估該書內容有關此標題的深度（或者重要性），給與恰當的值。值的範圍一如模糊邏輯，限定在0和1之間，1是最高值，0則是最低值。

（3）計算上述兩個數據的平均值（作法是將上述的（1）和（2）的值相加後，再除以2，即可求得平均值），並且根據所採用的級數，採取最相近級數的值，即為該標題的權值。[註 31]

舉例來說，以標題「天主教」而言，若是一本宗教方面書籍，其該方面篇幅佔50% ，則其在篇幅方面的值為0.5；接著編目人員評估該書內容，給其深度方面的值為0.8。則該標題的權值計算如下：(0.5+0.8) /2 = 0.65，因此當採用20級數時，權值即為0.65。

作者所建議的這個標題權值給與方式有如下的特點：

（1）一般而言，資料的權值分布可能會類似金字塔，權值小者在底部，而權值大者（如1）在頂端。此種資料的分布情形，對資料檢索而言是最適宜的。

（2）使用者操作容易，且擁有完全的主控權。使用者可以採用從權值1開始搜尋，然後逐步降低檢索的權值來擴大資料的範圍，因此使用者可以自行控制資料的搜尋和過濾範圍。

（3）適合祇須少數完全相關資料的主題檢索情境。因為計算公式混合篇幅比例和內容深度評估，會大幅降低標題權值為1的資料數目，除非該書全部篇幅都屬於該標題，而且內容深度很完整。這種標題權值為1很難達成的情形，對祇須要非常少數完全相關資料的主題檢索情境是非常有助益的，可以避免資料超載的情形。

（4）標題權值的計算公式混合客觀因素（篇幅比例）和主觀因素（內容深度評估），可以達成主客觀因素間的良好平衡。主客觀因素的混合，既可以避免因為編目人員間的主觀認定不同，造成標題權值的差異過大（因為有客觀的篇幅比例來沖淡差異性）；又給與編目人員主動調整權值的彈性，畢竟編目人員的專業認知，在可預見的將來仍是無法取代的。如此一來，也可避免有些書籍雖然篇幅比例高，但皆是泛泛之言，卻可以有高權值的情形發生。

（5）標題權值的計算公式簡易且易於施行。權值是由兩個數據的平均而來，計算方式簡易。此外篇幅比例的數據很容易求得，同時內容深度評估的數據，祇要館內訂定一套簡易的評估準則，也不難求得。

五、結語

主題分析可以說是目前圖書著錄人員用來揭露資料內容的主要工具，同時根據研究顯示，主題檢索也是使用者在線上檢索時最常使用的方法，因此其重要性是不言而喻的。

為了探討目前主題分析作業的效益和主題檢索的便利性，作者利用國家圖書館的NBINET來進行簡易的實驗，在四個主題（天主教、宗教、心理學、電腦）檢索個案分析中，實驗數據顯示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。

更進一步細究目前主題分析的作業方式，可以發現圖書著錄人員在設定標

題時，祇能在兩極化的要與不要中做抉擇，毫無彈性可言。這些限制造成一些潛在的問題，如：標題數目偏低、主題複分不足、無法滿足不同的檢索需求等。

作者認為目前主題分析作業方式上的這些缺失，藉由改採較具彈性的模糊邏輯觀念來取代僵硬的二元邏輯概念，可以得到適當的解決。因為基本上模糊邏輯是以0-1的權值區間（即[0,1]）來取代古典二元邏輯的非0即1，此時編目人員可以根據某些標準來設定標題的權值，如0、0.1、0.5、0.8、1等，來區分不同標題在該書的重要性或是相關性。因此模糊邏輯能使編目人員在給標題時有更大的彈性，能更貼切傳達資料在各探討主題的相對重要性。

另一方面，使用者也可以根據自身的檢索需求來設立資料過濾的門檻（threshold），例如祇有標題權值達到0.8（含以上）的資料才要。因此模糊邏輯能減少資訊負載，避免資訊超載，使資料檢索的效率更佳。

文獻分析顯示，雖然在1980年代以後，隨著學術圈興起研究模糊邏輯的熱潮，也有不少關於模糊邏輯在檢索方面的應用研究，或是模糊檢索系統的研究發表。可惜似乎由於這些研究太偏重電腦的技術層面，或是未能貼切圖書館實際的主題分析作業，因此目前均未能影響或是落實到圖書館的主題分析作業上。

為能實際導入模糊邏輯於圖書館實際的主題分析作業上，作者初步建議一套圖書館編目人員和使用者皆可輕易了解和運用的標題權值給定標準和方法。首先，作者建議權值的級數採用10或是20較為恰當；其次在標題權值計算方式方面，提議混合客觀因素（篇幅比例）和主觀因素（內容深度評估）來達成主客觀因素間的良好平衡。因為此種混合，既可以避免因為編目人員間的主觀認定不同，造成標題權值的差異過大（因為有客觀的篇幅比例來沖淡差異性）；又給與編目人員主動調整權值的彈性，畢竟編目人員的專業認知，在可預見的將來仍是無法取代的。

最後，由於現行的機讀編目格式在欄位的安排上，並沒有針對標題權值的設計，因此機讀編目格式必須要做小幅度的調整來導入模糊邏輯概念，不過這種調整在技術上是非常容易的。

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 55。

註 2：陳昭珍，「電子資訊的組織模式」，圖書館學刊12 期（民 86 年 12 月），頁 163-164。

註 3：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 4：同註1，頁11。

註 5：同註1，頁27-28。

註 6：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 7：同註6，頁41-48。

註 8：同註6，頁49-73。

註 9：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10：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11：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 31-38。

註12：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13：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100-123。

註14：L.A. Zadeh, "Fuzzy Sets," *Inform. Contr.* 8 (1965): 338-353.

註15：有關模糊邏輯的一些早期研究和實驗計畫，請參閱Kandel and Yager在1979年所發表的一篇文章。A. Kandel and R.R. Yager, "A 1979 Bibliography on Fuzzy Sets, Their Applications, and Related Topics," in *Advances in Fuzzy Set Theory and Applications*, ed. M.M. Gupta, R.K. Ragade, and R.R. Yager (Amsterdam: North Holland, 1982), 621-744.

註16：C.C. Lee, "Fuzzy Logic in Control Systems: Fuzzy Logic Controller," *IEEE Trans. Syst. Man. Cybern.* 20(2) (1990): 404-435.

註17：D. A. Buell, “A Problem in Information Retrieval with Fuzzy Sets,” JASIS 36(6) (1985): 398-401.

註18：G. Bordogna, P. Carrara, and G. Pasi, “Query Weights As Constraints in Fuzzy Information Retrieval,” Information Processing & Management 27(1) (1991): 15-26.

註19：D. Lucarella and R. Morara, “FIRST: Fuzzy Information Retrieval System,” Journal of Information Science 17 (1991): 81-91.

註20：M. H. Kim, J. H. Lee, and Y. J. Lee, “Analysis of Fuzzy Operators for High Quality Information Retrieval,” Information Processing Letters 46 (1993): 251-256.

註21：G. Bordogna and G. Pasi, “A Fuzzy Linguistic Approach generalizing Boolean Information Retrieval: A Model and Its Evaluation,” JASIS 44(2) (1993): 70-82.

註22：S.-M. Chen and J.-Y. Wang, “Document Retrieval Using Knowledge-Based Fuzzy Information Retrieval Techniques,” IEEE Trans. Systems, Man, and Cybernetics 25(5) (1995).

註23：該書前半部（第1- 8章）是模糊邏輯的基本概念和操作介紹，前半部（第9- 15章）是模糊邏輯各種應用的介紹。正如前述，該書主要從數學中的「集合」角度來切入，並且對如何從一般的集合推展到模糊集合，有很鮮明的對照和清楚的闡釋。T. J. Ross, Fuzzy Logic with Engineering Applications（N.Y.: McGraw-Hill, 1995）.

註24：同註20，頁251。

註25：同註11，頁26。

註26：同註13，頁101。

註27：書目記錄筆數的統計資料，取自國家圖書館有關NBINET簡介的網頁，http://nbinet.ncl.edu.tw/screens/opacmenu_chia.html，更詳細的統計資料請自行參閱該網頁。

註28：本文中所有的四次主題檢索--天主教、宗教、心理學和電腦，其時間均在西元2001年6月29日。

註29：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。

註30：有關資訊負載和資訊超載二者的定義與相互關係，請參閱蘇諼的「談資訊超載對圖書資訊專業的影響」一文之頁68。另外主題複分的功能則在頁71。蘇諼，「談資訊超載對圖書資訊專業的影響」，中國圖書館學會會報65期（民 89 年 12 月），頁67-74。

註31：當計算出來的平均值剛好在前後兩個級數中間時，編目人員可根據館內政策自取兩者之一。