教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 54 , no. 1 (2017) : 69-95

以開放資料的教師學術專長彙整表為基礎之學科標準分類分析

Analyses of the Standard Classification of Fields Based on the Directory of Faculty Expertise Open Data

林 頌 堅* Sung-Chien Lin*

Assistant Professor

E-mail：scl@cc.shu.edu.tw

English Abstract & Summary see link at the end of this article

教育資料與圖書館學　54 : 1 (2017) : 69-95
DOI:10.6120/JoEMLS.2017.541/0046.RS.AM

以開放資料的教師學術專長彙整表為基礎之學科標準分類分析

林頌堅

摘要

本研究以教育部提供的開放資料做為分析資料，利用學術專長文字資料的相似性，探討目前對大學校院系所進行分類所採用的學科標準分類，並提出改善的建議。使用的技術包括估計學術專長資料相似性的Word2Vec文字比對技術，以及分析分類性質的階層集群分析、多維尺度分析、輪廓試驗、近似學門分布和系所相似性統計等評估方法，並以資訊視覺化方法呈現研究的結果。研究結果指出目前的學科標準分類在分類結構、分類架構與資料品質上都必需要改善，才能符合教育統計、政策制訂與學術交流的需求。

關鍵詞：學科標準分類，開放資料，教師學術專長彙整表，輪廓試驗，Word2Vec

緒　論

近年各級政府為因應人民的需求與世界潮流，紛紛成立政府開放資料平台，公開許多有價值的調查資料，供民眾檢索與加值利用。在這個大潮流下，教育部也開放了許多教育相關的資料，除了原先可從教育部統計查詢網（https://stats.moe.gov.tw/）檢索多種教育相關的統計資料以外，目前在政府資料開放平台（http://data.gov.tw/）上，教育部也提供了許多原始的調查資料，可供民眾加值利用，例如「大學問」網站（http://www.unews.com.tw/）便是利用政府開放資訊，提供大學技專校系介紹以及職涯發展，做為升學時的參考。

對系所依據其主題劃分成領域、學門與學類等學科分類，有助於了解系所在學術領域的定位，擬定教學與研究計畫，同時也能提供具有相近學術專長的系所交流。政府資料開放平台上的大學校院教師學術專長彙整表，目前共有101、102、103及105年度的調查資料。每一筆教師資料包含「學校代碼」、「學校公私立」、「學校體系」等多個欄位資料。欄位資料的「領域名稱」、「學門名

世新大學資訊傳播學系助理教授
* 本文通訊作者：scl@cc.shu.edu.tw

2016/10/13投稿；2017/01/03修訂；2017/02/15接受

稱」與「學類名稱」是教育部統計處參照聯合國教育科學文化組織（UNESCO）「國際教育標準分類」（International Standard Classification of Education, ISCED）訂定的，並參酌專家學者意見，加以彙整修編完成（教育部，2007）。在教師學術專長彙整表上，依據教育部於民國96年七月頒布的「中華民國教育程度及學科標準分類（第4次修正）」，每一個大學校院系所或學程[1]都被指定一個領域、學門與學類名稱，領域包括教育、人文及藝術等八個領域，各領域下分為若干個學門，每個學門再細分為學類。[2]例如國立政治大學教育學系在彙整表上的「領域名稱」、「學門名稱」與「學類名稱」分別是「教育領域」、「教育學門」與「綜合教育學類」，而同一大學的中國文學系則被指定為「人文及藝術領域」、「人文學門」與「中國語文學類」。此外，某些系所因無法歸入各領域、各學門或各學類，則設定為「其他」，103年度的調查資料共有四個系所單位設定為「其他」。「學科標準分類」上各層次的學科分類都是基於主題（subject matter）的相似性聚合而成，也就是建立在各種應用到某類特定問題或為了某些特定目的的知識上，希望在此前提之下，這個分類系統可以提供進行教育統計和比較不同國家的教育系統的用途（UNESCO, 2013）。

因此，若是要整體地瀏覽、檢視或利用目前大學校院裡研究主題的知識領域分布情形，簡單的直覺做法可依據教師學術專長彙整表上學科標準分類系統的階層式結構進行。然而此一分類系統大多由依據各系所名稱分派指定（教育部，2007，頁II），這樣的分類是否能符合系所教學與研究主題的現況？再加上學術研究進展快速，新的研究主題不斷出現，而舊有的主題則不斷地調整與融合，研究人員間大量而快速地交流各自的研究主題知識。此一分類系統的修訂已接近十年，是否能跟得上學術研究主題的變化？[3]

所以本研究嘗試從教師學術專長彙整表上的其他資料著手，提供對學科標準分類系統的另一個應用、驗證與思考的方向。在教師學術專長彙整表上，每位教師的學術專長都是各自依據本身的教學與研究主題填寫，雖然每位教師提供的學術專長資料大都在35個字以內，描述本身的教學與研究專長，然而以系所為單位加以彙整後，可透露出各系所整體所著重的教學與研究主題。本研究假設主題相同或相近的系所彼此將會有相似的學術專長資料；也就是說在同一

---

1 ISCED上分類的基本單位為programme，在教師學術專長彙整表則有系所、學程等單位。為了敘述方便，以下以系所做為統稱。

2 教師學術專長彙整表以及其依循的「中華民國教育程度及學科標準分類（第4次修正）」的領域、學門和學類分別對應到ISCED學科標準分類上的大分類（broad fields）、中分類（narrow fields）及小分類（detailed fields）。雖然教師學術專長彙整表的學科分類系統與原先學科標準分類有些許不同，但大多一致，本文將此學科分類系統稱為「學科標準分類」。

3 根據教育部統計處（2016）網站上的記錄，中華民國教育程度及學科標準分類的最近修訂版本為第4次修正版，修訂日期為民國96年（西元2007年）七月。該次修正係參考ISCED 1997年版。第5次修正版則正研修中。

學門內的系所應該比不在同一學門內的系所在學術專長欄位上的文字資料更為相似。本研究以學門及系所為分析單位，從學術專長資料的相似性，探討以下的問題：

1.目前尚未有研究對學科標準分類系統上的學門關連進行分析與討論。本研究利用在學術專長上的文字資料相似性，估計學門之間的關連，分析學科標準分類系統上各學門之間的關連與集群，並與原本分類系統上由相關學門組成的領域進行比較。

2.本研究假設相同學門的系所大都應該具有較相近的學術專長，但實際上仍有某些系所的學術專長與學門內其他系所較不相似。因此，本研究將測量與比較同在一個學門內各系所學術專長的一致性，並分析學門內較不一致的系所，其學術專長資料與那些學門較接近。

3.較不一致系所的學術專長資料與本身學門內多個系所相似性較小。本研究將從系所之間的相似性統計，找出各學門較不一致的系所，並探討其與學門內多個系所相似性較小的原因，嘗試從這些訊息中提出改善分類系統的建議，使其較能夠符合學術發展的現況。

本研究並嘗試利用圖形將分析的結果視覺化，提供分析與解釋的參考。

二、文獻探討

目前國內外都未曾有類似利用政府開放資料與學科標準分類系統進行大學校院系所分類系統建置與分析的研究發表。實際上，系所的學科分類可視為一種根據共同主題的領域界分（field delineation）方式，而目前領域界分的研究中，大多根據期刊分類系統（journal classification systems）的主題類別做為研究對象，因此本研究借鏡於期刊主題類別進行分析的相關研究，做為探討系所學科分類的參考。

目前相關研究中，最常被使用來界定研究領域的期刊分類系統是Web of Science和Scopus等資料庫所提供的期刊分類系統，其原因是這兩種資料庫都有數量極大的期刊書目資料，為此兩者皆建置了相當完整的分類系統，提供使用者根據期刊主題進行檢索，並為許多研究者所熟悉。在WoS上有兩個分類系統，一個為具有250個類別的類別系統（a system of categories），另一個則是包含約150個研究領域的研究領域系統（a system of research areas）。此外，另一個分類系統僅包含科學與社會科學，稱為ESI（Essential Science Indicators; Wang & Waltman, 2016）。根據Leydesdorff與Rafols（2009）的說明，ISI（也就是WoS的前身）的主題類別是綜合了引用模式（citation patterns）、期刊題名與專家意見而產生，且也利用類別與期刊在引用資料與被引用資料相似程度的演算法，來協助指定期刊的主題類別（Pudovkin & Garfield, 2002）。Scopus的期刊分類系統名

為ASJC（All Science Journal Classification），分為兩個層級，下層有304個類別，上層則分為27個大類別，但未曾有文獻提到其分類系統的建構方式（Wang & Waltman, 2016）。

然上述期刊分類系統原本用途為了書目發現（bibliographic disclosure），較不適用於分析科學傳播內的潛藏結構（latent structures; Leydesdorff & Rafols, 2009），Glänzel與Schubert（2003）特別針對研究評鑑（research evaluation）的需求，提出15個主類別、67個次類別及一個多領域科學（multidisciplinary science）的分類系統，可用於指定期刊及論文的類別。

此外，許多研究認為，以專家意見產生的分類系統在客觀性與一致性等方面不足，因此利用集群演算法（clustering algorithms），根據期刊的書目資料相似性，將期刊區分成主題相關的集群，建立資料驅動（data-driven）的分類系統。資料驅動的分類系統使用的期刊資料大多為期刊之間的交互引用（cross citations）、共被引（co-citations）與書目耦合（bibliographic coupling）等引用資料為基礎的資訊，也有利用期刊內容上出現的詞語等文字資訊做為期刊特徵，或整合引用與文字資訊，Janssens、Zhang、De Moor與Glänzel（2009）比較了上述各種資訊進行期刊集群的成效。相似性的測量方式則包括著名的餘弦、Jaccard、Pearson相關係數（Pearson's correlation coefficient）等測量方式，Boyack、Klavans與Börner（2005）比較5種利用交互引用以及3種利用共被引的期刊相似性測量方式，共計8種不同方式，在區域準確性（local accuracy）、結構準確性（structural accuracy）、可擴展性（scalability）以及最後的集群品質等成效。

正如Boyack等（2005）的想法，既然有多種不同的期刊分類系統，就有必要對期刊分類系統的各種性質進行評估與比較。最直覺的評估方式是由專家來判斷，但由於沒有任何人能具備足夠專業知識來評估所有學科上的期刊分類情形，則需要相當多的專家參與，因此專家評估很明顯地有極大的困難。目前較通用的做法是根據期刊書目資料的統計與比較來進行評估。Janssens等（2009）將集群結果的性質評估方法分成內部與外部兩種評估方式，本研究認為在期刊分類系統也可比照這個區分方式。

外部評估需要利用另一個現有的標準分類系統進行比較，測量兩個系統上類別成員的相同程度，例如Boyack等（2005）和Klavans與Boyark（2006）使用的局部準確性。這個測量方法利用WoS的類別系統做為標準，統計屬於相同WoS次學科（subdiscipline）的期刊是否在被評估期刊分類系統內能被指定為同一類別。另一個例子是Janssens等（2009）運用ESI的22個領域做為標準，以Jaccard相似性測量計算與新產生的期刊分類系統兩者間期刊與論文資料項目的相同程度；Thijs、Zhang與Glänzel（2015）同樣使用Jaccard相似性作為指標，進行期刊分類系統性質的評估，但他們的比較標準是Glänzel與Schubert（2003）

的15個主類別。很明顯地，外部評估方法最大的困難便是需要事先確定一個良好的分類標準。

內部評估只使用期刊分類系統上期刊書目資料的統計資訊，不需要另一個分類系統做為評估標準。內部評估使用的統計資訊包括各類別本身的資料分布以及發生在類別內與類別之間的資料分布差異。使用各類別本身的資料分布的分類系統評估指標，針對每個類別的書目資料進行統計。例如，Janssens等（2009）統計詞語在ESI每個主題類別的TF-IDF值，以具有較高TF-IDF值的詞語做為該類別的描述語，然後以人為方式，從描述語的語意判斷有較高異質性（heterogeneity）的類別。C.-M. Chen（2008）建議利用類別內所有期刊之間的平均距離評估分類系統上的每個類別，較小的平均距離表示該類別成員期刊之間彼此較靠近，可能較有專一性（specificity）。

另一方面，著名的輪廓（silhouette）測量與模組性（modularity）則是屬於運用發生在類別內與類別之間的資料分布差異。依據輪廓測量的定義，在分類系統的每一種期刊都可以估算得到一個介於–1和1之間的輪廓值。若該期刊與類別內其他期刊之間的平均距離小於該期刊與任何其他類別上期刊的平均距離，它的輪廓值將為正數，代表該期刊被分配到適當的類別，並且輪廓值愈接近1，表示該分配愈適當（C. Chen, Ibekwe-SanJuan, & Hou, 2010; Rousseeuw, 1987），而類別內所有期刊的平均輪廓值可用來表示該類別內成員獲得適當分類的情形，統計期刊分類系統內所有期刊的平均輪廓值則可了解整個系統的分類品質，較大的平均輪廓值表示分類品質較佳。模組性是應用網路分析（network analysis）的概念，針對整個期刊分類系統的分配所進行的測量。將分類系統上每一種期刊視為網路上的節點，它們之間的關係視為節點之間的連結線，當期刊分類系統的分配會使得同一類別內期刊彼此有密集的連結而在不同期刊類別之間的連結較少時，便會得到較大的模組性（P. Chen & Redner, 2010; Newman & Girvan, 2004）。Janssens等（2009）曾分別以交互引用與文本方式為距離估算的基礎，測量22個ESI類別的平均輪廓值；同一研究中，他們也使用了平均輪廓值和模組性比較ESI以及引用、文本與混合等不同期刊距離測量方法所產生的分類系統的品質。除了輪廓值與模組性以外，Wang與Waltman（2016）認為任一期刊引用同一類別所屬的期刊或被這些期刊所引用的頻次必然較其他類別所屬期刊之頻次來得高，因此他們採用期刊之間直接引用（direct citation）發生在類別內與類別之間的分布比例，比較Web of Science和Scopus兩個資料庫所提供期刊分類系統的準確性，並利用這項資訊發現與本身被指定的類別只有較弱連結的期刊，或與未被指定的類別有較強連結的期刊。

最後，利用圖1來總結上述期刊分類系統性質評估方式的討論。另外，也將上述討論提及的指標列於表1。表1提供每一種指標的說明，除了書目資料評

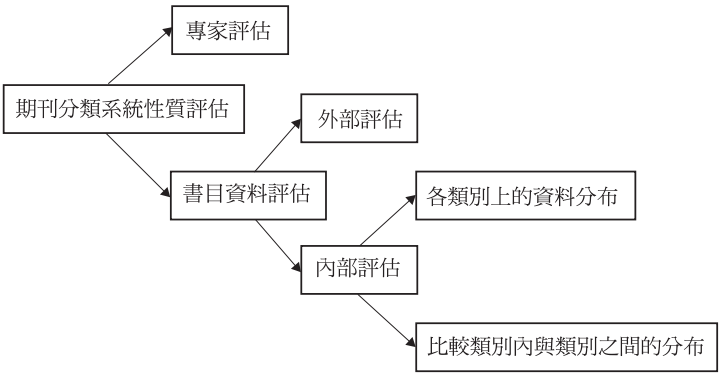

圖1 各種期刊分類系統性質評估方式分析架構

表1 各種期刊分類系統評估指標所屬評估方式、分析單位與計算方式

指標	書目資料評估方式	分析單位	計算方式簡述
局部準確性	外部評估	類別	受試分類系統內各分類成員在標準分類系統為同類別的比例
Jaccard 相似性	外部評估	類別	比較受試與標準分類系統間類別成員相同情形
類別描述語	內部評估：各類別本身的資料分布	類別	挑選各類別上 TF-IDF 值較大的詞語
類別內期刊間的平均距離	內部評估：各類別本身的資料分布	類別	計算各類別成員期刊彼此間的平均距離
輪廓值	內部評估：在類別內與類別之間的資料分布差異	期刊	該期刊在所屬類別的平均距離與在其他類別的最小平均距離之差異
模組性	內部評估：在類別內與類別之間的資料分布差異	期刊分類系統	類別內連結佔全部連結的比例與隨機連結的期望值之差異
期刊間的直接引用	內部評估：在類別內與類別之間的資料分布差異	期刊	發生在類別內及類別間的引用各占全部連結的比例

估的方式以外，還包括它們採用的分析單位，並簡要地敘述計算方式。

另外，就各種評估指標的計算而言，可分成以下幾種類型：

1. 以期刊的類別比例分布評估分類系統：表1上的外部評估類指標都以這種方式為計算基礎，這些指標的原理是良好的分類系統所產生的結果應該和標準一致。局部準確性測量內的期刊同被指定為標準分類系統的某一個類別的比例。Jaccard相似性則是將受試分類系統與標準分類系統的類別都視為期刊的集合，測量每一對受試類別和標準類別的期刊交集（intersection）在它們的聯集（union）上的比例。

2. 以期刊之間的距離與相似性評估分類系統：將期刊視為空間上的點，將分類系統視為以點在空間上的密集程度來劃分成類別的機制，較密集的區域代表這些點的距離較小，也就是具有較大的相似性，應成為同一類別，空間上較稀疏的部分則視為類別的邊界。好的分類系統使得彼此之間距離較小的期刊同

在一個類別內，距離較大的期刊則應被指定為不同的類別。表1上，類別內期刊之間的平均距離與輪廓值便都是以這種方式為基礎的評估指標。在計算期刊之間的距離或相似性時可採用期刊的內容文本或文獻引用資訊為基礎，甚至兩者並用。

3. 以期刊之間的連結評估分類系統：將期刊視為網路上的節點，彼此間有關係的期刊以連結相連。分類系統可根據節點間連結線的疏密程度進行評估，良好的分類系統盡量使彼此間具有連結的節點分配在同一類別內。模組性便屬於這類指標。Wang與Waltman（2016）則是將兩種期刊間的文獻引用視為它們之間的連結，測量引用發生在同一類別內的比例。此外，也可根據期刊之間的距離（或相似性）來定義它們之間的連結，假定距離較小（相似性較大）的期刊之間具有連結相連。

4. 根據類別描述語的語意關係評估期刊分類系統：最為特別的是採用期刊的內容文本產生每一個類別的描述語，例如Janssens等（2009）根據詞語的TF-IDF值選擇描述語。

三、研究資料與研究方法

本研究從政府資料開放平台上取得103年的大學校院教師學術專長彙整表（http://data.gov.tw/node/27931）做為分析資料，共計有93,368筆教師資料。由於教育部所提供的大學校院教師學術專長資料，若為非系所或學程的學校單位則大多沒有提供學科標準分類系統的各階層學科資料，所以本研究僅使用具有學科資料的系所或學程教師資料。考量學術專長較符合系所主要的教學研究方向，只選用專任教師的資料，並且去除專長資料填寫「無」的教師。最後共使用43,460筆教師資料，來自3,233個公私立大學校院系所或學程。表2提供103年大學校院教師學術專長彙整表的各項教師統計資料，包含教師總數、有無提供學科資料的教師人數、專兼任教師人數、專長填寫「無」或非填寫「無」的教師人數等。

表2　103年大學校院教師學術專長彙整表各項教師統計資料

統計項目	人數
教師總數	93,368
無提供學科資料的教師人數	13,061
提供學科資料的教師人數	80,307
兼任教師人數	45,483
專任教師人數	47,885
專長填寫「無」的教師人數	44
專長非填寫「無」的教師人數	93,324

在103年大學校院教師學術專長彙整表的領域及學門分類情形以及包含的學類數、系所數與專任教師數，可參見表3。

表3 103年大學校院教師學術專長彙整表領域及學門分類資料

領域名稱	學門名稱	學類數	系所數	教師人數
教育領域	教育學門	10	150	1,558
人文及藝術領域	藝術學門	10	123	1,152
人文及藝術領域	人文學門	12	310	4,217
人文及藝術領域	設計學門	5	169	1,683
社會科學、商業及法律領域	社會及行為科學學門	10	167	1,860
社會科學、商業及法律領域	傳播學門	8	69	689
社會科學、商業及法律領域	商業及管理學門	10	467	5,926
社會科學、商業及法律領域	法律學門	3	47	573
科學領域	生命科學學門	7	100	1,174
科學領域	自然科學學門	7	81	1,337
科學領域	數學及統計學門	3	49	775
科學領域	電算機學門	5	146	1,987
工程、製造及營造領域	工程學門	14	560	9,323
工程、製造及營造領域	建築及都市規劃學門	4	46	525
農學領域	農業科學學門	11	72	917
農學領域	獸醫學門	1	10	129
醫藥衛生及社福領域	醫藥衛生學門	9	251	5,048
醫藥衛生及社福領域	社會服務學門	4	65	750
服務領域	民生學門	9	315	3,434
服務領域	運輸服務學門	3	23	264
服務領域	環境保護學門	2	7	90
服務領域	軍警國防安全學門	1	2	22
其他	其他	2	4	27

從表3的數據可發現，各學門的大小極不均勻：有些學門具有相當多的學類、系所與教師，例如「工程學門」、「商業及管理學門」、「醫藥衛生學門」、「人文學門」以及「民生學門」等，都有超過3,000位教師，200個以上的系所，但「獸醫學門」、「環境保護學門」以及「軍警國防安全學門」等較小學門都只有10個以內的系所，不到200位教師。另有四個系所的領域與學門資料為「其他」。

㈠各系所間與各學門間的相似性估計

本研究假定同一學門下各系所的學術專長應具有彼此相似的特性，也就是同一學門各系所的學術專長資料平均上應比不同學門間的系所更為相似。為了測量各系所之間的學術專長相似性，本研究將各系所專任教師的學術專長資料彙整成一組代表系所的文字資料，然後比對文字資料的相似性。

進行文字資料的相似性比對時，目前的自然語言處理方法大多基於「詞袋模型」（bag-of-word model; Turney & Pantel, 2010），比對文字資料上出現詞

語種類與次數的相似性。詞袋模型將每筆文字資料轉換為一組向量，向量上每個元素的數值代表一種詞語在文字資料上的權重（weight）。文字資料間的比對便轉變成為向量間的相似性比對。然而詞袋模型有幾項較嚴重的問題（Le & Mikolov, 2014）：1.詞袋模型假定詞語的出現都是獨立的，缺乏詞語在文本上下文的脈絡。2.詞袋模型沒考慮多義詞及同義詞等語言現象，每個詞語都對應向量上的一個特定元素，缺乏語法和語意訊息。3.由於文字資料上詞語的種類必然相當多，導致向量的維度（dimension）相當大，但向量上的數值相當稀疏（sparse）。此外，在本研究中，除了少數英文與數字資料，文字資料上中文詞語之間缺乏明顯的邊界，學術專長的文字資料內卻常包含許多當前重要研究主題的「未知詞」（unknown words），目前的自動斷詞（word segmentation）系統很難完全正確地切分出所有的未知詞。為避免斷詞結果錯誤，導致相近似的學術專長卻產生相當不同的向量，除了具有明顯邊界的少數英文與數字資料，本研究選擇以字（character）為文字資料的比對單位。然而以字為比對單位，若是使用缺乏上下文脈絡的詞袋模型，在進行文字資料比對時，勢必將損失更多文字資料中蘊含的訊息。

為解決或減輕上述的詞袋模型缺乏上下文脈絡、未考慮語法和語意訊息與向量維度過於龐大等問題，本研究採用Word2Vec做為文字資料的比對方式（Mikolov, Chen, Corrado, & Dean, 2013），使得學術專長相近似的系所或學門，在文字資料比對時能獲得較大的相似性。Word2Vec是一種「類神經網路」（artificial neural networks）的文本處理程序，能將輸入的文字資料上每個出現的詞語轉換成對應的特徵向量。不同於詞袋模型，Word2Vec本身便具有隱含上下文脈絡的關係，轉換出來的特徵向量在空間向量的計算上能保留每種詞語的語法或語意訊息，而且Word2Vec產生的特徵向量維度通常遠小於詞袋模型的特徵向量維度。本研究與一般Word2Vec應用不同的在於中文部分是以字為處理單位。

經過Word2Vec程序後，每一個出現在學術專長欄位上的中文字、數值以及英文詞都會轉換為一個特徵向量，本研究以Gensim提供的python函式庫進行Word2Vec的計算。

接下來，根據學術專長的文字資料，找出對應的特徵向量，組合產生代表每一系所的向量，並以字詞的出現次數做為系所向量上的加權，使得出現次數較多的字詞，其對應的特徵向量對合成的系所向量有較大的影響，然後以各學門所屬系所的向量進行平均，產生代表學門的向量。圖2說明了上述的概念。

在Word2Vec程序中，通常以餘弦測量（cosine measure）估計文字資料之間的相似性。因此，產生代表系所的向量之後，可利用餘弦測量估計系所之間在學術專長上的相似性，也可將代表學門學術專長的向量輸入餘弦測量，估計學門之間的相似性。

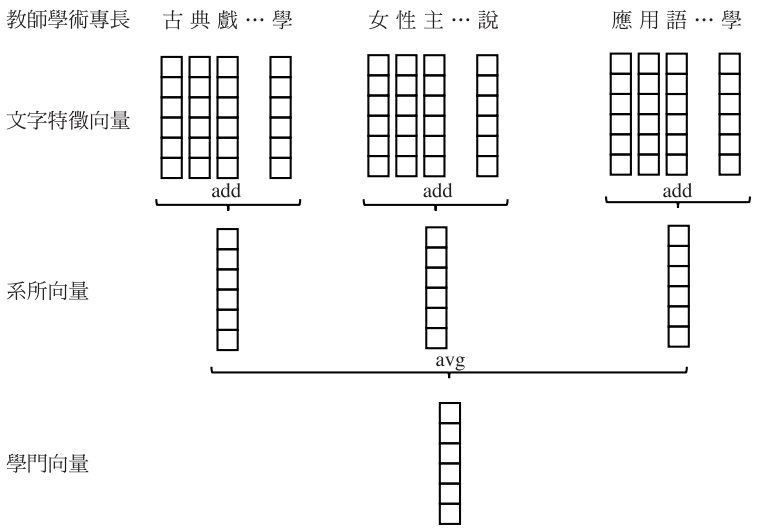

圖2 利用學術專長資料文字資料產生系所向量與學門向量概念示意圖

㈡分析的指標與圖表

本研究首先根據學門之間在學術專長資料上的相似性，分析學科標準分類系統上各學門之間的關連和集群，並觀察彼此接近的學門在分類系統上層的「領域」異同。在這裡，將利用凝聚性階層集群（agglomerative hierarchical clustering, AHC）演算法產生的樹形結構圖（dendrogram）和多維尺度法（multidimensional scaling, MDS; Kruskal & Wish, 1978）的散布圖（scatter plot），透過視覺化方式檢視學門之間的關連。計算凝聚性階層集群演算法與多維尺度法時，利用前一小節的說明，先訓練產生代表學門的向量，然後以餘弦測量計算學門之間相似性。

接下來計算各學門的平均輪廓值，以內部評估方式探討各學門內系所學術專長的一致性高低。以Janssens等（2009）測量ESI類別平均輪廓值的相同概念，本研究利用系所之間的學術專長相似性，計算各系所的輪廓值。系所i的輪廓值s_i的計算方式是將該系所與本身學門內其他系所之間的相似性平均值（a_i）減去該系所與任何其他學門上系所的相似性平均值（b_i）的值除以兩者中的最大值（$\max(a_i, b_i)$）。

$$s_i \stackrel{\text{def}}{=} \frac{a_i - b_i}{\max(a_i, b_i)}$$

如果系所i的輪廓值s_i大於0，表示該系所的學術專長與同一學門的系所相比，與其他學門更相近似；愈接近1，分配愈適當。

然後對每個學門以其所屬系所的輪廓值加以平均。若某一學門的平均輪廓值為較大的正數，則學門內各系所的學術專長較一致，彼此的學術專長相近

似。反之，若該學門的平均輪廓值小於或相當接近於0，則表示學門內可能有某些系所的學術專長與學門內其他系所較不相似，導致該學門較容易受到某些近似學門的混淆。

本研究進一步分析各學門內較不一致的系所，其學術專長資料與那些學門較接近。在先前計算某個系所的輪廓值時，可一併得知該系所的最相似學門，代表該系所與這個學門內的系所有最接近的學術專長。最相似學門可能是該系所本身所屬學門，也有可能是其他學術專長資料較相似而容易混淆的學門，後者即為本研究所定義的近似學門。本研究將分析每個學門的近似學門。為了方便解釋分析的結果，本研究將利用熱力圖（heatmap）呈現在每個學門上的近似學門分布情形。

最後，本研究計算每個學門內每一對系所之間的相似性，找出與學門內其他系所不一致的系所。一對系所之間的相似性小於某一個相當小的預設閾值（threshold），這對系所便可認定為不相似。若某些系所與其他系所大多不相似，可進一步了解這些系所的學術專長與學門內其他系所不同的原因，提供修正資料錯誤或重新指定這些系所的學門的參考，使得分類系統的結果較能符合學術發展的現況。按照統計學上繪製盒鬚圖（box-whisker plot）界外值（extreme）的概念，本研究將一對系所之間不相似的閾值定義為學門內所有相似性的第一四分位數（the first quartile）減去1.5倍的四分位距（interquartile range, IQR），相似性小於該閾值者都視為極端值。

四、研究結果

㈠各學門之間的關連

圖3是以學術專長資料的相似性，針對學科標準分類系統的各學門進行集群分群產生的樹形結構圖。

依據集群分群產生的樹形結構圖，可將學門分為三個群組：在圖3上，由左到右，第一群組從「自然科學學門」到「醫藥衛生學門」，第二群組從「建築及都市規畫學門」到「電算機學門」，第三群組從「藝術學門」到「運輸服務學門」。第二群組可再細分為兩個小群，第三群組則包括三個小群。圖3上可觀察到不少彼此相似性較高的學門，例如第一群組內的「獸醫學門」、「醫藥衛生學門」與「生命科學學門」，又如第三群組內的「社會及行為科學學門」、「教育學門」與「社會服務學門」等。事實上，這些彼此相似的學門在原本的學科標準分類上屬於不同領域。「獸醫學門」、「醫藥衛生學門」與「生命科學學門」分別屬於「農學領域」、「醫藥衛生及社福領域」與「科學領域」；「社會及行為科學學門」、「教育學門」與「社會服務學門」則分別屬於「社會科學、商業及法律領域」、「教育領域」與「醫藥衛生及社福領域」。雖然這些學門分別屬於不同領域，但屬於它

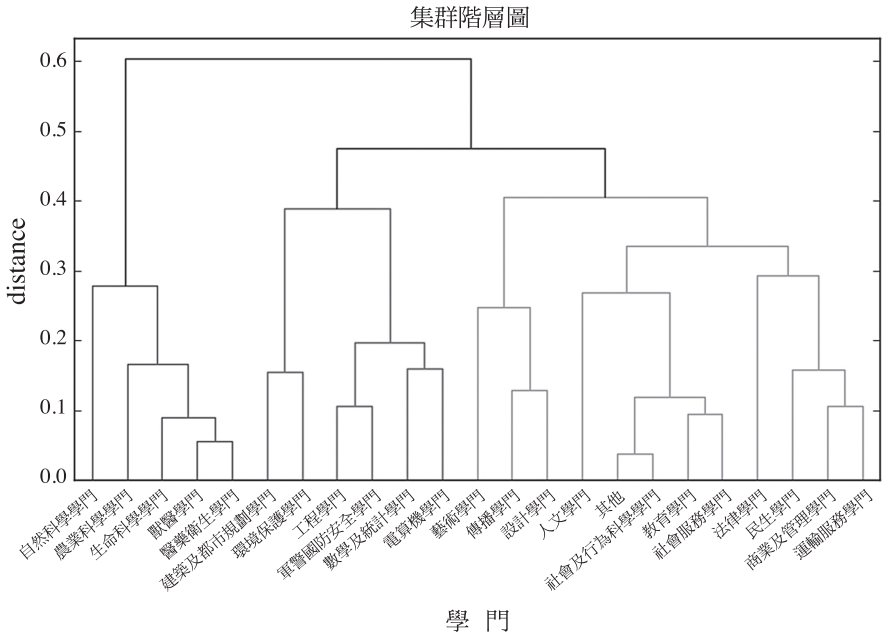

圖3 各學門進行集群分群產生的樹形結構圖

們的系所在學術專長上具有相當近似的文字資料，所以進行集群分析時，首先被聚集起來。

在學科標準分類系統內，有一些原先屬於同個領域的學門被歸屬在同個群組內，包括：「農學領域」的兩個學門都在第一群組內；「工程領域」的學門則是聚集在第二群組；「人文及藝術領域」與「社會科學、商業及法律領域」的各學門都在第三群組。但也有部分領域的學門分散在不同的群組：「科學領域」的「自然科學學門」與「生命科學學門」在第一群組內，「數學及統計學門」與「電算機學門」則在第二群組；「服務領域」的學門分布在第二和第三群組，前者包括「環境保護學門」與「軍警國防安全學門」，後者則有「民生學門」與「運輸服務學門」；「醫藥衛生及社福領域」的「醫藥衛生學門」與「社會服務學門」分別在第一群組和第三群組內。

圖4是將各學門在學術專長上的相似性輸入多維尺度法所產生的散佈圖，使得相似性較大的學門能映射在圖形上距離較相近的點。集群分析上同在一個群組的學門，在散佈圖上也映射在彼此鄰近的位置。並可明顯地看出第一群組的5個學門與其他兩個群組的學門之間有明顯的邊界，但第二群組和第三群組之間則有重疊。第二群組和第三群組之間的重疊，一部分來自於「建築及都市規劃學門」（第二群組）與「設計學門」（第三群組）之間較接近的距離。在「設計學門」裡有相當多位教師具有「建築設計」、「景觀設計」和「住宅規劃與政策」等等與「建築及都市規劃學門」極為有關的學術專長，因此這兩個學門的

學術專長很相似，而映射到圖形上較相鄰的位置。再者，第二群組的「電算機學門」包含的相關系所為資訊管理、網路與資訊技術、多媒體設計等，極大多數的教師學術專長中包含「視覺傳達設計」、「供應鏈」、「資料探勘」、「知識管理」、「電子商務」、「資訊系統」、「影音處理」等等文字，而這些文字也分別常見於第三群組的「設計學門」、「運輸服務學門」、「商業及管理學門」、「傳播學門」等各學門的教師學術專長。可見電腦與網路科技的應用已滲入多個學門，透露出學術知識的交流正使得學科之間界限愈來越模糊，造成第二和第三兩個群組之間邊界重疊。

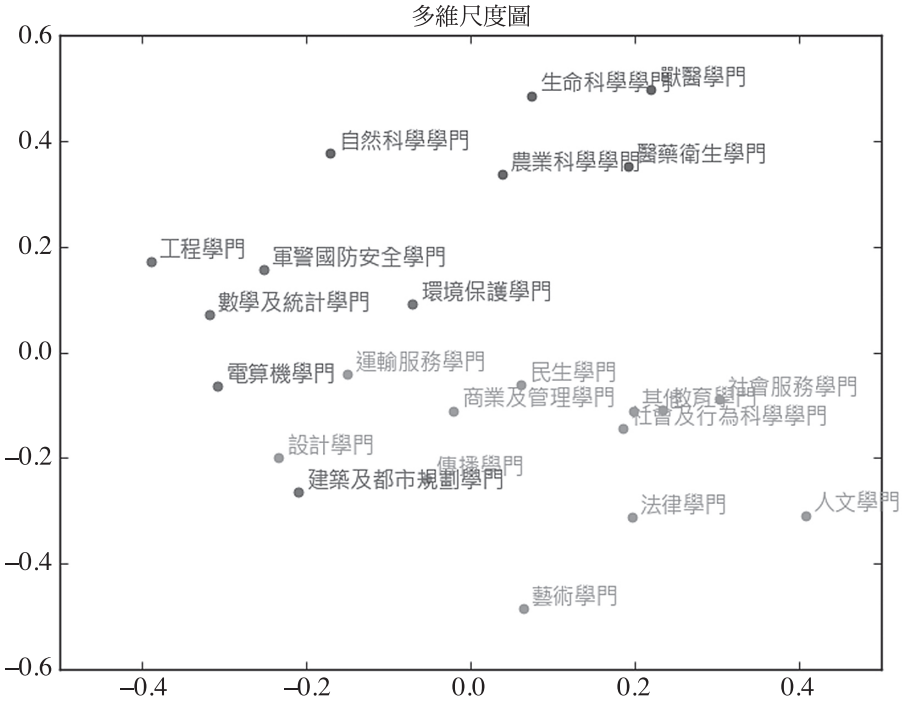

圖4　各學門進行多維尺度分析產生散佈圖

除了集群分析上相同群組的學門會在散佈圖上彼此映射到鄰近區域外，在學科標準分類上某些屬於同一領域的學門，在圖上也有較近的距離。例如同屬於「服務領域」的四個學門（環境保護學門、軍警國防安全學門、民生學門與運輸服務學門），雖分別被歸屬在第二群組和第三群組內，但還是可以發現這四個學門在散佈圖上的位置彼此相當接近。

㈡各學門的分類一致性與相互影響

為了解同一學門內各系所在學術專長上的一致性情形，接下來本研究測量各系所的輪廓值，然後計算各學門的平均輪廓測量。各學門間的平均輪廓測

量結果如圖5所示。圖5上依據各學門獲得的平均輪廓值大小由左至右排列，呈現各學門的學術專長一致性情形，左邊的長條表示平均輪廓值為正的學門，右邊則是輪廓測量得到負值的學門。從圖5可發現，平均輪廓值較大的學門有「法律學門」、「軍警國防安全學門」、「建築及都市規劃學門」、「數學及統計學門」、「電算機學門」、「人文學門」等；7個學門的輪廓值小於0，包括「其他」、「醫藥衛生學門」、「社會及行為科學學門」、「工程學門」、「農業科學學門」、「教育學門」、「民生學門」等。此外，「環境保護學門」、「藝術學門」、「自然科學學門」、「傳播學門」、「運輸服務學門」、「設計學門」的輪廓值雖為正數，但相當接近0，也都是系所的學術專長較不一致的學門。

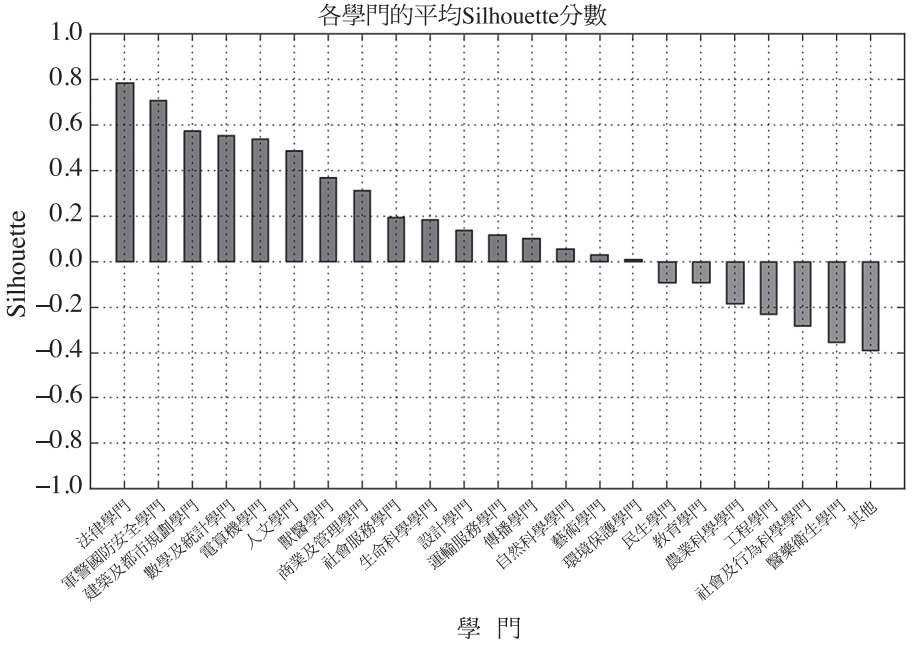

圖5 按平均輪廓分數排列呈現各學門學術專長一致性

接下來，運用圖6的熱力圖呈現各學門的近似學門分布。且為了能使熱力圖更加容易解讀，本研究利用凝聚性階層集群分析的結果，將學術專長較相似的學門對映在熱力圖上較接近的位置。

在熱力圖的每個橫列呈現一個學門，橫列上每個方塊顏色深淺表示近似學門的比例，較深顏色代表有較多系所的近似學門分布在直行上的學門。以第二列的「農業科學學門」為例，在熱力圖上可觀察到該學門較明顯的近似學門有「生命科學學門」和「環境保護學門」，其中並且在「農業科學學門」中，「生命科學學門」是相當多系所的近似學門。在圖6熱力圖上，顏色較深方塊大多集中在對角線附近區域，也就是學門的近似學門往往便是在集群分析中經常被歸為同

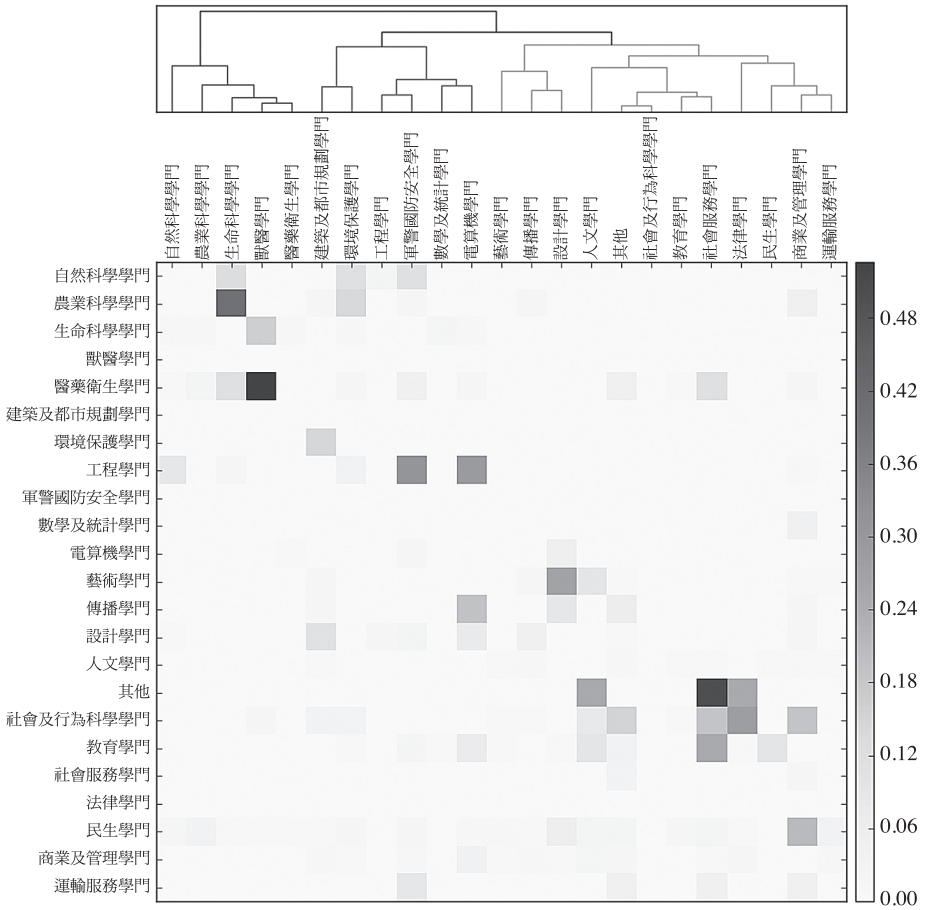

圖6　各學門的近似學門分布熱力圖

一群組的學門，也就是學術專長較相似的學門經常互為近似學門。下面謹就前述輪廓測量結果的幾個系所學術專長較不一致的學門加以說明：

「醫藥衛生學門」的近似學門主要是「獸醫學門」，也有少數分布落在「生命科學學門」與「社會服務學門」。從前面的集群分析結果可觀察到「醫藥衛生學門」、「獸醫學門」與「生命科學學門」等共同構成第一群組，並在多維尺度分析產生的散佈圖上彼此相鄰。再仔細觀察學科標準分類系統上各系所被指定的學類，近似學門落於「獸醫學門」的「醫藥衛生學門」系所大多為「醫學學類」、「牙醫學類」與「醫學技術及檢驗學類」，近似學門為「生命科學學門」的系所則大多為「藥學學類」。另外，「社會服務學門」與「醫藥衛生學門」在學科標準分類中同屬於「醫藥衛生及社福領域」，但在集群分析結果上，這兩個學門分別屬於第一群組和第三群組，在「醫藥衛生學門」中近似學門為「社會服務學門」的系所則大多為「護理學類」與「復健醫學學類」。UNESCO（2013）的國際教育標準分類（ISCED）中對於學科標準分類的文件 *ISCED-F* 2013 特別針對「醫藥衛生學

門」不是和學術專長資料較接近的「獸醫學門」歸類於同一領域，而是和「社會服務學門」共同組成為「醫藥衛生與社福領域」的原因提出說明：「醫藥衛生學門」與「社會服務學門」都是以人類的生活照護與福祉為學習目的，具有關注相同的對象，所以共同歸屬於「醫藥衛生與社福領域」，便於進行統計及國家政策分析；反之，「醫藥衛生學門」與「獸醫學門」雖然採用相似理論知識，但因關注對象與學習目的不同，在此一分類系統中則被歸入兩個不同的領域。

「社會及行為科學學門」具有較多的近似學門，這些學門在學科標準分類系統上分屬不同領域，但它們在集群分析結果上，都被歸入第三群組：近似學門為「法律學門」的系所大多屬於「政治學類」、「公共行政學類」與「國際事務學類」；近似學門落在「社會服務學門」的系所以「心理學類」為主；落在「商業及管理學門」的系所則大多為「經濟學類」；落在「人文學門」的系所大多為「區域研究學類」；最後部分系所的近似學門為「其他」，這些系所許多是「社會學類」。

此外，「工程學門」近似學門較明顯的有「軍警國防安全學門」和「電算機學門」：近似學門為「軍警國防安全學門」的系所包括「機械工程學類」與「土木工程學類」，「電算機學門」的系所則大多是「電資工程學類」。

其他學術專長較不一致的學門，「農業科學學門」的主要近似學門分別為「生命科學學門」，「社會服務學門」和「商業及管理學門」的主要近似學門則分別是「教育學門」和「民生學門」。上述學術專長較不一致的學門與其主要近似學門在集群分析中也都被指定為同一群組。

特別要說明的是，學門欄位填寫「其他」的四個系所：國立成功大學老年學研究所和樹德科技大學人類性學研究所的近似學門是「社會服務學門」，國立臺灣科技大學人文社會學科則為「人文學門」，另外，高雄市立空中大學有四位教師登錄系所資料為「市立空大」，在本研究中也將他們的資料視為一個系所單位，其輪廓測試結果，該單位的近似學門是「法律學門」。若是這些系所考慮修改其學科標準分類和教師學術專長彙整表上的資料，上述結果可做為參考。

㈢各學門中與大多系所不一致的系所

接下來，根據相似性的統計，確認學門內與大多系所不一致的系所。相似性小於閾值數量最多的是「商業及管理學門」，這個學門裡共有467個系所，但有8,242對系所彼此之間的相似性小於閾值，佔所有可能相似性的比例達7.57%（8242/(467*466/2)）。另外，「運輸服務學門」、「生命科學學門」和「數學及統計學門」相似性小於閾值的情形，比例也很高。

歸納相似性經常小於閾值的系所與學門內大多數系所不一致的原因如下：

1.系所的學術專長相當專殊。例如「數學及統計學門」的精算數學相關系所、「傳播學門」的博物館學相關系所、「法律學門」的專利相關系所。這些系所的學術專長偏向於某些特定主題，因此與同一學門的其他系所不同。

2.只有相當少專任教師人數的系所。在103年度的教師學術專長彙整表上，有48個系所的專任教師只有1人，這些系所大都為學程。人數較少的系所容易偏向少數主題的學術專長，與該系所所屬學門具有多個主題的其他系所有較大差異。

3.某些系所的資料填寫認知不同於大多數系所。例如：某些系所提供的學術專長資料是以英文撰寫，與目前大多數系所提供的中文資料不同。也有些系所提供的學術專長資料並非現象、方法與技術等主題特定的知識，而是如「人文社會」、「自然科學」、「管理類」、「民生學群」等範圍較廣泛的專長，因此也和其他系所有很大差異。

4.科際整合產生的新創系所。因應新的社會現象與問題產生，大學校院設立了許多跨領域的系所，使多個來自不同領域的學門合作進行研究與教學。這些新創系所的師資，往往來自傳統上極不相關的領域，然而在目前學科標準分類與教師學術專長彙整表上，受限於資料的組織結構，卻只能將這些系所歸屬於其中一個學門。例如文化創意產業相關系所同時需要多位具有原本屬於「商業及管理學門」、「人文學門」、「藝術學門」、「傳播學門」、「設計學門」等學術專長的教師，這些系所卻只能被指定為一個學門，而彙整得到的學術專長也與任何單一學門的系所有差異。又如「醫管學類」的醫務與健康產業管理相關系所被指定為「商業及管理學門」，這些系所除了具有「商業及管理學門」方面學術專長的教師，還需要有「醫藥衛生學門」與「環境保護學門」等學術專長的教師，才能因應醫務與健康產業管理的特殊需求，但造成這些系所彙整出來的學術專長與其他「商業及管理學門」的系所不同。

五、結　論

本研究以政府開放資料的教師學術專長彙整表做為分析資料，利用Word-2Vec文字資料比對技術，測量系所之間與學門之間在學術專長資料上的相似性，檢測分析目前的系所分類系統。在探討分類系統的分類成效時，首先以集群分析和多維尺度分析探討學門之間的集群與關連；接著針對每一個學門，以輪廓指標測量同在一個學門內的各系所學術專長的一致性，探討學科標準分類系統的分類成效，並分析各學門容易混淆的近似學門；最後，從系所之間的相似性找出各學門較不一致的系所，探討其相似性較小的原因。本研究並以視覺化圖表方式呈現上述的成果。

本研究的結果如下：

1.學門之間集群分析的結果與學科標準分類系統所指定的領域有一些差異，但從多元尺度分析的結果仍可發現多數領域的所屬學門彼此之間仍有關連。在樹形結構圖上，可明顯看出所有的學門分為三個群組，有些領域所屬的

學門集中在同一個群組上，但也有些領域的學門分散在不同群組。屬於同一群組的學門在散佈圖上會聚集在一起。第一群組的學門與其他兩個群組間的分隔較明顯，但第二和第三群組的學門之間有重疊。而某些在同一領域但集群分析歸屬於不同群組的學門，在散佈圖上也會映射到相鄰近的位置。

2.大抵來說，許多學門在輪廓測試的檢測結果並不好；換言之，這些學門內有許多系所的學術專長與本身學門的系所不一致，反而容易混淆為其他學門。根據熱力圖所呈現的近似學門分布情形，容易混淆的近似學門大多與本身學門在集群分析中屬於同一個群組。

3.各學門內與其他系所不一致的系所，除了系所本身具有較專殊的學術專長以外，其不一致的原因還包括：師資人數過少，使得彙整後的系所學術專長較傾向某些主題；系所提供的學術專長資料與多數系所的認知有差距；因應科際整合而新創的系所，其學術專長兼具多個不同領域與學門的主題。

針對以上結果，本研究對系所分類系統、開放資料品質與進一步的研究課題進行以下的討論與建議：

為了進行教育統計的需求，學科標準分類系統採用系所傳授與研究的知識主題做為組織架構（organization scheme），並按系所名稱決定該系所的主題。然而學科知識主題廣泛、多元複雜（heterogonous）且容易混淆（ambiguous），對同一系所的主題，若分類者的觀點（perspectives）不同，可想而知必然會產生不同的分類結果。相當多研究指出，以適當的資料驅動演算法決定主題能產生比主觀判斷更為一致的結果。本研究認為主題相近的系所通常有相似的學術專長資料，因此建議利用學術專長資料的相似性做為系所分類系統的評估標準，並利用文字比對進行相似性測量，研究結果證實這種做法相當可行。未來可進一步探討根據學術專長資料做為判斷學科分類系統的組織架構，以及提出更準確的比對演算法。

除此之外，目前學科標準分類系統採用由上而下（top-down）的階層式組織結構（organization structure）。這種分類結構的類別之間彼此不重疊，但對於新創的跨領域系所很難指定適當且唯一的學科分類。UNESCO（2013）在*ISCED-F* 2013 提出的做法是先決定系所的所屬領域，再將其指定為領域中專門收納跨領域的學門與學科。教育部（2007）的學科標準分類則大多指定無法歸類的新興系所為學門中編號尾數為99的其他學類，例如許多文化創意產業相關科系被指定為「其他商業及管理學類」（編號3499）、「其他人文學類」（編號2299）或「其他設計學類」（編號2399）。然而這些分類方式無法明確表示該系所所有涉及的領域與學門，很難達到原本*ISCED-F* 2013與教育部所想要達到的教育統計的目的，同時也不利於根據學科分類進行系所間的交流與整合。若可採用多標籤分類（multi-label classification）的組織結構，在必要情況下，將系所指定到多個學

科，將具有較大的彈性，能充分表達多元複雜的學科知識主題。但是多標籤分類的組織結構較複雜，容易造成使用者較大的認知負荷，且不易讓使用者對整個學科分類系統形成認知結構。因此，若要將多標籤分類應用在學科分類系統上，也需仔細考量其中的利弊得失，審慎判斷其適用性。

另外，從本研究的研究過程與結果分析，可了解資料正確對於開放資料運用成效的重要。目前教師學術專長彙整表上的學術專長文字資料是大多由教師各自填寫後彙整而成。由於未經後續的資料品質控管，資料的格式與內容都沒有標準化。從填寫資料的編碼（如：中文或英文、字型的全形或半形、專長項目之間的分隔），到學術專長的定義、範圍等等，資料提供者往往有不同的認知，造成資料應用上的困難。教育部既然已經開放這些資料，本應責無旁貸提升資料品質，以便於民眾加值應用。教育部應制定相關資料提供規範，提供建議範例，讓資料提供者在填寫資料前便了解正確的資料格式與內容，確保資料品質。彙整各院校各系所的資料之後，也應就資料格式與內容進行檢核，發現可能的錯誤，提報相關訊息給資料的管理者與提供者。在此，可發展能協助資料提供或偵測錯誤的文本處理技術，以減輕資料管理者與提供者的負擔，提升資料品質。

最後，本研究利用Word2Vec文字資料比對技術，並以字為比對單位，測量系所之間與學門之間在學術專長資料上的相似性。如此一來，固然解決了新詞或斷詞系統不完善引起的缺失，但也產生了一些過度匹配的問題，例如「資」是教師學術專長中相當常見的用字，出現於「資料」、「資訊」、「資源」、「師資」、「投資」、「資本」等詞上，這些詞則使用在不同學科的學術專長，例如「師資」最常被使用在「教育學門」，「投資」和「資本」則會常見於在「商業及管理學門」的教師學術專長，以字為比對單位較容易使原本學術專長不相關的系所或學門產生關連。因此，未來可嘗試利用斷詞系統所產生的詞串為輸入Word2Vec，比較字與詞兩種文字資料比對單位，對於中文的文字資料比對技術進行深入的研究。更進一步來說，Word2Vec文字資料比對方式在學科分類上的成效也是另一個需要探討的問題：Word2Vec考慮訓練文本的上下文脈絡，將每一個字或詞對應到一組特徵向量。適當的上下文脈絡範圍使特徵向量之間的相似性能精確地描述字或詞的語法及語意關係，未來可針對上下文脈絡的範圍進行比較分析，而傳統詞袋模型，如TF-IDF的比對方式，與Word2Vec的比較也需要深入研究。另一個可探討的問題是本研究基於專任教師是系所研究與教學的主力，匯集專任教師的學術專長便可表現系所研究與教學的知識主題，因此比對系所單位時並沒有使用兼任教師的學術專長資料。然而從研究結果發現，各大學開設的學程由於其跨學門與跨領域的特性，師資大多來自多個領域或學門的兼任教師，專任教師的人數並不多，本研究的假設情形使得這些單位的知

識主題容易偏向少數專任教師的學術專長。未來的研究或可在系所單位加入兼任教師的學術專長，以解決學程的專任教師人數較少的問題，揭露學程跨學門與跨領域的特性。

參考文獻

教育部（2007）。中華民國教育程度及學科標準分類（第4次修正）。檢索自http://stats.moe.gov.tw/files/bcode/96bcode_book.pdf

教育部統計處（2016）。教育程度及學科標準分類。檢索自http://depart.moe.edu.tw/ED4500/cp.aspx?n=77627C0EE4283293&s=85E1E406503C665B#

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351-374. doi:10.1007/s11192-005-0255-6

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, *61*(7), 1386-1409. doi:10.1002/asi.21309

Chen, C.-M. (2008). Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. *Journal of the American Society for Information Science and Technology*, *59*(14), 2296-2304. doi:10.1002/asi.20935

Chen, P., & Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics*, *4*(3), 278-290. doi:10.1016/j.joi.2010.01.001

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357-367. doi:10.1023/A:1022378804087

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, *45*(6), 683-702. doi:10.1016/j.ipm.2009.06.003

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, *57*(2), 251-263. doi:10.1002/asi.20274

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning* (pp. 1188-1196). Retrieved from http://www.jmlr.org/proceedings/papers/v32/le14.html

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362. doi:10.1002/asi.20967

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (arXiv:1301.3781v3). Retrieved from https://arxiv.org/pdf/1301.3781v3.pdf

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. doi:10.1103/PhysRevE.69.026113

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113-1119. doi:10.1002/asi.10153

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. doi: 10.1016/0377-0427(87)90125-7

Thijs, B., Zhang, L., & Glänzel, W. (2015). Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes. *Scientometrics*, *105*(3), 1453-1467. doi:10.1007/s11192-015-1641-3

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141-188. doi:10.1613/jair.2934

UNESCO. (2013). *ISCED Fields of Education and Training 2013 (ISCED-F 2013)*. doi:10.15220/978-92-9189-150-4-en

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347-364. doi:10.1016/j.joi.2016.02.003

Journal of Educational Media & Library Sciences 54 : 1 (2017) : 69-95
DOI:10.6120/JoEMLS.2017.541/0046.RS.AM

Research Article

Analyses of the Standard Classification of Fields Based on the Directory of Faculty Expertise Open Data

Sung-Chien Lin

Abstract

This paper presents a series of analyses of the Standard Classification of Fields which was applied to the classification of all departments in universities based on measuring similarity between text data of the faculty expertise directory from open data provided by the Ministry of Education of Taiwan, and suggests some possible directions for improvement of the directory and the classification system. The analysis techniques included the Word2Vec text matching technique to estimate the similarity of faculty expertise, the methods to expose properties of the classification system such as hierarchical clustering analysis, multidimensional scaling analysis, silhouette testing, distribution of fields with similar expertise set, and statistics of the similarity between departments, and a variety of information visualizations to illustrate the analysis results. The results of this study show that in order to meet requirements from educational statistics, policy making, and academic exchanges, the organization structure, organization scheme, and data quality of the Standard Classification of Fields should be improved.

Keywords: *The Standard Classification of Field, Open data, The Directory of Faculty Expertise, Silhouette test, Word2Vec*

SUMMARY

The Ministry of Education of Taiwan collects academic expertise of all faculty members in universities in Taiwan and publishes a Directory of Faculty Expertise every year for higher education system statistics. In the directories, each department in the universities is assigned to a broad field (BF), a narrow field (NF), and a detailed field (DF) according to a hierarchical classification system which is a modified version of the International Standard Classification of Education (ISCED 1997) developed by the UNESCO. The field assignment of a department mainly depends on the title and the subjects taught by the department (Ministry of Education, 2007). The aim of this study is to suggest some possible directions for improvement of the directory and the classification system. Thus we performed a series of analyses on the classification system based on the

Assistant Professor, Department of Information and Communications, Shih Hsin University, Taipei, Taiwan
E-mail: scl@cc.shu.edu.tw

assumption that the expertise data of a department can represent its subjects and the text of expertise data of the departments in the same field should be more similar to each other than those of departments in the other fields.

(1) We analyzed the relationships and clusters among all of the NFs in the classification system and then compared the results with the BFs which the NFs originally belonged to.

(2) We evaluated the coherence for all NFs based on the similarities between expertise data of the departments and visually presented the confusing NFs.

(3) We identified the departments that had expertise data that were less similar to those of many other departments in the same NF and explored the reasons why they were incoherent to others.

The faculty expertise directory used in this study was the version made available in 2014 (http://data.gov.tw/node/27931). The total number of the records in the directory was 93,368. However, those records from part-time faculty members, lacking field information, or displaying "None" as their expertise data were excluded. As a result, we analyzed 43,460 records from 3,233 departments, which included 8 BFs, 22 NFs, and 148 DFs in addition to the "Others" field.

Before analyzing the Directory of Faculty Expertise classification system, we estimated the similarity between the expertise data using a technique of artificial neural networks, Word2Vec. It was proposed by Mikolov, Chen, Corrado, & Dean (2013). The technique of Word2Vec can transfer all terms occurred in text of the input data into feature vectors which preserved the syntactic and semantic characteristics of the terms. Since most of the expertise data used in this study were provided in Chinese, we used Chinese characters instead of Chinese words as the Word2Vec input unit due to the difficulty of Chinese word segmentation, particularly when it is applied to the expertise data full of technical terms. The vectors for representing departments were generated by summed up all the corresponding feature vectors of all Chinese characters, English terms, and numerical values appearing in the expertise data of the departments. Moreover, a vector for a NF was generated by averaging the vectors of the departments assigned to it. The similarity of expertise data between two departments or two NFs were then measured with the cosine value between the two corresponding vectors.

We used the techniques of cluster analysis and multi-dimensional scaling to reveal the relationships and clusters among NFs. The results of cluster analysis showed that all NFs were divided into 3 groups. There were some BFs with NFs that were all clustered in the same groups. For example, the two NFs of the BF "Agriculture" both located in the first group, all of the NFs in the BF "Engineering" located in the second group, and the NFs of the BFs "Humanities and Arts" and

"Social Sciences, Business and Law" were all in the third group. However, the NFs of the remainder BFs were split into different groups. The NFs in the same groups as the results of the cluster analysis were usually mapped into the positions close to each other based the results of multi-dimensional scaling. Moreover, there was an obvious separation between the first group of NFs and the other two groups, but the second and the third group were overlapped with each other. The overlay were caused by that the expertise data of the NFs "Architecture and Urban Planning" and "Design", respectively in the second and third groups, were very similar. In addition, the expertise of computers had become common for a few NFs in the third group, which originally was unique to the NF "Computer" in the second group.

The coherence of NFs was evaluated using the silhouette test (C. Chen, Ibekwe-SanJuan, & Hou, 2010; Rousseeuw, 1987) based on the similarities among the expertise data of departments. Firstly, the silhouette score of each department was estimated. Silhouette scores are between 1 and –1. If a department had a higher positive silhouette score, it indicated that the department was more consistent with its assigned NF. Then, the degree of coherence for an NF was obtained by averaging the silhouette scores of all departments assigned to it (Janssens, Zhang, De Moor, & Glänzel, 2009). In this study, the results of coherence evaluation were less than satisfactory. The averaging silhouette scores of many NFs were negative values. In other words, in these NFs there were a lot of departments of which the expertise data were more similar to those of departments in the other NFs. Therefore, these NFs were easy to be confused with other NFs. We generated a heatmap to observe the confusion among NFs. The heatmap showed that the confusing NFs of an NF and the NF itself were usually observed in the same group based the results of cluster analysis. For example, the NFs "Veterinary" and "Life Sciences" were both of the confusing NFs of "Medicine" and these three NFs were all members in the first group.

Finally, we selected departments with the highest numbers of similarities that were lower than a pre-setting threshold and considered that they were inconsistent to their own NFs. In this study, the NF that had the highest number of similarities lower than the threshold was "Business and Administration", followed by "Transport Services", "Life Science", and "Mathematics and Statistics". In addition to having more unique expertise than other departments in the same NFs, the reasons for the departments being inconsistent are that the numbers of faculty members in those departments were too small, the format and the language used in the expertise data of the departments were very different, or the departments were set up for interdisciplinary learning and thus their expertise were related to many NFs.

In view of the results above, we present the following discussions and suggestions on the classification system of fields, the quality of open data, and further research.

First, the classification system of the current directory utilizes the academic subjects of the departments as the directory organization scheme. Thus the determination of the subjects of a department is based on the title. The results of this study has confirmed the feasibility of using the similarity of expertise data to analyze the classification system. It is desired to improve the precision of similarity estimation between expertise data and then to apply the estimation method to the organization scheme of the next classification system.

In addition, the current classification system also utilizes a top-down, hierarchical organizational structure and therefore the fields do not overlap with each other. But it is difficult to assign an appropriate and unique field to the departments with interdisciplinary background. A multi-label classification system is possible to fully express the diversity of subjects in those departments because it has a complex and flexible organizational structure. However, its complexity and flexibility also makes a larger cognitive load for users and it can be difficult for them to form a cognitive structure about the entire system. Therefore, it is necessary to carefully consider the feasibility of applying the multi-label classification system to the directory of faculty expertise.

According to lessons learned from the analysis process and the results, the quality of open data is worth more attention. The Ministry of Education is recommended to develop and provide guidelines and examples for data providers and check the format and contents of data to reduce possible errors. Moreover, it is possible to develop text processing techniques that can help prepare data or detect errors to decrease the burdens on data administrators and providers and improve the data quality.

The possible extensions of this study include further performance analysis of applying the Word2Vec technique to text similarity estimation and the study of multi-label classification focusing on the problems caused by the departments with interdisciplinary background. Using words or characters as input units or different length of context both affect the Word2Vec performance and it is worth further analysis to obtain better results. Comparison of the traditional TF-IDF approach and the Word2Vec is also needed. Developing and evaluating a multi-label classification system are still very challenging.

ROMANIZED & TRANSLATED REFERENCE FOR ORIGINAL TEXT

教育部（2007）。中華民國教育程度及學科標準分類（第4次修正）。檢索自http://stats.moe.gov.tw/files/bcode/96bcode_book.pdf【Ministry of Education. (2007). *Zhonghuaminguo*

jiaoyu chengdu ji xueke biaozhun fenlei (di si ci xiuzheng). Retrieved from http://stats.moe.gov.tw/files/bcode/96bcode_book.pdf (in Chinese)】

教育部統計處（2016）。教育程度及學科標準分類。檢索自 http://depart.moe.edu.tw/ED4500/cp.aspx?n=77627C0EE4283293&s=85E1E406503C665B#【Department of Statistics, Ministry of Education. (2016). Jiaoyu chengdu ji xueke biaozhun fenlei. Retrieved from http://depart.moe.edu.tw/ED4500/cp.aspx?n=77627C0EE4283293&s=85E1E406503C665B# (in Chinese)】

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351-374. doi:10.1007/s11192-005-0255-6

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, *61*(7), 1386-1409. doi:10.1002/asi.21309

Chen, C.-M. (2008). Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. *Journal of the American Society for Information Science and Technology*, *59*(14), 2296-2304. doi:10.1002/asi.20935

Chen, P., & Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics*, *4*(3), 278-290. doi:10.1016/j.joi.2010.01.001

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357-367. doi:10.1023/A:1022378804087

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, *45*(6), 683-702. doi:10.1016/j.ipm.2009.06.003

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, *57*(2), 251-263. doi:10.1002/asi.20274

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning* (pp. 1188-1196). Retrieved from http://www.jmlr.org/proceedings/papers/v32/le14.html

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362. doi:10.1002/asi.20967

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (arXiv:1301.3781v3). Retrieved from https://arxiv.org/pdf/1301.3781v3.pdf

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. doi:10.1103/PhysRevE.69.026113

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113-1119. doi:10.1002/asi.10153

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. doi: 10.1016/0377-0427(87)90125-7

Thijs, B., Zhang, L., & Glänzel, W. (2015). Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes. *Scientometrics*, *105*(3), 1453-1467. doi:10.1007/s11192-015-1641-3

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141-188. doi:10.1613/jair.2934

UNESCO. (2013). *ISCED Fields of Education and Training 2013 (ISCED-F 2013)*. doi:10.15220/978-92-9189-150-4-en

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347-364. doi:10.1016/j.joi.2016.02.003