檔案之自動化主題分類
——以《總裁批簽》為例

報告人：

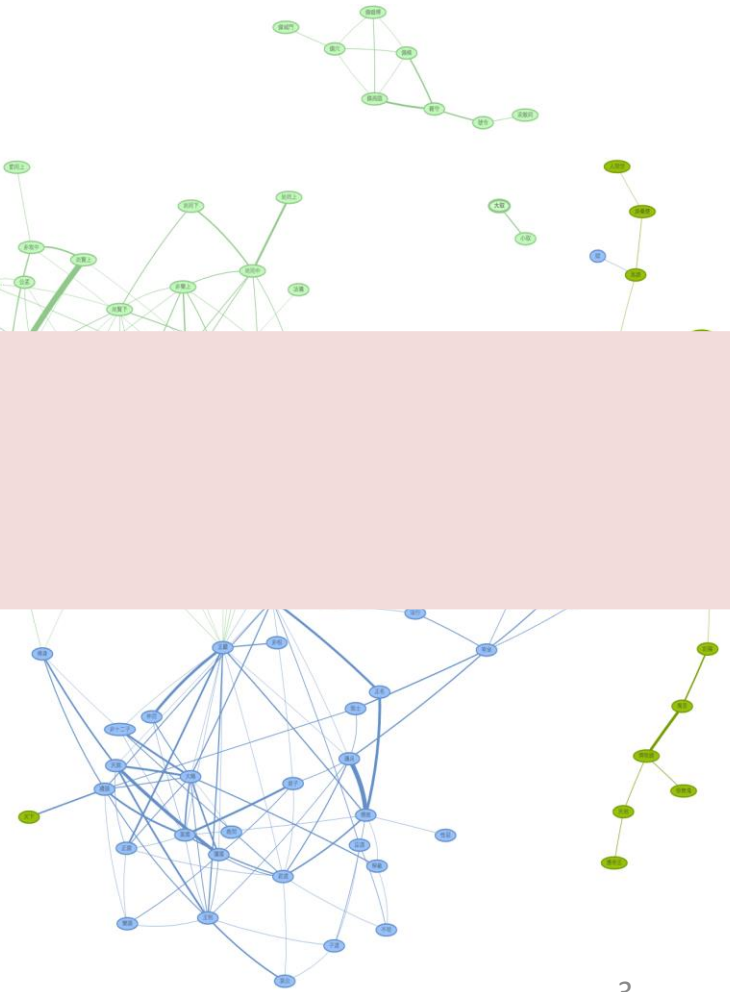
國立政治大學圖書資訊與檔案學研究所碩士生 吳承恩

國立政治大學圖書資訊與檔案學研究所教授兼所長、
圖書館副館長、圖書資訊學數位碩士在職專班執行長　林巧敏

報告大綱

- 前言
- 文獻回顧
- 研究設計與實施
- 分類結果與討論
- 結論與前瞻

前言



研究動機與目的

- 文件分類係指是依文件的內容主旨給予適當的類別標籤，進而進行分門別類的過程（郭俊桔，2018）。
- 以檔案而言，主題分類的目的不僅在於通過系統性地排列與組織以達有效管理；更藉由相同概念檔案的聚合，讓使用者透過關鍵字檢索，便可找到其所需要的資料，發揮檔案價值。
- 現階段檔案的主題分類多由檔案人員以檔案內容特性加以分類並賦予主題詞彙，或藉由參考資訊檢索技術所建立之詞庫的方式進行（林巧敏，2012）。
- 透過結合數位工具與自動分類技術的運用，可在短時間內處理大量的檔案，降低人力負荷，提升檔案整編與開放應用效益。
- 本研究便嘗試以《總裁批簽》檔案為例，從人文研究者的角度探討檔案運用自動化主題分類工具之可行性。

文獻回顧

① 自動化主題分類之技術發展
② 文本探勘工具與自動分類實務應用
③ 運用檔案材料之自動主題分類實務

自動化主題分類之技術發展（1/4）

- 自動分類及相關概念的形成，與自然語言處理（Natural Language Processing，NLP）技術有著密不可分的聯繫。1960年代開始，便有關於運用工具於自動分類技術的討論。
- 1992年TREC（text REtrieval conference）資訊評比會議所建立之提供共用測試集、相同評估準則與程序的模式，降低了進行相關研究的門檻。
- 現今根據不同的文件類型，相關的研究已包括主題檢索、資訊過濾、跨語檢索、全文影像辨識、語音與視訊、新事件偵測等眾多面向的應用（江玉婷、陳光華，1999；曾元顯，2014）。

自動化主題分類之技術發展（2/4）

- 自動分類研究可分為：

1. 基於規則式（Rule Based Methods）：使用引文內容建立各種模板進行比對後分類；
2. 基於統計式（Strtistics Based Methods）：使用統計排序與權重的方式進行分類；
3. 基於擷取式（Retrieval Based Methods）：強調使用人工智慧的理論執行分類。

- 自動分類研究流程：

1. 文件表徵之擷取；
2. 建立文件表徵；
3. 分類模型建立；
4. 自動分類評估的程序進行（郭俊桔，2018）。

自動化主題分類之技術發展（3/4）

- 層級群集方法：

1. 連鎖法：單一連鎖法（single linkage）、完全連鎖法（complete linkage）及平均連鎖法（average linkage）；
2. 最小變異數法（minimum variance method）

- 非層級群集方法：

1. K平均值演算法（K-means Method）
2. Cliques群集演算法（湯秋蓉，2009）

- 除上述舉例外，為因應解決語意相近但語詞不同的情況、短文的情況，同時，近年深度學習（deep learning）技術上的發展所產出之相關計算方式，亦值得關注。（李龍豪等，2016；郭俊桔，2018 ）。

自動化主題分類之技術發展（4/4）

國內圖資領域在自動分類方法上的探討：

1. 透過支援向量機（Support Vector Machine）結合詮釋資料建立分類模型並佐以統計方法修正的圖書自動分類（陳信源等，2009）；
2. 透過運用關鍵詞彙及參考書目比對進行分類之共現字（Co-word Analysis）及書目對（Bibliographic Coupling）方法（曾元顯，2011）
3. 因每一自動分類方法皆具有侷限性，而有統合多自動分類並使用多數決策略擬定圖書自動分類的方式（郭俊桔，2018）。

- 以現行相關研究的趨勢，仍顯示並沒有一種方法可有效處理各式文件以及應用；
- 雖然通過與過去機器學習方法的比較發現，以自動分類的文本分類而言，深度學習方法所得到的成果並不較佳（曾元顯，2014；2020）。

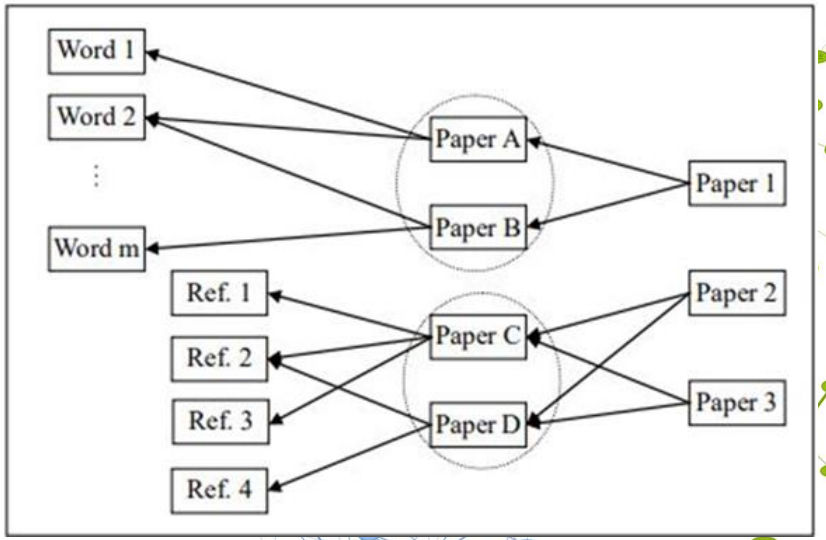
文本探勘工具與自動分類實務應用（1/2）

- 儘管牽涉自動分類領域的演算法相當豐富，但由於其大多僅為骨幹部分，在實際運用時大多需經過相當的增補以及修改才得以順利運行。
- 所幸，仍有部分工具的開發改善了此一情況，使更多研究者透過「用」的角度參與自動分類的研究領域之中，此些工具大多並非專為自動分類所設計，而較側重於文本內容探勘、文獻計量研究的形式，自動分類僅是其中的功能之一。

文本探勘工具與自動分類實務應用（2/2）

CATAR（Content Analysis Toolkit for Academic Research）

針對具有學術價值之文獻使用，並且使用者不再侷限於作者本人或技術人員。其功能分為兩種：概觀分析（overview analysis）及分解分析（breakdown analysis），其中分解分析便透過書目對及共現字的概念，通過對文件內容所擷取出詞彙或者引用文獻的計算，評估任意兩文件的相似度，從而進行文件歸類（cluster）（曾元顯，2011）。



$Sim(X, Y) = 2x|S(X) \cap S(Y)| / (|S(X)|+|S(Y)|)$

曾元顯、林瑜一（2011）

作者	篇名	研究文件	研究目的
湯秋蓉（2009）	自動化主題分析於圖書資訊領域之應用	臺灣地區圖書資訊學領域的學位論文與期刊文獻	研究主題
Chang, Y.H., Chang, C.Y., & Tseng, Y.H. (2010)	Trends of Science Education Research: An Automatic Content Analysis	《International Journal of Science Education》、《Journal of Research in Science Teaching》、《Research in Science Education》和《Science Education》四種期刊1990~2007年之文獻	研究主題、發展趨勢及貢獻者
陳淑貞（2010）	以自動化主題分析探索免疫學領域研究主題之發展	ESI資料庫免疫學領域1998~2008年之常被引文章	熱門主題
曾元顯、林綸一（2011）	內容探勘技術在教育評鑑研究發展趨勢分析之應用	Web of Science文獻	熱門主題、重要學者
曾元顯（2011）	文獻內容探勘工具—CATAR之發展與應用	Wok資料庫和科學教育與數位學習有關之11種期刊	主題趨勢、重要學者
汪耀華（2011）	如何利用文獻探勘技術分析全國碩博士論文-以「學習歷程檔案」為例	標題含「學習歷程檔案」之學位論文	研究主題、論文的被引用情形、年度篇數及趨勢、參考文獻被引用最多的作者排名及文獻
李濟福、陳志銘、曾元顯（2013）	數位學習領域主題分析之研究	SSCI期刊數位學習領域2000~2009年之文獻	研究主題架構
Tseng, Y.H., & Tsay, M.Y. (2013)	Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR	Web of Science的資訊科學與圖書學（IS & LS）領域之文章	發展適合圖書資訊學期刊之研究評估
Tseng, Y.H., Chang, C.Y., Tutwiler, M. S., Lin, M.C., & Barufaldi, J. P. (2013)	A scientometric analysis of the effectiveness of Taiwan's educational research projects	Web of Science中1990~2011年臺灣教育研究相關研究	主題發展趨勢
Yuan, Y. Y., Tseng, Y.H., & Chang, C.Y. (2014)	Tourism subfield identification via journal clustering	JCR中HLST（飯店、休閒、運動、旅遊）類別37種期刊2008 ~ 2012文獻	研究主題
Yuan, Y., Gretzel, U., & Tseng, Y.H. (2014)	Revealing the Nature of Contemporary Tourism Research: Extracting Common Subject Areas through Bibliographic Coupling	SSCI收錄2008年10種旅遊期刊之文獻	研究主題
邵軒磊、曾元顯（2018）	文字探勘技術輔助主題分析一以「中國大陸研究」期刊為例	期刊《中國大陸研究》1998 ~2015年刊載之論文	主題發展趨勢
原友蘭、曾元顯、何和鶯（2019）	運用自動內容分析技術探析觀光與旅遊領域研究主題與趨勢	SSCI收錄16本該領域期刊，1997～2016年	主題趨勢
郭知晴（2020）	自動化輔助主題分析臺灣明代研究之	《明代研究》	主題歸類、研究趨勢

CATAR實務應用

CATAR自動分類應用領域廣泛，包含科學教育、教育評鑑、數位學習、圖書資訊、教育研究、休閒旅遊、中國研究、免疫學、明代研究

多以期刊文獻

欄位	說明	欄位	說明
AU	作者	SO	期刊全名
TI	論文標題	CI	作者所屬國家
AB	論文摘要	TC	被引用的次數
DE	論文關鍵詞	PY	論文出版年
SC	論文所屬領域別	UT	WoK論文主鍵
IN	作者所屬機構	DP	作者所屬系所

表格：部分擷取自郭知晴（2020）

運用檔案材料之自動主題分類實務（1/2）

國內博碩士論文：
日治時期台灣法院檔案資料庫
台灣歷史數位圖書館（Taiwam History Digital Kibrary，THDL）所收古契書、古文書等

檔案管理局，2004年建置九二一地震數位檔案知識庫時，亦透過導入文本探勘技術試圖找出公文檔案中潛藏的內含知識，並以視覺化呈現知識探勘結果及各議題（主題）之間的因果關係，從而瞭解特定議題之處理流程，建立往後類似事件處理之決策標準（張文熙，2008）
在分群方法上係以斷詞後之中文詞彙的出現頻度為基礎，以統計方式判斷詞彙於該文件的重要性，並採取1. 依行政院檔案分類架構輔以人工建立分類規則；2. 運用非監督學習由電腦進行分群處理等兩種方式進行自動分類（張文熙，2008；蔣以仁，2009）。

發現檔案中包含之長名詞斷詞問題，對自動分類成效亦造成影響

運用檔案材料之自動主題分類實務（2/2）

2019年，德國廣播檔案館（German Broadcasting Archive）使用前德意志民主共和國廣播與電視廣播之影像資料，建構一自動影像內容分析與檢索系統，以方便使用者進行應用，主要包含有鏡頭邊界檢測、內容分類、人物識別、文本識別及相似性探索技術，並且透過使用者滿意度調查對其進行評估（Mühling, M等，2019）；

2020年，BinMakhashen, Galal M及Mahmoud, Sabri A兩位研究者針對數位圖書館所保存之電子手稿影像，提出一自動方案結合索引、分類內容檢索等，試圖改善使用者使用系統查找資訊之成效（BinMakhashen, G. M., & Mahmoud, S. A，2020）。

與檢索系統關聯，著重技術層面討論與影像辨識的結合運用

研究設計與實施

① 研究對象
② 研究工具
③ 分類表訂定

研究對象

主題性會較檔案原文更為明確

《總裁批簽》檔案目錄

包含題名、日期、內容描述、典藏號、典藏位置與製作單位等六欄位，主要會採用題名與內容描述兩較具內容主題性的項目（沒有主題欄位）

所謂《總裁批簽》指的是在1950年代中國國民黨透過改造確立由總裁——蔣介石為權力核心的組織結構中，由中央改造委員會與中央委員會評估屬「重要黨務」，並將之上呈予蔣介石批示文書的總稱，時間範圍包含1950年8月中央改造委員會成立至1975年4月蔣介石病逝，一共4361筆。

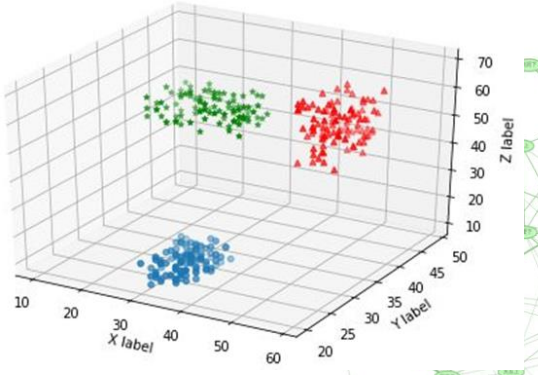
在資料的性質上，由於「重要黨務」的定義，往往取決於主事者個人主觀上的認定，未有一套明確的標準，使《總裁批簽》的內容主題，不僅反映黨務的性質，同時也具有國家事務的面向，因此《總裁批簽》的內容可說是紛亂雜陳，缺乏主題分類的情況不僅對管理者而言難以整理，對使用者來說更是查檢不易。（林果顯，2013）

研究工具（1/3）

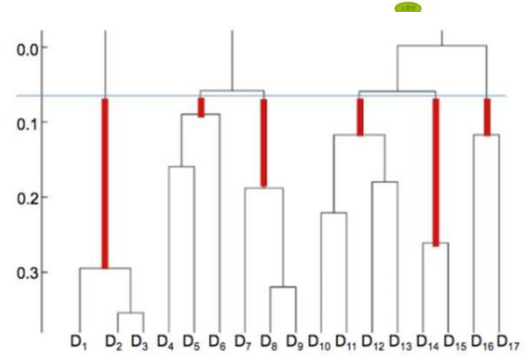
CATAR自動分類流程

$$\begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \\ \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix} \end{array}$$

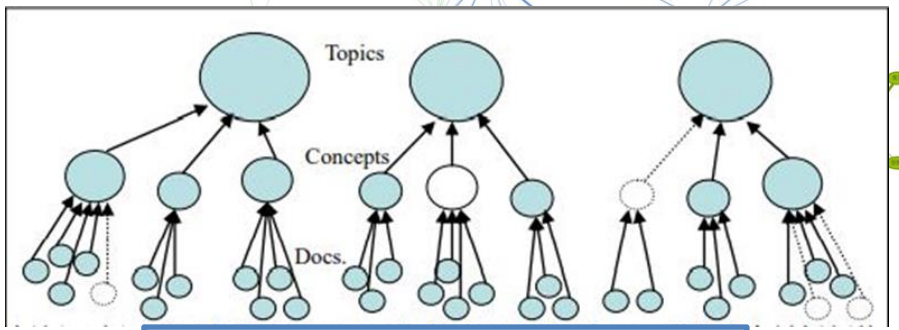
相似度矩陣



多維縮放



層次凝聚歸類法



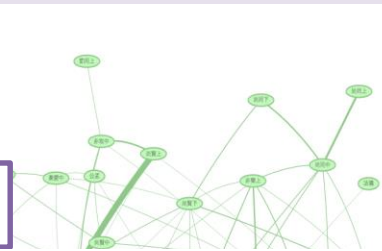
多階段主題歸類示意圖

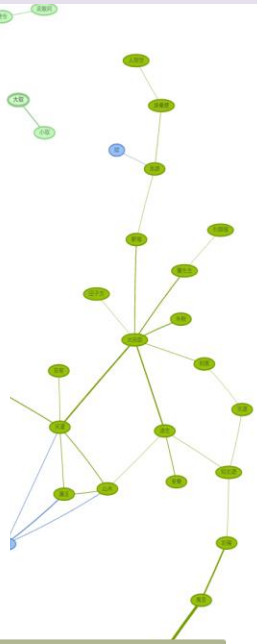
圖片來源：曾元顯、林瑜一（2011）

研究工具（2/3）

※為提升使用者的便利性

類別描述詞擷取





5(4):

- 129 : 475 Docs. : 0.1339 (中 共:88.1888, 大 陸:80.3610, 中 國:58.0049, 心 戰:49.5179, 廣 播:46.4286)
  - 15 : 439 Docs. : 0.4030 (中 共:70.3794, 大 陸:60.6717, 美 國:57.5867, 香 港:54.7048, 宣 傳:51.5505)
    - 2 : 265 Docs. : 0.5383 (中 共:49.6314, 大 陸:46.2291, 美 方:46.0000, 美 國:40.4504, 中 國:34.6718)
      - 2 : 749 : 119 Docs. : 0.1316(心 戰:56.5496, 大 陸:46.7741, 中 共:17.2428, 心 戰 工 作:14.7161, 美 方:13.0091)
      - 30 : 425 : 146 Docs. : 0.1747(中 國:45.2204, 美 國:33.7594, 香 港:22.7879, 中 共:19.3808, 左 舜 生:11.5138)
    - 29 : 2867 : 174 Docs. : 0.0318(政 府:28.5276, 來 台:25.5529, 同 志:11.9308, 反 共:9.1236, 鈞 座:8.6684)
  - 32 : 2067 : 36 Docs. : 0.0549(加 強:10.3353, 合 作:8.1909, 工 作:0.0636)

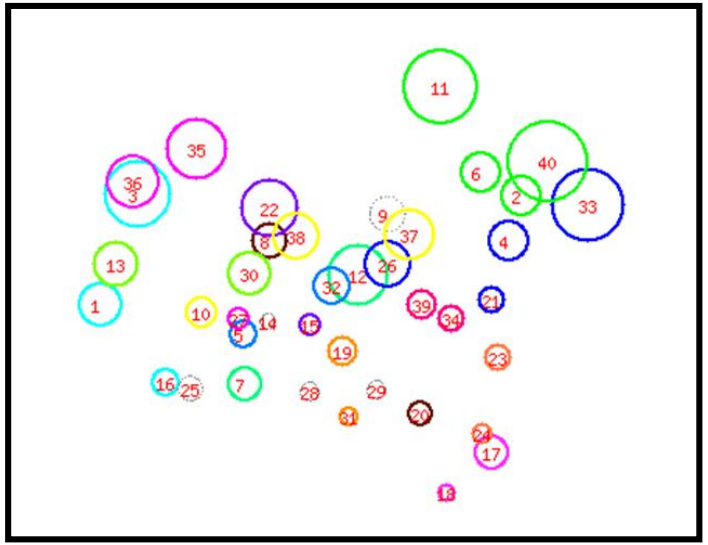
$$Co(T,C)=\frac{(TP\times TN-FN\times FP)}{\sqrt{(TP+FN)(FP+TN)(TP+FP)(FN+TN)}}$$

TP：詞彙 T 出現在類別 C 中的篇數
TN：其他類別沒出現詞彙 T 的篇數
FN：類別C不包含 T 的篇數
FP：T在其他類別中的篇數

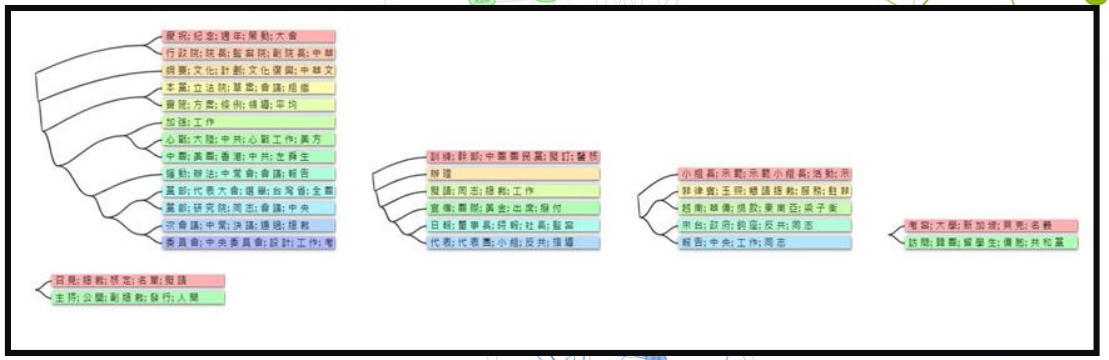
研究工具（3/3）

視覺化輔助工具

主題地圖



主題樹

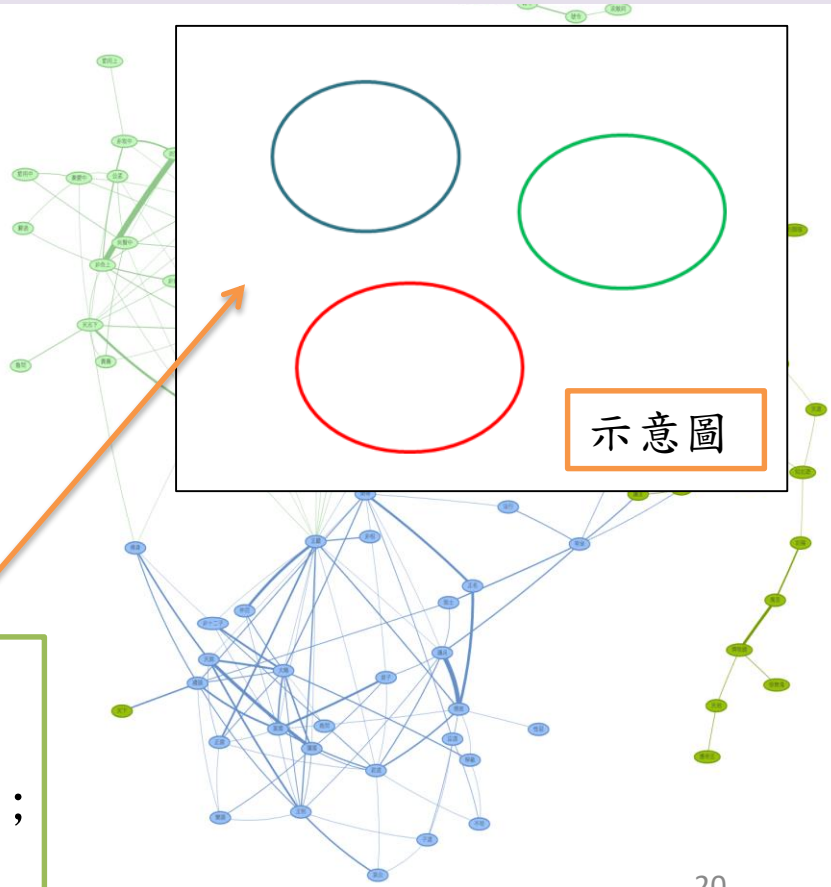


分類表訂定（1/4）

使用者操作：閾值（0～0.1）

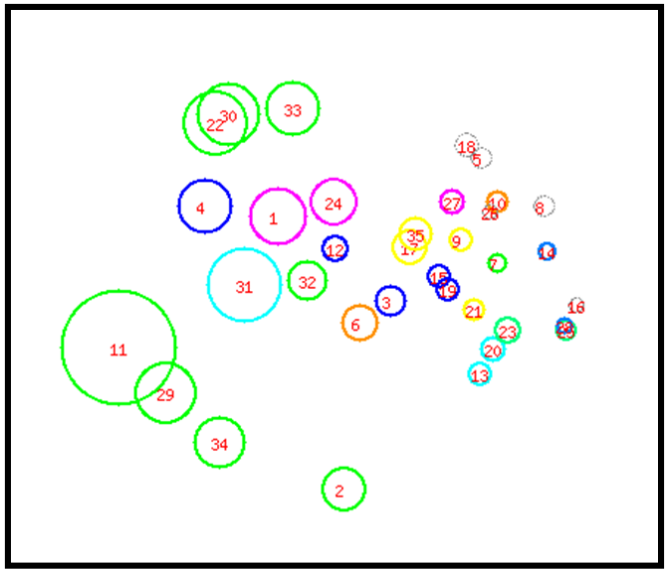
自動分類策略：
1. 符合分類原則；
2. 利用數位工具，迅速且（有效）

具體實施：
1. 類別數量不超過30類；
2. 透過主題地圖擇選主題分布具明確區隔者；
3. 結合主題樹衡量主題是否平均。

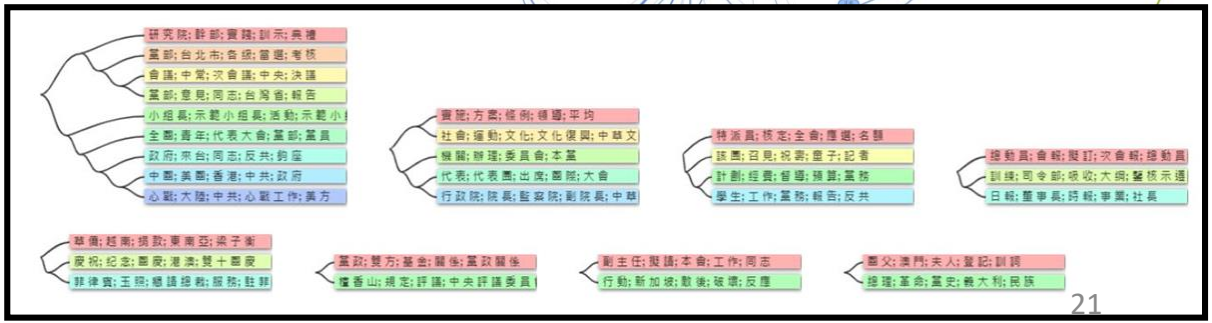


分類表訂定（2/4）

自動分類標的：內容描述項
閾值：0.03
層數：第四層



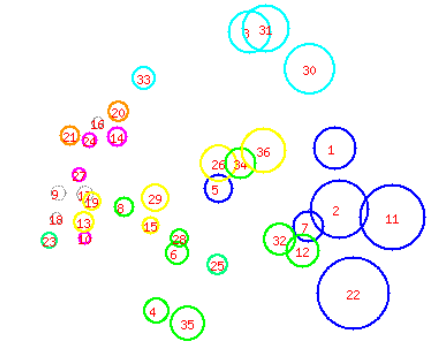
不足：
1. 主題分布仍不明確；
2. 主題詞數量仍不平均。

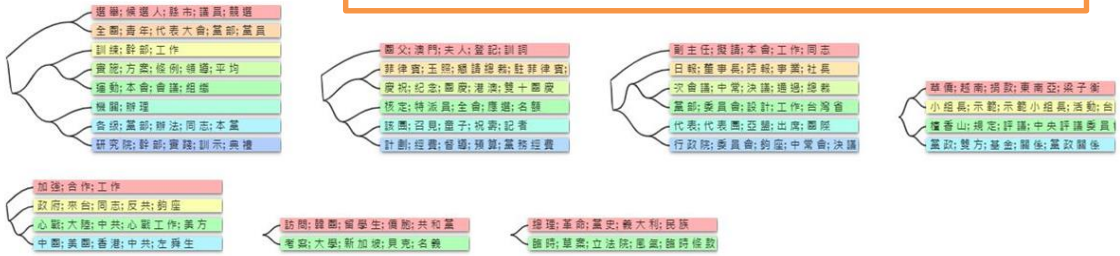


分類表訂定（3/4）

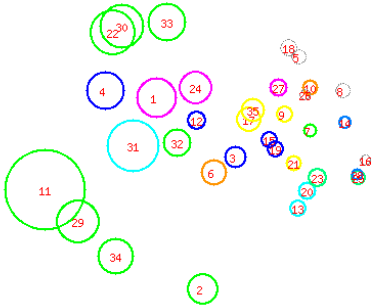
主題分布之群集關係較明確；
主題詞數量較為平均。

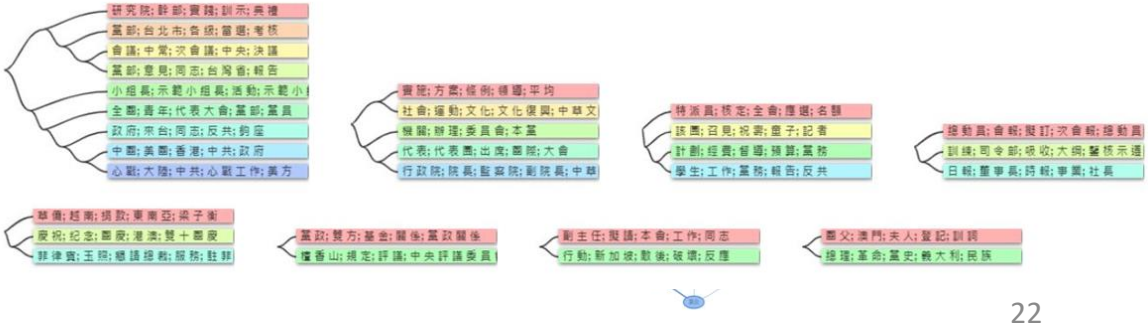
加入題名





僅內容描述

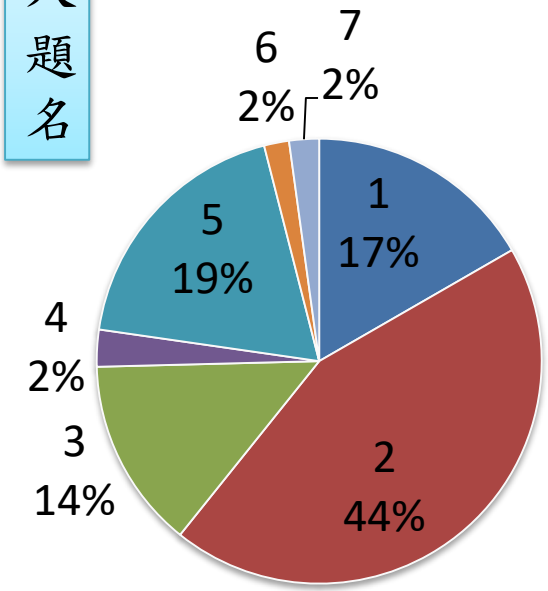




題名與內容描述比較

《總裁批簽》檔案目錄一共4361筆

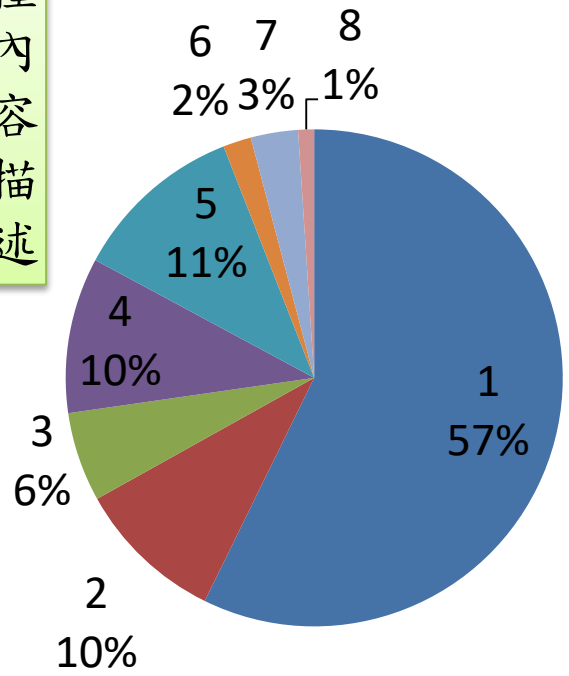
加入題名



類別	數量
1	423
2	1117
3	351
4	68
5	475
6	46
7	55

成功分類2535筆，比例約58%

僅內容描述

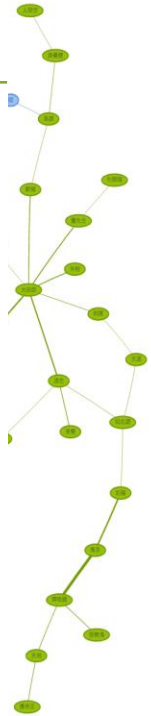


類別	數量
1	1436
2	242
3	146
4	254
5	281
6	46
7	77
8	26

成功分類2508筆，比例約57.5%

分類表訂定（4/4）

	主題詞（系統S4）	件數
1	0.0347（黨 部:99.3339, 辦 法:49.0425, 中 央:35.1331, 中 常:29.9633, 台 灣 省:25.0001）	423
2	0.1275（委 員 會:129.2451, 中 常 會:109.6520, 主 委:98.3668, 中 央:98.1195, 次 會 議:93.9111）	1117
3	0.0419（總 裁:34.3007, 黨 務:26.7953, 擬 請:24.9152, 海 外:12.6515, 中 共:11.4623）	351
4	0.0371（活 動:5.1490, 委 員 會:4.8008, 會 議:1.2683）	68
5	0.1339（中 共:88.1888, 大 陸:80.3610, 中 國:58.0049, 心 戰:49.5179, 廣 播:46.4286）	475
6	0.1749（國 大:15.2527, 臨 時:14.0000, 總 理:10.2667, 黨 史:9.0429, 風 氣:6.9561）	46
7	0.0699（留 學 生:11.1297, 大 學:9.8000, 新 加 坡:9.7385, 考 察:8.9966, 韓 國:8.6367）	55



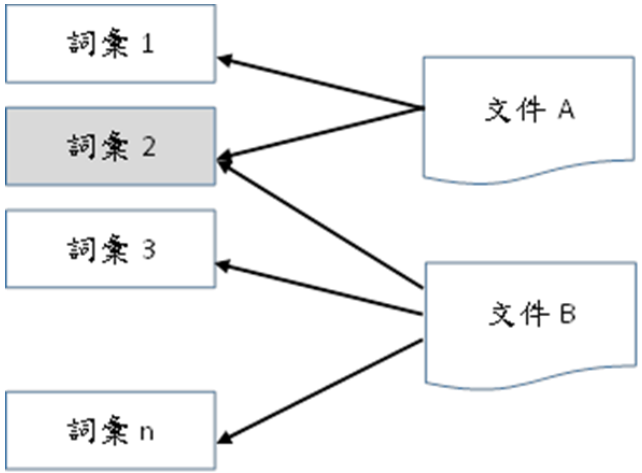
分類結果與討論

① 共現字分類討論
② 分類表主題名稱賦予
③ （分類表結構）

共現字分類討論

為什麼加入題名後的主題分布較明確？

共現字歸類



批簽題名結構：
公文字號 + 人名呈

題名包含：鄭彥棻

類別	數量
1	13
2	55
3	113
4	18
5	46
6	4
7	12

第3類總數：351件，名稱：海外黨務

圖片來源：郭知晴

分類表主題名稱賦予

類別內相似度下限值

主題詞代表性

	（分類名）	主題詞（系統S4）	內容簡陳
1	普通黨務	0.0347（黨 部:99.3339, 辦 法:49.0425, 中 央:35.1331, 中 常:29.9633, 台 灣 省:25.0001）	各地方及中央黨部事務
2	黨組織	0.1275（委 員 會:129.2451, 中 常 會:109.6520, 主 委:98.3668, 中 央:98.1195, 次 會 議:93.9111）	會議與人事
3	海外黨務	0.0419（總 裁:34.3007, 黨 務:26.7953, 擬 請:24.9152, 海 外:12.6515, 中 共:11.4623）	海外黨部以及外交事務
4	黨政關係	0.0371（活 動:5.1490, 委 員 會:4.8008, 會 議:1.2683）	調解黨務與政務
5	中國問題	0.1339（中 共:88.1888, 大 陸:80.3610, 中 國:58.0049, 心 戰:49.5179, 廣 播:46.4286）	中共、敵後工作
6	黨史／國大	0.1749（國 大:15.2527, 臨 時:14.0000, 總 理:10.2667, 黨 史:9.0429, 風 氣:6.9561）	次階層包含兩主題
7	宣傳	0.0699（留 學 生:11.1297, 大 學:9.8000, 新 加 坡:9.7385, 考 察:8.9966, 韓 國:8.6367）	交流、留學生

結論與前瞻

結論與前瞻

1. 以《總裁批簽》而言，題名可作為分類的參考依據；
2. 實務上透過設計關聯度底線，可提升整體分類成效。

1. 檔案運用數位工具進行自動分類在釐清主題分布具有可行性，可協助檔案人員進行研究出版時迅速對檔案內容有初步了解；
2. 檔案全文或者主題性詮釋資料電子化的必要性；
3. 為更進一步的提升運用範圍，可嘗試更多不同工具與方法，同時結合檔案人員、該領域之專家學者以及技術人員共同參與，可有效擬定分類策略與結果評估。

簡報參考文獻

江玉婷、陳光華（1999）。TREC現況及其對資訊檢索研究之影響。圖書與資訊學刊，29，36－59。

李龍豪、簡佑達、張俊彥、李宗諺、曾元顯（2016）。短文回應的主題自動歸類在行動教育活動上之應用初探。圖書資訊學研究，11：1，47－84。

林巧敏（2012）。檔案應用服務。文華圖書館管理。

林果顯（2013.11.22）。從總裁批簽看國民黨的對外宣傳。新史料、新視野：總裁批簽與戰後中華民國史研究，台北市，台灣。

郭知晴（2020）。自動化輔助主題分析臺灣明代研究之學位論文著作特性〔未出版之碩士論文〕。國立政治大學圖書資訊與檔案學研究所。

郭俊桔（2018）。使用多元決策略之圖書自動分類的研究。圖書資訊學研究，13：1，87－124。

陳信源、葉鎮源、林昕潔、黃明居、柯皓仁、楊維邦（2009）。結合支援向量機與詮釋資料之圖書自動分類方法。資訊科技國際期刊，3：1，2－21。

張文熙（2008）。我國檔案知識庫建置之檢討。檔案，7：1，44－57。

湯秋蓉（2009）。自動化主題分析於圖書資訊領域之應用〔未出版之碩士論文〕。國立師範大學圖書資訊學研究所。

曾元顯（2011）。文獻內容探勘工具——CATAR之發展與應用。圖書館學與教育科學，37（1），31－49。

曾元顯、林瑜一（2011）。內容探勘技術在教育評鑑研究發展趨勢分析之應用。教育科學研究期刊，56（1），1－32。

曾元顯（2014）。自動化資訊組織與主題分析近二十年來的研究與發展。教育資料與圖書館學，51：特刊，3－26。

蔣以仁（2009）。運用文本探勘建置九二一地震數位檔案知識庫。檔案，8：3，44－67。

BinMakhashen, G. M., & Mahmoud, S. A. (2020). Historical document layout analysis using anisotropic diffusion and geometric features. International Journal on Digital Libraries, 21(3), 329–342. https://doi-org.autorpa.lib.nccu.edu.tw/10.1007/s00799-020-00280-w

Mühling, M., Meister, M., Korfhage, N., Wehling, J., Hörth, A., Ewerth, R., & Freisleben, B. (2019). Content-based video retrieval in historical collections of the German Broadcasting Archive. International Journal on Digital Libraries, 20(2), 167–183. https://doi-org.autorpa.lib.nccu.edu.tw/10.1007/s00799-018-0236-z

Tseng, Y.H. (2020). The Feasibility of Automated Topic Analysis: An Empirical Evaluation of Deep Learning Techniques Applied to Skew-Distributed Chinese Text Classification. Journal of Educational Media & Library Science, 57: 1, 121-144.

敬請不吝賜教

E-mail：
adrian01072438@gmail.com