

第七章

政府檔案知識探索

Public Records Knowledge Discovery Technology

張文熙

Wen Hsi Chang

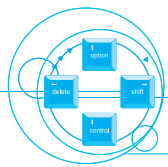
檔案管理局檔案資訊組 高級分析師

Senior Analyst, Archives Information Division,

National Archives Administration

壹、前言

爲使政府檔案資訊得以公開，檔案管理局成立之後即積極進行全國檔案資訊系統，其中一項工作爲建立全國檔案資訊單一查詢窗口，稱作全國檔案目錄查詢網（National Electronic Archives Retrieve, NEAR <http://near.archives.gov.tw>），結合巨量搜尋引擎技術，提供簡易查詢與進階查詢兩大功能；簡易查詢是指定部分欄位之關鍵字進行查詢，而進階查詢則提供機關名稱、系統流水號、檔號、案由、案名、主要發/來文者、文件產生日期、媒體型式、保存狀況、附件名稱、涉及之事項主題、涉及之人名、涉及之地點地名、涉及之相關時間、收文者、發文字號、收文字號等查詢點進行邏輯條件組合後再行查詢。全國各級機關依法送交所屬檔案目錄資料，截至九十三年一月已達八千八百餘萬多筆目錄資料，平均每一季成長量爲約二千萬筆，單筆資料量平均約爲 1.5KB，預計至九十三年底將超過二億筆以上。由於檔案目錄量日益巨增，民眾無論以任何欄位檢索組合均可能產生巨量的查詢結果，造成無效檢索比率大幅上揚。因此，如何提供有效的查詢結果，



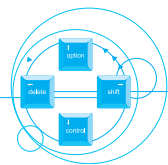
進一步考慮以資料採擷或知識探索的功能，提供檔案資料內容的附加價值，讓使用者充分取得所需要檔案的內涵，避免形成資訊科技資源浪費，並提高政府檔案資訊的地位，以期成為政府未來線上智庫的雛型。

政府機關較之於民間企業組織，就組織規模、業務內涵廣度及資訊量均毫不遜色，晚近在企業中出現所謂商業智能(Business Intelligence)觀念，成為一種決策工具，而所謂商業智能乃經由不同資訊來源資料分析而得，可以快速反應企業狀態的關鍵性指標，加強企業決策速度與品質，提升企業洞察力及遠見，影響企業市場創新與競爭力（註1）。由於政府機關業務運作均源於公文檔案，所以商業智能的觀念應用在政府檔案資訊，用來改善政府決策品質，似乎有其發展潛力。商業智能系統主要是將散落在各地的相關資料，包含內部 ERP(企業資源規劃)、SCM（供應鏈管理）、CRM（客戶關係管理）及外部資訊經 ETL(Extract, Transfer, Load) 擷取、轉置、載入工具或程序，有效地萃取、轉化成有用的資訊，集結於資料倉儲(Data Warehouse)，並透過 OLAP（線上即時分析）工具、EIS（主管資訊系統）、DSS（決策支援系統）、Data Mining（資料探勘）的應用，提供中高階主管與資料分析人員作動態報表查詢，決策判斷的參考依據，最後與企業入口網站整合，確實落實知識管理之應用（註2）。整理檔案的目的在於應用。政府檔案資訊除了來源不同於民間企業體之外，可能之應用的方式並不出其右，因此直觀判斷企業智能系統觀念，應也可套用在政府機關之運作。我國依據檔案法規定任何經過公務處理程序所產生的文件，均視為政府檔案（註3），政府檔案資料來源眾多而且複雜，如何建立一套具備企業智能的政府檔案知識探索(Knowledge Discovery)系統，如何讓所蒐集的資料產生決策支援的價值，哪些技術適用於政府檔案資訊，值得進行深入研究。

貳、政府檔案知識探索的目的

應用政府檔案知識探索技術的主要目的在解決相關檔案應用問題：

- 一、全國檔案目錄資料庫資料過於龐大，目前因應巨量資料檢索是以搜尋引擎功能縮短查詢回應時間，但卻無法進一步提供篩選適量且精要的資訊。基於商業智能之觀念，如能結合資料採擷技術，藉由全國檔案目錄擷取適用檔案目



錄之相關資訊，擷出檔案詮釋描述資料(Meta Data)間之內隱知識，進而規劃檔案知識管理系統可能應用模式，針對不同類型使用者需求提供相關資訊，減少使用者資料分析的時間，以便建立檔案知識庫雛型，進而提供公共建設相關業務決策支援之參考。將可使檔案管理工作向前推至知識管理之新局，作為強化保存數位檔案價值之有利因素。

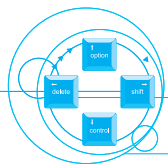
二、案目錄資訊內容可能隱含的知識，無法透過資料庫查詢語言(SQL)所能找到的事實(Facts)紀錄來呈現。主因源於檔案目錄描述檔案內容概要，限於著錄人力及背景知識，無法將檔案全文或含於字裡行間之內涵直接轉成詮釋資料，部分言及預算或計畫的檔案內容其中可能具備多維度知識（Multidimensional Knowledge）內涵，經轉置成多維度格式表示，可能可以線上分析處理(OLAP)之工具等資料探勘工具進行分析，提高檔案知識呈現的深度。規範出適用線上分析處理之政府檔案型態，是為政府檔案知識探索技術，所欲解決之問題。

三、順應時代發展趨勢，檔案全文影像資料及電子全文檔案勢必會越來越多，但缺乏有效的系統機制展現或擷取資料內涵，無法發揮檔案內含之價值。如能結合檔案目錄資訊所揭露之訊息加以整合，方能強化檔案目錄之價值。引入商業智能處理政府檔案資訊確有其誘因，但其所涉及資訊科學相關知識及技術甚廣，究竟其中何種技術適用於政府檔案處理之特定階段，或是必須改善政府檔案處理方式方能擷出相關知識，有待實驗證明。

四、檔案知識探索系統判斷何種問題可以由檔案目錄所揭露之訊息形成解答，或是作為決策參考，用以檢討檔案資料保存、典藏之政策，做為酌以調整檔案徵集的優先順序，以及鑑選作業之依據。

參、檔案知識探索相關技術分類

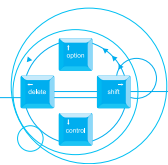
由於「知識探索」一詞定義並不清楚，楊新章與陳光華認為應該將「知識探



索」看待為「資料探勘」(Data Mining)與「文本探勘」(Textual Mining)的整合，亦即廣泛地自資料中擷取相關、適切、及時、有用的資料（註4）。現行政府檔案資訊係分兩大來源，一為檔案內文，另一則為檔案目錄之詮釋資料。檔案內文可能由影像檔案儲存，因非屬純文字資料，難以進行內容之資料探勘；而詮釋資料為系統後設所得，則多直接以文字型態貯存於電腦。由於大量的電子形態資料多數以資料庫管理系統進行管理，因此從資料庫中取得資料間內涵之知識為最可行之方式。在資料庫中發現知識(Knowledge Discovery in Databases, KDD)，簡稱「知識發現」，成為和資料探勘相互交換使用的術語。技術上，知識發現是一種運用科學方法來作資料探勘的應用，典型的知識發現處理模型則包含了擷取和準備資料，以及資料探勘後需要採取何種行動決策支援（註5）。目前許多領域可以進行知識探索的工作，包含資料探勘（Data Mining）、機器學習(Machine Learning)、統計(Statistics)、資訊檢索(Information Retrieval)、案例推理(Case-based Reasoning)、自然語言處理(Natural Language Processing) 等(註6)。

目前處理大量資訊的相關技術及其運用情況之相關學術文獻，很少單就政府檔案資訊處理可能運用情形及方式，惟僅能由分析一般性資訊組織與資訊整理之國際標準發展現況著手政府檔案知識探索。就資料應用而言，使用者並不在乎資料來源是否為結構化或非結構化資料，只要能夠提供所需的資料即可。由於檔案資料具有多樣性，必須使用不同的技術處理不同類型的資訊，然後將之包裝，使用者無須擔心技術的複雜性。檔案管理局所蒐集巨量檔案資料，在建立巨量資料搜尋引擎之後，接踵而來的使用者行為分析、文件集的統計分析，都是資料探勘所欲發展的議題。Merril 及 Tennuson 認為資料探勘就是一種學習，而學習是一種複雜的過程。學習又可以區分為四個等級：包含事實(Fact)、概念(Concept)、程序(Procedure)及原則(Principle)（註7）。當以檔案內容做為探勘對象是依據前述四個等級，所要說明的應該包含事實真相的簡單敘述、由一群檔案物件、特徵或是事件的集合形成概念，檔案敘述事件中所達成目標的連續動作，以及由一些真相所形成的通則及定律。Michael Berry 及 Gordon Linoff 定義資料探勘的周期(Virtual Cycle)為（註8）：

1. 定出企業問題所在。
2. 使用資料探勘技術轉換資料成為可以行動的資訊。



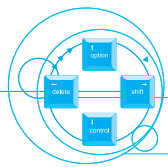
3. 依據資訊採取行動。

4. 進行結果評估。

所以系統的建立自然必須考量這四大階段應如何回應。而目前政府檔案資訊的結構，包含結構化的目錄資料，以及影像全文資料。雖然現階段目錄資料遠大於全文資料，但就系統功能面之規劃，必須兩方面都加以考量。資料探勘經常用來針對於沒有結構化的資料，建立一般化的模型（註9）。因此對於檔案全文的處理方式，應以資料探勘相關方法取代簡單檢索，則文字探勘(Text Mining)便是最重要的工具。目前大多數的文字探勘技術著重在字與字之間的距離或相似度，卻忽略文件本身的結構性，孫晉巖等提出以 Structured Cosine Similarity 結合 Spherical K-means Clustering 來表達過去查詢無法呈現的結構性，並提昇其準確度（註10）。

建立政府檔案知識探索系統目的在取出檔案知識，傳統的資料庫檢索對於事實性資料(Fact)相當成熟，雖然系統著重具備深層知識的挖掘功能，卻仍須保留一般檢索作業。適當結合全文查詢，含同音同義字，四聲相似等資料檢索，依據特定詞庫內容自動建立固定式文字索引，同時統計各索引之出現頻率，並提供索引之新增／刪除／重建等功能。配合詞庫管理，將重要的片語、複合語彙、同義字編列，和詞性規則整合，再以檔案專家之人工稽核結果，產出專業的詞彙典。系統提供自動統計斷詞處理，從檔案中自動學習斷出常出現之字詞。可以讓使用者匯入或擴充專業詞庫，強化及運用既有之知識庫，並提供維護更改詞庫或詞彙索引的功能，可使系統更實用。輔以布林代數條件、模糊關鍵字檢索與萬用字元的查詢，可強化與該類字詞有關資訊之搜集，再配合自然語言處理功能則有助於強化介面使用者親和性。由於公文檔案中多半存在主旨來說明檔案內容概要，但仍然有部分是有缺漏的，因此，針對內容產生適當的摘要值，可能有其必要。謝文泰等提以句子資訊量來產生文件摘要的模式以標點符號與連接詞為本的斷句法則將中文文章斷成適當的句子，接著以 N 元詞斷詞法配合標號，分析句子的組成結構，最後以句子的資訊量做為衡量句子重要性的依據，再按照使用者的需求，產生高可讀性以及高資訊含量的摘要（註11）值得作為參考。

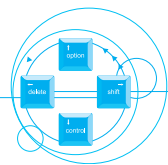
檔案知識的檢索結果如以知識概念地(Knowledge Map)來呈現，透過查詢之關鍵字詞迅速以將其所隸屬的整群知識概念進行資料檢索查閱，進而建立群式資料



索引，並依據詞庫進行特殊議題相關資料索引。知識地圖的目的是輔助需要知識的使用者能夠快速的定位知識的所在（註12）。將知識符號化，而以知識符號構成階層式知識架構，形成明瞭易懂的知識索引，以易於知識的搜尋與導覽（註13），是目前政府檔案管理作業上相當缺乏的一環，如能以此改善檔案管理分類查詢之機制應具有相當實用之價值。基本的知識地圖，則需要透過文件前置處理和關聯資料萃取來建構關聯。在文件前置處理過程中，先作有意義的斷句處理，以及文件中適當的詞性標記，最後對文件作詞目(Term)的萃取。再以詞性標記，對文件繼續從事段落與敘述文法規則的比對。用公文檔案範例句產生候選樣版節省設計時間，呈現關聯概念圖，把詞目當作實體概念，歸納建立有相同關連模式集成為節點，並考量詞目頻率，再作各種向量的統計與比對。用圖形顯示字詞主題間之依隨或因果關係，來產生知識資產，呈現某特定兩字詞間之關聯正確性高低，提供使用者協助決策研判。如能將推論與歸納分析結果均以動態圖形呈現，使用者可以在推論與歸納樹中，輕易地上溯及下溯，找出各相關主題間之可能成因與影響結果，從而獲得重要決策資訊，將提高檔案知識運用的可行性。

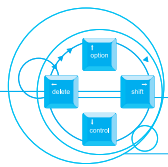
如檔案知識探索系統建立智慧型資訊代理系統來校準搜尋結果，使用者透過網際網路輸入自然問句時，由樹狀知識庫中所儲存的多個節點搜尋該問句之解答，可以作為當檔案中形成的議題解決之後，形成處理規則之檢索模式。林厚誼等結合文件探勘及概念空間(Concept Space)相關技術，以資策會電子商務技術實驗室之虛擬服務代理人，接受自然語言發問之問題，再列出相關答案的方式（註14）。檔案相關資料庫中，全文資料內容並沒有適當的系統進行分析取出其隱含的加值資料，知識探索或資料探勘自然是檔案應用發展的新希望。就資料探勘的功能而言不外六項（註15）：

1. 分類(Classification)
2. 估計(Estimation)
3. 預測(Prediction)
4. 親和性歸併(Affinity Grouping)
5. 分群(Clustering)
6. 母體描述(Description)



因此，檔案知識探索系統，所必須提供的基本功能考量也應以上述六大方向為基。資料探勘的策略可以粗分為監督式、非監督式及購物籃分析三類。當我們聽到一個個別的範例名稱時，往往會以自己所信任的概念來形成自己的分類模型。之後就會用此已發展的模型，來幫助我們辨識類似結構的物件。這類學習方式稱之為歸納法監督學習概念，簡稱監督式學習(Supervised Learning) (註16)。將輸入範例資料建立成決策樹，即是一種監督式學習的應用方式，易於了解及轉換成規則。建立訓練資料(Training Data)使得模型能正確分類，並能由測試集(Test Set)來驗證模型的正確性。分類、估計及預測屬於監督式學習的範圍。推論規則、類神經網路、及統計迴歸是常用的監督式學習技術。利用關聯法則(Association Rule)發展資料庫屬間有趣的關聯，如不處理數值屬性資料的 Apriori 演算法則屬之 (註17)。非監督式學習的分群技術則常用在測量一些範例的相似程度 (註18)。

檔案除了撰寫格式具有特殊性之外，應和一般文件的分析應用有相當程度的共通面。資料探勘技術在文件分析的應用層面，大致可分為詞彙分群(Term Clustering) 及文件分類(Document Classification) (註19)。檔案的管理作業中，分類是傳統作業的第一步，能對檔案文件進行自動的分類將是一項重要進步。文件分類的技術中，常用 Market-Basket D-T Table。文件分類的前處理過程中，有一個重要步驟，稱為特徵選擇(Feature Selection)，將相關的特徵選出，排除不相關的特徵 (註20)。並對特徵的重要性加以計算，再進行文件的分類。可以餘弦相似度(Cosine Similarity)的方式來進行分類計算，運用具有特徵重要性 D-T Table，表達成向量形式(Vector Representation) (註21)。自動階層式分群對檔案的內文以距離公式計算內容相似度，再將文件相似度大於某一臨界值 (Threshold) 的分類於同一群組中，即為同屬於某一概念(Concept)，並採階層方式建立索引。上層父代部分將含括下層子代之內容索引，所有之最底層子代，將擁有最清楚明確的分類，將已知/已有的內容，增加橫向的知識連接，應是檔案知識探索之基本作業。就非結構化的檔案內容文字，必須事先完成清洗等前處理的作業，方能提高探索結果的正確性及可用性。黃如玉等曾針對應用文件探勘法於產業新聞知識發掘，提出以資料庫知識發掘(Knowledge Discovery from Database, KDD)發展出文件探勘架構，主要是以全文資訊萃取(Information Extraction)作為前處理，將資訊轉化成特定結構的資訊，



再將現有的資料庫知識發掘技術作為後處理，形成完整的文件資料探勘系統（註22）。政府檔案資料，雖文件格式呈現和新聞資料不同，特別在公文檔案內容文義及表達方式，有比較清楚的定義，由而公文的撰寫有既定格式，內容組成近於半結構化，由於公文的產出流程較一般文件嚴謹，故理論上有助於機器的運作。

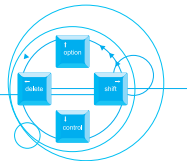
而對結構化檔案目錄資料，則應有針對缺漏值(Missing Value)部份進行資料處理(Data Processing)作業，以確保探勘結果的準確程度。貯倉法(Binning)是利用該欄位的鄰居以平緩一排序後的數值，先將排序後之數值分配至數個倉中，同一倉中之數值再以下列方式平緩（註23）：

1. 倉均值(Bin Means)：將倉中的所有數值以平均值取代。
2. 倉中值(Bin Medians)：將倉中的所有數值以中數取代。
3. 倉邊值(Bin Boundaries)：取倉中的最大值或最小值，以最接近該值的邊值來取代。結構化資料中難免會出現異常值，則可以 Null Instance Filter、Empty Value Filter 及 Sift Filter 等方法來加以濾除，或以人工填入合理值。

資料探礦之基本步驟，以政府檔案知識探索為目標，不論探礦對象為結構化或非結構化之資料，其由從資料清洗開始、進行整合、選取、挖掘、形成模式產生知識的流程大致可以確立（註24）。目前檔案管理局建置之全國檔案資訊系統中約有八千八百萬筆檔案目錄資料（至九十二年一月止），均屬檔案的描述資料（Metadata），由全國各機關所陸續上載，但資料的完整性不一。其中僅部分國家檔案，如「二二八事件檔案」，具有完整的目錄資料；但檔案全文本身為影像檔，且手稿居多，人力辨識困難，其餘大多數檔案目錄資料並無相對應之檔案電子全文資料。雖然，結構化資料與非結構化全文資料，並非一對一對應關係，但就系統開發而言，兩者均足以產生適當之附加資料價值，不可偏廢。

肆、檔案知識探索系統架構

建立檔案知識探索系統是希望藉由檔案知識探索的結果提供若干決策支援的

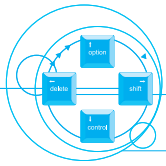


功能，並且能就探索結果取得特定議題未來發展趨勢預測及分析。楊新章與陳光華認為檔案知識管理系統功能應具備使用者管理、知識探索及知識分享模組，其知識探索模組之功能檔案目錄目其中有相當多相同的資料。因此，必先聚集相同檔案的目錄資料，才能進行後續的工作。如下：（註25）

1. 聚集相同檔案之目錄資料
2. 建構檔案資料之時間序列
3. 建立檔案資料之相互參照
4. 建立相關檔案資料之群集
5. 抽取重要事件與人物
6. 分析機關檔案分類表
7. 建構動態統計共現索引典(Statistical Co-occurrence Thesaurus)

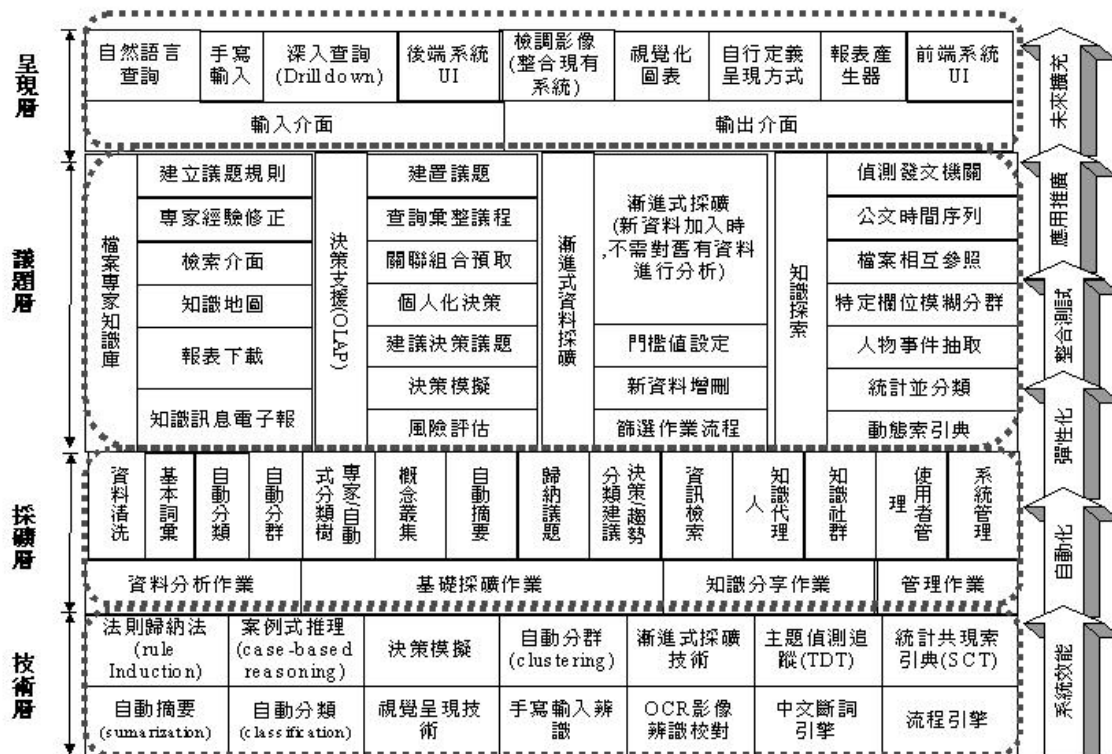
楊陳兩位所規劃之系統架構是以知識管理為基本模式，引入知識分享模組（Knowledge Share Module）利用推播技術（Push Technology）主動將檔案知識傳遞給使用者的方式。系統得依據日期（Current Date）、職掌（Duty）將所藏相關檔案資料主動傳遞給使用者；或結合群組關係、職務關連性主動傳遞相關檔案資料給其他使用者，相當具有實用性，惟推播技術必須建立在探索結果已然形成有用之知識才有意義。因此，必須針對知識探索結果可能應用方式、流程及非結構化的全文檔案內容進一步規劃，所以檔案知識探系統架構應加入資料現狀及作業目的定出取捨原則，以落實作業之可行性。根據前述資料及國外政府檔案資訊應用架構為思考出發點，初步規劃政府檔案知識探索系統可能的邏輯架構，如圖一所示：

如將探索政府檔案知識定位為增進決策品質作為目標，就系統作業架構及流程，可能涉及之相關問題，加以區分為技術層、採礦（探勘）層、議題層與呈現層四大層面再行討論。技術層提供採礦作業的基礎技術，包含法則歸納法、案例式推理、自動摘要、自動分類、決策模擬、自動分群、視覺呈現技術、手寫輸入辨識、OCR 影像辨識校對、漸進式採礦技術、主題偵測追蹤、中文斷詞引擎、統計共現索引典、流程引擎等。採礦層提供採礦作業流程，包含資料分析作業、基礎採礦作業、知識分享作業與管理作業等。議題層提供議題層次的知識發掘機制，



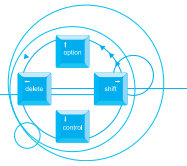
包含檔案專家知識庫、決策支援模組、漸進式採礦模組、知識探索模組。呈現層用以提供友善便利的使用者介面，包含自然語言查詢、手寫輸入、深入查詢、後端系統介面、整合現有系統提供影像檢調、視覺化圖表、自行定義呈現方式、報表產生器、前端系統介面等。

檔案知識探索必須由原始資料來源著手，由文件處理流程可以析出系統實作必須選用之相關技術，組合而成適用於檔案知識探索之應用系統。系統可依所採行資料探勘技術之分析技術，解譯文件中各個關聯意義，及於群組之間的相關性。先行向量化所搜集之文件，並依據現存政府機關檔案分類進行類別比較。依據在各分類中的隸屬度進行排序，並參考臨界值以決定文件分類。高排序且隸屬度大於臨界值的文件會被自動歸類到某些類別中，新增加的文件會再度用以進行遞增式學習，藉由評估函數之評量，以逼近文件分類之最佳的特徵值。



圖一 政府檔案知識系統邏輯架構

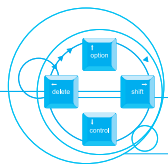
資料來源：檔案管理局檔案知識庫基礎建置案建議書 廖文彬繪



根據分類的規則設計專屬之自動代理人程式，作為資料搜索的依歸，透過自動代理人程式進行文章的蒐集與建立連結；自動代理人將自動作概念搜尋，搜尋對象定為相關文件或是人工登錄的文獻，將所有存於某分類中常出現的詞彙，或是概念相依密切的詞彙，帶入搜尋機制中進行相似度比較，將越相似的文獻於以蒐錄，並利用新值校正。建立代理人機程式(Software Agent)萃取出文獻中有意義的知識內容，透過所歸納的特徵向量，逐漸縮小搜尋資料的範圍，藉助資料探勘技術的特徵萃取技術完成。人機界面提供處理自然語言指令的代理人，以多代理人系統(Multi-Agent System)的方式運作，幫助使用者執行相關的指令及操作。代理人機制可分析使用者的查詢需求，對出現的字詞做語意階層擴張篩選，初步挑選出有意義的關鍵字組合。拆解對應至不同操作主題，再分配給適當的代理人做執行簡化操作程序。

資料探勘最終目的是期能由使用者經由應用系統所學習而得結果，採取適當的行動，諸如商業用途常見的選定促銷資訊寄送對象，物品架位排列或觸發其他科學研究等均屬之。因此，針對不同性質類型之政府檔案中之議題，設計實驗並驗證系統及方法之有效性，才能訂定出系統適用範圍。Sir Francis Bacon 在其著作 New Instrument (Novum Organum) 中說明科學方法的步驟為定義問題、制定假說、做實驗、做出結論及驗證結論，Richard J. Roger 及 Michael W. Geatz，對照科學方法將資料庫知識探索(Knowledge Discovery in Database, KDD)定義七大步驟，分別為訂定目標、建立目標資料集、資料前處理、資料轉換、資料探勘、解釋評估及採取行動(註26)。實驗設計所包含之建立目標資料集、資料前處理、資料轉換、資料探勘，應於系統實作時業已完成，實驗產生之結論及結論驗證則是用以解釋評估系統成效，並且轉化成可行的行動，是為研究最終目的。故此，經過解釋與評估決定一個學習模型是否可被接受，並且可應用在測試環境以外，方足以說明系統具有能力轉化成可行的行動模式。假若結果被證實為可接受的，那麼這個階段即可將學習得到的知識轉換成使用者可以了解的敘述。驗證系統方式可以用多種方式來進行包含：

- (一) 統計分析：用不同的屬性和資料範例建立出不同資料探勘模型，比較其結果是否存在重大的差異。

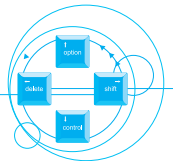


- (二) 試探性分析：以數值試探法決定哪些程度的資料是屬於被探索到的，或計算類別的相似性。
- (三) 實驗分析：如類神經網路及 **K-Means** 演算法用相同的資料及參數，每一次建立出來的模型也都會有些微的差異。建立如何選取不同屬性、資料範例和參數設定來取得明顯不同結果。
- (四) 人性分析：了解是否因為實驗程序未被良好控制，影響結果能否成功應用。

因此，建置檔案知識探索系統必須透過適當的方法來驗證其有效性，系統才有價值。如藉實驗評估模型集測試集之正確性以決定信賴區間、用統計評估來比較監督式學習模式的正確性、應用假說檢定決定數值型屬性的重要性、使用監督式學習法來評估非監督分群法的結果、評估具數值型輸出的監督式模型等種種手段，從而建立政府檔案探索作業參考程序及可行之系統架構。

伍、檔案知識探索系統未來發展範圍及限制

一、以檔案管理局檔案目錄資料庫為範圍是現階段進行政府檔案知識探索較能有效掌握的資源來源。並同時運用文字探勘(**Text Mining**)及資料採礦(**Data Mining**)作為檔案知識發掘之主要作業架構。期能以電腦化系統工具進行大量檔案資料文字內容探索、分析及進行機器學習作業，以求能夠從中發掘出檔案資料內涵，形成特定議題的知識模式辨識(**Pattern Recognition**)及處理規則(**Rules**)，得以作為決策作業支援參考。為避免範圍過大形成研究失焦，如設定以公共建設為主軸設定特定議題，縮小驗證範圍，證明運作之可行性，而非全面性漫無目的進行檔案知識發掘作業。但相對地，也可能因為範圍限制局限對檔案知識探索功能的觀察深度。另外，原始資料輸入的品質也可能影響知識探索的結果，但因為數龐大難以逐一要求。



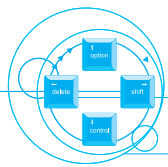
二、藉由自然語彙處理及文字資料分析之能力，可強化檔案知識概念的類聚。並以時間作為分析變因，解釋過去事件資料發生時序與檔案間之關係，透過現存檔案資料的相關資訊，系統並得以預測未來發展趨勢。但檔案所記錄的時間係以公文開始處理的時間為準，並未由描述檔案詮釋資料中清楚反應事件發件時間，可能造成預測失真。

三、檔案管理局目前僅部分檔案目錄資料有完整電子全文資料相對應，如九二一地震政府文書檔案資料庫。擇其作為資料來源範圍較小難度較低，所建置之檔案知識庫基礎應用模式，未來可以依循此作業模式擴大應用至不同主題類型之檔案數位內容。但極可能因為資料來源內容性質差異過大，作業模式無法套用，或產生預期錯誤。

四、前置處理之資料清洗(Cleansing)作業，目的在去除資料庫中雜訊(Noise)或無法辨識應用之資料，以增加正確性。為減少人工作業，影像進行辨識作業之人工校核作業，以抽驗為原則，必要時則全數捨棄，其餘則藉由資料庫資料清洗作業，篩選出可用之目錄詮釋資料(Metadata)及全文進行探勘作業，但很可能因此產生採礦結果偏差。

五、文字(Text)及資料(Data)採礦前，系統應建立基本詞彙資料庫。系統詞彙資料庫可針對處理文件新增詞彙部分，須具備抽取資料庫中文件內的新詞功能，以及中文基本斷詞能力。分析詞彙時應具備分析及計算出詞彙在文件中的重要程度，主題採礦部分能找出文件資料庫中重要的主題。但特定議題的專業領域基本詞彙取得不易，可能影響分析及探勘結果。

六、現有結構性資料檔案目錄(Metadata)及非結構性資料(全文資料)，系統得自動完成分類作業。分類結果須提供人工檢核機制，及修正自動分類法則之功能。於完成非結構性資料及結構性資料擷取後，可針對結構性資料及非結構性資料內容，再行資料分析或歸類作業，並提出基本統計分析。但人工檢核人力



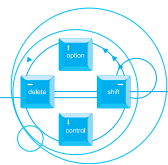
負擔過大，形成降低作業可行性。

七、以現行檔案資料庫所收之檔案文件內容分類(Classification)情形、描述現存資料庫內容檔案與檔案間 (Affinity Grouping) 不同類文件歸併相關性、提供內容類聚處理功能(Clustering)、檔案內容狀況描述等基本資料。內容繁複，不易印證所選用之相關演算方式正確率，必須涉及相當多的人力進行，實務作業困難度極高。

陸、結論及建議

進行政府檔案知識探勘的目的，在推動政府檔案資訊深層處理之觀念，藉以扭轉一般人對政府檔案文件僅止於保管典藏之應用觀念，進一步建構檔案知識庫基礎系統作業模式。將電腦自動化處理技術運用於政府資訊，提升檔案資訊價值，並且發展將非結構化政府資訊轉換為結構化政府資訊的機制及流程。同時，依實作結果分析目前檔案管理局所訂定政府檔案資訊詮釋資料(Metadata)的相關標準是否足以提供足夠的資訊產出附加價值，藉以檢討檔案目錄資料及電子檔案保存內容妥適性。透過系統運作進而了解政府檔案資訊內涵中，哪些議題適合監督式學習，哪些應以非監督式學習完成；哪些資料應建立模組的資料，哪些則是用來作為測試模組的資料，系統使用經驗得以讓其他有意進行政府檔案資訊的知識探勘之機關有參考的作業模式。筆者整理相關可能應用於政府檔案知識探勘技術形成下列結論：

一、知識探索模組擷取檔案文字中結構與非結構性資料內容，可對已定義字彙或概念、產生基本統計資料。並將同一機關所發之同一文偵測出來，聚集相同檔案之目錄資料。可以發文機關建構檔案資料在公文處理流程之時間序列。並偵測意涵相似之檔案並建立相互參照表，以建立檔案資料之相互參照關係。抽取特定機關之重要事件與人物及抽取特定時期之重要事件與人物，進



行統計與分析各機關所處理之檔案型態、數量、內容、其他參與機關等資訊，並建立分類。依使用者使用過程統計各檔案之共現圖型，得以動態統計方式建立檔案間共現索引典（Statistical Co-occurrence Thesaurus）。

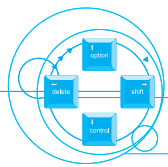
二、系統提供以由上至下的思考模式(Top-Down Manner)，允許系統自動由資料庫資料分析結果產生問題假設，並提出假說問題驗證(Hypothesis Testing)相關解答。也可以提供由下而上進行(Bottom-Up Manner)檔案知識探索(Knowledge Discovery)，得以直接或接間方式進行。直接型檔案知識探索指可以由系統自動對類聚（Clustering）結果自行解釋；間接型檔案知識探索則需提供不同類別資料關係形成之模式(Pattern)。足以建立政府檔案知識應用流程。

三、結合檔案資源與決策支援功能(Support Decision-Making)之關聯作業，依問題類型產生可用之決策參考資訊，產出相關統計圖表，便於決策者理解分析結果。能依時間序選取資料分析範圍，分析歷史資料提出未來事件採取行為之建議。

四、解決檔案知識資料擷取（Data Acquisition）、資料貯存管理介面、產出結果派送（Information Delivery）、專家規則分析（Drill-Down Analysis）、知識產出（Roll-Up Analysis）分析、現存資料庫概況分析(Drill –Through Analysis)、資料設限(如以時序或特殊概念)分析、報表產生器。提供使用者線上分析(Online Analytical Processing, OLAP)功能，快速呈現結構化資料間之關係。

五、應用代理程式技術(Agent Technology)可以取代使用者進行預先設計的工作產生預測或評估之結果。

六、發展以概念資料為主的檔案應用模式，可針對某事件相關會議紀錄及決議辦理情形相關，如以九二一重建委員會針對集合住宅所開的會議及相關處理進行探勘作業。



七、配合機器學習及統計分析功能需要，可重複匯入相關資料提高產生結果之正確性。可提供依資料分類樹架構瀏覽功能，包含檔案資料瀏覽及檔案知識規則(Rules)，可由使用者定義資訊呈現的型式、興趣等，建立個人使用規則並提供完整管理平台。

八、可由資料分析結果，歸納出檔案知識庫基本議題、提供決策建議分類及趨勢預測之分類，進而建立檔案專家知識庫模型。依據檔案資料內容，擷取、分析所形成檔案知識議題規則(Rules)，並能以知識地圖方式表示所含內容分類狀況。提供人工專家經驗針對知識庫內容新增、刪除及修改等功能。經由檔案資料內容所取出之摘要形成基本檔案知識，可提供高階主管進行事件處理程序參考，輸出結果可以視覺化工具表示，簡化高階主管理解時程。以案例推理(Case-Based Reasoning)與法則歸納法(Rule Induction)等，以整合型模式建立檔案知識庫系統。並提供適當之決策支援，藉由知識探索模組所得之概念群組作為主題，定義決策支援議題。

註釋：

註1 謝清佳、吳宗璠，資訊管理理論與實務，頁659（台北市：智勝，2003）。

註2 同註2，頁659-660。

註3 檔案管理局，檔案法，頁1（台北市：檔案管理局）。

註4 同註1，頁23。

註5 曾新穆、李建億，資料探勘，頁5（台北市：東華，民國92年）。

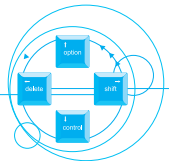
註6 丁一賢，游耿能，蕭如淵。「在非結構化文件中利用探勘方法以建構索引式知識地圖探討」，頁1-15。2002科際整合管理國際研討會。

註7 同註7。

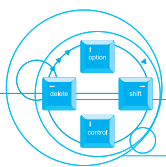
註8 Michael Berry and Gordon Linoff, Data Mining Techniques: for Marketing, Sales, and Customer Support: p.22. (New York: John Wiley & Sons, 1997)

註9 同註6，頁23。

註10 孫晉歲，苑守慈。「Ontology-Based Task-Oriented Voice Mining for Mobile B2E



- Application」頁297-304。第八屆國際資訊管理研究暨實務研討會。
- 註11 謝文泰，陳文銖，張履平。「以句子資訊量來產生文件摘要之模式」。頁661-666。
- 註12 同註5，頁4。
- 註13 Fu-Ren Lin， Chih-Ming Hsueh, “Knowledge Map Discovery in Virtual Communities of Practice”第十二屆國際資訊管理學術研討會，台灣大學，A-2。
- 註14 林厚誼，蔣岳霖，周世俊。「The Design and Implementation of Act E-service Agent Based on FAQ Corpus」。頁137-142。
- 註15 Michael Berry and Gordon Linoff, Data Mining Techniques: for Marketing, Sales, and Customer Support : p.51. (New York: John Wiley & Sons, 1997)
- 註16 同註7。頁8。
- 註17 同註7。頁51。
- 註18 同註7。頁52。
- 註19 歐陽彥正、葉建華、黃賢卿，「資料探勘的技術與應用」，檔案季刊第二卷第二期，頁15（台北市：檔案管理局，民國92年）。
- 註20 同註8，頁20。
- 註21 同註8，頁20。
- 註22 黃如玉、李惠明、陳武倚。「應用文件探勘法於產業新聞知識發掘」，頁333，（台北市：文化大學資訊管理研究所）。
- 註23 同註1，頁86。
- 註24 同註1，頁28。
- 註25 同註1，頁53-54。
- 註26 同註7，頁160。



檔案資訊資源管理
