知識探索及其於政府資訊之應用

Knowledge Discovery and Its Applications to Government Information

陳光華
Kuang-hua Chen
國立臺灣大學圖書資訊學系副教授
Associate Professor, Department of Library and Information Science, National Taiwan University

呂明香
Ming-Hsiang Lu
天主教輔仁大學圖書資訊學系講師
Lecturer, Department of Library and Information Science, Fu Jen Catholic University

摘　要

本文探討知識探索的內涵，並分別就結構化資料與非結構化資料，說明知識探索的相關技術。最後，以全國檔案目錄資料爲例，討論了幾項知識探索於政府資訊的應用，以期由時間歷線、交互參照、關連相關、或人物事件，探究檔案間可能的各種關係或未知的知識。

Abstract

This paper first discusses the coverage of knowledge discovery. Secondly, the related techniques of knowledge discovery for unstructured data and structured data are described, respectively. The last but not the least, this paper identifies a few possible applications of knowledge discovery for government information. The variant relationships and unknown knowledge could be discovered by using times series, cross reference, data association, or persons and events.

關鍵詞：資料探勘、政府資訊、知識探索、文本探勘
Keywords : Data Mining, Government Information, Knowledge Discovery, Text Mining

壹、緒論

全球資訊網（World Wide Web，簡稱WWW）已成爲資訊流通的重要管道，無論是個人、企業、政府都將各種資訊建置爲網頁，透過全球資訊網提供資訊需求者瀏覽與查閱。大量資訊傳播流通，帶來許多政治、社會與經濟發展的改革，如 e-government 的概念逐漸受到各界的重視，其中最重要的動力其實來自人民對於政府資訊公開化與透明化的需求。而政府資訊 e 化程度，也成爲國際間重要的國家競爭力評估指標。

美國布朗大學 2001 年全球各國政府資訊化的評比，台灣名列第二，僅次於美國；2002 年則更進一步躍升至第一。（註 1）這項評比結果令從事資訊工程與圖書資訊研究的學者，一則以喜，一則以憂。喜的是以往在硬體工業發達，受到忽視的軟體事業，似乎已受到政府應有的重視，其努力也收到良好成效。但憂的是，台灣各級政府資訊化的程度，真是如此卓越，足以領先眾多先進國家名列前茅，還是只是一種表像？

若是檢視目前各級政府的資訊網站，可以發現目前政府的資訊網站仍處於提供原生資料的階段，提供有限的全文檢索功能，或是於內部網站提供各部門公文線上簽核、管理等功能。事實上，各政府機關將比較多的心力用於版面構成、視覺呈現，這部分雖然是極爲重要的一環，對於資料的瀏覽與取用有很大的幫助，但是對於龐大的政府資訊，如何由縱向取得各時期相關的資訊，或是橫向瀏覽各機構不同卻相關的資訊，如何有效地組織整理政府資訊、如何處理老舊的政府資訊（Legacy Government Information）、如何有效地描述政府資訊等課題，就目前電子化政府發展階段而言，並沒有滿足上述的各項需求。

台北市政府自 1999 年起，已舉辦四次「網路新都金像獎」，針對市府各單位資訊網站進行評比。行政院研考會亦制訂「政府網站評鑑指標」，並從 2001 年起開始辦理「行政機關網站評獎」活動（http://www.a-site.nat.gov.tw）。這些評獎活動的目的不外乎，藉由評鑑各級政府機構網站，提升公務人員之資訊素養、加速政府資訊之數位化、提供及時的政府資訊、建置便民的資訊管道、發展有效的檢索機制、建構統一的資訊平台。

檔案是提供不同機關間業務聯繫與經驗交流，以及民眾瞭解與認識政府的絕佳管道。檔案管理局於民國 90 年 11 月 23 日成立後，即積極進行全國檔案資訊系統與機關檔案管理系統的建置工作，其中「機關檔案管理資訊網」（http://online.archives.gov.tw）收錄全國各機關上載的檔案目錄資料。爲了進一步讓一般民眾使用這些資料，檔案管理局亦建立「全國檔案目錄查詢網」（National Electronic Archives Retrieve, 簡稱 NEAR，http://near.archives.gov.tw），提供簡易查詢與進階查詢，進階查詢提供機關名稱、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。

NEAR 提供的功能，仍屬一般網站的檢索功能，並沒有進一步考慮資料探勘（Data Mining）、文本探勘（Text Mining）或知識探索（Knowledge Discovery）的功能。有鑑於 WWW 已成爲提供資訊的標準平台（de facto standard platform），如何在這樣的平台上，針對政府資訊的特性，發展適合的資料探勘或知識探索技術，是未來 e-government 重要的課題，也是提昇國家競爭力的重要指標。

學術界對於知識探索的研究與企業界對於知識探索的應用，已有一段日子了，也有一定的經驗，成功的應用案例亦不少見，應用知識探索於政府資訊是一項非常可行的作法。本文簡短地描述知識探索的內涵，分別就結構化資料與非結構化資料，說明知識探索的相關技術。最後，以檔案管理局收錄的全國檔案目錄資料爲例，探討多項可能的應用，以提供各界參考。

由於國內學界對於 Data Mining 與 Text Mining 的中文翻譯並沒有一致的通行術語，目前 Data Mining 有資料挖掘、資料探勘、資料採擷等翻譯散見於文獻，而 Text Mining 倒不多見學者將之翻譯，多數直接使用英文詞彙，若對照於 Data Mining，亦可能翻譯爲文本挖掘、文本探勘、或文本採

擷。本文爲了行文之便，將分別採用資料探勘與文本探勘作爲 Data Mining 與 Text Mining 的中文翻譯。至於本文題名所使用的「知識探索」則涵蓋資料探勘與文本探勘。此外，本文探討的皆爲文字型資料，然而文字型的結構化資料可用以描述音訊、視訊、圖像、圖片等非文字型資料。

本文結構如下所示：第貳節探討政府資訊的相關研究。第參節說明知識探索的技術，及其可能的應用。第肆節將以檔案管理局現有的檔案目錄資料爲例，討論如何以知識探索的技術擷取相關資訊。第伍節則是簡短的結論。

貳、文獻探討

基於自由取用無安全顧慮資訊的精神，政府資訊的開放與應用，除了供人民取得資訊外，協助政府公務人員，有效運用政府資訊，訂定政策並執行政策，是極爲重要的工作。這個趨勢已經受到全球各國的重視，也有眾多的學者專家進行相關的研究，政府機關亦投注經費建置資訊取用的機制。以下簡單介紹相關的研究。

陳瑩芳曾討論美國政府資訊指引服務（Government Information Locator Service，簡稱 GILS），GILS 可以說是政府資訊開放使用的重要一步，隨後，加拿大、澳洲、英國、日本等國亦建置類似的服務機制。（註 2）

美國聖地亞哥高速電腦中心（The San Diego Supercomputer Center，簡稱 SDSC）與學術機構以及政府機構合作，進行多項檔案數位化與檔案知識管理的計畫。（註 3）他們認爲永續性的數位檔案典藏方案，應該是以知識爲考量的基礎，而所謂的知識是資訊加上「脈絡與規則」（Context and Rules）。（註 4）脈絡與規則很可能是隱晦不明的，很可能必須由大量的資訊推導而得，很可能不直接包含於檔案文件中，必須透過其他方法得知，因此，SDSC 也規劃應用知識探索的技術，結合相關的標準（如 Topic Map），產生巨量檔案文件的知識。（註 5）

美國有關健康或稽核政府機關亦嘗試使用資料探勘技術，以找出醫藥資源浪費或濫用的型態。喬治亞州稽核局（Department of Audits, Georgia）使用簡單的規則，抽取資源濫用與其他資料的關連關係（Association），亦可發現違規事項的型態。威斯康辛州健康與家庭服務局則推估使用資料探勘的技術，每年可以節省 1 千萬到 2 千萬美元。（註 6）

數位政府研究中心（Digital Government Research Center）發展許多技術處理巨量政府資訊異質性、分散性、詞彙歧異性造成的問題。該中心執行的能源資料蒐集計畫（Energy Data Collection project，EDC），自美國聯邦政府機構取得大量的資料，建置 EDC 系統，該系統有一個整合的知識體系，提供標準的術語，以及術語之間的關係；可以整合不同資料庫的資料；協助建構查詢問句的功能，特別是支援決策的查詢問句。（註 7）

美國政府對於恐怖攻擊行動的偵防在 911 事件後更加積極，聯邦調查局運用了資料探勘與文本探勘的技術，從大量的新聞資訊、政府調查文件，擷取恐怖攻擊事件的型態、時間、地點，建立關聯資訊，期望由此預防可能發生的攻擊事件。（註 8）此外，美國國家標準與技術研究所（National Institute of Standards and Technology，NIST）推行一系列文件研究計畫與論壇，藉由推動文本探勘的研究，發展更先進的技術，以協助聯邦政府機構擷取重要的訊息。（註 9）

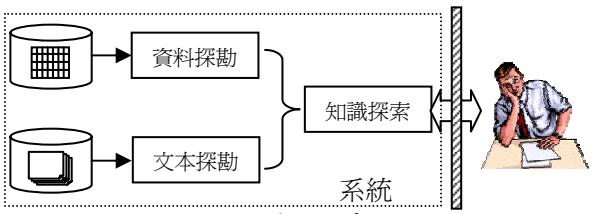
加拿大政府正在進行的 Government On-Line 計畫，揭櫫該國預計於 2005 年透過網路提供的政府服務項目。在全面推動各項計畫前，加拿大政府提供可觀基金，鼓勵各政府機關執行先導計畫，2001 年總計有 55 個計畫獲得補助，預算金額達六千萬加幣。其中加拿大統計局提出 Creation of a Data Exchange and Knowledge-Sharing Environment 計畫，係基於深信相關政府統計資訊，可以加速醫療、教育及法治等公共事務資訊有效且正確地傳遞，所以規劃建構社區醫療諮詢網、各級法院資訊交流網及各級教育行政聯絡網等網絡，使各相關機關在安全環境機制下，取用及互享所需統計資訊。（註 10）

新加坡政府自 1980 年全國電腦自動化計畫展開後，及其後的 IT2000 智慧島計畫，

到最近以“To be a leading e-Government to better serve the nation in the Digital Economy”爲目標建構六個電子化政府遠景，（註 11）將規劃重點置於加強提昇各級公務員資訊素養，使其能有效運用各項資訊科技改善作業流程、提升服務品質，然後將這些完善建構的政府 e 化資訊，透過單一便捷管道提供民眾運用，eCitizen 就是最好的例子（http://www.ecitizen.gov.sg/）。

參、知識探索

「知識探索」這個詞彙其實並不是十分清楚的，有些人將之視爲「資料探勘」，但是一般而言，資料挖掘是指由結構化資料找出資料的相關性，「相關性」爲何則必須再行定義。相對「資料探勘」而言，「文本探勘」指的是由非結構化的資料找出資料的相關性。吾人應該將「知識探索」看待爲「資料探勘」與「文本探勘」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，因爲從資料的應用面而言，使用者並不在乎結構化資料或是非結構化資料，只要能夠提供所需的資料即可。然而，從技術面而言，資料本身存在著多樣性，必須使用不同的技術處理不同類型的資訊，然後將之包裝，讓使用者不用擔心技術的複雜性，圖一是一個很好說明。



圖一　使用者與系統的分野

資料來源：作者繪製。

以下將分述結構化資料與非結構化資料，及相關的技術。

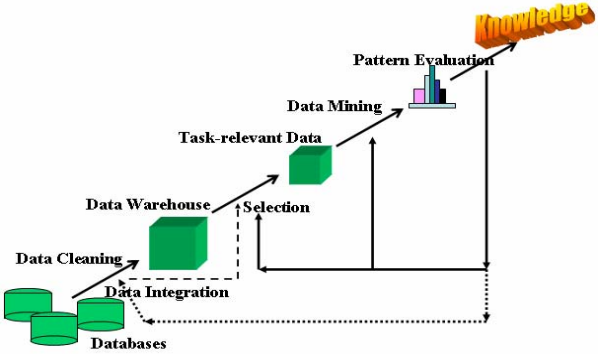
一、結構化的資料

從資料庫的觀點，所謂的結構化資料也就是依據「資料綱要」（Schema）建置的資料，換言之，即是機構爲了有效處理業務相關資料，使用 Entity-Relationship Model（ER Model）（註 12），建構了機構內部的 ER Diagram，然後運用資料庫管理系統（如微軟的 SQL Server，甲骨文的 Oracle，IBM 的 DB2）將 ER Diagram 轉換爲資料庫的 Schema，從此機構的各種資料皆有所屬，資料即具有結構性。如前述，結構化的資料是依據資料綱要，分門別類、按部就班被建置於資料庫，透過資料庫內建的檢索功能，使用者可以檢索相關的資料，既然如此，爲何還需要「資料探勘」呢？

ER Diagram 是系統分析師依據機構的業務需求，經過調查業務表單、訪問從業人員、觀察作業流程、審閱機構文件、體認機構文化，最後繪製而成，依據 ER Diagram 而建構的資料綱要反應的是「已知」的資料關係，眾多機構資料就依據資料綱要，陸續地被建置於資料庫系統。相對的，資料探勘是希望透過不斷累積的大量資料，找出資料「未知」的關係、知識與趨勢，這些不是資料綱要能夠提供的。當然，這種未知的關係、知識或趨勢，一旦被證實極爲有效，有助於機構決策人士做決策，亦有可能成爲新的資料綱要的一環。

企業界對於資料探勘的需求是非常迫切的，企業界感興趣的是消費者的消費行爲，消費市場的交叉分析，目標市場的確認等等導引企業獲利的重要資訊。所以亦有以 Business Intelligence 這個術語稱呼 Data Mining 的現象。

資料探勘有一套的程序，必須先做好資料的準備工作，圖二是一個典型的資料探勘的作業程序，包含去除雜訊、選擇資料、資料探勘、樣式評價。資料探勘的相關技術有：概念描述、關連、時間序列、回歸、歸類、分群、特異分析，以下分別敘述之。



圖二　資料探勘的程序

資料來源：Jiawei Han and Micheline Kamber , Data Mining: Concepts and Techniques, URL: http://www.cs.sfu.ca/~han/bk/1intro.ppt

（一）概念描述（Concept Description）

概念描述是概述、摘述、對比資料的特性，例如特定消費市場的特性，午間電視節目收視特性。透過這種方法，吾人可以知道關心或感興趣的資料（或對象）一般特性，從而訂出一些策略或施行辦法。

（二）資料關連（Data Association）

關連是試圖找出資料間的關連性，最有名的例子就是一項依據超級市場消費者同時採購物品的資料探勘研究，尿片與啤酒的關連性很高，很顯然，這種現象並非吾人可以事先知道的。但是資料探勘卻能揭示不爲人知的關係。

（三）時間序列（Time Series）與回歸分析（Regression Analysis）

時間序列與回歸分析用以找出資料的趨勢或演變，時間序列強調的是時間的歷線，回歸分析則不一定與時間相關。例如，企業界可利用過去 10 年第三季消費市場的消費情形，推估第一個第三季的消費情形。

（四）資料歸類（Data Classification）

歸類是找出一種歸類的方法，可以適當地將資料分到既定分類架構。例如依據脈搏、血壓等資料，將人分成高危險群、低危險群等分類，歸類之後，更重要的是可據以預測或推測。常用的歸類方法有決策樹（Decision Tree）、機器學習（Machine Learning）、類神經網路（Neural Network）。

（五）資料分群（Data Clustering）

歸類是在既有的分類架構上進行類別的判別，分群則是依據眾多資料的特性進行分群，無既有的分類架構。分群的指導原則是群內資料相似性高，群與群之間的相似性低。分群法使用完全連結（Complete Link）、單一連結（Single Link）、星狀連結（Star Link）、或線狀連結（String Link；Connected Components），以決定群內與群間的相似性。（註 13）

（六）特異分析（Outlier Analysis）

特異分析意圖找出與多數資料行爲模式不同的資料，這種分析方式對於企業界的應用，具有重要的意義。企業界運用特異分析技術可找出消費市場的罕見事件或特殊的消費行爲，因爲罕見事件可能是下一波企業獲利的重要啓示。

二、非結構化資料

所謂的非結構化資料是一般的文件，其內容包羅萬象，全賴文件作者行文遣字，無既定的規律，這類文件雖非「隱性知識」，但若欲以自動化的方法處理，卻也是相當大的挑戰。傳統上，從事文件研究的爲自然語言處理與計算語言學的研究人員，文件被視爲書面語（Written Language）的載體，這種書面語的實質內涵即是吾人認知的「文本」。另外不從語言角度入手，資訊檢索（Information Retrieval）的研究者將文件視爲是「裝載詞彙的袋子」（Bag of Words），嘗試用統計的方法，擷取文字的特徵以建構資訊檢索的模式。因此，自然語言處理與計算語言學的技術可說是深層處理（Deep Processing）的技術，而資訊檢索的技術是淺層處理（Shallow Processing）的技術。對於非結構化資料的文本探勘而言，自然應以深層處理爲宜。

文件有一定的結構，由上至下爲：文件，段落（Paragraph），句子（Sentence），子句（Clause），詞組（Phrase），詞（Word），字（Character）。從語意的角度，段落之上，還有論域（Discourse），論域是指具有特定主題的口語或書面語，論域可能由一個以上的段落構成。前述文件結構的每一層次都有相應的處理程序：段落層次有段落的辨識與論域的處理；句子層次（與子句層次）有句子的辨識，有述語參數結構（Predicate-Argument Structure）的處理，並進一步建構其邏輯正規形式（Logical Normal Form）；詞組有詞組的辨識；詞有詞的辨識（Segmentation），詞類標記（Tagging），詞幹處理（Stemming），詞義標記（Sense Tagging）。前述僅是進行文本探勘的前處理（註 14），真正進行文本探勘還必須伴隨著分類或推論等技術，以探勘未知的知識。以下說明相關的技術：

（一）文件結構的辨識

文件結構的辨識是處理非結構化資料的第一步，換言之，吾人必須給定文件每一成分所屬的結構標籤，如題名，作者，單位，機構，電子郵件，摘要，第一段落，其他段落，最後段落，圖片，表格等等結構標籤，至於到底有多少標籤，則由需求決定。文件

的類型當然會影響文件的結構，因此，使用所謂的適應性的系統（Adaptive System）是較爲可型的作法。（註 15）

（二）文件論域的辨識

一篇文件可能有數個論域，一個論域可能有數個段落，辨識文件的段落後，必須辨識哪些段落是屬於同一個論域，這就是所謂的文件論域的辨識。論域具有核心的主題，非結構化資料的文本探勘或多或少都牽涉到語意層面的處理，因此，確認該核心主題是重要的工作。有關文件論域的辨識可參閱註 16 所列的文獻。

（三）句子的辨識

句子的辨識是之後詞彙處理的基礎工作。句子辨識不如想像中的簡單，Palmer 與 Hearst 使用很複雜的處理程序以辨識英文句子，正確率爲 98.5%。（註 17）中文句子的特性又與英文不同，中文使用短句很頻繁，通常以逗號「，」結尾。句子辨識完成後，需以句子爲單位建立剖析樹（Parsing Tree）（註 18），隨著語法理論的不同，剖析樹有眾多形式，但是建構剖析樹之前，必須先處理詞組與詞彙。一旦建構完剖析樹，才能著手建構邏輯正規形式，而這是一切文本探勘中推論的基礎。

（四）詞組的辨識

詞組在英文中經常出現，通常一個相同動詞與不同的介副詞組合，就有不同的意義，這造成詞組的辨識是不可或缺的步驟。相對而言，中文詞組現象較少，倒是字組較多，如「一元復始」與「三陽開泰」。無論是詞組或是字組，其意義通常是個別的詞義或字義組合而兼有變化。目前使用多連（N-Gram），搭配統計模式以處理詞組或字組。（註 19）如果有完整的片語詞典，亦可採用詞典爲本（Dictionary-Based）的作法。

（五）詞彙的處理

中、日、韓等東方語系語言與西方的拉丁語系語言不同，東方語言有分詞（Word Segmentation）的問題，亦即東方語言詞彙間無西方語言詞彙間的空白標記。詞彙辨識完成，就是給定詞彙的詞類標記與詞義標記，另外，英文有詞幹處理的工作，中文則無。相關技術的詳細說明，可參閱註 15 所列參考文獻。

（六）文本歸類

文本歸類與資料歸類相似，皆是希望將文本給定一個既定的分類標籤。這在新聞機構的應用特別明顯，一般的新聞機構都有其分類體系，如政治、財經、體育、娛樂、文學、旅遊等分類標籤。目前已經有許多的歸類辦法，只要建構待歸類文本的特徵，再具之給定最適當的分類標籤即可。

（七）文本分群

將文件依據各自的特性分群，而無既定的分類體系，通常分群的作法是採用階層式分群法（Hierarchical Clustering），最後所形成的群是由相似性的門檻值（Threshold）決定的，而決定門檻值並不容易，除了經驗與實作外，還沒有系統化的方式。分群的另一問題是如何給定分群的標籤或是建議分群的標籤，以揭示各分群的特性，這是文本分群最難的工作。

（八）自動推論

自動推論是人工智慧研究的一環，其本身就是很重要的課題，自動推論需要一套推論機制（Reasoning Mechanism），接受各文本擷取的資料，並作爲推論的前提，運用推論引擎，以得到推論的結果。Swanson 做過一系列運用醫學文獻進行文本探勘的研究，事後亦證明這樣的作法是有效的。（註 20）

（九）主題偵測與追蹤（Topic Detection and Tracking，TDT）

TDT 嘗試由一群依時間排序的文件，偵測新事件（新主題）的出現，並追蹤該事件，換言之，TDT 將這群文件依照不同的主題，將相同主題的文件依時序串連，建構時間序列，如此可探知事件的發源、演變、以迄終止。有關 TDT 的研究可參閱相關的參考文獻。（註 21）

簡單描述結構化資料之資料探勘與非結構化資料之文本探勘，下一節將討論政府資訊的知識探索。

肆、政府資訊的知識探索

現代化的政府必須以服務民眾爲宗旨，爲人民謀福利，政府一切的施政作爲必須滿足人民的需求。現代化的政府另一項重要的特徵便是資訊的公開與流通，我國先後

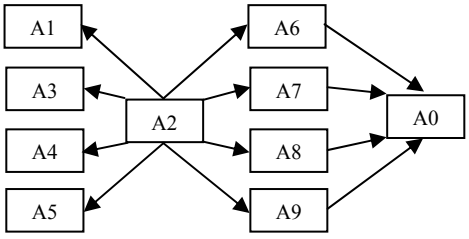
完成制訂「行政程序法」及其子法「行政資訊公開辦法」，以保障人民權益、提高行政效能與建構公正、公平與公開的程序文化。而爲進一步強化政府執政的透明度，符合民眾知的需求，「政府資訊公開法（草案）」並業已在立法院審議中。政府資訊公開的精神在於便利人民共享及公平利用政府資訊，增進人民對公共事務之瞭解、信賴及監督。（註 22）

檔案管理局爲檔案之中央主管機關，負責檔案政策、法規及管理制度之規劃，檔案目錄之彙整、國家檔案的徵集、移轉、整理、典藏、開放應用等作業之規劃與執行，等工作。（註 23）檔案管理局可謂是全國政府資訊的集散中心，爲了符合資訊公開法的精神，檔案管理局自應積極規劃檔案開放應用，一則服務人民，二則利於政府機關業務的推動。

檔案管理局目前建置有「機關檔案管理資訊網」，提供各級機關經由網際網路自動化管理及彙整檔案資訊，然後再透過「全國檔案目錄查詢網」，提供各界查詢。全國各機關則必須定期將所屬檔案的目錄資料（詮釋資料，Metadata）上載至「機關檔案管理資訊網」，爲了讓一般民眾使用這些資料，檔案管理局亦建立「全國檔案目錄查詢網」。全國檔案目錄資料成長極爲快速，目前已有 18,000,000 餘筆，其中 70%爲民國 91 年的資料，其餘 30%爲回溯資料，依此推估，一年之目錄資料量約爲 14,400,000 筆左右。然而其中有爲數眾多的「存查」資料，換言之，一年 14,400,000 筆的資料，其中有相當多的資料其實是不同機關處理同一機關的來文，實屬相同的資料。因此，知識探索的第一步應是聚集相同檔案的目錄資料。由於不同的機關對於來文會有不同的處理方式，因此，各機關的相同來文號檔案的目錄資料會有些許不同，這也反應各機關對於相同的政策或命令，可能的不同因應方式，所以分析同一群集檔案的異同，可以從橫的方向瞭解政策或命令的執行情形。

第二個可能的應用是建構檔案資料的時間序列。由時間的遞變觀察檔案資料的關係，可以建構歷史觀，亦可追蹤政策命令的執行情形。文獻計量學研究曾探討「依據文獻出版時期，做層次性描述的圖像，通稱爲時序圖」。（註 24）將文獻出版日期類比於公文檔案發文日期，亦可建構類似的時序圖。

第三個應用是建立檔案資料之相互參照。可仿效引用文獻分析的理論，藉由研究大量政府資料的相互參照， 可建構 Hub 與 Authority，請參見圖三，A0 爲 Hub ，A2 爲 Authority。（註 25）在這樣的應用環境下，hub 代表的意涵是：該機關的各項措施會參照很多相關公文、政策、或命令，顯示其發文皆有所本；authority 代表的意涵則是：該機關是主要的政策與命令的制訂者或規劃者。



圖三　Hub 與 Authority

資料來源：作者繪製。

第四個應用是相關檔案資料的群集。相關檔案資料的群集與相同檔案資料的群集是不相同的，這也是困難點所在，進行這項應用的知識探索時，必須定義關係爲何。例如，某項應用定義其相關爲「政策相關」，試圖找出各機關推動「政府出版品統一編號」的措施。這種動態即時的分群，非常接近傳統 search by surface form 的資訊檢索，但是卻不盡然相同，這種應用本質上主要是 search by concept。

第五個應用是抽取重要的人物與事件。從歷史的觀點，政府資訊中重要的人物與事件經常是具有舉足輕重的影響力，更重要是人物、時間與事件間交織的關係。在資訊擷取的研究領域，使用特定的樣版（Template）以擷取需要的資訊，這樣的樣版通常是環繞在人、事、時、地、物、關係等屬性，我們也可以將樣版視爲浮動的詮釋資料格式，這種格式是隨著應用而變動，而不是欄位固定的、權威控制的、靜態的詮釋資料格式。（註 26）

第六個可能的應用是建立統計共現索

引典（Statistical Co-occurrence Thesaurus）。政府機關其實應該建立專用的索引典，才能讓檔案目錄資料的著錄有一定的一致性。然而，這項工作卻是極耗費人力與物力的，若能夠運用文本探勘的技術，首先建立一個統計式的共現索引典，再據以編制更爲詳盡且詞彙關係完善的索引典，可以降低大量的人力與經費。

第七個應用是比較特殊的，這項應用與我國目前政府各機關分類號的設計有關，檔案管理局收錄全國各機關檔案目錄資料，各檔案皆有一檔案編號（檔號），檔號係由檔案年度號、分類號、案次號、卷次號及目次號依序組成之一組編號，其中分類號多數是由各機關依據業務特性自行制訂，例如：國立高雄師範大學的分類號如表一所示，秘書類又細分爲表二所示；而台北市立師範學院分類表如表三所示，秘書類又細分爲表四所示。由前述四表可以發現，性質如此相近之師範院校，其分類體系與分類號的位數亦不相同，這對於檔案的開放應用造成一定的影響，根本解決之道是制訂統一的分類表，全國各機關據以給定分類號。短期的解決的辦法，可以運用知識探索的技術，建構各機關分類號的對應。

表一　國立高雄師範大學分類號

類	名稱
0	研究發展類
1	教務類
2	訓導類
3	總務類
4	人事類
5	會計類
6	實習輔導類
7	學術類
8	進修推廣類
9	秘書類

資料來源：http://www.nknu.edu.tw/~tracylee/dochtml/doc_1_2.htm

一個可能的辦法是應用機器學習或是統計分析的技術，使用大量已給定檔號的檔案目錄資料，建構每一分類號的特徵，再使用特徵比對的方法，找出相近的分類號，如此可建構一個廣泛的分類號對照體系（Mapping Framework），稱之爲廣泛是因爲這個對照體系並非一個絕對的、一對一的對照表，而是一個可能的、需要系統者再次確認的對照表。接著，便可用該對照體系，處理各機關分類號的對照關係，有助於建構檔案的關聯性，並協助建構統一的分類體系。

表二　國立高雄師範大學分類號-秘書類

類	綱	目	節	項	類目名稱
9					秘書類
9	1				會議
9	1	1			會議協調
9	1	2			教育會議
9	2				視導
9	2	1			教育視導
9	2	2			軍訓視導
9	3				研考
9	3	1			管制考核
9	4				環保
9	4	1			環保業務
9	5				校史
9	5	1			校史
9	5	2			交接
9	9				其他
9	9	1			全校密件公文
9	9	2			高師大教育學術基金及校務基金
9	9	3			秘書室綜合業務

資料來源：同表一。

表三　台北市立師範學院分類號

類	名稱
01	秘書類
02	教務類
03	學生事務類
04	總務類
05	人事室
06	會計室
07	軍訓類
08	圖書類
09	進修暨推廣類
10	實習輔導類
20	研究教學與中心類
99	消費合作社類

資料來源：http://www.tmtc.edu.tw/~document1/document/document/form/form2.htm

表四　台北市立師範學院分類號-秘書類

單位	分類號（類、綱、目、節）	業務歸類
秘書類	01-00-01-00	綜合性業務
	01-00-02-00	重要決議及法令規章
	01-00-03-00	移交清冊及相關表件
	01-00-04-00	督導與評鑑
	01-00-05-00	校務中長程發展
	01-00-06-00	國際學術合作交流
	01-00-99-00	其他

資料來源：同表三。

伍、結論

基於政府資訊公開法的精神，人民有權取得無安全顧慮的政府資訊或文件，檔案管理局作爲檔案中央主管機關，負責檔案政策、法規及管理制度之規劃，檔案目錄之彙整、國家檔案的徵集、移轉、整理、典藏、開放應用，除了服務人民外，也必須服務全國各機關檔案從業人員，因此，檔案管理局要發展既利於人民的資訊應用，亦利於公務人員順利推展相關業務的資訊應用。

學術界對於知識探索的研究與企業界對於知識探索的應用，已有一段日子了，也有一定的經驗，成功的應用案例亦不少見，應用知識探索於政府資訊是一項非常可行的作法。本文簡短地描述知識探索的內涵，分別就結構化資料與非結構化資料，說明知識探索的相關技術。然而，本文並未觸及知識探索之前資料的準備工作，這部分的工作事實上是很重要的，但是卻與資料的特性有關，無法作一個簡短而實際的說明。

檔案管理局已成爲全國政府資訊的集散中心，無疑會累積越來越多的資料，對於從事政府政策或政府資訊的研究者而言，檔案管理局擁有的資料是隱含各種知識的寶庫。本文以全國檔案目錄資料爲例，討論了幾項知識探索於政府資訊的應用，可以從時間歷線、交互參照、關連相關、或人物事件，探究檔案間可能的各種關係或未知的知識，這些淺顯的例子，提供檔案從業人員參考，並期能收拋磚引玉之效。

註釋

註 1：這項調查主要係針對各國政府網站的服務內容與功能進行評估，包括網站內容的豐富完整程度、線上申辦服務項目、對電子郵遞意見的處理回復等。報告全文請參考其網 <http://www.insidepolitics.org> (Feb. 24, 2003)。

註 2：陳瑩芳，美國政府資訊指引服務之研究，國立台灣大學圖書資訊學碩士論文，民 88 年，13-17。

註 3：San Diego Supercomputer Center, SDSC, <http://www.sdsc.edu/> (Feb. 24, 2003).

註 4：“Persistent Digital Archives: A Knowledge-Based Approach,” Online 4, no. 25 (Feb., 2003), <http://www.npaci.edu/online/v4.25/persistent-archives.html> (Feb. 23, 2003).

註 5：同註 4。

註 6：Brian Robinson, “Digging for Data Treasure,” FCW.COM (Oct. 2, 2002) <http://www.fcw.com/civic/articles/2000/oct/civ-tech-10-00.asp> (Feb. 23, 2003).

註 7：José Luis Ambite, Yigal Arens, Eduard Hovy, Judith Klavans, and Andrew Philpot, “Scalable Access and Integration of Statistical Data for Digital Government,” in Proceedings of AFCEA Federal Database Colloquium and Exposition, San Diego, California, August 28-30, 2001. <http://www.cs.columbia.edu/digigov/ambite-afcea-database2001.doc> (Feb. 23, 2003).

註 8：Dennis M. Lormel, “Statement for the Record on Technology, Terrorism and Government Information,” (July 9, 2002) <http://www.fbi.gov/congress/congress02/idtheft.htm> (Feb. 23, 2003).

註 9：Institute of Standards and Technology, NIST (Feb. 21, 2003), <http://www.nist.gov/> (Feb. 23, 2003).

註 10：各機關提出的計畫共有 121 個，各項計畫內容及摘要，請參考其網站 <http://www.gol-ged.gc.ca/index_e.a

sp> (Feb. 24, 2003)。

註 11： 這六項計畫是：
1) Knowledge-Based Workplace
2) Electronic Services
3) Delivery Technology Experimentation
4) Operational Efficiency Improvement
5) Adaptive and Robust Infocomm Infrastructure
6) Infocomm Education National

註 12： R. Elmasri and S. Navathe, Fundamentals of Database Systems. (Redwood City, CA: The Benjamin Cummings Publishing Company Inc., 2000), 39-68

註 13： G. Kowalski and M. Maybury, Information Storage and Retrieval Systems, (Norwell, MA: Kluwer Academic Publishers, 2000), 147-150.

註 14： 不同的文本探勘目的需要的前處理工作並不相同，簡單的文本探勘目標可能比這些前處理工作還簡單，如基本的文件分類。

註 15： 陳光華，「資訊的組織與擷取」，圖書館學刊（國立臺灣大學）12 期（民 86 年 12 月）：134。

註 16： 陳光華，陳信希，「文件內容之分析-語料庫爲本的模型」，圖書館學刊（國立臺灣大學）11 期（民 85 年 12 月）：95-112。

Kuang-hua Chen, “Topic Identification in Discourse,” in Proceedings of the 7th Conference of the European Chapter of Association for Computational Linguistics (EACL95), (San Francisco, CA: Morgan Kaufmann Publishers, 1995), 267-271.

M. Hearst and C. Plaunt, “Subtopic Structuring for Full-Length Document Access,” in Proceedings of the 6th International ACM SIGIR Conference on Research and Development on Information Retrieval, (Baltimore, MD: ACM Order Department, 1993), 59-68.

註 17： D. Palmer and M. Hearst “Adaptive Sentence Boundary Disambiguation,” in Proceedings of the Conference on Applied Natural Language Processing, Stuttgart, Germany, Oct 1994, (San Francisco, CA: Morgan Kaufmann Publishers, 1994), 78-83.

註 18： 陳光華，「資訊檢索查詢之自然語言處理」，中國圖書館學會會報 57 期（民國 85 年 12 月）：141-153。

註 19： Yuen-Hsien Tseng, “Multilingual Keyword Extraction for Term Suggestion,” in Proceedings of the 6th International ACM SIGIR Conference on Research and Development on Information Retrieval, (New York: ACM Order Department, 1998), 377-378.

註 20： Don R. Swanson and N. R. Smalheiser, “Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease,” Neuroscience Research Communications, 15(1994):1-9.

註 21： J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. “Topic detection and tracking pilot study: Final report,” in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, Feb. 8-11, 1998 <http://www.nist.gov/speech/publications/darpa98/pdf/tdt2040.pdf > (Feb. 24, 2003).

註 22： 資訊公開法（草案）第一章總則第一條。

註 23： 檔案管理局組織條例第二條。

註 24： 何光國，文獻計量學 (台北市：三民，民 83)，219。

註 25： J. Kleinberg, "Authoritative sources in a hyperlinked environment," in Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms, 1998, <http://www.cs.cornell.edu/home/kleinber/auth.ps> (Feb. 23, 2003).

註 26： 陳光華，「資訊檢索技術之核心」，大學圖書館 3:4（民 88 年 1 月）：17-28。

參考書目

[1] 何光國。文獻計量學。臺北市：三民，民國 83 年。

[2] 陳光華，陳信希。「文件內容之分析-語料庫爲本的模型」。圖書館學刊（國立臺灣大學）11 期（民 85 年 12 月）：95-112。

[3] 陳光華。「資訊的組織與擷取」。圖書館學刊（國立臺灣大學）12 期（民 86 年 12 月）：121-147。

[4] 陳光華。「資訊檢索技術之核心」。大學圖書館 3:4（民 88 年 1 月）：17-28。

[5] 陳光華。「資訊檢索查詢之自然語言處理」。中國圖書館學會會報 57 期（民 85 年 12 月），141-153。

[6] 陳瑩芳。美國政府資訊指引服務之研究。碩士論文，國立台灣大學圖書資訊學研究所。民 88 年 12 月。

[7] "Persistent Digital Archives: A Knowledge-Based Approach." Online 4, no. 25 (Feb., 2003). <http://www.npaci.edu/online/v4.25/persistent-archives.html> (Feb. 23, 2003).

[8] Allan, J. J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. "Topic detection and tracking pilot study: Final report," In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, Feb. 8-11, 1998. <http://www.nist.gov/speech/publications/darpa98/pdf/tdt2040.pdf> (Feb. 24, 2003).

[9] Ambite, José Luis, Yigal Arens, Eduard Hovy, Judith Klavans, and Andrew Philpot. "Scalable Access and Integration of Statistical Data for Digital Government." In Proceedings of AFCEA Federal Database Colloquium and Exposition in San Diego, California, August 28-30, 2001. http://www.cs.columbia.edu/digigov/ambite-afcea-database2001.doc (Feb. 23, 2003).

[10] Chen, Kuang-hua. "Topic Identification in Discourse." In Proceedings of the 7th Conference of the European Chapter of Association for Computational Linguistics (EACL95), Dublin, Ireland, 1995, San Francisco, CA: Morgan Kaufmann Publishers, 1995, 267-271.

[11] Elmasri, R. and S. Navathe. Fundamentals of Database Systems. (Redwood City, CA: The Benjamin Cummings Publishing Company Inc., 2000), 39-68

[12] Hearst, M. and C. Plaunt. "Subtopic Structuring for Full-Length Document Access." In Proceedings of the 6th International ACM SIGIR Conference on Research and Development on Information Retrieval, New York: ACM Order Department, 1993, 59-68.

[13] Kleinberg, J. "Authoritative sources in a hyperlinked environment." In Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms, 1998. <http://www.cs.cornell.edu/home/kleinber/auth.ps> (Feb. 23, 2003).

[14] Kowalski, G. and M. Maybury. Information Storage and Retrieval Systems, (Norwell, MA: Kluwer Academic Publishers, 2000), 147-150.

[15] Lormel, Dennis M. "Statement for the Record on Technology, Terrorism and Government Information." (July 9, 2002). <http://www.fbi.gov/congress/congress02/idtheft.htm> (Feb. 23, 2003).

[16] Palmer, D. and M. Hearst. "Adaptive Sentence Boundary Disambiguation." In Proceedings of the Conference on Applied Natural Language Processing, Stuttgart, Germany, Oct 1994, San Francisco, CA: Morgan Kaufmann Publishers, 1994, 78-83.

[17] Robinson, Brian. "Digging for Data Treasure." FCW.COM (Oct. 2, 2002). <http://www.fcw.com/civic/articles/2000/oct/civ-tech-10-00.asp> (Feb. 23, 2003).

[18] Tseng, Yuen-Hsien. "Multilingual Keyword Extraction for Term Suggestion." In Proceedings of the 6th International ACM SIGIR Conference on Research and Development on Informaiton Retrieval, 1998, New York: ACM Order Department, 1998, 377-378.