# 專利文字之知識探勘：技術與挑戰
# Knowledge Discovery in Patent Texts: Techniques and Challenges

曾元顯 輔仁大學圖書資訊學系 教授
Yuen-Hsien Tseng
Professor
Dept. of Library & Information Science, Fu Jen Catholic University
tseng@lins.fju.edu.tw
現代資訊組織與檢索研討會，2004/11/19, 頁 111-123.

## 摘要：

專利文獻包含重要的研究成果，其揭露的技術方法也受到法律的保護，因此專利資訊的分析與運用，已逐漸成爲企業、機構持續生存、永續發展的重要措施。然而專利文件數量眾多、篇幅龐大、技術與法律用語並存、用詞冷僻，人工閱讀、分析耗時耗力，極需自動化的工具輔助分析。本文介紹文字探勘技術運用於專利分析的現況與挑戰。除了說明文字探勘的意涵，介紹相關的計畫與研究，文中也描述筆者自行發展的初步技術與方法，運用於美國專利的文字分析情形。目前專利分析仍有很多難題，需要領域專家、法律專才、資訊技術人員、專利分析人員、與嫻熟資訊組織與主題分析的圖書專業人員，共同合作，一起規劃，才比較容易設計出較佳的專利分析方法與工具。
**關鍵詞：**專利資訊、文字探勘、資訊組織、主題分析、摘要與歸類

## Abstract

Patent documents contain important research results. The technologies they reveal are protected by laws. They have been important information resources in nowadays knowledge economy activities. The analysis of patent documents and the use of the analyzed results have become an important issue for an enterprise to survive. However, patent documents are hard to read and analyze. Automatic tools for assisting patent analysis are in great demand. This article describes the development and challenges of patent analysis based on text mining techniques. Besides literature review, this article also reports our techniques. We conclude that there are still many challenges in current patent analysis. To develop useful tools, it would require the cooperation among domain experts, lawyers, technicians, and librarians.

**Keywords:** Patent Information, Text Mining, Information Organization, Subject Analysis, Summarization and Clustering

## 壹、前言

近年來由於國內經濟發展遲緩、產業結構變動、失業率攀高等問題，政府及學術界提出「知識經濟」規劃，做爲振興經濟的方案。知識經濟即是以知識爲基礎的經濟，然而知識不易定義、難以量化。在各種各類的知識當中，對個人、機構、產業、國家最具經濟價值、最具體而可評量者，非「專利」莫屬。專利的擁有，爲一種頗具威力的權力，是法律認可的獨占權之一，消極方面可用以代表個人、機構或國家的技術先進，積極方面可以保護、交易創新的知識，促進經濟與技術的進一步發展。

在知識經濟時代，鑑於專利對國家經濟、產業發展與技術研發扮演越來越重要的角色，專利資訊的分析與運用便成爲企業、機構知識管理的一環。事實上，根據世界智慧財產權組織（World Intellectual Property Organization, WIPO）的報告，專利文獻包含全世界每年 90%-95%的研發成果，有效運用專利資訊可縮短研發時間 60%，節省研究經費高達 40% [1]。

近年來，國內開始注重專利資訊的分析工作，相關單位逐步投入人才的培訓並進行專利分析的工作 [2-3]。其分析工作的主要步驟，筆者歸納大致如下：

一、選題：確立研究主題之專利分析範圍及目的。

二、篩選：選定專利資料庫、建立檢索策略、下載相關專利文獻。

三、轉換：切割資料、分欄擷取、正規化、資料整備，依結構化欄位資訊進行量化分析與圖表製作。

四、摘要：依照目的、方法、效用等分項製作專利閱讀分析摘要表，進行專利文字的內容分析。

五、歸類：依主題對專利領域做歸類（clustering）、對專利文件進行分類（classification）。

六、呈現：以多重分類表交叉分析專利文件，製作技術功效分佈矩陣或專利地圖。

七、解讀：判讀量化圖表與專利地圖，進行技術分析與趨勢預測。

上述「專利地圖」，在不同的單位有些許不同的認定，有的將結構化欄位資訊的統計量化圖表視爲專利地圖的一部份，有的則專門指內容分析後的主題交叉分析矩陣，才視爲專利地圖，而稱其他數據統計圖表爲專利圖（Patent Graphs）。依照劉尚志的描述，任何將專利資料轉換成技術競爭情報的分析結果，都稱爲專利地圖（Patent Maps）[4]。綜合而言，這些專利資訊的分析結果，可以用來評估與預測技術發展、規劃研發或技術發展項目、避免誤觸專利權而浪費研發資源、掌握企業發展動向及市場需求[5-7]。

然而專利文獻與一般學術或政府文獻有相當大的不同，其特色如下：

一、資訊源分散（各國專利局），使用者難以一次便蒐集到完整資訊；

二、查全導向（recall-oriented），某些應用下漏檢重要的專利，將來可能會付出很大代價；

三、專利數量眾多、篇幅龐大，閱讀、分析耗時耗力；

四、結構化、非結構化資訊合一；

五、專業領域、法律用語並存，閱讀、分析困難；

六、相同事物概念，常有不同用語描述，用以規避雷同、播種侵權地雷；

七、關聯性、衍生關係複雜；

有鑑於此，專利的分析人員常需具備資訊檢索、專業技術與法律規範的知識。然而，專利分析人員培養不易、經驗傳承困難、專利內容分析耗時費力。韓國智財局擬定 5 年製作出 120 個專利地圖[8]，顯示此項任務之工作繁重。因此，若要使專利分析的工作，普及到各個產業領域，自動化的電腦輔助分析，實有必要。

近年來文字資訊探勘的研究，在資訊檢索、自動摘要、自動分類、自動歸類、關聯分析上都頗有進展[9]，若能針對專利文件好好運用，對專利分析的工作，將有莫大的幫助。以日本國立情報學研究所( National Institutes of Informatics, NII )為例，在其 NTCIR( NII Test Collections for Information Retrieval )的計劃下，便舉辦專利資訊檢索（Patent Retrieval）的評比，以驗證目前資訊檢索在專利技術調查（technology survey）與前案檢索（invalidity search，搜尋前案使目前競爭者的專利失效），以及專利分析及專利地圖製作的各項技術[10]。

## 貳、文字知識探勘

**知識探勘**（knowledge discovery, KD）是擷取隱晦、有用、未被發掘、有潛在價值的規則、資訊或知識的一種過程 [11]。在實務上，此過程需要運用資訊組織與分析等探勘技術，透過與使用者的互動，來反覆探索資料庫或文件庫，以發現新的、有趣的訊息或規律，再經由人工解讀結果，讓發現的規律訊息變成有用的資訊或知識。

知識探勘的步驟大致分為：**資料蒐集**、**資料清理**、**資料轉換**、**探勘技術運用**、**結果呈現與解讀**。這些步驟與前面整理的專利分析步驟雷同，顯示專利分析本身即在進行知識探勘。那麼運用知識探勘的方法、技術與工具，便非常合理。

根據資料特性的不同，知識探勘可分為**資料探勘**（data mining, DM）與**文字探勘**（text mining, TM）。前者處理結構化（structured）資料，即每筆資料有共同欄位可記錄於資料庫者，而後者處理非結構化（unstructured）資料，即每筆資料沒有共通的結構性可言，經常為長短不一、記載訊息的自由文字。

由於資料特性的不同，資料探勘（DM）與文字探勘（TM）在每個步驟上幾乎都有所差異。以探勘技術而言，主要有：關聯分析（association）、分類（classification）、歸類（clustering）、摘要（summarization）、概略化（generation）、預測（prediction）、序列分析（sequence analysis）等，這些技術名稱在 DM 與 TM 中雖然相同，目的也一樣，但是技術細節卻不一樣。TM 運用的技術，幾乎都跟詞彙的頻率與出現篇數有關，但這兩項資訊在 DM 中極少用到。此外，DM 與 TM 的應用方向也大不相同。DM 主要運用於大型資料庫上，提供資料庫管理

系統額外的資料分析與統計功能；而 TM 主要運用在大量的文件庫上，供作資訊搜尋、訊息過濾、事件關聯、趨勢預測、犯罪分析、案例追蹤、知識萃取、知識管理、決策輔助等之用。DM 在傳統資料庫的運用上已算相當成熟，TM 最近才在各領域受到重視，如[12]，本文針對 TM 在專利文字上的運用，做初步的探討。

圖書館學在資訊組織與主題分析的理論與實務上，有長期而重要的貢獻。而資訊組織與主題分析採用的步驟與方法，與文字探勘有很多不謀而合之處。就專利文件的分析上，如前所述，其運用到資訊檢索、權威控制、內容摘要、歸類與分類等，幾乎都是圖書館學探討的內涵。因此，文字探勘跟資訊組織與主題分析，幾乎可說是同義詞，只是文字探勘更重視流程的連貫性與自動化技術的運用。

## 參、相關計畫與研究

近幾年國內外開始重視專利分析與專利地圖製作的工作，如前所述日本、韓國、新加坡以及國內國科會科資中心等技術資訊中心，都陸續提供這樣的服務[13]。這些單位主要以人力進行專利分析，還沒看到專利內容主題自動分析工具的運用。國科會科資中心運用過國內的 Patent Guider [14]工具，而韓國科資中心（KISTI）運用過其本國開發的 PIAS 系統[8]，這兩系統屬於「專利圖」的製作工具，亦即主要功能爲提供結構化欄位資料的統計與製圖。在商業化產品方面，有 Knowlegist 等產品協助分析專利內文[15]，透過自然語言剖析與自動摘要技術，擷取並摘要出單篇專利的應用方向、主要目標、專利方法、特色功效、重要主題等，方便使用者的快速深度分析。惟其成效水準，由於還未找到客觀公正的評估報告，目前難以得知。

眞正對非結構化專利文字進行內容分析與專利地圖自動建構的研究活動（research activities），目前所知，僅有日本 NTCIR 的專利檢索評比。NTCIR 3 舉辦第一次的專利檢索評比，其評比項目有二：一是給定一篇新聞報導（當作查詢來源），檢索相關專利以進行技術調查（technical survey）；二是參與者自行設計、提出的任意項目，讓主辦單位可以調查、瞭解專利相關的檢索需求與專利處理任務。此外，跨語言的專利檢索也有需要，但因爲是第一次進行評比，爲鼓勵更多人參與，先從最簡單的技術調查任務開始。此任務假設公司經理看到一篇有趣的新聞報導，想要瞭解相關的專利技術，因此將此新聞剪下順便加上註解，交給檢索者進行專利檢索。主辦單位準備二年的日文專利全文、五年的日文專利摘要以及五年的日本專利的英文摘要，檢索題目同時給出日、中、韓、英四種語文，以利跨語言專利檢索。NTCIR 3 的專利檢索總共有 8 隊參與，有 2 隊進行跨語言專利檢索的評估、其他隊伍則進行跨資料庫（用新聞查專利）、相關回饋、長詞索引、摘要檢索等各種專利檢索策略的評估與比較。至於參賽者自行提出、設計的自由項目部份，只有兩隊參加，一隊試驗各種文句對列的規則擷取方式，另一隊則針對專利宣告的部份進行結構分析的實驗，以便提升宣告的可讀性[16]。

專利檢索的目的有數種：技術調查、前案檢索、銷售/購買專利等，相同的

主題但不同的目的會有不同的相關專利，因此可能需要不同的檢索模式與技巧。在某些情況下，光檢索出相關的專利還不夠，需要進一步的分析如產生專利地圖以便釐清各個專利間的關係。因此 NTCIR 4 的專利檢索評比，其項目從技術調查改爲前案檢索及專利地圖製作。

在學術研究方面，國內相關研究稀少，大都爲法律、企管、經濟領域針對其關心的議題進行專利方面的研究[17-22]。國外則有 SIGIR 2000 的專利檢索工作坊[23]，可算是 2002 年 NTCIR 3 專利檢索評比的會前會。另外還有 ACL 2003 的專利文件處理的工作坊[24]，其內容部份爲 NTCIR 3 專利評比的結果報告，部份則爲自然語言處理技術運用於專利文件自動分析的計畫說明與目前成果。

在專利內文的結構分析方面，Shinmori 等人針對日文專利的宣告（claim）部份，做自然語言的處理，將長句斷成短句，以增加其可讀性[25]。其利用構詞分析器（morphological analyzer）、語法分析器（lexical analyzer）、文法剖析器（grammatical analyzer）、修辭結構分析器（RST tool）以及針對專利文件自行整裡的線索詞（cue phrase）、57 個前後文無關的文法（Context-free grammars）等資源進行專利宣告的分析與處理。在 100 筆宣告中，人工評估其文句結構分析的正確率達 80.85%，顯示其方法具有相當的準確性。Sheremetyeva 也報告自然語言處理技術運用於美國專利宣告的計畫，此計畫將繁複難懂的法律宣告文句轉換成數個短句，便利後續的機器翻譯與人工閱讀[26]。雖然其初步的成效良好，但和 Shinmori 等人都用到相當多不易自動取得的資源，如辭典、文法規則等。

在專利的主題分析方面，Lamirel 等人運用類神經網路的自我組織圖（Self-Organization Map, SOM）技術來自動偵測專利的主題[27]。他們從美國專利的摘要中，擷取出關於方法（use）、效用（advantage）、標題（title）等內容的文字片段，再分別運用 SOM 技術進行無監督式的自我學習與自動組織，將相似的主題詞彙與文件歸類在一起，並以二維圖示顯示出來，提供專業人員進一步的分析與解讀。其評估結果顯示，按段落分別建出的自我組織圖，比不分主題段落而用全文產生的自我組織圖，更能提供有用的分析資訊。

在專利的主題分類方面，新聞文件自動分類的研究很多，但針對專利文件做分類的研究非常少。Fall 等人[28]以美國專利文件建立一份測試集，包含 46324 篇訓練文件，28926 篇測試文件，共 114 個類別，451 個子類別。由於專利全文長度差異極大，他們以四種方式擷取內文來代表原文件：一、標題；二、宣告（Claim）；三、標題、發明人（Inventor）、應用領域（Application）、摘要（Abstract）、描述（Description）等五部份，每個部份的前 300 個字（words）（筆者註：其中描述的前 300 個字相當於敘述發明背景（Background of the Invention）的部份）；四、標題、發明人、應用領域、摘要。其運用 Naïve Bayes、KNN、SVM 等分類器做實驗，但不論哪種分類器，都發現將專利全文拿來做分類的效果最差，而取前 300 字內容的效果最好，且 SVM 分類器效果最好，但計算最耗時，KNN 次之，僅差 SVM 二到三個百分點。

## 肆、專利文字之探勘技術與挑戰

介紹過國內外相關的計畫與研究後，本節簡略描述筆者進行的專利文字探勘研究，透過範例的說明或展示，提供大家參考，希望收到拋磚引玉的效果，期能有更多國內的研究投入，提升專利分析的深度與成效。

筆者長期進行文件索引、檢索、關聯、分類、歸類、摘要等自動化資訊組織與主題分析的研究[29-38]，根據針對文字進行知識探勘的經驗，發現專利文件非常需要這些技術的應用，但也發現其中的問題與新的挑戰。茲列舉數項如下：

一、專利全文前置處理：欄位剖析與資料整備

美國專利全文為 HTML 格式，內容記載專利號（Patent Number）、申請日期（Filing Date）、公告日期（Date of Patent）、發明人（Inventor）、申請人（Assignee）、美國分類號（UPC）、國際分類號（IPC）、引用參考資料（References Cited）、相關的美國專利（U.S. Patent Documents）、專利名稱（Title）、專利摘要（Abstract）、專利宣告（Claim）、專利說明（Description）等項目。其中專利說明以文字詳細描述該發明創作，通常包含下列項目：發明領域（Field of the Invention）、發明背景（Background of the Invention）、發明摘要（Summary of the Invention）、圖式簡述（Brief Description of the Drawings）、發明細節（Detailed Description of the Invention）等。這些資料都必須從半結構化的 HTML 檔裡剖析出來，尤其專利宣告與說明的部份，更必須從自由文字中，分段擷取如發明領域、發明摘要等段落，瞭解其待解問題及解決方法，以便進行後續如摘要、分類、歸類等處理。

美國專利全文沒有用 XML 標示，導致程式無法輕易剖析理解其內容，所幸其 HTML 的標示對短欄位資料的部份還算規則，可以用特殊的網頁剖析器（wrapper）透過正規表達式（regular expression）的字串比對方式，擷取出內容。至於專利說明的全文部份，比較沒有規則，但還是可以在自由文字（free text）中比對大寫關鍵字的文句，擷取出發明領域、發明背景、發明摘要、圖式簡述、發明細節等段落，而自動產生其專利說明的內容目次（Table of Content）。至於專利宣告的部份，雖然也是自由文字，但分項說明，且寫法很傳統：獨立宣告項先寫，其後緊跟著依附宣告項。獨立宣告項為主要的專利標的，依附宣告項通常用來補充、說明獨立宣告項的細節，因此可以根據宣告寫作的習慣，自動擷取出獨立宣告項。

筆者透過電腦程式的字串處理功能，很快的完成從美國專利網站抓取其 HTML 全文、擷取各項短欄位的內容、自動產生專利說明內容目次、以及自動辨識獨立宣告項等工作，並將結果轉入到關聯式資料庫中。目前已有轉入 20 萬筆專利（約 13 giga bytes 資料）到關聯式資料庫的經驗，且發現失敗率非常小。顯示這部份的作業，困難度沒有太大，其中的挑戰，是程式設計者字串比對規則的歸納與整合能力。

然而，若考慮到不同專利之間的資料一致性，則仍有下列問題待解決：一、結構化欄位如發明人、申請人等資料常有同義異名的現象；二、技術名稱岐異且定義不一；三、主題分類不一致、不完整；四、文獻引用參考格式混亂等。除了第四點用字串比對可以勉強處理到一定程度外，其餘三點都難以有效進行自動化的處理。因此人工監控的（人名、機構名）權威控制、同義詞彙的指定、專業術語的人工分析與歸納，對資料的整備、雜訊的消除，還是扮演重要的角色。

二、分段摘要

前面提過，人工分析需要依照專利的目的、方法、效用等分項製作閱讀分析摘要表，而在自動化分類的研究方面也發現分段落各取前 300 字的效果最好，顯示對冗長的專利說明文字，需要進行段落切割與摘要，以提高專利的可讀性以及後續人工或自動處理的能力。

前一階段取得的專利說明內容目次，可獲得：發明領域、發明背景、發明摘要、發明細節等段落，透過自動摘要的技術，可將其分別對應到該專利的應用領域、待解問題、解決方法與獲得效益，整理成便利應用的形式。

自動摘要的作法，大抵可分爲「摘錄」( extraction ) 與「摘要」( abstraction ) 兩種。「摘錄」的結果爲文件中重要文句的重組。相對的，「摘要」的結果則不限於文件中的文句。由於「摘要」所需資源較多，目前以「摘錄」爲主的研究佔較多數。筆者以自動摘錄常用的技術，對專利的說明文字進行：「斷句」( 將文字按句子或更小的子句單位分段 )、「評句」( 評估句子的重要性 )、「選句」( 選擇少量具內容代表性的重要句子來代表原文 ) 與「組句」( 將可能離散四處的重要句子組合成或呈現成可讀性高的摘要 )。在最重要的「評句」方面，使用到多種方法，包括：評估句子包含重要詞彙（如關鍵詞、標題詞、線索詞等）的情況、評估句子在文件中的位置（文件頭、文件尾、段落頭、段落尾、內容首次出現）等。初步的實驗發現，專利文件的自動摘要技術與一般文件不同，特別是線索詞、文句位置對某些主題段落如宣告與發明背景特別重要。

圖一展示一份美國專利的標題、發明背景說明及其自動摘要後的結果。我們運用 MS Word 內建的自動摘要功能，與筆者自行發展的摘要技術做比對。圖一藍色斜體底線字爲 MS Word 做出來的摘要，有四句；紅色粗體底線字爲我們目前做出來的摘要，有三句。Word 選到的第一與第二個重要的句子，雖然能點出關聯式資料庫的功用，但對領域專家而言，這只是常識，而其第三、四句才較能反映發明背景的重點。至於我們選到的第一、二句，則直指過去方法的缺點，第三句則指出爲何會進行這項發明的理由。從原文的敘述中可以發現，事實上有不少句子在說明發明背景，可是細節範圍不同，例如：提到關聯式資料庫（relational database）只是將範圍限定於資料處理領域，提到資料探勘（data mining）則是界定其資料處理類型，而提到目前資料庫的資料探勘有哪些缺點才是本專利眞正的發明背景。大範圍的背景界定對非領域專家也許有用，但專利分析專家應該會直接想看較精細的背景說明。因此，我們的摘要結果應該比較符合

專家所需。

雖然目前大部分的發明背景已有不錯的摘要結果，然而我們發現其他段落敘述重點的方法差異頗大。如何針對不同的主題段落，發展出不同的摘要方法，以便能夠簡短有效的指出其方法、特徵與效用，實乃一大挑戰。

TITLE

Vertical implementation of expectation-maximization algorithm in SQL for performing clustering in very large databases.

BACKGROUND OF THE INVENTION

Relational databases are the predominate form of database management systems used in computer systems . *Relational database management systems are often used in so-called " data warehouse " applications where enormous amounts of data are stored and processed .* In recent years , several trends have converged to create a new class of data warehousing applications known as data mining applications . Data mining is the process of identifying and interpreting patterns in databases , and can be generalized into three stages .

Stage one is the reporting stage , which analyzes the data to determine what happened . Generally , most data warehouse implementations start with a focused application in a specific functional area of the business . These applications usually focus on reporting historical snap shots of business information that was previously difficult or impossible to access . Examples include Sales Revenue Reporting , Production Reporting and Inventory Reporting to name a few .

Stage two is the analyzing stage , which analyzes the data to determine why it happened . As stage one end-users gain previously unseen views of their business , , they quickly seek to understand why certain events occurred ; for example a decline in sales revenue . After discovering a reported decline in sales , data warehouse users will then obviously ask , " Why did sales go down ? " Learning the answer to this question typically involves probing the database through an iterative series of ad hoc or multidimensional queries until the root cause of the condition is discovered . *Examples include Sales Analysis , Inventory Analysis or Production Analysis .*

Stage three is the predicting stage , which tries to determine what will happen . As stage two users become more sophisticated , they begin to extend their analysis to include prediction of unknown events . For example , " Which end-users are likely to buy a particular product " , or " Who is at risk of leaving for the competition ? " It is difficult for humans to see or interpret subtle relationships in data , hence as data warehouse users evolve to sophisticated predictive analysis they soon reach the limits of traditional query and reporting tools . Data mining helps end-users break through these limitations by leveraging intelligent software tools to shift some of the analysis burden from the human to the machine , enabling the discovery of relationships that were previously unknown .

Many data mining technologies are available , from single algorithm solutions to complete tool suites . Most of these technologies , however , are used in a desktop environment where little data is captured and maintained . Therefore , most data mining tools are used to analyze small data samples , which were gathered from various sources into proprietary data structures or flat files . On the other hand , organizations are beginning to amass very large databases and end-users are asking more complex questions requiring access to these large databases .

*Unfortunately , most data mining technologies cannot be used with large volumes of data . Further , most analytical techniques used in data mining are algorithmic-based rather than data-driven , and as such , there are currently little synergy between data mining and data warehouses .* Moreover , from a usability perspective , traditional data mining techniques are too complex for use by database administrators and application programmers , and are too difficult to change for a different industry or a different customer .

One analytic algorithm that performs the task of modeling multidimensional data is " cluster analysis." Cluster analysis finds groupings in the data , and identifies homogenous ones of the groupings as clusters . If the database is large , then the cluster analysis must be scalable , so that it can be completed within a practical time limit .

**In the prior art , cluster analysis typically does not work well with large databases due to memory limitations and the execution times required . Often , the solution to finding clusters from massive amounts of detailed data has been addressed by data reduction or sampling , because of the inability to handle large volumes of data .** However , data reduction or sampling results in the potential loss of information .

**Thus , there is a need in the art for data mining applications that directly operate against data warehouses , and that allow non-statisticians to benefit from advanced mathematical techniques available in a relational environment .**

圖一：專利文件自動摘要比較。

三、主題偵測與歸類

自動歸類（clustering）可運用於專利檢索結果的重新組織，將相似的專利集合在一起，方便使用者瀏覽。自動歸類亦可運用於專利分析，偵測領域的主題及其分佈情況。歸類作法可分爲詞彙歸類（term-based clustering）與文件歸類（document- based clustering）。詞彙歸類是計算各詞彙之間的相似度（例如各詞彙共同出現在同一文件的統計量），將相似的詞彙連結在一起。其優點是歸類後的類別標題可由這些詞彙直接讀取，而其缺點是詞彙之間的非遞移性（non-transitivity）容易造成錯誤的歸類結果。例如：「圖書館」與「數位內容」相似、「數位內容」與「電腦動畫」相似，則「圖書館」與「電腦動畫」容易被歸類在一起。文件歸類則計數任意兩文件的主題相似度（或詞彙重複程度），再將相似的文件連結成一類。其優點是比較能夠偵測文獻的主題分佈，缺點是類別的主旨標題不易自動選定。在專利地圖的製作上，由於需要偵測主題分佈與類別的主旨標題，可能會同時用到詞彙歸類與文件歸類的方法，並且利用專利全文、專利段落（前述之發明背景、發明摘要等段落）的全文，以及專利段落的摘要等，分別進行歸類，以找出最佳的歸類方法與策略。

筆者事先即發展過一些方法，以驗證上述的想法是否可行。過去文獻的作法雖然提出很多文件歸類的作法，然而比較少提到類別標題如何決定。我們根據文獻上的技術，自行加入一些想法：先依 complete link 方式將文件歸類，然後用相關係數[39]算出各個詞彙與各個類別的相關程度，以取得類別標題詞。圖二爲 2003 年 5 月 20-22 日中央社新聞文件自動歸類的部份結果。藍色底線部份爲文件標題，其他則爲節點，內有節點編號、此節點下各文件間的最低相似度、以及此節點的主題。這裡以數個詞代表一個主題，且每個詞之後爲其與該類別的相關係數。圖二顯示歸類的效果還不錯，能夠細微區分 SARS 感染與台灣加入 WHO 兩個主題。但同樣的方法運用於英文專利全文時，發現文件太長、內容主題太多，歸類結果不盡理想。顯然切割專利的主題段落，再依此段落的全文或摘要各自進行自動歸類，才有可能得到滿意的結果。

527:0.0932 (病例:0.7167, 香港:0.6491, 感染:0.4912, ＳＡＲＳ:0.2205)
- 425:0.1030 (死亡:0.7924, 香港:0.7815, 北京:0.7688, 病人:0.7192)
  - 40:0.2817 (北京:1, 累計:0.8628, 疑似病例:0.8135, 診斷:0.8135)
    - 13:0.3454 (診斷:1, 新增:1, 疑似病例:1)
      - 34： 香港僅增一例WHO：擬撤旅遊警告
      - 92： 大陸稱疫情平緩WHO：勿遽下結論
    - 147： 大陸SARS通報新低
  - 47:0.2697 (全球:0.5687,急性:0.5245,呼吸道:0.4888,症候:0.4888)
    - 90：SARS敲警鐘WHA催生全球防疫網
    - 116： ＷＨＯ公布臺灣煞死率全球第三
- 120:0.1912 (世衛組織:0.5687,病例:0.3759,感染:0.2576,ＳＡＲＳ:0.1156)
  - 35： 星新病例追查感染源
  - 152： ＷＨＯ：菲切斷傳染鏈除煞

675:0.0818 (ＷＨＯ:0.5783, 臺灣:0.3431, )
- 161:0.1649 (連戰:0.4514, ＷＨＯ:0.4322, 臺灣:0.2564, )
  - 41:0.2812 (政治:0.6592,連戰:0.5332,主席:0.5028,ＷＨＯ:0.3729)
    - 9:0.3562 (連戰:0.4338,ＷＨＯ:0.30,疫情:0.21,臺灣:0.18)
      - 47： 臺灣入WHO連戰促北京去政治化
      - 52：CNN專訪連戰嚴譴北京打壓
    - 108： 連戰促扁疫情危急勿政治操弄
  - 48： 吳伯雄：參與世衛五重點動員海內外推動
- 224:0.1428 (陳建仁:0.6592,ＷＨＯ:0.3729,臺灣:0.2212,ＳＡＲＳ:0.1422)
  - 16:0.3427 (陳建仁:0.81,ＷＨＯ:0.30, 臺灣:0.18,ＳＡＲＳ:0.12)
    - 113：WHA未邀陳建仁報告疫情
    - 161： 湯姆森：WHO專家赴臺滅火 陳建仁：WHA拒我遺憾激動
  - 127： 世衛宣布全臺旅遊示警朝野同促醫護抗疫補網

圖二：2003 年 5 月 20-22 日中央社新聞歸類分析的部份結果。

## 伍、結語

文字探勘技術在專利分析的運用，有很多值得探討的課題。上一節筆者介紹了其運用的可行性，但也指出還有許多挑戰待克服。專利文件的關鍵詞擷取、關聯詞建構、摘要、歸類、分類、專利地圖建構、近似專利檢索等方法與技術，都需要根據專利文件的特性做調整。而這需要領域專家、法律專才、資訊技術人員、專利分析人員、與嫻熟資訊組織與主題分析的圖書專業人員，共同合作，一起規劃，才比較容易設計出較佳的專利分析方法與工具。

日本政府於 2002 年 7 月發表的「智財權策略大綱」中提及，大學院校、官方機構及企業團體在主導研究發展的初階段，即應活用國內外專利資訊，以制訂研發策略、評選研發課題。顯示專利分析對國家經濟、產業前景、科學技術發展的重要性，而其相關的研究，近幾年各國才開始投入。國內這方面的起步似乎慢了一些，但只有幾年之差，積極的投入，有很大的機會可以迎頭趕上，甚至領先超越。

## 參考文獻：


[1] 夏文龍，”專利對產業界的價值”，智慧財產權管理，16，1998 年，頁 20-21。

[2] 陳碧莉，”專利地圖簡介”，國科會專利地圖與專利分析人才培訓計劃授課講義，台北市：行政院國科會，1999 年。

[3] 劉尚志，專利分析與資料檢索，專利分析及專利運用人才基礎培訓搬授課講義，台北市：行政院國科會，2000 年。

[4] Shang-Jyh Liu, “Patent Map – A Route to a Strategic Intelligence of Industrial Competitiveness,” 第一屆亞太專利地圖研討會，台北，2003 年 10 月 29 日, 頁 1-1 到 2-13。

[5] Campbell, “Patent Trends as a Technological Forecasting Tool”, World Patent Information , Vol. 5, No. 3, 1983, pp. 137-143.

[6] 陳達仁，黃慕萱，專利資訊與專利檢索，文華圖書管理資訊股份有限公司出版，2002。

[7] 謝寶煖，”專利與專利資訊檢索”，大學圖書館，第二卷第四期，1998 年，頁 111-127。

[8] Young-Moon Bay, “Development and Applications of Patent Map in Korean High-Tech Industry”, 第一屆亞太專利地圖研討會，台北，2003 年 10 月 29 日,頁 3-1 到 3-23。

[9] José María Gómez Hidalgo and José María, “Text Mining and Internet Content Filtering,” ECML/PKDD-2002 Tutorial, http://ecmlpkdd.cs.helsinki.fi/hidalgo.html.

[10] NTCIR-4 Patent Retrieval Task, http://www.slis.tsukuba.ac.jp/~fujii/ntcir4/cfp-en.html。

[11] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurasamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/ The MIT Press, 1996.

[12] 陳光華，呂明香。「知識探索及其於政府資訊之應用」。檔案季刊第 2 卷第 2 期（民國 92 年 6 月），頁 38-49。

[13] 麥富德、黃楓台、簡國明、王永銘、陳秋燕，碳奈米管專利地圖及分析，行政院國家科學委員會科學技術資料中心編印，2002 年 4 月。

[14] 連穎科技，”Patent Guider”, http://www.learningtech.com.tw/products/function.htm。

[15] Invention Machine Corporation, http://www.invention-machine.com/

[16] Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Akihiko Takano, " Overview of Patent Retrieval Task at NTCIR-3," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan.

[17] 黃慕萱，「遺傳工程學之專利計量學研究」，國科會 92 學年度研究計畫，2003/08/01～2004/07/31。

[18] 賴奎魁，「利用專利共同引証分析建立專利分類架構之研究」，國科會 92 學年度研究計畫，2003/08/01～2004/07/31。

[19] 楊志海，「科學園區廠商的專利決定因素--一般化分析法」，國科會 91 學年度研究計畫，2002/08/01～2003/07/31。912415H032013.pdf

[20] 劉尚志，「基因科技引伸專利保護，授權，侵權暨國際爭端解決之研究」，國科會 91 學年度研究計畫報告，2001/06/01 ～ 2002/05/31。902420H009002.pdf。

[21] 賴奎魁，「利用專利地圖分析探討影像感測業專利策略之研究」，國科會 89 學年度第一期研究計畫報告，1999/08/01～2000/07/31。892416H224027.pdf。

[22] 馬難先，「落實產學合作－建立智財權管理與專利應用機制計畫」，國科會 88 學年度研究計畫報告，1999/02/01～1999/12/31。883011P398001.pdf。

[23] ACM SIGIR 2000 WORKSHOP on PATENT RETRIEVAL, http://research.nii.ac.jp/ntcir/sigir2000ws/, Athens, Greece, July 28, 2000.

[24] ACL-2003 Workshop on Patent Corpus Processing, http://www.slis.tsukuba.ac.jp/~fujii/acl2003ws.html, 12 July 2003, Sapporo, Japan.

[25] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa and Makoto Iwayama, “Patent Claim Processing for Readability - Structure Analysis and Term Explanation,” Proceedings of ACL Workshop on Patent Corpus Processing, 12 July 2003, Sapporo, Japan.

[26] Svetlana Sheremetyeva, “Natural Language Analysis of Patent Claims,” Proceedings of ACL Workshop on Patent Corpus Processing, 12 July 2003, Sapporo, Japan.

[27] “Intelligent Patent Analysis through the Use of a Neural Network: Experiment of Multi-Viewpoint Analysis with the MultiSOM Model,” Proceedings of ACL Workshop on Patent Corpus Processing, 12 July 2003, Sapporo, Japan, pp. 7-23.

[28] C. J. Fall, A. Torcsvari, K. Benzineb, G. Karetka, “Automated Categorization in the International Patent Classification,” ACM SIGIR Forum, Vol. 37 , No. 1, 2003, pp. 10 - 25.

[29] Yuen-Hsien Tseng, "Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text", Journal of the American Society for Information Science and Technology, Vol. 52, No. 5, April 2001, pp. 378-390.

[30] 曾元顯, 林瑜一, " 模糊搜尋、相關詞提示與相關詞回饋在 OPAC 系統中的成效評估", 「中國圖書館學會會報 61 期」, 1998 年 12 月, 第 61 期, 頁 103-125.

[31] Yuen-Hsien Tseng, Da-Wei Juang and, Shiu-Han Chen "Global and Local Term Expansion for Text Retrieval," to appear in the Proceedings of the Fourth NTCIR

Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, June 2-4, 2004, Tokyo, Japan.

[32] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, Nov. 2002, pp. 1130-1138.

[33] 曾元顯, "文件主題自動分類成效因素探討", 「中國圖書館學會會報」, 2002 年 6 月, 第 68 期, 頁 62-83.

[34] Yuen-Hsien Tseng and William John Teahan, "Verifying a Chinese Collection for Text Categorization," Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '04, July 25 - 29 Sheffield, U.K., 2004, pp.556-557.

[35] Yuen-Hsien Tseng and Da-Wei Juang, "Document-Self Expansion for Text Categorization," Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '03, July 28 - Aug. 1, Toronto, Canada, 2003, pp.399-400.

[36] 曾元顯, "中文手機新聞簡訊自動摘要", 第十六屆自然語言與語音處理研討會, 台北, 2004 年 9 月 2-3 日, 頁 177-189.

[37] Yuen-Hsien Tseng, " Content-Based Retrieval for Music Collections," Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99, Aug. 15-19, Berkeley, U.S.A., 1999, pp.176-182.

[38] 曾元顯, "數位文件之資訊組織與主題分析自動化之技術與應用", 「台北市立圖書館館訊」, 2002 年 12 月, 第 20 卷, 第 2 期, 頁 23-35.

[39] 曾元顯, 莊大衛, "文件自我擴展於自動分類之應用", 第十五屆計算機語言學研討會, 2003 年 9 月 18-19 日, 頁 129-141.