視覺化專利檢索系統

蔡鴻文、陳吉麟*

一、前言

專利公報是各國專利主管機構對外揭露專利申請人所提的專利申請內容文件。一般專利公報主要有四部分，如首頁、申請專利範圍、該發明創作之敘述、圖式。經濟部智慧財產局已建置本國與外國專利說明書之全文資料庫，各界可以很快速地檢索專利技術全文資料影像檔，查詢申請案件狀態及權利異動。傳統的專利檢索系統主要基於布林邏輯運算，由使用者輸入檢索關鍵字，檢索系統根據邏輯運算在特定欄位（例如專利說明書、摘要或請求專利範圍）尋找符合檢索關鍵字之內容，再將檢索到的專利列成檢索結果清單提供給使用者。傳統布林邏輯運算單就文字本身比對可能發生檢索範圍受限的問題，舉例而言，當使用者輸入關鍵字「鋰電池」，檢索結果僅會呈現包含「鋰電池」關鍵字之專利，與「鋰電池」相關的「充電電池」則不會被篩選出來。

考量前述問題，近年來不少專利檢索系統逐漸導入了模糊檢索的功能，例如歐洲專利局 Espacenet 提供的 Smart search 功能。部分平臺的模糊檢索功能建立於文字相似度，透過建立各文字對應之特徵向量，比對不同文字在特徵空間上的相似程度以進行分類。舉例而言，「鋰電池」及「充電電池」在資料集內可能呈現較高的相似度，而「鋰電池」及「大腸桿菌」則在資料集內則可能呈現較低的相似度。部分平臺亦導入機器學習系統，使演算法透過文件之間的規則自行建立文字之間相似度。

儘管模糊檢索已逐漸普及，然而現行專利檢索系統在資料的呈現方式上仍然採用過去的檢索結果清單。換而言之，使用者仍然必須逐一點選清單上的各個專利以確定與搜尋目標的相關程度。相較於傳統布林檢索，這種問題在模糊檢索的情況下更為嚴重，因為專利檢索系統會提供大量相似但不全然符合使用者檢索目標的專利，令使用者在逐一閱讀檢索結果清單的負擔加重。另外，在

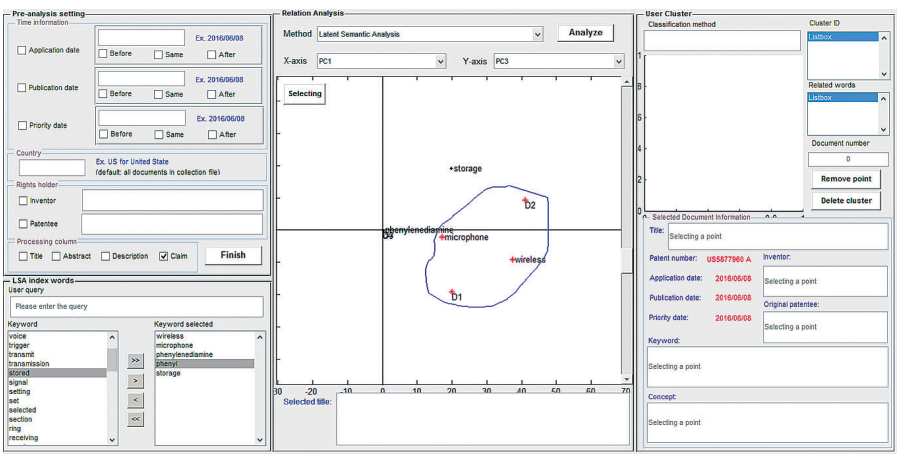
* 蔡鴻文，國立臺灣科技大學專利研究所副教授；陳吉麟，專利師（108 級）。

部分的檢索情境，使用者可能獲得一兩篇專利或文章而想檢索與之相關的其他專利。在這種情境下，傳統上使用者必須自行思考歸納與現有專利或文章相關之關鍵字，再以該些關鍵字執行檢索。如此的處理模式除了耗費人力之外，更導致不同的使用者獲得的檢索結果不盡相同，不利於資料的統計分析。

有鑑於此，我們提出了一種新型的視覺化專利檢索系統，採用模糊檢索的概念並以圖像特徵空間的方式加以呈現。使用者可以輕易的從圖像特徵空間上點與點之間的距離判斷專利與專利或專利與文字間的相似程度，對於相似程度過低的專利，使用者可以選擇忽略以節省檢索時間。以下我們以 MATLAB 軟體作為示範說明。

二、軟體介面

本文所示範之視覺化專利檢索系統以 MATLAB 的 GUI 模式呈現，其介面如圖一所示。軟體介面可區分為四大區域，分別為左上角的分析前設置區（Pre-analysis setting）、左下角的關鍵字區（LSA index words）、中間的關聯性分析區（Relation Analysis）及右方的使用者選取區（User Cluster）。視覺化專利檢索系統可以接收使用者提供的幾篇現有專利構成的資料集，計算資料集之中最為相關或出現字頻最高的單字推薦給使用者，使用者可以從中選取出關鍵字後，基於這些關鍵字執行相關性分析（本文將以潛在語意分析（Latent Semantic Analysis, LSA）作為示範），轉換後的關鍵字及專利文件會以點群聚的方式分布在圖像特徵空間，藉此找出相似的專利。圖二為本文所例示的視覺化專利檢索系統分析流程，總共可以區分為六個步驟，我們將依序說明。

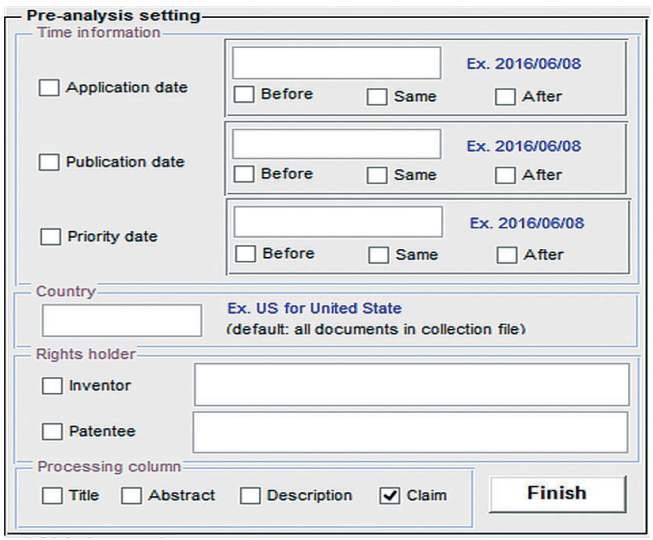


圖一：視覺化專利檢索系統介面

圖二：視覺化專利檢索系統分析流程

1. 專利資料庫資料擷取

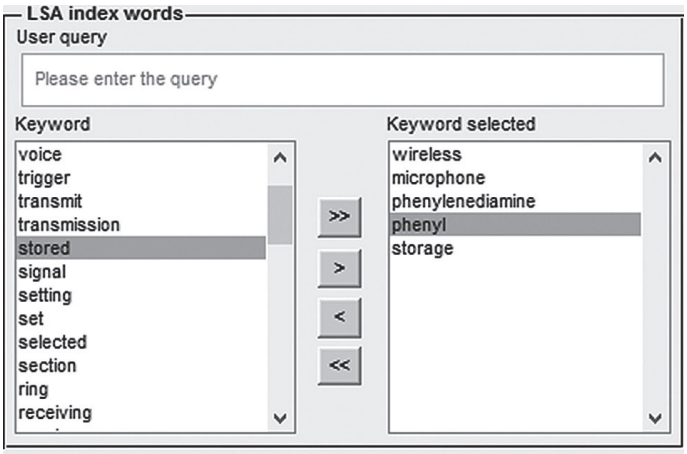
圖三是視覺化專利檢索系統介面的分析前設置區（Pre-analysis setting）的擷取圖。如同一般專利檢索流程，視覺化專利檢索系統提供初步的專利篩選，如申請日、公開日、優先權日、申請國、發明人及申請人。使用者輸入檢索項目及欄位後，按下 finish，視覺化專利檢索系統從資料庫中篩選出符合要求的數筆資料。本文將使用四筆專利文件輔助說明，分別為 D1: US 9131292 B2（Wireless microphone system）、D2: US20030087667 A1（Wireless microphone system, voice receiving apparatus, and wireless microphone）、D3: US20020106089 A1（Audio trigger devices）及 D4: WO2000003704 A1（Macrophage scavenger receptor antagonists）。四筆專利資料中，D1 與 D2 與無線麥克風相關；D3 為聲音裝置；D4 與巨噬細胞受器相關。在這裡，我們可以預先猜想 D1 與 D2 會得到較高的相似度，D3 相較於 D1 與 D2 的相關性較低，D4 相較於 D1 與 D2 的相關性最低。於本文中，程式僅能根據四篇專利資料內的文字進行判斷，而未能預先作此假設。



圖三：視覺化專利檢索系統之分析前設置區

2. 高頻字擷取與去除

圖四是視覺化專利檢索系統介面的關鍵字區（LSA index words）的擷取圖。視覺化專利檢索系統從四篇專利資料構成的資料集擷取內容文字，計算出高頻字並去除無關字（Stopwords，例如 and、it、this、however……等），最後於關鍵字區給予推薦。



圖四：視覺化專利檢索系統之關鍵字區

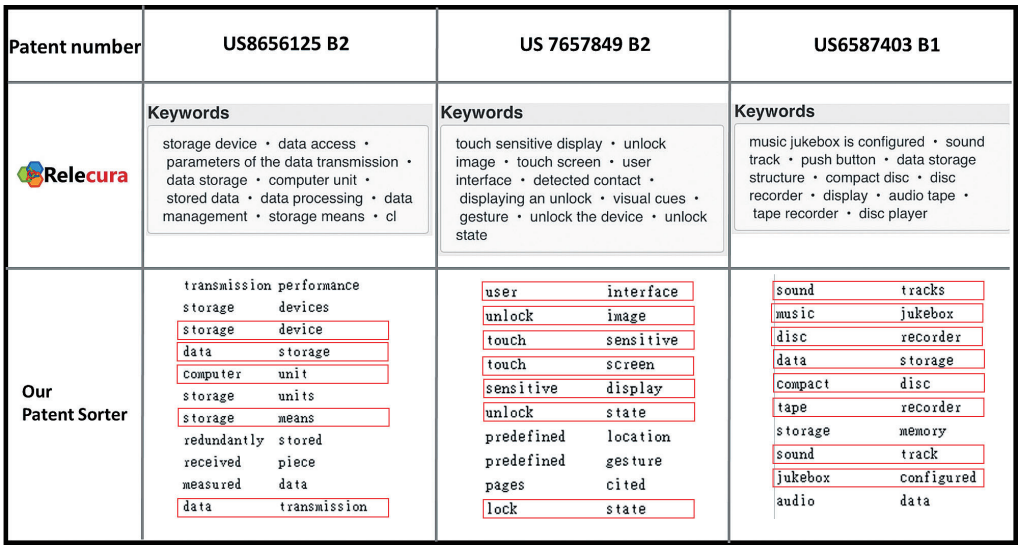
3. 關鍵字選取

使用者可以基於關鍵字區所列出的推薦字進行選取，以選出最相關的關鍵字，例如將圖四位於左方推薦欄位的關鍵字選取移動至右方的選取欄位。前述步驟節省了使用者自行從四篇專利文件中歸納出相關文字的麻煩，僅要從系統推薦的關鍵字中進一步選取即可。此外，使用者也可以選擇在關鍵字區輸入檢索句（query）的方式執行搜尋，此方案就如同傳統的檢索模式。

視覺化專利檢索系統可以給予組合字（由二至多個單字組合而成之單詞）的推薦，經過我們的測試，其執行結果與現有市面上搜尋引擎的關鍵字結果比較相去不遠，更甚者，我們的系統可以挑選出該些文字的複數型或分詞。舉例而言，在 US8656125B2 的資料比對中，storage device 及其複數型 storage devices 被挑選出來。

4. 建立關鍵字與專利文件 TF-IDF 矩陣

於步驟四中，我們基於專利文件及使用者選出的關鍵字建立出文件—文字關係矩陣，即各關鍵字在各文件的出現字數。舉例而言，如圖六所示，共有文



圖五：視覺化專利檢索系統與現有搜尋引擎之關鍵字推薦結果比較圖

	文件1	文件2	文件3
關鍵字1	1	0	3
關鍵字2	0	5	2
關鍵字3	2	1	6

圖六：文件—文字關係矩陣

件 1 至 3 等三份文件，我們希望找出這些文件中是否分別包含關鍵字 1、關鍵字 2 或關鍵字 3。對文件 1 而言，其內容出現了 1 次關鍵字 1、0 次關鍵字 2 及 2 次關鍵字 3。藉此歸納，我們可以在所有的專利文件之中尋找我們在步驟三選取的所有關鍵字，並判斷各關鍵字在各文件的出現次數。

其後，矩陣數值以 TF-IDF 公式（如圖七）加權處理以凸顯重要關鍵字。於 TF-IDF 公式之中，TF 表示關鍵字出現的頻率；idf 則可以表示為 log（總文件數量 N/ 出現某單字的文件數量 dfx），因此當文件中關鍵字出現越多次，TF 越高，加權越高（重要性高）；而當關鍵字出現過的文件數越多，idf 越低，加權越低（重要性低）。舉例而言，文字「microphone」在文件 D1 大量出現，因此極有可能為重要的關鍵字，因此 TF 分數高；文字「the」在文件 D1 大量出現，雖然 TF 分數高，但文字「the」在其他文件也大量出現，因此 idf 分數低，總體而言文字「the」的重要性較低。

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

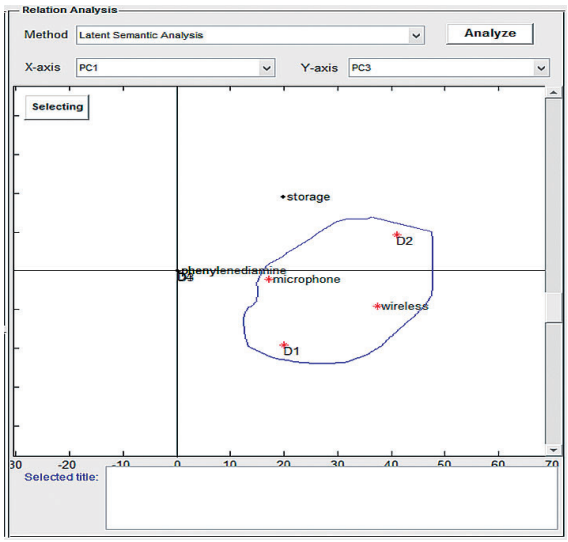
圖七：TF-IDF公式

TF 機制有助於我們找出高頻出現的文字，並假設該些文字屬極為重要性之關鍵字。idf 機制有助於我們刪除在步驟二中未被刪除的 Stopwords，並且透過此鑑別機制，該些文字（例如前述的文字「the」）可以被列入步驟二的 Stopwords 清單，以利於在未來的分析流程中直接在步驟二即將其排除。儘管 TF-IDF 建立於文字出現頻率與其重要性正相關之基本假設，而該假設並非在所有情況下皆正確。然而在多數專利文件下 TF-IDF 仍然是非常具有參考價值。理由在於專利文件的明確性要求，在內部文件（專利說明書、申請專利範圍與摘要）之中，撰稿人多採用一致性的用語以避免區別解釋，而此一致性的用語導致該文字在文件中大量地出現。另外，為符合通常知識的解讀，撰稿人在選擇部分文字時會參考外部文件（例如字典或其他專利文件）對於特定文字的用字遣詞，這也導致相關技術的專利具有相同或類似用語的可能性大大提升。

5. LSA 分析建立語意空間

圖八是視覺化專利檢索系統介面的關聯性分析區（Relation Analysis）的擷取圖。經過 TF-IDF 加權後的矩陣，我們將其做奇異值分解得到其特徵值及特徵向量，藉此我們將關鍵字及專利文件映射到特徵空間。圖八呈現出本文所例示的四筆專利資料以專利說明書欄位做文字處理後，映射到特徵空間的結果。特徵空間上的每一個資料點可以代表一個關鍵字（例如圖八上的 wireless）或一份專利文件（例如圖八上的 D2，即 US 20030087667 A1）。使用者可以圈選方式將這些群聚的資料分類在一起。在資料分布上，具有較重要特徵的資料點會遠離原點，可以看見 D1 及 D2 都具有較明顯的特徵且形成群聚，而與其他較無關的 D3 及 D4 文章則集中在原點。

在資料呈現上，視覺化專利檢索系統提供當游標移動到特徵空間上的資料點或予以點擊，即在使用者選取區（User Cluster）呈現該專利的詳細內容，例如摘要或圖式。如此一來，節省了使用者在操作專利檢索系統必須逐一查看檢索結果清單上的專利並自行歸納所花費的時間與心力。



圖八：視覺化專利檢索系統之關聯性分析區

6. 相似度分析

最後，藉由計算出資料點在特徵空間中的距離，便可以轉化為關鍵字或專利相似程度的相關程度參數。我們可將基於文字相關度處理得到的參數、基於 IPC 或其他分類號處理得到的參數、引用文獻相關參數……等總和比較得出專利文件的相關度。該些相關度也可以用作特徵空間上的 X 軸或 Y 軸。如圖八所示，關聯性分析區上方的跳出式選單（pop menu）呈現 X 軸為 PC1，Y 軸為 PC3，即表示基於文字分析所獲得之特徵值。該些軸向亦可以置換為專利的引用數量或公告時間，令使用者只針對有興趣範圍的資料進行探勘。

三、結語

本文提出了一種新型的專利檢索方案，系統提供使用者關鍵字推薦，使用者可以就系統所推薦的關鍵字進行選擇（或自行輸入其他關鍵字），節省了使用者基於現有專利或文章自行歸納關鍵字的麻煩。此外，使用者可以採用較直覺的圖像化方式選擇有興趣的檢索結果進行詳細閱讀，以避免逐一點選傳統檢索結果清單的繁雜。並且，視覺化專利檢索系統所呈現出的特徵空間雖然是基於文字的出現頻率產生，但在使用者預先選取有技術含意的關鍵字後，呈現在特徵空間上的關鍵字與專利分布一定程度反映出了其技術關聯度，而相當於形成「技術分布」。

因此，使用者可以直接在此抽象形成的「技術分布」空間上尋找相關聯的專利與其他關鍵字，而無須將專利內容自行歸納為關鍵字後重新執行檢索。並且，視覺化專利檢索系統提供圈選機制，可以將聚落內的專利或關鍵字打包輸出或下載，令使用者清楚了解其所打包的專利可能與哪些關鍵字相關。

本文僅以較簡單直觀的 LSA 分析機制呈現視覺化專利檢索系統，但在資料探勘的語意分析領域，諸多學者已經提供許多更為優異的演算法，該些演算法在未來亦可以採用模組化之方式納入視覺化專利檢索系統之關聯性分析區上的座標軸，以提供使用者更多的語意分析工具選擇。

致謝

本研究感謝科技部「可自動技術分類之多國語言專利系統研究（MOST 105-2633-H-011-001）」、「巨量專利文件之智能化技術分類與管理系統（MOST 109-2622-H-011-002-CC3）」、「中英文專利文獻之智能化計量分析與技術地圖方法論研究與系統開發（MOST 109-2410-H-011-021-MY3）」計畫經費補助。

參考文獻

張瑞芬、張力元、吳俊逸、樊晉源（2012）。《專利分析與智慧財產管理：以資訊技術與知識管理方法為手段》，臺北：華泰文化出版。

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing and Management,* 39, 45-65.

Broekstra, J., Klein, M., Decker, S., Fensel, D., Harmelen, F.-V., & Horrocks, I. (2002). Enabling knowledge representation on the Web byextending RDF Schema. *Computer Networks,* 39, 609-634.

Cheng, K. S, Young, G. H., & Wong, K. F. (1999). A study on word-based and integral-bit Chinese text compression algorithms. *Journal of the American Society for Information Science,* 50(3), 218-228.

Chen, Dar-Zen, Lin, Wen-Yau & Huang, & Mu-Hsuan. (2007). Using essential patent index and essential technological strength to evaluate industrial technological innovation competitiveness. *Scientometrics,* 71(1), 106-116.

Hou, J.-L., & Lin, F.-H. (2006). A hierarchical classification mechanism for organization document management. *International Journal of Advanced Manufacturing Technology,* 28(3), 417-427.

Jenkins, C., Jackson, M., Burden, P., & Wallis, J. (1999). Automatic RDF metadata generation for resource discovery. *Computer Networks,* 31, 1305-1320.

Trappey, A.J.C., Trappey, C.V., Chiang, T.A., & Huang, Y.H. (2013). Ontology-based neural network for patent knowledge management in design collaboration. *International Journal of Production Research,* 51(7), 1992-2005.