國立交通大學

資訊科學系

碩 士 論 文

以時間序列將專利文件視覺化的研究

Time Series Visualization of Patents

研 究 生：王瓊婉

指導教授：柯皓仁　教授

楊維邦　教授

中 華 民 國　九 十 四　年 六 月

以時間序列將專利文件視覺化的研究
Time Series Visualization of Patents

研 究 生： 王瓊婉　　Student：Chiung-Wan Wang

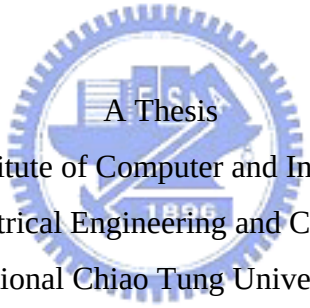
指導教授： 柯皓仁　　Advisor： Hao-Ren Ke

楊維邦　　Wei-Pang Yang

國 立 交 通 大 學

資 訊 科 學系

碩 士 論 文



A Thesis

Submitted to Institute of Computer and Information Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer and Information Science

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

Times Series Visualization of Patents

Student: Chiung-Wan Wang Advisor: Dr.Hao-Ren Ke, Dr. Wei-Pang Yang

Institute of Computer and Information Science

National Chiao Tung University

ABSTRACT

With the coming of economic-based knowledge, we pay much attention to intellectual property (IP) content and the amount of patent documents increases quickly. Patent documents are the repository of how technology advances and, more importantly, show how language supports the change.[2] However, increasing patent documents makes the reading complicate and requires advanced information technology to assist the investigation of patents. In this paper, we propose a patent trend analysis system, which combines text mining and statistic test methods. We want to show the technology trend in a specific field. In the patent trend analysis system, firstly, we make definitions about concepts and apply text mining to extract important concepts from patents. Second, we use statistic test to check whether the retrieved concepts are significant in a specific time interval. Finally, we visualize the analysis result.

Keywords: Patent trend analysis, patent map, text mining, patent analysis

以時間序列將專利文件視覺化的研究

Time Series Visualization of Patents

研究生：王瓊婉　　指導教授：柯皓仁博士，楊維邦博士

國立交通大學資訊科學研究所

摘要

隨著知識經濟時代的來臨，智慧財產權倍受重視，專利的申請件數是以驚人的速度在增加，但以人工進行專利分析是件耗時耗力的工作，如何快速地從專利文獻中獲得有用的情報，成為現今相當重要的議題。

本論文主要在提出一套系統，輔佐使用者進行專利分析，呈現特定領域的技術與經營脈絡。首先，使用文字探勘的方法，擷取出專利文件中的重要概念，再採用統計檢定的方法，檢定詞與時間的關係，定義我們所謂的趨勢，最後將分析結果以視覺化的界面呈現給使用者。

關鍵字：專利地圖、趨勢分析、文字探勘、專利分析

誌謝

當得知能順利畢業那一刻，心中除了感謝還是感謝。

首先要感謝的是楊維邦老師與柯皓仁老師兩年來的循循善誘，在兩位老師的悉心指導下，除了在課業與論文上的督促、獨立思考的訓練外，更讓我瞭解許多待人處事的道理。

接著再感謝實驗室的學長、同學、學妹們，因為有你們的陪伴，讓我的碩士生涯變得多采多姿，也因為有你們的幫忙，讓我能克服研究與課業上的難關。尤其特別要感謝忠億學長在學業上的傾囊相授與不遺餘力的幫忙，鎮源學長在研究思想上的啓蒙，夙賢學長在專利研究上的指引以及政璋、世彥不厭其煩聽我抱怨、解決我所遇到的問題。

另外，要感謝在我低潮期，一直鼓勵、支持我的好友—妙卉、穎瑜、鈺佩、淑伶、亭如、宥瑩還有與我互相扶持的好室友兼好同學—古典，真的很謝謝妳們的包容與關懷。

最後，要感謝我的家人，因為有你們的支持，我才能安心無虞的做研究，這篇論文才能順利完成。僅以此論文研究結果獻給我的家人與朋友。

July, 2005

目錄

表目錄

圖目錄

方程式目錄

第一章　簡介

第一節　研究動機與目的

自 1990 年初，由於全球經濟發展速度減緩，再加上各國對知識配置、生產和使用的依賴程度日益增加，使得許多工業國家意識到全球經濟已從工業時代過渡到知識經濟時代[27]。知識經濟時代與工業時代最大的不同是：企業賴以生存的立基點已不再是有形的產品，而是無形的資訊或知識。但知識是無形的，難以定義與量化。在各類各式的知識中，專利對政府、企業、產業而言是最具經濟價值的產物[29]。根據世界智慧財產局組織 (World Intellectual Property Organization, WIPO)的報告，專利說明書包含了世界上 90~95%的研發成果，其它技術文獻 (如論文或期刊) 僅含有 5~10%的研發成果。專利資訊是各種產業中最具有指標、最具商業價值的技術。掌握住專利，不但掌握了業界的發展動向，亦是技術發展的指標。企業若能善加利用專利情報，可以縮短研發時間及節省研究經費。

專利文件是以驚人的幅度快速在成長，而專利文件所衍生出來的專利資訊更是不斷增長，並且對企業、國家發展的影響是日益漸增。然而目前專利文件的晦澀拗口難以閱讀又是眾所皆知，這是由於專利文獻獨特的文法結構，以及特定用語遣辭與一般文章大相逕庭所造成的結果。正因為專利文獻的數量龐大與不易閱讀的特性，想要對專利文獻進行深入的分析，就必須花費大量的時間與人力。舉例而言，韓國的智財局擬定五年製作 120 個專利地圖 (Patent Map)[29] ，可見專利分析工作的繁重。但在技術迅速變動的今天，對新公告的專利說明書進行深度分析已成為長期且持續的工作，如何降低專利分析的時間是相當重要的課題。利用電腦技術設計一套合適的工具，輔佐使用者從專利文

件中萃取有用的情報與資訊，以進行專利分析是刻不容緩的工作。

專利的主要功用可分為：進行權利分析以保障權益、評估與預測技術發展、規畫研發或技術發展項目、掌握企業發展動向及市場需求等四項 [27] ，其中後三者皆與專利趨勢分析具有密不可分的關係，可見在專利分析中，專利趨勢分析是相當重要的一環。

現今大部份專利趨勢分析大都以傳統苦力式的人工閱讀方式進行，再經由專家解釋整合才能向外發佈。目前電腦技術已逐漸應用於專利欄位的統計量化分析，例如大部份專利資訊系統會將專利文件予以分類，而後統計歷年申請專利數、歷年申請公告數、歷年專利成長率、歷年專利權人國家數、歷年專利權人所屬國數、歷年專利權人數量、歷年發明人數量、歷年發明人成長率等。這些專利資訊系統藉著量化統計分析出有用的情報，但對於使用者而言，這些分析只能讓使用者得到廣泛的趨勢，若使用者想進一步瞭解技術情況時，這些資訊的幫助是相當有限的。

專利本身包含了大量的資訊，要以快速且經濟的方式分析獲取有用的情報，除了針對欄位進行統計分析外，有更多重要資訊是隱含在龐大的文字裡面。統計專利中各個欄位的資訊而得到以統計數字為基礎的專利分析資料，僅是概括的趨勢情報。本論文企圖從專利文件的龐大文字敘述中，擷取出專利的主題內容，藉著專利文件的分析希望能呈現技術與經營的脈絡，輔佐使用者進行專利分析。

第二節　研究方法與目標

本論文所提出的專利分析系統，主要架構是結合文字探勘 (Text Mining)、統計檢定方法與專利地圖 (Patent Map)的觀念，提供一個回饋式 (Feedback)的介面，並結合使用者的專業，期使本系統的服務能更貼近使用者的需求。

近年來，文字探勘在各個領域逐漸受到重視，主要用於文件庫中，目前逐漸與專利分析做結合 [14][18]，從專利文件裡挖掘對使用者有用的情報。本論文所提出的專利分析系統，是採用文字探勘來擷取重要的概念(Concept)，而對概念主要有兩種不同的假設：

- 針對各篇專利文件，若一些詞出現在同一篇文章同一個句子的頻率超過一定的比例，則這些詞具有同樣語意，即為該篇專利的重要概念 。
- 針對整個專利文件庫的字而言，以字的相似度分成 *N* 個群，每個群視為候選的重要概念。

本論文利用文字探勘的方式，希望能分析專利文件中的文字，擷取重要的概念後，再採用統計檢定的方法針對先前所挖掘的概念，檢定在某特定時間內，哪些概念已曾熱烈的被討論或是才在萌芽，並把分析結果地圖化—以專利地圖的形式指引使用者，輔佐使用者進行專利分析，當使用者接觸新的領域時能快速的知道自己的位置，掌握正確方向與里程。

第三節　論文架構

本論文的第二章將介紹目前專利分析的研究、相關文字探勘的方法，以及

與趨勢相關的統計檢定；在第三章則介紹本論文所提專利系統的架構及處理程序；第四章透過個案分析驗證處理程序的合理性；最後在第五章總結本論文，並且提出結論以及探討未來研究的方向。

第二章　相關研究工作

本章介紹與本論文相關的研究工作。關於「專利趨勢分析」的相關研究主要分為以下三方面：

- ◆ 專利地圖：[8][14][15][18]
- ◆ 文字探勘：[1][3][5][6][18][20]
- ◆ 統計檢定：[11][16]

圖 2-1 是依照年份與技術所整理與本論文相關的研究發展，斜體部分為本論文所參考的方法。

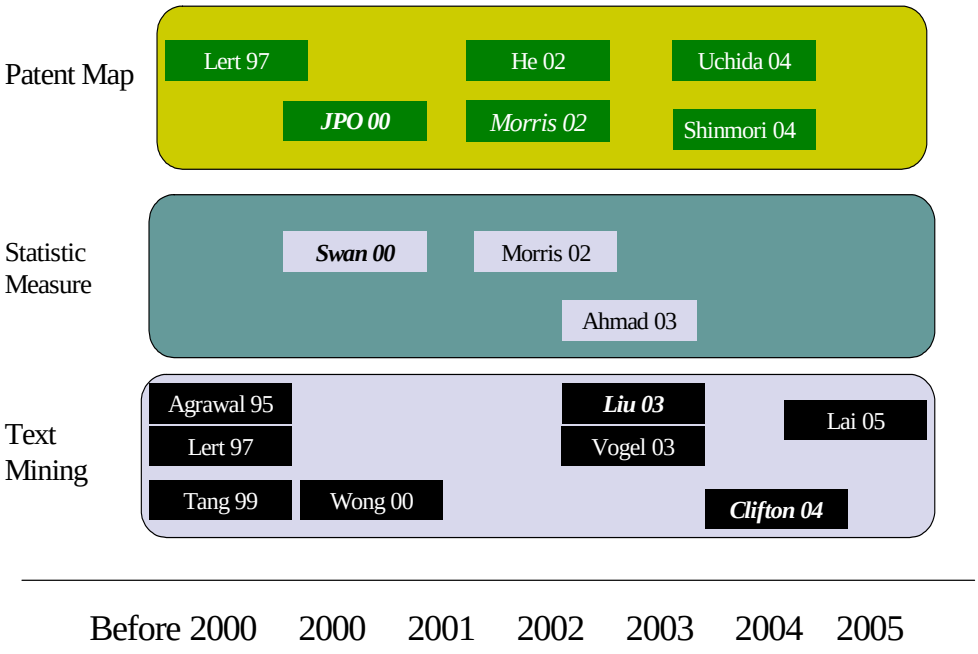


圖 2-1 相關研究工作發展

本論文是在提出一套系統，輔佐使用者進行專利分析，藉著分析特定領域的專利文獻，將技術脈絡以視覺化的方式將分析結果呈現給使用者，所以相關研究把重點放在專利分析的現況、如何利用文字探勘的技術，挖掘出足以代表專利的概念、及利用統計檢定方法檢定概念與時間的關係。

本章第一節先介紹專利的特性以及專利地圖的現況，第二節為介紹擷取重要概念的相關研究，接著第三節說明與趨勢有關的統計檢定方法。

第一節　專利地圖

2.1.1 專利特性

專利一詞出於拉丁文-Litterae Patents，意思是公開信件，是中古世紀君主授與權利及恩典的文件，在英國伊莉莎白一世時，就已經有專利的授與，目的是希望刺激收入並促進新興產業的發展[27]。到今天，專利是一個國家與發明人間的法律協定，國家授與發明人在一定時間享受排他性之製造、利用、販賣發明物行為的權利。正由於專利文件結合技術文件與法律文件的特性，相較於其它科技文獻而言，專利具有以下的特性[29]：

1、資訊源分散，各國有各國的專利系統，即使有歐洲專利局(European Patent Office)整合在歐洲方面的專利，但使用者難以一次蒐集到完整資訊。
2、查全導向 (Recall-Oriented)：在某些應用情況下，漏檢重要專利可能要付出很大的代價。
3、專業用語、法律用語並存，使得專利的閱讀極為困難。
4、相同事物的概念，常用不同的用詞描述，用以規避雷同、散播侵權地雷。

5、結構化與非結構化資訊並存。

由於專利查全導向的特性，目前市面上所能看到的系統多著重於專利資料庫的整合，最近幾年，才開始出現關於計算語言學 (Computing Linguistics)及自然語言 (Natural Language)於專利分析上的應用[21][22]。

2.1.2 專利架構

一般而言，專利說明書具有相當固定的欄位，其中比較重要的有：

1、專利編號 (Patent No)：當專利文件申請通過後，申請單位所給予一個獨一無二的編號。

2、專利標題 (Title)：專利的名稱全文，需符合發明的主題。

3、摘要 (Abstract)：以簡明扼要的文字，說明該發明或創作的構造或方法，通常與標題出現於專利說明書的首頁中。

4、宣告 (Claim)：該項技術所宣告的權利範圍，具有法律效果，向國家訴求保護核心所在，相較其它欄位，宣告中的資訊都是條列式的。

5、說明 (Description)：專利中描述技術最完整欄位，包含了細部的技術說明，是所有欄位中資訊量最多的欄位。

2.1.3 專利地圖 (Patent Map)

所謂的專利地圖是對特定領域的專利文件進行分析處理，並對所搜集到的專利資訊視覺化 (Visual Representation Of Related Patent Information) [8]，主要目的在於瞭解競爭對手的情況以及找尋本身機會。

以專利地圖來表示的專利資訊主要分為二類，一為「經營圖」，偏向於申請專利獲准件數為主之統計，分析各個國家、公司、發明人，相關技術佔有、競爭之情形，同時亦對各個專利被引用之情形、專利年齡（即專利期限）、技術生命週期等做各種專利經營面之分析。另一類為所謂之「技術圖」，此乃針對各篇專利加以詳細解讀，將各個專利申請的主要技術內容，加以剖析成技術研發人員更能了解的技術語言及層次之各種技術分析，若能妥善分析，對於所謂迴避設計 (Design Void)或想從事改良的研發人員是極為珍貴的資訊。

圖 2-2 為經營圖的一個例子：技術生命週期圖，此圖是出自於日本特許廳(JPO)在 2000 年介紹專利地圖所舉的例子 [8]，以申請人數與申請件數相對變化情形分析某個領域的興衰起迭，通常這類的管理圖是有一定分析的技巧，依不同時期的不同特性 [27] 區分為：

- 第一階段：技術萌芽期，申請件數與申請人數均較少。
- 第二階段：技術成長期，在這階段產業的技術可能有所突破或廠商對於市場價值看好，廠商持續投入，專利的申請件數與申請人數會出現急遽上升的情況。
- 第三階段：技術的成熟期，在這時期，其它廠商投入意願降低，只剩下少數的人繼續研發此一技術，因而申請件數與申請人數漸漸不再成長。
- 第四階段：技術退，淘汰期，在這階段，廠商抽取投資於研發的資源，退出這領域的廠商逐漸增加，使得申請件數與申請人數雙雙遞減。

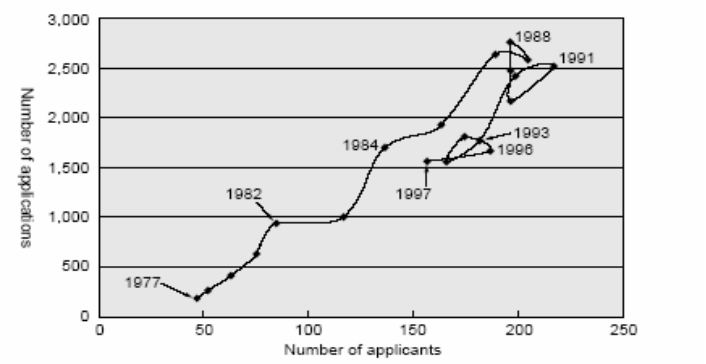


圖 2-2 技術生命週期圖 [JPO00]

即使技術生命週期圖的分析有軌跡可循,但要確切的劃分各時間仍需仰賴專家的幫忙。

圖2-3為技術圖的一種，稱為技術功效矩陣 [8]，所謂的技術功效矩陣是以特定的專利目的，彙整相關專利，分解其技術手段與達成功效，製成矩陣型態之統計表［24］。技術功效矩陣以各個專利資料所要達成的功效作為橫軸，將專利資料中所使用的技術手段作為縱軸。圖2-3 是與流體床焚化爐 (Fluid Bed Incinerators)有關的技術功效矩陣，此矩陣將技術分為四類：流體床燃燒法的特徵 (Fluid Bed Combustion Characteristics)、燃燒溫度控制 (Secondary Combustion Temperature Control)、燃燒混合控制 (Secondary Combustion Mixing Control)以及燃燒保留時間 (Secondary Combustion Retention Time)。而此技術主要的功效訴求為戴奧辛的分解 (Dioxin Decomposition only)、成本的降低、提高維護等。在每個方格中的一個圓點代表的是一篇專利說明書。

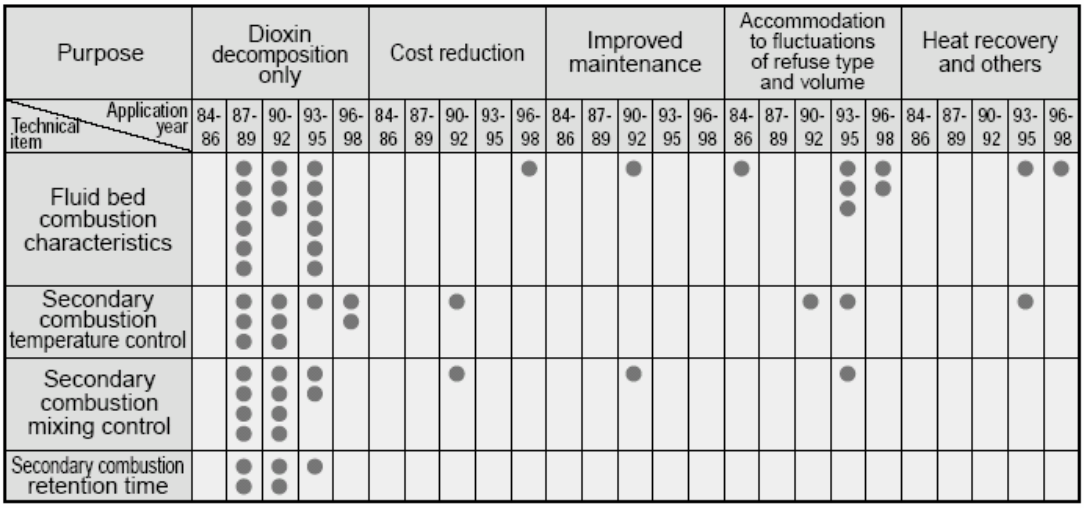


圖 2-3 技術功效矩陣表 [JPO00]

技術功效矩陣分析對公司而言是個相當重要的分析工具，研發者可從技術功效矩陣避免誤踩地雷，開拓研發的處女地，或研判某技術的潛在實力，因而可將研發的資源發揮至最大效用［26]。但其製作過程卻相當繁瑣。專利工作一開始，通常公司的法務部門會檢索相關的原始專利文章，並分配給研發工程師，以人工的方法填寫專利摘要表 (此表通常包含專利目的、達成功效以及技術手段等項目)；再由專利工程師根據所有的專利摘要表填寫技術功效矩陣，一份技術功效矩陣的完成需要大量的人力與時間。

現今專利資訊系統在專利地圖製作方面，大都能實作「經營圖」 [8][26]，目前專利地圖的研究主要在如何實作技術圖 [11][14][18]，讓使用者快速且經濟取得專利技術的資訊，以進行迴避設計、進行研發，讓使用者掌握更多專利技術。

在 [11]中，提出以循序規則探勘演算法 (Sequential Pattern Mining)，先探勘出足以代表專利文件的片語 (Key Phrase)，而後計算每年被公開的專利文件有多少專利篇數的標題或摘要包含這片語，並提供一套 Shape Definition

Language (SDL)查詢語言給使用者，可以讓使用者去選擇趨勢的走向，若使用者的 SDL 是要查詢成長的趨勢，系統更會將符合這趨勢的片語以折線圖的方式呈現予使用者。

在專利文件中，宣告 (Claim)是最重要的欄位，它代表著專利文件向法律訴求保護的技術核心所在。在 [15] 中，作者利用先前所做的研究，採用對 Claim 進行可讀性分析所擷取的名詞片語，以自動化的方式建立起類似技術功效矩陣的專利地圖，再經由專家研判。大部份專家對此研究是抱持肯定的態度，但覺得所建立的專利地圖結果太過粗糙。作者在結論中也提出，以目前的技術要做到完全自動化地產生技術功效矩陣圖是有所困難的，在最後作者針對自己所下標題 “Can Claim Analysis Contribute toward Patent Map Generation?” 以他們目前對宣告的分析結果，所回答的答案是 no。

考量到技術圖對使用者的便利但受限於自動化所產生的技術圖往往與使用者的需求有所落差，本論文企圖呈現的是半自動化的專利技術圖。

第二節　文字探勘 (Text Mining)

2.2.1 文字探勘簡介

隨著數位時代的來臨，許多知識隱涵在大量文字裡面，如網頁、新聞、專利文件等。文字探勘希望從龐大的文字中，擷取有意義的資訊，它結合了語意分析（Semantic Analysis）技術來對非結構性的文章進行分析，然後以有意義的方式對文章中所包含的概念，進行叢集 (Clustering)與分類 (Classification)，並

且製作索引 (Index)，以及結構化的方式呈現，讓使用者快速地利用索引來找到所需的資訊，藉以大幅降低閱讀所需的人力與提高閱讀的效率。

文字探勘包含了自動摘要 (Summarization)、自動分群 (Clustering)、自動分類 (Classification)等技術，在本論文中主要以文字探勘 (Text Mining)進行重要概念的萃取。

2.2.2 關聯式探勘 (Association Mining)

Clifton 在 2004 年針對新聞文件作主題偵測 (Topic Identification)時，採用 Apriori 演算法輔佐主題偵測的工作[5]，Clifton 藉著自然語言處理 (Natural Language Processing)，先標註文件中的人名、地名、組織名稱，每一篇文章視為一筆交易 (Transaction)，文章中的人名、地名與組織名稱視為項目 (Item)，目的在挖掘 (Mining)出那些人名、地名與組織名稱常一起出現於同一篇文件(即可能有關係的詞)，再進一步以 Apriori 所挖掘出來的項目集 (Itemset)進行分群的動作，以判斷主題 (Topic)。Clifton 在本篇文章的想法是：常常一起出現的詞 (Co-occurrence Words)會提供有用的資訊 。

Liu 在對網站內容進行比較時，在探討該如何呈現網頁內容時提出與 Clifton 相近的看法，Liu 在[12]中對了「概念」作了以下的定義：在網頁中，若有某些關鍵字的組合常常出現於同一個句子中，這些組合或許就是網頁想要強調、具有特別意義的資訊 。與 Clifton 不同的是，Liu 認為每篇網頁存在獨特想要強調的概念，如果把每一篇網頁視為一筆交易，會無法擷取出某篇網頁獨一無二的概念。所以 Liu 在進行關聯式探勘時，是以一篇網頁中所有句子的集合視為交易集合 (Transaction Set)，每個句子視為一筆交易，擷取出一篇網頁中重要的概

念。

第三節　統計檢定

本論文將 [16]中提的統計檢定方法應用在專利文件趨勢分析上，[16]是在辨識新聞文件中的主題，針對龐大的新聞文字進行主題偵測與追蹤 (Topic Detection and Tracking)，Swan 在 [16]中，利用列聯表 (Contingency Table)，如表 2-1 及方程式 2-1，進行統計檢定假設，希望能找出在語料庫 (Corpus)中有趣且具有意義的主題，在表 2-1 中f_0是代表被檢定的新聞語料庫中的一個字，a為在t_0時間裡所發表的文章中，曾出現f_0的文章篇數；b為在t_o裡的文章中，不含f_0的篇數；c為不在t_0的時段中，f_0曾出現的篇數；d則為不在t_0的文章數中也不包含f_0的篇數。方程式 2-1 中的N則是整個語料庫的文章篇數。

Swan在 [16]中，採用表 1 來檢定f_0與t_0是否具有關聯，在Swan的假設中，若有特別的事件發生時，在那個時間點與這事件有關的詞發生的頻率會異於往常，用 χ^2 檢定算出來的分數會較高；然而因 χ^2 是在測試兩個不同的變數有無顯著性的關聯，當f_0從未在t_o出現時，所得的分數也會偏高。所以使用 χ^2 時，必須限制a的大小，才會得到比較有意義的結果。

在 [16]中也提及其他的統計值，如 Fisher Exact Test 及 Expected Mutual Information Measure (EMIM)，但 Swan 在實驗結果中發現，Fisher Exact Test 所花的時間成本遠遠高於 χ^2，且 Fisher Exact Test 所得到的結果與 χ^2 排序的結果相似，所以在檢定詞與時間的關聯性時，Swan 是採用 χ^2 。

	f_0	$\overline{f_0}$
$t \in t_0$	a	b
$t \notin t_0$	c	d

表 2-1　列聯表 (Contingency Table)

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

方程式 2-1　卡方 (Chi-Square)檢定

第三章　專利趨勢分析系統

本論文提出一套輔佐使用者進行專利分析的工具，著重於提供使用者呈現技術名詞的專利技術圖。我們採用文字探勘技術挖掘出對使用者有意義的概念，再以統計檢定定義所謂的趨勢。本章在第一節描述系統的主要架構；第二到第四節詳細介紹模組的內部功能。

第一節　系統架構

圖 3-1 為系統架構圖，系統共分為三大模組：

1、文字分析 (Textual Analysis)：針對所要分析的專利說明書進行斷詞切字等前置處理，以利後續模組工作的發展。

2、統計分析 (Statistic Analysis)：經過前置處理的資料為這模組的輸入，利用文字探勘挖掘出有用的資訊，建構概念資料庫，最後利用統計檢定對這資料庫進行分析。

3、視覺化模組 (Visualization Module)：這模組主要的功能為提供分析結果予使用者，並透過使用者介面，讓使用者修正系統所挖掘的概念，使呈現出來的資訊更能體切使用者的需求。

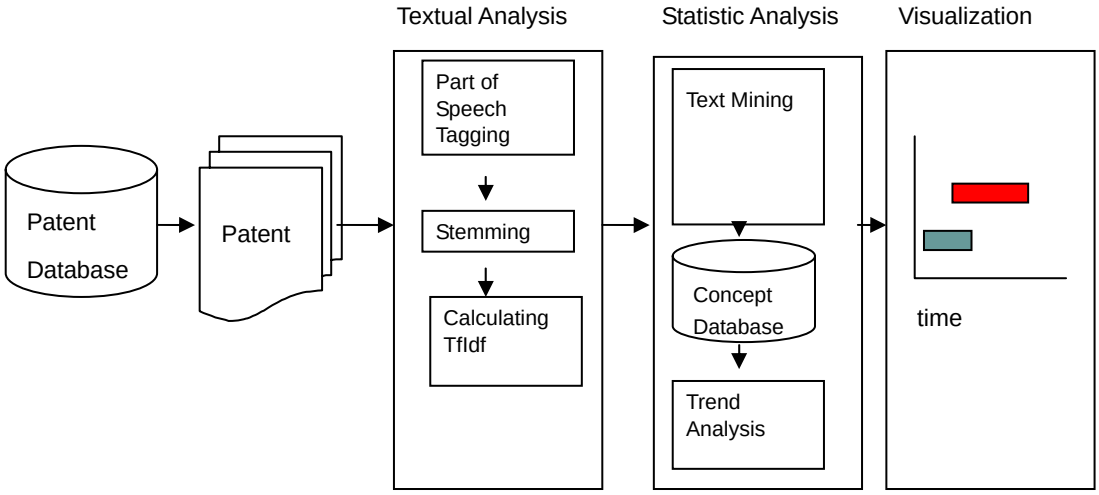


圖 3-1 系統架構圖

第二節　文字分析 (Textual Analysis)

第一個模組進行前置處理的工作，分別為針對處理資料進行格式轉換、斷詞切字及常用字 (Stop Words)的刪除、詞性還原 (Stemming)、以及 TF*IDF 的計算。藉著這些工作的進行，過濾一些雜訊，增進之後處理的效能及效果。

3.2.1 格式轉換

本論文系統所處理的專利文件來源為美國專利暨商標管理局資料庫 (USPTO)中的專利說明書。USPTO 裡的專利說明書皆為 html 格式的檔案，但 html 為一非結構性語言，所以在分析之前須做前置處理。首先當我們下載專利文件後，會先將 html 檔剖析轉換 (Parse)成 xml 檔以利後續分析。系統中 xml 檔儲存的資訊是專利編號、專利標題、申請日 (Apply Date)、公告日 (Issue

Date)、參考文獻 (Reference)、摘要、宣告以及專利說明。

3.2.2 斷詞切字及常用字的處理

處理完格式問題，後續動作是採用 NLPROCESSOR v. 3.8 [23]，對專利文件中的文字進行斷詞切字 (Tagging Of Speech) 的工作，標註 xml 檔裡面文字的詞性，並且為減輕後面計算的負荷量，會建立一個常用字的列表將專利文件中，不具資訊量的文字刪除。

3.2.3 形態還原（Stemming）

為避免有些字因為詞性的不同被視為不同的字，在前置處理中，我們會採用 Porter 所提出斷詞切字的演算法 [24]，進行詞性還原的工作，將字彙做形態轉換，避免意思相同的字卻只因字形而被看作是不同的字，最主要的目的是希望讓挖掘出來的結果更有意義。

Stemming：

- 去掉複數形式：maps->map, speeches->speech
- 去掉動詞變化：used->use, classified->classify

3.2.4 TF*IDF 的計算

TF (Term Frequency)為計算詞在文章中的出現頻率，IDF (Inverse Document

Frequency)則將詞曾出現的文章在語料庫中的分佈情況加以考量。TF*IDF的假設情況是，若某個詞在一篇文章出現的次數極高且又集中在某些文章出現，代表這個詞對這篇文章而言相當具有意義，帶有比較高的資訊量。在實作上我們設定一個門檻值 (Threshold)，只留下大於這門檻的值。方程式 3-1 為我們所採用的TF*IDF公式，為避免某些專利文章過長，致使每個字的TF都較其它篇專利文件高，分母是在做正規化 (Normalize)化。w_{ik}代表$Term_k$在$Document_i$的重要性，其中tf_{ik}代表$Term_k$在$Docuemnt_i$出現的頻率，n_k代表$Term_k$在語料庫中出現的文章篇數，t代表$Documnt_i$中相異字的字數，N為整個語料庫的文章篇數。

$$w_{ik} = \frac{tf_{ik} \log(N / n_k)}{\sqrt{\sum_{k=1}^{t} (tf_{ik})^2 [\log(N / n_k)]^2}}$$

方程式 3-1　計算 TF*IDF 公式

第三節　統計分析

這階段分為兩大部份，第一部份為利用文字探勘的技術，擷取能代表專利的重要概念，第二部份則是針對所擷取出來的詞，衡量詞與時間是否具有顯著性的關係。

3.3.1 概念的擷取

在概念的擷取上，我們針對不同的需求，作了不同的定義：

- 針對各篇專利文件，若一些詞在同一篇文章同一個句子中共同出

現的頻率超過一定的比例，則這些詞具有同樣語意，即為該篇的重要概念。

- 針對整個專利文件庫的字而言，分成特定 N 個群，希望每個群中的字義相似，每個群視為候選的重要概念。

針對不同定義的概念，我們實作了兩個不同的方法，企圖結合兩種不同定義的概念，提供使用者更多資訊 。

3.3.1.1 各篇專利所強調的概念

針對找出各篇專利所強調的概念，我們採用的是 Aprior 演算法，因為考量到專利文件習慣以詭譎的字眼描述其技術，所以不考量字的順序關係。有些字常一起出現在特定的視窗(Windows)，可能就是在描述相同事物的詞。當這些詞一起出現的頻率超過一定的次數可能就是作者想要強調的概念。

在實際應用上，我們做了以下的假設：一篇文章視為一個交易集合，以一個句子視為一筆交易，當有數個字出現在同一篇專利文章同一句子的比率超過某個門檻，即視為代表專利重要的概念。例如一篇專利說明書有四個句子 T1、T2、T3、T4，經過前置處理後剩下 A、B、C、D、E 五個字。從字曾出現在哪些句子，可以建立一個交易集，其中每筆交易表示每個句子包含哪些字，如表 3-1 中的第一筆交易 T1 代表這個句子包含 A、B、C 這三個字，第二筆交易 T2 代表這個句子包含 B、C、D 這三個字，其他交易依此類推。

句子	句子上出現的字
T1	A,B,C
T2	B,C,D
T3	C,E
T4	A,C

表 3-1　以一個句子為一筆交易的交易集

假設門檻值為 50%，現在有四筆交易，表示最小支持度 (Minimum Support) 為 2，從表的交易經過關聯規則探勘可以得到一些項目集，若項目集的支持度高於預設的最小支持度，則可稱為頻繁項目集 (Frequent ItemSet)。出現在同一個頻繁項目集中的詞，就是我們所定義可能代表專利文件的概念。如表 3-2 所示，在表 3-2 的例子中可以得到 2 個頻繁項目集。

群編號	頻繁詞群	支持度
001	A、C	2
002	B、C	2

表 3-2　經常出現於同一個句子的頻繁項目集

但以頻繁詞群代表專利的重要概念會產生一些問題，因 Apriori 演算法的特性，長度為 n 的頻繁項目集其任意長度的子集也一定是頻繁項目集，如果把所有大於最小支持度的頻繁項目集都提供給使用者，所提供的資訊會有所重覆。所以在這步驟我們會進行長詞優先的處理，如果長度為 n-1 的頻繁項目集已被包含在長度為 n 的頻繁項目集，我們只會提供長度為 n 的頻繁項目集。

3.3.1.2 語料庫的概念

針對找出整個語料庫所探討的概念，我們採用的方法是對專利文件資料庫的字加以分群，利用 K-means 演算法[7]把語料庫裡的字區分為 *K* 群，每個群代表者這語料庫中的概念。主要的想法是，如果兩個字各自的鄰居都是非常相似的字，則這兩個字的意涵會有一定相似的語意關係，所以現在字的向量是與鄰近的字相關，主要相關處理步驟如下：

1、選定待分群的字：若把整個語料庫的字都進行分群的處理，計算負荷量會變得相當大。同時，有些字所夾帶的資訊量並不大，勉強分群的結果提供給使用者的資訊也相當有限。考量專利文件的特性，在「說明」欄位的文字是專利文件中對技術描述最完整的欄位，「摘要」欄位則是對專利內容做概括性的描述。一般而言，名詞帶給使用者較多的資訊，所以我們所選定分群的詞為說明及摘要中的名詞，「宣告」欄位是整個專利說明書的核心所在，在宣告中的用詞不僅在保護自己的發明同時也在闡述自己的技術，所以除上述的名詞之外，在宣告裡的動詞、形容詞皆為我們所選定的字。

2、字的表示方式：針對所選定的字，以向量表示。向量的維度為整個語料庫的字濾掉常用字及進行形態還原(Stemming)後相異字的個數，當某個字與我們所選定的字$Term_i$在同一個句子上，不是常用字且矩離長<=d的字時，即會在代表$Termi$的向量上，填入符合條件的字的TF*IDF。方程式 3-2 為$Term_i$的向量表示法，n為整個語料庫去除掉常用字經過形態還原後相異字的個數，而且我們會建立一個視窗，如果$Term_j$與$Term_i$曾一起出現在這個視窗裡，則w_j為$Term_j$的tfidf，否則w_j=0。

$$Term_i = (w_1, w_2, w_3, ..., w_n)$$

方程式 3-2　$Term_i$在K-means中的表示法

3、以 K-means 進行分群。

A、隨機挑 k 個字，將之分別視為 k 個族群的群中心。

B、定義群中心後，對每一個字的向量 x，計算二向量間的餘弦 (Cos) 定義相似程度，尋找與之最接近的群中心，並將 x 代入該族群。

C、確認分群結果是否已趨於穩定，若未穩定，重新定義中心點，若已穩定則進入 D。

D、傳回分群結果。

4、當分群處理結束後，再計算群與時間的關聯度。

3.2.2 統計檢定

本論文針對趨勢所做的統計檢定，主要是參考 Swan 所用在新聞文件上進行主題偵測與追蹤的公式-方程式 3-3，測試詞在某個時間點顯著性的程度 [16]。

在方程式 3-3 中，t代表在語料庫裡各篇專利說明書向USPTO提出申請的申請日，f_0代表專利中出現我們現在所要檢定的重要概念，$\overline{f_0}$ 為專利不曾出現我們所要檢定的重要概念，a為在t_0這段時間曾出現f_0這概念的專利篇數，b為在t_0這段期間不曾出現f_0這概念的專利篇數，c為不在t_0這段期間曾出現f_0這概念的專利篇數，d為不在t_0這段期間申請的專利也未出現f_0這概念的專利篇數。

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

方程式 3-3　卡方 (Chi-Square)檢定公式

	f_0	$\overline{f_0}$
$t \in t_0$	a	b
$t \notin t_0$	C	d

表 3-3　列聯表 (Contingency Table)

藉著公式的處理是在協助我們分析哪些概念在某個時間點曾熱烈的被討論，以呈現現今語料庫中的熱門討論議題，若某個概念在特定的時間點曾被熱列的討論，即會得到一個偏高的 χ^2 分數。表 3-4 及表 3-5 為利用 χ^2 檢定詞與時間的例子，依表 3-4 的資料，利用卡方檢定，得到表 3-5 的結果。在所舉的例子(見表 3-4)，因在一月至三月時，共 18 篇文件，地震出現在 10 篇的文件中，其它時間區段，地震分別各出現於 5 篇文件，所以檢定出來地震在一~三月所得到的分數明顯的較其它時間區段偏高，表示地震對於這語料庫而言，在一~三月間熱烈被討論。針對一~三月這時間區段而言，地震的 χ^2 分數是 10.78，颱風的 χ^2 分數是 1.49，故在這段期間，地震比颱風受到更熱烈的討論，更能代表一~三月的討論主題。

我們企圖依上述的檢定方法，找尋已在某個時間點被熱烈討論的概念，呈現概念的脈絡予使用者，一覽特定領域討論主題的概況。

	地震	颱風	總篇數
一月~三月	10	5	18
四月~六月	5	11	18
七月~九月	5	5	18
十月~十二月	5	5	18

表 3-4 卡方(Chi-Square)檢定的舉例資料

	一月~三月	四月~六月	七月~九月	十月~十二月
地震	10.78	0.84	0.84	0.84
颱風	1.49	13.40	1.49	1.49

表 3-5 舉例資料的檢定結果

第四節 視覺化模組

本節討論的是本系統的介面與功能，圖 3-2 為使用者決定專利地圖的介面。在這個介面上，使用者可以選擇他們想要分析的時間區段，當按下分析按鈕 (Analysis)時，系統會把分析的結果以視覺化的方式呈現給使用者。在圖 3-2 的畫圖區，是呈現所分析的結果，橫軸為時間，縱軸為 χ^2 的分數，每個長條圖的上面為這個群的命名。

本系統僅以輔助的角度提供使用者進行專利分析，使用者可依據系統所提供的概念，自行修改更符合使用者需求的概念，以使得呈現出來的資訊更符合

使用者需求。

以圖 3-2 為例，目前使用者感興趣的時間為 1993 年 7 月 13 號到 2003 年 1 月 14 號，在這段期間具有顯著性的有六個群，而使用者若對群的命名不滿意，本系統也提供一個修改的介面，見圖 3-3。使用者可以在右邊的 Concepts 上找到群的名字，在群的名字上按 F2，即可以使用者認為較恰當的名字予以命名。同時，我們也給予使用者權限變更分群的結果，使用者可以將目前較不適當的成員刪除。

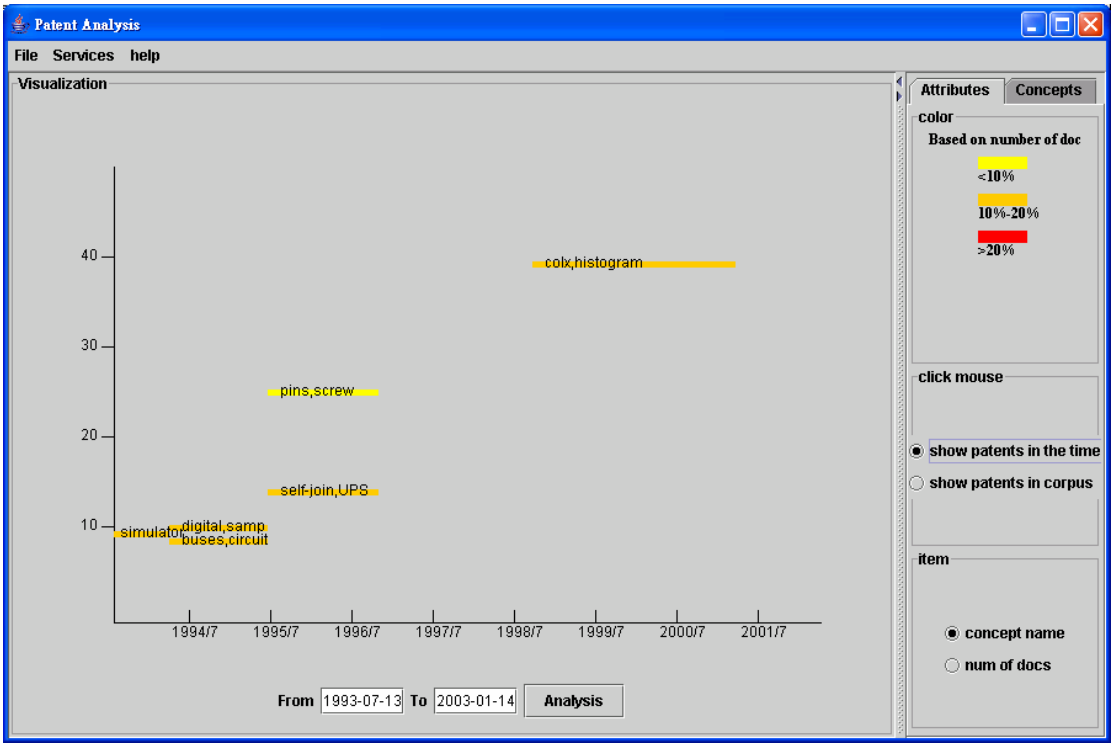


圖 3-2 專利分析的介面

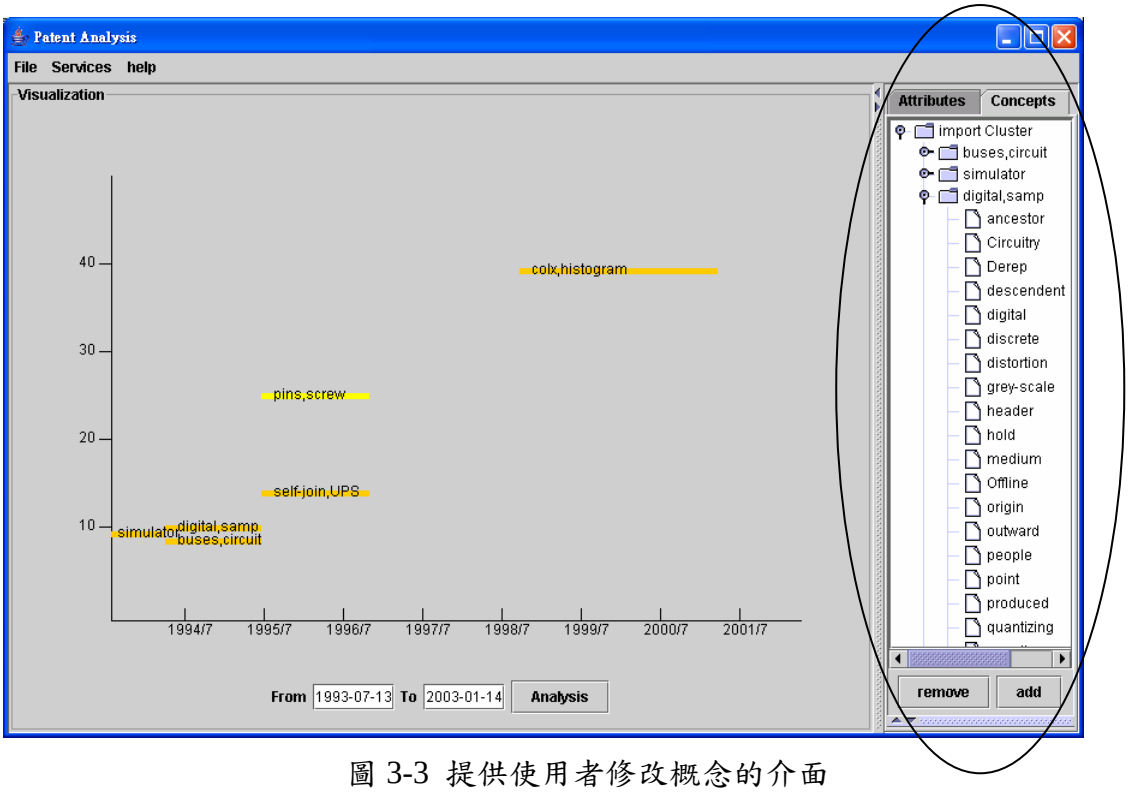


圖 3-3 提供使用者修改概念的介面

第四章　個案探討(USPTO)

本章將以個案探討方式，介紹本論文所提專利趨勢分析系統的實作細節，以及目前所得到的結果。第一節裡說明資料的來源和一些必要的前置處理，使其成為符合系統需要的資料型態，第二節討論以不同方式擷取概念所得到的結果，第三節則呈現所分析的趨勢結果，第四節是結果分析。

第一節　實驗資料

4.1.1 實驗資料來源

本論文的實驗資料來自於美國專利暨商標管理局(USPTO)的資料庫，以專利號碼 6847978 為主並搜集與此相關的專利文件，共 159 篇專利說明書，判斷相關的依據是以有無被此專利引證 (Citation)為主。引證關係如表 4-1，以二層的引證關係，共得到了 159 篇專利說明書。

6847978	5588150
6847978	5630120
6847978	5668987
6890658	5806061
6847978	6012054
6847978	6052689
6847978	6065007
6847978	6092062
6847978	6108658
6847978	6178449
6847978	6205441
6847978	6327587
6847978	6343288

5580150	4714995
5580150	4881166
5580150	5058000
5580150	5142470
5580150	5161158
5580150	5239577
5580150	5247664
5580150	5257366

表 4-1　專利引證 (Citation)關係表

在搜集的專利說明書裡，全部的專利說明書都與專利編號 6847978 有直接或間接的引證關係，專利編號 6847948 的專利說明書主要在改善對資料庫執行結構性查詢語言(SQL)後，資料庫內部執行方案 (Execution Plan)選擇問題，為提昇資料庫的執行效率，這篇專利發明人提出利用統計方法，使系統依統計的結果選擇一個較佳的執行方案。而與專利編號 6847978 有直接或間接引證關係的專利說明書，大部份也與資料庫、結構性查詢語言有關，有少部份是在討論記憶體、網路及電路等。

所下載的專利說明書是在 1976 年 10 月 4 日到 2001 年 12 月 26 日間向 USPTO 提出申請的，但其分佈頻率卻是相當分散，1990 前申請的專利說明書只有 31 件，有 128 件的專利說明書是在 1990 年以後申請。

4.1.2 前置處理

在 USPTO 所下載的專利說明書是 html 格式，html 檔案為非結構性的語言，在處理上比較不方便，所以在前置處理的時候，會先把所下載的 html 檔轉換成 xml 檔的格式，xml 檔所儲存的資料為專利編號、申請日、公告日、參考文獻、摘要、宣告及說明。圖 4-1 是我們所下載的專利說明書，圖 4-2 為轉換的 xml 檔。

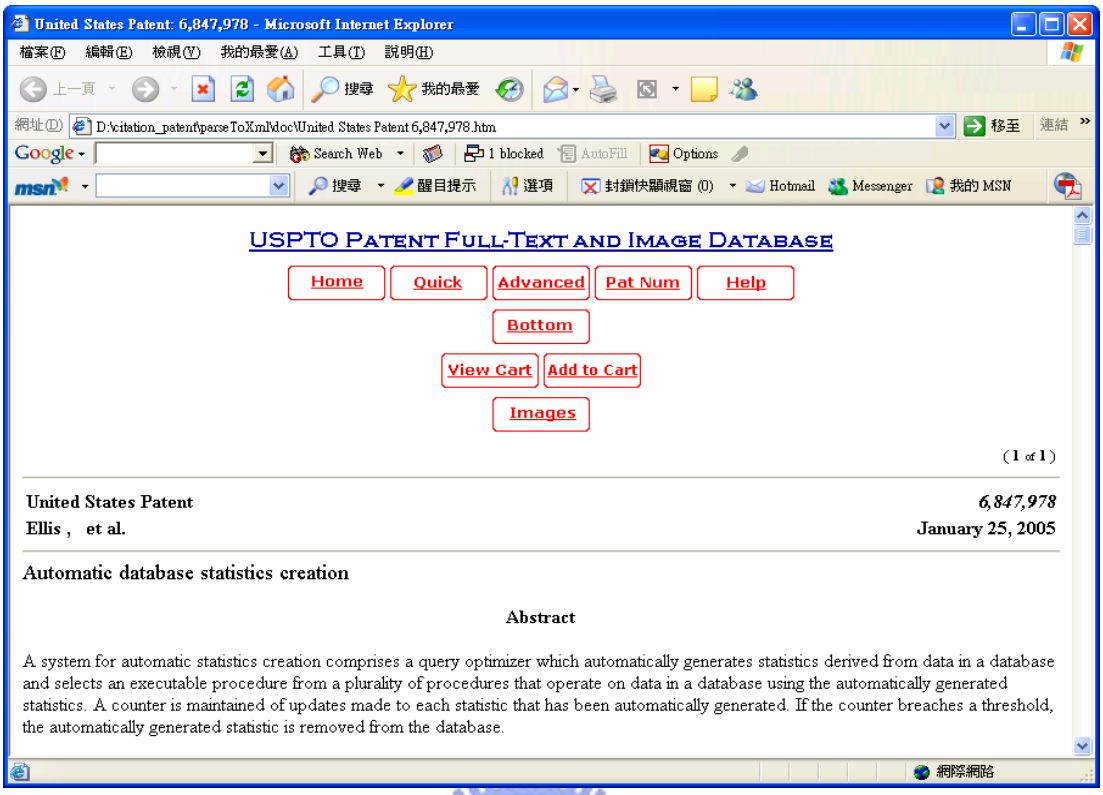


圖 4-1 USPTO 裡的專利資料 (專利編號：6847978)

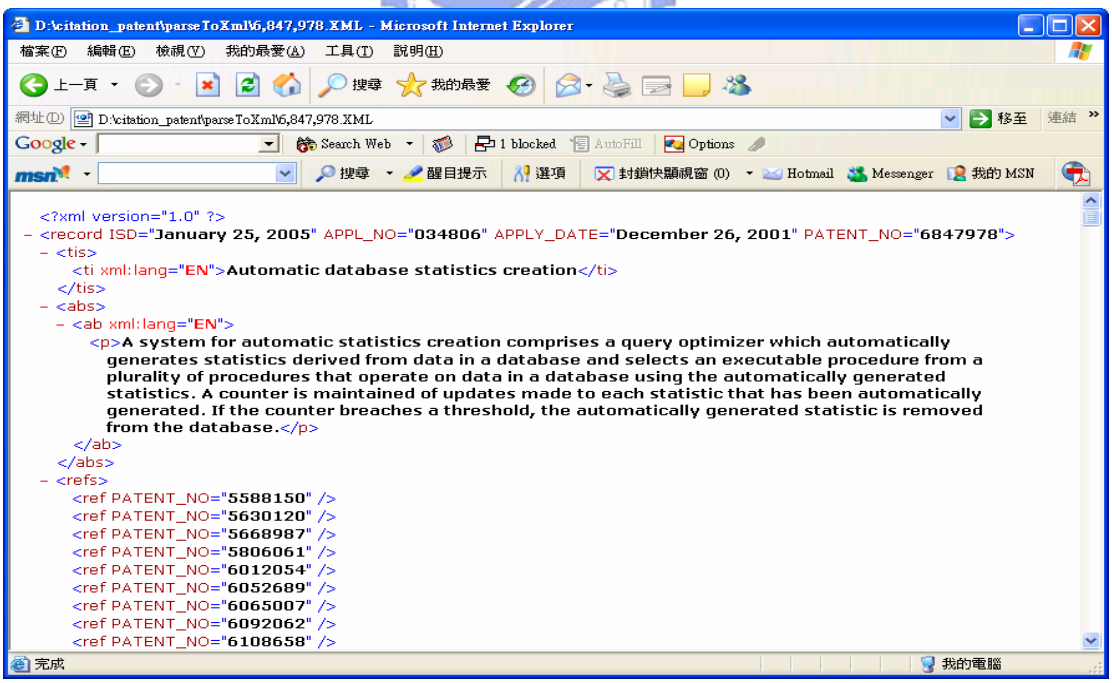


圖 4-2 經過轉換後的 xml 檔 (專利編號：6847978)

經過格式轉換的處理後，接下來的工作分別為：

1、採用工具 NLPROCESSOR v. 3.8 針對已轉換成功的 xml 檔進行斷詞切字的工作

2、建立一張常用字表，將一些不重要的雜訊移除，如”a”、”the”之類的字。

3、採用 Porter Stemming Algorithm 解決詞性不同會被視為不同字的問題，進行形態還原的動作。

4、計算每個字的 Tf*Idf ，並設定一個 Tf*Idf 的門檻，將小於門檻的詞刪除。

5、建立一個關連式探勘所需要的交易集 (如圖 4-3)，在圖 4-3 每一行代表本篇專利的一個句子，以 ”,” 區分句子裡的詞

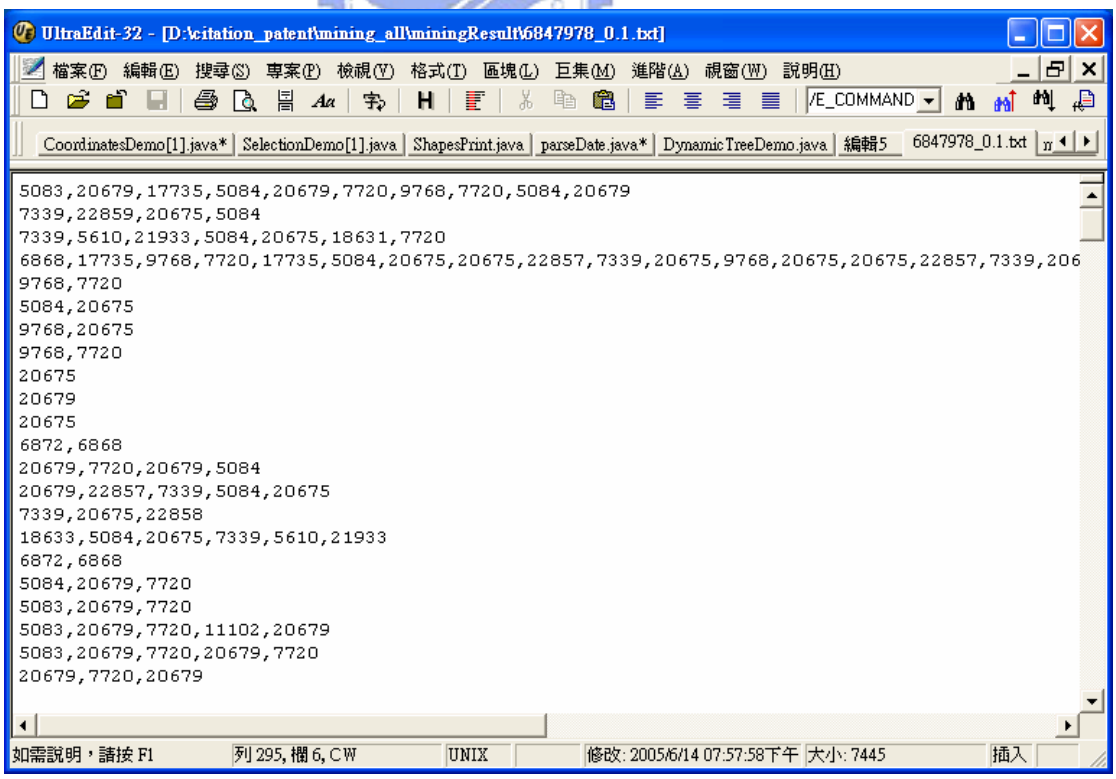


圖 4-3 關連式探勘所需要的交易集

第二節　重要概念的擷取

在本論文針對概念的擷取實作了兩個方法，方法一採用關連式探勘企圖找出各篇專利說明書在探討的重要概念，方法二是利用 K-means 演算法對整個語料庫進行分群的動作。

4.2.1 關連式探勘

採用關連式探勘的方法，找出代表各篇專利重要概念，其想法來自於[12]，Liu 認為若有些字常常一起出現在某個句子，這些組合就是作者想要強調的概念。

將每一篇專利說明書當作一個交易集，每一個句子視為一筆交易，而其中的項目是經過前置處理的詞。表 4-1 是專利編號 6847978 的專利說明書在門檻為 0.1，經過關聯式探勘處理所探勘出來的部份結果。由於專利編號 6847978 是在探討資料庫中有關結構查詢語言實際執行的情形，此專利發明人提出利用統計的方法輔佐資料庫自動去選擇效率較佳的執行程序，表 4-1 所列長度為 1 的頻繁項目集裡的”system”、”step”雖然與本篇專利所提出的技術無直接關連，但長度為 2 的頻繁項目集的詞還蠻符合本篇專利所探討的主題。

1-頻繁項目集	statistic, automatically, query, table, generated, data, step, system, plan, execution, database
2-頻繁項目集	generated statistic, plan database, plan execution, plan statistic, execution statistic, database statistic

表 4-1　關連式探勘結果 (support:0.1)-6847978

表 4-2 為另一篇專利編號 5345585 在門檻為 0.1 所挖掘出來的結果，這篇專利也是與結構性查詢語言有關的主題，藉著 KBZ 這演算法，改良資料庫目前結合 (Join)運算的順序。

1-頻繁項目集	sequence, scheme, optimization, KBZ, current, cost, Algorithm, order, join
2-頻繁項目集	sequence join, scheme current, scheme order, scheme join, optimization join, KBZ Algorithm, KBZ join, current join, current order, cost order, cost join, Algorithm join, order join,
3-頻繁項目集	scheme order join, current order join

表 4-2　關連式探勘結果 (support:0.1)-專利編號 5345585

依前兩個例子來看，採用關連式探勘對找出各篇專利所在探討的觀念具有一定的幫助，但主要的缺點則是所挖掘出來的概念彼此之間的意涵可能會有重疊 (Overlap)的傾向，如專利編號 5345585 所挖掘出來的概念”sequence join”與

概念”order join”。

進行關聯式探勘，當採用最小支持度門檻為 0.1 時，針對這 159 篇專利說明書，總共挖掘出 3,500 個不同的概念，進行長詞優先處理後仍然有 1,057 個不同的概念，進行長詞優先處理後平均每篇專利的概念有 6~7 個概念。表 4-2 為所探勘出來各個長度的頻繁交易集在進行長詞優先處理後個數的差異，藍色為進行長詞優先前，紫色為進行長詞優先後的結果，原本長度為 1 的頻繁項目集的個數有 764 個，進行長詞優先後剩下 243 個，因在此語料庫中頻繁項目集的長度最大值為 6，所以不論有無進行長詞優先，頻繁項目集長度為 6 的詞皆是 35 個。

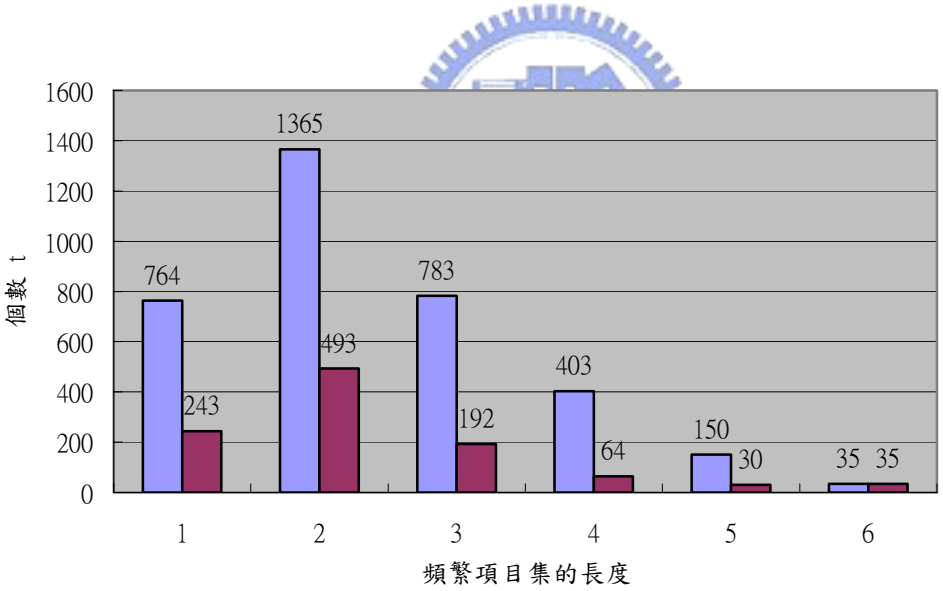


表 4-2　頻繁項目集各長度詞的個數統計

4.2.2 詞的分群

詞的分群在文字探勘處理中，一直是個專業的研究主題，在本論文是採用

K-means 對語料庫中的詞做分群，鑑於時間成本考量以及使用者並不會對語料庫中的每個詞都感興趣，在進行 K-means 之前，會先進行挑選的動作。所挑選的規則如下：

1、宣告中的形容詞、動詞及名詞。

2、摘要及說明的動詞與名詞。

上述挑選的規則因考量在宣告中的每個用語都有撰寫者的涵義在裡面，以及動詞與名詞對使用者而言比較有意義，依這規則，我們將 1,233 個字分佈在 50 個群。在這 50 個群裡面有 10 個群所包含的字大於 100 個，有 14 個群所包含的字介於 50~100 個之間，14 個群包含的字在 6~50 之間，而小於 5 個字的群有 8 個。雖然目前會有過大或過小的群，但某些群所包含的詞的意思還是十分相近，如表 4-4 中的各個詞幾乎出現在不同的專利文件裡，但所代表的意思似乎都與時間有關；而表 4-5 中的詞雖然雜帶有雜訊，但應可歸類為與影像處理有關。

群 1	Ck, duration, microseconds, time-length

表 4-4　K-means 的分群結果 1

群 2	colors, color-component, counting, image, interpolation, page, pixel, region, replicas, screen, subparts

表 4-5　K-means 的分群結果 2

4.2.3 比較分析

依前述實作兩個不同定義概念的結果，發現到當關連式探勘適合擷取各篇專利說明書所強調的概念與語意，經過關連式探勘所擷取的概念雖然會有語意重疊的情況發生，會挖掘出過多的詞，但所挖掘的結果還蠻符合各篇專利所探討的主題。而 K-means 所分出來的群是整個語料庫所在探討，比較廣泛的概念。

在進行趨勢分析時，受限於專利領域裡，不同的專利說明書作者習慣以不同的用語形容其技術核心，如某一篇專利文件 A 的重要概念包含”SQL”，另一篇專利文件 B 作者針對查詢語言(Query)的結合 (Join)運算進行改良，B 的作者在這份專利文件中比較常出現的是 ”Join”，在一般的認知上對 ”Join” 做改良其實相當於對 ”SQL” 的執行程序進行改善，但在專利文件 B 中， ”SQL”幾乎從未出現，所以若在計算 ”SQL” 這個詞出現在文章的篇數時，專利文件 B 卻往往不會被算進去，會使得結果有所偏差。在進行趨勢分析的時候，會以 K-means 所得到的概念進行趨勢分析，以關連式探勘的結果提供給使用者各篇專利所強調的概念。

第三節　趨勢分析與介面

本節是在介紹前述處理所擷取出來的概念，在經過本論文的趨勢分析得到的結果。目前群的命名先以人工暫定，使用者可以在介面上得知群的組成分子，再結合本身的專業對群的命名做更改。在這個專利地圖上，橫軸代表的是時間，而縱軸代表的是卡方的分數。

圖 4-4 為使用者查看 1993-7-13 到 2003-1-14 中間熱門的趨勢所得到的結果，得到 6 個群，在這 6 個群中，有 5 個群各自出現的專利篇數佔整個資料庫的 10%~20%。本語料庫中的專利說明書，大部份都是在探討有關 ”資料庫” 方面的改良，但在 1994 與 1995 年間的時候，與影像處理有關的詞在這段期間卻是具有顯著性，由此現象，我們或許可以猜測在那段時期左右，曾興起一波影像處理與資料庫結合的研究熱潮。

群的命名是以人工的方式,在群裡挑選兩個比較能代表這群大致上所討論議題的詞，但除此之外使用者也可依本身的專業在圖 4-5 的介面上，更改群的命名，或影響分群的結果，使分析的結果能更貼切使用者需求。

除此之外，使用者可以在呈現概念的長條圖上，按左鍵列出在這分析時間點包含這概念的專利文件或這語料庫中包列這概念的專利文件，圖 4-6 為使用者在右邊的平面，mouse click 的區塊，選擇”show patents in the time”的選項後在左邊的平面長條塊”cok, histogram”按右鍵所得到的結果。系統會呈現這段時間點上，包含這概念的專利文件編號，使用者可依據這專利編號，進一步查看這專利特別強調的概念，也就是關聯式探勘所得到的結果。

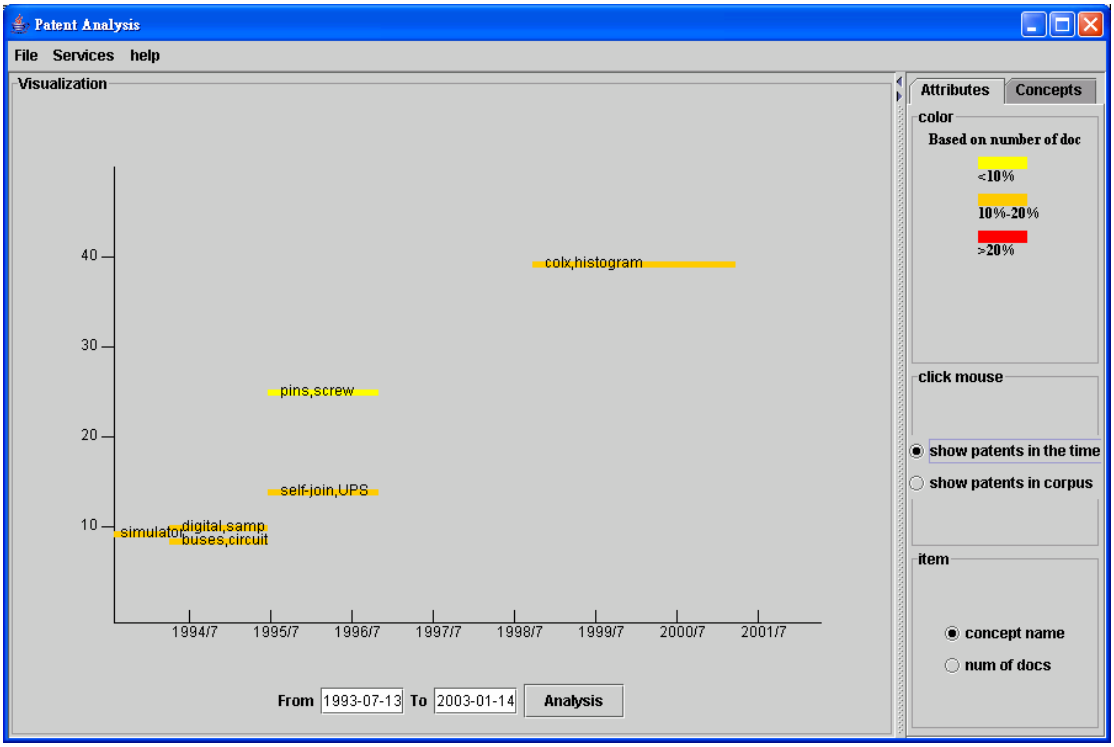


圖 4-4 呈現專利趨勢分析結果的介面

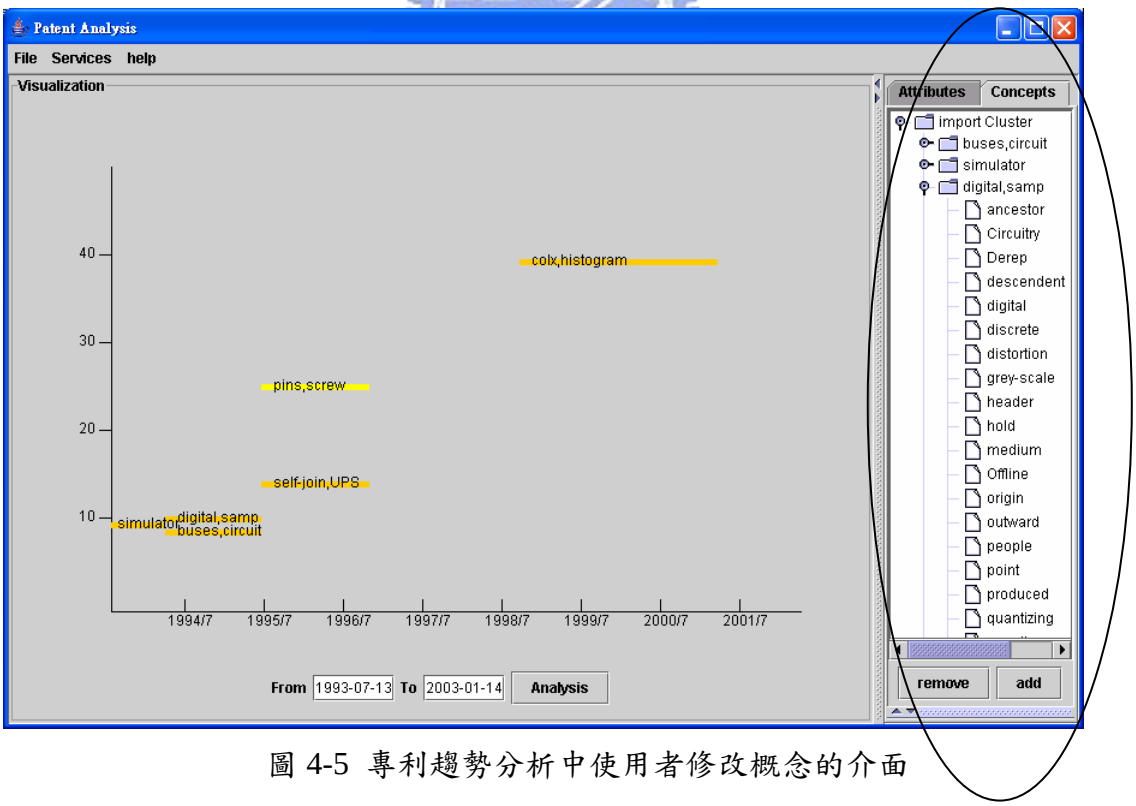


圖 4-5 專利趨勢分析中使用者修改概念的介面

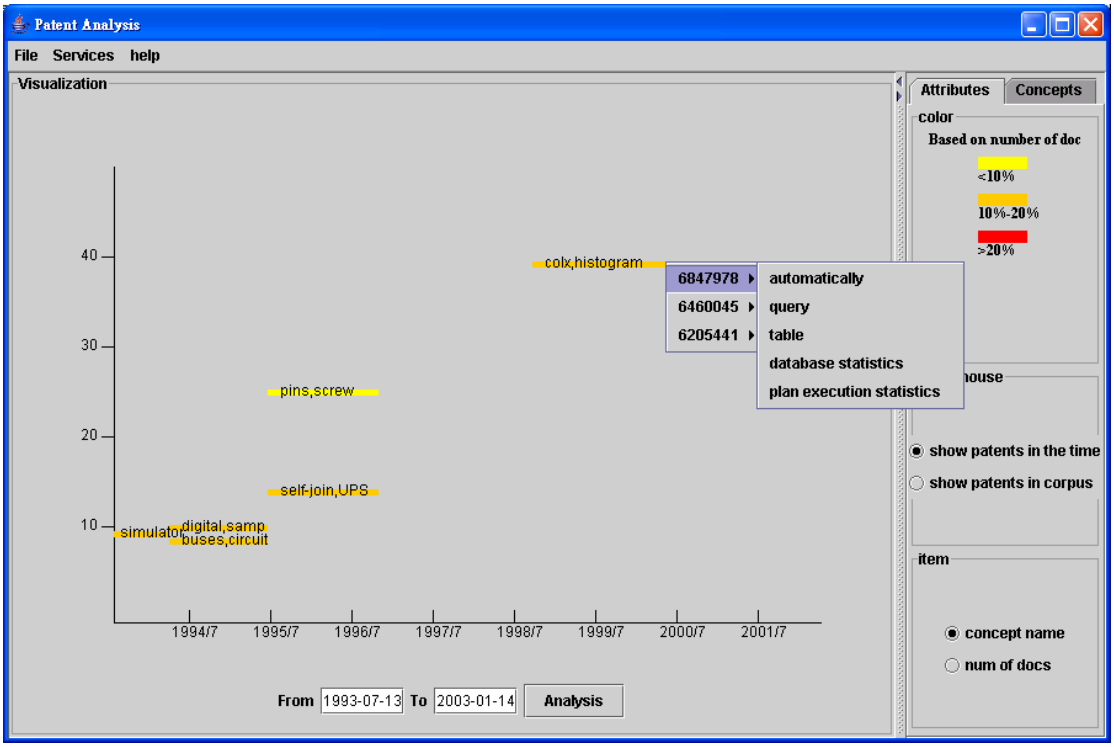


圖 4-6 呈現包含選取概念的專利文件

第四節　結果分析

相較於挖掘出代表各篇的概念再做趨勢分析而言，先對相同意義的字做分群再進行專利趨勢分析似乎能得到更佳的效果。現在趨勢分析的結果深受分群結果、分群命名所影響，雖然採取的 K-means 針對語料庫進行分群的工作，在有些分群中，所探討的議題還蠻貼近的，然而 K-means 同時也會產生過大或過小的群，過大的群會導致不同意思的字被集中在一起，對我們分析結果造成相當大的困擾。

日後工作需要再改善目前現在分群結果，如能改良分群的效果再進行專利趨勢分析，相信可以得到一個更佳的分析結果。

第五章　結論與未來研究方向

本章總結本論文並提出未來的研究方向，第一節討論與本篇論文所實作系統相關的議題，第二節則探討未來相關的研究方向。

第一節　結論

本論文提出一套專利趨勢分析系統，輔佐使用者進行專利分析，而能快速的取得情報。針對此項目標，本系統主要提供以下的功能：

1、利用文字探勘的方式將代表各篇專利的概念挖掘出來，使用者在進一步閱讀前，能先對專利主題有個大概的認識。
2、將專利中的概念與時間軸相結合，以專利地圖的方式呈現給使用者，可讓使用者一覽特定領域概念的演進情況。
3、提供互動式的介面，使用者能以本身的專業對我們所分群的結果作更改。

雖然礙於語料庫的選定、或對詞分群的效果不彰，本論文所作的趨勢分析並未得到一個相當良好的效果。但考量到使用者的需求，一個能呈現專利技術的專利地圖對使用者而言是個相當方便的工具，使用者可以快速的取得情報。

第二節　未來研究方向

依實作本系統的心得，我們認為下面有幾點值得作為專利趨勢分析以後改善的方向：

1、降低關聯式探勘所挖掘概念的數量：根據實驗結果，發現到採用關聯式探勘針對每篇專利文章挖掘出來的概念還蠻符合各篇專利所在探討的主題，但以此方法進行探勘所產生最大的問題是，所挖掘出來的詞似乎過多了，個別文章所探勘出來的概念彼此之間似乎有重疊的情況，如果能有適當的方法解決這種情況，所探勘出來的詞應該更有意義，可以做為各篇專利文章的關鍵字。

2、知識本體(Ontology)的建立：在研究中發現，專利趨勢分析最大的問題是不同的發明人習慣以不同的詞去描述相同概念的技術，所以常會發生二篇專利所挖掘出來的概念技術應該是有相似的地方，但依作者寫法不同，在計算 A 概念出現頻率時常會低估 A 概念”實際”出現在這語料庫的次數。用不同的詞語表現相同的技術的現象在專利中是相當普及的，如果在進行專利分析時，能請專家進立相關領域的知識本體，再建立每篇專利的關鍵字，則不管在進行專利趨域分析或專利分類工作，對專利分析的幫助是相當大的。

3、詞的分群：在本系統的實作過程中，曾利用 K-means 來對語料庫中的字分群，但未達到一個良好的效果，日後工作還是需要加強在詞的分群，如果能找到一個適當的分群法，對目前的分析結果會有一定程度的提昇。

4、統計檢定值：目前針對趨勢的定義是以統計方法 χ^2 檢定詞與時間的關係，若有概念在一段期間頻繁地被提起，我們即認為對那段時間而言，

這些概念是被熱列討論的技術,所以時間的畫分會影響到我們所做的檢定。對 χ^2 而言，畫分一個適切的時間進行檢定是相當重要的。但各個領域技術的生命週期都不相同,要找到一個最適切的切割區段似乎不容易。之後要對專利文件進行趨勢分析,也許可以選擇一個更適切的統計檢定值。

5、即時性的回饋系統：在本系統，我們提供給使用者一個互動式的介面，使用者可以藉著本身的專業更改群的名稱;除此之外,使用者也可以對群內的成員進行更名、刪除的動作,但當使用者執行這動作後,系統除了立即改變目前介面上的內容外,該如何重新分析目前的資料,將資料結果立即呈現給使用者,是以後系統發展的一大重點,除此之外,系統該如何利用使用者更名、刪除的資訊以提昇分群的準確度也是日後的一大課題。

以上幾點都是目前的專利分析系統尚未完成的工作，若能一一完成，相信對專利趨勢分析的結果有很大的助益。

參考文獻

[1] R. Agrawal and R. Srikant, Mining-Sequential Patterns, "In Proceedings of the International Conference on Data Engineering, Taipei, Taiwan, March 1995

[2] K. Ahmad and A. Althubaity," Can Text Analysis Tell us Something about Technology progress?, " *The ACL-2003 Workshop on Patent Corpus Processing*, Tokyo, 2003.

[3] J. Allan, R. Papka, and Victor L avrenko, "On-line new event detection and tracking, " *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, MElboume, Australia, 1998.

[4] Y.M. 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

[5] C. Clifton, R. Cooley and J. Rennie, "TopCat: Data Mining for Topic Identification in a Text Corpus," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 16, No. 8, pp. 969-964, August 2004.

[6] Y. He and S. C. Hui, "Mining a Web Citation Database for author co-citation analysis," *Information Processing and Management*, vol. 38, No. 4, pp. 491-508, 2002

[7] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2000

[8] Japan Patent Office Asia-Pacific Industrial Property Center, JIII, " Guide Book for Practical Use of "Patent Map for Each Technology Field", " 2000

[9] K.-K Lai and S.-J Wu, "Using the Patent co-citation approach to establish a new patent classification system," *Information Processing and Management*,

vol. 41, No. 2, pp.313-330, 2005

[10] L.S. Larkey, “A Patent Search and classification system,” *Proceedings of the forth ACM conference on Digital Libraries on Digital libraries*, California, pp. 197-187, 1999

[11] B. Lent, B. Agrawal and R. Srikant, “Discovering Trends in Text Databases,” *Proc. 3 rd Int Conf. On Knowledge Discovery and Data Mining*, California, 1997

[12] B. Liu, Y. Ma and P.S. Yu, “Discovering Business Intelligence Information by Comparing Company Web Sites,” *Web Intelligence*, pp. 105-125, Springer, 2003

[13] T. Mori, M. Kikuchi and K. Yoshida, “Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR systems,” In *Proceedings of the 19th International Conference on Computational Linguistics*pp, pp.688--694, (2002).

[14] S. Morris, C. DeYong, S. Salman and D. Yemenu, “DIVA: a visualization system for exploring document databases for technology forecasting,” *Computer & Industrial Engineering* , vol. 43, No. 4, pp. 841-862, 2002

[15] A. Shinomori and M. Okumura, ”Can Claim Analysis Contribute toward Patent Map Generation?” *Working Notes of NTCIR-4*, Tokyo, 2004

[16] R. Swan and J. Allan, ”Automatic generation of overview timeliness,” *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athen, Greece, 2000.

[17] A. Tang, “Text Mining: The state of the art and the challenges,” *Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data mining*, pp.65-70, Beijing, 1999, pp.65-70

[18] H. UCHIDA and A. MANO, “ Patent Map Generation using

Concept-based Vector Space Model, ” *Working Notes of NTCIR-4*, Tokyo, 2-4 June, 2004

[19] D. Vogel, “Using Generic Corpora to Learn Domain-Specific Terminology, “*Workshop on Link Analysis for Detecting Complex Behavior*, August 27, 2003

[20] P.C. Wong and W. Cowley and H. Foote and E. Jurrus and J. Thomas, “*Visualizing Sequential Patterns for Text Mining*, ” *Proceedings IEEE Information Visualization*, Salt Lake City, Utah, Oct 8 - Oct 13, 2000

[21] ACL Workshop on Patent Corpus Processing, Sappora, Japan, 2003. Available at http://acl.ldc.upenn.edu/acl2003/patent/index.htm.

[22] ACM SIGIR Workshop on Patent Retrieval, Athens, Greece, 2000. Available a t http://research.nii.ac.jp/ntcir/sigir2000ws/.

[23] NLPROCESSOR v 3. 8. Available at http://www.infogistics.com/demos/

[24] Porter Stemming Algorithm. Available at http://tartarus.org/~martin/PorterStemmer/

[25] 尹居中，” 人工膝關節專利分析” Available at http://designer.mech.yzu.edu.tw/article/articles/design/(2000-06-01)%20%A4H%A4u%BD%A5%C3%F6%B8%60%B1M%A7Q%A4%C0%AAR.htm

[26] 連穎科技，http://old.learningtech.com.tw/

[27] 陳達仁、黃慕萱， 專利資訊與專利索引 ，文華圖書館理，民 91

[28] 國立交通大學智慧財產權中心，「專利檢索/專利分析」訓練課程

[29] 曾元顯， “專利文字之知識探勘：技術與挑戰” 現代資訊組織與檢索研討會，2004