國立交通大學

資訊科學與工程研究所

碩 士 論 文

專 利 文 件 之 自 動 分 類 研 究

Automatic Classification of Patent Documents

研 究 生：林蘭綺

指導教授：梁 婷 教授

中 華 民 國 九 十 五 年 六 月

專利文件之自動分類研究

研究生：林蘭綺　　指導教授：梁婷 博士

國立交通大學資訊科學與工程研究所

摘要

專利文件分類是專利文件分析一項重要的工作。目前在重新調整分類結構和文件分類都需要以人工的方式進行，因此提高自動化分類的正確性，將有效地幫助專利研究人員快速地完成工作。在分類階段，以往利用 tf-idf 計算詞彙的權重，進行向量空間模型分類法。在本論文中，我們加入 Entropy 的概念提出新的修正詞彙權重計算方法，以降低因過少的詞彙群組數對文件分類造成的負面影響。我們分別取以主類別(語料 1)和子類別(語料 2)分類的美國專利文件進行分類研究，經過權重修正之後，在語料 1 中，當減少至 200 個詞群數時，調和平均值可由 0.387 提升至 0.735；而語料 2 在 8836 至 2000 之間的詞群數，調和平均值從 0.35 可提升至 0.42。

Automatic Classification of Patent Documents

Student: Lan-Chi Lin Advisor: Tyne Liang

Institute of Computer Science and Engineering

National Chiao Tung University

Abstract

Patent document classification is the work of first importance for the patent document analysis. Instead of patent document analysis by the human manually, the automatic technology of patent document classification can make processes dealt with by patent analysts more effective and fast. In the past document classification phases, it computed the weights of words by tf-idf and classified by vector space model. In this thesis, we proposed new method of counting term weight by adding the concept of Entropy to reduce the negative effects which fewer term clusters result in. We used two patent document sets classified by class(corpus 1) and subclass(corpus 2), respectively. After modifying the weight of words in patent documents, the result showed that the F-score of corpus 1 can yield 0.735 from 0.387 with 200 term clusters. The F-score of corpus 2 was also up to 0.42 from 0.35 when the number of term cluster ranged from 2000 to 8836.

致謝

首先我要由衷感謝的是我的指導教授梁婷老師，在碩士班的修業期間，老師對我的包容和教誨讓我這份論文總算可以發表出來，除了不斷的指引論文的方向外，老師也時常提供他個人專業的經驗教導，讓我在語言表達，上台報告，寫作技巧上也增進不少，使我受益匪淺。感謝我的爸爸媽媽多年辛苦的栽培，給予默默的支持，讓我能夠順利完成學業。

我也要感謝實驗室的伙伴們，包括吳典松學長、朱俊榮學長、鄭建富學長、陳淳齡學姐、劉正義學長、龔自良學長、以理、怡嘉、立泓、傳堯、曉茹、守益、善均、盛興，一起努力，藉由互相的鼓勵，使我能更加積極的面對漫長的研究生活。

最後我要感謝我的男友瑜宏和室友們穎劭、怡君、周彥，他們在我碩士研究這段時間裡給予的支持和照顧，一起分享快樂、難過的事，陪我一路走來。

目錄

圖目錄

表目錄

第 1 章　　緒論

專利文件記載專業技術發展的成果，為技術革新和知識資產的具體象徵，是極為珍貴的技術文獻。技術創作人可以藉由申請專利或權利認證，便可享受專利權保護的權利。根據「世界智慧財產權組織（World Intellectual Property Organization，WIPO）」的調查，專利文件涵蓋全球每年大約 90%-95%的發明或創造。

美國為世界上專利的主戰場，圖 1-1為近 20 年間美國專利局每年核准的專利文件數量，從圖中可看出在 1985 年時只有 7 萬多篇，2004 年為有 18 萬多篇，成長相當迅速。而台灣在美核准的專利文件數量也不斷在增加，如表 1-1所示，已是 10 年前 3.5 倍的成長，所占的比例也越來越高。除了美國，根據台灣智慧財產局的統計顯示，如表 1-2，申請與核准數量也不斷在生長。

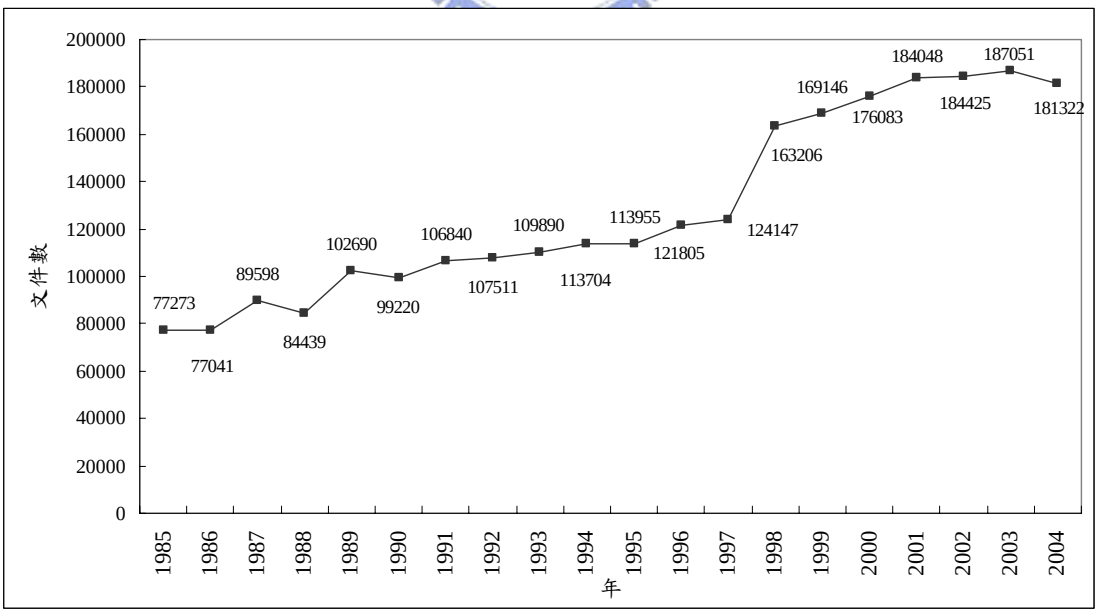


圖 1-1：近 20 年間美國專利局每年核准的專利文件數量

年份	台灣核准專利數	該年總專利數	所占比例(%)	世界排行
1995	2087	113955	1.83	7
1996	2419	121805	1.99	7
1997	2597	124147	2.09	7
1998	3805	163206	2.33	5
1999	4526	169146	2.68	4
2000	5806	176083	3.3	4
2001	6545	184048	3.56	4
2002	6730	184425	3.65	4
2003	6676	187051	3.57	4
2004	7207	181322	3.97	4

表 1-1：近 10 年台灣在美核准的專利文件數量

年度	申請	核准
84	43,461	29,707
85	47,055	29,469
86	53,164	29,356
87	54,003	25,051
88	51,921	29,144
89	61,231	38,665
90	67,860	53,789
91	61,402	45,042
92	65,742	53,034
93	72,082	27,717

表 1-2：近 10 年台灣智慧財產局專利申請及核准數統計數量

專利分析的主要目的在於收集、分析相關研發的專利研究，瞭解相關競爭產品的專利發展技術。分析專利的初始階段工作包括進行專利類別、相關詞彙、時間、地區...等確定範圍。初級資料數量可能非常大，且有許多的內容並非和計畫內容有直接的關係，必須再經由研究人員判斷，過濾出和計畫有關係的專利次級資料。

篩選出次級資料後，開始進行專利資料的解讀，包括專利之目的、欲解決的問題、研發技術及專利的涵蓋範圍，此步驟稱為專利分析（專利地圖）。此外還

有許多其他的數據分析，如技術領域歷年趨勢分析、發明人或申請公司歷年件數趨勢分析、國家歷年件數趨勢分析...等，可以讓研究人員有更整體的概念。

專利文件範圍含蓋各類專業技術領域，因此需要將特定主題或相關技術領域的專利文件做妥善的歸類。現有的專利分類是根據發明創作的技術內容建立組織結構和索引。希望透過有效的資料管理，合理區分專利的內容和技術範疇，讓專利審查人員或是一般專利的閱讀者能夠利用完備且快速的查詢系統，精確地查詢和檢索出來相關的專利文件，以方便爾後的應用，尤其在專利文件繁多的情形下，便利性顯得更加重要。

由於各國技術發展的差異性，因而產生各種以不同應用角度所建立的分類系統，現有主要專利分類系統包括：

- 國際專利分類，IPC[1]（International Patent Classification）

  為「世界智慧財產權組織（World Intellectual Property Organization，WIPO）」所採用，每五年就會重新調整一次分類架構，為最多國家所使用。分類為階層結構，分成五個不同的等級：8 個部（Section）、120 個主類（Class）、628 個次類（Subclass）、69000 個主目（Main Group）和次目（Sub Group）。台灣和大陸的專利也是以此分類為依據。

- 美國專利分類，USPC[2]（United States Patent Classification）

  為「美國專利局」在 1830 年創立，是目前歷史最悠久的專利分類系統，每隔一段時間就有新的分類系統，為動態調整之分類系統。USPC 之分類表（Class Schedule）中紀錄 400 多個主類別（Class）和近 16 萬個次類別（Subclass）。

- 歐洲專利分類，ECLA[3]（European Patent Office Classification）

  「歐洲專利局（European Patent Office）」內部用來檢索的分類系統，ECLA 分類在 IPC 的五階分類上做了七階的延伸，並且讓每個目

[1] www.ipc.org
[2] www.uspto.gov/main/patents.htm

（Group）的專利文獻量保持在一百文件以內。

- 日本專利分類，JPC[4]（Japanese Patent Classification）

  由「日本專利局」以 IPC 記號後加上展開記號或分冊識別記號，稱為 FI（File Index），另外還有日本專有的 F-Term（File Forming Term），以 FI 為基礎再多角度劃分，如用途、材料…等，目前 F-Term 有 2900 左右的技術主題分類。

因為專利文獻的數量龐大又文件篇幅冗長，加上涉及主題廣泛，所要處理的詞彙相當繁多，因此在進行專利文件分類之前，透過詞彙分群技術的應用將可提高文件內容的相似度，並且減少文件表示法的索引詞量。在本篇論文中，我們分別取以主類別(語料 1)和子類別(語料 2)分類的美國專利文件進行分類研究，採用 Baker and McCallum[1998]提出的詞彙分群方法 Distributional Clustering。並且提出改良的詞彙權重計算，加入亂度（Entropy）的概念，考慮詞彙在類別之間分佈的情況，進行向量空間模型分類法，修改原來完全依賴頻率（tf-idf）的作法，將改善詞彙群組數過少時所造成的缺陷，以利於詞彙分群發揮最大的作用，更加減輕文件分類過程中處理大量詞彙的負擔，和減少執行時間，達到使用者檢索上的便利性。

實驗結果顯示，詞彙分群方法在不影響文件分類正確率的情形下，語料 1 可由 2995 減少至 1500 個詞群數，皆是接近 0.78 的調和平均值；而語料 2 在 8836 至 2000 之間的詞群數，均有高於 0.35 的調和平均值，確實可有效減少詞彙群組數。而經過權重修正之後，在語料 1 中，當減少至 200 個詞群數時，調和平均值可由 0.387 提升至 0.735；而語料 2 在 8836 至 2000 之間的詞群數，調和平均值從 0.35 可提升至 0.42。此外，我們減少至 600 個詞群數時，語料 2 由 0.2 提高至 0.35，達到在未修正權重前和使用 2000 個以上的詞群數相同效果。

[3] http://www.european-patent-office.org/index.en.php
[4] www.jpc.de

第 2 章　　相關研究

2.1. 文件分類

分類的研究中主要探討有：文件的表示型態，相似度文件計算，分類方法，分類的結構...等。文件的表示型態中，向量空間模型為最常見的文件表示法，一篇文件以一個特徵向量表示，其文件特徵為文件的單一詞彙或是關鍵詞彙[Zhang, 2005]、雙連字串、三連字串...等，作為文件的表示，向量中的數值(權重)代表特徵在文件中的重要性，而計算詞彙權重的參考依據其特徵在文件中的頻率、出現的文件數、文件長度、該特徵的長度...等。

Hammouda 和 Kamel[2004]提出一個新的文件表示結構"DIG（Document Index Graph）"，運用在網頁內容上。DIG 為節點和邊集合表示的有向圖，節點代表文件內的單一詞彙，而二個詞彙在文件中有前後連續出現順序關係則建立連線，句子在 DIG 中會形成一條路徑，如此反覆直到所有文件都建立完成。DIG 上記錄連續詞彙的長度和次數，可看出文件之間相同連續詞彙的運用，有利於偵測文件之間相似的連續詞彙。其計算文件之間的相似度除了考慮出現的頻率外，還加入彼此相同的連續詞彙的長度，以增加連續詞彙的重要性。作者也比較使用單一詞和連續詞彙做為特徵對分類的影響，實驗結果顯示，二項同時搭配比只依據單一詞可以提昇 7.4%~60.6%的調合平均值和減少 9.1% ~ 64.6%的亂度。

許多分類法都因應不同的情形而產生。規則式的分類法為決策樹（Decision trees），產生容易了解的樹狀結構，可以對分類整理出一條條簡單的規則，但決策樹分類法只適合用於非連續數值的分類上。類神經網路（Neural Network）是類似於大腦神經網路的方式運作，透過反覆訓練的學習，直到對於輸入都能正確對應到所需要的輸出，但學習過程非常耗時。貝氏（Bayesian）分類法是假設各屬性彼此間是互相獨立的，來用機率預測分類的結果，因為其計算過程較其他分

類容易，所以更適用於龐大的資料上。Instance-based 的 *k*-Nearest Neighbor（*k*NN）Classification，優點在於不用事先訓練的過程，且在訊練資料少的情形下，仍有不錯的準確度，但在進行分類時反而需要花費較多的時間。

其他還有利用資料探勘（Data Mining）技術的分類法[Fung et al. 2003; Wang et al., 2005]。Fung 等人[2003]提出 Frequent Itemset Hierarchical Clustering（FIHC）的方法，取出文件的高頻項目組（Frequent Itemset），利用文件之間具有相同的高頻項目組，根據高頻項目組的支持度對文件進行樹狀階層式的分群，再由下而上計算兩兩子群組之間的相似度，合併相似度高於門檻值的子群組，對其樹狀分類結構做修剪，直到最後留下的合理的分類數量。而 Wang 等人[2005]則先將文件中的高頻項目組和支持度建立詞彙圖（Term Graph），每個節點為高頻項目組中的其中一個詞彙，高頻項目組裡的詞彙節點就會建立連線，形成一個圓圈，每個連線都會記錄其支持度。再依據所建立的詞彙圖，利用計算網頁等級的 PageRank，把詞彙圖中每個節點當作一個網頁的概念，計算文件和類別之間的相似度進行分類，其各類的正確率介於 83~100 % 之間。

分類的結構多是平行式或樹狀階層式。Lertnattee 等人[2004]提出一個以多方面分類（Multidimensional Text Classification，MDTC）的概念，因為平行或階層式的分類架構都可轉換成多方面分類。實驗結果以不同分類法：Nearest Neighbor Algorithm、Naïve Bayesian Algorithm、Centroid-Based Algorithm 進行比較，除了比其他分類架構有較高的正確率外(平行式：68%，階層式：68%、71%，多方面式：72%)。

2.2. 詞彙分群

詞彙分群是文件分類中一個基本的工作，以減少文件向量的維度。常見的方法有利用人工建立的索引詞庫，和統計式的分群法則。統計式的分群法包含隱含語意索引和 Distributional Clustering。

以索引詞庫為依據[Lin, 1997]，包含同義、反義和上/下義…等關係，將認為

是同義或相似度高的關係詞彙形成群組，這樣的群組都帶有相同的語義概念。一般詞彙的索引詞庫如 WordNet [Miller, et al., 1990]，包含了將近 200,000 個英文字義及其語意關係。其他專業的索引詞庫，如 SNOMED[5]。文件中的文字可藉由索引詞庫所給予的詞彙關係，彼此以相近的語義連接起來，如此一來也可分辨出一詞多義、一義多詞的情形，但其缺點是需要一個跟所訓練語料相關的詞典，若其訓練語料的太過專業或含蓋的主題太過廣範，甚至索引詞庫太久沒有更新，都會使此方法無法發揮最好的效用。Kang 和 Lee[2005]運用文件中的詞彙建立詞彙鏈（Lexical Chain），詞彙之間以在 WordNet 中有本身、同義詞、上/下位詞和附屬關係而建立連結，透過連結的關係調整原來的權重計算，所連結的關係越多，整個詞彙鏈所提高的權重也會越多，增加在文件中的重要性，改善文件分類的正確性。

隱含語意索引（Latent Semantic Indexing，LSI）[Deerwester et al., 1990] 可描述詞彙和文件意義的整體關聯性，利用詞彙在訓練文件出現頻率的向量空間矩陣計算縮減文字的維度。Dhillon 等人[2001]提出利用此方法配合 Spherical k-means 群集演算法，大幅提升群集的效率。

Baker and McCallum [1998]提出 Distributional Clustering（DC），利用詞彙在每個類別之間分佈的情形，合併相似的機率分佈曲線，進行詞彙分群。在有 20 個類別新聞語料的分類實驗中，實驗結果可將 50,000 個詞彙分成 50 個群組，只讓正確率減少 2%，大大的減少執行時間和資源浪費。和其他四種現有用來減少文字向量空間的演算法做比較，有同樣是進行詞彙分群方法的隱含語意索引、Class-based clustering 和以特徵詞選取為概念的 MI（Mutual information）和 Markov-blanket method，DC 表現都比其他的方法效果都來得好，但其結果只適用於和訓練語料相同主題範圍的文件分類上。Chen 等人[2005]針對此方法在群組過後分配的不平均的結果進行修正，群組數減少時，正確率隨之減少的情形減緩，實驗在 100 個群組數時，原來的正確率已不到 0.6，經過修正之後，正確率提升至 0.7，提高 10.6%的正確率。

[5] SNOMED：Systematized Nomenclature of Human and Veterinary Medicine，是為了滿足醫學資料處理的需求而產生的索引詞典，包含大約 150,000 個索引詞彙。

此外，Mandhani 等人[2003]提出文件和詞彙同時進行群組的方法，稱為 Rowset Partitioning and Submatrix Agglomeration（RPSA），利用所有文件的向量所組的矩陣 M，根據矩陣內的權重值尋找相似的詞彙及文件，這些相似的詞彙和文件為子矩陣 S，必需符合 S 的密度（其詞彙權重的平均）大於 M 的密度，子矩陣 S 內所代表的詞彙和文件就可形成一個群組，以階層性的方式進行群組，形成文件和詞彙階層性的分類結構。用 4 個不同大小和不同類別數的語料和 k-means 分群法比較，衡量標準為純度（Purity），RPSA 的實驗結果在其中 3 個語料提昇 3~30%的平均純度，而第 4 個語料雖然比 k-means 來得低，但不到 1%平均純度的差異。

2.3. 專利文件

2.3.1. 美國專利文件

美國專利文件是以發明創作經「美國專利局（United States Patent & Trademark Office，USPTO）」審核通過後所發行的證書，目前涵蓋從1971年起所有美國專利內容，約有200多萬份專利文件，其包含三種種類的專利證書：「發明專利（Utility Patent）」、「新式樣專利（Design Patent）」和「植物專利（Plant Patent）」，表 2-1列出各種專利類型及其簡短說明。美國專利分類主類別編號“D01”至“D32”為新式樣之分類；主類別編號“D99”為新式樣雜類之分類；主類別編號“PLT”代表植物專利文件之分類；發明專利的主類別編號從“2”到“987”，其編號中間並不連續設有空號，是為往後增加新類別時所使用，且大部份的專利都為發明專利。

專利類型	說明
發明專利 （Utility Patent）	創造或發現一種新的、有用的程序（Process）、機械（Machine）、合成物品（Article of Manufacture）、製造方法（Composition of Matter）、有用的改良或是具有某種實用價值。授予自申請日起20年的專利年限。
新式樣專利 （Design Patent）	發明一種具有創新性（Novelty）、對產品原創的外觀上特殊（Nonobvious）設計或有美化的作用（Ornamented or Aesthetic in Nature），並不一定要具有實用功能。授予自申請日起14年的專利年限。
植物專利 （Plant Patent）	創造或發現無性生殖、有性生殖培養任何獨特的或新種的植物。授予自申請日起20年的專利年限。

表 2-1：美國專利的種類

美國專利文件範例和各欄位說明如附錄一 （為了節省篇幅，內容均為節省略過）。美國專利文件全文為超文件標示語言（HyperText Markup Language，HTML）格式。

UPC 是美國專利文件的分類標準，為全世界專利分類最細最多的分類，都是專業人員以人工方式進行分類。總共包含 468 種主要類別（Class），表 2-2 顯示部分專利的主類別，每一主分類可再細分出許多的子分類（Subclass），子分類也是以階層式的分類，約有 160,000 個，包含主類別階層達 14 層之多。專利文件內以“主類別/子類別”的方式表示所屬類別，其中圖 2-1顯示主類別“2”的部分子類別，如“2/2.16”表示分類類別為“Apparel/Having an insulation layer”。專利文件可同時屬於多個主分類和子分類。

Class 2 - APPAREL

Class 4 - BATHS, CLOSETS, SINKS, AND SPITTOONS

Class 5 - BEDS

Class 7 - COMPOUND TOOLS

Class 8 - BLEACHING AND DYEING; FLUID TREATMENT AND CHEMICAL MODIFICATION OF TEXTILES AND FIBERS

Class 12 - BOOT AND SHOE MAKING

Class 14 - BRIDGES

Class 15 - BRUSHING, SCRUBBING, AND GENERAL CLEANING

Class 16 - MISCELLANEOUS HARDWARE

Class 19 - TEXTILES: FIBER PREPARATION

Class 23 - CHEMISTRY: PHYSICAL PROCESSES

Class 24 - BUCKLES, BUTTONS, CLASPS, ETC.

…

Class 930 - PEPTIDE OR PROTEIN SEQUENCE

Class 935 - GENETIC ENGINEERING: RECOMBINANT DNA TECHNOLOGY, HYBRID OR FUSED CELL TECHNOLOGY, AND RELATED MANIPULATIONS OF NUCLEIC ACIDS

Class 987 - ORGANIC COMPOUNDS CONTAINING A Bi, Sb, As, OR P ATOM OR CONTAINING A METAL ATOM OF THE TO 8TH GROUP OF THE PERIODIC SYSTEM

Class D01 - EDIBLE PRODUCTS

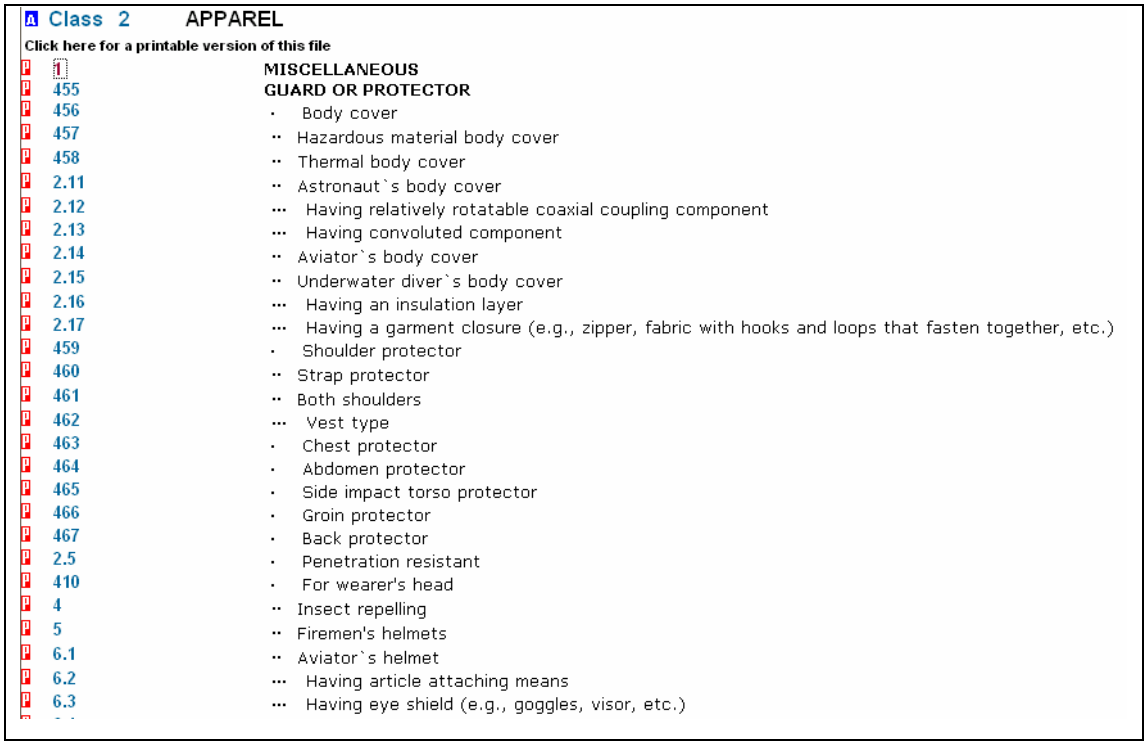
…

Class D34 - MATERIAL OR ARTICLE HANDING EQUIPMENT

Class D99 - MISCELLANEOUS

Class PLT - PLANTS

表 2-2：美國專利主類別部分範例



Class 2 APPAREL

Click here for a printable version of this file

1	MISCELLANEOUS
455	GUARD OR PROTECTOR
456	· Body cover
457	·· Hazardous material body cover
458	·· Thermal body cover
2.11	·· Astronaut`s body cover
2.12	··· Having relatively rotatable coaxial coupling component
2.13	··· Having convoluted component
2.14	·· Aviator`s body cover
2.15	·· Underwater diver`s body cover
2.16	··· Having an insulation layer
2.17	··· Having a garment closure (e.g., zipper, fabric with hooks and loops that fasten together, etc.)
459	· Shoulder protector
460	·· Strap protector
461	·· Both shoulders
462	··· Vest type
463	· Chest protector
464	· Abdomen protector
465	· Side impact torso protector
466	· Groin protector
467	· Back protector
2.5	· Penetration resistant
410	· For wearer's head
4	·· Insect repelling
5	·· Firemen's helmets
6.1	·· Aviator`s helmet
6.2	··· Having article attaching means
6.3	··· Having eye shield (e.g., goggles, visor, etc.)

圖 2-1：主類別“2”的子類別部分範例

2.3.2. 專利文件分類研究概況

專利文件的分類和一般的文件分類不同導因於專利文件具有下列特性：

1. 涵蓋的主題範圍廣大。
2. 龐大的文件數，且每一篇文件的篇幅冗長。
3. 專利文件有固定的結構，如標題，摘要，請求項...等。
4. 階層式的分類架構，且分類細，子分類之間的相似度高。
5. 每一篇專利文件可屬於不同階層和多個分類。
6. 所使用的用詞非常專業，一般的詞典無法涵蓋所有的詞彙，專業的詞典也只能對應部份類別文件的專業詞彙。
7. 存在許多的詞彙是作者自創的，使用現有的詞典無法辨別。

專利文件研究中，有許多方面的研究，其中有針對專利文件中的欄位的研究探討[李駿翔, 2003; Richter and MacFarlane, 2005]。

李駿翔[2003]利用資料探勘的文字知識發掘技術和向量空間模型，以tf-idf計算詞彙的權重，再以餘弦函數計算測試文件和各個類別之間的相似程度，決定其分類類別。語料選擇「基因轉殖生物」技術之相關美國專利文件資料，總共408筆，測試專利文件中不同的欄位內容對於分類結果的影響，其中以欄位“Title”搭配“Summary”或“Description”的分類結果可達到46%的正確率，但其中欄位“Description”的資料量是欄位“Summary”的數倍，二者的正確率卻相差不遠。而其它欄位“Title”搭配“Abstract”包含的資料量過少，欄位“Abstract”通常只是概要性的描述無法作為有效依據；欄位“Title”搭配“Claim”內容為了能擴大其專利發明權利的解釋範圍，所敍述的內容會過於含糊和不夠詳細。

Richter and MacFarlane[2005]除了利用文件的詞彙外，加入 metadata 的資訊來提高專利文件的正確率，作者主要加入專利文件內其他的欄位資訊，如發明人、國際分類編號…等欄位，收集專利合作條約（Patent Cooperation Treaty，PCT）2001 ~ 2002 年的專利文件，以“Gazette Classification”的 6 個分類，tf-idf 計算詞彙權重搭配 *k*NN 分類方法，正確率可從 70.8%提昇至 75.4%。

還有許多專利的研究上是希望詞彙的數量，減少執行的負擔。Chakrabarti 等人[1997, 1998]建立一個階層性的專利分類系統，提出 Fisher’s discriminant method 的選擇特徵詞（Feature Words）的計算方法，主要考慮詞彙在文件中的平均變異數比率，以去除許多大量會干擾分類的詞彙，再以貝式定理（Bayesian Algorithm）做階層性分類。在美國專利文件資料庫中抽取 12 個在第二層子類別作分類，平均分類在第一層子類別下，每個類別均有 307 ~ 361 之間不等的文件數，所需的詞彙只占原來的文章 12%~18%之間，平均的召回率為 66%。

Kin等人[2005]把專利文件分類的進行過程分成二個主要的步驟：特徵選取（Feature Selection）和文件分類方法。選用欄位“Abstract”和欄位“Description”的文件內容為分類依據。在特徵選取中詞彙的權重考慮詞彙的頻率（TF）、詞彙出現的文件數（TF-ICF）和詞彙出現的類別數（TF-ICF），均需設定門檻值來

篩選詞彙權重的範圍。比較文件分類的方法：*k*NN、MEM（Maximum Entropy Modeling）和SVM（Support Vector Machine），實驗結果發現SVM在較少資料量時會表現比其他二種分類方法好，卻較費時。*k*NN的表現都比MEM好，因此作者較推薦使用*k*NN的分類方法。

另外其他研究，Larkey[1998, 1999]建立一個專利文件查詢和分類系統工具，有提供線上使用者介面，包含自然語言（Natural Language）的查詢系統和欄位選項等，使用者可輸入專利文件的發明名稱、專利編號、發明人...等或輸入關鍵詞彙查詢相關文件。在美國專利文件中擷取主類別“395”的子類別“2.09”下的所有專利文件做為測試語料，抽取欄位“Title”、“Abstract”、“Summary”的前 20 行和“Claim”的內容，利用專利文件欄位內容出現的單一詞及名詞片詞出現的頻率，計算詞彙的重要性，再經過 k-nearest-neighbors（kNN）Algorithm 方法找出最相近 k 篇專利文件決定分類結果，如此一來就無需先訓練資料庫，但由於類別太過相近，又每一個子類別的文件數太少，因此正確性只有在 25% ~ 32%之間。實驗結果說明雖然名詞片詞可以提高專利文件搜尋相似文件的正確性，但對於分類卻沒有多大的幫助。

Winnow 是一種錯誤學習分類法，適合用於大量文件和龐大的文件特徵數量上，Koster 等人[2002]利用 Winnow 分類演算法用在歐洲專利分類上，取出歐洲專利局(European Patent Office，EPO)專利的欄位“Abstract”（平均 129 個詞）和全文（平均 4580 個詞）比較，並且和 Rocchio 分類法做比較，不論是單一類別（文件為 16,000 篇，16 個主類別）或多類別（文件為 10 萬篇以上，44 個主類別，549 個次類別）分類，Winnow 比 Rocchio 分類法在只取欄位“Abstract”和全文上都可提高 1~2 的調合平均值而達到 99~100%。

第 3 章　　專利文件分類

3.1. 語料處理

我們收集以UPC（US Patent Classification）為分類的美國專利文件語料，由「美國專利局（USPTO）」網站http://www.uspto.gov/下載，檔案格式為超文件標示語言，必須進一步解析超文件標示語言結構，擷取文件內容。由於專利文件的篇幅過大（甚至有20萬字以上），李駿翔[2003]的實驗結果認為專利文件中欄位“Title”搭配欄位“Summary”或“Description”是執行分類較有效的，因此我們擷取欄位“Title”和“Summary”以節省系統空間，其描述專利文件內容的部分包括發明的主題以及專利發明的內容概述。

由於美國專利文件的分類標準分工仔細，我們分別各取一份以主類別（語料1）為分類和以子類別（語料 2）為分類的語料來做測試：

類別編號	主類別（Class）	數量
345	Computer graphics processing and selective visual display systems	100
360	Dynamic magnetic information storage or retrieval	100
367	Communications, electrical: acoustic wave systems and devices	100
704	Data processing: speech signal processing, linguistics, language translation, and audio compression/decompression	100
705	Data processing: financial, business, practice, management, or cost / price determination	100
706	Data processing: artificial intelligence	100
712	Electrical computers and digital processing systems: processing architectures and instruction processing (e.g., processors)	100
713	Electrical computers and digital processing systems: support	100
718	Electrical computers and digital processing systems: virtual machine task or process management or task management / control	100

表 3-1：語料 1 的類別及數量統計

語料 1 是以主類別做為分類類別，主要都是和電腦科學相關主題隨機取 9 個類別。各選取最新的文件 100 篇，總共選取 900 篇，如表 3-1列出的類別編號、類別主題及其數量，由於分類類別的數量有限。此語料的文件均只歸屬於一個類別。

子類別編號	子類別（Subclass）	數量
1	DATABASE OR FILE ACCESSING	1457
2	Access augmentation or optimizing	1529
3	Query processing	2538
4	Query formulation, input preparation, or translation	1347
5	Query augmenting and refining	1197
6	Pattern matching access	1007
7	Sorting	577
8	Concurrency	682
9	Privileged access	643
10	Distributed or remote access	2995
100	DATABASE SCHEMA OR DATA STRUCTURE	1461
101	Manipulating data structure	1245
102	Generating database or data structure	1891
103R	Object-oriented database structure	800
103Y	Object-oriented database structure processing	87
103X	Object-oriented database structure network	26
103Z	Object-oriented database structure reference	26
104.1	Application of database or data structure	2076
200	FILE OR DATABASE MAINTENANCE	1031
201	Coherency	796
202	Recoverability	645
203	Version management	861
204	Archiving or backup	687
205	File allocation	608
206	Garbage collection	381

表 3-2：語料 2 的類別及數量統計

語料 2 是以主類別“707（DATA PROCESSING: DATABASE AND FILE MANAGEMENT OR DATA STRUCTURES）”下所有子分類做分類類別，有 25 個子類別，如表 3-2所示，總共有 12489 篇文件（原來總共有 13882 篇，去除欄位不完整及類別標示錯誤的文件）。

此語料內的文件所屬類別的分類數至少為 1（可多個），在這裡的類別架構雖是階層式，但文件的類別卻沒有固定在哪一層，也可能多層同時存在，如專利編號為“5455945”的文件屬於子類別“2”、“4”、“7”和“104.1”，子類別“2”、“7” 和“104.1”在主類別“707”的第二層，子類別“4”是在第三層。文件類別數分佈如圖 3-1所示，顯示在 12489 篇文件中有 5355 篇是只屬於單一類別的，約占了 43%；同時屬於二個類別的文件有 3706 篇，約占 30%。

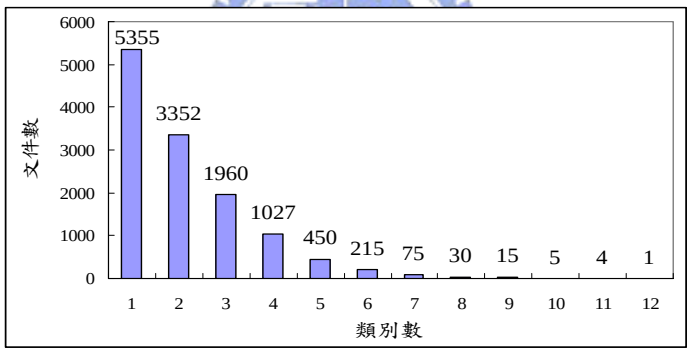


圖 3-1：語料 2 文件中所屬的類別數分佈

語料 1 文件均只屬於一個類別，不需由調適語料對不同詞彙群組數決定相似類別的門檻值，將資料分為訓練語料和測試語料，依序占全部資料 60%和 40%。而語料 2 分成訓練語料、調適語料和測試語料，依序占全部資料 60%、20%和 20%，有調適語料的目的是為了不希望最後決定的參數值會過度倚賴訓練語料的內容，以彌補資料稀疏性（Sparseness Data）的問題。先以訓練語料為詞彙分群的依據，依不同的詞彙群組數結果分別以調適語料去做測試，調適語料測驗結果決定相似類別的門檻值。

語料 1	訓練語料	測試語料		全部
	540	360		900
語料 2	訓練語料	調適語料	測試語料	全部
	7496	2494	2499	12498

表 3-3：訓練語料、調適語料和測試語料的文件數

針對專利文件中欄位“Title”和“Summary”的內容，抽取出句子中名詞和動詞的詞彙，因此必須要進行詞性標記，使用由清華大學張俊盛教授自然語言實驗室所提供的工具，這個工具是以 CoNLL2002 為語料所訓練出來的。專利文件前置處理的過程如下：

1. 詞根還原。
2. 進行詞性標記。
3. 抽取名詞和動詞詞彙。
4. 刪除停用字[6]（Stop Word）和數字。
5. 刪除低頻詞（只出現一次）。

前置處理各個步驟執行後的結果，如表 3-4所示。

步驟	語料 1		語料 2		
	訓練語料	測試語料	訓練語料	調適語料	測試語料
起始相異詞數	7783	5726	18871	10767	11923
詞態還原後	5212	4197	15111	8703	8432
除去非動詞和名詞後	4137	3091	11312	6104	7554
除去停用詞和數字後	3866	2870	9370	5790	6447
除去低頻詞後	2995	—	8836	—	—

表 3-4：前置處理結果

3.2. 詞彙分群

因為專利文件主題含蓋範圍太廣，詞彙在一般意思和專業領域的義意上會有差異，因此我們選擇統計式的。非監督式的分群法的無法界定合適群組的程度，又我們主要的目的是對文件做分類，因此選擇監督式的 Distributional Clustering

（DC）分群法，這種分群法是一種機率分佈曲線為依據的分群法，針對詞彙在類別之間的分佈進行群組。

DC 先計算每一個詞彙在各個類別之間的機率分佈，再依類別分佈曲線計算詞彙之間的相似度，以進行分群。步驟如下：

i. 統計詞彙在每個類別的出現機率為該詞彙的類別機率分佈曲線。

ii. 利用 Kullback-Leibler (KL) divergence 計算二個詞彙的類別機率分佈（Probability Distributions）的差異性。

KL-divergence [Pereira et al., 1993] 定義 $D(P(C|w_t) \| P(C|w_s))$ 如公式 3-1所示，以詞彙 w_s 的類別機率分佈曲線為基礎，計算 w_t 和 w_s 類別機率分佈曲線的差異性。

$$D(P(C \mid w_t) \| P(C \mid w_s)) = -\sum_{j=1}^{|C|} P(c_j \mid w_t) \log \left(\frac{P(c_j \mid w_t)}{P(c_j \mid w_s)} \right) \qquad \text{公式 3-1}$$

$P(c_j\|w_t)$：群組 w_t 類別 c_j 所占的比例
$P(C\|w_t)$：群組 w_t 的類別機率分佈曲線
C：所有的類別
$\|C\|$：所有類別的數量
$P(w_t)$ ：w_t 所占的出現的機率
$N(w_t)$ ：w_t 群組成員數

在 Baker and McCallum[1998]提出 DC 的過程中，為了修正 KL-divergence 在計算二個曲線差異性時，任取一個曲線為基礎會造成計算結果不對稱的問題，也就是 $D(P(C|w_t) \| P(C|w_s)) \neq D(P(C|w_s) \| P(C|w_t))$，因此加入加權平均的概念，對 KL-divergence 做修改。以二個曲線合併的結果為基礎，分別計算和此合併曲線的差異性，再以各別的權重分配二個結果進行加總。計算 w_t 和 w_s 之間類別機率分佈的差異性 $Dist(w_s, w_t) = Dist(w_t, w_s)$，如公式 3-2 所示，其中公式包含 $P(C| w_s \vee w_t)$，代表二個詞彙合併之後的類別機率分佈，如公式 3-3所示。

[6] http://dvl.dtic.mil/stop_list.html

$$Dist(w_t, w_s) = Dist(w_s, w_t)$$
$$= \frac{P(w_t)}{P(w_t + w_s)} D(P(C \mid w_t) \| P(C \mid w_t \vee w_s)) + \frac{P(w_s)}{P(w_t + w_s)} D(P(C \mid w_s) \| P(C \mid w_t \vee w_s)) \quad \text{公式 3-2}$$

$$P(C \mid w_t \vee w_s) = \frac{P(w_t)}{P(w_t) + P(w_s)} P(C \mid w_t) + \frac{P(w_s)}{P(w_t) + P(w_s)} P(C \mid w_s) \quad \text{公式 3-3}$$

iii. 依照所算出兩兩詞彙機率分佈的差異性，選擇其中可以適當被合併的詞彙。

iv. 進行合併，計算合併為公式 3-3。

v. 重覆以上的步驟，直到剩下合理的詞彙群組數。

在公式 3-2中，若 $P(w_t)$所占的比例比 $P(w_s)$大得多時，$Dist(w_s, w_t)$所得到的結果幾乎等於 $D(P(C|w_t)||P(C|w_s \vee w_t))$，而 $D(P(C|w_s)||P(C|w_s \vee w_t))$的差異性容易被忽略掉，造成 w_t 此群組有不合理的合併而過度膨脹。Chen 等人[2005]為修正此錯誤，加入群組成員數的考量形成 $Global\_Dist(w_s, w_t)$，如公式 3-4所示，改善在群組過後分配的不平均的結果。

$$Global\_Dist(w_t, w_s) = \frac{N(w_t) + N(w_s)}{2\sum_{k=1}^{|M|} N(w_k)} \times Dist(w_t, w_s) \quad \text{公式 3-4}$$

圖 3-2顯示語料 1 中 200 個詞彙群組數的分配結果，其中以“[# 1]”代表群組成員只有一個的詞彙群組個數，有 55 個詞彙群組的成員是只有一個的，約占有27.5%詞彙群組、2%詞彙，而群組“datum”的成員有 907 個詞彙。詞彙群組“datum”所囊括的詞彙非常的多，約占 32%以上的詞彙，分佈是相當不平均。

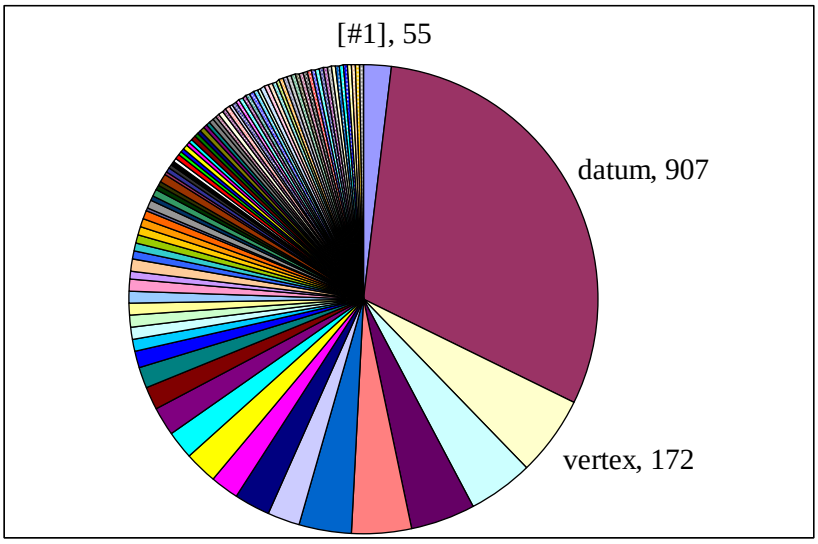


圖 3-2：語料 1 中 200 個詞彙群組數的分配結果

在圖 3-3裡，列出和類別“345”和“706”相關性（*P*(類別|詞彙群組) ）最大的二個詞彙群組：“vertex”、“scene”及“keyword”、“query”（以實線表示）以及成員最多的詞彙群組“datum”（以虛線表示）的類別機率分佈情形。詞彙群組“vertex”完全只會在類別“345”裡出現，詞彙群組“scene”會出現在類別“345”和類別“704”中，以類別“345”出現的機會來得大，在其他的類別機乎不會出現，而詞彙群組“datum”則平均分散在各類別之間。

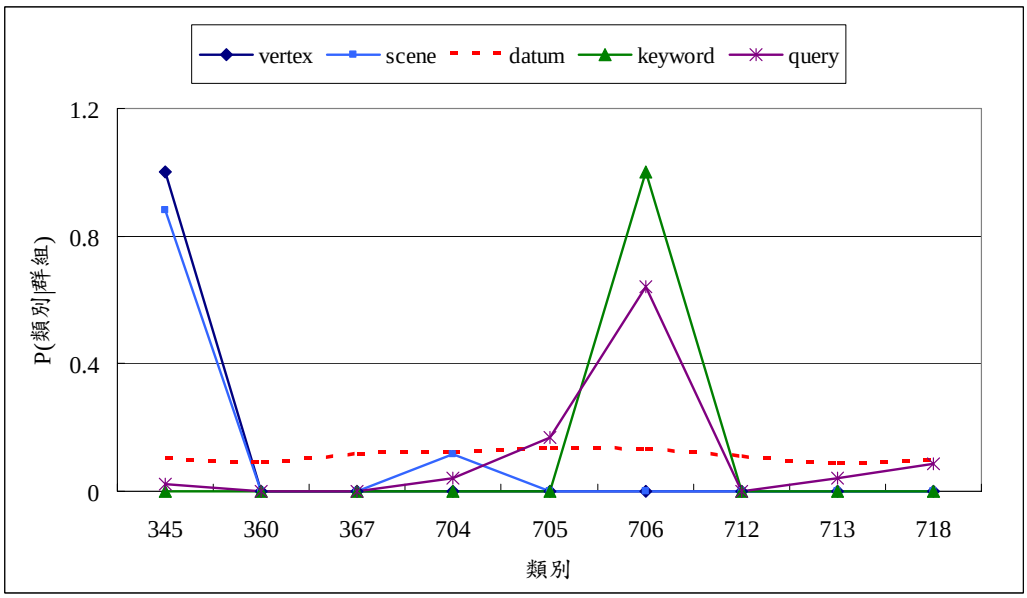


圖 3-3：語料 1 部份詞彙群組的類別機率分佈

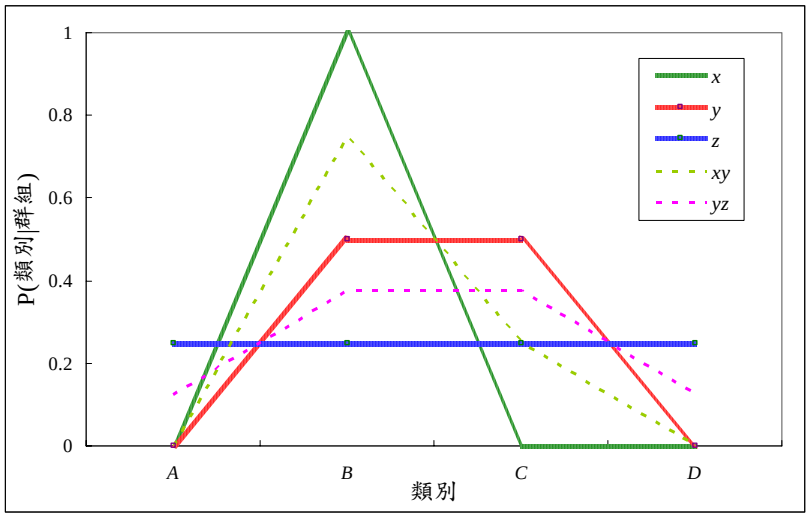


圖 3-4：群組機率曲線分佈圖範例

圖 3-4為詞彙群組{*x*、*y*、*z*}在類別{*A*、*B*、*C*、*D*}中的機率曲線分佈圖，而 *xy* 表示 *x* 和 *y* 合併的結果，*yz* 表示 *y* 和 *z* 合併的結果。其中 *x* 代表和類別 *B* 極為相關的詞彙，對分類來說是最有鑑別力；而 *y* 代表只會出現在類別 *B* 和 *C* 之間的詞彙，對分類也有會幫助的效果；*z* 代表在每個類別都會出現的詞彙，是在分類中最不重要的詞彙。一般都會認為 *y* 的重要性是在 *x* 和 *z* 之間，但以曲線的分佈來看，*y* 和 *z* 對於 *yz* 曲線距離比 *x* 和 *y* 對於 *xy* 曲線距離來得小，若{*x*、*y*、*z*}要做分群，首先一定會先選擇 *y* 和 *z*。在合併的過程中，詞彙群組會漸漸趨向曲線較平均的詞彙群組，這也是為什麼分群結果中會有少數幾個詞彙群組所占的詞彙特別的多，且此少數詞彙群組都是分散較平均的。

詞彙群組“vertex”內有 172 個成員，包含 vertex、aliasing、avatar、blend、blur、bone、cameras、CG、clipping、contour、dispersion、illusion、intensity、keyframe、landscape、mosaic、plane、polygon、ray、rectangle、rendering、shade、stencil、stereogram、texels、texture、tile、trim、voxels…等，都是和類別“345：Computer graphics processing and selective visual display systems”主題常見的詞彙，在一般的詞典來說可能不容易找出 “blur”（模糊化）和 “blend”（混合，混色）之間的關係，甚至“keyframe”（關鍵格)、“texels”（貼圖畫數）、“voxels”（立體像數）…等是在這主題領域中常見的詞彙但在一般的詞典卻找不到，而這些詞彙對於類別

“345”這個主題來說都是相當有鑑別力的詞。

詞彙群組“keyword”內有 125 個成員，包含“keyword”、Bayesian、categorization、centroid、classifier、clustering、collector、concentrator、documentation、dictionary、grammar、hierarchically、lexicon、noun、paraphrase、searching、sentence、sql、summarization、svm、synonym、thesaurus、tokenizing、verb…等，都是和類別“706：Data processing: artificial intelligence”主題常見的詞彙，其他還有像“evaluator”、 “English”、“topic”…等看起來跟“706”的主題沒有直接聯想的關係，因和其他類別的相關性差異大，其他的類別出現這些詞的關係相對於“706”來得更小。利用詞彙分群中 Distributional Clustering 方法能辨識出許多和類別主題相關的詞彙。

而詞彙群組“datum”很平均的分散在各個類別之中，詞彙群組“datum”主要的詞彙為 datum、invention、method、information、process、object、computer、device、user、set、second、memory、control、program、store、input、apparatus、output、processor、unit、base、comprise、display…等，都是在主類別“707”下較無直接相關的詞彙。

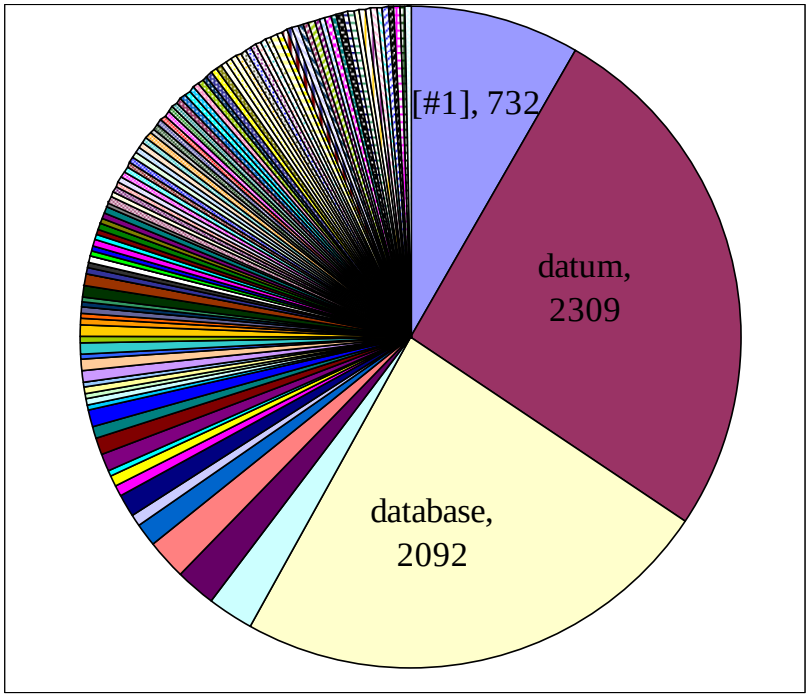


圖 3-5：語料 2 中 2000 個詞彙群組數的分配結果

圖 3-5顯示語料 2 中 2000 個詞彙群組數的分配結果，其中也以“[# 1]”代表詞彙群組成員只有一個的詞彙群組個數，有 732 個詞彙群組是只有一個成員的，占了 37%的詞彙群組、約占 8%詞彙，而以群組“datum”的成員有 2309 個詞彙，約占 26%詞彙，詞彙群組“database”的成員有 3092 個，約占 35%詞彙，和語料 1 的情形相同，詞彙群組之間分配得相當不平均。語料 2 的群組中，以詞彙群組“datum”和”database”所占的比例最大，在圖 3-6中顯示此二個詞彙群組的類別機率分佈曲線，且出現在大部份的類別機率都差不多，符合前面所說的，有少數重要性較低的幾個群組所占的比例會較大。

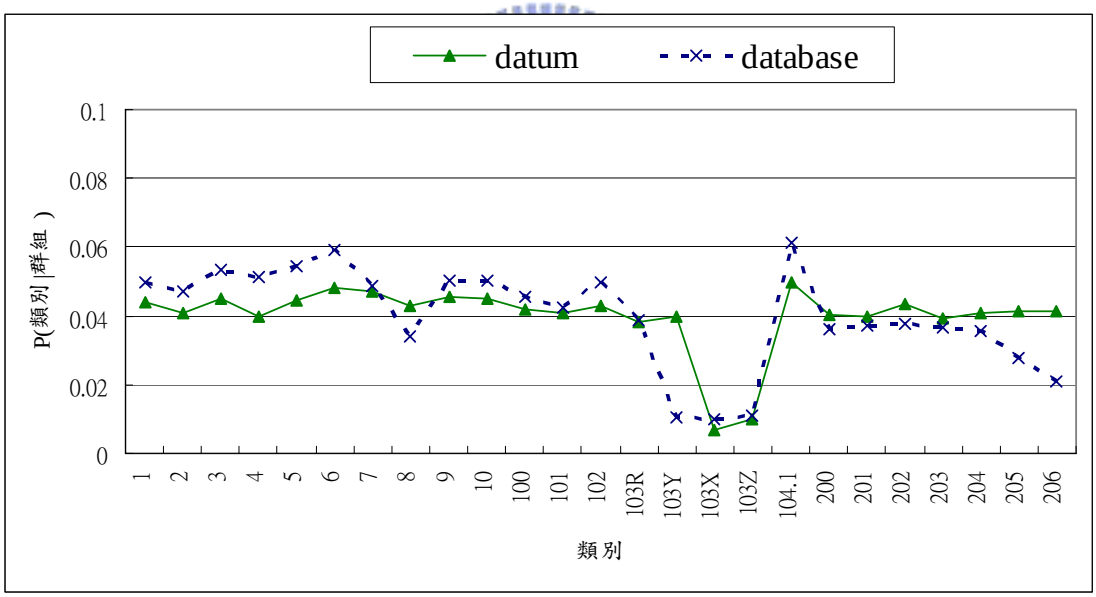


圖 3-6：語料 2 部份詞彙群組的類別機率分佈

3.3. 文件分類

接下來的步驟要進行文件的分類，所採用的方法是向量空間模型（Vector Space Model，VSM）。對資料庫內的每一文件建立一代表向量，向量裡的數值就顯示詞彙對於其文件的重要性。採用 tf-idf（Term Frequency / Inverse Document

Frequency）來計算出每一文件裡詞彙權重。

資料庫裡包含 M 篇文件 { doc_1，doc_2，…，doc_i，…，doc_M }，總共有 N 個彼此不相同的詞彙群組代表詞彙 { $word_1$，$word_2$，…，$word_j$，…，$word_N$ }，dw_{ij} 為每一個詞彙($word_j$)對每一篇文件(doc_i)的權重，文件 doc_i 的向量表示為 $\overrightarrow{doc_i}$ = < dw_{i1}，dw_{i2}，dw_{i3}，…，dw_{ij}，…，dw_{iN} >。其中權重 (dw_{ij}) 的計算 tf-idf 如公式 3-5 所示。

$$dw_{i,j} = \frac{\log(tf_{i,j}) \times \log\left(\frac{N}{df_{i,j}}\right)}{\sqrt{\sum_{j}^{|N|}\left(\log(tf_{i,j})^2 \times \log\left(\frac{N}{df_{i,j}}\right)^2\right)}}$$

公式 3-5

tf_{ij}：詞彙(w_j)在文件(d_i)裡出現的次數
df_j：出現詞彙(w_j)的文件總數量
cw_{kj}：詞彙(w_j)對於類別(c_k) 的權重
dw_{ij}：詞彙(w_j)對文件(doc_i)的權重
$|C_k|$：類別 (C_k) 的文件數
N：相異詞彙數量

藉由屬於該類別的文章向量取得 $class_k$ 的向量表示，如公式 3-6所示計算 cw_{kj} 來表示詞彙 (w_j) 對於類別 (c_k) 的權重。類別也可以用向量來表示 $\overrightarrow{class_k}$ = < cw_{k1}，cw_{k2}，cw_{k3}，…，cw_{kj}，…，cw_{kN} >。

$$cw_{k,j} = \frac{\sum_{i \in C_k} dw_{i,j}}{|C_k|}$$

公式 3-6

$$sim(doc_i, class_k) = \frac{\overline{doc_i} \bullet \overline{class_k}}{\| \overline{doc_i} \| \| \overline{class_k} \|}$$

公式 3-7

最後再以餘弦函數（Cosine Measure）計算測試文件的向量空間和各類別向量空間之間的相似性來決定相近的類別，如公式 3-7所示。相似度的值範圍介於 0~1 之間，其值越大相似度越高，為 1 時表示二者向量數值完全相同。

測試文件只有歸屬於一個類別時，選擇計算結果相似度最高的類別；歸屬於多個類別時，無法判斷文件屬於多少類別，但至少會有一個類別，因此相似度最高的會被認定是測試文件所屬的一個類別，而剩下非相似度最高的類別再次參考與相似度最高類別之間差異度，重新調整其相似度數值，認為更新過後的相似度高於一個門檻值的類別也是此測試文件的類別。

需要重新調整相似度數值的原因是，若文件歸屬類別數量少時，文件和歸屬類別的相似度會比和非歸屬類別的相似度高得多，程度相差較大；若文件歸屬類別數量較多時，此文件和歸屬類別的相似度都會偏低。若在一開始就直接定一個門檻值的話，歸屬類別數量較多的文件，可能和類別的相似度都低於門檻值，而無法預測其類別。因此和相似度最高的類別做比較，調整本身的相似度數值，當和相似度最高的類別的差異越小時，被認為測試文件類別的機會就會越高。

如表 3-5的範例所示，文件 *doc_x* 為分類在較少的類別中，對某一類別的相似度就會較高，但 *doc_y* 為分類在較多的類別中，和每一個類別的相似度都偏低，若直接設定門檻值，*doc_y* 可能無法預測其多類別，透過和最高類別的相似度差異性調整之後，*doc_y* 的多類別預測為{*classB*，*ClassC*}。

相似度調整前(門檻值 = 0.8)					
	classA	*classB*	*classC*	單一類別預測	多類別預測
doc_x	0.3	0.5	0.85	{*classC*}	{*classC*}
doc_y	0.3	0.6	0.5	{*classB*}	Nan
相似度調整後(門檻值 = 0.8)					
	classA	*classB*	*classC*	單一類別預測	多類別預測
doc_x	0.3 / 0.85 = 0.35	0.5 / 0.85 = 0.59	相似度最高類別	{*classC*}	{*classC*}
doc_y	0.3 / 0.6 = 0.5	相似度最高類別	0.5 / 0.6 = 0.83	{*classB*}	{*classB, classC*}

表 3-5：相似度值調整範例

表 3-6顯示語料 2 中 600 個詞彙群組時，進行十次實驗平均，相似度調整後對於分類的結果。可以看得出來在不論是在權重計算修正前後，相似度的調整都

對於分類的結果有幫助，均提升 0.1 以上的調和平均值。

相似度調整前	
權重計算修正前	權重計算修正後
0.216	0.27
相似度調整後	
權重計算修正前	權重計算修正後
0.304	0.415

表 3-6：相似度值調整調和平均值結果

3.4. 權重計算修正

在[Jing et al., 2002]中有提到，當一個詞彙在某少數類別裡的文件出現的特別多次，在其他的類別出現的次數少很多，這種情形下應該要提高他的權重，也就是他的重要性相對要增加，但因為tf-idf比較倚重在詞彙在文件的出現頻率上，比較類別之間的頻率重要性相對無法被突顯出來。

本論文提出新的權重公式希望把類別之間的頻率的重要性也考慮進去，將原來詞彙對於類別的權重加入亂度（Entropy）概念對於詞彙的影響，如公式 3-8 所示，是之前詞彙 (w_j) 對於文件 (c_k) 權重 (cw_{kj}) 的計算再乘上 ($1-AdaptiveEntropy(w_j)$)，當一個亂度高的詞彙，期望他的對分類權重也相對降低。以往亂度公式中因資訊以bits編碼，故log以基底2之對數表示，但這裡希望其值介面0~1之間，因此其基數為cf，加入此詞彙在出現類別所占的比例 ($cf / |C|$)，修正亂度公式的計算如公式 3-9 所示，在這裡利用前面做詞彙群組已經計算好的每一個詞彙在各個類別機率的分佈，$P(c|w)$，接著，由於當有一個 *word_x* 只會出現在其中二個類別而且在這二個類別分配出現的機率都是 $P(c|w) = 0.5$，有另一個 *word_y* 會在其中五個類別，同樣的在這五個類別分配出的機率為 $P(c|w) = 0.2$，在出現的類別都分佈的相當平均，以亂度計算二個詞彙的結果皆為1，但很容易看得出來 *word_x* 比 *word_y* 來得有鑑別力，因此在這裡加入 ($cf / |C|$) 以增加鑑別力。

$$cw_{k,j} = \frac{\sum_{i \in c_k} dw_{i,j}}{|C_k|} \times \left(1 - AdaptiveEntropy\left(w_j\right)\right)$$ 公式 3-8

$$AdaptiveEntropy\left(w_j\right) =$$
$$\left(-\sum_{k \in cf_j} \frac{p\left(c_k \mid w_j\right)}{\sum_{k \in cf_j} p\left(c_k \mid w_j\right)} \times \log_{|cf_j|}\left(\frac{p\left(c_k \mid w_j\right)}{\sum_{k \in cf_j} p\left(c_k \mid w_j\right)}\right)\right) \times \frac{|cf_j|}{|C|}$$ 公式 3-9

cf_j：詞彙 w_j 出現的類別

$|C|$：類別的數量

前面已經看過各個類別依照詞彙群組所顯示相關的詞彙群組，也看過因為權重調整讓文件分類的調和平均值變高，接下來要觀察詞彙群組權重調整前後的差異。

圖 3-7為語料 1 中 200 詞彙群組數，類別“345”的前 10 個經過權重調整後差異較大的詞彙群組，差異大的有詞彙群組“datum”，由於經過詞彙分群 tf 值過份擴張，(原來在訓練語料裡 tf 的值最大為 326，經過群組後，tf 的值最大為 3815)，藉由 tf-idf 計算後，所給予的權重也相對變大許多，但此詞彙群組的亂度大（在前面圖 3-3顯示此詞彙群組平均分散在各類別之間），經過權重調整之後，詞彙群組“datum” 權重調整變小。而詞彙群組“vertex”、“image”…等，其亂度小，經過亂度的調整後，權重值雖然變小，但和其他的詞彙群組權相比較，比其原來的權重值來得有影響力。

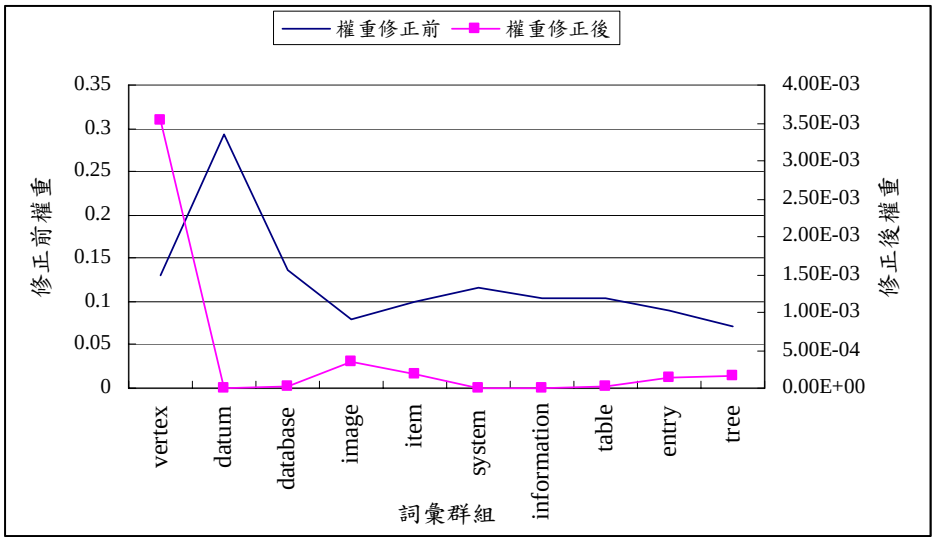


圖 3-7：語料 1 部份詞彙群組的權重比較

3.5. 評估計算標準

本論文所採用的衡量標準是在資料檢索領域中用來評估效能的標準：精確率（P）、召回率（R）和調和平均值（F），計算方法如表 3-7範例。

Doc_X	
系統分類結果	{ *Class1*、*Class2*、*Class3* }
正確分類結果	{ *Class3*、*Class4* }
精確率	P = 1 / 3 = 0.33
召回率	R = 1 / 2 = 0.5
調和平均值	F = 2 / (1 / 0.33 + 1 / 0.5) = 0.4

表 3-7：精確率、召回率和調和平均值計算範例

第 4 章　　實驗與分析

4.1. 詞彙分群實驗

此實驗是為測試詞彙分群對於專利文件分類的效果，針對不同的詞彙群組數做測試，以了解需要多少個詞彙群組數在實際的過程式可以減少分類系統的負擔，同時又不至於失去詞彙的鑑別力和文件分類準確度。

圖 4-1為語料 1 詞彙分群實驗測試結果曲線圖，在未分群(2995)減少到 1500 之間的詞彙群組數都是相當穩定接近 0.78 的調和平均值，對於詞彙分群詞彙群組數在未減少調和平均值的情形下至少降低一半資料庫文字的向量空間。最高的調和平均值在於詞彙群組為 2000 個左右，可見在 2000~2995 之間，詞彙分群可以減少雜訊的發生。

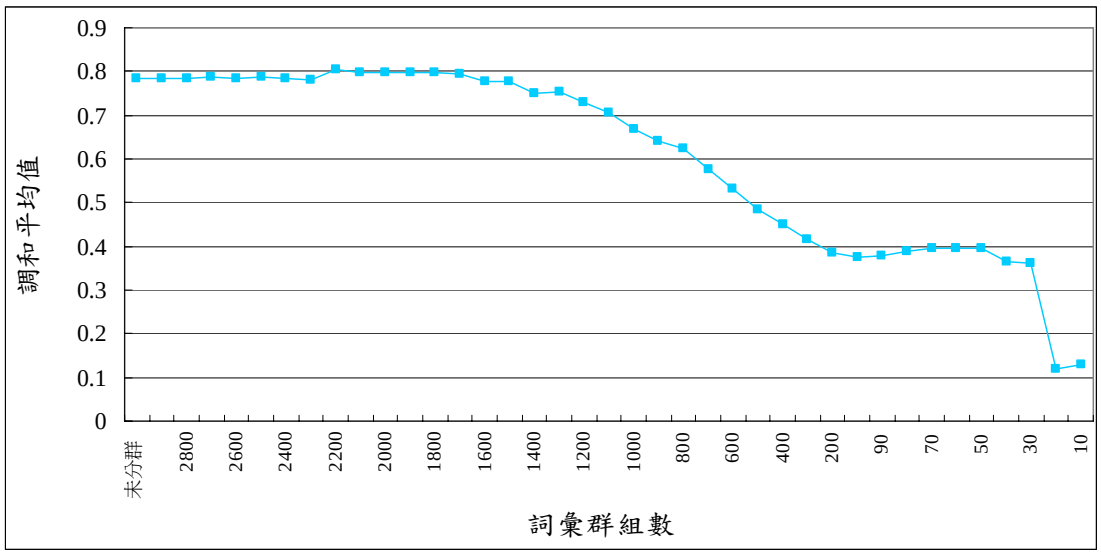


圖 4-1：語料 1 詞彙分群實驗測試結果曲線圖

圖 4-2為語料 2 詞彙分群的測試結果，由調適語料對不同詞彙群組數決定相

似類別的門檻值，再對測試語料測試調和平均值，因為分類的類別相似度相當高又各類別文件數分佈相當不平均，所以相對的調和平均值比語料 1 來得差，圖中曲線詞彙群組數落在未分群(8836)到 2000 之間測試語料調和平均值均高於 0.35，對於詞彙分群結果來說在未減少調和平均值的情形下至少降低至約四分之一資料庫文字的向量空間緯度。

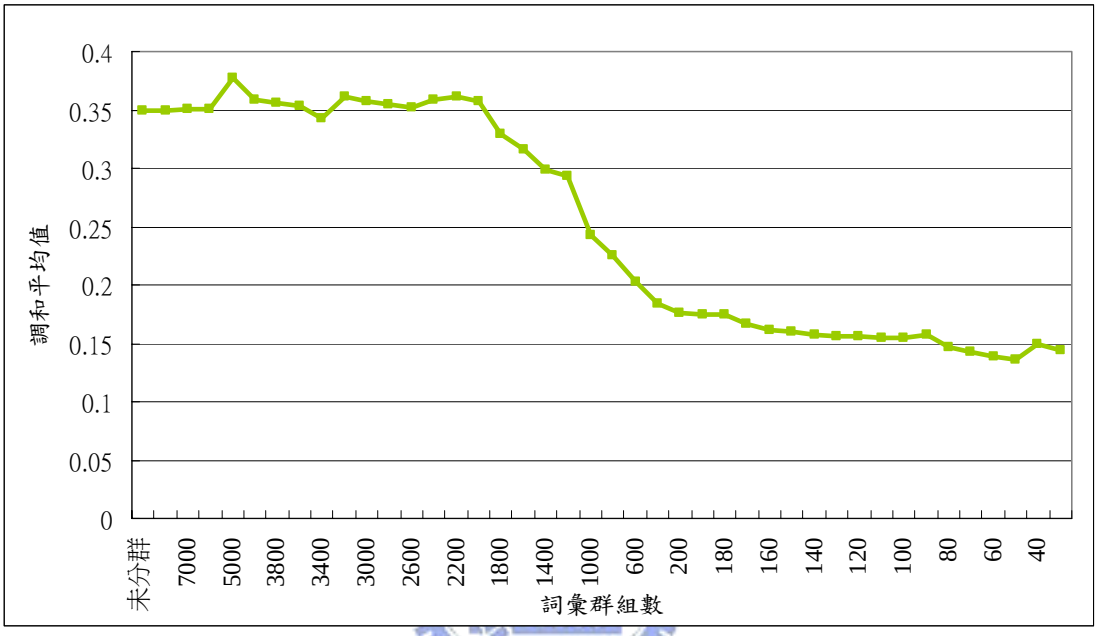


圖 4-2：語料 2 詞彙分群實驗測試結果曲線圖

4.2. 權重計算修正實驗

此實驗為觀察權重計算公式修正後對於專利文件的是否有影響，能否改善文件分類，針對不同的詞彙群組數做測試及其結果分析。

表 4-1為語料 1 權重計算修正實驗測試和原來未修正前的做比對，原來未修正前在少於 1500 個詞彙群組數調和平均值開始減少，在修正詞彙權重公式後調和平均值下降的情形開始漸緩，若以執行速度為主要考量，200 個詞彙群組的情形下調和平均值為 0.73，提升 86%的調和平均值，可見詞彙權重公式修正後的結果是有用的。表 4-2為語料 1 中 200 個詞彙群組數各類別的檢索精確率和檢索召

回率，雖然減少 5%的召回率，但大幅提升 41%的精確率。

詞彙群組數	權重修正前	權重修正後	詞彙群組數	權重修正前	權重修正後
無分群	0.784	0.791	1000	0.669	0.786
2900	0.784	0.791	900	0.641	0.786
2800	0.784	0.794	800	0.623	0.784
2700	0.786	0.791	700	0.576	0.786
2600	0.784	0.791	600	0.531	0.786
2500	0.786	0.786	500	0.484	0.778
2400	0.784	0.763	400	0.451	0.755
2300	0.781	0.768	300	0.418	0.763
2200	0.804	0.761	200	0.387	0.735
2100	0.799	0.784	100	0.374	0.696
2000	0.799	0.784	90	0.379	0.694
1900	0.799	0.784	80	0.389	0.699
1800	0.796	0.786	70	0.397	0.691
1700	0.794	0.789	60	0.394	0.668
1600	0.779	0.789	50	0.394	0.663
1500	0.779	0.786	40	0.364	0.653
1400	0.75	0.783	30	0.361	0.64
1300	0.753	0.786	20	0.12	0.512
1200	0.73	0.786	10	0.128	0.338
1100	0.707	0.791			

表 4-1：語料 1 權重計算修正實驗測試結果(調和平均值)

權重修正前						權重修正後					
類別	TP	TP+FN	TP+FP	召回率	精確率	類別	TP	TP+FN	TP+FP	召回率	精確率
345	18	40	18	0.45	1	345	32	40	44	0.8	0.727
360	23	40	25	0.575	0.92	360	36	40	53	0.9	0.679
367	3	40	4	0.075	0.75	367	34	40	36	0.85	0.944
704	7	40	11	0.175	0.636	704	18	40	20	0.45	0.9
705	8	40	8	0.2	1	705	32	40	34	0.8	0.941
706	40	40	262	1	0.153	706	37	40	72	0.925	0.514
712	5	40	5	0.125	1	712	34	40	43	0.85	0.791

713	7	40	7	0.175	1	713	14	40	22	0.35	0.636
718	20	40	20	0.5	1	718	30	40	36	0.75	0.833
平均				0.364	0.829	平均				0.774	0.742

TP：系統預測且正確的個數
TP+FN：正確的個數
TP+FP：系統預測的個數

表 4-2：語料 1 中 200 個詞彙群組數各類別的精確率和召回率

表 4-3為語料 2 權重計算修正實驗測試和原來未修正前的做比對，2000 個詞彙群組數以上由約 0.35 的調和平均值由 0.37 提升至 0.42，增加 5%~21%的調和平均值，經過權重計算公式修正之後都比原來未修正前的調和平均值略好一點，到 600 個詞彙群組數以上調和平均值有 0.35 以上（修改之前最少需要在 2000 個詞彙群組數以上），增加 70%的調和平均值。

詞彙群組數	調適語料 權重修正前 / 後		測試語料 權重修正前 / 後		詞彙群組數	調適語料 權重修正前 / 後		測試語料 權重修正前 / 後	
未分群	0.454	0.484	0.349	0.418	800	0.325	0.448	0.225	0.354
8000	0.454	0.482	0.349	0.426	600	0.29	0.43	0.203	0.35
7000	0.454	0.483	0.351	0.415	400	0.259	0.381	0.184	0.314
6000	0.452	0.476	0.35	0.408	200	0.228	0.321	0.176	0.259
5000	0.452	0.478	0.377	0.409	190	0.236	0.335	0.175	0.261
4000	0.451	0.478	0.359	0.394	180	0.233	0.336	0.175	0.26
3800	0.451	0.477	0.355	0.393	170	0.238	0.334	0.167	0.265
3600	0.452	0.477	0.354	0.412	160	0.251	0.327	0.161	0.264
3400	0.452	0.477	0.343	0.406	150	0.232	0.323	0.16	0.256
3200	0.454	0.474	0.361	0.41	140	0.231	0.326	0.157	0.255
3000	0.451	0.473	0.358	0.398	130	0.223	0.322	0.156	0.263
2800	0.448	0.475	0.355	0.416	120	0.221	0.321	0.155	0.257
2600	0.449	0.472	0.352	0.409	110	0.241	0.321	0.155	0.26
2400	0.448	0.467	0.359	0.408	100	0.2	0.319	0.155	0.255
2200	0.445	0.467	0.362	0.383	90	0.16	0.323	0.157	0.245
2000	0.438	0.466	0.357	0.384	80	0.159	0.312	0.147	0.246
1800	0.435	0.463	0.33	0.389	70	0.167	0.3	0.143	0.236
1600	0.431	0.467	0.316	0.388	60	0.171	0.301	0.138	0.233
1400	0.42	0.461	0.299	0.39	50	0.162	0.283	0.136	0.231
1200	0.396	0.458	0.293	0.38	40	0.152	0.251	0.15	0.2

1000	0.359	0.456	0.242	0.373	30	0.157	0.262	0.144	0.2

表 4-3：語料 2 權重計算修正實驗測試結果(調和平均值)

表 4-4為語料 2 的 2000 個詞彙群組數各類別的精確率和召回率，其中類別“103Y”、類別“103X”、類別“103Z”都是文件數較少且分類較細的類別，以致精確率或和召回率都特別的低。而精確率無法提高是因為無法預先知道文章所屬的類別數，一旦猜測的類別數增加，召回率就會相對下降。經過權重修正後，精確率和召回率均提昇 2%～4%。

權重計算修正後對於二份語料的分類皆能有效提昇分類的調和平均值，尤其詞彙群組數越小時所提昇的效果越大，讓詞彙分群更能達到作用，對於減少系統的資源耗費有相當的幫助。

權重修正前						權重修正後					
類別	TP	TP+FN	TP+FP	召回率	精確率	類別	TP	TP+FN	TP+FP	召回率	精確率
1	206	808	636	0.255	0.324	1	112	808	248	0.139	0.452
2	129	604	232	0.214	0.556	2	110	604	462	0.182	0.238
3	258	732	522	0.352	0.494	3	528	732	771	0.721	0.685
4	138	601	422	0.23	0.327	4	185	601	541	0.308	0.342
5	114	633	226	0.18	0.504	5	195	633	479	0.308	0.407
6	77	365	109	0.211	0.706	6	120	365	317	0.329	0.379
7	55	186	73	0.296	0.753	7	41	186	79	0.22	0.519
8	59	239	89	0.247	0.663	8	76	239	229	0.318	0.332
9	68	344	146	0.198	0.466	9	52	344	94	0.151	0.553
10	436	1505	844	0.29	0.517	10	494	1505	725	0.328	0.681
100	228	811	902	0.281	0.253	100	281	811	688	0.346	0.408
101	177	656	703	0.27	0.252	101	178	656	256	0.271	0.695
102	284	1029	902	0.276	0.315	102	322	1029	458	0.313	0.703
103R	125	308	346	0.406	0.361	103R	69	308	262	0.224	0.263
103Y	4	53	7	0.075	0.571	103Y	1	53	14	0.019	0.071

103X	2	12	3	0.167	0.667	103X	2	12	34	0.167	0.059
103Z	1	12	1	0.083	1	103Z	1	12	23	0.083	0.043
104.1	316	706	923	0.448	0.342	104.1	110	706	352	0.156	0.313
200	125	525	574	0.238	0.218	200	86	525	248	0.164	0.347
201	75	404	125	0.186	0.6	201	76	404	294	0.188	0.259
202	65	301	103	0.216	0.631	202	102	301	306	0.339	0.333
203	61	311	71	0.196	0.859	203	99	311	224	0.318	0.442
204	84	327	90	0.257	0.933	204	85	327	95	0.26	0.895
205	77	310	134	0.248	0.575	205	93	310	260	0.3	0.358
206	39	90	46	0.433	0.848	206	23	90	78	0.256	0.295
平均				0.27	0.458	平均				0.29	0.486

表 4-4：語料 2 中 2000 個詞彙群組數各類別的精確率和召回率

4.3. 語料 2 多階層分類實驗

因為語料 2 為多階層的分類，總共有 3 個階層，之前的實驗都是以第 3 為分類，在此我們要來觀察在其他階層分類的結果。將第三層子類別歸納於第二層子類別下，如表 4-6，再將第二層子類別歸納於第一層子類別下，如表 4-5所示。

子類別編號	子類別（Subclass）	數量
1	DATABASE OR FILE ACCESSING	9098
100	DATABASE SCHEMA OR DATA STRUCTURE	4612
200	FILE OR DATABASE MAINTENANCE	2740

表 4-5：語料 2 第一層子類別及數量統計

子類別編號	子類別（Subclass）	數量
1	DATABASE OR FILE ACCESSING	1457
2	Access augmentation or optimizing	1529
3	Query processing	4624
7	Sorting	577
8	Concurrency	682
9	Privileged access	643

10	Distributed or remote access	2995
100	DATABASE SCHEMA OR DATA STRUCTURE	1461
101	Manipulating data structure	1245
102	Generating database or data structure	1891
103R	Object-oriented database structure	807
104.1	Application of database or data structure	2076
200	FILE OR DATABASE MAINTENANCE	1031
201	Coherency	2452
205	File allocation	904

表 4-6：語料 2 第二層子類別及數量統計

表 4-7為第一層分類的結果，原來調和平均值約在 0.76，在詞彙群組數少於 2000 時調和平均值開始下降。在權重修正後，詞彙群組數為 600 時，調和平均值都還有 0.76 以上。表 4-8顯示在 600 個詞彙群組數時的分類情形，明顯提升召回率，由 0.57 提高至 0.7。

詞彙群組數	調適語料 權重修正前	後	測試語料 權重修正前	後	詞彙群組數	調適語料 權重修正前	後	測試語料 權重修正前	後
未分群	0.63	0.683	0.741	0.766	600	0.574	0.667	0.705	0.761
8000	0.631	0.683	0.743	0.767	400	0.596	0.665	0.702	0.76
7000	0.631	0.668	0.744	0.765	200	0.454	0.665	0.709	0.758
6000	0.633	0.669	0.744	0.762	180	0.439	0.661	0.694	0.756
5000	0.633	0.669	0.743	0.761	160	0.434	0.658	0.7	0.755
4000	0.64	0.672	0.749	0.76	140	0.425	0.655	0.709	0.754
3500	0.639	0.671	0.749	0.761	120	0.411	0.654	0.706	0.749
3000	0.638	0.67	0.748	0.761	100	0.397	0.649	0.702	0.748
2500	0.631	0.669	0.747	0.762	80	0.402	0.642	0.703	0.746
2000	0.61	0.669	0.738	0.761	60	0.398	0.635	0.705	0.742
1500	0.527	0.67	0.712	0.761	40	0.397	0.611	0.676	0.706
1000	0.519	0.669	0.708	0.762	20	0.404	0.56	0.684	0.655
800	0.533	0.667	0.709	0.761					

表 4-7：語料 2 第一層子類別分類結果(調和平均值)

權重修正前						權重修正後					
類別	TP	TP+FN	TP+FP	召回率	精確率	類別	TP	TP+FN	TP+FP	召回率	精確率
1	2336	3968	2774	0.583	0.898	1	2921	3968	2457	0.729	0.828
100	1727	2643	2776	0.657	0.972	100	2004	2643	1506	0.762	0.869
200	744	1985	693	0.417	0.904	200	947	1985	845	0.531	0.819
平均				0.571	0.922	平均				0.697	0.839

表 4-8：語料 2 第一層子類別 600 個詞彙群組數各類別分類結果

表 4-9為第二層分類的結果，在未分群時，原來 0.46 的調和平均值可提升到 0.54，在 600 個詞彙群組數時，由原來的 0.31 提升至 0.49。表 4-10顯示在 600 個詞彙群組數時的分類情形，主要也是一樣提升召回率，由 0.25 提高至 0.49。

詞彙群組數	調適語料 權重修正前 / 後		測試語料 權重修正前 / 後		詞彙群組數	調適語料 權重修正前 / 後		測試語料 權重修正前 / 後	
未分群	0.525	0.601	0.467	0.543	600	0.46	0.482	0.313	0.487
8000	0.526	0.599	0.467	0.539	400	0.461	0.457	0.308	0.451
7000	0.524	0.596	0.467	0.523	200	0.459	0.425	0.3	0.419
6000	0.523	0.593	0.461	0.525	180	0.453	0.397	0.3	0.418
5000	0.523	0.583	0.461	0.514	160	0.431	0.397	0.297	0.362
4000	0.52	0.579	0.467	0.515	140	0.423	0.389	0.282	0.327
3500	0.521	0.553	0.462	0.52	120	0.423	0.373	0.281	0.292
3000	0.516	0.56	0.461	0.513	100	0.402	0.37	0.279	0.28
2500	0.5	0.557	0.454	0.512	80	0.368	0.345	0.273	0.279
2000	0.488	0.55	0.434	0.513	60	0.388	0.324	0.264	0.273
1500	0.482	0.541	0.405	0.5	40	0.382	0.318	0.271	0.265
1000	0.475	0.519	0.351	0.5	20	0.381	0.314	0.261	0.265
800	0.482	0.495	0.337	0.5					

表 4-9：語料 2 第二層子類別分類結果(調和平均值)

權重修正前						權重修正後					
類別	TP	TP+FN	TP+FP	召回率	精確率	類別	TP	TP+FN	TP+FP	召回率	精確率
1	198	808	479	0.245	0.413	1	282	808	730	0.349	0.386
2	117	604	232	0.194	0.504	2	584	604	1676	0.967	0.348
3	285	1246	892	0.229	0.32	3	733	1246	2064	0.588	0.355
7	53	186	76	0.285	0.697	7	97	186	219	0.522	0.443
8	58	239	99	0.243	0.586	8	152	239	348	0.636	0.437
9	68	344	130	0.198	0.523	9	102	344	185	0.297	0.551
10	451	1505	1842	0.3	0.245	10	992	1505	1727	0.659	0.574
100	236	811	818	0.291	0.289	100	282	811	577	0.348	0.489
101	187	656	515	0.285	0.363	101	243	656	341	0.37	0.713
102	289	1029	659	0.281	0.439	102	371	1029	675	0.361	0.55
103R	98	562	464	0.174	0.211	103R	211	562	495	0.375	0.426
104.1	247	706	779	0.35	0.317	104.1	342	706	793	0.484	0.431
200	120	525	582	0.229	0.206	200	229	525	623	0.436	0.368
201	232	1075	309	0.216	0.751	201	465	1075	827	0.433	0.562
205	102	389	156	0.262	0.654	205	172	389	375	0.442	0.459
平均				0.257	0.402	平均				0.492	0.482

表 4-10：語料 2 第二層子類別 600 個詞彙群組數各類別分類結果

4.4. 分類法比較

這個章節要比較不同的分類法的專利文件分類結果：*k*NN 分類法以及 GIS 分類法。在前面使用的是為向量空間模型分類法。向量空間模型分類法和 *k*NN 分類法是常用的分類方法，此二個分類方法主要不同於在分類時所要比對的特徵表示（representative）的個數，向量空間模型分類法的分類類別個數為比對相似度依據，每一個特徵表示就代表一個分類類別，而 *k*NN 分類法為 Instance-based

的分類法，用到特徵表示個數為訓練語料的文件數，每一特徵表示為一個文件的表示。

*k*NN 分類法的概念是選擇和測試文件最相似的 k 篇文章，由這 k 篇文章來決定測試文件的類別，Kin [2005]的實驗裡認為 k = 10 的情形下已有足夠的資訊來辨識類別，由於文件屬於多個類別，因此被選取的 10 個文件中選擇同時 n 篇文件以上所屬的類別為測試文件的類別，n 太小會選擇過多的文件，n 太大會有許多文件的類別太少甚至沒有。利用調適語料決定 n 的大小，在這裡測試結果 n = 3，也就是只要有類別包含這被選取 10 篇中任意 3 篇文件，該類別就會被認為是測試文件的類別。

A generalized instance set（GIS）algorithm [Lam, 1998] 的概念是將數個相似文件的特徵表示結合成一個特徵表示表示，所以 GIS 分類法的特徵表示個數會少於（或等於）*k*NN 分類法，和 *k*NN 分類法、向量空間模型分類法之間的特徵表示個數關係為：|*D*|＝ kNN 分類法≧GIS 分類法≧向量空間模型分類法＝|*C*|，|*D*|表示訓練語料的文件數，|*C*|表示分類類別的個數。

利用 GIS 的概念，把分類完全一樣的文件形成一個類別，如表 4-11中 D_1、D_2 和 D_4 三篇文章的類別為{*A*、*B*、*C*}，就將此三篇文件分為一類，而 D_3、D_5 和 *D6* 的類別和其他的類別不同，就各自形成一個類別，本來只有{*A*、*B*、*C*}三個類別，GIS 的概念下就形成{{*A*、*B*、*C*}、{*A*、*B*}、{*A*}、{*B*、*C*}}四個類別，在分類比較時需要 4 個特徵表示個數，測試文件只需和此 4 個 GIS 類別的特徵表示比較，最相似的類別就為測試文件的類別。

訓練文件	分類類別
D_1	{*A*、*B*、*C*}
D_2	{*A*、*B*、*C*}
D_3	{*A*、*B*}
D_4	{*A*、*B*、*C*}
D_5	{*A*}
D_6	{*B*、*C*}

表 4-11：訓練文件-GIS 類別 範例

以語料 2 做測試，表 4-12為不同的分類法和不同程度的詞彙分群個數之間調和平均值的比較，*k*NN 分類法的特徵表示個數都遠比 CIS 分類法或是向量空間模型分類法來得多。向量空間模型分類法在 600 個詞彙群組數時表現得比其他分類好來得差，但權重修正後調和平均值由 0.215 提昇至 0.338，不僅改善最多，調和平均值也比其他二者分類法高。

此三種分類法比較上來看，以表 4-12中 600 個詞彙群組數無做權重修正的調和平均值數據顯示，當其分類法的特徵表示代表的文件數越多，對於詞彙群組數的大小越會影響分類的調和平均值，尤其當詞彙群組數低於一個門檻之後，調和平均值會急遽下降（特徵表示代表的文件數越多，調和平均值下降越快）。而當詞彙群組數影響越大時，詞彙權重的修正的效果越大。

	分類法		
	***k*NN**	**GIS**	**向量空間模型分類法**
特徵表示個數	7496	1636	25
詞彙群組數	**詞彙權重修正前 / 後**		
無分群	0.321 / 0.316	0.339 / 0.349	0.345 / 0.376
2000 詞彙群組數	0.318 / 0.330	0.300 / 0.350	0.348 / 0.366
600 詞彙群組數	0.274 / 0.313	0.243 / 0.300	0.215 / 0.338

表 4-12：分類法和詞彙分群之間的比較(調和平均值)

表 4-13顯示不同的分類法和不同程度的詞彙分群個數之間平均每篇文件的執行秒數，GIS 分類法和向量空間模型分類法在 600 詞彙群組數時比起沒有分群能減少一半的執行時間，而 *k*NN 分類法因為要比對的特徵表示個數太多，經過詞彙分群還是不能有效的節省時間，僅減少原來 5%的時間。

詞彙群組數	**分類法**		
	***k*NN**	**GIS**	**向量空間模型分類法**

無分群	8.025	1.142	0.323
2000 詞彙群組數	7.853	0.905	0.259
600 詞彙群組數	7.68	0.463	0.151

表 4-13：分類法和詞彙分群的執行時間(秒)

第 5 章　結論和未來工作

在本論文中，主要針對實際美國專利文件分類的情形，以主類別和次類別為分類類別的語料個別進行測試，對於專利分類上利用 Distributional Clustering 有效的減少文字的向量空間維度，以主類別為分類類別的語料可以在 2995 到 1500 之間的詞彙群組數都有接近 0.78 的調和平均值，在沒有降低調和平均值下，減少一半的文字維度，以次類別為分類類別的語料詞彙群組數 8000 到 2000 之間未減少調和平均值減少至四分之一文字的向量空間維度。

在上述執行詞彙分群之後，若選擇更少的詞彙群組數就會造成調和平均值急速下降，因此對於分類的權重計算公式做修正，以改善因為過少的詞彙群組數造成 tf 值快速增加的問題，提高在少量的詞彙群組數之調和平均值，權重修正過後，以主類別為分類類別的語料在 200 個詞彙群組數情形下調和平均值為 0.73，提升 86% 的調和平均值，以次類別為分類類別的語料到 600 個詞彙群組數以上調和平均值均有 0.35 以上，提升 70%的調和平均值，同時也減少一半的執行時間，對於龐大的專利文件語料大大有效節省系統資源。

以語料 2 做測試，我們的分類法和 *k*NN 分類法、CIS 分類法在詞彙分群個數之間比較，我們的分類法在無分群和 2000 個詞彙群組數時表現其他二者來得好，600 個詞彙群組數時表現得比其他分類法來得差，但權重修正後調和平均值還是比其他二者好。當其分類法的特徵表示代表的文件數越多，詞彙群組數的大小對於分類的調和平均值影響越大；當詞彙群組數影響越大時，詞彙權重的修正的效果越大。

在處理專利文件過程中，有少數的文件的類別標記是不存在的，語料 2 中

13882 篇的專利文件中發現 25 篇有｛707/102.1、707/103、707/104、707/107.1、707/500、707/501、707/501.1、707/513、707/514、707/516、707/707}的類別標記是不存在的，這樣的錯誤比例相當低，但不能排除其他的文件類別標記錯誤產生，或是因為人工分類方式，不同的專家主觀造成的差異因素，而影響分類結果的正確性。

在實際應用在專利的分類系統中，尚有許多需要改進的地方，包括未探討各個欄位對於分類的影響，若能把對分類有幫助的欄位（如發明人、申請公司...等）加入文件分類參考的依據，對於分類也許可再提高正確性或加速分類進行。對專利內容的詞彙做篩選或做詞彙語意的延伸，如剔除與主題無關的詞彙和加入其他在文件中沒出現的相關詞彙，加強確認詞彙對於主題的重要性。因為專利文件的數量龐大，系統的效率更顯重要，不論是執行的時間、所佔用的系統資源...等，都關係於日後使用的方便性。

在可正確進行專利文件分類之後，希望可以自動地調整分類類別階層結構，新的主題產生時能自動給予新的分類項目並且分配到類別結構中正確的位置，當其中一個類別裡的文件數過多時，能再自動做細分類，將該類別內專利文件重新分配，做到完全的專利文件分類自動化。

參考文獻

Baker, L. D., McCallum, A.K., Distributional clustering of words for text classification, *Proceedings. of SIGIR*, pp. 96-103, 1998.

Belkin, N. J., Croft, W. B., Information filtering and information retrieval: two sides of the same coin?, *Communications of the ACM*, 35, 12, 29–38, 1992.

Chakrabarti, S., Dora, B., Agrawal, R., Raghavan, P., Using taxonomy, discriminants, and signatures for navigating in text databases, In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB)*, pp. 446-455, 1997.

Chakrabarti, S., Dora, B., Agrawal, R., Raghavan, P., Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies, In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB)*, 163-178, 1998.

Chen, W., Chang, X., Wang, H., Zhu, J., Yao, T., Automatic Word Clustering for Text Categorization Using Global Information, First Asia Information Retrieval Symposium (AIRS), pp.1-6, 2004.

Deerwester, S.C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A., Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 1, pp. 391-407, 1990.

Dhillon, I. S., Modha, D. S., Concept Decompositions for Large Sparse Text Data Using Clustering, *Machine Learning*, Vol. 42, No. 1, pp.143—175, 2001.

Fung , B., Wang , K., Ester, M., Hierarchical Document Clustering Using Frequent Itemsets, *Proceedings of the SIAM International Conference on Data Mining*, 59-70, 2003.

Hammouda, K., Kamel, M., Document similarity using a phrase indexing graph model, *Knowledge and Information Systems*, Vol. 6, No. 6, pp. 710-727, 2004.

Jing, L.; Huang, H.; Shi, H., Improved Feature Selection Approach TFIDF in Text Mining, *Proceedings International Conference on Machine Learning and Cybernetics*, Vol. 2, pp. 944-946, Beijing, 2002.

Kang, B. Y., Lee, S. J., Document indexing: a concept-based approach to term weight estimation, *Information Processing and Management*, 41(5): 1065-1080, 2005.

Kin, J. H., Huang, J. X., Jung, H. Y., Choi, K. S., Patent Document Retrieval and

Classification at KAIST, In *Proceedings of NII-NACSIS Test Collection for IR Systems Workshop*, pp 6-9, 2005.

Koster, C. H. A., Seutter , M., Beney, J., Classifying Patent Applications with Winnow, *Proceeding Benelearn Conference*, pp. 19-26, 2002.

Lam, W., Using a generalized instance set for automatic text categorization. In *Proceedings of the 21th annual international ACM SIGIR*, pp. 81-89, 1998.

Larkey, L. Some issues in the automatic classification of U.S. patents, In *Working Notes for the AAAI-98 Workshop on Learning for Text Categorization*, 1998.

Larkey, L. S., A patent search and classification system, In *the Fourth ACM Conference on Digital Libraries*, pp. 79-87, 1999.

Lavelli, A., Sebastiani, F., Zanoli, R., Distributional Term Representations: An Experimental Comparison, In *Proceedings of the Thirteenth ACM conference on Information and knowledge managment*, pp. 615-624, 2004.

Lertnattee, V., Theeramunkong, T., Multidimensional Text Classification for Drug Information, *IEEE Transactions on Information Technology in Biomedicine*, Vol. 8, No. 3, pp. 306-312, 2004.

Lin, D., Using syntactic dependency as local context to resolve word sense ambiguity, In *Proceedings of ACL/EACL-97*, pp. 64–71, 1997.

Ma, L., Chen, Q., Cai, L., An Adaptive System for Online Document Filtering, *IEEE International Conference*, Vol. 5, pp. 4712- 4717, 2003.

Mandhani, B., Joshi, S., Kummamuru, K., A Matrix Density Based Algorithm to Hierarchically Co-Cluster Documents and Words, *Proceedings of the international conference on World Wide Web*, pp. 511-518, 2003.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J., WordNet: An On-line Lexical Database, *International Journal of Lexicography*, Vol. 3, No. 4, 1990.

Pereira, F., Tishby, N., Lee, L., Distributional clustering of English word, *Porceedings of the 31st Annual Meeting of the Association for Conputational Linguistics*, pp. 183-90, 1993.

Richter, G., MacFarlane, A., The impact of metadata on the accuracy of automated patent classification, *World Patent Information*, Vol. 27, pp. 13-26, 2005.

Shah, C., Chowdhary, B., Bhattacharyya, P., Constructing Better Document, Vectors Universal Networking Language (UNL), *Proceedings of International Conference*

on Knowledge-Based Computer Systems (KBCS), 2002.

Wang, W., Do, D. B., Lin, X., Term Graph Model for Text Classification, *Advanced Data Mining and Applications*, pp. 19-30, 2005.

Wu, C., Liu, C. L., An exploration of topic-dependent feature-weighting for summary extraction. *Proceedings of the 2003 National Computer Symposium (NCS), Taiwan*, pp. 18-19, 2003

Zhang, Y., Heywood, N. Z., Milios, E., Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora, *Proceedings of the 7th annual ACM international workshop on Web information and data management WIDM*, pp. 51-58, 2005.

李駿翔，應用資料探勘分類技術於專利分析之研究，中原大學資訊管理研究所碩士論文，2003。

附錄一　美國專利文件範例

和欄位說明

United States Patent 6,734,864
Abgrall **May 11, 2004**

Re-generating a displayed image

Abstract

The present invention is a method and apparatus to re-generate a displayed image. Graphic information is retrieved from a first storage. A graphic controller uses the graphic information to generate the displayed image. The first storage is accessible to a processor and the graphic controller. The graphic information is stored in a second storage which is accessible to the processor. The stored graphic information is retrieved from the second storage. The stored graphic information is written to the first storage to cause the graphic controller to re-generate the displayed image.

Inventors: **Abgrall; Jean-Paul** (San Jose, CA)
Assignee: **Phoenix Technologies Ltd.** (Milpitas, CA)
Appl. No.: **247645**
Filed: **September 19, 2002**

Current U.S. Class: **345/537**; ***345/418***
Intern'l Class: G06F 013/00
Field of Search: 345/536-538,559,522,501,418,555,667

References Cited [Referenced By]

U.S. Patent Documents

5121345	Jun., 1992	Lentz.
5128995	Jul., 1992	Arnold et al.
5131089	Jul., 1992	Cole.

Parent Case Text

This is a continuation of Application Ser. No. 09/336,255 filed Jun. 18, 1999 now U.S. Pat. No. 6,542,160.

Claims

What is claimed is:

1. A method for regenerating an image comprising:

displaying the image using a display adapter;

retrieving image information from the display adapter, the image information being used by the display adapter for displaying the image;

專利編號（Patent Number）：每一篇專利文件都會給予一個編號。

核准日期（Date of Patent）：專申請核准的日期。

發明名稱（Title）：一個簡短明確符合發明的主題。

發明摘要（Abstract 或 Abstract of the Disclosure）：簡明扼要地描述發明的技術內容，以 150 字為限。

發明人（Inventors）：發明人的名字。

受托人（Assignee）：財產保管人的名字。

申請日期（Filing Date）：專利文件申請的日期。

美國分類編號（UPC）：依 UPC（United States Patent Classification）標準的分類號碼，依機能為分類，分類仔細，是世界上專利文件分類最細的標準。

國際分類編號（IPC）：依 IPC（International Patent Classification）標準的分類號碼。

參考引證（References Cited）：所引用的參考資料。

請求項（Claims）：明確定義發明者所要求保護之發明範圍。

詳細說明（Detailed Description）：專利發明的實例詳細說明。

發明背景（Background of the Invention）：內容為發明相關領域和背景技術之說明，描述所涉及相關技術的領域，和目前的技術缺點以及待解決之問題。

圖示之簡單說明（Brief Description of the Drawings）：簡單的描述所附圖示之內容。

發明總覽（Summary of the Invention）：發明專利的內容概述或是描述要求項的內容。