第一章　緒論

第一節　研究背景

多元貝氏定理(multimembership Bayesian)，簡稱MMB，早期曾經在醫療方面的應用為伊利諾理工學院開發的「MEDAS」醫療專家系統，利用醫療專家提供的知識以及病人的病徵計算出可能的病症，目的是希望在醫生診斷的過程中，藉由病人目前的單一或多個病徵，再綜合該病人本身的歷史病歷，計算出新病人所患病症(多個)的個別機率值，並儘可能推論出更多的病症給醫生作為決策時的參考，以便協助醫生診斷(Lee, Evens, Carmony, Trace, and Naeymi-Rad, 1991)。

隨著網際網路的迅速發展，MMB開始運用於網站分類系統(web site classification system, WSCS)，此時期的重要研究則是於中文網站階層式分類推論。網站階層推論的方式是將目標網站的網站詞集送進推論引擎中，而推論引擎透過這些詞集來計算該目標網站隸屬各個類別的機率值。最後針對不同的階層以及所有的字詞在各階層的機率值，計算出目標網站可能隸屬的階層(駱思安，2005)。

網際網路的發展不只是網站的數量增加，電子郵件的使用也不斷增加，然而電子信箱內太多的電子郵件，讓使用者在整理歸類上產生困擾，因此MMB在郵件分類推論的方式是將測試郵件的郵件詞集送進推論引擎中，而推論引擎透過這些詞集來計算該郵件隸屬各個類別的機率值。最後針對不同的類別，以及所有的字詞在各類別的機率值，計算出測試郵件可能隸屬的類別(王瑄榕，2008)。

資訊科技日益發達，數位化檔案增加的速度越來越快，所以檔案自動分類的重要性也漸漸增加，近年來MMB的應用莫過於自動分類，因此在文件自動分類上，文件分類推論的方式一樣是將測試

文件的文件詞集送進推論引擎中，而推論引擎透過這些詞集來計算該文件隸屬各個類別的機率值。當然，為了改良MMB分類器在文件分類的準確度，在分類推論之前，將各字詞於訓練文件群中的出現次數依照次數排列，最後利用四分位數的方式，將出現次數過少的字詞過濾掉，也就是個字詞都有自己的四分衛門檻值；以及利用字詞貢獻差異度判斷，在進行分類推論時不考慮那些不具分類特徵或不具代表性的字詞，運用這兩種方法後，MMB自動分類的準確率有明顯的提升(周政淇，2009)。

已知MMB在各領域進行過研究與探討，在分類領域上駱思安(2005)將其應用於階層式類別做研究，類別階層彼此的差異性從上至下遞減；而王瑄榕(2008)與周政淇(2009)則以單一階層類別進行分類並且使用字典輔助，前者類別彼此差異性大，後者則差異性小，原因為後者周政淇(2009)僅於資訊類的領域以及使用專業字典進行分類研究。

而近期的MMB分類研究，不是將類別階層縮小，就是使用專業字典處理雜訊問題。當然，MMB應該不只侷限於小範圍分類，而且並不是每一種類別都有相關的專業字典可以應用，所以本研究會探討MMB文件自動分類在階層式類別且不使用專業字典的情況下與周政淇(2009)所使用的方式相互比較，並加以改良以提升分類準確率。近期的MMB文件分類的方式有以下兩種(周政淇，2009)：

一、以四分位數的方式篩選掉出現次數過少的字詞。

經該實驗測試，配合該研究的知識庫，以刪除出現次數少於第三四分位數的字詞時，分類正確率為88%。

二、運用字詞貢獻差異度判斷篩選不重要的字詞。

經該實驗測試，配合該研究的知識庫，將字詞貢獻差異度的門檻值設定為0.35時，分類正確率為95%。

第二節　　研究目的

已知目前已提出的改良方式，經過測試後都有得到良好的結果，但是字詞的篩選與字詞貢獻差異度在判斷上，可能會因為使用的訓練與測試資料來源不同，導致結果也會不同。再加上MMB只考慮字詞有無出現，而不考慮文章中字詞出現的次數，畢竟字詞的出現次數會影響該字詞的重要性，而這一點MMB是完全不考慮的。

目前在字詞篩選上，已有四分位數的方式來篩選出現次數少的字詞，但是卻有可能因此排除重要的關鍵字，導致分類準確率降低；而字詞貢獻差異度判斷，則是設定一個固定的門檻值，讓全部的類別做篩選使用。不過，不同的訓練與測試資料的來源，在設定的門檻值上會產生難題。加上目前數位化文件產生極快，其內容已經不是只談單一領域而是多領域的應用，很難說這兩種判斷方式是否還能在未來適用於文件分類上。

為了能更提升MMB推論之文件自動分類的準確率，且在不使用任何字典的輔助下，本研究的目的詳細敘述如下：

一、文件分類前的字詞篩選

周政淇在2009年使用四分位數的方式篩選出現次數少的字詞，雖然該實驗得到蠻高的正確率，不過目前字詞篩選的方法仍然有突破的空間，畢竟字詞的篩選在文件分類前的準備是很重要的，而且每篇文章的字詞出現次數是不可預期的。所以，本研究會針對字詞的出現次數進行研究，來找尋一個適合MMB分類法的字詞篩選法。

二、動態產生字詞貢獻差異度的門檻值

從周政淇(2009)的實驗得知字詞貢獻差異度的處理會決定

MMB分類法準確率的高低，進行MMB分類前必須先對特徵貢獻度做處理比對。不過以固定的字詞貢獻差異度門檻值來判斷字詞所具備的分類特徵或代表性，會不適用於大多數的文件與類別。所以，本研究會針對動態產生門檻值進行研究與測試。

第三節　研究限制

本論文的研究範圍與研究限制如下：

一、研究對象

研究對象主要是從Elsevier Electronic Subscriptions SDOL電子期刊全文資料庫上下載總計2266篇文件，並依照該電子期刊所定義之類別，把收集來的文件分成4個大類別和底下細分出的22個小類別，關於類別定義請參照本論文第五章。

二、研究限制

(一)文件種類

本研究所收集的文件皆為英文的文件，不考慮其他語系或者英文與其他語言夾雜的文件，單純的對英文文件進行內容擷取與分析。

(二)不考慮詞性分析器的分析錯誤

使用詞性分析器(GENIA tagger)進行詞性分析。但是，由於詞性分析不在本研究之討論範圍，故本研究不考慮分析後的詞性是否為正確和錯誤原因。

(三)以詞性為名詞的字詞，進行知識訓練與分類推論

根據以往的文件分類研究，得知使用名詞分析後的準確率最高。所以，本研究排除訓練與測試文件中名詞以外的詞性，只保留名詞。

(四)不考慮原始文件分類定義的錯誤

全部文件皆依據Elsevier Electronic Subscriptions SDOL定義其類別，故本研究不考慮文件原始定義類別錯誤的問題，完全以該網站定義為主。

(五)文件內容擷取

文件內容擷取的部分，主要針對文件中的純文字部分進行擷取與分析，非純文字部分如圖、圖表、數學公式等等皆不為文件內容分析的依據。

第四節　研究流程

本研究的步驟為以下幾點：一、確定研究動機與目的。二、文獻探討，進行與研究相關的文獻探討。三、研究方法，確定可以達到研究目的的方法。四、系統建構與開發，依據研究方法為基礎開發系統。五、系統測試與實驗結果。六、結論與建議。如圖1-1。

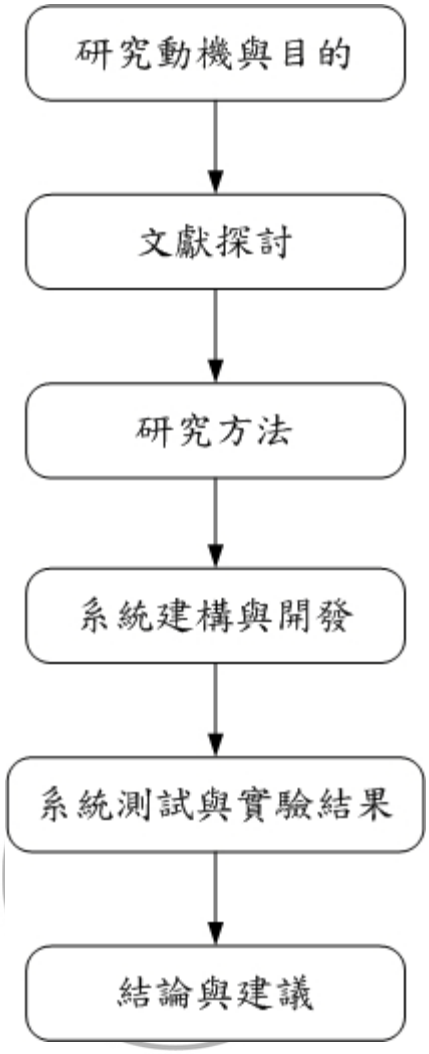

圖 1-1　研究流程圖

第二章　文件分類之文獻探討

第一節　文件分類

「文件分類」的意思是將一份文件依據其內容分類到一個或多個事先定義好的文件類別。早期的文件分類通常是經過專家的閱讀後，以人工判斷文件應該隸屬的類別。然而隨著網際網路的日新月異，網路上的資料猶如宇宙爆炸般迅速增加，如果還是仰賴以往的人工分類方式，則必須耗費相當多的時間去閱讀、篩選及過濾，來擷取該網站、文件、資料等的特徵與屬性，進而為此分類。而人工分類不僅耗時，也有可能因分類人員的認知不同等因素，造成分類錯誤。所以，在這資訊的時代，為了省時、提升效率、降低錯誤，於是許多專家學者紛紛投入文件自動分類的研究。

「文件自動分類」的意思是指利用電腦精確及快速的運算能力，依照某種數學模式、演算法，從測試文件或資料中自動擷取有代表性的特徵或關鍵字，並將其與知識庫中的知識相互比對，最後將性質相近的資料或文件聚集在一起完成分類工作。藉著文件自動分類來提高文件分類的正確性與一致性，以便於往後使用者能夠快速、正確的檢索到真正需要的資訊。

文件自動分類最早為Maron於1961年提出的想法。Maron 認為，人工分類文件時必須從文件的內容找到分類的線索，稱為關鍵詞(key words)。如果電腦也可以從文件中「自動」找出這些關鍵詞，那麼就可以做到所謂的自動分類(Maron, 1961)。

第二節　文件分類方法

目前大部分運用在文件分類上的方法。

一、支持向量機(support vector machines)

支持向量機(SVM)是以統計學習理論(statistical learning theory)為基礎發展出來的機器學習系統，屬於監督式學習的方法(Vapnik, 1995)。監督式學習是一個機器學習的方法，是由訓練資料中學到或建立一個模式，並依此模式推測新的例子。此概念是為找出一個超平面(hyperplane)來將分別屬於兩個類別的資料分開(林昕潔，2006)，如圖2-1。

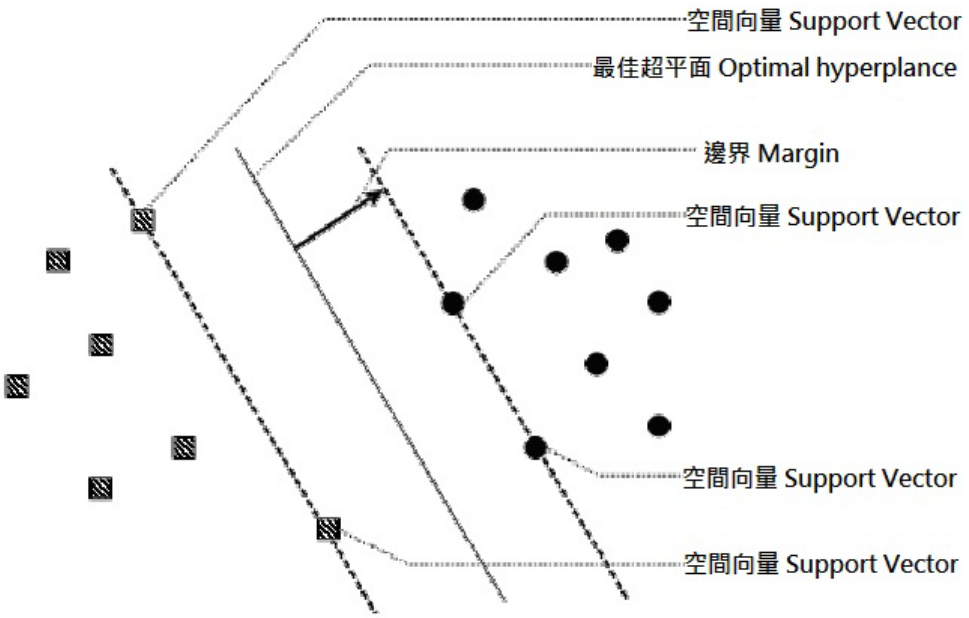

圖 2-1　支持向量機示意圖

資料來源：林昕潔(2006)，以SVM與詮釋資料設計書籍分類系統，國立交通大學資訊科學與工程研究所未出版之碩士論文。

這個超平面與最近的資點之間的距離稱為邊界(margin)，其邊界值越大越能將這些訓練資料明確的分開。也就是當有新資料且未知類別，能夠藉此超平面正確判斷新資料所屬的類別，找出一個最大邊界值的超平面(maximum-margin hyperplane)或稱為最佳超平面(optimal hyperplane)，而空間中最靠近最佳超平面的資料點稱為支持向量(support vector)。

假設有n個資料點$\{(w_1,c_1), (w_2,c_2),...,(w_i,c_i),\ldots, (w_n,c_n)\}$。其中$i=1,2,\ldots,n$且$c_i \in \{+1,-1\}$。$c_i$用+1與-1來表示資料點$X_i$的隸屬類別。用這些$(w_i,c_i)$作為訓練資料，找出最佳超平面來建構SVM，最佳超平面公式(2-1)如下：

$$w \times x - b = 0 \tag{2-1}$$

w代表邊界margin，b為一常數。

假設有兩個資料點x_1與x_2，分別為+1類別與-1類別，有兩個與最佳超平面平行且通過x_1與x_2的超平面p_1與p_2。計算邊界$|w|$的最小值，以知p_1：$w \times x_1 - b = +1$；p_2：$w \times x_2 - b = -1$，所以是$w \times (x_1 - x_2) = 2$，最後可得公式(2-2)。

$$\frac{w}{|w|} \times (x_1 - x_2) = \frac{2}{|w|} \tag{2-2}$$

另外，因為支持向量support vector為最佳超平面之訓練資料，因此不可能有其他訓練資料存在於p_1與p_2之間，則可得到條件式$w \times x_i + b \geqq 1$ or $w \times x_i - b \leqq 1$，可以寫成$c_i (w \times x_i) - b \geqq 1$，$1 \leqq i \leqq n$。

所以要解決$|w|$的最小值，必須同時符合上述條件式之條件，稱作二次規劃最佳化問題(quadratic programming optimization problem, QP)。

另外，林昕潔(2006)提到SVM 經學習之後，對於未知類別的新資 料，可依照$c=\{+1,\text{if } w\times x+b\geqq 0\}$ or $c=\{-1,\text{if } w\times x+b\leqq 0\}$的規則分 類(林昕潔，2006)。

由上式可以發現，SVM的優點在對未知 類別之資料進行分類時，僅是進行簡單的向量運算，並不會佔用太多運算時間。

二、貝氏分類法(naive Bayes classification)

貝氏分類法是根據貝氏定理作為推論的基礎，來推論出未分類的資料可能屬於或最接近的類別。因為貝氏分類法能夠容易且快速的實作，所以經常應用在文件分類上。主要是透過貝氏定理去計算出該資料隸屬於各個類別的機率值，藉此機率值的高低來判斷隸屬的類別。

傳統在文件分類上的應用，先利用多群已經事先分類的訓練文件來建立推論知識庫，然後依據測試文件內容來比對知識庫裡的推論知識，透過貝氏定理來計算出分類結果(Meena and Chandran, 2009)。

貝氏分類法在處理簡單的機率分類上，可以被定義為一個獨立的特徵模型，因為貝氏定理中每一個特徵假設都具有很強的獨立性(Guo, 2010)。

周政淇(2009)提到關於貝氏分類的方式。假設x為某個特徵值、C為某個類別。$P(x)$為x這個特徵值出現的機率；$P(C)$代表當我們任意取出特徵值時又恰巧為類別C的機率，$P(x)$與$P(C)$又稱為事前機率。接著根據條件機率，將其帶入以下的貝氏定理公式(2-3)做計算：

$$P(C|x)=\frac{P(C)\times P(x|C)}{P(x)} \tag{2-3}$$

$P(C|x)$代表當x特徵值出現的時候，又隸屬於類別C的機率，又可以稱為是事後機率；$P(x|C)$則代表當資料隸屬於類別C時，特徵值x發生的機率。以上僅說明貝氏分類的基本運用，而如果運用在文件分類上，類別不可能只有一個。

在多類別的文件分類上，假設有k個彼此互斥的類別$\{C_1, C_2, \ldots, C_k\}$，則可得特徵$x$出現時的事前機率公式(2-4)。

$$\begin{aligned} P(x) &= P(x \cap C_1) + P(x \cap C_2) + \ldots + P(x \cap C_k) \\ &= P(C_1) \times P(x|C_1) + P(C_2) \times P(x|C_2) + \ldots + P(C_k) \times P(x|C_k) \end{aligned} \quad (2\text{-}4)$$

同樣，根據條件機率，可以得到運用在k個類別的貝氏定理如公式(2-5)：

$$P(C_i|x) = \frac{P(C_i) \times P(x|C_i)}{P(C_1) \times P(x|C_1) + P(C_2) \times P(x|C_2) + \ldots\ldots + P(C_k) \times P(x|C_k)} \quad (2\text{-}5)$$

$P(C_i|x)$所得到的值代表特徵值x出現時，又剛好為類別C_i的機率。最後，比較各類別$C_1, C_2, \ldots, C_k$經過計算後的$P(C_i|x)$值，可得特徵值x最有可能隸屬的類別(周政淇，2009)。

Yong, Hodges, and Bo (2003)提出的分類應用，主要用於網路上的文件。透過自然語言處理(NLP)來分析與處理網路收集來的文件，NLP(natural language processing)就是要讓電腦「懂」人類的語言。將分析好的資料分成兩個部分，分別是訓練文件群與測試文件群。訓練文件為已經分類完成的文件，從已知隸屬類別的訓練文件來建構推論知識的知識庫。測試文件意指尚未分類或待分類的文件，將其輸入分類器，透過分類演算法以及知識庫的依據，計算出該文件可能隸屬某類別的機率值，比較各類別計算後的機率值高低，來決定最後的分類結果

(Yong, Hodges, and Bo, 2003)。如圖2-2。

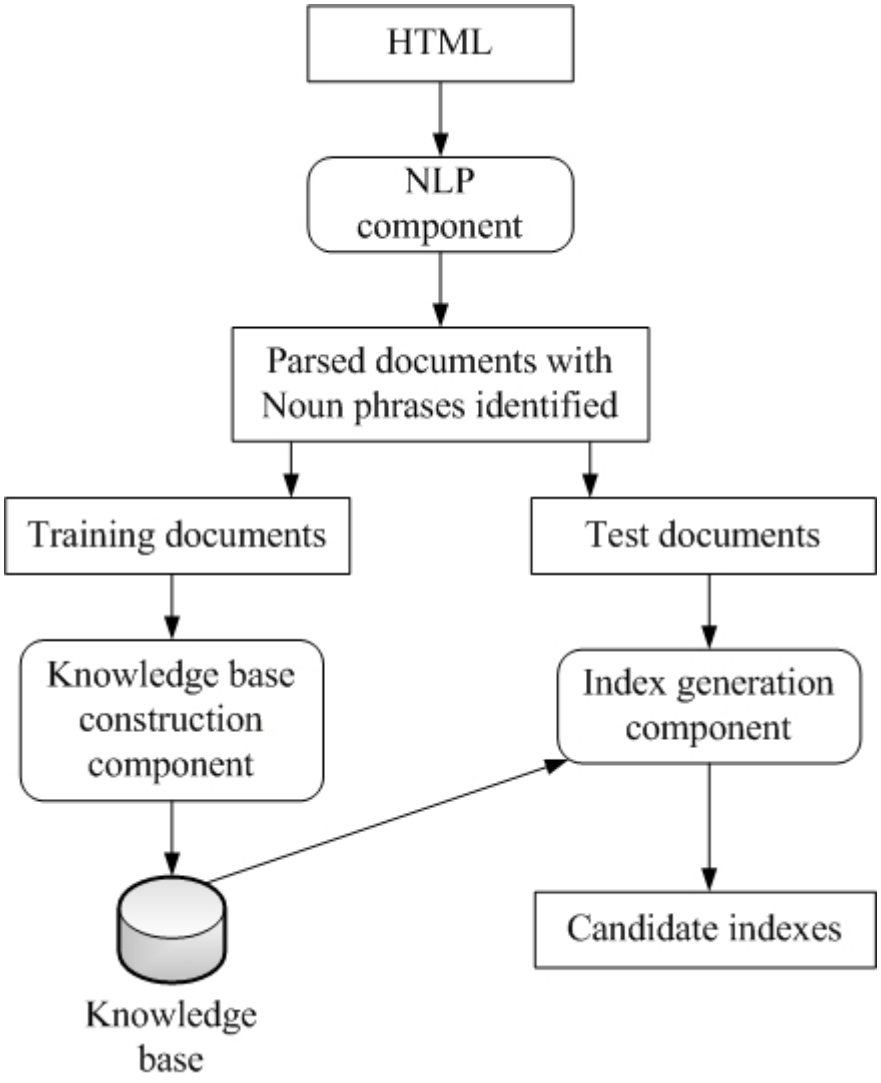

圖 2-2　貝氏文件分類示意圖

資料來源：W. Yong, J. Hodges, & T. Bo (2003). Classification of web documents using a naive Bayes method. *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on* (pp. 560-564). USA: Mississippi State University.

針對多類別的文件分類，Li and Jain (1998)提出了文件分類方法的比較，其研究比較了決策樹(decision tree)、最近鄰居

分群(nearest neighbor classifier)等方法，最後提出貝氏定理運用在文件分類上是一種簡單、實用及準確的方法。貝氏分類法推論方法如下，假設有m個類別$C(c_1, c_2, ..., c_m)$，文件D為未分類之文件，該文件D裡出現d'個字詞$W=\{W_1, W_2, ..., W_{d'}\}$。接著可得貝氏分類公式如公式(2-6)或公式(2-7)。

$$C_{NB}^{*} = \arg\max_{cj \in C} P(c_j) \prod_{i=1}^{d'} P(w_i \mid c_j) \tag{2-6}$$

$$\prod_{i=1}^{d'} P(w_i \mid c_j) = \frac{P(w_1) \times P(c_j \mid w_1)}{P(c_j)} \times \frac{P(w_2) \times P(c_j \mid w_2)}{P(c_j)} \times ..... \times \frac{P(w_{d'}) \times P(c_j \mid w_{d'})}{P(c_j)} \tag{2-7}$$

為了將所有字詞都考慮進來，所以$\prod_{i=1}^{d'} P(w_i \mid c_j)$代表針對該文件$D$裡出現的$d'$個字詞$W=\{W_1, W_2, ..., W_{d'}\}$，在類別$C_j$中出現機率的連乘積；$P(C_j)$代表任意文件又屬於$C_j$的事前機率，$C_{NB}^{*} = arg\ max_{cj \in C}$代表經過$m$個類別$C(c_1, c_2, ..., c_m)$的個別計算後，取最大機率值的類別當作該文件$D$的隸屬類別。

另外，同樣也是貝氏分類法的運用，在Kourd, Bensaid, and Rachidi (2004)的文章中，將貝氏定理應用於阿拉伯文章分類。其分類方法如下：

假設$D=\{W_1, W_2, ..., W_n\}$，W_n表示出現的字詞，D則為這些字詞的集合。

假設C_i為編號為i的類別。

假設$dosc_i$為尚未進行分類或確認，但是已經被列為編號i類別中的文件、$|Example|$為已經訓練過的文件數量。

$P(C_i) = dosc_i / |Example|$，$P(C_i)$表示亂數取出的文件屬於$C_i$類別的機率，稱做事前機率。如公式(2-8)

$$P(C_i \mid D) = [P(C_i) \times P(D \mid C_i)] / P(D) \qquad i = 1,2,...C \tag{2-8}$$

其中，$P(C_i|D)$表示當D這些字集出現時，又屬於C_i類別的機率稱為事後機率。在計算D在各類別C_i的機率後，取機率最高的當作該文件隸屬的類別。不過在這篇文章最後的測試結果，發現貝氏定理運用在阿拉伯文的文件分類上準確率不高。

在一般情況下，分類文件中的文字通常會表示為特徵或向量...等，來當作運算法的計算因素。通常不會只有一個詞彙，甚至可能幾十萬字。因此，減少非重要的特徵、空間或雜訊相對的重要。所以在進行分類之前篩選特徵，可以提升分類準確率與降低計算成本(Kim and Chang, 2007)。

貝氏分類法的優點在於其分類準確率優於其他的分類技術、推論引擎能夠在知識庫經過訓練之後很快的趨於穩定且貝氏分類法只需掃描一遍文件就可以建立模型(林傑斌，劉明德，2002)。

貝氏分類法的缺點在於只有考慮字詞出現的有無而不考慮出現頻率，因此在進行分類時，如果沒有搭配關鍵字提取...等方法，就必須考慮整篇文件的字詞，而其中絕對包含著不具分類特徵或代表性的字詞，最後的分類結果可能會降低準確率。

因此貝氏分類的應用，在分類前常常會搭配各種的關鍵字提取方式，來保留有意義的特徵或關鍵字，提升分類的準確率。

三、基因演算法(genetic algorithms)

基因演算法是一種模擬生物學上遺傳演化的過程。主要是利用基因演化和物競天擇的機制，來找出最佳特徵參數。基因演算法使用選擇(selection)、交配(crossover)和突變(mutation)等運算元(operator)繁衍出成功世代的各種模型。一般被用來當做搜尋最適解，亦即藉由演化的機制，為問題尋找最適當的解(莊惠美，1999)。而在文件分類的問題上，便是要找出最佳的

類別特徵組合。

基因演算法的典型運作方式如圖2-3。首先，隨機產生數條染色體，然後計算各條染色體的適應值。其次，選擇適應值較高的染色體進行重組。重組後得到的就是新染色體也就是後代。當然，如果新染色體未能符合需求，則必須繼續計算，直到最佳染色體出現為止(Wang, Hua, and Bai, 2008)。

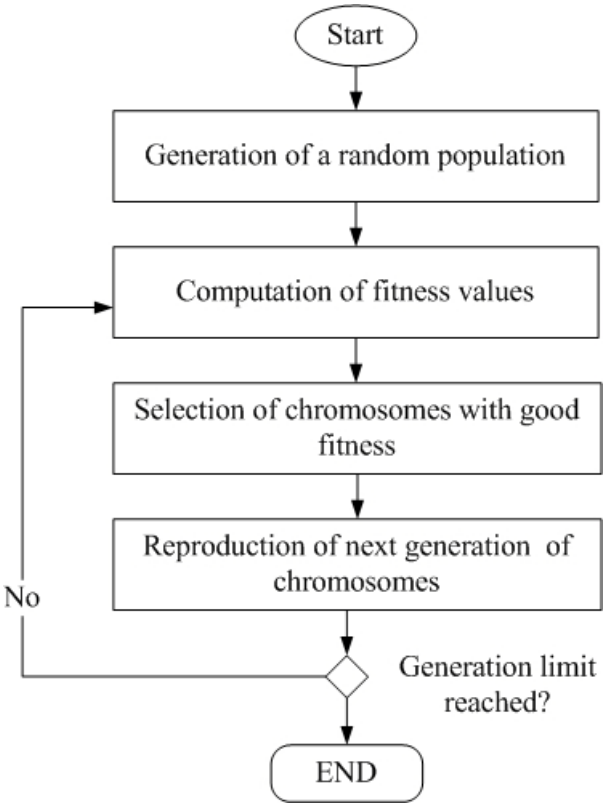

圖 2-3　基因演算法示意圖

資料來源：X. Wang, Z. Hua, & R. Bai (2008). A hybrid text classification model based on rough sets and genetic algorithms. *Software Engineering, Artificial Intelligence Networking, and Parallel/Distributed Computing, 2008. SNPD '08. Ninth ACIS International Conference on* (971-977). Zibo: University of Technol.

染色體依照不同的制訂方式通常為一串數值(浮點數、二進位值...等)，而重組染色體的方法有兩種，分別是交配(crossover)和突變(mutation)(莊惠美，1999)。

(一)交配(crossover)：包含單點交配、雙點交配與N點交配。

1.單點交配(one-point crossover)：選取兩個染色體當作父母，選取一個位置來進行交配，例如：選取位置2。

parent 1: 0 0 0 0 0 0

parent 2: 1 1 1 1 1 1

則得到兩個子代：

child 1: 0 1 0 0 0 0

child 2: 1 0 1 1 1 1

2. 雙點交配(two-point crossover)：選取兩個染色體當父母，選取兩個位置來進行交配，例如：選取位置3與位置4。

parent 1: 0 0 0 0 0 0

parent 2: 1 1 1 1 1 1

則得到兩個子代：

child 1: 0 0 1 1 0 0

child 2: 1 1 0 0 1 1

3. *N*點交配(N-point crossover)：選兩個染色體當父母，選取*N*個位置來進行交配，例如：變換偶數位置。

parent 1: 0 0 0 0 0 0

parent 2: 1 1 1 1 1 1

則得到兩個子代：

child 1: 0 1 0 1 0 1

child 2: 1 0 1 0 1 0

(二)突變(mutation):

一串染色體如下：

1 1 1 1 1 1

經過突變後：

0 1 1 1 1 1

而在文件分類上的應用，Liu, Lu, and Lee (2000)提到當使用基因演算法來解決問題時，最重要的是為要解決的問題選擇合適的表示法。其文章在判斷是否停止演進也就是是否找到最佳解時，是運用的是向量的方式。首先，針對每個類別都定義特徵集合，特徵集合為所有在此類別文件的特徵。假設類別*c*的特徵集合中集合有*n*個特徵，類別*c*中的每一篇文件內有*k*個特徵(*n*≦文件數 × *k*)。如此一來，每一篇文件的向量表示就可以一致。(每一個位置代表一個在特徵集合的一個字詞，即該字詞的有出現「1」或沒出現「0」)，如前述每一篇文件有*k*個特徵也就是有*k*個位置為1。在染色體處理上，將類別*c*裡*n*個特徵以符點數的方式轉換成染色體形式當作該特徵的權重$W_1, W_2, \ldots, W_n$。每一個W_i代表該位置與字詞的權重，權重值的範圍定義為-1與1之間。當一個文件向量出現時，則計算權重總和，因為每一篇文件有出現*k*個特徵，而且權重值在1與-1之間，所以每一文件權重總和介於*k*與-*k*之間，我們就將此文件歸類為本類類別;如果權重總和介於-*k*與0之間，則視為非本類別。經實驗證明能夠有效且正確的制定適當的特徵及其權重，而且在不用建構大型分類模型下採用分散式的方法，確實提升分類準確率。

同樣也是運用在文件分類上，Hossaini, Rahmani, and Setayeshi (2008)結合k-means algorithm來建立基因演算法分類器。所謂的k-means algorithm就是一個聚類演算法，把*n*個對象根據

他們的屬性分為*k*個分割，$k < n$。並且試圖找到數據中自然聚類的中心，以圖2-4來說明表示如下：

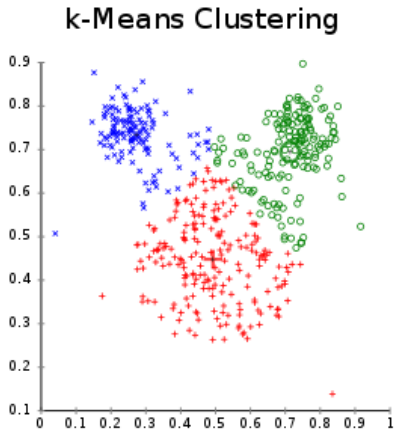

圖 2-4　k-means 演算法示意圖

資料來源：Wikipedia (2010). *Cluster analysis mouse.svg* [Online]. Available: http://en.wikipedia.org/wiki/File:ClusterAnalysis_Mouse.svg [2010, October 12].

可以略顯看三個群集，以文件分類的觀點來說，文件群(群集) 的中心代表較接近某類別的群集，而文件(個體)越接近中心，就代表該文件(個體)越可能屬於該類別。k-means algorithm計算如公式(2-9)。

$$V = \sum_{i=1}^{k} \sum_{Xj \in Si}^{k} (x_i \times u_i)^2 \qquad (2\text{-}9)$$

該篇文章依照根據訓練文件群在k-means algorithm的分布預先定義聚類的中心的近似的類別，然後其基因演算法運作方式為將每一篇文件皆轉成染色體表示，以「選擇」的方式，隨機抓取訓練文件與測試文件群，並計算該文件群的中心值。若

計算中未達接近標準時，則重新需「選擇」群集，繼續計算，直到文件群出現最接近某類別中心值的類別，則當作該測試文件的隸屬類別。

在特徵選擇上，Wang, Hua, and Bai (2008)利用基因演算法來提升SVM分類器的準確率。首先，將全部字詞以染色體表示，然後隨機選擇染色體群集後，利用SVM分類器進行分類。若未達適應標準(分類標準)，則利用基因演算法(交配、變異等)的方式，將以分類過的染色體群集重新產生新的世代(染色體群集)後，繼續以SVM分類器進行分類，直到符合適應標準(分類標準)。其適應值計算如公式(2-10)。

$$fitness = W_A \times SVM\_accuracy + W_F \times \sum_{i=1}^{nf} (C_i \times F_i)^{-1} \quad (2\text{-}10)$$

W_A代表SVM分類器的準確率權重。SVM_accuracy代表SVM分類器的準確率。W_F代表特徵數量的權重。C_i代表特徵i的花費成本。F_i則代表特徵i有出現「1」或沒出現「0」。

因此對於個體(染色體)是否貢獻度高、數量多少、特徵選擇成本低，該公式皆可以產生出具有高價值的適應值。

同樣利用基因演算法運用在特徵選擇上，Zhao, Wang, and Li (2010)提到基因演算法雖擁有隨機性，但收斂速度較慢，而k-means雖收斂快但在特徵選擇上較弱，於是將k-means與基因演算法結合運用，來補足彼此的優缺點。關於k-means的運算公式，如前提。結合運用的方法如下：

(一)選擇一個適應值最佳的個體M_i，作為初始聚類中心。

(二)在現有的聚類中心隨機選擇一個個體K_j與個體M計算相似度(k- means)。

(三)如果與聚類中心的相似度小於門檻值(0.45)，則以個體K_j為

中心建立一個新的群集(變異)；否則，把個體K_j放入最大相似度的群集(交配)。

(四)重複步驟 2和3，直到所有的個體在分群上不再有變動(達到演化上限)。

(五)保留每個群集的中心，刪除距離群集較遠的其他個體，然後與初始聚類中心M_i和保存最好的聚類中心K_j構成新群集(M+ K個)。

而這最後的新群集就是挑選完畢的最佳特徵。此特徵選擇的優點在於解決了基因演算法收斂慢的問題也解決了k-means特徵選擇不佳的問題，最重要的是實驗結果成功的提高了分類精確度和效率。

四、決策樹(decision tree)

決策樹是一種較為簡單的歸納式學習法且實作容易，所以應用層面廣泛，這裡僅介紹在文件分類方面的應用。

在決策樹樹狀圖中，每一個內部節點代表某個屬性的測試，向下的每一個分支代表此問題的一個可能值或多個可能值群。最後每一個樹葉節點所代表的是每一個對應的類別(韓歆儀，2003)。以一個簡單的決策樹樹狀圖來作基本概念說明，如圖2-5。

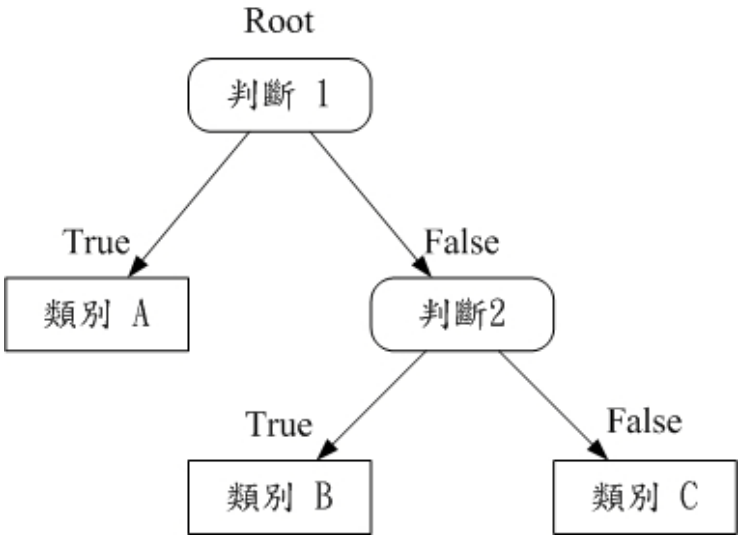

圖 2-5　決策樹示意圖

資料來源：韓歆儀(2003)，應用兩階段分類法提昇SVM法之分類準確率，成功大學工業管理科學研究所未出版之碩士論文。

如圖2-5所示，以文件分類方面的角度來說明，決策樹的根部(root)在圖頂端，當一份文件從根部進入後，每一次的判斷或決策(判斷1、判斷2)都會決定進入下一層哪一個子節點，直到文件到達最後的樹葉節點(類別A、類別B、類別C)為止，也就是代表該文件所隸屬目標類別。

決策樹的運行方式就是不斷的重複判斷或決策，直到文件到達最底層的樹葉節點。因此未到達樹葉節點以前，樹會不停的成長。不過，從眾多的文獻中了解，當樹的節點越多，測試或判斷的動作也越多，當然效率就越低。所以，在建立決策樹時，通常會選擇一些演算法，做適當的修剪(pruning)來提高效率。韓歆儀(2003)提到決策樹修剪有兩種較常用方法：

(一)預先修剪(pre-pruning)

預先修剪的目的是要提前停止樹的成長。當樹停止生長時，末端的節點會成為樹的樹葉，以文件分類來說就是

目標類別。通常制定一個門檻值來停止樹的成長，意即當分支的節點滿足已設定好的門檻值，就停止該分支的成長。

(二)事後修剪(post-pruning)

事後修剪顧名思義，就是當樹完整的建造後，再將其分支移除。主要是透過該分支的錯誤率(error rate)來決定是否移除，而最末端未被移除的節點就成為了新的樹葉。

現今學者認為以上兩中修剪方法，因為預先修剪所需要的門檻值設定，可能會過於主觀，所以在建立決策樹上較常用事後修剪的方法。

第三節　特徵選擇

一、TF-IDF

TF-IDF(term frequency - inverse document frequency)代表一個字詞出現在某文件的頻率越高且在其他文件中的頻率越低的話，該字詞就越具有代表性(Salton and Buckley, 1988)。在文件分類上，運用於關鍵字的擷取，並過濾與分類無關的雜訊，通常需要另外搭配分類法(例：貝氏分類法、支持向量機、類神經網路等等)，而不是直接用於分類推論。TF-IDF在資訊檢索上是常用的加權技術，主要用來過濾分類無關的文本雜訊。

TF(term frequency)詞頻為字詞在文章中出現的頻率，先統計字詞t_i在文件d_j出現的總次數$n_{i,j}$，然後再以字詞t_i出現的總次數$n_{i,j}$除以文件d_j內全部的字數$n_{i,j}$，得到的答案就是詞頻(劉德舜，2009)。如公式(2-11)。

$$Tf_{i,j} = n_{i,j} / \sum_k n_{k,j} \tag{2-11}$$

IDF(inverse document frequency)逆向文件頻率為關鍵文件出現頻率之倒數，目的為判斷字詞在文章的重要程度，當該字詞在很多文件中都有出現，則表示該字詞的代表性低；反之當該字詞出現在少數的文件中，則代表該字詞的代表性高。運算方法為全部文件的總數量除以包含字詞t_i的文件數，接著再取其對數，如公式(2-12)。

$$Idf_i = \log \frac{|D|}{|\{t_i \in d : d\}|} \tag{2-12}$$

|D|：全部的文件總數；$|\{t_i \in d : d\}|$：包含字詞t_i的文件數($n_i \neq 0$的文件數目)

得知TF與IDF後，兩者相乘的值就是字詞ti的加權值(weight)，如公式(2-13)。

$$TfIdf_{i,j} = Tf_{i,j} \times Idf_i \tag{2-13}$$

經過TF-IDF的計算後，得到各字詞的加權值，刪除過濾掉常見或者加權值較低的字詞，保留具有代表性與對分類有幫助的關鍵字。

二、Information Gain

Information Gain(IG)屬性選擇法，主要是以測 量資訊量多寡來計算各個類別的資訊量，並計算出該訓練集合的平均資訊量，也就是所謂的熵(entropy)來表達該集合中資料的複雜度。而以文件分類中特徵選取的角度來看，是根據字詞在各類別中有出現和無出現的機率計算該字詞的重要性。

林昕潔(2006)提到在進行分類時，定義訓 練資料的集合為S，假設類別的數量為m，則以$C_1,C_2,\ldots,C_m$代表這m個類別。假設d_i是屬於 類別C_i的訓練資料總數，pi是任一訓 練資料屬於C_i的機率，如公式(2-14)。

$$p_i = \frac{d_i}{|S|} \tag{2-14}$$

對一個給定的資料，可以以下列公式求出它的期望值，如公式(2-15)。

$$I(d_1, d_2, \ldots, d_m) = \sum_{i=1}^{m} p_i \log_2(p_i) \tag{2-15}$$

假設屬性A具有v個不同的值$a_1,a_2,\ldots,a_v$。然後用A將S劃分為

v個子集合$S_1,S_2,\ldots,S_v$，其中S_j 所包含的樣本屬於S且在A上的值為a_j。假設S_{ij}是子集合S_j中類別C_i的樣本數，根據由A劃分成之子集合的熵值為公式(2-16)。

$$E(A)=\sum_{j=1}^{v}\frac{S_{1j},S_{2j},\ldots,S_{mj}}{S}I(S_{1j},S_{2j},\ldots,S_{mj}) \tag{2-16}$$

其中$S_{1j},S_{2j},\ldots,S_{mj}$為第$j$個子集合的權重，並且等於子集合(即$A$值為$a_j$)中的樣本數除以$S$中的樣本個數。而熵值越小，則子集合劃分的純度越高。最後，在A上的分支的IG為公式(2-17)。

$$Gain(A)=I(d_1,d_2,\ldots,d_m)-E(A) \tag{2-17}$$

在Yang and Pedersen(1997)將IG值經過正規劃後，將門檻值設定為90%，也就是將90%以下的字詞過濾後，分類準確率最高。

三、*χ2 - statistic* ：

χ2 在文件分類中用於選擇出現次數適當的特徵。依據四種發生狀況來設定特徵參數，TF(W_x,C_i,1,1)、TF(W_x,C_i,1,0)、TF(W_x,C_i,0,1)以及TF(W_x,C_i,0,0)。「0」表示未出現；「1」表示有出現。例：TF(W_x,C_i,1,1) 表示 W_x 在類別 C_i 中出現；TF(W_x,C_i,1,0)代表特徵W_x在類別C_i以外的類別出現，依此類推。經過*χ2 - statistic* 公式計算後，保留值較高的特徵，如公式(2-18)。

$$\chi^2(c,w)=\frac{N\times(p\times s-q\times r)^2}{(p+r)\times(p+q)\times(q+s)\times(r+s)} \tag{2-18}$$

p表示w和c同時出現；q、r表示不是w出現就是c出現；s代表w和c皆不出現，而N代表總文件數。

在Han Joon Kim等(2007)的文章中，發現將 $\chi 2$ - *statistic* 篩選門檻值定為0.02時進行貝氏分類時效率最佳。

第三章　多元貝氏定理之文獻探討

多元貝氏定理(multimembership Bayesian) 簡稱MMB，意旨當已發生條件不只一個的時候，就可以使用多元貝氏定理。假設欲求特徵(W_1,W_2,...,W_j)發生且又屬於類別(C_1,C_2,...,C_n)的機率，這時就可以利用多元貝氏定理來解決這個問題。談到多元貝氏定理首先要介紹其核心P_{ij}與$\overline{P}_{ij}$，計算公式如公式(3-1)與公式(3-2)。

$$P_{ij} = P(W_j \mid C_i) = \frac{\text{特徵 } W_j \text{於類別} C_i \text{中發生之數量}}{\text{類別} C_i \text{發生樣本數}} \quad (3\text{-}1)$$

i代表類別編碼；j代表特徵編碼；而公式(3-1)則代表在類別C_i的情況下特徵W_j存在的機率為何。根據特徵W_j在類別C_i樣本當中所發生之次數來標示「K_1」($K_1 \geqq 1$)，若W_j沒有發生則標示「0」。

$$\overline{P}_{ij} = P(W_j \mid \overline{C}_i) = \frac{\text{特徵 } W_j \text{於類別} \overline{C}_i \text{中發生之數量}}{\text{類別} \overline{C}_i \text{發生樣本數}} \quad (3\text{-}2)$$

i代表類別編碼；j代表特徵編碼；而公式(3-2)則代表在類別$\overline{C_i}$(非C_i)發生的情況下特徵W_j存在的機率為何。根據W_j在$\overline{C_i}$樣本當中所發生之次數來標示「K_2」($K_2 \geqq 1$)，若W_j並沒有出現則標示為「0」。有了P_{ij}與$\overline{P}_{ij}$這兩個重要的元素之後，帶入多元貝式定理的計算公式如公式(3-3)。

$$P(C_i \mid W_1, W_2, ..., W_n) = \frac{P(C_i) \times P(W_1 \mid C_i) \times ... \times P(W_n \mid C_i)}{P(C_i) \times P(W_1 \mid C_i) \times ... \times P(W_n \mid C_i) + (1 - P(C_i)) \times P(W_1 \mid \overline{C}_i) \times ... P(W_n \mid \overline{C_i})} \quad (3\text{-}3)$$

公式(3-3)代表在特徵W_1, W_2 ,..., W_n都發生的情況下，而且屬於類別C_i的機率。若以下有論及多元貝氏定理公式的，皆以此算式的變數為基準。而多元貝氏定理，以下簡稱MMB，曾應用於醫療診斷、網站分類、郵件分類和文件分類上。

第一節　MMB 於醫療診斷之應用

在1991年時，美國伊利諾理工學院發展的MEDAS醫療診斷專家系統利用MMB的推論當作醫生決策的參考，希望能藉著一些病人過去的病例，推論出新病人患病的機率，以協助醫生能更快速的診斷。MEDAS應用的MMB公式如公式(3-3)，表3-1為變數名稱。

表 3-1　MMB 醫療診斷之變數名稱說明

代號	意義
C_i	編號為 i 病症
$\overline{C_i}$	編號為非 i 病症
W_j	編號為 j 的病徵

MMB的推論知識源自醫生、專家或病人之訪談，其知識庫中是由P_{ij}如公式(3-1)和$\overline{P}_{ij}$如公式(3-2)所組成。P_{ij}代表患有C_i病症的情況下W_j病徵發生的機率為何；$\overline{P}_{ij}$代表患有$\overline{C_i}$ (非C_i)病症的情況下W_j病徵發生的機率為何。公式(3-3)則是表示W_1,W_2 ,...,W_n這些病徵發生時又患有C_i病症的後天機率值；$P(C_i)$表示患有C_i病症的先天機率值；$P(\overline{C_i})$表示患有$\overline{C_i}$ (非C_i)病症的先天機率值。特徵可能有性別、血型、血球比率、其他病症...等等。當然，最後取機率最高者來決定該病患所可能患病的病症(Lee, Evens, Carmony, Trace, and Naeymi-Rad, 1991)。

第二節　MMB 於網站分類之應用

在2004年時，將MMB運用在中文商業網站分類的推論方法，藉著該網站類別內之網站的某些重要的關鍵字，推論出新網站屬於該類別的可能性為何。其MMB推論知識庫的知識產生如圖3-1。

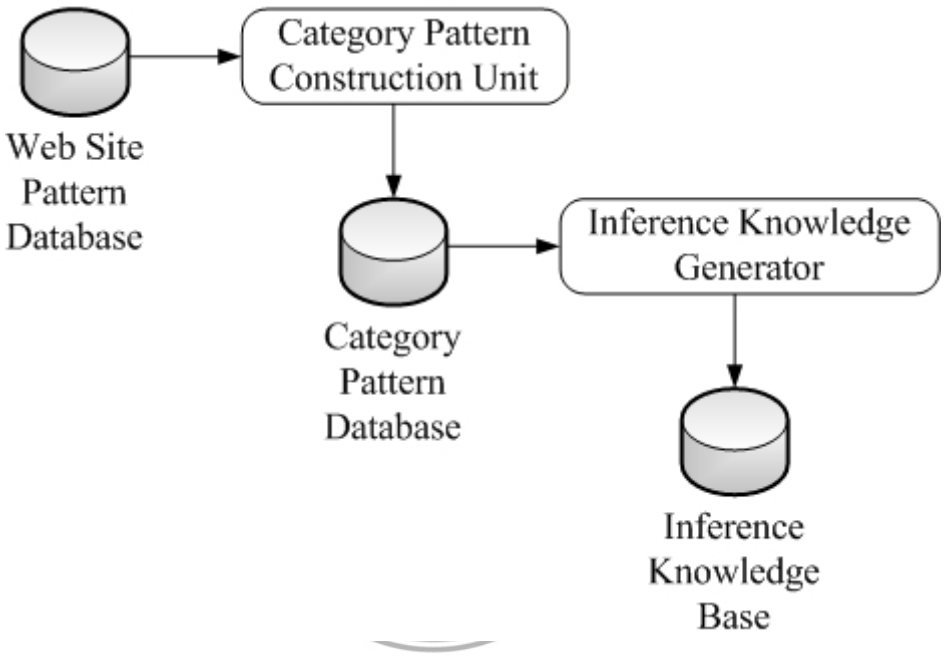

圖 3-1 多元貝氏定理之網站分類知識產生示意圖

資料來源：C. Y. Lee, H. Wu, & C. C. Yu (2004). Decision on classifying Chinese commercial web sites by Bayesian approach. *Paper presented at the 4th Annual Hawaii International Conference on Business, Honolulu*.

Web Site Pattern Database裡存著數個網站的訓練字詞樣本，例如：D_1網站字詞為W_1、W_2和W_3；D_2網站字詞為W_1、W_3和W_4；D_3網站字詞為W_4、W_5和W_6而D_4網站字詞為W_1、W_4和W_6。Category Pattern Database裡存放著網站類別，例如：C_1為書店、C_2為銀行。假設D_1與D_2屬於C_1的類別、D_3與D_4屬於C_2類別。知識推論的方式，例

如：字詞W_1若有在類別C_1中的網站D_1出現標示為(+)，若無則標示為(-)。可以得到表3-2。

表 3-2　網站字詞出現記錄

Category patterns	*Web site patterns*	*The common feature set*					
		W_1	W_2	W_3	W_4	W_5	W_6
C_1	W_1	+	+	+	-	-	-
	W_2	+	-	+	+	-	-
C_2	W_3	-	-	-	+	+	+
	W_4	+	-	-	+	-	+

而Inference knowledge database裡存放著各字詞W_j的P_{ij}如公式(3-1)和$\overline{P}_{ij}$如公式(3-2)。P_{ij}代表屬於C_i類別的情況下W_j字詞出現的機率為何；$\overline{P}_{ij}$代表屬於$\overline{C_i}$ (非C_i) 類別的情況下W_j字詞出現的機率為何。表3-3為MMB網站自動分類之變數名稱說明。

表 3-3　MMB網站自動分類之變數名稱說明

代號	意義
C_i	編號為 i 網站類別
$\overline{C_i}$	編號為非 i 網站類別
W_j	編號為 j 的字詞

承接表3-2的結果再搭配P_{ij}如公式(3-1)和$\overline{P}_{ij}$如公式(3-2)，可得表3-4：

表 3-4　網站字詞 W_1、W_2、W_3、W_4、W_5與 W_6的 P_{ij}與 $\overline{P}_{ij}$值

	C_1		C_2	
	P_{ij}	$\overline{P}_{ij}$	P_{ij}	$\overline{P}_{ij}$
W_1	1.00	0.50	0.50	1.00
W_2	0.50	0.00	0.00	0.50
W_3	1.00	0.00	0.00	1.00
W_4	0.50	1.00	1.00	0.50
W_5	0	0.50	0.50	0.00
W_6	0	1.00	1.00	0.50

MMB推論公式搭配所得到各字詞W_j的P_{ij}與$\overline{P}_{ij}$值，如公式(3-3)，表示W_1,W_2 ,…,W_n這些字詞出現時又屬於有C_i類別的後天機率值。$P(C_i)$表示屬於C_i類別的先天機率值；$P(\overline{C_i})$表示屬於$\overline{C_i}$(非C_i)類別的先天機率值。

透過上述即可計算出當網站D_x裡包含的字詞W_j都出現時，屬於類別C_i的機率有多少(Lee , Wu , and Yu, 2004)。

再來是2005年到2006年之間提出的MMB中文網站階層式分類推論，此分類系統包括三大模組，知識建構模組、推論引擎模組和知識學習模組。知識建構的流程如圖3-2。

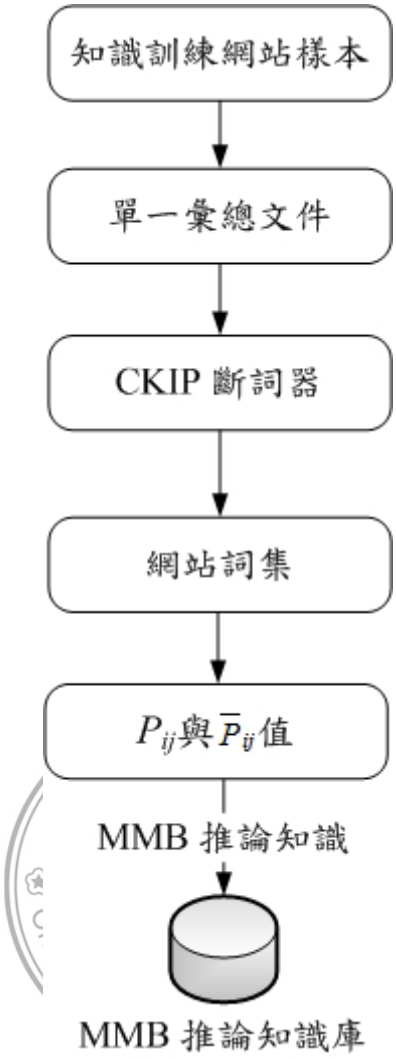

圖 3-2　多元貝氏定理之網站知識建構示意圖

資料來源：駱思安，李中彥，徐俊傑(2005)，以多重關係貝式演算法建構中文網頁自動分類系統，中華民國資訊學會通訊(IICM)。

首先將訓練網站樣本內所有的網頁文字彙總為單一文件後，並使用中研院開發的CKIP斷詞器進行斷詞，將斷詞後的結果去除名詞以外的字詞，在去除贅詞(例：加加油)、一字詞(例：「書籍」取代「書」)以及同義詞(例：「書籍」取代「書本」)後，彙總為網站詞集，接著以每個名詞在每個類別出現的次數推算出P_{ij}公式(3-1)和$\overline{P}_{ij}$公式(3-2)，並將其存入MMB推論知識庫。變數名稱如表3-3。

其中，P_{ij}代表在C_i類別裡的所有知識訓練網頁樣本當中字詞W_j出現的機率，根據W_j在C_i類別裡的知識訓練網站樣本當中所出現之樣本數來標示「K_1」($K_1 \geqq 1$)，若W_j沒有出現，則標示「0」；$\overline{P}_{ij}$ 代表在$\overline{C_i}$ (非C_i)類別裡的所有知識訓練網站樣本當中，W_j會出現的機率，根據W_j在$\overline{C_i}$ 類別裡的知識訓練網站樣本當中所出現的樣本數來標示K_2」($K_2 \geqq 1$)，若W_j沒有出現則標示「0」即可。

接著，知識推論的流程如圖3-3。

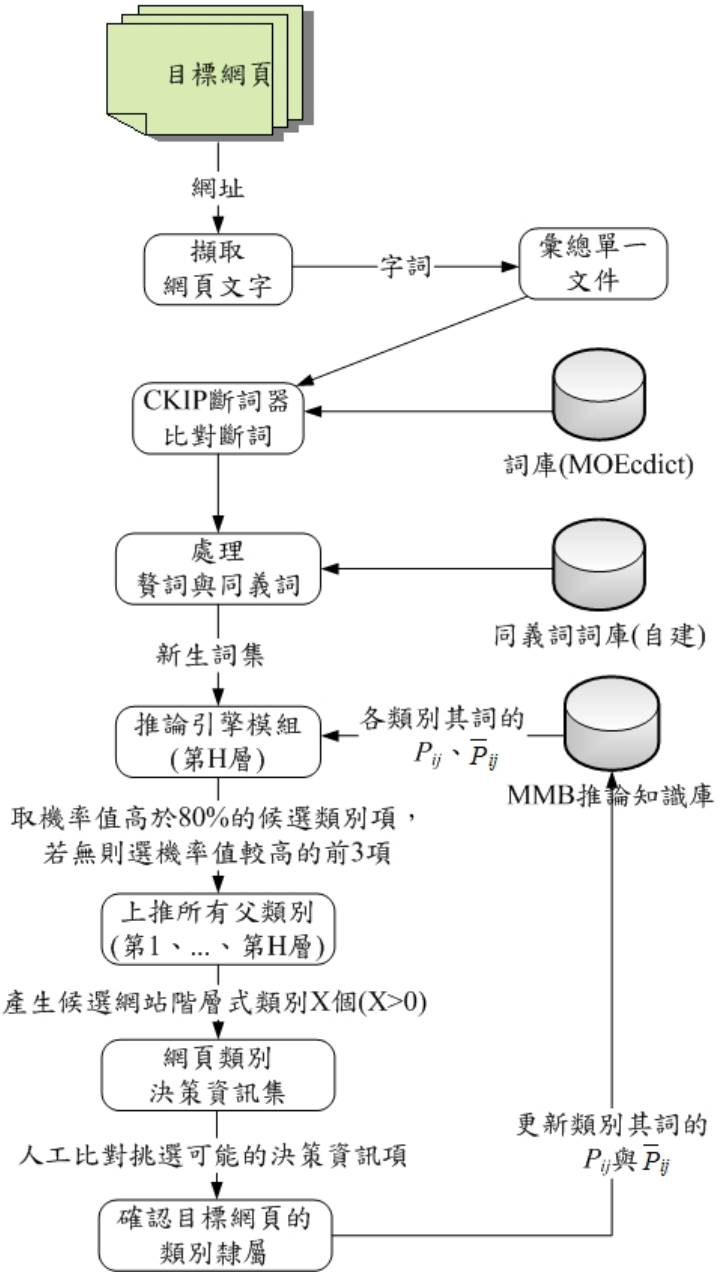

圖 3-3　多元貝氏定理之網站知識推論示意圖

資料來源：駱思安，李中彥，徐俊傑(2005)，以多重關係貝式演算法建構中文網頁自動分類系統，中華民國資訊學會通訊　(IICM)。

一開始的步驟跟知識建構相同，將目標網頁內容的所有文字彙

總成單一文件，再以CKIP斷詞器配合詞庫來斷詞，斷詞後去除名詞以外的字詞，在去除贅詞、一字詞以及同義詞後，彙總為網站詞集。再來就是與知識建構不同的地方，將網站詞集送進知識推論引擎中，來推論目標網頁隸屬於網頁分類目錄架構。假設目前在某一個第H層類別，同一時間內，推論引擎模組也會到類別第H層推論知識庫抓取詞彙的P_{ij}與$\overline{P}_{ij}$，來當做推論引擎模組的推論依據，以此類推H-1、H-2.....。最後，取P_{ij}與$\overline{P}_{ij}$高於門檻值80%或前三高的類別，若高於門檻值但機率值卻過於接近或、三高低於門檻值、或假設高於門檻值有兩個類別其機率值分別為84%與85%的情況，則採用人工比對的方式挑選最有可能的類別項，最後在事先建構好的WSCS階層式網站分類目路架構下賦予該網站的階層式類別隸屬。而推論引擎依據MMB演算法為基礎進行推論，MMB演算法的公式如公式(3-3)。

$P(C_i \mid W_1, W_2, \ldots, W_n)$代表目標網站包含有詞$W_1, W_2, \ldots, W_n$且屬於類別$C_i$的後天機率值(posterior probability)，而$P(C_i)$則是目標網站屬於類別C_i的先天機率(prior probability)，此分類系統將$P(C_i)$的值預設為0.5，也就是一開始在沒有任何預設立場的情況下，分類系統判別的目標網站屬於第H層各類別$(C_1, C_2, \ldots, C_m)$的個別機率值均相等。其中，$1<i<m<\infty$；$1<j<n<\infty$。

最後，透過知識學習，依照不同的網站樣本透過知識學習元件，不斷的更新知識庫裡的P_{ij}、$\overline{P}_{ij}$與字詞，才能適應任何的改變取得優勢(駱思安，李中彥，徐俊傑，2005)。

然而，駱思安，李中彥與徐俊傑(2005)的研究發現，以MMB為基礎的網站分類的確可以達到高準確率的分類結果。不過，決定最後分類結果的門檻值卻有個很大的問題。雖有提到若高於門檻值80%且機率值接近的類別，則利用人工的分是進行分類比對，但是如果兩個類別的機率值分別是79%與81%，而該研究處理方式為保

留81%的類別而不去考慮79%的類別，在加上MMB分類法的缺點之一就是沒有考慮到字詞出現次數，而只是單純的考慮出現的有無，這樣就有爭議性，也有分類不準確的風險。

第三節　MMB 於郵件分類之應用

以MMB為分類推論基礎的應用，在2008年時應用於中文電子郵件分類上。電子郵件的便利性讓使用人數不斷增加，然而電子信箱內太多的郵件，將讓使用者浪費時間整理歸類，因此將運用MMB在過濾中文電子郵件或分類郵件，而其知識建構的流程如圖3-4。

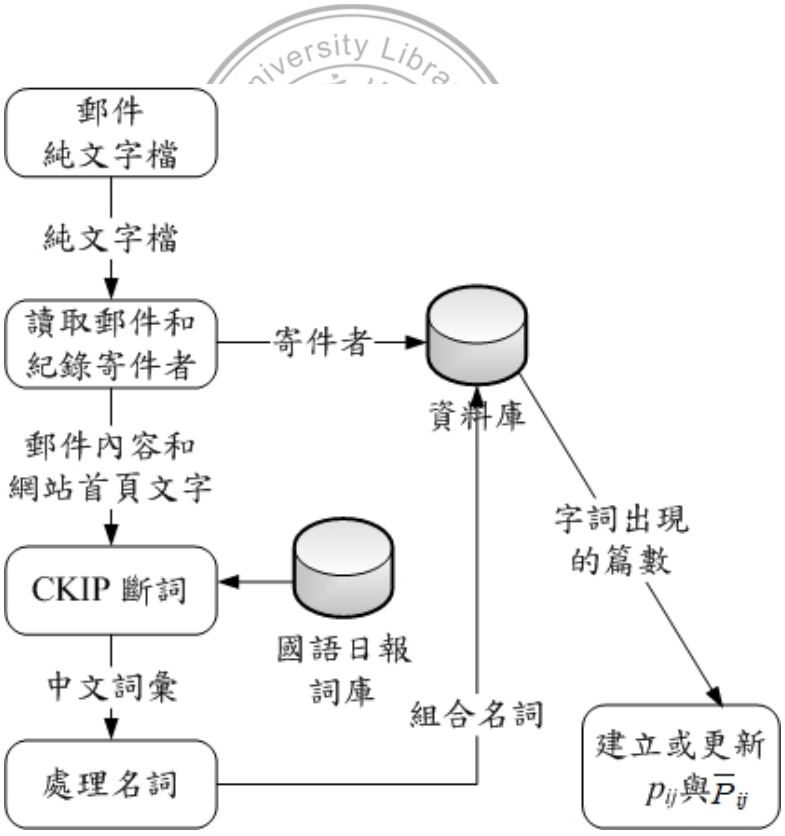

圖 3-4　多元貝氏定理之郵件知識建構示意圖

資料來源：王瑄榕(2008)，應用多元貝氏理論於中文郵件分類及知識建立，中國文化大學資訊管理研究所未出版之碩士論文。

知識建構的方法其實與MMB中文網站階層式分類推論(駱思安，李中彥，徐俊傑，2005)相似，首先讀取信件內的網址和寄件者，並存入資料庫；信件內容與信件內網址內容轉存為純文字檔案，以CKIP斷詞器搭配國語日報的詞庫，取出名詞和組合名詞後存入資料庫並計算這些名詞的P_{ij}如公式(3-1)和$\overline{P}_{ij}$如公式(3-2)。表3-5為MMB郵件自動分類之變數名稱說明。

表 3-5　MMB 郵件自動分類之變數名稱說明

代號	意義
C_i	編號為 i 郵件類別
$\overline{C_i}$	編號為非 i 郵件類別
W_j	編號為 j 的字詞

其中，P_{ij}表示在類別C_i裡的所有知識訓練郵件樣本當中W_j出現的機率，根據W_j在C_i類別裡的知識訓練郵件樣本當中所出現之樣本數來計算的，若W_j沒有出現則標示「0」；$\overline{P}_{ij}$代表在類別$\overline{C_i}$(非C_i)裡的所有知識訓練郵件樣本當中W_j會出現的機率，根據W_j在類別$\overline{C_i}$裡的知識訓練郵件樣本當中所出現的樣本數來計算的，若W_j沒有出現則標示「0」即可。

最後是分類推論部分，如圖3-5。

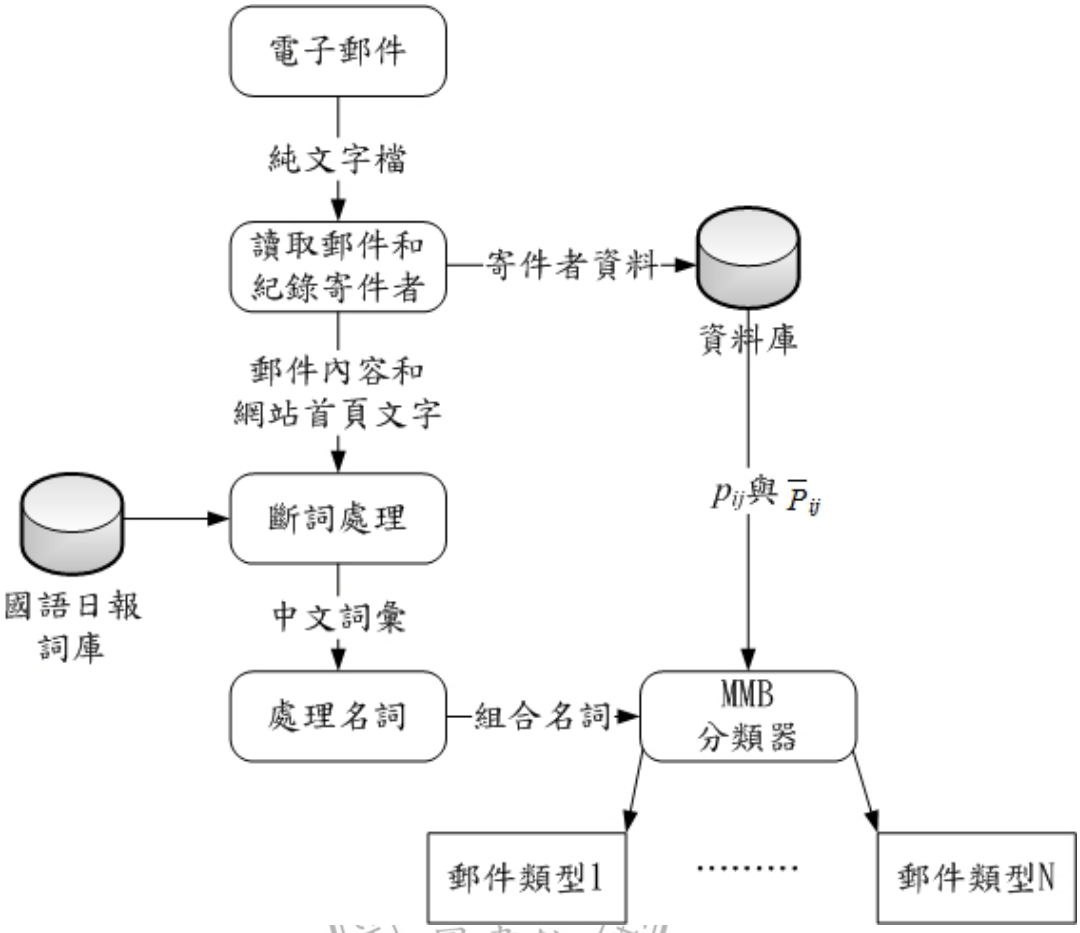

圖 3-5　多元貝氏定理之郵件知識推論示意圖

資料來源：王瑄榕(2008)，應用多元貝氏理論於中文郵件分類及知識建立，中國文化大學資訊管理研究所未出版之碩士論文。

分類器分類的方式則將想要分類的信件，經過斷詞、保留名詞與組合名詞處理後，依結果裡各個名詞，從MMB資料庫中取出該名詞或組合名詞在某類別所代表的P_{ij}與$\overline{P}_{ij}$後，輸入至MMB分類器來分類出郵件隸屬的類別。而推論引擎依據MMB演算法為基礎進行推論，MMB的演算公式與說明如公式(3-3)。$P(C_i \mid W_1, W_2, \ldots, W_n)$代表目標郵件包含有詞彙$W_1, W_2, \ldots, W_n$後屬於類別$C_i$的機率值，而$P(C_i)$是目標郵件屬於類別$C_i$的機率值。最後依照不同的郵件樣本，不斷的更新MMB資料庫裡的P_{ij}、$\overline{P}_{ij}$與字詞，來提升未來在分類上

的準確率(王瑄榕，2008)。

第四節　MMB 於文件分類之應用

時代不斷的進步，科技日益發達，數位化檔案增加的速度越來越快，所以檔案自動分類的重要性也慢慢增加，MMB近年來的應用莫過於自動分類，因此在MMB文件自動分類上，大致上知識推論的基礎都是相同的，同樣是利用字詞在各類別訓練文件中出現的文件數計算出P_{ij}與$\overline{P}_{ij}$，接著帶入MMB知識推論來分類。以周政淇(2009)的MMB自動分類系統來說，首先是知識建構流程如圖3-6。

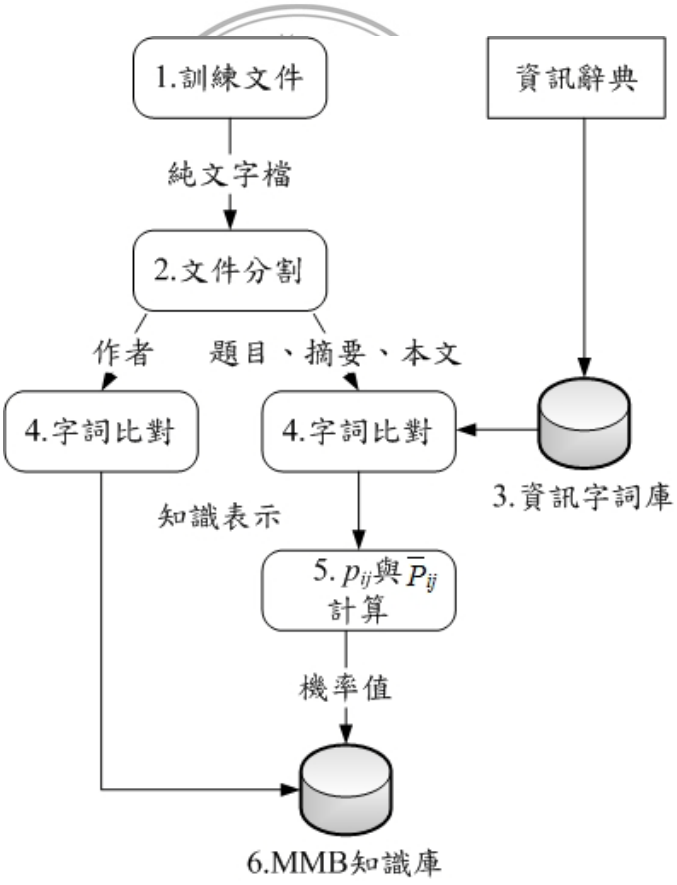

圖 3-6　多元貝氏定理之文件知識建構示意圖

資料來源：周政淇(2009)，以多元貝氏定理建構文件分類系統，中國文化大學資訊管理研究所未出版之碩士論文。

將訓練文件轉換成純文字檔後，將文件分割成「作者」和「題目、摘要、本文」，作者的部分則直接存入MMB知識庫，而「題目、摘要、本文」則要經過篩選，因為此系統所處理的項目皆為資訊類相關的文章，所以在字詞篩選的時候，特別針對資訊字來做比對與處理。處理過後，將剩下的資訊字詞經過計算求得個資訊字詞在各類別的P_{ij}如公式(3-1)和$\overline{P}_{ij}$如公式(3-2)並將其存入MMB知識庫，當作推論知識的依據。變數名稱如表3-6。

表 3-6　MMB 文件自動分類之變數名稱說明

代號	意義
C_i	編號為 i 文件類別
$\overline{C_i}$	編號為非 i 文件類別
W_j	編號為 j 的字詞

其中，P_{ij}表示在類別C_i裡的所有知識訓練文件樣本當中W_j出現的機率，根據W_j在類別C_i裡的知識訓練文件樣本當中所出現之樣本數來計算的，若W_j出現則標示「1」；若W_j沒有出現則標示「0」。$\overline{P}_{ij}$代表在類別$\overline{C_i}$(非C_i)裡的所有知識訓練文件樣本當中，W_j會出現的機率，根據W_j在類別$\overline{C_i}$裡的知識訓練文件樣本中所出現的樣本數來來計算的，若W_j出現則標示「1」，若W_j沒有出現則標示「0」即可。

最後是知識推論，針對題目、摘要、本文推論，如圖3-7。

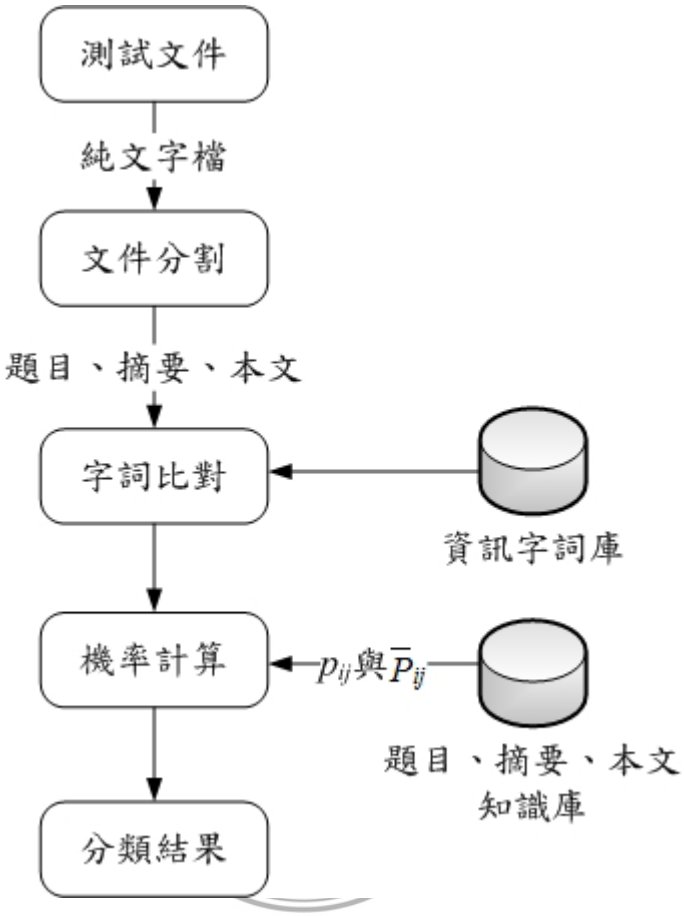

圖 3-7　多元貝氏定理之文件知識推論示意圖

資料來源：周政淇(2009)，以多元貝氏定理建構文件分類系統，中國文化大學資訊管理研究所未出版之碩士論文。

此為較一般MMB推論的流程，將測試文件的題目、摘要、本文等進行切割後，利用資訊字詞庫進行比對。最後將比對完的資訊字詞對照在各別MMB知識庫中的P_{ij}與$\overline{P}_{ij}$值，並將這些資訊字詞的P_{ij}與$\overline{P}_{ij}$值帶入MMB推論公式中後得到分類結果。

當然，為了改善分類的準確率，做了兩種改善方法，如下：

(一)加入門檻值得推論流程

加設此門檻值的目的在於去除掉一些出現次數較少的資訊字詞，在此這裡是取資料庫中各個字詞所有的出現次數以四分位數為其篩選門檻值，而該分類正確率為依照第一四分位數、中位數、第三四分位數的順序遞增，也就是表示出現次數的多寡往往會影響該字詞是否為關鍵字。如圖3-8。

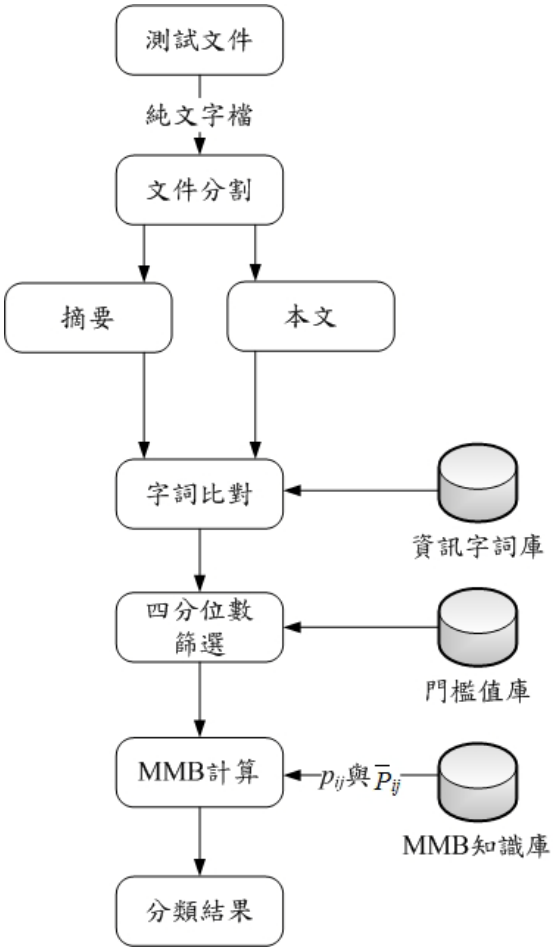

圖 3-8　多元貝氏定理之文件知識推論(字詞篩選)示意圖

資料來源：周政淇(2009)，以多元貝氏定理建構文件分類系統，中國文化大學資訊管理研究所未出版之碩士論文。

(二)字詞貢獻差異度篩選的推論流程

當某一個資訊字詞在該類別中的P_{ij}與$\overline{P}_{ij}$太接近造成不具分類特徵，重要性也較低，P_{ij}表示在該類別中此資訊字詞出現機率，而$\overline{P}_{ij}$表示此資訊字詞在其他類別出現的機率，當這兩個數值太過接近時就代表資訊字詞不具有分辨類別的特徵，故在機率計算時不考慮P_{ij}與$\overline{P}_{ij}$接近的字詞，如圖3-9。

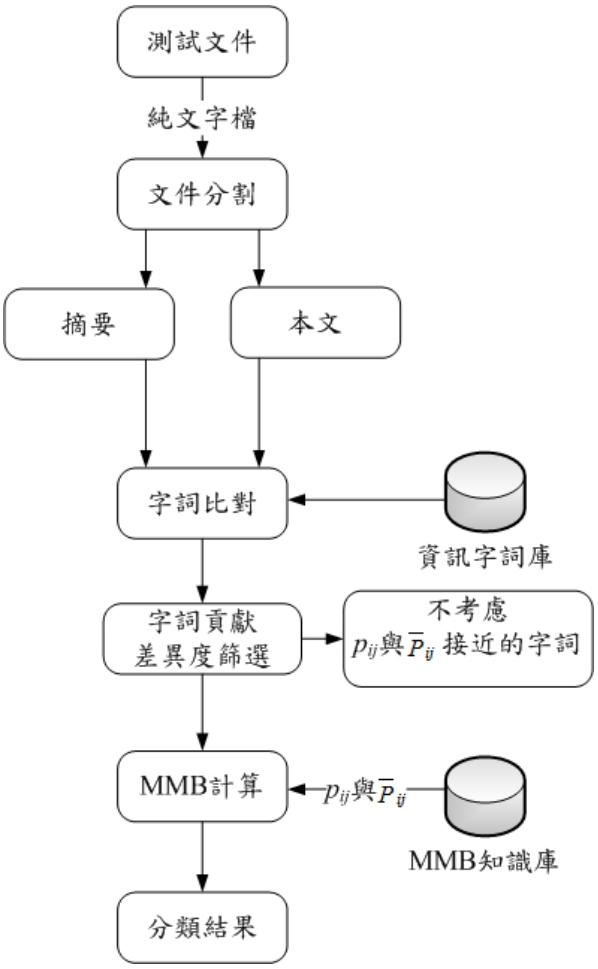

圖 3-9　多元貝氏定理之文件知識推論(差異度篩選)示意圖

資料來源：周政淇(2009)，以多元貝氏定理建構文件分類系統，中國文化大學資訊管理研究所未出版之碩士論文。

經該實驗證明，運用這兩種方法後，MMB自動分類的準確率有明顯的提升。雖然周政淇(2009)提出兩種方法並且提升了MMB分類的準確率。但是在資訊字詞篩選上，雖然解決了MMB分類法沒有考慮字詞出現次數的爭議問題，但是用第四分位數去篩選字詞，還是有可能把重要的關鍵字刪除；再來字詞貢獻差異度處理，雖然不考慮不具特徵性與代表性的字詞，但是不同的類別或不同的文件進行分類，很難決定一個固定的門檻值來進行篩選。

第四章　研究方法與系統架構

本研究主要的系統架構分別為知識建構模組與知識推論模組。其基本的系統架構圖如圖4-1。

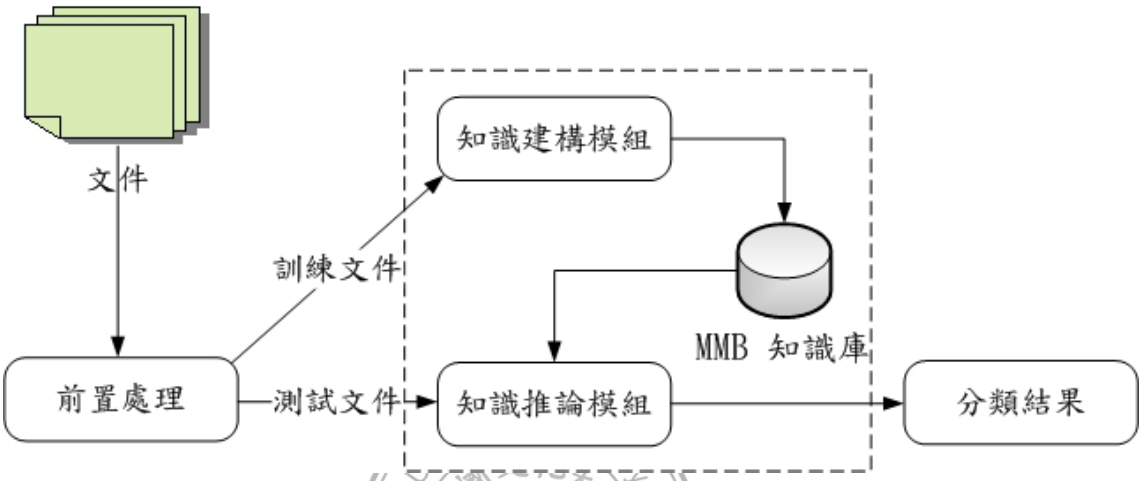

圖 4-1　系統架構圖

依圖4-1所示，文件會先透過「前置處理」後分成訓練與測試文件。虛線部分為文件分類器，是由知識建構模組與知識推論模組。「知識建構模組」利用訓練文件內的名詞建立MMB知識庫，供知識推論時使用；而測試文件則是利用本身的名詞透過「知識推論模組」，依據MMB知識庫中各個類別的知識進行計算，便可以推論出測試文件的分類結果。

一、前置處理

前置處理如圖4-2。

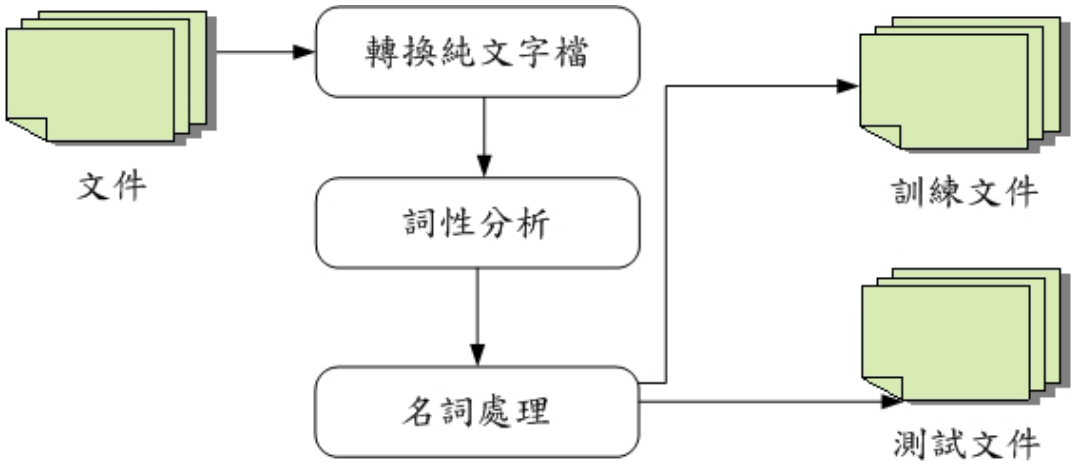

圖 4-2　前置處理流程圖

依據Elsevier Electronic Subscriptions SDOL電子期刊全文資料庫的定義將收集來的文件分門別類後，將文件檔案轉換為純文字檔，並利用詞性分析器進行字詞的詞性分析。分析後進行保留名詞、組合名詞以及將複數名詞還原為單數名詞，依照MMB分類的相關研究發現，保留名詞並且刪除其他詞性的詞彙，可以節省分類資料庫的空間，而且可以達到提升分類資料庫的品質與分類準確率(駱思安，2004)；至於單複數名詞的處理，畢竟兩者的意思相同，例如：apple與apples同樣代表的都是蘋果。最後，將處理完的文件分成訓練文件與測試文件。

二、知識建構模組

知識建構模組是要從訓練文件中，找出屬於各類別裡所有文建的類別特徵，並利用這些特徵建構知識庫。知識建構模組如圖4-3。

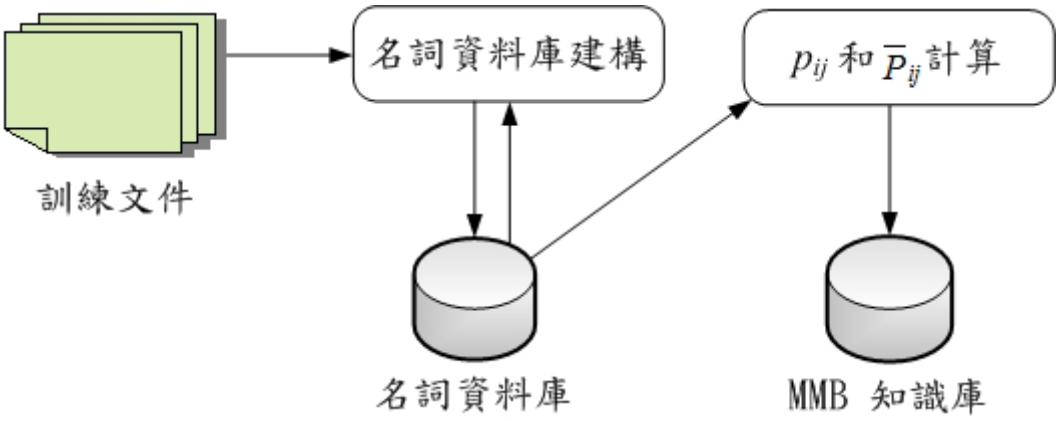

圖 4-3　知識建構模組流程圖

比對測試文件中各個名詞是否存在於名詞資料庫。若該名詞不存在，則將該名詞寫入名詞資料庫。另外，本研究認為只出現在一篇文件的名詞的重要性極低，故全部刪除。

當名詞資料庫完成以上所說的步驟後，讀取所有的名詞資料，並根據MMB機率公式計算出分別是$P(W_j \mid C_i)$(以P_{ij}代表)和$P(W_j \mid \overline{C_i})$(以$\overline{P}_{ij}$代表)。$P_{ij}$代表在$C_i$類別裡的所有訓練文件樣本中名詞$W_j$出現的機率。根據$W_j$在$C_i$類別裡的知識訓練文件樣本當中所出現之次數來標示「$K_1$」($K_1 \geqq 1$)，若$W_j$沒有出現則標示「0」；而$\overline{P}_{ij}$代表在$\overline{C_i}$類別裡的所有知識訓練文件樣本中名詞$W_j$會出現的機率。根據$W_j$在$\overline{C_i}$類別裡的知識訓練文件樣本當中所出現的次數來標示「$K_2$」($K_2 \geqq 1$)，倘若$W_j$並沒有出現則標示為「0」。計算$P_{ij}$值如公式(3-1)和計算$\overline{P}_{ij}$值如公式(3-2)所示。

舉例來說，假設類別C_1有三篇文件$D_{1,1}$、$D_{1,2}$、$D_{1,3}$。而$D_{1,1}$出現名詞W_2、W_3、W_4、W_5、W_7、W_8、W_9；$D_{1,2}$出現名詞W_1、W_2、W_4、W_5、W_6以及$D_{1,3}$出現名詞W_1、W_3、W_4、W_5、W_6、W_8。則如表4-1。

表 4-1　P_{ij} 值計算表

	$D_{1,1}$	$D_{1,2}$	$D_{1,3}$	P_{ij}
W_1	0	1	1	0.67
W_2	1	1	0	0.67
W_3	1	0	1	0.67
W_4	1	1	1	1
W_5	1	1	1	1
W_6	0	1	1	0.67
W_7	1	0	0	0.33
W_8	1	0	1	0.67
W_9	1	0	0	0.33

計算時，並沒有考慮其出現在文件中的次數，而是考慮該名詞是否有出現在該類別的文件中。以表4-1為例的W_1在類別C_1出現兩篇文件，總文件為三篇，所以W_1在類別C_1的P_{ij}值為0.67。

另外，$\overline{P}_{ij}$ 的計算，假設類別C_2有三篇文件$D_{2,1}$、$D_{2,2}$、$D_{2,3}$。而$D_{2,1}$出現名詞W_1、W_2、W_4、W_5、W_6；$D_{2,2}$出現名詞W_1、W_3、W_4、W_5、W_6、W_8；$D_{2,3}$出現名詞W_2、W_3、W_4、W_5、W_7、W_8、W_9。

另一個類別C_3有兩篇文件$D_{3,1}$、$D_{3,2}$。而$D_{3,1}$出現名詞W_1、W_2、W_4、W_5、W_6；$D_{3,2}$出現名詞W_1、W_3、W_4、W_5、W_6、W_8。則$\overline{P}_{ij}$如表4-2。

表 4-2　$\overline{P}_{ij}$ 值計算表

	$D_{2,1}$	$D_{2,2}$	$D_{2,3}$	$D_{3,1}$	$D_{3,2}$	$\overline{P}_{ij}$
W_1	1	1	0	1	1	0.8
W_2	1	0	1	1	0	0.6
W_3	0	1	1	0	1	0.6
W_4	1	1	1	1	1	1
W_5	1	1	1	1	1	1
W_6	1	1	0	1	1	0.8
W_7	0	0	1	0	0	0.2
W_8	0	1	1	0	1	0.6
W_9	0	0	1	0	0	0.2

以上表所示，W_7在非C_1類別出現了一篇文件，而非C_1類別的文件為五篇，所以W_7在類別C_1的$\overline{P}_{ij}$為0.2，依此類推。最後，將與P_{ij}與$\overline{P}_{ij}$值整合可得C_1的MMB推論知識，如表4-3。

表 4-3　類別 C_1 的 MMB 推論知識表

	P_{ij}	$\overline{P}_{ij}$
W_1	0.67	0.8
W_2	0.67	0.6
W_3	0.67	0.6
W_4	1	1
W_5	1	1
W_6	0.67	0.8
W_7	0.33	0.2
W_8	0.67	0.6
W_9	0.33	0.2

每一篇訓練文件經過以上的計算方法，皆可計算出每一個名詞在各類別的P_{ij}與$\overline{P}_{ij}$，而我們便可利用這兩個機率值進行知識推論模組，進而得到測試文件所屬各類別的機率。

三、知識推論模組

知識推論模組為針對測試文件進行分類推論的動作，利用MMB知識庫找出訓練文件中所有名詞的P_{ij}與$\overline{P}_{ij}$值後，將其帶入MMB推論公式推算出所屬各類別的機率，如圖4-4。

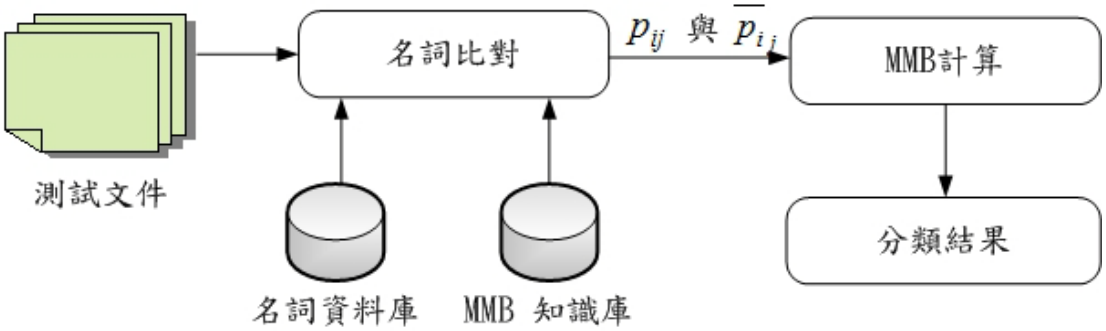

圖 4-4　MMB 公式的推論流程圖

測試文件的內容透過名詞資料庫與MMB知識庫交叉比對後的，將各個名詞與其在各類別的P_{ij}與$\overline{P}_{ij}$值帶入MMB推論公式中，便可得該測試文件的分類結果。MMB公式如公式(3-3)。

C_i為類別、W_j為名詞、$P(W_j|C_i)$(以P_{ij}代表)和$P(W_j|\overline{C_i})$(以$\overline{P}_{ij}$代表)。$P(C_i|\ W_1, W_2, \ldots, W_n)$是該測試文件中所有的名詞$W_1, W_2, \ldots, W_n$時屬於類別$C_i$的後天機率。$P(C_i)$=0.5為該測試文件屬於類別$C_i$的先天機率；$(1-P(C_i))$為該測試文件不屬於類別$C_i$的先天機率。

因此知識推論的流程是利用測試文件中的名詞，先比對名詞資料庫後，在將比對出來的名詞資料與MMB知識庫進行比對，若比對成功則取出此名詞的P_{ij}和$\overline{P}_{ij}$值，最後代入MMB推論式計算後，就可以得到該文件屬於各類別的機率。

(一)自動調適出現次數門檻

基於周政淇(2009)的研究，四分位數門檻值的實驗結果顯示字詞的出現次數會影響字詞的重要性。所以，在此將針對測試文件中名詞出現次數進行自動調整，本研究將出現次數較多的字詞列為關鍵字。至於常用字的部分，因為常用字出現次數多，出現的文章數也多，因此對分類的準確率沒益處也無壞處。每篇測試文件的起始出現次數門檻值為0，表示只對出現次數大於0的名詞進行MMB計算。當結果為無法分類，則該測試文件重新計算且出現次數門檻值加1，依此類推門檻值為0、1、2、...、最大出現次數。若達最大出現次數且無法成功分類時則降低評估門檻，評估門檻範圍為1、0.95、0.9、...、0.7。如圖4-5。

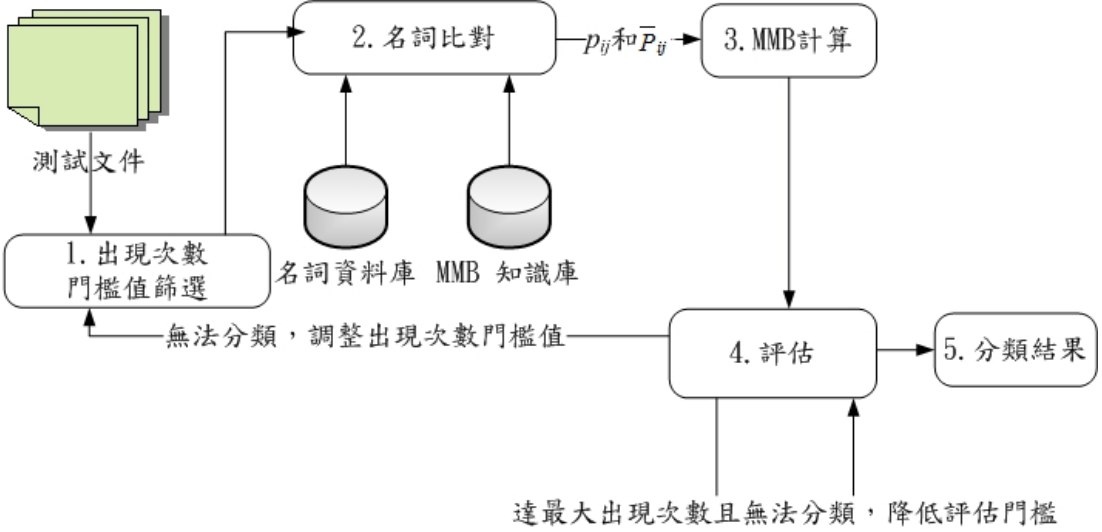

圖 4-5　自動調適出現次數門檻

(二)自動調適字詞貢獻差異度門檻

周政淇(2009)提到當該名詞在某類別中的P_{ij}與$\overline{P}_{ij}$值太過接近時，代表該名詞對該類別的重要性較低。該研究只利用固定門檻值進行分類，而我們則是當無法分類時，則該測試文件重新計算，每一次的重新計算差異度門檻值就會加0.05，依此類推。因為MMB公式中$P(C_i)$=0.5，所以門檻值範圍為0、0.05、0.1、...、0.5。若字詞差異度門檻值達0.5且無法成功分類時則降低評估門檻，評估門檻範圍為1、0.95、0.9、...、0.7。推論流程如圖4-6。

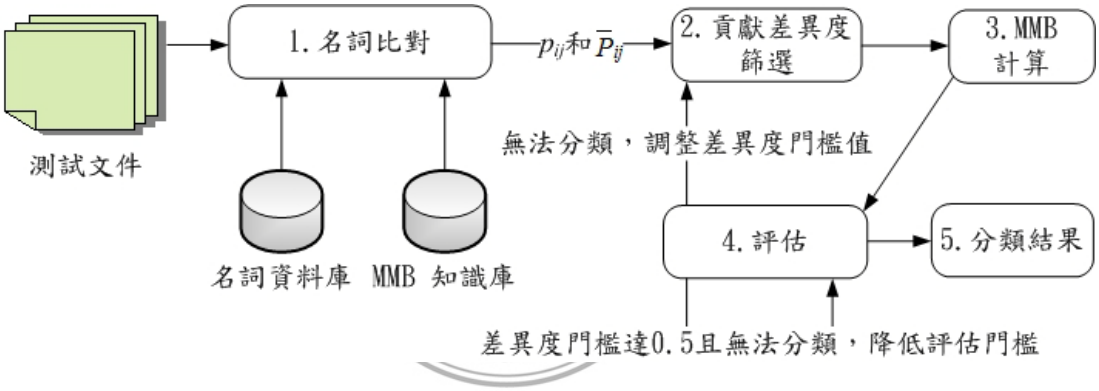

圖 4-6　自動調適字詞貢獻差異度篩選

(三)自動調適出現次數與字詞貢獻差異度門檻

將「自動調適出現次數門檻」與「自動調適字詞貢獻差異度門檻」聯合應用。當第一階段評估為無法分類時，則該測試文件重新計算且出現次數門檻值加1依此類推。當結果為無法分類，則該測試文件重新計算且出現次數門檻值加1，依此類推門檻值為0、1、2、…、最大出現次數。若達最大出現次數且無法成功分類時則降低評估門檻，評估門檻範圍為1、0.95、0.9、...、0.7。

通過第一階段評估後，將刪選過後的字詞之P_{ij}與$\overline{P}_{ij}$值

代入第二階段計算，當無法分類時則該測試文件重新計算且字詞差異度門檻值加0.05，依此類推差異度門檻值為0、0.05、0.1、…、0.5。第二階段評估主要為拉開類別間MMB機率值的差距，差距為0.05≦K≦0.5。如圖4-7。

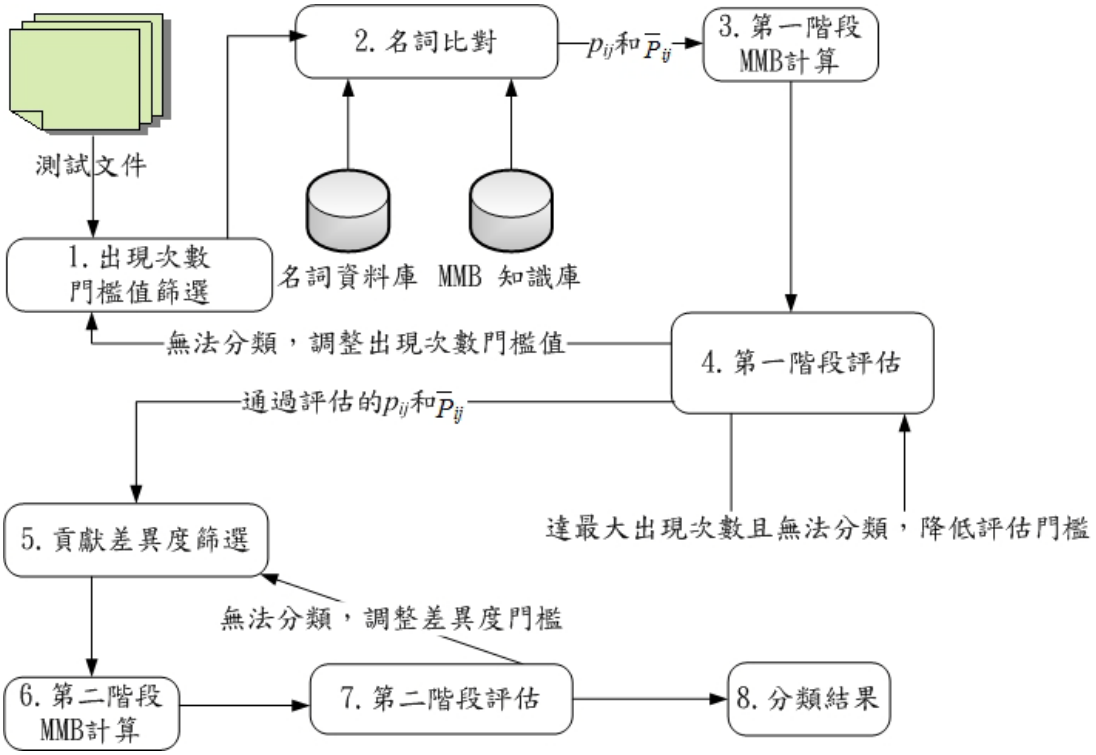

圖 4-7　自動調適出現次數與字詞貢獻差異度門檻

第五章　系統實作與實驗結果

第一節　系統實作

本研究的文件樣本是從Elsevier Electronic Subscription SDOL電子期刊全文資料庫下載總計2266篇文件，並依照該電子期刊所分類的4個大類別與22個小類別將收集來的文件分門別類，其文件的檔案格式皆為PDF格式。相關類別表格如表5-3。

一、前置處理

將所有PDF文件轉換成TXT文字檔後，刪除題目、作者、摘要、註解與參考文獻，只保留本文的部分。因為根據以往周政淇於2009年發表的MMB文件分類的相關研究，得知以本文作為訓練樣本以及測試樣本，可以得到較好的分類結果。接著，將已轉換為TXT檔案格式的文件，透過詞性分析器GENIA tagger，進行字詞的詞性分析。圖5-1為分析後的文件。

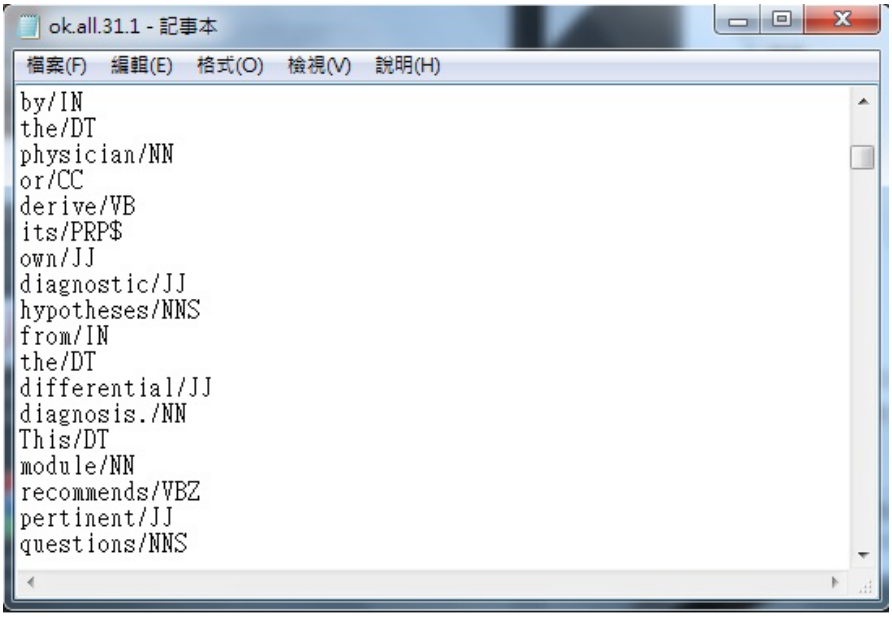

圖 5-1　詞性分析後的文件(字詞/詞性)

再來是名詞處理，也就是處裡詞性為名詞的字詞，如圖5-1，當詞性為「NN」or「NNS」則讀取該字詞。當詞性為「NNS」則表示該字詞為複數名詞，在還原單數的同時將連續的名詞合併為複合名詞，最後在考慮是否可以還原成單數名詞，其還原規則如表5-1。

表 5-1　複數還原單數

不規則型			規則型	
複數	單數		複數	單數
men	man		-ments	-ment
oxen	ox		-ies	-y
children	child		-ses	-s
…	…		-xes	-x
mice	mouse		…	…
indices	index		-oes	-o

完成複數名詞還原後，依照訓練文件量:測試文件量將文件分成7:3與8:2，並依照比例建立不同的知識庫，來探討訓練量與測試量之間的關係。文件量分布如表5-2與表5-3。

表 5-2　文件比例 7:3 之樣本分布表

大類別	小類別	訓練文件(篇)	測試文件(篇)
Physical Sciences and Engineering	Chemical Engineering	69	29
	Chemistry	61	26
	Computer Science	83	35
	Earth and Planetary Sciences	54	23
	Energy	67	28
	Engineering	57	25
	Materials Science	102	43
	Mathematics	70	30
Life Sciences	Agricultural and Biological Sciences	83	35
	Biochemistry, Genetics and Molecular Biology	70	30
	Environmental Science	85	37
	Neuroscience	71	30
Health Sciences	Medicine and Dentistry	73	31
	Nursing and Health Professions	71	31
	Pharmacology, Toxicology and Pharmaceutical Science	71	30
	Veterinary Science and Veterinary Medicine	70	30
Social Sciences and Humanities	Arts and Humanities	71	31
	Business, Management and Accounting	70	30
	Decision Sciences	71	30
	Economics, Econometrics and Finance	71	30
	Psychology	71	30
	Social Sciences	71	30
總計		1582	674

表 5-3　文件比例 8:2 之樣本分布表

大類別	小類別	訓練文件(篇)	測試文件(篇)
Physical Sciences and Engineering	Chemical Engineering	78	20
	Chemistry	70	17
	Computer Science	94	24
	Earth and Planetary Sciences	62	15
	Energy	76	19
	Engineering	66	16
	Materials Science	116	29
	Mathematics	80	20
Life Sciences	Agricultural and Biological Sciences	94	24
	Biochemistry, Genetics and Molecular Biology	80	20
	Environmental Science	98	24
	Neuroscience	81	20
Health Sciences	Medicine and Dentistry	83	21
	Nursing and Health Professions	82	20
	Pharmacology, Toxicology and Pharmaceutical Science	81	20
	Veterinary Science and Veterinary Medicine	80	20
Social Sciences and Humanities	Arts and Humanities	82	20
	Business, Management and Accounting	80	20
	Decision Sciences	81	20
	Economics, Econometrics and Finance	81	20
	Psychology	81	20
	Social Sciences	81	20
總計		1807	449

二、名詞資料庫建構

比對測試文件中各個名詞是否存在於名詞資料庫。若該名詞不存在，則將該名詞寫入名詞資料庫。以文件比例8:2的資料庫為例，如表5-4。

表 5-4　名詞資料庫

名詞 ID	名詞
1	addition
2	algorithm
…	…
17283	chart diagram
…	…
28939	baby food
…	…
43267	workplace characteristics

最後，建立「文件與名詞之關係」。記錄訓練文件名稱、名詞、出現次數。如表5-5。

表 5-5　文件與名詞之關係

文件	名詞	出現次數
b.a.22	chart diagram	2
…	…	…
b.b.38	addition	6
…	…	…
c.d.14	baby food	2
…	…	…
d.f.81	workplace characteristics	2
…	…	…
a.c.18	algorithm	50

上表5-5為例，在訓練文件「b.a.22」中，名詞「chart

diagram」的名詞總共出現了「2」次。其他資料依此類推。

三、P_{ij}與$\overline{P}_{ij}$值計算

以文件比例8:2的資料庫為例，在機率計算處理中，將計算出所有名詞的P_{ij}與$\overline{P}_{ij}$值。首先要找出所有訓練文件中名詞在每一個類別所出現的次數，如表5-6。

表 5-6　類別之字詞出現文章數

名詞	隸屬小類別	出現 文章數
addition	Chemical Engineering	25
addition	Chemistry	18
…	…	…
addition	Social Sciences	49
algorithm	Chemical Engineering	12
algorithm	Chemistry	5
…	…	…
algorithm	Agricultural and Biological Sciences	11
…	…	…
chart diagram	Chemical Engineering	0
…	…	…
workplace characteristics	Social Sciences	1

以表5-6為例，名詞「algorithm」在類別Agricultural and Biological Sciences出現11篇，而該類別內總共有94篇訓練文件(參考表5-3)，最後用公式(3-1)計算P_{ij}。分母為94篇，分子不為0的數量為11篇。故名詞ID 2在類別Agricultural and Biological Sciences的P_{ij}值為0.11702 (取自小數點後第5位數)。

接著$\overline{P}_{ij}$的計算，名詞「algorithm」在類別Agricultural and

Biological Sciences以外出現172篇，而　該類別以外的訓練文件數量總共有1713篇。利用公式(3-2)分母為1713篇，分子不為0的數量為172篇。故名詞「algorithm」在小類別Agricultural and Biological Sciences的$\overline{P}_{ij}$值為0.10041 (取自小數點後第5位數)。持續以上面的方式計算全部的字詞，可得到P_{ij}與$\overline{P}_{ij}$值之資料，如表5-7。

表 5-7　P_{ij}與$\overline{P}_{ij}$值計算結果

名詞 ID	隸屬小類別	P_{ij}	$\overline{P}_{ij}$
addition	Chemical Engineering	0.32051	0.40023
addition	Chemistry	0.25714	0.40242
…	…	…	…
addition	Social Sciences	0.60494	0.38702
algorithm	Chemical Engineering	0.15385	0.09890
algorithm	Chemistry	0.07143	0.10248
…	…	…	…
algorithm	Agricultural and Biological Sciences	0.11702	0.10041
…	…	…	…
chart diagram	Chemical Engineering	0.00000	0.00058
…	…	…	…
workplace characteristics	Social Sciences	0.01235	0.00000

四、知識推論程式

MMB知識庫件利完成後，依據不同的需求撰寫不同功能的推論程式。本研究依照文件比例7:3與8:2，以及大小類別，分別建立了「自動調適出現次數門檻」、「自動調適字詞貢獻差異度門檻」和「自動調適出現次數與字詞貢獻差異度門檻」三種功能的程式。

另外，為了找出更好的改良方式，本研究也依照周政淇(2009)所提出的兩中改良方法，建立了「四分位數門檻」與「字詞貢獻差異度門檻」兩種功能的程式。程式撰寫畫面如圖5-2。

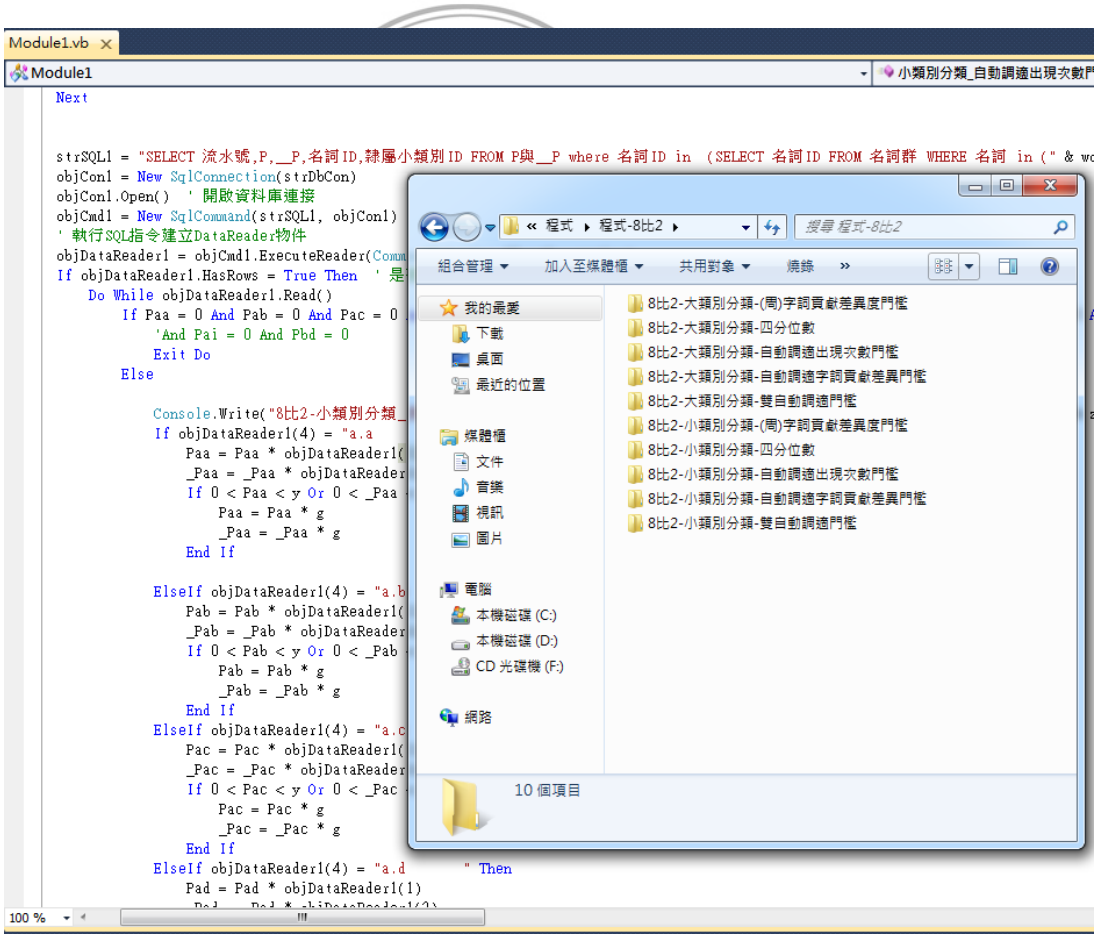

圖 5-2　程式撰寫畫面之示意圖

推論程式的執行畫面如圖5-3。

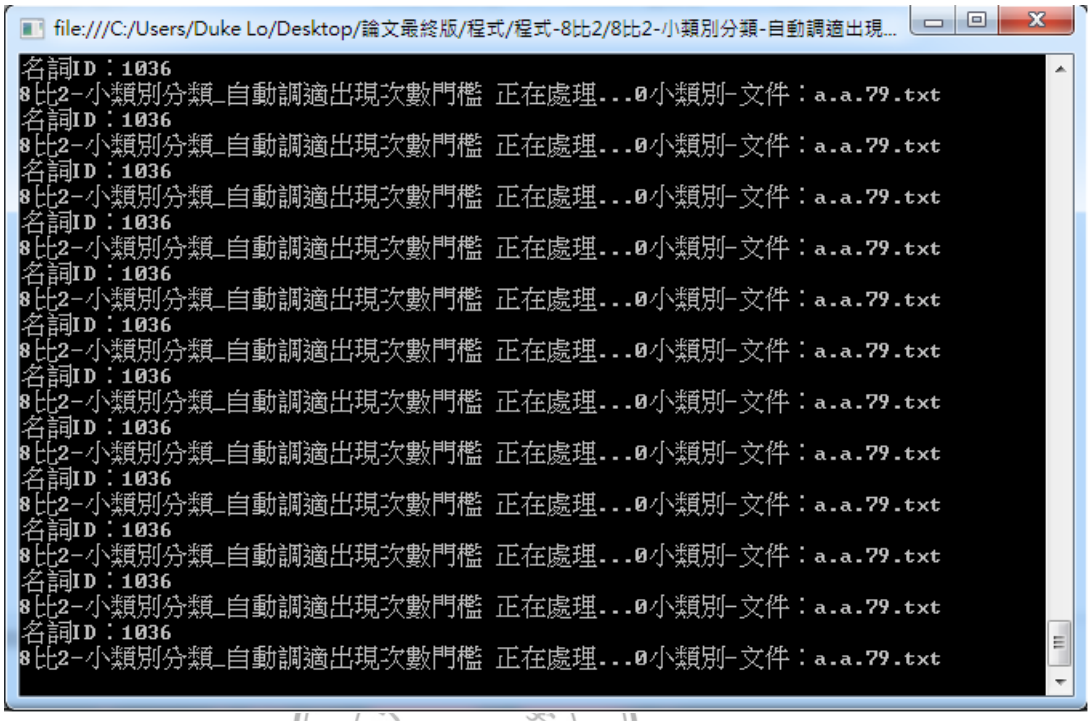

圖 5-3　程式執行畫面之示意圖

圖5-3主要是在處理文件比例8:2於小類別中執行自動調適出現次數門檻。

程式執行結果之畫面如圖5-4。

圖 5-4 程式執行結果之示意圖

所有程式執行後，皆以純文字檔來產生分類的結果。圖5-4為文件比例8:2於小類別中，自動調適出現次數的分類結果。

另外，關於周政淇(2009)提出的字詞貢獻差異度，本研究實作發現，分類結果幾乎是五個類別以上皆成立。如圖5-5。

圖 5-5　字詞貢獻差異度執行結果之示意圖

如圖5-5，本研究在字詞貢獻差異度的實驗上僅列出前十名的機率值，如同上一段所說成立的類別很多。不過，本研究還是比照其他方法的審核方式，與機率值最高的類別相差在0.05以內，則將其列為分類正確。

第二節　實驗結果

本研究分別利用資料量比例7:3，大約為訓練文件1582篇與測試文件674篇；資料量比例8:2，大約為訓練文件1807篇與測試文件459篇，來進行MMB文件自動分類之實驗，並討論「大類別與小類別」各自的分類狀況，以及與周政淇(2009)所提出的兩種改良方法之比較。訓練文件的分布狀況如表5-2與表5-3。

首先，根據實驗的結果得知單純使用MMB公式進行推論，不管是大類別或小類別的分類中，分類結果的正確率幾乎為0.00%。所以可得知沒有經過改良以及雜訊過多，會導致準確率下降，甚至無法分類。因為MMB公式的特性就是當文章中某個字詞的P_{ij}為0時，則屬於該類別機率就等於0。以下分別以大小類別呈現實驗結果。

一、大類別

(一)大類別分類之自動調適出現次數門檻

大類別分類之自動調適出現次數門檻，如表5-8。

表 5-8　大類別分類之自動調適出現次數門檻

訓練文件:測試文件	7:3				8:2			
大類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Physical Sciences and Engineering	563	239	217	90.79%	642	160	144	90.00%
Life Sciences	309	132	102	77.27%	353	88	66	75.00%
Health Sciences	285	122	94	77.05%	326	91	60	65.93%
Social Sciences and Humanities	425	181	164	90.61%	486	120	109	90.83%
平均正確率				83.93%				80.44%

針對字詞的出現次數以自動調適的方式篩選，可以過濾一些不重要的字詞，使MMB的分類正確率大幅提升。

(二)大類別分類之自動調適字詞貢獻差異度門檻

大類別分類之自動調適字詞貢獻差異度門檻，如表5-9。

表 5-9　大類別分類之自動調適字詞貢獻差異度門檻

訓練文件:測試文件	7:3				8:2			
大類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Physical Sciences and Engineering	563	239	191	79.92%	642	160	130	81.25%
Life Sciences	309	132	102	77.27%	353	88	60	68.18%
Health Sciences	285	122	70	57.38%	326	91	50	54.95%
Social Sciences and Humanities	425	181	176	97.24%	486	120	116	96.67%
平均正確率				77.95%				75.26%

(三)大類別分類之自動調適出現次數與字詞貢獻差異度門檻

大類別分類之自動調適出現次數與字詞貢獻差異度門檻，如表5-10。

表 5-10　大類別分類之自動調適出現次數與字詞貢獻差異度門檻

訓練文件:測試文件	7:3				8:2			
大類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Physical Sciences and Engineering	563	239	207	86.61%	642	160	138	86.25%
Life Sciences	309	132	77	58.33%	353	88	51	57.95%
Health Sciences	285	122	79	64.75%	326	91	55	60.44%
Social Sciences and Humanities	425	181	155	85.64%	486	120	103	85.83%
平均正確率				73.83%				72.62%

由以上研究可得知，兩種方法聯合使用效果較第一種「自動調整出現次數門檻」差。因為通過第一階段「自動調適」已保留能產生分類結果的字詞，而第二階段「自動調適字詞貢獻差異度門檻」，因為前階段所保留的字詞皆可能再執行調整字詞貢獻差異門檻時，將原本重要的字詞被判斷為不重要的字詞，因此導致分類的正確率降低。

(四)大類別分類之四分位數門檻

大類別分類之四分位數門檻，如表5-11。

表 5-11　大類別分類之第三四分位數門檻

訓練文件:測試文件	7:3				8:2			
大類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Physical Sciences and Engineering	563	239	89	37.24%	642	160	65	40.63%
Life Sciences	309	132	19	14.39%	353	88	13	14.77%
Health Sciences	285	122	36	29.51%	326	91	23	25.27%
Social Sciences and Humanities	425	181	26	14.36%	486	120	21	17.50%
平均正確率				23.88%				24.54%

大類別分類之四分位數門檻的分類結果如圖5-6。

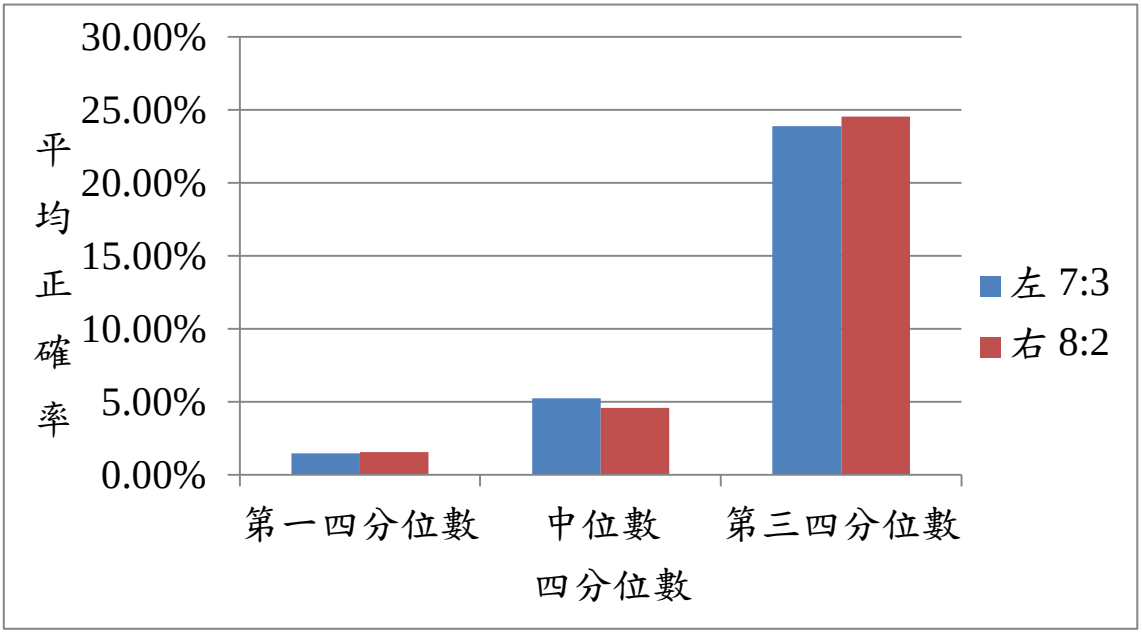

圖 5-6　大類別分類之四分位數門檻

以周政淇(2009)所提出的用四分位數門檻篩選的MMB分類法進行的實驗。

(五)大類別分類之字詞貢獻差異度門檻

大類別分類之字詞貢獻差異度門檻，如表5-12。

表 5-12　大類別分類之字詞貢獻差異度門檻

訓練文件:測試文件	7:3				8:2			
字詞貢獻差異度門檻值	0.05				0.10			
大類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Physical Sciences and Engineering	563	239	199	83.26%	642	160	125	78.13%
Life Sciences	309	132	104	78.79%	353	88	63	71.59%
Health Sciences	285	122	71	58.20%	326	91	45	49.45%
Social Sciences and Humanities	425	181	172	95.03%	486	120	120	100.00%
平均正確率				78.82%				74.79%

大類別分類之字詞貢獻差異度門檻的分類結果如圖5-7。

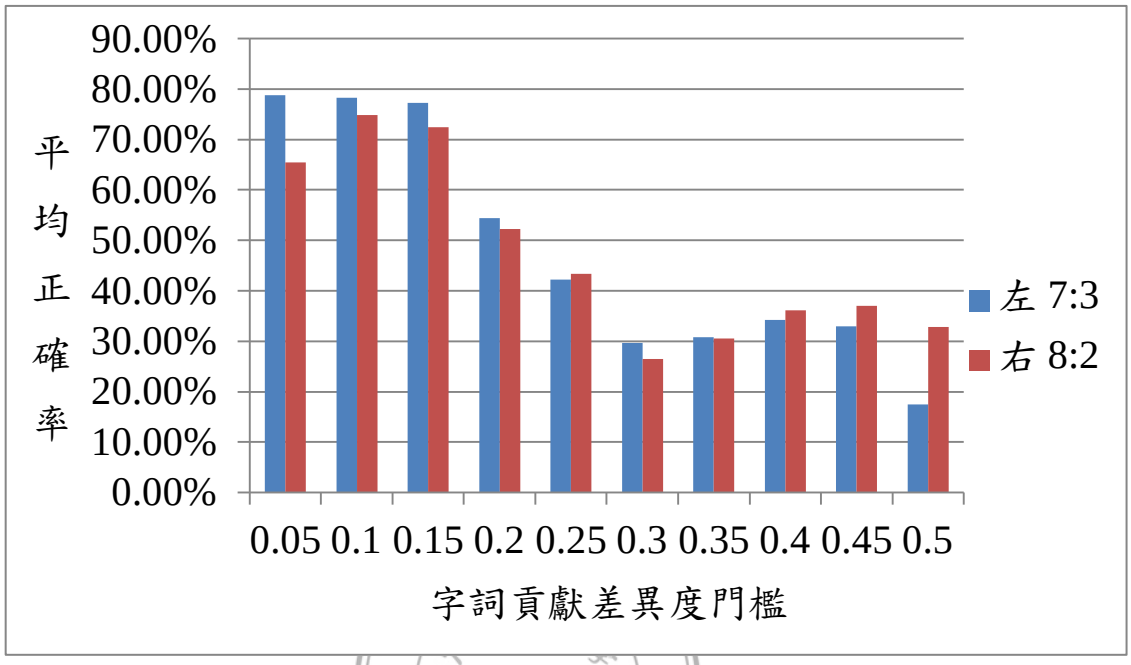

圖 5-7 大類別分類之字詞貢獻差異度門檻

以上的實驗是以周政淇(2009)所提出的固定字詞貢獻差異度門檻MMB分類法來進行實驗。

接著探討大類別跨領域的分類狀況，如表5-13。

表 5-13　大類別分類之跨類別分析

研究類別	訓練文件:測試文件	7:3			8:3		
	項目(大類別)	正確分類	跨類別	比率	正確分類	跨類別	比率
本研究	自動調適出現次數門檻	577	7	1.21%	459	379	0.53%
	自動調適字詞貢獻差異度門檻	539	186	34.51%	356	119	33.43%
周政淇之方法	字詞貢獻差異度門檻值(0.05 | 0.10)	546	114	20.88%	353	218	61.76%

表5-13僅列出列出分類正確率較好的方法，可以清楚的看出以字詞貢獻差異度為主軸的方法，有蠻多跨領域的情況。當超過4個類別以上的分類結果，則此訓練文件的跨類別成立，也就是分類的鑑別度較低，而又以周政淇提出的方法出現的比率更高。

二、小類別

(一)小類別分類之自動調適出現次數門檻

小類別分類之自動調適出現次數門檻，如表5-14。

表 5-14　小類別分類之自動調適出現次數門檻

訓練文件:測試文件	7:3				8:2			
小類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Chemical Engineering	69	29	18	62.07%	78	20	11	55.00%
Chemistry	61	26	17	65.38%	70	17	11	64.71%
Computer Science	83	35	25	71.43%	94	24	18	75.00%
Earth and Planetary Sciences	54	23	19	82.61%	62	15	14	93.33%
Energy	67	28	19	67.86%	76	19	11	57.89%
Engineering	57	25	20	80.00%	66	16	13	81.25%
Materials Science	102	43	35	81.40%	116	29	22	75.86%
Mathematics	70	30	18	60.00%	80	20	12	60.00%
Agricultural and Biological Sciences	83	35	28	80.00%	94	24	23	95.83%
Biochemistry, Genetics and Molecular Biology	70	30	18	60.00%	80	20	10	50.00%
Environmental Science	85	37	29	78.38%	98	24	19	79.17%
Neuroscience	71	30	22	73.33%	81	20	13	65.00%
Medicine and Dentistry	73	31	12	38.71%	83	21	7	33.33%
Nursing and Health Professions	71	31	19	61.29%	82	20	16	80.00%
Pharmacology, Toxicology and Pharmaceutical Science	71	30	18	60.00%	81	20	9	45.00%
Veterinary Science and Veterinary Medicine	70	30	22	73.33%	80	20	13	65.00%
Arts and Humanities	71	31	25	80.65%	82	20	18	90.00%
Business, Management and Accounting	70	30	27	90.00%	80	20	18	90.00%
Decision Sciences	71	30	19	63.33%	81	20	11	55.00%
Economics, Econometrics and Finance	71	30	22	73.33%	81	20	14	70.00%
Psychology	71	30	23	76.67%	81	20	14	70.00%
Social Sciences	71	30	22	73.33%	81	20	19	95.00%
平均正確率				70.60%				70.29%

(二)小類別分類之自動調適字詞貢獻差異度門檻

小類別分類之自動調適字詞貢獻差異度門檻，如表5-15。

表 5-15　小類別分類之自動調適字詞貢獻差異度門檻

訓練文件:測試文件	7:3				8:2			
小類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Chemical Engineering	69	29	17	58.62%	78	20	10	50.00%
Chemistry	61	26	10	38.46%	70	17	6	35.29%
Computer Science	83	35	18	51.43%	94	24	14	58.33%
Earth and Planetary Sciences	54	23	4	17.39%	62	15	5	33.33%
Energy	67	28	13	46.43%	76	19	8	42.11%
Engineering	57	25	5	20.00%	66	16	6	37.50%
Materials Science	102	43	14	32.56%	116	29	12	41.38%
Mathematics	70	30	7	23.33%	80	20	6	30.00%
Agricultural and Biological Sciences	83	35	29	82.86%	94	24	19	79.17%
Biochemistry, Genetics and Molecular Biology	70	30	3	10.00%	80	20	2	10.00%
Environmental Science	85	37	25	67.57%	98	24	17	70.83%
Neuroscience	71	30	13	43.33%	81	20	10	50.00%
Medicine and Dentistry	73	31	2	6.45%	83	21	1	4.76%
Nursing and Health Professions	71	31	23	74.19%	82	20	17	85.00%
Pharmacology, Toxicology and Pharmaceutical Science	71	30	10	33.33%	81	20	5	25.00%
Veterinary Science and Veterinary Medicine	70	30	4	13.33%	80	20	1	5.00%
Arts and Humanities	71	31	23	74.19%	82	20	17	85.00%
Business, Management and Accounting	70	30	21	70.00%	80	20	13	65.00%
Decision Sciences	71	30	19	63.33%	81	20	11	55.00%
Economics, Econometrics and Finance	71	30	21	70.00%	81	20	12	60.00%
Psychology	71	30	21	70.00%	81	20	16	80.00%
Social Sciences	71	30	27	90.00%	81	20	19	95.00%
平均正確率				48.04%				49.90%

(三)小類別分類之自動調適出現次數與字詞貢獻差異度門檻

小類別分類之自動調適出現次數與字詞貢獻差異度門檻，如表5-16。

表 5-16　小類別分類之自動調適出現次數與字詞貢獻差異度門檻

訓練文件:測試文件	7:3				8:2			
小類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Chemical Engineering	69	29	11	37.93%	78	20	6	30.00%
Chemistry	61	26	4	15.38%	70	17	2	11.76%
Computer Science	83	35	21	60.00%	94	24	15	62.50%
Earth and Planetary Sciences	54	23	14	60.87%	62	15	10	66.67%
Energy	67	28	19	67.86%	76	19	11	57.89%
Engineering	57	25	14	56.00%	66	16	10	62.50%
Materials Science	102	43	26	60.47%	116	29	17	58.62%
Mathematics	70	30	7	23.33%	80	20	5	25.00%
Agricultural and Biological Sciences	83	35	23	65.71%	94	24	19	79.17%
Biochemistry, Genetics and Molecular Biology	70	30	10	33.33%	80	20	6	30.00%
Environmental Science	85	37	16	43.24%	98	24	11	45.83%
Neuroscience	71	30	15	50.00%	81	20	10	50.00%
Medicine and Dentistry	73	31	1	3.23%	83	21	1	4.76%
Nursing and Health Professions	71	31	8	25.81%	82	20	9	45.00%
Pharmacology, Toxicology and Pharmaceutical Science	71	30	7	23.33%	81	20	6	30.00%
Veterinary Science and Veterinary Medicine	70	30	10	33.33%	80	20	7	35.00%
Arts and Humanities	71	31	20	64.52%	82	20	14	70.00%
Business, Management and Accounting	70	30	16	53.33%	80	20	11	55.00%
Decision Sciences	71	30	11	36.67%	81	20	4	20.00%
Economics, Econometrics and Finance	71	30	20	66.67%	81	20	12	60.00%
Psychology	71	30	17	56.67%	81	20	7	35.00%
Social Sciences	71	30	18	60.00%	81	20	16	80.00%
平均正確率				45.35%				46.12%

(四)小類別分類之四分位數門檻值

小類別分類之四分位數門檻值，如表5-17。

表 5-17　小類別分類之第三四分位數門檻值

訓練文件:測試文件	7:3				8:2			
小類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Chemical Engineering	69	29	1	3.45%	78	20	1	5.00%
Chemistry	61	26	0	0.00%	70	17	0	0.00%
Computer Science	83	35	2	5.71%	94	24	2	8.33%
Earth and Planetary Sciences	54	23	2	8.70%	62	15	2	13.33%
Energy	67	28	2	7.14%	76	19	2	10.53%
Engineering	57	25	0	0.00%	66	16	0	0.00%
Materials Science	102	43	1	2.33%	116	29	3	10.34%
Mathematics	70	30	4	13.33%	80	20	4	20.00%
Agricultural and Biological Sciences	83	35	0	0.00%	94	24	1	4.17%
Biochemistry, Genetics and Molecular Biology	70	30	4	13.33%	80	20	4	20.00%
Environmental Science	85	37	0	0.00%	98	24	1	4.17%
Neuroscience	71	30	0	0.00%	81	20	0	0.00%
Medicine and Dentistry	73	31	2	6.45%	83	21	2	9.52%
Nursing and Health Professions	71	31	0	0.00%	82	20	0	0.00%
Pharmacology, Toxicology and Pharmaceutical Science	71	30	2	6.67%	81	20	2	10.00%
Veterinary Science and Veterinary Medicine	70	30	5	16.67%	80	20	5	25.00%
Arts and Humanities	71	31	0	0.00%	82	20	0	0.00%
Business, Management and Accounting	70	30	0	0.00%	80	20	0	0.00%
Decision Sciences	71	30	0	0.00%	81	20	0	0.00%
Economics, Econometrics and Finance	71	30	0	0.00%	81	20	0	0.00%
Psychology	71	30	0	0.00%	81	20	0	0.00%
Social Sciences	71	30	0	0.00%	81	20	0	0.00%
平均正確率				3.81%				6.38%

小類別分類之四分位數門檻的分類結果如圖5-8。

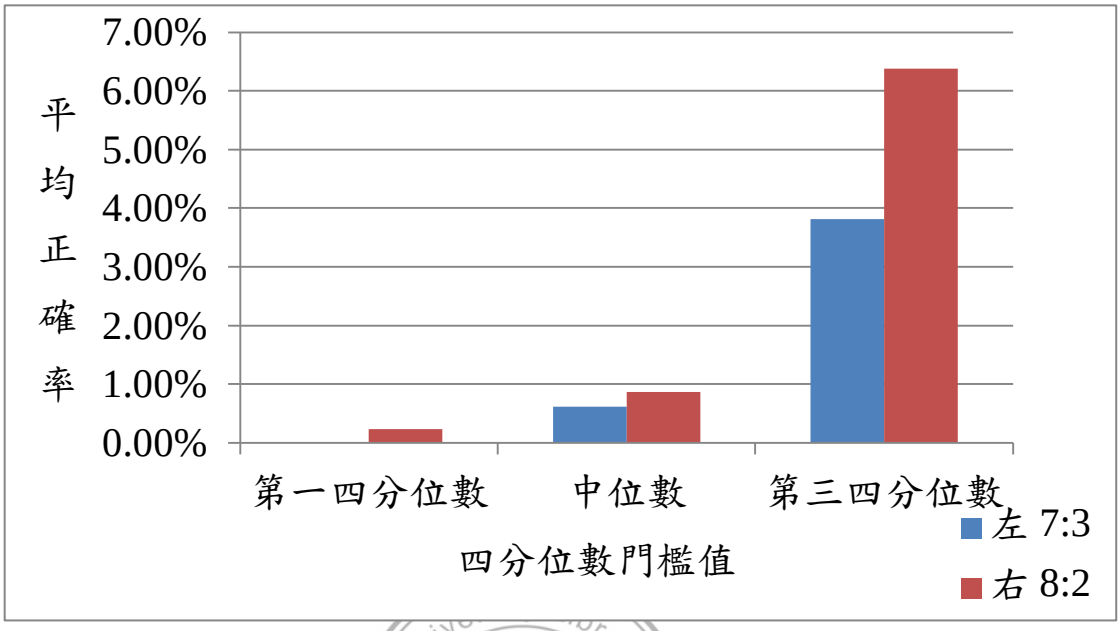

圖 5-8　小類別分類之四分位數門檻值

以上的實驗是以周政淇(2009)所提出的用四分位數門檻篩選的MMB分類法進行的實驗。

(五)小類別分類之字詞貢獻差異度門檻

小類別分類之字詞貢獻差異度門檻，如表5-18。

表 5-18　小類別分類之字詞貢獻差異度門檻

訓練文件:測試文件	7:3				8:2			
字詞貢獻差異度門檻值					0.30			
小類別	訓練文章	測試文章	正確分類	正確率	訓練文章	測試文章	正確分類	正確率
Chemical Engineering	69	29	26	89.66%	78	20	20	100.00%
Chemistry	61	26	23	88.46%	70	17	17	100.00%
Computer Science	83	35	35	100.00%	94	24	24	100.00%
Earth and Planetary Sciences	54	23	22	95.65%	62	15	15	100.00%
Energy	67	28	28	100.00%	76	19	19	100.00%
Engineering	57	25	25	100.00%	66	16	16	100.00%
Materials Science	102	43	39	90.70%	116	29	27	93.10%
Mathematics	70	30	27	90.00%	80	20	15	75.00%
Agricultural and Biological Sciences	83	35	35	100.00%	94	24	24	100.00%
Biochemistry, Genetics and Molecular Biology	70	30	7	23.33%	80	20	3	15.00%
Environmental Science	85	37	33	89.19%	98	24	22	91.67%
Neuroscience	71	30	30	100.00%	81	20	19	95.00%
Medicine and Dentistry	73	31	21	67.74%	83	21	11	52.38%
Nursing and Health Professions	71	31	30	96.77%	82	20	20	100.00%
Pharmacology, Toxicology and Pharmaceutical Science	71	30	29	96.67%	81	20	20	100.00%
Veterinary Science and Veterinary Medicine	70	30	26	86.67%	80	20	16	80.00%
Arts and Humanities	71	31	31	100.00%	82	20	20	100.00%
Business, Management and Accounting	70	30	30	100.00%	80	20	19	95.00%
Decision Sciences	71	30	30	100.00%	81	20	20	100.00%
Economics, Econometrics and Finance	71	30	30	100.00%	81	20	20	100.00%
Psychology	71	30	30	100.00%	81	20	20	100.00%
Social Sciences	71	30	30	100.00%	81	20	20	100.00%
平均正確率				91.58%				90.78%

小類別分類之字詞貢獻差異度門檻值的全部門檻的測試結果，如圖5-9。

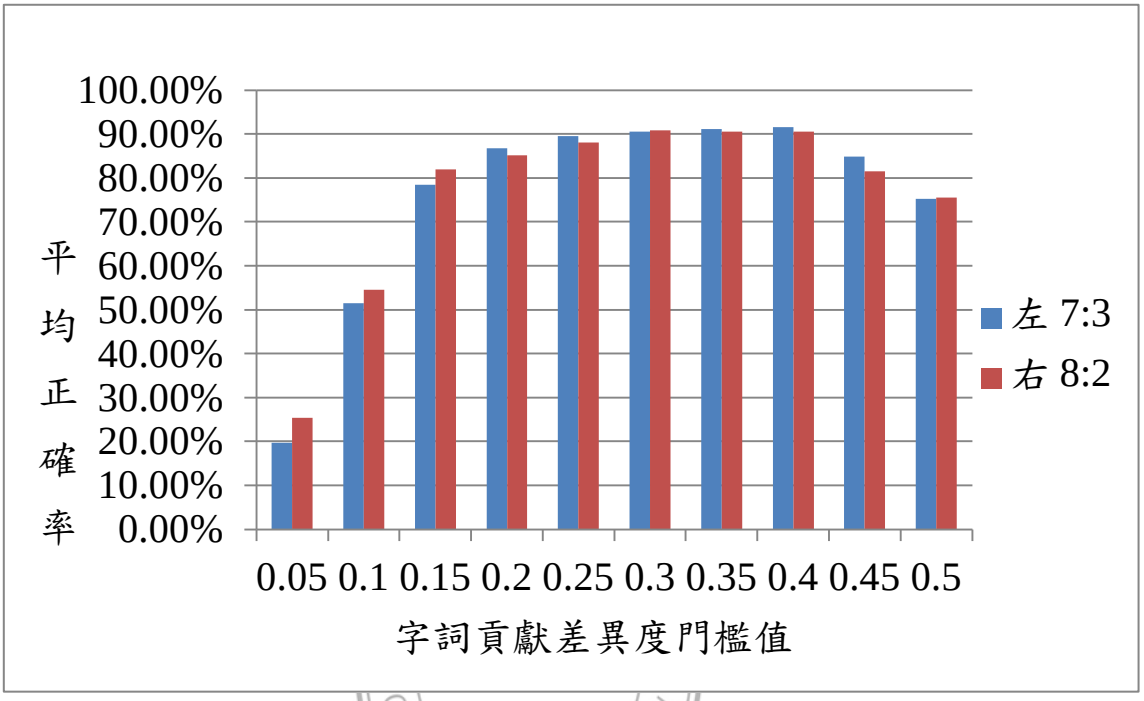

圖 5-9　小類別分類之字詞貢獻差異度門檻

以上同樣是以周政淇(2009)所提出的固定字詞貢獻差異度門檻MMB分類法進行實驗。

以同樣的審核方式，當超過4個類別以上的分類結果，則此訓練文件的跨類別成立。表5-19的狀況如同先前大類別分類之跨類別分析，而又以小類別分類的最為嚴重，幾乎每篇測試文件都有跨領域的現象，更能說明字詞貢獻差異度會使的分類的鑑別度下降。

表 5-19　小類別分類之跨類別分析

研究類別	訓練文件:測試文件	7:3			8:3		
	項目(小類別)	正確分類	跨類別	比率	正確分類	跨類別	比率
本研究	自動調適出現次數門檻	477	58	12.16%	316	29	9.18%
	自動調適字詞貢獻差異度門檻	281	251	89.32%	192	173	90.10%
周政淇之方法	字詞貢獻差異度門檻值 (0.40 | 0.10)	617	617	100.00%	406	406	100%

第六章　結論、貢獻與未來建議

第一節　結論與貢獻

本研究提出的三種改良方式為「自動調適出現次數門檻」、「自動調適字詞貢獻差異度門檻」和「自動調適出現次數門檻與字詞貢獻差異度門檻」。以下為較佳的實驗結果比較表：

一、大類別綜合比較

大類別綜合比較如表6-1。

表 6-1　大類別綜合比較

訓練文件:測試文件		7:3	8:2
項目(大類別)		平均正確率	平均正確率
MMB		0.00%	0.00%
本研究	自動調適出現次數門檻	83.93%	80.44%
	自動調適字詞貢獻差異度門檻	77.95%	75.26%
	自動調適出現次數門檻與字詞貢獻差異度門檻	73.83%	72.62%
周政淇之研究	四分位數門檻之第三四分位數	23.88%	24.54%
	字詞貢獻差異度門檻值 (0.05 | 0.10)	78.82%	74.79%

首先，可以清楚的看出單純使用MMB與加入改良後的方法有明顯的差異，也就是說單純使用MMB已經無法針對現今數位化檔案進行準確的分類。

接著是以字詞的出現次數來篩選的實驗，本研究的「自動調適出現次數門檻」在此表現較周政淇(2009)所提出的「四分位數門檻」優異，可能為自動調適出現次數門檻是依照出現次數漸進式調整，屬於小幅度調整；而四分位數門檻之所以表現較差之原因，原因為四分位數門檻調整幅度大，進而將有用的字詞刪除。

另外，在大類別分類中，本研究與周政淇(2009)提出的「字詞貢獻差異度門檻」的分類正確率無太大的差別，最高的正確率都在75.00%左右。

最後，在大類別分類中，以本研究的「自動調適出現次數門檻」為佳；而資料量比例則是7:3較佳，原因大類別彼此差異性小且識別度高，所以測試文件的多寡將直接影響正確率。

二、小類別綜合比較

小類別綜合比較如表6-2。

表 6-2　小類別綜合比較

訓練文件:測試文件		7:3	8:2
項目(小類別)		平均正確率	平均正確率
僅使用 MMB		0.00%	0.00%
本研究	自動調適出現次數門檻	70.60%	70.29%
	自動調適字詞貢獻差異度門檻	48.04%	49.90%
	自動調適出現次數門檻與字詞貢獻差異度門檻	45.35%	46.12%
周政淇之研究	四分位數門檻之第三四分位數	3.81%	6.38%
	字詞貢獻差異度門檻值 (0.40 | 0.40)	---	---

小類別分類的準確率之所以較大類別分類時低，原因為大類別彼此關係差異性大，比較能分辨出類別；而小類別則是相反，同篇文章跨不同類別的可能性變高，所以在篩選時也可能刪除原本隸屬類別的特徵而保留其他類別的特徵。

與周政淇(2009)的實驗結果相比較，證實不同的訓練與測試文件樣本來源，會影響MMB分類的正確率。而套用到本研究的知識庫，在小類別分類中，兩種資料量比例皆在字詞貢獻

差異度門檻值為0.40時，正確率為91.00%左右；相較於本研究則只有50.00%左右。在第五章末提到「字詞貢獻差異度門檻」所有的分類結果幾乎都跟圖5-5的示意圖一樣，證明固定的門檻值並不能有確切的識別度；而本研究的分類實驗，目的在把類別機率值的差距拉開，藉此提升結果的識別度。

而周政淇的研究僅以資訊類別的英文文件進行研究，加上有資訊類的專業字典作為輔助，只考慮專業字典中有出現的字詞，相反的雜訊(不重要)的字詞就可以輕鬆的排除。

而本研究進行分類的類別是4個各自獨立的大類別，以及底下細分出的22個小類別。因為範圍之廣而沒有專業字典輔助，只要是正確的名詞，就把其列為重要特徵，間接導致雜訊(非重要、常見名詞)過多，進而降低分類的正確性。所以，從以上的結果數據，可以得知獨立性高的類別與相對應的專業字典可以使MMB分類的正確性提升是肯定的。

本研究的實驗結果在小類別分類時，以「自動調適出現次數門檻」表現最好，準確率達70.60%。而大類別分類時也是以「自動調適出現次數門檻」表現最好，準確率皆達83.93%。證明當分類的類別數量少，而且各自的獨立性強，較能正確判斷測試文件隸屬的類別。

第二節　未來建議

從這些研究的結果，將提出一些可以有效提升MMB文件分類準確率的建議：

一、各自獨立性高的類別

獨立性高代表某些字詞在某類別的機率值較高，而在其他類別則較低。這樣就能在MMB分類計算上，更能正確分辨出應該隸屬的類別，而不是多個類別的機率值相近或者同時成立。類別的鑑別度高，想必能讓MMB的分類準確率提升不少。

二、關鍵字的提取或將其與MMB公式結合，產生新的推論公式

本研究使用的「自動調適出現次數門檻」就是將關鍵字詞定義為出現次數較多的字詞，以漸進的方式篩選過濾。而本論文第二章有提到的TF-IDF，就是依據字詞的出現次數、出現文章數給予權重，權重的高低讓常用字與關鍵字彼此區隔。如果能將此方法或概念與MMB結合產生新公式，應該能提升分類準確率。

參 考 文 獻

一、中文部分

王瑄榕(2008)，應用多元貝氏理論於中文郵件分類及知識建立，中國文化大學資訊管理研究所未出版之碩士論文。

周政淇(2009)，以多元貝氏定理建構文件分類系統，中國文化大學資訊管理研究所未出版之碩士論文。

林昕潔(2006)，以SVM與詮釋資料設計書籍分類系統，交通大學資訊科學與工程研究所未出版之碩士論文。

林傑斌，劉明德(2002)，資料採掘與OLAP理論與實務，文魁資訊。

莊惠美(1999)，以智慧型計算方法探討文件分類，屏東科技大學資訊管理系研究所未出版之碩士論文。

劉德舜(2009)，Blog分類之研究，中國文化大學資訊管理研究所未出版之碩士論文。

駱思安(2004)，以Web services建構中文網站階層式分類推論系統，中國文化大學資訊管理研究所未出版之碩士論文。

駱思安，李中彥，徐俊傑(2005)，以多重關係貝式演算法建構中文網頁自動分類系統，中華民國資訊學會通訊(IICM)。

韓歆儀(2003)，應用兩階段分類法提昇SVM法之分類準確率，成

功大學工業管理科學研究所未出版之碩士論文。

二、英文部分

Guo Q. (2010). An effective algorithm for improving the performance of naive Bayes for text classification. *Computer Research and Development, 2010 Second International Conference on* (pp. 699-701). Shanghai : Shanghai University.

Hossaini, Z., Rahmani, A. M., & Setayeshi, S. (2008). Web pages classification and clustering by means of genetic algorithms: a variable size page representing approach. *Computational Intelligence for Modelling Control & Automation, 2008 International Conference on* (pp. 436-440). Iran: Islamic Azad University.

Kim, H. J., & Chang, J. (2007). Integrating incremental feature weighting into naÃve Bayes text classifier. *Machine Learning and Cybernetics, 2007 International Conference on* (pp. 1137-1143). Hong Kong.

Kim, S. B., Han, K. S., Rim, H. C., & Myaeng, S. H. (2006). Some effective techniques for naive Bayes text classification. *Knowledge and Data Engineering, IEEE Transactions on, 18*(11), 1457-1466.

Kourd, M. El, Bensaid, A., & Rachidi, T. (2004). Automatic Arabic document categorization based on the naïve-Bayes algorithm.

Workshop on Computational Approaches to Arabic Script-based Languages, COLING- 2004, University of Geneva, Geneva, Switzerland. USA: PA.

Lee, C. Y. (2004). On using Bayesian approach in recognizing Chinese electronic bookstore web sites. *The Tenth ISSAT International Conference (Reliability and Quality in Design)* (pp. 205-209). USA: Las Vegas, Nevada.

Lee, C. Y., Evens, M., Carmony, L., Trace, D. A., & Naeymi-Rad, F. (1991). Recommending tests in a multimembership Bayesian diagnostic expert system. *Computer-Based Medical Systems, Proceedings of the Fourth Annual IEEE* (pp. 28-35). USA: Baltimore, MD.

Lee, C. Y., Wu, H., & Yu, C. C. (2004). Decision on classifying Chinese commercial web sites by Bayesian approach. *Paper presented at the 4th Annual Hawaii International Conference on Business, Honolulu*. USA*:* Honolulu.

Li, Y. H., & Jain, A. K. (1998). Classification of text documents. *Computer Journal, 41*(8), 537-546.

Liu, C. H., Lu, C. C., & Lee, W. P. (2000). Document categorization by genetic algorithms. *Systems, Man, and Cybernetics, 2000 IEEE International Conference on, 5,* 3868-3872.

Lo, S. A., Lee, C. Y., & Hsu, C. C. (2006). Automatically classify web

sites by multimembership Bayesian approach. *Information Technology: New Generations,2006. ITNG 2006. Third International Conference on* (pp. 580-583). USA: Las Vegas.

Maron, M.E. (1961). Automatic indexing: An experimental inquiry. *Journal of the ACM*, *8*(3), 404-417.

Meena, M. J., & Chandran, K. R. (2009). Naïve Bayes text classification with positive features selected by statistical method. *Advanced Computing, 2009. ICAC 2009. First International Conference on* (pp. 28-33). India: Coimbatore.

Quinlan, J. R. (1992). C4.5: Programs for machine learning. *Morgan Kaufmann.*

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Journal of Information Processing and Managemen*t, *24*(5), 513-523.

Department of information science of faculty of science of University of Tokyo (2006). *GENIA tagger* [Online]. Available: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/ [2006, October 20].

Vapnik, V. (1995). *The nature of statistical learning theory*. NY Springer.

Wang, X., Hua, Z., & Bai, R. (2008). A hybrid text classification model based on rough sets and genetic algorithms. *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD '08. Ninth ACIS International Conference on* (pp. 971-977). Zibo: University of Technol.

Wikipedia (2010). *Cluster analysis mouse.svg* [Online]. Available: http://en.wikipedia.org/wiki/File:ClusterAnalysis_Mouse.svg [2010, October 12].

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 412-420). San Francisco.

Yong, W., Hodges, J., & Bo, T. (2003). Classification of web documents using a naive Bayes method. Tools with Artificial Intelligence, *2003. Proceedings. 15th IEEE International Conference on* (pp. 560-564). USA: Mississippi State University.

Zhao, W., Wang, Y., & Li, D. (2010). A new feature selection algorithm in text categorization. *Computer Communication Control and Automation (3CA), 2010 International Symposium on* (pp. 146 -149). Changchun: Jilin Agric. University.