使用決策樹演算法之電子公文自動分類

葉介山
靜宜大學資訊管理學系
jsyeh@pu.edu.tw

廖偉敦
靜宜大學資訊管理學系
holytia.holy@msa.hinet.net

卓政儒
靜宜大學資訊管理學系

摘要

本研究運用資料探勘技術中之決策樹演算法，針對電子公文進行自動分類，本研究並建置一電子公文自動分類的雛型系統。本研究採用類似 ID3 之決策樹演算法，與原始 ID3 演算法作為分類方法不同之處在於，本研究將輸入的屬性經過簡化步驟，提升系統判斷的效率。分類方法所需的各屬性來自於事先已分類過之電子公文，系統會找出符合期望的關鍵字作為原始屬性，並自動整理出各屬性與分類目標之間的關係。主要分類依據為電子公文的主旨部份，從新進文件的主旨中找出與已知的屬性資料相符合的部份，再根據各屬性與分類目標之間的關係經過 ID3 演算法得出不同的權重。本雛型系統並可透過系統回饋機制(feed-back)，以不斷地提升準確度，以符合本研究的期望。未來該系統應用到圖書資訊管理、行為分析、資訊過濾等各方面。

關鍵詞：資料探勘、電子公文、ID3、決策樹、專家系統。

1. 前言

自民國 88 年起，行政院研究發展考核委員會推動機關公文電子交換作業以來，電子公文往來已成為各政府機關及學校單位文件往返之標準作業。然而，電子公文之分文至今仍採取人工作業，分文作業之正確與否完全仰賴分文者之經驗及訓練。人工方式管理，在準確度與效能上的表現雖然都有一定的水準，卻依然存在著明顯的缺陷。例如：人工的作業方式有一定的工作時間，當文件數量眾多的時候出錯的比率也隨之上升。此外，人工的文件管理方式必須要對該作業流程有某種程度上的了解與認識才能夠步上軌道。

由於以**可擴充之標示語言 XML** (eXtensible Markup Language) 定義公文電子交換之共同傳輸檔案格式，及公文電子交換和處理之技術規範，使得電子公文自動分類之可行性提高。本研究運用資料探勘技術中之決策樹演算法，針對電子公文進行自動分類，本研究並建置一電子公文自動分類的雛型系統。本研究採用類似 ID3 之決策樹演算法，與原始 ID3 演算法作為分類方法不同之處在於，本研究將輸入的屬性經過簡化步驟，提升系統判斷的效率。分類方法所需的各屬性來自於事先已分類過之電子公文，系統會找出符合期望的關鍵字作為原始屬性，並自動整理出各屬性與分類目標之間的關係。主要分類依據為電子公文的主旨部份，從新進文件的主旨中找出與已知的屬性資料相符合的部份，再根據各屬性與分類目標之間的關係經過 ID3 演算法得出不同的權重。本雛型系統並可透過系統回饋機制(feed-back)，以不斷地提升準確度，以符合本研究的期望。

本文章的架構如下：第一節為前言；第二節將探討背景與相關研究；第三節將介紹本研究所提出電子公文自動分類方法；實驗結果與分析則於第四節中探討；第五節為結論與建議。

2. 背景與相關研究

本節將簡述電子公文、決策樹演算法及文件自動分類等相關之背景及研究。

2.1 電子公文

機關公文電子交換，係指將文件資料透過電腦及電信網路，予以傳遞收受者。各機關對於適合電子交換之機關公文，於設備、人員能配合時，應以電子交換行之[4]。公文電子交換之共同傳輸檔案格式係參考以下十二個評估準則： 1.規範共識程度；2.產品可獲得性；3.規範穩定性；4.規格完整性；5.技術成熟性；6.業界使用狀況；7.使用自由度；8.作業之效率性；9.資料之再用性；10.格式之可擴充性；11.系統之整合性；12.文件之呈現方式等[5]。 而 XML (eXtensible Markup Language, 可擴充之標示語言)既符合以上評估準則，並據以定義公文電子交換之共同傳輸檔案格式，及公文電子交換和處理之技術規範。行政院研究發展考核委員會所公佈之電子公文交換類別如表 1。其中又以「令」、「函」、「公告」及「簽」等類別較為常見。

表 1 電子公文交換類別

代碼	內容	代碼	內容
1	令	7	會銜公文會辦單
2	函	8	公文時效統計
3	公告	9	公文欄位轉換格式表
4	開會通知單	A	檔案目錄傳輸格式表
5	簽	B	機關檔案分類表傳輸格式表
6	簽稿會核單		

以圖 1 中「函」之 XML DTD 檔為例，發文機關、函類別、發文字號、附件、主旨及段落等資訊皆可提供收文者公文分發之依據。本研究則以發文機關及主旨標籤中，抓取出所需的分析項目以進行電子公文分類。

```
<?xml version="1.0" encoding="BIG5"?>
<!-- 91_2.dtd 函 2002.1.1 -->
<!ENTITY % 基本標籤 SYSTEM "基本標籤.ent" >
%基本標籤;
<!ENTITY % 交換用標籤 SYSTEM "交換用標籤.ent" >
%交換用標籤;
<!ELEMENT 函 (發文機關+,函類別,地址,聯絡方式+,受
文者,速別?,密等及解密條件?, 發文日期,發文字號+,附件?,
主旨,段落*,正本,副本?,署名*)>
<!-- /函 -->
```

圖 1 函之 XML DTD 檔

2.2 決策樹演算法

在資料探勘(Data Mining)技術中，決策樹(Decision Tree)為分類法(Classification)中重要的方法之一。建立決策樹的演算法步驟如下：

1. T 為訓練範例的集合。
2. 選擇一個最能區分 T 集合中訓練範例之屬性。
3. 將選擇之屬性建立成一樹節點，在往下建立子聯結，而選擇之屬性在各聯結有一唯一的值。從每一子聯結再進一步將範例切割成子類別。
4. 對步驟三的子類別而言：
   甲、如果子類別中的所有範例都滿足先前定義的準則，或是已無其他屬性可選擇，則詳細描述決策樹路徑之分類。
   乙、如果子類別中的範例不滿足先前定義的準則，且至少還有一個屬性可用於切割樹的路徑，則讓 T 為該子類別的集合，並回到步驟二。

屬性的選擇的順序將決定其決策樹的大小，在選擇上，其主要的目標是減少數樹的高度和節點數。目前最被廣泛採用之決策樹演算法包括有 ID3/C4.5[10, 11]、CHAID (Chi-Square Automatic Interaction Detector) [9]及 CART(Classification and Regression Tree)[7]等。其中以 Quinlan 於 1986 年提出之 ID3 演算法最廣為被研究，且為 C4.5 的前身。ID3 演算法如下[8]：

Algorithm: Generate_decision_tree (ID3). Generate a decision tree from the training data.
Input: The training samples, *samples*; the set of candidate attributes, *attribute-list*.
Output: A decision tree.

Proecdure:

1. Create a node *N*
2. if samples are all of the same class, *C* then
3. return N as a leaf node labled with the class *C*
4. if attribute-list is empty then
5. return *N* as a leaf node labeled with the most common class in the *samples*// majority voting
6. select *test-attribute*, the attribute among attribute-list with the highest information gain;
7. label node *N* with *test-attribute*
8. for each known value a_i of the *test-attribute* // partition the samples
9. grow a branch from node N for the condition *test-attribute* = a_i
10. let s_i be the set of samples in samples for which *test-attribute* = a_i // a partition
11. if s_i is empty then
12. attach a leaf labeled with the most common class in tsamples
13. else
14. attach the node returned by Generate_decision_tree(s_i, *attribute-list - test-attribut)*

2.3 文件自動分類

在文件自動分類相關之研究方面，楊允言[1]根據次數、集中度、廣度三項條件，從訓練資料得到具有分類價值的關鍵詞，並以向量模式、機率模式，和不同的分類比重方式來做自動分類實驗。顧皓光[3]使用網路蒐集程式將搜尋引擎上的文件取回，利用這些已經具備分類特性的文件當作訓練文件，並建構一個可以模擬人工的向量空間模型。再由測試資料決定系統正確率。該研究並利用 Web 文件提供了超文件連結的特性及 HTML TAG 標籤加註的功能，藉以提昇系統分類能力。謝儒誠[2]依照資料集的組成狀況，做適當的分門別類，因而不需訓練資料庫。該研究以給定的關鍵字作為文件之屬性，利用 Jaccard 係數測量文件之間的相似度，最後在採用 complete-link 演算法來做分群。

3. 電子公文自動分類

為能使電子公文分類更為詳實準確，我們必須先找出可以作為分類依據的關鍵字，之後將該關鍵字與樣本母體作一個交叉比對，以確定該關鍵字在所有樣本中的分布，而主要的分類方法是依照即時性的決策樹產生之後來分析出分類的目標，當所有的流程完成之後接著就是進行整個系統的準確度提升與校正，此四個步驟的分列如下：

3.1 關鍵詞建立

在樣本母體中，任何的個別樣本公文都是事先以人工方式分類到一個目標(即：承辦單位)。以學校為例：公文之承辦單位有教務處、總務處等。每一個樣本公文都給予 ID 編號以區別之，分析的樣本公文主旨內容儲存於 Sample_context 欄位之中，Sample_sender 則紀錄來文機關，而 Class_id 則是記錄該樣本公文被分類到之分類目標之單位代

號。而關鍵詞的建立，除了可直接加入關鍵詞的作法之外，亦可以從該系統中目前所擁有的樣本集合中，依照分類目標的不同設立篩選條件來找出可用或者是想建立的關鍵詞。關鍵詞建立的過程有以下步驟：

A. 選定分類目標
B. 系統根據所選定的分類目標擷取資料
C. 設定所需要的字詞長度與篩選倍率
D. 使用者選擇關鍵字
E. 完成關鍵字建置

依此產生的關鍵字由於本身就是從已經存在的樣本中產生，所以具有相當的可靠性與可倚賴性，若是設立的篩選度較高，就有可能找出所謂的**絕對關鍵詞**(Absolute-keyword)，絕對關鍵詞是擁指有絕對的鑑別度，具有絕對性的關鍵詞出現在文件的主旨中則此文件有很高或者是絕對的機率會被分配到特定的分類目標中，當關鍵詞建立完成之後就要開始進行關鍵詞與樣本依存關係的比對與紀錄。

3.2 關鍵詞與樣本分類對照集合

當關鍵詞找出來之後我們意識到一個問題，該關鍵字是不是只屬於特定的分類目標？根據經驗，當文件中依據分類找出關鍵字之後有很高的機率該關鍵字也會出現在別的分類目標中。但如果該關鍵字出現在其他分類目標的次數並不大於出現在產生該關鍵字的分類目標中，因此更可以確定選出的關鍵字具有鑑別度，足以作為分類的依據之一。

為了確認樣本與關鍵字之間的分布關係，必須作一個整合性的統計，以了解該關鍵字具體的分布情形。一般來說關鍵字分散的情況越小越好，如果只出現在一個分類目標，即表示該關鍵字為一絕對關鍵字。鑑別度的確立並不是隨著出現的存在分類目標越多就會隨之降低，主要的鑑別度判定，依賴的是包含該關鍵字的樣本屬於那些分類目標的集中情形。

舉例來說，假設有關鍵字 A 存在於 150 個樣本中，關鍵字 B 存在另外的 150 不同的樣本中，而存在 A 的 150 個樣本有超過 70%都出現在甲分類目標，而存在 B 的 150 個分類樣本平均分散到 5 個不同的分類目標中，由此可以得出關鍵字 A 的鑑別度遠大於關鍵字 B，因為可以確定含有關鍵字 A 的新文件有很高的機率會被分到分類目標甲中，而含有關鍵字 B 的樣本卻至少會出現 5 個不同的分類目標而且機率幾乎相等。

在這個步驟之中主要做的動作就是做統計整理，當資料被整理出來以後會直接被寫入資料庫中，以方便日後分析的時候可以用來做分類的依據，每個被找出的關鍵字會個別的與樣本建立關係資料，所以可能會出現某些樣本擁有許多關鍵字同時存在的情形，除了個別的關鍵字具有鑑別度之外，關鍵字與關鍵字之間的組合也有可能會影響到分類的結果，因此需要將該結果作一個完整的紀錄。

所建立關鍵詞與樣本電子公文分類對照表可示意如表 2。其中，K_i 代表關鍵詞，D_i 代表樣本電子公文。例如：第一列代表樣本電子公文 D_1 包含有關鍵詞 K_1、K_2 及 K_3，且樣本電子公文 D_1 被分文至承辦單位 A。

表 2 關鍵詞與樣本電子公文對照表

	K_1	K_2	K_3	K_4	K_5	K_6	K_7	承辦單位
D_1	1	1	1	0	0	0	0	單位 A
D_2	1	0	0	0	0	0	0	單位 A
D_3	0	0	1	1	1	0	0	單位 A
D_4	0	0	0	0	0	1	0	單位 B
D_5	0	0	0	0	0	0	1	單位 B
D_6	0	0	1	0	0	0	1	單位 B
D_7	0	0	0	0	1	0	1	單位 C
D_8	0	1	0	0	0	0	1	單位 C
D_9	0	0	0	0	0	0	1	單位 C

根據以上表格，我們即可進入資料探勘的步驟。

3.3 以決策樹建立電子文件分類

資料探勘策略(data mining strategy)可以被分類為監督式(supervised)及非監督式(unsupervised)。監督式學習藉由輸入的屬性來預測輸出的屬性，並藉以建立模型。其中輸入屬性既是所謂的獨立變數(independent variable)，而輸出屬性既是所謂的相依變數(dependent variable)其結果是由一個以上的輸入屬性所決定。

在眾多資料探勘策略中，**分類(Classification)**具有以下三個特性：

1. 是監督式的學習；
2. 相依變數是類別的(categorical)；
3. 其建立之模型可將任何實例(instance)分類到一明確的類別(category)。

而電子公文分類問題既是非常適合利用分類(Classification)的策略來解決。決策樹(Decision Tree)在資料探勘中監督式學習上是很常用的架構。其中以 Quinlan 於 1986 年提出之 ID3 演算法最廣為被研究。ID3 演算法的主要以訓練資料來建立決策樹，然而，當訓練的資料隨著時間而變的龐大的時候，原本能夠快速完成的工作也會逐漸延長分析時間，為了改善這個問題，我們立基於原始的 ID3 演算法基礎上，作一個小幅度的簡化，以減少需分析的節點與屬性，同時儘量讓系統的準確度不至於偏差過多。

簡化後的 ID3 演算法變成能夠即時性執行分析，提供詳盡的資訊給使用者做參考，同時卻又不會因為關鍵字增多而喪失了效率與便捷性，為了找出文件的分類目標，主要的依據是靠訓練文件中已經存在的關鍵字，本計劃中的作法是將新的分析文件中所擁有的關鍵字抽出之後，再根據這些關鍵字在樣本母體中的分布情形去推算資訊需求值與期望資訊數值，如此一來雖然執行的時候決策樹的產生是即時性的計算，但是不用經過其他無用處的判斷，而且實際上的運作可以證明簡化後的 ID3 演算法所提供的分類資訊準確度並不比原本的差。

當原始 ID3 演算法執行的時候，必須將樣本母體中所有的屬性作一個集合判斷，而簡化過的 ID3 演算法則是將沒有用到的關鍵字抽掉，只留下將判斷的文件中所擁有的關鍵字。以表 2 為例：假設新的樣本中只有 K_1、K_2、K_3、K_5 這四個關鍵字，則系統所得到的新分析表如表 3。

表 3 新關鍵詞與樣本電子公文對照表

	K_1	K_2	K_3	K_5	承辦單位
D_1	1	1	1	0	單位 A
D_2	1	0	0	0	單位 A
D_3	0	0	1	1	單位 A
D_6	0	0	1	0	單位 B
D_7	0	0	0	1	單位 C
D_8	0	1	0	0	單位 C

簡化後的 ID3 演算法所需要分析的屬性減少，而且只需要找存在的關係表比較即可，大大降低分析時間提高執行效率。

3.4 系統建立以及校正測試

為了提升分類的準確度，必須擁有充足的訓練資料與回饋機制，系統的建立採用 C#.Net 作為程式語言，後端資料庫系統採用 SQL Sever 2000，作業環境為 Windows 作業系統。透過網路與遠端資料庫作連線，當系統建立完成之後除了已經存在的樣本母體之外還會有新的文件進入以供分類，為了提升準確性，本系統將關鍵字建立機制獨立出來，讓使用者能夠在新進入的文件數量累積到一定程度之後再次新加入關鍵字以提高準確度。至於關鍵字的數量依照文字使用的習慣來看，可用的關鍵字數量到達某種程度之後就會趨近於飽和，增加的速度會遠遠小於訓練時期。因此，本系統的成熟度可以從新關鍵字增加的數量來做預測，當新加入的關鍵字明顯變少的時候，就表示存在於資料庫內的關鍵字資料已經趨近完備。

4. 實驗結果與分析

以電子公文分類而言，本系統可得有以下五種不同的結果：

i. 無法分類資料

無法分類的資料主因在於，新分析的文件中並不含有已經存在的關鍵字，此時的作法有三，其一：將該文件做人工分類之後直接丟入樣本母體中，期待下次的關鍵字抽取動作可能找出該樣本所含有的關鍵字，因此樣本母體中本身並不是每個樣本都會有對應到的關鍵字存在，樣本母體存在的意義純粹只是提供一個可以產生關鍵字的來源；其二：直接由使用者新增關於該文件的關鍵字進入資料庫，並且稍後將該關鍵字與其他樣本做關聯分析與比較，這個做法比較適合使用於具有絕對關鍵字的新文件，依人工的方式機動性加入絕對關鍵字，可以大幅度提升文件分類的效率，缺點是通常新增上去的絕對關鍵字使用的頻率極低，相對於其他由系統產生選出的關鍵字而言，雖然具有相當的鑑別度，卻不一定符合使用效益；其三：將該文件人工處理後棄置，不寫入樣本母體，這個做法是為了因為文件中可能存在的特例而影響到整個資料的判斷，所謂的特例是指使用頻率很低甚至是不曾出現的文件分類，如果該文件主旨中不含有可供分析價值的關鍵字亦同特例處理，在關鍵字抽取的過程中，不難發現出現最頻繁的字眼通常是所謂的問候語或是敬語，此類的字詞雖然出現頻率極高，但是並沒有參考的價值，因此在抽取關鍵字的時候出現該類的字詞通常都是捨棄不用居多，如果某文件主旨組成大部份為此類字詞，則該文件並沒有成為樣本母體之一的價值，因此採用捨棄的動作以確保資料庫中的資料是必須與具有鑑別度的。

ii. 擁有絕對關鍵字存在

當進行文件分析的時候最好的情況就是擁有絕對關鍵字的存在，當文件中含有絕對關鍵字的時候等於宣告分析完成，具有絕對關鍵字的文件有明確的分類目標，而且只要未來新的樣本母體中其他的分類目標不具有相同的關鍵字，則該關鍵字的鑑別層級絕對是最高的。

iii. 關鍵字組合存於唯一分類目標

很多時候樣本中會出現兩個或兩個以上的關鍵字，而該關鍵字的組合經過系統比對的結果確定只出現在一個分類目標的時候，系統會默認該文件最適合的分類目標只有一個，從文字使用角度來解讀，很多時候文字表達的詞與詞之間具有連帶性，指的是在特定的情況下，使用 A 詞通常也會出現 B 詞，而 A、B 如果都被選為關鍵字的時候，雖然 A、B 分開的時候分別代表的是不同的分類目標，但是當 A 與 B 同時出現在一個文件中則該文件就會被分類到 A、B 重疊且唯一的分類目標。

iv. 擁有多數關鍵字

此類情況是指上述的第 3 種情形之例外狀況，當 A、B 重疊的區域中擁有兩個或兩個以上的分類目標的時候，預設性系統動作會統計出所有擁有該關鍵字組合的樣本分布情形，以提供使用者一個準確度較高的判斷依據，利用已經分類過後的文件作訓練材料時有很高的機會該樣本最適合的分類目標也會出現在建議範圍之中。

v. 無法確實分類

第 5 種情形是最複雜而且分類準確度最低的一種，雖然新的文件中具有已經存在的關鍵字，但是卻沒有辦法分析出可行的分類目標，因此預設系統動作會將所有存在於該文件的關鍵字做樣本母體分佈統計，將所有分類目標可能的機率列出，提供使用者做參考。

本系統以靜宜大學之電子公文為實驗之資料庫。其中樣本數量約 1500 件來做分析，分類目標數為 20 個單位。可提供分類的關鍵詞數量為 295 個，經由系統從樣本中自動篩選出來給使用者做選擇，提供分析的新文件數量約為 1012 件，已知分類範圍為 20 個單位。

進行分類測試結果，有 132 件具有單一分類目標存在，658 件具有兩個或兩個以上的分類目標存在，其餘 220 件文件無法被分類，132 件之中有 11 件檢驗出被分類到錯誤目標，正確率為 0.9167，錯誤率為 0.0833，658 件中正確分類目標出現在第一位的有 431 件，未出現正確分類目標的有 142 件，其餘 85 件為有出現在分類建議列表中但非為首要建議，該情形中將出現於首要分類目標的歸類為正確，有出現但非首選的視為模糊，其餘為錯誤，所得機率如下，在 658 件中正確率為 0.6551，錯誤機率為 0.2158，模糊機率為 0.1291，整體 1012 件中獲得的分析正確率為 0.5454，錯誤率為 0.1511，模糊率為 0.0839，無法分類機率為 0.2173。

首先從正確與錯誤的比例來看，正確率高出錯誤率有兩倍左右，被分類正確的文件大多屬於已知關鍵字較多的分類類別，因此分類結果較為集中，而分類錯誤的文件則呈現不同的分布模式，被分類錯誤的文件散佈於每個分類目標中，並不會因為關鍵字的多少而有所差別，導致該結果的原因判斷為關鍵字抓取的數量不足，以及文件中可提供分類的資料過少，導致分析結果出現誤差。改良的方法主要還是繼續擴充樣本資料庫，抓取鑑別度較高的關鍵字詞作為分析的依據才能有效的降低錯誤率，提高正確分析率。無法分類的機率主因在於可提供分析的資訊過少，主旨可能由無意義的字詞所組成，或者是文件主旨中的關鍵詞並未被建立入資料庫中，後者的問題可以經由擴充資料庫來解決，但是前者目前沒有任何可行的方法，但是在公文系統中，該類的文件出現頻率頗高，模糊率雖然佔的比例不高，卻是最好改進的部份，因為本身已經含有正確的分析結果，改進方法除了增加可判別的關鍵字之外，也可以藉由減少不同分類目標之間的共用關鍵字來改善分析結果模糊的比例。

5. 結論與建議

電子公文為各政府機關及學校單位文件往返之標準作業。然而，電子公文之分文至今仍採取人工作業，分文作業之正確與否完全仰賴分文者之經驗及訓練。本研究運用資料探勘技術中之決策樹演算法，針對電子公文進行自動分類，本研究並建置一電子公文自動分類的雛型系統。本系統雖然只應用於靜宜大學電子公文系統的分類上，而且處於訓練期，但從本系統目前的表現來看，將來可以應用到其他電子文件管理領域，例如：電子郵件的監控、圖書資料的查詢與管理、網路搜尋開發等許多不同的應用。

6. 致謝

本研究為為國科會計畫 (NSC 94-2815-C-126-003-E)之相關成果。本研究並感謝靜宜大學文書組及計算機及通訊中心提供電子公文測試資料。

參考文獻

[1] 楊允言。1993。文件自動分類及其相似性排序。國立清華大學。新竹。

[2] 謝儒誠。2002。資料探勘技術運用於文件自動分群之研究。中央警察大學。桃園。

[3] 顧皓光。1997。網路文件自動分類。國立台灣大學。台北。

[4] 行政院研究發展考核委員會。2000。機關執行公文電子交換作業重點。http://cww.rdec.gov.tw/mis/doc/edoc/method2.htm。

[5] 行政院研究發展考核委員會。2002。文書及檔案管理電腦化作業規範(九十一年修正版)。http://cww.rdec.gov.tw/mis/eg/news/news1.htm。

[6] C. Allen, D. Kania and B. Yaeckel, Internet World Guide to One-To-One Web Marketing. John Wiley & Sons Inc, 1998.

[7] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Wadsworth, 1984.

[8] J. W. Han and M. Kamber, Data Mining: Concepts And Techniques. San Francisco: Morgan Kaufmann Publishers, 2001.

[9] G. V. Kass, “An exploratory technique for investigating large quantities of categorical data,” Applied Statistics, vol. 29, pp. 119-127,

1980.
[10] J. R. Quinlan, “Induction of Decision Trees,” Machine Learning, Vol. 1, pp. 81-106, 1986.
[11] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

附錄

系統畫面

提供使用者輸入 SQL Sever IP 以及使用者帳號與密碼的功能，只要具有權限，就能夠透過使用者介面擷取遠端資料庫中的資料進行分類與分析的工作，登入後系統會自動去抓取遠端所需要的資料，自動進入工作畫面。

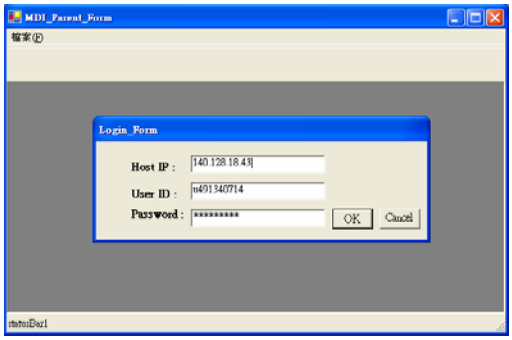

圖 2 系統遠端登入畫面

登入後使用者可以根據左手邊的樹狀結構點選來選擇自己想要了解的資料，上下兩個不同的母節點代表的分別是遠端資料與本機資料，為了保持資料的一致性，當系統完成工作之後本機所處理完成的資料會自動更新上遠端 Sever，不需要使用者自己動手更新。

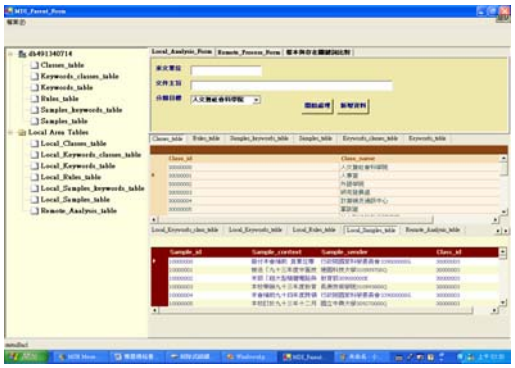

圖 3 系統操作畫面

當新的文件透過輸入畫面輸入完成之後進行分析如果有可以分析的元素則會另外開啟一個分析結果視窗給使用者參考，由於該系統目前還是處於訓練時期，因此自動新增的功能並不開放給系統自行使用，而是讓使用者透過選單將文件新增或不新增上樣本資料庫。

圖 4 新文件分析結果

設定好擷取關鍵字的來源與字詞長度及篩選倍率之後，系統會將符合條件的字詞列出在左側給使用者做選擇，被使用者選上的關鍵字可以點選到右邊按下 Done 的按鈕就可以將資料新增到資料庫裡面，當然因為使用者有可能重複選到已經存在的關鍵字，因此系統裡面本身就有檢查機制以避免資料重複造成資料誤判，影響分析結果。

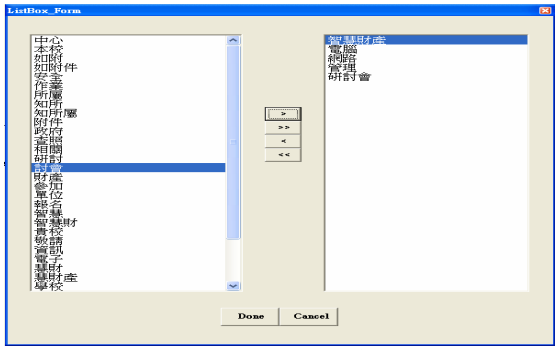

圖 5 關鍵字擷取畫面

完成關鍵字的選擇之後就要將關鍵字與現存的所有樣本作關聯分析，由於這個部份並不需要使用者的介入，因此只會出現一個等待畫面，因為這個步驟所做的工作比較大量而且不常使用，所以作業方法直接寫入在本機端，並不採用遠端處理，只將處理完的資料傳上資料庫。

圖 6 自動建立關聯