結合知識地圖之公部門陳訴文件自動化分案系統

鄭麗珍
東吳大學資訊管理學系

賴美惠
東吳大學資訊管理學系

摘要

電子化政府是政府部門於便捷的網路環境上提供便民服務，讓民眾可以不用出門便可處理各項業務。而為了提供民眾申訴或表達意見的管道，在政府部門網站內提供「首長信箱」功能。為表示對於民眾意見的重視，因此必須快速且正確地將民眾意見分案至相關單位進行處理及回應。本研究透過訪談方式，發現在這些分案專家的腦中似乎有張各部門工作職掌的知識地圖一般，可以快速且準確的做好分案工作。過去關於文件分案的研究，都忽略這一塊，而直接用文字探勘的技術來做分類。基於上述緣由，本研究嘗試提出二階段的分案處理模式建立文件分案自動化的機制，減少人工作業的流程及成本，提升政府的服務效率。本文所提之二階段分案處理模式，首先運用訓練資料及文字探勘技術來建立知識地圖，接著透過新進文件與知識地圖的比對進行文件分案的預測。經實驗證明，本文所提出之分案模式可以準確且有效率的達致預期的目標。

關鍵字：知識管理、知識地圖、文字探勘、關聯規則

An automatic classified system for public opinion in E-government based on knowledge map

Li-Chen Cheng
Department of Computer Science & Information Management, Soochow University

Mei-Hui Lai
Department of Computer Science & Information Management, Soochow University

Abstract

E-government means using information and communications technologies to improve the transparency, efficiency, and effectiveness of public institutions. The government can provide many convenient online services to serve the civil well. Especially, many local governments in Taiwan have used "Mayor's Email-box" to handle citizen complaints or opinions. Quick responses to those complaints or opinions are needed in e-government services. Therefore, to classify the citizens' opinions quickly and correctly has become an import work for the government. Through interview with the domain experts, we found that "knowledge map" can improve the efficiency of classification system which was ignored before.

This study proposes a framework which contains two phase works. First, the knowledge map should be built up based on the association rules discovered from the training data. Next, the unclassified documents will be categorized according to the knowledge map.

To verify the proposed framework efficiency and accuracy, extensive experiments are conducted using real data sets. The experimental results indicate that the proposed method is computationally efficient, and can effectively categorize documents.

Key words: Knowledge management, Knowledge map, Text mining, Association Rules

壹、導論

近年來，網際網路技術快速發展，網路頻寬由低速的窄頻升級到高速的寬頻、光纖，再加上可攜式裝置的普及，促使了網路的使用者大幅增加；同時隨著資訊科技的演進與產業競爭環境的發展，對於客戶關係管理（Customer Relationship Management, CRM）來說，建立一套完整的客戶資訊系統是最重要的，其目的在於為了管理與老顧客的關係，發展出適合顧客個別需要的產品服務，提高顧客最高的忠誠度、留住率與利潤貢獻度，並同時有效率選擇性的吸引好的新顧客。對於政府部門而言，民眾便是它的顧客，同樣也需要有良好的客戶關係管理，來提升民眾對政府部門的滿意度。

根據客戶關係管理理論及運用新興資訊技術，於是許多企業開始利用網際網路提供各種網路服務，方便顧客無時間、地點限制的進行網路服務，如：購物、技術諮詢、問題反應…等。同樣地，政府部門也開始有了以網際網路提供便民服務的想法。行政院研考會於民國86年訂定了「電子化/網路化政府中程（87至89年）推動計畫」，期望讓政府資訊及服務更加方便，民眾隨時隨地都可取得。范錚強（1999）綜合其他學者說法，認為電子化政府可提供全天候的服務，並以人民的需求為目標，由機關供給導向的個別服務轉換成以人民需求為導向的整合服務。

電子化政府是政府部門自民國85年起即開始陸續規劃、推動的計畫，運用資訊科技推動各項服務，而其中最重要的是於便捷的網路環境上提供便民服務，包括線上申請、線上繳稅…等等，讓民眾可以不用出門便可處理各項業務。而為了提升民眾對政府的滿意度，政府單位必須傾聽民眾的心聲，故提供民眾反應意見的需求因而產生，於是政府單位均於單位網站內提供「首長信箱」功能，讓人民可以填寫問題或是陳訴意見。為避免回應過慢造成民怨，有必要快速且正確地將民眾意見分案至相關單位進行處理及回應。

在一般企業網站中常使用「聯絡我們」這項功能來讓使用者提出問題，藉由此功能來讓企業了解使用者的需求或問題，進而來提升與顧客之間良好的關係；過去許多研究利用一般企業之FAQ功能，設計自動回覆的機制，以系統自動處理顧客所提出的問題或是意見。此自動機制除節省人力成本外，也能在最短時間內給顧客回應結果；但由於一般企業客訴問題具有一定的特性，多半的範圍僅限於產品種類、應用範圍等，通常由單一窗口直接回覆即可，處理流程較簡單，故可蒐集過去的FAQ來綜整歸納回覆的內容，以產生相關對應的回應內容規則。

反觀政府部門，雖然各機關網站上都提供「首長信箱」功能，來做為與民眾間溝通的橋樑。然民眾陳情或反應之文件繁雜，常需分門別類後指派給不同職掌的單位負責處理。政府部門為了要在短時間內正確地回覆民眾的意見，目前大多依靠專業人員來進行首長信箱內容的問題分辨及轉送相關權責單位處理，此專業人員必須經過長期訓練、熟悉各單位工作職掌，才能進行正確地分案。另外，也有少數政府單位有建立自動處理的機制，但前提是填寫者須在填寫的同時決定問題的分類屬性，系統依填寫者所選擇屬性

來自動分案至相關單位處理（郭瓊蓉 2005）。惟此機制對於填寫者來說並不方便，一方面填寫者可能無法正確判斷問題分類屬性（雖然系統已經設定大項分類讓填寫者選擇，但因公務部門牽涉範圍廣大，且各單位所管轄業務不同，並非一般民眾能清楚分辨的）。

透過了解現有「首長信箱」分案流程，發現目前專人處理這些案件時，這些專家的腦中似乎有張各部門職掌工作的知識地圖一般，所以可以快速且準確的做好分案工作。本研究提出兩階段的演算法來解決此一困難的問題。第一階段先將訓練資料利用資訊檢索概念，將這些陳情的文件特徵詞挑選出來，並利用文字探勘技術，找出單位與文件特徵詞之間有意義的關聯規則，希望能依此關聯建立各單位的知識地圖。第二階段，主要負責自動分案的功能。當新的文件經過斷詞等處理後，利用與知識地圖比對的演算法，以預測出可能的所屬單位。這個首長信箱文件的自動分案機制，期望能以建立自動化分案之機制，減少人工作業流程，降低人工處理成本，縮短處理時間，在最短時間內給予民眾回應，同時提升政府部門積極服務之印象。

本研究所提出之二階段分案模式將透過某行政單位的資料來進行實驗驗證。實驗驗證包含幾個部分：首先，針對第一階段的知識地圖，透過自動建立與專家檢視的互相驗證來檢驗此地圖的準確度；接下來，透過各種參數的調整與實驗，找出適合建置地圖的參數與分案文件的權重，以檢驗自動分案的準確度；最後，將本文所提出之系統與另一知名分類演算法支持向量機（Support Vector Machine, SVM）進行預測結果比較，本研究的演算法有不錯的表現。

本文總共分為五章，其架構如下所述，第二章是文獻探討，將回顧知識管理與文字探勘之相關研究；第三章為研究架構，包含參數上的定義及提出之演算法的解說；第四章為實驗設計與結果，會針對本文所提出之演算法，來進行準確性的實驗；第五章則為本研究結論說明與後續研究建議。

貳、文獻探討

本研究所提出的演算法中需要先建立知識地圖，接下來新進文件將透過與知識地圖的比對，提供系統進行自動分案之依據。本研究所利用之觀念、技術有「關聯規則」、「文件分類」、「知識地圖」等，以下將簡略　述過去相關文獻。

一、關聯規則（Association Rule）

資料挖礦（Data Mining）以自動或半自動的方式，從給定的資料庫中萃取出隱含且有用的模式。資料挖礦的模式很多，而關聯規則是最常被應用的模式之一。其中以Agrawal et al.（1993）首先提出Apriori演算法最為有名，主要從日積月累的交易資料中擷取關聯規則來顯示銷售商品之間的關聯性，作為未來行銷參考，其著名的商業應用即為購物籃分析（Market-Basket Analysis）。

關聯規則分析（Association Rule Analysis）主要是從龐大銷售商品資料庫資料中，

探索銷售商品間有趣的關係或相關性。關聯規則之定義描述如下，令$I=\{i_1,i_2,...,i_m\}$是所有相異物品項目（Item）的集合。T是指一筆交易（Transaction）內物品項的集合。而D則是所有交易記錄T的集合。若產生的關聯規則為發生X情況下，有很大機會發生Y（If X then Y），則規則表示形式為$X \Rightarrow Y$，X為前提項目組（Antecedent Itemset），Y為結果項目組（Consequent Itemset），X和Y皆為I的子集合，且$X \cap Y=\varnothing$。關聯規則的成立，必須滿足決策者所訂定之最小支持度（Minimum Support Threshold）和最小信賴度（Minimum Confidence Threshold）。關聯規則要有意義，其支持度與信賴度必須大於或等於所訂定之最小門檻值。因此，支持度（Support）與信賴度（Confidence）為關聯規則是否具有顯著意義的衡量指標。

應用關聯規則在文字探勘的研究，其基本觀念是把每篇文件（document）視為傳統關聯規則中資料庫內的一筆交易，而文件內的字詞（term）視同交易內的商品項目。目前應用關聯規則在文字探勘上可以分為兩大方向，首先有一部分的學者主要應用關聯規則在特徵選取階段，主要著力於字詞間的關聯性，以各種角度探討哪些字詞會一起出現，這之間的關係將代表特定的意涵。另一部分的研究，是以關聯規則為基礎的文件分類研究。

在字詞關聯性的研究上，Feldman et al.（1997）提出改良式的文件關聯規則探勘技術- 最大關聯規則（Maximal Association Rules），以此來計算特徵詞彙同時發生的頻率，找出文件中的重要資訊。Haddad et al.（2000）將關聯規則技術所探勘出文件中詞彙的關聯性，應用於資訊檢索系統可以提升檢索效能。後續許多中外的學者，都利用不同的角度著力於改良關聯規則演算法，希望更有效地找出文件內字詞的關聯性。在中文的資料檢索中，許中川等（2001）利用中文的新聞文件，嘗試了解關聯規則這樣的研究架構，是否適合應用在中文的資料檢索上，並提出許多有用的建議。侯建良等（2004）利用字詞關係為基礎，將客戶閱讀文件做分群處理。邱登裕等（2006）利用找出文件中詞彙的關聯性來建立知識地圖，以作為分群的基礎。陳良駒等（2009）發現利用關聯規則找出的字詞關聯性優於傳統的共詞分析，更適合用來建立分群的模型。

另一部分的學者，主要是延續Zaïan在1993年提出的文件分類法，這一類的方法主要探討的是文件與文件間所屬特定類別的字詞之間關聯性。有的學者改良關聯規則的演算法本身以提升分類的速度（李卓銘 2006），有的學者主要在規則建立後，以不同門檻值設定調整支持度與信賴度來篩選出有用的分類規則（魏莉斐 2006；陳育民 2008）。過去的研究，關聯規則所找出來的規則，會依照信賴度排序以示其不同的重要性。

本研究主要是以文件分類法為藍本，利用不同的門檻值找出有用的分類規則，並以每個規則的信賴度計算出相對應的權重，當作建立知識地圖的基礎。在文件分類比對部分，加入文件結構的概念，以調整文件中不同的位置的字詞之權重，例如：加重出現在主旨的特徵詞，這些都可以提升文件分案的正確性。

二、文件自動分類

在資訊爆炸的時代，由於文件資料的量不僅多且種類也很多元，因此將文件做適當

的分類管理是一件很重要的事，可以方便使用者依所需的類別擷取資料。文件分類主要是指是將文件事先依照其內容，指派到事先定義好的類別的過程。例如：新聞文件可依其報導的內容，歸類為政治、財經、等類別，而影響文件分類績效之相關因素，包括：特徵選擇、特徵詞彙刪減、前置摘要處理、分類器選擇、分類架構、文件標示原則、類別選擇、分類不一致等原因（曾元顯 2002）。

目前有許多學者提出各種演算法，希望可以提供更有效的分類結果。一般而言文件分類的模型可以分為三階段，特徵選取、文件表示與分類歸納階段 （許中川 2001）。在特徵選取方面，是文件分類最基本動作，主要是要挑出重要的特徵字詞作為基礎，後續都將以這些特徵字詞來代表文件做後續的處理，有許多文獻都著墨在這方面的研究，最常見的有利用卡方檢定的方式或是詞頻—逆向文件頻率（Term Frequency - Inverse Document Frequency, TF-IDF）等方法，其目的都希望找出與特定類別相關的字詞當作基底。在文件表現階段，常見的有布林模型與TF-IDF這兩大類別為主。最後在分類歸納階段，有許多文獻結合各種機器學習的演算法來建立分類模型，常見有的利用貝式分類法（Bayesian classification）、支持向量機（Support Vector Machine, SVM）、關聯規則式分類…等方法應用在文件分類的問題上（Joachims et al. 1998；許昌偉 2005；郭瓊蓉 2005；李卓銘2006；Kim et al. 2006；Hao et al. 2007；林昕潔等 2008；陳育民 2008）。其中，支持向量機與關聯規則式分類是目前兩種最常見的分類方式，本研究主要提出的演算法即是以關聯規則式分類為基礎。

三、知識地圖

近來知識管理在企業組織內蔚為風行，希望透過知識管理將企業組織內重要的技術知識累積傳承下去，提升組織的競爭力。而所謂知識是一種合理的信念或想法，是增加實體的能耐有效地行動的基礎（Huber 1991；Nonaka 1994）。利用知識可提升組織的競爭力，因此知識可被視為組織內部的一種無形資（Alavi et al. 2001）。如何管理這項無形資產，便開始提出許多管理的方法。其中Allee（1999）認為知識管理是將組織的隱性知識轉化為顯性知識以利知識的分享、更新及補充。Zack（2002）則提出知識管理是攫取知識、編輯知識、發展知識分類方法、發展散播知識、教導員工創新、分享及使用知識的過程。另Davenport et al.（1998）指出知識管理應包括有專家系統、人工智慧、知識庫及知識管理科技應用等技術；其分為兩個層面研究，一是由組織面來看知識在組織中的價值，並探討如何創新、保存組織中有價值的知識；另一層面則強調運用資訊科技來發掘未知知識的技術與運用，使知識管理有更豐富的內涵與價值。

Davenport et al.（1998）認為知識管理的具體工作有三，一為建構知識地圖或超文件，透過知識地圖可以讓組織中的知識項目和彼此關係清楚地顯現；二為塑造知識管理的文化，鼓勵知識的分享；三為建構知識的基礎建設，包括利用資訊系統與人員之間的互動及協同合作來提升知識的力量。

McCagg et al.（1991）認為知識地圖、概念圖與語意網路都是以用視覺化概念、知識與關聯之表達方式，其中知識地圖是一種視覺化表達知識來源與關係的工具（Kang

et al. 2003）。知識地圖（Knowledge Map）又可稱為知識分佈圖，是用來描繪知識及其分布的狀況，以協助組織利用知識進行對組織有用的策略（Davenport et al. 1998；Zack 2002），也就是一種知識的視覺化指南，告訴人們知識的所在位置，顯示有哪些資源可以利用，提供可以按圖索驥的優點。因此，知識地圖是搜尋知識的概念，從知識地圖找到知識的所在位置，同時將知識與知識間建立連結，讓人們可以輕易地找到所需要的專業知識。

知識地圖主要找出知識間的關係，將此關係以視覺化的方式呈現，以資訊檢索的角度來看，知識地圖可以視為以特徵式的方式進行文件分類的基礎（Lin et al. 2006）。透過文字探勘的技術所找出的字詞關聯性所建立的知識地圖亦可作為文件分群的依據（邱登裕等2006）。

參、研究架構

本研究首先以訓練資料建立知識地圖，接著透過比對的方式，為新文件找出所應負責處理的單位。系統架構如圖1所示，主要分為知識擷取及文件分案二個模組，在知識擷取模組主要為訓練階段，先利用文件處理模組找出文件的特徵詞，再使用知識地圖建立模組，利用關聯規則技術，找出特徵詞與單位之關聯，進而建立知識地圖；而文件分案模組則是實際應用階段，先進行新進文件處理模組找出新進文件的特徵詞，再透過分案處理模組將文件與知識地圖做比對，進而決定應送往哪一個單位處理，詳細流程說明如后。

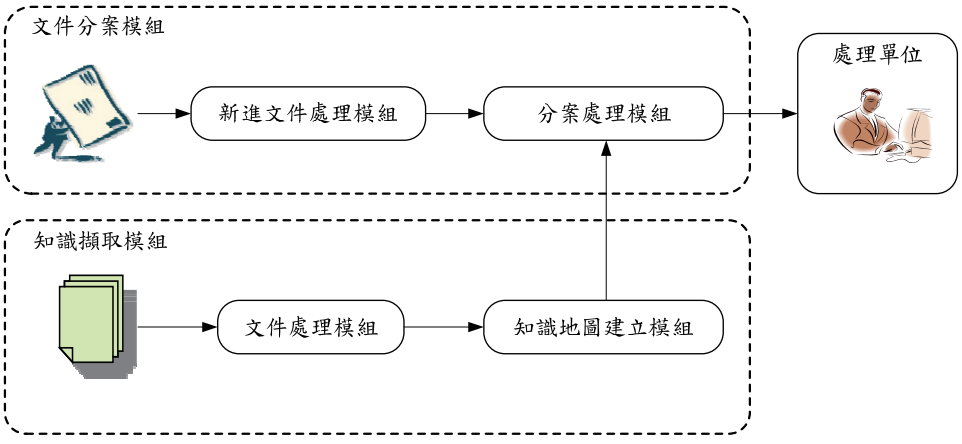

圖1：系統架構圖

一、知識擷取模組

知識擷取模組主要功能是為建立自動分案之知識地圖，作為未來新進文件進行分案時，文件比對的依據，透過新進文件與事先建立好的知識地圖比對，可預測出分送處理

的單位，以減少人工作業。此模組包含文件處理模組及知識地圖建立模組，各模組之處理流程分別由以下細項說明。

（一）文件處理模組

文件處理模組是整個演算法最基本的工作，將從各文件中擷取出有用的特徵字詞，作為建立特徵詞表的基準。主要為三個步驟，分別為文件前處理、文件特徵詞處理及文件表現，處理流程如圖2。

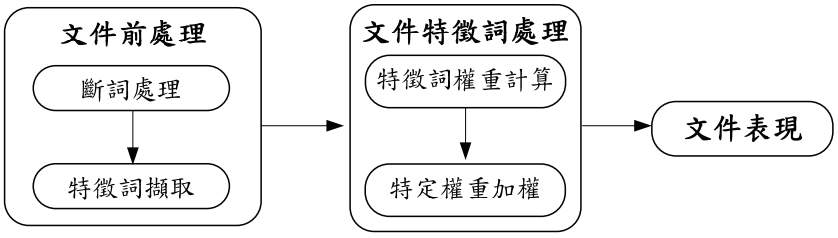

圖2：文件處理模組流程圖

1. 文件前處理

這部分主要包含文件斷詞處理與特徵詞擷取兩個重要的步驟。中文的文件前處理與歐美語系的文件有很大的不同之處，一般英文文件只要以空白當區隔將文件分解成一個個的詞（word）即可做後續的處理，本研究實驗對象是以中文文件為主，中文沒有空白可以斷開詞彙，目前最常用工具就是透過中研院所研發的CKIP中文斷詞系統，來進行中文文件斷詞處理。接下來透過詞性合併的規則希望擷取出有意義的特徵詞。

（1）斷詞處理

斷詞系統先將內容切分成一組組的詞彙，並給予每組詞彙詞類標記，例如詞類標記Na/Nb/Nc分別是普通名詞、專有名詞及地方詞，且均可統一歸類為N（名詞）。接著將各文件中標點符號等不具有語意的符號過濾掉，接下來在斷詞技術所擷取出文件中所有的字詞內容中，過濾掉不必要的停用詞（stop words），因為這些停用詞通常是和語義無關的字詞。在中文的特徵詞彙中，名詞（Nouns）與動詞（Verbs）所代表的意義最重要，較能代表文件中重要的概念，因此，本研究僅保留名詞與動詞的單字詞，其他詞性的字詞均予以忽略。

（2）特徵詞擷取

接下來將前一階段所產出之結果，整理合併出可以代表文件的特徵詞表。根據斷詞系統所產生的結果，本研究雖然僅保留名詞與動詞的詞組，根據過去文獻的建議若不做部分詞性組合，將會產生很多無意義的字詞（邱登裕等 2006）。考慮本研究的文件內容特色，發現當「護理人員」及「站務人員」兩個詞經斷詞系統後，分別得到護理（Na）人員（Na）及站務（Na）人員（Na），在此種情形下，若不經詞性組合，所得到的「人員」這個詞對於本研究後續的知識地圖建立上並無貢獻，因此參考林厚誼等人（2002）所提出之詞性合併規則，僅就連續的名詞作詞性合併。本研究之詞性合併規則，如下：

詞性組合	範例
N＋N	幼兒（N）＋疫苗（N）＝幼兒疫苗（N）

此外，為能使特徵詞對於未來分案時具有決定處理單位之特性，本研究先請專家針對各文件以人工方式判斷出所屬的負責處理單位，將單位名稱加入文件之特徵詞表中。

檢視特徵詞表內容，發現部分專有名詞（如疾病名稱、藥品名稱）因出現頻率有限，在知識地圖建立模組階段，可能無法找出相關有效的關聯規則，為避免此問題，將蒐集各機關網站、FAQ問答集等內容，依各機關屬性建立其特定名詞之特徵詞表，以特定名詞取代出現頻率較少之專有名詞，例如以疾病取代糖尿病、H1N1，以藥品取代壯陽藥、諾美婷等。

2. 文件特徵詞處理

並非所有的詞彙都是重要的詞彙，所以要透過權重篩選的方式，以保留重要的特徵詞。特徵詞權重計算主要包含了兩個步驟，首先是利用TFIDF加權模式計算各特徵詞權重，最後根據特徵詞出現的位置與事先設定的特定特徵詞作權重加乘。

一般而言，文件詞彙權重計算模式有TF加權模式、TFIDF加權模式 (Salton 1983)等等。TF是指詞頻（Term Frequency），表示字詞出現的次數，以字詞出現在某一文件中的頻率代表該字詞的重要性，如果出現頻率越多則表示越重要。TFIDF是一種統計方法，用以評估某字詞對於資料庫中的其中一份文件的重要程度。其計算公式如下：

$$tf_{ij} \times idf_i = tf_{ij} \times \log \frac{N}{n_i} \qquad (1)$$

其中tf_{ij}為字詞t_i在文件j出現的次數，n_i為資料庫中含有字詞t_i出現的文件篇數，N為資料庫總文件數。其公式的精神在於字詞的重要性隨著它在各文件中出現的次數成正比增加，但同時會隨著它在資料庫中出現的頻率成反比下降。

考量本研究之資料來源為民眾申訴或提供意見之內容，文件的長短不一，單用詞頻來計算權重會受到每篇文件字詞多寡所影響。為能真正找出各文件內重要的特徵詞，希望這些詞可以具鑑別力，本研究選擇以TFIDF加權模式來訂定特徵詞之權重，因為其考量到詞頻以及詞彙出現文件的多寡。

另一方面，本研究也結合文件結構的觀念，來加強權重。文件中的特徵詞出現的位置不同，也將給予不同的權重。在各政府機關網站，提供民眾填寫意見之格式大部分為主旨和內容兩段，歸納一般人撰寫的習性，主旨通常為文件的重要概念，位於主旨內的特徵詞可代表文件的重要性更高，故加重位於主旨特徵詞之權重；另若於文件內容已有指出機關或單位名稱，則該文件分案至該機關或單位之機率越高，為提升正確分案率，故增加此類機關或單位名稱特徵詞之權重。

3. 文件表現階段

特徵詞的權重可代表著該特徵詞在整篇文件所占的重要性，特徵詞權重越高越可表達文件概念，根據本研究的資料特性，每篇文件都挑出特徵詞權重前10高的特徵詞以代

表文件。將所有文件都整理以特徵詞表現，以此當作建立知識地圖的基礎，透過編碼轉換以方便下階段關聯規則探勘。

（二）知識地圖建立模組

文件處理模組將所有文件都轉換成以特徵詞表現，並存入文件資料庫。此階段將文件資料庫的各文件資料，利用資料探勘技術產生關聯規則，透過設定關聯規則最小支持度及信賴度門檻值，以挑選出真正有效的關聯規則；從有效的關聯規則中，找出各單位與特徵詞之關聯。由於機關單位本來就有既定概念階層，可以依照資料來源已知的單位，亦可建立多階層的知識地圖。接著，以同一機關單位為基準，在此單位之下所整理出來的規則，依照各規則的信賴度值排序，並正規化設定各自的權重，進而建立以單位為中心之知識地圖，未來即可由分案模組的演算法來決定應分案至哪個機關或單位來處理民眾意見，處理流程如圖3。

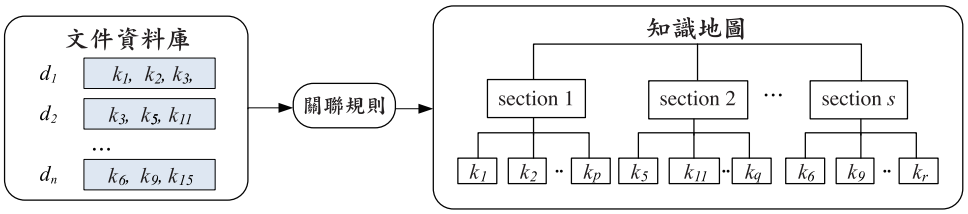

圖3：知識地圖建立模組流程圖

二、文件分案模組

文件分案模組是為新進文件選擇出最正確的處理單位，可分為新進文件處理模組及分案處理模組，各模組詳細流程詳述如下。

（一）新進文件處理模組

本模組主要工作是進行新進文件的處理，整個過程與訓練資料的文件處理模組一樣。經過中研院的中文斷詞系統進行文件斷詞，再依本研究提出的詞性合併規則，進行詞組合併，以產生有意義的詞組。找出文件的特徵詞並計算其權重。本模組再依權重給定規則結合文件結構，依照特徵詞出現的位置加權或部分特定特徵詞的權重加權，依照權重排序，挑選前10名之特徵詞來代表文件的重要概念。

（二）分案處理模組

本模組工作為將新文件分案至應處理之單位。經過前一步驟，新進文件處理模組會產生出有用的特徵詞，接著與知識擷取模組所建立之知識地圖比對，透過比對函數計算權重值總和最高者為處理單位，文件特徵詞與知識地圖比對說明如圖4。

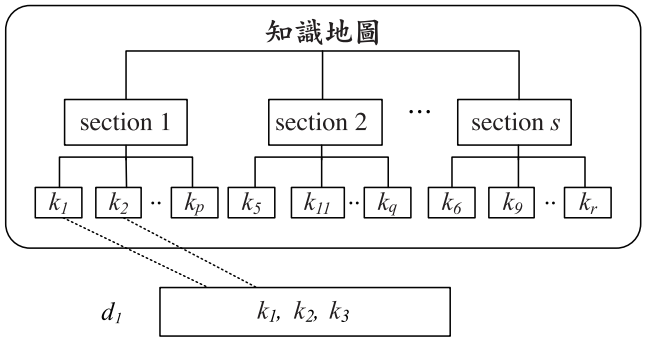

圖4：文件特徵詞與知識地圖結構說明

本研究研究依照特徵字出現在文件與知識地圖交集的詞組，理論上在新文件中越重要的詞組出現頻率高，且出現在某個知識地圖的重要點上，就代表屬於這單位的可能性越高。參考過去黃國禎（2007）的研究，本研究提出一個演算法，其中權重值$w(S_k)$越大代表新文件d_i與某單位S_k的歸屬度越高，代表傾向屬於某單位S_k。演算法如下

Input :
document database: d_i
Knowledge_map: M
Output: section number // 部門編號
Method:
For each section S_k in M
For each keyword k_j in d_i
For each keywords k_p in section S_k
If ($k_j == k_p$)
$w(S_k) = w(S_k) + w(k_j)*w(k_p)$

Select the maximum value of $w(S_k)$ //挑選部門權重最大者

利用迴圈判斷新文件d_i的每個特徵詞k_j是否存在於某單位S_k的知識地圖裡，若有則$w(S_k)$會利用公式計算知識地圖對應該個節點的權重乘積，其中$w(k_j)$為代表其特徵詞k_j在文件d_i的權重，$w(S_p)$ 代表其特徵詞k_p在該單位section S_k的權重，將兩個權重相乘，代表在新文件中越重要的詞組，且出現在某個知識地圖的重要點上，就代表新文件屬於這單位的可能性越高。將地圖各節點走完一遍，挑出最大的節點累積權重值，則代表新文件最可能分派到該單位。

肆、實驗設計與結果

本研究是期望能針對政府部門對於民眾所提出的意見，給予最快速的處理，並回覆處理結果，因此本研究利用某機關的民眾投書意見資料當作訓練資料，因政府部門所負責職掌種類繁多，為能有效驗證本研究效果，初期先以與民生問題相關的單位為主，有交通單位、衛生單位、環保單位、社福單位及教育單位等五個單位，各100篇共有500篇

文件來進行實驗，並將資料以7:3的比例區分為訓練資料和測試資料。

本研究主要使用工具分別為中研院CKIP中文斷詞系統，來進行文件內容的斷詞處理，斷詞後詞性合併則是自行撰寫程式來進行詞性的合併，整理後的特徵詞則存入Microsoft SQL Server 2008資料庫內，同時利用其內建之Analysis Service來進行關聯規則之產出及計算其特徵詞之權重。

一般文件比對的研究，最常使用的基本指標就是召回率（Recall Rate）、準確率（Precision Rate）及F-measure，本研究亦採用此指標來檢驗文件自動分案演算法的研究結果之精準度，其計算公式如（2）、（3）、（4），公式內變數關係圖如圖5，其中A為人工分案的結果，S為系統自動分案的結果，A∩S為A和S的交集結果，當召回率與精確度的值愈高時，其F-measure的值也愈高，這表示其演算法的效能也愈好。

$$Precision\ rate = \frac{A \cap S}{S} \tag{2}$$

$$Recall\ rate = \frac{A \cap S}{A} \tag{3}$$

$$F-measure = \frac{2 * Precision * Recall}{Prcision + Recall} \tag{4}$$

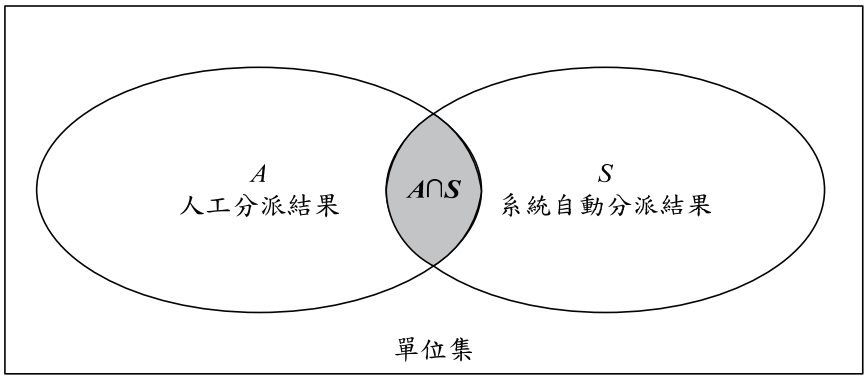

圖5：Precision rate與Recall rate關係圖

本文所提之分案系統建立的實驗分兩階段，第一階段先找出一個建立知識地圖的最佳模式，由於知識地圖主要的步驟是將斷詞之後的特徵詞，利用關聯規則做後續的處理，如何找出具代表性的特徵詞與如何找出適當的參數來整理關聯規則，都是這階段實驗的重點。（結果見實驗一）

第二階段的實驗重點，則在如何結合文件結構在文件分派上，本研究的文件結構有文件主旨與內文等，將利用不同的參數實驗出最佳的參數組合，找出最好的實驗結果。（結果見實驗二）

最後為驗證本文所提之系統分案準確性與演算法之效果，本文所提出之二階段分案系統亦與目前文件分類最有效的支援向量機演算法做比較。（結果見實驗三）

第一階段 知識地圖處理階段

本研究所提出的演算中，最重要的部分在於知識地圖的建立，也就是本研究系統架構中的知識擷取模組，包含資料的蒐集，資料前處理、透過各項參數設計，利用資料探勘軟體建立關聯規則，最後產出對應的知識地圖，知識地圖建立的完整性將影響後續文件自動分案預測的準確性。

關聯規則的產出在建立知識地圖階段是最重要的工作，關聯規則有2個重要參數最小支持度（minimum support）及最小信賴度（minimum confidence），同時在產生關聯規則時，會一併計算出規則的信賴度（confidence），此值可做為分案模組新進文件特徵詞與知識地圖比對參考之權重值。

考量民眾一般在填寫意見時，大多以口語化的詞句陳述，因此相同意義的詞句對於不同人便會有不同的表達方式，此情形將造成經斷詞系統斷詞及詞性合併後的特徵詞不盡相同，為避免重要的特徵詞因此未選入單位的特徵詞集合，故實驗將設定較低的支持度；目前設定最小支持度為支持個數2，最小信賴度為0.7，所產生的關聯規則簡略摘要於表1，權重值則是為該規則信賴度透過正規化方式所轉換得之，以代表該規則對於該單位之重要性。

表1：關聯規則摘要

單位類別	關　聯　規　則	信賴度（Confidence）	權重
交通單位	停車,機車->交通單位	0.727	0.00758
環保單位	廢棄,機車->環保單位	1	0.018108
衛生單位	食物->衛生單位	0.8	0.022743
社福單位	身心障礙,補助->社福單位	0.75	0.021253
教育單位	補習班->教育單位	0.875	0.026926

產生之關聯規則經整理所有規則，參考其信賴度，依據單位類別分別產生單位之知識地圖，如圖6，由特徵詞和其所代表之權重組成知識地圖的節點，作為第二階段文件分派階段的比對依據。

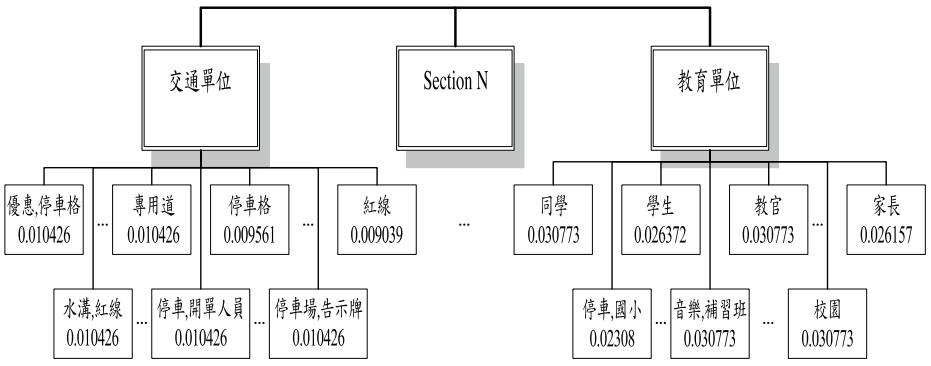

圖6：知識地圖

為驗證此知識地圖之有效性，特請專家針對訓練資料，將每篇文件之特徵詞標示出來，然後與知識地圖進行比對，套用公式（2）、（3）計算知識地圖與專家特徵詞之召回率及準確率，比對結果如表2。

表2：專家特徵詞與知識地圖比對結果

分派單位	交通單位	環保單位	衛生單位	社福單位	教育單位
召回率Recall Rate	93.80%	91.70%	90.30%	86.70%	95.20%
準確率Precision Rate	78.90%	78.60%	80.00%	72.20%	80.00%

經過與專家比對的結果，本研究的知識地圖確有其有效性與正確性。也發現一件有趣的事，專家因為靠上下文語意來判斷文件歸屬藉以找出重要的特徵詞，所以他所找出來的特徵詞多半是很具關鍵性或是單一字詞為主，但是本研究可以找出專家意想不到的字詞關係。在建立知識地圖階段發現部分特徵詞若與其他特徵詞組合，可能產生出完全不一樣的單位關聯結果，例如，一般提到「機車」可能直接會連想到交通單位，但是其實也可能出現在環保單位的意見投書。過去的研究常受限很多詞彙共同出現在多個單位，而文件分類不準確。因為其只考慮「機車」這個特徵詞出現的頻率，但是這樣很難區分文件所屬的單位。但是本研究可以突破這樣的限制，利用關聯規則可以發現若文件中僅出現「機車」很明確應屬於交通單位，但是文件若同時有提到「廢棄」與「機車」則很明確應屬於環保單位的職責，此實驗結果對於提高分案的準確性，是相當有助益的。

第二階段 文件分派階段

在前一階段的知識地圖經過與專家驗證與確定之後，接下來就是要驗證本研究所提出的文件分派的演算法之效果。這階段的實驗將針對不同的信賴度、特徵詞出現位置給予權重加成及與其他文件分類技術進行實驗，以提升本研究的準確性及有效性。

實驗一：不同信賴度對準確性的影響

本實驗透過不同信賴度所建立的知識地圖來做文件分派的實驗，希望能得到一最佳的信賴度，除能得到較佳的準確性外，也希望所建立的知識地圖中每個節點都具有該單位的代表性，降低自動分案作業所耗資源。本實驗針對所產出關聯規則之信賴度，分別依0.5、0.6、0.7及0.8進行規則篩選，須符合上述標準之規則，才列入知識地圖內。

首先觀察在不同信賴度下各單位的召回率表現，召回率代表正確分案數量與人工分案數量之比率，較高的召回率表示系統自動分案的正確率較高且涵蓋正確率高。圖7橫座標為信賴度由0.5，0.6，0.7，0.7做改變，縱座標為召回率。由圖7所呈現的結果分析，不同的信賴度在不同單位均有不同的表現，環保單位在較低的信賴度下有較佳的表現，大部分單位則是在較高的信賴度有較佳表現，以整體而言以信賴度0.7有較好的表現，其召回率均達0.86以上。

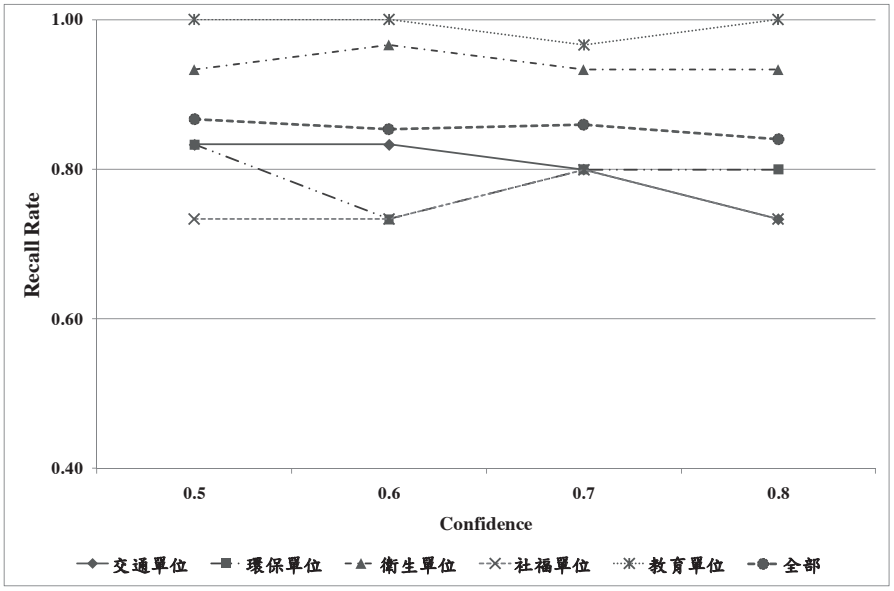

圖7：不同信賴度之召回率表現圖

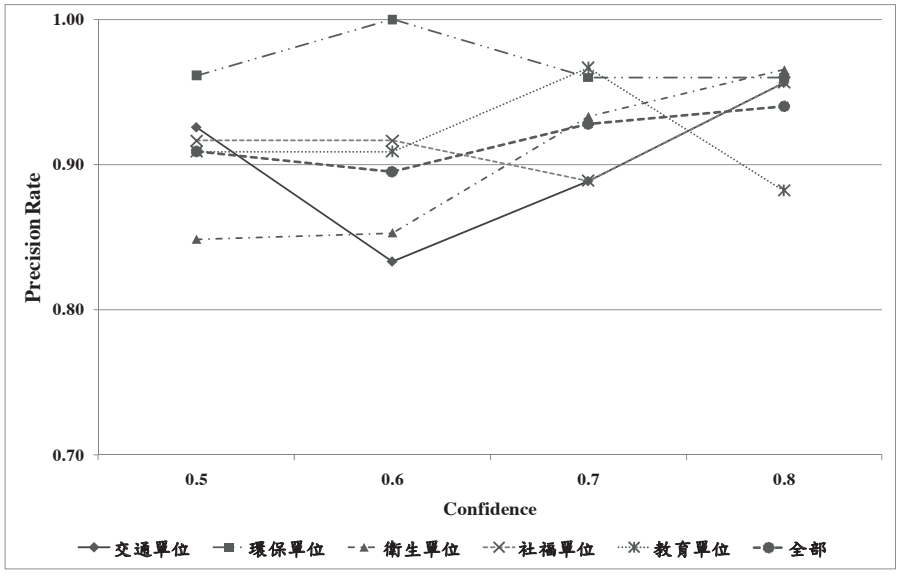

圖8：不同信賴度之準確率表現圖

圖8為在不同信賴度下各單位的準確率表現，準確率代表正確分案數量與系統自動分案數量之比率，較高的準確率表示系統自動分案的正確率較高。同樣地，不同信賴度在不同單位的表現各有高低，但整體而言，以信賴度0.7及0.8有較好的表現，其準確率均達0.92以上。

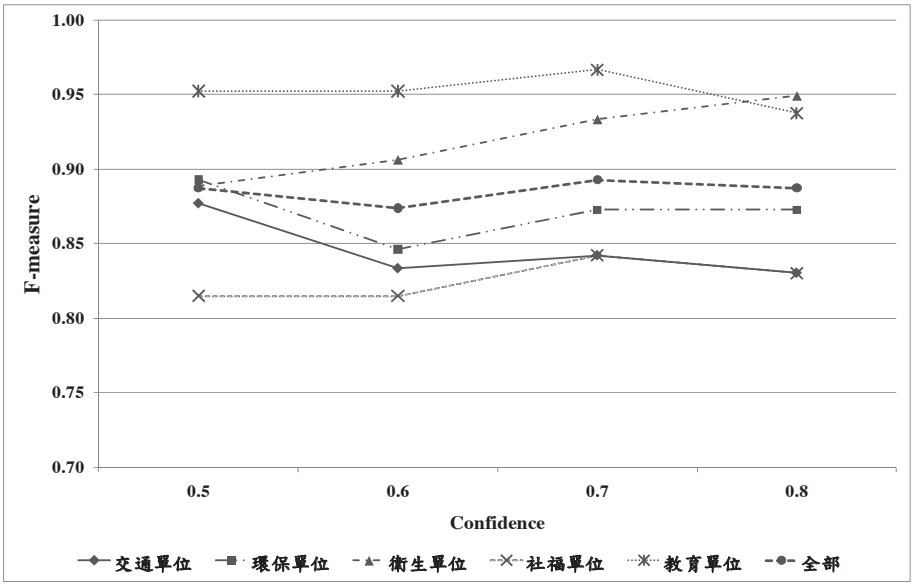

圖9：不同信賴度之F-measure表現圖

圖9為在不同信賴度下各單位的F-measure表現，F-measure為召回率及準確率的綜合評估值，整體而言，以信賴度為0.7時為最佳值。

實驗二：特徵詞出現位置給予不同權重加成方式對準確性的影響

在實驗一得到最小信賴度為0.7時有最佳的準確性，在本實驗即以此信賴度為基礎，進行依特徵詞出現位置給予不同權重之加成，找出最佳的加成方式，來提升分案的正確性。

在分析文件結構時，發現主旨通常可代表該文件的重要概念，若特徵詞出現於主旨，則表示該特徵詞可代表文件的重要性更高，因此本實驗將針對出現於主旨之特徵詞權重分別加成1.2及1.5倍，來觀察加成對分案的成效。另若在文件內已有指出機關或單位名稱時，該文件分案至該機關或單位的機率越高，因此本實驗同時對於單位名稱之特徵詞權重分別以1.2及1.5倍加成。

本實驗對於特徵詞出現位置及是否為機關單位名稱給予不同權重加成方式，交叉計算，分別計算其召回率、準確率及F-measure值，來找出最佳的加成方式，實驗結果如圖10，其中主旨加成表示特徵詞出現於主旨即給予權重加成，而機關加成則表示該特徵詞為機關單位名稱時給予權重加成。

由圖10可發現無論何種加成方式均對提高準確率是有幫助的，單純為主旨或機關名稱加成1.2或1.5倍對於準確性提升並無差別，但若在交叉加成計算時，則以主旨加成1.2倍、機關加成1.5倍及主旨與機關分別加成1.5倍時有較佳的準確性，因此實驗結果得到主旨權重加成至少為1.2倍，而機關名稱權重則需加成1.5倍，機關名稱之特徵詞權重加成影響準確性較位於主旨的特徵詞權重加成為大。

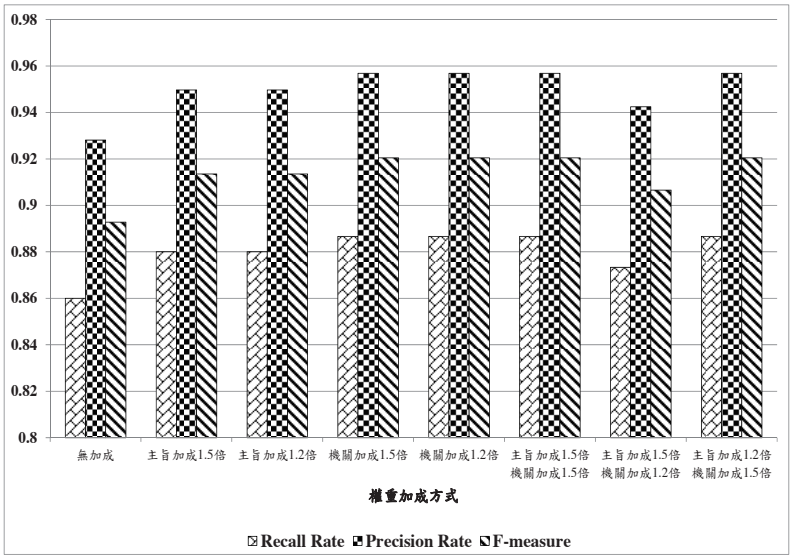

圖10：權重加成方式比較圖

實驗三：與其他分類技術之比較

由文獻探討得知，最常拿來作文件分類的技術為支持向量機（SVM）與關聯規則，因此本實驗將本研究結果與支持向量機分類結果進行比較。經統計在信賴度為0.7時，知識地圖所包含的特徵詞數量約有110個，故本實驗以本研究在信賴度為0.7，未進行任何權重加成方式的分類結果與SVM在110個特徵詞分類結果之召回率、準確率及F-measure進行比較，如圖11。

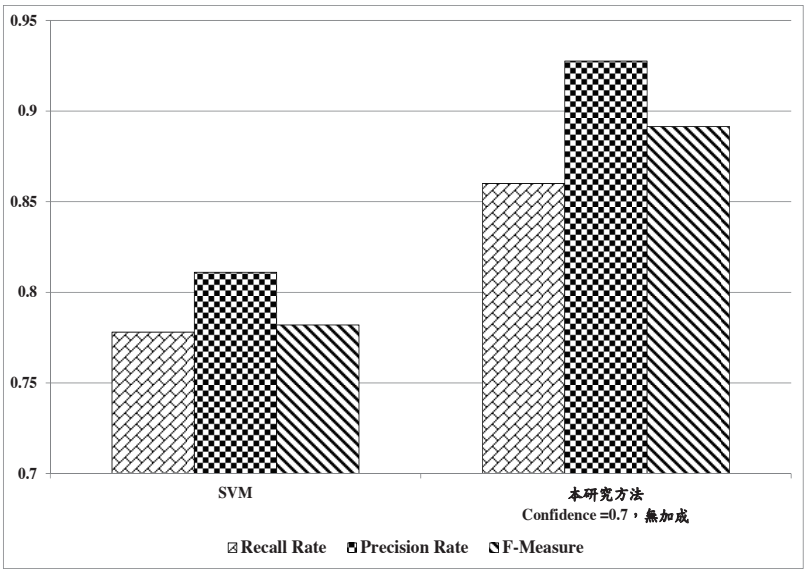

圖11：SVM與本研究結果比較圖

由圖11分析，本研究所提出之方法相較於SVM，無論在哪一個評估值，都有較佳的表現，因此利用關聯規則找出特徵詞與單位之相關性，對於本研究來說是較佳的方式。

伍、結論

本研究提出一個智慧型知識地圖建立的機制，來協助政府部門以自動化的方式來進行分案處理民眾的意見，使民眾意見能迅速分送至正確的處理單位，並迅速予民眾回應，以提升政府部門的處理績效及節省處理分送案件的人力。

本研究所提出的架構主要分兩大部分，首先利用文字探勘的技術找出投書文件中，文件特徵詞與單位之間的關聯性，找出特定樣態的關聯規則，以建立知識地圖。第二階段結合文件結構的概念，在文件的標題與特定的關鍵字上作用權重加成，提升系統自動分案的準確率。本研究利用真實資料來作實驗，實驗設計先驗證知識地圖建立的正確性，接下來透過不同參數的調整，找出最佳的組合作文件分配實驗；最後與目前最常用的分類法SVM作一比較，從實驗結果來看，本研究所提出之方法確實符合預期的目標，同時，驗證了特徵詞與單位關聯性對於自動分案機制是有幫助的。過去，很多文件分類，都致力找出屬於特定類別的字詞，當字詞分別混雜在許多類別時，會降低分類的準確性，本研究所提的知識地圖法，可以解決過去文件分類這部分的問題。而結合文件的概念，可以提高自動分案的準確性。

政府機關組織龐大，且組織內分層負責，各司其職，本研究所提的知識地圖建立方式，亦可以建立階層式的知識地圖。當原始資料所分類的單位階層更細，可以到單位內的承辦科室，則知識地圖可以結合既有的組織階層，建立多階層式的知識地圖。惟因受限於實驗資料的取得與原始資料分類的限制，僅分案至最上層之單位，故僅做到單位級的知識地圖。將來研究延伸，可再蒐集更多資料進行建立更多單位的知識地圖，亦可再進行階層式分案分析，不止建立單位的知識地圖，更可向下延深至單位內部的分層知識地圖，直接分案至單位內的承辦科室。

在研究初期現況訪談時，發現常有一封投書內包含多種類別之意見，為能使用系統自動分案，通常要求民眾在一陳訴信件內只能描述一個類別的意見，如有多類別之意見須分開填寫，造成民眾的不便，若自動分案系統能針對多類別自動分案，將可減少民眾在填寫意見選擇類別之困擾，並提高便利性。另外，未來也希望可增加系統回饋機制，由於時間環境變遷，不同時間點可能會有不同關注焦點，也會有不同的流行語詞，因此文件內的特徵詞會不斷變更、新增，如何讓系統能自動增加特徵詞的回饋機制，這都是可作為未來思考的方向。

陸、致謝

感謝兩位匿名審查委員無私的付出，提供許多的寶貴建議使本論文之內容更臻完美；本研究承蒙國科會專案部分經費贊助（計畫編號：NSC 98-2410-H-031-001與NSC 99-2410-H-031-051），謹致謝忱。

柒、參考文獻

1. 中央研究院中文計算語言研究小組，中文詞知識庫小組，http://godel.iis.sinica.edu.tw/CKIP/ 。
2. 李卓銘，2006，利用關聯式法則將中文文件分類，私立淡江大學資訊工程學系碩士論文。
3. 林昕潔，葉鎮源，陳信源，黃明居，柯皓仁，楊維邦，2008，『使用SVM與詮釋資料之圖書自動分類』，2008資訊科技國際研討會（AIT 2008），朝陽科技大學資訊學院主辦。
4. 林厚誼、蔣岳霖、周世俊，2002，『The design and implementation of Act e-Service Agent Based on FAQ Corpus』，TAAI。
5. 邱登裕、潘雅真，2006，『結合資訊檢索與分群演算法建構知識地圖』，資訊管理學報，第十三卷．第S期137～160頁。
6. 侯建良、楊綠淵，2004，『以文件關聯性為基礎之企業知識客服管理模式』，資訊管理學報，第十一卷．第四期：205～228頁。
7. 范錚強，1999，『先進國家網路申辦之推動現況』，研考雙月刊，第二十三卷．第一期：15～21頁。
8. 許中川、陳景揆，2001，『探勘中文新聞文件』，資訊管理學報，第七卷．第二期：103～122頁。
9. 許昌偉，2005，『資料探勘應用於文件分類技術之研究-以網路新聞分類為例』，私立銘傳大學資訊管理學系碩士論文。
10. 郭瓊蓉，2005，文件分類於電子化政府之應用：以政府機關市長信箱民眾陳情案件為例，國立中山大學資訊管理學系研究所碩士論文。
11. 陳育民，2008，利用關聯式法則改善文件分類準確度-結合其他分類器，私立淡江大學資訊工程學系碩士論文。
12. 陳良駒，陳日鑫，2009，『以關聯規則與共詞分析探討軍事新聞文件群集效果』，第20屆國際資訊管理學術研討會，中華民國資訊管理學會主辦。
13. 曾元顯，2002，『文件主題自動分類成效因素探討』，中國圖書館學會會報，第68期：62～83頁。
14. 黃國禎、朱蕙君、曾秋蓉、黃國豪、黃繼緯、林農堯，2007，『具自我調適功能之線上課程問題自動回覆系統』，電子商務學報，第九卷．第三期：599～624頁。
15. 魏莉斐，2006，有趣性度量結合詞彙權重之文件分類研究，國立臺灣科技大學資訊管理系碩士論文。
16. Agrawal, R., Imieli ski, T., and Swami, A.“Mining association rules between sets of items in large databases,”*ACM SIGMOD Record* (22:2) 1993, pp 207-216.
17. Alavi, M., and Leidner, D.“Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues,”*MIS quarterly*, 2001, pp.107-136.

18. Allee, V. "The art and practice of being a revolutionary," *Journal of Knowledge Management* (3), 1999, pp.121-131.
19. Davenport, T. H., and Prusak, L., *Working knowledge: How organizations manage what they know*, Boston Harvard Business School Press, 1998.
20. Feldman, R., Klosgen, W., Ben-Yehuda, Y., Kedar, G., and Reznikov, V. "Pattern based browsing in document collections," *Principles of data mining and knowledge discovery*, 1997, pp 112-122.
21. Haddad, H., Chevallet, J., and Bruandet, M. "Relations between terms discovered by association rules," 2000.
22. Hao, P., Chiang, J., and Tu, Y. "Hierarchically SVM classification based on support vector clustering method and its application to document categorization," *Expert Systems with Applications* (33:3), 2007, pp.627-635.
23. Huber, G. "Organizational learning: The contributing processes and the literatures," *Organization science* (2:1), 1991, pp.88-115.
24. Joachims, T. "Text categorization with support vector machines: Learning with many relevant features," *Machine Learning: ECML-98*, 1998, pp.137-142.
25. Kang, I., Park, Y., and Kim, Y. "A framework for designing a workflow-based knowledge map," *Business Process Management Journal* (9:3), 2003, pp 281-294.
26. Kim, S., Han, K., Rim, H., and Myaeng, S. "Some effective techniques for naive bayes text classification," *IEEE Transactions on Knowledge and Data Engineering*, 2006, pp.1457-1466.
27. Lin, F., and Hsueh, C. "Knowledge map creation and maintenance for virtual communities of practice," *Information Processing & Management* (42:2), 2006, pp.551-568.
28. McCagg, E., and Dansereau, D. "A convergent paradigm for examining knowledge mapping as a learning strategy," *The Journal of Educational Research* (84:6), 1991, pp.317-324.
29. Nonaka, I. "A dynamic theory of organizational knowledge creation," *Organization science* (5:1), 1994, pp.14-37.
30. Salton, G., and McGill, M. *Introduction to modern information retrieval*, McGraw-Hill, New York, 1983.
31. Zack, M. "Developing a knowledge strategy," *The strategic management of intellectual capital and organizational knowledge*, 2002, pp.255-276.