應用模糊資訊檢索對文件做多重分類之研究

蔡嘉嘉 (Chia-Chia Tsai)、曾守正 (Frank S.C. Tseng)

國立高雄第一科技大學 資訊管理系
National Kaohsiung First University of Science and Technology
Department of Information Management
imfrank@ccms.nkfust.edu.tw

摘　要

網際網路的普及造就了文件資料大量流通，所以對於日積月累的文件資料，如果不加以分類整理，將會為人們帶來相當多的困擾。在真實世界裡，一份文件的內容可能涉及多個不同的議題，或是各事先定義的類別之間並不完全獨立，使得將每份文件只歸類到單一特定類別的作法，並不見得合理。在本論文中，我們將以模糊集合理論 (Fuzzy Set Theory) 為基礎，透過「模糊資訊檢索分類」 (Fuzzy Information Retrieval Categorization) 以文件做為分析目標，將每份文件進行合理的多重分類。將文件同時歸屬於多類，不僅可以提高文件檢索的效率，後續更可以進一步建立文件倉儲 (Document Warehouse)，以便對該文件進行文件探勘 (Text Mining) 做準備。對於分類的結果，本研究將透過分類方法之效能評估，來驗證其正確率 (Precision Rate) 和回現率 (Recall Rate) 相關指標。

關鍵字：多重分類問題、模糊資訊檢索分類、正確率、回現率

一、緒論

在現代科技工作中，對文件內容做資訊擷取 (Information Retrieval) 是一項很重要的工作。如進行某項課題的研究，查詢某項專利內容及時間、範圍等。而文件檢索 (Document Retrieval) 的工作量是浩瀚的，系統必須在合理的時間內滿足使用者的需求，因此現代文件檢索都要借助計算機來建立文件檢索系統。但是，實際應用中所需檢索的訊息也就是使用者的資訊需求，往往具有一定的模糊性 (Fuzziness)。

過去的決策規則是建立於二值邏輯 (Two-Valued Logic) 之上，一個描述字 (Descriptor；或稱索引字, Index-Term) 只能「屬於」或「不屬於」一篇文件的描述。雖然傳統二值邏輯，對於理解和實作而言，皆顯得相當直觀且容易。但是，真實世界所存在的多種模糊性 (Fuzziness)，使它產生一些明顯的缺失：

1. 自然語言文件中，常有「非常」、「有些」等的模糊語意量化詞 (Fuzzy Linguistic Quantifier)。因此，描述字對於文件的描述程度應有所差異才是。
2. 無法對於查詢搜尋樣型 (Query Search Pattern) 中的描述字，以不同的重要性加以處理。如：「CASE (有點重要) and Hypertext (非常重要)」。

由於文件之內容常跨多個主題，使得文件的分類含有不確定性，分類結果常因人而異，所以本研究利用模糊理論之演算，來規範文件分類的不確定性或大約的推理，方便人們的資訊處理，減少人為的差異。本論文以模糊集合理論 (Fuzzy Set Theory)[5][20][19] 為基礎，運用「模糊資訊檢索分類」(Fuzzy Information Retrieval Categorization) [21] 以文件做為分析目標，來將每份文件進行合理的多重分類。將文件同時歸屬於多類，不僅可以提高文件檢索的效率，後續更可以進一步建立文件倉儲 (Document Warehouse)[13]，以便對該文件進行文件探勘 (Text Mining) 做準備。對於分類的結果，本研究將透過分類方法之效能評估，來驗證其正確率 (Precision Rate) 和回現率 (Recall Rate) 相關指標。

本論文的結構組織如下：第二節對相關研究做一番說明；第三節說明如何運用「模糊資訊檢索分類」來對文件做多重分類的問題與解法；第四節說明「模糊資訊檢索分類」的應用案例；第五節對分類方法之效能評估做一番討論；最後總結並討論未來可能的研究方向。

二、相關研究

2.1 模糊集合理論

根據定義，一個模糊集合可視為一個以模糊界限來分群所成的集合。假設存在一定義域 (Universe of Discourse) U，在此 $U=\{u_1, u_2, \ldots, u_n\}$，則一個針對 U 之模糊集合 A 將如定義 2.1 所述，其中μ_A可視為對應模糊集合 A 的歸屬函數 (Membership Function)，而 $\mu_A(u_i)$ 可視為 u_i ($u_i \in U$) 隸屬於模糊集合 A 的歸屬程度 (Membership Degree)，其值介於 [0, 1] 之間。

定義 2.1：針對一個定義域 (Universe of Discourse) $U = \{u_1, u_2, \ldots, u_n\}$而言，一個模糊集合 (Fuzzy Set) A 定義為

$$Fuzzy\ Set\ A = \{(u_i, \mu_A(u_i)) | u_i \in U\}$$

假設有兩個針對定義域 U 之模糊集合 A 和 B (在此 $U = \{u_1, u_2, \ldots, u_n\}$) 分別表示 *Fuzzy Set* $A = \{(u_i, \mu_A(u_i)) | u_i \in U\}$和 *Fuzzy Set* $B = \{(u_i, \mu_B(u_i)) | u_i \in U\}$。則模糊集合 A 和 B 的聯集運算和交集運算之定義分別如定義 2.2 及定義 2.3 所示。

定義 2.2：針對一個定義域 (Universe of Discourse) $U = \{u_1, u_2, \ldots, u_n\}$而言，模糊集合 $A = \{(u_i, \mu_A(u_i)) | u_i \in U\}$和模糊集合 $B = \{(u_i, \mu_B(u_i)) | u_i \in U\}$的聯集運算定義為

$$A \cup B = \left\{ \left(u_i, \mu_{A \cup B}(u_i) \right) \middle| \mu_{A \cup B}(u_i) = Max\left(\mu_A(u_i), \mu_B(u_i) \right), u_i \in U \right\}$$

定義 2.3：針對一個定義域 (Universe of Discourse) $U=\{u_1, u_2, \ldots, u_n\}$而言，模糊集合 $A = \{(u_i, \mu_A(u_i))| u_i \in U \}$和模糊集合 $B = \{(u_i, \mu_B(u_i))| u_i \in U \}$的交集運算定義為

$$A \cap B = \left\{ \left(u_i, \mu_{A \cap B}(u_i) \right) \middle| \mu_{A \cap B}(u_i) = Min\left(\mu_A(u_i), \mu_B(u_i) \right), u_i \in U \right\}$$

最後，我們介紹模糊集合理論中α-截集 (α-*level set or* α-*cut*) 的概念[20][21]。α-截集是在模糊集合與普通集合相互轉化中的一個重要概念，在模糊決策中也經常用到。針對一個模糊集合 A 的α-截集 A_α 如定義 2.4 所示。

定義 2.4：針對一個定義域 (Universe of Discourse) $U=\{u_1, u_2, \ldots, u_n\}$而言，一個模糊集合 $A = \{(u_i, \mu_A(u_i))| u_i \in U \}$的$\alpha$-截集 A_α定義為

$$A_\alpha = \left\{ u_i \middle| \mu_A(u_i) \geq \alpha, u_i \in U \right\} \quad \alpha \in [0, 1]$$

普通集合 A_α是對原來的模糊集合 A 的歸屬度先確定一個限定值α $(0 \leq \alpha \leq 1)$ 之後，再把歸屬度 $\mu_A(u_i) \geq \alpha$ 的元素挑選出來而得。在 A_α的定義中要注意兩點：

1. A 是模糊子集，但 A_α是普通集合；
2. A_α的直觀意義是 u_i對 A 的歸屬度達到或超過α的就算 u_i是 A 的元素。

而α-截集性質為[21]：

1. $(A \cup B)_\alpha = A_\alpha \cup B_\alpha$
2. $(A \cap B)_\alpha = A_\alpha \cap B_\alpha$
3. 若 α_1、 $\alpha_2 \in [0, 1]$，且 $\alpha_1 \leq \alpha_2$，則 $A_{\alpha_1} \supseteq A_{\alpha_2}$。

由上述第 3 點可知：α-截值越低，A_α越大；α-截值越高，A_α越小。當$\alpha = 1$ 時，A_α最小。

所以，為了迎合使用者的主觀意識及提供多重分類的彈性，模糊訊息檢索分類引用了α-截集的觀念可做為分類過程中的一個模糊項臨界值 (Threshold)。這個臨界值用以滿足限制條件的最低程度。

2.2 分類的觀念和回顧

2.2.1 分類的定義及目的

分類就是將資料 (Data) 或物件 (Object) 做某種方式的歸類，其主要作用在於將性質相近的資料或物件，放在同一個地方，使得人們要從眾多資料中查詢到所需的資料時，能夠更有效率且迅速地取得。尤其在此資訊爆炸的時代，如果沒有做分類的話，要從龐大的資料庫內尋找資料將是使用者的一大困擾，因此分類的重要性更形突顯。

2.2.2 分類的方法及型態

根據 [7][8] 的研究，分類的型態有很多種，而 Lance and Williams (1967)[8] 則將分類的問題型態以一樹狀結構圖表示，如圖 2.1 所示。圖 2.1 的樹狀圖定義了不同種類的分類問題型態，圖中的每一個節點分別說明如下：

1. 互斥(Exclusive) 與非互斥 (Non-Exclusive) (或稱重疊 (Overlapping))：前者是指一資料只能被分到一類別中。後者是指容許同一資料被分到多個類別。例如：在圖書分類中，一本名為「自然語言與資訊檢索」的書，可以被分到「自然語言」類，也可以被分到「資訊檢索」類。
2. 監督式 (Supervised) (或稱非固有 (Extrinsic)) 及非監督式 (Unsupervised) (或稱固有 (Intrinsic))：前者是指事先設定好類別，然後將欲分類的資料歸到與之相似程度較高的類別。例如：大家所知道的圖書分類法，這種方法稱為分類 (Classification)。後者是指事先不設定有幾類，只要欲分類的資料被視為相似，就被分為同一類。
3. 階層式 (Hierarchical) 與非階層式 (Non-Hierarchical) (或稱平面式 (Partitional))：前者是指分類具有階層性。例如：“電腦通訊” 這個類別又包含了 “程式設計”、 “電腦遊戲” 等類別，而 “電腦遊戲” 類別又包含了 “遊戲軟體”、“遊戲週邊” 等類別。以分群 (Clustering) 的觀點來看，階層式的分類有一個很大的好處，可以根據其樹狀結構得知資料是在什麼階段被合併 (Merge) (從樹葉開始往上看到樹根) 或被分開 (從樹根開始往下看到樹葉)。後者是指分類不具階層性，即平面式的分類。例如：將人劃分為男人與女人，即可以視為平面式的分類。以某個角度來看，可以將階層式的分類視為平面式分類的特殊情形。

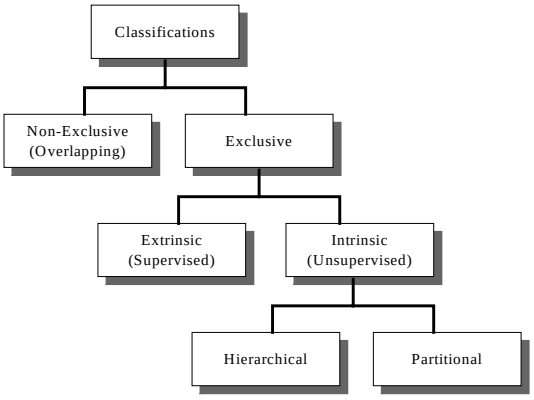

圖 2.1：分類型態之樹狀圖

本論文所採用的分類方法：是融合上述之非互斥 (亦即做重覆分類)、監督式 (亦即本身就有類別) 與階層式所進行的研究。

2.2.3 分類與分群

分類 (Classification) 與分群 (Clustering) 的區別說明如下：

1. 分類 (Classification)：在分類的做法上，首先要定出類別並試圖用某些特性來描繪此類別。這些特性通常必須具備辨識能力

(Discrimination Value)，亦即它常常在某一類中出現，而幾乎不會在其他類別中出現。如何決定類別，也許非常主觀，直覺認為要這樣分，也可能是先瀏覽過欲分類的資料，然後訂出類別。另外，也可以先做分類之後，利用因素分析 (Factor Analysis) 重新決定類別[6]。

2. 分群 (Clustering)：分群即是將一些具有共同特性的文件群集在一起，目標是將相關的文件組合在一起[1]。最直覺的做法就是將資料依序讀入，與現有的類別比較相似度 (Similarity)，相似度如果大於某一定程度，即將此文件歸於此類，否則就自成一類，直到所有的資料都已分類完畢。當然，我們必須把原始資料用某種方法做適當地表示，例如：選定一些特性 (Property) 或屬性 (Attribute) 來描述這些資料，如此也才有衡量相似性的依據。分群在資訊檢索方面佔很重要的地位，也有許多這方面的研究成果發表[9][4]。對一個檢索系統而言，若此系統將資料事先做分群，以樹狀結構來儲存，當一個查詢來時，就不必做線性搜尋 (Linear Search) 而只要做樹狀搜尋 (Tree Search)，可以省去許多搜尋時間。

2.3 模糊資訊檢索分類系統架構

2.3.1 模糊文件檢索架構

所謂的「文件」(Documents)，指的是一個資訊體，其內容可以包含文字 (Text)、圖形 (Graphics)、動畫 (Animation)、語音 (Voice)、視訊 (Video) 等多媒體。然而，由於不同的媒體有不同的特性，因此在本論文中所探討的仍是以文字為主體的文件。

文件檢索的流程大致可以以圖 2.3 表示。其中，查詢 (Query) 是使用者對系統所下的命令 (Command)，至於其命令的格式則因系統之不同而有差異，通常可包含搜尋樣型 (Search Pattern)、文件資料庫 (Document Base) 之名稱、符合文件數之上限以及輸出裝置 (Output Device) 等。所謂的「搜尋樣型」，指的是由描述字和邏輯運算子「NOT」、「AND」及「OR」所組成的片語，這是一個查詢的最基本部份，不可或缺。而其他部份則選擇性地可有可無。

查詢在經過語法、語意分析後，系統便開始從文件資料庫中尋找符合搜尋樣型的文件，最後回應給使用者一份符合文件的清單。文件資料庫中的文件，由於數量龐大，一般並不適合做全文檢索 (Full-Text Retrieval)。因此，原始文件必須經過一個抽象化的過程，而以簡化的文件描述 (Document Description) 代表原始文件。在抽象化的過程中，原始文件通常是以對其內容具有描述性或代表性的描述字 (Descriptor；或稱關鍵詞，Keyword) 來組成它的文件描述，如圖 2.4 所示。

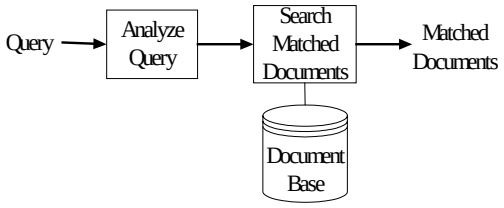

圖 2.3：文件檢索的概略流程

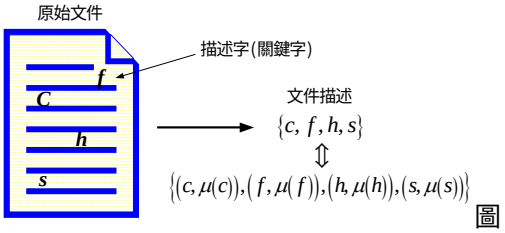

圖 2.4：文件抽象化成文件描述

至於模糊文件檢索的第一步，就是要從改進文件描述著手。文件描述中的任一描述字，對於文件的描述程度，應該根據文件之實際內容而有所不同，而不是一律視為平等，如圖 2.4 中的文件描述應改成{(*c*, $\mu(c)$), (*f*, $\mu(f)$), (*h*, $\mu(h)$), (*s*, $\mu(s)$)}，其中$\mu(c)$為 *c* 對該文件的描述程度，其餘類推，也就是說文件描述將是一個描述字的模糊集合。

在過去，關於模糊檢索的研究，為了解決傳統二值邏輯法的缺失，多是屬於數學模型的研究[10][17][11][12][14][15]；這些研究提供了一些幫助，使我們得以數學的觀點了解模糊檢索；不過就實務觀點而言，有些仍難以實作。以下子節我們將以模糊文件檢索為基礎，提出應用模糊資訊檢索分類法來進行文件多重分類的系統架構，如圖 2.5 所示。

1.3.2 模糊資訊檢索分類架構

模糊資訊檢索分類架構是根植於模糊文件檢索的基礎上，它按照查詢訊息特性，應用模糊集合理論 (Fuzzy Set Theory) 和方法來來預先對文件進行分類，從而提高文件檢索的效率。模糊資訊檢索分類法的步驟為：

1. 首先，建立關鍵詞-文件模糊矩陣 (Keyword-Document Fuzzy Matrix) 及關鍵詞-類別模糊矩陣 (Keyword-Category Fuzzy Matrix)；
2. 其次，建立文件-類別模糊矩陣 (Document-Category Fuzzy Matrix)，並利用該矩陣建立文件-多重類別模糊矩陣 (Document-Multi-Category Fuzzy Matrix)；
3. 第三步則根據 *α-cut* 來判別文件可被多重歸屬的類別。而模糊文件檢索的查詢句改為新加入文件，原為檢索的模型就可成為分類的模型。圖 2.5 為本研究模糊資訊檢索分類的系統架構。而在模糊資訊檢索分類系統架構中之模糊分類機制 (Fuzzy Classifier Machine) 部份則是本研究極欲探討之中心主題，如圖 2.5 陰影部份所示。

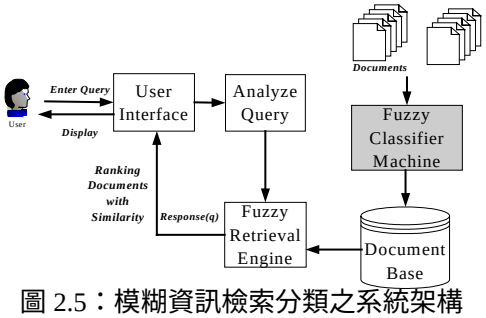

圖 2.5：模糊資訊檢索分類之系統架構

本研究運用已預先完成的人工分類概念，延伸模糊文件檢索模型來進行文件模糊多重分類。利用已分類資料當作學習範本，並且假設人工分類類別都是正確的。進行模糊多重分類時，系統會依照新文件的描述項 (關鍵詞)、描述程度和各個訓練類別所得的類別分類知識進行比對，相似度較高的類別為文件的類別。以下章節將說明本研究模糊資訊檢索分類系統架構之詳細步驟以及如何運用於文件多重分類的問題與解法。

三、文件多重分類問題

本研究利用模糊資訊檢索分類法來將一個查詢和文件內容之間的匹配過程，按照查詢指令特點，並運用模糊集合理論和方法來對文件資料進行多重分類，從而提高文件檢索的效率。而在模糊分類好之後，後續更可以進一步建立文件倉儲 (Document Warehouse)，以便對該文件進行文件探勘 (Text Mining) 做準備，探勘出文件所蘊含的訊息。圖 3.1 為本研究模糊資訊檢索分類架構中模糊分類機制 (Fuzzy Classifier Machine) 之運作流程，也是本研究欲探討之中心主題。

3.1 模糊資訊檢索分類步驟

模糊資訊檢索分類系統可分成前端類別關鍵字選取與後端模糊多重分類兩部份架構，參考圖 3.2。下面各小節將分述此模糊分類機制之方法與步驟。

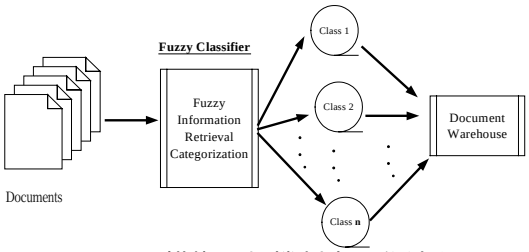

圖 3.1：模糊分類機制之運作流程

3.2 關鍵詞的選取

從文件自動擷取片語知識，包括專有名詞，人名，地名，lexical templates 等，對多數資訊檢索應用，如文件摘要、分類、過濾等都是相當重要的研究課題。此外，對中文等東方語言而言，因為書寫時詞與詞間缺乏邊界標示，高效率片語知識擷取技術需求更加殷切。由於任一特定文件集合其重要關鍵詞多不會收錄在辭典中，因此發展中文關鍵詞抽取很難藉助辭典，必須有其他突破性的作法。

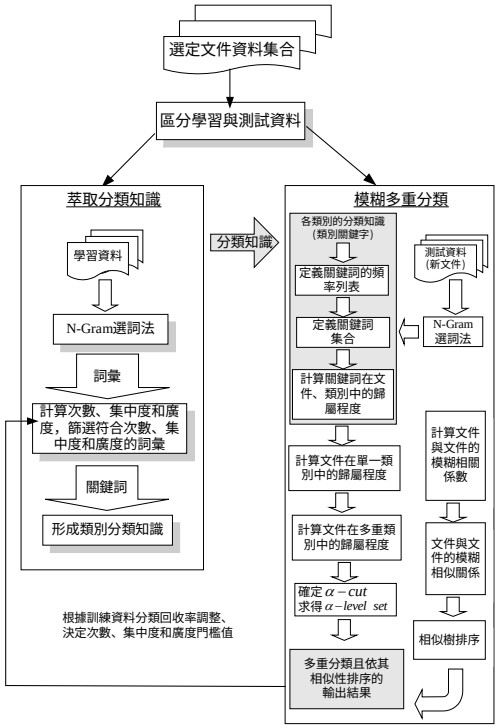

圖 3.2：文件多重分類步驟

本研究擬採用 *N-Gram* 選詞方法，這項技術已經應用在文件關鍵詞判定 (Keyword Identification) 以及文件分類 (Document Classification)[24]上。因為 *N-Gram* 選詞法是完全不需要辭典且全部依靠語言的統計結果來決定的統計檢測方法，透過檢測文件集合中所有字串前後相依程度，語意完整度及在集合內之重要性，多數中文關鍵詞或片語，特別是人名以及專有術語都可有效擷取出，並且擷取出的術語無長度限制。舉例來說，假設一個有分類價值的專有名詞 AB，其中 A、B 為中文字，因為電子詞典沒有這個詞，一被斷詞系統拆成 A 與 B 兩個單字詞以後，原本可以是有用的關鍵詞，可能因此而不被視為關鍵詞。再假設另一個有分類價值的專有名詞 ABCD，其中 A、B、C、D 為中文字，但是電子詞典中沒有這個詞，其中 AB 與 CD 因為分佈很平均以致不能成為關鍵詞，然而 BC 經過 2-Gram 篩選後成為關鍵詞，目的還是達到了，並沒有漏掉有用的訊息。

而在統計句子內字與字的相鄰機率時，若是以每兩個字為一個單位切開，就叫做 *Bi-Gram* (*2-Gram*)，每三個字為一組，就叫做 *Tri-Gram* (*3-Gram*)，其餘類推。*N-Gram* 就是文件中任意 *N* 個連續字元，如「中國社會」此一字串，當 *N* 為 2 時將可產生「中國」、「國社」、「社會」三個索引詞。如此可排除或降低「字元法」中類似「中國」與「國中」的字串順序問題，也可省去「詞彙法」中維護詞庫的煩惱。在 *Bi-Gram* (*2-Gram*) 執行步驟中：先統計大量的語料，統計句子內字與字的相鄰機率。在斷詞的過程中，找出最大的兩字相鄰機率值，視此兩字為一詞，並且切割句子為前後兩部份，繼續此方法，直到任一相鄰兩字的機率值

小於一個設定值。若僅採用此一 *Bi-Gram* (*2-Gram*) 選詞方法其特點是不需使用辭典，也不需用到任何法則的簡單方法；缺點是僅能找出兩字詞和單字詞，且正確率不是很高。

本研究擬利用 Bi-Gram 做基礎，繼續找出 N-Gram，以進行關鍵詞的選取，不僅語料充足，且可以擷取的詞彙知識比傳統豐富，包括任意長度以及不同語言層次的資料串 (Data Stream) 機率，如字串，詞類串，語意串等，因此關鍵詞的判定將可以有效實施。而對於分類系統而言，一個具有分類價值的 N-Gram 關鍵詞應該滿足下列三個條件[23]：

1. 次數要夠：所選出的字詞並非都是有意義的詞，通常不具意義的字詞出現的次數不會多，如果定一界限值 (Threshold)，去掉出現次數低於此界限值者，則那些無意義的字詞大都會被摒除在外。根據[25]實驗所得的結果，界限值定為 5 時得到最高的回現率。
2. 集中度 (Conformity)：一個有分類價值的關鍵詞，應該要集中出現在某一類或某幾類中，而不是平均出現在各類中。因此，第二步利用 Shannon 所提出熵 (Entropy) 的公式來做篩選，以符合集中度的要求。對於一個 N-Gram T_i，衡量 T_i 的集中度 Entropy 值公式如下：

$$H_i = -\sum_{j=1}^{k} p_{ij} \log p_{ij} \quad , j \text{ 代表類別}$$

其中，$p_{ij} = \dfrac{d_{ij}}{\sum_{j=1}^{k} d_{ij}}$ ，d_{ij} 表示類別C_j中，出現T_i的文件數。

當平均分布在各類時，所得到的集中度值最大 $H_j = logN_{class}$ (N_{class} 代表類別數)，相對的，若只出現在一類中，則 $H_j = 1log1 = 0$，因此每一詞彙的 H 值應介於 0 (最集中) 與 $logN_{class}$ (最分散) 之間，至於其臨界值則視訓練資料之不同，得經過多次的試驗來決定，沒有一標準可遵循。根據[24]實驗中，Entropy 界限值訂定為 *log*2 (= -1/2*log*1/2-1/2*log*1/2)，大於界限值之 N-Gram 則予以捨棄。而實驗結果發現將 Entropy 界限值定為 *log*2 是假設一個平均分佈在兩類中的詞彙能讓它通過，但是後來發現關鍵詞跨類的現象很普遍，所以，對單一分類而言界限值定為 *log*2 可能太過於寬鬆。但對於本研究屬於重複分類的問題，Entropy 的界限值定為 *log*2 就可以了。
3. 廣度 (Uniformity)：在某類中出現頻率高的關鍵詞，如果是出現在此類中許多篇文件中，則它愈具有分類價值；反之，若只集中在此類的某一、兩篇文件中，則原因可能只是一突發事件，或是特定撰稿者的特殊寫作風格所致，所以此關鍵詞較無分類價值。因此訂定一個公式來篩選 N-Gram 關鍵詞，以符合廣度的要求[23]。對於一個 N-Gram T_i，衡量 T_i 的廣度公式如下：

$$Value(T_i) = \max_j \left(\frac{d_{ij}}{t_{ij}} \times \frac{d_{ij}}{\sum_{j=1}^{k} d_{ij}} \right)$$

其中，d_{ij} 表示類別C_j 出現 T_i 的文件數，
t_{ij} 表示 T_i 出現在類別 C_j 的次數。

顯然 *Value* 的值愈大，此 N-Gram 愈具有分類的價值。假設有一個 N-Gram，它在 A 類出現 12 次且分佈在 6 篇文件中，以及在 B 類出現一次且分佈在一篇文件中，則此 N-Gram 的廣度依照公式計算如下：

$$Value = max\left(\frac{6}{12}\times\frac{6}{7}, \frac{1}{1}\times\frac{1}{7}\right) = max(0.43, 0.14) = 0.43$$

其臨界值與集中度一樣也是端視訓練資料不同，得經過多次的試驗才決定。本研究根據[24]來訂定廣度界限值為 0.2，小於界限值 0.2 則予以捨棄。

3.3 定義關鍵詞的頻率列表

假設一個文件檔案有 n 個文件 D_1，D_2，…，D_n 組成，每個文件 D_r 由 m 個描述項 (關鍵詞) K_1，K_2，…，K_m 來描述。我們採隨機選擇的方式來挑選文件 (其一般性越高越好) 並統計每一個描述字在文件中出現的平均次數，最後將所有的描述字和其出現的次數記錄成一個頻率列表 (Frequency List)，以 *FL* 表示。

定義 3.1：

$$FL = \left\{ \left(fk_i, N_{FL}\left(fk_i \right) \right) \middle| i = 1, ..., m_f \right\}$$

fk_i:頻率列表中第i個描述字

$N_{FL}(fk_i)$:描述字 fk_i 在 N 封文件中出現的平均次數

m_f : 頻率列表中描述字總數

在各事先定義類別訓練組文件的內容中，所篩選出來的各類別的關鍵字，可聯集成一個關鍵字集合 (Keyword Set)，以 KS 表示。

定義 3.2：

$$KS = \{k_1, k_2, \ldots, k_m\}, \quad k_i \in KS \quad (i = 1, 2, \ldots, m)$$

3.4 定義模糊關鍵詞集

此一步驟我們利用模糊理論中的歸屬函數$\mu_{Dr}(k_i)$和$\mu_{Cm}(k_i)$來判斷每一個關鍵詞 k_i 在文件和各分類類別中重要性的程度。我們採用一種模糊關鍵詞集 (Fuzzy Keyword Set, FKS)[22]的方法來作文件 D_r 的表示，它是一種用模糊數值來表達關鍵詞在文件中的重要程度。其計算方式如下：

定義 3.3：

$$FKS(D_r) = \left\{ \left(k_1, \mu_{D_r}(k_1)\right), \left(k_2, \mu_{D_r}(k_2)\right), ..., \left(k_m, \mu_{D_r}(k_m)\right) \right\}$$

其中

$$\mu_{Dr}(k_i) = \frac{n_{D_r}(k_i)}{n_{D_r}(k_i) + n_{FL}(k_i)}; \quad r = 1, 2, ..., n$$

$n_{D_r}(k_i)$:關鍵字k_i 在該文件 D_r 中出現的次數

$n_{FL}(k_i)$:關鍵字k_i 在頻率列表中出現的次數

於是 n 個文件和 m 個描述項的關係可用一個 $n \times m$ 階矩陣 M_1 來表示，為公式 3.1。

$$M_1 = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & \cdots & k_m \end{array} \\ \begin{bmatrix} \mu_{D_1}(k_1) & \mu_{D_1}(k_2) & \cdots & \mu_{D_1}(k_m) \\ \mu_{D_2}(k_1) & \mu_{D_2}(k_2) & \cdots & \mu_{D_2}(k_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{D_n}(k_1) & \mu_{D_n}(k_2) & \cdots & \mu_{D_n}(k_m) \end{bmatrix}_{n\times m} \end{array} \quad \text{(公式 3.1)}$$

同理，我們運用 *FKS* 的方法來作類別 C_m 的表示，以模糊數值來表達關鍵詞在類別中的重要程度。其計算方式如下：

定義 3.4：

$$FKS(C_m) = \left\{\left(k_1, \mu_{C_m}(k_1)\right), \left(k_2, \mu_{C_m}(k_2)\right), ..., \left(k_m, \mu_{C_m}(k_m)\right)\right\}$$

其中

$$\mu_{C_m}(k_i) = \frac{n_{C_m}(k_i)}{n_{C_m}(k_i) + n_{FL}(k_i)}; \quad m = 1, 2, ..., k$$

$n_{C_m}(k_i)$：關鍵字 k_i 在該類別 C_m 中出現的次數

$n_{FL}(k_i)$：關鍵字 k_i 在頻率列表中出現的次數

於是 k 個類別和 m 個描述項的關係可用一個 $k \times m$ 階矩陣 M_2 來表示，為公式 3.2。

$$M_2 = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & \cdots & k_m \end{array} \\ \begin{bmatrix} \mu_{C_1}(k_1) & \mu_{C_1}(k_2) & \cdots & \mu_{C_1}(k_m) \\ \mu_{C_2}(k_1) & \mu_{C_2}(k_2) & \cdots & \mu_{C_2}(k_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{C_k}(k_1) & \mu_{C_k}(k_2) & \cdots & \mu_{C_k}(k_m) \end{bmatrix}_{k\times m} \end{array} \quad \text{(公式 3.2)}$$

完成此一步驟後，從訓練組文件建立分類模型便已完成。而接下來便是要利用分類模型將未分類的文件加以分類。

3.5 確定文件所屬類別之歸屬函數

此步驟主要確定未分類文件所屬類別之歸屬函數並判斷未分類文件所應該被歸屬的類別。首先利用 N-Gram 選詞法將未分類文件做關鍵字篩選，並根據定義 3.2 中所求出的各類別關鍵字取聯集，定義出一組關鍵字集合。

假設關鍵字集合為{gk_1, gk_2, …, gk_m}，則利用定義 3.3 和 3.4 中之公式分別求出關鍵字 gk_i 在未分類文件 D_r 和事先定義類別 C_m 的 *FKS* 表示法，可得到定義 3.5 和 3.6，其中 $\mu_{Dr}(gk_i)$ 和 $\mu_{Cm}(gk_i)$ 分別表示關鍵詞 gk_i 在未分類文件 D_r 及事先定義類別 C_m 中的重要程度。

定義 3.5：

$$FKS(D_r) = \left\{\left(gk_1, \mu_{D_r}(gk_1)\right), \left(gk_2, \mu_{D_r}(gk_2)\right), ..., \left(gk_m, \mu_{D_r}(gk_m)\right)\right\}$$

其中

$$\mu_{Dr}(gk_i) = \frac{n_{D_r}(gk_i)}{n_{D_r}(gk_i) + n_{FL}(gk_i)}; \quad r = 1, 2, ..., n$$

$n_{D_r}(gk_i)$：關鍵字 gk_i 在該文件 D_r 中出現的次數

$n_{FL}(gk_i)$：關鍵字 gk_i 在頻率列表中出現的次數

定義 3.6：

$$FKS(C_m) = \left\{\left(gk_1, \mu_{C_m}(gk_1)\right), \left(gk_2, \mu_{C_m}(gk_2)\right), ..., \left(gk_m, \mu_{C_m}(gk_m)\right)\right\}$$

其中

$$\mu_{C_m}(gk_i) = \frac{n_{C_m}(gk_i)}{n_{C_m}(gk_i) + n_{FL}(gk_i)}; \quad m = 1, 2, ..., k$$

$n_{C_m}(gk_i)$：關鍵字 gk_i 在該類別 C_m 中出現的次數

$n_{FL}(gk_i)$：關鍵字 gk_i 在頻率列表中出現的次數

則未分類文件 D_r 與類別 C_m 之間的相關程度即以歸屬函數 $\mu_{Cm}(D_r)$ 來加以表示。其計算方式為公式 3.3：

$$\mu_{C_m}(D_r) = \frac{\sum_{i=1}^{m} \mu_{D_r}(gk_i)\mu_{C_m}(gk_i)}{\sum_{i=1}^{m} \mu_{D_r}(gk_i)}; \quad \begin{array}{ll} D_r, & r = 1, 2, ..., n \\ C_m, & m = 1, 2, ..., k \\ gk_i, & i = 1, 2, ..., m \end{array} \quad \text{(公式 3.3)}$$

如果把所有未分類文件 D_1，D_2，…，D_n 歸屬於各分類類別 C_m 的程度列成矩陣，就得到矩陣 M_3，為公式 3.4。

$$M_3 = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} \mu_{C_1}(D_1) & \mu_{C_2}(D_1) & \cdots & \mu_{C_k}(D_1) \\ \mu_{C_1}(D_2) & \mu_{C_2}(D_2) & \cdots & \mu_{C_k}(D_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{C_1}(D_n) & \mu_{C_2}(D_n) & \cdots & \mu_{C_k}(D_n) \end{bmatrix}_{n\times k} \end{array} \quad \text{(公式 3.4)}$$

公式 3.4 中，列是對應類，行對應文件。

3.6 計算文件於多重類別的歸屬度

當每一個未分類文件和各個分類類別之間的相關程度都以模糊理論中的歸屬函數表示後，接下來我們即利用模糊集合理論的交集運算 (參考定義 2.3) 來計算每一個未分類文件在多重分類類別的歸屬度。若將 $C_m \cap C_n$ (m、n =1, 2, …, k) 的歸屬函數用矩陣形式表示，則得到矩陣 M_4，為公式 3.5。

$$M_4 = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 \cap C_2 & C_1 \cap C_3 & \cdots & C_{k-1} \cap C_k \end{array} \\ \begin{bmatrix} \mu_{C_1 \cap C_2}(D_1) & \mu_{C_1 \cap C_3}(D_1) & \cdots & \mu_{C_{k-1} \cap C_k}(D_1) \\ \mu_{C_1 \cap C_2}(D_2) & \mu_{C_1 \cap C_3}(D_2) & \cdots & \mu_{C_{k-1} \cap C_k}(D_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{C_1 \cap C_2}(D_n) & \mu_{C_1 \cap C_3}(D_n) & \cdots & \mu_{C_{k-1} \cap C_k}(D_n) \end{bmatrix}_{n\times C_n^2} \end{array} \quad \text{(公式 3.5)}$$

其中 $\mu_{Cm \cap Cn}(D_r)$ 是表示文件 D_r 在模糊集 $C_m \cap C_n$ 上的歸屬函數，其計算方式為公式 3.6。

$$\mu_{Cm \cap Cn}(D_r) = min\left\{\mu_{C_m}(D_r), \mu_{C_n}(D_r)\right\} \quad \text{(公式 3.6)}$$

然而，模糊多重分類原則應是使每個文件至少分到一類中去，所以模糊訊息檢索分類法引用了 α-截值的觀念做為分類過程中的一個模糊項臨界值 (Threshold)。這個臨界值用以滿足限制條件的最低程度且提供多重分類的彈性。經由觀察公式 3.5 的運算結果，可得出 $max[\mu_{Cm \cap Cn}(D_r)]$，在代入公式 3.7，即可確定 α-截值 (α-*cut*)，其計算方式為：

$$\alpha < \min_{m,n}\left\{max\left[\mu_{Cm \cap Cn}(D_r)\right]\right\} \quad \text{(公式 3.7)}$$

並且得到普通集合，即公式 3.8。

$$S_{C_m} = \left\{D \middle| \mu_{C_m}(D) \geq \alpha\right\} \quad (m = 1, 2, ..., k) \quad \text{(公式 3.8)}$$

於是可以根據普通集合 S_{Cm} (i = *1*, *2*, …, k) 來進行分類。

四、模糊資訊檢索分類的應用案例

Problem： 假設一個文件檔案有一組文件共 15 份，要用“計算機”，“應用數學”，“資訊科技”，“自動控制” 4 個描述項來表示，試討論分類情況。

Solution： 取定定義域 $U = \{D_1, D_2, \ldots, D_{15}\}$ 並按描述項初步分為 4 類。首先根據定義 3.3 和定義 3.4 求出 $\mu_{Dr}(k_i)$ 和 $\mu_{Cm}(k_i)$，(r =1, 2, …, 15；m =1, 2, 3, 4；i =1, 2, 3, 4)，即分別求出描

述字 (或關鍵詞) k_i在文件D_r及類別C_m中的重要程度。

定義 3.3：

$$FKS(D_r)=\{(k_1,\mu_{D_r}(k_1)),(k_2,\mu_{D_r}(k_2)),...,(k_m,\mu_{D_r}(k_m))\}$$

其中

$$\mu_{Dr}(k_i)=\frac{n_{D_r}(k_i)}{n_{D_r}(k_i)+n_{FL}(k_i)};\quad r=1,2,...,n$$

$n_{D_r}(k_j)$:關鍵字k_j在該文件D_r中出現的次數
$n_{FL}(k_i)$:關鍵字k_i在頻率列表中出現的次數

定義 3.4：

$$FKS(C_m)=\{(k_1,\mu_{C_m}(k_1)),(k_2,\mu_{C_m}(k_2)),...,(k_m,\mu_{C_m}(k_m))\}$$

其中

$$\mu_{C_m}(k_i)=\frac{n_{C_m}(k_i)}{n_{C_m}(k_i)+n_{FL}(k_i)};\quad m=1,2,...,k$$

$n_{C_m}(k_j)$:關鍵字k_i在該類別C_m中出現的次數
$n_{FL}(k_i)$:關鍵字k_i在頻率列表中出現的次數

再利用公式 3.3 算出歸屬函數，假設得到了公式 3.4 中的矩陣 M_3 為

$$\mu_{C_m}(D_r)=\frac{\sum_{i=1}^{m}\mu_{D_r}(k_i)\mu_{C_m}(k_i)}{\sum_{i=1}^{m}\mu_{D_r}(k_i)}\quad\begin{matrix}D_r,\ r=1,2,...,n\\ C_m,\ m=1,2,...,k\\ k_i,\ i=1,2,...,m\end{matrix}\quad\text{(公式 3.3)}$$

$$M_3=\begin{matrix}\\ D_1\\ D_2\\ \vdots\\ D_{15}\end{matrix}\begin{matrix}\begin{matrix}C_1 & C_2 & C_3 & C_4\end{matrix}\\ \begin{bmatrix}\mu_{C_1}(D_1) & \mu_{C_2}(D_1) & \mu_{C_3}(D_1) & \mu_{C_4}(D_1)\\ \mu_{C_1}(D_2) & \mu_{C_2}(D_2) & \mu_{C_3}(D_2) & \mu_{C_4}(D_2)\\ \vdots & \vdots & \vdots & \vdots\\ \mu_{C_1}(D_{15}) & \mu_{C_2}(D_{15}) & \mu_{C_3}(D_{15}) & \mu_{C_4}(D_{15})\end{bmatrix}_{15\times4}\end{matrix}\quad\text{(公式 3.4)}$$

$$M_3=\begin{matrix}\\ D_1\\ D_2\\ D_3\\ D_4\\ D_5\\ D_6\\ D_7\\ D_8\\ D_9\\ D_{10}\\ D_{11}\\ D_{12}\\ D_{13}\\ D_{14}\\ D_{15}\end{matrix}\begin{matrix}\begin{matrix}C_1 & C_2 & C_3 & C_4\end{matrix}\\ \begin{bmatrix}0.63 & 0.60 & 0.76 & 0.46\\ 0.69 & 0.14 & 0.27 & 0.33\\ 0.06 & 0.29 & 0.53 & 0.98\\ 0.58 & 0.51 & 0.62 & 0.76\\ 0.66 & 0.37 & 0.52 & 0.39\\ 0.34 & 0.45 & 0.64 & 0.72\\ 0.60 & 0.56 & 0.69 & 0.61\\ 0.37 & 0.21 & 0.40 & 0.65\\ 0.63 & 0.33 & 0.45 & 0.54\\ 0.32 & 0.40 & 0.57 & 0.87\\ 0.46 & 0.34 & 0.52 & 0.59\\ 0.63 & 0.42 & 0.55 & 0.51\\ 0.42 & 0.47 & 0.64 & 0.73\\ 0.44 & 0.31 & 0.47 & 0.69\\ 0.49 & 0.39 & 0.55 & 0.63\end{bmatrix}\end{matrix}$$

由於在本例中所有文件都有 $\mu_{C2}(D)<\mu_{C3}(D)$，即 $C_2\subseteq C_3$。故只須分為 C_1、C_3和 C_4 三類，由 M_3代入公式 3.5 和公式 3.6 算得 M_4。

$$M_4=\begin{matrix}\\ D_1\\ D_2\\ \vdots\\ D_{15}\end{matrix}\begin{matrix}\begin{matrix}C_1\cap C_3 & C_1\cap C_4 & C_3\cap C_4\end{matrix}\\ \begin{bmatrix}\mu_{C_1\cap C_3}(D_1) & \mu_{C_1\cap C_4}(D_1) & \mu_{C_3\cap C_4}(D_1)\\ \mu_{C_1\cap C_3}(D_2) & \mu_{C_1\cap C_4}(D_2) & \mu_{C_3\cap C_4}(D_2)\\ \vdots & \vdots & \vdots\\ \mu_{C_1\cap C_3}(D_{15}) & \mu_{C_1\cap C_4}(D_{15}) & \mu_{C_3\cap C_4}(D_{15})\end{bmatrix}_{15\times C_{15}^2}\end{matrix}\quad\text{(公式 3.5)}$$

$$M_4=\begin{matrix}\\ D_1\\ D_2\\ D_3\\ D_4\\ D_5\\ D_6\\ D_7\\ D_8\\ D_9\\ D_{10}\\ D_{11}\\ D_{12}\\ D_{13}\\ D_{14}\\ D_{15}\end{matrix}\begin{matrix}\begin{matrix}C_1\cap C_3 & C_1\cap C_4 & C_3\cap C_4\end{matrix}\\ \begin{bmatrix}\underline{0.63} & 0.46 & 0.46\\ 0.27 & 0.33 & 0.27\\ 0.06 & 0.06 & 0.53\\ 0.52 & 0.58 & 0.62\\ 0.52 & 0.40 & 0.40\\ 0.34 & 0.34 & \underline{0.64}\\ 0.60 & \underline{0.60} & 0.61\\ 0.37 & 0.37 & 0.40\\ 0.45 & 0.54 & 0.45\\ 0.32 & 0.32 & 0.57\\ 0.46 & 0.46 & 0.52\\ 0.55 & 0.51 & 0.51\\ 0.42 & 0.42 & 0.60\\ 0.44 & 0.44 & 0.47\\ 0.49 & 0.49 & 0.55\end{bmatrix}\end{matrix}$$

由 M_4得 $max_{\mu C1\cap C3}=0.63$, $max_{\mu C1\cap C4}=0.60$、$max_{\mu C3\cap C4}=0.64$。再代入公式 3.7，求得$\alpha<0.60$。

$$\alpha<\min_{m,n}\{max[\mu_{Cm\cap Cn}(D_r)]\}=\min_{m,n}\{0.63,0.60,0.64\}\ \text{(公式 3.7)}$$

$$\alpha<0.60$$

故應滿足 $\alpha<0.60$，我們取定 $\alpha=0.59$，再按矩陣 M_3，可得文件分類

$$S_{C_1}=\{D_1,D_2,D_5,D_7,D_9,D_{12}\}$$
$$S_{C_3}=\{D_1,D_4,D_6,D_7,D_{13}\}$$
$$S_{C_4}=\{D_3,D_4,D_6,D_7,D_8,D_{10},D_{11},D_{13},D_{14},D_{15}\}$$

這裡的分類中，一個文件可以同時屬於多類，但是這種情形對於文件的訊息檢索分類是合理的，因此所得結果同實際相符。

五、分類方法之效能評估

為了評估分類方法的效能，我們使用的度量值為：正確率 (Precision Rate) 和回現率 (Recall Rate)[2][18]。正確率是使用者每一次查詢之後，系統提供檢索結果中正確的資料量佔檢索結果出來的資訊總數比率，用來評估系統擷取的精確度。回現率是使用者每一次查詢之後，系統回傳擷取結果裏正確的資料量佔符合查詢要求的資訊總數比率，用來評估系統擷取的廣泛程度。

在此以 C_m 類別來說明這兩個度量值的意義。圖 5.1 為分類方法之效能評估方式。

$$precision=\frac{\alpha}{\alpha+\beta}$$
$$recall=\frac{\alpha}{\alpha+\delta}$$

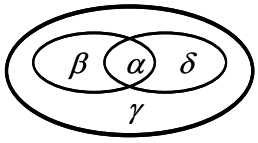

圖 5.1：分類方法之效能評估方式

假設所有未分類之文件的數目為 N，當分類完成後，可能發生的情況有下面四種：

1. 有α篇被歸類於 C_m 類別，且歸類正確。
2. 有β篇被歸類於 C_m 類別，但歸類錯誤。
3. 有γ篇不被歸類於 C_m 類別，且歸類正確。
4. 有δ篇不被歸類於 C_m 類別，但歸類錯誤。

且$(\alpha+\beta+\gamma+\delta)=N$，則 C_m 類別的回現率與正確率如下表示：

$$\text{正確率 } P(C_m)=\frac{\alpha}{\alpha+\beta}=\frac{(\text{擷取之與使用者需求相關文件數})}{(\text{擷取之總文件數})}$$

$$\text{回收率 } R(C_m)=\frac{\alpha}{\alpha+\delta}=\frac{(\text{擷取之與使用者需求相關文件數})}{(\text{與使用者需求相關文件數})}$$

當類別回現率 $R(\cdot)$或正確率 $P(\cdot)$越高，則表示分類方法的效能越佳。當然，高回現率與高正確率是本研究所期望的，但兩者很難同時滿足，合理的情形是在某個回現率的範圍內，儘量提高正確率[16]。

六、結論與未來研究方向

到目前為止，雖已有許多處理文件分類問題的方法相繼被提出，但是，這些方法均是為

處理單一分類問題所設計的，且並不適用於處理多重分類問題。因此，本研究以模糊理論來做多重分類的原因如下：

1. 較適合自然語言：人類之語言大都是表示模糊觀念，所以用模糊數學來做文件分類，理論上是很適用的。
2. 不確定性：文件之分類本來是不確定的，就算用人工分類，分類結果也會因人而異，我們並不能硬性說某一篇文章就是屬於某一類，而用模糊之歸屬度來說明文件所屬之類別是較為合理。且應用於多重分類，可以增加文件資料分類之正確性。
3. 一致性：人工分類常常因為人員的異動，同樣一篇文章有不同標準的分類，若利用模糊內積的計算方式來讓電腦自動分類就不會有這樣的問題。

本研究使用模糊訊息檢索分類應用於文件多重分類問題的方法，在處理文件多重分類問題時，比起用傳統處理文件單一分類問題的方法來處理文件多重分類問題更加的適合。

根據 Survey.com (http://www.survey.com) 的分析結果顯示：其實企業所需要的商業智慧 (Business Intelligence, BI) 大約只有 20% 是由存放在傳統關聯式資料庫中的結構化資料所推導出來的。其餘 80% 左右的商業智慧必須要到各式各樣的商業文件中去找尋。目前企業界對於這些文件的管理上也僅止於文件本體的管理，對於文件的內容仍然是以人為閱讀的方式來吸收，效率不彰且可能流於以偏蓋全。因此，在資料倉儲與資料採擷 (Data Mining) 已經普遍為企業界所認同與採行之際，學術界應當加緊腳步邁向下一個挑戰，那就是「文件倉儲」(Document Warehouse) 與「文件採擷」(Text Mining) 的深入研究，以協助企業進一步掌握整體的商業智慧，提昇整體的競爭力。

致謝

本研究部分承蒙國科會計劃補助，計劃編號 NSC 91-2416-H-327-005，特此感謝。

參考文獻

[1] Baeza-Yates, R. and B. Ribeiro-Neto, "*Modern Information Retrieval*," Addison-Wesley, 1999.

[2] Benkhalifa, M., A. Bensaid and A. Mouradi, "Text Categorization Using the Semi-supervised Fuzzy C-algorithem," *NAFIPS Int'l Fuzzy Info. Processing Soc.*, 1999, pp. 561-565.

[3] Blosseville, M.J. *et al.*, "Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together," *ACM Trans. on Info. Sys.*, July 1992, pp. 51-58.

[4] Can, F. and E. Ozkarahan, "A Dynamic Cluster Maintenance System for Information Retrieval," *Proc. the 10th Annual Int'l ACM SIGIR Conf.*, 1987, pp. 123-131.

[5] Dubois, D. and H. Prade, "*Fuzzy Sets and System: Theory and Applications*," Academic Press, New York, 1980.

[6] Heaps, H.S., "*Information Retrieval – Computational and Theoretical Aspects*," Academic Press, New York, 1978.

[7] Jain, A.K. and R. C. Dubes, "*Algorithms of Clustering Data*," Prentice-Hall, Inc., 1988.

[8] Lance, G. N. and W. T. Williams, "A General Theory of Classificatory Sorting Strategies: II. Clustering Systems," *Computer Journal 10*, 1967, pp. 271-277.

[9] Miyamoto, S., "*Fuzzy Sets in Information Retrieval and Cluster Analysis*," Kluwer Academic Publishers, May 1990.

[10] Murai, M.M. and M. Shimbo, "A Fuzzy Document Retrieval Method Based on Two-Valued Indexing," *Fuzzy Sets And Systems*, Vol. 30, 1989, pp. 103-120.

[11] Radecki, T., "Fuzzy Set Theoretical Approach to Document Retrieval," *Infor. Processing and Management*, Vol. 15, 1979, pp. 247-259.

[12] Radecki, T., "Mathematical Model of Information Retrieval Systems Based on the Concept of Fuzzy Thesaurus," *Info. Processing and Management*, Vol. 12, 1976, pp. 313-318.

[13] Sullivan, D., "*Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales*," John Wiley & Son, Inc., 2001..

[14] Tahani, V., "A Conceptual Framework for Fuzzy Query Processing --A Step toward Very Intelligent Database Systems," *Info. Processing and Management*, Vol. 13, 1977, pp. 289-303.

[15] Tahani, V., "A Fuzzy Model of Document Retrieval Systems," *Info. Processing and Management*, Vol. 12, 1976, pp. 177-187.

[16] van Rijsbergen, C. J., "*Information Retrieval*," 2nd Ed., London, 1979.

[17] Yager, R.R., "A Logical On-Line Bibliographic Searcher: an Application of Fuzzy Sets," *IEEE Trans. Systems, Man, & Cybernetics*, Vol. 10, No.1, 1980, pp. 51-53.

[18] Yang, Y. and C.G. Chute, "An Example-Based Mapping Method for Text Catalogization and Retrieval," *ACM Trans. on Info. Sys.*, Vol. 12, No. 3, July 1994, pp. 252-277.

[19] Zadeh, L.A.,, "Fuzzy Sets," *Information and Control*, Vol. 8, 1965, pp. 338-353.

[20] Zimmermann, H.-J., "*Fuzzy Set Theory – and Its Applications*," 2nd revised edition, Kluwer Academic Publishers, 1991.

[21] 吳萬鐸與吳萬釗編著，*模糊數學與計算機應用*，台北市，儒林圖書，9 月，1993.

[22] 李孟瑜、曾秋蓉，*智慧型自動化網路客服系統之研究*，台灣區網際網路研討會，2001。

[23] 陳淑美，*財經新聞自動分類之研究*，國立台灣大學圖書館學研究所碩士論文，1992。

[24] 楊允言，*文件自動分類及其相似性排序*，國立清華大學資訊科學研究所碩士論文，1993。

[25] 謝清俊、陳淑美、楊允言、陳克健，*Auto classification of Texts*，如何利用大型語料庫作研究研討會，計算語言學會，9 月，1992。