網路文件自動分類
Automatic Network Documents Classification

顧皓光　　莊裕澤
Hao-Kuang Ku　　Yuh-Jzer Joung
r4725012@im.ntu.edu.tw　　*joung@ccms.ntu.edu.tw*
國立台灣大學資訊管理研究所
Department of Information Management
National Taiwan University, Taipei, Taiwan

摘要

本論文提出一個適合網路文件自動分類的模型，藉以幫助使用者處理這些資訊。首先，我們使用網路資源蒐集程式將蕃薯藤搜尋引擎上的文件取回，接著使用這些已經具備分類特性的文件當作訓練文件，在機器中建立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。另外，鑑於Web文件提供了超文件連結的特性、HTML TAG標籤加註的功能，我們充分利用這兩項特性，設計實驗方法，藉以提昇系統分類能力。

實驗結果顯示我們提出的方法在13個基本類別條件下，可以正確辨識文件集合中71.2%的文件，辨識錯誤的文件有24.3%，另外有4.5%的文件是屬於無法辨識的文件。

1、序論

隨著網路的蓬勃發展，現在有許多人願意花比以往多很多的時間上網路，不管是寄送EMail 也好、工作也好、當作休閒娛樂也罷，網路已漸漸成爲很多人生活的一部份。在這過程當中，人們也正不停的將現實世界裡現有的資訊，諸如報紙、雜誌、書籍、政府文件等等，都丟到網路上，也不停的在網路上生產新的資訊。

就以WWW 來說，根據統計[7]，WWW 大約以每年十倍的驚人速率成長，其中包括各種種類的資訊，如政治、經濟、新聞、教育、娛樂等等。隨著各種資訊的大量成長，Internet儼然已經成爲了一個知識的大寶庫。

網路的普及使得網路上的資訊迅速增加，如果不能有系統的儲存整理這些資訊，日後的查詢將變得十分困難。所謂儲存整理的概念，最容易被大家所了解的就是全文檢索技術，這種技術是以字串比對爲核心，圍繞在其周圍的還有布林運算、語意典等等的技術，一般常見的系統均是以此爲主。

雖然說目前的全文檢索技術可以說很完備了，但使用者除了比對字串之外，還有一個很大的需求是對資料的瀏覽。我們都有一個經驗，在圖書館找資料的時候，最後找到的資料往往不是一開始找到的，而是以瀏覽的方式從一個主題中整理得來。而爲了達到瀏覽的功能，需要加入分類功能，所謂分類功能便是事先以人工將資訊分類，再利用系統整理儲存，提供使用者自行尋找所需的資訊。這是分類應用在資訊檢索上最明顯的一個例子。

除了瀏覽之外，「文件分類」還可以協助檢索系統縮小全文檢索查詢範圍。以往全文檢索系統的比對方法是將查詢句和文件集合內每一篇文件逐一比對，從中尋找出最大交集的部份，整個過程費時又費力，效果也不見得很好。試想若加入了文件分類觀念，各個文件均有特別的類別意義，查詢時，可打破原本各個文件一視同仁的情況，僅對部份高度相關的文件子集合進行更精密的檢索，將可加速查詢速度與提高查詢正確率。這是分類應用在資訊檢索上第二個例子。

更有甚者，文件分類系統還可以幫助使用者管理日常文件，去除垃圾信件或是顯示緊急的文件，在資訊檢索上稱之爲資訊過濾或是資訊代理人。

有鑑於「文件分類」在資訊檢索領域中如此重要，引起我們研究分類的動機，希望能由此研究累積中文資訊檢索的知識。然而，「文件分類」範圍廣泛，絕不是個人一朝一夕可以窮究，我們特別注意其中和自動化有關的部份，也就是「文件自動分類」的議題，尤其是「文件自動分類」在中文上應用。

另外，以往研究「文件自動分類」時，大多是以紙本資料爲實驗對象，隨著網路文件日益增加，未來資訊檢索研究樣本將更爲豐富，本研究也針對此趨勢，將「文件自動分類」的研究延伸到網路文件。總結起來，本研究的研究目的在探討資訊檢索領域中文件自動分類模型在新興網路文件的應用。

1.1 文件分類在網路資源檢索的使用模型

如何「分類」呢，和字串比對這類0 與1 的概念不同，分類牽涉到更深層有關認知的部分。以往資訊科技能提供的支援有限，不敢奢望能藉由電腦來幫我們分類，因此大多是由人工進行，不過現在資訊增加的速度遠超過人工分類所能負荷的程度，自動分類的需求於是因應而生。

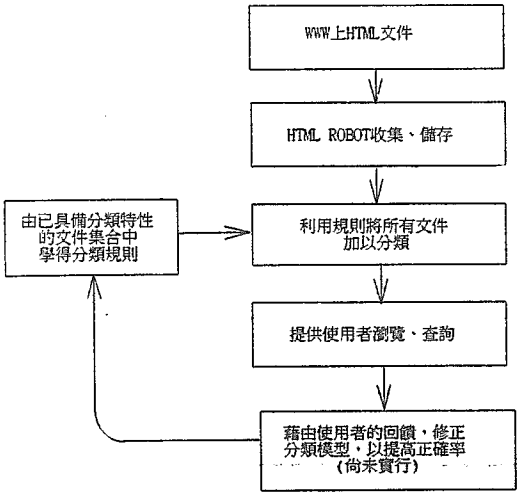

圖1-1　運用模型處理整體網路文件

所謂的「自動」分類，是指一篇新的文件經過分類模型的處理，系統可以自行決定此文件的類別，整個過程中完全不需要人工介入。原本人工判斷部份全由分類模型取代，分類模型的優劣決定了系統的品質，模型愈好，愈能正確判斷文件的類別，反之相反。一旦模型建立完成，便可以使用此模型處理許許多多未曾分類過的文件，達到自動分類的目的，如圖1-1。

1.2 本論文所採用的方法

自動分類的相關研究在西方已行之有年[13] [16] [17][8][9][18]，國內從民國八十年開始陸續有研究生投入這方面的研究 [1] [3] [4]，屢屢有突破，整理前人精華，可以發現自動分類的理論根據是：系統可依照文件內容自動將性質相近的文件存放於鄰近的地方，以便日後的查詢，查詢的方式是衡量新文件和舊有群集的關係。

就這部份來看，若是進行自動分類實驗，需要一個已經具備分類性質的文件集合當作學習範本，系統根據此範本的資訊勾勒出文件集合中各個類別的分佈關係。另外爲了驗證系統的正確性，必須在原有的文件集合中抽取一部份當作測試資料，由於我們已經知道測試資料的「真正」分類結果，只要比對系統的分類結果和真正的分類結果，就可以客觀的衡量分類正確能力，如圖1-2。

圖1-2 解釋了本研究的幾個主要工作：

一、 尋找具分類特性之文件集合：蕃薯藤網站即具備此特性，成爲我們的實驗對象。

二、 製作網路文件蒐集器：爲取得蕃薯藤網站資訊，特別針對WWW通訊協定，製作網路文件蒐集器，蒐集實驗所需的資料。

三、 區分訓練資料和測試資料：前面提及，分類模型便是一個學習模型，但爲了測試學習模型的正確性，必須在同一組資料中區分訓練資料和測試資料，以訓練資料架構出學習模型，再由測試資料驗證正確率。

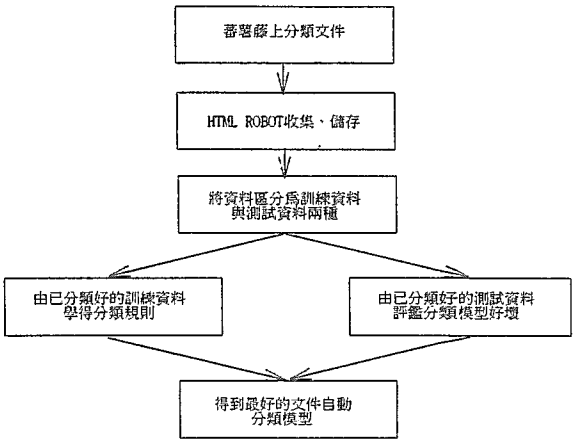

圖 1-2　論文使用方法

四、 文件自動分類：整個實驗討論重點便是「文件自動分類」模組。我們運用已有文件自動分類技術，再加上HTML語言特性，以蕃薯藤資料爲基本，做一系列自動分類的研究。相對於以往的研究，我們的實驗資料取自公開系統，內容涵括食衣住行各種資訊需求，範圍較大；另外一個特點是包含網路HTML語言特性。

五、 架構出一個適合瀏覽功能的搜尋引擎：以前述模組爲核心，建立一個可以兼具分類和字串比對的系統。

本論文的架構如下：第二章主要針對本實驗採用的技術和方法作一個回顧。第三章討論實驗資料，並預估可能造成的限制。第四章討論文件分類實驗的各個步驟，接著進行我們的實驗。第五章總結。

2、文獻探討

文件自動分類的工作奠定在資訊檢索，而其根本的問題在如何表達一篇文件。其中被引用最多的是Salton從1971開始使用的「向量空間模型」（Vector Space Model：簡稱VSM ）[16] [12] [14]。Salton提出以文件索引向量（Index Vector）爲主，並以關鍵詞顯著值（Term Significances）計算結果當作評估索引優劣的方法。

索引向量的定義如下：文件（D）的內容可以以關鍵詞（T）的向量空間線性組合方式代表$D = (T_1, T_2, \ldots, T_i)$。

若將屬性加權，索引向量就改寫成$D = ([T_1, W_1], [T_2, W_2] \ldots, [T_i, W_i])$：所謂的「加權」代表該關鍵詞（屬性）在文件上的顯著值（重要性）。

VSM 模型的數學基礎是矩陣代數（Matrix Algebra），核心是「關鍵詞－文件矩陣」（Term－Document Matrix ）。採用向量代表各個文件，不僅可以方便表現出各文件之間的關係，同時也可以計算彼此相似程度。另外，用這種「關鍵詞－文件」

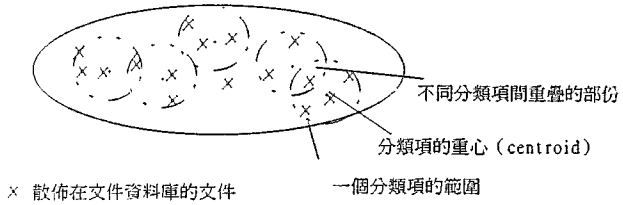

圖 2-1 文件自動分類的概念

矩陣，可以在機器中方便的表達關鍵詞與文件的觀念。

文件自動分類法的基本觀念是：計算出文件和詞彙頻率的相關性，利用這組相關性將一群群文件依照性質或內容之不同予以分開，將同性質的資料彙集一起，日後若有新文件進入，就可以判斷出該文件的類別屬性，如圖2-1。

如何由原始的文件資料整理得出分類結果？可分成六個步驟， [10] [16]。

一、選定文件集合。

二、以人工的方式訂定各個分類項，例如「教育」、「娛樂」、「休閒」等等等各個不同的類別。再依次對第一步驟中蒐集的文件進行人工分類，確定文件是屬於各個類別中的哪一個。

三、用斷詞程式擷取出每篇文件的詞彙，建立第一階段「文件詞彙」VSM 模型，接著在合併這個矩陣中的資料成爲第二階段「類別詞彙」VSM 模型。這個矩陣中包含了文件資料庫中所有的詞彙，可是並不是所有的詞彙都這麼重要，必須使用篩選方法去除不必要的詞彙，剩下的詞彙對分類的鑑別力變得很大，我們稱其爲關鍵詞。

排除人工斷詞，常見的自動斷詞方法有長詞優先、法則式、機率式方法、N-Gram取詞方式[2] [3] [5] [6] [15] [11]。

篩選關鍵詞必須考慮：

(1) 關鍵詞出現次數：大致來說，出現一次兩次的詞都會被去掉，除非它是很專門、很新潮的東西；另外，出現太多次也先去掉，例如英文的「the 」，中文的「的」。至於說這個界限值訂多少才好，要經過實驗才能決定的。

(2) 集中度（Conformity）：集中度運用到熵（Entropy）的觀念。「一個具有分類價值的關鍵詞，應該是集中出現在某幾類文件中，而不是平均分佈在各類」。若是一個關鍵詞只集中出現在某幾個分類之中，則這個關鍵詞愈能代表這幾個類別；反之，若是一個關鍵詞分佈出現在很多個分類之中，則這個關鍵詞的鑑別能力很低。

衡量關鍵詞T_i的集中度icf（Inverse Cluster Frequency）的公式如下：

$$icf = -\sum_{Ck=1}^{N_{Class}} P_{CkTi} \log P_{CkTi} = \sum_{Ck=1}^{N_{Class}} P_{CkTi} \log \frac{1}{P_{CkTi}}$$

$$P_{CkTi} = \frac{FD_{CkTi}}{\sum_{Ck=1}^{N_{Class}} FD_{CkTi}}$$

其中

C_k: 類別k（Class k）

N_{Class}: 文件集合中類別總數

FD_{CkTi}: C_k中出現T_i的文件數目

P_{CkTi}: T_i在C_k中出現的機率

至於集中度的界限值大小爲何，並沒有一定的標準，往往需要經過多次的實驗反覆推敲得知。

(3) 廣度（Uniformity）：一個具有分類價值的關鍵詞，應該平均出現在某一分類的各篇文件中，而不是只集中在幾篇文件內。計算方法同樣運用熵的概念，只是現在將範圍縮小到某一個類別。

衡量關鍵詞T_i在某類別C_k的公式如下：

$$-\sum_{Dj=1}^{N_{Class-Doc}} P_{DjTi} \log P_{DjTi} = \sum_{Dj=1}^{N_{Class-Doc}} P_{DjTi} \log \frac{1}{P_{DjTi}}$$

$$P_{DjTi} = \frac{F_{DjTi}}{\sum_{Dj=1}^{N_{Class-Doc}} F_{DjTi}}$$

其中

$N_{Class-Doc}$: 類別C_k的文件總數

P_{DjTi}: T_i在文件類別C_k的文件D_j中出現的機率

四、接下來要考慮如何指定每個關鍵詞和類別間的權重給定方式。常見的有下列五種方法：

	說明
頻率統計	出現過幾次，就記錄幾次
原始比重	根據此關鍵詞在各類的分佈情形來決定
第一種標準化比重	各關鍵詞的向量長度拉長至單位向量中
第二種標準化比重	每一類文件數與其所代表向量長度之間存在非線性的關係
詞彙比重分析法	同時考慮集中度和廣度

五、建立矩陣模型。一個具有m個類別和n個關鍵詞的VSM「關鍵詞－類別」矩陣，可以用下式表達。

$$\begin{array}{c} \\ \text{n個詞彙} \end{array} \begin{array}{c} \text{m個類別} \\ \begin{bmatrix} W_{11} & W_{21} & \cdots & W_{m1} \\ W_{12} & W_{22} & \cdots & W_{m2} \\ \cdots & & & \\ W_{1n} & W_{2n} & \cdots & W_{mn} \end{bmatrix} \end{array}$$

W_{ki}表示關鍵詞T_i在類別C_k內的權重。

六、指定類別：指定的方法主要根據向量模式，由類別矩陣和文件的向量進行內積運算，內積值最大者決定文件類別。

3、實驗資料分析

以往的資訊檢索實驗大多是配合全文檢索的目的而設計，強調每一篇文件中詞彙的特性。我們的實驗是以自動分類爲主，不再強調每一篇文件中詞彙的特性，而是這篇文件中有無類別的特性。

目前，並沒有一組具分類特性的標準資料供我們實驗。參諸以往論文的實驗文件，均是作者手邊能取得的資料，例如「經濟日報」和「國科會技術報告」。雖然這些資料已有分類特性，仍欠缺我們需要的HTML標註(tag)特性。如何取得已經經過人工分類，又兼具HTML標註特性的資料呢？我們發現有許多搜尋引擎內部已經有許多經由人工處理過的文件資料，以國內著名的搜尋引擎「蕃薯藤」來說，內部資料全部皆經過人工的判斷整理分類，共分成十二大類，每類內又有詳細的分類與整理，十分符合實驗的要求。更重要的是，它儲存的文件均有HTML標註，完全符合實驗需要。

我們使用Robot 程式將其完全拷貝至實驗機器中。以上工作告一段落之後，再加以整理，得出至民國86年3月，蕃薯藤統計資訊，如表3-1。

蕃薯藤內部共分爲十二個大項，每個大項中，又有詳細的分類，依此繼續。表中所謂子目錄就是指內部又有多少個分類項。文件個數是指使用者能夠看到的文件。而有關中文特性的統計，中文字共有1907831 個，常見字出現了1904249次，罕見字出現了3582次。

觀察這些資料，發現有兩點問題。

- 從比例上，可以發現各個類別占全體資料的比例差距十分不均勻。
- 重複分類的情形很多，所謂重複分類的意思是說，「同一篇文件重複出現在不同的類別之中」。例如「工商產業」這個大類中有一個「資訊服務業」子分類，同樣在「電腦網路」這個大類中，也有一個「資訊服務業」子分類。這兩個子分類的內容完全相同。類似的例子不甚枚舉。

理論上來說，各個類別的文件數目不平均，很有可能影響分類的正確率。文件多的類別，因爲母體夠大，詞彙出現的機會高，容易被辨識：相反的，文件少的類別，很多詞彙可能很重要，但是因爲蒐集的文件數目少於一定的標準值，很多有意義的詞彙反而被忽略了。

就重複分類的情形，對一個以提供瀏覽爲主的資料系統而言，重複分類是無可避免。重複分類的概念和關鍵詞選取的集中度原則相左，根據文件自動分類的基本原則，好的關鍵詞應該出現在少數幾個類別，而且是愈少愈好。如果允許重複分類，只會使得這個詞彙成爲關鍵詞的機會降低，目前重複分類的情形嚴重，自然會影響正確率。

雖然有上述兩點問題，但是我們仍然嘗試以目前的分布情形和比例做實驗，因爲這樣能夠反應我們進行網路文件分類實際將遭遇的困難。

表 3-1 蕃薯藤資訊統計

	子目錄個數	文件個數	檔案大小（KB）	所佔比例%（以檔案大小計）
人文	5	50	237	0.85
工商產業	100	2792	6738	24.00
生活資訊	18	307	888	3.16
自然科學	16	142	392	1.40
育樂休閒	92	3126	7352	26.19
社會文化	27	524	1372	4.89
社會科學	17	251	669	2.38
國家	33	449	1085	3.86
教育	125	962	2396	8.53
電腦網路	105	1537	5253	18.71
醫藥保健	16	345	912	3.25
藝術	17	225	783	2.79
總和	583	10710	28077	100

另外，經過實驗之後，發現上述所蒐集的資料可能稍嫌「不足」，這裡所謂的「不足」不是指量的不足，而是指資料在統計意義上的不足。原因如下：

一、設計網站主頁時，大多以精簡爲主，如果需要詳細的資料就要連結進網站內其他的文件。因爲此造成了每一個文件的內容都不大。以蕃薯藤來說，平均一篇文件中僅有178個中文字，再扣除掉「歡迎光臨某某某的網頁」、「你是第幾個光臨的客人」、「歡迎你寄信給我」等非常普通的句子，剩下可能不到120 字。我們怎麼期待120個字能提供什麼訊息？比對楊允言曾經做過的經濟日報實驗[3]，每篇新聞報導有500個中文字之多，差距實在很大。

二、網站主頁多愛用公司名稱、人名等模糊的專有名詞來代替普遍出現的名詞。舉例來說，台大的首頁中出現了多次「椰林」，一般人可以很快的由「椰林」引申出「台大」，而且覺得很有意思。但是很可惜的，我們的VSM 模型還不行。

三、網路文件，尤其是HTML文件，容易用圖形代替文字，就使用者和網頁設計者而言，寧願使用圖形也不願使用文字是自然的情況，但是對於以文字爲操作主體的VSM 模型來說，豐富的圖形反而降低了辨識的可能性。

爲了解決上述三項問題，我們希望能夠多拿一些和這篇文章相關的資料回來，何謂相關的資料？既然我們蒐集的網路文件是網站的文件，這些文件一定有連結會指向網站內部其他的文件，就意義上來看，網站內部的文件因爲範圍較大，在字數上一定比網站主頁多出很多，使用的詞彙必定也比較精確，圖形的部份也可以因此得到一些解決。在這個想法下，我們決定把蒐集回來一萬多篇網路主頁內所指向的文章全部抓回來，每個相連的文章組合成爲一個單一個文章，再以一個單獨的文件視之。如同圖3-1所示，每個區域內都是一個不同的網站，蕃薯藤指

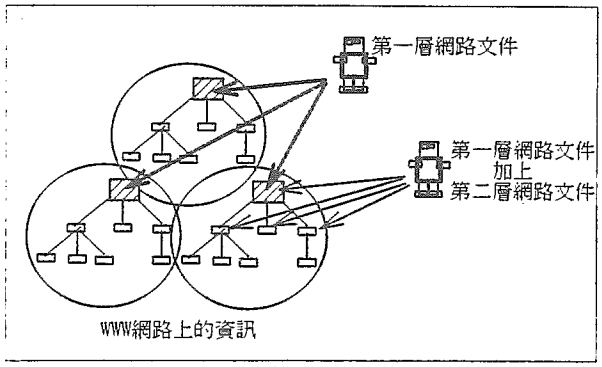

圖 3-1 網路文件之階級性

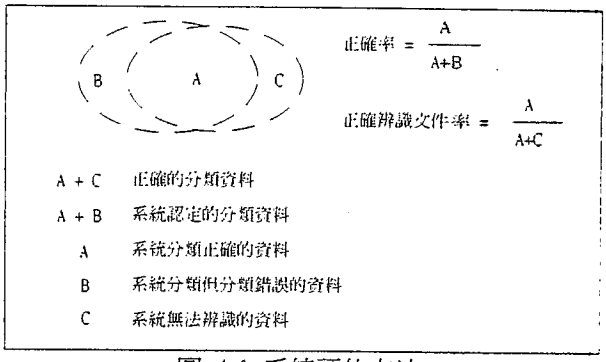

圖 4-1 系統評估方法

向的網站主頁只是該網站的一個進入點，這些資料被我們稱之爲「第一層網路文件」，如果要擴大選取的範圍，便要向下蒐集相關的資料，合併稱之爲「第一層網路文件加第二層網路文件」。

如此做得到的效益是：

一、因爲加入了很多相關的內文，字數大幅提高，從收集完畢的資料中可以看出，由原本平均一篇178 個中文字，提升到收集後平均一篇800 個中文字。

二、原本考慮網站主頁大多使用公司名稱，人名，專有名詞的問題，因爲加入了很多相關的內文，這些內文使用的詞彙和網站主頁比較起來應該較爲「平實」，較能符合系統的需求。

4、文件自動分類實驗

主要實驗方法是在以往的分類模型中加入網路文件(Web documents)的特性，試圖驗證自動分類的有效性，尤其是較大範圍的中文文件。所謂以往的分類模型，便是前述文獻探討中提到的文件自動分類技術，包括建立矩陣，依照次數、集中度、廣度的條件選取關鍵詞、給定關鍵詞比重。網路文件特性是指彼此連結和HTML中TAG資訊。實驗的觀念已經具備，比較有問題的地方是，如何給定各個參數的界限值，這些界限值往往牽一髮而動全身，彼此間又互相關連，很難從幾次實驗中就找到最佳解，需要進行多次的實驗才能得知。至此，文件自動分類的學習過程便告一段落，接下來進入測試的階段。我們實驗中有二項衡量的標準：正確率與正確辨識文件率，如圖4-1。

如何決定正確率與正確辨識文件率呢？我們的作法是先以隨機的方式將蕃薯藤的資料分爲兩組，訓練資料與測試資料，彼此的比例爲80%與20%，以訓練資料訓練建立模型，再以測試資料帶入模型中，比較其結果和原本人工分類結果，計算出正確率與正確辨識文件率。

實驗分成三個主要階段，第一階段是承襲以往前人的實驗技巧，測試蕃薯藤網站第一階層十二大類的正確率。我們調整了各個類別的分布〖註：我們將「職業運動」從「育樂休閒」中刪除，把「人物」從「育樂休閒」中移出單獨看成一個類別（因此新的資料總共分爲十三大類），「社會文化」中出現了很多次「女性」和「兩性關係」，我們也做了一些改變；另一個比較大的變動是移除「工商產業」中的「資訊電腦業」。〗、測試了各種比重給定方法、還改變了實驗的詞彙，諸多驗證後發現：採用原始比重方法效果最好：

	訓練資料	測試資料
正確率	75.9%	71.5%
正確辨識文件率	74.7%	70.1%

據我們了解，這是第一個從事中文網路文件分類的研究，因此上述的數據無從比較優劣。但是爲了證實我們的分類系統是否夠完善，我們以楊允言的實驗資料[3]來驗證我們的系統。楊允言實驗的文件經過精密篩選，除了不會有分佈不均的問題外，各個類別間的差異也比較大。因此，理論上應該比網路上這種比較一般而普遍性的文件容易分類。而事實上，經過我們的實驗發現，在相同的實驗資料下，使用我們的系統正確率可以從楊允言原本的67%提升到80%，可見我們的系統的確適合處理文件自動分類，也比楊允言的系統來的好。但和百分之八十比起來，目前的71.5%似乎就沒有特別突出。這個部分，如同前面解釋的，可以歸因於以往的實驗大多在操作十分專業的語料庫，各個類別之間有比較明顯的界線。另外，網路HTML文件有其特殊的性質，容易以圖形代替文字，這也可能影響正確率。

第二個階段開始考慮網路文件的特性，由於我們認爲現階段所使用的蕃薯藤網站主頁文件內容不夠豐富，於是擴大蒐集文件範圍至「第二層網路文件」，實驗結果卻出乎原本預料之外反而比較差，這引發我們重新思考「究竟什麼樣的文件集合才適合當作處理網網站主頁的訓練資料呢，是原本網站主頁構成的訓練資料還是由更詳細的內部資料構成訓練資料呢」。

	訓練資料	測試資料
正確率	65.2%	62.5%
正確辨識文件率	60.7%	55.4%

第三個階段真正接觸HTML語言特性，我們將注意的焦點放在討論HTML TAG所能帶來的資訊。實驗結果發現，如果僅僅擷取TAG內的詞彙當作實驗的資料，效果不會比較好；如果加重TAG內詞彙的權重，正確率比以往實驗方法高一些，除此之外，指定一篇文件爲某個類別的信心也比較高。

	訓練資料	測試資料
正確率	77.6%	74.5%
正確辨識文件率	75.8%	71.2%

經由我們的初步實作，發現利用上述模型的確可以幫助完成一個分類檢索系統，另外發現，建構VSM模型時所取得的一些資料可以順便幫助資訊檢索使用，例如詞彙在各個類別的分布特性將可以應用在提昇索引效能。

5、結論與展望

我們以蕃薯藤網站內全部資料爲主體，運用VSM模型的技術，進行文件自動分類的實驗，取詞的方式是採用自動斷詞(長詞優先斷詞方法)，參考的詞庫是中研院八萬目詞。

文件自動分類一直是很多人的夢想，早期是因爲熱衷於人工智慧，企盼藉由學習分類的功能以模倣人類判斷、比較、類歸的功能，晚期則是鑒於資訊充斥，自動分類可以大量取代傳統人工分類，整理更多資訊。

我們不敢否定人工分類的價值，相反的，還要特別強調人工分類的正當性，時時觀察人腦是如何進行分類。實驗使用的VSM模型是眾多模擬方式中最常被使用的一種，它是企圖藉由使用文件和關鍵詞的向量空間模型在機器中重現文件的特性，再利用篩選詞彙成爲關鍵詞、關鍵詞比重給定方法兩種技巧達到文件自動分類的目的。雖然達到一定成效，不過若要稱其有智慧還有一段距離。並且，從整個研究經驗所得，我們相信VSM模型並不是那麼容易瞭解的東西，還有更多隱含因素等待我們去了解。

經由本論文發現下列幾點：

一、整合以往的實驗方法對大範圍且具有網路特性的中文文件進行分類實驗，就理論層次來看，本實驗印證了「VSM模式的確可以代表並區別類別的內容」這項基本假設，就實驗結果來看，此模型可以正確分類百分之七十的測試文件。

二、以往認爲所有文件（網站主頁和內部文件）都可以使用同一個VSM模型進行分類，實驗結果發現不是如此，針對網站主頁的特殊性質可能要和處理內部文件的VSM模型有所區分。

三、如果考慮HTML內TAG 能夠提供的幫助後，分類正確率和正確辨識文件率會提昇。

最後，我們還做了一個實驗，把這篇論文的初稿放入系統辨識文件類別，很高興系統告訴我們這篇文件是電腦網路類別。

參考文獻：

[1] 陳淑美. 財經新聞自動分類研究.台大圖書館學研究所碩士論文. 民國81年.

[2] 彭載衍. 中文詞彙歧異之研究－斷詞與詞性標示. 清華資訊科學研究所碩士論文. 民國82年.

[3] 楊允言. 文件自動分類及其相似性排序. 清華資訊科學研究所碩士論文. 民國82年.

[4] 蔣俊霞. 中文文件自動分類之探討. 淡江資訊工程研究所碩士論文. 民國83年.

[5] 劉孟達. 中文詞及自動產生及中文拼字檢查. 中正資訊工程研究所碩士論文. 民國84年

[6] 洪振超. 網路首頁資源之國語語音檢索及其動態語言模型技術. 台大資訊工程研究所碩士論文. 民國85年.

[7] 果芸. 網際網路發展之回顧與展望. 資訊與電腦雜誌1996年11月號. 民國85年.

[8] M.Aboud., C.Chrisment., R.Razouk., F.Sedes. and C.Soule-Dupuy. Quering a hypertext information retrieval system by the use of classification. *Information Processing & Management*, vol. 29, no. 3, pp. 387-396. 1993.

[9] W.Bruced Croft. and Howard R. Turtle. Retrieval strategies for hypertext. *Information Processing & Management*, vol. 29, no. 3, pp. 313-324. 1993.

[10] Budi Yuwono., Savio L. Y. Lam., Jeffy H. Ying. and Dik L.Lee. A World Wide Web Resource Discovery System. In: Proc. of the 4th Int. World Wide Web Conf. 1995.

[11] Eugene Charniak. *Statistical language learning.* Massachusetts Institute of Technology. 1993.

[12] H. S. Heaps. *Information retrieval: computational and theoretical aspects.* Academic Press. 1993.

[13] Jacqueline W.T. Wong., W.K. Kan. and Glibert Young. ACTION : Automatic classification for full-text document. *ACM-SIGIR*, vol. 30, no. 1, pp. 26-41. 1996.

[14] R. N. Oddy., S. E. Robertson., C. J. van Rijsbergen. and P. W. Williams. Information Retrieval Research. Butterworths. 1981.

[15] Richard Sproat and Chilin Shih. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, vol. 4,no. 2,pp. 336-349. 1990.

[16] Salton, G. and McGill, M. J. *Introduction to modern information retrieval.* McGraw Hill Book Co. 1983.

[17] Valery I, F., Nick I, K. and Jacob, S. One Approach to classification of users and automatic clustering of documenta. *Information Processing & Management*, vol. 29, no. 2,pp. 187-195. 1993.

[18] Jacques Savoy. An Extended Vector-Processing Scheme for Searching Information in Hypertext. *Information Processing & Management*, vol. 32, no. 2, pp. 155-170. 1996.