中文文件自動分類之研究

A Study of Document Auto-Classification in Mandarin Chinese

†楊允言　‡謝清俊　⋆陳淑美　‡陳克健

† 中央研究院資訊科學研究所研究助理
‡ 中央研究院資訊科學研究所研究員
⋆ 國立臺灣大學圖書館館員

摘　　要

本論文中，我們提出利用雙連字串(Bigram)替代關鍵詞的方法，來做中文文件自動分類的實驗。其目的，是要讓電腦來幫忙做中文文件分類，減輕人的負擔。

我們從工商時報民國80年7月到81年1月間取樣出來的2306篇財經類新聞報導，包括產業、企業、機械、電機、資訊五大類，共24小類，先以人工將之分類，並分爲訓練資料(2095篇)及測試資料(211篇)兩部分，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，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，並比較其結果。實驗結果，測試資料有67%左右的正確率(召回率)，若取前三名有80%的正確率；至於訓練資料則有97%的正確率。

在文中，我們探討了關鍵詞的篩選以及文件自動分類的方法，採用向量模式時，並討論了標準化的方法；同時，我們針對電腦與人工在做分類以及相似性排序時的不同點提出簡單的比較與討論，讓我們了解之間的差異。

1　緒言

根據統計，自從1971年開始，平均每2.3年，線上資料庫的數量就增加一倍，而這些線上資料庫內的資料，則以更快的速度增加中[Smi89,p1]。如果沒有適當地存放這些資料，以後當我們要找尋所需的資料時，可能會遇到類似海底摸針的窘境。文件自動分類，就是將文件做某種方式的排列，使性質相近的文件被放在相同或靠近的地方，以便當人們要從眾多文件中查詢到其所需時，能有效率且迅

速地得到。

傳統的分類工作需要利用大量的人力，這樣不僅要耗去很多時間，並且，人工分類一直存在一個問題，即不同的人會做出不同的結果，不管是建索引或是判定文件的相似性等等，其一致性並不高[Sal86]。既然如此，如果我們利用電腦來幫人們做這件事，上述的兩個問題，就可以得到解決。

利用電腦來做文件自動分類的實驗，從1960年代就已開始[Mar61] [BoBe63]，並陸陸續續有相關論文發表，如[Kwo75][HaZa80]，最近這一、兩年則又引起了較多的關注，如[Lar92][BHMP92][Jac92][Jac93][Lew92]···等等。一般而言，利用電腦來做文件自動分類，不論哪種方式，大致包括以下的步驟：

1. 選定文件(Document)集合，並選定文件替代品(Profile)，例如只採用題目、摘要等等，做爲訓練及測試資料。並選定類別，即欲將文件分爲哪些類。

2. 從文件替代品中，根據訂定的篩選規則，找出所要的關鍵詞。

3. 類別的指定，可以採用向量模式(Vector Space Model)或是機率模式(Probability Model)來計算。

向量模式的做法是，假設關鍵詞有T_1,T_2,···,T_m共m個，文件分爲C_1,C_2,···,C_n共n類，我們以向量$X_j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$表示類別$C_j$，對一文件D，我們以向量$Y = (y_1, y_2, \ldots, y_m)^T$表示，關鍵詞$T_i$若出現在類別$C_j$中，則$x_{ij} > 0$，否則$x_{ij} = 0$，關鍵詞$T_i$若出現在文件D中，則$y_i > 0$，否則$y_i = 0$，將向量Y分別與向量$X_j$做內積運算，若其值最大，則表示文件D屬於此類。

，如何決定x_{ij}值是一個研究的主題，x_{ij}爲何，有何意義、x_{ij}如何選取，x_{ij}的值是否要加權(Weight)，x_{ij}值要不要做正規化(Normalization)、y_i的值是以二元關係(也就是說非0即1)或是以文件D的替代品(Profile)中詞彙k_i出現的次數爲準，···等等不一而足。　　若將向量模式的方法視爲分數(Scores)的相加，則機率模式的方法也可以簡單地視爲是分數的相乘。假設一篇文件D的替代品(Profile)中出現r個關鍵詞k_{i1}、k_{i2}、···、k_{ir}，則此文件屬於類別C_j的機率爲

$$P(C_j|k_{i1}, k_{i2}, \ldots, k_{ir})$$

根據貝氏定理，上式可化簡爲

$$\begin{aligned} & \frac{P(C_j) \times P(k_{i1}, k_{i2}, \ldots, k_{ir}|C_j)}{P(k_{i1}, k_{i2}, \ldots, k_{ir})} \\ = \; & r \times P(C_j) \times P(k_{i1}, k_{i2}, \ldots, k_{ir}|C_j) \\ = \; & r \times P(C_j) \times P(k_{i1}|C_j) \times P(k_{i2}|C_j, k_{i1}) \times \cdots \times P(k_{ir}|C_j, k_{i1}, k_{i2}, \ldots, k_{i,r-1}) \end{aligned}$$

假設k_{i1}、k_{i2}、$\cdots$、k_{ir}兩兩互相獨立(Independent),則上式可變成

$$r \times P(C_j) \times P(k_{i1}|C_j) \times P(k_{i2}|C_j) \times \cdots \times P(k_{ir}|C_j)$$

其中$r = 1/P(k_{i1}, k_{i2}, \ldots, k_{ir})$爲一常數[Mar61][HaZa80]。

如果將這些機率值取對數(Logarithm)之後再做運算,相乘變成相加,則我們又可將機率模式與向量模式看做是同一回事。

以上所提到的,都是針對英文及法文所做的,而我們清楚,中文與西方語言有相當大的差異。[Che92]開始處理中文的文件自動分類,但是其中的一個困難處是關鍵詞必須由人工所選取的。因此我們採用中文的雙連字串(Bigram)來取代「關鍵詞」,並且發現,利用雙連字串可以得到相同的結果。

本篇論文的研究方法與步驟,簡單敘述如下:

1. 選定資料(新聞報導)的範圍及分類系統。我們採用日本產經新聞所定的分類法,共挑出五大類、24小類。

2. 以工商時報爲對象,取民國80年7月至81年1月,每8天取樣一天,將當天合於選定範圍的新聞報導做爲樣本,共有2306篇,將之劃分爲訓練資料(2095篇)及測試資料(211篇)。

3. 根據次數、集中度以及廣度三條件,從訓練資料中找出雙連字串,並指定此雙連字串在各類的分類比重。

4. 分別利用向量模式及機率模式,不同的分類比重給定方法來做自動分類,並針對錯誤部分做進一步的探討。

2 實驗方法

2.1 資料選取及類別選定

本實驗所採用的文件是工商時報80年7月到81年1月的新聞報導,每8天取樣1天,共抽樣22天爲樣本。將前六個月,共20天的新聞報導(文件)爲訓練資料,最後一個月,共2天的新聞報導當做測試資料。這些文件(新聞報導)是存放在電腦內,以BIG-5碼儲存。

認否概一面方司公？丙墊、稅漏逃
證蒐券證元大赴今局查調
針對調查局約談大元證券涉嫌逃漏稅及從事丙種墊款一案，台灣證券交易所指出，將於十日上午派員赴大元了解案情，並針對報載事項要求該券商出具書面說明，同時查核公司的帳冊，以及經紀部門營運是否正常。
由於調查局於昨日早上鎖定大元證券，追查該券商涉嫌逃漏證所稅與丙種墊款情事，並約談公司高級主管。交易所稽核完昨日以電話查詢，大元證券向交易所的說明是僅該公司總經理鄧國泉被調查局約談，對調查局所指陳的逃漏稅及墊丙事項，大元證券一概否認。

圖2-1 新聞報導的例子

此外，這些資料中，也會有一些雜訊(noise)，這些雜訊主要來自兩方面：第一是標題，標題有右至左以及左至右；第二則是錯字。這些雜訊，自然多少會影響到後來關鍵詞的選取工作。至於存放電腦中的新聞報導，是20字一行，經過處理後，平均一篇報導有31.13行；此外，2095篇訓練資料中，一共有943961個中文字(3569個不同的中文字)，852387個中文Bigram(40085個不同的中文Bigram)。

至於類別，我們採用日本產經新聞的分類方式，採用的原因請參閱[Che92, p30]。該分類法共分22大類、150小類。我們取其中五大類來做實驗，而這五大類中，又共分爲45小類，我們將新聞報導數小於10篇的類別刪去，因爲數量太少時，很難去做些有意義的統計或計算；這樣子剩下24小類。如此，本實驗所選定的類別即爲此24類，訓練資料共有2095篇屬於此24類的新聞報導，測試資料則有211篇新聞報導。

2.2 關鍵詞(雙連字串)的選取

篩選雙連字串(Bigram)，我們設了三個條件：(1)次數(2)集中度(3)廣度。

我們採用雙連字串，而沒有加入斷詞系統，主要是基於下列幾點考量：

1. 用雙連字串代價比較低；
2. 針對專有名詞以及縮寫詞，目前的斷詞系統並不能有效地解決；

3. 根據[Che92],我們發現,利用雙連字串當做關鍵詞與人工挑選關鍵詞所得的實驗結果,正確率(召回率)相去不遠。

與人工選取關鍵詞相比較,人工挑選出來的關鍵詞可能都比較具有意義(就人的觀點而言),然而,人工挑選永遠會遇到一個問題,就是不同的人會做出不同的結果,可能比較不客觀;另外,人覺得有分類意義的詞彙,不見得都是很有分類價值的詞彙。我們希望能夠做到完全自動化。

就分類系統而言,一個具有分類價值的關鍵詞,應該滿足下列三條件:

1. 次數要夠:雙連字串並非都是一個有意義的詞,任意相鄰的兩個中文字即可形成一個雙連字串;通常,一個不具意義的雙連字串出現的次數不會多,如果定一個界限值,出現次數低於此界限值者就去掉,則那些無意義的雙連字串大部分都會被摒除在外了。

2. 集中度:一個有分類價值的雙連字串,應該要集中出現在某一類或某幾類中,而不是平均分佈在各類中。

3. 廣度:在某一類頻頻出現的雙連字串,如果它出現在這類中許多篇文件裡,則它愈具有分類的價值,相反地,若此雙連字串,雖然出現次數夠多,但是卻只集中在某幾篇文件中,這種雙連字串的出現,原因可能爲某一突發事件,或是撰稿者特殊的寫作風格所致,而這種雙連字串,其分類的價值相對上就小多了。

接下來,便根據這三個標準,逐一來篩選此分類系統所需要的雙連字串。

關鍵字串的篩選方法,第一步,以出現次數爲篩選的標準。

在[XCYC92]的實驗中,在做五大類的分類實驗,採用1,2,3,5,10,15,20等不同界限值,而界限值定爲5時,得到較好的結果,因此以5次爲篩選基準。在40085個雙連字串中,共有17825個雙連字串符合此條件而留下來。

第二步,利用Entropy公式來做篩選,以符合集中度的要求。對於一個雙連字串T_i,T_i的Entropy值爲:

$$H_i = -\sum_{j=1}^{24} p_{ij} \log \frac{1}{p_{ij}}$$

其中,$p_{ij} = \frac{d_{ij}}{\sum_{j=1}^{24} d_{ij}}$,$d_{ij}$爲類別$C_j$中出現$T_i$的文件數。

H_i的值介於0(最集中)與$\log 24$(最分散)之間,如果T_i只出現在某一類中,則H_i的值爲0,若平均分散在各類,即$p_{ij} = \frac{1}{24}$,H_i的值爲$\log 24$。我們定的Entropy

界限值爲$\log 2$ $(= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2})$,大於界限值之雙連字串則予以捨棄。$\log 2$的意思是:一個詞彙平均分佈在兩類中,只要比這種情形還不平均的,就會被留下來。

第三步,要訂一個公式來篩選雙連字串,以符合廣度的要求。對一個雙連字串T_i,T_i的廣度定爲:

$$Value(T_i) = \max_j(\frac{d_{ij}}{t_{ij}} \times \frac{d_{ij}}{\sum_{j=1}^{24} d_{ij}})$$

其中,d_{ij}爲類別C_j出現T_i的文件數,t_{ij}爲T_i出現在類別C_j的次數。此界限值定爲0.2,小於界限值0.2則予以捨棄。

經過這三個步驟的篩選,共得到了5579個雙連字串符合分類系統的要求,這5579個雙連字串,我們將之視爲關鍵詞。

2.3 由雙連字串找到多字串

我們另外做了一個實驗,從雙連字串中找出多字串(N-gram)。簡而言之,這裡所找的多字串$A_1A_2\ldots A_N$,其中A_1A_2及$A_{N-1}A_N$必定是原先的關鍵詞,而多字串中間的雙連字串則不一定是關鍵詞。而這個N-gram,也許只是一個M字詞$(M \geq N)$的子字串。尋找多字串的目的,在於刪減多餘的雙連字串。以機率模式的角度來看,在機率模式中,公式是假設各關鍵詞互相獨立(Independent)而推導出來的,假設一個詞彙ABCD被切成AB、BC、CD三個關鍵詞,則這三個關鍵詞可能是相依的(Dependent)。用多字串可使理論更周延[Yan93]。

我們根據前面找出來的雙連字串關鍵詞,逐步接成三字串(Trigram),四字串(4gram),...,一直到十四字串,把可能成爲關鍵詞的N-gram都挑出來,接著再以長詞優先的原則,經過次數、集中度、廣度三條件的篩選,最後得到4711個多字串$(2 \leq N \leq 14)$[YaZC93]。

五字串候選者(部分)			八字串候選者
大盤成交量	正所得稅法	股上週股價	各規格雞種每台斤
及合作金庫	交量無法放	股及金融股	例每千股無償配發
引市場實戶	交量萎縮至	股無償配發	券副總經理陳文鋒
引投資性買	各規格雞種	股價平均下	益證券公司自營部
日均線反壓	向交易所申	股價平均上	商時報股價平均下
加權指數上	在投資性買	股價平均則	商時報股價平均上
可逢低承接	自辦融資融	活期儲蓄存	商時報股價平均則
可無償配發	呈多頭排列	苯乙烯單體	資證券及其結匯辦
司六月份營	呈相對強勢	時報股價平	僑及外國人投資證
失望性賣壓	例每千股無	破十日均線	證證券公司自營部

表2-1 部分的五字串候選者及八字串候選者

結果請參看表2-2, 表2-2列出14字串到雙連字串候選者數量以及最後成爲關鍵詞的數量。這樣子做, 從原來5579個雙連字串關鍵詞, 下降到有4711個多字串關鍵詞。

N-gram	候選者數量	關鍵詞數量	N-gram	候選者數量	關鍵詞數量
十四字串	1	1	七字串	11	6
十三字串	1	1	六字串	25	15
十二字串	2	0	五字串	50	30
十一字串	2	1	四字串	216	152
十字串	2	1	三字串	711	443
九字串	4	1	雙連字串	5579	4053
八字串	10	7			

表2-2 N-gram候選者及關鍵詞數量

2.4 類別的向量表示法

在資訊檢索過程中, 一篇文件通常以此一文件所包含的關鍵詞的存在向量(Term Vector)來代表此一文件, 因此代表某一類的向量, 也可以用此類中所有文件的關鍵詞的存在與否來決定, 唯其分量值, 不應考慮以0表示不存在1表示存在此一關鍵詞。

在機率模式中，只採用原始分類比重；至於向量模式，除了原始分類比重以外，還採用兩種標準化的方式。

1.原始分類比重: 原始分類比重的給定，基本上是根據此關鍵詞在各類的分佈情形。假設關鍵詞T_i在類別C_j中的分類比重爲x_{ij}，則

$$x_{ij} = \frac{\frac{t_{ij}}{len_j}}{\sum_{j=1}^{24} \frac{t_{ij}}{len_j}}$$

其中，t_{ij}爲關鍵詞T_i在類別C_j中出現的次數，len_j爲訓練資料中屬於類別C_j的新聞報導總數。

2.第一種標準化: 在原始分類比重的方式中，以向量X_j來表示類別C_j，在第一種標準化中，以向量$U_j = (u_{1j}, u_{2j}, \ldots, u_{mj})^T$來表示類別$C_j$，其中，$\|U_j\| = 1.00$，即$U_j$爲單位向量，且$u_{ij} = \frac{x_{ij}}{\|X_j\|}$。

此種標準化的想法是，將代表類別的向量變爲單位向量，以使每一類能夠「公平競爭」。

3.第二種標準化: 在第二種標準化中，以向量$V_j = (v_{1j}, v_{2j}, \ldots, v_{mj})^T$來表示類別$C_j$。假設$\|V_j\| = \sqrt{DN_j}$，其中，$DN_j$爲訓練資料中屬於類別$C_j$的新聞報導數。

檢視第一種標準化的方法，我們發現到有一個很嚴重的問題，亦即每一類的文件數目並不一樣，基於此，如果我們硬將代表每一類別的向量都弄成一樣長度，對文件數量多的類別並不公平。代表此類的向量長度會因文件數的增加而遞增，我們不知道確切的函數關係，但是這關係不會是線性關係，於是利用根號函數來逼近。

我們將利用這三種給定分類比重的方式來做文件自動分類實驗。

3 自動分類實驗結果

在本節中，我們將列出，分別採用向量模式和機率模式，其分類的實驗結果。

3.1 向量模式與機率模式實驗結果

類別	訓練資料召回率				測試資料召回率			
	文件數	原始分類比重	第一種標準化	第二種標準化	文件數	原始分類比重	第一種標準化	第二種標準化
J0201	56	92.45%	92.45%	92.45%	6	50.00%	66.67%	50.00%
J0202	11	100.00%	100.00%	100.00%	2	0.00%	0.00%	0.00%
J0203	64	100.00%	100.00%	100.00%	8	62.50%	87.50%	75.00%
J0204	59	96.61%	98.31%	96.61%	5	40.00%	60.00%	40.00%
J0205	40	100.00%	100.00%	100.00%	5	40.00%	40.00%	40.00%
J0206	285	93.68%	84.91%	95.44%	24	83.33%	58.33%	87.50%
J0207	699	96.71%	82.69%	95.42%	61	90.16%	70.49%	86.89%
J0208	33	100.00%	100.00%	100.00%	8	100.00%	100.00%	100.00%
J0209	24	100.00%	100.00%	100.00%	0	?.??%	?.??%	?.??%
J0211	116	87.83%	87.83%	89.57%	14	21.43%	28.57%	21.43%
J0301	112	84.91%	82.08%	93.40%	6	16.67%	16.67%	16.67%
J0302	32	93.10%	96.55%	93.10%	2	0.00%	50.00%	0.00%
J0303	211	94.29%	94.76%	94.76%	14	71.43%	85.71%	71.43%
J0305	36	97.06%	97.06%	94.12%	25	60.00%	64.00%	76.00%
J1008	40	100.00%	94.74%	100.00%	6	33.33%	33.33%	33.33%
J1009	24	90.00%	90.00%	85.00%	0	?.??%	?.??%	?.??%
J1012	44	100.00%	97.73%	97.73%	5	80.00%	80.00%	80.00%
J1103	20	90.00%	95.00%	95.00%	1	0.00%	0.00%	0.00%
J1105	12	100.00%	100.00%	100.00%	0	?.??%	?.??%	?.??%
J1201	30	93.33%	96.67%	93.33%	5	60.00%	60.00%	60.00%
J1202	58	85.96%	84.21%	87.72%	4	25.00%	75.00%	50.00%
J1203	29	91.67%	91.67%	91.67%	4	0.00%	0.00%	33.33%
J1204	29	88.46%	88.46%	88.46%	4	0.00%	0.00%	0.00%
J1205	31	96.55%	96.55%	93.10%	2	50.00%	50.00%	50.00%
總計	2095	94.57%	88.50%	94.86%	211	64.29%	60.95%	67.14%

表3-1 5579個雙連字串在向量模式中各種方法的分類實驗結果

	訓練資料召回率			測試資料召回率		
	原始分類比重	第一種標準化	第二種標準化	原始分類比重	第一種標準化	第二種標準化
5579個雙連字串	94.57%	88.50%	94.86%	64.29%	60.95%	67.14%
4711個N-gram	94.53%	86.73%	94.58%	59.90%	56.04%	61.35%

表3-2 5579個雙連字串與4711個N-gram在向量模式中實驗結果的比較

	訓練資料召回率			測試資料召回率		
	第一名	前二名	前三名	第一名	前二名	前三名
向量模式	94.86%	99.56%	99.95%	67.14%	76.67%	82.86%
機率模式	97.23%	99.90%	99.95%	59.52%	74.76%	79.52%

表3-3 向量模式與機率模式實驗結果的比較

請參看表3-1 、表3-2及表3-3 。簡單說,根據實驗的結果,我們可以發現,利用第二種標準化方式,測試資料的召回率可以達到67.14%,如果取前三名,則更可以達到82.86%;至於採用4711個N-gram關鍵詞,則沒有得到較好的結果,無論在向量模式或是機率模式的情形下;此外,就測試資料而言,我們利用機率模式所得的結果比向量模式稍差一些。

3.2 其它相關的實驗結果

1. 用每篇新聞報導的前面幾行當做文件替代品

   通常一篇報導,重要訊息應該會傾向集中在報導的前面幾行;事實上,人工在做分類的工作時,通常也都是只看前面數行就決定了此篇新聞報導的類別。如果是這樣,在做分類時,是不是可以只取每篇新聞報導的前面數行當做文件替代品(Profile),在不影響分類正確率的前題下,減少電腦記憶儲存的負荷及計算時間?

   現在,我們做一個實驗,只取前20行、前15行以及前10行來做看看,比較正確率有甚麼改變,請參看表3-4 。

所取用的行數	關鍵詞數量	訓練資料召回率			測試資料召回率		
		第一名	前二名	前三名	第一名	前二名	前三名
整篇報導	5579	94.86%	99.56%	99.95%	67.14%	76.67%	82.86%
前20行	4008	92.46%	98.49%	99.32%	62.65%	72.37%	77.82%
前15行	3451	85.34%	93.90%	95.57%	62.55%	76.83%	80.69%
前10行	2445	82.72%	91.81%	94.13%	60.54%	73.56%	79.69%

表3-4 只取前幾行來做分類所得的實驗結果(向量模式、第二種標準化)

實驗結果,就測試資料而言,整篇報導與前20行比較起來,還有將近五個百分點的差距。或許這樣的結果表現出了一個事實,即我們收集到的這些新聞報導,其訊息可能平均地分佈在報導中,並沒有特別集中在前幾行。

2. 減少訓練資料數量所得實驗結果

我們認爲訓練資料嚴重不足,這部分的實驗,就是想要試著來估計大概要多少篇新聞報導才達到飽和;如果沒有辦法估計出來,也想試著以具體的數據來說明目前的訓練資料確實不足。

目前所用的訓練資料,時間是從80年7月到80年12月;現在,先扣除7月份的訓練資料然後來做實驗,再扣除8月份的訓練資料來做實驗,一直扣除到只剩下12月一個月份的訓練資料來做實驗,讓我們來比較其正確率的變化,至於測試資料則同樣是用81年1月。請參看表3-5的實驗結果,同樣是採用向量模式、第二種標準化的方法。

所用訓練資料月份	關鍵詞數量	訓練資料召回率			測試資料召回率		
		第一名	前二名	前三名	第一名	前二名	前三名
7～12月	5579	94.86%	99.56%	99.95%	67.14%	76.67%	82.86%
8～12月	5085	94.67%	99.41%	99.85%	61.98%	74.14%	79.09%
9～12月	4344	96.09%	99.74%	99.87%	61.69%	77.39%	82.76%
10～12月	3379	97.09%	99.90%	100.00%	59.77%	71.26%	78.93%
11～12月	2379	98.16%	99.54%	100.00%	53.82%	67.18%	74.81%
12月	1297	99.62%	100.0%	100.00%	53.64%	67.43%	71.65%

表3-5 減少訓練資料數量所得的實驗結果(向量模式、第二種標準化)

就訓練資料而言,當訓練資料量愈少,正確率(召回率)會愈高是很合理的,因爲資料量愈少,一個關鍵詞跨多類的機會相對減少。做此實驗,是希望能找到一個點,到了這個點以後,就算再增加訓練資料的數量,也不會提高測試資料的正確率,顯然我們並沒有找到這個點,所以訓練資料應該確實是不足夠的。

3. 將出現關鍵詞的文件數定界限值

在選取關鍵詞時,一共定了三個標準,分別是詞彙次數要夠、分佈特別集中於某些類以及在同一類中要儘量分散在各篇報導中。其中最後一項定0.2爲界限值,雖然有因此而去除掉一些詞彙,但是可能標準定得太鬆了一些,導致實際上有許多出現5次且集中在一篇新聞報導的詞彙,剛好在界限值的邊緣而被「抓」進來成爲關鍵詞。

不管界限值0.2是不是定得太鬆,其實還可以用此詞彙出現的文件數來定界限值,如果定了界限值,又能不影響到分類結果的正確率,則可以減少關鍵詞的數量。

我們將文件數的界限值分別定爲1(就是沒有設限)、2、3及5,實驗結果請參看表3-6。

文件數界限值	關鍵詞數量	訓練資料召回率			測試資料召回率		
		第一名	前二名	前三名	第一名	前二名	前三名
1	5579	94.86%	99.56%	99.95%	67.14%	76.67%	82.86%
2	5432	94.47%	99.47%	99.90%	66.67%	77.14%	83.33%
3	4843	93.53%	99.17%	99.85%	64.59%	75.12%	82.30%
5	3375	91.06%	98.62%	99.44%	62.44%	73.17%	78.05%

表3-6 定文件數爲界限值所得的實驗結果(向量模式、第二種標準化)

從上面的實驗結果告訴我們,當在挑選關鍵詞,其實我們還可以定文件數的界限值,出現在兩篇或兩篇以上的新聞報導中才要,只出現在一篇新聞報導中的可以捨棄,正確率幾乎沒有甚麼差別。

4 錯誤分析

實際來檢視電腦分錯的例子,大致上,我們可以歸納爲兩點:(1)類別相近或是報導本身模稜兩可;(2)此篇報導的關鍵詞數量很少。

關於第一點,若我們直接看類別名稱,就可以發現一些較相近的類別,例如J0206金融、J0207證券;例如J1103電子材料、J1203硬體···等等。

現在我們將類別C_j以向量$X_j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$表示,如果關鍵詞$T_i$在類別$C_j$的分類比重不是0,則$x_{ij}$=1,否則$x_{ij}$=0,類別$C_i$與$C_j$的相似性,我們用$X_i$與$X_j$的餘絃(Cosine)來表示,即$\frac{X_i \cdot X_j}{\|X_i\| \|X_j\|}$。我們將相似性超過15.0%的類別列出:

J0207(證券)	J0303(業績財務)	0.39
J0206(金融)	J0207(證券)	0.36
J0204(貨幣流通)	J0205(物價)	0.20
J1202(電腦)	J1203(硬體)	0.19
J0207(證券)	J0209(商品行情)	0.18
J1103(電子材料)	J1203(硬體)	0.16
J0207(證券)	J0301(經營總論)	0.15

對於某些類別相近這樣的問題,若要增進自動分類的正確率,大致有幾種方法:

1. 重新調整類別，將較相似的類別合併。

2. 採用重覆分類，如果一篇新聞報導，經過電腦計算後，有兩類的分數都很高，則電腦將此篇報導同時指定在這兩類。當然，如果這樣的話，人工在指定類別時也要考慮將某些報導同時分爲兩類甚至三類。

另外，讓我們來考慮，電腦與人工在做分類及挑選關鍵詞時，兩者之間的差異。挑選關鍵詞時，人所挑出的關鍵詞，大致上具有分類的意義，而電腦所挑出的關鍵詞，大致上具有分類的價值，但是不一定都具有意義（就人的觀點而言）。一個有分類意義的詞彙，可能因爲次數太少而使其分類的價值減低，也可能因爲出現在太多不同類別，而失去分類的價值。反過來說，有分類價值的詞彙，都特別集中在某一類或某幾類的詞彙，在本實驗中，因爲是用雙連字串，找出來的雙連字串，可能是詞彙的一部分，但也可能因爲包含功能詞（Function Word）或是附著語（Bound Word）而看起來不具甚麼意義。若要改進這些詞彙的「品質」，利用斷詞系統以及詞性標示等技巧來過濾掉這些「不好」的雙連字串是可能的解決方式。

至於分類的做法方面，電腦在做分類時，是以關鍵詞當做線索，而人在做分類時，可能是以某一句話或是某幾句話當做線索，這樣說來，電腦做分類的結果，可能沒有辦法達到和人一模一樣，因爲重點的這幾句話如果沒有出現關鍵詞，電腦就可能會分錯。基本上，人在做分類時，利用的是「概念」，而比較不是因爲看到了某些關鍵詞而決定要將之分到哪一類。在大部分的情形下，這段概念包含一些關鍵詞，但是有些情形下，概念中沒有任何的關鍵詞。另外一點，一篇新聞報導可能會提到很多東西，人很容易看出甚麼是主題甚麼是次要的部分，但是我們顯然還沒有賦予電腦這個能力。通常，主題會寫在一篇報導較前面的部分，然而也可能整篇都在談主題。

當然，電腦分類的最大好處是每次做的結果都會一樣，有一致性，人工做分類的話，不同的人會做出不同的結果，甚至同一人在不同時間做也可能會有不同的結果。

5 結論及未來方向

針對這些新聞報導，實驗結果顯示，在我們的做法中，向量模式比機率模式的結果好一些，第二種標準化（代表此類別的向量長度正比於此類別訓練資料數量開根號）方式比原始分類比重結果好一些。

另外，利用5579個雙連字串關鍵詞，會比採用4471個N-gram關鍵詞所得的結果還要好些；但是就人的觀點而言，我們會覺得4711個N-gram關鍵詞比較有意義，且就理論層面而言也較健全些。

就測試資料而言，召回率大約在67%，平均三題對兩題，結果並不盡理想，究其主要原因，應是受限於訓練資料的不足，另外，有些問題出在人所指定的類別本身，以及有些報導本身確有模稜兩可的情形發生。　　67%的召回率（正確率），尚不足以完全取代人工，然而電腦分出來的結果有絕對的一致性，不像人工分類的結果常因人因時而會有所差異。　　若取分數最高的前三名，召回率有83%左右。如果電腦先找出前三名，交給人來看，再由人來做分類的工作，這樣的話，應該多少還是能夠減輕人工的負擔。

未來，我們希望能夠再做下列的工作：

1. 增加訓練資料數量：在前面曾經提到過，有些錯誤發生的原因應該歸咎於訓練資料數量的不足，實際上，有些類別只有十幾篇，要用這十幾篇新聞報導來代表這一類別，並不是很妥當。我們希望再繼續做些取樣，訓練資料增加，應該可以得到更好的結果。

2. 加入自然語言的知識：基本上，到現在爲止我們的做法，主要是用統計的方法，並沒有善用電腦在自然語言處理上所獲致的成果，例如斷詞、自動詞性標示（Automatic Tagging），例如Mutual Information …等等。
爲了提高關鍵詞的品質，加入斷詞應該是有必要的，但是得克服諸如專有名詞、縮寫詞…等等的問題。另外，規則運算式似乎也可以考慮加入，例如「昨日升值」、「前日升值」兩個詞彙，就分類的意義應該是相同的，那麼，可以考慮以「？日升值」的方式來儲存此詞彙。

3. 利用類神經網路的技巧來給定分類比重：我們想引進類神經網路的技巧，經由訓練來決定分類比重，看看能不能得到不同的結果。

4. 重覆分類：現實生活中，很容易發現一些東西難以歸類，本實驗中，有些新聞報導確實也有類似的情形，因此重覆分類可能較符合實際需要。

5. 類別選定：我們的目的，在於利用電腦來取代人工，以節省人力。若要取代人工，則得先設法使電腦分類結果與人工分類結果的一致性提高，但是經由分析，有些錯誤應該是出在類別選定的問題，如果根據電腦分類的結果，再回頭檢視人所規定的類別，做個折衝，在人覺得有意義的條件下，重新調整類別，召回率提高，才更有可能達到目標。

誌　　謝

本論文的完成，要感謝中國時報社所提供的工商時報新聞報導資料；另外，中央研究院資訊所助研究員簡立峰博士以及清華大學資訊所張俊盛教授提供了不少寶貴的意見，在取一併誌謝。

附　　錄

五大類及24小類類別名稱：

類別		訓練資料篇數
J02	經濟、產業	
J0201	產業總論	56
J0202	財政	11
J0203	稅制	64
J0204	貨幣流通	59
J0205	物價	40
J0206	金融	285
J0207	證券	699
J0208	保險	33
J0209	商品行情	24
J0211	業者動態	116
J03	經營、企業	
J0301	經營總論	112
J0302	企業組織	32
J0303	業績財務	211
J0305	商品流通	36
J10	機械、器具、設備	
J1008	產業機械、半導體	40
J1009	事務機械	24
J1012	四輪配備	44
J11	電子、電機	
J1103	電子材料	20
J1105	家庭電器	12
J12	情報、通信	
J1201	資訊總論	30
J1202	電腦	58
J1203	硬體	29

J1204	系統軟體	29
J1205	通訊設備	31

參考文獻

[Che92] 陳淑美.財經新聞自動分類之研究.台大圖書館學研究所碩士論文,台北.1992年12月.

[XCYC92] 謝清俊,陳淑美,楊允言,陳克健.Autoclassification of Texts.如何利用大型語料庫作研究研討會.計算語言學會,台北.1992年9月.

[Yan93] 楊允言,張俊盛,陳克健.文件自動分類及其相似性排序.清華大學資訊科學研究所碩士論文,新竹.1993年6月.

[BHMP92] Blosseville M.J., Hebrail G., Monteil M.G. and Penot N. " Automatic Document Classification : Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together ". in SIGIR '92 : Proc. of the 15th Ann. International ACM *SIGIR Conf. on R. and D. in Inform. Retr.*. pp51-57. Denmark : Copenhagen. Jun.21-24 1992.

[BoBe63] Borko, Harold and Bernick, Myrna. " Automatic Document Classification ". *J. of the ACM* 10(1) pp151-162. 1963.

[HaZa80] Hamill, Karen A. and Zamora Antonio. " The Use of Titles for Automatic Document Classification ". *JASIS* 31(6) pp396-402. Nov.1980.

[Jac92] Jacobs, Paul S. " Joining Statistics with NLP for Text Categorization ". in *Third Conference on Applied Natural Language Processing — Association for Computational Linguistics* pp178-185. Italy : Trento. 31Mar.-3Apr. 1992.

[Jac93] Jacobs, Paul S. " Using Statistical Methods to Improve Knowledge-Based News Categorization ". *IEEE Expert* 8(2) pp13-23. Apr. 1993.

[Kwo75] Kwok, K.L. "The Use of Title and Cited Titles as Document Representation for Automatic Classification ". *Inform Proc. and Manag.* 11 pp201-206. 1975.

[Lar92] Larson, Ray R. " Experiments in Automatic Library of Congress Classification ". *JASIS* 43(2) pp130-148. 1992.

[Lew92] Lewis, David D. " An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task ". in *SIGIR '92 : Proc. of the 15th Ann. International ACM SIGIR Conf. on R. and D. in Inform. Retr.*. pp37-50. Denmark : Copenhagen. Jun.21-24 1992.

[Mar61] Maron, M.E. " Automatic Indexing : an Experimental Inquiry ". *J. of the ACM* 8 pp404-417. 1961.

[Sal86] Salton, Gerard. " Another Look at Automatic Text-Retrieval Systems". *Comm. of the ACM* 29(7) pp648-652. Jul.1986.

[Smi89] Smith, Stephen Ray. *An Advanced Full-Text Information Retrieval System.* PhD thesis, Computer Science Dept.; Univ. of Alabama in Huntsville. 1989.