階層式文件自動分類之特徵選取研究

柯淑津
東吳大學資訊科學系
ksj@volans.cis.scu.edu.tw

陳振南
銘傳大學資訊管理系
jnchen@mcu.edu.tw

摘要

文件分類（Text Categorization）是指針對一組事先設定好的類別集，透過特徵選取的作法，將自然語言文件標上適當的主題類別。文件分類的應用範圍非常廣泛，包括：電子郵件與新聞過濾、資訊檢索、自動索引、以及詞彙語意解析等等。

有關文件分類的研究，常由文件內容中抽取重要的特徵（feature）來代表這個文件，而特徵抽取的來源包羅萬象，可以簡單地從文件作者、出版機構著手，或是由蘊含豐富資訊的語言結構來作為抽取文件特徵的依據。先前的研究通常只由歸屬同類別的文件選出特徵集，很少將類別間是否具相關性納入考慮，而且當選完特徵後通常不再加以變動。這樣的作法對於線性分類或許是可行的，若是應用於階層式分類便顯得不恰當。

本研究提出一個適用於階層式文件自動分類系統的特徵選取方法，經初步選完特徵集後，再依各特徵與相近類別間所具的分類意義做適當的調度。我們以『財經記事』的新聞資料進行分類實驗，結果驗證系統的強健性。另外，也得到下列幾個結論：（1）少的特徵數目有利於分類的進行，（2）階層式分類優於線性分類，（3）適當的特徵選取將更凸顯階層式分類的效能。

1. 簡介

自動文件分類（Text Categorization）是指針對一組事先設定好的類別集，透過自動化的作法，將自然語言文件標上適當的主題類別。文件分類的應用範圍非常廣泛，譬如電子郵件與新聞過濾（E-mail and News filtering）、資訊檢索（Information Retrieval）、自動索引（Automatic Indexing）、詞彙語意解析（Word Sense Disambiguation）等等。

在網際網路的蓬勃發展下，資訊的傳播沒有國度、時間的限制。Internet 持續地累積多樣化的資訊，已經形成一個巨大、分散的多媒體。然而這些大批的文件資料若是未能做妥善的整理，對個人或者資訊服務系統而言都將造成資訊氾濫。文件分類的技術正是解決這個難題的利器。現行著名的網路資訊服務系統，如：Yahoo (http://www.yahoo.com/)以及 Virtual Library (http://vlib.stanford.edu/) 便提供這類的服務。他們將蒐集到的網路文件組織成適當的結構，像是依照文件的地理區域、出版時間、出版機構、或是內容主題等，對文件加以分類（林頌堅，1998）。透過這些分類資料，使用者可以根據其需求，自系統中快速地選取相關資訊。

一般而言文件分類的建構方式可分為兩種：（1）由機器自動學習，（2）以人工對文件進行主題標示。由機器自動學習的作法中可歸納為兩類，督導式（Supervised）及非督導式（Unsupervised）學習。督導式的學習是由使用者或一些專家先對部分文件進行分類，然後再將分類好的文件作為自動分類系統的訓練資料。通常這種督導式學習的分類效果較非督導式為佳（Lewis, 1996）。至於，以人工對文件進行主題標示的工作，則常需仰賴所謂的分類專家，如圖書館員或是專業領域中的專家學者，以其專業知識對文件進行分類。這樣的作法，雖然可以得到較準確的結果，卻要付出相當的時間與人力。面對現在資訊爆炸的時代，我們急需一個良好的自動處理技術與工具來進行文件分類工作。

本研究提出一個適用於階層式文件自動分類系統的特徵選取方法，經初步選完特徵集後，再依各特徵與相近類別間所具的分類意義做適當的調度。我們以『財經記事』的新聞資料進行分類實驗，結果驗證系統的強健性。

本文的其他部分結構如下：在第二節中介紹先前有關文件分類的研究，第三節是我們對於資料所做的一些觀察，第四節提出適用於階層式文件分類的演算法，接著是實驗描述與結果討論，最後，提出結論與探討未來的研究方向。

2. 先前有關文件分類的研究

先前有關文件分類的研究，常由文件內容中抽取重要的特徵（feature）來代表這個文件，而特徵抽取的來源包羅萬象，可以簡單地從文件作者、出版機構著手（Blosseville, et al., 1992; May, 1997），或是由蘊含豐富資訊的語言結構：語彙(Lexical)、語法

(syntactic)、或是語意(semantic)等資訊來作為文件特徵抽取的依據。

文件的語彙資訊是語言結構中最容易抽取的特徵內容，常見的有：字、詞、片語等單位，有些研究利用語詞出現在文件中的頻率值(tf, term frequency) (Frakes and Baeza-Yates, 1992; Witten, Moffat and Bell, 1994)做為文件的特徵(Salton 和 McGill, 1983)，較常見的是除了頻率值外，再加上語詞本身的重要性這個考量，即是以各語詞的 tf x idf (inverse document frequency, Witten, Moffat and Bell, 1994)值所組成的向量做為文件的特徵(Salton 和 Buckley, 1988)。另外，有些研究人員認為所有的語詞都併入特徵值的處理並不恰當，他們建議以統計方法 χ^2 檢定來選取重要的語詞當作文件的特徵(Watanabe et a1., 1996; Ng, Goh and Low, 1997)。

在文獻中以語詞所含的語意代替語詞本身來設定文件特徵的作法有下列幾個，Liddy 提出的 DR-LINK 系統（Liddy et al., 1993）利用朗文機讀字典將文件中的每個語詞轉換成主題碼(SFC-Subject Field Code)，若有歧義情形發生時再依句子中 SFC 的分佈狀況等資訊，設定出合適的 SFC 碼（Liddy, Paik and Yu, 1994）。最後，以經正規化後的 SFC 向量來表示文件特徵。另外，Schütze 等人提出的隱含語意索引(LSI - Latent Semantic Indexing)，將文件看成為空間上的一個特徵向量，藉著分析語詞共生模式(word co-occurrence pattern)，再利用奇異值分析(SVD – Singular Value Decomposition)的技巧，將高維度的向量轉化成為一個具較低維度的向量(Schütze, Hull and Pedersen, 1995)，他們的實驗證實了這個方法的有效性，尤其是在減少計算量的部分。

在文件分類的處理中以語詞為單位來粹取文件特徵的研究方法，很明顯地存在著下述幾個缺點：同義字問題、一詞多義問題、參數數量問題、以及多字詞問題等等。另外，Yang 和 Chute 他們觀察到以同義詞典為主的分類方法，往往因為一般的同義詞典所涵蓋的字不足以應付各種不同領域需求(Yang and Chute, 1994)，因此，利用同義詞典將語詞轉化成為主題(Subject)的有關研究，相較於直接用語詞當作文件特徵的作法，並無法得到較佳的精確度。他們認為存在於文件中的自由文體與同義詞典中的控制詞彙間的詞彙漏洞(vocabulary gap)，可以利用人類知識來加以彌補。一者是利用先前由人工對應過的訓練資料，或者由相關性回饋(relevant feedback)的技巧來收集資訊。

過去有關分類的研究，證實分類結果的精確度與召回率常隨著類別個數增加而降低（Apte, Damerau and Weiss, 1994; Yang, 1996）。而階層式分類由樹根開始，在每一節點只需考慮往其子節點細分，因此，在處理過程的每一階段，所需面對的類別個數較少，這是階層式分類往往有較佳效果的原因之一。另外，隨著樹的階層數（level）遞增，所處理的文件常設定在愈來愈窄的特殊範疇（specific domain），此時詞彙的歧義程度較易規範（D'Alessio et al., 1998），這是階層式分類的另一個優點。

3. 階層分類的特徵詞彙調整

經觀察『財經記事』新聞資料後，我們發現存在著這個現象：樹狀結構中距離相近的類別共用特徵詞彙。相似的類別往往會共用特徵詞彙，這種情形尤其常出現在階層式分類系統中，被歸屬在同一大類別下的幾個細層類別，往往具有相當高的相似度。因此，它們會共用特徵詞彙，這種現象若不加以處理，將導致細層類別不易區分。如表一所示，『央行』、『交易』、以及『利率』等詞彙，同時以高頻率出現在金融篇的幾個中類別裏。

當文件被歸到階層式分類系統中的某一節點後，往下進行細層分類時，這些共用詞彙應被視為該節點的停用字（stopword）。各細層類別需靠它們之間相異的特徵詞彙來彼此競爭，因此，這些共用詞彙應自其特徵集中移除。

表一　共用詞彙分佈在細層類別的頻率值

詞彙	語料中出現頻率	中類別	中類別出現頻率
台幣	346	金融	65
		外匯	95
外匯	260	銀行	62
		外匯	65
央行	797	金融	392
		銀行	142
		外匯	77
交易	773	金融	126
		外匯	22
		股票	221
利率	844	金融	168
		銀行	337
		外匯	19

4. 階層式文件分類演算法

線性分類系統將各個類別之間的關係當成彼此完全獨立，而這種假設在階層式分類系統是不恰當的。因為在階層架構中擁有同一父節點的兄弟節點間的關係顯然較其他節點更為密切。因此傳統適用於線性分類的特徵選取方式，直接搬移到階層式分類系統，並不完全可行。

4-1 特徵詞彙選取

起始特徵詞彙選取

階層式分類系統的特徵選取處理，我們區分為葉節點(leaf node)與非葉節點(non-leaf node)兩個部分。對於葉節點的部分，我們利用訓練資料透過詞頻統計的方式，為每個分類選取適當的特徵詞彙集，並且依據每個特徵 f 與類別 c 間的關係強度，設定權重值 W(f, c)。本研究採用 tf-idf 來計算權重值（如公式１、２所示）。這些特徵分為正項特徵與負項特徵兩類，其中，正項特徵詞彙以高頻率出現於歸屬類別 c 的文件中，他們擁有正值權重。而負項特徵詞彙在類別 c 的文件中並不出現，而且以高頻率出現在與 c 擁有相同父節點的其他兄弟節點類別中，因此他們在類別 c 的權重值為負數。

至於，非葉節點 p 的特徵選取，我們以其所有子節點 c 特徵詞彙的聯集當為 p 的候選特徵集。節點 p 與候選特徵集合中的特徵詞彙 f 間的權重值 W(f, c) 為 p 的所有子節點 c 與 f 的權重值 W (f, c)的總和，如公式３所示。

$$W(f, c) = tf_{f,c} \times idf_f, \qquad \text{(當 } c \text{ 屬葉節點時)} \qquad (1)$$

$$idf_f = \log(\frac{T}{df_f} + 1), \qquad (2)$$

$$W(f, c) = \sum_{a \in c\text{的子節點}} W(f, a) \qquad \text{(當 } c \text{ 屬非葉節點時)} \qquad (3)$$

$tf_{f,c}$ ： 詞彙 f 出現在類別 c 中的頻率值，

df_f ： 詞彙 f 出現的類別數，

T ： 所有類別總數。

透過觀察我們瞭解階層式分類的特徵集，需依特徵所具的分類意義而做調整。對於分類意義的量化，我們以 Lin 在 1997 年的論文中所提出的分支比率－BR(Branch Ratio ）來決定（Lin, 1997），如公式 4 所示。當 BR 值愈小，代表該特徵之高權重普遍來自各個子節點，因此，此特徵對所有子節點而言不具分類意義。所以需將此特徵自其子節點的特徵集合中移除。反之，BR 值愈大時，表示此特徵值主要來自某個特定子節點，因此，我們必須保留此特徵不加以更動。

$$\mathrm{BR}(f,p)=\frac{\underset{c\in p\text{的子節點}}{\mathrm{MAX}}\ \mathrm{W}(f,c)}{\underset{c\in p\text{的子節點}}{\mathrm{SUM}}\ \mathrm{W}(f,c)} \tag{4}$$

4-2 相關函數計算

對於文件與類別的相關程度，我們仍以 tf-idf 的方式來計算，如公式 5 所示。其中，將文件 d 所含詞彙 f 與類別 c 計算權重 W(f, c)，再加總所有的 W(f, c)，即得它們的相關強度 R(c, d)。

$$\mathrm{R}(c,d)=\sum_{f\in d}\mathrm{W}(f,c)\times idf_f \tag{5}$$

至於當階層樹中各個類別的特徵詞彙集訓練完成後，面對一個待分類的文件 d，我們要如何決定 d 的歸屬類別呢？首先，我們由階層樹的樹根開始，往下計算位居第二階層的各個節點類別 c 與 d 的相關度 R(c, d)。此時，若是所有的 R(c, d)皆無法高過預先設定的門檻值 θ，則停止往下層分類的工作，並將文件 d 分派給父節點。若存在節點類別 c 使得 R(c, d)高過 θ，則將文件 d 分派給擁有最大 R(c, d)值的節點類別 c。然後照相同方式，繼續往下層節點進行分類。

5. 實驗設計及結果說明

5-1 實驗資料

本研究以『財經記事』中所含的財經新聞標題（見附錄）進行實驗，其內共含 132606 則新聞標題，其來源為民國八十一年間中國時報、工商日報、聯合報、民生報等各

報社之新聞標題。這些標題事先經人工標示採三層式分類，共分為金融、產業等九大類別，大類別下細分為 39 個中類別，中類別下又分小類別。本研究僅以大、中兩層類別進行實驗，其編碼方式以及類別內容等詳見附錄。

經審視資料後，發現有少數標題完全重複，另外，大部分的標題僅設定一個類別，少部分標題給定多種分類。在整個實驗中，為方便效能評估我們將重複標題去除，並且僅採用單一類別之標題，最後剩下 119845 則新聞標題。我們將這些標題以隨機方式取出 20% 作為測試語料，其餘的標題當訓練語料。

5-2 實驗設計

為驗證本研究提出的階層分類特徵選取方法之效能，特設計一連串實驗，以『財經記事』的新聞資料進行分類。其中，第一組實驗採線性方式分類，直接以第二層的中類別進行自動文件分類。而第二組實驗則為階層式分類，先將文件由根節點開始，分至合適的大類別，再由此大類別往下分派給所屬的中類別。每組實驗我們皆以不同的特徵個數, k, 來設定特徵詞彙集，並比較 k 值大小所造成的效果差異。

5-3 實驗結果

在效能部分，本研究採資訊檢索領域中最常用的精確率（Precision Rate）及召回率（Recall Rate）進行評估。在線性分類的部分，當選取 25 個特徵詞彙進行分類實驗時，我們可以得到 59.1%的精確率以及 77.8%的召回率。而隨著特徵詞彙數量的降低，我們發現實驗的效果愈來愈見提升（請詳見表二）。當每個類別只選取 10 個特徵詞彙時，可達到 61.0%的精確率及 81.0%的召回率。

階層式分類的結果列於表三，同樣地，我們可以發現小的特徵數，可以達到較佳的效果。當特徵詞彙的選取由 25 個降至 10 個時，召回率可以由 95.4%提升至 98.1%，同時，精確率也由 63.8%提升至 66.0%。這種現象是因為小的特徵集所含詞彙的岐義性較少，這結果同時也驗證了 Lewis 在 1992 年對路透社新聞進行分類的研究經驗（Lewis, 1992）。另外，值得注意的是在效果方面，階層式分類優於線性分類。而且，這種現象對不同大小的特徵集皆有一致的結果，這似乎驗證本研究所提出的特徵選取方法具強健性。

6. 結論與未來研究方向

本文提出一個適用於階層式文件自動分類系統的特徵選取方法，並且經由實驗證實它的強健性。另外，也得到下列幾個結論：（1）少的特徵數目有利於分類的進行，（2）階層式分類優於線性分類，（3）適當的特徵選取將更能凸顯出階層式分類的效果。

未來我們將嘗試以詞彙概念代替詞彙本身作為特徵的選取單位，探討這種作法對文件分類系統所帶來的衝擊。另外，將特徵出現在文件的位置併入計算權重值的考量，這些研究方向對分類效果應該會有更上一層的空間。

表二　線性分類的實驗結果

特徵詞彙數	召回率	精確率
10	81.0 %	61.0 %
15	80.0 %	59.4 %
20	79.8 %	59.0 %
25	77.8 %	59.1 %

表三　階層式分類的實驗結果

特徵詞彙數	召回率	精確率
10	98.1 %	66.0 %
15	96.5 %	64.4 %
20	96.4 %	63.7 %
25	95.4 %	63.8 %

致謝

本實驗獲得行政院國科會贊助（計畫編號：NSC 88-2213-E-031-003），特此致謝。

參考文獻

1. Apte, C., F. Damerau and S. M. Weiss, "Automated Learning of Decision Rules for Text Categorization," ACM Transactions on Information Systems, 12 (3), 1994, pp. 233-251.
2. Blosseville, M., G. Hebrail, M. Monteil, and N. Penot, "Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together," In Proceedings of the 15th Annual International ACM

SIGIR Conference on Research and Development in Information Retrieval (SIGIR-92), 1992, pp. 51-58.

3. D'Alessio, S., K. Murray, R. Schiaffino, I. College and A. Kershenbaum, "The Effect of Topological Structure on Hierarchical Text Categorization," In Proceedings of the Sixth Workshop on Very Large Corpora, 1998, pp. 66-75.
4. Frakes, W. B., and R. Baeza-Yates, Information Retrieval Data Structures & Algorithms, Edited by Frakes, W. B., and R. Baeza-Yates, Prentice Hall, New Jersey, 1992.
5. Lewis, D. D., "Feature Selection and Feature Extraction for Text Categorization," In Proceedings of Speech and Natural Language, 1992, pp. 212-217.
6. Lewis, D. D., "Challenges in Machine Learning for Text Classification," In Proceedings of the Ninth Annual Conference on Computational Learning Theory, 1996, pp. 1.
7. Liddy, E. D., W. Paik, and E. S. Yu, "Text Categorization for Multiple Users Based on Semantic Features from a Machine-Readable Dictionary," ACM Transactions on Information Systems, 12 (3), 1994, pp. 278-295.
8. Liddy, E. D., W. Paik, E. S. Yu and K. A. McVearry, "An Overview of DR-LINK and its Approach to Document Filtering," In Proceedings of the Human Language Technology Workshop, Princeton, N.J., 1993.
9. Lin Chin-Yew, Robust Automated Topic Identification, PhD Dissertation, University of Southern California, 1997.
10. May, A. D., "Automatic Classification of E-mail Message by Message Type, Journal of the American Society for Information Science," 48 (1), 1997, pp. 32-39.
11. Ng, H. T., W. B. Goh, K. L. Low, "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization," In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-97), 1997, pp. 67-73.
12. Oard, D. W., "Adaptive Filtering of Multilingual Document Streams," In Fifth RIAO Conference on Computer Assisted Information Searching on the Internet, 1997.
13. Salton, G. and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, NY, USA, 1983.
14. Schütze, H., D. A. Hull and J. O. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem," In Proceedings of the 18th

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-95), 1995, pp. 229-237.

15. Watanabe, Y., M. Murata, M. Takeuchi, and M. Nagao, "Document Classification Using Domain Specific Kanji Characters Extracted by Method," In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), 1996, pp. 794-799.
16. Witten, I. H., A. Moffat and T. C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, International Thomson Publishing Company Press, New York, 1994.
17. Yang, Y. and C. G. Chute, "An Example-Based Mapping Method for Text Categorization and Retrieval," ACM Transactions on Information Systems, 12 (3), 1994, pp. 252-277.
18. Yang, Y., "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval," In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-94), 1994, pp. 13-22.
19. Yang, Y., "An Evaluation of Statistical Approaches to MEDLINE Indexing," In Proceedings of the AMIA, 1996, pp. 358-362.
20. 林頌堅，"自動化文件分類在資訊服務上的應用，21世紀資訊科學與技術的展望，1998, pp. 255-280.

附錄　財經記事資料介紹

財經記事為卓越出版社出版的財經新聞資料，每則共分為三部分：第一部份為分類碼、日期及報社等，第二部份為編者，通常是記者；第三部份為新聞標題。附表一為取自財經記事的部分範例。財經記事採階層式分類，共分為大類別、中類別、及小類別三層。其中，大類別共分為金融、產業等九類，附表二為大類別的類別種類及文件分佈情形。大類別之下又細分為 39 個中類別，其分類情形如附表三。由附表三，我們可以看出平均每則新聞標題約含 60 個位元組，即約 30 個中文字長。另外，由最長及最短兩個欄位，我們可以發現文件的長度變化很大，長度短於 5 個中文字的標題通常是不完整的文件。原本財經記事的分類，在中類別下以序號細分為小類別，但本研究僅以大、中兩個類型進行實驗。附表四及附表五分別為『總體篇』下的『經濟』及『稅賦』兩個中類別往下細分的情形。

附表一 財經記事部分範例

[1]HX.7104;810106;經;16(M);1;	[2]任文姍,江東峰	[3]日本區域性股市與店頭市場抬頭
[1]HF.01;810106;經;26(M);1;	[2]陳嬿妮	[3]經部商業司及多處機構全力推動,81 年商店自動化帶來市場新契機
[1]HB.9005;810106;聯;09(M);1;	[2]編輯部	[3]積極尋求與獨立國協會員國發展關係,中共與烏克蘭及塔吉克建交
[1]HDl.20;810106;聯;06(M);1;	[2]吳媛華	[3]3 分之 1 女性專任太太,鼓勵已婚婦女回到就業市場,勞委會擬妥草案報院
[1]HGb.01;810112;中晚;04(M);1;	[2]初馨怡	[3]新銀第 1 年可吸收存款 1000 億,佔資金市場 35 分之 1 金融環境將質變,產生消費者的銀行
[1]HB.9010;810112;中晚;03(S); 1;	[2]社評	[3]大陸企業來台投資
[1]J.60;810112;自早;02(M);1;	[2]社論	[3]妥善因應對韓關係的可能變動
[1]J.2007;810112;自晚;03(S); 1;	[2]彭琳淞	[3]兩岸條例延而難立
[1]J.01;810112;自晚;03(M);1;	[2]黃煌雄	[3]給黨團成員的一封公開信
[1]HJd.11;810112;工;01(M);1;	[2]陳志峻	[3]第 2 波擴大適用加值稅小店舖選定,鐘錶眼鏡等 9 行業須開立發票,決自 4 月 1 日起,展開強力輔導
[1]HX.12;810117;貿;05(L);1;	[2]編輯部	[3]我輸加鞋類退居第 3,業者應多警惕(加拿大)(附表 1:1990 年加拿大鞋類進口主要供源)
[1]J.10;810118;聯晚;02(M);1;	[2]社論	[3]民主化的嚴峻考驗:評立法院正副院長選舉
[1]HN.20;810118;聯晚;15(M);1;	[2]方紫苑	[3]我思・我捐系列(5):慈善有夢夢當圓,聯合勸募,加油!

附表二 財經記事大分類表

類別名稱	文件篇數	類別名稱	文件篇數	類別名稱	文件篇數
公營事業篇	2017	金融篇	12404	貿易篇	3308
其他	38432	國際篇	15742	農業篇	1777
服務業篇	12669	產業篇	16960	總體篇	29293

附表三 財經記事中分類表

大分類	中分類					
類別名稱	代碼	類別名稱	文件數	文件長度（位元組）		
				平均	最長	最短
公營事業篇	HDn	公營事業	2017	59	250	6
其他	HBp	人口、移民	223	55	162	8
	HDm	公共建設	3154	65	212	8
	HEs	郵政,電信	465	58	164	8
	HEm	大眾傳播	79	46	236	10
	HJ	財政	908	60	218	10
	HN	社會	2280	54	240	8
	J	政府,政治	15086	57	248	0
	L	教育	2368	59	251	10
	RA	醫療衛生	2383	61	201	8
	T	科技	758	61	157	6
	TD	環境	2927	60	229	6
	HFm	企業管理	3280	41	246	8
	W	公司檔案	956	51	215	7
	Hp	人物檔案	2564	54	234	8
	HPt	人事動態	770	51	218	8
	WG	集團企業	231	56	158	8
服務業篇	HF	服務業	8165	54	247	6
	HE	交通運輸業	2766	63	237	8
	HEt	觀光旅遊	1738	53	232	8
金融篇	HG	金融	3166	59	235	4
	HGb	銀行	2874	60	214	10
	HGe	外匯	564	56	157	4
	HGs	股票	4758	62	251	6
	HGt	租賃	1042	57	186	8
國際篇	HX	國際政經	15742	50	250	4
產業篇	HDo	各項產業	16960	57	240	4
貿易篇	HFt	貿易	3308	64	248	4
農業篇	S	農業	1250	60	206	8
	SD	林,牧,漁,礦	527	58	169	10
總體篇	HB	經濟	17404	60	251	4
	HBc	消費	1296	47	182	4
	HD	土地	1536	64	240	8
	HDi	工業	1966	62	238	8
	HDl	勞工	2858	59	201	10
	HJd	稅賦	2949	58	241	7
	HJt	關稅	384	62	176	14
	HFc	商標,智慧財產	900	57	181	8

附表四　總體篇經濟（HB）之細層分類

分類碼	類別名稱	分類碼	類別名稱
HB.01	綜合動態(利益輸送)	HB.20	經濟研究機構
HB.02	管理法令政策	HB.30	經濟辭彙
HB.03	統計數據,指標	HB.40	經濟建設計劃(六年國建)
HB.0301	經濟成長	HB.50	技術合作
HB.0302	國民生產毛額	HB.60	整廠輸出
HB.0303	國民所得	HB.70	自由化國際化
HB.0304	國民儲蓄	HB.80	經濟犯罪
HB.0305	物價	HB.85	地下經濟,走私,洗錢
HB.04	景氣循環	HB.8501	地下工廠(攤販)
HB.05	經濟史	HB.90	大陸政經
HB.06	生產力(經濟生產變化)	HB.9001	綜合動態
HB.10	投資,海外合作發展基金	HB.9002	貿易
HB.1001	對外投資	HB.9003	產業(含企業集團)
HB.1002	僑外投資(華僑)	HB.9004	商業(含金融,服務業)
HB.1003	投資環境介紹	HB.9005	政治
HB.1005	創業頭資(V.C)	HB.9010	兩岸經濟交流(貿易)
HB.1007	企業購併(事件性)	HB.9020	其他交流
HB.15	中小企業(自創品牌)	HB.9030	民運

附表五　總體篇稅賦（HJd）之細層分類

分類碼	類別名稱	分類碼	類別名稱
HJd.01	綜合動態	HJd.30	地方稅
HJd.02	管理法令政策	HJd.3001	土地稅
HJd.05	賦稅改革委員會	HJd.3003	房屋稅
HJd.10	稅捐稽徵(含機關)	HJd.3005	契稅
HJd.11	統一發票	HJd.3007	加值型營業稅
HJd.13	逃漏稅	HJd.3009	牌照稅
HJd.20	國稅	HJd.3011	印花稅
HJd.2001	個人所得稅(綜合所得稅)	HJd.40	規費(行政費用)
HJd.2003	營利事業所得稅	HJd.4001	商港建設費
HJd.2005	遺產,贈與稅	HJd.4003	工程受益費
HJd.2007	貨物稅	HJd.4005	都市建設捐
HJd.2009	證券交易稅,證券交易所得		