應用本體論於企業工作日誌自動化分類

陳榮靜 博士
朝陽科技大學 資訊管理系
e-mail：crching@cyut.edu.tw

徐幸謙 研究生
朝陽科技大學 資訊管理系
e-mail：s9654608@cyut.edu.tw

摘要

隨著資訊技術的普及，企業與機構過多的資訊導致數位文件不斷迅速累積，自動化文件分類已受到普遍重視。因此出現專門文件分類機制，主要有 4 個步驟：(1)資料選取與文字處理，(2)關鍵詞擷取，(3)特徵選擇，(4)分類器選擇。因此我們將以此機制為基礎，並利用 TFIDF 資訊檢索方法，加上本研究所提出的利用本體論的結合，以進行相關詞彙深度之量測，最後援用監督式的類神經網路(Supervised Learning Network)作為學習訓練的過程，即可分別得到資料最後的分類結果，之後使用準確度、召回率跟 F1 分別計算評估分類的成效。

關鍵詞：本體論、TF-IDF、文件分類法、類神經網路

Abstract

With the popularization of information technology, business and institutions have too much information leading to the rapidly accumulated digital documents. Automated text categorization has attracted universal attention. Thus the emergence of specialized file classification system, there are four main steps: (1) data selection and word processing, (2) Key words capture, (3) feature selection, (4) classifiers. Therefore, we will use this mechanism as the basis for information retrieval using TFIDF, multiplied by the depth of this research. The vocabulary carries out the correlation of the measure. Finally, Supervised Learning Network is used to train the processing documents. We can get classification results, after use the classification accuracy assessment of the effectiveness calculation.

Keywords: Ontology, TFIDF, Automatic Classification, Neural Networks

1. 前言

近幾年來，隨著網路技術的普及，企業與機構的資訊也相對地大量成長中，工作日誌是每個公司行號都會使用到的知識管理系統。如果只單純的用人工分類日誌的歸類，就會變成沒有效率與浪費時間，也有可能將日誌歸類錯誤，為了幫助使用者自資料庫中精確存取文件，所以提供一個簡單、易用而且有效的方式給使用者，是本研究致力達成的目標，因此分類對知識的管理與儲存就顯得非常重要。

文件分類的功能通常是讓使用者容易快速瀏覽尋找需求的文件，透過文件分類的功能可以明顯的知道每一篇文件的所在類別位置，因此使用者依據自己的需求就可以快速的找尋到各類別的文件，或者使用者輸入欲查詢的關鍵字，並且限制欲查詢的類別範圍，而得到查詢相關類別的文件。文件分類對使用者是相當方便的功能，因此透過文件分類可以方便的管理公司文件，讓員工方便的快速找詢文件增快工作效率。

本研究之目的，主要針對自動化資訊分類運作作為研究主題，並實作出系統，資料探勘中決策樹結合類神經網路的分類方式，建構最佳的日誌分類模型。在此類神經下可以針對不同階層內容的文件進行準確也分類處理，並進而讓使用者能快速且準確地歸類到真正的資訊位置。

在第二節探討說明關於本論文所使用到的技術文獻；第三節介紹本研究架構設計；第四節說明本研究的實驗結果；最後，提出了本研究的結論。

2. 文獻探討

目前為止已經有多種技術被拿來應用於文件分類上，如 SVM、決策樹、貝氏分類法以及類神經網路都有其各自相關文章分類系統已被發表，每項技術都有其優缺點，下列將分別介紹目前較有名與常運用的分類方法。

2.1 分類方法

2.1.1 支援向量機

支持向量機(Support vector Machine, SVM)[10][11][12]指的是用在機器學習的演算法，SVM 採用的方式可以有效用來表示文件與關鍵詞彙之間的關聯性，也可以代表文件之間的相似度，在早期的文件分類方法中都有著顯著的貢獻存在。

支援向量機(SVM)基本原理透過線性分割超平面(The linear separating hyper-plane)處理非線性的分類範圍，對於一群資料而言，我們可以透過 SVM 分類器將資料分成兩個類別。在典型的分類問題中，我們通常會定義以下基本的表示方法：x_i：是一個向量，手來描述某筆 N 維資料的樣式(Pattern)或是屬性(Attribute)，

$$x_i \in R^N, i=1,2,3,\ldots m \quad (1)$$

y_i：稱為標註(Label)或是目標(Target)，通常用｛±1｝表示（假設我們的分類目標為兩類），+1 和-1 皆表示兩不同的類別(Class)。

$$y_i \in \{\pm 1\}, i=1,2,3,\ldots m \quad (2)$$

而就資料分群而言，我們已知有相關系統已有不錯的效果。而如果在正確的使用前提之下，其方式幾乎都有達到很好的分類效果，然而，SVM 的優勢在於應用上較為容易。但也由於 Training instance 的數量呈線性成長，這樣的方式的缺點就是相當花費時間。

2.1.2 決策樹

決策樹(Decision Tree)[13][14]是一個樹狀結構(tree structure)的流程圖，其吸引人之處在於決策樹具有規則，決策樹的節點與節點之間的分支(branch)為經過條件測試後之結果，選擇一開始從根部到每一個葉部都有一套獨特的路徑，而這些路徑就是用來表達資料分類的規則。

決策樹常用於預測模型的演算法，他代表的是對象屬性與對值之間的一種映射關係。透過行為或資料，幫助我們把複雜的數據表示轉換成相對簡單的值觀結構。相對於其他數據挖掘算法，決策樹易於理解和實現，人們在通過解釋後都有能力去理解決策樹所表達的意義。

相對於其他數據挖掘算法，決策樹在以下幾個方面擁有優勢：

- 決策樹易於理解和實現。
- 對於決策樹，數據的準備往往是簡單或者是不必要的。
- 能夠同時處理數據型和常規型屬性。
- 是一個白盒模型。
- 易於通過靜態測試來對模型進行評測。
- 在相對短的時間內能夠對大型數據源做出可行且效果良好的結果。

但決策樹仍然有許多缺點現，在於無法考量所有相關的屬性，當分類種類過多時，會造成推論過程受資料缺失的影響。

2.1.3 自然貝氏分類法

自然貝氏分類法是以貝氏定理(bayes' theorem)為基礎[15]，以機率、統計學為原理概念發展出來，其主要的觀念是決定未分類別的資料應將分到哪一個類別中的分法。

自然貝氏法(Naïve Bayes)分類器做了一個簡化的假設，對於結果來說每個參數之間出現的機率是互相獨立的。Naive Bayes 分別計算在給予的文件下，文件歸屬於每一類別的機率，並將文件的類別指定給機率最高的類別[8]如下公式：

$$P(d'|c_k) = \prod_{j=1}^{m} P(d_j|c_k) \quad (3)$$

$P(d'|c_k)$表示在特定類別c_k裡找到文件 d'的可能，d 表示文件向量，$d=(d_1,\ldots,d_m)$，m 為文件向量的維度，N 為類別數目，c_k表示第 k 個類別。

不過大部分情況下參數出現的機率並不是互相獨立的，對預測的結果來說，此一簡化的假設也耗損了一些精確度。

2.1.4 TFIDF

文章的基本組成單位是字詞，而行自動索引的實驗中，為統計詞彙的出現頻率，發現除卻高頻與低頻者，所留下的中頻（middle-frequency）字詞，多半是比較有意義的，因而提出「關鍵字詞適度詞頻論」(resolving power of significant words)。 (Luhn, 1958)。而引發日後諸多學者如：Sparck Jones (1972), Salton & McGill (1983)等人投入自動文件處理的興趣。

文件分類方法中最較為常用的方法就是 TF-IDF 方式[9]，此方式是計算出現在個別文件中字彙的頻率，以及該字彙在所有訓練文件中所出現的頻率，前者數值越高代表該字彙越

能代該份文件，而後者數值越低則代表該字彙在整個分類過程中越有代表性。將兩者相乘之後，則可以得到該字彙的重要性排序，再透過一個門檻值的設定，便產生出足以代表該份文件的字彙集合。

2.1.5 類神經網路

為了在語音及影像辨認獲至與人腦相似的功能，自 1940 年起，科學家即著手從事此方面的研究，仿造最簡單的神經元模式，開始建立最原始的類神經網路(Artificial Neural Network ANN)，歷經 40 年的發展，類神經的研究工作雖曾一度陷入低潮，近幾年又再度復甦，並且結合了生理，心理，電腦等科技而成為新的研究領域。

類神經網路已被研究多年，以人類頭腦的思考邏輯方的原理而發展出來的稱為類神經網路[16][17]。Freeman 及 Skapura 所發表的類神經定義，類神經網路(Neural Networks)是指模仿生物神經網路的資訊處理系統，利用大量相連的人工神經元(Neuron)來模擬生物神經網路的能力。隨著類神經的研究一直發展，Boger 等人[2]採用類神經網路的理論發展一套用資訊過濾與用關鍵詞彙選擇的應用系系。此研究是利用類神經網路將使用者的設定資料與電子郵件檔給予向量化，並在兩者間建立關係，經由類神經網路分析兩向量，進一步探索電子郵件中重要的詞彙，進而應用於郵件間的相關性計算。

在應用領域上，類神經網路結構是一個層狀的前饋結構，分為輸入層(input layer)，隱藏層(hidden layer)，與輸出層(output layer)，每一層都包含許多處理單元。簡單來說，類神經網路是一個必須教他正確的輸入輸出範例，然後它會依照曾經學習過的範例提出新範例的評估值，理論上範例越多就越精確。

在類神經網路的基本原理中，最有代表性的就是倒傳遞類神經網路[18]，倒傳遞網路乃採用最陡降法(Gradient Steepest Descent Method)的原理，將輸出與預期學習值相當才算收斂而停止，倒傳遞類神經網路屬於監督式學習網路，因而適合診斷、預測等應用。且學習精度高，本研究將以類神經網路分類法中倒傳遞網路學習模式來預測。

表 1 文件自動方法比較之優缺點[7]

文件分類方法	優點	缺點
Support vector Machine	計算容易、快速分類	分類準確度底
Decision Tree	分類規則容易解釋	類別過多時分類準確度下降
Naïve Bayes	計算容易、快速分類	簡化假設使準確度不足
TF-IDF	實現方便	缺乏對語義的考慮
Neural Network	經由學習訓練可接近人類專家分類結果	需耗時進行訓練，需要一定數量之訓練樣本

2.2 類神經網路運用在文件分類上相關研究介紹

(1) 以類神組網路中的倒傳遞網路做為分類器，經由這樣學習的方式，可以使文件分類系統更容易地應用到其他的領域。該研究採用階層混合式的專家模組，在一個事先定義好的階層架構下定義較小的分類問題，然後藉由已訓練好的分類器對測試樣本中的新文件做自動化分類的動作。[3]

(2) 利用類神經網路技術建立新的文件分類方法 GCS(Growing Cell Structures)，並使用語義學的微特徵來作為分類規則，進行新聞文件分類的實驗。[19]

(3) 透過倒傳遞類神經網路建構一個考慮個人偏好因素之多重文件分類模式，實驗結果證實文件分類模式可依據不同分類者之個人偏好，給予同一份文件不同的個人化分類預測結果，且對每位分類者來說，分類效用之表現皆是可接受的。[4]

(4) 以並聯式的倒傳遞網路來進行文件的分類學習，實驗結果證實他們提出的分類模組比起單一式倒傳遞網路分類模組有較高的精確度[5]

(5) 應用倒傳遞類神網路分類技術為基礎，結合 sunger 資料庫中的 Cancer Gene Census 資訊，針對 PubMed 資料庫類的癌症基因醫學文獻，發掘其癌症疾病類別中子類別與基因之關聯性。[6]

2.3 CKIP

詞是最小有意義且可以自由使用的語言

單位。任何語言處理的系統都必須先能分辨文本中的詞才能進行進一步的處理，例如機器翻譯、語言分析、語言了解、資訊抽取。因此中文自動分詞的工作成了語言處理不可或缺的技術。基本上自動分詞多利用詞典中收錄的詞和文本做比對，找出可能包含的詞，由於存在歧義的切分結果，因此多數的中文分詞程式多討論如何解決分詞歧義的問題，而較少討論如何處理詞典中未收錄的詞出現的問題（新詞如何辨認）。

由於網際網路產生大量資訊但缺乏有效的自動化分析方法及技術足以快速處理。為了達到智慧型的資訊處理、知為本的訊息處理，中研院資訊所和語言所成立一個跨所合作的中文詞知識庫小組(Chinese Knowledge Information Processing group, CKIP)[1]，共同合作建構中文自然語言處理的資源與研究環境，其進行研究主要有三個方向：智識擷取、知識表達及知識應用。其中詞知識庫提供語料庫標記訊息(Tagging information)。句子的結構是語義析的必要訊息，利用自動剖析系統將語法規律和斷詞後的文本做比對，找出可能的簡短語句結構，使用機率式無語境規律的模型(Probabilistic Context-free Grammar)並加入結構中詞彙搭配關係機率解決結構歧義。CKIP 斷詞系統還包括自動抽取新詞建立領域用詞，具有新詞辨識力並附加詞類標記的功能。此系統含有約十萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據此一詞彙庫及定量詞、重疊詞等構詞規律，來解決分詞歧義問題。

3. 架構設計

本研究架構，如圖 1，第一步收集工作日誌資料，第二步建立本體論，第三步結合本體論資料分析、第四步建立類神經網路模組，第五步分類結果。

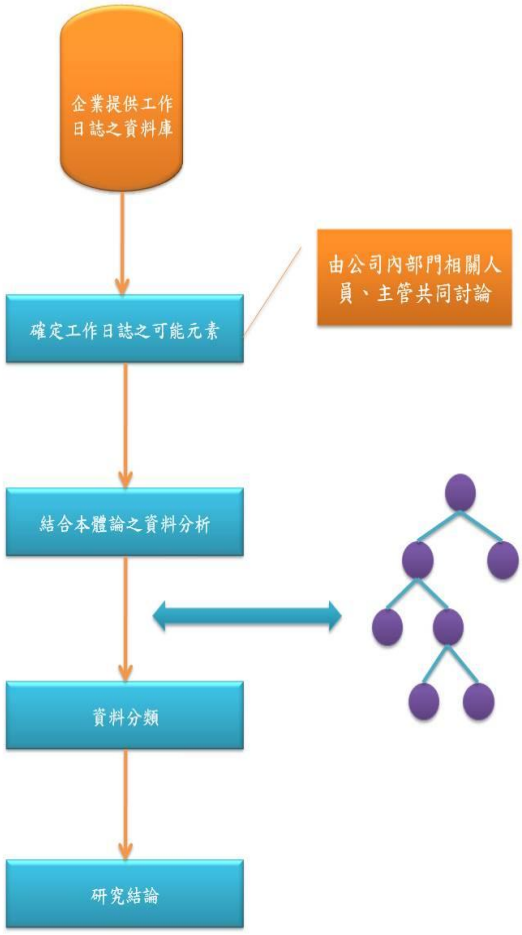

圖 1 文件分類自動流程

3.1 日誌文件前處理

在處理企業日誌資料分析前，必須將資料文字表示成適用在結構化資料處理的格式，而轉換成此表示方法需進行下列步驟：

(1) 移除 HTML 標籤：

由於有些標籤是與資料分析無關的內容，所以事先將 HTML 的標籤移除。

(2) 中文斷詞：

中文語系不如印歐語系，字詞之間均有空白作為區隔。因此，必須先將中文句子斷成一個個獨立的詞目，並且足以代表該文件上的特徵意義，顯得有些困難。但目前中研院資訊所和語言所成立一個跨所合作的中文詞知識庫小組，所研發的斷詞程式有不錯的成效，因此，結合 CKIP 斷詞系統將日誌文件加以斷詞。

(3) 詞性選取：

利用 CKIP 之詞性標記(part-of-speech tagging, POS)並依據杜海偷(1999)的研究結

果，刪除一般性詞目外，保留名詞與動詞，本研究再將研究結果為單字詞的部份，刪除單一字元的詞目，一般而言，單一字元的詞目大多是無法表達一個完整的概念，並不具備有關鍵字詞的特質，所以先將單一詞目的字詞刪除，此法應可擷取斷詞不當提升關鍵字詞擷取的品質。

3.2 統計式詞彙特徵選取法

日誌文件在經過資料淨化(data cleansing)的過程後，為選取各類日誌之關鍵字，在關鍵字的選取方法上，有相當多的論文文獻發表，如詞庫比對、斷詞、剖析、雙連字串、甚至是機率統計透過對文件內容之分析，累積充分之統計參數後，再擷取統計參數符合指定條件之片語作為關鍵字。

但是因為所處理的文件為企業內部的專業術語，所以需要具有領域知識作為輔助，所以在前置作業中使用的詞庫需有專家詞庫來彌補自然語言處理系統自動斷詞，不如人工建置詞典來得精準。根據楊允言有關中文文件自動分類之研究中內容所提，經由人作業進行關鍵字的選取是較具意義的，因此本研究則由負責各部門工作日誌分類之主管同仁進行關鍵字之挑選，同時對事件類別進行判定。

另外，在分割研究期間上，將文件一分為二的方法，將各部門劃分為 4 個分類項目，每個分類項目中各分一半，當作訓練跟測試實驗之用

3.3 建立本體論

為選取各類日誌之關鍵字，在關鍵字的選取方法上，有相當多的論文文獻發表，如詞庫比對、斷詞、剖析、雙連字串、甚至是機率統計透過對文件內容之分析，累積充分之統計參數後，再擷取統計參數符合指定條件之片語作為關鍵字。

但是因為所處理的文件為企業內部的專業術語，所以需要具有領域知識作為輔助，所以在前置作業中使用的詞庫需有專家詞庫來彌補自然語言處理系統自動斷詞，不如人工建置詞典來得精準。根據楊允言有關中文文件自動分類之研究中內容所提，經由人作業進行關鍵字的選取是較具意義的，因此本研究則由負責各部門工作日誌分類之主管同仁進行關鍵字之挑選，同時對事件類別進行判定。

3.4 結合本體論資料分析

建立好本體論之後，首先先利用傳統的字詞頻率加權法，來找出每個定義好的概念字詞在工作日誌中出現的重要性，以 P_i 表示為工作日誌的第 i 個，W_{ij} 表示為將概念字詞在工作日誌中的權重，其中 i 為第 i 個工作日誌，j 為第 j 個概念字詞，因此一個工作日誌可以表示如公式(4)：

$$P_i=\{W_{i1}, W_{i2}, \ldots, W_{in}\} \quad (4)$$

接著，我們使用了字詞頻率的加權法，首先先計算概念字詞在工作日誌中出現的頻率如公式(5)。

$$tf(i,j)=\frac{w_{i,j}}{\sum_{k=1}^{t} w_{i,k}} \quad (5)$$

接著計算逆向文件頻率，其中 N 為文件總筆數，n_j 則包含徵 j 的件筆數，如公式(6)

$$idf(j)=\log\frac{N}{n_j} \quad (6)$$

再計算每個概念字詞的 TFIDF 字詞頻率權重，如公式(7)，得到每個概念字詞在工作日誌的權重。

$$W_{ij}=TF_{ij}\times IDF \quad (7)$$

然而，這次傳統的 TFIDF 的方法，本研究加入了本體論的概念，不只利用字詞概念出現在工作日誌出現的頻率權重，且考慮字詞概念在本體論的中深度所代表不同的含義，因此，我們依概念字詞出現在本體論的深度給予不同的權重，以表示概念字詞在不同深度所代表不同的含義，如公式(8)。其中 DF_j 表示概念字詞 j 在本體論深度的權重。

$$W_{ij}=TF_{ij}\times IDF_j\times df_j \quad (8)$$

透過以上的計算後，我們得到了每一個工作日誌中概念字詞的權重，這個權重表示了字詞出現頻率及本體論中深度的含義，接著將資料輸入至類神經網路，以進行分類。

3.5 建立類神經網路模組

本研究是以監督式的類神經網路作為學習訓練的過程，因此，採用監督式的類神經網路中的倒傳遞網來做資料分類。倒傳遞網路(BPN)是目前最具代表性與應用最普遍的監督式學習類神經網路，BPN 可將訓練的範例一直反覆學習，不斷調整各層的連接權重，其學習過程以一次一個訓練範例進行，經過許多學習週期後，直到輸出與預期學習值相當才算收斂而停止。

圖 2 是類神經網路架構，包含輸入層、隱藏層、輸出層，各層說明如下：

- 輸入層：

用以表現網路輸入變數，其處理單元數目依問題而定，且一般輸入層的轉換函數採用線性函數。此階層將 TFIDF 運算後的結果再計算深度所得到的結果，將資料匯入類神經網路輸入層中。

- 隱藏層：

負責處理輸入向量與輸出向量間之交互影響，且應用上常採用一層到二層的隱藏層，單元數目無標準方法決定。

- 輸出層：

為輸出向量輸出之神經元，用以表現網路之輸出變數，其處理單元數目依問題而定。本階段在隱藏層中向量與輸出向量間的交互影響下，將輸出得到分類的結果。

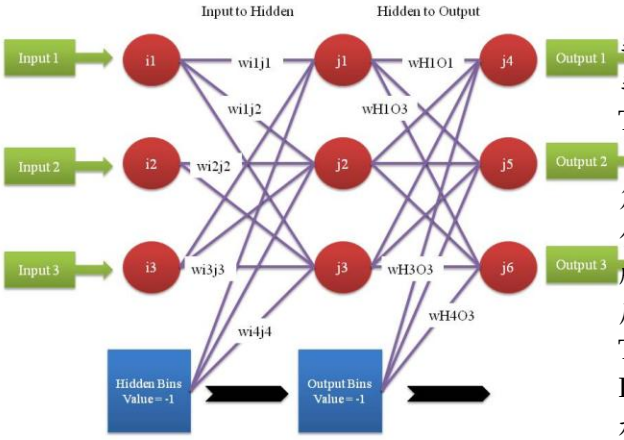

圖 2 類神經基本架構

4. 實驗結果分析

本小節將介紹 TF-IDF 與字詞對於一個文件集的其中一份文件的重要程度，TF 是使用[9]提出來的概念，其中 TF 指的是某一個給定的詞語在該文件中出現的次數，以出現的頻率代表該字詞的重要性，所以詞頻越高，它的重要性就越高，其 Term Frequency(TF)定義公式如下所示：

$$tf(i,j) = \frac{w_{i,j}}{\sum_{k=1}^{t} w_{i,k}} \quad (9)$$

以上式子中 W_{ij} 表示文件 i 對特徵 j 的 TF 值，t 表示文件 i 的特徵總數。tf(i,j)越大表示特徵 j 對文件 i 越重要。

對於 IDF 定義為文件出現得頻率之倒數，是一個詞語普遍重要性的度量。簡單而言，就是判斷字詞在此文章中的重要程度，可以由總文件數目除以包含該詞語之文件的數目，再將得到的商取對數，其詳細 Inverse Document Frequency(IDF)定義的公式如下所示：

$$idf(j) = \log \frac{N}{n_j} \quad (10)$$

其中，N 指全部文章的數量，而分母指有出現關鍵字的文章數量。

然後，計算每個概念字詞的 TF 與 IDF 相乘之後，所得的字詞頻率權重，定義的公式所示如下：

$$W_{ij}=TF_{ij}\times IDF_J \quad (11)$$

換句話說，某一特定文件內的高詞語頻率，以及該詞語在整個文件集合中的低文件頻率，可以產生出高權重的 TF-IDF，所以使用 TF-IDF 當作衡量標準。

透過下列的實際範例中詳細說明，假如一篇文件以調整為關鍵字，計算 TF-IDF 的權重值，用下表整理說明各函數所得之值如下表 1 所示。其中，關鍵字「server」出現了 2 次，那麼「server」一詞在該文件中的出現次數即為 TF，而「調整」出現的文章佔全部的比重則為 IDF，將 TF 跟 IDF 相乘後的結果，即為調整對於這篇文章的重要程度 TF-IDF 值。

表 2 計算調整的 TF-IDF 實際範例

關鍵字	調整
文章內容	盡快完成新 APS Server 上線. HPLP1000R 建立 APS 測試環境. 1. 欲建置的 Server 因 Exchange 測環境需求,轉而建置 Exchange Test Server.
TF	0.0689655172413793103

IDF	1.1405361839811078967
TF-IDF	0.0786576678607660617

4.1 類神經之資料訓練

在文件分析階段，首先計算文件中本體論概念出現的頻率，並且依據不同的詞彙類別給予不同的權重，每個概念可能出現於不同的類別，其權重的分配例如：頻寬這個概念在客服中心以及網路維運部門都有相關聯，但頻寬這個概念對於網路維運比較具有較深的關聯特徵，所以在網路維運中此一關鍵字的權重為 0.7，而客服中心的權重值為 0.5

類神經訓練方式主要是以 MATLAB 工具作為訓練的分類結果，測試分類的集合主要是從 1852 筆的文件中分成一半，以進行分類測試，藉此得知本研究方法的可行性與實用性，同時會將本方法與傳統文件分類方法來做比較，以驗證本研究方法，在針對訓練文件分類的實驗下，是否具有一定程度的功效與分類正確性。

4.2 實驗評估標準

在文件分類完成後，我們需要對利用類神網路進行分類後的結果進行評估，來驗證本研究方法是否有效，本研究採用兩種評估方法，一種是資訊檢索評估測量作為分類結果之評估標準另一種為 f-measure。

回收率主要計算真正的正確分類數之中，資訊擷取之正確資料項；而精確度為系統判斷下，分類到相關文章的比例
回收率與精確率之公式如下：

表 3 回收率與精確度示意圖

	相關	**不相關**
分類到	A	B
未分類到	C	D

$$\text{回收率} = \frac{A}{(A+C)} \tag{12}$$

$$\text{精確度} = \frac{A}{(A+B)} \tag{13}$$

從一個資料庫的文件集合中分類文件時，可以把文件分成四種組合，如表 2：
A→系統分類到的相關文件。
B→系統分類到不相關文件。
C→相關但是系統沒有分類到的文件。
D→不相關而且也未被系統分類到的文件。

而 F-measure 之公式如下：

$$F = \frac{2PR}{P+R} \tag{14}$$

公式中 R 為回收率，P 為精確度。

4.3 測試結果

在本研究中，以一份文件只會歸屬到一個分類的情形，而不會產生一份文件對應多個類別情況，因此所有比較皆建立在此基礎上，以下將就本研究方法與
字頻率跟 TF-IDF 方法進行比較。

將文件總數取出 1052 份作為訓練之用，而各個分類項目是以企業內部的部門別作為主分類項目，共可分為 4 個類別，各分類詳細名稱如表 3 所列，表 3 同時為每一個分類項目編號，以作為分類過程中之識別用。

表 4 本實驗之各分類項目名稱及文件數目

分類編號	**分類項目名稱**	**訓練文件數目**	**測試文件數目**
1	網路維運	223	150
2	客服中心	326	250
3	資訊技術	276	250
4	行政管理	227	150
		1052	800

在所有收集到的 1852 份文件中，本實驗使用了其中的 1052 份文件作為訓練之用，其餘的則作為測試之用，訓練方式則根據本體論建構、類神經網路訓練來進行，直到為每一個文件各自歸類之後，便可進行實驗的評估。而 TF-IDF 分類方法則如上一節所示，先計算單字在某文件的出現頻頻率，再計算逆向文件頻率，並使用此方法來進行測試的動作。

以同樣的測試資料(共 800 份文件)針對這三個方法進行比較，便可得到表 4、表 5 和表 6 的結果：

表 5 分類準確度列表

分類項目	**字詞頻率**	**TF-IDF**	**本研究分類方法**
1	68%	69.33%	71.33%
2	99.2%	98.8%	97.6%
3	70%	72%	72%
4	77.33%	81.33%	83.33%

表 6 分類召回率列表

分類項目	字詞頻率	TF-IDF	本研究分類方法
1	68%	69.33%	71.33%
2	59.60%	59.60%	59.60%
3	40%	43%	42%
4	77.33%	81.33%	83.33%

表 7 F-measure 列表

分類項目	字詞頻率	TF-IDF	本研究分類方法
1	68.00%	69.33%	71.33%
2	74.46%	74.35%	74.01%
3	50.58%	54.00%	52.73%
4	77.33%	81.33%	83.33%

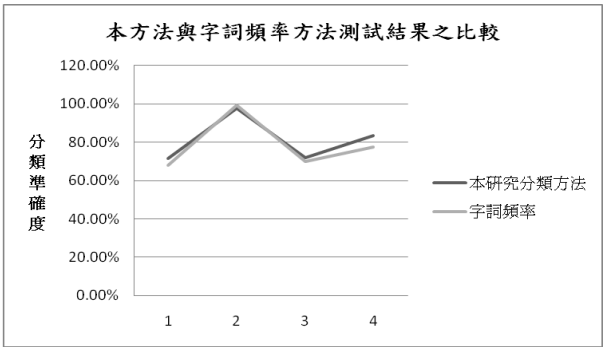

圖 3 本方法與字詞頻率方法測試結果之比較

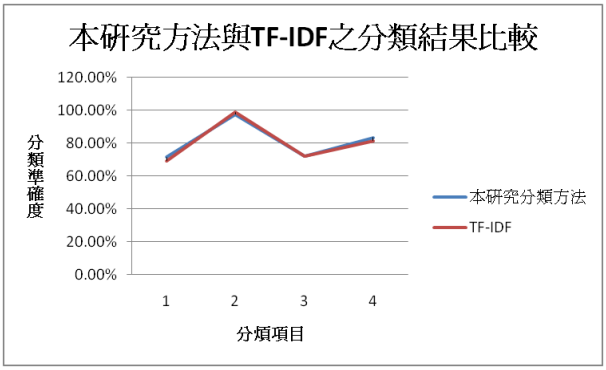

圖 4 本研究方法與 TF-IDF 之分類結果比較

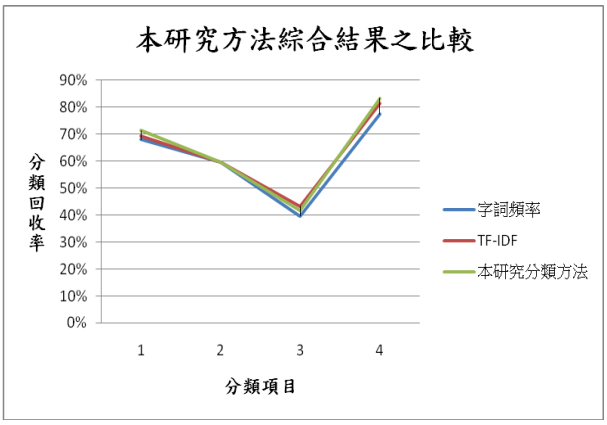

圖 5 本研究方法之召回率綜合比較結果

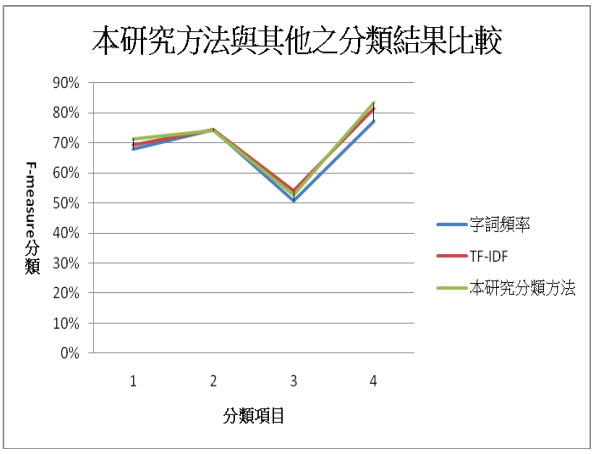

圖 6 本研究方法以 F-measure 評估結果

由上面圖表所示，可以知道，本方法比傳統字詞頻率分類方法，分類的準確度隨之提高，在針對第 1、3、4 文件的分類上，有著很明顯的改善效果。

在測試過程中，可以發現當日誌文件對工作內容描述越詳細，使用本研究方法後之分類以資訊檢索評估跟 F-measure 方法，如圖 4、5、6 所示，明顯的提昇。在實際分類過程中，

如果一份未知的文件可以比對到該分類的較深層的字彙，便讓其越偏向該分類，這樣的做法將有助於提升分類的正確性。在分類過程中，錯誤的資料在到了本體論結合資料分析後無法找到正確的詞彙，再經由類神經網路運算處理後所得出的結果，會產生出評估結果偏低的情況，因為這些評估方法是由歸類出來的文件中正確的比例來計算的，當無法找到正確關鍵字的情況下也就會讓評估結果也偏低。

5. 結論

在本篇論文中，我們提出了以深度關鍵字

詞來判斷文件內容是否自動歸類到正確的部門。在我們的實驗結果顯示以深度關鍵字詞歸類文件類別，確實能符合正確的歸類文件的類別。而以深度關鍵字詞所歸類的文件類別也確實比一般 TF-IDF 為佳且更符合公司所適用。

本研究資料來源是根據某公司提供的資料庫並經由撈取所蒐集 4 個部門文件，由於樣本類別 4 個部門略嫌少量，期望未來可以蒐集更多的樣本部門類別，以利於更進一步的探討與研究。

未來隨著日誌文件的增加，會產生越來越多無法符合關鍵字詞的文件，這時把這些無法符合關鍵字詞的文件交付給領域專家進行知識的驗證，確定關鍵的字詞後，則進行本體論的新增。新增本體論的詞彙包含其他部門的專業用語，提供使用者方便日誌快速的分類，免除手動歸類文件的困擾，增加工作效率。

隨著文件的新增與刪除，管理知識庫的工作也越來越煩雜。因此如何把文件日誌上的內容擷取相關的知識，建立知識彼此間的關聯，將所獲取的知識建構成知識樹。知識樹除了記錄知識實體外也記錄知識實體間的關係，這樣能提供使用者輕鬆分類文件。

參考文獻

[1] 中研院平衡語料庫詞類標記集 24，http://ckipsvr.iis.sinica.edu.tw/category_list.doc

[2] 許正欣，語意網上自動化建構本體論之研究，天主教輔仁大學資訊管理學系碩士論文，2004。

[3] 陳彥呈和蔣榮先，基於基層式類神網網路之自動新聞文件分類方法，第十四屆計算語言學研討會論文集，pp. 89-97，2001。

[4] 陳麒偉，考慮個人偏好因素之多重文件分類方法，國立成功大學資訊管理碩士論文，2008。

[5] 曾祥泰，以類神網路為基礎的中文文件分類研究，國立交通大學資訊工程碩士論文，1998。

[6] 李俊宏，類神經網路與癌症基因統計資訊應用於白血病醫學文件分類之研究，工程科技與教育學刊，2009，第六卷，第三期，pp. 239-256。

[7] 林家誼，應用類神經網路文件自動分類技術建構電子化知識文件管理系統，國立清華大學工業工程與工程管理碩士論文，2004。

[8] 吳文峰，中文郵件分類器之設計及實作，逢甲大學資訊工程學系碩士班碩士論文，2002。

[9] Salton, G., Wong, A., and Yang, C.S., “A Vector Space Model for Automatic Indexing”, Communication of the ACM, 18(11), 1975,pp.613-620.

[10] Jiu-Zhen liang, “SVM multi-classifier and Web document classification,” Proceeding of 2004 International Conference on Machine Learning and Cybernetics, vol. 3, pp. 1347-1351, 2004.

[11] Chin-Ming Chen, Hahn-Ming Lee, and Ming-Tyan Kao, “Multi-class SVM with negative data selection for Web page classification,” Proceeding of 2004 International Conference on Neural Networks, vol..3, pp. 2047-2052, 2004.

[12] Harris Drucker, Donghui Wu, Vladimir N. Vapnik, “Support Vector Machines for Spam Categorization,” IEEE Transactions on Neural Networks, Vol. 10, NO. 5, 1999.

[13] Yalin Wang, Haralick R., and Phillips I.T., “Xone content classification and its performance evaluation,” Proceeding of sixth International Conference on Document Analysis and Recognition, PP. 540-544, 2001.

[14] Jiawei Han, Micheline Kamber, “Data Mining : Conceptsand Techniques,” MORGAN KAUFMANN PUBLISHERS, 2001.

[15] Yong Wang, Hodes J., and Bo Tang, “Classification of Web documents using a naïve Bayes method,” Proceeding of 15th IEEEE International Conferenceon Tools with Artificial Intelligence, pp.560-564, 2003.

[16] Farkas J, “Towards classifying full-text using recurrent neural networks,” Conference on Electrical and Computer Engineering, vol. 1,pp. 511-514, 1995.

[17] Farkas J, “Neural networks and document classification,” Conference on Electrical and Computer Engineering, vol. 1, pp. 1-4, 1993.

[18] Jennifer Farkas, “Document classification and recurrent neural networks,” Proceedings, of the 1995 conference of the Centre for Advanced Studies on Collaborative research, page 21, 1995.

[19] W.G Baxt, “Use an Artificil Neural

Network for Data Analysis in Clinical Decision-Making: the Diagnosis for Acute Coronary Occlusion," Neural Computation, 1990, Vol. 2, No. 4, pp. 480-489.