# 以自組織映射圖進行計算語言學領域術語視覺化之研究
# Visualizing the Terms of Computational Linguistics with Self-Organizing Maps

林頌堅
Sung-Chien Lin

世新大學資訊傳播學系
Department of Information and Communications, Shih-Hsin University
scl@cc.shu.edu.tw

**摘要** 本論文的研究利用自組織映射圖(SOM)技術將計算語言學相關術語對應到二維圖形，使得術語之間的關係可以在映射圖中加以呈現，提供使用者做爲資訊檢索以及了解重要研究主題的輔助工具。在本論文中，我們所探討的問題有(1)發展SOM技術應用到術語資訊視覺化的方法，(2)評估SOM技術應用到術語資訊視覺化的成效，(3)利用研究結果分析計算語言學中重要的研究主題與主題之間的關係。在SOM技術的應用中，首先從論文資料中抽取重要的術語，接著以術語之間的共現關係做爲基礎，建立每一個術語的特徵向量。再以術語特徵向量做爲輸入資料，進行SOM訓練以及將術語映射到圖形上。對於這項技術在應用上的成效評估，由於映射節點的距離關係在視覺上要需要符合術語間的相關性。因此，我們建議以特徵向量的距離與節點位置的距離之間的相關係數做爲成效評估的標準。最後，對於計算語言學領域的術語所進行的實驗中可以發現大多數相關的術語都可以映射到相鄰近的節點上，而術語所映射節點的位置也可以大致表現主題之間的關係。這個結果表示SOM技術適合應用於術語資訊視覺化。

## 1 緒論

本論文是一個將計算語言學相關術語(terms)對應到二維圖形的研究，其目的是希望能夠蘊含在術語之間的資訊加以視覺化(visualization)。從論文所抽取出來的術語可以表示研究問題、方法、理論與技術等論文相關的主題，若是針對某一研究領域所發表的論文進行術語抽取並加以統計，所得到的高頻術語便是這個領域的重要主題[1]。因此，這些從論文抽取出來的術語將有助於了解這個領域所發展的研究課題或是進行資訊的檢索。爲了進一步幫助使用者從大量的文件資料庫中搜尋相關的資訊來解決所面對的研究問題以及提供他們對於這個領域研究所產生的知識結構(knowledge structure)有完整的認識，可以將這些術語整理成階層式(hierarchical)組織或網路式(network)組織，來闡明術語之間的關係。在資訊檢索的技術與應用上，索引典(thesaurus)便是將某一特定領域的相關術語與它們之間的關係整理成一個階層式與網路形式的組織[2]。在索引典的結構裡，將每一個術語作爲網路中的節點，而以相關術語之間的關係作爲相應節點之間的連結。近來，許多研究提出各種術語組織的自動化方法，這些方法多以統計的叢集(clustering)技術爲組織術語的方法，將關聯性較強的術語放到相同的集合中，並且利用術語在文句中的共現(co-occurrence)關係作爲術語之間的關聯[3, 4]。利用叢集所形成集合便可以了解術語之間的關聯性，並且在同一集合中的術語往往經常共同出現在主題相關的論文中，因此這些術語集合可以呈現這個研究領域的研究主題。然而，除了利用叢集技術所形成的集合來對於術語之間的關聯進行分析之外，若能夠將術語以及它們之間的關聯呈現在圖形中，提供瀏覽與深入探索，對於檢索相關資訊與分析領域的知識結構勢必更有幫助。

『資訊視覺化』(information visualization)是以二維或三維的圖形來表現一組資料之間的可能關係，目的是輔助人們認知原本的資料間不易察覺的關係，作爲決策判斷或探索新知的依據[5]。在過去，資訊視覺化常被應用於高維的數值資料，然而由於電子文件的數量大幅增加，對於組織大量文件以及方便而有效的全文檢索介面的需求越來越大，已經有許多學者著手進行文字資訊視覺化的探討。文字資訊視覺化的目標是將每一個文字資料對應到圖形上某一位置上的一點，使得文字資料之間的相關程度(relevance)可以用圖形上點與點之間的距離加以表示，兩點間的距離愈近便表示所代表的兩筆文字資料愈相關。使用者便可以直覺地將圖形上表示的距離作爲資料間的關聯，進而了解資料的整體分布情形。因此，在文

字資訊視覺化研究中常見的做法是首先設定文字資料的特徵向量(feature vectors)，再以特徵向量來估算資料兩兩間的相關程度，接著利用映射技術將文字資料對應到圖形上，盡量使圖形上點與點的距離之間的關係保持術語相關程度間的關係。常使用的映射技術有統計導向與類神經網路導向兩類[6]。在統計導向的方法中，將所有資料間的相關程度組合成一個矩陣，每一筆資料對所有資料的相關程度對應到矩陣中的一行與一列，換言之矩陣上的每一個元素便是兩筆資料間的相關程度。接著便利用統計技術，如SVD(singular value decomposition)[7]、PCA (principal component analysis)或是MDS (multidimensional scaling) [6, 8]等，找到一組轉換矩陣將原先的矩陣加以分解與轉換，使得重要的距離訊息得以保留在新產生的矩陣中。而以轉換矩陣作爲將資料映射到圖形的依據。

另一方面，自組織映射圖(self-organizing maps, SOM)則是在應用類神經網路導向的方法到文字資訊視覺化處理中常採用的技術[9]。顧名思義，SOM是一種以資料驅動(data-driven)的非監督式學習(unsupervised learning)方法，利用資料的特徵向量作爲訓練資料，訓練一組排列成方陣的節點，從反覆的訓練過程中讓產生的映射圖反應資料之間的關係[10]。在SOM技術中，每一節點都是一個向量，向量的維度與資料特徵向量的維度相同。在經過多次的訓練過程後，所有的資料都依照其特徵向量與節點的相似程度，映射到某一個節點上，而且節點間愈接近者相似程度愈高。因此，相關程度接近的資料會映射到同一節點或鄰近的節點上，而且所投射節點之間的相對距離可以表示資料的相關程度大小，距離愈大相關程度愈小。SOM的優點包括了可以將高維資料的距離關係，以自組織的型式保留在二維的映射圖中，並且MDS等統計導向方法大多需要極大量的運算資源，且在新增資料時，無法利用先前的計算結果，在實作方面，SOM技術較容易達成。因此，近年來有相當多文字資訊視覺化的研究採用SOM作爲映射技術。

在本論文的研究中，我們嘗試將計算語言學術語的關係視覺化，利用SOM將術語之間的相關程度映射到圖形上。因此，本論文的研究問題包括：(1) 發展SOM技術應用到術語資訊視覺化的方法，(2)評估SOM技術應用到術語資訊視覺化的成效， (3)利用研究結果分析計算語言學中重要研究主題之間的關係。

本論文其餘的章節組織如下，第2節中將簡介SOM技術，並回顧利用SOM技術處理文字資訊的研究；第3節說明本研究如何利用SOM技術，將計算語言學相關術語進行資訊視覺化處理的方法，並提出成效評估的方法；第4節則是對此一研究相關實驗的結果與說明；最後的第5節是本論文的結論與未來進一步研究的建議。

## 2 相關研究

SOM是一種非監督式的類神經網路[10]，在資料的叢集與視覺化上，應用十分廣泛。SOM的特色包括了它的類神經網路型態(topology)與訓練模式。在SOM中，由一組反映輸入資料的節點所構成，而這些節點排列成矩陣的型態，每一個節點與其他四個節點相連接，此一結構便是所謂的特徵映射圖(feature map)。事實上，每一個節點都代表一個特徵向量，向量的維度與資料項的特徵向量維度相同。在輸入資料之後，便重複訓練過程，調適節點的特徵向量，使得特徵映射圖可以反映輸入的資訊項之間的關係。SOM與一般『向量量化』(vector quantization)在訓練過程中最大的不同是，每次的訓練時，不僅只調適節點中與輸入資料最相接近的特徵向量，而且還同時調適了在特徵映射圖上鄰近範圍(neighborhood)內所有節點的特徵向量。因此，在經過多次的訓練之後，可以使特徵向量接近的資料映射到相同或是鄰近的節點上，使得圖形具有組織化的結構，而且將原本資料在高維特徵向量的距離或接近程度表示到SOM的節點的距離。通常用來衡量節點間距離的方式爲式(1a)中的歐幾里德距離(Euclidean distance)或式(1b)中的Manhattan距離等。

$$d(n_w, n_c) = ((x_w - x_c)^2 + (y_w - y_c)^2)^{1/2} \tag{1a}$$

$$d(n_w, n_c) = | x_w - x_c | + | y_W - y_c | \tag{1b}$$

式(1)中，$n_w$是特徵向量與某次輸入的特徵向量最接近的節點，在SOM的訓練過程中稱爲『獲勝者』(the winner)，$n_c$則是映射圖上另一節點，$(x_w, y_w)$與$(x_c, y_c)$分別是節點$n_w$與$n_c$在映射圖上的座標。在訓練時，每個節點調適的幅度與這個節點跟獲勝者間的距離有關，距離愈近的節點獲得調適幅度愈大；反之，較遠的節點則調適幅度較小。

在SOM訓練的另一項特色是以訓練次數的多寡來控制每次訓練獲勝者的鄰近範圍以及特徵向量調適的幅度，使得隨著訓練次數增加，鄰近範圍與調適幅度愈來愈小，而保證SOM的訓練結果可以收斂。舉例而言，在第$\tau$+1次的訓練中，對某一節點$n_c$調整的方式如式(2)所示。

$$f_c(\tau+1) \overset{def}{=} f_c(\tau) + h(\tau, d(n_w, n_c))[f_i - f_c(\tau)] \tag{2}$$

式子中，$f_c(\tau)$是表示第$\tau$次的訓練後，節點$n_c$的特徵向量，$f_i$是輸入資料的特徵向量，$h(.)$是一個訓練次數$\tau$與節點和獲勝者之間的距離$d(n_w, n_c)$有關的調適函數，爲節點$n_c$的特徵向量此次訓練的調適幅度，如上所述，當訓練次數愈多，或者距離$d(n_w, n_c)$愈大，$h(.)$所得到的值愈小。

SOM的訓練過程如下。首先，根據輸入資料的數量與特徵向量的維度設定節點的個數與特徵向量的維度，並對每一個節點隨機產生一個特徵向量。在輸入資料後，開始進行多次的訓練。在SOM的每一次訓練中，首先從輸入的資料中隨機選取一個資料，再從節點中選出與訓練資料的特徵向量最相似者，也就是獲勝者。接著如式(2)所示，根據調適函數$h(.)$計算出的調適幅度，調整獲勝者與其鄰近節點的特徵向量，使其愈加相似於訓練的特徵向量。當SOM訓練完成後，便依據術語特徵向量與節點特徵向量的接近程度，將術語映射到圖形上。

在利用SOM技術對文字資料進行叢集或視覺化的研究中，可以依據處理的對象分爲文件與術語兩類。在以文件爲處理對象的SOM研究，大多將輸入的每一筆文件表示成一個以索引詞(index terms)的出現次數爲基礎的特徵向量[11, 12]，因此，索引詞的出現情形較爲接近的文件可以映射到同一節點或鄰近的節點上。爲了使文件的特徵向量可以表示語意訊息，Wermter與Hung利用WordNet的語意階層關係，計數具有相近概念術語的出現次數作爲向量的特徵值，以SOM技術對Reuters新聞語料進行文件分類(text classification)的研究[13]。Kohonen等人則先對術語進行SOM的叢集，使得具有相關語意的術語，映射到同一節點上。再以叢集後的節點作爲基礎，計數節點對應的所有術語出現在文件資料中的次數總和作爲向量的特徵值，作爲資料縮減的技巧來處理極大量的新聞群組(newsgroups)線上文字資料[14]。此外，在文件叢集的應用中，由於以索引詞爲基礎的特徵向量維度非常高，一般的二維映射圖較難表示文件資料間所具有複雜的主題關係，因此，Merkl認爲需要表現出主題間的階層關係，可以利用階層式自組織映射圖(hierarchical self-organizing feature maps)，訓練一組多層的映射圖，使得位置在上層的映射圖之節點表示文件資料中較廣泛的主題，而以下層的映射圖之節點表示較特定概念的主題[12]。

在利用SOM處理術語的研究上，則有Ritter與Kohonen對於英語術語[15]和Ma等人對漢語及日語術語[16]叢集的研究。在術語特徵向量的設定上，Ritter與Kohonen以術語的出現(occurrences)及前後各一個術語的上下文關係(contexts)作爲特徵[15]；Ma等人則利用術語的共現次數爲基礎作爲向量的特徵[16]。

在目前利用SOM技術所進行文字資料叢集或資訊視覺化的研究，其實驗結果可以看出主題相近的文件或術語可以被映射到相同或鄰近的節點，在視覺呈現上，符合人們的認知，這些研究可以證明SOM技術應用於文字資訊視覺化的可行性。然而，從這些研究中卻也可以發現大多數研究在說明實驗結果時，多半以叢集的結果與主題的相關程度進行討論，在客觀的評估方法上也都以傳統資料分類的檢全/檢準(recall/precision)爲標準[16]，甚少討論所得到的實驗結果在不同主題間的關係。但在資訊視覺化的研究中，藉由圖形表示文件或術語之間的分布，是相當重要的目標。在進行這方面的研究時，也應該根據這方面的要求，設計一套合適的評估方法。

## 3 研究設計

本研究是應用SOM技術的初步研究，因此除了提出術語進行資訊視覺化處理的方法之外，如何評估其結果也是重要的研究問題。此外，在現階段的研究中，本論文採用一般的SOM技術作爲探討的對象，先以一般常用的型態與訓練模式做爲SOM的應用，來了解這項應用的可行性。更爲先進與複雜的技術如階層式自組織映射圖[12]，可在後續的研究中進行。以下首先說明以SOM進行術語資訊視覺化的方法，接著提出評估資訊視覺化成效的方法。

### 3.1 以SOM進行術語資訊視覺化的方法

在利用SOM進行術語資訊視覺化的方法中，首先進行術語抽取(term extraction)，從輸入的論文題名、摘要與參考文獻的題名等文字資料，抽取出計算語言學領域中重要的中英文術語[1]。判斷一個出現在文字資料中的字串是否是與這文字資料主題相關的術語可以從字串的『單元完整性』(unithood)與『主題代表性』(termhood)的兩方面著手[17]，單元完整性是指做爲術語的字串是否爲語言結構(linguistic structure)上的完整單位，如詞(words)或詞組(phrases)，而主題代表性則是指此一術語能否代表文字資料的主題並與其他主題區別。在本研究中將以統計訊息爲主，配合若干經驗法則(heuristic rules)來達到這兩項要求。首先將論文資料輸入，建立一個PAT-tree資料結構[18]，接著從PAT-tree檢取所有出現在論文資料中的字串，並計算字串在所有論文的出現總次數、字串在論文資料中的平均出現頻次和標準差(standard

deviation)以及字串前後接字的複雜度等統計資訊。其中，字串前後接字的複雜度(如式(3a, b))，加上停用詞(stop words)不能出現在字串首尾的經驗法則，用來檢測字串的單元完整性。

$$C_{1S} \overset{def}{=} -\sum_{a} \frac{F_{aS}}{F_S} \log(\frac{F_{aS}}{F_S}) \tag{3a}$$

$$C_{2S} \overset{def}{=} -\sum_{b} \frac{F_{Sb}}{F_S} \log(\frac{F_{Sb}}{F_S}) \tag{3b}$$

式(3a)和(3b)中，字串*S*的前後接字複雜度分別以$C_{1S}$和$C_{2S}$表示，*a*和*b*則代表字串*S*在論文資料中任一個可能的前接字和後接字，$F_S$、$F_{aS}$和$F_{Sb}$分別是字串*S*、*aS*和*Sb*的出現總次數。當字串的前後接字複雜度較小時，表示此一字串需與其前面或後面的某一字串共同構成新的字串，才能表示語法和語意上的一個單元。因此，當前後接字複雜度愈大，愈有可能表示一個完整的術語。而所檢出的高頻字串中，字串首尾經常是介詞、連詞或定詞等停用詞，因此我們過濾掉首尾爲停用詞的字串，使得過濾後的術語句有單元完整性的要求。但停用詞出現在中間的字串，如 “part of speech”，只要出現次數夠多、頻率夠高仍爲重要的術語。在另一方面，字串在所有論文的出現總次數、平均出現頻次和標準差則用來表示術語的主題代表性，出現總次數愈大的術語表示這個術語在領域中常被使用而具有重要意義，術語的平均出現頻次和標準差則可表示這個術語在論文中的使用情形，平均出現頻次愈大的術語，即有可能在許多論文中出現多次，是這些論文的重要術語；而術語的出現頻次標準差較大則表示此術語在某些特定論文出現較多次，對這些論文相當重要。所以這三項統計訊息可以作爲檢驗術語是否符合主題代表性的依據。因此，本研究即整合上述的訊息做爲判斷字串是否爲計算語言學領域中重要術語。

接著，對上述步驟所抽取出來的每一個術語設定一個特徵向量來訓練SOM。爲了產生合適的SOM，相關術語所設定的特徵向量必須相接近。如此一來，當把術語映射到SOM時，相關術語將映射到同一節點上或鄰近的節點中，所形成圖形便具有相關術語的距離將較非相關術語的距離小的特性。本研究以術語對每一個術語的共現關係的估算值做爲這個術語的特徵向量，如式(4)表示術語$t_i$的特徵向量$f_i$。

$$f_i = [o_{i,1}, \ldots, o_{i,k}, \ldots, o_{i,N}]^T \tag{4}$$

在式(4)中，假定術語抽取步驟中共得到*N*個術語，因此每一個術語的特徵向量都是一個*N*維的向量。在術語$t_i$的特徵向量$f_i$中，第*k*個元素$o_{ik}$是術語$t_i$與另一術語$t_k$共現關係的估算值。當比較術語$t_i$與$t_j$的相關程度時，可以比較這兩個術語與其他術語$t_k$之間的共現情形。一旦當$t_i$與$t_k$共同出現在某一些論文資料時，同時$t_j$也經常出現在這些論文資料時，術語$t_i$與$t_j$可能相關於同一個特定的主題，這兩個術語便可能相關。如果$t_i$與$t_j$有許多共同的共現術語時，$t_i$與$t_j$的特徵向量便很接近而表示兩個術語間具有較大的相關程度。以數學的方式來表示上述的說明，當我們以歐幾里德距離作爲兩個術語特徵向量之間距離的估算方式時，當兩個特徵向量具有愈多相近的元素，在特徵向量所在的*N*維空間的距離愈小，表示兩個術語的相關程度愈大；反之特徵向量之間相異的元素愈多，距離愈大，兩個術語的相關程度便愈小。

在兩個術語$t_i$與$t_k$的共現關係上，也就是上述特徵向量$f_i$中的元素$o_{ik}$之值，可以利用近來資訊檢索常使用的『隱含語義分析』(latent semantic analysis, LSA)技術[19]來進行估算，使得某些相關術語卻較少共同出現的問題可以減輕。其估算方法如下，我們首先建立『術語-文件矩陣』(term-document matrix)，以每一個抽取出來的術語對應到矩陣中的一行(row)，矩陣中的每一列(column)則對應到一筆論文資料，在矩陣中第*i*行第*p*列的元素，其值爲第*i*個術語在第*p*筆論文資料中出現的次數。接著對於『術語-文件矩陣』進行奇異值分解(singular value decomposition)，求得一組維度較小的新術語向量。比方說新向量的維度爲$\delta$，新的術語向量組便是所有維度爲$\delta$的向量組中，內積的估算值與原先『術語-文件矩陣』的內積誤差最小的向量組之一，術語間共現關係便以這組向量兩兩之間的向量內積值作爲估算值。而且對於缺乏共同出現的術語，此一共現關係的估算方法具有適當的補償效果，使得相關術語的特徵向量較爲接近。因此，本研究所產生的特徵向量可以作爲SOM技術的輸入，所得到的結果將比由『術語-文件矩陣』所估算的共現關係爲佳。

接下來，便對於每一個術語所產生的特徵向量進行SOM訓練。本研究中所採用的調適函數如式(5)所示，

$$h(\tau, d(n_w, n_c)) = e^{-\frac{\tau \times [d(n_w, n_c)^2 + 1]}{\alpha}} \tag{5}$$

在式(5)中，$\alpha$是一個預設的參數值，用來控制訓練次數和獲勝者鄰近範圍中進行調適的節點數量。如同第二節中所提到的，對於某一節點$n_c$，調適函數*h*(.)所產生的調適幅度與訓練次數$\tau$和這個節點與獲勝者$n_w$之間的距離$d(n_w, n_c)$有關。在本研究中採用歐幾里得距離做爲$d(n_w, n_c)$的計算方式。在式(5)中，可以發現在每次訓練中，愈接近『獲勝者』的節點($d(n_w, n_c)$值愈小)，獲得的調整幅度愈大，愈遠離則幅度愈

小；而『獲勝者』是調整幅度最大的節點。而且隨著訓練次數增加，調適的節點數量以及調適幅度都愈來愈小。因此，可以保證在經過多次的訓練之後，所產生的SOM會收斂。

### 3.2 資訊視覺化成效的評估

利用SOM技術進行資訊視覺化的目的是希望當資料被映射到圖形上時，它們的關係仍然可以盡量保持原先在高維特徵向量之間的關係，如此一來，可以從SOM產生的圖形認知原先的資料關係。也就是說，假設任何兩對術語($t_1$, $t_2$)和($t_3$, $t_4$)，每一個術語的特徵向量分別是$f_1$、$f_2$、$f_3$和$f_4$，如果在特徵向量上的距離關係是$d(f_1, f_2) > d(f_3, f_4)$。在經過術語資訊視覺化的過程後，我們希望當術語映射到節點$n_1$、$n_2$、$n_3$和$n_4$時，可以發現$n_1$、$n_2$、$n_3$和$n_4$在圖形的位置上，其歐幾里得距離具有$d(n_1, n_2) > d(n_3, n_4)$的關係。

所以，在比較應用SOM進行資訊視覺化的成效時，可以先計算出每一對術語在特徵向量的距離，在將術語映射到圖形後，再以所映射的節點計算術語在圖形上的距離，最後再計算這兩種距離的相關係數(correlation coefficients)，做爲資訊視覺化成效的評估標準，相關係數較小，表示SOM的結果較不理想；相關係數愈大，則表示SOM所產生的圖形保留愈多原先在高維特徵向量上的關係，可以從圖形上認知術語的叢集以及分離的關係，進而探索研究主題彼此之間的關係。

## 4 結果與討論

本論文以第一屆(1988)到第十四屆(2001) ROCLING研討會的235篇論文資料做爲分析計算語言學主題的素材，從這些論文的題名、摘要及參考文獻的題名中，抽取重要的術語，並將術語的關係視覺化。進行術語抽取時，本論文字串出現總次數的閾值設定爲20次，平均頻次和標準差的總和設爲2.5，前後接字的複雜度則設爲0.5，結果共得到229個術語。

接著將所抽取出來的229個術語，利用LSA技術估算彼此間的共現關係，建立各個術語的特徵向量。最後以術語的特徵向量進行SOM訓練，在本研究中，我們以20×20個節點進行實驗，測試訓練次數以及第3節式(4)的參數$\alpha$之影響結果。在實驗中，參數$\alpha$分別設定爲250、150、 50和25，每一個不同的$\alpha$值，進行三次試驗，記錄訓練次數0 (初始)、10、50、100與200等各次的相關係數。取三次試驗中第200次訓練獲得較佳結果的試驗，也就是$\tau$=200時相關係數最大者，進行比較。實驗的結果所產生的相關係數，如表1各欄所示。

表1 以自組織映射圖進行術語資訊視覺化的實驗結果

| 訓練次數$\tau$ | $\alpha$=250 | $\alpha$=150 | $\alpha$=50 | $\alpha$=25 |
|---|---|---|---|---|
| 0 | 0.07 | 0.06 | 0.07 | 0.08 |
| 10 | 0.54 | 0.52 | 0.44 | 0.24 |
| 50 | 0.36 | 0.52 | 0.44 | 0.34 |
| 100 | 0.30 | 0.49 | 0.42 | 0.32 |
| 200 | 0.29 | 0.50 | 0.41 | 0.32 |

從表1的結果，我們可以看到幾個現象。(1) 初始的時候，映射圖仍未組織化，術語映射到圖上的各個節點上，其距離與特徵向量的距離無關，因此，相關係數不高，各欄均在0.06至0.08之間，顯示此時除了少數的相關術語映射到相同的節點上，大多數的術語的相關程度未能映射到圖形中。(2) 經過幾次訓練之後，映射圖上節點的特徵向量已經依照某種規則排列，此時的實驗結果獲得較大的相關係數，顯示若干相關的術語已經被映射到鄰近的節點中，比方說，以$\alpha$=150一組的數據爲例，在訓練次數超過10次之後，相關係數約爲0.50。(3) 各欄的資料也表示，訓練次數相當大時，本研究提出的SOM技術可以收斂。(4) 如第3節中所提到參數$\alpha$可以控制調適的節點數量，$\alpha$值愈大，調適的節點數量愈多。從實驗中，我們發現$\alpha$值過大，在訓練的過程中較不穩定；但較小的$\alpha$值，卻很容易收斂到較爲次佳的結果。在本研究的實驗中，以$\alpha$值爲150所得到的結果，較令人滿意。(5) 然而必須加以說明的是在SOM的訓練模式中，是以輸入的特徵向量對映射圖進行組織化，並不是對資訊視覺化的評估條件進行最佳化。因此，相關係數並不會呈現單調遞減的情形。而且，相關係數雖然可以提供客觀的評估標準，然而所得到的結果還需要進一步呈現來加以詮釋，才能看出SOM技術運用在術語資訊視覺化的成效。

因此，除了以相關係數來衡量資訊視覺化的成效之外，最爲重要的仍是經實際產生的映射圖所表達的訊息，我們將上面實驗中所得到較佳的結果之一，$\alpha$值爲150、訓練次數50次所得到的映射圖，呈現在

圖1中。從圖1中，我們可以發現大多數相關的術語都被映射到同一節點或是鄰近的節點上，比方說。在映射圖下方，所包括的術語大多與語言學研究相關，如最左邊的 “syntax”、“functional”、“syntactic” 、“semantic” 、“lexical” 、“semantics” 、“lexicon” 以及“verb”。以及較右邊的“剖析”、“名詞”、“結構” 、“語法” 、“動詞” 、“詞類” 、“語意” 以及“詞彙”。又如在橫軸的16，縱軸10到12的地方可以發現這裡的術語都與語言模型的研究相關，如“bigram”、“language model”、“language modeling” 、“language models” 、“clustering” 、“class based”以及“n gram”。因此，在映射圖上可以發現主題相關的術語會形成叢集，我們可以依據圖1的相關術語分布情形，將幾個較大主題叢集表示成圖2。

除了相關的術語會映射在相近的節點上，從圖1與圖2也可以顯示在映射圖上距離很接近的主題具有相關性，比方說，『機器翻譯』(machine translation)相當接近於『剖析器與文法規則』(parser and grammars)與『語法與語意』(syntax and semantics)的研究，表示語法、語意、文法規則以及剖析器經常應用在機器翻譯的研究。『語音處理』中各個主題，包括『語音合成』(speech synthesis)、『語音辨認』(speech recognition)、『語言模型』(language models)等主題，彼此間也很接近。另外，映射圖上方的『斷詞』(word segmentation)、『未知詞偵測』(unknown word detection)與『詞類標示』(part-of-speech tagging)等相鄰近的情形，可以推測這些主題之間有相關性。圖形上『資訊檢索』(information retrieval)相關的主題，除了『斷詞』以及『語言模型』之外，還有『摘要』(summarization)。整體的圖形看來，偏左偏下的部份與語言學研究相關，而右上則是各種的技術應用與系統製作的研究，如『資訊檢索』和『語音處理』等各種主題便在圖形的右方。

然而，由於術語的數目相當龐大，特徵向量的維度也相當高，事實上，也有若干的術語映射結果並不理想，比方說，的“pat”與“tree”等術語所表示的PAT-tree是資訊檢索中重要而常用的技術[18]，但在這個映射圖上並沒有和位於橫軸19，縱軸17處的『資訊檢索』主題相鄰。此外，整個圖形中最明顯的現象是中英文同義或相關的術語雖然在圖形上它們的位置已經相當接近，但仍然可以認爲是分離。比方說，圖1中分布在圖形橫軸12到18，縱軸8到10處的三個同義的術語，“語音辨認”、“語音辨識”和“speech recognition”。這個現象表示即便我們利用參考文獻的題名做爲輸入資料以及LSA來進行補償，但中英文的資料仍然有區別，在論文資料中缺乏共現關係，使得中英文同義或相關的術語在圖形上相近但無法映射到同一節點上。

## 5 結論

本論文的研究利用自組織映射圖(SOM)技術將計算語言學相關術語對應到二維圖形，使得術語之間的關係可以在映射圖中加以呈現，提供使用者做爲資訊檢索以及了解研究領域的重要主題的輔助工具。在本論文中，我們所探討的問題有(1)發展SOM技術應用到術語資訊視覺化的方法，(2)評估SOM技術應用到術語資訊視覺化的成效，(3)利用研究結果分析計算語言學中重要的研究主題與主題之間的關係。在SOM技術的應用中，本研究首先從論文資料中利用字串出現的統計訊息以及經驗法則，抽取重要的術語。接著以術語之間的共現關係做爲基礎，建立每一個術語的特徵向量，以特徵向量之間的距離表示術語的相關性，愈相關的術語，特徵向量間的距離愈小。再以術語特徵向量做爲輸入資料，進行SOM訓練並將術語映射到圖形上，使得特徵向量距離相近的術語映射到同一節點或鄰近的節點上。如此一來，利用術語在映射圖上的分布情形，便可以輔助使用者認知研究主題之間的關係。對於這項技術的成效評估，我們建議將特徵向量的距離與節點位置的距離進行相關係數的計算，以所得到的相關係數大小做爲成效評估的標準。最後，對於計算語言學領域，以ROCLING論文集的論文資料做爲研究對象，進行術語資訊視覺化的實驗。在經過若干次的訓練之後，映射圖逐漸組織化。因此，術語特徵向量的距離與所映射節點位置的距離之相關係數增加。並且從實際產生的圖形中可以觀察出，大多數相關的術語都可以映射到相鄰近的節點上，在映射圖上所形成的叢集與計算語言學的主題相關，而這些叢集在圖形上的位置也可以確實地表現計算語言學主題之間的關係。值得一提的是，本研究所提出的方法並不會因爲論文資料數量增多，而增加時間與記憶體等計算資源的需求。由於這方法需要較多計算資源的階段是在SOM的訓練過程。而我們以術語的特徵向量做爲SOM的訓練資料，在特定領域中，術語的數目極爲有限，其數量並不會隨論文數目增多而快速成長，而且可以在術語抽取的階段，藉由參數的設定，只選取出現次數較多並較重要的術語，所以可以控制所需的計算資源，因此這個方法具有可升級性(scalability)。這些都顯示了SOM技術應用到術語資訊視覺化的可行性。

在進一步研究的建議上，除了進一步了解SOM在術語視覺化的能力與極限之外，比方說在產生映射圖之後，可以進一步自動將節點再度叢集、歸類，使得使用者更能解讀領域內的主題與趨勢。此外，階層式自組織映射圖等更先進的映射圖型態與技術可以表現出術語的概念階層(conceptual hierarchy)，將可以提供更有效的資訊組織工具。而如何應用本研究發展出來的成效評估方法，使得SOM更有效率、結果

更有用將也是發展的目標之一。再者，可以利用術語與論文之間的關係，產生論文的特徵向量，將論文映射到節點上，根據論文發表的年代，觀察研究主題的發展趨勢，對於了解研究領域的知識結構將有幫助。

## 參考文獻


[1] 林頌堅, “基於自然語言處理技術的研究主題抽取與分析,” Proceedings of ROCLING XV, pp. 231-256.

[2] P. Srinivasan, “Thesaurus Construction,” Information Retrieval—Data Structures & Algorithms, edited by W. B. Frakes and R. A. Baeza-Yates, Prentice-Hall, Inc., pp. 161-218.

[3] H. Chen, T. Yim, D. Fye, and B. Schatz, “Automatic Thesaurus Generation for an Electronic Community System,” Journal of the American Society of Information Science, Vol. 46, No. 3, pp. 175-193.

[4] Y-H. Tseng, “Automatic Thesaurus Generation for Chinese Documents,” Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, pp.1130-1138.

[5] S. K. Card, J. D. Mackinlay, and B. Shneideman “1 Information Visualization,” Readings in Information Visualization— Using Vision to Think, Morgan Kaufmann, pp. 1-34.

[6] S. Huang, M. O. Ward, and E. A. Rundensteiner, Exploration of dimensionality reduction for text visualization. Technical Report TR-03-14, Worcester Polytechnic Institute, Computer Science Department, 2003.

[7] T. K. Landauer, D. Laham, and M. Derr, “From Paragraph to Graph: Latent Semantic Analysis for Information Visualization,” Proceedings of the National Academy of Science of the USA, Vol. 101, pp. 5214-5219.

[8] A. Flexer, “On the Use of Self-organizing Maps for Clustering and Visualization,” Intelligent Data Analysis, Vol. 5, pp. 373-384.

[9] X. Lin, “Visualization for the Document Space,” Proceedings of IEEE Visualization 1992, pp. 274-281.

[10] T. Kohonen, Self Organizing Maps, Springer Verlag.

[11] X. Lin, D. Soergel, and G. Marchionini, “A Self-organizing Semantic Map for Information,” Proceedings of SIGIR 1991, pp. 262-269.

[12] D. Merkl, “Exploration of Text Collections with Hierarchical Feature Maps,” Proceedings of SIGIR 1997, pp. 186-195.

[13] S. Wermter and C. Hung, “Selforganizing Classification on the Reuters News Corpus,” Proceedings of COLING 2002.

[14] T. Kohonen, S. Kaski, K. Lagus, and T. Honkela, “Very Large Two-Level SOM for the Browsing of Newsgroups,” Proceedings of ICANN 1996, pp. 269-274.

[15] H. Ritter and T. Kohonen, “Self-organizing Semantic Maps,” Biological Cybernetics, 61, pp. 241-254.

[16] Q. Ma, M. Zhang, M. Murata, M. Zhou, and H. Isahara, “Self-organizing Chinese and Japanese Semantic Maps,” Proceedings of COLING 2002.

[17] K. Kageura and B. Umino, “Methods of Automatic Term Recognition-A Review,” Terminology, Vol. 3, No. 2, pp. 259-289.

[18] G. H. Gonnet, R. A. Baeza-Yates, and T. Snider, “New Indices for Text: PAT Trees and PAT Arrays,” Information Retrieval—Data Structures & Algorithms, edited by William B. Frakes and Ricardo Baeza-Yates, Prentice-Hall, Inc., pp. 66-101.

[19] S. Deerwester, S. T. Dumais, G. W. Furnas, Thomas K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” Journal of the American Society for Information Science, 41(6), pp. 391-407.

圖1 利用SOM技術對計算語言學術語進行資訊視覺化的結果

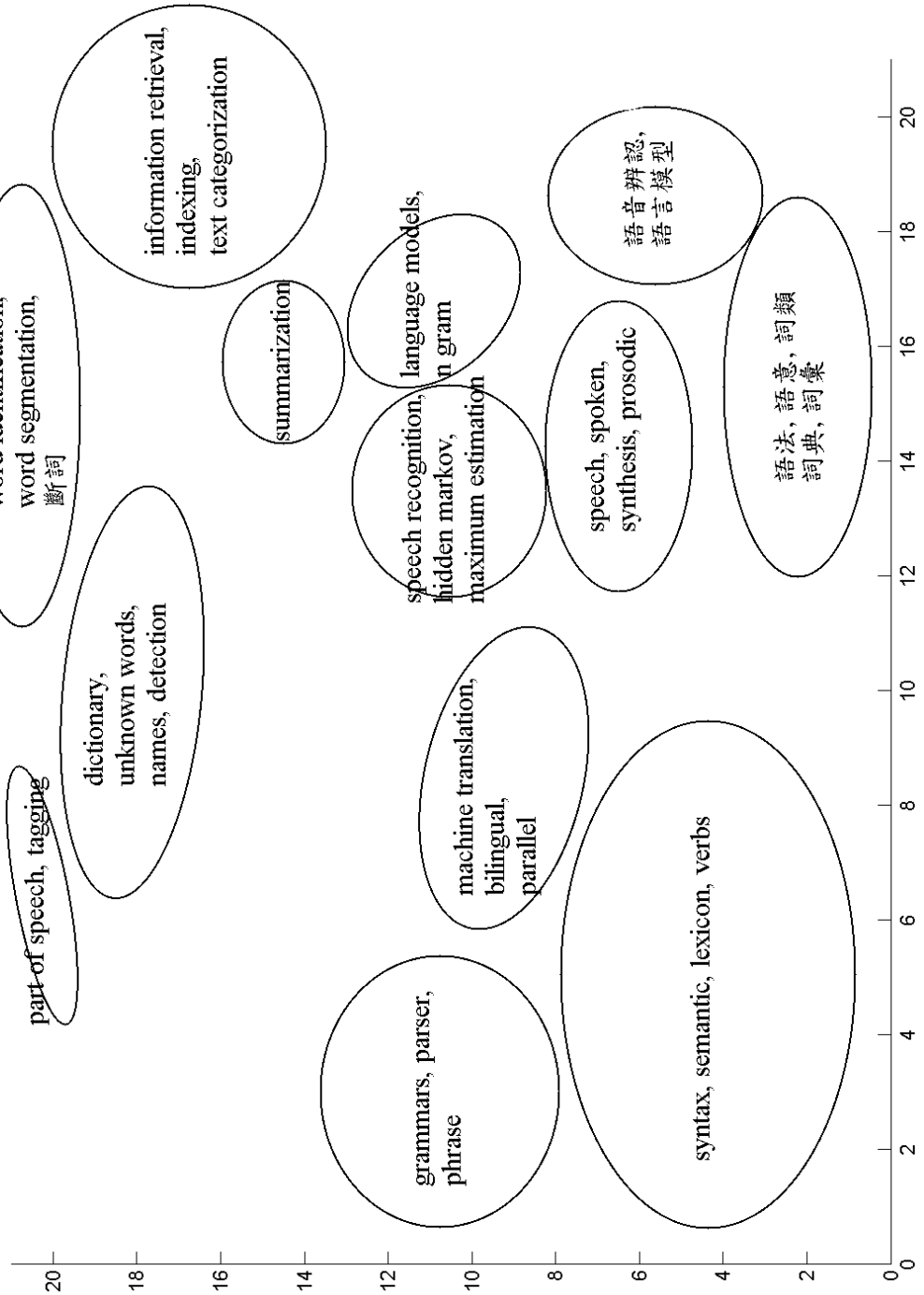


圖 2　計算語言學術語主題叢集的分布情形