自動化研究主題探勘方法及其在計算語言學之應用

林頌堅

助理教授
世新大學資訊傳播學系

摘要

由於科學研究的規模日益龐大而且研究的工作也愈來愈複雜，研究人員與科技管理人員需要一套能夠有效地探勘研究主題的方法。過去我們針對這個問題提出一系列文本處理與文字資訊探勘的技術，其中主要爲關鍵語詞抽取技術以及資訊視覺化技術。關鍵語詞抽取技術以研究領域中的論文文字資料做爲輸入，自動化抽取關鍵語詞來代表領域中的重要主題。資訊視覺化技術則將這些語詞和它們之間的關係呈現在二維的圖形，提供使用者可以透過產生的圖形了解該領域的重要主題和它們的發展情形。其餘還包括了語詞共現估計、主題相關程度計算以及論文映射等技術。本論文將這些技術整合起來並應用到國內的計算語言學領域，確認這個領域研究與發展的重點。結果發現計算語言學早期著重於各種語言知識的計算理論化，以因應機器翻譯的需求；中期和近期則有語音處理和資訊檢索等更多的應用出現，而應用的技術則傾向採用具有強健與容易實作等特性的統計導向方法。

關鍵詞：研究主題探勘，文本處理，文字資訊探勘，資訊視覺化，計算語言學

緒　論

學術領域的研究發展趨勢可說是一種「創新的擴散」(diffusion of innovation)(Rogers, 1983)的傳播現象。學術研究的創新是在於新研究問題的提出、新研究方法與技術的運用以及新理論的發現等等。一個學術領域的研究和發展之趨勢可以從相關論文的發表數量之變化情形來加以了解。當某一主題相關的論文增加時，可以認爲在這個學術領域正有許多研究資源與學術工作者投入此主題的研究，所以有較多成果發表；反之，在領域中愈來愈少論文所提及的主題，則可能表示這個主題正在沒落。在主題開始發展的時候，只有極少數的相關論文發表；經過一段時間的努力，取得令人矚目的研究成果後，此刻有較多的研究計畫開

展，投入較多的資源，而且也吸引較多的學術工作者發表相關論文，經由他們的交流與互動，形成研究這些主題的社群(community)，這個主題相關研究的正式與非正式的傳播也將透過這個研究社群進行。在持續投入資源後，將使得研究成果獲得飛躍似的進展，也促成社群發行期刊與並舉辦研討會，學術工作者的交流將更加密切，一方面使得研究社群更加組織化，同時並造成相關主題論文的大幅增加。但在歷經一段時間的發展後，如果沒有獲得重大的研究成果，主題相關的論文數量成長將會趨緩，甚至消退。因此某一特定主題相關論文的累積數量呈現類似Logistic曲線的成長(Price, 1963; Crane, 1972)。因此，論文數目增長的原因，除了新增期刊、研討會或者電子出版，使得傳播管道拓寬將會造成之外，最重要的原因可說是有創新主題的論文之發表，爲研究發掘新的研究問題或者提供新的理論或方法，促使在一段時間內有更多研究人員對這些新的問題進行研究或者採用這些理論與方法，投入相關的研究工作(Tabah, 1996)，產生更多的研究成果，而記載這些成果的論文在期刊或者研討會獲得發表的可能性比較高。

對於研究人員而言，掌握研究領域中正在發展的主題，進行相關的研究，較容易取得學術資源，也較容易獲得發表的機會。另一方面，制定科技政策、擁有資源分配權力的科技管理人員則必須廣泛地了解各個學術領域內各項主題的發展趨勢並且熟悉本國的優勢，才能作出合適的科技發展計畫，使得國家的人力、預算與精密儀器等有限的研究資源能夠充分發揮(行政院國家科學委員會科學技術資料中心，2003)。雖然，近年來科學研究取得了很多重大的成果，解決了不少關於自然、社會與人類的問題，但隨著科學研究愈發的進步，研究的問題愈來愈困難，而且愈來愈複雜。現今的科學研究走向需要大量人力與物力資源的「大科學」(big science)以及從不同面向解決問題的科際整合研究(interdisciplinary research)。研究專殊化的結果則使得研究人員要能夠解決他們所面對的問題往往需要長時間的訓練與學習。從期刊與論文的數量與篇幅等研究資源快速成長的現象(Meadows, 1998)便足以說明在科技研究與管理上所面對的困難與複雜性。這些的趨勢造成了研究人員與科技管理人員愈來愈難以掌握他們所需要面對的研究主題。

要解決上述問題，可以利用電腦從大量論文資料中自動發掘學術領域的研究趨勢。傳統上，論文是學術領域研究的主要產品，也是要了解研究領域的發展與最新趨勢時所必須取得的資源。研究人員在閱讀學術論文時，他們從論文的題名、摘要和內文等文字資料解讀與理解論文所陳述的研究問題以及進行研究時所使用的方法、理論與技術等主題。電腦科學家基於這個想法，利用電腦強大的儲存與處理能力，根據語言知識或統計上的特徵，模擬人類的認知及處理資訊的能力，進行自動化分析，找出論文中可能包含的主題，提供資訊檢索、資訊過濾、文件分類、摘要等技術(Salton & McGill, 1983)，便於研究人員從大量的學術出版品中，有效率的取得主題相關的所需資訊。科技管理人員也常利用SCI、SSCI等書目引文資料庫，藉由論文發表數、論文被引數、影響係數(impact factor)等書

目資料，進行標準量化指標的統計分析，調查各學術領域的發展現象，做爲考核研發投入與制定科技政策的參考(行政院國家科學委員會科學技術資料中心，2003)。因此，整合這兩種研究所得到的理論與技術，利用電腦技術分析學術領域的研究主題發展趨勢應屬可行。進一步來說，利用電腦繪圖技術將重要的分析結果顯示出來，勢必有助於對學術領域研究有更深更廣的認知與了解。書目計量學研究(bibliometrics)也以論文、期刊、作者和語詞等在論文資料中的共現關係(co-occurrences)(Börner, Chen & Boyack, 2003)，透過「資訊視覺化」(information visualization)技術(Card, Mackinlay & Shneiderman, 1999)以各種映射(mapping)與繪圖技術將論文資料投射到二維或三維的空間上，並在螢幕上顯示，提供使用者對論文之間的關係產生視覺認知，進行更有效率的互動。一般而言，經過資訊視覺化處理後，可以使得主題相關的論文在圖形上的距離較近，形成叢集(cluster)；因此，將學術領域中所有發表的論文映射到圖形上，形成的各個叢集便可以表現此一學術領域內所有相關的研究主題與它們整體的分布情形。如果將圖形按照論文的發表時間依序顯示，各時期對應的圖形中，論文分布較密集的主題便是這個時期此一學術領域較爲重視的研究方向，將連續數個時期所產生的圖形相互比對，便可以了解領域內各主題相關研究的消長情形。因此，將某一學術領域所發表的論文進行資訊視覺化不僅可以提供方便有效的檢索介面，還可做爲分析此一學術領域研究發展趨勢的工具。研究人員便可依據所產生的圖形，針對他的研究問題，瀏覽相關的研究主題，深化目前的研究工作，也可以藉由對研究領域的全盤認知，激發研究的靈感，產生創新的研究。科技管理人員也可以透過學術領域研究發展趨勢的分析結果，找出具有優勢與研究潛力的研究主題，鼓勵研究人員投入，並且支援研究所需的經費與設備。

本研究的目的在整合一系列先前所提出的文件處理與文字資訊探勘技術(林頌堅，2003a；林頌堅，2004)，以研究領域中的論文文字資料做爲輸入，分析論文資料中的關鍵語詞，自動化產生領域相關的重要主題，並且利用資訊視覺化技術將這些主題之間的關係呈現在一個二維的圖形上，使得使用者可以透過這個圖形了解該領域的重要主題和它們的發展情形。這些技術包括關鍵語詞抽取技術、語詞共現估計方法、利用自組織映射圖的領域主題視覺化技術以及論文投射技術等。過去的研究領域主題探勘方法多以利用引用文獻(citations)中的作者、期刊或論文本身等資訊的共現關係，這些資訊在統計分析時需要足夠多的數量，統計上才有意義。因此，目前的這方面的研究多從SCI或SSCI等論文引用資料庫檢索相關的資料，而且分析的領域必須是規模較大或較爲活躍的領域，才能取得較多數目的論文引用資料進行分析。大多數新興的研究領域在論文引用資料庫中出現的數目卻仍然不足以進行可靠的統計分析。另一方面，利用作者或期刊等共同被引用的關係所產生的結果，在解讀上較爲困難。使用這些結果來進行分析的研究人員或科技管理人員本身必須相當熟悉這個研究領域的期刊與作者，才能對所呈現的

圖形解讀出有意義的資訊，因此這些方法無法爲最需要幫助的新手研究人員提供有效的幫助。我們認爲使用語詞作爲分析與結果呈現的對象，將可以解決上述的問題。相對於其他由共被引資料所產生的現象而言，語詞在論文文字資料中的共現現象明顯地次數較多，換言之，在利用語詞的共現關係估算語詞的相關程度時可以獲得統計上較爲可靠的結果。並且語詞本身便具有明確的語意，許多的語詞從字面上便可以了解它們所代表概念之間的關係，不僅在結果的解讀時較爲容易，而且最爲重要的是在技術發展的階段，可以檢視相關結果，加以修正或調整，發展更有效率的技術。因此，本研究所使用的技術將以語詞爲統計分析的對象，並作爲結果的呈現，提供使用者準確而且較容易解讀的結果。此外，這些技術與整合方法都是特別考慮目前台灣的學術研究現況而設計的。台灣的許多研究領域的期刊或者學術研討會接受以中文或英語發表的論文，因此以論文的文字資料輸入時，系統必須能夠同時處理中文或英語的資料，而且在訓練自組織映射圖時，能夠估算出這些語詞的相關程度。這些問題都將在本研究所提出的技術中加以克服，第二節將對這些技術做一詳細的描述。

爲了驗證這些研究主題探勘技術在應用上的可行性，我們將以台灣的計算語言學研究進行分析。計算語言學（computational linguistics）領域是透過電腦的運算來分析、探索及模擬口語與書面語等語言傳播過程中人類的語言表達以及理解的一門研究領域（Hausser, 2001）。選擇計算語言學做爲探討學術領域研究發展趨勢的起點，是著眼在這個學科具有許多特色，包括這個領域的快速發展、高度的科際整合研究、基礎理論與實務應用並重、中文獨特的研究等。顧名思義，計算語言學的研究主要整合了語言學（linguistics）與電腦科學（computer science）兩個學科的理論與技術，其目的在於利用電腦的運算來分析、探索與模擬口語和書面語等語言傳播過程中人類的語言表達以及理解。計算語言學的研究人員提出各種計算化（computational）的語言理論，並利用以這些語言理論製作出的系統應用到各種文字和語言資料的處理上，發展有效而方便的人機互動（human-computer interaction）方式與資訊處理技術。近年來，由於文字與語言資料的快速增加，輸入、處理與使用這些資料的需求也隨之快速成長，計算語言學研究在這個方面獲得了很豐碩的成果，提出各種機器翻譯（machine translation）、語音處理（speech processing）和文本處理（text processing）技術。而且計算語言學不僅在實務性的技術有很大的突破，在應用這些技術的過程中，語言理論獲得了相當的進展。而且這些研究對中文而言更爲重要，中文本身具有非常多與西方語言不同的語言特性，無法直接利用其他語言已經發展出來的理論與技術，但台灣的計算語言學界也針對中文，發展出許多獨特的理論與技術。因此在分析研究領域的主題時，台灣計算語言學研究是一個相當值得研究的對象。第三節將報告以台灣的計算語言學爲分析對象，利用本論文所提出的方法所得到的研究成果，使讀者更進一步了解計算語言學研究的發展以及本研究所使用技術的可行性。

二、研究主題探勘方法與相關技術

本研究所提出之研究主題探勘方法的處理過程如圖1所示。首先，蒐集研究領域中相關期刊與研討會出版的論文資料，建立論文資料庫，並從論文的文字資料中抽取關鍵語詞，以關鍵語詞代表研究領域中的重要主題，並以關鍵語詞在論文資料中與其他語詞的共現關係建立相對應主題的特徵，做爲產生這個領域之主題關係圖的資訊。最後，再依據論文中的語詞出現情形，將論文映射到主題關係圖上，以論文投射的數目估計各個主題在領域上的分布情形，研究人員並可以依序分析與比較不同年代主題關係圖上的分布情形，對分析領域的發展情形有更深的認識。以下分別說明各步驟的細節。

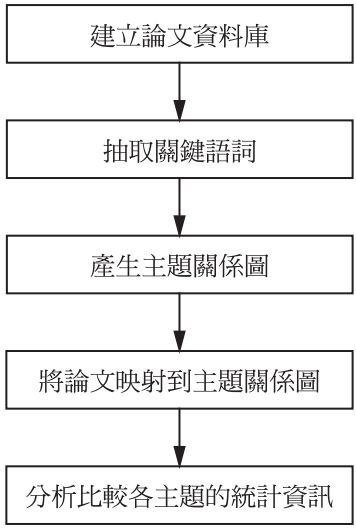

圖1　研究主題探勘方法處理流程

(一)建立論文資料庫

進行領域研究主題探勘的第一步驟是蒐集研究領域內的相關論文資料。對於研究領域的定義，本研究認爲，一個研究領域是由研究社群所面對研究問題上的所有相關概念，用來表現這些概念的符號以及承載這些符號的文本(text)等部分共同構成。問題的相關概念包括了問題本身、解決這個問題所使用的方法與技術和由研究成果所歸納的理論等，文本則是透過研究社群進行傳播的實體，研究社群的成員可以透過相關的文本上記錄的符號來了解問題、過程與成果等研究上的相關資訊。研究領域中最常見的文本是論文，而最常見的符號便是論文中的文字資料，論文中出現的符號還包括圖表、演算法、數學式與化學反應式等，文本還包括其他形式的書籍、技術手冊、專門詞典與索引典。當然若考慮研究社群中非正

式傳播的範疇，領域內的文本還包括研究社群的成員之間交換的相關書信與口語等。近年來由於電子媒體的進展，文本的種類更擴及了各種資料庫和影片等。既然文本中包含了研究問題所需要的資訊，而且相關研究社群的成員利用特定的符號來表示研究上的概念，因此，如果要分析相關研究社群所重視的研究主題，可以對一個研究領域蒐集相關的文本，對符號加以辨認與處理，找出統計上具有明顯特徵的符號做爲研究領域中重要主題的代表。本研究便依據這個想法，針對在特定研究社群發表的論文中包含的主題資訊進行分析，論文蒐集的範圍包括由這個研究社群所認可的期刊或者所舉辦的研討會上發表的論文，而以論文中的關鍵語詞做爲主題探勘的對象。

本研究以所蒐集論文的題名、作者姓名、發表年代、摘要和參考文獻的題名等資料建立資料庫。其中題名和摘要等文字資料將用來抽取關鍵語詞來代表研究領域的重要概念；而其餘的論文資料如作者姓名和論文發表年代等資料則將可用於進一步的資料分析。論文的題名和摘要是論文內容的濃縮，研究人員可以透過其中包含的文字資訊，對論文的主題有初步的認識。以本論文爲例，從本論文的題名和摘要，讀者可以了解本論文是有關利用自動化方法呈現某一特定研究領域的研究主題方面之研究。因此，如果將出現在題名和摘要中的關鍵語詞辨認出來，必然可以代表該論文的主題。許多資訊檢索系統便是利用這個概念，提供使用者輸入相關語詞來檢索所需要的論文資料（Salton & McGill, 1983）。本論文也援用相同的想法，從題名與摘要中抽取關鍵語詞。

特別值得一提的是，除了從題名與摘要中抽取關鍵語詞，在本論文使用的分析方法中還可以加入參考文獻的題名。參考文獻的引用是論文的作者利用前人的研究成果支持本身的研究，或以前人的研究爲基礎，進一步對研究問題進行探討、對研究方法與技術加以改進以及對理論的修正（Garfield, 1979）。因此，論文的主題與論文中出現的參考文獻的主題之間必然有某種程度的相關，從前面的論述中，又可以知道論文的題名與論文本身的主題相關，所以論文與它引用的參考文獻題名可能有相關的主題。加入參考文獻的題名來進行關鍵語詞的抽取，將使得文字資訊更豐富，可以增加關鍵語詞在統計上的可靠性。除此之外，目前國內的許多研究領域同時接受以中文或英語書寫的論文。這個研究主題探勘方法在建立語詞的相關性時，必須特別考慮不同語言但表示同一概念的語詞。由於許多論文可能只有中文的題名與摘要，作者並未提供英語的題名與摘要；反之，以英語撰寫成的論文則僅有英語題名與摘要，缺乏中文題名和摘要。若是只有題名和摘要等論文文字資訊，要利用語詞在論文資料的共現現象做爲建立不同語言語詞之間的相關性相當困難。在論文的文字資訊中加入參考文獻的題名的技巧便可以減輕這個問題。在國內發表的論文在引用參考文獻時，往往會同時引用中文和英語的文獻。在論文文字資訊的處理上加入參考文獻的題名，即便是不同語言的語詞也可以產生共現現象，能夠估算出它們的相關程度。此外，對於相同概念但不同

形式的語詞，比方說，「text categorization」或「document classification」，不同的作者往往習慣於只使用其中的一種語詞，但在引用的參考文獻中，兩種不同的語詞都可能出現在題名上。因此，利用上述的技巧也可以解決這方面的問題。

(二) 抽取關鍵語詞

所謂的語詞在本論文中是由特定序列的中文字和英文詞所構成，指在論文中表示某一個特定概念的語言單位。語詞可以包含單一個詞，如「研究」或「research」，或一個詞組，如「研究領域」或「research domain」；但不可以是不具有意義的字或詞的片段，比方說「而將」或「is a」。由於語詞定義爲一個表示特定概念的語言單位，進一步來說，在論文的文字資訊中出現的關鍵語詞是作者用來表達此篇論文所探討的研究問題、所根據的理論和所使用的研究方法與技術等主題，當論文出現某一具有特定概念的語詞時，表示這篇論文涉及此一語詞所涉及的主題，而且出現愈多某一類主題的相關語詞時，這篇論文愈有可能和這個主題有關。以本論文的題名與摘要來舉例，本論文的研究主題表現在題名與摘要中經常出現的語詞，如「主題」、「研究領域」、「計算語言學」等，讀者可以知道我們所探討的問題是有關研究領域的主題探勘，並且以計算語言學領域做爲實際探討的對象。因此將論文中的關鍵語詞抽取出可以代表這篇論文的主題，而將一個研究領域發表的論文集合起來，對這些論文進行關鍵語詞抽取，並統計所有抽取出來的語詞，找出具有統計意義的語詞，可以代表這個領域中重要的主題。所以，在論文資料庫建立後，我們接著從論文的題名、摘要和參考文獻題名等文字資料中抽取關鍵語詞。

一般認爲在抽取關鍵語詞時，必須要先確認出文字資料中所有的詞，再利用詞的詞類(part of speech)以及語法結構(syntactic structure)，辨認出相關的詞組。對英語的文字資料來說，詞和詞之間有明顯的界限，詞的確認相當容易，而且根據詞的出現型態，可能的詞類十分有限，再加上英語有非常多語法結構的計算理論與資源，非常合適於上述的處理過程。然而，中文的文字資料沒有明顯的界限，因此在確認中文詞方面相當困難，所以許多學者認爲在處理中文文字資訊前，須要先經過斷詞處理(word segmentation)(Chen & Liu, 1992; Wu & Tseng, 1993; Sproat, et. al., 1996)。但斷詞處理的過程中很可能會產生錯誤，這些錯誤將會影響到後續的詞類標示(part-of-speech tagging)和詞組辨認(phrase recognition)等處理，而且在處理包含許多特有術語的論文文字資訊，產生斷詞錯誤的可能性更高。另外，在中文裡，具有同樣型態但不同詞類的詞相當常見，在詞類標示和詞組辨認較爲困難。在考慮這個探勘方法所處理論文資訊將可能包含中文及英文兩種語言，詞語抽取上必須使用中文或英文都適用的語詞抽取技術。本論文所使用的方法是將所有出現在文字資訊中的字串，一一擷取出來，作爲可能的候選詞語，再以字串的統計訊息與經驗法則做爲語詞篩選的依據(林頌堅， 2003a)。

爲了能夠完整而且有效率地擷取出論文資料中所有出現的字串與它們的統計訊息，我們利用PAT tree資料結構（Chien, 1997）來組織這些字串。在這個樹狀結構中，從第一層的某一節點到其下層的另一節點的路徑代表某一個出現在文字資訊中的特定字串，而且在下層的節點中還儲存了相對應的字串在論文資料中的出現總次數、出現文件數、在各論文中出現的次數等統計資訊。舉例而言，從第一層的節點到第L層的節點，長度爲L的路徑代表相對應的字串爲$c_1c_2\cdots c_{L-1}c_L$，第L層的節點中儲存了字串$c_1c_2\cdots c_{L-1}c_L$在論文資料中的統計資訊。而且很明顯的，在路徑上，所有以c_1爲開頭而長度小於L的字串c_1、c_1c_2一直到$c_1c_2\cdots c_{L-1}$都是這個字串的子字串（substring）。當論文文字資訊輸入PAT tree索引系統後，所有出現在論文文字資訊的字串與它的相關統計資訊都可以在所產生的PAT tree中擷取出來，作爲候選語詞進行進一步的篩選。

接下來，利用語詞的單元完整性（unithood）和主題相關性（termhood）（Kageura & Umino, 1996）來篩選所有從PAT tree擷取的字串。單元完整性是指做爲候選語詞的字串是否爲語法結構上的完整單位，如詞或詞組，本研究以式（1）所定義的前後接字複雜度和停用詞（stop words）不能出現在字串首尾的經驗法則，來檢測候選語詞的單元完整性。

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

$$C_{2S} \overset{def}{=} -\sum_{b} \frac{F_{Sb}}{F_S} \log\left(\frac{F_{Sb}}{F_S}\right) \qquad (1b)$$

式（1a）和（1b）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，表示此一字串前後出現的字種類有限且組合情形單純，可能表示這一字串須與其前面或後面的某一字串共同構成新的字串，才能作爲語法結構和意義上的完整單元。比方說，當我們在論文資料中看到一段字串「訊檢」，因爲這個字串前面出現的字種類相當有限，絕大多數的情形下是「資」字，而後面出現的字也以「索」字爲主，因此，這個字串相當有可能是「資訊檢索」這個語詞的一個部分，我們便可以將這個字串過濾去。反之，字串的前後接字複雜度愈大，則愈有可能表示一個完整的語詞。另外，論文資料裡具有高頻字串的首尾經常是介詞（prepositions）、連詞（conjunctions）或定詞（determiners）等停用詞，因此我們過濾掉首尾爲停用詞的字串，使得過濾後的語詞具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、平均出現頻次夠高則仍爲重要的語詞。

主題代表性則是指此一語詞能否代表論文的主題並與其他主題區別，本研究

以字串在所有論文資料的出現總次數、平均出現頻次和標準差來表示語詞的主題代表性。假若一個語詞具有很高的出現總次數，表示這個語詞在研究社群的傳播現象中常被研究人員使用，這個語詞很有可能代表了領域重要概念，以圖書資訊學而言，「圖書館」這個語詞代表了相當重要的概念，因此如果我們蒐集圖書資訊學相關期刊所發表的論文資料，進行統計分析，必然可以發現這個語詞在這些論文資料的出現總次數很高（林頌堅，2002）。因此，出現總次數高的字串極有可能是這個領域的關鍵語詞；反之，出現次數較低的字串極有可能不是一個語詞，而即便這個字串是一個語詞，它的重要性也不高，可以被排除。語詞的平均出現頻次和標準差則可表示這個語詞在出現的論文的重要性，平均出現頻次的估算方式是語詞的出現總次數除以出現的論文數。平均出現頻次愈大的語詞，表示這個語詞在出現的論文中多半具有較多的出現次數，代表這些論文的重要主題。在上述的例子中，「圖書館」在圖書資訊學領域相關論文中的平均出現頻次也相當高（林頌堅，2002）。語詞的出現頻次標準差則可以利用它的平均出現頻次、出現論文數與出現在各論文的次數進行估算，語詞的出現頻次標準差較大則表示此語詞的出現不均勻，集中在某些論文上。對這些論文來說，這個語詞便可能代表了其中的重要主題。所以，若是一個語詞的平均出現頻次和標準差都不高，它對所有出現這個語詞的論文可能也不具有重要性，可以被排除。

總結上述各點，在語詞抽取技術方面，本論文以PAT tree儲存所有出現在論文資料中的字串與其統計資訊，並每次從這個資料結構擷取一個字串做爲候選語詞，利用字串的統計資訊估算其單元完整性與主題相關性，濾去可能性較低的候選語詞。最後，將保留下的字串作爲關鍵語詞，代表這個研究領域的重要主題。

（三）定義語詞特徵與距離

抽取領域的關鍵語詞之後，在產生主題關係圖之前，我們首先定義語詞之間的距離，用來表示它們所代表主題之間的相關程度，使得語詞之間距離愈小，表示它們代表的主題之間愈相關；反之，當兩個語詞間有最大的距離時，表示它們的代表主題之間不相關。在本論文中，我們將每一個語詞表示爲是由一組特徵值所構成的特徵向量，以特徵向量間的歐幾里得距離（Euclidian distance）來估算語詞間的距離，並且利用語詞與其他語詞的共現關係做爲語詞特徵向量上的特徵值。兩個語詞之間的共現關係是指這兩個語詞出現於相同論文資料的情形，如果兩個語詞經常一起出現，亦即它們具有很強的共現關係，通常是主題相關的語詞。如果以式(2)來表示我們所抽取出來的第 i 個語詞的特徵向量fi，在特徵向量上的第一個特徵值 $o_{i,1}$ 便是這個語詞與抽取出來的第1個語詞之間的共現關係，第 k 個特徵值 $o_{i,k}$ 則是這個語詞與第 k 個語詞之間的共現關係等，依次類推。而且在特徵向量上愈大的特徵值，表示這個語詞與相對應的語詞的共現關係愈強，這兩個語詞之間愈相關。

$$f_i = \left[o_{i,1},\ldots,o_{i,k},\ldots,o_{i,N}\right]^T \qquad (2)$$

利用歐幾里德距離來估算兩個語詞特徵向量的距離來表示相對應主題之間的相關程度時，如果兩個特徵向量的距離愈相近，表示比較的兩個語詞在論文資料中和其它語詞的共現關係愈相似。比方說，當抽取出來的第 i 個和第 j 個語詞與第 k 個語詞都存在共現關係時，這兩個語詞的特徵向量 f_i 和 f_j，在第 k 個特徵值 $o_{i,k}$和$o_{j,k}$上都有較大的值；反之，若對第 k 個語詞都缺乏共現關係時，特徵值 $o_{i,k}$ 和$o_{j,k}$ 都將較小。因此，當計算兩個語詞的距離時，如果得到較小的值，這兩個語詞必然會有許多相同的相關語詞，因此可以推論這兩個語詞在代表的主題上有相關性。所以可以利用語詞特徵向量之間的歐幾里德距離來表示相對應主題之間的相關程度。然而實際上，在估算語詞間的共現程度時，經常發生相關語詞可能不在論文資料中一起出現的現象，使得這些語詞所代表的主題之間的相關程度被錯估。爲了減輕這個現象的影響，在本研究中將利用LSA（latent semantic analysis）技術來估算語詞間的共現程度。LSA技術是利用奇異值分解（singular value decomposition）的方式，將各個語詞的特徵向量轉換成維度較低的向量，來估計各個語詞之間的隱含語意結構（Deerwester et. al., 1990），使得即便不存在共現關係的相關語詞，依然可以估算出它們相對應主題的相關程度。在本論文中利用LSA技術估算語詞特徵向量間距離的詳細做法可以參考（林頌堅，2003b）。

（四）產生主題關係圖

在產生每一個語詞的特徵向量作爲訓練資料之後，便開始進行資訊視覺化，將抽取出來的語詞依據前述定義的語詞特徵向量與距離計算方式，映射到一個二維的圖形上。使得產生的圖形結果能夠反映計算語言學的知識組織現況，也就是說重要的主題都可以表現在圖形上，而且語詞的映射點之間的距離關係表示對應主題間的相關性，使得每一個語詞映射點鄰近範圍的映射點爲具有相同或相關主題的語詞。

在資訊視覺化的相關研究中，有許多方法可以將文字資料映射到而爲圖形上，比方說，「奇異值分解」（Landauer, Laham & Derr, 2004）、「主成分分析」（principal component analysis, PCA）、「多維尺度法」MDS（multidimensional scaling）（Huang, Ward, & Rundensteiner, 2003）以及「自組織映射圖」（self-organizing maps, SOM）（Lin, 1992; Flexer, 2001）等。在考慮運算資源的需求與新增資料可以相容於先前產生的結果等實作方面的問題，在本論文的研究中，我們選擇利用自組織映射圖技術來產生研究領域的主題關係圖（林頌堅，2004）。

SOM是一種非監督式訓練（unsupervised training）的類神經網路（artificial neural networks）（Kohonen, 1989），常用於資料的叢集分析（cluster analysis）和資訊視覺化等應用中（Flexer, 2001）。過去SOM在處理文字資料的應用也相當多，

一類以語詞做爲處理對象，目的在分析語詞之間的叢集關係（Ritter & Kohonen, 1989; Ma, et. al., 2002）；另一類則以文件爲對象，產生文件的映射圖，做爲資訊檢索應用的人機介面（Lin, Soergel & Marchionini, 1991; Kohonen, et. al., 1996; Merkl, 1997; Wermter & Hung, 2002）。SOM的運作概念是利用一組排列成方陣的節點（nodes），每一個輸入的資料項以它的特徵向量與圖形上所有節點的特徵向量進行比對，根據比對的結果，選擇與資料項特徵向量最相似的節點與在這個節點鄰近範圍內的節點進行調適，這便是所謂的SOM自組織訓練過程（Kohonen, 1989）。經過多次訓練後，SOM便可以表現資料項間的關係，使得特徵向量接近的資料項映射到同一節點或相接近的節點上。借助這樣的特性，SOM可以對具有高維度特徵向量的資料項進行叢集分析，而且可以將這些資料項映射到圖形上，使得原本複雜而難以認知的資料分布，透過SOM產生的二維圖形得以進行分析。本研究便是利用SOM技術將論文資料中抽取出的語詞映射到二維圖形上，使得具有相關主題的語詞在圖形上形成主題叢集來表現出研究領域的主題關係。

本研究採用一般的SOM訓練方法（Kohonen, 1989）。每次從訓練資料中隨機挑選一個語詞，以它的特徵向量進行訓練，直到到達預設的訓練次數或SOM不再改變爲止。每次的訓練過程包含選擇（selection）與調適（adaptation）兩個組織化步驟。首先以訓練的特徵向量與SOM各節點的特徵向量進行比對，計算它們與訓練特徵向量之間的歐幾里得距離，選擇一個最小距離的節點，在SOM的訓練法中此一節點稱爲「獲勝節點」（the winner node）。接著對獲勝節點以及在它周圍的節點進行調適，使它們的特徵向量與訓練特徵向量距離縮小，如式（3）所示。

$$f_c(\tau+1) \overset{def}{=} f_c(\tau) + h(\tau, d(n_w, n_c))[f_i - f_c(\tau)] \qquad (3)$$

式（3）中，$f_c(\tau)$是表示第τ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τ與節點和獲勝節點n_w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。節點特徵向量的調適幅度與它們與「獲勝節點」的距離大小有關，「獲勝節點」本身的調適幅度最大，而愈遠離「獲勝節點」的節點調適幅度愈小。另外，隨著訓練次數增加，調適的節點範圍以及調適幅度逐漸縮小。在本研究中，$h(.)$定義如式（4）。

$$h(\tau, d(n_w, n_c)) = e^{\frac{\tau \times [d(n_w, n_c)^2 + 1]}{\alpha}} \qquad (4)$$

在式（4）中，α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。在式（4）中，可以發現在每次訓練中，愈接近「獲勝節點」的節點（$d(n_w, n_c)$值愈小），獲得的調整幅度愈大，反之則愈小。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會收斂到組織化的狀態。

完成SOM訓練後，可以依據各個節點的特徵向量標示這個節點的相關主題，做爲這個領域的主題關係圖。如前所述，語詞特徵向量中的每一個特徵值代表了這個特徵向量所對應的語詞與另一個語詞之間的共現關係，特徵值愈大表示兩者的主題愈相關。因此，在訓練出來的SOM上，節點特徵向量中最大特徵值的相對應語詞是這個節點最相關的語詞，可作爲這個節點的主題之標示 。所以進行節點標示時，我們以特徵向量的特徵值最大者所對應的語詞，做爲這個節點的標示。使用者便可以透過映射在圖形上的關鍵語詞了解領域中重要的主題並且利用這些語詞在圖形上的距離了解主題之間的相關性。

(五)將論文資料映射到主題關係圖

在產生表示重要的研究主題以及它們之間關係的主題關係圖之後，本研究進一步探討研究領域在主題上的成果。我們的想法是以論文在各主題上的分布情形做爲研究領域在主題上的成果，研究主題上面的相關論文愈多，愈可能表示這個主題是研究人員所重視的研究。因此，接下來，本研究將論文資料庫中的各筆論文資料都映射到先前產生的主題關係圖，並且以關鍵語詞在論文資料的分布情形做爲該論文的特徵向量，與圖形上每一個節點的特徵向量進行比對，選取節點中歐幾里得距離距離最小的節點，做爲論文資料的映射點。以論文 p 爲例，論文資料特徵向量的定義，如式(5)所示。

$$f_p = \frac{\sum_{t \in p} F_{p,t} f_t}{\sum_{t \in p} F_{p,t}} \qquad (5)$$

在式(5)中，t 代表在論文 p 中出現的任一語詞，$F_{p,t}$是這個語詞在論文 p 的出現次數，所以式(5)的分母部分是論文 p 中所有出現語詞的次數總和。另外，f_t代表語詞 t 的特徵向量。這個定義中，論文的特徵向量與該論文中所有出現語詞的特徵向量有關，而且在論文資料中出現愈多次的語詞則愈能夠代表這筆論文資料，因此，以每一語詞出現的次數作爲加權，再以所有語詞的出現次數總和進行正規化。最後依據歐幾里德距離，選取距離最小的節點做爲論文資料的映射點，如此一來論文資料會映射到相關主題的節點，而且主題相關的論文資料在圖形上的分佈會大致相近，研究人員和科技管理人員便可以利用論文資料在圖形上的分布了解各主題的研究成果。

三、計算語言學的發展趨勢

提出本論文使用領域主題分析方法應用在台灣計算語言學研究的成果之前，首先我們先對計算語言學的研究範疇做一介紹。計算語言學研究主要著眼於利用

電腦強大的運算與儲存能力，分析、探索及模擬口語與書面語之表達以及理解等人類的語言傳播過程，整合語言學家和電腦科學家雙方面的知識與技術。語言學家的工作在提出可以計算化的語言理論，並利用根據這些語言理論製作出來的電腦系統，對實際發生的語言表達與理解進行驗證（王士元，1988；Huang, 2000）；而電腦科學家的任務除了根據語言理論設計與製作電腦系統驗證這些理論之外，還包括了將語言理論應用到人類與電腦互動的介面上，希望可以讓電腦透過自然語言直接與人們進行互動，並且可以處理大量以自然語言形式產生的資訊（Dale, 2000）。因此，計算語言學與傳統語言學研究中很大的不同是計算語言學家可以對日常生活中發生的各種形式、語體的大量語料（corpus）加以蒐集，分析蘊含在其中的語言現象，進行語音學（phonetics）、詞典編纂學（lexicography）、構詞學（morphology）、語法學（syntax）、語意學（semantics）、語用學（pragmatics）等語言理論的研究。除此之外，計算語言學研究還注重將語言理論應用到實際的技術發展上，使電腦系統可以輸出、輸入與處理語言形式的資料。這些技術系統應用於許多人機介面以及文字資訊處理中，比方說語音合成（speech synthesis）、語音辨認（speech recognition）、機器翻譯、資訊檢索（information retrieval）、資訊抽取（information extraction）和文件分類（text categorization）等（Lenders, 2001; Kay, 2003），提供人們更方便而有效率的人機介面以及處理巨量的文字資料。對於台灣的計算語言學研究而言，由於中文的電腦輸入較拼音文字不便，更需發展合適的電腦輸入系統。此外，中文在語音、書寫方面具有相當多的特色，比方說中文語音的一字一音節與書寫上中文的詞與詞間沒有間距等，都是處理中文的語言傳播上需要特別注意與處理的地方。台灣的計算語言學研究在這些地方也都有相當不錯的研究成果，每年有相當數量的論文在國際的期刊或研討會上發表以外，此外，研究人員還組成了一個研究社群，計算語言學學會（The Association for Computational Linguistics and Chinese Language Processing, ACLCLP）。這個社群除了發行通訊和期刊，從1989年起每年舉辦一次學術研討會ROCLING（計算語言學研討會，2004年起變更爲自然語言與語音處理研討會），但2002年因爲ACLCLP承辦國際的計算語言學學術研討會而沒有舉辦ROCLING研討會。而除了做爲研究人員間非正式傳播的管道以外，ROCLING所出版的論文集中也累積了許多國內計算語言學的研究成果，可以用於探勘這個研究領域的主題。

因此，我們便以計算語言學研討會ROCLING的會議論文資料來探勘這個研究領域的主題。本研究蒐集了從1988到2001年共十四屆的ROCLING研討會會議論文，共325筆論文資料。在建立ROCLING論文資料庫後，我們從論文文字資訊中抽取關鍵語詞。本研究將語詞的前後字複雜度之閾值設爲0.5，出現總次數設爲20次以上，平均出現頻次與標準差之和則需達到2.5以上，結果在論文資料中共抽取出229個關鍵語詞。

(一)計算語言學的主題關係圖

在爲抽取出來的229個關鍵語詞建立特徵向量之後，我們以這些特徵向量產生主題關係圖。在本研究中，我們將SOM的節點數目設爲20*20，共400個節點，α值設定爲150，進行50次的SOM訓練，結果所顯示的主題分布情形，如圖2所示。

在圖2中，圖形左下方的節點標示的語詞大多與語言理論主題相關，包括了「syntax」、「lexical」、「verb」、「語法」等等。計算語言學中和語言理論相關的基礎技術爲剖析技術，而剖析語言現象時，需要使用由各種語言知識編纂成的文法規則。因此，語言理論相關的節點也和這些主題相關節點的位置相當接近，比方在圖形左方的「grammars」、「grammatical」和「parsing」等。利用語言理論與剖析技術所發展的應用爲「machine translation」，節點位置在這兩群節點之間。再者，蒐集大量語言實例的資料庫稱爲「corpus」與「corpora」，利用這種資料庫進行語言知識分析的研究稱爲語料庫語言學，而應用來處理這些資料庫的技術如「part of speech tagging」和「word sense disambiguation」等，與這些主題相關的節點都分佈在圖形的上方。

語音處理與資訊檢索是計算語言學的主要應用。語音處理相關的節點分布在圖形的右方，包括「speech recognition」和「speaker」等。在語音辨認的技術中，常用「language model」、「語言模型」等語言資訊，作爲辨識時限制搜尋空間的資訊，以迅速找到正確字詞。因此，語言模型的建立與應用也是語音處理研究相當重要的主題。從圖形上也可以看出兩者相關節點的位置相當接近。最後，「information retrieval」、「summarization」、「text categorization」、「檢索」、「分類」等相關節點上分布在圖形的右上方，而且它們的處理對象「chinese text」和「document」也在圖形的附近。

從圖2中，我們也可以看到許多主題相關的節點交錯地排列在圖形上。比方說，「parsing」、「grammars」和「theory」以及「information retrieval」、「text categorization」和「summarization」等都是相關的主題。另外，class-based language model是一種以詞群爲基礎的語言模型，目的在減少語言模型所需的大量訓練語料，並獲得較準確的辨認結果。而「統計」和「系統」的相關節點在圖形上交錯，則表示在系統實做方面通常利用統計方法，以達成強健和容易製作等目標。由以上的討論證明本研究所產生的主題關係圖可以代表計算語言學領域的研究主題分布情形。

(二)ROCLING論文資料在主題關係圖上的分布

我們將第一屆(1988)到第十四屆(2001)的235筆ROCLING論文資料映射到主題關係圖，從論文的分布情形來了解台灣計算語言學研究人員在各主題上的成果，圖3是映射的結果。從圖3中，我們可以看到論文分布較多且較爲密集的主題包括：「語言」、「漢語」、「語法」、「lexical」、「theory」、「parsing」、「machine

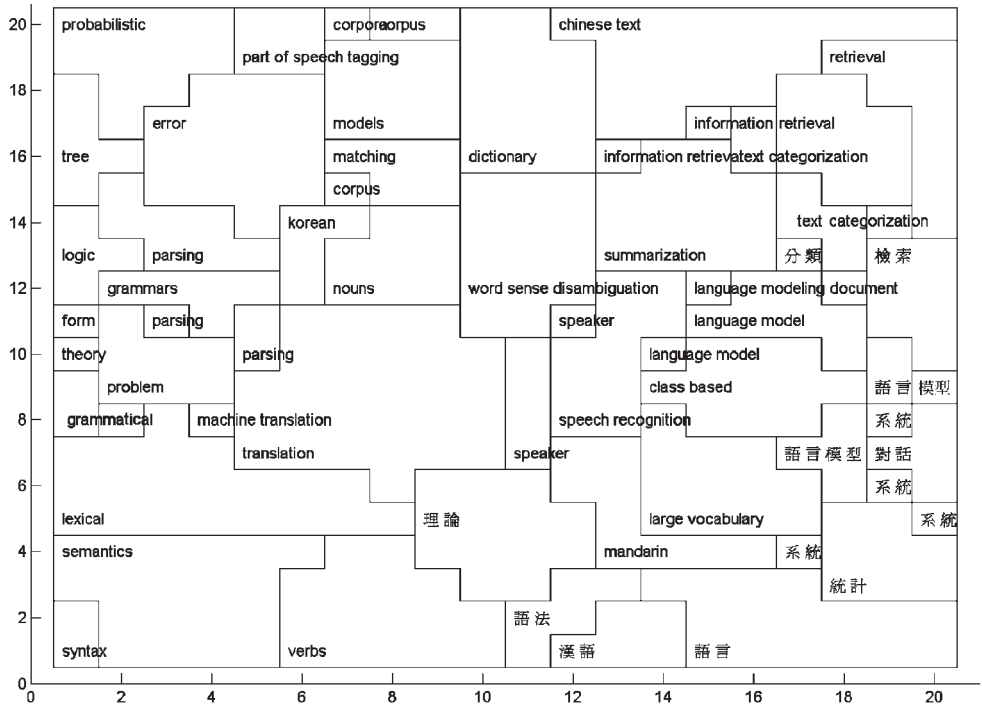

圖2　台灣計算語言學研究領域的主題關係圖

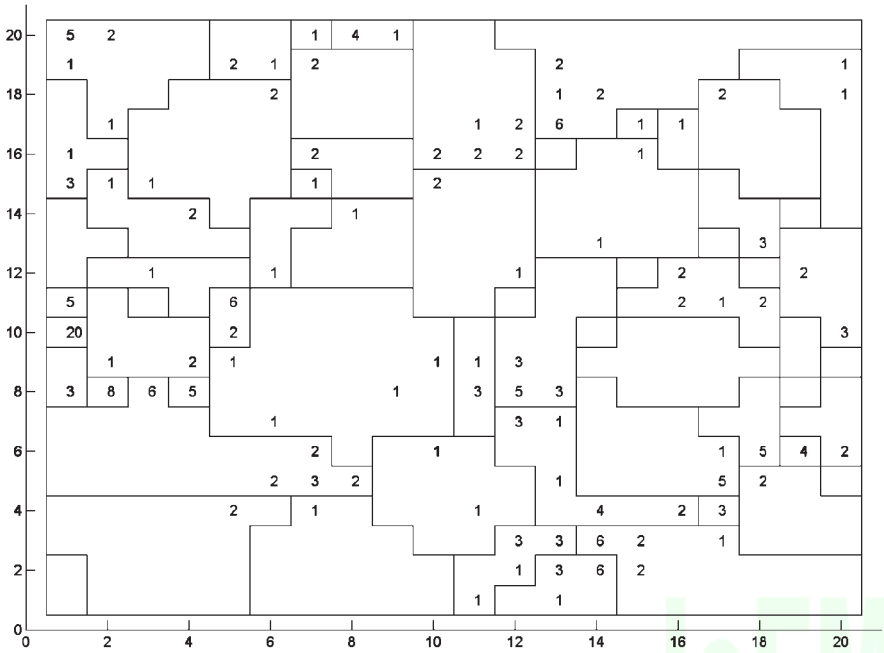

圖3　ROCLING（1988-2001）論文在SOM上分布情形

translation」、「corpus」、「dictionary」、「large vocabulary」、「speech recognition」、「language model」、「語言模型」、「Chinese text」和「檢索」等。所以，在過去台灣計算語言學研究包含了漢語語言理論的建立，包括以語料庫進行語言現象的分析與驗證及詞典和語法理論的開發，並利用研究的成果應用在剖析技術及機器翻譯的應用上；其次，則是大詞彙的語音辨認技術的開發與應用，包含語言模型等相關技術；最後，是中文文字資訊的檢索。

(三)計算語言學研究主題的發展

從ROCLING論文在主題關係圖的分布情形，我們瞭解了計算語言學在各主題上的研究成果，以下進一步探討各時期各主題上的論文分布情形，分析計算語言學各主題的研究發展趨勢。依照ROCLING論文的發表年代分爲三個時期，圖4到圖6則分別是1988到1992年、1993到1997年，及1998到2001年等三個不同時期的論文分布情形。

早期(1988~1992年)的ROCLING會議論文的映射點大都分布在主題關係圖左方及下方(圖4)，對照圖2上這些節點的標示，可以知道這些論文與「漢語」、「語法」、「parsing」、「grammars」、「translation」和「machine translation」等等研究主題相關，且有相當多論文投射在標示爲「theory」的節點上。上述結果很明顯地可說明，這時期的計算語言學研究因應於機器翻譯的應用及剖析技術的發展，而產出各種可用於剖析技術的文法規則，並著重於各種語言知識的計算理論化。另外，較少數的論文投射在圖形最右方，標示爲「對話」和「系統」的節點上，代表這時期也有少部分研究主題，嘗試進行有關對話系統方面的研究。

ROCLING的中期論文(1993~1997年)幾乎涵蓋了前述討論過的各種研究主題(圖5)，除了語言知識的計算理論化，編製文法規則、發展剖析與機器翻譯技術等等主題仍是計算語言學的研究範疇外，這時期較令人矚目的現象有二：1.語料庫和詞典是儲存大量知識的組織結構。由網際網路出現所帶來的大量語料，提供計算語言學領域可探索實徵而大量的語言現象，同時爲處理這些語料，研究人員發展了以機率爲主(probabilistic)的技術，如詞類標示和詞義區辨(word sense disambiguation)等，確認語料內的語法和語意訊息，發展與應用這些語言知識組織結構。2.語音處理技術已受到計算語言學領域的重視，特別是語言模型的建立與應用等技術，非常適合應用這個領域已發展出的統計導向自然語言處理技術。另外值得一提的是，由於大量文件的出現，包括資訊檢索、文件分類與摘要等文本處理的相關技術在此時期已開始萌芽。

在最後時期的四年(1998~2001年)中，ROCLING論文主要著重於語料庫與詞典等語言知識組織、語音處理技術與應用系統和文本處理技術等三個方向的研究發展(圖6)，特別在大量網路資源的需求和影響下，促成資訊檢索、文件分類和摘要等技術的快速發展，產生相當多實用的應用系統。早期的語言知識的計算理論

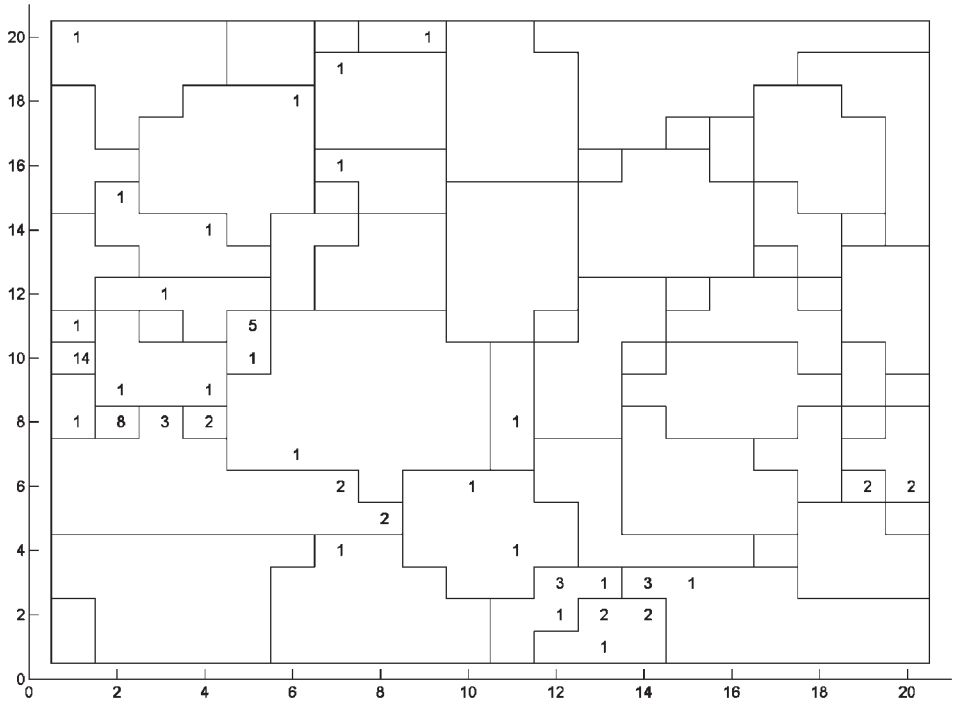

圖4　ROCLING（1988-1992）論文在主題關係圖上分布情形

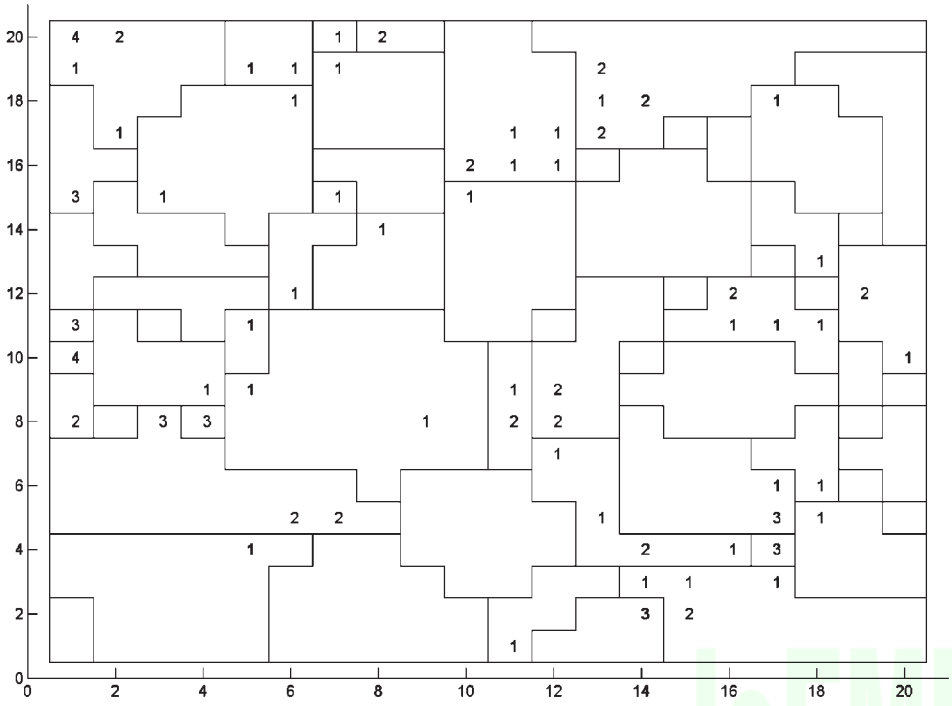

圖5　ROCLING（1993-1997）論文在主題關係圖上分布情形

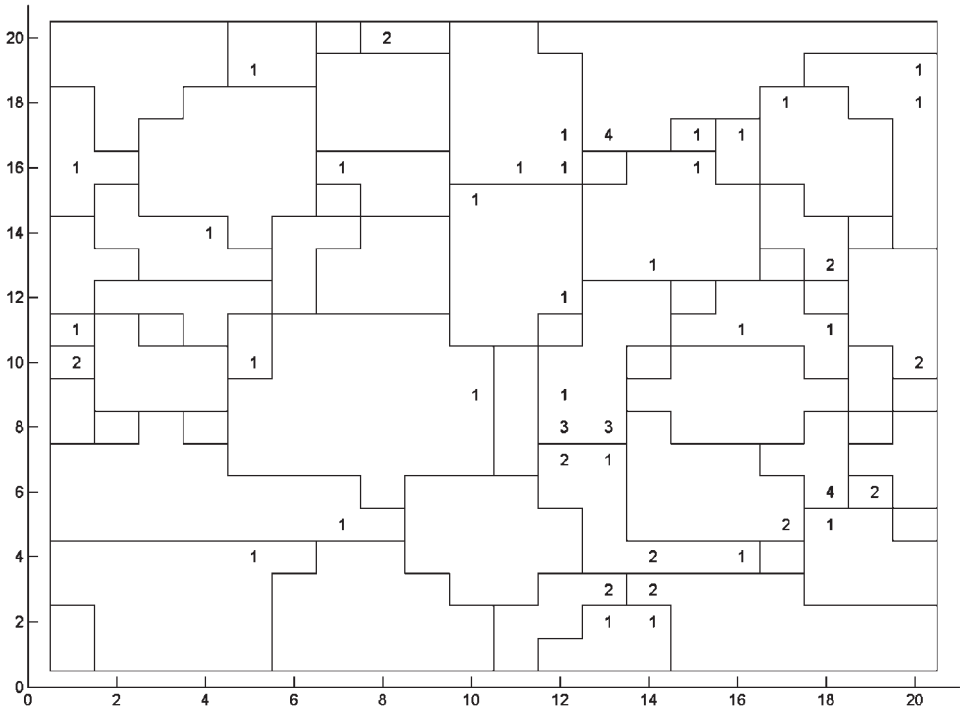

圖6 ROCLING(1998-2001)論文在主題關係圖上分布情形

化以及相關技術的研發在這個時期則較爲式微。

綜合上述分析與討論，在較早的時期因爲機器翻譯的需求，需要對各種語言的語言現象進行系統化的分析，並將分析的結果編製成同時適合人類了解與提供電腦應用的文法規則。而這樣的需求不僅產生了各種語言理論與自然語言處理技術，而且促使語言學和電腦科學的研究人員共同進行科際整合研究，開始建立了計算語言學領域。在中期，由於電腦與網路的普及應用，產生了較多的電子文件，提供自然語言學研究所需的語料。一方面，經過處理的語料進一步提供了相當多實徵的語言現象，作爲計算語言學理論與技術的驗證與修正；另一方面，了解應用先前發展的文法規則與剖析技術的限制，促使統計式語言模型和機率導向自然語言處理技術的發展，這些都更加符合電腦科學的知識與研究方法。因此，語料庫改變了計算語言學領域的知識內容與獲取知識的方法，而且更促進語言學和電腦科學的整合研究。此外，語音處理配合自然語言處理技術提供更方便而有效率的人機介面發展的可行性，使得許多語音信號處理的學者進入這個研究領域。而借助於他們對於在信號處理的統計知識與技術，也爲這個領域引入相當多的統計方法。最後時期計算語言學的研究更加實務導向，雖然語言理論的研究減少，而語音處理和語料庫的研究已是主要的研究方向。除此之外，網際網路的盛行，提供使用者透過網路取得各類資訊的需求，也使得文本處理技術成爲計算語言學的重要研究。

四、結　論

在本論文所提出來的研究主題探勘方法中，我們應用了各種的技巧與技術，從論文資料分析研究領域中重要的研究主題，並且進一步以資訊視覺化的方式，將主題間的關係以及領域研究的發展情形，呈現在圖形上，提供科技管理人員在擬定科技發展計畫的參考，並提高研究者在學術競爭中取得優勢的機會。這些技巧與技術包括：1.利用論文的題名、摘要和參考文獻的題名中出現的關鍵語詞爲分析單位，提供豐富的統計訊息並且方便成果的解讀；2.在考慮單元完整性和主題相關性的情形下，以統計訊息和經驗法則抽取中文和英語的關鍵語詞；3.利用關鍵語詞間的共現關係，定義語詞的特徵向量和語詞之間的距離，並以LSA技術取得語詞之間的隱含語意關係，有效地估測主題間的相關程度；4.以資訊視覺化中容易實作的自組織映射圖技術，產生可以代表研究領域中重要研究主題與其關係的圖形；5.利用論文資料中的語詞分布情形，將論文資料映射到主題關係圖中，以論文的數目表示研究社群在領域中各種主題的研究成果。

在應用以上技術後，我們將可以得到三種主要的輸出結果。1.主題關係圖表現出研究領域的重要主題以及各主題之間的關係，研究人員在進行研究的初期，可以依據這個圖形，了解研究所需的相關背景知識；在研究的中晚期，也可以從圖形上了解本身研究的定位，做爲更進一步發展的參考。2.論文在主題關係圖上的分布情形，表現研究領域在各主題上的累積成果。3.不同年代的主題發展情形，依據論文發表的年代，將論文映射到主題分布圖上，從各年代論文在各主題的分布，了解當時研究領域所注重的研究主題，可以做爲分析研究主題的增長與變化趨勢的參考資訊，提供科技管理人員了解與考核研究領域的成果，並做爲研究人員選擇具有發展潛力研究問題的參考。

另外，本論文將研究主題探勘方法實地應用到台灣的計算語言學研究中，從產生的結果，可以確認這個領域的重要研究主題以及這些主題在不同時期的發展情形，本研究並依據產生的結果，提出合理的解釋。因此，目前的結果初步地證明了這項工具應用於領域分析的可行性。

在過去，圖書資訊學的研究人員認爲「資訊檢索研究」和包括學術傳播、書目計量學、引文分析等主題的「領域分析」是這個領域的主要研究方向。然而他們也認爲這兩個研究之間缺乏整合，無法利用彼此的研究結果，深化這個領域的知識結構，此一問題逐漸成爲圖書資訊學在研究上的重大缺陷（Persson, 1994; White & McCain, 1998）。本研究利用資訊檢索與資訊視覺化技術進行學術領域的研究發展分析，嘗試提出一個整合資訊檢索語領域分析的研究方向。在未來的研究中將嘗試以更多領域做爲分析的對象，以了解此一技術的限制，並且可以利用學術論文以外的科技文獻進行產業資訊的分析，如技術手冊、專利宣告、專利說明書、相關網頁資料等，探討此一技術在實用上的價值。

參考文獻

Börner, Katy, Chen, Chaomei & Boyack, Kevin W. (2003). "Visualizing knowledge domains," *Annual Review of Information Science and Technology, 37*, 179-255.

Card, Stuart K., Mackinlay, Jock D. & Shneiderman, Ben (1999). "Information visualization," *Readings in Information Visualization—Using Vision to Think*, 1-34. Morgan Kaufmann.

Chen, Keh-jiann & Liu, Shing-Huan (1992). "Word identification for mandarin Chinese sentences," In *Proceedings of the 14th International Conference on Computational Linguistics* (COLING 92).

Chien, Lee-Feng (1997). "PAT-tree-based Keyword Extraction for Chinese Information Retrieval," *Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-58.

Crane, Diana (1972). *Invisible College—Diffusion of knowledge in scientific community*. 中譯爲劉珺珺、顧昕、王德祿譯(1988)。無形學院知識在科學共同体的擴散。北京市：華夏出版社。

Dale, Robert (2000). *Handbook of Natural Language Processing*. Marcel Dekker, Inc.

Deerwester, Scott, et. al. (1990). "Indexing by latent semantic analysis," *Journal of the American Society for Information Science, 41*(6), 391-407.

Flexer, A. (2001). "On the use of self-organizing maps for clustering and visualization," *Intelligent Data Analysis, 5*(5), 373-184.

Garfield, Eugene (1979). *Citation Indexing—Its Theory and Application in Science, Technology, and Humanities*, reprinted in 1983. Philadelphia, PA: ISI Press.

Hausser, Roland (2001). *Foundations of Computational Linguistics: Human-computer Communication in Natural Language*. New York: Springer-Verlag.

Huang, Shiping, Ward, Matthew O. & Rundensteiner, Elka A. (2003) *Exploration of Dimensionality Reduction for Text Visualization*. Technical Report TR-03-14, Worcester Polytechnic Institute, Computer Science Department.

Huang, Chu-Reng. (2000). "From quantitative to qualitative studies: Developments in Chinese computational and corpus linguistics," 漢學研究，第18卷特刊，473-509.

Kageura, Kyo & Umino, Bin (1996). "Methods of automatic term recognition—A review," *Terminology, 3*(2), 259-289.

Kay, Martin (2003). "Introduction," *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov, xvii-xx.

Kohonen, Teuvo (1989). *Self-Organization and Associative Memory* (3rd ed.). New York: Springer-Verlag.

Kohonen, Teuvo, et. al. (1996). Very Large Two-Level SOM for the Browsing of Newsgroups. *Proceedings of the 1996 International Conference on Artificial Neural Networks*, 269-274.

Landauer, Thomas K., Laham, Darrell & Derr, Marcia (2004). "From paragraph to graph: latent semantic analysis for information visualization," *Proceedings of the National Academy of Science of the USA*, 101, 5214-5219.

Lenders, Winfried (2001). "Past and future goals of computational linguistics," *Proceedings of ROCLING XIV*(第十四屆計算語言學研討會論文集), 213-236.

Lin, Xia (1992). "Visualization for the document space," *Proceedings of IEEE Visualization*

1992, 274-281.

Lin, Xia, Soergel, Dagobert & Marchionini, Gary (1991). "A self-organizing semantic map for information retrieval," *Proceedings of the 14th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 262-269.

Ma, Qing et. al. "Self-organizing Chinese and Japanese semantic maps," *Proceedings of the 19th International Conference on Computational Linguistics* (COLING 2002), 605-611.

Meadows, A. J. (1998). *Communicating Research*. San Diego: Academic Press.

Merkl, Dieter (1997). "Exploration of text collections with hierarchical feature maps," *Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 186-195.

Persson, Olle (1994). "The intellectual base and research fronts of JASIS 1986-1990," *Journal of the American Society for Information Science, 45*(1), 31-38.

Price, Derek. J. de Solla (1963). *Little Science, Big Science—and Beyond*, reprinted in 1986. Columbia University Press.

Ritter, H. & Kohonen, T. (1989). "Self-organizing semantic maps," *Biological Cybernetics, 61*, 241-254.

Rogers, Everett M. (1983). *Diffusion of Innovations* (3rd ed.). New York: Free Press.

Salton, Gerard & McGill, Michael J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sproat, Richard, et. al. (1996). "A stochastic finite-State word-segmentation algorithm for Chinese," *Computational Linguistics, 22*(3), 377-404.

Tabah, Albert N. (1996). *Information Epidemics and the Growth of Physics*. Ph. D Dissertation of McGill University, Montreal, Canada.

Wermter, Stefan & Hung, Chihli (2002). "Selforganizing classification on the Reuters News Corpus," *Proceedings of the 19th International Conference on Computational Linguistics* (COLING 2002), 1086-1092.

White, Howard D. & McCain, Katherine W. (1998). "Visualizing a discipline: An author co-citation analysis of information science, 1972-1995," *Journal of the American Society for Information Science, 49*(4), 327-355.

Wu, Zimin & Tseng, Gwyneth (1993). "Chinese text segmentation for text retrieval: Achievements and problems," *Journal of the American Society for Information Science, 44*(9), 532-542.

行政院國家科學委員會科學技術資料中心(2003)。台灣學術研發能量之總體表現。編者出版。

林頌堅(2002)。「基於高頻詞語的圖書資訊學研究領域分析之初步探討」。中國圖書館學會會報，69，138-154。

林頌堅(2003a)。「基於自然語言處理技術的研究主題抽取與分析」。*Proceedings of ROCLING XV*(第十五屆計算語言學研討會論文集)，頁231-256。

林頌堅(2003b)。「基於詞語抽取的圖書與資訊學刊研究主題分析」。圖書與資訊學刊，47，頁15-35。

林頌堅(2004)。「以自組織映射圖進行計算語言學領域視覺化之研究」。*Proceedings of ROCLING XVI*(第十六屆自然語言與語音處理研討會論文集)，頁69-77。

王士元(1988)。「電腦在語言學裡的運用」。*Proceedings of ROCLING I*(第一屆自然語言與語音處理研討會論文集)，頁257-287。

Journal of Educational Media & Library Sciences 42 : 4 (June 2005) : 523-544

An Automatic Method for Topic Exploration in a Subject Domain and Its Application on Computational Linguistics

Sung-Chien Lin
Assistant Professor
Department of Information and Communications, Shih-Hsin University
Taipei, R.O.C.
E-mail: scl@cc.shu.edu.tw

Abstract

Because the size of modern scientific research is larger than before and the task of research becomes even more complex, researchers and managers urgently need an effective method to explore important topics in research domains. In the past, we had proposed a series of technologies based on text processing and text mining to deal with such a problem. Using text information in papers of the examined domain as input, a technology for term extraction was proposed to select key terms in the text information to represent important topics in the domain. Another proposed technology for information visualization was used to present the terms and their relationships in two-dimensional graphs with a technology of information visualization. Users can easily browse the topics of the domain as well as their development through the generated graphs for decision making of research and management. In addition, the technologies include several techniques of estimating term co-occurrences, calculating degrees of relevance between topics, and mapping paper information to the topic graph. In this paper, an automatic method for topic exploration was proposed with the integration of the developed technologies and it was applied to the studies of computational linguistics in Taiwan to depict foci of research and development in the domain. The result shows that for the development of technologies of machine translation, the earlier studies in the domain emphasized the computational theorization of several linguistic knowledge, but in its mid and later periods, there were more applications emerging, such as speech processing and information retrieval, and a lot of statistical approaches were adopted as the technologies for their robustness and easy implementation.

Keywords: *Topic exploration; Text processing; Text mining; Information visualization; Computational linguistics*