基於自然語言處理技術的研究主題抽取與分析

Extraction and Analysis of Research Topics Based on NLP Technologies

世新大學資訊傳播學系
Department of Information and Communications, Shih-Hsin University
林頌堅
Sung-Chen Lin
Email: scl@cc.shu.edu.tw

摘要

本論文針對研究主題分析的問題，提出一系列以自然語言處理為基礎的技術，從學術領域中發表的論文資料中抽取重要的關鍵詞語，並將這些詞語依據彼此間共現關係進行叢集，以叢集所得到的詞語集合表示領域中重要的研究主題。研究主題分析在學術領域的應用上，可以提供研究人員一個清楚的梗概；在資訊檢索的過程中，則可以幫助使用者釐清資訊需求。我們並將所提出的方法應用到 ROCLING 研討會的論文資料上，抽取計算語言學領域的重要研究主題。結果顯示這個方法可以應用於國內學術領域的特殊環境，同時抽取出中文和英文的關鍵詞語，所得到的詞語叢集結果也可以表示領域中重要的研究主題。這樣的結果初步的驗證了本論文所提出方法的可行性。從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，抽取出來的詞語叢集中有許多與機器翻譯、語音處理和資訊檢索相關，在語言的計算模式上，語法模式與剖析、斷詞和統計式語言

模型的建立則是國內計算語言學家所關心的主題。

一、緒論

資訊檢索研究著重的問題是人與資訊之間的介面，近來的研究趨勢注重於使用者所具有的背景知識、在檢索過程中對問題的認知[Wilson, 1999]及資料的嫻熟程度(material mastery)[Bishop, 1999][Covi, 1999]。為了對一個學術領域的資訊傳播現象進行全面的了解，所謂的「領域分析」(domain analysis)藉由對學術領域內重要的學術活動，諸如研究、論文發表、會議參與等等進行分析，探討研究人員所使用或產生的知識組織、結構、合作模式、語言和通訊形式、資訊系統以及相關標準等[Hjørland and Albrechtsen, 1995] 。而研究主題分析可以說是領域分析的一項要務，了解重要的研究主題可以掌握領域中的知識組織，幫助使用者釐清資訊需求(information need)，迅速取得所需的資訊。此外，藉由有系統的方法抽取研究主題並加以分析，可以展示學術領域研究一個完整的面貌，提供新進學者在初期進入領域時的參考，也可以作為學術研究領域發展的指引(road map)，提供已經深入的研究人員擴展學術研究的範疇。

本論文提出一個自動化的研究主題抽取方法，從學術領域中發表的論文集合中選出關鍵詞語，再依據詞語彼此間出現在相同論文中具有特定意義的共現(co-occurrences)現象，辨認每一篇論文中可能具有的研究主題，作為分析這個領域重要研究主題的依據。我們認為論文的豐富詞彙訊息蘊含了研究主題。在論文發

表的過程中，作者藉由論文題名、摘要以及本文中的詞語將研究的問題、方法與結果等主題傳達給讀者，甚至論文所引用的參考文獻題名也包含許多與主題相關的詞語訊息；而讀者在閱讀論文時，便可以依據這些詞語判斷與本身研究興趣上的相關性，同時將這些資訊建構與融入個人的知識結構中[Harter, 1992]。以本論文做一例子，在本論文的題名、摘要和本文中包含了許多『學術領域』、『研究主題』、『論文』等等詞語，目的是希望讀者在閱讀時，可以從這些詞語的共同出現與使用，了解我們所研究的主題是從學術論文中抽取重要的研究主題，而有興趣的讀者在閱讀後，便可在研究與發表上加以利用。進一步地，在一個學術領域中，可以發現某些受到重視的研究主題相關的詞語在許多論文中出現。以計算語言學領域來看，便可以發現諸如『語料庫』、『剖析』、『資訊檢索』等等的詞語在許多論文中出現，這些都是這個領域中的重要研究主題。而且與研究主題相關的一組詞語會重複出現在許多論文中。因此，如果對學術領域出版的論文進行分析，選取具有代表主題意義的詞語，統計這些詞語間的共現現象，利用這些資訊將經常一起出現的一組詞語叢聚成一個集合，所形成的詞語集合可以視為是某一特定的研究主題。在分析某一論文的主題時，便可以估算代表各研究主題的詞語叢聚與該論文的相關性，作為判斷該論文是否具有此一主題的資訊。因此，本論文嘗試利用自然語言處理技術來分析學術領域中發表的論文，確認論文中出現的詞語，抽取蘊含在其中詞語的共現訊息，再進行詞語叢聚(term clustering)，作為辨認主題分析的資訊。

我們並將所發展出來的技術應用於國內計算語言學領域的主題分析。選擇以計算語言學作為研究對象的主要原因是這個領域具有科際整合研究(interdisciplinary research)的特色，並且成功地將發展出的理論和技術應用到學術研究與實際的系統和產品研發[Lenders, 2001]。參與這個領域研究的研究人員主要來自於語言學和計算機科學兩個學科，對於計算機科學家來說，主要的研究工作在於建構一個實用的電腦系統來處理有關自然語言的問題，比方說機器翻譯、字型辨認、語音辨認、資訊檢索等等。語言學家的工作則在於計算性理論的規範與應用，用來解釋自然語言的認知現象模式及模擬驗證的能力[王士元, 1988]。計算機科學家的工作需要依賴語言學家所形成的語言理論來建立合理而有效率的電腦系統；而語言學家則是利用計算機科學家所發展的計算理論與系統探究自然語言的規律[Huang, 2000]。在這個領域中的重要研究除了將計算機方法應用於自然語言理論的探討之外，最受矚目的研究還包括利用語料庫(corpus)所發展出來的語言理論及利用這些理論設計與發展各種實務系統[Church and Mercer, 1993]。所以，從這個領域所進行的學術活動，可以觀察語言學和計算機科學兩種不同學科的學者在互相激盪下產生的成果，也可以觀察到從理論的研究到技術發展，再到實務的應用，對研究主題分析是一項具有挑戰且有意義的研究。除此之外，另一方面則是我們對於這個領域的熟悉，將有助於研究方法的發展，對於所得到的初步結果做出合理的詮釋，並作為下一階段改進的參考。

在使用 ROCLING 一到十四屆的學術研討會論文資料，共 235 篇，我們共抽

取出 343 個關鍵詞語。研究主題叢集後得到 34 個代表重要研究主題的詞語集合。結果顯示所發展的詞語抽取法可以同時抽取出中文和英文的關鍵詞語，所得到的詞語叢集結果也可以表示領域中重要的研究主題。初步驗證了本論文所提出方法的可行性。從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，抽取出來的詞語叢集中有許多與機器翻譯、語音處理和資訊檢索相關，在語言的計算模式上，語法模式與剖析、斷詞和統計式語言模型的建立則是國內計算語言學家所關心的主題。

本論文其餘的章節架構如下：在第二節中首先說明一些相關研究及本論文所提出研究主題分析方法的概念與合理性，並且依據這些概念設計出利用一系列自然語言處理技術進行研究領域分析的方法。接著在第三節和第四節中分述這個方法中的核心技術：詞語抽取和研究主題叢聚。第三節中，我們提出了在多語環境下的關鍵詞語抽取方法，可以從中英文論文資料中取得代表研究主題的關鍵詞語。第四節則提出一個詞語叢聚方法，利用詞語的共現關係，將詞語進行多重叢聚來代表可能的研究主題；本節中並且說明研究主題與論文之間相關程度的計算方式。第五節中報告將此分析方法應用到國內計算語言學研究的結果。最後，第六節則是結論。

二、本論文提出的研究主題分析方法

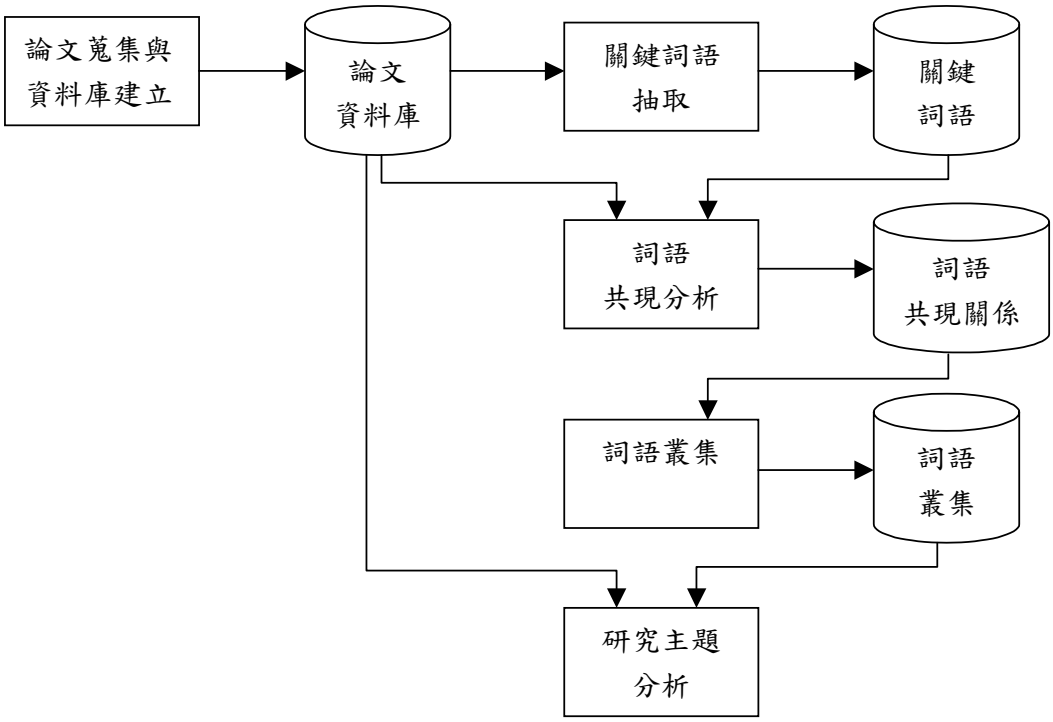
本論文希望發展一套研究主題抽取方法，可以從特定學術領域中出版的論文

中，抽取用來表達研究人員共識的重要研究主題，作為進一步分析或資訊檢索應用的資訊。在資訊檢索研究範疇中與這個問題相近的研究有主題偵測(topic detection)。主題偵測是希望從一序列來源各不相同的新聞中，偵測出與某些『事件』(events)相關的連續報導[Wayne, 2000]。目前研究人員認為『叢聚假說』(cluster hypothesis)可以適用於解決這個問題[Yang, Pierce and Carbonell, 1998][Hatzivassiloglou, Gravano and Maganti, 2000]，利用具有相關主題的文件具有相似的詞語分布情形，以文件叢聚(document clustering)技術，偵測新進文件是否與現有文件集合具有相似的詞語分布情形，將文件歸入相關事件的集合中；若文件與現有集合皆不相近，則視為是一個新事件，產生一個新的集合。因此，我們嘗試應用叢聚假說，探索領域中可能的研究主題。再者，主題偵測研究已應用專有名詞(proper nouns)等相關詞彙作為區隔不同新聞事件的重要訊息[Hatzivassiloglou, Gravano and Maganti, 2000]，本論文也將嘗試利用論文中與領域相關的詞語抽取出來作為分析的主要訊息。此外，主題偵測應用所謂的『新聞熱潮』(news bursts)現象，將時間訊息加入叢聚演算法，提昇偵測的結果[Yang, Pierce and Carbonell, 1998]。但是學術論文雖然有所謂『資訊流行』(information epidemic)[Tabah, 1996]的說法，也就是在某一項新的理論、研究方法或技術提出後，如果得到很大的成功，將可以吸引許多研究人員投入後續的研究中，造成一股相關研究發表的風潮，然而在實證研究中卻發現此一現象雖然存在但並不常見[Tabah, 1996]，所以在本論文並不考慮加入時間訊息。

在本論文中，我們利用相同研究主題的論文中具有相似詞彙訊息的概念，利用在論文中詞語的共現關係，找出詞語的叢聚情形來代表研究主題。以論文中出現的詞語取代整篇論文作為分析對象的主要原因是希望能獲得較可信賴的統計訊息。並非所有的研究領域都有足夠多的學術論文發表可供進行研究，較小的學術領域所出版的論文數量較為不足，以整篇論文進行分析，統計上不容易得到研究主題的分析結果。以關鍵詞語作為分析對象，可以獲得充足的統計訊息，克服文件數量較少的問題。某些論文具有多個研究主題也可藉由詞語的多重叢聚加以表示，進而探索研究主題之間的關係。此外，文件叢聚不易詮釋結果所代表的主題，詞語叢聚則容易直接由成員的語義進行解釋。

本論文方法的架構如圖一所示。首先對需要進行分析的學術領域蒐集相關論文資料，建立論文資料庫。資料庫中收錄的資料包括論文的題名、摘要和參考文獻的題名等作為詞語抽取與叢聚分析的資訊，論文作者和出版年等項目則用來作為後續研究主題的分析工作上。特別值得一提的是，國內的學術論文基本上是中、英語雙語並行，許多領域皆接受論文以中文或英文發表，然而並非所有的論文同時具有中、英文雙語的題名和摘要，無法單就某一種語言的文本進行分析。若只考慮以某一種語言發表的論文進行分析，而忽略另一種語言，有可能造成某些特殊的研究主題被遺漏的情形。若是分別處理各種語言的論文，缺乏分屬兩種語言的詞語在論文中的共現訊息，無法分析出這些詞語的相關性，在整合上有相當大的困難。因此需要考慮這個特殊的論文發表現象，提出可以同時分析兩種語言論

文的方法。本論文所提出的解決之道是加入論文中參考文獻的題名進行分析，通常參考文獻的題名與研究的理論、方法及技術等也有密切的關係，而且參考文獻的題名可能包含兩種語言，若能提出適當的多語詞語抽取方法，便可以統計分屬兩種語言的相關詞語的共現現象，整合兩種語言的詞語訊息，而得到較佳的研究主題分析結果。



圖一 本論文提出的研究主題抽取與分析方法

在建立好論文資料庫後，接著便利用多語的關鍵詞語抽取方法從論文資料中自動抽取領域中具有意義的詞語，統計詞語在論文中的共現關係，利用這些資訊將相關的詞語叢集成集合，用來代表某一個特定研究主題。在進行研究領域分析時，當詞語叢集與某一論文的相關性(relevance)足夠強時，可以假定該論文具有該詞語叢集所代表研究主題。下面的兩節中，將針對多語環境下的重要詞語抽取以

及詞語叢聚技術詳細說明。

三、多語環境下的關鍵詞語抽取

為了抽取可以代表學術領域研究主題的關鍵詞語，我們首先確認論文資料中重要的中英文詞組以及中文的多字詞，增強詞語的語彙訊息，再選擇具有代表研究主題意義的詞語，作為這一階段的結果。在學術論文中，常以詞組的形式表達重要的研究主題，比方在計算語言學領域的論文中，可以發現諸如英文的“language model”、“machine translation”或是中文的“語言模型”、“機器翻譯”等等。此外，中文的文本裡，詞與詞之間沒有明顯的界限，進行自然語言處理前，需要先進行分詞，確認文本內可能的詞。所以要進行研究主題分析，首要工作是從論文中確認重要的中英文詞組以及中文的多字詞。然而學術論文中經常有許多新的詞語出現，來代表新的概念、方法和技術，我們無法事先收錄各個領域裡所有可能的詞語來製作十分完整的詞典，進行斷詞。而且利用構詞律的規則式斷詞方法，需要處理同時中文和英文兩種語言的文本，難以整合應用。所以本論文採用統計式的處理方法[Chien, et. al., 1999]，以便同時解決中文的多字詞及中英文的詞組問題。

本論文所使用的方法如下：首先利用題名、摘要和參考文獻的題名等論文資料裡所有的文句建立一個 PAT-tree 資料結構，用來儲存所有出現在論文資料中的字串及它們所在的論文資料[Chien, 1997]。接著在 PAT-tree 中擷取可能的字串作為候選詞語，以統計訊息及經驗法則(heuristic rules)作為判斷字串是否為詞語的標準。

在本論文中，所使用的統計訊息包括字串在所有資料中的出現總頻次、字串在出現論文中的平均頻次和標準差(standard deviation)以及字串前後接字的複雜度。字串的出現總頻次代表該字串在領域中的重要性，出現頻次高表示這個字串在領域裡的論文經常出現而具有重要意義。字串在出現論文中的平均頻次和標準差用來表示字串對出現論文的重要程度，如式(1)

$$R_S \overset{def}{=} m_S + \sigma_S \tag{1}$$

在式(1)，m_S和 σ_S分別代表字串 S在出現論文中的平均頻次和標準差。當字串 S的平均頻次超過某一閾值時，表示此字串極有可能在許多論文中出現多次，是這些論文的關鍵詞語，應該被選取出來。或是雖然字串 S 在論文的平均頻次較低，但在某些論文中出現多次，是這些論文的關鍵詞語，也需要被選取出來，此時字串 S 會有一個較大的標準差 σ_S。因此，我們可以利用字串在出現論文中的平均頻次和標準差的總和 R_S代表字串對出現論文的重要程度，R_S值愈高的字串對出現論文愈重要。

字串前後接字的複雜度則可以判斷是否是一個完整的詞語或是其他詞語的部分，字串 S的前後接字複雜度 C_{1S}和 C_{2S}分別如式(2a)和(2b)所示

$$C_{1S} \overset{def}{=} -\sum_a \frac{F_{aS}}{F_S} \log(\frac{F_{aS}}{F_S}) \tag{2a}$$

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

式(2a)和(2b)中，a 和 b 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 的出現總頻次。以式(2a)前接字的情形來看，若是字串 S 有愈多種類的前接字，而且每一種前接字出現的次數越接近時，C_{1S}的值愈大，反之，當字串前只有一種前接字時，C_{1S}的值等於 0，或是有一個前接字出現的機會較其他大非常多時，則 C_{1S}的值接近於 0，表示該字串再加上這個前接字可能才是一個詞語。愈大的前接字複雜度代表該字串愈有可能是獨立的詞語而不是其他詞語的一部分；後接字的情形也是相同的道理。

通過上面條件的字串，再利用停用詞(stop words)不能出現在字串首尾的經驗法則，進一步過濾去不完整的詞語。在過去的經驗中，介詞和定詞等停用詞常出現在抽取出字串的首尾，如 “名詞+的”、“名詞+of”或“to+動詞”等詞組結構。但停用詞出現在字串的中間代表特定的詞組，例如“part of speech”，因此，將這種情形加以保留。

在確認論文資料中重要的中英文詞組以及中文的多字詞後，以這些詞語建立斷詞處理所需的詞典。我們使用長詞優先法則與詞語的出現總頻次將所有論文資料加以斷詞，確認所有在論文資料中出現的詞語。此時，我們分出來的詞語包括了一些中英文的詞組、詞和一些中文單字。在本論文中並非需要確認論文資料中所有可能的詞語，而是希望抽取所有可能代表研究主題的關鍵詞語，因此我們過濾具有以下情形的詞語。首先是中文資料中的單字(characters)，多半是一些介詞、

停用詞或是在上一步驟中無法組成詞的詞語，加以濾掉。其次，出現總頻次與式(1)之 R_S 值太小的詞語，也加以過濾，其理由如前面所述。剩下的詞語與它們在論文資料中的出現情形則是下一階段分析的對象。

四、研究主題叢聚

為了探索學術領域中重要的研究主題，本論文依據詞語在論文資料的共現關係，建立詞語之間的相關程度，將詞語進行叢聚，以一組叢聚的相關詞語作為一個研究主題。由於有些詞語可能包含在不同的研究主題中，本節中提出一個可以對詞語進行多重叢集的演算法。

首先，我們將上一階段抽取出來的詞語，利用可以進行多重叢聚的 cliques 叢集演算法[Kowalski and Maybury, 2000]進行詞語叢集。在選定最小相關程度的情形下，我們可以得到若干個詞語叢集，在這些詞語叢集中的詞語，彼此間的相關程度都在所選定的最小相關程度之上；而且詞語因它們與其他詞語相關程度的不同，可以叢集在多個集合中。本論文所使用的詞語相關程度的計算方式如下：我們先計算每一詞語在每一篇論文的題名、摘要和參考文獻的題名等資料中出現的頻次，作為詞語的特徵值運算的資訊。但因為只取用上述論文資料的資料量較小，詞語在其中出現的頻次不會太高，為了使低頻次的詞語差異不會太大，以詞語在每一篇論文資料的頻次的平方根作為一個特徵值。如此一來，對每一詞語有一組特徵向量(feature vector)，計算詞語間的相關程度便可以利用所對應特徵向量間夾

角的餘弦值(cosine value)來估算。如式(3)和式(4)分別表示詞語 A 的特徵向量和詞語 A 與 B 間的相關程度估算方式。

$$\vec{v}_A' \overset{def}{=} [\sqrt{f_{1,A}}, \sqrt{f_{2,A}}, \ldots, \sqrt{f_{N,A}}] \tag{3}$$

$$R(A,B) \overset{def}{=} \frac{\vec{v}_A \cdot \vec{v}_B}{\|\vec{v}_A\| \|\vec{v}_B\|} \tag{4}$$

式(3)中，$f_{i,A}$ 代表詞語 A 在第 i 篇論文資料中出現的頻次。式(4)中，分子部分是詞語 A 和 B 的特徵向量 $\vec{v}_A$ 和 $\vec{v}_B$ 內積(inner product)的值，分母部分則是兩個特徵向量長度 $\|\vec{v}_A\|$ 和 $\|\vec{v}_B\|$ 的乘積。經過 cliques 演算法所得到的結果是相當嚴格的，只有以詞語的共現關係所估算的相關程度在某一閾值以上的一對詞語才有可能叢集在一個集合內。然而，在研究主題中相同或相近的概念可能以不同詞語來表示，這些詞語不一定出現在相同的論文資料中，利用上述以詞語共現現象的相關程度估算方法將會得到很小的相關程度估算值，無法利用 cliques 演算法將這些詞語叢集起來。本論文採用以下兩種技術來解決上述的問題。

首先我們利用 LSI(Latent Semantics Indexing)技術對上述的特徵向量所形成的『詞語-特徵』矩陣 M 進行奇異值分解 (SVD, singular value decomposition)運算[Deerwester, et. al., 1990]，將矩陣 M 分解成三個矩陣，T_o、S_o 和 D_o，使得 $M = T_o S_o D_o'$。此處 T_o 和 D_o 為 M 的左、右奇異向量(singular vectors)所形成的矩陣，其大小分別為 $t \times r$ 和 $d \times r$，t 和 d 分別為詞語和特徵的數目，r 則為矩陣 M 的秩(rank)，而 S_o 為一個大小為 $r \times r$ 的對角線矩陣(diagonal matrix)，其對角線上的值為

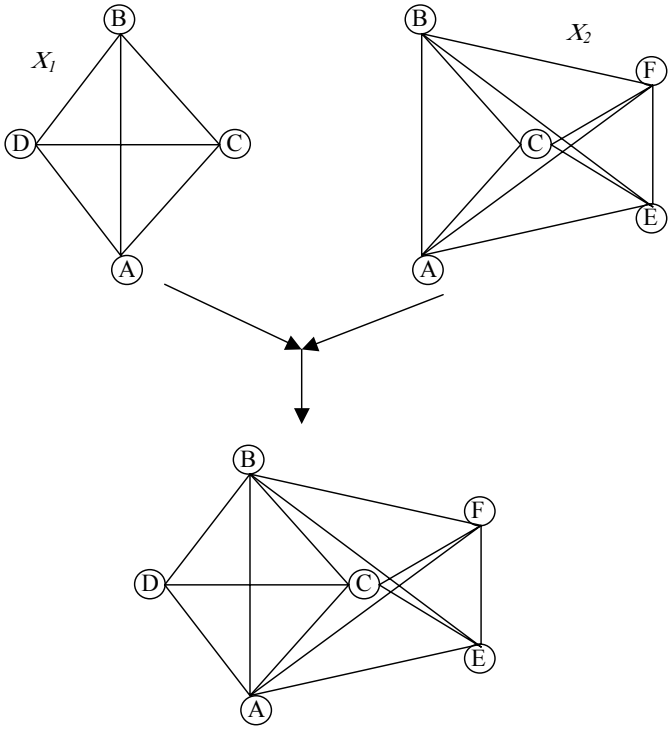
M 的奇異值(singular values)，且依據遞減的方式排列。若我們希望取得一個秩為 k 的矩陣 $\hat{M}$ ，$k<=r$，並使得 $\hat{M}$ 與 M 的最小平方差(least square error)最接近，可以取 S_o 對角線上的前 k 個奇異值，產生一個大小為 $k\times k$ 的新矩陣 S，同時 T_o 和 D_o 也分別取前 k 個行向量(column vectors)，形成矩陣 T 和 D，大小分別為 $t\times k$ 和 $d\times k$。矩陣 $\hat{M}$ 便可由 $\hat{M} = TSD'$ 計算得到。在使用 LSI 技術的檢索過程，當進行詞語的相關程度估算時，以 $\hat{M}\hat{M}'$ 來估算原先以 MM' 計算兩兩詞語特徵向量間的內積值，如式(5)所表示，

$$MM' \approx \hat{M}\hat{M}' = TSD'(TSD')' = TSD'DS'T' = TSS'T' = TS^2T' \qquad (5)$$

在式(5)中，由於矩陣 D 中的行向量彼此互為單位正交(othonormal)，$DD'=I$，而且 S 為對角線矩陣，$S'=S$，所以 $\hat{M}\hat{M}' = TS^2T'$。利用 SVD 取得隱含語義結構(latent semantic structure)的特性，使得原先因為共現關係較弱或是不存在，而相關程度較估算得很小的兩個相關詞語，可以獲得較大的估算值[Deerwester, et. al., 1990]。

其次，進行 cliques 叢集演算法後，我們對於所得到的結果依據它們成員間重疊的情形再次進行叢集。假設兩個叢集之間有三個以上的成員是相同的，而且其餘的成員間雖然沒有很強的詞語共現關係，但是也曾在某些論文資料中一起出現，我們即將這兩個叢集的詞語集合進行聯集，產生新叢集。如圖二所示， 在 A、B、C、D、E 和 F 六個相關詞語中，依據它們的共現關係進行 cliques 叢集，叢集成 X_1 和 X_2 兩個詞語集合。在這兩個叢集間，有三個詞語 A、B 和 C 是相同，而且

經過比對，叢集X_1剩下的成員 D 和叢集X_2剩下的成員 E 和 F 出現的論文資料有一些是相同的，我們便將X_1和X_2兩個叢集進行合併，使得所得到的叢集更具有研究主題的代表性。



圖二 將X_1和X_2兩個具有相同成員的叢集進行合併的示意圖

最後經過上述的叢集處理後，可以得到一些代表領域中重要研究主題的詞語叢集。在分析研究主題時，我們計算每一叢集與論文間的相關程度，計算方式依據為 LSI 的相關估計方式[Deerwester, et. al., 1990]，如式(6)計算叢集X對所有論文的相關程度。

$$R_X = \chi TSD' \tag{6}$$

式(6)中，χ為一個行向量，$\chi' = [e_1, e_2, \ldots, e_t]$，每一個元素代表一個特定詞語是

否出現在叢集X之中，換言之，如果第 i 個詞語包含於這個叢集中，則 e_i 的值為 1；否則若是這個叢集不包含這個詞語，e_i 的值為 0。式(6)所得到的結果 R_X也是一個行向量，大小為 $1\times d$，每一個元素所代表的值為詞語叢集X與所對應的論文之間的相關程度估算值。最後依據這個結果，將相關程度大的論文資料取出，作為研究主題相關的論文資料來進行分析。

五、國內計算語言學的研究主題分析的實驗結果

計算語言學研討會 ROCLING 是國內的計算語言學領域相當重要的學術活動。因此，ROCLING 的研討會論文集中論文資料，可以說是歷年來國內計算語言學領域學者的心血結晶，所蘊含的研究主題也是他們所關心的研究主題。因此，本論文將以 ROCLING 研討會的論文資料做為分析國內計算語言學研究主題的素材。

分析資料為從第一屆(1988)到第十四屆(2001)的 ROCLING 研討會論文，共 235 篇。進行詞語抽取時，首先抽取重要的多字詞及詞組，所設定的字串出現總頻次的閾值，較短的字串(2 或 3 字)設定為 15 次，較長的字串(4~5 字)則設定為 10 次，字串對出現論文資料的重要程度 R_S(平均頻次和標準差的總和)設為 2.5，前後接字的複雜度設定為 0.5。這些抽取出來的多字詞或詞組加入詞典後，對論文資料進行分詞，依據第三節的方法對所有詞語進行統計，過濾去不重要的詞語。最後的結果共得到 343 個關鍵詞語。由於篇幅所限，無法將所有的詞語一一列出，我們將

出現總頻次最高的前 50 個詞語及出現的總頻次列表於表一。

表一 關鍵詞語抽取所得到的前 50 個出現總頻次最高的詞語及總頻次

次序	詞名	出現總頻次	次序	詞名	出現總頻次
1	parsing	209	26	parser	80
2	speech	184	27	probabilistic	78
3	系統	175	28	動詞	78
4	sentences	141	29	語音	78
5	lexical	138	30	knowledge	74
6	mandarin	134	31	語法	74
7	speech recognition	132	32	chinese text	73
8	方法	131	33	語言	73
9	semantic	130	34	semantics	72
10	corpus	129	35	corpora	71
11	syntactic	107	36	used	71
12	recognition	106	37	國語	71
13	data	105	38	discourse	70
14	分析	104	39	處理	70
15	learning	102	40	dictionary	68
16	mandarin chinese	97	41	problem	65
17	sentence	97	42	分類	65
18	machine translation	92	43	corpus based	64
19	words	92	44	design	62
20	theory	87	45	information retrieval	62
21	rules	84	46	syntax	61
22	models	83	47	generation	60
23	phrase	83	48	語料庫	60
24	漢語	82	49	應用	60
25	classification	80	50	character	59

接著將取出來的詞語進行研究主題叢聚。進行詞語的 cliques 叢集時，我們分別以第四節中原先的詞語特徵向量與經過 SVD 處理的特徵向量，k 值為 30、60 及 120，進行相關程度估算。將相關程度的閾值設為 0.4，經過 cliques 叢集與叢集合

併後，所得到三個詞語以上的叢集的數目，如表二所示。

表二 不同相關程度估算方法進行研究主題叢集所得到的叢集數目

	Original feature vectors	SVD k=120	SVD k=60	SVD k=30
cliques 叢集	65	78	85	74
叢集合併	27	34	34	32

從表二中，可以觀察到經過 SVD 處理的 cliques 叢集數目較原先的特徵向量來得多，顯然 LSI 技術有助於捕捉詞語不共現卻相關的隱含語義結構，產生較多 cliques 叢集。因此，我們以經 SVD 處理 k 值為 60 的特徵向量進行詞語相關程度估算，將所得到的 34 個詞語叢集作為進一步的分析的對象，這 34 個詞語叢集列表於附錄一。

從詞語叢集的結果我們可以看到幾個現象。第一、若干叢集同時具有中文詞語與英文詞語，甚至包含縮寫與相同概念但不同詞名的詞語，比方說，叢集 12 包含了‘machine translation’、‘mt’、‘機器翻譯’等詞語；或是又如叢集 18 包含了 ‘word identification’、‘word segmentation’、‘斷詞’等詞語。可見將參考文獻的題名加入論文資料，可以獲得中文和英文兩種語言的詞彙訊息，而且利用詞語的共現關係可以將相關的詞語叢聚起來。第二、大部分的詞語叢聚都可以明顯地用來代表一個特定的研究主題。除了叢集 3、叢集 11 與叢集 29 由意義較廣泛的詞語形成之外，其餘叢集的詞語間都具有相關性，而且可以用來代表計算語言學領域中的特定研究主題。比方說，叢集 7 為語音辨認的相關詞語、叢集 9 則為文件分類的相關詞

語。因此，本論文所提出來的研究主題抽取方法的可行性便可以得到初步驗證。

表三 與語言的計算模式相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
23 語法模式與剖析	分析, 表達, 剖析, 格位, 訊息, 動詞, 結構, 詞類, 漢語, 語法, 語法模式, 語意, 模式, 關係	1989 "訊息為本的格位語法--一個適用於表達中文的語法模式" 1991 "連接詞的語法表達模式-以中文訊息格位語法(ICG)為本的表達形式" 1992 "漢語的動詞名物化初探--漢語中帶論元的名物化派生名詞"
18 斷詞	chinese text, chinese word segmentation, segmentation, unknown word, word identification, word segmentation, words, 斷詞	1994 "Chinese-Word Segmentation Based on Maximal-Matching and Bigram Techniques" 1995 "A Unifying Approach to Segmentation of Chinese and Its Application to Text Retrieval" 1997 "Unknown Word Detection for Chinese by a Corpus-based Learning Method" 1997 "Chinese Word Segmentation and Part-of-Speech Tagging in One Step" 1997 "A Simple Heuristic Approach for Word Segmentation"
22 統計式語言模型的建立	bigram, class based, clustering, entropy, language model, language modeling, language models, n gram	1994 "An Estimation of the Entropy of Chinese - A New Approach to Constructing Class-based n-gram Models" 1997 "Truncation on Combined Word-Based and Class-Based Language Model Using Kullback-Leibler Distance Criterion" 2001 "使用關聯法則為主之語言模型於擷取長距離中文文字關聯性"

由於篇幅的限制，本論文無法對所有抽取出來詞語叢集一一進行詳盡的報告，以下針對幾個主題較明確的詞語叢集進行說明。表三是與語言的計算模式相關的詞語叢集及相關論文的列表，論文前的數值是論文在 ROCLING 研討會中發表的年份。表三可以驗證早期的計算語言學多以規則式的語法模式與剖析為主，近來則較多發展統計式語言模型，而斷詞則是一直以來國內計算語言學領域相當重視的獨特問題。

此外從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，表四到表六分別列出與機器翻譯、語音處理和資訊檢索相關的集合。從表四的結果，說明機器翻譯是計算語言學最早的應用問題之一[Lenders, 2001]，而其發展從規則式的自動翻譯到統計式，近期的應用則是在跨語言檢索部分。

表四 與機器翻譯相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
12 機器翻譯	'bilingual', 'machine translation', 'mt', 'transfer', '機器翻譯'	1991 "Lexicon-Driven Transfer In English-Chinese Machine Translation" 1992 "A Modular and Statistical Approach to Machine Translation" (只有與叢集 12 相關) 1995 "THE NEW GENERATION BEHAVIORTRAN: DESIGN PHILOSOPHY AND SYSTEM ARCHITECTURE" 1996 "介詞翻譯法則的自動擷取" 2001 "統計式片語翻譯模型"
32 機器翻譯	'bilingual', 'machine translation', 'translation', '機器翻譯'	

在過去計算語言學所處理的對象多為書寫語言(orthographic languages)，近年來語音處理已經成為計算語言學相當重視的研究主題。從 ROCLING 的論文資料中所得到的結果可以分析成語言模型、聲學辨認以及語音合成三個研究主題(表五)。國內計算語言學較早進行研究的主題是語言模型和語音合成，近年在聲學辨認研究上，也有許多研究人員進入這個領域發表相關論文。在表五，另外還可將語音合成研究分成系統製作(叢集 30)與聲學訊息研究(叢集 31)兩個部分。

表五 與語音處理相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
13 語言模型	dictation, large vocabulary, 語言模型, 語音辨認	1993 "國語語音辨認中詞群雙連語言模型的解碼方法" 1994 "國語語音辨認中詞群語言模型之分群方法與應用" 1995 "應用於'音中仙'國語聽寫機之短語規則分析與建立" 1996 "國語語音辨認中多領域語言模型之訓練、偵測與調適" 1999 "國語電話語音辨認之強健性特徵參數及其調整方法" (只有與叢集 17 相關)
17 語言模型	國語, 語言模型, 語音辨認, 辨認	
7 聲學辨認	hidden markov, maximum, robust speech recognition, speech recognition	1998 "Speaker-Independent Continuous Mandarin Speech Recognition Under Telephone Environments" 1999 "國語電話語音辨認之強健性特徵參數及其調整方法" 2000 "具有累進學習能力之貝氏預測法則在汽車語音辨識之應用" 2000 "綜合麥克風陣列及模型調整技術之遠距離語音辨識系統"
30 語音合成	speech, synthesis, 文句翻語音, 合成, 系統, 音節, 國語, 連音, 語音, 輸入	1995 "以 CELP 為基礎之文句翻語音中韻律訊息之產生與調整" 1996 "時間比例基週波形內差--一個國語音節信號合成之新方法" 1996 "中英文文句翻語音系統中連音處理之研究" 1999 "台語多聲調音節合成單元資料庫暨文字轉語音雛形系統之發展" (只有與叢集 30 相關) 1999 "國語文句翻台語語音系統之研究" (只有與叢集 30 相關) 2001 "Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method", (只有與叢集 31 相關)
31 語音合成	mandarin text to speech, pitch, prosodic, speech, synthesis, 文句翻語音, 合成	

在計算語言學領域中，資訊檢索比起其他研究可說是一個較新的主題，然而由於網際網路與電子文件的發展使得這項應用成為相當具有潛力的研究主題。我們可以從表六中發現國內計算語言學在這方面的重要研究包括資訊檢索和文件分類。

表六 與資訊檢索相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
25 資訊檢索	csmart, databases, document, indexing, information retrieval, retrieval, text retrieval, 檢索	1995 "適合大量中文文件全文檢索的索引及資料壓縮技術" 1996 "尋易(Csmart-II):智慧型網路中文資訊檢索系統" 1997 "An Assessment on Character-based Chinese News Filtering Using Latent Semantic Indexing" 1999 "A New Syllable-Based Approach for Retrieving Mandarin Spoken Documents Using Short Speech Queries"
9 文件分類	document, hierarchical, text categorization, 分類, 文件, 文件分類, 特徵	1993 "中文文件自動分類之研究" 1999 "階層式文件自動分類之特徵選取研究" 2001 "基於階層式神經網路之自動文件分類方法" 2001 "適應性文件分類系統"
28 文件分類	document, text categorization, 分類, 文件分類, 文件自動, 關鍵詞	

六、結論

本論文針對研究主題分析的問題，提出一系列以自然語言處理為基礎的技術，從學術領域中發表的論文資料中抽取重要的關鍵詞語，並將這些詞語依據彼此間共現關係進行叢集，以叢集所得到的詞語集合表示領域中重要的研究主題。在本論文中，我們將所提出的方法應用到 ROCLING 研討會的論文資料上，抽取計算語言學領域的重要研究主題，結果顯示這個方法可以同時抽取出中文和英文的關鍵詞語，所得到的詞語叢集結果也可以表示領域中重要的研究主題。這樣的結果初步驗證了本論文所提出方法的可行性。從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，抽取出來的詞語叢集中有許多與機器翻譯、語音處理和資訊檢索相關，在語言的計算模式上，語法模式與剖析、斷詞和語言

模型則是國內計算語言學家所關心的主題。

在後續的研究上，除了進一步改善目前所提出來的方法，並且深入探討各研究主題的起源、發展與演變之外，我們將探索各個研究主題之間的相關性，並嘗試將結果以圖形化的方式加以呈現。另外，對於不同學術領域間的相關研究主題的發掘和分析，比方說資訊檢索同樣是圖書資訊學所關心的研究主題，如何利用自然語言處理技術來分析兩個領域間的共通與相異，是一項值得探討的研究。

致謝

本研究受國科會「國內計算語言學學術資訊交流之研究(I)」(編號NSC91-2413-H-128-004-)計畫案補助。另外，也感謝三位審查委員所提供的意見。

參考文獻

[Bishop, 1999] A. P. Bishop, “Document Structure and Digital Libraries: How Researchers Mobilize Information in Journal Articles”, Information Processing and Management, 35, p255-279.

[Chien, 1997] Lee-Feng Chien, “PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval”, SIGIR’97, p50-58.

[Chien, et. al., 1999] Lee-Feng Chien, Chun-Liang Chen, Wen-Hsiang Lu, and Yuan-Lu Chang, “Recent Results on Domain-Specific Term Extraction From Online Chinese Text Resources”, ROCLING XII, p203-218.

[Church and Mercer, 1993] K. W. Church and R. L. Mercer, “Introduction to the Special Issue on Computational Linguistics Using Large Corpora”, Computational Linguistics, 19(1), p1-24.

[Covi, 1999] L. M. Covi, “Material Mastery: Situating Digital Library Use in University Research Practices”, Information Processing and Management, 35, p293-316.

[Deerwester, et. al., 1990] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis”, Journal of the American Society for Information Science,

41(6), p391-407.

[Harter, 1992] S. P. Harter, “Psychological Relevance and Information Science”, Journal of the American Society for Information Science, 43(9), p602-615.

[Hatzivassiloglou, Gravano and Maganti, 2000] V. Hatzivassiloglou, L. Gravano and A. Maganti, “An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering”, SIGIR’2000, p224-231.

[Hjørland and Albrechtsen, 1995] B. Hjørland and H. Albrechtsen, “Towards a New Horizon in Information Science: Domain-Analysis”, Journal of the American Society for Information Science, 46(6), p400-425.

[Huang, 2000] Chu-Ren Huang, “From Quantitative to Qualitative Studies: Developments in Chinese Computational and Corpus Linguistics”, 漢學研究, 第十八卷特刊, p473-509.

[Kowalski and Maybury, 2000] G. J. Kowalski and M. T. Maybury, “Document and Term Clustering”, Information Storage and Retrieval Systems: Theory and Implementation, 2nd ed., Chapter 6, p139-163.

[Lenders, 2001] W. Lenders, “Past and Future Goals of Computational Linguistics”, ROCLING XIV, p213-236.

[Tabah, 1996] A. N. Tabah, Information Epidemics and the Growth of Physics, Ph. D. Dissertation of McGill University, Canada.

[Wayne, 2000] C. L. Wayne, “Topic Detection and Tracking in English and Chinese”, IRAL 5, p165-172.

[Wilson, 1999] T. D. Wilson, “Models in Information Behaviour Research”, Journal of Documentation, 55(3), p249-270.

[Yang, Pierce and Carbonell, 1998] Y. Yang, T. Pierce, and J. Carbonell, “A Study on Retrospective and On-Line Event Detection”, SIGIR’98, p28-36.

[王士元, 1988] “電腦在語言學裡的運用”, ROCLING I, p257-287.

附錄一 ROCLING 研討會論文資料所得到的詞語叢集

叢集編號	詞語
1	generation, generator, systemic, text generation
2	acquisition, explanation, generalization, learning
3	方法, 系統, 問題, 處理
4	initial, min, taiwanese, 台語, 台灣, 資料庫

叢集編號	詞語
5	atn, attachment, pp, preference
6	complexity, computational, gpsg, morphology
7	hidden markov, maximum, robust speech recognition, speech recognition
8	aspect, logic, temporal, tense
9	document, hierarchical, text categorization, 分類, 文件, 文件分類, 特徵
10	classifiers, decision, non, symbols
11	分析, 系統, 處理, 語言
12	bilingual, machine translation, mt, transfer, 機器翻譯
13	dictation, large vocabulary, 語言模型, 語音辨認
14	adaptation, maximum, robust speech recognition, 語音辨識
15	attachment, pp, preference, score
16	系統, 設計, 輸入, 鍵盤
17	國語, 語言模型, 語音辨認, 辨認
18	chinese text, chinese word segmentation, segmentation, unknown word, word identification, word segmentation, words, 斷詞
19	attention, conversation, discourse, elicitation, interaction
20	continuous, hidden markov, maximum, speech recognition
21	統計, 詞彙, 語言, 語料
22	bigram, class based, clustering, entropy, language model, language modeling, language models, n gram
23	分析, 表達, 剖析, 格位, 訊息, 動詞, 結構, 詞類, 漢語, 語法, 語法模式, 語意, 模式, 關係
24	adaptive, compression, scheme, 英文, 資料, 調整, 壓縮
25	csmart, databases, document, indexing, information retrieval, retrieval, text retrieval, 檢索
26	grammars, parser, parsing, sentence
27	continuous, large vocabulary, mandarin, speaker, speech, speech recognition, telephone
28	document, text categorization, 分類, 文件分類, 文件自動, 關鍵詞
29	方法, 系統, 設計, 應用
30	speech, synthesis, 文句翻語音, 合成, 系統, 音節, 國語, 連音, 語音, 輸入
31	mandarin text to speech, pitch, prosodic, speech, synthesis, 文句翻語音, 合成
32	bilingual, machine translation, translation, 機器翻譯
33	explanation, generalization, learning, parse
34	aspect, functional, lexical, lexical semantic, mandarin chinese, meaning, parsing, phrase, roles, semantic, semantics, syntactic, syntax, thematic, theory, verb, verbal, verbs