Computational Linguistics and Chinese Language Processing
Vol. 9, No.1, February 2004, pp.97-112

基於術語抽取與術語叢集技術的主題抽取

Topic Extraction Based on Techniques of Term Extraction and Term Clustering

林頌堅*

Sung-Chen Lin*

摘要

本論文針對主題抽取的問題，提出一系列以自然語言處理爲基礎的技術，應用這些技術可以從學術論文抽取重要的術語，並將這些術語依據彼此間的共現關係進行叢集，以叢集所得到的術語集合表示領域中重要的主題，提供研究人員學術領域的梗概並釐清他們的資訊需求。我們將所提出的方法應用到 ROCLING 研討會的論文資料上，結果顯示這個方法可以同時抽取出計算語言學領域的中文和英文術語，所得到的術語叢集結果也可以表示領域中重要的主題。這個初步的研究驗證了本論文所提出方法的可行性。重要的主題包括機器翻譯、語音處理、資訊檢索、語法模式與剖析、斷詞和統計式語言模型等等。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。

關鍵詞: 主題抽取、術語抽取、術語叢集

Abstract

In this paper, we propose a series of natural language processing techniques to be used to extract important topics in a given research field. Topics as defined in this paper are important research problems, theories, and technical methods of the examined field, and we can represent them with groups of relevant terms. The terms are extracted from the texts of papers published in the field, including titles, abstracts, and bibliographies, because they convey important research information and are relevant to knowledge in that field. The topics can provide a clear outline of the field for researchers and are also useful for identifying users' information

*世新大學資訊傳播學系 Department of Information and Communications, Shih-Hsin University, Taipei, Taiwan, R.O.C.
Email: scl@cc.shu.edu.tw

needs when they are applied to information retrieval. To facilitate topic extraction, key terms in both Chinese and English are extracted from papers and are clustered into groups consisting of terms that frequently co-occur with each other. First, a PAT-tree is generated that stores all possible character strings appearing in the texts of papers. Character strings are retrieved from the PAT-tree as candidates of extracted terms and are tested using the statistical information of the string to filter out impossible candidates. The statistical information for a string includes (1) the total frequency count of the string in all the input papers, (2) the sum of the average frequency and the standard deviation of the string in each paper, and (3) the complexity of the front and rear adjacent character of the string. The total frequency count of the string and the sum of its average frequency and standard deviation are used to measure the importance of the corresponding term to the field. The complexity of adjacent characters is a criterion used to determine whether the string is a complete token of a term. The less complexity the adjacent characters, the more likely the string is a partial token of other terms. Finally, if the leftmost or rightmost part of a string is a stop word, the string is also filtered out. The extracted results are clustered to generate term groups according to their co-occurrences. Several techniques are used in the clustering algorithm to obtain multiple clustering results, including the clique algorithm and a group merging procedure. When the clique algorithm is performed, the latent semantic indexing technique is used to estimate the relevance between two terms to improve the deficiency of term co-occurrences in the papers. Two term groups are further merged into a new one when their members are similar because it is possible that the clusters represent the same topic. The above techniques were applied to the proceedings of ROCLING to uncover topics in the field of computational linguistics. The results show that the key terms in both Chinese and English were extracted successfully, and that the clustered groups represented the topics of computational linguistics. Therefore, the initial study proved the feasibility of the proposed techniques. The extracted topics included “machine translation,” “speech processing,” “information retrieval,” “grammars and parsers,” “Chinese word segmentation,” and “statistical language models.” From the results, we can observe that there is a close relation between basic research and applications in computational linguistics.

Keywords: Topic extraction, term extraction, term clustering

1. 緒論

本論文提出一個自動化的主題抽取方法，利用論文中的詞彙訊息來抽取學術領域的主題。論文的題名、摘要、本文，甚至所引用的參考文獻題名等文字資料表達了研究的問

題、方法與結果，因此這些論文資料中的術語與研究主題非常相關。以本論文做一例子，在題名、摘要和本文出現許多『學術領域』、『主題』、『論文』、『抽取』等等術語，可以了解這個研究與從學術論文中抽取主題相關。所以抽取論文中的術語可以了解論文的主題。在一個學術領域中，受到重視的主題的相關術語會在許多論文中出現。以計算語言學領域爲例，許多論文包含了諸如『語料庫』、『剖析』、『資訊檢索』等等術語，因爲它們與這個領域的重要主題相關。而且進一步地，主題相關的術語會經常一起出現，具有較強的共現(co-occurrence)關係。因此，如果對學術領域相關的論文進行分析，選取具有高頻而代表主題意義的術語，利用共現資訊將相關的術語叢集成一個集合，所形成的術語集合便可以視爲是領域中重要的主題。在分析論文的主題時，便可以透過論文對各術語集合的相關性來進行評估。

因應學術論文較多獨特術語的特性，本研究在術語抽取(term extraction)的技術上，參考[Chien, 1997]、[Chien, ***et. al.***, 1999]和[Zhang, ***et. al.***, 2000]等統計方法，利用字串的頻次爲基礎的統計訊息，從論文中抽取多語的術語。在術語叢集(term clustering)上，則考慮同義詞和一詞多義的現象，利用 LSI (latent semantics indexing) [Deerwester, ***et. al.***, 1990] 和 clique 叢集演算法[Kowalski and Maybury, 2000]等技術，將經常共現的術語叢集起來。在應用上，我們使用 ROCLING 一到十四屆學術研討會的論文資料，進行術語抽取與術語叢集。研究結果初步驗證了這些技術用於主題抽取的可行性。

本論文其餘的章節架構如下：在第二節中扼要說明相關研究及所提出一系列之技術。接著在第三節和第四節中分述這個研究的核心技術：術語抽取和術語叢集。第四節中並且說明主題與論文之間相關程度的計算方式。第五節是應用這些技術到國內計算語言學領域的研究與結果。第六節則是本論文的結論。

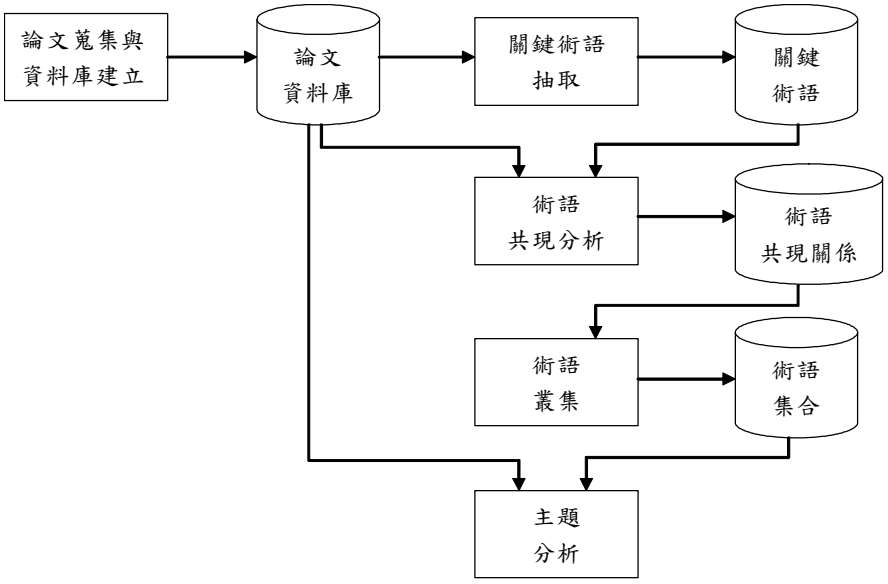
2. 本論文提出的主題分析方法

本論文希望發展主題抽取的技術，從相關論文抽取重要的主題。在資訊檢索研究的範疇中類似的研究有主題偵測(topic detection)。主題偵測希望從一序列來源各不相同的新聞中，偵測出某些『事件』(events)相關的連續報導[Wayne, 2000]。目前許多的研究利用『叢集假說』(cluster hypothesis) 來解決這個問題[Yang, Pierce and Carbonell, 1998][Hatzivassiloglou, Gravano and Maganti, 2000]，以文件叢集(document clustering)技術，利用相關文件具有相似的術語分布，統計新進文件的術語分布情形，將文件歸入相關事件的集合中。因此，本論文也嘗試應用叢集假說發展相關技術。再者，主題偵測研究已應用專有名詞(proper nouns)等術語作爲區隔不同新聞事件的重要訊息[Hatzivassiloglou, Gravano and Maganti, 2000]，因此本論文也將嘗試利用論文中的相關術語。此外，主題偵測應用所謂的『新聞熱潮』(news bursts)現象，將時間訊息加入叢集演算法，提昇偵測的結果[Yang, Pierce and Carbonell, 1998]。雖然學術論文有所謂『資訊流行』(information epidemics)的說法，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，我們利用術語在論文中的共現關係，找出術語的叢集情形來代表主題。

以論文中出現的術語取代整篇論文作爲分析對象的主要原因是希望能獲得較可信賴的統計訊息。由於較小的學術領域所出版的論文數量較爲不足，統計上不易得到滿意的分析結果。以術語作爲分析對象，因爲數量較多，可以獲得充足的統計訊息，克服文件數量較少的問題。具有多個主題的論文也可藉由術語的叢集，找出所有的相關主題，並且進而探索主題間的關係。此外，文件叢集不易直接詮釋結果所代表的主題，術語叢集則可以由成員的語意進行解釋。

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爲術語抽取與叢集分析的資訊，論文作者和出版年等項目則可以用來作爲後續的分析工作上。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都具有雙語的題名和摘要。若只針對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術語在論文中的共現訊息，無法分析出這些術語的相關性，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，提出可以同時獲得兩種語言的術語之方法。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，通常論文的主題與其他的文獻相關時會加以引用，因此參考文獻的題名與主題間也有密切的關係，加入參考文獻的題名可以增加分析的資訊，而且引用的參考文獻可能來自中英文兩種語言，若能利用適當的多語術語抽取技術，便可以統計分屬兩種語言的相關術語的共現現象，整合兩種語言的術語訊息，得到較佳的結果。



圖一 本論文的主題抽取方法

在建立好論文資料庫後，便利用第 3 節所描述的多語術語抽取方法從論文資料中自動抽取領域中具有意義的術語。接著以第 4 節的術語叢集技術統計術語在論文中的共現關係，將相關的術語叢集成集合，用來代表特定的主題。進行主題分析時，對於某一主題，可以根據術語集合與論文的相關程度，取出具有主題的論文。

3. 多語環境下的術語抽取

爲了抽取主題相關的術語，我們首先確認重要的中英文詞組(phrases)以及中文的多字詞，再選擇具有代表意義的術語，作爲這一階段的結果。在學術論文中，常以詞組的形式表達重要的主題，比方在計算語言學領域中，可以發現如英文的"language model"、"machine translation"或是中文的"語言模型"、 "機器翻譯"等等都是重要術語。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，需要先斷詞。然而學術論文中經常有許多新的術語出現，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術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，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。

在過去對於術語抽取的相關研究中，曾利用字串的『相對頻率』(relative frequency)、『互見資訊』(mutual information)和『上下文依附』(context dependency)等各種統計訊息[Su, *et. al.*, 1994][Chien, 1997][Zhang, *et. al.*, 2000]。字串的『相對頻率』是指該字串的出現頻次與語料中所有長度相同字串平均頻次的比值，可以測量字串的重要性，相對頻率愈大的字串愈重要，愈可能是一個術語[Su, *et. al.*, 1994]。『互見資訊』雖然有不同計算公式，但都是用來測量組成術語的字或詞彼此間的相互關係(association)，成員間『互見資訊』愈高的字串，愈有可能是一個術語[Su, *et. al.*, 1994] [Zhang, *et. al.*, 2000]。『上下文依附』則用來測量字串與上下文字詞間的依附程度，依附程度較大的字串可能是術語的一個部份，不應被抽取出來；反之，字串的依附程度較小，則可能代表是術語的邊界[Chien, 1997][Zhang, *et. al.*, 2000]。

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建立一個 PAT-tree 資料結構，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爲候選術語，以統計訊息及經驗法則(heuristic rules)作爲判斷是否爲術語的標準。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次數、字串的平均頻次和標準差(standard deviation)以及字串前後接字的複雜度。其中，字串的出現總次數、平均頻次和標準差等統計訊息和相關研究中的『相對頻率』作用相同，在於衡量字串的重要性。字串的出現總次數代表在領域中的重要性，總次數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。字串對於出現論文的重要程度則用字串的平均頻次和標準差來表示，如式(1)

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

m_S 和 σ_S 分別代表字串 S 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，是這些論文的關鍵術語，應該被選取出來。或者雖然字串 S 的平均頻次較低，但在某些論文中出現相當多次，是這些論文的關鍵術語，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。

前後接字的複雜度則和『上下文依附』的作用相同，可以判斷字串是否是一個完整的術語或是其他術語的部分，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

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

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

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 的出現總次數。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術語，所以前接字複雜度愈大代表該字串愈有可能是獨立的術語。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，進一步過濾不完整的術語。在過去的經驗中，介詞、連接詞和補語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，以這些術語建立斷詞處理所需的詞典。我們使用長詞優先法則與術語的出現總頻次將所有論文資料加以斷詞。論文資料經過斷詞處理後，將產生了一些中英文的詞組、詞和一些中文單字。在這一階段的目標是抽取論文中所有可能代表主題的術語，因此我們過濾具有以下情形的字串。首先是中文單字，多半是一些停用詞或是無法組成術語的片段。其次，出現總次數與式(1)之 R_S 值太小的術語，因爲對領域的重要性較低，也加以過濾。剩下的術語則是下一階段分析的對象。

4. 術語叢集

本論文依據術語在論文資料的共現關係，將術語進行叢集，以一組叢集的相關術語作爲一個主題。由於有些術語可能包含在不同的主題中，本節中提出一個可以對術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術語，利用 cliques 叢集演算法[Kowalski and Maybury, 2000]進行術語叢集。cliques 叢集演算法在選定最小相關程度的情形下，可以得到若干個術語集合，在集合中的術語，彼此間的相關程度都在所選定的最小相關程度之上，而且術語可以被歸類到多個集合，因此符合多重叢集的要求。本論文所使用的相關程度計算方式如下：我們先計算每一術語在每一筆論文資料中出現的頻次，作爲術語的特徵值。但因爲術語在論文資料裡出現頻次不高，爲了使低頻次的術語差異不會太大，以頻次的平方根作爲特徵值，並除以術語的總頻次進行正規化(normalization)。如此一來，對每一術語便有一組特徵向量(feature vector)，如式(3)表示某一術語 *A* 的特徵向量。

$$\mathbf{r}_{v_A} \stackrel{def}{=} \frac{[\sqrt{f_{A,1}}, \sqrt{f_{A,2}}, \ldots, \sqrt{f_{A,N}}]^T}{\sum_{i=1}^{N} f_{A,i}} \tag{3}$$

式(3)中，$f_{A,i}$代表術語*A*在第*i*篇論文資料中出現的頻次，分母的 $\sum_{i=1}^{N} f_{A,i}$ 是 *A* 的總頻次。術語間的相關程度便可以利用特徵向量的內積 (inner product)來估算。

經過 cliques 演算法與上述的相關程度計算方式所得到的結果是相當嚴格的，相關術語若要叢集在同一個集合中，所有術語彼此間的共現關係必須都很強。然而，在論文中相同或相近的概念可能以不同術語來表示，使得相關的術語不一定經常共同出現，利用上述的估算方法將會得到很小的相關程度，無法將這些術語叢集起來。爲此，本論文採用以下兩種技術來加以補救。

首先我們改用 LSI 技術估算術語間的相關程度。LSI 技術是利用奇異值分解 (SVD, singular value decomposition)對上述的特徵向量所形成的『術語-特徵』矩陣 *M* 進行分解[Deerwester, *et. al.*, 1990]，產生新的矩陣 $\hat{M}$ ，假設 $\hat{M}$ 的秩(rank)爲 *k*，*k* 小於或等於原先矩陣 *M* 的秩，則 $\hat{M}$ 是所有秩爲 *k* 的矩陣中，與 *M* 的平方差最小的矩陣。以術語在新矩陣 $\hat{M}$ 所對應的行向量(row vector)取代原先之特徵向量，當進行術語的相關程度估算時，便可以 $\hat{M}\hat{M}^T$ 來估算原先以 MM^T 計算兩兩術語特徵向量間的內積值。利用 SVD 可以取得隱含語義結構(latent semantic structure)，使得原先因爲共現關係較弱或是不存在的兩個相關術語，獲得較大的估算值[Deerwester, *et. al.*, 1990]。

其次，在進行 cliques 叢集演算法後，對於所得到的結果依據它們成員間重疊的情形進行合併。假設兩個集合之間有多個成員是相同的，這兩個集合很可能屬於同一主題，我們即將這兩個術語集合進行聯集，產生新集合。以數學式表示如下，C_1 和 C_2 爲兩個集合，如果$|C_1 \cap C_2| \geq c*Min(|C_1|, |C_2|)$，則合併成新的集合$C_3$，此處 $c*Min(|C_1|, |C_2|)$是兩個集合的最小相同成員數，*c* 是一個介於 1 與 0 間的實數，$Min(|C_1|, |C_2|)$是取出兩個數值中最小值的函數。

經過上述的叢集處理後，可以得到代表重要主題的術語集合。在確認主題的相關論

文方面，可以利用 LSI 的估計方式[Deerwester, *et. al*., 1990]，計算每一術語集合與論文間的相關程度。計算方式是將術語集合中每一個術語的特徵向量相加，再正規化成單位向量，即可求得主題與所有論文之間的相關程度估算值。最後依據將相關程度大的論文資料取出，作爲主題的相關論文。

5. 國內計算語言學的主題抽取之實驗結果

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，可以說是歷年來國內計算語言學領域學者的心血結晶，所蘊含的主題也是他們所共同關心的主題。因此，本論文將以第一屆(1988)到第十四屆(2001) ROCLING 研討會的 235 篇論文資料做爲分析國內計算語言學主題的素材。

進行術語抽取時，本論文根據字串長度將字串出現總次數的閾值作不同的設定，較短的字串(2 或 3 字)設定爲 15 次，較長的字串(4~5 字)則設定爲 10 次，平均頻次和標準差的總和 R_S 和前後接字的複雜度分別設爲 2.5 與 0.5。接著利用抽取出來的多字詞或詞組對論文資料進行斷詞，過濾不是術語的字串，並進行統計。結果共得到 343 個術語，表一是出現總次數最高的前 50 個術語及它們的出現次數。表一中列出的術語大多屬於概念較廣泛的術語。這些術語出現在較多論文資料中，因此出現次數較高。表一中有些是其他領域也常見的術語，比方說『系統』、『方法』、『分析』等等，但許多術語和計算機科學及語言學相關，如『parsing』、『data』、『speech』、『lexical』等等，或是本身即是計算語言學特有的概念，如『speech recognition』、『machine translation』等等。

表一 術語抽取所得到的前 50 個出現總次數最高的術語

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

接著進行術語叢集，首先利用 LSI 技術進行相關程度估算。由於無法以客觀而有系統的方式決定新矩陣之秩的大小[Deerwester, *et. al.*, 1990] ，因此在本論文中分別嘗試秩爲 30、60 及 120 的新矩陣，產生術語的特徵向量來估算相關程度。將 clique 叢集中相關程度的閾值與集合合併 *c* 值分別設爲 0.4 和 0.6，最後所得到三個術語以上的叢集的數目與未被叢集術語的數目，如表二所示，另外爲了檢驗應用 LSI 技術的優點，表二中也同時顯示未經 SVD 處理的特徵向量之叢集結果。

表二 進行術語叢集所得到的結果

	SVD Rank=120	SVD Rank=60	SVD Rank=30	Original feature vector
cliques 叢集後的集合數目	78	85	74	65
合併後的集合數目	44	34	32	32
未被叢集的術語數目	209	189	214	223

從表二中，可以觀察到經過 SVD 處理的特徵向量，不論秩的大小，其未被叢集的術語數目都較原先特徵向量者少，換言之，LSI 技術有助於捕捉術語不共現卻相關的隱含語義結構，因此較多的術語可以被叢集。其中秩值爲 60 的特徵向量所得到的結果，是未被叢集的術語數目最少者，因此我們將所得到的 34 個術語集合作爲進一步分析的對象，這 34 個術語集合列表於附錄一。

從術語集合的結果我們可以看到幾個現象。第一、若干集合同時包含中文術語與英文術語，甚至包含縮寫與相同概念但不同詞名的術語，比方說，集合 12 包含了'machine translation'、'mt'、'機器翻譯'等術語；或是又如集合 18 包含了 'word identification'、'word segmentation'、'斷詞'等術語。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，而且利用術語的共現關係與 LSI 技術可以將相關的術語叢集起來。第二、大部分的術語集合都可以明顯地用來代表一個特定的主題。除了集合 3、11 與 29 由概念較廣泛的術語形成之外，其餘集合的術語間都具有相關性，可以用來代表計算語言學領域中的特定主題。比方說，集合 7 爲語音辨認的相關術語、集合 9 則爲文件分類的相關術語。此外，對於沒有被叢集的術語加以檢視，發現術語未被叢集的原因，一是該術語與其他術語間的相關程度較小，而這些往往都是一些主題較廣泛的術語，或者是該術語的主題相當特殊，僅有少數論文進行探討，因此僅與少數術語發生共現關係，無法形成術語集合。因此，本實驗所得到的術語集合大多具有概念明確，容易進一步詮釋結果，而本論文所提出來的主題抽取方法的可行性，便可以得到初步驗證。

由於篇幅的限制，本論文無法對所有抽取出的術語集合一一進行詳盡的報告，以下針對幾個主題較明確的術語集合進行說明。表三是與語言的計算模式相關的術語集合及相關論文的列表，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爲主，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。

表三 與語言的計算模式相關的術語集合及相關論文

集合編號	術語	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 "訊息爲本的格位語法--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)爲本的表達形式" 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	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 "使用關聯法則爲主之語言模型於擷取長距離中文文字關聯性"

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 與機器翻譯相關的術語集合及相關論文

集合編號	術語	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)
32 機器翻譯	'bilingual', 'machine translation', 'translation', '機器翻譯'	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"

表五 與語音處理相關的術語集合及相關論文

集合編號	術語	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 爲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	

在過去計算語言學所處理的對象多爲書寫語言(orthographic languages)，近年來語音處理已經成爲計算語言學相當重視的主題。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(集合 30)與聲學訊息研究(集合 31)兩個部分。

表六 與資訊檢索相關的術語集合及相關論文

集合編號	術語	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	

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爲相當具有潛力的主題。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。

6. 結論

本論文針對主題分析的問題，提出一系列以自然語言處理爲基礎的技術，從學術領域中發表的論文資料中抽取重要的術語，並將這些術語依據彼此間共現關係進行叢集，以叢集所得到的術語集合表示領域中重要的主題。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，結果初步驗證了本論文所提出方法的可行性。

在後續的研究上，將進一步改善目前所提出來的方法，比方說，目前術語叢集的效果還不十分理想，對於術語叢集技術的改進將是下一階段努力的目標。此外，在本研究中有許多參數需要設定，未來需要參考各種客觀的參數調整法來達到較佳的結果。最主要的工作在於深入探討各主題的起源、發展與演變之外，我們將探索各個主題之間的相關性，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主題，兩個領域間共通與相異的分析相當值得探討。

致謝

本研究爲國科會計畫 NSC91-2413-H-128-004-『國內計算語言學學術資訊交流之研究(I)』之研究成果。另外，作者亦對三位審查者的寶貴意見與建議深表感謝。

參考文獻

Lee-Feng Chien, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," *Proceedings of SIGIR'97*, 1997, pp. 50-58.

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," *Proceedings of ROCLING XII*, 1999, pp. 203-218.

K. W. Church and R. L. Mercer, "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *Computational Linguistics*, 19(1), 1993, pp.1-24.

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), pp. 391-407.

V. Hatzivassiloglou, L. Gravano and A. Maganti, "An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering," *Proceedings of SIGIR'2000*, 2000, pp. 224-231.

G. J. Kowalski and M. T. Maybury, "Document and Term Clustering," *Information Storage and Retrieval Systems: Theory and Implementation, 2nd ed.*, Chapter 6, 2000, pp.139-163.

W. Lenders, "Past and Future Goals of Computational Linguistics," *Proceedings of ROCLING XIV*, 2001, pp. 213-236.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang, "A Corpus-based Approach to Automatic Compound Extraction," *Proceedings of ACL 94*, 1994, pp. 242-247.

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

C. L. Wayne, "Topic Detection and Tracking in English and Chinese," *Proceedings of IRAL 5*, 2000, pp.165-172.

Y. Yang, T. Pierce, and J. Carbonell, "A Study on Retrospective and On-Line Event Detection," *Proceedings of SIGIR'98*, 1998, pp. 28-36.

Jian Zhang, Jianfeng Gao, and Ming Zhou, "Extraction of Chinese Compound Words: An Experimental Study on a Very Large Corpus," *Proceedings of the Second Chinese Language Processing Workshop*, 2000, pp. 132-139.

附錄一 ROCLING研討會論文資料所得到的術語叢集

叢集編號	術語
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
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

叢集編號	術語
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