關鍵詞自動擷取技術與相關詞回饋

Automatic Keyword Extraction and Relevance Feedback

曾　元　顯
輔仁大學圖書資訊學系
Email: tseng@blue.lins.fju.edu.tw
「中國圖書館學會會報 59 期」

Nov. 4, 1997

摘要 | 前言 | 擷取方法的比較 | 中文關鍵詞擷取 | 相關詞回饋 | 結語

摘要：

關鍵詞自動擷取乃資訊檢索領域的基礎與核心技術。本文中我們比較關鍵詞擷取技術的幾種主要方法，說明每一種方法的優缺點、適用情況以及國內研究的現況。此外我們也簡介自行發展的關鍵詞擷取方法運用於中、英文相關詞回饋的情況。初步的分析顯示其錯誤率低（18%以下）、精確率高（50%以上）。其主要的缺點則是受限於查詢結果的多寡，而查詢結果的多寡則與該查詢主題的館藏量有關。除了充實該類主題的館藏量外，亦可透過索引典的自動建立與運用加以改進。

Abstract

Automatic keyword extraction is an important and fundamental technology in advanced information retrieval system. This article briefly compares several major keyword extraction methods, lists their advantages and disadvantages, and reports recent research progress in Taiwan area. Besides, this article also describes the application of a keyword extraction algorithm developed by the author in an information retrieval system for relevance feedback. The preliminary analysis shows that the error rate of extracting relevant keywords is as low as 18%, and the precision rate is over 50%. The main disadvantage of this approach is that the extraction results highly depend on the retrieval results, which in turn highly depends on the data hold by the database. Except collecting more data, this problem can be alleviated by the application of a thesaurus constructed by the same keyword extraction algorithm.

關鍵字：資訊檢索、關鍵詞擷取、相關詞回饋

壹、前言

過去大部份的書目檢索系統，受限於資料庫管理系統（DataBase Management System）特殊的索引製作方式，僅能以布林邏輯及右切截比對功能提供資料查詢，對於書目記錄的全文式（左右切截）檢索，則以建立關鍵詞庫的方式達成。然而此類關鍵詞庫，必須以人工或半人工的方式建立，除了耗費大量人力、時間之外，還必須經常維護更新，以反應書目資料的新增異動。

目前網際網路通達的程度與普及速度，使資料成長更為快速，各種檢索系統的使用情況更為頻繁。新一代資訊檢索系統，尤其是允許全文式查詢的系統，必須能夠運用更具效率的自動化技術，以提供簡易有效的檢索服務。然而此類自動化技術，如自動索引、索引典自動建立 [1]、自動摘要 [2]、自動分類 [3]、相關回饋 [4]、自動過濾 [5]、概念檢索 [6] 等，大部份都必須先進行關鍵詞擷取（keyword extraction）的動作，依此結果再進行其他的處理。因此，無論是書目性資料或網路上的全文資料，關鍵詞自動擷取都是資訊檢索系統的基礎與核心技術，其重要性將隨網路的發展而越來越明顯。

「關鍵詞自動擷取」是一種辨認有意義且具代表性片語或詞彙的自動化技術。由於用途的差別，不同的研究，對此問題的定義、採用的方法、運用的條件與擷取的成效也各有差異。例如，自然語言處理的領域將此問題定義為「斷詞」問題（word segmentation），其目的在掃瞄一段文句，將此文句斷開成各個可賦予詞類的片語或單字，以做為機器翻譯或瞭解語意的基礎 [7]。因此其運用條件是即使輸入單一個句子，亦必須將構成句子的各個詞彙斷出來。由於斷出來的字彙中包含組成句子的各種詞類，如名詞、動詞、代名詞、連結詞、介系詞等，這種結果對資訊檢索而言，並非必要。因此，底下的討論將只針對關鍵詞擷取應用於資訊檢索的領域。

貳、擷取方法的比較

從文獻的分析得知 [8-12]，關鍵詞擷取的技巧主要有三種方法。第一種為詞庫比對法：即利用已建立的詞庫，來比對輸入文件（或文句），將文件中出現在詞庫中的片語擷取出來。此種方法製作簡單，只要將詞庫中的每個詞，去比對是否出現在輸入文件中即可。其結果都是詞庫中的正確詞彙，但並不保證所有關鍵

詞都能被擷取出來。除此之外，其缺點還包括：需要耗費人力、時間維護詞庫以容納各個領域的專業用語與新生詞彙，無法應付未曾預料的人名、地名、機構名等專有名稱，且詞庫越大比對速度越慢。

第二種為文法剖析法：透過自然語言處理技術的文法剖析程式，剖析出文件中的名詞片語，再運用一些方法與準則，過濾掉不適合的詞彙。其結果幾乎也都是有意義的名詞片語，但大部份的剖析程式，需要藉助已經建立的詞典或語料庫[13]，因此其缺點也和詞庫比對法一樣。除此之外，有些文法剖析法甚至只能剖析合乎文法的完整文句，使得書目、標題等資料裡的關鍵詞無法被擷取出來。

第三種方法為統計分析法：透過對文件的分析，累積足夠的統計參數後，再將統計參數符合某些條件的片語擷取出來。最簡單的統計參數是計數詞彙發生的頻率，即詞頻，將詞頻落在某一範圍的詞彙取出。由於沒有用到詞庫或語料庫，會有擷取錯誤的情況發生，得到無意義或不合法的詞彙。此外，統計參數不足的關鍵詞無法被選到。然而其優點是較不受語文國別與句型的限制，而且可以擷取出未曾被詞庫、語料庫網羅的專業用語、新生詞彙與專有名稱等片語。

其他的方法還包括上述方法的綜合運用，或加入一些變化。例如，利用一些排版規則，將重要的片語取出，如標題項、條列項中的文字，或強調詞（大寫、字頭語、斜體、加粗、加底線、引號內的文句）等等[3]。可以想見，各個方法都有其優缺點，運用時需要針對不同的環境條件加以考量。

關鍵詞在本文中雖定義為有意義且具代表性的片語或詞彙，然而關鍵詞的認定牽涉到個人的主觀判斷，且相同的詞彙在不同的主題下，也有不同的認定。在此情況下，要比較各種方法的擷取成效，並不容易。不過一些文獻對此問題仍有初步的探討，其中 Arppe [14] 以文法剖析方式試驗其擷取成效，結果發現大約 80%-99% 的關鍵詞為名詞片語，而且雖然名詞片語的擷取準確率與召回率皆可達 95% 以上，然而具代表性的名詞片語不到總數的 50%，因此單純剖析出名詞片語後，仍需要依據其他特徵以過濾掉不要的詞彙。Godby [15] 則比較文法剖析法與統計分析法的優劣，發現統計分析法除了可以跟文法剖析法做得一樣好之外，亦具備簡單、不受語文國別與句法的限制、以及可同時過濾不具代表性片語的優點。

參、中文關鍵詞擷取

國內對中文關鍵詞自動擷取的問題也有研究。清大自然語言處理實驗室曾嚐試擷取關鍵詞作為書後索引（book index），其主要方法為運用電子字典協助斷出詞彙 [16]，再以統計方式配合自然語言處理技術剖析名詞片語，最後再設定過濾條件，篩選索引詞彙 [17]。在成效評估方面，以一本軟體使用手冊為對象，相對於人工製作的索引，其精確率與召回率可同時達到 63% 的程度。至於導致錯誤的主要來源有：斷詞錯誤（42%）、統計特徵不足（39%）、以及無法處理複雜語法結構（19%）。

中央研究院資訊科學研究所也有關鍵詞自動擷取運用在資訊檢索的研究。其主要作法乃先建構一種稱為 PAT-tree 的資料結構，再輔以詞頻等統計特徵擷取出關鍵詞 [18, 19]。PAT-tree 雖然在資訊檢索上有相當優良的特性，不過其建造過程需耗費相當長的時間，例如，建構 600 Mega bytes 的資料需要一個星期的時間 [20]。可以想見，此種方式的有效運用，必須要能改進 PAT-tree 的建構速度。

最近，我們也發展出一套關鍵詞擷取的技術，並且已實際運用在輔仁大學圖書館的 OPAC 線上書目檢索系統上 ，提供相關詞回饋的功能 [21]。其方法為統計分析法，運用統計詞頻的方式來斷出關鍵詞，沒有用到辭典、語料庫、或自然語言處理的技巧。因此具備擷取速度快、擷取的正確率高（82%-100%）、中英文均適用、擷取的詞彙沒有長度限制、可同時擷取廣義詞與狹義詞等特性 [22]。下一節將簡介此種擷取方法的運用情形。

肆、相關詞回饋

在資訊檢索領域中，有一種查詢模式稱為「相關回饋」（relevance feedback）。其施行的方式是在前一階段找到的文件中，挑取重要的特徵，再回饋給系統，以期找到更多相關的資料。此種特徵若是文件本身，則可稱為相關文件回饋，若為相關詞，則稱為相關詞回饋，或檢索詞提示（term suggestion）。相關回饋在資訊檢索中被認為對檢索成效助益甚大 [20]。研究顯示，在一些全文資料庫中，可提昇檢索成效 20% [23]，而在醫學書目資料庫 MEDLINE 中，可提昇 16% 的檢索成效 [24]。

過去對於相關回饋的研究中，以相關文件回饋的方式居多，使用者只要在查詢結果的顯示螢幕上點選相關的文件，送回系統即可。然而在全文檢索環境中，

要判斷哪些文件相關，需要對文件做相當程度的瀏覽，此種情形常常造成使用者額外的負擔。相較之下，相關詞回饋因為牽涉到的額外資訊較少，使用者較易判斷，因此是一種比較好的相關回饋方式。然而目前提供此種回饋方式的系統比較少，這是因為讓系統自動斷出有用的相關詞，比起讓系統只提供文件讓使用者判斷，是較為複雜而困難的工作。一些系統即使做到相關詞回饋，目前也還不甚理想。以擁有鉅量網頁著稱的 AltaVista 檢索引擎為例，其所提供的相關詞為英文單字詞，至於對區分文件能力更具效果、表達更精確、對檢索成效幫助更大的英文片語則尚未提供。

在輔大書目檢索系統中 [25]，使用者可以利用模糊搜尋方式下達檢索條件，系統會將檢索結果以每頁二十筆資料的方式分頁顯示，在此同時，系統也會顯示從該頁書名中擷取出的關鍵詞，由於同一頁的結果應該都是與檢索條件相近的書目，因此從中擷取出來的關鍵詞應該是與此次檢索主題相關的相關詞。

表一列出十個檢索主題查詢得出的相關詞結果。從錯誤詞數當中可以瞭解此關鍵詞擷取方法擷取錯誤的情形很少，錯誤比率最低 0% ，最高 18%（2/11）。另外擷取出的關鍵詞有一半以上與檢索主題相關，最低比率 50%（4/8），最高 100%（3/3）。這些與檢索主題相關的詞彙可概略分為廣義詞、狹義詞、相關詞等具備近一步查詢參考價值的詞彙。這裡所謂廣義詞、狹義詞、相關詞主要是指字面上的意義而言。例如，從「服裝設計」檢索主題得出的四個相關詞中，「服裝」、「設計」為廣義詞，「實用服裝」歸類為相關詞，而「服裝設計」與檢索詞完全一樣，可以與其他相關詞一起運用，但沒有進一步單獨引用的必要，因此沒有歸類為上述任一詞類。

由於相關詞是從檢索結果擷取出來的，而檢索結果大都與原檢索主題字串相近，因此擷取出的相關詞也大都跟檢索主題字串相近。然而仍然會有與原檢索字串差異較大的相關詞彙出現。例如「素食」主題中的「健康」、「長壽」、「營養」，以及「Prolog 與人工智慧」中的「專家系統」。對具備模糊搜尋的檢索系統而言，這類與原檢索字串差異較大的相關詞，比較能夠拓展檢索的範圍。而與檢索字串相近的相關詞，其檢索效果則近似重新排列檢索結果。然而不管是拓展檢索範圍或是近似重新排列結果，對使用者而言都能提供檢索上的方便性。

	檢　索　詞	擷取出的	錯誤詞數	擷取出的	廣義詞	狹義詞	相關詞

		總詞數		相關詞數			
1	服裝設計	8	1	4	2	0	1
2	證券投資	5	0	5	1	2	1
3	精神分析學	7	1	5	1	1	2
4	兒童心理	10	0	8	1	6	0
5	蝴蝶生態	6	0	4	2	0	2
6	素食	11	2	9	0	4	4
7	C 語言	13	1	12	1	5	5
8	Prolog 與人工智慧	13	0	9	2	2	5
9	public library services for children and young adults	16	1	11	10	0	1
10	subject searching in online catalog systems	3	0	3	2	0	1

表一：相關詞擷取結果

此種相關詞擷取方式的主要缺點在於其相關詞彙太依賴於檢索結果。如果檢索結果太少或是得不到任何結果，則相關詞彙跟著減少，甚至付諸缺如。例如「subject searching in online catalog systems」主題中，系統只找回五筆資料，可供擷取相關詞的資料太少，以致只得出三個相關詞。如要改進此項缺點，除了充實該類主題的館藏外，勢必事先建立索引典。因此，索引典的自動建立與運用，將是未來的工作目標。

伍、結語

關鍵詞自動擷取乃資訊檢索領域的基礎與核心技術。過去中文方面的研究較少，未來如要將中文資訊檢索的領域拓展到自動索引、索引典自動建立、自動摘要、自動分類、相關回饋、自動過濾、概念檢索等地步，則中文方面的基礎技術還要再投入更多的研究。

本文中我們介紹了關鍵詞擷取技術的數種方法，說明每一種方法的優缺點與適用情況。此外，我們也簡介自行發展的關鍵詞擷取方法運用於相關詞回饋的情況。初步的分析顯示其錯誤率低（0%-18%）、精確率高（50%-100%）。而「召回率」方面，則由於系統內所有相關詞的認定困難而無法取得。未來的工作將在既有的基礎上進行索引典的自動建立與運用，以進一步提昇中、英文相關詞回饋的成效。

參考資料

[1] Gerard Salton, “Automatic Text Processing: The Transformation, Analysis, and

Retrieval of Information by Computer" Addison-Wesley, 1989.

[2] Timonthy C. Craven, "An Experiment in the Use of Tools for Computer-Assisted Abstracting" ASIS 1996 Annual Conference Proceedings, Oct. 19-24, 1996. Also available at http://www.asis.org/annual-96/ElectronicProceedings/craven.html

[3] Bruce Krulwich, "Learning Document Category Descriptions through the Extraction of Semantically Significant Phrase" Workshop on Data Engineering for Inductive Learning, IJCAI-1995, Montreal, Canada, Aug. 20 1995. Also available at http://ai.iit.nrc.ca/DEIL/krulwich.ps.Z

[4] AltaVista, http://altavista.digital.com/

[5] Michael Mc Elligoot and Humphrey Sorensen, "An Evolutionary Connectionist Approach to Personal Information Filtering" Proc. Fourth Irish Neural Network Conference, pp. 141-146, Sept. 1994. Also available at http://odyssey.ucc.ie/pub/filtering/INNC94.ps

[6] C. Lin and H. Chen, "An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents" http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html, July 5, 1994.

[7] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese" Computational Linguistics, Vol.22, No. 3, pp.376-404, 1996.

[8] Burgin, R., Dillon, M. "Improving Disambiguation in FASIT," Journal of American Society for Information Science, 43(2), 1992, 101-114.

[9] Fagan, J. L. "The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval," Journal of American Society for Information Science, 40(2), 1989, 115-132.

[10] Jones, L. P., Gassie, E. W., & Radhakrishnan, S. "INDEX: The Statistical Basis for an Automatic Conceptual Phrase-indexing System," Journal of American Society for Information Science, 41(2), 1990, 87-98.

[11] Paijmans, H, "Comparing the Document Representation of Two IR Systems: CLARIT and TOPIC," Journal of American Society for Information Science, 44(7), 1993, 383-392.

[12] Zimin Wu and Gwyneth Tseng, "ACTS: An Automatic Chinese Text

Segmentation System for Full Text Retrieval," Journal of American Society for Information Science, 46(2), 1995, 83-96.

[13] 陳光華，"資訊檢索查詢之自然語言處理"，中國圖書館學會會報，第 57 期，85 年 12 月，頁 141 - 153 。

[14] Antti Arppe, "Term Extraction from Unrestricted Text," http://www.lingsoft.fi/doc/nptool/term-extraction.html, 1995.

[15] Jean Godby, "Two Techniques for the Identification of Phrases in Full Text," http://www.oclc.org/oclc/research/publications/review94/part1/twotech.htm .

[16] Jen-Nan Chen, Jyun-Sheng, Chang and Huey-Chyun Chen, "Using Word Segmentation Model for Compression of Chinese Text" http:// nlplab.cs.nhtu.edu.tw/~mathis/own/html/PAPER/JNL/95/cpcol/ CPCOL95.htm

[17] Mathis H. C. Chen, Tsong-Yi Tseng, Jason J. S. Chang, "Automatic Generation of Indices for Chinese Books," http://nlplab.cs.nthu.edu.tw/~mathis/own/html/ PAPER/JNL/96/cpcol/BookIdx.htm

[18] 簡立峰，"尋易系統（Csmart）與中文智慧型資訊檢索"，資訊傳播與圖書館學， 3 卷 2 期， 85 年 12 月，頁 28-37。

[19] Lee-Feng Chien, "PAT-Tree Based Keyword Extraction for Chinese Information Retrieval" ACM SIGIR 1997.

[20] William B. Frakes and Ricardo Baeza-Yates, *Information Retrieval: Data Structure and Algorithms*, Prentice Hall, 1992.

[21] 曾元顯，"新一代資訊檢索技術在圖書館 OPAC 系統的應用"，大學圖書館，1 卷 3 期，86 年 7 月。http://www.lins.fju.edu.tw/~tseng/papers/iropac/

[22] Yuen-Hsien Tseng, "Fast Keyword Extraction of Chinese Documents in a Web Environment," Information Retrieval Workshop for Asia Languages - 1997, Oct. 8-9, Japan, pp. 81-87.

[23] Harman, D. "Overview of the third Text REtrieval Conference (TREC-3)" Proceedings of the Third Text Retrieval Conference, 1994, pp.1-19.

[24] Padmini Srinivasan, "Query Expansion and MEDLINE" Information Processing & Management, Vol. 32, No. 4, 1996, pp. 431-443.

[25] 輔大書目資料檢索系統，http://xlib.fju.edu.tw/ 。