中文書後索引自動產生之研究
Automatic Generation of Indexes for Chinese Books

張俊盛　　曾聰義

台灣30043新竹市光復路
清華大學
資訊科學研究所

截至目前爲止，索引大都由人工產生，中文、英文書籍都是如此。在這篇論文中，我們將研究利用自然語言與統計的技術，來研究產生中文索引的可行性。實驗結果顯示，名詞組、動詞組分析，配合簡單的統計分析，具有不錯的索引產生的效率：召回率與正確率約爲63%左右，精確的召回率與正確率約爲37%左右，比起斷詞與統計分析的作法，可以有25%以上的提升。另外以比較低的統計量臨界值0.1，來篩選索引的時候，粗略召回率可達到87%，精確召回率也可達到約50%左右。這樣的技術，應該可以用來發展一個電腦輔助的索引產生系統。

The paper reports on a new approach to automatic generation of back-of-book indexes for Chinese books. Parsing on the level of complete sentential analysis is avoided because of the inefficiency and unavailability of a Chinese Grammar with enough coverage. Instead, fundamental analysis particular to Chinese text called word segmentation is performed to break up characters into a sequence of lexical units equivalent to words in English. The sequence of words then goes through part-of-speech tagging and noun phrase analysis. All these analyses are done using a corpus-based statistical algorithm. After these processes, the system produces candidates for back-of-book indexes which is finalized using common statistical analyses in IR research. Experimental results have shown satisfactory results.

關鍵字：中文自動索引 書後索引 自然語言處理 統計分析 搭配 斷詞 詞性標示 名詞片語標示

Keywords: automatic indexing, POS tagging, NP parser, collocation

中文書後索引自動產生之研究

張俊盛　曾聰義

台灣 30043 新竹市光復路
清華大學
資訊科學研究所

1. 動機與目的

在這個知識爆炸的時代，如何迅速有效地獲取知識是一件很重要的事，而書籍是獲得知識的重要來源之一；但是現代出版事業發達，書籍汗牛充棟，無法全部詳讀，只能針對知識的需求做選擇性的閱讀，而索引是提供選擇性閱讀及參考的工具之一。截至目前爲止，索引大都由人工產生，中文、英文書籍都是如此。在這篇論文中，我們將研究利用自然語言與統計的技術，來發展一個以電腦輔助人工產生中文索引系統的可行性。

2. 傳統作法

比較完整實際的自動索引的作法大致可分成兩類[1]，第一類是運用統計學的觀念，強調索引會具有「中等頻率」的特性，而且會集中分布在文件的某些部份上；第二類除了利用上述的作法外，還加上一些自然語言處理的技術。以下分別介紹。

2.1 機率式作法

最早提出機率式作法的是H. P. Luhn [2]，其方法如下：(1)先統計文件中所有詞的頻率。(2)去掉高頻率與低頻率的詞。(3)使用剩下的所謂中等頻率的詞，對每個句子計算其績分(score)。(4)取文件中具有中等頻率的詞且詞所在的句子有高績分的，當作索引。在這個作法中，有許多並不分明確的地方。具有中等頻率的詞也未必全可當作索引。

Sparck Jones 針對上述的缺點加以改進，並提出一個公式，做爲在文件中的每個索引候選項的評估標準：[3]

$$\frac{\text{在某一文件中出現的頻率}}{\text{在所有文件中出現的頻率}}$$

機率式自動索引只利用頻率來決定索引，有以下幾個缺點：

(1) 頻率並不能決定一切，並不是所有中等頻率且出現集中的詞都可以當作索引。例如介系詞通常出現頻率很高而且分佈分散，但由於作者寫作習慣或多人創作的因素，有時候也會有中等頻率且出現集中的特性。另外像書中的例題或引用別人的著作，因其用詞通常與書中的其他部分不同，也會造成中等頻率且出現集中的假象，使頻率分析發生錯誤。

(2) 由於沒有文法的觀念，無法事先過濾掉一些不可能做爲索引的詞，如介系詞、 連接詞等，增加系統計算的負擔。

(3) 索引通常不會僅由一個詞所構成，而是由詞組或片語（名詞片語、動詞片語等）構成，但是機率式作法欠缺文法觀念，很難正確地找出片語或詞組。

2.2 結合機率與語言模式的作法

Salton 提出一種結合機率與語言模式的作法[4]。先以剖析器對句子的文法結構進行分析，產生一簡單的剖析樹；再從剖析樹中挑選出適合做爲索引的片語或詞組，如下列幾種組合：(1)名詞 + 名詞 (2)形容詞 + 名詞 (3)過去分詞 + 名詞，或現在分詞 + 名詞。爲了增加正確性，他提出下列幾項附加原則：

(1) 利用非索引用字表或停止表(stop list)，來過濾文件中明確的虛字(Function words)。

(2) 將每一個詞都轉換成它的stem form，所有的統計與計算均以 stem form 進行。例如：computing, computed, computation 這三個詞都應轉換成 comput。

Salton 又以資訊檢索中常用的權重(weighting)的公式，做爲選擇索引的原則：

$$w = tf * \log(N/df)$$

其中，tf (term frequency) 是指某一文件中，某一個片語或詞組出現的頻率；df (document frequency) 則指某一個片語或詞組出現的文件數；w代表某一片語或詞組的權重，N則是所有的文件數。做爲索引的片語或詞組必須具有高的term frequency與相對低的document frequency。

Salton 的作法是這幾個作法中最完整的，但仍會面臨到以下幾個問題：

(1) 以目前的自然語言處理的技術而言，對無限制文章的句法剖析並不完全，就英文而言，根據 Grishman 的研究[11]，大約只有 70%左右的句子能剖析完全。在這種情況下，對剖析不完全的句子要如何處理，也是一個問題。

(2) 利用剖析器來分析句子的文法結構，剖析器所發生的錯誤將會遺留下來，使索引也發生錯誤；若選擇的剖析器太差，會嚴重影響到索引的正確性。

(3) 選擇索引的統計公式仍無法攬括所有的索引；也就是說，還是有些索引不具「高的 term frequency 與相對低的 document frequency」的特性。

3. 進行方式

由於這是一個相當新的研究題目，缺乏相關的文獻可供參考，所以我們先進行一些簡單的實驗，一方面做爲正式實驗的比較依據，另一方面也試探正式實驗應該走的方向。結果我們發覺用語言分析過濾非關鍵字，再用頻率分析加強並選取我們需要的索引，並且利用搭配來加大索引的結果，似乎是一條可行的路。另外在實驗中，我們也發現一些缺陷可供正式實驗時做改進。

3.1 準備工作

我們的實驗材料是 Lotus 1-2-3 for Windows 中文版的機讀資料與其人工索引。我們用電腦產生的索引與人工索引進行比對，但由於人工索引與電腦索引在性質上畢竟有差異，不利我們進行比對。所以我們要先對人工索引進行修改，我們所依據的原則是：

(1) 刪除階層結構。若有索引項是做說明之用或牽涉語意分析，直接刪除；否則就參考正文，將階層結構轉變成平列式的結構。

(2) 依據人工索引所提供的頁碼檢查該索引項是否出現在正文中，若完全沒有出現則刪除該索引項；若只有索引項中的部份文字出現，則更正該索引項。

(3) 刪除在人工索引中的「另見」的說明項。

(4) 除去重複的索引項。

以上所提到的就是我們修改 Lotus 1-2-3 人工索引的原則。經過這樣的修改，我們共到803 個索引項。另外我們也要對正文做些修改，由於電腦版的資料是中英對照的形式（參考附錄二），爲了處理方便起見，先把英文的部份拿掉。又爲了標明各句子的位置，以便依照詞出現的位置給予不同的權重，我們定義了一串代碼來表明某個句子來自第幾章第幾節第幾個段落。做這樣修改的原因是一方面可做爲頻率分析判別分布情形的依據；另一方面也可以對不同位置的詞給予不同的權重。

最後，我們要有一個評估程式，來比對電腦產生的索引項目與人工索引項目，計算精確率（precision rate) 與召回率 (Recall Rate)，以評估所產生索引的好壞。爲了表現所產生索引的特性，我們將比較的結果分成兩類：

(1) 完全吻合，電腦所產生的索引與人工索引完全相同。

(2) 粗略吻合，但人工索引中的主要關鍵詞彙在電腦所產生的索引中已出現者。

這種分類方法也有利於評估中文自動索引系統是否能完全自動化，或是只適合輔助人工產生索引。如果評估的結果第一類的召回率與正確率令人滿意的話，這個系統就可以完全自動化產生索引；如果第一類的召回率與正確率不高，而第二類的結果可以接受的話，我們可以將這套系統，發展成爲輔助人工產生索引的系統。

根據以上兩類比較結果，再配合精確率與召回率，我們將有四個數值來表示比較的結果：

(1)完全精確率(Perfect Percision Rate)，簡稱爲PP。

(2)完全召回率(Perfect Recall Rate)，簡稱爲PR。

(3)粗略精確率(Rough Precision Rate)，簡稱爲RP。

(4)粗略召回率(Rough Recall Rate)，簡稱爲RR。

四個數值的公式如下：　PP = a/c,　PR = a/m

RP = (a+b)/c,　RR = (a+b)/m

其中　a 是電腦產生的索引中，符合第一類比較結果的索引個數。

b 是符合第二類比較結果的電腦索引個數。

c 是電腦產生的索引總數。

m 則是人工索引的總數。

3.2 語言分析

在這項研究中，語言分析包含三項：斷詞、詞性標示與名詞片語標示。以下分別做簡單的說明。

在中文文件中，詞並不像英文以空白彼此分隔開來，而是許多字連續出現，直到標點符號出現爲止。所以要取出中文詞並不像英文那麼簡單，還要分析句子的文法結構，將最理想的詞的組合找出來，這個過程叫做「斷詞」。目前斷詞的正確率約可達到95%以上[5,6,7]。

詞性標示則是指將斷好詞的句子對每一個詞標上其詞性。由發展的較晚，而且難度又比斷詞高很多，所以目前正確率大約指達到80%左右 [5]。

名詞片語標示則是在做完上述兩種分析後將句子中的名詞片語標示出來。所謂名詞片語是由名詞、專有名詞或代名詞爲主體所形成的詞組，除了名詞、專有名詞與代名詞以外，還可包含數詞、量詞、定詞和形容詞等詞性 [9,10,11]。由於名詞片語標示受到上述兩種分析的影響很大，在斷詞或詞性標示產生錯誤的情況下，名詞片語標示就幾乎不可能標示正確，所以其正確率約只有 76%左右。

3.3 頻率分析

頻率分析是一種利用統計技術來選取關鍵詞的方法。主要的原理是因爲發現大部份的關鍵詞都有「出現頻率高」與「分布集中」的特性。像下列的公式就是根據這個原則來的。

$$w = tf * \log(N/df)$$

其中，tf (term frequency) 是指某一文件中，某一個片語或詞組出現的頻率 df (document frequency) 則指某一個片語或詞組出現的文件數；N則是所有的文件的數目。我們選取關鍵詞所用的頻率分析也是根據這個公式，另外我們還對章名、節名等重要的文字做加權。

3.4 搭配

搭配是指在語料中去統計兩個相鄰或相距一定單位的詞 A 與 B 的出現次數 W，如果 W 這個值很大的話，則表示這 A 與 B 常常會在一起出現 [8]。我們則嘗試用搭配來加大電腦索引的結構。因爲我們發覺索引常常不是一個名詞片語組成的，而進行標示比名詞片語更大結構又有困難存在，所以我們用搭配來完成這個目的。

4. 實驗

接著我們說明，利用前面所提到的幾種方法來進行實驗的情形與結果。我們也將其結果與（修改過）人工索引進行比對，以了解實驗的效果。

4.1 實驗方法

我們先對標示過代碼的正文進行語言分析，也就是說文件資料先要經過斷詞、詞性標示、名詞片語標示等語言分析的過程，以找出文件中的名詞片語結構。在文件經過語言分析取出名詞片語後，我們利用頻率分析來過濾不適當的名詞片語，選出比較關鍵性的片語來進行進一步的分析。接著我們利用搭配來對相鄰或相距一定距離的名詞片語做組合的工作，以產生比較大的結構。由搭配所產生的索引候選項再經過一次頻率分析，產生系統認定的索引。另外對於名詞片語中含有數詞、量詞、定詞等詞性的問題，由於這些名詞片語大都有一定的樣式，所以我們直接在取出名詞片語的地方做檢查，凡是有固定樣式的名詞片語，就把數詞、量詞、定詞等詞性去掉。

我們認爲像1-2-3手冊中的巨集指令與函數，這種有固定樣式的索引項，可以用另一種方法直接辨識，而不用經過頻率分析。因此我們製作了一個具有通用字元 (Wild Character) * 和 ？的辨識機構，* 可以代表任意數目的某些字元，？則代表任意字元，使用者可以輸入含有 * 和 ？的表示式，來選取一些固定樣式的索引項。

4.2 實驗結果

表一是實驗的結果，對於實驗的結果，我們要特別注意兩個地方，第一個是當召回率與正確率大約相等時，這時所顯示的是系統的最大效能，我們發覺粗略的召回率與正確率約爲 63%左右，精確的召回率與正確率約爲 37%左右，比起初步實驗的結果，所有的值大約都提升 25%以上。另一個要注意的是臨界值很低的時候，這時所表達的是系統的最大產能。我們可以發現臨界值爲 0.1 的時候，粗略召回率可達到 87%，精確召回率也可達到約 50%左右。

4.3 錯誤分析

我們將錯誤的原因分成三大類分別討論。第一類是以我們目前的程式而言，無法加以處理的索引項。這類的索引項大致可再分成幾種，一種是遞迴的名詞片語，如：

2維的範圍　　比較財務的選項　改變資料的外觀

這類的問題可以將名詞片語標示方法改成遞迴式的來解決，但由於此類的索引項並不多，而且遞迴式名詞片語標示的作法要比非遞迴式作法困難許多，以成本效益而言，並不值得這樣做；或許可以直接由判斷「的」字的存在，來解決這個問題，因爲中文的遞迴式名詞片語大都會有「的」字的存在。

第一類中，另一種無法處理的問題，是索引項的結構超過名詞片語、動詞片語的範圍，如在工作表中加入文字，以公式爲準則，使用相鄰標記替單一儲存格命名等等。這些索引項大都是正文中的標題。雖然我們的程式無法找出這樣的索引項，但是可以找到這些索引項中大部份的名詞片語，解決一部份的問題。還有一種情形是索引項並不是由名詞片語或動詞片語組成的，而是由動詞所組成的，如切換至和剪下。

臨界值	threshold	0.8	0.6	0.55	0.525	0.5	0.3	0.1
人工索引數目	m	803	803	803	803	803	803	803
電腦索引數目	c	615	768	787	790	1359	1624	1952
完全吻合數目	p	260	288	293	295	321	357	393
部分吻合數目	r	162	201	205	205	222	272	309
可接受者數目	p+r	422	489	498	500	543	629	702
粗略召回率	RR=(p+r)/m	52.6%	60.9%	62.0%	62.3%	67.6%	78.3%	87.4%
粗略正確率	RP=(p+r)/c	68.7%	63.7%	63.3%	63.3%	34.0%	38.7%	36.0%
精確召回率	PR=p/m	32.4%	35.9%	36.5%	36.7%	34.0%	44.5%	48.9%
精確正確率	PP=p/c	42.3%	37.5%	37.2%	37.3%	23.6%	22.0%	20.1%

表一 實驗結果

第二大類是語言分析所產生的錯誤，這類的錯誤有時候很難區分到底是由斷詞、詞性標示或是由名詞片語標示所造成的，所以乾脆歸爲一類，一起統計。以下是幾個錯誤（以＊標示）的例子：

|1-2-3|視 |窗|　|.CGM|檔|　　|2|維 |範圍||

[np　] *v [nc] [np　]*cl|　　|q[*nc|nc　]

第三類的錯誤是頻率分析錯誤所引起的。由於頻率分析要求索引項具有「出現頻率高且分布集中」的特性，使得電腦選出的索引項大都是具有中等頻率的，高頻與低頻的詞比較難被挑選出來；而在人工索引中有些常見的術語具有較高的頻率，如1-2-3 巨集和字型。這些術語在文件中幾乎處處可見，出現頻率很高，卻不具有分布集中的特性，所以在頻率分析時被過濾掉，造成錯誤。另外有些人工索引出現頻率卻偏低，如：.DBF，CTRL+END，CTRL+PGDN，系統資訊等。由於出現頻率太低，所以這些索引項也會在頻率分析時被去掉。其實這類的錯誤與臨界值有很大的關係：若將臨界值逐漸調低時，這類索引項就有機會會出現在電腦索引中。

經過統計發現，其中無法處理的錯誤大約佔了20%，語言分析錯誤大約佔了40%，頻率分析錯誤也大約佔了40%。下表是統計的結果。

錯誤類型	無法處理錯誤	語言分析錯誤	頻率分析錯誤
錯誤個數	97	211	200
錯誤比率	19.09%	41.54%	39.37%

表二 錯誤分析的結果

5. 結論

由於中文自然語言的研究還不到十分成熟的階段，中文自動索引更是幾乎沒有人從事這方面的研究，所以這篇論文是以一些比較基本的想法來研究中文自動索引的問題，主要目的是觀察這些基本想法的效率與建立一套基本的模組。我們從實驗結果可以得到幾點結論：

(1)中文自動索引是相當複雜的事，由於索引的成份包羅萬象，不是純粹由名詞或名詞片語所形成，有時還會牽涉到語意因素，所以不適合用單一方法來解決。

(2)文法剖析有助於找尋索引的基本組成單元，但不容易找到適當且完整的索引項，仍需其他方法的配合。

(3)搭配適合組合兩個單元成一個更大的單元，但是由於搭配不具有文法觀念，可能使產生的大單元不合需要。

(4)從實驗中可以發現，文法剖析只做到名詞、動詞片語的階段仍嫌不足，因爲索引中還是會出現少數超過名詞、動詞片語的結構。但是中文文法剖析要再進一步 分析更大的結構並不容易，錯誤也會大量出現，是否要如此做，值得商榷。

(5)由於中文文法剖析仍未成熟，所以文法剖析的錯誤比率仍嫌偏高。但是從實驗中我們發現簡單的語言分析仍不失爲一個好方法。

(6)頻率分析的錯誤也很多，這可能是我們使用的統計模組太簡單的緣故。但是頻率分析是一個相當基礎有效的方法，所以我們應該繼續改進。

6. 參考資料

[1] Susan Jones, "Text and Context : Document Processing and Storage", Springer-Verlag, London, 1991.

[2] H. P. Luhn, "A Statistical Approach to the Mechanized Encoding and Searching of Literary Information", IBM Journal of Research and Development, 1:4, October 1957, pp. 309-317.

[3] K. Sparck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", Journal of Documentation, vol. 28, no. 1, pp.11-21, March 1972.

[4] Gerard Salton, "Syntactic Approaches to Automatic Book Indexing", The Proceedings of ACL 88', 1988, pp 204-210.

[5] 彭載衍，"中文辭彙岐義之解決—斷詞與詞性標示"，清華大學碩士論文，臺灣新竹，1993.

[6] Lee, H. J. et al. Rule-Based Word Identification for Mandarin Chinese Sentences - A unification Approach . Computer Processing of Chinese and Oriental Languages . Vol 5 , no 2, pp. 97-118 , March 1991.

[7] T. H. Chiang,T. S. Chang,M.Y. Lin,and K. Y. Su,1992, Statistival Models for Word Segmentation and Unknown Word Resolution , ROCLING V , pp.147-175.

[8] Frank A. Smadja, "From N-gram to Collocations-An Evaluation of XTRACT", ACL, 1991, pp279-284.

[9] 盧士仁，"英文介詞組連繫問題之統計式作法"，清華大學碩士論文，臺灣新竹，1992, pp23-25.

[10] Kenneth Ward Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", ANLP 88', 1988, pp136-143.

[11] Grishman R., Macleod C., Sterling J. "Evaluating Parsing Strategies Using Standardized Parse Files", ANLP 92", 1992, pp156-161.