中文新聞之關聯詞推薦

黃謙順　　曾祺淋
文化大學資管所副教授　文化大學資管所研究生
cshwang@faculty.pccu.edu.tw　92707157@scenet.pccu.edu.tw

摘要

因為網際網路使用率越來越大眾化，以及科技技術的進步，資訊傳遞越來越容易，資訊量也日益增加，在大量的資訊中，使用者越來越困難找到有用的資訊。

目前大部份資訊檢索的方式，是透過全文搜尋，搜尋出符合使用者所輸入關鍵詞的資料，查詢的速度非常緩慢，而且使用者如果輸入太廣泛或錯誤的關鍵詞，還是會遺漏掉許多資訊，為了減少這樣的問題，則必需從詞與詞之間的關聯性著手，自動化建構關聯詞典。本研究首先利用類神經網路演算法，自動擷取出切合文意的關鍵詞，再以詞頻反轉頻率的權重計算公式，計算出關鍵詞之間的關聯權重，建構直接或間接關聯的詞典，並且提供使用者瀏覽或檢索的參考。

本研究隨機選取聯合新聞網的 500 篇新聞文件，實驗結果顯示平均每篇新聞有 3 個是符合人工定義關鍵詞，其餘雖然不符合人工定義，但約六成可被使用者所接受。對每篇文章，我們擷取前 20 個關鍵詞做為代表詞彙，來計算其關聯性。

關鍵詞：關鍵詞、類神經網路、倒傳遞模型、間接關聯詞。

1. 緒論

由於網際網路的蓬勃發展，以及資訊科技日益進步，有越來越多的資訊在網路上傳遞，對使用者來說，要取得資訊是非常容易及方便的，但是如何在大量的資訊中，擷取有用的資訊是一個相當重要的課題。

目前各大網站提供的搜尋方式，大部份是提供關鍵詞進行搜尋，例如使用者想找和汽車產業的相關資訊，會輸入關鍵詞“汽車”，即會出現文章中內含“汽車”關鍵詞的文章，如果使用者有興趣其中一篇文章，會直接點選及閱讀該篇文章，如果該篇文章並不是使用者想要的，則必須重新再查詢其他關鍵詞的文章。雖然這個方法已經解決使用者不需要一篇篇文章瀏覽，即可找出需要的資訊，但是如果使用者輸入太廣泛或錯誤的關鍵詞，還是無法幫助使用者，找到潛在有用的資訊。

一篇文章是由許多詞彙所組成的，在這些詞彙集合中有一些重要詞彙是可以被擷取出來，形成短文摘要或是文章的索引詞，這些重要的詞彙，一般稱之為「關鍵詞」，也就是全文重點關鍵的詞彙。

關鍵詞之間的關聯性判定，是以二詞共同出現的頻率為主，例如關鍵詞“裕隆”和“RFID”共同出現在文件中的頻率很高，所以可判斷其有直接關聯；而“永豐餘”和“RFID”同時出現的頻率也很高，判斷也有直接關聯，則“裕隆”、“RFID”及“永豐餘”這三個關鍵詞有關聯。因此“裕隆”及“永豐餘”有間接關聯，使用者可透過關鍵詞的不同關聯，找到同領域卻不同方向的資訊。

但是在大量的文字之中，人工擷取出關鍵詞及關聯詞是非常耗費時間及人力，因此本研究提出自動化擷取出關鍵詞，以及自動建構關聯詞彙庫。當使用者搜尋資訊時，除了會提供文章中切合文意的關鍵詞，同時也推薦有直接或間接相關的關聯詞。

本研究利用類神經網路的演算法，建立自動擷取出關鍵詞的訓練模型，再以詞頻反轉頻率的權重計算公式，計算出關鍵詞之間的關聯權重，取得直接及間接關聯詞。

2. 文獻探討

2.1 關鍵詞擷取

關鍵詞是文章有意義的最小組成單位，大部份的文件自動化處理，例如自動摘要、自動索引及自動分類等，都會先做關鍵詞擷取動作，再進行後續處理。可以說，關鍵詞擷取是所有文件自動處理的基礎與核心技術。

關鍵詞擷取的方法，可大略分為統計法、詞庫法、規則法或這三種方法的合併運用。在過去的文獻中，一般關鍵詞擷取的技術可以區分成三大類型 [13] [14]。

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），擷取文件中出現符合詞庫中的片語。

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。

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，再將統計參數符合門檻值的片語擷取出來。

其他的方法還包括上述方法的綜合運用，或運用不同的演算法。例如 Krulwich, B. and Burkey, C.

[3] 為了文章自動分類，利用經驗法則演算法，從文章中擷取出關鍵詞，作為分類的特徵值，然而其實驗結果卻擷取出大量且低精確度的關鍵詞。

Muñoz, A. [4] 提出無監督式學習方法來擷取二個字的關鍵詞，採用自適應共振理論網路(Adaptive Resonance TheoryNetwork, ART)，其結果也是擷取出大量且低精確度的關鍵詞。Steier, A. M., and Belew, R. K. [8]使用相互訊息函數 (Mutual information) 來計算關鍵詞特徵值，但其方法只能接受二個字的關鍵詞。

Turney, P.D. [9][10] 提出 Genex 架構，主要以遺傳基因演算法 (GeneticAlgorithm ,GA)擷取關鍵詞，結果平均每篇文章擷取出二個關鍵詞。 Witten, I.H., Paynter, G.W., Frank, E., Gutwin[12] 提出一個 Kea 實作架構，使用貝式(Bayesian)演算法，此演算法在 Turney, P.D.[11]中實驗證明出 Kea 和 Genex 有大約相等的效率。

2.2 詞彙權重計算

在資訊儲存與檢索的範疇而言，索引辭典是記錄詞彙之間階層或語意的關係，做為使用者檢索資料時，可透過索引辭典推薦相似概念的字或詞。

一般索引辭典是記錄同義詞，還有反義詞、廣義詞、狹義詞、相關詞等，用以擴展或縮小檢索詞彙的主題範圍。詞意相關的索引辭典，必須由人力維護，當文章數量越來越多，則要有更多的人力及時間才能維護索引辭典。為了能夠有效及自動化建立索引辭典，我們以詞彙共同出現的關係，來做為詞彙之間的關聯。

Aas, K. and Eikvil, L.[1]整理出各種不同詞彙權重的計算公式，利用詞彙權重來評估文章的相似度，如下述：

N 是指文件的總數目，M 是指斷詞後的詞彙總數，n_i 是指詞彙 i 出現的文件數。

(1) Boolean：最簡單的計算方式，如果該詞彙出現在這篇文章，則權重值為 1，反之則為 0。

f_{ik} 是指詞彙 i 出現在文章 k 中的次數。

$$w_{ik} = \begin{cases} 1 & \text{if } f_{ik} > 0 \\ 0 & \text{otherwise} \end{cases}$$

(2) word frequency weighting：詞彙 i 出現在文章 k 中的次數。

$$w_{ik} = f_{ik}$$

(3) TF×IDF weighting：詞頻反轉文件頻率。

$$w_{ik} = f_{ik} * \log(\frac{N}{n_i})$$

(4) tfc-weighting：考慮到不同的文章長度，計算該詞的詞頻反轉文件頻率在文章中的比例。

$$w_{ik} = \frac{f_{ik} * \log(\frac{N}{n_i})}{\sqrt{\sum_{j=1}^{M}[f_{j_k} * \log(\frac{N}{n_j})]^2}}$$

(5) ltc-weighting：類似 tfc-weighting，為避免高詞頻的影響，而調整詞頻。

$$w_{ik} = \frac{\log(f_{ik} + 1.0) * \log(\frac{N}{n_i})}{\sqrt{\sum_{j=1}^{M}[\log(f_{jk} + 1.0) + \log(\frac{N}{n_j})]^2}}$$

(6) Entropy：熵權重，是一個複雜的權重計算方法，。

$$w_{ik} = \log(f_{ik} + 1.0) * \left((1 + \frac{1}{\log(N)}) \sum_{j=1}^{N} \frac{f_{ij}}{n_i} \log(\frac{f_{ij}}{n_i}) \right)$$

前述所提出的詞彙權重計算公式，主要是做為文章分類時，計算文章的相似度。本研究利用這些詞彙權重公式，來計算詞彙間在中文新聞文章中的相似度。

2.3 類神經網路

類神經網路是一種模擬生物神經系統的處理系統。生物神經系統由許多神經元相互連結，而每個神經元都有輸出及輸入訊號和其他神經元相連及傳遞消息。目前類神經網路可分為下類三種：

第一種監督式學習網路，從問題領域中提供訓練範例，包含輸入資料及輸出資料。並且從網路中學習輸入資料與輸出資料的內在對映規則。常應用於預測或分類上。例如債券分級[2]和破產預測[6]。本研究所使用的倒傳遞神經網路(Back -Propagation Network) 即是屬於此類型。

第二種無監督式學習網路，從問題領域中取得只有輸入資料的訓練範例，並從網路中學習輸入資料的內在聚類規則，以應用於新的案例。例如自組織映射圖網路(Self-Organizing Map，SOM)、自適應共振理論網路 (Adaptive Resonance Theory Network，ART) 。

第三種聯想式學習網路，以狀態變數值為訓練範例，學習範例中的記憶規則，然後應用於只有不完整狀態值，而需推論完整狀態的新案例，這種網路可以應用於擷取應用與雜訊過濾。例如霍普非爾網路(Hopfield Neural Network)以及雙向記憶網路(Bi-directional Associative Memory)等屬之。

3. 系統架構

本研究系統架構分成二個模組，第一模組是自動化擷取出關鍵詞彙，第二模組是利用上一個模組的關鍵詞彙，計算詞彙間的相似度，建立關聯詞典。

3.1 自動化關鍵詞擷取

關鍵詞擷取步驟描述如下：

(1) 建立訓練模型：先由人工方式來定義訓練文件的關鍵詞，再利用定義出的關鍵詞來建立一個訓練模型。相關流程請見圖一。

(2) 擷取關鍵詞：利用上一步驟的模型，擷取出測試文件中的關鍵詞。相關流程請見圖二。

表 1 詞性規則的過濾

1.	名詞
2.	形容詞＋名詞
3.	名詞＋名詞
4.	名詞＋動詞
5.	動詞＋名詞

中文斷詞是利用中研院的中文斷詞系統[5]來做斷詞，再將斷出的詞彙透過詞性規則過濾，擷取出候選詞，如表 1。

上述作業擷取出來的候選詞，必須再計算三項特徵值，做為訓練資料的特徵，如表 2。這三項訓練特徵值的描述如下：

(1) 詞彙出現的權重：本因素考慮詞彙在每一篇文章上出現的位置不同，而有不同的重要性，所以詞彙出現不同的位置，則設有不同的權重。例如出現在抬頭，則權重為 w_1，出現在第一段，則權重為 w_2，其他地方的權重為 w_3，計算方法如公式(1)所示。其中 f_{ik}^1 是詞彙 i 在文件 k 中，出現在抬頭的詞頻，f_{ik}^2 是詞彙 i 在文件 k 中，出現在第一段的詞頻，f_{ik}^3 是詞彙 i 在文件 k 中，出現在其他位置的詞頻。

$$PW_{ik} = f_{ik}^1 \times w_1 + f_{ik}^2 \times w_2 + f_{ik}^3 \times w_3 \quad (1)$$

(2) 相對詞長：表示詞彙長度除以文章中所有詞彙的平均長度。

(3) TF×IDF：詞頻反轉文件頻率，有兩個基本假設：一個詞出現在一份文件中次數越多則越重要；若在所有蒐集文件中出現次數越多則越不重要，因為表示這詞無法代表這份文件的特性，其計算方法如公式(2)所示，其中 f_{ik} 是詞彙 i 在文件 k 的詞頻，N 為總文件數，n_i 為至少出現一次詞彙 i 的文件數。

$$TFIDF_{ik} = f_{ik} * \log(\frac{N}{n_i}) \quad (2)$$

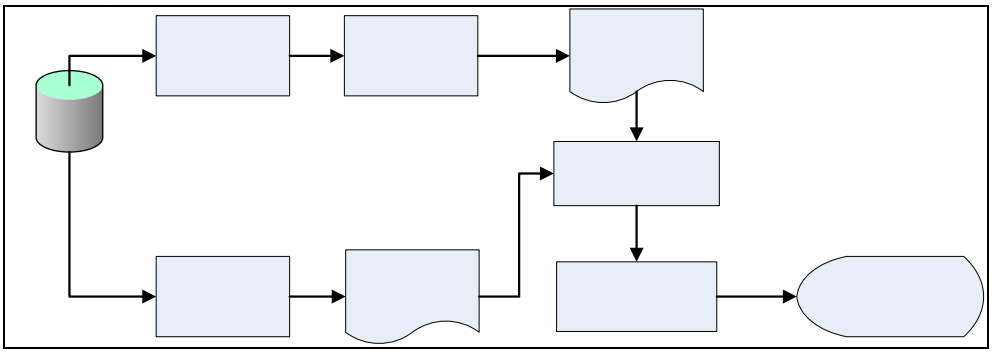

圖 1 建立訓練模型

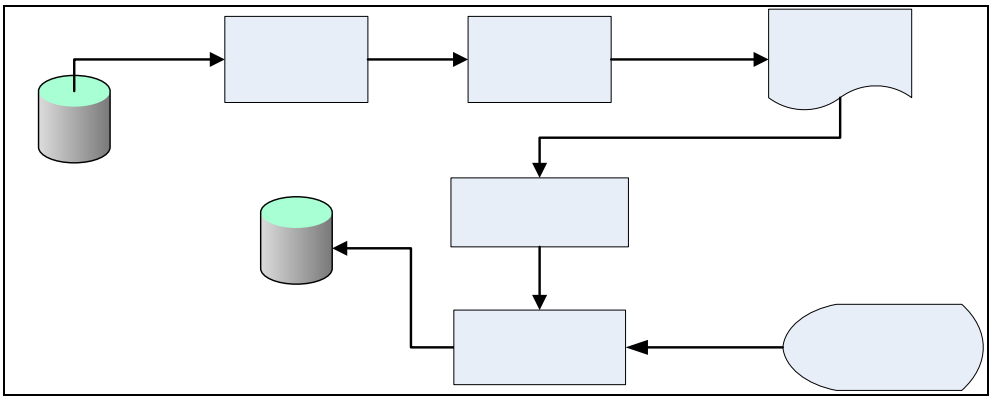

圖 2 擷取關鍵詞

表 2 訓練資料的特徵

特徵名稱	描述
詞頻	詞彙出現在抬頭、首段及其他位置的詞頻加權總和。
相對詞長	詞彙的長度除以文章中的所有詞彙平均長度。
TF×IDF	詞彙的 TF×IDF 值。

同時由專家人工定義每篇新聞的關鍵詞，再將之前中文斷詞系統擷取出來的候選詞彙，過濾掉長度小於 2 的詞彙，計算上述訓練特徵，加上是否為關鍵詞，整理成訓練資料，透過類神經網路的倒傳遞方法訓練資料，建立出關鍵詞的訓練模型。

第二步驟是以其他新聞文件作測試，先利用中文斷詞系統擷取出候選詞，計算每個詞彙的上述三項特徵，利用第一步驟的關鍵詞訓練模型，來擷取出關鍵詞。

3.2 關聯詞推薦

關聯詞推薦步驟描述如下：

(1) 計算詞彙權重：將擷取出的關鍵詞，計算其在每篇文章的權重，建立詞彙權重檔。

(2) 計算詞彙相似度：將上述步驟的詞彙權重檔，計算兩兩詞彙在測試文章中的相似度，擷取出有關聯的詞彙。

上述相關流程請見圖 3。

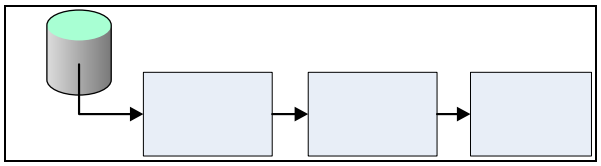

圖 3 關鍵詞推薦

新聞的內容並不長，以及詞頻也不高，所以本研究採用 Aas,K. and Eikvil, L.[1] 所提出的 tfc-weighting 方法，計算各詞彙在測試新聞文章中的權重，計算公式如(3)，N 是指文件的總數目，M 是指斷詞後的詞彙總數，n_i 是指詞彙 i 出現的文件數。

$$w_{ik} = \frac{f_{ik} * \log(\frac{N}{n_i})}{\sqrt{\sum_{j=1}^{M}[f_{j_k} * \log(\frac{N}{n_j})]^2}} \qquad (3)$$

計算完詞彙權重後，本研究以向量空間來表示文章及詞彙所構成的二維度空間，例如有 n 篇新聞文章，及所有文章中總共出現有 m 個詞彙，則建立 $n*m$ 的二維度空間，如圖 4。

以二維度空間來計算詞詞彙間的關聯性，關聯性測量方式以 Salton,Gerard [7]提出的向量關聯性計算公式，如(4) ，w_{ik} 是指詞彙 i 在文章 k 中的權重。

$$sim(T_i, T_j) = \sum_{k=1}^{n} w_{ik} \times w_{jk} \qquad (4)$$

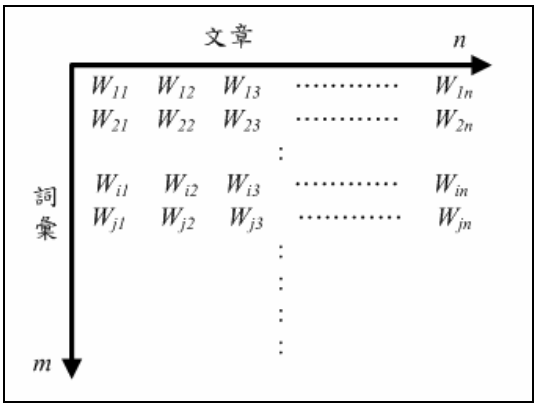

圖 4 向量空間

4. 實作結果

本研究以聯合新聞網 11~12 月的財經新聞，隨機選取 500 篇做為研究對象，分成 400 篇為訓練文件及 100 篇為測試文件。

本研究中的第一項訓練特徵，詞彙出現的權重，為加重出現在抬頭或首段的關鍵詞的權重，設定出現在抬頭，則權重為 2，出現在第一段，則權重為 1.5，其他地方的權重為 1。

有關自動擷取關鍵詞，本研究採取二種方式來評估模型成效，第一種是 Ian H. Witten 等人在 Kea 模型中所提出的方法，從文章中擷取出關鍵詞，計算有多少是符合人工定義的關鍵詞，其主要原因如下：

(1) 此方法比用精確率及召回率更容易被使用者理解。

(2) 精確率及召回率可能會誤導使用者，為了追求高精確率而犧牲了召回率，或追求高召回率而犧牲了精確率。

(3) 本方法符合使用者常以文章所擷取出關鍵詞數量來衡量。

表 3 就是以每篇測試文章取出前 5、10、15、20 個關鍵詞，統計有多少個是符合人工定義的關鍵詞，並計算每篇實際符合的關鍵詞數量。Kea 模型以英文期刊為研究對象，擷取出每篇的前 5、10、

15、20 個詞彙中平均有 0.93、1.39、1.68、1.88 個符合人工定義的關鍵詞，雖然和本研究的測試文件的語言及內容不同，無法客觀的比較，但是本研究模型的實驗結果顯示比 Kea 的較佳。

表 3 關鍵詞數量

符合人工定義	擷取關鍵詞數量			
	5	10	15	20
平均	1.98	2.7	2.99	3.1

第二種則是以最常見的精確率(Precision)及召回率(Recall)來評估，結果請見圖 5，統計訓練出實際符合關鍵詞的精確率及召回率，結果表示出在擷取詞彙數量越多，則召回率越高，精確率越低。

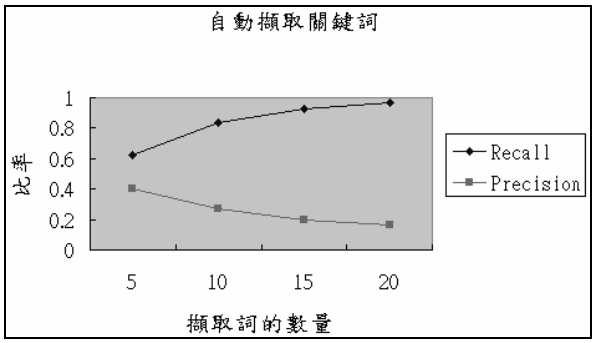

圖 5 自動擷取關鍵詞的精確率及召回率

目前 500 篇新聞分給 30 個有財經背景的研究生評估，大約有 58%的關鍵詞可接受。

訓練模型所擷取出的關鍵詞，在前 20 個關鍵詞的召回率高達 0.97，因此本研究採取每篇文章的前 20 個關鍵詞，來計算其關聯性。

依據詞彙之間的關聯性，取出關聯權重高於 0.5 的直接及間接關聯詞，如表 4，表 4 中的“乙烯”是生產原料，和“石化業”、“石化產業”及“大陸石化”有直接關係，而“石化業”一詞又對應到“輕油裂解廠”、“投資計畫”、“年產能”，也可以說是“乙烯”及“輕油裂解廠”在中文新聞中有間接關係。

表 4 直接及間接關聯詞

查詢詞彙	關聯詞		權重	
	直接	間接	直接	間接
乙烯	石化業	輕油裂解廠	0.89	0.99
乙烯	石化業	投資計畫	0.89	0.73
乙烯	石化業	年產能	0.89	0.73
乙烯	石化產業	聯合採購	0.75	0.89
乙烯	大陸石化	兩岸石化業	0.63	0.75
人民幣升值預期	外匯存底	熱錢流入	0.66	0.72
人民幣貸款	台塑寧波三期	法國巴黎銀行	1.00	1.00
人民幣貸款	台塑寧波三期	聯貸案	1.00	0.87
人民幣貸款	台塑寧波三期	融資	1.00	0.57
人民幣貸款	中資銀行	台商融資	0.67	0.67
人民幣貸款	中資銀行	吸收外匯存款	0.67	0.67
人民幣貸款	中資銀行	外匯存款準備金	0.67	0.54
人民幣貸款	中資銀行	聯貸案	0.67	0.59
中華電信	通訊業者	道路使用費	0.58	0.67
中華電信	高速公路	遠東電子	0.57	0.75
中華電信	高速公路	電子收費系統	0.57	0.75
台西區	台塑集團	雲林離島新興區	0.76	0.59
台西區	台塑集團	新日鐵	0.76	0.55
台西區	台塑集團	嘉義縣政府	0.76	0.77
台西區	台塑集團	養生文化村	0.76	0.78
台西區	煉鋼廠	彰化大城海埔地	0.55	0.83
台塑集團	新日鐵	鋼鐵廠	0.55	0.96

5. 結論

本研究的實作結果，得知以下結論：

(1) 中文新聞文章內容不長，關鍵詞的特徵不明顯，較不容易擷取出來，在本研究雖然有較高的召回率，卻是低精確率。

(2) 本研究所擷取出來的雖然不是都符合人工定義關鍵詞，但每篇文章至少有六成的詞彙可以被使用者接受。

(3) 本研究以詞彙之間共同出現的頻率，計算詞彙關聯性，讓使用者在短時間即可瀏覽到中文新聞的簡易索引及關聯。沒有和查詢詞彙共同出現在文章中，但是卻和查詢詞彙的直接關聯詞共同出現在文章中，也可以作為使用者查詢的參考。

6. 未來研究方向

本研究所擷取出的關鍵詞及關聯詞，未來會繼續研究如何應用在新聞文件檢索的功能。建立出新聞文件關鍵詞彙索引，以圖示顯示不同關鍵詞之間的關係，協助使用者快速查詢所需要的新聞文件。

參考文獻

[1] Aas, K., Eikvil, L.: Text Categorisation: A Survey. Norwegian Computing Center, Oslo 1999 .

[2] Dutta, S. and Shekhar, S., Bond rating: A non-conservative application of neural networks, IEEE International Conference on Neural Networks-San Diego, Vol.2 , pp443-450, 1988.

[3] Krulwich, B., and Burkey, C. , Learning user information interests through the extraction of semantically significant phrases. In M. Hearst and H. Hirsh, editors, AAAI 1996 Spring Symposium on Machine Learning in Information Access. California: AAAI Press.

[4] Muñoz, A., Compound key word generation from document databases using a hierarchical clustering ART model. Intelligent Data Analysis, 1 (1), Amsterdam: Elsevier.1996.

[5] Ma, Wei-Yun and Keh-Jiann Chen, Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff, Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171, 2003.

[6] Odom, M, Sharda, R., Aneural network model for bankruptcy prediction,IEEE INNS IJCNN,Vol.2,PP.163-168,1990.

[7] Salton,Gerard , Automatic text processing:the transformation, analysis, and retrieval of information by computer,Addison-wesley publishing Company, Inc,1989.

[8] Steier, A. M., and Belew, R. K. , Exporting phrases: A statistical analysis of topical language. In R. Casey and B. Croft, editors, Second Symposium on Document Analysis and Information Retrieval, pp. 179-190, 1993.

[9] Turney, P.D., Extraction of Keyphrases from Text: Evaluation of Four Algorithms. National Research Council, Institute for Information Technology, Technical Report ERB-1051,1997.

[10] Turney, P.D., Learning to Extract Keyphrases from Text. National Research Council, Institute for Information Technology, Technical Report ERB-1057,1999.

[11] Turney, P.D. Learning algorithms for keyphrase extraction. Information Retrieval, 2, pp.303-336, 2000.

[12] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C. and Nevill-Manning, C.G., KEA: Practical automatic keyphrase extraction. Proceedings of Digital Libraries 99 (DL'99), pp. 254-256. ACM Press,1999.

[13] 曾元顯，關鍵詞自動擷取技術與相關詞回饋，中國圖書館學會會報，1997 年，12 月，第五十九期，頁 59-64。

[14] 曾元顯，關鍵詞自動擷取技術之探討，中國圖書館學會會訊，1997 年，9 月，第 106 期，頁 26-29。