圖書館學刊 第十一期 (民國85年12月) ，頁95-112

文件內容之分析 - 語料庫為本的模型

Content Analysis — A Corpus Based Model

陳 光 華

Kuang-Hua Chen

國立臺灣大學圖書館學系暨研究所講師

Instructor

Department and Graduate Institute of Library Science

National Taiwan University

陳 信 希

Hsin-Hsi Chen

國立臺灣大學資訊工程學系教授

Professor

Department of Computer Science and Information Engineering

National Taiwan University

【摘　要】

一般資訊檢索的研究著重於檢索模型的建構、查詢的回饋機制、檢索行為的探討、檢索系統的執行效能。本文則把研究的重心回歸資訊或文件本身，希望對資訊的內容有一個初步的瞭解。本文根據三個因素：1）詞彙的重複，2）詞彙的重要性，3）共容語意，提出一個基於真實語料的文件內容分析的模型。這樣的模型著重於文章中名詞／動詞與名詞／名詞之間的配對關係。本文也說明如何使用文件分析模型進行文件切分與文件主題辨識的研究，同時討論相關實驗的結果。

【ABSTRACT】

An important step to understand text is to build the discourse structure through cohesion and coherence. However, to build the discourse structure in turn depends on the full understanding of texts, so that many efforts on this line are not automatic and not successful. A corpus-based model based on 1) repetition of words, 2) importance of words, and 3) collocational semantics for texts is proposed in this paper. It focuses on association norms of noun-noun relations and noun-verb relations defined on discourse level and sentence level, respectively. According to this model, a text partition algorithm is proposed to determine the boundaries of discourse structures and a topic identification algorithm is also presented. The results of a series of experiments show that the proposed model is promising.

關鍵詞 Keywords:

言談分析；資訊檢索；自然語言處理

Discourse Analysis; Information Retrieval; Natural Language Processing

一、緒論

一般資訊檢索的研究著重於檢索模型的建構、查詢的回饋機制、檢索行為的探討、檢索系統的執行效能。本文則把研究的重心回歸資訊或文件本身，希望對資訊的內容有一個初步的瞭解。因為這項工作非常困難，少有新的突破，從事這方面研究的學者專家並不多見。本文提出一種基於真實語料訓練而得的模型，希望能夠對這項研究課題有些微的貢獻。

在正常的情形下，文件並非僅僅是一系列句子的並排，而是組織完善、有中心意念的文字鋪陳，提供讀者閱讀、欣賞、獲得資訊、或是與作者溝通等等的功能。言談分析（Discourse Analysis）的目的在於探究文件是如何組織，並且企圖瞭解文件意涵的結構。這樣的結構通常跨越數個句子，組成一個意涵單位，而被稱為言談段落（Discourse Segment, DS）。許多學者專家提出各種不同的看

法，嘗試建構言談段落裡句子的關係以及言談段落彼此間的關係。例如，修辭言談結構（Rhetorical Discourse Structure, RDS）（註1）著重於結構的 intention 與 attention，以反應言談段落裡句子的組合意涵。而言談表示理論（Discourse Representation Theory, DRT）（註2）提出所謂的言談表示結構（Discourse Representation Structure, DRS），用以反應言談段落裡句子的語法上合成的關係。

雖然已經投入許多研究的努力，然而對於言談結構的看法，學者專家並沒有達到一致的結論。本文則提出一種基於語料庫研究的言談模型，透過語料庫的研究，認為文件是由一系列的事件（Event）構成，而文件中的名詞與動詞則是這些事件的核心角色。（註3）因此可以將名詞與動詞的關係建構於述語參數結構（Predicate-Argument Structure）；而名詞與名詞的關係則是建構於言談的層次。至於這種關係的強弱，可以藉由語料庫涵蓋的真實語料，透過統計式語言模型，建構適當的評分方式訂定之。為了便於本文的討論，筆者稱這種關係為「關聯正則」（Association Norm）。

若能夠建構適當的言談模型，計算（名詞，名詞）與（名詞，動詞）的關聯正則，不僅可用以切分文件，形成意涵獨立的言談段落，也能夠用以決定言談段落的主題。對於前者，可以基於某個跨度（Span），計算跨度內各句子所有關聯正則的組合積分，可得到這個跨度內句子形成言談段落的分數，當分數越高，代表構成言談段落的可能性越高。對於後者，可以計算言談段落中，每一個名詞與其餘名詞以及動詞關聯正則的組合積分，分數越高的名詞，越有可能成為言談段落的主題。

本文於以下各節中，詳細說明筆者如何建構關聯正則，並用之處理上述兩種應用。第二節筆者主要以計算的觀點，回顧言談分析相關的研究。第三節正式地建構名詞與名詞、名詞與動詞的關聯正則。第四節討論如何計算潛在的言談段落的分數，並且用來處理文件的切分，同時進行若干實驗。第五節說明如何自動判定言談段落的主題，同樣也進行相關的實驗，並且評估模型的效能。

二、言談分析相關研究的回顧

許多學術論文曾經討論切分文件的種種策略，例如，Youmans（註4）記錄文件某跨度內作者引入新詞彙的數目，然後根據這樣的統計數據，決定言談段落的邊界。這種作法的缺點是，僅僅考慮詞彙重複出現的因素。Morris 與 Hirst（註5）提出五種不同的相近關係（Thesaural Relation），企圖找出詞彙上的關連，然後使用這些關連性找出文件的結構。Morris 與 Hirst 的作法有下面的缺點，僅僅找出詞彙間有沒有前述的相近關係，然而卻不規範關係的強弱；另外並沒有提出自動的程序，而是使用人力建構言談結構。Hearst（註6）則提出稱為 TextTiling 比較新穎的演算法。TextTiling 使用詞頻（Term Frequency, TF）與逆向文件頻率（Inverse Document Frequency, IDF），將文件切成一片片馬賽克（Tile）。然而整個演算法只考慮名詞，忽略其餘類型的詞彙，同時也忽略詞彙共現的關係。

文件主題的自動辨識則少有學術論文討論。然而，對於許多實際的研究，主題的自動辨識相當重要。例如，照應詞（Anaphora）的解決與資訊檢索的應用。自動辨識所得的主題可充當 DRT 理論探討的宇集（Universe）的成員，亦即照應詞所指先行詞（Antecedent）的候選集合。至於有關資訊檢索的應用，由文件自動決定的主題可作為非控制式的關鍵詞，作為控制式索引詞彙的輔助。

三、詞類的關聯性

言談結構中的資訊流動可以用事件與參與其中的詞彙描述，以下面的一段文字作例子：

> 安排與小華開個會議吧，明天下午三點半，我們在他的辦公室見面，也請大明一起來。

這段文字描寫「開會」的事件，參與這事件的詞彙，如「安排」、「見面」、「會議」、「辦公室」彼此關係相當的密切。這種密切的關係，可以藉由詞彙間的共容語意（Collocational Semantics）求得。（註7）以下筆者會逐步建立分析文件內容的模型。

首先筆者將說明詞彙的重要性，每一個詞彙在文件中扮演的角色不同，不能視為同樣地重要。Salton（註8）在1986年提出逆向文件頻率的觀念，相當適用於詞彙重要性的度量。以下是IDF的定義：

$$IDF(W) = \log((P - O(W))/O(W)) + c$$

這裡 P 代表訓練語料庫文件的數目，$O(W)$ 代表含有詞彙 W 的文件數目；而 c 只是一個門檻值，用於處理前一項為負值時的情形。筆者使用 LOB Corpus 訓練模型所需的統計參數。LOB Corpus 是一份具有一百萬詞的英國英語語料庫，總共有500篇文章，15種不同的文章型態，因此在本文 P =500。（註9）只要詞彙 W 在語料庫中一半以上的文件出現過，則該詞彙的 $\log((P - O(W))/O(W))$ 值即為負數，這也代表 W 並不重要，亦即 W 對訓練語料庫中的文章不具有鑑別性，出現與否對讀者或資訊尋求者不具任何意義。下面列出這些不重要的詞彙

名詞

time (-3. 68)	way (-1. 92)	year (-1. 71)
man (-1. 47)	day (-1. 12)	part (-0. 76)
people (-0. 75)	thing (-0. 73)	hand (-0. 54)
life (-0. 51)	fact (-0. 40)	place (-0. 40)
work (-0. 35)	end (-0. 12)	case (-0. 09)
point (-0. 05)		

動詞

make (-5. 01)	take (-3. 56)	give (-2. 95)
come (-2. 45)	find (-2. 30)	see (-2. 26)
know (-2. 20)	say (-2. 18)	go (-2. 11)
seem (-1. 30)	show (-1. 20)	think (-1. 18)
use (-1. 07)	get (-1. 06)	become (-0. 95)
bring (-0. 73)	put (-0. 68)	leave (-0. 62)
look (-0. 48)	call (-0. 43)	tell (-0. 41)
keep (-0. 32)	hold (-0. 18)	ask (-0. 23)
begin (-0. 08)		

有了詞彙重要性的度量之後，可以考慮兩個詞彙在訓練語料庫中共同出現時，如何計算此次詞彙共現關係的強度。筆者定義下面兩個式子分別計算名詞與名詞以及名詞與動詞一次共現關係的強度：

$$SNV(N_i, V_j) = IDF(N_i) \bullet IDF(V_j) / D(N_i, V_j)$$
$$SNN(N_i, V_k) = IDF(N_i) \bullet IDF(N_k) / D(N_i, N_k)$$

其中 SNV 代表名詞與動詞的共現強度； SNN 則代表名詞與名詞的共現強度；至於 $D(X,Y)$ 表示 X 詞彙與 Y 詞彙之間的距離，目前以 X 與 Y 之間的詞彙數目表示距離。要注意的是當 $i = k$ 時， $SNN(N_i, N_k)$ 等於 0 。引入距離的因素是考慮一般文件中，相關的事件會出現在相近的位置，否則讀者容易遺忘，因此作者總是盡可能地將之妥善安排。

距離計算的方式如下，以文件的段落為單位，為每一個名詞與動詞設定一個編號，正如下面這一段文章所示：

With so many problems$_1$ to solve$_2$, it would be a great help$_3$ to select$_4$ some one problem$_5$ which might be the key$_6$ to all the others, and begin$_7$ there. If there is any such key-problem$_8$, then it is undoubtedly the problem$_9$ of the unity$_{10}$ of the Gospel$_{11}$. There are three views$_{12}$ of the Fourth Gospel$_{13}$ which have been held$_{14}$.

詞彙 X 與 Y 的距離 $D(X,Y)$ 就可以用以下的方式計算：

$$D(X,Y) = \text{abs}(C(X) - C(Y))$$

這裡的 abs 為絕對值函數， $C(X)$ 代表詞彙 X 的編號，如 C(begin) =7，而 C(problem) =9，所以 D(begin, problem) = 2。

現在可以計算訓練語料庫中，相同配對的詞彙，所有共同出現的次數，作為這一組詞彙配對的關聯正則，以下列二式表示。

$$ANV(N_i, V_j) = \sum SNV(N_i, V_j)$$

$$ANN(N_i, N_k) = \sum SNN(N_i, N_k)$$

ANV 代表名詞與動詞的關聯正則； *ANN* 代表名詞與名詞的關聯正則。這裡筆者要特別強調，當詞彙出現的次數少，其 *IDF* 值會比較高。然而，根據上面兩個式子，這類詞彙被累加的機會也比較少，因此， *IDF* 與詞彙出現的頻率有互補的作用，最終關聯正則的高低，端視這兩個因素之間的抵換效果。英國英語語料庫LOB Corpus則用以訓練各詞彙配對的關聯正則。表一簡要說明LOB Corpus的各項統計數據。

表一、 LOB Corpus 的統計數據

文件數	500
段落數	18, 678
句數	54, 297
名詞數	23, 399
動詞數	4, 358
名詞 / 動詞配對數	422, 945
名詞 / 名詞配對數	3, 476, 842

求得詞彙間的關聯正則後，可用以計算新的文件中，某個名詞與其餘名詞、動詞的關係，以求得該名詞在文件的重要程度。假設文件某個段落有 m 個名詞與 n 個動詞。對於名詞 $N_i(1 \le i \le m)$ 在這段落的聯結強度（Connective Strength, *CS*），筆者有以下的定義：

$$CS(N_i) = PN \cdot CSNN(N_i) + PV \cdot CSNV(N_i)$$
$$CSNN(N_i) = \sum_k ANN(N_i, N_k) / D(N_i, N_k)$$
$$CSNV(N_i) = \sum_k ANN(N_i, V_k) / D(N_i, V_k)$$

CSNN 與 *CSNV* 為名詞 N_i 分別與段落中其餘的動詞以及名詞的聯結強度， *PN* 與 *PV* 則是 *CSNN* 與 *CSNV* 的權重參數， $PN + PV = 1$。讀者可以發現，基於相同的原因，前述的式子仍然引入了距離的因素。筆者使用消去內插法

（Deleted Interpolation，註10），以下列的數學式子計算 PN 與 PV 值。

$$S_N = \sum_i \frac{PN \times CSNN(N_i)}{PN \times CSNN(N_i) + PV \times CSNV(N_i)}$$

$$S_V = \sum_i \frac{PV \times CSNV(N_i)}{PN \times CSNN(N_i) + PV \times CSNV(N_i)}$$

$$PN = \frac{S_N}{S_N + S_V} \qquad PV = \frac{S_V}{S_N + S_V}$$

訓練語料庫LOB Corpus被分成3:1兩部份，PN 與 PV 的初始值設定為0.5。第一回訓練過程，使用3/4語料庫，產生新的 PV 與 PN 值。第二回訓練過程，使用1/4的語料庫重複計算 PV 與 PN 值，直到 PN 與 PV 值收斂。最後，PN 與 PV 的收斂值分別為0.675844與0.324156。

行文至此，整個模型的數學架構已經完成，讀者如果比較仔細的話，應該可以發現這個模型與本文第二節討論的其他學者專家的模型，有何重大的區別與優點，筆者將之條列如下：

- 詞彙的頻率不是唯一考慮的因素。
- 考慮動詞的重要性。
- 考慮距離的因素。
- 透過真實的語料計算詞彙的關係。
- 度量詞彙間的關係，不再只是「有」或是「沒有」關係。

筆者將運用這個模型，在下面兩節中分別討論文件切分與主題辨識的應用。

四、文件切分

學者專家對於言談的共識是：「言談中的句子具有緊密結合的關係」。至於如何描述這種緊密結合的關係則眾說紛紜，各有不同的作法，更遑論提出自動

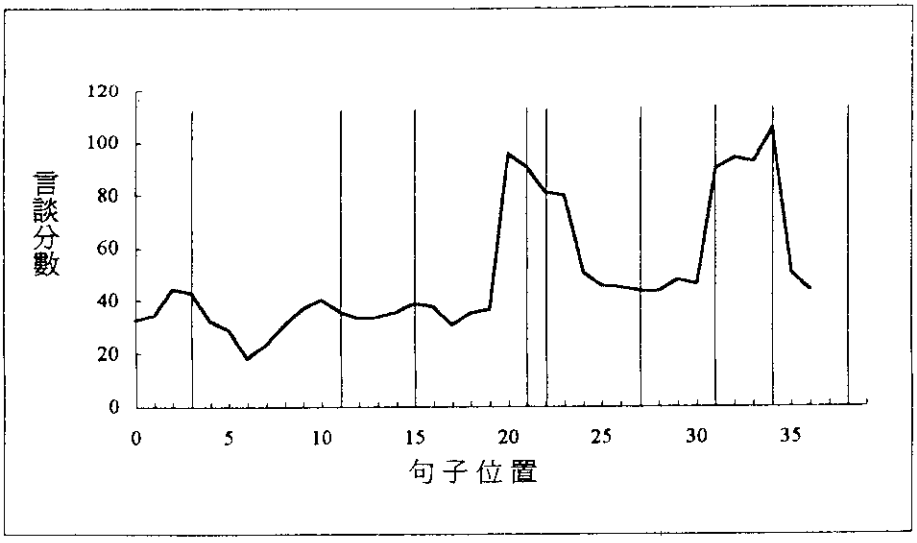
計算的程序。根據本文第三節提出的模型，筆者採用如下的作法。假設有一可能形成言談段落的句子序列包含 n 個名詞，構成言談段落的分數為這些名詞的聯結強度的總和除以 n。以數學式子表示如下：

$$score(DS) = \sum_{i=1}^{n} CS(noun_i)/n$$

現在要將文件切分為緊密結合的段落，可以使用大小為 w 的移動窗，由文件第一個句子逐步往後移動，一次一個句子，並且計算一次移動窗內這些句子構成言談段落的分數。如此，能夠求得句子位置與言談分數的相對關係，有了位置／分數關係圖，找出言談段落邊界的工作就比較容易了。以這裡提出的作法，當移動窗內的句子分數比較高時，形成言談段落的可能性就越高，這時移動窗的位置就是可能的言談段落的邊界。筆者以實際語料求得的關係圖展示於圖一，圖中的垂直線是語料的實際段落邊界，可看到垂直線與位置／分數曲線的高峰有相當多的重疊現象，這也表示模型正確率相當高。

移動窗的大小 w 是這個計算模型的參數，如何適當選定 w 也是重要的考量因素。w 不能太小，因為這樣與言談段落為句子組成的看法互相矛盾；也不能太大，這樣對於言談邊界的判定會比較不準確。筆者選定 w = 3, 4, 5, 與 6 四種不同的情形，針對 10 篇由 LOB Corpus 選出的文章（參見表二）進行實驗，實驗結果與五位邀請的讀者進行比對。

五位受邀的讀者，在不給予任何的提示下，逕行切分 10 篇文章（註 11），圖二下半部是第一篇文章的實驗結果，而上半部則是五位讀者對同一文章的切分結果。可發現曲線的高峰處與讀者切分處有明顯重合的現象，尤其是第22句的位置，五位讀者都認為應該切分，而實驗的結果也同樣在該處有一高峰。若是再仔細探究圖二，會發現一些有趣的現象。第一，總共有15個位置被五位讀者認為是言談段落的邊界，然而其中有9個位置僅有一人認為是邊界，可見大家看文章時的心靈運轉是各不相同。第二，如果有某個位置被多數人認為是邊界，則這個位置的確越有可能是言談的邊界，實驗的結果也肯定了這項觀察，大家看文章時心靈的運轉雖然各不相同，但是也有相似之處。



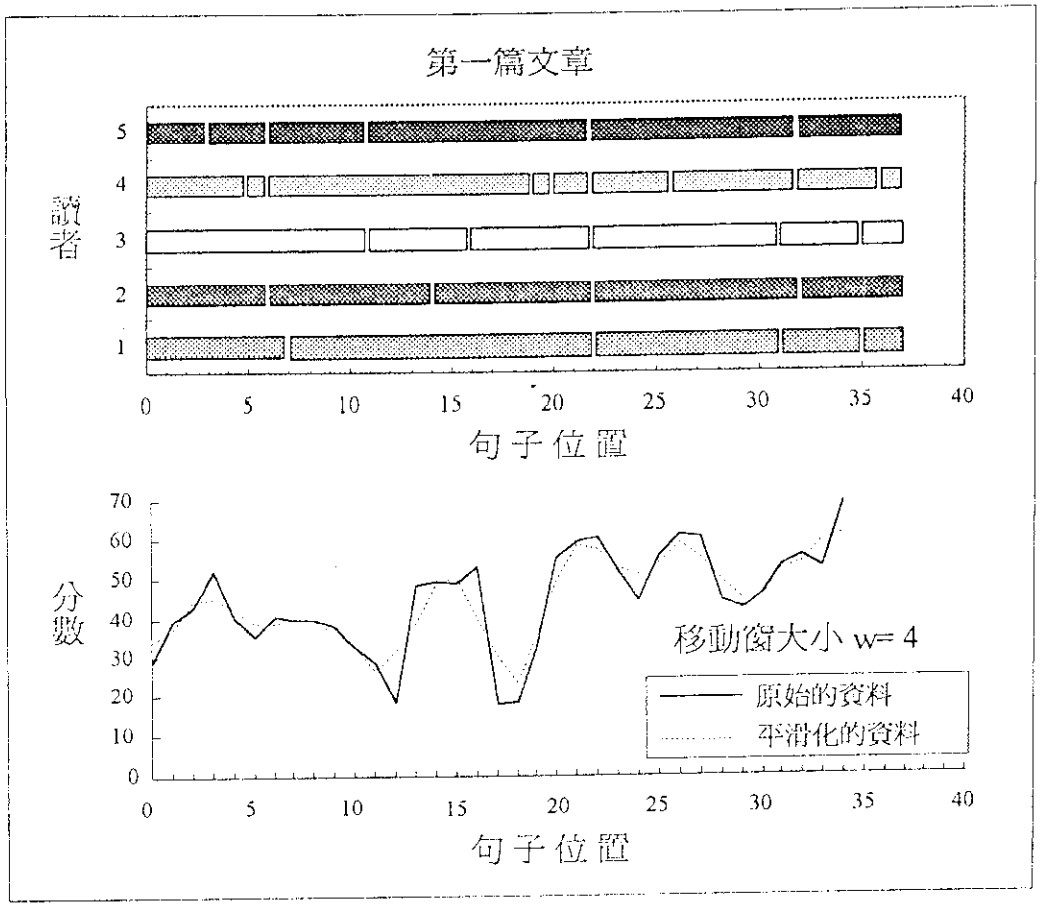
圖一、預測的段落分佈與實際段落分佈的比較

筆者接著比較五位讀者與模型的精確率（Precision）與召回率（Recall），表三非常詳細地列出實驗的結果。精確率在此處表示判定為邊界的位置，而其確實為邊界位置的比率；而召回率則是模型或讀者判定的真正邊界的數目除以文章中真正邊界的數目。通常任何一種模型或是系統企圖追求高精確率與高召回率，然而這兩項評估標準常常有抵換的效果。表三可以看到選擇不同的移動窗，對實驗結果的影響；也可以觀察讀者判定的一致性。

當移動窗的大小 w 等於 3 時，精確率與召回率都是最好，這符合一般的觀察，然而比較有趣的情形是，w 等於 6 時的精確率與召回率都比 w 等於 5 時的精確率與召回率要好。要注意的是，6剛好是3的倍數，這顯示最佳移動窗大小 w 的倍數，通常也有比較好的實驗結果。至於讀者判斷的精確率與召回率，在一般的情形下，比模型的精確率與召回率的表現要好。但是，只要選擇適當的 w 值，模型的表現就與一般讀者的表現相去不遠（請參考表三的最後一列）。

表二、實驗文章的各項資料

	語料庫索引標記	段落數	句數	詞數
第一篇文章	G31 : 002 - G31 : 061	11	38	670
第二篇文章	G31 : 123 - G31 : 196	10	40	897
第三篇文章	A11 : 004 - A11 : 058	9	18	618
第四篇文章	K21 : 004 - K21 : 184	33	77	2282
第五篇文章	J61 : 002-J61 : 192	14	63	2774
第六篇文章	H21 : 002 - H21 : 193	15	97	2210
第七篇文章	G71 : 004 - G71 : 170	13	88	2017
第八篇文章	G61 : 002 - G61 : 123	7	54	1564
第九篇文章	J71 : 003 - J71 : 046	5	16	481
第十篇文章	J71 : 048 - J71 : 077	3	12	333



圖二、第一篇文章的實驗結果與五位讀者對同一文章的切分結果

表三、讀者與模型的精確率與召回率的比較

文章	評估標準	讀者 1	讀者 2	讀者 3	讀者 4	讀者 5	W = 3	W = 4	W = 5	W = 6
1	精確率	0.50	0.75	0.40	0.38	0.60	0.43	0.50	0.20	0.33
	召回率	0.50	0.75	0.50	0.75	0.75	0.75	0.75	0.25	0.50
2	精確率	0.60	0.40	0.67	0.27	0.67	0.50	0.50	0.20	0.20
	召回率	0.60	0.40	0.80	0.60	0.40	0.60	0.60	0.20	0.20
3	精確率	0.75	1.00	1.00	0.43	1.00	0.67	0.50	0.00	0.00
	召回率	1.00	0.67	0.67	1.00	0.67	0.67	0.33	0.00	0.00
4	精確率	0.50	0.62	0.42	0.19	0.50	0.29	0.23	0.00	0.17
	召回率	0.63	1.00	0.63	0.75	0.50	0.25	0.25	0.00	0.13
5	精確率	0.45	0.78	0.67	0.15	0.45	0.60	0.50	0.20	0.40
	召回率	0.71	1.00	0.86	0.43	0.71	0.43	0.29	0.14	0.28
6	精確率	0.69	0.70	0.64	0.44	0.64	0.23	0.33	0.17	0.28
	召回率	0.90	0.70	0.90	0.70	0.90	0.20	0.30	0.10	0.20
7	精確率	0.40	0.67	0.50	0.41	0.40	0.60	0.60	0.40	0.40
	召回率	0.57	0.57	0.71	1.00	0.29	0.43	0.43	0.28	0.28
8	精確率	0.88	0.71	0.63	0.35	0.88	0.57	0.50	0.33	0.40
	召回率	0.88	0.63	0.63	0.75	0.75	0.50	0.38	0.25	0.25
9	精確率	0.20	1.00	0.50	0.40	0.33	1.00	0.50	0.50	0.50
	召回率	0.50	1.00	0.50	1.00	0.50	1.00	0.50	0.50	0.50
10	精確率	0.50	0.50	0.50	0.67	0.50	1.00	1.00	1.00	1.00
	召回率	0.67	0.33	0.33	0.67	0.67	0.33	0.33	0.33	0.33
平均	精確率	0.547	0.713	0.593	0.369	0.597	0.589	0.516	0.300	0.368
	召回率	0.696	0.705	0.657	0.765	0.614	0.516	0.416	0.205	0.267

五、 主題辨識

Brown and Yule 曾經說明：「雖然只有讀者或是作者而不是文件有主題的概念」(註12)，但是企圖建構模型，辨識或是判斷文件主題，仍是資訊檢索學者專家努力的目標。對於這個問題，筆者嘗試藉由第三節討論的模型，提出一種新的作法。首先檢視一段取材於 LOB Corpus 的真實語料：

There is a whole group of theories which attempt to explain the problems of the Fourth Gospel by explanations based on assumed textual dislocations. The present state of the Gospel is the result of an accident-prone history. The original was written on a roll, or codex, which fell into disorder or was accidentally damaged. An editor, who was not the author, made what he could of the chaos by placing the fragments,

or sheets, or pages, in order. Most of those who expound a theory of textual dislocation take it for granted that the Gospel was written entirely by one author before the disturbance took place but a few leave it open to suppose that the original book had been revised even before the upheaval.

這段文字討論的是錯簡的問題（Dislocation Problem），而這兩個名詞 dislocation 與 problem 和本段文字中其餘的動詞（explain, fell, placing, suppose）與名詞（theories, explanations, roll, codex, disorder, order, disturbance, upheaval）顯然有極密切的關係。基於這樣的觀察，筆者提出下面的看法：

主題與言談中的事件有緊密結合的關係

這種緊密結合的關係，用前面提及的聯結強度（CS）可決定言談中哪些名詞可能是主題。由於言談段落中，同一個名詞可能不只出現一次，必須進一步處理。假設同一個名詞 N 在某個言談段落出現次，第一次出現，其淨聯結強度（Net CS, NCS）等於 $NCS(N_{o(1)}) = CS(N_{o(1)})$，$O(K)$ 表示第 k 次出現時的編號（參見第三節有關編號的說明）；若第 i 次出現，則 $NCS(N_{o(i)}) = NCS(N_{o(i-1)}) + (1 - NCS(N_{o(i-1)}))CS(N_{o(i)})$，要注意的是，$C(N_{o(i)}) < C(N_{o(2)}) < C(N_{o(3)}) < .. < C(N_{o(k-1)}) < C(N_{o(k)})$。

筆者仍然使用 LOB Corpus 作為測試語料，由 LOBT-D1 、 LOBT-F1 、 LOBT-G1 、 LOBT-H1 、 LOBT-K1 、 LOBT-M1 、與 LOBT-N1 這幾類文章抽取七篇進行實驗（為了方便，以 D01 、 F01 、 G01 、 H01 、 K01 、 M01 、與 N01 表示）。每一篇文章都經過一位語言學專家事先決定每一段落的主題集合，稱之為「判定的主題」。而這裡提出的模型，用以計算每一段落各名詞的 NCS，具有最高 NCS 的前 20% 的名詞形成「計算的主題」。表四記載筆者用以分析模型效用的幾項指標，實驗結果則展示於表五。先看最後標記「+」、「-」、與「?」的三列，「+」表示這篇文章有多少段落的主題被模型正確地計算得到；「-」表示計算錯誤的段落數目；「?」表示無法判斷的段落數（註 13）。若是扣除無法決定的段落，本文提出的模型的正確率為 80/131=61.07% 。

第一列到第六列對應表四的一到六列，這六列到底透露出什麼訊息。以D01這篇文章為例，第一列表示平均每一段有21.59個名詞，因此，企圖從其中隨機選取名詞作為主題是不可能，這項統計數字說明主題的自動辨識是相當困難的工作。第二列記載的是語言學專家判定的主題，在計算後所得的排名。（註14）比較第一列與第二列，可以看出專家所判定的主題在平均21.59個名詞中，平均排名為 4.56 ，以這麼困難的問題而言，這樣的結果算是相當不錯。第三列、第四列、與第五列分別是所有名詞的平均頻率、判定主題的平均頻率、與計算主題的平均頻率，這三項數字主要是探究一般人閱讀時的行為，是否受文章中經常出現的詞彙影響，而認為這些詞彙可能是文章的重點。由這些統計數字可看出，無論是專家判定的主題或是模型計算的主題，其平均頻率都大於所有名詞的平均頻率，這也符合一般讀者的閱讀行為。

第六列是另一項有趣的統計數字。有時讀者會有疑問：「為何作者在此處分段？」筆者希望藉由數字來討論這個問題。第六列的數字表示本段的主題在前一段落的平均排名。結果顯示，本段主題的排名的確是落在後面，充分反應作者在寫作的過程中，重新起段的行為並非是隨意的，應是經過思考的結果。

表四、模型效用的幾項指標

1	候選名詞的平均數目	Σ # of nouns in basic form in paragraph i / # of paragraphs
2	判定主題的平均排名	Σ rank of assumed topic in paragraph i/# of paragraphs
3	候選名詞的平均頻率	Σ # of nouns / Σ # of nouns in basic form in paragraph i
4	判定主題的平均頻率	Σ occurrences of assumed topic / # of paragraphs
5	計算主題的平均頻率	Σ occurrences of computed topic / # of paragraphs
6	本段主題在前一段落的平均排名	Σ rank of topic in the previous paragraph / (# of paragraph - 1)

表五、實驗的結果

	D01	F01	G01	H01	K01	M01	N01
1	(21.59, 9.96)	(10.57, 18.42)	(62.43, 18.42)	(19.77, 8.39)	(31.71, 23.80)	(15.22, 6.44)	(12.21, 6.73)
2	(4.56, 5.98)	(5.25, 5.51)	(7.29, 10.35)	(4.55, 4.13)	(7.08, 16.02)	(2.61, 2.11)	(3.68, 3.87)
3	(1.32, 0.88)	(1.39, 0.89)	(1.21, 0.56)	(1.33, 0.82)	(1.11, 0.39)	(1.11, 0.32)	(1.06, 0.25)
4	(2.61, 1.60)	(1.27, 1.21)	(2.57, 1.18)	(2.46, 1.62)	(1.77, 1.05)	(1.50, 0.69)	(1.28, 0.60)
5	(3.33, 1.97)	(2.39, 1.84)	(3.43, 1.40)	(2.91, 1.56)	(1.86, 0.99)	(1.48, 0.50)	(1.29, 0.52)
6	(6.29, 7.84)	(5.48, 5.09)	(19.67, 16.64)	(5.71, 6.06)	(17.23, 18.51)	(7.92, 6.28)	(9.36, 6.62)
+	12	13	6	12	9	13	15
-	6	15	1	10	4	5	10
?	0	0	0	0	1	9	9

註：＋表示這篇文章有多少段落的主題被模型正確求得

－表示計算錯誤的段落數

？表示無法判斷的段落

六、結語與討論

本文提出一種建構於語料庫的新模型，這個模型考慮了幾項重要的因素：

- 詞彙的重複（Repetition of Words）
- 詞彙重要性（Importance of Words）
- 共容語意（Collocational Semantics）

詞彙重複的考量表現於連結強度的計算上；詞彙的重要性是以逆向文件頻率表示；而共容語意則是以關聯正則表示。根據前述的模型，筆者將之運用於文件的切分與主題的辨識等研究。

文件的切分有其實際的用途，Hearst（註 15）曾經詳細討論文件的切分對資訊檢索的影響，使用文件切分後的言談段落，作為查詢比對的單位，精確率與召回率都有顯著地提昇。除此之外，文件的切分還有其餘應用。在通訊網路交通如此方便的時代，資訊的交流非常頻繁，雙語或甚至多語的資訊與文件，越來越

多。因此，文件中兩種語言句子對列的工作，變得非常重要。從事這項研究的學者專家發現（註16），如果能夠採用值得信賴的文件定位點（如文件的段落處），對列的精確率與召回率可以提昇大約15%。採用文件切分的方法，正可以製造更多的定位點（言談段落的邊界處），有效提昇雙語文件對列的結果。

對於文件主題的辨識，這裡提出的作法是考慮與其餘名詞以及動詞有很強的聯結強度的名詞組成主題集合。實驗的結果顯示正確率為61.07%，但目前尚未看到其餘相關研究的結果報告，因此無法比較。另一項有趣的現象是主題移轉（Topic Shift）。一般作者撰寫文章時，新起一段是否代表主題有所轉移，針對這個問題，筆者特別計算新段落的主題與前一段落的關係。根據實驗所得的數據，新段落主題與前一段落的關係明顯偏低，證實了主題移轉的現象確實存在。

文件內容的分析一直是個困難的研究課題。筆者提出的分析模型是以計算的角度出發，希望這樣的觀點對文件分析的研究能有些微的貢獻。筆者也將根據不同的語料進行更大規模的實驗，以檢驗該計算模型的調適性。

致謝

筆者感謝台灣大學圖書館學系陳雪華教授的寶貴意見，以及資訊工程學研究所博士候選人，李御璽，提供的協助。

註 釋

註 1： B. Grosz and C. Sidner, "Attention, Intentions, and the Structure of Discourse," Computational Linguistics, 12 : 3 (1986) : 175-204.

註 2： H. Kamp, "A Theory of Truth and Semantic Representation," in J. Groenendijk, T. Janssen, and M. Stokhof, Eds. Formal Methods in the Study of Language, 1 (Mathematische Centrum, 1981) .

註 3： 基本上，語料庫是語言素材的集合，包括報紙、小說、文學作品等等。至於要蒐集那類型的語言資料，端賴語料庫建構者的籌畫與爾後的使用方式。由收集的語言種類，語料庫可分為單語語料庫、雙語語料庫、多語語料庫；以收集的方式，可分為平衡式語料庫與非平衡式語料庫，也就是考量語料型態分佈的情形；由整理的方式，可分為原始語料庫、標記語料庫、與樹狀語料庫。

註 4： G. Youmans, "A New Tool for Discourse Analysis : The Vocabulary-Management Profile," Language, 67 (1991) : 763-789.

註 5： J. Morris and G. Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Texts," Computational Linguistics, 17 : 1 (1991) : 21-48.

註 6： M. Hearst, TextTiling : A Quantitative Approach to Discourse Segmentation, Sequoia 20000 93/24 (Berkeley : University of California, 1993) .

註 7： F. Smadja, Extracting Collocations from Text. An Application : Language Generation, Ph. D. Dissertation (New York : Columbia University, 1991) .

註 8： G. Salton, "On the Use of Term Associations in Automatic Information Retrieval," Proceedings of COLING-86, (Bonn, 1986) : 380-386.

註 9： S. Johansson, The Tagged LOB Corpus : Users' Manual, (Bergen : Norwegian Computing Centre for the Humanities, 1986) .

註 10： F. Jelinek, "Markov Source Modeling of Text Generation," in J. Skwirzynski, Eds. The Impact of Processing Techniques in Communication , (Nijhoff, Dordrecht, 1985) .

註 11：Passonneau 與 Litman 以及 Hearst 三位學者分別在下面兩篇學術論文中，採用相同的方式評估實驗的結果。

R. Passonneau and D. Litman, "Intention-Based Segmentation : Human Reliability and Correlation with Linguistic Cues, " Proceedings of the 31st Annual Meeting of ACL, (Columbus, 1993) : 148-155.

M. Hearst, "Multi-Paragraph Segmentation of Expository Text, " Proceedings of the 32nd Annual Meeting of ACL, (Las Cruces, 1994) : 9-16.

註 12：G. Brown and G. Yule, Discourse Analysis, (Cambridge : Cambridge University Press, 1983) .

註 13：因為專家判定的主題是代名詞，而本文提出的模型，並沒有考慮代名詞，這或許是有待改進的地方。然而困難的地方是代名詞是一種照應詞（Anaphora），本身就有必要先決定代名詞到底指的是哪一個名詞。另外，代名詞出現的頻率相當高，如何適當的對待代名詞，仍有待進一步的努力。

註 14：每一段落經由本文模型的計算後，根據NCS的數值，會得到一份所有名詞的排名表列，排名第一的當然是模型「計算的主題」，若是專家「判定的主題」排名越前面，代表模型的效用越高。值得注意的是，排名的數值越少，代表排名越高。

註 15：M. Hearst and C. Plaunt, "Subtopic Structuring for Full-Length Document Access, "Proceedings of SIGIR-93, (Pittsburgh, 1993) : 59-68.

註 16：K. H. Chen and H. H. Chen, "A Part-of-Speech-Based Alignment Algorithm, " Proceedings of COLING-94, (Kyoto, 1994) : 166-171.