第三章　自動新詞擷取方法

本文提出的新詞自動擷取方法總共包含三個部份，分別為一.新詞自動擷取方法，二.舊有詞庫的應用與三.淘汰錯誤且過時的字詞。透過第一和第二部份並搭配 Google 新聞資料，將能有效建立一新聞類專業詞庫，不需依靠人工即能達到一定的正確率。由於隨著時間的變化，新聞資料將不停的累積，有些新聞用詞也會過時而不再被使用，且所建立的新聞類專業詞庫也會隨著時間而膨脹，進而影響中文斷詞系統的效率。因此經由第三部份來淘汰不適當的字詞，有效避免詞庫過大的問題。

3.1　新詞自動擷取演算法

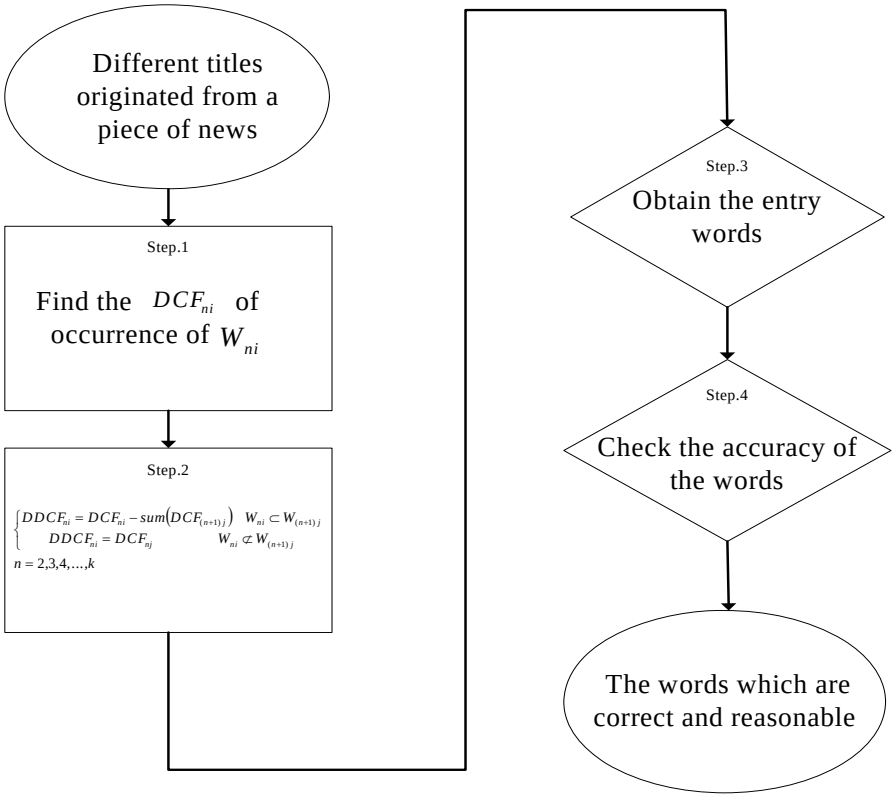

圖3-1　新詞自動擷取架構圖

圖 3-1 為本文的新詞自動擷取架構圖，新詞自動擷取總共有四個步驟，將新聞資料套入本文提出的方法，經過四個步驟後將自動擷取出新詞。

Step.1

先從 Google 提供的新聞服務中自動提取多則經由 Google 分類好之同一新聞事件的標題，統計出每一字詞的詞頻，並把原本的詞頻轉換成新的詞頻DCF_{ni}，轉換方式如下。

$$DCF_{ni} = \frac{m!}{2!(m-2)!} \qquad (3\text{-}1)$$

其中DCF_{ni}代表第第 i 個 n 字詞W_{ni}的重複組合頻率(duplicate combination frequency)，m代表W_{ni}原先的詞頻。

如圖 3-2 當台灣出現兩次時其$DCF_{2(台灣)} = \frac{2!}{2!\times 0!} = 1$，假如出現三次則$DCF_{2(台灣)} = \frac{3!}{2!\times 1!} = 3$，$DCF_{ni}$對於原先的詞頻有放大的效果，相對於其他詞頻的計算方式，本文提出的DCF_{ni}可以透過訂定門檻值來降低錯誤的多字詞。

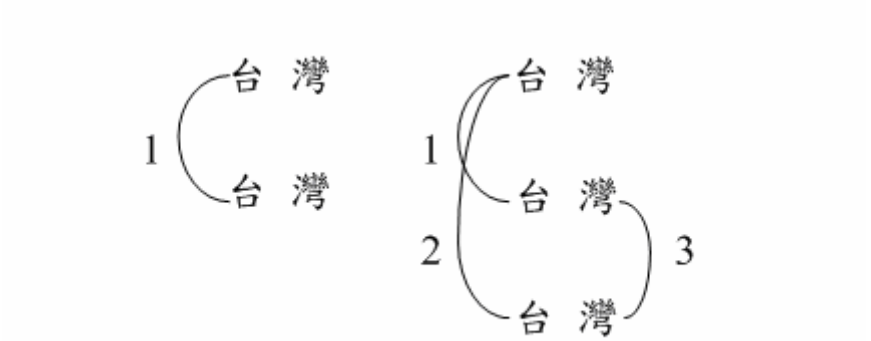

圖3-2　　DCF_{ni}計算方式

實際範例如表 3-1 所示，系統將自動擷取被 Google 分類好之單一新聞事件的所有標題，由於某些新聞媒體會定出完全相同的標題，所以系統會先濾除完全相同的標題。如表 3-2 至表 3-5 所示，接下來系統會將所有標題切成二、三、四、五字詞並統計其詞頻，詞頻將會被轉換成 DCF_{ni}。

表3-1　同一新聞事件之不同新聞標題與來源

新聞標題	新聞來源
雲南兩飛機相撞墜毀一飛行員跳傘具體傷亡不詳	東方網
中共兩架軍機在雲南相撞墜毀	蕃薯藤新聞
中共空軍驚發惡性事故兩戰機空中相撞一副團長死亡	新浪網
新手追撞前機兩殲七墜毀	臺灣蘋果日
雲南相撞的戰機證實為兩架殲七副團長已犧牲	澳洲日報
雲南兩軍機相撞墜毀	中國報
兩機雲南撞毀 1 名飛行員跳傘傷亡不詳	香港商報
中共軍機操練傳出墜毀失事飛機疑為殲七-II 戰機	東森新聞報
中國・雲南軍機相撞一人死一飛行員跳傘受傷	星洲日報
傳殲 10 試飛雲南撞機	成報
雲南兩飛機空中相撞墜燬山頂傷亡情況正在調查	深圳僑報
滇兩殲七機相撞副團長犧牲	香港文匯報
多維簡訊：雲南兩飛機相撞墜毀 1 飛行員跳傘逃生	多維新聞網
彝族聖藥稱能治艾滋雲南秘方開價 50 億美元	多維新聞網

表3-2　　二字詞詞頻與 DCF_{ni}

二字詞	詞頻	DCF_{ni}
雲南	10	45
南兩	4	6
兩飛	3	3
飛機	4	6
機相	5	10
相撞	9	36
撞墜	5	10
墜毀	5	10
一飛	2	1
飛行	4	6
行員	4	6
員跳	4	6
跳傘	4	6
傷亡	3	3
亡不	2	1
不詳	2	1
中共	3	3
兩架	2	1
軍機	4	6
南相	2	1
戰機	3	3
機空	2	1
空中	2	1

中相	2	1
撞一	2	1
副團	3	3
團長	3	3
兩殲	2	1
殲七	4	6
犧牲	2	1
南撞	2	1

表3-3　　三字詞詞頻與DCF_{ni}

三字詞	詞頻	DCF_{ni}
雲南兩	4	6
南兩飛	3	3
兩飛機	3	3
飛機相	2	1
機相撞	5	10
相撞墜	5	10
撞墜毀	4	6
一飛行	2	1
飛行員	4	6
行員跳	4	6
員跳傘	4	6
傷亡不	2	1
亡不詳	2	1
雲南相	2	1
南相撞	2	1
機空中	2	1
空中相	2	1
中相撞	2	1
相撞一	2	1
副團長	3	3
兩殲七	2	1
軍機相	2	1
雲南撞	2	1

表3-4　　四字詞詞頻與DCF_{ni}

四字詞	詞頻	DCF_{ni}
雲南兩飛	3	3
南兩飛機	3	3
兩飛機相	2	1
飛機相撞	2	1
機相撞墜	3	3
相撞墜毀	4	6
一飛行員	2	1
飛行員跳	4	6
行員跳傘	4	6
傷亡不詳	2	1
雲南相撞	2	1
機空中相	2	1
空中相撞	2	1
軍機相撞	2	1

表3-5　　五字詞詞頻與DCF_{ni}

五字詞	詞頻	DCF_{ni}
雲南兩飛機	3	3
南兩飛機相	2	1
兩飛機相撞	2	1
飛機相撞墜	2	1
機相撞墜毀	3	3
一飛行員跳	2	1
飛行員跳傘	4	6
機空中相撞	2	1

Step.2

如表 3-6 當二字詞裡有被包含在三字詞時,則把二字詞的詞頻減去三字詞的詞頻並以此數目當成新的詞頻的差異重複組合頻率(difference duplicate combination frequency)$DDCF_{ni}$，如果二字詞並沒有被包含在任何三字詞時，則此二字詞的詞頻將不會被更改，同樣的三字詞新的詞頻則是利用四字詞的詞頻來重新計算，以此類推。運算規則如下

$$\begin{cases} DDCF_{ni} = DCF_{ni} - sum\left(DCF_{(n+1)j}\right) & W_{ni} \subset W_{(n+1)j} \\ DDCF_{ni} = DCF_{nj} & W_{ni} \not\subset W_{(n+1)j} \end{cases} \quad (3\text{-}2)$$
$$n = 2,3,4,...,k$$

如式 3-2W_{ni}為第 i 個 n 字詞、而$W_{(n+1)j}$代表第 j 個(n+1)字詞，當W_{ni}被包含在$W_{(n+1)j}$時，其$DDCF_{ni}$將會不同於原本的DCF_{ni}，假設W_{ni}沒有被任何$W_{(n+1)j}$所包含到，則其$DDCF_{ni}$將會等於原本的DCF_{ni}。如式(3-3)二字詞“雲南”被包含在三字詞“雲南兩”、“雲南相”與“雲南撞”則其$DDCF_{ni}$將會是 37。表 3-6 至表 3-8 所示為所有

字詞的 $DDCF_{ni}$ 與 DCF_{ni}。

$$\begin{aligned} & DDCF_{2(\text{雲南})} \\ &= DCF_{2(\text{雲南})} - \left(DCF_{3(\text{雲南兩})} + DCF_{3(\text{雲南相})} + DCF_{3(\text{雲南撞})}\right) \\ &= 45 - (4 + 2 + 2) = 37 \end{aligned} \tag{3-3}$$

表3-6　　二字詞 $DDCF_{ni}$ 與 DCF_{ni}

二字詞	$DDCF_{ni}$	DCF_{ni}
雲南	37	45
南兩	-3	6
兩飛	-3	3
飛機	2	6
機相	-2	10
相撞	13	36
撞墜	-6	10
墜毀	4	10
一飛	0	1
飛行	-1	6
行員	-6	6
員跳	-6	6
跳傘	0	6
傷亡	2	3
亡不	-1	1
不詳	0	1
中共	3	3
兩架	1	1

軍機	5	6
南相	-1	1
戰機	3	3
機空	0	1
空中	-1	1
中相	-1	1
撞一	0	1
副團	0	3
團長	0	3
兩殲	0	1
殲七	5	6
犧牲	1	1
南撞	0	1

表3-7　　三字詞 $DDCF_{ni}$ 與 DCF_{ni}

三字詞	$DDCF_{ni}$	DCF_{ni}
雲南兩	3	6
南兩飛	-3	3
兩飛機	-1	3
飛機相	-1	1
機相撞	5	10
相撞墜	1	10
撞墜毀	0	6
一飛行	0	1
飛行員	-1	6
行員跳	-6	6
員跳傘	0	6
傷亡不	0	1
亡不詳	0	1
雲南相	0	1
南相撞	0	1
機空中	0	1
空中相	-1	1
中相撞	0	1
相撞一	1	1
副團長	3	3
兩殲七	1	1
軍機相	0	1
雲南撞	1	1

表3-8　四字詞 $DDCF_{ni}$ 與 DCF_{ni}

四字詞	$DDCF_{ni}$	DCF_{ni}
雲南兩飛	0	3
南兩飛機	-1	3
兩飛機相	-1	1
飛機相撞	-1	1
機相撞墜	-1	3
相撞墜毀	3	6
一飛行員	0	1
飛行員跳	-1	6
行員跳傘	0	6
傷亡不詳	1	1
雲南相撞	1	1
機空中相	0	1
空中相撞	0	1
軍機相撞	1	1

Step.3

經過第二步驟處理後的字詞，假如新的詞頻 $DDCF_{ni}$ 大於門檻值 R_n 時，則將被當成是正確並合理的字詞。規則如下

$$\begin{aligned} &IF \quad DDCF_{ni} \geq R_n \quad THEN \quad W_{ni} \quad IS \quad correct \\ &ELSE \quad W_{ni} \quad IS \quad wrong \\ &n = 2,3,4,...,k \end{aligned} \tag{3-4}$$

如表 3-9 假如 R_n 設定為 1 時將保留 $DDCF_{ni} \geq 1$ 的字詞

表3-9　　$DDCF_{ni} \geq R_n$ 之字詞

雲南	37
飛機	2
相撞	13
墜毀	4
傷亡	2
中共	3
兩架	1
軍機	5
戰機	3
殲七	5
犧牲	1
雲南兩	3
機相撞	5
相撞墜	1
相撞一	1
副團長	3
兩殲七	1
雲南撞	1
相撞墜毀	3
傷亡不詳	1
雲南相撞	1
軍機相撞	1

如表 3-9 經過第三步驟所產生出來的二字詞經過觀察通常正確率是相當高的，但產生出來的三字詞與四字詞就不一定完全是正確的，很容易出現由一個正確的二字詞配上一個一字詞所組成的三字詞例如“雲南兩”、“兩殲七”等，或者是由兩個正確的二字詞例如“軍機相撞”或一個正確的三字詞配上一個一字詞所組成的四字詞。為了解決此一問題，必須要再經過步驟四來消去這些可能不合理的三字詞與四字詞。

Step.4

把所有第三步驟所產生出來的字詞再做一次比對，由於第三步驟所產生的二字詞正確率相當高，所以將使用這群二字詞來過濾掉不正確的三字與四字詞。當二字詞被包含在三字詞或四字詞時且此二字詞的重複次數大於三字詞或四字詞一定倍數 M 時，則將這三字詞或四字詞判斷成是不正確的並消去此三字詞或四字詞。規則如下

$$\begin{aligned} &IF \quad W_{ni} \subset W_{(n+a)j} \quad AND \quad DDCF_{ni} \le M \times DDCF_{(n+a)j} \\ &THEN \quad W_{(n+a)j} \quad IS \quad correct \quad ELSE \quad W_{(n+a)} \quad IS \quad wrong \\ &where \quad n = 2, a = 1,2, and \quad M \ \ is \ \ a \ \ constant \end{aligned} \tag{3-5}$$

如表 3-10 雲南兩與相撞墜毀的 $DDCF_{ni}$ 皆為 3，而雲南的 $DDCF_{ni}$ 為 37，則將可藉由此規則來判定雲南兩與相撞墜毀是否為正確的字詞。

表3-10　　少字詞取代多字詞

$W_{(n+a)j}$	$DDCF_{(n+a)i}$	W_{ni}	$DDCF_{ni}$
雲南兩	3	雲南	37
機相撞	5	相撞	13
相撞墜	1	相撞	13
相撞一	1	相撞	13
兩殲七	1	殲七	5
雲南撞	1	雲南	37
相撞墜毀	3	(相撞,墜毀)	(13,4)
傷亡不詳	1	傷亡	2
雲南相撞	1	(雲南,相撞)	(37,13)
軍機相撞	1	(軍機,相撞)	(5,13)

表3-11　　系統輸出之正確字詞

正確字詞	$DDCF_{ni}$
雲南	37
飛機	2
相撞	13
墜毀	4
傷亡	2
中共	3
兩架	1
軍機	5
戰機	3
殲七	5
犧牲	1
副團長	3

重複同樣的方法，再利用剩下的三字詞去過濾四字詞，把錯誤的四字詞降至最低。規則如下

$$\begin{aligned} &IF \quad W_{ni} \subset W_{(n+a)j} \quad AND \quad DDCF_{ni} \le M \times DDCF_{(n+a)j} \\ &THEN \quad W_{(n+a)j} \quad IS \quad correct \quad ELSE \quad W_{(n+a)} \quad IS \quad wrong \\ &where \quad n = 3, a = 1, and \quad M \ \ is \ \ a \ \ constant \end{aligned} \tag{3-6}$$

如表 3-11，假如 M=0.1，最後則輸出所有正確並合理的字詞。

3.2 舊有詞庫的應用

對於某些字詞其在單一新聞事件中的詞頻很低。例如字詞W在每一新聞事件中的詞頻都很低，但卻經常出現在不同的新聞事件，這類字詞無法經由本文提出的新詞自動擷取方法所提取出來。為解決此一問題本文將利用舊有詞庫來擷取這些新聞領域相關但詞頻較低的字詞。Shan He和Jie Zhu [13]提出使用Etropy的方法，來擷取中文新詞。而本文將Entropy的方法用於擷取舊有詞庫中曾出現，但是在新聞中詞頻出現較低的字詞。並加入所要建立的新聞類專業詞庫。

$P(w_iW)$和$P(Ww_i)$分別代表一字詞w_i出現在多字詞W左邊和右邊的機率。V_L和V_R則是所有出現在W左邊和右邊之w_i的數目。$H_L(W)$和$H_R(W)$分別是其Entropy值。

$$\begin{aligned} H_L(W) &= \sum_{i=1}^{V_L} -P(w_iW)\times\log P(w_iW) \qquad ,\sum_{i=1}^{V_L} P(w_iW)=1 \\ H_R(W) &= \sum_{i=1}^{V_R} -P(Ww_i)\times\log P(Ww_i) \qquad ,\sum_{i=1}^{V_R} P(Ww_i)=1 \end{aligned} \tag{3-7}$$

當$H_L(W)$和$H_R(W)$兩個值都很高時，代表連接W的字詞w_i種類很多，W將被視為一單獨出現的正確字詞，反之當$H_L(W)$或$H_R(W)$其中一個數值很低時，代表W被包含在其他字詞裡面，不是一個完整且正確的字詞。

如表 3-12 假設“雲南”為舊有詞庫裡的字詞，經由比對新聞資料後就能計算出其$H_L(W)=0.7782$(表 3-13)和$H_R(W)=0.6388$(表 3-14)，並可依據$H_L(W)$和$H_R(W)$值來判斷其是否為一完整且正確的字詞。

表3-12　　範例: w_iW 與 Ww_i

新聞標題	w_iW	Ww_i
雲南兩飛機相撞墜毀一飛行員跳傘具體傷亡不詳		雲南兩
中共兩架軍機在雲南相撞墜毀	在雲南	雲南相
雲南相撞的戰機證實為兩架殲七副團長已犧牲		雲南相
雲南兩軍機相撞墜毀		雲南兩
兩機雲南撞毀 1 名飛行員跳傘傷亡不詳	機雲南	雲南撞
中國・雲南軍機相撞一人死一飛行員跳傘受傷	・雲南	雲南軍
傳殲 10 試飛雲南撞機	飛雲南	雲南撞
雲南兩飛機空中相撞墜燬山頂傷亡情況正在調查		雲南兩
多維簡訊：雲南兩飛機相撞墜毀 1 飛行員跳傘逃生	：雲南	雲南兩
彝族聖藥稱能治艾滋雲南秘方開價 50 億美元	滋雲南	雲南秘

表3-13　　w_iW 之 $P(w_iW)$ 與 $H_L(W)$

w_iW	次數	$P(w_iW)$	$-P(w_iW)\times\log P(w_iW)$
在雲南	1	1/6	0.1297
機雲南	1	1/6	0.1297
・雲南	1	1/6	0.1297
飛雲南	1	1/6	0.1297
：雲南	1	1/6	0.1297
滋雲南	1	1/6	0.1297
Total	6	1	0.7782

表3-14　　Ww_i之$P(Ww_i)$與$H_R(W)$

Ww_i	次數	$P(Ww_i)$	$-P(Ww_i)\times\log P(Ww_i)$
雲南兩	4	4/10	0.1592
雲南相	2	2/10	0.1398
雲南撞	2	2/10	0.1398
雲南軍	1	1/10	0.1
雲南秘	1	1/10	0.1
Total	10	1	0.6388

本文將利用舊有詞庫裡的字詞去比對所有已收藏之新聞資料，將$H_L(W)$和$H_R(W)$皆很高的字詞擷取出來並加入所要建立的新聞類專業詞庫。

3.3　淘汰過時與錯誤的字詞

由於新聞資料龐大，藉由本文提出的方法將能擷取出大量的新詞，但當詞庫擴大到一定的程度時，對於中文斷詞系統將會嚴重影響其執行效率。由於新詞的擷取是透過統計的方法去計算詞頻並擷取詞頻較高的字詞，無法達到百分之百的正確率。所建立的詞庫龐大如果依賴人工去過濾錯誤的字詞，將耗費大量資源。而某些正確的字詞因為時代的變遷也許在未來將不在出現，這些錯誤或不合時宜的字詞將佔據詞庫很大的空間，因此必須制定一有效的方法來解決這個問題。Google news 提供的新聞服務，除了把同一新聞事件的不同新聞標題集合在一起，也提供了新聞標題出現的時間，而考慮時間這個因素將能有效的過濾那些錯誤或不合時宜的字詞，避免詞庫的無限膨脹。

本文所要建立的新聞類專業詞庫，除了字詞以外也記錄了每一字詞出現的時間點，因此將能藉由考慮每一字詞出現的時間點來淘汰不適合的字詞。如圖 3-3，假設現在時間為t，字詞W_{ni}第一次出現的時間為t_1，第二次出現的時間為t_2，以此類推，t_j為字詞W_{ni}第j次出現的時間點

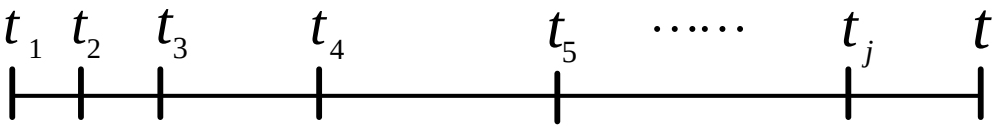

圖3-3　　字詞出現時間點

制訂一規則如下

$$T_1 = t - t_j$$
$$T_2 = \frac{t_j - t_1}{j} \tag{3-8}$$

其中T_1代表字詞W_{ni}已經多久沒出現過，而T_2代表字詞W_{ni}出現的平均間隔時間。考慮T_1與T_2值並結合模糊規則推論，用來判斷詞庫裡的字詞是否應當被淘汰掉，將是一個合理的方法。如表 3-15 使用模糊規則將T_1與T_2模糊化並分為短、中、長，而輸出分為五個等級。表 3-16 代表字詞的可靠度，可以用來決定哪些字詞是必須保留下來而哪些字詞則必須淘汰掉。表 3-17 由於T_1與T_2分別被模糊化為三個等級，總共有九種組合會出現。如表 3-16 這九種情況的字詞會被分為五個等級。

表3-15　　輸入之模糊推論

Linguistic variable	Representation
Short	S
Moderate	M
Long	L

表3-16　　字詞可信度等級

Linguistic variable	Representation
Very Lowly Confidence Degree	VLCD
Lowly Confidence Degree	LCD
Moderately Confidence Degree	MCD
Highly Confidence Degree	HCD
Very Highly Confidence Degree	VHCD

表3-17　　字詞可信度之分類方式

字詞之可信度		T1		
		S	M	L
T2	S	VHCD	HCD	MCD
	M	HCD	MCD	LCD
	L	MCD	LCD	VLCD