第三章 研究設計

有關於試題分類的研究，並沒有專門的研究可供參考，然而在文件分類、新聞分類及網頁分類的研究方面，卻都可以找到很明確的研究方向。因此在實驗設計中也以新聞分類及網頁分類研究轉化成試題分類研究為參考。實驗步驟如下所述:

第一節 確定研究素材

首先確定研究的素材採用具有公信力的法人機構,「中華民國電腦教育發展協會」所研發的 PreMOUS (MOCC 標準級) 及 MOCC 專業級的認證試題, 科目包含微軟的 Office 應用程式 Word、Excel、PowerPoint 及 Access 等認證。

本研究採用其中的 Word 試題作為素材，PreMOUS 共有 91 題 MOCC 專業級共有 30 題。在差異上，PreMOUS 屬於單一能力項目型的題型，因此試題內容較為單純，而 MOCC 專業級屬於綜合能力項目類型的題型，因此在難度上 PreMOUS 必須具備基礎的實作能力，而 MOCC 專業級必須擁有全面及應用能力，其認證層次上有所差異。

PreMOUS 共有 24 個能力項目, 主要是參考國際認證 MOUS –Core (Microsoft Office User Specialist) 的能力項目分類所命製的試題。共有 24 項目的能力，條列如表 3-1 所示：

表 3-1　PreMOUS Word 的能力項目分類

能力項目碼	能力項目
1	版面設定
2	字型設定
3	段落設定
4	框線及網底、頁面框線設定
5	分隔設定
6	欄位設定
7	加入項目符號
8	建立表格
9	表格內容設定
10	在文件中插入圖片
11	使用定位點
12	插入符號
13	加入文字方塊
14	插入日期
15	段落首字放大
16	設定文件頁首及頁尾
17	使用繪圖工具
18	取代文字功能
19	文字藝術師
20	使用拼字檢查
21	插入 Excel 物件
22	文字格式轉換成表格
23	直書/橫書
24	使用複製、貼上功能

第二節 試題分類處理

PreMOUS Word 2000 試題，主要是參考電教協會撰寫的標準教材「PreMOUS Word 2000 認證主題式指定精選教材」的試題分類如表 3-2 所示：

表 3-2　試題檔案列表

編號	能力項目	試題檔名	題數
1	版面設定	WD2k0101A.doc ~ WD2k0106A.doc	6
2	字型設定	WD2k0201A.doc ~ WD2k0211A.doc	11
3	段落設定	WD2k0301A.doc ~ WD2k0311A.doc	11
4	框線及網底、頁面框線設定	WD2k0401A.doc ~ WD2k0406A.doc	6
5	分隔設定	WD2k0501A.doc ~ WD2k0502A.doc	2
6	欄位設定	WD2k0601A.doc ~ WD2k0603A.doc	3
7	加入項目符號	WD2k0701A.doc ~ WD2k0702A.doc	2
8	建立表格	WD2k0801A.doc ~ WD2k0803A.doc	3
9	表格內容設定	WD2k0901A.doc ~ WD2k0915A.doc	15
10	在文件中插入圖片	WD2k1001A.doc ~ WD2k1005A.doc	5
11	使用定位點	WD2k1101A.doc ~ WD2k1104A.doc	4
12	插入符號	WD2k1201A.doc	1
13	加入文字方塊	WD2k1301A.doc ~ WD2k1302A.doc	2
14	插入日期	WD2k1401A.doc	1
15	段落首字放大	WD2k1501A.doc ~ WD2k1502A.doc	2
16	設定文件頁首及頁尾	WD2k1601A.doc ~ WD2k1606A.doc	6
17	使用繪圖工具	WD2k1701A.doc	1
18	取代文字功能	WD2k1801A.doc	1
19	文字藝術師	WD2k1901A.doc	1
20	使用拼字檢查	WD2k2001A.doc	1
21	插入 Excel 物件	WD2k2101A.doc ~ WD2k2102A.doc	2
22	文字格式轉換成表格	WD2k2201A.doc	1
23	直書/橫書	WD2k2301A.doc	1
24	使用複製、貼上功能	WD2k2401A.doc ~ WD2k2403A.doc	3
		總計	91

PreMOUS 及 MOCC 標準級的試題，共分為三個部份：

(一) 試題資料檔：未完成的 Word 文件，檔案格式為 .DOC。

(二) 試題答案檔：經由作答需求的要求，所完成的標準答案檔，檔案格式為 .DOC。

(三) 作答需求檔：敘述如何處理試題資料檔，以完成試題答案檔的內容的步驟及要求。

在分類的意義上，試題資料檔及試題答案檔比較不具意義，而作答需求檔中，所包含的一些功能、專有名詞及作答動作的字詞，具有明顯特徵的意義，以 WD2k0207a.txt 的作答需求為例：

> 請將第一段字元格式為：「法治的國家」設定中文字體為新細明體，粗體，斜體，大小為 28 點，色彩為粉紅色。「民主」設定中文字體為新細明體，粗體，斜體，大小為 36 點，色彩為藍色。

由以上的例子可以看到許多如：「新細明體」、「粗體」、「斜體」….等專有的名詞，包含在文件之中，可以作為某一些能力項目的特徵關鍵字 (Feature Terms, FT)。然而也有許多與分類較無關係的特徵字詞，如：「法治」、「國家」…等，這些特徵字詞，也有可能在分類上造成干擾。

第三節 收集基礎詞庫及建立特徵關鍵字詞庫

由於試題主要的成份便是文字，因此要找出試題的特徵，也就是試題內所內含的專有名詞，動詞或形容詞是最具有參考價值的部份，這些字詞，本研究稱之為「特徵關鍵字詞 (Feature Terms, FT) 」。要從試題中擷取具有特徵意義的字詞，則必須有可靠的字詞剖析程序。

圖 3-1 為建立特徵關鍵字詞庫 (Feature Term Base, FTB) 的架構圖，此架構中的字詞剖析器必須仰賴「基礎字詞庫」的大量字詞比對樣本試題的文字內容，將比對到的字詞列入特徵關鍵字詞庫 (FTB)。

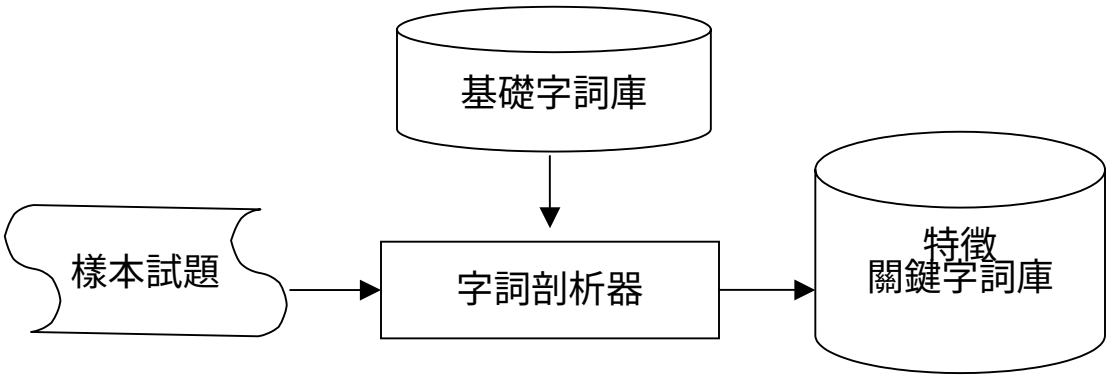

圖 3-1 建立特徵關鍵字詞庫的架構圖

一、基礎字詞庫

基礎字詞庫在此研究定義為:「由兩個以上的中文字組成，廣泛而未經分類且確實在實際用途上可使用的任何字詞」。目前開發詞庫的單位有「中華民國計算語言學學會」，大約有八萬目左右的中文詞知識

庫，而各廠商亦有其應用範圍內的詞庫。在此研究中直接採用 Windows 98/2000 的注音輸入法的相關字詞庫，當作本研究的基礎字詞庫，較為接近電腦應用軟體領域的範圍。相關字詞庫如圖 3-2 所示。

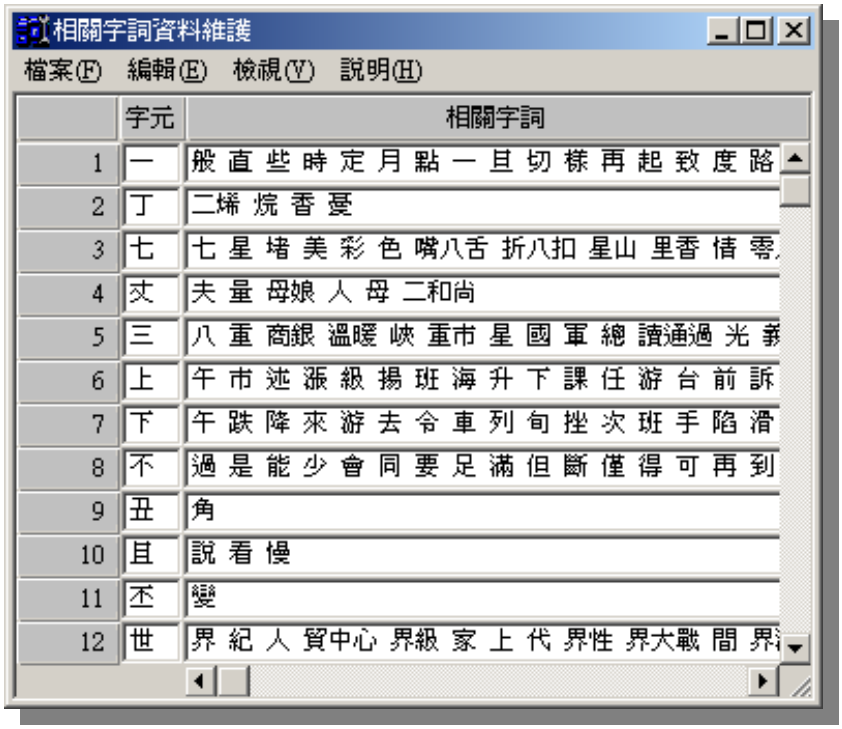

圖 3-2 相關字詞庫

該詞庫經整建後，可以得到 44,579 筆的字詞，足以取得常用字詞及電腦應用軟體領域的相關字詞。

二、字詞剖析器

字詞剖析器共分為四大部份分別有去除英數字及標點符號、斷詞處理、綴字處理及剩餘字詞處理，分別說明如下：

(一) 去除英數字及標點符號

由於英數字在中文的 Word 2000 較不具意義，因此在本研究不採納，並且將其視為與標點符號同一性質的資料而予以去除。

(二) 詞庫式斷詞處理

詞庫式斷詞法在基礎字詞庫完整的情況下，可以達到很好的斷詞品質及準確的擷取必要的 FT。主要使用的方法簡單的來說，是使用詞庫去比對試題，並將比對到的字詞擷取出來。

雖然說 Windows98/2000 的相關字詞庫已經可以擷取到很多字詞，但不能保證其足以涵蓋 Microsoft Word 2000 的所有領域。在詞庫式斷詞法處理之後，可能還會殘留一些字句，而這些字句可能仍然具有特徵意義，因此後續還要作綴詞處理及剩餘字詞處理，以免因為漏掉一些關鍵字詞，而使分類的品質降低。

(三) 綴字處理

因此在使用詞庫式斷詞法處理之後，接下來就要處理綴字，綴字大部份是一些介詞 (如:在、將、…)、代名詞(如:你、我、他…)或助動詞 (如:是、的…)，而這些文字並沒有特徵意義存在，但有

意義的詞句，往往被這些字詞隔開，因此一律予以去除之後，關鍵字詞自然就會出現，接著便進行剩餘字詞處理。

(四) 剩餘字詞處理

當完成綴字處理後，剩下一個獨立中文字的部份都予以去除，而二字以上的字詞，列入特徵關鍵字詞庫 (FTB) 中。雖然說這些關鍵字詞不見得有意義，但也有可能是沒列到基礎詞庫中，但卻有意義的關鍵字，不應當任意排除。

三、特徵關鍵字詞庫 (Feature Term Base, FTB)

透過字詞剖析器，便可以產生 FTB。FTB 為試題分類最主要的核心，其內容記載關鍵字詞在那個能力項目的出現的頻率，藉由字詞出現頻率 (Term Frequency) 的高低，評斷關鍵字詞在那個分類特徵最明顯，藉以判讀其能力項目分類。

第四節 字詞權重函數 (Term Weighted Function, TWF)

計算出該關鍵字詞在各能力項目中的權重，可以使用 TWF 由三方面來思考:

1. TF (Term Frequency): 關鍵字詞出現在能力項目的次數。
2. IDF (Inverse Document Frequency)：關鍵字詞所分佈的能力項目總數。
3. WIDF (Weighted Inverse Document Frequency)：考慮各類別字詞出現頻率差異的 IDF。

因為本研究的分類為能力項目，因此文獻探討中，使用文件類別 *d*，在此均改為能力項目 *a*。

字詞出現頻率 TF (Term Frequency) ，*TF(a,t)* 也就是關鍵字詞 *t* 出現在能力項目 *a* 上的頻率，可以判斷該關鍵字詞在該類文件的重要性。做為字詞權重函數可定義為：

$$W(a,t)=TF(a,t) \qquad (3\text{-}1)$$

其中

TF(a,t) ：能力項目 *a* 中出現關鍵詞 *t* 的權重

*TF(a,t)*值越大，代表出現於能力項目 *d* 的次數越多，因此其對於 *d* 能力項目的重要性也越大。然而，*TF* 並無法反映出完整的現象，舉例而言，關鍵字詞 *t* 在所有能力項目的 *TF* 值都大於 0，則對所有的能力項目而言這個關鍵字詞都具有重要性，但相對的該關鍵字詞就變得不

容易辨別；反之關鍵字詞 *t* 即使只有在唯一的一個能力項目 *a* 中出現一次，這樣反而會應該較容易辨別。因此，關鍵字詞是否普遍分佈程度應該要被考慮。

逆文件頻率 IDF (Inverse Document Frequency)主要是用來考慮關鍵字詞普遍分佈的程度，IDF*(t)*所表示的意義是關鍵字詞 *t* 是出現在幾個能力項目，其計算公式為:

$$IDF(t) = N/af(t) \qquad (3\text{-}2)$$

其中

N：代表文件的總數

af(t)：代表含有關鍵詞 t 的能力項目總數

而 *IDF(t)*的數值越小的話，代表關鍵字 *t* 的分佈越普遍，則其特徵意義就相對越低。舉例而言，若能力項目有 24 個，而 *IDF(t)*也等於 1/24，則代表所有能力項目都有出現關鍵字詞 *t*,則關鍵字詞 *t* 毫無特徵意義可言；反之 *IDF(t)*等於 1 的時候, 代表只有一個能力項目出現過關鍵字詞 *t*，其特徵意義就特別明顯。

*WIDF(a,t)*是採用關鍵字詞 *t* 出現在各能力項目次數的總和倒數，來表達其普遍性。*IDF(t)*固然可以考慮到普遍性，但是對於關鍵字詞都有出現，然而分佈的數量卻不平均的狀況便無法考慮。比如說，關鍵

字詞 t 在所有能力項目都出現一次，但在能力項目 a_n 卻出現 10 次，以這樣的情況而言使用 *IDF* 運算，則會將 t 在 a_n 特徵意義給衰減掉，因此將 *IDF* 改定義為出現在各能力項目次數總合的倒數，如此特徵意義就會更加明顯。

$$W(d,t)=TF(d,t) \cdot IDF(t) \qquad (3\text{-}3)$$

其中

$W(d,t)$：關鍵詞 t 在 d 文件類別的權數

$TF(d,t)$：關鍵詞出現率(Term Frequency)

$IDF(t)$：逆文件頻率(Inverse Document Frequency)

因此，將採用文獻探討中所論述的公式，而字詞權重函數採用下列六種，以比較最佳的運算方式:

(一) $W(a,t)=TF(a,t) \cdot IDF(t)$ (3-4)

(二) $W(a,t)=IDF(t) = N/af(t)$ (3-5)

(三) $W(a,t)=TF(a,t) \cdot IDF(t)$ (3-6)

(四) $W(a,t)=TF(a,t) \cdot IDF(t)^2$ (3-7)

(五) $W(a,t)=WIDF(a,t) = \dfrac{1}{\sum_{i \in D} TF(i,t)}$ (3-8)

第五節 權重向量矩陣

關鍵字詞權重計算，主要是參考權重向量矩陣，而權重向量矩陣都是透過特徵關鍵字詞庫來演算所產上的向量表。我們以列為關鍵字詞 T，以行能力項目 A ，能力項目與關鍵字詞之交為其權重，該權重值可能是，如此可以構成表 3-3 的架構:

表 3-3 權重向量矩陣

	A_1	A_2	A_3	…	A_j
T_1	W_{11}	W_{12}	W_{13}	…	W_{1j}
T_2	W_{21}	W_{22}	W_{23}	…	W_{2j}
T_3	W_{31}	W_{32}	W_{33}	…	W_{3j}
⋮	⋮	⋮	⋮	…	⋮
T_i	W_{i1}	W_{i2}	W_{i3}	…	W_{ij}

輸入一道試題 Q，經過字詞剖析程序以後，得到 m 組的關鍵字詞，而各關鍵字詞的出現頻率 fr_1~fr_m，而構成輸入試題的權重向量矩陣 Q:

$$Q = [fr_1 \quad fr_2 \quad fr_3 \quad ... \quad fr_m] \qquad (3\text{-}9)$$

其中

Q:輸入試題的權重向量

fr_m：輸入試題第 m 個關鍵字詞的出現頻率

承 3-9 式，使用輸入試題的關鍵字詞共有 m 個，當使用來篩選權重向量矩陣，令 j 為總能力項目數，則得到權重向量矩陣 M 為

$$M = \begin{bmatrix} W_{11} & W_{12} & W_{13} & \cdots & W_{1j} \\ W_{21} & W_{22} & W_{23} & \cdots & W_{2j} \\ W_{31} & W_{32} & W_{33} & \cdots & W_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{m1} & W_{m2} & W_{m3} & \cdots & W_{mj} \end{bmatrix}$$

(3-10)

其中

M：使用輸入試題關鍵字詞篩選而成的權重向量矩陣

W_{mj}：第 m 個關鍵字詞在第 j 個能力項目的權重

承 3-10 式，若輸入試題的關鍵字詞 i 在能力項目 k 的權重 tw_{ik}，關鍵字詞 i 在能力項目 k 個出現頻率為 fr_{ik}，則，得公式如下:

$$tw_{ik} = fr_i \cdot W_{ik} \qquad (3\text{-}11)$$

其中

tw_{ik} ： 輸入試題的關鍵字詞 i 在能力項目 k 的權重

fr_i ：輸入試題的關鍵字詞 i 的出現頻率

W_{ik}：關鍵字詞 i 在能力項目 k 的權重

承(3-11)式，令輸入試題在能力項目 k 的字詞權重為 S_k，則

$$S_k = \sum_{i=1}^{m} fr_i \cdot W_{ik} = \sum_{i=1}^{m} tw_{ik} \qquad (3\text{-}12)$$

其中

S_k：輸入試題在能力項目 k 的字詞權重

m：輸入試題的關鍵字詞總數

由 3-9、3-10、3-11，得到最後可的道試題Q在各能力項目的權重矩陣 M_Q

$$M_Q = [S_1 \quad S_2 \quad S_3 \quad ... \quad S_j] = Q \cdot M \qquad (3\text{-}13)$$

其中

M_Q：經權重運算後，得到的試題 Q 的權重矩陣

S_k ：單一試題在能力項目 k 的權重

最後判讀 M_Q 矩陣中的各元素中找到最大值 D，公式如 3-14 所示。

$$D = Max(M_Q) \tag{3-14}$$

其中

D ：所判讀到的能力項目權重最大值

M_Q：Q試題經過權重運算後的權重矩陣

假設 $D=S_n$，表示 S_n 為 M_Q 矩陣中的最大值，則可以得到 n 為試題 Q 的能力項目分類。

向量矩陣內的向量值必須透過字詞權重運算公式產生，使用權重向量矩陣作能力項目分類的架構如圖 3-3：

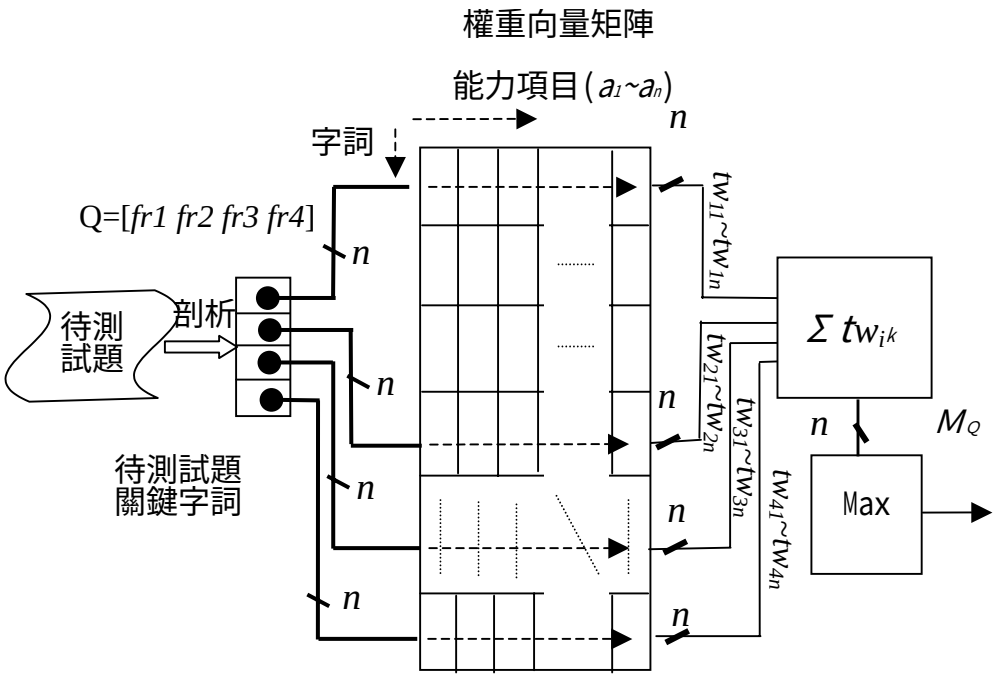

圖 3-3 權重向量矩陣的運算模式

參考圖 3-3, 待測試題經過剖析之後, 假設得到 4 組關鍵字詞 t_1~t_4, 假設能力項目有 n 項，則 $t1$ 與 $W_{11},W_{12},W_{13},\ldots,W_{1n}$ 相乘得 $tw_{11},tw_{12},tw_{13},\ldots,tw_{1n}$，以同樣的方法，分別得到 $tw_{21},tw_{22},tw_{23},\ldots,tw_{2n}$、$tw_{31},tw_{32},tw_{33},\ldots,tw_{3n}$ 及 $tw_{41},tw_{42},tw_{43},\ldots,tw_{4n}$ 的結果，最後再取得各能力項目的加總 $tw_1,tw_2,tw_3,\ldots,tw_n$，也就是

$$S_i = \sum_{j=1}^{4} tw_{ji} \quad , \quad 1 \le i \le n \qquad (3\text{-}15)$$

最後再由 S_1~S_n 找出最大值，如 S_8 為最大值，則第 8 項便是判讀的能力項目。

舉例而言，讀取 Wd2k0901a.txt 的試題，試題內容如下：

> 請將表格第一列的前 5 格合併，而設定 ” 2001 一月 ” 文字，字體大小 60 點，儲存格左右對齊。

得到該試題所使用的關鍵字詞及其出現的次數如下表:

表 3-4　Wd2k0901a.txt 各關鍵字的出現次數

關鍵字詞	出現次數
字型	1
表格	1
背景	1
粗體	1
設定	2
黃色	1
填滿	1
對齊	1
標題	1

表 3-5 經過篩選後的權重向量表

關鍵字詞	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
字型		.03							.01				.05	.04		.04			.05					
表格								.24	.09		.02											.10		.05
背景									.02	.01														
粗體		.03							.01						.06									
設定	.14	.05	.04	.09	.03		.03		.07	.02	.05		.02	.04			.04	.09			.03			
黃色				.02					.01															
填滿				.04					.01				.02			.01	.04							
對齊			.08						.03		.02		.02											.03
標題		.06	.02	.05	.03	.06	.03		.02			.09	.02	.04					.05			.10		

則運算結果得:

$$D = Max(S) = 0.32$$

而 0.32 為第 9 項『表格內容設定』能力項目的權重，另外能力項目第 1 項及第 8 項，分別列於第二及第三的順位，屬於其他較為突出的能力項目，亦可解釋為隱含的能力項目。

第六節　回歸率

由上一節的計算，可以得到 D=0.32，其對應的能力項目為 9，然而分類的結果並不保證絕對正確。為了得知其分類的效率為何，可藉由比對特徵關鍵字詞庫試題來源原先的能力項目分類，以了解所有的試題是否都可以回歸到本身的試題分類。分類正確的試題數量與總試題數量之比，即是回歸率(Recall Rate)。

$$RR = \frac{C_{right}}{C_{total}} \tag{3-16}$$

式中

RR：回歸率

C_{right}：回歸到本身能力項目的試題數量

C_{total}：總試題數量

藉由回歸率可以得知分類機制的優劣，也可以作為比較各 *TWF* 優劣的參考指標。