第三章　研究方法與工具

本章係依據前述研究目的及相關文獻與理論，依序說明本研究之方法及工具，研究方法包括中文斷詞、潛在語意分析與相似試題之分類。

第一節 研究方法

一、中文斷詞

在利用潛在語意分析分析題庫中試題的相似性之前，因為處理的為中文試題，因此必須先做斷詞，本研究所使用的的斷詞法為反向最大匹配法，中文辭典的詞彙最長為 10 個字，英文辭典則由題庫中統計所產生，斷詞之流程如下：

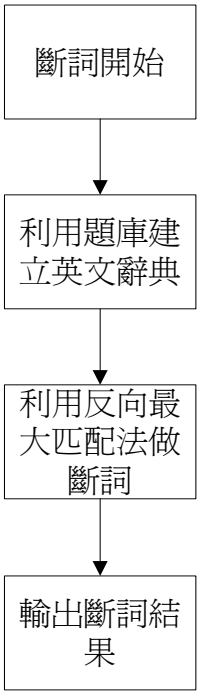

圖 3-1 斷詞流程圖

二、比較潛在語意分析與向量空間模型

本研究使用之題庫為行政院勞工委員會所編製的「電腦軟體應用技能檢定丙級學科」92 年度和 93 年度兩個版本的題庫共 1000 題單選題，由於尚無研究對此題庫之試題作相似度之分析，因此無法得知使用潛在語意分析所得到的結果是否較佳，故本研究使用向量空間模型（Vector Space Model ,VSM）和潛在語意分析判斷此題庫試題間之相似度，以比較何者效果較佳。

三、潛在語意分析的流程

斷詞完成後，再來是將斷詞的結果，利用潛在語意分析的方法求試題間的相似度，其流程如下：

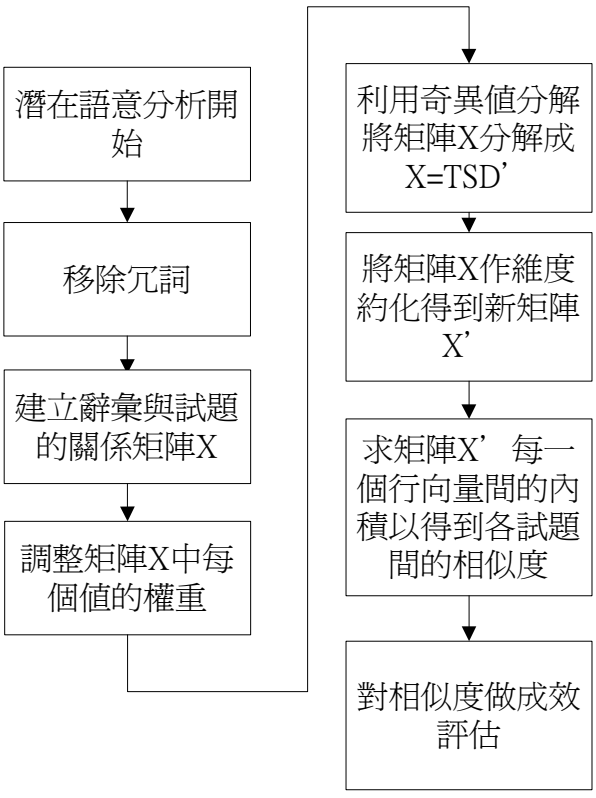

圖 3-2 潛在語意分析流程

以下說明圖 3-2 的每個步驟：

1. 移除冗詞

冗詞(common words)是指在文件中出現頻率較高，但不具重要意義的詞，如標點符號、數字、語助詞、介系詞等，英文的「the、is、by」等，和中文的「的、了、是」等，都屬於冗詞。

在英文的研究中，已建立冗詞的索引表（Fox,1990），但中文目前尚未有較標準的冗詞索引表，因此本研究參考英文冗詞索引表的內容，與題庫中出現頻率較高的辭彙，以建立中文冗詞索引表，然後依據此索引表移除文件的冗詞。

2. 建立辭彙與試題的關係矩陣 X

移除冗詞後，接著統計其餘詞彙在試題中出現的次數，以建立詞彙與試題的關係矩陣 X，矩陣 X 每一列為這個詞彙在每個試題中出現的次數，每一行為這個試題中出現的詞彙次數，如表 2-1。

觀察矩陣 X，會發現建立起來的矩陣為稀疏矩陣，為了節省儲存的空間，本研究使用 Harwell-Boeing sparse matrix 來儲存此矩陣（Duff, Grimes, &Lewis, 1992），Harwell-Boeing sparse matrix 利用三個陣列來儲存稀疏矩陣中非零的資料和位置，設此三個陣列為 VALUES、ROW_IND 與 COL_PTR，VALUE 記錄矩陣中非零的值，ROW_IND 記錄列的索引值（row indices），COL_PTR 記錄行的指標（column pointer ），其儲存的方式如下：

VALUES：儲存矩陣中非零的值，其儲存方式為先存第 1 行非零的值，之後再存第 2 行，依序存到第 n 行。

ROW_IND：記錄矩陣中非零的值在矩陣的第幾列。

COL_PTR：記錄矩陣中每一行的第一個非零的值在陣列 VALUES 中的位址。

以下舉例說明 Harwell-Boeing sparse matrix 的儲存方式：

假設有一矩陣 X 為：

$$\begin{pmatrix} 1 & -3 & 0 & -1 & 0 \\ 0 & 0 & -2 & 0 & 3 \\ 2 & 0 & 0 & -4 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 5 & 0 & -5 & 0 & 6 \end{pmatrix}$$

可求得其 Harwell-Boeing sparse matrix 如表 3-1：

表 3-1 Harwell-Boeing sparse matrix 範例

ID	1	2	3	4	5	6	7	8	9	10	11
VALUES	1	2	5	-3	4	-2	-5	-1	-4	3	6
ROW_IND	1	3	5	1	4	2	5	1	3	2	5
COL_PTR	1	4	6	8	10						

以下說明如何將 Harwell-Boeing sparse matrix 轉換成原始矩陣：

假設要求第 4 行的值，ID 4 的 COL_PTR 為 8，ID 5 的 COL_PTR 為 10，由定義可知 VALUES 中第 8 和第 9 的-1 和-4 即為原始矩陣中第 4 行的值，再來比較相對應的 ROW_IND，可知 -1 和-4 在原始矩陣中是在第 1 列和第 3 列，依此方法即可將 Harwell-Boeing sparse

matrix 轉換成原始矩陣。

3. 調整矩陣 X 中每個值的權重

建立詞彙和試題的關係矩陣後，由於此矩陣只統計每個詞彙在試題中出現的次數，因此出現頻率為此系統判斷詞彙重要性的唯一指標，如此對於有些詞彙雖然經常出現，但其重要性並不高，如「下列、何者、屬於」等詞彙，或者雖然出現次數不多，但對試題占有重要影響者，如「剪貼簿、試算表、列印」等詞彙，會有誤判的情形，因此在做奇異值分解前，須先調整矩陣中每個值的權重。

權重的調整通常由 local 權重和 global 權重兩部份所組成(Dumais,1991)， local 權重指的是詞彙在試題的重要性，基本上詞彙在單一試題中出現次數愈多，表示其在此試題的重要性高，global 權重指的是詞彙在題庫的重要性，詞彙在題庫中出現次數愈多，表示其在整個題庫的重要性低，求得 local 和 global 的權重後，再利用兩者的積來調整詞彙的權重，即將矩陣 X 的每個值換成

L(i,j)×G(i)，

L(i,j)是詞彙 i 在試題 j local 的權重。

G(i)是詞彙 i global 的權重。

L(i,j)可以用以下三種方式求得(Salton & Buckley, 1988; Harman, 1992)：

(1) binary: $L(i,j) = x(tf_{ij}) = \begin{cases} 0, tf_{ij} = 0 \\ 1, tf_{ij} > 0 \end{cases}$

(2) term frequency(tf): $L(i,j) = tf_{ij}$ ，tf_{ij} 表示詞彙 i 在文件 j 出現的次數。

（3）log：$L(i,j)=\log_2(tf_{ij}+1)$，取 log 是讓 tf_{ij}之間的值不會相差太大，加 1 是為了避免$\log_2 0$的情況。

G(i)可以用以下四種方式求得(Salton & Buckley, 1988; Dumais,1991)：

（1）normal：$G(i)=\frac{1}{\sqrt{\left(\sum_j tf_{ij}^{\ 2}\right)}}$

（2）inverse document frequency(idf)：$G(i)=\log_2(\frac{n}{\sum_j x(tf_{ij})})+1$，n 為文件的數量。

（3）idf squared(idf2)：$G(i)=\log_2(\frac{n}{\sum_j (x(tf_{ij}))^2})+1$

（4）entropy：$1+\sum_j \frac{p_{ij}\log_2(p_{ij})}{\log_2 n}, p_{ij}=\frac{tf_{ij}}{gf_i}$，$gf_i$是詞彙 i 在所有文件中出現次數的總和。

本研究嘗試分析各種 local 權重與 global 權重的組合，以找出效果最好的組合。

4. 利用奇異值分解將矩陣 X 分解成 X=TSD'

奇異值分解的部份，我們使用的工具為 General Text Parser(GTP)（Giles, Wo, & Berry, 2001），此工具能快速的將矩陣 X 分解成 T、S、D 三個矩陣。

5. 將矩陣 X 作維度約化得到新矩陣 X'

維度約化是將奇異值分解後所得到的三個矩陣T_r、S_r、D_r，保留 k 個維度後相乘以得到新的矩陣 X' ，如此可消除語意空間中的雜訊（Landauer, Foltz, &Laham, 1998）。由於要保留多少個維度 k，並沒有理論上的最佳值(Wang & Nie,2003)，必須經由實驗的方式找出不同文件集最佳的 k 值，因此我們比較不同的 k 值以找出成效最佳者。

6. 求矩陣 X' 每一個行向量間的內積以得到各試題間的相似度

要得到兩個試題間的相似度，可計算矩陣中行向量(column vector)的內積值（inner product），計算方式如下（Bellegarda,2000）：

$$\mathrm{sim}(d_i, d_j) = \cos(v_i, v_j) = \frac{v_i v_j^{t}}{\| v_i \| \| v_j \|}$$

d_i、d_j為試題i和試題j，$sim(d_i, d_j)$為兩試題的相似度，且$0 \le sim(d_i, d_j) \le 1$，$v_i$、$v_j$為兩試題在矩陣X' 的行向量。

7. 利用精確率(precision rate)與召回率(recall rate)做成效評估

研究者先分析題庫找出相似試題，再將使用 LSA 所求得的相似試題，取相似度最大的 N 個（N 設為 1000，並以 50 為單位），分析不同 N 值的精確率(precision rate)與召回率(recall rate)。

精確率是利用 LSA 的方法所找出的相似試題，真正為相似試題的比率，召回率為題庫中所有相似的試題，利用 LSA 的方法，被找出的比率。

精確率和召回率的公式如下：

$$精確率=\frac{R_a}{A}$$

$$召回率=\frac{R_a}{R}$$

R_a 為利用 LSA 所找出的相似試題中，為正確的相似試題的數量。

A 為利用 LSA 所找出的相似試題的數量。

R 為題庫中所有相似試題的數量，相似與否由研究者事先判斷之。

四、相似試題之分類

本研究使用之題庫為行政院勞工委員會所編製的「電腦軟體應用技能檢定丙級學科」92 年度和 93 年度兩個版本的題庫共 1000 題單選題，研究者並依試題間相似的程度，將其分為「完全相同」、「非常相似」、「部份相似」與「些微相似」四類。

完全相同的試題有三類：(一)敘述方式完全相同的試題、(二)敘述方式不同但題意相同的試題、(三)部分辭彙不同，但意義相同的試題，以下舉例說明：

(一)敘述方式完全相同的試題，例如：

92 年度第 727 題：「下列何者是多人多工的作業系統？」

93 年度第 726 題：「下列何者是多人多工的作業系統？」

或

92 年度第 886 題：「電腦病毒的侵入是屬於」

93 年度第 885 題：「電腦病毒的侵入是屬於」

(二)敘述方式不同但題意相同的試題，例如：

92 年度第 591 題：「電腦執行數值運算的速度受到下列何者影響？」

93 年度第 590 題：「下列何者會影響電腦執行數值運算的速度？」

或

92 年度第 634 題：「一般編寫程式的流程爲」

93 年度第 633 題：「編寫程式的一般流程爲何？」

(三)部分辭彙不同，但意義相同的試題，例如：

92 年度第 876 題：「預防電腦病毒，下列敘述何者有誤？」

92 年度第 940 題：「避免電腦中毒的方法，下列何者不正確？」

或

92 年度第 576 題：「二進制 1011，1001，1100，0011 以十六進制表示爲」

92 年度第 615 題：「二進制數值 1101001 轉換爲十六進制時，其值爲」

非常相似的試題有兩類：(一)敘述方式完全相同，但有一個關鍵詞不同的試題、(二)敘述方式不同但意義相似，且有一個關鍵詞不同，但此關鍵詞意義相似的試題，以下舉例說明：

(一)敘述方式完全相同，但有一個關鍵詞不同的試題

在「電腦軟體應用技能檢定丙級學科」的試題中，因為軟硬體的升級或改版，所以將 92 年和 93 年的試題合併成一個題庫時，有部分試題敘述方式相同，但有一個關鍵詞不同，此類試題組合研究者視為「非常相似」，例如：

92 年度第 832 題：「在 Windows 98 中，在中文輸入法中，要切換全型和半型輸入，預設值爲按下列那一個按鍵？」

93 年度第 831 題：「在 Windows XP 中，在中文輸入法中，要切換全型和半型輸入，預設值爲按下列那一個按鍵？」

或

92 年度第 692 題：「在 Word 中，若對已被選取之文件，先按住[Ctrl]鍵再做拖曳，表示執行」

93 年度第 691 題：「在 Word XP 中，若對已被選取之文件，先按住[Ctrl]鍵再做拖曳，表示執行下列那一個動作？」

(二)敘述方式不同但意義相似，且有一個關鍵詞不同，但此關鍵詞意義相似的試題，例如：

92 年度第 954 題：「對於電腦病毒的防治方式下列何者是錯誤的？」

93 年度第 948 題：「對於「防治電腦病毒」的敘述中，下列何者正確？」

或

92 年度第 463 題：「SET 是目前公認 Internet 上的電子交易安全標準，下列哪一公司未參與 SET 之發展？」

92 年度第 963 題：「下列何者不是一個完整的安全電子交易 SET 架構所包括的成員之一？」

部份相似的試題有兩類：(一) 敘述方式完全相同，但有二個關鍵詞不同的試題、(二) 敘述方式不同但意義相似，且有二個關鍵詞不同，但此關鍵詞意義相似的試題，以下舉例說明：

(一) 敘述方式完全相同，但有二個關鍵詞不同的試題，例如：

92 年度第 831 題：「在 Windows 98 中，在中文輸入法中，要切換不同的輸入法，預設值爲按下列那一個按鍵？」

93 年度第 831 題：「在 Windows XP 中，在中文輸入法中，要切換全型和半型輸入，預設值爲按下列那一個按鍵？」

或

93 年度第 876 題：「關於「預防電腦病毒的措施」之敘述中，下列何種方式較不適用？」

93 年度第 948 題：「對於「防治電腦病毒」的敘述中，下列何者正確？」

（二）敘述方式不同但意義相似，且有二個關鍵詞不同，但此關鍵詞意義相似的試題，例如：

92 年度第 113 題：「十進制 60.875 以二進制表示爲」

92 年度第 115 題：「二進制 1011，1001，1100，0011 以十六進制表示爲」

或

92 年度第 244 題：「Windows 98 的檔案總管與我的電腦的功能表中的選單，不同的是檔案總管多了那一個功能表？」

92 年度第 323 題：「在 Windows98 中，如果要在檔案總管中讓所有檔案名稱以大寫來顯示，則應在那一個功能表中設定？」

些微相似指的是同一主題的試題，如「Windows XP」、「記憶體」、「電腦病毒」等主題，以下舉例說明：

93 年度第 795 題：「在 Windows XP 下，下列何者爲預設的字型」

93 年度第 857 題：「在 Windows XP 中，要關閉一個作用中的視窗，可以使用以下那一個按鍵」

92 年度第 95 題：「一般說來，下面那一種記憶體速度最快？」

93 年度第 613 題：「一般而言，我們說 PC 有 4MB 主記憶體，指的是」

或

92 年度第 441 題：「何者可能會感染電腦病毒？」

92 年度第 467 題：「下列何者不是電腦病毒的分類之一？」

上面這三組試題，主題依序為「Windows XP」、「記憶體」、「電腦病毒」，研究者在判斷時，將這類試題視為些微相似。

第二節 研究工具

在研究工具方面，本研究使用之題庫為行政院勞工委員會所編製的「電腦軟體應用技能檢定丙級學科」92 年度和 93 年度兩個版本的題庫共 1000 題單選題，研究者並依試題間相似的程度，將其分為「完全相同」、「非常相似」、「部份相似」「些微相似」四類。

另外研究者再自行開發中文斷詞系統與潛在語意分析系統，中文斷詞系統中所用的中文辭典為台灣師範大學資訊工程研究所陳柏琳教授所建立的中文辭典，共 72645 個詞彙；潛在語意分析系統中奇異值分解的部份，則使用 Giles 所發展的 GTP (General Text Parser) (Giles, Wo, Berry, 2001)。GTP 是一個以 Java 發展而成的套裝軟體，除了可做奇異值分解外，另外也有資料檢索的功能。