第三章　系統架構

想要正確解析一筆書目資料，第一步就是要盡可能了解每一個字所代表的意義。但是畢盡知識是無窮盡的，就算知道了每一個字的意義，有時還是必須知道整個句子的結構才有辦法作出正確的解析。因此必須要有一個有彈性的做法，可以隨時輕易加入新格式，並且兼顧語意及語法來解析後設資料。為了達成這個目標，我們以基因序列紀錄書目資料結構，並且對於一些字義也作了判斷。使得本系統可說是針對語意及語法兩方面來解析每一筆資料。系統中使用 BLAST 作字串比對。使用 BLAST 可說是一項意外的收穫，因為 BLAST 不僅可以快速正確比對基因序列，而且可以輕易在資料庫中加入新的基因序列。這項優點讓整個系統建立的過程事半功倍。以下針對系統作更進一步的詳述。

3.1 系統流程

系統流程如圖 3.1 所示，主要可分為四個步驟：

- 步驟一： 盡可能辨別書目資料中每一個字所代表的意義，並且將書目資料轉換成基因格式，例如本系統會將‘Jewell, M (2002) Title’轉換成‘ARMIYK’這樣的基因格式，詳細的基因序列轉換規則會在 3.3 節中介紹，至於字義解析在 3.2 節中有進一步的討論。

- 步驟二： 對於由‘A’所包圍的區域只留下一個‘A’當代表，例如將“ AQAQAAARRSW ”以“ ARRSW ”來取代，因為系統辨別的是區間的位置，所以重複出現的基因對於系統的判斷上並沒有幫助，反而會對下個步驟的處理造成困擾，因此將

這些冗餘部份先行刪除。

- 步驟三： 將第二步驟處理過後的基因序列輸入 BLAST，作蛋白質序列比對，從建立好的資料庫中找出最相似的序列。因為 BLAST 是以 Blosum62 作為預設的計分表(scoring) table ，這個計分表是根據蛋白質的特性以及演化的程度而設定的[18]。所以為了本系統的需求我們也重新設計一個計分表以使本系統準確率達到較佳效果，對於重新設計的計分表在 3.4 節會有解說。3.5 節會介紹我們以人工為本系統建立樣板資料庫的方法。

- 步驟四： 以步驟三得到的基因序列為依據，將原始書目資料根據此序列樣板作欄位解析。找出每一個區域的起訖點，並且賦予每一個區域它所代表的欄位。確切作法在 3.6 節進一步解說。得到這些後設資料後，可將有價值的部分經過人工確認後存入資料庫。例如論文出處這項資料，如果它日收集到可觀的程度也可作知識庫使用。

經過上訴四個步驟的運算，書目資料的後設資料就可以正確的被解析出來，圖 3.2 是本系統的解析結果範例。

3.2 知識背景

在資訊檢索的領域中，知識時常被當作解決問的工具。但是想要擁有所有知識是很困難的。本系統在可獲得的知識中，捨棄了標題名稱以及論文名稱這兩樣知識。以期刊名稱為例，雖然我們已經蒐集超過 8400 筆期刊名稱，可是在本系統上的實際測試卻只能用來辨認百分之四十的期刊名稱。而且蒐集的速度也比不上新增的速度。期刊名稱已經是如此更何況是論文名稱。既然如此決定還是忍痛不用這些知

識，以免系統準確度隨著日子一天天經過而遞減。使用這兩類知識可能還會造成其它困擾。例如論文名稱及期刊名稱難免會有些共用的字，這會產生模稜兩可的情形，而且論文名稱及期刊名稱通常是並列的，這會使得無從分辨這兩個欄位。因此本系統只使用了下列五類知識：

- 中文姓氏

這項知識主要是從外交部所提供的護照外文拼音參考資料中取得，外交部提供的資料有四百多筆中文姓氏英文拼字，而扣除同音字以及一些拼字在英文具有意義的姓氏如“I”、“Sun”，…等，有 200 多筆資料，不過還有一些常見的姓氏英文拼字並非如同外交部所提供的資料，再加上這些資料總共約有 300 筆左右的姓氏英文拼字資料。

- 月份

所有月份的英文單字都在蒐集的知識庫中，以及任何有可能是月份的縮寫格式。

- 數字

辨別是不是數字非常簡單，所有程式幾乎都有提供這類函式，但是從知識的領域只能判斷是不是數字，而無法得知真正代表的意義。不過有些特殊的格式還是可以作些猜測，本系統將任何以 19 或 20 開頭的四位數字都當成年份，例如 1990、2001、…等，不過這些數字也有可能是代表頁數，但是畢盡只是少數。

- 關鍵字

有些文字在書目資料中經常出現，而且也代表某些特殊意義，本系統將這些文字列入蒐集的範圍之內，有‘page’、‘volume’、‘number’、

‘issue’等關鍵字，任何跟它們有關的縮寫及變形也列入本系統蒐集之列，例如很多書目資料是以‘pp’代表‘pages’，這些都會成為本系統的基本知識。

3.3 基因序列轉換規則

人體組織是由蛋白質構成，而蛋白質是由二十種氨基酸所組成，為了可以清楚表達蛋白質的合成基因，一般都是以一個英文字母代表一種氨基酸，如此就可以得到我們常見的蛋白質的基因序列。本系統也充分利用了這個特性，將這二十個字母分別對應到一種欄位或一種特殊字義，以 3.2 節所提到的基本知識為基礎，將書目資料轉為基因序列，以下就是轉換的規則：

A

對應到姓氏，也就是作者所在的區域。

R

對應到標點符號“，”或“；”。

N

對應到數字。

D

對應到標點符號“ . ”。

C

對應到關鍵字“ issue ”。

Q

對應到“，”、“；”、“.”、“(”、“)”以外的標點符號。

E

對應到 issue 值。

G

對應到標點符號“〃”。

H

對應到 page 值。

I

對應到標點符號“(”。

L

對應到期刊名稱。

K

對應到標點符號“)”。

M

對應到月份。

F

對應到 volume 值。

P

對應到關鍵字“ page ”。

S

對應到關鍵字“ number ”。

T

對應到論文名稱。

W

對應到 number 值。

Y

對應到年份。

V

對應到關鍵字“ volume ”。

基因格式轉換的過程中，系統將辨識出字義的字轉換成相對應的基因，無法辨識的就轉成 NULL，經過轉換之後書目資料就成為一串基因序列。不過“ T ”、“ L ”、“ E ”、“ F ”、“ H ”、“ W ”這幾個字將不會出現在基因格式之中，因為基本知識庫中沒有這些欄位的知識。不過這些字將會用在 template，因為在建立 template 資料庫時，會將“T”及“L”以人工的方式加入書目資料基因格式中，將“N”以人工方式轉換成“E”或“F”或“H”或“W”。

3.4 計分表(Scoring table)

為了將比對的結果量化，BLAST 使用計分表紀錄字母間互相配對的分數，常見的計分表有 Blosum 及 Pam，它們都是根據氨基酸特

性而設計的，在 Blosum 及 Pam 後面會有數字代表基因演化的程度，例如 BLAST 預設使用的計分表是 Blosum62，經過演化時間越長的基因序列，比對時就要使用後面數字越大的計分表，如此才使能得到最佳的比對結果，但是這樣的計分表並不適用於本系統，因此經過重新設計的計分表如表 3.1 所示。

從圖 3.3 可看出，對角線上的節點都給予正分，代表相同字母互相對應時給予正面評價，其中以姓氏相互對應給予最高得分，這是因為作者欄位在書目資料中佔據的區域相對較大，對於比對的影響也比較大，相對的如果是錯誤的比對是在作者欄位上也要給予最多的負分，同樣的論文名稱及期刊名稱也是如此。

3.5 樣板(template)

圖 3.4 顯示樣板資料庫所儲存的基因資料庫。每一筆資料的開頭是以符號“>”為起點。原本符號“>”後面應該要接著基因名稱及敘述，然後接下來幾行再存入基因序列。但是在這個系統中，“>”後面不再接著基因名稱，而是存入該筆基因序列所對應的正確書目蛋白質序列。如此 BLAST 在比較所有的基因之後，輸出最相近的序列時，會將該筆書目蛋白質序列所對應的正確蛋白質序列，輸出在蛋白名稱的位置。此時就可以擷取下來利用，解析出正確的後設資料。建立樣板資料庫主要分成五個步驟：

- 步驟一：將書目資料轉換成 brief form。
- 步驟二：在 brief form 中論文名稱的位置插入“T”。
- 步驟三：在 brief form 中期刊名稱的位置插入“L”。
- 步驟四：尋找 brief form 中“N”的位置，再找出書目資料中相

同位置的數字。判斷此處“N”所代表的意義，將 brief form 中的“N”以正確的的基因替換。

- 步驟五：將 brief form 存入樣板資料庫中基因序列的位置，將所對應的正確蛋白質序列存入基因名稱的位置。

圖 3.2 中的 brief form 與樣板序列不同之處在於樣板序列比 brief form 多了“T”和“L”，“T”代表論文名稱，“L”代表期刊名稱，它們都被插入正確的位置。而第一個“N”被轉為代表 volume 值的“F”，第二個“N”則被轉換成代表 number 值的“W”， 而這樣的轉換動作目前還是須由人工來執行。

3.6 Pattern Extraction

在書目資料基因序列轉換過程中，系統建立一份與書目資料相同長度的陣列來儲存書目蛋白質序列(如圖 3.5 所示)。我們以 B_i 代表去除冗餘基因後，書目蛋白質序列中第 i 個基因，以 C_i 代表書目資料中的第 i 個 token。因此 B_i 與 C_i 呈現一個對應一個的關係，不過 B_i 中的值有可能會是 NULL。經過 3.1 節前三個步驟處理過後可得到樣板序列，我們同樣以 T_i 代表樣板序列中第 i 個基因。Pattern extraction 可分成下列四個步驟(i、j、k、m、n 初始值都是 0)：

- 步驟一：由 T_k 目前位置往後找出第一個 p，使得 T_p 為“A”、“T”、“L”其中之一。將 k 設為 p。

- 步驟二：由 B_j 目前位置往後找出第一個 p，使得 $B_p=T_k$，如果 p 不存在則找出第一個 p 使得 B_p=NULL。將 j 設為 p。

- 步驟三：接續步驟二 B_j 的位置往前找，找出第一個 m 使得 $B_m=T_{k-1}$。由步驟二 B_j 的位置往後找，找出第一個 n 使得 $B_n=T_{k+1}$。

如此 C_{m+1} 至 C_{n-1} 的 tokens 就成為一組後設資料。重複步驟一、步驟二、步驟三，直到 T_k 的最後一個位置。

- 步驟四：找出所有 B_i=“N”，從樣板序列中找出 B_i 所代表的意義，再對應到 C_i 中，得到後設資料。以及找出 B_i=“Y”，則 C_i 代表年份。找出 B_i=“M”，則 C_i 代表月份

以圖 3.2 為例，要找出作者群的位置可先從樣板序列中找出“A”的相對位置。“A”在樣板序列中在第一個位置，後面跟著“RG”。因此只要在圖 3.5 中找到“A”的位置，那麼“A”前面所有的字，以及“A”後面一直到“RG”之前，可連接成一區域。這區域就是作者群的位置。同樣要尋找著作名稱也是如此，先從樣板序列找出“T”的相對位置，再從 3.5 圖找出第一個 NULL 的位置，然後根據前述的方法左右觀察就可以找到著作名稱的區域。接下來以同樣的方法找出期刊名稱的區域。在處理“N”之前，我們知道“N”可能是代表頁數的“H”，也可能是代表 volume 值“F”，也有可能是代表 issue 值的“E”，或者是代表 number 值的“W”，所以根據得 3.2 節步驟三所得到的樣板序列可以得知，第一個“N”是“F”也就是 volume 值，第二個“N”是“W”也就是 number 值，最後從 3.5 圖找出“M”及“Y”的位置，就可找出月份及年份，這樣就可正確抽出所有 pattern。

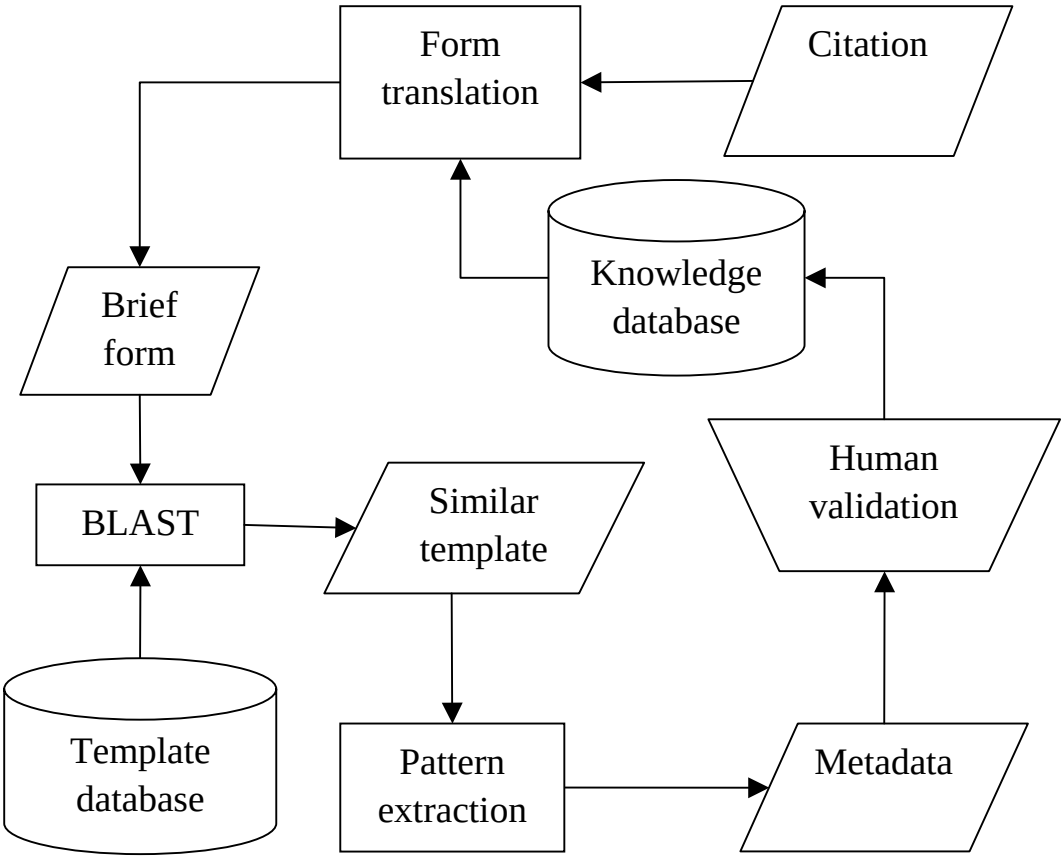

圖 3.1 系統流程圖

圖 3.2 解析結果輸出

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -4
R -9  4  0 -2 -3 -4  0 -2  0 -6 -2 -6 -2 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -9  0  9  1 -3  0  0  0  1 -6 -3 -6 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -9 -2  1  4 -3  0  2 -1 -1 -6 -4 -6 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C -9 -3 -3 -3  6 -3 -4 -3 -3 -6 -1 -6 -2 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -9 -4  0  0 -3  4  2 -2  0 -6 -2 -6 -2 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -9  0  0  2 -4  2  4 -2  0 -6 -3 -6 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G -9 -2  0 -1 -3 -2 -2  9 -2 -6 -4 -6 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -9  0  1 -1 -3  0  0 -2  4 -6 -3 -6 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -9 -6 -6 -6 -6 -6 -6 -6 -6  9 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -4
L -9 -2 -3 -4 -1 -2 -3 -4 -3 -6  4 -6 -2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -9 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6  9 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -4
M -9 -2 -2 -2 -2 -2 -2 -3 -2 -6  2 -6  6  0 -2 -2 -2 -2 -2 -2 -2 -2 -2 -4
F -9 -3 -3 -3 -2 -3 -3 -3 -1 -6  0 -6 -2  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -9 -2 -2 -1 -3 -1 -1 -2 -2 -6 -3 -6 -2 -4  6 -1 -1 -4 -3 -2 -2 -1 -2 -4
S -9 -1  1  0 -1  0  0  0 -1 -6 -2 -6 -2 -2 -1  6  1 -3 -2 -2  0  0  0 -4
T -9 -1  0 -1 -1 -1 -1 -2 -2 -6 -1 -6 -2 -2 -1  1  9 -2 -2  0 -1 -1  0 -4
W -9 -3 -4 -4 -2 -2 -3 -2 -2 -6 -2 -6 -2  1 -4 -3 -2  4  2 -3 -4 -3 -2 -4
Y -9 -2 -2 -3 -2 -1 -2 -3  2 -6 -1 -6 -2  3 -3 -2 -2  2  9 -1 -3 -2 -1 -4
V -9 -3 -3 -3 -1 -2 -2 -3 -3 -6  1 -6 -2 -1 -2 -2  0 -3 -1  6 -3 -2 -1 -4
B -9 -1  3  4 -3  0  1 -1  0 -6 -4 -6 -2 -3 -2  0 -1 -4 -3 -3  9  1 -1 -4
Z -9  0  0  1 -3  3  4 -2  0 -6 -3 -6 -2 -3 -1  0 -1 -3 -2 -2  1  9 -1 -4
X -9 -1 -1 -1 -2 -1 -1 -1 -1 -6 -1 -6 -2 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

圖 3.3 系統計分表

```
>ARGTRGLRVFRCEQERPHQHRYDIXK
ARGRGRVNRCNQNRPNQNRYDIK
>ARGTRGLRVFRCEQERPHQHRYD
ARGRGRVNRCNQNRPNQNRYD
>ARGTRGLRWIFKQHQHRMYD
ARGGRNINKQNQNRMYD
>ARGTRGLDIXK
ARGRGQRRIKRDIK
>ARGTRGLD
ARGRGQRCGGRD
>ARGTGRLRVFRSWRYD
ARGGRRRQCRVNRSNRYD
>ARGTGRLRVFRSWRMYD
```

圖 3.4 樣板資料庫

Yieh-Ran Huang and Jan-Ming Ho, “Distributed Call Admission Control for a Heterogeneous PCS Network”, to appear IEEE Trans. On Computers, vol. 51, no. 11, Nov. 2002.

	Q		A			Q		A	R	G										G
R					D			R	V	D	N	R	S	D	N	R	M	D	Y	D

圖 3.5 書目資料基因轉換