第四章 實驗結果

4.1 測試資料產生

本實驗所需的資料可分為兩類，一類是著作表列，另一類是樣板資料庫。著作表列的來源一共有兩種，第一種是不包含中文姓氏的著作表列，這是隨機從 D-Lib 的著作中選取，擷取論文後面的參考文獻來當作我們的測試資料。第二種是有包含中文姓氏的著作表列，這是來自中央研院資訊科學所網頁。網頁上一共有一千三百多筆資料。這些資料的組成是由資訊科學所內，所有研究員、副研究員、以及助理研究員，在成員網頁中所記載的個人著作表列。蒐集這些資料後，在沒有經過修飾下存入資料庫中供作實驗。

我們以包含中文姓氏的著作表列來建立本系統的樣板資料庫，根據第三章所提及的方法來建立資料庫。首先將上述一千多筆書目資料逐筆建立它們的樣板序列，但是並非每一筆書目資料的樣板序列都納入樣板資料庫中。一開始樣板資料庫是空的，所以第一筆書目資料理所當然進入樣板資料庫中。接下來的每一筆資料則是從該筆資料之前所建立的樣板資料庫，找出最相近的樣板序列來解析該筆書目資料。如果解析結果正確就繼續往下一筆作下去，否則就將該筆書目資料的樣板序列也加入樣板資料庫中。接著往下繼續做下一筆，整個過程以人工一筆一筆審核，一筆一筆建立。經過這樣的程序，整個樣板資料庫增加為二百二十多筆資料。

以這兩百多筆不同格式的資料作為樣板資料庫，剩下的一千一百多筆資料為著作表列資料。經過實驗發現其準確率可以達到 85.4%，準確率的算法是如下:

$$\text{Precision} = \frac{\text{Number of tokens which are parsed correctly}}{\text{Number of tokens in the ciation}},$$

為了追求更高的準確率，以上述的方法再一次增加樣板資料庫數量。在第二回合建立樣板資料庫的過程結束後，又得到了三十五筆樣板序列資料。整個樣板資料庫增為二百六十筆，著作表列資料於是剩下一千多筆書目資料。以這樣的結果再作一次實驗得到準確率提升到91.72%。這二百六十筆資料就成為本系統的預設樣板資料庫。

4.2 實驗設計

很明顯樣板資料庫決定了本系統的準確度。為了估計樣板資料庫對於本系統精準度的影響，我們使用 cross-validation 來驗證。首先將資料分成 k 組樣本，C_1，…，C_k，每一組資料的筆數都相同。接著建立資料組

$D_i = D - C_i$，where $D = \bigcup C_i$

然後以 D_i 為樣本，將 D_i 中每一筆資料都轉入樣板資料庫。並且使用資料組 C_i 作為測試資料。計算以 D_i 為樣板資料庫來解析 C_i 資料組所得到的準確率 $f_{D_i}()$。當所有 $f_{D_i}()$ $(1 \le i \le k)$都經過計算後，將其結果取平均得到 k cross-validation 試驗的準確率。試驗中使用含中文姓氏的測試資料(附錄 A)以及不含中文姓氏的資料(附錄 B)各一百筆資料作 cross-validation。表 4.1 紀錄當 k 值分別為 10、5、2 時，測試含中文姓氏著作表列，所得到各組資料的準確率。圖 4.1、圖 4.2、及圖 4.3 則是這一百筆資料在 k 值分別為 10、5、2 時的準確率分佈圖。X 軸為書目資料編號，代表第 n 筆資料，n 從 1 到 100。Y 軸代表第 n 筆資料的準確率。圖 4.4、圖 4.5、圖 4.6 則是 OpCit 所得到的準確率分佈圖。

計分表對於系統準確率也是有所影響。實驗中一共測試了三種計分表。第一種是在 BLAST 程式中所預設的 BLOSUM62，如圖 4.7。第二種是齊頭式計分法，也就是不分欄位，對應到的就給正分，沒有對應到的就給負分，圖 4.8。第三種是本系統目前所使用的計分表，圖 3.3。這是經由經驗判斷後所調整出來，只是相對上比較好，還不是最好的計分表。以這三種計分表實驗所得到的準確率見表 4.3。

既然樣板資料庫完整的程度對於系統的準確率扮演著關鍵的角色，因此也設計了一個實驗來觀察樣板資料庫完整度對於準確率的影響。並且觀察本系統以及 OpCit 系統，在含有中文姓氏的著作表列與不含中文姓氏的著作表列上的表現。所得到結果以圖 4.9 表示。

為了測試姓氏資料庫對系統準確度的影響，我們分別以包含中文姓氏的著作表列，和不包含中文姓氏的著作表列作測試。測試不包含中文姓氏的著作表列時，本系統的中文姓氏資料庫完全無法發生作用，形同不使用姓氏資料庫。同時在這個實驗中也測試了 OpCit 在這兩種測試資料的準確度。實驗中本系統與 OpCit 系統所使用的樣板資料庫又可以分成兩類，一種是系統預設的樣板資料庫，另一種則是以人工的方式賦予每一筆資料正確的樣板。結果以直方圖表現於圖 4.10。

以 dynamic programming 求最佳解，通常可能的結果會有好幾個。BLAST 的結果輸出不只給我們最佳解，還輸出更多可能的配對。BLAST 的輸出先後是依照可能性排列。為了證明 BLAST 的輸出順序誤差不會有太大，在這裡設計了一個實驗。以含中文性的著作表列及不含中文姓氏的著作表列為測試資料，先求得以 BLAST 第 i 順位輸出的 template 作解析所得到的準確率 P_i，再求出 Top_i 的準確率。

$$Top_i = \max(P_1, P_2, ..., P_i),$$

實驗中求出了 Top_1、Top_2、Top_3、Top_{10} 的結果，將結果畫為直方圖，如圖 4.13 及圖 4.14 所示。

4.3 結果

比較表 4.1 與表 4.3，可看出樣板資料庫與計分表對於系統準確度都具有影響力。但是從影響的程度來看，完整的樣板資料庫與不完整的 template 資料庫對於系統準確率有 20%之多，而好的計分表與壞的計分表對系統準確率的影響卻只有 1%左右。相較起來計分表對於系統準確度影響程度比樣板資料庫小很多。樣板資料庫的完備程度是影響本系統準確度的最主要關鍵，只要樣板資料庫愈完備，整個系統的準確度就愈能提升。從上述一千書目資料隨機選取一百筆作測試，在理想的情況下，也就是樣板資料庫夠完備，本系統的解析結果可得到高達 91.2%的準確率，其準確率散佈情況如圖 4.9 所示。以同樣的資料在 OpCit 所建立的 ParaCite[19]實驗性質網站上接受測試，所得到的準確率只有 64.4%，準確率分佈情況為圖 4.10。從圖中可看出 OpCit 對於這一類資料完全無法百分之百正確解析。最主要的原因是因為含中文姓氏的著作表列，在作者的部分時常會被 OpCit 誤認為著作名稱。雖然 OpCit 建立的樣板數目比本系統多，但是該系統卻無法正確分辨作者欄位所在位置，而嚴重影響到整個系統下面欄位的判斷結果，造成低準確率。不過最主要的原因還是 OpCit 在找尋樣板時，經常無法找到最適當的樣板，甚至找不到樣板，而造成無法解析的狀況。而在我們的系統中 BLAST 卻始終可以幫我們找到樣板序列，這是使用 BLAST 得到好處之一。

圖 4.11 左邊數來第一組直方圖，代表在 OpCit 的系統中我們給予 OpCit 正確的樣板資料庫，測試含有中文姓氏的資料，以及不含中文

姓氏的資料所得到的準確率。第二組是使用 OpCit 預設的四百筆樣板資料庫。第三組是本系統使用正確樣板資料庫所得到的準確率。第四組則是本系統使用預設的兩百六十筆樣板資料庫所得到的準確率。從圖可看第三、第四組直方圖中，左邊直方圖比右邊直方圖高。表示本系統對於含中文姓氏的資料解析能力較佳，相反的 OpCit 系統則是對於不含中文姓氏的資料解析能力佳。從整體表現來看本系統平均高度比 OpCit 系統高，因此準確率上較佔優勢。

圖 4.12，橫軸代表樣板料庫的完整度，縱軸代表系統在該完整度下使用不同系統測試不同資料所得到的準確率。可發現本系統在含有中文姓氏的著作表列上比較佔有絕對優勢，而 OpCit 在不含中文姓氏的著作表列上則和本系統的準確率不相上下，不過當樣板資料庫完整度較低時，本系統還是佔有優勢，因為 BLAST 的容錯能力相當高，在困難的環境下還是可以找到相似的樣板序列，而 OpCit 因為則在樣板資料庫完整度較低時會無法找到相似的樣板造成很大的損失。

在理想情況下(圖 4.9)系統還是沒有正確的解析每一筆資料。這不禁令人懷疑是不是 BLAST 的結果出了問題。圖 4.13、4.14 橫軸表示使用不同完整度樣板資料庫，縱軸表示其準確率。可以從圖中了解，造成錯誤的原因並不是由 BLAST 引起，因為 Top_1、Top_2 的誤差不大於 2%，而 Top_2 和 Top_3 之間的差別更小於 1%，至於 Top_3 和 Top_{10} 之間的差距最多只有 0.04%。最重要的是，沒有在 Top_1 以後新增完全正確解析的結果。也就是如有最佳結果一定會在一開始就由 BLAST 找出，所以 BLAST 在這裡成功地扮演它的角色。

由於著作表列解析出來的後設資料，通常會拿來作其它用途的輸入。因此解析品質相對上就比較重要，必須要達到一定的品質才有再利用的價值。從表 4.4、表 4.5 可看出，如果以準確率 70%為分界點，

準確率 70%以上稱之好，準確率 70%以下稱之不好。OpCit 在解析品質的表現上就不如本系統。在 10 cross-validation 本系統有 85%的解析達到好的程度，而 OpCit 只有 26%達大好的程度。如果大部分的資料都沒有辦法達到好的品質，想要拿來作為其他用途的輸入實在是不大可能。

分析本系統錯誤的原因，有百分之一是因為 BLAST 找不到樣板序列。BLAST 會找不到樣板序列的主因是因為 brief form 的長度太短，而導致比對時分數太低，BLAST 為了品質的考量而捨棄所有的序列不作輸出。另外有百分之十五的錯誤，雖然我們得到了正確的樣板序列，但是得到的樣板序列所能提供的資訊卻是模稜兩可的，往往無法找到欄位正確的位置，所以還是造成錯誤。這是本系統最大的致命傷。最後剩下的百分之八十四的錯誤，則是因為資料庫中沒有正確的樣板序列，所以沒有辦法提供正確的樣板序列供作解析。

表 4.1: 本系統 k cross-validation 各組資料結果

Value of k	data set	Percent correct	Percent unparsed
10			
	E_1	93.93	0
	E_2	96.66	0
	E_3	85.32	0
	E_4	97.87	0
	E_5	100.0	0
	E_6	89.51	0
	E_7	77.50	20
	E_8	84.46	0
	E_9	91.86	0
	E_{10}	80.95	0
	Average	89.8	
5			
	F_1	89.63	0
	F_2	94.81	0
	F_3	94.75	0
	F_4	75.95	10
	F_5	74.11	0
	Average	85.85	
2			
	G_1	84.75	0
	G_2	55.64	4
	Average	70.19	

表 4.2: OpCit k cross-validation 各組資料結果

Value of k	data set	Percent correct	Percent unparsed
10			
	E_1	19.39	80
	E_2	95.64	0
	E_3	45.15	40
	E_4	50.00	50
	E_5	30.00	70
	E_6	25.00	70
	E_7	33.21	30
	E_8	7.98	80
	E_9	0	100
	E_{10}	0	100
	Average	30.63	
5			
	F_1	57.62	40
	F_2	42.88	50
	F_3	27.50	70
	F_4	12.45	75
	F_5	0	100
	Average	28.13	
2			
	G_1	10	90
	G_2	2	98
	Average	6	

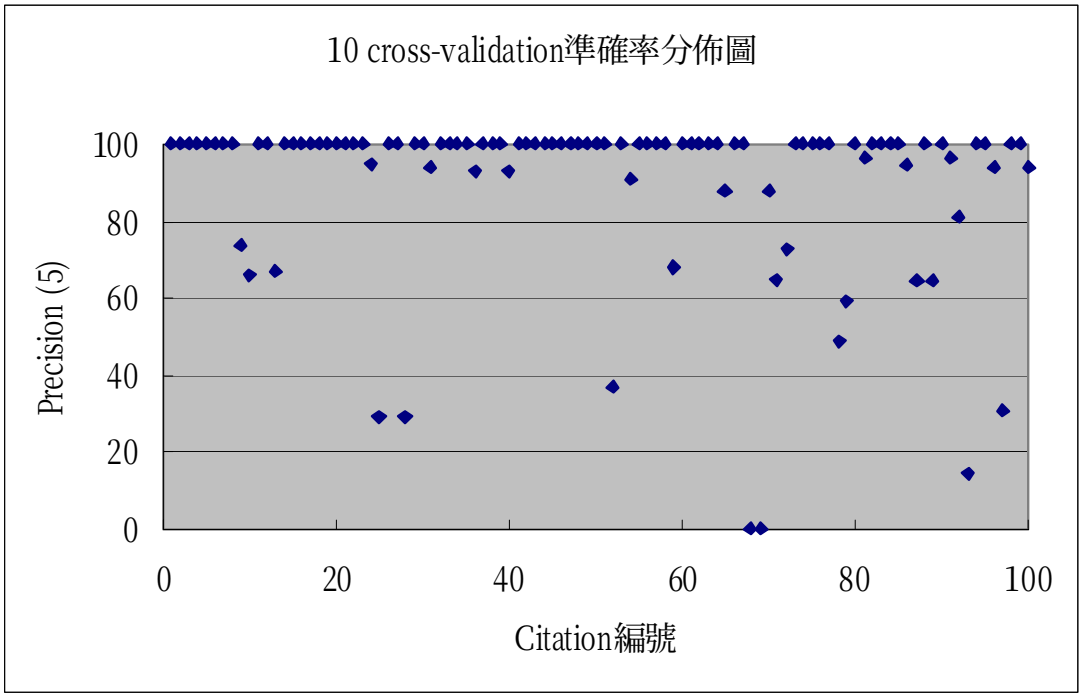

圖 4.1： 本系統以 10 cross-validation 測試含中文姓氏著作表列準確率分佈圖

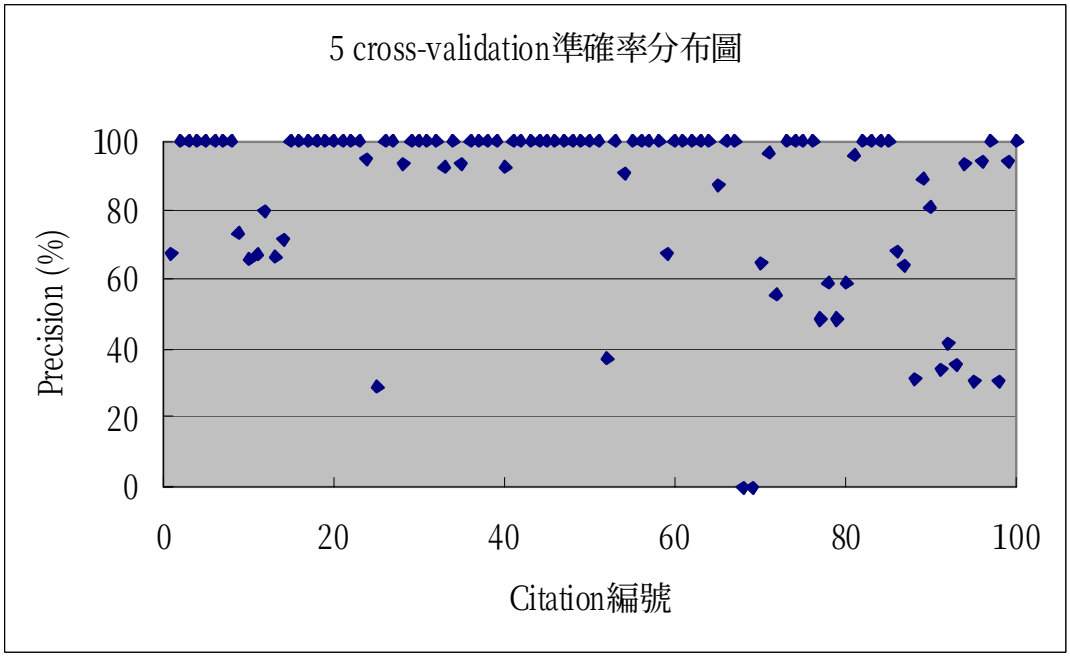

圖 4.2：本系統以 5 cross-validation 測試含中文姓氏著作表列準確率分佈圖

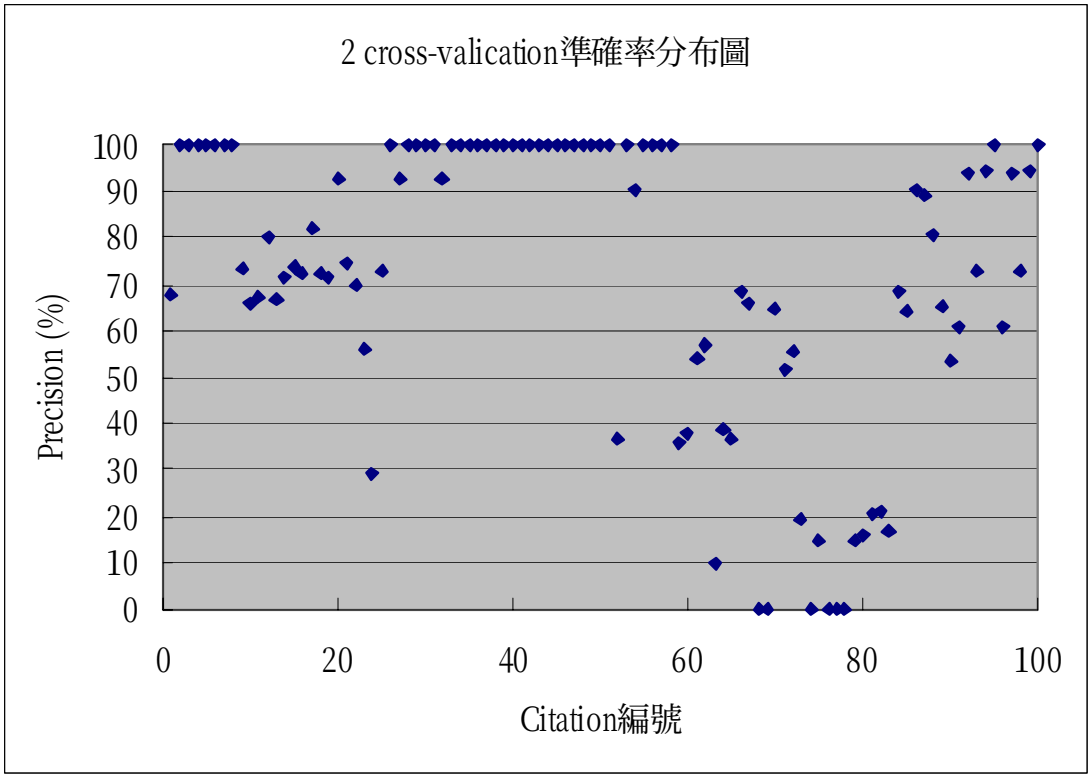

圖 4.3： 本系統以 2 cross-validation 測試含中文著作表列準確率分佈圖

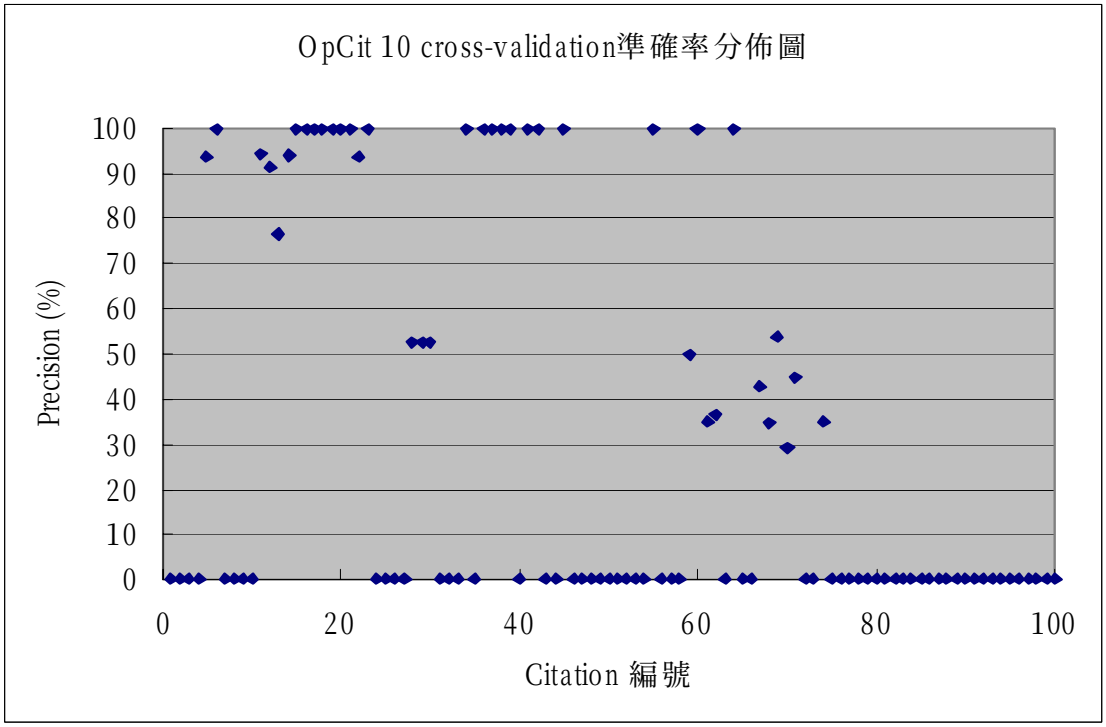

圖 4.4： OpCit 以 10 cross-validation 測試含中文姓氏著作表列準確率分佈圖

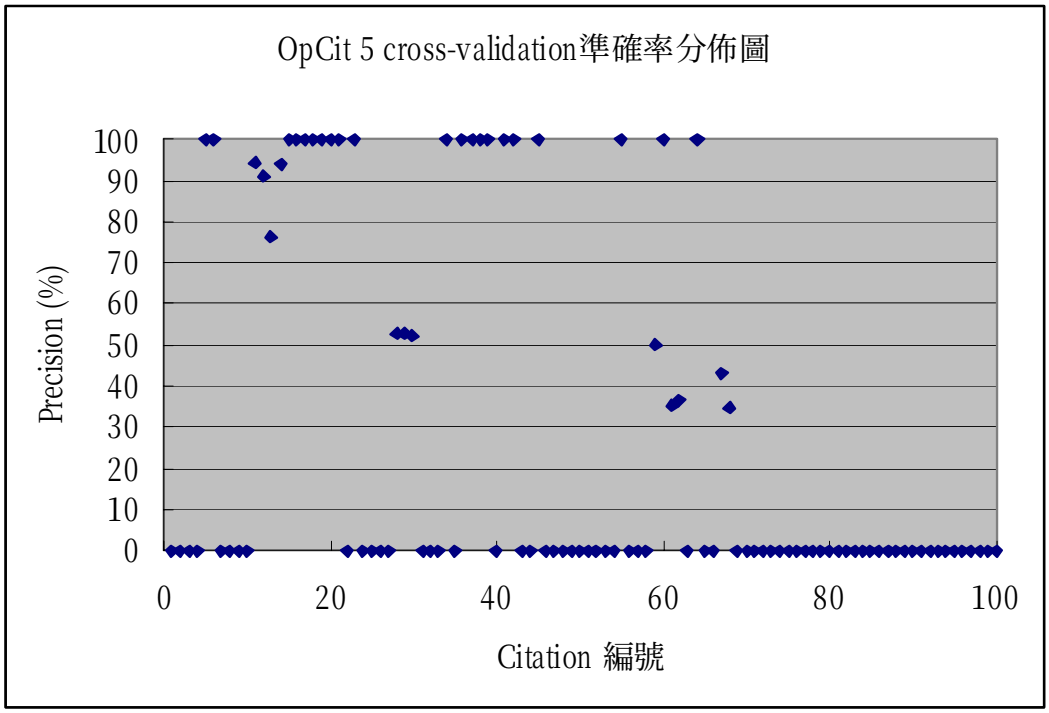

圖 4.5： OpCit 以 5 cross-validation 測試含中文姓氏著作表列準確率分佈圖

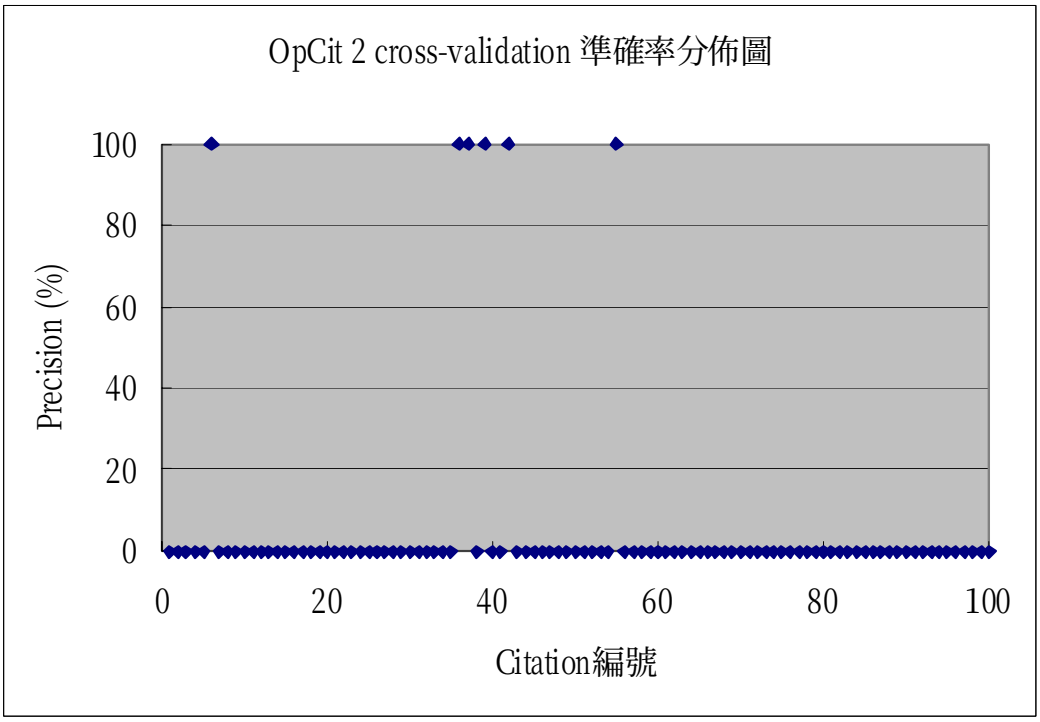

圖 4.6： OpCit 以 2 cross-validation 測試含中文姓氏著作表列準確率分佈圖

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-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 4.7： BLOSUM62

	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	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
R	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
N	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
D	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
C	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
Q	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
E	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
G	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
H	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
I	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
L	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
K	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
M	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
F	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
P	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-1	-4
S	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-1	-4
T	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-1	-4
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-1	-4
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-1	-4
V	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-1	-4
B	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-1	-4
Z	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-1	-4
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	-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 4.8： 二分法計分表

表 4.3： 使用不同計分表所得到的準確率

計分表	Precision (%)
系統計分表	91.72
2 分法計分表	91.48
BLOSUM62	90.67

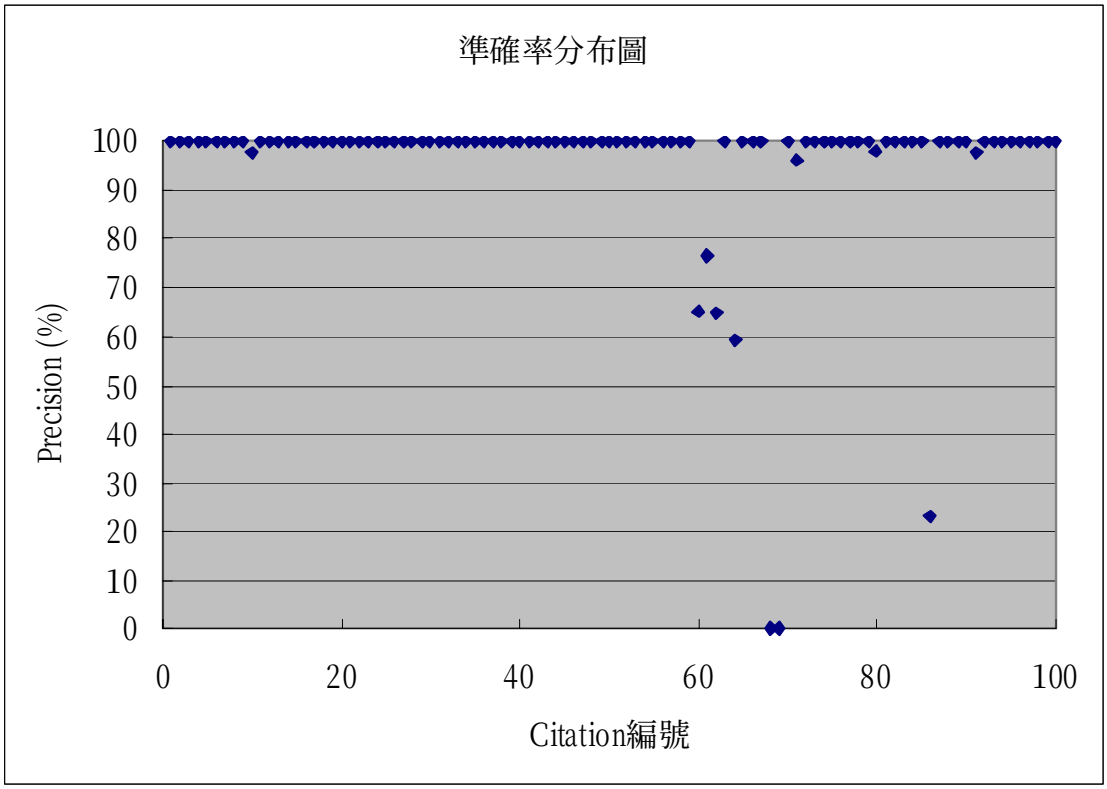

圖 4.9： 本系統理想狀況下解析中文著作表列的準確率分佈圖

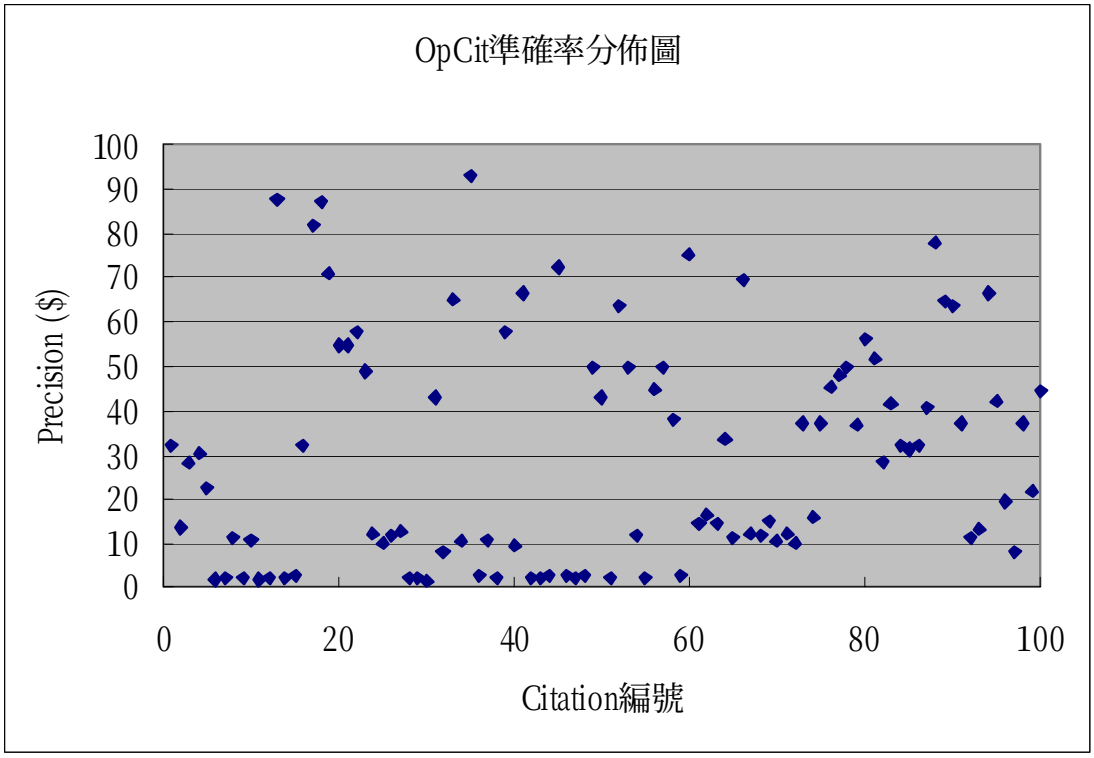

圖 4.10： 用含中文姓氏著作表列測試 OpCit 系統所得到的準確率分佈圖

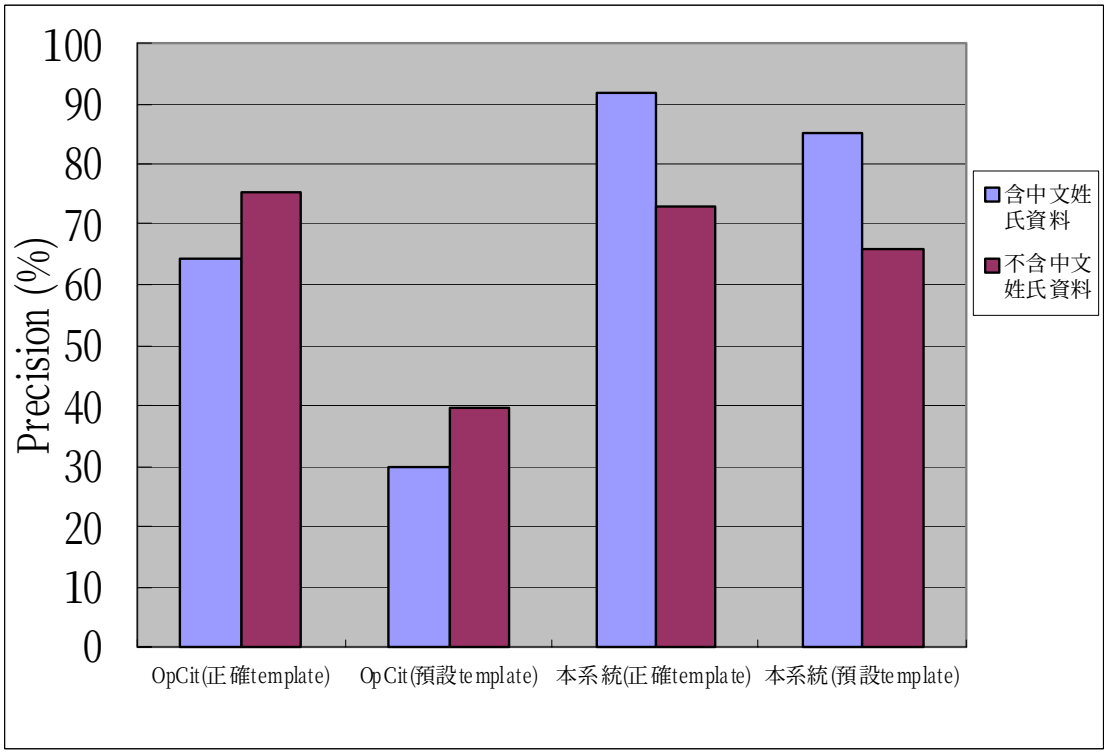

圖 4.11： 姓氏資料庫對系統資料庫的影響

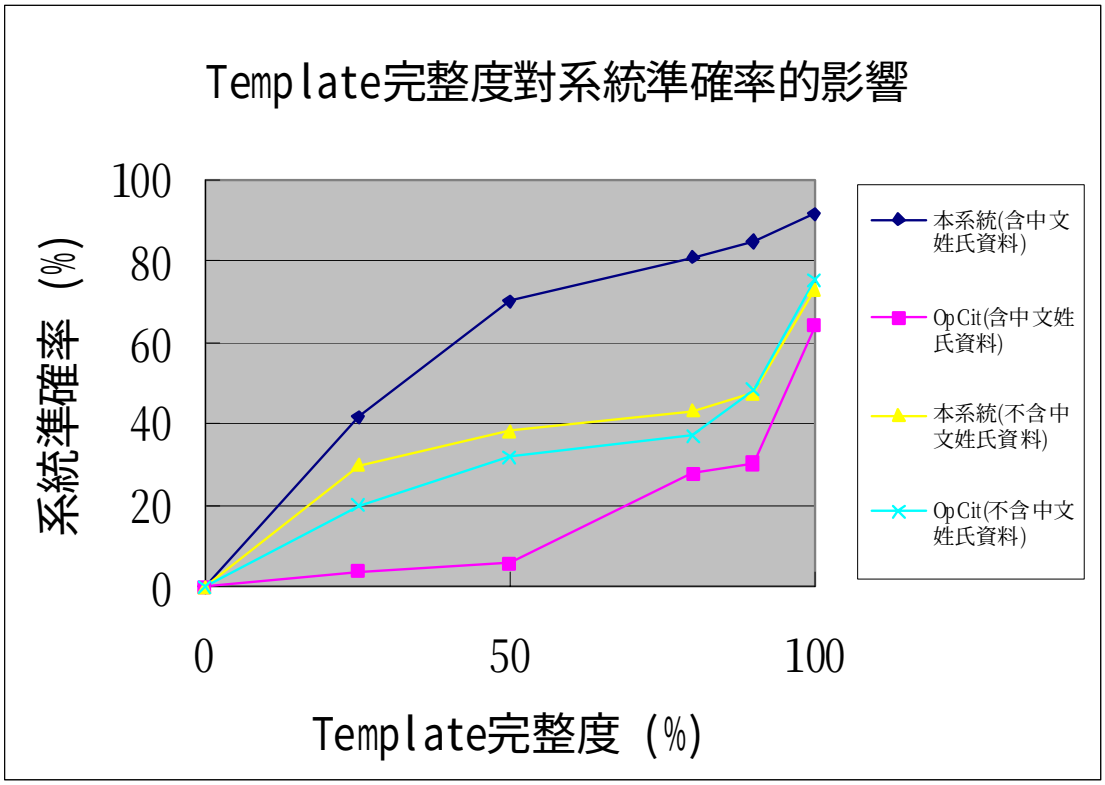

圖 4.12：以不同完整度的樣板資料庫測試含中文姓氏著作表列以及不含中文姓氏著作表列

表 4.4： 本系統解析結果品質

k 值	描述	百分比%
10		
	完全正確	71
	好(大於 70%)	85
	不好(小於 70%)	15
5		
	完全正確	59
	好(大於 70%)	76
	不好(小於 70%)	24
2		
	完全正確	37
	好(大於 70%)	52
	不好(小於 70%)	48

表 4.5： OpCit 解析結果品質

k 值	描述	百分比%
10		
	完全正確	21
	好(大於 70%)	26
	不好(小於 70%)	74
5		
	完全正確	21
	好(大於 70%)	26
	不好(小於 70%)	74
2		
	完全正確	6
	好(大於 70%)	6
	不好(小於 70%)	94

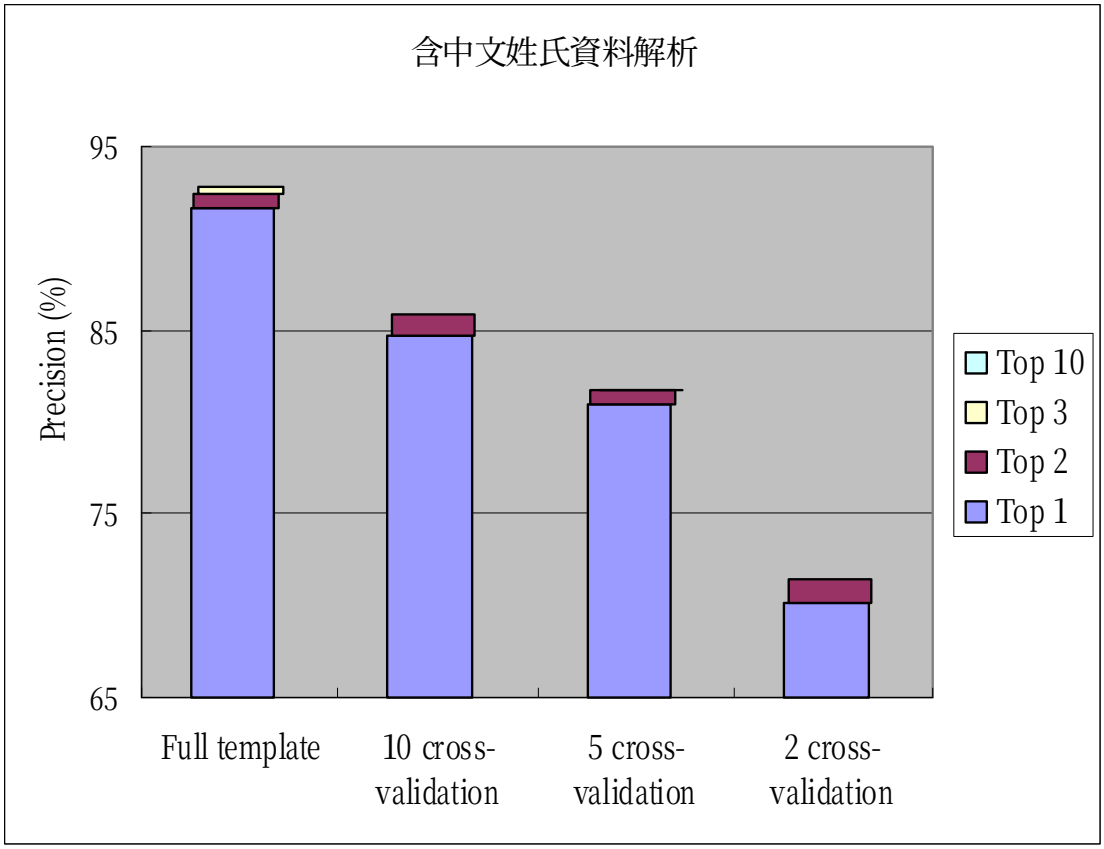

圖 4.13：在含中文姓氏的著作表列中求 Top_i

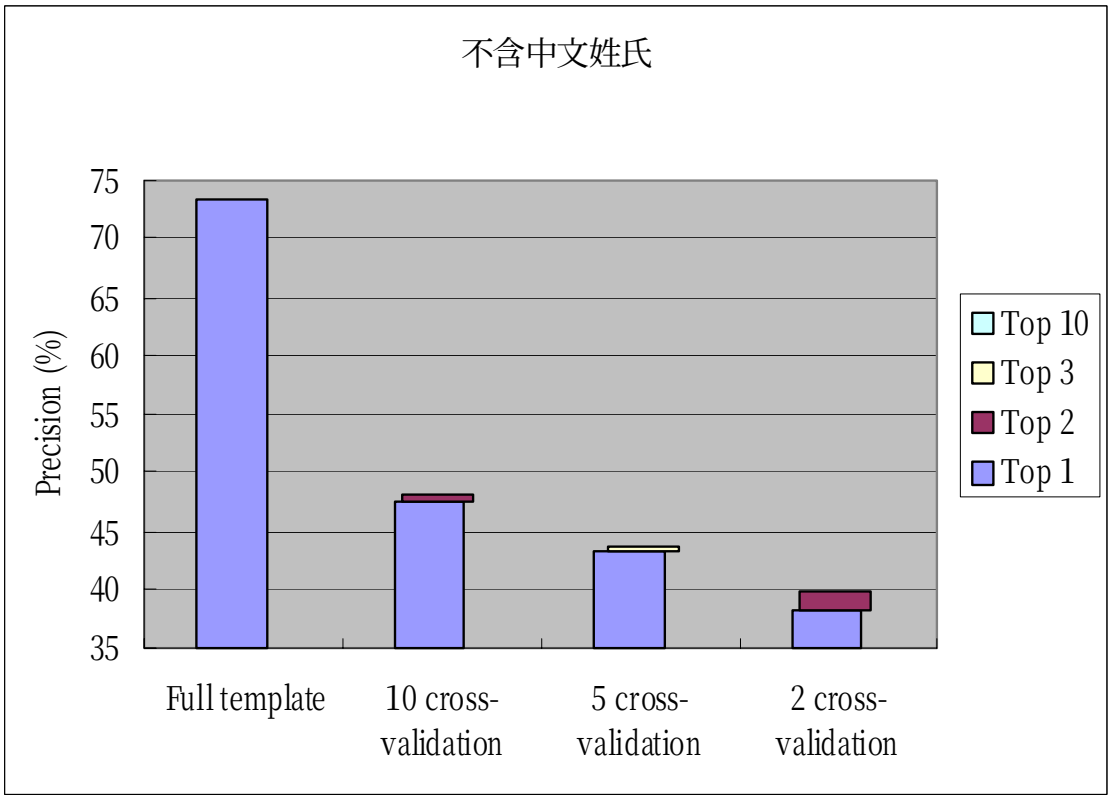

圖 4.14：在不含中文姓氏的著作表列中求 Top_i