

專題報告

中文版reCAPTCHA驗證碼在典籍數位化之應用

研究起源

- 中國歷史悠久，已出版的中文書數量已難以估計，而大量的中文典籍需要數位化並進行查詢、保存需要耗費很高的成本。
- 雖然可以自動掃描中文書籍以進行數位化，然而現今光學辨識技術的正確率並不高(例如:因、困)，需要再用大量人工去校對。

千里之行
始於足下

龍

長安張珂

發想

- 為了節省時間及避免需要大量專業人員來校對，透過網路上的瀏覽者在使用中文recaptcha驗證器時識別文字的過程來校對文字並幫助中文書籍的數位化。



+



起點



圖片切割

圖片發送

使用者辨認

統計程序

達80%標準?

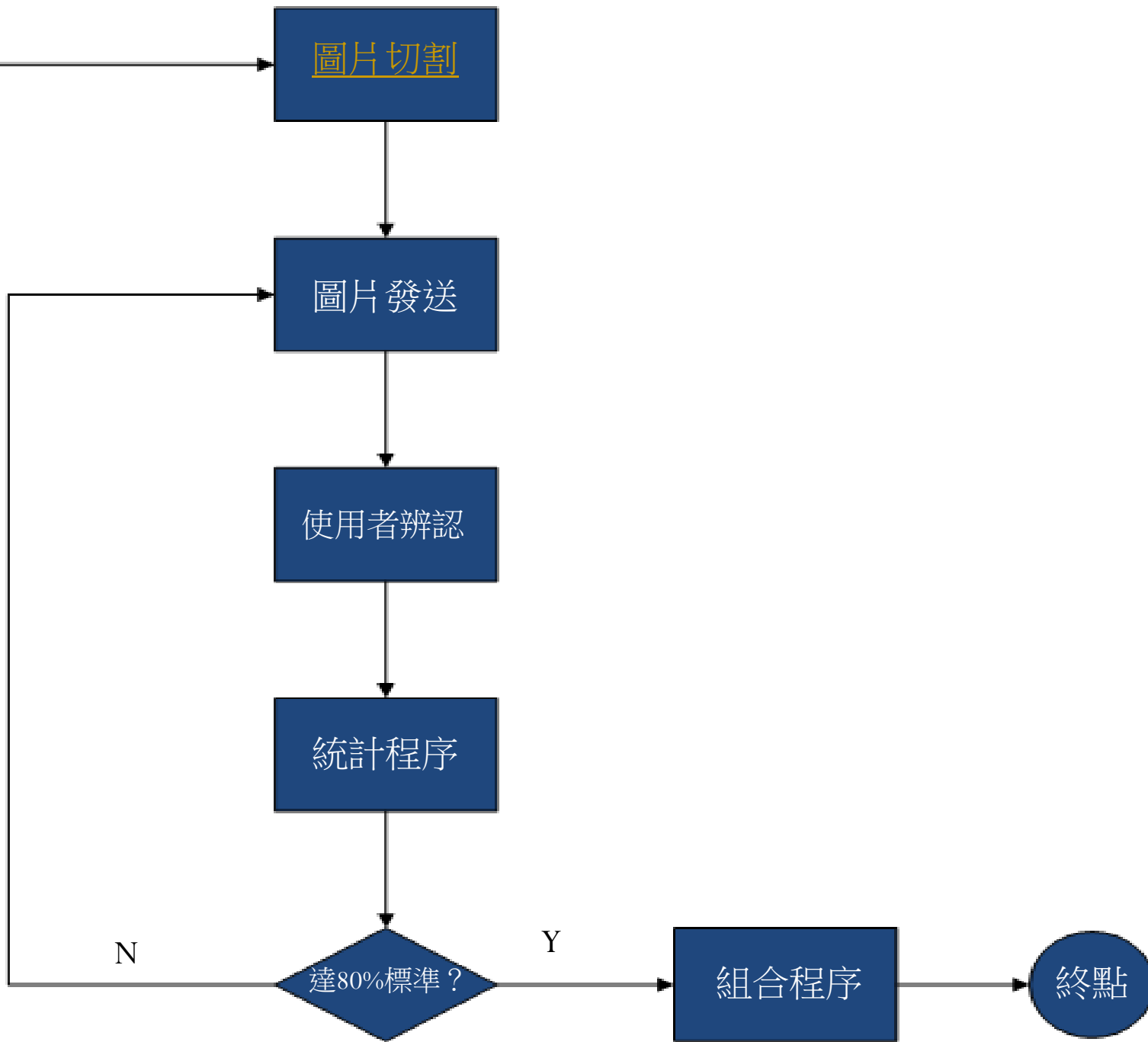
Y

N

組合程序

終點

流程圖



圖文切割

The image shows a software window titled "切割圖片" (Image Cutting) with a text document on the left and a grid of character cutouts on the right. A large blue arrow points from the text document to the grid.

Text Document Content:

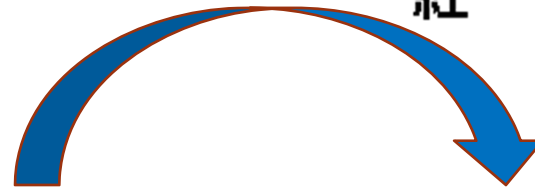
大學轉學說明會
時間：九月十日今天中午 12 :
地點：5 6 1 6 教室
對象：大一新生
(現場贈送各校轉學考資料手冊

Character Grid Content:

大 1_1.bmp	學 1_2.bmp	轉 1_3.bmp	學 1_4.bmp	說 1_5.bmp	明 1_6.bmp	會 1_7.bmp	時 2_1.bmp	間 2_2.bmp	:	2_3.bmp
九 2_4.bmp	月 2_5.bmp	十 2_6.bmp	日 2_7.bmp	今 2_8.bmp	天 2_9.bmp	中 2_10.bmp	午 2_11.bmp	12 2_12.bmp	:	2_13.bmp
10 2_14.bmp	地 3_1.bmp	點 3_2.bmp	:	5 3_4.bmp	6 3_5.bmp	1 3_6.bmp	6 3_7.bmp	教 3_8.bmp	室 3_9.bmp	
對 4_1.bmp	象 4_2.bmp	:	大 4_4.bmp	一 4_5.bmp	新 4_6.bmp	生 4_7.bmp	(5_1.bmp	現 5_2.bmp	場 5_3.bmp	
贈 5_4.bmp	送 5_5.bmp	各 5_6.bmp	校 5_7.bmp	轉 5_8.bmp	學 5_9.bmp	考 5_10.bmp	資 5_11.bmp	料 5_12.bmp	手 5_13.bmp	
冊 5_14.bmp	,	每 5_16.bmp	人 5_17.bmp	限 5_18.bmp	拿 5_19.bmp	一 5_20.bmp	份 5_21.bmp) 5_22.bmp		

存進資料庫

癸明
學問
欲學
塵紅



大	學	轉	學	說	明	會	時	間	:
1_1.bmp	1_2.bmp	1_3.bmp	1_4.bmp	1_5.bmp	1_6.bmp	1_7.bmp	2_1.bmp	2_2.bmp	2_3.bmp
九	月	十	日	今	天	中	午	12	:
2_4.bmp	2_5.bmp	2_6.bmp	2_7.bmp	2_8.bmp	2_9.bmp	2_10.bmp	2_11.bmp	2_12.bmp	2_13.bmp
10	地	點	:	5	6	1	6	教	室
2_14.bmp	3_1.bmp	3_2.bmp	3_3.bmp	3_4.bmp	3_5.bmp	3_6.bmp	3_7.bmp	3_8.bmp	3_9.bmp
對	象	:	大	一	新	生	(現	場
4_1.bmp	4_2.bmp	4_3.bmp	4_4.bmp	4_5.bmp	4_6.bmp	4_7.bmp	5_1.bmp	5_2.bmp	5_3.bmp
贈	送	各	校	轉	學	考	資	料	手
5_4.bmp	5_5.bmp	5_6.bmp	5_7.bmp	5_8.bmp	5_9.bmp	5_10.bmp	5_11.bmp	5_12.bmp	5_13.bmp
冊	,	每	人	限	拿	一	份)	
5_14.bmp	5_15.bmp	5_16.bmp	5_17.bmp	5_18.bmp	5_19.bmp	5_20.bmp	5_21.bmp	5_22.bmp	

DataBase(待辨認區)

起點

圖片切割



將圖片切割放入DB的待辨認區

圖片發送

使用者辨認

統計程序

達80%標準?

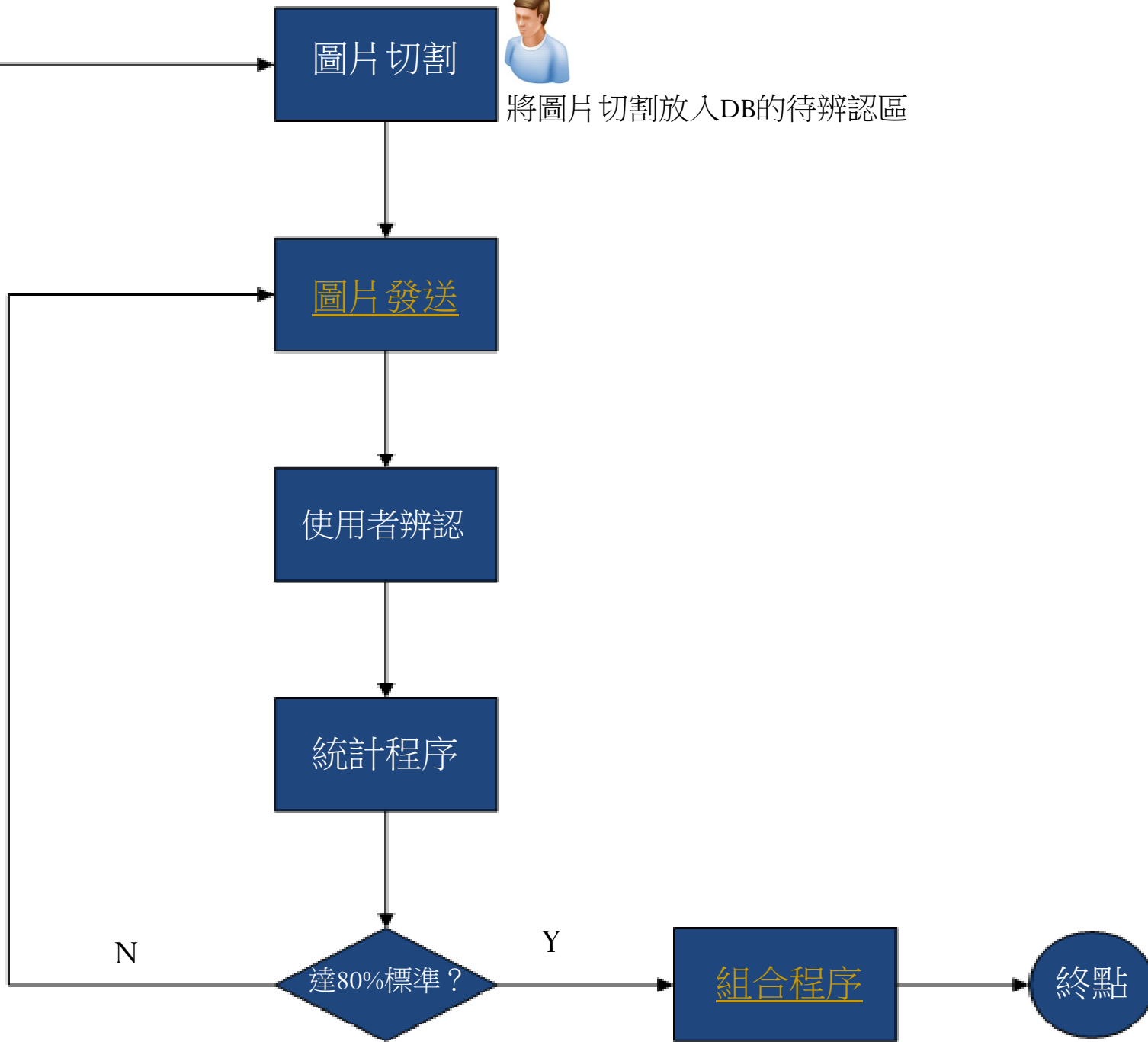
Y

組合程序

終點

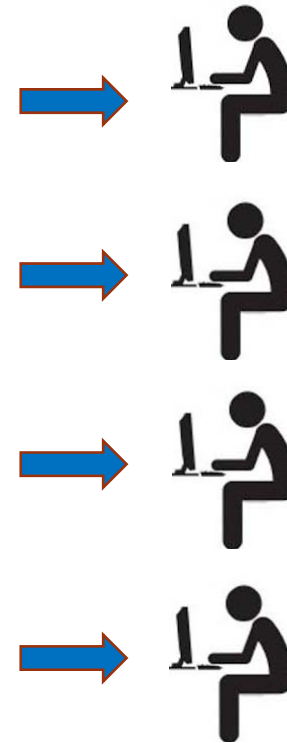
N

流程圖

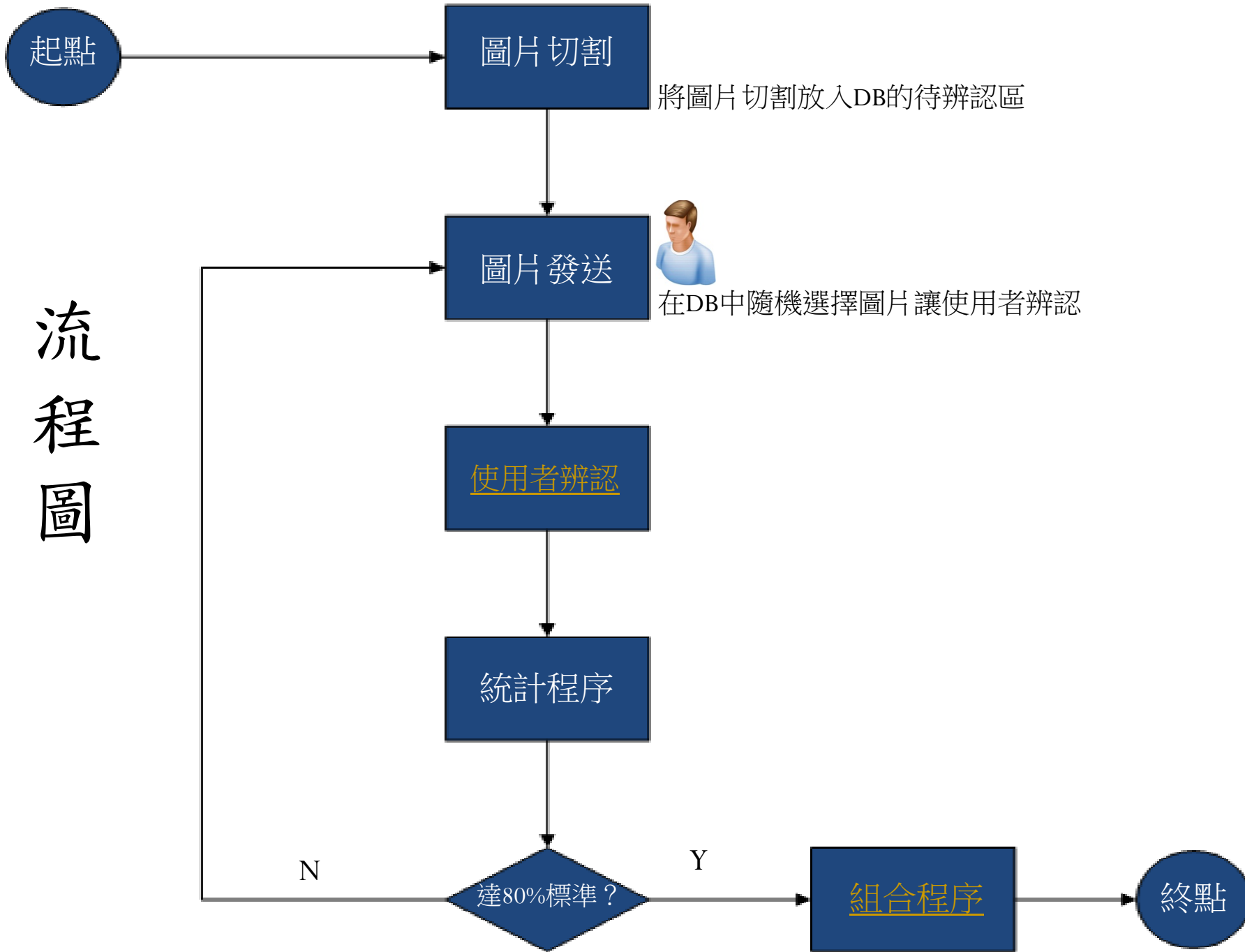


圖片發送

從待辨認區隨機抽出資料
顯示至驗證器上供使用者判斷




流程圖



使用者辨認

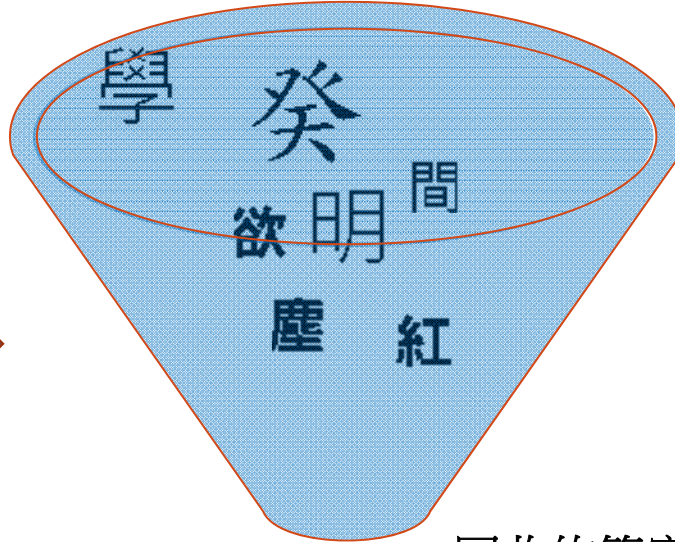
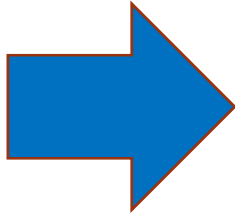
[首頁](#) | [瀏覽留言](#) | [發表留言](#)

	標題:馬肥ㄅㄨㄛˊㄅㄨㄛˊ之專題報告
	☺ YAYAYAYAYAYAYAYAYAYAYA
	宋宇軒 2012/12/26 下午 03:42:15

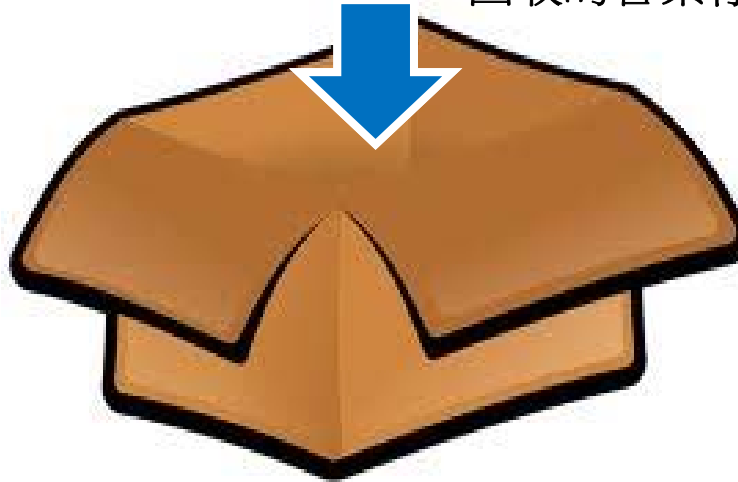
	標題:路過
	☺ 中文RECAPTCHA超讚
	馬振宇 2012/12/26 下午 01:03:53

	標題:
	2012/12/26 上午 05:20:23

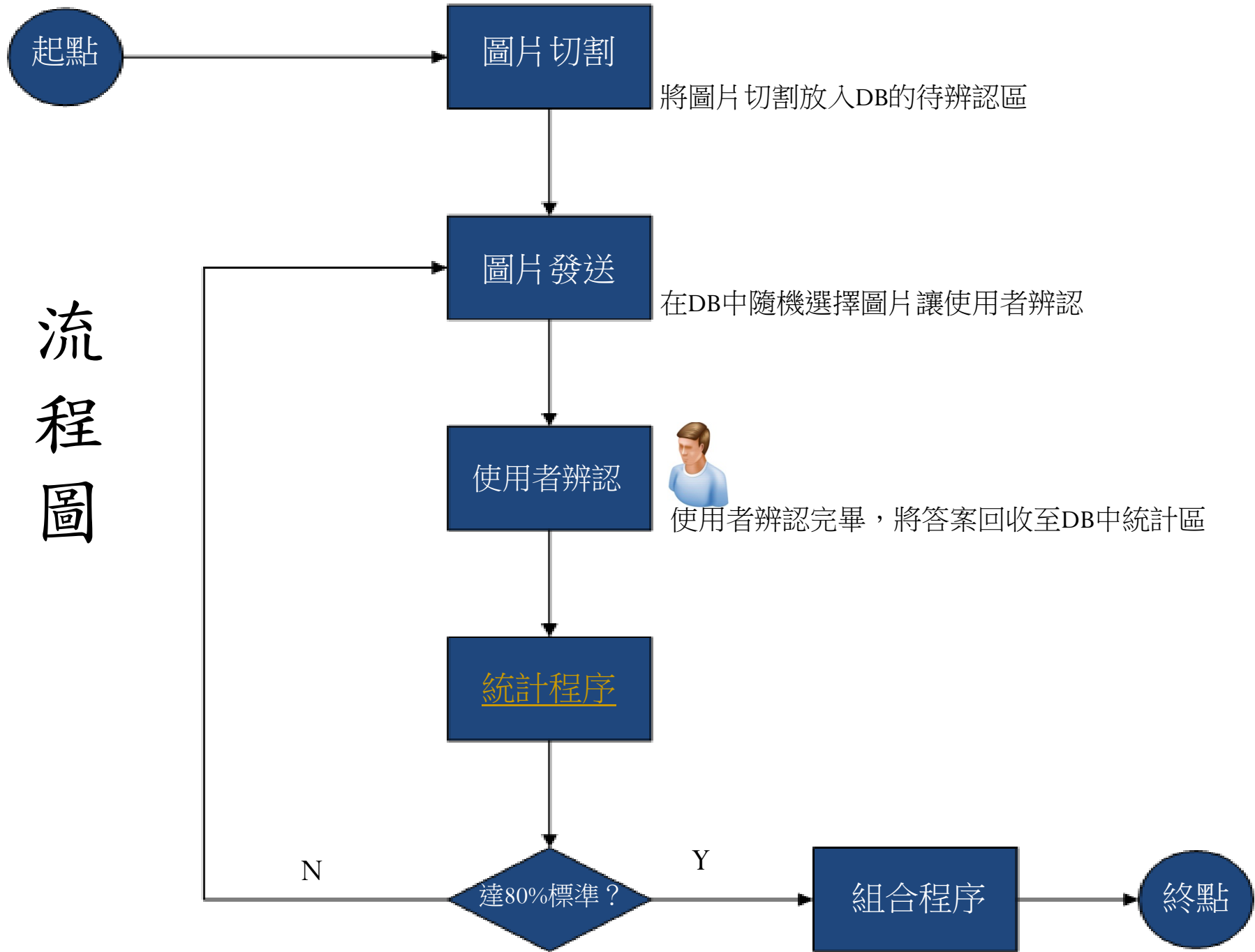
回收答案



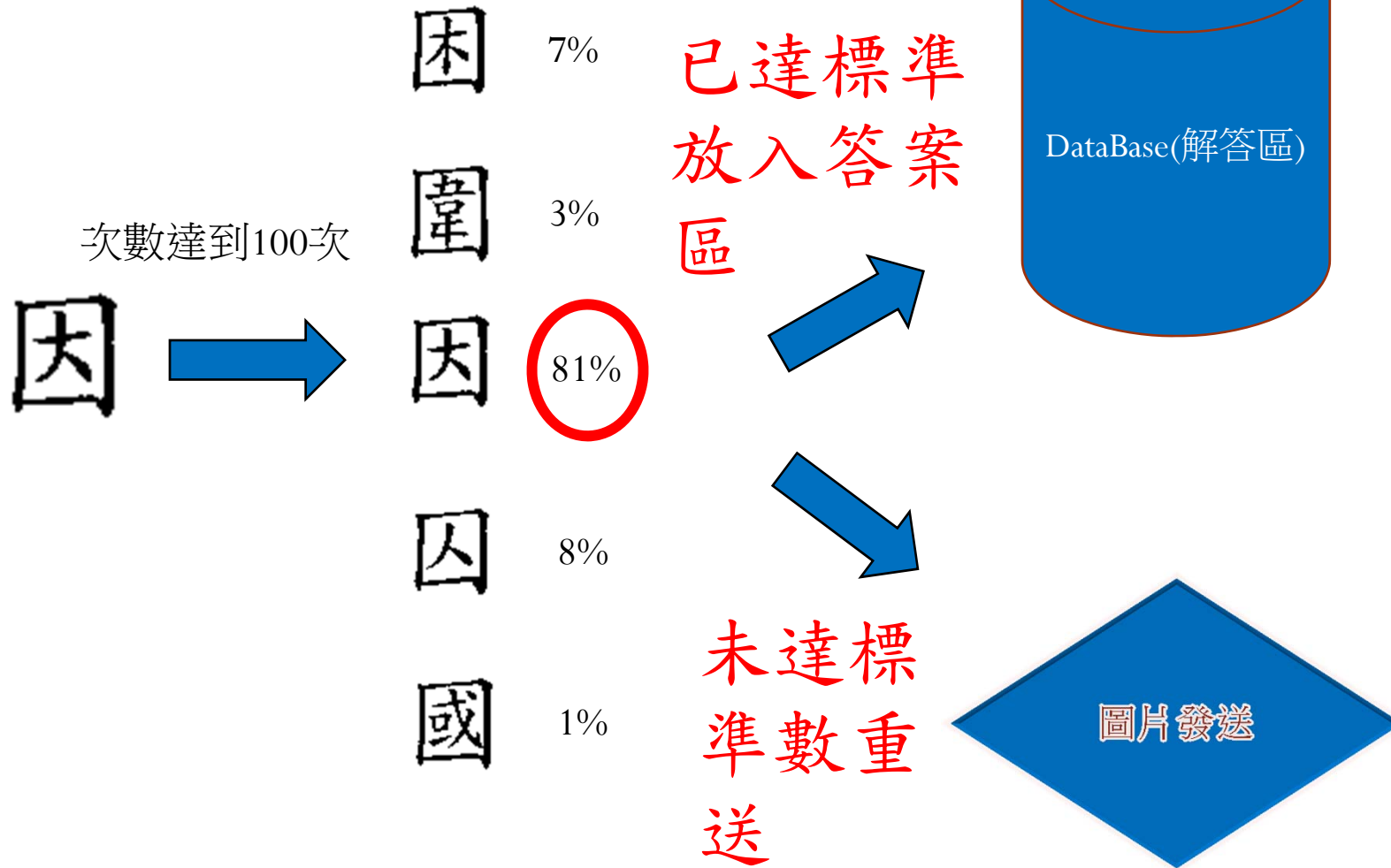
回收的答案存放至待統計區域



流程圖



統計



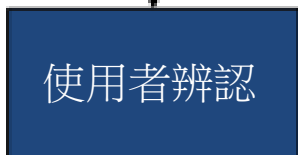
流程圖



將圖片切割放入DB的待辨認區



在DB中隨機選擇圖片讓使用者辨認



使用者辨認完畢，將答案回收至DB中統計區



回收圖片達到需求數量後進行統計答案



N

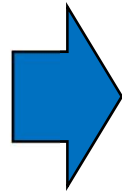
Y

將該圖片重新發送已取得答案



組合

大學轉學說明會
時間：九月十日 中午十二時
地點：5616教室
對象：大一新生
(現場贈送各校轉學考資料手冊，每人限拿一份)



The screenshot shows two windows. The top window, titled 'Form1', contains the text from the left side of the image. On the right side of this window, there is a label '第幾頁:' followed by a text input field containing '11'. Below the input field are two buttons: '組合' (Combine) and '匯出' (Export). A blue arrow points from the '匯出' button to the bottom window. The bottom window is a Notepad application titled '新文字文件.txt - 記事本'. It contains the same text as the form above.

流程圖

