應用於資訊檢索的中文 OCR 錯誤詞彙自動更正

Error Correction of Chinese OCR texts for Information Retrieval

曾元顯
輔仁大學圖書資訊學系 教授
Yuen-Hsien Tseng
Professor
Dept. of Library & Information Science, FJU
tseng@lins.fju.edu.tw
「中國圖書館學會會報」，72 期，2004 年 6 月，頁 23-31.

摘要：

本文描述一套全自動的方法來更正中文 OCR 文件的錯誤。文件的重複詞彙，不論正確或錯誤，都可被擷取出來，並歸類在一起。查詢時可將對應的錯誤詞彙提示出來，供使用者參考選用，以降低 OCR 錯字對檢索成效的影響。實驗顯示，這套全自動的方法，可以讓查全率增加 15.01%，但查準率會下降 4.13%。但是若有其他資源可用，例如與 OCR 文件主題類似的乾淨文件，則查準率的下降可以輕易改善。

Abstract:

A fully automatic error correction method is proposed for use in Chinese OCR text retrieval. Terms, correct or not, in OCR texts are extracted and then clustered to get confusion pairs. Query terms can then first match these confusion pairs to get various versions of the same terms for better searches that reduce the vocabulary mismatch problem. This article describes how this can be done for Chinese texts, where no delimiters exist between words. The proposed method made an improvement in gross micro-average recall by 15.01% and a reduction in micro-average precision by 4.13% in a Chinese OCR test collection. Compared to no error correction at all, these results are encouraging. If other resources are available, the degradation in precision can be alleviated. A direct application of this technique would be term suggestion in an interactive retrieval environment. Another application is to strengthen the recognition capability of OCR techniques.

關鍵詞：錯字更正、詞彙歸類、中文處理、混淆字表、資訊檢索

Keywords: Error correction, term clustering, Chinese processing, confusing table, information retrieval

壹、導言

電腦網路的發達，使得資訊的出版、傳播、與取得過程更加便利。雖然未來的資料以全數位形式出現是可預期的事情，然而現今紙本資料，仍然記錄了全世界非常多的資訊。要將（回溯性）紙本資料數位化，以提供網路化的資訊服務並非易事。

一套將紙本資料數位化的流程，是將紙本資料掃瞄成影像檔，再利用光學文字辨識（Optical Character Recognition, OCR）軟體辨識影像檔來取得數位文字，以便提供全文的檢索與影像的取用。如果原始文件可產生高品質的影像和 OCR 文字，採用這個途徑將會是成本效益極佳的選擇。

然而 OCR 文件常常含有辨識錯誤的詞彙，導致其提供的檢索品質可能降低。過去的研究顯示，OCR 辨識錯誤的情形，對影像品質良好的文件，並不嚴重，對檢索成效的影響也不大 [1]。然而圖書館的館藏，常常是年代較為久遠、印刷或紙質較差的紙本資料，其 OCR 的結果，常常是詞彙錯誤率較高的數位文件，對檢索結果的影響也較顯著。因此，一個值得研究的課題，是如何降低 OCR 的辨識錯誤、提升 OCR 文件的檢索成效。

OCR 系統中，有時也會有事先建好的詞庫與錯誤詞彙的更正模組，以自動校正 OCR 辨識錯誤的詞彙。然而既使是最好的 OCR 系統，辨識錯誤總是難以避免。這可能是事先建構的詞庫，無法涵蓋任一領域、任一文件所使用的詞彙所致。相對的，檢索系統中則總是會建構文件特定（collection- specific）的索引詞彙，而這些特定文件的全域性（global）索引詞彙，如果好好利用，有可能可以進一步幫助更正更多的 OCR 錯誤。

基於上述的想法，本文提出一種方法，利用檢索系統裡自動辨識的關鍵詞彙或索引詞彙，以及檢索系統製作的索引與詞頻統計資訊，進行詞彙歸類（clustering），將可能的錯誤詞彙與正確詞彙關聯起來，以便在檢索時，系統可以根據使用者提供的正確詞彙，提示出可能的錯誤詞彙，而達到錯字偵測與更正的目的。當錯誤詞彙被找出後，即可掃瞄原 OCR 文件，將錯誤詞彙都改成正確詞彙。

本文的內容安排如下：下一節將介紹以往相關的研究，簡略介紹 OCR 技術的錯字更正方法，以及資訊檢索領域相關的作法；第三節介紹本文提出的方法及其特性；第四節則以實際的 OCR 文件進行測試，以驗證其成效；最後一節則討論本方法可能的應用，以及其優缺點。

貳、相關研究

數位文件中，出現錯字的原因有很多種，包括：(一）人工輸入的錯字：如同音（辨識、辯士、便是)、同形（「籍由」應為「藉由」)、別字、疏漏（漏打一個字)、冗字（多打一個字，或該刪掉而未刪)；(二）機器辨識的錯字：影像文字辨識錯誤、語音辨識錯誤；(三）字碼轉換的錯字：繁簡體轉換錯誤、字碼間轉換錯誤（如 CCCII 碼轉成 BIG5 碼，有時會對應不到或對應錯誤)。而且不同語文的錯誤情況可能有很大差異，例如英文有拼字錯誤的問題，中文則非拼音文字，基本上沒有此問題。根據 Kukick [2] 的歸納，不同語文、不同原因以及不同應用的錯字偵測與更正，其所需要的技術有相當大的不同。因此後續的討論，僅限於中文 OCR 文字的錯誤更正，特別是在 OCR 文件的檢索系統中，進行的錯字偵測與更正。

全文影像辨識的軟體中，常有一套錯字更正的後處理系統，以更正依據文字影像辨識出來的錯誤詞彙。影像辨識核心系統，掃瞄全文影像後，將可能的文字區塊辨識出來，然後分析文字走向，釐清一行行文字，再分割出一個個字元，最後將分割出來的字元影像，做筆跡細化處理（thinning）以及連筆或斷筆的判斷，再依據字元的形狀，辨識出其數位字碼。由於這個過程是一個個字元獨立運算，同型態的字元容易造成錯誤辨識，例如「已經」的「已」辨識成「己」。後處理系統，則運用詞庫、混淆字表（confusion table，亦即型態近似容易辨識錯誤的字元對照表)、詞彙統計模型、以及辨識系統錯誤統計模型等資源與技術，來更正錯誤的詞彙 [3]。

例如，透過詞庫，我們可以知道沒有「己經」這個詞，同時透過混淆字表，我們知道「已」與「已」容易互相辨識錯誤，因此，我們可以將「己經」改成「已經」。然而，由於中文詞彙常有新詞產生，且混淆字表僅涵蓋型態較相似的字元，不夠相似、詞庫又沒有的詞彙，就很難被上述方法更正。

Taghva 等人曾提出一套更正英文 OCR 錯字的方法 [4]。首先將索引中的非停用詞依據詞庫分成「中心詞彙」（centroids）與「錯誤詞彙」（misspellings)，中心詞彙為詞庫中出現的索引詞，而錯誤詞彙則不在詞庫中出現的索引詞。接著利用 agrep 這個相似字串比對的工具軟體 [5]，將錯誤詞彙依照近似程度歸類到中心詞彙。如果一個錯誤詞彙只歸類到一個中心詞彙去，則以後碰到該錯誤詞彙就更正成該中心詞彙。如果一個錯誤詞彙被歸類到多個中心詞彙去，則利用下列兩種方法來決定其最後的中心詞彙。(一）利用局部資訊：找出錯誤詞彙出現的文件，統計該篇文件每個詞彙出現的次數，若其對應的中心詞彙出現次數為 0，則刪除該中心詞彙。若這樣做就可得到唯一的中心詞彙，則將該錯誤詞彙更正成此中心

詞彙。（二）利用混淆字表：若經局部資訊更正後，還有多個中心詞彙，則查詢混淆字表，將不可能發生的情況排除。例如，表一中的 transporation 經局部資訊更正後，還是歸類在 transpiration 與 transportation 兩個中心詞彙中，經查詢類似表二的混淆字表後，得知 t 可能會被省略而不被辨認出來（表二最後一行），但 i 不可能被辨識成 o（混淆字表中從沒出現）。因此，transporation 應該更正成 transportation。

表一：利用局部資訊與混淆字表來更正錯誤詞彙（資料來源：Taghva 等人論文）

錯誤詞彙	中心詞彙	局部資訊更正	混淆字表更正
ariation	aviation variation	variation	
downwar	downwarp downward	downward	
ountain	fountain mountain	mountain	
transporation	transpiration transportation	transpiration transportation	transportation

表二：混淆字表範例（資料來源：Taghva 等人論文）

次數	正確	錯誤
48	e	c
25	c	e
18	m	rn
16	t	r
2	t	

Taghva 等人利用一份檢索測試集來測試上述方法的效果。此檢索測試集包含 71 道查詢，204 篇文件。雖然文件篇數少，但總共有 9300 頁，而且乾淨文件（人工輸入）與 OCR 文件都各一份。這 71 道查詢，平均每道查詢有 5 個詞，以布林邏輯組合，進行布林查詢。從乾淨文件中，總共檢索出 632 篇文件，而從 OCR 文件中，總共檢索出 617 篇，檢出率為 97.6%，遺漏 15 篇。若將 OCR 文件經上述詞彙更正處理後，則總共檢索出 624 篇，檢出率為 98.7%，遺漏 8 篇。顯然其錯誤詞彙更正方法有些許效果。但其效果沒有很明顯，其原因可能是文件的長度很長，以致查詢詞在文件中重複出現，而只要有一次是正確的形式出現，即可被檢索出來。

參、本文提出的方法

要運用 Taghva 等人的方法於中文文件，必須先從 OCR 文件中取出詞彙，而且不管此詞彙是正確或錯誤，都要擷取出來，然後依據詞庫將其分成中心詞彙與錯誤詞彙。但是中文不像英文，詞彙間有空格隔開，中文斷詞本身就是一件不容易的事，更何況要斷出錯誤詞彙。

所幸，我們觀察到，即便是 OCR 辨識錯誤的詞彙，也會重複出現。光憑這一點特性，我們就可以利用筆者之前發展的關鍵詞擷取方法，擷取出最大的重複詞（maximally repeated strings）[6]。此方法僅利用主題詞彙容易重複的特性，即可擷取關鍵詞，不論其正確與否。

此外，Taghva 等人對英文詞彙的歸類方式，係依據 agrep 工具進行相似比對，將相差只有一小部份的英文詞彙歸類在一起。例如英文詞彙平均有五個字母，將相差兩個字母的英文詞彙歸類在一起。然而中文詞彙有很高的比例是由 2 個、3 個少數字元組成，其歸類方式勢必要做些許調整。因此，我們提出一套針對中文特性的改良方法，分為三個主要步驟，顯示在圖一當中。

步驟一：重複詞擷取

1. 以參考資料 [6] 中的演算法擷取每篇 OCR 文件中的重複詞彙，並累計每個詞彙的出現篇數。

步驟二：詞彙歸類、混淆字表產生

1. 從步驟一獲得的詞彙中挑出所有的 n 字詞，其中 n>=3。
2. 對於具有相等長度的任意兩個 n 字詞，比較其組成的字元。
3. 如果此兩詞彙只有相差一個字元，就將其歸類在一起，並且以出現篇數高者為「中心詞彙」，出現篇數低者為「錯誤詞彙」。
4. 將上述兩詞彙不同的字元取出置於混淆字表中，並累積其出現次數。

步驟三：混淆字表排序

1. 將混淆字表中的項目按照其出現次數由高到低排序。

圖一：中文 OCR 文件錯誤詞彙之偵測與更正之全自動步驟

為了說明步驟一中的作法，圖二顯示一實施範例。圖二中每個英文大寫字母代表一個中文字，並且冒號後面的數字，代表該字串出現的次數。此方法僅憑輸入字串：「BACDBCDABACD」，即可全自動獲得 CD 出現 3 次，BACD 出現 2 次的結果，而沒有用到其他如詞庫、字典等額外資源。

輸入文字: “ BACDBCDABACD ”, 假設 門檻值 = 1
步驟一 : 將輸入字串拆解成雙連字（bi-grams)，並累計次數，得到一序列：
L = (BA:2 AC:2 CD:3 DB:1 BC:1 CD:3 DA:1 AB:1 BA:2 AC:2 CD:3)
步驟二: 將次數都高於門檻值的相鄰雙連字合併回來，合併過者丟掉，高於門檻但無法與其左右鄰合併的則留住, 重複此步驟, 直到無法合併為止 :
第一次 :合併 L 成 L1= (BAC:2 ACD:2 BAC:2 ACD:2)
丟掉: (BA:2 AC:2 CD:3 DB:1 BC:1 DA:1 AB:1 BA:2 AC:2 CD:3)
留住 : (CD:3)

第二次 : 合併 L1 成 L2 = (BACD:2 BACD:2)
丟掉 : (BAC:2 ACD:2 BAC:2 ACD:2)
留住 : (CD:3)

第三次 : 合併 L2 成 L3 = ( )
丟掉 : ( )
留住 : (CD:3 BACD:2)

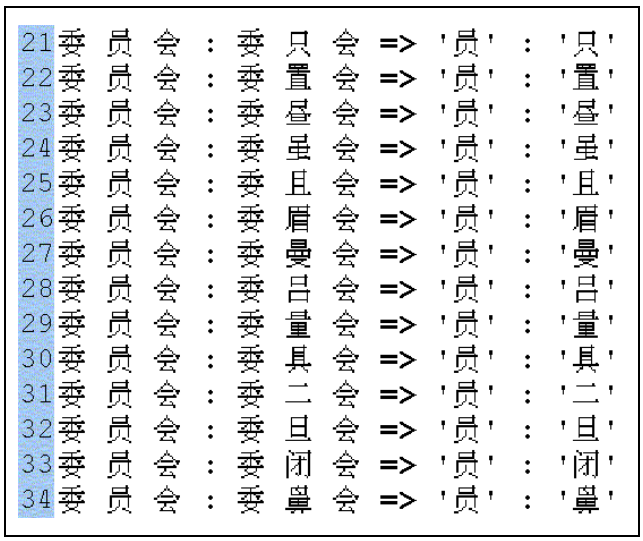
圖二：擷取重複字串的方法示意圖。

取出所有文件的重複詞後，接下來要將詞彙歸類。由於印刷的中文字長寬一致，不太可能發生如英文字中有刪除、省略的辨識情況 (如表二中最後一列的現象)，因此我們只比較長度一樣的詞彙，來找出可能的錯字。

當能夠正確斷出正確與錯誤的詞彙，且兩個詞只差一個字的時候，我們可以像 Taghva 等人那樣，運用詞庫與混淆字表來斷定那個是正確、那個是錯誤。例如，如圖三所示，「委員會」與「委具會」在第二個字不同，且其混淆字元為「員」與「具」，當詞庫告訴我們「委員會」為一個詞，而「委具會」不是，且混淆字表顯示「員」容易錯成「具」，那麼我們就可以將「委具會」改成「委員會」。

為此，我們曾向 OCR 廠商要過其混淆字表，亦即其系統認為最容易辨識錯誤的字元對（pair)，共 4720 對，如圖四左邊所示，箭頭後面的數字，代表相似程度。然而根據圖一的步驟，筆者從 8438 篇 OCR 文件中產生如圖四右邊的結果，冒號後面的數據，代表該混淆字出現的次數。圖四左右兩相比較，兩者有非常不同的混淆字統計排序。這顯示，廠商的混淆字表乃根據其系統特性或根據其自我的實驗環境產生，而非根據實際文件的 OCR 辨識狀況產生，因此跟實際的情況有很大的出入。另外，圖一的步驟產生了很多 OCR 廠商的混淆字表中沒有的情況，顯示很多辨識錯誤都超過現有 OCR 軟體能夠更正的能力。例如圖三中的「委員會」與「委二會」,「員」與「二」兩個字外型差距很大，混淆字表裡不太可能蒐

納進去，因而也就不太可能被傳統的 OCR 系統所更正。



圖三：利用圖一步驟產生的詞彙歸類結果，最左欄為中心詞彙，其右邊是錯誤詞彙，箭頭後面是其混淆字元。

圖四：混淆字表範例。左圖是 OCR 廠商給的，右圖是根據圖一的步驟產生的。

基於上述原因，這裡提出的方法沒有依賴廠商的混淆字表來更正錯誤，而是根據一個可被自動化處理的簡單原則，亦即圖一中步驟二的第 3 點：出現篇數高者視為正確，出現篇數低者視為錯誤。此原則根據一項假設：文件的 OCR 辨識率沒有很低，使得辨識正確的狀況高過辨識錯誤的狀況。

後面，將以實際的 OCR 文件來驗證上述的各種假設與更正步驟。

肆、實驗驗證

為瞭解 OCR 的辨識錯誤，對主題檢索的影響，我們曾製作了一份 OCR 文件檢索測試集，包括 8438 篇 OCR 文件及其原始影像，30 道主題檢索題目，以及每一道題目的相關文件判斷 [7,8]。這些文件是 1950 年到 1976 年間來自中國大陸、台灣、香港的新聞剪報，繁簡體都有，其中以簡體字佔大多數。這些剪報以 300 dpi（dot per inch）的解析度掃瞄成影像檔，再利用「丹青」辨識軟體，轉成數位文字，據我們估計其平均正確率約 69%。為製作適合此文件集的查詢主題，我們以同年代探討中國、台灣、香港等題目的 100 篇期刊、論文為對象，取其標題作為查詢主題的參考來源，最後刪除、改寫出 30 道查詢題目。圖五顯示一道查詢主題的例子。最後，我們找 3 位大學部、研究所同學，對每一篇文件，進行其與每一道查詢主題的相關判斷，做成這份適合主題檢索實驗的 OCR 文件測試集。

然而為評估本文提出的 OCR 錯字更正成效，我們沒有直接用這些查詢主題。由於 OCR 錯誤，會導致字彙不匹配問題（查詢詞與索引詞不一致的問題），進而可能導致檢索遺漏（漏檢了某些原來應該符合的文件），我們便以單字詞（single-term query）布林查詢的召回能力（recall capability），來驗證 OCR 錯誤更正的成效。

```
<topic>
  <num> 02 </num>
  <title>  麥克馬洪線  </title>
  <description>中共與印度間對於麥克馬洪線之爭論  </description>
  <narrative>
  相關文章內容包含中共或印度方面對於麥克馬洪線的看法或主張，若文章內容僅是對於中印邊境間之戰爭情形做報導則視為完全不相關。
  </narrative>
</topic>
```

圖五：OCR 檢索測試集之查詢主題範例。

為了準備一組客觀無偏 (unbiased) 的單字查詢詞，筆者以圖一的步驟，對這 8438 篇文件進行處理，獲得 3 萬多對詞彙（term pairs）。接著將詞頻高者（即中心詞彙）蒐集起來做成一個詞庫，利用這個詞庫對這 30 道查詢主題的全部內容進行斷詞處理，如此獲得了 20 個單字查詢詞，以模擬使用者可能下達的查詢詞彙。由於這些詞彙來自之前設計好的查詢主題，不是我們刻意安排或選擇的，應該能夠客觀反應出我們要測試的成效。

我們以這 20 個單字查詢詞，對 8438 篇 OCR 文件進行布林精確比對的檢索

（Boolean search），結果如圖六範例所示。圖六顯示在 8438 篇 OCR 文件中，兩個正確詞彙與其錯誤詞彙出現篇數的統計。其中：

第一欄：自動發現的「正確詞彙」（亦即圖一中提到的「中心詞彙」）。

第二欄：第一欄詞彙的文件篇數。

第三欄：自動發現的「錯誤詞彙」。

第四欄：第三欄詞彙的文件篇數。

第五、六欄：「正確詞彙」與「錯誤詞彙」的「混淆字元」。

第七欄：混淆字元在所有(正確詞彙與錯誤詞彙的)「詞彙對」出現的次數（例如：若「灣」、「瀉」是從「解放台灣、解放台瀉」，以及「台灣問題、台瀉問題」得來的，那麼這欄位的數字就是 2）。

第八欄：包含「錯誤詞彙」但不含「正確詞彙」的文件篇數。亦即，如果只用「正確詞彙」進行查詢，會漏檢的文件數量。

第九欄：同一「正確詞彙」的所有「錯誤詞彙」在第八欄中的文件篇數總和。此欄的數字代表用「正確詞彙」查詢不出來，但用其所有的「錯誤詞彙」能夠查出來的文件篇數。

1.	中	日	关	系	:	5	:	中	苏	关	系	:	3	:	日	:	苏	:	7	:	3	:	6
2.	中	日	关	系	:	5	:	中	捷	关	系	:	2	:	日	:	捷	:	3	:	2	:	6
3.	中	日	关	系	:	5	:	中	日	关	东	:	2	:	系	:	东	:	1	:	1	:	6
4.	解	放	台	湾	:	138	:	解	放	台	湛	:	6	:	湾	:	湛	:	3	:	6	:	76
5.	解	放	台	湾	:	138	:	解	放	台	沿	:	5	:	湾	:	沿	:	2	:	5	:	76
6.	解	放	台	湾	:	138	:	解	放	台	泻	:	1	:	湾	:	泻	:	2	:	1	:	76
7.	解	放	台	湾	:	138	:	解	放	台	海	:	9	:	湾	:	海	:	2	:	9	:	76
8.	解	放	台	湾	:	138	:	解	放	台	褐	:	1	:	湾	:	褐	:	1	:	1	:	76
9.	解	放	台	湾	:	138	:	解	放	台	渴	:	28	:	湾	:	渴	:	4	:	27	:	76
10.	解	放	台	湾	:	138	:	解	放	台	溜	:	24	:	湾	:	溜	:	3	:	24	:	76
11.	解	放	台	湾	:	138	:	解	放	台	门	:	2	:	湾	:	门	:	2	:	2	:	76
12.	解	放	台	湾	:	138	:	解	放	台	沁	:	2	:	湾	:	沁	:	1	:	1	:	76

圖六：兩個查詢詞及其相對應的錯誤詞彙。

把每個詞以「精確比對模式」拿來查詢後，這 20 個詞總共查出 3238 篇文件。以這 20 個詞的「錯誤詞彙」查詢，則總共可得出額外的 572 篇相關文件，但會額外得出 164 篇不相干的文件（像「中日關係」的詞彙用其歸類在一起的「中蘇關係」去查，就會得到另外 3 篇事實上沒有包含「中日關係」的文件，如圖六第一列所示）。因此，查全率（recall）最多會進步 572/(572+3238)=15.01%。查準率（precision）會退步 164/(164+572+3238) = 4.13%。整體而言，查詢 OCR 文件的成效有明顯的提升。

以上的成效沒有利用到任何人工維護的資源，如辭典、語料庫等，是全自動的方法做到的。如果有額外的資源可用，則可以輕易的提升其成效，亦即降低查準率退步的情況。

一個簡易的改進方法如下：假若我們有完全乾淨的文件，且這些文件跟 OCR 文件的領域、用詞差不多，則可以將圖六中第三欄的「錯誤詞彙」拿來與這些正確的乾淨文件做比對，如果這些「錯誤詞彙」出現在正確文件中，則我們可以得知，他們其實不是「錯誤詞彙」，而可以從上述的「詞彙對」中去除。例如：「中蘇關係」與「中捷關係」（中國與捷克的關係），若能從其他乾淨文件中比對到這兩個詞，則我們可以從圖六中將其刪除，查詢「中日關係」時，就不會用「中蘇關係」與「中捷關係」去查，使得查準率不會因此降低下來。

另外，從這些錯誤的「詞彙對」所得到的「混淆字元」，我們也可以知道這「混淆字元」是錯的（例如「日」與「蘇」、「日」與「捷」）。用這些錯的混淆字元可以回來更正更多 OCR 文件的「詞彙對」， 使得精確率的降低更為輕微，甚至完全消失。例如，如果有「美蘇衝突」與「美日衝突」這樣的「詞彙對」被圖一的步驟歸類在一起，則已知「日」與「蘇」不可能是混淆字元的情況下，這個「詞彙對」可以被刪除，因而不會造成錯誤的檢索。

伍、結語

本文提出的方法，與 OCR 技術運用的方法不同，其最大差異在於沒有用到 OCR 辨識原有的混淆字表，以及任何需要人工維護的詞庫與字典，而只用文件索引檔裡自動擷取的重複詞彙及其出現篇數。從檢索系統的索引檔中，我們可以比較出哪些長度相同的多字詞只差一個字，而將其列為可能的錯誤「詞彙對」（term pairs），在檢索時便能互相提示，達到詞彙更正以及提升檢索成效的的目的。我們的實驗顯示，查全率粗估可以提升 15.01%，而查準率則會受影響，退步 4.13%。但是若有其他資源，如類似主題的乾淨文件可用，則查準率的退步情況可被大幅改善。

本方法的一項直接應用，是運用於查詢提示：當使用者輸入某個查詢詞時，系統自動提示文件中相對於此查詢詞的所有錯誤詞彙，使用者可因此瞭解文件中真正記錄的詞彙，並挑選適當的詞彙進行檢索，使得 OCR 的錯誤對檢索成效的影響降到最低。另外，由於本方法與傳統 OCR 技術不同，也可運用本方法來加強 OCR 軟體的錯字更正能力。

由於本方法沒有用到任何需要人工維護的資源, 因此可立即適用於各個領域的文件，而無須額外付出任何代價。當然，如果有領域資源或知識可用，也可以用進來，以進一步提升其成效。

本方法可能的弱點，是錯誤詞彙必須重複出現，才能被擷取出來，進而被更正。我們觀察到，OCR 錯誤詞彙真的會重複出現，尤其是主題詞彙。但我們對個別文件的分析，也觀察到，有些影像品質較差的文件，其 OCR 後，錯誤率很高，錯誤的一致性也不明顯。可以說，一旦影像品質很差之後，幾乎所有的方法都會失效，這是筆者目前覺得最難以克服的部份。

誌謝：

本研究由國科會研究計劃補助，計劃編號： NSC 90-2413-H-030-004-。

參考文獻：

[1] Yuen-Hsien Tseng, "Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text", Journal of the American Society for Information Science and Technology, Vol. 52, No. 5, April 2001, pp. 378-390.

[2] Karen Kukich, “Automatically Correcting Words in Text,” ACM Computing Survey, Vol 24, No. 4, December 1992.

[3] 陳舜德，雜訊通道為本之文字辨識後處理系統，清華大學資訊科學研究所博士論文，1996 年 7 月。

[4] Kazem Taghva, Julie Borsack, Allen Condit, Srinivas Erva, "The Effects of Noisy Data on Text Retrieval," Journal of the American Society for Information Science, Vo.45. No. 1, 1994, pp.50-58.

[5] Wu, S., & Manber, U. “Fast Text Searching with Errors,” Technical Report, Dept. of Computer Science, Univ. of Arizona, Tucson, June, 1991.

[6] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, Nov. 2002, pp. 1130-1138.

[7] 蔡孟竹, 曾元顯, " 中文 OCR 文件檢索測試集之製作與應用", 「教育資料與圖書館學」, 第 40 卷, 第 3 期, 2003 年 3 月, 頁 325-344.

[8] Yuen-Hsien Tseng and Douglas W. Oard, "Document Image Retrieval Techniques for Chinese" Proceedings of the Fourth Symposium on Document Image Understanding Technology, Columbia Maryland, April 23-25th, 2001, pp. 151-158.