教育資料與圖書館學，40:3 (March 2003):325-344

中文OCR文件檢索測試集之製作與應用

蔡孟竹
碩士生
曾元顯
副教授
輔仁大學圖書資訊學系

摘要

本文描述一套中文OCR檢索測試集的建構過程及其實際的檢索應用。我們克服回溯性資訊需求難以獲得的困難，擬定出30道模擬使用者需求的查詢主題。爲獲得眞實的OCR文件，我們以OCR軟體將8439篇全文影像轉換成數位檔案，並評估其辨識率在7成上下。爲了求得每一道查詢主題的相關文件，我們邀請三位人員分別檢視並判斷每一篇文件是否跟查詢主題相關。經由Kendall和諧係數的統計驗證，這三位判斷者在20道查詢主題上，相關判斷的結果非常一致，顯示標準答案(即相關文件)有足夠的共識。最後，以12種檢索策略來比較OCR文件的檢索成效，我們發現辨識率降低到7成的情況下，檢索成效差不多也降低到7成左右。

關鍵詞：光學文字辨識，資訊檢索，測試集，成效評估，中文檢索

前　言

目前資訊檢索系統所能檢索的文件，多是以文字符號為基礎的數位化文件。相對而言，許多儲存於傳統媒體之紙本資料，並無法直接被現有之檢索系統處理與利用。而若欲以人工方式，將儲存於傳統媒體之文件轉化為數位化(symbolic)之檔案，實為一曠日廢時之工作。幸好科技進步，可利用掃描器將傳統紙本文件掃描成影像檔，並以光學文字辨識系統(Optical Character Recognition，以下簡稱OCR)將影像檔中之文字辨識出來，即可成為可供檢索利用的數位化OCR文件。

然而OCR系統的辨識正確率並非百分之百，先前研究指出，要使OCR系統能達到良好的成效，被掃描的原始文件不可有太多污點，例如原始打字稿，或極乾淨清

晰之影印本等，且該文件要盡可能具有簡單的版面配置、段落格式，以及非常普遍的字型。OCR系統之成效往往會隨著輸入影像的品質低劣，而出現較多的錯誤文字或錯亂章節段落等「雜訊」(noise)，故OCR文件也簡稱雜訊文件(註1)。在此情形下，檢索OCR文件是否有效，無疑是值得關注及研究的議題。國外已有這樣的研究(註2)，但針對中文的OCR文件檢索研究則較為稀少(註3)。究其原因，大體上可歸結於實驗環境的缺乏。這促成了本研究建立「中文OCR文件檢索測試集」的動機，希望透過這項資源的建立，讓中文OCR文件檢索的研究得到較好的發展。

二、相關研究

一般情形下，使用者進行資訊檢索時，是將資訊需求表達成查詢問句，輸入檢索系統，系統把可能滿足使用者需求之文件列表輸出給使用者，使用者檢視這些結果以挑出其判斷為符合資訊需求的文件。資訊檢索系統之評估就模擬這樣的程序，以測試集(test collection)做為系統測試評估的環境。因此，測試集通常包含三個部分：一群待檢索文件組成的文件組(document collection)、數道代表使用者資訊需求之查詢主題(query topic)，以及由人工決定的文件與查詢主題間的相關判斷結果(relevance judgment)(註4)。各種創新的不同檢索方法，透過同一測試集的試驗，彼此間的成效優劣將有比較的相同基礎，如此的比較結果才有意義，真正優秀的檢索方法才能正確的突顯出來。過去資訊檢索研究的進展，一大部分要歸功於檢索測試集的製作與應用，才得以有所成果。因此，以下簡略回顧幾個重要檢索測試集的概況，作為我們發展OCR測試集的參考。

(一)Cranfield II 研究

資訊檢索測試集之架構，可謂起源於Cyril W. Cleverdon於1960年代進行的Cranfield II 研究。Cleverdon以航空動力學領域為主，比較33種不同索引方法的檢索

註1 Tapas Kanungo & Philip Resnik, "The Bible, Truth, and Multilingual OCR Evaluation," In *Proceedings of SPIE Conference on Document Recognition and Retrieval VI, San Jose, Canada, January 27-28, 1999*, p.87; 及Tapas Kanungo, Gregory A. Marton, & Osama Bulbul, "Performance Evaluation of Two Arabic OCR Products," In *Proceedings of the 27th Applied Imagery Pattern Recognition Workshop on Advances in Computer Assisted Recognition, Washington, DC, October 14-16, 1998*, pp.76-77.

註2 針對國外語文(英文、法文、德文、西班牙文、阿拉伯文等)的OCR文件檢索，相關的研究請參閱文末附錄。

註3 Yuen-Hsien Tseng, "An Approach to Retrieval of OCR Degraded Text," *Journal of Library Science*, (National Taiwan University), (Dec. 3, 1998): 153-168.

註4 Justin Zobel, "How Reliable Are the Results of Large-Scale Information Retrieval Experiments?" In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 24-28, 1998*, p.307;
Ellen M. Voorhees, "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness," *Information Processing and Management*, 36: 5 (September 2000): 697-698; 及
Gordon V. Cormack, Christopher R. Palmer, & Charles L. A. Clarke, "Efficient Construction of Large Test Collections," In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 24-28, 1998*, p.282.

成效。為了進行研究，Cleverdon需要先建立一個可供檢驗成效的實驗環境。Cleverdon經過種種程序，集合了1400篇文獻組成之文件組、279個查詢問題，以及此279個查詢問題和1400篇文獻之間的相關判斷結果，形成了一個測試集，做為Cranfield Ⅱ研究所需之測試環境，並採正規化回收率（normalized recall）為主要的評估指標。Cranfield Ⅱ研究首開運用測試集做為資訊檢索系統評估環境的先例，對於資訊檢索系統評估之發展有重大之影響(註5)。

(二)TREC

自1992年開始，美國每年舉辦的文件檢索會議（Text REtrieval Conference，以下簡稱TREC），已成為資訊檢索領域極受矚目之焦點。其建構了大型測試集，多種測試項目，以及不同之評估方式，提供了不同資訊檢索系統間的標準測試評估環境，且舉辦論壇來討論及分享結果(註6)。TREC之測試集是依據Cranfield研究的概念擴展而來，故其測試集主要也分為文件組、查詢主題，以及相關判斷三部分(註7)。以下對其測試集之組成做一簡述。

1.文件組

TREC的文件組內容以報紙文章為主，共有約200萬篇文件，5GB之資料量，經壓縮後儲存於五片光碟(註8)。文件組之文件均以SGML(Standard Generalized Mark-Up Language)加以標示，以利資訊檢索系統進行處理。

2.查詢主題

為了因應不同測試項目之需求，歷屆TREC建構了許多查詢主題(topics)。自第5屆起，查詢主題的結構均由Title、Description及Narrative三個欄位構成，如圖1顯示的查詢主題範例。Title欄位以至多3個字的長度，給這個查詢主題一個名稱。Description欄位則常以一個句子來敘述查詢主題的內容。Narrative欄位則提供必要的說明，讓判斷者可以精確的判斷哪些文件跟此主題相關(註9)。

查詢主題是由TREC主辦單位聘請的專家，根據其本身的興趣專長建構而成。首

註5 Michael Keen, "Cyril W. Cleverdon," *The Journal of Documentation*, 54: 3 (June 1998): 269;
Cyril W. Cleverdon, "The Significance of the Cranfield Tests on Index Languages," In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, 1991*, p.7;及
Gerard Salton, "The State of Retrieval System Evaluation," *Information Processing and Management*, 28: 4 (1992): 442-444.

註6 陳光華、江玉婷，〈中文資訊檢索測試集之設計與製作〉，資訊傳播與圖書館學，6：3(民國89年3月)：62。

註7 Ellen M. Voorhees & Donna Harman, "Overview of the Sixth Text REtrieval Conference (TREC-6)," *Information Processing and Management*, 36: 1 (January 2000): 8.

註8 Ellen M. Voorhees & Donna Harman, "Overview of the Ninth Text Retrieval Conference (TREC-9)," <http://trec.nist.gov/pubs/trec9/papers/overview_9.pdf>.
Ellen M. Voorhees, "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness," *Information Processing and Management* 36: 5 (September 2000): 699.

註9 Ellen M. Voorhees & Donna Harman, "Overview of the Eighth Text REtrieval Conference (TREC-8)," <http://trec.nist.gov/pubs/trec8/papers/overview_8.pdf>.

先，查詢主題之建構者提出數道查詢主題供候選之用；繼之，利用主辦單位發展出來的PRISE檢索系統在文件組中試行檢索，預估每道查詢主題可能的相關文件數量；最後，篩選出數道具有適當相關文件數量之查詢主題供TREC使用。

```
<top>
<num> Number: 406
<title> Parkinson's disease
<desc> Description:
   What is being done to treat the symptoms of Parkinson's disease
and keep the patient functional as long as possible?
<narr> Narrative:
   A relevant document identifies a drug or treatment program
utilized in patient care and provides an indication of success or
failure.
</top>
```

圖1　TREC之查詢主題範例

(資料來源：<http://trec.nist.gov/data/topics_eng/topics.401-450.gz>)

3.相關判斷

相關判斷(Relevance Judgments)是由查詢主題之建構者來進行，對於查詢主題與文件之相關關係採用二元化的尺度，也就是只分為相關與不相關(註10)。由於TREC文件組的數量十分巨大，一篇文章依篇文章的判斷，相當耗時費力，因此採用Pooling的方法來輔助相關判斷的進行。此方法把參與TREC評比的檢索系統，將其檢索結果依系統評估的相關程度由大到小排序，再抽取各系統查詢結果的前n篇文件，合併形成一個Pool，在去除Pool中重覆的文件後，此Pool就做為該查詢主題之相關文件候選組，送回給該查詢主題的建構者進行相關判斷。至於未被列入查詢主題相關文件候選組的文件，則一律假設與查詢主題不相關(註11)。

在建構者完成相關判斷後，列表整理查詢主題與文件之相關關係。此相關判斷結果，就如同一套標準答案，指出每個查詢主題有哪些文件與其相關。故檢索系統的檢索結果，就可與此相關判斷結果做比較，因而得以評估系統之成效。

(三)NTCIR

由於資訊檢索之蓬勃發展，以及TREC之成功，鼓舞了更多測試集的出現。例如日本的NTCIR資訊檢索評估會議，提供了中、英、日、韓等語文測試集，並舉行不同的競賽測試項目，來評估各參賽隊伍的資訊檢索系統成效(註12)。其中文測試集，是由台灣大學陳光華教授與江玉婷小姐等人建構，包括了13萬篇從網路下載之報紙文

註10 "Text REtrieval Conference (TREC) Data - English Relevance Judgements," <http://trec.nist.gov/data/rel-judge_eng.html>.

註11 江玉婷、陳光華，〈TREC現況及其對資訊檢索研究之影響〉，圖書與資訊學刊，29期（民國88年5月）：43-44。

註12 NTCIR Workshop, 〈http://research.nii.ac.jp/ntcir/workshop/work-en.html〉

章所組成之文件組，50道查詢問題，以及採四點式尺度之相關判斷結果(註13)。今年更擴增至38萬篇中文文件。NTCIR之出現，鼓舞了東亞資訊檢索研究者，提昇資訊檢索系統評估概念，提供交流管道，並增進資訊檢索系統之成效(註14)。

三、測試集之製作

近年來資訊檢索之課題之一：「OCR文件檢索」，尤其中文OCR文件檢索，至目前為止，並沒有合適之測試集可供應用。有鑑於此，本研究參考了其他測試集之架構，建立了「中文OCR文件檢索測試集」，其內容包含三個部分：

文件組：有文件影像，OCR文件，以及「部分乾淨文件」三個單元。

查詢主題：30道查詢主題。

相關判斷：文件與查詢主題間之相關判斷結果。

本節分別就此三個部分，對於其製作過程及組成內容加以陳述。

（一）文件組之建立

1.資料來源

OCR文件檢索是本研究關注之焦點，故須先尋得文件影像，以供OCR系統辨識輸出OCR文件，而使資訊檢索系統得以在其中進行檢索。

根據行政院大陸委員會所出版的《臺灣地區大陸研究及資料單位簡介》一書介紹，輔仁大學中國社會文化研究中心(以下簡稱社文中心)，其館藏在研究大陸問題過程中，為具有特殊參考價值的重要資料來源。其收藏有自1949年中共政權掌握大陸後，大陸、香港及台灣的報紙、廣播抄稿、人名機構卡片，及其他印刷物品。社文中心並對這些報紙的文章進行人工篩選、剪貼、標示、分類等工作，整理成一篇篇之剪報，全部大約有七、八十萬篇之多。在這些剪報中，已有約60萬篇剪報掃描成影像檔。

社文中心對於剪報內容概分為三大類，其下尚可按細目分類共計100多項以上，可見其內容之豐富多元。故研究者與該中心研究人員討論之後，決定選擇內容較易為一般人士所了解之軍事及外交兩類之新聞，年代範圍涵蓋1950~1976年，共22108篇，文字以中文簡體字為主，少數有中文繁體字的剪報，做為文件影像的初步來源。圖2顯示一剪報影像的範例。

受限於相關判斷實施等程序需耗費大量人力及時間，研究者擬對文件影像之數目加以限制。在便於實行之考量下，採集群(Cluster)抽樣法，將之前選出的文件影像，按其原始檔案編號順序分成45個集群後，隨機抽出23個集群，共11108篇文件影像，做為供OCR系統辨識的文件影像來源。

註13　江玉婷，中文資訊檢索測試集設計與製作之研究，國立臺灣大學圖書資訊學研究所，碩士論文，民國88年7月。

註14　陳光華，〈資訊檢索系統的評估—NTCIR會議〉，在國立臺灣大學圖書資訊學系四十週年系慶學術研討會，台北市，2001年11月16日，頁84。

2.文件組－文件影像與OCR文件單元之建立

我們請四位助理人員，運用市售之OCR系統，即「丹青文字辨識系統」，將11108篇文件影像辨識為純文字檔形式之OCR文件。而該OCR系統之採用係考量資源之可得性，以及系統功能之適用性。當時曾評估過三種系統，試用過後，以丹青文字辨識系統較能符合當時所需。四位助理人員在兩個工作天內，完成此項數位影像轉數位文字之程序。

57.2.28.台盟总部和北京、上海支部昨日集会 P10

紀念台灣人民"二·二八"起义十周年

謝雪紅号召台灣同胞为和平解放台灣作出更大的貢献

新华社27日訊 台灣民主自治同盟总部和北京市支部今天为紀念台灣人民"二·二八"起义十周年举行座談会。

座談会由台灣民主自治同盟主席謝雪紅主持。她在会上致詞說：当我們今天紀念"二·二八"起义十周年的时候，我們台灣人民应該更加緊密地团結起来，为和平解放台灣作出更大的貢献。

副主席李純青也在会上講了話。他說，爭取和平解放台灣不仅对台灣人民有利，而且对世界和平也是有利的。不管采取什么方式，台灣終归是要解放的。他說，只有台灣人民和祖国人民一道，只有全中国人民的力量，台灣才能得到解放，台灣人民才能回到祖国的怀抱。

出席座談会的除在北京的部分盟員外，还有在北京的台灣同胞和从日本、东南亚等地归国的台灣同胞共七●多人。許多人都很激动地在会上發了言，他們一致認为，"二·二八"起义英勇斗爭的革命历史，更加坚定了他們解放台灣的信心。他們表示，要在自己的岗位上，努力工作、学習，为爭取和平解放台灣貢献出自己的力量。

本报上海27日專电 台灣民主自治同盟上海市支部为紀念台灣人民"二·二八"起义十周年，今晚召集在上海的盟員和部分台灣同胞举行座談会。台盟上海支部委員林田烈作了題为"齐心协力爭取和平解放台灣"的报告，他在結合国内外形势說明和平解放台灣政策的精神和意义后說，和平解放台灣的号召已經在国内外包括蔣介石集团内部發生了重大影响，和平解放台灣的可能性正在不断增長。林田烈号召台盟成員和在上海的台灣同胞要和全国人民一道更緊密地团結在中国共产党的周圍，以积極的劳动热情响应增产节約的号召，为爭取和平解放台灣、实現祖国的完全統一而奋斗。

到会者对現在处在美国操制下的故乡亲人非常怀念，他們在發言中表达了台灣同胞盼望早日回到祖国怀抱的心情，并一致痛斥美国非法控制台灣的罪行，坚决反对所謂"兩个中国""台灣独立"和双好屢文鞍泰美国主子之命組織所謂"台灣共和国臨时政府"等試圖分裂中国領土的陰謀活动。到会者紛紛表示要繼承"二·二八"的光輝革命傳統，做好本崗位工作，以实际行动爭取早日解放台灣。

圖2 輔仁大學中國社會文化研究中心剪報影像範例

(資料來源：輔仁大學中國社會文化研究中心)

為使文件影像與OCR文件之內容可相互對照，各OCR文件之主檔名部分均比照其辨識時使用之文件影像檔名。例如，若文件影像檔名為123.tif，則其經OCR系統辨識所得之OCR文件檔名即為123.txt。受限於文件影像本身掃描品質、所含格式段落，以及OCR系統辨識能力等之限制，11108篇文件影像經OCR系統辨識後，僅得到8439篇有效的OCR文字檔案。其餘影像檔案，OCR軟體不是讀不進去，不然就是辨識不出任何文字。圖3為一份OCR文件的範例。

本研究就以此8439篇OCR文件，以及與其內容相對應之8439篇文件影像，分別做為文件組之OCR文件單元，以及文件影像單元。

OCR系統在進行文字辨識過程時，會將文件影像中所含中文簡體字自動轉為相對應之中文繁體字，故所有OCR文件均以BIG-5編碼之繁體中文形式儲存。然為便利世界各地人士利用，這些OCR文件也經文字編碼轉換軟體，轉換成為GB編碼之簡體中文形式。

3.文件組－部分乾淨文件單元之製作

為探討OCR文件之檢索成效，若文件組中有乾淨文件組成之單元，做為對照之用，則更為理想。在文件數量龐大，以及人力、時間等因素限制下，無法就文件組全部影像建立其正確無誤的數位文字檔案，我們僅建立一部分的乾淨文件。我們挑選那些被判定為與某個查詢主題相關的文件，共899篇，進行人工重新輸入文字的作業。

完成899篇乾淨文件後，以這些乾淨文件，取代OCR文件組中對應的899篇OCR

文件，而形成了文件組之「部分乾淨文件」單元，一樣也是8439篇文件。

4.文件組之內容

文件組中OCR文件及部分乾淨文件此兩單元，是由文件組之文件影像單元經由OCR辨識或部分人工輸入建立而成，故此三個單元事實上可視為三種不同剪報表現形式的版本，涵蓋了一致的內容範圍。

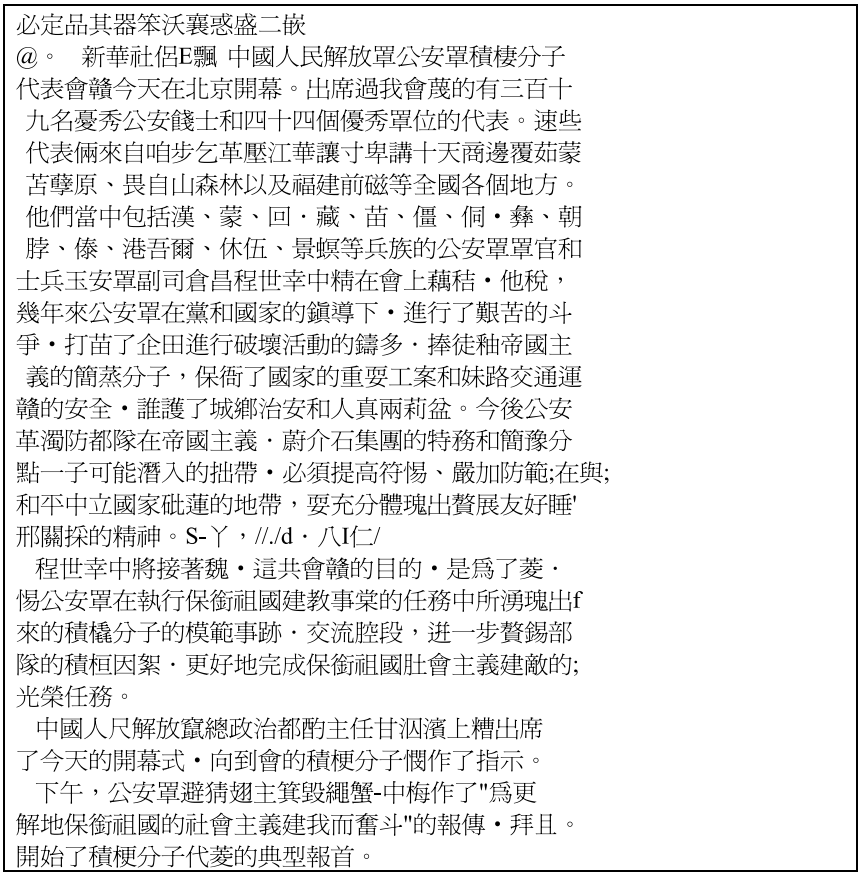
必定品其器笨沃襄惑盛二嵌
@。　新華社侶E飄 中國人民解放罩公安罩積棲分子
代表會贛今天在北京開幕。出席過我會蔑的有三百十
九名憂秀公安餞士和四十四個優秀罩位的代表。速些
代表倆來自咱步乞革壓江華讓寸卑講十天商邊覆茹蒙
苫孽原、畏自山森林以及福建前磁等全國各個地方。
他們當中包括漢、蒙、回・藏、苗、僵、侗・彝、朝
脖、傣、港吾爾、休伍、景螟等兵族的公安罩罩官和
士兵玉安罩副司倉昌程世幸中精在會上藕秸・他稅，
幾年來公安罩在黨和國家的鎮導下・進行了艱苦的斗
爭・打苗了企田進行破壞活動的鑄多・捧徒釉帝國主
義的簡蒸分子，保衙了國家的重耍工案和妹路交通運
贛的安全・誰護了城鄉治安和人真兩莉盆。今後公安
革濁防都隊在帝國主義・蔚介石集團的特務和簡豫分
點一子可能潛入的拙帶・必須提高符惕、嚴加防範;在與;
和平中立國家砒蓮的地帶，耍充分體瑰出贅展友好睡'
邢關採的精神。S-丫，//./d・八I仁/
程世幸中將接著魏・這共會贛的目的・是爲了菱・
惕公安罩在執行保銜祖國建教事棠的任務中所湧瑰出f
來的積橇分子的模範事跡・交流腔段，迸一步贅錫部
隊的積桓因絮・更好地完成保銜祖國肚會主義建敵的;
光榮任務。
中國人尺解放竄總政治都酌主任甘泅濱上糟出席
了今天的開幕式・向到會的積梗分子憫作了指示。
下午，公安罩避猜翅主箕毀繩蟹-中梅作了"爲更
解地保銜祖國的社會主義建我而奮斗"的報傳・拜且。
開始了積梗分子代菱的典型報首。

圖3　OCR文件範例（繁體中文形式）

以文件內容類別分析，在各單元的文件中，外交類共有4785篇文件，軍事類則有3654篇文件，外交類所佔比例略多於軍事類。

就文件影像單元而言，每篇文件影像均為4133×2720個像素，解析度為300 DPI之TIFF格式黑白影像。平均每篇文件影像之檔案大小為105 KB，全部8439篇文件影像之檔案大小約850 MB。

就OCR文件單元而言，平均每篇文件之檔案大小為3KB，全部8439篇文件約為24 MB。字數方面，平均每篇文件約1293字，以字數較長之文件為主，全部8439篇文件的總字數為10,915,454字。

就部分乾淨文件單元而言，平均每篇文件之檔案大小為3KB，全部8439篇文件

約為24 MB。字數方面，平均每篇文件約為1302字，以字數較長之文件為主，全部8439篇文件之總字數為10,985,333字。事實上，由於與OCR文件單元相較，本單元僅對其中之899篇文件以人工方式重新建立，故各項數據均十分接近。此兩個單元之統計數字，摘要顯示於表1。

表1　文件組中文字之統計

單元	文件總數	字數平均數	字數中位數	字數標準差	字數總和
OCR文件	8439	1293.454	1164	886.864	10915454
部分乾淨文件	8439	1301.734	1172	872.855	10985333

5.OCR文件之辨識率(精確率與召回率)

資訊檢索系統是針對文件的內容文字來檢索，故內容文字的辨識正確率如何影響檢索成效是值得探討的。受限於人力，在完成測試集之相關判斷後，研究者僅先就被判斷為相關的899篇OCR文件，觀察其內容文字之辨識率。進行的方式是以人工輸入的文件，與其相對應的OCR文件做比較，計算同樣一篇新聞在這兩種版本上每個字的相同與差異情況，把每一篇新聞兩種版本的「相同總字數」除以「OCR文件的總字數」，得到該篇新聞的辨識「精確率」。加總所有的精確率除以899後，平均每篇OCR文件之精確率為77.5%。至於「召回率」的計算，則是把每一篇新聞兩種版本的「相同總字數」除以「人工輸入文件的總字數」。其平均的召回率為75%。圖4顯示這899篇OCR文件精確率與召回率的分佈情形。表2顯示一篇假設文件的計算範例。這種計算方法忽略文字的位置與順序，只能算是一種降低人力成本的概略估算。

表2　OCR文件精確率、召回率計算範例

字	一	七	了	二	人	入	力	下	上	又	三	八	刀	合計
乾淨文件	19	1	11	1	25	2	9	4	4	1				78
OCR文件	22	2	8	1	19	5	6	2	4		1	4	1	75
相同字數	19	1	8	1	19	2	6	2	4	0	0	0	0	62

精確率=62/75=82.67%，召回率62/78=79.49%

圖4的橫軸為按召回率由低至高排列之文件編號，非原始編號。精確率由於上下起伏劇烈，導致畫面不易閱讀，在此以其對數平滑曲線顯示。

事實上，我們也曾進行過另外一種計算其辨識精確率的方式。在運用OCR軟體進行文字辨識時，曾觀察一組1300篇文件的辨識情形，把OCR軟體對每一篇文件辨識後報告出來的「辨識總字數」扣去OCR軟體報告出來的「可疑字數」，再除以「辨識總字數」，得到每一篇的精確率。這些文件的精確率分佈圖形與圖4近似。加總所有的精確率除以文件數後，平均每篇OCR文件之精確率為69%(註15)。這個方法就沒

註15 Yuen-Hsien Tseng & 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.

有辦法估算召回率了，因為系統辨識出來的字不一定是對的，所以其報告出來的「辨識總字數」不一定是全部正確字的總字數。同理，用此方法得到的精確率也不是非常準，只能說是個估計數字，因為系統認為辨識對的，不一定真得對。總之，我們用上述兩種方法估計出來的數據，可以透露此OCR文件單元的辨識率，平均在70%上下。但個別文件辨識的程度則如圖4所示，最高到最低的數據相差很大。

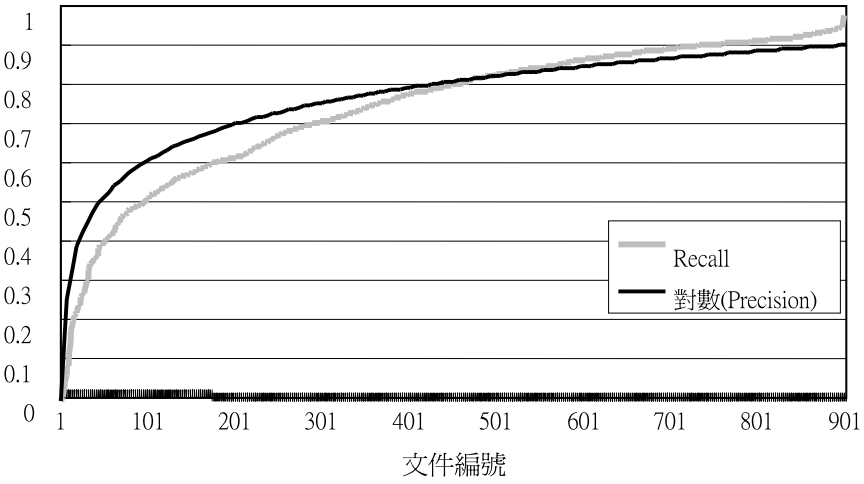

圖4　OCR文件召回率與精確率

(二)查詢主題之建構

1.資訊需求之搜尋

查詢主題的擬定，須事先蒐集跟文件組內容相關的資訊需求。而資訊需求最好是來自真實使用者的經驗，使查詢主題可以反映真實狀況。由於社文中心並沒有對該中心剪報使用者記錄其資訊需求與尋求過程，對本研究而言，少了一個直接獲取真正使用者資訊需求的途徑，使得必須另尋其他可行方法。對於建構回溯性資料的資訊檢索測試集，資訊需求的蒐集是令人較感困擾的部分。

所幸我們觀察到，本測試集的文件組部分，係以中國大陸1950年至1976年間之軍事、外交為主題。過去這一段時間，許多社會科學方面的學者、專家會將其自身，或民眾所關心的這方面議題，以觀察、評論，或追蹤報告等形式，撰寫為論文，並登載於期刊或發表於會議。而欲完成這些論文，這些學者、專家必然有其真實之資訊需求，並應經歷一番相關資訊的搜尋過程。故這些論文就其題名部分而言，實包含了這方面真實且值得探討之主題，可視為實際之資訊需求。因此研究者便收集這些論文題名，以建構出查詢主題。

欲收集論文題名，利用圖書館一向重視的索引，可說是最好的選擇。在比較過各索引之收錄年代、內容範圍等特性之後，研究者決定利用政大國際關係研究中心所編輯的「大陸問題專論索引」來收集論文題名。「大陸問題專論索引」之內容分為黨政、人權、外交、軍事等13個類別。故研究者從該索引之軍事、外交兩個類別

著手，收集符合本研究文件組涵蓋年代的論文，共收集了80筆論文題名，做為初步查詢主題建構之依據。

2.資訊需求之篩選

由於收集到的80筆論文題名，可能與文件組內容間毫無關聯，因此我們利用Crystal檢索系統進行檢索(註16)，初探在文件組中各論文題名可能找出相關文件的數目，以對此80筆論文題名進行篩選。

本研究預計建構30道查詢主題，故根據檢索結果以人工檢視，篩選出可能相關文件數在前30名之論文題名，做為查詢主題之建構依據。

3.查詢主題之組成內容

本研究之查詢主題內容係以前述方式篩選出30筆論文題名，參考TREC之查詢主題結構加以轉化改寫而成，主要以Title、Description，及Narrative等欄位組成。

查詢主題之轉換改寫過程，均由研究者與實際進行相關判斷之三位判斷者，經反覆、充分的討論來進行。如此可讓查詢主題具有一定之明確與詳盡程度，且可增加相關判斷者對於查詢主題的認知程度，使本研究之相關判斷結果能更為正確及客觀，進而提高此測試集整體的可靠性。

<top>
<num> Number：08
<title> 南沙群島
<desc> Description：
中共方面對於南沙群島主權之主張
<narr> Narrative：
相關文章內容包含中共方面對於南沙群島主權之主張或看法，若文章內容僅提及其他國家對南沙群島之侵略情形則視爲完全不相關。
</top>

圖5 查詢主題範例（中文形式）

每一道查詢主題包括以下欄位內容，如圖5所示。

<top>：起始標記，代表查詢主題之開始。

<num>：編號，以兩位數字組成，以供識別。查詢主題彼此間之順序並未具有任何特殊意義。

<title>：標題，以一個名詞或名詞片語組成，是查詢主題中對於資訊需求字數最簡短，但也是最廣義的描述。意即標題所隱含之意義、範圍可能會大於或相等於查詢主題中<desc>或<narr>欄位內容所包含的。<title>欄位的內容係由研究者及相關判斷者共同討論，待查詢主題中之<desc>欄位內容建構完成後，從中挑選並改寫而成。

<desc>：陳述，以一句句子敘述資訊需求，為查詢主題中最主要之欄位。此欄位之內容係由篩選出的30筆論文題名，經研究者及相關判斷者共同討論，保留各筆論文

註16 曾元顯、林瑜一，〈模糊搜尋、相關詞提示與相關詞回饋在OPAC系統中的成效評估〉，中國圖書館學會會報，61期(1998年12月)：103-125。

題名所含之意旨，並加以適度之修飾轉化，改寫成語意明確、清晰之問句形式。

<narr>：說明，以數句句子來說明相關或不相關文件所具備之條件，有助於釐清查詢主題中<desc>欄位之內容。<narr>欄位的內容係待相關判斷者依據<desc>欄位內容與文件組之文件做完相關判斷後，由研究者及判斷者共同分析相關判斷的過程，而撰寫出來的相關判斷輔助說明。

</top>：結束標記，代表查詢主題之結束。

本研究依上述過程共建立了30道查詢主題，依其內容類別劃分，有17道查詢主題屬於外交類，而其餘之13道則屬於軍事類。表3顯示這30道主題各欄位的字數統計。

表3　查詢主題各欄位字數統計

欄位	最小字數	最大字數	平均字數	中位數	標準差	變異係數
<title>	2	11	4.467	4	2.193	49.097%
<desc>	7	25	14.800	14	3.537	23.899%
<narr>	23	80	52.533	54	15.878	30.225%
查詢主題總字數	40	107	71.800	72	18.115	25.230%

為提供更廣泛的應用，例如英、中跨語言資訊檢索(Cross Language Information Retrieval)，經社文中心研究人員的協助，這30道查詢主題均翻譯成英文形式，如圖6所示，以擴展其可能的應用範圍。

<top>
<num> Number: 08
<title> Spratly Islands (Nansha Islands)
<desc> Description:
The PRC's sovereignty claim over the Spratly Islands.
<narr> Narrative:
The articles should deal with the PRC's claims over the Spratly Islands.　Articles dealing only with the invasion of the Spratly Islands by other countries are irrelevant.
</top>

圖6　查詢主題範例（英文形式）

(三)相關判斷之實施

1.相關判斷者

為求增加相關判斷結果之客觀性，避免僅憑單一判斷者之判斷結果過於主觀之疑慮，每一道查詢主題均以三位人員進行相關判斷。三位判斷者分別具有學科背景專長或檢索專長，與真實檢索環境中可能使用者自行檢索，或由圖書館員(資訊中介者)代為檢索之情形相似。表4列出這三位判斷者的特質。

表4　三位相關判斷者之特質

相關判斷者	特　質	說　　明
A	學科專長	歷史系大三學生，具有與查詢主題及文件組內容相符之學識背景
B	學科專長	歷史系大三學生，具有與查詢主題及文件組內容相符之學識背景
C	檢索專長	圖資系畢業之圖書館員，從事讀者服務工作，具有豐富之資訊檢索經驗

2.相關判斷評量尺度及程序

本研究將查詢主題與文件之間的相關關係，依程度由高至低分為完全相關、部分相關，及完全不相關，並分別給予2分、1分及0分之相關分數。

由於若以Pooling方式來輔助相關判斷，須採用多個資訊檢索系統來進行，在難以取得多個合適之中文資訊檢索系統的情形下，我們採用「完整判斷」(exhaustive judgment)的方式進行相關判斷的工作，亦即每一篇文件都拿來檢視並判斷其會與哪一道查詢主題相關，從而得到每一道查詢主題的相關文件。

在進行相關判斷時，為避免OCR文件的錯誤段落與文字造成影響，人工檢視的是影像文件而非其OCR文件。三位判斷者分別利用電腦螢幕觀看文件影像，逐一與30道查詢主題之<desc>欄位內容比對，進行相關判斷，並按其所判定之相關程度給與相關分數。此完整判斷的工作相當耗時，在兩個月內，三位判斷者共做了3 ×30 × 8439 = 759,510次相關判斷。

3.相關判斷結果之建立

完成相關判斷後，基於每位判斷者的結果均同等重要的原則，將三位判斷者對同一篇文件給出的相關分數加總，做成該文件對於某查詢主題的最終相關分數，如表5所示。每篇文件之相關分數介於0(三位判斷者均判斷該文件與查詢主題不相關)至6(三位判斷者均判斷該文件與查詢主題完全相關)。對每一道查詢主題，將其相關分數不為0的文件集結、列表，就成為可供資訊檢索系統評估檢索成效之相關判斷結果。

目前評估資訊檢索系統成效之指標，多採二元化之相關判斷尺度(相關、不相關)為計算之基礎。然而本研究為求接近真實狀況中相關判斷具有程度上的差異，遂採三點式之尺度(完全相關、部分相關、完全不相關)；且將三位相關判斷者之分數加總成為文件之相關分數。這些介於0至6的分數雖無法直接套用於二元尺度的成效計算，但依需求目的之不同，採用適當的門檻值，仍然可將相關分數轉化為二元化之值，即0(低於門檻值者)與1(高於門檻值者)。

表5　查詢主題相關文件列表範例

查詢主題編號	文件編號	相關分數			相關分數
		相關判斷者A	相關判斷者B	相關判斷者C	
17	0007619	2	2	2	6
17	0007620	1	1	1	3
17	0008326	1	0	1	2
17	0053678	2	2	1	5
17	0054802	2	1	1	4
17	0054803	2	2	2	6
17	0054804	2	2	1	5

4.查詢主題與文件間之相關關係

30道查詢主題的所有相關文件，共包含了930篇文件，但因有31篇文件與一道以

上之查詢主題存在相關關係，扣除重複後僅有899篇不同之文件。相關判斷程序實施時，係以文件影像為之，故只要與此899篇文件影像內容相對應之OCR文件，即視為與查詢主題存在相關關係。

就各查詢主題相關文件之數量觀之，最多的為125篇，最少的則只有4篇，中位數為16，代表有半數之查詢主題相關文件數量在16以下，而標準差達到34.756，也顯示各主題相關文件數量十分不均，如表6所示。由於本研究的文件內容為新聞剪報，涵蓋時間長達25年，故若查詢主題內容屬發生時間持續較久或熱門之議題，則可能會有較多文件與查詢主題相關。相對的，若查詢主題內容不屬於以上情況，則該查詢主題相關文件數量多半顯得較少。

表6　查詢主題相關文件數量及平均相關分數

查詢主題編號	內容類別	相關文件數	相關文件組平均相關分數
01	外交	44	3.318
02	外交	22	4.409
03	外交	24	3.625
04	外交	7	4.714
05	外交	125	3.128
06	軍事	61	2.377
07	外交	28	2.643
08	外交	9	3.889
09	軍事	90	3.022
10	外交	10	5.100
11	外交	13	4.538
12	外交	25	4.000
13	外交	4	5.750
14	外交	5	4.400
15	外交	9	2.333
16	外交	12	5.667
17	外交	7	4.429
18	軍事	66	3.318
19	軍事	31	4.742
20	軍事	11	4.545
21	軍事	7	3.000
22	軍事	12	4.250
23	軍事	93	3.591
24	軍事	5	5.000
25	外交	5	4.400
26	軍事	19	3.263
27	軍事	13	4.846
28	軍事	22	4.500
29	軍事	28	5.143
30	外交	123	2.805
中位數		16	4.325
平均數		31	4.025
最大值		125	5.750
最小值		4	2.333
標準差		34.736	0.945

5.相關判斷的一致性

由於檢索系統之成效數據是依據相關判斷結果加以計算，因此相關判斷的結果是否可靠、值得信賴，是非常重要的問題。本研究之相關判斷結果是否值得信賴，可由三位判斷者之判斷結果是否具有「一致性」(consistency)及「穩定性」(stability)來考量。

相關判斷是一種主觀的心智活動，無法非常客觀地衡量，判斷者的個人特質也可能會影響判斷的結果。判斷結果可能會全面高估或低估，甚至個人前後判斷寬鬆不一，形成時而高估，時而低估之情形。因此，若直接比較三位判斷者做出來的結果，即文件之相關分數，可能會有許多歧異出現。由於相關分數屬於順序尺度，直接將三位判斷者評定的分數差異，視為判斷結果之不同，並無太大意義，而應該重視相關分數順序上的意義。此道理就如同有A、B、C三位閱卷者對同樣30份考卷做評分，每位閱卷者對同一份考卷所評定之分數可能會有所不同，但若此閱卷者本身均具有一致且穩定之評分原則，且彼此間的評分結果具有其一致性，則將此30份考卷依A閱卷者之評分由高至低做排序，此等級順序應與分別依B及C閱卷者的評分所排列之順序極為一致。

對每一道查詢主題而言，想要瞭解三位判斷者相關判斷的一致性，可將他們判斷的相關文件按相關分數排序，並以統計推論驗證其等級順序是否一致。以編號14之查詢主題為例，表7列出三位判斷者對於各文件所評定之相關分數，以及依相關分數由高至低之排序等級。而若遇有數篇文件之相關分數相同時，則以等級平均法處理，即這些文件之排序等級以其原來等級之平均數表示(註17)。例如對編號0056549_01、0056555與0056557三篇文件而言，相關判斷者A所評定之相關分數均為2，原始等級應是1、2與3，但因分數相同，故取1、2與3之平均數2，做為該三篇文件依相關分數排序之等級。

表7　編號14之查詢主題相關文件組文件相關分數與排序等級

文件編號	相關判斷者 A		相關判斷者 B		相關判斷者 C	
	相關分數	排序等級	相關分數	排序等級	相關分數	排序等級
0056549_01	2	2.0	2	2.0	2	2.0
0056555	2	2.0	2	2.0	2	2.0
0056557	2	2.0	2	2.0	2	2.0
0150056	1	4.5	1	4.0	0	5.0
0150080	1	4.5	0	5.0	1	4.0

在統計量數中，Kendall和諧係數W是用來計算三組(含)以上資料之一致程度，例如K個評分者評N個作品時，此K個評分者所做評分之一致程度，適用於研究評分者間之信度。而Kendall和諧係數W，其值介於0與1之間，W值越大，表示彼此間之一

註17　顏月珠，實用無母數統計方法(台北市：陳昭明，民國75年)，頁261-264。

致性越高；若為0則表示缺乏一致性(註18)。故本研究採用Kendall和諧係數W來計算三位判斷者相關判斷結果之一致性。使用SPSS軟體，將資料輸入，利用其無母數檢定中之Kendall's W檢定，計算結果如表8所示。

表8　編號14之查詢主題相關文件組相關判斷結果

個　　數	3
Kendall's W 檢定（Kendall 和諧係數）	0.957
卡方	11.489
漸近顯著性（P）	0.022

表9　查詢主題相關文件組相關判斷結果一致性檢定

查詢主題編號	相關文件組文件數	W	P
01	44	0.533	0.008*
02	22	0.645	0.006*
03	24	0.593	0.012*
04	7	0.235	0.635
05	125	0.694	0.000*
06	61	0.539	0.002*
07	28	0.610	0.005*
08	9	0.668	0.042*
09	90	0.628	0.000*
10	10	0.907	0.004*
11	13	0.541	0.078
12	25	0.581	0.013*
13	4	0.333	0.392
14	5	0.957	0.022*
15	9	0.541	0.112
16	12	0.333	0.443
17	7	0.719	0.044*
18	66	0.829	0.000*
19	31	0.655	0.001*
20	11	0.333	0.440
21	7	0.692	0.053
22	12	0.493	0.132
23	93	0.582	0.000*
24	5	0.889	0.031*
25	5	0.619	0.115
26	19	0.718	0.003*
27	13	0.496	0.120
28	22	0.668	0.004*
29	28	0.789	0.000*
30	123	0.608	0.000*

（*表示達到0.05顯著水準）

註18　黃國光，SPSS與統計原理剖析(台北市：松崗，民國89年)，頁9-25-9-27；及張勝溢，SPSS/PC進階篇(台北市：碁峰，民國82年)，頁5-21。

由表8可知，Kendall和諧係數W值為0.957，P=0.022<0.05，表示達到統計上0.05之顯著水準，意即三位判斷者相關判斷結果具有顯著之一致性。

利用同樣之統計量數與方法，可分別計算30道查詢主題三位判斷者所做相關判斷之一致性。由表9可知，共有20道查詢主題其相關判斷結果一致性可達到統計上0.05之顯著水準(P<0.05)。進一步觀察相關判斷結果未達到顯著一致性之其他10道查詢主題，其判斷發生歧異之情形，大多出現在完全相關與部分相關此兩種情況（相關分數2與1）。此外，這些查詢主題的相關文件數量最多僅有13筆，在文件數較少的情形下，雖僅有些微文件發生相關判斷不一致的情形，但反映在統計計算與檢定上，就顯得有較大之不一致。綜合來說，此三位判斷者對於各查詢主題相關文件之判斷結果，應具有相當高的一致性。

四、應用測試集之檢索實驗

在完成「中文OCR文件檢索測試集」之製作後，實際運用此測試集來進行檢索實驗(註19)。

(一)實驗設計

本檢索實驗之目的，在於初探相較於乾淨文件，OCR文件帶有雜訊情形下之檢索成效。以「中文OCR文件檢索測試集」之OCR文件單元，以及部分乾淨文件單元，配合查詢主題、相關判斷結果，構成實驗環境。我們設計出數種檢索策略，在不同的文件環境下，採向量檢索模式為基礎之Crystal檢索系統來進行檢索實驗。實驗時所使用之查詢問句，分別以測試集中內30道查詢主題之<title>欄位內容，形成長度較短之30道查詢問句；以及將<title>、<desc>及<narr>三個欄位內容加以結合，形成長度較長之30道查詢問句。故各檢索策略將產生30個檢索結果。

(二)評估方式

對於檢索實驗結果，利用TREC之工具軟體－trec_eval，以「平均精確率」此指標來評估檢索成效。由於trec_eval係採用二元化之相關關係(相關、不相關)來計算指標之測量值，故須將相關判斷結果加以轉化。在此做法為：只要文件之相關分數不為0，該文件就視為與查詢主題相關。

此外，參照TREC分析檢索結果之做法，將各檢索策略所有30個檢索結果之平均精確率，再加以平均，就成為該檢索策略檢索結果之平均精確率。

(三)實驗結果

本實驗所採之檢索策略及其檢索成效，如表10所示。

綜合OCR文件，以及部分乾淨文件之檢索結果可以發現，在平均精確率評估下成效最佳之檢索策略(分別為編號11及12)，均是使用較長之查詢問句，以及1-gram結

註19 在註15的論文中，我們也用這個測試集來實驗，但那時有一篇文件損毀，所以實際是以8438篇文件進行各種實驗。

合2-gram之索引詞模式。而使用此索引詞模式之檢索策略，無論對於較長或較短之查詢問句，均較僅使用單一n-gram(1-gram或2-gram)索引詞模式之檢索策略，表現出較佳之成效。

表10 檢索策略之組成及其檢索成效

檢索策略編號	文件組	查詢問句組成欄位	索引詞模式	平均精確率
01	OCR文件	<title>	1-gram	0.3509
02	部分乾淨文件	<title>	1-gram	0.5118
03	OCR文件	<title>	2-gram	0.4044
04	部分乾淨文件	<title>	2-gram	0.5880
05	OCR文件	<title>	1-gram and 2-gram	0.4164
06	部分乾淨文件	<title>	1-gram and 2-gram	0.5964
07	OCR文件	<title>+<desc>+<narr>	1-gram	0.3963
08	部分乾淨文件	<title>+<desc>+<narr>	1-gram	0.5736
09	OCR文件	<title>+<desc>+<narr>	2-gram	0.4582
10	部分乾淨文件	<title>+<desc>+<narr>	2-gram	0.6390
11	OCR文件	<title>+<desc>+<narr>	1-gram and 2-gram	0.4757
12	部分乾淨文件	<title>+<desc>+<narr>	1-gram and 2-gram	0.6588

此外，使用較長之查詢問句有助於提昇平均精確率，在OCR文件下提昇之幅度(14.2%)則大於部分乾淨文件(10.5%)。在使用較長查詢問句時之最佳成效，部分乾淨文件較OCR文件相對提昇了38.5%。而在使用較短查詢問句時之最佳成效，部分乾淨文件則較OCR文件相對提昇了43.2%。

換個角度來看，在OCR文件之正確性約為70% 之情形下，使用較長查詢問句時之最佳檢索策略，其成效可達到部分乾淨文件之72.2%(0.4757 / 0.6588)，而使用較短查詢問句時之最佳檢索成效則可達到部分乾淨文件之69.8%(0.4164/0.5964)。

值得注意的是，本實驗係以部分乾淨文件單元來模擬全部是乾淨文件之環境。但在部分乾淨文件單元中，相關的文件均是乾淨文件，而不相關的文件都是有雜訊之OCR文件，故以部分乾淨文件單元檢索時所能達到之平均精確率，事實上會比文件全部是乾淨時還要高。因為不相關的文件比較沒有正確的詞彙來影響檢索結果(註20)。也就是說，OCR文件與乾淨文件相較，所能達到之檢索成效，在程度上應可高於本實驗之結果。而根據本實驗之結果可以推測，OCR文件帶有之雜訊確實會影響檢索成效，而用OCR文件檢索所能達到之檢索成效，在程度上略等於OCR文件之正確程度。

五、結論

根據IBM之估計，全世界一年花費約2兆5千萬美元在將儲存於傳統媒體之非數位化文件，以人工鍵入方式轉化為數位化之文件。雖耗費了大量時間與金錢，其所

註20 曾元顯，〈回溯性資料數位化服務之規劃與建置〉，二十一世紀資訊科學與技術國際學術研討會，2001年11月29-30日，頁255-274。

能處理之文件量卻僅佔總數之5%(註21)。在此情形下，利用掃描器將文件掃描成影像檔，再經由OCR系統之處理而成為數位化之OCR文件，就成為一迅速且價廉之選擇。

然而，經由OCR系統所辨識輸出之OCR文件，通常帶有著雜訊，故OCR文件在各種資訊檢索系統所能達到之檢索成效，是不可忽視之議題。有鑑於此，本研究建立了「中文OCR文件檢索測試集」，並利用以進行檢索實驗，發現帶有雜訊之OCR文件的確會對檢索成效發生影響。

由於人力等因素限制，本研究仍未臻完善，但仍希望能藉此帶動各相關研究之進行，而使得中文資訊檢索的領域有更廣闊之發展，並讓研究者有更豐富之資源可供檢索及應用。事實上，縱使在乾淨文件中，仍可能因人工疏忽等問題而存在錯誤的文字，故OCR文件檢索之發展，對於乾淨文件之檢索成效也將有所助益。

誌謝

感謝社文中心關秉寅前主任、狄神父主任，以及康芳菁、李青玲、鄭明賢等人多年的協助，本文才得以完成。

本文由國科會計畫補助，計畫編號：NSC 88-2418-H-001-011-B8908、NSC 88-2418-H-001-011-B9003 、NSC 88-2413-H-030-017-、NSC 89-2413-H-030-006-。

附錄

W. B. Croft, S. M. Harding, K. Taghva, & J. Borsack, "An Evaluation of Information Retrieval Accuracy with Simulated OCR Output," *The 3rd Symposium of Document Analysis and Information Retrieval, 1994*, pp.115-126.

Kazem Taghva, Julie Borsack & Allen Condit, "Evaluation of Model-based Retrieval Effectiveness with OCR Text," *ACM Transactions on Information Systems*, 14 : 1 (1996) :64-93.

K. Taghva, J. Borsack, A. Condit, & S. Erva, "The Effects of Noisy Data on Text Retrieval," *Journal of the American Society for Information Science,* 45 : 1 (1994) :50-58.

Kazem Taghva, Julie Borsack & Allen Condit, "Results of Applying Probabilistic IR to OCR Text," *Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval July 3 - 6, 1994,* Dublin Ireland, pp. 202-211.

K. Taghva, J. Borsack, & A. Condit, "Effects of OCR Errors on Ranking and Feedback Using the Vector Space Model," *Information Processing and Management,* 32 : 3 (1996) : 317-327.

Amit Singhal, Gerard Salton, & Chris Buckley "Length Normalization in Degraded Text Collections," *Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996,* pp. 149-162.

Elke Mittendorf, Peter Schauble & Paraic Sheridan, "Applying Probabilistic Term Weighting to OCR Text in the Case of a Large Alphabetic Library Catalogue," *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval July 9-13, 1995,* Seattle, WA, USA, pp. 328-335.

Daniel Lopresti & Jiangying Zhou, "Retrieval Strategies for Noisy Text," *Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996,* pp. 255-269.

註21 "Document Analysis and Recognition," <http://www.almaden.ibm.com/cs/dare.html>

Claudia Pearce & Charles Nicholas, "TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data," *Journal of the American Society for Information Science,* 47：4 (1996)：263-275.

W. Harding, B. Croft, & C. Weir, "Probabilistic Retrieval of OCR Degraded Text Using N-Grams," in *Research and Advanced Technology for Digital Libraries*, Carol Peters & Costantino Thanos, (ed), 1997, pp. 345-359. http://ciir.cs.umass.edu/info/psfiles/irpubs/ir-115.ps.gz

Andreas Myka & U. Guntzer, "Fuzzy Full-Text Searches in OCR Databases," In Nabil R. Adam et al. (eds), *Digital Libraries - Research and Technology Advances*, LNCS 1082, 1996, pp. 131-145.

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* (Previously known as *Journal of the American Society for Information Science, JASIS*), 52：5 (2001)：378-390.

Kareem Darwish & Douglas W. Oard, "Term Selection for Searching Printed Arabic," to appear in the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '02.

Journal of Educational Media & Library Sciences, 40:3 (March 2003):325-344

Construction and Application of an Chinese OCR Test Collection for Information Retriveal

Mung-Chu Tsai
Graduate
Yuen-Hsien Tseng
Associate Professor
Dept. of Library & Information Science, Fu Jen Catholic University
Taipei, Taiwan, R.O.C.

Abstract

This article describes the process of constructing a Chinese OCR test collection and the application of this collection in an retrieval experiment. We have overcome the difficulty of obtaining past information need for retrospective data and created 30 query topics that simulate real user needs. To obtain real OCR documents instead of simulated ones, we have converted 8439 full-text images into 8439 OCR text files. An evaluation of the OCR documents reveals an average of 70% of recognition accuracy. To obtain the relevant documents for each query, we invited 3 judges to examine each of 8439 images and give relevance score to each document for each topic. According to Kendall's statistical coefficient, highly consistent judgments are obtained in 20 query topics. Finally in our experiment with 12 search strategies, our results show that the retrieval effectiveness of OCR documents decrease to 70% when the recognition accuracy is about 70%.

Keywords : *OCR; Information retrieval; Test collection; Effectiveness evaluation; Chinese document retrieval*