資訊檢索的績效評估
Performance Evaluations for Information Retrieval

陳光華 Kuang-hua Chen
國立台灣大學圖書資訊學系
Department of Library and Information Science,
National Taiwan University
e-mail:khchen@ntu.edu.tw

【摘要 Abstract】

資訊檢索在目前網際網路廣泛運用的環境下，成為眾所矚目的研究課題。多數的研究者將大部分研究經費與精力投注在資訊檢索系統的研究，但有關資訊檢索系統績效評估的相關議題，卻沒有受到國內學者相對的重視。本文主要說明目前國際上正努力推動的資訊檢索評估的研究與活動，且不拘限傳統的文件檢索模式，而以更廣泛的角度看待資訊檢索評估，並期盼國內的學者能夠積極參與資訊檢索系統的評估。

Information retrieval (IR) has been a hot research topic since the Internet was broadly used for various applications. Many researchers devote themselves into the study on the information retrieval systems, but few efforts are made on the performance evaluations for IR systems by Taiwan's researchers. This article describes the related activities and researches of IR evaluation and regards IR as a broad term for other related research topics rather than the traditional "document retrieval" only. Finally, the author suggests IR researchers in Taiwan actively participate the global activities for IR evaluation.

關鍵詞：評估；資訊檢索；CLEF；NTCIR；TREC

Keywords：CLEF；Evaluation；Information Retrieval；NTCIR；TREC

一、緒論

長久以來，人類的資訊需求從未間斷，只是隨著時代的變遷與科技的精進，使取得資訊的方式有所不同。電腦在 1940 年後，開始進入人類文明史，美國海軍相關機構隨即著手研發資訊檢索系統，美國學者 Bush (1945)極為有名的文獻 “As we will think”更是指導著從事資訊檢索的研究者，朝更卓越的前景邁進。當然，早期的資訊檢索系統應該稱為「文件檢索系統」，然而，想要很精確地搜尋相關的文件，到現在仍是很大的挑戰；同時，如何公平地評估資訊檢索系統的績效（performance），使得研究者可以更清楚地瞭解技術的優劣、績效的瓶頸，從而改進資訊檢索系統，並提供研究發展的方向，更是吾人關心的課題。因此，可以清楚地理解評估扮演的角色其實是雙重的：一是評估系統績效，二是引導研究方向。

資訊檢索評估的歷史可上溯至 1950 年左右。早期的評估是在「正規化環境」下進行，使用量化或質化的方法，企圖衡量不同檢索技術、檢索模式、索引語言之相對績效。1966 年 Cleverdon 進行的 Cranfield II 計劃，以文件集（Document Set）、查詢問題（Question）及相關判斷（Relevance Judgment）構成一組測試集（Test Collection），並訂定一套績效測量準則，評估多種索引語言。(Cleverdon, 1967) Cranfield II 研究採用的實驗模式與評估方法，在資訊檢索評估的研究領域具有里程碑的時代意義，直至今日仍佔有舉足輕重的地位。然而，前述的正規化環境，可說是在實驗的環境下進行，與實際檢索環境差距甚遠，因而，使評估的結果與實用性受到許多質疑。

美國國防部高等研究計劃署（Defense Advanced Research Projects Agency，簡稱 DARPA）與美國國家標準暨技術局（National Institute of Standards and Technology，簡稱 NIST），在 1992 年共同舉辦了文件檢索會議（Text REtrieval Conference，簡稱 TREC），TREC 透過大型測試集的建構，伴隨測試項目、測試程序、評估準則的訂定，以及舉辦論壇提供參與者討論與分享結果。(Harman, 1993)它首創了前所未有的大型測試集，使測試環境得以更接近實際情況，對檢索技術發展與系統績效提昇，具有相當重要的貢獻，同時，也成為從事資訊檢索評估相關機構或機制的仿效對象。

目前，國際上除了 TREC 外，還有二個從事資訊檢索評估的合作機制： CLEF（Cross-Language Evaluation Forum）與 NTCIR（NII Test Collection for Information Retrieval）。CLEF 是由歐洲各國的學者專家合作建構的評估機制，主要的負責人是義大利的 Carol Peters；NTCIR 則是由日本國立情報學研究所（National Institute of Informatics）主辦，韓國科學技術資訊研究所（Institute of Science and Technology Information）與國立台灣大學協助籌辦。由於筆者自 2000 年開始即參與 NTCIR 的工作，因而本文除了探討資訊檢索評估對於資訊檢索研究的影響之外，亦將著重於 NTCIR 在資訊檢索評估扮演的角色及相關的研究成果。

事實上，資訊檢索的研究已由早期文件檢索（Document Retrieval）逐步地進入更深入且廣泛的範疇。目前廣義的資訊檢索涵蓋文件檢索，文件過濾（Document Filtering），文件摘要（Document Summarization），主題辨識（Topic Identification），資訊擷取（Information Extraction，簡稱 IE），標題生成（Title Generation），問題答詢（Question Answering，簡稱 QA），以及橫跨這些研究議題的跨語言、跨媒體、與跨文化的檢索技術。因而，資訊檢索的評估也從早期文件檢索的評估，邁向上述各類型資訊服務系統的評估，本文也將說明 NTCIR 在新類型資訊檢索評估的工作與成果。

本文結構說明如後。第二節將簡述資訊檢索評估的歷史與有關學術文獻；第三節說明 NTCIR 的組織與工作；第四節探討目前資訊檢索評估的重要項目，第五節討論常用

的資訊檢索的評估準則；第六節說明 NTCIR 推動資訊檢索評估所獲致的成果；第七節則是簡短的結論。

二、文獻探討

從事資訊檢索研究的學者專家為了證明他們的系統有很好的執行績效，通常會進行一連串的實驗來驗證，而實驗所需的測試資料，則依個別目的而分別蒐羅建立。由於不同研究機構研發的資訊檢索系統，很可能使用不同的測試資料，因此，很難說服其他研究人員，A 系統比 B 系統好，或是 C 技術比 D 技術好。為了避免各說各話，有些個別的測試集應運而生，如 ADI、MEDLARS、TIME、CACM、CISI、NPL、INSPEC、ISILT、UKCIS、UKAEA、LISA 等(Salton, 1972; Sparck Jones & van Rijsbergen, 1976; Fox, 1983; Harman,1996)，這些公開的測試集提高了資訊檢索評估的公平性，至少，不同的系統使用相同的測試集，使彼此有了對話與討論的基礎。

但是，上述測試集的規模都不大，且測試集的文件同質性甚高。例如，Cranfield II 的測試集有 1400 篇文件以及 200 多個問題，文件主題為太空動力學相關學術文獻，請參見表 1 有關前述測試集的統計數據。這些和實際檢索環境相去甚遠的測試集，讓學者專家對於使用它們進行評估而獲致的結果產生疑慮。

為了回應各方的批評與責難，NIST 與 DARPA 展開合作，於 1992 年開啟了 TREC 計畫。而 TREC 也形成一種評估機制與論壇，且成為後續類似評估機制的開創者。基本上，TREC 是個一次持續一整年的資訊檢索評估活動，有籌備工作、公布評估項目、參與者送回檢索結果、工作人員進行評估、送回評估結果、舉辦年度論壇會議等程序。年度論壇會議包含討論各參與團隊之資訊檢索系統的優缺點，探討新的語言處理技術與可用的語言資源，新的資訊檢索技術與跨語言的模式，發展新的評估項目並聽取參與者的意見等過程。然後，依循相同模式，隨即進行下一年度的活動。在這樣的評估機制下，測試集的建置是極為重要的一環，不同的評估項目需要不同的測試集，且隨資訊檢索研究的深化與廣化，測試集的形式也日趨多元化，以適應不同的資訊檢索研究。表 2 則詳列近期較為著名的測試集的統計數據，相較於表 1，可以明顯看出，文件的數量增加非常多，而查詢問題的字數也變多。

查詢問題字數變多的原因與查詢內容的變革有密切關係。早期測試集的查詢問題內容，雖然是以自然語言的方式陳述，但卻十分簡短，包含的訊息相當少；TREC 首創多欄位的「查詢主題」(Topic rather than Query)，以呈現不同層次的資訊需求，後來發展的測試集也紛紛仿效。另外，查詢問題所涉及的面向也愈來愈多元化，除了主題相關之外，也逐漸加入檢索目的、檢索背景等其他層面的描述。

至於測試集在整個評估活動所扮演的角色可以圖一為例，這是一個典型的文件評估項目的示意圖，可以看出測試集，檢索系統，評估人員與評估準則之間的互動關係。其中值得注意的是測試集中「相關判斷」的建構方式。早期測試集的文件數量不多，因此評估者（Assessor）可以逐篇閱讀所有文件，然後判斷其與查詢問題的相關性；惟近期測試集的文件數量皆相當龐大，評估者實無法逐一閱讀每篇文件，因此，TREC 發展了所謂的“Pooling Method”。Pooling Method 的假設是：「與查詢問題（或查詢主題）相關的文件應該會被多數的資訊檢索系統檢索出來」，利用此方法的主要精神是希望透過眾多不同的資訊檢索系統與不同的檢索技術，盡量網羅所有可能的相關文件，減少人工判斷的負擔。匯集眾多資訊檢索系統的檢索結果構成一個「相關文件候選集」(Pool)，而真正相關的文件，應該包含於這個 Pool。因此，運用 Pooling Method，可以為每一個查

詢主題建構一個 Pool，通常一個 Pool 會有 1000 篇上下的文件，評估者僅需閱讀 Pool 裡的文件，然後判斷 Pool 裡每一篇文件與查詢主題的相關性。

TREC 評估機制開創了大量的文件集，結構化的查詢主題，務實的相關判斷，使得資訊檢索的評估技術突飛猛進，也得到從事資訊檢索研究者的認同，紛紛參與 TREC 的各項評估項目活動，並以評估的結果撰寫學術論文，宣示自己的研究成果。同時，隨著 TREC 擬定的評估項目，如跨語言檢索、問題答詢、資訊過濾，也促使資訊檢索研究者投入該評比項目的研究，使得評估引導研究方向，成為目前資訊檢索研究的趨勢。最顯著的例子便是阿拉伯語的文件檢索與跨語言資訊檢索研究，而這些研究成果也陸續發表在 SIGIR 以及相關資訊檢索期刊與會議。

TREC 的成功，也讓後續著重於其他語言的資訊檢索評估機制（CLEF 與 NTCIR）有了依循的基礎，而這些評估機制的攜手合作，更促進資訊檢索評估受到資訊檢索研究者的普遍重視，並擴大引導資訊檢索研究的方向。

表 1：早期測試集

測試集	文件數	文件 平均字數	查詢 問題數	查詢問題 平均字數	主題領域	語文
Cranfield II	1,400	53.1	225	9.2	太空動力學	英文
ADI	82	27.1	35	14.6	文獻學	英文
MEDLARS	1,033	51.6	30	10.1	醫學	英文
TIME	423	570	24	16.0	世界情勢	英文
CACM	3,204	24.5	64	10.8	ACM 通訊	英文
CISI	1,460	46.5	112	28.3	資訊科學	英文
NPL	11,429	20.0	100	7.2	電子、電腦、物理、地理	英文
INSPEC	12,684	32.5	84	15.6	物理、電子、控制	英文
ISILT	800	N/A	63	N/A	文獻學	英文
UKCIS	27,361	182	193	N/A	生化	英文
UKAEA	12,765	N/A	60	N/A	核子科學	英文
LISA	6,004	N/A	35	N/A	N/A	英文
Cystic Fibrosis	1,239	49.7	100	6.8	醫學	英文

表 2：近期測試集

測試集	文件數	文件 平均字數	查詢 問題數	查詢問題 平均字數	主題領域	語文
OSHUMED	348,566	250	101	10	N/A	英文
TREC (TREC-1~6)	1,754,896	481.6	350	105.8	多主題	英文
AMARYLLIS	336,000	N/A	56	N/A	多主題	法文
NTCIR1	300,000	N/A	100	N/A	多主題	日文
CIRB010	132,173	N/A	50	N/A	新聞（多主題）	中文
CIRB030 (NTCIR4-Chinese)	381,681	N/A	110	N/A	新聞（多主題）	中文
NTCIR4-Japanese	596,058	N/A	60	N/A	新聞（多主題）	日文
NTCIR4-Korean	254,438	N/A	60	N/A	新聞（多主題）	韓文

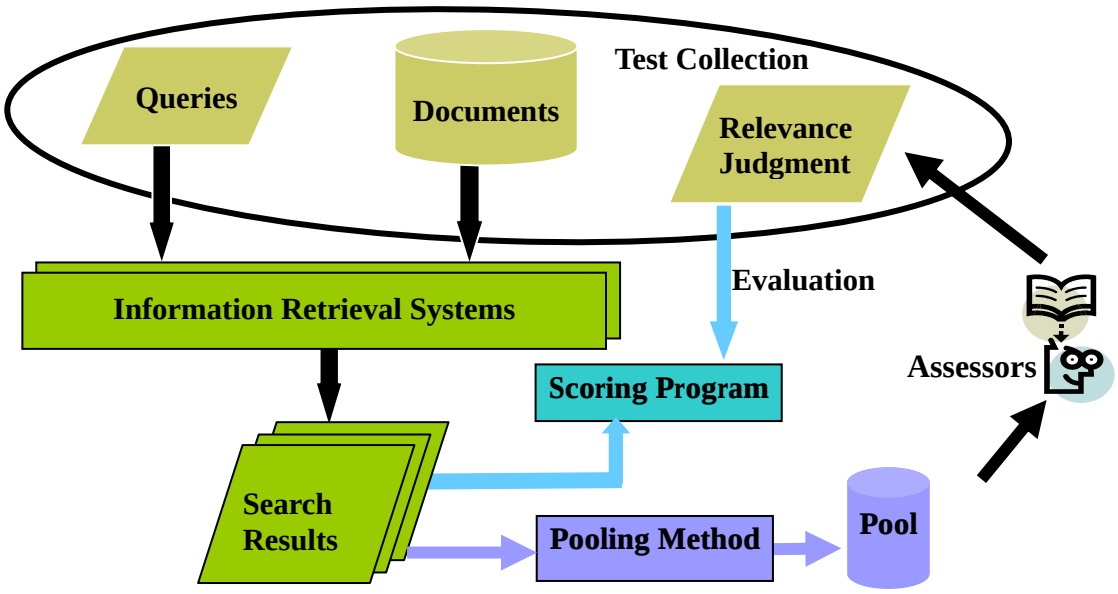


圖 1：文件檢索評估項目

三、NTCIR 評估機制

NTCIR 由日本國立情報學研究所（National Institute of Informatics，簡稱 NII）主辦，韓國科學技術資訊研究所（Institute of Science and Technology Information）與國立台灣大學則是協助籌辦。要瞭解 NTCIR 這個評估機制，必須先說明 NII 這個近年來在學術界享有盛譽的研究機構。

NII 的前身是學術情報中心（National Center for Science Information Systems，簡稱 NACICS），2000 年 4 月更名為國立情報學研究所。長期以來，日本的文部省便認為進行資訊學的研究是非常必要的課題，因此亟思構想成立一個資訊學研究機構。然而，成立一個全新的資訊學研究機構，需要極為龐大的經費支應，於是轉而規劃以學術情報中心為基礎，再加以擴大，便成為現在的「國立情報學研究所」。NACICS 時期，主要的任務是提供資料庫、書目系統和網際網路服務，以及進行與這些服務有關的研究。現在的 NII 除了承續 NACICS 原有的服務與研究外，還增加了更廣泛與新穎的研究內容。從機構的更名開始，就慢慢增加了不同領域的新進研究者，因此，雖然研究範圍變廣，但與以前比較起來，目前在目錄、資料庫、網路方面的研究卻更為活躍，同時，基礎研究與軟硬體設備也更加強。

NII 的資訊學研究（Informatics），分成七大領域：資訊學基礎研究、資訊技術基礎研究、軟體研究、資訊媒體研究、智慧系統研究、人類社會資訊研究、學術資訊研究。以往 NACICS 的目的在於提供研究者所需的學術資訊，以及如何讓資訊的流通更為順暢，因此是以圖書資訊學、軟體研發、網路為主要工作項目。但是，NII 並不特別著重某一類的研究，不論資訊科學或人文社會科學，各方面的資訊都包含在其研究範圍之內，是朝科際性方向發展。簡言之，NACICS 以提供服務為主，所以，為了提供服務而做研究；NII 則是以做研究為主，提供的服務則有如對研究成果的一種測試。

目前 NII 有六十多位研究員，及數位客座研究員、博士後研究員。其中，具有圖書資訊學背景的只有 3 人。但有些研究者原本是資訊工程背景，加入 NACSIS 之後，其主要研究反而變成是圖書資訊學領域的主題。另外，其他許多不同背景的人，也都是在加入 NII 之後，開始接觸參與圖書資訊學相關的研究，如資訊服務、數位圖書館等議題。

由此可以發現圖書資訊學與資訊工程學研究人才的交流，在 NII 的學術環境下，是極其自然，且互蒙其利。

資訊檢索作為資訊服務的核心，當然是 NII 極為注重的課題，也有許多的研究員從事資訊檢索的研究，當 TREC 在英文資訊檢索領域獲得空前的成就之後，NII 也想在日文資訊檢索方面扮演同樣的角色，所以積極籌劃第一屆 NTCIR，並於 1999 年順利完成。第一屆 NTCIR 純然是日本國內的活動。但 NII 的研究員 Noriko Kando（神門典子）博士，想要將 NTCIR 提升為國際性的資訊檢索評估活動，以結合亞洲地區（至少是東亞地區）的研究者的力量，籌劃跨語言的資訊檢索評估活動，期能進一步影響此地區的資訊檢索研究，從而扮演更積極的角色。筆者便在這樣的情形之下，由第二屆開始，加入 NTCIR 評估機制，提供中文單語資訊檢索評估與英、中跨語資訊檢索評估項目。韓國則於第三屆 NTCIR 開始加入。自此，NTCIR 開始走向國際化，緊密結合韓國與台灣的研究人員，NTCIR 所需的經費則多數由 NII 提供的。

NTCIR 做為亞洲地區最重要的資訊檢索評估機制，提供各種單語、雙語、多語、以及跨語資訊檢索評估項目，受到國際上資訊檢索研究的學者的重視，參與的國家地區，以及研究團隊的數目持續成長，NTCIR 的評估結果也成為各研究者發表學術論文的重要依據。

NTCIR 也與 TREC 和 CLEF 密切合作，例如，TREC 協助 NTCIR 製作英文檢索問題，NTCIR 協助 TREC 製作或是翻譯中文檢索問題，NTCIR 也使用 CLEF 製作的檢索問題。TREC 甚至放棄舉辦中文單語與雙語資訊檢索的評估項目，並鼓勵有關研究的學者專家轉而參加 NTCIR 舉辦的中文資訊檢索評估項目；CLEF 也舉辦中文檢索歐洲語言文件的評估項目，而 NTCIR 則協助提供中文問題。這樣的合作使得資訊檢索評估的影響力擴大，更加促使資訊檢索評估機制引導資訊檢索的研究方向。

NTCIR 是一項長期性的、跨國的合作研究計畫，對於資訊檢索的研究、資訊檢索評估的研究、專利檢索的研究、自動摘要的研究、以及問題答詢的研究具有重要的影響力。在第二屆 NTCIR 資訊檢索評估會議之後(Kando, 2001)，來自日本、韓國、和台灣的研究者便開始討論更複雜，且更接近實際檢索環境的跨語言資訊檢索(Cross-Language Information Retrieval，簡稱 CLIR）的評估項目，並為了第三屆 NTCIR 資訊檢索評估會議籌備工作，組織一個 CLIR 執行委員會負責各項工作的推動，該執行委員會由 9 個分別來自於日本、韓國和台灣的研究者組成，成員每年在日本聚會，討論 CLIR 任務的詳細內容、制訂計畫和擘畫整個研究藍圖。這種模式也將繼續沿用於後續的 NTCIR 的各個評估項目的進行。

至於，CLEF 則是近年來歐洲國家加速統合工作的一項成果。在歐盟的推動下，歐洲各國進行了許多的合作研究計畫，如電子化政府與跨語言文化的研究。其中跨語言的研究更是其中的重點，源因於歐洲各國有各式的語言，如何使得語言不再成為資訊交流與促進發展的障礙，對歐盟各國而言，皆是非常重要的課題。在此環境下，CLEF 便仿效 TREC 的運作模式，投入跨語言資訊檢索的評估工作，和 TREC 與 NTCIR 二個評估機制相同，CLEF 有許多評估項目的執行委員會（由主席與委員組成），由一位總攬大局，類似秘書長或執行長的人物；另有一指導委員會（Steering Committee）由來自各國的學者專家組成，提供 CLEF 發展方向的意見與看法。

四、評估項目

NTCIR，CLEF，以及 TREC 選擇的評估項目，通常都是資訊檢索研究領域中，重要且具有發展潛力的研究課題。下文將說明 TREC、NTCIR、與 CLEF 舉辦的評估項目，並以 CLEF 為例，詳細說明各評估項目。

（一）TREC

TREC 自 1992 年開始，已經舉辦非常多的評估項目，有最傳統的文件檢索，也就是俗稱的 Ad Hoc Task。所謂的 Ad Hoc 就是一般使用者認知的資訊檢索，有一個固定的文件集（如光碟資料庫），不同的使用者有不同的檢索問題，因此問題是變動的，而文件是固定的。相對的 Routing Task，則是文件是變動的，而檢索問題是固定的，因此 Routing Task 就是所謂的資訊過濾，這時的檢索問題通常是以使用者的 Profile 形式展現，其內容是使用者固定的資訊需求。其餘的檢索評估項目，羅列如下。

1. 資料庫合併檢索（Database Merging）：評估資訊系統檢索不同資料庫，以及合併檢索結果的能力。
2. 高查準率檢索（High Precision）：著重於高「查準率」的資訊檢索。
3. 互動檢索（Interactive）：企圖模擬使用者與資訊檢索系統互動的情形。
4. 多語言檢索（Multilingual）：反映網際網路上資訊檢索的實際情形。
5. 自然語言檢索（Natural Language Processing）：由於自然語言是使用者最熟悉的查詢語言，因此本項目是用以評估資訊檢索系統處理自然語言的能力。
6. 問題答詢（Question Answering）：將檢索結果的單位縮小為句子、片語、詞彙，而非文件，以更接近使用者的資訊需求。
7. 巨量文件（Very Large Corpora）：評估檢索系統處理巨量文件的能力，這在網際網路的環境下，是非常重要的課題。
8. 基因資訊檢索（Genomics）：關注於生物資訊的檢索，尤其是近年基因研究蓬勃發展，使基因資訊的檢索成為重要研究課題。
9. 強健型資訊檢索（Robust Retrieval）：TREC 特別設計本評估項目，用以評估資訊檢索系統對於困難查詢問題的處理能力。
10. 網頁檢索（Web）：真正使用網際網路的網頁為標的所進行之文件檢索工作。

（二）NTCIR

NTCIR 自 1999 年開始籌辦，目前進入第 5 屆，舉辦的評估項目有傳統的日文、中文、韓文、英文的單語 Ad Hoc 項目；對於 NTCIR 而言，最重要的項目應該是跨語言資訊檢索，若以 C、J、K、E 分別代表中文、日文、韓文、英文，則有 C→CJKE、J→CJKE、K→CJKE、E→CJKE 等極為複雜的檢索項目，提供給資訊檢索研究者非常具有挑戰性的研究課題；另外尚有 Pivot Language Information Retrieval，這個項目是模擬在語言資源不足的情形下，進行的跨語資訊檢索，例如，要進行 C→K 跨語資訊檢索，但是沒有中韓雙語詞典，只好借用中英詞典以及英韓詞典，此時，英文就視為是 Pivot Language。其他評估項目羅列如下。

1. 問題答詢挑戰（Question Answering Challenge, QAC）：與 TREC 的 QA 類似，是單語 QA，不同的是 QAC 較困難，它提供一種特別的 QA 評估項目，Context QA 是由一系列的問題構成，每個問題都是環環相扣，資訊檢索系統必須具有照應詞解析（Anaphora Resolution）的能力，才能夠有效處理這類的問題。

2. 網頁檢索（Web）：與 TREC 的 Web 檢索項目類似。
3. 自動摘要（Summarization）：為文件進行摘要，僅 NTCIR 提供。在美國另有一 SUMMAC（SUMMarization Conference）評估機制提供文件自動摘要的評估，因此 TREC 並沒有這個評估項目。
4. 專利檢索（Patent Retrieval）：NTCIR 與日本智慧財產局合作的專利檢索評估項目，其目的是為了提昇專利檢索的品質與績效。

（三）CLEF

CLEF 自 2000 年開始籌辦，是歐洲各國共同合作進行的一項長期性的研究計畫，企圖透過評估資訊科技，來確認何種技術可以有效的降低前述的語言障礙，讓使用者可以應用自己的母語取得其他語言的資訊或文件，然後再運用（半）自動翻譯的技術，閱讀相關資訊或文件。以第五屆（2004 年）CLEF 為例，其舉辦的評估項目如下所示。

1. ImageCLEF

   影像檢索對使用者而言，是能最直接了當獲得答案的方式，因為影像是眾人皆可理解，且本質上不存在語言的阻礙（其實，不全然沒有）。然而，針對影像內容為本的檢索（content-based）仍有實質上的困難，因此，目前的影像檢索多數還是使用文字檢索技術，來搜尋影像的標題（Caption），以及其他詮釋資料（Metadata）。所以，在使用者看不到的技術層面部分，仍有跨語言的問題有待處理，這也是 CLEF 提出 ImageCLEF 項目的主要原因。今年比較特別的方法，是有研究團隊整合了文字檢索技術以及影像檢索技術，發現要比單獨使用文字檢索技術的效果更好。同時，本次影像檢索使用的是醫學的 X 光片影像以及圖書館所典藏的數位圖片。前者，對於後續醫學領域的應用有很重要的影響，深具意義。

2. Ad Hoc Information Retrieval

   跨語言檢索技術通常使用雙語辭典、類比語料庫和機器翻譯系統，以翻譯使用者下達的查詢問句（Query）。然而，並非所有的語言配對都能有雙語辭典、類比語料庫或機器翻譯系統，因此 CLEF 也有 Pivot Language Information Retrieval，亦即使用英語作為 Pivot Language 的模式。這種方式在歐洲這樣多語言的環境下，是非常務實的作法，然而其間涉及兩次的 Disambiguation 動作，是必須慎重處理的。今年的評估結果顯示，Pivot Language Information Retrieval 的檢索效能還算不錯，可以繼續發展。

3. QA@CLEF

   就資訊檢索的服務而言，QA 才是使用者真正想要享用的服務。然而一個實際可行的 QA 系統卻尚未見成熟，因此，TREC、CLEF、NTCIR 等評估機制仍然持續地進行這方面的探討與研究。CLEF 比較特殊的是很早就提出 CLQA（全稱？），相對的 TREC 仍然是單語 QA，而 NTCIR 將於 2004 年 9 月啟動的 NTCIR5 提出 CLQA 項目。今年 QA 的成果尚稱不錯，原因是將問題限制為較單純的問題，如「日本電影羅生門的導演是誰？」、「2000 年當選美國總統的候選人是誰？」等，轉而將挑戰放在跨語的技術上。

4. CL-SDR（Cross-Language Spoken Document Retrieval）

   CL-SDR 主要是針對有背景雜訊的語音文件，進行檢索效能的評估。使用的語音文件是由 TREC 準備的 557 小時美國英文新聞，包括 ABC，CNN，PRI，以及 VOA 等新聞機構。下表為今年參與評估的團隊以及評估的結果。Primary 指的是英語單

語檢索，且無類比語料庫；或是法德至英語的跨語檢索，亦無類比語料庫。Secondary 指的是使用類比語料庫。而參與團隊僅有二隊。比較單語檢索與跨語檢索，語音檢索的效能降低甚多，顯現還有很大的發展空間；此外，由參與團隊數量觀之，這個項目目前仍需學術界與產業界更多的投入，才可能快速提升語音檢索的效能。

5. GIRT（German Indexing and Retrieval Test database）

   GIRT 評估項目的理論基礎是探討使用專門領域上下文（domain-specific context）的訊息，以進行文件檢索的效能。今年使用 GIRT-4 German/English 社會科學資料庫，該資料庫有多種語言控制詞彙（德國／英語，德國／俄語）可供利用，因此在本質上與傳統資訊檢索領域使用的自由詞彙有很大的不同，應用的檢索技術亦有不同，著重於索引典與標題表的應用。

6. iCLEF（Interactive Cross-Language Information Retrieval）

   iCLEF 嘗試模擬實際檢索環境下，使用者與檢索系統的互動情形，以增進資訊檢索的效能。實際觀察一個使用者檢索文件的行為，很少是一次檢索便能滿足其資訊需求，通常的情形是一個巡迴的檢索（One Session），亦即經過查詢問句的不斷修正，而逐步完成檢索工作。iCLEF 將查詢問句設計為一系列的問題，參與團隊必須依照既有的順序進行檢索，以模擬使用者與檢索系統互動的情境。

五、評估準則

任何的評估機制都必須建構公平合理的評估程序，並採用適切的評估準則以及績效評分（Performance Scoring）。就傳統的文件檢索而言，最常用的評分方式就是查全率（Recall）與查準率（Precision），以及結合二者的 F1-measure，計算式如下所示。

$$\mathrm{P} = \mathrm{Precision} = \frac{\#\,\text{relevant retrieved}}{\#\,\text{retrieved}}$$

$$\mathrm{R} = \mathrm{Recall} = \frac{\#\,\text{relevant retrieved}}{\#\,\text{relevant}}$$

$$\mathrm{F1} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}}$$

但是這種計算方式僅適用於無排序的搜尋結果（Non-Ranked Search Results），顯然並不適用於目前依「相關程度」排序的搜尋結果。TREC 使用 Buckley (1991)發展的 trec_eval 評分程式用以評估排序的搜尋結果，而 trec_eval 也成為標準的評估程式，圖 2 是 trec_eval 的評分結果。

```
Queryid (Num):          1
Total number of documents over all queries
    Retrieved:        740
    Relevant:          42
    Rel_ret:           40
Interpolated Recall - Precision Averages:
    at 0.00         1.0000
    at 0.10         1.0000
    at 0.20         1.0000
    at 0.30         0.9333
    at 0.40         0.8095
    at 0.50         0.7857
    at 0.60         0.7568
    at 0.70         0.6977
    at 0.80         0.4789
    at 0.90         0.3585
    at 1.00         0.0000
Average precision (non-interpolated) for all rel docs(averaged over queries)
                    0.7104
Precision:
  At    5 docs:     1.0000
  At   10 docs:     0.9000
  At   15 docs:     0.9333
  At   20 docs:     0.8000
  At   30 docs:     0.7667
  At  100 docs:     0.3700
  At  200 docs:     0.1900
  At  500 docs:     0.0780
  At 1000 docs:     0.0400
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:          0.6905
```

圖 2：trec_eval 評分程式的執行結果

圖 2 可分成幾個部分，“Queryid (Num)”為檢索問題的識別編號；“Total number of documents over all queries”的 Retrieved 說明該資訊檢索系統針對此檢索問題共檢索出 740 篇文件，Relevant 是此檢索問題真正相關的文件數有 42 篇，Rel_ret 說明該檢索系統檢索出 40 篇相關文件。因此，如果採用前述無排序的查全率與查準率的算法，查全率為 40/42=0.9524；查準率為 40/740=0.0541。以下分別說明“Interpolated Recall-Precision Averages”、 “Average precision”、 “Precision: At X docs”、以及“R-Precision”等四項針對有排序的搜尋結果的評分方式。

- “Interpolated Recall-Precision Averages”是所謂的 11-point Precision，以內差法的方式估計在固定的 Recall 下相對的 Precision，其計算基礎是以“Precision: At X docs”等數據計算而得（請參見下文），請參考表 3，圖 3 則分別以 Raw Data 與 Interpolated Data 繪製 P-R 曲線。
- “Average precision”是以下列方式計算而得，其意涵是平均每篇相關文件被檢索出時的 Precision。

$$\text{Average Precision}_j = \text{AP}(j) = \frac{\sum_{i=1}^{r_j} \frac{i}{\#Doc_j(i)}}{r_j}$$

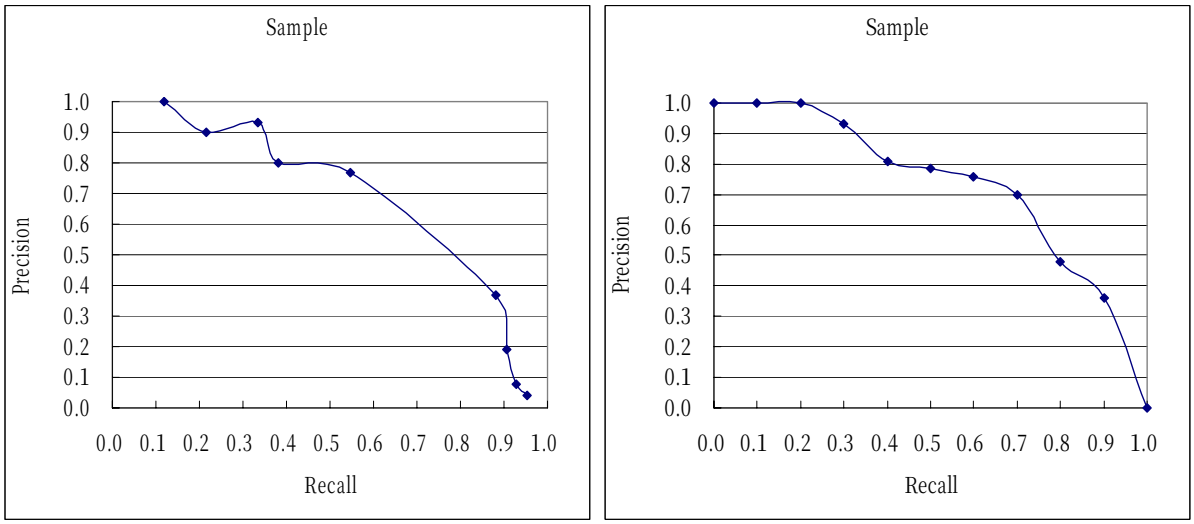
r_j 表示資訊檢索系統針對編號j的查詢問題，共檢索出的相關文件數。

$\#Doc_j(i)$ 表示資訊檢索系統針對編號j的查詢問題，在第i篇相關文件被檢索出時，總共被檢索出的文件數。

- “Precision: At X docs”表示在檢索出 X 篇文件時的 Precision。例如 “At 15 Docs: 0.9333”表示前 15 篇文件有 14 篇相關，所以 Precision 是 14/15=0.9333。
- “R-Precision”則是表示在檢索出第 R 篇文件時的 Precision，R 是查詢問題真正相關的文件數。以圖 2 的情形為例，R 是 42，而檢索出第 42 篇文件時的 Precision 為 0.6905，換句話說，前 42 篇文件中有 42*0.6905=29 篇相關文件。

表 3：11-Point Precision 的計算

Doc Retrieved	Precision	Relevant Retrieved	Raw Data		Interpolated Data	
			Recall	Precision	Recall	Precision
5	1.0000	5	0.119048	1.0000	0.0	1.0000
10	0.9000	9	0.214286	0.9000	0.1	1.0000
15	0.9333	14	0.333333	0.9333	0.2	1.0000
20	0.8000	16	0.380952	0.8000	0.3	0.9333
30	0.7667	23	0.547619	0.7667	0.4	0.8095
100	0.3700	37	0.880952	0.3700	0.5	0.7857
200	0.1900	38	0.904762	0.1900	0.6	0.7568
500	0.0780	39	0.928571	0.0780	0.7	0.6977
1000	0.0400	40	0.952381	0.0400	0.8	0.4789
					0.9	0.3585
					1.0	0.0000



Raw Data　　Interpolated Data

圖 3：原始數據與內差數據

一般而言，問題答詢會依評估項目所要求的搜尋結果的不同而採用不同的評分方式，目前的問題答詢評估項目有

- 資訊系統必須送回5個排序的答案（5 Ranked Answers），第一個正確答案的序號倒數（Reciprocal Rank）即為該問題的分數，整體評分為所有問題得分的平均數（Mean Reciprocal Rank，簡稱MRR）。
- 資訊系統僅送回一個答案，評分方式為Precision。

自動摘要的評分方式比較困難，與評估項目的形式有密切關係，且評估者主觀因素較強。例如 SUMMAC 的 Categorization Task，評估人員必須閱讀摘要，據以設定該摘要屬於哪一個 Topic，如果無法決定則設定為「其他」，然後計算評估者正確分類的比例（也就是 Precision）。NTCIR 舉辦的自動摘要評估項目，則要求評估者閱讀摘要後，依"Rank"與"Revision"二項準則評分。Rank 的評分基礎是「內容完整性」與「內容可讀性」；Revision 則是依據需要花費多少的「刪除、插入、取代」等編輯作業才能將自動摘要改成具有可讀性、完整性的摘要。(Fukusima & Okumura, 2002)

其他的評分方式還有 Discounted Cumulative Gain、Weighted Mean Reciprocal Rank 等，請參閱參考文獻。(van Rijsbergen, 1975; Järvelin & Kekäläinen, 2000; Eguchi et al., 2002; Buckley & Voorhees, 2004)

六、NTCIR 的成果

NTCIR 已經舉辦了四屆，第五屆的 NTCIR 正在進行中，這幾年的評估成果已經得到許多有關資訊檢索研究的經驗，下文將擇要說明這些成果。

要說明 NTCIR 的成果，第一個切入點就是參與 NTCIR 的團隊與國家的成長，表 4 詳列歷年來參與 NTCIR 各評估項目的團隊數量，以及國家地區的分佈。由表 4 可以發現由第一屆開始，參與 NTCIR 的團隊與國家數量呈現穩定的成長，其中尤以隸屬於大學的研究機構更是積極地加入 NTCIR 的評估活動，其數量由第二屆的 20 個大學研究機構，到第三屆成長為 44 個。

表 4：NTCIR 歷年參與團隊數據分佈

	NTCIR 1	NTCIR 2	NTCIR 3	NTCIR 4
Period	1998-11/1999-09	2000-06/2001-03	2001-08/2002-10	2003-04/2004-06
#Groups	28	36	64	76
#Countries	6	8	9	10

跨語言資訊檢索可說是 NTCIR 最關注的項目，因此，我們最想知道的一件事是單語檢索績效與跨語檢索績效的比較，以及中文、日文、韓文、英文等語言的檢索績效的差異，表 5 詳列了這些統計數字。(Chen et al., 2002) 對中文而言，中文單語資訊檢索的績效為 0.2086，英中跨語資訊檢索的績效為 0.1443，下降了 31%；日中跨語資訊檢索的績效為 0.1010，下降了 52%；韓中跨語資訊檢索的績效為 0.0258，下降了 88%。綜合上述數據，英中跨語資訊檢索是做的最好的跨語資訊檢索，而日中與韓中跨語資訊檢索都還有很長的路要走。

就單語資訊檢索而言，英文最好，韓文次之、日文再次之，而中文最差，可能的原因是中文比較困難；有可能是中文的資訊檢索仍未出現很好的處理模式，有待更進一步的研究；當然也有可能是中文的評估人員比較嚴謹，使得相關文件較少，從而降低資訊檢索的績效。

表 5：Mean Average Precision (MAP)

Topic \ Doc	C (#run/#grp)	E (#run/#grp)	J (#run/#grp)	K (#run/#grp)
C	0.2086 (34/15)	0.1055 (3/1)	0.0891 (4/2)	N/A
E	0.1443 (16/6)	0.3476 (29/15)	0.1602 (11/5)	0.1134 (6/2)
J	0.1010 (5/3)	0.2149 (1/1)	0.3029 (30/13)	N/A
K	0.0258 (2/1)	0.3148 (2/1)	N/A	0.3084 (17/8)

對於資訊檢索的研究者經常使用的檢索技術、檢索模式、與索引單位有關的作法的言，就 NTCIR 歷年的評估結果分析，有如下的初步結論。

- 相較於過去，許多團隊使用機率模式（probabilistic model），而且有良好的檢索績效。
- 倒置檔（inverted file）是最常使用的索引檔案結構。
- tf/idf-based 索引詞彙權重計算方式，仍然是最多團隊使用的方法。
- 查詢詞彙的擴張（Query expansion）有很好的檢索績效。
- 停用詞表可以有效提升系統績效，是非常好的語言資源。
- 中文單語檢索時，索引單位以字（Character）或 N-gram 較佳；跨語檢索時，索引單位以詞較佳。
- 在參與 BLIR（Bilingual CLIR）和 MLIR（Multi-lingual CLIR）評估項目的 15 個團隊，有 7 組採用以詞典為本（Dict-based）的翻譯方式；有 3 組採用機器翻譯為本（MT-based）的翻譯方式；2 組採用以語料庫為本（Corpus-based）的翻譯方式；2 組採用 MT+Dict-based 翻譯方式以及 1 組 MT+Corpus-based 翻譯方式。
- 在 E→C 評估項目採用以詞典為主的翻譯方式，有良好的績效。
- 若僅採用詞典為主的翻譯方式，使用所有可能的翻譯詞（select-all）的效果似乎勝於使用前 X 個翻譯詞（select-X）方法。
- 在 E→J 評估項目採用機器翻譯的方式，有良好的績效。

七、結論

資訊檢索研究對於目前我們所處的資訊環境有很重要的貢獻，有關的研究不僅是學術界的熱門研究議題，也是產業界亟需開發及掌握的關鍵技術，而如何公平地進行資訊檢索績效的評估，便成為大家關注的課題。相較於國外的研究團隊積極地參與各項評估計畫，連亞洲地區的日本、韓國都有為數眾多團隊參與，國內的研究團隊卻顯得比較保守，至今曾經參與的機構僅有 3 個。

就個別的資訊檢索研究人員而言，NTCIR、CLEF 與 TREC 對於從事資訊檢索研究的學者與專家，其實有非常大的助益，從前無法或是沒有能力進行大規模的績效評估的系統，可以藉由前述三個機制代為評估，只要參與適當的評估項目即可。除了可以節省大筆評估系統所需要的經費，同時可以獲得具有公信力的評估結果，在撰寫學術論文時，更可以引用這些評估結果，以證明資訊檢索系統的優越性與技術的可行性。

就廣大的資訊檢索研究領域而言，透過大規模資訊檢索的績效評估，我們可以知道

何種檢索模式比較好，何種索引單位比較恰當，何種跨語檢索技術比較實用，Query Expansion 應如何進行。這些珍貴的評估結果，可以避免資訊檢索的研究人員走入錯誤的道路，或是重覆前人的錯誤。此外，透過評估論壇的進行，參與團隊可以交換研究經驗與意見，甚至發展新的研究議題或方向，對於整個資訊檢索領域的研究，都是極為正面的影響。

從筆者從事資訊檢索評估數年的經驗而言，國內參與人數與團隊過少，實在是非常可惜的一件事。事實上，國外的許多團隊是抱著研究的心態參與資訊檢索績效的評估，因此，即使評估結果非常不好，他們也知道經過這樣的評估機制，已經證實所使用的技術或模式是不好的，應該加以檢討。這對於發展新技術，證實新構想的成立與否，是非常好的驗證程序。國內從事資訊檢索研究的學者其實很多，若能夠從評估新檢索技術與新檢索模式的角度，而非競賽的角度來看待這些評估活動，相信將會有更多的研究人員願意參與資訊檢索的績效評估。

參考文獻

Buckley, C. (1991). *Trec_eval IR Evaluation Package* [Computer software and manual]. Retrieved from ftp://ftp.cs.cornell.edu/pub/smart

Buckley, C. & Voorhees, E. M. (2004). Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval* (pp. 25-32). New York: ACM Press.

Bush, V. (1945). As We Will Think. *Atlantic Monthly, 176*(1), 641-649.

Chen, K.-H. et al. (2002). Overview of CLIR Task at the Third NTCIR Workshop. In *Proceedings of the Third NTCIR Workshop Part II: Cross-Lingual Information Retrieval Task* (pp. 1-38). Tokyo: NII.

Cleverdon, C. (1967). The Cranfield Tests on Index Language Devices. *Aslib Proceedings, 19*(6), 173-194.

Eguchi, K. et al. (2002). Overview of Web Retrieval Task at the Third NTCIR Workshop. In *Proceeding of the third NTCIR Workshop Part VI: Web Retrieval Task* (pp. 1-24). Tokyo: NII.

Fox, E. (1983). *Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts* (Tech. Rep. TR 83-561). Ithaca, NY: Cornell University, Computing Science Department.

Fukusima, T. & Okumura, M. (2002). Text Summarization Challenge 2: Text Summarization Evaluation at NTCIR Workshop. In *Proceeding of the third NTCIR Workshop Part V: Text Summarization Challenge 2* (pp. 1-6). Tokyo: NII.

Harman, D. (1993). The First Text REtrieval Conference (TREC-1). *Information Processing and Management, 29*(4), 411-414.

Harman, D. (1996). Panel: Building and Using Test Collections. In *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 335-337). New York: ACM Press.

Järvelin, K. & Kekäläinen, J. (2000). IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval* (pp. 41-48). New York: ACM Press.

Kando, N. (2001). Overview of the Second NTCIR Workshop. In *Proceedings of the Second*

NTCIR Workshop on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization (pp. 51-72). Tokyo: NII.

Salton, G. (1972). A New Comparison between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART). *Journal of the American Society for Information Science, 23*(1), 75-84.

Sparck Jones, K. & van Rijsbergen, C. J. (1976). Information Retrieval Test Collections. *Journal of Documentation, 32*, 63-73.

van Rijsbergen, C. J. (1975). *Information Retrieval*. London: Butterworth & Co.