中文資訊檢索測試集之設計與製作

The Design and Implementation of the Chinese IR Benchmark

陳光華
Kuang-hua Chen
國立臺灣大學圖書資訊學系助理教授

江玉婷
Yu-ting Chiang
國立臺灣大學圖書資訊學研究所碩士

【摘要　Abstract】

在國內資訊檢索研究已日趨受到重視，卻缺少合適的測試評估機制之背景下，本文參考國外各測試集的結構、特性與建構經驗，設計中文資訊檢索測試集的方法與程序，發展華文世界第一套中文資訊檢索測試集，實際地進行測試集的規劃建置工作。本研究所建立的測試集包括 132,207 篇新聞文件、50 個查詢主題、以及文件與查詢主題間的相關判斷，平均每個查詢主題有 16.32 篇相關文件。研究結果顯示，以統計抽樣的觀點，本測試集的文件數量具有一定的效度；查詢主題呈現詳盡且多樣化的查詢需求，能反映真實的檢索情況；由三位判斷者進行的相關判斷具有顯著的一致性，推斷它們是具有可信度的。本測試集不僅具有可應用性，並能作爲進一步資訊檢索與評估研究之基礎。

The research and development of information retrieval has made considerable progress recently. However, there is not any applicable test mechanism for system evaluation in the Chinese research society. This paper reports our research on the design and implementation of the first Chinese information retrieval benchmark. According to the framework and contents of the existing foreign benchmarks, we develop a methodology to establish the Chinese IR benchmark. An IR benchmark consists of three parts: document set, topic set, and relevance judgments. Our document set contains 132,207 documents collected from news web sites, the topic set contains 50 topics transformed from real users' information needs, and each topic has on the average16.34 related documents as a result of the relevance judgments. The results of our research show that the quantity of document set is valid from the viewpoint of sampling statistics. The topics reveal multiple kinds of information need, and they also reflect certain real retrieval environment. Besides, the judgments given by three judges have exhibited significant consistency, so we conclude their reliability. Although the benchmark is in its first edition, it possesses a complete structure and medium scale. On this basis, it is readily feasible to expand this benchmark's current scale to a proper large one in the near future.

【關鍵詞　Keywords】：

資訊檢索測試集；資訊檢索評估；文件集；查詢主題；相關判斷

IR Benchmark（IR Test Collection）；IR Evaluation；Document Set；Topic；Relevance Judgment

一、前言

在知識爆炸的現代，如何從不斷遽增的龐大資訊中快速且精準地進行搜尋篩選，已是眾所關切的焦點課題。因此，在資訊檢索系統扮演的角色日趨重要的背景之下，許多學者紛紛投入相關研究，期盼能發展更好的檢索技術與檢索系統，協助人們快速有效地掌握所需資訊。

資訊檢索系統不論在設計、研發、運作等各階段，評估均是其中不可或缺的重要環節。透過此程序，研究者能藉以驗證系統效益、比較各種檢索技術的優劣，以作爲改進之參考，使資訊檢索系統的運作及效能更臻完善。資訊檢索系統評估的研究發展，自 1950 年代至今，已有四十年以上的歷史。（註1）早期此方面相關的實證研究，大多是在規範化的環境（Laboratory Environment）中進行測試（Test），透過一些量化或質化的準則，衡量不同技術或不同系統間檢索效益之優劣。最早採用此評估模式的是1966年Cleverdon所進行的Cranfield II 計劃，它以文件集（Document Set）、查詢問題（Question）及相關判斷（Relevance Judgment）構成一組測試集（Test Collection）作爲測試的基礎資料，並訂定一套效益測量準則（Effectiveness Measurement），以評估多種索引方式之優劣。（註2）Cranfield 研究採用的實驗模型與測試方法，在系統評估的領域中一直廣受仿效與援用，直至今日仍佔有舉足輕重的開創性地位。然而，早期的測試集規模通常不大，與真實檢索環境間存在頗大的差距，因此植基於其上所發展的檢索系統，在實際運作時往往無法達到良好的效益。（註3）

1992 年，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）共同舉辦了文件檢索會議（Text REtrieval Conference，簡稱 TREC），透過大型測試集的建構，以及測試項目、測試程序、評估準則的訂定，提供不同檢索系統與檢索技術之間的標準評比環境，並舉辦論壇提供參與者討論及分享結果。（註4）它首創了前所未有的大型測試集，使測試環境得以更接近真實的情況，對檢索技術的發展與系統效益的提昇具有相當重要的貢獻。

影響資訊檢索系統效益的因素十分廣泛而複雜，系統評估工作亦應考量到各個層面，並不能僅依據單純的量化準則。無可否認的，如同 Cranfield II 及 TREC 這般的測試機制，的確在許多方面都有其侷限與爭議性，但是至目前爲止，它們確實是少數能得知系統可能效益的具體可行方案，對資訊檢索系統的研究與發展來說，還是具有十分重大的意義。

在今日資訊檢索研究蓬勃發展之際，各界紛紛意識到建立一致性評比環境的必要性。目前除了 TREC 之外，已有一些針對不同語言設計的類似機制嘗試開始運作，如 NTCIR（NACSIS Test Collection for IR Systems）計劃（註5）與 IREX（Information Retrieval and Extraction Exercise）計畫（註6）分別建立了日文測試集，AMARYLLIS 計畫則建立了以法文爲主的測試集（註7）。

反觀國內，目前亦有許多中文檢索系統，但合適的測試機制卻一直付之闕如。雖然 TREC 已建構一個小型的中文測試集（註8），不過由於只有參與 TREC 的單位才能使用測試資料，且其文件集之特性及使用文字之方式與台灣地區有所不同（註9），並不適用於國內資訊檢索系統的評估。因此，建構一個合適的中文資訊檢索測試集的確是迫切亟需的。本研究乃實際地經由蒐集文件資料、建構使用者查詢需求、進行相關判斷等程序，建立一個結構完整的中文資訊檢索測試集，期能提供給眾多的研究發展人員，作爲有效測試系統效益的基礎，實質地解決目前缺乏適當測試集的情形，也希望本研究建構測試集的方法與程序能作爲後續研究的參考依據。

以下本文將對測試集作進一步探討，並提出中文資訊檢索測試集實作之方法與成果；第二節文獻分析介紹測試集的發展演進及數個現行的重要測試集研究；三、四、五節說明測

試集建置的具體方法與步驟，並分析探討其特性及可用度；第六節除了簡要的總結之外，並提出對未來研究工作方向的建議。

二、文獻分析

使用者進行資訊檢索的一般模式，是將欲查詢的問題（Question）形成查詢問句（Query）輸入檢索系統（註10），系統在文件集中進行檢索，將可能符合需求的文件輸出給使用者。資訊檢索系統測試便是希望模擬這樣的程序，因此測試集通常會包括一組文件集、查詢問題以及二者之間的相關判斷。（註11）換句話說，我們可將測試集視爲系統測試的基礎資料，參與測試的系統必須在其上運作，依據所訂定的查詢問題，以文件集作爲檢索的對象，並將測試集提供的相關判斷結果視爲標準答案，以此進行檢索效益的評比。

早期的測試集大部分是爲了個別的測試計劃而建立，除了前述之 Cranfield II 之外，還有 ADI、MEDLARS、TIME、CACM、CISI、NPL、INSPEC、ISILT、UKCIS、UKAEA、LISA 等（註12）（註13）（註14）（註15）（註16），它們各依不同的測試目的、測試對象而有不同的組成架構，但共有的特性是測試集的規模均不大，且同質性頗高。舉例來說，Cranfield II 實驗所使用的測試集由 1400 篇文件、200 餘個查詢問題組成，文件範圍限定於太空動力學的領域，且文件長度均頗爲相似。（註17）由於這些測試集的規模及特性與真實的檢索環境差異頗大，因此依據它們進行的系統測試，效度（註18）受到許多質疑。（註19）1980 年代之後陸續發展的一些測試集如 OHSUMED、Cystic Fibrosis、BMIR-J2 等（註20）（註21）（註22），雖然有些規模稍大，但大體來說其形式還是與早期的測試集相似，也有著上述的缺失。

要建構一個測試集是很耗費時間及人力的，尤以相關判斷爲甚。再以 Cranfield II 研究爲例，若要將每個查詢問題逐一與每篇文件比對，必須執行數十萬次的相關判斷，其間所需花費之代價可想而知。因此，早期的測試集發展並不十分熱絡，也往往無法達到很大的規模，一旦有較完整的測試集出現，就算不盡符合系統評估之需求，通常還是會被許多人重覆利用，如 Cranfield II 測試集中的子測試集就廣受援用（註23）。另外，也有一些研究將多個測試集結合起來，如 SMART 系統評估計劃即採用六個不同主題領域的測試集進行實驗。（註24）

研究者對於測試集的需求是十分急切的，若能建立一可因應不同測試目的及需求的通用性「可攜式」（Portable）測試集，無疑將對資訊檢索研究產生相當大的助益。Sparck Jones 與 Van Rijsbergen 認爲，理想的測試集除了必須具備一定的規模之外，在文件及查詢問題的內容、型態、取得來源等方面要有相當程度的異質性，以反映真實的檢索環境，但是在測試集內部，也應包含一些同質性高的子測試集，提供特殊目的的測試之用。（註25）

1992 年，在美國舉行的 TREC 建立了一個不同於以往的大規模測試集，其文件集及查詢問題的結構特性亦與先前的測試集有顯著的差異，可說是爲資訊檢索系統評估測試的研究開創了一個新的里程碑，在此之後陸續發展的測試集，大部分均仿效其架構與模式。以下將 TREC 測試集的各部分組成要素作一簡述：(註26)

1. 文件集（Document Set）

   目前 TREC 已蒐羅約二百萬篇的各類型文件，且每年持續地擴展增加。TREC 使用標準通用標誌語言（Standard Generalized Markup Language，簡稱 SGML）及文件型態定義檔（Document Type Definition，簡稱 DTD）爲每篇文件加上標記（Tags），以利系統進行各種剖析（Parsing）工作。

2. 查詢主題（Topic）

   TREC 的查詢問題形式與早期的測試集有顯著的不同。它模擬使用者的資訊需求，以各種方式、各種角度陳述，並利用結構化的欄位呈現，稱之爲查詢主題（Topic）。TREC 每年建構 50 個新的查詢

主題，並將之循序編號，以便於利用辨識。至 TREC-7（1998）爲止，已有 400 個不同的查詢主題。每年 TREC 會根據先前的測試結果或當時的特殊需求，將查詢主題的結構與呈現方式作適度的修正，使其能發揮最佳的測試效能。TREC-1 與 TREC-2 的查詢主題所包含的欄位多達 10 個，十分詳細且複雜；而近年的查詢主題則有簡化的趨勢，主要以<title>、<description>及<narrative>三部分爲主，呈現不同詳簡層次的資訊需求。

3. 相關判斷（Relevance Judgment）

TREC 採取二元化的相關判斷方式，即將所有文件分爲相關與不相關二個層次，只要文件中一部分與查詢主題有關聯即視爲相關。對於 TREC 這樣的大型測試集來說，要逐一將每個查詢主題與每篇文件作詳盡的相關比對，所須耗費的工程可想而知，因此 TREC 採用 Pooling Method 輔助相關判斷的進行：在參與測試評比的系統均能提供相關性排序功能的前提下，抽取各系統送回之測試結果的前 n 篇文件，合併形成一個 Pool，視之爲該查詢主題的相關文件候選集，去除集合中重覆的文件後，再送回給該查詢主題的原始建構者進行相關判斷。利用此方法的主要精神是希望透過多個不同的系統與不同的檢索技術，盡量網羅所有可能的相關文件，減少人工判斷的負荷。

目前正積極發展的其他語文測試集，均承襲 TREC 查詢主題多欄位化的設計概念，有些也將之擴展變化，例如日本的 BMIR 測試集即加入了功能性標記的欄位（註27）。由於測試集發展的規模愈來愈大，以 Pooling 輔助相關判斷的方法廣受延用，除此之外，有些測試集也同時利用系統評比的結果對測試集的標準答案作修正（註28）。另外值得注意的是，許多測試集在進行相關判斷時，均傾向以多元化的相關判斷取代 TREC 所採用的二元化模式。

表一整理了自 1960 年代至今的重要測試集基本資料（註29），從中可看出測試集規模的演變與部分特性。早期測試集的文件大部分由題名、摘要、關鍵詞等簡短的書目性資料組成，主題領域也多屬專門。近年來測試集的發展主要以 TREC 爲標竿，逐漸趨向前述之「可攜式」目標，雖然還未到達理想的境界，但無論在規模、組成特性等各方面均較以往大幅增進，包含多主題的全文文件以及詳盡的查詢問題，且正持續擴展之中。

無疑地，資訊檢索測試集對資訊檢索研究的價值，已受到相當程度的重視與肯定。但是從 Cranfield II 研究開始，還是不斷有學者對測試集、測試方法、評估準則等各方面的有效性提出質疑。在文件集方面，由於早期的文件集大多只有數千至數萬篇，且其中有許多僅包含文件摘要部分，因此主要受到的批評是規模太小、文件的同質性過高，無法反映真實的檢索環境，使測試的結果較無意義與代表性，各系統間的顯著差異也較難顯現（註30）（註31）（註32）。但從表一可看出，近年來發展的幾個較大規模的測試集在此方面已較以往改進不少。

在理想的情況下，測試集的查詢問題應爲真實環境中的使用者資訊需求，但一般來說，由於蒐集這些需求並不容易，且爲了使實驗測試能獲得較佳的控制，查詢問題通常會以人工模擬建立，或是對使用者原始的需求作部分修飾，如 Cranfield II 與 TREC 的查詢主題均是由上述方法建構而成。因此，測試集中的查詢問題常被認爲過於人工化，使得系統測試的效度產生疑慮。（註33）（註34）（註35）

在查詢問題的內容方面，大多數測試集雖然是以自然語言的方式陳述，但卻十分簡短，所包含的訊息相當少，因此有不少學者認爲它們過於簡化使用者的需求。（註36）（註37）（註38）近來 TREC 首創的查詢主題，以多個欄位呈現不同層次的資訊需求，可說是一大突破，後來發展的測試集也紛紛仿效這樣的模式。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，也逐漸加入其他層面的描述。（註39）

表一　各測試集之基本資料

測試集	文件數	文件集大小 (MB)	文件平均字數	查詢問題數	查詢問題平均字數	查詢問題平均相關文件數	主題領域	相關判斷層次		語文
								相關	不相關	
Cranfield II	1,400	1.6	53.1	225	9.2	7.2	太空動力學	4	1	英文
ADI	82	0.04	27.1	35	14.6	9.5	文獻學	N/A		英文
MEDLARS	1,033	1.1	51.6	30	10.1	23.2	醫學	2	2	英文
TIME	423	1.5	570	24	16.0	8.7	世界情勢	N/A		英文
CACM	3,204	2.2	24.5	64	10.8	15.3	ACM 通訊	N/A		英文
CISI	1,460	2.2	46.5	112	28.3	49.8	資訊科學	N/A		英文
NPL	11,429	3.1	20.0	100	7.2	22.4	電子、電腦、物理、地理	N/A		英文
INSPEC	12,684	N/A	32.5	84	15.6	33.0	物理、電子、控制	2	1	英文
ISILT	800	N/A	N/A	63	N/A	8.4	文獻學	1	1	英文
UKCIS	27,361	N/A	182	193	N/A	57	生化	2	2	英文
UKAEA	12,765	N/A	N/A	60	N/A	N/A	核子科學	2	1	英文
LISA	6,004	3.4	N/A	35	N/A	10.8	N/A	N/A		英文
Cystic Fibrosis	1,239	N/A	49.7	100	6.8	6.4-31.9	醫學	6	1	英文
OSHUMED	348,566	N/A	250	101	10	17/19.4	N/A	2	1	英文
BMIR-J2	5,080	N/A	621.8	60	102.2	10.6/28.4	經濟、工程	2	1	日文
TREC (TREC-1~6)	1,754,896	~5GB	481.6	350	105.8	185.3	多主題	1	1	英文
AMARYLLIS	336,000	201	N/A	56	N/A	N/A	多主題	N/A		法文
NTCIR	300,000	N/A	N/A	100	N/A	N/A	多主題	2	1	日文
IREX	N/A	N/A	N/A	N/A	N/A	N/A	多主題	2	1	日文

相關原本就是較主觀且模糊的概念，相關判斷更會因判斷者、判斷情境等諸多因素而可能產生很大的差異，加上進行相關判斷時往往宥於時間人力等種種限制，無法作十分周詳的考量，通常只能採取一些可行性較高的權宜方案。因此，相關判斷在測試集中一直是最受爭議的部分。歷來學者對於測試集中相關判斷部分的質疑，主要可歸納爲相關層面、相關測量尺度、相關判斷者、相關判斷的完整性等幾個議題。

在進行檢索系統評估時，宥於使用者相關層面的複雜、模糊與不確定性，大部分的研究者僅論及主題相關層面，前面所介紹的測試集大部分亦採用此觀點。但畢竟相關無法單純由客觀的主題因素決定，許多學者主張在相關判斷時應納入如情境相關等多層面的考量。

相關與不相關之間爲一連續地帶，相關程度很難清楚地劃分，不同使用者間的認知也往往有相當大的差異（註40）（註41）。但是由於測試集中文件與查詢問題的相關程度必須有較爲客觀且明確的定義，採用抽象的排序或連續尺度是較爲困難的，因此現行測試集大多採用類別尺度。在必須考量實施可行性的前提下，測試集所採用的測量尺度是否能準確地反映實際的相關程度差異，是值得進一步探討的。

一般認爲資訊需求者是最具資格進行相關判斷的人（註42），因此理論上相關判斷應由原始的查詢問題建構者進行。但對依據真實使用者需求構成查詢問題的測試集來說，如此實施的困難度較高，所以大多數的相關判斷是由一位或多位次判斷者（Secondary Judges）進行。至於如何結合不同判斷者的意見以形成最後的相關判斷結果，則有許多不同的做法，例如 IREX 以第三者參考其他人的相關判斷進行

最後的決策工作，Reid 等人則提出加權式計算方法結合不同的判斷決策。（註43）

不同判斷者所產生的相關判斷結果，通常也有相當程度的歧異產生。TREC 的實驗顯示，不同的相關判斷者之間有高達 71%的不一致狀況。Saravecic 歸結先前的研究結果發現以下現象：（1）判斷者的主題專長與查詢主題愈接近，判斷的一致性愈高；（2）較缺乏查詢主題知識的判斷者，愈容易將文件判斷爲相關（意即判斷結果愈爲鬆散）；（3）判斷爲不相關的一致性通常高於判斷相關的一致性。（註44）（註45）在如此不穩定的相關判斷之下，測試集的有效性是否受到影響呢？TREC 的實驗結果顯示，相關判斷的差異並不會影響系統效益優劣排序的穩定度（註46），Burgin、Kazhdan、Cleverdon 及 Lesk & Salton 的研究亦得到相似的結論。（註47）（註48）但是，Harter 則認爲這樣的測試集有效性仍是值得質疑的。（註49）

相關判斷的完整性指的是查詢問題在文件集中真正相關的文件被判斷爲相關的程度。求全率（Recall）（註50）是目前系統測試的重要準則，理想中測試集應找出文件集中所有可能的相關文件，才能精準地計算求全率，但是相關判斷的工作非常耗費人力、時間，且判斷者的不同認知會產生不同判斷結果，使得可信度（註51）受到影響，因此要獲致一個相當完整的相關文件集合是十分不容易的，測試集漏失真正相關文件的高比例（註52）使得評估系統效益時計算求全率的意義令人質疑。（註53）但亦有學者認爲，吾人可以透過事前對可能遺漏的相關文件數量的預測，減低評估時的偏差。（註54）

三、文件集

本研究發展中文資訊檢索測試集的方法，主要參照現有測試集的實施經驗，考量各種不同作法的優缺點與可行性，並根據中文資訊檢索系統的型態與特性，實際設計一套建構測試集的模式，依各步驟的不同需求，選擇採用適當的研究方法與工具，以建立一套可因應不同評估需求的通用性測試集。在評估對象方面，考慮以檢索文字式資料爲主、以單篇文件爲最小檢索單位、能夠計算文件與輸入系統的查詢問句間之相關性、並提供相關排序輸出的一般資訊檢索系統；評估方式則考量以相關爲主的效益測量方法，在測試集中提供所需的標準答案。本節及以下二節即依文件集、查詢主題、相關判斷三個部分，說明中文資訊檢索測試集的建構方法、程序以及有關的分析討論。

本研究在一年的時間內，自WWW網站下載大量的新聞報導全文（註55）。選擇新聞文件的原因主要是目前網際網路上許多新聞性網站均提供大量的全文式新聞，且多數具有免費與正當的存取管道，因此在資料取得上實施的困難度較低。再者，網際網路上的資料傳播更新十分迅速，內容大多極爲新穎，主題分佈也非常廣泛，以其作爲文件集的主要組成元件，應能即時反映目前語言文字的使用情形與特性，如此不僅可以測試出資訊檢索系統是否能適應時代的走向及需求，也較能切合一般資訊檢索系統或搜尋引擎的設計目的與應用對象。文件下載來源主要爲中時電子報（包括中國時報、工商時報與中時晚報）、中央日報、中華日報等三個新聞網站中的報紙新聞電子版部分，這些網站均提供綜合性主題的新聞全文，且文件長度不致過短，大致上符合本研究對文件集構成之要求。

爲了使系統易於對文件進行辨識與處理，文件格式應具一致性，在測試時才不致因爲文件中的其他雜訊使系統的檢索結果受到影響。因此，本研究將取得的 html 文件整理爲一致的純文字格式，刪除新聞報導正文之外的資訊。另外，並考慮文件原始的結構與特性，將之加上標記，使每篇文件具有相同的格式與資料項目：除了各文件原有的新聞標題與新聞內容之外，並統一加註文件來源識別碼與新聞報導日期，如圖一所示。但本研究僅針對文件的呈現形式作統一的處理，並不對新聞內容作任何更改。

表二說明本文件集的來源與數量，目前共有 132,173 篇文件，約佔 200MB。文件內容包括政治、財經、社會綜合、生活、體育、藝文、國際、資訊科技等多元性主題。

文件集規模愈大，愈能接近真實的檢索環境，測試集本身的效度也愈高。因此，吾人可從統計抽樣的觀點檢視文件集的效度，意即依據取得的文件樣本進行之測試，結果能推估實際情況的有效程度。(註56)(註57)真實檢索環境中的文件數量通常十分龐大，因此我們可假定它是一個無限大的母體，其分布應趨近於常態分配，如此可進一步推得一個比較保守的應取樣本數 n： $n \cong \frac{[Z(\alpha/2)]^2}{4b^2}$ (b 爲容許誤差值)。以本測試集的 132,173 篇文件而言，若將容許誤差設爲 0.5%，信賴區間可高達 99.9%以上；而若以估計中慣常採用的信賴區間 95%或 90%所得的結果來推斷，如此的文件集規模應已能達到相當高的效度。(註58)

```
<doc>
<id>chinatimes_focus_0005660</id>
<data>05071999</date>
<title>解決高鐵融資 尋求第三管道</title>
<text>
<p>
【記者羅兩莎台北報導】據負責台灣高速鐵路聯合貸款的主辦銀行表示，高鐵融資問題目前仍卡在銀行團、交通部高鐵局以及台灣高鐵公司「三方合約」內容的訂定。在銀行團和交通部一直未能就相關歧見達成共識之下，三大主辦銀行原則決定，將尋求行政院經建會等第三管道與交通部協調，以儘早解決銀行團和交通部之間對融資問題的歧見。
<p>
高鐵案將向國內銀行融資二千八百多億元，這項聯貸案確定由交銀、台銀和中國國際商業銀行共同主辦。不過，由於高鐵是國內首宗ＢＯＴ案，潛在風險究竟有多高，銀行無從評估。三大主辦銀行與交通部和台灣高鐵公司訂定貸款合約時，重點亦著重在風險控制以及債權確保。
<p>
據主辦銀行主管表示，銀行當然希望債權確保不會有問題，譬如，在三方合約中訂定，由政府出面保證萬一將來台灣高鐵公司蓋不下去時，政府可以出面買下，負責把工程完成等。
<p>
但是三大主辦銀行經多次與交通部協商，前述問題均未達成共識。
</text>
</doc>
```

圖一　文件標記範例

表二　文件集數量統計

中國時報	工商時報	中時晚報	中央日報	中華日報	總數
38,163	25,812	5,747	27,770	34,728	132,173
28.8%	19.5%	4.4%	21.0%	26.3%	(200MB)

四、查詢主題

查詢主題之建構主要有以下三個程序：

1. 查詢需求之徵集

爲了加強測試集與真實檢索環境的相似度，本研究希望經由徵集真實環境中使用者的查詢需求，獲致查詢主題建構的參考來源，再將其修正轉化成正式的查詢主題。我們透過網路問卷的方式進行調查，共徵得 405 個查詢需求。問卷內容由封閉式與開放式的問題組成，收集使用者資訊需求的類別、主題、詳細內容、及各種相關資訊。實施此方法的基本假定爲：使用者均能對其特定之查詢問題作清楚且詳盡的陳述。

2. 查詢需求之篩選

由於蒐集而來的問卷答卷品質並不整齊，對問題的敘述詳簡各異，問題形式與所涵蓋之主題範圍也不一定適合作爲測試集的查詢主題，因此我們分三階段對其進行篩選的工作，找出 50 個最合適的查詢需求。

第一階段以人工檢視的方式考慮填答者對查詢需求之陳述方式與需求主題之適

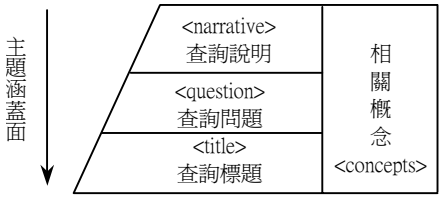
切性，過濾敘述不清、不夠詳盡、或過於主觀的需求，在需求主題方面則將範圍過廣、與文件集主題不符、型態特殊、或變動過大的即時性問題刪除。第二階段利用龍捲風全文檢索軟體，考慮可能相關文件之數量，判斷查詢需求的主題範圍是否過於廣泛或過於狹窄，另外也透過觀察檢索所得之前 n 篇文件與查詢需求之相關情形，初步預測查詢需求之難易度。第三階段以人工檢視的方式，考慮的層面包括需求之事件主題的相似性，以及需求敘述的詳簡及清晰程度，選擇最適當的 50 個查詢需求。各次篩選結果如表三所示。

表三　查詢需求之篩選

	篩選方式	刪除數量	剩餘數量
第一次篩選	人工檢視	163	242
第二次篩選	以全文檢索軟體輔助	173	69
第三次篩選	人工檢視	19	50

3.　查詢主題之建構

此部分主要以前述篩選產生的 50 個查詢需求爲藍本，依據所訂定的建構原則，將問卷回答內容轉化成標準一致的格式。每個查詢主題均以 <title> 、 <question> 、 <narrative>及<concepts>等四個欄位呈現查詢需求的內容，各欄位之特性、意義與建構依據如表四所示。查詢主題中<title>欄位所涵蓋的主題範圍最廣，其次是<question>欄位，<narrative>欄位雖然敘述最詳盡，卻也是其中最爲特定的，而<concepts>欄位中的詞彙則可能涉及上述各層次的主題，其間之關係如圖二所示。



圖二　查詢主題之欄位關係

表四　查詢主題各欄位說明

欄位	中文名稱	內容	組成語法	建構依據
<title>	查詢標題	對查詢主題的簡單描述	名詞 或名詞片語	查詢需求主題
<question>	查詢問題	利用簡短語句傳達查詢需求的主要內容	一至二個句子	查詢需求
<narrative>	查詢說明	對查詢問題的進一步解釋、專有名詞的釋義與澄清、相關與不相關資訊之列舉、對相關文件的特殊需求與限制	數個句子	查詢需求
<concepts>	相關概念	與查詢主題中各層次敘述相關的詞彙	一至數個詞彙	相關與不相關文件之詞彙

在釐清查詢主題之結構與意義之後，將先前篩選出的 50 個查詢需求依據規範編輯轉化爲正式的查詢主題，並使其能具有相近的呈現模式及一定的詳盡程度。建構查詢問題時必須遵照下列之主要原則：(1)儘量使用清楚、易懂、且常用詞彙表達，艱澀難以理解之詞彙應儘量避免出現，或必須在查詢主題中加以解釋。若爲人名、事件、專有名詞等，應選擇正式且一般普遍常用的名稱；（2）句子陳述應符合正確之文法，並加上標點符號。內容應以簡潔直述的方式呈現，在能充分傳達所需資訊的前提下，應刪除不必要的贅語，另外，陳述中也應避免出現隱喻、暗示、及不正式的用語。

除了上述原則之外，各欄位也分別有其建構規範，主要依據所蒐集之各層面的使用者需求加以轉化。較特別的是在<concepts>相關概念的部分，我們假定在去除一般常用的詞彙之後，相關文件裡使用率高的詞彙傾向是與查詢主題有關的概念，而常出現在不相關文件中的詞彙則應避免列舉出來。因此本研究透過一中文斷詞程式，分析與查詢主題十分相關及十分不相關文件中所使用的詞彙及頻率，作爲相關概念關鍵詞列舉的主要來源。建構完成的正式查詢主題如圖三所示。

<topic>
<number>01-011</number>
<title>金融機構合併。</title>
<description>
查詢我國政府單位鼓勵金融機構合併之各項措施。
</description>
<narrative>
財政部等相關單位爲健全金融市場、改善金融體質，推動了一連串鼓勵銀行、證券商及保險公司等金融機構合併的措施。相關文件內容包括各項具體的獎勵優惠辦法、施行細節、法令中明定之規範條文、以及各界對相關政策的討論與評估。若文件中只陳述金融機構合併之個案，視爲不相關。
</narrative>
<concepts>
金融機構、合併、銀行合併、租稅優惠、租稅減免、稅前盈餘、低利融資、促進產業升級條例、財政部、經濟部、央行、中央銀行、增值稅、印花稅、證交稅。
</concepts>
</topic>

圖三　查詢主題範例

最後產生的 50 個查詢主題可概略劃分爲 9 個新聞類別，其中以社會綜合類最多（佔 28%），其次爲生活類與科技資訊類。查詢主題的平均總字數約爲 169 字，與其他測試集相較，相近欄位的字數相差並不大，但本研究 50 個查詢主題之間的變異程度較低，意即各查詢主題陳述文字的多寡分布較平均，並沒有特別簡略或特別長的查詢主題，相關統計值如表五所示。

本研究所建構的查詢主題是依據真實環境中的使用者查詢需求修正轉化而來，並以多種不同的形式與詳簡層次呈現。經由查詢需求的篩選與建構準則與規範的訂定，我們可較爲確保其在測試評估時之適用性，查詢主題的內容品質與陳述的明確性亦能得到一定的控制。在測試的功能上，這樣的查詢主題不僅能反映實際情況，且能展現多種不同的查詢問題形態。

透過個別候選文件集中文件與查詢主題的相關情形，我們也可觀察不同查詢主題應用在測試上的不同意義與特性。就本測試集初步進行相關判斷的結果來看，候選文件中真正與查詢主題相關的比例多在 0.2-0.3 之間。然而，查詢主題的平均文件相關度僅能反映片斷的狀況，仍需將以下二方面的資訊納入考量：（1）查詢標題與查詢問題之關係：由於查詢問題是判斷者主要依據的判斷基準，因此以查詢標題所產生的候選文件集，真正相關文件的比例也會應有多寡程度的差別，若查詢標題與真正的查詢問題相差太大，將會使平均相關度變得很低；（2）查詢主題與文件集之關係：在各個查詢主題的候選文件集中，相關度並不會以固定的模式分布，可能有多種不同的情況產生，例如常態分布、隨機分布與兩極化分布的相關判斷結果，其內在的意義亦是可深入探討的。

另外，從三位判斷者相關判斷結果的一致情形，也可判斷查詢主題在下列方面可能具有的不同表現：（1）查詢主題是否提供詳盡的資訊？（2）查詢主題是否表達清晰？（3）查詢主題的專指性如何？

透過以上資訊，可以幫助吾人研判查詢主題可能的難易度，並推估其不同特性所造成判斷困擾的程度。不過，由於查詢主題難易度仍是一個模糊的概念，不僅牽涉的因素多且複雜，在不同的觀點與基礎下，對它的解釋也會有所不同，目前並無法清楚界定。但是系統仍可以藉此觀察每個查詢主題所展現的不同特質，或是從中選擇合適的查詢主題進行測試。

表五　各測試集查詢主題長度比較

	欄位	最小字數	最大字數	平均字數	標準差	標準差/平均字數
本研究	<title>	3	13	6.52	2.23	0.34
	<question>	12	37	23.64	5.92	0.25
	<narrative>	57	141	93.90	20.43	0.22
	<concepts>	26	74	44.68	11.58	0.26
	Total	103	244	168.74	27.77	0.16
TREC Chinese Topic 1-54	<title>	4	29	12.30	5.58	0.45
	<desc>	6	35	17.48	7.40	0.43
	<narr>	31	174	81.54	30.28	0.37
	Total	53	204	111.32	31.36	0.28
TREC-6 Topic 301-350 （中文翻譯）	<title>	3	13	6.80	2.28	0.34
	<desc>	7	87	30.14	16.88	0.56
	<narr>	26	217	94.56	42.15	0.45
	Total	64	237	131.5	42.03	0.32

五、相關判斷

執行相關判斷的主要目的是建立查詢問題與文件集中文件的關聯程度，而此階段工作是以下列假定爲前提：（1）使用者的資訊需求能透過相關的概念得到滿足；（2）相關判斷者能依據查詢主題做出客觀而正確的判斷，不受當時外在環境或個人內在因素的影響；（3）相關判斷結果是穩定而不易變動的，判斷者不需在不同的時間對同一組查詢問題與文件重覆進行多次的判斷；（4）相關判斷者能將文件與查詢主題之間的關聯性，量化區分爲數個不同的相關等級或類別；（5）個別查詢主題與文件間所形成之相關現象是相互獨立的，不會受到其他判斷結果的影響。系統評估必須依據相關判斷的結果進行測試，始能進一步得知其可能的效益，因此相關判斷可說是測試集的關鍵部分。首先，應建立一些判斷的準則與程序，包括選取相關判斷者、決定判斷尺度及訂定判斷規範等。另外，爲了縮小欲進行相關判斷的文件數量，必須針對每個查詢主題進立相關文件候選集。在相關判斷進行完畢之後，則根據各判斷者的判斷結果，計算每篇文件的總相關分數。茲將主要步驟分述如下：

1. 相關判斷實施之準則與規範

本測試集的查詢主題數量眾多，並且經過重重的篩選以及結構重組建構而成，而每個查詢主題必須進行判斷的文件亦不少，在此種種限制下，50 位查詢需求提供者很難能夠完全配合研究的進行。本研究因而改採以次判斷者（Secondary Judges）進行相關判斷。另外，爲了增強判斷結果的信度，不致因單一判斷者的特殊認知或可能產生的錯誤影響測試集的有效性與客觀性，對每一個查詢主題安排三位次判斷者進行相關判斷，並分別將其定位爲具主題專長、檢索專長以及一般使用者的角色。在相關判斷中，若除去純爲個人的特殊觀點或特殊判斷情境等主觀因素，具有此三種不同外在角色特性的判斷者，應能反映一般檢索情境中可能造成相關認知不同的情況，換句話說，透過三位次判斷者的判斷結果，可以推測某文件在真實情況中相關狀況是具有一致性或爭議性，使判斷結果能有一定的效度。另外，由於詳盡的相關判斷耗日費時，所有的判斷工作不可能以一人之力爲之，因此假定具有同一判斷背景的判斷者在客觀性前提下，對查詢主題的認知觀點是相似的，而相關判斷的結果可以、也只能反映該背景的判斷特性。透過對判斷者背景的控制與整合，本測試集的相關判斷結果應是較爲可信與客觀的。

本研究在相關層面的考量以主題相關爲主，重視文件與查詢主題之間較爲具體、可形諸文字的主題關係。在這樣的概念基礎上，判斷者應客觀地將查詢主題與文件內容作相關性連結，也因此我們在研究中以不同次判斷者進行的相關判斷工作，其間的信度與一致性應不致過低。在判斷的決策層級方面，由於測試集中文件與查詢問題的相關程

度必須有較爲客觀而明確的定義，採用排序或連續尺度是較困難的，而若單純區分爲相關與不相關二類，又較不實際。因此本研究採用多元式類別測量尺度，將相關程度分爲非常相關、相關、部分相關與不相關四個等級。雖然它們嚴格來說是屬於類別尺度，但就相關的程度來說，它們仍隱含順序的概念，即非常相關＞相關＞部分相關＞不相關，因此我們亦根據其相關的程度分別給予 4 至 0 的相關分數。

2. 相關文件候選集之建立

由於相關判斷非常耗費人力與時間，而文件集的文件數量眾多，要逐一對每篇文件進行判斷是不太可能的。在儘可能兼顧相關判斷可行性與完整性的考量前提下，本研究利用查詢主題各欄位間主題涵蓋面相互隸屬的特性，對每個查詢主題建立一相關文件候選集（文件數量介與 30 篇與 200 篇之間），再針對候選集中的每篇文件以人工進行相關判斷。進行方式是利用「龍捲風」全文檢索軟體，根據查詢主題中主題意義最廣的欄位進行檢索，並配合使用各種檢索技巧與策略（如詞彙擴展），期能儘量完整地蒐羅所有可能相關的文件。運用此法要達到基本的信度與效度，必須建立在二個重要假定之上：（1）查詢標題所提供之資訊，在主題意義上能完全涵蓋該查詢主題中對資訊需求的所有陳述；（2）檢索系統或檢索者具有優良的檢索能力，能找出文件集所有可能與所依據之欄位內容相關的文件。建構產生的 50 個相關文件候選集，平均約有 94 篇文件，如表六所示。

3. 相關判斷之實施

在進行相關判斷時，每位判斷者必須詳細閱讀並了解查詢主題，並以<question>欄位作爲主要的判斷依據，逐一檢視候選文件集中每篇文件的內容，將其指派到判斷者認爲適當的相關類別。判斷者必須在一段連續的時間內完成一個查詢主題的判斷工作，以儘量確保判斷標準前後的一致性。同一集合中文件的呈現順序，則依據文件識別碼排列。實驗中，18 位判斷者共耗費約 230 小時進行了近 15,000 次的相關判斷工作。（註 59）

表六　相關文件候選集數量統計

文件數	頻率（查詢主題數）		
31-50	14	平均數	93.82
51-100	15	最大值	198
101-150	12	最小值	30
151-200	9	標準差	47.137
總計	50	總數	4691

4. 相關分數之結合

測試集必須建立一個查詢主題與文件相關程度的表列，即俗稱的「標準答案」，使系統能在同一基準上進行效益的比較與評估。因此，相關判斷工作實施完畢之後，尚必須結合各判斷者的判斷結果，爲每篇文件建立標準統一的相關分數，再決定如何解釋此分數的意義，以及如何應用其進行系統評估。本研究以較直觀的想法結合多個判斷結果，主要有下列基本原則：（1）每個判斷者對於整體相關分數有相等的貢獻；（2）每個相關類別對相關判斷決策具有等同的地位，因此單純以前述給定的相關分數進行計算，不另作加權；（3）個別判斷結果是獨立的，結合時不因其分布狀況的不同而有不同的計算方式。

依據上述想法，本研究將三位判斷者對同一篇文件所做的判斷結果，以下列公式結合，計算其與該查詢主題的相關程度值：

$$R = \frac{(X_A + X_B + X_C)/3}{3}$$

其中 X 爲各判斷者對文件所給的類別等級，A,B,C 則爲三位判斷者之代號，若所得的值愈接近 1，表示二者愈相關，反之則愈不相關。

就系統評估功能的觀點，在獲得每一文件的相關度之後，還必須對這些值賦予意義，以配合使用一些效益評估準則。以目前最常用的

求全率與求準率爲例，進行效益計算時必須將文件劃分爲相關與不相關二類，因此本研究亦將測試集中的相關判斷結果進一步作二元化的區分，取得一個合理可靠的相關度門檻值，將具有某個相關度以上的文件定義爲相關文件。實際作法爲利用 Kappa 一致性係數（註60）分析此二不同的相關區分方式在判斷結果中的一致性表現。若判斷者對於相關與不相關的認知具有較高的一致性，即表示此區分方法是較適當的，因此我們可依據其決定一個相關度門檻值，將相關度大於等於此值之文件視爲相關，反之則視爲不相關。經實驗檢測後，本研究將相關度門檻值訂爲 0.556，50 個查詢主題平均有 16.34 篇相關文件，佔候選文件集的 17.4%，如表七所示。就 TREC 的二元式相關判斷標準來看，其作法是只要文件與查詢主題有部分相關即視爲相關之文件，相較之下本研究的二元式相關劃分是較爲嚴格的。若使用類似 TREC 的基準，本研究相關度門檻值應降爲 0.333，平均相關文件則會提升到 25.22 篇，佔候選文件集的 26.9%。

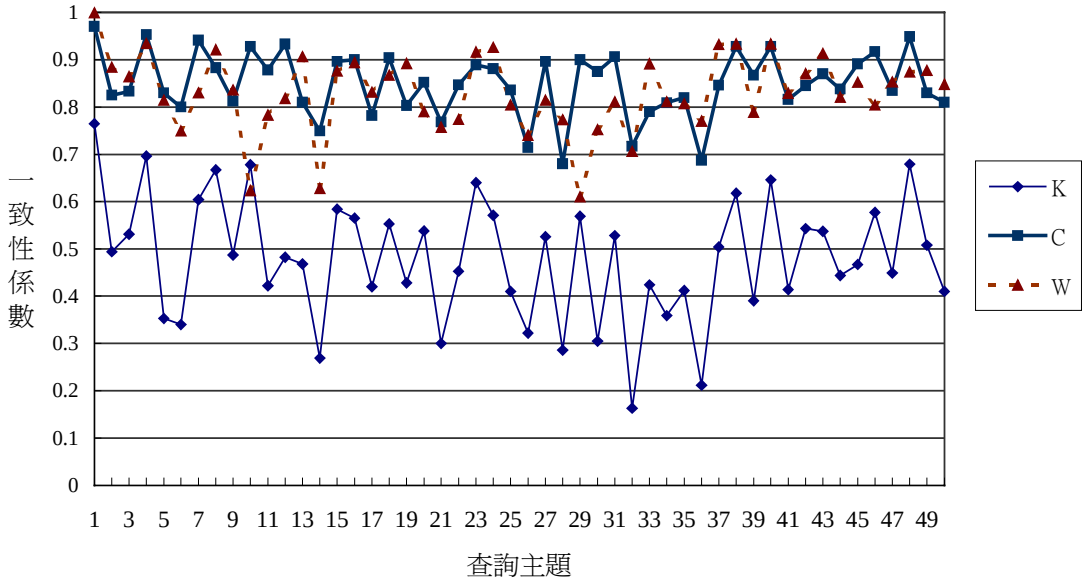
表七　二元式相關劃分

相關文件數	相關度門檻值	
	0.556	0.333
	頻率(查詢主題數)	
0-10	12	6
11-20	23	15
21-30	11	12
31-40	4	11
41-50	0	3
51-60	0	1
61-70	0	2
相關文件數總和	817 (17. 4%)	1261 (26.9%)
平均值	16.34	25.22
最大值	39	68
最小值	3	4
標準差	8.442	14.403
標準差/平均值	0.517	0.571

測試集中查詢主題相關文件數量的多寡會受到許多因素的影響，包括文件集的範圍與數量、查詢主題的訂定、以及相關判斷的標準等等，而評估測試的目的與規範亦是重要的考量點。由於本研究並無涉及評比機制的設計，因此若單與現行的測試集比較，就此規模大小初步來看，這樣的相關文件量尚屬恰當，而相關文件數在候選文件集中所佔的比例亦與 TREC 頗爲接近（請參見表一之數據資料）。但是，細部觀察各查詢主題的相關文件數發現，它們的分布並非十分集中，在候選文件集中所佔的比例亦有不同，如此在依測試時不同的需求，各查詢主題應能發揮其不同的功能特性。

影響相關判斷的因素非常多，不同的判斷者在個人知識、智力、認知狀態、判斷經驗等的背景下，會產生不同的相關判斷結果是十分正常的。但也由於牽涉到的因素如此複雜、不確定，在本研究的相關判斷實驗中，並無法對各種變因做很嚴格的控制，相關判斷結果完全一致的比例（即三位判斷者均將文件指派到同一個類別的情形）並不是很高，僅佔所有判斷情形的 7%左右。然而，爲了使測試集能夠客觀地測試資訊檢索系統的效益，相關判斷仍必須有某種程度以上的一致性，才足以顯示判斷結果是具有可信度、沒有偏頗的。因此，本研究利用三種不同的統計量對相關判斷結果進行檢測與驗證：包括前述之 Kappa 一致性係數 (*K*)、Kendall 一致性係數 (*W*)（註61）、以及考慮兩兩相關分數間距離的一致度 (*C*)（註62）。

圖四爲 50 個查詢主題在三種統計量之下的表現與分布狀況。雖然它們計算一致性的統計原理不盡相同，但圖中曲線大致上的起伏情形仍頗爲相似，*W* 值與 *C* 值大致上均在 0.7 以上，平均值也達到 0.8，變異數則不到 0.1，由這樣的結果可推測判斷不同的現象應大多出現在相鄰的二個類別，判斷者對於相關等級的認知也多具有一定程度的共識。另外，本研究亦進一步利用統計公式計算 Kendall (*W*) 與 Kappa (*K*) 的顯著性，結果發現兩者的表現均低於顯著水準 (α)，顯示三位判斷判斷結果的一致性已具有顯著意義，可初步研判本研究相關判斷的施行具有一定的信度，足以作爲系統評估依循之客觀基準。



圖四　相關判斷一致性分析

六、結論與建議

本研究已實際完成一組包含文件集、查詢問題以及相關判斷的測試集，也初步驗證了其建構程序是可行的。與其他測試集相較，本測試集的規模在中等以上，文件集與查詢主題均盡量接近真實之檢索環境，而相關判斷亦結合多位判斷者進行，減低了判斷結果可能出現的偏差情形。在各界急於研發中文資訊檢索系統的今日，預期此測試集之出現，應能解除國內無從取得中文測試資料的現狀，使檢索系統的發展能有更高的可行性，也期望它能成爲後續相關研究的基礎。

資訊檢索評估所涉及的層面相當廣泛且多元，而建立一個合適有效之測試集的困難點，除了在具體實施時必須耗費大量的時間與人力之外，測試集實際應用的效能與可行性，及其是否能兼顧反映真實檢索情境與系統評估的客觀性等需求，均是目前爭議性頗高、有待進一步探討的複雜課題。因此，未來測試集的建構與應用仍有很大的發展空間，在此就以下方面提供幾點淺見，作爲進一步研究之參考：

（一）　測試集之改進與擴展

在文件集方面，若能進一步擴展文件集規模，將會增進測試集之效度。此外，亦可透過蒐集更多不同性質與類型的文件，使文件集的適用範圍與測試功能更能符合多樣化的檢索環境與檢索需求。

在查詢主題方面，可加入非主題式的陳述，如查詢需求形成的原因、背景、特殊需求情境等方面，使其能成爲以使用者爲出發點的測試機制，並更接近真實檢索環境的使用者需求。若能深入分析各查詢主題具有的評估功能並加以標示，系統可依據這樣的資訊，判斷處理該查詢主題所需採用的檢索技術，或根據個別需求挑選具有特定功能的查詢主題進行訓練測試，如此可將各種影響變因作較佳的控制以擴展測試集的評估效能。（註63）另外，由於查詢主題的難易度將直接影響系統效益測量的結果，它常是研究者十分希望獲知的訊息。難易度本身是較爲模糊且主觀的概念，影響它的因素亦相當多，訂定一個準確的難易度指標是相當困難的，但吾人仍可經由查詢主題的功能特性、整體相關判斷結果的不一致程度、各參與測試的系統檢索結果的差異情形、文件候選集中相關與不相關文件使用詞彙的

差異程度等各方面加以推斷。

在相關判斷準則方面，目前所使用的測量尺度大多採用類別式尺度，主要原因在於其與連續性或順序性尺度相比，較能展現其客觀的意義以及明確性。然而不論相關類別的多寡，判斷者在將文件歸類時往往有不確定的想法，尤其在選擇相鄰類別時，猶豫情形是十分常見的。所以若僅單純將文件指派給某單一類別，判斷者對決策的確定程度便隱含於其中，這將使判斷結果有所偏差，無法完全反映判斷者對文件與查詢問題間關聯程度的真正認知。爲此，可讓判斷者加入判斷的信心值（即將文件指派於特定類別的確定程度），或允許其同時指派多個相關類別，如此應能反映更詳盡、更可信的相關認知結果，亦可稍微彌補類別尺度方式在表現連續性相關概念時的不足之處。另一方面，利用團體決策的方式進行測試集的相關判斷亦是一個值得嘗試的方案，若能透過共同討論的過程達成認知的一致性，可使判斷的考慮較爲周詳客觀，並能減少個人對查詢主題可能產生錯誤或偏差認知的機會。另外，利用本研究實施相關判斷所產生的各項實驗數據，應能獲得一些有價值之資訊：例如可研究判斷者的背景特性是否會使其與其他判斷結果間造成顯著的差異情形，或是進一步探討判斷者對查詢主題的認知是否有所不同，這些訊息，均可作爲再次選擇判斷者時的參考。

測試集設計建置的最終目的，無非是希望它能具有高度評估效能，因此亟待進行的下一步工作，即是實際運用它進行系統測試，檢驗其評比的功能與效度。進行驗證的重點可包括以下部分：（1）依據測試集所進行的評估結果，是否容易顯現系統間應有的差異？（2）與其他測試集相較，依據本測試集所計算之系統效益評估值（如求準率與求全率）之結果分布，是否類似於一般狀況，不致過高或過低？（3）進行實驗測試不同的相關判斷結果，是否影響系統效益的排序狀況？

（二） 檢索系統評估之探討

測試集必須配合使用一套效益測量方法，始能對系統進行測試與評估。目前系統評估大多以計算求全率與求準率爲主，此測量準則已行之有年，成爲系統間相互比較的標準。但是，其所考量的因素仍然有限，其效益計算方式的本質上也存在著一些重要的缺失。例如，它們僅將文件劃分爲相關與不相關二類，就相關判斷的連續性本質來說，這樣的二分法其實是很不合乎實際情況的。因此，若能考慮相關高低程度之差異，應較符合真實情境中使用者對相關的認定方式。本文基於對每篇文件計算之相關度 (R)，並參考 Reid 等人的想法（註64），建議可採用加權式（Weighted）的效益計算方法：

$$\text{求準率} = \frac{\text{檢索所得文件之相關度總和}}{\text{檢索所得文件之潛在相關度總和}}$$

$$\text{求全率} = \frac{\text{檢索所得文件之相關度總和}}{\text{文件集所有文件之相關度總和}}$$

此公式主要仍依循求全率與求準率原始精神，但以文件相關度作爲主要的運算單位。在求準率公式中，分母「檢索所得文件之潛在相關度總和」可視爲系統對檢索所得文件自行給定的相關分數，其值必須與本測試集所定義的相關度 (R) 一樣介於 0 和 1 之間。若系統並無區分檢索所得文件的相關程度，則可將所有文件均看作是完全相關的（給定相關值爲 1），在此情況下，求準率的分母則可看作是檢索到的總文件數。

除了系統效益計算方法的改進之外，未來對整個評估模式、評估程序、評估項目的設計，也應以多元化的層面考量。例如，近年來許多研究者均倡導互動式系統的評估，以反映真實的檢索現況。由於資訊檢索評估方法對資訊檢索系統之設計與發展方向影響十分深遠，建構一符合現實情況與需求的評估標準，實爲目前的當務之急。

（三）建置一致性的評比環境

多個不同的檢索系統的評估若能在標準的條件與環境之下，運用一致的測試集與效益測量準則進行，將會使評估結果更具意義。例如 TREC 每年所舉行的評估會議，不僅提供了

一個以測試集爲基礎的評估環境，更爲資訊檢索研究者開發一個可供相互討論、經驗交流的開放式論壇，歷年來實施的成果，也確實對檢索系統的發展帶來了深遠的影響。

因此，在目前中文測試集已可獲取的情況下，若能植基於此，進一步建立系統評比的標準環境，並由專職單位統籌規劃相關事宜，使其成爲推動系統評估的常設機制，相信中文資訊檢索系統測試評比的風氣能有顯著的提升，並能加速檢索技術之發展與改進。

註 1： 黃慕萱，「檢索系統評估之發展—理論與實務」，中國圖書館學會學報 59 期 (民國 86 年 12 月)，頁 109。

註 2： Cyril W. Cleverdon, “The Cranfield Tests on Index Language Devices,” Aslib Proceedings 19, no. 6 (1967): 173-194.

註 3： Donna K. Harman, “Evaluation Issues in Information Retrieval,” Information Processing and Management 28, no. 4 (1992): 439.

註 4： Donna K. Harman, “The First Text REtrieval Conference (TREC-1),” Information Processing and Management 29, no. 4 (1993): 411-414.

註 5： K. Kageura and others, eds., “NACSIS Corpus Project for IR and Terminological Research,” In Natural Language Processing Pacific Rim Symposium '97, Phuket, Thailand, December 2-5, 1997, 493.

註 6： “IREX (Information Retrieval and Extraction Exercise) Homepage,” <http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html> (Oct. 31, 1998)

註 7： Alan F. Smeaton and Donna K. Harman, “The TREC Experiments and Their Impact on Europe,” Journal of Information Science 23, no. 2 (1997): 173.

註 8： Ellen M. Voorhees and Donna K. Harman, “Overview of the Fifth Text REtrieval Conference (TREC-5),” In The Fifth Text REtrieval Conference (TREC-5), Gaithersburg, Maryland, November 20-22, 1996, ed. Ellen M. Voorhees and Donna. K. Harman, <http://trec.nist.gov/pubs/trec5/papers/overview.ps> (Aug. 26, 1998)

註 9： TREC 的中文語料主要是以大陸地區新華社與人民日報的新聞文件爲主。

註 10： 本文論及查詢問題、查詢問句與查詢主題此三個在字面上十分相似的詞彙，但其實質含義是有所不同的。查詢問題（Question）所指的是使用者根據其資訊需求（Information Needs or Request）所作出的問題陳述。大部分的測試集均會建構或蒐集一些查詢問題，並以自然語言的方式呈現。所謂的查詢問句（Query）是依據使用者的查詢問題，經由思考、分析、處理後所輸入檢索系統之詞彙或語句。檢索系統在實施測試時，通常會自測試集所提供的查詢問題中，以人工或自動的方法抽取查詢問句進行檢索。查詢主題（Topic）則是 TREC 首先提出之特殊用語，用以表示測試集中的查詢問題，與其他測試集的不同點在於它是以多欄位的方陳述各種不同層次的查詢需求。查詢問題與查詢主題常被混用，但一般通常會形式類似 TREC 的資訊需求陳述稱爲查詢主題。

註 11： 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, 397.

註12： Donna K. Harman, “Panel : Building and Using Test Collections,” In Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, August 18-22, 1996, 337.

註13： Karan Sparck Jones and C. J. van Rijsbergen, “Information Retrieval Test Collections,” Journal of Documentation 32 (1976): 63-73.

註14： Gerard Salton, “A New Comparison between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART),” Journal of the American Society for Information Science 23, no. 1 (1972): 75-84.

註15： Edward A. Fox, “Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts,” (Technical Report TR 83-561, Cornell University: Computing Science Department, 1983), <http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR83-561> (Nov. 30, 1998)

註16： William M. Shaw, Robert Burgin, and Patrick Howell, “Performance Standards and Evaluations in IR Test Collections: Vector-Space and Other Retrieval Models,” Information Processing and Management 33, no. 1 (1997): 15-36. <http://ruby.ils.unc.edu/~howep/perform/hypergeom.html> (Dec. 3, 1998)

註17： 同註 2。

註18： 與測試集有關的效度概念包括樣本效度（Sampling Validity）、內在效度（Internal Validity）、及外在效度（External Validity）。樣本效度所指的樣本足以代表母群體的程度，就文件集來說，可由樣本與母群體二者之間結構的相似性以及母體比例來決定樣本效度的高低。內在效度的意義爲研究處理過程中影響應變項的程度，應用在本研究中，即測試集測試結果是否能反映不同系統效益的優劣。外在效度所指的則是研究發現代表真正現象的程度，在此意謂著測試的結果是否能成功地預測及推論到其他情境中。

註19： David Bawden, User-oriented Evaluation of Information Systems and Services. (Aldershot : Gower, 1990), 87-88.

註20： William Hersh, “OHSUMED: An Interactive Evaluation and New Large Test Collection for Research,” In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, July 3-6, 1994, 192-201.

註21： William M. Shaw, Judith B. Wood, Robert E. Wood, and Helen R. Tibbo, “The Cystic Fibrosis Database: Content and Research Opportunities,” Library and Information Science Research 13 (1991): 347-366.

註22： Tsuyoshi Kitani and others, eds., “BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems,” In Proceedings of IPSJ SIG Notes, DBS-114-3, 1998, 15-22.

註23： Karan Sparck Jones, “The Cranfield Tests,” In Information Retrieval Experiment, ed. Karan Sparck Jones (London; Boston: Butterworths, 1981), 276.

該子測試集包含 200 篇文件及 42 個查詢問題。

註24： Gerard Salton, “The State of Retrieval System Evaluation,” Information Processing & Management 28:4 (1992): 446.

註25： 同註 13，頁 67。

註26： Ellen M. Voorhees and Donna K. Harman, “Overview of the Seventh Text REtrieval Conference (TREC-7),” In The Seventh Text REtrieval Conference (TREC-7), Gaithersburg, Maryland, November 9-11, 1998, edited by Ellen M. Voorhees and Donna K. Harman, <http://trec.nist.gov/pubs/trec7/papers/overview_7.ps> (June. 6, 1999)

註27： Tsuyoshi Kitani and others, eds., “Lessons form BMIR-J2: A Test Collection for Japanese IR Systems,” In Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24-28 August 1998, 345-346.

註28： AMARYLLIS 參考各系統送回的檢索結果，對先前建立的相關判斷進行修正。若某文件在原始相關判斷中雖不被視相關，但在測試時被一半以上的系統檢索出來，或是在每個系統送回的前 10 篇文件之中，則將其修正爲相關文件；若某文件在原始相關判斷中被視爲相關，但在各系統送回的檢索結果中均未出現，則修正爲不相關文件。

註29： 此表參考相關文獻彙整分析而成，但由於各文獻所載之數據資料稍有出入，表中所列僅爲其近似值。字數的計算在英文及法文中是詞（Term）爲單位，在日文中則以字元(Character)爲單位。另外，在查詢問題的平均相關文件數部分，有些測試集因相關判斷尺度的不同而有多個數值。表中的 N/A (Not Available)表示該項資料未能獲取。

註30： 同註 19。

註31： 同註 13，頁 60-61。

註32： 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, 397.

註33： Karen Sparck Jones, “Reflections on TREC,” Information Processing and Managements 31, no. 3 (1995): 310。

註34： Pia Borlund and Peter Ingwersen, “The Development of a Method for the Evaluation of Interactive Information Retrieval Systems,” Journal of Documentation 53, no. 3 (1997): 226.

註35： Gerard Salton, “The State of Retrieval System Evaluation,” Information Processing & Management 28:4 (1992): 443.

註36： 同註 34。

註37： Robert N. Oddy, “Laboratory Tests: Automatic Systems,” In Information Retrieval Experiment, ed. Karan Sparck Jones (London; Boston: Butterworths, 1981), 161.

註38： Don R.Swanson, “Historical Note: Information Retrieval and the Future of an Illusion,” Journal of the American Society for Information Science 39, no. 2 (1988): 95.

註39： 例如 Borlund 與 Ingwersen 建構包含需求情境的查詢主題以測試互動式的檢索系統，NTCIR 的查詢主題中亦納入了一些如檢索目的、檢索背景等項目，使系統測試能考量

更多的層面，更符合真實的檢索情況。

註40： Michael B. Eisenberg and X Hu, "Dichotomous Relevance Judgments and the Evaluation of Information Systems," In Proceedings of the 50th Annual Meeting of the American Society for Information Science, 24, 1987, 66-70.

註41： Joseph W. Janes, "The Binary Nature of Continuous Relevance Judgements: A Study of Users' Perceptions," Journal of the American Society for Information Science 42, no. 10 (1991): 754-756.

註42： Tefko Saracevic, "The Concept of 'Relevance' in Information Science: A Historical Review," In Introduction to Information Science, ed. Tefko Saracevic (N. Y.: Bowker, 1970), 120.

註43： Jane Reid and Stefano Mizzaro, "On the Consensus between Relevance Judges in a Multi-media Context," In Proceedings of the 6th Mira Workshop, Dublin, October 28-30, 1998, <http://www.dcs.gla.ac.uk/mira/workshops/dublin/procs/mr.pdf> (Nov. 5, 1998)

註44： Tefko Saracevic, "Relevance: A Review of and a Framwork for the Thinking on the Notion in Information Science," Journal of the American Society for Information Science 26 (1975): 341-342.

註45： 同註 43。

註46： Ellen M. Voorhees, "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness," In Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24-28 August 1998, 315-323.

註47： Robert Burgin, "Variations in Relevance Judgments and the Evaluation of Retrieval Performance," Information Processing and Management 28, no. 5 (1992): 619-627.

註48： Stephen P. Harter, "Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness," Journal of American Society for Information Science 47, no. 1 (1996): 40.

註49： 同上註，頁 37-49。

註50： Precision 與 Recall 在轉譯爲中文的用詞上，一直頗爲紛雜，沒有適當且統一的選擇，如 Recall 譯爲回現率、回收率、再現率，Precision 譯爲精確率、準確率等。大陸學者王崇德先生將其譯爲「查全率」與「查準率」，在字面上能頗爲適當地反映 Precision 與 Recall 所代表的意涵。然而，「查」字意指文件的檢索，使得它在其他方面的應用性較爲侷限（如在中文斷詞結果的評估中便不能使用）。故在此建議以「求準率」及「求全率」表示 Precision 與 Recall。

註51： 在統計的意義上，可信度是指實驗結果的一致性（Consistencies）或穩定性（Stability），亦可稱爲可靠性（Trustworthiness）。

註52： Peter Wallis and James A. Thom, "Relevance Judgements for Assessing Recall," Information Processing and Management 32, no. 3 (1996): 273-286.

註53： 早期 Cranfield II 實驗採用逐一比對的方式進行了十分詳盡的相關判斷，經 Harter 的再次檢驗後，推斷它仍可能遺漏了七千多篇相關文件。TREC 曾對其相關判斷的完整性作了一個實驗性評估，發現若將 Pool 之大小設爲 100，平均每個查詢主題會遺漏約 1 篇

真正相關的文件。

註54： 同註 32，頁 307-314。

註55： 本研究文件下載的工作自 1998 年 5 月 11 日至 1999 年 5 月 10 日止，共約一年的時間，

註56： 李卓偉，統計學 (台北市：智勝文化，民國 82 年)，頁 6-59~6-60。

註57： 木本晴夫 ，「情報 索 評 用 － － － 構築 提案」。情 研報 FI-32-1 (1993)：4。

註58： 假設樣本數量爲 n，某一查詢主題的相關文件數量與母體之比例爲 p，就區間估計的概念，若欲使估計之 p 值與真實 p 值之誤差不大於 b，且有 (1-α) 的信賴水準，則應取樣本數 $n \approx \frac{[Z(\alpha/2)]^2 \cdot p(1-p)}{b^2}$。但是，由於我們並無法得知 p 的先驗預估值，則由不等式 $p(1-p) = -p^2 + p = -(p-\frac{1}{2})^2 + \frac{1}{4} \le \frac{1}{4}$ 可得出一個比較保守的應取樣本數 $n \cong \frac{[Z(\alpha/2)]^2}{4b^2}$。

註59： 相關文件候選集之數量共有近 5,000 篇，而每篇文件被判斷 3 次，因此相關判斷的總次數約爲 15,000 次。

註60： Kappa 一致性係數（Kappa Coefficient of Agreement, K）是屬於無母數統計的範疇，適用於類別尺度變數，主要目的是探討不同測量者對一組不同物件分類結果的一致狀況。應用在本研究中，即爲判斷者在同一查詢主題的候選文件集中相關判斷結果之一致性。Kappa 一致性統計量的形成有一基本假設：判斷者在有意識的情況下所進行的判斷，其一致性不應低於隨機指派的結果。公式主要是計算判斷者實際判斷一致的次數比例 P (A)，與判斷者可能達成的最大一致比例（定義爲 1）之間的比值，其中並以預期在隨機指派可能形成的一致比例 P (E) 進行校正：$K = \frac{P(A) - P(E)}{1 - P(E)}$。若文件數量較大，隨機指派所得到的 K 值應會接近標準常態分布，因此透過 K 之變異數，我們可以算出常態分布之統計值 z（$z = \frac{K}{\sqrt{\text{var}(K)}}$），進一步檢測其顯著之一致性。

註61： Kendall 一致性係數（The Kendall Coefficient of Concordance, W）是一種衡量多種關聯變數之間一致性的度量方法，它主要考慮的是變數之間順序關係的強度，亦即檢定不同判斷者判斷結果之間是否具有某一順序的關聯性，其統計量以 W 表示。此檢定屬於無母數統計的範疇，適用於順序尺度以上的資料。進行 Kendall 檢定必須先將判斷結果加以排序，並給予等級。就本研究來說，雖然判斷者並非對所有文件的相關程度進行排序，但由於這些類別隱含著順序尺度的意義，我們可將判斷者給予每篇文件的相關分數加以排序，產生候選文件集中的等級變數，但基本前提是同一判斷者對相關分數的給定標準必須前後一致，如此形成的順序等級才是具有意義的。Kendall 檢定的統計原理是，若不同判斷者的判斷結果彼此之間並無關聯，那麼其等級應具隨機性，其總和應是接近的；反之，若彼此之間有關聯，其總和則會有明顯的區別。換句話說 Kendall 統計量主要是定義 k 種實際等級總和之變異，與完全具一致性時之等級總和之變異的

比值（在本研究中 k=3），計算公式爲：$W=\frac{12\sum \overline{R_i}^2-3N(N+1)^2}{N(N^2-1)-(\sum T_j)/k}$， R_i 爲等級變數，N 爲排序物件的個數。計算出的 W 值介於 0 於 1 之間，若 W 愈趨近 1，表示不同判斷結果的一致性愈高，愈趨近於 0 則表示一致性愈低。此外，我們尙可求出自由度爲 N-1 之卡方值$\chi^2 = k\,(N-1)\,W$，檢測此統計量的顯著性。

註62： 一致度 (*C*)主要是考量判斷結果在相關意義上的接近程度。根據此想法，我們計算每一文件兩兩判斷結果間相關分數距離的總和，並除以可能的最大距離（在本研究中此值爲 6），表示其不一致度，最後的一致度 (*C*) 則以 1 減去不一致度，計算公式爲：

$$C=1-\frac{|X_A-X_B|+|X_B-X_C|+|X_C-X_A|}{6}$$

註63： 日本近年來發展的 BMIR-J1 及 J2 測試集，在查詢主題中標示基本、數字範圍、語法、語意及詞彙知識五種類別，即初步揭示了查詢主題所具備的測試功能。

註64： Jane Reid and Stefano Mizzaro, “On the Consensus between Relevance Judges in a Multi-media Context,” In Proceedings of the 6th Mira Workshop, Dublin, October 28-30, 1998, <http://www.dcs.gla.ac.uk/mira/workshops/dublin/procs/mr.pdf> (Nov. 5, 1998)