圖書與資訊學刊
29(May '99)36-59
ISSN 1023-2125

TREC 現況及其對資訊檢索研究之影響

The TREC and Its Impact on IR Researches

江 玉 婷　陳 光 華
Yu-ting Chiang　Kuang-hua Chen

國立臺灣大學圖書資訊學系
Department of Library and Information Science
National Taiwan University
E-mail : ytchiang@steelman.lis.ntu.edu.tw
khchen@steelman.lis.ntu.edu.tw

【摘要　Abstract】

測試集為一在規範化環境中測量資訊檢索效益的機制，研究者常藉著它來改進資訊檢索系統。然而，過去所使用的測試集規模通常很小，無法有效模擬真實的資訊檢索環境。自 1992 年起，一個稱為 TREC（Text REtrieval Conference）的評估會議每年在美國舉行。它是一個大型的實驗室測試機制，可以比較不同檢索系統的效能，並提供一個論壇讓參與者相互切磋討論。今日，TREC 在資訊檢索領域中已成為備受矚目的標準測試環境。本文試圖對 TREC 作一個概覽性的介紹，包括它所使用的測試集、測量準則、測試項目及過去的發展概況。另外，也透過一些相關文獻的分析，對 TREC 在資訊檢索評估領域中的價值及所扮演的角色作一簡要的評述。

Test Collection is a mechanism which is used to explore relative benefits of different retrieval system in a normalized environment, and researchers traditionally use it to improve their retrieval systems. However, for many cases in the past, the scale of Test Collection was quite small, thus it can not simulate the real state of information retrieval task. Since 1992, a series of annual benchmarking evaluation exercises, called TREC (Text REtrieval Conference), have launched in the USA. TREC experiments were designed to allow large-scale laboratory testing, compare the effectiveness and performance of different information retrieval techniques, and provide a forum for research groups to discuss their work together. Today, TREC has become the major experimental effort in the IR field. In this paper, we present an overview of TREC, the test collection, the test measures, the main tasks and specific 'tracks', and the development during the past six years. Besides, we also briefly review the performance and the evaluation of IR.

關鍵詞　Keyword

評估　資訊檢索　測試集　TREC

Evaluation; Information retrieval; Test collection; Text REtrieval Conference

壹、前　言

在資訊檢索的領域中，檢索系統評估對於系統的研究、設計與發展一直有其顯著的影響力。早期對檢索系統評估最著名的研究是 Cleverdon 在 1950 年代末期開始進行的 Cranfield 計劃，它開創了以測試集（Test Collection）配合測量準則（Measures）來評估系統的模式。所謂測試集，是一在規範化環境中測試系統效能的機制，包括測試問題（Queries）①、測試文件集（Document Set）、及相關判斷（Relevance Assessment）②等三個部分。其研究設計的概念是假設在給定的查詢問句與文件集中，某些文件是與查詢問句相關的。系統的目的是檢索出相關的文件，並拒絕不相關的文件，因此採用回收率（Recall）及精確率（Precision）作爲測量準則。Cranfield 研究首開規範化系統評估之先河，其評估模式亦成爲後世普遍採用的標準。③

然而，由於早期所採用的測試集規模均不大，且大多數是使用同質性較高的文件集（例如，Cranfield 的第二期研究只包含了 331 個查詢問句及 1400 篇文件④，而文件集中的文件長度亦十分相似），因此它與真實的檢索環境之間存在著相當大的差異。植基於這樣的測試集所發展出來的檢索系統，在實際運作時往往受到極大的限制，成效並不佳。⑤不過，要建構大型的測試集必須耗費相當可觀的人力、時間與經費，這對於大多數從事資訊檢索研究的單位來說是不堪負荷的，因此長久以來，大規模的測試集一直付之闕如。

爲了促進資訊檢索的研究與應用的發展，美國國防部高等研究計劃局（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）。透過所發展出的大型測試集，制定各種測試項目、測試程序及測量準則，組合成一評估檢索系統的機制。TREC 在 1992 年舉辦了第一屆，其後持續在每年年底舉辦會議，至今已進行了七屆。除了與會者依據大會提供的測試集送回各測試項目的資料以進行評比之外，尚有一爲期三天的研討會，與會者可以在會中發表系統的架構、評估結果，並相互討論切磋。

TREC 的主要目標有下：⑥

一、以大型測試集爲基礎，鼓勵文件檢索的研究。

二、經由開放式的論壇，使與會者能交換研究的成果與心得，以增進學術界、產業界與政府的交流互通。

三、經由對真實檢索環境的模擬與實質的論證，加速將實驗室研究技術轉移爲可運作的系統。

四、發展適當且具應用性的評估技術，供各界遵循採用。

五、發展多種不同的測試項目，希望能在一致的模式中對各種檢索技術進行評估。

本文希望能對 TREC 作一個概覽性的介紹，包括所使用的測試集、測量準則、測試項目及過去的發展概況。另外，也透過一些相關文獻的分析，對 TREC 在資訊檢索評估領域中的價值及所扮演的角色作一簡要的評述。

貳、TREC 測試集

TREC 的評估機制基本上是依據 Cranfield 研究的概念擴展而來，因此其測試集亦包含文件集、主題及相關判斷三個主要部分，分別介紹如下：

一、文件集（Document Set）

TREC 文件集所收錄的主要是新聞性文件及雜誌期刊。⑦它們分散儲存在數片光碟中，每片

約含爲 1GB，目前 TREC 的文件集已儲存了有五片光碟之多⑧，約近二百萬篇文件，這樣的文件數量，與早期僅含數千至數萬篇文件的小型文件集相比，的確有十分顯著的進步，除了文件集的規模之外，文件的異質性亦爲一大特色，尤其在文件的長度方面，雖然大多介於 300 至 400 字之間，但也有些多達數百頁，⑨茲將 TREC 使用過的文件集整理如表一：⑩

表一：TREC 文件集

Volume	Revised	Sources	Size (MB)	Docs	Median # Mean #	Terms/Doc Terms/Doc
1	March 1994	Wall Street Journal, 1978-1989	267	98,732	245	434.0
		Associated Press newswire, 1989	254	84,678	446	473.9
		Computer Selects Articles, Ziff-Davis	242	75,180	200	473.0
		Federal Register, 1989	260	25,960	391	1315.9
		Abstracts of U.S. DOE publications	184	226,087	111	120.4
2	March 1994	Wall Street Journal, 1990-1992（WSJ）	242	74,520	301	508.4
		Associated Press newswire（1988）（AP）	237	79,919	438	468.7
		Computer Selects articles, Ziff-Davis(ZIFF)	175	56,920	182	451.9
		Federal Register（1988）（FR88）	209	19,860	396	1378.1
3	March 1994	San Jose Mercury News, 1991	287	90,257	379	453.0
		Associated Press newswire, 1990	237	78,321	451	478.4
		Computer Selects articles, Ziff-Davis	345	161,021	122	295.4
		U.S. patents, 1993	243	6,711	4445	5391.0
4	May 1996	The Financial Times, 1991-1994（FT）	564	210,158	316	412.7
		Federal Register, 1994（FR94）	395	55,630	588	644.7
		Congressional Record, 1993（CR）	235	27,922	288	1373.5
5	April 1997	Foreign Broadcast Information Service (FBIS)	470	130,471	322	543.6
		Los Angeles Times （1989, 1990）	475	131,896	351	526.5
Routing Test Data		Foreign Broadcast Information Service (FBIS)	490	120,653	348	581.3

資料來源：Ellen M. Voorhees and Donna K. Harman, "Overview of the Sixth Text REtrieval Conference（TREC-6）," in The Sixth Text REtrieval Conference（TREC-6）held in Gaithersburg, Maryland, November 19-21, 1997, ed. Ellen M. Voorhees and Donna K. Harman, <http://trec.nist.gov/pubs/trec6/t6_proceedings.html>

這些文件均依據標準通用標誌語言（Standard Generalized Mark-Up Language，簡稱 SGML）及資料型態定義檔，簡稱 DTD（Data-Type-Definition）檔加上標籤（Tag），以便系統進行簡易的剖析（Parsing）工作，如圖一所示⑪。NIST 在進行文件標記時的準則是，在能將文件作一致性解碼的前提下，儘可能地保留其原始的結構，因此不同資料庫中的文件標記，在細節上可能有些許不同，但也由於儘量保持文件原貌的原則，且利用人工逐一檢查文件內容，在實施上有其困難性，因此文件中存在著許多訛誤。NIST 採用自動機制檢查一些控制字元、特殊符號等部分，但

避免在其內容上作更動（如拼字錯誤即忽略），⑩

二、主題（Topics）

TREC 不同於一般的測試集採用傳統的查詢問句作測試，而是模擬使用者的資訊需求，以各種形式、各種角度陳述出來，並以結構化的欄位來呈現，稱之爲主題（Topics）。TREC-1 及 TREC-2 共有 150 個主題，之後 TREC 每屆均建構 50 個新的主題，將之作循序編號，以便於利用辨識。至 TREC-6 爲止，已有 350 個不同的主題。

㈠主題的組成結構

每屆 TREC 會根據先前的測試結果或當時特別想探究的問題，改變主題的結構與特性，以期發揮主題測試的最佳效能。

```
<DOC>
<DOCNO>FT911-3</DOCNO>
<PROFILE>AN-BE0A7AAIFT</PROFILE>
<DATE>910514
</DATE>
<HEADLINE>
FT 14 MAY 91/International Company News : Contigas plans DM900m east German project
</HEADLINE>
<BYLINE>
By DAVID GOODHART
</BYLINE>
<DATELINE>
BONN
</DATELINE>
<TEXT>
CONTIGAS, the German gas group 81 percent owned by the utility Bayernwerk, said yesterday that it intends to invest DM900m (Dollars 522m) in the next years to build a new gas distribution system in the east German state of Thuringia. ...
</TEXT>
</DOC>
```

圖一：TREC 之文件標示

資料來源：Ellen M. Voorhees and Donna K. Harman, "Overview of the Sixth Text REtrieval Conference (TREC-6)," in The Sixth Text REtrieval Conference (TREC-6) held in Gaithersburg, Maryland, November 19-21, 1997, ed. Ellen M. Voorhees and Donna K. Harman. <http://trec.nist.gov/pubs/trec6/t6_proceedings.html>

TREC-1 與 TREC-2 所使用的主題（Topic 1-150）包含了複雜且詳細的欄位，每一欄位都從不同的功能觀點來陳述資訊需求，除了自然語言式的長、短敘述（Description, Narrative）之外，亦有相關概念（Concepts）、特殊名詞定義（Definition）等，如圖二所示。TREC 對於使用者資訊需求的多樣化呈現方式，鼓勵了研究者探討如何分析資訊需求，選擇、結合主題中各欄位，並從中擷取有意義的資訊。⑪⑫如 Callan 與 Croft 所做的實驗顯示，對於一些具有較長需求陳述的欄位（如 Narrative），需要經過複雜的處理程序才能達到較好的效果，若不當的擷取，不如

完全將它忽略。[9]

但是，由於 TREC 主題產生的方式是先以某資訊需求爲主線（如 Narrative 欄位的內容），再依據其在文件集中檢索所得的相關文件，擴展其他的欄位，因此一般認爲這樣的主題多少會受到文件集特性的影響。[10]

<top>
<head> Tipster Topic Description
<num> Number: 037
<dom> Domain: Science and Technology
<title> Topic: Identify SAA components
<desc> Description:
Document identifies software products which adhere to IBM's SAA standards.
<narr> Narrative:
To be relevant, a document must identify a piece of software which is
considered a Systems Application Architectural （SAA） component or one which
conforms to SAA.
<con> Concept（s）:
1.SAA
2.OfficeVision
3.IBM
4.Standards, Interfaces, Compatibility
<fac> Factor（s）:
<def> Definition（s）:
OfficeVision - A series of integrated office automation applications from
IBM that runs across all of its major computer families.
Systems Application Architecture （SAA） - A set of IBM standards that provide
consistent user interfaces, programming interfaces, and communications
protocols among all IBM computers from micro to mainframe.
</top>

圖二：TREC-1 與 TREC-2 之主題

資料來源："TREC-1 routing topics," <URL: http://trec.nist.gov/data/topics/trec6/topics.1-50.txt>

由於 TREC-1 與 TREC-2 測試主題的結構過於複雜，因此 TREC-3 對其作了一些調整。首先，主題的長度明顯地較之前短，而在欄位的選擇上亦有更動，將呈現資訊需求的欄位刪減爲 Topic、Description 以及 Narrative 三項，如圖三所示[11]。但是，與會者認爲這樣的需求陳述，與使用者一般在檢索時所給予系統的相比，還是太長太複雜了。因此，TREC-4 將主題縮得更短，欄位也作更大幅度的簡化，只留下 Description 欄，希望能藉此更真實地模擬實際環境中的情況，如圖四所示[12]。然而，經 TREC-4 測試之後發現，如此短的測試主題並無法達到預期的效益，並導致了一

<top>
<num> Number: 177
<title> Topic: English as the Official Language in U.S.
<desc> Description:
Document will provide arguments supporting the making of English the standard language of the U.S.
<narr> Narrative:
A relevant document will note instances in which English is favored as a standard language. Examples are the positive results achieved by immigrants in the areas of acceptance, greater economic opportunity, and increased academic achievement. Reports are also desired which describe some of the language difficulties encountered by other nations and groups of nations, e.g., Canada, Belgium, European Community, when they have opted for the use of two or more languages as their official means of communication.
Not relevant are reports which promote bilingualism or multilingualism.
</top>

圖三：TREC-3 之主題

資料來源："TREC-3 adhoc topics," <http://trec.nist.gov/data/topics/trec6/topics.151-200.txt>

些處理上的困難。因此，TREC-5 又將主題調整回與 TREC-3 主題相似的結構，但其平均長度較 TREC-3 爲短。[20] TREC-6 的主題結構及長度大致上則與 TREC-5 相去不遠，表二列出了 350 個主題的主要欄位及長度[21]。

㈢主題的建構方式

雖然 TREC 主題的內容是有關使用者資訊需求的陳述，但是它是以模擬的方式建立的並非實際搜集而來，爲了使每次建構出的 50 個主題在描述方式及詞彙運用等方面能有某種程度的一致性，自 TREC-2 開始，每屆的測試主題均由一至二人建構發展。

爲了使主題難易適中，且能反映真正的資訊需求，TREC 設立了一個特殊的篩選程序，透過

<top>
<num> Number: 217
<desc> Description:
Reporting on possibility of and search for extra-terrestrial life/intelligence.
</top>

圖四：TREC-4 之主題

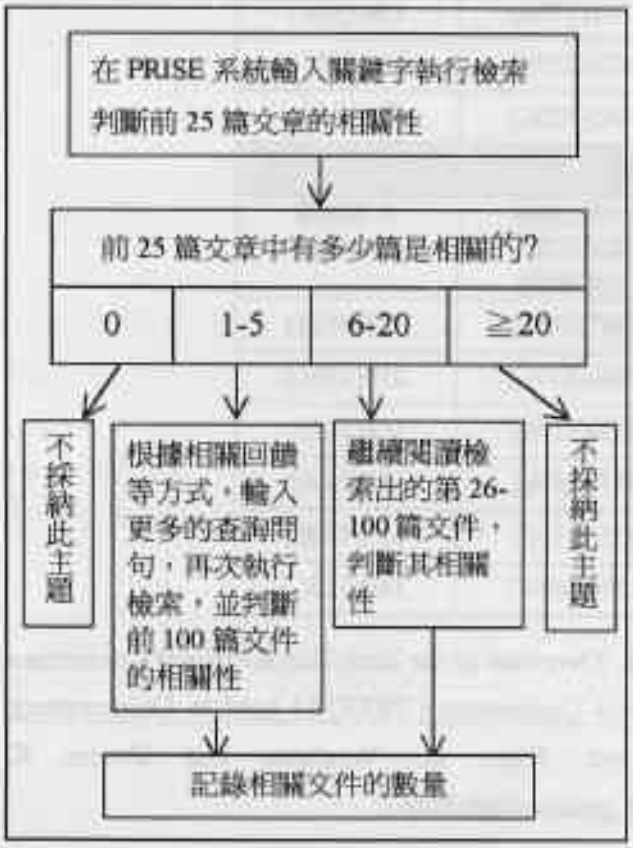
資料來源："TREC-4 adhoc topics," <http://trec.nist.gov/data/topics/trec6/topics.201-250.txt>

表二：主題各欄位長度比較

	欄　位	字數（包含停字）		
		最小字數	最大字數	平均字數
TREC-1 (51-100)	Total	44	250	107.4
	Title	1	11	3.8
	Description	5	41	17.9
	Narrative	23	209	64.5
	Concepts	4	111	21.2
TREC-2 (101-150)	Total	54	231	130.8
	Title	2	9	4.9
	Description	6	41	18.7
	Narrative	27	165	78.8
	Concepts	3	88	28.5
TREC-3 (151-200)	Total	49	180	103.4
	Title	2	20	6.5
	Description	9	42	22.3
	Narrative	26	146	74.6
TREC-4 (201-250)	Total	8	33	16.3
	Description	8	33	16.3
TREC-5 (251-300)	Total	29	213	82.7
	Title	2	10	3.8
	Description	6	40	15.7
	Narrative	19	168	63.2
TREC-6 (301-350)	Total	47	156	88.4
	Title	1	5	2.7
	Description	5	62	20.4
	Narrative	17	142	65.3

資料來源：Ellen M. Voorhees and Donna K. Harman, "Overview of the Sixth Text REtrieval Conference （TREC-6）," in The Sixth Text REtrieval Conference （TREC-6）held in Gaithersburg, Maryland, November 19-21, 1997, ed. Ellen M. Voorhees and Donna K. Harman, <http://trec.nist.gov/pubs/trec6/t6_proceedings. html>

系統的初步檢索結果，預測其在文件集中可能的相關文件數量。主題建構者首先根據目標文件集所涵蓋的主題範圍，模擬約 100 個描述使用者資訊需求的候選主題，利用 NIST 的 PRISE 系統在對應的文件集中檢索，再依據檢索出文件的數量，篩選出 50 個最後的主題。一般而言，若檢索出的文件太多、太少或是含意較模糊、難以判斷者，將會被刪除。利用此法選出後，評估者會再次檢視每個主題，若有不符合建構準則或要求者，再加以修改。必須一提的是，此處所進行的檢索僅作爲篩選主題的參考，與之後進行的相關判斷無關。圖五爲 TREC-6 的主題篩選程序㉘。



圖五：TREC-6 之主題篩選程序

資料來源："Topic creation," <http://trec.nist.gov/presentations/TREC6/11.html>

㈢主題的難易度

爲了使主題充分發揮其測試的功能，並對其作良好的控制，建構者均嘗試在事前對主題的難易度作預測，然而，這是一個頗爲困難的工作。NIST 即針對這方面做了一項實驗，探討建構者對主題難易度評估的有效性。首先在主題對文件集的相關判斷還未進行之前，請九位人員預測 50 個主題的難易度，將其分爲難、中、易三群。當正式的測試結果及相關判斷產生出來後，則依據其平均精確率進行難易度評量，視其爲該主題真正的難易度。二者經相關係數的計算之後，預測與實際的主題難易度呈現相當不一致的狀況。此實驗顯示了主題的難易度評估工作的確十分困難，而目前也尚未找到形成其難易度的確實原因。因此，要如何建立難易適中的主題，仍是有待研究的。㉙

三、相關判斷

相關判斷在測試集的建構過程中是最困難、最花時間的，但它卻是最重要的一部分。早期的 Cranfield 研究規模雖不大，但其相關判斷卻進行了 50 餘萬次。對於 TREC 這樣的大型測試集來說，要將每個主題在文件集中進行詳盡的相關判斷，所須耗費的工程可見一斑。

TREC 的相關判斷主要是根據主題的 Narrative 欄位進行，對相關與否的判斷原則，是只要文件部分與主題相關即可（即使只是數句），並不要求文件的每個部分均與主題相關。㉚

在如此大的測試集中，若要將每個主題與每篇文件逐一作相關判斷，相當不可行。因此，TREC 採用了 pooling 的方式進行，亦即針對每個主題，從各系統所送回的測試結果中，抽取出一定數量的文件（通常爲 100 篇），合併

形成一個 pool，將之視爲該主題可能的相關文件集合，將此 pool 中重覆的文件去除後，再給該主題的原始建構者進行相關判斷，平均來說，在集合中真正的文件數量約爲各系統送回總數的 20%[3]。此作法是意圖利用不同的系統，不同的檢索技術，輔助縮小相關判斷的範圍，以減少判斷者的負荷，另外，由於參與測試的系統均提供相關排序的輸出（排在前面的爲較相關的文獻），因此 pooling 技術在此環境中得以展現其功效。

表三：Pooling 與實際相關文件對照表

Adhoc			
	各系統送至 Pool 內之文件總數	Pool 中實際的文件數（去除重覆）	實際相關文件數
TREC-1	8800	1279(39%)	277(22%)
TREC-2	4000	1106(28%)	210(19%)
TREC-3	2700	1005(37%)	146(15%)
TREC-4	7300	1711(24%)	130 (8%)
TREC-5	10100	2671(27%)	110 (4%)
TREC-6	8480	1445(42%)	92(6.4%)
Routing			
	各系統送至 Pool 內之文件總數	檢索出的相關文件數 (去除重覆)	實際相關文件數
TREC-1	2200	1067(49%)	371(35%)
TREC-2	4000	1466(37%)	210(14%)
TREC-3	2300	703 (31%)	146(21%)
TREC-4	3800	957 (25%)	132(14%)
TREC-5	3100	955 (31%)	113(12%)
TREC-6	4400	1306(30%)	140(11%)

資料來源：Ellen M. Voorhees and Donna K. Harman, "Overview of the Sixth Text REtrieval Conference（TREC-6）," in The Sixth Text REtrieval Conference（TREC-6）held in Gaithersburg, Maryland, November 19-21, 1997, ed. Ellen M. Voorhees and Donna K. Harman.<http://trec.nist.gov/pubs/trec6/t6_proceedings.html）

Pooling 技術的有效性可以由各系統間檢索結果的重覆（overlap）程度來衡量。表三列出了 TREC-1 至 TREC-6 進行相關判斷時，pool 的大小、重覆的情形以及相關文件數量的比較㉖。從表中可以看出，主題的平均相關文件數正逐年地下降，而真正與主題相關的文件在 pool 所佔的百分比亦逐漸下降。

由於相關判斷在測試集中扮演極重要的角色，所以對於相關判斷品質的評估也是不可忽視的課題，其中完整性（Completeness）及一致性（Consistency）常是研究者感興趣的二項指標。

相關判斷的完整性是指主題在文件集中所有真正相關的文件被判斷爲相關的情形。TREC 針對 TREC-2 及 TREC-3 相關判斷完整性的結果作了一個實驗性評估，將各參與系統所送回結果的第 101-200 篇文件集合起來，形成一個新的 pool，再據之進行相關判斷。結果在每個測試項目中，平均發現一篇新的相關文件，也就是說，若將 pool 的大小設爲 100，平均每個主題會遺漏約一篇真正相關的文件。對於具有較多相關文件的主題來說，可能遺漏的相關文件也愈多，因此 TREC 希望建構一些具有較少相關文件的主題，以減少真正相關文件的遺漏數量，此想法可以從表三中看出一些端倪。㉖

一致性是指不同人員在進行相關判斷時，對相關及不相關文件認知的差異程度而言。TREC 爲探討其相關判斷的一致性，進行了多次的實驗。在 TREC-3 的一致性實驗中，研究者發現二組不同的相關判斷一致性高達 80%，這可能是由於實驗的主題本身比較明確的緣故。㉘ TREC-4 所進行的實驗，則是利用原本被判斷爲相關的文件及不相關的文件所形成的二個 pool，給予另外二位人員分別再次進行相關判斷，結果顯示二組判斷中有 71%不一致狀況；有 13.3%在相關的認知上不一致，有 58.4%則是在不相關認知上產生差異。其中，有 30%的文件原本被評定爲相關，而被其後的二位人員評定爲不相關，至於原先判定爲不相關的文件被評定爲相關的比例，則不到 3%。㉙

相關判斷在測試集中一直是較受爭議的部分，因爲相關原本就是主觀的概念，而相關判斷更會因判斷者、判斷情境、判斷時間等因素而可能產生很大的差異。在如此複雜易變的情況下，如何能對系統作有效的測試呢？對於相關判斷的差異所可能造成測試結果偏頗的問題，已有不少的文獻探討，如 Harter 與 Sweason 均對於 Cranfield II 相關判斷的過程與結果提出質疑，推斷其共遺漏了七千多篇相關文件，這無疑地對回收率產生相當大的影響，亦可能會使精確率改變。㉚

TREC 對於這個問題所持的態度是，測試集的重要功能在於增進資訊檢索的研究以及系統效能的改進，其目的並不是要獲致一個絕對的系統效益測量值，因此相關判斷的重點並不在於是否具有一致性，重要的是它是否能正確地反映不同檢索技術的相對價值。有關此一想法，在 TREC 之前已有一些研究對於早期的小型測試集進行實驗，包括 Burgin（1992）㉛、Kazhdan（1979）、Cleverdon（1970）及 Lesk & Salton（1968），其目的均在驗證不同的相關判斷在測試系統間相對效益時所產生的影響。這些研究利用相似的方法，針對各種不同型態的相關判斷進行測試，以回收率及精確率評估測試結果。他們歸結出相似的結論：相關判斷的差異並不會影響系統績效優劣排序的穩定度。㉜不過，由於它們均是以小規模的測試集作爲驗證對象，並不能任意對大型測試集作相同的推論。

TREC 爲了證實在大型測試集中，相關判

斷在測試時的有效性，也進行與上述相似的實驗，結果顯示，在四組不同的相關判斷中，所產生的系統效益的排序雖非完全一致，但仍具有相當程度的穩定性。㉘

從另一角度看，由於不同使用者所產生相關判斷的結果通常有著很大的差異，因此檢索系統在設計時對於不同使用者相關判斷異動的反應最好是遲鈍的㉙，也就是說，系統執行的效益不應受到使用者間相關判斷差異的影響。

參、系統評量準則㉚

TREC 主要是以回收率及精確率來作爲系統評量準則㉛：

$$\text{回收率} = \frac{\text{檢索到的相關文件數}}{\text{所有的相關文件數}}$$

$$\text{精確率} = \frac{\text{檢索到的相關文件數}}{\text{檢索到的文件數}}$$

參與測試的系統依據此二準則產生的測試結果，通常以下列幾種圖表呈現：

一、摘要統計表（Summary Statistics Table）

㈠測試項目的名稱、測試主題的數量。

㈡送回 TREC 的文件數、主題相關的文件數以及檢索出爲相關的文件數。

二、回收率與精確率對應表（Recall Level Precision Averages Table）

㈠ 11 點特定的回收率（0, 0.1, 0.2,……1.0）所對應的平均精確率（11-Point Precision）。

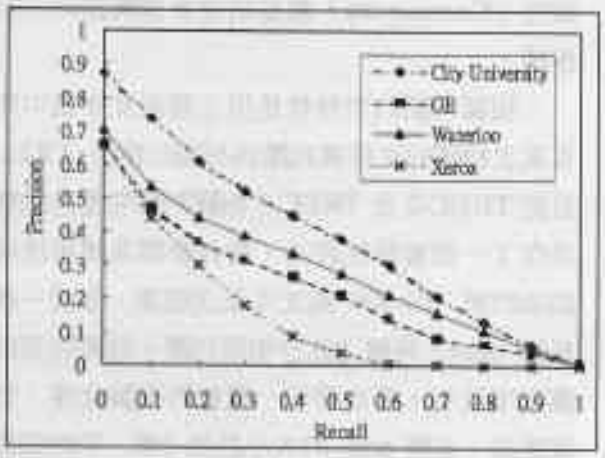
㈡所有相關文件檢索出來時的平均精確率。

三、文件數與精確率對應表（Document Level Averages Table）

㈠在 9 個特定文件數量被檢索出來時，所對應的精確率。

㈡檢索出R篇文件時的精確率。R等於該主題中所有相關文件的數量（R-Precision）。

四、回收率／精確率圖（Recall／Precision Graph）：此爲系統間比較時最常用到的對照圖，利用 11 點回收率繪製。各系統的檢索結果可以畫在同一圖中進行比較，如圖六所示㉜。



圖六：回收率與精確率對應圖

五、平均精確率長條圖（Average Precision Histogram）：以同一測試項目中各系統平均精確率的中位數爲基準，記錄個別系統在每個主題中的相對效益。

肆、測試程序與測試項目

一、主要測試項目（Main Tasks）

TREC 的核心測試項目爲 Routing Task 與 Adhoc Task，茲將其運作程序分述如下：

㈠ Routing Task ㉝

此測試項目假設使用者總是問相同的問

題，而不斷有新的文件加入檢索，其概念類似圖書館中的資料選粹服務（SDI），依據使用者的需求檔（Profile）進行新進資訊的搜尋及過濾。爲此，TREC 使用已做好相關判斷的舊主題，在一個全新的文件集中進行測試，主要目的爲探討利用已知的測試主題檢索新文件的能力。

在正式進行 Routing Task 測試之前，TREC 先給予參與的系統一組主題，及其在文件集中的相關判斷結果，以作爲訓練（Training）之用。參加者可以透過各種方式，從這些主題中產生欲輸入系統的查詢問句（即圖六中的 Q1 ⑨），並依檢索結果不斷反覆進行系統參數調整，以期能產生最佳系統效益。訓練完成之後，TREC 會自先前提供訓練的主題中選出約 50 個進行正式的測試（即 Routing Topic）。系統依據先前訓練的結果，從主題中產生查詢問句（即圖七中 Q2），在指定的新文件集上進行測試（即 Routing Document）。此階段所輸出的檢索結果就是 Routing Task 的正式結果。

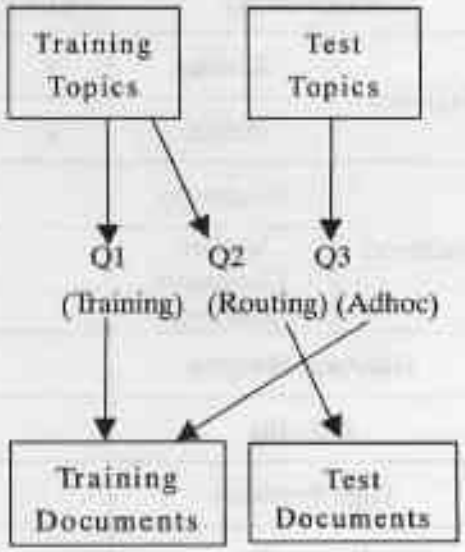
在 Routing Task 中，取得新的文件集常是比較困難的步驟，因此 TREC 選擇 Routing Topic 的方法是先搜集新的文件集，再據之選擇較適合的主題。倘當進行訓練與測試時所採用的文件集在特性上有較大的差異時，可能會影響測試的結果。例如 TREC-6 利用與訓練資料同質性較高文件集進行測試，精確率較 TREC-5 提昇了約 9%（TREC-5 與 TREC-6 使用幾乎相同的測試主題），但是有些人則認爲，若訓練與測試的文件集同質性過高，亦會使訓練的結果太過侷限。⑩

Routing Task 在 TREC-1 至 TREC-3 是屬於基本的測試項目，但是到了後幾年，此一測試項目變爲非必要測試，參加系統也漸漸減少。TREC-7 將停止舉辦 Routing Task，而以 Filtering Track 取代。⑫

⇨ Adhoc Task ⑬

此測試項目的概念是模擬圖書館中使用者搜尋資訊的行爲：圖書館的館藏是已知且固定的，而使用者可能提出的問題則是未知的。因此，與 Routing Task 恰恰相反，Adhoc Task 的目的在探討系統使用新主題檢索固定文件集的效益。

TREC 先給予參加者一組與 Routing Task 相同的訓練主題與文件集，而在正式測試之時，系統必須在這些訓練文件上，對 50 個新主題產生查詢問句（即圖七中之 Q3），輸出正式的測試結果。



圖七：TREC 主要測試項目運作圖

資料來源：Donna K. Harman, "The TREC Conference," in Readings in Inform-ation Retrieval, ed. Karen (Sparck Jones and Peter Willett San Francisco: Morgan Kaufmann Publishers, Inc., 1997.), 247.

圖七呈現了 Routing Task 與 Adhoc Task 的主要進行流程。TREC 主要利用新舊文件集

及新舊主題的組合來進行訓練與測試。在二個測試項目中均利用舊的文件集及主題進行訓練，產生最初的查詢問句 Q1，而 Q2 及 Q3 則是根據訓練時所得的經驗及結果而建立的。將 Q2 及 Q3 輸入系統，便分別產生 Routing Task 與 Adhoc Task 正式測試結果。

二、特殊測試項目（Tracks）

TREC 爲了使其測試機制亦能對資訊檢索中一些特殊的議題及新的檢索技術進行有效的評估，TREC-3 開始在 Routing Task 及 Adhoc Task 之外，非正式地進行了一些其他的測試。自 TREC4 起，正式產生了一些特殊的測試項目，稱爲 "Track"。每個 Track 有其測試程序與指導原則，與會者可以選擇參加其中的一個或多個 Tracks。TREC4 首先舉辦了多語、互動、資料庫合併、混淆、過濾等五項 Tracks [6]，而其後的 TREC-5、TREC-6 及 TREC-7 也都建立了多個 Tracks，有的是新發展出來的，有的則承襲舊有，但在測試程序上有時會有所更動。表四將歷年來所舉辦過的測試項目作一整理。

表四：TREC 的測試項目

Tasks/Tracks		TREC1	TREC2	TREC3	TREC4	TREC5	TREC6	TREC7
Main Tasks	Routing	V	V	V	V	V	V	
	Adhoc	V	V	V	V	V	V	V
Confusion	Confusion				V	V		
	Spoken Document Retrieval						V	V
Database Merging					V	V		
Filtering					V	V	V	V
High Precision							V	V
Interactive					V	V	V	V
Multilingual	Cross Language						V	V
	Spanish			V	V	V		
	Chinese				V	V		
Natural Language Processing					V	V		
Query								V
Very Large Corpus							V	V

以下即對 TREC 歷年來舉辦過的 Tracks 作簡單的介紹：[48]

(一)混淆（Confusion）：[49]探討經由語音辨識及光學辨識等處理程序所產生之可能具有雜訊的混淆性資料，對於系統檢索效益的影響。TREC 將文件中的訓練經由刪除、取代等程序，產生具有數種錯誤率的不同文件版本，測試系統利用已知款目在混淆資料中的檢索能力。TREC-6 另外舉辦了語音文件檢索項目（Spoken Document Retrieval，簡稱 SDR），欲單獨探討系統對由語音轉而來之文件的處理能力。測試時模擬新聞廣播，以不同的方法產生三種不同文件版本，系統亦在其中執行已知款目的檢索。[50] 此項目主要在 Adhoc Task 中進行。

(二)資料庫合併（Database Merging，簡稱 DM）：[51]探討系統對不同資料庫的檢索，並將檢索結果作合併排序的能力。測試時以系統在單一資料庫中的檢索效益作爲比較基準。此項目主要在 Adhoc Task 中進行。

(三)過濾（Filtering）：[52]探討系統利用過濾方式檢索相關文件的效益。系統必須針對每一文件作是否相關的判斷，即作二元式分類（Binary Classifiction），將文件分爲相關與不相關兩類。它與 Routing Task 主要的不同點在於，系統不必將檢索結果排序，而對檢索效益的評估除了回收率及精確率之外，也測量其效用 u_t：

$$u_t = u_{u1}Ai + u_{u0}Bi$$

其中，Ai 爲系統送回之相關文件數量，B_i 爲系統送回之不相關文件數量，u_{u1} 爲使用者接受到相關文件所獲得的價值（爲一正值），而 u_{u0} 則爲使用者接受到不相關文件所獲得的價值（爲負值）。此項目所進行的程序與使用的測試集主要與 Routing Task 所用的相同。[53]

(四)高精確率（High Precision，簡稱 HP）：[54]探討系統的使用者介面及其效益。TREC 在測試時給予每位使用者 5 分鐘的時間檢索出前 10 篇相關的文件。此項目主要在 Adhoc Task 中進行。

(五)互動（Interactive）：[55]除了測試檢索結果之外，也將檢索中人與系統互動的過程考慮在內。TREC 每年均發展出不同的測試程序，嘗試找出一個最佳的評估互動系統的方法，例如讓使用者在某特定的時間中儘可能檢索出最多的相關文章。但 TREC 至今似乎仍未發展出理想的評估方式，有些人認爲這些測試方法並沒有考慮到使用者的特質及認知對檢索結果所可能產生的影響。[56]此項目主要在 Adhoc Task 中進行。

(六)多語（Multilingual）：[57]探討檢索系統處理非英語文件或主題的效益。至 TREC-7 爲止，共舉辦了西班牙文、中文以及跨語言（Cross Language，簡稱 CLIR）等三個多語的項目。其中跨語言項目是探討系統利用某一語文的主題去檢索另一語文文件集的能力，而 TREC 所提供的文件集爲內容近似但並非完全平行的語料。[58]訓練及測試時所使用之非英語的主題及文件集都是另外取得的，例如在中文部分使用的是新華社的新聞資料庫。此項目主要在 Adhoc Task 中進行。

(七)自然語言處理（Natural Language Processing，簡稱 NLP）：[59]探討目前自然語言處理技術在資訊檢索領域所能

達到的效益，測試時並與非自然語言檢索的結果相比較。

(六)查詢問句（Query）：⑩探討查詢問句此一變因對於系統檢索效益的影響。參加者必須處理不同型式的主題，例如以自然語言敘述的資訊需求、特別短的查詢問句等。

(六)大型語料庫（Very Large Corpus，簡稱VLC）：⑪探討系統處理特別大量資料的能力。測試所用的資料除了包括所有TREC曾使用過的文件集之外，亦加上另外搜集的資料庫，構成一大型的語料。TREC-6使用了20GB，約750萬篇的文件來測試，而TREC-7的目標是使用達100GB的語料。測試時依據檢索出前20篇文件的精確率、回應時間及檢索花費等項目來評估。此項目主要在Adhoc Task中進行。

三、一般指導原則⑫

除了個別項目的測試程序及方法之外，TREC亦訂定了一些一般性的指導原則，如知識庫建構、產生查詢問句的方法等，允許多樣化方法的使用，以期能反映實際的運作環境。

TREC的測試又可分爲AB二類，A類使用整個文件集，B類則是使用較精簡的文件集，參加者可以選擇其中一種測試方式，亦可以二者都加入。系統必須送回正式測試時所檢索出的前1000篇相關文件；另外，依各測試項目的不同，TREC亦允許參賽者使用不同的方法或檢索技術產生一至多組的測試結果，讓TREC進行評估。

TREC允許參加者利用多種方式從主題抽取真正輸入檢索系統的查詢問句，並大致上區分爲二種方式：一爲自動產生（Automatic），即系統以完全自動的方式從主題中抽取出查詢問句；另一則爲人工產生（Manual），除了完全以人來建立查詢問句之外，亦允許人與系統進行互動，或對查詢問句作多次的修正。實證性的研究顯示，互動式查詢問句的產生是很有效的，而適當地建立及使用索引典和片語索引也會增進檢索效益，⑬而自動抽取查詢問句的方式，雖然效益不致過低，但同一個方法是否適用於不同的主題？是否能依不同主題的特性（如長度）來產生查詢問句？這些都是值得探討的課題。

TREC-5的Adhoc Task將產生的查詢問句分爲短問句（Short Queries）和長問句（Long Queries）二個部分，短問句是使用自動方式抽取主題中任一欄位，長問句則可選擇使用自動或人工的方式來產生。⑭TREC-6又進一步將測試區分爲短問句（Short Queries）、長問句（Long Queries）、及超短問句（Very Short Queries），短問句抽取主題中的Description欄位，長問句針對整個主題來產生，而超短問句則只處理Title欄位的資料。TREC規定使用自動產生查詢問句的系統選擇至少一種方式，以利各系統間的比較，而人工的方式則無限制。表五將歷年來Adhoc查詢問句的抽取方式及欄位來源作了簡單的整理⑮。

表五：Adhoc 查詢問句的產生方式以及欄位比較表

Topic Components					TREC-5		TREC-6		
	TREC-1	TREC-2	TREC-3	TREC-4	Short Query	Long Query	Very Short Query	Short Query	Long Query
Title	V	V	V	V		V	V		V
Description	V	V	V	V	V	V		V	V
Narrative	V	V	V			V			V
Concepts	V	V							
Queries Construction	a/m	a/m	a/m	a/m	a	m/(a)	(a)	a	m/(a)

資料來源：Karan Spark Jones, "Summary Performance Comparisons TREC-2, TREC-3,TREC-4, TREC-5, TREC-6," in The Sixth Text REtrieval Conference (TREC-6) held in Gaithersburg, Maryland, November 19-21, 1997, ed. Ellen M. Voorhees and Donna K. Harman <http://trec.nist.gov/pubs/trec6/t6_proceedings.html>

伍、發展概況及重要結果

TREC 自 1992 年以來，每年的參加單位在逐漸增加中（見表六），包括世界各地十餘個國家的團體（主要來自美加地區），其中學校等學術研究單位與產業界的研發單位所佔的比例相當[63]，此顯示這樣一個測試標準機制在資訊檢索研究中的重要性，不論是在學術界抑或是產業界，都愈來愈受到認同。不過，由於參加單位必須自行利用 TREC 所提供的測試集進行訓練和測試，需要具有強大的計算能力，儲存空間的電腦設備，以及可將大量資料建立索引的機制，因此要參加 TREC 亦受到某種程度的限制。參加者的動機大多是希望能在大規模測試集中評估系統或驗證新技術的效益，另外亦有一些參加者的主要目的是欲觀察此領域的發展趨勢與新技術的研發。[64]

檢索系統效益主要受到環境參數和系統參數的影響。在 TREC 中，環境參數所指的是其整個測試的機制而言，包括測試集、測量準則及測試項目；而系統參數則是指系統本身所使用的技術，如索引模式、產生查詢問句的方法，訓練學習的程序，相關排序的準則等。[65]

在 TREC 中出現過的技術十分多樣化，而這些檢索技術的發展與增進，主要可歸結爲二個因素：其一爲因應 TREC 測試環境及轉變的趨勢，其二是受到其他系統所使用的技術的影響。表七將歷年來 Adhoc Task 中所關注的重要研究領域及發展演變做了一個整理：[66]

從表七中可以觀察到以下幾點：

一、系統所使用的檢索技術是相互影響的。成功的技術在參加測試的各團體中將會快速地散布。例如 Smart 與 Okapi 的加權演算法就是最好的例子。

二、許多新技術是因應當時環境的轉變而產生，如 subdocuments/passages 檢索是由於當時對於全文文件或特別長文件的處理技術仍不成熟。但是後來由於有效的詞彙加權技術的出現，使得該技術在 TREC-4 及 TREC-5 時顯得式微，到了 TREC-6，爲了增進相關回饋的效能，系統又開始採用此技術。

三、由於不同型態的技術可以分別在不同的方面產生效益，因此可能會同時各自分線發展，例如自動與人工的詞彙擴展方式即是很好的例子。

但是本文不擬對系統使用的檢索技術作深入的探討，而希望能從 TREC 測試環境對於系統效益的影響此一角度來分析。以下將歷屆 TREC 中所產生的重要事件及影響作一簡單的回顧：

一、TREC-1：測試集在規模上由傳統的數 MB 巨幅地增加到 2GB，且文件的長度也比以往增加許多，再加上主題的呈現方式對於參加者來說亦甚爲陌生，參加者必須因應這樣的改變，對其原先的系統做大幅度的調整與重建，因此 TREC-1 所產生的結果只被視爲一個初步的試驗。[5]

二、TREC-2：由於系統經過一年的適應期，其整體表現比起 TREC-1 有相當幅度的增進，所以 TREC-2 被視爲評估檢索系統效益的一個標準點。值得注意的是，系統自動產生查詢問句的方式與人工產生的效益相當，這無疑鼓勵了系統對自然語言處理介面的研究。[6]

三、TREC-3：除了自動詞彙擴展、段落檢索等技術的發展之外，TREC-3 開始非正式地舉辦一些特殊測試項目，如西班牙文、互動式檢索等。另外，TREC-3 Adhoc 主題較先前簡化，但並不會減低系統的效益。一般認爲，這樣的主題與使用者平常輸入系統的查詢問句比較起來，還是太長了。[7]

四、TREC-4：正式引進五個 Tracks，並針對其特性發展各自的測試標準與程序。最值得注意的改變是，Adhoc 主題大幅變短，欄位也減至一個，但系統依據這樣的主題所產生的檢索效益普遍不佳，推測可能是由於短問句所涵蓋的線索不足，使得系統無法有效地從中抽取出查詢問句。而 TREC-4 的參加者多半採用人工產生查詢問句的方式，也顯示了主題的特性對系統效益所產生的影響。[8]

表六：歷屆 TREC 參加單位數統計表

TREC-1	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
25	31	33	37	38	51

表七：Adhoc Task 中所採用之重要技術及其發展

	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
weighting algorithm	baseline for most systems beginning of Okapi weighting experiments	Okapi perfects BM25algorithm	new SMART weighting algorithm new INQUERY weighting algorithm	use of Okapi/SMART weighting algorithm by other groups	adaptations of Okapi/SMART weighting algorithm in most system
passages/sub documents	use of subdocuments by PIRCS system	heavy use of passages/ subdocuments			use of passages in relevance feedback passages
automatic expansion		beginning of expansion using top X documents	heavy use of expansion using top X documents	beginning of more complex expansion schemes	more sophisticated expansion experiments by many groups
manual expansion		beginning of manual expansion using other sources	Major experiments in manual editing/user -in-the-loop	continued user -in-the-loop experiments	extensive user -in-the-loop experiments
data fusion		initial use of "data fusion"	continued use of "data fusion"	continued use of "data fusion"	more complex use of "data fusion"
initial topics				Beginning of more concentration on initial Topic	continued focus on initial topic, including title

資料來源：Ellen M. Voorhees and Donna Harman, "Overview of the Sixth Text REtrieval Conference (TREC-6)," in The Sixth Text Retrieval Conference (TREC-6) held in Gaithersburg, Maryland, November 19-21, 1997, ed. Ellen M. Voorhees and Donna K. Harman, <http://trec.nist.gov/ pubs/trec6/t6_proceedings.html>

五、TREC-5：由於受到 TREC-4 主題結構的影響，許多系統均調整其參數以適應較短的主題，這使得系統可以自主題本身產生比以往更多的資訊，而非單純地從主題的詞彙中抽取關鍵詞。TREC-5 將主題調整回類似 TREC-3 的結構，但是系統的效益與 TREC-3 比較起來卻呈現了下降的情況，推測是主題本身的困難度較高所致。㉕而從此屆開始，TREC 允許系統可互動地產生查詢問句（如利用相關回饋的方式產生），如此一來，影響系統效益的因素就變得更爲

複雜，而人工與自動產生查詢問句的方法以何種比例結合較佳，亦成爲值得探討的課題。另外，由於 Routing Task 中訓練與測試文件集的同質性過低，導致系統效益大爲降低。[22]

六、TREC-6：在 Adhoc Task 中增設了超短主題（Very Short Topic）的測試，令人驚訝的是，利用它所產生的檢索結果比抽取 Description 欄位的短主題來得好。推測其原因，可能是 TREC-6 主題的相關文件較少，而系統透過 Title 欄中的少數幾個詞彙就幾乎可以檢索到所有的相關文件。因此，短主題利用 Description 欄中的資訊來產生查詢問句，可能只是增加了系統檢索的雜訊。另外，平均長度只有 2.6 個字的超短主題，似乎比較接近使用者實際搜尋資訊的模式，因此從很少的線索來建立查詢問句的方式，雖然可能無法導致非常高的績效，但仍漸漸開始受到重視。

陸、TREC 的影響與評價

TREC 對於文件檢索的影響可以由三個方面來看：[23]

一、在測試集方面：TREC 測試集至今已有 5GB 的文件集以及 350 個具有相關判斷的主題，並已被整個文件檢索的研究社群所廣泛採用。這樣一個大規模測試集的可得性，使得研究者得以發展出更符合真實運作環境的檢索系統。也由於此測試集的發展目標爲模擬真實的檢索環境，檢索系統亦可測試新技術在真實環境中可能的成效，增進原有技術，提昇檢索效益。目前 TREC 測試集正被資訊檢索的研究社群所廣泛採用，有些未能真正參加 TREC 的團體，亦使用此測試集來發展其檢索策略。

二、在測試項目方面：TREC 持續地致力於研究發展許多新的測試項目，使得不同的檢索技術均能在一致的測試環境中進行評估，也將傳統的文件檢索研究擴展至新的領域，如中文、西班牙文、跨語檢索等項目，使得對非英語文件的檢索亦能有一個可供評估測試的園地，而語音文件檢索則將語音辨識的研究與文件檢索作了一個初步的結合。

三、在會議及論壇方面：TREC 的舉行使得研究者能透過系統測試以及相互間的觀摩切磋，使系統的檢索技術得到改良，並獲致更高的檢索效益。如 SMART 系統在六年內績效成長爲二倍，就是一個很好的例子。會議的影響力亦使許多好的技術得以在資訊檢索界迅速傳播，甚至成爲基本的檢索技術之一。另外，許多 TREC 的參與者也開始習慣利用此會議的進行作爲其研究發展的趨動力。

但是，作爲一個系統評估的機制，TREC 在許多方面也受到了質疑：[24]

一、在測試集方面：測試集的可信賴度是很重要的，唯有如此，才能正確地評估出不同系統及檢索技術的價值。對於 TREC 測試集的批評普遍是認爲其文件集、主題及相關判斷均太過人工化。[25] 由於它們並非從現實的環境中取得，因此使得系統測試的成效仍有某種程度的不準確性。其中，對於相關判斷有效性的質疑是最多的。除了先前討論過有關相關判斷一致性的問題之外，Wallis 與 Thom 認爲 pooling 的方法不能獲致絕

對的回收率，因此對於某些要求回收率的主題來說，其測試結果會有失公正。[29] Beaulieu 等人認爲二元式（Binary）的相關判斷對真實的使用者來說是不太實際的。[22] Harter 則認爲 pooling 相關判斷技術雖然可行性高，但單一固定的相關判斷卻忽略了使用者、系統等環境變數，而使其在真實環境中應用的有效性受到質疑。[30]

二、在測量準則方面：TREC 僅以回收率和精確率作爲主要的測量準則，因此嚴格來說它只是以某個角度來在測量系統的效益，而並沒有真正到達「評估」的層次，因爲若僅以此角度進行檢索系統的評估，未免有失偏頗，真正的系統評估尚包括了許多層面：如系統回應時間、親和力、顯示格式、使用者對檢索的主觀滿意度等等。[29] Show 等人認爲，使用回收率和精確率並不能測量出系統的真正威力。[4]而 Harter 則認爲 TREC 測試結果分析的考慮層面不夠廣，應針對不同的主題進行分析，使得系統在不同情況下所產生的不同效益能夠顯現出來。[31]

三、在測試程序方面：雖然 TREC 也針對互動式檢索進行測試，但其他測試項目均缺乏原始資訊需求者或使用者的介入。[32]另外，資訊需求是動態的，會受到外在如其他人、其他資料的影響而隨時可能有所改變，這是批次（Batch）的測試方式無法考慮到的。[4]這些缺失，使得依據 TREC 的測試機制所設計發展的檢索系統，很容易忽略人的因素。

柒、結　語

影響系統檢索效益的因素十分廣泛且複雜。雖然 TREC 所提供的測試機制在許多方面都仍有其侷限性與爭議性。但它至少提供了一個前所未有的大型測試環境，以及系統間相互比較、討論發表的園地。無可否認的，TREC 的出現已使得檢索系統的發展得以更接近實際可行的環境，對系統效益的提昇也具有很重要的貢獻。

爲了使測試機制更能發揮其效益，並促進檢索技術的發展，TREC 每年都試著對其測試集、測試項目及測試程序作調整改進，以期能發展出一個更完善的評估體系，使測試的結果真正有其參考價值，並能符合檢索環境變動的潮流。

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重要性。目前除了 TREC 之外，已有一些針對不同語言設計的相似機制開始嘗試運作，如 NTCIR（NACSIS Test Collection for IR Systems）與 IREX（Information Retrieval and Extraction Exercise）計畫是以日文測試集爲主[33][34]，AMARYLLIS 計畫則是以法文測試集爲主[35]。

反觀國內，許多中文檢索系統也正在急遽地發展之中，但是目前這樣的測試標準與機制卻仍付之闕如。在這方面，TREC 的確作了一個很好的示範。希望在不久的將來，國內也能出現一個可供系統評比的測試環境，以輔助中文檢索系統的發展。

（收稿日期：1998 年 9 月 4 日）

註　釋

註①：其意義即為使用者需求的陳述。

註②：查詢問句在文件集中的相關文件，即俗稱之「標準答案」。

註③：Stephen P. Harter and Carol A. Hert, "Evaluation of Information Retrieval Systems : Approaches, Issues, and Methods," Annual Review of Information Science and Technology (ARIST) 32 (1997), p.8.

註④：同上註。

註⑤：Alan F. Smeaton and Donna Harman, "The TREC Experiments and Their Impact on Europe," Journal of Information Science 23:2 (1997), p.170.

註⑥：Ellen M. Voorhees and Donna K. Harman, "Overview of the Sixth Text REtrieval Conference (TREC-6)," in The Sixth Text Retrieval Conference (TREC-6) held in Gaithersburg, Maryland, November 19-21, 1997, <http：//trec.nist.gov/pubs/trec6/t6_proceedings.html>

註⑦：文件集中的文件不一定是連續性的集合，且格式不一，甚至不一定具有標題。

註⑧：第一片及第二片文件集是由 TIPSTER 計畫所發展。

註⑨：同註⑤, p.171。

註⑩：同註⑥。

註⑪：同註⑥。此文件是從 The Financial Times 中擷取出來的。

註⑫：Donna K. Harman, "Overview of the Fourth Text REtrieval Conference (TREC-4)," in The Fourth Text REtrieval Conference (TREC-4)held in Gaithersburg, Maryland, November 1-3, 1995, <http：//trec.nist.gov/pubs/trec4/ t4_proceedings.html>

註⑬：TREC-1 routing topics, <http：//trec.nist.gov/data/topics/trec6/topics.1-50.txt>

註⑭：J. P. Callan, J. Broglio, and W. B. Croft, "TREC and TIPSTER Experiments with INQUERY", Information Processing & Management 31:3 (May-June, 1995), pp.327-343.

註⑮：N. J. Belkin, ed al., "Combining the Evidence of Multiple Query Representations for Information Retrieval , Information Processing & Management 31:3 (May-June, 1995), pp.431-448.

註⑯：James P. Callan and W. Bruce Croft, "An Evaluation of Query Processing Strategies Using the TIPSTER Collection," in The 16th ACM-SIGIR International Conference on Research and Development in Information Retrieval, Pittsburgh, USA, June-July 1, 1993, p.354.

註⑰：同註⑯。

註⑱："TREC-3 adhoc topics," <http：//trec.nist.gov/data/topics/trec6/topics.151-200.txt>

註⑲："TREC-4 adhoc topics," <http：//trec.nist.gov/data/topics/trec6/topics.201-250.txt>

註⑳：Ellen M. Voorhees and Donna K. Harman, "Overview of the Fifth Text REtrieval Conference (TREC-5)," in The Fifth Text Retrieval Conference (TREC-5) held in Gaithersburg, Maryland, November 20-22,1996, <http：//trec.nist.gov/pubs/trec5/t5_proceedings.html>

註㉑：同註⑤，主題長度的計算以具主題意含的欄位爲主，如主題編號等不算在內。

註㉗："Topic creation," <http：//trec.nist.gov/presentations/TREC6/11.html>

註㉘：同註⑥。

註㉙：Karen Sparck Jones, "Reflections on TREC," Information Processing and Managemtns 31:3 (1995), p.294.

註㉚：同註⑤, p.171。

註㉛：同註⑥。

註㉜：同註⑫。

註㉝：同註⑤, p.171。

註㉞：同註⑳。

註㉟：黃慕萱，資訊檢索中「相關」概念之研究 (台北市：台灣學生，民國 85 年)，頁 123-124。

註㊱：Robert Burgin, "Variations in Relevance Judgements and The Evaluation of Retrieval Performance," Information Processing and Management 28:5(1992), pp.619-627.

註㊲：Stephen P. Harter, "Variations in Relevance Assessments and The Measurement of Retrieval Effectiveness," Journal of American Society for Information Science 47:1 (1996), p.40.

註㊳：Ellen. M. Voorhees, "Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness," in The 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24-28 August 1998.
<http：//www.itl.nist.gov/div894/894. 02/works/papers.html#sigir98.dvi>

註㊴：同上註.

註㊵："Evaluation Techniques and Measures," In The Sixth Text Retrieval Conference (TREC-6) held in Gaithersburg, Maryland, November 19-21,1997, ed. Ellen M. Voorhees and Donna. K. Harman.
<http：//trec.nist.gov/pubs/trec6/t6_proceedings. html>

註㊶：但是有些測試項目有其特殊的測量準則。

註㊷：本圖之原始資料取自 TREC-6 Routing Task Results, <http：//trec.nist.gov/pubs/ trec6/ t6_proceedings.html>

註㊸：同註㉜.

註㊹：Donna K. Harman, "The TREC Conference," in Readings in Information Retrieval, ed. Karen Sparck Jones and Peter Willett (San Francisco：Morgan Kaufmann Publishers, Inc., 1997), p. 247.

註㊺：同註㉒.

註㊻：同註⑤。

註㊼：同註⑥。

註㊽：同註⑳。

註㊾：同註⑮。

註㊿：依據 Track 名稱的字母順序排列。

註(51)：同註㉒。

註(52)：同註⑤。

註(53)：同註㉒。

註㊽：同註㊷。

註㊾：David D. Lewis, "The TREC-5 Filtering Track," in The Fifth Text REtrieval Conference (TREC-5) held in Gaithersburg, Maryland, November20-22, 1996, ed. Ellen M. Voorhees and Donna. K. Harman, <http://trec.nist.gov/pubs/trec5/t5_proceedings.html>

註㊿："What are the TREC-7 tracks?" <http://trec.nist.gov/faq.html#Q9>

註(51)：同註⑥。

註(52)：同註③，p.26.

註(53)：同註㊷。

註(54)：同註⑥。

註(55)：同註㊹。

註(56)：同註㊷。

註(57)：同註㉑。

註(58)：同註⑤。

註(59)：同註⑨。

註(60)：同註㊹。

註(61)：Karan Spark Jones, "Summary Performance Comparisons TREC-2, TREC-3,TREC-4, TREC-5, TREC-6," in The Sixth Text REtrieval Conference (TREC-6)held in Gaithersburg, Maryland, November 19-21, 1997, ed. Eellen M.Voorhees and Donna. K. Harman, <http://trec.nist.gov/pubs/trec6/t6_proceedings.html>
表中的 a 代表系統自動建立查詢問句，m 代表人工建立查詢問句，而（a）則表示參加者可自行選擇是否由系統自動建立。

註(62)：同註㊹，p.248.

註(63)：同註⑤, p.172.

註(64)：同註㊹，pp.299-305.

註(65)：同註⑤。

註(66)：同註㊹，p.252.

註(67)：同註㊹，p.252.

註(68)：Donna K. Harman, "Overview of the Third Text REtrieval Conference (TREC-3)," in Overview of the Sixth Text REtrieval Conference held in Gaithersburg, Maryland, November 2-4, 1994., ed. Donna K. Harman, <http://trec.nist.gov/pubs/trec3/t3_proceedings.html>

註(69)：同註㊷。

註(70)：同註㊸。

註(71)：同註㊸。

註(72)：Donna K. Harman, "The Text REtrieval Conferences (TRECs) : Providing a Test-Bed for Information Retrieval Systems," Bulletin of the American Society for Information Science 24:4 (1998), pp.12-13.

註(73)：同註⑤, pp.26-27。

註㉖：同註⑧, p.310。

註㉗：Peter Wallis and James A. Thom, "Relevance Judgements for Assessing Recall," Information Processing and Management 32:3 (1996), pp.273-286.

註㉘：Micheline Beaulieu, Stephen E. Robertson, and Edie M. Rasmussen, "Evaluation Interactive System in TREC," Journal of the American Society for Information Science 47:1 (1996), pp.85-94.

註㉙：Stephen P. Harter, "Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness," Journal of American Society for Information Science 47:1 (1996), pp.37-49.

註㉚：同註⑨，頁 115。

註㉛：William M. Shaw, Robert Burgin and P. Howell, "Performance Standards and Evaluations in IR Test Collections : Vector-Space and Other Retrieval Models," Information Processing and Management 33:1 (1997), pp.15-36.

註㉜：同註⑩.

註㉝：同註⑭.

註㉞：同註②, p.16。

註㉟：NTCIR Project(NACSIS Test Collection for IR systems) Homepage,
<http://www.rd.nacsis.ac.jp/~ntcadm/index-en.html>

註㊱：IREX(Information Retrieval and Extraction Exercise)Homepage,
< http://cs.nyu.edu/cs/proteus/irex/index-e.html>

註㊲：AMARYLLIS, < http://www.inist.fr/accueil/profran.htm>