# 中文檢索評估系統可行性研究
# Proposing a Prototype for Chinese Corpus Test Collection

吳美美 Mei-Mei Wu
國立臺灣師範大學社會教育學系副教授

Division of Library and Information Science
Department of Adult & Continuing Education
National Taiwan Normal University

【摘 要】

本研究主要在探討建置一個中文檢索評估系統的可行性。數位化成爲必要的趨勢之後，世界上各重要語言都受到其使用國家的重視與贊助而開始進行研究。1992 年起美國國家標準暨技術局(NIST)和美國國防部尖端研究企畫署的資訊科技室(DARPA)開始組織一系列的文獻檢索會議(Text REtrieval Conference, TREC)，每年舉辦一次評比研討會議，至今已有 TREC 9 的召開。日本在籌畫五年之後，亦在今年(1999)開始推展日文的檢索評比研討活動。本文提出建置一個中文檢索評估系統可行性研究，並期望政府重視支持而能建設一個良好的中文檢索研究環境，以便鼓勵更多的研究者參與中文資訊檢索的研究。

【Abstract】

Chinese language retrieval evaluation has caught the attention of researchers studying cross-language retrieval techniques internationally. The importance of Chinese language in information retrieval ( IR) has been recognized, however, the evaluation tool and research environment is limited. Given the importance of Networked Information Retrieval (NIR) and the accelerating development of Digital Library (DL), a Chinese corpus test collection and a set of evaluation criteria for IR evaluation is desperately needed for the IR research and development community. The purpose of this study is to propose a prototype of CTREC. Main tasks and tracks, test collections, topics, measurement procedures and evaluation criteria are proposed. Some solutions for implementing CTREC are also suggested.

關 鍵 詞 ：中文文獻檢索；中文資訊檢索；中文檢索評估系統；資訊檢索評估
Keywords：Prototype, Chinese Corpus Test Collection ,Text Retrieval Evaluation, Information Retrieval Evaluation

# 壹、背景

## 一、研究源起

美國國家標準暨技術局(National Institute of Standards and Technology, NIST, 以下簡稱NIST)和美國國防部尖端研究企畫署的資訊科技室(Information Technology Office, Defense Advanced Research Project Agency, DARPA, 以下簡稱DARPA)1992 年起開始組織一系列的文獻檢索會議(Text REtrieval Conference, TREC, 以下簡稱TREC)活動，主要的目的在透過提供大型的語料庫以及統一的測試程序，有系統地整理分析其評鑑的結果，以促進網路資訊檢索技術的迅速發展。 TREC 在 1996 年更加入了中文的語料，以便於進行跨語系檢索(cross-language retrieval)的研究。 TREC的舉辦吸引了全球從事資訊檢索研究人員的注意，尤其國內有許多資訊檢索的研究者對於 TREC 的評鑑系統深感興趣。不過由於 1996 年 TREC 所加入的中文語料，僅限於新華社及人民日報的簡體字新聞語料庫，對於有志參加 TREC 中文檢索系統的研究者而言，限制很大，間接也可能限制了未來中文在網路社會的正常發展。而 1998 的 TREC-7 更由於語料範圍上的限制，暫時終止中文部分的檢索評鑑活動。

爲確保中文資訊檢索系統之檢索品質，因應數位化圖書館對於網路資訊檢索的需求，以及提供研究者良好的測試環境，一套中文文獻評估系統實有其必要。本研究擬呼籲建立一套中文資訊檢索評估雛形系統，一方面希望能藉此引起國內專家學者的討論，共同參與研議，以期形成一個良好的中文檢索研究環境；一方面希望鼓勵中文資訊檢索研究者熟悉 TREC 的參與方式，進而參與國際的 TREC 活動。在未來若能嘗試和美國 NIST TREC 結合，使其成爲國際通用的評估系統，則中文的繁簡字有可能在國際的資訊交換系統中並存並行，中文獨特的象形、會意、望文生義的文字特色，才不會因爲科技的原因而莫名式微。

## 二、研究目的與研究問題

美國國家標準局自 1992 年起組織了「文獻檢索會議」(TREC)的活動之後，促進網路資訊檢索技術的迅速發展，在 1996 年更開始加入中文的特別評比項目(track)，資料庫來源包括新華社(Xinhua)的 24,988 篇及人民日報(People's Daily news)的139,801篇文章(Kwok, 1997, p.35)，其皆爲簡體字系統，對於研究中文文獻檢索系統研究的研究者及中文使用可能產生長遠的影響。本研究主要的目的可分成三方面來說，其一，成立一個「中文檢索評估系統」 (Chinese Text Retrieval Conference, CTREC, 以下簡稱CTREC)可以鼓勵更多中文資訊檢索研究者進行相關的檢索技術的研究；其二，可以促進中文資訊檢索研究者熟悉 TREC 的參與方式，進而參與國際的 TREC ，使中文資訊檢索的研究廣爲國際網路資訊檢索的社群接受，進而促進跨語系的檢索(cross-language retrieval)成效；第三，CTREC還可嘗試和美國NIST TREC結合，使其成爲國際通用的評估系統，則中文的繁簡字有可能在國際系統中並存並行。此外，建置一套中文資訊檢索效能評估系統尚有三個重要

的原因：

1.網路資訊檢索(Network Information Retrieval, NIR)日益普及，需有一套效能評估系統來確保檢索之品質；

2.數位化圖書館(Digital Library, DL)正處於大量蘊釀的階段，優良的中文化資訊檢索系統有其必要性，值得鼓勵；

3.中文化資訊檢索系統的研究人員正需要有一套效能評估系統來協助其研究發展的工作。

以美國NIST TREC而言，目前的檢索評估的範圍，仍以文獻檢索的評估爲主，而一個文獻檢索評估系統的要素主要包含二部分：合適的語料庫(data sources)，包括測試語料(corpus)、檢索問題(topics)，以及一套評鑑準則(evaluation criteria)。因此爲了解建置一個中文文獻檢索評估雛形系統的可行性，本研究提出之研究問題有四：

(一)可資使用的中文語料有那些？ 可否獲取？透過那些管道？

(二)評鑑的準則是什麼？如何決定？

(三)TREC是如何運作的？有那些可資借鏡之處？

(四)如果要形成一個CTREC ，要如何開始推動？

上述的研究問題可藉由下列的方法來回答問題，詳細的研究方法請參考本文第三部分。

(一)文獻探討及擷取相關研究者的經驗。

(二)專家討論會(panel discussion)：邀訪檢索引擎設計者及研究者座談。

(三)實地調查訪問(field study and interview)：針對一般系統使用者進行了解。

(四)系統建置的準備：了解適合建置一套評鑑雛形系統的可用軟體及系統架構。

# 貳、文獻探討

## 一、TREC概覽

資訊檢索研究一直著重於索引效率與檢索方式的改善。過去有關資訊檢索系統的評估研究，主要是在實驗室中進行，使用小型的語料庫、人工形成的檢索問題(topics)，以及非資訊需求者的相關判斷等方式。大型語料庫和大型的測試用檢索問題一直都是資訊檢索研究者的呼籲(如Sparck Jones & Rijsbergen, 1975, 1976；Sparck Jones, 1995; Harman et al., 1996, p.5等)。1992年11月TREC-1在美國國家標準暨技術局(NIST) 和美國國防部尖端研究企畫署的資訊科技室(DARPA)的贊助下開始進行，是1950年Cranfield資訊檢索評鑑研究實行以來，第一次各系統可以在相同文獻集合上，使用相同評估方法以比較各不同系統的檢索結果。TREC在資訊檢索研究領域中最著名的是其所使用的文獻集合十分龐大，而參與的系統與技術亦十分廣泛。TREC到了1999已進行到TREC-8，參與團體來自世界各地，只要能向計畫委員會提出系統功能說明及欲參與的評比項目(task)、特定檢驗項目(tracks)，經由計畫委員會評量之後即可加入。TREC的出現對資訊檢索領域的發展有很大的影響，藉著提供大量文獻集合及檢索問題等，可分析比較不同系統所採用的檢索技術，而最重要的是能達到公開討論、技術轉移的目的。

根據Smeaton & Harman (1997)，TREC成立的宗旨有下列五項：

(1)鼓勵大量文獻資料庫的資訊檢索研究。
(2)透過公開的學術討論，加強產業界與政府間的溝通。
(3)加速研究成果與產業應用間的技術轉移與交流速度。
(4)提供一個可以呈現資訊檢索技術最新發展之環境。
(5)改善評估技巧。

TREC最初的目標在鼓勵利用大量文獻集合以進行資訊檢索的研究，其提供一個設計完整的文獻集合以供檢索研究，並透過友善性的評估討論，鼓勵參與團體間的互動。由於NIST的加入，使擁有商業性檢索產品的團體也加入TREC，促進了研究實驗室與商產業間的技術轉移。以下簡單列出其主要的發展經過：

1989 美國DARPA進行Tipster Text program。
1992 DARPA與NIST合作舉行TREC會議，爲Tipster Text Program之部分計畫。
1993 舉辦TREC-2。
1994 舉辦TREC-3。
1995 舉辦TREC-4。
1996 舉辦TREC-5 (有中文語料加入)。
1997 舉辦TREC-6 (有中文語料加入)。
1998 TREC-7開始展開，中文評估部分因故取消。
1999 舉辦TREC-8。

## 二、TREC語料

TREC的測試資料包括：(一)語料集(corpus)；(二)檢索問題集(topics)。此外，相關判斷及所謂「正確答案」亦應考慮在內。以下以TREC-2爲例，說明測試資料如后(Harman, 1995b,c)：

### (一)語料集(corpus)

語料集(corpus)必須反映在TREC理論應用中，包括不同類型、長度、書寫形式、編輯層次、詞彙等。這些語料的處理方式有三：(1)用SGML標示語言予以規範處理；(2)在大結構上一致，小結構則保持彈性；(3)以保持資料原貌爲原則。語料集(corpus)以CD-ROM形式儲存，每一片儲存量爲1 gigabyte。TREC的語料庫內容包括有：

Disk 1
- WSJ—Wall Street Journal (1987，1988，1989)
- AP—AP Newswire (1989)
- ZIFF—Computer Select disks (由Ziff-Davis出版)
- FR—Federal Register (1989)
- DOE—來自DOE出版品的簡短摘要

Disk 2
- WSJ—Wall Street Journal (1990，1991，1992)
- AP—AP Newswire (1988)
- ZIFF—Computer Select disks (由Ziff-Davis出版)
- FR—Federal Register (1988)

Disk 3
- SJMN—San Jose Mercury

News (1991)

· AP—AP Newswire (1990)

· ZIFF—Computer Select disks (由Ziff-Davis出版)

· PAT—U. S. Patents (1993)

另根據 TREC-5 綜覽[1] 的統計資料顯示，各屆 TREC 的語料和檢索問題集(topics)的分佈如表 2.1 ，語料內容及內容量如表 2.2 。

**表 2.1 各屆 TREC 的語料和檢索問題的分佈[2]**

| Document and Topic Sets for TREC 1-5 | | | |
|---|---|---|---|
| TREC | Task | Documents | Topics |
| TREC-1 | ad hoc<br>routing | disk 1 & 2<br>disk 2 | 51 - 100<br>1 - 50 |
| TREC-2 | ad hoc<br>routing | disk 1& 2<br>disk3 | 101 - 150<br>51 - 100 |
| TREC-3 | ad hoc<br>routing | disk 1 & 2<br>disk 3* | 151 - 200<br>101 - 150 |
| TREC-4 | ad hoc<br>routing | disk 2 & 3<br>CS+FR | 201 - 250<br>assorted |
| TREC-5 | ad hoc<br>routing | disk 2 & 4<br>FBIS-1 | 251 - 300<br>assorted |
| TREC-6 | ad hoc<br>routing | disk 4 & 5<br>FBIS-2 | 301 - 350<br>assorted |
| *re-use of disk3 forced by lack of new data | | | |

表2.2 TREC的語料內容及內容量 [3]

| | Size (magabytes) | # Docs | Median # Term/Docs | Mean # Term/Docs |
|---|---|---|---|---|
| Disk 1 | | | | |
| *Wall Street Journal*,1987-1989 | 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 |
| Disk 2 | | | | |
| *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 |
| Disk 3 | | | | |
| *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 |
| Disk 4 | | | | |
| *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 |
| Routing Test Data | | | | |
| Foreign Broadcast Information Service (FBIS) | 470 | 130,471 | 322 | 543.6 |

## 三、TREC檢索問題

在設計TREC時，必須提供使用者需求陳述，TREC中所使用的檢索問題(topics)均是模仿使用者需求，由資訊檢索者所設計。每一個檢索問題(topic)均以相同標準加以格式化，以便於自動找尋檢索問題，其格式包括：開始與結束的標記、檢索問題(topic)的編號、標題、簡短敘述、敘事體，以供對文獻判斷相關時的描述、概念部分，列出與該檢索問題(topic)有關的概念、定義部分、因素部分等，如下列例子[4]所示：

<num>Number:103

<dom> Domain: Law and Government

<title>Topic: Description:
Document will report those proposed or enacted changes to U.S. federal, state, or local welfare laws and regulations which are propounded as reforms.

<smry> Summary:
Document will reort those proposed or enacted changes to U. S. federal, state, or local welfare laws and regulations which are propounded as reforms.

<narr>Narrative:
A relevant document will reveal attempts by any U. S. jurisdiction to change the legal or regulatory context of existing welfare programs, or to add new programs, which are described or labeled as a reform. For the purposes of this topic, a welfare program must revolve around a transfer payment for which the recipient had not previously made contributions, other than through general tax payments. Examples include General Assistance, Aid for Families with Dependent Children (AFDC), public housing, Medicaid, income subsidies (rent and fuel subsidies, earned income tax credit, food stamps, etc.), child care programs (day care subsidies, nutrition programs, Head Start, etc.), and vocational training programs. NOT relevant are those programs which could be viewed as the payment of insurance benefits to those who had previously made contributions (either directly or indirectly through a family member or employer), such as Medicare, Social Security survivor's benefits, and uneploy with the financing or administration of existing programs.

<con> Concepts(s):
1.welfare recipient, public assistance, family assistance, welfare benefit, relief
2.welfare reform, welfare dependency, reform movement, workfare, homeless, welfare culture
3.General Assistance, Aid for Families with Dependent Children, AFDC, Medicaid, earned income tax credit, food stamps, Head Start, job training, public housing
<fac> Factor(s):
<nat> Nationality: US
</fac>
<def> Definitions(s):
Welfare program: although the above distinction between a "pure" transfer payment program and a putative insurance program is debatable, distinction is made to focus the data search on those programs which are clearly in the welfare category, as opposed to the wide variety of social programs supported by U. S. government entities.

以 TREC-5 爲例，150 個檢索問題(topics)中選出 50 個檢索問題(topics)，依檢索結果適合的相關筆數做爲選擇的依據。檢索結果太多和太少都不在獲選之列。檢索問題(topics)由執行相關判斷的同一人來研擬。訓練用的檢索問題(topics)同時要提供一套相對應的相關文獻集。以下是 TREC 1-6 檢索問題結構比較：

**表 2.3 TREC1-6 檢索問題結構比較**

| | | TREC-1 | TREC-2 | TREC-3 | TREC-4 | TREC-5 | TREC-6 |
|---|---|---|---|---|---|---|---|
| Field | Topics 1-50 | Topics 51-100 | Topics 101-150 | Topics 151-200 | Topics 201-250 | Topics 251-300 | Topics 301-350 |
| <top>, </top> | X | X | X | X | X | X | X |
| <head> | X | X | X | | | | |
| <num> | X | X | X | X | X | X | X |
| <dom> | X | X | X | | | | |
| <title> | X | X(3.8) | X(4.9) | X(6.5) | | X(3.8) | X |
| <desc> | X | X(17.9) | X(18.7) | X(22.3) | X(16.3) | X(15.7) | X |
| <smry> | | X | X | | | | |
| <narr> | X | X(64.5) | X(78.8) | X(74.6) | | X(63.2) | X |
| <con> | X | X(21.2) | X(28.5) | | | | |
| <fac> | Optional | Optional | Optional | | | | |
| <def> | Optional | Optional | Optional | | | | |
| <nat> | Optional | Optional | Optional | | | | |
| <time> | Optional | Optional | Optional | | | | |

[說明]

1. <num> =number; <dom> =domain, <desc> =description, <smry> =summary, <con> =concept(s), <fac> =factor, <def> =definition, <nat> = narrative.
2. X 表示具備此一欄位，( ) 內數字爲檢索問題平均長度。

## 四、TREC 評比項目

TREC 的執行方式主要分爲兩個部分，一、隨機測試(ad hoc)和二、固定檢索測試(routing retrieval)；各系統須就這兩種不同的檢驗方式進行測試；另有三、特定檢驗項目(tracks)活動，各屆 TREC 提供的評比項目(task)會略有不同，由參與者自行選項參加。將三種方法說明如后：

### (一)隨機測試 (ad hoc)

**目的**：檢驗各系統使用常見檢索問題(topic)來檢索大量文獻資料庫的表現。檢測既有的文獻資料庫是否能滿足各種新的檢索問題。

**方法**：將大量的異質文獻集合，送至各參與系統中，以供其編製索引或局部安裝於系統中。之後將由測驗用檢索問題(test topic)中產生的檢索詞彙集(query set)送至各系統中，使其就這些問題來檢索文獻集合，並就各問題的檢索結果依相關性排序，將最相關的前 1,000 筆資料文獻送回 NIST 作相關判斷，以檢驗其平均回現率與精確率。

### (二)固定檢索測試(routing retrieval)

**目的**：檢驗各系統在利用既定的檢索詞彙集(standing query)檢索新文獻資料庫時的表現。利用既定檔(profile)過濾新進的文獻。

**方法**：將做過相關判定的訓練用檢索問題(training topic)與大量的文獻集合送出給各系統，這些主題檔可使用於產生真正檢索時的檢索詞彙(query)。接著利用這些問題集就訓練用的文獻資料庫(training documents)來對系統施行訓練。待訓練完成後，各系統將被給予同樣由檢索詞彙集(training topic)中找出的固定問題集(routing query)，以其檢索新加入的測試用文獻語料庫(test documents)，並將結果送回 NIST 作相關判斷，檢驗其平均回現率與精確率。

### (三)特定檢驗項目(tracks)

允許參與者針對特定的檢驗項目進行測試，檢驗項目每年各有不同，過去有：語音文獻的檢索、多重語文文獻檢索、中文檢索，以及測試用在互動檢索上新的評鑑方法等。在 TREC 中所使用的資訊檢索技術範圍幾乎涵蓋所有資訊檢索技術，包括：

- 大量統計及詞彙加權(term-weighting)、段落與章節(passage)檢索
- 結合一種以上獨立檢索的結果
- 附屬文獻檢索分數(scores)結合計算文獻排序
- 基於先前相關評估的檢索
- 擴展
- 自然語言處理、統計上的片語檢索
- Bayesian 推理網路
- 超文獻檢索
- 或然率索引編製
- 布林與軟布林(soft boolean)檢索
- 納入語法結構的檢索
- 問題類別
- 基於同義字詞彙來擴展問題
- 由已知相關文獻來擴展問題
- 邏輯回歸技術
- 分散式問題處理
- 串列檢索
- 概念辨識
- 賦予標記或剖析

・字典式

・詳細索引與檢索

## 五、TREC 評鑑方法

TREC 的評鑑最主要的部分是相關判斷，其主要的技術包括集中排序混合法(Pooling method)，以及完整性和一致性的檢驗方法，分別說明如后：

### (一)集中排序混合法(Pooling method)

將各系統產生的最前 100 篇相關文獻集中排序，去除重複，再由一位研究者來判斷、決定相關的文獻。由於各系統都是利用排序的方法呈現檢索結果，選用最前 100 篇則無大困難。平均各年度的重合率在 TREC1-5 分別是 39%, 28%, 37%, 24%, 27% ，相關文章則分別是 22%, 19%, 15%, 8%, 4%。

### (二)完整性(completeness)及一致性(consistency)檢驗

1.完整性的檢驗

第二組 100 篇再用集中排序混合法(pooling method)重整一次，發現平均每 100 篇中大約僅有一篇相關文獻，因此可說第一次的排序混合法具有完整性的特質。

2.一致性的檢驗

TREC-4之後，對於49個隨機測試(ad hoc)的檢索問題(topics)，每個檢索問題(topic)隨機選出 200 篇相關文獻(若相關文獻不滿 200 篇，則全部相關文獻使用之)，並隨機取樣 200 篇不相關的文獻將之合併，另交由兩名相關判斷者來進行相關判斷。檢索結果是獲得頗一致的判斷（71.7% 的一致性）。在 14,968 篇文獻中，獲判有1992(13.3%)是一致的相關，8742(58.4%)是一致的不相關，總共有 71.7% 的一致性。和以前的研究相較，這樣高的一致性真不多見。不只是和原始判斷者相比，有 30% 是原始判斷者判斷爲相關而爲後來者判斷爲不相關。而原始判斷者判斷爲不相關，卻爲後來者判斷爲相關者，則少於 3% 。整體而言，相關判斷的一致性程度應可被接受。

# 參、研究方法與實施

## 一、研究方法簡述

本研究之研究程序分爲兩階段：第一階段--準備期：測試語料及環境的準備。了解 TREC 執行的程序各項標準的內容，包括 TREC 簡史、TREC 系統描述(一般狀況、執行方法)、TREC資料庫組成、TREC評鑑要項、TREC成效評估等。並舉行專家討論會(panel discussion)，邀集資訊檢索系統開發者、研究者舉行小型專家研討會，以了解其對擬進行的評估系統之資料庫內容、檢索詞彙(query)類型及評估標準的期望看法。第三爲實地調查訪問：(一)訪問系統使用者，以了解(1)資訊檢索系統一般使用者之資訊需求，(2)蒐集使用者的檢索問題(topics)，以及(3)使用者常用的資料種類；(二)訪問系統設計者以了解其對於評估系統的看法以及其設計檢索系統的特色。

第二階段則決定雛形系統的內容，包括：資料庫來源選擇 (資訊需求分析、資訊類別分析)、測試樣本 (query 的來源及數量)、資料共用(shareware)的可能性、測試的項目的範圍

等。此外並定義語料庫內容格式及研擬中文檢索評估系統使用辦法。

## 二、專家討論會

先後舉行兩次專家討論會，獲得CTREC的初步架構之共識。

## 三、實地訪談

實地訪談有參與TREC經驗之學者專家，以進一步了解TREC實施細節及考量的問題。他們實際參加TREC的經驗對於語料的收集、檢索問題的合適性、實施的辦法與執行的細節等項目提供相當多寶貴的意見。對於CTREC所需之軟硬體規格也有進一步之了解。

## 四、檢索問題(topic)的收集

訓練訪談員，實際對檢索系統的使用者進行觀察及訪談的工作。徵求使用者的同意利用前問卷收集其檢索問題(topic, search request, or search question)，接著觀察並記錄使用者的檢索詞彙(search terms)，於檢索結束之後並進行檢索結果的相關判斷後問卷，及訪談，以了解其對於該次檢索的收穫及滿意程度。收集的部分檢索問題於4.2中說明。

最後便是依據上述分析結果，進行系統分析與系統設計的工作。

# 肆、中文資訊檢索系統會議(CTREC)雛形研議

## 一、語料

目前國內有可能成爲測試用語料的語料來源列表如表4.1，其他可能使用的語料，可參考Hsu, Chang & Su (1997)包括有文字資源(text corpora)、電子字典(electronic dictionaries)、WWW網路資源(World Wide Web resources)等。

本研究並參考相關的契約書，研擬和語料提供單位簽約的契約樣本。

## 二、有關檢索問題（topic）

### (一)CTREC檢索問題結構可暫定如后：

<top>, </top>

<num>

<title>

<desc>

<narr>

### (二)CTREC檢索問題

理想的CTREC檢索問題應由有實際資訊需求的使用者產生。本研究產生經由實際收集有實際資訊需求的檢索者所提出的檢索問題共96題。真正在CTREC中使用的檢索問題，應依據語料庫中檢索到相關文獻的多寡，再選出可用的檢索問題（相關文獻太多，或太少均不適於使用）。預計選用30個訓練用檢索問題，評鑑用的新檢索問題30個。訓練用的檢索問題還應附有相關判斷的結果。可能的檢索問題則可以包括：釣魚臺的問題、日本北方領土的問題、有關公務人員之特別權力關係、有關兩稅合一、有關有線電視、有關司法羈押權的資料、有關貨幣貶值或是金融風暴的資料(以俄羅斯爲主)、有關黑松企業產業環境與行銷、有關可口可樂之行銷策略、有關我國反傾銷政

| (一) 中央研究院 Csmart 測試語料庫 | | |
|---|---|---|
| 語料庫名稱 | 資料筆數 | 資料量 |
| 中央社全文新聞(1991.01-1998.05) | 700,000 筆 | 約 900MB |
| 中央社英文全文新聞(1997.01-1998.05) | 13,000 筆 | 約 25MB |
| 中央社摘要新聞(1996.04-1998.05) | 220,000 筆 | 約 83MB |
| 教育部國語會國語辭典 | 160,000 筆 | 約 32MB |
| 國立成功大學圖書館書目 (中英) | 160,000 筆 | 約 25MB |
| 工業研究院 IT IS 產業資料庫(中英) | 330,000 筆 | 約 320MB |
| USENET News tw.bbs.* (1996.12-1998.05) | | more than 2GB |
| 華夏文摘(1991-1997), GB code | | 約 24MB |
| 聯合報 | | |
| 中國時報 | | |
| 佛學論文集 | | |
| SWETS 書目資料 | | |
| 美國圖書館學會英文權威檔 | | 約 2GB |
| (二) STICNET 科技資訊網路：國科會科資中心研究計畫及研究報告資料庫 | | |
| 語料庫名稱 | 資料筆數 | 資料量 |
| 博士論文摘要暨碩士論文目錄(1987-1995/02) | 53,072 筆 | 約 52MB |
| 研究報告摘要(1970-1997/07) | 53,932 筆 | 約 91MB |
| 研究計畫摘要(1986/07-1998/03) | 37,539 筆 | 約 57MB |
| 科技期刊論文摘要(1988-1998/01) | 117,092 筆 | 約 170MB |
| 人文期刊論文摘要(1991-1997/06) | 28,328 筆 | 約 11MB |
| 學術會議論文摘要(1988-1997/09) | 37,087 筆 | 約 53MB |
| 國科會研究獎勵論文摘要(1985-1996/07) | 33,851 筆 | 約 37MB |
| 科技簡訊(1985-1997/12) | 185.096 筆 | 約 321MB |
| 國科會補助出席國際會議論文(1991-1997/09) | 9,004 筆 | 約 16MB |
| 科技政策(1985/01-1998/04) | 14,851 筆 | 約 26MB |
| (三)新聞局光華雜誌 | | |
| (四)聯合報 | | |
| (五)國家圖書館中文期刊論文索引影像資料庫 | | |

表4.1 可能使用之語料庫

策與實施案例、緬甸加入東南亞國協的效應、有關共同基金之研究、有關有線電視設立過程及相關有線電視法、南向政策與對外直接投資等。不適用於做檢索問題(topics)者包括「太專門」、「太特定」、或「太一般性」的問題，選擇用於評估系統的檢索問題時，應予避免。

### 三、評鑑量標及評鑑方法

TREC 利用 Salton 在麻省理工學院(MIT) SMART 實驗室所開發的一套評鑑軟體，來檢驗各系統的評鑑結果。至於評鑑的主要依據，則仍是由同一位檢索者來進行相關判斷，而以計算出來的回現率和精確率，做爲各系統的評比依據。CTREC 對於評鑑量標的決定，應更爲審慎，最好組成一個工作小組來共同討論決定。尤其以過去 TREC 進行過八、九次的經驗，多方檢討的結果而言，評鑑量標的選擇仍是 TREC 運作的最大困難，譬如 Don Dabney (Harmon et al., 1996)即提出建議，亦即除了主題的相關之外，應加上以下九點評鑑的考量：新穎、撰寫者的權威、來源出版可靠、文字良好、有新想法、有觀點、有閱讀之樂趣、適合讀者之程度、有足夠之背景資料。CTREC 應針對系統開發者和使用者分別發展適用的評量量標。

## 伍、結語與具體建議

### 一、結語：中文 CTREC 的可行及必要

本文共討論了 CTREC 所需使用的測試語料、檢索問題、評估標準和評估的方法，以及會議的實施辦法，並預先設計了一個CTREC的首頁，從方法上而言，CTREC 是可行的。此外，由兩次專家會議及專家訪談可知CTREC的執行已刻不容緩。問題是這組織是如何形成？誰要擔負這個協調及執行的工作？從兩次的專家討論會，獲得的訊息是令人振奮的，與會的學者專家們多是樂觀其成且十分樂於參與規畫與執行的工作。爲推展CTREC的運行，具體建議如后。

### 二、具體建議

結合國內主要研究單位成立 CTREC 推動小組，並組成工作小組，來推動CTREC的實施，具體的建議：應即刻著手規畫中文檢索評估系統建置的進行，以下依推動的方式、語料、檢索問題、評比項目、其他注意事項敘述之：

**(一)推動的方式**

1.成立推動小組：邀請具學術影響力之學者贊助並推動此項活動。

2.成立執行小組：規畫及協調整體活動運作方式以 4, 5 人爲原則。

3.成立工作小組：規畫會議的各項細節工作，如

①推廣宣傳方面：(1)進行活動推廣宣傳 (2)規畫及執行學生獎勵辦法。

②技術方面：(1)檢索問題的研擬及相關判斷的進行(2)評比項目的規畫 (3)評估結果的執行（評鑑者的人數儘可能少，應具公信力）等。

4.在 CTREC 常設機構的考量上，學會的架構比學校好（因爲學校也參加比賽），而

舊學會比新學會好。因此建議先輪流主辦，固定運作組織，但成員可以置換。

5.CTREC 可能的常設機構－以下各政府單位與財團法人可供參考選擇：中央標準局、科資中心、國家圖書館、中研院資訊所。由於學校計算機中心能掌握的資源有限，人員異動性也大，可能並非合適的主辦單位。

6.經費來源：由推動小組提研究計劃(proposal)爭取經費。中華民國計算語言學會(ROCLING) 爲經費來源之一，應向其理事會申請。其他贊助單位可邀請廠商或透過基金會。另學生獎勵來源，可透過教育部軟體創作比賽、程式設計競賽。

**(二)語料**

1.目前收集、建整中的語料頗好、可接受，但有些語料單位比較樂於和以學校爲代表的研究群組定約。

(1)可考慮連繫無此嚴格要求之語料單位，透過相關學會提供語料服務，如要求中華民國計算語言學會語料興趣小組(ROCLING sigcorpus)加強收集語料。

(2)對於堅持只和學術單位簽約之語料提供單位，考慮由學校和學會合作舉辦活動。

2.可考慮酌增網際網路資料(Internet data)，並建議進一步研議，可設計在特定檢驗項目(tracks)中應用。

**(三)檢索問題**

可由收集的真實的檢索問題中，選出30個訓練用檢索問題，專家對檢索問題的設計有如下的建議：

1.訓練(training)用的檢索問題可以先測試一次，對於會產生過多或過少檢索結果的問題均不宜使用。

2.檢索問題的設計考慮分級，配合評比項目(task)的設計，從簡單的檢索問題 (topics)和評比項目(task)開始，以吸引團隊參加。

3.配合資訊檢索技術的發展趨勢，將檢索技巧的應用，加入檢索問題的規畫設計中。

4.檢索問題應注意合適性，太專門、特定的問題（如特定人名、公司名等）或太一般性的問題均不適合。

例如：a. **有關釣魚台與日本北方領土問題**：此題並不十分適合，兩者似乎無關係。

b. **有關季辛吉的資料有關黑松企業產業環境與行銷有關可口可樂之行銷策略**

以上三個檢索問題太特定，檢索出來的文獻太少，會影響評估結果。

此外，太專門和太過一般性的檢索問題會導致其產生的平均回現率 / 精確率(recall/precision)數字過於異常。例如：**有關企業併購**：此題太廣，可加以處理，可修飾爲：**美國企業併購台灣企業**，在題目處理有較合適的複雜程度。**有關有線電視設立過程及相關有線電視法**是不錯的題目。

5.檢索問題敘述(narrative)部分要說明清

楚。

6.Internet 上的檢索詞彙(query)品質不佳，其平均長度約只有一點多字，因而不能採用。

**(四)評比項目**

1.從特定檢驗項目(simple task)開始，以吸引團隊參加。

2.特定檢驗項目(track)的範圍包括原有之正式評比項目外，加強以下各項：加強創意評比項目，如Internet資料的應用、繁簡字、跨語系…等。若可能也可考慮速度、空間的評比，不過在技術執行上應先克服困難。

3.CTREC 應逐年調整其特定檢驗項目(tracks)。

**(五)其他注意事項**

其他注意事項包括：(1)先吸引至少 10 隊參加，國外團隊亦可。最好有人確定參加，再予以公開宣傳。(2)第一年試行，先以國內團隊爲主，鼓勵配合學生團隊參加。（注意時間表的安排配合）。(3)若要開放其他地區參加，要考慮到內碼問題以及繁簡、中英、東亞語系檢索等，如果有跨語系需要，需有其他考量：如會議使用語言(official language)應爲中文，提供翻譯服務。

國內在資訊檢索系統設計與製作的單位或研究群，日益增多，文獻檢索評估雛形系統若有可行，可以進一步發展，並建議由學術團體或政府相關單位繼續開發運作，有可能成爲中文文獻檢索系統評估的常設機置，並可進一步參與國際的 TREC 使跨語系檢索(cross-language retrieval)研究的環境早日形成。

誌謝：本報告爲財團法人資訊工業策進會八十七年度轉委託研究計畫部分研究結果。作者感謝清華大學語言研究所劉英享同學、師大資訊研究所吳松輝同學等協助收集資料、整理訪問稿，楊曉雯小姐協助校對並提供意見。一併誌謝。

## 附註

[1] http://trec.nist.org/pubs/trec5/papers/overview.ps

[2].http://trec.nist.gov/data/intro_eng.html

[3].http://trec.nist.gov/pubs/trec5/papers/overview.ps

[4].http://trec.nist.gov/data/topics/trec6/topics.101-150.txt

[5].Yu-Ling Una Hsu, Jing-Shin Chang, Keh-Yih Su(1997). "Computational tools and resources for linguistic studies". Computational Linguistics and Chinese Language Processing. vol.2, no.1, Feb. 1997, pp.1-40.

## 參考文獻

1.Allan, J. (1996). Incremental relevance feedback for information filtering. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 270-

278). New York: ACM.

2.Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation*, 53 (1), 8-19.

3.Beaulieu, M., Robertson, S., & Rasmussen, E. (1996). Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, 47(1), 85-94.

4.Belkin, N. J., & Callan, J. P. (1993). Effect of multiple query representations on information retrieval system performance. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 339-346). New York, NY: ACM.

5.Belkin, N. J., & et al. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3), 731-438.

6.Belkin, N. J. (1997). Remarks on the presentation of the Gerard Salton Award for Excellence in Research in Information Retrieval to Tefko Saracevic, July 1997. *SIGIR Forum*, 31(2), 14-15.

7. Blair, D. C. (1996). STAIRS Redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47(1), 4-22.

8.Blair, D. C., & Maron, M. E. (1990). Full-text information retrieval: Further analysis and clarification. *Information Processing & Management,* 26(3), 437-447.

9.Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3), 225-250.

10.Buckley, C., Allan, J., & Salton, G. (1995). Automatic routing and retrieval using SMART - TREC-2. *Information Processing & Management*, 31(3), 315-326.

11.Caid, W. R., Dumais, S. T., & Gallant, S. I. (1995). Learned vector-space models for document retrieval. *Information Processing & Management,* 31(3), 419-429.

12.Callan, J. P., Croft, W. B., & Broglio, J. (1995). TREC and TIPSTER experiments with INQUIRY. *Information Processing & Management,* 31(3), 327-343.

13.Charoenkitkarn, N. (1996). *The effect of markup-querying on search pattern and performance in large-scale text retrieval.* Unpublished doctoral dissertation, University of Toronto.

14.Damashek, M. (1995). Gauging similarity with n-grams: Language independent categorization of text. *Science*, 267(5199), 843-848.

15.Ellis, D. (1996). The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science*, 47(1), 23-36.

16.Ellis, D. (1994). Paradigms in information retrieval research: Information retrieval research and the con-

cept of a paradigm. In A. Kent (Ed.), *Encyclopedia of Library and Information Science* (Vol. 54, pp. 275-291). New York: Marcel Dekker.

17.Evans, D. A., & Lefferts, R. G. (1995). CLARIT-TREC experiments. *Information Processing & Management,* 31(3),385-395.

18.Gardiner, D. A. (1996). *Verb-based relations and conceptual proximity in information retrieval.* Unpublished doctoral dissertation, University of Minnesota.

19. Harman, D. (1992). Evaluation issues in information retrieval. *Information Processing & Management,* 28 (4),439-440.

20.Harman, D. K. (1993). The First Text Retrieval Conference (TREC-1) Rockville, MD, USA, 4-6 November, 1992. *Information Processing & Management,* 29(4), 411-414.

21.Harman, D. (1993). Overview of the first text retrieval conference. *Proceedings of the 14th National Online Meeting* (pp. 181-187). Medford, NJ: Learned Information.

22.Harman, D. (1993). Overview of the first TREC conference. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 36-47). New York: ACM.

23.Harman, D. (1994). Overview of the Third Text REtrieval Conference (TREC-3). *Proceedings of the 3rd Text Retrieval Conference* (NIST SP 500-226). Gaithersburg, MD: NIST.

24.Harman, D. (1995a). Overview of the Fourth Text REtrieval Conference (TREC-4). *Proceedings of the 4th Text Retrieval Conference* (NIST SP 500-236). Gaithersburg, MD: NIST.

25.Harman, D. (1995b). The 2nd Text Retrieval Conference (TREC-2). *Information Processing & Management,* 31(3),269-270.

26.Harman, D. (1995c). Overview of the 2nd Text Retrieval Conference (TREC-2). *Information Processing & Management,* 31(3), 271-289.

27.Harman, D. (1996). Panel: Building and using test collections. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335-337). New York: ACM.

28.Harman, D., & et al. (1996). Report on building and using test collections panel, SIGIR 1996. *SIGIR Forum,* 30(2), 5-10.

29.Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science,* 47(1), 37-49.

30.Hersh, W., Pentecost, J., & Hickam, D. (1996). A task-oriented approach to information retrieval evaluation.

*Journal of the American Society for Information Science,* 47(1), 50-56.

31. Hsu, Y.-L. Una, Chang, J. S., & Su, K. Y. (1997). Computational tools and resources for linguistic studies. *Computational Linguistics and Chinese Language Processing.* 2(1), 1-40.

32.Hull, D. A. (1996). Stemming algorithms: A case-study for detailed evaluation. *Journal of the American Society for Information Science,* 47(1), 70-84.

33.Jacobs, P. S., Krupka, G. R., & Rau, L. F. (1993). A Boolean approximation method for query construction and topic assignment in TREC. *Proceedings of the Second Annual Symposium on Document Analysis and Information Retrieval* (pp. 191-200). Las Vegas, NV: Univ. Nevada.

34.Jones, S., Gatford, M., Do, T., & Walker, S. (1997). Transaction logging. *Journal of Documentation*, 53(1), 35-50.

35.Kaszkiel, M. Z. J. (1997). Passage retrieval revisited. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 178-185). New York: ACM.

36.King, M. (1996). Evaluating natural language processing systems. *Communications of the ACM,* 39(1), 73-79.

37.Kwok, K. L. (1997). Comparing representations in Chinese information retrieval. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 34-41). New York: ACM.

38.Lancaster, F. W., Ulvila, J. W., Humphrey, S. M., Smith, L. C., Allen, B., & Herner, S. (1996). Evaluation of interactive knowledge-based systems: Overview and design for empirical testing. *Journal of the American Society for Information Science,* 47(1), 57-69.

39.Ledwith, R. (1992). On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. *Information Processing & Management,* 28(4), 451-455.

40.Lewis, D. D., & Sparck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM*, 39(1), 92-101.

41.Losee, R. M. (1996). Evaluating retrieval performance given database and query characteristics: Analytic determination of performance surfaces. *Journal of the American Society for Information Science*, 47(1), 95-105.

42.Mettler, M., & Nordby, F. (1995). TREC routing experiments with the TRW/Paracel fast data finder. *Information Processing & Management,* 31(3), 379-384.

43.Robertson, S. E. (1997). Overview of the Okapi Projects. *Journal of Documentation,* 53(1), 3-7.

44.Robertson, S. E., & Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Information Processing & Management*, 28(4), 457-466.

45.Robertson, S. E., Walker, S., & Beaulieu, M. (1997). Laboratory experiments with Okapi: Participation in the TREC program. *Journal of Documentation,* 53(1), 20-34.

46.Robertson, S. E., Walker, S., & Hancock-Beaulieu, M. M. (1995). Large test collection experiments on an operational, interactive system - Okapi at TREC. *Information Processing & Management,* 31(3), 345-360.

47.Salton, G. (1992). The state of retrieval system evaluation. *Information Processing & Management*, 28(4), 441-449.

48.Salton, G. (1995). Performance of text retrieval systems. *Science*, 268(5216), 1418-1419.

49.Salton, G. (1997). Expert systems and information retrieval. *SIGIR Forum*, 31(1), 39-42.

50.Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 138-146). New York: ACM.

51.Saracevic, T. (1997). USERS LOST: Reflections on the past, future, and limits of information science. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1-2). New York: ACM.

52.Saracevic, T. (1997). USERS LOST: Reflections on the past, future, and limits of information science. *SIGIR Forum*, 31(2), 16-27.

53.Schutze, H., Hull, D. A., & Pedersen, J. O. (1995). Comparison of classifiers and document representations for the routing problem. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval* (pp. 229-237). New York: ACM.

54.Smeaton, A. F., & Harman, D. (1997). The TREC experiments and their impact on Europe. *Journal of Information Science,* 23(2), 169-174.

55.Sparck Jones, K. (1995). Reflections on TREC. *Information Processing & Management,* 31(3), 291-314.

56.Sparck Jones, K. & van Rijbergen, C. (1975). Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection, British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.

57.Sparck Jones, K. & van Rijsbergen, C. (1976). Information retrieval test collections. *Journal of Documentation,* 32(1), 59-75.

58.Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing & Management,* 31(3),397-417.

59.Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management,* 28(4), 503-516.

60.Tague-Sutcliffe, J. (1995). *Measuring information: An information services perspective.* San Diego, CA: Academic Press.

61.Tague-Sutcliffe, J. (1996a). Information retrieval experimentation. In A. Kent (Ed.), *Encyclopedia of Library and Information Science* (Vol. 57, pp. 195-209). New York: Marcel Dekker.

62.Tague-Sutcliffe, J. M. (1996b). Some perspectives on the evaluation of information retrieval systems. *Journal of the American Society for Information Science,* 47(1), 1-3.

63.Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management,* 28(4), 467-490.

64.TREC-2 Program Committee. (1993). Report on TREC-2 (Text REtrieval Conference) 30 August - 2 Setember, Gaithersburg, *USA. SIGIR Forum,* 27(3), 14-18.

65.Veerasamy, A., & Belkin, N. J. (1996). Evaluation of a tool for visualization of information retrieval results. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 85-93). New York: ACM.

66.Viles, C. L. (1996). *Maintaining retrieval effectiveness in distributed, dynamic information retrieval systems.* Unpublished doctoral dissertation, University of Virginia.

67.Voorhees, E. M., & Harman, D. (1996). Overview of the Fifth Text REtrieval Conference (TREC-5). *Proceedings of the 5th Text REtrieval Conference* (NIST SP 500-238). Gaithersburg, MD: NIST.

68.Yang, J.-J. (1993). *Use of genetic algorithms for query improvement in information retrieval based on a vector space model.* Unpublished doctoral dissertation, University of Pittsburgh.

69. Zobel, J., Moffat, A., Wilkinson, R., & Sacksdavis, R. (1995). Efficient retrieval of partial documents. *Information Processing & Management,* 31(3), 361-377.