行政院國家科學委員會專題研究計畫成果報告

智慧型知識擷取技術與應用研究(III)：語料庫之設計與製作(III)

Design and Implementation of Corpora (III)

計畫編號：NSC 88-2213-E-002-035

執行期限：87 年 8 月 1 日至 88 年 7 月 31 日

主 持 人：陳光華　國立臺灣大學圖書資訊學系

研究助理：江玉婷　國立臺灣大學圖書資訊學系

一、中文摘要

在國內資訊檢索研究已日趨受到重視，合適的測試評估機制卻十分缺乏的背景下，本研究實際進行測試集的規劃與建置工作。測試集建構工作主要包括蒐集整理文件、建立查詢主題、以及進行相關判斷三個部分。本研究建立的文件集來源為新聞網站中的五種電子報，共有 132,207 篇文件。查詢主題是透過網路問卷實際徵集查詢需求，並進行三次的篩選之後，修正建構而成，共完成 50 個查詢主題。相關判斷的部分則是先對每個查詢主題建立一相關文件候選集，再針對候選集中的每篇文件以人工進行相關判斷，每一查詢主題由三位次判斷者同時進行，最後，則依據判斷結果計算並定義文件的相關程度。經由研究結果的分析顯示，本測試集有完整的架構及一定的規模，未來的研究應可以此為基礎，作進一步的擴展與改進

關鍵詞：標竿測試集、資訊檢索、相關判斷、查詢主題

Abstract

The research and development of information retrieval (IR) has made much progress recently. However, there's not any applicable mechanism for system evaluation in the Chinese research society. This project aims at the design and the implementation for Chinese information retrieval benchmark. Generally speaking, a benchmark consists of a set of documents, a set of topics, and a set of relevance between documents and topics. Accordingly, our task is also separated into three parts. The document set is downloaded from various electronic news sites, and totally 132,207 documents are collected. To build the topics, we investigate the real user information needs by using a questionnaire, and then modify them to be the formal topics. As to relevance judgment, we first set up a pool of candidate documents for each topic, and then invite three persons to judge the relevance. Finally, we combine the judgments and offer a relevance measure for each document in the pool. The result of our research shows that the benchmark possesses a complete structure and medium scale, and we may further expand and improve it based on existing framework in the future.

Keywords: Benchmark, Information Retrieval, Relevance Judgment, Topic.

二、計畫緣由與目的

人類的科技文明不斷的推進，其中資訊的交流與取得的方式扮演重要的角色，語言、文字承載著資訊透過傳輸管道散佈，使得人們取得所需的資訊，因而對於語言文字的理解是獲得資訊的不二法門；同時為了加速資訊的取得，發展檢索資訊的技術更是重要的工作。為了協助人類更有效的取得資訊，無數的學者專家研究發展不同技術，希望達到前述的目的。計算語言學家由語言文字本身入手，發展分析、理解、產生自然語言的技術，使得電腦成為終端使用者（End Users）的前端處理系統（Front-End Processing Systems），為使用者初步分析、過濾資訊，降低使用者的資訊負載。資訊檢索學家則從資訊取得的方式入手，研究更有效率的檢索方式，使得使用者可以快速地取得適切、合用的資訊，避免面臨眾多資訊時徬徨無依的困境。

為了評估學者專家的研究成果，通常採用一定的評估程序給予適當的評價。計算語言學的研究通常使用語料庫作為語言的素材，配合各種的統計模式據以建構語言模型。接著為了驗證語言模型的優劣，使用訓練語料（Training Corpus）或陌生語料（Unseen Corpus）進行模型的評估作業，前者稱為封閉性測試（Closed Test），後者稱為開放性測試（Open Test），然後根據評估的結果進行語言模型的修正。這裡吾人可以發現語料庫扮演訓練以及評估的雙重角色，然而就目前實際的研究方法而言，語料庫是以訓練為主，評估為輔。

資訊檢索系統的評估一直是個重要的研究課題，長期以來相關學派（Relevance School）與效用學派（Utility School）各自有獨到的見解。不過目前幾乎所有的資訊檢索系統均以相關學派發展的求全率（Recall）與求準率（Precision）為主要的評估準則；而效用學派似乎逐漸式微。進行求全率與求準率的評估作業，首先必須建構標竿測試集（Benchmark）。標竿測試集可分為三部份：第一是查詢主題；第二是相關判斷；第三是文件集。文件集在資訊檢索領域的重要性不僅其本身即為使用者欲檢索的對象，同時在系統發展的過程，文件集具有評估系統的功能。在這裡文件集主要是用以評估資訊檢索的績效而非訓練（當然它亦用以訓練檢索系統，但是就方法論而言主要是評估）。

觀察目前計算語言學與資訊檢索研究的發展，兩者的研究息息相關，彼此的交流也越來越蓬勃，由發表於 ACM 的資訊檢索學術會議（SIGIR）與 ACL 的計算語言學學術會議（ACL、COLING、EACL、ANLP）的學術論文可以清楚看到這個趨勢，因而語料庫的運用也成為這兩個學科領域的共同重視的研究方法。

本計畫的主要目的是建構適用於評估資訊檢索的標竿測試集，同時亦可用於訓練語言模型、建構計算機制的語料庫。適用於這兩種用途的文件語料，其文件數量必須夠大，評估的成果或訓練的模型可信賴度才可能滿足需求。鑑於文件語料的蒐集、整理、與標記工作需耗費大量時間，本計畫將先建構測試集的雛形，分析測試集的特性（包括文件長度、文件主題、標記種類、描述文件的方式），並制訂標準流程。

三、研究方法

本計畫主要的目標是建立一個可實際應用的資訊檢索系統標竿測試集，首先要確立測試集的主題。以 TREC 的經驗為例，經過多年的努力，目前擁有 400 個不同的主題，如第 51 個為 Airbus Subsidies；第 52 個為 South African Sanctions。TREC 主題成長的模式為逐次增加 50 個，設計者為每次 50 個的主題蒐羅文件語料，同時為每個主題做出文件語料的相關判斷。TREC 的作法顯示由小規模的測試集做起，每次著重於幾個領域的文件，簡化整個作業流程。本計畫採用相同的模式，初期將建構 50 個查詢主題。

主題確立後則必須進行初步的使用者需求分析，亦即使用者通常使用何種方式檢索文件，使用者檢索的方式會影響標記主題的格式，以及描述主題的用語。這部份將利用問卷調查的方式，分析使用者的需求。分析使用者需求之後，根據分析結果制訂主題的構成元件，TREC 的欄位有編號欄位、領域欄位、題名欄位、描述欄位、摘要欄位等等。此外必須確認哪些欄位是必要欄位，哪些可以省略。接著必須為每一欄位制訂標記，若以 TREC 為例，其為文件加上的標記如圖一所示。

<top>
<head> Tipster Topic Description
<num> Number: 051
<dom> Domain: International Economics
<title> Topic: Airbus Subsidies
<desc> Description:
Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<smry> Summary:
Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<narr> Narrative:
A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.
<con> Concept(s):
1. Airbus Industrie
2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3. federal subsidies, government assistance, aid, loan, financing
4. trade dispute, trade controversy, trade tension
5. General Agreement on Tariffs and Trade (GATT) aircraft code
6. Trade Policy Review Group (TPRG)
7. complaint, objection
8. retaliation, anti-dumping duty petition, countervailing duty petition, sanctions
<fac> Factor(s):
<def> Definition(s):
</top>

圖一、TREC 的主題標記

由圖一可以發現 TREC 共使用<top>、<head>、<num>、<dom>、<title>、<desc>、<smry>、<narr>、<con>、 <fac>、<def>、</top>等 12 個標記。本計畫依據 TREC 的作法並參考 TEI Header 的標記方式，制訂適用於中文的標記集。

接下來重要的工作即是蒐集大量的文件資料，本計畫的作法是從網際網路上持續且大量地下載各種主題的文件。對於搜集文件、建立測試集的過程中，利用網際網路上的資料可以有下列助益：

a 時下許多資訊檢索系統都是應用於網際網路，其檢索機制的設計也多是針對網際網路上文件的

特性。因此，採用網際網路上的資料來作為測試集的主要組成元件，比較切合資訊檢索系統的應用對象。

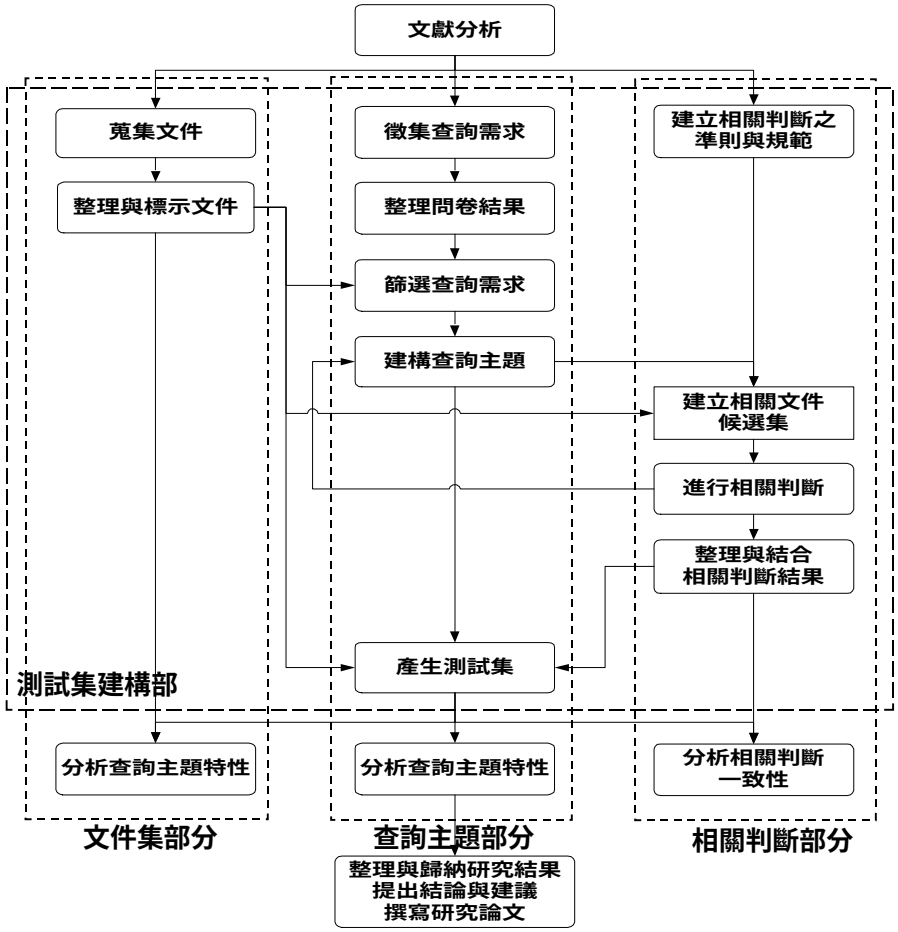
- 網際網路上的資料傳播速度快，且大多具有正當的存取管道，得以比較容易地取得資料，因而可以降低蒐集文件的困難。
- 網際網路上的資料新穎，能夠即時反映目前語言文字的使用情形，因此可以測試出資訊檢索系統的是否能適應時代的走向及需求。

此外在蒐集文件的同時，必須使用先前制訂的標記集進行文件的整理、組織、與標示的工作。

對於製作與各主題相關文件的相關判斷是本計畫比較困難的部份，這牽涉了主觀判斷的問題，有些學者對於所謂的相關有極為不同的看法。相關可以分為主題相關、情境相關、心理相關等多種情形，為了降低爭議，本計畫所指的相關為「主題相關」，並假設使用者能夠判定文件與主題是否具有相關性，在這個假設之下，本計畫聘請多位工讀生進行文件相關的判讀。

四、研究成果

本研究之實施流程如圖二所示。下文將依文件集、查詢主題、相關判斷三部份分別說明研究成果。



圖二、研究實施流程

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，文件來源主要為中時電子報、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之新聞全文，且文件長度不致過短，大致上符合文件集構成之要求。

從網站搜集而來的文件多為 html 格式，除了新聞內容本身之外，有些尚具有其他頁框內容或相關連結等資訊，呈現形式也十分多樣化。但是由於本研究所欲建立的測試集是以文件檢索為目的，為了使系統易於對文件進行辨識與處理，文件格式應具有某種程度的劃一性，如此在測試時也不致因為文件中的其他雜訊，而使系統的檢索結果受到影響。因此，我們將下載的文件整理成純文字檔，並刪除新聞報導之外的資訊，另外也考慮文件原始的結構與特性，將之加上標記(Tag)，使每篇文件均具有相同的格式與資料項目。我們僅針對文件的呈現形式作統一的處理，對新聞內容則不作任何更改。

文件標記的工作是主要以程式進行，在欲標示之文件內容前後加上開始標記（Start Tag）與結束

標記（End Tag），表示方式分別為「<標記名稱>」與「</標記名稱>」，各標記所包含的文件範圍是可以重疊的。圖三即為一經處理後文件範例，各標記項目與意義如下：

- <doc> ：標記文件的開頭與結尾，文件中所有資訊均涵蓋在內。
- <id> ：為文件識別碼，由「報紙名稱＋主題類別＋文件編號」組成。報紙名稱以英文示之，共有五種，分別為 chinatimes（中國時報）、commercial（工商時報）、express（中時晚報）、cdn（中央日報）以及 cdns（中華日報）。主題類別是主要依循該報紙對新聞的分類與名稱，各個網站對文件的分類方式與類別不盡相同。另外，我們就同一報紙與同一類別中的文件，再給予編號，預設值為七位數。因此，以此三個部分所組成的編碼在文件集中是唯一、不會重覆的，具有識別的功能，而透過此編碼我們也可得知文件的來源與類別。
- <date> ：標記新聞文件之日期，依據 ISO8601 之著錄規範，編碼方式為「〔西元年份〕（4 碼）-〔月〕（2 碼）-〔日〕（2 碼）」。如此可看出圖三的文件為 1999 年 1 月 8 日之新聞報導。
- <title> ：標記新聞文件之標題。
- <text> ：標記新聞文件之內文。
- <p> ：標記新聞文件之段落。

查詢需求以網路問卷的型式徵集，共計七題。第一題列出九個新聞主題，請使用者選擇所欲查詢的資訊類別。此處類別的區分與前述對文件集的分類方式稍有不同，原因是我們希望所列出的類別必須能讓填答者容易區辨、不致混淆，而前者則是欲在原有的分類基礎下進行整合歸類，二者的目的與考量點有所不同。第二題至第七題均是開放式問題，以不同的方式與角度詢問查詢需求的內容：第二題先請填答者說出欲查詢的主題（如事件、人物等）；第三題請填答者進一步敘述其在該主題中感興趣的部分；第四題則詢問在此主題中，有那些不是填答者想知道的資訊，三、四兩題的目的均是希望能藉以誘導填答者提出較明確、主題範圍較窄的問題；第五題請其列舉出一些關鍵詞彙，研究者可由這些相關概念進一步了解問題，也希望能獲致更多有關的資訊；第六題詢問提出問題的動機與目的，以窺探填答者對問題的認知與需求層面；第七題請填答者依據前幾題的回答，將其問題再做一次清楚的摘要陳述，如此可幫助填答者與研究者逐步釐清需求問題。最後，則請填答者提供一些基本的背景資料。

調查實施共回收了 405 份有效問卷。由於查詢需求是經由網路問卷徵集而來，未進行深入訪談，因此填答者的答卷品質並不整齊，對問題敘述的詳簡各異，主題與問題的形式也不一定適合作為測試集的查詢主題，因此，我們必須先對這些需求進行篩選的工作。我們最終的目標是產生 50 個查詢需求，所以我們亦預定篩選出 50 份查詢需求，再據之加以修正轉化為正式的查詢主題。此部分我們分三階段進行，第一階段純以人工針對查詢需求的內容陳述與特性作判斷，第二階段利用網路上的新聞搜尋引擎輔助篩選，第三階段則再次以人工檢視選擇。經由上述三次的篩選，餘下 50 個查詢需求。總刪除比例為 87.7%，各類別中刪除比例較高的有體育類、財經類以及科技資訊類，另外，娛樂類中的有關旅遊的查詢需求刪除比例亦較大。

查詢主題組成結構主要是參考國外各測試集的作法，並考慮測試的功能與需求而訂定的，目的是模擬使用者的查詢需求，並使查詢主題能展現各種形式、詳簡不同的需求內容。每個查詢主題均是由數個欄位結合而成，並使用不同的標記加以識別，其格式為「＜欄位名稱〉〔欄位內容〕</欄位名稱>」，「<欄位名稱>」為開始標記，「</欄位名稱>」則為結束標記。主要呈現查詢主題內容的欄位為 <title>、<question>、<narrative>、<keywords>，每部分均包含不同層面的主題資訊。另外，亦有 <topic>與<number>二個些識別性欄位，請參見圖四。茲將各欄位表示方式、意義與結構語法分述如下：

- <topic>：查詢主題標記。功能為識別查詢主題的開頭與結尾。
- <number>：查詢主題編號。主要由數字組成，其編碼結構為「00-000」。前 2 位數字的功能為識別不同時期或不同功能類型之查詢主題，本次研究之 50 個查詢主題均編碼為 01。後 3 位數字則為查詢主題之數字編號，其順序並不具備特殊意義，但本研究將主題相近之查詢主題集中排列。
- <title>：查詢標題。主要是由名詞或名詞片語組成，是對查詢主題較簡單的描述，它並不會顯示其中所有的相關概念，但必須在意義上是能夠涵蓋它們的，亦即，此欄位通常會表現出該查詢主題中最具概括性與代表性的主題。因為其意義通常較廣，若僅以此欄位進行檢索，會得到一些不相關的資訊，但一般使用者在檢索時輸入系統的查詢問句卻往往與這樣的模式十分接近。
- <question>：查詢問題。以語句的形式陳述所欲查詢的資訊內容，原則上以一至二個句子組成。其與<title>欄位的不同處除了內容的呈現方式之外，所展現的查詢資訊需求層次也不同。如前所述，<title>欄位表現概括的概念，而<question>欄則是確實地傳達的查詢需求。
- <narrative>：查詢說明。主要由數個語句組成，是該查詢主題中對查詢需求最詳盡的描述如對<question>欄位內容的進一步解釋、專有名詞的釋義與澄清、相關與不相關資訊項目之擴展與列舉、以及對相關文件的特殊的需求與限制等。此欄位形成的主要概念是模擬真實檢索情況中，使用者所提出的對查詢需求的說明，以使問題更具體化，其在查詢主題中扮演支援<title>與

<question>**欄位的角色。**

- <concepts>**：相關概念。由一至數個關鍵詞組成，所有與查詢主題中各層次敘述有關的詞彙都可能包括在內，這些詞彙有些會在查詢主題的其他欄位中出現，有些則不會。**

綜合上所述，可知查詢主題中，<title>**、**<question>**、**<narrative>**、**<concepts>**四個欄位具有主題上的意義：**<title>**欄所涵蓋的主題範圍最廣，其次是**<question>**欄，**<narrative>**欄位雖然敘述最詳盡，卻也是其中最為特定的，而**<concepts>**欄中的詞彙則可能涉及上述各層次的主題。**

<doc>
<id>chinatimes_economy_0003302</id>
<date>1999-01-08</date>
<title>98 **年進出口皆大幅衰退** </title>
<text>
<p>
【記者謝錦芳台北報導】財政部七日公布去年全年海關進出口貿易統計，進出口皆出現少見的衰退幅度。出口比前年衰退九．四%，創下民國四十四年以來最嚴重的衰退；進口衰退八．五%，去年貿易出超五十九億美元，創下民國七十三年以來新低紀錄，影響所及，去年經濟成長率恐怕無法達到五%。
<p>
財政部統計長許國忠指出，受到亞洲金融風暴影響，去年我國對亞洲出口衰退幅度高達十八．五%，我國對日本貿易逆差創下一七六．九億美元的歷史新高。在亞洲景氣低迷之際，唯獨我國對歐洲出口仍有六．七%的成長率。若景氣在下半年逐漸復甦，今年出口成績可以由負轉正。
</text>
</doc>

圖三、標記的文件

<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>
財政部等相關單位為健全金融市場、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，視為不相關。
</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>

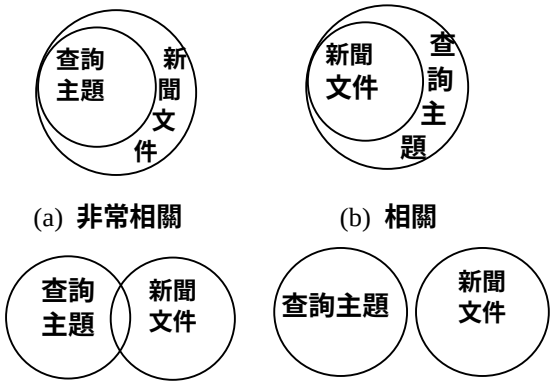
圖四、查詢主題

相關判斷的建構，首先必須確定相關的層次，我們將相關分為下列四個層次：

- **非常相關（相關分數為** 3**）：文件與查詢需求非常符合，查詢主題中提出的每個部分都能在該文件中找到相關的陳述。請參見圖五**(a)**。**
- **相關（相關分數為** 2**）：整篇文件內的絕大部分**

的陳述不超出查詢主題的需求範圍，文件中沒有與查詢問題不相關的部分。其與「非常相關」之主要不同點在於，查詢主題某些部分的需求在此等級的文件中可能會找不到答案。請參見圖五(b)。

- 部分相關（相關分數為 1）：文件中只有一部分符合查詢主題的需求，意即其中有些部分是不相關的。但判斷者可自行依據相關部分的程度多寡，判定文件是若應歸為「相關」或「不相關」，例如，若文件中僅有極少部分是不相關的，則可判定為「相關」。請參見圖五(c)。
- 不相關（相關分數為 0）：非以上三類者屬之。請參見圖五(d)。



(a) 非常相關　(b) 相關　(c) 部分相關　(d) 不相關

圖五、 查詢主題與文件之相關關係示意

相關判斷之實施必須對候選文件集的文件逐一進行判斷，共有近五千篇，而每篇文件會被判斷三次，因此判斷的總次數約為一萬五千次。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>欄位作為主要的判斷依據，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為適當的相關類別。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，則依據文件中的識別碼排列。18 位判斷者共耗費了約 230 小時完成所有的相關判斷工作。

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。但是，三位判斷者指派文件的類別可能會有多種情形，單以相關與不相關二個類別來看，各判斷者所認為相關的文件集合不盡相同，由此可以想見，若有四個相關類別，其判斷結果將會更複雜。因此，在相關判斷工作實施完畢之後，我們必須結合各判斷者的判斷結果，為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。

我們取三個判斷結果之平均數，作為結合後的相關分數。為了較易從分數上辨別一文件相關程度的高低，我們將算得的平均數再除以判斷的最高分數值（即非常相關的情況），使最後的相關分數介於 0 與 1 之間。相關值愈接近 1 者，為表示該文件愈相關，反之，則愈不相關。以下為文件相關度 (R) 之計算公式：

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

其中 X 為各判斷者對文件所給的類別等級，A，B，C 則為三位判斷者之代號。

為了評估判斷者對於相關判斷的一致性，本研究使用 Kappa（K），Kendal（W），及一致度（C）三項係數計算一致性。Kappa 統計量是考慮判斷的類別，Kendall 考量的則是順序關係（Siegel, 1988），「一致度」的公式如下所示，

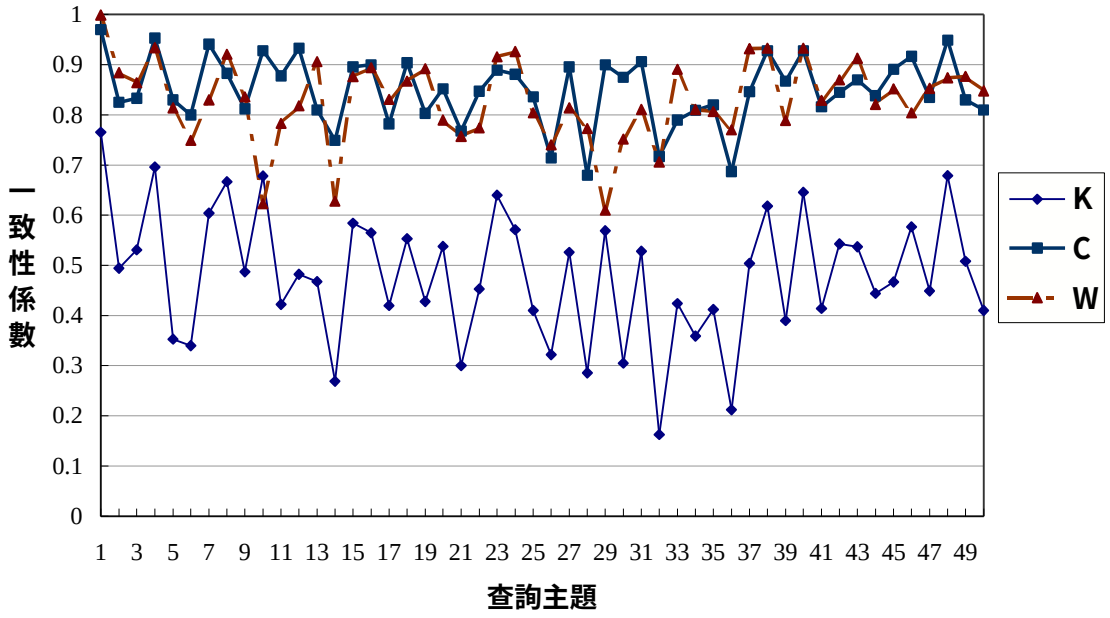
$$C = 1 - \frac{|X_A - X_B| + |X_B - X_C| + |X_C - X_A|}{6}$$

其目的是希望能反映不同相關程度所顯現的相關意義，也就是說，我們認為不一致情形應考慮判斷結果在相關意義上的接近程度。計算結果如圖六所示。我們同時分析各項係數的顯著性，結果顯示各判斷者的一致性相當高，亦即本研究所建構的標竿測試集實用程度很高。

五、結論

本研究已實際建構完成一包含文件集、查詢問題以及相關判斷的完整測試集，也初步驗證了此建構程序是可行的。與現行其他測試集相較，本測試集的規模已在中等以上，在文件集與查詢主題方面，均盡量使其能接近真實之檢索環境，提高其測試的效度，而相關判斷的部分，亦結合多位判斷者進行，減低了判斷結果可能出現偏差機率。在各界急於研發中文資訊檢索系統的今日，預期此測試集之建置與出現，應能稍微解除目前國內中文完全無從取得測試資料的現狀，使中文資訊檢索系統的發展能有更高的可行性，也期望它能成為後續相關研究的基礎。未來的研究可著重於下列幾項工作：

- 進一步擴展文件集規模，以增進其效度。
- 在查詢主題中加入非主題式的陳述。
- 分析查詢主題之難易度。
- 加入判斷者指派相關類別的信心值。
- 研究團體判斷的可行性。
- 研究不同背景判斷者對判斷結果的影響。
- 測試集有效性測試。
- 研擬適當的系統效益評估方法。



圖六、相關判斷一致性分析

六、參考文獻

"AMARYLLIS Homepage." <http://www.inist.fr/accueil/profran.htm> (Oct. 29, 1998)

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

"MIRA (Evaluation Frameworks for Interactive Multimedia Information Retrieval Application) Homepage." <http://www.dcs.gla.ac.uk/mira> (Nov. 5, 1998)

"NTCIR Project (NACSIS Test Collection for IR Systems) Homepage." <http://www.rd.nacsis.ac.jp/~ntcadm/index-en.html> (Oct. 31, 1998)

"Test Collections." <ftp://ftp.cs.cornell.edu/pub/smart/ > (Dec. 5, 1998)

"Test Collections." <http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/> (Dec. 5, 1998)

"Text REtrieval Conference (TREC) Homepage." <http://trec.nist.gov/> (Dec. 7, 1998)

Beaulieu, Micheline, Stephen E. Robertson, and Edie M. Rasmussen. "Evaluation Interactive System in TREC." Journal of the American Society for Information Science 47, no. 1 (1996): 85-94.

Belkin, N. J., J. A. Shaw, Edward A. Fox, and P. Kantor, eds. "Combining the Evidence of Multiple Query Representations for Information Retrieval." Information Processing & Management 31, no. 3 (1995): 431-448.

Borlund, Pia. "Simulated Information Needs: A Practical Approach." In Proceedings of the 6th Mira Workshop, Dublin, October 28-30, 1998. <http://www.dcs.gla.ac.uk/mira/workshops/dublin/procs/borlund/> (Nov. 5, 1998)

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

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

Cleverdon, Cyril W. "The Significance of the Cranfield Tests on Index Languages." In Proceedings of the 14th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Chicago, IL, October 13-16, 1991, 3-12.

Eisenberg, Michael B. and X Hu. "Dichotomous Relevance Judgements and the Evaluation of Information Systems." In Proceedings of the American Society for Information Science Annual Meeting, 24, 1988, 66-70.

Ellis, David. "The Dilemma of Measurement in Information Retrieval Research." Journal of the American Society for Information Science 47, no. 1 (1996): 23-36.

Harman, Donna K. "The Text REtrieval Conferences (TRECs): Providing a Test-Bed for Information Retrieval Systems." Bulletin of the American Society for Information Science 24, no. 4 (1998): 11-13.

Harter, Stephen P. "The Cranfield II Relevance Assessments: A Critical Evaluation." Library Quarterly 41, no. 3 (1971): 229-243.

Harter, Stephen P. "Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness." Journal of American Society for

Information Science 47, no. 1 (1996): 37-49.

Janes, Joseph W. “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.

Kageura, K. 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, 493-496.

Kitani, Tsuyoshi 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, August 24-28, 1998, 345-346.

Lesk, M. E. and Gerard Salton. “Relevance Assessments and Retrieval System Evaluation.” Information Storage and Retrieval 4, no. 4 (1969): 343-359.

Matsui and others, eds. “Test Collection for Information Retrieval Systems form the Viewpoint of Evaluation System Functions. In Proceedings of International Workshop on Information Retrieval with Oriental Languages, 1996, 42-47.

Over, Paul. “Presentation on The TREC Interactive Track: An Overview.” In Proceedings of the 6th Mira Workshop, Dublin, October 28-30, 1998. <http:// www.itl.nist.gov/div894/894.02/works/presentations/dublin98/index.htm> (Nov. 5, 1998)

Pao, M. L. “Term and Citation Retrieval: A Field Study.” Information Processing and Management 29, no. 1 (1993): 95-112.

Parsons, Simon and E. H. Mamdani. “Qualitative Dempster-Shafer Theory.” In Proceedings of the III Imacs International Workshop on Qualitative Resoning and Decision Technologies, Barcelona, June 1993.

Reid, Jane 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)

Robertson, S. E. and Micheline Beaulieu. “Research and Evaluation in Information Retrieval.” Journal of Documentation 53, no. 1 (1997): 51-57.

Salton, Gerard. “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.

Salton, Gerard. “The State of Retrieval System Evaluation.” Information Processing and Management 28, no. 4 (1992): 441-449.

Saracevic, T. “Relevance Reconsidered.” In Proceedings of COLIS 2: Second International Conference on Conceptions of Library and Information Science: Integration in Perspective, Copenhangen-Denmark, October 13-16, 1996, 201-218.

Saracevie, T., P. B. Kantor, A. Y. Chamis, and D. Trivision. “ A Study of Information Seeking and Retrieving: Part I. Background and Methodology.” Journal of the American Society for Information Science 39, no. 3 (1988): 161-176.

Schamber, L. “Relevance and Information Behavior.” In Annual Review of Information Science and Technology (ARIST), 29, edited by Martha E. Williams, 3-48. New York: Interscience Publishers, 1994.

Shaw, William M., 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.

Siegel, Sidney. Nonparametric Statistics of the Behavioral Sciences. New York: McGraw-Hill, 1988.

Sparck Jones, Karan and C. J. van Rijsbergen. “Information Retrieval Test Collections.” Journal of Documentation 32 (1976): 59-75.

Sparck Jones, Karan. Information Retrieval Experiment. London; Boston: Butterworths, 1981.

Spink, Amanda and Howard Greisdorf. “Partial Relevance Judgements During Interactive Information Retrieval: An Exploratory Study.” In Proceedings of the 59th American Society for Information Science Annual Meeting, 33, 1997, edited by Candy Schwartz and Mark Rorvig, 111-122.

Tague-Sutcliffe, Jean M. “Some Perspectives on the Evaluation of Information Retrieval Systems.” Journal of the American Society for Information Science 47, no. 1 (1996): 1-3.

Voorhees, Ellen. M. “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.