教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 46 , no. 2 (Winter 2008) : 183-210

跨語言資訊檢索中查詢問題特性於檢索效益之影響

Influences of Query Characteristics to Retrieval Performance in Cross-Language Information Retrieval

陳光華　Kuang-Hua Chen
Associate Professor
E-mail: khchen@ntu.edu.tw

吳恬安　Tien-An Wu
Graduate Student
E-mail: r95126015@ntu.edu.tw

English Abstract & Summary see link at the end of this article

教育資料與圖書館學　46 : 2 (Winter 2008) : 183-210

跨語言資訊檢索中查詢問題特性於檢索效益之影響

陳光華*

副教授
國立臺灣大學圖書資訊學系
E-mail: khchen@ntu.edu.tw

吳恬安

研究生
國立臺灣大學圖書資訊學系
E-mail: r95126015@ntu.edu.tw

摘要

本研究分別由查詢主題(Topic)的文本特性、語言特性，及欄位特性等三個面向，檢視第六屆NTCIR資訊檢索評估會議之跨語言檢索評估項目的資料，試圖找出查詢主題特性對檢索效益之影響。研究結果發現查詢主題的文本特性僅有「主題類別」對檢索效益具影響力，涵蓋地域、時間空間明確性，條件敘述明確性及數量等對檢索效益沒有顯著差異；語言特性方面，以韓文表現最佳，表示在CLIR機制中，選用不同語言的問題集與文件集對於檢索效益的確會造成影響；欄位特性方面，本研究發現不論選用標題(Title)、簡短資訊需求(Description)、詳細資訊需求(Narrative)、相關概念(Concepts)為索引來源，不會造成明顯檢索效益的差異，但部分欄位的檢索效益確實比較好；在研究團隊與欄位特性及語言特性的交互作用的檢定，本研究發現研究團隊與檢索欄位在檢索效益上有交互作用，也就是說研究團隊的不同與檢索欄位的不同對檢索效益會造成影響，而在語言組合的選擇上則無明顯的交互作用。

關鍵詞：資訊需求，查詢問句，查詢問題，檢索效益，查詢主題

---

* 本文主要作者兼通訊作者。

2008/11/05投稿；2008/12/23修訂；2008/12/30接受

緒　論

在資訊爆炸時代如何快速有效地檢索資訊，以最省力的方法得到檢索結果，是資訊檢索研究領域重視的問題與挑戰；網際網路的全球化特性，則促使跨越語言的檢索需求更加地迫切。為了評估資訊檢索技術的有效性、優越性，及跨語性，以公正、公開的方法進行評估，並得到研究領域的認同，則是資訊檢索評估研究的重要課題。1966年Cleverdon進行的Cranfield Test II計劃，以測試文件、測試問題及相關判斷構成一組測試集（Test collection），並訂定一套效益測量準則以評估多種索引語言，在資訊檢索評估的研究領域具有里程碑的劃時代意義，直至今日仍佔有舉足輕重的地位，成為後續資訊檢索評估的學習模型及改良依據。

一般而言，使用者進行資訊檢索的模式，是依據其資訊需求（Information need）轉化為查詢問題（Question），然後將查詢問題建構為查詢問句（Query），以輸入資訊檢索系統，接著資訊檢索系統在文件集合進行檢索，將可能符合需求的文件輸出給使用者。資訊檢索的評估便是希望模擬這樣的程序，因此測試集應包括文件集（Document set）與問題集（Question set）。當然，若涉及語言的因素，文件集與問題集則會分別包含不同語言的文件與問題。此外，為了判斷檢索所得文件的相關性，還會有答案集（Answer set），也就是文件與問題二者之間的相關判斷。換句話說，我們可將測試集視為資訊檢索系統評估的基礎資料，參與評估的資訊檢索系統必須依據所訂定的查詢問題，以文件集作為檢索對象，並將測試集提供的答案（相關判斷）視為標準答案，以進行檢索效益的評估。對於實際用於廣泛檢索評估之測試集，並為研究者接受，作為學術性資訊檢索實驗之評估依據，研究者會稱之為標竿測試集（Benchmark），有關資訊檢索評估的詳細程序與資訊檢索測試集的建構，請參見陳光華、江玉婷（2000）一文。

Crandfield Test II計畫及後續的資訊檢索評估研究使用的查詢問題，雖以自然語言的方式陳述資訊需求，卻十分簡短，包含的訊息有限，因此不少學者認為它們過於簡化使用者的資訊需求，與實際情形相去甚遠，且不易進行相關判斷。1992年由DARPA（Defense Advanced Research Projects Agency）與NIST（National Institute of Standards and Technology）共同舉辦的文件檢索會議（Text REtrieval Conference，簡稱TREC），首先採用結構化的查詢主題（Topic），以自然語言的方式描述詳簡不一的資訊需求。結構化的查詢主題與其他評估研究使用的查詢問題的不同處在於，TREC以多欄位方式陳述各種不同層次的資訊需求。因而，研究者通常會將類似TREC多欄位形式的資訊需求陳述稱為查詢主題而非查詢問題。

1999年起由日本國立情報學研究所主辦，國立臺灣大學參與協辦的NTCIR

（NII-NACSIS Test Collections for IR Systems, http://research.nii.ac.jp/ntcir/）資訊檢索評估會議，著重於東亞語言資訊檢索的評估工作，迄今已舉辦了七次資訊檢索評估會議，也建構了許多測試集。測試集的查詢主題仿效TREC的做法，皆由多個欄位組成，包括標題（Title）、簡短資訊需求（Description）、詳細資訊需求（Narrative），及相關概念（Concepts）等欄位，並描述相關詞彙的定義、背景知識、檢索目的、相關判斷的準則等，以顯示使用者不同層次的資訊需求。

資訊檢索評估會議或計畫的主辦單位在建構查詢主題時，是依據可能來自不同主題領域的真實使用者需求，以自然語言修正後，寫入固定的欄位結構而成。參與檢索系統評估的研究團隊，可選擇使用特定的欄位進行資訊檢索實驗，最後再將檢索結果交由主辦單位，以相關判斷的結果評比各研究團隊研發之檢索系統的檢索效益。從上述流程可知，查詢主題是參與評估會議或計畫的研究團隊進行資訊檢索的依據，但查詢主題可能因各種不同的特性，存在著個別差異，進而導致各檢索主題的難易度不同，使得某些查詢主題檢索效益的表現明顯高於其他查詢主題。因此，若能分析查詢主題對於資訊檢索效益的影響，便可規劃更健全的資訊檢索評估會議或計畫。為進行這項研究工作，本研究使用第六屆NTCIR CLIR評估項目（Cross Lingual Information Retrieval Task）的查詢主題（Kishida, Chen, Lee, Kuriyama, Kando, & Chen, 2007），分析其不同面向的特性：㈠文本特性，㈡語言特性，㈢欄位特性，試圖探討查詢主題的各項特性與檢索效益之相關性。

本文結構如下：第一節說明研究背景及資訊檢索的評估工作；第二節探討資訊查詢問題的相關研究；第三節說明本研究的設計；第四節討論本研究的分析結果；第五節為結論。

二、查詢問題之相關研究

在資訊檢索過程中，使用者會依據其資訊需求，建構查詢問題，當使用者真正將其查詢問題送入資訊檢索系統時，是以查詢問句的形式，逐一地輸入查詢詞彙，然後等待檢索結果的產生。但並不是每個查詢問句都能夠順利地得到適當的檢索結果，有些查詢問題可能得到上萬筆檢索結果，有些查詢問題卻只有為數極少的回覆，使用者就會逐步地修正其查詢問句。然而，資訊檢索系統的檢索結果不一定會符合檢索者的需求，使用者也會逐步修正其查詢問題，因為還牽涉到檢索者認知的判斷。

前文提及之查詢問題、查詢問句與查詢主題此三個在字面上十分相似的詞彙，其實質含義是有所不同的。查詢問題指的是使用者根據其資訊需求（Information need 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。所謂的查詢問句是依據使用者的查詢

問題，經由思考、分析、處理後，輸入於檢索系統之詞彙或語句。查詢主題則是TREC首先提出之特殊用語，用以表示測試集的查詢問題，與其他測試集的不同點在於它是以多欄位方式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，但一般會將形式類似TREC的資訊需求陳述稱為查詢主題，本文在行文遣詞時並不特意區分查詢問題與查詢主題，亦會交替使用這二個詞彙。參與資訊檢索評估的檢索系統，在實際進行資訊檢索實驗時，會模擬使用者產生查詢問句的情境，由標竿測試集提供的查詢問題或查詢主題中，以人工（Interactive mode）或自動（Automatic mode）的方法產生查詢問句進行檢索。

Lancaster（1968）使用700,000篇文件與300個查詢問題評估MEDLARS檢索系統，研究結果將檢索失敗歸納為五項因素：㈠文件本身，㈡索引語言，㈢查詢問題的陳述，㈣檢索策略，㈤使用者的相關判斷。Fidel & Soergel（1983）提出影響資訊檢索的面向，包括檢索環境、使用者、查詢問題、檢索系統，與檢索者等面向。因此，許多早期的資訊檢索研究已提出「查詢問題」是影響檢索效益的因素之一。雖然，許多因素均可能導致檢索結果的差異，但在檢索者與檢索系統等條件皆受控制的情形下，影響不同查詢問題的檢索結果的因素，就是查詢問題本身，也就是查詢問題的不同，影響著檢索結果的不同。

許多學術文獻曾探討查詢問題的特性，及查詢問題對於檢索結果的影響。黃怡如（1999）將查詢問題的研究分為兩類，一類是理論性研究，純粹探討查詢問題的本質，以助於對問題的了解，進而提昇檢索效益。另一類則是實證性研究，以統計檢定查詢問題類型的影響，再為查詢問題的各種性質建立操作性定義並予以量化。

純粹探討查詢問題本質的理論性研究，有Lehnert（1978）、Derr（1982）、Saracevic & Baxter（1983）、Graesser & Murachver（1985）等人的研究。Lehnert（1978）將查詢問題初步分為五類，Why Questions、How Questions、Yes/no Questions、Occurrence Questions，及Component Questions，並進一步將問題擴展為13個類別，Causal Antecedent、Causal Consequent、Concept Completion、Disjunctive、Enablement、Expectation、Feature Specification、Goal Orientation、Instrumental/ procedural、Judgmental、Quantification、Request、與Verification。

Derr（1982）基於哲學及認知科學的分類概念，分析資訊中介者和參考館員實際面臨的問題，將「問題」的結構分為主題（subject）與疑問（query）兩部分，主題指被檢索事物的客觀範疇，疑問指被檢索事物的主觀表達，也因此使用者必定存在著一前提假定（Conceptual presupposition），再進一步形塑出疑問為何。而依據前提假定的種類可將查詢問題分為三大類：範疇類問題、文獻類問題，及使用者服務類問題。

Saracevic & Baxter（1983）參考學者提出的查詢問題特性，綜合為五個一般

化特性。範疇（Domain），表示查詢主題的領域及附加限制；清晰度（Clarity），表示查詢主題用語的邏輯、語法與內容；專指度（Specificity），表示查詢詞彙的專指程度；複雜度（Complexity），表示查詢主題包含的檢索概念數；前提假定表示查詢該主題需具備的知識架構及查詢需求。此分類依據也成為後續多項實證研究的基礎。

Graesser & Murachver（1985）承襲Lehnert的理論，將查詢問題分為功能與陳述要素兩個面向，功能面向包括Why、How、Enable、Consequence of、When、Where、Significance of；陳述要素則包括企圖、事件，與狀態，配對後有21種不同的組合。

實證方面的研究包括Saracevic & Kantor（1988）採用Saracevic & Baxter提出的五種問題特性，將查詢問題依據明確的測量標準分類，測試檢索結果受查詢問題特性之影響。研究結果顯示低清晰度、低專指度、高複雜度及多個前提假定的問題可檢索到較多的相關文件；另外，低專指度及高複雜度的問題，檢索結果的精確率表現最好，但各特性在回收率的表現上則無顯著影響。

Keyes（1996）仿效Lehnert提出的概念式分類，將查詢問題分為確認型（verification）、因果型（causation）、概念完成型（concept completion）、聯想型（association）與非聯想型（disassociation）五種。研究結果顯示，五種類型的問題在檢索結果上並無顯著影響。

陳明君（1999）分析檢索者之檢索背景和查詢問題對於檢索技巧及檢索結果的影響，參考Saracevic & Baxter提出的範疇、複雜度、專指度與前提假定四種特性將使用者在PsycLIT執行的查詢問題分類。研究結果發現檢索背景和查詢問題的確對於檢索技巧的運用及檢索結果造成影響，在查詢問題方面，主要影響在於問題的複雜性，複雜度越高則使用者需運用的檢索技巧越多，同時檢索次數也越多。而查詢問題的範疇、專指性和前提假定個數對於檢索技巧及檢索結果並沒有顯著的影響。

王怡人（2004）將Grolier線上百科使用者的查詢問題分為開放性、封閉性兩種特性，另有事實問題及看法問題兩種問題類型，也考慮查詢問題涉及的概念數量，研究中要求受測者以檢索查詢及瀏覽查詢兩種方式進行問題查詢。研究結果顯示，開放性看法問題較封閉性事實問題之檢索成功率較高，在瀏覽查詢中，問題概念數量與檢索表現成正比關係；但在檢索查詢中，問題概念數量與檢索表現則無特定關係。

參照於前述之各項相關研究，本文提出之研究屬於實證性研究，將分析第六屆NTCIR會議跨語言資訊檢索評估項目的50個查詢主題及各研究團隊的檢索結果，探討查詢主題的特性對於檢索結果的影響，Nelson（1995）雖曾做過類似的研究，但主要是探討查詢主題的長度的效應，本研究將分析查詢主題的文

本特性、語言特性，及欄位特性等面向，比Nelson（1995）更為完整而深入。Nelson（1995）主要是分析TREC-3 Routing與Ad Hoc二項評估項目各50個查詢主題的長度對於檢索結果的影響。Nelson列出168個停用詞後，記錄各個查詢主題在各欄位的字數及加總字數，分析內容長度對檢索結果的影響。研究結果顯示，查詢主題總字數與各欄位字數兩者與查詢主題的平均查準率（Mean Average Precision，簡稱MAP）皆為正向相關。換言之，查詢主題越長，文字描述越多，檢索效益越好。此外，第7、8、9屆TREC資訊檢索評估會議舉辦的Query Track（Buckley, 2000），其目的也在探討查詢主題的特性對於檢索的影響，但侷限於英文查詢主題與英文文件檢索的探討。

三、查詢問題分析之研究設計

本研究使用第六屆NTCIR CLIR評估項目的50個查詢主題為研究對象，並向NTCIR檢索評估會議主辦機構申請，使用參與該項評估項目的各國研究團隊的檢索結果作為分析查詢主題的依據。CLIR評估項目使用的文件集由新聞文件組成，選用新聞文件的原因在於新聞文件的主題分佈較為廣泛，且能即時反映目前語言文字的使用情形與特性，因此可測試出資訊檢索系統能否適應時代的走向及需求，也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（陳光華、江玉婷，2000）。另一方面，新聞報導的深度與廣度適合一般社會大眾閱讀，此項設計也可減少相關判斷時，因判斷人員的背景知識不足而導致判斷錯誤。

㈠第六屆NTCIR CLIR評估項目

CLIR是跨語言資訊檢索的評估項目，測試集採用的文件包括中文（Chinese, C）、日文（Japanese, J）、韓文（Korean, K）及英文（English, E）四種語言，參與評比的隊伍可選擇檢索主題的語言與檢索文件的語言。第六屆NTCIR的CLIR評估項目共有20個團隊參與，各研究團隊總共從事C-CJK、C-C、C-J、E-C、E-J、E-K、J-C、J-J、J-K、K-C、K-J，及K-K等12種不同語言組合的檢索項目。每個團隊可從查詢主題的各個欄位選擇使用標題（Title, <T>）、簡短資訊需求（Description, <D>）、詳細資訊需求（Narrative, <N>），或相關概念（Concepts, <C>）等欄位，並以各自研發的資訊檢索系統進行檢索，但規定每個參賽團隊都必須執行所謂的「指定檢索項目」（Mandatory run）。指定檢索項目包括由查詢主題的Title欄位擷取詞彙及由Description欄位擷取詞彙以產生查詢問題的檢索結果，每次的檢索結果稱為一個Run。此外，各團隊也可自行選擇從其他單一欄位或從多個欄位擷取詞彙以產生查詢問題。因此每個團隊交出的檢索結果至少有二個Run，一為Title Run（T Run），另一為Description Run（D Run），

其他欄位或欄位組合的可能檢索結果還包括DN Run、C Run，或TDNC Run等。

參與檢索評估的研究團隊選擇檢索欄位、檢索主題與檢索文件的語種之後，總計產生152個Runs，其標記方式為「研究團隊代號－問題集語言－文件集語言－選擇欄位－檢索項目序號」，例如LIPS-C-C-D-03表示研究團隊LIPS，選擇中文問題集與中文文件集，從Description欄位篩選詞彙的第03號檢索結果。

研究團隊送出檢索結果後，由相關判斷人員（Assessors for relevance judgments）給予每筆檢索結果與查詢主題相關程度的分數，相關判斷的層次又可分為四種：高度相關（S）、相關（A）、部分相關（B）與不相關（C）。經過判斷者人工比對檢索出的文件與查詢主題的相關程度後，每一題皆採用Average precision的計算方式，然後計算所有題目的平均表現，就是前文所提的Mean Average Precision（MAP）。計算相關程度的方式又可分為較嚴格的Rigid relevance（S與A視為相關）及Relaxed relevance（S、A與B視為相關）兩種，因此Rigid relevance會低於Relaxed relevance的分數，每題共計產生152組檢索結果與兩種計算方式的分數。檢索結果的分數介於0～1之間，越接近1則該筆結果的表現越佳，也就是檢索出的文件與查詢主題相關度越高。本文後續的討論與分析是使用Rigid relevance計算出的分數。

㈡影響檢索效益之因素

本研究將探討查詢問題的文本特性、欄位特性、語言特性三個面向，對於跨語言資訊檢索效益的影響，除了分析並探討可能影響檢索效益的因素，亦將討論研究團隊的檢索技術與前述面向的交互效用。

1. 文本特性

本研究首先分析50個查詢主題的文本特性，繼而比對各查詢主題與相關判斷得到的分數，試圖找出文本特性與檢索效益的相關性，亦即分析檢索效益高的查詢主題具有的文本特性，及分數低的查詢主題具有的文本特性。

本研究參考Saracevic & Baxter（1983）提出的查詢問題的範疇、專指度、複雜度與前提假定等四種文本特性，並考量新聞文件集的資料特性，設計發展查詢主題的屬性類別。新聞文件的查詢主題涉及地區、時間、空間、主題等要素，敘述文字明確與否、條件限制個數，都可能對檢索結果產生不同程度的影響。從Saracevic & Baxter提出的查詢問題之特性，本研究發展出五個查詢主題的文本特性：⑴涵蓋地域，⑵時間明確性，⑶空間明確性，⑷主題類別及⑸條件敘述明確性與數量，說明如下，並請參見表1。

表 1 本研究引申 Saracevic & Baxter 問題屬性之定義

Saracevic & Baxter		本研究
查詢問題之特性	定 義	查詢問題之文本特性
範疇 Domain	查詢主題的領域及附加限制	查詢主題的涵蓋地域與主題類別
專指度 Specificity	查詢詞彙的專指程度	查詢主題的時間、空間、條件敘述明確與否
複雜度 Complexity	查詢主題的檢索概念數	查詢主題包含的條件敘述個數
前提假定 Conceptual presupposition	查詢該主題需具備的知識架構及查詢需求	查詢主題的主題類別與條件敘述

- 涵蓋地域：全球範疇、區域國家、單一國家（3 Categories）

  因新聞討論的主題分佈廣泛，包含全球範疇、國家間的區域性問題，或只發生於單一國家的議題，因此將地域屬性分為全球範疇、區域國家、單一國家三個類別。
- 時間明確性：明確、不明確（2 Categories）

  新聞文件主題涉及事件發生的時間，在陳述資訊需求時可能限制尋找特定時間的事件，或廣泛找出所有相關的文獻，在此將時間特性區分為明確及不明確兩種。
- 空間明確性：明確、不明確（2 Categories）

  新聞文件主題涉及事件發生的空間，在陳述資訊需求時可能限制尋找特定空間的事件，或廣泛找出所有相關的文獻，在此將空間特性區分為明確及不明確兩種。
- 主題類別：政治、財經、社會綜合（含教育、環保）、生活（含醫藥）、科技資訊、藝文、體育、娛樂（含旅遊）（8 Categories）

  主題類別參考江玉婷（1999）針對新聞文件蒐集使用者資訊需求所進行的實證研究，將新聞文件分為 8 大類別，在此用來區分本文的查詢主題，每個查詢主題只歸屬於單一主題類別。
- 條件敘述明確性及數量

  每一個查詢主題皆包含 <REL> 欄位，內容為該查詢主題的條件敘述，分為相關、部分相關及不相關三種查詢需求層次，讓判斷人員決定查詢所得之文件與查詢主題的相關程度高低。陳明君（1999）的研究，將查詢問題包含概念的多寡視為該問題的複雜度，問題包含的概念越多，則複雜程度越高。本研究採用這個觀點，計算每個查詢主題包含相關、部分相關及不相關三種條件敘述的個數。同時，本研究亦考量這些條件敘述的明確性，分為明確與不明確。

2. 語言特性

參與評比的研究團隊選擇使用的「問題集」與「文件集」語言組合，可能是

影響表現分數的因素之一，觀察不同語言組合的檢索效益，可更深入分析以下四點。

- 各種語言的組合情形
- 各種語言的檢索效益
- 各種語言問題的表現
- 各種語言文件的表現

3. 欄位特性

參與評比的研究團隊可自行選擇欄位作為索引詞彙的來源，因此也可能是造成表現分數差異的因素之一，本研究分別討論以下兩項結果。

- 檢索欄位的使用情形
- 檢索欄位的檢索效益

4. 研究團隊表現

參與評比的研究團隊使用之檢索技術與前述三面向的交互作用，也可能是造成檢索效益差異的因素之一，本研究分別探討以下二項。

- 檢索技術的檢索表現
- 檢索技術與其他因素之交互作用

四、查詢問題特性之分析

第六屆CLIR跨語言資訊檢索評估項目，共有20個來自世界各國的研究團隊參與各項檢索項目。在C、J、K、E等4種語言的可能配對之下，研究團隊總計送出分屬12種語言組合的152個檢索結果（Runs），但其中有4組結果不合理，可能源自程式之誤，遂將其排除不列入分析，因此研究團隊的總數量為19，檢索結果共計148組。本研究後續的分析，皆以95%的信心水準（1-α），也就是5%顯著水準（α），進行各項的統計檢定。

㈠查詢主題文本特性之分析

本研究將50個查詢主題根據前文所述之涵蓋地域、時間明確性、空間明確性、主題類別、條件敘述明確性及數量五個文本特性，區分其所屬之類別，並以Rigid relevance計算相關程度之方式比較各文本特性對於檢索效益之影響。

1. 涵蓋地域：全球範疇、區域國家、單一國家（3 Categories）

- 全球範疇：⑴沒有明確指出涵蓋地域，為一世界性的趨勢及議題。⑵牽涉範圍廣泛，事件發生地點超過一個洲或區域。
- 區域國家：⑴明確指出屬於同一洲之事件。⑵牽涉範圍超過一個國家，但不至於構成全球性議題。
- 單一國家：⑴明確指出屬於單一國家事件。⑵提及的人物或事件可界定

歸屬於單一國家。

在50個查詢主題中，全球範疇的查詢主題最多，有37題，佔全部題數的74%，平均表現為0.2589；另有8題區域國家的查詢主題，平均表現為0.2398；最少的是單一國家的查詢主題，只有5題，平均表現為0.2133，基本統計資料請參考表2。透過ANOVA單因子變異數分析，檢定涵蓋地域的三個類別對於檢索效益是否存在差異。檢定結果顯示涵蓋地域的不同類別並未達顯著水準（$F = 0.284$, $df = 2$, $p = .754$），亦即沒有證據顯示不同類別的涵蓋地域範疇會對檢索效益造成影響，請參考表3的統計檢定表。

表2 涵蓋地域—觀察值摘要

類別	題數	%	MAP均數	MAP標準差
全球範疇	37	74	.2589	.1415
區域國家	8	16	.2398	.1187
單一國家	5	5	.2133	.1031
總　和	50	100	.2513	.1334

表3 三種涵蓋地域類別之單因子變異數分析摘要

變異來源	df	SS	MS	F
組間	2	.0104	5.208E-03	.284
組內	47	.8620	1.834E-02	
總　和	49	.8730		

2. 時間明確性：明確、不明確（2 Categories）

- 明確：在查詢主題中明確指定事件發生的時間。
- 不明確：在查詢主題中沒有明確指定事件發生的時間。

在50個查詢主題中，有40題查詢主題的時間不明確，平均表現為0.2601；另外10題為時間明確的查詢主題，平均表現為0.2160，如表4所示。

再以獨立樣本T檢定驗證時間明確性對於檢索效益是否存在差異。檢定結果顯示該2樣本之母體變異數可視為相同，而查詢主題中時間敘述的明確性對於檢索效益沒有顯著差異（$t = .933$, $df = 48$, $p = .356$），亦即查詢主題之中對於時間的敘述明確與否不會造成檢索效益的不同。

表4 時間明確性—觀察值摘要

類別	題數	%	MAP均數	MAP標準差
時間不明確	40	80	.2601	.1376
時間明確	10	20	.2160	.1150
總　和	50	100	.2513	.1334

3. 空間明確性：明確、不明確（2 Categories）

- 明確：在查詢主題中明確指定事件發生的空間。
- 不明確：在查詢主題中沒有明確指定事件發生的空間。

在50個查詢主題中，有36題空間敘述不明確的查詢主題，平均表現為0.2726，另外14題為空間敘述明確的查詢主題，平均表現只有0.1966，請參見表5。

表5　空間明確性—觀察值摘要

類別	題數	%	MAP均數	MAP標準差
空間不明確	36	72	.2726	.1381
空間明確	14	28	.1966	.1062
總　和	50	100	.2513	.1334

從獨立樣本T檢定的檢定結果，可看出查詢主題中對於空間敘述的明確性對於表現分數沒有顯著差異（$t = 1.852$, $df = 48$, $p = .070$）。前述結論是以95%信心水準進行檢定的結果，若以90%信心水準進行檢定，結論便不相同，這顯示空間的明確性有某種程度的影響，端視研究者採取的顯著水準而定。

4. 主題類別

- 政治：查詢主題涉及國家或國際間的政治議題。
- 財經：查詢主題涉及財經議題。
- 社會綜合（含教育、環保）：查詢主題涉及社會議題，教育與環保議題也隸屬於社會綜合類別。
- 生活（含醫藥）：查詢主題論及生活議題，醫藥相關文件也隸屬於生活類別。
- 科技資訊：查詢主題論及科技及資訊議題。
- 藝文：查詢主題論及藝文活動。
- 體育：查詢主題論及體育競賽、活動或人物。
- 娛樂（含旅遊）：查詢主題論及娛樂設施、娛樂活動、影藝新聞及旅遊資訊。

在50個查詢主題中，最多為社會綜合類的查詢主題，共有16題，次多為科技資訊類的查詢主題12題。其中以科技資訊類的查詢主題表現最佳，平均分數為0.3740，表現最差則為社會綜合類的查詢主題，平均分數僅0.1961，而在50個查詢主題中，並不包含藝文、體育類的查詢主題，如表6所示。

表6　主題類別—觀察值摘要

主題類別	題數	%	MAP均數	MAP標準差
社會綜合	16	32	.1961	.1247
科技資訊	12	24	.3740	.1194
政治	8	16	.2060	.1022
財經	7	14	.2027	.0910
生活	4	8	.2714	.1254
娛樂	3	6	.2617	.1554
總　和	50	100	.2513	.1334

再以ANOVA單因子變異數分析，檢定六個不同主題對於表現分數是否存在差異。檢定結果發現，六個不同類別的主題對於表現分數有顯著差異（$F = 3.825$, $df = 5$, $p = .006$），請參見表7；變異數同質性Levene檢定結果，假設變異數相同（$p = .754 > .05$），進一步利用Scheffe法進行事後多重比較，結果顯示，組別間的差異的確存在，科技資訊的查詢主題（$m = .3740$）表現分數高於社會綜合查詢主題（$m = .1961$, $p = .016$），但其他類別主題的表現分數並無顯著差異。

表7 不同主題類別之單因子變異數分析摘要

變異來源	df	SS	MS	F
組間	5	.264	5.287E-02	3.825**
組內	44	.608	1.382E-02	
總 和	49	.873		

**p < .01

5. 條件敘述明確性及數量

界定一個明確的條件敘述情形包括：提及明確的專有名詞、有範圍限制的敘述句，及使用某些特定詞彙，例如「定義」、「案例」及「政策」。此外，條件敘述還可分為相關、部分相關及不相關三個層次。將三個層次明確的條件敘述個數加總，可反映該查詢主題的複雜度，條件敘述個數越少，則複雜度越低；條件敘述個數越多，則複雜度越高。

在50個查詢主題中，最多為包含4個條件敘述的查詢主題，共有22題，其中又以包含6個條件敘述詞彙的查詢主題表現最佳，平均分數為0.3857；表現最差則為包含8個條件敘述的查詢主題，平均分數僅有0.0915，但此兩項皆只有一組檢索結果，數據代表性不足，請參見表8。

表8 條件敘述總個數—觀察值摘要

條件敘述總個數	題數	%	MAP均數	MAP標準差
8個	1	2	.0915	–
7個	2	4	.3378	.0604
6個	1	2	.3857	–
5個	4	8	.1676	.0976
4個	22	44	.2799	.1577
3個	12	24	.2415	.1057
2個	8	16	.2107	.1045
總 和	50	100	.2513	.1334

若進一步考量敘述明確與否，則表現最佳為包含4個明確條件敘述詞的查詢主題，平均分數為0.2844，從平均分數看來，一個查詢主題中包含多於或少於4個明確的條件敘述詞彙，皆會降低檢索效益的表現，請參見表9。

另一方面，包含4個不明確條件敘述詞的查詢主題表現最佳，檢索效益的

平均分數為0.3570，若少於或多於4個不明確的條件敘述詞彙，也會降低檢索效益的表現，請參見表10。

表9　明確條件敘述個數—觀察值摘要

明確條件敘述個數	題數	%	MAP均數	MAP標準差
6個	1	2	.0915	–
5個	1	2	.0421	–
4個	3	6	.2844	.2236
3個	12	24	.2286	.1346
2個	14	28	.2805	.1211
1個	13	26	.2468	.1394
0個	6	12	.2835	.1021
總　和	50	100	.2513	.1334

表10　不明確條件敘述個數—觀察值摘要

不明確條件敘述個數	題數	%	MAP均數	MAP標準差
5個	3	6	.3116	.1239
4個	2	4	.3570	.0658
3個	6	12	.2557	.1656
2個	15	30	.2816	.1211
1個	17	34	.2225	.1170
0個	7	14	.1965	.1798
總　和	50	100	.2513	.1334

從平均分數雖可看出表現分數之高低，但還要進一步以相關分析檢驗條件敘述個數與表現分數之間是否存在正負向關係。本研究進行以下三組相關分析以檢定檢索結果，三組相關分析結果皆未達顯著水準：

- 條件敘述總個數v.s.表現分數（Pearson r = – .262, p = .066）。
- 明確條件敘述個數v.s.表現分數（Pearson r = – .193, p = .180）。
- 不明確條件敘述個數v.s.表現分數（Pearson r = .264, p = .064）。

也就是說一個查詢主題具有的條件敘述個數多寡，與檢索效益不存在顯著的正向或負向關係，雖然有一些趨勢，如不明確條件敘述個數有些許正相關的傾向。只能說明一個查詢主題若包含不多不少的不明確條件敘述詞彙，或許可以產生較佳的檢索表現，本研究的資料分析結果顯示，這裡的不多不少是4個。

為進一步觀察上述統計分析結果，從148組檢索表現的平均分數，排序出前20%表現最佳的查詢主題，在這十個查詢主題中，最常見的特色為全球範疇、時間空間皆不明確、科技資訊類的查詢主題，並包含2個不明確條件敘述、2～3個明確條件敘述，如表11所示。以同樣的方式觀察表現差的十個查詢主題，最常見的特色為全球範疇、時間空間皆不明確、社會綜合主題，並包含0～1個不明確的條件敘述、3個明確的條件敘述，如表12所示。

表11 前20%表現佳的查詢主題之文本特性分布

Topic	MAP均數	涵蓋地域	時間	空間	主題類別	條件敘述個數	
						明確	不明確
21	0.5535	全球範疇	不明確	不明確	科技資訊	1	3
74	0.5025	全球範疇	不明確	不明確	科技資訊	4	0
53	0.4788	全球範疇	不明確	不明確	社會綜合	3	1
75	0.4509	全球範疇	不明確	不明確	科技資訊	2	2
58	0.4413	全球範疇	不明確	不明確	科技資訊	2	2
59	0.4407	全球範疇	不明確	不明確	科技資訊	2	2
42	0.4276	全球範疇	不明確	不明確	科技資訊	3	1
80	0.4035	區域國家	明確	明確	財經	0	4
77	0.3976	單一國家	不明確	不明確	娛樂	0	2
36	0.3966	全球範疇	不明確	不明確	生活	3	0

表12 後20%表現差的查詢主題之文本特性分布

Topic	MAP均數	涵蓋地域	時間	空間	主題類別	條件敘述個數	
						明確	不明確
26	0.1416	區域國家	不明確	明確	政治	2	0
17	0.1252	全球範疇	不明確	不明確	生活	3	1
37	0.1225	全球範疇	明確	明確	社會綜合	2	1
27	0.1102	區域國家	不明確	明確	政治	3	1
100	0.0923	全球範疇	明確	不明確	娛樂	3	0
103	0.0915	全球範疇	不明確	不明確	社會綜合	6	2
105	0.0794	全球範疇	不明確	不明確	社會綜合	1	1
18	0.0590	全球範疇	不明確	不明確	社會綜合	1	3
110	0.0556	全球範疇	不明確	不明確	社會綜合	4	0
19	0.0421	全球範疇	不明確	明確	社會綜合	5	0

從表現佳與表現差的查詢主題具有之特性，可再次驗證前述查詢主題分析的結果，亦即涵蓋地域、時間明確性、空間明確性、明確/不明確條件敘述個數，此四項查詢主題特性，對於檢索表現皆不具顯著的影響力。

㈡語言特性之分析

1. 各種語言的組合情形

CLIR檢索項目包括單語言（X-X）、雙語言（X-Y）或多語言（X-XYZ）的組合，連字符號前的語言代表查詢問題的語言（問題集的語言），連字符號後的語言代表檢索文件的語言（文件集的語言），以下為第六屆NTCIR CLIR項目研究團隊送交之檢索結果的語言組合情形，C-CJK、C-C、C-J、E-C、E-J、E-K、J-C、J-J、J-K、K-C、K-J、K-K（其中C＝繁體中文，J＝日文，K＝韓文，E＝英文）。

由表13可知單語言（X-X）檢索結果最多為J-J的43組，其次為C-C的32組，

之後為K-K的20組。雙語言（X-Y）檢索結果最多為C-J的17組，其次為E-J的9組。多語言（X-XYZ）只有2個檢索結果。

另外，在單一語言中，以K-K的表現分數最高，平均分數為0.3846。在雙語言中，以J-K的表現分數最高，平均分數為0.3043，多語言只有一種組合C-CJK，平均分數為0.0644，同時也是12個語言組合中表現最差的一組。

表13　不同語言組合之檢索效益

語言別	No. of Runs	%	MAP均數	MAP標準差
單語言				
J-J	43	29.1	.2554	.0591
C-C	28	18.9	.2249	.0624
K-K	20	13.5	.3846	.0950
雙語言				
C-J	17	11.5	.2140	.1003
E-J	9	6.1	.2666	.0388
E-C	8	5.4	.1472	.0401
E-K	5	3.4	.2789	.0288
J-K	5	3.4	.3043	.0302
K-J	5	3.4	.2516	.0111
J-C	4	2.7	.1162	.0638
K-C	2	1.4	.1070	.0682
多語言				
C-CJK	2	1.4	.0644	.0084
總合	148	100	.2511	.0947

2. 各種語言的檢索效益

如前文所述，在CLIR評估項目中，參與評估的研究團隊可選用不同語言的文件集與問題集進行檢索，本研究分析各語言組合的表現，企圖找出語言差異對於檢索效益之影響，也就是當研究團隊使用不同語言的文件集或問題集是否會影響檢索效益的表現。從各語言組合檢索效益的檢定結果，可看出選用不同語言的文件集與問題集對於表現分數確實有顯著差異（$F = 13.599$, $df = 11$, $p = .000$），請參見表14。

變異數同質性Levene檢定的結果，假設變異數不相同（$p = .009 < .05$），再透過Games-Howell事後成對比較檢定顯示，K-K表現最佳，平均分數顯著高於C-CJK、C-C、C-J、E-C、E-J、E-K、J-C、J-J、K-C、K-J等十組的平均分數，在十二種語言組合中的表現最佳；相較之下，C-CJK在十二種語言組合中是表現最差的組合，平均分數顯著低於C-C、C-J、E-C、E-J、E-K、J-J、J-K、K-J、K-K等九組的平均分數。

十二種語言組合的平均分數由高至低依序為：K-K、J-K、E-K、E-J、J-J、K-J、C-C、C-J、E-C、J-C、K-C、C-CJK，很顯然地，相同的問題語言，單語檢索效益比雙語檢索高，雙語檢索效益比多語檢索效益高。

3. 各種語言問題集的表現

我們將12種檢索語言組合依據問題集之語言分為四組，找出表現最佳的問題集。採用中文問題集包括C-CJK、C-C、C-J，共有47組檢索結果，平均表現為0.2142；英文問題集包括E-C、E-K、E-J，共有22組檢索結果，平均表現為0.2260；日文問題集包括J-C、J-J、J-K，共有52組檢索結果，平均表現為0.2494；韓文問題集包括K-C、K-J、K-K，共有27組檢索結果，平均表現為0.3394。ANOVA統計檢定顯示不同語言的問題集對於檢索效益確實有顯著差異（$F = 13.467, df = 3, p = .000$），請參見表15。

表14 不同語言組合之變異數分析摘要

變異來源	df	SS	MS	F
組間	11	.690	.0628	13.599***
組內	136	.628	.0046	
總　和	147	1.318		

***p < .001

表15 不同語言問題集之變異數分析摘要

變異來源	df	SS	MS	F
組間	3	.289	.0963	13.467***
組內	144	1.029	.0071	
總　和	147	1.318		

***p < .001

依據變異數同質性Levene檢定的結果，母體變異數不相同（p = .006 < .05），再使用Games-Howell事後成對比較檢定顯示，選用韓文的問題集，平均表現分數（m = .3394）顯著高於日文問題集（m = .2494, p = .004）、英文問題集（m = .2260, p = .001），及中文問題集（m = .2142, p = .000）。顯示在四種問題集的語言中，以韓文問題集的檢索效益表現最佳，日文問題集次之，中文問題集表現最差。

4. 各種語言文件集的表現

將12種檢索語言組合依據文件集之語言分為四組，找出表現最佳的文件集。多語文件集包括C-CJK，共有2組檢索結果，平均表現為0.0644；中文文件集包括C-C、E-C、J-C、K-C，共有42組檢索結果，平均表現為0.1942；日文文件集包括C-J、E-J、J-J、K-J，共有74組檢索結果，平均表現為0.2470；韓文文件集包括E-K、J-K、K-K，共有30組檢索結果，平均表現為0.3536。

經過統計檢定顯示選用不同語言的文件集對於表現分數確實有顯著差異（$F = 31.519, df = 3, p = .000$），請參見表16。依據變異數同質性Levene檢定的結果，母體變異數相同（$p = .096 > .05$），再使用Scheffe事後成對比較檢定顯示，選用韓文文件集，平均表現分數（$m = .3536$）顯著高於日文文件集（$m = .2470, p = .000$）、中文文件集（$m = .1942, p = .000$），及多語文件集（$m = .0644, p$

＝.000）；另一方面，選用日文文件集的平均表現分數（m＝.2470）顯著高於中文文件集（m＝.1942, p＝.005），及多語文件集（m＝.0644, p＝.010）。

表16　不同語言文件集之變異數分析摘要

變異來源	df	SS	MS	F
組間	3	.522	.1740	31.519***
組內	144	.796	.0055	
總　和	147	1.318		

***p＜.001

上述統計檢定結果可看出在四種文件集語言中，以韓文文件集的表現最佳，日文文件集次之，多語文件集表現最差。

㈢欄位特性之分析

1. 檢索欄位的使用情形

第六屆CLIR評估項目的148組檢索結果，使用之檢索欄位可為標題（Title, <T>）、簡短資訊需求（Description, <D>）、詳細資訊需求（Narrative, <N>）、相關概念（Concepts, <C>），或欄位的組合。總計各研究團隊產生的檢索結果包括T Run、D Run、N Run、C Run、DN Run及TDNC Run六種，其中以N Run表現最佳（m＝.4631），其次為C Run（m＝.3921）、DN Run（m＝.2705），表現最差為TDNC Run（m＝.2339），請參見表17。但N Run及C Run皆只有一組檢索結果，數據代表性不足，因此在後續分析中，此兩組檢索結果將略而不計。

表17　研究團隊選擇之檢索欄位—觀察值摘要

執行檢索欄位	No. of Runs	%	MAP均數	MAP標準差
D	61	41.2	.2460	.0866
T	57	38.5	.2506	.0879
N	1	0.7	.4631	–
C	1	0.7	.3921	–
DN	13	8.8	.2705	.0667
TDNC	15	10.1	.2339	.1477
總　和	148	100	.2511	.0947

2. 檢索欄位的檢索效益

如前文所述，參與評估的研究團隊可選擇使用不同的欄位進行檢索，本研究分析<D>、<T>、<DN>、<TDNC>等欄位或欄位組合的表現，檢視檢索欄位的差異是否會影響檢索的表現。雖然表17顯示各欄位之檢索效益有些許差異，但應用ANOVA統計檢定（排除<N>與<C>欄位），顯示並沒有顯著效果，請參見表18，可看出使用不同的查詢主題欄位進行檢索對於檢索效益沒有顯著影響（F＝.386, df＝3, p＝.763）。

表18 檢索欄位之變異數分析摘要

變異來源	df	SS	MS	F
組間	3	.010	.0034	.386
組內	142	1.242	.0087	
總 和	145	1.252		

㈣研究團隊之分析

1. 研究團隊檢索表現

如上文所述，總計有19個研究團隊從結構化的查詢主題中選擇用以檢索的欄位或欄位組合，總共產生6種不同的欄位或欄位組合情形；另外，研究團隊也可選用不同語言的文件集與問題集進行檢索工作，總共產生了12種不同的語言組合。各研究團隊檢索效益的表現如表19所示。使用ANOVA進行統計檢定，可看出研究團隊檢索效益的表現的確存有顯著差異（$F = 9.707$, $df = 18$, $p = .000$）。變異數同質性Levene檢定結果顯示，變異數不相同（$p = .000 < .05$），再使用Games-Howell事後成對比較檢定，顯示KLE表現最佳（$m = .3707$），其次為UniNE（$m = .3345$）及TSB（$m = .3200$），請參見表20。

本研究檢定各研究團隊的檢索效益之後，發現研究團隊之間的檢索效益的確有顯著差異，因此接下來將分析研究團隊與檢索欄位以及檢索語言的交互作用，探討檢索欄位與語言組合的表現差異，是否是因為研究團隊的不同而產生。

表19 研究團隊之檢索效益—觀察值摘要

研究團隊	No. of Runs	%	MAP均數	MAP標準差
AINLP	2	1.4	.0835	.0306
BRKLY	16	10.8	.2058	.0806
CCNU	2	1.4	.2618	.0089
CYUT	2	1.4	.0644	.0084
HUM	15	10.1	.2112	.1083
I2R	4	2.7	.2539	.0668
IASL	2	1.4	.1070	.0068
ISQUT	3	2.0	.1066	.0139
JSCCL	4	2.7	.2515	.0419
KLE	8	5.4	.3707	.0958
NCUTW	10	6.8	.2188	.0332
NICT	35	23.6	.2850	.0588
OASIS	2	1.4	.0624	.0123
OKSAT	5	3.4	.1809	.0176
pircs	4	2.7	.2096	.0486
TSB	12	8.1	.3200	.0209
UniNE	15	10.1	.3345	.0694
WTG	4	2.7	.1537	.0364
YLMS	3	2.0	.2882	.0261
總 和	148	100	.2511	.0947

表20　研究團隊檢索效益之變異數分析摘要

變異來源	df	SS	MS	F
組間	18	.758	.0421	9.707***
組內	129	.560	.0043	
總　和	147	1.318		

***p < .001

2. 交互作用對檢索效益影響之分析

各研究團隊可選擇檢索6種欄位組合（<T>、<D>、<N>、<C>、<DN>、<TDNC>）、12種語言組合（C-CJK、C-C、C-J、E-C、E-J、E-K、J-C、J-J、J-K、K-C、K-J、K-K）。經由兩兩交互作用的二因子變異數分析可知，研究團隊與語言組合對於檢索效益之交互作用效果不顯著（$F(12,106)=.746$, $p=.703$），請參見表21；而研究團隊與檢索欄位對於檢索效益則有顯著交互作用（$F(25, 99)=1.768$, $p=.026$），請參見表22。

表21　研究團隊與語言組合之二因子變異數分析摘要

變異來源	df	SS	MS	F
組間				
研究團隊	16	.376	.0235	10.949***
語言組合	9	.306	.0340	15.830***
交互作用	12	.019	.0016	.746
組內（誤差）	106	.228	.0021	
總　和	145	1.252		

***p < .001

表22　研究團隊與檢索欄位之二因子變異數分析摘要

變異來源	df	SS	MS	F
組間				
研究團隊	18	.738	.0410	10.895***
檢索欄位	3	.006	.0021	.569
交互作用	25	.166	.0067	1.768*
組內（誤差）	99	.373	.0038	
總　和	145	1.252		

***p < .001, *p < .05.

進一步以單因子變異數分析檢定檢索欄位在研究團隊的單純主效果，結果發現，研究團隊使用<T>為檢索欄位之57組檢索結果，檢索效益有顯著差異（$F=4.576$, $df=18$, $p=.000$），其中以KLE的表現最佳（$m=.3445$, $sd=.1119$）；研究團隊使用<D>為檢索欄位之61組檢索結果，檢索效益亦有顯著差別（$F=5.046$, $df=18$, $p=.000$），以UniNE的表現最佳（$m=.3407$, $sd=.0809$）；研究團隊使用<TDNC>為檢索欄位之15組檢索結果，檢索效益亦有顯著差異（$F=14.425$, $df=3$, $p=.000$），以KLE的表現最佳（$m=.4789$）；研究團隊使用<DN>

為檢索欄位之13組檢索結果，檢索效益則無顯著的差異（$F = 1.371$, $df = 4$, $p = .325$）。

另一方面，本研究分析個別研究團隊的檢索結果，檢驗是否在某些特定欄位的表現較好。分析結果顯示，HUM使用<T>欄位（$m = .2688$）或<D>欄位（$m = .2460$）都比使用<TDNC>欄位組合（$m = .0266$）的檢索表現優異；ISQUT使用<T>欄位（$m = .1146$）比使用<D>欄位（$m = .0905$）的檢索表現優異；YLMS使用<T>欄位（$m = .3182$）比使用<D>欄位（$m = .2731$）的檢索表現優異；其他研究團隊選擇不同欄位後，並沒有明顯的表現差異。此項檢定結果與前文所述之四種檢索欄位的平均分數由高至低的排序：<DN>、<T>、<D>、<TDNC>相符，也就是說，各研究團隊選擇不同的檢索欄位，但在檢索表現上是一致的。

再以同樣的方式檢定各個研究團隊在特定語言組合是否有比較好的表現，統計檢定的結果如下所示。

- BRKLY執行J-J（m = .2716）優於C-C（m = .2176）與J-C（m = .1162）；
- I2R執行C-C（m = .3116）優於E-C（m = .1962）；
- pircs執行C-C（m = .2513）優於E-C（m = .1679）；
- UniNE執行K-K（m = .4270）優於J-J（m = .2941）且優於C-C（m = .2822）；
- WTG執行C-C（m = .1852）優於E-C（m = .1222）；
- NICT執行K-K（m = .4032）優於J-K（m = .3043）、E-K（m = .2789）、J-J（m = .2763）、K-J（m = .2516）、C-J（m = .2432）、E-J（m = .2375）；
- KLE執行K-K（m = .4437）優於J-J（m = .2597）；

其他研究團隊選擇不同語言組合後，並沒有明顯的表現差異。檢定結果與上文中十二種語言類別的平均分數由高至低的排序：K-K、J-K、E-K、E-J、J-J、K-J、C-C、C-J、E-C、J-C、K-C、C-CJK大致相符，也就是說，雖各研究團隊選擇了不同的語言組合，但不同語言組合在檢索效益的表現上是一致的。

五、結　　論

本研究分別由查詢主題的文本特性、語言特性，及欄位特性等三個面向，檢視第六屆NTCIR CLIR檢索評估項目的資料，試圖找出查詢主題特性對檢索效益之影響。為避免這些因素受研究團隊本身檢索技術的制約，導致分析結果的謬誤，本研究亦探討這些面向與研究團隊的交互作用。

在查詢主題文本特性方面，以涵蓋地域、時間明確性、空間明確性、主題、條件敘述明確性及數量五個文本特性，分析50個查詢主題的內容。研究發現僅有「主題類別」對檢索效益有影響，其中又以科技資訊類的表現顯著優於國際事件類；其他四項文本特性對檢索效益並無顯著影響。

在語言特性方面，第六屆NTCIR CLIR有中日韓英四種語言，研究團隊提

交的檢索結果共有12種不同語言的組合。研究發現問題集與文件集皆以韓文的表現最佳，表示在CLIR機制中，選用不同的問題集與文件集對於檢索結果的確會造成影響，但導致不同文件集/問題集在檢索效益產生差異的因素，仍有待進一步確認，可能是研究團隊在跨語言檢索的過程，使用的翻譯機制不同，因此造成不同語言的檢索效益產生差異；也可能是某些語言文件集的特性，如相關文件特別多，造成該語言的檢索效益特別高。另外，CLIR提供的問題集是以英文為中間語言，再翻譯為中文、日文與韓文，轉譯成不同語言的過程可能產生完整性的不足或詮釋的不同，進而造成不同語言檢索效益的差異。

在欄位特性方面，本研究發現不論使用標題（Title, <T>）、簡短資訊需求（Description, <D>）、詳細資訊需求（Narrative, <N>）、相關概念（Concepts, <C>），或各欄位的組合，對於檢索效益沒有顯著影響。一般而言，<T>欄位的長度小於<D>欄位，<D>欄位的長度小於<N>欄位，也就是說依據第六屆NTCIR CLIR的檢索資料的分析結果，本研究認為查詢問題的長度對於檢索效益沒有顯著的影響，這和Nelson（1995）的研究結論「查詢主題的內容長度與檢索結果平均查準率為正相關」並不相同。然而，本研究也顯示四種檢索欄位或欄位組合的檢索效益，由高至低的排序為<DN>、<T>、<D>、<TDNC>，即使其間的差異並不顯著。

在研究團隊與欄位特性及語言特性的交互作用的檢定，本研究驗證研究團隊與檢索欄位在檢索效益上有交互作用，也就是說研究團隊的不同與檢索欄位的不同對檢索效益會造成影響；而在語言組合的選擇上則無明顯的交互作用。

資訊檢索本身是一個複雜的課題，影響檢索效益的因素非常多，本研究以實證的方式，分析實際的跨語言資訊檢索系統產生的檢索結果，探討查詢問題對於檢索效益的影響，其結論仍有其侷限性，還有許多研究議題需要從事資訊檢索的學者與專家共同的努力。

誌　　謝

本研究得到國科會專題研究計畫「資訊檢索評估中檢索問題特性之探究」的補助，計畫編號為NSC95-2413-H-002-012。作者感謝二位匿名審查委員的建議與意見。

參考文獻

王怡人（2004）。國立台灣大學學生使用線上百科全書之資訊尋求行為：*Grolier Multimedia Encyclopedia*為例。未出版之碩士論文，國立臺灣大學圖書資訊學研究所，台北市。

江玉婷（1999）。中文資訊檢索測試集設計與製作之研究。未出版之碩士論文，國立臺灣

大學圖書資訊學研究所，台北市。

陳光華、江玉婷（2000）。中文資訊檢索測試集之設計與製作。資訊傳播與圖書館學，*6*（3），61-80。

陳明君（1999）。檢索背景與查詢問題對檢索技巧及檢索結果之影響研究。未出版之碩士論文，國立臺灣大學圖書資訊學研究所，台北市。

黃怡如（1999）。終端使用者與系統互動前後查詢問題、檢索概念與檢索詞彙變化之研究。未出版之碩士論文，國立臺灣大學圖書資訊學研究所，台北市。

Buckley, C. (2000). *The TREC-9 query track*. Retrieved December 24, 2008, from http://trec.nist.gov/pubs/trec9/papers/query_track.pdf

Derr, R. L. (1982). A classification of questions in information retrieval by conceptual presupposition. *Proceedings of the 45th Annual Meeting of the American Society for Information Science, 19*, 69-71.

Fidel, R., & Soergel, D. (1983). Factors affecting online bibliographic retrieval: A conceptual framework for research. *Journal of the American Society for Information Science, 34*(3):163-180.

Graesser, A. C., & Murachver, T. (1985). Symbolic procedures of question answering. In A.C. Graesser, & J. B. Black (Eds.), *The psychology of questions* (pp. 15-88). Lawrence Hillsdale, NJ: Erlbaum Associates.

Keyes, J. G. (1996). Using conceptual categories of questions to measure differences in retrieval performance. *Proceedings of the 59th Annual Meeting of the American Society for Information Science, 33*, 238-242.

Kishida, K., Chen, K. H., Lee, S., Kuriyama, K., Kando, N., & Chen, H. H. (2007, May). Overview of CLIR task at the sixth NTCIR workshop. In *Procedings of NTCIR-6 Workshop Meeting*, Symposium held at the Tokyo, Japan.

Lancaster, F. W. (1968). *Evaluation of the MEDLARS demand search service*. Bethesda, Md.: National Library of Medicine.

Lehnert, W. G. (1978). *The process of question answering: A computer simulation of cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nelson, M. J. (1995). The effect of query characteristics on retrieval results in the TREC retrieval tests. In *Proceedings of the 23rd annual conference of the canadian association for information science (CAIS 95)* (pp. 156-163). Canada: University of Alberta.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving II: Users, questions, and effectiveness. *Journal of the American Society for Information Science, 39*(3), 177-196.

Saracevic, T., & Baxter, M. A. (1983). On a method for studying the structure and nature of requests in information retrieval. *Proceedings of the 46th Annual Meeting of the American Society for Information Science, 20*, 22-25.

Journal of Educational Media & Library Sciences 46 : 2 (Winter 2008): 183-210

Influences of Query Characteristics to Retrieval Performance in Cross-Language Information Retrieval

Kuang-Hua Chen*

Associate Professor
E-mail: khchen@ntu.edu.tw

Tien-An Wu

Graduate Student
Department of Library and Information Science
National Taiwan University
Taipei, Taiwan
E-mail: r95126015@ntu.edu.tw

Abstract

This study investigated the influences of structured query on retrieval performance from textual, linguistic, and fielded characteristics. The search runs of NCTIR-6 CLIR Task have been used as the targeted data for this study. The results concluded that only subject out of other textual characteristics demonstrated significant effects on retrieval performance; linguistic characteristics showed great impacts on retrieval performance; fielded information did not show significant effects on retrieval performance but some fielded information performed better than others. This study also made clear that there existed interactions between fielded information in queries and participating teams (consequently the retrieval techniques used by these teams) but no significant interactions between linguistic characteristics and participating teams.

Keywords: *Information need; Query; Question; Retrieval performance; Topic*

SUMMARY

1. Research Design and Data

The research on information retrieval evaluation has become more and more important in recent years. TREC, CLEF, and NTCIR are the top information retrieval evaluation conferences in the world with focuses on English, European languages, and Asian languages, respectively. In general, IR evaluation is carried out based on the "benchmark" test collections which consist of document set, topic set (question set), and answer set (relevance judgments). Apparently, topic set is used by participating teams to generate queries for search tasks. Hereafter, the terms question, topic, and query may be used interchangeably. Therefore, it is likely that the characteristics of topics have impacts on the retrieval performance. This study uses data created in NTCIR6 Cross-Language Information Retrieval (CLIR) task to investigate the im-

* Principal author for all correspondence.

pacts of textual characteristics, linguistic characteristics, and fielded characteristics of topics on the retrieval performance for Cross-Language Information Retrieval.

The 50 topics which were created by task organizers for NTCIR6 CLIR task and the search runs which were submitted by many research teams to NTCIR6 CLIR task are used to examine the impacts. NTCIR6 CLIR task used news articles as the document set. The reasons to use news articles are manifold. Some major reasons are the variability of topics and focuses, the coverage of language patterns, and the similarity of search contents. Another important consideration is that the assessors could easily and correctly judge the relevance of news articles.

(1) NTCIR6 CLIR Task

The document languages and topic languages of NTCIR6 CLIR task are Chinese (C), Japanese (J), Korean (K), and English (E). The participating teams could determine the combination of document languages and topic languages while carrying out their own search tasks. Each topic consists of <Title> (<T>), <Description> (<D>), <Narrative> (<N>), and <Concepts> (<C>) fields. Basically, the participating teams could choose the fields (in topics or questions) to carry out their search tasks and submit their search runs. However, each participating team has to submit two mandatory runs which are <Title> run (<T> run) and <Description> run (<D> run). Other submitted runs could be any combinations of <T>, <D>, <N>, and <C>. Each search run has an identifier (shown in the following) composed of the team's ID, topic language, document languages, topic fields for search, and Priority.

• Group's ID-Topic Language-Document Language-Topic Field-pp

The 'pp' is two digits used to represent the priority of the run. It is used as a parameter for pooling. The participants have to decide the priority for each submitted run on the basis of each language pair. "01" means the highest priority. For example, a participating team, LIPS, submits 3 runs for C-->CJE searches. The first is a T run, the second is a D run, and the third is a DN run. Therefore, the Run ID for each run is LIPS-C-CJE-T-01, LIPS-C-CJE-D-02, and LIPS-C-CJE-DN-03, respectively. The 152 search runs submitted by 20 participating teams in NTCIR6 CLIR task are categorized into C-CJK, C-C, C-J, E-C, E-J, E-K, J-C, J-J, J-K, K-C, K-J, and K-K language combinations.

Relevance judgments are carried out by humans. Each document for each topic is assigned one of four categories of relevance: "Highly Relevant (S)", "Relevant (A)", "Partially relevant (B)", and "Irrelevant (C)". Each kind of relevance is assigned a relevance score. "Highly relevant" is 3, "Relevant" is 2, "Partially relevant" is 1, and "Irrelevant" is 0. Therefore, we have two sets of relevance judgments. One is "Rigid Relevance"; the other is "Relaxed Relevance". The "Highly Relevant" and "Relevant" are regarded as relevance in "Rigid Rel-

evance". The "Highly Relevant", "Relevant", and "Partially Relevant" are regarded as relevance in "Relaxed Relevance". This study uses "Rigid Relevance" for analyses and discussions.

(2) Topic Characteristics v.s. Retrieval Performance

This study investigates the impacts of textual, linguistic, and fielded characteristics of topics on the retrieval performance of cross language information retrieval. In order to avoid the influences of retrieval systems that created these search runs, we also examine the interactions among characteristics and retrieval systems.

We first analyze the textual characteristics for 50 topics and compare the rigid relevance score to figure out the influence of textual characteristics. We propose 5 textual characteristics for this study. These characteristics are Geographic Coverage, Temporal Clarity, Spatial Clarity, Subject Category, and Number and Clarity of Restrictions.

The combinations of topic languages and document languages could have impacts on the retrieval performance. Examining the different combinations helps analyze the influences coming from topic languages, document languages, and the combinations of the two.

Since the participating team could determine the fields for search tasks, we would like to discuss the application of different fields and the influences coming from different fields.

Different participating teams would adopt different approaches to developing their retrieval systems. The different technologies involved in retrieval systems have different influences on the retrieval performance. This study would also like to investigate the influences coming from different retrieval systems and the interactions among characteristics and retrieval systems.

2. Analyses of Topic Characteristics

This study investigates 152 search runs submitted by 20 participating teams. The number of language combinations of these submitted runs is 12. However, four runs submitted by one team are unreasonable due to the various problems. As a result, only 148 search runs submitted by 19 participating teams are used for investigation and analysis. In addition, all the significant tests used in this study are based on 5% significant level (α=5%).

(1) Analysis for Textual Characteristics

For the Geographic Coverage, 74% of the topics are international and the average retrieval performance is 0.2589; 16% of the topics are regional and its performance is 0.2398; 10% of the topics are domestic and its performance is 0.2133. No evidence shows Geographic Coverage has significant influence on the retrieval performance based on ANOVA test (F=0.284, df=2, p=.754).

As to the Temporal Clarity, temporal description of 80% of the topics is not

clear. The retrieval performance is 0.2601. The remaining 20% of the topics are clear in temporal description and the average performance is 0.2160. No evidence shows Temporal Clarity has significant influence on the retrieval performance based on T test (t=.933, df=48, p=.356).

Considering Spatial Clarity, 36 out of 50 topics are not clear in Spatial Clarity. The retrieval performance is 0.2726. The other 14 topics are clear in Spatial Clarity. The retrieval performance is 0.1966. No evidence shows Spatial Clarity has significant influence on the retrieval performance based on T test (t=1.852, df=48, p=.070).

As to the Subject Category, 16 out of 50 topics are social and 12 topics are information technology related. IT-related topics outperformed other topics. The average performance is 0.3740. The retrieval performance of Social topics is 0.1961. Subject Category has significant influence on the retrieval performance based on ANOVA test (=3.825, df=5, p=.006). We post test Subject Category using the Scheffe method. Significant differences exist between IT-related topics and Social topics, but no significant differences between other subject categories.

Considering the Number of Restrictions, 22 topics (44%) containing 4 restrictions show the best performance. The performance decreases while number of restrictions decrease or increase. Considering number and clarity together, containing four clear restrictions or four unclear restrictions, show good performance. In general, topics with unclear restrictions show better performance than topics with clear restrictions do. However, there exists no significant difference.

(2) Analysis for Linguistic Characteristics

The 152 search runs are categorized into 12 language combinations which are C-CJK, C-C, C-J, E-C, E-J, E-K, J-C, J-J, J-K, K-C, K-J, and K-K. For the single language search, the most submitted runs are J-J runs, C-C runs the second, and K-K runs the third. For the bi-lingual search, the most submitted runs are C-J runs (17 runs) and E-J runs the second (9 runs). For multilingual searches, only 2 C-CJK runs are submitted. K-K runs perform the best in single language search (0.3846); J-K runs perform the best in the bi-lingual search (0.3043); the performance of the multilingual search (C-CJK runs) is 0.0644.

According to the results of the ANOVA test, significant differences exist among submitted runs with different language combinations (F=13.599, df=11, p=.000). Using the Games-Howell method to carry out post tests, the performance of K-K runs is significantly better than other runs except J-K runs. The increasing order of performance for language combinations is K-K, J-K, E-K, E-J, J-J, K-J, C-C, C-J, E-C, J-C, K-C, and C-CJK. Obviously, for the same topic language, the performance of single language search is better than that of bi-lingual search; the performance of bi-lingual search is better than that of multilingual search.

(3) Analysis for Fielded Characteristics

The field combinations of submitted runs are <T> run, <D> run, <N> run, <C> run, <DN> run, and <TDNC> run. The <N> run shows the best performance (0.4631); the <C> run is the second (0.3921); the <DN> run is the third (0.2705); the <TDNC> run is the worst (0.2339). However, there are only 1 <N> run and 1 <C> run. We would not use them to do further analysis. According to result of the ANOVA test, there exists no significant difference in performance among submitted runs with different field combinations (F=.386, df=3, p=.763). However, there exists a tendency of precedence of fields in retrieval performance, i.e., <DN>, <T>, <D>, and <TDNC>.

(4) Interaction among Characteristics and Retrieval Systems

The search data used in this study are generated by 19 participating teams. According to results of the ANOVA test, there exists significant difference in performance among runs which were submitted by different participating teams (F=9.707, df=18, p=.000). Further investigation is necessary to analyze the interactions. According to 2-way ANOVA, the interaction between retrieval systems and linguistic characteristics is not significant ($F(12,106)$=.746, p=.703); the interaction between retrieval systems and fielded characteristics is significant ($F(25, 99)$=1.768, p=.026).

3. Conclusions

This study investigated the influences of structured queries on retrieval performance from textual, linguistic, and fielded characteristics. The search runs of NCTIR-6 CLIR Task have been used as the targeted data for this study. In order to avoid the influences of retrieval systems which created these search runs, we also examined the interactions among characteristics and retrieval systems.

The results concluded that only subject categories out of other textual characteristics demonstrated significant effects on retrieval performance; linguistic characteristics showed great impacts on retrieval performance; fielded information did not show significant effects on retrieval performance but some fielded information performed better than others. This study also made it clear that there existed interactions between fielded information in queries and retrieval systems but no significant interactions between linguistic characteristics and participating teams.

ROMANIZED & TRANSLATED REFERENCES FOR ORIGINAL TEXT

王怡人[Wang, Yi-Ren](2004)。國立台灣大學學生使用線上百科全書之資訊尋求行為：*Grolier Multimedia Encyclopedia*為例[*Guoli Taiwan Daxue xuesheng shiyong xianshang baikequanshu zhi zixun xunqiu xingwei: Grolier Multimedia Encyclopedia wei li*]。未出版之碩士論文，國立臺灣大學圖書資訊學研究所，台北市[Unpublished master dissertation, Graduate Institute of Library and Information Science of National Taiwan University, Taipei]。

江玉婷[Chiang, Yu-Ting] (1999)。中文資訊檢索測試集設計與製作之研究[*Zhongwen zixun jiansuo ceshiji sheji yu zhizuo zhi yanjiu*]。未出版之碩士論文，國立臺灣大學圖書

資訊學研究所，台北市[Unpublished master dissertation, Graduate Institute of Library and Information Science of National Taiwan University, Taipei]。

陳光華、江玉婷[Chen, Kuang-Hua, & Chiang, Yu-Ting] (2000)。中文資訊檢索測試集之設計與製作[Zhongwen zixun jiansuo ceshiji zhi sheji yu zhizuo]。資訊傳播與圖書館學[*Zixun Chuanbo yu Tushuguanxue*]，*6*(3)，61-80。

陳明君[Chen, Ming-Chun] (1999)。檢索背景與查詢問題對檢索技巧及檢索結果之影響研究[*Jiansuo beijing yu chaxun wenti dui jiansuo jiqiao ji jiansuo jieguo zhi yingxiang yanjiu*]。未出版之碩士論文，國立臺灣大學圖書資訊學研究所，台北市[Unpublished master dissertation, Graduate Institute of Library and Information Science of National Taiwan University, Taipei]。

黃怡如[Huang, Yi-Ju] (1999)。終端使用者與系統互動前後查詢問題、檢索概念與檢索詞彙變化之研究[*Zhongduan shiyongzhe yu xitong hudong qianhou chaxun wenti, jiansuo gainian yu jiansuo cihui bianhua zhi yanjiu*]。未出版之碩士論文，國立臺灣大學圖書資訊學研究所，台北市[Unpublished master dissertation, Graduate Institute of Library and Information Science of National Taiwan University, Taipei]。

Buckley, C. (2000). *The TREC-9 query track*. Retrieved December 24, 2008, from http://trec.nist.gov/pubs/trec9/papers/query_track.pdf

Derr, R. L. (1982). A classification of questions in information retrieval by conceptual presupposition. *Proceedings of the 45th Annual Meeting of the American Society for Information Science, 19*, 69-71.

Fidel, R., & Soergel, D. (1983). Factors affecting online bibliographic retrieval: A conceptual framework for research. *Journal of the American Society for Information Science, 34*(3):163-180.

Graesser, A. C., & Murachver, T. (1985). Symbolic procedures of question answering. In A.C. Graesser, & J. B. Black (Eds.), *The psychology of questions* (pp. 15-88). Lawrence Hillsdale, NJ: Erlbaum Associates.

Keyes, J. G. (1996). Using conceptual categories of questions to measure differences in retrieval performance. *Proceedings of the 59th Annual Meeting of the American Society for Information Science, 33*, 238-242.

Kishida, K., Chen, K. H., Lee, S., Kuriyama, K., Kando, N., & Chen, H. H. (2007, May). Overview of CLIR task at the sixth NTCIR workshop. In *Procedings of NTCIR-6 Workshop Meeting*, Symposium held at the Tokyo, Japan.

Lancaster, F. W. (1968). *Evaluation of the MEDLARS demand search service*. Bethesda, Md.: National Library of Medicine.

Lehnert, W. G. (1978). *The process of question answering: A computer simulation of cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nelson, M. J. (1995). The effect of query characteristics on retrieval results in the TREC retrieval tests. In *Proceedings of the 23rd annual conference of the canadian association for information science (CAIS 95)* (pp. 156-163). Canada: University of Alberta.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving II: Users, questions, and effectiveness. *Journal of the American Society for Information Science, 39*(3), 177-196.

Saracevic, T., & Baxter, M. A. (1983). On a method for studying the structure and nature of requests in information retrieval. *Proceedings of the 46th Annual Meeting of the American Society for Information Science, 20*, 22-25.