圖書資訊學研究5：1 (December 2010)：51-74

學科知識與檢索訓練對檢索效益的影響

The Impact of Domain Knowledge and Search Training on Search Effectiveness

劉英享

Ying-Hsang Liu

澳洲查理斯特大學資訊研究學院 資訊管理講師
Lecturer in Information Management
School of Information Studies
Charles Sturt University

【摘要 Abstract】

本文從資訊檢索互動觀點，以實驗法研究使用者特徵（使用者學科知識及檢索訓練）對檢索效益的影響。本研究將受試者分為搜尋新手、學科專家、搜尋專家及醫學圖書館員等四種類型，以取自對檢索系統難度頗高的文獻檢索會議2004年基因體學檢驗項目文件集之檢索問題，實際搜尋以MEDLINE資料庫近十年書目紀錄所建構的資訊檢索實驗系統，並以文獻檢索會議所提供的相關判斷作為評估檢索效益基準。實驗系統有兩種不同的查詢介面，差別在能否運用醫學標題表搜尋。研究發現學科專家使用醫學標題表所得之檢索結果，比搜尋專家更好，顯示醫學標題表在檢索技術性問題時，對學科專家最有幫助。

To what extent do MeSH terms improve search effectiveness of different kinds of users? We observed four different kinds of information seekers using an experimental information retrieval system: 1. search novices; 2. domain experts; 3. search experts and 4. medical librarians. The information needs were a subset of the relatively difficult topics originally created for the Text REtrieval Conference (TREC). Effectiveness of retrieval was based on the relevance judgments

投稿日期：2010.10.4；接受日期：2010.11.2
email: 劉英享yingliu@csu.edu.au

provided by TREC. Participants searched either using a version of the system in which MeSH terms were displayed or another version in which they had to formulate their own terms. The results suggest that MeSH terms are more helpful in terms of precision for domain experts than for search experts. We speculate that this is because of the highly technical nature of the topics; only the domain experts had the knowledge to understand and therefore make use of the MeSH terms. The results advance our understanding of the usefulness of controlled vocabulary in interactive information retrieval systems.

[關鍵字 Keywords]

資訊檢索互動；使用者特徵；醫學標題表；檢索評鑑

Interactive information retrieval; User characteristics; MeSH terms; Information retrieval evaluation

壹、前言

資訊檢索互動研究近幾年在國際圖書資訊學界逐漸成型，資訊檢索互動情境研討會（Information Interaction in Context Symposium, IIiX）（Belkin & Kelly, 2010; Borlund, et al., 2008; Ruthven, et al., 2006） 可為代表。這一系列研討會探討情境在資訊檢索互動的重要性，讓資訊檢索系統、互動式資訊檢索及資訊行為等領域研究者有交流機會。檢索評鑑研究向來是備受關注的重要議題之一，實驗室和實際測試評鑑研究取向如何互補長短，使用者工作任務等情境因素在檢索搜尋與檢索行為的角色，均為近年研究發展趨勢（國內有關資訊檢索互動研究方法引介、實證研究與實務應用，參見吳美美，1993，1998，2001；林珊如，1995；唐牧群、鄧雅文、鄭瑋、謝宜瑾，2008；黃慕萱，1997）。

從歷史角度來看， 控制詞彙（controlled vocabulary）和自由詞彙（free-text）檢索效益的比較，和資訊檢索研究歷史一樣悠久，兩者孰優孰劣至今仍無定論（Sparck Jones, 1981; Svenonius, 1986, 2000; Vickery, 1971）。鑑於現代搜尋引擎普遍運用自由詞彙，以檢索為目的所建構的控制詞彙，是否符合經濟效益，為當今重要的研究議題之一。

從實務的角度來看， Calhoun（2006）所撰寫的美國國會圖書館委託研究報告，建議採用自動索引，因為人工索引不符合成本效益；美國加州大學圖書館書目服務工作小組則認為控制詞彙難以瞭解及使用

（Bibliographic Services Task Force, 2005）。Byrd 等人（2006）與Mann（2006）主張必須維護人工索引，理由是學術圖書館需要有系統且詳盡的文獻檢索機制。最近，美國國會圖書館書目控制工作小組有鑑於變動的資訊環境、資料分享經濟模式及使用者的高度期望，建議將國會標題表再製利用，認清電腦運算方法在主題分析實務的運用潛力（On the record: Report of the library of congress working group on the future of bibliographic control, 2008）。

一般來說，資訊檢索系統研究人員認為，以各種檢索技術自動構建全文索引，並不像人工索引需要龐大資源，就能夠檢索到相關文件（Croft, 1989; Salton, 1986）。主張人工索引研究者則認定人類自然語言缺乏系統，無法充分表述文件的內容（Svenonius, 1986, 2000）。控制詞彙索引法的目的，在於支持使用者的搜尋任務，儘量減少自然語言的固有歧義性（如 Nelson, Johnston, & Humphreys, 2001）。

先前有關控制詞彙索引法的評鑑研究，以醫學標題表和自動索引技術為例，都在實驗室環境下以批次模式（batch mode）進行研究（如 Salton, 1972; Savoy, 2005），使用者並沒有實際使用系統搜尋文件。總結來說，以多種檢索技巧（retrieval techniques）與檢索模型（retrieval models）組合，與控制詞彙比較，兩者隨機搜尋（ad-hoc search）的檢索表現，可相互匹敵。

自動與人工索引目的都在幫助使用者找到所需資訊，使用者特徵（user characteristics）如何影響檢索結果，在檢索評鑑研究中也是重要的切入點。和一般觀察所得有些出入，使用者特徵相關研究指出：1. 使用者學科知識或特定主題知識，和檢索結果無關（Allen, 1991; Pao, et al., 1993）；2. 資料庫檢索經驗也和檢索結果無關（Fenichel, 1981; Howard, 1982; Sutcliffe, Ennis, & Watkinson, 2000）。

有鑑於詞彙控制相關議題在資訊檢索實務與檢索評鑑的重要性，本文從資訊檢索互動（interactive information retrieval）觀點，以實驗法研究使用者特徵（使用者學科知識及檢索訓練）對檢索效益（search effectiveness）的影響，試圖回答以下研究問題：

一、醫學標題表對使用者是否有幫助？

二、醫學標題表對不同類型的使用者是否有不同程度的幫助？

本研究觀察四種不同類型的使用者：1. 搜尋新手（Search Novice, SN）；2. 學科專家（Domain Expert, DE）；3. 搜尋專家（Search Expert, SE）；4. 醫學圖書館員（Medical Librarian, ML），搜尋以MEDLINE資料庫近十年的書目紀錄所建構的資訊檢索實驗系統。檢索問題（search

topic）取自對檢索系統難度頗高的文獻檢索會議（Text REtrieval Conference, TREC）2004年基因體學檢驗項目文件集（Hersh, et al., 2006; TREC, 2005）。

根據以上研究問題及相關實證研究發現，本研究提出以下研究假設：

假設一：查詢使用醫學標題表的檢索結果會比不使用來得更好。

假設二：檢索的結果視不同類型使用者而有所不同：

甲、學科專家和醫學圖書館員使用醫學標題表所得檢索結果，會比搜尋新手和搜尋專家更好。

乙、醫學圖書館員使用醫學標題表所得檢索結果，會比學科專家更好。

貳、文獻探討

文獻表述（text representation）的探索，幾乎和資訊檢索研究一樣有長久歷史，因為文件的意義為作者與使用者之間如何相匹配的關鍵。控制詞彙研究者認定人類自然語言缺乏系統，無法充分表述文件的內容（如Svenonius, 1986, 2000）。控制詞彙主要透過統一詞彙間的變異，制定詞彙間複雜的語義關係，以解決詞彙變異的問題。例如，醫學標題表為當代最先進的控制索引語言，由美國國家醫學圖書館專業索引人員維護，並運用在MEDLINE資料庫中的文件索引（Aronson, et al., 2004; Humphrey, Rogers, Kilicoglu, Demner-Fushman, & Rindflesch, 2006; Nelson, et al., 2001）。

然而，有些研究指出，複雜的詞彙關係對檢索效益的提升並無實質的幫助。Sparck Jones（1981）全面檢視1958年至1978年索引法發現 ，不同索引語言的檢索表現可達到相同水平，簡單與複雜的索引法，檢索表現相當。由於檢索評鑑實驗室的研究，重點是對不同的自動檢索技巧比較，而非不同的控制索引法（Sparck Jones, 2005），至今吾人對控制詞彙或控制索引法的瞭解仍然有限。此觀察與Tsay（2004）針對1977至2000年主題索引的書目計量研究結果，有關主題索引相關研究報告1992年之後有驟減的趨勢相符。

從使用者搜索行為角度來看，Jenuwine與Floyd（2004）以研究者檢索經驗為基礎，運用不同檢索策略，比較醫學標題表和自由詞的檢索表現。研究結果指出兩者有互補之處，在進行檢索時應交互運用。Wacholder與Liu（2006, 2008）對控制索引法與自然語言處理所產生的查

詢詞彙（query term）進行比較研究，在互動式的檢索環境，讓使用者瀏覽詞彙並回答預設問題 。此系列研究建議可用使用者的查詢詞彙點擊資料（click-on data）和檢索效益，來評估不同查詢詞彙的有用性。

一般來說，資訊檢索系統研究者認為，以各種檢索技術自動構建全文索引，並不像人工索引需要龐大資源，就能夠檢索到相關文件（Croft, 1989; Salton, 1986）。實證研究指出，在實驗室控制環境下，比較人工和自動索引系統，如果自動索引系統建置得當，其檢索表現可與人工索引法相比（如Salton, 1969, 1972; Srinivasan, 1996）。由於這方面研究都在實驗室環境中進行，使用者並不參與實際搜尋，從資訊檢索互動角度而言，控制索引法在互動式的資訊檢索環境，並沒有足夠且適當的評鑑研究。

從資訊搜尋與資訊檢索互動認知的角度來看，研究關注使用者從事檢索活動時的知識狀態和知識結構與文獻表述法間的聯繫（Belkin & Kwasnik, 1986; Ingwersen, 1996; Ingwersen & Jarvelin, 2005）。為因應傳統資訊檢索模式認為使用者具清晰的資訊需求，並能夠清楚以查詢詞彙表達需求的假定，Belkin與Kwasnik提出以詞彙聯想圖（word association map）表述使用者的知識狀態，檢索系統可據以選擇不同的檢索策略。

資訊檢索系統與使用者互動的相關研究，研究者試圖建立使用者搜尋模型，考慮使用者情境、檢索問題與資訊蒐尋歷程（information seeking process）之間的互動等（Lin & Belkin, 2005）。相關研究指出使用者特徵對檢索結果的影響，使用者個別差異（individual differences）很大（Bellardo, 1985; Fenichel, 1981; Saracevic & Kantor, 1988）。使用者的學科知識、檢索訓練或資料庫使用經驗，一直是研究者關注的變項，實證研究結果指出：1. 使用者學科知識或特定主題知識，和檢索結果的精確率（precision）和回現率（recall）無關（Allen, 1991; Pao, et al., 1993）；2. 資料庫檢索經驗也和檢索結果的精確率和回現率無關（Fenichel, 1981; Howard, 1982; Sutcliffe, et al., 2000）。這些研究主要限制在於使用少數的檢索問題，檢索問題差異性（search topic variability）對檢索效益的影響過大。

綜合而言，檢索評鑑研究一直受到資訊科學研究社群的重視，從最早資訊組織中控制詞彙效益研究、資訊檢索系統檢索技巧實驗室研究，到資訊檢索互動研究，都不斷推陳出新。然而，近年研究顯示控制詞彙的效益評鑑相關議題，並未得到應有重視 。雖然控制詞彙的目的主要幫助領域專家，實證研究卻發現，使用者特徵（如學科知識和檢索訓練），和檢索結果的精確率和回現率無關。本研究以資訊檢索系統實驗

室研究為基礎，融合資訊檢索互動觀點，以醫學標題表為例，探討控制詞彙和不同類型使用者之間的關係。

參、研究方法

本研究觀察四種不同類型的使用者，1. 搜尋新手（SN）；2. 學科專家（DE）；3. 搜尋專家（SE）；4. 醫學圖書館員（ML），搜尋以MEDLINE資料庫近十年的書目紀錄所建構的資訊檢索實驗系統。檢索問題取自對檢索系統難度頗高的2004年文獻檢索會議基因體學檢驗項目文件集（Hersh, et al., 2006; TREC, 2005），檢索效益以TREC提供的相關判斷為基準。

一、受試者

參與本實驗研究的受試者，以其學科知識與檢索訓練程度的高低，區分為以下四種使用者類型：

表1　以學科知識與檢索訓練區分的四種使用者類型

使用者類型（Searcher Type）	學科知識	檢索訓練
搜尋新手 Search Novice (SN)	−	−
學科專家 Domain Experts (DE)	+	−
搜尋專家 Search Experts (SE)	−	+
醫學圖書館員 Medical Librarians (ML)	+	+

註：加號（+）與減號（−）分別表示使用者學科知識與檢索訓練程度之「高」「低」。

這四種使用者類型的操作型定義如下：

1. 搜尋新手（SN）： 在生物醫學領域沒有進階的相關知識，非主修生物的大學生，未經正規檢索訓練課程培訓。雖然很多大學生為網路重度使用者，但他們並沒有對線上資料庫檢索有深入的瞭解。

2. 學科專家（DE）：生物醫學領域研究生，未經正規檢索訓練課程培訓。
3. 搜尋專家（SE）：在生物醫學領域沒有進階知識，或未曾在大學主修生物或擁有生物醫學領域碩士以上學位的圖書資訊學碩士研究生。修過線上檢索或其他相關正規資訊檢索課程，。
4. 醫學圖書館員（ML）：受過生物醫學領域正規教育或醫學圖書館兩年以上工作經驗的醫學圖書館員。具線上檢索專長。

二、搜尋系統

本研究所使用的資訊檢索實驗系統，提供MeSH+與MeSH−兩種不同的查詢介面（search interface）。 MeSH+版本可以讓檢索者使用醫學標題表，檢索結果也會一併顯示標題； MeSH−版本則不提供此項功能，其餘設定則完全相同。為防止醫學標題表可能引起受試者的特別注意，本研究把 MeSH+與MeSH−查詢介面分別稱為「System Version A」 與「System Version B」（如圖1所示）。

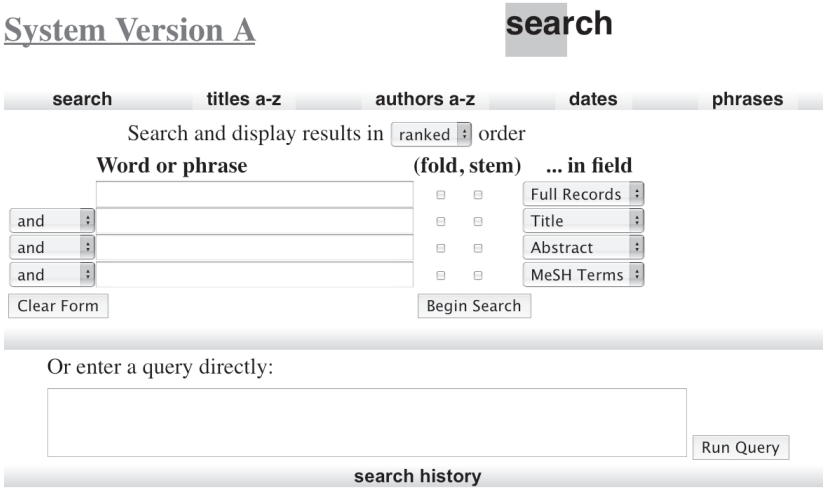

圖1　以綠石數位圖書館軟體建置的資訊檢索實驗系統查詢介面

TREC Genomics Track 2004 測試集語料取自MEDLINE資料庫，從1994到2003年間擷取4,591,108筆書目紀錄。本實驗去除不包含醫學標題表或摘要的書目紀錄，使用75.0%的語料，共3,442,321 筆。資訊檢索實驗系統以綠石數位圖書館軟體建置（Greenstone, 2006），系統預設成以布林邏輯為基礎的排序系統（Boolean-based ranked system），並使用TF

× IDF加權規則排序檢索結果。搜尋引擎使用 MGPP（MG++） 搜尋和壓縮演算法（為MG [Managing Gigabytes] 重新實作）（Witten, Moffat, & Bell, 1999）。MeSH+版本中的主題標目並沒有做特別處理，一併加入系統索引中。

本實驗研究刻意不執行某些系統功能，如能讓使用者利用醫學標題表階層結構的相關功能，例如超連結的主題標目、Explode擴展狹義詞功能和自動查詢擴展（query expansion；見Hersh, 2008; Lu, Kim, & Wilbur, 2009）。雖然系統具備這些功能可能增進醫學標題表的有用性，但也會導入查詢介面與查詢詞處理（query processing）層面的因素，導致實驗結果無法偵測控制詞彙的檢索效益。

三、實驗設計

本研究為 4 × 2 × 2 因子實驗設計（factorial design），考慮四種不同類型的使用者、兩種查詢介面，並控制檢索問題配對（search topic pair）。本研究共招募32位受試者，每一類型使用者各8位，每人搜尋8題檢索問題（MeSH+與MeSH−版本各4題）。為考量檢索問題差異性，本研究透過 4 × 4 希臘—拉丁方陣平衡實驗設計（Graeco-Latin square balanced design; Fisher, 1935; Kirk, 1995），為使每一題檢索問題由同一類型使用者檢索兩次以上，共需20題檢索問題（隨機組成10對）。根據此研究設計，每一檢索問題共被檢索16或20次。不同類型使用者、查詢介面版本與檢索問題配對分配表，如圖2所示。

四、檢索問題

本研究的檢索問題取自TREC Genomics Track 2004（TREC, 2005）測試集，本來的目的在對不同的檢索技巧進行檢索效益評估（圖3為檢索問題樣本）。檢索問題的主題涵蓋基因體學重要主題，為生物學家從事研究，搜尋資料的典型問題。每一題除了編號（ID）之外，還包括標題（Title）、資訊需求（Need）與情境（Context）等欄位。

標題為簡短查詢問題，資訊需求欄位提供生物學家有興趣的資訊，而情境欄位則提供文件相關判斷的描述。資訊需求與情境欄位原本目的在於提供更多檢索詞彙，作為實驗室環境系統測試之用。

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SN	DE	SE	ML	DE	SN	ML	SE	SE	ML	SN	DE	ML	SE	DE	SN
38	12	29	50	38	12	27	45	29	50	27	45	42	46	9	36
12	38	50	29	12	45	38	27	50	29	29	27	46	36	42	9
29	50	12	38	27	38	45	12	27	45	45	50	9	42	36	46
50	29	38	12	45	27	12	38	45	27	50	29	36	9	46	42
42	46	32	15	9	36	30	20	2	43	1	49	2	43	33	23
46	42	15	42	36	9	20	30	43	1	49	2	43	2	23	33
32	15	42	46	30	20	9	36	1	49	2	43	33	23	2	43
15	32	46	32	20	30	36	9	49	2	43	1	23	33	43	2

圖2　4×4 希臘—拉丁方陣實驗設計

註：第一行數字1-16為受試者編號；SN、DE、DE及ML為不同類型的使用者：SN=搜尋新手，DE=學科專家，SE=搜尋專家，ML=醫學圖書館員；陰影和非陰影區塊分別表示MeSH+與MeSH−查詢介面；區塊內數字代表檢索問題編號（見TREC, 2005）；10 對檢索問題編號如下：（38, 12）、（29, 50）、（42, 46）、（32, 15）、（27, 45）、（9, 36）、（30, 20）、（2, 43）、（1, 49）、（33, 23）。

ID: 39
Title: Hypertension
Need: Identify genes as potential genetic risk factors candidates for causing hypertension.
Context: A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.

圖3　檢索問題樣本

由於基因體學大部分為技術性問題，本研究考量使用者是否能夠理解這些檢索問題，特別是對生物醫學領域沒有進階知識的使用者。有鑑於檢索問題本來用在實驗室評鑑研究，並沒有考慮使用者，本研究依照以下步驟，從50題中挑選出20題：

1. 諮詢有生物醫學背景的檢索專家與神經科學研究生，幫助判斷出一般使用者可以理解的題目。檢索問題中使用術語，如特定的基因、途徑和機制等被排除在外；

2. 使用醫學標題表瀏覽器 （MeSH Browser [2003 MeSH], 2004），確保檢索主要概念可對應到標題表。例如：第39題可以對應到主題標目高血壓（hypertension）及危險因素（risk factors）；
3. 刪除相關判斷平均精確率平均值（Mean Average Precision, MAP）和P10（Precision at Top 10）過低的題目，因為這些題目對檢索系統而言難度太高；
4. 從測試集共50題檢索問題，共選出20題， 隨機選擇組成10對，作為實驗條件（參見圖2）。

五、研究實施

本研究主要在美國東北部一所公立研究型大學進行，有部分醫學圖書館員受試者在其工作場所進行。以張貼佈告欄、寄發一般信件或電子郵件與線上討論群等方式招募受試者，受試者在進行檢索活動之前，需填寫一份背景問卷。經短暫的系統使用、檢索問題概念分析（concept analysis）及醫學標題表使用訓練之後，分配到其中一個實驗條件（見圖2），並搜尋分配的檢索問題。在完成每一檢索問題後，填答一份搜尋後問卷，並對預先選取的兩筆書目紀錄進行相關判斷。一共完成8題之後，進行簡短的後訪談。

在搜尋檢索問題前，受試者需針對檢索問題進行概念分析。分配到MeSH+版本系統時，可以使用醫學標題表網路瀏覽器查詢主題詞（MeSH Browser [2003 MeSH], 2004），醫學標題表網路瀏覽器是由美國國家醫學圖書館建置，主要用來幫助使用者查詢相關詞彙，可顯示詞彙階層結構，選取適當的主題標目。使用MeSH−版本時，受試者必須自己發想相關的檢索詞彙，在檢索結果中也不會顯示標題（見圖4）。

<table>
<tr><td rowspan="3">背景問卷</td><td rowspan="3">系統使用相關訓練</td><td>概念分析</td><td>實驗系統</td><td rowspan="3">搜尋後問卷</td><td rowspan="3">搜尋後訪談</td></tr>
<tr><td>MeSH Browser</td><td>MeSH+ 版本</td></tr>
<tr><td>發想檢索詞彙</td><td>MeSH− 版本</td></tr>
</table>

圖4　研究實施概覽

每一檢索問題以10分鐘為限，並以最後一次的搜尋結果評估檢索表現，若於規定時間內無法完成，則請受試者選定自認最滿意的一次檢索結果。在每次檢索過程中，系統會自動記錄檢索詞彙、排序的檢索結果及花費時間等。

為使受試者能夠積極表現，本研究設有獎勵措施。除了參與實驗可得20美元之外，依照檢索結果前十篇文件的精確率（以8題檢索問題平均計算），最多可多得18美元。實驗結果指出，每位受試者平均多得4.4美元（最大與最小值分別為8與2）。

六、資料分析

因本研究主要聚焦實驗室評鑑研究和使用者檢索表現間的轉移性，以文獻檢索會議2004基因體學文件集（TREC, 2005）的相關判斷作為檢索效益評估的基準，精確率和回現率作為評鑑量標，並以trec_eval程式計算檢索結果（Buckley, 1999）。為確保TREC彙總的相關判斷集的全面性及有效性，特別進行可靠性測試，結果指出TREC相關判斷集適用於本研究（詳見Liu, 2009分析，並參見Hersh, et al., 2004與Zobel, 1998有關以彙總法 [pooling method] 所得相關判斷的潛在偏差）。

本研究基於因子實驗設計，考量檢索系統、使用者和檢索問題等因素， 建構固定效應線性模型（linear fixed-effects model） 如下：

$$y(i,j,k) = m + s(i) + r(j) + t(k) + e(i,j,k)$$

其中，

$y(i,j,k)$ = 精確率與回現率；系統i，使用者j，檢索問題配對k。

m = 檢索結果的平均精確率與回現率。

$s(i)$ = 系統效應i；其中i = 1（MeSH−版本），2（MeSH+版本）。

$r(j)$ = 使用者效應j，其中j = 1（搜尋生手，SN），2（學科專家，DE），3（搜尋專家，SE），4（醫學圖書館員，ML）。

$t(k)$ = 檢索問題配對效應, 其中k = 1至10檢索問題配對。

$e(i,j,k)$ = 隨機誤差觀察$y(i,j,k)$。

為使精確率與回現率能夠趨近常態分布，本研究進行資料平方根數據轉換（Fox, 1997; Hull, 1993; Tague-Sutcliffe, 1992, p. 485），以滿足變異數分析（analysis of variance, ANOVA）的基本要求。除檢視QQ Plot（quantile-quantile plot）之外，也使用Shapiro-Wilk 檢定資料是否呈常態分布。檢定結果無法駁回轉換資料呈常態分布的虛無假設。

綜合而言，因子實驗設計能夠在受試者不多的情況下，分離檢索系統、使用者與檢索問題配對等因素對檢索結果的影響。資料轉換是為了滿足變異數分析的基本要求，統計分析與圖表繪製以統計軟體R進行（R Development Core Team, 2009）。

七、受試者學科知識與檢索訓練

本研究四種類型使用者各招募8位，學科知識程度以大學及研究所主修生物醫學相關課程數目為計算標準。一般來說，如本研究所預期，學科專家DE的學科知識程度最高（曾修習大學部相關課程中位數為15，研究所課程中位數為7.5），醫學圖書館員ML次之，但館員的生物醫學相關知識比學科專家低很多，大部分還是以大學部訓練為主（如圖5、圖6所示）。

就檢索訓練而言，以曾經修習線上檢索相關課程為標準，醫學圖書館員的程度最高，搜尋專家次之，絕大部分的搜尋生手與學科專家沒有受過相關訓練（見圖7）。綜合而言，本研究所招募的各類型使用者的學科背景與檢索訓練，符合原先的實驗設計。

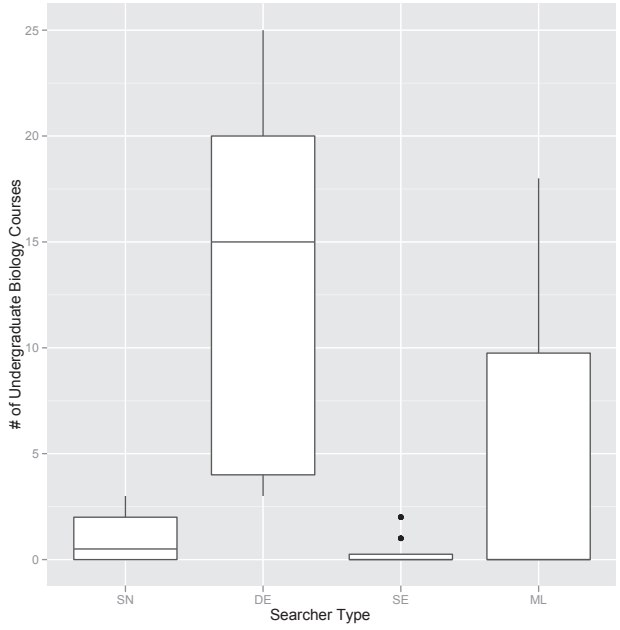

圖5　曾修習大學程度生物課程數盒狀圖

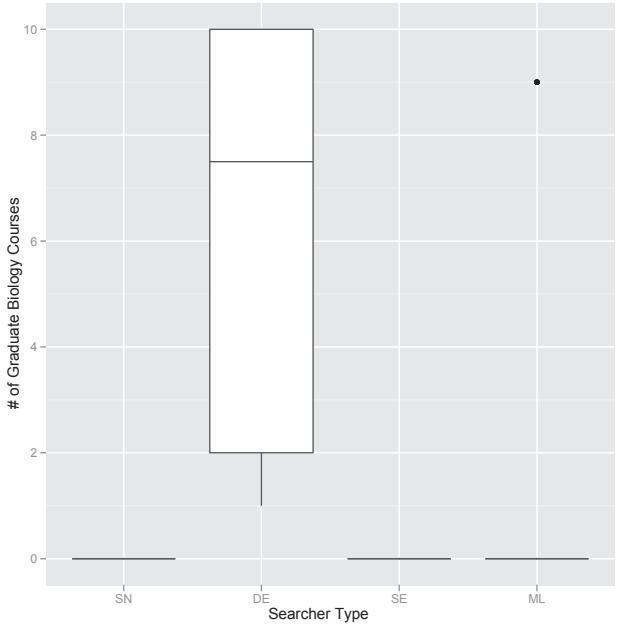

圖6　曾修習研究所程度生物課程數盒狀圖

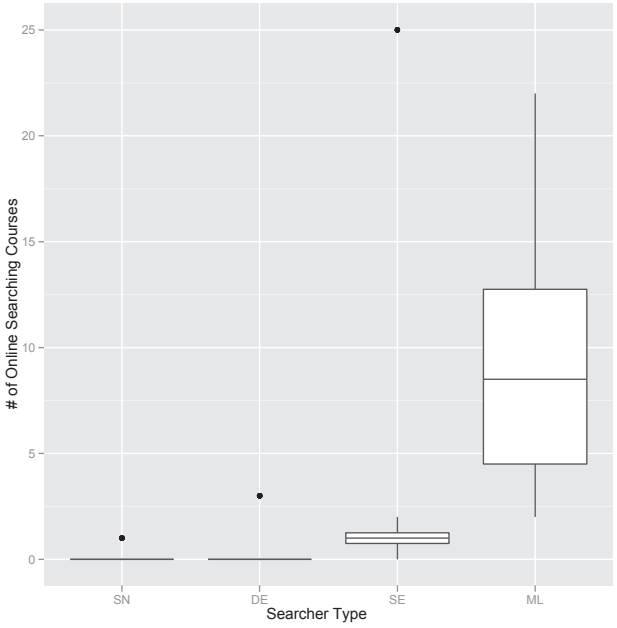

圖7　曾修習線上檢索相關課程數盒狀圖

肆、研究結果

一、醫學標題表使用概觀

根據本研究系統查詢紀錄的檢索欄位使用分析，結果指出受試者的確在檢索過程中運用醫學標題表（如表2所示）。在所有檢索問題當中，使用標題表的檢索題數，佔所有題數的39.1%。進一步分析指出，不同類型的使用者和檢索欄位的運用，在統計上有顯著關係（Fisher's Exact Test, $p < .001$）。使用者的檢索訓練表現在標題使用上，受過愈多檢索訓練的使用者，愈可能使用標題。本實驗的學科專家雖然不曾使用醫學標題表，但經簡短訓練之後，在進行檢索時約有三分之一的問題使用標題，此項結果也確認實驗材料與實施辦法的有效性。

如研究實施圖4所示，受試者在使用MeSH+系統進行查詢前，必須針對檢索問題進行概念分析，可使用MeSH Browser查找和檢索問題相關的醫學標題表，作為查詢時運用檢索詞彙參考之用。表2顯示僅有近四成在檢索過程中使用MeSH，反應使用者的實際檢索行為。至於為何有六成沒有使用MeSH欄位進行檢索，本實驗研究無法提供明確答案。如果實驗系統設計和Wacholder與Liu（2006, 2008）類似，迫使受試者瀏覽並使用醫學標題表，可增進實驗本身的內在效度（internal validity），但相對地未能反應使用者實際檢索行為。

表2　MeSH+版本醫學標題表欄位使用情形

使用者類型	使用題數（%）	未使用題數（%）	題數總計（%）
搜尋生手SN	2 (6.3%)	30 (93.8%)	32 (100.0%)
學科專家DE	10 (31.3%)	22 (68.8%)	32 (100.0%)
搜尋專家SE	16 (50.0%)	16 (50.0%)	32 (100.0%)
醫學圖書館員ML	22 (68.8%)	10 (31.3%)	32 (100.0%)
總計	50 (39.1%)	78 (60.9%)	128 (100.0%)

註：每一類型使用者各有8位使用者，每位以MeSH+版本搜尋4題。

二、檢索效益

比較MeSH+與MeSH−兩種資訊檢索實驗系統版本，使用者的檢索表現指出，兩者在統計上檢索結果的精確率（one-way ANOVA, F (1, 254) = 0.01, $p = 0.94$, $p > .05$）與回現率（one-way ANOVA, F (1, 254) = 0.30, p =

0.58, $p > .05$）並沒有顯著差異。因此，研究結果無法支持研究假設一：查詢使用醫學標題表會比不使用所得的檢索結果更好。

整體來說，四種不同類型的使用者檢索結果表現類似，在統計上檢索結果的精確率（one-way ANOVA, F (3, 252) = 1.86, p = 0.14, $p > .05$）與回現率（one-way ANOVA, F (3, 252) = 1.66, p = 0.18, $p > .05$）在統計上並無顯著差異（見表3）。

表3　四種不同類型使用者檢索表現

使用者類型	平均精確率 Mean Precision	*N*	平均回現率 Mean Recall	*N*
搜尋生手SN	0.29	64	0.21	64
學科專家DE	0.40	64	0.15	64
搜尋專家SE	0.30	64	0.15	64
醫學圖書館員ML	0.35	64	0.23	64
總計	0.34	256	0.18	256

但比較不同類型使用者在不同系統版本的檢索表現，發現兩者之間以精確率而言，在統計上有相當顯著的差異（one-way ANOVA, F (7, 248) = 3.48, p = 0.001, $p < .01$）。學科專家DE和搜尋專家SE在使用MeSH+版本時，以檢索結果精確率而言，在統計上有相當顯著的差異（Tukey's HSD, mean difference = .31, $p < .01$）（見表4，陰影部分標示在統計上有顯著差異的檢索結果）。進一步分析，DE和SE之間的差異對檢索效益的影響，以精確率而言，其效益值（Effect Size）相關係數r = 0.46（Borenstein, 2009）。 研究結果顯示，學科專家與搜尋專家間的主要差異，在於生物醫學相關學科知識程度的高低（參見圖5、圖6）。兩者在進行檢索任務時，學科知識和MeSH+系統版本產生互動效應，導致精確率在統計上有相當顯著的差異，進一步反應在效益值上。

表4　不同檢索介面版本與使用者類型的檢索表現

項目	MeSH+			MeSH−		
使用者類型	平均精確率 Mean Precision	平均回現率 Mean Recall	*N*	平均精確率 Mean Precision	平均回現率 Mean Recall	*N*
搜尋生手SN	0.36	0.21	32	0.23	0.20	32
學科專家DE	0.51	0.15	32	0.29	0.15	32
搜尋專家SE	0.21	0.16	32	0.38	0.13	32
醫學圖書館員ML	0.28	0.22	32	0.42	0.24	32
總計	0.34	0.19	128	0.33	0.18	128

整體而言，本研究假設「檢索的結果視不同類型使用者而有所不同」未能獲得支持。本研究先前假設二（甲）：「學科專家和醫學圖書館員使用醫學標題表所得檢索結果，會比搜尋新手和搜尋專家更好」，未能獲得研究結果支持。若考慮不同類型使用者與不同系統版本間的互動關係，研究結果則支持以下修正後的假設：「學科專家使用醫學標題表所得檢索結果，會比搜尋專家更好」。

更進一步來說， 本研究先前假設二（乙）：「醫學圖書館員使用醫學標題表所得檢索結果，會比學科專家更好」也未能獲得研究結果的支持（見圖8）。在使用 MeSH+版本時，就精確率而言，學科專家DE所得的結果較醫學圖書館員ML高，但兩者在統計上沒有顯著的差異。醫學圖書館員用MeSH作為輔助工具，精確率反而降低，可能和檢索問題理解程度相關。本研究的檢索問題偏向基因體學專門技術性問題，和醫學圖書館員平常遇到的參考問題有些差距，可能造成使用MeSH，精確率反而降低（參見Liu, 2009針對受試者對檢索問題理解程度的進一步分析）。

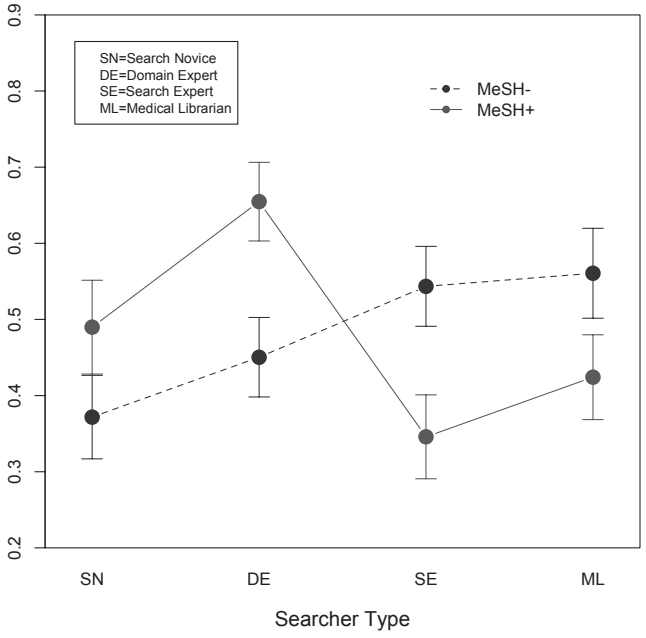

圖8　使用者類型與系統介面版本檢索表現折線圖

三、搜尋時間

受試者在本研究中對搜尋任務相當專注，大部分用完10分鐘預設的時間。圖9使用者搜尋花費時間密度直方圖顯示，以550至660秒（約9至10分鐘）的時間區段密度最高。進一步分析指出，使用MeSH+或MeSH−介面搜尋所花費的時間，在統計上並無顯著的差異（one-way ANOVA, F (1, 254) = 2.77, p = 0.10, p > .05）。然而，不同使用者類型所花費的時間，在統計上有顯著的差異（one-way ANOVA, F (3, 252) = 3.47, p = 0.02, p < .05）；學科專家DE所花費的時間顯著比搜尋專家SE多（Tukey HSD, mean difference = 71.86, p < .05）。

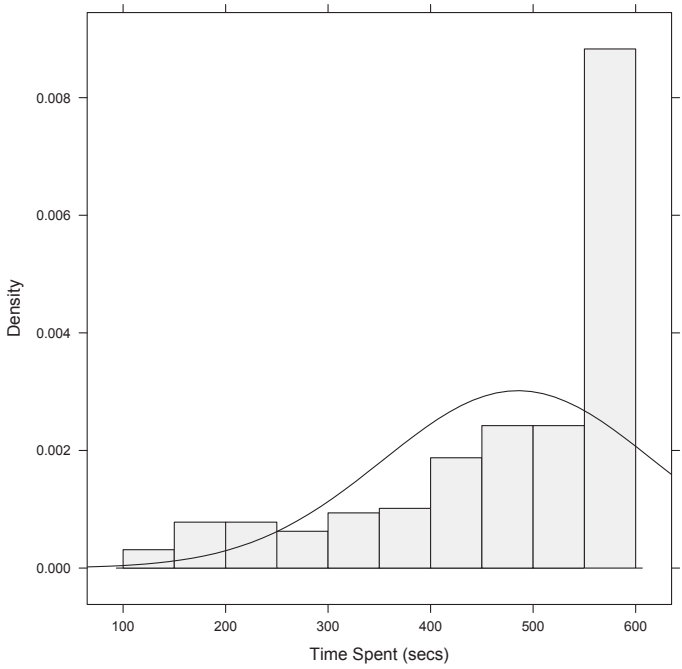

圖9　搜尋花費時間密度直方圖（N=256）

四、研究結果小結

以檢索效益與花費時間而言，在整體結果、系統檢索介面與使用者類型等方面的結果，如表5所示。整體結果指出，檢索結果的精確率與回現率相當低，分別只有 .34與 .18，每一檢索問題約8分鐘。兩種檢索介面對檢索效益及花費時間，並沒有顯著差異。研究結果無法支持研

究假設：「查詢使用醫學標題表會比不使用所得的檢索結果更好」。相同地，不同類型使用者的檢索表現相當，但學科專家明顯比搜尋專家花更多的時間。更重要的是，學科專家使用醫學標題表所得檢索結果，比搜尋專家更好，顯示醫學標題表在檢索技術性問題時，對學科專家最有幫助。

表5　檢索效益與花費時間提要

項目	檢索效益	花費時間
整體結果	平均精確率 = .34 平均回現率 = .18	平均 = 485.3 秒／題
系統檢索介面	MeSH+ = MeSH–	MeSH+ = MeSH–
使用者類型	SN = DE = SE = ML	DE >> SE
使用者與系統介面	DE-MeSH+ >> SE-MeSH+ DE-MeSH+ >> SN-MeSH–	DE-MeSH+ >> SE-MeSH–

註：>> 符號表示在統計上小於.05顯著水準；= 符號表示在統計上沒有顯著差異；SN = Search Novices 搜尋生手；DE = Domain Experts 學科專家；SE = Search Experts 搜尋專家；ML = Medical Librarians 醫學圖書館員。

伍、結論

本文從資訊檢索互動觀點，以實驗法研究使用者特徵（使用者學科知識及檢索訓練）對檢索效益的影響，研究發現搜尋技術性的檢索問題，需要有進階的學科知識來妥善運用醫學標題表 。具體來說，檢索標目對學科專家提升精確率最有幫助，此項研究結果和先前相關研究抵觸（如Allen, 1991; Pao, et al., 1993）。可能的解釋為先前研究的受試者學科背景知識較為類似， 另一種解釋可能是TREC檢索主題的內在複雜性，只有相關領域的專家具備足夠知識解讀問題。Blair（2002）指出窮舉搜尋（exhaustive searches）極為困難，本實驗的檢索問題屬於此類，須在有限時間內搜尋龐大的測試文獻集。

醫學圖書館員在研究進行中指出，本次測試的檢索問題並非館員工作時遇到的典型檢索問題，此類難題在實際的工作環境中，通常需要更多的時間解決。此外，本研究的檢索任務（search task）需要找出最佳的檢索詞彙，以最後一次檢索結果作為評估標準，和一般使用者的檢索行為可能不同。因此，我們不應把本研究所得之醫學標題表的有用性，推論到其他比較不具技術性的檢索問題，或其他不同的檢索任務上。

本實驗發現若使用者具備相關領域的進階知識，在搜尋技術性的問題時，可受益於醫學標題表。先前有關控制詞彙索引法的評鑑研究，以醫學標題表和自動索引技術為例，都在實驗室環境下以批次模式進行研究（如Salton, 1972; Savoy, 2005），使用者並沒有實際使用系統搜尋文件。研究發現自動索引技術與控制詞彙比較，兩者隨機搜尋的檢索表現，可相互匹敵。在本實驗使用者實際參與搜尋，研究發現支持此項結論。為增進我們對控制詞彙在互動式檢索系統的瞭解，TREC互動檢索項目（interactive track）中使用者的個別差異（searcher differences）對檢索結果的影響（Lagergren & Over, 1998），值得進一步研究。

致謝

本研究獲美國國家科學基金會資助，計畫編號0414557，計畫主持人為Michael Lesk與Nina Wacholder。筆者感謝匿名受試者分享專業知識，Nina Wacholder、Michael Lesk、Paul Kantor、Nick Belkin及單中杰（Chung-Chieh Shan）諸位教授在研究進行期間的討論與建議，以及劉魯（Lu Liu）先生在系統技術上的支援。本文部分結果曾發表於美國資訊科學學會2008年年會（Liu & Wacholder, 2008），與會者及三位期刊匿名審查委員寶貴意見一併致謝。

參考文獻

吳美美（1993）。言談分析和資訊檢索互動研究。**教育資料與圖書館學，30**，340-350。

吳美美（1998）。「邏輯相關」和「適切相關」——中文資訊檢索系統評鑑量標初探。**圖書館學與資訊科學，24**(2)，44-64。

吳美美（2001）。**中文資訊檢索系統使用研究**。臺北市：學生書局。

林珊如（1995）。從終端使用者資訊檢索行為談圖書館的資訊服務政策及角色：實證研究之探討。**圖書與資訊學刊，13**，22-39。

唐牧群、鄧雅文、鄭瑋、謝宜瑾（2008）。佛學數位圖書館詞彙建議介面之評估研究。**大學圖書館，12**(1)，111-127。

黃慕萱（1997）。檢索系統評估之發展——理論與實務。**中國圖書館學會會報，59**，109-126。

Allen, B. (1991). Topic knowledge and online catalog search formulation.

Library Quarterly, 61, 188-213.

Aronson, A. R., Demner, D., Humphrey, S. M., Ide, N. C., Kim, W., Liu, H., et al. (2004). Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations. In E. M. Voorhees & L. P. Buckland (Eds.), *The thirteenth text retrieval conference proceedings (TREC 2004)*. Retrieved April 11, 2009, from http://trec.nist.gov/pubs/trec13/papers/nlm-umd-ul.geo.pdf

Belkin, N. J., & Kelly, D. (2010). *Proceeding of the third symposium on information interaction in context*. New York: ACM.

Belkin, N. J., & Kwasnik, B. H. (1986). Using structural representation of anomalous states of knowledge for choosing document retrieval strategies. In F. Rabitti (Ed.), *Proceedings of the 9th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 11-22). New York: ACM.

Bellardo, T. (1985). An investigation of online searcher traits and their relationship to search outcome. *Journal of the American Society for Information Science, 36*, 241-250.

Bibliographic Services Task Force. (2005). *Rethinking how we provide bibliographic services for the University of California*. Retrieved September 17, 2006, from http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf

Blair, D. C. (2002). The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. *Information Processing & Management, 38*, 273-291.

Borenstein, M. (2009). Effect sizes for continuous data. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221-235). New York: Russell Sage Foundation.

Borlund, P., Schneider, J., W., Lalmas, M., Tombros, A., Feather, J., Kelly, D., et al. (2008). *Proceedings of the second international symposium on information interaction in context.* New York: ACM.

Buckley, C. (1999). Trec_eval IR evaluation package [Computer software]. Retrieved November 16, 2007, from ftp://ftp.cs.cornell.edu/pub/smart/

Byrd, J., Charbonneau, G., Charbonneau, M., Courtney, A., Johnson, E.,

Leonard, K., et al. (2006). *A white paper on the future of cataloging at Indiana University*. Retrieved September 26, 2010, from http://www.iub.edu/~libtserv/pub/Future_of_Cataloging_White_Paper.doc

Calhoun, K. (2006). *The changing nature of the catalog and its integration with other discovery tools*. Retrieved August 14, 2006, from http://www.loc.gov/catdir/calhoun-report-final.pdf

Croft, W. B. (1989). Automatic indexing. In B. H. Weinberg (Ed.), *Indexing: The state of our knowledge and the state of our ignorance* (pp. 86-100). Medford, NJ: Learned Information.

Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different types of experiences. *Journal of the American Society for Information Science, 32*, 23-32.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Greenstone. (2006). Greenstone Digital Library Software (Version 2.70). Hamilton, New Zealand: Department of Computer Science The University of Waikato.

Hersh, W. R. (2008). *Information retrieval: A health and biomedical perspective* (3rd ed.). New York: Springer.

Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., & Kraemer, D. F. (2004). *TREC 2004 genomics track overview. The Thirteenth Text REtrieval Conference (TREC-13)*. Retrieved November 16, 2007, from http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf

Hersh, W. R., Bhupatiraju, R. T., Ross, L., Roberts, P., Cohen, A. M., & Kraemer, D. F. (2006). Enhancing access to the Bibliome: The TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration*. Retrieved from http://www.j-biomed-discovery.com/content/1/1/3

Howard, H. (1982). Measures that discriminate among online searchers with different training and experience. *Online Review, 6*, 315-327.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. *Proceedings of the ACM SIGIR Conference, 16*, 329-338.

Humphrey, S. M., Rogers, W. J., Kilicoglu, H., Demner-Fushman, D., & Rindflesch, T. C. (2006). Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology, 57*, 96-113.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50.

Ingwersen, P., & Jarvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Dordrecht, The Netherlands: Springer.

Jenuwine, E. S., & Floyd, J. A. (2004). Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. *Journal of Medical Library Association, 92*, 349-353.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. *Proceedings of the ACM SIGIR Conference, 21*, 164-172.

Lin, S.-J., & Belkin, N. (2005). Validation of a model of information seeking over multiple search sessions. *Journal of the American Society for Information Science and Technology, 56*, 393-415.

Liu, Y.-H. (2009). *The impact of MeSH (Medical Subject Headings) terms on information seeking effectiveness*. Unpublished PhD dissertation, Rutgers: The State University of New Jersey, New Brunswick, NJ.

Liu, Y.-H., & Wacholder, N. (2008). Do human-developed index terms help users? An experimental study of MeSH terms in biomedical searching. *Proceedings of the American Society for Information Science and Technology Annual Meeting, 45*, 1-16.

Lu, Z., Kim, W., & Wilbur, W. (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval, 12*(1), 69-80.

Mann, T. (2006). *The changing nature of the catalog and its integration with other discovery tools. Final report. March 17, 2006. Prepared for the Library of Congress by Karen Calhoun: A critical review*. Retrieved from http://www.guild2910.org/AFSCMECalhounReviewREV.pdf

MeSH Browser (2003 MeSH). (2004). Retrieved November 16, 2007, from http://www.nlm.nih.gov/mesh/2003/MBrowser.html

Nelson, S. J., Johnston, W. D., & Humphreys, B. L. (2001). Relationships in Medical Subject Headings (MeSH). In C. A. Bean & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 171-184). Dordrecht, The Netherlands: Kluwer Academic.

On the record: Report of the library of congress working group on the future of bibliographic control. (2008). Retrieved April 11, 2009, from http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

Pao, M. L., Grefsheim, S. F., Barclay, M. L., Woolliscroft, J. O., McQuillan, M., & Shipman, B. L. (1993). Factors affecting students use of MEDLINE. *Computers and Biomedical Research, 26*, 541-555.

R Development Core Team. (2009). R: A language and environment for statistical computing (Version 2.9.1). Vienna, Austria: R Foundation for Statistical Computing.

Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N. J., Tombros, A., & Vakkari, P. (2006). *Proceedings of the 1st international conference on information interaction in context*. New York: ACM.

Salton, G. (1969). A comparison between manual and automatic indexing methods. *American Documentation, 20*(1), 61-71.

Salton, G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Sciences, 23*, 75-84.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*, 648-656.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science, 39*, 197-216.

Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Information Processing & Management, 41,* 873-890.

Sparck Jones, K. (1981). Retrieval system tests 1958-1978. In *Information Retrieval Experiment* (pp. 213-255). London: Butterworths.

Sparck Jones, K. (2005). Epilogue: Metareflections on TREC. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and evaluation in*

information retrieval (pp. 421-448). Cambridge, MA: The MIT Press.

Srinivasan, P. (1996). Optimal document-indexing vocabulary for MEDLINE. *Information Processing & Management, 32*, 503-514.

Sutcliffe, A. G., Ennis, M., & Watkinson, S. J. (2000). Empirical studies of end-user information searching. *Journal of the American Society for Information Science, 51*, 1211-1231.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science, 37*, 331-340.

Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, MA: The MIT Press.

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

Text Retrieval Conference. (2005). *TREC 2004 genomics track document set* [Data file]. Available from NIST TREC 2004 Genomics Track Web site (http://trec.nist.gov/data/t13_genomics.html).

Tsay, M. Y. (2004). Literature growth, journal characteristics, and author productivity in subject indexing, 1977 to 2000. *Journal of the American Society for Information Science and Technology, 55*, 64-73.

Vickery, B. (1971). Document description and representation. *Annual Review of Information Science and Technology, 6*, 113-140.

Wacholder, N., & Liu, L. (2006). User preference: A measure of query-term quality. *Journal of the American Society for Information Science and Technology, 57*, 1566-1580.

Wacholder, N., & Liu, L. (2008). Assessing term effectiveness in the interactive information access process. *Information Processing & Management, 44*, 1022-1031.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). Managing gigabytes: *Compressing and indexing documents and images* (2nd ed.). San Francisco: Morgan Kaufmann.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 307-314). New York: ACM.