天主教輔仁大學圖書資訊學系碩士班

指導老師：曾元顯

查詢擴展之詞彙篩選應用於主題檢索之研究

Term-Selection for Query Expansion in Topic Information Retrieval

研究生：蔡育欽

中華民國九十四年六月三十日

摘要

幫助使用者滿足其資訊需求是資訊檢索技術發展之目標。使用者所輸入的查詢詞彙對檢索結果有著直接且顯著的影響，本研究將以自動化的方式進行查詢擴展進而提升檢索成效。查詢擴展隨著擴充詞之來源不同將之分爲局域擴展與全域擴展，局域擴展使用的擴充詞彙來自初次檢索結果中的關鍵詞彙，而本研究中全域擴展使用的是事先建立之共現索引典中的詞彙。目前局域擴展在眾文獻與本研究中皆可證實其穩固之成效，而全域擴展於之前的研究中顯示其成效不夠穩固。故本研究之目的希望建立自動的篩選規則將索引典中與查詢主題相關的詞彙加入初始查詢，提升全域擴展之檢索成效。

我們使用日本 NTCIR 單語檢索的資料建立實驗所需之環境，利用其提供的多欄位描述方式進行主題檢索。研究中我們先以人工方式篩選關聯詞彙，並根據人工篩選之經驗與文獻之分析結果，提出四種自動化篩選詞彙之策略，並以不同檢索模式與不同擴展方式交叉驗證其檢索成效。除量化分析外，研究也針對查詢主題之描述與篩選出的關聯詞兩者進行觀察分析，了解其對檢索成效之影響。

實驗結果顯示當查詢詞品質較高時，不同檢索模式的成效差異較小，而當查詢詞品質較不一致時，以機率模式檢索成效較高。且當擴展模式爲先全域擴展再局域擴展其檢索成效高於單獨使用局域擴展或全域擴展。全域擴展的檢索實驗中，人工的篩選結果有著不錯的成效，足可證實共現索引典於查詢擴展具有相當幫助，而自動化的篩選方式則以計算關聯詞對主題之強度的方式成效較佳，但整體成效幫助有限。由主題需求描述之觀察，發現當主題需求描述使用了較多”多意義詞彙”時，不僅難以查詢擴展提升成效，且檢索成效通常較差。觀察關聯詞對檢索成效之影響，選出的詞彙精確率越高，檢索結果越好，而回收率高但選出的詞彙與主題的關聯度較低時，檢索成效容易變差。

Abstract

The primary purpose to develop Information Retrieval is to satisfy users' information need. The influences of query words on the quality of the search results are direct and apparent; therefore, query expansion which suggests more related terms to users is often adopted to improve the effect of information retrieval. According to the different origins of expanded terms, query expansion is divided into two categories: Local Expansion in which the expanded terms are generated from the key words of the first retrieval results; and Global Expansion in which the expanded terms are from the established Co-occurrence Thesaurus. The effectiveness of local expansion is robust according to various earlier researches and this study; however, the effect of global expansion has not brought into full play. Hence, the aim of this study is to heighten the effect of global expansion by establishing an automatic term-selection rule to add related terms from thesaurus into the first retrieval.

The application of Japanese NTCIR Single Language Information Retrieval provides a multi-column description in topic-based information retrieval. In this study, on the base of the experiences of manual selection and the outcomes of literature review, we propose four strategies for automatic global term selection which have been evaluated by distinct retrieval models and various expanded approaches. Apart from

the quantitative analysis, this study focuses on the relationship between the descriptions of retrieval topic and the selected related terms as well.

The experiment results show that the difference among varied retrieval models is unobvious with the higher-quality query words; nevertheless, the effect of probability model is better with inconsistent quality query words. The better retrieval outcomes are taken place as the expansion models are developed from applying global expansion first and then local expansion later. In the experiment of global expansion, the usefulness of co-occurrence thesaurus in query expansion is proved by the effective of manual selection; on the other hand, automatic selection is good only when such selection is based on identifying the related degree between related term and topic but only has marginal total effect. The analysis of topic description indicates that the more “polysemy” be used the less effects be produced. In terms of the impact of related terms on the effect of retrieval, we discover that the better results are due to the higher accuracy of selected terms; however, the effectiveness reduces with high recall rate.

目次

圖目次

表目次

第一章 緒論

第一節 研究背景與動機

資訊檢索是現今爲滿足人類資訊需求十分重要的途徑。由於資料不斷的數位化，且大至全世界小至個人，不分種族、性別、年齡，皆有對資訊之需求。然而資訊量的增加將導致滿足資訊需求變的困難，因此唯有資訊檢索技術的提升方可快速的幫助使用者滿足其資訊需求。

爲因應廣大的使用者對資訊檢索之需求，世界各地資訊檢索之專家學者，無不積極開發更有效率、效能的檢索系統；然而開發過程中，經常面臨缺乏完整的實驗環境，對各自開發出來的系統進行公正與有效評估以利比較與掌控檢索系統之優劣。因此，由日本國立情報學研究所 （National Institute of Informatics，簡稱 NII）主辦的 NTCIR （NII Test Collection for Information Retrieval），即提供了一個完善的實驗環境供各家學者評估其開發的資訊檢索系統（陳光華，2004）。

資訊檢索系統所檢索出之資料是否能夠滿足使用者的需求，其最直接之因素莫過於檢索者所提供的查詢詞彙了。當使用者發生資訊需求時，必須將其所欲查詢之資訊轉化成文字的描述，然而此一過程往往隨著使用者的知識程度與背景環境之不同而產生變化，導致使用了錯誤或不適當之詞彙表達其資訊需求，使得檢索結果不符合需要。傳統上有索引典可以幫助使用者解決用詞的問題，比如以「參見」幫助使用者使用正確的詞彙，因此索引典一直以來便是圖書館必備之工具書，然而索引典的編制卻需要大量的人力與時間方可完成。而近代檢索技術也發展出以自動化的方式輔助使用者使用正確的查詢詞彙稱之爲查詢擴展。而自動化的技術可分類爲局域擴展（Local Expansion）與全域擴展（Global Expansion）兩者，而以後者較接近傳統處理之方式。

在 NTCIR4 競賽中曾元顯以全域擴展之方式進行查詢擴展之研究，然而當時所使用的篩選方式太過嚴謹，導致所擴充的詞彙數量太少，而無法有效提升檢索的成效（Tseng，2004）。本研究即是希望建立自動過濾關聯詞之規則並檢驗其是否可穩定運作於各個檢索模式（Retrieval Model），以獲得最佳的檢索結果。

第二節 研究目的

在資訊檢索的相關研究中，已經提出不少能夠有效的提升檢索效能的方法與技術，而最明顯能影響檢索的兩個因素，其一爲使用者所提供之查詢詞彙，另一爲檢索系統所使用的檢索模式。

根據 Hui Fang 研究指出使用者所提供之查詢詞彙是最直接影響系統成效的因素，輸入較多查詢詞的檢索成效會比輸入較少查詢詞的成效來的高，輸入較多需求描述的檢索也比輸入較少需求描述的檢索成效來的高，且不論是使用何種檢索模式，其岐異性並不大（Hui Fang，2004）。因此查詢詞彙對檢索效果較具有影響力，如果系統可以導引使用者輸入較多高品質的查詢詞，可直接使檢索結果的品質更優良。

除使用者所提供之查詢詞彙會影響檢索成效外，另一重要因素，也是資訊檢索非常重要的核心因素，就是檢索模式。目前資訊檢索的型態大致可分爲：布林式(Boolean)檢索、向量式(Vector-Based)檢索以及機率式(Probability-Based)檢索等（Ricardo，1999）；其中傳統上最常被廣泛使用的爲布林式檢索，對於使用者的查詢句，布林式檢索提供查詢詞彙和各文件之間的交集（AND）、聯集（OR）、差集（NOT）的運算，將使用者所要查詢的範圍加以縮小或擴大，以使查詢出來的結果更加符合使用者的要求。布林檢索的速度快，但缺少相關程度之類的計算而無法依照相關度來排序。因此本研究將使用向量模式與機率模式配合中文之環境實驗各檢索模式之成效。

查詢擴展使用局域擴展的方式已是相當普遍且有效的技術（Mitra, 1998），而全域擴展的技術在曾元顯的實驗與各研究中皆顯示並不十分穩固（Tseng，2004）（Xu, 1996）。因此本研究希望透過建立自動化的過濾規則提升使用者初次輸入的查詢詞的品質，並實驗各種檢索模式的效能，以檢驗全域擴展之規則適用

於各檢索模式並達到最佳的搭配方式提昇檢索的成效。

本研究之目的如下：

一、希望建立有效的過濾規則應用於 Global Feedback

二、評估不同查詢擴展與檢索模式對檢索成效之差異

第三節 研究貢獻

在查詢擴展的研究中，大多以局域擴展與全域擴展兩者爲主，且大部分研究中顯示使用局域擴展確實對檢索成效有幫助。無論使用何種資料集、不管是以自動或是由使用者介入的方式使用局域擴展，其檢索成效皆有明顯的提升（Harman, 1988）（Mitra, 1998）。

而使用全域擴展者，其成效並不十分顯著與穩定。在 NTCIR4 的實驗中，以自動化的方式進行，其成效並不十分顯著，雖有進步但卻不明顯。但某些查詢主題以人工方式篩選詞彙進行全域擴展，其成效進步了九成之多（Tseng，2004）。這表示了全域擴展如果可以有效的篩選擴充詞彙則可預期其成效將會顯著提升，但是如何篩選出相關的詞彙才是真正影響檢索成效的因素。

因此本研究即是希望提出各種可能之方式進行篩選的實驗，使全域擴展可以穩定的輔助使用者找出更多優良品質的擴充詞彙，以協助使用者檢索資訊。

第四節 詞彙定義

一、主題檢索（Topic Information Retrieval）

本研究中主題檢索之檢索方式非一般以標題或是分類號的方式描述需求，而是採用類似 TREC 所創的多欄位的「查詢主題」（Topic rather than Query），藉以多面向的欄位以描述需求。一般的主題檢索對需求之描述過短，其中包含的訊息可能不足，因而本研究採類似 TREC 的方式，對需求以多元的方式描述以及對檢索目的、檢索背景等其他層面之敘述以表達使用者需求，而 NTCIR 提供了中文主題檢索所需要的資料。

在 NTCIR3 資料集中，即以 Title 描述查詢需求之主題，通常爲名詞或是名詞片語，並對查詢主題做簡單描述；Description 則使用簡單的一至二個句子描述需求的主要內容；Narrative 則是使用數個句子描述查詢問題與專有名詞的定義，也描述了相關與不相關之範疇或其他特殊限制；Concept 則使用數個詞彙描述查詢主題中各層次相關的詞彙。

二、關聯詞（Related Terms）

本研究之關聯詞有別於一般人工建構的索引典，是採取自動建構的方法，即是計算單一文件中任意兩個詞彙個別出現與共同出現的句子數，再套用資訊檢索常用的相似度計算公式，求得兩詞彙之間的關聯強度。計算出每一篇文件重要詞彙之間的關聯強度後，累積關聯強度超過某個門檻值的關聯詞，即可完成整個文件資料庫的關聯詞庫。（曾元顯，2001）

產生關聯詞之步驟是先根據 NTCIR3 所提供之主題描述，並利用關鍵詞擷取演算法，透過統計其詞頻的方式自動分析出關鍵詞彙，將由主題描述中找出關鍵詞，再根據關聯詞庫找出與關鍵詞關聯強度高於門檻之詞彙，將這些與關鍵詞關聯度較高之詞彙視爲關聯詞。同時本研究之詞頻皆爲該詞出現之文件數目，而非

該詞出現之總次數。

三、查詢擴展（Query Expansion）

資訊檢索系統常利用查詢擴展（Query Expansion）的方式補足使用者所提供查詢詞的不足，以達到檢索成效的提升。查詢擴展對資訊檢索領域來說，是一種幫助檢索者在檢索過程中可以修正檢索詞彙且降低檢索者在檢索系統中繁雜的步驟。而查詢擴展依產生擴充詞的方式，粗略可分爲人工與自動，「人工」就是傳統圖書館中索引典的運用；「自動」則以擴充詞彙之來源不同而分成局域擴展與全域擴展，前者從第一次檢索結果或是相關的文件擷取相關詞彙作爲擴充詞彙，後者則以全部文件，經過自動化的整理組織之後而產出之擴充詞彙。

四、局域擴展（Local Expansion）

在本文中是指將第一次檢索結果的前 N 篇經過相關排序之後，抽取出前 M 個關鍵詞作爲擴充詞彙加入原先的查詢。也就是擴充詞的來源爲部份文件之集合。舉例如下：

例如：由原始查詢「一九九八年諾貝爾物理學獎」經過自動的斷字斷詞及產生關鍵詞後做查詢，並由查詢結果的前 6 篇文件中抽取 15 個詞彙作爲擴充詞彙加入初始的查詢，作爲第二次查詢的查詢詞彙（見表 1-1）。

表 1-1：局域擴展擴充詞彙之範例

原查詢		一九九八年諾貝爾物理學獎
原查詢詞彙	1 gram	一, 九, 九, 八, 年, 諾, 貝, 爾, 物, 理, 學, 獎
	2 gram	一九, 九九, 九八, 八年, 年諾, 諾貝, 貝爾, 爾物, 物理, 理學, 學獎
	關鍵詞	諾貝爾物理學獎, 諾貝爾, 物理, 物理學

擴充詞彙	崔琦，諾貝爾物理學獎，獲得，發現，研究 科學家，教授，朱棣文，三位，獲得諾貝爾物理學獎，諾貝爾獎，得諾貝爾物理學獎，物理學，華裔科學家，獲頒

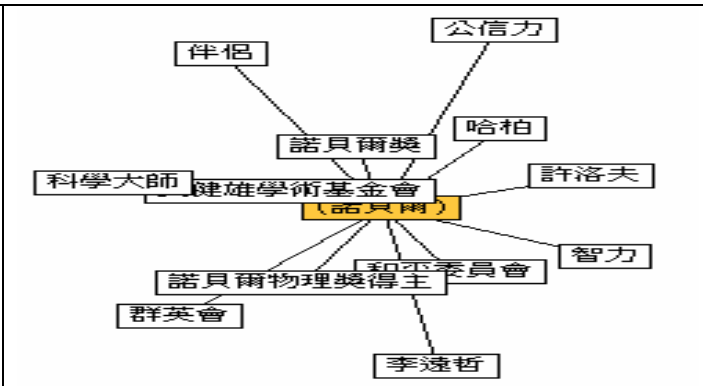
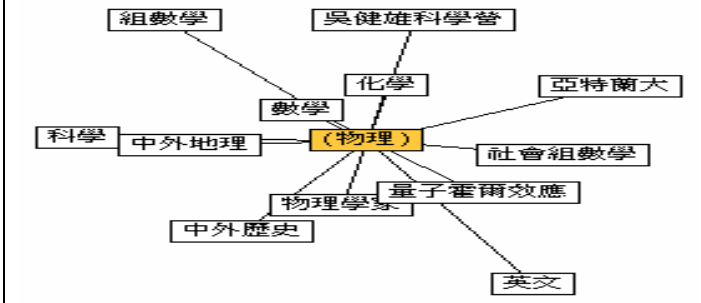
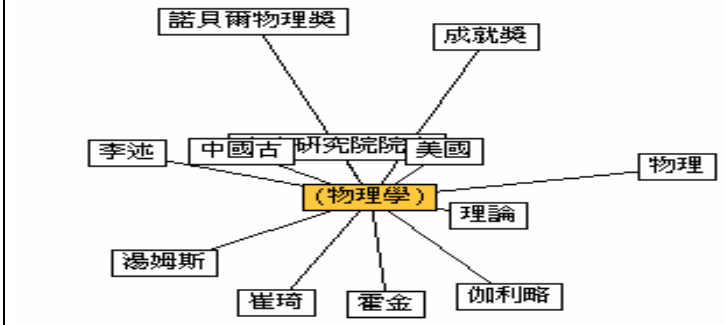
五、全域擴展（Global Expansion）

運用所有資料集之文件，建立一索引典之架構以定義出詞與詞之間的關係，並找出與查詢詞有關的詞作爲擴充詞彙，加入初始查詢。本研究中使用的方式即建立一共現索引典（或稱關聯詞庫），以該索引典中所提供之詞彙作爲擴充詞的來源，達成查詢擴充之目的。舉例如下：

由原始查詢「一九九八年諾貝爾物理學獎」經過自動的斷字、斷詞及產生關鍵詞後（Tseng，2002），將關鍵詞的關聯詞拿來作爲擴充詞彙：（見表 1-2）

表 1-2：全域擴展擴充詞彙之範例

關鍵詞	關聯詞（圖以 2D 方式呈現，中心爲關鍵詞，各分枝爲關聯詞）
諾貝爾物理學獎	

諾貝爾	 公信力 伴侶 諾貝爾獎 哈柏 科學大師 許洛夫 智力 諾貝爾物理獎得主 群英會 李遠哲
物理	 組數學 吳健雄科學營 化學 亞特蘭大 數學 科學 中外地理 （物理） 社會組數學 量子霍爾效應 中外歷史 英文
物理學	 諾貝爾物理獎 成就獎 李遜 中國古 美國 物理 （物理學） 理論 湯姆斯 崔琦 霍金 伽利略

第五節 研究問題

查詢擴展的技術對資訊檢索有著相當大的助益，本研究將以全域擴展之方式進行查詢主題的擴展之研究。研究將探討如何建立自動化篩選關聯詞的策略使關聯詞回饋的技術能夠獲得最大的發揮及提昇資訊檢索的成效，再搭配各種不同的檢索模式，以驗證其是否可以運作穩定。因此本研究將細部探討以下問題：

一、建立自動的關聯詞過濾規則

查詢擴展使用全域擴展的方式時，如果關聯詞庫中的詞彙未經過濾就直接送入，恐會對要檢索的主題產生主題漂移（Topic Drift）的情況。因此必須透過有效的篩選策略使主題不致偏移又能達到查詢擴展之目的。

在 NTCIR4 的競賽之中，使用的過濾規則爲（Tseng，2004）：

1.關聯詞的詞頻要低於關鍵詞的詞頻。

2.在規則一的限制下，關聯詞於同一查詢主題之下，出現於不同的關鍵詞中。

同時符合以上兩者的詞彙會被當作擴展的詞彙，將被加入原始查詢達到查詢擴展之目的；然而這樣的條件組合方式卻使得所擴充的詞彙大量縮減，平均每題只有 4-5 個擴充詞彙且有部分查詢主題甚至沒有擴充詞，如此一來便失去了查詢擴展之意義。

爲此，本研究即爲使全域回饋能發揮最大功效，擬建立出不使主題產生偏斜又能達到查詢擴展之目的的篩選策略。

二、驗證 Global feedback 效能是否穩定運作

在眾研究中，Local Feedback 已是穩固且有效的查詢擴展方式，而 Global

Feedback 之規則是否可以穩定的運作，本研究將針對此進行交叉驗證，即在不同擴展方式與檢索模式下，Global Feedback 之效能是否穩固。

本研究將比較下列不同的擴展方式：

- 不使用查詢擴展
- Local Feedback（L）
- Global Feedback（G）
- 先使用（G）再進行（L）

使用到的檢索模式有：

- ByteSize Normalization
- Pivoted Normalization Method
- BM11
- BM25
- BM25m（修改自 BM25）
- Language Model（Dirichlet prior retrieval method）

本研究將針對不同的查詢擴展方式以及不同的檢索模式進行實驗，以比較出不同模式、不同查詢擴展對檢成效之影響。

第二章 文獻探討

第一節 查詢擴展

使用適當的查詢詞彙表達資訊需求對一般檢索者是不容易的，且隨著知識背景不同使得彼此對相同主題的描述使用了不同的詞彙，而往往造成使用者必須不斷的修正自己所選用的詞彙以取得其欲檢索之資料。查詢擴展即是爲幫助檢索者重新修正查詢詞彙的方法，以更接近檢索者欲檢索的主題。而其運作的方式主要有互動與自動二者：互動方式，例如以提供友善的介面供檢索者選擇其需要的相關詞彙以重建查詢；而自動的方式則是不需要使用者介入即由系統自動加入相關的詞彙進行檢索。先前研究已對互動式檢索多所探討（葉佳昀，2004）（曾元顯，1998），因此本研究將對自動的查詢擴展方式進行相關文獻分析，並依照擴充詞彙來源分爲局域、全域與混何兩者之查詢擴展模式進行探討。

一、局域擴展（Local Expansion）

局域擴展主要目的在於透過初次檢索的結果，擷取其中與查詢相關之資訊，使其與原始查詢合併作遞迴式查詢。此法極類似傳統圖書館的檢索策略－引用文獻滾雪球法，其先決條件是事先掌握一篇或是數篇相關文章，利用這些相關文章尋找更多相關的文章，因此相關文章就像是珍珠或雪球一樣愈滾愈大。一般而言，引用文獻滾雪球法是由精確率反向追求回收率（黃慕萱，1996）。而現代化的資訊檢索方式如相關資訊回饋（Relevance Feedback），一般有兩種方式，一是在初次查詢句中加入更多相關的詞彙，二是調整部份查詢詞的權重（Ricardo，1999）。

在初次查詢句中加入更多相關的詞彙是相當有效且簡單的技術，其主要意義便是由初次檢索結果中由使用者判斷爲相關的文件中，取其文件之特徵如關鍵詞等，作爲擴充詞彙加入初始之查詢，藉以提高檢索成效；但由於需要人工方式介

入，無法自動完成。因此以 Blind Relevance Feedback 方式自動完成相關回饋，其主要意義便是將初次檢索的前 N 篇文件皆假設爲相關的情況下進行相關回饋，以初次檢索所得到的文件之中取其前 N 篇相關分數較高者或是適當的 N 篇文件，由這 N 篇文件中取其可用之資訊例如摘要或關鍵詞，加入下一次的檢索中，增強檢索時需要的資訊（Mitra，1998）。相關回饋對資訊檢索的成效非常有幫助，甚至在部分全文資料庫中可提升檢索成效 20%（曾元顯，1997），其基本的假設在於某些和查詢句相關之文件，透過相關排序後，相關的文件會排名在前面，擷取這些文件中重要的特徵回饋給系統去補充和增強查詢句之不足；然而在某些情形下違背了相關回饋基本的假設，將導致無法提昇檢索成效甚至降低了檢索的成績。

局域擴展的研究除了使用相關資訊回饋外，也有研究使用文字分析（Context Analysis）爲回饋的方式。其與相關回饋的最大不同便是以段落（Passage）取代文件，主要原因在於解決文件長度過長以及單一篇文件中包含多個主題之情形，而其具體作法爲：（Mitra，1998）

1. 將文件切割成段落，在該研究中以 300 字爲一個段落，作爲原始資料。由初次檢索的結果依相關排序取其前 N 的段落。
2. 由這 N 個段落中的詞彙形成個別主題。透過共現性的計算，找出與查詢詞彙相關的詞形成主題之概念。
3. 在個主題中取較爲相關的詞彙 M 個，加入初始查詢。

由於局域擴展的擴展詞是由第一次的檢索的結果而來，故擴展的詞彙來自局部的文件集。如果第一次的檢索中未檢索到重要的文件，將使相關回饋所提供的擴充詞無法發揮效果，因此也有研究對此進行探討，其針對自動查詢擴展的成效提升進行實驗，以 Blind Relevance Feedback（BRF）進行查詢擴展的方式，使用 TREC3 至 TREC6 的資料集進行測試，透過人工與自動化的方式進行擴充詞彙的

過濾並比較其檢索成效。其研究結果顯示以人工過濾查詢詞彙將使成效提升 7%-22%，而以其自動化之方式進行過濾可使成效提升 6%-13%，而其自動化步驟爲：（Mitra，1998）

1. 由初始檢索的 50 篇文件中取 20 篇作爲回饋的文件。
2. 將每一篇文件根據新的相似度計算公式重新排序，
3. 在新的排序的文件中以 Rocchio（1971）的方式進行查詢擴展。
4. 將以重新建構的查詢詞彙檢索出最後的相關文件。

查詢擴展使用 BRF 的方式進行，其整體成效確實有明顯之進步；但 Sakai 研究指出 BRF 也有導致部分查詢主題成效降低之情形。BRF 的擴展方式雖然使整體成效的平均值獲得顯著的提升，但觀察個別查詢主題約有三分之一個查詢主題使用了 BRF 的擴展技術後發生成效降低的情形。一般而言 BRF 運作良好大多是使用了大文件集之情形，因其可提供較多相關之文件作爲回饋。除非某些特殊之主題在初次檢索的結果中，因相關文件分部較不集中，而導致檢索成效降低。Sakai 使用對查詢需求之複雜度進行分析，並根據不同之查詢複雜度給予不同的擴展範圍，即是檢索結果前 n 篇文件的前 m 個詞彙，以不同需求複雜度給於不同的（n,m）值。（Sakai，2000）

1. Fixed Local Feedback，不管需求之複雜度，一律以檢索結果前 10 篇前 20 個詞作爲擴充詞彙。
2. Ideal Local Feedback 隨著不同需求調整 n, m 的值，取最佳化之結果。
3. Number of case particles in the request，以平假名出現次數爲複雜度之計算依據。
4. Number of search terms in the initial query，以描述需求之字數最爲複雜度之計算依據。
5. The highest score in the initial ranked output: 以檢索結果之最高成績做

爲複雜度之計算。

6. Top n` documents score: 以檢索結果前 n`篇文件之成績做爲複雜度之計算。

並依據每個策略之數據定義出不同複雜度之群組，並設定各群組以不同的 n, m 值進行查詢擴展。然 Sakai 的實驗結果並沒有明顯的使查詢結果獲得更多的提升；即使以很精密的方式使用 BRF 與一般使用方式相比並沒有顯著提升，且不論以精密方式或是一般方式做 BRF 的擴展相較於不擴展有很大的提升，因此查詢擴展使用局域擴展不論其使用人工經驗或細緻的科學方法對檢索之成效皆有顯著的提升。

二、全域擴展（Global Term Expansion）

全域擴展指的是在檢索前即組織全部文件之資訊，待查詢時提供與查詢字串相關之資訊，以補足初始查詢之不足。簡而言之即是以全部文件所形成之資訊作爲查詢擴展。以傳統方式即是類似以索引典等輔助查詢。傳統圖書館在檢索資料、查找書籍時，索引典爲必備之工具，用以在短時間之內取得所需的資料（鄭恆雄，1984）。透過索引可以使用權威控制確保詞彙使用的一致性，並清楚的描述詞彙與詞彙之關係，例如上位詞（Board Term, BT）、下位詞（Narrow Term, NT）、關聯詞（Related Term, RT），作爲詞彙的擴張，當查詢結果不足時，可以使用關聯詞找到相似主題的資料，或是使用上位詞擴大主題涵蓋的範圍；反之，當資料過多不知如何取捨時，則使用下位詞縮小範圍，但索引典的製作曠日費時，在資訊爆炸的今日已無法滿足資訊需求者，若能夠透過自動建構索引典的方式，則能應付大量的數位文件，並幫助使用者檢索資料。

自動化的方式大多利用共現性的特徵建構索引典以提供查詢擴展所需之資訊。根據 Salton 所提之架構在檢索前即建立好詞對之關係，建立關聯矩陣，以詞

頻爲基礎計算出各詞彙之相似度，並依相似度進行詞彙歸類，形成主題類別。

$T_j = (d_{1j}, d_{2j}, ..., d_{nj})$　詞彙在所有文件的權重

$$sim(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ij} d_{ik}}{\sqrt{\sum_{i=1}^{n} d_{ij}^2 \sum_{i=1}^{n} d_{ik}^2}}$$

而歸類時有兩種方式：

1. Complete-link

   先將每個詞彙視爲單獨的類別，並計算兩兩之相似度，如果達到一定標準者，則結合爲同一類別，不斷重複至兩個類別的相似度最低。將會產生多個類別，每個類別詞彙少。

2. Single link

   先將每個詞彙視爲單獨的類別，並計算兩兩之相似度，如果達到一定標準者，則結合爲同一類別，不斷重複至兩個類別的相似度最高。其特性會產生的類別較少，但每個類別詞彙較多。

以 Salton 之方式所建立的索引典最爲查詢擴展時，其檢索成效約提昇 10%-20%，但歸類的運算量過大，若文件太多將耗費大量時間。（曾元顯，2001）

Kwon 則使用既有人工編製完成的索引典爲基礎，由其中建立出詞彙與詞彙的關聯程度，以篩選出關聯程度較高者爲擴充詞彙；再根據延伸布林邏輯運算，計算出 Query 與 Documents 之相似度並進行排序。傳統的索引典含有豐富的關聯資訊，使用者可清楚找出上位詞（BT）、下位詞（NT）、同義詞（SYN）與關聯詞（RT）；但以此傳統的作法並不包含詞彙與詞彙之間的關聯程度易導致較差的檢索成效，因而在某些檢索系統則以詞彙與詞彙之間的連結次數作爲彼此的關聯程度。Kwon 則是根據不同的關聯情形而使用不同的公式，計算出彼此的關聯

程度。其做法則是先使用現成人工編製完成的索引典爲基礎，並以某特殊領域之文件爲資料集，計算每索引典中的詞彙在資料集出現之次數，並依據不同關聯狀況計算詞彙之關聯程度，不同關聯狀況之關聯程度公式如下：（Kwon，1994）

- BT：

  $$SR(A,B) = nf(\log_2 \frac{P(A,B)}{P(A)P(B)})$$

  where

  $$P(A,B) = \frac{freq(A,B)}{N}$$

  $$P(A) = \frac{freq(A)}{N}$$

  N = total number of words

  nf : normalize function[0,1]

- NT： $A \supset B$

  $$SN(A,B) = nf(\log_2 \frac{P(A,B)}{P(B)})$$

- SYN：

  $$SS(A,B) \approx 1$$

而其中下位詞使用了非對稱的方式計算兩詞彙之關聯程度，不同於一般使用對稱之方式計算相似度，且獲得不錯之成效。

三、整合不同查詢擴展

Harman 以三種不同查詢擴展方式進行實驗，其分別爲相關回饋、字尾擴展、索引典擴展進行實驗。根據其實驗結果顯示如果不過濾相關回饋所延展的擴充詞將導致檢索成效降低 2.8%，而使用擴充詞彙有個數限制的方式可以提升成效，當限制在 20 個擴充詞時有最好的成效，具體提升了 1.8%，而使用人工過濾則有

提升 8.0%，使用人工並加上個數限制則僅提升 4.1%，因此經過過濾的擴充詞彙則較會提升成效；而結合三種查詢擴展之成效實驗中，以同時結合三種成效最佳，且任兩種的組合皆比單獨使用一種來的好（Harman，1988）。

陳光華的實驗將使用者建構的查詢問句，在檢索前先以同義詞典或索引典進行詞彙擴展，再使用擴展後的查詢問句進行檢索，其進行五種實驗以探討索引典與同義詞典對檢索效益之影響，其實驗如下：

1. 基礎檢索
2. 同義詞典擴展（以共現性自動建構）
3. 索引典擴展（人工編纂）
4. 同義詞典擴展，索引典加權
5. 同義詞典擴展，索引典加權與二次擴展

歸納其結論得知同義詞典擴展，索引典加權的成效最好，其次爲同義詞典擴展，但同義詞典擴展後的查詢詞數量最好在 8~10 個詞彙之間，而以人工索引典擴展除了狹義詞擴展外其他方式的詞彙關係擴展成效均比基礎檢索來的好，唯同義詞典擴展，索引典加權與二次擴展只比基礎檢索來的好。歸納導致成效下降之因素最重要者莫非是雜訊之因素，不論使用何種擴展之方式都將容易導致過度擴展導致主題漂移之現象，因爲多次擴展或是錯誤擴展後容易產生較不相關的詞彙。因此如果能找出規則，以規則濾除雜訊詞彙，去蕪存菁，預期可大大的減少偏移之情形，保持查詢擴展之成效（陳光華，2001）。

Mandala 運用三種索引典做查詢擴展之實驗，其結合三種不同類型之索引典，將三種索引典中詞彙之相似度取其平均值作爲擴展詞彙排序依據。各索引典及細節如下：

1. Hand-crafted thesaurus based（WordNet）

詞彙網路（WordNet）是以人工發展之索引典，於 1990 年普林斯頓（Princeton）所發展的英語詞網。並計算其中兩兩字彙的最短距離作爲相似度。

2. Co-occurrence-based automatically constructed thesaurus based
   主要是計算資料集中兩兩詞彙共同出現在一篇文件裡的機率作爲相似度。
3. Head-modifier-based automatically constructed thesaurus based
   此方式是基於語言學之概念將詞彙歸類，而非以統計共同出現之文件。主要假設爲兩詞彙出現在相似的文法或文章脈絡中，則兩詞彙視爲相關詞。並依據不同的文法套用不同的機率公式取得兩兩詞彙之相似度。

最後將三種索引典之相似度結合。以各別相似度正規劃後，再取三種相似度之平均值，以平均相似度排序，並根據相似度排序做查詢擴展。

Cui 使用 Local Feedback 方式建立索引典運用於 Global Feedback。Cui 的實驗大致上以初次檢索的 Query Log 計算出查詢詞及其關聯詞之相似度，經過時間的累積後形成索引典，再根據此索引典做爲全域擴展的來源。其檢索紀錄包含有查詢詞與被點選文件編號，並假設被點選之文件與查詢詞有某種程度的相關性，也許實際上不見得相關，但可被視爲似乎是相關的。這樣的假設雖沒有使用者判斷來的精確但卻比 PRF 來的好，且可以自動化進行。Cui 花費兩個月的時間由 Encarta(http://encarta.msn.com) 記錄了約 4 百多萬筆的檢索紀錄。圖 2-1 表示了檢索紀錄之資訊，並由其機率之公式計算出查詢詞彙與文件詞彙共同出現之機率如圖 2-2。（Cui，2002）

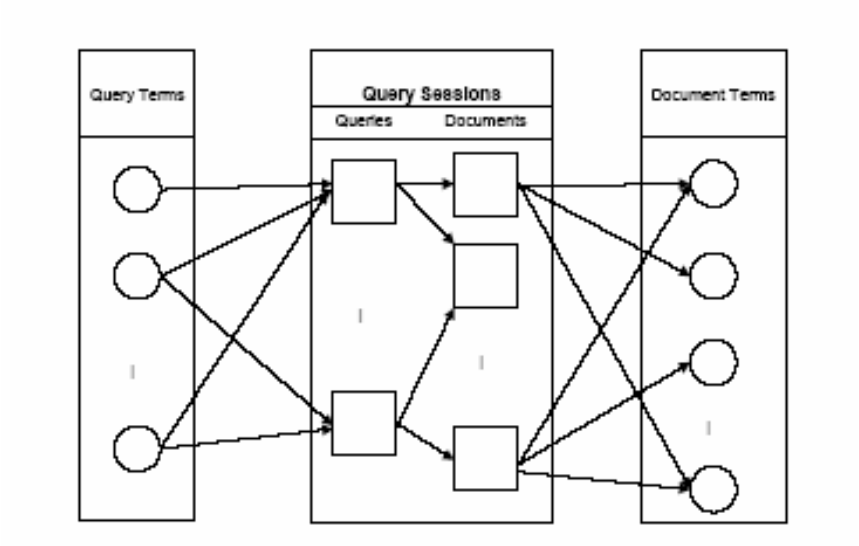


圖 2-1：Cui 實驗中所使用的檢索紀錄結構

資料來源：Cui, Hang., Wen, J. R., Nie, J.Y., & Ma, W.Y.（2002）. Probabilistic query expansion using query logs. Proceedings of the 11th international conference on World Wide Web, 325-332.

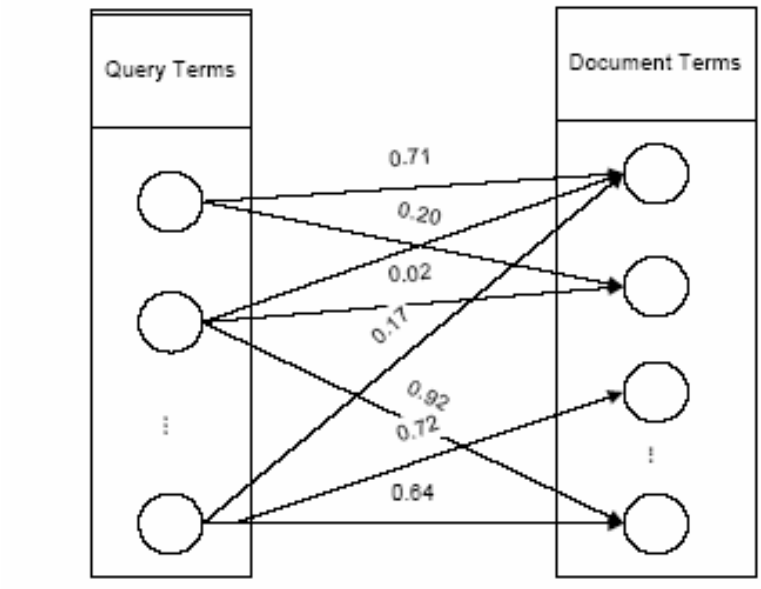


圖 2-2：Cui 實驗中查詢詞彙與文件詞彙共現機率示意圖

資料來源：Cui, Hang., Wen, J. R., Nie, J.Y., & Ma, W.Y.（2002）. Probabilistic query expansion using query logs. Proceedings of the 11th international conference on World Wide Web, 325-332.

當進行查詢擴展時只需要依據其機率公式結合文件詞彙出現之機率，並加以排序選出是當的詞彙作爲關聯詞即可。其結合機率之公式如下：

$$CoWeight_z(w_j^{(d)}) = \ln(\prod_{w_t^{(q)} \in Q}(P(w_j^{(d)} \mid w_t^{(q)}) + 1))$$

其實驗結果中，30 個查詢主題，每題選則排序前 50 個詞彙作爲擴充詞，相關判斷的結果判斷正確的比例成長了 32%，其檢索結果平均精確率相較於傳統的 Local Feedback 成長了 39%，因此其進步非常顯著。

第二節 檢索模式

檢索模式是檢索系統的核心，透過檢索模式將使用者輸入的查詢詞彙與系統中的文件進行比對，以幫助使用者在大量文件中，利用自動化的方式找到其所需要的資料。而傳統的布林模式既簡單又快速，但由於布林檢索語法的輸入對一般使用者不易且其比對模式採二元方法導致檢索結果不夠精確；而現代檢索系統多採向量式或統計式檢索模式，其皆可避免使用者輸入麻煩的查詢語法，而改以以自然語言的方式輸入查詢詞彙，並能計算文件與查詢詞彙的相似程度，以此最為排序的準則，對使用者而言相當簡便。研究者將對布林模式、向量模式與機率模式之文獻進行探討。

一、布林模式（Boolean Model）

布林模式的檢索方式是一種傳統且簡單的檢索方法，在過去幾年被很多的圖書資訊系統採用。其透過集合理論（Set theory）與布林代數（Boolean algebra）運算，因為此種檢索方法不僅速度快，而且在實現上亦十分容易。通常可透過一些關鍵詞與邏輯運算元（Logical operators）所組成的聯集（Disjunction）與交集（Conjunction）布林語句（Boolean expression），表示使用者想要檢索的資訊需求。由於集合的概念相當直覺，很容易讓使用者所理解與掌握，在今日仍有許多大型系統採用此類的檢索模式。（Ricardo，1999）

然而布林模式主要缺點便是它屬於一種非 0 及 1 的二分法（Binary decision criterion），也就是每筆被檢索的資料只有兩種狀態：相關或不相關，而沒有分數或相似程度的機制在內，造成檢索結果的不精確。

舉例而言以指令【$k_a \wedge (k_b \vee \neg k_c)$】將檢索出圖 2-3 中灰色區域，此為一標準的布林表示法，並代表要搜尋的內容為包含關鍵字k_a且包含關鍵字k_b或不含k_c的資料，其中 $\wedge$、$\vee$、$\neg$ 分別代表布林運算AND、OR及NOT，布林檢索的優點

爲實現方法簡單、速度快，且查詢語法十分明確，但缺點則是檢索的結果往往容易產生兩極化的現象：高失敗率(search failure)或高溢檢率(Information Overloading)。（Ricardo，1999）

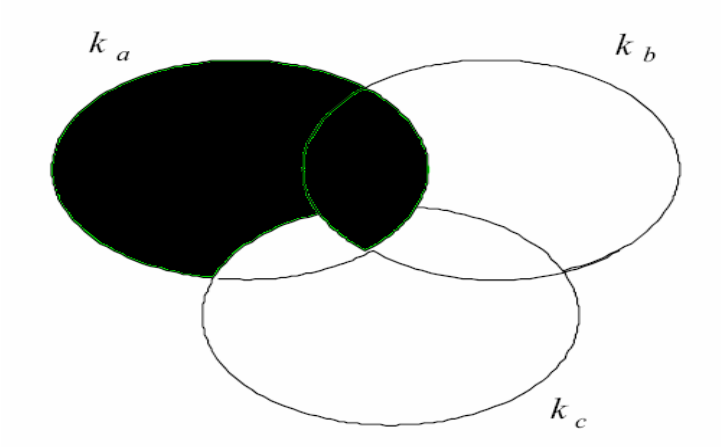


圖 2-3：布林邏輯範例示意圖

資料來源：Ricardo,B.Y.,&Berthier,R.N.（1999）.Modern Information Retrieval. New York：Addison Wesley.

二、向量模式（Vector Space.Model）

向量模式的檢索模型是由 Salton 等人在 1971 所提出，爲提升檢索系統之效能及解決布林模式的諸多限制，其作法爲：（Ricardo，1999）

1. 將使用者的詢問句及資料庫中的文件轉換成維度（Dimension）相同的向量表示法。

   $q=[W_{1,Q},W_{2,Q},\ldots W_{n,Q}]$，代表查詢向量

   $dj=[W_{1,dj},W_{2,dj},\ldots W_{n,dj}]$，代表文件向量

   將文件與查詢句轉換爲向量之後，就可以用量化方式處理，並計算其相似

度。例如最常被使用的Cosine運算。

2. 以Cosine來計算兩向量的夾角，其值介於 0 到 1 之間。公式如下：

$$sim(q,d_j)=\frac{d_j \bullet q}{|d_j|\times|q|}=\frac{\sum_{i=1}^{t} w_{i,dj}\times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,dj}^2}\times\sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

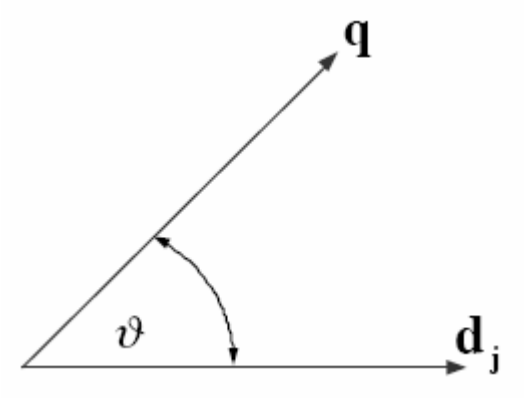


圖 2-4：向量模式示意圖

資料來源：Ricardo,B.Y.,&Berthier,R.N.（1999）.Modern Information Retrieval. New York：Addison Wesley.

藉由Cosine運算得出相似度的值，當兩向量夾角爲 0 時，其Cosine值爲 1 最大值，而當兩向量夾角爲 90 度垂直時，其Cosine值爲 0 亦即相關度爲 0；最後再將所有文件以相似度加以排序，得出相似程度的排名。

3. 索引項目權重的計算對檢索成效有很大的影響，適當的權重，有助於提升檢索的正確率。一般常用tf-idf，其公式如下：

$$tf_{i,dj}=\frac{C_{i,dj}}{\sum_{k=1}^{t} C_{k,dj}}$$ 代表i詞在文件j中出現頻率

$$idf_i=\log\frac{N}{n_i}$$ N代表文件筆數，n_i詞i出現的文件筆數

$$w_{i,di}=tf_{i,dj}\times idf_i$$

而根據Buckley所建議的權重公式如下：

$w_{i,dj} = (0.5 + 0.5tf_{i,dj}) \times idf_i$　可以產生平滑的作用

三、機率模式（Probability Model）

以機率的概念處理資訊檢索的問題。先建立一組相關文件作爲理想解答來當作訓練文件，透過遞迴的方式計算資料間分布的關係，使得查詢句對應到文件的機率計算能達到最大，最後得到一組收斂的參數。當資料集夠大時，該參數將可代表一般情況下理想的分布情形。因此機率模型一開始是以猜測的方式取得初步的理想解答，系統再利用此資訊改善理想解答，經重複多次處理後接近實際的理想解答（Ricardo，1999）。

在此假設R定義爲與查詢Q相關的文件集，$\overline{R}$ 爲R的補集（Complementary Set），則P(R|d_j)爲文件d_j相關機率，而P($\overline{R}$|d_j)視爲文件d_j不相關的機率，則相似度爲：

$$sim(d_j, q) = \frac{P(R \mid \overrightarrow{d_j})}{P(\overline{R} \mid \overrightarrow{d_j})}$$

由貝式法則（Baye's rule）拆解爲：

$$sim(d_j, q) = \frac{P(\overrightarrow{d_j} \mid R) \times P(R)}{P(\overrightarrow{d_j} \mid \overline{R}) \times P(\overline{R})}$$

其中 P(R)爲文件中隨機取到相關文件的機率。由於對所有文件而言，P(R)或 P($\overline{R}$) 都是一樣的，所以公式簡化爲：

$$sim(d_j, q) \cong \frac{P(\overrightarrow{d_j} \mid R)}{P(\overrightarrow{d_j} \mid \overline{R})}$$

四、不同檢索模式對成效之影響

Hui Fang 之實驗以 7 個基本且合理的規則驗證各檢索模式。由於傳統以來常用的各檢索模式雖然有效，但其公式的產生皆非完整之理論結果，也因此 Hui Fang 等分析出基本的規則，透過這些規則驗證各檢索模式之公式，穩定的檢索系統必會符合規則，反之則不一定。假若能夠符合這些規則則代表公式本身是可信賴的，反之當公式無法達到這些基本的假設，則應修改之。以下爲規則條件：

1. Term Frequency Constraints （TFCs）

   用以確保檢索模式可以給查詢詞出現次數較多的文件高分（TFC1），或是文件包含更多查詢詞彙者成績較高（TFC2），或是確保高詞頻查詢詞次數的變化對成績的影響力低於低詞頻次數的變化（TFC2）。

2. Term Discrimination Constraint（TDC）

   檢測TF（詞頻）與IDF（出現文件數）對成績的影響力

3. Length Normalization Constraints（LNCs）

   當查詢詞詞頻相同時，要使長文件分數低於較短文件的分數（LNC1），同時又要能避免過度降低長文件的分數（LNC2）。

4. TF-LENGTH Constraints （TF-LNC）

   調整TF與文件長度對成績之影響

研究中被檢驗的模式有：

向量模式使用 Pivoted Normalization Method，其爲目前最佳的向量模式之公式，公式如下：

$$\sum_{w \in q \cap d} \frac{1 + ln(1 + ln(c(w,d)))}{(1-s) + s\frac{|d|}{avdl}} \cdot c(w,q) \cdot ln\frac{N+1}{df(w)}$$

Okapi Method-new，其爲目前最佳的機率模式之公式，公式如下：

$$\sum_{w \in q \cap d} \left(ln\frac{N - df(w) + 0.5}{df(w) + 0.5} \times \frac{(k_1+1) \times c(w,d)}{k_1((1-b) + b\frac{|d|}{avdl}) + c(w,d)} \times \frac{(k_3+1) \times c(w,q)}{k_3 + c(w,q)} \right)$$

語言模式（Language Model）使用 Dirichlet Prior Method，其爲目前最佳的語言模式之公式，公式如下：

$$\sum_{w \in q \cap d} c(w,q) \cdot ln(1 + \frac{c(w,d)}{\mu \cdot p(w|C)}) + |q| \cdot ln\frac{\mu}{|d| + \mu}$$

同時根據其驗證結果顯示原有之 Okapi Method 無法通過，而其他檢索模式大多可以通過。

進一步以長查詢與短查詢進行檢索成效測試，其結果顯示原有之 Okapi Method 之檢索成效最差，而其餘三者成效均差不多。其中更加顯示出所有的長查詢不論使用何種檢索模式均比短查詢成效來的好。如表 2-1 所示，使用者輸入較多關鍵詞(（lk）均比輸入較少關鍵詞（sk）好，使用較長的需求描述（lv）也比較短的需求描述（sv）好，因而可以明顯的解釋出影響資訊檢索的兩大因子：第一便是使用者所輸入的查詢詞彙，第二則是檢索系統所使用的檢索模式（Hui Fang，2004）。

表 2-1：Hui Fang 查詢長度與檢索模式於各資料集之檢索結果

		AP	DOE	FR	ADF	Web	Trec7	Trec8
lk	Piv	0.39	0.28	0.33	0.27	—	—	—
lk	Dir	0.38	0.28	0.32	0.25	—	—	—
lk	Okapi	0.38	0.27	0.28	0.33	—	—	—
lk	Mod-Okapi	0.39	0.28	0.28	0.33	—	—	—
sk	Piv	0.23	0.18	0.19	0.22	0.29	0.18	0.24
sk	Dir	0.22	0.18	0.18	0.21	0.30	0.19	0.26
sk	Okapi	0.23	0.19	0.23	0.19	0.31	0.19	0.25
sk	Mod-Okapi	0.23	0.19	0.23	0.19	0.31	0.19	0.25
lv	Piv	0.29	0.21	0.23	0.21	0.22	0.20	0.23
lv	Dir	0.29	0.23	0.24	0.24	0.28	0.22	0.26
lv	Okapi	0.03	0.07	0.09	0.06	0.23	0.08	0.11
lv	Mod-Okapi	0.30	0.24	0.25	0.23	0.28	0.26	0.25
sv	Piv	0.19	0.10	0.14	0.14	0.21	0.15	0.20
sv	Dir	0.20	0.13	0.16	0.16	0.27	0.18	0.23
sv	Okapi	0.08	0.08	0.08	0.09	0.21	0.09	0.10
sv	Mod-Okapi	0.19	0.12	0.16	0.14	0.25	0.16	0.22

資料來源：Hui Fang, Tao Tao, ChengXiang Zhai.（2004）. A Formal Study of Information Retrieval Heuristics. Proceedings of the 27th annual international conference on Research and development in information retrieval, 50-55.

第三節 檢索成效評估

本研究分爲兩大實驗分別進行實驗與評估，在關聯詞自動過濾規則之建立上將使用相關評估的方式進行規則之驗證與比較；而在不同檢索模式的成效實驗上，將使用 NTCIR 所提供的文件進行成效評估。

一、關聯詞自動過濾規則之評估

1. 精確率(precision ratio)與回收率(recall ratio)：

精確率與回收率將根據以下資訊進行評估，以最具代表性的 2X2 表格來說明：

表 2-2：精確率、回收率和雜訊比之２乘２表格

	相關	不相關	總計
被檢索出	TP	FP	TP+FP
未被檢索出	FN	TN	FN+TN
總計	TP+FN	FP+TN	TP+FP+FN+TN

資料來源：S. E. Robertson, “The Parametric Description of Retrieval Test,” Journal of Documentation 25:1 (1969).P.3.

TP（True Positive）：代表相關文章被檢出的筆數。

FP（False Positive）：代表不相關文章被檢出的筆數。

FN（False Negative）：代表未被檢出之相關文章筆數。

TN（True Negative）：代表正確回絕之不相關文章筆數。

回收率(recall ratio)：

意指檢出之相關文獻佔所有相關文獻的比例。公式如下：

$$\text{回收率（R）} = \frac{TP}{TP+FN} = \frac{\text{檢索所得之相關文章筆 數}}{\text{資料庫中所有相關文章 筆數}}$$

精確率(precision ratio)：

意指相關文獻於檢索出文章中所佔的比例。公式如下：

$$\text{精確率（P）} = \frac{TP}{TP+FP} = \frac{\text{檢索所得之相關文章筆 數}}{\text{檢索所得之所有書目筆 數}}$$

F-VALUE

一般而言，精確率與回收率常呈現反比的關係，也就是以較嚴格的過濾規則常會得到較高的精確率，但此時其回收率也較低；而我們爲了實驗出較好的規則會時常更動不同的方式與參數的設定，必須結合以上兩種數據作爲準則，F-VALUE 即爲此種評估方式。公式如下：

$$\text{F-VALUE（F）} = \frac{2PR}{P+R}$$

二、檢索模式的成效評估

以上的傳統評估方式較適用於無排序的檢索結果（Non-Ranked Search Results），然今日之搜尋引擎除了重視檢索到的資料外，更在意其與主題相關之

程度作爲排序依據，因此而使用 Non-Interpolated Average Precision Rate（NAP）較爲理想，其公式如下：（Ricardo，1999）

$$\text{NAP} = \frac{\sum_{1}^{N} \frac{i}{Rank}}{N}$$

舉例來說，在檢索結果的文件中，實際與主題相關的文件被排名在第一名、第三名、第五名、第六名，則 NAP 的值約爲 $0.7\bar{3}$ 故使用 NAP 測量值除了評估檢索的文件數之外還兼顧了排序的正確性。

$$\text{NAP} = \frac{\frac{1}{1}+\frac{2}{3}+\frac{3}{5}+\frac{4}{6}}{4} = 0.7\bar{3}$$

除 NAP 之外，TREC 使用 Buckley 發展的 TREC_EVAL 程式，用以更精確的評估排序的搜尋結果。至此之後 TREC_EVAL 資訊檢索系統常用的評估工具。

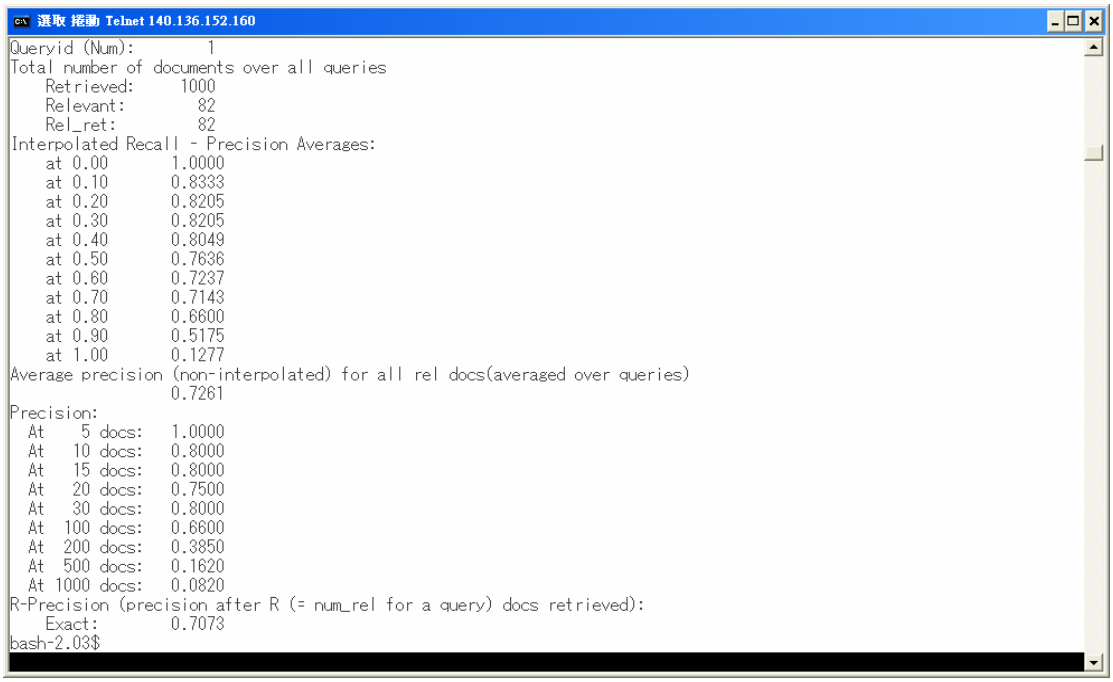


圖 2-5：TREC_EVAL 執行結果畫面

利用 TREC_EVAL 工具，對於評估檢索結果有相當大的幫助，其中可以得到的重要數據如下：（陳光華，2004）

- “Interpolated Recall - Precision Averages”是所謂 11-point Precision，以內差法的方式估計在固定的 Recall 下其相對的 Precision 值。
- “Average precision for all rel docs”是平均每篇相關文件被檢索出時的 Precision值，其公式如下：

$$Average\,\Pr ecision = AP(j) = \frac{\sum_{i=1}^{rj} \frac{i}{\#Doc_j(i)}}{r_j} = NAP$$

r_j：系統在查詢主題編號j所檢索出相關文件數。

$\#Doc_j(i)$：系統對查詢主題編號j中，在第i篇相關文件被檢索出時，總共被檢索出的文件數。

- “Precision At X docs”表示檢索出 X 篇文件時的 Precision。
- “R-Precision”表示檢索出 R 篇文件時的 Precision，R 是真正相關的文件數。

TREC_EVAL 個數據之細節整理於表 2-3

表 2-3：TREC_EVAL 各項數據之意義

欄位		數據	意義
Queryid (Num)		1	檢索題號。
Total number of documents over all queries	Retrieved	1000	檢索系統檢索出的文件數。
	Relevant	82	與查詢主題相關的文件數。

	Rel_ret	82	系統檢索出與主題先關的文件數。
Interpolated Recall - Precision Averages	At 0.00	1.000	Recall 爲 0 時，Precision 爲 1。
	At 1.00	0.1277	Recall 爲 1 時，Precision 爲 0.1277。
	Average precision (non-interpolated) for all rel docs(averaged over queries)	0.7261	平均精確率爲 0.7261。
Precision	At 5 Docs	1.000	前 5 篇的 Precision 值是 1。
	At1000 Docs	0.0820	前 1000 篇的 Precision 值是 0.0820。
	R-Precision (precision after R (= num_rel for a query) docs retrieved)	0.7073	前 R 篇時的 Precision 值。（此題 R 爲 82）

因此 TREC_EVAL 確實提供相當有效且可靠之數據，作爲檢索系統評估時非常具有代表性的評估方式。

第三章 研究方法與設計

第一節 研究方法

本研究將使用「實驗研究法」(Experimental Method）進行，藉助實驗程序以發現因果關係或比較各個變項（Variables）之結果，進行多次的實驗並觀察實驗結果，尋求現象的因果關係。其基本原則爲藉著操控（X）變項以瞭解（Y）變項的改變情形，其基本原則如下：(葉至誠，2003）

- 當一變項（X）改變時，另一變項（Y）是否也會跟著變動。
- 及是否只有變項（X）改變時，才會造成（Y）變項的改變。

同時根據實驗法之特徵：

- 實驗本身提供有意義的評估結果
- 同時運用必要的程序以控制已知的變異來源。
- 根據實驗方式以進行統計分析。
- 同一時間上實驗許多因素。

因此本研究將以實驗法進行，運用 NTCIR3 的資料集作爲測試，對各個變項的變異方向與變異量進行瞭解，再使用 NTCIR4 的資料作爲測試，收集各個變項的變化對測試結果所發生的效應。

第二節 研究流程與架構

本研究的流程與架構如圖 3-1 所示：

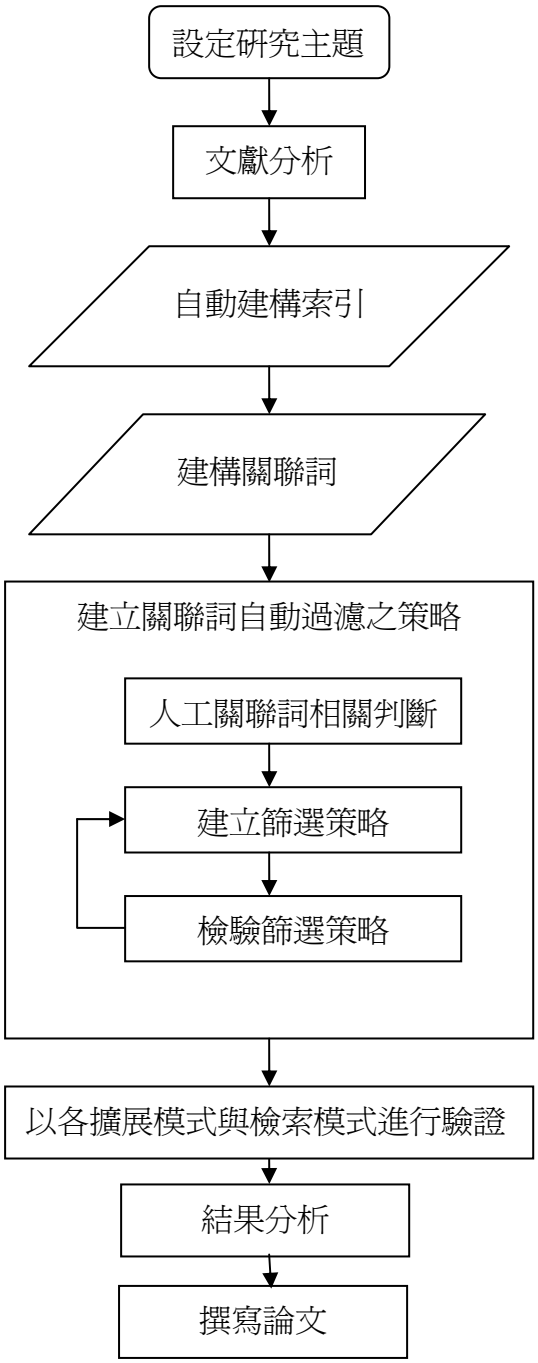


圖 3-1：研究流程圖

各步驟內容如下：

一、選定主題與研究方向

以學習過程與文獻之調查，決定欲進行研究之方向，並透過互動討論決定主題之細節，並開始思考研究目的與貢獻並開始相關文獻之搜集。

二、文獻分析

首先根據 NTCIR4 之研究結果進行相關文獻探討。並逐步模擬實驗，以實際瞭解檢索系統之細節，最後開始針對主題相關的研究進行文獻蒐集與分析。文獻蒐集之範圍包含有查詢擴展、檢索模式、評估方式三大主題。

三、自動建構索引

由於實驗需建立自動化的關聯詞過濾規則，因此透過 NTCIR3 的資料集做爲訓練之資料，並將其訓練成果運用於 NTCIR4 以驗證結果之可行性；故首先應該建立符合實驗需求之實驗環境，因此實驗將以 2002 所舉行之 NTCIR3 跨語言檢索競賽之資料做爲訓練資料並利用威知資訊公司所提供之 WebGenie 試用版軟體建立索引。

四、建關聯詞

由 WebGenie 建索引時也同時建構好共現索引（關聯詞庫），運用其 API（Application Program Interface）介面以曾元顯老師所寫的 PERL 程式將關聯詞庫由系統匯出，研究者再整理出可供使用者進行相關判斷之格式與內容。

五、建立關聯詞自動過濾之策略

1. 人工關聯詞相關判斷

全域擴展對檢索成效之影響爲本研究之重點，爲建立自動化之規則以提升檢索之成效，故將所有關聯詞以人工方式過濾之，透過人工的方式過濾出對檢索成效有幫助的關聯詞，期建立出能正確提升檢索結果的關聯詞篩選規則，以作爲建立自動化過濾規則之解答。

2. 建立篩選策略

透過人工過濾關聯詞之經驗，提出自動化規則之假設。透過實驗，瞭解各變項對檢索成效之影響進行假設。

3. 檢驗篩選策略

以人工過濾之關聯詞爲正確之資料集，對假設之自動化過濾規則進行實驗驗證，驗證方式以精確率、回收率、F-VALUE 之數據作爲驗證之依據，如果成效不理想則重複上一步驟，進行規則之假設，直到得到理想之成效爲止。

六、以各擴展模式與檢索模式進行驗證

爲確認全域擴展之成效，除以 NTCIR3 做爲訓練與測試外，更輔以 2004 年 NTCIR4 跨語言檢索競賽之資料集作爲測試資料集，將關聯詞自動化過濾之規則應用於 NTCIR4 之競賽，透過測試資料集之幫助，對全域擴展的成效進行測試。

此外，測試方式爲瞭解全域擴展之詞彙的篩選能夠穩定運作於各種情況，因此以不同檢索模式搭配進行各項實驗。

七、結果分析與撰寫論文

第三節 研究設計

本研究之目的爲利用全域擴展的方式，提升檢索成效。透過自動化過濾之方式，由系統自動產生的詞彙中篩選出與主題較爲相關的關聯詞，進而使檢索成效獲得提升。

一、自動建構索引

本次實驗主要採用的訓練資料集爲日本 NTCIR 會議，由其中的單語檢索競賽所使用的中文資訊檢索測試集 1.1 版（CIRB011, Chinese Information Retrieval Benchmark version 1.1）和 2.0 版（CIRB020, Chinese Information Retrieval Benchmark version 2.0）;而 CIRB011 皆下載自五個新聞網站於 1998 年 5 月至 1999 年 5 月間的報導，這些新聞包括：中國時報、工商時報、中時晚報、中央日報以及中華日報。CIRB020 則下載自聯合新聞網站於 1998 年 1 月至 1999 年 12 月底的報導。選擇新聞文件作爲評估的資料集，主要是取決於新聞文件內容主題豐富且符合現實生活的環境，其評估出來的結果可信度較高。表 3-2 詳列文件集的數量與大小（NII, 2003）（陳光華，2001）。

表 3-1：CIRB011 之文件來源與比例分佈

新聞媒體	文件數	百分比
中國時報(Chinatimes)	38,163	28.8%
工商時報(Chinatimes Commercial)	25,812	19.5%
中時晚報(Chinatimes Express)	5,747	4.4%
中央日報(Central Daily News)	27,770	21.0%
中華日報(China Daily News)	34,728	26.3%
Total	132,173	(200MB)

資料來源：陳光華（2001）。資訊檢索系統的評估－NTCIR會議，台灣大學圖書資訊學系四十週年系慶研討會，69-73。

表 3-2：CIRB011 與 CIRB020 之文件含概年代與數量

資料集	文件數量
CIRB011 (1998-1999): Chinese	132,173
CIRB020 (1998-1999): United Daily News (1998-1999): Chinese	249,508

資料來源：NII. （2003）. README for Topics and Relevance Assessments of NTCIR-3 CLIR Test Collection - <Formal runs>. Retrieved March, 15, 2005. From http://research.nii.ac.jp/ntcir/permission/READMEforTOPICS_FormalRun.htm

NTCIR 文件格式，包括：文件來源識別碼、新聞報導日期、新聞標題、新聞內容等，爲使系統易於辯識，每篇文件都具有相同的格式與資料項目。文件採用 BIG5 編碼與 XML 型態的標記，將標記加入文件之後便可以辨識文件的特殊區域。文件範例如圖 3-2 所示。

```
<DOC>
<DOCNO>cdn_chi_19980508_0002</DOCNO>
<LANG>CH</LANG>
<HEADLINE> 張榮恭強調江澤民所提不能視爲統一時間表 </HEADLINE>
<DATE>1998-05-08</DATE>
<TEXT>
<P>
針對外傳中共將在全國對臺工作會議中以「促進中國完全統一」做爲未來五年的重大政治任務，
中國國民黨大陸工作會主任張榮恭昨日則表示，這項說法只不過是一種政治口號和目標而已。
</P>
<P>
據香港「星島日報」引述北京的消息來源報導，中共「國家主席」江澤民預定於即將召開的全國對臺工作會議中，
在「江八點」的基礎上提出可操作性的「促進中國完全統一」的新對臺政策，
並將把「完全統一」列爲中共今後五年的重大政治任務。
</P>
<P>
張榮恭表示，中共已故元老鄧小平在八十年代曾說希望實現中國統一，
而中共領導人在一九九o年的全國對臺工作會議中也講過類似「號召統一大業」的話，
因此，不必解讀爲中共對統一的具體時間表，而僅是一種政治號召。
</P>
<P>
不過，他認爲，
中共預定五月十一日起在北京舉行全國對臺工作會議，
身兼中共中央對臺工作小組長的江澤民極有可能在這次會議上發表有關對臺問題的重要談話。
</P>
</TEXT>
</DOC>
```

圖 3-2：NTCIR 資料集文件標記範例

本實驗將採用威知資訊公司所提供之 WebGenie 試用版軟體建立索引，再以 Perl 撰寫程式匯出關聯詞。

首先必須先將測試資料集轉換成資料庫之格式，由於測試文件集本身是 XML 標記格式，無法直接使用 WebGenie 來建立索引，因此必須先撰寫轉檔程式，將 XML 以 Regular Expansion 的方式轉換成 CSV 檔案格式並匯入資料庫（見圖 3-3），其欄位可分爲：

- DocID　代表文件編號
- DocDate　代表文件日期
- DocTitle　代表文件標題
- DocContent　代表文件內容

待轉檔完成後再以 WebGenie 建立索引及產生關聯詞庫。

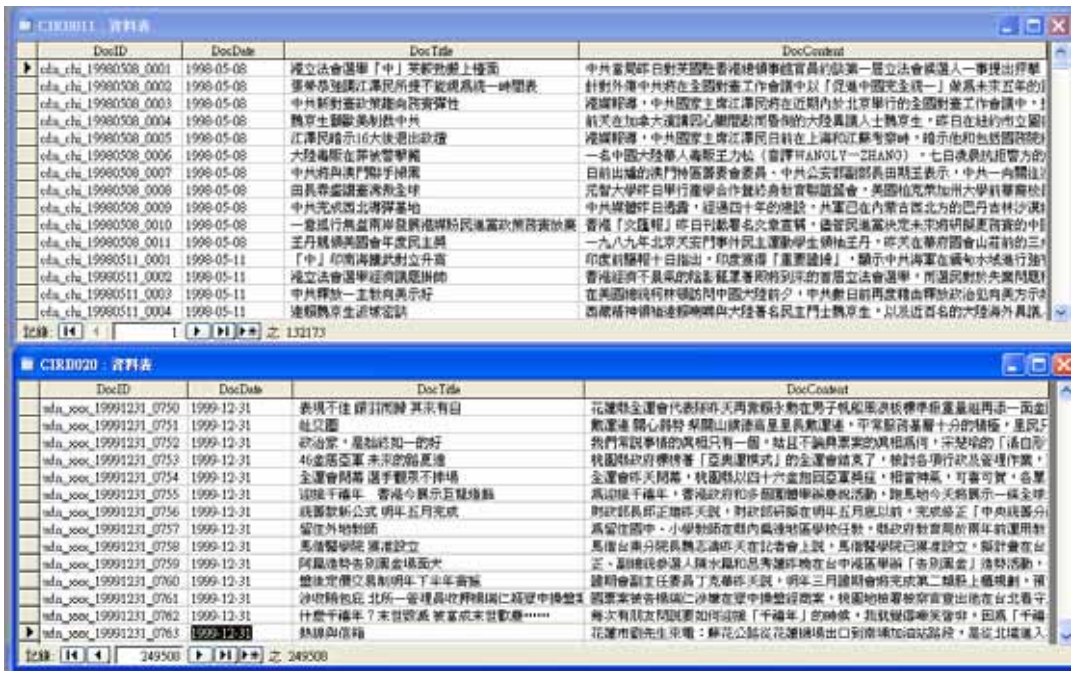


圖 3-3：NTCIR 資料集轉檔後資料庫範例

二、 建構關聯詞

建完索引之後，根據威知資訊所提供之 API 文件將關聯詞庫匯出，並整合由曾元顯教授以 Perl 程式所開發之檢索系統搭配使用。透過修改原系統之關聯詞過濾模式，由程式輸出每個查詢主題所使用到的所有未經過濾的關聯詞，並匯整至資料庫中以方便統計計算，資料庫共設計有 7 個欄位（見圖 3-4），其欄位可分爲：

- ID　　關聯詞 ID
- TOPIC　查詢主題
- Query　關鍵詞
- Wdf　　關鍵詞詞頻
- KW　　關聯詞
- Df　　關聯詞詞頻

- Related　　人工相關判斷，0 表示不相關，1 表示相關

查詢2：選取查詢

ID	TOPIC	Query	wdf	KW	df	related
1921	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	風暴	14416	1
1922	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	世紀	16110	0
1923	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	紀錄	16559	0
1924	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	因素	17858	0
1925	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	現象	19489	0
1926	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	中心	42950	0
1927	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	經濟	51873	0
1928	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	政府	81626	0
1929	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰 現 象	308	台灣	92557	0
1930	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰	454	秘魯人	10	0
1931	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰	454	西北太平洋	11	1
1932	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰	454	赤道東太平洋	13	1
1933	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰	454	大氣總署	14	1
1934	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰	454	NINO	15	0
1935	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰	454	蒜頭產量	15	0
1936	何謂反聖嬰現象及其與聖嬰現象的比較與影響	聖 嬰	454	藤森	17	0

記錄: 7 之 8144

圖 3-4：彙整各查詢主題之關聯詞資料庫

三、 人工關聯詞相關判斷

以人工對關聯詞進行相關判斷，判斷將根據關聯詞與查詢主題是否相關爲依據來判斷，實驗將以二分法將所有關聯詞分類爲與查詢主題相關或是不相關。由於使用 NTCIR 的資料集之故，其包含的主題包括有政治、財金、社會綜合、生活、體育、娛樂、國際、資訊、科技等新聞資料。並不需要太過專業的知識即可完成大部分的相關判斷，只要少部份需要自行尋找相關資料並對主題進行瞭解之後再進行相關判斷。

初步判斷原則將以人工方式對查詢主題之需求瞭解後，以較嚴格的方式進行相關關聯詞的選擇，並假設如果選擇了該關聯詞之後是否可能會提升檢索成效。實驗將以此判斷結果爲依據，對自動過濾規則的假設進行驗證。

然而經過第一次人工過濾之後，發現檢索成效提升有限，經過深入逐題調查後，歸納整理出兩大因素：

- 實驗環境與真實環境之差異
- 判斷者本身之知識背景

因此以修正人工初步判斷之方式，目的在於希望能夠有更好的檢索成效，使平均精確率提升。

修正方式爲先將關聯詞做第一次過濾，利用 NTCIR 所提供之答案集作爲過濾之資訊，判斷如果關聯詞有出現在答案集之中則再進行人工之判斷；反之，如果關聯詞不在答案集之中則先過濾掉，即直接認爲是不相關的詞。目的是希望瞭解 NTCIR 所提供之答案集被判斷爲與主題相關之因素，並提升檢索之精確率。然而發現這樣的修正方式雖然提升精確率，卻可能影響到回收率或真實的相關判斷，因此實驗計畫將同時保留初步判斷結果（簡稱人工一）與修正後的判斷結果（簡稱人工二）。

四、提出自動過濾規則

經過關聯詞的人工相關判斷之後，在其過程之中逐漸累積出一些個人經驗；並加以文獻分析之過程，根據其他學者之研究，整理出以下策略做爲過濾規則：

策略 一 (S1)：以關聯詞與關鍵詞之詞頻爲基礎之過濾規則

策略 二 (S2)：關聯詞遞迴擴展

策略 三 (S3)：依據關聯詞與主題之強度

策略 四 (S4)：依據關聯詞之檢索結果

策略 一 (S1)：以關聯詞與關鍵詞之詞頻爲基礎的過濾規則

經過關聯詞的判斷之後，以 NTCIR4 所使用之規則，進行測試與觀察。

R1.關聯詞的詞頻要低於關鍵詞的詞頻。

R2.在規則一的限制下，關聯詞於同一查詢主題之下，出現於不同的關鍵詞中。

只有同時符合已上兩規則之關聯詞會被加入原查詢據進行全域擴展。

表 3-3：人工初次判斷結果與 NTCIR4 之規則分析

代號	人工判斷相關	規則一	規則二	次數
A	0	0	0	2488
B	0	0	1	1247
C	0	1	0	2976
D	0	1	1	354
E	1	0	0	233
F	1	0	1	145
G	1	1	0	502
H	1	1	1	199

註：欄位”人工判斷相關”的值以 0 代表不相關，1 代表相關。欄位”規則一”與”規則二”的值以 0 代表不符合規則，1 代表符合規則。

由表 3-3 發現以 NTCIR4 所使用之規則可以正確的判斷出不相關的關聯詞，及代號（A+B+C）其正確率達 96.6%，但同時也將許多相關的詞也誤判爲不相關，如代號（E+F+G）其誤判率達 81.6%。明顯的顯示出 NTCIR4 所使用之規則太過於嚴格導致相關的詞彙都被判定爲不相關了。

且初步實驗後發現有不少實際相關的關聯詞被過濾掉的原因是關鍵詞的詞頻過低，而無法通過 R1，自然也就沒有辦法通過 R2 的規則，所以通常低頻關鍵

詞的關聯詞被判斷爲與主題相關的機率較高，而高頻關鍵詞的關聯詞被判斷爲相關的機率較低，因此根據表 3-3 可推測，低頻關鍵詞比較具有主題的判別性，故首先應將關鍵詞分爲低頻與高頻兩者，在進行後續的關聯詞過濾。

並以人工之經驗歸納出幾點現象可做爲關聯詞過濾規則假設之依據：

1. 部分主題不擴張成績較好

   據實驗觀察部分主題進行查詢擴張成效並不理想，例如 NTCIR3 中有一查詢主題是關於「大學學術追求卓越計劃」其斷出關鍵詞「學術」與「計畫」，但此二關鍵詞皆是詞頻較高的詞，因此容易導致擴展結果偏離主題。

2. 關聯詞選到同義詞，對查詢結果最有幫助

   由實驗觀察中發現如果關聯詞選到了同義詞則使檢索結果大爲提升。例如查詢主題「台灣加入 WTO 後各產業可能面對的問題」，在進行 Global Feedback 時，由人工選詞選到了 WTO 的同義詞「世界貿易組織」與「世貿組織」，使得成效獲得大幅提升。反觀 Local Feedback 選出的詞彙中大多是高頻詞造成主題之嚴重偏離，使成效大打折扣（如表 3-4）。

表 3-4：同義關聯詞之範例與其檢索結果

擴展方式	送入的查詢詞	BS
不使用擴展	台灣,灣加,加入,入 WTO,WTO,後各,產業,業可,可能,能面,面對,對的,的問,問題	0.0803
Global	WTO,產業,台灣,農民,入會案,農業,開放市場,**世界貿易組織**,產業,農委會,農產品,台灣,國際組織,關稅,立場,WTO,**世貿組織**,競爭力	0.2480
Local	台灣,ＷＴＯ,入ＷＴＯ,WTO,台灣加,大陸,美國,入 WTO,中國大陸,台灣加入ＷＴＯ,中共,連戰,石廣生,談判,經濟	0.0279

G -> L	WTO,產業,台灣,農民,入會案,農業,開放市場,世界貿易組織,產業,農委會,農產品,台灣,國際組織,關稅,立場,WTO,世貿組織,競爭力,WTO,ＷＴＯ,農業,農委會,入ＷＴＯ,組織,大陸,開放,中共,因應,諮商,對策,因應對策,減少,芒果	0.2669

3. 關聯詞選到廣義詞（詞頻高於關鍵詞）

   廣義關聯詞與主題相關，對檢索有幫助；反之，當廣義關聯詞與主題不相關，則會導致主題漂移，但對檢索成效所產生之影響並不大。

4. 關聯詞選到狹義詞（詞頻低於關鍵詞）

   狹義關聯詞與主題相關，對檢索較有幫助；反之，當狹義關聯詞與主題不相關，導致嚴重主題漂移。如表 3-5 所示，查詢主題「有中新一號衛星及評論」使用 Global Feedback 以人工選詞時，選錯了「中華衛星」與「華衛」使得成效大幅滑落，而調查出「中華衛星」與「華衛」皆是「衛星」一詞之關聯詞，且其爲高頻詞，但其實兩詞與查詢主題並不相關，可見篩選狹義詞的重要性將大於廣義詞的篩選。

表 3-5：狹義關聯詞之範例與其檢索結果

擴展方式	送入的查詢詞	BS
不使用擴展	有中,中新,新一,一號,號衛,衛星,星及,評論	0.7505
Global	衛星,新加坡電信,**中華衛星**,太空計畫室,衛星通訊,衛星,通訊衛星,**華衛**,太空科技,發射場	0.1688
Local	衛星,一號,新一號,中新一號,中新一號衛星,發射,租用,中華衛星一號,操控,調查局,衛星操控,建立,啓用,工作,時間	0.7549
G -> L	衛星,新加坡電信,中華衛星,太空計畫室,衛星通訊,衛星,通	0.2251

	訊衛星,華衛,太空科技,發射場,一號,中華衛星一號,發射,衛星,新一號,中新一號,十二,中新一號衛星,號衛星,發射場,頻段,實驗,國科會,美國,科學	

因此根據上述之現象，表示出關聯詞的篩選應首重同義詞的選擇，其次爲狹義詞的篩選，最後爲廣義詞的篩選，並建立以下 7 點規則：

定義：

R：關聯詞集合

r：關聯詞

k：關鍵詞

df()：詞頻，則 df(r)即爲關聯詞詞頻，df(k)爲關鍵詞詞頻。

策略一之規則：

規則一(R1)：df(r) < df(k)

若關聯詞詞頻低於關鍵詞詞頻，則將該關聯詞加入原始查詢。

規則二(R2)：R(r1) = R(r2)

若關聯詞於同一查詢主題之下，出現於不同的關鍵詞中，則將該關聯詞加入原始查詢。

規則三(R3)：R1 & R2

同時符合規則一與規則二之關聯詞，將作爲擴充詞彙。

規則四(R4)：IF df(k)<T:R1 ELSE:R2

若關鍵詞頻低於某一門檻，則以規則一的方式篩選關聯詞；否則以規則二的方

式篩選關聯詞。

規則五(R5)：IF df(k)<T:R1’ ELSE:R2

若關鍵詞頻低於某一門檻，則以相似於規則一的方式篩選關聯詞，相似於規則一的方式定義於下；否則以規則二的方式篩選關聯詞。

R1’：df(r)<N*df(k);1<N

此爲相似於規則一的方式，即如果關聯詞詞頻低於 N 倍的關鍵詞詞頻，而 N 必須大於 1，使得過濾的門檻降低，以容納更多的關聯詞。

規則六(R6)：IF df(k)<T:R1 ELSE:R3

類似規則四，僅關鍵詞高於某一門檻時，以較高之門檻過濾關聯詞。即若關鍵詞頻低於某一門檻，則以規則一的方式篩選關聯詞；否則以規則三的方式篩選關聯詞。

規則七(R7)：IF df(k)<T:R1’ ELSE:R3

類似規則五，僅關鍵詞高於某一門檻時，以較高之門檻過濾關聯詞。即若關鍵詞詞頻低於某一門檻，則以相似於規則一的方式篩選關聯詞；否則以規則三的方式篩選關聯詞。

策略 二 (S2)：關聯詞遞迴擴展

根據關聯詞彼此關聯的特性，以關聯詞遞迴擴展的方式過濾與篩選之策略。即假設由查詢詞所擴展出的關聯詞，兩兩關聯詞之間有某種程度的關聯性，以此補足上述策略一規則二（R2）過於嚴謹導致擴充詞彙太少的缺點，希望透過此方式將原本不符合 R2 的好詞也作爲擴充詞。以觀察 R2 的方式所篩選出之關聯

詞發現有許多與主題相關的關聯詞，因爲沒在其他關鍵詞中出現，故無法被選出作爲擴充詞，例如查詢主題「一九九八年諾貝爾物理學獎」爲例，以 R2 之方式可以找出不少相關度高的詞彙，如「量子」、「崔奇」等，然而有些詞彙可能因爲詞頻過低之影響，無法達到 R2 之規則，如「量子霍爾效應」，因此將以本策略之概念，以關聯詞彼此相關之想法地回擴展關聯詞，做法如下：

1. 由查詢主題（TOPIC）擷取關鍵詞（KW）
2. 由關鍵詞擴展出關聯詞（RT）
3. 取出符合 R2 之關聯詞，並視爲新的關鍵詞（KW’），不符合者視爲候選詞（nRT）
4. 由新的關鍵詞擴展出新的關聯詞（RT’）
5. 取出符合 R2 之新的關聯詞
6. 將上步驟與候選詞取交集者，即爲新的擴充詞彙

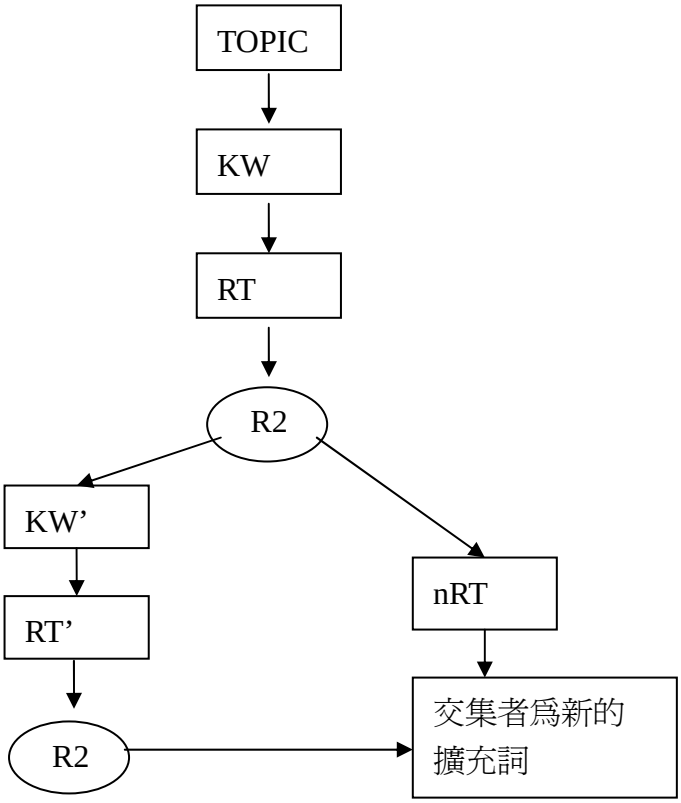


圖 3-5：策略二流程圖

透過上述之方式，將可發現「量子霍爾效應」被「崔奇」與「量子」共同關聯到（見圖 3-6），同時又是原關鍵詞「物理」的關聯詞，故雖然不符合策略一 R2 之規則，但經上述方式也可擴展出相關的關聯詞，即是「量子霍爾效應」。

表 3-6：策略二各階段篩選之詞彙以「一九九八年諾貝爾物理學獎」爲例

查詢主題	一九九八年諾貝爾物理學獎	
關鍵詞	諾貝爾物理學獎,諾貝爾,物理,物理學	
關聯詞	KW'（符合 R2)	諾貝爾物理獎得主,**量子**,法國,教授,華人,得主,科學,**崔琦**,因子,理論,吳健雄科學營,世界,學者,化學,亞特蘭大,美國,丁肇中,華裔科學家,台灣

	nRT	革命性,定律,譯名,知識,年輕人,中外地理,普大,經濟,成就獎,年輕科學家,香港,電機,院士,電信,防曬劑,霍金,研討會,評審委員,醫師,中心,伴侶,伽利略,固態,**量子霍爾效應**,一九二年代…等
新的關聯詞（符合R2)	經濟,研討會,醫師,物理學,專家,作家,哈柏,**量子霍爾效應**,科學家,諾貝爾物理獎,實驗室,知識,朱棣文,香港,中心,醫學,學生,瑞典,大會,吳健雄學術基金會,博士,諾貝爾獎,物理,院士,固態,論文,馬來西亞,人類,記者會,中國,政府,李遠哲,會議,環境,活動…等	
新的擴充詞彙	知識,諾貝爾物理獎得主,朱棣文,物理,量子,法國,哈柏,教授,華人,經濟,香港,院士,得主,研討會,中心,醫師,科學,固態,量子霍爾效應,崔琦…等	

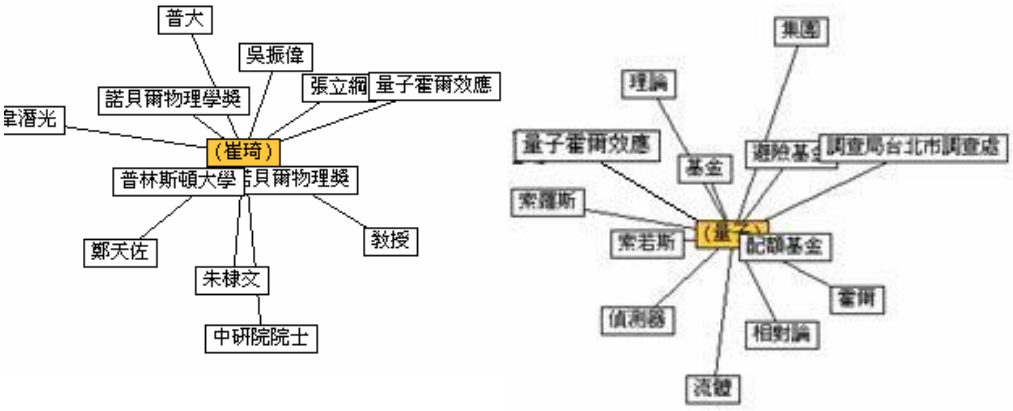


圖 3-6：策略二關聯詞範例

策略 三 (S3)：依據關聯詞與主題之強度

策略三以關聯詞與主題之強度作爲排序篩選的依據，即計算出關聯詞對整個

查詢主題之強度，而非對個別關鍵詞之強度。Qiu 與 Frei 即曾利用相似索引典以擴展整個查詢主題爲目標，而非擴展個別獨立的查詢詞並獲得不錯之成效（Qiu，1993）。因此，本研究即以此基本之概念，計算關聯詞與主題之間的強度作爲篩選之依據。上述策略關聯詞之篩選皆以二元之概念進行，也就是關聯詞與主題不是相關就是非相關，因而無法對其進行相關強度之排序，因此以計算關聯詞與主題之強度來篩選關聯詞。做法如下：

1. 由查詢主題（TOPIC）擷取關鍵詞（KW）
2. 由查詢主題（TOPIC）擷取二字詞（2-gram）
3. 由關鍵詞擴展出關聯詞（RT）

計算每一關聯詞與主題之強度

各種計算強度之公式如下：

- Dice Coefficient（Dice）：

$$S1_r = \sum_{k \in Q} \frac{2 \times df\ (r \cap k)}{df\ (r) + df\ (k)}$$

- 非對稱(nDice)取關聯詞爲基：

$$S2_r = \sum_{k \in Q} \frac{df\ (r \cap k)}{df\ (r)}$$

- Dice Coefficient（Dice）配合 IDF：

$$S3_r = \sum_{k \in Q} \frac{2 \times df\ (r \cap k)}{df\ (r) + df\ (k)} \times IDF$$

- 非對稱(nDice)取關聯詞爲基配合 IDF：

$$S4_r = \sum_{k \in Q} \frac{df\ (r \cap k)}{df\ (r)} \times IDF$$

$$IDF = \frac{\log\left(\frac{N}{df(r)}\right)}{\log(N)}$$

r : 關聯詞

k: 查詢詞

df(r): 關聯詞詞頻

df(k): 查詢辭詞頻

df(r∩k): 關聯詞與查詢詞共同出現詞頻

4. 根據排序選出關聯詞

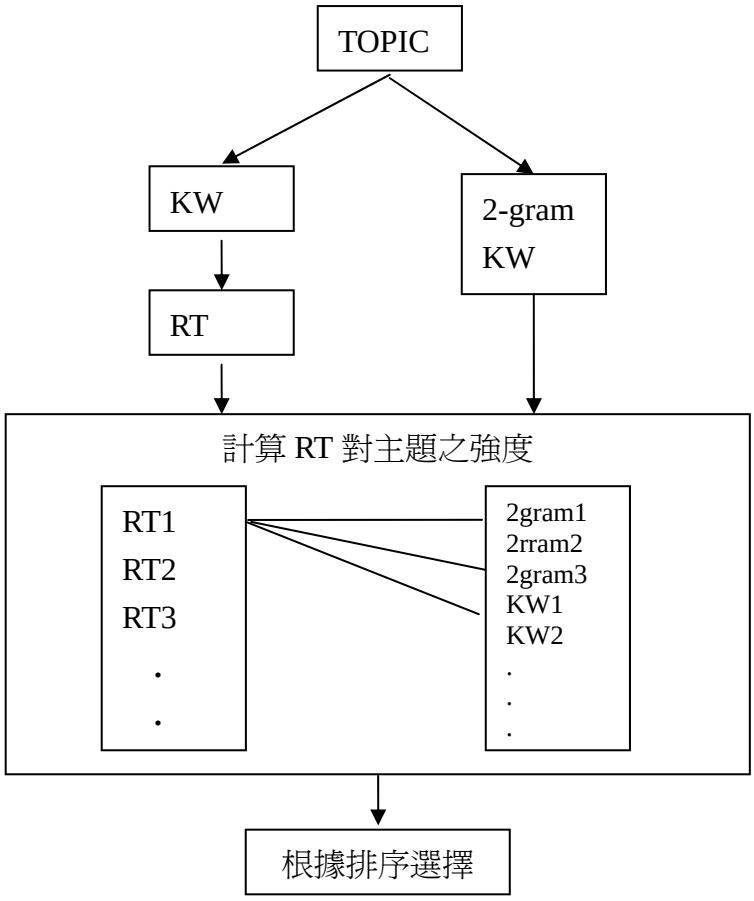


圖 3-7：策略三流程圖

表 3-7：策略三關聯詞篩選之詞彙以「一九九八年諾貝爾物理學獎」爲例

查詢主題	一九九八年諾貝爾物理學獎
2-gram	一九，九九，九八，八年，年諾，諾貝，貝爾，爾物，物理，理學，學獎
關鍵詞	諾貝爾物理學獎，諾貝爾，物理，物理學
關聯詞 (Top 10)	韋潛光，量子霍爾效應，崔琦，許洛夫，吳健雄學術基金會，朱棣文，瑞典皇家科學院，吳健雄科學營，丁肇中，粒子物理

策略 四 (S4)：依據關聯詞之檢索結果

策略四主要概念在於以關聯詞之檢索結果判讀其是否與主題相關。以人工判斷之過程中，對於部分難以人工明確分辨爲相關或不相關的關聯詞，經常採取的策略是直接以關聯詞進行檢索，並以檢索結果決定其是否與主題相關。而策略四即是以此概念自動化進行，過程如下：

1. 將關聯詞送入檢索
2. 取檢索結果前 n 篇的 Title
3. 比對 Title 是否有出現任何查詢詞
4. 有則將該關聯詞作爲擴充詞彙

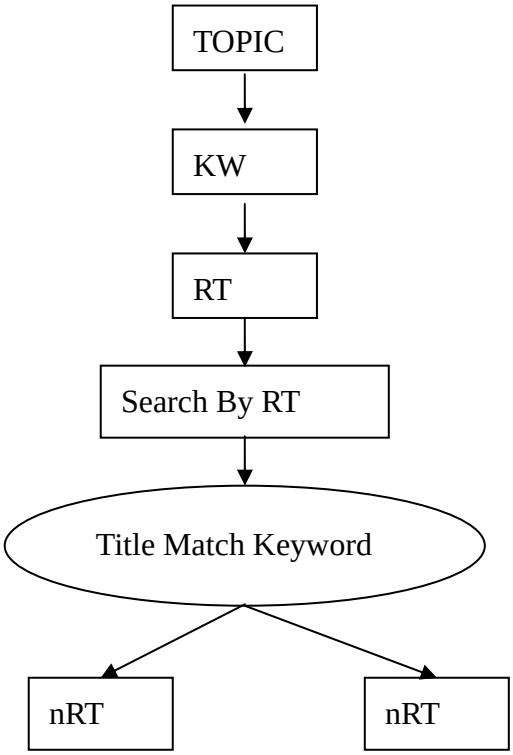


圖 3-8：策略四流程圖

策略四最不同於局域擴展在於，局域擴展使用整個主題之檢索結果進行擴展；而策略四則是以個別關聯詞之檢索結果作爲判斷相關之依據，關聯詞之檢索結果中，如果越多關於主題的，則可視爲與主題具有相關，且對檢索結果較有幫助。

五、檢驗自動過濾規則

檢驗方式則分別採用相關評估法，即回收率(recall ratio)、精確率(precision)及 F-VALUE 三測量值，這三個測量值是評估假設規則是否有效。檢驗將以人工篩選之關聯詞爲答案，比對以上四種假設何者最能找出人工判定爲相關的關聯

詞。

透過以上的驗證方式相較於 TREC_EVAL 可以比較快得到檢驗結果。因爲所有資料均記載於資料庫中，只需要以 SQL 語法即可得到所需要的所有數據，但必須保證人工篩選出之詞彙送入 TREC_EVAL 後可以有良好之成績；而使用 TREC_EVAL 必須要先進行漫長之檢索實驗，將實驗結果送入 TREC_EVAL，耗時往往需要數個鐘頭，相較於 SQL 語法只需要數分鐘的方式，本實驗將以 SQL 方式進行大量參數測試。

由初步規則之實驗結果歸納部分規則之特性，其中表示規則一的回收率最高，但相對的精確率爲最低；而規則二在提升精確率時造成回收率與規則一相比足足少了一半；規則三則符合預期的結果，精確率最高。以 F-Value 判斷，R7 的成效預期應該最好。

表 3-8：篩選方式之相關判斷評估結果

規則	TP	FN	FP	TN	P	R	F
R1	657	384	3375	3728	0.163	**0.631**	0.259
R2	331	710	1615	5488	0.170	0.318	0.221
R3	174	867	379	6724	**0.315**	0.167	0.218
R7(T=1500,N=1.5)	499	542	1644	5459	0.232	0.479	**0.313**

六、以各擴展模式與檢索模式進行驗證

本實驗將以查詢擴展與檢索模式之組合進行研究。透過各種組合搭配以觀察 Global Feedback 之規則是否可以穩定的運作。

查詢擴展：

- 不使用查詢擴展
- Local Feedback（L）
- Global Feedback（G）
- 先使用（G）再進行（L）

檢索模式：

- ByteSize Normalization
- Pivoted Normalization Method
- BM11
- BM25
- BM25m
- Language Model（Dirichlet prior retrieval method）

實驗將進行 6x4 共 24 次檢索實驗，並將檢索結果送入 TREC_EVAL 進行評估。並根據每個查詢主題之成績與全體之成績進行觀察。

表 3-9：ByteSize 與 BM11 於不同擴展方式之檢索結果

擴展模式	ByteSize	BM11
不使用查詢擴展	0.2360	0.2339
Local Feedback	0.2796	0.3046
Global Feedback（人工）	0.2890	0.2845
Global Feedback（Tseng，2004）	0.2432	X
G -> L	0.3267	0.3204

第四章 實驗數據分析

本研究將於此章節分析各關聯詞篩選策略之結果與數據探討，利用 TREC_EVAL 與相關判斷評估各篩選策略之成效。並於數據收集完成進行分析探討後開始撰寫論文，本章共分 3 節，第一節描述各關聯詞篩選策略之評估；第二節描述各篩選策略搭配不同檢索模式之成效;第三節描述局域擴展與全域擴展搭配不同檢索模式之成效。

第一節 關聯詞篩選策略之評估

本節將分析各篩選策略之成效以 ByteSize 檢索模式作爲測試。由第三章已知各策略之運用方式，本節將實際運用並評估其結果，並實驗各參數以取得各策略之最佳結果與各策略之比較分析。

一、策略一 (S1)：以關聯詞與關鍵詞之詞頻爲基礎之過濾規則

利用詞頻之門檻篩選關聯詞，利用關鍵詞、關聯詞之詞頻決定是否選擇該關聯詞作爲擴充詞彙。

規則一 (R1)：df(r) < df(k)，關聯詞詞頻要低於關鍵詞詞頻。

規則二 (R2)：R(r1) = R(r2)，關聯詞要是同一主題內兩關鍵詞之關聯詞。

規則三 (R3)：R1 & R2，規則一與規則二之交集。

規則四 (R4)：IF df(k)<T:R1 ELSE:R2，若關鍵詞低於某門檻，使用規則一方式過濾；否則即使用規則二方式過濾。

規則五 (R5)：IF df(k)<T:R1’ ELSE:R2，若關鍵詞低於某門檻，則使用類似規則

一（R1’）之方式過濾；否則使用規則二方式過濾。

R1’：df(r)<N*df(k);1<N<2，關聯詞詞頻要低於 N 倍的關鍵詞詞頻。

規則六 (R6)：IF df(k)<T:R1 ELSE:R3，若關鍵詞低於某門檻，使用規則一方式過濾；否則即使用規則三方式過濾。

規則七 (R7)：IF df(k)<T:R1’ ELSE:R3，若關鍵詞低於某門檻，則使用類似規則一（R1’）之方式過濾；否則使用規則三方式過濾。

關鍵詞門檻（T）以人工之經驗取 10、50、100、200、300、400、500、1000、1500、2000、2500、3000、3500、4000、5000、10000、20000 做爲測試門檻。而 R1’之關鍵詞詞頻擴展倍數（N）取 1.2、1.4、1.6、1.8、2 作爲測試。

策略一搭配不同門檻共有 207 種組合，以程式自動篩選擴充詞彙進行檢索並計算成效共約花費 2 天時間完成所有數據之搜集。表 4-1 紀錄各規則之最佳檢索成效，其紀錄各規則之數據乃根據檢索成效 TREC_EVAL 中平均精確率之數值，取平均精確率最高者爲該規則之成效，及其相對應之相關判斷（F-Value）。由表得知規則五爲最佳過濾方式，即“若關鍵詞詞頻低於 50，則關聯詞詞頻要低於 1.4 倍的關鍵詞詞頻；否則使用規則二方式過濾”。

根據規則四到規則七除規則六其關鍵詞詞頻門檻皆以 50 爲最佳門檻，而規則六若以 50 爲關鍵詞詞頻門檻其平均精確率爲 0.2503 與其他規則差距並不顯著，由如圖 4-1 顯示關鍵詞詞頻門檻明顯以 50 爲最佳，而門檻越高則平均精確率越低，且門檻在 10000 以後平均精確率呈現穩定較差的成績。而關鍵詞詞頻擴展倍數以 1.4 最佳，然各倍數之平均精確率差距不顯著；而在相關判斷以招回率差異較大，擴展倍數越大其招回率也越高，如表 4-2。

表 4-1：策略一各規則之相關判斷與 TREC_EVAL 之比較

規則	相關判斷	TREC_EVAL
R1：df(r) < df(k)	0.259	0.2064
R2：R(r1) = R(r2)	0.222	0.2498
R3：R1 and R2	0.218	0.2491
R4(50)：IF df(k)<50:R1 ELSE:R2	0.226	0.2507
R5(50,1.4) ：IF df(k)<50:R1’ ELSE:R2 R1’ : df(r)<1.4*df(k)	0.226	0.2518
R6(200)：IF df(k)<200:R1 ELSE:R3	0.251	0.2508
R7(50,1.4)：IF df(k)<50:R1’ ELSE:R3 R1’ : df(r)<1.4*df(k)	0.233	0.2512

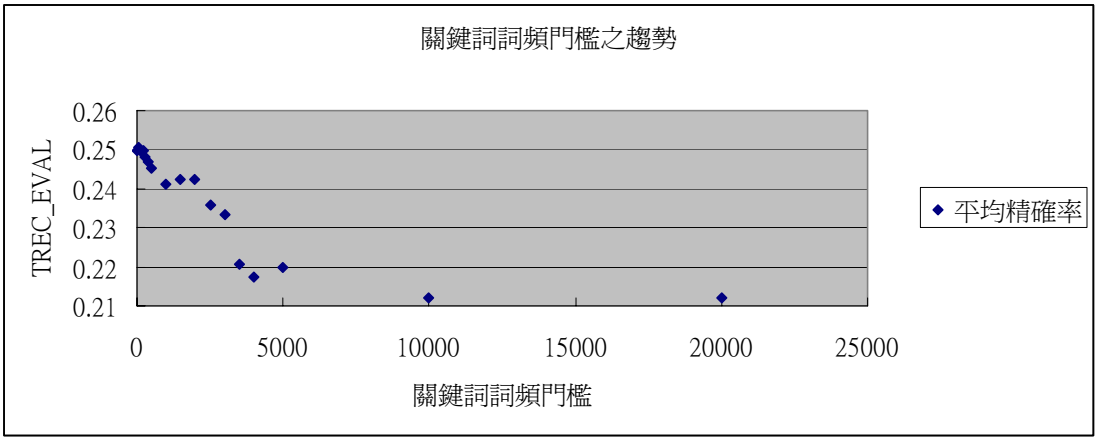


圖 4-1：關鍵詞詞頻門檻之趨勢

表 4-2：關鍵詞詞頻擴展倍數（N）於相關判斷與 TREC_EVAL 之成效

N 倍	Presison	Recall	F-Value	TREC_EVAL
1.2	0.1746	0.4625	0.25	0.2379
1.4	0.176	0.4778	0.2532	0.2372
1.6	0.1747	0.4841	0.2528	0.2364
1.8	0.174	0.49	0.2525	0.2363
2	0.174	0.497	0.2523	0.2351

同時根據策略一之數據顯示，相關判斷與 TREC_EVAL 之數據不具顯著相關性。

由統計分析軟體 SPSS 整理出表 4-3 之相關分析，其顯示精確率與 TREC_EVAL 爲正相關，招回率與 TREC_EVAL 爲負相關，F-Value 與 TREC_EVAL 爲負相關。唯此相關分析因 TREC_EVAL 與 F-Value 之數據差異不大，故對此結論採保守之態度。

表 4-3：F-Value 與 TREC_EVAL 相關係數表

相關

		P	R	F	TREC
P	Pearson 相關	1.000	-.725**	.129	.489**
	顯著性 (雙尾)	.	.000	.063	.000
	個數	207	207	207	207
R	Pearson 相關	-.725**	1.000	.394**	-.896**
	顯著性 (雙尾)	.000	.	.000	.000
	個數	207	207	207	207
F	Pearson 相關	.129	.394**	1.000	-.375**
	顯著性 (雙尾)	.063	.000	.	.000
	個數	207	207	207	207
TREC	Pearson 相關	.489**	-.896**	-.375**	1.000
	顯著性 (雙尾)	.000	.000	.000	.
	個數	207	207	207	207

**. 在顯著水準爲0.01時 (雙尾)，相關顯著。

二、策略二 (S2)：關聯詞遞迴擴展

策略二以策略一中規則二之概念做遞迴式的擴展，並配合關聯次數與詞頻做爲過濾門檻。策略二簡而言之即以 R2：“關聯詞要是同一主題內兩關鍵詞之關聯詞”之擴展，將符合 R2 之關聯詞視爲關鍵詞後再進行一次 R2 之規則，企圖使關鍵詞增多後關聯詞也隨之增多。

由初步概念之實驗篩選過程中發現，遞迴擴展之使用將許多好的關聯詞視爲關鍵詞進行二度擴展，其中由於關鍵詞數量太多而發現部份關聯詞同時被許多關鍵詞所關聯之詞彙不一定爲好的查詢詞彙，甚至大多數爲廣義詞，容易造成主題漂移之現象。故以人工之經驗做關聯次數之門檻以篩選出好的詞彙。以查詢主題「一九九八年諾貝爾物理學獎」爲例，「活動」一詞會被「台灣, 諾貝爾物理獎得主, 學者, 得主, 世界, 法國, 教授, 吳健雄科學營, 華人」所關聯，且關聯次數高達 9 個，而「活動」一詞送入檢索可能會產生主題漂移，由表 4-4 可見部份關聯詞及其關鍵詞。故根據人工觀察數個查詢主題後以關聯次數門檻爲 4 時可選出較多好的詞彙，而門檻爲 4 以上詞彙如：「物理」、「人類」、「學生」、「活動」將視爲不好的擴充詞彙。

表 4-4：關聯詞所屬關鍵詞及其關聯次數

關聯詞	關鍵詞	關聯次數
知識	理論, 科學	2
朱棣文	崔琦, 諾貝爾物理獎得主	2
物理	崔琦, 諾貝爾物理獎得主, 丁肇中, 化學, 量子, 因子, 亞特蘭大, 科學, 吳健雄科學營	9
量子霍爾	崔琦, 量子	2

效應		
固態	崔琦，量子	2
人類	華裔科學家，世界，理論，因子，科學	5
經濟	華人，理論，學者	3
學生	學者，化學，理論，科學，教授，吳健雄科學營，華人	7
活動	台灣，諾貝爾物理獎得主，學者，得主，世界，法國，教授，吳健雄科學營，華人	9
醫學	因子，科學	2
諾貝爾物理獎	崔琦，得主	2
日本	法國，教授，得主，學者	4
作品	法國，教授，得主，科學	4
空間	科學，理論	2
中國	法國，華人，學者	3
華裔	華人，崔琦，丁肇中	3

策略二除了以關聯次數爲門檻外加以關聯詞詞頻作爲篩選之依據，根據觀察如果關鍵詞詞頻皆高於門檻，此關聯詞大多爲不好的關聯詞；反之，若所有關鍵詞中至少有一詞詞頻低於門檻，關聯詞則不至於太差。關聯詞的門檻則參考人工判讀爲相關的關聯詞之平均值而來，根據統計結果顯示相關的關聯詞其平均詞頻約 3000，標準差約 8000，故門檻取平均值正負 0.5 倍標準差作爲門檻，因此門檻值約 3000 與 7000。同樣以查詢主題「一九九八年諾貝爾物理學獎」爲例，例如「經濟」會被「華人(4245)」、「理論(4439)」、「學者(12269)」等關鍵詞所關聯，但因其詞頻皆高於 3000，故在此將被視爲不好的擴充詞彙。由表 4-5 可見被門檻

過濾出不好的擴充詞彙如：「知識」、「經濟」、「空間」、「中國」等。

表 4-5：關聯詞所屬關鍵詞與關鍵詞詞頻門檻之紀錄

關聯詞	關鍵詞(詞頻)	詞頻低於 3000
知識	理論(10812), 科學(4439)	無
量子霍爾效應	崔琦(47), 量子(217)	
固態	崔琦(47), 量子(217)	
經濟	華人(4245), 理論(4439), 學者(12269)	無
醫學	因子(1063), 科學(10812)	
諾貝爾物理獎	崔琦(47), 得主(2699)	
日本	法國(10301), 教授(13328), 得主(2699), 學者(12269)	
作品	法國 10301, 教授(13328), 得主(2699), 科學(10812)	
空間	科學(10812), 理論(4439)	無
中國	法國(10301), 華人(4245), 學者(12269)	無
華裔	華人(4245), 崔琦(47), 丁肇中(31)	

表 4-6 顯示使用策略二遞迴擴展關聯詞同時使用關聯次數與詞頻限制有最佳之成效爲 0.2505，即關聯次數門檻爲 4 同時詞頻門檻爲 3000 有最佳檢索效果。故策略二篩選關聯詞時關聯詞被關鍵詞關聯次數必需低於 4 且此 4 關鍵詞之詞頻中至少要有一詞詞頻低於 3000。

表 4-6：策略二之相關判斷與 TREC_EVAL 之比較

策略二各種實驗參數	相關判斷	TREC_EVAL
遞迴擴展不以參數限制	0.237	0.2417
遞迴擴展使用關聯次數限制(4)	0.237	0.2475
遞迴擴展使用詞頻限制(3000)	0.251	0.2457
遞迴擴展使用詞頻限制(7000)	0.241	0.2418
遞迴擴展使用詞頻與關聯次數限制(4,3000)	**0.249**	**0.2505**
遞迴擴展使用詞頻與關聯次數限制(4,7000)	0.241	0.2474

三、策略三 (S3)：依據關聯詞與主題之強度

策略三以關聯詞與主題之強度作爲排序篩選的依據，即計算出關聯詞對整個查詢主題之強度，而非對個別關鍵詞之強度。並嘗試以多種公式計算強度取檢索結果最佳者。公式選用一般常見之對稱計算公式 Dice Cofficient（S1），並嘗試以 IDF 調整強度之計算（S2）；另根據部分文獻也嘗試使用非對稱方式計算強度（S3），同樣以 IDF 調整強度之計算（S4），並進行交互比較，結果記錄於表 4-7。

各種計算強度之公式如下：

- Dice Coefficient（Dice）：

$$S1_r = \sum_{k \in Q} \frac{2 \times df\ (r \cap k)}{df\ (r) + df\ (k)}$$

- 非對稱(nDice)取關聯詞爲基：

$$S2_r = \sum_{k \in Q} \frac{df(r \cap k)}{df(r)}$$

- Dice Coefficient（Dice）配合 IDF：

$$S3_r = \sum_{k \in Q} \frac{2 \times df(r \cap k)}{df(r) + df(k)} \times IDF$$

- 非對稱(nDice)取關聯詞爲基配合 IDF：

$$S4_r = \sum_{k \in Q} \frac{df(r \cap k)}{df(r)} \times IDF$$

$$IDF = \frac{\log(\frac{N}{df(r)})}{\log(N)}$$

當關聯詞之強度計算完成之後的詞彙選用以以下兩方式實驗：

- 每主題排序前 n 名之關聯詞爲擴充詞彙，n 取 5, 10 ,20, 30, 40 爲測試。
- 將所有主題之關聯詞做正規化使其強度在 0 與 1 之間，取強度門檻 t 篩選詞彙，門檻取 0.1, 0.2, 0.15, 0.27, 0.5 爲測試。

正規化之方式：

$$S_n = \frac{X - Min}{Max - Min}$$

X 爲關聯詞原先之強度，Sn 爲關聯詞正規化之強度

表 4-7 之強度公式實驗以每主題強度排序前 5 名的關聯詞作爲擴充詞彙進行

查詢擴展之檢索實驗，由其評估結果可得知最佳公式應爲不使用 IDF 調整之非對稱公式，平均精確率爲 0.2524。同時也可發現 Dice Cofficient 使用 IDF 調整時，不論相關判斷與平均精確率其成效皆優於不使用 IDF 之公式；然而非對稱之公式有著相反的結果，即有使用 IDF 調整的公式成效較差。故 IDF 對於對稱之計算公式有幫助，但對於非對稱之公式有負面影響。以人工觀察對稱與非對稱之排序結果可明顯發現對稱的方式所排序之詞彙中，廣義詞較多而產生主題漂移。例如查詢主題「一九九八年諾貝爾物理學獎」以對稱方式計算前 5 名的關聯詞爲：崔琦, 得主, 物理, 朱棣文, 諾貝爾獎；而以非對稱方式計算前 5 名的關聯詞爲韋潛光, 量子霍爾效應, 崔琦, 許洛夫, 吳健雄學術基金會。故對稱之方式易找出類似得主, 物理, 諾貝爾獎等造成主題漂移之詞彙而對檢索成效造成負面影響。

觀察詞彙的選用實驗記錄於表 4-8, 4-9，其顯示如果以每主題排序前 n 名之關聯詞爲擴充詞彙時，取前 5 名的關聯詞擴充，其成效最好；若選超過前 20 名以後則明顯變差。另如果以將所有主題之關聯詞做正規化使其強度在 0 與 1 之間，取強度門檻 t 篩選詞彙，當 t 爲 0.2 時成效最佳；門檻若低於 0.1 則導致強度太差的關聯詞被選取導致負面之檢索成效。

人工觀察兩選詞之方式發現每主題排序前 n 名之篩選方式不論關聯詞是否適當皆選入 n 個詞，其排序的競爭只與該主題之關聯詞，因此有些主題並無真正適當關聯詞者也會送入 n 個關聯詞，有些主題有大量適當之關聯詞卻只能送出排序較前幾個而已；而正規化之方式則將所有關聯詞一起排序競爭，因此容易發生有些主題一個關聯詞沒有，因其並無真正適當的關聯詞，而某些主題也能夠送入較多真正適合的關聯詞，進而使查詢擴展的詞彙品質一致。然正規化之方式其計算時間較久，因爲必須先取得所有查詢之關聯詞及其強度，因而需要大量運算時間，且成效相較於排序篩選前 n 名之方式並無顯著提升，故策略三以非對稱方式計算取排序前 5 名之方式運作於後續之實驗。

表 4-7：強度計算公式之相關判斷與 TREC_EVAL 之比較

TOP-5	相關判斷	TREC_EVAL
Dice	0.141	0.2185
Dice (IDF)	0.166	0.2386
nDice	**0.136**	**0.2524**
nDice(IDF)	0.128	0.2453

表 4-8：以篩選前 N 名爲排序方式之檢索成效

N	**5**	10	20	30	40
nDICE(排序)	**0.2524**	0.2375	0.2405	0.2284	0.2180

表 4-9：以正規化門檻值 T 爲排序方式之檢索成效

T	0.1	0.15	**0.2**	0.27	0.5
nDICE(正規化)	0.2457	0.2522	**0.2540**	**0.** 2526	0.2468

四、策略四 (S4)：依據關聯詞之檢索結果

策略四主要概念在於以關聯詞之檢索結果自動判斷其是否與主題相關。大致方式以關聯詞之檢索結果取前 n 篇之題名與查詢詞比較，如果比對到查詢詞則該關聯詞即視爲相關，n 取 5, 10, 15, 20 做爲測試。

根據表 4-10 可見最佳結果爲關聯詞之檢索結果取前 15 篇題名與關鍵詞比對成效最好。且此策略整體之平均精確率皆不高，且呈現負成長；但其相關判斷皆高於之前策略，主要因爲此策略之招回率較高，使相關判斷成績比其他策略來的高，而在精確率不高的情形下可能也導致 TREV_EVAL 的平均精確率低之原因。

表 4-10：策略四門檻 n 之相關判斷與 TREC_EVAL 之比較

N	相關判斷	TREC_EVAL
比對篇數:5	0.292	0.2090
比對篇數:10	0.314	0.2080
比對篇數:15	0.320	0.2134
比對篇數:20	0.317	0.2116

五、小結

總結以上四種關聯詞篩選策略之最佳結果：

策略一：以詞頻篩選關聯詞，若關鍵詞詞頻低於 50，則關聯詞詞頻要低於 1.4 倍的關鍵詞詞頻；否則使用規則二方式過濾。

規則二：關聯詞遞迴擴展時，關聯次數門檻爲 4，同時詞頻門檻爲 3000。

規則三：計算關聯詞之強度以非對稱方式計算並取排序前 5 名之關聯詞。

規則四：以關聯詞之檢索結果取前 15 篇題名與關鍵詞比對作爲篩選方式。

第二節 篩選策略之成效

本節將比較各擴展方式對主題檢索之影響。本節針對局域擴展、全域擴展使用人工選詞、全域擴展使用自動選詞之檢索結果進行比較，並紀錄數據於表 4-11。

一、局域擴展之結果

局域擴展使用 Blind Relevance Feedback 取檢索結果前 6 篇前 15 個關鍵詞。檢索結果可發現 42 個查詢主題中有 29 題成效獲得提升，而有 13 題成效下降低，此數據與文獻所提一致。即使最終之平均精確率有提升，然而仍有 13 個查詢主題成效不升反降，此應與 Blind Relevance Feedback 本身初次檢索結果未必與主題相關所導致，但以最終之檢索成效而言，Local Feedback 使用 Blind Relevance Feedback 仍是穩定有效的策略。

二、全域擴展使用人工篩選之結果

本次人工判斷有兩種篩選結果，人工一之篩選過程大致上以 Description 之資料與關聯詞進行判斷，即在並非真正了解需求而做出之初步判斷；而人工二則參考 NTCIR3 所提供正確相關判斷之文件進行參考，以篩選關聯詞。且本研究對人工二之篩選結果應接近真正資訊需求者所篩選之詞彙，因實際資訊需求者對於查詢需求、查詢目標有明確的認知，此認知即以 NTCIR3 所提供正確相關判斷之文件類似，故人工二以相關文件作爲關聯詞篩選之判斷應接近實際使用者之判斷。

觀察實驗數據顯示人工二之數據優於人工一，且人工判斷關聯詞之檢索結果爲最佳，但仍有部分查詢主題無法透過詞彙擴展獲得更好之成效，主要爲該主題之關聯詞與需求無關，也有可能是關聯詞多爲高頻詞，造成雜訊過多而導致成效降低。

三、全域擴展使用自動化篩選之結果

觀察所有最佳自動化策略之檢索結果，策略三檢索成效最佳，策略四相關判斷最高；而策略一至策略三對半數查詢主題有正面幫助，而策略四有三分之二查詢主題成效不升反降。觀察各策略對查詢主題之影響，策略四之平均成長率最高，即該策略對成效有提升之主題其程度優於其他策略。

四、小結

本次研究中，以查詢擴展之成效而言，對檢索最有幫助者爲全域擴展使用人工篩選詞彙，其次爲局域擴展使用 Blind Relevance Feedback，而自動化的全域擴展尚有進步之空間。自動化關聯詞擴展之方式以策略三成效最佳，策略一與策略二其次，而策略四成效最差。

表 4-11：各擴展方式與篩選策略之成效

			擴充詞數量	成長題數	負成長題數	平均成長率	負平均成長率	相關判斷	TREC_EVAL
不擴展		Basic							0.2355
局域擴展		Local	630	29	13	0.085	0.025	X	0.2876
全域擴展	自動策略	策略一	1944	22	20	0.096	0.008	0.226	0.2518
		策略二	1510	21	21	0.066	0.013	0.249	0.2505
		策略三	210	22	20	0.048	0.046	0.136	0.2524
		策略四	2233	18	24	0.131	0.022	0.320	0.2134
	人工選詞	人工一	1041	27	15	0.025	0.005	X	0.2929
		人工二	1219	33	9	0.025	0.005	X	0.3299

第三節 局域擴展與全域擴展搭配不同檢索模式

本節將針對不同擴展模式與不同檢索模式進行分析。實驗將觀察查詢擴展對不同檢索模式之變異情形，以找出最佳擴展模式與檢索模式之搭配。

一、檢索模式之比較

本實驗以向量模式、機率模式、語言模式進行比較，其中向量模式採用 ByteSize Normalization（BS）、Pivoted Normalization Method（PC），機率模式採用 BM11、BM25、BM25m，而語言模式使用 Dirichlet prior retrieval method（LM）。

表 4-12 紀錄各檢索模式之檢索結果，其中 LM 之檢索結果與其他差異甚大，且檢索結果成績幾乎是 0，可能所引用的公式有誤或程式錯誤造成，在此不列入分析（見表 4-14），故只對 BS、PC、BM11、BM25、BM25m 進行分析整理。

各檢索模式之實驗結果中，BM25 與 BM25m 最佳，其次爲 BM11、BS，最差爲 PC。此外，不論何種檢索模式其最大值與最小值差異甚大，因此不論何種檢索模式皆因不同查詢擴展方式影響其成效甚大，而觀察期變異系數可發現以 PC 最易受影響，其次爲 BM11，而 BS、BM25、BM25m 最穩定。

表 4-12：各檢索模式之檢索結果分析

檢索模式	平均數	最大值	最小值	變異系數
BS	0.2753	0.3431	0.2134	0.1513
PC	0.2212	0.284	0.1705	0.1804
BM11	0.2779	0.3562	0.2136	0.1655
BM25	0.2820	0.3543	0.2184	0.1564
BM25m	0.2820	0.3544	0.2183	0.1582

二、擴展模式之比較

實驗各種不同擴展方式配合不同各種檢索模式，其中發現全域擴展與局域擴展搭配皆優於單獨使用局域擴展或全域擴展。最佳檢索成效爲全域擴展以人工選詞搭配局域擴展其平均值爲 0.3382；同時不使用人工介入，以自動化篩選全域擴展之詞彙，其檢索成效以策略三搭配局域擴展成效較佳，其平均值爲 0.2959。而各擴展方式之成效不易受檢索模式不同而影響其成效，全域擴展中人工選詞之方式較自動選詞之方式穩定，而以局域擴展最易受不同檢索模式而影響成效（見表 4-13）。

表 4-13：各擴展方式之檢索結果分析

查詢擴展	平均數	最大值	最小值	變異系數
不擴展	0.2288	0.2422	0.1921	0.0914
局域擴展（Local）	0.2821	0.3080	0.2087	0.1480
策略一	0.2380	0.2514	0.1891	0.1150
策略二	0.2362	0.2523	0.1793	0.1349
策略三	0.2480	0.2571	0.2224	0.0587
策略四	0.2068	0.2184	0.1705	0.0989
人工一	0.2822	0.2932	0.2441	0.0759
人工二	0.3197	0.3329	0.2746	0.0791
策略三+Local	0.2959	0.3141	0.2467	0.0961
人工二+Local	0.3382	0.3562	0.2840	0.0912

三、小結

由查詢擴展與檢索模式之實驗分析中可明顯發現，整體而言查詢擴展對檢索之成效影響較大，檢索模式對檢索成效影響較小，此與文獻分析所得有一致之結論，即檢索結果常依賴於使用者所輸入之詞彙。

而詳細觀察其檢索成效之數據可知自動選詞之方式不論使用局域或全域擴展，易受不同檢索模式之影響，而人工選詞則較不受不同檢索模式之影響（見表 4-13 與表 4-14）。即較好的擴充詞彙不論使用何種模式都可得到較好的成效，如全域擴展使用人工選詞，而詞彙的品質越差則易受檢索模式之影響，如局域擴展使用 Blind Relevance Feedback 在檢索模式 BM25m 得到 0.3080 的高檢索成效，在 PC 只得到 0.2087。

而觀察全域擴展與局域擴展兩者之擴展成效，全域擴展使用第二次人工判斷之結果最好，而第一次人工判斷不及局域擴展，自動化全域擴展方式成效較差。第二次之判斷結果，主要是參考了 NTCIR3 所提供之查詢主題相關文件，與描述查詢主題之所有欄位，故判斷結果較精確可靠。而第一次的判斷結果只依據描述查詢主題中 Description 此一欄位之內容，便進行關聯詞的相關判斷，而此檢索成效較差主要是因爲無法直接以 Description 就知道檢索的真正需求，例如主題「中新一號衛星及評論」容易找出關聯詞是關於其他衛星或是所有與衛星相關的詞彙，同時這類詞彙也確實是有與描述相關，但其 Narrative 欄位，卻又補充只接受與中新一號這顆衛星相關的評論，因此初次選擇的關聯詞內容範圍超過主題所描述之需求，導致成效變差。而以第一次人工判斷之結果與局域擴展使用 Blind Relevance Feedback 之結果可推論，當使用者對某領域的知識尚不足或根本不懂的情形下，使用提示詞得方式給使用者挑選詞彙做查詢擴展，容易找出不相關的辭而導致主題檢索成效變差，例如第一次人工選詞。因此，如果以自動化局域方式補充或修飾其檢索詞彙成效應該會比較好。反之，使用者對該領域已經有基礎

之知識或是該領域專家，以提示辭的方式由其自行修飾查詢詞，成效應該比較理想，例如第二次人工選詞。

此外，Pivoted Normalization Method（PC）之成效與其他模式相比，成效較差，該模式所採用之公式與係數均未經過最佳化調整，乃直接採用 Hui Fang 於文獻中所提及之公式與係數，且在其文獻中被證明具有可適用性（Hui Fang，2004）。然在本實驗結果發現中文環境中卻有著相當大的差異，PC 的成效並不如其他模式，且不論是長查詢或短查詢其成效皆較差，故係數若經最佳化調整應可取得較一致之成效。

表 4-14：擴展策略與檢索模式於 NTCIR3 成效之比較

擴展方式	BS	BM11	BM25m	BM25	PC
Basic	0.2355	0.2329	0.2415	0.2422	0.1921
Local	0.2876	0.3028	0.3080	0.3035	0.2087
策略一	0.2512	0.2482	0.2502	0.2514	0.1891
策略二	0.2505	0.2482	0.2509	0.2523	0.1793
策略三	0.2524	0.2511	0.2571	0.2570	0.2224
策略四	0.2134	0.2136	0.2183	0.2184	0.1705
人工判斷一	0.2929	0.2877	0.2930	0.2932	0.2441
人工判斷二	0.3299	0.3286	0.3329	0.3327	0.2746
策略三+L	0.2960	0.3092	0.3133	0.3141	0.2467
人工判斷二+L	0.3431	0.3562	0.3544	0.3543	0.2840

第五章 關聯詞與主題分析

本研究將於此章節分析關聯詞與查詢主題對檢索成效之影響，以人工觀察個別主題之需求與其擴展之關聯詞對檢索成效的影響進行分析。本章共分 2 節，第一節描述各查詢主題特性之分析；第二節描述關聯詞之分析。

第一節 查詢主題特性分析

本研究主要目的雖在於查詢擴展詞彙之篩選，然在整體實驗之過程發覺查詢主題本身即具有某些因素，而這些因素深深影響著檢索之成效。實驗過程中可明顯發現部分查詢主題總是有著較高或較低的檢索成效，不論是運用何種查詢擴展方式或是何種檢索模式，即部份查詢主題本身之需求難易度對檢索成效有穩定且直接之影響，本節主要目的在於以人工之方式觀察並歸納部份查詢主題之因素及其對檢索成效之影響。

一、查詢主題觀察設計

以下爲觀察查詢主題之步驟：

1. 蒐集欲觀察之主題
2. 初步觀察主題並歸納設計主題屬性
3. 查詢主題屬性之定義與判斷
4. 查詢主題分析

各步驟之細節如下：

1.蒐集欲觀察之主題

本研究欲觀察造成查詢主題成效不佳之因素，故蒐集檢索成效較好的主題與

檢索成效較差的主題，以及查詢擴展成效較好與較差之主題，以了解查詢擴展對具有某些因素之主題較有幫助。

2. 初步觀察主題並歸納設計主題屬性

將上述步驟所蒐集之查詢主題進行初步之人工觀察，用以歸納比較各主題之屬性。人工觀察主題之欄位有 Title，Description，Concept，Narrative 等四個主要欄位。

3. 查詢主題屬性之定義與資料

將觀察後所彙整能夠代表主題之屬性，先給予合理之定義使各屬性有一致之範圍與意義，再以人工提供適當之資料，使各屬性得以進一步分析利用。

4. 查詢主題分析

將各屬性之資料加以分析，觀察各屬性對檢索成效之影響，並歸納出具體之影響因素。

二、查詢主題蒐集

利用各擴展實驗之結果，觀察各主題之檢索成效與擴展成效，找出主題成效之分佈於表 5-1，由表中可歸納出以下主題之特性：

- 最容易檢索：1, 6, 9, 10, 17, 23, 25
- 最不容易檢索：32, 45, 47, 50
- 擴展成效最好：12, 32, 35
- 擴展成效最差：21, 37, 38, 45
- 擴展導致成效更差：45
- 擴展導致成效變好：12, 13
- 即使擴展成效最好成績還是最差：32

故將此篩選之主題進一步以人工觀察並分析其影響成效之因素。

表 5-1：各策略中查詢主題檢索結果之分佈

	最容易檢索	最不容易檢索	擴展成效最佳	擴展成效最差
Basic	1, 3, 9, 10, 17	12, 13, 32, 47, 50	X	X
Local	6, 9, 10, 17, 23	18, 32, 45, 47, 50	12, 32, 35, 36, 43	2, 18, 25, 42, 45
策略一	6, 9, 14, 23, 25	7, 32, 45, 47, 50	11, 12, 13, 32, 47	7, 21, 36, 38, 40
策略二	1, 6, 9, 14, 25	7, 32, 45, 47, 50	5, 21, 36, 40, 50	12, 13, 32, 43, 47
策略三	6, 9, 10, 17, 25	7, 32, 45, 47, 50	12, 22, 32, 35, 43	36, 37, 38, 45, 47
策略四	1, 6, 10, 23, 25	21, 36, 45, 47, 50	12, 13, 32, 35, 43	9, 21, 36, 37, 45
人工一	1, 6, 10, 23, 25	18, 32, 45, 47, 50	2, 12, 13, 32, 35	5, 15, 21, 38, 45
人工二	1, 6, 9, 14, 17	32, 33, 45, 47, 50	12, 13, 32, 35, 43	33. 37, 38, 39, 45

三、查詢主題之屬性定義

初步觀察上述所篩選之主題，加以人工歸納假設，而定出以下主題之屬性：

討論範圍：觀察主題所討論之範圍，所涉及之內容範圍與面向。

需求限制：由 Narrative 欄位觀察需求之細節。

專有名詞：觀察查詢主題所使用的專有名詞，可能爲人名、地名、機構名等。

多意義詞彙：觀察容易有多種意思、多種涵義之詞彙，易對檢索結果造成負面影響之詞彙，例如：國際組織，該詞可能包含有多種國際組織之名稱。

需求文字長度：查詢主題之字數。

Concept 數目：紀錄 Concept 之數目。

以「國際合作解決環境問題」查詢主題爲例，屬性之資料見表 5-2。

表 5-2：查詢主題之屬性舉例以「國際合作解決環境問題」爲例

TopicID	45
Title	國際合作解決環境問題
Concept	共同合作，環境問題，環境議題，污染，國際組織
Description	查詢國際共同合作爲解決環境問題諸如空氣、水、土壤與自然的污染之相關報導
Narrative	相關內容應說明爲了解決環境問題國際間彼此的相互合作，舉凡空氣污染、水污染、土壤污染、生態破壞等等皆在範圍之內。文中至少要提及一個國家及其所面對之無法獨力解決的具體環境議題。對於沒有明確指出國家名稱的報導則視爲部分相關。
討論範圍	廣

需求限制	至少要提及一個國家及其所面對之無法獨力解決的具體環境議題
專有名詞	無
多意義詞彙	國際組織、國家、 污染、 環境問題、 具體環境議題
需求文字長度	176
Concept 數目	5

四、主題之觀察結果與分析

- 觀察最容易檢索的查詢主題，可隱約發現其有共同之特性：

只要與目標有關的都是相關。此類題型著重在目標或主題的所有相關討論，通常這類型查詢主題之目標明確、不易混淆、也無嚴格之需求限制，因此可輕易檢索出較高的成效。例如查詢主題「漢代文物大展」只要與目標有關的都是相關，舉例如下。其中需求只要是關於故宮博物院所展覽關於"漢代文物大展"者皆視爲相關，討論的目標明確，且不容易導致不相關，因此此查詢主題不論以何種檢索模式或擴展策略，其皆有不錯之成效。（範例如表 5-3）

表 5-3：查詢主題「漢代文物大展」各屬性之分析

TopicID	1
Title	漢代文物大展
Concept	漢代，文物大展，故宮博物院，歷史
Description	查詢故宮博物院所舉辦之千禧漢代文物大展相關內容
Narrative	台灣的故宮博物院是著名的典藏中國寶物的博物館，有關漢朝的典藏品展現了西元前 206 年到西元 220 年，中國漢朝的

	強盛與偉大。對於故宮博物院所舉辦之千禧漢代文物大展之說明，例如展出的文物種類、對於展出文物之介紹、展出時間、故宮的籌畫過程、合作單位等，以及展出後的成果與民眾的反應視爲相關。非本次展覽內容之漢代文物介紹，以及其他展覽活動之介紹視爲不相關。
討論範圍	廣
需求限制	無
專有名詞	故宮博物院
多意義詞彙	歷史、千禧、反應
需求文字長度	218
Concept 數目	4

此外就是需求的描述沒有不相關範圍之說明。一般查詢主題會告知哪些範圍是相關，哪些是部分相關，哪些是不相關，而通常沒有不相關部份之陳述者，其檢索之成效也易取得較好之結果。

需求的描述沒有不相關範圍之說明舉例如下，其中 Narrative 的描述中無不相關之部分，且討論的範圍廣泛，因此只要幾乎有討論到都可被視爲部分相關，故此題得檢索結果回收率相當高，排序的結果也不錯。（範例如表 5-4）

表 5-4：查詢主題「金大中總統對亞洲的政策」各屬性之分析

TopicID	23
Title	金大中總統對亞洲的政策
Concept	金大中，總統，國家政府，亞洲政策，經濟危機，經濟改革，中國，台灣，日本

Description	有關金大中總統對亞洲的政策之文章
Narrative	金大中先生在 1997 年年底的總統大選中被選出並在 1998 年 2 月 15 日擔任南韓總統，而金大中政府正式就任。而金總統視新政府爲「國家政府」，他上訴將焦點集中在解決嚴重經濟危機的人們。其開始結構性的調整，如金權統治改革。若一文章描述金大中總統對亞洲的政策，則其爲相關。若一文章並未描述他對亞洲的政策，而是其對外國的姿態或計畫，則爲部分相關。
討論範圍	廣
需求限制	無
專有名詞	金大中、國家政府政策、金權統治改革
多意義詞彙	亞洲、 中國、 台灣、 日本、總統、人們
需求文字長度	230
Concept 數目	9

- 觀察最不容易檢索之題型，明顯與最容易檢索有難度上的區別，其中可簡單歸納出以下特性：

查詢主題有特殊需求之限制，即除了本身有描述相關與不相關外，用以界定相關的範圍過於嚴格，而導致查詢結果成效較低。例如查詢主題「國際合作解決環境問題」，其需求描述了相關的範圍與程度，但描述中有提到需求限制如"… 至少要提及一個國家及其所面對之無法獨力解決的具體環境議題…"，即相關的文獻還必須要滿足此限制，才是使用者真正之需求，此一類型之主題通常檢索之成效較差。（範例如表 5-5）

表 5-5：查詢主題「國際合作解決環境問題」各屬性之分析

TopicID	45
Title	國際合作解決環境問題
Concept	共同合作，環境問題，環境議題，污染，國際組織
Description	查詢國際共同合作爲解決環境問題諸如空氣、水、土壤與自然的污染之相關報導
Narrative	相關內容應說明爲了解決環境問題國際間彼此的相互合作，舉凡空氣污染、水污染、土壤污染、生態破壞等等皆在範圍之內。文中至少要提及一個國家及其所面對之無法獨力解決的具體環境議題。對於沒有明確指出國家名稱的報導則視爲部分相關。
討論範圍	廣
需求限制	至少要提及一個國家及其所面對之無法獨力解決的具體環境議題
專有名詞	
多意義詞彙	環境問題、污染、國際組織
需求文字長度	176
Concept 數目	5

多意義詞彙與主題混淆，即查詢主題本身之需求描述使用了過多的”多意義之詞彙”，造成主題之目標混淆，增加檢索難度。例如查詢主題「日韓貿易」，其需求描述相關與不相關使用了許多多意義之詞彙如：產品名稱、交易量、特色、相關背景等。其中”產品名稱”指的是相關產品之”名稱”，而名稱爲一代名詞；”交易量”則爲以文字表達對數字之需求；”特色”與”背景”則是模糊與不明確之詞

彙，故無法單以文字上之共現因素就能處理得宜，導致查詢擴展不易進行。（範例如表 5-6）

表 5-6：查詢主題「日韓貿易」各屬性之分析

TopicID	47
Title	日韓貿易
Concept	貿易，進口，出口，日本，韓國，衝突，問題，貿易協商
Description	查詢說明或預測日韓貿易形態的相關報導
Narrative	相關內容應描述日韓之間的國際貿易形態，包括產品名稱、交易量、特色及相關背景。對於只提到貿易前景及評論，而沒有確切產品名稱和交易量的報導則視爲不相關。
討論範圍	廣
需求限制	沒有確切產品名稱和交易量的報導則視爲不相關
專有名詞	
多意義詞彙	產品名稱、交易量、特色、相關背景
需求文字長度	122
Concept 數目	8

- 影響擴展成效好壞之因素：

1. 關鍵詞與主題之關聯

當關鍵詞本身與查詢主題之關聯較強，或關鍵詞能滿足查詢之需求，由關鍵詞所擴展之關聯詞對主題檢索成效較有幫助；反之，當關鍵詞大多與主題較無關聯時，關聯詞對檢索之幫助也較小，甚至容易造成負面影響。

因此做了關鍵詞檢索成效之實驗，查詢擴展只擴充關鍵詞，即檢索之詞彙爲 2-gram 與關鍵詞加權，對照 2-gram 與關鍵詞不加權，以測試關鍵詞是否對檢索有幫助，如果關鍵詞對檢索成效有正面幫助，則擴展後之關聯詞也應對檢索有幫助。實驗結果使用了關鍵詞加權之檢索成效成長了 0.002，即幫助不顯著。觀察結果發現關鍵詞對主題之相關性非絕對，即有些關鍵詞對檢索成效有幫助，有些則導致負面影響。

以查詢主題「控訴戰爭罪惡」而言，其關鍵詞有”世界”、”日本”、”訴訟”、”世界大戰”，其中僅”世界”一詞較無關，而大多數關聯詞皆較有關，對檢索成效有正面幫助。而查詢主題「人體複製禁令」，其關鍵詞爲”人體”、”禁令”、”政府”，其需求爲複製之法令，而關鍵詞中並無”複製”之關鍵詞彙，因此找出之關聯詞僅少數與複製有關，且關聯度低，對檢索之成效造成負面影響。（範例如表 5-7、表 5-8）

2. 關聯詞本身

如果所擴展之關聯詞較多屬於高頻詞，或是多意義之詞彙，則容易造成主題漂移；若擴展詞彙中關聯詞多爲低頻詞，或是單一意義之詞彙，對檢索結果幫助最大。

以查詢主題「控訴戰爭罪惡」而言，其關聯詞有”小額訴訟”、”慰安婦”、”二次大戰”、”受災戶”、”官司”、”律師”、”法官”、”法院”、”日本”、”日本神戶”、”戰爭”、”中國”，其中”小額訴訟”、”慰安婦”等皆與需求有直接關係、除”日本”、”受災戶”、”日本神戶”、”中國”較無關外，其餘皆與主題有低度相關。而查詢主題「人體複製禁令」，其關聯詞爲”胚胎”、”基因”、”器官”、”細胞”、”疾病”、”醫學”、”美國”、”問題”、”英國”、”美國”、”政府”，其中無與複製禁令直接相關之詞彙，而”胚胎”、”基因”、”

器官”、”細胞”等詞彙僅低度相關，剩餘詞彙皆爲多意義之詞彙，易造成主題漂移使檢索成效有負面影響。（範例如表 5-7、表 5-8）

表 5-7：查詢主題「控訴戰爭罪惡」各屬性之分析與其擴充詞彙

TopicID	35
Title	控訴戰爭罪惡
Concept	二次世界大戰，民事訴訟，日本戰罪，判決
Description	查詢在日本的二次世界大戰戰罪訴訟
Narrative	相關內容應描述有關日本在二次大戰期間戰罪的民事訴訟案件，包括新的訴訟案件、過程、判決結果、對於判決的輿論。戰罪法庭的資訊則視爲不相關。
討論範圍	較小
需求限制	
專有名詞	二次大戰
多意義詞彙	日本、輿論
需求文字長度	108
Concept 數目	4
關鍵詞	世界、日本、訴訟、世界大戰
擴充之詞彙	小額訴訟、慰安婦、二次大戰、受災戶、官司、律師、法官、法院、日本、日本神戶、戰爭、中國

表 5-8：查詢主題「人體複製禁令」各屬性之分析與其擴充詞彙

TopicID	37

Title	人體複製禁令
Concept	人體複製，官方政策，法律，人體複製禁令
Description	查詢政府或國際上對於人體複製禁令所做的努力
Narrative	相關內容應描述政府或國際上對於人體複製禁令所做的努力，包括限制或禁止人體複製實驗的官方政策、法案或提案。對於人體複製倫理的意見及非人體的動物複製則視爲不相關。
討論範圍	較小
需求限制	非人體的動物複製視爲不相關。
專有名詞	
多意義詞彙	人體複製、官方政策
需求文字長度	125
Concept 數目	4
關鍵詞	人體、禁令、政府
擴充之詞彙	胚胎、基因、器官、細胞、疾病、醫學、美國、問題、英國、美國、政府

- 查詢擴展後使檢索成效便的更差

成效變差之情形主要有兩種，一爲檢索成效排序變差，另一爲回收率變差。依據研究者之觀察，如果關聯詞與主題低度相關，易使排序變差，造成類似主題漂移之情形，影響檢索之成效。而另一種則是回收率差，通常是因爲關鍵詞本身與查詢需求關聯較低，故關聯詞多屬低度相關或是無關的詞彙，因而導致查詢結果無法滿足需求，只找到少數相關文件。例如查詢主題「國際合作解決環境問題」，關鍵詞爲”土壤” 、” 環境問題”、” 環境”、” 空氣”，而其需求主要爲”解

決方法”，故關鍵詞本身與主題之關聯度低，因而關聯詞多與主題無直接相關，導致成效變差。（範例如表 5-9）

表 5-9：查詢主題「國際合作解決環境問題」各屬性之分析與其擴充詞彙

TopicID	45
Title	國際合作解決環境問題
Concept	共同合作，環境問題，環境議題，污染，國際組織
Description	查詢國際共同合作爲解決環境問題諸如空氣、水、土壤與自然的污染之相關報導
Narrative	相關內容應說明爲了解決環境問題國際間彼此的相互合作，舉凡空氣污染、水污染、土壤污染、生態破壞等等皆在範圍之內。文中至少要提及一個國家及其所面對之無法獨力解決的具體環境議題。對於沒有明確指出國家名稱的報導則視爲部分相關。
討論範圍	廣
需求限制	至少要提及一個國家及其所面對之無法獨力解決的具體環境議題
專有名詞	
多意義詞彙	環境問題、污染、國際組織
需求文字長度	176
Concept 數目	5
關鍵詞	土壤、環境問題、環境、空氣
關聯詞	有機氯農藥、煉銅廠、亞太開發公司、裕台化工、土壤液化問題、農藥汙染、環檢、RCA 桃園廠、除草劑、爐石、地下水汙染、化學肥料、農藥、環保署、空氣、汙染、環保局、

	能源使用效率、台灣環境、垃圾分類、產業界、空氣、生態、空調系統、汙濁、空氣品質、土壤、氣溫、汙染、環保局、工廠、環保

- 查詢擴展使成效變好

查詢擴展使成效變好通常是關聯詞符合需求，即關聯詞不僅要與 Description 更要能擴展出 Narrative 之描述爲最佳。以查詢主題「曼谷亞運」爲例，其關鍵詞有”曼谷”、”亞洲運動會”、”亞洲”、”運動”、”運動會”，而擴展出比賽項目（”撞球”、”登山車”）、比賽結果（”獎牌”、”金牌”、”銀牌”），等符合需求的關聯詞，因此直接對檢索成效有非常大的提升。

表 5-10：查詢主題「曼谷亞運」各屬性之分析與其擴充詞彙

TopicID	12
Title	曼谷亞運
Concept	亞洲運動會，曼谷，台北，台灣，選手
Description	查詢曼谷亞洲運動會的相關報導
Narrative	亞洲運動會是亞洲最大的運動賽會，第十三屆亞運在曼谷舉辦。曼谷亞洲運動會的時間、地點，中華代表隊在各項競賽中的表現與所獲的獎牌數，中華隊得獎選手的簡介，以及各國的比賽成績等視爲相關。其他國家選手的介紹視爲不相關。
討論範圍	廣
需求限制	
專有名詞	曼谷、台北、台灣

多意義詞彙	亞洲運動會、選手
需求文字長度	140
Concept 數目	5
關鍵詞	曼谷、亞洲運動會、亞洲、運動、運動會
關聯詞	登山車、中華民國國旗、中華亞運代表團、總領隊、中華奧會主席黃大洲、亞運會、熱身賽、撞球、獎牌、銀牌、代表隊、曼谷亞運、代表團、中華隊、金牌、亞運、泰國、選手、運動員、體委會、亞運、選手、中華代表團、獎牌

- 即使擴展成效很好成績還是最差

此類型之主題，明顯是關聯詞與需求不符而導致檢索成效提升不足。以查詢主題「一九九八中國洪難救助」爲例，其需求爲人道救助及撫恤之相關文件，然而其關鍵詞”洪水”、”中國”、”大洪水”、”國大”中，僅”國大”較無關聯，其餘爲低度相關，故關聯詞也難有更多關於人道救助及撫恤之詞彙，僅能選擇更多關於災情之報導，希望能增加回收率。

表 5-11：查詢主題「一九九八中國洪難救助」各屬性之分析與其擴充詞彙

TopicID	32
Title	一九九八中國洪難救助
Concept	人道救援，中國九八洪難
Description	查詢一九九八年中國大洪水期間相關的人道救助及撫恤
Narrative	相關內容應包括在一九九八年中國大洪水期間的人道救助及撫恤工作。對於只描述水患災情、其他水災及防洪措施的文件均視爲不相關。

討論範圍	廣
需求限制	
專有名詞	
多意義詞彙	洪難、救助
需求文字長度	105
Concept 數目	2
關鍵詞	洪水、中國、大洪水、國大
關聯詞	哈爾濱市、九江市、受災人口、吉林省、決口、溫家寶、黑龍江省、長江大堤、洞庭湖、嫩江、松花江、長江水位、險情、長江流域、大洪水、哈爾濱、大堤、湖北省、洪災、支流、洪峰、大慶、暴雨、武漢、湖北、防線、華中、湖南、防洪、水災、雨水、重慶、堤防、農田、水位、豪雨、長江、災情、災民、受災人口、堤壩、長江流域、大堤、支流、河流、死亡人數、防洪、洪水、長江

第二節 關聯詞之分析

一、關聯詞於資料集與答案集之分布

觀察由 NTCIR3 所提供之答案集，可發現選擇答案集中關聯詞出現機率高於資料集之機率者，其成效較好。因此如果找到的關聯詞其於答案集之詞頻較高但在資料集之詞頻較低，則明顯爲該主題之特徵詞。以下將分析各策略之篩選結果中關聯詞在資料集與答案集之分布情形。

以下以關聯詞之包涵率來觀察關聯詞在資料集與答案集之分布情形。包涵率之計算爲：

$$\text{包涵率} = \frac{\text{關聯詞在答案集之文件數}}{\text{關聯詞在資料集之文件數}}$$

各包涵率之分布於圖 5-1、圖 5-2、圖 5-3、圖 5-4、圖 5-5。

由圖與包涵率之比例假設，當包涵率大於 0.1 爲高度相關，當包涵率介於 0.1 與 0 之間爲低度相關，當包涵率等於 0 爲不相關，並根據包涵率之計算結果彙整於表 5-12。

策略一與策略二高度相關的關聯詞雖少，但不相關的關聯詞較策略四與人工二少，即雖然高度相關的詞不很多，不相關的詞也不很多。策略三不相關的詞彙最少，整體而言高度相關的詞彙比例極高，但總詞彙數較低，高度相關的詞彙僅人工二或策略三之半數而已，故詞彙擴展無法使檢索結果有效提升。策略四高度相關的詞最多，但明顯發現不相關的詞彙非常多，不相關的詞佔了將近二分之一，故成效最差。而人工二之結果顯示高度相關的詞彙僅比策略四少一個，可算是非常好的結果，且不相關與部分相關的詞比例較策略一、策略二、策略三少，故成效最好。

表 5-12：各策略包涵率之分析

	高度相關	低度相關	不相關
策略一	34	676	114
策略二	50	1050	371
策略三	70	47	57
策略四	163	988	972
人工二	162	329	480
全部詞彙	214	4252	3678

註：包涵率大於 0.1 爲高度相關，包涵率介於 0.1 與 0 之間爲低度相關，包涵率等於 0 爲不相關。

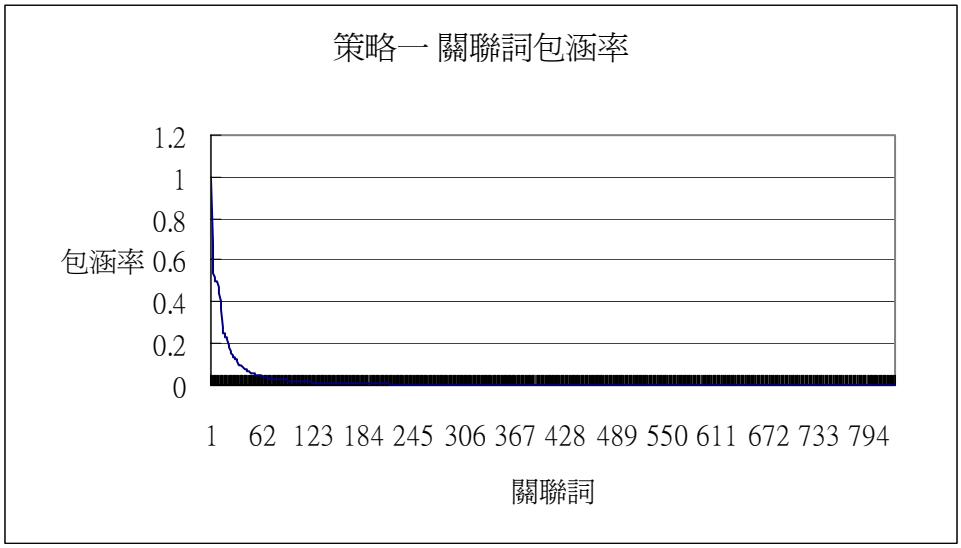


圖 5-1：策略一之關聯詞包涵率

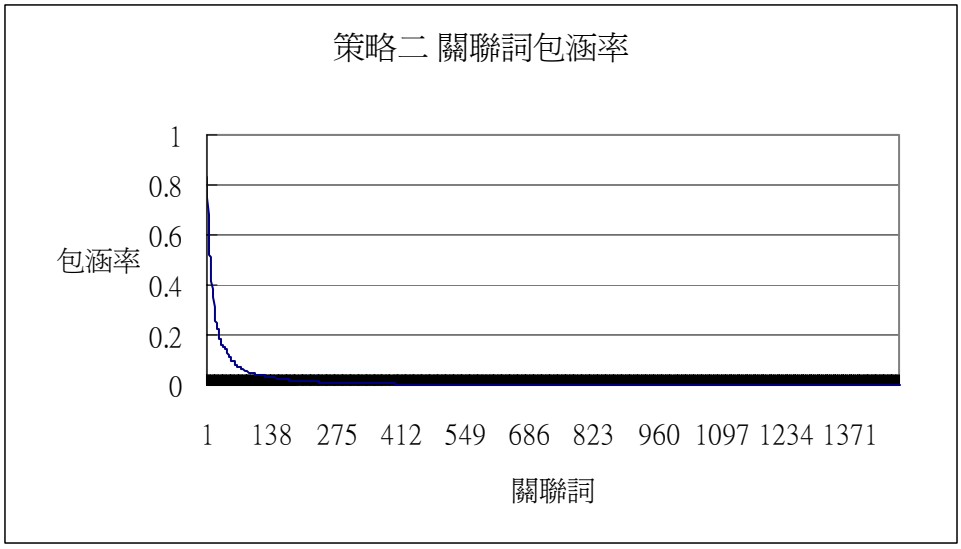


圖 5-2：策略二之關聯詞包涵率

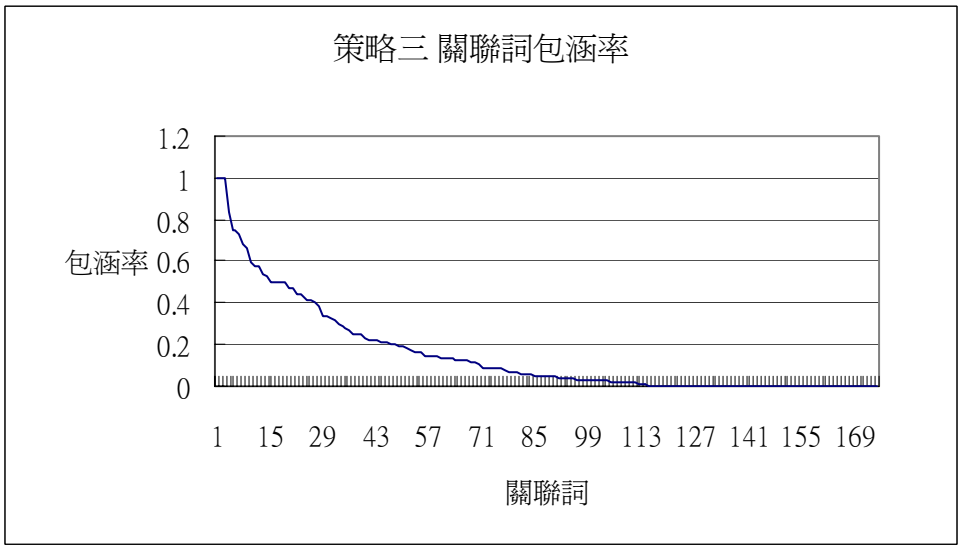


圖 5-3：策略三之關聯詞包涵率

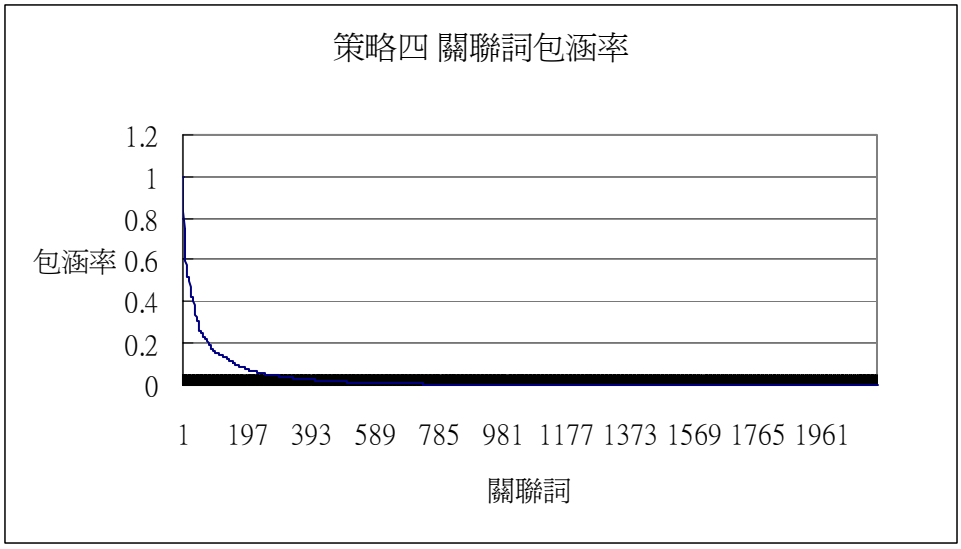


圖 5-4 策略四之關聯詞包涵率

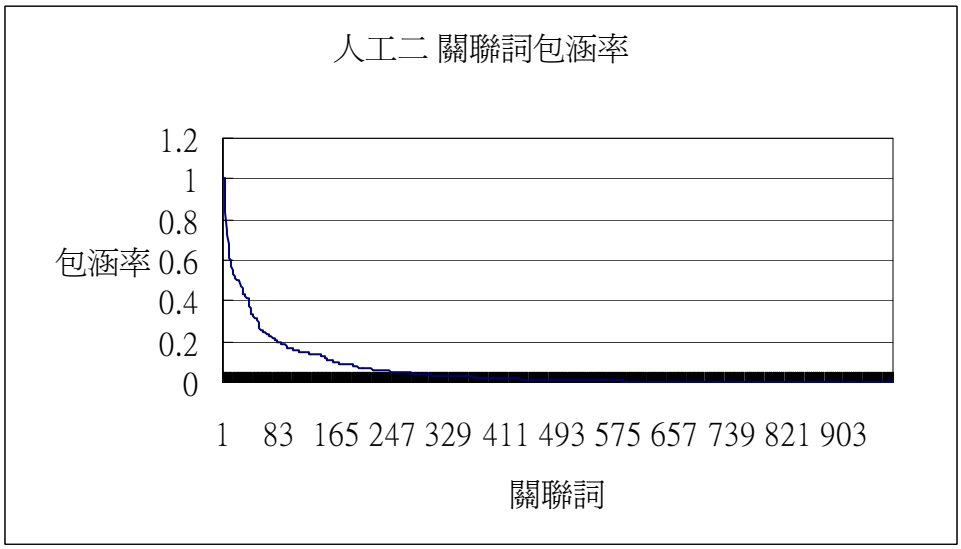


圖 5-5：人工二之關聯詞包涵率

二、關聯詞之重疊性

由實際檢索結果之成效，發現策略一、策略二、策略三檢索結果相當接近，

其平均精確率不超過 0.01，而其中策略一與策略二所使用之詞彙以人工觀察有大多數是相同之詞彙，故猜測此重複之詞彙使檢索結果一致。

以下進行各詞彙之重複情形進行分析，比較兩兩策略之詞彙重複比例，計算方式如下：

$$\text{重複比例} = \frac{2 \times \text{重複詞彙數}}{\text{兩策略之總詞彙數}}$$

計算結果紀錄於表 5-13

而觀察重複率之數據可知策略一與策略二重複率最高，將近 7 成，即策略一與策略二所篩選之詞彙有 7 成比例是一樣的詞彙。因此策略一與策略二所重複之詞彙對檢索成效影響較大，非重複之詞彙對檢索結果較無影響，推測僅使用兩策略重複之詞彙也可達到類似之檢索成效。

策略三與所有其他策略重複比例皆不達 2 成，而其確與策略一與策略二成效一致。重複率低即意味策略一與策略二對於策略三有不一致之關聯詞彙，卻有一致之成效，可推測若將兩者結合，以策略二補足策略三回收率之不足，以策略三補足策略二之精確率，預期可使檢索成效獲得提升。

策略四之詞彙與人工較爲接近，達到 3 成之比例，然而兩者之成效差距極大，可推斷策略四與人工不一樣的詞彙中，多數爲不相關的詞彙。

表 5-13：各策略關聯詞之重複比例

	策略一	策略二	策略三	策略四	人工一	人工二
策略一	1.0000	0.6969	0.0448	0.2214	0.2552	0.2345
策略二	0.6969	1.0000	0.0402	0.3500	0.2716	0.2568
策略三	0.0435	0.0395	1.0000	0.1240	0.1519	0.1566
策略四	0.2139	0.3403	0.1240	1.0000	0.3379	0.3705
人工一	0.2468	0.2652	0.1539	0.3461	1.0000	0.7187
人工二	0.2245	0.2488	0.1572	0.3729	0.7071	1.0000

第六章 結論與建議

本研究旨在研究如何透過查詢擴展之技術增加資訊檢索之成效，協助使用者資訊搜尋，讓使用者檢索出更多符合本身需求之資料。實驗計畫以局域或全域方式進行查詢擴展，並以各種檢索模式觀察其檢索成效。全域擴展使用以曾元顯教授所提供之關聯詞擴展技術，建立篩選詞彙之策略。並於查詢擴展使用篩選出之關聯詞彙，並觀察其成效，以建立最佳篩選策略。最後則針對查詢主題與關聯詞進行觀察與分析，以了解查詢主題需求與檢索詞彙對檢索成效之影響。本研究於此章歸納前述實驗與觀察之結論以及對研究之建議。

第一節 結論

本研究以局域或全域查詢擴展進行實驗，並利用不同檢索模式進行檢索與評估；此外，並針對主題與關聯詞進行觀察，具體結論如下：

一、查詢擴展成效

1. 檢索模式

本研究以 ByteSize Normalization、Pivoted Normalization Method、BM11、BM25、BM25m、Dirichlet prior retrieval method 等模式進行檢索測試，以 TREC_EVAL 工具計算平均精確率作爲實驗結果。根據數據明顯可知各檢索模式差異不大，尤其當查詢擴展所用擴展詞彙品質較高時，各模式差異較小；而當查詢擴展使用擴充詞彙品質不穩定時，各模式差距較大。

2. 局域擴展

局域擴展使用 Blind Relevance Feedback，雖然簡單卻可大幅提升檢索成效，不論搭配何種方式進行二度擴展，其檢索成效皆能有所提升，故再次驗證出 Blind

Relevance Feedback 爲穩定、有效、簡單的擴展策略。

3. 全域擴展

全域擴展使用人工篩選詞彙，可有最佳的檢索成效，尤其再搭配 Blind Relevance Feedback 進行二度擴展，成效可再大幅提升。而全域擴展使用策略二與策略三最佳，兩者差距並不明顯。策略二以關聯詞遞迴擴展之概念，擴展出可能有多意義之詞彙，而策略三以計算關聯詞與主題之強度進行篩選。

4. 本研究之最佳檢索策略

透過上述實驗結果發現，全域擴展以人工介入選詞再以 Blind Relevance Feedback 進行二度擴展，檢索模式以機率模式成效最佳。而不以人工介入方式，全域擴展以策略三篩選詞彙再以 Blind Relevance Feedback 進行二度擴展，檢索模式以機率模式成效最佳。

二、查詢需求與關聯詞對檢索之影響

1. 查詢需求分析

觀察不同難易度的主題，可知影響主題成效最直接的因素爲，描述需求時所使用的詞彙。當對於描述需求所使用的詞彙意義較複雜時，很難查詢擴展出相關的詞彙，而當描述需求所使用的詞彙意義較單純時，則較能夠以查詢擴展來增加檢索成效。此外，如果需求範圍廣泛，卻又嚴格限制相關範圍，同樣會使檢索難度增加。而全域擴展，所使用詞彙是根據關鍵詞而來，故當關鍵詞與主題需求關聯較強時，則詞彙對檢索較有幫助。反之，若關鍵詞對主題需求無關或相關度低時，則易使排序變差或是主題漂移。

2. 關聯詞分析

觀察各策略所擴展出的關聯詞彙，若篩選策略精確率越高則包涵率同樣較高，其檢索成效也較好，如策略三。而回收率高但包涵率低，則易對檢索產生負面效應，使檢索成效變差，如策略四。以查詢擴展而言，當精確率與回收率無法同時顧及時，應該選擇高精確率之策略，也就是寧願使成效增加較小幅度，而並避免使主題因擴展導致成效變差。

第二節 建議

一、研究之建議

1. 關聯詞彙之相關判斷

以人工判斷關聯詞與查詢主題是否相關。初步人工判斷的結果主要是根據 Description 欄位進行人工判斷，然而 Description 並不能真正代表主題需求，即人工判斷與主題之需求已經存有差距，造成判斷結果的不精確與檢索成效不理想。此外，人工進行相關判斷，由於各主題的關聯詞對主題強度不一，故人工選詞易發生不公平的情形。且某些低度相關之關聯詞因與同主題的其他關聯詞相較下強度小，而不視爲相關；某些主題因多數關聯詞爲低度相關，即使沒有較好的關聯詞可以被擴展，也只能選擇一些低度相關之關聯詞彙，故對關聯詞選擇不夠精細與一致，而對策略產生亦會造成不良的影響。關聯詞之相關判斷除了應該以查詢需求做主要依據外，更應該以關聯詞多重屬性，例如做詞性標記、相關程度等作爲判斷依據，俾方便後續篩選策略與其他研究之利用。

2. 全域擴展之擴充詞彙來源

本研究使用曾元顯教授所提供自動建立關聯詞之工具，進行詞彙篩選研究。查詢擴展詞彙篩選之目的，主要是希望擴展詞彙可提升檢索成效，並以自動化的方式進行爲目標。本研究建議，應該以多種詞彙來源進行測試，以進一步了解篩選策略的特性，觀察是否可穩定運作於各種詞彙來源。故對本研究所進行之篩選策略若能與人工篩選結果相近，則建議應多利用不同詞彙來源進行測試，不論是全域、局域、或是其它各索引典等，皆可進行詞彙篩選之測試以了解篩選策略的穩定性。

3. 關聯詞篩選策略

關聯詞篩選策略使用了許多參數，而這些參數多爲人工的觀察及經驗，如果策略用了非驗證之參數，易導致詞彙篩選策略不夠穩定。例如使用了詞頻限制，而詞頻係根據 NTCIR3 之資料集而訂定，一旦運作於其他較小或較大之資料集，則詞頻參數無法發揮作用，而造成篩選策略不穩定性。故建議研究篩選策略應盡量使用依存性低或非絕對參數，俾於後續研究之複製與未來之參考。

二、後續研究之建議

1. 擴充詞篩選策略

在目前本研究所採取策略中，雖然策略一、策略二、策略三檢索成效一致，然而經關聯詞分析後證實策略一與策略二過於相似，而策略三與所有策略相差較大，策略四成效較差。以目前進行的策略與人工篩選結果相比，並無較有效的策略。故建議後續研究應持續開發較有效力的篩選策略，並嘗試結合不同策略進行演算，以整合不同策略之優點。

2. 查詢擴展之詞彙對檢索成效之影響

在本研究所觀察的實驗中，大多是透過篩選出的詞彙對應檢索成效來進行分析，但對於單一詞彙對檢索成效並未充分理解，例如哪些詞彙對排序有幫助、哪些詞彙對回收率有幫助、哪些詞彙易造成負面效應等，可於後續研究再加以分析，不僅對詞彙篩選有所助益，相信也可提供資訊檢索者與系統開發者正面幫助。

參考書目：

卜小蝶（1996）。圖書資訊檢索技術。台北市：文華。

莊雅蓁（1999）。中文查詢句擴展問句之研究，碩士論文，國立台灣大學圖書資訊學系，台北市。

陳光華（2004）。資訊檢索的績效評估。2004 年現代資訊組織與檢索研討會論文集，129-136。

陳光華（2001）。資訊檢索系統的評估－NTCIR會議，台灣大學圖書資訊學系四十週年系慶研討會，69-73。

陳光華、莊雅蓁（2001）。資訊檢索之中文詞彙擴展。資訊傳播與圖書館學，8-1，60-70。

葉至誠，葉立誠（2003）。研究方法與論文寫作。台北市：商鼎文化。

葉佳昀（2004）。中文互動式檢索輔助功能之效益評估-以關聯提示詞爲例，碩士論文，私立輔仁大學圖書資訊學系，台北縣。

曾元顯（2001）。共現索引典之自動建構、評估與應用，台灣大學圖書資訊學系四十週年系慶研討會，87-105。

曾元顯（1997）。關鍵詞自動擷取技術與相關詞回饋。上網日期：2005 年 03 月 11 日。網址：http://blue.lins.fju.edu.tw/~tseng/papers/feedback.htm。

曾元顯，林瑜一（1998）。模糊搜尋、相關詞提示與相關詞回饋在 OPAC 系統中的成效評估。中國圖書館學會會報，61，103-125。

黃慕萱（1996）。資訊檢索。臺北市：臺灣學生。

鄭恆雄（1984）。中文資料索引及索引法。台北市：文史哲。

Cui, Hang., Wen, J. R., Nie, J.Y., & Ma, W.Y.（2002）. Probabilistic query expansion using query logs. Proceedings of the 11th international conference on World Wide Web, 325-332.

Harman, D.（1988）. Towards Interactive Query Expansion. Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, 322-326.

Hui Fang, Tao Tao, ChengXiang Zhai.（2004）. A Formal Study of Information Retrieval Heuristics. Proceedings of the 27th annual international conference on Research and development in information retrieval, 50-55.

Kwon,O.W, Kim C.M. & Choi, K.S. （1994）.Query Expansion Using Domain-Adapted, Weighted Thesaurus in an Extended Boolean Model. Proceedings of the third international conference on Information and knowledge management, 140-146.

Mandala, R., Tokunage, t., & Tanaka H.（1999）. Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 191-197.

Mitra, M, Singhal, A., &. Buckley, C.（1998）. Improving Automatic Query Expansion. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 206-214.

NII. （2003）. README for Topics and Relevance Assessments of NTCIR-3 CLIR

Test Collection - <Formal runs>. Retrieved March, 15, 2005. From http://reseaarch.nii.ac.jp/ntcir/permission/READMEforTOPICS_FormalRun.htm

Qiu, Y., & Frei, H. P.（1993）. Concept Based Query Expansion. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 166-168.

Ricardo,B.Y.,&Berthier,R.N.（1999）.Modern Information Retrieval. New York：Addison Wesley.

Robertson, S. E. （1969）. The Parametric Description of Retrieval. Test. Journal of Documentation, 25 （1）, 3.

Sakai, T., Kajiura, M., & Sumita, K.（2000）. A first step towards flexible local feedback for ad hoc retrieval. Proceedings of the fifth international workshop on on Information retrieval with Asian languages, 95-102.

Tseng, Y. H.（2002）. Automatic Thesaurus Generation for Chinese Documents. Journal of the American Society for Information Science and Technology, 1130-1138.

Tseng, Y. H., Juang, D. W., & Chen, S. H.（2004）. Global and Local Term Expansion for Text Retrieval. Proceedings of the Forth NTCIR Workshop on Evaluation of Information Retrieval. Retrieved March 23, 2005, from http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/CLIR/NTCIR4WN-CLIR-TsengY.pdf.

Xu, J., & Croft, W. B.（1996）. Query Expansion Using Local and Global Document Analysis. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 4-9.