資訊檢索之中文詞彙擴展

Expansion of Chinese Words in Information Retrieval

陳光華

Kuang-hua Chen

國立臺灣大學圖書資訊學系副教授

Associate Professor, Department of Library and Information Science, National Taiwan University

莊雅蓁

Ya-chen Chuang

國立臺灣大學圖書資訊學系碩士

Graduate Student, Department of Library and Information Science, National Taiwan University

關鍵字(Keywords)： 資訊檢索(Information Retrieval)，查詢問句擴展(Query Expansion)，索引典(Thesaurus)，同義詞(Synonym)

【摘要】本研究主要探討議題有三：一，自動建構之同義詞典對資訊檢索之輔助效益；二，以何種索引典詞彙關係來擴展查詢問句可得到最佳的效益；三，同義詞典與索引典整合輔助檢索的效益分析。限於詞彙資源取得不易，本研究採用實驗文件資料庫爲基礎，以進行查詢問句擴展的實驗。首先，蒐集原始查詢問句，再以不同的詞彙來源，包括以同義詞典及索引典分別擴展查詢問句，以及整合兩者再擴展查詢問句，建構多組不同的查詢問句擴展模式。實驗結果的效益評估，由人工進行相關判斷，再依判斷所得計算檢索結果的求準率。研究顯示，以同義詞典詞彙群內詞彙數量較少的層次來擴展查詢問句，可得到較好的檢索效益；不過以索引典各種詞彙關係來擴展查詢問句時，檢索結果沒有顯著的差異。整體而言，以整合所有詞彙關係的擴展模式有較好的檢索效益，尤其以同義詞典擴展，再輔以索引典加權後的檢索效益可略爲提升。但如再以索引典進行二次擴展時，檢索效益反而降低。實驗亦發現自動建構的同義詞典內容，受斷詞品質的優劣所影響，因此，對查詢問句擴展的檢索效益而言，字串比對方式亦是重要的影響因素。

【Abstract】This thesis aims at three important issues for query expansion: whether the automatic constructed synonym dictionary could enhance the retrieval effectiveness, which relationship of thesaurus has the best performance, and the effectiveness of the integration of synonym dictionary and thesaurus. In the experiments of query expansion, the queries are expanded in different models, including expanding by either synonym dictionary or thesaurus or both. Finally, performance is evaluated in precision. The results show that query expansion using second level of synonym dictionary has better performance. Though the effects of different relationships prescribed in the thesaurus are similar, expanding by union of all relationships shows better performance. The model of first expanding query by synonym dictionary then modifying it by thesaurus has improved retrieval performance slightly, but the performance is decreased in further expansion. We also find that the correctness of word segmentation has a great impact on the quality of synonym dictionary. The mode of string mapping is another important factor.

一、前言

目前使用者與各種資訊檢索系統之間的溝通模式，多數是以詞彙爲主，其檢索過程是使用適當的詞彙來導引概念，亦即資訊的檢索是用詞彙來代表概念。但概念與詞彙之間的關係時常混淆，因爲概念與詞彙的關係並非都是一對一的，如同義詞（Synonym），表示多個詞彙可代表一個概念；而同形異義詞（Homographs），則指一個詞彙可以代表多個不同的概念。除此之外，不同的人使用同一詞彙時，也可能賦予它不同的意義。因此，在檢索時，需建構概念與詞彙間的明確關係，才能有效提升檢索效益。（註 1）除了利用關鍵詞進行全文檢索（Full-Text Search）外，有些資訊檢索系統在資訊組織與整理的過程，尙針對文件的內容進行分析，給予文件資料檢索標識（如主題詞彙或分類號），並使用索引詞彙來表示文件內容。然而相對地，資訊使用者提出查詢問句後，也同樣必須經過概念分析與將概念轉換成詞彙的過程，再藉著與資訊檢索系統的互動來找尋所需的資訊。亦即，資訊使用者與資訊檢索系統之間，是藉由索引詞彙與檢索詞彙之間的對映來達到擷取與過濾資訊的目的。因此，詞彙是資訊使用者與各種資訊處理系統溝通之重要元素。（註 2）

資訊檢索系統中，文件與查詢問句所具有的相同概念可藉由詞彙顯現，而詞彙本身的語意關係（如關聯詞、狹義詞及廣義詞）亦在概念空間內相互連結。因此，詞彙不僅表現出文件的內容，同時也表達出檢索者的資訊需求。一般而言，如果檢索者可以將問題用正確且適當的詞彙表達，則系統也可以將之對映到相同概念的索引詞彙，如此一來，檢索結果應能滿足使用者的資訊需求。換言之，理想的檢索過程應該是一個「直接精準的處理過程」（One-Shot Process），也就是以極準確的方式處理「提出問題→選擇檢索詞彙→進行檢索→提供答案」的程序。（註 3）

事實上，在文件描述和查詢問句陳述等不同階段，都有相當程度的不確定性。例如，詞彙的選擇可能是隨機的，因爲實在無法預知使用者到底會選擇那些詞彙，所以，即使使用非常嚴謹的索引典，一本書還是可以用很多詞彙來描述，而通常索引者和檢索者不見得會使用相同的詞彙來描述同一本書。

此外，檢索過程中亦很難確認詞彙的複雜性，以及相關詞彙間細微的區分。通常檢索者必須經過一連串的過程才能確定該使用那些檢索詞彙，這些過程包括對系統文件描述方式的認知、對索引語言的認知，以及經由詞彙的語意關係匯集更多詞彙意義等。這表示使用者在下達查詢問句時，必須盡可能將所有相關詞彙列出，才能檢索到足夠的資訊。但以目前線上檢索系統而言，一方面系統很少提供這種功能，一方面大部分的使用者不知道應盡可能將相關詞彙列出才能提高檢索品質，而且即使系統提供線上索引典以輔助查詢問句的建構，使用者通常也不知道有這樣的功能可利用。（註 4）另外，使用者對該學科領域的認知或許不深，就算是該學科領域的學科專家，也可能不願意花時間將所需詞彙全部鍵入系統。總之，使用者在檢索資訊時，通常只提供大約可敘述其需要的詞彙即停止，而且幾乎都是很短的查詢問句，當然得到的檢索結果也較差。因此，實有迫切需要於系統上加強協助使用者建構查詢問句的功能。

理想的資訊檢索系統，除了能找出完全符合搜尋條件的文件外，還應檢索出在意義或概念上接近的文件，以解決因用語不同而造成的檢索效益不彰問題。此外，一般使用者期望的是系統能提供最簡易的檢索程序，然後得到令人滿意的檢索結果，所以，資訊檢索系統內部機制的精進與提升，實爲資訊檢索研究的重心。爲了讓使用者有效地檢索出所需的文獻資料，本論文希望以查詢問句擴展（Query Expansion, QE）的方式，建構符合使用者資訊需求的檢索機制，以增進中文檢索系統之檢索效益。

本文將探討下列幾項議題：

- 利用文獻分析探討查詢問句擴展的相關研究與技術。
- 利用同義詞典進行中文查詢問句擴展的實驗，探討利用同義詞典擴展查詢問句之檢索效益。
- 利用索引典進行中文查詢問句擴展的實驗，探討利用索引典擴展查詢問句之檢索效益。
- 整合同義詞典與索引典進行中文查詢問句擴展的實驗，探討整合不同詞彙資源擴展查詢問句之檢索效益。

研究期望建構與領域無關（Domain-Independent）

的查詢問句擴展機制，但因國內尚缺乏架構較完備的通用性索引典，故採用科學技術資料中心出版的「科技索引典」爲實驗對象，以發展出具有適應性的模式。

除了使用索引典外，由於同時要進行透過同義詞典擴展查詢問句的實驗，故必須準備一部同義詞典，但同樣的，由於現今科技領域並無同義詞典形式的詞彙資源，因此決定採用作者先前自動建構的同義詞典（註 5），做爲輔助工具，探求其輔助效益。

本文的結構如下所示，首先探討資訊檢索與查詢問句的技術，尤其特別著重於文獻所載查詢問句擴展的相關研究。接著詳細說明在中文資訊檢索環境下，透過索引典以及同義詞典進行查詢問句擴展的實驗。然後，提出初步的實驗結果，並分析其結果，以探討索引典與同義詞典對檢索效能的影響。最後則提出我們的結論以及未來的研究方向。

二、 相關研究

查詢問句的擴展通常都以使用者提供的檢索詞彙爲基礎，當原始查詢問句的檢索效益不好時，則可以追加更多的詞彙來改善。至於查詢問句的擴展，可以利用相關回饋（Relevance Feedback）或是使用知識架構（Ontology）的元知識（Atom Knowledge）來進行。相關回饋指以初次檢索結果爲基礎的查詢問句擴展，不過其效益隨原始查詢問句、排序的公式及相關詞彙的數量、初次檢索結果品質而改變；然以知識架構爲基礎的查詢問句擴展並不依賴檢索結果，而普遍多以統計或是以語料庫爲基礎。同樣的，索引典中詞彙的同義詞也可用以擴展查詢問句，所謂的同義詞可指十分相關的詞彙，亦可謂於文法上或語意上完全可相互取代的詞彙。Gauch 與 Smith 於 1993 年的研究指出，以線上索引典修飾查詢問句，對檢索結果有所提升。（註 6）

查詢問句擴展的主要方法有三種：以查詢問句爲基礎的查詢問句擴展、以語料庫爲基礎的查詢問句擴展、及以語言分析特性爲基礎的查詢問句擴展。若以使用者的角度來看，查詢問句擴展技術可分爲半自動與自動。半自動的查詢問句擴展技術要求使用者對檢索所得的文件進行相關判斷後再回饋給系統修飾原始查詢問句，而使用者與系統的互動，將持續到使用者滿意其檢索結果後才結束。半自動查詢問句擴展的主要優勢爲減輕使用者重新組織查詢問句的負荷，但其缺點爲使用者在得到良好檢索結果之前，必須花時間與精神來判斷文件的相關度。半自動化查詢問句擴展也被稱爲互動式查詢問句擴展（Interactive QE）或使用者輔助查詢問句擴展（User-aided QE）。另一種半自動查詢問句擴展的技術爲結合可瀏覽的索引典，使用者先給予系統一組查詢問句，系統則提供相關詞彙清單提供使用者挑選以取代或增加原有的查詢問句，不過這種方式的效益仍有賴於使用者對該學科領域的認知以及使用者與系統的良好互動。

自動化的查詢問句擴展技術則不需依賴使用者的相關判斷，通常是以語言分析或詞彙共現等技術爲基礎。Qiu 與 Frei 認爲自動化查詢問句擴展的技術可分爲四類：詞彙共現法（Use of Co-occurrence Data）、文件分類法（Use of Document Classification）、語法分析法（Use of Syntactic Context）及相關資訊回饋法（Use of Relevance Information）。詞彙共現法的查詢問句擴展方法是依詞彙與其他詞彙間的相關程度來聚集相關詞彙，並且區隔不相關詞彙；文件分類法則將文件先依文件分類演算法進行分類，認爲出現在同一類文件的詞彙彼此間有某程度的相關；而語法分析法中，詞彙關係是利用語言學知識與詞彙共現的統計分析來建立，於查詢問句擴展的應用上即於原查詢問句中加入最相關的詞彙；相關資訊回饋法則包括相關回饋技術。（註 7）

利用相關回饋在原始查詢問句中追加詞彙的查詢問句擴展是一常見的技術，且具有相當的效益。然而，正如前文所言，對於大部分的使用者而言，要提供相關回饋所需要的相關判斷資訊實爲困難。在這種情況下，可採行 Ad Hoc 與 Blind Feedback 擴展查詢問句，此方法使用準相關回饋（Pseudo Relevance Feedback），而非實際要求使用者輸入：利用原查詢問句檢索出之一組文件，不經使用者判斷即假定所有文件皆爲相關，而這些假定的相關文件即經由相關回饋的程序建構新的查詢問句，再利用其做進一步的檢索。此方法有一明顯的缺點：若假定之相關文件清單中，實際上不相關的文件占大部分，那麼加入原始查詢問句的擴

展詞彙與原檢索主題並不相關，則擴展後查詢問句的檢索品質會變差。（註 8）

雖然相關回饋能顯示驚人的效果，但也顯露當加入太多擴展詞彙之後所導致的失敗，而且隨著文件資料庫的不同或是文件清單排序方式的不同，也會有不同的結果。對於利用相關回饋資訊所進行的自動查詢問句擴展，新加入詞彙的數目亦是決定檢索效益的重要因素。Harman 發現從相關回饋所產生的詞彙清單擷取詞彙時，追加 20 個詞彙，檢索效益有所提升，但超過 20 個後就降低了，這表示每一個候選詞彙清單都存在一個理想的切點（Cut-off Point）。Harman 使用 Cranfield 1400 測試資料庫進行實驗，發現切點是介於 20 到 30 之間。然而 Buckley 指出使用不同的測試集（TREC）、不同的檢索系統，以及不同的相關回饋加權演算法時，其最佳的擴展詞彙數目則是全部皆加入，因此認爲最合適的切點是全面考慮文件集、查詢問句、系統與使用者而得的，理想的參數不可能以自動的方式決定。（註 9）

索引典爲查詢問句中詞彙的重要來源，索引典可提供概念詞彙之同義詞、近同義詞、反義詞、廣義詞、狹義詞及相關詞等，但是索引典通常是由學科專家或索引專家挑選的控制詞彙所組成，未受訓練或不熟悉的使用者很難靈活運用索引典。一般而言，索引典仍是資訊檢索的重要詞彙資源，具有豐富的資訊與價值，然而，目前缺乏適切的技術與方法妥善應用索引典，因此爲求有效地使用檢索詞彙，仍不能忽略索引典在查詢問句擴展上的運用。

Stiles 是最早提出利用相關詞彙來改進檢索效益理論的學者之一。Doyle 進一步提出無論人工或機器判斷出來的相關詞彙都應加入檢索。Courtial 及 Pomian 則強調在科學與技術領域中使用者需要的檢索詞彙及概念常超出傳統索引典中所能提供的資源。（註 10）

有一研究發現，依據統計分析的詞彙關係擴展查詢問句，將會降低檢索效果。Harman 指出，利用自動建構的索引典來擴展查詢問句有不佳的檢索結果，但若加上相關回饋機制則有提升的效果。

爲了擴展查詢問句，使用者輸入的檢索詞彙可經由除去詞尾、加入同義詞或利用索引典中詞彙的關係進行擴展。其中從索引典挑選相關詞彙的順序爲父節點→兄弟節點→子節點，利用上一層的概念取代原檢索詞彙的方式，不但應用於人工的詞彙擴展，亦可運作於自動化詞彙擴展系統。然而 Crouch 發現從索引典挑選新的檢索詞彙加入查詢問句中，比利用索引典的相關詞彙取代原檢索詞彙有較佳的檢索結果。Harman 則主張索引典所提供的相關詞彙經由使用者挑選之檢索結果較佳。（註 11）

Harman 的實驗是以自動程序模擬使用者的挑選動作，企圖建構一組理想的擴展詞彙，他指出無論以任何方式或任何資源所擷取之候選詞彙清單（例如相關回饋、除去詞尾、同義詞、索引典），皆可經二次的挑選以進一步產生第二次詞彙清單，使得詞彙清單中詞彙的相關程度更爲提升。實驗方法爲平均第一次挑選 20 個詞彙，第二次挑選的詞彙數目是從初次挑選的 20 個中抽取 12 個，結果顯示，只利用第二次所擷取出之詞彙清單進行查詢問句擴展有顯著的檢索效益提升。因此 Harman 認爲，若使用者可以具備相同的詞彙挑選能力，則半自動之查詢問句擴展的檢索效益於理論上遠超過自動查詢問句擴展。（註 12）

前述學者專家提出的方案皆係應用於英文的資訊檢索環境，本文則以中文資訊檢索環境爲實驗的平台，利用索引典與同義詞典之詞彙資源，改善使用者建構的查詢問句過短而提供予檢索之資訊過少所造成之檢索效益不彰的情形，並討論索引典與同義詞典在中文查詢問句擴展扮演的角色及其效能。

三、查詢擴展的設計與實驗

本研究的實驗設計方法是將使用者建構的查詢問句，在進行檢索之前，先經過同義詞典與索引典的詞彙擴展處理，再將經過修正擴展後之查詢問句於文件資料庫中進行檢索。爲了分析索引典與同義詞典對於檢索效益之影響，同時探討結合兩種資源對於檢索效益整體的影響，因此設計以下五種不同的實驗：

- 基礎檢索（BL）
- 同義詞典擴展（SE）
- 索引典擴展（TE）
- 同義詞典擴展，索引典加權（SETW）
- 同義詞典擴展，索引典加權與擴展（SETWE）

基礎檢索是以使用者建構之原始查詢問句直接進

行檢索，所得之檢索結果做爲評比的基準，以分析不同的詞彙資源與不同的詞彙擴展設計對於檢索效益的影響。由於本研究使用之同義詞典是利用實驗之文件資料庫建構而成，因此由同義詞典擴展之實驗，可了解同義詞典建構方法之優劣及對檢索效益提升之成效。以索引典進行擴展之實驗，可再細分爲數個以不同詞彙關係擴展查詢問句之實驗模組，主要目的除了比較索引典與同義詞典對於檢索效益提升之效能，也能夠了解不同詞彙關係對於檢索結果的個別影響。以同義詞典擴展再以索引典加權之實驗，是以同義詞典擴展之後，將新的查詢問句利用索引典之詞彙資源以彙，減低同義詞典建構時可能包含之雜訊。以同義詞典擴展再以索引典加權與擴展之實驗，則是於同義詞典擴展之後，以檢索詞彙加權的方式強調詞彙的相關度，再利用索引典中詞彙關係，加入更多相關詞彙。

本研究設計之各項實驗，目的在於尋找最佳之查詢問句擴展模式，希望運用各種可能提升檢索效益的方法，建構整體上表現最好的查詢問句。對於各查詢擴展形式的檢索效益，本研究僅採用求準率評估各實驗的效益，因爲求全率（Recall）是實際檢出的文件數目與文件集合中所有真正相關文件數目的比率，故實驗語料本身必須具備標準答案，但本研究使用的語料缺乏此條件，無法計算求全率，而求準率（Precision）的定義是指實際檢出的文件有多少比率是真正相關的文件，本研究可在取得檢索結果之相關候選文件後，以爲判斷檢索出的文件是否與查詢問句相關，再計算求準率，因此本研究以求準率做爲檢索效益的指標。

（一）基礎檢索（BL）

基礎檢索進行檢索之步驟爲：

1. 將原始查詢問句（Q_O）中的檢索詞彙進行 bi-gram 處理，轉換爲查詢問句向量。
2. 將查詢問句向量與文件向量進行相關係數（Sim(Q,D)）計算。
3. 將檢出文件以 Sim(Q,D)相關係數值排序，選取前 50 篇進行相關判斷
4. 依據相關判斷結果計算求準率以評估檢索效益。

（二）同義詞典擴展（SE）

使用自動建構之同義詞典做爲詞彙擴展之依據，本同義詞典共分成四個層次，同義詞典擴展的實驗即分別用不同層次之相似詞彙群擴展原始查詢問句之檢索詞彙：

- 以第二層詞彙群擴展
- 以第三層詞彙群擴展
- 以第四層詞彙群擴展

以同義詞典進行查詢問句擴展之步驟爲：

1. 將原始查詢問句（Q_O）中之檢索詞彙，一一於同義詞典中進行字串比對。
2. 於同義詞典中搜尋出相符詞彙之後，將該詞彙所屬之詞彙群中其他成員全部加入原查詢問句，形成以同義詞典擴展後之查詢問句（Q_S）。
3. 將擴展後之查詢問句（Q_S）中的檢索詞彙進行 bi-gram 處理，轉換爲查詢問句向量。
4. 將查詢問句向量與文件向量進行相關係數（Sim(Q,D)）計算。
5. 將檢出文件以 Sim(Q,D)相關係數值排序，選取前 50 篇進行相關判斷。
6. 依據相關判斷結果計算求準率以評估檢索效益。

（三）索引典擴展（TE）

系統首先將使用者的查詢問句與索引典進行對映，再依據所對映的詞彙於索引典中搜尋不同詞彙關係之相關詞彙，再將該相關詞彙加入原查詢問句進行擴展。以索引典擴展的實驗中，每一查詢問句以五種不同的詞彙關係進行擴展：

- 狹義詞擴展（Narrower Term, NT）
- 等同詞擴展（USE, USE FOR）
- 關聯詞擴展（Related Term, RT）
- 廣義詞擴展（Broader Term, BT）
- 聯合擴展（NT, BT, RT, USE）

等同詞擴展的目標是包含索引典中敘述同一概念的所有可能詞彙。相同地，其他以不同詞彙關係擴展的實驗依次爲利用索引典中狹義詞、關聯詞以及廣義詞等詞彙關係進行擴展。最後的聯合擴展則將所有不同詞彙關係的擴展以聯集方式整合。

本研究中以索引典進行查詢問句擴展時，會使用敘述語（Descriptor）與款目詞（Entry Term）。另外，擴展階層以兩層爲限，亦即本研究只取用距離檢索詞

彙之深度向上及向下各爲二層以內的詞彙。對於一些檢索詞彙，有些詞彙關係並不存在於索引典中，如等同詞、狹義詞、廣義詞以及關聯詞，故可能有些查詢問句的某種詞彙關係擴展結果是無擴展的狀態。

以索引典進行查詢問句擴展之步驟爲：

1. 將原始查詢問句（Q_O）中之檢索詞彙，一一於索引典中進行字串比對。
2. 於索引典中搜尋相符詞彙之後，依據實驗設計，搜尋與該詞彙具某種詞彙關係之相關詞彙，將之加入原查詢問句，形成以索引典擴展後之查詢問句（Q_T）。
3. 將擴展後之查詢問句（Q_T）中的檢索詞彙進行 bi-gram 處理，轉換爲查詢問句向量。
4. 將查詢問句向量與文件向量進行相關係數（Sim(Q,D)）計算。
5. 將檢出文件以 Sim(Q,D)相關係數值排序，選取前 50 篇進行相關判斷。
6. 依據相關判斷結果計算求準率以評估檢索效益。

（四）同義詞典擴展，索引典加權（SETW）

本實驗利用索引典修正同義詞典擴展之後可能產生的雜訊，其假設爲，若由同義詞典擴展後之查詢問句中的檢索詞彙亦被索引典所收錄，則可保證該詞彙之正確性，而未收錄於索引典之其他檢索詞彙來自自動建構的同義詞典，因此應加權出現於索引典的檢索詞彙。對於詞彙的加權值須如何設定，目前仍未有可供依循的標準，故本研究將檢索詞彙之權重初步設定爲：原查詢問句與擴展後查詢問句之所有檢索詞彙權重皆爲 1，可於索引典中對映成功之檢索詞彙加權後之權重爲 2。另外，對於應以同義詞典之何種層次擴展查詢問句，本組實驗則依據以同義詞典擴展（第二組）實驗結果，選擇檢索效益最佳者進行同義詞典的擴展，再以索引典加權。以同義詞典進行查詢問句擴展，再以索引典加權之步驟爲：

1. 將原始查詢問句（Q_O）中之檢索詞彙，一一於同義詞典中進行字串比對。
2. 於同義詞典中搜尋相符詞彙之後，將該詞彙所屬之詞彙群中其他成員全部加入原查詢問句，形成以同義詞典擴展後之查詢問句（Q_S）。
3. 將 Q_S 中所有檢索詞彙，一一於索引典中進行字串比對。
4. 若 Q_S 中之檢索詞彙亦收錄於索引典中，則該檢索詞彙之權重爲 2。
5. 將以同義詞典擴展並以索引典加權後之查詢問句（Q_{SM}）中的檢索詞彙進行 bi-gram 處理，轉換爲查詢問句向量。
6. 將查詢問句向量與文件向量進行相關係數（Sim(Q,D)）計算。
7. 將檢出文件以 Sim(Q,D)相關係數值排序，選取前 50 篇進行相關判斷。
8. 依據相關判斷結果計算求準率以評估檢索效益。

（五）同義詞典擴展，索引典加權與擴展（SETWE）

本組實驗基本程序與第四組實驗相同，但進一步利用索引典的詞彙關係進行第二次擴展，亦即針對索引典加權的詞彙再進行擴展。第二次擴展的索引典詞彙關係，則參考以索引典擴展（第三組）的實驗結果，選擇檢索效益最佳者，而第二次擴展的詞彙爲索引典所收錄，其權重亦設定爲 2。以同義詞典進行查詢問句擴展，再以索引典加權與擴展之步驟爲：

1. 將原始查詢問句（Q_O）中之檢索詞彙，一一於同義詞典中進行字串比對。
2. 於同義詞典中搜尋相符詞彙之後，將該詞彙所屬之詞彙群中其他成員，全部加入原查詢問句，形成以同義詞典擴展後之查詢問句（Q_S）。
3. 將 Q_S 中所有檢索詞彙，一一於索引典中進行字串比對。
4. 若 Q_S 之檢索詞彙亦收錄於索引典，則檢索詞彙之權重爲 2，產生擴展並加權之查詢問句（Q_{SM}）。
5. 將與比對成功之詞彙具某種詞彙關係之相關詞彙加入 Q_{SM}，該相關詞彙之權重皆設爲 2，產生二次擴展後之查詢問句（Q_{SM2}）
6. 將以同義詞典擴展並以索引典加權與擴展後之查詢問句（Q_{SM2}）中的檢索詞彙進行 bi-gram 處理，轉換爲查詢問句向量。
7. 將查詢問句向量與文件向量進行相關係數（Sim(Q,D)）計算。
8. 將檢出文件以 Sim(Q,D)相關係數值排序，選取前

50 篇進行相關判斷。

9. 依據相關判斷結果計算求準率以評估檢索效益。

四、實驗材料與評估準則

本研究使用的文件爲國科會科資料中心提供之「中華民國科技研究報告摘要資料庫」，收錄國內大專院校、研究機構與公民營企業等單位資助之研究計劃成果報告摘要。所使用的資料內容爲民國七十一年起至民國八十年間提出並執行完成之研究計畫成果報告之書目摘要，共計 11,875 筆。文件長度約 200 至 500 字不等，文件類別分爲數理、工程、醫學、農業及人文社會等五大類。文件資料庫之文件類別與文件數量如表一所示。

索引典使用的是國科會科資中心編訂之科技索引典。本索引典係參考日本科技資訊中心的 JICST 索引典（JICST Science and Technological Thesaurus）、美國國家醫學圖書館的 MeSH、英國國家標準局的 ROOT 等國際上知名索引典之優點，中英文並列，其中科技專門詞彙之中文譯名則採用教育部公布之各類科技詞彙。內容採三層架構，所涵蓋的主題共分 19 類，大類之下又細分爲 188 中類及 1,011 小類，每一小類均包括詳細之主題範圍說明。科技索引典涵蓋之主題範圍如表二所示。至於同義詞典，正如前文提及，是先前以自動程序建構而成的。

爲了進行效能的評估，實驗使用的查詢問句均經特別設計。目前無論是網際網路上的搜尋引擎使用者，或是專門資料庫檢索系統之使用者，大部分以多個檢索詞彙來表示查詢的主題。本研究則以多詞組合方式表達查詢問句，以充分表達使用者資訊需求的三至五個檢索詞彙，來建構檢索實驗之原始查詢問句。

爲了避免查詢問句涵蓋類別之文件數量過低，導致無效的檢索結果，我們依據實驗文件資料庫中文件類別分佈情形，先挑選文件數量較多的類別，在特定類別之下，由各學科領域之大學部及研究所學生提供特定檢索主題，再以三個以上的詞彙表達檢索主題，組成一原始查詢問句。經由初步之檢索測試，多數查詢問句的檢索主題之相關文件數量過低，因此篩選後，只保留其中 12 個查詢問句做爲實驗的查詢問句。表三爲原始查詢問句之檢索詞彙與主題敘述。

表一、文件資料庫之文件類別與文件數量

科資中心分類號	學 科 主 題	文件數量
AA	科技總論	87
CA	數學	194
CB	物理學	243
CC	化學	352
CD	地球科學	337
CE	生物科學	803
EA	醫學	1646
GA	農業科學	1050
IA	資訊工程	394
IB	工業工程與管理	104
IC	能源	5
ID	核子工程	76
IE	電子電機工程	726
IF	機械工程	657
IG	土木工程	499
IH	環境科學與工程	157
II	礦業工程	42
IJ	材料科學工程	308
IK	化學工程	458
PA-PH, SA-SM	人文社會科學	694
分類號爲舊版		2009
無分類號		1072

表二、科技索引典主題範圍

科技總論	生物科學	能源	環境科學與工程
數學	醫學	核子工程	礦業工程
物理學	農業科學	電子電機工程	材料科學與工程
化學	資訊工程	機械工程	化學工程
地球科學	工業工程與管理	土木工程	人文社會科學總類

表三、原始查詢問句

QUERY 1	主題敘述	探討有關建築物結構之防震或耐震設計
	檢索詞彙	建築物,結構體,防震,耐震
QUERY 2	主題敘述	染色體分析對於診斷先天畸形,智能不足,習慣性流產,惡性腫瘤等疾病之研究
	檢索詞彙	染色體,癌,流產,畸形
QUERY 3	主題敘述	探討防洪工程與洪水預報作業所必須之降雨量與河川流量變化觀測,逕流分析等技術
	檢索詞彙	洪水,預報,降雨,河川流量,逕流
QUERY 4	主題敘述	探討有關紅樹林沼澤生態環境之研究,包含紅樹林區中各別生物生態之研究
	檢索詞彙	紅樹林,沼澤,生態
QUERY 5	主題敘述	有關以射出成型法製作塑膠製品之相關材料及技術的研究
	檢索詞彙	射出成型,塑膠,塑膠加工,模具,拔模
QUERY 6	主題敘述	有關電力系統的控制與穩定技術研究, 包括軟硬體
	檢索詞彙	電力系統,穩定器,控制器
QUERY 7	主題敘述	有關電腦輔助教學多媒體的應用與軟體設計
	檢索詞彙	電腦輔助,教學,多媒體,軟體,
QUERY 8	主題敘述	有關台灣地區颱風行進路徑及強度等預報之研究
	檢索詞彙	颱風,天氣預報,風速,路徑
QUERY 9	主題敘述	有關廢水及重金屬污泥之處理技術
	檢索詞彙	廢水,重金屬,污泥,危害性廢棄物,
QUERY 10	主題敘述	有關流動流體之對流熱傳遞,增強熱傳性能之研究,任何對流系統組成因素（包括流體種類,流動型態,界面形狀,方向性等等）對提升熱傳遞性能的研究.
	檢索詞彙	流動流體,熱傳遞,對流
QUERY 11	主題敘述	機器人或機械手臂控制系統設計之研究
	檢索詞彙	機器人,座標系,自由度,伺服
QUERY 12	主題敘述	食物營養成分對動物體內脂質含量（如血清膽固醇量,血清脂質,肝脂質等）或膽固醇代謝之影響
	檢索詞彙	膽固醇,脂質,脂質含量

用以評估資訊檢索系統之檢索效益，必須以是否符合使用者之資訊需求爲基礎，包括檢索結果之文件數量、文件內容與資訊需求的相關程度、花費的檢索時間等等，最重要的則爲檢索出來的文件是否爲使用者所需要，是否能滿足使用者的資訊需求，目前的資訊檢索效益的評估通常使用求全率（Recall）與求準率（Precision），其計算公式如下：

$$\text{Recall} = \frac{\text{檢索所得文件中真正相關的文件數}}{\text{文件集中真正相關的文件數}}$$

$$\text{Precision} = \frac{\text{檢索所得文件中真正相關的文件數}}{\text{檢索所得之文件數}}$$

本研究對於查詢問句擴展機制的檢索效益評估上，由於實驗環境並非既有的測試文件資料，除了缺乏實驗所需的查詢問句外，亦缺乏評量檢索效益的標準答案，也就是不知道文件集合中真正相關的文件爲何，因此無法計算求全率。故本研究在取得檢索結果之相關候選文件之後，以人爲判斷檢索出的文件是否與查詢問句相關，再計算求準率做爲檢索效益評估的依據。以人爲判斷文件相關程度的方式一直受到爭議，因爲相關的判斷常因諸多因素產生很大的差異，相關判斷（Relevance Judgment）牽涉到四個因素：

- 判斷的情況：包括判斷的時間是否充裕、文件集合的種類和大小、文件集合中的排列順序，以及相關性的定義。
- 文件本身：包括文件集合的主題、內容、形式，以及內容是否被縮減過（如摘要）。
- 資訊需求的描述：即查詢問句的內容、詳細程度、表達方式等。
- 判斷者：包括判斷者的專業知識、經驗、態度等。

由於判斷的過程十分複雜，因此部分學者提出相關判斷不能只依靠個人，而必須集合一群人的力量，

才夠客觀。（註 13）

本研究以主題相關爲相關判斷之原則，且以人爲判斷評估文件與查詢問句是否相關，因此可能會因爲相關判斷者的背景知識與主觀價值而導致偏差及不一致，故採用多位相關判斷者對相同的檢索結果執行相關判斷的方式，希望能修正可能的缺失。表四說明參與相關判斷的人士的背景資料。

表四、相關判斷者背景說明（註 14）

角　色	背 景 說 明	人數
主題專長	具有與檢索主題相近之學科背景或專長	8
檢索專長	圖書資訊學研究所碩士班研究生（含研究者本人），具有一般性學科知識以及豐富之資訊檢索經驗。	2
一般使用者	圖書資訊學系大四學生以及圖書資訊學系畢業生。	3

查詢問句擴展實驗中，共有 12 個查詢問句以 13 種不同的形式來擴展查詢問句，每次的檢索結果擷取前 50 篇文章進行判斷，總計共有 7,800 篇文件。由於每一個查詢問句擴展結果皆有三位相關判斷者進行相關判斷，亦即每一篇文件將被判斷三次，因此判斷的總次數爲 23,400 次。

本研究採用的評估方法並非於實驗進行之前先建立標準答案，亦即實驗文件資料庫並無建立正確的相關文件集合，而是將檢索實驗後之檢索結果，以人工判斷的方式挑選出相關的文件，亦即實驗檢出的文件中，有多少文件是相關判斷者認爲相關之文件。相關判斷者對每篇文件需給予分數，以表示文件與查詢問句的相關程度，再依此計算各檢索結果的求準率。給分的標準爲：相關判斷者認爲相關的文件，分數爲檢索系統所計算的相關係數值 Sim(Q,D)，因爲相關係數反映出文件與查詢問句的相關程度；反之，判斷者認爲不相關的文件，分數一律爲 0。

對於某一查詢問句以某種形式擴展後之檢索結果的 50 篇文件，三位相關判斷者所認爲相關的文件不盡相同，可能有意見一致的部分，亦有不一致的部分。然而每位相關判斷者對於各檢索結果的檢索效益分數有相等的貢獻，故取三位相關判斷者之判斷結果的平均數，做爲各檢索結果之檢索效益分數。以下爲每次檢索結果求準率（P）的計算公式：

$$P = \frac{[R_1]+[R_2]+...[R_m]}{n \times m}$$

其中，n 表示進行判斷之相關候選文件數，m 表示相關判斷人數，R_m 表示每位相關判斷者給予每一篇相關候選文件分數的總和。

五、結果與分析

經由分析前述各項實驗結果，各種查詢問句擴展的方式的效能說明如下。

（一）字串比對方式與檢索效益分析

本研究進行查詢問句擴展實驗，無論是以同義詞典或是以索引典爲擴展詞彙來源，皆必須以搜尋的方式找出與原始查詢問句相同之詞彙，再擷取與該詞彙相關聯之其他詞彙加入原查詢問句以進行擴展。本實驗以字串比對的方式於同義詞典與索引典中搜尋詞彙，又考量不同的字串比對方式對查詢問句擴展之檢索效益可能的影響，故分別以「部分字串比對」及「完全字串比對」兩種不同的方式進行同義詞典擴展與索引典擴展兩組查詢問句擴展實驗：

部分字串比對：部分字串比對的原則爲被搜尋詞彙之部分字串與原詞彙相同者，即被視爲比對成功。舉例說明，原詞彙爲「塑膠」，以部分字串比對方式則「工程塑膠」、「塑膠薄膜」、「塑膠」皆視爲比對成功。

完全字串比對：完全字串比對的方式則是指被搜尋詞彙與原詞彙必須完全相同才視之爲比對成功，以相同的例子來看，只有「塑膠」才會被擷取。

由檢索結果大致趨勢分析可知，同義詞典擴展及索引典擴展之實驗，完全字串比對較部分字串比對有較高的檢索效益。

1. 同義詞典擴展以完全字串比對表現較佳

   由於同義詞典之詞彙的擷取與詞彙群的產生，皆由實驗文件資料庫之詞彙共現資訊所得，自動建構過程中無人爲的介入，故一方面同一詞彙群的主題概念可能較爲分散，一方面以斷詞程式所擷取之詞彙，所提供的專門詞彙比較少，因此若以部分字串比對擴展，可能將更多不同主題的詞彙

加入原始查詢問句，反而混淆了原本的檢索主題，使檢索結果更難符合使用者需求。完全字串比對的方式則依據使用者提供的詞彙搜尋相同詞彙，擴展後的查詢問句則包含該詞彙在實驗文件資料庫中共現之詞彙，不會包含過多主題相差太遠的雜訊，所以相較之下，同義詞典擴展以完全字串比對方式有較佳的檢索效益。

但是有少數查詢問句無論以第二層、第三層或第四層詞彙群擴展皆是部分字串比對的表現較好，其受原始查詢問句之檢索詞彙詞義涵蓋面以及同義詞典詞彙分群品質所影響，亦即，由於該原始查詢問句的檢索詞彙字串較短或詞義較廣，如「耐震」以部分字串比對方式可找出「耐震能力」、「耐震性」、「耐震特性」等詞彙，若以完全字串比對則無法取得這些詞彙。從同義詞典詞彙分群品質角度來看，由於同義詞典建構過程中，受停用字表、斷詞、相關係數計算等因素之影響，使得詞彙群的品質並不一致，有些詞彙群會包含較多雜訊，因此以部分字串比對擴展之檢索結果爲佳者，是由於比對成功之詞彙所屬詞彙群品質較好，包含相關且專指的詞彙；反觀以完全字串比對擴展之檢索結果爲佳者，是由於擴展的詞彙群品質不佳且包含太多雜訊，若以部分字串比對擴展則加入比完全字串比對擴展更大量無關的詞彙，反而更混淆檢索主題。

2. 索引典擴展以完全字串比對表現較佳

以索引典擴展，字串比對的對象爲索引典中所有詞彙，故以部分字串比對的方式搜尋的步驟爲：

- 初步先擷取比對成功之所有詞彙。
- 擷取與比對成功詞彙之相關詞彙。

第一步驟比對成功的詞彙，通常彼此皆具備某種詞彙關係，也就是在第一步驟時就可能把第二步驟才連結到的相關詞彙都擷取出來，這是因爲這些詞彙的部分字串是與原詞彙相同的。

索引典是由人工建構，整體結構與詞彙品質皆經過嚴謹的控制，收錄了該學科領域通用及專精的詞彙，而使用者提供的原始查詢問句，以較簡單且較具概括性概念的詞彙爲主，因此若以部分字串比對的方式搜尋，通常比對成功的機會很高，但會有許多詞彙主題概念與原詞彙不相關的情形。如本研究使用之第 10 題查詢問句中的原始檢索詞彙「對流」，以部分字串比對的方式可於索引典中搜尋「對流熱傳遞」、「自然對流」、「混合對流」等與檢索主題十分相關的詞彙，但亦搜尋出「對流雲」、「積雲對流」、「對流層」等雖部分字串相同但主題相差極大的詞彙，也因此擴展出如「大氣圈」、「大氣傳播」、「雲」等更多與主題不相關的詞彙，因此以部分字串比對的方式反而加入更多雜訊，而且也因爲索引典詞彙之豐富，令錯誤擴展的詞彙數量十分龐大，即使其中包括與檢索主題相關的詞彙，也因爲包含於爲數龐大的查詢問句中使得該詞彙的重要性無法突顯。故相較之下，完全字串比對的檢索效益較好。

（二）同義詞典擴展

本實驗針對 12 題原始查詢問句以同義詞典做了 6 種不同形式的擴展，在不同擴展模組檢索結果的相對比較之下，大致以同義詞典第二層的詞彙群擴展表現最佳，且檢索效益大都比基礎檢索高：以完全字串比對方式，第二層表現最好，於 12 個查詢問句中有 7 題的檢索效益最高，其次是第四層，最後爲第三層；以部分字串比對方式，亦是第二層表現最好，於 12 個查詢問句中有 8 題的檢索效益最高，其次是第三層，最後爲第四層。

無論完全字串或部分字串比對，檢索效益較高者皆爲詞彙群內詞彙數量最少的第二層，亦即詞彙群再次合併令詞彙群包含過多雜訊反而分散詞彙群之主題概念。由於同義詞典的建構方式爲利用詞彙於文件資料庫的共現資訊，計算詞彙間相關係數以聚引相關詞彙，因此在每次的詞彙群合併過程，詞彙群之間的相關係數皆比前一次合併所計算之相關係數有大幅度的下降，此即反映詞彙群內詞彙之間的相關程度隨著詞彙群規模的擴大而愈形微薄，故以詞彙間相關程度較強的第二層擴展應比以相關程度較弱的第三層擴展之檢索效益更好，以此類推，同義詞典擴展之效益評比結果，應爲第二層>第三層>第四層。值得注意的是完全字串比對第四層的表現反而比第三層爲佳，此情形與擴展詞彙數量與檢索效益成反比的假設不符。

以同義詞典擴展之檢索效益，可能與所擴展的詞彙數量有密切的關係，故分別調查各查詢問句以三個層次擴展中檢索效益最佳者之擴展詞彙數量，但發覺擴展的詞彙數量較少，檢索效益不一定較高；同樣地，擴展的詞彙數量較多，檢索效益不一定比擴展數量少者差。此由於受實驗文件資料庫本身文件類別的分佈情形影響，文件資料庫有些主題的文件數量較多，可擷取之詞彙共現資訊較豐富，使得相關主題的詞彙群內詞彙數量較多，擴展後的檢索詞彙數量也因此較多，又因爲文件豐富使得檢出相關文件的機率較大，檢索效益即較高，故本實驗擴展詞彙數量與檢索效益的關係，受文件資料庫本身的特性所影響。表五是以同義詞典擴展之檢索結果，使用求準率比較檢索效益。

（三）索引典擴展

我們使用索引典做了 10 種不同形式的查詢問句擴展。在不同擴展模組檢索結果的相對比較之下，以整體大致上的趨勢分析，部分字串比對以廣義詞的擴展表現最佳，於 12 個查詢問句中有 9 題的檢索效益最高，狹義詞、等同詞及關聯詞擴展之檢索效益相同，較廣義詞略差，最後爲聯合擴展。

完全字串比對以聯合擴展之檢索效益較高，於 12 個查詢問句中有 5 題的檢索效益最高，其次依序爲關聯詞、廣義詞、狹義詞，最後爲等同詞。以完全字串比對的方式擴展則會出現無擴展的情形，其中以等同詞較多無擴展的結果，狹義詞則皆有擴展。

部分字串比對的結果，出現多種詞彙關係擴展的檢索效益相同的情形。12 題查詢問句中，有 7 題四種以上的詞彙關係擴展檢索效益完全相同，此可能因爲以部分字串比對的方式擴展的詞彙數量較爲龐大，且如之前所提，於擴展的第一步驟就把第二步驟以詞彙關係擷取出來的詞彙全都先取出，亦即部分字串比對的擴展不只擴展某單一詞彙關係，而是已包含各種不同詞彙關係之詞彙，使得擴展後的結果近似於聯合擴展。再分別觀察各詞彙關係所擴展的檢索詞彙，最少有 1 個詞彙、最多有 51 個詞彙與其他詞彙關係的擴展結果不同，占擴展後檢索詞彙數量的 5%~50%。雖然不同詞彙關係擴展後的詞彙組合差異很大，但因爲這些與其他詞彙關係擴展結果不同的詞彙，並不存在於文件資料庫，在檢索時文件與查詢問句相關係數的計算上，也被忽略不予處理，所以這些不同的詞彙即使數量多卻沒有影響檢索結果的能力，因此會發生檢出的文件相同、文件相關係數相同、檢出文件排序相同等情形，使得檢索效益亦相同。表六是索引典各詞彙關係擴展查詢問句之檢索結果。

（四）同義詞典擴展，索引典加權

以同義詞典擴展再利用索引典加權之實驗，則參考第一階段以同義詞典不同層次擴展查詢問句之實驗結果，選擇同義詞典擴展實驗中表現最佳之第二層詞彙群，並以檢索效益較高的完全字串比對方式進行第四、五組之查詢問句擴展實驗，期能得知利用由人工建構之索引典的詞彙資源，調整以同義詞典擴展後檢索詞彙的詞彙權重，是否能提升檢索效益。

以完全字串比對方式利用同義詞典第二層詞彙群，將原始查詢問句擴展後，每個檢索詞彙之權重皆相等，設定權重爲 1，將擴展後之查詢問句以完全字串比對方式在索引典搜尋詞彙，比對成功之詞彙即將其權重調整爲 2，再以加權後之查詢問句進行檢索。

同義詞典擴展再利用索引典加權之實驗，目的在探討以人工建構之索引典是否能降低同義詞典所包含的雜訊對檢索的不良影響，故以第四組實驗結果與第二組實驗中第二層擴展的實驗結果相互比較分析：12 題查詢問句其中 8 題加權之後的檢索效益有所提升。

與基礎檢索比較，則 12 題查詢問句中，有 8 題高於基礎檢索之檢索效益。相較於以同義詞典第二層詞彙群擴展之實驗結果，12 題中只有 5 題的檢索效益比基礎檢索爲佳，經由索引典的加權、調整檢索詞彙權重之後，即增加爲有 8 題的檢索效益高於基礎檢索。由該角度分析得知，利用索引典之詞彙資源修正以同義詞典擴展可能的雜訊，對檢索效益有正面的提升。表七是以同義詞典擴展、再以索引典加權之檢索結果，以求準率比較檢索效益。

（五）同義詞典擴展，索引典加權並擴展

以同義詞典擴展，再利用索引典加權並擴展的實驗，亦參考第一階段索引典各種詞彙關係擴展查詢問句的實驗結果，選擇表現最佳者進行第五組實驗。第一階段的實驗結果，選擇較好的完全字串比對方式。

雖然以聯合擴展的檢索效益較佳，但以廣義詞及關聯詞擴展的檢索效益與聯合擴展相較之下，並無顯著差距，故第五組的查詢問句擴展實驗設計，分別以索引典的廣義詞、關聯詞與聯合擴展進行第二次擴展。

同義詞典擴展，再利用索引典加權並擴展的實驗結果，與基礎檢索及只以同義詞典第二層詞彙群擴展之檢索結果比較，檢索效益略有提升。而本組實驗在不同擴展模組檢索結果的相對比較之下，以聯合擴展的表現最佳，其次爲以廣義詞擴展，最差爲以關聯詞擴展，但三者亦只有些微的差距。將本組實驗結果與第四組實驗之結果相互比較，整體而言，無論是以廣義詞、關聯詞或是聯合等不同詞彙關係進行第二次的擴展，檢索效益反而降低。由於第二次擴展的詞彙在索引典的架構中雖與原詞彙雖具有某種詞彙關係，但可能與檢索主題不相關，若加入許多不相關的詞彙且詞彙權重爲 2，則該雜訊的干擾程度更高，使得檢索效益比第二次擴展前更差。表八是以同義詞典擴展，以索引典加權與擴展之檢索結果。

六、結論與建議

（一）結論

目前英文資訊檢索研究有較完備的實驗文件資料庫可做爲研究所需的實驗對象，包括一定數量的文件集合、查詢問句以及標準答案，但中文資訊檢索的環境則缺少標準的實驗文件資料。目前雖有陳光華與江玉婷建構的中文資訊檢索測試集（註 15），但由於索引典之查詢問句擴展爲本研究重要主題之一，受限於索引典的類型，無法選用該中文資訊檢索測試集。查詢問句擴展之實驗，係針對具有較多文件數量的文件類別設計查詢問句，但由於實驗文件資料庫文件類別分佈過於分散，且涵蓋範圍廣大，即使某一類別的文件數量到達二三百篇，相關於同一主題的文件數量仍嫌太少。故查詢問句的限制較大，導致查詢問句數量略嫌不足，只能針對個別情形探討可能因素，而無法以足夠的樣本進行實驗結果的整體性趨勢分析。

本研究以字串比對的方式於同義詞典與索引典搜尋詞彙，故分別以「部分字串比對」及「完全字串比對」兩種不同的方式進行同義詞典擴展與索引典擴展兩組查詢問句擴展實驗，由檢索結果大致趨勢分析可得，同義詞典擴展以及索引典擴展之實驗，完全字串比對較部分字串比對有較高的檢索效益。

以同義詞典擴展之檢索效益，無論完全字串比對或是部分字串比對，表現最佳者皆爲詞彙群內詞彙數量最少的第二層。詞彙群內詞彙數量愈多，擴展的詞彙愈多，則雜訊愈多，使得檢索效益降低，故同義詞典擴展之效益評比結果，應爲第二層>第三層>第四層。以同義詞典擴展之實驗結果顯示，擴展後的詞彙數量約 8~10 個詞彙有較好的檢索結果。

以索引典各詞彙關係擴展查詢問句，各種詞彙關係擴展之檢索效益並沒有顯著的差異。但以整體的趨勢分析，部分字串比對以廣義詞的擴展表現最佳，完全字串比對以聯合擴展表現最佳。以狹義詞擴展的檢索結果，與其他詞彙關係比較之下都較差，因狹義詞的主題概念更分歧，反而模糊了擴展後查詢問句的檢索主題。使用部分字串比對方式擴展後查詢問句的詞彙數量皆十分龐大。實驗結果顯示，擴展後詞彙數量過多，其檢索效益有明顯地降低情形。

以同義詞典擴展再利用索引典加權與擴展之實驗結果亦顯示，索引典加權之後的查詢問句，檢索效益不但高於基礎檢索，且較以同義詞典擴展但未加權的檢索結果更好，故利用索引典之詞彙修正以同義詞典擴展可能的雜訊對檢索效益有正面的提升。但若加權之後，再以索引典做第二次的擴展，雖比基礎檢索或是未加權擴展的第二層詞彙群檢索結果佳，但與只利用索引典加權之檢索結果相比較，檢索效益反而降低。

（二）建議

中文資訊檢索的研究涉及許多議題，包括斷詞、索引、字串比對、檢索、評估方法等，而關於以查詢問句擴展提升檢索效益的研究，除了擴展機制與形式的設計之外，有更多複雜課題必須解決，本研究僅針對實驗中未能克服的相關事項提供以下建議。

各種詞彙資源需要建立適當的整合模式才能發揮輔助檢索的功能，本研究已初步探討詞彙資源整合的形式以及詞彙擴展的模式，若能進一步深入研究，尋求最佳的整合方式，可做爲未來擴展查詢問句的重要參考。一般整合不同詞彙資源，通常先建構分屬不同資源的詞彙間互動與連結的方式，可能的方式有聯合

（Union）、連結（Chaining）與對映（Mapping）。聯合爲將原始查詢問句分別在不同資源進行查詢問句擴展後，再整合所有出現於各組合中的所有詞彙，形成擴展後的查詢問句。聯合有兩種方法可結合不同資源：加權與不加權。加權指各組合皆附帶權重，兩組詞彙交錯者其權重相加。連結策略則依次使用不同的資源，即原始查詢問句先利用一資源擴展，將已進行一次擴展的查詢問句利用另一資源再次進行擴展。對映則將原始查詢問句分別利用不同資源進行擴展，再擷取二組擴展後之查詢問句重疊的部分，以完成查詢問句的擴展。（註 16） 以上述三種方式爲基礎，還能發展出更多不同的整合模式，值得進一步探討。

資訊科技的發展與目前網際網路成長的速度，使得資料量成長更爲快速，資訊的需求亦大爲提升，因此各種資訊檢索系統的使用情況更爲頻繁，使用者對於資訊檢索系統介面與效率的要求亦更多元。就中文資訊檢索研究而言，查詢問句擴展是增進檢索效益，滿足使用者需求的重要議題，且仍有極大的發展空間，因此，未來中文資訊檢索的研究，期能在標準的評估環境之下，致力於參考既有相關研究的理論與技術，並考量中文詞彙特殊的語文性質，以利用查詢問句擴展機制發展高效率的中文檢索系統。

表五、以同義詞典擴展之檢索結果

	基礎檢索	部分字串比對			完全字串比對		
		第一層	第二層	第三層	第一層	第二層	第三層
QUERY 1	0.0801	0.1210	0.1025	0.0636	0.0791	0.0616	0.0623
QUERY 2	0.0931	0.0566	0.0566	0.0544	0.1081	0.1100	0.0470
QUERY 3	0.0864	0.0962	0.0914	0.0914	0.0957	0.0927	0.0942
QUERY 4	0.0345	0.0566	0.0521	0.0565	0.0605	0.0573	0.0590
QUERY 5	0.0399	0.0345	0.0336	0.0309	0.0413	0.0409	0.0419
QUERY 6	0.0456	0.0554	0.0243	0.0341	0.0450	0.0422	0.0323
QUERY 7	0.0511	0.0233	0.0408	0.0317	0.0448	0.0432	0.0484
QUERY 8	0.0530	0.0558	0.0574	0.0490	0.0501	0.0487	0.0470
QUERY 9	0.0512	0.0433	0.0489	0.0483	0.0471	0.0475	0.0483
QUERY 10	0.0610	0.0622	0.0761	0.0947	0.0946	0.0896	0.0871
QUERY 11	0.0806	0.1133	0.1097	0.1093	0.0692	0.0688	0.0701
QUERY 12	0.1361	0.0833	0.0822	0.0761	0.0816	0.0775	0.0761

表六、以索引典擴展之檢索結果

	基礎檢索	部分字串比對					完全字串比對				
		狹義詞	等同詞	關聯詞	廣義詞	聯合	狹義詞	等同詞	關聯詞	廣義詞	聯合
QUERY 1	0.0801	0.0583	0.0583	0.0414	0.1056	0.0583	0.0529	無擴展	0.0917	0.0889	0.0635
QUERY 2	0.0931	0.0541	0.0541	0.0541	0.0541	0.0125	0.0969	0.0991	0.0988	0.1070	0.0979
QUERY 3	0.0864	0.0944	0.0944	0.0944	0.0944	0.0944	0.0930	無擴展	0.0942	0.0861	0.0984
QUERY 4	0.0345	0.0338	0.0338	0.0439	0.0450	0.0424	0.0460	無擴展	0.0479	0.0462	0.0387
QUERY 5	0.0399	0.0407	0.0415	0.0415	0.0407	0.0070	0.0415	0.0372	0.0357	0.0391	0.0416
QUERY 6	0.0456	0.0460	0.0460	0.0460	0.0460	0.0447	0.0479	0.0455	無擴展	0.0469	0.0483
QUERY 7	0.0511	0.0451	0.0455	0.0455	0.0451	0.0461	0.0386	0.0448	0.0436	無擴展	0.0386
QUERY 8	0.0530	0.0408	0.0408	0.0408	0.0408	0.0408	0.0453	0.0422	0.0452	0.0447	0.0455
QUERY 9	0.0512	0.0386	0.0386	0.0386	0.0386	0.0235	0.0396	0.0345	0.0478	0.0482	0.0429
QUERY 10	0.0610	0.1056	0.1056	0.1056	0.1056	0.1029	0.1006	0.0936	0.0839	0.0778	0.0998
QUERY 11	0.0806	0.1127	0.1107	0.1107	0.1126	0.1107	0.0880	無擴展	無擴展	0.0806	0.0884
QUERY 12	0.1361	0.0822	0.0822	0.0822	0.0822	0.0822	0.0837	0.0841	0.0902	0.0868	0.0859

表七、以同義詞典擴展再以索引典加權之檢索結果

	基礎檢索	同義詞典第二層擴展 （完全字串比對）	以同義詞典擴展 再以索引典加權
QUERY 1	0.0801	0.0791	0.0514
QUERY 2	0.0931	0.1081	0.0877
QUERY 3	0.0864	0.0957	0.0773
QUERY 4	0.0345	0.0605	0.0559
QUERY 5	0.0399	0.0413	0.0609
QUERY 6	0.0456	0.0450	0.1001
QUERY 7	0.0511	0.0448	0.0986
QUERY 8	0.0530	0.0501	0.1124
QUERY 9	0.0512	0.0471	0.1107
QUERY 10	0.0610	0.0946	0.1067
QUERY 11	0.0806	0.0692	0.0942
QUERY 12	0.1361	0.0816	0.0893

表八、同義詞典擴展，索引典加權並擴展之檢索結果

	基礎檢索	以同義詞典擴展 再以索引典加權	第二次擴展		
			關聯詞	廣義詞	聯合
QUERY 1	0.0801	0.0514	0.0583	0.0568	0.0470
QUERY 2	0.0931	0.0877	0.0860	0.0845	0.0845
QUERY 3	0.0864	0.0773	0.0690	0.0733	0.0741
QUERY 4	0.0345	0.0559	0.0526	0.0527	0.0518
QUERY 5	0.0399	0.0609	0.0574	0.0599	0.0616
QUERY 6	0.0456	0.1001	無擴展	0.1028	0.1053
QUERY 7	0.0511	0.0986	0.0951	0.0832	0.0898
QUERY 8	0.0530	0.1124	0.0997	0.1053	0.1178
QUERY 9	0.0512	0.1107	0.1058	0.1097	0.1097
QUERY 10	0.0610	0.1067	0.0994	0.1024	0.1135
QUERY 11	0.0806	0.0942	無擴展	0.0826	0.0942
QUERY 12	0.1361	0.0893	0.0941	0.0934	0.0898

註釋

註 1： 李連揮，「索引典與索引方法」，圖書館學與資訊科學 3：2 (民國 66 年 10 月)：頁 48。

註 2： 陳佳君，「檢索詞彙來源與檢索詞彙效益之研究」，國立台灣大學圖書館學研究所碩士論文，民國 84 年，頁 2。

註 3： 同註 2，頁 2。

註 4： 黃慕萱，「線上索引典顯示格式之設計探討」，中國圖書館學會會報 第 53 期 (民國 83 年 12 月)：頁 131。

註 5： 陳光華，莊雅棻。「應用於資訊檢索的中文同義詞之建構」，中國圖書館學會會報（出版中）。

註 6： Susan Gauch and John B. Smith, “An Expert System for Automatic Query Reformation,” Journal of the American Society for Information Science 44, no.3 (1993)：133.

註 7： Yonggang Qiu and H.P. Frei, “Concept Based Query Expansion,” In Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Pittsburgh, PA, USA, Jun 27-Jul 1 1993 , 160-161.

註 8： Mandar Mitra, Amit Singhal and Chris Buckley, “Improving Automatic Query

Expansion," In Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Melbourne, Australia, August 24-28, 1998, 206-207.

註 9： Magennis, Mark. "Expert Rule-based Query Expansion."(1995) <http://www.dcs.gla.ac.uk/ir/publications/papers/Prostscript/magennis95.ps.gz> (Accessed: 30 May 2000)。

註 10： Hsinchun Chen, Tobun D. Ng, Joanne Martinez, and Bruce R. Schatz, "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System." Journal of the American Society for Information Science 48, no.1 (1997): 17-31.

註 11： D. Harman, "Towards Interactive Query Expansion," In Proceedings of the 11th Annual International ACM-SIGIR Conference on Research & Development in Information Retrieval, Grenoble, France, 1988, 322-323.

註 12： 同註 9。

註 13： 吳忻萍，「以隱藏語意索引爲基礎之中文資訊檢索」，國立台灣大學資訊管理學研究所，碩士論文，民國 86 年，頁 24。

註 14： 陳光華，江玉婷。「中文資訊檢索測試集之設計與製作」。資訊傳播與圖書館學 6：3（民國 89 年 3 月）：頁 61-80。

註 15： 同註 14。

註 16： Richard C. Bodner and Fei Song, "Knowledge-Based Approaches to Query Expansion in Information Retrieval" In Advances in Artificial Intelligence (New York: Springer, 1996): 151-152.