國立交通大學

資訊學院 數位圖書資訊學程

碩士論文

以成語涵義為基礎之中文成語檢索系統

Chinese Idiom Information Retrieval System Based on the Idiom Semantics

研 究 生：張正霖
指導教授：黃明居 博士

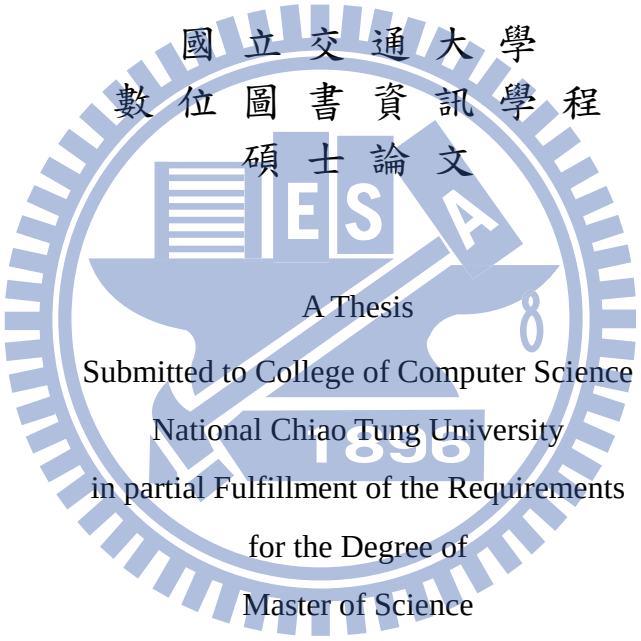
中華民國九十八年九月

以成語涵義為基礎之中文成語檢索系統

Chinese Idiom Information Retrieval System Based on the Idiom Semantics

研 究 生：張正霖　　Student：Chen-Lin Chang

指導教授：黃明居 博士　　Advisor：Dr. Ming-Jiu Hwang



國 立 交 通 大 學

數 位 圖 書 資 訊 學 程

碩 士 論 文

A Thesis

Submitted to College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Digital Library

September 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年九月

以成語涵意為基礎之中文成語檢索系統

研 究 生：張正霖　　指導教授：黃明居　博士

國立交通大學電　資訊學院　　數位圖書資訊學程碩士班

摘要

目前成語檢索系統的查詢功能主要包括：單一關鍵詞釋義查詢、字詞查詢、類別查詢、首字部首查詢、以及首字拼音查詢。但使用者查詢成語時，往往僅知其「成語涵義」，不知成語的字詞，使得使用者較無法查得所需成語。本研究建置一套以成語涵義為基礎的成語檢索系統-MIRS(Meaning of Idiom Retrieval System)來解決上述成語檢索系統之問題。MIRS 包括：成語資料前置處理，查詢問句處理，檢索處理，以及結果顯示等四大模組。使用者輸入簡單的口語化查詢問句，利用擴展查詢與增加關鍵詞權重方法，即能更精準找出成語。系統提供查詢結果的關鍵詞統計與分類，讓使用者透過層面分類查詢(Facet Query)與修訂查詢(Revised Query)功能亦可有效找到成語，另一方面，本系統引進 Web 2.0 概念，讓使用者提供同義詞和成語釋義的建議資料，進而更提升系統查詢效益。由系統的評估發現，本系統所提供的功能，讓使用者選擇最適合的檢索方法，不但查詢功能更友善，而且結果更精準。

關鍵字：成語檢索、資訊擷取、關鍵詞權重、查詢擴展、層面分類、修訂查詢

Chinese Idiom Information Retrieval System Based on the Idiom Semantics

Student：Chen-Lin Chang　　Advisor：Dr. Ming-Jiu Hwang

Degree Program of Computer Science

National Chiao Tung University

Abstract

At present, the search functions of Chinese idiom retrieval systems include single keyword search, searching for character, searching for category, and searching for radical, pinyin, stroke number in first character of idiom. Users are requested to input these query items into idiom retrieval systems for search. However, users always remember the meaning of Chinese idiom but not idiom text when they want to search. The aim of this study was to construct a Chinese idiom information retrieval system base on the meaning of idiom (Meaning of Idiom Retrieval System, MIRS), and to solve these questions described as above. MIRS contains four models which are pre-processing of idiom content, query processing, and retrieval processing, and exhibition of query outcome. User inputted oral questions with a simple query, MIRS can more accurately find the idioms by handling query extension and increasing keywords weight. System also can effectively find the idiom by counting and classifying keywords of searched results, and then working "Facet Query" and "Revised Query". In addition, MIRS also builds the conception of Web 2.0 that users can provide synonyms and recommend meanings information, in order to increase the efficiency of MIRS. According to the evaluation of MIRS, we found that MIRS is friendly to use and gives users one choice of adaptable retrieval systems to acquire precise queried information.

Keyword: idiom search, information retrieval, term weight, query extension, facet query,

revised query

誌　　謝

感謝兩年來指導教授黃明居老師在各方面的悉心教導與照顧，讓我們能一窺研究領域中學問的奧秘；另外，還要感謝擔任口試委員的柯皓仁老師和孫春在老師，您對於論文中未周全的部份給予許多的建議與指導，使論文能更嚴謹，由衷的感謝。

此外還要感謝峰智學長、有凱同學、文雄同學，以及學弟妹，都曾在學術上給我建議，使得我的論文得以順利的完成。

最後，還要感謝我親愛的家人與朋友長久以來的支持與鼓勵，讓我在學習的道路上無後顧之憂，使我能專心致力於研究，並得以順利完成學業。

目　　錄

圖 目 錄

表 目 錄

一、 緒論

本章描述本研究的動機與目的，以及希望解決的問題。1.1 節說明研究動機與本研究希望達成的目的；1.2 節為研究內容，以資訊檢索為基礎，進一步拓展應用到成語檢索，需要解決的幾個關鍵點；1.3 節為研究的範圍與限制；1.4 節則說明本論文各章節的內容架構。

1.1. 研究動機與目的

隨著全球資訊網路(World Wide Web, WWW)的興起，已有不少電腦輔助學習系統出現在網際網路中，例如，教育部成語典[1]、漢典[2]，這些系統提供成語學習者方便且快速的查詢功能。

目前成語檢索系統功能主要分為兩種模式，第一種是關鍵字比對，將使用者所輸入的關鍵字與成語字詞、釋義進行比對，例如教育部成語典，如圖 1-1 所示，包含成語字詞檢索；第二種是以分類為基礎的查詢方式，常見有類別檢索、首字筆劃查詢、首字部首查詢及首字拼音查詢等，例如漢典，如圖 1-2 所示。但使用者查詢成語時，往往僅知其「成語涵義」，不知成語的字詞，目前的系統查詢功能較無法讓使用者清楚描述其資訊需求，也就無法快速地查到所需成語。如何符合“僅知「成語涵義」，不知成語字詞，而可找尋適當成語”之使用者需求，及提供友善的查詢介面以協助使用者找到所需成語，是一個值得深究的課題，亦是本研究主要的動機所在。

[1] 教育部成語典 http://dict.idioms.moe.edu.tw/
[2] 漢典 http://www.zdic.net/

圖 1-1 教育部成語典查詢功能

圖 1-2 漢典查詢功能

本研究目的將建置提供一個僅知成語意涵的檢索系統，使用者只須輸入簡單的口語化問句或關鍵字來描述想要查詢的成語，即可精準地查得所需成語，除此之外，本研究目的包括：

1. 深入探討資訊檢索相關研究，以及整合現有技術建置一套以成語涵義為基礎的成語檢索系統。

2. 提出「成語涵義」中文成語檢索系統之模組架構，使系統效能最佳。

3. 結合中研院CKIP之中文斷詞系統，提出一套適合中文成語檢索之查詢擴展模式，使系統之查詢更加精準。

4. 設計適合國人使用之檢索機制與使用者介面，讓使用者能夠快速找尋所需之成語資訊。

系統針對成語釋義部份，利用現行關鍵字檢索技術(Indexing)，製作可供快速比對(matching)的索引資料，其後使用者輸入描述查詢語句，經由中央研究院中文詞知識庫小組(Chinese Knowledge Information Processing Group, CKIP)[3]所研發之中文斷詞系統(包含未知詞擷取與標記)[4]斷詞，將關鍵詞進行同義詞查詢擴展(Query Extension)，依照詞性給與權重，便可和索引資料進行比對計算文件得分供排序使用，接著將查詢結果進行關鍵詞統計與分類，讓使用者可以透過層面分類查詢(Facet Query)與修訂查詢(Revised Query)快速找出所需成語資料，達成資訊檢索的目的。

1.2. 研究內容

1.2.1. 資訊檢索

在面對大量雜亂無章的資料時，一個可以幫助使用者快速取得所需要資訊方式，就是資訊檢索，簡單的說，就是將使用者的需求資訊與已經儲存的資訊進行比對，使檢索結果能夠滿足使用者的資訊需求。由此定義來看，若資訊是指大量的文件集合，那麼文件檢索即是依據使用者定義之規範，由大量文件集合中篩選並輸出使用者所需的文件，而整個系統中則包含使用者、資訊以及檢索系統三個項目。其一般模型如圖 1-3 所示[1]:

此圖以文字資訊為例，說明整個資訊檢索過程，左邊為文件資訊的產生，先把大量的文件做處理並形成集合，然後用適當的方式加以組織，將文字資訊集合轉化為對應的代表特徵；另一邊為具有特定資訊需求目的的使用者，將其資訊需求以適當方式表達，然後轉換為查詢。接著將文字資訊的代表特徵與使用者之查詢進行比對，找出使用者所需資訊，計算並評估該檢索結果，再回饋給使用者或檢索系統，使其能分別針對文字資訊的代表特徵或使用者之查詢加以修正。

[3] 中文詞知識庫小組(CKIP) http://rocling.iis.sinica.edu.tw/CKIP/
[4] 中文斷詞系統 http://ckipsvr.iis.sinica.edu.tw/

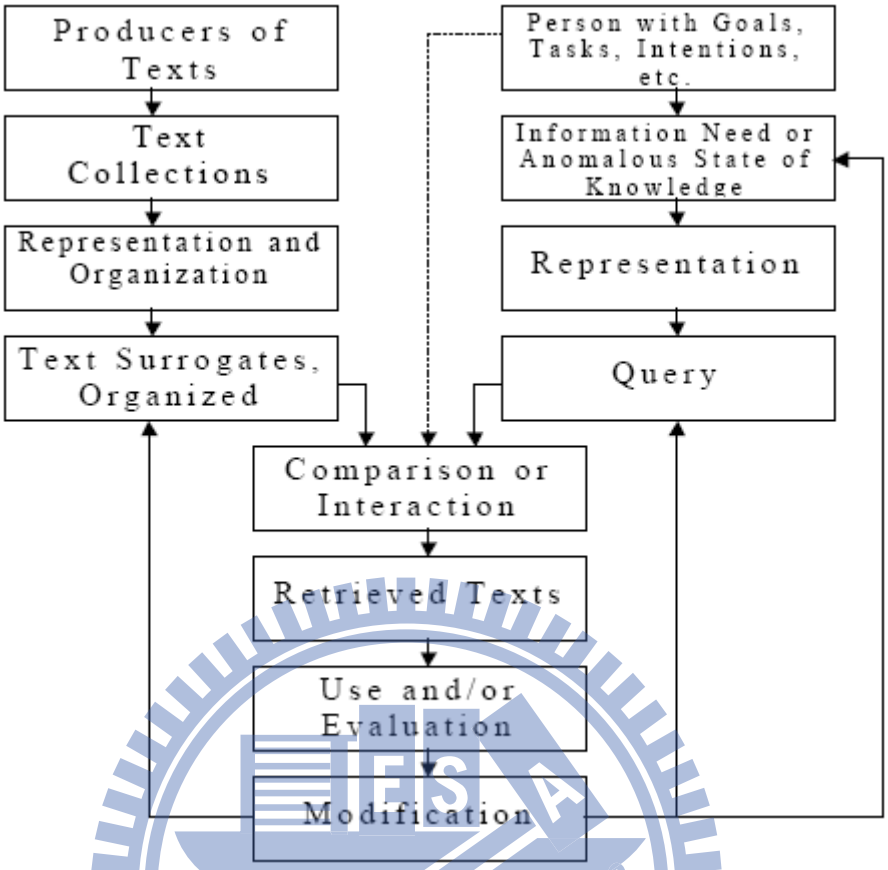


圖 1-3 資訊檢索一般化模型

1.2.2. 文件組織

如何將資訊以適當的形式表示，以利資訊檢索的應用，是所有資訊檢索系統都必須面對的問題。以本研究主題成語檢索系統為例，如何將成語資訊轉化為適當的代表形式，是一個重要的研究課題，其中一個被廣泛應用的方法就是所謂的索引(Indexing)，將資訊以適當的形式表示與組織，其所含的意義為「分析資訊內容、決定資訊特徵、並以特徵形式代表資訊的整個過程」[2]。

影響索引效果的主要因素，包括索引的詳盡性(Indexing Exhaustiveness)與索引詞的明確性(Term Specificity)。所謂的索引的詳盡性是指索引能夠反應某一文件內容主題的詳盡程度，索引愈詳盡則使用愈多的索引詞彙，以描述文件內容主要及次要主題；索引詞的明確性則是反應索引詞的廣義及狹義程度，當我們使用較廣義的索引詞，就比較不易辨識相關與不相關的文件差異[3]。隨著索引技術的進步，相關領域的革新除了由人工索引改變為自動索引外，另一個重要的突破是所使用的特徵不

再侷限於詞彙。例如可代表文件資訊特徵的 Signature files[4]。

本研究在成語文件檢索整個過程中的主要幾個關鍵點在於：

1. 如何組織所有的成語資訊，並萃取其代表性特徵？
2. 如何提供更友善的查詢功能介面，讓使用者能更完整描述出其資訊需求？
3. 如何將使用者的查詢語句與系統內儲存的代表特徵進行比較，以檢索出使用者真正需求的資訊？

1.3. 研究範圍與限制

受限於研究者可以取得資源有限，為了避免研究問題過於複雜，且保持研究問題單純性等考量，本研究限制如下：

1. 處理資訊形式限定為文字檢索。雖然資訊儲存形式有很多，例如：文字、影像、音訊、視訊等，不同的資訊形式需要不同的檢索方式，但目前成語資料主要是以文字資訊形式存在，所以僅處理目前最普遍、成熟的文字資料形式。
2. 處理語言形式限定為中文成語。各國語言基於字詞與文法差異，在進行前置處理(Pre-processing)時，需做不同的斷詞處理。本研究不討論各國語言間的特性，故選擇成語收錄所佔比例最高的「教育部成語典」中文資料作為實驗對象。
3. 需搭配斷詞系統使用。在產生文件代表特徵前，需對待分析文件進行關鍵資訊的萃取，以利後來資料分析，此過程仰賴對文字的處理技術，如：斷詞、建立詞幹、詞頻統計...等。關於文章斷詞處理，已是一個專門的研究，故將斷詞交由中研院CKIP斷詞系統執行。
4. 本研究的各種實作系統及實驗環境，皆以網際網路作業環境為例。

1.4. 研究流程

本研究進行流程，由於觀察到目前成語檢索系統所面臨資訊查找不易，在發現該問題之後，經由文獻探討與分析，找出所面臨的研究問題，確定研究目的與範圍

後，擬定系統功能需求。然後開始進行系統系統分析與設計，建置一資訊系統解決該問題，並與目前成語檢索系統功能進行比較，最後撰寫研究報告提出結論與貢獻。研究流程如圖 1-4 所示。

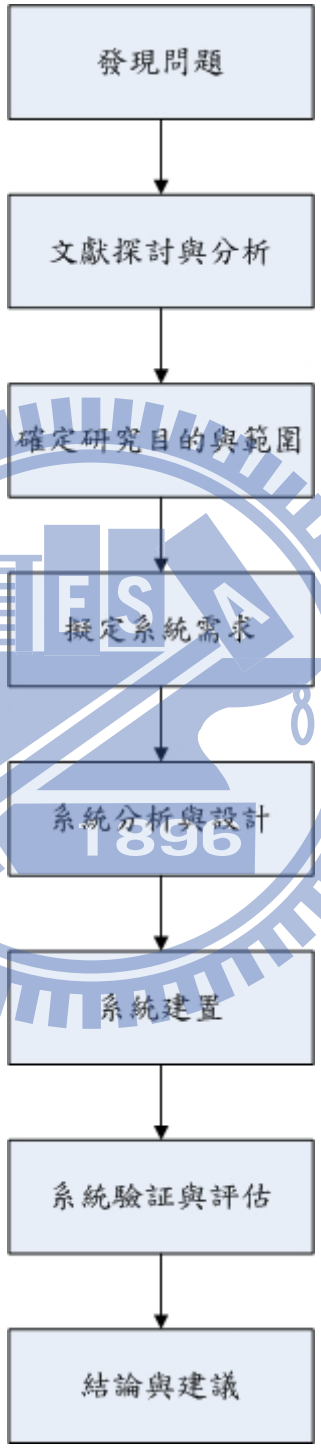


圖 1-4 研究流程圖

1.5. 論文大綱

本論文主要分為五個章節，整理架構流程如下詳述：

1. 第一章為緒論，透過研究背景與目的，說明目前成語檢索系統所遭遇的問題，並訂定研究限制以及研究流程。

2. 第二章為文獻探討與系統功能描述，首先介紹資訊檢索模式、資訊檢索系統的詞彙擴展，並描述系統功能需求和所使用的程式工具。

3. 第三章為系統架構分析與設計，詳細描述本系統架構、資料處理方法、檢索功能設計、以及如何解決所遭遇的問題。

4. 第四章為系統展示與評估，展示系統的功能，說明是否符合系統功能需求，並與目前最常使用的成語檢索系統功能進行比較，提出本系統優點。

5. 第五章為結論與未來研究方向，為本研究之總結，說明本研究之貢獻，並提出未來可繼續研究之議題。

二、 文獻回顧與系統功能需求

本章內容分為三個部份，3.1 節就研究所需，說明相關研究；3.2 描述本系統擬定之功能需求；3.3 節為建置本系統所用到的程式工具。

2.1. 文獻回顧

圖 2-1 為本研究文獻回顧整體分佈圖，共分為資訊檢索、檢索模式進展，與詞彙查詢擴展相關文獻等四大部份。

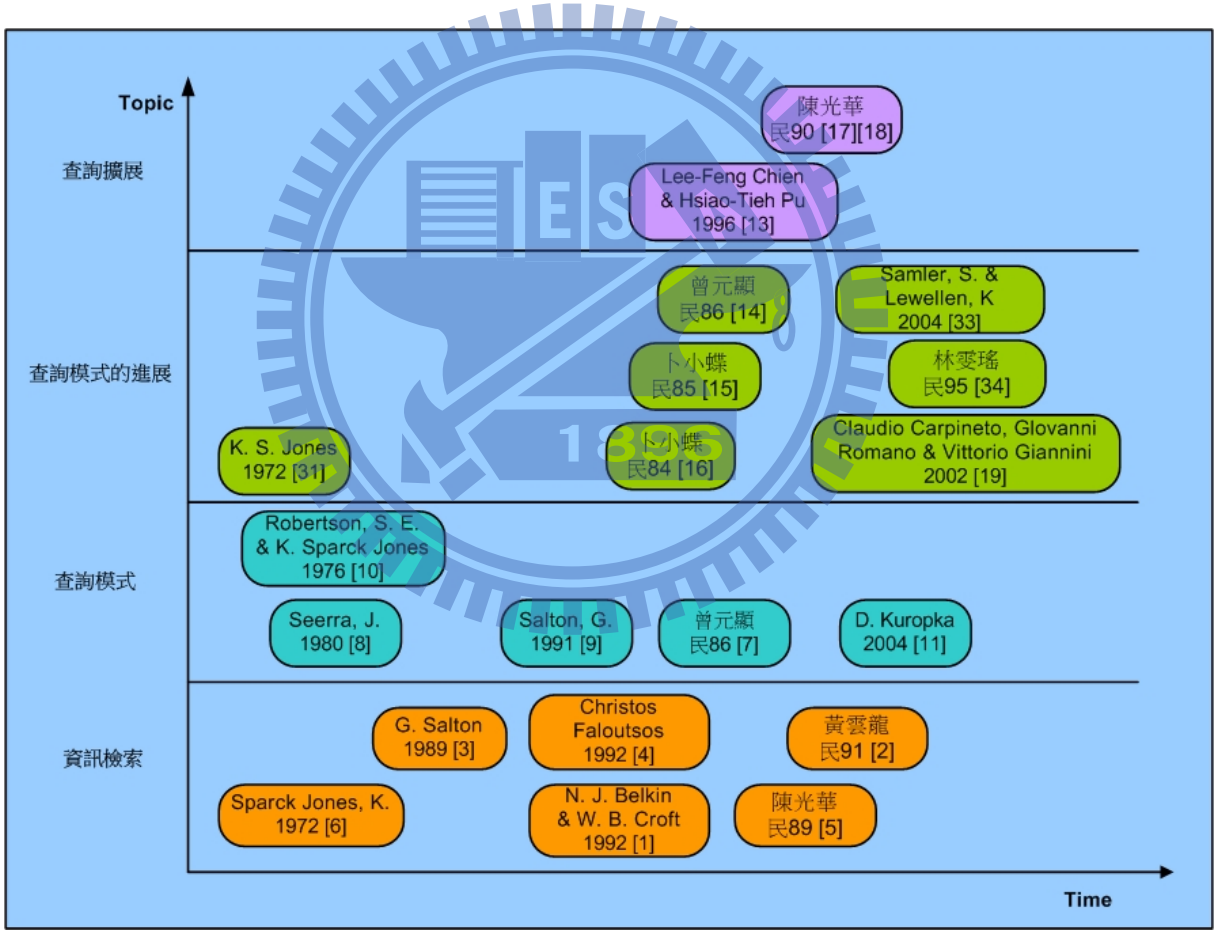


圖 2-1 文獻整理

2.1.1. 資訊檢索

Salton 對於「查詢」定義為：「a set of criteria specified by the user which describes

the kind of information desired」[3]，在資訊檢索過程中，就是由使用者輸入部份。即由使用者制定出一系列規範，用來明確描述其需求的資訊。在使用者產生資訊需求後，必須要能夠完整地向系統表達此資訊需求，若系統無法充份地瞭解使用者真正的資訊需求為何，則查詢效果將大受影響。因此要如何將使用者的資訊需求，忠實完整地轉化成系統能夠接受的查詢形式，一直是一個很重要的研究議題。

在網際網路應用益形重要的背景之下，資訊檢索愈來愈受到資訊科學研究者的重視。目前資訊檢索大致可分為下列三種型式[5]：

1. 全文檢索(Full-Text Retrieval)

根據使用者輸入的檢索詞彙，從文本或資料庫中，不限定資料欄位，自由地萃取出訊息的技術。基本上，全文檢索是針對文字型資料而設計的檢索方式，目前技術多數仍侷限於 TF×IDF(詞彙頻率乘反向文件頻率)的基本計算模式[6]。

2. 內容檢索(Content-Based Retrieval)

全文資訊檢索是使用者依據查詢主題之詞彙，來擷取相關的文件，在全文文件集中，運用與內容(context)相關的屬性特徵，做為檢索過程中文件內容的識別因子(context identifiers)，即所謂的內容檢索，亦即提供使用者以內容查詢的方式(query by content)，檢索所需的資源或資訊。例如，使用者可以使用特定的影像，檢索資料庫中類似的影像。

3. 詮釋資料檢索(Metadata Retrieval)

詮釋資料檢索的最大特色是可以達到一致性的檢索，利用詮釋資料來描述文字、影像、音訊、視訊等各種資料類型的內容或特色，進而達成協助檢索的目的。因此，檢索所得的資料，就可能包含多種資料類型。

2.1.2. 檢索模式

傳統的資訊檢索模式通常採用布林邏輯模式，布林邏輯模式在實作上較為簡單，也可針對不同欄位資料或相同欄位多個關鍵詞做布林邏輯運算，以縮小檢索範圍。然而其最大缺點在於檢索結果雖然都是符合條件的檔案或記錄，卻無法區別出個別文件對此次檢索的重要程度。此外，目前網路普及，一般使用者皆可從家裡、

辦公室、或任何方便的地方連線上網查詢資料，使用者若沒有受過相當的訓練，或具備足夠的經驗，很難利用布林邏輯方式，擬出比較有效的檢索策略來進行資料檢索。於是檢索技術的發展方向逐漸把使用者端的檢索複雜度，移向檢索主機這一端，或是設計更便利的人機介面模式，讓使用者的檢索環境越來越簡單，但還維持一定、甚至更好的檢索效能(Retrieval Efficiency)[7]。

圖 2-2 是整個檢索模式發展的歷史。左列標示集合論的(set-theoretical)、代數學的(algebraic)、機率的(probabilistic)分別是代表(Boolean model)[8]、向量空間模式(Vector Space Model)[9]和機率模式(Probabilistic Model)[10]下所擁有的著名演算法。在每個演算法間如果有箭號連接時，箭頭所指的演算法是屬於改進來源的演算法產生[11][12]。這三種模式分別介紹如下：

1. 布林邏輯模式

是利用集合中的交集和聯集來表示。例如查詢中關鍵字如果用布林表示 q = (information retrieval, chinese idiom, similar) = (1, 1, 0)，其中集合中有 1 則表示文件中有存在該關鍵字，0 就是代表沒有該關鍵字，所以從這個例子中關鍵字包含 information retrieval 和 chinese idiom 二個關鍵字，文件內容符合檢索詞之間的布林運算者才取出，不符合者即捨去。

2. 向量空間模式

轉換文件及查詢語句到向量空間後比對相似度，利用餘弦夾角(cosine)，將文件中關鍵字用向量表示並依權重表示出其中的差異程度。例如查詢 q = (information retrieval, chinese idiom, similar) = (3, 0, 1)，其中跟布林模式最大的差別在於向量中的數字是文件關鍵字所出現的次數。可概略稱為「近似字串查詢」、「容錯查詢」、或是「模糊搜尋」(Fuzzy Search)、「近似自然語言查詢」或「自然語言查詢」。模糊搜尋：即容錯式、全文式、非控制字彙、近似字串(Proximity)、允許利用近似自然語言的方式表達檢索字串與條件的檢索模式。此種模式大大降低資訊檢索的複雜度，對不明確自己檢索主題的使用者幫助尤其顯著。

3. 機率模式

利用機率的方式解釋查詢詞彙與相關文件的不確定性，典型是用 TF/IDF 演算法。TF 的目的是顯示出關鍵字在文件中所出現的次數，並以此來代表該關鍵字對文件的重要程度，IDF 的目的是看在所有文件中有該關鍵字的文件個數來分析這個關鍵字是否有代表性。優點有考慮多個文件間的交互關係和部份關鍵字所具的代表性。缺點是因為要計算所有文件中每個關鍵字相互關係，其中的計算複雜度相當高。並且所選擇的文件集合會影響結果。機率模式亦可做到向量模式的查詢效果，兩者不同處在基本假設與運算模式。

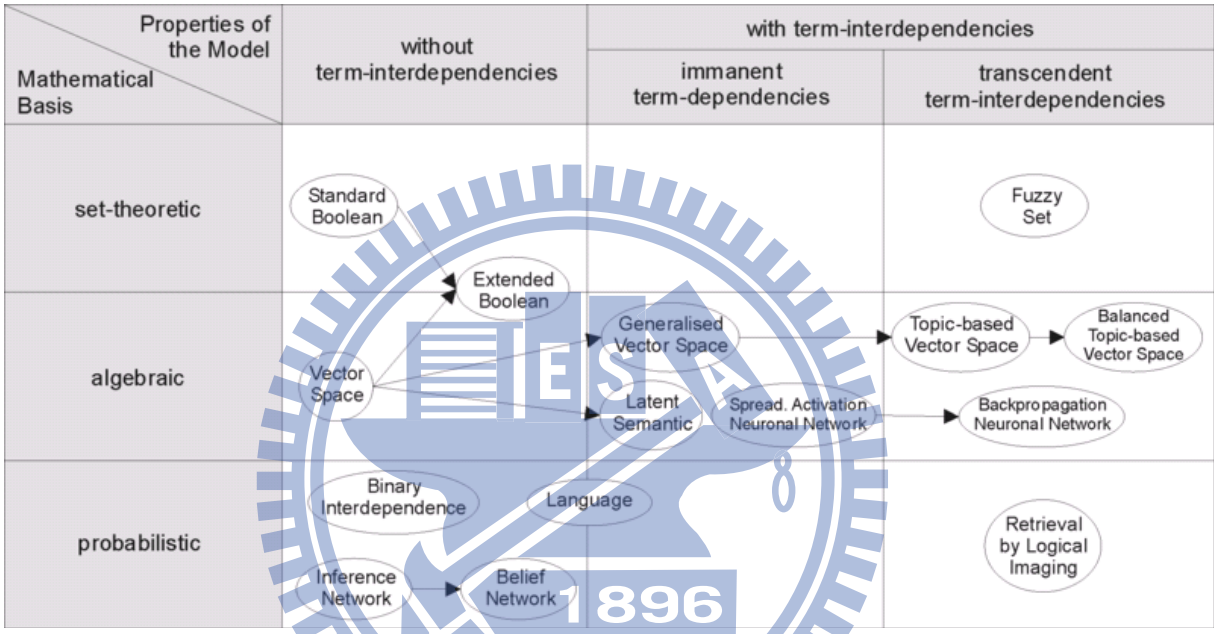


圖 2-2 文件比對的模型

資料來源：Dominik Kuropka [11]

2.1.3. 查詢模式的進展

資訊檢索技術歷經數十年的發展，累積了不少經驗與成果，在以使用者為導向(User Based)的趨勢下，各種查詢模式被提出來[13][7]：

1. 布林邏輯模式(Boolean model)

為最簡單的檢索模式，單純使用布林邏輯判斷文件中是否有檢索詞彙存在，對需求明確的檢索非常有效。然一般使用者較難以利用此種模式表達複雜的查詢而且缺乏程度上比對。

2. 重要性排序(Ranking)

文件檢索結果按符合程度大小排序，以顯示文件的重要性，加快檢索結果的檢視與利用，此為布林邏輯模式難以達到的重要功能。

3. 模糊搜尋(Fuzzy search)

即容錯式、全文式、非控制字彙、近似字串 (proximity)、允許利用近似自然語言的方式表達檢索字串與條件的檢索模式。此種模式大大降低資訊檢索的複雜度，對不明確自己檢索主題的使用者幫助尤其顯著。

4. 相關回饋(Relevance feedback)

使用者在將檢索詞鍵入系統後，會出現數筆相關資料，使用者依其資訊需求對每筆資料進行相關性評估，再回饋給系統。通常使用者會挑選重要的特徵，此種特徵若是文件本身，則可稱為相關文件回饋，若為相關詞，則稱為相關詞回饋，或檢索詞提示(term suggestion)[14]。系統便會根據這些使用者列入「相關」的資料，再重新做更進一步精確的檢索，使檢索結果的準確性大大提升。相關資訊回饋法最常用的技術是查詢句擴充法(Query Expansion, QE)。

5. 資訊過濾(Information filtering)

此模式主要是透過電腦來進行自動抽取、分類、摘要等工作；找出文件的關鍵詞並加以自動分類，再以自動摘要的技術將文件相關內容做一整理，提供給使用者瀏覽，其過濾的方式因應用範圍於不同領域、資訊系統所涵蓋的資訊特質不同，所採用的過濾方式也不同，其過濾技術大致可分為內容式資訊過濾(Content-based Information Filtering)及協力式資訊過濾(Collaborative Information Filtering)兩種，除了可在大型網路資源搜尋系統上使用，亦可和網頁瀏覽器結合運用[15][16]。

6. 語音檢索(Query by voice)

以語音為主的檢索，採用的技術主要是語音識別與其相關處理技術。由文字介面轉換為較為自然的口語語音介面，降低文字輸入的困難度，可以配合其他檢索模式一起運用。

7. 同音查詢(Approximate)

中文字存在許多同音字，此模式可以查出查詢值內中文字的「同音字」及「破音字」。

8. 自然語言檢索(Query by natural language)

對話式查詢仍由系統主導話題與使用的語句，自然語言檢索是運用人工智慧的方式，分析使用者在欄位中所鍵入的關鍵字，並能依據相關程度的多寡，排序呈現檢索結果。允許使用者以不限定的自然詞語、句法與系統溝通，因此使用者的負擔更輕。

2.1.4. 詞彙查詢擴展

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，同時也表達出使用者的資訊需求。一般而言，如果使用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，那麼檢索結果應能滿足使用者的資訊需求。但概念與詞彙的關係並非都是一對一，如同義詞(synonym)用來表示多個詞彙都具有相同概念。通常使用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。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，一般只提供可敘述其需要的詞彙即停止，幾乎是很短的查詢問句，所以得到的檢索結果也較差。因此，系統有必要提供加強協助使用者建構查詢問句的功能[17]。

以向量空間模型為基礎的資訊檢索系統，若文件與查詢問句相關，但其所使用的索引詞彙並非是檢索詞彙而是情境相同的詞彙，則該文件無法被檢索出來。常見的解決方法即是擴展查詢問句，亦即加入更多相關詞彙以提供足夠的資訊。查詢問句的擴展通常都以使用者提供的檢索詞彙為基礎，原始查詢問句的檢索效益如果不

好則可以追加更多詞彙。查詢問句擴展可以利用相關回饋或知識架構：相關回饋是以初次檢索結果為基礎的查詢問句擴展，但其效益是隨著原始查詢問句、排序的公式及相關詞彙的數量、初次檢索結果的品質而改變；以知識架構為基礎的查詢問句擴展並不依賴檢索結果，普遍多以統計或是以語料庫為基礎[18]。所以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檢索效益不彰問題。

2.2. 系統需求說明

2.2.1. 系統角色定位

在本系統的規範中，一個成語包含一個描述的詮釋資料(XML)檔案(成語資料和詮釋資料是一一對應的)，系統以詮釋資料來代表一個成語，檢索時主要是針對詮釋資料，而系統的主要工作有：

1. 將成語資料轉換為詮釋資料。
2. 將詮釋資料進行索引。
3. 提供可以快速且精準地查詢到該成語的能力。

由於使用者有不同的層級，為了讓他們更容易完成工作，系統需要提供不同角色使用者不同的功能介面：

1. 系統管理者功能：維持系統的運作、系統參數的調整。
2. 成語專家功能：確認成語資料的正確性並匯入系統。
3. 一般使用者功能：透過我們提供的介面來檢索系統的資料。

2.2.2. 系統工作流程

系統有三個工作流程：成語檢索流程、成語索引建立流程，以及同義詞建立流程，分別詳述如下：

1. 成語檢索流程

圖 2-3 為系統的檢索流程。使用者透過使用者介面輸入欲檢索的查詢問句，經過斷詞處理，擷取關鍵詞，轉換為系統可接受形式，之後進行查詢擴展，產生代表此查詢需求的關鍵詞集合，接著進行關鍵詞權重處理，並傳送至檢索機制。在檢索機制中將這些關鍵詞集合與文件特徵集合進行比對，找出符合使用者需求的文件資料，並依據關鍵詞權重計算文件的得分，將查詢結果按得分大小排序呈現給使用者，使用者可進行判斷是否要修改查詢語句重新送出查詢。

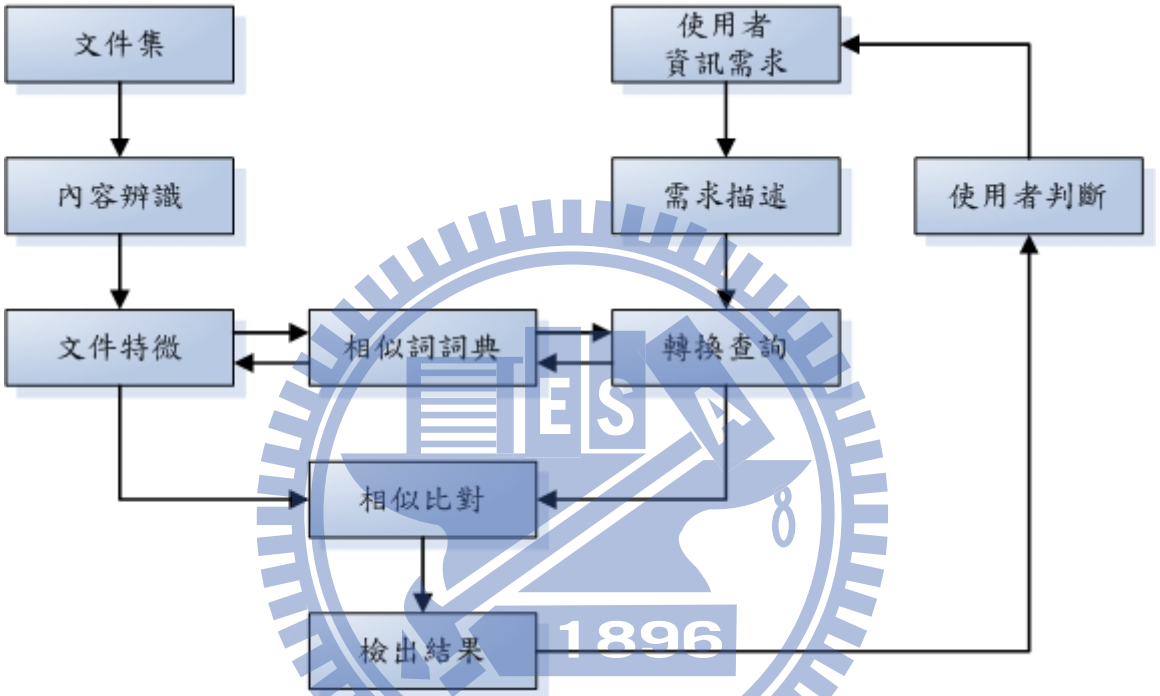


圖 2-3 成語檢索流程圖

2. 成語索引建立流程

圖 2-4 為成語索引建立流程。一個成語資料要進入成語資料庫，它必須依循系統標準工作流程：首先使用者可針對成語釋義提供建議資料，在資料確認步驟，成語專家角色人員會確認該資料是否正確，確認無誤後執行匯入，系統將進行關鍵詞擷取、資料轉換、索引建立並將其儲存到系統中，之後檢索和顯示都會依照該 XML 格式來進行。另一方面，系統也接受透過管理介面直接建立成語索引資料。

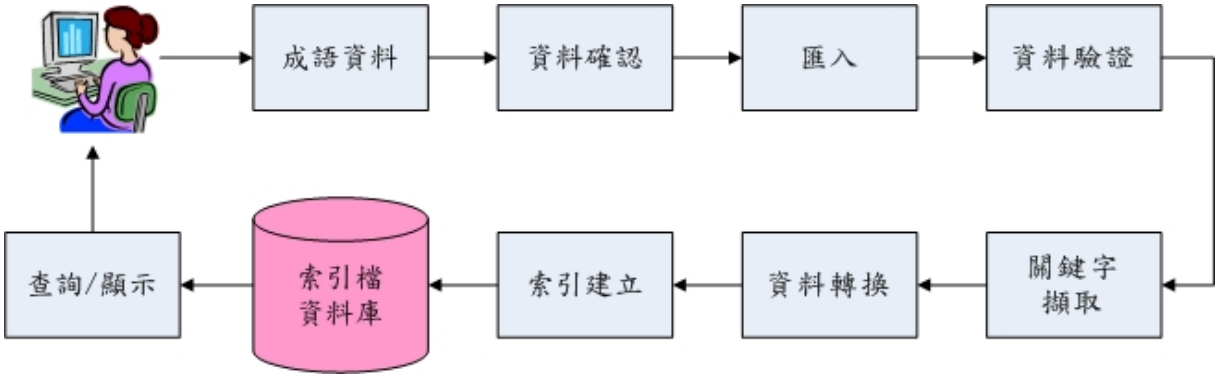


圖 2-4 成語索引建立流程圖

3. 同義詞建立流程

圖 2-5 為同義詞建立流程。同義詞資料要進入成語資料庫，它也必須依循系統標準流程。使用者可針對查詢問句所擷取的關鍵字提供同義詞建議，在資料確認步驟，成語專家角色人員將確認該資料是否正確，確認無誤後執行匯入，系統進行資料驗證並將其儲存到資料庫中，之後可提供查詢擴展時使用。另一方面，系統也接受透過管理介面直接建立同義詞資料。

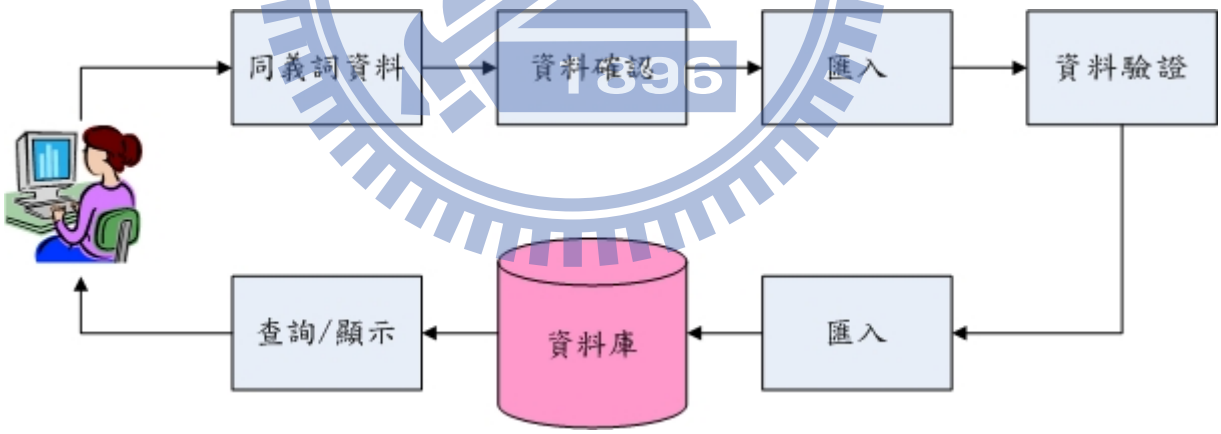


圖 2-5 同義詞建立流程圖

2.2.3. XML 文件處理

就先前所說，系統有一項主要的工作是在處理 XML，更嚴格來要講，是處理 XML 能夠真正拿來使用前的前置工作。成語資料在轉換為 XML 檔之前，必需先進行資料的確認，所以系統需提供成語資料維護，以及 XML 檔維護功能。處理 XML

檔的過程包含：關鍵字擷取、驗證、資料轉換、索引建立與刪除、檢視。這些過程在於確認 XML 檔的正確性，並且完成詮釋資料的索引建置。

2.2.4. 檢索功能

查詢模式的設計會影響到檢索效益，為了要讓檢索效益更為提升，需要多方面檢索技術與功能來相互配合，底下為本系統所規劃的查詢模式功能：

1. 向量空間檢索模式

   系統查詢模式以向量空間模型為基礎，利用計算查詢問句與文件中對應詞彙的權重來評估其相關性。

2. 重要性排序

   將文件檢索結果按符合程度大小(相關性)排序。

3. 相關詞回饋

   使用者提供給檢索系統的查詢句中，通常句子的長度都偏短，如此能提供之資訊便相對的減少，容易造成檢索時產生混淆的情況[19]。為了解決此類問題，系統將以查詢問句擴展方式，從相關文件內多找些同主題中常會出現的詞，以補充查詢句過短之缺點，並強調不同查詢詞的重要性，藉此改善檢索效果[20]。

4. 資訊過濾

   在一個數位化極度發展的時代裡，要能夠在浩瀚資訊裡擷取出所需要的資訊，指引出使用者所需之最相關資訊仰賴資訊過濾技術的發展。本系統將利用關鍵字比對方式，將文件內容(Content)加以分析比對，從大量的資訊中，過濾出重要訊息資訊並加以自動分類，提供給使用者瀏覽。

2.2.5. 評分設定功能

評分是搜尋引擎中很重要的一個概念。對於檢索的結果，需要按一定的順序回傳給使用者。因此，需要有一種機制來對檢索結果進行排序，以便將更精確的結果回傳給使用者。文件檔案的得分都與查詢詞有關[21]，為了加強檢索效能，將更精

準的查詢結果呈現給使用者，查詢時需能強調不同查詢詞的重要性，底下介紹本系統規劃影響評分的設定功能：

1. 停用字設定

查詢問句及文件內容經過斷詞後，需對其結果予以過濾，將一些比較不適合作為候選關鍵詞的詞類過濾掉，更進一步減少候選關鍵詞的範圍，以便更準確地找出更適合的特徵詞。

2. 文件特徵擷取設定

為了降低特徵詞數，簡化計算，提升查詢效率，系統需提供擷取查詢問句與文件內容的特徵詞的功能設定。

3. 同義詞權重設定

查詢詞需要適當地分配權重，在文件得分式子中需可改變每個查詢詞之權重，比較重要的詞給予較高的權重，相反地，對於不重要的詞給予較低的權重，如此，期望對於這些含有具代表性查詢詞比較少的相關文件在計算對查詢句之相似分數時能夠有所提昇。

2.3. 使用的程式工具

2.3.1. Java Servlet 和 JSP

本系統有大量工作需要在伺服器端完成，像是成語文件的斷詞切字、索引檔建立與更新等，這些工作都需要執行複雜且十分消耗系統資源的程式。有許多方法都可以在網頁伺服器上進行複雜動作，像是 CGI、PHP 等等，與傳統 CGI 和許多其他類似 CGI 技術相比，Java Servlet 具有更高的效率，更容易使用，功能更強大。Servlet 另一項優勢在於由標準 Java 所寫成，而 Java 語言特性之一就是跨平台，也因此 Servlet 自然擁有跨平台的能力。只要支援 Servlet 語言的網頁伺服器都能順利的執行 Servlet 程式。

目前使用 JSP 開發 Web 網頁程式，Tomcat 為較最佳選擇。Tomcat 是一種可以

執行 Java Servlet 及 Java Server Pages 技術的網頁應用伺服器。Tomcat 為 Apache Software Foundation 下的子 Project-Jakarta 所開發及維護。透過 Http Request 命令伺服器執行 Servlet 或 JSP 程式。Request 除了指示執行要求外，還能傳資料或參數，例如傳送上傳檔案內容。執行結果則透過網路傳回給呼叫的瀏覽器。由於能夠直接執行 Java Class，我們不但能使用 Java 大部份功能，更重要的是能夠使用許多已經開發完整的套件，例如處理 XML 的 DOM4J 和處理索引資料的 Jakarta Lucene。Servlet 和 JSP 都是在伺服器上執行 Java Class，Servlet 程式執行之前必須經過編譯，而產生的 class 檔案則存放在伺服端電腦的固定資料夾下，當客戶端使用者對伺服器提出請求時，則會自動執行被請求的 Servlet 程式，最後再回應給客戶端所需資料。JSP 方便的多，我們能直接在網頁中加入 Java 程式碼，所以在 JSP 中編寫靜態 HTML 更加方便，不必再用 println 語句來輸出每一行 HTML 代碼。更重要的是，借助內容和外觀的分離，頁面製作中不同性質的任務可以方便地分開：比如，由頁面設計專家進行 HTML 設計，同時留出供 Servlet 程式師插入動態內容的空間。

2.3.2. Apache Lucene

在建置系統時需要一個高性能全文索引引擎工具包，它要可以方便的嵌入到各種實際應用中實現全文搜索/索引功能，故選擇 Apache Lucene。

Lucene 是 Apache 軟體基金會 Jakarta 項目組的一個子項目，是一個使用 Java 語言開發且是開放原始碼的全文檢索引擎工具，提供資訊檢索所需要的重要功能：建立索引(index)和檢索(retrieval)，它可以方便的嵌入到各種應用中實現全文索引以及檢索功能。Lucene 的索引讓搜尋效率比傳統逐字比較大大提高。Lucene 提供一組解讀、過濾、分析檔、編排和使用索引的 API，它的強大之處除了高效和簡單外，最重要的是使用者可以隨時應自己需要自訂其功能[22]。

底下介紹 Lucene 的查詢語法與評分機制：

1. Lucene 查詢語法

Lucene 其檢索設計支援多種檢索模式，如布林檢索、短語檢索、模糊檢索、限制檢索等，它提供 API 供用戶自行建構檢索模式，同時它也支援檢索語句語法分析，即僅從檢索語句中按照某種語法給出檢索要求，而查詢分析器會解析這些語法，自動構造成相應的檢索類型。

對查詢語法做簡單的總結：

- 語法分析
  - 支持停用字過濾，可擴展
  - 詞幹還原：Porter Stemming 演算法
  - 可擴展語法分析
- 檢索模型：向量空間模型
  - 默認為布林權值，可自定義
  - 基於向量內積的相似度計算方法

2. Lucene 評分機制

評分機制是對檢索結果按某種標準進行評估，然後按分數值的高低來對結果進行排序。檔案的得分是在使用者進行檢索時即時計算出來的。所有檔案的得分應當都與使用者輸入的關鍵字有關係，而且是即時運算的結果。所謂得分，可以簡單理解成是某個關鍵字在某檔檔案中出現的頻率。圖 2-6 是 Lucene 用於計算某個關鍵字在對應於某文件中的得分公式[21][23]。

$$\sum_{t\ in\ q} tf(t\ in\ d) * idf(t) * boost(t.field\ in\ d) * lengthNorm(t.field\ in\ d) * coord(q,d) * queryNorm(q)$$

圖 2- 6 Lucene 評分公式

在 Lucene 得分公式中，已經包含了影響檔案評分的各種因素。表 2-1 中詳細介紹每一種因素對搜尋結果評分的影響作用。

表 2-1 Lucene 得分公式的解釋

因素	在公式中的作用描述
tf(t in d)	搜尋項 t 在文件檔案 d 中出現的頻率
idf(t)	搜尋項 t 在文件檔案中的反向頻率
boost(t.field in d)	域的加權因數(boost)，它的值在索引過程中進行設置
lengthNorm(t.field in d)	域的標準化值(normalization value)，即在某一域中所有項的個數。通常在索引時計算該值並將其存儲到索引中。

coord(q, d)	協調因數(Coordination factor)，該因數的值基於文檔中包含查詢的項的個數
queryNorm(q)	在給出每個查詢條目的方差和後，計算某查詢的標準化值

2.3.3. Apache Solr

Solr 是一個基於 Lucene 的 Java 搜索引擎伺服器。Solr 提供了層面搜索、醒目顯示並且支援多種輸出格式(包括 XML/XSLT 和 JSON 格式)。它易於安裝和配置，而且附帶了一個基於 Http 的管理介面。Solr 已經在眾多大型的網站中使用，較為成熟和穩定。可以使用 Solr 表現優異的基本搜索功能，也可以對它進行擴展從而滿足企業需要。因為 Solr 包裝並擴展了 Lucene，所以 Solr 的基本上沿用了 Lucene 的相關術語。更重要的是，Solr 建立的索引與 Lucene 搜索引擎庫完全相容。通過對 Solr 進行適當的配置，某些情況下可能需要進行編碼，Solr 可以閱讀和使用構建到其他 Lucene 應用程式中的索引[24]。

Solr 對外提供標準的 Http 介面來實現對資料索引的新增、修改、刪除、查詢。在 Solr 中，用戶通過向部署在 Servlet 容器中的 Solr Web 應用程式發送 Http 請求來啟動索引和搜索。Solr 接受請求，確定要使用的適當 SolrRequestHandler，然後處理請求。透過 HTTP 以同樣的方式返回結果，圖 2-7 為 Solr 運作的原理。可以向 Solr 索引 Servlet 傳遞四個不同的索引請求：

1. add/update：允許向 Solr 增加或更新索引文件。直到提交後才能搜索到這些新增和更新。

2. commit：告訴 Solr，將上次提交以來所所有變更生效。

3. optimize：優化 Lucene 的索引檔以改進檢索性能。如果更新比較頻繁，則應該在使用率較低的時候安排優化。一個索引無需優化也可以正常地運行。優化是一個耗時較多的過程。

4. delete：可以透過 id 或查詢來指定。按 id 刪除：刪除具有指定 id 的文檔；按查詢刪除：刪除查詢返回的所有文件。

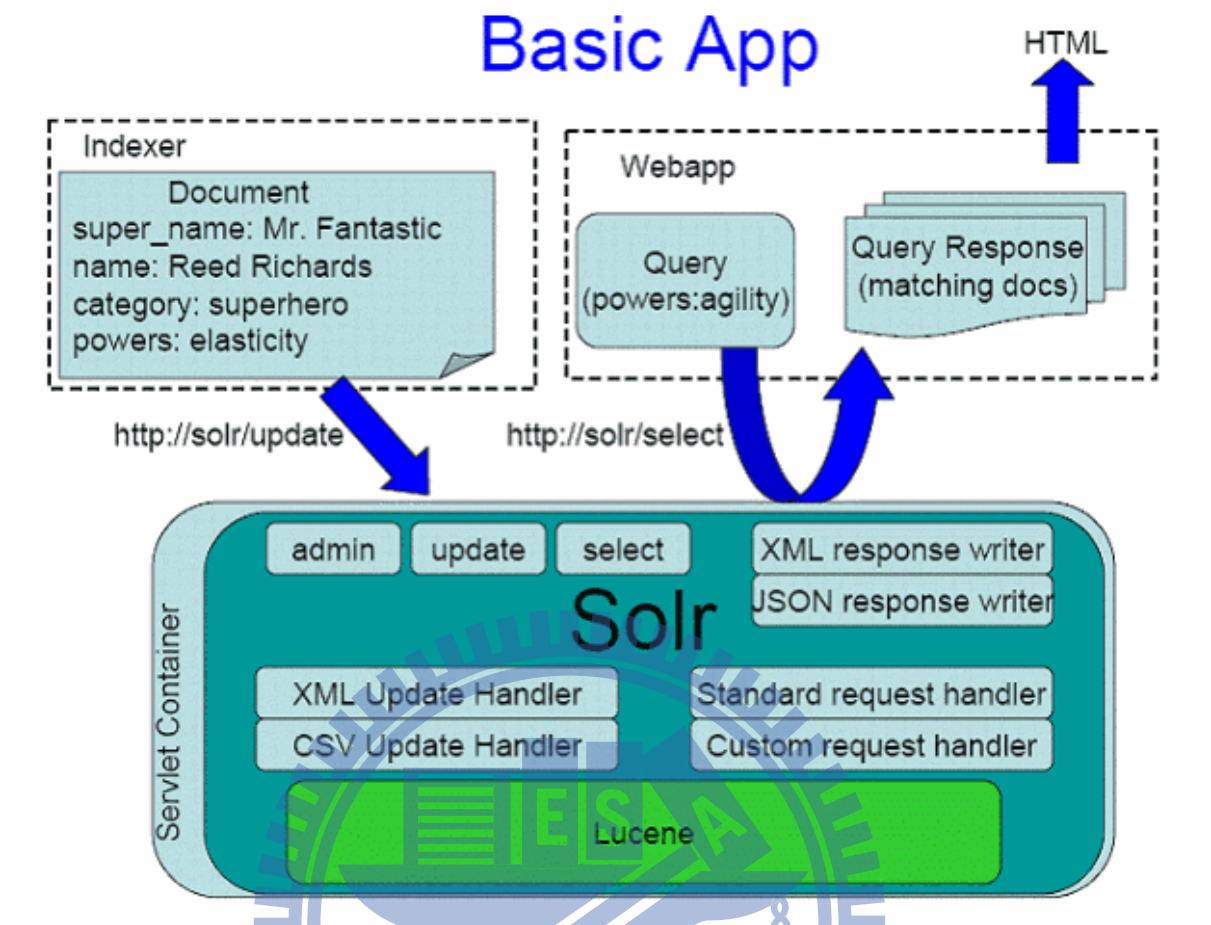


圖 2-7 Solr 運作原理

資料來源：http://wiki.apache.org/Solr/FrontPage

2.3.4. Ajax

Ajax 全稱為「Asynchronous JavaScript and XML」(非同步 JavaScript 和 XML)，是一種創建互動式網頁應用的網頁開發科技。在傳統 Web 應用程式中，通常使用同步的互動過程，這種方式下，使用者透過某個觸發行為來提交表單，向伺服器發出請求。伺服器接收到請求後，向使用者回傳一個 HTML 頁面，使用者端都要進行頁面的跳轉或者更新。這種方式多少會讓使用者感覺到不連貫。伺服器在處理請求時，使用者多數時間處於等待的狀態，螢幕中網頁的內容也是一片空白[25]。

Ajax 技術的出現在一定程度上彌補了傳統 Web 應用程式的不足。Ajax 在使用者與伺服器之間扮演了重要的角色，使用者不必使用頁面更新或跳轉的手段來從伺服器取得資料。Ajax 是用 JavaScript 語言來撰寫，主要用來完成使用者與網頁應用程式之間的互動過程採用非同步的方式進行，而不必一直等待某一個操作所需要得

到的回應已經全部傳回。

使用 Ajax 不但可以向伺服器端發送請求，而且還可以接收伺服器回傳的資料，其最大優點，就是能在不更新整個頁面的前提下維護資料。這使得 Web 應用程式更為迅捷地回應使用者，並避免了在網路上發送那些沒有改變過的資訊。

2.3.5. DOM4J

DOM4J 是一個易用的、開源的程式套件，用於 XML，XPath 和 XSLT。它應用於 Java 平臺，採用了 Java 集合框架並完全支援 DOM，SAX 和 JAXP。DOM4J 使用來非常簡單，只要你瞭解基本 XML-DOM 模型，就能使用，是一個非常優秀的 Java XML API，具有性能優異、功能強大和易用的特點[26]。

2.3.6. HttpClient

HTTP 是現在 Internet 上使用得最多、最重要的協定，越來越多 Java 應用程式需要直接透過 HTTP 協定來訪問網路資源。雖然在 JDK 的 java.net 包中已經提供了訪問 HTTP 協議的基本功能，但是對於大部分應用程式來說，JDK 庫本身提供的功能還不夠豐富和靈活。HttpClient 是 Apache Jakarta Common 下的子項目，可以用來提供高效能、最新、功能完善來支持 HTTP 協定的用戶端開發工具，並且它支援 HTTP 協定最新的版本和建議[27]。

2.3.7. HTML Parser

HTML Parser 是一個對現有 HTML 進行分析的快速即時的解析器，讓我們可以從 HTML 中提取所需資訊。不同於 XML 這種對格式要求非常嚴格的標記語言，HTML 在推出時並沒有對其格式進行嚴格定義，比如 HTML 中標籤並不一定要成對出現，但是又要求流覽器能儘量的正確顯示其所要表達出來的內容。流覽器經過多年發展其適應能力越來越強，很多格式非常糟糕的 HTML 檔都能顯示得令人滿意。不過如果我們需要精確獲取 HTML 中包含的資料，這恐怕比顯示一個 HTML 更令人頭疼。HTML Parser 是解決此問題程式工具包[28]。

2.3.8. SolrJ

前面小節提到 Solr Server 對外提供 Web Service 的溝通方式，接受 XML 格式資料。當要把 Solr 跟商業邏輯整合時，希望有更加友好的介面和工具。SolrJ 應運而生，並且被官方包含在 Solr1.3.0 的發佈程式包中。SolrJ 是 Java 程式師與 Solr Server 最好的對話工具包，它將所有 Solr 對外介面封裝成了規整、精巧、靈活的工具包，讓程程式員可以免去直接解析、格式化 XML[29]。

三、 系統分析與設計

本章描述 MIRS(Meaning of Idiom Retrieval System)的架構設計與各步驟所採用的方法。3.1 節介紹系統設計模式與整體架構；3.2 節介紹成語資料的前置處理(Preprocessing)與索引建立；3.3 節介紹如何將查詢問句轉換為關鍵詞集合；3.4 節介紹如何將關鍵詞進行權重調整、轉換為 Lucene 查詢語句，並透過 Lucene 搜尋引擎檢索出文件並返回給使用者。

3.1. 系統設計

本研究所開發之 MIRS，其系統設計示意圖，如圖 3-1 所示，主要分為四大模組：前置處理模組、查詢問句處理模組、文件檢索處理模組、及查詢結果顯示。首先分析成語資料，定義成語的詮釋資料(Meta-Information)，清楚地描述成語的釋義、典源、典故等資訊，將成語資料進行前置處理，產生詮釋資料檔，利用 Solr Web 建立索引資料供檢索使用；另一方向，將使用者輸入的查詢問句進行處理，擴展查詢，產生代表此表問句的關鍵詞集合；接著進行文件檢索的處理，賦予關鍵詞權重，交由 Solr Web 進行檢索，最後將檢索結果返回給使用者。

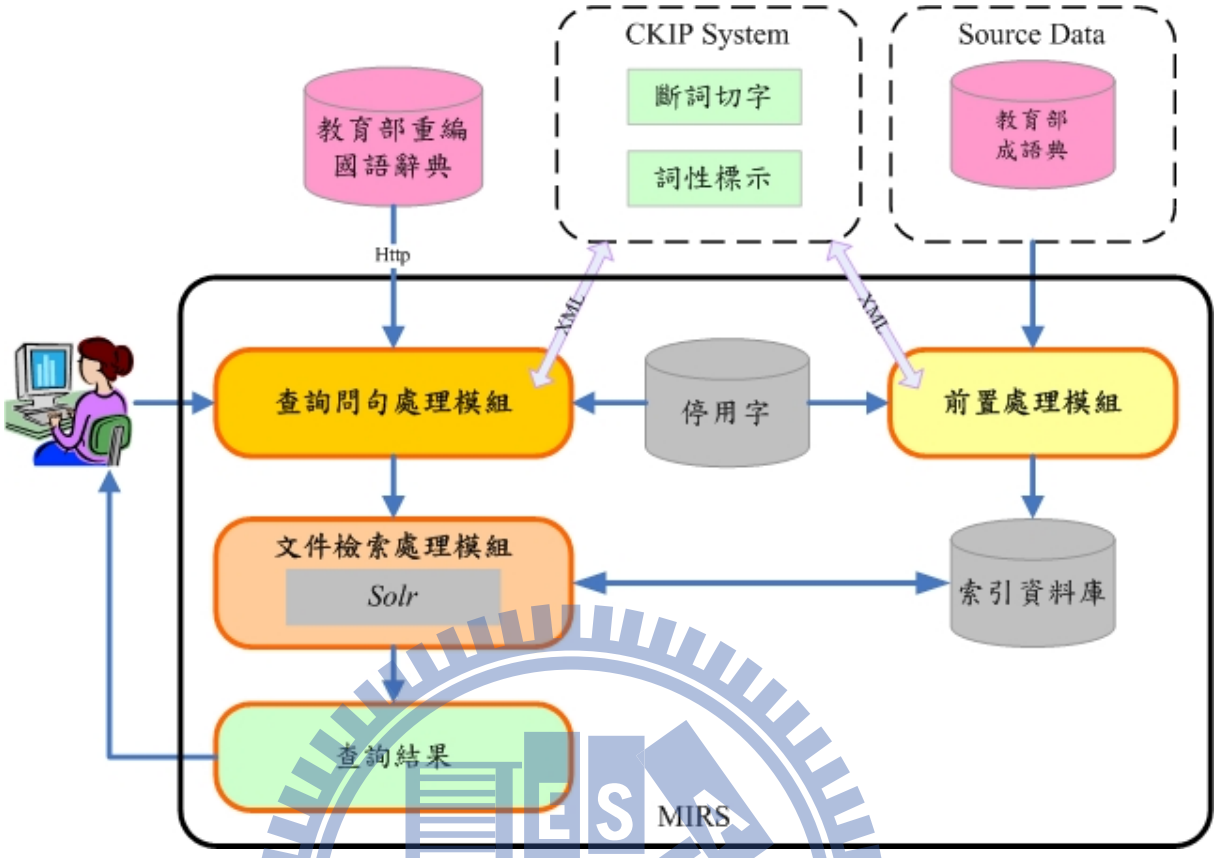


圖 3-1 系統設計示意圖

3.1.1. 設計架構與模式

MIRS 的建立是採用三層式(展示層、商業邏輯層、資料存取層)架構，商業邏輯層則使用 MVC 設計模式，將系統分成 Model, View, 及 Control 三種元件，如圖 3-2 所示，分別介紹如下：

1. 三層式架構

(一) 展示層(Presentation Layer , PL)

為Client端的使用者透過圖形使用者介面(Graphic User Interface , GUI)程式(如：IE Browser)，來與Web Server進行互動，在三層式架構中展示層相當於使用者介面，提供使用者輸入資料的介面。

(二) 商業邏輯層(Business Logic Layer , BLL)

負責在Web Server上執行處理邏輯運算，此層接受從Client端的請求

(Request)服務，並且將程式處理結果傳回到Client 端，它的目的是用來作為使用者與資料庫之間的溝通橋樑，商業邏輯層相當於Web Server，提供程式處理運算的平台。

(三) 資料存取層(Data Access Layer , DAL)

主要是提供資料給企業邏輯層來處理，Client 端無法直接對資料庫進行存取，必須經由企業邏輯層與資料存取層的連結才能對資料進行存取，也因此提高的系統的安全性。

2. MVC 設計模式

MVC 最早架構在 Smalltalk-80 語言中被使用來處理使用者介面，之後此架構成為一個被廣泛運用來處理使用者介面的框架(Framework)[30]。使用 MVC 的做法是軟體元件分為 Model(模型)、View(呈現)、Controller(控制子)等三個獨立元件，以下就是 MVC 模式中的三種元件的任務：

(一) Model 元件

它的任務是維持應用程式的資料或狀態。它同時管理了由資料來源處提取及儲存資料的任務。當資料變更時，並通知 View 元件展示這些資料。

(二) View 元件

包含展示邏輯。它把 Model 元件所包含的資料展示給使用者。它也讓使用者得以與系統進行互動，並通知 Controller 使用者的動作。

(三) Controller 元件

管理了這整個流程。它實例化 Model 與 View，並協調這兩個元件彼此間的動作。依照應用程式的需求，它可以一次實例化多個 View，並把它們與同一個 Model 一齊協同工作。它傾聽使用者的動作，並依照商業邏輯去指示 Model 工作。

把資料代理(Model)與資料呈現(View)分開來。使用相同的資料得以以不同的方式呈現出來。Model 和 View 元件可以獨立進行更動，而介面卻仍

然保持不變。這增加了系統的可維護性與可擴充性。把資料呈現(View)與應用程式行為(Controller)分開來。這使得 Controller 得以在執行時期依 Model 去製造出合適的 View。把應用程式行為(Controller)與資料代理(Model)分開來，這讓使用者的請求得以經由 Controller 去對映到 Model 裡應用程式層級的功能。

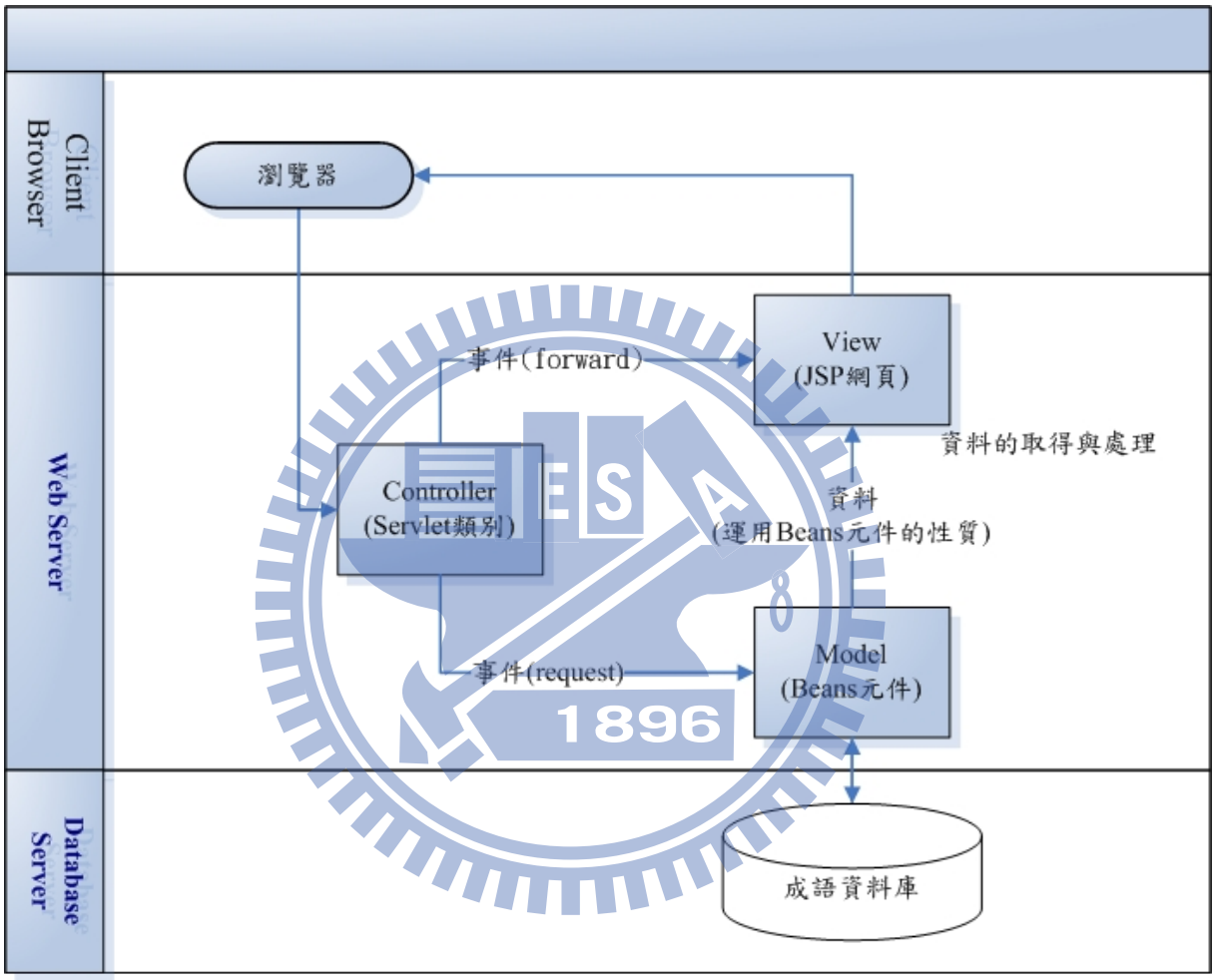


圖 3-2 系統設計模式圖

3.1.2. 權限控管

MIRS 每一 Model 元件都對應到唯一的事件碼(Event ID)，所有提交給 Controller 元件的需求中均包含此事件碼，Controller 會驗證目前使用者是否有權限執行該事件，其流程如圖 3-3 所示。

在權限控管上是利用角色管理原則(Role-based Access Control)來建立權限。每一個角色可以定義允許執行的事件碼。以角色為依據的權限控管機制中，管理者管

理不同角色所需的權限及每個使用者所擁有的角色。當某使用者被刪除時，管理者無須額外處理權限變換相關事宜，可減少管理者管理成本。

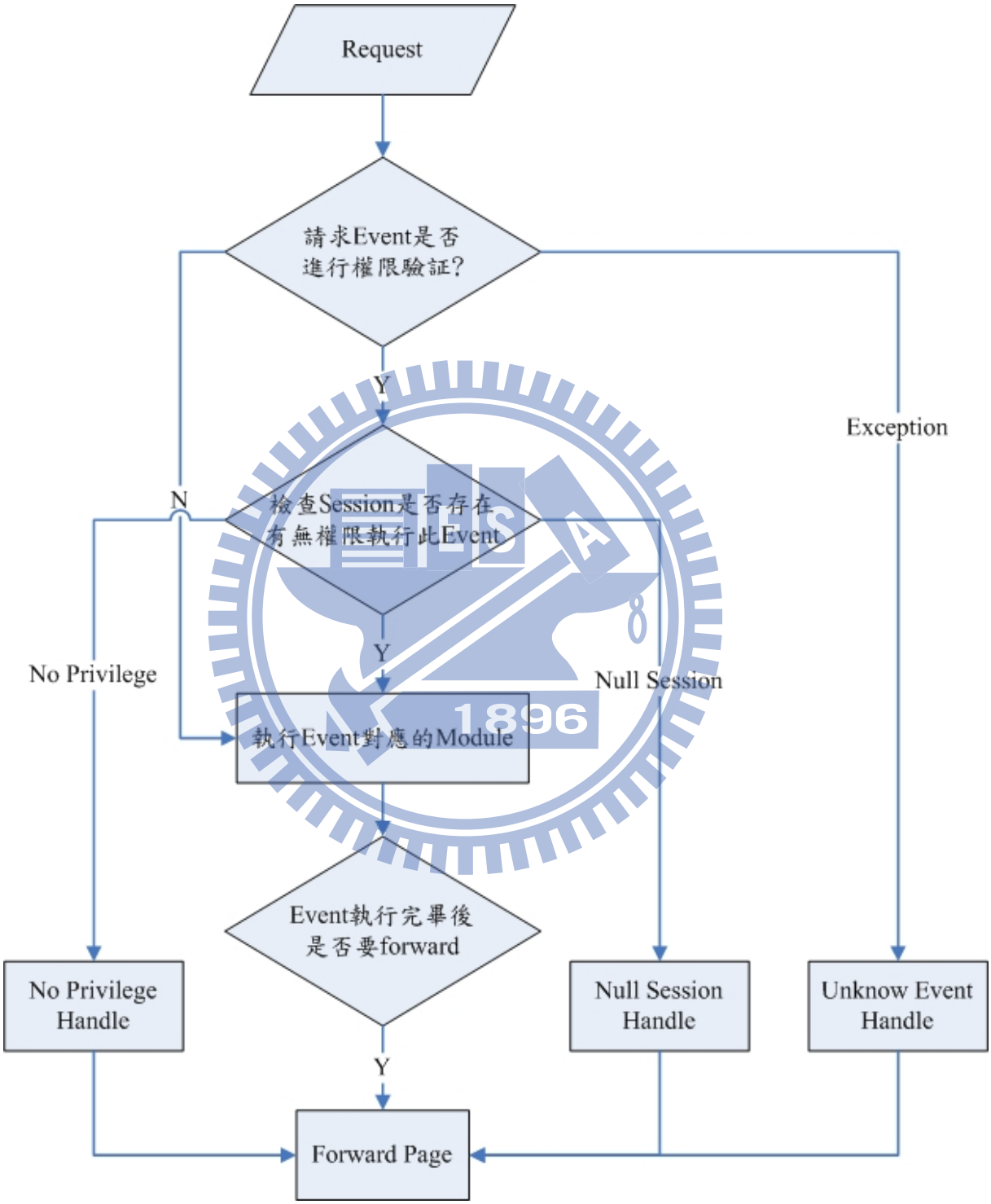


圖 3-3 權限控管流程圖

3.2. 前置處理模組

建立索引前之首要工作在於將雜亂無序的敘述資料，經過整理，成為系統所需資料，若無此項程序，則系統較不易處理資料，且易造成檢索成效不彰之結果，因此，資料前置處理便顯得格外重要。MIRS 所進行之前置處理程序包含成語資料的準備，斷詞切字、詞性標示、去除標點符號、去除停用字、關鍵字合併、文件特徵挑選及產生詮釋資料等步驟，其流程如圖 3-4 所示。

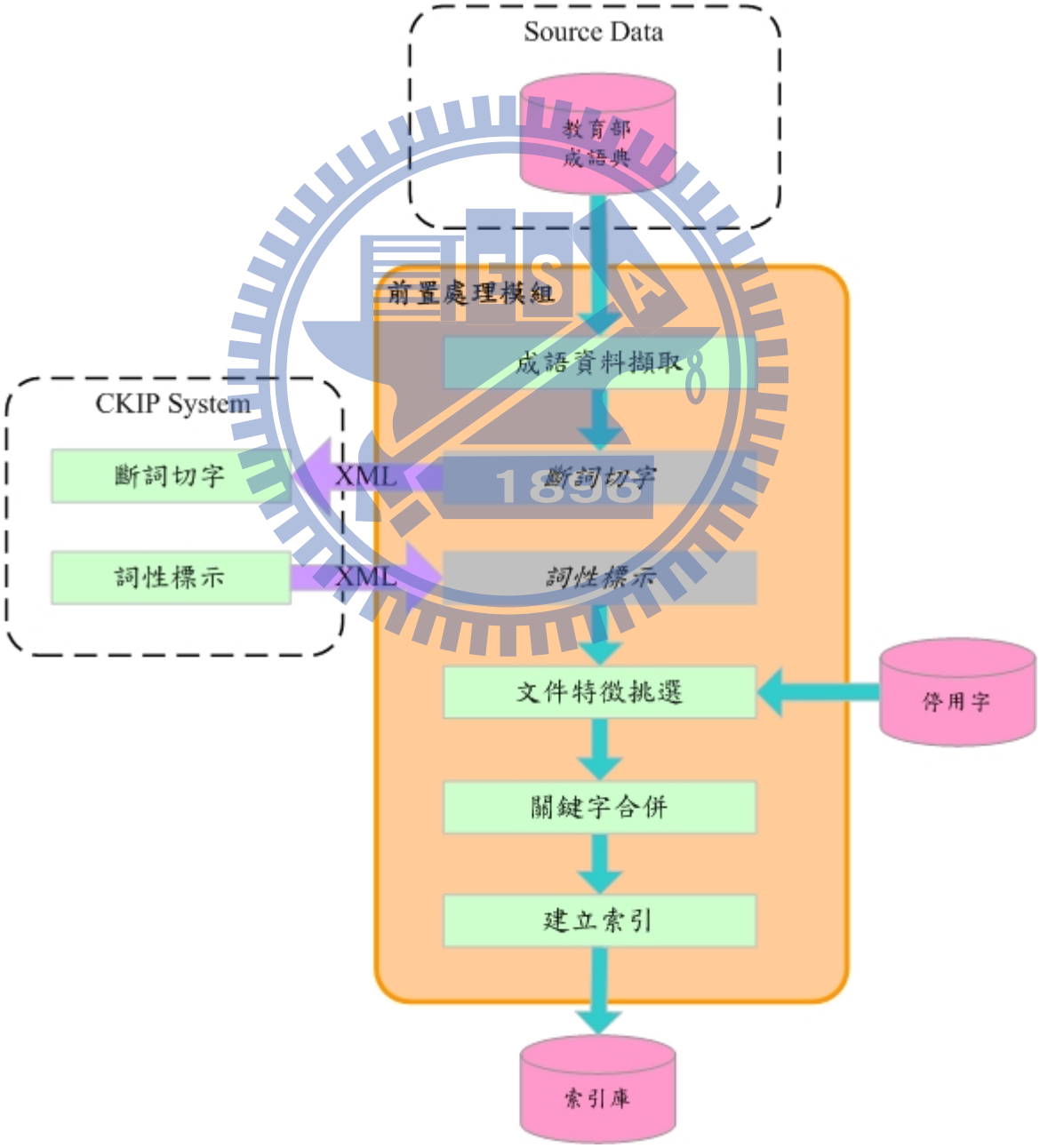


圖 3-4 前置處理流程圖

3.2.1. 成語詮釋資料格式

成語文件是藉由詮釋資料來描述整份文件內容，這些詮釋資料是透過自動處理機制分析，拆解，以及擷取方式得到。此詮釋資料是為了產生成語文件內容基本描述，並給予一個唯一識別碼來表示，例如，此文件的識別碼(id)，成語標題(title)，釋義(interpretation)等等。我們依據教育部成語典的成語資料內容，總共定義 10 個標籤屬性，說明如下：

- id：系統給予唯一識別碼。
- title：成語標題。
- interpretation：成語完整的釋義描述，包含釋義、典源及典故等資訊。
- interpretationKey：釋義描述的關鍵詞集合。
- interpretationAlt：成語釋義的語義資訊。
- interpretationAltKey：釋義及語義資訊的關鍵詞集合。
- creationDate：文件建立日期。
- filepath：檔案相對位置。
- filename：檔案名稱。
- rating：自定演算法檢索時文件的得分。

3.2.2. Solr 的建立

1. Solr 安裝和配置

關於 Solr 的安裝和配置，這裡有兩篇非常好的文章[5]，下面說明需要注意的地方。

[5] 使用 Apache Solr 實現更加靈巧的搜索
http://www.ibm.com/developerworks/java/library/j-solr1/
http://www.ibm.com/developerworks/java/library/j-solr2/

Solr 安裝並不困難，下載 Solr 的 zip 包後解壓縮將 dist 目錄下的 war 檔改名為 solr.war，接著複製到$TOMCAT_HOME\webapps 目錄並設置 Solr 主位置。以本系統為例，是在 Tomcat 裡配置 java：comp/env/Solr/home 一個 JNDI 指向 Solr 主目錄，建立$TOMCAT_HOME/conf/Catalina/localhost/Solr.xml 文件，文件內容下圖 3-5 所示。

```
<?xml version='1.0' encoding='utf-8'?>
<Context docBase="solr" path="/solr" reloadable="true" debug="0"
crossContext="true">
<Environment name="solr/home" type="java.lang.String" value="c:/tomcat/tomcat6/solr"
override="true" />
</Context>
```

圖 3-5 JNDI 配置內容

這個 Solr 主位置，裡面存在兩個重要目錄：conf 和 data。其中 conf 目錄存放最重要的兩個配置檔 solrconfig.xml 和 schema.xml，前者是該索引服務的一些屬性，比如索引檔存放目錄(預設為與 conf 同一層的 data 目錄)，還有緩存，分佈之類的一些設置；後者是索引庫的欄位定義，欄位類型，欄位名，處理方式。詳細說明請參考 Solr Wiki[6]。

Solr 在預設的情況下只能搜尋英文，不能識別中文，Solr 內核支援 UTF-8 編碼，所以修改$TOMCAT_HOME\conf 裡的 server.xml 配置文件，在 8080 埠定義的後面加上 URIEncoding="UTF-8"，如圖 3-6 所示，如此才能識別 UTF8 編碼的中文輸入。另外，向 Solr Post 請求時需要轉為 UTF-8 編碼，返回的查詢結果也需要進行一次 UTF-8 轉碼。檢索資料時對查詢的關鍵字也需要轉碼，並用"+"符號連接。

```
<Connector port="8080" protocol="HTTP/1.1" connectionTimeout="20000" redirectPort="8443"
URIEncoding="UTF-8" />
```

圖 3-6 Tomcat 連接埠定義

[6] Solr Wiki http://wiki.apache.org/solr/SchemaXml

2. Solr schema 定義

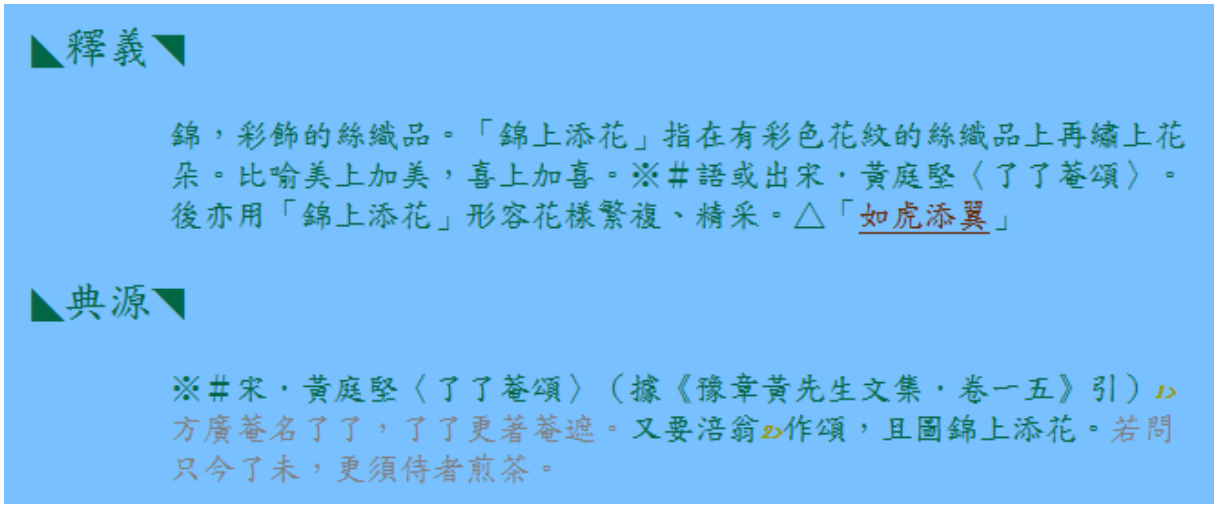
完成 Solr 的安裝後，接著依照先前介紹的成語詮釋資料格式定義，修改 schema.xml 文件內容，其結構如圖 3-7 所示。

```
<field name="id" type="text_ws" indexed="true" stored="true"/>
<field name="title" type="text" indexed="true" stored="true"/>
<field name="interpretationKey" type="text" indexed="true" stored="true" multiValued="true"
omitNorms="true"/>
<field name="interpretationAltKey" type="text" indexed="true" stored="true" multiValued="true"
omitNorms="true"/>
<field name="creationDate" type="date" indexed="true" stored="true"/>
<field name="filepath" type="text_ws" indexed="true" stored="true"/>
<field name="filename" type="text_ws" indexed="true" stored="true"/>
<field name="rating" type="sint" indexed="true" stored="true"/>
<field name="interpretationAlt" type="text" indexed="false" stored="true" multiValued="true"/>
<field name="interpretation" type="text" indexed="false" stored="true" multiValued="true"/>
```

圖 3-7 Solr schema.xml 範例

3.2.3. 成語資料的建立

MIRS 實驗資料是採用教育部成語典的內容，首先我們發現到教育部成語典對於成語資訊內容有兩種不同的顯示格式，如圖 3-8 及圖 3-9 所示，成語內容解釋介於兩兩段落標題之間，我們依此特點利用網頁解析器(HTML Parser)剖析網頁文件，分別將成語的釋義、典源及典故等資訊內容擷取出來，並儲存到資料庫中供後續系統的分析與建置使用。



釋義

錦，彩飾的絲織品。「錦上添花」指在有彩色花紋的絲織品上再繡上花朵。比喻美上加美，喜上加喜。※＃語或出宋．黃庭堅〈了了菴頌〉。後亦用「錦上添花」形容花樣繁複、精采。△「如虎添翼」

典源

※＃宋．黃庭堅〈了了菴頌〉（據《豫章黃先生文集．卷一五》引）1)方廣菴名了了，了了更著菴遮。又要涪翁2)作頌，且圖錦上添花。若問只今了未，更須侍者煎茶。

圖 3-8 教育部成語典成語資料顯示格式

教育部重編國語辭典修訂本

條目	早生貴子
注音一式	ㄗㄠˇ ㄕㄥ ㄍㄨㄟˋ ㄗˇ
漢語拼音	zǎo shēng guì zǐ
釋義	祝人夫婦早生男孩。初刻拍案驚奇．卷六：「賈門信女巫氏，情願親誦白衣觀音經卷專保早生貴子，吉祥如意。」

圖 3-9 教育部重編國語辭典成語顯示格式

3.2.4. 斷詞切字

詞彙是最小有意義且可以自由使用的語言單位。任何語言處理系統都必須先能辨別文本中的詞才能進行進一步處理，例如機器翻譯、語言分析、語言了解、資訊抽取。然而中文詞的結構，有單字詞、多字詞等多種不同型態，且中文文件中詞與詞的界線不明；不像英文，極大多數都是一個詞(Word)，就是一個意義單位(Meaning Unit)，因此中文處理較英文困難，中文自動分詞的工作也就成了語言處理不可或缺的技術。基本上自動分詞多利用詞典中收錄的詞和文本做比對，找出可能包含的詞，由於存在歧義的切分結果，因此多數的中文分詞程式多討論如何解決分詞歧義的問題，而較少討論如何處理詞典中未收錄的詞出現的問題(如何辨認新詞)。根據統計，一般的文章中約有百分之三到百分之五的未知詞，因此一個演算法的未知詞

識別能力對於其分詞與標記的正確率將有很大的影響。中央研究院中文詞知識庫小組(Chinese Knowledge Information Processing Group, CKIP)所研發之中文斷詞系統(包含未知詞擷取與標記)提供了一個解決方案，可以線上即時分詞功能。為一具有新詞辨識能力並附加詞類標記的選擇性功能之中文分詞系統，輸出結果如圖 3-11 所示。分詞依據為此一詞彙庫及定量詞(Quantifier)、重疊詞(Reiterative)等構詞規律及線上辨識的新詞，並解決分詞歧義問題。[31]。

本研究將透過中研院資訊科學所詞庫小組所開發之 CKIP 斷詞系統，做為斷詞處理的工具，此系統採用線上服務模式，使用一 API 呼叫，資料的交換方式為 XML，採用 TCP Socket 連線傳送驗證資訊及文本到 Server 進行斷詞，伺服器經過處理後經由原連線傳回斷詞結果。

3.2.5. 詞性標示

本研究所採用之斷詞系統除斷詞功能外，亦可輸出簡化之詞類標記。此系統提供之線上斷詞系統的內部處理採用中研究資訊科學所詞庫小組所編列的中研院平衡語料庫詞類標記集之「簡化詞類」，而線上斷詞服務則採用「精簡詞類」。

經由斷詞系統處理過的文件以 XML 格式傳回結果，對每個擷取出來的詞以括弧標示詞性，每個詞之間以全形空白隔開。

形容容貌美麗、姿態曼妙的女子↵

圖 3-10 前置處理實例[7]

將圖 3-10 的原文輸入中研院 CKIP 中文斷詞系統之後，進行斷詞處理並標示詞類，輸出結果如圖 3-11 所示。

形容(VT)　容貌(N)　美麗(VI)　、(PAUSECATEGORY)　姿態(N)　曼妙(Vi)　的(T)　女子(N)↵

圖 3-11 前置處理實例－斷詞切字與標示詞性結果

[7] 摘錄自 教育部成語典 成語「國色天香」的釋義內容

CKIP Client 專案[8]是將上述斷詞流程依不同程式語言實作，提供使用者方便使用 CKIP 中文斷詞服務。我們利用此工具將成語的釋義內容經由 CKIP 中文斷詞系統分析，將結果存入資料庫，記錄詞彙名稱以及詞性等訊息。

3.2.6. 停用字

停用字是指「電腦檢索中的虛字、非檢索用字」，這些字在文章中並沒有語意，通常用來平順語意的詞。這是為了節省儲存空間和提高搜索效率，搜索引擎在索引頁面或處理搜索請求時會自動忽略某些字或詞，這些字或詞即稱為停用詞，通常包括介系詞、指示代名詞、連詞、助詞等，如：的、是、之、我們。某些停用字在文件資料中出現的頻率極高，若是以頻率來計算詞彙的重要程度，那麼有些停用字會因此而被突顯出來，這些出現頻率較高但不具檢索價值的詞彙，將對文件中的有效資訊造成雜訊干擾，因此將它們歸納於停用字一覽表(Stop word List)中，所以在前置處理及搜索引擎檢索之前都要對所索引的資訊進行消除雜訊的處理。在文件內容中適當地減少停用詞出現的頻率，可以有效地幫助我們提高關鍵字密度，所以在文件標題標籤中避免出現停用詞能夠讓所優化的關鍵字更集中、更突出。停用字的擇定一來不可太寬鬆，以免降低分類的成效，但又不能太少，以免遺漏重要資訊，亦會影響檢索結果。表 3-1 為本研究部分停用字範例。

表 3- 1 停用字範例

次數	詞彙	詞性	次數	詞彙	詞性	次數	詞彙	詞性
4664	比喻	Vt	349	為	Vt	221	好	Vi
3696	形容	Vt	335	知	Vt	207	言	Vt
2284	參	Vt	318	是	Vt	193	說	Vt
998	有	Vt	302	沒有	Vt	185	小	Vi
646	無	Vt	290	使	Vt	175	見	Vt
416	大	Vi	246	死	Vi	173	作	Vt
400	做	Vt	238	指	Vt	165	生	Vt

3.2.7. 文件特徵挑選

特徵挑選一直是影響著檢索效率的一項重要環節，一般構成文件敘述資料的詞

[8] CKIP Client 專案 http://ckipclient.sourceforge.net/

彙數量相當大，這代表敘述資料的向量之維度(dimension)相當大，但真正重要的詞彙卻並不是那麼多，因此，若無經過特徵選取過程，則原始敘述向量會產生許多冗餘資料(redundancy)。特徵選取的目標是要從原有特徵中挑選出最佳的部分特徵，使其辨識率(recognition rate)能夠達到最高值，其對系統最主要目的有兩個，第一，簡化系統運算時間、增進系統效率。第二，使敘述向量能更具體地代表該文件之意義。由於敘述資料中，普遍存在許多詞彙對於敘述資料的整體意義並無太大影響，若剔除這些對敘述資料重要性小的詞彙，對敘述資料表現其意義時，並無太大影響，但卻可省去大部分運算量與使用空間。因此，良好的特徵選取程序，可以降低特徵詞數，提升效率，而不致大幅降低檢索效果，甚致還有可能因為過濾掉干擾檢索效益的雜訊詞(noisy term)，而提升檢索效果。

根據中研院詞性標記表，將詞性分為四十七種詞類，其中以名詞及動詞所代表的訊息最具意義且比例最高。另外，副詞常用在動詞和形容詞前面，表示程度、範圍、時間等，主要用作修飾形容詞和動詞，如：我***很***高興。我***常常***到那兒去。為了讓更貼近使用者需求的檢索結果排序在更前面，因此，副詞亦在我們分析的範圍內。所以本論文只保留詞性為名詞(Nouns)、動詞(Verbs)、形容詞(Adjective)與副詞(Adverb)的字詞，其餘字詞予以剃除。此外，亦將標點符號等不具有語意的詞剃除。圖 3-12 為圖 3-11 的內容去除標點符號及非名詞、動詞副詞的字詞之後的結果。

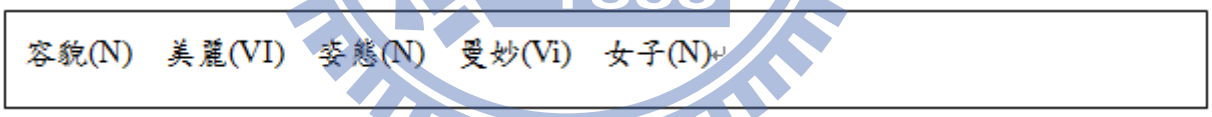


圖 3-12 前置處理實例－刪除停用字

特徵挑選方法的種類繁多，其中詞頻(TF)與反向文件頻率(IDF)之內積(TFIDF)[32]是滿常用的一種方式。另外，我們分析成語的釋義資料，平均一個成語的解釋大約只有 16.5 個字，扣除停用字及保留詞性為名詞、動詞、形容詞與副詞，最後統計出構成敘述資料的平均長度只有 8，也就是說大約只有 4 個詞彙，即表現敘述資料的向量之維度(Dimension)並不大，再者 Lucene 關鍵詞權重計算公式裡，包含了 TFIDF 的計算，加上目前教育部所收集的成語大約有四萬多筆，資料量也不算太大，因此我們選用詞性挑選及去除停用字來做為文件特挑選的處理。

3.2.8. 關鍵字合併

由於本系統在斷詞方面是採用中研院 CKIP 斷詞系統來進行處理，CKIP 在未知詞的偵測上具有相當的成效，對於專有名詞皆能有效斷出該詞彙，但對於新新人類用語的辨識上卻效果不佳，以字詞「美眉」為例：CKIP 系統將斷成「美(VH)」與「眉(Na)」的單字詞。「美眉」為一具有特殊代表性及意義的詞彙，相較於 CKIP 斷出的詞彙將因此失去其資訊含量，這樣的斷詞結果將影響後續資訊檢索模組中計算詞彙權重的誤差。中文的意義單位，在文言文中很多都是單字詞，所以在當時字(character)確實是意義單位；但在現代的中國語文當中，單字詞的型態已經很少，現代的中國語文慢慢發展以雙字詞為主的一種語言型態。

另外，否定副詞表示對情況、行為或性狀的否定的副詞，代表否定的意義。當使用者輸入一查詢問句「不高興」，CKIP 的斷詞結果為「不(ADV)」與「高興(Vi)」，若以此關鍵詞進行檢索將將會找到含「高興」字詞的文件，而造成檢索的誤差。常見的否定副詞有「不，弗，毋，勿，未，非，否」等。其中又以「不」為最常見，我們整理規則如下[33]：

1. 「不」「弗」。都表示一般的否定，「不」的用法較寬。可以否定動詞與形容詞，可以帶賓語或不帶賓語。「弗」在先秦時代只用在及物動詞的前面，而且不再出現賓語。
2. 「毋」「勿」。常用在祈使句中，表示禁止或勸阻。用法與「不」「弗」一樣。「勿」字後面的動詞帶賓語的非常少見。
3. 「未」。表示情況沒有出現。是對已然的否定。
4. 「非」。用於名詞性謂語前面。用在判斷句中。也可以表示對行為和性質的否定。

由以上規則得知，在特徵挑選的詞性裡，否定副詞只用在動詞、名詞前面，分析成語釋義的斷詞結果顯示，否定副詞接著動詞或名詞佔了 94%，其他 6%為副詞。

綜合以上我們得到一個結論，為了加強詞彙得鑑別力，針對單字詞以及否定副詞的部份，系統需要進行關鍵字合併。

3.2.9. XML 檔案索引建立

對 XML 檔作索引是這個系統很重要的工作，因一般使用者要透過查詢和檢索

XML 文件。底下介紹產生 XML 檔以又及建立索引的步驟：

1. 釋義資料萃取

由於教育部成語典及與重編國語辭典成語資料中，釋義欄位內容包含釋義、典源、典故以及舉例說明等的資訊內容，為避免這些資訊造成的雜訊干擾情況發生，因此，在進行斷詞之前，必須先將成語內容給予適當的處理，首先，藉由標點符號用法，將內容重新斷句，之後透過人工方式將涵義資訊挑選出來，並更新至資料庫。

2. 產生 XML 檔案

Solr 建立索引和查詢時得先進行字串的分詞，在向索引庫增加全文檢索類型的索引時，Solr 會先用空格進行分詞，然後把分詞結果依次使用指定的篩檢程式進行過濾，最後剩下的結果才會加入到索引庫中以備查詢。所以我們將萃取出來的成語釋義資訊，透過 CKIP 斷詞系統進行斷詞切字，並將結果儲存在系統中，接著利用特徵挑選的規則取出代表此成語文件的特徵詞，並對每個關鍵詞以空白隔開建立 XML 檔，如圖 3-13。

```
<add>
<doc>
<field name="idiomID">25731</field>
<field name="title">黑白顛倒</field>
<field name="interpretation">是非善惡錯亂不明。如：沒有證據怎可如此黑白顛倒，誣陷別人呢,</field>
<field name="interpretationKey">是非 善惡 錯亂 不明 沒有 證據 怎 可 如此 黑白 顛倒 誣陷 別人</field>
<field name="interpretationAlt">是非善惡錯亂不明。</field>
<field name="interpretationAltKey">是非 善惡 錯亂 不明</field>
<field name="filepath">/idiom/www/mandarin/fulu/dict/cyd/47/</field>
<field name="filename">cyd47496.htm</field>
</doc>
</add>
```

圖 3-13 成語詮釋資料格式

3. 索引建立

將需要索引的資料組裝成 XML 格式後，接著使用 HttpClient 工具程式將資料提交到 Solr 的 Http 介面，例如：http://localhost:8080/solr/update，也可以參考 Solr 安裝程式所包含提交 XML 檔建立索引的工具包 post.jar[9]來實現。本系統

[9] Solr 測試範例目錄位置 apache-solr-1.3.0/example/exampledocs

是利用 post.jar 發送 Http 請求給 Solr Web 建立索引[10]。

3.3. 查詢問句處理模組

進行檢索前需將查詢問句透過一連串的處理，取得代表此問句的關鍵詞組，再交由搜尋引擎進行檢索。系統處理查詢問句的程序包括：斷詞切字、詞性標示、去除標點符號、去除停用字、關鍵字合併、問句特徵挑選及關鍵詞擴展等步驟，其流程如圖 3-14。其中關鍵字合併與問句特徵挑選的處理與先前介紹前置處理的方法一樣，所以不再說明。

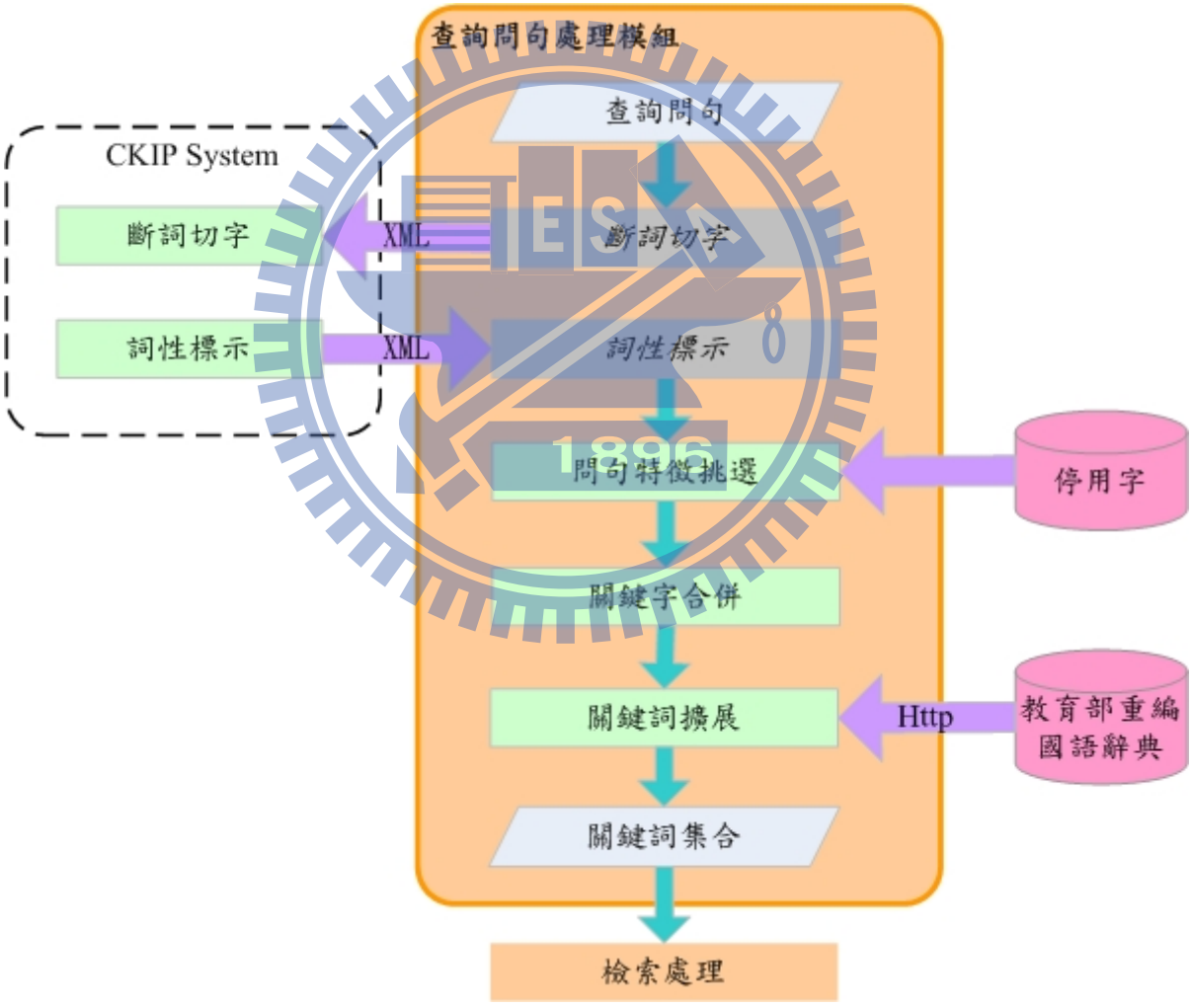


圖 3-14 查詢問句處理流程圖

[10] 建立索引檔指令 java -Durl=http://localhost:8080/solr/update -Ddata=files -jar post.jar *.xml

3.3.1. 關鍵詞擴展

MIRS 利用「教育部重編國語辭典修訂本」網站內容來建立詞彙同義辭典，該網站提供了查詢詞彙同義詞的功能，如圖 3-15 所示。在前置處理時，系統已儲存大量的詞彙，於是我們利用 HttpClient 及 Html Parser 程式包工具，對每一個詞彙送出查詢同義詞的 Request URL，並解析返回的網頁內容，取得同義詞資料，並儲存於系統中，其處理流程步驟介紹如下：

1. 利用 HttpClient 程式包工具送出查詢詞彙同義詞的 Request URL，如圖 3-16 所示。

2. 接著利用 Html Parser 程式包工具解析該網站返回的網頁內容，若找到關鍵字「相似詞」，則代表送出的詞彙在教育部重編國語辭典網站裡有同義詞的資料，如圖 3-15 所示，於是我們取出「相似詞」後面的同義詞資料。

3. 若是找不到關鍵字「相似詞」，但回應的網頁內容含有多筆查詢詞彙內容，如圖 3-17 所示，則以此詞彙為鍵值嘗試比對網頁內容的字詞是否等於關鍵詞，如圖 3-17 的第一筆資料「美麗」，解析網頁內容取得 Ukey 及 RecNo，若有取得 Ukey 及 RecNo，則重新送出 Request URL 如圖 3-18 所示，接著再執行步驟 2。

4. 最後利用標點符號將此詞彙中每一個同義詞拆解出來，並儲存到系統中建立同義詞典，供查詢擴展使用。

圖 3-15 教育部重編國語辭典詞彙相似詞查詢結果

```
http://dict.revised.moe.edu.tw/cgi-bin/newDict/dict.sh?idx=dict.idx&cond=關鍵詞
&pieceLen=50&fld=1&cat=&imgFont=1
```

圖 3-16 查詢詞彙相似詞 URL

圖 3-17 教育部重編國語辭典詞彙查詢結果

```
http://dict.revised.moe.edu.tw/cgi-bin/newDict/dict.sh?cond=關鍵字
&pieceLen=50&fld=1&cat=&ukey=Ukey&serial=1&recNo=RecNo&op=f&imgFont
```

圖 3-18 查詢詞彙相似詞 URL

取得代表查詢問句的特徵詞，接著從系統同義詞典中取出此特徵詞的同義詞，若詞典不存在此詞彙的同義詞，則利用前段介紹的方法，嘗試由教育部國語辭典取得同義詞資料。另外，該詞彙若執行過此步驟，則需加以註記，避免下次同一詞彙出現，再次執行而浪費系統資源。透過此關鍵詞擴展的處理，得到代表此查詢問句的關鍵詞集合，就可以送到文件檢索程序進行處理。

3.4. 檢索處理模組

文件檢索處理的程序包括：關鍵詞權重處理、轉換為 Solr 查詢語句等步驟，將此查詢語句提交給 Solr，讓 Lucene 進行檢索處理，最後將查詢結果返回給使用者，其流程如圖 3-19 所示。

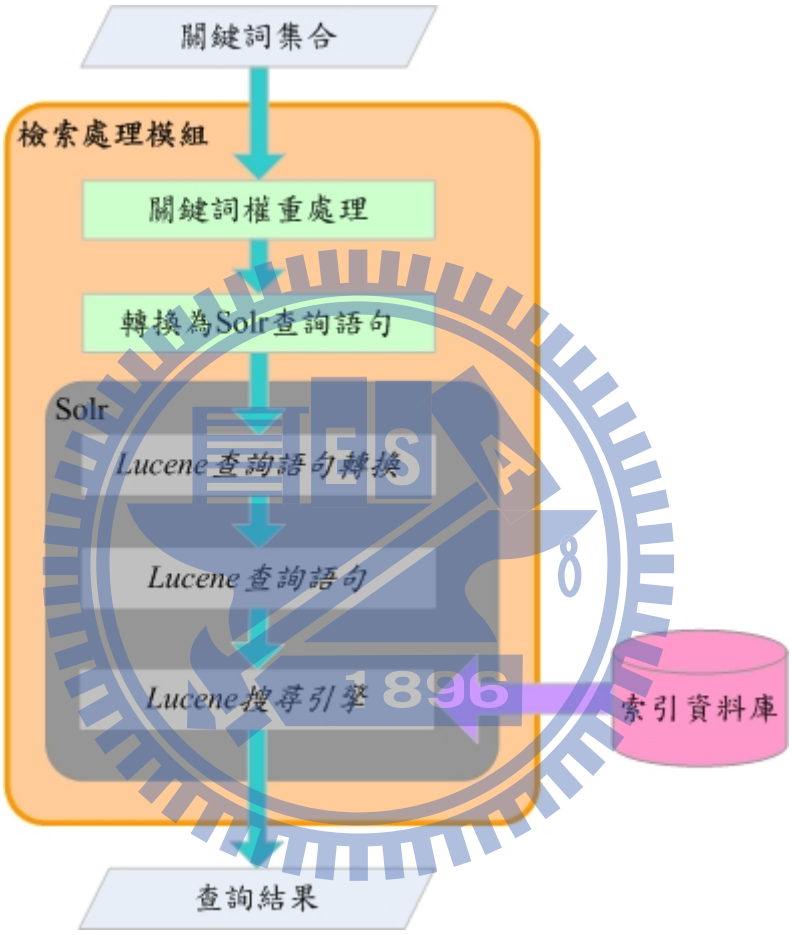


圖 3-19 文件檢索處理流程圖

3.4.1. 關鍵詞權重處理

除內置的得分演算法外，Solr 的檢索運算符號中提供了增加關鍵詞權重的方法，"^"符號可以控制相關度檢索。此作法的目的是將關鍵詞得分乘以"^"符號後的數值，以這個新數值作為關鍵詞在文件檔案中的得分，我們將在下一小節詳細介紹此用法。

當系統進行檢索時，文件內容的關鍵詞若跟查詢問句中的關鍵詞直接匹配，我

們希望此文件的排序在前面，而透過關鍵詞擴展才匹配到的文件相對的排在後面，所以系統在設計上提供了一個全域的設定，能針對查詢問句中的同義字權重乘上一個常數，這個常數是小於 1 的。另外詞彙的同義字也有有相似度之分，例如「女子」在教育部國語辭典網站中的相似詞有：「女人」、「女性」、「處女」、「女兒」等，其中「處女」這個詞相似度比較低，所以權重相對的應該要比較低，所以系統上提供設定同義字相似度的欄位，此欄位值需列為計算關鍵詞權重的因素。

3.4.2. 轉換為 Solr 查詢語句

在處理查詢問句時，Solr 提供了很多參數來擴展它自身的強大功能，當系統接收到使用者所需入查詢問句的關鍵詞集合和參數內容後，此程序會依照這些輸入內容組合成 Solr 查詢語句，並提交給 Solr 的 Http 介面進行檢索，表 3-1 為本系統用到 Solr 參數的使用方法。

首先假設資料欄位有：title、interpretationAlt、interpretationAltKey，預設是搜尋 title 這個欄位，如果剛好要搜尋的欄位就是 title，則就不需要指定搜尋欄位名稱。查詢規則(參數 q)內容如下：

1. 搜尋特定欄位(非預設欄位)

   在查詢詞前加上該欄位名稱加 ":"符號，例如：

   例如：title:國色 interpretationAlt:形容女子容美麗

2. 搜尋聯集結果

   在詞與詞間空格或加上大寫 "OR"，例如：

   interpretationAltKey:容貌 OR 相貌 OR 長相
   interpretationAltKey:容貌 相貌 長相

3. 搜尋交集結果

   在詞與詞間加上大寫"AND"或"+"，例如：

   +interpretationAltKey:美麗 + interpretationAltKey:漂亮

interpretationAltKey:美麗 AND interpretationAltKey:漂亮

4. 排除查詢

在要排除的詞前加上"-"號。例如:美麗 -醜陋，搜尋結果不會有包含醜陋的詞的結果在內。如底下代表搜尋結果不包含「男子」詞的結果。

美麗 -男子

5. 群組搜尋

使用"()"來包含一個群組。如希望搜尋在 interpretationAlt 欄位內同時有"女子"及"美麗"，例如：

(女子 AND 美麗)

6. 增加權重

如要搜尋"女子 容貌"，但因回傳太多筆資料內有"女子"或"容貌"的結果，想要把有包含"容貌"的資料往前排，可使用"^"符號在後面加上愈增加的權重數，例如"2"，則可以這樣做："女子 容貌^2"，會同時搜尋含有女子或容貌的結果，並把容貌這個詞加權，所以搜尋時會先判斷容貌這一個詞在搜尋結果中的比重，甚至一筆資料內"容貌"出現過兩次以上的就更加會有優先權。

表 3- 2 Solr 語法參數說明

參數	說明
q	查詢字串，必要參數。
version	代表 Solr 版本。
start	回覆第一筆記錄在完整找到結果中的偏移位置，0 開始，一般分頁用。
rows	指定回應結果最多有多少筆記錄，配合 start 來實現分頁。
indent	代表輸出的 xml 要不要縮行，預設為開啟 on。
sort	排序，格式：sort = <field name> + <desc|asc> [,<field name> + <desc|asc>]。範例：(score desc, title asc)表示先"score"降冪，再"title"升冪，系統預設為相關性降冪。

fl	表示索引顯示那些 field(*表示所有 field, score 是 Solr 的一個匹配程度)。
q.op	表示 q 中查詢語句中各條件的邏輯操作 AND、OR。
hl	是否高亮度。
hl.fl	高亮欄位。
hl.simple.pre	高亮前面的格式。
hl.simple.post	高亮後面的格式。
facet	是否啟動層面統計。
facet.field	層面統計的欄位。

3.5. 查詢結果顯示

3.5.1. 層面分類與修訂查詢

為了讓使用者能快速有效查詢或瀏覽資訊，將資訊做適當分類，是一可行的改善方向[34]。層面查詢功能比關鍵字檢索功能更具有檢索效益，因為透過層面的分析，並以目錄型式呈現，可以有效地引導使用者找出他想到的資料，使其不至於因不夠準確的資訊需求而迷失在浩瀚的資訊海中[35]。

我們將查詢結果文件中的關鍵詞加以統計、自動分類，提供給使用者挑選，進行修訂查詢，此外系統也保留查詢歷史，可瀏覽此次登入使用所有查詢紀錄，藉此瞭解檢索策略，方便使用者進行查詢字串與條件的調整，這些方法將協助使用者一步一步逼進所要查找的資訊。

3.5.2. 引進 Web 2.0 擴充同義詞

查詢擴展需要用到同義詞，雖然我們利用教育部國語辭典網站資料來建立同義詞典，但沒有任何辭典能包含所有中文字詞，尤其在現今資訊發達的世界裡，每天都有大量的中文新詞被創造出來，新詞在各個領域中不斷的出現，並且在文章中所出現的比例愈來愈高，甚至新詞往往是文章中最關鍵最重要的詞彙。為了擴充系統的同義詞，我們引進 Web 2.0 概念，讓使用者可以提供同義詞的建議資料。另外，本系統也允許使用者提供成語釋義資料建議的功能。

為了減少使用者提交次數，讓使用者與伺服器的互動性和親和性上有更大的提

升，系統設計時使用 Ajax 技術，讓使用者直接在查詢結果畫面上進行資料建議的操作。

四、 系統功能展示與評估

本章內容分為兩個部份，5.1 節介紹系統實作環境，其次展示系統實作結果；5.2 節分析目前主要的成語檢索系統功能，列出其優缺點。

4.1. 系統實作環境

本研究系統的實作環境如圖 4-1 所示，系統作業平台為 Windows Vista，使用 Jakarta Tomcat 6.0 作為處理 Servlet 和 JSP 的 Web Server 引擎，系統開發工具主要為昇陽公司的 Java JDK 1.5 和很多開源工具庫，並使用 Oracle 10g 為資料庫伺服器。

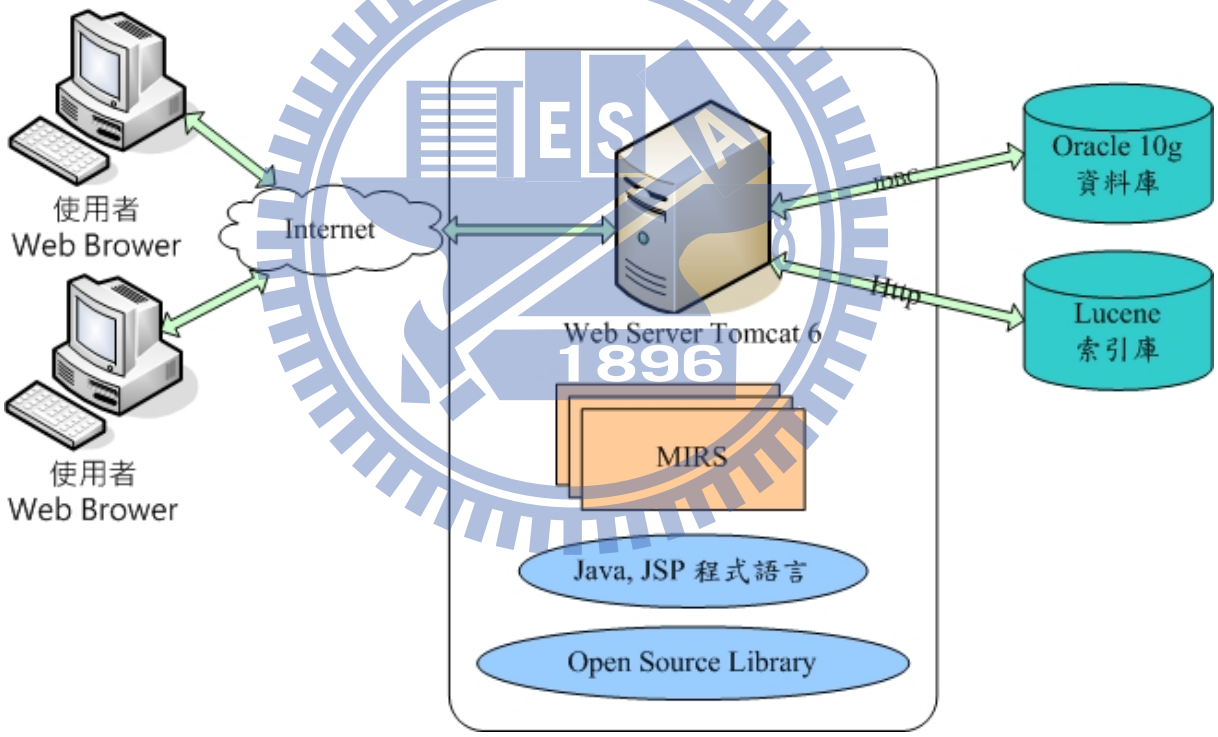


圖 4-1 系統實作環境圖

MIRS 因應網路使用環境特性選以 Java 搭配 Oracle 資料庫以及 JSP 程式為開發系統之作業環境，具有使用方便且相容性高的優點，因此不但可以使整個系統開發過程更為順暢，同時亦兼顧了後續維護的可行性。圖 4-1 中，使用者以瀏覽器進入 MIRS 系統，向系統提出資訊需求，經過程序處理，更新或取得後端資料庫資料。

4.2. 系統實作結果

MIRS 依據前面章節所描述的系統設計，結合資訊檢索的技術，實作出以成語涵義為基礎的檢索系統，MIRS 主要功能分為：檢索功能、管理者功能，以及成語料維護功能等三大部份，如圖 4-2 的所示。

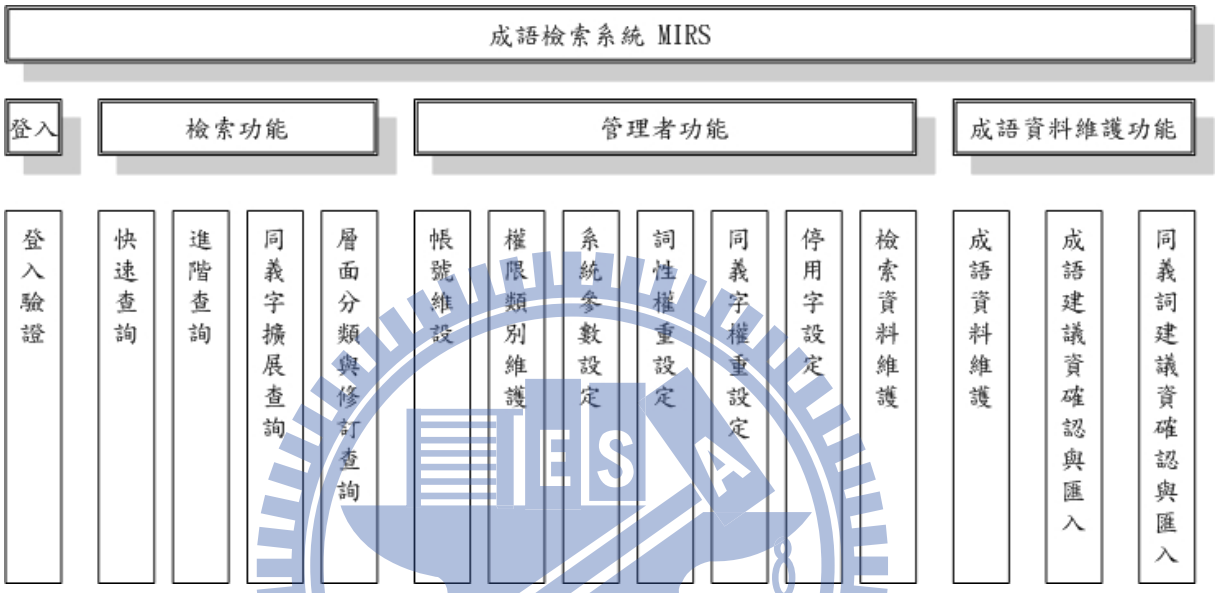


圖 4-2 系統功能圖

4.2.1. 檢索功能

使用者在遠端透過瀏覽器進入 MIRS，系統直接帶出快速查詢畫面，如圖 4-3 所示，網頁的內容分割為兩個區塊：功能選單、輸入條件區。系統主要提供的檢索功能包括：快速查詢、進階查詢、查詢擴展，以及層面分類查詢，說明如下：

圖 4-3 系統執行畫面

1. 快速查詢

(一) 輸入條件

系統提供了精確與模糊的查詢模式，精確查詢是將查詢問句中的關鍵詞用交集(AND)串連，模糊查詢則是用聯集(OR)。查詢條件提供了是否進行斷詞，以及是否進行同義詞擴展查詢的選項。查詢欄位則包括成語標題及釋義。這些查詢條件在此功能中均給定預設值，方便使用者直接輸入查詢語句即可，以加速檢索的效能。

(二) 查詢結果

圖 4-4 為查詢結果，畫面顯示出查詢問句轉換為 Solr 查詢語句的字串、關鍵詞權重，以及成語資料結果。系統也保留使用者先前的查詢策略方便修正檢索條件進行再次檢索，如圖 4-5 所示。

圖 4-4 查詢結果畫面

圖 4-5 查詢記錄畫面

(三) 提供建議資料

圖 4-6 為使用者提供同義詞建議的畫面。系統列出查詢字串的所有關鍵詞，讓使用者挑選想要針對那一個關鍵詞提供建議。圖 4-7 為提供成語解釋建議的畫面。使用者直接在查詢結果上進行操作，讓動作更快速，資料更直觀。

圖 4-6 提供同義字建議功能畫面

圖 4-7 提供成語釋義建議功能畫面

2. 進階查詢

圖 4-8 為進階查詢的功能畫面，除擁有快速查詢所有功能外，還可以下拉選單指定不同的邏輯操作運算元（and、or、not、+、-）和欄位來組合複雜的檢索條件、指定同義詞權重、每頁筆數、是否要高亮度顯示結果、將查詢結果依指定欄位降幕或升幕排序。

圖 4-8 進階查詢畫面

3. 查詢擴展

圖 4-9 為查詢「形容國家處境困難的時刻」的結果，透過畫面中的查詢字串可看出，系統將查詢問句中的關鍵詞進行擴展。

圖 4-9 查詢結果畫面

4. 層面分類查詢

圖 4-9 中，系統將查詢結果進行關鍵詞的統計與分類。此查詢預期得到成語為「國步維艱」，但結果並不符合期望，此時可利用層面分類查詢進行過濾，當勾選「國家」這個關鍵詞並執行 Filter，系統將列出包含「國家」這個關鍵詞的結果，我們會發現到成語「國步維艱」將排在第一位。

4.2.2. 系統管理功能

本研究所建置之系統管理者功能包括：帳號及權限維護、系統參數設定、詞性權重設定，同義詞權重設定、停用字設定、以及檢索資料維護，說明如下：

1. 帳號及權限維護

可以依不同角色，組合出不同的角色權限。例如圖 4-10 為管理者角色權限，圖 4-11 為成語專家角色權限，由圖中可看出左列的功能項目，會隨著不同角色

而展現出不同的選項。

圖 4-10 管理者角色權限

圖 4-11 成語專家角色權限

2. 系統參數設定

這些參數設定包括：Solr Web 介面運行的網址、Solr Home 主位置、斷詞系統服務位址、斷詞系統登入的密碼及密碼、特徵挑選所擷取的詞性、進行關鍵詞擴展的詞性，以及全域同義詞權重設定等，如圖 4-12 所示。

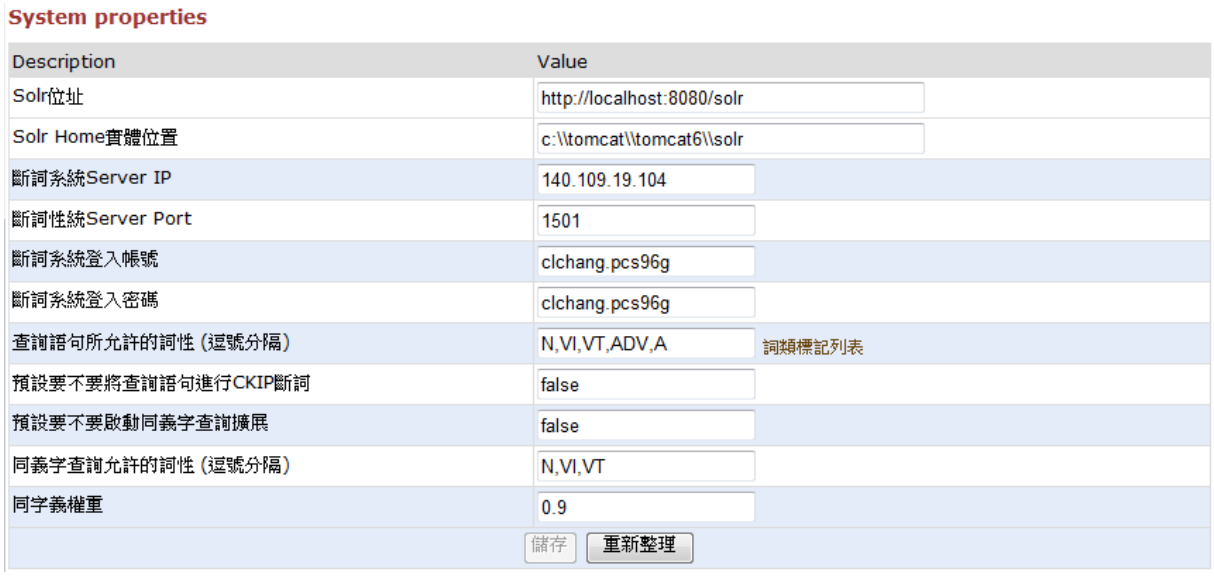


圖 4-12 系統參數設定畫面

3. 詞性權重設定

透過此功能，可依據查詢問句所擷取的詞性，給與不同權重的設定。

4. 同義字權重設定

透過此功能，可進行詞彙同義詞資料維護，以及權重設定，如圖 4-13 所示。

圖 4-13 同義詞權重設定畫面

5. 停用字設定

設定查詢問句以及建立索引將過濾掉的停用字，其功能畫面如圖 4-14 所示。前一章節提到查詢問句及文件內容的特徵挑選均要去除停用字，且 Solr 在索引及查詢會依指定的篩檢程式(StopFilter)進行過濾，其停用字以文件檔案形式儲存在特定的位置[11]。為了單一維護機制，當進行停用字的維護，則會同時更新 Solr 的停用字設定。

[11] Solr 停用字檔案位置 solr.home\conf\stopwords.txt

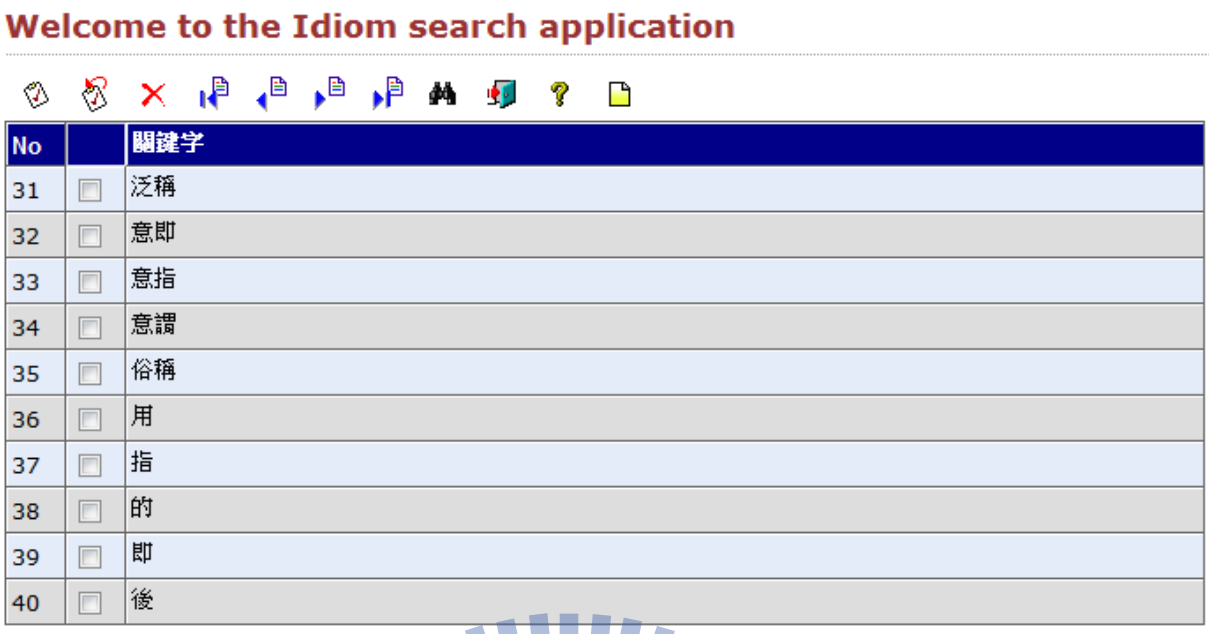


圖 4-14 停用字維護畫面

6. 檢索資料維護

索引資料維護的功能包括：建立索引檔、刪除索引檔、Commit 交易、Optimize 索引檔資料，如圖 4-15 所示。

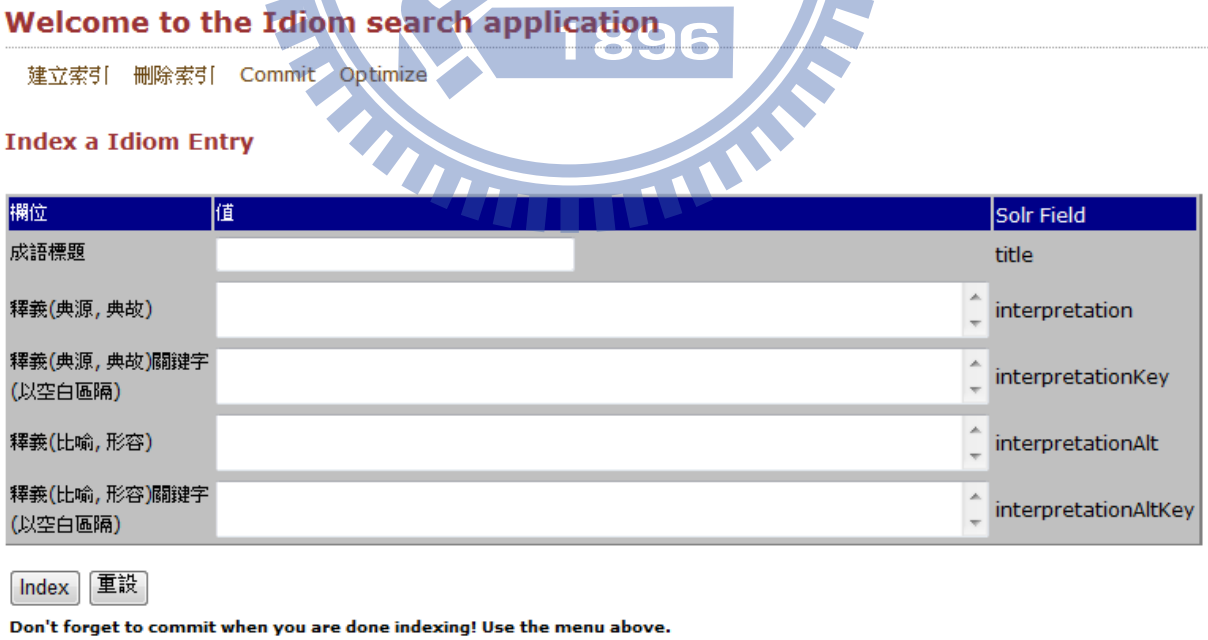


圖 4-15 索引維護功能畫面

4.2.3. 成語資料維護功能

本研究所建置之成語專家角色功能包括：成語資料的維護、建議資料確認與匯入，說明如下：

1. 成語資料的維護

若發現成語的解釋不恰當，除了可透過使用者的建議資料進行更新，管理者也可利用此功能更新成語的解釋，如圖 4-16 所示。當成語資料更新時，才能執行重新斷詞，接著才能執行重新建立索引的動作。

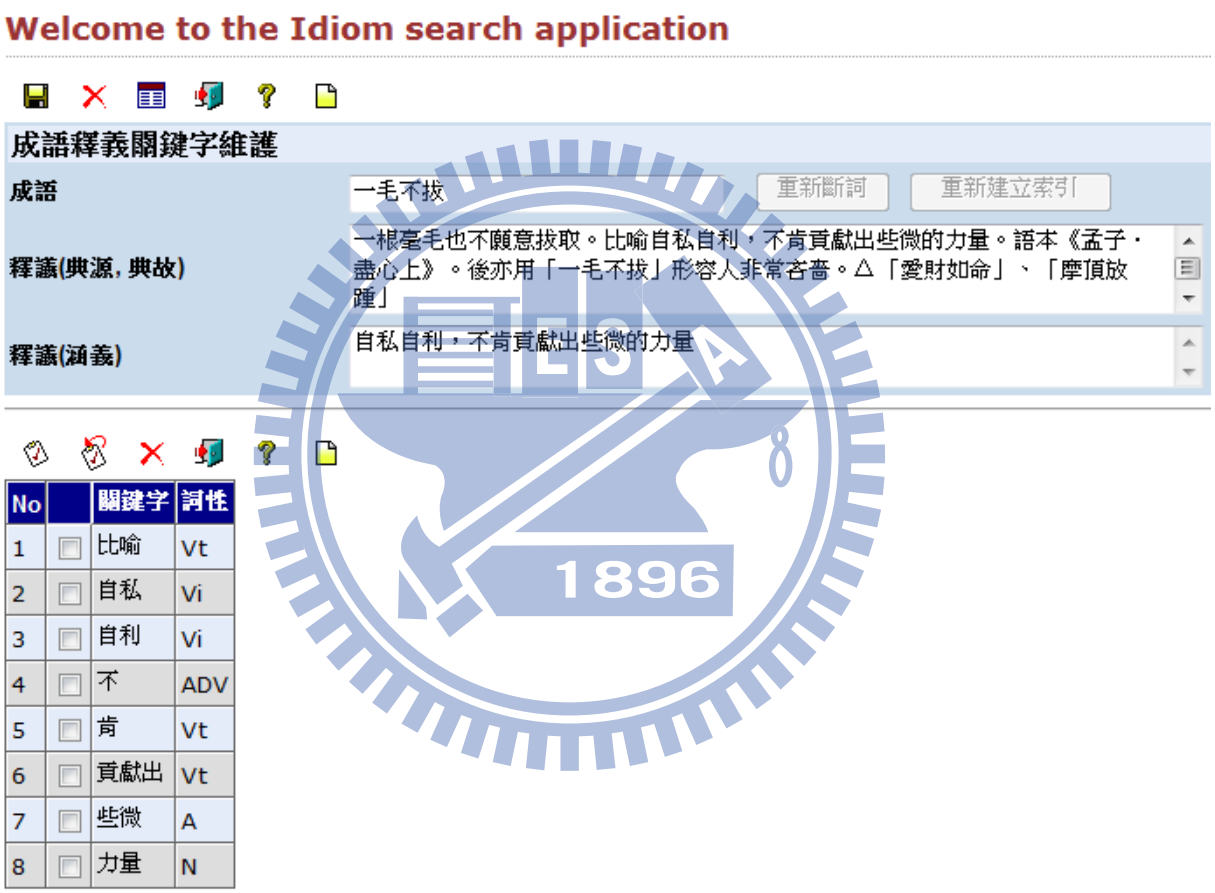


圖 4-16 成語資料維護畫面

2. 建議資料確認與匯入

為了避免不相關的建議資料直接進到系統而影響檢索的效益，使用者提供的建議資料會暫存在介面資料表裡，等待專業人員確認，執行匯入後才會真正進到系統裡，供檢索使用。建議的資料包括成語釋義資料與同義詞資料，圖 4-17 是關鍵詞「女人」同義詞的匯入結果。

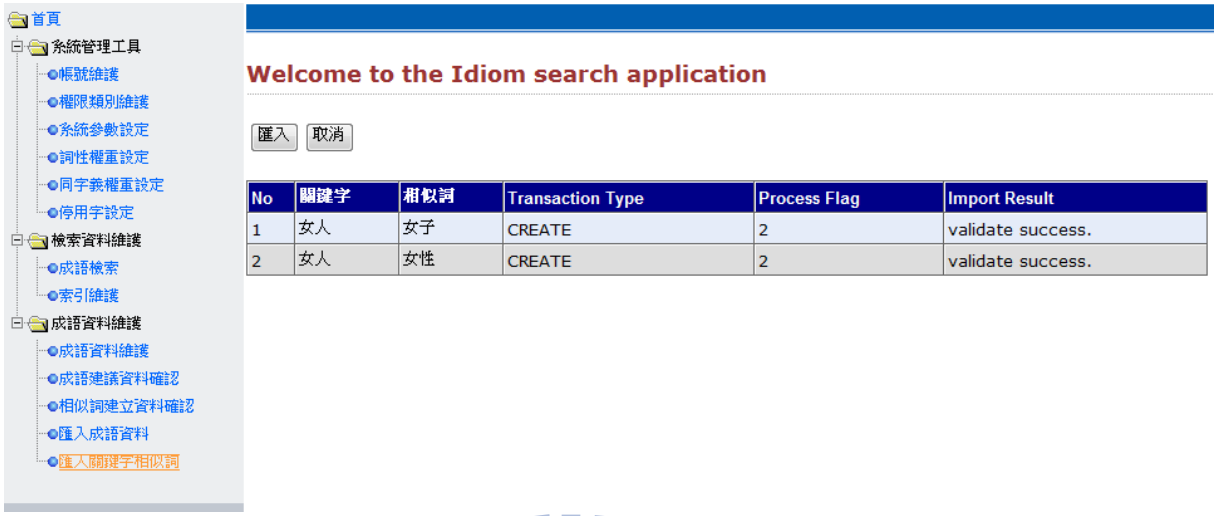


圖 4-17 同義詞匯入驗證畫面

4.3. 檢索效能評估

4.3.1. 效益評估方法

資訊檢索系統一般的效益評估方法多是以準確率(Precision Rate)與召回率(Recall Rate)來作為評斷的標準，準確率和召回率的定義為：

$$準確率 = \frac{為相關文件且被檢索出的文件數}{檢索出文件數}$$

$$召回率 = \frac{為相關文件且被檢索出的文件數}{總相關文件數}$$

由定義的式子可知，當準確率愈高，則代表所檢出的相關文件有越高的比率是正確的；當召回率愈高時，代表此系統能搜尋出愈多相關的文件。

4.3.2. TopN 篇準確率

本研究所使用之文件為教育部成語典，我們擷取含有釋義資料的成語，共計 22309 筆。為了有效進行效能的評估，實驗用的查詢問句均經特別設計。我們將以口語化問句來表示查詢的主題，以充分表達使用者資訊需求的二至五個檢索詞彙，

來建構檢索實驗之原始查詢問句。為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，我們依據教育部成語典的類別以及實驗文件資料庫中文件類別分佈情形挑選特定檢索主題，再以兩個以上的詞彙表達檢索主題，組成一原始查詢問句。表 4-1 為原始查詢問句。

表 4-1 原始查詢問句

編號	主題類別	查詢問句
1	行動	比喻人的言行與舉動
2	生活	形容生活富裕
3	時間	形容聲勢驚人
4	景象	眼前景象引發內心情緒
5	時候	比喻到了關鍵時刻
6	行為	形容手法、技巧高明
7	才智	形容才華洋溢
8	文章	形容文章內容生動有趣
9	言語	形容非常清楚明白
10	時機	比喻掌握機會
11	身分	形容人身分卑微
12	品德	比喻人格高尚品行高潔
13	心理	形容心理擔憂恐懼
14	情感	形容情感深厚
15	自然	比喻時機成熟，事情自然成功
16	處境	形容處境險惡
17	社會	比喻社會風俗習慣
18	風景	形容風景優美秀麗
19	美麗	形容容貌美麗的女子
20	情勢	形容事情很危急
21	工作	形容努力完成任務
22	勢力	比喻實力相當，不相上下
23	思想	形容人精神恍惚不清的樣子
24	聲音	形容美妙的聲音

25	影響	形容身心痛苦疲憊
26	教育	比喻學識淵博的人
27	人物	形容優秀傑出的人才
28	計謀	所有計謀與方法都用盡了
29	意志	形容人意志堅強
30	數量	比喻數量很多
31	事理	形容內心的感受
32	政治	比喻壞人橫行作惡
33	狀態	形容令人震驚
34	軍事	比喻完成事前準備工作
35	官場	形容貪汙受賄，破壞法紀
36	文教	對於事理能領悟通曉
37	做事	形容做事謹慎周密

由於本研究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，因此無法計算召回率。故本研究在取得檢索結果之相關候選文件之後，以人為判斷檢索出的文件是否與查詢問句相關，再計算準確率做為檢索效益評估的依據。本研究以成語解釋相關為相關判斷之原則，採用多位相關判斷者執行相關判斷。評估方法是將檢索實驗後之檢索結果，由使用者挑選主觀認定為相關的文件「檢出且相關文件數量」，再依此計算各檢索結果的準確率。另外使用者只會查看排名排 TopN 篇文件，通常只看 Top10 名的文件，因此我們非在意 Top10 篇文件是否顯著改善檢索效果。我們將以查詢擴展前後結果之 Top10、Top20、Top30 篇文件的準確率來觀察改善情況。

表 4-2 為本實驗查詢擴前後查詢結果的精確度，表格中編號為查詢問句編號，QE 為是否有進行查詢擴展，最後一列為平均精確度。圖 4-18、圖 4-19 及圖 4-20 分別為 Top10、Top20 及 Top30 進行查詢擴展前後的精確率比較，我們可以發現，在 Top10、Top20、Top30 我們利用查詢擴展分別改善了 8%、11%、9%的精確度，實驗結果表明，本文所利用的查詢擴展方法是有效的。

表 4-2 檢索結果精確度

	Top10 Precision	**Top20 Precision**	**Top30 Precision**

編號	QE = N	QE = Y	QE = N	QE = Y	QE = N	QE = Y
1	0.82	0.84	0.64	0.64	0.57	0.75
2	0.98	0.88	0.63	0.63	0.5	0.75
3	0.92	0.94	0.84	0.84	0.88	0.91
4	0.8	0.84	0.86	0.86	0.83	0.87
5	0.5	0.74	0.39	0.39	0.37	0.45
6	0.76	0.96	0.63	0.63	0.64	0.73
7	0.58	0.82	0.67	0.67	0.72	0.75
8	0.7	0.72	0.56	0.56	0.51	0.43
9	0.86	0.76	0.82	0.82	0.77	0.73
10	0.3	0.86	0.34	0.34	0.34	0.57
11	0.9	0.92	0.57	0.57	0.49	0.68
12	0.66	0.92	0.62	0.62	0.67	0.81
13	0.88	0.84	0.69	0.69	0.79	0.84
14	0.64	0.96	0.41	0.41	0.33	0.49
15	0.56	0.58	0.61	0.61	0.54	0.54
16	0.74	0.78	0.65	0.65	0.67	0.65
17	0.9	0.86	0.84	0.84	0.62	0.86
18	0.98	0.9	0.82	0.82	0.65	0.88
19	0.92	0.96	0.95	0.95	0.9	0.87
20	0.34	0.58	0.26	0.26	0.26	0.43
21	0.52	0.64	0.4	0.4	0.39	0.48
22	0.94	0.94	0.68	0.68	0.54	0.71
23	0.92	0.92	0.86	0.86	0.77	0.75
24	0.56	0.64	0.6	0.6	0.48	0.47
25	0.64	0.88	0.63	0.63	0.69	0.56
26	0.96	0.94	0.88	0.88	0.86	0.91
27	0.86	0.82	0.78	0.78	0.71	0.75
28	0.96	0.72	0.66	0.66	0.6	0.55
29	0.72	0.94	0.61	0.61	0.63	0.94
30	0.6	0.8	0.67	0.67	0.69	0.73

31	0.98	0.94	0.89	0.89	0.83	0.79
32	0.88	0.72	0.73	0.73	0.55	0.73
33	0.66	0.82	0.41	0.41	0.39	0.75
34	0.3	0.38	0.29	0.29	0.29	0.3
35	0.64	0.74	0.59	0.59	0.55	0.57
36	0.68	0.9	0.62	0.62	0.57	0.71
37	0.84	0.86	0.82	0.82	0.73	0.71
平均	**0.74**	**0.82**	**0.65**	**0.76**	**0.6**	**0.69**

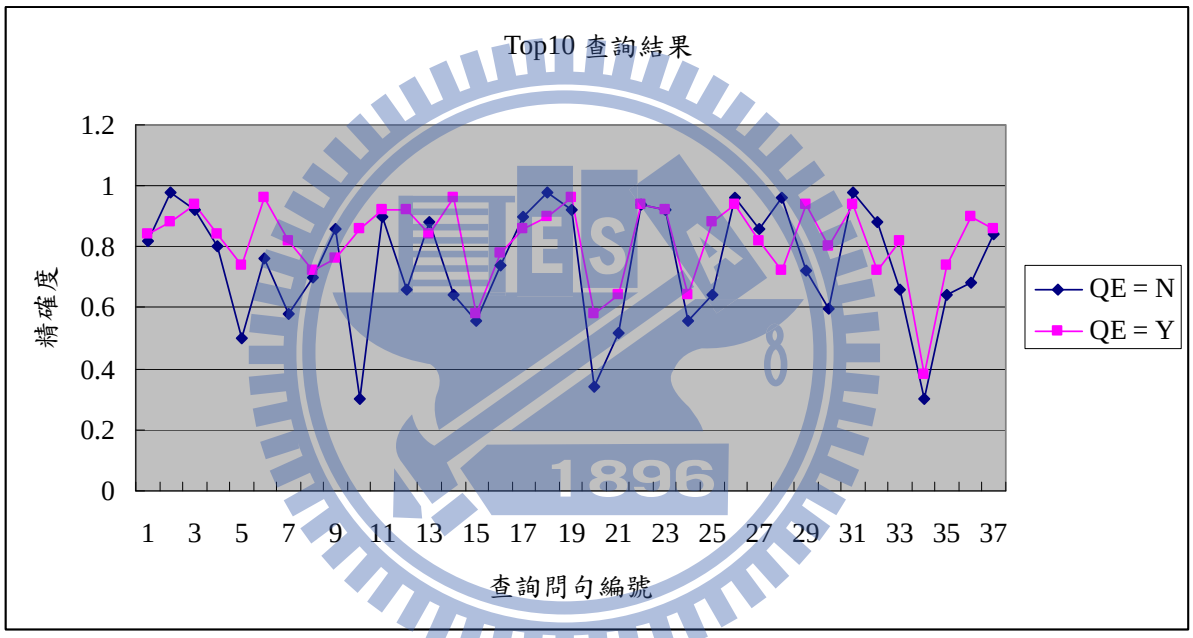


圖 4-18 Top10 查詢擴展前後精確率比較

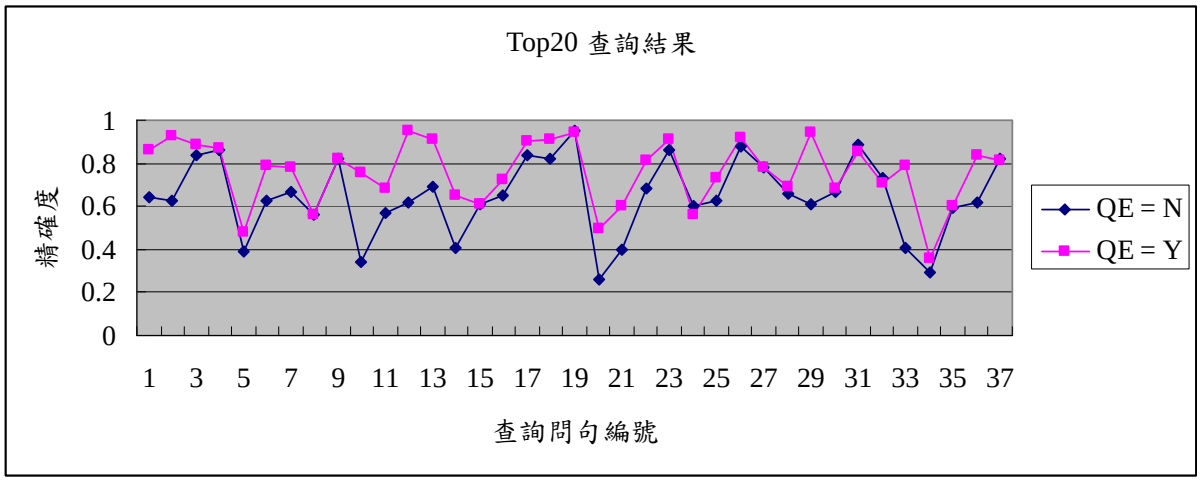


圖 4-19 Top20 查詢擴展前後精確率比較

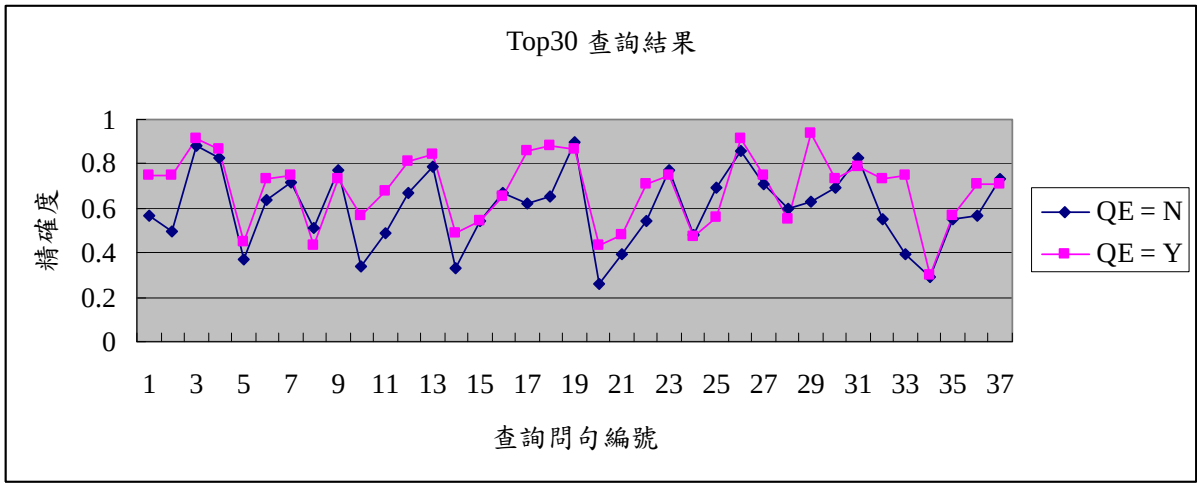


圖 4-20 Top30 查詢擴展前後精確率比較

4.4. 系統功能評估

我們將 MIRS 與目前兩岸最廣泛使用的成語檢索系統「教育部成語典」與「漢典」相比較，優點可以分成以下三個部份：查詢功能、查詢結果、以及動態地建立同義詞典。如文獻回顧所述，向量檢索模式，可以依重要性排序，目前只有 MIRS 運用此模式設計。此外，檢索系統是否能夠提供修訂查詢，甚至加入相關回饋以及整合不同查詢功能的特色，運用前次的檢索結果，逐步改善檢索成效，也是影響成語檢索系統設計的重點之一，因此列入以下分析。各系統功能分類的結果如表 4-2 所示。

表 4-3 成語檢索系統功能分類表

	檢索模式	查詢功能
MIRS	向量模式	口語化輸入、層面分類、修訂查詢、查詢擴展、權重調整
成語典	布林邏輯模式	單一關鍵詞查詢、類別查詢
漢典	布林邏輯模式	單一關鍵詞查詢、部首查詢、拼音查詢

4.4.1. 查詢功能

1. 可輸入口語化查詢問句

為了更能讓系統親近於使用者，更進一步邁向人機化介面的操作功能，因此 MIRS 設計口語化的查詢輸入，使用者可以用慣用的語氣、詞句輸入要查找成語的口語化描述問句或片語即可進行檢索，提供了在搜尋條件上幫助使用者更快、更方便的輸入。

2. 以層面分類進行檢索

如果資訊檢索系統提供多次檢索的功能，使用者可以針對先前檢索的結果，修改查詢資料，將結果中較符合需求的結果調整到較前的次序，提高檢索的準確性。MIRS 利用層面分類將查詢結果中的關鍵詞加以統計與分類，讓使用者可針對查詢結果中感興趣之關鍵字再做進一步的查詢，也可修改前次的查詢條件，嘗試使檢索的結果愈來愈符合使用者的需求。

4.4.2. 查詢結果

1. 查詢問句擴展

「教育部成語典」與「漢典」均是透過提供搜尋引擎的站台，以單一關鍵詞「全文檢索」的方式，將符合使用者輸入之資料搜尋出來；然而，單單以搜尋符合關鍵字的資料回應將會遺失許多資訊，例如，若使用的索引詞彙並非檢索詞彙而是檢索詞彙的同義詞，則該文件將無法被檢索出來。一個可能的解決方案即是擴展查詢問句，目前也只有 MIRS 運用此技術來擴展使用者查詢詞彙，以增進檢索效益。

2. 以字詞權重為基礎的檢索

除了擴展查詢詞彙，MIRS 也運用詞性加權的判斷機制，依不同的詞性給予不同程度加權，使用者也能自行決定同義詞權重，將檢索的結果依文件的重要性排列以達到較佳的檢索效益。

4.4.3. 動態地建立同義詞典

索引典及同義詞典是提升資訊檢索系統效能的重要資源，MIRS 在處理查詢問句

時，利用教育部國語辭典動態的取得同義詞資料，透過這樣的機制，可以動態地建構適用於 MIRS 的同義詞典。此外，也引進 Web 2.0 概念讓使用者提供同義詞的建議，籍此以網羅現代重要之詞彙。

五、 結論與未來工作

本章總結整篇論文並提出未來可能的研究方向。5.1 節主要提出本研究的結論及對成語檢索所帶來的貢獻；5.2 節則是在論文研究過程中，受限於本身或環境的限制而無法完全討論的相關議題，由系統的角度，提出不足或是可擴充之處做為未來的研究工作。

5.1. 研究結論與貢獻

本論文針對現存的成語檢索系統作了一番比較，說明其查詢功能效益缺失之處，利用資訊檢索技術結合中研究 CKIP 斷詞系統，建置一套基於成語涵義的檢索系統以求改進，讓使用者能以符合自然語言習慣的口語化方式來描述資訊需求，相較其他已知的成語檢索系統更具親和力。系統依據使用者提交的查詢問句擷取關鍵詞組，進查詢行擴展，並依據詞性、同義詞給與不同權重，提高了檢索效益。另外，系統提供查詢結果的關鍵詞統計與分類，讓使用者透過層面分類查詢與修訂查詢功能來修改提交的查詢條件，以便獲得更加貼近所需的文件資源。如此，將可快速的找到所需資料，有效減少成語查找時間，不僅僅有管理和搜尋功能，其對於推廣、發揚，甚至教學研究，都存在者極大的幫助。

總結來說，本研究最主要有下面三點貢獻：

1. 首創第一個提供可輸入類自然語言口語化查詢語句的成語檢索系統。
2. 可針對文件評分機制進行權重調整，以提高檢索的精準度。
3. 找出可透過「教育部重編國語辭典」取得同義詞來進行查詢擴展的方法。

5.2. 未來工作

總結目前的研究結果與心得，有很多相關的議題仍有待進一步研究。在評估過程中，我們也發現系統檢索效能仍然有許多可改進的地方。對於本研究之未來發展與應用可朝以下方向進行：

1. 考慮文件其他特徵

特徵詞的篩選方法很多，本系統依據停用詞、詞性來過濾詞彙。其他像文件篇數(document frequency, DF)、詞頻與反相篇數(TF x IDF)等來過濾詞彙，都是值得探討的議題。例如篇數過少的詞，只出現在一、兩篇文件上，這類的詞卻很多。刪除這些詞，可以大幅降低特徵詞數。另外，雖然大部份的研究都著重在文件的關鍵字，但文件特徵除了組成內容的字詞外，尚有很多值得研究與探討的議題，應該可以挖掘出更多描述此文件涵義的特徵。

2. 關鍵詞查詢擴展與權重

系統已設計關鍵字同義詞權重的欄位，可進行此權重值該如何賦予。不同的詞在不同的查詢問中的重要性並不相同，因此，查詢擴展還要考慮的另外一個問題是擴展之後各個詞在新查詢中的權重分配，另外也可以再利用索引典加權與擴展，以便提高檢索效益，儘量地返回給用戶端相關度更高的檢索結果，以此控制由於查詢擴展而造成的檢索結果過多的問題。

3. 文件得分演算法的改進

系統查詢結果文件的評分是利用 Apache Lucene 程式庫的演算法，未來期望能找出一個更適合應用在成語的資料探勘演算法，整合出一個更為適用的成語檢索系統，以進一步提升檢索的精確度與強健性(Robustness)，在未來的研究上都是重要的課題之一。

4. 結合語意分析處理

近年來，語意分析在文件擷取上也有廣泛的研究，並且對於檢索效能有顯著的改善。由於本研究僅處理針對查詢問句進行詞彙處理，造成無法對文件做更深一層的語意分析，建議在未來可結合本體論(Ontology)架構，透過語意文法的分析及詞彙的從屬關系，從隱藏在結構資料尋找更高層次的語意關鍵字，找出具有代表性語意關聯性，以及相似度具有代表關鍵字詞，用來輔助字詞上語意判斷，此做法勢必可提升檢索的精確度，此外，在自然語言理解方面，目前的系統並未具備推理能力，在許多情況下，詞語的組合可能引申另外的意義。這些會遭遇到但仍無法解決的問題，有待未來持續地研究。

參考文獻

[1] N. J. Belkin and W. B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?", Communications of the ACM, Vol. 35, No. 12, pp. 29-38, 1992 December.

[2] 黃雲龍，張佑任，「中文全文資訊檢索之效能評量初探」，南華大學資訊管理學研究所，資訊管理研究學刊，民國 91 年。

[3] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer", Addison-Wesley Pressed, 1989.

[4] Christos Faloutsos, "Signatures Files", In Information Retrieval: Data Structures and Algorithms, pp. 46-65, 1992.

[5] 陳光華，「數位圖書館中權威控制系統的設計」，政治大學圖書與資訊學刊，34 期，頁 51-71，民國 89 年 8 月。

[6] Sparck Jones, K. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", Journal of Documentation, Vol. 28, No. 1, pp. 11-21, 1972.

[7] 曾元顯，「新一代資訊檢索技術在圖書館 OPAC 系統的應用」，大學圖書館，1 卷 3 期，頁 82-93，86 年 7 月。

[8] Seerra, J., "The Boolean model and random sets," Computer Graphics and Image Processing, Vol.231, No.12, pp. 99-126, 1980.

[9] Salton, G., "Developments in automatic text retrieval", Science, Vol. 253, pp. 974-979, 1991.

[10] Robertson, S. E. and K. Sparck Jones, "Relevance weighting of search terms", Journal of the American Society for Information Sciences, Vol. 27, No. 3, pp. 129-146, 1976.

[11] D. Kuropka, "Modelle zur Repräsentation natürlichsprachlicher Dokumente - Information-Filtering und -Retrieval mit relationalen Datenbanken", In series: Advances in Information Systems and Management Science, 10th issue, Logos Verlag, Berlin, 2004, ISBN 3-8325-0514-8.

[12] http://en.wikipedia.org/wiki/Information_Retrieval, Information Retrieval, available at 2009.09.23.

[13] Lee-Feng Chien and Hsiao-Tieh Pu, "Important Issues on Chinese Information Retrieval", Computational Linguistics and Chinese Language Processing, Vol. 1, No. 1, pp. 205-221, 1996 August.

[14] 曾元顯，「關鍵詞自動擷取技術與相關詞回饋」，中國圖書館學會會報，59 期，頁 61-62，民國 86 年。

[15] 卜小蝶，圖書資訊檢索技術，文華圖書館，台北市，民國85年。

[16] 卜小蝶，「Interne 資源收集與整理的方法探討」，資訊傳播與圖書館學，2 卷 1 期，頁 78-79，民國 84 年 9 月。

[17] 陳光華、莊雅蓁，「資訊檢索之中文詞彙擴展」，資訊傳播與圖書館學，8 卷 1 期，頁 59-75，民國 90 年 9 月。

[18] 陳光華、莊雅蓁，「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，頁 93-107，90 年。

[19] Claudio Carpineto, Glovanni Romano and Vittorio Giannini, "Improving Retrieval Feedback with Multiple Term-Ranking Function Combination", ACM Transactions on Information Systems, Vol. 20, No. 3, pp. 259-290, 2002 July.

[20] http://morris.lis.ntu.edu.tw/wikimedia/index.php/Relevance_feedback_相關回饋, Relevance feedback 相關回饋, available at 2009.09.23.

[21] 李剛、征服Ajax+Lucene構建搜索引擎，文魁，台北市，民國95年。

[22] http://lucene.apache.org/java/docs/index.html, apache lucene, available at 2009.09.23.

[23] http://lucene.apache.org/java/2_0_0/api/org/apache/lucene/search/Similarity.html, Lucene Similarity, available at 2009.09.23.

[24] http://lucene.apache.org/solr/, apache solr, available at 2009.09.23

[25] http://zh.wikipedia.org/wiki/AJAX/, ajax, available at 2009.09.23.

[26] http://www.dom4j.org/, dom4j, available at 2009.09.23.

[27] http://hc.apache.org/httpclient-3.x/, httpclient, available at 2009.09.23.

[28] http://htmlparser.sourceforge.net/, html parser, available at 2009.09.23.

[29] http://wiki.apache.org/solr/Solrj/, solrj, available at 2009.09.23.

[30] http://zh.wikipedia.org/wiki/MVC, MVC available at 2009.09.23.

[31] http://ckipsvr.iis.sinica.edu.tw/, 中研院中文斷詞系統。

[32] K. S. Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", Journal of Documentation, Vol. 28, No. 1, pp. 11-20, 1972.

[33] http://web2.tcssh.tc.edu.tw/school/guowenke/books/gh/new_page_11.htm, 台中二中全球資訊網。

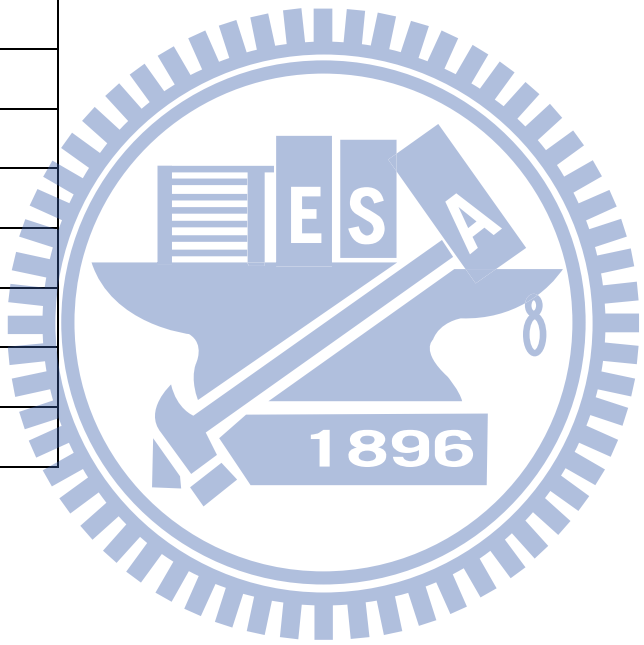
[34] Samler, S. and Lewellen, K, "Good taxonomy is key to successful searching", EContent, Vol. 27, No.7/8, S20, 2004.

[35] 林雯瑤，「層面分類的概念與應用」，教育資料與圖書館學，44卷1期，頁153-171，民國95年。

附錄一 「簡化詞類」與「精簡詞類」對照

簡化詞類	精簡詞類
A	A
Caa	C
Cab	POST
Cba	POST
Cbb	C
D	ADV
DE	T
Da	ADV
Dfa	ADV
Dfb	ADV
Di	ASP
Dk	ADV
FW	FW
I	T
NAV	NAV
Na	N
Nb	N
Nc	N
Ncd	N
Nd	N
Nep	DET
Neqa	DET
Neqb	POST
Nes	DET
Neu	DET
Nf	M
Ng	POST
Nh	N

SHI	Vt
T	T
VA	Vi
VAC	Vt
VB	Vi
VC	Vt
VCL	Vt
VD	Vt
VE	Vt
VF	Vt
VG	Vt
VH	Vi
VHC	Vt
VI	Vi
VJ	Vt
VK	Vt
VL	Vt
V_2	Vt



附錄二 中研院平衡語料庫詞類標記集

簡化標記	對應的CKIP詞類標記[12]	
A	A	/*非謂形容詞*/
Caa	Caa	/*對等連接詞，如：和、跟*/
Cab	Cab	/*連接詞，如：等等*/
Cba	Cbab	/*連接詞，如：的話*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
Da	*Daa*	/*數量副詞*/
Dfa	Dfa	/*動詞前程度副詞*/
Dfb	Dfb	/*動詞後程度副詞*/
Di	Di	/*時態標記*/
Dk	Dk	/*句副詞*/
D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Neu	*Neu*	/*數詞定詞*/.
Nes	*Nes*	/*特指定詞*/
Nep	*Nep*	/*指代定詞*/
Neqa	*Neqa*	/*數量定詞*/
Neqb	*Neqb*	/*後置數量定詞*/
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
I	I	/*感嘆詞*/
P	P*	/*介詞*/
T	Ta, Tb, Tc, Td	/*語助詞*/

[12] 斜體詞類，表示在技術報告#93-05中沒有定義，即後來增列的。

VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/
VI	VI1,2,3	/*狀態類及物動詞*/
VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/
DE	/*的，之，得，地*/	
SHI	/*是*/	
FW	/*外文標記*/	