結合線上翻譯服務的跨語言專利檢索系統

鄧舜元　Shun-Yuan Teng
華梵大學資訊管理學系暨研究所
Department of Information Management
Huafan University
wells0609@gmail.com

邊國維　Guo-Wei Bian
華梵大學資訊管理學系暨研究所
Department of Information Management
Huafan University
gwbian@cc.hfu.edu.tw

摘要

本論文提出一個結合線上翻譯服務網站的跨語言專利檢索系統(Cross-language Patent Retrieval System)，利用適合處理不同語言的 bi-gram 索引方法，透過檢索引擎處理多語言的專利文件集，並結合網路翻譯服務系統，利用查詢翻譯的方法，將原始的查詢加以翻譯，再進行專利檢索。本系統進一步利用檢索結果，根據專利分類體系之中的 International Patent Classification (IPC)分類，可將專利文件相關的 IPC 分類列出。

目前本系統可以處理英文與日文的單語專利檢索、以及日文檢索英文專利文件與英文檢索日文專利文件兩種跨語言檢索。使用者可以選擇查詢集來源，編輯與修改查詢，選擇三種不同的查詢集翻譯方式，選擇三種不同的翻譯服務網站進行查詢翻譯，並且可以選擇檢索的欄位及專利文件集的種類，進行跨語言的專利檢索。

Abstract

In this paper, we introduced a cross-language patent retrieval system which combined the various free web translators on the internet. The bi-gram indexing method was used to deal with the multilingual patent documents, and the query translation method was used to translate the query form the source language to the target language.

Currently, this system provides the functions of the monolingual and cross-language patent retrieval in English and Japanese. The users can input the queries and use the different translation systems to process the query translation. The different fields of the query topics and various patent document sets are selected to perform the cross-language patent retrieval from Japanese to English, and vice versa.

關鍵詞：跨語言檢索、專利檢索、Bi-gram、查詢翻譯

Keywords: Cross-Language Patent Retrieval, Bi-gram, Query Translation.

一、緒論

專利文件是極為重要的科技訊息來源，長期以來一直受到研發者或企業經營者的重

視。專利文件是目前唯一完全公開技術並能使用法律來保障專利發明人權益的一種方式，正因為其揭發技術方法能迅速反映最新科技動態及研究成果，因此專利的質與量是目前衡量國家創新能力的重要指標，由於產業界的國際競爭越趨激烈，全球企業莫不積極藉由專利的保護，維持技術領先優勢與市場利益。

根據世界智慧財產權組織（World Intellectual Property Organization, WIPO）[1]報導，專利文件包含全世界 90%～95%之研發成果，而其它的技術文件（論文或期刊等）中只僅含 5%～10%之研發成果，STN International [2]也指出有 70%~90%的專利資訊，根本沒有在其他的期刊或者雜誌發表過。此外 WIPO 還指出在研究工作中若能善於應用專利文件的話可以得到縮短 60%研發時程、同時減少 40%研發經費之效益。因此，閱讀與分析專利文件成為極為重要而且為不可或缺的一項工作，而使用專利檢索是分析專利文件中極為重要的一環，因為如果檢索出來的結果不正確，那麼依照錯誤的結果所做的分類、分析以及所有的數據、圖表等，都會無法正確的反應出隱含在專利文件中的知識，由此可知專利檢索對於企業或研發者都是很重要的一項工作。

當企業或研發者在開發新產品或申請專利時候，一定會先檢索目前的現況，才能預先知道已存在的研究有哪些、驗證產品開發計畫是否有誤、是否重複研發、是否抵觸他人之專利侵權，這樣才可以節省金錢和時間上的浪費，並能有效的推展研究發展的工作。

由於專利係採屬地主義，如果專利要受到保障的話，就必須要和各國來申請專利，當企業或研發者在檢索專利的時候，期望可以使用同樣的關鍵字來檢索美國，日本，跟台灣等國的專利，並取得相關專利文件資訊，不過目前大部分的專利檢索系統並不提供跨語言的檢索方式，所以使用者必須以三個不同語言，分別到三個不同的系統查詢，才能找到所有相關的資料。但問題是，並不是所有的使用者都具備足夠的語言能力，可以使用不同的語言來檢索專利，所以如果一開始能在界定的資料範圍內，提供了涵蓋兩種以上的語言，那麼系統就可以成為一個跨語言的專利檢索系統，讓使用者使用自己的語言，也可以檢索到英文或日文的專利。

本研究提出的跨語言專利檢索，其目的結合網路各種免費的翻譯資源，開發一個能處理多個語言的跨語言專利檢索統；由於取得資料上得限制，目前我們的系統處理的文件資料包含日文與英文專利文件集。本論文第二節介紹相關的研究，第三節說明系統架構與檢索程序，第四節介紹實驗與結果，最後為結論及未來的研究方向。

二、相關研究

依我國專利法的定義是為鼓勵、保護、利用發明與創作，以促進產業發展，我國專利主分為：發明專利（提供新的做事方式或對某一問題提出新的技術解決方案的產品或方法）、新型專利（對舊事物的形狀、構造或裝置提出新的技術性創作）、新式樣專利（在事物的外觀上追求美感的新創作）三者。

專利文件是經申請並通過審查後所授予的一種權利，全世界現有 100 多個專利局公佈專利文件，每年平均公佈 100 多萬件，它既是法律文件，又是重要的技術情報。據統計有 90%發明成果的技術內容只有在專利文件中才能找到，而且專利文件還具有對發明創造說明詳盡和公佈最早的特點，透過檢索查詢專利文件，可以把握市場科技開發方向，並且可以參考他人研究成果，節省研發經費與縮短投入的時間，同時也為廣大企業在國內外貿易中瞭解有關產品技術狀況，對預防侵權提供幫助，並且研擬市場競爭策略。

表 1.專利檢索的項目

項目	檢索時機	檢索目的
專利現況檢索	在進入某一研究領域或開發新產品之前	大量檢索出相關專利、了解目前的專利概況、並在了解之後做出正確的判斷。
可專利性的檢索	有新構想擬申請專利時	對申請專利的內容和技術做一新穎性的確認，調查有無相關前案有助於專利申請的通過。
侵權的檢索	為技術、產品引進或輸出入時進行	在一項新技術或新產品進入市場之前應進行有無侵權的檢索，以避免構成對他人專利權的可能侵犯。
專利有效性檢索	在異議或舉發別人的專利是否有效而進行	檢索出相同的技術或文獻以證明別人的專利無新穎性，防止競爭者佔領某一技術領域。
技術預測檢索	為預測未來的發展	能正確的運用專利加速開發創造。
具體專利技術檢索	為解決技術問題上	專利資料中之有關技術背景與問題，常比期刊或書籍中記載要來的詳細。

由於專利檢索有其重要性，在每個階段檢索目的都不同，表 1 整理了一般企業及研發者使用專利檢索，其檢索的項目與時機，並且說明其檢索的目的。

專利文件的分類通常採用 IPC 分類, IPC 分類表目前由世界智慧財產權組織負責出版，每五年修訂一次（如表 2），而現在使用的是第八版。從 2000.1.1 至 2005.12.31 使用的第七版 IPC 碼，在專利文獻上表示為：Int. CL.7；第七版共有 8 個部（Section）、120 個主類（Class）、628 個次類（Subclass）、69,000 個主目（Group）及次目（Subgroup）。一個完整之分類碼必須由代表部、主類、次類、主目或次目之符號結合構成（如圖 1），我們使用一個範例來說明 H04L 12/44 所代表的意義如下：

部： H 電學
類： H04 電氣通信技術
次類： H04L 數位資訊之傳輸，例如電報通信
主目： H04L 12/00 數據交換網路
次目： H04L 12/44 星形或樹狀網路

表 2 國際專利分類表版本

版本	有效期間
第一版	1968/09/01～1974/06/30
第二版	1974/07/01～1979/12/31
第三版	1980/01/01～1984/12/31
第四版	1985/01/01～1989/12/31
第五版	1990/01/01～1994/12/31
第六版	1995/01/01～1999/12/31
第七版	2000/01/01～2005/12/31
第八版	2006/01/01 生效

（資料來源：國際專利分類檢索系統(第 8 版)使用指南）

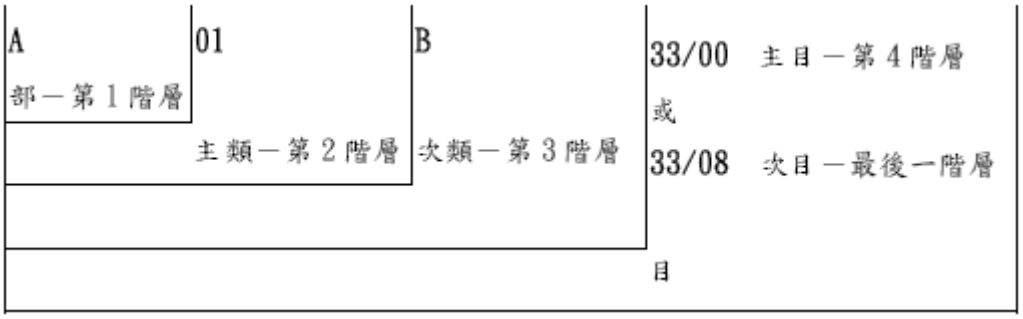

圖 1 完整之分類號組成

（資料來源：國際專利分類檢索系統(第 8 版)使用指南）

跨語言資訊檢索（Cross-Language Information Retrieval, CLIR）是使用某一種語言來查詢另外一種語言的文件，不過由於語言上的差異，通常都需要將查詢（Query）轉換成跟文件一樣的語言。目前大多數的使用者會在網際網路上使用搜尋引擎來查詢所需要的資料，當我們輸入中文的查詢字，執行檢索後我們可以發現結果可以包含其他語系與 Query 有關的相關資訊，這就是因為搜尋引擎會自動將您的輸入的 Query 翻譯成其他的語系並執行檢索的動作，由此可知跨語言的資訊檢索可以讓使用者方便使用自己熟悉的語言來檢索其他語系的文件。在跨語言資訊檢索相關的研究中，大部分採用的方法可歸納成文件翻譯（Document Translation）和查詢翻譯（Query Translation）兩種，兩種技術的目標都是要將查詢和文件的語言轉為一致。

使用文件翻譯的檢索方式必須將所有文件都翻譯和 Query 相同的語系，優點是文件與查詢都是使用相同的語言，使用者可以直接閱讀，缺點是翻譯所有的文件必須耗費大量的時間。

查詢翻譯需先將 Query 翻譯成和文件相同的語言，目前在跨語言資訊檢索中被廣泛的使用，此方法的好壞取決於 Query 是否被正確翻譯，而翻譯的方法有幾種被提出；有字典翻譯（Dictionary-based translation）方法[4]，語料庫翻譯（Corpus-based translation）方法[5]，混和（Hybrid）方法[6]，網路翻譯擷取（web-based translation extraction）方法[7]；由於網際網路上的資源眾多，很多的專家學者利用此優勢，使用網路查詢後再使用機率統計其結果，最後選擇最佳的翻譯當結果， Zhang 等人[8]指出使用網路擷取翻譯方式可以降低詞彙涵蓋度的問題。

綜觀以上方法，主要的目標都是將查詢和文件轉化成相同的語言，再進行資訊檢索，查詢的文字中某些關鍵字詞若無法被正確地翻譯，將會影響跨語資訊檢索的準確性。

在跨語言資訊檢索中，大部分的亞洲語言並不像英文一樣在每個單詞間都有分隔符號，因此斷詞這個步驟就顯得格外的重要，Shi & Nie[9]針對亞洲語系的斷詞使用不同方法，指出在處理日文斷詞方法採用 bi-gram 加上 uni-gram 可得到更好的效用。

NTCIR（NACSIS Test Collections for IR）計畫[10]是由日本國家科學資訊系統中心（National Center for Science Information Systems, NACSIS）所策劃主辦的，其目的是希望能建立一個大型日文標竿測試集，作為資訊檢索與自然語言處理研究的基礎資料。NTCIR 從 1999 年開始舉辦，至今已經邁入第七屆，從第三屆（2001-2002）開始舉辦了第一次的專利檢索評比，提供大型的文件集，包含二年的日文專利全文、五年的日本專

利摘要及五年日本專利的英文摘要，檢索題目有英、日、中、韓等四種語言，作為跨語言的專利檢索。由於專利檢索有不同的目的：技術調查（technology survey）、前案檢索（invalidity search）、專利地圖（patent map）等，假使查詢的主題相同但不同的檢索目的就會出現不同相關專利的結果，需要不同的檢索模式與技巧， NTCIR 的專利檢索從技術調查（technology survey）、前案檢索（invalidity search）、專利地圖（patent map）、專利分類（patent classification）到今年的專利採礦（patent Mining）、專利翻譯（patent translation）每年都會有不同的任務。

根據 NTCIR-6 有關專利檢索的研究[11, 12, 13]，要提昇專利檢索的精確度，除了原本的查詢欄位外，必須加入其他的相關欄位，甚至把整份專利文件都當檢索的條件，都可以增加檢索的查全率（recall）及查準率（precision）[11]。

三、系統描述

本系統架構如圖 2 所示，首先將專利文件集資料，經過模組程式過濾不需要的特殊字元、控制碼…等後，採用 bi-gram 的方式來處理日文文件，建置索引資料庫。查詢集利用三種線上翻譯系統翻譯為目的語言，檢索模型使用 TF-IDF（term frequency-inverse document frequency）的方法將檢索到的文件評分並加以排序，分類模組程式將檢索的結果進一步作 IPC code 自動分類。圖 3 為處理日文查英文之跨語言專利檢索的過程，先將日文查詢集經線上翻譯系統翻譯成英文，將翻譯過後的查詢集進行詞彙與斷詞的處理，最後進行檢索作業；圖 4 是處理英文查日文之跨語言專利檢索的過程。

我們使用 Lucene[14]作為專利檢索的搜尋引擎，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）。Lucene 針對軟體開發人員提供一個簡單易用的工具包，可以建立完整的全文檢索引擎，Lucene 除了有 JAVA 的版本之外，也陸陸續續的被開發成其他不同的版本，如 C#、C++、Delphi、Perl、Python、Ruby 和 PHP 等，本實驗所建立的專利檢索系統採用 C#的版本。

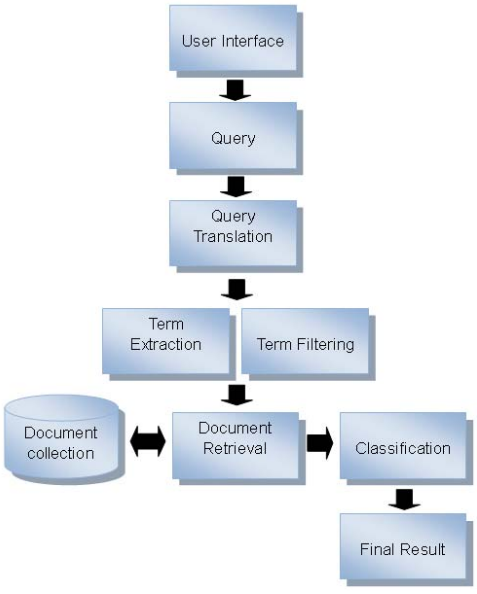

圖 2 系統架構圖

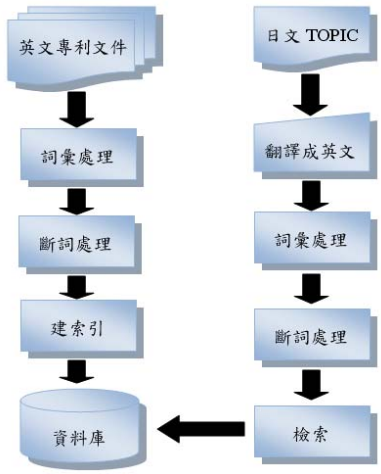

圖 3　日文查英文之跨語言專利檢索處理

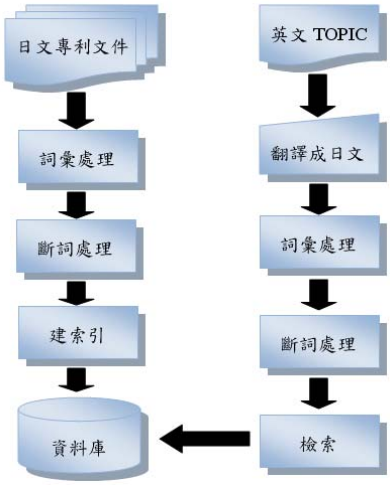

圖 4　英文查日文之跨語言專利檢索處理

3.1 詞彙處理

處理亞洲語系文件的首要工作是詞彙處理，因為大部分的亞洲語言並不像英文一樣在每個單詞間都有分隔符號，一般採用 N-gram 的技術處理不同的亞洲語言[15,16,17]，其中使用 bi-gram 的方法又比使用 uni-gram 的方式好，因此在本實驗中對於日文的斷詞方式採用 bi-gram 的方法，表 3 為日文句子採用 uni-gram 及 bi-gram 斷詞後的結果。

在資訊檢索時，標點符號、特殊字元與停用字（Stop Word）都是無意義的，因此在建立索引和檢索之前需把查詢集與文件集內的這些相關字元給去除。由於使用 Lucene 建立索引時，系統會自動將標點符號與停用字給過濾掉，因此我們只要注意特殊字元的處理即可。

表 3　　日文採用 uni-gram 與 bi-gram 斷詞後結果

原始句	チリ産い貝の接着タンパク質の合成
uni-gram	チ|リ|産|い|貝|の|接|着|タ|ン|パ|ク|質|の|合|成
bi-gram	チリ|リ産|産い|い貝|貝の|の接|接着|着タ|タン|ンパ|パク|ク質|質の|の合|合成

3.2 查詢翻譯處理

本系統的查詢翻譯是將原始語言的查詢集利用不同的線上翻譯網站將其翻譯成目標語言，再進行單語專利檢索；例如：英文的查詢集經過線上翻譯系統翻譯成日文後，再進行日文專利檢索。由於不同的線上翻譯系統所翻譯的結果均不相同，因此我們採用了 Google Translation[18]、Yahoo Babel Fish[19]及 Excite[20]三種不同的線上翻譯系統來彌補翻譯不足的問題。

在實驗中，我們使用的線上翻譯系統並非完全針對資訊檢索的查詢集(Topics)而設計的，所以若我們直接把查詢集的文件傳送到線上翻譯系統，發現傳回來的結果與文件格式會有所錯誤，表 4 列出直接使用線上翻譯系統時可能產生之的問題，這些錯誤必須人工檢視翻譯後的結果，並修正其錯誤，方可進行後續的檢索處理。

我們採取另一種的方式可以有效的讓這些錯誤發生的問題降低。首先先將查詢集的文件轉成 XML 的文件格式，因為 XML 的文件具有欄位的特性，我們將每個欄位分別傳送到翻譯系統，再按欄位依序取回翻譯的結果，雖然這樣可以正確的取回結果，不過發現部份結果會帶有 HTML tag，由於我們不需要這些 HTML tag，將這些 HTML tag 去除後，得到我們需要的翻譯結果。

本系統使用三種線上翻譯系統進行查詢翻譯，分別為 Google Translation、Yahoo 線上翻譯、與 Excite 線上翻譯，分別介紹於下:

Google Translation 是一個免費的線上翻譯網站，提供多種語系的翻譯，其中也包含使用英文對日文的翻譯服務，Google 線上翻譯系統有別於其他翻譯系統的作法，是採用統計式的作法，由電腦進行網頁比對找出翻譯機率，當作文件翻譯之用。

表 4　　直接使用線上翻譯可能產生之格式錯誤

翻譯後格式錯誤結果之範例	錯誤說明
<TOPIC> <TOPIC-ID> 100 </ TOPIC-ID> PB <TITLE> sound transmission in mobile communications processing system </ TITLE>	翻譯後內容與 Tag 的位置不正確
<TOPIC-ID> 101 </ TOPIC-ID> <TITLE> Artificial boundary-derived lipids and proteins and biological hybrid by RIPOSOMUWAKUCHIN </ TITLE>	Tag 內有多餘的空白產生
<topic-id> 100 < /トピック- ID を> <title> dtmf （デュアルトーンマルチ周波数）伝送方式は、移動体通信システム< /タイトル>	Tag 的文字也被翻譯了
A. B hepatitis, B. genetic engineering techniques, C. vaccine, vaccination	傳回的結果會增加一些不必要的 Html Tag

表 5 維基百科與 Google、Yahoo 翻譯比較表

	人名	專有名詞
	日文/英文對照	日文/英文對照
維基百科	ヘルベルト・フォン・カラヤン / Herbert von Karajan	航空交通管制 / Air traffic control
Google 翻譯	ヘルベルトフォンカラヤン / Herbert von Karajan	航空管制 / Air traffic control
Yahoo 翻譯	ハーバートフォン Karajan / [heruberuto] phone Karajan	航空管制 / Flight control

在實際使用上， Google Translation 對於專有名詞、人名、術語等的翻譯上有其獨特的地方，我們在維基百科（Wikipedia）上隨機選擇一個人名與專有名詞來作比較，在維基百科上的日文與英文別為ヘルベルト・フォン・カラヤン、Herbert von Karajan，我們使用Google翻譯的結果為ヘルベルトフォンカラヤン、Herbert von Karajan 而 Yahoo 翻譯的結果為ハーバートフォン Karajan、[heruberuto] phone Karajan，其結果如表5所示。

Yahoo 翻譯網站提供多國語系的翻譯，它是採用 Alta Vista 和 Systran 合作提供的翻譯服務「Babel Fish」。Yahoo 日本網站使用的翻譯服務與 Babel Fish 是不同的技術，而且僅提供英文、中文、日文及韓文的翻譯服務。由於這兩個網站都有提供英日相互翻譯的功能，我們選擇使用 Yahoo 翻譯作為我們系統的其中一種翻譯系統。

Excite 翻譯網站提供的翻譯語系較上述兩種系統為少，它僅提供日文對英文、英文對日文、日文對中文、中文對日文、日文對韓文、韓文對日文等六種翻譯方式，但此網站是多數人在翻譯日文時推薦使用的線上翻譯網站，因此也納入作為其中一種的翻譯系統。

3.3 分類處理

本系統的分類處理採用下列步驟決定 IPC 分類碼：

(1) 將專利文件集建立索引

(2) 使用 Topics (Query)進行檢索

(3) 對步驟二取出之前 3000 份專利文件分別抽取相對應之 IPC code

(4) 步驟三得到之 IPC code 分別使用下列公式計算 Score

Score(IPC)= Σ(專利文件對於 Query 的相似度分數)

例如：從檢索結果中取出的 Top 3000專利文件，當中含有“A61B_5_02”這個 IPC code 的專利文件編號分別為 PATENT-US-GRT-2000-06152884 、PATENT-US-GRT-1993-05181521與 PATENT-US-GRT-1998-05772600，對於 Query 的相似度分數分別為21.264、17.724,、20.125，則：

Score (A61B_5_02) = 21.264 + 17.724 + 20.125 = 59.113

(5) 根據得分高低排序輸出 IPC codes

四、實驗

實驗資料來源是採用 NTCIR 提供的文件集 (Document Sets)與查詢集(Topics)，資料包含 1993 年到 2002 年的 Unexamined Japanese patent applications、USPTO patent data、Patent Abstracts of Japan 與 NTCIR-1 與 NTCIR-2 CLIR task Test Collection 等相關專利文件，表 6 為這些文件集的數量與所佔儲存容量。

我們使用 Lucene 建立文件集的索引，由於日文的文件集都是採用日語語系的 EUC 文件編碼方式，造成讀取文件與建立索引時有亂碼產生的問題，由於 Lucene 支援 UTF-8 的編碼方式，所以預先將所有的日文文件集使用工具將 EUC 編碼轉換成 UTF-8 編碼，去除標點符號、特殊字元與停用字，再使用 bi-gram 的斷詞處理，接下來使用 Lucene 建立索引檔，建立完索引檔後，便可以使用 Lucene 提供的檢索功能檢索資料。

表 7 所列的是使用日文專利文件集建立索引所花的時間及所佔儲存容量大小，表 8 是使用英文專利文件集建立索引所花的時間及所佔儲存容量大小，建立索引時所使用機器為：Pentium 4 2.66GHz，記憶體大小為 1GB，作業系統為 Microsoft Windows XP。

表 6 NTCIR-7 的文件集數量

類別	語言	文件數	容量（MB）
NTCIR-1	日文	332,918	312
	英文	187,080	218
NTCIR-2	日文	403,240	600
	英文	134,978	200
Unexamined Japanese patent applications	日文	3,496,252	96,768
Patent Abstracts of Japan	英文	2,543,488	4,102
USPTO patent data	英文	1,315,470	53,351

表 7 NTCIR-7 的日文文件集索引

類別	文件數	term 數量	Index 容量	Index 時間
NTCIR-1	332,918	1,596,747	312 MB	6.98 小時
NTCIR-2	403,240	2,021,914	710 MB	9.56 小時
Unexamined Japanese patent applications	3,496,253	7,596,840	46,445 MB	308.08 小時

表 8　　NTCIR-7 的英文文件集索引

類別	文件數	term 數量	Index 容量	Index 時間
NTCIR-1	187,080	1,033,575	211 MB	4.11 小時
NTCIR-2	134,978	785,607	227 MB	2.78 小時
Patent Abstracts of Japan	2,543,488	3,191,893	676 MB	22.07 小時
USPTO patent data	1,315,470	2,653,186	1942 MB	27.69 小時

在檢索之前，首先把查詢集的文件均轉成 UTF-8，再將檔案格式改為 XML 的格式，接下來將查詢集裡的特殊字給去除，最後得到我們所需的查詢集。系統畫面如圖 5 所示，主要步驟包括：

(1) 首先載入查詢集

(2) 根據要翻譯的方式，選擇日翻英、英翻日或者不翻譯

(3) 接下選擇使用的線上翻譯工具

(4) 系統根據使用者選擇的檢索內容、檢索的專利文件集（日文專利集或英文專利集）、檢索的欄位（TITLE、ABSTRACT、Free Text）、檢索結果的顯示筆數，顯示最後的檢索結果

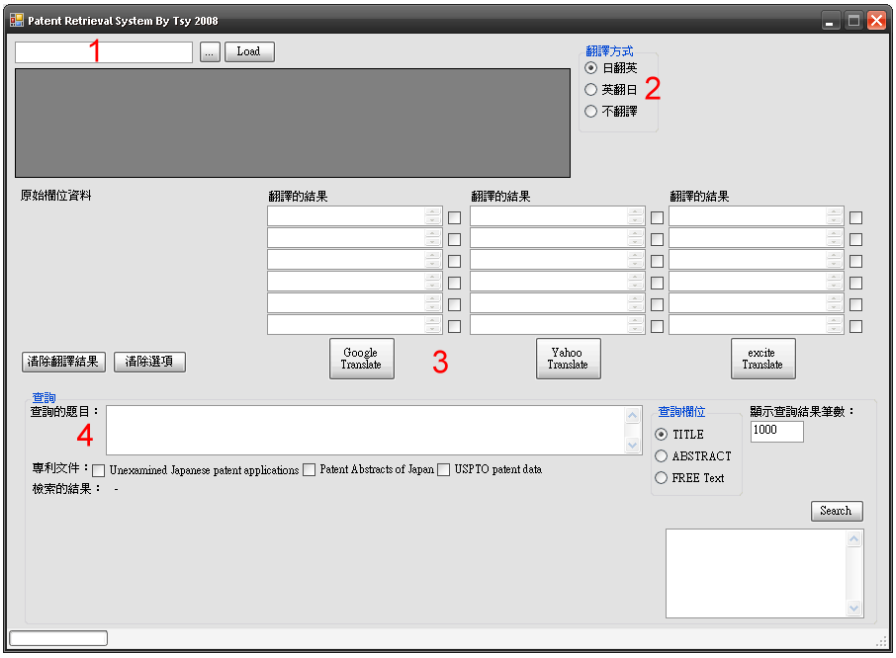

圖 5 系統主畫面

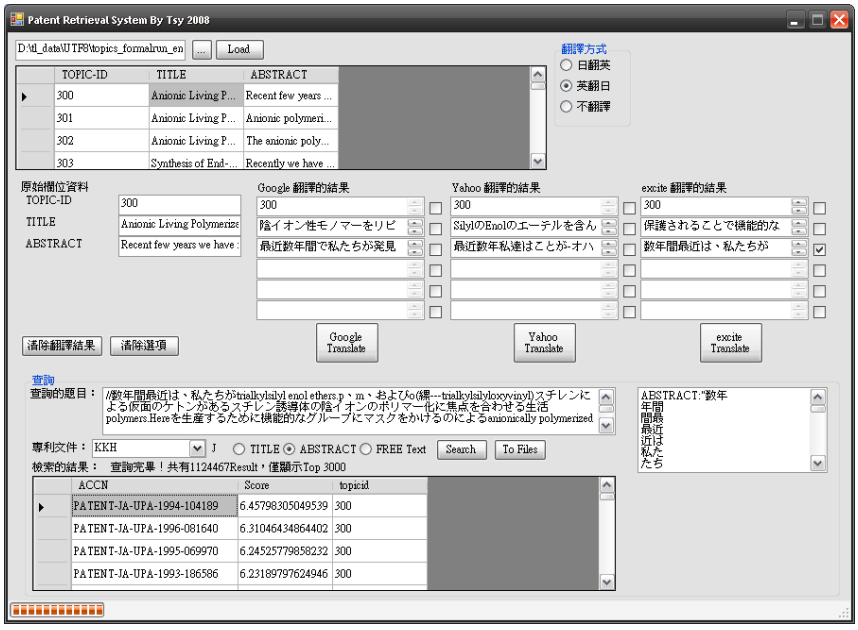

圖 6　英文查詢集檢索日文專利文件集之範例

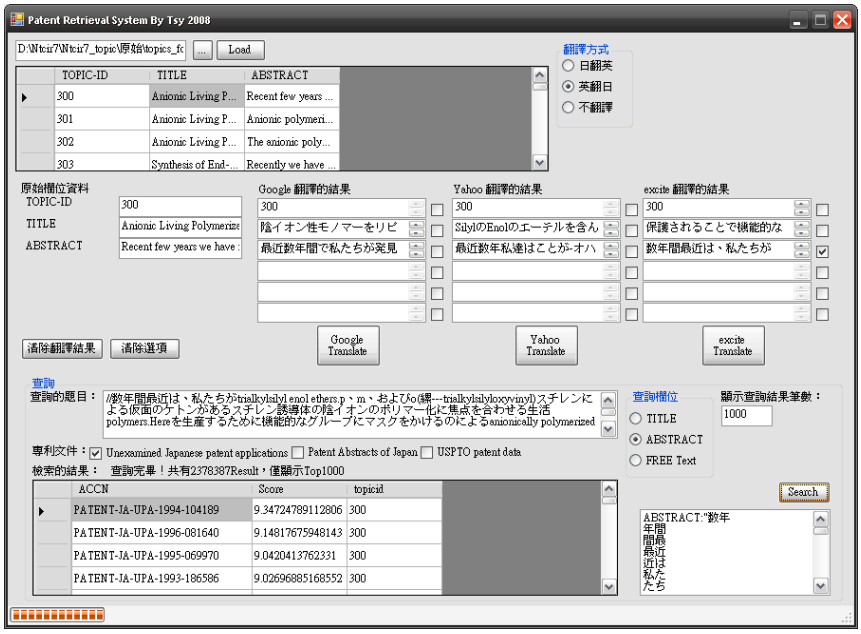

圖 7 日文查詢集檢索英文專利文件集之範例

由於要提昇專利檢索的精確度，除了原本的查詢欄位外，必須加入其他的相關欄位，甚至把整份專利文件都當檢索的條件，都可以增加檢索的查全率（recall）及查準率（precision）[11]，所以我們在設計系統時，也將全欄位納入我們的檢索條件內，圖 6 為使用英文查詢集來檢索日文專利文件，首先載入英文查詢集，選擇英翻日，選擇查詢題目編號為 300 的題目，接者使用 Excite 線上翻譯系統，使用 ABSTRACT 為檢索條件，專利文件集則選擇日文專利文件集，檢索欄位為 ABSTRACT，結果顯示筆數為 1000 筆，點選 search 則會顯示檢索結果，其中 ACCN 為專利文件編號、Scroe 為適合度分數；圖 7 為使用日文查詢集來檢索英文專利文件，使用日文查詢集，日翻英方式、查詢題目編號 1393、Google 線上翻譯、TITLE 為檢索條件、Free Text 為檢索欄位、英文專利文件集、顯示筆數為 1000 筆。

表 9 日文查詢集編號 1393 的檢索結果

原始 Query 的 Title	鉛直遮水壁の封じ込め効果に関する透水土槽実験（その3）－透水土槽実験と事後解析－
翻譯後的結果	Vertical wall impervious soil permeability effects of the containment tank experiment (3) - and subsequent laboratory analysis permeable earth tank --
檢索結果 Top 1 的內容	<DOC><DOCNO>PATENT-PAJ-G-H08-222168</DOCNO><TEXT><PATDOC><JPPAT><SDOBI LA="E"><B110>10057788</B110><B121>PATENT ABSTRACTS OF JAPAN</B121><B130>A</B130><B140>19980303</B140><B190>JP</B190><B210>08222168</B210><B220>19960823</B220><B511> B01F 3/00 </B511><B512> G01N 33/24 </B512><B541>EN</B541><B542>SOIL TANK FOR EXPERIMENT</B542><B711>KAJIMA CORP</B711><B721>IKEZOE KATSUJI</B721><B721>UEKI MUTSUO</B721><B721>NOMURA KEIGO</B721><B721>TAMAI TATSURO</B721><B721>SHIRAI SHUNSUKE</B721></SDOBI><SDOAB LA="E"><SEC><P>PROBLEM TO BE SOLVED: To constitute a soil tank in such a manner that the tank is released from the nonuniformity of lower part compression by an arch action, etc., that sand and soil are uniformly compressed even if a tunnel, etc., are built in experiment soil and that the state in the actual sand and soil is embodied. </P><P>SOLUTION: This soil tank is formed with an upper cell 2, a middle cell 3 and a lower cell 4 by films having flexibility and impervious property, has a means for introducing pressure water from outside into the respective cells and has a means for supporting the sand and soil when the sand and soil are sealed into the middle cell 3. The structure to execute soil tank experiment in the sand and soil sealed into the middle cell 3 by sealing the sand and soil into the middle cell 3 and introducing the pressure water into the respective cells, i.e., the upper cell 2, the middle cell 3 and the lower cell 4 is obtd.</P><P>COPYRIGHT: (C)1998,JPO</P></SEC></SDOAB><SDODR LA="E"><EMI ID="00000001" HE="089" WI="066" TI="AD" IMF="TIFF"></EMI></SDODR></JPPAT></PATDOC></TEXT></DOC>

表 9 為使用日文 Topic 編號為 1393 的 TITLE 當作 Query 的題目，經過 Google Translation 將日文翻譯為英文，查詢英文專利文件的結果我們取出 Top 1 的內容並顯示。表 10 為使用日文查詢集編號 305 時，三種線上翻譯系統的結果比較，以使用 Google Translation 的準確度較佳。

由於專利文件的資料相當的龐大，我們花了很多的時間在轉換文件編碼，處理原始資料上一些沒用的資訊，如特殊字、標點符號…等，當這些處理完畢後，才能開始建立索引。而在查詢翻譯處理中，使用的線上翻譯系統並非完全針對資訊檢索的查詢集(Topics)而設計的，在翻譯結果上會產生格式上的錯誤，我們採取的方式可以有效解決這些格式問題。

根據專利分類體系之中的 International Patent Classification (IPC)進行自動分類，因此當完成檢索之後，需將檢索結果加入 IPC code 並加以評分，得到最後之結果。分類子系統畫面如圖8所示，主要的步驟包括：

(1) 載入檢索結果
(2) 選擇要加入 IPC code 的種類
(3) 進行比對並加入 IPC code

例如載入編號300的檢索結果，IPC code 的分類方式為 PAJ＆USPTO，點選執行後產生加入 IPC code 後的結果（如圖9）；其中 topicid 為檢索題目的編號、IPC 為 IPC code、Score 為分數、IPC-rank 為順序。

表 10　　日文查英文項目，日文查詢集編號 305 的翻譯結果

	末端に官能基を有するポリマーの合成〔VIII〕官能基を保護したω-ハロ化合物とアニオンリビングポリスチレン、ポリイソプレンの反応
Google	Functionalized end of synthetic polymers VIII] [ω-functional group to protect the compound and ANIONRIBINGUPORISUCHIREN Hello, polyisoprene reaction
Yahoo	ω-[haro] chemical compound and the anionic living polystyrene which protect the synthetic (viii) functional group of the polymer which possesses the functional group in end, the reaction of the polyisoprene
Excite	ω-hello the reaction of the compound, the anion living polystyrene, and polyisoprene that protects synthesis VIII functional group of Polymer that has the functional group in the end.

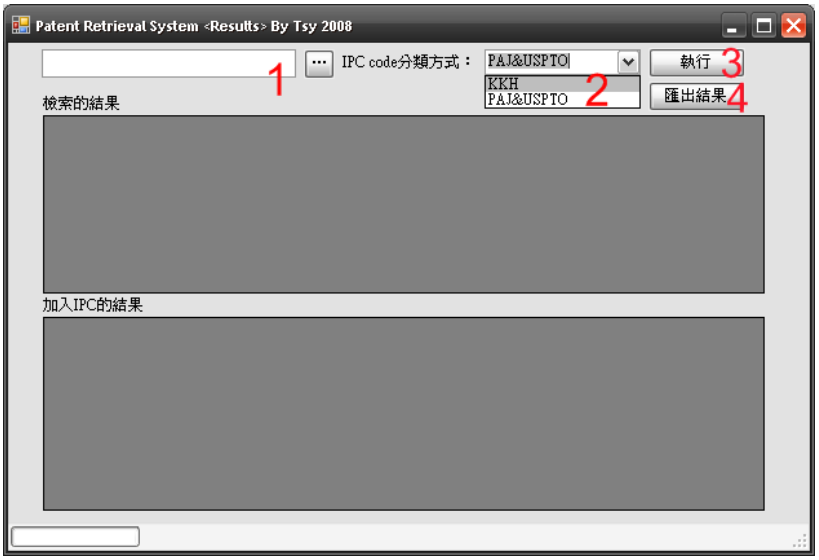

圖 8　　分類子系統的主畫面

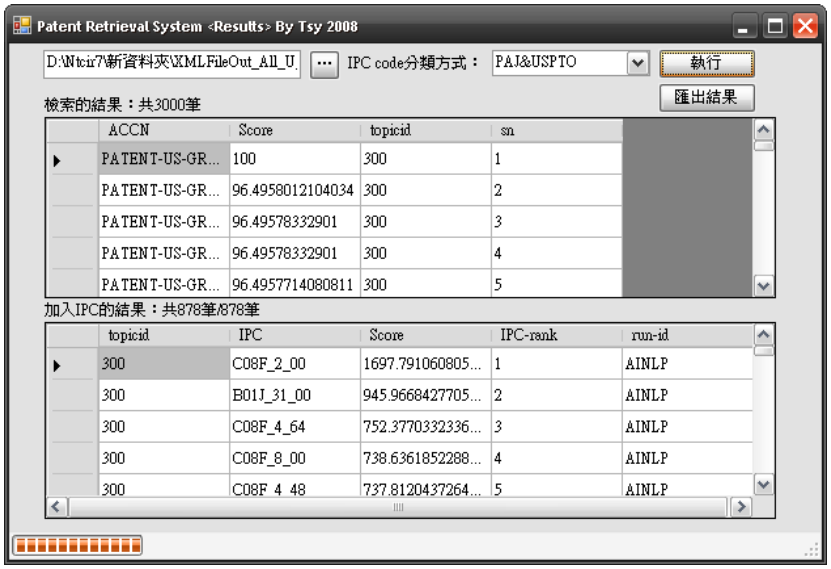

圖 9　　檢索結果加入 IPC code 之範例

五、結論

本論文提出一個 CPRS 跨語言專利檢索系統（Cross-language Patent Retrieval System）的架構，採用適合處理不同語言的 bi-gram 索引方法，透過 Lucene 檢索引擎處理多語言的專利文件集（Document Sets）與查詢集（Topics），並結合網路翻譯系統，利用查詢翻譯的方法，將原始的查詢加以翻譯，再進行專利檢索。並且建置一個能處理多語專利文件的跨語言專利檢索系統，進一步利用檢索結果，根據專利分類體系之中的 International Patent Classification (IPC)分類，得到相關的之 IPC 分類。

目前我們的系統可以處理英文與日文的單語檢索、以及日文檢索英文專利文件與英文檢索日文專利文件兩種跨語言檢索。使用者可以選擇三種不同的翻譯方式來翻譯查詢集，並且可以選擇檢索的欄位及專利文件集的種類，進行跨語言專利檢索。

致謝

本論文的完成感謝日本 NTCIR（NACSIS Test Collections for IR）提供專利文件，以及華梵大學先進製造研究中心提供部份的研究經費支援。

參考文獻

[1] WIPO, http://www.wipo.int/portal/index.html.

[2] STN International, http://www.stn-international.de.

[3] 國際專利分類檢索系統(第 8 版)使用指南， URL: http://newweb.tipo.gov.tw/ch/MultiMedia_FileDownload.ashx?guid=5dc74ecb-4be5-42c7-ada2-dcd37ad908fb

[4] Ballesteros, L. and Croft, W.B. “Dictionary–based Methods for Cross-Lingual Information Retrieval.” Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 791-801, 1996.

[5] Yang, C.C. and LI, K.W. “Mining English/Chinese Parallel documents from the world wide web”, Proceedings of the 11th international World Wide Web Conference, Honolulu, Hawaii, May, 188-192, 2002.

[6] Bian, G..W. and Chen H.H. “The Study of Query Translation and Document Translation in a Cross-Language Information Retrieval System” Ph.D. Dissertation, National Taiwan University, Taipei, Taiwan, 1999.

[7] Cheng, C.C.; Shue, R.J.; Lee, H.L.; Hsieh, S.Y.; Yeh, G.C. and Bian, G.W. “AINLP at NTCIR-6: Evaluations for Multilingual and Cross-Lingual Information Retrieval”, Proceedings of NTCIR-6 Workshop, Japan, 2007.

[8] Zhang, Y.; Vines, P. and Zobel, J. “Chinese OOV translation and post-translation query expansion in chinese--english cross-lingual information retrieval”, ACM Transactions on Asian Language Information Processing, Vol.4, No.2, June, 55-77, 2005.

[9] Shi, L. and Nie, J.Y. “Using Unigram and Bigram Language Models for Monolingual and Cross-Language IR", Proceedings of NTCIR-6 Workshop, 2007.

[10] Makoto, I.; Atsushi, F. and Noriko, K. “Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task”, Proceedings of NTCIR-6 Workshop, 2007.

[11] Fujii, A. “Integrating Content and Citation Information for the NTCIR-6 Patent Retrieval Task”, Proceedings of NTCIR-6 Workshop, 2007.

[12] Mase, H. and Iwayama, M. “NTCIR-6 Patent Retrieval Experiments at Hitachi”, Proceedings of NTCIR-6 Workshop, 2007.

[13] Tseng, Y.H. ; Tsai, C.Y. and Juang, D.W. “Invalidity Search for USPTO Patent Documents Using Different Patent Surrogates”, Proceedings of NTCIR-6 Workshop, 2007.

[14] Lucene, http://lucene.apache.org/java/docs/index.html

[15] 鄭貞信，「英中日韓文的跨語言檢索之比較」，華梵大學資訊管理學系碩士論文，民國九十六年。

[16] Kwok, K-L and Dinstl N. “NTCIR-6 Monolingual Chinese and English-Chinese Cross Language Retrieval Experiments using PIRCS”, Proceedings of NTCIR-6 Workshop, Japan, 2007.

[17] Su, C.Y.; Lin, T.C. and Wu, S.H. “Using Wikipedia to Translate OOV Term on MLIR”, Japan, Proceedings of NTCIR-6 Workshop, 2007.

[18] Google Translation, http://www.google.com.tw/translate_t

[19] Yahoo Babel Fish, http://babelfish.yahoo.com/

[20] Excite, http://www.excite.co.jp/world/english/