

The Establishment of Retrieving System for ancient Books of Chinese Medicine and Solution for the Infrequent Chinese Characters

Ching-Wen Chen, Shi-Xian You, Yi-Cheng Xiao
Computer Sciences and Information Engineering Department, Diwan College
of Management

Abstract

This study established a mining system for ancient Books of Chinese Medicine and Solution for the Infrequent Chinese Characters with JAX、ASP.NET、Java、XML ontology technologies. The functions of this mining system included mining with author name, book name, keywords and book classification.

The problems of the ancient Chinese characters are full of rare and infrequent characters. Most of these infrequent characters are not supported by current coding system inside the computer. We modified the fonts of ancient Chinese Medicine from the Committee on Chinese Medicine and Pharmacy, Department of Health, Executive Yuan. These fonts, which did not matched with modern Unicode system, were built during 1995. We revised all codes of the 4112 infrequent characters with Unicode' s and established the pictures for the left characters which could not be found the related code.

Keywords : Chinese Medicine, AJAX, Unicode, Infrequent Chinese Characters

建立中醫典籍檢索平台與中醫古難字 解決方案

陳擎文 游士賢 蕭耀晟

致遠管理學院資訊工程學系

摘 要

這個研究乃是採用AJAX、ASP.NET、Java、XML網路本體論來建立中醫典籍探勘平台並提出一個中醫古難字解決方案。系統以AJAX技術建立中醫典籍探勘平台，系統包括『以年代查詢、以書名查詢、以作者查詢、以性質分類查詢、以關鍵字查詢』。

中醫古字的問題，主要是中醫古籍充滿太多的缺難字與罕用字，這些罕用字在目前的編碼系統中（例如：big5和Unicode）大多仍是沒有支援編碼的。本研究參考行政院衛生署中醫藥委員會的94年版『中醫藥造字檔』，該造字檔乃是民國八十四年底DOS、Windows 95系統交接時期所設計的產品，並需配合漢書軟體使用，已較為古老。故本研究以該造字檔為依據，並參考『行政院主計處電子處理資料中心建置「CNS11643中文標準交換碼全字庫」』，將4112個中醫罕用字逐一校對目前Unicode所支援的編碼，並將剩餘的編碼以字型圖片顯示。

關鍵詞：中醫古籍、AJAX、編碼、缺難字、罕用字

壹、前言

傳統中醫學習者在學習中醫文獻時，所遇到的最大難題就是，有太多的艱澀詞彙難以理解，並且有很多罕見的字型，造成學習中醫過程的困難。故中醫古字的問題，主要是中醫古籍充滿太多的缺難字與罕用字，這些罕用字在目前的編碼系統中（例如：big5和Unicode）大多仍是沒有支援編碼的。雖然大部分的罕用字可能是錯字，但是本研究希望推動一項對幾千年來大部分重要中醫藥典籍的總校對工作，希望能提出一個可行又方便的中醫罕用字解決方案。

在中文電腦中，中文字的顯示，仰賴於中文內碼的支援，如：BIG-5(大五碼)、Unicode(萬國碼)，我們所使用的中文字，其組合方式，是由BIG-5或者Unicode提供內碼的轉換對應，才能正常顯示出我們所輸入所顯示的中文字。

由於中國文字在歷史上，或者應用上都有相當長的歷史記載，鑒於如此，在中文的演進從最早的象形文字，一直到我們現在所使用的正體文字以及簡體文字，在歷史上已有四五千年的記載，也因為使用的時間長，加上以前的社會沒有電腦文書，也沒有一定的文字規範，有相當多的古籍作者可能在抄寫文字時，不小心寫錯字，或者自行造字，都是電腦缺字的主要原因。

電腦發展至今，已經有相當多的文字編碼，如倚天碼、零壹碼、IBM5550碼、BIG-5碼、倉頡碼、公會碼，等等的

編碼（謝清俊，1996），而在系統專屬的編碼更是五花八門，幾乎發展一套電腦硬體，就會對應一種編碼，如：IBM碼、王安碼、大同碼…等，或者因為機關的關係，自行發展了內碼系統，如警政署、戶政機關等，在內碼繁多的情況下，要統一是非常困難的，為了解決這些問題，我們需要在一台伺服器上，同時安裝這些字型內碼，在林林總總的字碼中，找出相互之共通性而排除，並以Unicode為統一編碼，讓使用者不必再額外安裝其他字型檔。

在本研究中，我們必須要收集文字各編碼的相關區段，來提供我們字形造字，並存於伺服器中。並採用系統中央造字，以及採用程式設計的方式，來進行文字圖片製圖的動作。至目前為止，我們共利用程式創造了2673個中文字，且持續累加中。

本研究乃是參考行政院衛生署中醫藥委員會的中醫藥造字檔（行政院衛生署中醫藥委員會，2008），該造字檔乃是民國八十四年的DOS、Windows 95系統交接時期所設計的產品，並需配合漢書軟體使用，已較為古老，也無法對應到現代較為普遍使用的Unicode萬國碼。故本研究以該造字檔為依據，並參考『行政院主計處電子處理資料中心建置CNS11643中文標準交換碼全字庫』（行政院主計處電子處理資料中心，2008），然後再將之整合到Unicode編碼。

目前可用來組字的部件有字集包含

《中文電腦基本用字》(8532字)、五大碼(Big5, 13053字)，辭典方面有《中文大辭典》(49905字)、《漢語大字典》(54640字)、《說文解字》(9335字)、《康熙字典》(46000字)、《辭海字典》(16534字)。字形結構的登錄，是參考《康熙字典》的部首(214個)，依據橫連、直連、包含三個法則，將字形拆分成部件，再將部件拆分成字根(謝清俊，1997)。目前可用來組字的部件除了Big5用字13051個外，本研究還提供了4112個字，其中包含1931個Big5找不到的古文字。

構字式，乃是字體本身組成是利用好幾個部件組成而成。例如字形「許」拆成「言 午」，「張」拆成「弓 長」，「李」拆成「木 子」。部件可為兩個以上，如「糊」拆成「米 古 月」，「哲」打成「扌 斤 口」，但是絕大多數的情形都只使用兩個部件，即是構字式。

根據《漢字部件規範》，「漢字部件」是指由筆畫組成的具有組配漢字功能的構字單位，簡稱「部件」。例如「賴」字的左邊為「束」，右邊的「刀 貝」並不在Big5中，字形組合的方式是『賴=束 (刀貝)』，其中同時用到兩個拆分符號，所以無法採用構字式，當缺乏適當的部件，如「賴」字，則可完全摒除拆分符號，用部件序來表示(莊德明、謝清俊、林晰等，1998)。

有些字連一個部件都拆不出來，在中文大辭典共有49905個字，約有上千個

字像圖形，幾乎無法拆分，這時候就得利用缺字序號來識別它。缺字序號的型式無法利用上述規則表達的缺字，本研究部分罕用字乃是採用內建編碼再配合文字圖示的方法。

貳、研究方法與步驟

本研究在處理中醫古難字的研究方法與解決步驟，如下所述：

1. 步驟1：我們應用了AJAX即時更新的技術，在使用者端載入一篇中醫古文時，系統將會在文中找尋 Unicode 或者 Big5 碼內沒有的文字，並進入 Database 中找尋相對應之內碼，利用 Java 即時產生 XML 並傳回圖片數據給予使用者端的 AJAX 介面，或稱為 JavaScript 來處理並顯示出這個中醫古難字，免除以往使用者仍需要下載「文字編碼更新檔」才能觀看，且利用AJAX的特性，我們可以不斷的重覆載入新文章並即時導入資料庫中來進行查詢字庫的動作。Ajax是利用使用者與伺服器端的資料傳送技術，給流覽器可以為用戶提供更為自然的流覽。在Ajax之前，Web技術是讓用戶的動作與伺服器的“思考時間”同步(提交→等待→重新顯示範例)。Ajax提供與伺服器非同步請求的技術(用戶請求→回應的迴圈中解脫出來)。借助於Ajax，可以在用戶單擊按鈕時，使用JavaScript和DHTML立即更新UI，並向伺服器發出非同步請求，以執行更新

建立中醫典籍檢索平台與中醫古難字解決方案

或查詢資料庫。當請求返回時，就可以使用JavaScript和CSS來相應地更新UI，而不是刷新整個頁面。利用Ajax技術的方式，第一次查詢時IE會主動查詢，而第二次以後，對於相同資料則會置放於IE快取中，所以速度就加快了。Ajax的核心是JavaScript物件 XMLHttpRequest，XMLHttpRequest使您可以使用JavaScript向伺服器提出請求並處理回應，而最重要的是，讓用戶不知道瀏覽器正在與伺服器通信

2. 步驟2：以下為大五碼的內碼位置配置表

0x8140 0x8DFE(使用者造字區)

0x8E40 0xA0FE(使用者造字區)

0xA140 0xF9FE(大五碼主要字集區)

保留：0xA3C0 ~ 0xA3FE

保留：0xC6A1 ~ 0xC8FE

0xFA40 0xFEFE(使用者造字區)

0x8140 ~ 0x8DFE 與 0x8E40 ~ 0xA0FE 並不需要去進行處理，因為 Java 或 Win32 API 並不會使用到這一個區塊的內碼段，若使用者沒有下載或安裝「中醫藥委員會的字型檔」，有些文字將會顯示為『 』空白字串，而使用者若使用16-Hex Editor 觀看文字時將會發現這些空白並非內碼為 0xA140 的空白字串，而是因為 Unicode 與 Big5 皆沒有這些內碼字符的原故。

大陸之『中華醫典』軟體是目前國際間最大的中醫典籍商業軟體，其處理方法說明如下：

他們有自己的一套字型檔，而使用

到的內碼區段也是存在於使用者造字區 0x8140 ~ 0x8DFE 與 0x8E40 ~ 0xA0FE，所以與「中醫藥委員會」的內碼區段重覆了，若想要進行「中醫藥委員會」的字型檔做相容，勢必採以人工作業方式進行「中華醫典」與「中醫藥委員會」的對應資料表，並利用查表方式進行查詢，此時將可相容於「中醫藥委員會」的字型檔了。

但不管如何，在沒有一個標準出現以前，這些都只是治標不治本，只會造成更多的字碼混亂，任何的人，或者字型檔的開發者，勢必還是需要進行移植的動作，麻煩且費時。

根據上述的這幾種方法，我們決定採用以下方式來處理中醫古難字：

(A)我們必須將4000多字的古難字轉換成Unicode並建立其圖檔。

(B)將 0x8140 ~ 0x8DFE 與 0x8E40 ~ 0xA0FE 的字型改用 Unicode

(C)若不存在於Unicode，則用CNS圖片或者自行製作圖片

(D)若上述方法都處理了，還是發生缺字，則採用以下方式來處理：

由於缺字的字碼，多屬於大小不定、格式複雜的文字，為了統一管理，我們將採用圖像檔來做轉換，並利用文字原始內碼來存檔，日後讓程式能讀取內碼後進入 Database 內找出圖片並顯示，本系統利用Unicode編碼共有4112字，格式如下：

Unicode 4 位數編碼不需要圖片 (2181字)

Unicode 5 位數編碼全字庫文字
(227字)

找不到編碼自製文字 (1704字)

3. 步驟3：系統採用Java Zhcode的繁簡對應來做轉碼動作，但大五碼只有 0xA140 ~ 0xF9FE 可以完整對應，使用者自行造字區段則無法對應，為了讓系統能夠對應，則需要搭配 Win32 API 以及外掛DLL並配合 JNI 才能有效處理這些區段，另外為了讓使用者能夠在最短的時間內查詢到相關文字，我們將一些常用的字庫，大約2000字，建立主索引(Index)，並將關鍵字與其對

應之 Contents 進行關聯，若使用者輸入的關鍵字並不存在於這些索引內，系統將會立即採用動態模式進行查詢，並重新建立其索引。

參、研究成果與討論

一、中醫古難字的處理結果

本研究主要目的是希望在雜亂且綜多的文字編碼中，建立一個可以共通的文字字庫，並搭配 Unicode 與自建圖像資料庫來開發，只要使用者電腦有支援 Unicode 則將不會產生亂碼問題。



圖1. 沒有安裝字型檔之前的文件

建立中醫典籍檢索平台與中醫古難字解決方案



圖2. 安裝字型檔後顯示出的文字

為了能夠幫助使用者在沒有安裝字 字，方便使用者進行導讀。
型檔的電腦，也能快正確的導入相關文



圖3. 將記事本文字複製到 Dreamweaver

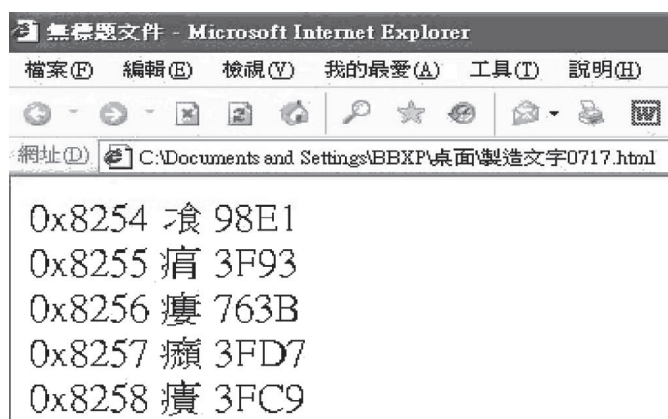


圖4. 網頁瀏覽中醫難字與罕字

由於很多的中醫難字及罕字需要另外安裝字型檔，在我們製作的過程中已經預先安裝完成，並經由 Dreamweaver 輸入這些文字後，瀏覽到網頁即能正常

讀取，此時我們再將瀏覽器視窗利用電腦的視窗擷取圖像功能，擷取圖片至 PhotoImpact。

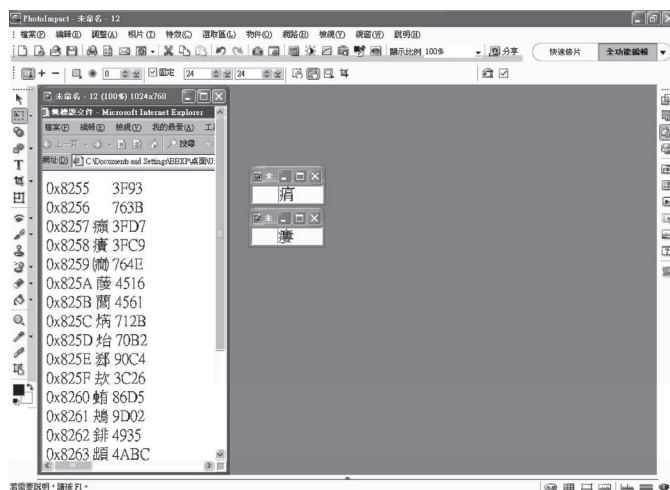
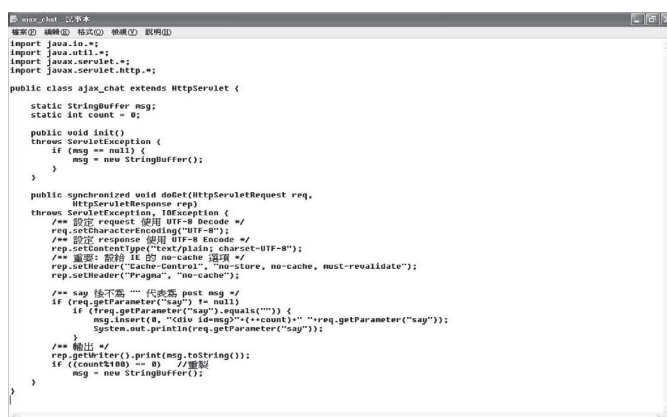


圖5. 將古難字與罕字從圖片中取出成小圖示

在 PhotoImpact 中，我們可以將一整張網頁的圖像內的文字利用 PhotoImpact 擷取的元件，進行文字截取的動作，並

將截取出來的文字小圖示，依照編碼為檔名並進行存取。

建立中醫典籍檢索平台與中醫古難字解決方案



```
src/ajax_chat/ajax_chat.java
import java.io.*;
import java.util.*;
import javax.servlet.*;
import javax.servlet.http.*;

public class ajax_chat extends HttpServlet {

    static StringBuffer msg;
    static int count = 0;

    public void init()
        throws ServletException {
        if (msg == null) {
            msg = new StringBuffer();
        }
    }

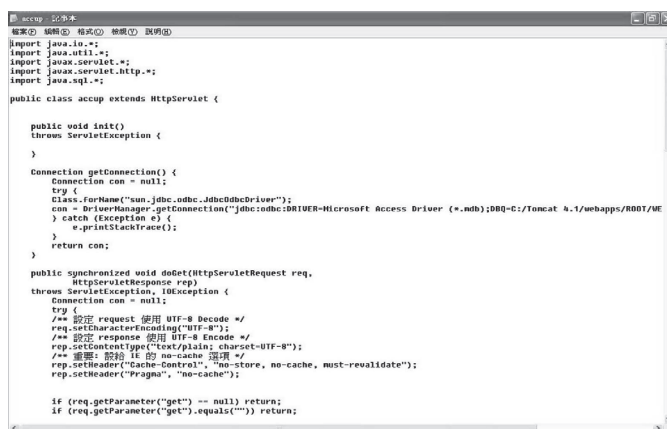
    public synchronized void doGet(HttpServletRequest req,
        HttpServletResponse rep)
        throws ServletException, IOException {
        /** 設定 request 使用 UTF-8 encode */
        req.setCharacterEncoding("UTF-8");
        /** 設定 response 使用 UTF-8 encode */
        rep.setContentType("text/plain; charset=UTF-8");
        /** 重覆: 設給 IE 的 no-cache 選項 */
        rep.setHeader("Cache-Control", "no-store, no-cache, must-revalidate");
        rep.setHeader("Pragma", "no-cache");

        /** say 送不高, 代表當 post msg */
        if (req.getParameter("say") != null) {
            if (req.getParameter("say").equals("")) {
                msg.insert(0, "div id=msg" + (++count) + " " + req.getParameter("say"));
                System.out.println(req.getParameter("say"));
            }
        }
        /** 輸出 */
        rep.getWriter().print(msg.toString());
        if (count % 100 == 0) // 重製
            msg = new StringBuffer();
    }
}
```

圖6. Encoding System Program

我們必須經由Java來設定程式導出後的網頁語言為Unicode，並且設定瀏覽器中不要啟用快取選項，讓資料保證每次都是最新的資訊，且為了避免使用者因

為網路緩慢或中斷的問題，造成資料的傳送不完整與遺失，才會進行這樣的設定。



```
src/acup/acup.java
import java.io.*;
import java.util.*;
import javax.servlet.*;
import javax.servlet.http.*;
import java.sql.*;

public class acup extends HttpServlet {

    public void init()
        throws ServletException {
    }

    Connection getConnection() {
        Connection con = null;
        try {
            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
            con = DriverManager.getConnection("jdbc:odbc:DRIVER=Microsoft Access Driver (*.mdb);DBQ=C:/Tomcat 4.1/webapps/ROOT/WE");
        } catch (Exception e) {
            e.printStackTrace();
        }
        return con;
    }

    public synchronized void doGet(HttpServletRequest req,
        HttpServletResponse rep)
        throws ServletException, IOException {
        Connection con = null;
        try {
            /** 設定 request 使用 UTF-8 encode */
            req.setCharacterEncoding("UTF-8");
            /** 設定 response 使用 UTF-8 encode */
            rep.setContentType("text/plain; charset=UTF-8");
            /** 重覆: 設給 IE 的 no-cache 選項 */
            rep.setHeader("Cache-Control", "no-store, no-cache, must-revalidate");
            rep.setHeader("Pragma", "no-cache");

            if (req.getParameter("get") == null) return;
            if (req.getParameter("get").equals("")) return;
        }
    }
}
```

圖7. Search the Database Program

取得文字編碼後，進入資料庫進行查詢文字編碼的動作，若此編碼為Unicode或Big5中沒有的文字，系統將會

自動從圖像檔中找出相對應之編碼的圖片。

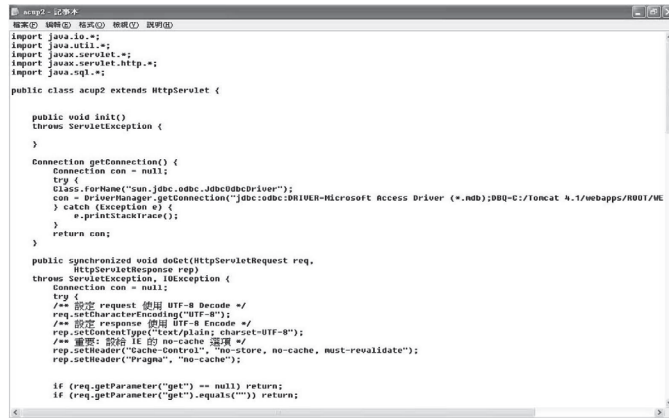


圖8. Convert the Data To XML

為了讓AJAX能夠順利導讀，我們必須將找到的文字從Database取出的文字以及從圖像檔中找到的圖片數據，轉為

XML資料並傳送給AJAX來接收，並藉由AJAX的分析，顯示出來。

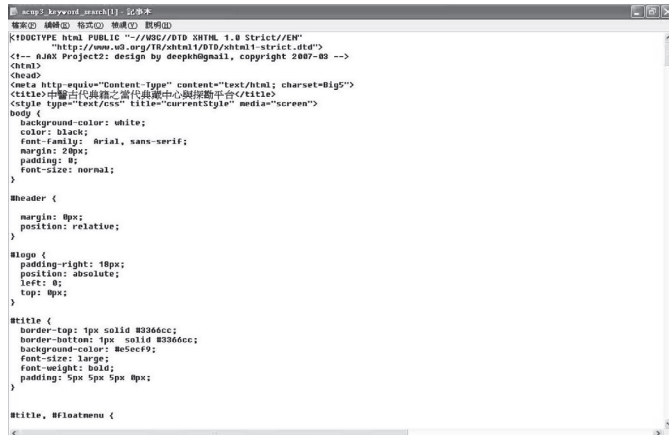


圖9. AJAX Source Program

AJAX是建立在使用者端的網頁語法上，並經由AJAX的作用，能即時的與後端程式與研JAVA進行聯繫，並經由後端程式語言將處理好的文字與圖片連結，

傳給AJAX做分析與處理，最後將成果顯示於使用者端的網頁上，在過程中使用者絲毫感覺不出來網頁正與後端程式進行溝通與查詢的動作。

二、中醫古籍的即時檢索平台

系統以AJAX技術建立中醫典籍探勘平台，系統包括『以年代查詢、以書名查詢、以作者查詢、以性質分類查詢、以關鍵字查詢』，相關的網址可以參閱：

中醫古籍以科別分類：

<http://210.70.108.141:8080/acup3.htm>

中醫古籍以作者查詢：

http://210.70.108.141:8080/acup3_author_search.htm

中醫古籍以書名查詢：

http://210.70.108.141:8080/acup3_book_search.htm

中醫古籍以年代分類：

http://210.70.108.141:8080/acup3_year_search.htm

中醫古籍關鍵字查詢：

http://210.70.108.141:8080/acup3_keyword_search.htm

中醫古籍簡體版查詢：

http://210.70.108.141:8080/acup3_1_chs.htm

在處理線上關鍵字查詢方面，由於考慮時效問題，本系統採用index方式做查尋，先將約2000常見關鍵字建檔於資料庫中，並將這些關鍵字所關聯到的context 做一個表並附上一個專屬ID，若使用者輸入關鍵字不存在於此index時，再動態用上面的方式動態再產生一個新的表。最後輸出只要依關鍵字去抓取這個表所關聯到的 context 就可以了。

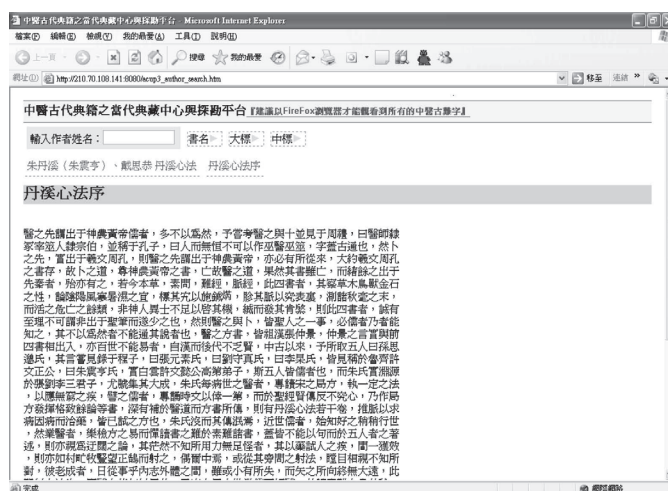


圖10. 以科別分類

明道學術論壇 5(1)：18-33(2009)

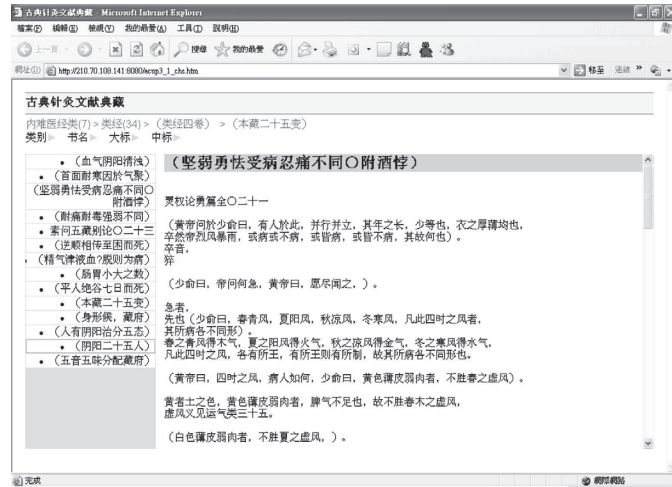


圖11.以科別分類(簡體版)

使用者可以經由我們的設計的程式，共分兩套一套是繁體中文版，另外一套則是採用簡體中文版，使用Java 的

Zhcode 即時做翻譯動作，並由AJAX即時導入Web頁面內。

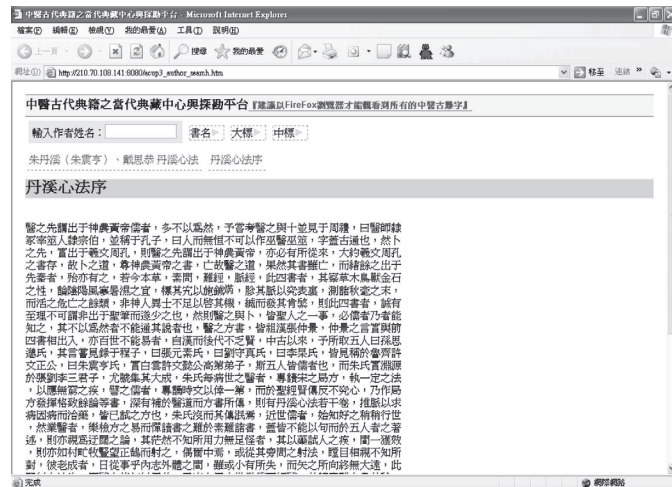


圖12.輸入作者姓名、書籍、年代查詢

建立中醫典籍檢索平台與中醫古難字解決方案

使用者可以藉由系統輸入姓名、書籍、年代來查詢古籍資料，並藉由AJAX即時導入文章，以及將古難字藉由圖片方式呈現在使用者畫面上，而不會有亂碼文字造成。

本研究所處理的古難字，基本上有兩種顯示方式，如下所述：

1.以圖片顯示古難字，例如下面圖片所顯示的『𪛗』字

分類▶ 書名▶ 大標▶ 中標▶

針灸推拿類(7) 針灸大成(14) 針灸大成 卷三

靈光賦 針灸大全

黃帝岐伯針灸訣，依他經裏分明說，三陰三陽十二經，更有兩經分八脈，靈光典註極幽深，偏正頭痛瀉列缺，睛明治眼努肉攀，耳聾氣閉聽會間，兩鼻𪛗針禾膠，鼻塞不聞迎香間，滯氣上壅足三里，天突宛中治喘痰。

2.以罕見unicode字碼顯示，例如下面圖

片所顯示的『𪛗』字

分類▶ 書名▶ 大標▶ 中標▶

針灸推拿類(7) 針灸大成(14) 針灸大成 卷三

靈光賦 針灸大全

黃帝岐伯針灸訣，依他經裏分明說，三陰三陽十二經，更有兩經分八脈，靈光典註極幽深，偏正頭痛瀉列缺，睛明治眼努肉攀，耳聾氣閉聽會間，兩鼻𪛗針禾膠，鼻塞不聞迎香間，滯氣上壅足三里，天突宛中治喘痰。

雖說可以顯示古難字，但在微軟公司的系列產品（例如：word、I.E.網頁瀏覽器），在支援罕見unicode字碼會有錯誤發生，也就是會出現『□』字，例如

會將『𪛗』字變成『□』顯示了，如下圖所示。相信這是目前微軟公司的系列產品目前的一些小瑕疵，在日後的版本即可改善這些缺點了。

針灸推拿類(7) 針灸大成(14) 針灸大成 卷三

靈光賦 針灸大全

黃帝岐伯針灸訣，依他經裏分明說，三陰三陽十二經，更有兩經分八脈，
靈光典註極幽深，偏正頭痛瀉列缺，睛明治眼努肉攀，耳聾氣閉聽會間，
兩鼻齕口針禾膠，鼻塞不聞迎香間，滯氣上壅足三里，天突宛中治喘痰。

但是上述的『罕見unicode字碼』
(例如『齕』字)，若以FireFox網頁瀏
覽器，或是PDF格式檔案、Flash格式檔
案來瀏覽，則皆能正確顯示無誤。

本研究的中醫古籍除了提供可以線
上瀏覽檢索查詢外，也在每個網頁提供
下載每本古籍的電子書，包括：PDF、
Flash、Word等三種格式檔案。

此外，本研究也提供類似搜尋引擎
效果的關鍵字查詢子系統，當使用者輸
入關鍵字後，系統將會藉著AJAX的即時
特性，進入 Database中查詢相關索引關
鍵字，如果沒有在系統預設內建中，將
會採用動態模式進行查詢，並加入新索
引。

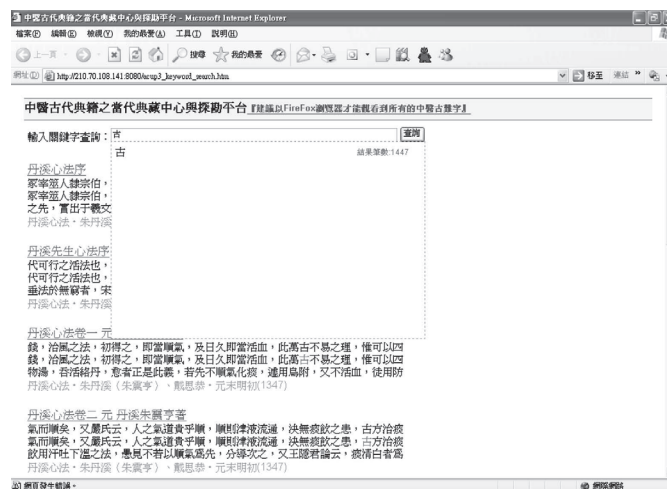


圖13. 輸入關鍵字查詢

建立中醫典籍檢索平台與中醫古難字解決方案

本系統目前已經建立中醫古籍共74本，包括有：醫學發明、金匱鉤玄、十四經發揮、備急千金要方、千金翼方、經效產寶、外臺秘要、太平惠民和劑局方、小兒藥證直訣、婦人良方、神農本草經、審視瑤函、溫疫論、類經、景岳全書、針灸大成、濟陰綱目、針灸問對、女科證治準繩、外科正宗、針灸節要、針灸聚英、素問、靈樞、難經、脈經、針灸甲乙經、神農本草經、傳青主女科、吳鞠通醫案、溫病條辨、理瀉駢文、女科切要、血證論、胎產指南、重訂產孕集、張氏醫通、廣溫疫論、盤珠集胎產症治、類證治裁、溫熱逢源、醫方集解、沈氏尊生書、雜病源流犀燭、外科大成、三時伏氣外感篇、臨證指南醫案、溫熱論、女科經綸、溫熱病篇、外感溫熱篇、時病論、本草求真、女科百問、傷寒心要、傷寒標本心法類萃、傷寒直格、素問玄機原病式、素問病機氣宜保命集、黃帝素問宣明論方、傷寒心鏡別集、儒門事親、蘭室秘藏、內外傷辨、此事難知、脾胃論、諸病源候論、針灸資生經、本草問答。

此資料庫具有擴充性，目前正以每年40本的速度新增下去。

四、結論

在中醫古籍文字缺字的解決方案，在國內外皆沒有一套完整的解決方案，有鑑於此我們提出了以文字圖片方式來顯示缺字，並利用系統程式自動製造這

些缺字圖片，有效節省人工的造字或製圖，以人工方式，則需要好幾個月，甚至幾年的時間，才有以程式造字的產量，且以人工方式容易造成編碼的錯誤，並不利於程式的執行與解析。

利用文字圖片來顯示缺字編碼，是目前不需要額外安裝字型就可以讀取的方式，但是使用者並無法自行輸入這些文字，像是：倉頡、新注音…等等，使用者只能顯示這些文字編碼，而無法再進一步的轉述張貼至其他地方。

在中醫古籍查詢系統方面，我們則進行了很多的古籍分類，也將這些分類按照使用者的閱讀習慣製作了不同的版本，讓使用者可以很方便的選擇自己所習慣的模式來閱讀，且搭配Unicode古難字補足系統，在文章的顯示方面，已經可以把全文顯示完整，而不會有缺字情形發生。

以目前的中醫古籍查詢系統，除了使用AJAX模式導入文章，另外也提供每本古籍的三種格式電子書，可以自由下載，方便儲存與列印瀏覽。

本研究成功地提出一個中醫典籍古難字的解決方案，與建立中醫古籍的檢索引擎與探勘平台，希望能對中醫做點棉薄貢獻。

五、致謝

本研究之經費來自於國科會(編號 NSC 95-2422-H-434 -001 -)補助之專題研究計畫，謹此致謝。

六、參考文獻

- 行政院衛生署中醫藥委員會（2008），中醫藥造字檔下載及安裝使用說明，2008年8月20日，取自<http://www.ccmp.gov.tw/public/public.asp?selno=525&relno=525&level=C>
- 台灣行政院主計處電子處理資料中心（2008），CNS11643中文全字庫，2008年8月20日，取自<http://www.cns11643.gov.tw/web/index.jsp>
- 莊德明、謝清俊、林晰（1998），中央研究院古籍全文資料庫解決缺字問題的方法，第二次兩岸古籍整理研究學術研討會，30-36，北京。
- 謝清俊（1996），電子古籍中的缺字問題，第一屆中國文字學會學術討論會，23-29，天津。
- 謝清俊（1997），談古籍檢索的字形問題，珍藏文獻整理與資訊科技應用研討會，15-21，台北。