建立中醫古籍檢索系統暨 Unicode 古難字補足系統

蕭燿晟
致遠管理學院 資訊工程學系
laymice@gmail.com

陳擎文
致遠管理學院 資訊工程學系
ccw@dwu.edu.tw

ABSTRACT

This research imposes the technologies of ASP.NET, AJAX, XML and SQL Server to establish the supplement system of Unicode for the ancient rare Chinese characters as well as the search engine and retrieved system of Chinese medicine ancient book.

In the research, we establish a set of Character Array for the Unicode writing. Then, the algorithm of sequential search is applied to find the corresponding characters. The Unicode of that character is examined whether it belongs to the character code of user's font. It is substituted for the default picture when Unicode of that character in the code of user's font. On the contrary, it is displayed on-line with mode of picture-creating.

The search engine and retrieved system of Chinese medicine ancient book are established with the methods of ASP.NET, AJAX, XML and SQL Server to achieve the real time effect of retrieve. The whole system could display all the rare characters of ancient books without the messy codes or deficient characters.

Keyword: Chinese Medicine, XML, AJAX, Unicode

摘要

本研究提出一個中醫典籍古難字的解決方案，並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。系統乃是採用 ASP.NET、AJAX、SQL Server 及 XML 來建構線上 Unicode 古難字補足系統及古書籍查詢系統。

在研究中，我們將現有的 Unicode 文字，建立一套 Character Array，並利用搜尋法，從文章內找出對應之文字，並檢查此 Unicode 是否為使用者造字區之內碼，若是則採用系統預設建立好的文字圖庫圖片取代，若系統中沒有內建圖片，則採用線上製圖模式。中醫古籍查詢系統，則採用 XML 與 SQL Server 及利用 ASP.NET 與 AJAX 的相互結合，來達成快速閱讀古籍的效果，並搭配 Unicode 古難字補足系統，將可以完整呈現古籍中的所有古難字，而不會有亂碼或缺字的情形發生。因應不同需求，系統提供 6 種不同功能的檢索平台，也提供簡體字的檢索與瀏覽。

本研究改良行政院衛生署中醫藥委員會的 94 年版『中醫藥造字檔』，且使用者不需要另外再安裝任何字型檔即可完整瀏覽古難字或罕字。

關鍵字：中醫，典籍罕字，XML、AJAX、Unicode 編碼

一、 前言

1.1 中醫古難字 Unicode 之線上補足系統

1.1.1 文字起源與電腦編碼

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。

不管在業界或者行政單位，如：警政署、戶政機關，在文字的缺碼處理，尤為重要，一旦有任何文字缺字或亂碼，將會造成這項公文失去法律效力，而無法正常執行公權力。

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，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，要統一是相當困難的，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。

在這個研究中，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。

執行編碼的補完計畫，我們會在系統中安裝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」』

1.1.2 中文字的組字原則

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。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，其中包含括 1931 個 Big5 找不到得古文字。

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。

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。

1.1.3 缺字的發生

在任何非英語系國家的電腦中，缺字問題必然會發生，其發生原因，多半是人名、古籍資料，人的命名為了生辰八字，有可能會自行造字，而這樣的字，以往的時代是沒有的，所以樣的情況下，電腦自然不可能收錄進去，而在古籍方面，中文字並沒有像英文只有 26 個英文字母，而是有相當多的文字所組合而成，在人工蒐集上，能力有限，加上有相當多的古文字，至今仍舊不斷被發現出來，或者被創造出來，在電腦的支援上，遠不及發現與創造的速度，自然而然就產生了缺字問題。

缺字問題對於一般使用者並不會造成太大的困擾，但是對於古籍閱讀者，或者警政機關、戶政機關，在處理資料時，或者公司間的文件交換，就會造成很大的困擾，一字之差都可能會造成相當重的嚴重性，以戶政機關來講，若有一個字出現錯誤，則該文件將不具有法律效力，當然被通知者也有權利不需要繳費或受罰。

1.2 中醫古籍查詢系統

1.2.1 中醫古籍查詢系統與 AJAX

本研究所建立的中醫古籍查詢系統，乃是採用 ASP.NET、AJAX 及 XML 網路本體論，在系統中，我們採用文件預知功能，在讀取到相關分類時，能夠將可能開啟的書籍預先載入，讓系統能夠在最順暢的情況下運作，讓使用者感受不到系統間的資料傳輸與交換，同時對於龐大的資料集，也能夠降低傳輸的時間消耗。

若發現有資料沒有事先經由系統載入，則會藉由 AJAX 以非同步方式，進入系統中，並呼叫 ASP.NET 來解析 XML，並將解析好的文字，再經由 AJAX 來接收資料，最後顯示於使用者螢幕上。

由於 AJAX 有使用瀏覽器暫存功能，只要瀏覽過一次古籍資料，系統將會自動寫入暫存，待下次使用者在度開始時，速度將會是原來的好幾倍快。

另外在許多情況下，我們需要利用到搜尋引擎這種工具，但在中醫古籍查詢系統，想要藉著 Google 或者 Yahoo 的搜尋引擎來查詢，是相當不方便，且資訊是相當不精確的，在此我們也製作了中醫古籍的關鍵字查詢系統，每當使用者輸入一個關鍵字後，將會使用 AJAX 進入資料庫中，找尋相關的關鍵字，並同時顯示出這些關鍵字所擁有的資料筆數，再經由使用者決定好最終關鍵字後，將會進入完整的資料頁面。

1.2.2 與電子書的比較

電子書形式多樣，常見的有 TXT 格式，DOC 格式，HTML 格式，CHM 格式，PDF 格式等。這些格式大部分可以利用微軟 Windows 作業系統自帶的軟體打開閱讀。至於 PDF 等格式則需要使用其他公司出品的一些專用軟體打開，這其中有著名的免費軟體 Adobe Reader。

支持電子書的軟體一般都支持「查找」、「書籤」、「筆記」等擴展功能，這使得用戶可以更專注於內容本身，而不必考慮其他附帶問題如筆記本電腦丟失，忘記資料等。而且「查找」功能更是可以

在極短時間內完成傳統讀圖書者需要十幾秒甚至更久來完成的資料查找。

在傳統網頁電子書中，我們並沒有辦法達成像在電腦端開啟檔案一樣的快速，且沒有像電腦一樣的預覽功能，使用者常常需要上一頁下一頁的不斷切換畫面，在查找資料上是相當不方便的。

有鑑於此我們的中醫古籍查詢系統，就針對這個問題來做改善，我們增加了文件的預覽導讀，讓使用者可以在最快速的時間預覽過想要看的主題，我們也將這些古籍資料建立一個好的分類制度，使用者只需要依循分類，就可以找到屬於自己想要閱讀的文章。

在工具方面，中醫古籍查詢系統也不像 PDF、Word 這樣麻煩，需要另外安裝導讀軟體，我們只需要作業系統有支援瀏覽器，無論 Internet Explorer 或者 FireFox …等等瀏覽器，均可以正常讀取網頁，並可經由滑鼠來選擇分類，或由鍵盤輸入關鍵字查詢相關文章。

在標籤方面，中醫古籍查詢系統，也會在以即時的方式顯示出使用者目前所在位置，所點選的分類為何，讓使用者不會因為不小心將滑鼠移開，而不知道文件的所在位置，又要重頭找起。

1.2.3 使用 XML 儲存的優勢

XML 為「可擴展標示語言」(eXtensible Markup Language)，其規格由 W3C 所制定，並於 1998 年 2 月成為推薦規格，現金有許多的廠商採用，視為關鍵性技術，如：Adobe、IBM、Microsoft，而常見的支援軟體則有：Navigator、Internet Explorer、RealPlayer

使用一個標準的格式，可以使系統發展者不用費心設計資料交換的格式，而專注於更重要的問題上，例如：資料結構、資料類別、如何使用這些資料。

XML 是 SGML（Standard Generalized Markup Language，即 ISO 的標準通用標示語言，ISO 8879:1986 和其 1997 WebSGML 附件 AnnexesJ、K 及 L）的簡化版本。SGML 系統是非常複雜的，因為它具有許多的機制，以便提供各種的語法。SGML 是被發展用來解決編輯及保存內容龐大複雜且互相連結的技術文件，而 XML 只取用 SGML 系統中的文件結構的核心部份。

XML 規格制定的初衷是為了簡化程式設計的困難度，目前有相當多款的工具程式，可以讀取並修改 XML 文件檔，也能夠採以樹狀結構方式來讀取文件，是相當方便的一種格式。

我們使用標準的格式來儲存古籍資料，將有利於未來的資料交換，無論在手機系統、PDA 系統、Linux 系統下，皆能很快的交換資料。

二、 研究方法

2.1 中醫古難字 Unicode 之線上補足系統

2.1.1 Unicode 編碼及組字規則

Unicode 的編碼方式與 ISO 10646 的通用字符集（Universal Character Set，UCS）概念相對應，目前實際應用的 Unicode 版本對應於 UCS-2，使用 16 位的編碼空間。也就是每個字元占用 2 個位元組。這樣理論上一共最多可以表示 2^{16} 即 65536 個字元。基本滿足各種語言的使用。實際上目前版本的 Unicode 尚未填充滿這 16 位編碼，保留了大量空間作為特殊使用或將來擴展。

Unicode 的實現方式不同於編碼方式。一個字元的 Unicode 編碼是確定的。但是在實際傳輸過程中，由於不同系統平臺的設計不一定一致，以及出於節省空間的目的，對 Unicode 編碼的實現方式有所不同。Unicode 的實現方式稱為 Unicode 轉換格式（Unicode Translation Format，簡稱為 UTF）。

例如，如果一個僅包含基本 7 位 ASCII 字元的 Unicode 文件，如果每個字元都使用 2 位元組的原 Unicode 編碼傳輸，其第一位元組的 8 位始終為 0。這就造成了比較大的浪費。對於這種情況，可以使用 UTF-8 編碼，這是一種變長編碼，它將基本 7 位 ASCII 字元仍用 7 位編碼表示，占用一個位元組（首位補 0）。而遇到與其他 Unicode 字元混合的情況，將按一定演算法轉換，每個字元使用 1-3 個位元組編碼，並利用首位為 0 或 1 進行識別。這樣對以 7 位 ASCII 字元為主的西文文檔就大大節省了編碼長度。類似的，對未來會出現的需要 4 個位元組的輔助平面字元和其他 UCS-4 擴充字元，2 位元組編碼的 UTF-16 也需要通過一定的演算法進行轉換。

由於 Unicode 有這樣的特性，所以當我們在製作文字內碼反查，及查找資料時，會造成速度上的緩慢，為了加快文章的文字辨識速度我們將一些罕見的古難字，建立一個字元索引(Index)，並在程式執行時，導入系統中，接著利用字元陣列(Character Array)，再依照循序搜尋法(Sequential Search)方式，導入一篇文章中，找尋出以建立在字元陣列中的文字，並進入預設建立的字元圖庫中，找尋是否有相對應的文字圖片，若沒有則呼叫 ASP.NET 內所內建的繪圖函數，即時繪圖存取並修改字元陣列索引，最後將文章中相關文字利用圖片取代，並顯示於使用者端。

簡體中文的對照上，我們則依據 Big5 為譜，與 GB2312 為簡體內碼對照轉換，至於中醫古難字的圖片方面，則維持原來的圖片，而不再做另外的轉換。

為了解決正體中文與簡體中文的字碼不一情

況，我們在轉換完成後，必須再轉換為 Unicode 來觀看，讓使用者的系統語言只需要支援 Unicode 即可。

2.1.2 線上 Unicode 古難字補足系統研究步驟

步驟 1：將 Unicode 文字與 Big5 文字，以及所安裝字型之文字，全部列出並導入 TXT 記事本中，再經由程式來找出相同的文字，並取代掉，只留下唯一的一個。

步驟 2：確定文字為唯一值時，再經由程式將 Big5 文字與 Unicode 文字，逐一刪除，最後留下所安裝之字形檔的文字。

步驟 3：經由程式讀取這些文字的內碼，再啟動 ASP.NET 內建的繪圖函式庫，來繪出這些文字的圖片，但前提是這台繪圖所用電腦必須有安裝這些字形檔，否則繪出會是空白圖片，繪出後將圖片以內碼編號為檔名。

步驟 4：將文章導入系統中，找出非 Big5 與 Unicode 的內碼文字，並將這些內碼文字以文字圖片方式顯示於使用者端。

步驟 5：若使用者選擇使用簡體中文導讀，系統則利用 Big5 碼與 GB2312 碼為轉換要點，轉換後將 GB2312 再轉為 Unicode 方便使用者閱讀。

2.2 中醫古籍查詢系統

2.2.1 傳統查詢書籍與非同步查詢書籍比較

在傳統的書籍查詢網站，我們只能利用框架(Frame)，或者利用單頁連結以達成查詢效果。

這樣的方式，在查詢上是淺顯易懂，但是並不利使用者在閱讀時的連結，以及網路的傳輸速度亦有影響，然而這些頁面的轉換，也造就了一些不必要的頻寬浪費，並非最佳的閱讀方式。

在這項研究中，我們採用了 AJAX 的同步傳輸，以及 AJAX 的非同步傳輸，雙向並用，在使用者一進入中醫古籍查詢系統時，將會利用 ASP.NET 進入 XML 資料集中，查詢在分類中最高的點閱書籍，使用 XML 的格式，傳輸給 AJAX 系統，並利用 AJAX 來分析這些文件的相關意義，最後把分析好的文章顯示在使用者畫面上。

而當使用者點擊到沒有預先載入的文章時，也將採用 AJAX 的非同步傳輸，再度進入 ASP.NET 系統中，來取得相關的書籍或文章。

在頻寬管理上，比傳統的省資源，在於傳統查詢方式，有太多的 HTML 語法重複傳輸，或者在文章中有太多重複的圖片傳輸，而造成了頻寬上的浪費，而使用 AJAX 的同步與非同步傳輸，在系統中只需要傳回相關文章，並不需要將這些圖片、語法重複的傳輸給使用者，亦不會造成頻寬的浪費。

同時為了讓使用者能夠藉著輸入某些關鍵字，而找到相對應之文章，我們也製作了關鍵字查詢系統，在功能方面我們參考 Google 的即時搜尋關鍵字之 AJAX 技術，進而自己研發，由於 Google 採用的關鍵字屬於一般大眾共同使用，在中醫等專業領域內，是相當不適用的一套工具。

為此我們必須量身訂做一套屬於中醫的關鍵字查詢系統，希望能做出類似 Google 卻是專門用來檢索中醫資料的搜尋引擎，我們將一些中醫名詞，先行導入資料庫中，並利用這些中醫名詞進入已建立好的這些 XML 文件中，查詢是否有相關名詞的文章，若有則累計到資料庫的關鍵字。

當然這些關鍵字是相當不夠用的，我們必須開發一套自動增加索引(Index)的程式，來應付廣大的關鍵字資訊，一旦使用者輸入的關鍵字，在系統資料庫內並查詢不到時，我們將開啟動態索引模式，利用使用者的關鍵字，進入資料集中查詢，若有相關資料時，我們將這個關鍵字寫入資料庫中，待下位使用者再度查詢相同關鍵字時，能夠即時的顯示出這個關鍵字有多少筆資料。

2.2.2 中醫古籍查詢系統之研究步驟

步驟 1：將所有古籍資料利用 XML 格式儲存，未來可以提供給手機、PDA、Linux 來讀取資料，也便於未來的資料修改。

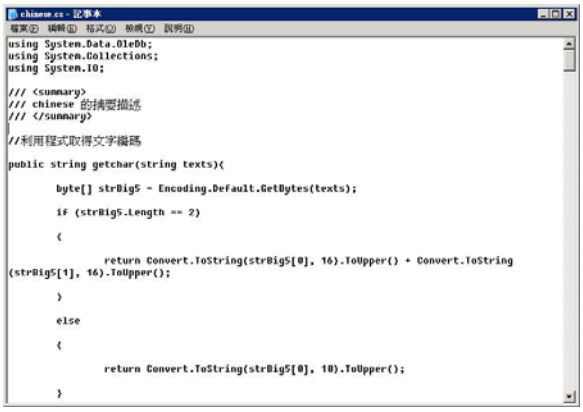
步驟 2：將古籍資料依據作者、書籍、年代、大標、中標來分類，使用者可以依據作者、書籍，或者年代來查詢。

步驟 3：搜集中醫名詞，將中醫名詞導入 SQL Server 並進入 XML 文件集中查詢是否有相關資料，並將資料筆數儲存在資料庫中。

步驟 4：當使用者輸入的關鍵字為資料庫中所沒有的，則進行動態搜尋，找出相關的資料，並進入將索引重新調整並加入新關鍵字。

三、 研究過程

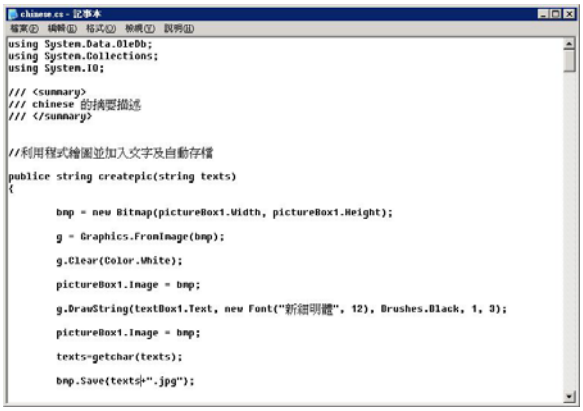
3.1 線上 Unicode 古難字補足系統



```
using System.Data.OleDb;
using System.Collections;
using System.IO;

/// <summary>
/// chinese 的摘要描述
/// </summary>

//利用程式取得文字編碼

public string getchar(string texts){

        byte[] strBig5 = Encoding.Default.GetBytes(texts);

        if (strBig5.Length == 2)

        {

                return Convert.ToString(strBig5[0], 16).ToUpper() + Convert.ToString
(strBig5[1], 16).ToUpper();

        }

        else

        {

                return Convert.ToString(strBig5[0], 10).ToUpper();

        }
```

圖 1 利用程式取得文字內碼

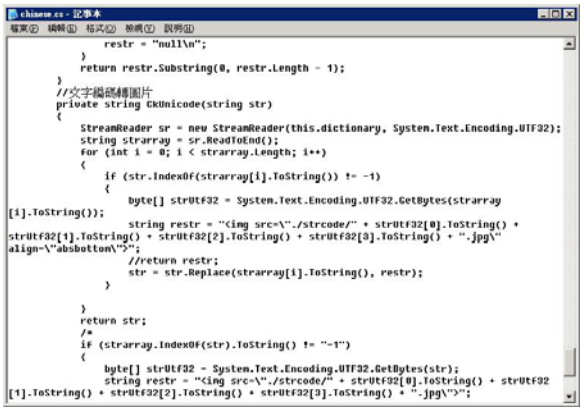
將取得文字編碼的程式，寫成一個函數(Function)並設為公用，提供所有需要呼叫到編碼的程式使用。



```
using System.Data.OleDb;
using System.Collections;
using System.IO;

/// <summary>
/// chinese 的摘要描述
/// </summary>

//利用程式繪圖並加入文字及自動存檔

public string createpic(string texts)
{

        bmp = new Bitmap(pictureBox1.Width, pictureBox1.Height);

        g = Graphics.FromImage(bmp);

        g.Clear(Color.White);

        pictureBox1.Image = bmp;

        g.DrawString(textBox1.Text, new Font("新細明體", 12), Brushes.Black, 1, 3);

        pictureBox1.Image = bmp;

        texts=getchar(texts);

        bmp.Save(texts+".jpg");
```

圖 2 利用編碼製作圖片

在文字圖片資料庫中沒有的文字，我們將取得編碼後，利用程式線上製圖的方式，及時製作一張圖片，並儲存於伺服器系統中。



```
            restr = "null\n";
        }
        return restr.Substring(0, restr.Length - 1);
    }
    //文字編碼轉圖片
    private string CkUnicode(string str)
    {
        StreamReader sr = new StreamReader(this.dictionary, System.Text.Encoding.UTF32);
        string strarray = sr.ReadToEnd();
        for (int i = 0; i < strarray.Length; i++)
        {
            if (str.IndexOf(strarray[i].ToString()) != -1)
            {
                byte[] strUtf32 = System.Text.Encoding.UTF32.GetBytes(strarray
[i].ToString());
                string restr = "<img src=\"./strcode/" + strUtf32[0].ToString() +
strUtf32[1].ToString() + strUtf32[2].ToString() + strUtf32[3].ToString() + ".jpg\"
align=\"absbottom\">";
                //return restr;
                str = str.Replace(strarray[i].ToString(), restr);
            }

        }
        return str;
        /*
        if (strarray.IndexOf(str).ToString() != "-1")
        {
            byte[] strUtf32 = System.Text.Encoding.UTF32.GetBytes(str);
            string restr = "<img src=\"./strcode/" + strUtf32[0].ToString() + strUtf32
[1].ToString() + strUtf32[2].ToString() + strUtf32[3].ToString() + ".jpg\">";
```

圖 3 利用搜尋將文章中的相關字碼取代

利用循序搜尋法，將文章中的相關字碼，取代為圖片，這些被取代的字碼將是 Unicode 及 Big5 內碼中所沒有的文字，有利於一般沒有安裝其他文字內碼的使用者查詢。

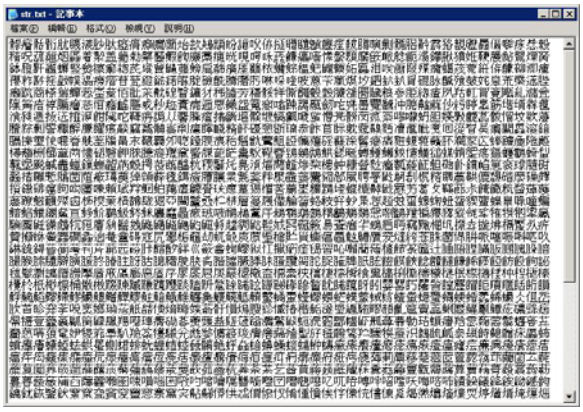


圖 4 文字編碼儲存方式(I)

圖 4 文字編碼儲存方式(II)

在中文字中，2Byte 即為一個中文字，我們利用 C 語言的 char 陣列方式，將這些需要轉換為圖片的文字，儲存在一個記事本中，再經由程式的讀取，轉為 char 陣列，讓程式來查詢與讀取。

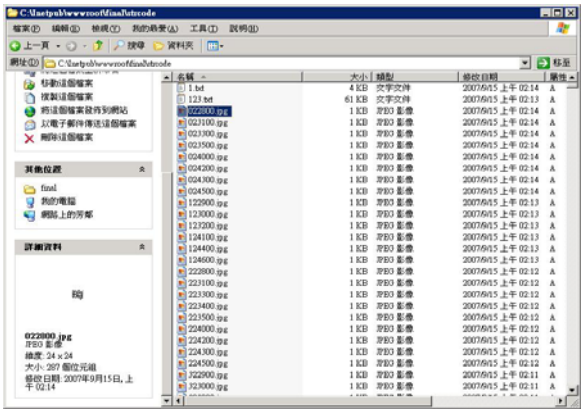


圖 5 文字圖庫目錄(I)

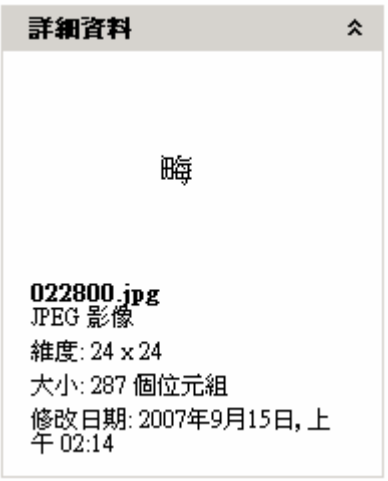


圖 5 文字圖庫目錄(II)

文字的圖片我們採用 Unicode 的內碼來當作圖片的檔名，並按內碼順序排列，且可隨時動態性增加，圖 5(II)是文字圖片的例圖。

3.2 中醫古籍查詢系統

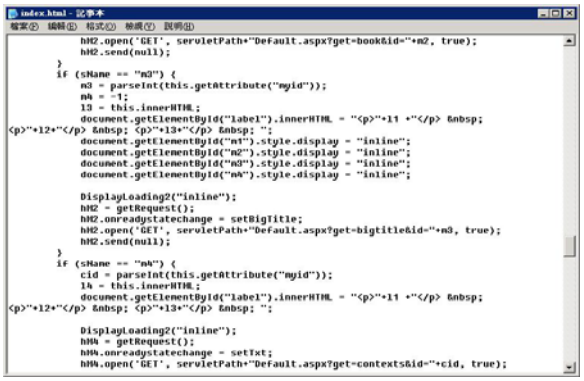


圖 6 網頁 AJAX

在網頁與 ASP.NET 互相傳輸資料的媒介，我們選擇使用 AJAX (Asynchronous JavaScript and XML) 並採用同步傳輸，與非同步傳輸同時運作，讓使用者感受不到頁面的切換，與資料的傳遞。

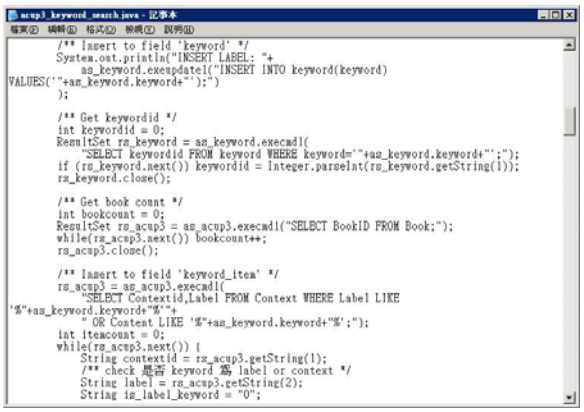


圖 7 關鍵字查詢系統

我們藉由 SQL Server 的資料庫索引，來進行關鍵字的搜尋動作，找到資料庫關鍵字的相關資料時，使用者端將會顯示出：相關關鍵字、相關筆數，並可以提供使用者選擇其他關鍵字，選擇完畢後即可查詢，或者繼續選擇書籍名稱，或作者名稱以及年代，來進行更細部的查詢動作。

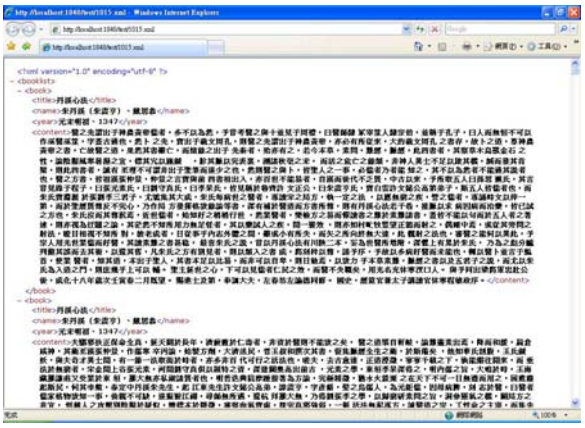


圖 8 XML 格式儲存資料(I)

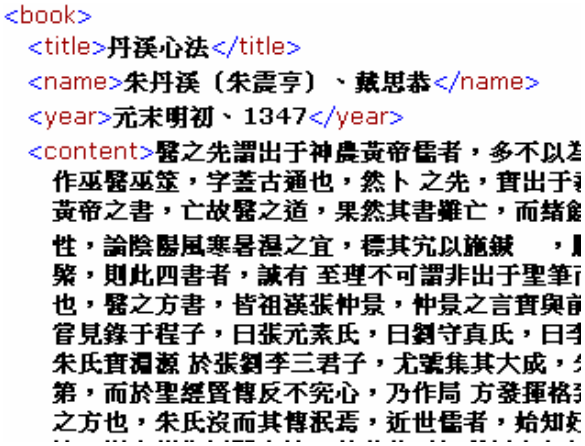


圖 8 XML 格式儲存資料(II)

在古籍資料方面，我們採用 XML 標準格式來儲存，在未來的資料交換上，皆可以有很好的效果，而不會受限於系統的不同，而無法將資料交換。

四、 研究成果

3.1 線上 Unicode 古難字補足系統之研究成果

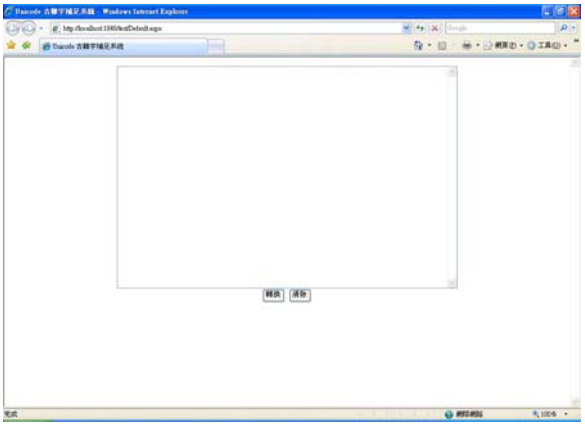


圖 9 線上 Unicode 古難字補足系統

進入線上 Unicode 古難字補足系統時，會先看到一個大的輸入資料欄位，使用者輸入資料後，將會進入到系統中來查詢文字內碼。

圖 10 線上 Unicode 古難字補足系統 使用圖

使用者輸入後文章後，並按下轉換，系統將會自動且快速的將轉換完成的文章顯示於使用者的瀏覽器畫面上。

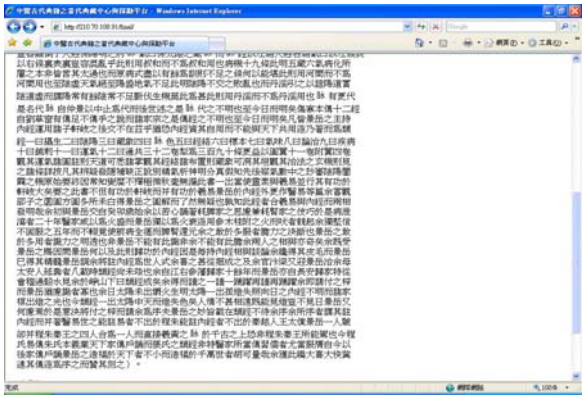


圖 11 文字內碼轉換結果圖(I)

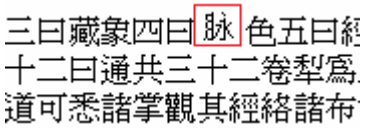


圖 11 文字內碼轉換結果圖(II)

文章導入系統之後，系統將會自動找出相對應之內碼文字，並取代為圖檔方式顯示於使用者端，讓這整篇文章並沒有缺字問題發生，圖 11(II)中的紅色框起來文字，則是被取代的文字。

3.2 中醫古籍查詢系統 研究成果

本研究屬於國科會數位典藏計畫，所有的研究成果可以經由以下網址來進行檢索所有的中醫古籍：

以科別分類：

http://tung.dwu.edu.tw/ajax/AncientBook_Classify.html

以書名檢索

http://tung.dwu.edu.tw/ajax/AncientBook_BookName.html

以作者名稱檢索

http://tung.dwu.edu.tw/ajax/AncientBook_AuthorName.html

以年代檢索

http://tung.dwu.edu.tw/ajax/AncientBook_Year.html

以關鍵字查詢的搜尋引擎

http://210.70.108.141:8080/acup3_keyword_search.htm

以簡體字檢索

http://210.70.108.141:8080/acup3_1_chs.htm

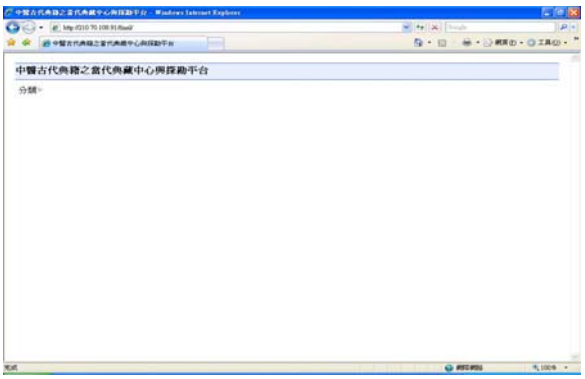


圖 12 中醫古籍查詢系統首頁

當網頁載入完成之後，系統已經將分類的數據全部載入，並等待使用者將滑鼠移到相關分類時，展開並繼續往下層點選書籍與選擇其他標題與內容分類。

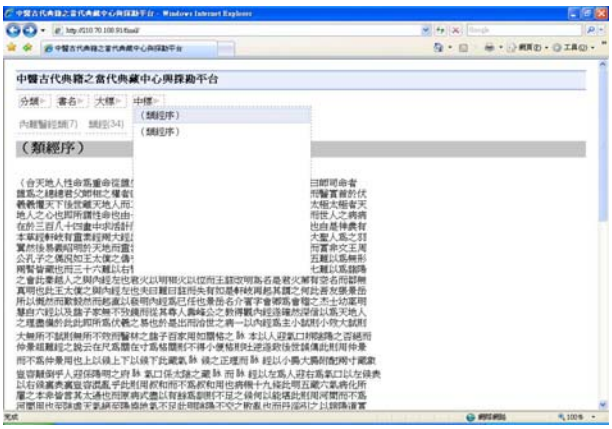


圖 13 滑鼠移過所有分類

當滑鼠移至其他分類時，將會自動載入其分類的內容並搭配線上 Unicode 古難字補足系統，將文章內容快速搜尋一次，將完整的文章顯示於使用者端，而不會有缺字情形發生。

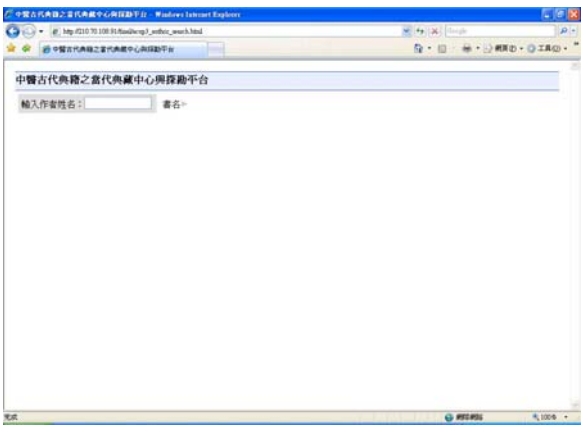


圖 14 作者資料查詢

我們可以利用此系統來進行作者、書籍名稱、年代，來查詢相關書籍內容，若使用者沒有輸入任何關鍵字來進行查詢，系統則自動轉換至書籍查詢系統，使用者仍然可以使用滑鼠輕移來達到查詢與閱讀效果。

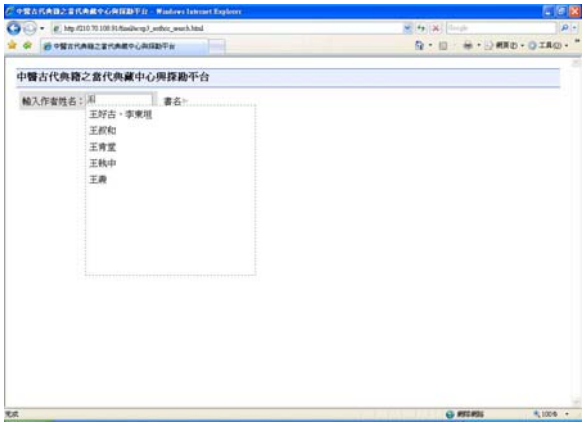


圖 15 作者查詢系統操作圖(I)

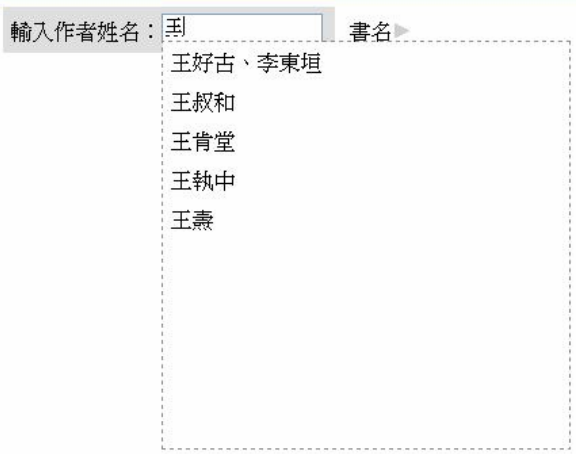


圖 15 作者查詢系統操作圖(II)

當使用者輸入任何關鍵字時，系統將會自動進入資料庫中查詢此關鍵字的相關資訊，並提供使用者來選擇關鍵字或繼續輸入其他關鍵字，在例中，我們輸入「王」為查詢關鍵字。

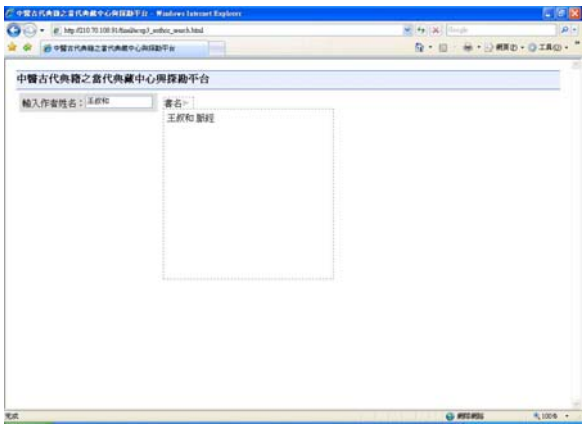


圖 16 輸入完整作者名稱後顯示出相關書籍(I)

圖 16 輸入完整作者名稱後顯示出相關書籍(II)

當使用者輸入完成資料後，右邊將會出現相關的書籍名稱，供給使用者選擇，並繼續往下瀏覽其他書籍的內容，在例中我們以「王叔和」為搜尋關鍵字。

圖 17 選擇想閱讀的相關書籍

當使用者選起完成之後，將會即時出現該文章內容，並可以隨時閱讀其他章節，不管在時間上，與換頁上及資料傳輸上我們的搜尋系統，都比過去的連結方式，要來得快速且資料傳輸量大幅縮小。

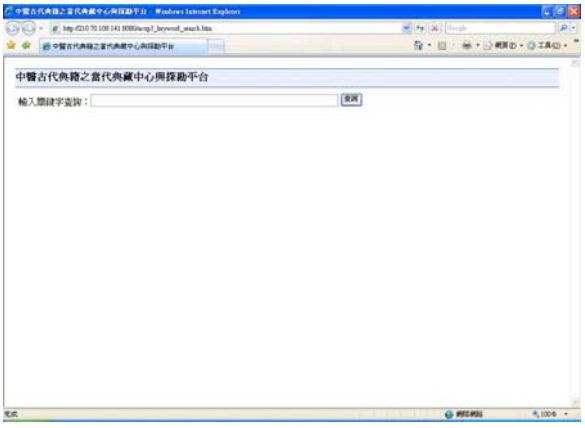


圖 18 以關鍵字查詢文章

利用關鍵字查詢中醫古籍資料，讓使用者在查詢古書籍時，能夠在最快最精確的情況下搜尋到相

關的文章，使用者只要輸入相關詞，系統將會自動進入資料庫中，查詢是否有關聯字詞，使用者仍可以自行輸入，亦可使用滑鼠來選擇系統自動找尋的字詞，來進行查詢。

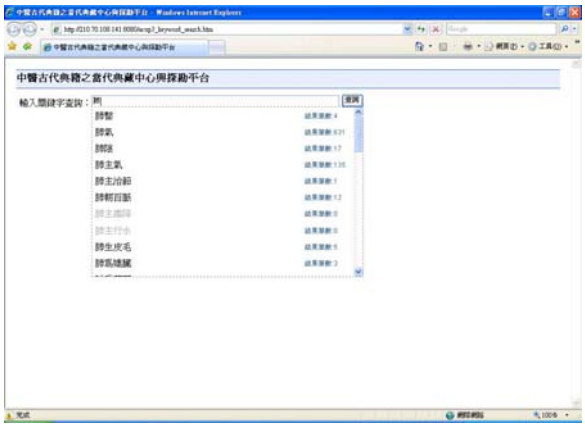


圖 19 輸入關鍵字後顯示出相關名詞(I)

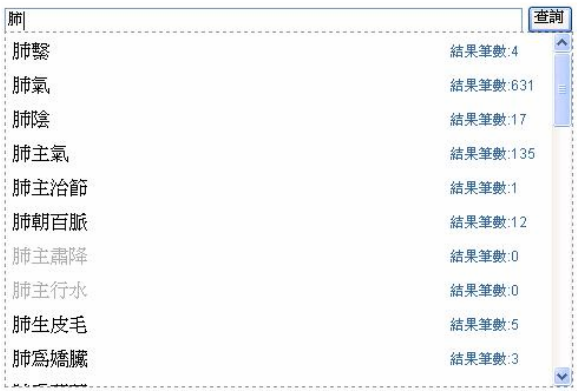


圖 19 輸入關鍵字後顯示出相關名詞(II)

當使用者輸入關鍵字之後，系統將會立刻進入 ASP.NET 中並向 SQL Server 查詢此關鍵字是否有相關的文章，若關鍵字沒有存在於資料庫中，則採用動態索引的方式建立。

圖 20 關鍵字輸入完畢查詢結果圖(I)

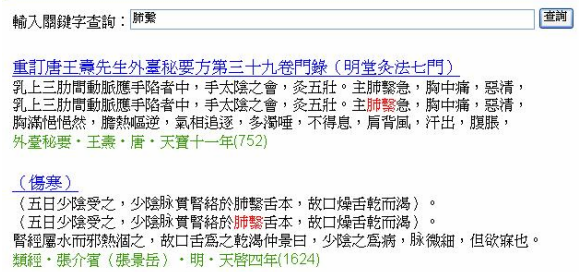


圖 20 關鍵字輸入完畢查詢結果圖(I)

當使用者輸入完成關鍵字後，並按下 Enter 將再度使用 AJAX 進入 ASP.NET 並向建立好的 XML 查詢相關關鍵字之文章標題，並顯示於使用者畫面中，再經由使用者來選擇想要觀看的文章。

圖 21 中醫古籍詳細文章

使用者按下全文觀看之後，系統將會導入詳全文給使用者觀看，此時使用者可以選擇另開視窗，或者直接使用原視窗開啟。

圖 22 以表格方式排版(I)

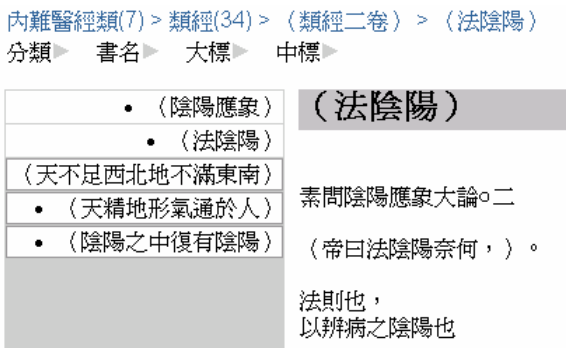


圖 22 以表格方式排版(II)

使用表格方式排版頁面，更符合大多數使用者的閱讀習慣，使用者可以利用滑鼠點擊來選擇欲閱讀文章，此排版方式與 **圖 12** 的不同在於 **圖 12** 當使用者滑鼠移開後，使用者要選擇下一章節或其他書籍時，並沒有一個良好的索引，使用者容易造成錯亂，或者找不到文章的窘境，但是在快速導讀時，卻又以 **圖 12** 來得方便。

本系統中也可以自由的切換版本語言，若使用者所要檢索的語言是簡體版之中醫古籍查詢系統，則系統可利用 Big5 碼與 GB2312 碼的轉換關聯，轉換後，再將 GB2312 再轉為 Unicode 方便使用者閱讀。

圖 23 簡體版的中醫典籍檢索系統(I)

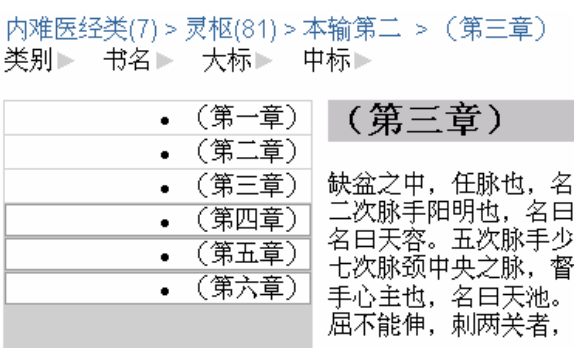


圖 23 簡體版的中醫典籍檢索系統(II)

使用者選擇使用簡體中文來閱讀時，在古難字的處理方面，仍舊採用 Big5 與 Unicode，這是由於古難字的相關文字，不管在 Big5 或者 GB2312 皆無收錄，所以並無須要進行轉碼動作。

五、 結論

在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，由於使用 AJAX 模式導入文章，使用者將無法列印相關文章，必須將其文章複製到如 Word 、 記事本，才能夠進行列印，在未來我們也將針對這部分著手進行相關研究，找出一個可行且便捷的方式，讓使用者可以自由的取得這些寶貴的古籍文獻。

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。

六、 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。

七、 參考資料

[1] 周亞民、黃居仁，“數位典藏缺字問題的解決”， 景文技術學院資訊管理學系

[2] 邱菊梅、陳郁夫，“中文電腦缺字研究”， 中國語文研究所，玄奘人文社會學院

[3] 莊德明、謝清俊、林晰，中央研究院古籍全文資料庫解決缺字問題的方法，1998 年五月。

[4] 陳玉芳(1999)，「政府機關使用 CNS 11643 中文

[5] 謝清俊，談古籍檢索的字形問題，1997 年四月。

[6] 謝清俊，漢字的字形與編碼，1996 年十月

[7] Chou, Y.M. and Huang, C.R.(2006), Hantology:Linguistic Resources for Chinese Language Processing and Studying, to appear in Proceedings of Language Resources and Evaluation, Genoa ,Italy, May, 24-26.

[8] 標準交換碼全字庫概況」調查，行政院主計處政府機關資訊通報 ，11 月，頁 13-17。

[9] 中醫藥委員會，http://www.ccmp.gov.tw/

[10] 全字庫，
http://www.cns11643.gov.tw/web/index.jsp