漢字構形資料庫在網路上的應用

[1]李宏益　　[2]林德潤　　[2]王祥安
[1]元智大學資訊管理學系
[2]中央研究院資訊科學研究所
s941712@mail.yzu.edu.tw, {soar, sawang}@iis.sinica.edu.tw

摘要

從古至今，漢字的形、音、義的不同，加上不斷的演變而產生了數量極多的字。電腦處理漢字的方式，是透過對應的交換碼來顯示正確漢字，由於編碼空間有限，無法表示所有的漢字，所以造成缺字問題，為了完整保存與呈現古籍中的所有漢字，所以中央研究院文獻處理實驗室發展漢字構形資料庫以延伸的編碼集表達現有系統無法處理的漢字。

本研究以此資料庫為基礎，發展了適合於網際網路環境上使用的缺字處理技術，讓缺字能夠在網頁中被記錄與呈現。我們也發展線上的漢字構形資料庫，讓使用者能更有效率的使用，以學習及了解古漢字之美。

關鍵字：缺字系統、構字式、漢字構形資料庫

1. 前言

早期電腦系統發源於英語系國家，因此最早的編碼系統僅包含數字、26 個英文字母的大寫與小寫、標點與其他的特殊符號，西方文字可以有限字母表示所有的字。然而亞洲地區大多為表意文字，尤其漢字的字集依古今的變異、學術與應用環境的差異等，而有字數、字形、字音以及字義上的變化。因為字數不斷增加，其數量已超過早期編碼系統所能記錄的數量，造成編碼空間對亞洲語系是不足夠的。電腦的編碼系統是採取一個字對一個交換碼，才能在電腦上顯示出來，所以電腦在處理漢字資料時，會因為字形沒有對應的交換碼而無法顯示出來，在儲存古籍的系統中情況特別嚴重。以目前常被使用的繁體中文編碼 BIG5 為例，僅收納了一萬三千多個常用字，仍有許多漢字無法被表示。

為了解決缺字問題，一般治標的方法是在交換碼的使用者造字區內，選一個碼位，並製造所缺的字形。這種做法雖然可以解決電腦上不能顯示的缺字，但可能會衍生如下的問題：

1. 大幅增加資料登錄的工作；當鍵入資料時，遇到缺字要先以一種特殊符號暫時替代，等造好字之後，還得必須重新校對原文，整理改所有缺字，過程非常繁雜。
2. 造字的不易管理；若缺字數量龐大，加上所造的字無法依交換碼的字序排列，會導致這些新造的字不易查核比對。
3. 造字區的空間不足；通常交換碼中允許造字的空間在數十字至數百字之間。以 Big5 而言，造字空間是最大的了，也只有 5809 字，超越此字數後，將造成碼位的重疊或相同。
4. 造成資訊共享的障礙；使用者造字所定的單行碼，並不能確保在別台電腦系統上的位碼也是相同的，破壞了交換碼的通用性，使電子文件無法與大家共享。

目前中央研究院文獻處理實驗室發展漢字構形資料庫收納、整理中國古今漢字，以漢語大字典為主，將各種字集的漢字資料收納，並建立各種字集的資料庫，內容是由人工逐字所建立，系統功能強大，但是使用者必須安裝在自己的電腦上。漢字構形資料庫使用者比較偏向專業的人士，如：歷史學家、漢學家、考古學家、博物館管理者、圖書館管理者、政府相關機構。然而一般使用者，如：學生、藝術家，對於漢字構形資料庫所提供的強大功能並不需要太多，而且操作方式複雜，會令他們失去對古漢字的興趣。

在資訊科技與網路的快速發展下，人們更容易透過網際網路取得資訊，若能將漢字構形資料庫應用在網際網路上，更能推廣中國的漢字文化，也讓使用者可隨時隨地透過網路來學習與了解漢字構形的發展。

我的漢字構形資料庫應用於網路上，主要是為了解決一般使用者不需要長時間使用漢字構形資料庫，並且希望能快速看到擁有相似部件字的古字、字形演變，以及異體字。希望讓一般使用者能對古漢字有更多的接觸，以求我們的中國文化，能被欣賞、傳播還有傳承。

2. 相關理論與技術

2.1 漢字構形資料庫

漢字構形資料庫是一個記錄漢字形體知識的資料庫，這些知識包括：

1. 古今漢字的字形演變
2. 古今漢字的字形結構
3. 不同漢字間的使用關係

目前電腦處理漢字的諸多問題，例如缺字、異體字等，主要原因在於電腦裡的漢字知識嚴重不足。有鑑於此，中央研究院資訊所文獻處理實驗室

自 1993 年起，即先由字形著手，建置漢字構形資料庫。

漢字構形資料庫早期收錄的字形是以楷書的現代印刷字體為主，其後陸續增加小篆、金文、甲骨文、楚系文字等古漢字。主要的特色有如下：

1. 銜接古今漢字以反映字形源流的演變。
2. 收錄不同歷史時期的異體字表，以表達不同漢字在各個時期的歷史層面的使用關係。
3. 記錄不同歷史時期的漢字結構，以呈現漢字因義構形的特點。
4. 使用構字式解決古今漢字的編碼問題。

漢字構形資料庫收錄不同歷史時期的漢字，除了要作字形、字義的銜接外，還要依據不同的形體來作構形分析。字形的銜接是依據字形的演變，在電腦中使用相同的編碼位置，編入不同的字型。字義的銜接是參考字義的隸屬，在電腦中使用不同的編碼位置，編入異體字表。不同歷史時期的漢字構形分析，雖然不見得合乎構形理據，但是在字形的檢索上也有一定的根據，這也符合漢字『以義構形』的依據。圖 1即漢字構形資料庫的組成，以及比較重要的資料表。

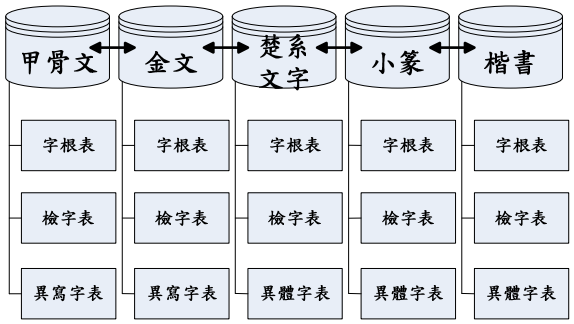

圖1 漢字構形資料庫的架構

2.2 構字式的處理技巧

漢字構形資料庫中有以特別規則表示每個漢字的方式，是「構字式」。其基本的概念是以數量較少的部件去形成數量極多的漢字。構字式是由部件、構字符號，依照構字規則所行組成一個字的代表式。構字式經過處理後，可從所屬字型檔取得正確的漢字。

構字符號目前有三大類，十三種分別為：⿰、⿱、⿸、⿻、⿲、⿳、⿴、⿵、⿶、⿷、⿺、⿹、⿼，說明如表 1，表 2為範例。

表1 構字符號表

類別	符號	說　明
連接符號	⿰	當部件的連接順序由左至右
	⿱	當部件的連接順序由上至下
	⿸	當部件的連接順序由外至內
部件序	⿺	按部件書寫順序輸入，前後以起始符號(⿺)和終止符號(⿻)包夾
	⿻	
方便符號	⿻	二個相同部件直連
	⿳	三個相同部件直連
	⿻	二個相同部件橫連
	⿲	三個相同部件橫連
	⿼	三個相同部件呈三角狀排列
	⿲	四個相同部件橫連
	⿳	四個相同部件直連
	⿼	四個相同部件呈四角狀排列

表2 構字式範例

使用符號	字例	構字式
⿰	加	力⿰口
⿱	凴	馮⿱几
⿸	閃	門⿸人
⿺、⿻	解	⿺角刀牛⿻
⿻	炎	⿻火
⿼	龘	⿼龍

3. 系統架構

本章節將介紹利用漢字構形資料庫在網路上的應用及完整系統的架構。

3.1 漢字構形資料庫應用於網路上系統架構

目前系統是使用 IIS 伺服器並配合 SQL Server 資料庫。採用 ASP.Net 語言來開發。

系統架構主要分成兩個部份，一個是利用漢字構形資料庫的資料讓使用者在網路上檢索的功能，一個是在網頁上呈現缺字的方式。系統架構圖如圖 2。

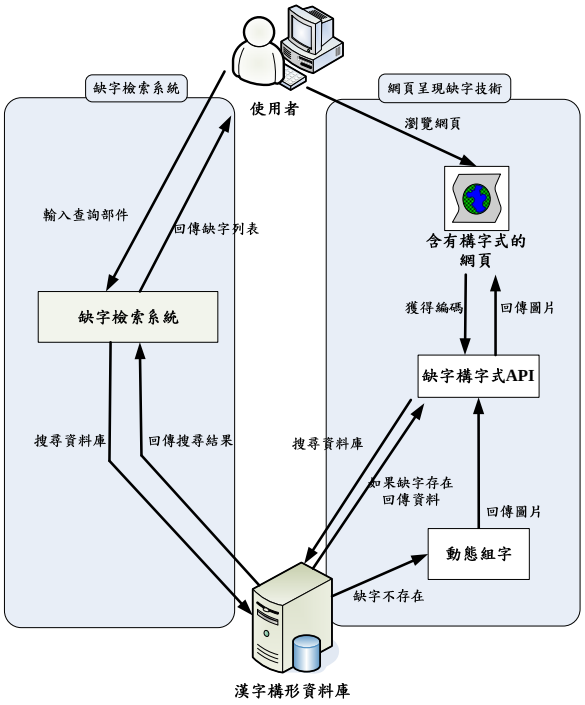

圖2 系統流程圖

3.1.1 缺字檢索系統

缺字檢索系統主要包含部件檢字、製作字形圖片 字形演變以及異體字表。

3.1.2 部件檢字

部件檢字中，使用者輸入欲查詢文字的部件進行檢索，檢索後會傳回字形、構字式、構字組合、注音等資訊。如圖 3。

輸入的部件會被拆解成字根，進入漢字構形資料庫的楷書資料庫作查詢。主要用到楷書資料庫的字根資料表及檢字表。透過檢字表可查出部件擁有的字根，接著利用字根資料表將字根依編號作排序，再傳回檢字表搜尋含有相同字根的資料。

圖3 部件檢字

3.1.3 製作字形圖片

當部件檢字後，使用者可以透過製作字形圖片，來設定字形大小、字形顏色及字體，來取得一張透明底圖的字形圖片，如圖 4。

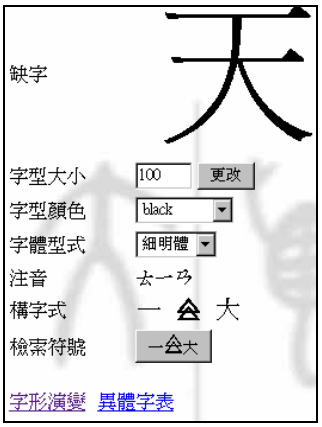

圖4 字形圖片

3.1.4 字形演變

在製作字形圖片時，使用者可以點選字形演變來檢視該文字在不同時期的文字演變。資料庫中收錄了五個時期的字集，包含了甲骨文、金文、楚系文字、小篆及楷書，如圖 5。讓使用者可以了解從古至今的文字演變，以及文字出現的時期與收藏於目前字典的何處。使用者不但可以藉此學到古字，亦可欣賞古漢字美之所在。

圖5 字形演變

3.1.5 異體字表

在漢字構形資料庫的異體字表其實是包含異寫字以及異體字，異寫字是同一個字因不同寫法，而造成形體差異，如圖 6「中」在金文字集中的異寫字；異體字是音與義都相同，如圖 7為「拾」的異體字。

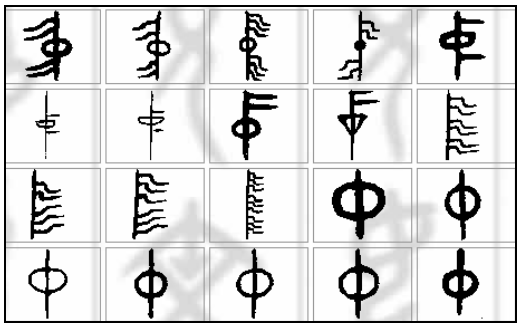

圖6 異寫字表

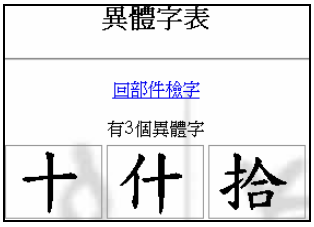

圖7 異體字表

3.2 網頁呈現缺字技術

發展這項技術的目的是為了處理在網頁上呈現缺字的問題。我們採用的策略則是將構字式轉換成缺字圖形。其優點是能夠在任何瀏覽器中使用，而且使用者不需要額外安裝字形檔。

3.2.1 構字式處理 API

構字式處理 API 會先判斷頁面上的構字式，並將構字式送到漢字構型資料庫，若能搜尋到符合的構字式，將會依據使用者設定的字形大小與字形顏色，把該構字式轉換成一張透明底圖的缺字字形圖片回傳至頁面的適當位置。

3.2.2 動態組字程式

漢字構型資料庫目前仍由中央研究院資訊所文獻處理實驗室持續維護與新增，以因應持續進行的漢字歷史文獻之數位化工作。因此，有些漢字所對應的構字式尚未輸入漢字構型資料庫中。在處理這些漢字的時候，我們會利用動態組字的函式即時產生字形圖片。

動態組字是一套按照構字式繪製漢字字形圖片的自由軟體。例如"火⿰中⿱天"是資料庫中沒有儲存的構字式，透過動態組字的函式，便可以即時產出"⿰⿱"的字形圖片。

4. 結論與未來規劃

本文在說明漢字構形資料庫在網路上的應用，目前的應用有缺字檢索系統，包含部件檢字、製作字形圖片、字形演變和異體字表；以及網頁呈現缺字技術，包含構字式 API 和動態組字程式。往後將會繼續維護和發展新的技術，我們希望缺字系統能提供更完善的功能給使用者，讓使用者可以進一步了解漢字在各個時期的寫法及演變，並且欣賞各種字體之美。

目前會使用這項技術與系統的使用者侷限在一些專業領域人士，為了讓更多的使用者能接觸到這些資訊，未來會在原本的架構上，重新規劃系統流程，讓操作上能更為靈活且與使用者更多互動；增加其他檢索方式，如字音、部首、筆劃等方式，並連結現有的線上辭典和字庫，成為一個古今漢字檢索的整合網站。

誌謝：

此研究計畫由台灣行政院國家科學委員會的數位典藏與數位學習國家型科技計畫(TELDAP)資助，計畫編號：NSC 96-3113-H-001-010、NSC 96-3113-H-001-011、及 NSC 96-3113-H-001-012。

參考文獻

[1] T. J. Lin, J. W. Huang, Christine Lin, H. Y. Li, H. A. Wang, C. Y. Chiu, “A Mechanism for Solving the Unencoded Chinese Character Problem on theWeb”, ECDL, Aarhus, Denmark, 2008.

[2] D. M. Juang, J. H. Wang, C. Y. Lai, C. C. Hsieh, L. F. Chien, and J. M. Ho, "Resolving the Unencoded Character Problem for Chinese Digital Libraries", JCDL, Denver, Colorado, USA, p311-p319, June 7-11, 2005.

[3] C. C. Hsieh, L. L. Wu, Y. M. Chou, “A Missing Characters Description Language for Han Characters”, Int. Computer Symposium, Taipei, Taiwan, Dec, 2004.

[4] 黃俊瑋, 林金龍, 黃國倫, “缺字系統整合動態組字之應用”, TANET 2007 台灣網際網路研討會, 台北, 2007.

[5] 謝清俊, “電子古籍中的缺字問題”, 第一屆中國文字學會學術討論會, 1996.

[6] 莊德明, 謝清俊, “漢字構形資料庫的建置與應用”, 漢字與全球化國際學術研討會, 2005.

[7] 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”, 第二次兩岸古籍整理研究學術研討會, 北京大學, 北京, 1998.

[8] 莊德明, “中文電腦缺字解決方案”, 全國技專院校圖書館自動化規劃第七屆研討會, 屏東, 2001.

[9] 全字庫, http://www.cns11643.gov.tw/

[10] 漢字構形資料庫, http://www.sinica.edu.tw/~cdp/cdphanzi/

[11] 中央研究院資訊科學研究所文獻處理實驗室, http://www.sinica.edu.tw/~cdp/index.html

[12] 數位典藏國家型計畫/數位典藏技術發展組 http://daal.iis.sinica.edu.tw/Chinese/System/inde

x.htm
[13] 缺字系統, http://char.ndap.org.tw/