# 行政院國家科學委員會專題研究計畫 成果報告

## 萬國碼漢字檢字系統改善計畫<br>研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 99-2221-E-029-035-
執行期間：99年08月01日至100年12月31日
執行單位：東海大學資訊管理學系

計畫主持人：林正偉

計畫參與人員：碩士班研究生-兼任助理人員：黃筱婷
大專生-兼任助理人員：黃箴嫥
大專生-兼任助理人員：高璟屏
大專生-兼任助理人員：吳雅靜
大專生-兼任助理人員：粘尹芝
博士班研究生-兼任助理人員：王敦威

報告附件：出席國際會議研究心得報告及發表論文

公開資訊：本計畫可公開查詢

中華民國 101年03月29日

中文摘要：萬國碼(Unicode) 6.0 版已經收錄超過 75,000 個漢字，預計在不久的未來將成長到 10 萬個漢字。然而，除了一開始收錄在中日韓認同表意文字區(CJK Unified Ideograph)的 20,902 個漢字外，後來新收錄的 5 萬多個漢字只能在少數的系統上使用。隨著科技進步，有愈來愈多的系統可以支援 32 位元的 Unicode 系統，也有少數的字型廠商也開始推出支援 7 萬多個漢字的字型檔。可以預見的是，在全球化的數位世界裡，將會有更多的資訊處理使用 Unicode 作為基本字集。

然而，新收錄的那 5 萬多個漢字多半是一些生僻的漢字，大部份的中文輸入法無法輸入這些字，許多情形只能使用內碼輸入。當使用者要使用某一個特定字形的漢字而無法用一般的中文輸入法輸入時，他需要使用一個漢字檢字系統來查詢該字是否收錄在 Unicode 中，若有，內碼是多少，若無，則該字是一個缺字，使用者必需使用缺字系統來處理。

在 98 年計畫『萬國碼漢字檢字系統』裡，我們設計了一個演算法來計算兩個漢字之間的距離，作為相似字形評估比較的基礎。透過字形相似度的建立，不需要精確地描述一個漢字的外形、讀音、意義，使用者透過聯想的方式，利用一般的輸入法輸入一個外形相近的字，就可以進行檢字作業。例如，使用者可以用『壺』來檢索『壼』或『壷』。

本計畫進一步提供使用者透過組成部件進行檢字作業。傳統部件檢字需要輸入精確的部件，然而，有些漢字的組成部件並不容易輸入，如上述的『壷』，兩個組成部件中只有上位部件『士』是容易輸入的，下位部件則無法輸入。本計畫之成果可以提供模糊部件的使用，例如，使用者可以簡單地輸入『士亞』兩個部件，即可檢索到『壷』這個字。本計畫之成果，已經整合到國際電腦漢字與異體字知識庫(http://chardb.iis.sinica.edu.tw)之部件檢索系統，提供各界使用。

中文關鍵詞：萬國碼、漢字、部件檢字、模糊部件

英文摘要：Unicode 6.0 has encoded more than 75,000 Han characters in its repertory, and in the near future, it is expected that Unicode will encode more than 100,000 Han characters. However, most systems equipped with their default font display and input method subsystems support only 20,902 Han characters encoded in Unicode CJK Unified Ideograph in the first release of Unicode 1.1. Nowadays, there are some systems supporting the use of 32-bit Unicode characters, and thus Han characters encoded in Unicode CJK Unified Ideograph Extensions A to D. When suitable font files are installed, users can use all characters encoded in the Unicode repertory. However, most widely-used Han character input methods are incapable of looking up these newly-encoded Han characters. Due to the Unicode unification mechanism, a Han character that has several similar shapes might have been assigned to several unique codepoints, each of which identifies a shape； or, a Unicode Han character might also represent many characters, which have similar shapes and are used in different regions. A Unicode Han character shown on a user’s screen or printed on a paper is probably different when different font files are used. As a result, it is hard for a generic user to look up a Han character encoded in Unicode.

In the previous NCS project in 2009, we had developed a Unicode Han character lookup system. An algorithm is devised to calculate the similarity of the shapes of two Han characters. Thus, users can input a similar-in-shape Han character to look up a Han character whose shape he or she does not exactly know. For example, users can use ’壺’ to look up ’壸’ or ’壷’.

Sometimes, it is convenient to look up Han characters via components. Traditionally, Han character lookup via components requires good knowledge of exact components. However, it is hard to input certain components. For example, users cannot input the bottom component of ’壷’ to look up this character. In this project, we had enhanced this system by supporting lookup via similar components. As a result, users can easily input ’士亞’ to look up ’壷’. We had integrated this subsystem into International Encoded Han Character and Variants Database (at http://chardb.iis.sinica.edu.tw).

英文關鍵詞： Unicode； Han Character； Character Lookup； Component；

# 行政院國家科學委員會補助專題研究計畫 □期中進度報告 ■期末報告

## 萬國碼漢字檢字系統改善計畫

計畫類別：■個別型計畫　□整合型計畫
計畫編號：NSC 99－2221－E－029－035－
執行期間：99 年 8 月 1 日至 100 年 12 月 31 日

執行機構及系所：東海大學資訊管理學系

計畫主持人：林正偉
共同主持人：
計畫參與人員：黃筱婷、王敦威、黃箴嫥、高璟屏、吳雅靜、粘尹芝

本計畫除繳交成果報告外，另須繳交以下出國報告：
□赴國外移地研究心得報告
□赴大陸地區移地研究心得報告
■出席國際學術會議心得報告及發表之論文
□國際合作研究計畫國外研究報告

處理方式：除列管計畫及下列情形者外，得立即公開查詢
　　　　　□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

中　華　民　國　101 年 3 月 20 日

# 行政院國家科學委員會補助專題研究計畫　期末報告

# 萬國碼漢字檢字系統改善計畫

## 中文摘要

萬國碼(Unicode) 6.0 版已經收錄超過 75,000 個漢字，預計在不久的未來將成長到 10 萬個漢字。然而，除了一開始收錄在中日韓認同表意文字區(CJK Unified Ideograph)的 20,902 個漢字外，後來新收錄的 5 萬多個漢字只能在少數的系統上使用。隨著科技進步，有愈來愈多的系統可以支援 32 位元的 Unicode 系統，也有少數的字型廠商也開始推出支援 7 萬多個漢字的字型檔。可以預見的是，在全球化的數位世界裡，將會有更多的資訊處理使用 Unicode 作為基本字集。然而，新收錄的那 5 萬多個漢字多半是一些生僻的漢字，大部份的中文輸入法無法輸入這些字，許多情形只能使用內碼輸入。當使用者要使用某一個特定字形的漢字而無法用一般的中文輸入法輸入時，他需要使用一個漢字檢字系統來查詢該字是否收錄在 Unicode 中，若有，內碼是多少，若無，則該字是一個缺字，使用者必需使用缺字系統來處理。

在 98 年計畫『萬國碼漢字檢字系統』裡，我們設計了一個演算法來計算兩個漢字之間的距離，作為相似字形評估比較的基礎。透過字形相似度的建立，不需要精確地描述一個漢字的外形、讀音、意義，使用者透過聯想的方式，利用一般的輸入法輸入一個外形相近的字，就可以進行檢字作業。例如，使用者可以用『壸』來檢索『壺』或『壷』。

本計畫進一步提供使用者透過組成部件進行檢字作業。傳統部件檢字需要輸入精確的部件，然而，有些漢字的組成部件並不容易輸入，如上述的『壷』，兩個組成部件中只有上位部件『士』是容易輸入的，下位部件則無法輸入。本計畫之成果可以提供模糊部件的使用，例如，使用者可以簡單地輸入『士亞』兩個部件，即可檢索到『壷』這個字。本計畫之成果，已經整合到國際電腦漢字與異體字知識庫(http://chardb.iis.sinica.edu.tw)之部件檢索系統，提供各界使用。

關鍵詞：萬國碼、漢字、部件檢字、模糊部件

## Abstract

Unicode 6.0 has encoded more than 75,000 Han characters in its repertory, and in the near future, it is expected that Unicode will encode more than 100,000 Han characters. However, most systems equipped with their default font display and input method subsystems support only 20,902 Han characters encoded in Unicode CJK Unified Ideograph in the first release of Unicode 1.1. Nowadays, there are some systems supporting the use of 32-bit Unicode characters, and thus Han characters encoded in Unicode CJK Unified Ideograph Extensions A to D. When suitable font files are installed, users can use all characters encoded in the Unicode repertory. However, most widely-used Han character input methods are incapable of looking up these newly-encoded Han characters. Due to the Unicode unification mechanism, a Han character that has several similar shapes might have been assigned to several unique codepoints, each of which identifies a shape; or, a Unicode Han character might also represent many characters, which have similar shapes and are used in different regions. A Unicode Han character shown on a user's screen or printed on a paper is probably different when different font files are used. As a result, it is hard for a generic user to look up a Han character encoded in Unicode.

In the previous NCS project in 2009, we had developed a Unicode Han character lookup system. An algorithm is devised to calculate the similarity of the shapes of two Han characters. Thus, users can input a similar-in-shape Han character to look up a Han character whose shape he or she does not exactly know. For example, users can use "壺" to look up "壸" or "壷".

Sometimes, it is convenient to look up Han characters via components. Traditionally, Han character lookup via components requires good knowledge of exact components. However, it is hard to input certain components. For example, users cannot input the bottom component of "壷" to look up this character. In this project, we had enhanced this system by supporting lookup via similar components. As a result, users can easily input "士亞" to look up "壷". We had integrated this subsystem into International Encoded Han Character and Variants Database (at http://chardb.iis.sinica.edu.tw).

Keywords: Unicode; Han Character; Character Lookup; Component;

## 一. 研究背景

就中文資訊處理而言，缺字問題[24]-[27]一直是一個很嚴重的問題。由於早期的電腦漢字資訊交換碼，不論是繁體中文(or 正體字)的業界標準 BIG5 [17]、簡體中文(or 簡化字)的大陸官方標準 GB 2312-80 [21]、日本漢字的 JIS X 0208:1997 標準[19]或是韓國漢字的 KS X 1001:1992 標準[20]均採用雙位元組的編碼方式，碼位有限，無法完全收錄所有的漢字，造成某些使用頻率很低的罕用字和異體字沒有被收錄到字集裡面。

自 1993 年 Unicode 1.0 發表以來，國際標準組織 ISO/IEC JTCI/SC2/WG2 下的 IRG (Ideographic Rapporteur Group)工作小組仍持續不斷的搜集、檢視、收錄不同時期、區域的被使用到的漢字到 ISO-10646 (Universal Multiple-Octet Coded Character Set，簡稱 UCS) [11]系列國際標準中，並由 Unicode Consortium 發佈到 Unicode [12]的新版本中。除了一開始收錄在中日韓認同表意文字區(CJK Unified Ideograph)的 20,902 個漢字外，到了 Unicode 6.0 版，在中日韓認同表意文字區新增擴展 A 至 D 區(CJK Unified Ideograph Extension A to D)，總收錄超過 75,000 個漢字，包含康熙字典[1]、漢語大字典[2]等字書所錄漢字及中、日、韓、越、港、澳、星、馬等地所使用的漢字[14]-[23]，預計在未來將達到總數 10 萬個漢字以上。其中，中日韓認同表意文字區與擴展 A 區的碼點在 Unicode BMP (Basic Multilingual Plane)以內，可以使用 16 位元的 UCS-2 編碼。而中日韓認同表意文字擴展 B 區至 D 區及未來新增擴展區的碼點在 Unicode BMP 以外，必需使用 32 位元的 UCS-4 編碼。

由於使用 Unicode 具有兼容多國文字的好處，可以預見的是，在全球化的數位世界裡，將會有更多的資訊處理使用 Unicode 作為基本字集。由於漢字集是一個開放的字集，不斷有新的漢字因為不同的理由在不同的地區被發明、使用，缺字問題將永遠存在。即使如此，對漢字資訊處理而言，使用收錄了 7 萬餘個漢字的 Unicode 仍可大幅減少缺字的情形。

然而，目前大多數的系統仍是 16 位元的 Unicode 系統，只能處理碼點在 Unicode BMP 之內的字元，如收錄在中日韓認同表意文字區及擴展 A 區的漢字，而無法處理 Unicode BMP 以外的 32 位元 Unicode 字元，如收錄在中日韓認同表意文字擴展 B 至 D 區的漢字。事實上，目前絕大多數的系統，如 Microsoft Windows 系列的大多數版本、Linux、FreeBSD 等作業系統的預設版本等，並沒

有安裝中日韓認同表意文字擴展 A 區中漢字的字形檔，因此無法正常的顯示和列印這一區中的漢字。而大部份的中文輸入法也不支援這些漢字的輸入。整體來說，目前大部份的系統仍是只支援最初的 20,902 個漢字。而收錄在中日韓認同表意文字擴展 A 區至 D 區的 5 萬多個漢字只能在少數的系統上顯示，使用內碼輸入。

這種情形隨著科技的進步已經慢慢改變，目前 Microsoft Vista 與 Server 2008 作業系統已經可以支援 32 位元的 Unicode 字元，也有少數字形廠商開始提供字形檔。然而，雖有部份的輸入法宣稱可以輸入全部的 7 萬多個漢字，實際上，這些輸入法可以很容易輸入的仍是最初的 20,902 個漢字。對於大多數的使用者而言，Unicode 6.0 版新收錄的這 5 萬多個漢字多半是一些生僻的罕用字或是另一個漢字的異體字。相反的，真正需要使用到這些罕用字或異體字的使用者，如處理古籍、佛經等特定領域的使用者，仍缺乏一個真正可以讓他們輸入這些罕用字的輸入法。當他們使用某一個特定字形的漢字而無法用一般的輸入法輸入時，他需要使用一個漢字檢字系統來查詢該字是否收錄在 Unicode 中，若有，內碼是多少，若無，則該字是一個缺字，使用者必需使用缺字系統來處理。

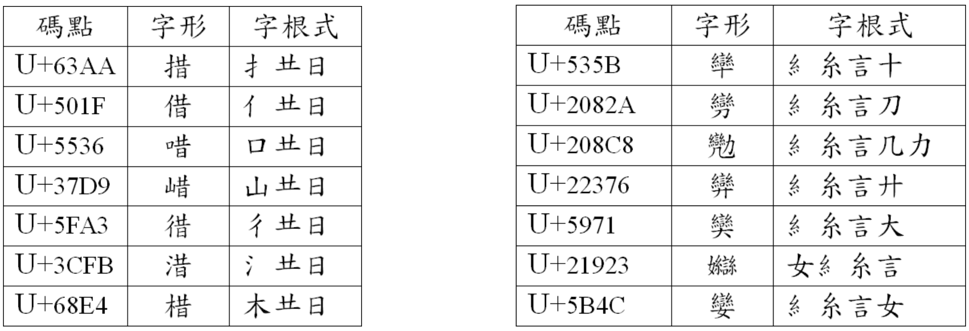
在 98 年度的計畫『萬國碼漢字檢字系統』中，我們設計了一個漢字檢字系統，使用者只需利用一般的輸入法輸入一個外形相近的字即可進行檢字。我們設計了一個演算法來計算兩個漢字之間的距離，作為相似字形評估比較的基礎。透過字形相似度的建立，不需要精確地描述一個漢字的外形、讀音、意義，使用者透過聯想的方式，利用一般的輸入法輸入一個外形相近的字，就可以進行檢字作業。例如，使用者可以用『壺』來檢索『壼』或『壸』。

本計畫進一步提供使用者透過組成部件進行檢字作業。傳統部件檢字需要輸入精確的部件，然而，有些漢字的組成部件並不容易輸入，如上述的『壸』，兩個組成部件中只有上位部件『士』是容易輸入的，下位部件則無法輸入。本計畫之成果可以提供模糊部件的使用，例如，使用者可以簡單地輸入『士亞』兩個部件，即可檢索到『壸』這個字。本計畫之成果，已經整合到國際電腦漢字與異體字知識庫[32]之部件檢索系統，提供各界使用。

## 二. 相關研究

傳統上，漢字檢字系統是以部首、筆畫、注音、部件等相關屬性及其組合來進行檢字，電腦漢字輸入法的設計也是基於類似的原理。教育部異體字典[5]提供部首、筆畫兩種檢索方式，但沒有提供 Unicode 的碼點對照。漢字構形資料庫[24]、數位典藏 web 缺字系統[30]另外提供了構字式[25]與部件檢索，但目前尚未支援 Unicode。行政院主計處電子處理資料中心建置、維護的全字庫[16]支援部首筆畫、注音、倉頡碼、部件等多種查詢方式，但只有提供收錄在國家標準 CNS-11643 [15]的漢字的資料。

一個漢字的外形在不同的地區可能是不同的，如臺灣、大陸與日本三地分別使用『吳』、『吴』與『呉』三種字形。由於 Unicode 採取字形認同原則，不同外形的漢字有時被分到不同碼點，有時則共享同一個碼點，如『吳』、『吴』與『呉』三字有三個不同的碼點 U+5433、U+5434 與 U+5449，而『祦』、『祦』共用 U+7966。在電腦系統上，有時甚至字形檔不同，字形就不同，這使得許多 Unicode 漢字的屬性如筆畫、部件與使用者的認知有差異。更糟的是許多 Unicode 漢字只有字形描述，而讀音、意義等屬性並不完全。因此，要使用上述的系統，使用者對於漢字編碼與漢字結構需有一定程度的了解，並不適合一般使用者。對於 Unicode 漢字編碼原則與其造成 Unicode 漢字檢索困難的詳細說明，可以參照本團隊 98 年度執行之國科會計畫『萬國碼漢字檢字系統』的結案報告與發表論文[34]。



| 碼點 | 字形 | 字根式 |
|---|---|---|
| U+63AA | 措 | 扌 龷日 |
| U+501F | 借 | 亻 龷日 |
| U+5536 | 唶 | 口龷日 |
| U+37D9 | 㟙 | 山龷日 |
| U+5FA3 | 徣 | 彳龷日 |
| U+3CFB | 㳻 | 氵 龷日 |
| U+68E4 | 棤 | 木龷日 |

| 碼點 | 字形 | 字根式 |
|---|---|---|
| U+535B | 卛 | 糹 糸言十 |
| U+2082A | 𠠪 | 糹 糸言刀 |
| U+208C8 | 𠣈 | 糹 糸言几力 |
| U+22376 | 𢍶 | 糹 糸言廾 |
| U+5971 | 奱 | 糹 糸言大 |
| U+21923 | 𡤣 | 女糹 糸言 |
| U+5B4C | 孌 | 糹 糸言女 |

圖 1: Unicode 漢字的字根構字式

| 篡 ＝ | ⺮ | 目 | 大 | 厶 |
|---|---|---|---|---|
| ↓轉換 | ↓不變 | ↓不變 | ↓不變 | ↓取代 |
| 纂 ＝ | ⺮ | 目 | 大 | 糸 |

| 篡 ＝ | ⺮ | 目 | 大 | 厶 | |
|---|---|---|---|---|---|
| ↓轉換 | ↓刪除 | ↓不變 | ↓不變 | ↓不變 | ↓增加 |
| 瞌 ＝ | | 目 | 大 | 厶 | 皿 |

圖 2: 『篡』轉換成『纂』或『瞌』

從漢字的構成來說，除了最基本不可被分解的字根之外，字根和字根可以組成部件，而大部份的漢字是由一些字根與部件依某些規則組合而成，如說文解字提到的會意、形聲兩種方式。我們在識別某一個漢字是什麼字時，也常常會利用部件的觀念。漢字構形資料庫[24]利用這個觀念，設計出構字式[25]的概念來描述一個漢字的組成方式。當一個漢字的組成部件拆解到最基本的字根時，可以得到它的字根式，如圖 1 所示。如此，一個漢字可以用它的組成部件來進行檢索。如利用『金』與『芬』來檢索『𨧼(U+289FC)』。

使用部件檢索的最主要困難仍在於許多部件、字根在一般的系統上無法輸入，如『丂』、『㠯』、『豖』、『疒』、『𧘇』等等，更不用說『乌』、『𤰔』、『𦥑土』等罕用字根。如『壺』字的下位部件『亞』，並沒有被編碼，無法輸入與顯示，若只是輸入上位部件『士』，則有上千個候選字。『𠗦(U+205E6)』、『𥁝(U+2505D)』、『𨳇(U+28CC7)』等漢字都有主要部件無法輸入的問題，因此難以檢索。

在 98 年度計畫『萬國碼漢字檢字系統』中，為了簡化檢字作業，我們設計了一個相似字形的檢索系統，透過計算兩個漢字構字式之間的編輯距離來評估兩個漢字的相似度。使用者只需利用一般的輸入法輸入一個外形相近的字即可進行檢字。

兩個漢字的相似度以字根式之間轉換的編輯距離來估計，如圖 2 所示。編輯距離[35][36]透過增加、刪除和取代三種字元操作，將一個字串轉換為另一個字串，是一種常被用來估計兩個字串之間相似度的方法。

考慮到字根操作的特性，我們將『增加字根』、『刪除字根』的操作成本設定為字根的筆畫數，反應出字根本身的複雜度。對於『取代字根』的操作，若兩個字根不相似，如用『乂』取代『口』，其操成本設為兩個字根的筆畫之和，相當於先刪除後再新增。若是相似字根之間的取代操作，如用『東』取代『柬』，其操作成本設定為小於兩者筆畫和的一個數字，這將使得編輯距離採用取代操作，以得到較小的成本。如此，使用者可以利用相似字『馮』、『盂』、『門』，很容易地在國際電腦漢字與異體字知識庫[32]檢索到『𠗦』、『𥁝』、『𨳇』這兩個漢字。詳細說明可以參考本團隊 98 年度所進行的國科會計畫『萬國碼漢字檢字系統』之結案報告與發表論文[34]。

然而有些漢字難以使用相似字形進行檢索，如『𣡆(U+23846)』。這些時候，可以使用『壺』與

| | | | | |
|---|---|---|---|---|
| (a) | 臣官 = | 臣 | 宀 | 吕 |
| | ↓轉換 | ↓取代 | ↓刪除 | ↓取代 |
| | 𦣤 = | 𦣝 | | 吕 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (b) | 十中亞 = | | 十 | 中 | | 一 | 卝 | 一 |
| | | ↓增加 | ↓刪除 | ↓刪除 | ↓增加 | ↓刪除 | ↓刪除 | ↓刪除 |
| | 𡈋 = | 巿 | | | 𤰃 | | | |

圖 3: 直接進行相似部件檢索的得到好的結果(a)與不好的結果(b)

『桑』兩個部件來進行檢索會很方便。如上所述，使用部件檢索的最主要困難仍在於許多部件、字根無法輸入。如『𦣤(U+268E4)』、『𡈋(U+2120B)』等字，所使用之部件『𦣝』、『吕』、『巿』、『𤰃』在一般的系統上均無法輸入。其中，『𦣝』可以使用相似部件『臣』,『吕』的相似部件是『吕』，無法輸入，所幸，『吕』是『官』字的下位部件，其上位部件『宀』只有三畫，因此可用『臣』與『官』來檢索『𦣤』，如圖 3(a)所示。然而，但『巿』與『𤰃』兩者均無可以直接輸入的相似部件可供檢索。一般人可能會將『巿』拆解成『十』與『中』，而『𤰃』則可能用『田』或『亞』。無論那種方式，透過編輯距離的計算，均不能得到滿意的結果，如圖 3(b)所示。

## 三.相似部件檢字系統之設計

在執行 98 年國科會計畫『萬國碼漢字檢字系統』時，我們注意到某些部件之間的相似度計算難以用編輯距離評估。比如『良』與『皀』在『鄉』、『塈』這類漢字裡是相似部件，然而『良』的兩個部件是『丶』與『艮』，而『皀』的兩個部件為『白』與『七』，從編輯距離的觀點來看，並不相近。

構字式所使用之字根約 1,100 個，若不考慮文字學的學理考究，其中有多個其實可以再拆解成更小的組成部件，如『亞』、『異』、『東』等。為了解決上述問題，我們重新審視構字式的字根，依照使用者可能會使用的方式，將可再分解的字根，進行『無理據』的分解。這裡，『無理據』指的是不依照文字學的理論，單從一般使用的角度來拆解，如表 1 所示。

表 1: 『無理據』的字根分解(部份列表)

| | |
|---|---|
| 亞 | 口卝一 |
| 異 | 甲口口一八 |
| 東 | 十中中小 |
| 良 | 丶日𠤎 |

在相似字形檢索裡，使用者輸入的相似漢字必需是已被 BIG5 或 Unicode 等編碼標準所收錄的漢字。這些漢字均是已知的，因此可以事先將使用者可以輸入的漢字與 Unicode 收錄的 7 萬 5 千餘個漢字進行編輯距離的計算，將結果儲存在資料庫裡。使用者進行相似字檢索時，系統只需從資料內讀取出事先計算好的結果即可。

在部件檢索裡，由於使用者輸入的部件組合可以有無限多重，無法事先計算。因此，每一次的檢索，都必需進行線上的編輯距離計算。若要將整個 Unicode 收錄的 7 萬 5 千餘個漢字一一進行比較，相當費時，因此，有必要進行事先的過濾。

假設使用者輸入一個由 n 個部件形成的組合『$R_1^1R_2^1\cdots R_n^1$』，對於每一個部件 $R_i^1$，我們可以透過已有的相似字根資料庫，找到它所有的 $X_i$ 個相近部件 $R_i^2$, $R_i^3$,…, $R_i^{Xi}$，那麼總共會有共有 $X_1*X_2*X_3*\cdots*X_n$ 種可能的相似部件組合，我們可以用一個特殊的正規表示式(regular expression)『$[R_1^1R_1^2R_1^3\cdots R_1^{X1}][R_2^1R_2^2\ R_2^3\cdots R_2^{X2}]\cdots[R_n^1R_n^2R_n^3\cdots R_n^{Xn}]$』來表示。若要進行 Query Expansion，一一檢索每種可能的相似部件組合，當相似部件的數量多時，計算量會成指數成長。因此，我們需要一個有效的演算法來找出可能的候選漢字。

對於這個問題，我們團隊過去在處理中文異體字域名時，比如說想要用『臺灣』這個詞找出由『[臺台][灣湾]』這個 regular expression 所產生的任何一個異體字詞，利用異體字資料，我們可以系統性地建造一個特殊的函數來產生異體字詞索引，可以有效地降低計算量[33]。在這裡，我們建造另一個專門用來處理相似字根的索引函數。利用這個索引函式，可以在構字式資料庫裡很快地找到那些漢字的構字式具有類似於 $R_1^1R_2^1\cdots R_n^1$ 的部件組合。

我們首先將所有最小字根跟據相似字根進行分組，如表 2 所示。

表 2: 最小字根分群(部份列表)

| 群組字根 | 代表字根 |
|---|---|
| 乙乙⺄… | 乙 |
| ㄥ厶夕纟 幺乡… | 厶 |
| 𠕁冊册田𠕋𦉬𦉫… | 冊 |

如此，每一個 Unicode 漢字可以建立最小字根式的索引，如表 3 所示，儲存在資料庫中。

表 3: 最小字根式索引之建立(部份列表)

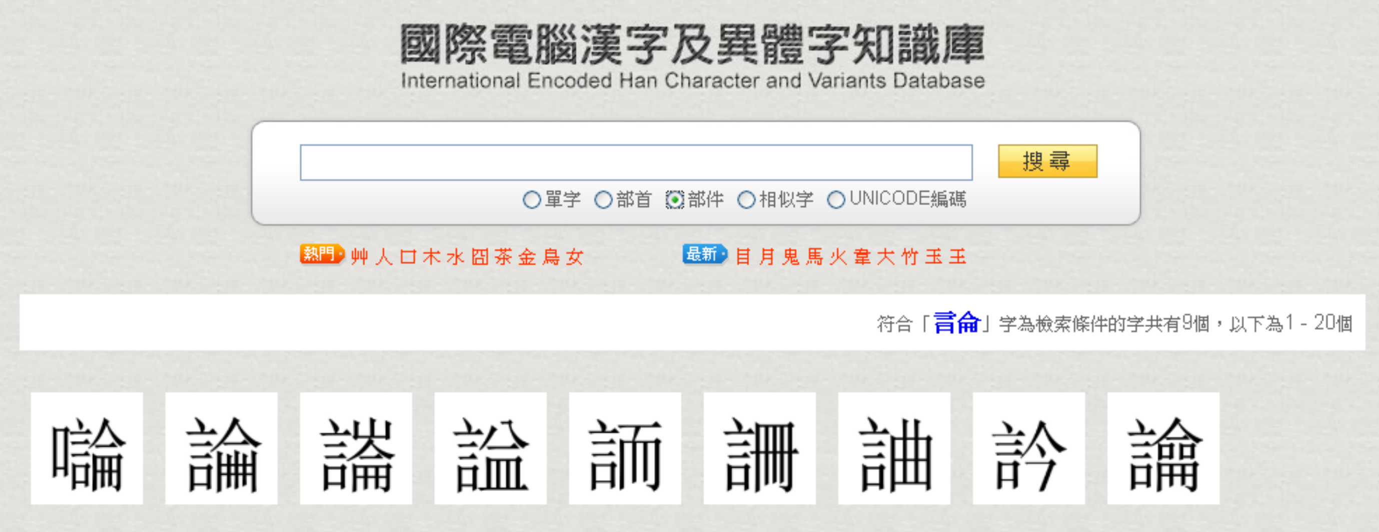
| Unicode 漢字 | 最小字根式 | 最小字根式索引 |
|---|---|---|
| 𧫲(U+27AF2) | 言戶𦉫 | 言戶冊 |
| 𧪺(U+27ABA) | 言人一𠕁 | 言人一冊 |
| 𧭆(U+27B46) | 言人一口口冊 | 言人一口口冊 |

當使用者輸入『言』與『册』兩個部件進行所引時，根據最小字根分群的結果，使用代表字根『言』與『冊』，由資料庫中選出最小字根式索引含有這兩個最小字根的所有漢字，接下來再一一計算『言册』與各個漢字之間的編輯距離，作為排序的基礎。

本研究的成果已應用在國際電腦漢字與異體字知識庫[32]之相似字檢索系統，具體成果如圖 4 所示。

## 四.結論與未來展望

在這個研究裡，我們擴展了原有 Unicode 漢字檢字系統的功能，對於部首、筆畫、注音等屬性不清楚，甚至連字形都不甚確定的漢字，使用者可以輸入數個相似部件，進行漢字檢索。我們利用構字式的概念，將 Unicode 所有漢字分解成最小字根所組合而成的字根式，再利用字根群組代表字建立索引。當使用者輸入相似部件進行檢索時，相似部件同樣進行分解成最小字根，再利用字根群組代表字從資料中挑出可能的候選字，一一進行編輯距離的計算，作為排序的結果。實驗結果說明我們的系統可以提供使用者相似部件的檢索。



國際電腦漢字及異體字知識庫

International Encoded Han Character and Variants Database

搜尋

○單字 ○部首 ●部件 ○相似字 ○UNICODE編碼

熱門 艸 人 口 木 水 囧 茶 金 烏 女

最新 目 月 鬼 馬 火 韋 大 竹 玉 王

符合「言侖」字為檢索條件的字共有9個，以下為1 - 20個

圖 4: 在國際電腦漢字與異體字知識庫利用『言侖』進行相似部件檢索

除了繼續調整個別最小字根的拆解和字根相似度，未來我們將朝向更多樣化的字根操作進行研究，如字根前後交換的操作，在漢字構成之中是常有的，『𪀚』和『䳘』的構字式都是『鳥我』，兩者都是『鵝』的異體字，而『鵝』的構字式是『我鳥』。

本研究的成果已應用在國際電腦漢字與異體字知識庫[32]之相似字檢索系統，系統與演算法相關細節已整理，投稿到國際會議與期刊上，準備發表。

### 致謝

本研究感謝中央研究院資訊科學研究所范紀文、林峰生兩位先生協助進行將計畫成果整合到國際電腦漢字與異體字知識庫。

## Reference:

[1] Zhang Yushu, Chen Tingjing et al. 1716. The KangXi Dictionary. Zhonghua Bookstore, ISBN 962-231-006-0, 1989. [康熙字典, 張玉書等編著, 1716, 中華書局, 1989]

[2] Hanyu Da Zidian, compiled by Hanyu Da Zidian editorial committee. 1986. Sichuan Cishuan Publishing (Chengdu), ISBN 780-543-001-2, etc. [漢語大字典, 1986, 四川辭書出版社, 成都]

[3] Dai Kan-Wa Jiten, revised edition, compiled by Morohashi Tetsuji, 1986. Taishukan Shoten. [大漢和辭典修訂版, 諸橋轍次編著, 1986, 大修館書店, 日本]

[4] Dae Jaweon (Korean) Dictionary, first edition, 1988, Samseong Publishing. [大字源, 1988, 韓國]

[5] Dictionary of Chinese Character Variants, compiled by Mandarin Promotion Council of Taiwan, version 2 was published in Aug 2001 on Web site: http://140.111.1.40/. [異體字典第二版, 2001, 中華民國教育部國語推行委員會]

[6] A Complete Set of Simplified Chinese Characters, published in 1986 by the Committee of National Language and Chinese Character of China. [簡化字總表, 1986, 中華人民共合國國務院]

[7] CCCII - Chinese Character Code for Information Interchange. 1987. Council for Culture Affairs of Taiwan. [中文資訊交換碼(CCCII), 1994, 資訊應用國字整理小組, 中華民國行政院文化建設委員會]

[8] Chinese Character Code for Information Interchange (CCCII) Character Variant Table. 1994. Council for Culture Affairs of Taiwan. [中文資訊交換碼(CCCII)異體字表, 1994, 資訊應用國字整理小組, 中華民國行政院文化建設委員會]

[9] Ken Lunde. 1999. CJKV Information Processing. O'Reilly, ISBN 1-56592-224-7.

[10] Tanaka Yuuichi, Tanimura Eiji, Furuya Yukio, and Matsouka Eiji. Sanseido's Unicode™ Kanji Information Dictionary. Sanseido. ISBN4-385-13690-4. [漢字情報辭典, 三省堂, 日本].

[11] ISO/IEC 10646-1:2000(E). 2000/10. International Standard - Information technology -- Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane.

[12] The Unicode Consortium, "The Unicode Standard", http://www.unicode.org/unicode/standard/standard.html.

[13] Unihan database, http://www.unicode.org/Public/3.1-Update/Unihan-3.1.1.txt

[14] CNS 11643-1986 Standard Interchange Code for Generally-Used Chinese Characters. 1986. Chinese National Standard. [中文標準交換碼, 1986, 中華民國中央標準局]

[15] CNS 11643-1992 Chinese Standard Interchange Code. 1992. Obsoletes CNS 11643-1986. Chinese National Standard. [中文標準交換碼, 1992, 中華民國中央標準局]

[16] 全字庫, 中華民國行政院主計處電子處理資料中心. http://www.cns11643.gov.tw/web/index.jsp.

[17] BIG5-1984, or simply BIG5, a de facto standard for Traditional Chinese character. [大五碼, 1984, 業界標準]

[18] BIG5-2003, 2003, in appendix of CNS-11643 expansion. [大五碼-2003, 2003, 中文標準交換碼 CNS 11643擴編附錄]

[19] JIS X 0208:1997 7-Bit and 8-Bit Double Byte Coded Kanji Sets for Information Interchange. Japanese Standards Association. 1997. [7位元和8位元雙字節情報交換用符號化漢字集, 1997, 日本標準]

[20] KS X 1001:1992 Code for Information Interchange (Hangul and Hanja). Korean Industrial Standard. 1992. Original designated KS C 5601-1992. [朝鮮語與漢字資訊交換碼, 1992, 韓國標準]

[21] GB 2312-80: Code of Chinese Graphic Character Set for Information Interchange Primary Set. 1981. Technical Standards Press, People's Republic of China. [GB 2312-80:資訊交換用漢字編碼字元集-基本集, 中華人民共合國]

[22] GB 18030-2000: Information technology - Chinese Ideograms Coded Character Set for Information Interchange - Extension for the Basic Set. 2000. Technical Standards Press, People's Republic of China. [GB 18030-2000: 資訊技術-資訊交換用漢字編碼字元集-基本集的擴充, 中華人民共合國]

[23] HKSCS-2004, Hong Kong Supplementary Character Set. 2004. [香港增補字符集, 2004, 中文界面諮詢委員會, 香港特別行政區政府]

[24] 莊德明、謝清俊等. 漢字構形資料庫. http://www.sinica.edu.tw/~cdp/cdphanzi/.

[25] 莊德明, 2007, "構字式的處理技巧," http://www.sinica.edu.tw/~cdp/service/documents/T960419.pdf

[26] 莊德明、謝清俊, 2005/01, “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, 台北.

[27] Der-Ming Juang, Jenq-Haur Wang, Chen-Yu Lai, Ching-Chun Hsieh, Lee-Feng Chien, and Jan-Ming Ho, 2005/06, “Resolving the Unencoded Character Problem for Chinese Digital Libraries,” Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2005), pages 311-319, Denver, Colorado, USA.

[28] 莊德明、鄧賢瑛, 2008/08, “漢字檢字入口網站的規畫及應用,” 第四屆中國文字學國際學術研討會, 山東煙台.

[29] 賴震宇, 漢字智慧型編碼網路工具集, http://icstoolkit.openfoundry.org

[30] 數位典藏技術發展組, 數位典藏web缺字系統, http://140.109.18.63/word/s.html

[31] Chen-Yu Lai, Jan-Ming Ho, You-Qiao Wang, Zhi-Zhueng Huang, 2004, “A composite approach to handle missing characters on Web interface”, ICDAT2004.

[32] 電腦漢字字形及異體字詞彙整合知識庫, http://chardb.iis.sinica.edu.tw/.

[33] Jeng-Wei Lin, Jan-Ming Ho, Li-Ming Tseng, and Feipei Lai, 2008/11, “Variant Chinese Domain Name Resolution,” ACM Transactions on Asian Language Information Processing, Vol. 7, No. 4, P.1-P.29.

[34] Jeng-Wei Lin and Feng-Sheng Lin, 2011/10, “An Auxiliary Unicode Han Character Lookup Service Based on Glyph Shape Similarity,” the 11th IEEE International Symposium on Communications & Information Technologies, Hangzhou, China.

[35] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals”. Soviet Physics Doklady 10 (1966):707-710

[36] Gonzalo Navarro, 2001, “A Guided Tour to Approximate String Matching”, ACM Computing Surveys, Vol.33. No.1, pp.31-88.

# 國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：100年10月18日

| | | | |
|---|---|---|---|
| 計畫編號 | NSC 99－2221－E－029－035－ | | |
| 計畫名稱 | 萬國碼漢字檢字系統改善計畫 | | |
| 出國人員姓名 | 林正偉 | 服務機構及職稱 | 東海大學資訊管理學系<br>助理教授 |
| 會議時間 | 100年10月12日至<br>100年10月14日 | 會議地點 | 中國・杭州 (Hangzhou, China) |
| 會議名稱 | (中文)<br>(英文) The 11th IEEE International Symposium on Communications and Information Technologies (ISCIT 2011) | | |
| 發表論文題目 | (中文)<br>(英文) An Auxiliary Unicode Han Character Lookup Service Based on Glyph Shape Similarity | | |

一、參加會議經過

本人這次出國行程的主要目的是參加在中國杭州(Hangzhou, China)舉行的第11屆IEEE國際通訊與資訊科技研討會(The 11th IEEE International Symposium on Communications and Information Technologies ，簡稱ISCIT 2011)，並發表學術論文。在該會議上發表的論文，均經過嚴格的審查，並被收錄在IEEE Xplorer上面。本次研討會計有來自多國的學者投稿232篇文，共接受106篇，接受率為45%，是一個具有相當高水平的研討會。會議於2011年10月12日至14日在杭州金溪酒店會議廳舉行。本次會議共安排了2場keynote演講，與20個technical sessions，一個special session，與一個industrial forum，主題包含Cognitive Networking, Green Networking, Cooperative Networking, Coding and Modulation, Cross-layer Air Interface, Vehicular Networks, Smart Grid, Multimedia Services, Network Management, Artificial Intelligence and Applications等多項重要的相關議題。

本人這次發表論文的題目為『An Auxiliary Unicode Han Character Lookup Service Based on Glyph Shape Similarity』，安排於 12 日下午的 Technical Session-1 進行口頭發表。這個場次共有 6 篇論文發表，本人被安排在第 6 位，在場多位學者對於這個議題均深感興趣，席間多次提問，休息時間，亦有多名學者與本人交換意見。本人從這些提問與意見的對話中，收穫良多，對日後的研究重點有相當的助益。

除了發表論文之外，本人也全程參與會議議程。會議期間，本人聆聽了許多來自不同領域的學者的研究心得。休息期間，本人也就所得與論文作者討論，受益不淺。特別是同樣來自台灣的暨南大學黃育銘教授，給予本人未來的研究方向很多寶貴的意見，收穫良多。

二、與會心得

本人在這次會議中，認識許多來自不同國家、不同地區、及不同領域的學者，深感收穫甚多。以下本人僅就這次行程中的會議過程與內容，作一簡短的心得報告。

1. 多媒體網路、無線感知網路、多重異質網路與同儕對等式網路是網路領域近年來十分搶眼的議題，在這次的會議中，本人也深深感受到這些領域被許多專家學者重視，仍有多項關鍵技術待進一步的了解與分析，尋求解答。

2. Prof. Tariq S. Durrani 在他以『Going Green and Global for a Golden Future?』為題的 Keynote Speech 特別提到了未來，Green 節能會是重要的議題，與會學者專家與本人均深感同意。

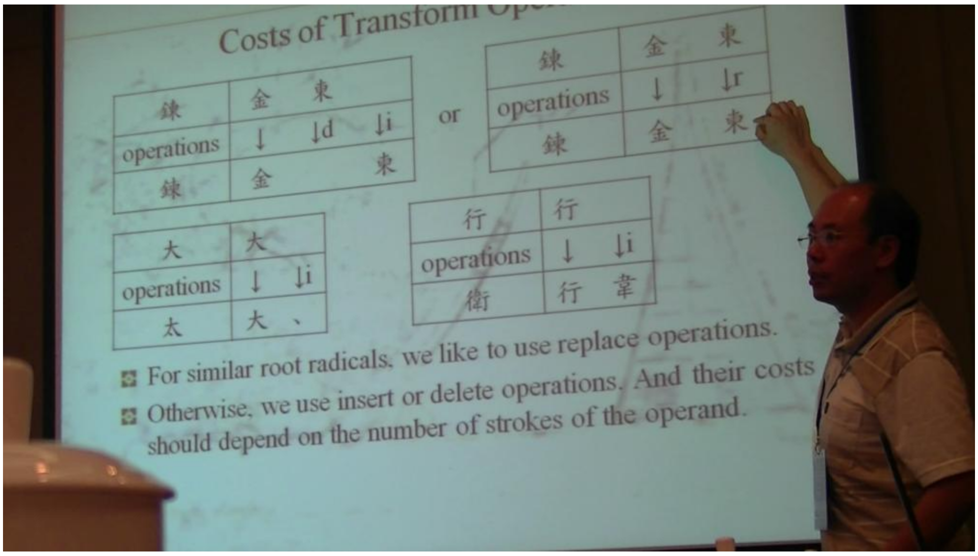
三、攜回資料名稱及內容

1. 一份研討會論文集 CD。

## 四、其他

1. 會議網址：http://www.iscit2011.org/

2. 大會議程與發表論文如附件所示。



Costs of Transform Op

| | |
|---|---|
| 鍊 | 金 東 |
| operations | ↓ ↓d ↓i |
| 鍊 | 金 東 |

or

| | |
|---|---|
| 鍊 | 金 東 |
| operations | ↓ ↓r |
| 鍊 | 金 東 |

| | |
|---|---|
| 大 | 大 |
| operations | ↓ ↓i |
| 太 | 大 、 |

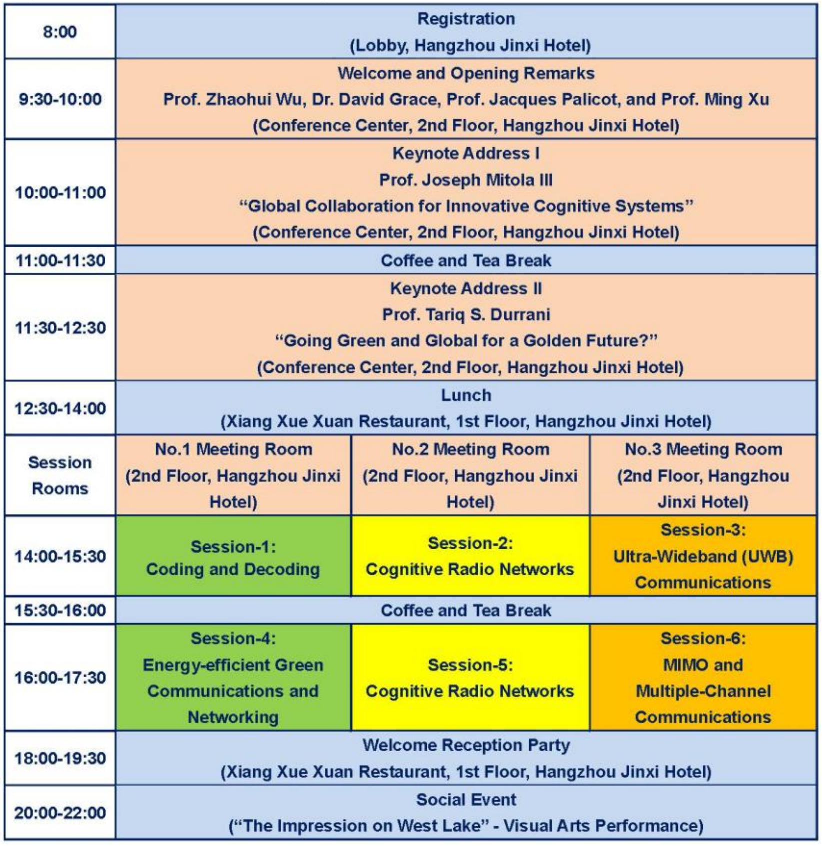
| | |
|---|---|
| 行 | 行 |
| operations | ↓ ↓i |
| 衛 | 行 韋 |

- For similar root radicals, we like to use replace operations.
- Otherwise, we use insert or delete operations. And their costs should depend on the number of strokes of the operand.

照片 1：本人於會場進行口頭報告

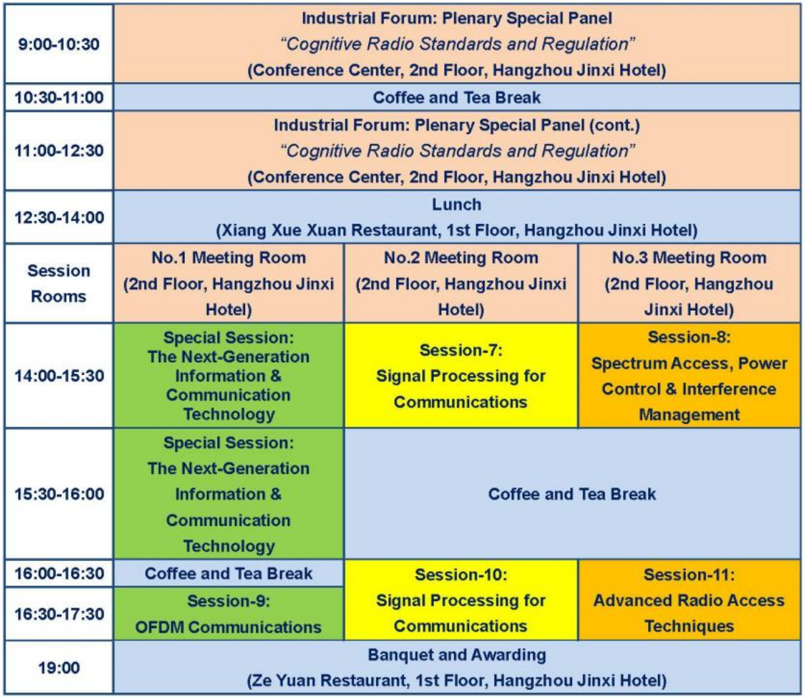
附件一：會議議程

Day 1: 2011/10/12 (Wednesday)



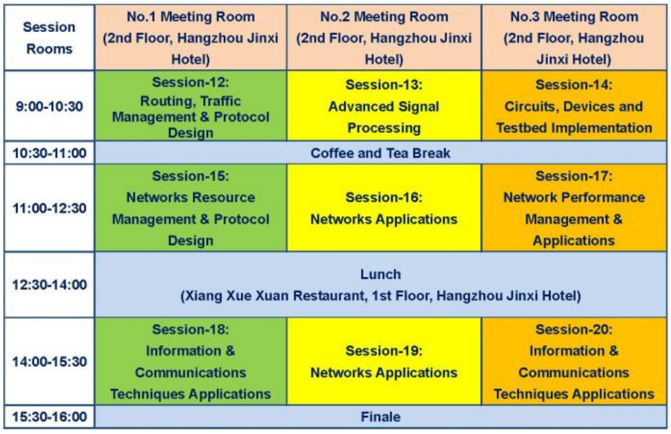
| | | | |
|---|---|---|---|
| 8:00 | Registration<br>(Lobby, Hangzhou Jinxi Hotel) | | |
| 9:30-10:00 | Welcome and Opening Remarks<br>Prof. Zhaohui Wu, Dr. David Grace, Prof. Jacques Palicot, and Prof. Ming Xu<br>(Conference Center, 2nd Floor, Hangzhou Jinxi Hotel) | | |
| 10:00-11:00 | Keynote Address I<br>Prof. Joseph Mitola III<br>"Global Collaboration for Innovative Cognitive Systems"<br>(Conference Center, 2nd Floor, Hangzhou Jinxi Hotel) | | |
| 11:00-11:30 | Coffee and Tea Break | | |
| 11:30-12:30 | Keynote Address II<br>Prof. Tariq S. Durrani<br>"Going Green and Global for a Golden Future?"<br>(Conference Center, 2nd Floor, Hangzhou Jinxi Hotel) | | |
| 12:30-14:00 | Lunch<br>(Xiang Xue Xuan Restaurant, 1st Floor, Hangzhou Jinxi Hotel) | | |
| Session Rooms | No.1 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) | No.2 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) | No.3 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) |
| 14:00-15:30 | Session-1:<br>Coding and Decoding | Session-2:<br>Cognitive Radio Networks | Session-3:<br>Ultra-Wideband (UWB) Communications |
| 15:30-16:00 | Coffee and Tea Break | | |
| 16:00-17:30 | Session-4:<br>Energy-efficient Green Communications and Networking | Session-5:<br>Cognitive Radio Networks | Session-6:<br>MIMO and Multiple-Channel Communications |
| 18:00-19:30 | Welcome Reception Party<br>(Xiang Xue Xuan Restaurant, 1st Floor, Hangzhou Jinxi Hotel) | | |
| 20:00-22:00 | Social Event<br>("The Impression on West Lake" - Visual Arts Performance) | | |

Day 2: 2011/10/13 (Thursday)



| | | | |
|---|---|---|---|
| 9:00-10:30 | Industrial Forum: Plenary Special Panel<br>*"Cognitive Radio Standards and Regulation"*<br>(Conference Center, 2nd Floor, Hangzhou Jinxi Hotel) | | |
| 10:30-11:00 | Coffee and Tea Break | | |
| 11:00-12:30 | Industrial Forum: Plenary Special Panel (cont.)<br>*"Cognitive Radio Standards and Regulation"*<br>(Conference Center, 2nd Floor, Hangzhou Jinxi Hotel) | | |
| 12:30-14:00 | Lunch<br>(Xiang Xue Xuan Restaurant, 1st Floor, Hangzhou Jinxi Hotel) | | |
| Session Rooms | No.1 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) | No.2 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) | No.3 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) |
| 14:00-15:30 | Special Session:<br>The Next-Generation Information & Communication Technology | Session-7:<br>Signal Processing for Communications | Session-8:<br>Spectrum Access, Power Control & Interference Management |
| 15:30-16:00 | Special Session:<br>The Next-Generation Information & Communication Technology | Coffee and Tea Break | |
| 16:00-16:30 | Coffee and Tea Break | Session-10:<br>Signal Processing for Communications | Session-11:<br>Advanced Radio Access Techniques |
| 16:30-17:30 | Session-9:<br>OFDM Communications | | |
| 19:00 | Banquet and Awarding<br>(Ze Yuan Restaurant, 1st Floor, Hangzhou Jinxi Hotel) | | |

Day 3: 2011/10/14 (Friday)



| Session<br>Rooms | No.1 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) | No.2 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) | No.3 Meeting Room (2nd Floor, Hangzhou Jinxi Hotel) |
|---|---|---|---|
| 9:00-10:30 | Session-12: Routing, Traffic Management & Protocol Design | Session-13: Advanced Signal Processing | Session-14: Circuits, Devices and Testbed Implementation |
| 10:30-11:00 | Coffee and Tea Break | | |
| 11:00-12:30 | Session-15: Networks Resource Management & Protocol Design | Session-16: Networks Applications | Session-17: Network Performance Management & Applications |
| 12:30-14:00 | Lunch<br>(Xiang Xue Xuan Restaurant, 1st Floor, Hangzhou Jinxi Hotel) | | |
| 14:00-15:30 | Session-18: Information & Communications Techniques Applications | Session-19: Networks Applications | Session-20: Information & Communications Techniques Applications |
| 15:30-16:00 | Finale | | |

| | |
|---|---|
| Session I: Coding & Decoding | |
| No.1 Meeting Room, Hangzhou Jinxi Hotel<br>Wednesday, October 12, 14:00-15:30<br>Chair: Jinsong Wu, Bell Laboratories | |
| S1-1: | Three-Dimensional Combined Diversity Coding and Error Control Coding: Code Design and Diversity Analysis<br>Jinsong Wu (Bell Laboratories, P.R. China); Pei Xiao (University of Surrey, United Kingdom); Qingchun Chen (Southwest Jiaotong University, P.R. China); Steven D Blostein (Queen's University, Canada) |
| S1-2 | An Improved LDPC Decoding Algorithm Based on Min-sum Algorithm<br>Yue Cao (Northeastern University, P.R. China) |
| S1-3 | An Improved Decoding Algorithm for Refined Clipping Noise in IDMA System with Limited Peak Transmission<br>Yier Yan (Guangzhou University, P.R. China); Xueqin Jiang (Chonbuk National University, P.R. China); Moon Ho Lee (Chonbuk National University, Korea) |
| S1-4 | A Simple Algorithm for a High Code Rate LDPC Parity Matrix Design<br>Sekson Timakul (King Mongkut's Institute of Technology Ladkrabang, Thailand); Somsak Choomchuay (King Mongkut's Institute of Technology Ladkrabang, BKK, Thailand) |
| S1-5 | A Secure Arithmetic Coding Algorithm Based on Integer Implementation<br>Yuh-Ming Huang (National Chi Nan University, Taiwan); Yin-Chen Liang (National Chi Nan University, Taiwan) |
| S1-6 | An Auxiliary Unicode Han Character Lookup Service Based on Glyph Shape Similarity<br>Jeng-Wei Lin (Tunghai University, Taiwan); Feng-Sheng Lin (National Taiwan University, Taiwan) |

# An Auxiliary Unicode Han Character Lookup Service Based on Glyph Shape Similarity*

Jeng-Wei Lin
Department of Information Management
Tunghai University
Taichung, Taiwan (R.O.C)
jwlin@thu.edu.tw

Feng-Sheng Lin
Institute of Information Science
Academia Sincia
Taipei, Taiwan (R.O.C)
skyrain@iis.sinica.edu.tw

*Abstract*—**Most legacy computer systems only well support input and display of 20,902 Han characters (Hanzis for short) encoded in Unicode 1.0. In 2010, Unicode 6.0 has encoded 75,616 Hanzis. However, it is not easy to use these newly encoded Hanzis, even in the latest computers. Most of these newly encoded Hanzis are rarely used in daily lives. Some are only used in ancient literature or individual Sinospherical countries. Users may have confusion of their glyph shapes, pronunciations, meanings, and usages. Most Chinese IMEs (input method editors) require users to have good knowledge of Hanzis. As a result, users cannot input these Hanzis. We present an auxiliary Unicode Hanzi lookup service based on glyph shape similarity. One can key in a similar Hanzi by any IME to look up the wanted Hanzi. Each Unicode Hanzi is decomposed as a glyph expression. The similarity of glyph shapes of two Hanzis is calculated based on a derived edit distance on their glyph expressions. As a result, the system provides users a convenient way to look up unfamiliar Hanzis.**

***Keywords: Unicode; Han character lookup; Hanzi; glyph expression, edit distance.***

## I. INTRODUCTION

In 1991, Unicode had encoded 20,902 Han characters used in Traditional and Simplified Chinese, Japanese, and Korean, in the basic block - CJK Unified Ideographs - in its first version 1.0. We note that Han characters are commonly referred to as Hanzi in Chinese, Kanji in Japanese, and Hanja in Korean. In this paper, Hanzi is used for short. Later, as shown in TABLE I, the number of encoded Hanzis increased as new versions of Unicode released. In 2010, Unicode 6.0 had totally encoded more than 74,000 Hanzis [1][2], including those used in Hang Kong, Vietnam, Singapore, and so on, and in ancient literature. Most of these newly encoded Hanzis are rarely used in daily lives. In some sense, this great repertory solves the unencoded Han character problem[1] largely. However, it is not easy to use these newly encoded Hanzis. Earlier systems typically use 2-byte Unicode and only install the fonts of the initial 20,902 Hanzis encoded in the basic block. Users cannot use the Hanzis encoded in the CJK Unified Ideographs Extensions A~D. New systems, such as Microsoft Windows 7 and Linux, are able to handle 4-byte Unicode and install suitable fonts. However, most Chinese IMEs (input method editors) require users to have good knowledge of Hanzis. A user has to know the pronunciation, shape, or meaning of a wanted Hanzi. If the Hanzi has several character variants or similar characters, the user further has to know which one is wanted and whether it is encoded in Unicode. It is very likely that the user is unfamiliar with a rarely used Hanzi, nor the encoding schemes of Unicode. As a result, the user fails to input these newly encoded Hanzis.

### A. *Principles of Unicode Unification for CJK Ideographs*

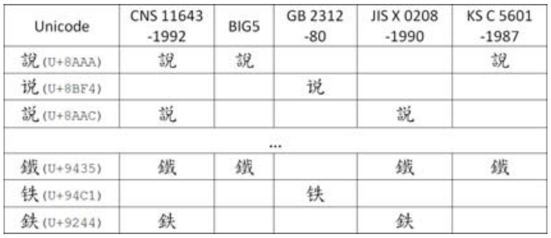
In different time periods and regions, a Hanzi may vary in its shape, pronunciation, meaning, and usage due to many reasons. Two shape-distinguishable Hanzis are said to be character variants of each other if they have the same meaning and are pronounced the same. Some variations are small, while others are significant. Among the character variants of a Hanzi, some are frequently used in one region, and others are in another region; some were only used in ancient literature, and new ones were coined now and then. According to Principles of Han Unification [1][3], these character variants are probably (1) assigned to different codepoints, as shown in Fig. 1, (2) unified to share a same codepoint, as shown in Fig. 2, or (3) unencoded at this time. For case (1), the use of Han character variants may introduce side effects, as discussed in [4]. For case (2), users usually have to use some font technologies to display these character variants. However, if a proper font is

TABLE I. HANZIS ENCODED IN UNICODE 6.0

| **Block** | **Codepoint Range** | **Number of Characters** | **Comment** |
|---|---|---|---|
| CJK Unified Ideographs | `04E00-09FA5`<br>`09FA6-09FCB` | 20,902<br>38 | common |
| Extension A | `03400-04DB5` | 6582 | rare, historic, some in current use |
| Extension B | `20000-2A6D6` | 42,711 | |
| Extension C | `2A700-2B734` | 4,149 | |
| Extension D | `2B740-2B81D` | 222 | |
| Compatibility | `0F900-0FAD9` | 474 | duplicates, unifiable variants |
| Compatibility Supplement | `2F800-2FA1D` | 542 | |

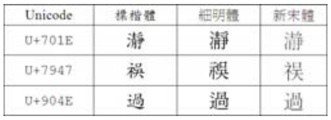
[1] The unencoded Han character problem occurs when a wanted Han character is not encoded in a computer system and thus cannot be used for information processing. An unencoded Han character in one encoding is possibly encoded in another encoding. Many Simplified Chinese characters unencoded in BIG5 [6] are encoded in GB-2312 [7]. It is also possible that an unencoded Han character at this moment will be encoded someday in the future. Before that, however, users usually have to replace the wanted Han character by one of its encoded character variants, to use a scanned image, or to use a self-made font associating with a private code point. In particular software applications, there exist sophisticated plug-ins for users to search and display some of unencoded Han characters [12].

*This work was supported in part by National Science Council, Taiwan (R.O.C.), under Grant NSC 99-2221-E-029-035 and NSC 100-2631-H-001-012.



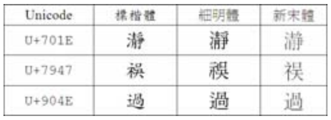
| Unicode | CNS 11643<br>-1992 | BIG5 | GB 2312<br>-80 | JIS X 0208<br>-1990 | KS C 5601<br>-1987 |
|---|---|---|---|---|---|
| 說 (U+8AAA) | 說 | 說 |  |  | 說 |
| 说 (U+8BF4) |  |  | 说 |  |  |
| 説 (U+8AAC) | 説 |  |  | 説 |  |
| ... |  |  |  |  |  |
| 鐵 (U+9435) | 鐵 | 鐵 |  | 鐵 | 鐵 |
| 铁 (U+94C1) |  |  | 铁 |  |  |
| 鉄 (U+9244) | 鉄 |  |  | 鉄 |  |

Figure 1. Hanzi variants are assigned to different codepoints. An empty cell means the specified standard does not encoded the character. CNS 11643 [5] and BIG5 [6], GB 2312 [7], JIX 0208 [8], and KS C 5601 [9] are regional standards of character set encodings used for Traditional Chinese, Simplified Chinese, Japanese and Korean respectively.



| Unicode | 標楷體 | 細明體 | 新宋體 |
|---|---|---|---|
| U+701E | 瀞 | 瀞 | 瀞 |
| U+7947 | 祇 | 祇 | 祇 |
| U+904E | 過 | 過 | 過 |

Figure 2. Unified Hanzi variants share a same codepoint. It is usually required to use certain font technologies to display these character variants.



| Unicode | 標楷體 | 細明體 | 新宋體 |
|---|---|---|---|
| U+701E | 瀞 | 瀞 | 瀞 |
| U+7947 | 祇 | 祇 | 祇 |
| U+904E | 過 | 過 | 過 |

Figure 3. Several Hanzis have the same shape or very similar shapes.

not selected, a user may mistake an encoded Hanzi for unencoded. For example, a user may mistake 祇 for unencoded if the user only installs 標楷體 and 細明體. It is also possible several irrelevant Hanzis have the same shape or very similar shapes. They usually had been assigned distinct codepoints, as shown in Fig. 3.

Usually, a user types a sequence into a Chinese IME to pick up a wanted Hanzi. If not found, the user may type another sequence or switch to another IME. Sometimes, the user may doubt whether the wanted is encoded in Unicode. Unihan database provides Unicode Hanzi lookup services via radical-stroke index [10]. International Encoded Hanzi and Variants Database additionally supports search by components [11]. These services also require users to have good knowledge of the wanted Hanzi. Furthermore, the radical and number of strokes of a Hanzi are probably different from the user's guess due to the unification process.

In this paper, we present a Unicode Hanzi lookup service, in which to look up a Hanzi, a user can simply input a similar Hanzi by any IME. IMEs typically emphasize a high precision rate of a Hanzi lookup by a short sequence. In contrast to IMEs, a Hanzi lookup service for rarely used Hanzis should have a higher recall so that users can find the wanted Hanzi easily. We use a glyph expression to describe the glyph shape of a Hanzi [12]. The similarity of two Hanzis is estimated by the edit distance [13] between the two glyph expressions.

The paper is organized as follows. We describe our method and related issues in Section II. We present the resultant Hanzi lookup service in Section III. Finally, we conclude the paper and give some future direction in Section IV.

## II. The Method To Measure The Similarity

### A. Glyph Expressions

In general, a Hanzi is either an atomic glyph unit, referred to as a root radical, or composed of several root radicals. For example, Hanzi 說 consists of two components, 言 and 兌. It can be further decomposed into four root radicals, 言, 八, 口, and 儿, as shown in Fig. 4. We use "言八口儿" as the reduced glyph expression of 說. Fig. 5 shows more examples.

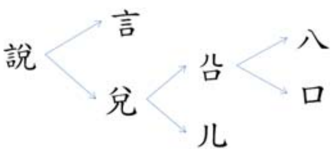
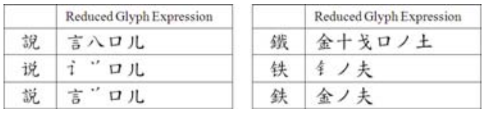


Figure 4. Hanzi 說 consists of four root radicals, 言, 八, 口, and 儿.



|  | Reduced Glyph Expression |
|---|---|
| 說 | 言八口儿 |
| 说 | 讠丷口儿 |
| 説 | 言丷口儿 |

|  | Reduced Glyph Expression |
|---|---|
| 鐵 | 金十戈口丿土 |
| 铁 | 钅丿夫 |
| 鉄 | 金丿夫 |

Figure 5. More examples of reduced glyph expressions.

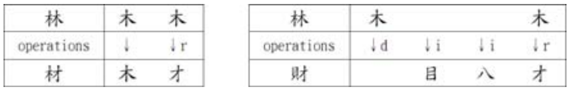
### B. Similarity of Two Hanzis

In this study, we estimate the similarity of glyph shapes of two Hanzis by the edit distance [13] of their corresponding reduced glyph expressions. The edit distance is defined as the cost of edit operations required to transform one expression into the other. We note that the edit distance between expressions can be easily computed by a dynamic algorithm. As shown in Fig. 6, we can transform one Hanzi into another, by insertion, deletion, and replacement of root radicals. Thus,

$$\text{Cost}(\text{林}\rightarrow\text{材})=\text{Cost}_r(\text{木}, \text{才}).$$

And

$$\text{Cost}(\text{林}\rightarrow\text{財})$$
$$=\text{Cost}_d(\text{木})+\text{Cost}_i(\text{目})+\text{Cost}_i(\text{八})+\text{Cost}_r(\text{木}, \text{才}).$$



| 林 | 木 | 木 |
|---|---|---|
| operations | ↓ | ↓r |
| 材 | 木 | 才 |

| 林 | 木 |  |  | 木 |
|---|---|---|---|---|
| operations | ↓d | ↓i | ↓i | ↓r |
| 财 |  | 目 | 八 | 才 |

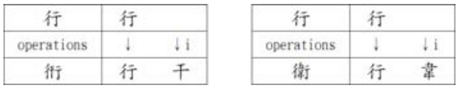
Figure 6. Examples of transformation between reduced glyph expressions. i, d, and r beside ↓ denote insertion, deletion, and replacement respectively.

It is trivial that 衎 is more similar to 行 than 衛, as shown in Fig. 7. The cost of insertion or deletion of a root radical is thus defined as the number of strokes in the root radical.

However, if we transform 鍊 into 鍊 by first deleting 東 and then inserting 東, the cost is high. In this case, the transformation is done by replacing 東 with 柬, as shown in Fig. 8. In other words, there are two cases.

*1)* For similar root radicals, we want to use replacement edit operation.

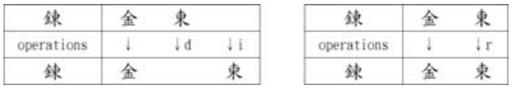
*2)* For dissimilar root radicals, we want to use insertion or deletion edit operations.



| 行 | 行 | |
|---|---|---|
| operations | ↓ | ↓i |
| 衎 | 行 | 干 |

| 行 | 行 | |
|---|---|---|
| operations | ↓ | ↓i |
| 衛 | 行 | 韋 |

Figure 7. Examples of edit distances



| 錬 | 金 | 東 | |
|---|---|---|---|
| operations | ↓ | ↓d | ↓i |
| 鍊 | 金 | | 柬 |

| 錬 | 金 | 東 |
|---|---|---|
| operations | ↓ | ↓r |
| 鍊 | 金 | 柬 |

Figure 8. Examples of edit distances

As a result, if two root radicals A and B are similar, inequality (1) must hold.

$$\mathrm{Cost}_r(A, B) < \mathrm{Cost}_d(A) + \mathrm{Cost}_i(B) \qquad (1)$$

Otherwise, if two root radicals A and B are dissimilar,

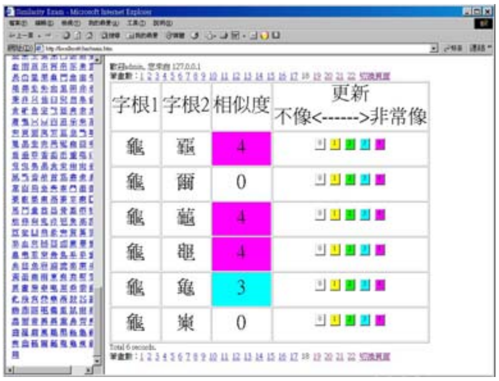
$$\mathrm{Cost}_r(A, B) = \mathrm{Cost}_d(A) + \mathrm{Cost}_i(B) \qquad (2)$$

Hence, we have to define the similarity of two root radicals before we define these cost functions.

## III. Unicode Hanzi Lookup Service

### *A. Similarity Between Root Radicals*

In [12], there are totally 1,151 root radicals. We have built a web-based system to manually set the similarity of two root radicals. As shown in Fig. 9, the similarity of two root radicals A and B, Sim(A, B), are labeled to 0~4, where 0 indicates they are dislike, and 4 indicates they are very similar.



| 字根1 | 字根2 | 相似度 | 更新<br>不像<------>非常像 |
|---|---|---|---|
| 龜 | [illegible] | 4 | 0 1 2 3 4 |
| 龜 | [illegible] | 0 | 0 1 2 3 4 |
| 龜 | [illegible] | 4 | 0 1 2 3 4 |
| 龜 | [illegible] | 4 | 0 1 2 3 4 |
| 龜 | [illegible] | 3 | 0 1 2 3 4 |
| 龜 | [illegible] | 0 | 0 1 2 3 4 |

Total 6 records.

Figure 9. Setting similarity between root radicals

### *B. Implementation*

In this study, the cost of insertion or deletion of a root radical A is defined as its number of strokes, as shown in (3).

$$\mathrm{Cost}_i(A)=\mathrm{Cost}_d(A)=\mathrm{Strokes}(A) \qquad (3)$$

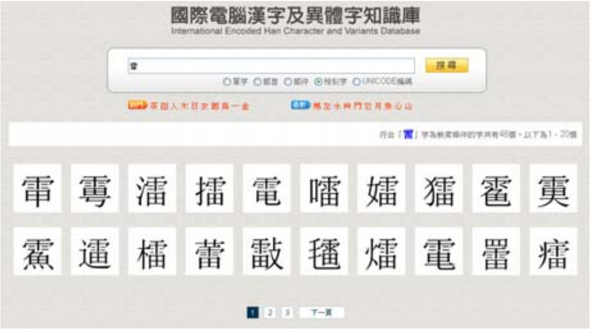
The cost to replace a root radical A with B is defined as (4). This definition obeys (1) and (2) when the total number of strokes of A and B is more than four. In general, we think replacement between similar complex root radicals is more significant, such as 東 and 柬.

$$\mathrm{Cost}_r(A, B)=\begin{cases} 5-\mathrm{Sim}(A,B) & \text{, if } \mathrm{Sim}(A,B)>0 \\ \mathrm{Cost}_d(A)+\mathrm{Cost}_i(B) & \text{, else} \end{cases} \qquad (4)$$

According to [12], every Hanzi encoded in Unicode CJK Unified Ideographs and Extension A and B is described by its glyph expression in [11]. These glyph expressions are further decomposed into reduced glyph expressions.

The similarity between each pair of Hanzis encoded in Unicode is calculated by a dynamic programming algorithm [13]. For a Hanzi, we sort the resultant similarities between it and all the others and store into database 100 Hanzi candidates whose edit distances to the Hanzi are smaller than the others.

Currently, we had integrated this service into [11]. Fig. 10 shows a lookup result when a user searches similar Hanzis by Hanzi 雷, which the user can easily key in by any IME.



國際電腦漢字及異體字知識庫
International Encoded Han Character and Variants Database
雷
搜尋
擂
電
蕾
電
1 2 3 下一頁

Figure 10. Similar Hanzi lookup by 雷

The user can further look up a Hanzi shown on the returned page for more information, as shown in Fig. 11. Thus, the user can make sure which one is wanted and its Unicode codepoint.

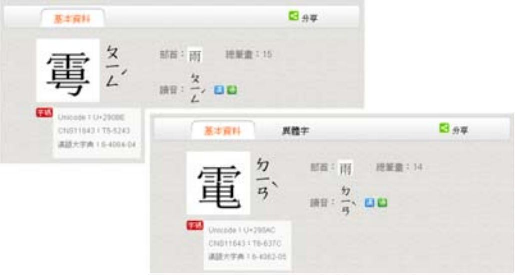


Figure 11. More information about the returned similar Hanzis

### C. Evaluations and Discussions

In the beginning, we tried to solve this problem by OCR (Optical Character Recognition) software. However, for the total 70,195 Hanzis printed on image files using standard fonts, only 3,000 or so of them were successfully recognized by a commercial Chinese OCR product. Most of recognized Hanzis were frequently used. We note that most OCR software is not trained to recognize so many different Hanzis. Furthermore, OCR software depends on corpuses of phrase to correct its recognition result heavily. However, in that experiment, the Hanzis were printed one-by-one and thus not semantically meaningful.

In this study, we adopt the concept of edit distance between the reduced glyph expressions of two Hanzis. We examined the resultant similarities between Hanzis by human. We hired several students to pick up a most-similar Hanzi in the list of the 100 returned candidates. They recorded the position of that Hanzi in the list. As shown in Fig. 12, they almost could pick up a most-similar Hanzi in the list. In fact, for 86% of all Hanzis, they could pick up a most-similar Hanzi in the first 20 candidates. However, there were 1,000 or so Hanzis that they failed to pick up a most-similar Hanzi in the list. For example, 鄉 and 鄕 were considered as non-similar according to their edit distance, as shown in Fig. 13.

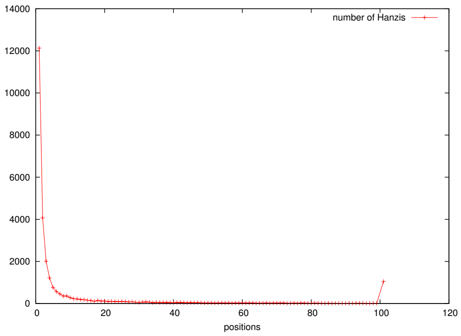


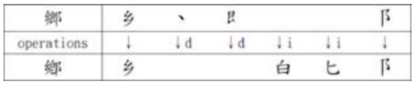
Figure 12. This figure shows that most-similar Hanzis picked up by human are in the list of the 100 returned candidates.

## IV. Conclusions

At the time this paper was written, Unicode has encoded more than 70,000 Hanzis in its repertory. It is expectedly that the number will increase to more than 100,000 in the near future. However, most of these Hanzis are rarely used in daily lives. As a result, it is not easy for generic users to use these Hanzis in computers. A reason is that people do not know how to input these Hanzis into computers by traditional IMEs, which usually require users to have good knowledge of a wanted Hanzi. In this paper, we present an auxiliary Unicode Hanzi lookup service. In contrast to IMEs, the Hanzi lookup service for rarely used Hanzis should have a higher recall so that users can find a wanted Hanzi easily. People can key in a similar Hanzi by any IME to look up the wanted Hanzi in this service.

We observed that two similar Hanzis usually can be easily transformed into each other by insertion, deletion, or replacement of some radicals. We calculate the cost of these operations to estimate the similarity of two Hanzis.

Currently, we had not yet considered the relative position between root radicals. 加 and 召 are recognized as similar. Sometimes it is required and sometimes not. In addition, some complex root radicals probably can be further decomposed. For example, 艮 can be viewed as the combination of 日 and 厶. We note that this kind of decomposition is somehow unreasonable according to the construction principal of Hanzis. However, it probably helps the calculation of the similarity of Hanzis. We will further study these problems in the future.



| 鄉 | 乡 | 丶 | 艮 | | | 阝 |
|---|---|---|---|---|---|---|
| operations | ↓ | ↓d | ↓d | ↓i | ↓i | ↓ |
| 鄕 | 乡 | | | 白 | 匕 | 阝 |

Figure 13. Edit distance between 鄉 and 鄕

## References

[1] The Unicode Standard, The Unicode Consortium, version 6.0, 2010. Available: http://www.unicode.org/.

[2] International Standard - Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane, ISO/IEC 10646-1:2000(E), 2000.

[3] Ken Lunde, CJKV Information Processing, O' Reilly, 1999.

[4] Jeng-Wei Lin, Jan-Ming Ho, Li-Ming Tseng, and Feipei Lai, “Variant Chinese Domain Name Resolution,” ACM Transactions on Asian Language Information Processing, Vol. 7, No. 4, 2008.

[5] Chinese Standard Interchange Code, Chinese National Standard, CNS 11643-1992, 1992.

[6] BIG5-1984 (simply BIG5), a de facto standard for Traditional Chinese character, standardized in appendix of CNS-11643 expansion, referred to as BIG5-2003.

[7] Code of Chinese Graphic Character Set for Information Interchange Primary Set, Technical Standards Press, People's Republic of China, GB 2312-80, 1981.

[8] 7-Bit and 8-Bit Double Byte Coded Kanji Sets for Information Interchange, Japanese Standards Association, JIS X 0208:1997, 1997.

[9] Code for Information Interchange (Hangul and Hanja), Korean Industrial Standard, KS X 1001:1992, 1992. Original designated KS C 5601-1992.

[10] Unihan Database. Available at: http://www.unicode.org/charts/unihan.html

[11] International Encoded Han Character and Variants Database. Available at: http://chardb.iis.sinica.edu.tw/.

[12] Der-Ming Juang, Jenq-Haur Wang, Chen-Yu Lai, Ching-Chun Hsieh, Lee-Feng Chien, and Jan-Ming Ho, “Resolving the Unencoded Character Problem for Chinese Digital Libraries,” 5th ACM/IEEE Joint Conference on Digital Libraries, Denver, Colorado, USA, 2005.

[13] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” Soviet Physics Doklady 10:707–710, 1966.

# 國科會補助計畫衍生研發成果推廣資料表

日期:2012/03/29

| | |
|---|---|
| 國科會補助計畫 | 計畫名稱: 萬國碼漢字檢字系統改善計畫 |
| | 計畫主持人: 林正偉 |
| | 計畫編號: 99-2221-E-029-035- 學門領域: 自然語言處理與語音處理 |

無研發成果推廣資料

# 99 年度專題研究計畫研究成果彙整表

計畫主持人：林正偉 計畫編號：99-2221-E-029-035-

計畫名稱：萬國碼漢字檢字系統改善計畫

| 成果項目 | | | 量化 | | | 單位 | 備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等） |
|---|---|---|---|---|---|---|---|
| | | | 實際已達成數（被接受或已發表） | 預期總達成數(含實際已達成數) | 本計畫實際貢獻百分比 | | |
| 國內 | 論文著作 | 期刊論文 | 0 | 0 | 100% | 篇 | |
| | | 研究報告/技術報告 | 0 | 0 | 100% | | |
| | | 研討會論文 | 0 | 0 | 100% | | |
| | | 專書 | 0 | 0 | 100% | | |
| | 專利 | 申請中件數 | 0 | 0 | 100% | 件 | |
| | | 已獲得件數 | 0 | 0 | 100% | | |
| | 技術移轉 | 件數 | 0 | 0 | 100% | 件 | |
| | | 權利金 | 0 | 0 | 100% | 千元 | |
| | 參與計畫人力（本國籍） | 碩士生 | 1 | 1 | 100% | 人次 | |
| | | 博士生 | 1 | 1 | 100% | | |
| | | 博士後研究員 | 0 | 0 | 100% | | |
| | | 專任助理 | 0 | 0 | 100% | | |
| 國外 | 論文著作 | 期刊論文 | 0 | 0 | 100% | 篇 | |
| | | 研究報告/技術報告 | 0 | 0 | 100% | | |
| | | 研討會論文 | 1 | 1 | 100% | | |
| | | 專書 | 0 | 0 | 100% | 章/本 | |
| | 專利 | 申請中件數 | 0 | 0 | 100% | 件 | |
| | | 已獲得件數 | 0 | 0 | 100% | | |
| | 技術移轉 | 件數 | 0 | 0 | 100% | 件 | |
| | | 權利金 | 0 | 0 | 100% | 千元 | |
| | 參與計畫人力（外國籍） | 碩士生 | 0 | 0 | 100% | 人次 | |
| | | 博士生 | 0 | 0 | 100% | | |
| | | 博士後研究員 | 0 | 0 | 100% | | |
| | | 專任助理 | 0 | 0 | 100% | | |

| | |
|---|---|
| 其他成果<br>(無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。) | 無 |

| 🟥 | 成果項目 | 量化 | 名稱或內容性質簡述 |
|---|---|---|---|
| 科教處計畫加填項目 | 測驗工具(含質性與量性) | 0 | |
| | 課程/模組 | 0 | |
| | 電腦及網路系統或工具 | 0 | |
| | 教材 | 0 | |
| | 舉辦之活動/競賽 | 0 | |
| | 研討會/工作坊 | 0 | |
| | 電子報、網站 | 0 | |
| | 計畫成果推廣之參與（閱聽）人數 | 0 | |

# 國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

   ■達成目標

   □未達成目標（請說明，以 100 字為限）

   □實驗失敗

   □因故實驗中斷

   □其他原因

   說明：

2. 研究成果在學術期刊發表或申請專利等情形：

   論文：□已發表 ■未發表之文稿 □撰寫中 □無

   專利：□已獲得 □申請中 ■無

   技轉：□已技轉 □洽談中 ■無

   其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

   使用 Unicode 已經是發展國際化軟體的趨勢，但由於漢字字數太多，許多基礎的工作並未完善，需要許多文字專家和電腦專家合作進行資料庫的整理與開發。中文是我國國民使用的主要文字，本計畫之成果為相關基礎建設之一，可有效降低國人使用 Unicode 漢字的困難。