國際電腦漢字及異體字知識庫

本院覽號

05T-1041205

公告日期

智財權狀態

know-how

摘要

本系統提供使用者利用字形相關資訊或直接用電腦編碼查詢到所要需的漢字，進而得到此字體的相關屬性以及相關異體字，首先使用者需先對漢字構形以及中文編碼有基本的認識，以下是簡略的介紹。 【部件與構字式】：漢字可以說是由許許多多的小部件所組成，部件就像是小小的建築積木，每一個漢字都是由數個部件堆砌而成；根據中央研究院文獻處理實驗室的統計，所有的基本部件總數為1316個，也就是說，每一個漢字都可以由這些基本部件來組成；當一個漢字用一組部件來表示的時候，這一組部件我們稱之為構字式。 稍微了解電腦資料處理的人都應該知道，在電腦開始使用的時候，所有的字都只用 1byte 來儲存， 1byte 包含 8bits，每個 bits 都只能表示 on/off， 也就是 1byte 只能表示 0000 0000 到 1111 1111 的編碼範圍， 只有 256 個編碼空間，這對中文而言，是不夠的。 我們知道中文字在目前常見的電腦上是由兩個位元組(two bytes) 所編碼組成的。 最常見的編碼方式有台灣地區所通行的 Big5 編碼，及大陸地區所使用的 GB 編碼。而且開頭的位元組幾乎都是大於 128 的數值，也就是所謂 non-ASCII 碼的範圍(ASCII 是指小於 128 的編碼)。

技術優勢

雖然常用的 Big5 已經使用 2bytes 來表示中文字，但是 2bytes = 16bits = 2^16 = 65536 個編碼空間， 以 Big5 的標準而言，為了要和 ASCII 能夠相容，只能使用兩萬多字， 現存的中文字最少在七萬以上，造成許多字在 Big5 的系統下， 無法使用。在加上中文標準繁多，卻又沒有最後的標準規格， 各家廠商所實做產品也就未必相容。 最明顯的例子就是日文平假片假名， 在這些中文編碼中並不是每個都包含， 當遇到所謂的「Big5 日文」時，就會產生許多問題。 為了解決編碼字數不足的問題，我們可以使用國際標準ISO/IEC 10646-1:1993廣用多八位元編碼字元集(unicode)，此為一套用來表示、傳輸、交換、處理、儲存、輸入和表達等多用途的全球編碼標準。目前因為有Unicode Consortium組織的全力推廣與實作介紹，故得到全球各大廠商與資訊界的重視。這套編碼字元集，幾乎已包括了全球已定義好完整字集的各種語言文字，並且仍在持續擴充中。

應用範圍

中文字處理

創作人

何建明

中央研究院
ACADEMIA SINICA