THE DESIGN OF A CROSS-REFERENCE DATA BASE FOR CHINESE CHARACTER INDEXING†

*Chung-tao Chang**
*Jack. K. T. Huang***
*Chen-chau Yang****
*Ching-chun Hsieh*****

1. Introduction

There are many Chinese data processing systems available nowadays. But, the data processing capability of each system is quite limited. One of the reasons that led to this situation is the lack of powerful indexing tools to identify characters stored in computer.

Usually, the search keys of Chinese character that a computer can offered are limited to only one kind. And this search key is directly related to the Chinese input method or the collating sequence of the Chinese character internal code used by the computer system. For example, if a computer use phonetic code of Chinese characters as an input method, then he may provide you to search characters by using the phonetic code of the character as the search key. You know what might happen if you want to use the stroke count of character to find it. It will not work at all unless the computer has stored all the stroke count information in his memory. Most possibly, the computer

---

† Paper presented at the International Workshop on Chinese Library Automation, Taipei, Feb. 14-19, 1981.

* Professor, Dept. of Electronic Engineering, National Taiwan Institute of Technology.
** Professor, Dept. of Computer Science, Ming Chuan College.
*** Director, Software Center.
**** Professor, Dept. of Electronic Engineering, National Taiwan Institute of Technology.

use stroke count as the collating rule to assign the internal code of characters so that they can provide you the information needs to search characters by stroke count. Since, in most Chinese data processing systems, the input method also closely related to the collating rule for assigning internal code, they only allows you to search character by one kind of key. This situation is untolerable to public servicing system, because user may have a wide spectrum of different problems that must use different keys to do their jobs. On the other hand, it is of no good to prohibit user to use a popular method with which he is familiar to input or process Chinese character just because the computer system can not serve it.

The main theme of this paper is try to present a way of solving this problem.

2. The characteristics of Chinese character indexing

The Chinese character indexing problem is defined as the problem of identifying, finding or addressing a Chinese character by its proprietary information, such as its pronunciation, stroke count, radical telegraph code etc. It is quite clear that the experience of processing alphabets can not be applied here directly. When we are searching for a character, we usually search it by some information associated with the character. Therefore, it is not a simple code match problem, but an associative search problem. As a consequence, we must store all the necessary information provided the capability of doing associative search.

3. The constituent of the Cross-Reference Data Base

The information which constitutes the Cross-Reference Data Base is the proprietary information of characters, which is popular and likely to be used as a key for indexing characters.

The information we collected includes the following items of a character.

(1) CCCII[1]
(2) Radical
(3) Stroke count
(4) Stroke sequence
(5) Component expression (or radical expression)[4]
(6) Phonetic codes include: Kuo-Yu, Wade-jile, Yale, Pinyin and Liu,[2]
(7) Telegraph code
(8) Three-corner code[3]
(9) Four-corner code
(10) Internal codes of various Chinese data processing systems.

The purpose of having the internal codes is try to provide corss-reference service of those informations to the existing data processing system. And the data-base is named after it.

4. The structure of the data base

The block diagram of the Cross-Reference Data Base for Chinese Character indexing is shown in Figure-1. In Figure-1, besides the related files of those informations described in the last section, there are files for character strock images. It is integrated all together as a whole in order to achieve system efficiency for both input, processing and output.

5. The Data Structure of Files

The data structure of each record in those files are shown in Figure 2. It is estimated that the total space required for this data base of approximately 50000 characters and 15000 variant

forms is less than 15 M bytes. Among all the files, the smallest one is the files for pronunciation indexing. It is only 100K byte that can cover all the popular pronunciation systems. This part is bared on the conparactive study of phonetic systems by D.J. Liu.[2] Just because its effectiveness, a part of the system has been implemented by some companies in minicomputer or even in micro-computer.

6. Concluding Remark

This data base has not been completed yet. A part of the work has been published in the comprehensive dictionary of Chinese character index.[4] The purpose of this data base is two-fold: first, to provide service of code conversions and cross-referenceing and second to provide a base for further studies on Chinese characters. Serve to the public will be arranged before the end of 1981.

7. Acknowledgement

This research is a part of the work conducted by the Chinese Character Analysis Group which is support by two private non-profit fundations, namely 明德 and 元智 . It is their support that make this research possible.

Reference

1. Hsieh Ching-chun et., The Design and Application of the Chinese Character Code for Information Interchange, this workshop, 1981.
2. D.J. Liu, A Comparative Study of Romanization Systems, this workshop 1981.
3. Jack K.T. Huang et al., The Digitized Chinese Dictionary, Taipei, System Publication, 1979.
4. Dah-jen Liu et al., The Comprehensive Dictionary of Chinese Character Index, New York, Asian Associates, 1979.

檔案結構圖

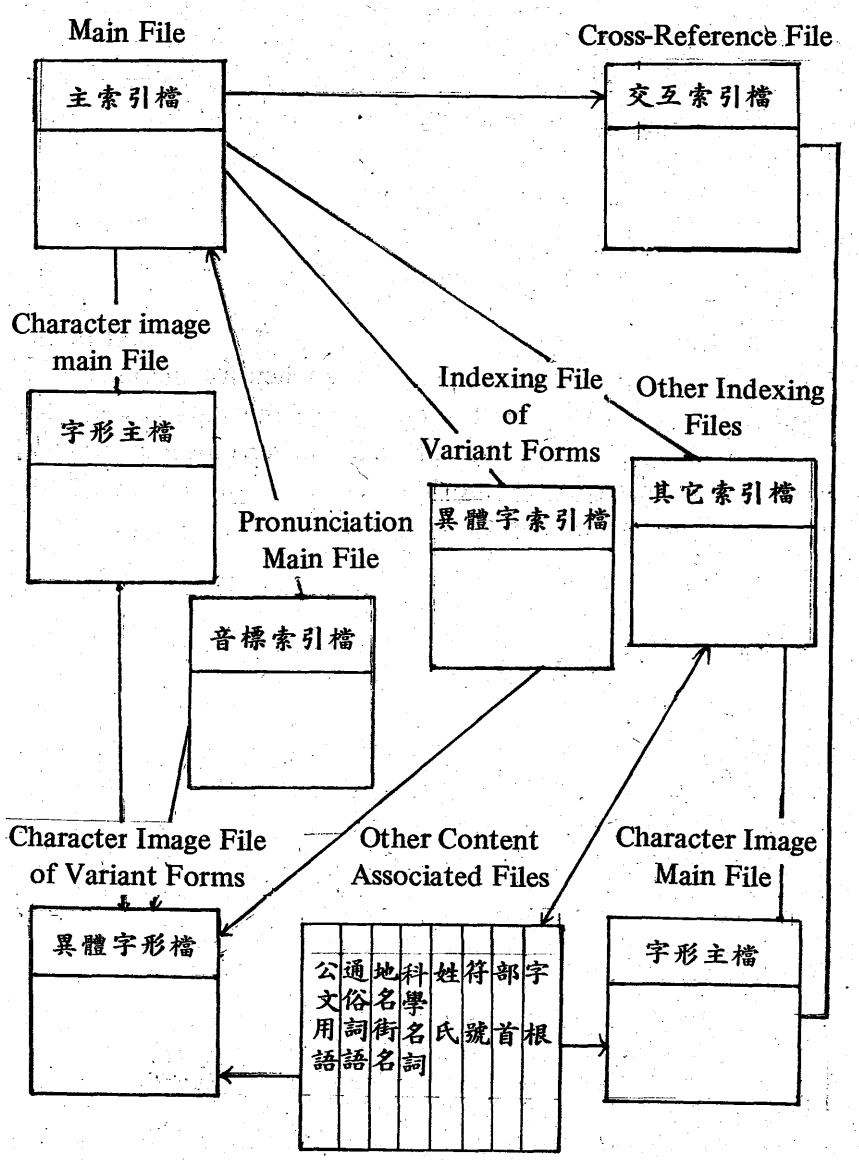
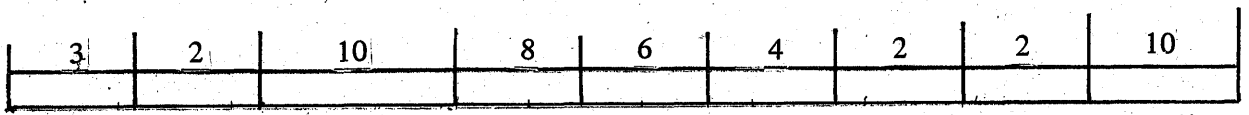


Figure 1. The Block Diagram of the Cross-Reference Data Base for Chinese Indexing

主索引檔 **Main Index File**



47 Bytes/Record

235 M Bytes/50000 characters

字形主檔 **Stroke Image Main File**

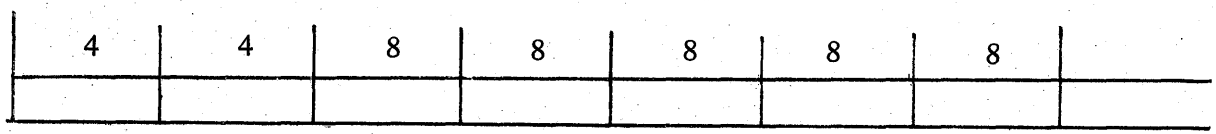
132 Byte/Record

6.6 M Bytes/50000 ch.

異體字形檔 **Stroke Image File For Variant Forms**

133 Byte/Record 2 M Bytes/15000 ch.

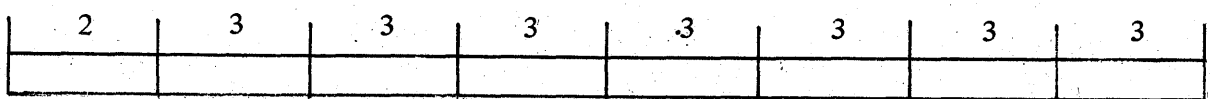
音標檔 **Pronunciation File**



48 Byte/Record

0.1 M Bytes/2100 pronunciations

異體索引檔 **Index File For Variant Forms**



29 Byte/15000 ch.

0.435 M Bytes/15000 ch.

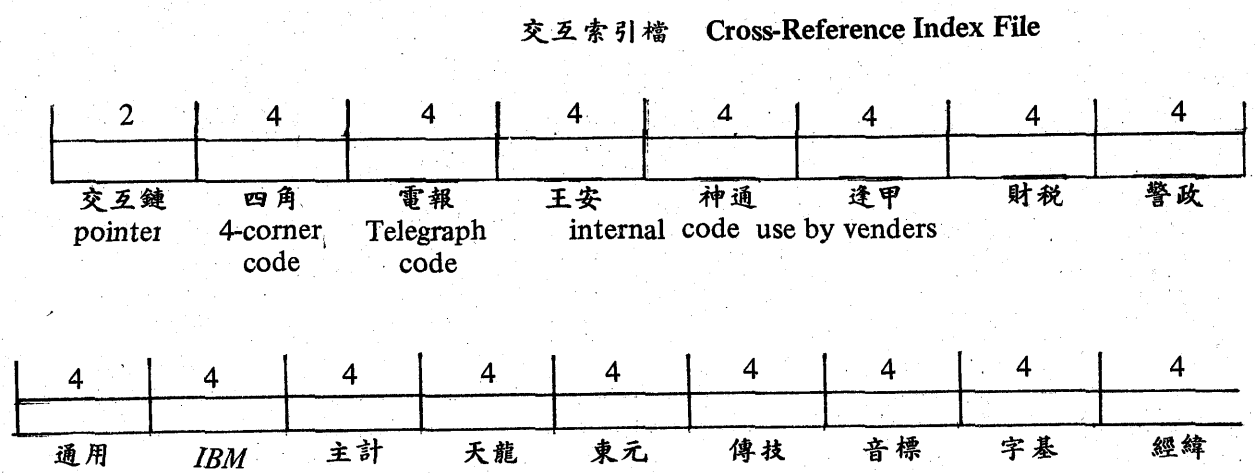


Figure 2: Data Structure of Each Files