A CHINESE CHARACTERS CODING SCHEME FOR COMPUTER INPUT AND INTERNAL STORAGE

Chorkin Chan, Computer Centre, University of Hong Kong, Hong Kong

Abstract

A coding scheme for inputting Chinese characters by means of a conventional keyboard has been developed. The code for each Chinese character is composed of two strings of keys, one corresponds to the spelling and the other the ideographic property of the character. Each code requires no more than seven keys (average five and a half keys) and 99.5% of the ten thousand characters in a dictionary 'XianDai HanYu CiDian' have unique codes. Each input code can be packed into 32 bits for internal representation.

Introduction

Over the last few years, encoding Chinese characters has become a very active subject of research. Numerous papers have appeared, mainly written in Chinese (hence difficult to be referenced in English), proposing various kinds of inputting schemes. Unfortunately, most of these papers offered only the ideas without accompanying implementation and experimentation. This paper presents a coding scheme of Chinese characters based on their ideographic properties as well as their spellings so that a conventional typewriter keyboard can be used for inputting purposes. This scheme has been implemented at the University of Hong Kong using an IBM 3031 under VM/CMS. Without a proper output device to display the Chinese characters, when the code of a Chinese character is entered, the address of that character (where it can be found) in a dictionary 'XianDai HanYu CiDian' is displayed. This is awkward but still sufficient to prove the correctness of the code recognition procedure.

The Coding Scheme for Inputting

In this scheme, a code for a Chinese character consists of two strings of symbols concatenated together. One string of three symbols corresponds to the ideographic radicals the character is composed of.The other of no more than four symbols is the spelling of the character. Corresponding to each of the ten thousand characters in the dictionary 'XianDai HanYu CiDian', with the exception of twenty six pairs, there exists a unique code in this scheme. In other words, this coding scheme is 99.5% unique. Furthermore, among these pairs of characters sharing the same codes in a pair-wise manner, the non-uniqueness of eighteen pairs can be easily removed by deleting one member of each pair from the vocabulary because they are either 'dead' characters appearing in ancient classics only or they can be replaced by other characters of equivalent meaning. The non-uniqueness of the remaining eight pairs can be removed also by either changing the ideographic pattern or the spelling of one of the members in each pair. Thus, by means of these remedial measures, this coding scheme offers a unique code to each of the ten thousand characters found in this dictionary. The list of characters sharing the same codes is in Table 1 together with the suggested remedies to overcome the problem of non-uniqueness.

The Spelling of Chinese Characters

There are two standard systems to spell Chinese characters, one in terms of the Latin alphabets and the other in terms of Mandarin Pin Yin symbols. By means of the former; a maximum of five alphabets are normally required to spell a Chinese character. However, since the alphabet 'G' (except when it is the leading alphabet) always appear with 'N' as 'NG', one can replace 'NG' with 'G' and reduce the maximum number of alphabets required from five to four. By means of the latter, no more than three symbols are required to spell a Chinese character. This can be an important saving but in this paper, spellings are in terms of Latin alphabets just because a conventional terminal keyboard does not have Mandarin Pin Yin keys.

It is not always obvious whether one should read certain Chinese characters with or without a curling tongue, i.e., whether one should spell with 'C' or 'CH', 'S' or 'SH' and 'Z' or 'ZH'. This is particularly difficult to those whose mother tongue is not Mandarin. In order to be more forgiving, this coding scheme allows one not to differentiate 'C' from 'CH', 'S' from 'SH' and 'Z' from 'ZH' so that, for example, 'SHAO' can be spelled as 'SAO'. As a consequence, there will be three additional pairs of characters sharing the same codes in a pair-wise manner as listed in Table 2. Fortunately, the non-uniqueness so engendered can be easily eliminated by deleting one member of each pair because of its rare occurrence. For the same reason, this coding scheme also allows one to confuse a leading 'N' with a leading 'L'. For example, 'LUAN' can be spelled as 'NUAN' and vice versa. No non-uniqueness is introduced as a result of this

Table 1: Pairs of Chinese Characters Sharing the Same Codes

Spelling	Radical Composition	Char. 1	Char. 2	Suggested Remedy	Justification
AN	5	厂	广	delete char. 2	same meaning
BI	9TE	蔽	弊	write char. 2 as 痲	it means a defect
BO	M	尟	跛	delete char. 1	same meaning
DIAO	V	鳥	弔	delete char. 2	replaced by 弔
DUN	KB	呑	囤	delete char. 1	uncommon
E	KGX	噩	婀	delete char. 1	uncommon
E	T2-	吪	囫	delete char. 2	uncommon
FU	VM	怫	弗	spell char. 2 as FO	so is 佛
GU	K2	咕	固	delete char. 1	uncommon
JIA	FDK	架	枷	write char. 2 as 钾	metalic shackle
JIAN	JY	樫	检	delete char. 1	uncommon
JING	D6=	刭	劲	write char. 1 as 倒	human activity
JUAN	PKL	睊	胃	delete char. 2	uncommon
LIAN	-Y	磏	裣	delete char. 1	uncommon
LING	KYR	呤	囹	delete char. 1	uncommon
MAO	HOP	瑁	氅	delete char. 2	uncommon
PANG	87	雱	旁	write char. 1 as 霈	that's original
QI	I	屺	岂	write char. 2 as	that's original
SHAO	2DK	邵	邵	delete char. 2	replaced by 劭
SI	I	厶	丝	write char. 1 as 私	that's original
XIAO	8EL	宵	霄	write char. 2 as 肴	being celestial
YI	F?	杙	栧	delete char. 1	uncommon
YI	?	乙	弋	delete char. 2	uncommon
YU	O	鱼	禺	spell char. 2 as OU	so is 偶
YUN	2K;	陨	郧	delete char. 2	uncommon
ZHANG	27	幛	障	write char. 1 as 缦	made of fabric
ZHEN	JPX	榛	葬斗	delete char. 1	uncommon
ZI	X.X	訾	吡	delete char. 1	replaced by 訾
ZI	;.X	赀	觜	delete char. 1	replaced by 资

relaxation because the complete code consists of the radical string as well as the spelling string. Over the ten thousand characters in 'XianDai HanYu CiDian', this coding scheme requires an average of 2.5 alphabets to spell a Chinese character.

The Radical Composition of Chinese Characters

One traditional method of looking up a Chinese character in a dictionary is first to identify a radical in the graphic representation of the character. There are hundreds of different standard radicals used in a dictionary and there are rigid rules to apply in order to identify one. The number of Chinese characters identified to a single radical is numerous. Even a combination of the spelling and the identifying radical together is not sufficient to yield a unique code for a Chinese character.

An experiment was conducted in which each of the ten thousand characters mentioned above was decomposed into a string of as many as eight radicals. In order to do so, a total of four hundred and fifty six radicals were employed. These radicals were grouped into fifty sets according to their common graphical properties. Each set is then associated with a key of a conventional keyboard. Table 3 lists all these radicals, their groupings and their associations with the keys of a keyboard. Human engineering aspects were considered when the set-key association was determined. The radical string for a Chinese character consists of the keys corresponding to the first three radicals composing the character. In case the character is decomposed into less than three radicals, blanks are used as fillers to make up a string of three keys. For instance, the character 待 is decomposed into 91T and the radical string for 彳 is I . In this coding scheme, the grouping of radicals into sets is of paramount importance. On the one hand, they are grouped according to their common graphic properties into as few sets as possible. On the other hand, care is exercised to assure the uniqueness (or almost uniqueness) of the code-character correspondence.

The Coding Scheme for Internal Representation

For data processing purposes, it is necessary to arrange the Chinese characters into a collating sequence which is a direct result of their internal representation in computer memory. Hence, when one is designing the internal codes, besides minimizing the length of the codes, one should also observe that the collating sequence that follows is logical and practical. This paper attempts to derive the internal codes logically from the input codes which, in turn, are logically related to the spellings and graphical properties of the Chinese characters. When a new character is created in the future with a unique input code, this scheme guarantees that the internal code will also be unique and a logical place in the collating sequence for it is assured.

The maximum number of keys used for an input code is seven. Storing seven symbols, in general, requires seven bytes. We recall that three symbols out of the seven serve to indicate which sets of radicals the Chinese character is composed of. Since there are fifty sets of radicals altogether, there are a total of 125,000 possible combinations. Seventeen bits will be sufficient to represent these combinations. The remaining four alphabetic symbols used to represent the spelling have the following properties:- The first symbol can be any alphabet from A to Z (except V). Five bits would suffice to represent it. The second symbol can be a blank, A, E, H, I, M, N, O, R, U or V, a total of eleven possibilities. Four bits would suffice. The third symbol can be a blank, A, E, G, I, N, O, or U, a total of eight possibilities. Three bits would suffice. The fourth symbol can be a blank, A, G, I, N, O, or U, a total of seven possibilities. Three bits would suffice.

Thus the spelling can be packed into fifteen bits. Combining with the seventeen bits required for the radicals, a code in these scheme requires only thirty two bits of memory space.

As a consequence of this internal representation, the collating sequence would be such that where a character should appear in the sequence first depends on the spelling of the character. The order of two characters of the same spelling depends on the keys used in the radical strings for the two characters.

Table 2: Conflicts Introduced by not Differentiating 'C' from 'CH', 'S' from 'SH' and 'Z' from 'ZH' as Leading Symbols in Spelling Chinese Characters

Spelling	Radical Composition	Char. 1	Char. 2	Suggested Remedy	Justification
CU/CHU	72Y	篨	卒	delete char. 1	uncommon
SA/SHA	7M	飒	箑	delete char. 2	uncommon
SI/SHI	G2	什	佀	delete char. 1	replaced by 似

Table 3: Grouping of Radicals

Key	Radicals in Sets
A	钅(金) 𠂉 乍
B	山 心 凶 [illegible] 去 屯 凵
C	马
D	刀 (刂) 𠃌 [illegible] 力 刃 办 为 与 [illegible] 与 专 コ 乛
E	小 ⺌ 瓜 爪 水 米 氺
F	木 术
G	亻 [illegible] 住 [illegible] [illegible] [illegible] [illegible]
H	王 圭 主 玉 丰 手 毛 生 丰 非 韭 [illegible] 耒 [illegible] 丟 奉
I	纟 幺 糸 丝 厶 巛 夕 乡 [illegible]
J	才 寸 扌 [illegible] 朿 斗 求
K	口 囗 四 凸
L	月 [illegible] 丗
M	儿 几 兀 九 丸 尢 尤 [illegible] 宂 心 [illegible] 见 [illegible] 风 龙 光 兆 兜 [illegible] [illegible] 走 足 是 [illegible]
N	冂 门 匚 [illegible] 凵 戊 戌 几 [illegible] 戕 門
O	田 由 甲 申 电 [illegible] 卑 [illegible] 里 果 早 毌 [illegible] 曲 [illegible] 禺 更
P	目 罒 皿 自 旦 且 其 身 鼎
Q	衤 [illegible] [illegible] 氏 民 長 艮 [illegible] 农 长
R	火 冫 丶 ㄑ 氺 小
S	饣 (食) [illegible]
T	女 乂 叉 丈 夂 夕 [illegible] 夂 [illegible] 支 父 [illegible] 攵 史 吏 叓
U	氵

Key	Radicals in Sets
V	鸟 [illegible] 乌 弓 弔 [illegible] 弗
W	丬 [illegible] 豕 [illegible] 习 羽 [illegible] 勿 彡
X	[illegible] 乚 七 北
Y	人 八 [illegible] 今 余 分 侖 [illegible] [illegible] 䒑 兼
Z	之 巳 巴 己 已 [illegible] [illegible] 也 [illegible] 甩
1	丨 忄 卜 丬 [illegible] [illegible] [illegible] 州
2	十 古 卉 卓 卩 [illegible] [illegible] [illegible] [illegible] [illegible] 巾 半 市 书 [illegible] [illegible] 韦 卌 [illegible]
3	士 吉 喜 [illegible] 生
4	疒 [illegible]
5	厂 𠂇 广 [illegible] [illegible] 尸 户 [illegible] [illegible] [illegible] [illegible] 手 卢 虍 厌 [illegible]
6	大 太 犬 夭 天 [illegible] [illegible] 久 矢 夬 [illegible] 头 [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] 莫 [illegible] 夹 [illegible] [illegible] [illegible] 入 央 夹 夷
7	⺮ 亠 立 方 亥 言 章
8	[illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] 六 [illegible] [illegible] [illegible] [illegible] 興 [illegible] [illegible]
9	艹 廾 [illegible] 廿 甘 [illegible] [illegible] [illegible] [illegible] 开
0	日 毋 毌
-	一 丁 丂 不 石 歹 [illegible] 可 匡 匽 [illegible] [illegible] [illegible] 丽
=	工 干 亍 于 五 云 [illegible] 耳 平 [illegible] 豆 至 亞
乛	乛 龴 了 子 孑 [illegible] 乃 及 予 [illegible] 承 皮 甬
!	中 四 串 革 堇 事 東 禹

Key	Radicals in Sets
;	贝 用 冉 肉 内 甫 舟 骨
:	彳 亻
'	而 丙 两 面 鬲 酉 百 首 页 雨 覀
"	灬 黑
,	丿 禾 舌 重 垂 熏 手 白 自 臼 牛
<	车 朱 东 七
.	虫 止 业 业 丑 [illegible] [illegible] 业
>	子 书 聿 隶 肃 肃 尹 尹 士
/	讠(言) 乚 己
?	乙 弋 七 戈 戋 戋 弋 戈 气 电 飞 我 幾 戋 (乙)

The Next Step

In order to evaluate the effectiveness of this coding scheme, the author plans to experiment with different users and measure their coding efficiencies as a function of training and experience as well as their reaction towards using this scheme. The acceptance of the users is the ultimate measure of success of any invention. The design of the set-key association in Table 3 is somewhat arbitrary. Since it has a subtle impact on the collating sequence, more research in this area is necessary.

Acknowledgement

The author is indebted to Professor T.C. Chen for his constructive suggestions and criticisms. The author is also grateful to Mr. T.H. Tse for his assistance and discussions.