Unicode Han Character Lookup Service Based on Similar Radicals*

Jeng-Wei Lin[1] and Feng-Sheng Lin[2]

[1]Department of Information Management, Tunghai University
[2]Institute of Information Science, Academia Sinica
jwlin@thu.edu.tw, skyrain@iis.sinica.edu.tw

Abstract

Unicode 6.1 (2012) had encoded more than 74,000 Han characters. This great repertory could solve the problem of unencoded Han characters to a significant extent. However, most information systems today still only support input and display of the first 20,902 encoded Han characters in Unicode 1.0 (1991). Even in latest systems, designed to support 32-bit Unicode and with suitable fonts installed, it is not easy to use these newly encoded Han characters. We note that many of these newly encoded Han characters are rarely used in users' everyday life. An ordinary user may have confusions of their glyph shapes, pronunciations, meanings, and usages. IMEs (input method editors) for Han characters usually require users to have good knowledge of wanted Han characters. It is not unusual users try but fail to input unfamiliar Han characters. In this paper, we present an auxiliary Unicode Han character lookup service by radicals. One can use any Han character IME to key in one or more radicals to look up a wanted Han character. Every Unicode Han character is decomposed as a glyph expression of radicals. The similarity between the glyph expression and user input is estimated by a derived edit distance algorithm. The most similar Unicode Han characters are returned. As a result, the system provides users a convenient way to look up unfamiliar Unicode Han characters.

Keywords: *Unicode; Han character lookup; glyph expression; radicals; edit distance*

1. Introduction

In 1991, the first version of Unicode, Unicode 1.0, was released. It had encoded 20,902 Han characters used in Traditional and Simplified Chinese, Japanese, and Korean in the basic block - CJK Unified Ideographs [1]-[4]. These Han characters were encoded previously in different local charset encodings, such as CNS-11643 [5], BIG5 [6], GB-2313 [7], JIS X 0208 [8], KS C 5601 [9], and so on, and used in computer systems in different East Asia countries. We note that Han characters are commonly referred to as Hanzi in Chinese, Kanji in Japanese, and Hanja in Korean. The abbreviation CJKV is also frequently used since the writing systems of earlier Vietnamese are based on Han characters. From then on, as new Unicode versions were released, many more Han characters were encoded in batch in different CJK Unified Ideographs Extensions A to D, as shown in Table 1. Unicode 6.1, recently published in 2012, had totally encoded more than 74,000 Han characters, including characters listed in many important dictionaries, such as Kang-Xi Dictionary [10]. The number of encoded Unicode Han characters has been increased significantly. We note that due to the composition rules of Han characters, it is impossible to encode all Han characters. Many Han characters are still not encoded in Unicode. As well, new Han characters are coined now and forever. Nevertheless, this great repertory could solve the problem of unencoded Han characters to a significant extent. However, it is not easy to use these newly encoded Han characters, even in

* This work was supported in part by National Science Council, Taiwan (R.O.C.), under Grant NSC 99-2221-E-029-035 and NSC 101-2631-H-001-006.

a latest computer system that supports 32-bit Unicode and installs suitable fonts.

The problem of unencoded Han character occurs when a user wants to use a Han character in a computer system, but the wanted Han character is not encoded in the charset encodings adopted by the computer system, or more specifically, operating system or application software. As a result, the user cannot use the wanted Han character for information processing. The user usually has to 1) replace the wanted Han character by one of its encoded character variants, which has the same (or very similar) meaning and pronunciation, but has a different glyph shape [11]-[13], 2) use a digitized image of the wanted Han character, or 3) use a private code point associating with a self-made font. Using the third approach, different communities had created many private and incompatible extensions of Han character charset encodings. For example, HKSCS [14], for Cantonese, is an extension of BIG5, developed by Hong Kong Government. In particular application software, there also exist sophisticated patches or plug-ins for users to search and use some unencoded Han characters [15][16]. We note that a Han character unencoded in one charset encoding is possibly encoded in another charset encoding. Many Simplified Chinese characters unencoded in BIG5 are encoded in GB-2312, and vice versa. It is possible that an unencoded Han character at this moment will be encoded in the future.

Earlier systems typically only support 2-byte Unicode, and thus the use of Han characters encoded in Unicode CJK Unified Ideographs and Extension A. Thus, people of these systems cannot use Han characters encoded in other CJK Unified Ideographs Extensions, i.e., B to D. In fact, in most systems, only the fonts of the initial 20,902 Han characters are installed. New systems, such as Microsoft Windows 7 and Linux, support 4-byte Unicode and install suitable fonts. People of these new systems can use all encoded Unicode Han characters. However, it is still not easy for ordinary users to use these newly encoded Han characters in CJK Unified Ideographs Extensions A to D.

Since the number of encoded Han characters is very large, people usually have to use a Han character IME (Input Method Editor) to input Han characters. A user typically keys into a Han character IME a sequence of ASCII alphabets and digits, which associates a certain combination of the radicals, pronunciations, meanings, and other properties of a wanted Han character. If the user has confusions about the wanted Han character and cannot key into the IME the correct sequence, the user will fail to input the wanted Han character. In fact, most Han character IMEs only well support the input of the initial 20,902 Han characters when this paper was written. If the user tries several IMEs and all fail, the user may finally doubt whether the wanted Han character is encoded in Unicode, even though the wanted is indeed encoded.

As described above, users of a Han character IME should have good knowledge of Han characters. The sentence seems trivial. However, an ordinary user typically knows 6,000 or so Han characters. This is a small portion of the huge set of Han characters. In fact, most of the Han characters newly encoded in CJK Unification Ideographs Extensions A to D are rarely

Table 1. Han Characters Encoded in Unicode 6.1

Block	Version	Year	Range of Codepoint	Number of Characters	Comment
CJK Unified Ideographs	1.0	1991	04E00-09FA5	20,902	common
	4.1~		09FA6-09FCC	39	
Extension A	3.0	1999	03400-04DB5	6,582	rare, historic, some in current use
Extension B	3.1	2001	20000-2A6D6	42,711	
Extension C	5.2	2009	2A700-2B734	4,149	
Extension D	6.0	2010	2B740-2B81D	222	
Compatibility			0F900-0FAD9	474	duplicates, unifiable variants
Compatibility Supplement			2F800-2FA1D	542	

used in everyday life, as shown in Table 1. In other words, it is very likely that an ordinary user is unfamiliar with these newly encoded Han characters. As a result, the user cannot input these Han characters even if they are indeed encoded in Unicode.

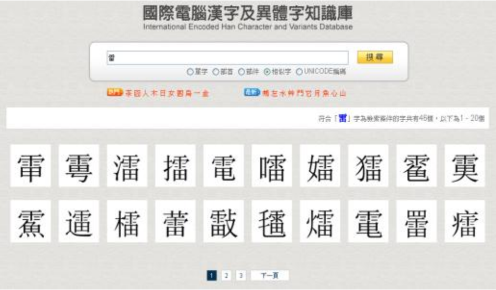

國際電腦漢字及異體字知識庫
International Encoded Han Character and Variants Database
雷
搜尋
○單字 ○部首 ○部件 ●相似字 ○UNICODE編碼
符合「雷」字為檢索條件的字共有45個，以下為1 - 20個
1 2 3 下一頁

Figure 1. Similar Han character lookup by 雷

Saturations could be worse. Even when a user keys into the IME a correct sequence of a Han character, the displayed glyph shape is probably one of its character variants and differs from the user’s expectation. We note that a Han character may several character variants [11], 12]. Due to the Principles of Unicode Unification for CJK Ideographs [1], the character variants of a Han character may 1) be unencoded, 2) share the same code points, or 3) have been assigned to different code points. In the second case, the glyph shape shown on papers or screens depends on the installed fonts. For example, 吳 (U+5433[1]), 吴 (U+5434), and 呉 (U+5449) are encoded as three different code points, while 祦 and 祦 shares a same code point U+7966. Similarly, 溫 (U+6EAB) and 温 (U+6E29) have different code points, while 蝹 and 蝹 share a same code point U+8779. The use of Han character variants may have several side-effects in information processing [13]. However, a user may sometimes prefer the use of a particular variant, for example, when a sinologist is writing an article talking about Han characters. If the user does want to use a particular variant and cannot accept the displayed variant, the user may falsely consider the wanted variant is not encoded. As a result, even for a sinologist, it is not easy to use these newly encoded Unicode Han characters.

Han character IMEs typically emphasize a high precision rate of a Han character lookup by a short sequence. In contrast to IMEs, a Han character lookup service for rarely used Han characters should have a higher recall so that users can find the wanted Han character easily. In International Encoded Han Character and Variants Database [11], we had implemented an auxiliary Unicode Han character lookup service, in which a user can simply input a similar Han character by any IME to look up a Han character [17]. For example, people can use 雷 (U+96F7) to look up □ (U+29098), □ (U+290BE), and □ (U+290AC), as shown in Figure 1. We use a glyph expression to describe the glyph shape of a Han character [15][16]. The similarity of two Han characters is estimated by the edit distance [18] between the two glyph expressions, i.e., the minimal cost to transform from one glyph expression into another via insertions, deletions, and replacements of radicals.

Sometimes users may prefer to look up Han characters via one or more similar radicals. For example, people may also want to use “雨甲”, “雷　”, and “雷七” to look the above three Han characters □, □, and □, respectively. To look up some singular Han characters, such as □ (U+21542), however, it is hard for users to guess a similar Han character. On the other hand, the user may try “士四亞凶” or “士亞亞凶” to look up □. In this paper, we

[1] In this article, the Unicode characters are identified by their positions, or code points. The notation U+12AB, for example, indicates the character at the position 12AB (hexadecimal) in the Unicode table. The code point of a Unicode character appears right after the first appearance of the character or in the tables and figures.

	Reduced Glyph Expression
說	言八口儿
说	口儿
説	言 口儿

	Reduced Glyph Expression
壺	士 一
壼	士 一 一
壷	士

Figure 2: Some examples of reduced glyph expressions

Root Radical 1	Root Radical 2	Total Strokes	Similarity
厂		4	4
厂		4	4
厂	尸	5	3

Figure 3: Some examples of similarities between root radicals

present an extended service that supports Han character lookups via several similar radicals in International Encoded Han Character and Variants Database.

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

2. The Method to Measure the Similarity

Based on the successful experience in our previous study [17], we again use the edit distance algorithm to estimate the similarity of a Han character and a sequence of radicals specified by the user. In this section, we briefly review the result of our previous study. Next, we discuss the difference between the two Han character lookup services, one via a similar character and another via a sequence of similar radicals. Then, we present how to look up Han characters via similar radicals.

2.1. Glyph Expressions

In general, a Han character is either an atomic glyph component, referred to as a root radical, or composed of several root radicals. For example, Han character 說 consists of two components, 言 and 兌. It can be further decomposed into four root radicals, 言, 八, 口, and 儿. We use "言八口儿" as the reduced glyph expression of 說.

In [17] we had decomposed all Unicode Han characters into reduced glyph expressions according to Chinese Glyph Structure Database [15]. Figure 2 shows more of them. Totally, there are 1,151 root radicals in use. In addition, we had built a similarity table for each pair of root radicals, as shown in Figure 3. We hired several students to manually label the similarity between each pair of root radicals. The more similar two root radicals, the higher similarity.

2.2. Similarity of Two Han characters

In [17], we estimate the similarity of glyph shapes of two Han characters by the edit distance [18] 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. Three edit operations are used: 1) insertion of a root radical, 2) deletion of a root radical, and 3) replacement from a root radical with another. We note that the edit distance between two reduced glyph expressions can be easily computed by a dynamic program algorithm. As

馬	馬
	↓i ↓
碼	石 馬

棋	木 其
	↓d ↓
其	其

林	木 木
	↓ ↓r
村	木 寸

釘	金 丁
	↓d ↓i ↓r
計	言 十

Figure 4: Examples of transformation between reduced glyph expressions. Labels i, d, and r beside ↓ denote insertion, deletion, and replacement

大	大	
	↓	↓i
太	大	

大	大	
	↓	↓i
夸	大	

Figure 5: Examples of edit distances

鍊	金	柬
	↓	↓r
鍊	金	東

or

鍊	金	柬	
	↓	↓d	↓i
鍊	金		東

Figure 6: Examples of edit distances

shown in Figure 4, we can transform one Han character into another. Thus,

Cost(馬→碼) = Cost$_i$(石),

Cost(棋→其) = Cost$_d$(木),

Cost(林→村) = Cost$_r$(木, 寸),

and

Cost(釘→計) = Cost$_d$(金) + Cost$_i$(言) + Cost$_r$(丁, 十).

It is obviously that 大 and 太 are more similar than 大 and 夸, as shown in Figure 5. Thus, the cost of insertion or deletion of a root radical is proportional to the number of strokes in the root radical.

However, if we transform 鍊 into 鍊 by first deleting 柬 and then inserting 東, the cost is high. In this case, the preferred transformation is replacement from 柬 with 東, as shown in Figure 6. In other words, when two root radicals are similar, replacement edit operation is preferred. As a result, if two root radicals A and B are similar, inequality (1) must hold.

Cost$_r$(A, B) < Cost$_d$(A) + Cost$_i$(B) (1)

Otherwise, equation (2) holds so that a replacement from A with B is not different from the combination of a deletion of A and an insertion of B.

Cost$_r$(A, B) = Cost$_d$(A) + Cost$_i$(B) (2)

2.3. Han Character Lookup via Radicals

When a user wants to look up an uncertain Han character via a similar one, the user can choose any Han character IME to input the similar one. However, situation is different in Han character lookup via radicals. In fact, many root radicals are unencoded in Unicode, such as , , and so on. In most information systems, people cannot input these unencoded root radicals. As a result, people have to use similar radicals. For example, users may use 田 and 臣 as a replacement of and , respectively.

We also notice that among the 1,151 root radicals, there are several groups of related root radicals, such as the group of 馬, , , , and . However, in most systems, people can only input 馬.

If we do not insist that every root radicals should be correct and complete in the view point of sinology, we can furthermore decompose these root radicals, as shown in Figure 7. Thus, the systems only has to recognize that and are similar.

Many other root radicals can be furthermore decomposed in the same way. For example, we can furthermore decompose into 口 一, and into 一, as shown in Figure 7. Thus, we have reviewed all 1,151 root radicals again. We have furthermore decomposed root radicals, if suitable, and constructed a smaller set of basic radicals. Finally, there remain about 300 basic radicals. Most of these basic radicals have less than 7 strokes.

馬	
	一
…	…

	口　一
	一
	一

聿	
	巾
	土
疌	十

	禾　戈
更	田乂
革	廿中十
	口止

Figure 7: Furthermore decomposition of root radicals without consideration of correctness and completeness in the view point of sinology.

3. UNICODE HAN CHARACTER LOOKUP SERVICE

3.1. Similarity Between Basic Radicals

Since we have furthermore decomposed many root radicals into smaller basic radicals, the similarities between all basic radicals have been revised. In this study, the similarity of two basic radicals A and B, Sim(A, B), is labeled from 0 to the sum of strokes of A and B, as shown in (3), where 0 indicates they are not considered as similar at all, and a larger value indicates that they are very similar.

$$0 \leq \mathrm{Sim}(A, B) \leq \mathrm{Strokes}(A) + \mathrm{Strokes}(B) \qquad (3)$$

3.2. Design of Cost Functions

In this study, the cost of insertion or deletion of a basic radical A is defined as its number of strokes, as shown in (4).

$$\mathrm{Cost}_i(A) = \mathrm{Cost}_d(A) = \mathrm{Strokes}(A) \qquad (4)$$

The cost to replace a basic radical A with B is defined as (5).

$$\mathrm{Cost}_r(A, B) = \mathrm{Strokes}(A) + \mathrm{Strokes}(B) - \mathrm{Sim}(A, B) \qquad (5)$$

In other words,

$$\mathrm{Cost}_r(A, B) = \mathrm{Cost}_d(A) + \mathrm{Cost}_i(B) - \mathrm{Sim}(A, B) \qquad (6)$$

If Sim(A, B) is 0, i.e., A and B are considered as very dissimilar, the cost to replace A with B equals to the cost of a deletion of A and an insertion of B. If Sim(A, B) is larger than 0, $\mathrm{Cost}_r(A, B)$ is smaller than $\mathrm{Cost}_d(A)+\mathrm{Cost}_i(B)$. Thus, the edit distance algorithm will prefer to use the replacement operation. This definition obeys (1) and (2).

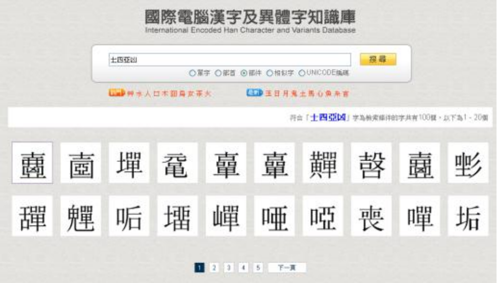
3.3. Implementation

Since many root radicals have been furthermore decomposed, all Unicode Han characters are decomposed into their reduced glyph expressions of basic radicals. For example, the glyph expression of □, 士　凶, is furthermore decomposed into 士　　一凵丨乂.

When a user inputs a query sequence of Han radicals, the sequence is also decomposed into a reduced sequence of basic radicals. We note that since not all Han radicals are encoded in Unicode, the user usually input similar Han radicals. For example, if a user inputs 士四亞凶, the sequence is furthermore decomposed into 士四一　一凵丨乂. As shown in Figure 8, the system can estimates the similarity between the input sequence and a Unicode Han character.

士四亞凶	士	四	一		一	凵	乂
	↓	↓r	↓d	↓	↓	↓	↓
□	士				一	凵	乂

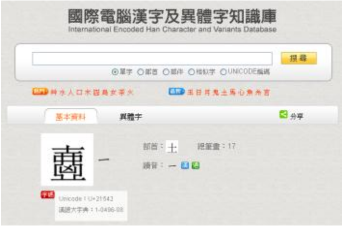
Figure 8: Edit distance between an input sequence 士四亞凶 and a Han character𡕃.


國際電腦漢字及異體字知識庫
International Encoded Han Character and Variants Database

士四亞凶 搜尋
○單字 ○部首 ● 部件 ○相似字 ○UNICODE編碼

符合「士四亞凶」字為檢索條件的字共有100個，以下為1 - 20個

Figure 9: Han character lookup via similar radicals "士四亞凶".


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

𡕃 一
部首：士 總筆畫：17
讀音：一
Unicode：U+21543
漢語大字典：1-0496-08

Figure 10: More information about the returned Han character.

The lookup service computes the edit distance between the reduced input sequence and the reduced glyph expressions for all Unicode Han characters. It returns the top 100 similar Unicode Han characters. We have integrated this service into International Encoded Han Character and Variants Database [11]. Figure 9 demonstrates the result of lookup via radicals 士四亞凶. The user can further look up a Han character shown on the returned page for more information, as shown in Figure 10. Thus, the user can make sure which one is wanted and its Unicode codepoint.

4. Conclusion

At the time this paper was written, Unicode had encoded more than 74,000 Han characters in its repertory. It is expectedly that the number will increase to more than 100,000 in the near future. However, it is not easy for ordinary users to use these newly encoded Han characters in computers. A reason is that people do not know how to input these Han characters into computers. Traditional IMEs usually require users to have good knowledge of a wanted Han character. However, most of these newly encoded Han characters are rarely used in daily life.

Based on our previous study, in this paper, we present an extended Unicode Han character lookup service via one or more similar radicals. In contrast to IMEs, the Han character lookup service for rarely used Han characters should have a higher recall so that users can find a wanted Han character easily. In the extended Unicode Han character lookup service, users can use any IME to input a sequence of several encoded radicals, each of which is similar to one part of a wanted Han character. It is not necessary for the user to specify all parts of the wanted. The similarity of the user input and each Unicode Han character is estimated by the edit distance of their corresponding reduced glyph expressions. Unicode Han characters similar to the user input are returned. The service is integrated in International Encoded Han Character and Variants Database. As shown in Figure 9 and 10, it provides users a convenient way to look up a wanted but unfamiliar Han character. Currently, we had not yet considered the relative position between radicals. 加 and 召 are recognized as similar. However, 鵝 and 䳘 are recognized as dissimilar. We will further study these problems in the future.

References

[1] The Unicode Standard, version 6.1. **(2012)**. Available at 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] K. Lunde, CJKV Information Processing, 2nd edition, O' Reilly **(1999)**, ISBN ISBN 978-0-596-51447-1.
[4] Unihan Database, http://www.unicode.org/charts/unihan.html.
[5] Chinese Standard Interchange Code, Chinese National Standard, CNS 11643-1992, **(1992)**.
[6] BIG5-1984, usually simply BIG5, a de facto standard for Traditional Chinese character **(1984)**, standardized in appendix of CNS-11643 expansion **(2003)**, 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] Z. Yushu, C. Tingjing, et. al., Kang Xi Dictionary, **(1716)**, Zhonghua Bookstore **(1989)**, ISBN 962-231-006-0.
[11] International Encoded Han Character and Variants Database, http://chardb.iis.sinica.edu.tw/.
[12] K. Yuen, J. -W. Lin, J. -H. Chan and Y. -S. Zhao, "The Study of Unicode Han Characters and Their Variants -- with Dictionary", **(2011)**, ISBN 978-986-02-6986-4.
[13] J. -W. Lin, J. -M. Ho, L. -M. Tseng and F. Lai, "Variant Chinese Domain Name Resolution", ACM Transactions on Asian Language Information Processing, vol. 7, no. 4, **(2008)**.
[14] Hong Kong Supplementary Character Set, simply HKSCS, **(1999)**.
[15] 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, **(2005)** Denver, Colorado, USA.
[16] Chinese Glyph Structure Database. Available at http://cdp.sinica.edu.tw/cdphanzi/.
[17] J. -W. Lin and F. -S. Lin, "An Auxiliary Unicode Han Character Lookup Service Based on Glyph Shape Similarity", 11th IEEE International Symposium on Communications & Information Technologies, **(2011)** Hangzhou, China.
[18] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", Soviet Physics Doklady, 10:707–710 **(1966)**.

Authors

Jeng-Wei Lin received the B.S. degree in Computer Information Science from National Chiao Tung University, Hsinchu, Taiwan, in 1994, and the M.S. and Ph.D. degrees in Computer Science and Information Engineering from National Taiwan University, Taipei, Taiwan, in 1996 and 2005, respectively. Since 2005, he joined the Department of Information Management, Tunghai University, Taichung, Taiwan. From 1996 to 2005, he was a research assistant with the Institute of Information Science, Academia Sinica, Taipei, Taiwan. His current research interests include multimedia systems, P2P networking, cloud computing, and Chinese information processing.

Feng-Sheng Lin received the B.S. and M.S. degrees in Computer Science and Information Engineering from National Taiwan University, Taipei, Taiwan, in 2009 and 2011, respectively. Then he joined the National Digital Archive Program as a research assistant in Institute of Information Science, Academia Sinica, Taipei, Taiwan. His research interests include multimedia streaming, bio-informatics, and Chinese information processing.