NaXi Pictographs Information Processing System

Guo Hai, Zhao Jing-ying*
Department of Computer Science and Engineering
DaLian Nationalities University, DaLian 116600
China
guohai@dlnu.edu.cn

Abstract: - This Paper describes a NaXi pictographs information processing system we developed for different applications. It sets a basis for computerization of NaXi pictographs. Basic NaXi pictographs information processing Font, IME (Input Method Editors) and corresponding application software have been realized. This sets an end to the history that the NaXi pictographs can't be handled with computers. It will greatly promote the computerization process of the minority languages.

Key-Words: - NaXi pictographs; information processing; outlines font; IME (Input Method Editors); embedded font; WEFT

1 Introduction

NaXi pictograph belongs to the Naxi language of yi language branch of Tibetan-Burmese languages which was created by the ancestor of Naxi. It's credited as "the only living ancient hieroglyph" and still being used in writing lection and composition and in the field of communication, therefore, Naxi pictograph has a special role in the history of human character. Since hieroglyph appeared in the early stage of the development of characters, through a deep study of it we can get something about the evolutionary history of human characters and human culture which will make a great contribution to research on the origin of modern characters. Many experts and scholars at home and aboard have have been working on Naxi pictograph for a long time, among which Harvard and Yunnan Academy of Social Sciences lead a dominant role. But the problem that a lot of literature and ancient books can't be processed efficiently makes it urgent to realizing an information processing system for Naxi pictograph.

The traditional hand-drawing method of processing Naxi pictograph has low efficiency and can't guarantee the stability and standard of publication printing, while the problem can be solved by the output of computer's standard font, therefore, out project designs and develops a Naxi pictograph information processing system for the first time which ends the processing history of Naxi pictograph without computerization.

2 The Principle of Naxi Pictograph Information Processing

2.1 The principle of Naxi Pictograph Outline Fonts

Since outline font can be zoomed in and out at will without distortion, it is widely applied in printing, typesetting, character processing and art creating, and Windows we uses everyday is displayed with outline font too. This type of font is composed of Bézier curve, and the TrueType font uses quadratic Bézier curve, the PostScript font uses cubic Bézier curve, so the reducibility of PostScript font is better than the TrueType font. A Bézier curve is controlled by three points, as is shown in Figure1, when the middle control point changes the whole curve changes too. If an outline word consists of n+1 points (P0, P1 … Pn), it needs $\sum_{i=0}^{n}\binom{i}{n}t^i(1-t)^{n-i}P_i$ Bézier curves to form the font.

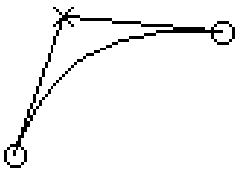

Fig.1 Bézier Curve

A PostScript outline is formed of a group of cubic Bézier curves which can be described in the following formula:

* Corresponding author: Jing-ying Zhao Email: dalianzjy@gmail.com

$$x = a_X * t^3 + b_X * t^2 + c_X * t + d_X \quad (1)$$

$$y = a_y * t^3 + b_y * t^2 + c_y * t + d_y \quad (2)$$

A TrueType outline is formed of a group of quadratic Bézier curves and every curve is defined by three control points. For a quadratic Bézier curve's three control points are (Ax, Ay),（Bx, By）and（Cx, Cy）, then:

$$P_X = (1-t)^2 A_X + 2t(1-t)B_X + t^2 C_X \quad (3)$$

$$P_y = (1-t)^2 A_y + 2t(1-t)B_y + t^2 C_y \quad (4)$$

By modifying the parameter t from 0 to 1, all values of p defined by A, B and C can be generated, from which the quadratic Bézier curve is obtained.

Font is defined as a combination of a set of outline curves and every curve includes three or more points. The illustration using Bézier outline curve of Naxi pictograph "elephant" is shown in Fingure2. The outline is constituted of a set of instructions which pictured the character's exterior.

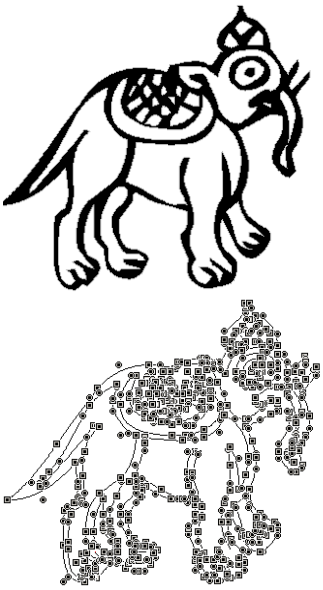

Fig.2 the Outline Curve delineation of Naxi Pictograph "elephant"

All the parameter information of outline fonts is stored in a Naxi pictograph table "glyf". Because of the high complexity and similarity of Naxi pictograph, the precision of outline curve delineation is greatly higher than the dot matrix font delineation. Based on the two outline font technique, we developed Naxi pictograph TrueType font and PostScript font which basically met the requirement of processing basic model of Naxi pictograph with computer.

2.2 The Feature Extraction Method of Naxi Pictograph Outline Font

The outline font is filled with a series of outline curves and delineating the outline accurately is the key point for the success of a font. It can be seen from Figure3 that the Naxi pictograph is complicated, various and also with many strokes which make it difficult for truly reducing the outline of Naxi pictograph. According to these characteristics, our project proposes a dual-mode transformation algorithm for extracting the Naxi pictograph outline points and uses some rules to check whether a pixel point is in the outline.

Rule1: if the center of a pixel point is in the outline, the point is lightened and becomes a part of the outline curve.

Rule2: if an outline line passes the center of a pixel, the pixel is lightened.

Rule3: if a scanning line located in the center of two adjacent pixels (horizontal or vertical) intersects with the on-Transition and the off-Transition and the points on this line haven't lightened by rule1 and rule2, the left endpoint is lightened when this line is horizontal, while the right endpoint is lighted when it's vertical.

Rule4: rule3 can be used only in the case that two contour surfaces still intersect the scanning line in two-way, but this doesn't mean these pixels are "stubs". By checking the scanning line-segment formed a square with crossing scan line-segment, whether they intersect each other through two contour surfaces can be verified. There is the possibility which is small but exists that more than one contour intersect with discontinuity point, so it's necessary to control some characters outline using grid-fitting.

The method discussed above is used in the secondary development of Pgaedit based on Linux which effectively prevents the generation of discontinuity point, and the accuracy of outline point extraction of Naxi pictograph reaches 99.99%

3 The design Goal of Naxi Pictograph Information Processing Platform

The Naxi Pictograph processing platform is developed for special groups, including units of printing and publication, Naxi pictograph research institutes and art designing companies, so it's not a universal software platform. These reasons make it definite that our development must meet the requirement of different customers. The analysis of application level of this project can be summarized as Table1.

According to the analysis above, our project has designed various Naxi pictograph processing

systems which can meet specific requirement of different customers. For professional customer doing publication and printing, we developed Naxi pictograph standard edition with TrueType and PostScript outline fonts, and three input methods including internal code input method, Latin input method and English input method, which can fully satisfy the needs of printing and typesetting. As for the Naxi pictograph research institutes, we developed Naxi pictograph TrueType outline font and Naxi Pinyin input method. Considering that the artistic designing units just need the font style outline, their system only have Naxi pictograph TrueType outline fonts and Naxi pictograph English input method. At last, we developed web embedding fonts and PDF document creating technology of Naxi pictograph for ordinary customers.

Table1 the Analysis of Customer Requirement of Naxi Pictograph Information Platform

customer group	customer requirement
Printing Publication	high accuracy, fast input, output, various convenient input method
research institutes	general output accuracy, advanced input method
artistic designing	various of output font styles, input method easily to learned
ordinary user	offer web pages and electronic document browsing

4 The Font Library

Font library is the basis and also the core of the system, accurate font and correct code set solid foundation for developing the complete system. The Naxi pictograph font library includes Naxi pictograph TrueType outline font and Naxi pictograph PostScript font. TrueType outline font is composed of quadratic Bézier curves and PostScript is built with cubic Bézier curves. Since a quadratic Bézier curve can be transformed to a cubic Bézier curve, we develop TrueType outline font first and then realize the transformation by expert tools.

The outline font development can be divided into four steps: font and script designing, digital fitting, modifying and font generating. Our project uses the original script of A dictionary of Naxi pictograph sound-indication which is transformed to standard bit map by scanning, and then the describing points are extracted for digital fitting, after that the Naxi pictograph outline font is integrated and created. This kind of technique can guarantee the veracity of original script and the generated outline font has the features of excellent reducibility and vector property as well. The Naxi pictograph outline font generated is shown in Figure3.

5 The Input Method System

5.1 The Input Method Principle

As Windows is the most widely used operation system at present, this paper focuses on the input system for Naxi pictograph based on Windows. The input method based on Windows transforms standard ASCII string into Naxi word or string using some particular coding rule. With different application program the user can not design a transformation program himself, due to which the task of inputting Naxi pictograph should be taken by the Windows system administration.

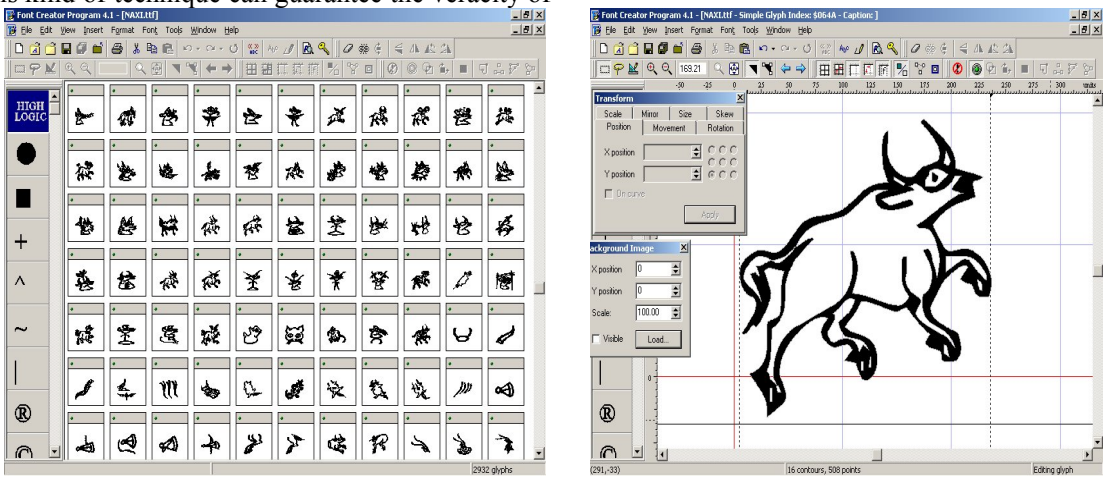

Fig.3 the Naxi Pictograph Outline Font

As shown in Figure4, at first, the keyboard event of Naxi pictograph input system is received by the Windows file use.exe, then use.exe transfers the event to the Input Method Manager (IMM), after that the IMM conveys the event to the input method editor which translates the keyboard event to its corresponding Naxi character (or string) with reference to user's encoding dictionary, when this is done the translated event is propagated back to use.exe and then to the executing application, until now the whole input process of Naxi pictograph is finished[4-5].

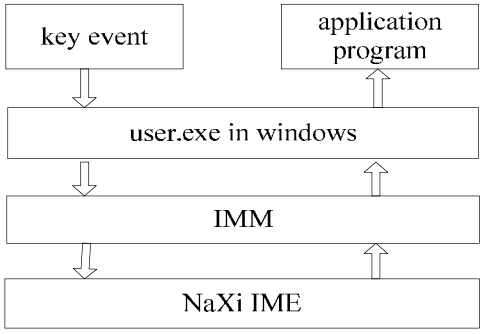

Fig.4 the Input Method Principle of NaXi Pictograph

5.2 The Scheme Design and Optimization of Pinyin Encoding

Current input method includes two types: Pinyin and font. The font input method can be divided by strokes which needs the user has the ability to write. Unfortunately, writing Naxi pictograph is so hard for the ordinary user that makes the shape code unsuitable for Naxi pictograph being inputted to the computer. The phonetic code input method just requires the user know how to read the character, therefore, Pinyin input method is more suitable for Naxi pictograph.

Early dictionaries of Naxi pictograph sound-indication employ International Phonetic Alphabet to mark Naxi character, while the computer uses Latin letters as input code, the conversion between Naxi Phonetic and Latin Pinyin becomes necessary. Table2 lists the mapping between International Phonetic Alphabet and Latin Pinyin in detail. When designing the input method of Naxi pictograph one can encounter the problem of too long encoding due to the characteristics of Naxi pronunciation. For instance, the Naxi character "" can be coded as ssoxiqssoddassa, it needs fifteen English characters to map this single one character which will result in very low efficiency when input an article. Research shows that the initials repetition phenomenon is common in Naxi pictograph pronounced coding, through the method of designing code with simplified initials the Pinyin input method of Naxi pictograph can greatly reduce the coding length and improve the coding efficiency. Let's take the character "" as an example, its pronounced coding is ssoxiqssoddassa, after the simplification of pronunciation it becomes to soxiqsodasa, the length reduces to eleven from fifteen. Naxi pictograph consists of 2120 characters and the average coding length is twelve bits which is reduced to eight bits after the simplification of pronunciation, so the input speed is also increased.

Table2 Naxi pinyin coding

IPA	Latin letter	IPA	Latin letter	IPA	Latin letter	IPA	Latin letter
p'	p	**p**	b	**b**	bb	**f(ɯ)**	f(w)
t	d	**t'**	t	**d**	dd	**n**	n
l	l	**k**	g	**k'**	k	**g**	gg
ŋ	ng	**h**	h	**ʧ**	**j**	**ʧ'**	q
dz̧	jj	**ȵ**	ni	**ɕ(ʑ)**	x(y)	**z̧**	r
ʂ	sh	**dz̧**	rh	**tʂ**	zh	**tʂ'**	ch
z	ss	**s**	s	**dz**	zz	**ts'**	c
ts	z	**i**	i	**u**	u	**y**	iu
a	a	**o**	o	**ə**	e	**v**	v
ɯ	ee	**ər**	er	**e**	ei	**æ**	ai
ie	iei	**iæ**	iai	**ia**	ia	**iə**	ie
uei	ui	**uæ**	uai	**ua**	ua	**uə**	ue

5.3 The Basic Structure of Naxi Pictograph IMM-IME

The IMM-IME structure provides various input methods for applications, each tread of an application can keep an active input window. The processing order of other messages won't be disturbed by inserting the Naxi pictograph message to message circle. The head file immdev.h should be included for using these new features. The detailed working principle of Naxi pictograph Pinyin input method is shown in Figure5.

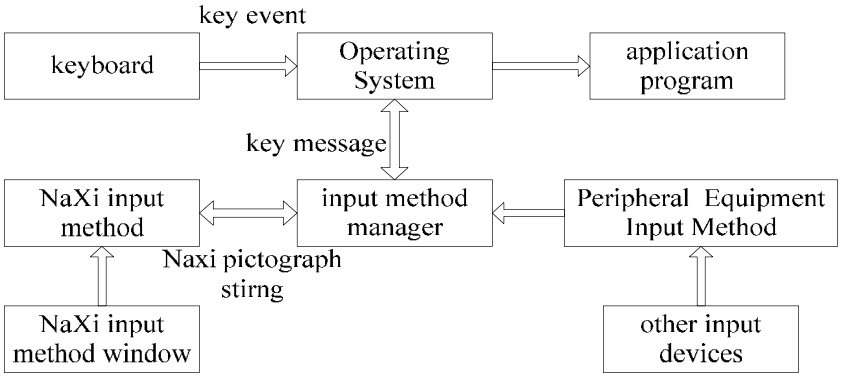

Fig.5 Working Principle of Naxi Pictograph Pinyin Input method

5.4 The Realization of Naxi Pictograph IME

The Naxi pictograph Input Method Editor (IME) is realized as a form of Dynamic Link Library (DLL) which is located under Windows System or Systems directory. The only difference from ordinary DLL is that the input method uses .IME as its suffix. IME need to offer two units: IME conversion interface and IME user interface. The former interface can be realized though a group of exported functions of IME model which is called by IMM, and the latter one through a group or windows which receive messages and also supply IME user interface.

IME includes IME conversion interface and user interface. The conversion interface is composed of several interface functions of which the concrete interface and detailed function is regulated by the rules of the development interface of IME. IMM completes the transform function by calling a corresponding interface function. IME user interface consists of a number of relevant users' windows which can receive and handle messages from IMM and also serve as interactive interface with users.

5.4.1 IME Conversion Interface Function Realization

IME Conversion Interface Function Realization

Of all the functions called by IMM, there are four important ones for Naxi pictograph Pinyin input method that should be realized with first consideration

Imelnquir: When a user selects the Naxi pictograph input method, this function will be called first by IMM to get relevant information about this input method. The function should return initial information of IME, set every attribute of the method under the IMEINFO structure and name the window class of user interface.

ImeConfigure: This function will be called by IMM when a user sets attributes of the input method through control panel and system icon. It can show attribute setup dialogue for user to set options of Naxi pictograph input methoud.

ImeProcessKey: This function will be called by IMM when a keyboard event needs to be handled. A keyboard event will be preprocessed by this function, then, according to returned value the system makes a decision with the consideration of specific context whether this event should be transferred to IME. If the returned value is true it means the keyboard message should be conveyed to the IME and ImeToAsciiEx will be called in a minute. On the contrary, if the returned value holds false it indicates there is no need for the IME to process the keyboard message, in this case, Naxi IMM directly transfers the message to the application.

ImeToAsciiEx: According to the context of Naxi input method, this function generates conversion result using conversion engine of IME and puts relevant character message to specified buffer area. Returned value is the number of messages, if this number is larger than the length of the buffer the system will turn to hMsgBuf item of the context of Naxi input method to read the message. This function together with function ImeProcessKey construct the main body of the conversion engine of IME based on keyboard input method.

5.4.2 IME User interface Realization.

The user interface of Naxi input method mainly includes user interface window (window class and window procedure), user interface window components (status window, writing widow, and candidate window of window class and window procedure), setup window of the input method, soft keyboard, indicator window of the taskbar (icon, mail, tool tips) and hot key of the input method. During the process of programming the user interface window of Naxi input method the main task is dealing with the IME messages from the default IME window, WM_IME_SETCONTEXT, WM_IME_ COMPOSITION and WM_IME SELECT are relatively important and should be processed first. Through the process discussed above, we develop outline fonts and input method for the Naxi pictograph. Figure6 shows the usage situation of this input method in word.

5.5 Evaluation of Naxi Pictograph Pinyin Input Method

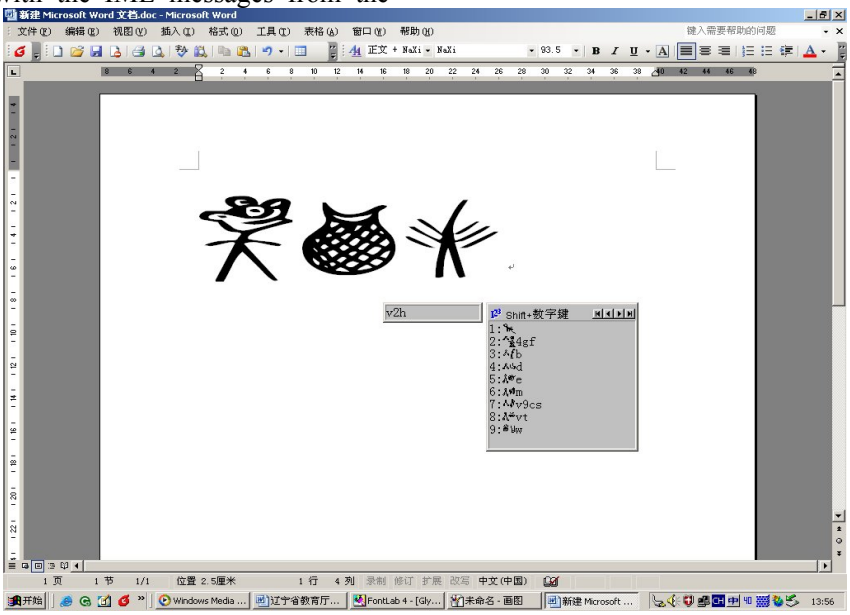

Fig.6 Naxi Pictograph Pinyin Input Method

In order to evaluate the efficiency of Naxi pictograph input method proposed by this paper and the original method,we develop an evaluating system for Naxi pictograph input method. Its flow procedure is shown in Figure7. The evaluating system employs the Windows message mechanism to automatically converse the Naxi Pinyin text to Naxi pictograph text under the conditions of activating the Naxi input method with and without optimization. After getting the evaluation result, we can calculate the conversion accuracy.

6 Web Embedding Application

Aiming at the ordinary customers' requirement, our project also made some study on the application of Naxi pictograph and developed Web embedding and PDF embedding technology.

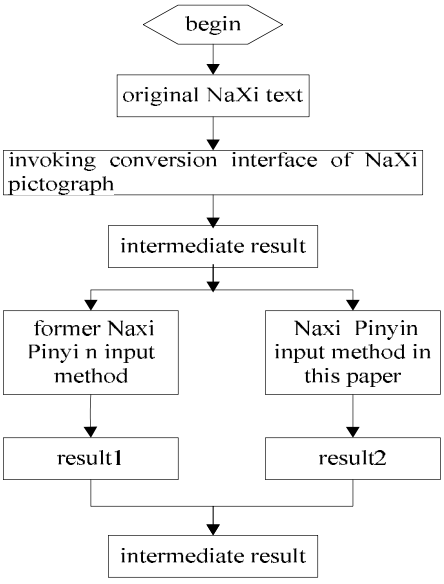

Fig.7 Working flow of Naxi Pictograph Pinyin Evaluation System

6.1 The Classification of Web Embedding

The lattice font has been gradually eliminated at the beginning of 90s in last century with the updating of operating system, the new outlines font has replaced it. The mainstream outlines font includes:

①PostScript Type 1 font, brought forward by Adobe corporation years ago, applied in publication typesetting system, belongs to the first generation of outlines font.

②TrueType font, brought forward by Apple and Microsoft. Because it has so many merits, it has been used in a variety of operating system on Mac and Pc, belongs to the second generation of outlines font.

③OpenType font, Put forward by Adobe and Microsoft as a new generation of outlines font standard. It fuses the merits of Type 1 and TrueType, belongs to the third generation of outlines font.

The TrueType font embedding technology of NaXi pictographs can be divided into embedded OpenType and True Doc. Put forward by Microsoft, Embedded OpenType technology can compresse TrueType font into Eot file, and then embeds it into HTML web pages. While TrueDoc brought forward by Netscape and Bitstream can compress TrueType font into TrueDoc file and then embed it into HTML web pages. The primary difference between the two can be seen in Table 3.

In the experiment, we choose five Naxi texts to the evaluating system, the results demonstrate that the optimized Naxi Pinyin input method achieved higher average conversion accuracy than the original method.Tabel4 shows details.

Table3. The classification of web embedding fonts technology

Transplant Type	Supported font	Supported browsers	Transplant Development tool
Embedded OpenType	TrueType OpenType	IE Navigator	WEFT (free)
TrueDoc	TrueType PostScrip Type 1	Navigator	HexWeb Typograph($149)

Table 4 Evaluation results Comparison of two Naxi Pinyin Input Methods

Conversion Accuracy	text1	text2	text3	text4	text5	Average Accuracy
Original Method	98.6%	97.4%	96.6%	95.8%	96.4%	96.9%
Optimized Method	99.1%	98.1%	98.4%	96.1%	97.8%	98.1%

7 Web Embedding Application

Aiming at the ordinary customers' requirement, our project also made some study on the application of Naxi pictograph and developed Web embedding and PDF embedding technology.

6.2 The Classification of Web Embedding

The lattice font has been gradually eliminated at the beginning of 90s in last century with the updating of operating system, the new outlines font has replaced it. The mainstream outlines font includes:

①PostScript Type 1 font, brought forward by Adobe corporation years ago, applied in publication typesetting system, belongs to the first generation of outlines font.

②TrueType font, brought forward by Apple and Microsoft. Because it has so many merits, it has been used in a variety of operating system on Mac and Pc, belongs to the second generation of outlines font.

③OpenType font, Put forward by Adobe and Microsoft as a new generation of outlines font standard. It fuses the merits of Type 1 and TrueType, belongs to the third generation of outlines font.

The TrueType font embedding technology of NaXi pictographs can be divided into embedded OpenType and True Doc. Put forward by Microsoft, Embedded OpenType technology can compresse TrueType font into Eot file, and then embeds it into HTML web pages. While TrueDoc brought forward by Netscape and Bitstream can compress TrueType font into TrueDoc file and then embed it into HTML web pages. The primary difference between the two can be seen in Table 4.

6.3 The Principal of Web Embedding Technology of NaXi Pictographs

Web embedding technology of NaXi pictographs uses CSS2 calling compressed OpenType font files to embed NaXi pictographs into web page. Downloading compressed fonts of NaXi pictographs downloaded to clients' temporary directories lets clients browse web pages with NaXi pictographs correctly without installing NaXi pictographs font.. The principle is that a client sends requirement for browsing, and then HTTP server sends HTML files to client browser. Client sends the information of web pages contained NaXi pictographs to server through CSS2, the server send this information to EOT database. EOT calls the TrueType font files stored in HTTP server, then HTTP server sends this inmessage back to the browser. The NaXi pictographs will be deleted autonomously after the client closes the browser. In this way fonts copyright can be better protected from piracy. By calling CSS2 several times in a web page, not only NaXi pictographs but Chinese, Mongolian, Tibetan and other minority languages can be displayed in the same web page. The principle is shown in Fig.8.

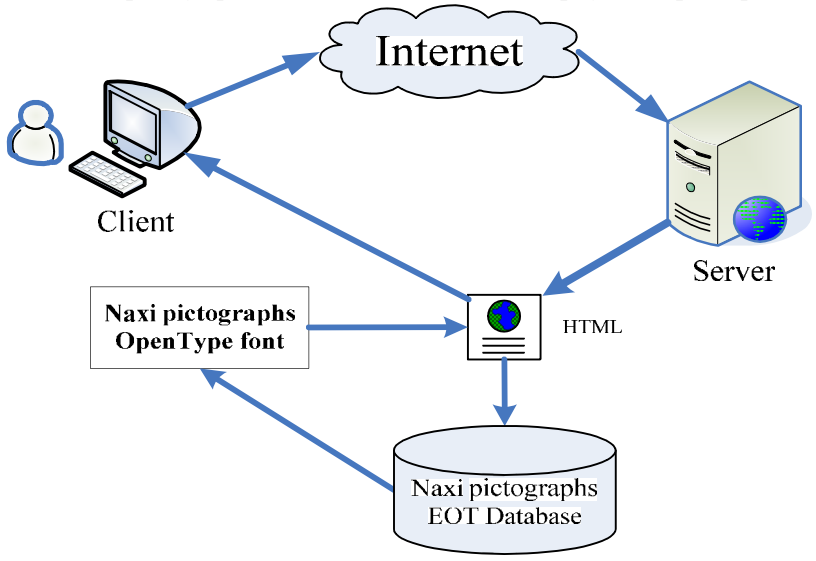

Fig.8 The principle of web embedding tochnology of NaXi pictographs

6.4 Realization of Web Embedding Fonts Technology of NaXi Pictographs

6.4.1 The environment of developing web embedding fonts of NaXi pictographs

The development environment is an environment of making web pages, EOT files, and CSS tables, it is mainly composed of DreamWeaver, WETF (Web Embedding Fonts Tool), Font Creator and etc. Free BSD installing Apache is adopted as the server. Firstly this structure is totally compatible with Microsoft IIS system. Secondly, it provides more useful functions, faster operating speed and better stability than Microsoft IIS system.

6.4.2 The application of CSS2 in web embedding fonts of NaXi pictographs

pictographs TrueType font into web pages possible. From the way that CSS is inserted, there are three

kinds of CSS: inline mode list, embedded mode list and exterior mode list, and the inline mode list and embedded mode list are broadly used in web embedding fonts technology of NaXi pictographs. The key sentence of CSS is @font-face, which has defined the name, type, thickness and other information of a planted font.

6.4.3 The generation of database of NaXi pictographs planted fonts

The substance in creating font database is to compress NaXi pictographs TrueType fonts into OpenType fonts. There are many ways in generating NaXi pictographs planted databases, this paper adopts Microsoft WETF (Web Embedding Fonts Tool). Before planting the validity of NaXi pictographs TrueType fonts in the development environment are checked by WEFT. WEFT displays the font validity in graph, there are three hints as follows:

①Implies that fonts can be planted

②Implies that fonts can not be planted, possible errors exist or not allow to

③Implies that the font belongs to the core fonts of Window, don't need to be planted

After the success of font-checking, add the needed NaXi pictographs into EOT files, then

calling NaXi pictographs font while browsing will not be a problem.

6.4.4 The embedding of NaXi pictographs font

You can input NaXi pictographs by using NaXi pictographs input method developed by the information system lab, then create a web page containing NaXi pictographs with DreamWeaver. Then insert codes as follows between <head> and </head>:

```
<title> NaXi pictographs TrueType font Web planting website</title>
<meta http-equiv="Content-Type" content="text/html; charset=gb2312">
<STYLE TYPE="text/css">
<!-- /* $WEFT -- Created by: GuoHai (guohai@ourcampus.net) on 2003-4-4 -- */
 @font-face {
  font-family: NaXi;
  font-style: normal;
  font-weight: normal;
src: url(NAXI.eot); }
-->
</STYLE>
```

Embedding function @font-face provides with four parameters: you can use font-family to name this font in the current webpage, this project is defined as NaXi; font-style can be anyone of normal, italic or oblique, commonly defined as normal; font-weight can be normal, bold, bolder, lighter or other legal thickness value; FontURL is a URL pointing to OpenType files, normally adopts absolute pathway. After saving the NaXi pictographs webpage, you can test it by transfering it to the HTTP server. The testing Figure can be seen in Fig.9. Up to now, the preliminary process of planting NaXi pictographs is officially finished.

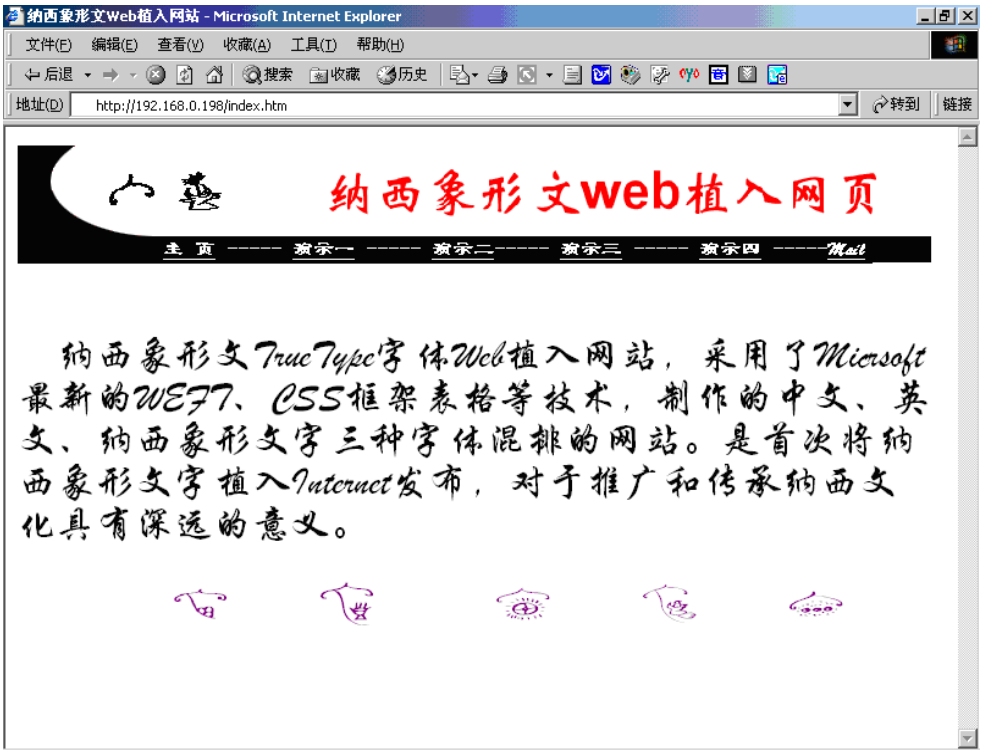

Fig.9 Web page of NaXi pictographs

7 Conclusion

Our project developed a complete Naxi pictograph information processing platform. It ends the history of Naxi pictograph processing without computer, provides valuable reference for creating other minority languages information processing system and also plays a significant role in promoting the computerization process of minority characters in China.

Acknowledgement

This work is supported by the pre-research project for national defense under the grant No 60803096.

References:

[1] Joseph Francis Charles Rock,A Nakhi-English encyclopedic dictionary, I.M.E.O, Rome,1963.

[2] Liu Yongkui, Guo Hai, Lu Guiyan,Li Hongyan,"Input technology and information

processing of NaXi pictograph",Journal of Computational Information Systems,v3, n1, February, 2007, p361-368.

[3]Guo Hai,Che Wengang,Nie Juan,Li Bin etc,"Web embedding fonts technology of Naxi pictographs",Jisuanji Gongcheng/Computer Engineering,v31,n17,Sep 5, 2005, p203-204+207.

[4] Guo Hai,Zhao Jing-ying,"Development of the NaXi Pictographs Information Processing System",Control & Automation,v22,n22, 2006, p122-124.

[5] Xie Qian,Jiang Li,Wu Jian,etc,"Research on Chinese Linux input method engine standard",Jisuanji Yanjiu yu Fazhan/Computer Research and Development, v43, n11, November, 2006, p1965-1971.

[6] Tseng Chun-Han, Chen Chia-Ping,"Chinese input method based on reduced Mandarin phonetic alphabet",INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP, v 2, INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP, 2006, p 733-736.

[7]Tanimoto Yoshio,Nanba Kuniharu,Rokumyo Yasuhiko, etc,"Evaluation system of suitable computer input device for patients", Proceedings of the Third Workshop - 2005 IEEE Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2005, Proceedings of the Third Workshop - 2005 IEEE Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2005, 2007, p 369-373.