

网页中即时动态显示和输入冷僻字的方法

夏立宁 高玉军 唐英敏 吕肖庆

(中国文字字体设计与研究中心 北京 100871)

(北京北大方正电子有限公司 北京 100085)

(北京大学计算机科学技术研究所 北京 100871)

【摘要】与其他国家的语言文字一样,字典里的字词,绝大部分都是不常用的。以目前公认的《汉语大字典》为样本,它共收单字五万多。但是,实际通用的汉字并不很多,那些不常见的汉字都被称为冷僻字。伴随着中文信息化技术的发展,越来越多的资料、公文、报刊等都数字化了,使得冷僻字的输入、显示及其传输问题变得越来越突出。本文提及的网页中冷僻字的动态显示和输入,可实现客户端无需安装本地字库和输入法即可进行在线编辑和显示,解决了冷僻字在网络上的编辑、录入和传播等问题。

【关键词】冷僻字,网络嵌入式字库, EOT, 在线输入法。

The Display And Input Of Rarely-used Chinese Characters Processing In Internet

Tang Yingmin, Xia Lining, Gao Yujun

(Institute of Computer Science and Technology, Peking University, Beijing 100871)

(Center for Chinese Font Design and Research, Beijing 100871)

(Beijing Founder Electronics Co. Ltd., Beijing 100085)

Abstract: Nowadays, rarely-used Chinese characters need to be displayed in e-book, e-mail and internet pages more and more. But lots of them still processed as images in current network age. And traditional input method can't be used to input these words to a document. With the problem of display and input rarely-used Chinese characters becomes more outstanding. This paper introduces a solution for display and input rarely-used Chinese characters processing in internet.

Keywords: rarely-used Chinese characters, Network embedded font, EOT, Online input method

1、引言

汉字的数量并没有准确数字,大约将近十万个,日常所使用的汉字只有几千字。据统计,1000个常用字能覆盖约92%的书面资料,2000字可覆盖98%以上,3000字时已到99%,简体与繁体的统计结果相差不大。在汉字计算机编码标准中,早期的国家标准GB2312-80只有6763个汉字,GB18030-2000^[1]收录了27484个汉字,而ISO-10646^[2]收录的汉字已超过七万。虽然ISO-10646收录汉字已超过7万,但还有很多字正在等待审定。计算机系统中常用的字库大部分还只有两万多字,所以冷僻字势必会长期存在。随着网络技术的发展,冷僻字的存在给电子邮件、电子书等电子信息媒介带来了很大不便,因为这些媒介里包含的冷僻字必须通过一定的手段传输到客户端的电脑上才能使阅读者看到这些冷僻字,否则,客户看到的仍是不完整的文档。人名、地名是使用冷僻字较多的领域,涉及到人名、地名的地方,如

果里面包含了冷僻字则会带来很多的麻烦,比如:身份证的办理,人名、地名的显示等问题。人们所熟知的一代证不是通过计算机系统制作的,遇到冷僻字用造字的办法解决这个问题,只要把姓名、住址打印上去就行;实行网络制证,采用网络传输以后,这就带来了一些问题,字库里没有的字,计算机就打不出来,公民就领不到身份证。

目前,电脑里显示的字为 TrueType 格式,均需预先安装后才能使用和显示,但这些字库通常不包含冷僻字,即便是有了这些冷僻字字库,如何正确、方便、快捷的输出和定位到这些冷僻字也是一直困扰冷僻字应用的一大瓶颈。冷僻字在线输入法通过笔顺、笔划、部首、拼音等多途径结合方式,可以方便的对所有冷僻字进行简单易行的查找和定位。

2、 基本方法和原理

网页中即时动态显示和输入冷僻字的方法实现了以下功能:

- 1) 客户端无需预先下载冷僻字字库和输入法
- 2) 在此基础上用户可以随意的输入、编辑、显示、检索冷僻字

即时显示冷僻字主要采用了动态生成冷僻字 EOT 字库的技术,对于用户输入的冷僻字通过调用后台的功能模块,将此部分冷僻字转换成可供网页调用的 EOT 嵌入字库,同时当前页面即时调用此 EOT 字库,这样用户的输入就即时的显示出来了。

冷僻字的输入主要采用的是笔顺、笔画和部首检索的方式,用户无需知道当前的冷僻字的读音即可根据字形按照其笔画和部首来进行查找和检索。笔顺检索时,可以按照冷僻字的笔画顺序来一步步选择输入的字母,不同的字母代表了不同的笔画,当输入完毕时即可找到所需的字。部首检索时,根据冷僻字的部首,可以找到含有该部首所有的字,用户只需从中挑出自己想输入的字即可。这种带有智能引导的输入法模式给用户带来了最简单方便的输入途径。另外,由于输入法本身也含 EOT 冷僻字字库,所以在输入的时候输入法界面上的冷僻字也是即时可见的。

进而,可以在该技术基础上建立一个冷僻字的使用和交流的平台,对于需要使用和显示冷僻字的用户,可以在此平台上进行检索并输入自己需要的冷僻字,然后通过即时动态生成的 EOT 字库将这些冷僻字正常的显示在其他任何网页上。而在线输入法又为此功能提供了必不可少的支持。

3、 冷僻字网络嵌入式 EOT 字库

3.1 EOT 字库

IE 浏览器里定义了一种网页嵌入字库格式——EOT (Embedded OpenType),这种字库格式解决了在网络中使用冷僻字的问题。EOT 是一种压缩字库格式,体积较小,适合在网络上传输,而且可以根据自己的需要做成仅含一部分字符的字库,无需每次都把所有字符都嵌入到字库里,这样就更进一步增加了它的灵活性,减少了文件的大小。同时,EOT 字库内嵌了安全机制,它在生成的时候和域名进行了绑定,只有事先绑定的域名才可以使用和显示该 EOT 字库。

当含有 EOT 的页面被客户端浏览器打开的时候,客户端浏览器会在后台自动下载页面中引用的 EOT 字库,下载完毕以后,页面上的冷僻字就能正常的显示出来了,改变了往常使用图片代替冷僻字的方法。既美观,使用又方便,还可以支持搜索。

3.2 冷僻字 EOT 的动态转换方法

3.2.1 EOT 动态转换概述

虽然微软网站提供了一种生成 EOT 字库的小工具，但是如果每次都使用此工具进行手动生成，然后再更新到服务器上，还是一件很麻烦的事情，而且不能即时显示需要看到的冷僻字。为了解决这个问题，可以在后台增加一个可以即时生成 EOT 字库的模块，当前端用户输入冷僻字后，后台会自动调用此模块，将所输入的冷僻字转换成 EOT 并在网页中进行调用。这样，用户输入完之后就可以很快看到自己输入的冷僻字了。

图 1 是即时显示冷僻字的一个简单示意图

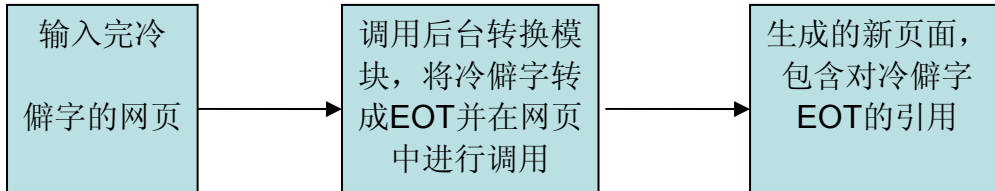


图 1

3.2.2 EOT 动态转换过程

为了即时生成 EOT 字库，需要在后台增加一个即时生成 EOT 字库的模块，在这个模块里面提供了一个函数，通过调用此函数就可以直接生成 EOT 字库了。在调用函数时，唯一需要传入的就是需要嵌入的字符。在这个模块里面已经预先设置好了需要绑定的 URL 地址，即当前网页的地址。有了传入的字符，模块还需要结合预先安装的冷僻字字库来生成所需的 EOT 字库。这个 EOT 字库就是网页中所调用的 EOT 字库。

3.2.3 EOT 的网页调用

有了所需的 EOT 字库，如果想在页面上即时的调用此字库，还需要即时的给页面加入调用语句，当生成了 EOT 字库后，模块里面会记录它的路径，然后在当前页面里加入类似如下格式的调用语句：

```
<STYLE TYPE="text/css">
@font-face
{
font-family:冷僻字嵌入式 EOT 字库;
src: url(lengpizi.eot);
}
</STYLE>
```

有了这些语句，网页中的冷僻字就可在不预先安装冷僻字字库的情况下正常的调用和显示。

3.2.4 动态生成的 EOT 字库在其他网页上的调用

如果用户需要将这些冷僻字在其他的页面上正确的显示，也可以“定制”这些冷僻字的 EOT 字库，用户只需把需要显示这些冷僻字的页面 URL 地址填入到指定的位置，然后输入所有可能用到的冷僻字，提交这些信息后，后台通过调用生成 EOT 的模块，把这些字符打包成 EOT 字库，并通过用户给定的域名进行授权，只有用户允许的域名下的网页才能调用此 EOT 字库。同时，反馈给用户一个下载链接，用户点击此链接后下载生成的 EOT 字库，这个 EOT 字库用户就可以用在他希望的那些域名下的网页上了，以后浏览这些网页的其他用户也可以正常的看到这些冷僻字。

4、冷僻字在线输入法

4.1 在线输入法介绍

EOT 技术很好的解决了在线显示冷僻字的问题，但是如果输入冷僻字又该怎么办呢？无法输入就无法进行录入工作和检索工作。在线输入法很好的解决了这个问题。

在线输入法区别于传统 IME 机制的输入法。它基于浏览器，可以支持常见的 IE 浏览器，FireFox 浏览器以及 NetScape 浏览器^[3]等。而且在线输入法不依赖于当前计算机的地区语言，可实现多文种，如：少数民族文、古文等。在线输入法也不需要在本地进行安装即可在线录入。有了在线输入法，普通用户均可在网页上直接录入冷僻字，其操作界面类似于常见的传统输入法，使用方便。使用完毕后，关闭当前的网页即可自动关闭在线输入法。

图 2 是方正超大字库在线输入法输入冷僻字的示例。

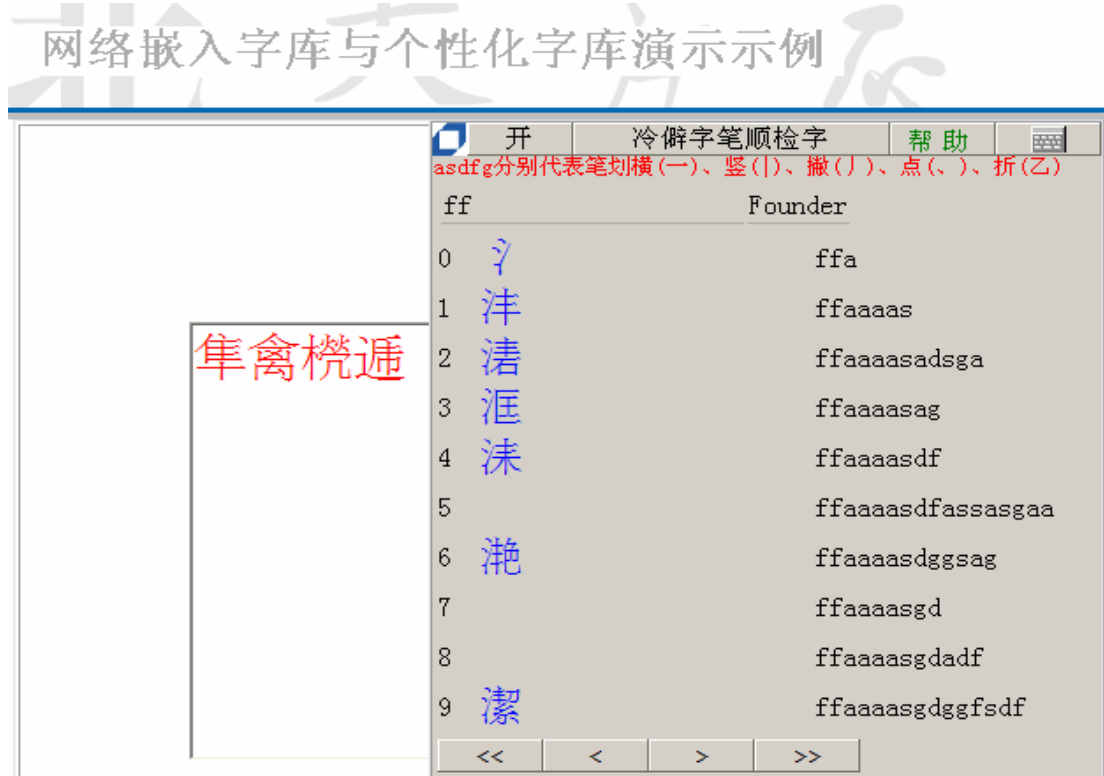


图 2

4.2 在线输入法的技术原理

在线输入法是通过 HTML、JavaScript、EOT、码表的相互配合来实现的。首先，用 HTML 语言描述出在线输入法的界面并保存成 HTML 文件，然后把描述界面的 HTML 文件引入到 JavaScript 程序中，这些 JavaScript 程序是可以被 Web 浏览器直接解释和运行的。同时，JavaScript 程序里有对键盘键位的响应，网页文本录入焦点的确定等功能。码表作为输入法的重要组成部分，使用 IE 中的免费 ActiveX 控件 TDC(Tabular Data Control)来进行符合规则的码表数据的绑定和检索。

4.3 冷僻字在线输入法的使用

当用户打开含有在线输入法的网页的同时就会动态生成在线输入法，使用输入法进行冷僻字的输入时，可采用冷僻字在线笔顺检字、部首检字、拼音检字以及三者相结合的方式

进行输入。用户可以根据自己对三种输入条件的熟悉程度进行选择。

4.3.1 笔顺检字

在笔顺检字里，asdfg 五个字母分别代表笔画里的横(一)，竖(|)，撇(丿)，点(丶)，折(乙)。图 3 演示了“乳”字的输入方式。

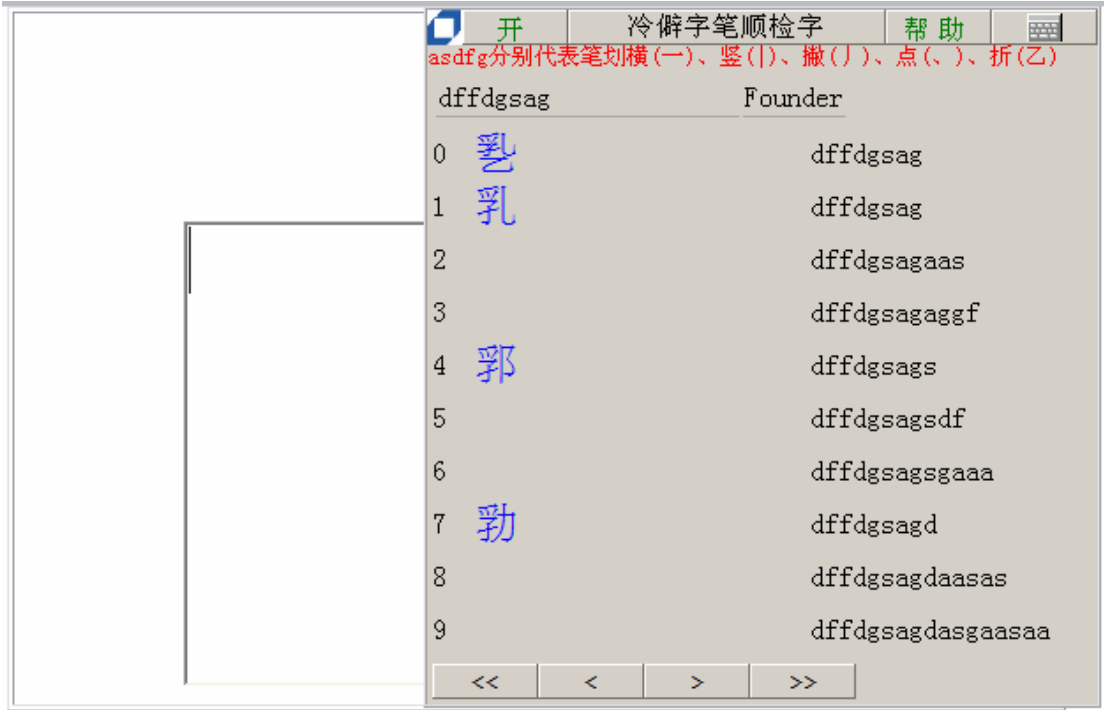


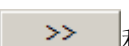
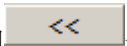


图 3

输入前，需要先选中编辑框，然后根据“乳”字的笔画及其顺序，可以知道该字的笔画依次是：撇、点、点、撇、折、竖、横、折，所以通过输入 dffdgsag 序列即可得到“乳”字，然后根据输入法界面中的提示可以知道第 0 个字即是想要的字，这时按数字 0 键或空格键，“乳”字就被输入到编辑框中。

如果不了解当前字的某一笔画应该对应横(一)，竖(|)，撇(丿)，点(丶)，折(乙)中的哪个了，这时还可以通过点击  和  按钮来向后、向前翻页，直到找到想输入的字，然后进行选择即可。 和  分别表示最后一页和第一页。

4.3.2 部首检字

在部首检字方式中，可以先按照部首的笔顺来输入部首部分，当找到需要的部首后，输入法会列出含有该部首的所有汉字，用户可以从中找到最终想要的字。如果列出的汉字过多不好查找，还可以通过输入剩余笔画数来缩小查找范围。另外，在输入部首时，除了通过笔顺来输入部首外还可以通过输入部首的拼音来输入部首或者通过输入部首的笔画数然后再进行下一步的查找来输入部首。这样一来，就可以通过部首、笔顺、拼音三者相结合的方式快速准确的找到所需的字。这种检字方法的优点是用户不需要认识这个字，输入法能够比较准确的缩小找字范围，使检字变得更加快捷。

4.3.3 拼音检字

拼音检字的使用方法比较简单，用户只需要输入汉字的拼音，然后通过选择拼音对应的

音调，然后，输入法会把该读音的汉字全部列举出来，用户只要从中选择自己想输入的字即可。这种方法的缺点是用户需要预先知道该字的读音，并且，由于相同读音的汉字可能会有很多，所以在查找时会比较费时，输入法不能很好的缩小查找范围。

总之，以上三种检字方式用户可以根据自己的喜好来手动选择适合自己的一种，或者对于不同的字选择不同的检字方式，这三种方式的结合使得检字变得更加方便易行，每一个人都不用担心在使用在线输入法时遇到找不到字的情况。

4.4 冷僻字嵌入字库与输入法的结合

当使用冷僻字在线输入法的时候会发现，不论输入什么冷僻字，输入法里都会在线即时的显示所输入的冷僻字，这是为什么呢？因为在输入法里也同样调用了冷僻字 EOT 字库，所以当动态产生在线输入法的同时，会自动调用冷僻字 EOT 字库进行基础显示，从而避免了客户端未安装冷僻字字库而使输入法界面显示空白的现象。当用户输入的时候输入法同样是通过调用冷僻字 EOT 字库来进行界面上冷僻字的显示。这样就保证了即使客户端未安装冷僻字字库也能正常使用冷僻字在线输入法。否则，未安装冷僻字字库的客户端将无法看到在线输入法界面上的冷僻字，也就无法正常输入和选择了。

5 结束语

随着网络的发展，冷僻字在网络上的使用和传播越来越显的重要。目前常用的方法只是通过放置图片来显示，更不要说输入了。本文介绍的即时动态显示和输入冷僻字的方法给冷僻字在网络上的应用提供了很好的途径，不但可以输入和显示冷僻字，还可以通过检索来查找含有冷僻字的内容。普通用户也无需每人都安装上冷僻字字库就能像使用常用汉字那样使用冷僻字了，这样一来，既简化了用户的使用，提高了效率，还方便了冷僻字的网络传输和交流。

在日本，他们规定了起名字的用字范围，不能随便起名，所以他们的名字都差不多。但我国人口众多，父母为了防止孩子和别人重名重姓，就故意在名字里加入了冷僻字，这样一来就给孩子以后的社会活动和交往带来了些许麻烦。据统计，目前涉及到姓名、住址方面的冷僻字，字库里没有的大概有 4600 多个。现在，完全可以把这些冷僻字通过本文所说的方法进行必要的处理，那样，就不用再担心遇到冷僻字了。冷僻字的出现，说明了还是有许多地方需要用到他们，尽管有很多麻烦，但不能为了逃避而不使用他们，应该做的是通过技术手段来克服这些困难，这样，技术才能更好的服务于人民，才能向着更高新更全面的方向发展。

参考文献

- [1] 《信息技术、信息交换用汉字编码字符集、基本集的扩充》(GB18030-2000)[S]，北京：中国标准出版社，2001
- [2] International Standard ISO/IEC 10646:2003 First Edition. Information technology – Universal Multiple-Octet Coded Character Set (UCS)[S], 2003
- [3]高玉军，刘慧杰，吕肖庆，唐英敏，尹江红. 小篆文本的在线编辑技术