

数码输入法字码本的自动获取技术

周克兰, 张玉华

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘要:为了科学地评测某种数码输入法的性能,首先必须获得该输入法的字码本。文中提出了通过技术分析自动获取字码本的两种方法:一种方法是使用文件监测程序 Filemon 和十六进制文件查看器 UltraEdit 分析经过压缩加密的码本文件,自动获取解密后的数码输入法的字码本;另一种方法是利用功能测试工具 WinRunner 自动获取字码本,创建可修改和可复用的测试脚本模拟汉字输入析出码本,再编写程序处理成字码本所需的格式。自动获取数码输入法的字码本对测试输入法的性能指标起到非常重要的作用。

关键词:数码输入法;字码本;自动获取;模拟汉字输入;功能测试工具

中图分类号:TP391.1

文献标识码:A

文章编号:1673-629X(2009)11-0031-04

Technology of Auto-Get Character Database in Numerical Input Method

ZHOU Ke-lan, ZHANG Yu-hua

(School of Computer Science and Technology, Suzhou University, Suzhou 215006, China)

Abstract: To evaluate the numerical input method scientifically, first of all, must get its character database. Supply two ways to auto-get the character database. One is using the file monitoring tool Filemon and the hexadecimal file editor Ultraedit to analyze the encrypted database file of an input method, then get the decrypted character database automatically. The other is using WinRunner, one of the function testing tools, to create the changeable and reusable testing scripts to simulate the Chinese character inputting process and then get the corresponding Chinese characters sequence, and last write a program to convert the characters sequence to character database. It is highly important to auto-get the character database of numerical input method to evaluate the performance of it.

Key words: numerical input method; character database; auto-get; simulate Chinese character input; function testing tool

0 引言

随着各种电子产品的数字化,特别是手机的进一步普及,使得在这些数码产品中输入汉字成为一个迫切的需求,数码输入法成为一个热门的研究课题,涌现出各种数字编码方案的输入法。

如何评价数码输入法编码方案的优劣?可以依据 GB/T 18031《信息技术 数字键盘汉字输入通用要求》的性能指标进行评价。GB/T 18031^[1]中定量的输入系统性能指标有两个:汉字输入平均码长和重码字词键选率。国家标准给出的指标是当前应达到的最低要求。

(1) 汉字输入平均码长。

定义:在输入给定的测试样本时,测得的输入每个

汉字的平均击键次数。

计算公式:平均码长 = 输入样本的击键次数/测试样本总字数(键/字)。

GB/T 18031 给出了通用键盘和数字键盘汉字输入平均码长的指标,如表 1 所示。

表 1 GB/T 18031(数字键盘)给出的指标

| 编码类型 | 平均码长(键/字) |
|---------|-----------|
| 逐字段输入 | <6 |
| 字、词混合输入 | <4 |

(2) 重码字词键选率。

定义:在输入给定测试样本过程中,通过重码选择键确认的汉字字数与测试样本总字数的百分比。

计算公式:重码字词键选率 = (重码选择键确认的字数/测试样本总字数) × 100%。

GB/T 18031 给出的数字键盘重码字词键选率的指标,如表 2 所示。

作为编码发明者和输入系统设计者,要设计出科

收稿日期:2009-03-05;修回日期:2009-06-06

基金项目:国家教育部博士点基金(20060285008)

作者简介:周克兰(1965-),女,硕士,副教授,研究方向为中文信息处理。

学的汉字编码方案,必须对流行的各种输入法产品进行客观评测和研究,从而推出符合国家规范的更好的输入法产品。要科学地评价某个数码输入法,必须对该输入法的性能指标进行测试,为了对数码输入法进行性能测试,必须获取它们各自的字码本,因此,研究输入法码本的自动获取技术有着非常重要的意义^[2-7]。

表 2 GB/T 18031(数字键盘)给出的指标

| 输入方式 | 重码字、词键选率(%) |
|---------------|-------------|
| 逐字段笔画、部件码输入 | <8 |
| 字、词混合笔画、部件码输入 | <10 |

1 字码本自动获取技术

为了对数码输入法进行性能测试,首先必须获取它们各自的字码本。往往只能通过购买或网上下载后获得各种数码输入法软件,这些数码输入法软件只供用户使用,而它们的字码本文件往往经过压缩和加密,用户不可能直接获得其字码本,必须通过技术分析而得到^[4,8,9]。

字码本的获取方式通常分为四种方式:第一种方式是直接获取,有些输入法可以通过该输入法的研究机构获得,例如,纵横输入法^[10]的字码本;第二种方式是由输入法软件提供的字码本 word 文件转换而来,例如,几何码的字码本就可以根据系统提供的字码本的 word 文件转换而来^[11];第三种方式,通过分析输入法软件的各个文件找出码本文件,再分析码本文件结构,编写程序解析出字码本;第四种方法,利用功能测试工具 WinRunner 创建可修改和可复用的测试脚本模拟汉字输入析出码本,再编写程序处理成字码本所需的格式。这四种获取码本方式的特点是:前两种方法比较简单,但是只适用于个别输入法,后两种方法比较复杂,但通用性较强,特别是第四种方式,适用于所有数码输入法。文中主要介绍后两种字码本自动获取技术。

2 分析输入法码本文件自动获取字码本

数码输入法软件的字码本文件均经过压缩和加密,需要通过分析输入法软件的各个文件找出码本文件,再分析码本文件结构,编写程序解析出字码本。下面以数码输入法 A 为例说明此种解析码本的技术。

为了研究数码输入法 A 的特点,笔者分析了数码输入法 A 软件^[9-11]的文件,找出字码本文件,通过分析字码本结构解析出笔者所需格式的字码本文件。下面详细介绍数码输入法 A 的字码本解析方法。

●所有的资源:

数码输入法 A 安装程序;

数码输入法 A 说明书。

●解析码本采用的工具:

文件监测小程序,如 Filemon;

十六进制文件查看器,如 UltraEdit;

自己写一些小程序辅助分析。

●数码输入法 A 字码本的解析流程:

使用 Filemon 监测出数码输入法 A 的码本文件,发现在 C:\WINDOWS\SYSTEM 下有 13 个文件有可能为码本文件。

使用 UltraEdit 查看这些文件的内容。这时需要很多的假设,然后进行验证。笔者很快发现“GWB-SMGBK.RCM”和“GLBHGBK.RCM”中的内容很有规律。其中的“GWBSMGBK.RCM”的内容中每八个字节就有一个汉字,可以假设这八个字节中的其他六个字节包含这些汉字的输入码。

此时可以对照说明书中的字编码进行分析。为了便于对照,笔者写了一个小程序把上述文件转换成一个文本文件,同时能从说明书中得到一个汉字编码,如图 1 所示。然后对比它们的异同,寻找规律。

| 九键六码: csv | | 九键六码 | |
|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| 文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H) | 文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H) | 文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H) | 文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H) |
| "D544C8100000", "啊" | "啊", "65212" | "啊", "65212" | "啊", "65212" |
| "A938C8180000", "阿" | "阿", "52162" | "阿", "52162" | "阿", "52162" |
| "579DC8180280", "埃" | "埃", "25488" | "埃", "25488" | "埃", "25488" |
| "E7B1CF2040A0", "挨" | "挨", "715488" | "挨", "715488" | "挨", "715488" |
| "DC5CC8081100", "哎" | "哎", "6727" | "哎", "6727" | "哎", "6727" |

原始数据

已知数据

图 1 原始数据和已知数据的对照

已知数码输入法 A 的基本码元为 1 到 9,通过对八个字节中的六个字节仔细分析,发现使用 3 个 bit 表示一个编码。0 表示没有编码,1-7 表示相应编码。其中第一个字节的 1 到 3 位表示第一码,第一个字节的 4 到 6 位表示第二码,第二个字节的 1 到 3 位表示第三码,第二个字节的 4 到 6 位表示第四码,第一个字节的 7 到 8 位和第二个字节的第 8 位表示第五码,第三个字节的 6 到 8 位表示第六码。

对于数码输入法 A 的每一位编码还可能是 8 和 9,这是 3 个 bit 不能表示的。实际上剩下的两个字节就是一个修正值,如图 2 所示。每两个比特代表一个

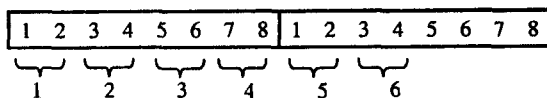


图 2 编码修正解析图

修正值,0 和 1 表示无须修正,2 表示加 1,3 表示加 2。

根据上述分析结果再编写程序把“GWBSMGBK.RCM”的相关内容转换为文本文件,将文本文件和数码输入法 A 说明书中的码表对比,发现正是数码输入法 A 字码本文件,从而完成数码输入法 A 字码本的解析。

3 利用功能测试工具 WinRunner 自动获取字码本

Mercury Interactive 的 WinRunner 软件为企业提供一个强大的功能测试工具。通过捕获、检测和重放用户对企业 Web 应用程序的互动操作,WinRunner 可自动执行功能测试。因此可以辨认错误,确保应用程序顺利部署,并且能够维持其长时间的可靠运行。

文中设计了一种通用的自动获取字码本的方法,该方法利用功能测试工具 WinRunner 创建可修改和可复用的测试脚本模拟汉字输入析出码本,再编写程序处理成字码本所需的格式。

下面以数码输入法 B^[9-11]为例说明此种字码本自动获取的技术。

解析原理:

通过模拟键盘敲击动作,动态穷举出所有字符出来。比如数码输入法 B 中汉字“奈”,需要输入的键有,“1”,“8”,“*”,“1”四个键,所以要模拟四次击键动作。

击键语句为 invoke-application(“c: \ \ sw. exe”, mychar+96, “”, SW_SHOW);

sw. exe 是一个模拟键盘的应用程序,mychar,就是需要输入的数字码,如“1”,“8”等,后两个参数不需要修改。

解析流程:

一个编码所有的汉字:

先打开一个临时文本文件,输入该编码,一一列举所有的当前页的汉字,Ctrl+s 保存,对它进行判断,判断所得当前页是否为有效页,将有效页存入目标文本文件,再列举第二页,重复上述过程,直至该编码列举结束。

整个过程:

有个 bitmap 文件,用来存储已有的编码信息,用 0,1 表示该编码是否存在,比如 23:1,表示编码 23 有相应的汉字,103:0,表示 103 无对应的汉字,该文件的目的在于减少不必要的析码,如果读到 103,发现 103 不存在,则 1031,1032,1033,... 对应的编码均不存在,所以遍历到 103XXX 的时候,直接就跳过去,不再对它进行处理。遍历之前,先读 bitmap,获取目前所取得的汉字编码信息,然后遍历,每个编码的处理见上。

程序中的自定义函数说明:

public function inputchar(number, pagenum); 输入指定编码 number 的第 pagenum 页。

public function inputcharslow(number, pagenum); 同上,不同在于只是时间上放慢,降低输入的出错率。

public function readchar(); 读临时文件中的当前页的内容,以“|”分隔。

运行前需要准备的是:

在桌面上新建一个文本文件“tempfile. txt”,将“sw. exe”保存在 C 盘中,将 bitmap. txt, character. txt 放入 D 盘中,也可以放入任意的分区中。

运行时:

打开 tempfile. txt,点击 winrunner 中的 run from top 按钮,再激活 tempfile 文件。

程序中使用常用函数说明:

set_ window (“tempfile - 记事本”,1); 激活临时文件,用时 1 秒。

file_ printf (“d: \ \ character. txt”, “\ r \ n” & num & “: \ r \ n”); 将 num 写入 character 文件中。

type (“<kCtrl-L- >”); 模拟 Ctrl+ 空格的动作。

wait(1); 等待一秒钟。

file_ open(); 打开文件。

file_ getline(); 读文件中的一行。

file_ close(); 关闭文件。

substr(); 获取字符串子串。

主程序代码:

```
load_dll(“User32.dll”);
extern int keybd_event(char, char, long, long);
file_ open(“d: \ \ character. txt”, 2);
# Program Manager
set_ window (“Program Manager”, 1);
set_ window (“tempfile-记事本”, 1);
public bitmap[];
bitmap[0] = 1;
# if the program goes wrong, read bitmap[] from the bitmap. txt,
set the num to last digital processed
readbitmap(bitmap);
For(num=0; num<1000000; num++)
{
    ancestor = int(num/10); bitmap[num] = bitmap[ancestor];
    if(bitmap[num] == 0) continue; # must not exist, go to next number!
    inputchar(num, 1);
    currentpage = readchar();
    firstchar = getfirstchar(currentpage);
    currentpage = cuttailfirstpage(firstchar, currentpage);
```

```

report_msg("1page:"&currentpage);
if(length(currentpage) == 0)
{
    bitmap[num] = 0;
    # report_msg("fdfsafa");
}
else # more than one character
{
    bitmap[num] = 1;
    lastpage = "";
    hasnextpage = 1;
    firstpage = 1;
    pagenum = 1;
    while(hasnextpage)
    {
        if(length(currentpage) < 20)
        {
            if(firstpage)
            {
                file_printf("d: \\ character.txt", "\\ r\\ n"&num&": \\ r\\ n");
                firstpage = 0;
            }
            file_printf("d: \\ character.txt", currentpage&" - - -");
            hasnextpage = 0;
        }
        else
        {
            pagenum ++;
            if(firstpage)
            {
                file_printf("d: \\ character.txt", "\\ r\\ n"&num&": \\ r\\ n");
                firstpage = 0;
            }
            file_printf("d: \\ character.txt", currentpage&" - - -");
            lastpage = currentpage;
            inputchar(num, pagenum);
            currentpage = readchar();
            currentpage = cuttail(firstchar, currentpage);
            # report_msg ( pagenum&"page:"&currentpage&": "&length
            (currentpage));
        } # end of else
    } # end of while
} # end of else
}

```

使用此种自动获取字码本的方法对几种比较流行的数码输入法进行了解析,再编写程序将析出的码本处理为所需的格式,经过与各数码输入法的字码本比对,发现两者一致,从而验证了此自动获取字码本方法的通用性。

4 结束语

文中重点论述了通过技术手段获取字码本两种方式:一种是通过 Filemon 和 UltraEdit 软件分析输入法软件的各个文件从而找出码本文件,再分析码本文件结构,编写程序解析出字码本;另一种方法是利用功能测试工具 WinRunner,创建可修改和可复用的测试脚本模拟汉字输入析出码本,再编写程序处理成字码本所需的格式。采用上述字码本的获取方法同样可以获得各输入的词码本。由于各输入法词码本的词库差异很大,采用这样的词码本测试所得的各输入法的性能指标缺乏可比性,一般采用统一的词库,针对几个不同的词库样本,采用各输入法的词组编码规则,利用各输入法的字码本编程自动生成各输入的词码本。研究字码本自动获取技术为进一步研究自动测试数码输入法性能指标打下了基础。

参考文献:

- [1] 中华人民共和国国家标准. GB/T 18031-2000 数字键盘汉字输入通用要求[S]. 2000.
- [2] 罗桂琼,费红晓,戴戈. 基于反序词典的中文分词技术研究[J]. 计算机技术与发展, 2008, 18(1): 80-83.
- [3] 蔡增玉,刘书如,张建伟,等. 汉字模糊有穷自动机的研究[J]. 计算机技术与发展, 2008, 18(3): 89-91.
- [4] 周克兰,吕强,张玉华,等. 试论汉字数字输入法评价[J]. 中文信息学报, 2007, 21(1): 67-73.
- [5] Zhu Qiaoming, Li Peifeng, Gu Ping, et al. A Chinese Mobile Phone Input Method Based on the Dynamic and Self-study Language Model[C]//The 1st International workshop on Embedded Software Optimization (ESO2006). Seoul, Korea: EUC Workshops, 2006: 836-847.
- [6] Wu Xian, Zhu Qiaoming, Li Peifeng, et al. The Pretreatment of Chinese Character Database Based on ISO10646[C]//Proceedings of the Fourth International Conference on Computer and Information Technology (CIT 2004). Wuhan, China: [s. n.], 2004: 1134-1140.
- [7] Li Peifeng, Gu Pin, Zhu Qiaoming. A Dynamic and Self-study Language Model Oriented to Chinese Characters Input[C]//Proceedings of the Seventh International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2006). Las Vegas, USA: [s. n.], 2006: 311-318.
- [8] 张玉华,周克兰. 基于规则库的汉字输入法自动评测系统的设计[J]. 中文信息学报, 2004, 18(4): 50-54.
- [9] 周克兰,张玉华. 数码输入法功能的分析与研究[J]. 苏州大学学报:工科版, 2004, 24(1): 45-47.
- [10] 苏州大学纵横汉字信息技术研究所. 计算机纵横汉字输入系统教程[M]. 苏州:苏州大学出版社, 2008.
- [11] 南京虎光数码科技有限公司. 几何数码打字王 V5.0[CP/CD]. 南京:江苏电子音像出版社, 2002.