

水书水字可视化输入中的模式匹配

戴 丹, 陈笑荣

(贵州大学 计算机科学与信息学院, 贵州 贵阳 550025)

摘 要:在我国贵州南部至今还使用一种古老的民族宗教典籍“水书”。在计算机中输入水书并对水字进行编辑,将有利于保存这一文化瑰宝,便于水书的流传、学习和研究。介绍了水书水字可视化输入法,模式匹配问题及匹配算法,并给出模式匹配算法在水书水字输入法中的具体实现。为了加快在水书内存码表中查找匹配的水字编码的速度,采用了哈希表并构造了哈希函数,同时解决冲突。最后进行了输入字符串的分析。实验结果表明,使用哈希表和哈希函数大大加快了水书水字可视化输入中的模式匹配速度。

关键词:水书;水字;输入法;码表;模式匹配

中图分类号:TP319

文献标识码:A

文章编号:1673-629X(2011)09-0187-03

Pattern Matching in Visual Input Method of Shui Character

DAI Dan, CHEN Xiao-rong

(School of Computer Science and Information, Guizhou University, Guiyang 550025, China)

Abstract: Shui script is an ancient ethnic and religious scripture currently used in the south of Guizhou province in China. If Shui character can be directly input computer and handled by the edit software, it will benefit to protect the cultural treasure and provide grate convenience to study and research Shui script. The problem and algorithms of pattern matching are introduced, the concrete implementation is provided in input method of Shui character. In order to mend the speed of looking up codes in shared code table, the hash tables are used and the hash functions are produced. At the same time, the conflict is resolved. At last, the input strings are analyzed. The result shows that using hash tables and hash function mends the speed of pattern matching.

Key words: Shui script; Shui character; input method; code table; pattern matching

0 引言

全国约40多万人口的水族主要聚居在中国西部贵州黔南布依族、苗族自治州。水族人自己祖先创造的文字和用这种文字写成的典籍统称为“泐(音le)虽”。“泐”在水语里含有“书”和“字”两种含义。所以人们一直把这种文字记录成册的书籍统称为“水书”^[1]。水书是一种类似甲骨文和金文的古老文字符号,也是除了东巴文之外又一存活的象形文字。在水族人的社会生活中,至今还起着很重要的作用,特别是在丧葬、营建、出行、过节、占卜、农事等活动中发挥着指导规范的作用^[2]。水书蕴藏着水族的语言、文字、天象、历法、宗教等方面的丰富资料,具有重要的学术价值。开发水书水字可视化输入法,使用计算机来对水书进行保存和研究具有重要的意义。

在水文输入法中,启动输入法实例时需要将输入

码对照表即码表文件装入内存,并从中查找与输入码对应的水文^[3]。当用户在写作窗口中输入水字的外部编码字符串时,系统将对输入的外部编码进行分析,再根据分析结果确定对应的水字字符串编码。据此,可以迅速找到每一个外码对应的水字信息,然后将对应的编码相同的那些水书文字显示到候选窗口中让用户选择他所需要的特定水字。

1 模式匹配

对用户输入的外部编码字符串进行分析,根据分析结果确定对应的水字字符串的过程其实就是一个在内存码表中查找与输入码匹配的编码的过程。这是一个字符串匹配问题,也就是在给定的字符串序列中查找另一个或若干个字符串序列的问题,因此,又将字符串匹配称为模式匹配。其中,给定的字符串称为文本串,可记为 $T = t_1 t_2 \cdots t_n$;要查找的字符串称为模式串,记为 $P = p_1 p_2 \cdots p_m$ 。 T 和 P 都定义在同一个字符集合 Σ 上, n 和 m 为自然数。

单字符串的模式匹配就是在在一个文本串 $T =$

收稿日期:2011-02-28;修回日期:2011-06-05

基金项目:贵州省科学基金项目(黔科合J字[2010]2093号)

作者简介:戴 丹(1979-),女,贵州贵阳人,硕士,讲师,主要研究方向为网络与数据库、信息处理、图形图像。

$t_1 t_2 \cdots t_n$ 中找到某个特定模式串 $P = p_1 p_2 \cdots p_m$ 的所有出现位置集合^[4]。若要在一个大型字符串集合中搜寻若干个符合条件的字符串,就是字符串集合的模式匹配问题。字符串集合的模式匹配问题可描述为求解 $T = t_1 t_2 \cdots t_n$ 中所有出现任意 P_i 的位置。即根据建立一个有限的符号集合上的特征串和文本,找到文本中与特征串完全相等的子串的所有出现位置^[5]。

2 匹配算法

实现字符串匹配有很多算法,典型的有蛮力算法、KMP 算法、BM 算法、BOM 算法等^[6]。

2.1 蛮力法

最简单的字符串匹配算法是蛮力法。它用双重循环来求解简单的字符串匹配。外层循环移动窗口,内层循环检查窗口内字符串是否与模式串相等。其时间复杂度最差,最差性能为 $O(n * m)$, 最好性能为 $O(n)$ 。

2.2 KMP 算法

KMP 算法的匹配过程如图 1 所示。

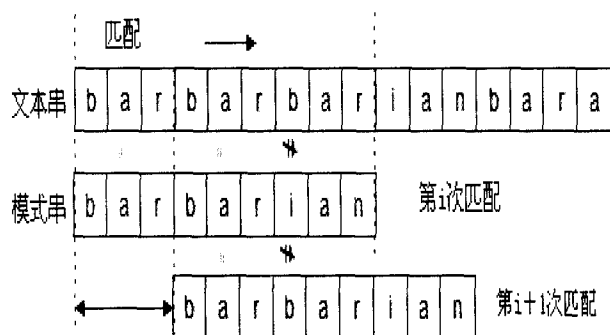


图 1 KMP 算法匹配过程

这是一种改进的模式匹配算法,对模式串的最长前缀做了预处理。生成一个函数 $next[i]$, 其中 $1 \leq i \leq m$, 表示在第 i 位发生不匹配的时候,可以计算已经比较过的前缀串 $p_1 p_2 \cdots p_{i-1}$ 中是否有一个最大的 k 使得 $p_1 p_2 \cdots p_{k-1} = p_{i-k+1} p_{i-k+2} \cdots p_{i-1}$, 如果有,那么有 $next[i] = k$, 比较下一位的时候可以将模式串 P 直接向后移动 $i - next[i]$ 位,并从模式串的第 k 位开始比较。如果不存在这样的 k ,有 $next[i] = l, next[l] = 0$ ^[6]。

KMP 算法预处理过程的时间复杂度为 $O(m)$, 进行匹配的时间复杂度为 $O(n)$ 。因此,该算法的时间复杂度为 $O(n + m)$ ^[7]。

3 水书水字可视化输入法中的模式匹配

3.1 水书水字输入法

水书水字输入法其实是一个动态链接库,其后缀名为 ime。输入法主要有三个窗口:状态窗口、写作窗

口及候选窗口^[8]。运行时界面如图 2 所示,状态窗口显示当前的水字输入法的状态,并可显示当前是水字输入还是英文输入。用户通过敲击键盘在写作窗口中输入水书文字的外部编码,该编码对应的水书文字就会显示在候选窗口中,用户通过加减号控制候选窗口翻页,显示不同的水字信息,敲击相应数字可选择需要的水书文字,确认之后就可在文字编辑或其它软件中输入显示被选择的水书文字。

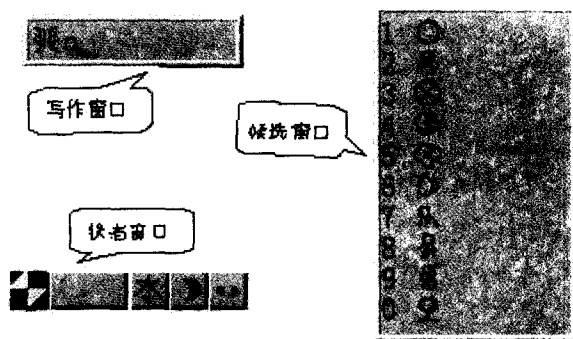


图 2 水书水字输入法界面

输入法启动时,首先装入码表文件,码表文件给出了每个输入码对应的水字,如输入码 aa 对应的水字为 𐤀𐤁;输入码 gho 对应的水字为 𐤆𐤇𐤈。当用户敲击键盘时,根据装入的码表文件中的编码,对输入的字符串进行分组,同时根据编码查找对应的水字,并分页显示在候选窗口中。为了加快在内存码表中查找匹配编码的速度,采用了哈希表。

3.2 哈希表和哈希函数的确定

3.2.1 哈希表

哈希表是根据关键码值(Key)而直接进行访问的数据结构。其中映射函数为哈希函数,存放记录的数组为哈希表^[9],算法理想复杂度为 $O(1)$ 。哈希表的性能主要取决于两个因素:哈希函数和冲突处理。因此,在建造哈希表时不仅要设定一个好的哈希函数,而且要设定一种处理哈希冲突^[10]的方法。

外部码表文件读入内存之后,为了快速查找用户在写作窗口中输入的输入码对应的水字字符,根据码表文件的内容分别构造两个用来保存外部编码及其对应的水字字符的哈希表 BMTab 和 SSTab,如图 3 及图 4 所示。其中,外部编码信息保存在表 BMTab 中,水字字符信息保存在表 SSTab 中,wKey 是 WORD 类型的编码标识号;SZBJ 是字符串类型的外部编码值;wLen 是 WORD 类型的相同外部编码的水书文字的数目;lpSW 是一个指向相同外部编码水文字串的指针;lpBJTab 是指向内存中保存编码信息表的起始地址的指针;而 lpShuiWen 是指向内存中保存水文字符信息表的起始地址的指针。这里的每一个表都是连续存储区,每一条记录都按顺序存储。

lpBJTab	
wKey	SZBJ
1	aa
2	aaa
3	aaaa
4	ah
5	bb
K	vvv
k=1	vvvv

图 3 内存中保存编码信息的表 BMTab

lpShuiWen		
wKey	wLen	lpSW
1	2	编码为 aa 的所有水字字符 ⓪ 㒹
2	1	
		编码为 aaa 的所有水字字符 㒺

图 4 内存中保存水字字符信息的表 SSTab

3.2.2 哈希函数的构造及冲突解决

水书水字输入法使用二十个英文字母作为码元,与水书水字中的基本笔画和部件相对应,实现了一字一码的编码,从而确定了水书文字的标准键盘键位。因此,其外部编码是以英文字母 a 到 z 开头的编码。

在 BMTab 表中,MAX_EACH_BM_NUM 规定了以某一个英文字母(比如 o)开头的外部编码的个数的最大值。当外部编码被装入内存时,将外部编码以从 a 到 z 开头的编码分成 26 组依次存放在 BMTab 中,每组的空间大小都为 MAX_EACH_BM_NUM。这就得到了一个 26 行 MAX_EACH_BM_NUM 列的二维数组。

如果用户在写作窗口中输入字符串 st,st 的首字符是 cha,那么可以构造一个哈希函数:

$$f(c)=lpBJTab+(cha-a)\times MAX_EACH_BM_NUM \times sizeof(BM)$$

其中,BM 是 BMTab 表中每一条记录的大小。函数结果为以 cha 为首字符的一组编码在 BMTab 中的物理地址。这里所有首字符相同的输入字符串的散列地址相同,因此,哈希函数会产生冲突。处理的方法就是再一次在首字符相同的那组编码中查找是否有编码与 st 匹配。这是一个顺序查找过程。因为首字符相同的编码数目很少,查找的速度很快。

在找到匹配的编码后,就可以从表 BMTab 中得到该编码对应的编码标识号 wKey。再根据 wKey,构造 SSTab 表的哈希函数:

$$f(wKey)=lpShuiWen+wKey\times sizeof(SHUIWEN)$$

其中,SHUIWEN 是 SSTab 中每一条记录的大小。其结果是每一个编码在 SSTab 表中的物理地址。因此,每一个编码对应的水字字符信息可以迅速被找到,

由其中的 lpSW 就可以找到对应的编码相同的一组水书水字,这组水书水字就能显示到候选窗口中让用户来选择。

当然,水书水字的使用频度是有差异的,有的使用频度高,有的使用频度低,而且各个输入者对它们的使用频度也很不一样,所以在词库中统计了词频信息,并将使用频度最高的字、词优先显示在候选窗口中,使输入法更加智能化。

3.3 输入字符串分析

这里说的过程是输入的字符串刚好为一个编码的情况。实际输入时比这个要复杂得多。因为用户在输入字符的时候是一个连续的过程,一次就有可能输入一个词组甚至是一个句子,所以用户每输入一个新的字符就要分析一次。在水书水字输入法中,如果用户一次输入多个编码,他所选择的水书文字是首先显示到写作窗口中,直到他所输入的编码对应的水书水字都选择完了,才全部输出到字处理软件等应用系统中。

分析过程比较复杂,但可以给出简单的说明如下:

- ①将指针 lpBM 指向一个编码的首字符,设置偏移量 wOff 的初值为 0;
- ② 每输入一个字符,就把偏移量的值加 1;
- ③ 检查从 lpBM 到 lpBM+ wOff 的字符串是否是一个编码或一个编码的一部分。如果是,等待下一个输入,wOff 的值再加 1,再继续③。如果不是,从 lpBJ 到 lpBJ+ wOff-1 的字符串为一个编码,转④;
- ④将编码保存到一个缓冲区中,再将 lpBM 移到 lpBM+wOff 的位置,开始下一个编码的判断。

4 实验结果

以 KMP^[11]算法为对照,对文中哈希算法的效率进行测试。对于模式串匹配算法来说,评定一个算法好坏的一个重要指标是运算时间。所以总的运算时间与匹配成功的次数相结合能够较好地体现算法之间的优劣。分别取模式串长度为 2 及 4 对同一个水书水字输入法码表进行实验。分别记录不同模式串在文本中成功匹配的总次数和总的运行时间。实验环境为 Pentium 1.6 GHz,512MB 内存,Windows XP。实验平台为 Microsoft Visual C++ 6.0。实验结果如表 1 所示。

表 1 实验结果

模式串长度	KMP 算法		本算法	
	成功次数	时间 (ms)	成功次数	时间 (ms)
2	19456	67	19456	52
4	9560	48	9560	37

由上面的实验结果可以看出,对于同一文本,在成
(下转第 193 页)

效模拟曲柄滑块机构的运动,实现了以下功能。

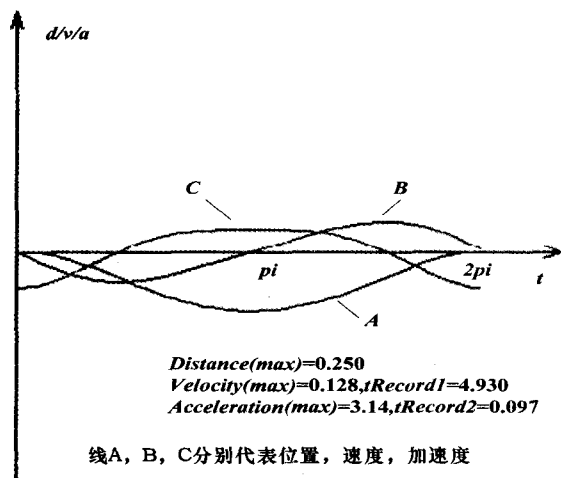


图5 滑块位置、速度、加速度图

1) 构建了一个较逼真的虚拟环境,实现了实时调节速度的功能,能够在不同速率下观察机构的运动。

2) 实时的界面输出机构运动参数,这些运动参数也为曲柄滑块机构的设计提供了参考。

3) 根据用户输入的参数生成滑块的位置、速度、加速度图^[9-12]。根据图像验证仿真的正确性。

随着虚拟仿真技术的产业化,该仿真思想也为用户开发新的可视化效果更好的仿真系统提供了有价值的参考。

参考文献:

[1] 于 辉,赵经成,付战平,等. EON 入门与高级应用技巧

(上接第 189 页)

功匹配相同次数的情况下,本算法实现了运行效率的进一步提高。

5 结束语

实现了水书水字的可视化输入。要解决水书文字输入码到内码的转换问题,必须考虑输入字符串的模式匹配。为了加快从内存码表中查找匹配编码的速度,采用了哈希表和哈希函数。给出了具体的哈希表和哈希函数以及在实际的水字输入法中输入的字符串的分析过程。结果表明输入法中对字符串的匹配效率有明显改善。

参考文献:

- [1] 吴正彪,祖 民. 守护精神的家园[M]. 北京:作家出版社, 2006.
- [2] 罗 刚. 计算机水书语料库建设的探讨[J]. 黔南民族师范学院学报, 2008(4): 67-69.
- [3] 戴 丹,董 芳. 内存映射文件及其在水文输入法码表共

[M]. 北京:国防工业出版社, 2008.

- [2] Tan Jianrong, Wang Zheng, Liu Zhenyu. Stable programmed manifold solver for virtual prototyping motion simulation[J]. Chinese Journal of Mechanical Engineering, 2006, 19(1): 76-80.
- [3] 刘 诚,付宜利. 基于 EON 的交互式虚拟装配仿真系统的设计与实现[J]. 东北林业大学学报, 2009, 37(8): 109-111.
- [4] Kallmann M. Object Interaction in Real-Time Virtual Environment[D]. Switzerland: Swiss Federal Institute of Technology, 2001.
- [5] 方传磊,苏群星,刘鹏远,等. EON 中基于 Script 的功能扩展研究[J]. 科学技术与工程, 2008, 8(3): 799-801.
- [6] 李世停,朱 波,陈力生,等. 基于 Script 交互控制的船用核动力装置虚拟维修研究[J]. 船海工程, 2006(1): 43-45.
- [7] 曾昭德,王 政,黄镇昌. LabView 在平面四杆机构运动分析与仿真的应用[J]. 现代制造工程, 2005(3): 112-114.
- [8] 张建中,高 宁. 连杆虚拟样机运动仿真[J]. 计算机技术与发展, 2008, 18(11): 191-193.
- [9] 张 扬,陈再良. 基于 TurboC 程序的曲柄滑块机构的运动仿真[J]. 机械设计与制造, 2006(2): 133-134.
- [10] 林水雄,余伟铭,刘 峰. 基于 MATLAB 及 Pro/E 对曲柄导杆滑块组合机构的仿真[J]. 机械设计与制造, 2009(3): 86-88.
- [11] 李娟玲,张建峰. 基于 C 语言的平面连杆机构的运动分析[J]. 机械研究与应用, 2006, 19(5): 117-118.
- [12] 乔沙林. 几种常用连杆机构输入输出位移方程的一般公式及微机辅助运动学分析[J]. 机械, 1988, 15(3): 19-22.
- 享中的应用[J]. 贵州工业大学学报, 2007(4): 69-71.
- [4] 何 畏. 快速精确字符串匹配算法研究[D]. 合肥:合肥工业大学, 2010.
- [5] 李 雪. 大规模特征串匹配技术的研究[D]. 北京:北京邮电大学, 2008.
- [6] Knuth D E, Morris J H, Pratt V R. Fast Pattern Matching in Strings[J]. SIAM Journal on Computing, 1977, 6(2): 323-350.
- [7] 欧 鬼,吴纯青. 几种字符串匹配算法的分析和比较[J]. 微处理机, 2007(4): 59-61.
- [8] 刘政治,李 炜,吴建国. 基于 IMM-IME 的汉字键盘输入法编程技术研究[J]. 计算机技术与发展, 2006, 16(12): 43-45.
- [9] 郑卫斌,张德运. 基于哈希表的高性能 URL 过滤器研究[J]. 小型微型计算机系统, 2005, 26(2): 178-180.
- [10] 马如林,蒋 华,张庆霞. 一种哈希表快速查找的改进方法[J]. 计算机工程与科学, 2008, 30(9): 66-68.
- [11] Baeze-Yates R, Navarro G. New and faster filters for multiple approximate string matching[J]. Random Structures & Algorithms, 2002, 20: 23-49.