中 文 信 息 学 报
第 13 卷 第 6 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 6

⑨

关于汉字字符串排序算法

61-64 TP391

钟 诚

广西大学计算机与信息工程学院 南宁 530004

摘要 分析汉字字符串分组排序算法，在讨论基选择的基础上，给出将字符串映射成整数和处理映射冲突数据的改进的有效方法。

关键词 汉字字符串 排序算法 映射 笔画排序，拼音排序

Research on the Chinese-Character Strings Sorting Algorithms

Zhong Cheng

School of Computer and Information Engineering Guangxi University Nanning 530004

Abstract Based on Paper[1], the choice of base, method of mapping chinese-character strings into integers and collision problem are researched, and an improved chinese-character strings sorting technique is given.

Keywords Chinese-Character String Sorting Algorithm Mapping

一、汉字字符串分组排序算法的核心思想

文献[1]基于分组原理和映射方法，给出一种处理汉字字符串的排序算法，并分析了其时空复杂性。该算法的特点是，引入映射方法，从而可以加快汉字字符串排序的速度。

汉字字符串分组排序算法的关键是，若按汉字的拼音排序，则取基为 27，即将任一个汉字字符串转换成一个 27 进制的整数；若按汉字的笔画排序，则取其为 31，即将任一个汉字字符串转换成一个 31 进制的整数。然后，将汉字字符串大小顺序关系的比较映射转换成数值数据的大小关系的比较。最后，通过排序这些 27 或 31 进制的整数数据序列来完成汉字字符串的排序。

二、将汉字字符串映射转换成数值数据的排序方法

2.1 基的选择

在实际应用中，汉字字符串中的字符可以是字典表中的任一字符，即对于任给的一个汉字

受国家自然科学基金(59665002)、广西自然科学基金(9712010)和广西教委科研基金(S96308)资助
本文于 1998 年 8 月 4 日收到

字符串，它除了含有常规的汉字之外，还可能含有英文字母或者其他字符的特殊符号。

例如，科技论文的题目就可能既有汉字又有英文字符和其他特殊符号，而且有些题目是以特殊符号开头的。当按论文题目进行索引排序时，就需要同时处理中英文字符和其他特殊符号。再如，对于多个中外学者合作署名撰写论的场合，就既有汉字字符又可能有法文或德文西班牙文字符等，当按姓名进行索引排序时，也需要同时处理各种中西文字符。

若选取27或31作为基进行数据转换会使得排序算法的应用范围有限(实际上它只能处理由纯汉字或纯英文字符组成的字符串的排序)。因此，基的选择值得考虑。

需要说明的是，不管字典表的大小是多少，也不管字符串由何种字符构成，这些待排序的(汉字)字符串在计算机内存中都可以用0或1组成的序列表示。这样，除了按中文姓名笔画顺序对纯汉字字符串排序的特殊场合外，一般可以取基d=2。

2.2 汉字字符串顺序关系映射转换成数值数据大小关系的讨论

假设取基 $d=27$ 或31，则对于任意给定的一个长度为 m 的汉字字符串 $c[1..m]$，可将其按如下公式映射转换成惟一的整数 x：

$$x = asc(c[1])27^{m-1} + asc(c[2]27^{m-2} + \cdots\cdots + asc(c[m-1]27^1 + asc(c[m])27^0 \quad (1)$$

或者

$$x = asc(c[1])31^{m-1} + asc(c[2]31^{m-2} + \cdots\cdots + asc(c[m-1]31^1 + asc(c[m])31^0 \quad (2)$$

其中，函数 $asc(c[i])$ 表示求字符 $c[i](i=1,2,3,\cdots,m)$ 对应的映射数字(0～26或者0～30)。

当字符串较长时，m 值较大，从而使得 27^{m-1} 或 31^{m-1} 的值很大，这可能会超出计算机所能不示整数的范围；另一方面，计算 27^i 或 $31^i(i=0,1,2,\cdots\cdots,m^{-1})$ 也需要较多时间。

若考虑将字符串转换成实数，则由于实数在计算机内存中是近似表示的，经过运算后可能就会成误差，这样两个不同的字符串(尤其是绝大多数字符都相同，只有个别字符不同的两个字符串)转换后可能会得到相同的数据，从而影响字符串排序的正确性。

综上所述，可见按式(1)或(2)进行映射转换的方法只适宜于较短的字符串的排序，而且排序速度较慢。

那么，可否找到一种方法既能处理较长的字符串，又能快速地将字符串映射成整数，从而可能快速地完成字符串的排序？图灵奖获得者 Karp 和 Rabin 在文献[2]中提出的将字符串映射转换成整数的方法值得我们参考：

设 d 为字典表的大小，将任一长度为 m 的字符串 $c[1..m]$ 按如下映射关系转换成整数 x：

$$y = asc(c[1]d^{m-1} + asc(c[2])d^{m-2} + \cdots\cdots + asc(c[m-1])d^1 + asc(c[m])d^0 \quad (3)$$

$$x = \mathrm{HASH}(y) = y \bmod p \quad (4)$$

这里的 p 是某个适当大的素数。在实际进行转换时，可能下列方法求出 x 的值(它可避免直接计算大整数 d^{m-1} 的问题)：

```
x:=0
for i:=1 to m do
  x:=(x * d+asc(c[i])) mod p;
```

设将 n 个字符串按式(3)进行映射转换得到的 n 个整数记为 $y_1, y_2, \cdots, y_{n-1}, y_n$；而将 $y_1, y_2, \cdots, y_{n-1}, y_n$ 按式(4)进行映射得到的 n 个整数记为 $x_1, x_2, \cdots, x_{n-1}, x_n$。

显然，式(3)的映射是一个单调上升的内射函数，因此对所得到的整数数据序列 $y_1, y_2,$

$\cdots, y_{n-1}, y_n$ 排序就相当于对 n 个字符串进行排序。

记 $max\_y = MAX(y_1, y_2, \cdots, y_{n-1}, y_n)$。为了使得对整数数据序列 $x_1, x_2, \cdots, x_{n-1}, x_n$ 排序相当于对 n 个字符串进行排序，要求式(4)最好也是单调上升的内射函数。

只要所选择的 p 是小于等于 max_y 的最大的那个素数，就可能使得按式(4)映射得到的整数数据序列 $x_1, x_2, \cdots, x_{n-1}, x_n$ 也是单调上升的。但是，式(4)的映射却不是一个内射函数，这表明存在映射冲突问题，即可能会出现两个不同的汉字字符串映射转换后得到相同的整数 x。尽管如此，按式(4)进行数据转换时产生映射冲突的可能性是极小的[2]；另一方面，即使发生冲突，我们也可以通过对那些具有相同的映射函数值的字符串再做一次常规的字符串比较(逐个比较字符的方法)，以确定出这些字符串的正确排序位置。从而使得本方法仍然具有实际的应用意义。

三、映射冲突部分数据的排序方法

文献[1]将 n 个整数 $x_1, x_2, \cdots, x_{n-1}, x_n$ 按如下关系进行映射排序：

$$position = \mathrm{int}((x_i - min) * k/(max - min)) + 1 \qquad (5)$$

其中：$i = 1, 2, 3, \cdots, n$；int 表示取整；max 和 min 分别为这 n 个整数中的最大元素和最小元素；$k = a * n$，a 为某个大于零的正常数。由离散数学理论并用反证法容易证明式(5)的映射不是一个内射函数，这表明对于两个不同的整数 x_i 和 x_j $(i \neq j; 1 \leqslant i, j \leqslant n)$ 按式(5)进行映射时，可能会出现如下的情况：

$$\mathrm{int}((x_i - min) * k/(max - min)) + 1 = \mathrm{int}((x_j - min) * k/(max - min)) + 1 = pos$$

即将 x_i 和 x_j 排序定位到散列表同一桶位置 pos 上，从而产生排序冲突。

设这些映射冲突部分的数据共有 s 个，若采用顺序插入排序方法对其进行排序，则需约 $s^2/4$ 次数据比较操作和 $s^2/4$ 次数据移动操作，故速度较慢。事实上，对于这部分数据可以采用更快速的方法进行排序。这是因为映射落入散列表同一桶位置上的那些数据是分布均匀的[3~6]。对于这些均匀分布的数据，可采用二次映射方法[5,6]或者改进的 QUICKSORT 方法[7]进行快速排序。从而，使得整个字符串的排序能更快地完成。

四、结束语

排序和查找约占据计算机处理数据 25% 的工作，而且排序有助于查找[8]。因此，研究快速的排序技术具有重要的意义。汉字字符串的排序有其特殊性。设计广义意义下通用的汉字字符串快速排序方法及其并行算法是一个值得进一步研究的课题。

参 考 文 献

[1] 周建饮. 关于汉字的分组排序算法及其复杂性. 中文信息学报，1996，10(3)：58－64

[2] Karp R M, Rabin M O. Efficient randomized pattern-matching algorithms. IBM J of Research and Development. 1987, 31(2): 249－260

[3] Cormen T H, Leiserson C E, Rivest R L. Introduction to algorithms. Cambridge, MA: MIT Press, 1990

[4] 周建饮，赵志远. 排序和查找理论及算法. 北京：科学出版社，1993

[5] 杨大顺，陶明华，丁青. 二次链接桶排序法. 计算机研究与发展，1996，33(120)：881－886

[6] 钟诚. 基于散列和归并技术的有效并行排序方法. 计算机工程与科学，1998，20(4)：42－45

[7] David R Musser. Introspective sorting and selection algorithms. Software-Practice and Experience, 1997, 27(8): 983 - 993

[8] Kunth D E. The art of computer programming, vol. 3: sorting and searching. Reding, MA: Addison Wesley Publishing Company, 1973(中译本：管纪文、苏运霖. 计算机程序设计技巧，第 3 卷：排序与查找. 北京：国防工业出版社，1984)

(上接 46 页)

(1)"所"字本身一般不重读；

(2)"所"字后继成分的词组基本按原读法，不会受"所"音节的影响。

五、结束语

本文就改善合成语音的自然度及其重要性进行了讨论，从正确界定连续文本中的韵律短语入手，分析了虚词的语法属性和韵律特性，着重研究了结构助词"的""地""得""所"在韵律短语界定中的作用，得到一组相应的规则。事实上，其它虚词在韵律短语界定中也有相应的规则，作者将另文讨论。在我们的汉语文语转换系统文本分析模块中，引入上述韵律短语界定的规则，给出较为适当的韵律短语分界，一定程度地改善了合成语音的自然度，赋予了语句发音的节奏感，获得了比较好的效果。

参 考 文 献

[1] 蔡莲红等. 汉语文语转换中的语言学处理. 中文信息学报，1995，9(1)

[2] Chiu - yu Tseng. Investigating mandarin Chinese prosody through speech database. Second International Workshop on East - Asian Language Resources and Evaluation. Taiwan, 1999, 5: 65 - 68

[3] 王厚峰等. 汉语句子结构分析算法研究. COMMUNICATIONS OF COLIPS, 1997, 17(2)

[4] 叶军. 停顿的声学征兆. 见：第三界全国语音学研讨会论文集，北京. 1996, 21 - 22

[5] 俞士汶等. 现代汉语语法信息词典详解. 北京：清华大学出版社，1998

[6] 朱德熙. 语法讲义. 北京：商务印书馆，1982