数据结构与算法“字符串”教学设计

北京大学信息科学技术学院　赵海燕

1．字符串在课程中的定位和前测知识点

字符串（String）是零个或多个字符的顺序排列所组成的复合数据结构，是一种特殊的线性表，其基本组成元素是单个字符。大多数程序设计语言都支持字符串这种数据类型。

字符串一章在主要介绍字符串的的基本概念，字符串的存储表示和类定义，字符串的运算，侧重于对于字符串的变长特性的处理；在此基础上，重点介绍了字符串的模式匹配。

前测知识点要求如下，可以视情况给学生补充：

（1） ASCII、UNICODE 等字符编码知识；

（2） C/C++程序设计语言中的字符串相关知识。

2．学习目标

（1） 理解字符的编码原则和编码顺序；

（2） 熟练掌握字符串的常用运算，理解 C++标准串运算的实现以及 String 类中如何处理字符串的变长特性；

（3） 学会如何快速进行字符串的模式匹配。

3．知识点和学时分配

理论授课 4 学时，建议安排实验 6-8 学时。

以下内容是本课程要求的基本教学内容，在授课中必须完全涵盖，授课时可以根据学生的状况、教师的科研背景等在某些方面进行扩展和对学生进行引导，以扩大适当学生的涉猎面。计算机专业的学生可以适当介绍除KMP之外的其他无回溯算法。

各知识点建议授课时间如下：

字符串的基本概念	20 分钟
字符串的存储和实现	60 分钟
字符串模式匹配	150 分钟

4．重点和难点

本章重点包括：

（1） 字符编码；

（2） 字符串类 class String 的存储结构；

（3） 朴素字符串模式匹配算法；

（4） 无回溯的 KMP 模式匹配算法。

其中，难点在于

（1） 字符串类 class String 对于字符串长度变化所作的适应性处理；

（2） 特征向量的概念与计算。

5．授课提示

下面是字符串一章的重点和难点内容的讲授注意事项。

（1）字符编码

组成字符串的基本单位是字符，字符可以是所使用的字符集中的数字、文本字符或者特定的符号。字符在计算机中是通过 0、1 组成的字节来表示的，这种表示特定字符集的“字

符”的“字节”即是该字符的编码。

根据要表示的字符集的大小和用途，字符有多种编码方式。常用的编码包括英文的ASCII 编码，中文的 GB2312-80 编码，繁体中文的 BIG5 编码，以及通用文字符号编码标准UNICODE。

授课时需强调字符编码方式的不同取决于具体的应用需要。字符串是一种元素为字符的特殊的线性表，字符编码的不同仅仅表示组成线性表的元素不同，并不改变字符串概念和操作的本质。

（2）字符串类 class String 的存储结构

字符串的一个显著且难以回避的特点是其长度动态变化，选择串的存储结构时需要考虑串的变长特点。静态长度的顺序存储很难适应诸如串的拼接、查找、置换等运算所导致的串长度的改变。

Class String 类采用一种动态变长的存储结构来存储字符串，根据运算的需要动态地改变字符串的长度，使得编程人员摆脱自己维护静态定长数组的重任。

讲授时宜采用字符串赋值或拼接等运算为例来介绍 class String 类的存储机制。

（3）朴素字符串模式匹配算法

本课程中所涉及的字符串模式匹配仅限于精确匹配中的单选情况：给定一个由字符或符号组成的字符串目标对象 T 和一个字符串模式 P，模式匹配的目的是在目标字符串 T 中搜索与模式 P 完全相同的子串，并返回 T 和 P 匹配的第一个子串的首字符位置。

模式匹配的一个简单方法就是把模式 P 的字符依次与目标 T 的相应字符进行比较。从首字符开始，依次将两个串对应位置上的字符进行比较。当某次比较失败时，则把模式 P 相对于 T 右移一个字符位置，重新开始下一趟匹配。如此不断重复，直到某趟配串成功返回；或是比较到目标串的结束也没有出现“配串”的情况，则匹配失败。

授课时需引导学生分析朴素匹配算法的时间代价，并找出其最佳、最差、及平均情况下的开销，同时引导学生分析和总结朴素算法之所以时间代价较大的症结所在，最好辅以实际的匹配例子来说明，能够采用动画的话更好。

（4）KMP 算法中特征向量的含义与计算

朴素模式匹配算法存在的一个问题是，一旦某趟匹配中发生失配，无论模式的具体情况如何，都采用模式右移一位的方式开始下一趟的匹配，这可能导致很多冗余的比较。Knuth、Morris、Pratt等人发现在模式匹配失配后，模式 P 相对于目标 T应该右移的位数存在且与目标串无关，仅依赖于模式本身，他们对朴素的模式匹配算法进行了改进得到了KMP算法，其基本思想为：预先处理模式本身，分析其字符分布状况，并为模式中的每一个字符计算失配时应该右移的位数，这即为字符串的特征向量。

授课时须强调特征向量的计算过程本身即是一个模式匹配的过程，采用相同的匹配方法。另外，应该着重介绍一下计算特征向量时的优化考虑，以帮助学生了解算法设计的精微之处。建议采用图示或动画来辅助说明特征向量的计算。

6．课后练习和实习

作为教学的重要一环，课后练习和上机实习帮助学生巩固课堂所学理论知识，训练学生创新思维能力训练、工程实践能力。

字符串部分可以安排 4 道书面作业，可酌情布置有关字符串的综合训练，以帮助学生充分巩固课堂所学知识，课程网站上字符串综合训练题目如下：

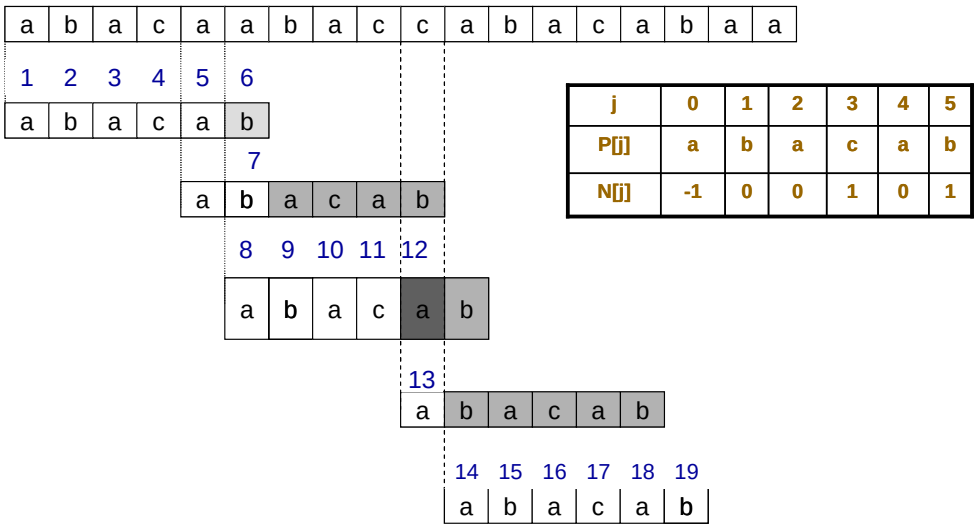
http://www.jpk.pku.edu.cn/pkujpk/course/sjjg/report/HWTest/2007Proj/P1_Edline/P1.doc

7．教学案例

以 KMP 模式匹配算法为例。在计算了特征向量之后， KMP 模式匹配算法基于特征分析进行快速的模式匹配，在每趟匹配过程中如果发现某次失配时，不再单纯地把模式右移一位，而是根据当前的字符的特征数来决定模式右移的位数，具体实现参见下面的算法，其中第 3 个参数传递的是模式的特征向量。

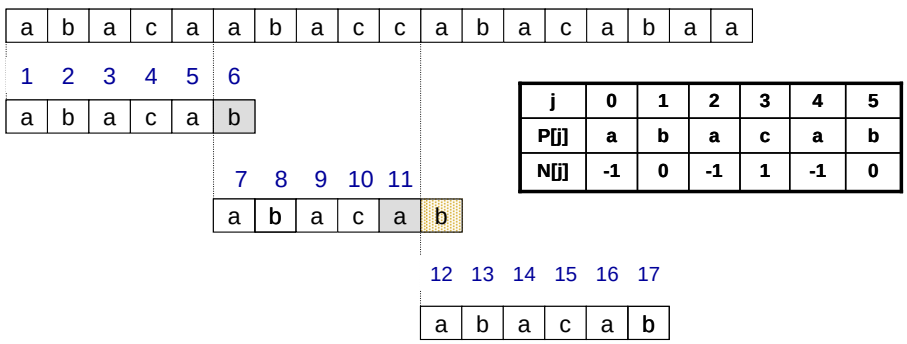
```
int KMPStrMatching(const String & T, const String & P,    int *N) {
      int i = 0;                                          // 模式的下标变量
      int j = 0;                                          // 目标的下标变量
      int pLen = P. length ( );                           // 模式的长度
      int tLen = T.length( );                             // 目标的长度
      if (tLen < pLen)                                    // 如果目标比模式短，匹配无法成功
            return (-1);
      while ( i < pLen    &&    j < tLen)    {            // 反复比较对应字符来开始匹配
            if ( i == -1    ||    T[j] == P[i])
                  i++,    j++;
            else i = N[i];
      }
      if ( i > = pLen)
            return (j – pLen+1);
      else return (-1);
 }
```

整个匹配的过程如下面的两个图示，前者中所用的特征向量未经优化，所以总体的比较次数为 19 次，而后者基于优化后的特征向量，消除了冗余的比较，比较次数缩减为 17 次。



j	0	1	2	3	4	5
P[j]	a	b	a	c	a	b
N[j]	-1	0	0	1	0	1

采用没有优化的特征向量



j	0	1	2	3	4	5
P[j]	a	b	a	c	a	b
N[j]	-1	0	-1	1	-1	0

采用优化后的特征向量

授课时除了采用动画来演示匹配过程之外，建议进一步引导学生对其进行算法复杂性分析。假设目标串长为 n, 模式串长为 m.上述 KMP 算法的复杂性主要体现在 while 循环语句中。由于 j 只增不减的特性，循环体中的 j++; 语句的执行次数最多为 n 次，所以与此在同一语句中的运算 i++也不会超过 n 次。i 的初值为 0，其间能使之减少的语句只有 i = N[i];，因为 N[i] < i，所以每执行一次该语句 i 值至少减 1，但是一旦使 i = -1，则下一步 if 语句的条件必成立，进而执行 i++, j++; 故循环体中 i = N[i]的执行次数不会超过 i++, j++;语句的执行次数加 1。所以整个循环体的执行次数至多为 2n+1 次，即其时间代价与目标串的长度成线性关系。由于计算 next 数组的算法本身采用的也是 KMP 模式匹配，因此其时间代价也与模式长度成正比，为 O(m)。所以整个 KMP 匹配的时间代价为 O(n+m)。

8．总结

字符串应用非常广泛的，是非数值处理中主力。本章讨论了字符串这一数据结构，首先介绍了字符串的基本概念和抽象数据类型，然后描述了字符串的存储结构和类定义，最后重点讨论了字符串模式匹配。

参考文献：

1. 张铭，王腾蛟，赵海燕，《数据结构与算法》，高等教育出版社，2008 年 6 月。——普通高等教育“十一五”国家级规划教材。
2. 张铭、赵海燕、王腾蛟、宋国杰 、高军，北京大学“数据结构与算法”教学设计，《计算机教育》2008 第 20 期。获得“英特尔杯 2008 年全国计算机教育优秀论文评比”一等奖。
3. 北京大学《数据结构与算法》精品课程网站（2008 年北京市“精品课程”暨国家“精品课程”）， http://www.jpk.pku.edu.cn/pkujpk/course/sjjg/