数据结构与算法（四）

张铭 主讲

采用教材：张铭，王腾蛟，赵海燕 编写
高等教育出版社，2008. 6（“十一五”国家级规划教材）

https://pkumooc.coursera.org/bdsalgo-001

主要内容

- 字符串基本概念
- 字符串的存储结构
- 字符串运算的算法实现
- 字符串的模式匹配
  - 朴素算法
  - KMP 快速模式匹配

字符串示例

- s1=“123”
- s2=“ABBABBC”
- s3=“BB”
- s4=“BB ”
- s5=“Hello World!”
- s5=“”

4.1 字符串基本概念

- 字符串，特殊的 线性表，即元素为 字符 的线性表
- n（≥ 0）个字符的有限序列，n ≥ 1时，一般记作

S：“$c_0c_1c_2\ldots c_{n-1}$”

  - S 是串名字
  - “$c_0c_1c_2\ldots c_{n-1}$”是串值
  - c_i 是串中的字符
  - N 是串长（串的长度）：一个字符串所包含的字符个数
    - 空串：长度为零的串，它不包含任何字符内容（注意与 空格串 “ ” 的区别）

字符串是一种特殊的线性结构

- 数据对象
  - 无特殊限制
  - 串的数据对象为字符集
- 基本操作
  - 线性表的大多以“单个元素”为操作对象
  - 串通常以“串的整体”作为操作对象
- 线性表的存储方法同样适用于字符串
  - 应根据不同情况选择合适的存储表示

字符/符号

- 字符 (char) ：组成字符串的基本单位
- 取值依赖于字符集 Σ（同线性表，结点的有限集合）
  - 二进制字符集：Σ = {0,1}
  - 生物信息中 DNA 字符集：Σ = {A,C,G,T}
  - 英语语言： Σ = {26个字符，标点符号}
  - ……

字符编码

- 单字节（8 bits）
  - 采用 ASCII 码对 128 个符号进行编码
  - 在 C 和 C++ 中均采用

- 其他编码方式
  - GB
  - CJK
  - UNICODE

字符编码顺序

- 为了字符串间比较和运算的便利，字符编码表一般遵循约定俗成的“偏序编码规则”

- 字符偏序：根据字符的自然含义，某些字符间两两可以比较次序
  - 其实大多数情况下就是字典序
  - 中文字符串有些特例，例如“笔划”序

字符串的数据类型

- 因语言而不同
  - 简单类型
  - 复合类型

- 字符串常数和变量
  - 字符串常数 (string literal)
    - 例如： “\n”, “a”, “student”...
  - 字符串变量

子串(Substring)

- 子串 定义

  假设 s_1，s_2 是两个串：

  $s_1 = a_0a_1a_2\ldots a_{n-1}$

  $s_2 = b_0b_1b_2\ldots b_{m-1}$

其中 $0 \le m \le n$, 若存在整数 i $(0 \le i \le n-m)$，使得 $b_j = a_{i+j}$，$j = 0,1,\ldots,m-1$ 同时成立，则称 串 s_2 是串 s_1 的 子串，s1 为串 s2 的主串，或称 s_1 包含串 s_2

- 特殊子串
  - **空串是任意串的子串**
  - 任意串 S 都是 S 本身的子串
  - 真子串：非空且不为自身的子串

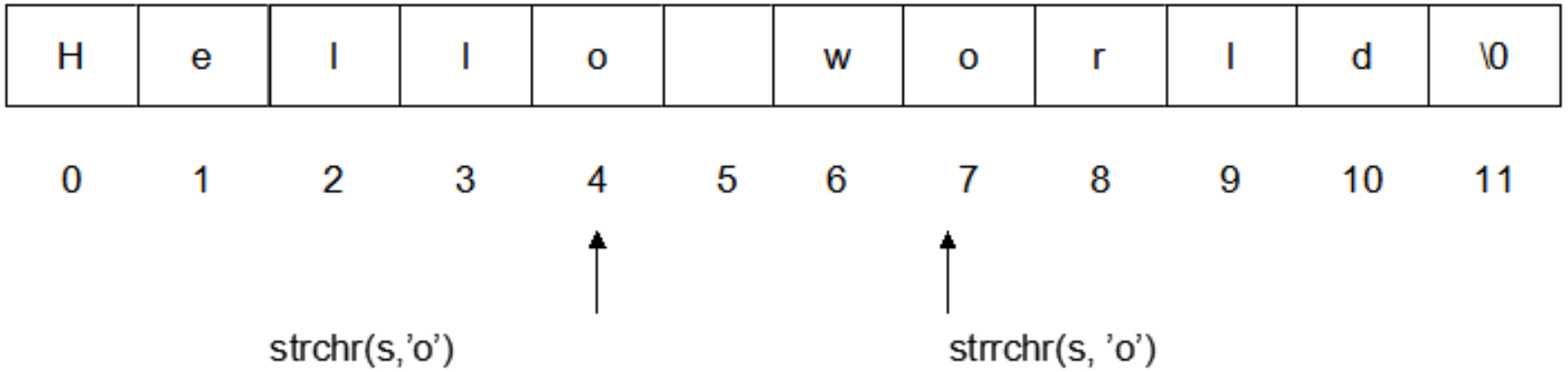
字符串的基本运算

C 标准函数库需要 #include <string.h>

- 求串长 int strlen(char *s);
- 串复制 char *strcpy(char *s1, char*s2);
- 串拼接 char *strcat(char *s1, char *s2);
- 串比较 (注意)
  - int strcmp(char *s1, char *s2);
  - 看 ASCII 码，s1>s2，返回值 > 0；两串相等，返回 0
- 定位 char *strchr(char *s, char c);
- 右定位 char *strrchr(char *s, char c);
- 求子串 char *strstr(const char *str1, const char *str2);

定位函数示例

- 字符串 s：



- 寻找字符 o, strchr(s,'o') 结果返回 4
- 反方向寻找 r, strrchr(s,'o') 结果返回 7

String抽象数据类型

C++标准字符串类库

```
#include <string>
using namespace std;
```

- 字符串类（class String）
  - 适应字符串长度动态变化的复杂性
  - 不再以字符数组 char S[M] 的形式出现，而采用一种动态变长的存储结构

C++ String 部分操作列表

操作类别	方法	描述
子串	substr ()	返回一个串的子串
拷贝/交换	swap ()	交换两个串的内容
	copy ()	将一个串拷贝到另一个串中
赋值	assign ()	把一个串、一个字符、一个子串赋值给另一个串中
	=	把一个串或一个字符赋值给另一个串中
插入/追加	insert()	在给定位置插入一个字符、多个字符或串
	append () / +=	将一个或多个字符、或串追加在另一个串后
拼接	+	通过将一个串放置在另一个串后面来构建新串
查询	find ()	找到并返回一个子序列的开始位置
替换/清除	replace ()	替换一个指定字符或一个串的字串
	clear ()	清除串中的所有字符
统计	size () / length()	返回串中字符的数目
	max_size ()	返回串允许的最大长度

得到字符串中的字符

- 重载下标运算符[ ]

```
char& string::operator []  (int  n);
```

- 按字符定位下标

```
int string::find(char c,int start=0);
```

- 反向寻找，定位尾部出现的字符

```
int string::rfind(char c, int pos=0);
```

思考

- 1. 判断哪些是”software”的子串
  - 空串、software、soft、oft...
  - fare、sfw...
- 2. 若字符串 s =“software”，则其子串的数目为？

数据结构与算法

谢谢聆听

国家精品课“数据结构与算法”

https://pkumooc.coursera.org/bdsalgo-001

张铭，王腾蛟，赵海燕

高等教育出版社，2008. 6。“十一五”国家级规划教材