字符串相关算法

主讲：李跃

主要内容

- KMP算法
- 字符串HASH
- 最小表示法
- TRIE树
- AC自动机
- 有限状态自动机DFA
- 后缀数组

什么是串

- 由字母、符号组成的线性表
- 与串有关的问题及算法在实际应用中非常广泛
- 例如 文本搜索

与串有关的概念

- 长度
- 字符集
- 前缀
- 后缀
- 字典序

模式匹配

- 主串（text）
- 模版串（pattern）
- 朴素算法
- KMP算法
- Maxtrix67 blog KMP算法详解
- http://www.matrix67.com/blog/archives/115

KMP重点： next函数的意义

（1）next[0]= -1 意义：任何串的第一个字符的模式值规定为-1。

(2)next[j]=k 意义：模式串T中下标为j的字符，如果j的前面k个字符与开头的k个字符相等,且T[j] != T[k] （1≤k<j）。即T[0]T[1]T[2]。。。T[k-1]==T[j-k]T[j-k+1]T[j-k+2]…T[j-1]且T[j] != T[k].（1≤k<j）；

KMP重点：next函数的意义

- next[i]=k 表示i的前有k个字符与0--(k-1)相同，如
- s []= “ a b c a b c a b c \0”
- next[]= { -1 0 0 0 1 2 3 4 5 6 }

例题

- 字串寻址
- http://222.196.33.254/JudgeOnline/showproblem?problem_id=1139

Hash是什么？

- 将某个对象对应到一个关键值，然后通过关键值归类，放入到一个表中（哈希表），今后可以根据关键值迅速查找
- Hash可以用来判重和统计数目

字符串Hash

- Hash中最常用的是字符串Hash
- 将一个字符串对应到一个整型数值，插入到哈希表
- 对应方法有很多种，甚至可以根据问题的特殊性自己构造，常用的有Rabin-Karp，ELFHash

Rabin-Karp

- 如果字符串中可能出现的字符有k个，则可以将字符串对应到k进制数
- 例如，如果字符串只可能为小写字母组成，则acm就对应到0×26^2+2*26+12
- log(2^63)/log(26) =13.403001373861878677 19
- 当字符串长度不超过13的时候，用long long作关键值类型，加上字符串长度作为限制，每个字符串唯一对应关键值
- 当字符串长度超过13的时候，就要进一步验证

LEFhash算法

- ELFhash: 黑书 P96　1.4.3

```
int ELFhash(char *key)
{
    unsigned long h=0;
    while(*key)
    {
        h=(h<<4)+*key++;
        unsigned long g=h&0Xf0000000L;
        if(g) h^=g>>24;
        h&=~g;
    }
    return h%MOD;
}
```

Hash冲突的处理

- 不同的字符串可能映射到同一个key值。
- （1） 开放地址法。
- （2）拉链法。

最小表示法

- 2003年 冬令营 周源 论文

TRIE树

- 又称字典树
- 可用于字典中单词的查找
- 优点：节省查找时间
- 缺点：字符集太大时 空间耗费大

AC自动机

- 多模式匹配
- AC自动机=Trie树+KMP
- AC自动机算法详解

  http://www.cppblog.com/mythit/archive/2009/04/21/80633.html

用途与实现

- 有一个单词集 以及一个文本 要求找出每个单词出现的次数及位置
- 计数问题
- 实现：
- 1.构建TRIE树
- 2.通过一次BFS连后向边
- 例题：hdu 2222
- http://acm.hdu.edu.cn/showproblem.php?pid=2222

有限状态自动机DFA

- 常见于与字符串相关的计数问题
- 多于组合计数、DP、矩阵相乘相结合
- 矩阵相乘须二分相乘

- 例题：http://222.196.33.254/JudgeOnline/showproblem?problem_id=1585

后缀数组

- 2004年论文　许智磊：《后缀数组》
- 后缀：
- 后缀数组SA：
- Rank数组
- Height数组：
- 例题：　POJ 2774

练习

- POJ上的字符串题目：
- kmp
  2752 2406 1961 2185
- Hash & tire树
  1200 2503 3007 2001 2513
- AC自动机
  1204
- 后缀数组
  1743 2774 3450 3617 3623 3415 3080