

第9章 符号表

- 实现符号表的简单方法
- 用散列表实现符号表
 - 开散列
 - 闭散列
 - 散列函数及其效率
 - 闭散列的重新散列技术
- 应用举例



第9章 符号表

- 在算法设计中用到的集合,往往不做集合的并、交、差运算,而常要判定某个元素是否在给定的集合中,并且要不断地对这个集合进行元素的插入和删除操作。
- 以集合为基础,并支持SetMember、SetInsert和 SetDelete三种运算的抽象数据类型叫做符号表。
- ADT符号表的应用
 - ■公司的客户字典
 - 一个地区的固定/移动电话号码字典
 - 软件开发中的数据字典
 - 网上的在线字典
 - 互联网/企业网/局域网网管中的IP地址字典等等

2



■9.1 实现符号表的简单方法——用固定数组

■ 用数组实现符号表的结构定义如下:

```
typedef struct atab *Table;
typedef struct atab{
   int arraysize;
   int last;
   SetItem*data;
}Atab;
```



■创建一个定长数组大小为size的空符号表

```
Table TableInit(int size)
{
    Table T=malloc(sizeof *T);
    T->arraysize=size;
    T->last=0;
    T->data=malloc(size*sizeof(SetItem));
    return T;
}
```



■符号表的成员查询运算

```
int TableMember(SetItem x,Table T)
{
  int I;
  for(i=0;i<T->last;i++)
  if(T->data[i]==x) return 1;
  return 0;
}
```



■符号表的元素插入运算

```
void TableInsert(SetItem x,Table T)
{
  if(! TableMember(x,T) && T->last<T->arraysize)
    T->data[T->last++]=x;
}
```



■符号表的元素删除运算

```
void TableDelete(SetItem x,Table T)
{
  int i=0;
  if (T->last>0){
    while(T->data[i]!=x && i<T->last) i++;
    if(i<T->last && T->data[i]==x)
     T->data[i]=T->data[--T->last];
  }
}
```



■9.1 实现符号表的简单方法——用固定数组

- 优点:结构简单,实现操作的逻辑简单。
- 缺点:
 - 所表示的集合的大小受到数组大小的限制。
 - 三个基本操作在最坏情况下都需要*O(n)*。
 - 通常集合元素并不占满整个数组,因此,空间没有得到 充分利用。



■ 9.2 用散列表实现符号表

- ■实现符号表的另一个重要技巧是散列技术。
- 用散列来实现符号表可以使符号表的每个运算所需的平均时间是一个常数值,在最坏情况下每个运算所需的时间 正比于集合的大小。
- 散列有两种形式:
 - 开散列: 它将符号表元素放在一个潜无穷的空间里, 能处理任意大小的集合。
 - 闭散列: 它使用一个固定大小的存储空间,所能处理的集 合大小不能超过其存储空间大小。



→ 9.2.1 开散列

- 基本思想:
 - 把符号表中的所有元素散列(hashing),即映射到一个桶数组(散列表)的桶中。当有多个不同元素被散列到同一个桶时,这些元素用链在一个表里。期望散列能均匀,使得当桶数组的规模与符号表的规模同阶时,桶数组的每一个桶中大致有常数个元素,从而对符号表的三个基本操作都只需要常数时间。
 - 这里的问题是如何散列即如何构造散列(映射)函数去达到设想的期望?



→ 9.2.1 开散列

- ■开散列的数据要素
 - ■按照设想,开散列首先需要拥有一个桶数组,桶数组的规模与符号表的规模同阶,桶数组中的每一个桶有一个指针各指向一个链表。其次需要拥有一个散列(映射)函数,它把符号表中的元素分别散列(映射,分散)到各桶所对应的链表中。



→ 9.2.1 开散列

- ■散列函数的例子
 - 假设符号表的元素是字符串x,桶数组的规模为101,那么,下面是散列函数的一个具体例子:

```
int hash1(char* x)
{
  int len=strlen(x), hashval = 0;
  for(int i=0;i< len ;i++)
    hashval += x[i];
  return hashval % 101;
}</pre>
```



■用开散列实现的符号表结构OpenHashTable的定义

```
typedef struct open *OpenHashTable;
typedef struct open{
  int size; //桶数组的大小
  int (*hf)(SetItem x); //散列函数
  List *ht; //桶数组
}Open;
```



■ 创建一个空的开散列表,其桶数组的大小为nbuckets, 散列函数为hashf(x)

```
OpenHashTable HTInit(int nbuckets,int(*hashf)(SetItem x)))
 int i;
 OpenHashTable H=malloc(sizeof *H);
 H->size=nbuckets;
 H->hf=hashf;
 H->ht=malloc(H->size*sizeof(List));
 for(i=0;i<H->size;i++)
 H->ht[i]=ListInit()
 return H;
```



■ 开散列表的成员查询函数HTMember(x,H),根据元素x的散列函数值确定存储该元素的桶号,然后调用相应的表定位函数返回查询结果。

```
int HTMember(SetItem x,OpenHashTable H)
{
  int i=(*H->hf)(x) % H->size;
  return (ListLocate(x,H->ht[i])>0);
}
```



■ 开散列表的元素插入运算HTInsert(x,H),根据元素x的散列函数值确定存储该元素的桶号,然后在该桶的表首插入元素x。

```
void HTInsert(SetItem x,OpenHashTable H)
{
  int i;
  if (HTMember(x,H)) Error("Bad Input");
  i=(*H->hf)(x) % H->size;
  ListInsert(0,x,H->ht[i]);
}
```



■ 开散列表的元素删除运算HTDelete(x,H),根据元素x的散列函数值确定存储该元素的桶号,然后调用相应的表元素删除函数删除元素x。

```
void HTDelete(SetItem x,OpenHashTable H)
{
  int i;
  i=(*H->hf)(x) % H->size;
  if (k=ListLocate(x,H->ht[i]))
  ListDelete(k,H->ht[i]);
}
```



→ 9.2.2 闭散列

- ■与开散列的相同点和区别
 - ■相同点: 把符号表中的所有元素散列(hashing)即映射 到一个桶数组(散列表)的桶中。
 - ■区别:桶数组(散列表)的桶直接用来存放符号表元素,而且一个桶只存放一个元素。出现多个元素散列到同一个桶时(这叫冲突),需要按照确定的规则在桶数组中进行探测,找还没有存放元素的桶(这叫空桶),然后将发生冲突的后到元素放入空桶,解决冲突,实现散列。



→ 9.2.2 闭散列

- ■闭散列引发的问题
 - (1) 需要一个探测函数c(i),i=0,1,2,...,size-1: c(0)=0,而且对于任意的0≤j≤size-1,{(j+c(0))% size, (j+c(1))% size, ...,(j+c(size-1))% size}是 {0,1,2,...,size-1} 的重排,保证不会漏测。
 - (2) 需要对ht[]的每一个桶的状态加以标记:
 - state[k]=0表示ht[k]桶存放着元素。
 - state[k]=1表示ht[k]桶一直是空桶。
 - state[k]=2表示ht[k]桶现在是空桶但曾经存放过元素



■用闭散列实现的符号表结构HashTable的定义

```
typedef struct hashtable *HashTable;
typedef struct hashtable{
  int size; //桶数组大小
  int (*hf)(SetItem x); //散列函数
  SetItem *ht; //桶数组
  int *state; //占用状态数组
}Hashtable;
```



■ 初始化桶数组ht和state,将每个桶都设置为空桶。 创建一个空的散列表

```
HashTable HTInit(int divisor, int (*hashf)(SetItem x))
  int i;
  HashTable H=malloc(sizeof *H);
  H->size=divisor;
  H->hf = hashf;
  H->ht=malloc(H->size*sizeof(SetItem));
  H->state=malloc(H->size*sizeof(int));
  for (i=0;i<H->size;i++)
   H->state[i]=1;
  return H;
```



■ 在散列表H的桶数组中查找元素 x,并返回它在桶数组中的位置。

```
int FindMatch(SetItem x, HashTable H)
{
    int I,j,k;
    j=(*H->hf)(x);
    for (i=0;i<H->size;i++){
    k=(j+HashProb(i))%H->size;
    if (H->state[k]==1) break;
    if(!H->state[k] \&\& H->ht[k]==x) return k;
    return H->size;
int HashProb(int i)
  return i;
```



■ 返回散列表H的桶数组中可存储元素x的未占用桶单元位置k。

```
int Unoccupied(SetItem x, HashTable H)
 int I,j,k;
 j=(*H->hf)(x);
 for (i=0;i<H->size;i++){}
   k=(j+HashProb(i)) % H->size;
   if (H->state[k]) return k;
 return H->size;
```



■ 闭散列表通过调用函数FindMatch实现成员查询

```
int HTMember(SetItem x,HashTable H)
{
   int i=FindMatch(x,H);
   if (i<H->size && H->ht[i]==x) return 1;
   return 0;
}
```



■闭散列表的元素插入运算

```
int HTInsert(SetItem x, HashTable H)
  int i;
  if(HTMember(x,H)) Error("Bad Input");
  i=Unoccupied(x,H);
  if(i<H->size){
    H->state[i]=0;
    H->ht[i]=x;
  else Error("table full");
```



■闭散列表的元素删除运算

```
int HTDelete(SetItem x,HashTable H)
{
  int i=FindMatch(x,H);
  if (i<H->size && H->ht[i]==x)
    H->state[i]=2;
}
```



- 上述删除算法HTDelete的缺点:
 - 在执行了大量元素删除运算后,大量的桶的状态标记为
 - 2,即大量的桶的状态标记既非 1 也非 0,使得运算 FindMatch 中的循环次数大大增加,从而导致运算 FindMatch的速度大大减慢。因此人们提出HTDelete的

一种改进版本HTDelete1。

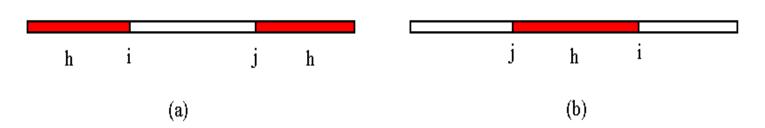


■ 改进HTDelete的基本思想:

- 动机是希望腾出的空桶(记为ht[i])不仅仅可供新元素插入,而且能为提高还在桶数组中的元素(比如y= ht[j])的查找速度作出贡献:减少查找y的探测次数。
- 容易理解,如果不作任何改进,查找y的探测次数会等于插入y时的探测次数。如果插入y时x已在桶ht[i]中且与x发生冲突而增加了插入的探测次数,那么,现在x不存在了,只要将x腾出的桶ht[i]让y顶替,就可以使得将来查找y时减少探测次数。
- 因此,改进HTDelete的途径是在当前的桶数组中找能顶替x的y。如果找不到,就按HTDelete的原版处理;如果找得到,在用y顶替x之后还可以引起连锁反应。



- 改进HTDelete的细节——找能顶替x的y
 - 假设被删除元素x位于桶单元ht[i]。现考察一个非空单元 ht[j]中的元素y,其散列函数值设为h=hf(y),则按从h出发 的线性探测,只要i比j离h近即可使得在顶替后找y的探测次 数减少。
 - ①当i<j时,若i<h≤j,则不可用元素ht[j]顶替ht[i];若h≤i<j或 i<j<h,则可用元素ht[j]顶替ht[i]。如下图(a)。
 - ②当j<i时,若j<h≤l,则可用元素ht[j]顶替ht[i],如下图 (b);否则不可。
 - 这里以线性探测为前题,以顶替后减少探测次数为目标。





■ 改进HTDelete的 函数——HTDelete1

```
Void HTDelete1SetItem x, HashTable H)
  int i=FindMatch(x);
  if (i<H->size) // &&H->ht[i]==x可以去掉
  for (;;) { //找ht[i]的顶替元素
    H->state[i]=2; //先按无顶替元素处理
    int j; //从(i+1)%size开始线性搜索可顶替元素
    for (j=(i+1)\%H->size; !state[j]; j=(j+1)\%H->size)
     { int h=(*H->hf) (H->ht[i]);
      if ((h \le i \& \& i \le j) || (i \le j \& \& j \le h) || (j \le h \& \& h \le i)) break;
        //ht[j]可顶替ht[i],跳出内for循环,做顶替工作
        // ht[j]不可顶替ht[i],继续线性搜索可顶替元素
    if (H->state[j]) break; //跳出外for循环
    H-> ht[i]=ht[j]; H->state[i]=state[j]; //做顶替工作
    i=j; //为构成循环找ht[j]的顶替元素
```



→ 9.2.3 散列函数及其效率

- 若能选择一个好的散列函数,将集合中的n个元素均匀地散列到B个桶中,那么每个桶中平均有n/B个元素。使得在开散列表中,HTInsert,HTDelete和HTMember运算都只要O(n/B)的平均时间。当n/B为一常数时,每一个符号表运算都可在常数时间内完成。
- 因此:对于开散列表,关键在于选择一个好的散列函数。



→ 9.2.3 散列函数及其效率

- 几种计算简单且效果较好的散列函数构造方法:
 - (1)除余法: h(k)=k%m
 - (2)数乘法: h(k)= [B(ka-[ka])]
 - (3)平方取中法: h(k)= \[\langle k^2/c \] %B
 - (4)基数转换法
 - (5)随机数法: h(k)=random(k)



→ 9.2.4 闭散列的重新散列技术

■ (1) 二次重新散列技术 它选取的探查桶序列为: $h(x),h_1(x),h_2(x),...,h_{2i}(x),h_{2i+1}(x),$ 其中, $h_{2i}(x) = (h(x) - i^2)\%B$ $h_{2i-1}(x) = (h(x) + i^2)\%B$

■ (2) 随机重新散列技术

它选取的探查桶序列为: $h_i(x) = (h(x) + d_i)\%B$, i=1,2,...,B-1。 其中, d_1, d_2, \dots, d_{B-1} 是1,2,...,B-1的一个随机排列。

- (3) 双重散列技术
 - 这种方法使用2个散列函数h和h'来产生探索序列:

$$h_i(x) = (h(x) + ih'(x))\% B$$

其中 i=1, 2, ..., B-1。要求h'(x)的值和B互素。



9.3 应用举例:字符串频率统计问题

- P199
- P201习题9.1

34



本章小结

- 定长数组实现符号表
- 用散列表实现符号表
 - 开散列
 - 闭散列
 - 散列函数及其效率
 - 闭散列的重新散列技术