如何读懂后缀数组代码

北京 省选

张若天 me@zrt.io

2018 年 1 月 20 日

清华大学 交叉信息研究院

```
void mk_sa(int n,int m){
    int *x=t,*y=t2;
    for(int i=0;i<m;i++) c[i]=0;
    for(int i=0;i<n;i++) c[x[i]=s[i]]++;
    for(int i=1;i<m;i++) c[i]+=c[i-1];
    for(int i=n-1;i>=0;i--) sa[--c[x[i]]]=i;
    for(int k=1;k<=n;k<<=1){
        int p=0;
        for(int i=n-k;i<n;i++) y[p++]=i;
        for(int i=0;i<n;i++) if(sa[i]>=k) y[p++]=sa[i]-k;
        for(int i=0;i<m;i++) c[i]=0;
        for(int i=0;i<n;i++) c[x[y[i]]]++;
        for(int i=1;i<m;i++) c[i]+=c[i-1];
        for(int i=n-1;i>=0;i--) sa[--c[x[y[i]]]]=y[i];
        swap(x,y);
        p=1;x[sa[0]]=0;
        for(int i=1;i<n;i++){
            x[sa[i]]=y[sa[i]]==y[sa[i-1]]&&y[sa[i]+k]==y[sa[i-1]+k]?p-1:p++;
        }
        if(p>=n) break;
        m=p;
    }
}
```

第 0 行解读

```
void mk_sa(int n,int m){
```

n: 需要排序的字符串长度，字符串使用 0 到 n-1。

m: 初始时字符集大小。

第 2-5 行解读

```
for(int i=0;i<m;i++) c[i]=0;
for(int i=0;i<n;i++) c[x[i]=s[i]]++;
for(int i=1;i<m;i++) c[i]+=c[i-1];
for(int i=n-1;i>=0;i--) sa[--c[x[i]]]=i;
```

x[i] 存储了第一关键字 (s[i] 的值)。

第 4 个 for 反着循环为了排序的稳定性。(虽然这次不重要)

sa 最后存的是通过 1 个字符辨别的顺序。

第 6 行解读

```
for(int k=1;k<=n;k<<=1)
```

k 倍增，代表位置 i 比较 (x[i],x[i+k]) 组成的二元组。

第 8-9 行解读

```
for(int i=n-k;i<n;i++) y[p++]=i;
for(int i=0;i<n;i++) if(sa[i]>=k) y[p++]=sa[i]-k;
```

y 存储了第二关键字的顺序。

就是第二关键字排第 i 的为 y[i]。

大于等于 n-k 的排到前面，因为 i+k 大于等于 n。

其他元素第二关键字为 x[i+k]，他们原来 (x[i] 时) 的顺序是 sa[i]，所以现在是 sa[i]-k。

第 10-13 行解读

```
for(int i=0;i<m;i++) c[i]=0;
for(int i=0;i<n;i++) c[x[y[i]]]++;
for(int i=1;i<m;i++) c[i]+=c[i-1];
for(int i=n-1;i>=0;i--) sa[--c[x[y[i]]]]=y[i];
```

第一关键字是 x[i]。

依然是四个 for 的基数排序 (在第二关键字 y 基础上对第一关键字 x 排序)，不过和第一次比 x[i] 变成了 x[y[i]]。(第一次相当于 y[i]=i)。

比如把 a 数组 3,1,2,5,4 排成有序，y 可以等于 2,3,1,5,4。这样访问 a[y[i]] 相当于有序了。

第 14-行解读

```
swap(x,y);
p=1;x[sa[0]]=0;
for(int i=1;i<n;i++){
    x[sa[i]]=
        y[sa[i]]==y[sa[i-1]]
            &&
        y[sa[i]+k]==y[sa[i-1]+k]
            ?
            p-1:p++;
}
if(p>=n) break;
m=p;
```

这里 x,y 互换了。

然后重新计算 x 数组 (下回合的第一关键字)。

x[sa[0]]=0, 位置 sa[0] 的关键字为 0。

Questions?

谢谢大家！

Email: me@zrt.io

QQ: 401794301

zrt.io

LaTeX