

Strings?

Starria 扯淡时间

§ 0.1.写在前面

Wrong Answer

>这里是Starria/dsfz_ltx

>今天的内容是【字符串】

>光老师说应该选自己不擅长的讲，然后我就相信了他的鬼话

>当时我报上去的是{有限自动机理论;SAM;PAM;鏹字符串}

>非常遗憾，这些我今天一个也不讲

>前面那个我说不明白，中间两个除了我大家都会，后面那个我没学会

>如讲课过程中有错漏之处…请不要大意地指出。

>~~这个ppt是我昨天晚上10点开始做的，我保证它不会没有错~~
~~（雾）~~

Compile Error Time Limit Exceeded

§ 0.2. 目录

Wrong Answer

>这次讲的东西主要源于近年集训队论文，并包含一些大家较为陌生的字符串算法（所以很多自动机就不讲了）。内容比较杂，各小节间也没有依赖关系（所以可以选听/不听x

1. 后缀数组(DC3)
2. 三种后缀数据结构(?)的相互转化
3. 后缀平衡树

>应该是讲不满的。所以与当前内容相关/无关的问题都可以问。
（不然最后剩个2h就很尴尬）

Compile Error Time Limit Exceeded

§ 1.1. Warming-up

Wrong Answer

>如果要放个例题的话...

“给出一个字符串，求它的sa,rk和height数组。”

>（光速逃

>大家可以进入一下状态了。

Compile Error Time Limit Exceeded

§ 1.1. Warming-up

Wrong Answer

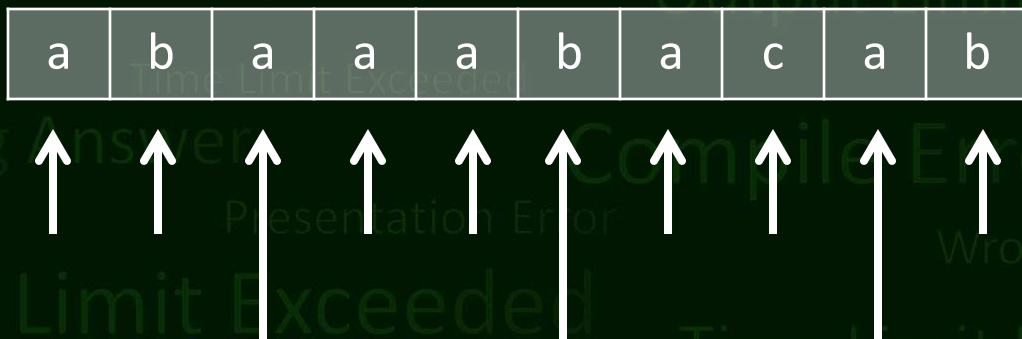
- > DC3是一个线性复杂度构造后缀数组的算法。
- > 那么为什么有 $O(n)$ 的算法 我们还一直在学 $O(n\log n)$ 的倍增呢
- > 因为有人说DC3比倍增…要慢一点?
- > 不过还有人说有的题DC3可过而倍增不可…
- > 接下来的过程将解答一个问题: 为什么DC3既不叫DC2也不叫DC4

Compile Error Time Limit Exceeded

§ 1.2. 一个过程

Wrong Answer

- > 首先我们有一个串 $S[1..n]$ 。(本例中 $n=10$)
- > 我们把该串的所有后缀按编号是否整除3分为两类。



- > 这两类我们分别处理。

Compile Error Time Limit Exceeded

§ 1.2. 一个过程

Wrong Answer

- > 把 $S[1..n]$ 和 $S[2..n]$ 拿出来，接到一起。
- > 如果这两个串中的任意一个长度不被3整除，在后面补0。
- // 如果前面那个串长度被3整除，就补3个0，不然边界不好处理。

a	b	a	a	a	b	a	c	a	b	0	0	b	a	a	a	b	a	c	a	b
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

- > 接下来对这个长为 $2n$ 的串做三个字符一组的基数排序。
- ~~不要以为不用倍增就不需要基数排序~~

Compile Error Time Limit Exceeded

§ 1.2. 一个过程

Wrong Answer

a	b	a	a	a	b	a	c	a	b	0	0	b	a	a	a	b	a	c	a	b
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

2	1	3	4	5	2	6
---	---	---	---	---	---	---

>这是排好后的样子。

>容易发现这2/3个合并后的元素分别对应了

$S[1..n], S[4..n], S[7..n], S[10..n], S[2..n], S[5..n], S[8..n]$ 。

>也就是我们求出了长度模3不为0的后缀的rk...

>等等，这显然没排完吧

Time Limit Exceeded

§ 1.2. 一个过程

Wrong Answer

a	b	a	a	a	b	a	c	a	b	0	0	b	a	a	a	b	a	c	a	b
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

2	1	4	5	6	3	7
---	---	---	---	---	---	---

- > 递归地对新串DC3，以保证rk互不相同。
- > 长度 $\%3==0$ 或1的后缀可以由字符 $S[i]+S[i+1..n]$ （已求好的后缀）双关键字表示。
- > 当然， $\%3==0$ 或2的后缀也可以三关键字表示。
- > 我们再基数排序一发，所有后缀间的rk就求出来了。

Compile Error Time Limit Exceeded

§ 1.3.复杂度?

Wrong Answer

>时间复杂度?

> $T(n)=T(2/3n)+O(n)$

>之前我不会DC3的原因就是一直觉得这东西带log。

>空间复杂度?

>显然线性…可能多几倍，看着开吧

>代码复杂度?

>别问我，我这辈子也不打算写它

Time Limit Exceeded

§ 1.3.复杂度?

Wrong Answer

- >有了这个东西，SA的理论复杂度就是线性的了
- >SA党可以与SAM党分庭抗礼辣（而且还比SAM少了个 Σ ）

- >总结：在第一小节，我们学习了一个几乎没用的算法。
- >难度还是很亲民的。（可能神犇们已经不想听了）
- >预告：
 - >接下来的一小节应该是更没有用的，而且可能大家都知道。
 - >第三小节或许有点用？

Compile Error Time Limit Exceeded

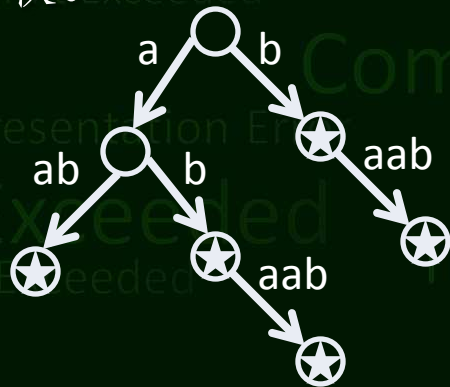
§ 2.1. 一个古老的数据结构

- > 后缀树。
- > 就是那个传说中比SA和SAM不知低到哪里去的东西。
- > lzz（应该是）冬令营时，为了表明US的数据结构科技落后，说了这样一句话
- > “我们在写后缀自动机的时候，他们还在写后缀树”
- > 为方便本小节之后内容的介绍，我们先说一下后缀树。

§ 2.1. 一个古老的数据结构

- > 它的实质就是对所有后缀建一棵trie，然后把同类节点合并，也就是把链缩起来。
- > 这个东西的节点数不超过 $2n$ ，因为每次加一个串最多会添加一个叶子和一个树杈。

abaab



§ 2.1. 一个古老的数据结构

- >构造方法是从前向后添加 $S[i..n]$ 。
- >用 $hd[i]$ 表示这个后缀与之前添加的后缀的最长公共前缀，也就是这一次分叉的位置。
- > $hd[i-1]$ 去掉最前面的一个字符以后，必然是 $hd[i]$ 的前缀。
- >维护指向 $h[i]$ 去掉最前面字符后的子串末尾的 $s[i]$ 。
- >从 $s[i-1]$ 开始找 $hd[i]$ ，找到以后加一个叉。
- > $s[i]$ 从父亲转移。
- >复杂度是线性的。
- >这个结构是可以识别 S 所有子串的。

§ 2.2.ST和SA的故事

Wrong Answer

- >前置知识结束。
- >其实这些后缀数据结构本质很相似。
- >//还有一堆针对模式串的前缀数据结构，比如KMP自动机（我没说错，kmp是个自动机
- >比如，后缀树字典序dfs一遍，就得到了后缀数组的rk。
- >这个也是线性的SA求法。（只是很暴力（不知为何觉得比DC3简单
- >然后，如果先建了SA，两个后缀的LCP可以 $O(1)$ ，这一步就是在求后缀树上的hd。于是可以避开s指针线性建后缀树。

Compile Error Time Limit Exceeded

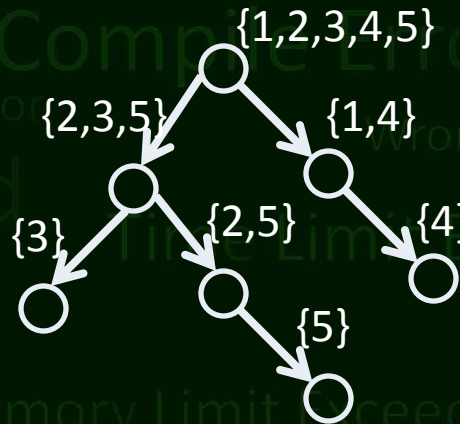
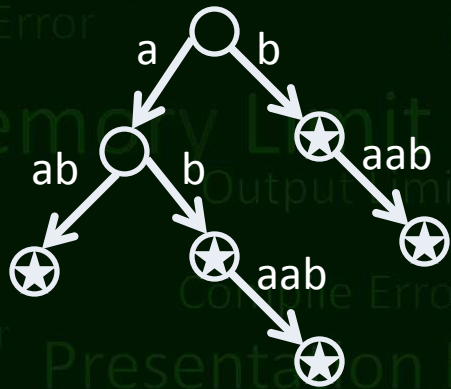
§ 2.3.ST和SAM的故事

- > “ST最多有 $2n$ 个节点”
- > “SAM最多有 $2n$ 个节点”
- > 真是巧?
- > 前几页有个abaab的后缀树。现在我们把baaba在后缀自动机上的parent树画一下。

§ 2.3.ST和SAM的故事

Wrong Answer

- > “ST最多有 $2n$ 个节点”
- > “SAM最多有 $2n$ 个节点”
- > 真是巧?
- > 前几页有个abaab的后缀树。现在我们把baaba在后缀自动机上的parent树画一下。



Compile Error Time Limit Exceeded

§ 2.3.ST和SAM的故事

Wrong Answer

>S串的ST是S' 串SAM上的parent树。

>好了，现在我们可以用SAM写后缀树…（逃

>当然，以后缀树为基础写SAM也挺有趣的。这个我没仔细研究，之前看别人的blog，忽然发现从这里推SAM的实现完全不涉及Right集合…

>这可能就是不知道多长时间前neither讲SAM的时候说的“一种不用Right集合的方法”？

Compile Error Time Limit Exceeded

§ 2.3.ST和SAM的故事

Wrong Answer

>S串的ST是S' 串SAM上的parent树。

>好了，现在我们可以用SAM写后缀树…（逃

>当然，以后缀树为基础写SAM也挺有趣的。这个我没仔细研究，之前看别人的blog，忽然发现从这里推SAM的实现完全不涉及Right集合…

>这可能就是不知道多长时间前neither讲SAM的时候说的“一种不用Right集合的方法”？

>其实是在这里看到的

Compile Error Time Limit Exceeded

§ 2.4.SA和SAM的故事

Wrong Answer

- > 显然对着SAMdfs也能线性求SA。
- > 然而我好像真没见过拿着SA建SAM的操作…

ceeded

Runtime Error

Output Limit Exceeded

Time Limit Exceeded

Wrong Answer

Compile Error

Output

entation Error

Presentation Error

Wrong Answer

Memory Limit Exceeded

Output Limit Exceeded

Time Limit Exceeded

ong Answer

Compile Error

Presentation Error

Memory Limit Exceeded

Runti

Compile Error

Time Limit Exceeded

§ 2.5.还有个东西叫SC

Wrong Answer

- > 它叫仙人掌不是因为有环，而是因为带刺。
- > 和后缀树长得非常像。
- > 据说唯一好处是常数比较小（不资道和SAM比怎么样）

> 总结：扯淡的一节就到这里

> 预告：接下来的一节是一个动态维护串的简单数据结构。

Compile Error Time Limit Exceeded

§ 3.1.一道经典题

Wrong Answer

>//顺便说一句，火星那个题是假的后缀平衡树。

Compile Error Time Limit Exceeded

§ 3.1.一道经典题

Wrong Answer

>bzoj2555 substring

> “支持在S串末尾动态加字符，询问某个串在S中的出现次数”

>SAM统计答案是需要dfs的。

>我会SAM+LCT!

Compile Error Time Limit Exceeded

§ 3.1.一道经典题

Wrong Answer

>bzoj4768 wxh loves substring

> “支持在S串末尾动态增删字符，询问某串在S中的出现次数”

>反正我是不会可持久化LCT。

>还是用一些更简单的东西吧。

Compile Error Time Limit Exceeded

§ 3.2. 后缀平衡树

Wrong Answer

- > 后缀平衡树。
- > 又是一个听起来很厉害实际很傻的知识点。
- > 就是用平衡树维护后缀的rk。
- > 每个节点是一个后缀。
- > 建树：对着后缀数组建一发。
- > 显然它支持在串首添加/删除字符。

Compile Error Time Limit Exceeded

§ 3.2. 后缀平衡树

Wrong Answer

- > 考虑添加一个字符的复杂度。
- > 必然要做 $\log n$ 次比较。
- > 二分 LCP 复杂度是 \log 的。
- > 把每个后缀看作 $c[i] + S[i+1..n]$ 的形式，我们只需要比较已经在树里的两个后缀大小关系。
- > 这个用重量平衡树是 $O(1)$ 的。
- > 可以选择写个 Treap 或者 SGT。

Compile Error Time Limit Exceeded

§ 3.3.真·例题

Wrong Answer

>clj论文里的例题。

>“支持双端插入删除、正中间插入删除，询问某个串出现次数”

>常识数据范围。

>分块就算了吧。

Compile Error Time Limit Exceeded

§ 3.3.真 · 例题

Wrong Answer

- > 双端插入删除？
 - > 维护两个后缀平衡树，每个维护一半的主串。
 - > 如果有一个删没了就重构。
 - > 跨越的部分和询问总长度同级，拿出来KMP。
 - > 正中间插入删除？
 - > 维护两组后缀平衡树，每组维护一半的主串。
 - > 保证分隔点是正中间。
- 在修改后微调中间的几个节点。

Compile Error Time Limit Exceeded

§ 3.4.顺便提一句

Wrong Answer

- >大家都知道有道题叫Sone2。
- >在文本串任意位置增删字符，询问固定模式串与文本串的某一段后缀的LCP的max与最大值出现次数。
- >好像完全包含了这道题呢。
- >Border Tree加一些黑科技。
- >就是因为试图看一些类似于这种的自己看不懂的paper，我才没做完课件（

Compile Error Time Limit Exceeded

§ 4.Reference

Wrong Answer

朱泽园 《多串匹配算法及其启示》

陈立杰 《重量平衡树和后缀平衡树在信息学奥赛中的应用》

王悦同、徐毅、徐子涵 《后缀仙人掌》

王鉴浩 《浅谈字符串匹配的几种方法》

Compile Error Time Limit Exceeded