区间计数题目讲解

fizzydavid（杨骏昭）

清华大学

2022 年 12 月 7 日

目录

题目大意

- 给定一个长度为 n 的整数数组 $a_1, a_2, \ldots, a_n$，其中 $a_i \in [1, n]$，且 1 到 n 每个数字**最多出现两次**。
- 定义一个**可重**集合 $S \subseteq \{1, 1, 2, 2, \ldots, n, n\}$ 是合法的当且仅当存在一个区间 $[l, r]$ $(1 \le l \le r \le n)$，使得 S 恰好包含 $\{a_l, a_{l+1}, \ldots, a_r\}$ 这些数字。你需要求出存在多少个不同的合法可重集合。

暴力一 直接模拟

- `std::set<std::multiset<int>>`

暴力二 集合哈希

- 可重集合哈希，每个数字对应一个随机生成的哈希值。
- 一个集合的哈希值等于其中每个数字对应的哈希值的和。
- 如果一个数字出现多次，那么求和的时候也要计入多次。

暴力二 集合哈希

- 可重集合哈希，每个数字对应一个随机生成的哈希值。
- 一个集合的哈希值等于其中每个数字对应的哈希值的和。
- 如果一个数字出现多次，那么求和的时候也要计入多次。
- `std::unordered_set<unsigned long long>`

暴力二 集合哈希

- 可重集合哈希，每个数字对应一个随机生成的哈希值。
- 一个集合的哈希值等于其中每个数字对应的哈希值的和。
- 如果一个数字出现多次，那么求和的时候也要计入多次。
- `std::unordered_set<unsigned long long>`
- 时间复杂度 $O(n^2 \log n)$ 或 $O(n^2)$。
- 哈希表中如果元素太多会由于 cache 变慢，可以考虑每次只算相同长度的区间，这样空间复杂度可以做到 $O(n)$ 。

判断相同的另一种做法

- 考虑两个不相交的区间 $[l_1, r_1], [l_2, r_2]$，用除了哈希以外的方法判断是否可重集合是否相同？
- 令 b_i 为 a_i 上一次在 a 中出现的位置的下标，那么相同的条件是
  - $r_1 - l_1 = r_2 - l_2$。
  - $[l_2, r_2]$ 中的元素都是第二次出现。
  - 对于任意 $i \in [l_2, r_2]$，应该有 $b_i \in [l_1, r_1]$。

判断相同的另一种做法

- 考虑两个不相交的区间 $[l_1, r_1], [l_2, r_2]$，用除了哈希以外的方法判断是否可重集合是否相同？
- 令 b_i 为 a_i 上一次在 a 中出现的位置的下标，那么相同的条件是
  - $r_1 - l_1 = r_2 - l_2$。
  - $[l_2, r_2]$ 中的元素都是第二次出现。
  - 对于任意 $i \in [l_2, r_2]$，应该有 $b_i \in [l_1, r_1]$。
- 为了简便，对于第一次出现的位置，令 $b_i = -2i$（或其他非法值），定义 $bl(l, r) = \min_{i=l}^{r} b_i$，$br(l, r) = \max(0, \max_{i=l}^{r} b_i)$，那么上述条件变为

$$br(l_2, r_2) - bl(l_2, r_2) - (r_2 - l_2) = 0$$

关键区间

Definition (关键区间)

若序列 a 中两个不相交的区间 $[l_1, r_1], [l_2, r_2] (r_1 < l_2)$ 对应的可重集合相同，那么称 $[l_2, r_2]$ 为关键区间。

关键区间

Definition (关键区间)

若序列 a 中两个不相交的区间 $[l_1, r_1], [l_2, r_2] (r_1 < l_2)$ 对应的可重集合相同，那么称 $[l_2, r_2]$ 为关键区间。

Definition (对应区间)

若 a_i 在序列 a 中第二次出现，则 $b_i = j(a_j = a_i, j < i)$，否则 $b_i = -2i$。令 $bl(l, r) = \min_{i=l}^{r} b_i$，$br(l, r) = \max(0, \max_{i=l}^{r} b_i)$。那么称 $[bl(l, r), br(l, r)]$ 是 $[l, r]$ 的对应区间。

关键区间

Definition (关键区间)

若序列 a 中两个不相交的区间 $[l_1, r_1], [l_2, r_2] (r_1 < l_2)$ 对应的可重集合相同，那么称 $[l_2, r_2]$ 为关键区间。

Definition (对应区间)

若 a_i 在序列 a 中第二次出现，则 $b_i = j(a_j = a_i, j < i)$，否则 $b_i = -2i$。令 $bl(l, r) = \min_{i=l}^{r} b_i$，$br(l, r) = \max(0, \max_{i=l}^{r} b_i)$。那么称 $[bl(l, r), br(l, r)]$ 是 $[l, r]$ 的对应区间。

- 对应区间的长度大于等于原区间长度。

关键区间

Definition (关键区间)

若序列 a 中两个不相交的区间 $[l_1, r_1], [l_2, r_2] (r_1 < l_2)$ 对应的可重集合相同，那么称 $[l_2, r_2]$ 为关键区间。

Definition (对应区间)

若 a_i 在序列 a 中第二次出现，则 $b_i = j(a_j = a_i, j < i)$，否则 $b_i = -2i$。令 $bl(l, r) = \min_{i=l}^{r} b_i$，$br(l, r) = \max(0, \max_{i=l}^{r} b_i)$。那么称 $[bl(l, r), br(l, r)]$ 是 $[l, r]$ 的对应区间。

- 对应区间的长度大于等于原区间长度。
- 对于相交的情况（不妨设 $l_1 < l_2 < r_1 < r_2$），可以转化为：区间 $[l_1, l_2 - 1]$ 是 $[r_1 + 1, r_2]$ 的对应区间。

利用关键区间计数

- 有多少个区间 $[l, r]$，它和某个比它更左的区间有着同样的可重集合？
- 区间 $[l, r]$ 被计数一次当且仅当至少满足下列之一：
  - 条件一：$[l, r]$ 是关键区间。
  - 条件二：存在某个后缀区间 $[l', r](l \leq l' \leq r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

利用关键区间计数

- 有多少个区间 $[l, r]$，它和某个比它更左的区间有着同样的可重集合？
- 区间 $[l, r]$ 被计数一次当且仅当至少满足下列之一：
  - 条件一：$[l, r]$ 是关键区间。
  - 条件二：存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。
- 考虑计数：满足条件一的个数 + 满足条件二的个数 - 同时满足条件一和二的个数

利用关键区间的优秀暴力

- 考虑 $O(n^2)$ 枚举 $[l, r]$ 寻找所有的关键区间。
- 两个条件都可以 $O(1)$ 计数，这里省略细节。
- 由于只有一半的数字是第二次出现，所以关键区间数量至多有 $n^2/8$，常数很小。

利用关键区间的优秀暴力

- 考虑 $O(n^2)$ 枚举 $[l, r]$ 寻找所有的关键区间。
- 两个条件都可以 $O(1)$ 计数，这里省略细节。
- 由于只有一半的数字是第二次出现，所以关键区间数量至多有 $n^2/8$，常数很小。
- 对于性质一的数据中，序列中数的位置被施加了扰动。随着区间长度变长，其成为关键区间的概率会近似指数地衰减（当长度小于某个常数时），所以当 n 大的时候枚举区间长度不用很长。
- 可以通过具有性质一的数据。

计数条件一

Recall (条件一)

$[l,r]$ 是关键区间。

- 较为经典的问题：维护连续段。
- 从右往左枚举左端点 l，对于每个右端点 r，线段树维护

$$c_r = br(l,r) - bl(l,r) - (r-l)$$

计数条件一

Recall (条件一)

$[l, r]$ 是关键区间。

- 较为经典的问题：维护连续段。
- 从右往左枚举左端点 l，对于每个右端点 r，线段树维护

$$c_r = br(l, r) - bl(l, r) - (r - l)$$

- 一定有 $c_r \geq 0$，计数关键区间个数等价于维护最小值个数。

计数条件一

Recall (条件一)

$[l, r]$ 是关键区间。

- 较为经典的问题：维护连续段。
- 从右往左枚举左端点 l，对于每个右端点 r，线段树维护

$$c_r = br(l, r) - bl(l, r) - (r - l)$$

- 一定有 $c_r \geq 0$，计数关键区间个数等价于维护最小值个数。
- $bl(l, r), br(l, r)$ 对于 r 都是单调的，可以用单调栈维护。

计数条件一

Recall (条件一)

$[l,r]$ 是关键区间。

- 较为经典的问题：维护连续段。
- 从右往左枚举左端点 l，对于每个右端点 r，线段树维护

$$c_r = br(l,r) - bl(l,r) - (r-l)$$

- 一定有 $c_r \geq 0$，计数关键区间个数等价于维护最小值个数。
- $bl(l,r), br(l,r)$ 对于 r 都是单调的，可以用单调栈维护。
- 使用区间加修改 c_r，均摊 $O(1)$ 次区间加。

计数条件二

Recall (条件二)

存在某个后缀区间 $[l', r](l \leq l' \leq r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 可能有多种符合要求的后缀，我们只想计数一次。
- 考虑计数“是否是最小的符合要求的后缀”。

计数条件二

Recall (条件二)

存在某个后缀区间 $[l',r](l \le l' \le r)$，使得 $[l',r]$ 为关键区间，且其对应区间右端点 $br(l',r) = l-1$。

- 可能有多种符合要求的后缀，我们只想计数一次。
- 考虑计数“是否是最小的符合要求的后缀”。
- 假设从右往左枚举左端点 l，维护 $c_r = br(l,r) - bl(l,r) - (r-l)$。

计数条件二 - 分析

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 考虑某一个固定的右端点 r，它可能有多个 l 满足 $[l, r]$ 是关键区间，也就是在从右向左枚举 l 的多个时刻中，c_r 达到了最小值 0。

计数条件二 - 分析

Recall (条件二)

存在某个后缀区间 $[l', r](l \leq l' \leq r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 考虑某一个固定的右端点 r，它可能有多个 l 满足 $[l, r]$ 是关键区间，也就是在从右向左枚举 l 的多个时刻中，c_r 达到了最小值 0。
- 考虑这些 $[l, r]$ 的对应区间的右端点 $br(l, r)$。随着 l 的不断变小，$br(l, r)$ 应该单调不减。

计数条件二 - 分析

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 考虑某一个固定的右端点 r，它可能有多个 l 满足 $[l, r]$ 是关键区间，也就是在从右向左枚举 l 的多个时刻中，c_r 达到了最小值 0。
- 考虑这些 $[l, r]$ 的对应区间的右端点 $br(l, r)$。随着 l 的不断变小，$br(l, r)$ 应该单调不减。
- 对于 $br(l, r)$ 相同的那些时刻 l，应该将这些 $[l, r]$ 对于 $[br(l, r) + 1, r]$ 满足条件二的次数只贡献一次。

计数条件二 - 例子

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 举例：$3, 4, 1, 2, 5, 4, 3, 2, 1$，考虑固定的 $r = 9$
- 当 $l = 9, 8, 6$ 时，$[l, r]$ 为关键区间，对应区间分别是 $[3, 3], [3, 4], [1, 4]$。

计数条件二 - 例子

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 举例：$3, 4, 1, 2, 5, 4, 3, 2, 1$，考虑固定的 $r = 9$
- 当 $l = 9, 8, 6$ 时，$[l, r]$ 为关键区间，对应区间分别是 $[3, 3], [3, 4], [1, 4]$。
- 其中 $([3, 3], [9, 9])$ 贡献了区间 $[4, 9]$ 的条件二计数，$([3, 4], [8, 9]), ([1, 4], [6, 9])$ 贡献了区间 $[5, 9]$ 的条件二计数。

计数条件二 - 例子

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 举例：$3, 4, 1, 2, 5, 4, 3, 2, 1$，考虑固定的 $r = 9$
- 当 $l = 9, 8, 6$ 时，$[l, r]$ 为关键区间，对应区间分别是 $[3, 3], [3, 4], [1, 4]$。
- 其中 $([3, 3], [9, 9])$ 贡献了区间 $[4, 9]$ 的条件二计数，$([3, 4], [8, 9]), ([1, 4], [6, 9])$ 贡献了区间 $[5, 9]$ 的条件二计数。
- 后者只应被计入一次。

计数条件二 - 数据结构

Recall (条件二)

存在某个后缀区间 $[l',r](l \le l' \le r)$，使得 $[l',r]$ 为关键区间，且其对应区间右端点 $br(l',r) = l-1$。

- 由于维护了 $br(l,r)$ 的单调栈，每次修改单调栈的时候可以知道哪些 r 对应的 $br(l,r)$ 增加了。
- 考虑对每个 r 维护一个“一血标记”，每次 $br(l,r)$ 增加的时候对 r 打上标记，计数的时候要把满足 $c_r = 0$ 的 r 的标记去除。

计数条件二 - 数据结构

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 由于维护了 $br(l, r)$ 的单调栈，每次修改单调栈的时候可以知道哪些 r 对应的 $br(l, r)$ 增加了。
- 考虑对每个 r 维护一个“一血标记”，每次 $br(l, r)$ 增加的时候对 r 打上标记，计数的时候要把满足 $c_r = 0$ 的 r 的标记去除。
- 区间加，区间打标记，区间询问有标记的最小值并把标记去掉。怎么维护？

计数条件二 - 维护标记

- 区间加，区间打标记，区间询问有标记的最小值并把标记去掉。怎么维护？
- 除了把最小值的标记去掉以外都是常规的线段树操作。
- 由于线段树区间维护了最小值，“把最小值的标记去掉” 也可以下传，只要判断左右区间的最小值是否是大区间的最小值。

计数条件一 + 条件二

Recall (条件一)

$[l, r]$ 是关键区间。

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 这样的 $[l, r]$ 一定有它的对应区间为为 $[2l - r - 1, l - 1]$。

计数条件一 + 条件二

Recall (条件一)

$[l, r]$ 是关键区间。

Recall (条件二)

存在某个后缀区间 $[l', r](l \leq l' \leq r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 这样的 $[l, r]$ 一定有它的对应区间为为 $[2l - r - 1, l - 1]$。
- 只要在之前计数条件二的计数中额外限定 $br(l, r) = l - 1$ 就行了。在单调栈中二分即可。

计数条件一 + 条件二

Recall (条件一)

$[l,r]$ 是关键区间。

Recall (条件二)

存在某个后缀区间 $[l',r](l \le l' \le r)$，使得 $[l',r]$ 为关键区间，且其对应区间右端点 $br(l',r) = l-1$。

- 这样的 $[l,r]$ 一定有它的对应区间为为 $[2l-r-1, l-1]$。
- 只要在之前计数条件二的计数中额外限定 $br(l,r) = l-1$ 就行了。在单调栈中二分即可。
- 总时间复杂度 $O(n \log n)$，空间复杂度 $O(n)$ 。期望得分 100 分。

计数条件二的另一种方式

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 考虑**从左往右枚举右端点** r，用线段树维护左端点 l 以及 $c_l = br(l, r) - bl(l, r) - (r - l)$。

计数条件二的另一种方式

Recall (条件二)

存在某个后缀区间 $[l',r](l \le l' \le r)$，使得 $[l',r]$ 为关键区间，且其对应区间右端点 $br(l',r) = l-1$。

- 考虑**从左往右枚举右端点** r，用线段树维护左端点 l 以及 $c_l = br(l,r) - bl(l,r) - (r-l)$。
- 按照单调栈中相同的 $br(l,r)$ 对 l 分段，每一段中应该统计是否存在一个 l 满足 $[l,r]$ 为关键区间。

计数条件二的另一种方式

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 考虑**从左往右枚举右端点** r，用线段树维护左端点 l 以及 $c_l = br(l, r) - bl(l, r) - (r - l)$。
- 按照单调栈中相同的 $br(l, r)$ 对 l 分段，每一段中应该统计是否存在一个 l 满足 $[l, r]$ 为关键区间。
- $br(l, r)$ 需要支持区间修改。怎么维护？

计数条件二的另一种方式

Recall (条件二)

存在某个后缀区间 $[l', r](l \le l' \le r)$，使得 $[l', r]$ 为关键区间，且其对应区间右端点 $br(l', r) = l - 1$。

- 考虑**从左往右枚举右端点** r，用线段树维护左端点 l 以及 $c_l = br(l, r) - bl(l, r) - (r - l)$。
- 按照单调栈中相同的 $br(l, r)$ 对 l 分段，每一段中应该统计是否存在一个 l 满足 $[l, r]$ 为关键区间。
- $br(l, r)$ 需要支持区间修改。怎么维护？
- 线段树维护区间最左和最右一段相同的 $br(l, r)$ 是什么以及是否有关键区间。维护中间段的答案的和。可以 $O(1)$ 区间信息合并。

其他算法

- 也可以使用一种专门维护连续段的数据结构：析合树。它很好地刻画了连续段的结构，我们可以使用这种结构对答案计数。这里具体细节略去，感兴趣的同学可以自行思考。
- 注：如果使用线性构造算法可以将总时间复杂度降到线性。

题目总结

- 思维难度和代码难度适中，主要考察选手对题目性质的分析与转化能力与在数据结构方面的综合解题能力。希望选手们能获得愉快的做题经历。

谢谢大家！

- Q & A
- 欢迎提问。