Solution for 2016 Multi-University Training Contest 9 - Generator

xyz2606

October 28, 2016

We consider the following process: a monkey typing at random, given sufficient time, produce any given string. How long might this be expected to take? We can compute this value by Gauss elimination easily. But there exists a more efficient algorithm [T.01]. In this article, we will introduce that algorithm first and then talk about the problem [16].

1 The Algorithm

In this section, we introduce the algorithm. Let $\sigma = (\sigma_1, \dots, \sigma_n)$ denotes a fixed finite sequence of alphabet A. Let $c_1, c_2, \dots$ be a sequence of i.i.d. random digits and $\Pr_{c_i}[c_i = a] = p_a$, $a \in A$. Assume $\sum_{a \in A} p_a = 1$.

Lemma 1.1. $\mathbb{E}_\sigma[T_\sigma] = \Pi_{i=1}^n \frac{1}{p_{\sigma_i}}$*, where* T_σ *denotes the time* σ *first occurs(We say that* σ *occurs at time* t *if* $(b_{t-n+1}, \dots, b_t) = \sigma$ *and we assume* $(b_{-n+1}, \dots, b_0) = \sigma$ *for convenient).*

Proof. The sequence b can be seen as a Markov chain: the states are the consecutive n charactors and the transitions are "Adding one random charactor to the end of the state and then deleting the first charactor". The transition matrix is doubly stochastic. So we have a uniform stationary distribution and hence we have

$$\mathbb{E}_\sigma[T_\sigma] = \Pi_{i=1}^n \frac{1}{p_{\sigma_i}}$$

which is just the expected time for a random walk to return to its initial state. □

Theorem 1.2. *Let* τ *be a binary string of length* n *and let* T_τ *be the least* j *such that* τ *occurs in* $(b_1, b_2, \dots, b_j)$*(note that here we no longer assume values of* b_i *for* $i \leq 0$*). Let* σ *be the longest proper suffix of* τ *that is also a prefix of* τ*. Then*

$$\mathbb{E}[T_\tau] = \mathbb{E}[T_\sigma] + \Pi_{i=1}^n \frac{1}{p_{\tau_i}}$$

.

Proof. First suppose σ is non-empty. (The empty case is just Lemma 1.1 because of the observation that the pre-assumed values for $b_i, i \leq 0$ have no affect on the first occurence of τ.) Let $\nu_1, \nu_2, \dots$ be the successive times that σ occurs in b. Let

$$Y_j = \begin{cases} 1, \text{ if } \tau \text{ occurs during } (\nu_j + 1, \dots, \nu_{j+1}), \\ 0, \text{ else.} \end{cases}$$

Because of σ's maximality, Y_i are independent from each other. And we have

$$\Pr\left[Y_j = 1\right] = \Pi_{i=|\sigma|+1}^{n} p_{\tau_i}$$

.

Let N be the first $j \geq 1$ such that $Y_j = 1$, then N has a geometric distribution and

$$\mathbb{E}\left[N\right] = \Pi_{i=|\sigma|+1}^{n} \frac{1}{p_{\tau_i}}$$

.

Next, observe that on $Y_j = 1$ we must have $\nu_{j+1} = \nu_j + (n - |\sigma|)$ by definition of σ. Thus

$$T_\tau = T_\sigma + \sigma_{j=1}^{N}(\nu_{j+1} - \nu_j)$$

.

The $\nu_{j+1} - \nu_j$ are also i.i.d., and by Lemma reflem:noBifix we have $\mathbb{E}\left[\nu_{j+1} - \nu_j\right] = \Pi_{i=1}^{|\sigma|} \frac{1}{p_{\sigma_i}}$. Finally, by Wald's first lemma,

$$\mathbb{E}\left[\sigma_{j=1}^{N}(\nu_{j+1} - \nu_j)\right] = \mathbb{E}\left[N\right]\mathbb{E}\left[\nu_2 - \nu_1\right] = \Pi_{i=1}^{|\sigma|} \frac{1}{p_{\sigma_i}} \Pi_{i=|\sigma|+1}^{n} \frac{1}{p_{\tau_i}} = \Pi_{i=1}^{n} \frac{1}{p_{\tau_i}}$$

. □

2 The Problem

In this section, we consider the following problem [16]: Given N strings of length L each and the array p as in the previous section, output the expected time when all the N strings occur in b.

By

$$max_{i=1}^{k} R_i = \sum_{s \in [k]} (-1)^{|S|+1} min_{i \in S} R_i$$

,

we convert the problem into: Given N strings of length L each and the array p, output the expected time when any one of the N strings first occurs in b.

Let the N strings be $s_1, \dots, s_n$. Let F_i be the event [s_i occurs the earliest among all the N strings]. Let x_i be $\Pr\left[F_i\right]$. We have

$$\sum_{j=1}^{n} x_j (\mathbb{E}\left[T_{s_j} | F_j\right] + \mathbb{E}\left[T_{s_i} - T_{s_j} | F_j\right]) = \mathbb{E}\left[T_{s_i}\right]$$

for any $i = 1, \dots, n$. The answer we need is just $\sum_{j=1}^{n} x_j \mathbb{E}\left[T_{s_j} | F_j\right]$. We denote this by ans.

Let $\sigma_{i,j}$ be the longest string which is a suffix of s_j and a prefix of s_i. Then $\mathbb{E}\left[T_{s_i} - T_{s_j} | F_j\right]$ is just $\mathbb{E}\left[T_{s_i} - T_{\sigma_{i,j}} | F_j\right]$.

For any prefix p of s_i,

$$\mathbb{E}\left[T_{s_i}\right] = \mathbb{E}\left[T_p\right] + \mathbb{E}\left[T_{s_i} - T_p\right]$$

.

So finally, for any i, we have the following equation:

$$\sum_{j=1}^{n} x_j (\mathbb{E}\left[T_{s_j} | F_j\right] + \mathbb{E}\left[T_{s_i}\right] - \mathbb{E}\left[T_{\sigma_{i,j}}\right]) = \mathbb{E}\left[T_{s_i}\right]$$

subtract $\mathbb{E}\left[T_{s_i}\right]$ from both sides, we have

$$\sum_{j=1}^{n} x_j (\mathbb{E}\left[T_{s_j} | F_j\right] - \mathbb{E}\left[T_{\sigma_{i,j}}\right]) = 0 \tag{1}$$

$$\sum_{j=1}^{n} x_j (\mathbb{E}\left[T_{\sigma_{i,j}}\right]) = ans \tag{2}$$

.

Together with $\sum_{i=1}^{n} x_i = 1$, we have $n + 1$ equations for $n + 1$ variables $x + 1, \dots, x_n$ and ans. Now we can use Gauss elimination to solve the problem.

References

[16] . 2016 multi-university training contest 9 for acm icpc, 2016.

[T.01] McConnell T. The expected time to find a string in a random binary sequence, 2001.