PhD Foundation: Reservoir Random Sampling

ZHANG Shiqi

July 26, 2019

1 Calculus Foundation

1.1 Basic Definition

Series: the sum of the terms of an infinite sequence of numbers.

n_{th} **partial sum** S_n**::** the sum of the first n terms of the sequence, that is, $\sum_{k=1}^{n} a_k$.

Convergent Series:: A series is convergent if the sequence of its partial sums $\{S_1,\ S_2,\ S_3, \dots\}$ tends to a limit.

1.2 Harmonic Series

Definition: The harmonic series is the divergent infinite series:

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \dots$$

Partial Sum: The finite partial sums of the diverging harmonic series are called harmonic numbers.

$$H_n = \sum_{k=1}^{n} \frac{1}{k}$$

Rate of Divergence: The harmonic series diverges very slowly. This is because the partial sums of the series have logarithmic growth.

$$\sum_{k=1}^{n} \frac{1}{k} \approx \ln k$$

2 Reservoir Sampling: Algorithm R

The sampling problem is to randomly select n out of N records by one pass. In [1], the author presents the reservoir sampling algorithm R. The basic idea is to update a size-n reservoir.

1. Initially, the first n records in N will be stored in reservoir.
2. The $(t+1)_{st}$ record$(t+1 > n)$ can be reserved with a probability of $\frac{n}{t+1}$. It can be implemented by generating a random integer r.
3. If reserved, that is $r \leq n$, it will replace the old record based on the random integer r.

2.1 Complexity Analysis

The total running time is $O(N)$ since it has to scan all records in a file. Now we focus on analyzing the I/O time complexity.

The total number of records that will be loaded into memory is,

$$n + \sum_{n \le t < N} \frac{n}{t+1}$$

Based on the properties of harmonic series in 1.2, the equation is approximately equal to

$$n(1 + H_N - H_n) \approx n(1 + \ln \frac{N}{n})$$

Therefore, the I/O time complexity is $O(n(1 + \ln \frac{N}{n}))$

3 Discussion: Does each record have equal probability of being chosen?

Theorem 1. Suppose the reservoir size is n and total record size is N. Each record has equal probability (i.e. $\frac{n}{N}$) of being chosen for the reservoir.

Proof. Let's prove the correctness by induction.

(1). When there are less or equal than n records, then the probability of being chosen is 1.

(2). While the $(n+1)_{st}$ record is coming, the new record has the $\frac{n}{n+1}$ probability of being chosen. The i_{th} old record ($i \le n$) has the $(\frac{1}{n+1} + \frac{n-1}{n} * \frac{n}{n+1})$ probability, that is $\frac{n}{n+1}$. The first part represents the new record is discarded and the second part represents the new record is kept but not replace i_{th} position. Thus all $(n+1)$ records have same probability.

(3). While the $(n+2)_{st}$ record is coming, the new record has the $\frac{n+1}{n+2}$ probability of being chosen. All $(n+1)$ previous records have $\frac{n}{n+1}$ probability. Thus, the i_{th} old record ($i \le n$) has the $(\frac{1}{n+2} + \frac{n}{n+1} * \frac{n-1}{n} * \frac{n+1}{n+2})$ probability, that is $\frac{n}{n+2}$.

(4). By induction, when there are N records, the probability of being chosen is $\frac{n}{N}$ for each record.

□

References

[1] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985.