Lecture 7: Reservoir Sampling

CMSC 5705 Advanced Topics in Database Systems

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

November 8, 2010

In this lecture, we will discuss a classic sampling algorithm called *reservoir sampling*. This algorithm takes a random sample set of the desired size in only one pass over the underlying dataset. This feature makes the algorithm ideal for stream environments where every item can be processed only once.

Problem (Random sampling)

Given a set S of items, compute a random sample set of size k.

Reservoir

algorithm `reservoir`(k, S)
/* take k random samples from the dataset S */

1. initialize an array *samples* of size k
2. **for** $i = 1$ to $n = |S|$
3. $\quad o =$ the i-th item
4. $\quad$ **if** $i \leq k$ **then**
5. $\quad\quad samples[i] = o$
6. $\quad$ **else**
7. $\quad\quad$ generate a random integer from 1 to x
8. $\quad\quad$ **if** $x \leq k$ **then**
9. $\quad\quad\quad samples[i] = o$

Example

Let $S = \{59, 100, 2, 30, 63, ...\}$, and $k = 3$.

- The first k items are directly added to the sample set. So $samples = (59, 100, 2)$.
- Given the 4th item, the algorithm generates a random integer x from 1 to 4. Assume that the generated $x = 4$. As $x > k$, the item is ignored.
- Given the 5th item, again, the algorithm generates x randomly, but now from 1 to 5. Assume that $x = 2$ this time. Hence, the item is added to *samples*, and replaces the 2nd value there. Hence, *samples* becomes $(59, 63, 2)$.
- The remaining items are processed in the same manner.

The reservoir algorithm is very efficient: it spends $O(1)$ time per item. Next, we will show that the algorithm is correct, namely:

1. (equal likelihood) Every item of S has the same probability of being sampled.
2. (independence) For any two items o_1, o_2, the events they are sampled are independent from each other.

The second statement is obvious – the decision we made to sample an incoming item is not based on the results of sampling the preceding items. Next, we focus on proving the first statement.

Proof of correctness

Theorem

After $n \geq k$ items in S have been processed, each of those items is sampled with probability s/n.

Proof

We prove the theorem by induction. Basic step: for $n = k$ the statement is obviously correct. Inductive step: assuming the correctness for $n = m$, next we show that the statement is also correct for $n = m + 1$.

Proof of correctness (cont.)

Proof (cont.)

Our discussion distinguishes the $(m+1)$-th object o and any of the first m objects o':

- o is sampled if and only if the random number x generated for o falls in the range from 1 to s. Hence, o is sampled with probability $s/(m+1)$.
- o' is sampled (after processing o) if and only if (i) it was sampled after processing the first m items, and (ii) the random number x generated for o is not equivalent to the index value of o in the array *samples*.

  By our inductive assumption, (i) happens with probability s/m. (ii) occurs with probability $m/(m+1)$. As the two events are independent, the probability that they happen simultaneously equals $\frac{s}{m} \cdot \frac{m}{m+1} = \frac{s}{m+1}$.

Playback of this lecture

- Reservoir algorithm.
- $O(1)$ time per item.
- One pass.