Data mining: homework 2

Edo liberty

The setup is as follows. We have a universe of N items $A = \{a_1, \ldots, a_N\}$ and m subsets $S_i \subset A$, $i \in \{1, \ldots, m\}$. We assume that given a set S_i we can iterate over its elements one by one. The exercise will deal with approximating the size of different unions of these sets.

1. In case you wanted to give an ε approximation, w.p. $1 - \delta$, to the size of $S_1 \cup S_2$. How would your ability to approximate the zero'th frequency moment in streams help you with that? (We assume here that $O(|S_1| + |S_2|)$ running time is acceptable, and that ε and δ are both constants)

2. Assuming it is only possible to compute the second frequency moment of streams, one can still give an ε approximation, w.p. $1 - \delta$, to the size of $S_1 \cup S_2$. How?

3. Assume now that you are tasked with designing an algorithm. Your algorithm is allowed to preprocess the sets S_i in any amount of time and produce any data structure. It should then be able to take as input a set of indexed $I \in \{1, \ldots, m\}$ and produce an ε approximation of $|\cup_{i \in I} S_i|$ with probability at least $1 - \delta$. The aim is to create an algorithm which runs in time $o(\sum_{i \in I} |S_i|)$, i.e., the solution from question 1 is not the answer. It is assumed that for all i, $|S_i| \in \omega(1)$.

   - describe the preprocessing stage and its resulting data structure. (before I is given)
   - describe the estimation process. (after I is given)
   - prove your algorithm's correctness.
   - give the space usage of your data structures.
   - give the runtime complexity your estimation process.