Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II

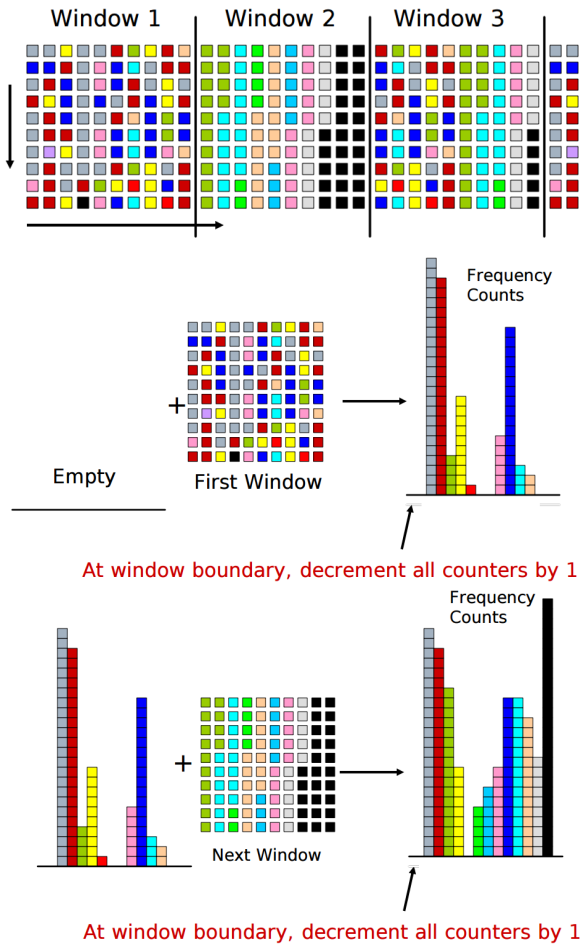
SS 2019

Exercise 6: Correlation Clustering and Stream 1

Exercise 6-1 Lossy Counting

Before stream clustering, let's take a look at a more fundamental task in stream: count the occurrence of objects in a stream and output objects with a count larger or equal to some given threshold: $minSup \times L$, where L is the length of the stream up to now and $minSup$ is the given threshold (minimum support).

Lossy Counting is one of the basic algorithms that solve this problem. Given the windows size as $w = \frac{1}{\epsilon}$, the lossy counting algorithm works as the follows: cut stream into windows, process one window a time and prune histogram entries with 0 counts at each window boundary. The illustration is given below:



Please prove that

(a) the maximum count error (the maximum difference between the real count and the estimated count) of the lossy counting algorithm is ϵL

(b) the memory consumption, i.e., the number of entries stored in the histogram, is $O(\frac{1}{\epsilon}\log(\epsilon L))$. (Optional)

Let W be the current window id. For each $i \in [1, W]$, let d_i denote the number of entries in the histogram H which corresponds to window $W - i + 1$.

Thus, the item in the stream corresponding to such entry must occur at least i times in window $B - i + 1$ through W; otherwise, it would have been deleted. Since the size of each window is w, we have:

$$\sum_{i=1}^{j} id_i \leq jw \quad for\ j = 1, 2, \ldots, W$$

Now we want to prove: $\sum\limits_{i=1}^{j} d_i \leq \sum\limits_{i=1}^{j} \frac{w}{i}$

By induction:

- For $j = 1$, this is true.
- Assume it is true for $j = 1, 2, \ldots, p-1$, then for $j = p$, adding the first inequality for $j = p$ to all $p - 1$ instances of the second inequality gives us:

$$\sum_{i=1}^{p} id_i + \sum_{i=1}^{1} d_i + \sum_{i=1}^{2} d_i + \cdots + \sum_{i=1}^{p-1} d_i \leq pw + \sum_{i=1}^{1} \frac{w}{i} + \sum_{i=1}^{2} \frac{w}{i} + \cdots + \sum_{i=1}^{p-1} \frac{w}{i}$$

(Here, the second inequality is used for $p - 1$ times with j varies from 1 to $p - 1$. We can do this because we assume the second inequality is true for all $j \in [1, p - 1]$.)

Then we get $p \sum\limits_{i=1}^{p} d_i \leq pw + \sum\limits_{i=1}^{p-1} \frac{(p-i)w}{i} = p \sum\limits_{i=1}^{j} \frac{w}{i} \Rightarrow \sum\limits_{i=1}^{p} d_i \leq \sum\limits_{i=1}^{j} \frac{w}{i}$

(The process is:

$$\begin{aligned} pw + \sum_{i=1}^{p-1} \frac{(p-i)w}{i} =& \frac{pw}{p} + \frac{1w}{1} + \frac{2w}{2} + \cdots + \frac{(p-1)w}{p-1} + \sum_{i=1}^{p-1} \frac{(p-i)w}{i} \\ =& \sum_{i=1}^{p} \frac{pw}{i} \end{aligned}$$

)

Thus the memory consumption at window W is $|H| = \sum\limits_{i=1}^{W} d_i \leq \sum\limits_{i=1}^{W} \frac{w}{i} = \frac{1}{\epsilon}\log W = \frac{1}{\epsilon}\log \epsilon L$

(Here the inequality $\sum\limits_{i=1}^{W} \frac{1}{i} \leq \log W$ is used, as it is the harmonic series.)