COMP90056 Stream Computing and Applications

Assignment 2 Part 1, 2019 Sem2

Xiuge Chen 961392

Part 1

Given a insertion-only stream σ_i, Misra-Gries sketch with parameter k is a majority algorithm that returns a candidate set S_i, which contains $k-1$ potential items have frequency $> \frac{m_i}{k}$. As proved in lecture note 8, the following inequality holds for all elements of S_i:

$$f_{i,x} - \frac{m_i - \tau_i}{k} \leq \hat{c}_i(x) \leq f_{i,x}, \forall x \in S_i \tag{1}$$

where $\hat{c_i}(x)$ is the estimated frequency of item x in S_i, $f_{i,x}$ is the actual frequency of item x in stream σ_i, m_i is the sum of actual frequencies of all items appeared in stream σ_i (also could be called as $F_{i,1}$), τ_i is the sum of estimated frequencies of all items in S_i.

According to the algorithm described in the part 1 of spec, two k-Misra-Gries summaries (S_1 and S_2) of two streams, σ_1 and σ_2, could be further combined into a new candidate set S_{new} that contains $k-1$ potential elements have frequency $> \frac{m_1+m_2}{k}$ over stream $\sigma_1 \cdot \sigma_2$, and the inequality (1) still holds for all elements of the new S.

To proof the inequality holds for all elements of S, let's firstly focus on the easier side, the upper bound of $c_n\hat{e}w(x)$.

For every element $y \in S_{new}$, its actual frequency $f_{new,x}$ in $\sigma_1 \cdot \sigma_2$ should be the sum of its frequencies in σ_1 and σ_2, since two streams are concatenated together and they are both insertion-only. For elements that appear in only one stream, this assertion could still be applied by considering their actual and estimated frequencies in another stream as 0. Using the corresponding rule described in the combining algorithm:

$$\hat{c}_{new}(y) = \hat{c}_1(y) + \hat{c}_2(y) \leq f_{1,x} + f_{2,x} = f_{new,x}$$

$$Consequently,\ \hat{c}_{new}(y) \leq f_{new,x}$$

As for the lower bound of $\hat{c}_{new}(y)$, it is suffice to show that by considering the maximum difference between estimated and actual frequency of y, $(f_{new,y} - \hat{c}_{new}(y))$, after combining two sketches. As shown above, $\hat{c}_{new}(y)$ could only be smaller than $f_{new,y}$, thus no overestimation will be made. Also, according to the property of Misra-Gries sketch and combination algorithm, $\hat{c}_{new}(y)$ will be smaller than $f_{new,y}$ if and only if $\hat{c}_{new}(y)$ is decremented either while obtaining individual Misra-Gries results S_i (subtracting one at new item arrival), or while combining two results together (subtracting by the counter of k^{th} largest items in S_{new}). Therefore, the difference between estimated and actual frequency of $y \in S_{new}$ is the sum of decrements in obtaining S_1 $(f_{1,y} - \hat{c}_1(y))$, decrements in obtaining S_2 $(f_{2,y} - \hat{c}_2(y))$, and decrements in merging them ($M = \hat{c}_{new}(k^{th}\ largest\ item\ K)$ before decremented $= \hat{c}_1(K) + \hat{c}_2(K)$). Again, if y only appears in one stream, then its decrements in another stream $(f_{other,y} - \hat{c}_{other}(y))$ is just simply 0, the assertion still applies. Thus:

$$\begin{aligned} f_{new,y} - \hat{c}_{new}(y) &= (f_{1,y} - \hat{c}_1(y)) + (f_{2,y} - \hat{c}_2(y)) + M \\ &\leq \frac{m_1 - \tau_1}{k} + \frac{m_2 - \tau_2}{k} + M \\ &= \frac{(m_1 + m_2) - (\tau_1 + \tau_2)}{k} + M \end{aligned} \tag{2}$$

Note that $\tau_1 + \tau_2$ is the sum of all estimated frequencies in S_1 and S_2, while τ_{new} is just the sum of estimated frequencies in S. During the production of S_{new}, all estimation are decremented by the k^{th} largest item's estimation M, so that only the top $(k-1)$ largest items will be remained in S_{new} with estimation $(\hat{c}_1(x) + \hat{c}_2(x) - M)$, other items will be left out. Therefore:

$$\tau_1 + \tau_2 = \textstyle\sum_{i \in (S_1 \cup S_2)} (\hat{c}_1(i) + \hat{c}_2(i))$$

$$\tau_{new} = \textstyle\sum_{i \in (S_{new})} (\hat{c}_1(i) + \hat{c}_2(i) - M)$$

Combining the results above together and let the k^{th} largest item be K, M is equal to $\hat{c}_1(K) + \hat{c}_2(K)$ as defined above, thus:

$$\begin{aligned} \tau_1 + \tau_2 &= \tau_{new} + (k-1) \times M + \sum_{i \in (S_1 \cup S_2) - S_{new}} (\hat{c}_1(i) + \hat{c}_2(i)) \\ &= \tau_{new} + k \times M + \sum_{i \in (S_1 \cup S_2) - (S_{new} \cup \{K\})} (\hat{c}_1(i) + \hat{c}_2(i)) \end{aligned} \tag{3}$$

$$\text{since } \textstyle\sum_{i \in (S_1 \cup S_2) - (S_{new} \cup \{K\})} (\hat{c}_1(i) + \hat{c}_2(i)) \geq 0 \text{ (insertion-only stream)}$$

$$\text{thus } \tau_{new} + k \times M \leq \tau_1 + \tau_2$$

If the result size of $S_1 \cup S_2$ is just $k-1$ (no need to decrease counts), then M and $\sum_{i \in (S_1 \cup S_2) - S_{new}} (\hat{c}_1(i) + \hat{c}_2(i))$ could all be considered as 0, $\tau_{new} = \tau_1 + \tau_2$, the inequality above still applies. So that:

$$M \leq \frac{\tau_1 + \tau_2 - \tau_{new}}{k} \tag{4}$$

Because the two streams are insertion-only and concatenated, so that the total frequencies of $\sigma_1 \cdot \sigma_2$ could be calculated by adding up the total frequencies of σ_1 and σ_2, thus $m_{new} = m_1 + m_2$. By replacing M, it is easy to show that:

$$\begin{aligned} f_{new,y} - \hat{c}_{new}(y) &\leq \frac{(m_1 + m_2) - (\tau_1 + \tau_2)}{k} + M \\ &\leq \frac{(m_1 + m_2) - (\tau_1 + \tau_2)}{k} + \frac{\tau_1 + \tau_2 - \tau_{new}}{k} \\ &= \frac{(m_1 + m_2) - \tau_{new}}{k} \\ &= \frac{(m_{new}) - \tau_{new}}{k} \\ \textit{Consequently, } \hat{c}_{new}(y) &\geq f_{new,y} - \frac{(m_{new}) - \tau_{new}}{k} \end{aligned} \tag{5}$$

Therefore, inequality (1) holds for all item $y \in S_{new}$.