Hypersistent Sketch Mathematical Analysis

1 Persistence Estimation

1.1 Error Bound

THEOREM 4.1 Let $\hat{p}_i$ is the estimated persistence of our method.We have

$$p_i \leq \hat{p}_i \leq T \tag{1}$$

PROOF.Let the threshold of L_1 be M_1, the threshold of L_2 be M_2.The number of time Windows in S is T.When $\hat{p}_i > M_1$, the item is transmitted to L_2, $\hat{p}_i > M_1 + M_2$, the item is transmitted to L_3. Correspondingly, the number of hash functions in L_1,L_2,L_3 is d_1,d_2 and d_3.So $\hat{p}_i$consists of three parts.$\hat{p}_i = \hat{p}_i^1$+$\hat{p}_i^2$+$\hat{p}_i^3$.respectively,where $\hat{p}_i^1$,$\hat{p}_i^2$ and $\hat{p}_i^3$ are estimated persistence in L_1,L_2 and L_3.Next, we find the upper and lower boundaries of $\hat{P}_i$. For $\hat{p}_i^j$,$(j = 1, 2, 3)$,each item arrived can cause the mapping,make $\hat{p}_i^j$,$(j = 1, 2, 3)$ increased, so the $\hat{p}_i^j \geq p_i^j(j = 1, 2, 3)$.For each time window, at most $\hat{p_i}^j$ increases by 1, so $\hat{p}_i \leq T$.

To sum up, $p_i \leq \hat{p}_i \leqslant T$.

THEOREM 4.2 For convenience, we set $\hat{p}_i < M_1$ as small data streams, when $\hat{p}_i \leq M_1 + M_2$ as medium data streams, $M_1 + M_2 \leq \hat{p}_i$.
For the medium data streams, let $\varDelta_j p_i = \varDelta_j p_i^1$+$\varDelta_j p_i^2$. Where $\varDelta_j p_i^1 = C_j^1[h_j(e_i)]$− p_i^1, $\varDelta_j p_i^2 = C_j^2[h_j(e_i)] - p_i^2$. Let E_t be the set of all items that differ significantly in the time window T. p_i is the set of time windows where e_i occurs. $\bar{P}_i = \{1, 2, ..., T\} - P_i$. Let

$$I_{i,j,t} = \begin{cases} 1, \text{ if } \exists e_k \in E_t, i \neq k \bigvee h_j^{1,2}(e_i) = h_j^{1,2}(e_k) \\ 0, \qquad\qquad\qquad \text{Otherwise} \end{cases}$$

$$E[\varDelta_j p_i] = E[\varDelta_j^1 p_i] + E[\varDelta_j^2 p_i]$$

$$= \sum_{t \in \bar{P}_i} \left[1 - \left(1 - \frac{1}{l_1} \right)^{|E_t|} \right] \times \left[1 - \left(1 - \frac{1}{l_1} \right)^{|E_t|'} \right] \tag{2}$$

Where $|E_t|'$ is the set of all items in $\hat{p}_i > M_1 + M_2$. In practice, the number of such items is relatively small. We let

$$\|p\|_1 = \sum_{i=1}^{N} p_i = \sum_{t=1}^{T} |E_t| = \|E\|_1$$

$\|p\|_1^1 = \sum_{p_i > M_1} p_i$ is the sum of items whose estimated persistence is greater than M_1. Therefore, we have

$$E\left[\varDelta_j p_i\right] \leq \sum_{t \in \bar{p}_i} \frac{|E_t|}{l_1} \times \frac{|E_t|^1}{l_2} \leq \frac{\|p\|_1 \times \|p\|_1^1}{l_1 \times l_2}$$

It is known from the increment of the limit function, we can deduce

$$E\left[\varDelta_j p_i\right] \geq \sum_{t \in \bar{p}_i} \left[1 - \left(\frac{1}{e} \right)^{\frac{|E_t|}{l_1}} \right] \times \left[1 - \left(\frac{1}{e} \right)^{\frac{|E_t|\prime}{l_2}} \right].$$

1.2 Comparison with Related Work

THEOREM 4.3 Let $\hat{p}_i^{On-Off}$ be the estimated persistence of the On-Off Sketch(OO). For the same memory condition, we filter the entire data stream by assigning a different number of counters and d_i for small, medium, and large data streams. Let $\mathit{\Delta}_j^{OO} p_i = C_j[h_j(e_i)] - p_i$, where $p_i = \min_{1 \leq j \leq d}(C_j[h_j(e_i)])$.

For small data streams, we use low-byte storage, meaning the number of counters is higher. From the expression

$$E[\mathit{\Delta}_j p_i] = \sum_{t \in \bar{P}_i} \left[1 - \left(1 - \frac{1}{l} \right)^{|E_t|} \right], \tag{3}$$

we can deduce that

$$E\left(\mathit{\Delta}_j p_i\right) \leq E\left(\mathit{\Delta}_j^{OO} p_i\right). \tag{4}$$

For the medium stream, we have:

$$1 - \left(1 - \frac{1}{l_2} \right)^{|E_t|'} < 1.$$

Therefore, we conclude:

$$E\left(\mathit{\Delta}_j^{OO} p_i\right) > 1 - \left(1 - \frac{1}{l_1} \right)^{|E_t|} > E\left(\mathit{\Delta}_j p_i\right). \tag{5}$$

For the large stream, the analysis is similar to the above. To sum up, the above inequality holds true.

2 Finding Persistent Items

2.1 Comparison with Related Work

Our method is better than the On-Off method.

THEOREM 4.4 We let $\hat{p}_i = B\left[h_1\left(e_i\right)\right]\left[e_i\right]$, and the On-Off Sketch is denoted as $\hat{p}_i^{\text{on-off}}$. We have:

$$p_i \leq \hat{p}_i \leq \hat{p}_i^{\text{on-off}} \leq T \tag{6}$$

PROOF. The On-Off method uses an alternative way to record persistence. Our method filters use cold-item filter based on it, so that the probability of hash collision brought by other items will be reduced when the estimator is recorded. We assume in the initial state there is $\hat{p}_i = \hat{p}_i^{\text{On-Off}}$. When a new item comes, there will be the following cases:

Case 1: When e_i arrives, if e_i is in $B\left[h_1\left(e_i\right)\right]\left[e_i\right]$, whether it is insertion or collision, if the number of counters is the same, from the perspective of probability, the above equation is still valid. That is, when e_i arrives under normal circumstances, it can be considered that the above equation remains valid. If e_i

is not in $B\left[h_1\left(e_i\right)\right]\left[e_i\right]$, whether or not insertion and hash collision occur has no effect on the above equation.

Case 2: When $e_j (i \neq j)$ comes, if $C_1\left[h_1\left(e_i\right)\right] = C_1\left[h_1\left(e_j\right)\right]$, this will cause errors due to collisions. However, the method in this paper filters many cold items, so the flow when reaching the record will be much less, meaning $\left|e_j\right| < \left|e_j^{\text{On-Off}}\right|$, resulting in a smaller $\hat{p}$. Otherwise, there is no change.

Finally, we can conclude that the equation holds for all cases.