F Visualization on Size Distribution

A better-estimated distribution brings a better compression rate. Here, we visualize the distribution on $|S|$ to verify that our method is a better estimator.

The distribution of our model is just the $P(S;r)$ defined in section 5.1. For the distribution of single-item-based method, let $q_i = \Pr(i \in S)$, then the $Q(S) = \Pi_{i \in S} q_i \cdot \Pi_{j \in V-S}(1-q_j)$.

The exact size distribution of $P(S;r)$ and $Q(S)$ can be calculated by ordinary generating function efficiently.

F.1 Calculation of the size distribution

F.1.1 Ordinary Generating Function. We define the ordinary generating function(OGF) of a size distribution as follows.

$$F(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots$$

where the a_i is the $\Pr(|S| = i)$, for $i \in \mathbb{N}$.

F.1.2 Calculation of the Size Distribution of Single-item-based Model. For $i \in V$, let $q_i = Pr(i \in S)$. The probability of a $S \subseteq V$ for the single-item-based model is

$$Q(S) = \Pi_{i \in S} q_i \cdot \Pi_{j \in V-S}(1-q_j)$$

We can calculate the OGF of $Q(S)$ by the following steps.

(1) F(x)=1
(2) For $i = 1, \dots, K$,

$$F(x) := F(x) \times ((1-q_i) + q_i x)$$

With $O(K)$ steps, we can obtain the size distribution of $Q(S)$.

F.1.3 Calculation of the Size Distribution of binary tree Model. For nodes l in the binary tree, we can recursively calculate the OGF $F_v(x)$ of $P(S;v)$. The notations are the same as section 5.1.

For a leaf node l, the OGF $F_l(x) = x$. For an internal node v, let v's left and right child be u and w, we have

$$F_v(x) = p(0;v)F_u(x) + p(1;v)F_w(x) + p(2;v)F_u(x)F_w(x)$$

With $O(2K+1)$ steps, we can obtain the OGF $F_r(x)$ of the root node r.

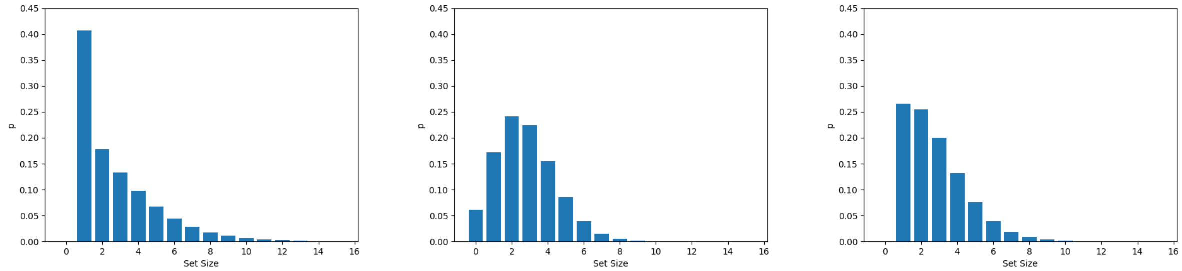
F.2 Results

The figure 10 and figure 11 visualizes the size distribution of ground truth distribution $P^*(S)$, $Q(S)$ and $P(S;r)$ on the Tmall dataset. As illustrated in the figures, the $Q(S)$ has a positive value on $|S| = 0$, while there is zero probability for $|S| = 0$ in $P^*(S)$ and $P(S;r)$.

Both the ground truth size distribution and our model's size distribution have a similar fast-decaying shape. While the size distribution of the single-item-based model has a different bell shape like Gaussian.

The visualization of the size distribution further confirmed that our binary tree model matches the pattern of real-world data better and thus brings us a lower compression rate.

Received 1 Feb 2025

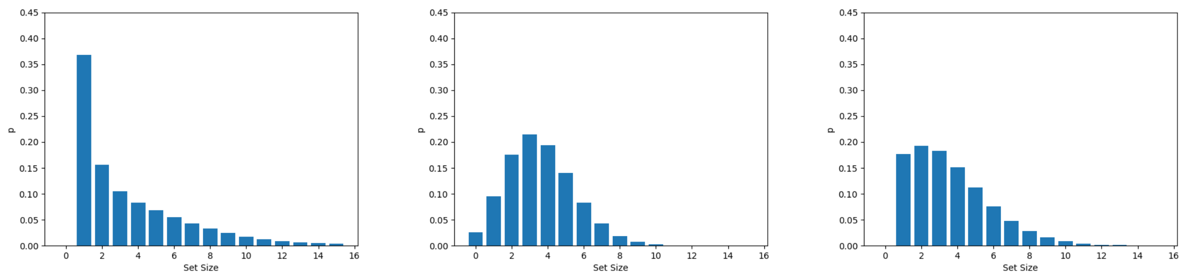


(a) Size distribution in dataset $\mathcal{S}$

(b) Size distribution of single-item-based

(c) Size distribution of Ours

Figure 10: The size distribution on the Tmall dataset. We only present the proportion of $|S| \leq 16$. Figure (10a): The empirical size distribution in the Tmall dataset. Figure (10b): The size distribution of the single-item-based baseline model learned by the Tmall dataset. Figure (10c): The size distribution of the binary tree model learned by the Tmall dataset.



(a) Size distribution in dataset $\mathcal{S}$

(b) Size distribution of single-item-based

(c) Size distribution of Ours

Figure 11: The size distribution on the HKTVmall dataset. We only present the proportion of $|S| \leq 16$. Figure (11a): The empirical size distribution in the HKTVmall dataset. Figure (11b): The size distribution of the single-item-based baseline model learned by the HKTVmall dataset. Figure (11c): The size distribution of the binary tree model learned by the HKTVmall dataset.