Big Data Computing

Written test 07/06/2018
(SOLUTIONS of EXERCISES)

Problema. Let P be a set of N *bicolored points* from a metric space, partitioned into k clusters $C_1, C_2, \ldots, C_k$. Each point $x \in P$ is initially represented by the key-value pair $(\mathrm{ID}_x, (x, i_x, \gamma_x))$, where ID_x is a distinct key in $[0, N-1]$, i_x is the index of the cluster which x belongs to, and $\gamma_x \in \{0,1\}$ is the color of x.

1. Design an efficient *2-round MapReduce algorithm* that for each cluster C_i checks whether all points of C_i have the same color. The output of the algorithm must be the k pairs (i, b_i), with $1 \le i \le k$, where $b_i = -1$ if C_i contains points of different colors, otherwise b_i is the color common to all points of C_i.

2. Analyze the local and aggregate space required by your algorithm.

N.B. *For full score, your algorithm must require $o(N)$ local space and $O(N)$ aggregate space.*

Soluzione.

1. The 2-round MapReduce algorithm is the following.

   Round 1:

   - *Map phase*: map each pair $(\mathrm{ID}_x, (x, i_x, \gamma_x))$, into the intermediate pair $(\mathrm{ID}_x \bmod \sqrt{N}, (x, i_x, \gamma_x))$ (for simplicity, assume that $\sqrt{N}$ is an integer).
   - *Reduce phase*: for each key ℓ independently, with $0 \le \ell < \sqrt{N}$, gather all intermediate pairs with key ℓ and let P_ℓ be the subset of points represented by these pairs. For every cluster C_i such that $C_i \cap P_\ell \ne \emptyset$ produce the pair $(i, b_i(\ell))$, where i is the key and $b_i(\ell) = -1$, if $C_i \cap P_\ell$ contains points of different colors, otherwise $b_i(\ell)$ is the color common to all points of $C_i \cap P_\ell$.

   Round 2:

   - *Map phase*: identity mapping.
   - *Reduce phase*: for each $1 \le i \le k$ independently, gather the at most $\sqrt{N}$ pairs $(i, b_i(\ell))$, and return (i, b_i) $b_i = 0$ if all $b_i(\ell)$'s are 0, $b_i = 1$ if all $b_i(\ell)$'s are 1, and $b_i = -1$ in all other cases.

2. In Round 1, there are at most $\sqrt{N}$ input pairs mapped to the same key ℓ, hence to the same reducer, while in the reduce phase of Round 2 at most $\sqrt{N}$ pairs are gathered for each i. Therefore, the required local space is $M_L = \Theta\left(\sqrt{N}\right)$. Also, the total number of pairs produced by the map and reduce phases of Round 1 are $\Theta(N)$. Therefore, the required aggregate space is $M_A = \Theta\left(\sqrt{N}\right)$.

$\square$

Problema. Let T be a dataset of N transactions over a set I of d items and suppose that *every itemset* $X \subseteq I$ *is closed* (i.e., every superset $Y \supset X$ has smaller support). Let $X_1, X_2, \ldots$ be the sequence of itemsets by nonincreasing support, including the empty itemset (X_1) which has support 1. For a given $K > 1$, let $s = \text{Supp}_T(X_K)$ and $s' = \text{Supp}_T(X_{K+1})$ and assume that $1 > s > s' > 0$.

1. Show that for every itemset X_j of support s' (hence, $j \geq K+1$), and for every item $a \in X_j$, the itemset $X_j - \{a\}$ belongs to $\{X_1, X_2, \ldots, X_K\}$.

2. Derive an upper bound to the number of itemsets of support exactly s' and, from this, an upper bound to the number of Top-$(K+1)$ frequent itemsets. (**Hint:** *note that the previous point implies that any itemset of support* s' *is obtained by adding an item to some* X_i *with* $1 \leq i \leq K$.)

Soluzione.

1. Since every itemset is closed, we have that $\text{Supp}_T(X_j - \{a\}) > \text{Supp}_T(X_j)$, therefore $\text{Supp}_T(X_j - \{a\}) > s'$ which implies that $\text{Supp}_T(X_j - \{a\}) \geq s$. Since $\{X_1, \ldots, X_K\}$ are all itemsets of support $\geq s$, the itemset $X_j - \{a\}$ must be one of them.

2. Since each X_i, with $1 \leq i \leq K$, induces at most d itemsets of support s', this implies that the itemsets of support exactly s' are at most dK. Clearly, the *Top*-$(K+1)$ *frequent itemsets* are all itemsets of support $\geq s'$, that is, all itemsets of support $\geq s$ plus those of support exactly s', which are at most $K + dK = (d+1)K$.

□