MinHash sketches

As discussed in class, for a min-wise independent family H, we can associate a sketch

$$s(X) = \langle min\ h_1(X), min\ h_2(X), \dots min\ h_k(X)\rangle$$

with each set X in the given data collection, where $h_1, h_2, \dots, h_k$ are independently chosen at random from H. Consider now any two sets A and B, with their sketches $s(A)$ and $s(B)$. Can you compute a sketch for $A \cup B$ using just $s(A)$ and $s(B)$ in $O(k)$ time? Can you prove that it is equivalent to compute $s(A \cup B)$ from scratch directly from $A \cup B$?

SOLUTION

We have the two sketches of A and B:

$$s(A) = \langle min\ h_1(A), min\ h_2(A), \dots min\ h_k(A)\rangle$$
$$s(B) = \langle min\ h_1(B), min\ h_2(B), \dots min\ h_k(B)\rangle$$

We want to compute the sketch for $A \cup B$:

$$s(A \cup B) = \langle min\ h_1(A \cup B), min\ h_2(A \cup B), \dots min\ h_k(A \cup B)\rangle$$

For each $i \in 1, 2, \dots k$ we will take $min\ \{\ min\ h_i(A),\ min\ h_i(B)\ \}$ as $min\ h_i(A \cup B)$. This procedure require a constant number of operation (exactly one min) and must be repeated k times, so the time needed is $O(k)$.

In order to prove that what we obtain is exactly the sketch of $A \cup B$ we exploit the following lemma.

Lemma 1. *For each $i \in \{1, 2, \dots k\}$ it holds $h_i(A \cup B) = h_i(A) \cup h_i(A)$.*

Proof. take any x, then all the following are equivalent:

$$\begin{aligned}
&x \in h_i(A \cup B)\\
&\exists y \in A \cup B\ .\ h_i(y) = x\\
&\exists y \in A\ .\ h_i(y) = x \quad OR\ \ \exists y \in A\ .\ h_i(y) = x\\
&x \in h_i(A) \qquad OR \qquad x \in h_i(B)\\
&x \in h_i(A) \cup h_i(B)
\end{aligned}$$

□

Finally.

Theorem 2. *For each $i \in \{1, 2, \dots k\}$ it holds $min\ h_i(A \cup B) = min\ \{\ min\ h_i(A),\ min\ h_i(B)\ \}$.*

Proof. take any $i \in \{1, 2, \dots k\}$, then

$$\begin{aligned}
&min\ h_i(A \cup B) = && [using\ lemma 1]\\
&min\ (h_i(A) \cup h_i(B)) = && [trivial\ property\ of\ min]\\
&min\ \{min\ h_i(A),\ min\ h_i(B)\}
\end{aligned}$$

□