1 Hashing

Let $U$ be an universe and $S$ be a set. We would like to store $S$ in a hash table $T$. Suppose $h$ is function from $U$ to $T$ that is one-one on $S$. Then we can use $h$ to store $S$ as follows: Given an element $k \in S$, we store it at location $T[h(x)]$. In future, if we want to search for a key $a$, then we compare $a$ with $T[h(a)]$.

Now the question is how to find such a $h$. Observe that if we allow $|T| = |U|$, then this is trivial. However we want $|S| \approx |T|$. Let $|U| = N$, $|S| = m$, we first consider the case when $|T| = t = m^2$. Our goal is find a function $h$ from $U$ to $T$ such that for every $x \neq y \in S$, $h(x) \neq h(y)$. The solution is simple: Pick a random function. We show that a random function from is one-one on $S$ with good probability.

Consider the following collision pair set:

$$C = \{ \langle x, y \rangle \mid h(x) = h(y), x \neq y, x \in S, y \in S \}$$

Clearly $h$ is one-one on $S$ if and only if $C = \emptyset$. If we randomly pick $h$, what is the expectation of $|C|$?

For every $x$ and $y$ in $S (x \neq y)$, consider the following random variable

$$C_{xy} = \begin{cases} 1 & \text{if } h(x) = h(y) \\ 0 & \text{else} \end{cases}$$

Clearly,

$$|C| = \sum_{x, y \in S, x < y} C_{xy}$$

Thus $C$ is random variable that denotes the number of colliding pairs.

$$E(C_{xy}) = \Pr_{h \in H} ( h(x) = h(y) ) = \frac{1}{t}$$

$$E(|C|) = \sum_{x, y \in S, x \neq y} \frac{1}{t} = \left( \frac{N}{2} \right) \frac{1}{t}$$

$$if \ t = m^2, E(C) = \frac{m(m-1)}{2} \frac{1}{m^2}$$
\[ E(C) \leq \frac{1}{2} \]

Using Markov inequality, we obtain that

\[ \Pr(C \leq 1) \geq \frac{1}{2} \]

Thus, if we randomly pick a hash function then probability that there is no collision is at least half. If we randomly pick 100 hash functions then at least one of them is one-one on \( S \) is probability bigger than \( 1 - 1/2^{100} \).

Thus the time required to find \( h \) is \( O(m) \), the space required to store \( S \) in the size of \( T \) which is \( m^2 \). The query time is \( O(1) \). Next we see how two reduce \( |T| = O(N) \). For this we use a two-stage hashing. We will ignore the space required to store the hash function.

If \( t = m \),

\[ E(|C|) \leq \frac{m}{2} \]

By Markov inequality,

\[ \Pr(|C| \geq m) \leq \frac{1}{2} \Rightarrow \Pr(|C| < m) \geq \frac{1}{2} \]

Thus for a random \( h \), expected number of colliding pairs is less than \( m \), with probability at least half. Pick such function \( h \). This is our hash function in first stage.

For \( i \in T \), let \( M_i \) is the set of elements from \( S \) that are mapped to \( i \) via \( h \). Let \( |M_i| = m_i \). Now, store \( M_i \) is in a secondary table of size \( m_i^2 \).

The total size of the table is given by

\[ m + \sum m_i^2 + \text{space to store all hash functions} \]

Observe that

\[ |C| = \sum_{i \in T} \left( \frac{m_i}{2} \right), \]

and

\[ \sum_{i \in T} \left( \frac{m_i}{2} \right) = \sum \frac{m_i^2}{2} - \sum \frac{m_i}{2} \]

Thus

\[ 2|C| = \sum m_i^2 - \sum m_i, \]

Since \( |C| = N \) and \( \sum m_i \leq m, \sum m_i^2 \leq 3m \).

Thus the total space used (ignoring the space used for hash functions) is \( O(m) \).

In general, we would like to answer the following question? Let \( H \) be a family of hash functions from \( U \) to \( T \). Let \( S \) be any subset of \( U \). If we randomly pick a function from \( H \), then “how much one-one” is \( h \) on \( S \)? We now answer this question.
Let $|U| = N, |S| = m, |T| = t$. An element $i \in S$ is said to be unique with respect to $h$ if $h^{-1}[h(i)] \cap S = \{i\}$

We would like to know how many unique elements can $S$ have?. Let

\[ X_i = \begin{cases} 
1 & \text{if } i \text{ is unique} \\
0 & \text{else} 
\end{cases} \]

\[ E(X_i) = \Pr(X_i = 1) = 1 - \Pr(X_i = 0) \]

\[ \Pr(X_i = 0) = \Pr(\exists j \neq i \text{ and } j \in S, \ h(i) = h(j)) \]

\[ \Pr(X_i = 0) \leq \sum_{j \in S, j \neq i} \Pr(h(i) = h(j)) \text{ by union bound} \]

\[ \leq \frac{m - 1}{t} \]

Therefore

\[ \Pr(X_i = 1) = 1 - \frac{m}{t} + \frac{1}{t} \]

\[ E(X_i) \geq 1 - \frac{m}{t} \]

Let $X =$ the number of unique elements of $S$. i.e.

\[ X = \sum_{i \in S} X_i \]

\[ E(X) \geq m - \frac{m^2}{t} \]

In particular if we take $T = km$, then $E(X) = (1 - 1/k)m$. 

3