HASH TABLES

1. Introduction

2. Dictionary data type

Definition 2.1. Let $\mathcal{X}$ be a set of *items* and let $\mathcal{U}$ be a set of *keys*. Consider an abstract data type D which dynamically stores a collection of pairs (k,x) where $k \in \mathcal{U}$ and $x \in \mathcal{X}$ in such a way that D does not store two pairs having the same key at the same time. Moreover, we assume that D supports the following operations.

> INSERT$\big(D,(k,x)\big)$
> Adds pair (k,x) into D if there is no other pair stored in D with k as a first entry.
>
> DELETE$\big(D,k\big)$
> Removes a pair with k as a first entry from D if such pair is stored in D.
>
> SEARCH$\big(D,k\big)$
> Returns x if a pair (k,x) is stored in D. Otherwise returns *nil*.

An abstract data type with these properties and interface is called *an associative array* or *a dictionary*.

Definition 2.2. Let $\mathcal{X}$ and $\mathcal{U}$ be sets. *Dictionary problem for $\mathcal{X}$ and $\mathcal{U}$* is the task of designing a dictionary with $\mathcal{X}$ as the set of items and $\mathcal{U}$ as the set of keys.

3. Hash functions

In this section we introduce the important notion of a hash function and we discuss some probabilistic properties of such functions.

Definition 3.1. Let $\mathcal{U}$ be a set. *A hash function* is a mapping $h : \mathcal{U} \to \{0,1,...,m-1\}$ where $m \in \mathbb{N}_+$.

Definition 3.2. Let $h : \mathcal{U} \to \{0,1,...,m-1\}$ be a hash function. *A collision* is a pair of keys $k_1, k_2 \in \mathcal{U}$ such that $k(k_1) = h(k_2)$.

Definition 3.3. Let X be a set and let $n \in \mathbb{N}_+$. Then a set

$$X^{\wedge n} = \big\{(x_1,...,x_n) \in X^n \,\big|\, \forall_{1\le i<j\le n}\, x_i \neq x_j\big\}$$

is called *the antisymmetric cartesian power of X*.

Definition 3.4. Let $\mathcal{U}$ be a measurable space. We consider $\mathcal{U}^{\wedge n}$ as the measurable subspace of the product space $\mathcal{U}^n$. Suppose that P is a probability distribution on $\mathcal{U}^{\wedge n}$. Let $h : \mathcal{U} \to \{0,1,...,m-1\}$ be a measurable hash function for some $m \in \mathbb{N}_+$. Assume that the following assertions hold.

(1)

$$P\big((k_1,...,k_n) \in \mathcal{U}^{\wedge n} \,\big|\, h(k_i) = l\big) = \frac{1}{m}$$

for every element $i \in \{1,...,n\}$ and every $l \in \{0,1,...,m-1\}$.

(2)

$$P\big((k_1,...,k_n) \in \mathcal{U}^{\wedge n} \,\big|\, h(k_i) = h(k_j)\big) \le \frac{1}{m}$$

for every pair of distinct elements $i,j \in \{1,...,n\}$.

Then h is *a simple uniform hashing with respect to P*.

Example 3.5. Let $\mathcal{U} = [0, m]$ for some $m \in \mathbb{N}_+$. Then $\mathcal{U}$ is a measurable space with respect to Borel algebra $\mathcal{B}\big([0, m]\big)$. We define a hash function $h : \mathcal{U} \to \{0, 1, ..., m-1\}$ by formula

$$h(x) = \lfloor x \rfloor$$

Then h is a simple uniform hashing with respect to the normalization of n-dimensional Lebesgue measure on $[0, m]^{\wedge n}$.

Example 3.6. Let $\mathcal{U} = \{0, 1, ..., m^2 - 1\}$ for some $m \in \mathbb{N}_+$. Then $\mathcal{U}$ is a measurable space with respect to the power algebra $\mathcal{P}\big(\{0, 1, ..., m^2 - 1\}\big)$. Consider $\mathcal{U}^{\wedge n}$ as a probability space with respect to the uniform distribution P. We define a hash function $h : \mathcal{U} \to \{0, 1, ..., m-1\}$ by formula

$$h(x) = x \bmod m$$

For $i \in \{1, ..., n\}$ and $l \in \{0, 1, ..., m-1\}$ we have

$$P\big((k_1, ..., k_n) \in \mathcal{U}^{\wedge n} \,\big|\, h(k_i) = l\big) = \frac{m \cdot (m^2 - 1) \cdot (m^2 - 2) \cdot ... \cdot (m^2 - n + 1)}{m^2 \cdot (m^2 - 1) \cdot ... \cdot (m^2 - n + 1)} = \frac{1}{m}$$

Fix distinct $i, j \in \{1, ..., n\}$ and $l \in \{0, 1, ..., m-1\}$. Note that

$$P\big((k_1, ..., k_n) \in \mathcal{U}^{\wedge n} \,\big|\, h(k_i) = l, h(k_j) = l\big) = \frac{m \cdot (m-1) \cdot (m^2 - 2) \cdot ... \cdot (m^2 - n + 1)}{m^2 \cdot (m^2 - 1) \cdot ... \cdot (m^2 - n + 1)} = \frac{1}{m \cdot (m+1)}$$

Hence

$$P\big((k_1, ..., k_n) \in \mathcal{U}^{\wedge n} \,\big|\, h(k_i) = h(k_j)\big) = \sum_{l=0}^{m-1} P\big(h(k_i) = l, h(k_j) = l \,\big|\, (k_1, ..., k_n) \in \mathcal{U}^{\wedge n}\big) =$$

$$= \frac{m}{m \cdot (m+1)} = \frac{1}{m+1} \leq \frac{1}{m}$$

Thus h is a simple uniform hashing with respect to P.

4. Hash tables with chaining as a solution to dictionary problem

In this section we present the solution to the dictionary problem and discuss its efficiency.

Definition 4.1. Let $\mathcal{U}$ and $\mathcal{X}$ be sets. Let $h : \mathcal{U} \to \{0, 1, ..., m-1\}$ be a hash function for some $m \in \mathbb{N}_+$. We consider an m-element array D_h such that $D_h[l]$ is a linked list storing values from $\mathcal{U} \times \mathcal{X}$ for every $l \in \{0, 1, ..., m-1\}$. We describe dictionary operations.

INSERT$\big(D_h, (k, x)\big)$
Inserts pair (k, x) to the linked list $D_h[h(k)]$ as its new head.

DELETE$\big(D_h, k\big)$
Deletes a pair with first entry k from the linked list $D_h[h(k)]$.

SEARCH$\big(D_h, k\big)$
Searches for the pair with the first entry k in the list $D_h[h(k)]$. If such pair is found, then returns its second entry. Otherwise returns *nil*.

Then D_h together with these operations is a solution of dictionary problem for $\mathcal{U}$ and $\mathcal{X}$. We call it *the hash table with collisions resolved by chaining for h*.

Suppose that $\mathcal{U}$ and $\mathcal{X}$ are sets. Let $h : \mathcal{U} \to \{0, 1, ..., m-1\}$ be a hash function. Consider the hash table D_h. Fix $l \in \{0, 1, ..., m-1\}$ and $n \in \mathbb{N}_+$. Suppose that pairs $(k_1, x_1), ..., (k_n, x_n)$ for $(k_1, ..., k_n) \in \mathcal{U}^{\wedge n}$ and $x_1, ..., x_n \in \mathcal{X}$ are consecutively inserted to initially empty D_h. After these sequence of insertions is performed the length of the linked list stored in $D_h[l]$ is equal to the cardinality of the set $\big\{i \,\big|\, h(k_i) = l\big\}$. We denote the function

$$\mathcal{U}^{\wedge n} \ni (k_1, ..., k_n) \to \Big|\big\{i \,\big|\, h(k_i) = l\big\}\Big| \in \mathbb{N}$$

by $coll_l$.

Theorem 4.2. *Let $\mathcal{U}$ be a measurable space and let $\mathcal{X}$ be a set. Let $h : \mathcal{U} \to \{0, 1, ..., m-1\}$ be a measurable hash function and fix $n \in \mathbb{N}_+$. Then the following assertions hold.*

(1) *The function $coll_l : \mathcal{U}^{\wedge n} \to \mathbb{N}$ is measurable for every $l \in \{0, 1, ..., m-1\}$.*

(2) *If h is a simple uniform hashing with respect to some probability distribution P on $\mathcal{U}^{\wedge n}$, then*

$$\mathbb{E}\, coll_l = \int_{\mathcal{U}^{\wedge n}} coll_l \, dP = \frac{n}{m}$$

for every $l \in \{0, 1, ..., m-1\}$.

Proof. Suppose that X_i is the indicator function of the measurable set $\{(k_1, ..., k_n) \in \mathcal{U}^{\wedge n} \,|\, h(k_i) = l\}$. Then

$$coll_l = \sum_{i=1}^{n} X_i$$

This proves that $slot_l$ is measurable. If in addition h is a simple uniform hashing with respect to some probability distribution P on $\mathcal{U}^{\wedge n}$, then

$$\mathbb{E}\, coll_l = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E} X_i = \sum_{i=1}^{n} P\big((k_1, ..., k_n) \in \mathcal{U}^{\wedge n} \,\big|\, h(k_i) = l\big) = \frac{n}{m}$$

□

Suppose that $\mathcal{U}$ and $\mathcal{X}$ are sets. Let $h : \mathcal{U} \to \{0, 1, ..., m-1\}$ be a hash function. Consider the hash table D_h. Fix $n \in \mathbb{N}_+$ and $i \in \{1, ..., n\}$. Suppose that pairs $(k_1, x_1), ..., (k_n, x_n)$ for $(k_1, ..., k_n) \in \mathcal{U}^{\wedge n}$ and $x_1, ..., x_n \in \mathcal{X}$ are consecutively inserted to initially empty D_h. After these sequence of insertions is performed the number of elements in the linked list stored in $D_h[k(k_i)]$ which precede (k_i, x_i) is equal to the cardinality of the set $\{j \in \{i+1, ..., n\} \,|\, h(k_j) = h(k_i)\}$. We denote the function

$$\mathcal{U}^{\wedge n} \ni (k_1, ..., k_n) \to \big|\{i \,|\, h(k_i) = l\}\big| \in \mathbb{N}$$

by $coll_{<i}$.

Theorem 4.3. *Let $\mathcal{U}$ be a measurable space and let $\mathcal{X}$ be a set. Let $h : \mathcal{U} \to \{0, 1, ..., m-1\}$ be a measurable hash function and fix $n \in \mathbb{N}_+$. Fix $i \in \{1, ..., n\}$. Then the following assertions hold.*

(1) *The function*

$$\mathcal{U}^{\wedge n} \ni (k_1, ..., k_n) \mapsto \#\{j \,|\, i \leq j \leq n \text{ and } h(k_i) = h(k_j)\} \in \mathbb{N}$$

is measurable.

(2) *If h is a simple uniform hashing with respect to some probability distribution P on $\mathcal{U}^{\wedge n}$, then*

$$\mathbb{E}\#\{j \,|\, i \leq j \leq n \text{ and } h(k_i) = h(k_j)\} = \int_{\mathcal{U}^{\wedge n}} \#D_h[l] \, dP = \frac{n}{m}$$

for every $l \in \{0, 1, ..., m-1\}$.

Proof. Suppose that X_i is the indicator function of the measurable set $\{(k_1, ..., k_n) \in \mathcal{U}^{\wedge n} \,|\, h(k_i) = l\}$. Then

$$\#D_h[l] = \sum_{i=1}^{n} X_i$$

This proves that $\#D_h[l]$ is measurable. If in addition h is a simple uniform hashing with respect to some probability distribution P on $\mathcal{U}^{\wedge n}$, then

$$\mathbb{E}\#D_h[l] = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E} X_i = \sum_{i=1}^{n} P\big((k_1, ..., k_n) \in \mathcal{U}^{\wedge n} \,\big|\, h(k_i) = l\big) = \frac{n}{m}$$

□