Chapter 8

Dictionaries & Hashing

You manage a library and want to be able to quickly tell whether you carry a
given book or not. We need the capability to insert, delete, and search books.

Definition 8.1 (Dictionary). A dictionary is a data structure that manages
a set of objects. Each object is uniquely identified by its key. The relevant
operations are:

• search: find an object with a given key
• insert: put an object into the set
• delete: remove an object from the set

Remarks:

• There are alternative names for dictionary, e.g. key-value store, associ-ative array, map, or just set.
• If the dictionary only offers search, it is called static; if it also offers insert and delete, it is dynamic.
• For our purposes, we will ignore that we actually have a set of ob-
jects, each of which is identified by a unique key, and just talk about
the set of keys. With regard to the library example, books are glob-
ally uniquely identified by a key called ISBN. Whenever we say we
insert/delete/search a key, we can just drag the key’s object along.
• The classic data structure for dictionaries is a binary search tree.

8.1 Search Trees

Definition 8.2 (Binary search tree). A binary search tree is a rooted tree
(Definition 2.7), where each node stores a key. Additionally, each node may
have a pointer to a left and/or a right child tree. For all nodes, if existing, the
nodes in the left child tree store smaller keys, and those in the right child tree
store larger keys.

The classic data structure for dictionaries is a binary search tree.

Algorithm 8.3 Search Tree: Search
Input: key k, root r of search tree
Output: k if it is in the tree, else ⊥.
1: If r contains key k, return k
2: If k is smaller than the key of r, set r to left child and recurse
3: If k is larger than the key of r, set r to right child and recurse
4: Return ⊥.

Remarks:

• The symbol ⊥ (“bottom”) signifies null or undefined.
• The cost of searching in a binary search tree is proportional to the
depth of the key, which is the distance (Theorem 2.15) between
the node with the key and the root.
• There are search trees called splay trees that keep frequently searched
keys close to the root. There may be keys with linear depth in a splay
tree, but on average the cost of a search is logarithmic in the number
of keys.
• Using balanced search trees, we can maintain a dictionary with worst-
case logarithmic depth for all keys, and thus worst-case logarithmic
cost per insert/delete/search operation.
• Is there a way to build a dictionary with less than logarithmic cost
and with keys that cannot be ordered?

8.2 Hashing

Definition 8.4 (Universe, Key Set, Hash Table, Buckets). We consider a uni-
verse U containing all possible keys. We want to maintain a subset of this
universe, the key set N ⊆ U with |N| = n, where |N| ≪ |U|. We will use a
hash table M, i.e. an array M with m buckets M[0], M[1], ..., M[m − 1].

Remarks:

• The standard library of almost every widely used programming lan-
guage provides hash tables, sometimes by another name. In C++,
they are called unordered_map, in Python dictionary, in Java HashMap.
• The translation from virtual memory to physical memory uses a piece
of hardware called translation lookaside buffer (TLB), which is a hard-
ware implementation of a hash table. It has a fixed size and acts like
a cache for frequently looked up virtual addresses.
• Compilers make use of hash tables to manage the symbol table.

Definition 8.5 (Hash Function). Given a universe U and a hash table M, a
hash function is a function h : U → M. Given some key k ∈ U, we call h(k)
the hash of k.

Algorithm 8.3 Search Tree: Search
Input: key k, root r of search tree
Output: k if it is in the tree, else ⊥.
1: If r contains key k, return k
2: If k is smaller than the key of r, set r to left child and recurse
3: If k is larger than the key of r, set r to right child and recurse
4: Return ⊥.
8.2. HASHING

Remarks:

- A hash function should be efficiently computable, e.g. \( h(k) = k \mod m \) for a key \( k \in \mathbb{N} \).
- If we use ISBN \( \mod m \) as our library hash function, can we insert/delete/search books in constant time?!!
- What if two keys \( k \neq l \) have \( h(k) = h(l) \)?

**Definition 8.6** (Collision). Given a hash function \( h : U \to M \), two distinct keys \( k, l \in U \) produce a collision if \( h(k) = h(l) \).

Remarks:

- There are competing objectives we want to optimize for when hashing. On the one hand, we want to make the hash table small since we want to save memory. On the other hand, small tables will have more collisions. How likely is it to get a collision for a given \( n \) and \( m \)?

**Theorem 8.7** (Birthday Problem). If we throw a fair \( m \)-sided dice \( n \leq m \) times, let \( D \) be the event that all throws show different numbers. Then \( D \) satisfies

\[
\Pr[D] \leq \exp \left(-\frac{n(n-1)}{2m}\right)
\]

Proof. We have that

\[
\Pr[D] = \frac{m}{m} \cdot \frac{m-1}{m} \cdot \ldots \cdot \frac{m-(n-1)}{m} = \prod_{i=0}^{n-1} \frac{m-i}{m}
\]

\[
= \prod_{i=0}^{n-1} \left(1 - \frac{i}{m}\right) = \exp \left(\sum_{i=0}^{n-1} \ln \left(1 - \frac{i}{m}\right)\right)
\]

We can use that \( \ln(1+x) \leq x \) for all \( x > -1 \) and the monotonicity of \( e^{x} \):

\[
\Pr[D] = \exp \left(\sum_{i=0}^{n-1} \ln \left(1 - \frac{i}{m}\right)\right) \leq \exp \left(\sum_{i=0}^{n-1} \frac{i}{m}\right) = \exp \left(-\frac{n(n-1)}{2m}\right)
\]

- Theorem 8.7 assumes totally random hash functions — for non-random distributions of hashes, we might have more collisions. In particular, if we fix a hash function, then we can always end up with a hash set \( N \) that suffers from many collisions. E.g., if many books have an ISBN that ends in 000, then ISBN mod 1,000 is a terrible hash function.
- Maybe we can use modulo, but with a different \( m \)?
- However, for any hash function there are bad key sets.
- On the other hand, for every key set there are good hash functions! How do we efficiently pick a good hash function, i.e. one that is likely to distribute hashes well?

**Definition 8.8** (Universal Family). Let \( \mathcal{H} \subseteq \{ h : U \to M \} \) be a family of hash functions from \( U \) to \( M \). If for all pairs of distinct keys \( k \neq l \in U \), the probability of a collision is \( \Pr[h(k) = h(l)] \leq \frac{1}{m} \) when we choose \( h \in \mathcal{H} \) uniformly, then \( \mathcal{H} \) is called a universal family (of hash functions).

Remarks:

- In other words: if we choose a hash function from a universal family, how likely is it to get a collision is \( \approx \frac{m}{n^2} \).
- We cannot just pick a random function from \( U \) to \( M \) because there are \( |M|^{|U|} \) many, so we need \( |U| \log |M| \) bits to encode such a random function. That is even more bits than keys in our huge universe \( U \).

**Lemma 8.9.** For prime \( m \), \( a \in \{0, \ldots, m-1\} \), \( \delta \in \{1, \ldots, b-1\} \), and \( b \leq m \), the linear function \( f_{\delta}(a) := a \cdot \delta \mod m \) is a bijection.

Proof. Suppose there are integers \( a, a' \) with \( 0 \leq a < a' < m \) such that \( f_{\delta}(a) = f_{\delta}(a') \mod m \). Then, there exists an integer \( k \) such that \( a' - a = b(n - a) = \delta \cdot \delta \cdot \delta \mod m \), which implies that \( m \mid \delta \cdot (a' - a) \). Since \( \delta \) is relative prime to \( m \), we have \( m \mid (a' - a) \), thus \( a' = a \cdot \delta \mod m \). \( \Box \)

**Theorem 8.10** (Universal Hashing). Let \( m \) be prime and \( s \in \mathbb{N} \). Let \( U = \{0, \ldots, b-1\}^{s} \) and let \( M = \{0, \ldots, m-1\} \) with \( b \leq m \). For a key \( k = (k_{0}, \ldots, k_{s-1}) \in U \) and coefficient tuple \( a = (a_{0}, \ldots, a_{s-1}) \in \{0, \ldots, m-1\}^{s} \), define

\[
h_{a}(k_{0}, \ldots, k_{s-1}) := \sum_{i=0}^{s-1} a_{i} \cdot k_{i} \mod m.
\]

Then \( \mathcal{H} := \{ h_{a} : a \in \{0, \ldots, m-1\}^{s} \} \) is a universal family of hash functions.

Proof. Let \( (k_{0}, \ldots, k_{s-1}) = k \neq k' = (k'_{0}, \ldots, k'_{s-1}) \in U \). Using \( k_{i} = k'_{i} - k_{i} \), we get

\[
h_{a}(k') - h_{a}(k) = \sum_{i=0}^{s-1} a_{i} \cdot k'_{i} - \sum_{i=0}^{s-1} a_{i} \cdot k_{i} = \sum_{i=0}^{s-1} a_{i} \cdot \delta_{i} \mod m.
\]

The terms with \( \delta_{i} = 0 \) are 0, and we can ignore them. Let \( X := \{i \in \{0, \ldots, m-1\} : \delta_{i} \neq 0\} \) be the non-empty set of the indices of the non-zero terms.
with $x = |X|$. There are $m^x$ possibilities to choose the coefficients $a_i : i \in X$.

How many of these possibilities will have a conflict, i.e., a sum of 0? If we choose $x = 1$ coefficients arbitrarily, then the last coefficient can always render the sum 0, as according to Lemma 8.9 the last coefficient can add any value $(0, \ldots, m - 1)$ to the sum. In other words, $m^{x-1}$ sums are 0. Therefore, our chance of randomly picking an $x$ that produces a collision is $\frac{m^x-m}{m^x} = \frac{1}{m}$.

**Remarks:**

- Theorem 8.10 gives us a general method for picking hash functions from a universal family in an efficient manner. We simply choose a prime number $m$ and uniformly at random some factors $a_0, \ldots, a_k$.
- Thus, we can represent our hash function as the tuple $(m, a_0, \ldots, a_k)$.
- In practice, hash tables perform really well, and if we detect that we need to rebuild our table with a new function — this is called rehashing.

### 8.3 Static Hashing

How can we state the tradeoff more precisely?

**Definition 8.11** (Number of Collisions). Given a hash function $h : U \to M$ and a key set $N \subseteq U$, define the number of collisions that $h$ produces on $N$ as

$$C(h, N) := \{(k, l) \subseteq N : k \neq l, h(k) = h(l)\}.$$  

**Lemma 8.12** (Space vs. Collisions). Let $b$ be an upper bound on the number of collisions we want a hash function $h_b$ to produce on a given key set $N$ of size $|N| = n$. If we sample from a universal family, we can find an $h_b$, that satisfies $C(h_b, N) \leq b$ and $m = \lceil \frac{n(n-1)}{2b} \rceil$ by sampling a constant number of times in expectation.

**Proof.** There are $\binom{n}{2}$ pairs of distinct keys in $N$, and each of those produces a collision with probability at most $1/m$ since $h$ is chosen from a universal family. Together, using the linearity of expectation we get

$$E[C(h_b, N)] \leq \frac{\binom{n}{2}}{m} = \frac{n(n-1)}{2m}.$$

The Markov inequality states that for any random variable $X$ that only takes on non-negative integer values, we have $Pr[X \geq k] \leq E[X] / k$. Hence,

$$Pr[C(h_b, N) \geq 2 \cdot E[C(h_b, N)]] \leq \frac{1}{2},$$

and so

$$Pr[C(h_b, N) < 2 \cdot E[C(h_b, N)]] \geq \frac{1}{2}.$$

If we choose $m$ such that $2 \cdot E[C(h_b, N)] \leq b$, then we only need to sample 2 hash functions in expectation. Solving for $m$, we get

$$2 \cdot E[C(h_b, N)] \leq b \iff \frac{n(n-1)}{m} \leq b \iff \frac{n(n-1)}{b} \leq m.$$

**Remarks:**

- According to Theorem 8.12, if we want no collisions, we set $b = 1$ and choose $m = \lceil \frac{n(n-1)}{2} \rceil = n(n-1)$.
- Similarly, if we can tolerate $n$ collisions, we find that a hash table of size $m = n - 1$ suffices.
- Algorithm 8.13 defines perfect static hashing, which applies the result of Theorem 8.12.

**Algorithm 8.13** Perfect Static Hashing

**Input:** fixed set of keys $N$

**Output:** Primary hash table $M$ and secondary hash tables $M_i$

**Function:** $N_i := \{k \in N : h(i) = i\}$

**Function:** $n_i := |N_i|

1. $M := \text{a table with } n \text{ buckets}$

2. **repeat**

3. $h := \text{hash function } N \to M$ (sampled from universal family)

4. **until** $C(h, N) < n$

5. **for** $i \in M$ **do**

6. $M_i := \text{a table with } n_i(n_i - 1) = 2^n$ buckets

7. **repeat**

8. $h_i := \text{hash function } N_i \to M_i$ (sampled from universal family)

9. **until** $C(h_i, N_i) < 1$

10. **end for**

11. **return** $(M, h_i)_{i \in \{0, \ldots, m-1\}}, (h_i)_{i \in \{0, \ldots, m-1\}}$

**Remarks:**

- In a first stage (Lines 1 to 4), we find a hash function $h$ with at most $n$ collisions in linear space according to Theorem 8.12.
- In a second stage (Lines 5 to 10), we find a hash function $h_i$ per bucket $i$ without collisions by using an amount of space that is quadratic in the number of keys in the bucket $n_i$ as per Theorem 8.12.

**Theorem 8.14** (Perfect Static Hashing). Algorithm 8.13 returns a collision-free data structure of with a total size $(M$ and all $M_i$) less than $3n$.

**Proof.** The data structure is collision-free because of Line 9. Bucket $i$ of size $n_i(n_i - 1) = 2^n$ holds $n_i$ keys that produced $\binom{2^n}{2}$ collisions. Since the total number of collisions is less than $n$ (Line 4), and each bucket is of size twice its collisions, the total size of all buckets is less than $2n$. Because of Line 1, the size of $M$ is $n$, so together the total size of the data structure is less than $n + 2n = 3n$.

**Remarks:**

- We now have a hashing algorithm that can be built in linear space and expected linear time, and offers worst-case constant time search for a static set $N$.
- But what about a dynamic dictionary?
8.4 Collisions

Definition 8.15 (Hashing with Chaining). In hashing with chaining, every bucket $M[j]$ stores a pointer to a secondary data structure that manages all keys $k$ with $h(k) = j$. Insertion, search, and deletion of $k$ are all relegated to those data structures. In the simplest implementation, we use a simple linked list for each bucket.

Remarks:
- Algorithm 8.13 is an instance of hashing with chaining with the $M_i$ being the secondary data structures managing the buckets.
- The Java standard library uses hashing with chaining to resolve collisions.

Definition 8.16 (Load Factor). The fraction $\alpha = n/m$ is called the load factor of the hash table.

Remarks:
- The performance of all three operations (insert/delete/search) depends on the load factor for all collision resolution strategies.
- Hashing with chaining allows for a load factor $\alpha > 1$ since the size of the table is the number of secondary data structures, performance deteriorates with growing $\alpha$.
- If we use linked lists as secondary structures and use a hash function chosen from a universal family, the cost for an unsuccessful search is $1 + \alpha$ in expectation.
- If we use one of the strategies of this section and $\alpha$ grows too large, we should rehash with a bigger $m$ in order to remain efficient. In the Java standard library, if a hash table surpasses a load factor of 0.75, it is rehashed into a hash table with twice the size of the old one.

Definition 8.17 (Hashing with Probing). In hashing with probing, keys are stored directly in the hash table. The sequence $(h_i(k) \mod m)_{i \geq 0}$ is called the probing sequence of $k$, and each step of the iteration is a probe.

Remarks:
- Algorithm 8.18 defines how to search for a key in hashing with probing. Line 5 is a successful search, and Lines 7 and 11 are the two cases of unsuccessful searches.
- To insert a key, we adapt Algorithm 8.18: with an unsuccessful search in Line 7 we insert in the empty bucket. Therefore, the cost of an insert is roughly the cost of an unsuccessful search. An unsuccessful search in Line 11 triggers a rehash.

Algorithm 8.18 Hashing with Probing: Search

Input: key $k$ to search for
Output: $i$ if found, else $\perp$

Function: parametrized hash function $h_i$

1: $i := 0$
2: while $i < m$ do
3: $j := h_i(k) \mod m$
4: if $M[j] = k$ then
5: return $M[j]$
6: else if $M[j] = \perp$ then
7: return $\perp$
8: end if
9: $i := i + 1$
10: end while
11: return $\perp$

Remarks:
- Table 8.19 describes three different types of hashing with probing, each together with the expected time that a successful or unsuccessful search takes in expectation. More generally, linear probing uses some linear function $h_i(k) = h(k) + ci$ for some $c \neq 0$, and quadratic probing uses some quadratic function $h_i(k) = h(k) + ci + di^2$ with $d \neq 0$. As long as we guarantee that $h_i(k)$ is integer for all $i \in [m]$, the constants $c$ and $d$ can be rational.

<table>
<thead>
<tr>
<th>Type</th>
<th>$h_i(k)$</th>
<th>$\approx$ cost successful</th>
<th>$\approx$ cost unsuccessful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear probing</td>
<td>$h(k) + i$</td>
<td>$\frac{1}{m} + \frac{1}{\sqrt{m}}$</td>
<td>$(1 + \frac{1}{\sqrt{m}})\alpha$</td>
</tr>
<tr>
<td>Quadratic probing</td>
<td>$h(k) + i^2$</td>
<td>$\frac{1}{m} + \frac{1}{\sqrt{m}}\alpha$</td>
<td>$1 + \frac{1}{\sqrt{m}}\alpha$</td>
</tr>
<tr>
<td>Double hashing</td>
<td>$h_1(k) + i \cdot h_2(k)$</td>
<td>$\frac{1}{m} + \frac{1}{\sqrt{m}}\alpha$</td>
<td>$\frac{1}{m} + \frac{1}{\sqrt{m}}\alpha$</td>
</tr>
</tbody>
</table>

Table 8.19: Different types of hashing with probing together with the expected number of probes per search, where $\alpha$ is the load factor of the table. For hashing with probing, we need $\alpha \leq 1$ since we must have $n \leq m$. Each of $h, h_1, h_2$ is a hash function drawn from a universal family.

- Table 8.19 describes three different types of hashing with probing, each together with the approximate time that a successful or unsuccessful search takes in expectation. More generally, linear probing uses some linear function $h_i(k) = h(k) + ci$ for some $c \neq 0$, and quadratic probing uses some quadratic function $h_i(k) = h(k) + ci + di^2$ with $d \neq 0$. As long as we guarantee that $h_i(k)$ is integer for all $i \in [m]$, the constants $c$ and $d$ can be rational.

- Table 8.19 describes three different types of hashing with probing, each together with the approximate time that a successful or unsuccessful search takes in expectation. More generally, linear probing uses some linear function $h_i(k) = h(k) + ci$ for some $c \neq 0$, and quadratic probing uses some quadratic function $h_i(k) = h(k) + ci + di^2$ with $d \neq 0$. As long as we guarantee that $h_i(k)$ is integer for all $i \in [m]$, the constants $c$ and $d$ can be rational.

- Table 8.19 describes three different types of hashing with probing, each together with the approximate time that a successful or unsuccessful search takes in expectation. More generally, linear probing uses some linear function $h_i(k) = h(k) + ci$ for some $c \neq 0$, and quadratic probing uses some quadratic function $h_i(k) = h(k) + ci + di^2$ with $d \neq 0$. As long as we guarantee that $h_i(k)$ is integer for all $i \in [m]$, the constants $c$ and $d$ can be rational.
whole table. Assume \( m = 7 \) buckets and \( h(k) = 0 \), then the probing sequence of \( k \) is \((0, 1, 4, 2, 2, 4, 1)\) — buckets 3, 5, 6 do not appear.

- Double hashing does not suffer from either version of clustering. One can show that if the hash functions \( h_1 \) and \( h_2 \) used in double hashing are independently drawn from a universal family, then double hashing performs as well as an idealized hash function that assigns hashes uniformly at random.

8.5 Worst Case Guarantees

So far, the cost of all operations for dynamic key sets has been given in expected time cost. There are algorithms that allow us to do better and give us worst case guarantees on some of the operations. Two widely known possibilities to achieve this are called dynamic perfect hashing and cuckoo hashing.

Algorithm 8.20 Cuckoo Hashing: Insert

1. Cuckoo hashing uses two hash tables, and as such provides two possible buckets \( M_i[h_1(k)] \) or \( M_i[h_2(k)] \) for each key \( k \).
2. When inserting a new key \( k \), and one of the two buckets is empty, we simply place \( k \) there.
3. If both buckets are occupied by other keys, the new key is anyway inserted in one of its two possible buckets, “kicking out” the key \( k' \) that currently resides in this bucket.
4. The kicked out key \( k' \) moves to its other bucket, potentially kicking out the currently resident key \( k'' \); this process is repeated recursively until an empty bucket is found.
5. If this recursion loops or takes too long (logarithmic in the table size), the hash table is rebuilt using two new hash functions.

Remarks:

- Search and delete only need to check two buckets to figure out whether a given key is in the table, and so those operations are worst case constant time.
- One can show that the expected insert cost in cuckoo hashing is constant as long as the load factor \( \alpha \) is below 0.5.
- Cuckoo hashing gets its name from cuckoo birds: they lay their eggs into the nests of other birds, and once the cuckoo chicks hatch, they push the other eggs/chicks out of the nest.
- The idea behind cuckoo hashing is to use the “power of two choices”, which can be roughly described as: if you can choose between two resources and use the one that is less busy, you gain efficiency.
- To adapt perfect static hashing to a dynamic setting where we can also handle inserts and deletions, all we have to do is choose the size of \( M_i \) twice as large as in Algorithm 8.13, and rehash appropriately: Whenever \( C(h_i, N_i) > 0 \) for some bucket \( i \), we rehash that bucket until there are no collisions. Once some bucket reaches \( n_2^2 = |M_i| \), due to insertions, we rehash the entire table. This leaves us with expected constant time insert and delete, and worst case constant time search.

To keep the table linear-sized, we rehash everything after every \( m \) updates (inserts or deletes).

Chapter Notes

Dictionaries based on search trees are useful for providing additional operations such as nearest neighbor queries or range queries, where we want to find all keys in a certain range. Binary search trees were first published by three independent groups in 1960 and 1962 (for references, see Knuth [9]). The first instance of a self-balancing search tree that guarantees logarithmic cost for insert/search/delete is the AVL-tree, named so after its inventors Adelson-Velski and Landis [1]. For multidimensional keys, e.g. geometric data or images, there are specialized tree structures such as kd-trees [2] or BK-trees [3].

Hashing has a long history and was initially used and validated based on empirical results. One of the first publications was Peterson’s 1957 article [11] where he defined an idealized version of probing and empirically analyzed linear probing. Universal hashing was introduced two decades later by Carter and Wegman in 1979 [4]. Perfect static hashing was invented in 1984 by Fredman et al. [7] and is sometimes also referred to as FKS hashing after its inventors. Its dynamization by Dietzfelbinger et al. took another decade until 1994 [6]. A comprehensive study on perfect hashing by Czech et al. was compiled in 1997 [5]. Cuckoo hashing is a comparatively recent algorithm; it was introduced by Pagh and Rodler in 2001 [10].

There have been a number of other developments regarding hashing since the late 1970s; for an overview, see Knuth [9], in particular the section on History at the end of chapter 6.4. For a neat visualization of hashing with probing, see [8] online.

The power of two choices paradigm has found widespread application and analysis in load balancing scenarios. It was initially studied from the perspective of a balls-into-bins game where we want to minimize the maximum number of balls in any bin, and to do this we can pick two random bins and put the next ball into the least full of the two bins. Richa et al. [12] compiled an excellent survey on the earliest sources and numerous applications of this paradigm.

This chapter was written in collaboration with Georg Bachmeier.

Bibliography


