

Lesson 11

Hashtables

Wholeness of the Lesson

Hashtables provide even faster access to elements in a collection than a BST, but at the price of losing the sorted status of the elements. If maintaining order among the elements of a collection is not required, hashtables are the most efficient data structure for storing elements in memory, when insertion, deletion, and lookup operations are needed. Hashtables give concrete expression to the ability of pure intelligence to know any one thing instantaneously; this ability, or quality of intelligence, is known as *Ritam Bhara Pragma*.

The Hashtable ADT

1. A Hashtable is a generalization of an array in which any object can be used as a key instead of just integers. A Hashtable has *keys* which are used to look up corresponding *values*. A typical example is to store records from a database in memory. A key field from the database is often used as a key in the hashtable, and the corresponding record is the value in the hashtable.
2. Two basic operations: (usually also have a `remove(Object key)` operation)
`void put(Object key, Object value);`
`Object get(Object key)`
3. Simple Example from the Labs (this is an elementary implementation of Hashtable ADT)

User's view

Char key	String value
'a'	"Adam"
'b'	"Bob"
'c'	"Charlie"
'w'	"William"

```
//insert into table
table.put('c', "Charlie");

//retrieve from table
table.get('c'); //returns "Charlie"
```

Implementation:

`put(c, s):`

- obtain array index: $c \rightarrow (\text{int})c \rightarrow i = (\text{int})c - 'a'$
- insert new `Entry(c, s)` into `table[i]`

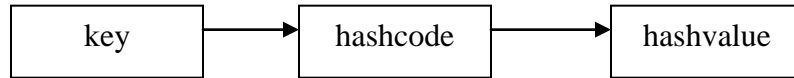
`get(c):`

- obtain array index: $c \rightarrow (\text{int})c \rightarrow i = (\text{int})c - 'a'$
- `Entry e = table[i]; return e.value`

Pattern:

KEY (non-number) \rightarrow HASHCODE (number) \rightarrow HASH VALUE (array index)

4. Basic steps in creating a hashtable data structure:
 - a. Devise a way of converting keys to integers so that different keys are mapped to different integers. This is what Java's `hashCode()` function is for – must be overridden by every class (since `hashCode()` is a method of the `Object` class).
 - b. Devise a way of converting hashcodes to smaller integers (called *hash values*) that will be the indices of a smaller array, called the *table*. To do this, you usually need to decide on the `tableSize`, which is the length of the array. A typical way (note: Java does it differently – see below) of making hashvalue from hashcode is by the formula

$$\text{hashvalue} = \text{hashcode} \% \text{tableSize}$$


- c. When a `put` operation is invoked, the key `k` and value `val` are stored in an `Entry` instance. The `hashCode` of the key is computed, and the hashvalue `h` of the `hashCode` is computed. This hashvalue `h` is the index in the table (array) `table`. Idea is to then insert the `Entry e` into this slot in the array `table[h] = e`. In practice, however, since multiple values may be hashed to the same slot `h`, each table slot stores a `LinkedList` (rather than storing `Entry` objects), so `Entry` objects are inserted into the linked list located at position `h`:

```
table[h].add(e) .
```

- d. When a `get` operation is called, passing in a key `k`, the `hashCode` and hashvalue of the key are computed to determine the slot `h`. Idea is that the value of the `Entry` in `table[h]` is returned. In practice, however, since a `LinkedList` is stored in `table[h]` (rather than `Entry` objects), this list is searched until the `Entry` with key `k` is found, and then the value `val` of that `Entry` is returned.

- e. **Example:** Java provides a `HashMap` class

```
HashMap h = new HashMap();  
h.put("14156", record1);  
h.put("21415", record2);  
h.get("14156"); //returns record1
```

Overriding the hashCode() Method

1. Any implementation of the Hashtable ADT in Java will make use of the `hashCode()` function as the first step in producing a hash value (or table index) for an object that is being used as a key.
2. Default implementation of `hashCode()` provided in the `Object` class is not generally useful.

Example: We wish to use pairs (firstName, lastName) as keys for `Person` objects in a hashtable. (See Demo)

Demo illustrates that default `hashCode` method is not useful. By default, it simply gives a numeric representation of the memory location of an object. If two `Pair` objects, created at different times, are equal (using the `equals` method), we would expect them to have the same `hashCodes`, so that, after hashing, they are sent to the same table slot. But default `hashCode` method does not take into account the fields used by `equals` method, so equal `Pair` objects may be assigned different slots in the table.

3. **Conclusion:** *Whenever `equals` is overridden in a class, `hashCode` must also be overridden.*

Creating Good Hash Codes When Overriding hashCode ()

There are two general rules for creating hash codes:

- I. (Primary Hashing Rule) Equal keys must be given the same hash code (otherwise, the same key will occupy different slots in the table)

If $k1.equals(k2)$ then $k1.hashCode() == k2.hashCode()$

- II. (Secondary Hashing Guideline) Different keys should be given different hash codes (potential danger: in the worst case, if every key is given the same hash code, then all keys are sent to the same slot in the table; in a more common case, unexpected regularities in the keys can result in poor distribution of keys in the table).

When this is not feasible, the hash codes should at least be distributed as evenly as possible (this means that one integer occurs as a hash code approximately just as frequently as any other)

Rule I is required; Rule II is a strong recommendation, though in some cases it may not be feasible.

Note: If Rule II is completely ignored, it is possible that all hashCodes for a collection of keys are identical. In that case, all keys are hashed to the same table slot. Since only one table slot is used, the hashtable is no more efficient than a linked list.

Example. *How Java overrides hashCode in the String class:* Any Java String is converted to an integer via hashCode () by this formula:

Given a String s of length k+1

$$\begin{aligned} s.hashCode() \text{ equals } & 31^k * s.charAt(0) + \\ & 31^{k-1} * s.charAt(1) + \\ & 31^{k-2} * s.charAt(2) + \dots + \\ & 31^0 * s.charAt(k) \end{aligned}$$

Since every character in the String is taken into account, equal Strings must have equal hashCodes. Because of the formula, it is highly unlikely that two distinct Strings will be assigned the same hashCode (though it's possible)

Example. *Overriding hashCode in the Person-Pair example.* We must take in account the same fields in computing hashCode as those used in overriding equals. The fields in Pair are Strings, and Java already provides hashCodes for Strings. So we make use of these and combine them to produce a complex hashCode for Pair.

```
public int hashCode() {  
    int result = 17; //seed  
    int hashFirst = first.hashCode();  
    int hashSecond = second.hashCode();  
    result += 31 * result + hashFirst;  
    result += 31 * result + hashSecond;  
    return result;  
}
```

Creating a Hash Value from Object Data

(From Effective Java, 2nd Ed.)

You are trying to define a hash value for each instance variable of a class. Suppose `f` is such an instance variable.

- If `f` is boolean, compute `(f ? 1 : 0)`
- If `f` is a byte, char, short, or int, compute `(int) f`.
- If `f` is a long, compute `(int) (f ^ (f >>> 32))`
- If `f` is a float, compute `Float.floatToIntBits(f)`
- If `f` is a double, compute `Double.doubleToLongBits(f)` which produces a long `f1`, then return `(int) (f1 ^ (f1 >>> 32))`
- If `f` is an object, compute `f.hashCode()`

Formula for creating your hashCode function

Step 1. Use the table above to produce a temporary hash of each variable in your class.

Example: You have variables `u`, `v`, `w`. Produce (using the chart above) temporary hash vals `hash_u`, `hash_v`, `hash_w`.

Step 2. Combine these temporary hashes into a final hashCode that is to be returned

Example:

```
int result = 17;
result += 31 * result + hash_u;
result += 31 * result + hash_v;
result += 31 * result + hash_w;
return result;
```


Sample Code: Creating Your Own Hashtable

```
public class MyHashtable {  
    private static final int INITIAL_SIZE = 20;  
    private int tableSize;  
    private LinkedList[] table;  
    public MyHashtable() {  
        this(INITIAL_SIZE);  
    }  
    public MyHashtable(int tableSize) {  
        this.tableSize = tableSize;  
        table = new LinkedList[tableSize];  
    }  
}
```

```

// FIRST TRY (needs to be fixed -- see SECOND_TRY BELOW)
public void put(Object key, Object value){
    //disallow null keys
    if(key==null) return;

    //get the "big" integer corresponding to the object
    //assumes key is not null
    int hashCode = key.hashCode();

    //compress down to a table slot
    int hash = hash(hashCode);

    //put the value and the key into an Entry object
    //which will be placed in the table in the
    //slot (namely, hash)
    //allows a null value
    Entry e = new Entry(key,value);

    // now place it in the table
    if(table[hash] == null){
        table[hash] = new LinkedList();
    }
    table[hash].add(e);
}

```

Big Problem: Suppose a client class attempts these put operations:

```

    put(key, "Bob")
    put(key, "Dave")

```

Suppose the hashvalue for key is 5. In the approach above, there will be two Entries placed in the list in slot 5 – [key, “Bob”] and [key, “Dave”]. Then there will be unpredictable results when a get (key) operation is performed .

```

// SECOND_TRY
public void put(Object key, Object value){
    //disallow null keys
    if(key==null) return;

    //get the "big" integer corresponding to the object
    //assumes key is not null
    int hashCode = key.hashCode();

    //compress down to a table slot
    int hash = hash(hashCode);

    //create the entry
    Entry e = new Entry(key,value);

    boolean keyAlreadyInUse = false;
    if(table[hash] != null) {
        for(Object ob : table[hash]) {
            Entry ent = (Entry)ob;
            if (ent.key.equals(key)) {
                keyAlreadyInUse = true;
                ent.value = value; //update value for this Entry
            }
        }
    }
    //we handled case keyAlreadyInUse==true in loop
    if(!keyAlreadyInUse) {
        // now place it in the table
        if(table[hash] == null){
            table[hash] = new LinkedList();
        }
        table[hash].add(e);
    }
}

```

```

public Object get(Object key){
    //null key not allowed
    if(key==null) return null;
    //get the "big" integer corresponding to the object
    int hashCode = key.hashCode();

    //compress down to a table slot
    int hash = hash(hashCode);

    //if slot given by hash not yet in use, return null
    if(table[hash] == null) return null;

    //now look for the desired Entry
    Entry e = null;
    for(Iterator it = table[hash].iterator(); it.hasNext()){
        e = (Entry)it.next();
        if(e.key.equals(key)) {
            return e.value;
        }
    }
    return null;
}

```

```
public String toString(){
    String n = System.getProperty("line.separator");
    StringBuilder sb = new StringBuilder();
    for(int i = 0; i < table.length;++i){
        if(table[i] != null){
            Entry next = null;
            for(Iterator it = table[i].iterator; it.hasNext()){
                next = (Entry)it.next();
                sb.append(next + n);
            }
        }
    }
    return sb.toString();
}

private int hash(int bigNum) {
    return bigNum % tableSize;
}
```

```
private class Entry{
    private Object key;
    private Object value;
    Entry(Object key, Object value){
        this.key = key;
        this.value = value;
    }
    public String toString(){
        return key.toString()+"->"+value.toString();
    }
}
}
```

Java's Implementation of Hashtables

1. Pre-j2se5.0: HashMap and Hashtable (HashMap is preferred; Hashtable is "legacy")

Example:

```
HashMap map = new HashMap();
map.put("Bob", new Employee("Bob", 40000, 1996, 10, 2));
Employee emp = (Employee)map.get("Bob");
```

Feature	java.util.HashMap	java.util.Hashtable	MyHashtable
Allows null key	Yes	No	No
Allows null values	Yes	No	Yes
Allows duplicate keys	No	No	No (after correction)
Synchronized (for safe multithreading)	No	Yes	No

2. j2se5.0 version is parametrized:

```
HashMap<String, Employee> map =
    new HashMap<String,Employee>();
map.put("Bob", new Employee("Bob", 40000, 1996, 10, 2));
Employee emp = map.get("Bob"); //no downcasting required
```

MAIN POINT

The most common implementation of hashtables uses separate chaining; each hashvalue is an index in an array of Lists; all objects with colliding hashvalue i are stored in the i th list in the array. The solution to the problem of avoiding hashvalue collisions (namely, by using a List in each of the array slots) illustrates the principle of the second element: Using a List in each array slot provides a way to "harmonize" objects that were apparently in conflict because of identical hashvalues.

Hashtable Application #1: Removing Duplicates

1. The most common use of hashtables is as in-memory look-up tables. For example, Employee records from a database could be stored by using Employee ID as key and the entire Employee record as value.
2. Another application of hashtables is for “bookkeeping” purposes. A simple example is an efficient procedure for removing duplicates from a list. The “naïve” way to remove duplicates is to use nested loops: For each element e in the list, use an inner loop to look at all elements preceding e in the list to see if e has occurred before; if so, remove this second occurrence of e .
3. A more efficient approach is to do the following: Create an auxiliary hashtable H . For each e in the list, check to see if e is a key in H . If so, remove e from the list. If not, add the entry $\langle e, e \rangle$ to H .
4. For a list having 1000 elements, the second procedure requires roughly 2000 steps of execution, whereas the first procedure requires on the order of 1,000,000 steps.

Hashtable Application #2: The Set ADT

1. Mathematically, a *set* is (roughly) a collection of objects. Two sets are said to be equal if they have the same elements.

For example:

$$\{1, 1, 3\} = \{1, 3\} = \{3, 1\}$$

because all have the same elements.

A set does not impose an ordering of elements (the set may contain elements that have their own natural order, like integers, but the set itself does not impose an order)

2. We can represent the mathematical notion of a set as an ADT called `Set`. In order to faithfully represent the properties of the mathematical idea, the `Set` ADT must have the following characteristics:
 - A. It does not allow duplicate elements
 - B. Iterating through a set does not guarantee any special order on its elements (in particular, there is no guarantee that the order in which an `Iterator` will provide elements corresponds to the order in which the elements were added in the first place)
 - C. Its overridden `equals()` method declares two `Sets` to be equal if and only if they have the same elements
3. Java provides a `Set` interface (and in j2se5.0, the parametrized version `Set<E>`) as part of the `Collections` API. It also provides an implementation of `Set` in the class `HashSet`, which observes the rules described above and is based on the `HashMap`. Using a `HashMap` naturally prevents duplicate elements from being added in the `Set`. `HashSet` is far more limited than it should be for a serious user of `Sets`. We will implement our own.

The MySet Interface

The following is a Java interface:

```
public interface MySet extends Iterable {

    //adds an Object to the Set
    //do not allow null and prevent duplicates
    public boolean add(Object ob);

    //removes an Object
    public boolean remove (Object ob);

    //returns true if ob is an element
    public boolean contains(Object ob);

    //returns true if the set has no element
    public boolean isEmpty();

    //returns the number of elements in the set
    public int size();

    //returns true if every element of this set is also
    //an element of the argument passed in
    public boolean isSubset(MySet set);

    //returns true if every element of the set passed in is
    //also an element of this set
    public boolean isSuperset(MySet set);

    //overrides the Object equals method, using the
    //criterion:
    //    two sets are equal iff they have same elements
    public boolean equals(Object ob);

    //returns an Iterator instance
    public Iterator iterator();

    //should be static
    //sets the isUnmodifiableFlag to true
    public void makeUnmodifiable(MySet set);

    //override hashCode
    public int hashCode();
}
```

MAIN POINT

The Hashtable ADT is a generalization of the concept of an array. It supports (nearly) random access of table elements by looking up with a (possibly) non-integer key. In the usual implementations, objects used as keys in a hashtable are “hashed”, producing hashcode (a numeric value) and hashvalue (numeric value reduced in size to be less than the table size). The hashvalue is an index in an array that can be used to locate or insert an object. Hashtables illustrate the principle of Do less and accomplish more – they provide an incredibly fast implementation of the main List operations.

Guidelines For Use of Common Data Structures

1. Array List

- Use When: Main need for a list is random access reads, relatively infrequent adds (beyond initial capacity) and/or number of list elements is known in advance. Sorting routines run faster on an ArrayList than on a Linked List.
- Avoid When: Many inserts and removes will be needed and/or when many adds expected, but number of elements unpredictable. Also: Maintaining data in sorted order is very inefficient.

2. Linked List

- Use When: Insertions and deletions are frequent, and/or many elements need to be added, but total number is unknown in advance. There is no faster data structure for repeatedly adding new elements than a Linked List (since elements are always added to the front).
- Avoid When: There is a need for repeated access to ith element as in binary search – random access is not supported.

3. Binary Search Tree

- Use When: Data needs to be maintained in sorted order. Faster than Linked Lists for insertions and deletions, but ordinary adds are slower. Provides very fast search for keys.
- Avoid When: The extra benefit of keeping data in sorted order is not needed and rapid read access is needed (Array List provides faster read access by index and hashtables provide faster read access by key)

4. Hashtable

- Use When: Random access to objects is needed but array indexing is not practical (recall example of Employee numbers – numbers in a very large range but relatively few Employees). Provides fastest possible insertion and deletion (faster than BST's).
- Avoid When: The order of data must be preserved (example: you want to find all employees whose salaries are in the range 60000..65000) or "find Max" or "find Min" operations are needed. Also, searching for values when keys are not known is slower than for other data structures because of hashtable overhead.

5. Set

- Use When: Objects need not be kept in a special order, duplicates should be disallowed, and there is no need for rapid lookup of individual set elements. Example: `keySet ()` in `HashMap` returns a `Set`.

CONNECTING THE PARTS OF KNOWLEDGE WITH THE WHOLENESS OF KNOWLEDGE

*Random access expanded from integer index to arbitrary index,
from "point" to "infinity"*

1. Arrays and ArrayLists provide highly efficient index-based access to a collection of elements.
2. The Hashtable ADT generalizes the behavior of an array by allowing non-integer keys (in fact, any object type can be used for a key), while retaining essentially random access efficiency for insertions, deletions, and lookups.

3. **Transcendental Consciousness:** TC is the home of all knowledge. The Upanishads declare "Know that by which all else is known" – this is the field of pure consciousness.

4. ***Wholeness moving within Itself:*** In Unity Consciousness, one sees that the "key" to accessing complete knowledge of any object is the infinite value of that object, pure consciousness, which is known in this state to be one's own Self. Knowing that level of the object, it then becomes possible to know any more relative level of the object as well.

