

6. Hash Tables

6.1. Motivation

Das *Collection* Interface lässt bewusst viele Freiheiten für verschiedene Implementierungen. Diese unterscheiden sich durch unterschiedliche Eigenschaften über das für alle *Collections* Verbindliche hinaus. So erlauben *Lists* Elemente gezielt zu positionieren und damit die Reihenfolge zu kontrollieren. *Queues* und *Stacks* wiederum erlauben rasches Auffinden von Elementen in Abhängigkeit der Reihenfolge, in der sie in die Datenstruktur eingefügt wurden. *Trees* und *Priority Queues* ordnen die Elemente selbständig nach einem Ordnungskriterium.

Keine dieser Datenstrukturen kann jedoch die *contains*-Operation im Durchschnitt mit Aufwand $O(1)$ ausführen. Wird keine der zuvor diskutierten Zusatzeigenschaften benötigt, sondern nur ein rasches Auffinden von Objekten anhand eines Schlüssels, sind *Hash-Tabellen* eine gute Lösung.

6.2. Lernziele

- Sie können die Grundidee und die Eigenschaften von *Hash Tables* im Vergleich mit anderen Datenstrukturen erklären und bewerten.
- Sie können *Hash-Funktionen* vorschlagen und ihre Eignung für konkrete Anforderungen beurteilen.
- Sie können Verfahren zur *Kollisionsbehandlung* vorschlagen und ihre Eignung für konkrete Anforderungen beurteilen.
- Sie können eine *Hash Table* effizient in Java implementieren.

6.3. Grundidee

Eine oft zu lösende Aufgabe ist eine Sammlung mehrerer Elemente zu unterhalten und darin einzelne Elemente anhand eines Schlüssels wiederzufinden (bzw. festzustellen, ob es überhaupt ein solches Element gibt). Beispielsweise möchte man zu Personen das Geburtsdatum speichern und anhand des Namens finden können.

Datenstrukturen wie geordnete Arrays und Bäume unterstützen das Wiederfinden, ohne dass man *alle* Elemente dafür einzeln betrachten muss. Dabei wird die *logische Anordnung* der Elemente in einer Sequenz oder einem Baum benutzt, um rasch die Menge der möglichen Speicherorte zu verkleinern. Das kann man sich vorstellen wie die Suche in einem alphabetisch geordneten Karteikasten.

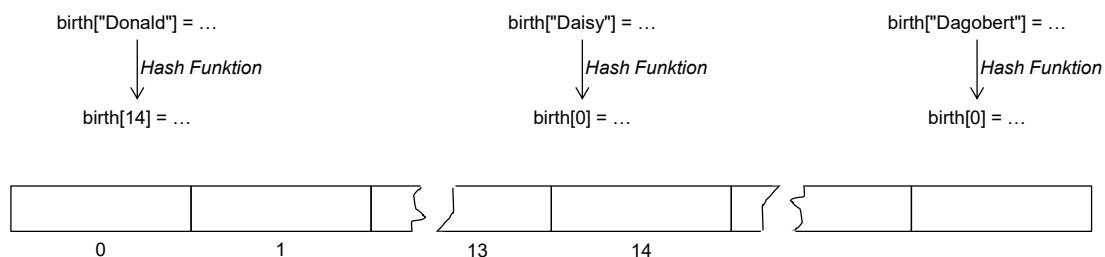
Nochmal deutlich schneller ist das Wiederfinden anhand eines Schlüssels, wenn der direkt auf den Speicherplatz verweist, an dem sich das gesuchte Objekt befindet. Das ist vergleichbar mit Gegenständen, die wohldefiniert in Schubladen eingeräumt sind, und bei denen man genau weiss, welche Schublade man aufziehen muss, um einen bestimmten Gegenstand zu finden.

In Programmiersprachen entsprechen *Arrays* eigentlich ganzen Schränken mit solchen Schubladen. Wenn man den Array-Index des gesuchten Objekts kennt, kann man es „mit einem Griff“ – also Aufwand $O(1)$ – finden, bzw. prüfen, ob es im Array enthalten ist.

Ideal wäre also, wenn die Programmiersprache z.B. zulassen würde zu schreiben:¹

```
birth["Donald"] = "9.Juni 1934";
if (birth["Donald"].equals(birth["Dagobert"])) ...
```

Objekte wie z.B. vom Typ *String* sind jedoch nicht geeignet, den Speicher eines Computers oder ein Array-Element direkt zu adressieren. Daher muss man das Objekt in eine Zahl transformieren, die dann als Index benutzt werden kann. Das leistet eine sogenannte *Hash Funktion*.



¹ Es gibt Programmiersprachen (z.B. *Perl* oder *Python*), die solche Schreibweisen ermöglichen. Tatsächlich handelt es sich dabei aber um *syntactic sugar* für Hash-Tabellen, also eine abkürzende Schreibweise für solche Datenstrukturen, um die es hier geht.

6.4. Datenstruktur Map

Das Interface *java.util.Map* des Java Collection Frameworks definiert eine abstrakte Datenstruktur, in der Schlüssel-Werte-Paare abgelegt werden können. Die Schlüssel können dabei beliebige Objekte sein, nicht nur Zahlen oder Namen. Java bietet damit den im Abschnitt 6.3 gezeigten Mechanismus an, wenn auch in der Library und nicht als Teil der Programmiersprache.

```
public interface Map<K,V> {
    V get(Object key);           // get value stored for key
    V put(K key, V value);       // store value with key as "index"
    V remove(Object key);        // remove value stored for key
    Set<K> keySet();              // get all keys currently stored in map (keys must be unique)
    Collection<V> values();      // get all values currently stored in map (this may be a bag)
    ...
}
```

Das *interface Map* kann auf verschiedene Arten implementiert werden, beispielsweise mit ausgeglichenen Bäumen (*class TreeMap*) oder mit dem oben beschriebenen *Hashing* (*class HashMap*).

Für *HashMaps* müssen die Schlüssel-Objekte zwei Bedingungen erfüllen:

1. Es muss feststellbar sein, ob zwei Objekte als *gleich* betrachtet werden sollen
2. Aus den für den Schlüssel relevanten Attributwerten muss ein *Hash-Wert* berechnet werden

Sowohl die *Gleichheit* zweier Objekte, als auch die *Hash-Werte* können nicht generell ermittelt werden, sondern hängen von Attributwerten und deren Bedeutung im Einzelfall ab. Java sieht deswegen in der Klasse *Object* zwei entsprechende Operationen vor, die dann für konkrete Objekte implementiert werden können.

```
boolean equals(Object obj) // obj is "equal to" this one
int hashCode();           // hash value of this object
```

Damit *HashMaps* richtig funktionieren, muss für zwei Objekte *a* und *b* grundsätzlich gelten:

```
a.equals(b)                a.hashCode() == b.hashCode()
```

6.5. Implementierung von Hash Tables

Das Beispiel im Abschnitt 6.3 zeigt die grundsätzliche Funktionsweise von solchen *Hash Tables*. Nach Berechnen eines Array-Index durch eine Hashfunktion können Elemente mit Aufwand $O(1)$ eingefügt und gefunden werden – jedenfalls so lange es keine Kollisionen gibt. Daraus ergeben sich zwei grundsätzliche Überlegungen zu *Hash Tables*:

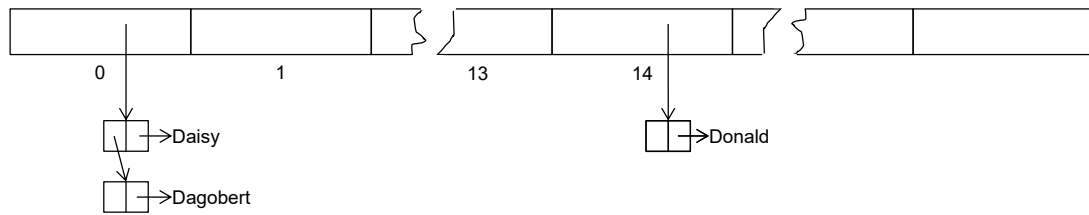
1. *Kollisionen*, wie beim Versuch, Daten mit dem Schlüssel „Dagobert“ einzufügen, können fast nie vollständig ausgeschlossen werden.² Oft hat man eine grosse oder gar prinzipiell unbegrenzte Menge von Schlüssel-Objekten (z.B. vom Typ *String*) und muss diese auf einen vergleichsweise kleinen Index-Bereich abbilden. Kollisionen – also gleiche Index-Werte für verschiedene Schlüssel – kann man dabei nicht systematisch ausschliessen. Man kann versuchen, Sie möglichst unwahrscheinlich zu machen, aber es braucht auch Strategien, um mit dann doch auftretenden Kollisionen umzugehen.
2. Als Anforderungen an *Hash Funktionen* ergeben sich folgende Punkte:
Sie müssen
 - a. schnell berechnet werden können und
 - b. Werte erzeugen, die möglichst gleichmässig auf dem gesamten Index-Bereich verteilt sind und
 - c. für Schlüssel, die logisch als gleich betrachtet werden (*equals*), denselben Wert erzeugen.

Im Rest dieses Kapitels werden verschiedene Verfahren zur *Kollisionsbehandlung* und dazu geeignete *Hash Funktionen* vorgestellt.

² Die Ausnahme bildet sogenanntes *perfektes Hashing*, bei dem man für eine im Voraus genau bekannte Menge von Elementen Hash-Werte so vergeben kann, dass es keine Kollisionen gibt. Diese Randbedingungen lassen sich in der Praxis selten erfüllen. Mögliche Anwendungen sind Speicher mit unveränderlichen Daten.

6.6. Eine einfache Strategie zum Umgang mit Kollisionen – Separate Chaining

Eine einfache Möglichkeit ist, mehrere verschiedene Objekte, die mit gleichem Array-Index abgelegt werden sollen, tatsächlich so zu speichern. Die Elemente des Arrays sind dann nicht Objekte sondern Listen von Objekten:



Diese Technik wird als *Separate Chaining* bezeichnet, weil die Elemente in (separaten) Ketten am Array hängen. Das Java Collection Framework benutzt Separate Chaining in der Klasse *HashMap*.³ Darin sind u.a. folgende Deklarationen enthalten:

```

public class HashMap<K,V> implements Map<K,V> {
    Node<K,V>[] table;
    ...

    static class Node<K,V> implements Map.Entry<K,V> {
        final K key;
        V value;
        Node<K,V> next;
        ...
    }
}
  
```

6.7. Wahrscheinlichkeit von Kollisionen

Bei rein zufälligen Verteilungen von *Hash-Werten* ist es sehr wahrscheinlich, dass einzelne Kollisionen auftreten, aber unwahrscheinlich, dass es viele gibt. Das sollen folgende Überlegungen illustrieren.

Gehen wir davon aus, dass das Array ca. doppelt so vielen Elementen Platz bietet, als darin enthalten sind. Es ist dann sehr wahrscheinlich, dass es einzelne Kollisionen gibt.⁴ Andererseits ist es sehr unwahrscheinlich, dass sehr viele Objekte am selben Ort landen und die Hash-Tabelle damit zu einer linearen Liste entartet. Diese beiden Aussagen sollen im Folgenden an Beispielen untersucht werden.

Die Wahrscheinlichkeit, dass es zu keinen Kollisionen kommt, wenn man in ein Array der Länge m insgesamt n Elemente einfügt, kann man mit folgender Formel berechnen:

Betrachten wir als Beispiel eine Tabelle mit 13 Plätzen, in der 7 Elemente gespeichert werden sollen. Die Wahrscheinlichkeit, dabei mindestens eine Kollision zu verursachen, ist 1 minus die Wahrscheinlichkeit für keine Kollision:

³ Seit Java 8 werden bei einer grossen Menge von Kollisionen anstelle von Listen Baumstrukturen benutzt.

⁴ Ein beliebtes Beispiel hierfür ist das sogenannte *Geburtsdays-Paradox*: Wenn 23 Personen in einem Raum sind – wie gross ist die Wahrscheinlichkeit, dass mindestens zwei davon am selben Tag eines Jahres Geburtstag haben? Übersetzt auf *Hash-Tabellen*: Wenn 23 Elemente in eine Tabelle mit 365 Plätzen eingefüllt werden – wie wahrscheinlich ist eine Kollision? Schätzen Sie erst, rechnen Sie dann.

Um Zahlen x in einer Hash-Tabelle der Länge m abzulegen, bietet es sich an, den Rest dieser Zahlen bei Division durch m – also $x \bmod m$ – als Index zu verwenden. Betrachtet wir beispielsweise 7 willkürlich ausgewählte Zahlen (man wähle z.B. 4-stellige beliebige Zahlenfolgen), die in eine Tabelle der Grösse 13 eingefüllt werden sollen:

Zahl x	$x \bmod 13$

Anzahl Kollisionen:

Max:

Durchschn. #Zugriffe:

0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	

6.8. Berechnung von *hashCode*

Die Ergebnisse der *hashCode*-Methode sollten möglichst zufällig über den ganzen Wertebereich des Typs *int* gestreut sein. Für Datentypen, deren Wertebereich ohnehin nicht grösser ist, als jener von *int*, verwendet man üblicherweise gerade den entsprechenden Wert selbst. Dies gilt für *boolean*, *byte*, *short*, *int*, *char* und *float*.

Für die 64-Bit Datentypen *long* und *double* ist eine *exclusive-oder*-Verknüpfung der beiden 32-Bit-Teile üblich, wie folgendes Beispiel aus der Klasse *Long* illustriert:

```
public int hashCode() {
    return (int)(value ^ (value >>> 32));
}
```

Um einen Hash-Wert über mehrere Variablen zu berechnen, die gemeinsam den Schlüssel-Wert ausmachen, könnte man einfach die Hash-Werte dieser einzelnen Werte addieren. Dabei erhält man jedoch für verschiedene systematische Variationen immer dieselbe Summe.

Als Beispiel betrachten wir den Datentyp *String*. *String*-Werte lassen sich als Folge von *char*-Werten auffassen. Ein trivialer *String*-Hash-Wert wäre also die Summe aller Zeichen-Codes (s. nebenstehende Tabelle).

Wort	Hash-Wert
AUS	65 + 85 + 83 = 233
USA	
SAU	
OHR	
WEM	
OFT	

Tatsächlich ist in der Klasse *String* ein etwas komplizierteres Verfahren implementiert:⁵

```
public final class String {
    /** The value is used for character storage. */
    private final char value[];

    /** Cache the hash code for the string */
    private int hash; // Default to 0

    ...

    public int hashCode() {
        int h = hash;
        if (h == 0 && value.length > 0) {
            char val[] = value;
            for (int i = 0; i < value.length; i++) {
                h = 31 * h + val[i];
            }
            hash = h;
        }
        return h;
    }
    ...
}
```

Damit ergeben sich besser verteilte Hash-Werte:

Wort	Hash-Wert
AUS	
USA	
SAU	

⁵ Das angegebene Programmfragment stammt aus Java 8. Seit Java 9 werden in der Klasse *String* intern Variablen vom Typ `byte[]` anstelle von `char[]` benutzt. Das Schema der Berechnung bleibt aber gleich.

6.9. Open Addressing

Separate Chaining zum Umgang mit Kollisionen ist ein einfaches Konzept (s. 6.6). Die Implementierung verursacht jedoch Aufwand z.B. in Form von zusätzlichem Speicher für die Listenverkettung. Ausserdem wird eine dynamische Speicherverwaltung vorausgesetzt, die es z.B. bei kleinen Geräten (*embedded Systems*, die z.B. mit *JavaCard* programmiert werden) nicht gibt.

Open Addressing vermeidet diesen Aufwand. Statt einer zusätzlichen Datenstruktur, werden noch freie Plätze in der Hash-Tabelle selbst benutzt. Kommt es zu einer Kollision, wird nach einer festen Regel ein neuer Index berechnet und die Suche dort fortgesetzt.

Es gibt verschiedene konkrete Umsetzungsmöglichkeiten von solchem *Open Addressing*. Zwei davon werden im Folgenden näher betrachtet: *Linear Probing* und *Double Hashing*.

6.9.1. Linear Probing ⁶

Eine einfache Regel zur Suche nach einem freien Platz ist, jeweils den nächsthöheren Index zu überprüfen. Erreicht man dabei die obere Grenze des Indexbereichs, setzt man die Suche bei 0 fort.

```
public void add(T elem) {
    int i = (elem.hashCode() & 0x7FFFFFFF) % size;
    while (array[i] != null) {
        i = (i + 1) % size;
    }
    array[i] = elem;
}
```

Beispiel

Fügt man Elemente mit den ersten Quadratzahlen als Hash-Wert in eine Tabelle der Länge 16 ein und zählt dabei jeweils, wie oft man Sondieren (weetersuchen) muss, ergibt sich folgendes Bild:

hashCode()	0	1	4	9	16	25	36	49	64	81
%16										
Kollisionen										

Hash-Tabelle:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Bei diesem Beispiel kann man sehr gut das sogenannte *Clustering* beobachten:

⁶ Auch: *lineares* oder *sequenzielles Sondieren*

6.9.2. Double Hashing

Um die beim *Linear Probing* beobachtete Cluster-Bildung zu vermeiden, muss der Mechanismus durchbrochen werden, der die Cluster systematisch wachsen lässt. Dazu müssen die beim Sondieren gemachten Schritte vergrößert werden. Die besten Ergebnisse erreicht man, wenn man die Schrittweite für verschiedene Elemente unterschiedlich wählt, d.h. in Abhängigkeit des Hash-Wertes:

```
public void add(T elem) {
    int i = (elem.hashCode() & 0x7FFFFFFF) % size;
    int step = ...?
    while (array[i] != null) {
        i = (i + step) % size;
    }
    array[i] = elem;
}
```

Die Regel zur Berechnung der Werte für `step` muss mit Sorgfalt gewählt werden. Zur Veranschaulichung, worauf es ankommt, hier einige Beispiele für `step`-Werte (die Tabellengröße sei 16):

step	Sondierungsfolge bei Start mit $i = 1$	Beobachtung
7	1,	
0	1,	
16	1,	
4	1,	
12	1,	
9	1,	

Alle Werte von `step` müssen zwei Bedingungen erfüllen:

Es gibt zwei bekannte Strategien, um dafür zu sorgen, dass diese Bedingungen immer erfüllt sind:

- Wähle für die Grösse der Tabelle N eine 2er-Potenz (2^m) und `step`
- Wähle für die Grösse der Tabelle N eine Primzahl und `step`

Eine beliebte Lösung ist, als Tabellen-Grösse eine Primzahl zu wählen und die Schrittweite zum Sondieren dann zu bestimmen als:

```
int step = 1 + (elem.hashCode() & 0x7FFFFFFF) % (size - 2);
```

Bemerkung:

Grundsätzlich kann man in obiger Berechnung auch $\% (size - 1)$ anstelle von $\% (size - 2)$ verwenden. Der Grund, warum man das oft nicht tut, ist folgender:

Wenn $size$ prim ist, ist es auch ungerade. Somit ist $size - 1$ gerade. Für jede gerade Zahl m gilt, dass $x \bmod m$ genau dann gerade ist, wenn x gerade ist. Hat man nun beispielsweise aus systematischen Gründen nur gerade Hash-Werte, so werden alle Schrittweiten ebenfalls gerade sein. In einem solchen Fall erhält man eine grössere Vielfalt an Schrittweiten, wenn m ungerade ist.

Auch bei Variante a) oben steht grundsätzlich nur die Hälfte der Zahlen (die ungeraden) für Schrittweiten zur Verfügung.

Beispiel:

Wir verwenden im Beispiel vereinfachend nur nicht-negative Hash-Werte. Obige Schrittweiten-Formel kann man dann für eine Tabellengrösse von 13 vereinfacht schreiben als:

```
int step = 1 + hashCode() % ;
```

Fügt man damit Elemente mit den ersten Quadratzahlen als Hash-Wert in eine Tabelle der Länge 13 ein und zählt dabei jeweils, wie oft man Sondieren (weitsuchen) muss, ergibt sich folgendes Bild:

hashCode()	0	1	4	9	16	25	36	49	64	81
%13										
step										
Kollisionen										

Hash-Tabelle:

0	1	2	3	4	5	6	7	8	9	10	11	12

6.9.3. Schnellere Implementierung

Sowohl beim *Linear Probing* als auch beim *Double Hashing* verursacht die *Rest-* oder *Modulo-Operation* (%) in der Sondierungs-Schleife den grössten Laufzeitaufwand, denn diese benötigt eine Division und dafür werden im Vergleich zu den sonstigen sehr einfachen Operationen viele Einzelschritte in der Hardware benötigt. Es lohnt sich also, nach Alternativen für die Anweisung `i = (i + step) % size;` zu suchen:

- Wenn `size = 2k` für ein beliebiges `k` – also `size` eine Zweierpotenz ist –, lässt sich die Modulo-Operation durch eine Bit-Maskierung in einem Zyklus berechnen:
`i = (i + step) & (size - 1);`
- Anstelle einer Modulo-Rechnung kann man den Überlauf auch durch einen Vergleich erkennen:
`i = i + step; if (i >= size) i -= size;`
- Im Prinzip ist ein Vergleich von `i` mit 0 schneller als einer mit einem beliebigen Zahlenwert `size`, denn letzterer wird eigentlich mit einer zusätzlichen Operation berechnet: `i - size >= 0`. Diese Überlegung legt als weitere Variante nahe, *rückwärts zu sondieren* und statt eines Überlaufs einen Unterlauf zu erkennen:
`i = i - step; if (i < 0) i += size;`

Die ersten beiden Implementierungs-Varianten berechnen dieselben Indexwerte, wie die ursprüngliche Variante mit %. Die damit erzeugten Hash-Tabellen sehen ebenfalls gleich aus.

Da die letzte Variante jedoch in der anderen Richtung sucht, werden die Elemente bei Kollisionen an anderen Plätzen abgelegt. Für die Daten aus dem Beispiel bei 6.9.2 ergibt sich damit diese Hash-Tabelle:

0	1	2	3	4	5	6	7	8	9	10	11	12

6.9.4. Load Factor ⁷

Es ist zu erwarten und auch in den bisher betrachteten Beispielen erkennbar, dass Kollisionen immer öfter auftreten, je voller eine Tabelle ist. Um in diesem Zusammenhang quantifizierbare Aussagen machen zu können, verwendet man den *Load Factor* λ als Messgrösse:

$$\lambda = \frac{\text{Anzahl Elemente in der Tabelle}}{\text{Tabellengrösse}}$$

Beispiel:

Für die beiden Beispiel-Tabellen in 6.9.1 und 6.9.2 gilt

$$\lambda_{6.9.1} = \quad \text{bzw.} \quad \lambda_{6.9.2} =$$

Für *Separate Chaining* gilt:

- Es gibt keine grundsätzliche Obergrenze für λ .
- Die durchschnittliche Länge der Listen ist gleich dem *Load Factor* λ .
- Um die Effizienz zu erhalten, empfiehlt sich: $\lambda < 1$.

Für *Open Addressing* gilt:

- Der Load Factor ist hier systembedingt begrenzt: $\lambda \leq 1$.
- Nur solange $\lambda < 1$ gilt, wird man bei jeder Sondier-Schleife irgendwann auf (allenfalls den) einen freien Platz treffen. Lässt man auch den Fall $\lambda = 1$ zu, muss man die Schleifenabbruchkriterien so formulieren, dass auch abgebrochen wird, wenn man alle Plätze einmal besucht hat.
- Um die Effizienz zu erhalten, empfiehlt sich bei *Linear Probing* $\lambda < 0.75$ und bei *Double Hashing* $\lambda < 0.9$.

Rehashing

Kennt man die Anzahl Elemente, die eingefügt werden sollen, von vornherein, kann man unter Berücksichtigung des angestrebten *Load Factors* ein entsprechend grosses Array erzeugen.

Kennt man die Anzahl der Elemente nicht, empfiehlt es sich, ein neues grösseres Array zu erzeugen und die Elemente dort hinein zu kopieren, sobald der *Load Factor* eine gegebene Grenze überschreitet. Das Verfahren ist von halbdynamischen Array-Listen her bekannt. Jedoch kann man bei *Hash Tables* nicht mehr einfach die Positionen im Array beibehalten, sondern muss die Indices mit dem Hash-Mechanismus neu berechnen. Diesen Prozess nennt man deswegen auch *Rehashing*.

6.9.5. Entfernen von Elementen

Entfernt man aus der im Kapitel 6.9.2 gefüllten Tabelle am Schluss das Element 9 wieder, erhält man diese Tabelle:

0	1	2	3	4	5	6	7	8	9	10	11	12
0	1	49	16	4		64		81		36		25

Sucht man in dieser Tabelle (mit der gleichen Sondierungs-Strategie wie in Kapitel 6.9.2) nach 49 oder 64 erhält man das Ergebnis

weil

Man darf also in solchen Tabellen keine Werte einfach entfernen. Stattdessen setzt man in solchen „frei gewordenen“ Plätzen eine spezielle Markierung (*Sentinel*) ein. Trifft man beim Sondieren nach einem Element (z.B. bei *contains(x)*) auf das Sentinel, dann

Soll ein neues Element eingefügt werden und man trifft beim Sondieren nach einem freien Platz auf das Sentinel, dann

⁷ Auch *Belegungsfaktor* oder *Belegungsgrad*

Implementierungstechnik für Sentinels

Benötigt man in einer objektorientierten Programmiersprache wie Java ein *Sentinel*, gibt es eine spezielle Implementierungstechnik: Man erzeugt ein fixes Objekt und setzt dieses (eigentlich eine Referenz darauf) als *Sentinel* ein. Dieses Objekt ist nie gleich irgendeinem tatsächlich in der Datenstruktur enthaltenen Objekt, aber auch nicht *null*. Per Referenzvergleich *o == sentinel* kann es klar identifiziert werden.

```
public class Table<T> {
    private final Object[] arr;
    private final Object sentinel = new Object();

    public Table(int size) {
        this.arr = new Object[size];
    }

    public void add(T elem) {
        ...
    }

    public boolean contains(T elem) {
        ...
    }

    public void remove(T elem) {
        assert elem != null;
        int i = (elem.hashCode() & 0x7FFFFFFF) % arr.length;
        int cnt = 0;
        while (arr[i] != null && !elem.equals(arr[i]) && cnt != arr.length) {
            i = (i + 1) % arr.length; // example uses linear probing for simplicity
            cnt++;
        }
        if (cnt != arr.length) arr[i] = sentinel;
    }
}
```