

Algorithmen und Datenstrukturen (ESE)
Entwurf, Analyse und Umsetzung von
Algorithmen (IEMS)
WS 2012 / 2013

Vorlesung 4, Dienstag 14. November 2013
(Mittlere Laufzeit, Assoziative Arrays aka Maps,
Hashverfahren)

Junior-Prof. Dr. Olaf Ronneberger
Image Analysis Lab
Institut für Informatik
Universität Freiburg

Blick über die Vorlesung heute

- Organisatorisches
 - Ihre Erfahrungen mit dem Ü3 (O-Notation)
- Laufzeiten
 - Beste, schlechteste und mittlere Laufzeit
- Assoziative Arrays aka Maps aka = also known as
 - Nutzen anhand eines typischen Beispielles
 - Alternative dazu: Sortieren
 - **Übungsaufgabe:** die weltweit häufigsten Ortsnamen finden ... einmal mit "Maps", und einmal mit Sortieren
- Hashverfahren

Erfahrungen mit dem Ü3 (O-Notation)

- Zusammenfassung / Auszüge Stand 14. November 09:00
 - 26 Abgaben nach Verlängerung
 - Für die meisten gut machbar (Kommentare hatten wir ja letzte Woche schon)

Feedback von Ihrem/r Tutor/in

- Machen Sie dazu svn update
 - ... spätestens bevor Sie ein neues Übungsblatt bearbeiten
 - Sie bekommen dann (hoffentlich) eine Datei `uebungsblatt_XY/feedback-tutor.txt`
 - Dort sehen Sie unter anderem
 - warum wo wie viel Punkte abgezogen
 - Verbesserungsvorschläge für Code / Beweise
 - Antworten auf Fragen aus Ihrer erfahrungen.txt
 - Und was Ihrem/r Tutor/in noch alles so einfällt ...

Verschiedene Laufzeiten



- Laufzeit hängt nicht immer ausschließlich von der Größe des Problems ab, sondern auch von der Beschaffenheit der Eingabemenge
- Daraus ergeben sich
 - **beste Laufzeit**
beste Laufzeitkomplexität für eine Eingabeinstanz der Größe n
 - **schlechteste Laufzeit**
schlechteste Laufzeitkomplexität für eine Instanz der Größe n
 - **mittlere oder erwartete Laufzeit**
gemittelte Laufzeitkomplexität für alle Eingabeinstanzen der Größe n

Beispiel 1 (if Abfrage)



- Eingabe: Feld a mit n Elementen $a[i] \in \mathbb{N} \quad 1 \leq a[i] \leq n$
- Ausgabe: Feld a mit n Elementen mit $a[0] \neq 0$

if (a[0]==0)	O(1)	O(1)	?
then		O(1)	
a[0]:=1;	O(1)		
else			
for i:=0 to n-1 do	O(n)	O(n)	
a[i]:=2;	O(1)		

- **beste Laufzeit:** $O(1) + O(1) = O(1)$
- **schlechteste Laufzeit:** $O(1) + O(n) = O(n)$

Mittlere Laufzeit



- Mittlere Laufzeit wird durch die Mittelung der notwendigen Schritte für **alle Eingabeinstanzen der Größe n** ermittelt.
- Jedes Element $a[i]$ kann n Werte annehmen.
- n Elemente: $n \cdot n \cdot \dots \cdot n = n^n$ verschiedene Eingabeinstanzen der Größe n
- $a[i]=0$ in n^{n-1} Instanzen
- $a[i] \neq 0$ in $n^n - n^{n-1} = n^{n-1} \cdot (n-1)$ Instanzen
- Gesamtschritte:
 n^{n-1} Instanzen \cdot 1 Schritt + $n^{n-1} \cdot (n-1)$ Instanzen \cdot n Schritte
 $= n^{n-1} + n^n \cdot (n-1)$
- **Mittlere Laufzeit:**
$$\frac{n^{n-1} + n^n \cdot (n-1)}{n^n} = \frac{1}{n} + n - 1 = \mathcal{O}(n)$$

(Schritte / Instanzen)

Beispiel 2 (Binär-Addition)



- Eingabe: n-stellige Dualzahl a
- Ausgabe: n-stellige Dualzahl a+1
- Zahl der Schritte des Algorithmus ergibt sich aus der Zahl der "0"- "1"- bzw. "1"- "0"-Wechsel

Problemgröße n	Eingabe	Operation	Ausgabe	Zahl der Schritte
10	1011100100	+1	1011100101	1
4	1011	+1	1100	3
8	11111111	+1	00000000	8 (=n)

- **beste Laufzeit:** 1 Schritt = $O(1)$
- **schlechteste Laufzeit:** n Schritte = $O(n)$

Mittlere Laufzeit



- ermittle die durchschnittliche Schrittzahl für alle Eingabeinstanzen der Größe n

n=1		
Eingabe	Ausgabe	Schritte
0	1	1
1	0	1

$$\text{Durchschnitt} \quad \frac{1 + 1}{2} = 1$$

n=2		
Eingabe	Ausgabe	Schritte
00	01	1
01	10	2
10	11	1
11	00	2

$$\text{Durchschnitt} \quad \frac{1 + 2 + 1 + 2}{4} = \frac{3}{2}$$

n=3		
Eingabe	Ausgabe	Schritte
000	001	1
001	010	2
010	011	1
011	100	3
100	101	1
101	110	2
110	111	1
111	000	3

$$\text{Durchschnitt} \quad \frac{7}{4}$$

$$\text{Mittlere Laufzeit} \quad 2 - \frac{1}{2^{n-1}} = \mathcal{O}(1)$$

Mittlere Laufzeit Fallunterscheidung



- Fallunterscheidung für Instanzen der Größe n

n-stellige Eingabe	Instanzen	Zahl der Schritte
xxx...x0	2^{n-1}	1
xxx...x01	2^{n-2}	2
xxx...x011	2^{n-3}	3
...
x0111...1	$2 = 2^1$	$n-1$
0111...1	$1 = 2^0$	n
111...1	1	n

- **Mittlere Laufzeit** (Schritte / Instanzen):

$$\frac{1 \cdot 2^{n-1} + 2 \cdot 2^{n-2} + \dots + (n-1) \cdot 2^1 + n \cdot 2^0 + n \cdot 1}{2^{n-1} + 2^{n-2} + \dots + 2^1 + 2^0 + 1} = \frac{\sum_{i=1}^n (i \cdot 2^{n-i}) + n}{\sum_{i=0}^{n-1} (2^i) + 1}$$

Zwischenrechnung



$$\frac{\sum_{i=1}^n (i \cdot 2^{n-i}) + n}{\sum_{i=0}^{n-1} (2^i) + 1} = ?$$

Nenner:

$$\sum_{i=0}^{n-1} (2^i) + 1 \quad \begin{array}{l} \text{geometrische} \\ \text{Reihe} \end{array} = 2^n - 1 + 1 = 2^n \quad \begin{array}{l} \text{intuitiv} \\ \text{klar} \end{array}$$

Zähler:

$$\sum_{i=1}^n (i \cdot 2^{n-i}) + n \quad \begin{array}{l} (a = 2a - a) \\ \end{array} = 2 \cdot \sum_{i=1}^n (i \cdot 2^{n-i}) - \sum_{i=1}^n (i \cdot 2^{n-i}) + n$$

$$= 1 \cdot 2^n + 2 \cdot 2^{n-1} + 3 \cdot 2^{n-2} + \dots + (n-1) \cdot 2^2 + n \cdot 2^1 \\ - 1 \cdot 2^{n-1} - 2 \cdot 2^{n-2} - \dots - (n-2) \cdot 2^2 - (n-1) \cdot 2^1 - n \cdot 2^0 + n$$

$$= 1 \cdot 2^n + 1 \cdot 2^{n-1} + 1 \cdot 2^{n-2} + \dots + 1 \cdot 2^2 + 1 \cdot 2^1 - n + n$$

$$= 2^n + 2^{n-1} + 2^{n-2} + \dots + 2^2 + 2^1 + 2^0 - 2^0 = 2^{n+1} - 1 - 2^0$$

Ergebnis



- Mittlere Laufzeit (Schritte / Instanzen)

$$\frac{2^{n+1} - 2}{2^n} = 2 - \frac{1}{2^{n-1}}$$

$$\lim_{n \rightarrow \infty} \left(2 - \frac{1}{2^{n-1}} \right) = 2 \quad \Rightarrow \mathcal{O}(1)$$

Normale vs. Assoziative Arrays

■ Normales Array

- Zugriff auf eine (große) Anzahl von gleichartigen Elementen über einen fortlaufenden **Index**

$A_0, A_1, A_2, A_3, A_4, A_5, \dots$

- Beispielanfrage: das Element mit Index 3 $A[3]$

■ Assoziatives Array

- Zugriff auf eine (große) Anzahl von gleichartigen Elementen über einen beliebigen sog. **Schlüssel** (Key)

$A_{\text{Europa}}, A_{\text{Amerika}}, A_{\text{Asien}}, A_{\text{Afrika}}, \dots$

- Beispielanfrage: das Element mit Schlüssel Amerika

- Die Schlüssel können auch (nicht fortlaufende) Zahlen sein $A["Amerika"]$

Und wofür braucht man das?

- In der Praxis ...
 - ... gibt es kaum ein größeres Programm wo man nicht irgendwo ein assoziatives Array braucht
- Unser Beispiel für diese Vorlesung
 - Eine Liste von Infos zu Ländern (von www.geonames.org)

countryInfo.txt

ISO	ISO3	ISO-Numeric	fips	Country	Capital	Area(in sq km)	Population	Continent	tld	CurrencyCode	CurrencyName
AD	AND	20	AN	Andorra	Andorra la Vella	468	84000	EU	.ad	EUR	Euro
AE	ARE	784	AE	United Arab Emirates	Abu Dhabi	82880	4975593	AS	.ae	AED	Dirham
AF	AFG	4	AF	Afghanistan	Kabul	647500	29121286	AS	.af	AFN	Afghani
AG	ATG	28	AC	Antigua and Barbuda	St. John's	443	86754	NA	.ag	XCD	Dollar
AI	AIA	660	AV	Anguilla	The Valley	102	13254	NA	.ai	XCD	Dollar
AL	ALB	8	AL	Albania	Tirana	28748	2986952	EU	.al	ALL	Lek
AM	ARM	51	AM	Armenia	Yerevan	29800	2968000	AS	.am	AMD	Dram
AO	AGO	24	AO	Angola	Luanda	1246700	13068161	AF	.ao	AOA	Kwanza
AQ	ATA	10	AY	Antarctica		14000000	0	AN	.aq		
AR	ARG	32	AR	Argentina	Buenos Aires	2766890	41343201	SA	.ar	ARS	Peso
AS	ASM	16	AQ	American Samoa	Pago Pago	199	57881	OC	.as	USD	Dollar
AT	AUT	40	AU	Austria	Vienna	83858	8205000	EU	.at	EUR	Euro
AU	AUS	36	AS	Australia	Canberra	7686850	21515754	OC	.au	AUD	Dollar
AW	ABW	533	AA	Aruba	Oranjestad	193	71566	NA	.aw	AWG	Guilder
AX	ALA	248		Aland Islands	Mariehamn		26711	EU	.ax	EUR	Euro
AZ	AZE	31	AJ	Azerbaijan	Baku	86600	8303512	AS	.az	AZN	Manat
BA	BIH	70	BK	Bosnia and Herzegovina	Sarajevo	51129	4590000	EU	.ba	BAM	Marka
BB	BRB	52	BB	Barbados	Bridgetown	431	285653	NA	.bb	BBD	Dollar
BD	BGD	50	BG	Bangladesh	Dhaka	144000	1,56E+08	AS	.bd	BDT	Taka
BE	BEL	56	BE	Belgium	Brussels	30510	10403000	EU	.be	EUR	Euro
BF	BFA	854	UV	Burkina Faso	Ouagadougou	274200	16241811	AF	.bf	XOF	Franc
BG	BGR	100	BU	Bulgaria	Sofia	110910	7148785	EU	.bg	BGN	Lev
BH	BHR	48	BA	Bahrain	Manama	665	738004	AS	.bh	BHD	Dinar
BI	BDI	108	BY	Burundi	Bujumbura	27830	9863117	AF	.bi	BIF	Franc
BJ	BEN	204	BN	Benin	Porto-Novo	112620	9056010	AF	.bj	XOF	Franc
BL	BLM	652	TB	Saint Barthelemy	Gustavia	21	8450	NA	.gp	EUR	Euro

...

Und wofür braucht man das?

- Uns interessieren die Spalten 5 (Land) und 9 (Kontinent)
- Welche Kontinente haben wie viele Länder?
- Wie würden Sie das lösen?
- Wir schauen uns zwei typische Lösungen dafür an
 - Mit Sortieren
 - Mit einem assoziativen Array aka Map

Lösung 1: Sortieren

■ Idee

- Wir sortieren die Zeilen (Spalte 5 und 9) nach Kontinent
- Dann stehen alle Länder im selben Kontinent in einem Block hintereinander
- Die Größe von jedem Block können wir dann zählen
- Und dann den größten Block finden

■ Nachteil

- Sortieren braucht Zeit $\Theta(n \log n)$ und geht nicht in $O(n)$
- Wir müssen **zweimal** über die ganzen Daten laufen

■ Vorteil

- Algorithmisch einfach
- Kommandozeile: geht mit einfachen Unix/Linux-Befehlen

Einfache Lösung mit Unix-Befehlen

- Daten liegen als reine Text-Datei vor
 - Spalten getrennt mit Tabulator
 - Kommentarzeilen beginnen mit Doppelkreuz (#)

- **grep** wählt bestimmte Zeilen aus

```
grep -v '^#' countryInfo.txt
```

↑ ↑
„nicht“ „Zeilenanfang“

- **cut** wählt bestimmte Spalten aus (Tab separiert)

```
cut -f5,9
```

↑
„Spalte 5 und 9“

Einfache Lösung mit Unix-Befehlen

- **sort** sortiert Zeilen

```
sort -t ` ` -k2,2
```

„Spaltentrenner“ „Tab“ (einfügen mit CTRL-v TAB)

Sortierschlüssel ist Spalte 2 bis 2

- **uniq** findet oder zählt eindeutige Schlüssel (Zeilen)

```
uniq -c
```

„zählen“

- **less** zeigt Datei Bildschirmweise an

- **head** gibt n Zeilen am Anfang der Datei aus

```
head -n30
```

Liste nach Kontinenten sortieren

```
grep -v '^#' countryInfo.txt | cut -f5,9 | sort -t' -k2,2 | less
```

```
Algeria    AF
Angola     AF
Benin      AF
Botswana   AF
Burkina Faso      AF
Burundi    AF
Cameroon   AF
Cape VerdeAF
Central African Republic      AF
Chad        AF
Comoros     AF
Democratic Republic of the Congo AF
Djibouti    AF
```

Länder pro Kontinent zählen

```
grep -v '^#' countryInfo.txt | cut -f9 | sort | uniq -c | sort -nr
```

58 AF

54 EU

52 AS

42 NA

27 OC

14 SA

5 AN

Lösung 2: Assoziatives Array / Map

■ Idee

- Ein assoziatives Array mit dem Kontinent als Index
- Der Wert ist entweder einfach ein Zähler, oder die Liste aller Länder in dem Kontinent

■ Vorteil

- Laufzeit $O(n)$
- Vorausgesetzt wir können in Zeit $O(1)$ auf ein Element des Arrays zugreifen, wie bei einem normalen Array
- Dass das geht, zeigen wir in der nächsten Vorlesung
- Aber erstmal sollen Sie verstehen, wozu und wie man assoziative Arrays benutzt !

Map in Java und C++ 1/3

- In **Java** und **C++** heißen die assoziativen Arrays **map**
 - Der Index heißt dort **key**, das Element heißt **value**
 - Eine **map** unterstützt u.a. die folgenden Operationen
 - **Einfügen** von Element **value** mit Schlüssel **key**
 - in Java `put(key, value)` ... in C++ `insert(key, value)`
 - **Zugriff** auf das Element mit Schlüssel **key**
 - in Java `get(key)` ... in C++ `operator[](key)` *A["abc"]*
 - **Löschen** des Elementes mit Schlüssel **key**
 - in Java `remove(key)` ... in C++ `erase(key)`
 - **Fragen** ob Element mit Schlüssel **key** da ist
 - in Java `containsKey(key)` ... in C++ `count(key)`

■ Effizienz

- Hängt von der Implementierung ab, es gibt insbesondere
 - In Java: `java.util.HashMap` und `java.util.TreeMap`
 - In C++: `__gnu_cxx::hash_map` und `std::map`
 - in C++11 ist die hash map `std::unordered_map`
- Was das genau ist, sehen wir in der nächsten Vorlesung
 - da bauen wir uns unsere eigene **map**
 - und analysieren sie dann

- Vorsicht bei folgendem Feature der map in C++
 - Nehmen wir an, wir haben eine `map` mit Schlüsseln vom Typ `std::string` und Elementen vom Typ `int`
`std::map<std::string, int> M;`
 - Dann kann man auf Elemente einfach mit dem `[]` Operator zugreifen wie bei einem normalen Array
`M["europe"] = 54;`
`M["asia"] = 52;`
 - Aber Vorsicht, wenn das Element vorher nicht in der `map` war, wird es durch `[]` angelegt, mit dem Defaultwert für den Typ von `value`, für `int` ist das `0`.
`if (M["venus"] > 0) ... // Inserts "venus" with value 0.`
`if (M.count("venus") > 0) ... // Only asks if "venus" is there.`

Übungsblatt 4

- Ermitteln Sie die häufigsten Städtenamen der Welt
- Quelle „allCountries.txt“ von geonames.org (ca. 8 Millionen Einträge)
- per Sortieren und per Hashmap

Literatur / Links

■ Assoziative Arrays

- In Mehlhorn/Sanders:

4 [Hash Tables and Associative Arrays](#) (das führt schon weiter)

- In Cormen/Leiserson/Rivest

12 [Hash Tables](#) (das ebenso)

- In Wikipedia

http://de.wikipedia.org/wiki/Assoziatives_Array

http://en.wikipedia.org/wiki/Associative_array

■ Map in Java und in C++

<http://download.oracle.com/javase/1.4.2/docs/api/java/util/Map.html>

- und ...HashMap bzw. ...TreeMap

<http://www.cplusplus.com/reference/stl/map/>