Informatik II: Algorithmen und Datenstrukturen SS 2015

Vorlesung 5a, Dienstag, 19. Mai 2015
(Universelles Hashing)

Prof. Dr. Hannah Bast
Lehrstuhl für Algorithmen und Datenstrukturen
Institut für Informatik
Universität Freiburg

UNI FREIBURG

Blick über die Vorlesung heute

UNI FREIBURG

- Organisatorisches
  - Ihre Erfahrungen mit dem Ü4 (HashMap, AOL Query Log)
  - Große Dateien: SVN und Unit Test
  - Nächste Woche ist Pfingsten = keine Vorlesungen
- Universelles Hashing
  - Motivation, Definition, zentraler Satz
  - Crash-Kurs Wahrscheinlichkeitsrechnung

Erfahrungen mit dem Ü4 1/4

UNI FREIBURG

- Zusammenfassung / Auszüge Stand 19. Mai 12:00
  - Für die meisten gut machbar, insbesondere zeitlich
  - Aufgabe 2 hat vielen gefallen … freut uns, so war's gedacht
  - Bei Aufgabe 1 (HashMap) haben einige erst intern die sprach-eigene HashMap / Dictionary benutzt … nicht Sinn der Sache
  - Mehr auf Leute ohne Programmiererfahrung eingehen

    Was wünschen Sie sich da konkret ?
  - Musterlösungen wären gut … gibt es schon die ganze Zeit !
  - Einige wollen gerne mehr Theorie, andere (mehr) mehr Praxis

Erfahrungen mit dem Ü4 2/4

UNI FREIBURG

- Zum Tempo der Vorlesung
  - Die allermeisten finden Tempo und Umfang genau richtig

    Positiv: Beispiele + genug Zeit, selber mitzudenken
  - Für einige könnte es ruhig etwas / deutlich schneller gehen

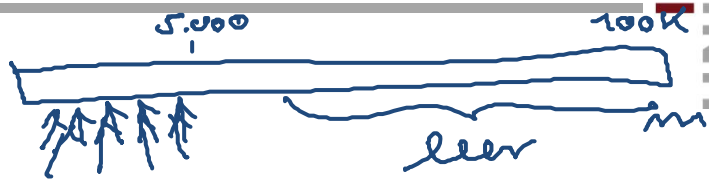
    Das sind tendenziell die, die am Mi / Do / Fr schon das Übungsblatt fertig haben

    "Schaue mir die Aufzeichnung mit erhöhter Geschwindigkeit an"

    Oder wie gesagt: machen Sie schon mal das Übungsblatt
  - Für einige ist auch jetzt schon das Tempo hoch

    Insbesondere bei den formaleren Sachen, wie O-Notation

Erfahrungen mit dem Ü4 3/4



- Hinweise zu den Lösungen
  - Die Summe der ASCII Codes war keine gute Hashfunktion:

    771.308 verschiedene Anfragen im AOL Query Log

    Also Hashtabelle in der Größenordnung sinnvoll

    Aber Summe der ASCII Codes selbst für einen String der Länge 50 gerade mal 5.000 → sehr viele Kollisionen am Anfang der Hashtabelle, während hinten alles leer
  - Viele haben dann die interne Hashfunktion benutzt

    Python: hash Java: hashCode C++: std::hash

Erfahrungen mit dem Ü4 4/4

UNI FREIBURG

- Häufigste Suchanfragen
  - Ergebnis von Aufgabe 2:

google	32163
yahoo	13646
ebay	13075
mapquest	8719
myspace	6877
…	

Das AOL Query Log ist wohlgemerkt von 2006

Da lief das Internet noch mit Dampf

Große Dateien

UNI FREIBURG

- Unit Tests
  - Unit Tests immer auf **kleinen** Beispielen und **schnell**

    Also z.B. nicht auf dem ganzen AOL Query Log

    Schreiben Sie ggf. eine kleine Beispieldatei, die können Sie dann auch nach Belieben gestalten
- SVN
  - **Niemals** große Dateien ins SVN einchecken

    Sonst 1 Punkt Abzug pro Byte der Datei … das wird teuer

    Von Ihren Unit Tests benötigte (kleine) Beispieldateien müssen dagegen ins SVN

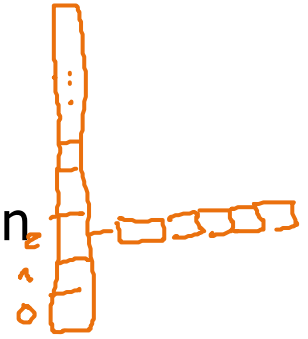
Universelles Hashing 1/12

UNI FREIBURG

- Zur Erinnerung
  - Wenn die Schlüsselmenge zufällig ist, tut es auch die einfach Hashfunktion $h(x) = x \bmod m$
  - Für eine beliebige Schlüsselmenge kann diese Funktion dagegen beliebig schlecht sein, zum Beispiel

    alle = 2 mod 10

    $h(x) = x \bmod 10$ und $x = 12, 42, 32, 72, 102, \ldots$



  - Allgemeiner: keine einzelne Hashfunktion kann für alle Schlüsselmengen gut sein

    Grund: h ist eine Funktion von U nach $\{0, \ldots, m-1\}$ und $|U| \gg m$

    Selbst im besten Fall werden so $|U| / m$ Schlüssel auf denselben Wert abgebildet

Universelles Hashing 2/12

UNI FREIBURG

- Idee
  - Eine Menge (Klasse) von Hashfunktionen zur Auswahl
  - Und zwar so, dass man leicht ein zufälliges Element aus dieser Menge wählen kann
  - Beispiel: $h(x) = (a \cdot x + b) \bmod m$ mit $a, b \in U$

    Das sind m^2 verschiedene Funktionen

    Wir sehen morgen, dass das keine gute Klasse ist
  - Beispiel: $h(x) = (a \cdot x + b) \bmod p \bmod m$ mit $a, b \in U$

    Ebenfalls $|U|^2$ verschiedene Funktionen

    Wir sehen morgen, dass das eine sehr gute Klasse ist

Universelles Hashing 3/12

UNI FREIBURG

- Was ist eine gute Klasse
  - Informal: eine Klasse H von Hashfunktionen ist dann gut, wenn es für jede Schlüsselmenge S viele Funktionen in H gibt, die S mit wenig Kollisionen abbilden

    Dann können wir nämlich einfach eine zufällige Funktion aus H wählen und hoffen, dass es gut klappt

    Wenn nicht, merken wir das und machen nach einer Weile einen Rehash, mit einer anderen Funktion aus H

    Wenn H die obigen Eigenschaft hat, wird das nicht oft passieren und "im Mittel" gut funktionieren

    Das machen wir jetzt in der Folge formaler

Universelles Hashing 4/12

UNI FREIBURG

- Definition ... erster Versuch (informal)
  - Sei H eine Menge von Hashfunktionen $U \to \{0, \ldots, m-1\}$
  - Dann gilt, dass für **jede** Schlüsselmenge $S \subseteq U$ ein Bruchteil von $|H| / C$ aller Hashfunktion in H "gut" sind
  - Dabei heißt eine Hashfunktion "gut" auf S, wenn es höchsten K Kollisionen gibt

  So könnte man das definieren, aber es ist aus mehreren Gründen umständlich:

  Man argumentiert über **alle Teilmengen** $S \subseteq U$, das macht Beweise schwierig

  Man hat **zwei Konstanten** C und K

Universelles Hashing 5/12

UNI FREIBURG

- Definition ... so machen wir es schließlich
  - Sei H eine Menge von Hashfunktionen $U \to \{0, ..., m-1\}$
    d.h. gut im Sinne der Folien vorher
  - H ist c-universell wenn für alle $x, y \in U$ mit $x \neq y$ gilt:

    $| \{h \in H : h(x) = h(y)\} | \leq c \cdot |H| / m$
  - Mit anderen Worten, wenn $h \in H$ zufällig gewählt, dann

    $\text{Prob}(h(x) = h(y)) \leq c \cdot 1 / m$

    Anders auf der Folie vorher, argumentieren wir jetzt nur noch über **Paare von Schlüsseln**, das ist viel einfacher

    Mit dieser einfachen Definition zeigen wir gleich, was wir eigentlich wollen: für jedes beliebige $S \subseteq U$ sind viele Funktionen aus H "gut" = wenig Kollisionen auf S

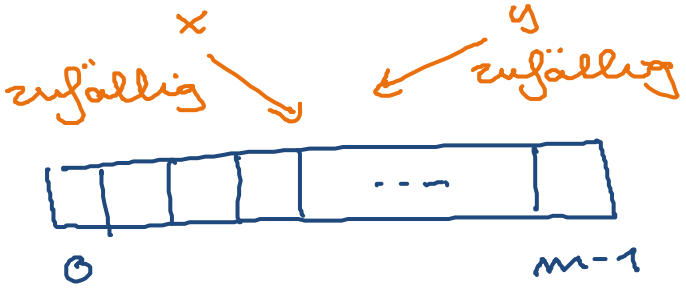
Universelles Hashing 6/12

UNI FREIBURG

- Vergleich mit zufälligem Werfen
  - Wenn man die Schlüssel x und y zufällig in die Hashtabelle wirft, dann ist die Kollisionswahrscheinlichkeit 1/m

    Mehr kann man von einer zufälligen Hashfunktion nicht verlangen … 1-universell ist also das Optimum

    Gut für das tiefere Verständnis: **beweisen** Sie, dass besser als 1-universell nicht geht

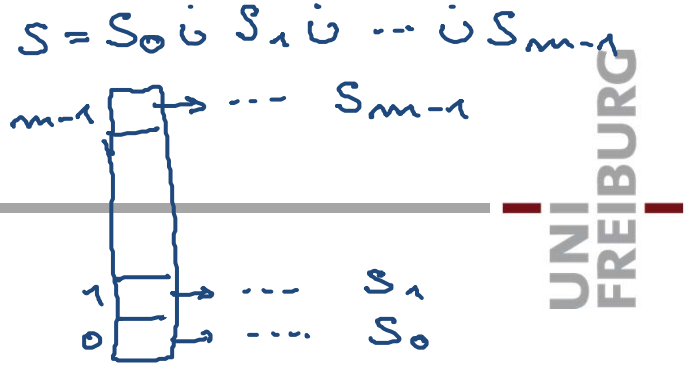


$$\Pr(\text{Kollision}) = \frac{1}{m}\cdot\frac{1}{m} + \frac{1}{m}\cdot\frac{1}{m} + \cdots$$

beide an Stelle 0 — beide an Stelle 1

$$= m \cdot \frac{1}{m}\cdot\frac{1}{m} = \frac{1}{m}$$ □

Universelles Hashing 7/12



- Zentraler Satz
  - Sei H eine c-universelle Klasse von Hashfunktionen
  - Sei S eine Menge von Schlüsseln und $h \in H$ zufällig gewählt
  - Sei S_i die Menge der Schlüssel x mit $h(x) = i$
  - Dann ist $E(|S_i|) \leq 1 + c \cdot |S| / m$ für alle i
  - Insbesondere: Falls $m = \Omega(|S|)$ gilt $E(|S_i|) = O(1)$

Bevor wir das beweisen, ein kleiner Auffrisch- bzw. Crash- Kurs in Wahrscheinl.keitsrechnung

Beispiel 1: $m = 10$, $|S| = 100$

IDEAL: $|S_i| = \frac{|S|}{m} = \frac{100}{10} = 10$

SATZ: $E(|S_i|) \leq 1 + 2 \cdot \frac{|S|}{m} = 21$, $c = 2$

Beispiel 2: $m = 100$, $|S| = 100$

IDEAL: $|S_i| = \frac{|S|}{m} = 1$

SATZ: $E(|S_i|) \leq 1 + 2 \cdot \frac{|S|}{m} = 3$, $c = 2$

Universelles Hashing 8/12

UNI FREIBURG

- Wahrscheinlichkeitsraum / Ereignisse
  - Wir beschränken uns hier auf den diskreten Fall
  - Wahrscheinlichkeitsraum Ω von sog. Elementarereignissen
  - Die haben Wahrscheinlichkeiten ... Bedingung $\Sigma_{e \in \Omega} \Pr(e) = 1$
  - Ereignis E = Teilmenge von Ω, Wahrsch. $\Pr(E) = \Sigma_{e \in E} \Pr(e)$
  - Zum Beispiel: zweimal würfeln, dann $\Omega = \{1,\ldots,6\}^2$

    (1,1) (1,2) (1,3) ⋮ (3,6) (4,6) (5,6) (6,6)

    Jedes e aus Ω hat dann Wahrscheinlichkeit $\Pr(e) = 1/36$

    E = beide Augenzahlen sind gerade, dann $\Pr(E) = 9 \cdot \frac{1}{36} = \frac{1}{4}$

    (2,2) (4,2) (6,2)
    (2,4) (4,4) (6,4)
    (2,6) (4,6) (6,6)

    9 Elem. ereignisse jedes $\Pr = \frac{1}{36}$

Universelles Hashing 9/12

UNI FREIBURG

- Zufallsvariable
  - ... weist einem Ausgang des Zufallsexperiments eine Zahl zu
  - Zum Beispiel: X = Summe Augenzahlen bei zweimal Würfeln
  - Sowas wie $X = 12$ oder $X \geq 7$ sind dann einfach Ereignisse
  - Beispiel 1: $\text{Prob}(X = 2) = \frac{1}{36}$ (1,1)
  - Beispiel 2: $\text{Prob}(X = 4) = \frac{3}{36} = \frac{1}{12}$ (1,3) (3,1) (2,2)
  - **Erwartungswert** ist definiert als $\mathbf{E}(X) = \Sigma\, k \cdot \Pr(X = k)$

    Intuitiv: gewichtetes Mittel der möglichen Werte von X, wobei die Gewichte die Wahrscheinlichkeiten der entspr. Werte sind

Beispiel X oben (Summe Augenzahlen)

$$E(X) = \underbrace{1 \cdot \Pr(X=1)}_{=0} + \underbrace{2 \cdot \Pr(X=2)}_{=\frac{1}{36}} + \cdots + \underbrace{12 \cdot \Pr(X=12)}_{=\frac{1}{36}}$$

Universelles Hashing 10/12

müssen nicht mal „unabhängig" sein

- Summe von Erwartungswerten
  - Für beliebige (diskrete) Zufallsvariablen $X_1, \ldots, X_n$ gilt

$$\mathbf{E}(X_1 + \ldots + X_n) = \mathbf{E}(X_1) + \ldots + \mathbf{E}(X_n)$$

X_1 = Augenzahl Würfel 1

X_2 = Augenzahl Würfel 2

$$E(X_1) \overset{\text{per Definition}}{=} \underbrace{1 \cdot \Pr(X_1=1)}_{=1/6} + \underbrace{2 \Pr(X_1=2)}_{=1/6} + \cdots + \underbrace{6 \cdot \Pr(X_1=6)}_{=1/6}$$

$$= \frac{1+2+3+4+5+6}{6} = \frac{\frac{1}{2} \cdot 6 \cdot 7}{6} = 3.5$$

$E(X_2)$ = genau dasselbe

$$E(\underbrace{X_1+X_2}_{X \text{ von der Folie vorher}}) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$$

Universelles Hashing 11/12

UNI FREIBURG

- Summe von Erwartungswerten, Korollar
  - Bei einem Zufallsexperiment tritt das Ereignis E mit Wahrscheinlichkeit p auf. Sei X die Anzahl der Auftreten von E bei n Ausführungen dieses Experimentes, dann ist **E**(X) = n · p

    Beispiel: **E**(Anzahl Sechser bei 60 mal Würfeln) = 10

Beweis:

$I_j = \begin{cases} 1, & \text{wenn } E \text{ bei Ausführung } j \text{ eintritt} \\ 0, & \text{sonst} \end{cases}$

↖ nennt man Indikatorvariable

Dann $X = \sum_{j=1}^{n} I_j$

$\Rightarrow E(X) = E\left(\sum_{j=1}^{n} I_j\right)$

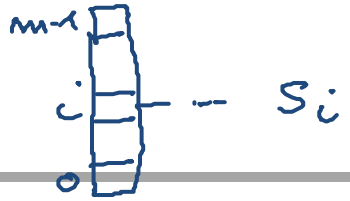
$\overset{\text{Satz}}{=} \sum_{j=1}^{n} \underbrace{E(I_j)}_{=p} = n \cdot p$ ∎

$E(I_j) = ?$

$E(I_j) = 0 \cdot Pr[I_j = 0] + 1 \cdot \underbrace{Pr[I_j = 1]}_{=p}$

$E(I_j) = Pr[I_j = 1] = p$ ∎

Universelles Hashing 12/12



$S_i = \{x \in S : h(x) = i\}$

- Beweis von $E(|S_i|) \leq 1 + c \cdot |S| / m$ für alle i

Annahme: $h \in H$ zufällig mit $\Pr(h(x) = h(y)) \leq c \cdot \frac{1}{m}$ $\forall x, y$ $x \neq y$ (VORR.)

Fall 1: $S_i = \emptyset \Rightarrow |S_i| = 0 \Rightarrow E(|S_i|) = 0$ ✓

Fall 2: $S_i \neq \emptyset \Rightarrow$ Sei x ein fixes Element aus S_i

$$I_y = \begin{cases} 1, & \text{falls } h(y) = i \\ 0, & \text{sonst} \end{cases}$$

Dann $|S_i| = 1 + \sum_{y \in S \setminus \{x\}} I_y$ (wegen $x \in S_i$)

$$\Rightarrow E(|S_i|) = 1 + \sum_{y \in S \setminus \{x\}} E(I_y) \quad (E(I_y) \leq c \cdot \frac{1}{m})$$

$$\leq 1 + c \cdot \frac{|S|}{m}$$ ∎

sogar $c \cdot \frac{|S|-1}{m}$

$E(I_y) = ?$ $y \in S \setminus \{x\}$ (Folie vorher)

$E(I_y) \overset{\downarrow}{=} \Pr(I_y = 1)$

$= \Pr(h(y) = i)$

$= \Pr(h(y) = h(x))$

$\leq c \cdot \frac{1}{m}$ nach VORR.

Literatur / Links

- Universelles Hashing
  - In Mehlhorn / Sanders:

    4 Hash Tables and Associative Arrays
  - In Wikipedia

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