Humboldt-Universität zu Berlin
Institut für Informatik
Lehrstuhl Logik in der Informatik
Prof. Dr. Nicole Schweikardt

9. Juli 2015

Big Data Analytics in Theorie und Praxis

Prof. Johann-Christoph Freytag, Ph.D. und Prof. Dr. Nicole Schweikardt
Sommersemester 2015

Theorie-Übungsblatt 5

Zu bearbeiten bis 16. Juli 2015

Aufgabe 1: **(25 Punkte)**

Sei $m \in \mathbb{N}_{\geq 1}$ und sei M eine Primzahl mit $M > m$. Sei H eine 2-universelle (d.h. paarweise unabhängige) Familie von Hashfunktionen h mit $h : \{1, \ldots, m\} \to \{0, \ldots, M-1\}$.

Sei außerdem $s \geq 2$, $t \in \{0, \ldots, M-1\}$ und $u_1, \ldots, u_s \in \{1, \ldots, m\}$.

Wir wählen zufällig, gleichverteilt ein $h \in H$ und betrachten für jedes $i \in \{1, \ldots, s\}$ die Zufallsvariable $Y_i := \begin{cases} 0 & \text{falls } h(u_i) > t \,, \\ 1 & \text{sonst} \,. \end{cases}$

Beweisen Sie, dass für alle $i, j \in \{1, \ldots, s\}$ mit $i \neq j$ die beiden Zufallsvariablen Y_i und Y_j unabhängig sind.

Zur Erinnerung:

(1) Zwei Zufallsvariablen X und Y heißen *unabhängig* genau dann, wenn für alle Werte a und b gilt, dass $\Pr(X = a \text{ und } Y = b) \;=\; \Pr(X = a) \cdot \Pr(Y = b)$.

(2) Eine Familie H von Funktionen von $\{1, \ldots, m\}$ nach $\{0, \ldots, M-1\}$ heißt *paarweise unabhängig* (oder *2-universell*), falls für alle $x_1, x_2 \in \{1, \ldots, m\}$ mit $x_1 \neq x_2$ und alle $y_1, y_2 \in \{0, \ldots, M-1\}$ gilt: Für ein zufällig, gleichverteilt aus H gewähltes h ist

$$\Pr\Big(\, h(x_1) = y_1 \;\text{ und }\; h(x_2) = y_2 \,\Big) \;=\; \frac{1}{M^2} \,.$$

Anmerkung: Zusammen mit Aufgabe 3 von Blatt 4 liefert diese Aufgabe die Rechtfertigung dafür, im Beweis zu Satz Ⓐ auf Seite 44 des Vorlesungsskripts zu verwenden, dass

$$\mathrm{Var}\left(\sum_{i=1}^{A} Y_i\right) \;=\; \sum_{i=1}^{A} \mathrm{Var}(Y_i)$$

für die dort betrachteten Variablen Y_i ist.

Auf der Rückseite finden Sie weitere Aufgaben.

Aufgabe 2: **(35 Punkte)**

Führen Sie den in der Vorlesung vorgestellten BJKST-Algorithmus für folgende Werte aus:

- $m := 101$ (in Worten: hundertundeins) und somit ist $U := \{0, 1, \ldots, 100\}$,
- $h(x) := a \cdot x + b \pmod{m}$ mit $a := 2$ und $b := 40$,
- $g(x) := a' \cdot x + b' \pmod{M}$ mit $a' := 3$, $b' := 5$, $M := 17$ und
- der Datenstrom $x_1, \ldots, x_{14}$ ist 17, 33, 17, 72, 12, 7, 11, 12, 69, 20, 1, 2, 28, 7.

Nehmen Sie für dieses Beispiel an, dass im BJKST-Algorithmus aus S Tupel entfernt werden, solange $|S| \geq 4$ ist.

Berechnen Sie für jedes Element x des Datenstroms die Werte $h(x)$, Nullen$(h(x))$ und $g(x)$. Geben Sie S immer an, nachdem Tupel zu S hinzugefügt wurden oder aus S entfernt wurden. Erwähnen Sie außerdem, wenn N sich ändert.

Was gibt der BJKST-Algorithmus für dieses Beispiel aus? Vergleichen Sie die Ausgabe mit der Anzahl der unterschiedlichen Elemente im obigen Datenstrom.

Aufgabe 3: **(40 Punkte)**

Seien $k, n, t \in \mathbb{N}_{\geq 1}$ mit $n = k \cdot t$.

Der in der Vorlesung vorgestellte Bloom-Filter wählt zufällig und unabhängig voneinander k Hash-Funktionen $h_1, \ldots, h_k$ mit $h_i : U \to \{1, \ldots, n\}$ für alle $i \in \{1, \ldots, k\}$.

In einem alternativen Ansatz wählen wir stattdessen zufällig und unabhängig voneinander k Hash-Funktionen $g_1, \ldots, g_k$ mit $g_i : U \to \{1+(i-1) \cdot t, \ldots, i \cdot t\}$ für alle $i \in \{1, \ldots, k\}$. Also wird der Wertebereich $\{1, \ldots, n\}$ in k Teilstücke der Länge t aufgeteilt und jede der k Hash-Funktionen bildet ihre Eingaben auf Werte von jeweils einem dieser Teilstücke ab. Außer bei der Wahl der Hashfunktionen ändert sich nichts im Vergleich zu dem in der Vorlesung vorgestellten Bloom-Filter.

Vergleichen Sie diese neue Variante des Bloom-Filters mit dem in der Vorlesung vorgestellten Bloom-Filter. Bestimmen Sie dazu insbesondere auch für die neue Variante die Wahrscheinlichkeit, dass bei Eingabe $u \in U$, aber $u \notin S$, „u gehört möglicherweise zu S“ ausgegeben wird. Welcher der beiden Varianten sollte man in der Praxis den Vorzug geben?