Internet-Algorithmen

Sommersemester 2014

Prof. Dr. Georg Schnitger
Hannes Seiwert

Arbeitsgruppe Theoretische Informatik, Institut für Informatik

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Übungsblatt 3

Ausgabe: 19.05.2014
Abgabe: 26.05.2014 vor Vorlesungsbeginn

Aufgabe 3.1. (8)

Ähnlichkeit von Vektoren

Bezeichne $S = \{x \in \mathbb{R}^n \mid ||x|| = 1\}$ die $(n-1)$-dimensionale Sphäre. Für zwei Vektoren $x, y \in S$ sei $\phi(x, y) \in [0, \pi]$ der Winkel zwischen x und y. Dann ist $e(x, y) := 1 - \phi(x, y)/\pi$ ein Ähnlichkeitsmaß.

Für jeden Vektor $r \in S$ definieren wir die Funktion $h_r : S \to \{0, 1\}$ durch

$$h_r(z) = \begin{cases} 1 & \langle z, r \rangle \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Zeige, dass $\mathcal{F} = \{h_r \mid r \in S\}$ eine Klasse ähnlichkeitsbewahrender Hashfunktionen für das Ähnlichkeitsmaß e auf Vektoren der Sphäre S ist.

Hinweis: Warum können wir o.B.d.A. $n = 2$ annehmen?

Aufgabe 3.2. (4+4)

Bloom-Filter und Blockansatz

Gegeben sei ein Bloom-Filter B mit m Zellen, in den n Elemente aus einem Universum U eingefügt werden. Es werden k unabhängige Hashfunktionen $h_1, \ldots, h_k$ verwendet, die gleichverteilt nach $\{1, \ldots, m\}$ abbilden.

In der Vorlesung wurde die Wahrscheinlichkeit $p_x := q_{n,m,k}$ einer falsch positiven Antwort für ein gegebenes Element $x \in U \backslash B$ mit $Q = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k$ approximiert.

a) Wir vergleichen die tatsächliche Wahrscheinlichkeit p_+ mit der Näherung Q.

Zeige in einem Beispiel mit $|U| = 2$, $n = 1$ und $k = m = 2$, dass $Q < p_+$ gelten kann.

b) Wir untersuchen nun einen alternativen Bloom-Filter-Ansatz:

Die m Positionen von B werden in k Blöcke von jeweils m/k Positionen zerlegt und dann k Hashfunktionen gewählt, die jeweils nur für einen Block zuständig sind, d.h.

$$h_i : U \to \left\{\tfrac{(i-1)\cdot m}{k} + 1, \ldots, \tfrac{i \cdot m}{k}\right\}, \quad i = 1, \ldots, k$$

Bestimme die Wahrscheinlichkeit p_0, dass B an einer beliebigen, fixierten Position eine Null speichert. **Berechne** anschließend die Wahrscheinlichkeit p'_+ einer falsch positiven Antwort im neuen Ansatz.

Sinkt oder steigt p'_+ im Vergleich zu p_+ im konventionellen Ansatz? Für den Vergleich darf angenommen werden, dass die Approximation $p_+ \approx Q$ hinreichend genau ist.

Hinweis: Die Bernoulli-Ungleichung darf benutzt werden: $1 + xn \leq (1 + x)^n$ für $x \geq -1$ und $n \in \mathbb{N}_0$.

Aufgabe 3.3. (2+6) *Flusskontrolle mittels Bloom-Filter*

Wir nehmen an, dass eine Menge F von n Paketflüssen in einen zählenden Bloom-Filter mit m Zählern $Z[i]$, $i = 1, \ldots, m$ eingefügt wird. Es werden wie üblich k unabhängige Hashfunktionen $h_1, \ldots, h_k$ verwendet. Sei v_x die Anzahl an Paketen eines Flusses $x \in F$, die unseren Router durchlaufen. Wir geben als Schätzung der Häufigkeit $v'_x := \min\{Z[h_1(x)], \ldots, Z[h_k(x)]\}$ aus.

a) Wie ist der Zusammenhang zwischen $\text{prob}[v'_x > v_x]$ und der Wahrscheinlichkeit $p_+ := q_{n,m,k}$ einer falsch positiven Antwort in einem herkömmlichem (nicht-zählendem) Bloom-Filter?

b) Tatsächlich können wir unseren Ansatz für Anwendungen in der Praxis noch verbessern. Wie sollte eine Modifikation aussehen, die Überschätzungen der Häufigkeiten v_x für $x \in F$ erschwert? Insbesondere, wenn v''_x die Abschätzung deines neuen Verfahrens ist, dann soll stets $v_x \leq v''_x \leq v'_x$ gelten und die Ungleichung $v''_x < v'_x$ soll in einigen Fällen auftreten können.

Hinweis: Eine kurze Antwort genügt!