Karlsruher Institut für Technologie

Maximilian Katzmann
Hans-Peter Lehmann
Stefan Walzer

Übungsblatt 13 – Aktivsession

Randomisierte Algorithmik – Wintersemester 2023/2024

Folgende Aufgaben werden in der Aktiv-Session am 1.2.2024 gemeinsam bearbeitet. Gib eine Lösung zu einer *der beiden Aufgaben (deine Wahl) wie gewohnt bis zum 8.2.2024 über Ilias ab.*

Aufgabe 1 – Schnitte Schätzen mit Bloomfiltern

Seien $n, m, k \in \mathbb{N}$. Alice und Bob wollen abschätzen, wie ähnlich ihr Musikgeschmack ist. Seien X die n Lieblingslieder von Alice und Y die n Lieblingslieder von Bob. Zu schätzen ist $\gamma := \frac{|X \cap Y|}{n} \in [0, 1]$. Beide gehen folgendermaßen vor.

- Alice konstruiert einen Bloomfilter $A[1..m] \in \{0,1\}^m$ für X unter Verwendung von k Hashfunktionen $h_1, \ldots, h_k$.
- Bob konstruiert einen Bloomfilter $B[1..m] \in \{0,1\}^m$ für Y unter Verwendung *derselben* k Hashfunktionen.
- Alice und Bob tauschen ihre Filter aus und berechnen $\delta := \frac{|\{i \in [m] | A[i] \neq B[i]\}|}{m}$.
- Alice und Bob berechnen basierend auf δ eine Schätzung $\bar{\gamma}$ für γ.

Löse folgende Teilaufgaben:

(a) Diskutiere: Welche Vor- und Nachteile könnte das Verfahren im Vergleich zum direkten Austausch von X und Y haben?

(b) Gewinne Intuition: Welche Werte von δ erwartest du (in etwa) für die Extremfälle, in denen $\gamma = 1$ bzw. $\gamma = 0$ gilt?
Hinweis: Du darfst hier und im Folgenden davon ausgehen, dass den Bloomfiltern eine „optimale“ Konfiguration mit $\alpha k = \ln(2)$ zugrundegelegt wurde.

(c) Berechne $\mathbb{E}[\delta]$ als Funktion von γ. Du darfst hierbei Terme niederer Ordnung unter den Tisch fallen lassen, also z.B. $(1 - \frac{1}{m})^m \approx e^{-1}$ schreiben, ohne ein $o(1)$ mitzuführen.
Hinweis: Zunächst scheint es, als könnten andere Parameter (z.B. $n, m, k, \alpha, \varepsilon$) auch eine Rolle spielen. Deren Einfluss verschwindet aber in Termen niederer Ordnung.

(d) Diskutiere: Welche Konzentrationsschranke eignet sich, um zu beweisen, dass δ mit hoher Wahrscheinlichkeit nahe an $\mathbb{E}[\delta]$ liegt?

(e) Stelle die Gleichung aus (c) um, sodass ersichtlich wird, wie eine Schätzung $\bar{\gamma}$ für γ aus δ berechnet werden kann.

(f) Spekuliere: Welche Rolle spielt die Wahl von k (bzw. von ε) im vorliegenden Kontext?

Aufgabe 2 – Cuckoo Hashing: Verwandte Fragen und Modelle

Aus mathematischer Sicht ist es oft nebensächlich und ablenkend Cuckoo Hashing mit zwei getrennten Tabellen aufzuziehen. Betrachten wir daher stattdessen folgende Variante:

> Es gibt *eine* Tabelle der Größe m, es gibt *zwei* Hashfunktionen $h_0, h_1 : D \to [m]$ und ein Schlüssel $x \in D$ darf entweder in $h_0(x)$ oder in $h_1(x)$ platziert werden.

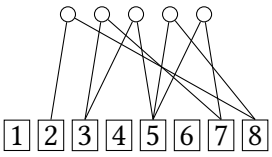
(a) Passe die insert-Methode aus der Vorlesung auf diese Variante an. Welcher Nachteil ergibt sich für die Anzahl der nötigen Hashfunktionsevaluationen?

Betrachte den bipartiten Graphen $B = (V_0, V_1, E_B)$ und den gewöhnlichen Graphen $C = (V, E)$, die für eine gegebene Schlüsselmenge $S \subseteq D$ folgendermaßen definiert sind.

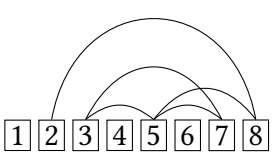
$$B = (S, [m], \{(x, h_0(x)) \mid x \in S\} \cup \{(x, h_1(x)) \mid x \in S\})$$
$$C = ([m], \{\{h_0(x), h_1(x)\} \mid x \in S\})$$

Beachte: In C sind sowohl Multikanten als auch Schleifen möglich. Beispiel: Für eine Schlüsselmenge $S = \{x_1, \ldots, x_5\}$ und Hashwerte wie folgt ergeben sich für B und C die Graphen wie illustriert.

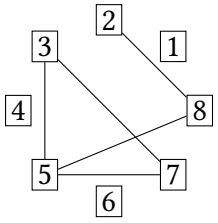
	$h_0(x)$	$h_1(x)$
x_1	2	8
x_2	7	3
x_3	3	5
x_4	5	8
x_5	7	5



B



C



auch C

(b) Zeige dass folgende Aussagen äquivalent sind.

(i) Alle Schlüssel aus S können in der Hashtabelle platziert werden.

(ii) Es gibt ein Matching der Größe $n = |S|$ in B.

(iii) C ist ein Pseudowald.

(iv) Man kann jede Kante aus C mit einer Richtung versehen, sodass jeder Knoten im entstandenen gerichteten Graphen $\vec{C}$ Eingangsgrad höchstens 1 hat.

Hinweis: Zur Definition eines Pseudowalds siehe `https://en.wikipedia.org/wiki/Pseudoforest`

Sei $G_{m,n}$ ein Zufallsgraph im Erdős-Renyi Modell mit m Knoten und n Kanten. Das heißt unter allen $\binom{\binom{m}{2}}{n}$ Graphen mit Knotenmenge $[m]$ und n Kanten ist $G_{m,n}$ uniform zufällig gewählt. Der Graph C hat große Ähnlichkeiten mit $G_{m,n}$.

(c) Mache dir die feinen Unterschiede zwischen der Verteilung von $G_{m,n}$ und C klar. Überlege dir zum Beispiel Urnenmodelle für die beiden Verteilungen. Wieviele Bälle sind in der Urne? Was bedeuten sie? Sampeln wir mit oder ohne Zurücklegen?