Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αλγοριθμική Επιστήμη Δεδομένων 2019 – 2020

Διδάσκοντες: Α. Παγουρτζής, Θ. Σούλιου, Δ. Φωτάκης

2η Σειρά Ασκήσεων

Άσκηση 1. Δίνεται ένα data stream με m στοιχεία $x_1, \ldots, x_m$, όπου το m είναι περιττός φυσικός και κάθε $x_i \in \{1, \ldots, n\}$ (για ευκολία, υποθέτουμε ότι $n \gg m$ και ότι κάθε στοιχείο εμφανίζεται μόνο μία φορά στο data stream – να σχολιάσετε όμως αν μια τέτοια υπόθεση είναι πράγματι εύλογη ή/και χωρίς βλάβη της γενικότητας). Για κάθε στοιχείο x_i, ορίζουμε $\text{rank}(x_i) = |\{j : x_j \leq x_i\}|$. Ο median M ενός τέτοιου data stream $x_1, \ldots, x_m$ είναι το στοιχείο x_i για το οποίο ισχύει ότι $\text{rank}(x_i) = m/2 + 1$.

Θεωρούμε τον εξής αλγόριθμο για τον προσεγγιστικό υπολογισμό του median σε ένα τέτοιο data stream: Για κάποιο $t \ll m$, επιλέγουμε τυχαίο δείγμα t στοιχείων από το data stream (ώστε κάθε στοιχείο να εμφανίζεται στο δείγμα με πιθανότητα t/m), υπολογίζουμε τον median M' στο τυχαίο δείγμα, και θεωρούμε ότι το M' αποτελεί μια καλή προσέγγιση του M.

Αληθεύει ότι υπάρχει $t = o(m)$, τέτοιο ώστε για κάθε τέτοιο data stream, με πιθανότητα τουλάχιστον $3/4$, ισχύει ότι $M' \in [M - n/10, M + n/10]$;

Αν η απάντησή σας είναι θετική, να αιτιολογήσετε κατάλληλα τον ισχυρισμό σας. Διαφορετικά, να δώσετε ένα αντιπαράδειγμα και να προτείνετε αποδοτικό αλγόριθμο για την προσέγγιση του median σε data streams αυτής της μορφής (και να αιτιολογήσετε τις εγγυήσεις απόδοσης του αλγορίθμου σας).

Άσκηση 2. Να λύσετε την Άσκηση **4.5.3** και την Άσκηση **4.5.4** από το βιβλίο LRU.

Άσκηση 3. Να λύσετε την Άσκηση **3.1.3** από το βιβλίο LRU.

Άσκηση 4. Λέμε ότι μια οικογένεια συναρτήσεων $\mathbf{H}$ είναι (d_1, d_2, p_1, p_2)-ευαίσθητη ως προς μια απόσταση d στοιχείων ενός συνόλου U αν για καθε $h \in \mathbf{H}$ και για κάθε $x, y \in U$, (i) αν $d(x, y) \leq d_1$, τότε $\text{Prob}[h(x) = h(y)] \geq p_1$, ενώ (ii) αν $d(x, y) \geq d_2$, τότε $\text{Prob}[h(x) = h(y)] \leq p_2$. Στο μάθημα, αποδείξαμε ότι η minhash οικογένεια συναρτήσεων είναι $(d_1, d_2, 1 - d_1, 1 - d_2)$-ευαίσθητη ως προς την Jaccard distance συνόλων, για κάθε $0 \leq d_1 < d_2 \leq 1$.

(α) Να βρείτε μια $(d_1, d_2, 1 - d_1/n, 1 - d_2/n)$-ευαίσθητη οικογένεια συναρτήσεων ως προς την απόσταση Hamming δυαδικών συμβολοσειρών μήκους n, για κάθε $n \geq 1$ και για κάθε $0 \leq d_1 < d_2 \leq n$.

(β) Να βρείτε μια $(d/2, 2d, 1/2, 1/3)$-ευαίσθητη οικογένεια συναρτήσεων ως προς την Ευκλείδεια απόσταση σημείων στο $\mathbb{R} \times \mathbb{R}$, για κάθε $d > 0$.

Άσκηση 5. Θεωρούμε τον αλγόριθμο Count-Min-Sketch που διατηρεί εκτιμήσεις της συχνότητας εμφάνισης των συχνά εμφανιζόμενων στοιχείων σε ένα data stream.

(α) Έστω C_1 και C_2 τα sketches που προκύπτουν για τα data streams σ_1 και σ_2, αντίστοιχα (θεωρούμε ότι έχουμε τις ίδιες παραμέτρους m και d, και τα ίδια hash functions $h_1, \ldots, h_d$). Υπάρχει τρόπος να συνδυάσουμε τα C_1 και C_2 ώστε το αποτέλεσμα να ταυτίζεται με το sketch C που προκύπτει (για τα ίδια m, d και $h_1, \ldots, h_d$) για την παράθεση των σ_1 και σ_2; Να αιτιολογήσετε κατάλληλα τον ισχυρισμό σας.

(β) Να προτείνετε (αποδοτική) μέθοδο χρήσης του Count-Min-Sketch ώστε να διατηρούμε εκτιμήσεις της συχνότητας εμφάνισης των συχνά εμφανιζόμενων στοιχείων σε data streams όπου έχουμε (i) τόσο εισαγωγές όσο και διαγραφές στοιχείων (turnstile model), ή (ii) χρονικό παράθυρο (δηλ. εκτιμούμε τις συχνότητες εμφάνισης μόνο των W τελευταίων στοιχείων του data stream). Για το (ii), να θεωρήσετε ότι το μέγεθος του χρονικού παράθυρου W είναι σημαντικά μεγαλύτερο από τη διαθέσιμη μνήμη.

Άσκηση 6. Να λύσετε την Άσκηση **7.3.2** από το βιβλίο LRU.

Άσκηση 7. Να λύσετε την Άσκηση **7.4.1** από το βιβλίο LRU.

Προθεσμία υποβολής και οδηγίες. Οι απαντήσεις θα πρέπει να υποβληθούν έως τις 28/5/2020, σε ηλεκτρονική μορφή. Για απορίες / διευκρινίσεις: στείλτε μήνυμα στη διεύθυνση `ads@corelab.ntua.gr`.