Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αλγοριθμική Επιστήμη Δεδομένων, 2018-19

Διδάσκοντες: Δ. Φωτάκης, Α. Παγουρτζής

1η Σειρά Ασκήσεων

Άσκηση 1. Λύστε τις ασκήσεις 6.3.1 και 6.4.2 του βιβλίου LRU.

Άσκηση 2. Εξετάστε την μέθοδο κατακερματισμού ανοιχτής διευθυνσιοδότησης (open addressing hashing) και:

(α) Εξηγήστε γιατί ο αναμενόμενος (μέσος) χρόνος επιτυχούς αναζήτησης είναι ίδιος με τον μέσο χρόνο εισαγωγής των στοιχείων στον πίνακα.

(β) Αποδείξτε ότι ο χρόνος αυτός φράσσεται άνω, εκτός από την ποσότητα $\frac{1}{\alpha}\ln\frac{1}{1-\alpha}+\frac{1}{\alpha}$, και από την ποσότητα $\frac{1}{1-\alpha}$, όπου $\alpha = n/m$ ο παράγοντας φόρτου.

Άσκηση 3. Απαντήστε στα παρακάτω ερωτήματα σχετικά με τον αλγόριθμο DGIM (που μετράει το πλήθος των '1' στα τελευταία k bits ενός stream ($k \leq N$, όπου N ένα δεδομένο μέγεθος 'παραθύρου').

(α) Ποιος είναι ο λόγος προσέγγισης του αλγορίθμου (στη χειρότερη περίπτωση); Μας ενδιαφέρει ο λόγος $\max(\frac{A}{C},\frac{C}{A})$, όπου C η πραγματική τιμή του μετρητή και A η τιμή που επιστρέφει ο αλγόριθμος.

(β) Μπορείτε να βελτιώσετε τον λόγο αυτό τροποποιώντας απλώς και μόνο την εκτίμηση για το πλήθος '1' στον παλαιότερο κάδο (η εκτίμηση στις διαφάνειες είναι 2^{i-1}); Ποιος είναι ο μικρότερος λόγος που μπορείτε να πετύχετε με αυτόν τον τρόπο;

(γ) Πώς μπορούμε να χρησιμοποιήσουμε ιδέες από τον DGIM για να εξοικονομήσουμε χώρο μνήμης στο πρόβλημα όπου το stream περιέχει ακεραίους και θέλουμε να υπολογίσουμε (προσεγγιστικά) το άθροισμα των k τελευταίων ($k \leq N$);

Άσκηση 4. 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 το 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 5. Να λύσετε τις ασκήσεις 4.5.1 και 4.5.3 από το βιβλίο LRU.

Άσκηση 6. Να λύσετε τις ασκήσεις 3.1.3 και 3.3.4 από το βιβλίο LRU.

Άσκηση 7. 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 Prob$[h(x) = h(y)] \geq p_1$, ενώ (ii) αν $d(x, y) \geq d_2$, τότε 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 $d > 0$.

Προθεσμία υποβολής και οδηγίες. Οι απαντήσεις θα πρέπει να υποβληθούν έως τις 24/4/2019, σε ηλεκτρονική μορφή.

Για απορίες / διευκρινίσεις: στείλτε μήνυμα στη διεύθυνση `ads@corelab.ntua.gr`.