Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αλγοριθμική Επιστήμη Δεδομένων 2024-25

Διδάσκοντες: Α. Παγουρτζής, Θ. Σούλιου

1η Σειρά Ασκήσεων

Άσκηση 1. Έστω ένα σύνολο U. Μια οικογένεια συναρτήσεων κατακερματισμού $\mathcal{H} = \{h : U \to [m]\}$ λέγεται *καθολική* αν

$$\forall x, y \in U,\ x \neq y : \quad \Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{m}$$

(*σημείωση: χρησιμοποιούμε τον συμβολισμό* $[m] = \{0, \dots m-1\}$)

Ισοδύναμα, για κάθε δύο διαφορετικές τιμές $x, y \in U$, υπάρχουν το πολύ $|\mathcal{H}|/m$ συναρτήσεις $h \in \mathcal{H}$ για τις οποίες $h(x) = h(y)$.

(α) Αποδείξτε ότι για $a \in [m] \setminus \{0\}, b \in [m]$ η οικογένεια συναρτήσεων $h_{a,b}(x) = (ax + b) \bmod m$ δεν έχει την ιδιότητα της καθολικότητας για $U = [m^k], k \geq 2$.

(β) Εξηγήστε αν και γιατί ισχύει ότι για πρώτο αριθμό $p > |U| = m^k, k \geq 2$ και για $a \in [p] \setminus \{0\}, b \in [p]$ η οικογένεια συναρτήσεων $h_{a,b}(x) = ((ax + b) \bmod p) \bmod m$ έχει την ιδιότητα της καθολικότητας.

(γ) Δώστε ένα συγκεκριμένο παράδειγμα της παραπάνω ιδιότητας, για $m = 7, p = 53, k = 2$ και πέντε ζεύγη κλειδιών (x_i, y_i) της επιλογής σας. Δείξτε για ποιες συναρτήσεις της οικογένειας τα ζεύγη κλειδιών που επιλέξατε έχουν την ίδια hash value και για ποιες έχουν διαφορετική.

(γ) Εξακολουθεί να ισχύει η ιδιότητα της καθολικότητας αν στο ερώτημα (β) θέσουμε $U = [p^2]$; Εξηγήστε την απάντησή σας.

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

(α) Εξηγήστε γιατί ο μέσος χρόνος επιτυχούς αναζήτησης, μετά από εισαγωγή n στοιχείων, είναι ίδιος με τον μέσο χρόνο εισαγωγής των στοιχείων στον πίνακα.

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

Υπόδειξη: ξεκινήστε με μια εκτίμηση για το αναμενόμενο πλήθος δοκιμών κατά την εισαγωγή του i-οστού στοιχείου υποθέτοντας uniform hashing.

Άσκηση 3. Να λύσετε τις ασκήσεις **6.3.1** και **6.3.4** του βιβλίου MMDS.

Άσκηση 4.

(α) Εξηγήστε συνοπτικά την ορθότητα της μεθόδου A-priori για το πρόβλημα της *εξόρυξης συχνών συνόλων στοιχείων* (frequent itemset mining) από δεδομένα τύπου 'καλαθιού αγορών' (market basket data). Πόσες διασχίσεις (passes) πραγματοποιούνται στην βάση δεδομένων;

(β) Εξηγήστε γιατί η χρονική πολυπλοκότητα του αλγορίθμου A-priori είναι *πολυωνυμική ως προς την έξοδο*.

(γ) Συζητήστε αν η πολυπλοκότητα του αλγορίθμου FP-Growth είναι επίσης πολυωνυμική ως προς την έξοδο.

Άσκηση 5.

(α) Εκτελέστε τον αλγόριθμο A-priori στο παρακάτω παράδειγμα. Υποθέστε ότι το κατώφλι στήριξης είναι $s = 5$.

```
{a  b  c  d}     {   b     d}   {a     c  d}     {   b     d}
{a        d}     {a         }   {a  b     d}     {   b  c  d}
{   b     d}     {a  b     d}   {a     c   }     {a     c   }
{a  b  c  d}     {   b      }   {   b  c   }     {a     c  d}
{   b  c   }     {   b     d}   {a  b  c  d}     {   b  c  d}
```

(β Βρείτε όλους τους κανόνες συσχέτισης που έχουν στήριξη (support) τουλάχιστον 5 και confidence τουλάχιστον 60%. Εξηγήστε με ποιον τρόπο θα αξιοποιήσετε την αντιμονοτονικότητα του confidence ως προς το δεξί μέρος (RHS) των κανόνων.

(γ) Εκτελέστε τον αλγόριθμο του Toivonen στο ίδιο παράδειγμα, θεωρώντας ότι το δείγμα είναι οι 8 εγγραφές των 2 επάνω γραμμών. Χρησιμοποιήστε ως κατώφλι στο δείγμα $s' = 2$ και $s'' = 1$ (θα κάνετε δύο διαφορετικές εκτελέσεις του αλγορίθμου). Τι παρατηρείτε;

Άσκηση 6. Να λύσετε τις ασκήσεις **7.2.2**, **7.2.3**, **7.2.5**, **7.3.2** και **7.4.1** από το βιβλίο MMDS.

Προθεσμία υποβολής και οδηγίες. Οι απαντήσεις θα πρέπει να υποβληθούν έως τις 27/4/2025, σε ηλεκτρονική μορφή. Συνιστάται *θερμά* να αφιερώσετε ικανό χρόνο για να λύσετε τις ασκήσεις μόνοι σας προτού καταφύγετε σε οποιαδήποτε *θεμιτή* βοήθεια (διαδίκτυο, βιβλιογραφία, συζήτηση με συμφοιτητές). Σε κάθε περίπτωση, οι απαντήσεις θα πρέπει να είναι *αυστηρά* ατομικές (δηλαδή όχι 'copy-paste'). Για απορίες / διευκρινίσεις: στείλτε μήνυμα στη διεύθυνση `ads@corelab.ntua.gr`.

Καλή επιτυχία!