Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αλγοριθμική Επιστήμη Δεδομένων

Διδάσκοντες: Α. Παγουρτζής, Θ. Σούλιου, Β. Νάκος

Εξέταση Ιουλίου 2021 – Διάρκεια Εξέτασης: **120′**

Για το 'άριστα' αρκεί να συγκεντρώσετε 10 μονάδες

1.	
2.	
3.	
4.	
5.	
Σ.	

Ονοματεπώνυμο: . . . . . .

Μετ. πρόγραμμα: . . . . . Α. Μ.: . . . . . .

Θέμα 1 (3 μον.)

(α) Εξηγήστε τη σχέση του προβλήματος *εξόρυξης συχνών συνόλων στοιχείων* (frequent itemset mining) από δεδομένα τύπου 'καλαθιού αγορών' (market basket data), με το πρόβλημα *εξόρυξης κανόνων συσχέτισης* (association rules mining) από δεδομένα τέτοιου τύπου.

(β) Εξηγήστε την ορθότητα της μεθόδου A-priori για το πρόβλημα αυτό. Πόσες διασχίσεις (passes) πραγματοποιούνται στην βάση δεδομένων;

(γ) Συζητήστε αν και γιατί η χρονική πολυπλοκότητα του αλγορίθμου A-priori είναι *πολυωνυμική ως προς την είσοδο* ή *πολυωνυμική ως προς την έξοδο*.

(δ) Εκτελέστε τον αλγόριθμο του Toivonen στο παρακάτω παράδειγμα. Υποθέστε ότι το κατώφλι στήριξης είναι $s = 4$, και ότι το δείγμα είναι οι 3 πρώτες εγγραφές. Χρησιμοποιήστε ως κατώφλι στο δείγμα $s' = 1$ και $s'' = 2$ (θα κάνετε δύο διαφορετικές εκτελέσεις του αλγορίθμου). Τι παρατηρείτε;

Βάση δεδομένων (transaction database):

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

Θέμα 2 (2 μον.)

Θεωρήστε τον αλγόριθμο των Flajolet-Martin για την εκτίμηση του πλήθος διαφορετικών στοιχείων που εμφανίζονται σε ένα data stream και υποθέστε ότι ο αλγόριθμος χρησιμοποιεί μία μόνο συνάρτηση κατακερματισμού (hash function) και ότι στο data stream εμφανίζονται m διαφορετικά στοιχεία, όπου το m είναι δύναμη του 2. Να εκτιμήσετε (ως συνάρτηση των m και c) την πιθανότητα (i) ο αλγόριθμος να επιστρέψει μια τιμή $\hat{m} < m/2^c$, και (ii) ο αλγόριθμος να επιστρέψει μια τιμή $\hat{m} > m \cdot 2^c$.

Θέμα 3 (2 μον.)

(α) Δίνεται η οικογένεια συναρτήσεων κατακερματισμού ($\{0, \ldots, 99\} \to \{0, 19\}$) με παράμετρο a που παίρνει ομοιόμορφα τυχαίες τιμές από το σύνολο $\{1, \ldots, 19\}$, και ορίζεται ως εξής:

$$h_a(k) = a \cdot k + 3 \bmod 20$$

Έχει η οικογένεια h_a την ιδιότητα της καθολικότητας ή όχι;

(β) Τι ισχύει για την οικογένεια συναρτήσεων

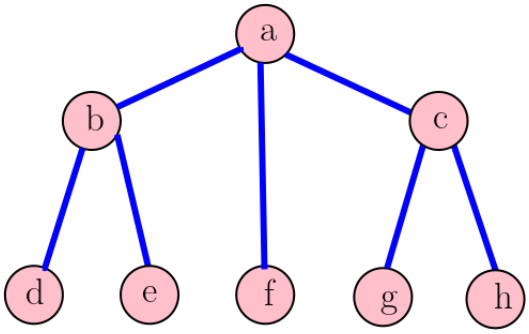
$$h_a(k) = (a \cdot k + 3 \bmod 101) \bmod 20$$

με παράμετρο a που παίρνει ομοιόμορφα τυχαίες τιμές από το σύνολο $\{1, \ldots, 100\}$;

Εξηγήστε τις απαντήσεις σας στα παραπάνω ερωτήματα.

Θέμα 4 (2 μον.)

(α) Βρείτε το edge betweeness κάθε ακμής του παρακάτω γράφου:



(β) Περιγράψτε έναν όσο το δυνατόν πιο αποδοτικό αλγόριθμο για τον υπολογισμό του edge betweeness αν γνωρίζουμε ότι ο γράφος εισόδου είναι δέντρο. Ποια είναι η πολυπλοκότητα του αλγορίθμου σας; Εξηγήστε.

Θέμα 5 (3 μον.)

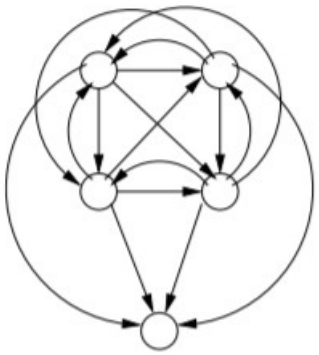
(α) Δώστε ένα παράδειγμα ενός συνόλου δεδομένων και μιας επιλογής k αρχικών κεντροειδών (centroids) έτσι ώστε όταν τα σημεία επανατοποθετούνται στο πλησιέστερο κεντροειδές τους στο τέλος, τουλάχιστον ένα από τα k αρχικά σημεία να επανατοποθετείται σε διαφορετική συστάδα (cluster). Στο παράδειγμα να φαίνονται τα αρχικά κεντροειδή και τα κεντροειδή που προκύπτουν σε κάθε βήμα. Ποιά η διαφορά του αλγορίθμου BFR από τους k-means αλγόριθμους και ποιά είναι τα βήματα του Cure;

(β) Το σχήμα που ακολουθεί είναι ένας πίνακας χρησιμότητας, ο οποίος αναπαριστά τις αξιολογήσεις, σε κλίμακα 1-5 αστέρων, οκτώ αντικειμένων, a έως h, από τρεις χρήστες A, B και C. Υπολογίστε τα ακόλουθα από τα δεδομένα αυτού του πίνακα:

1. Αντιμετωπίζοντας τον πίνακα χρησιμότητας ως boolean, υπολογίστε την απόσταση Jaccard μεταξύ κάθε ζεύγους χρηστών.
2. Επαναλάβετε το ερώτημα 1, αλλά χρησιμοποιήστε την απόσταση συνημιτόνου.
3. Αντιμετωπίστε τις αξιολογήσεις 3, 4 και 5 ως 1 και τις 1, 2 και κενές ως 0. Υπολογίστε την απόσταση Jaccard μεταξύ κάθε ζεύγους χρηστών.
4. Επαναλάβετε το ερώτημα 3, αλλά χρησιμοποιήστε την απόσταση συνημιτόνου.
5. Κανονικοποιήστε τον πίνακα αφαιρώντας από κάθε μη κενή εγγραφή τη μέση τιμή για τον χρήστη του.
6. Χρησιμοποιώντας τον κανονικοποιημένο πίνακα από το ερώτημα 5, υπολογίστε την απόσταση συνημιτόνου μεταξύ κάθε ζεύγους χρηστών.

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

(γ) Ας υποθέσουμε ότι ο Ιστός αποτελείται από μια κλίκα n κόμβων (κλίκα = σύνολο κόμβων με όλες τις δυνατές ακμές μεταξύ τους) και έναν επιπλέον κόμβο προς τον οποίο υπάρχει ακμή από κάθε κόμβο της κλίκας. Το Σχήμα που ακολουθεί δείχνει αυτό το γράφημα για την περίπτωση $n = 4$. Προσδιορίστε το PageRank κάθε σελίδας, ως συνάρτηση του n και του β.



Καλή επιτυχία!