ΑΝΑΦΟΡΑ ΕΡΓΑΣΙΑΣ 3 - ΥΛΟΠΟΙΗΣΗ ΤΕΛΕΣΤΩΝ

Δημήτριος Παγώνης

28 Μαΐου 2025

Μέρος 1 (Containment queries)

Απλή μέθοδος αναφοράς (naive)

Η συνάρτηση naive() υλοποιεί μια απλή μέθοδο αναφοράς για το πρόβλημα του containment. Για κάθε ερώτηση (query), διατρέχει όλες τις συναλλαγές του πίνακα και ελέγχει αν η κάθε συναλλαγή περιέχει όλα τα αντικείμενα της ερώτησης (δηλαδή αν το query είναι υποσύνολο της συναλλαγής). Αν μια συναλλαγή ικανοποιεί το ερώτημα, το id (θέση) της συναλλαγής προστίθεται στη λίστα αποτελεσμάτων για το συγκεκριμένο query. Στο τέλος, η συνάρτηση επιστρέφει για κάθε ερώτηση μια λίστα με τα ids των συναλλαγών που την περιέχουν ως υποσύνολο.

Επιπλέον, στη συνάρτηση naive(), για κάθε ζεύγος query και συναλλαγής, εφαρμόζεται ένας αλγόριθμος τύπου merge-join. Επειδή τόσο το query όσο και η συναλλαγή είναι ταξινομημένες λίστες, συγκρίνω τα στοιχεία τους ένα-ένα, προχωρώντας σε κάθε λίστα ανάλογα με το ποιο στοιχείο είναι μικρότερο. Αν όλα τα στοιχεία του query βρεθούν στη συναλλαγή με αυτόν τον τρόπο, τότε το query είναι υποσύνολο της συναλλαγής.

Exact signature file

Για τις ανάγκες του δεύτερου ερωτήματος δημιούργησα μια συνάρτηση που δημιουργεί για κάθε συναλλαγή μια δυαδική υπογραφή (bitmap), όπου κάθε bit αντιστοιχεί σε ένα αντικείμενο του καθολικού συνόλου. Το bit είναι 1 αν το αντικείμενο υπάρχει στη συναλλαγή και 0 διαφορετικά. Για κάθε ερώτηση (query), η μέθοδος ελέγχει αν όλα τα αντικείμενα του query υπάρχουν στη συναλλαγή, εξετάζοντας αν τα αντίστοιχα bits στο bitmap της συναλλαγής είναι ενεργοποιημένα (δηλαδή ίσα με 1). Αυτό γίνεται με bitwise πράξεις, που επιτρέπουν πολύ γρήγορο έλεγχο υποσυνόλου. Αν όλα τα bits που αντιστοιχούν στα αντικείμενα του query είναι 1, τότε η συναλλαγή περιέχει το query και προστίθεται στα αποτελέσματα. Η μέθοδος αυτή είναι ιδιαίτερα αποδοτική για μεγάλα σύνολα δεδομένων, καθώς μειώνει σημαντικά τον χρόνο αναζήτησης σε σχέση με την απλή γραμμική αναζήτηση.

Exact bitslice signature file

Η μέθοδος bitslice signature file δημιουργεί για κάθε αντικείμενο ένα bitmap (bitslice), όπου κάθε bit αντιστοιχεί σε μία συναλλαγή: το bit είναι 1 αν το αντικείμενο υπάρχει στη συγκεκριμένη συναλλαγή και 0 διαφορετικά. Για κάθε ερώτηση (query), η μέθοδος παίρνει τα bitmaps των αντικειμένων του query και υπολογίζει το bitwise AND μεταξύ τους. Το αποτέλεσμα είναι ένα νέο bitmap, όπου τα bits που είναι 1 αντιστοιχούν στις συναλλαγές που περιέχουν όλα τα αντικείμενα του query. Έτσι, με μία μόνο bitwise πράξη, εντοπίζονται γρήγορα οι συναλλαγές που ικανοποιούν το ερώτημα, καθιστώντας τη μέθοδο πολύ αποδοτική για μεγάλα σύνολα δεδομένων.

Inverted file

Η μέθοδος για την υλοποίηση της inverted file δημιουργεί για κάθε αντικείμενο μια ανεστραμμένη λίστα (inverted list), η οποία περιέχει τα ids των συναλλαγών όπου εμφανίζεται το συγκεκριμένο αντικείμενο. Οι λίστες αυτές είναι ταξινομημένες ως προς τα ids των συναλλαγών. Για κάθε ερώτηση (query), η μέθοδος παίρνει τις λίστες που αντιστοιχούν στα αντικείμενα του query και υπολογίζει την τομή τους με έναν αποδοτικό merge αλγόριθμο για ταξινομημένες λίστες. Οι συναλλαγές που ανήκουν στην τομή

είναι εκείνες που περιέχουν όλα τα αντικείμενα του query και επιστρέφονται ως αποτελέσματα. Αυτή η προσέγγιση επιτρέπει γρήγορη αναζήτηση containment queries, ειδικά όταν τα αντικείμενα είναι σπάνια ή οι λίστες είναι μικρές.

Εντολες εκτέλεσης κώδικα

```
python3 containment_queries.py <transactions.txt> <queries.txt> <qnum> <method>
```

Μέρος 2 (Relevance queries)

Ο κώδικας του relevance.py υλοποιεί δύο μεθόδους για την αποτίμηση ερωτημάτων σχετικότητας (relevance queries) σε ένα σύνολο συναλλαγών: τη μέθοδο inverted file και μια απλή (naive) μέθοδο. Αρχικά, διαβάζει τις συναλλαγές και τα queries από αρχεία και δημιουργεί ένα inverted file που για κάθε αντικείμενο καταγράφει σε ποιες συναλλαγές εμφανίζεται και πόσες φορές. Υπολογίζει επίσης για κάθε αντικείμενο έναν παράγοντα σπανιότητας (rarity factor), που βασίζεται στο πλήθος των συναλλαγών όπου εμφανίζεται. Για κάθε query, η naive μέθοδος υπολογίζει το relevance score κάθε συναλλαγής μετρώντας τις εμφανίσεις των αντικειμένων του query, ενώ η μέθοδος inverted file ενώνει τις inverted lists των αντικειμένων του query και υπολογίζει το relevance score μόνο για τις σχετικές συναλλαγές. Τα αποτελέσματα ταξινομούνται φθίνοντα ως προς το relevance score και εμφανίζονται τα k καλύτερα, ενώ το πρόγραμμα μετρά και εμφανίζει τον χρόνο εκτέλεσης κάθε μεθόδου.

Εντολή εκτέλεσης κώδικα

```
python3 relevance.py <transactions.txt> <queries.txt> <qnum> <method> <k>
```