ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

MultiDimensional Partitioning Framework Based on Query-Aware and Skew-Torelant Space-Filling Curves

Ριγγς Βασίλης

A.M. : E13163

Επιβλέπων Καθηγητής : Χρήστος Δουλκερίδης

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ, 2020

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

B+ Tree

- Ευρετήρια
- Tree-Based Indexing Δομές
- Δομή B+ Tree
- B+ Tree INSERT
- B+ Tree DELETE
- B+ Tree SEARCH
- B+ Tree RANGE-SEARCH
- Διπλότυπα Κλειδιά και Φύλλα Υπερχείλισης

Καμπύλες Πλήρωσης Χώρου

- Εισαγωγή
- Πολυδιάστατος Χώρος και Διάταξη
- Σχέση Καμπυλών Πλήρωσης Χώρου και B+ Tree

Πειραματική Αξιολόγηση

- Μεθοδολογία Παραγωγής Αποτελεσμάτων
-

Εισαγωγή

Τελευταία, η διαχείριση των δεδομένων παίζει έναν ολοένα και σημαντικότερο ρόλο στα Data Analytics καθώς η πρόσβαση στα δεδομένα περιορίζεται από τις συνιστώσες ενός υπολογιστικού συστήματος (κυρίως CPU, GPU και RAM). Επιπλέον έχουν αυξηθεί σημαντικά και οι συλλογές των δεδομένων προς ανάλυση. Στις περισσότερες περιπτώσεις, οι αναλύσεις αυτές χρειάζονται ένα μικρό κομμάτι των συνολικών δεδομένων με αποτέλεσμα οι συμβατικοί τρόποι αποθήκευσης των δεδομένων (αρχεία χωρίς, μερική, ή ολική διάταξη δεδομένων) να έχουν καταστεί ακατάλληλοι.

Τα συνολικά δεδομένα είναι αδύνατον να προσπελάζονται πλέον από σειριακές δομές μνήμης (αρχεία) ή δομές δεδομένων που λειτουργούν βάσει διάταξης στοιχείων. Αφενός μεν γιατί το μέγεθος των δεδομένων είναι τεράστιο με αποτέλεσμα ο χρόνος προσπέλασης γιναντώνεται, και αφετέρου, πολυδιάστατα δεδομένα, όπως οι συντεταγμένες π.χ. δεν μπορούν να διαταχθούν και να ταξινομηθούν. Αυτό, έχει ως συνέπεια να γίνουν έρευνες προς την κατεύθυνση τεχνικών που θα χειρίζονται πολυδιάστατα δεδομένα, ομαδοποιώντας, διατάσσοντας, και αποθηκεύοντας τα με αποδοτικό τρόπο.

Η τεχνική που προτείνει η παρούσα εργασία, ο σκοπός της οποίας είναι να μελετήσει και να αναπαράγει την ερεύνα που έχει γίνει από τους Shoji Nishimura & Haruo Yokota στο επιστημονικό τους paper «Quilts: Multidimensional Partitioning Framework Based on Query-Aware and Skew-Tolerant Space-Filling Curves» είναι το Data Skipping. Η τεχνική αυτή, όπως περιγράφεται στο paper, χωρίζει τα δεδομένα σε «κομμάτια» (στη συνέχεια θα αναφέρονται ως pages) με σκοπό την αποδοτικότερη ανάκτηση των δεδομένων, δηλαδή καλείται να ελαχιστοποιήσει τα pages που ανακτώνται μέσω ενός ερωτήματος εύρους (range query) και συνεπώς να ελαχιστοποιήσει τον αριθμό των δεδομένων που ανακτώνται για ανάλυση και μελέτη. Η αποδοτική ανάκτηση των δεδομένων απαιτεί έναν βέλτιστο τρόπο τμηματοποίησης αυτών σε pages βάσει της κατανομής των δεδομένων (data distribution) και των μοτίβων ερωτημάτων (query patterns). Συνεπώς, η τμηματοποίηση των δεδομένων με βέλτιστο τρόπο, αποτελεί ένα NP-hard problem.

Οι Nishimura & Yokota προτείνουν ένα framework σύμφωνα με το οποίο κάθε διαμέριση του πολυδιάστατου χώρου των δεδομένων αντιστοιχίζεται αμφιμονοσήμαντα με ένα αναγνωριστικό. Γίνεται πλέον εύκολο να αντιστοιχίσουμε δεδομένα με πολλές συνιστώσες ή χαρακτηριστικά (τα οποία ορίζουν τις διαστάσεις) με μονοδιάστατες μεταβλητές. Έτσι, δημιουργείται μια «ελαφριά» διάταξη στα δεδομένα καθώς μεταφέρονται από τις n διαστάσεις, στην μία διάσταση. Το framework, για να το πετύχει αυτό, επιστρατεύει την τεχνική των καμπυλών πλήρωσης χώρου (γνωστές και ως Space-Filling Curves). Οι καμπύλες πλήρωσης χώρου χρησιμοποιούνται για να αντιστοιχίσουν μια διαμέριση του ν-διάστατου χώρου σε ένα μοναδικό-για κάθε διαμέριση-αναγνωριστικό. Με τον τρόπο αυτό, κάθε ένα πολυδιάστατο στοιχείο <xi, yi, zi, ..ni>, από ένα σύνολο στοιχείων που βρίσκονται σε ένα αρχείο αποκτά ένα μονοδιάστατο κλειδί k που είναι το αναγνωριστικό της διαμέρισης του χώρου στην οποία βρίσκεται το στοιχείο αυτό. Στη συνέχεια, το στοιχείο εισάγεται σε μία δομή ευρετηρίου: INSERT<k, (xi, yi, zi, .. ni)> Οι δομές αυτές καλούνται να μειώσουν δραστικά το κόστος των πράξεων που εκτελεί ένα υπολογιστικό σύστημα, όπως κάποιες διακριτές πράξεις που είναι οι: INSERT, DELETE, SEARCH (ή FIND) και κάποιες περισσότερο περίπλοκες όπως η συνάρτηση RANGE_SEARCH (ερώτημα εύρους). Ως κόστος ορίζω τον αριθμό των λειτουργιών I/O του δίσκου που πραγματοποιεί η δομή για μία συγκεκριμένη πράξη.

Η δομή ευρετηρίου που χρησιμοποιώ στην παρούσα εργασία είναι ένα B+ Tree, ελαφρώς τροποποιημένο και ανήκει στην οικογένεια των Tree-Based Indexing δομών. Ένα B+ Tree παίρνει τη μορφή ενός ισορροπημένου δέντρου που παρουσιάζει μειωμένη απόδοση για τις INSERT και DELETE, και απαιτεί επιπλέον χώρο (σε αποδεκτά επίπεδα), αλλά παρουσιάζει αυξημένη απόδοση για τις SEARCH και RANGE_SEARCH.

Η παρούσα εργασία, βασιζόμενη στην εργασία των Nishimura & Yokota στο επιστημονικό τους paper, το «QUILTS», αποτελεί μία προσπάθεια για μελέτη και αξιολόγηση του προβλήματος της εύρεσης κατάλληλης καμπύλης πλήρωσης χώρου δοσμένης της κατανομής των δεδομένων (data distribution) και μοτίβου ερωτήματος (query pattern). Χωρίζεται σε δύο κύρια σκέλη. Στο πρώτο σκέλος γίνεται η παρουσίαση των τεχνικών, των δομών που χρησιμοποιώ και του προβλήματος που θέλω να επιλύσω. Στο δεύτερο σκέλος, παρουσιάζω και αξιολογώ τα ευρήματα μου.

B+ Tree

Space-Filling Curves

Στη Μαθηματική Ανάλυση, με τον όρο καμπύλη πλήρωσης χώρου (space filling curve) αναφερόμαστε σε καμπύλες που το σύνολο τιμών τους είναι ολόκληρο το μοναδιαίο τετράγωνο [0,1]^2 (ή ολόκληρος ο υπερκύβος [0,1]^n). Πολλές φορές αυτές οι καμπύλες αναφέρονται και ως καμπύλες Πεάνο, επειδή ο Τζιουζέπε Πεάνο ήταν το πρώτος μαθηματικός που ανακάλυψε μια τέτοια καμπύλη (παρότι και ο Ντάβιντ Χίλμπερτ είχε αναφερθεί σε παρόμοιες κατασκευές).

Γενικά, μια συνεχής καμπύλη σε έναν χώρο 2, 3 ή περισσοτέρων διαστάσεων αποτελεί ένα μονοπάτι ενός σημείου που κινείται με συνεχή τρόπο. Ο επίσημος μαθηματικός ορισμός (ο οποίος εξαλείφει την αοριστία της προηγούμενης φράσης) είναι ο παρακάτω (Τζόρνταν 1887):

«Μία καμπύλη (με σημεία που υποδηλώνουν την αρχή και το τέλος) είναι μια συνεχής συνάρτηση
r : [0,1] → X, όπου ο χώρος X μπορεί να είναι οποιοσδήποτε τοπολογικός χώρος.»

Συνήθως, ο χώρος X αναφέρεται σε Ευκλείδειους χώρους όπως ο R^2 (επίπεδη καμπύλη) ή ο R^3 (καμπύλη στο χώρο). Πολύ συχνά με τον όρο καμπύλη αναφερόμαστε στην εικόνα της αντίστοιχης συνάρτησης , δηλαδή στο σύνολο όλων των δυνατών τιμών (σημείων) που μπορεί να πάρει η συνάρτηση. Επιπλέον, μπορούμε να ορίσουμε και καμπύλες χωρίς αρχικό ή τελικό σημείο (π.χ. καμπύλες που εκτείνονται στο άπειρο) αν θεωρήσουμε ως πεδίο ορισμού το ανοικτό σύνολο (0,1) ή ολόκληρο το R.

Μία καμπύλη που γεμίζει το μοναδιαίο τετράγωνο στο επίπεδο μπορεί να θεωρηθεί ότι αντιστοιχεί σε μία συνεχή και επί συνάρτηση της μορφής r^-> :[0,1] → [0,1]^2. Γενικότερα, μία καμπύλη που γεμίζει τον μοναδιαίο υπερκύβο, αντιστοιχεί σε μία συνεχή και επί συνάρτησης μορφής r^-> : [0,1] → [0,1]^n

Στην Επιστήμη των Υπολογιστών, όπου τα δεδομένα μπορεί είναι ν-διάστατα στοιχεία διακριτά και απτά, με διακριτές αρχή, μέση και τέλος, όπου δεν υπάρχουν οι αφαιρετικές έννοιες του απείρου, των άρρητων αριθμών με δεκαδικά ψηφία που τείνουν στο άπειρο, παρά μόνο κάποιες απτές προσεγγίσεις και στρογγυλοποιήσεις, οι καμπύλες πλήρωσης χώρου, μοιάζουν περισσότερο σαν μία τεχνική παρά σαν ένα θεώρημα. Βέβαια, ισχύει στο ακέραιο ο επίσημος μαθηματικός ορισμός των καμπυλών πλήρωσης χώρου που αποτελεί σταθμό για την ανάπτυξη

πολλών κατασκευών τέτοιων καμπυλών από πολλούς ερευνητές και επιστήμονες στη σύγχρονη εποχή.