



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

**Υλοποίηση και Ανάλυση Χωρικών Ερωτήσεων στο MariaDB
και Spatial MariaDB**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Μουρατίδη Γεώργιου

Επιβλέπουσα : Αν. Καθηγήτρια Ακριβή Βλάχου, Πανεπιστήμιο Αιγαίου

Μέλη εξεταστικής επιτροπής: κος Κωστούλας, κα Κωνσταντίνου

Σάμος, [Σεπτέμβριος 2023]

Πρόλογος και ευχαριστίες

Ευχαριστώ την επιβλέπουσα καθηγήτρια κυρία Βλάχου Ακριβή για την καθοδήγηση που μου παρείχε κατά τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Θα ήθελα επίσης να ευχαριστήσω την οικογένεια μου για την στήριξη που μου παρείχε σε όλη την διάρκεια των σπουδών μου.

© 2023

του

ΜΟΥΡΑΤΙΔΗ ΓΕΩΡΓΙΟΥ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πίνακας περιεχομένων

1	Εισαγωγή	1
1.1	Εισαγωγή.....	1
1.2	Αντικείμενο διπλωματικής.....	1
1.3	Στόχος της διπλωματικής.....	2
1.4	Δομή της διπλωματικής	3
2	Ορισμός Προβλήματος	4
2.1	Χωρικά Δεδομένα	4
2.1.1	<i>Τι είναι τα χωρικά δεδομένα και γιατί είναι σημαντικά</i>	<i>4</i>
2.1.2	<i>Βασικές Χωρικές ερωτήσεις</i>	<i>5</i>
2.1.3	<i>Τεχνικές ή ευρετήρια για χωρικά δεδομένα</i>	<i>5</i>
2.2	Εντολή Explain	6
3	Σύστημα διαχείρισης βάσεων δεδομένων MariaDB.....	8
3.1	MariaDB	8
3.2	Spatial MariaDB	9
4	Ανασκόπηση Υπάρχουσας Βιβλιογραφίας.....	12
4.1	Σχετικές Εργασίες	12
4.1.1	<i>Efficient Processing of Spatial Joins using R-Trees</i>	<i>12</i>
4.1.2	<i>Αναζήτηση σε χωρό-κειμενικά δεδομένα</i>	<i>13</i>
4.1.3	<i>Collective Spatial Keyword Querying</i>	<i>13</i>
4.1.4	<i>Keyword Search in Spatial Databases: Towards Searching by Document</i>	<i>14</i>
4.1.5	<i>Efficient Processing of Top-k Spatial Keyword Queries</i>	<i>15</i>
4.1.6	<i>Spatial Keyword Query Processing: An Experimental Evaluation</i>	<i>15</i>
5	Μέθοδος για την Επίλυση του Προβλήματος.....	17
5.1	Εισαγωγή.....	17
5.2	Σχήμα Βάσης Δεδομένων	18
5.3	Ερωτήσεις	19
5.4	Διαφορετικές Τεχνικές Ευρετηριοποίησης	20
6	Πειραματική Μελέτη.....	22
6.1	Εισαγωγή.....	22
6.2	Περιγραφή Υλοποίησης.....	22
6.3	Ανάλυση Πειράματος.....	24

6.4	Συμπεράσματα Πειράματος	28
7	Συμπεράσματα	29
	Βιβλιογραφία.....	31

Λίστα Σχημάτων

Εικόνα 1: Διανυσματικοί τύποι δεδομένων.....	27
--	----

Λίστα Πινάκων

Πίνακας 1: Αποτελέσματα για $w = 5$	24
Πίνακας 2: Αποτελέσματα για $w = 10$	25
Πίνακας 3: Αποτελέσματα για $w = 20$	25
Πίνακας 4: Αποτελέσματα για $w = 30$	26
Πίνακας 5: Αποτελέσματα για $w = 20$ και $k = 5$	26
Πίνακας 6: Αποτελέσματα για $w = 20$ και $k = 10$	27
Πίνακας 7: Αποτελέσματα για $w = 20$ και $k = 20$	27

Ακρωνύμια

Χωρικά δεδομένα	Δεδομένα που προσδιορίζονται στον χώρο
Big Data	Δεδομένα μεγάλου όγκου
CSV	Comma Separated Values
Χωρο-κειμενικά δεδομένα	Δεδομένα που προσδιορίζονται στον χώρο και περιέχουν επιπλέον πληροφορία
Γεωγραφικές συντεταγμένες	Στοιχεία που χρησιμοποιούνται για την περιγραφή ενός αντικειμένου στον χώρο
Longitude	Γεωγραφικό μήκος
Latitude	Γεωγραφικό πλάτος
Ερώτημα	Βασικό εργαλείο των συστημάτων βάσεων δεδομένων
OGC	Open Geospatial Consortium
dataset	Συλλογή δεδομένων

Περίληψη

Η σημερινή σύγχρονη εποχή διακρίνεται από την συνεχή άνθιση του όγκου πληροφοριών και δεδομένων που αποθηκεύονται σε δομές αποθήκευσης και προσφέρονται στους χρήστες. Αυτή η αύξηση οφείλεται στην ψηφιοποίηση της πληροφορίας πριν και μετά τη δημιουργία του διαδικτύου. Ένας μεγάλος όγκος πληροφοριών αφορά τα γεωγραφικά δεδομένα της γης, γνωστά ως χωρικά δεδομένα. Η διπλωματική εργασία επιδιώκει την ανάπτυξη μεθόδων για τη διαχείριση, αποθήκευση και ανάκτηση χωρικών δεδομένων, ενσωματώνοντας χωρικές ερωτήσεις και κείμενο. Το αρχείο που χρησιμοποιείται αναφέρεται σε λέξεις-κλειδιά (keywords) και έχουν σημαντικό ρόλο στην αναζήτηση δεδομένων, προσδίδοντας επιπλέον χαρακτηριστικά σε αυτά. Η έννοια της διαχείρισης δεδομένων καλύπτει την αποθήκευση, εισαγωγή, τροποποίηση και αναζήτηση δεδομένων. Στη διπλωματική αυτή παρουσιάζονται μέθοδοι που επιτυγχάνουν αυτούς τους στόχους με αποτελεσματικό και αποδοτικό τρόπο.

Λέξεις Κλειδιά: χωρικά δεδομένα, χωρική ανάλυση, χωρικά ερωτήματα, γεωγραφικές συντεταγμένες, τύπος ερωτημάτων, μεγάλα δεδομένα

Abstract

The current modern era is characterized by the continuously increasing volume of information and data stored in storage structures and made available to users. The main reason for this growth is the digitization of information before and after the creation of the internet. A substantial amount of information pertains to geographical coordinates of the Earth, known as spatial data. This thesis aims to develop methods for the management, storage, and retrieval of spatial data, incorporating spatial queries and text. The texts used refer to keywords and play a significant role in data search, adding additional characteristics to them. The concept of data management encompasses storage, deletion, insertion, modification, and data retrieval. This thesis presents methods that achieve these goals efficiently and effectively.

Keywords: spatial data, spatial analysis, spatial queries, geographic coordinates, query type, big data

1

Εισαγωγή

1.1 Εισαγωγή

Η παρούσα διπλωματική αναφέρεται σε μια μελέτη που εξετάζει την εφαρμογή και την ανάλυση ερωτήσεων που σχετίζονται με τη γεωγραφική τοποθεσία σε βάσεις δεδομένων. Σε αυτήν την περίπτωση, το MariaDB και το Spatial MariaDB αποτελούν τα εργαλεία και τις πλατφόρμες με τις οποίες γίνεται η εφαρμογή αυτής της χωρικής ανάλυσης. Οι "χωρικές ερωτήσεις" αναφέρονται σε ερωτήματα που συνδέονται με τη γεωγραφική πληροφορία, όπως η απόσταση μεταξύ τοποθεσιών, η εντός-εκτός περιοχής κατάσταση ενός σημείου, και οι σχέσεις μεταξύ χωρικών αντικειμένων. Η "υλοποίηση" αφορά την πρακτική εφαρμογή αυτών των ερωτήσεων και λειτουργιών μέσω της δημιουργίας βάσεων δεδομένων και της ανάπτυξης κώδικα που επιτρέπει τη διαχείριση και την ανάλυση γεωγραφικών πληροφοριών.

Η "ανάλυση" αποσκοπεί στην εξαγωγή εργαλείων και γνώσης από τα δεδομένα, προκειμένου να ενισχυθεί η κατανόηση των χωρικών σχέσεων και να προκύψουν ενδιαφέροντα συμπεράσματα. Συνολικά, ο τίτλος αναδεικνύει τη σημασία της ενσωμάτωσης της γεωγραφικής πληροφορίας σε μια βάση δεδομένων, με τη χρήση των εργαλείων MariaDB και Spatial MariaDB, προκειμένου να επιτευχθεί αποτελεσματική διαχείριση και ανάλυση των χωρικών πτυχών των δεδομένων.

1.2 Αντικείμενο διπλωματικής

Η υλοποίηση και η ανάλυση χωρικών ερωτήσεων αναδεικνύουν έναν σημαντικό τομέα στον κόσμο των βάσεων δεδομένων, επιτρέποντας την αποτελεσματική διαχείριση και αξιοποίηση γεωγραφικών πληροφοριών. Οι χωρικές ερωτήσεις αναφέρονται σε ερωτήματα που απαιτούν την ανάλυση στοιχείων με βάση την χωρική τους τοποθεσία, προσφέροντας πλούσια εικόνα και κατανόηση του περιβάλλοντος. Η παρούσα εργασία εστιάζει στην ενδυνάμωση της δυναμικής των βάσεων δεδομένων MariaDB και Spatial MariaDB στον τομέα των χωρικών ερωτήσεων. Εξερευνώντας τις δυνατότητες που παρέχονται από τις χωρικές επεκτάσεις αυτών των συστημάτων, αποσκοπούμε στην εξαγωγή ενδιαφέροντος αναλύσεων και συμπερασμάτων από τα γεωγραφικά δεδομένα.

Στο πλαίσιο αυτό, αναλύουμε ποικίλες χωρικές ερωτήσεις που μπορούν να εκτελεστούν με τη χρήση των χωρικών λειτουργιών που παρέχονται. Αναπτύσσουμε και παρουσιάζουμε παραδείγματα ερωτήσεων που αναλύουν την απόσταση μεταξύ σημείων, την περιοχή εντός-εκτός πολυγώνων και τις σχέσεις γειτονίας μεταξύ χωρικών στοιχείων. Μέσω αυτής της εργασίας, αναδεικνύουμε τη σημασία της αποτελεσματικής διαχείρισης των χωρικών δεδομένων και προσφέρουμε μια περιεκτική εισαγωγή στην υλοποίηση και ανάλυση χωρικών ερωτήσεων μέσω των συστημάτων βάσεων δεδομένων MariaDB και Spatial MariaDB.

1.3 Στόχος της διπλωματικής

Ο στόχος της εργασίας είναι να διερευνήσει, να υλοποιήσει και να αναλύσει την εφαρμογή χωρικών ερωτήσεων στο πλαίσιο των συστημάτων βάσεων δεδομένων MariaDB και Spatial MariaDB.

Οι κύριοι στόχοι περιλαμβάνουν:

- **Υλοποίηση Χωρικών Λειτουργιών:** Η εργασία αποσκοπεί στην ενσωμάτωση των χωρικών επεκτάσεων και λειτουργιών που παρέχουν το MariaDB και το Spatial MariaDB. Αυτό περιλαμβάνει την δημιουργία πινάκων με χωρικά δεδομένα και την εισαγωγή δεδομένων με γεωγραφική σημασία.
- **Ανάλυση Χωρικών Ερωτήσεων:** Η εργασία αναλύει τη δυνατότητα εκτέλεσης διάφορων ειδών χωρικών ερωτήσεων, όπως εύρεση πλησιέστερου γείτονα, εύρεση στοιχείων εντός ενός πολυγώνου, υπολογισμό αποστάσεων κ.ά. Εφαρμογή σε Πρακτικά Σενάρια: Μέσω πρακτικών παραδειγμάτων και σεναρίων, η εργασία αποδεικνύει πώς οι χωρικές ερωτήσεις μπορούν να χρησιμοποιηθούν σε πραγματικές περιπτώσεις, όπως διαχείριση γεωγραφικών δεδομένων καταστημάτων, παρακολούθηση κίνησης, ανάλυση εδαφών κ.ά.
- **Συνεισφορά στη Γνώση:** Η εργασία αποσκοπεί να προσφέρει νέα εισαγωγή και κατανόηση στην ανάλυση χωρικών δεδομένων μέσω της χρήσης των συστημάτων MariaDB και Spatial MariaDB. Με την ανάδειξη των δυνατοτήτων των βάσεων δεδομένων σε αυτόν τον τομέα, προσφέρει νέες προοπτικές για την ανάλυση γεωγραφικών πληροφοριών.

Συνολικά, ο στόχος της εργασίας είναι να επιδείξει τη σημασία και την ευελιξία της ανάλυσης χωρικών ερωτήσεων μέσω των συστημάτων βάσεων δεδομένων MariaDB και Spatial MariaDB

και να προσφέρει αναλυτική κατανόηση για την αξιοποίηση της γεωγραφικής πληροφορίας σε διάφορους τομείς.

1.4 Δομή της διπλωματικής

Αρχικά μελετήθηκαν κάποια κείμενα που σχετίζονται με ερευνητικές προσπάθειες σχετικές με τον τρόπο λειτουργίας των χωρο-κειμενικών δεδομένων.

Ένα ενδιαφέρον Paper το οποίο περιγράφει τον τρόπο λειτουργίας ενός συγκεκριμένου τύπου ευρετηρίου χωρο-κειμενικών δεδομένων είναι η εργασία με τίτλο “A framework for efficient spatial web object retrieval” των Dingming Wu, Gao Cong και Christian S. Jensen. Στην παραπάνω μελέτη προτείνεται μια τεχνική ευρετηρίασης δεδομένων για χωρο-κειμενικά ερωτήματα κ πλησιέστερων γειτόνων αξιοποιώντας την δεντρική δομή δεδομένων R-tree για την ευρετηρίαση χωρικών δεδομένων και την τεχνική των ανεστραμμένων κειμένων για την ευρετηρίαση κι επεξεργασία της κειμενικής πληροφορίας.

Ο συνδυασμός των δύο αυτών τεχνικών οδηγεί στην υλοποίηση μιας μεθόδου ευρετηρίασης χωρο-κειμενικών δεδομένων με την χρήση της δομής δεδομένων που προκύπτει η οποία ονομάζεται IR-tree. Στη συνέχεια αυτού του ερευνητικού κειμένου γίνεται προσπάθεια

βελτιστοποίησης της προτεινόμενης δομής δεδομένων μέσω μεθόδων ενσωμάτωσης κειμενικών πληροφοριών κατά την διαδικασία υπολογισμού των Minimum Bounding Rectangles(MBR) που χρησιμοποιούνται στην κατασκευή της τελικής μορφής του IR-tree αλλά και μέσω μεθόδων ομαδοποίησης των εγγράφων που συσχετίζονται με τα χωρικά αντικείμενα.

Έτσι, προκύπτουν 4 τελικές δομές δεδομένων των οποίων η αποτελεσματικότητα μελετάται με βάση τις επιδόσεις τους σε συγκεκριμένους τύπους χωρο-κειμενικών επερωτήσεων που καθορίζονται από τους συγγραφείς.

2

Ορισμός Προβλήματος

2.1 Χωρικά Δεδομένα

2.1.1 Τι είναι τα χωρικά δεδομένα και γιατί είναι σημαντικά

Τα χωρικά δεδομένα παρέχουν μοναδικά πλεονεκτήματα στην ανάλυση γεγονότων και φαινομένων, καθώς επιτρέπουν την εξέτασή τους με βάση τη γεωγραφική τους διάσταση. Με τη χρήση χωρικών δεδομένων, μπορούμε να ανιχνεύσουμε τοπικές τάσεις και αλληλεπιδράσεις που διαφοροποιούνται ανάλογα με την περιοχή.

Επιπλέον αναφέρονται σε δεδομένα που περιλαμβάνουν πληροφορίες σχετικά με τη γεωγραφική θέση των αντικειμένων. Αυτά τα δεδομένα αποτελούνται από αντικείμενα που περιγράφουν σημεία, γραμμές, περιοχές και περιλαμβάνουν επιπλέον πληροφορίες, όπως ο χρόνος. Κρίσιμα στοιχεία για την αναπαράσταση των χωρικών δεδομένων είναι οι γεωγραφικές συντεταγμένες σημείων και διανυσμάτων. Το σημείο αποτελεί το βασικό στοιχείο, καθώς καθορίζει τη θέση ενός αντικειμένου στον χάρτη. Επιπλέον, δύο ή περισσότερα σημεία μπορούν να σχηματίσουν ένα διάνυσμα, όπως συμβαίνει με ένα κτήριο.

Τέλος, η σύνθεση πολλαπλών διανυσμάτων δημιουργεί τα πολύγωνα, όπως τα γεωγραφικά όρια μιας πόλης στον χάρτη. Ένας τρόπος που μπορούμε να διαχειριστούμε τα χωρικά δεδομένα είναι να τα αποθηκεύουμε και να τα επεξεργαζόμαστε με βάση συγκεκριμένα κριτήρια. Αυτό συμπεριλαμβάνει την εκτέλεση ερωτημάτων σε αυτά και την εξαγωγή αποτελεσμάτων όπως σημεία ή συλλογές σημείων. Τα ερωτήματα αναζητούν και απομακρύνουν τα χωρικά δεδομένα

βάσει γεωγραφικών περιοχών, αποστάσεων, λέξεων-κλειδιών ή κριτηρίων βαθμολόγησης. Παρακάτω παραθέτουμε κάποια παραδείγματα χωρικών ερωτημάτων:

Εμφάνισε όλα τα ξενοδοχεία στην πόλη Κοζάνη.

Εμφάνισε όλα τα ξενοδοχεία που βρίσκονται σε απόσταση 3 χιλιομέτρων.

2.1.2 Βασικές Χωρικές ερωτήσεις

Στο πεδίο των χωρικών δεδομένων, ένα αντικείμενο μπορεί να παριστάνει ένα σημείο, μια γραμμή, ένα πολύγωνο, ή ακόμη και να αποτελείται από πολλές γραμμές ή πολύγωνα. Η γεωμετρία αποτελείται από ένα ή περισσότερα στοιχεία, τα οποία δεν είναι αναγκαία του ίδιου τύπου. Για παράδειγμα, μπορεί να υπάρχει γεωμετρία που περιλαμβάνει σημεία και πολύγωνα ταυτόχρονα. Οι ερωτήσεις που ανατίθενται σε χωρικές βάσεις δεδομένων συνήθως επικεντρώνονται στα χωρικά χαρακτηριστικά των δεδομένων, όπως η θέση τους στο χώρο. Μερικές συνηθισμένες χωρικές ερωτήσεις περιλαμβάνουν:

- Ερώτηση για σημείο (point query): Κατά τη διάρκεια αυτής της ερώτησης, δίνεται ένα σημείο pt και αναζητούνται όλα τα αντικείμενα που περιλαμβάνουν το σημείο pt.
- Ερώτηση περιοχής (range query): Κατά τη διάρκεια αυτής της ερώτησης, δίνεται μια ακτίνα r και αναζητούνται όλα τα αντικείμενα που εντάσσονται εντός της ακτίνας r.
- Ερώτηση κατεύθυνσης (direction query): Κατά τη διάρκεια αυτής της ερώτησης, δίνεται ένα αντικείμενο o και μια κατεύθυνση R (όπως "βόρεια" ή "αριστερά") και αναζητούνται όλα τα αντικείμενα που βρίσκονται στην κατεύθυνση R σε σχέση με το o.
- Ερώτηση κοντινότερου γείτονα (Nearest Neighbor Query): Κατά τη διάρκεια αυτής της ερώτησης, δίνεται ένα αντικείμενο o και αναζητείται το κοντινότερο αντικείμενο προς το o.

Για την αποτελεσματική διαχείριση των χωρικών δεδομένων και των χωρικών σχέσεων, έχουν προταθεί διάφορες επεκτάσεις στα μοντέλα δεδομένων, τις γλώσσες ερωταποκρίσεων και τις δομές δεδομένων κατά την τελευταία δεκαετία.

2.1.3 Τεχνικές ή ευρετήρια για χωρικά δεδομένα

Μια σημαντική κατηγορία γεωμετρικών τελεστών που απαιτεί εξειδικευμένη υποστήριξη στον πραγματικό κόσμο είναι οι χωρικοί τελεστές αναζήτησης. Στην ανάκτηση και ενημέρωση των χωρικών δεδομένων, η τοποθεσία των αντικειμένων παίζει συχνά κρίσιμο ρόλο, παράλληλα με τα αλφαριθμητικά χαρακτηριστικά τους. Οι ερωτήσεις ανάκτησης σε χωρικές βάσεις δεδομένων απαιτούν συχνά γρήγορη εκτέλεση είτε ως ερωτήσεις σημείων είτε ως ερωτήσεις περιοχών. Για παράδειγμα, σε έναν k-διάστατο ευκλείδειο χώρο, όπου αποθηκεύονται γεωμετρικά αντικείμενα, μια ερώτηση περιοχής υπολογίζει τα αντικείμενα που τέμνουν ένα συγκεκριμένο πεδίο

αναζήτησης. Αντίστοιχα, μια ερώτηση σημείου εντοπίζει τα αντικείμενα που περιέχουν ένα συγκεκριμένο σημείο. Και οι δύο αυτές λειτουργίες απαιτούν γρήγορη πρόσβαση στα αντικείμενα της βάσης που εμπλέκονται.

Για τη διαχείριση χωρικών δεδομένων, χρειάζονται ειδικοί τύποι δεδομένων και ευρετηρίων. Συνήθως χρησιμοποιούνται πολυδιάστατα δέντρα για να αντιμετωπιστεί η πολυπλοκότητα αυτή. Οι μέθοδοι προσπέλασης πολυδιάστατων δεδομένων χωρίζονται συνήθως σε δύο κατηγορίες: μέθοδοι προσπέλασης σημείων και μέθοδοι προσπέλασης χωρικών δεδομένων. Οι μέθοδοι προσπέλασης σημείων αντιμετωπίζουν την αναζήτηση ενός σημείου σε δεδομένα σημείων, ενώ οι μέθοδοι προσπέλασης χωρικών δεδομένων ασχολούνται με αντικείμενα που έχουν χωρικές ιδιότητες όπως εμβαδόν και σχήμα, πέραν της τοποθεσίας τους.

Για να υποστηριχθούν οι λειτουργίες αυτές, χρειάζονται εξειδικευμένες μέθοδοι προσπέλασης πολυδιάστατων δεδομένων. Μεταξύ αυτών, βρίσκονται τα "ιεραρχικά δέντρα", που είναι δομές δεδομένων που οργανώνουν τα δεδομένα με τρόπο που διευκολύνει την αναζήτηση και πρόσβαση. Ανάλογα με την υλοποίηση, αυτά τα δέντρα περιλαμβάνουν quadtree, k-D tree, oct-tree και πολλά άλλα.

Συνοψίζοντας, οι χωρικοί τελεστές αναζήτησης αποτελούν ένα σημαντικό πεδίο στη διαχείριση γεωμετρικών δεδομένων. Μέσω εξειδικευμένων δομών και μεθόδων, μπορούμε να αντιμετωπίσουμε αποτελεσματικά τις προκλήσεις που προκύπτουν από την ανάκτηση, ενημέρωση και διαχείριση χωρικών δεδομένων σε πολυδιάστατους χώρους.

2.2 Εντολή Explain

Η εντολή EXPLAIN στη γλώσσα SQL χρησιμοποιείται για να αναλύσει τον σχέδιο εκτέλεσης μιας SQL εντολής. Παρέχει λεπτομερείς πληροφορίες σχετικά με τον τρόπο με τον οποίο ο βάσης δεδομένων σχεδιάζει και εκτελεί την εντολή, συμπεριλαμβανομένων των βημάτων που ακολουθούνται, των ευρετηρίων που χρησιμοποιούνται, των πινάκων που συμμετέχουν και άλλων σχετικών πληροφοριών.

Το κύριο πλεονέκτημα της χρήσης της εντολής EXPLAIN είναι ότι μας βοηθά να βελτιστοποιήσουμε τις ερωτήσεις μας SQL. Μας παρέχει μια καθαρή εικόνα του τρόπου με τον οποίο ο διαχειριστής της βάσης δεδομένων επεξεργάζεται τα αιτήματά μας και μπορεί να αποκαλύψει πιθανά προβλήματα απόδοσης και δυνητικές βελτιώσεις. Για να χρησιμοποιήσουμε την εντολή EXPLAIN, απλά προσθέτουμε το πρόθεμα "EXPLAIN" μπροστά από την SQL εντολή την οποία θέλουμε να αναλύσουμε. Μας επιστρέφει το σχέδιο εκτέλεσης που έχει σχεδιάσει ο διαχειριστής, μας δίνοντας πολύτιμες πληροφορίες για το πώς μπορούμε να βελτιώσουμε τις ερωτήσεις μας SQL.

EXPLAIN SELECT column1, column2 FROM table WHERE condition;

Μετά την εκτέλεση αυτής της εντολής, η βάση δεδομένων θα επιστρέψει μια λεπτομερή ανάλυση του σχεδίου εκτέλεσης για το συγκεκριμένο SQL ερώτημα. Οι πληροφορίες που επιστρέφονται μπορεί να ποικίλουν ανάλογα με το σύστημα διαχείρισης βάσεων δεδομένων που χρησιμοποιείτε, αλλά συνήθως περιλαμβάνουν πληροφορίες όπως ο τύπος εντολής, οι πίνακες που εμπλέκονται, τα ευρετήρια που χρησιμοποιούνται, οι τρόποι συνδυασμού δεδομένων και άλλα. Ανάλογα με τις πληροφορίες που επιστρέφονται, μπορούμε να εντοπίσουμε πιθανά προβλήματα απόδοσης, να βελτιστοποιούμε τις ερωτήσεις μας και να λάβουμε αποφάσεις που θα βελτιώσουν την απόδοση της βάσης δεδομένων μας.

Επίσης εντολή έχει βελτιωθεί σημαντικά με την πάροδο των ετών, προσθέτοντας νέες ικανότητες που καθιστούν τη χρήση της αποτελεσματικότερη. Αρχικά, η έξοδος της ήταν απλώς ένα κείμενο που περιέγραφε τα βασικά βήματα εκτέλεσης της εντολής. Με τον καιρό προστέθηκαν περισσότερες λεπτομέρειες όπως ο τύπος πρόσβασης, ο αλγόριθμος ταξινόμησης και άλλα στοιχεία που διευκολύνουν την ανάλυση.

Στη συνέχεια, η EXPLAIN άρχισε να υποστηρίζει και άλλους τύπους εξόδου όπως XML και JSON, βελτιώνοντας τη χρησιμότητά της. Σήμερα, μπορεί να αναλύει όχι μόνο απλές ερωτήσεις αλλά και σύνθετες δομές όπως συναρτήσεις.

Επιπλέον, παρέχει οπτικοποίηση των αποτελεσμάτων για ευκολότερη κατανόηση. Με αυτόν τον τρόπο, η EXPLAIN έχει εξελιχθεί σε ένα ισχυρό εργαλείο βελτιστοποίησης απόδοσης συγκρινόμενο με τις ανάγκες της κάθε εποχής.

3

Σύστημα διαχείρισης βάσεων δεδομένων MariaDB

3.1 MariaDB

Η MariaDB είναι μια ανοικτού κώδικα σχεσιακή βάση δεδομένων που αναπτύχθηκε ως διακλάδωση του συστήματος διαχείρισης βάσεων δεδομένων MySQL. Αναπτύχθηκε από τον Michael "Monty" Widenius, ένα από τους αρχικούς δημιουργούς του MySQL, ως μια εναλλακτική λύση με σκοπό να διατηρήσει την ανοικτότητα και την ελευθερία του κώδικα μετά την εξαγορά της MySQL AB από την Oracle Corporation. Ορισμένες από τις βασικές χαρακτηριστικές πληροφορίες για την MariaDB περιλαμβάνουν:

- **Ανοικτός Κώδικας:** Η MariaDB είναι πλήρως ανοικτού κώδικα, οπότε ο πηγαίος κώδικάς της είναι προσβάσιμος και μπορεί να τροποποιηθεί από την κοινότητα. Αυτό προάγει τη διαφάνεια, την ασφάλεια και τη συνεργασία.
- **Διακλάδωση του MySQL:** Η MariaDB δημιουργήθηκε ως διακλάδωση του MySQL με στόχο να παραμείνει συμβατή με τον κώδικα και τη σύνταξη του MySQL. Αυτό σημαίνει ότι οι εφαρμογές που αναπτύσσονται για το MySQL μπορούν συνήθως να εκτελεστούν χωρίς αλλαγές στη MariaDB.
- **Βελτιώσεις και Επεκτάσεις:** Η MariaDB προσφέρει πολλές βελτιώσεις και επεκτάσεις σε σχέση με την τελευταία έκδοση του MySQL που υπήρχε κατά την διακλάδωση. Αυτές οι επεκτάσεις περιλαμβάνουν νέα χαρακτηριστικά, βελτιωμένη απόδοση και επιπλέον υποστήριξη για διάφορες μηχανές αποθήκευσης.

- Συμβατότητα: Η MariaDB είναι συμβατή με τις περισσότερες εφαρμογές και βιβλιοθήκες που υποστηρίζουν MySQL. Αυτό σημαίνει ότι μπορείτε να μεταβείτε από το MySQL στη MariaDB χωρίς μεγάλες αλλαγές στον κώδικα σας.
- Βελτιστοποίηση Απόδοσης: Η MariaDB περιλαμβάνει βελτιώσεις στον κώδικα και στον τρόπο λειτουργίας που βελτιώνουν την απόδοση των ερωτημάτων σας και τη χειριστικότητα της βάσης δεδομένων.
- Ποικιλία Μηχανών Αποθήκευσης: Η MariaDB υποστηρίζει διάφορες μηχανές αποθήκευσης, όπως InnoDB (η προεπιλεγμένη), Aria, TokuDB, κ.α. Αυτό επιτρέπει στους χρήστες να επιλέξουν την κατάλληλη μηχανή αποθήκευσης για τις ανάγκες τους.
- Ενεργή Κοινότητα: Η MariaDB έχει μια ενεργή κοινότητα προγραμματιστών, συντηρητών και χρηστών που συνεισφέρουν στη συνεχή ανάπτυξη και βελτίωση του συστήματος.

Συνολικά, η MariaDB προσφέρει μια ισχυρή, αξιόπιστη και ανοικτού κώδικα λύση για σχεσιακές βάσεις δεδομένων με πολλά επιπλέον χαρακτηριστικά και βελτιώσεις σε σύγκριση με το MySQL.

3.2 Spatial MariaDB

Η MariaDB η οποία υποστηρίζει τους γεωμετρικούς τύπους δεδομένων που περιγράφονται στο γεωμετρικό μοντέλο που προτείνει ο διεθνής οργανισμός OGC. Θα διαχωρίσουμε τους τύπους σε δύο κατηγορίες. Στην πρώτη κατηγορία ανήκουν όσοι από αυτούς έχουν την δυνατότητα αποθήκευσης ενός χωρικού αντικειμένου και στην δεύτερη κατηγορία όσοι από αυτούς έχουν την δυνατότητα να αποθηκεύσουν πολλαπλά χωρικά αντικείμενα.

Τύποι με δυνατότητα αποθήκευσης μιας τιμής

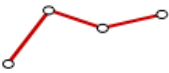
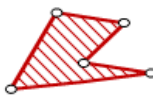
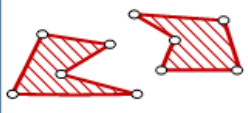
- Geometry : Ο τύπος αυτός αντιστοιχεί στην κλάση geometry που είναι η αρχική κλάση του γεωμετρικού μοντέλου της ιεραρχίας κλάσεων του OGC. Δεν μπορούμε να δημιουργήσουμε στιγμιότυπο αυτής της κλάσης απλά αποτελεί υπερ-κλάση των υπολοίπων στις οποίες κληρονομεί τις μεθόδους και ιδιότητες της. Επίσης όλες οι υπο-κλάσεις της που μπορούν να δημιουργήσουν στιγμιότυπα περιορίζονται σε αντικείμενα μηδενικών, ενός ή δύο διαστάσεων. Η κλάση αυτή έχει κάποιες βασικές ιδιότητες :
- Το σύστημα συντεταγμένων στο οποίο αναφέρεται το κάθε αντικείμενο (Spatial Reference Identifier, SRID), πρόκειται για ένα ακέραιο αριθμό που συνοδεύει κάθε τύπο δεδομένων για να το συσχετίζει με το Σύστημα Συντεταγμένων.
- Τις συντεταγμένες του αντικειμένου. Κάθε μη-μηδενικό αντικείμενο έχει τουλάχιστον ένα ζεύγος συντεταγμένων που αντιστοιχούν στο σύστημα .
- Τον χώρο που καταλαμβάνει, το όριο του και τον εξωτερικό χώρο. Κάθε αντικείμενο έχει κάποια θέση στον χώρο. Εξωτερικό χώρο εννοούμε όλο τον χώρο που δεν καταλαμβάνεται από το αντικείμενο.

- Το ελάχιστο τετράγωνο του που αποτελείται από ένα αντικείμενο που σχηματίζεται από τις ελάχιστες και μέγιστες συντεταγμένες του. ((MINX MINY, MAXX MINY, MAXX MAXY, MINX MAXY, MINX MINY))
- Το αν το αντικείμενο είναι απλό ή όχι. Κάθε μια από τις υπό-κλάσεις που υποστηρίζει αυτή την ιδιότητα την χειρίζεται και διαφορετικά όπως θα δούμε παρακάτω.
- Εάν το αντικείμενο είναι κλειστό ή όχι *isClosed*. Εάν το αντικείμενο είναι κενό ή όχι. Ένα αντικείμενο θεωρείται κενό εάν δεν έχει κανένα σημείο.
- Η διάσταση του που μπορεί να έχει τιμές -1 για κενό αντικείμενο, 0 για αντικείμενο χωρίς μήκος και εμβαδό, 1 για αντικείμενο με μήκος αλλά μηδενικό εμβαδό και τέλος 2 για αντικείμενο με μη-μηδενικό εμβαδό.
- **Point** : Αποθηκεύει γεωμετρικά σημεία. Τα αντικείμενα αυτά όπως γνωρίζουμε είναι μηδενικών διαστάσεων και αποτελούνται από ένα ζευγάρι συντεταγμένων. Η αναπαράσταση κειμένου που αντιστοιχεί σε ένα σημείο με τετμημένη και τεταγμένη ένα είναι `Point(1 1)`.
- **Linestring** : Πρόκειται για τον τύπο που χρησιμοποιείται για την αποθήκευση όλων των τύπων γραμμών. Είναι μιας διάστασης και η αναπαράσταση κειμένου που αντιστοιχεί σε μία γραμμή αποτελούμενη από δύο συνεχόμενα ευθύγραμμο τμήματα `Linestring(1 1,4 5,7 8)`.
- **Polygon** : Περιλαμβάνει όλων των ειδών τα πολύγωνα. Πρόκειται για αντικείμενα δύο διαστάσεων. Η παρακάτω αναπαράσταση αντιστοιχεί σε ένα πολύγωνο το οποίο διαθέτει ένα εξωτερικό και ένα εσωτερικό δακτύλιο που διαχωρίζονται μεταξύ τους με κόμμα. Πάντοτε όταν πρόκειται για πολύγωνο με δύο δακτυλίους αναφερόμαστε πρώτα στον εξωτερικό. `Polygon((1 1,5 1,5 5,1 5,1 1),(2 2,3 2,3 3,2 3,2 2))`.

Τύποι με δυνατότητα αποθήκευσης πολλαπλών τιμών

- **Multipoint** : Πολλαπλά γεωμετρικά σημεία
- **Multilinestring** : Πολλαπλές γραμμές
- **Multipolygon** : Διάφορα πολύγωνα
- **Geometrycollection** : Αυτός ο τύπος δεδομένων αποτελεί την μοναδική εξαίρεση διότι μπορούμε να αποθηκεύσουμε πολλαπλά δεδομένα ίδιου ή και διαφορετικών τύπων . Η μορφή της αναπαράστασης κειμένου σε αυτούς τους τύπους δεδομένων φαίνεται στα παραδείγματα που ακολουθούν.

`MultiPoint(1 1,3 3,5 5)` Αποτελείται από τρία διαφορετικά σημεία τα οποία διαχωρίζονται με κόμμα. `MultiLinestring((1 1,4 5),(1 2,3 4))` Αποτελείται από δύο ξεχωριστές γραμμές η κάθε μία εκ των οποίων αντιστοιχεί σε ένα μόνο ευθύγραμμο τμήμα. `MultiPolygon(((1 3,3 1,3 3,1 3,1 1)),((2 6,4 6,6 6,6 4,2 6)))` Αυτή η κλάση παρουσιάζει κάποια ιδιαιτερότητα και για να γίνει πιο κατανοητή πήραμε την πιο απλή περίπτωση στην οποία έχουμε δύο ξεχωριστά πολύγωνα τα οποία αποτελούνται από ένα εξωτερικό δακτύλιο μόνο. `Geometry Collection(Linestring(1 1,2 2),Point(4 4))` Ο τύπος αυτός περιλαμβάνει δύο ξεχωριστά γεωμετρικά δεδομένα διαφορετικού τύπου που αποτελούνται από ένα ευθύγραμμο τμήμα και ένα σημείο. Παρακάτω γίνεται και η οπτικοποίηση τους.

POINT		MULTIPOINT		Tree, Pole, Hydrant, Valve
LINESTRING		MULTILINESTRING		Road, River, Railway, Pipeline
POLYGON		MULTIPOLYGON		Cadastre, Park, Administrative Boundary
COLLECTION				Graphics, Markups

Εικόνα 1: Διανυσματικοί τύποι δεδομένων

Είδαμε λοιπόν ότι η MariaDB σε ό,τι αφορά τους τύπους δεδομένων ακολουθεί το μοντέλο του οργανισμού OGC χωρίς διαφοροποιήσεις.

1 https://sqlmentalist.files.wordpress.com/2012/06/image_thumb7.png?w=600&h=223

4

Ανασκόπηση Υπάρχουσας Βιβλιογραφίας

4.1 Σχετικές Εργασίες

Όπως έχει ήδη αναφερθεί, η πρόκληση της αναζήτησης χωρο-κειμενικών δεδομένων έχει επισκιάσει την επιστημονική κοινότητα τα τελευταία χρόνια, μετά την εισαγωγή των τεχνολογιών εντοπισμού τοποθεσίας και την ευρεία διάδοση των smartphones. Αποτέλεσμα αυτής της εξέλιξης είναι η ανάδυση διαφόρων μελετών, οι οποίες έχουν ως στόχο να προτείνουν αποδοτικές λύσεις, καθώς και η ανάπτυξη διαφόρων εφαρμογών που εφαρμόζουν αυτές τις λύσεις. Σε αυτήν την ενότητα, θα αναλύσουμε σύντομα μερικές από αυτές τις μελέτες, αναδεικνύοντας τις κύριες προσεγγίσεις και αποτελέσματα που έχουν προκύψει.

4.1.1 Efficient Processing of Spatial Joins using R-Trees

Η μελέτη από τους Tanin, Ooi και Papadias (2005) ασχολείται με την ανάπτυξη και αξιολόγηση ενός αλγορίθμου που χρησιμοποιεί τη δομή των R-trees για την αποτελεσματική επεξεργασία χωρο-κειμενικών αναζητήσεων σε βάσεις δεδομένων. Η μελέτη αναγνωρίζει τη σημασία της αποδοτικής αναζήτησης σε χωρο-κειμενικά δεδομένα, καθώς αυτά χρησιμοποιούνται σε πολλές εφαρμογές όπως γεωγραφικές πληροφοριακές συστάσεις, υπηρεσίες τοποθεσίας και πολλές άλλες. Στη μελέτη αυτή, παρουσιάζεται ένας αλγόριθμος που εκμεταλλεύεται την δομή των R-trees για την επεξεργασία των χωρο-κειμενικών αναζητήσεων. Ο αλγόριθμος αυτός αποδεικνύεται

ότι είναι αποδοτικός και αποτελεσματικός σε σχέση με άλλες μεθόδους. Κατά την εκτέλεση των πειραματικών αξιολογήσεων, τα αποτελέσματα επιβεβαιώνουν την απόδοση και την αποτελεσματικότητα του αλγορίθμου σε όρους του χρόνου εκτέλεσης και του κόστους χώρου. Συγκεκριμένα, ο αλγόριθμος αυτός δίνει τη δυνατότητα για την αναζήτηση και ανάκτηση χωρο-κειμενικών δεδομένων με υψηλή απόδοση, επιτρέποντας την γρήγορη απόκριση σε ερωτήματα που σχετίζονται με τη θέση και το κείμενο των δεδομένων.

Συνολικά, το paper αυτό αποτελεί μια σημαντική συνεισφορά στον τομέα των χωρο-κειμενικών βάσεων δεδομένων, παρέχοντας έναν αποδοτικό τρόπο αναζήτησης και ανάκτησης δεδομένων με χωρο-κειμενικά χαρακτηριστικά. Ο αλγόριθμος που προτείνεται αποτελεί μια εξαιρετική επιλογή για εφαρμογές που απαιτούν γρήγορη και αποδοτική επεξεργασία χωρο-κειμενικών αναζητήσεων.

4.1.2 Αναζήτηση σε χωρό-κειμενικά δεδομένα

Η μελέτη που ξεχωρίζει είναι αυτή του Felipe και συνεργατών (2008), η οποία ασχολείται εκτενώς με το ζήτημα της αναζήτησης χωρο-κειμενικών δεδομένων. Αυτή η μελέτη αποτελεί έναν από τους πιο δημοφιλείς πόρους στον τομέα, με τον στόχο να παρουσιάσει μια νέα προσέγγιση που ενισχύει την αποδοτικότητα της αναζήτησης χωρο-κειμενικών δεδομένων. Κύριες συνεισφορές αυτής της μελέτης περιλαμβάνουν: Ο καθορισμός του προβλήματος της top-k χωρο-κειμενικής αναζήτησης, εστιάζοντας στην εξαγωγή των κορυφαίων αποτελεσμάτων. Η πρόταση μιας αποδοτικής δομής για την αποθήκευση χωρικών και κειμενικών δεδομένων, η οποία υποστηρίζει την αποτελεσματική αναζήτηση.

Η παρουσίαση ενός αλγορίθμου που χρησιμοποιεί τη δομή αυτή για να απαντήσει αποτελεσματικά σε ερωτήσεις top-k χωρο-κειμενικής αναζήτησης. Η μέθοδος που προτείνεται βασίζεται στη δομή R-Tree, αλλά επεκτείνεται και ονομάζεται "Information Retrieval R-Tree". Αυτή η δομή περιλαμβάνει μια υπογραφή σε κάθε κόμβο του δέντρου, που βοηθά στην αποτελεσματική αναζήτηση των λέξεων-κλειδιών.

Μια τυπική δομή "Information Retrieval R-Tree" περιλαμβάνει τριάδες (ObjPtr, A, S), όπου το ObjPtr και το A ακολουθούν την δομή του R-Tree και το S αναπαριστά την υπογραφή του αντικειμένου. Οι διεργασίες εισαγωγής και διαγραφής κόμβων διεκπεραιώνονται στη δομή αυτή, εξασφαλίζοντας έτσι την χαμηλή πολυπλοκότητα του αλγορίθμου. Τα πειραματικά αποτελέσματα που παρουσιάζονται στο τέλος της μελέτης επιβεβαιώνουν την αποδοτικότητα του προτεινόμενου αλγορίθμου.

4.1.3 Collective Spatial Keyword Querying

Οι συγγραφείς της μελέτης (Cao et al, 2011) αναφέρουν ότι, μέχρι εκείνη την στιγμή, τα ερωτήματα που είχαν εξεταστεί επικεντρώνονταν κυρίως στην ανεύρεση μεμονωμένων αντικειμένων που ικανοποιούσαν ένα ερώτημα, αλλά όχι στην αναζήτηση ομάδων αντικειμένων που θα μπορούσαν να απαντήσουν συλλογικά σε ένα ερώτημα. Οι συγγραφείς ορίζουν ένα νέο

πρόβλημα, την ανάκτηση μιας ομάδας χωρικών αντικειμένων, όπου οι λέξεις-κλειδιά της ομάδας πρέπει να ικανοποιούν τις λέξεις-κλειδιά του ερωτήματος, ενώ τα αντικείμενα να βρίσκονται κοντά στη θέση του χρήστη, με ελάχιστες αποστάσεις μεταξύ τους. Πιο συγκεκριμένα, ορίζονται δύο περιπτώσεις του ερωτήματος:

Στην πρώτη περίπτωση, το στόχο είναι να βρεθεί μια ομάδα αντικειμένων που να ικανοποιεί τις λέξεις-κλειδιά του ερωτήματος, έτσι ώστε το άθροισμα των χωρικών αποστάσεων μεταξύ των αντικειμένων να είναι ελάχιστο.

Στη δεύτερη περίπτωση, ο στόχος είναι να βρεθεί μια ομάδα αντικειμένων που να ικανοποιεί τις λέξεις-κλειδιά του ερωτήματος, έτσι ώστε να ελαχιστοποιείται το άθροισμα της μέγιστης απόστασης μεταξύ των αντικειμένων και του ερωτήματος, καθώς και η μέγιστη απόσταση μεταξύ δύο αντικειμένων στην ομάδα.

Για την κατάταξη των δεδομένων, προτείνεται η χρήση της δομής IR-Tree, η οποία είναι μια παραλλαγή του R-Tree που χρησιμοποιεί ανεστραμμένα αρχεία. Κάθε κόμβος της IR-Tree περιέχει καταχωρήσεις της μορφής (ο, ο.λ, ο.di), ενώ το λεξιλόγιο των διακριτών λέξεων και οι ανεστραμμένοι δείκτες αρχείων βοηθούν στην αποτελεσματική αναζήτηση. Επίσης, παρουσιάζονται δύο συναρτήσεις για τον υπολογισμό του κόστους της αναζήτησης στην IR-Tree. Η μελέτη αναλύει την εφαρμογή των παραπάνω αλγορίθμων σε διάφορα σύνολα δεδομένων και παρουσιάζει τα σχετικά αποτελέσματα.

4.1.4 Keyword Search in Spatial Databases: Towards Searching by Document

Η μελέτη (Zhang et al, 2009) επικεντρώνεται στον τομέα της αναζήτησης χωρο-κειμενικών δεδομένων, και έχει ως στόχο να παρουσιάσει το ερώτημα m-closest keywords (mCK) query. Αυτό το ερώτημα απευθύνεται σε μια βάση δεδομένων με χωρικά αντικείμενα, όπου κάθε αντικείμενο συσχετίζεται με περιγραφικές πληροφορίες που αντιστοιχούν σε λέξεις-κλειδιά. Το ερώτημα mCK αποσκοπεί στην εύρεση των πλησιέστερων χωρικών αντικειμένων που ταιριάζουν με τις λέξεις-κλειδιά που ορίζει ο χρήστης. Αυτό το είδος ερωτήματος, το mCK query, μπορεί να εφαρμοστεί ευρέως σε χωρικές βάσεις δεδομένων.

Ένα παράδειγμα εφαρμογής είναι η προσθήκη γεωγραφικών ετικετών σε έγγραφα, συγκρίνοντας τα λέξεις-κλειδιά με άλλα έγγραφα που έχουν γεωγραφικές ετικέτες σε μια βάση δεδομένων. Για την αποτελεσματική αντιμετώπιση των ερωτημάτων mCK, η μελέτη παρουσιάζει τη δομή bR*-Tree, που αποτελεί επέκταση του R*-Tree. Αυτή η δομή επιτρέπει την αποθήκευση και αναζήτηση των λέξεων-κλειδίων και των αντίστοιχων χωρικών πληροφοριών με αποτελεσματικό τρόπο.

Οι κύριες συνεισφορές της μελέτης περιλαμβάνουν: Πρόταση ενός νέου spatial keyword query, του mCK query, με πολλές δυνατές εφαρμογές σε χωρικές βάσεις δεδομένων. Εισαγωγή της δομής bR*-Tree ως νέο τρόπο αναζήτησης για αποτελεσματική αντιμετώπιση των ερωτημάτων mCK.

Παρουσίαση αποτελεσματικών στρατηγικών αναζήτησης που μειώνουν τον απαιτούμενο χώρο αναζήτησης. Ορισμός δύο περιορισμών μονοτονικότητας (distance mutex και keyword mutex) για

διαγραφή κόμβων στο δέντρο. Διεξοδική πειραματική αξιολόγηση που επιδεικνύει την αποτελεσματικότητα και επεκτασιμότητα του προτεινόμενου αλγορίθμου.

Συνολικά, η μελέτη αναδεικνύει τη σημαντικότητα του ερωτήματος mCK query και παρουσιάζει μια ολοκληρωμένη προσέγγιση για την αποτελεσματική αντιμετώπισή του.

4.1.5 Efficient Processing of Top-k Spatial Keyword Queries

Η μελέτη (Joao et al, 2011) παρουσιάζει μια νέα προσέγγιση για τη βελτίωση της απόδοσης των ερωτημάτων top-k χωρο-κειμενικών αναζητήσεων, εισάγοντας τη δομή Spatial Inverted Index (S2I). Ο σκοπός της μελέτης είναι να προτείνει αυτόν τον νέο τρόπο ανάθεσης δεικτών, που μπορεί να βελτιώσει την απόδοση των ερωτημάτων που αφορούν την ταξινόμηση και ανάκτηση των πλησιέστερων αντικειμένων βάσει των χωρικών και κειμενικών πληροφοριών.

Η προτεινόμενη δομή, S2I, αντιστοιχεί κάθε όρο σε ένα σύνολο αντικειμένων που περιέχουν αυτόν τον όρο. Αντί να αποθηκεύει τους όρους σε μια απλή ανεστραμμένη δομή, το S2I χρησιμοποιεί συγκεντρωτικό R-Tree (aR-Tree) και αρχεία για να αποθηκεύει τους όρους ανάλογα με τη συχνότητα τους και τη χωρική εγγύτητα των αντικειμένων. Οι κύριες συνεισφορές της μελέτης περιλαμβάνουν: Προτείνεται ο νέος τρόπος ανάθεσης δεικτών, S2I, που χαρτογραφεί και αποθηκεύει τους όρους με βάση τη συχνότητα και τη χωρική εγγύτητα. Περιγράφονται αλγόριθμοι που εκμεταλλεύονται το S2I για την αποτελεσματική επεξεργασία ερωτημάτων χωρο-κειμενικών αναζητήσεων.

Παρουσιάζεται εκτενής πειραματική αξιολόγηση που δείχνει ότι η προτεινόμενη προσέγγιση υπερβαίνει τους παραδοσιακούς αλγορίθμους ως προς τον χρόνο ενημέρωσης, το κόστος και τον χρόνο απόκρισης.

4.1.6 Spatial Keyword Query Processing: An Experimental Evaluation

Η μελέτη (Chen et al, 2013) επικεντρώνεται στην αξιολόγηση των υπαρχόντων μεθόδων αναζήτησης σε χωρο-κειμενικά δεδομένα. Ο σκοπός της μελέτης είναι να διερευνήσει τις διαφορετικές τεχνικές που χρησιμοποιούνται για την εκτέλεση αναζητήσεων σε τέτοιου είδους δεδομένα και να παρουσιάσει αποτελέσματα σύγκρισης ανάμεσά τους. Η μελέτη επισημαίνει τα αποτελέσματα που προέκυψαν κατά τη διάρκεια της σύγκρισης, παρέχοντας νέες πληροφορίες που μπορούν να βοηθήσουν στην επιλογή της κατάλληλης μεθόδου αναζήτησης σε χωρο-κειμενικά δεδομένα. Η μελέτη κατατάσσει τις διάφορες τεχνικές αναζήτησης βάσει των ειδών των ερωτημάτων που μπορούν να χρησιμοποιηθούν:

Boolean kNN Query: Ανακτήστε τα k αντικείμενα που βρίσκονται πλησιέστερα στην τρέχουσα τοποθεσία του χρήστη, με την κειμενική περιγραφή τους να περιέχει τις λέξεις-κλειδιά "πίτσα και πορτοκαλάδα".

Top-k kNN Query: Ανάκτηση των k αντικειμένων με την υψηλότερη βαθμολογία, συνδυάζοντας την απόστασή τους από την τοποθεσία του ερωτήματος και τη συνάφεια της κειμενικής περιγραφής τους με τις λέξεις-κλειδιά "πίτσα και καπουτσίνο".

Boolean Range Query: Ανακτήστε όλα τα αντικείμενα που έχουν κειμενική περιγραφή που περιέχει τις λέξεις-κλειδιά "πίτσα και καπουτσίνο" και βρίσκονται σε απόσταση τριών χιλιομέτρων από την τοποθεσία του ερωτήματος. Οι συγγραφείς της μελέτης κατατάσσουν τις τεχνικές αναζήτησης βάσει του τρόπου με τον οποίο πραγματοποιείται η δεικτοδότηση των λέξεων-κλειδιών, καθώς και της αναζήτησης. Η σύγκριση πραγματοποιείται βάσει του χρόνου αναζήτησης και του κόστους σε χώρο.

Επιπλέον, η μελέτη περιλαμβάνει πειράματα για διαφορετικούς όγκους δεδομένων, προσφέροντας έτσι μια αντικειμενική αξιολόγηση των τεχνικών αναζήτησης και βοηθώντας στην επιλογή της καταλληλότερης μεθόδου για την πραγματοποίηση αναζητήσεων σε χωρο-κειμενικά δεδομένα.

5

Μέθοδος για την Επίλυση του Προβλήματος

5.1 Εισαγωγή

Στο παρόν κεφάλαιο, θα εξετάσουμε τα κριτήρια που καθορίζουν την χρησιμότητα και την επιτυχία της υλοποίησης των ιδεών που παρουσιάζονται στην παρούσα εργασία. Θα περιγράψουμε τα κριτήρια αξιολόγησης της εφαρμογής, όπως η ικανότητά της να μετατρέπει χωρικά δεδομένα κειμένου σε δισδιάστατες συντεταγμένες και μονοδιάστατα χωρικά δεδομένα στον χώρο. Ένα από τα κριτήρια αξιολόγησης της εφαρμογής είναι η ικανότητα μετατροπής των χωρικών δεδομένων κειμένου σε δισδιάστατες συντεταγμένες. Αυτό σημαίνει ότι η εφαρμογή πρέπει να μπορεί να αναγνωρίζει χωρικές πληροφορίες από το κείμενο και να τις αντιστοιχίζει σε συντεταγμένες που χρησιμοποιούν οι χάρτες. Για να αξιολογήσουμε την αποτελεσματικότητα της μετατροπής των δεδομένων κειμένου σε δισδιάστατα δεδομένα με χρήση ευρετηρίασης, είναι απαραίτητο να χρησιμοποιήσουμε συγκεκριμένα σύνολα δεδομένων και να προσδιορίσουμε λεπτομερώς τα χωρικά ερωτήματα που παίζουν σημαντικό ρόλο στην εργασία αυτή.

Ο τρόπος με τον οποίο θα προδιαγράψουμε τα χωρικά ερωτήματα εξαρτάται από τον συγκεκριμένο τύπο των δεδομένων και την εφαρμογή που χρησιμοποιείται. Είναι σημαντικό να αναλύσουμε τα δεδομένα μας και να προσδιορίσουμε ποιες χωρικές πληροφορίες μπορούν να εξαχθούν από αυτά. Για παράδειγμα, αν ασχολούμαστε με ένα σύνολο δεδομένων που περιέχει διευθύνσεις, μπορούμε να εξάγουμε χωρικές πληροφορίες, όπως οι γεωγραφικές συντεταγμένες των διευθύνσεων. Έτσι, θα μπορούσαμε να εφαρμόσουμε την ευρετηρίαση για να μετατρέψουμε αυτές τις διευθύνσεις σε δισδιάστατα δεδομένα, κατά προτίμηση σε έναν επιφανειακό χάρτη. Επιπλέον, η λεπτομερής προδιαγραφή των χωρικών ερωτημάτων είναι ουσιώδης για την αποτελεσματική μετατροπή των δεδομένων. Αυτό σημαίνει ότι πρέπει να καθορίσουμε ποιες

χωρικές πληροφορίες αναζητούμε ακριβώς και πώς θα τις αντιστοιχίσουμε στις δισδιάστατες συντεταγμένες. Αυτή η προδιαγραφή θα βοηθήσει στην αποφυγή ασάφειας και στην αποτελεσματική ανάπτυξη της εφαρμογής. Συνοψίζοντας, για να αξιολογήσουμε την αποτελεσματικότητα της μετατροπής των δεδομένων σε δισδιάστατα δεδομένα με ευρετηρίαση, πρέπει να ορίσουμε σαφώς τα χωρικά ερωτήματα που πρέπει να απαντηθούν και να εξετάσουμε πώς η εφαρμογή μας μπορεί να μετατρέψει τα δεδομένα σε δισδιάστατες συντεταγμένες που να χρησιμοποιούν ευρετηρίαση. Στην παρούσα εργασία, χρησιμοποιούμε ένα ειδικό είδος ερωτήματος για χωρικά κείμενα, που ονομάζεται "spatio-textual range query". Αυτό το είδος αλγορίθμου αρχικά αναπτύχθηκε για την επεξεργασία χωρικών ερωτημάτων και την αναζήτηση σημείων εντός ενός ακτίνας r από ένα αρχικό σημείο.

Στην παρούσα εργασία, θα χρησιμοποιήσουμε αυτό το είδος αλγορίθμου για την αναζήτηση δισδιάστατων χωρο-κειμενικών αντικειμένων. Τα ερωτήματα εύρους στα χωρικά δεδομένα κειμένου έχουν τους ίδιους περιορισμούς με τα κλασικά ερωτήματα εύρους στα χωρικά δεδομένα, αλλά επιπλέον περιλαμβάνουν περιορισμούς σχετικούς με το κείμενο. Στην παρούσα εργασία, θα αξιολογήσουμε την ακρίβεια των αποτελεσμάτων που προκύπτουν από τον αλγόριθμο που χρησιμοποιούμε για την απάντηση σε αυτά τα ερωτήματα. Επιπλέον, όπως θα αναλυθεί παρακάτω, οι δισδιάστατοι δείκτες (χωρικής και κειμενικής) χωρικών δεδομένων κειμένου με διάσπαση των αντικείμενων κάποιου συνόλου δεδομένων στις ομάδες. Για να επιλέξουμε τα υποψήφια αντικείμενα που μπορεί να ανήκουν στα αντίστοιχα αποτελέσματα ενός ερωτήματος, αναζητούμε κάθε αντικείμενο την ομάδα στην οποία ανήκει και επιλέγουμε όσο το δυνατόν λιγότερα αντικείμενα που περιλαμβάνουν το σύνολο των αντικειμένων.

Ο σκοπός είναι να βρούμε υποψήφια αντικείμενα που θα ανταποκρίνονται στα χαρακτηριστικά του ερωτήματος και να αξιολογήσουμε την ικανότητα του αλγορίθμου μας να επιλέγει όσο το δυνατόν λιγότερα αντικείμενα. Για να ελέγξουμε αν ένα υποψήφιο αντικείμενο ανήκει στο τελικό αποτέλεσμα, χρησιμοποιούμε τον αλγόριθμο brute-force που ελέγχει κάθε στοιχείο στο σύνολο δεδομένων.

5.2 Σχήμα Βάσης Δεδομένων

Η βάση δεδομένων αποτελείται από 4 πίνακες εκ των οποίων όλοι έχουν τις ίδιες στήλες όπως φαίνεται στον παρακάτω πίνακα.

ID	NAME	NUMBER_OF_REVIEWS	PRICE	LATITUDE	LONGITUDE
----	------	-------------------	-------	----------	-----------

Με την διαφορά ότι ο καθένας έχει διαφορετικά κλειδιά και τεχνική ευρετηριοποίησης όπως επίσης οι τρεις πρώτοι πίνακες έχουν «απλά» δεδομένα, ενώ ο τέταρτος αντί για τις στήλες LATITUDE, LONGITUDE έχει μία στήλη COORDINATES που είναι τύπου Point όπως φαίνεται παρακάτω.

ID	NAME	NUMBER_OF_REVIEWS	PRICE	COORDINATES
----	------	-------------------	-------	-------------

Ο πρώτος πίνακας δεν έχει κανένα κλειδί και καμία τεχνική ευρετηριοποίησης.

Ο δεύτερος πίνακας έχει κλειδί την στήλη ID και ευρετήριο στην στήλη PRICE.

Ο τρίτος πίνακας έχει κλειδί τις στήλες LATITUDE και LONGITUDE όπως και ευρετήριο στις στήλες αυτές

Ο τελευταίος πίνακας έχει κλειδί και ευρετήριο στην στήλη του την COORDINATES που είναι τύπου SPATIAL.

5.3 Ερωτήσεις

Η περιγραφή που δίνετε παρακάτω αναφέρεται σε διάφορα εργαλεία και τεχνικές που χρησιμοποιούνται στον προγραμματισμό για την αποτελεσματική επεξεργασία μεγάλων αρχείων δεδομένων στα αρχεία CSV. Η επιλογή μεταξύ FileReader, CSVParser και BufferedReader εξαρτάται από το ποιες ακριβώς λειτουργίες χρειαζόμαστε για να εκτελέσουμε στα δεδομένα μας. Κάθε ένα από αυτά τα εργαλεία έχει διαφορετική χρήση και λειτουργία.

- **FileReader:** Ο FileReader είναι ένας απλός τρόπος διαβάσματος χαρακτήρων από ένα αρχείο. Ωστόσο, είναι περιορισμένος σε βασικές λειτουργίες ανάγνωσης χαρακτήρων και δεν παρέχει ειδικές λειτουργίες για την ανάλυση της δομής των δεδομένων, όπως το CSV.
- **CSVParser:** Ο CSVParser είναι ειδικά σχεδιασμένος για την ανάλυση δεδομένων CSV. Αυτός ο parser μπορεί να διαχειριστεί τη διαδικασία της ανάλυσης των πεδίων, αγνοώντας τα διαχωριστικά και αντιμετωπίζοντας κενά πεδία ή πεδία που περιέχουν ειδικούς χαρακτήρες. Αυτό σας επιτρέπει να επεξεργαστείτε δομημένα δεδομένα όπως πίνακες ή λίστες.
- **BufferedReader:** Ο BufferedReader είναι ένα εργαλείο για την αποτελεσματική ανάγνωση δεδομένων από ένα αρχείο, καθώς χρησιμοποιεί buffer για να μειώσει τον αριθμό των φορτώσεων από τον δίσκο. Αυτό μπορεί να βελτιώσει την απόδοση κατά την ανάγνωση μεγάλων αρχείων.

Συνδυάζοντας τη χρήση της FileReader για τη σειριακή ανάγνωση και τη χρήση του CSVParser με το κατάλληλο CSVFormat, αποφεύγουμε το φόρτωμα όλων των δεδομένων στη μνήμη ταυτόχρονα, και αντ' αυτού να διαβάζουμε τα δεδομένα σειριακά και να τα επεξεργαζόμαστε κατάλληλα. Αυτό μπορεί να είναι εξαιρετικά αποδοτικό για μεγάλα αρχεία δεδομένων, καθώς δεν χρειάζεται να φορτώνουμε όλα τα δεδομένα στη μνήμη μαζί.

Αυτή η προσέγγιση βοηθάει στο να μειωθεί ο χρόνος που απαιτείται για το ανέβασμα και την επεξεργασία μεγάλων όγκων δεδομένων. Όπως φαίνεται και στον παρακάτω πίνακα 1.

Test Run	FileReader	CSVParser	BufferedReader
1000 εγγραφές	0.05 min	0.0003 min	0.07 min
25.463 εγγραφές	3 min	0.4 min	3.4 min

Πίνακας 1.

Για την σύνδεση στην βάση δεδομένων χρησιμοποιήθηκε η κλάση Connection το οποίο είναι ένα interface που δίνει τη δυνατότητα σύνδεσης με μία βάση δεδομένων. Το interface Connection είναι το βασικό σημείο επαφής μεταξύ του κώδικα της εφαρμογής και της βάσης δεδομένων και παρέχει τη δυνατότητα εκτέλεσης επερωτήσεων και λήψης αποτελεσμάτων.

```
String url = "jdbc:mariadb://localhost:3306/hotels";
String username = "user";
String password = "pass";
Connection conn = DriverManager.getConnection(url, username, password);
```

Για το ανέβασμα των δεδομένων στην βάση δεδομένων χρησιμοποιήθηκε μια υποκλάση της κλάσης java.sql.Statement που δίνει τη δυνατότητα εκτέλεσης επερωτήσεων SQL σε μία βάση δεδομένων. Το Prepared Statement είναι ένας προαπαρασκευασμένος (prepared) έλεγχος επερωτήσεων, ο οποίος επιτρέπει την επαναφορά και την εκτέλεση της ίδιας επερώτησης πολλές φορές, με διαφορετικά δεδομένα. Για να δημιουργήσετε ένα Prepared Statement, πρέπει να χρησιμοποιήσετε τη μέθοδο prepareStatement() της κλάσης Connection. Γενικά, η μέθοδος αυτή δέχεται ως παράμετρο ένα String με την επερώτηση SQL και επιστρέφει ένα PreparedStatement object. Με τους παραπάνω μεθόδους έχουμε δημιουργήσει τις απαραίτητες συνθήκες για να διαβάσουμε τα αρχεία και να τα ανεβάσουμε στην βάση δεδομένων με σκοπό την επεξεργασία και την ανάλυση τους.

5.4 Διαφορετικές Τεχνικές Ευρετηριοποίησης

Η MariaDB υποστηρίζει αρκετούς τύπους δεικτοδότησης για τη βελτιστοποίηση ερωτημάτων και τη βελτίωση της απόδοσης των λειτουργιών της βάσης δεδομένων. Στην συγκεκριμένη εργασία χρησιμοποιούμε τους εξής:

- **B-tree Index:** Αυτή είναι η προεπιλεγμένη μέθοδος δεικτοδότησης στη MariaDB. Είναι ένα ισορροπημένο δέντρο που επιτρέπει αποδοτικές λειτουργίες αναζήτησης, εισαγωγής και διαγραφής.
- **Spatial Index:** Οι χωρικοί δείκτες χρησιμοποιούνται για τη βελτιστοποίηση ερωτημάτων χωρικών δεδομένων. Είναι ειδικά σχεδιασμένοι για στήλες που περιέχουν γεωμετρικά δεδομένα (π.χ. σημεία, γραμμές, πολύγωνα). Οι χωρικοί δείκτες χρησιμοποιούν

εξειδικευμένους αλγόριθμους όπως τα R-trees για τη γρήγορη ανάκτηση σχετικών χωρικών δεδομένων.

- Primary Key Index: Ένας δείκτης πρωτεύοντος κλειδιού επιβάλλει τη μοναδικότητα και τους περιορισμούς μη μηδενικότητας στις ευρετηριασμένες στήλες. Κάθε πίνακας μπορεί να έχει μόνο έναν δείκτη πρωτεύοντος κλειδιού.

6

Πειραματική Μελέτη

6.1 Εισαγωγή

Το σύνολο δεδομένων που χρησιμοποιήσαμε για την υλοποίηση της εργασίας μας και τον έλεγχο της ορθότητας των αλγορίθμων που υλοποιήθηκαν αφορά τα δεδομένα κάποιων ξενοδοχείων. Συγκεκριμένα, το dataset περιέχει στοιχεία για περίπου 25463 ξενοδοχεία. Τα στοιχεία που αναφέρονται για κάθε ξενοδοχείο είναι το id του ξενοδοχείου, η τοποθεσία του (συντεταγμένες της θέσης του) οι κριτικές του και η τιμή του.

Έτσι, στο συγκεκριμένο σύνολο δεδομένων η τοποθεσία κάθε ξενοδοχείου αποτελεί την χωρική πληροφορία των δεδομένων και οι υπηρεσίες που παρέχονται από κάθε ξενοδοχείο αποτελούν την κειμενική πληροφορία των δεδομένων.

Οι αλγόριθμοι για την επεξεργασία χωρο-κειμενικών δεδομένων που αναπτύχθηκαν και οι οποίοι θα παρουσιαστούν σε επόμενη ενότητα μπορούν να εφαρμοστούν σε οποιοδήποτε σύνολο δεδομένων που έχει παρόμοια μορφή με αυτό που περιγράψαμε παραπάνω.

6.2 Περιγραφή Υλοποίησης

Ο πίνακας που αποθηκεύει τα χωρικά δεδομένα είναι ένας καίριος παράγοντας για την αποδοτική διαχείριση και ανάκτηση των γεωγραφικών πληροφοριών. Παρακάτω αναλύονται οι τέσσερις προσεγγίσεις αποθήκευσης στο πλαίσιο της βάσης δεδομένων MariaDB και με χρήση της επέκτασης Spatial MariaDB για την υποστήριξη χωρικών δεδομένων.

- Απλός πίνακας: Σε αυτήν την προσέγγιση, δημιουργούμε έναν απλό πίνακα. Οι στήλες Latitude και Longitude δηλώνονται ως απλοί πραγματικοί αριθμοί, χωρίς να υπάρχει ευρετήριο. Αυτή η προσέγγιση είναι απλή, αλλά δεν παρέχει αποτελεσματικό τρόπο ανάκτησης χωρικών ερωτήσεων, καθώς δεν υπάρχουν βελτιστοποιημένοι μηχανισμοί αναζήτησης.
- Πίνακας με ευρετήριο στο πρωτεύον κλειδί: Σε αυτήν την προσέγγιση, χρησιμοποιούμε έναν πίνακα με ευρετήριο στη στήλη που αποτελεί το πρωτεύον κλειδί (π.χ. price). Ωστόσο, οι στήλες Latitude και Longitude παραμένουν απλοί πραγματικοί αριθμοί. Αυτή η προσέγγιση μπορεί να βελτιώσει την ανάκτηση σε σχέση με τον πρώτο τρόπο, αλλά δεν είναι ακόμη βέλτιστη για χωρικές ερωτήσεις.
- Πίνακας με ευρετήρια στο Latitude και Longitude: Σε αυτήν την προσέγγιση, δημιουργούμε έναν πίνακα όπου οι στήλες Latitude και Longitude έχουν ευρετήρια. Αυτό βοηθά στην βελτίωση της ανάκτησης χωρικών ερωτήσεων, καθώς τα ευρετήρια επιτρέπουν πιο αποτελεσματική αναζήτηση.
- Ευρετήρια με βάση το Spatial MariaDB: Σε αυτήν την προσέγγιση, χρησιμοποιούμε τις δυνατότητες της επέκτασης Spatial MariaDB. Οι στήλες Latitude και Longitude δεν αποθηκεύονται ως απλοί αριθμοί, αλλά δηλώνονται ως τύπου point. Η επέκταση Spatial παρέχει βελτιστοποιημένους αλγορίθμους αναζήτησης και ευρετήρια που είναι προσαρμοσμένα για χωρικά δεδομένα, επιτρέποντας γρήγορη και αποτελεσματική ανάκτηση γεωγραφικών πληροφοριών.

Οι διαφορετικές προσεγγίσεις προσφέρουν διαφορετικά επίπεδα αποτελεσματικότητας και απόδοσης στην ανάκτηση χωρικών ερωτήσεων. Η τελική επιλογή εξαρτάται από τις απαιτήσεις της εφαρμογής και το επίπεδο της απόδοσης που απαιτείται για τις χωρικές ερωτήσεις.

Σε επόμενη φάση πραγματοποιούμε τη σύνδεση με τη βάση δεδομένων MariaDB χρησιμοποιώντας την κλάση DriverManager. Η σύνδεση περνάει τις παραμέτρους για τη διεύθυνση του διακομιστή, τη θύρα και τα διαπιστευτήρια χρήστη για την πρόσβαση στη βάση δεδομένων. Αμέσως μετά τη σύνδεση, εκτελείται μια ερώτηση SQL για να βρεθούν οι ελάχιστες και μέγιστες τιμές των γεωγραφικών συντεταγμένων (latitude και longitude) στους πίνακες που έχουμε δημιουργήσει. Οι τιμές αυτές αποθηκεύονται σε μεταβλητές.

Δίνουμε μια παράμετρος w που καθορίζει το μέγεθος του παραθύρου ερωτήσεων σε σχέση με την περιοχή των γεωγραφικών συντεταγμένων. Το πρόγραμμα εκτελεί έναν αριθμό τυχαίων ερωτήσεων, όπου κάθε ερώτηση αναζητά ξενοδοχεία που εμπίπτουν στο παράθυρο ερωτήσεων. Οι τυχαίες τιμές για το x και το y υπολογίζονται με βάση τις ελάχιστες και μέγιστες τιμές των γεωγραφικών συντεταγμένων και τον παράγοντα w.

Για κάθε τυχαία ερώτηση, εκτελείται η αντίστοιχη SQL ερώτηση στη βάση δεδομένων. Οι αποτελέσματα αποθηκεύονται σε ένα αντικείμενο ResultSet, και γίνεται αποθήκευση της συνολικής πλήθους αποτελεσμάτων και της πλήθους αποτελεσμάτων για κάθε ερώτηση.

Ο χρόνος εκτέλεσης καταγράφεται και προστίθεται στον συνολικό χρόνο εκτέλεσης. Μετά την εκτέλεση όλων των τυχαίων ερωτήσεων, υπολογίζονται οι μέσες τιμές για τον χρόνο εκτέλεσης, και το πλήθος αποτελεσμάτων που διαβάστηκαν.

Η μέση τιμή υπολογίζεται διαιρώντας τον συνολικό χρόνο, τον συνολικό αριθμό αποτελεσμάτων και τον συνολικό αριθμό πλειάδων αντίστοιχα με τον αριθμό των τυχαίων ερωτήσεων.

Η μέση τιμή των χρόνων εκτέλεσης και των πλήθους αποτελεσμάτων που διαβάστηκαν εκτυπώνονται στην οθόνη για παρατήρηση και ανάλυση των αποτελεσμάτων.

6.3 Ανάλυση Πειράματος

Με βάση την παραπάνω περιγραφή καταγράφουμε τα παρακάτω αποτελέσματα.

- Για $w = 5$

	Μέσος χρόνος(ms)	Μέση τιμή στο πλήθος αποτελεσμάτων
Χωρίς ευρετήριο	27.2	1949.9
Ευρετήριο στο price	22.6	1949.9
Ευρετήριο στο Lat,Long	11.3	1949.9
Spatial MariaDB	10.2	1949.9

Πίνακας 1: Αποτελέσματα για $w = 5$

- Για $w = 10$

	Μέσος χρόνος(ms)	Μέση τιμή στο πλήθος αποτελεσμάτων
Χωρίς ευρετήριο	43.0	11116.3
Ευρετήριο στο price	36.2	11116.3
Ευρετήριο	26.0	11116.3

στο Lat,Long		
Spatial MariaDB	21.4	11116.3

Πίνακας 2: Αποτελέσματα για $w = 10$

- Για $w = 20$

	Μέσος χρόνος(ms)	Μέση τιμή στο πλήθος αποτελεσμάτων
Χωρίς ευρετήριο	53.4	16332.1
Ευρετήριο στο price	40.5	16332.1
Ευρετήριο στο Lat,Long	37.8	16332.1
Spatial MariaDB	35.6	16332.1

Πίνακας 3: Αποτελέσματα για $w = 20$

- Για $w = 30$

	Μέσος χρόνος(ms)	Μέση τιμή στο πλήθος αποτελεσμάτων
Χωρίς ευρετήριο	66.9	22264.1
Ευρετήριο στο price	59.1	22264.1
Ευρετήριο στο Lat,Long	41.5	22264.1
Spatial MariaDB	39.4	22264.1

Πίνακας 4: Αποτελέσματα για $w = 30$

Αφού έχουμε τελειώσει με τα παραπάνω πειράματα προχωράμε στο επόμενο το οποίο είναι παρόμοιο με το προηγούμενο, όμως εδώ ορίζονται δύο παράμετροι αντί για έναν.

Η παράμετρος w αντιπροσωπεύει το ποσοστό μεγέθους του παραθύρου, ενώ η παράμετρος k αντιπροσωπεύει τον αριθμό των αποτελεσμάτων που θα ανακτηθούν σε κάθε ερώτηση.

- Για $w = 20$, $k = 5$

	Μέσος χρόνος(ms)	Μέση τιμή στο πλήθος αποτελεσμάτων
Χωρίς ευρετήριο	27.9	4.0
Ευρετήριο στο price	12.3	4.0
Ευρετήριο στο	18.7	4.0

Lat,Long		
Spatial MariaDB	9.8	4.0

Πίνακας 5: Αποτελέσματα για $w = 20$, $k = 5$

- Για $w = 20$, $k = 10$

	Μέσος χρόνος(ms)	Μέση τιμή στο πλήθος αποτελεσμάτων
Χωρίς ευρετήριο	29.1	8.0
Ευρετήριο στο price	13.6	8.0
Ευρετήριο στο Lat,Long	19.8	8.0
Spatial MariaDB	11.5	8.0

Πίνακας 6: Αποτελέσματα για $w = 20$, $k = 10$

- Για $w = 20$, $k = 20$

	Μέσος χρόνος(ms)	Μέση τιμή στο πλήθος αποτελεσμάτων
Χωρίς ευρετήριο	30.5	14.0
Ευρετήριο στο price	15.2	14.0
Ευρετήριο στο Lat,Long	21.4	14.0
Spatial MariaDB	15.3	14.0

Πίνακας 7: Αποτελέσματα για $w = 20$, $k = 20$

6.4 Συμπεράσματα Πειράματος

1. Παράμετρος w (Μέγεθος του Παραθύρου):

- Η παράμετρος w καθορίζει το μέγεθος του χωρικού παραθύρου που χρησιμοποιούμε για να καθορίσουμε μια τυχαία περιοχή για το εύρος της ερώτησής σας.
- Αυξάνοντας την τιμή του w , θα αυξήσετε το μέγεθος της εύρους μας, κάνοντας το δυνατότερο να περιλαμβάνει μεγαλύτερη γεωγραφική περιοχή.
- Μειώνοντας την τιμή του w , το παράθυρο θα μικραίνει, περιορίζοντας την γεωγραφική περιοχή που καλύπτεται από την ερώτησή σας.
- Η επιλογή της κατάλληλης τιμής για το w εξαρτάται από τις ανάγκες μας. Αν θέλετε να καλύψετε μεγάλες γεωγραφικές περιοχές, μπορεί να χρειαστεί να αυξήσουμε την τιμή του w .

2. Παράμετρος k (Αριθμός Αποτελεσμάτων):

- Η παράμετρος k καθορίζει πόσα αποτελέσματα θα επιστρέφονται από κάθε ερώτηση και καθορίζει τον αριθμό των αποτελεσμάτων που θα επιστραφούν.
- Αυξάνοντας την τιμή του k , θα λάβουμε περισσότερα αποτελέσματα από κάθε ερώτηση, αλλά αυτό μπορεί να αυξήσει τον χρόνο εκτέλεσης και την απαιτούμενη μνήμη.
- Μειώνοντας την τιμή του k , θα λάβουμε λιγότερα αποτελέσματα, με πιο γρήγορη εκτέλεση.

Επομένως, η παράμετρος w επηρεάζει την περιοχή που καλύπτεται από την ερώτησή μας, ενώ η παράμετρος k επηρεάζει τον αριθμό των επιστρεφόμενων αποτελεσμάτων. Ανάλογα με τις απαιτήσεις μας, πρέπει να επιλέξουμε τις κατάλληλες τιμές για αυτές τις παραμέτρους.

7

Συμπεράσματα

Στην ενότητα αυτή, παρουσιάζονται τα συμπεράσματα που απορρέουν από την ανάλυση των αποτελεσμάτων της πειραματικής μελέτης που περιγράφηκε παραπάνω, καθώς και γενικότερα από την εφαρμογή των προτεινόμενων αλγορίθμων στην παρούσα εργασία. Επιπλέον, παρουσιάζονται προτάσεις για μελλοντικές επεκτάσεις που μπορούν να ενσωματωθούν στην υπάρχουσα υλοποίηση προκειμένου να βελτιστοποιηθεί η λειτουργία της. Αρχικά, πρέπει να τονιστεί ότι οι εκτελέσεις του προτεινόμενου αλγορίθμου παρήγαγαν αποτελέσματα που συμπίπτουν με τα αποτελέσματα του brute-force αλγορίθμου, ο οποίος λειτουργεί απευθείας στο αρχικό σύνολο δεδομένων. Αυτό σημαίνει ότι η ευρετηρίαση των δεδομένων δεν επηρέασε την ακρίβεια των στοιχείων του συνόλου δεδομένων.

Επίσης, η μετατροπή των πολυδιάστατων χωρο-κειμενικών δεδομένων σε δισδιάστατα δεν αποσκοπεί στην απόκρυψη των πληροφοριών που παρέχει το αρχικό dataset, και ο αλγόριθμος χωρο-κειμενικής επερώτησης εύρους λειτουργεί σωστά και σύμφωνα με τις προδιαγραφές.

Μια αξιοσημείωτη παρατήρηση είναι ότι με τον συγκεκριμένο αλγόριθμο επιτυγχάνεται σημαντική μείωση στο πλήθος των δεδομένων που πρέπει να ελεγχθούν για να ανήκουν στο τελικό αποτέλεσμα. Αυτό επιτυγχάνεται μέσω των περιορισμών που εφαρμόζονται σε σχέση με τα spatial και textual partitions, καθώς και μέσω των ορίων των διαστημάτων αναζήτησης αντικειμένων. Αυτή η ενίσχυση της ταχύτητας αναζήτησης είναι σημαντική για την αποτελεσματικότητα του αλγορίθμου στον χειρισμό μεγάλων συνόλων δεδομένων. Επιπλέον, μέσω των διαγραμμάτων που παρουσιάζονται, επιβεβαιώνεται η υπεροχή του προτεινόμενου αλγορίθμου σε σχέση με τον brute-force αλγόριθμο.

Ο μέσος όρος του πλήθους των spatial partitions και των textual partitions στα οποία γίνεται αναζήτηση είναι σημαντικά μικρότερος από το αντίστοιχο πλήθος αντικειμένων που ελέγχονται

από τον brute-force αλγόριθμο. Επίσης, παρατηρείται ότι το πλήθος των υποψήφιων αντικειμένων σε κάθε πείραμα συμβαδίζει με το πλήθος των keywords της επερώτησης.

Τέλος, εισάγονται και ορισμένοι περιορισμοί της υλοποίησης. Η υλοποίηση αυτή αξιοποιεί κάποια από τα ευρήματα για την επιτάχυνση της αναζήτησης των αντικειμένων, κάτι που θα μπορούσε να βελτιώσει την απόδοση του αλγορίθμου. Με διαφορετικές υλοποιήσεις θα μπορούσε να επιταχυνθεί η αναζήτηση των υποψήφιων αντικειμένων. Τέτοιες βελτιώσεις θα μπορούσαν να μελετηθούν σε μελλοντικές επεκτάσεις της υλοποίησης.

Όπως φαίνεται, η διαχείριση και αναζήτηση πληροφοριών σε χωρο-κειμενικά δεδομένα αποτελεί έναν σημαντικό τομέα της Πληροφορικής και, ειδικότερα, της Επιστήμης των Δεδομένων. Το ενδιαφέρον σε αυτόν τον τομέα είναι υψηλό, τόσο σε ερευνητικό επίπεδο όσο και σε εργασιακό επίπεδο.

Η γρήγορη διαχείριση και επεξεργασία χωρο-κειμενικών δεδομένων έχει ένα ιδιαίτερο κοινό, το οποίο αποτελείται τόσο από ερευνητές όσο και από τους εφαρμογείς στην πράξη. Η δυνατότητα αποτελεσματικής διαχείρισης και επεξεργασίας χωρο-κειμενικών δεδομένων αποτελεί σημαντικό επίτευγμα, καθώς βοηθά στην απάντηση ερωτημάτων που συνδυάζουν τον χωρικό και τον κειμενικό προσανατολισμό. Αυτό το επίτευγμα δεν μόνο βοηθά τον απλό χρήστη στην αντιμετώπιση τέτοιων ερωτημάτων, αλλά και οδηγεί στη βελτιστοποίηση και επέκταση πολλών εφαρμογών.

Με την εισαγωγή προηγμένων χαρακτηριστικών και λειτουργιών στις εφαρμογές, επιτυγχάνεται η αύξηση της λειτουργικότητας και η ανάπτυξη πιο εξελιγμένων λύσεων. Είναι, λοιπόν, ζωτικής σημασίας η συνεχής ανάπτυξη και προώθηση των διαφόρων πεδίων που απαρτίζουν αυτόν τον τομέα.

Βιβλιογραφία

- Wu Dingming, Cong Gao & Jensen S. Christian : A framework for efficient spatial web object retrieval (2012), The VLDB Journal-The International Journal on Very Large Data Bases.
- Tsatsanifos George, Vlachou Akrivi : On Processing Top-k Spatio-Textual Preference Queries (2015), EDBT.
- Felipe, Ian & Hristidis, Vagelis & Rish, N.. (2008). Keyword Search on Spatial Databases. Proceedings - International Conference on Data Engineering. 656 - 665. 10.1109/ICDE.2008.4497474.
- Gkorgkas, O., Vlachou, A., Doukeridis, C. and Nørnvåg, K., 2015, March. Finding the Most Diverse Products using Preference Queries. In *EDBT* (pp. 205-216).
- E. Tanin, A. Harwood and H. Samet, "A distributed quadtree index for peer-to-peer settings," *21st International Conference on Data Engineering (ICDE'05)*, Tokyo, Japan, 2005, pp. 254-255, doi: 10.1109/ICDE.2005.7.
- Rigaux, P., Scholl, M. and Voisard, A., 2002. Spatial databases: with application to GIS. Morgan Kaufmann.
- Reese, G., 2000. Database Programming with JDBC and JAVA. " O'Reilly Media, Inc."
- Kenler, E. and Razzoli, F., 2015. *MariaDB Essentials*. Packt Publishing Ltd.
- Bartholomew, D., 2012. Mariadb vs. mysql. Dostopano, 7(10), p.2014.

Lu, H. and Ooi, B.C., 1993. Spatial indexing: Past and future. IEEE Data Eng. Bull., 16(3), pp.16-21.

Mahmud, S.H., Hossin, M.A., Jahan, H., Noori, S.R.H. and Bhuiyan, T., 2018, May. CSV-ANNOTATE: Generate annotated tables from CSV file. In 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 71-75). IEEE.

Keegan, P., Champenois, L., Crawley, G., Hunt, C. and Webster, C., 2005. Netbeans™ ide field guide: developing desktop, web, enterprise, and mobile applications. Prentice Hall Press.

Siahaan, V. and Sianipar, R.H., 2019. Learning MariaDB: A Comprehensive Approach to Developing Database-Driven Applications Using NetBeans. SPARTA PUBLISHING.

