Εθνικο Μετσοβιο Πολυτεχνειο

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΒΑΣΕΩΝ ΓΝΩΣΕΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Μάριου Β. Παπαμιχάλη

Επιβλέπων: Τίμος Σελλής - Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΒΑΣΕΩΝ ΓΝΩΣΕΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Μάριου Β. Παπαμιχάλη

Επιβλέπων: Τίμος Σελλής - Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11η Ιουλίου 2011.

.....	.....	.....
Τιμολέων Σελλής	Νεκτάριος Κοζύρης	Θοδωρής Δαλαμάγκας
Καθηγητής Ε.Μ.Π.	Αναπληρωτής	Ερευνητής Β΄
	Καθηγητής Ε.Μ.Π.	ΙΠΣΥΠ/Ε.Κ.'Αθηνά'

Αθήνα, Ιούλιος 2011.

.....

Μάριος Β. Παπαμιχάλης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός & Μηχανικός Ηλεκτρονικών Υπολογιστών

Περίληψη

Αντικείμενο της διπλωματικής εργασίας είναι η ανάπτυξη και η υλοποίηση αλγορίθμου που θα επιτρέπει online απαντήσεις σε πιθανοτικά ερωτήματα διαρκείας σχετικά με τη θέση μεγάλου αριθμού αβέβαιων θέσεων κινούμενων αντικειμένων. Τα δεδομένα καταφθάνουν με μεγάλο και δυναμικά μεταβλητό ρυθμό. Τα πιθανοτικά ερωτήματα διαρκείας θα τίθενται από διάφορους κινούμενους χρήστες που επιθυμούν να ενημερώνονται οποτεδήποτε στην περιοχή τους συμβαίνει κάποιο έκτακτο γεγονός (π.χ. επίσκεψη φίλου). Τέτοια γεγονότα καταγράφονται σε κεντρικό υπολογιστή αλλά με αβεβαιότητα ως προς την ακριβή γεωγραφική τους θέση. Οι κινούμενες συσκευές διαθέτουν δυνατότητα γεωγραφικού εντοπισμού (GPS) όμως το στίγμα κάθε αντικειμένου ποτέ δεν αποκαλύπτεται στον κεντρικό υπολογιστή. Ωστόσο, θεωρείται γνωστή η ευρύτερη περιοχή του συμβάντος. Η πιθανότητα εκδήλωσης του γεγονότος δεν θεωρείται ομοιόμορφη, αλλά μπορεί να ποικίλλει. Οι χρήστες μπορούν να υποβάλλουν τα χωρικά ερωτήματα διαρκείας για περιοχές ενδιαφέροντός τους, οπότε ο επεξεργαστής οφείλει να συνεκτιμήσει τις σχετικές πιθανότητες των προσφάτως καταγεγραμμένων συμβάντων και να δίνει τακτικά ενημερωμένες προσεγγιστικές απαντήσεις με κυμαινόμενη ποιότητα. Τέτοια στοιχεία θα μπορούσαν να αξιοποιηθούν σε εφαρμογές κοινωνικής δικτύωσης με κινητά τηλέφωνα, εκτίμηση περιβαλλοντικού κινδύνου σε φυσικές καταστροφές (λ.χ. διαρροή πετρελαίου), πρόγνωση μετεωρολογικών φαινομένων (λ.χ. τυφώνες) κ.ά.

Η εργασία επικεντρώνεται κυρίως στην ανάπτυξη τεχνικών δεικτοδότησης και κλαδέματος βάσει των οποίων θα μπορούμε να μειώσουμε το κόστος και τον χρόνο επεξεργασίας των δεδομένων χωρίς να μειωθεί αισθητά η ακρίβεια των αποτελεσμάτων. Ο αλγόριθμος που προτείνεται επιλέχτηκε να είναι προσεγγιστικός και παρέχει μία λύση στο πρόβλημα της αποτίμησης πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων. Με εφαρμογή των παραπάνω τεχνικών,πραγματοποιήθηκαν πειράματα σε συνθετικά δεδομένα πάνω στο οδικό δίκτυο της Αθήνας, από τα οποία προέκυψαν ενθαρρυντικά αποτελέσματα. Επιπροσθέτως, επιβεβαιώθηκαν οι αναμενόμενες επιδόσεις τους σχετικά με τους χρόνους εκτέλεσης και την ακρίβεια των προσεγγιστικών απαντήσεων. Συνολικό συμπέρασμα της εργασίας είναι ότι ο αλγόριθμος που δημιουργήθηκε για αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων είναι κατάλληλος για προβλήματα πραγματικού χρόνου, όπου η ακρίβεια μπορεί να θυσιαστεί για χάρη της γρήγορης απόκρισης.

Λέξεις κλειδιά : αβεβαιότητα, πιθανοτικά ερωτήματα περιοχής, κινούμενα αντικείμενα, ρεύματα δεδομένων.

Abstract

The purpose of this thesis is to develop and implement an algorithm that allows online answers to probabilistic continuous queries concerning uncertain positions of a large number of moving objects. Location data comes with large and dynamically variable rate. The probabilistic queries are submitted by different mobile users who want to be informed if there is an extraordinary event (e.g. a friend's visit) at any time in their spatial region of interest. Such events are recorded in a server with uncertainty regarding their exact geographical location. Mobile devices use GPS but the position of each object is never revealed to the server. The area of the incident is known, although the distribution of the event is not uniform, but may vary. The processor must consider the relative probabilities of newly recorded incidents and regularly provide approximate answers with varying quality. Such data could be exploited in applications of social networking with mobile phones, environmental risk forecasts in natural disasters, prediction of meteorological phenomena (e.g. hurricanes), etc.

This thesis focuses primarily on developing indexing and pruning techniques that can reduce the cost and time of data processing without particular loss on the accuracy of results. The proposed algorithm was chosen to be approximate and provides a solution to the evaluation of probabiblistic region queries for uncertain positions of moving objects. This method yielded promising results in a comprehensive experimental study conducted against systhetic datasets generated using the road network of Athens. In addition, the expected performance on the execution times and accuracy of the approximate answers was confirmed. The overall conclusion of this thesis is that the algorithm developed for evaluating probabilistic region queries for uncertain positions of moving objects is suitable for real-time problems, where some accuracy may be sacrificed for the sake of timely response.

Key words: uncertainty, probabilistic region queries, moving objects, data streams.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα κ. Κώστα Πατρούμπα για την υποστήριξη που μου προσέφερε καθώς και για την προθυμία του να με καθοδηγήσει στις απορίες και στα προβλήματα που συνάντησα. Ακόμα, ευχαριστίες οφείλω στον καθηγητή μου κ.Τίμο Σελλή για το ιδιαίτερο ενδιαφέρον που έδειξε για την εργασία. Τέλος, θα ήθελα να ευχαριστήσω τον αδελφό μου Φάνη για την βοήθεια που μου προσέφερε σε προσωπικό επίπεδο την περασμένη χρονιά.

Μάριος Παπαμιχάλης
Αθήνα, Ιούλιος 2011

Πρόλογος

Η σύγχρονη τεχνολογία παρέχει πολλές δυνατότητες για εποπτεία και παρακολούθηση κινήσεων, φαινομένων, προσώπων κτλ. Παραδείγματα τέτοιων εφαρμογών βρίσκουμε:

- Στον τουρισμό, λ.χ. ένα μουσείο μπορεί να παρέχει στους επισκέπτες του συσκευές, η οποίες έχουν την δυνατότητα να εντοπίζουν τη θέση του και να του παρέχουν οπτικοακουστικό υλικό με πληροφορίες από τα εκθέματα που βρίσκονται κοντά του.
- Στη διαφήμιση, λ.χ. ένας χρήστης μιάς συσκευής όταν περνάει έξω από ένα κατάστημα, αφού εντοπιστεί η θέση του, έχει τη δυνατότητα να λαμβάνει πληροφορίες για τα προϊόντα που πωλούνται, τυχόν εκπτώσεις κτλ.
- Στην προστασία του περιβάλλοντος, λ.χ. η παρακολούθηση της κίνησης των ζώων σε περιπτώσεις φωτιάς σε ένα δάσος.
- Στη διακίνηση εμπορευμάτων, λ.χ. ένας χρήστης μπορεί κάθε στιγμή να δει που βρίσκονται τα προϊόντα που έχει παραγγείλει στο internet κτλ.

Η υπάρχουσα τεχνολογία για εποπτεία οντοτήτων και φαινομένων παρέχει ευρύ πεδίο εφαρμογής στα κινούμενα αντικείμενα. Συλλέγονται γεωγραφικά στίγματα από ανθρώπους, άγρια ζώα, εμπορεύματα κτλ. και σε πραγματικό χρόνο παρέχονται έγκυρες απαντήσεις σχετικά με τη θέση, την κατάσταση και τη συσχέτιση των αντικειμένων. Ωστόσο τα στίγματα δεν είναι σχεδόν ποτέ απολύτως ακριβή για διάφορους λόγους όπως:

- Υπάρχει *εγγενές σφάλμα* λόγω μετρήσεων από τα συστήματα GPS, RFID κτλ.
- Υπάρχει *χρόνος υστέρησης*. Το κινούμενο αντικείμενο έχει μετακινηθεί λόγω της ταχύτητάς και της κατεύθυνσής του από την τελευταία φορά που έστειλε το στίγμα του.
- Προστασία *ιδιωτικότητας προσώπων*. Κανείς χρήστης δεν θέλει να αποκαλύπτει επακριβώς το ακριβές στίγμα του.

Η εργασία στοχεύει στη μελέτη και υλοποίηση μιας εφαρμογής που θα επιτρέπει online απαντήσεις σε *πιθανοτικά ερωτήματα διαρκείας* (*probabilistic continuous queries*) σχετικά με τη θέση μεγάλου αριθμού αβέβαιων θέσεων κινούμενων αντικειμένων. Τα πιθανοτικά ερωτήματα διαρκείας θα τίθενται από διάφορους κινούμενους χρήστες που εγγράφονται στην υπηρεσία και επιθυμούν να ενημερώνονται οποτεδήποτε στην περιοχή τους συμβαίνει κάποιο έκτακτο γεγονός (π.χ. επίσκεψη φίλου). Τέτοια γεγονότα καταγράφονται σε κεντρικό υπολογιστή (server) αλλά με *αβεβαιότητα* ως προς την ακριβή γεωγραφική τους θέση. Οι κινούμενες συσκευές διαθέτουν δυνατότητα γεωγραφικού εντοπισμού (GPS) όμως το στίγμα κάθε αντικειμένου ποτέ δεν αποκαλύπτεται στον κεντρικό υπολογιστή. Ωστόσο, θεωρείται γνωστή η ευρύτερη περιοχή του συμβάντος. Η πιθανότητα εκδήλωσης του γεγονότος δεν θεωρείται ομοιόμορφη, αλλά μπορεί να ποικίλλει. Οι χρήστες έχουν την δυνατότητα να υποβάλλουν τα χωρικά ερωτήματα διαρκείας για περιοχές ενδιαφέροντός τους, οπότε ο επεξεργαστής οφείλει να συνεκτιμήσει τις σχετικές πιθανότητες των προσφάτως καταγεγραμμένων συμβάντων και να δίνει τακτικά ενημερωμένες προσεγγιστικές απαντήσεις με κυμαινόμενη ποιότητα. Τέτοια δεδομένα θα μπορούσαν να αξιοποιηθούν σε εφαρμογές κοινωνικής δικτύωσης με κινητά τηλέφωνα, εκτίμηση περιβαλλοντικού κινδύνου σε φυσικές καταστροφές (λ.χ. διαρροή πετρελαίου), πρόγνωση μετεωρολογικών φαινομένων (λ.χ. τυφώνες) κ.ά. Ως βασική υπόθεση της εργασίας θεωρείται η παρακολούθηση πολλών κινούμενων αντικειμένων και ερωτημάτων τα οποία ανανεώνουν συχνά την περιοχή της θέσης τους και δημιουργούν *ρεύματα δεδομένων* (*data streams*):

- Τα στοιχεία καταφθάνουν σε μεγάλο και ενδεχομένως μεταβλητό ρυθμό σε πραγματικό χρόνο (online).
- Τα ρεύματα έχουν μεγάλο μέγεθος, πιθανόν απεριόριστο.

Η φύση του μοντέλου συνηγορεί σε επιλογή προσεγγιστικών αλγορίθμων που εξοικονομούν κυρίως χρόνο, συνεκτιμώντας το σφάλμα που μπορεί να γίνει αποδεκτό από την εκάστοτε εφαρμογή.

Διάρθρωση της εργασίας

Ο τόμος διαρθρώνεται σε πέντε κεφάλαια. Τα κεφάλαια 1, 2 και 3 παρέχουν θεωρητικές πληροφορίες για τα ρεύματα δεδομένων, τα κινούμενα αντικείμενα και την αβεβαιότητα στα συστήματα διαχείρισης ρευμάτων δεδομένων αντίστοιχα. Το κεφάλαιο 4 παρουσιάζει τον αλγόριθμο για την αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων. Στο κεφάλαιο 5 περιγράφονται τα πειραματικά αποτελέσματα του αλγορίθμου και τέλος στο κεφάλαιο 6 συνοψίζονται τα κυριότερα συμπεράσματα της εργασίας.

Στα κεφάλαια 1, 2 περιγράφεται η έννοια των ρευμάτων κινούμενων αντικειμένων. Αρχικά, τονίζεται η αδυναμία και η ανεπάρκεια των συστημάτων διαχείρισης βάσεων δεδομένων να επεξεργαστούν αποδοτικά δεδομένα που μεταβάλλονται δυναμικά με τον χρόνο. Παρουσιάζονται τα χαρακτηριστικά των ρευμάτων δεδομένων και οι επεκτάσεις που έγιναν πάνω στις συμβατικές βάσεις δεδομένων. Στη συνέχεια, αναφερόμαστε στη διαχείριση χωρικών δεδομένων και τέλος εστιά-

ζουμε στα κινούμενα αντικείμενα. Τα ρεύματα κινούμενων αντικειμένων είναι αυτά που θα μας απασχολήσουν στην ανάπτυξη της εφαρμογής που πραγματεύεται η παρούσα εργασία.

Στο κεφάλαιο 3 περιγράφεται η έννοια της αβεβαιότητας στα συστήματα διαχείρισης βάσεων και ρευμάτων δεδομένων. Το κεφάλαιο αυτό αποτελεί επισκόπηση της έρευνας που έχει γίνει μέχρι σήμερα πάνω στην αβεβαιότητα στα συστήματα διαχείρισης βάσεων δεδομένων και ρευμάτν δεδομένων. Ιδιαίτερη μνεία γίνεται σε διάφορα χαρακτηριστικά της αβεβαιότητας και στην εφαρμογή της πάνω σε ερωτήματα και τεχνικές βελτιστοποίησης.

Στο κεφάλαιο 4 παρουσιάζεται ο πρωτότυπος αλγόριθμος για την αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων. Ο αλγόριθμος αυτός δεν είναι αναλυτικός και δίνει προσεγγιστικές απαντήσεις. Δέχεται ως εισόδους δύο ρεύματα δεδομένων, ένα για τα κινούμενα αντικείμενα και ένα για τα κινούμενα ερωτήματα και δίνει ως έξοδο ένα ρεύμα δεδομένων αποτελούμενο από τα αντικείμενα που δίνονται ως απάντηση. Χαρακτηριστικό κομμάτι της λειτουργίας του είναι ότι βασίζεται πάνω σε τεχνικές κλαδέματος που μειώνουν κατά πολύ τόσο το χρόνο όσο και το κόστος επεξεργασίας.

Στο κεφάλαιο 5, αξιολογείται πειραματικά ο αλγόριθμος και παρουσιάζονται και σχολιάζονται διεξοδικά τα αποτελέσματα.

Τέλος, στο κεφάλαιο 6 εκτίθενται ορισμένα συμπεράσματα και πιθανές μελλοντικές προοπτικές της εργασίας.

Περιεχόμενα

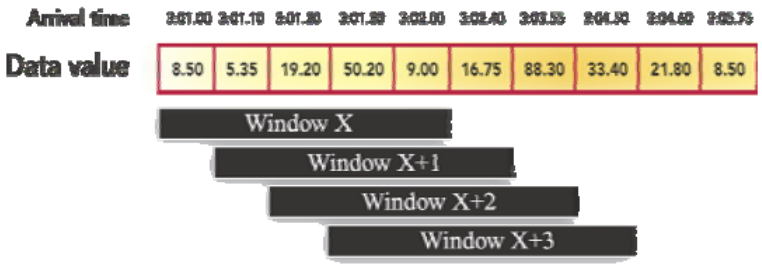
Κεφάλαιο 1

Ρεύματα Δεδομένων

1.1 Εισαγωγή

Τα τελευταία χρόνια προέκυψε η ανάγκη για έρευνα πάνω σε θέματα που προκύπτουν από ένα νέο μοντέλο επεξεργασία δεδομένων. Τα δεδομένα δεν λαμβάνουν τη μορφή στατικών δεδομένων όπως στις βάσεις δεδομένων αλλά αλλάζουν δυναμικά με τρόπο συνεχή, γρήγορο και χρονικά μεταβαλλόμενο με τη μορφή *ρευμάτων δεδομένων* (*data streams*). Η νέα αυτή κατηγορία δεδομένων έχει γίνει ευρέως αναγνωρισμένη και υποστηρίζει πλέον πολλές εφαρμογές ανάμεσα στις οποίες είναι : οικονομικές εφαρμογές, παρακολούθηση δικτύου, ασφάλεια, τηλεπικοινωνίες διαχείριση δεδομένων, διαδικτυακές εφαρμογές, δίκτυα αισθητήρων, κ.ά. Στις εφαρμογές αυτές, τα δεδομένα δεν παραμένουν αμετάβλητα, αλλά μεταβάλλονται με την πάροδο του χρόνου. Π.χ. σε οικονομικές εφαρμογές οι τιμές των μετοχών μεταβάλλονται όταν τελειώνει κάποια αγοροπωλησία ενώ σε ένα δίκτυο αισθητήρων οι τιμές αλλάζουν όταν αλλάζει κάποια θερμοκρασία. Στο μοντέλο ρευμάτων δεδομένων, τα δεδομένα μπορεί να παρουσιάζονται ως σχεσιακές πλειάδες, π.χ., μετρήσεις του δικτύου, τα αρχεία κλήσεων, επισκεψιμότητα ιστοσελίδας, δίκτυα αισθητήρων κτλ. Παρόλα αυτά, η συνεχής άφιξη τους σε πολλαπλά, χρονικά μεταβαλλόμενα, ενδεχομένως απρόβλεπτα και απεριόριστα ρεύματα (σχήμα 1.1) γεννούν νέα ερευνητικά προβλήματα, διαφορετικά από αυτά των βάσεων δεδομένων.

Για όλες τις εφαρμογές που προαναφέρθηκαν, δεν είναι εφικτό να χρησιμοποιήσουμε ένα παραδοσιακό σύστημα διαχείρισης βάσεων δεδομένων (ΔΒΜΣ) και να δουλέψουμε με αποτελεσματικότητα εκεί. Τα παραδοσιακά ΣΔΒΔ δεν είναι σχεδιασμένα για την ταχεία και συνεχή φόρτωση των συγκεκριμένων δεδομένων. Δεν υποστηρίζουν άμεσα τα ερωτήματα διαρκείας (continuous queries), τα οποία είναι χαρακτηριστικά των εφαρμογών των ρευμάτων δεδομένων. Επιπροσθέτως, τα παραδοσιακά ΣΔΒΔ σε μεγάλο βαθμό επικεντρώνονται σε ακριβείς απαντήσεις οι οποίες υπολογίζονται από σταθερά πλάνα ερωτημάτων. Αντίθετα ουσιαστικό χαρακτηριστικό των εφαρμογών αυτών αποτελούν οι προσεγγιστικές και οι προσαρμοστικές απαντήσεις στην εκτέλεση ερωτημάτων πάνω σε ρεύματα δε-



Σχήμα 1.1: Παράδειγμα ρεύματος δεδομένων.

δομένων.

Στο κεφάλαιο αυτό, παρουσιάζεται το μοντέλο των ρευμάτων δεδομένων. Αρχικά αναπτύσσονται οι λόγοι της ύπαρξης συστημάτων ρευμάτων δεδομένων και στη συνέχεια γίνεται μία επισκόπηση πάνω στα χαρακτηριστικά τους (μοντέλο, τελεστές), που έχουν παρουσιαστεί σε διάφορες εργασίες. Κατόπιν, γίνεται αναφορά στους διάφορους τύπους ερωτημάτων που χρησιμοποιούνται. Τέλος παρουσιάζονται διάφορα συστήματα ρευμάτων δεδομένων που είδη έχουν υλοποιηθεί.

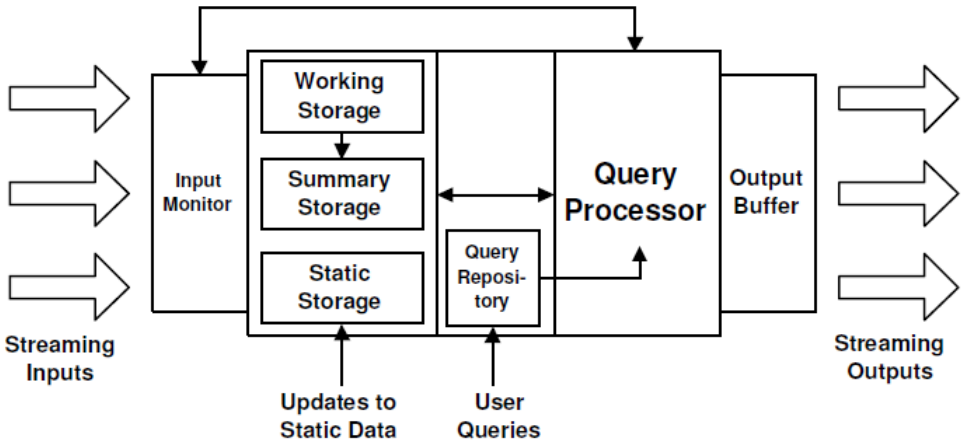
1.2 Μοντέλο ρευμάτων δεδομένων

Στα συστήματα διαχείρισης βάσεων δεδομένων ενδιαφερόμαστε κυρίως για γρήγορες και ακριβείς απαντήσεις στα ερωτήματα που τίθενται. Αυτό επιτυγχάνεται με κατάλληλο σχεδιασμό και βελτιστοποίηση της βάσης κάτι που σχετίζεται με αρκετούς παράγοντες. Δύο από αυτούς είναι τα χαρακτηριστικά του συστήματος που θέλουμε να μοντελοποιήσουμε και τα χαρακτηριστικά της μνήμης η οποία μας διατίθεται για αποθήκευση. Η τεχνολογία των ΣΔΒΔ παρά το γεγονός ότι έχει εξελιχθεί, από την δεκαετία του 60 μέχρι και σήμερα, και θεωρείται αξιόπιστη παρουσιάζεται ανεπαρκής ως προς την επεξεργασία εφαρμογών πραγματικού χρόνου. Αυτό οφείλεται κατά κύριον λόγο στον τρόπο χειρισμού των δεδομένων. Στα ΣΔΒΔ τα δεδομένα είναι στατικά και δεν αλλάζουν σε τακτά χρονικά διαστήματα ενώ τα ερωτήματα αναλύονται σε πλάνα ερωτημάτων με σκοπό την βελτιστοποίηση του τρόπου αποτίμησής τους. Ακόμα, τα δεδομένα αποθηκεύονται σε αποθηκευτικούς χώρους, όπως π.χ. ο σκληρός δίσκος. Η αρχιτεκτονική ενός συστήματος διαχείρισης ΒΔ ακολουθεί ένα συγκεκριμένο μοντέλο προσπέλασης των δεδομένων σύμφωνα με το οποίο ο χρήστης χρειάζεται να ανακτήσει δεδομένα. Έτσι υποβάλει ένα ερώτημα στο σύστημα για να αντλήσει τις ανάλογες πληροφορίες (pull-based model). Για δεδομένα που αλλάζουν με τρόπο δυναμικό, συνεχή, αδιάκοπο και σε πραγματικό χρόνο, η επεξεργασία αυτή καθίσταται ιδιαίτερα χρονοβόρα. Έτσι, παρουσιάστηκε η ανάγκη για ανάπτυξη ευέλικτων συστημάτων που να διαχειρίζονται με αποτελεσματικότητα τέτοια δεδομένα κάτι που οδήγησε στη δημιουργία του μοντέλου των ρευμάτων δεδομένων.

Στο μοντέλο ρευμάτων δεδομένων, τα δεδομένα εισόδου που πρέπει να επεξεργαστούν δεν είναι διαθέσιμα από πρόσβαση στο δίσκο ή τη μνήμη, αλλά φτάνουν ως ένα ή περισσότερα συνεχή ρεύματα δεδομένων. Τα ρεύματα δεδομένων

αποτελούνται από σχεσιακές πλειάδες όπως στις συμβατικές βάσεις δεδομένων. Οι πλειάδες αυτές συνοδεύονται από ένα πεδίο που αναφέρει την προέλευσή τους και ένα *χρονόσημο* (*timestamp*) που αναφέρεται στην χρονική στιγμή που έφτασε στο σύστημα. Τυπικά ως ρεύμα δεδομένων ορίζεται μία συνεχής, πραγματικού χρόνου, χρονικά ταξινομημένη μη φραγμένη ακολουθία δεδομένων ([12]). Ο χρόνος μετριέται είτε σε πραγματικό χρόνο, π.χ. 1 μέρα, είτε ως χρονόσημο, π.χ. 10 τελευταίες ανανεώσεις δεδομένων. Ένα σύστημα διαχείρισης ρευμάτων δεδομένων μοιάζει με εκείνο των βάσεων δεδομένων. Αντίθετα όμως, στις εφαρμογές ρευμάτων δεδομένων, τα στοιχεία προωθούνται στο σύστημα (push-based model) και εκείνο οφείλει να παράγει τα αποτελέσματα των υποβαλλόμενων ερωτημάτων από τους χρήστες. Το γεγονός ότι δεν είναι πλέον ο χρήστης εκείνος που δρομολογεί την ροή των δεδομένων, αλλά οι ίδιες οι πηγές, αποτελεί την ειδοποιό διαφορά με το πρότυπο των συμβατικών ΣΔΒΔ. Έτσι, ο χρήστης περιορίζεται σε «παθητικό» ρόλο του παρατηρητή, αρκούμενος να ρυθμίζει τις παραμέτρους του συστήματος και να υποβάλλει ερωτήματα. Τα ρεύματα δεδομένων παρουσιάζουν τα εξής χαρακτηριστικά ([12], [22]):

- Τα δεδομένα καταφθάνουν έγκαιρα, σε πραγματικό χρόνο (online) και τότε μπορούν επεξεργαστούν.
- Τα ρεύματα δεδομένων μπορεί να είναι απεριόριστα σε μέγεθος.
- Το σύστημα οφείλει να δίνει έγκυρες και συνεπείς απαντήσεις σε πραγματικό χρόνο ακόμη και αν το μέγεθος των ρευμάτων είναι απεριόριστο.
- Τα δεδομένα συνήθως δεν αποθηκεύονται και αντικαθίστανται από τα καινούργια.
- Χρησιμοποιείται αποκλειστικά η κύρια μνήμη και υπάρχουν ευέλικτες δομές για γρήγορες ενημερώσεις.
- Τα ερωτήματα αντιμετωπίζονται ως ερωτήματα διαρκείας και παρουσιάζεται μικρή διάρκεια εγκυρότητας απαντήσεων π.χ οι απαντήσεις των ερωτημάτων μπορεί να αλλάζουν ανάλογα με το χρόνο.
- Το σύστημα δεν έχει κανέναν έλεγχο πάνω στην εγκυρότητα καθώς επίσης και στη σειρά με την οποία τα δεδομένα φθάνουν ώστε να υποστούν επεξεργασία. Έτσι πρέπει να εξασφαλίζεται η κανονική λειτουργία του συστήματος σε περίπτωση που αντιμετωπίζονται τέτοιες ατέλειες.
- Έτσι, το μοντέλο δεδομένων και τα ερωτήματα θα πρέπει να επιτρέπουν τελεστές με βάση τη σειρά των δεδομένων και τον χρόνο που λαμβάνονται.
- Οι αλγόριθμοι που χρησιμοποιούνται είναι ενός περάσματος (one pass).
- Κάθε δεδομένο αφού υποβληθεί σε επεξεργασία είτε απορρίπτεται είτε αρχειοθετείται. Η αποθήκευση των δεδομένων δεν είναι εύκολη υπόθεση, εκτός αν ρητά αποθηκεύονται στη μνήμη, κάτι που μπορεί να γίνει μόνο εφόσον έχουμε μικρό όγκο ρευμάτων δεδομένων.



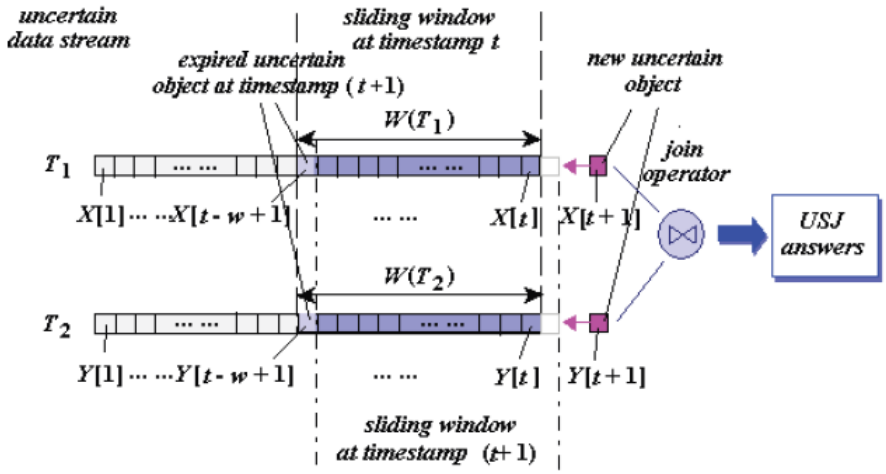
Σχήμα 1.2: Σύστημα διαχείρισης ρευμάτων δεδομένων.

- Η μη αποθήκευση του ρεύματος δεδομένων οδηγεί σε προσεγγιστικές (αλλά συνεπείς) απαντήσεις στην αποτίμηση ερωτημάτων που εμπλέκουν αποθήκευση μεγάλου όγκου δεδομένων (π.χ. δομές σύνοψης).
- Η γλώσσα των ερωτημάτων που τίθενται πάνω στα ρεύματα δεδομένων θα πρέπει να υποστηρίζει δομές και τελεστές βασιζόμενους σε όλα τα παραπάνω χαρακτηριστικά.

Ολοκληρώνοντας, πρέπει να τονιστεί τα ρεύματα που αποτελούν είσοδο στο σύστημα μπορεί να παράγονται από την επεξεργασία άλλων ρευμάτων δεδομένων. Συγκεκριμένα παραδείγματα εφαρμογών ρευμάτων δεδομένων είναι τα εξής:

- *Παρακολούθηση θέσης κινούμενων αντικειμένων.* Τα αντικείμενα αλλάζουν την θέση τους και την ταχύτητά τους στο χώρο ανάλογα με τον χρόνο. Διάφορα ερωτήματα μπορούν να απαντηθούν με βάση τη θέση τους, την κατάστασή τους ή και την συσχέτιση μεταξύ τους.
- *Δίκτυα αισθητήρων.* Καταγράφουν θερμοκρασίες, καιρικά φαινόμενα, δείκτες καιρού, αριθμό αντικειμένων ή ανθρώπων που πέρασαν από κάπου για κάποιο χρονικό διάστημα κτλ.
- *Οικονομικοί δείκτες.* Σε χρηματιστηριακές αγορές ή σε αγοροπωλησίες ή σε δημοπρασίες καταγράφονται οι τιμές των προϊόντων που συναλλάσσονται.
- *Τηλεπικοινωνίες.* Για καταγραφή συνδιαλέξεων ή πληροφοριών (π.χ. ιστορικό επισκεψιμότητας στο internet).
- *Ασφάλεια δικτύων.* Καταγραφή επισκεπτών για εκτίμηση συμφόρησης σε διάφορους κόμβους του δικτύου.

Οι τελεστές που χρησιμοποιούνται στα συστήματα ρευμάτων δεδομένων [13] αποτελούν προεκτάσεις των παραδοσιακών ΣΔΒΔ ώστε να μπορούν να ανταπεξέλθουν στις απαιτήσεις τους. Σε αυτές έχει προστεθεί η έννοια του χρόνου (π.χ. ερωτήματα σε δεδομένα τα τελευταία 10 λεπτά ή τις τελευταίες 10 χρονικές στιγμές). Οι κυριότεροι από αυτούς είναι οι εξής:



Σχήμα 1.3: Σύνδεση ρευμάτων δεδομένων [15].

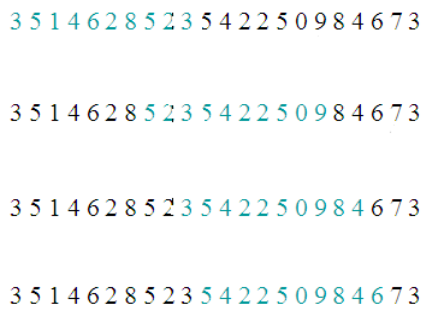
- *Επιλογή* (*Selection*) : Φιλτράρει δεδομένα με συγκεκριμένα χαρακτηριστικά.π.χ. τα λεωφορεία από όλα τα κινούμενα αντικείμενα στην πόλη.
- *Σύνδεση* (*Join*) : Συνδιάζει μεταξύ τους ρεύματα δεδομένων είτε ρεύματα με στατικά δεδομένα (σχήμα 1.3).π.χ. ποια λεωφορεία και τράμ από όλα τα κινούμενα αντικείμενα στην πόλη έχουν ταχύτητα μικρότερη από 10χμ.
- *Συνάθροισης* (*aggregation*) : Συσχετίζει αντικείμενα μεταξύ τους, π.χ. ο μέσος όρος των ταχυτήτων όλων των κινούμενων αντικειμένων.
- *Πολυπλεξία και αποπολυπλεξία* (*multiplexing and demultiplexing*) : Γίνεται ένωση και αποσύνθεση ρευμάτων, κατ' αντιστοιχία με τους τελεστές ένωσης και ομαδοποίησης στις παραδοσιακές βάσεις δεδομένων.
- *Τελεστές συχνότητας* (*Frequent item queries*) : Δίνει την συχνότητα ανίχνευσης τιμών που εμφανίζονται συχνά στο ρεύμα.

Πρέπει να τονιστεί ότι επιπρόσθετοι τελεστές στο μέλλον μπορεί να δημιουργηθούν με βάση τις ανάγκες που παρουσιάζονται στις εφαρμογές.

1.3 Ερωτήματα

Η φύση των ερωτημάτων σε συστήματα διαχείρισης ρευμάτων δεδομένων είναι ίδια με εκείνη των αντίστοιχων για βάσεις δεδομένων, διαφοροποιούνται όμως αρκετά ως προς τη σημασιολογία και τις λειτουργίες των τελεστών τους. Διακρίνονται σε δύο κατηγορίες ([12]):

- *Ερωτήματα διαρκείας* (*continuous queries*) : Τα ερωτήματα εκτελούνται συνεχώς και οι απαντήσεις τους ανανεώνονται σε κάθε χρονόσημο, με βάση τα καινούργια στοιχεία που καταφθάνουν.
- *Στιγμιαία ερωτήματα* (*one-time*) : έχουν την ίδια φύση με τα αντίστοιχα ερωτήματα των βάσεων δεδομένων. Δεν ανανεώνονται χρονικά και επεξεργάζονται τα τρέχοντα δεδομένα.



Σχήμα 1.4: Παράθυρο ρεύματος δεδομένων.

Τα ερωτήματα διαρκείας εφαρμόζονται όταν είναι απαραίτητο να υπάρχει πάντοτε διαθέσιμη η τρέχουσα απάντηση σε ένα ερώτημα, το οποίο αναφέρεται στα διαρκώς μεταβαλλόμενα δεδομένα του ρεύματος. Η απάντησή του συνήθως αποτελείται από ρεύμα δεδομένων εξόδου, το οποίο όμως τις περισσότερες φορές δεν αποθηκεύεται. Το ρεύμα δεδομένων εισόδου πολλές φορές είναι ανεξάντλητο, οπότε το μέγεθος της απάντησης είναι μεγάλο και η αποθήκευση δύσκολα μπορεί να πραγματοποιηθεί. Κάθε χρονική στιγμή, νέα στοιχεία και νέες πλειάδες καταγράφονται στα αποτελέσματα αντικαθιστώντας τις παλιές. Οι περιορισμοί στην αποθήκευση των στοιχείων είναι αναγκαία. Τα ερωτήματα διαρκείας παράγουν με τον χρόνο απαντήσεις που αφορούν τα δεδομένα μέχρι την τρέχουσα χρονική στιγμή. Διαγραφές και ενημερώσεις που συναντάμε συχνά στις βάσεις δεδομένων δύσκολα μπορούν να πραγματοποιηθούν. Ερωτήματα που απαιτούν δεδομένα και πληροφορίες που καταγράφηκαν στο παρελθόν δυσχεραίνουν ακόμα περισσότερο τα πράγματα. Για παράδειγμα, ο υπολογισμός του μέσου όρου των 1000 τελευταίων στοιχείων ενός ρεύματος δεδομένων θα απαιτούσε την καταγραφή των τελευταίων 1000 στοιχείων του στη μνήμη και την αναθεώρηση τις απάντησης κάθε χρονική στιγμή. Πιο περίπλοκα και πολύπλοκα ερωτήματα θα είχαν αντίστοιχα περισσότερες απαιτήσεις σε μνήμη. Η αντιμετώπιση των προβλημάτων αυτών έκανε απαραίτητη την ύπαρξη περιορισμών στις απαντήσεις και την παροχή προσεγγιστικών αλλά συνεπών απαντήσεων στα ερωτήματα. Τέλος, ως προς τα ερωτήματα διαρκείας μπορούμε να πούμε ότι χωρίζονται σε δύο κατηγορίες:

- Προκαθορισμένα ερωτήματα (predifined queries)
- Μη προκαθορισμένα ερωτήματα (ad-hoc queries)

Η διάκριση αυτή γίνεται για τον εξής λόγο : τα μεν προκαθορισμένα ερωτήματα είναι εκ των προτέρων γνωστά στο σύστημα πριν καταφτάσουν τα δεδομένα, τα δεδομένα γνωρίζουμε ότι πάντα θα είναι διαθέσιμα για επεξεργασία και το σύστημα μπορεί να κατανέμει καταλλήλως τους πόρους του, π.χ. την μνήμη του για την επεξεργασία. Τα μη προκαθορισμένα ερωτήματα αναφέρονται σε δεδομένα που πιθανόν να μην είναι διαθέσιμα, όπως π.χ. δεδομένα από το παρελθόν που δεν υπάρχουν πλέον, ενώ επίσης μπορεί να μην είναι διαθέσιμοι οι απαραίτητοι πόροι για την επεξεργασία τους, π.χ. μικρή χωρητικότητα μνήμης.

1.4 Παράθυρα σε ερωτήματα διαρκείας

Ένα σύστημα διαχείρισης βάσεων δεδομένων εξετάζει ένα συγκεκριμένο τμήμα του συνόλου των πλειάδων που καταφτάνουν ως είσοδος. Όπως αναφέρθηκε

και παραπάνω είναι ασύμφορο να αποθηκεύει όλα τα δεδομένα του ρεύματος εισόδου. Το εξεταζόμενο αυτό τμήμα του ρεύματος εισόδου λέμε ότι αποτελεί ένα *παράθυρο* (*window*) επί των πιο πρόσφατων στοιχείων του ρεύματος. Συγκεκριμένα, το παράθυρο περιλαμβάνει ένα τμήμα διαδοχικών πλειάδων στις οποίες τίθεται το ερώτημα. Π.χ. σε ένα δίκτυο αισθητήρων οι τιμές των θερμοκρασιών που καταγράφονται αφορούν τις τελευταίες μέρες ή το πολύ εβδομάδες, με τα παλιότερα δεδομένα να σβήνονται. Έτσι μπορούμε να πούμε ότι τα παράθυρα απομονώνουν ένα πεπερασμένο πλήθος στοιχείων από ένα μεγάλο, πιθανόν άπειρου μήκους ρεύμα δεδομένων. Τρία είδη των παραθύρων συναντάμε συνήθως στις περισσότερες εφαρμογές. Αυτά είναι ([12]):

- *Παράθυρα ορόσημου* (*landmark windows*).Τα παράθυρα έχουν ως σταθερή αφετηρία κάποιο χρονόσημο, αλλά το πέρας τους παρακολουθεί τη χρονική εξέλιξη των πλειάδων του ρεύματος. Επομένως, το νεότερο άκρο του παραθύρου προχωρεί παράλληλα με το χρόνο, ταυτιζόμενο με την παρούσα χρονική στιγμή, ώστε να καλύπτει συνεχώς την έλευση νέων στοιχείων. Το εύρος του παραθύρου αυξάνεται λοιπόν διαρκώς, όπως και ο αριθμός των πλειάδων που περιλαμβάνει. Λ.χ. «Υπολόγισε υπολόγισε το μέσο όρο των θερμοκρασιών που καταγράφτηκαν από την αρχή του έτους μέχρι σήμερα». Έτσι, δεν είναι προκαθορισμένο το μέγεθος μνήμης που απαιτείται.

- *Κυλιόμενα παράθυρα βάσει χρόνου* (*Time based sliding windows*). Έχουν αφετηρία και πέρας που κινούνται ταυτόχρονα παρακολουθώντας την χρονική εξέλιξη των στοιχείων που συρρέουν σύστημα. Έτσι, παλαιότερα δεδομένα απορρίπτονται και καινούργια εισέρχονται με κυμαινόμενο ρυθμό. Το εύρος των παραθύρων παραμένει σταθερό, όμως ούτε το πλήθος των πλειάδων ούτε φυσικά και τα περιεχόμενά τους διατηρούνται αμετάβλητα. Λ.χ. «Στο τέλος κάθε μέρας, υπολόγισε το μέσο όρο των θερμοκρασιών που καταγράφτηκαν». Δεν είναι προκαθορισμένο το μέγεθος μνήμης που απαιτείται.

- *Κυλιόμενα παράθυρα βάσει πλειάδων* (*Tuple based sliding windows*).Έχουν αφετηρία και πέρας που κινούνται ταυτόχρονα παρακολουθώντας τις πλειάδες που συρρέουν στο σύστημα (σχήμα 1.4). Λ.χ. «Υπολόγισε το μέσο όρο των 10 τελευταίων θερμοκρασιών που καταγράφτηκαν». Είναι προκαθορισμένο το μέγεθος μνήμης που απαιτείται.

Πρακτικά, με τη χρήση παραθύρων ο χρήστης έχει την δυνατότητα να μεταβάλλει την εμβέλεια των ερωτημάτων που θέτει. Τα παράθυρα αποτελούν σημαντικό ερευνητικό θέμα στα ρεύματα δεδομένων ως προς την αποδοτικότητά και την βελτιστοποίησή τους.

1.5 Προβλήματα ερωτημάτων

1.5.1 Μη φραγμένες απαιτήσεις μνήμης

Η αποθήκευση ολόκληρου του ρεύματος δεδομένων σε κάθε περίπτωση πρακτικά είναι αδύνατη, ιδιαίτερα για ρεύματα άπειρα σε μέγεθος. Η αποθήκευση

ανεξάντλητων ρευμάτων θα απαιτούσε τεράστιο αποθηκευτικό χώρο και εκτός από δαπανηρή το σύστημα δεν θα μπορούσε να δώσει απαντήσεις σε πραγματικό χρόνο από τις εισόδους που θα δεχόταν. Το κόστος ανάκτησης των δεδομένων από το δίσκο θα ήταν μεγάλο και δεν θα έδινε περιθώρια απόκρισης σε πραγματικό χρόνο. Εκτός από το δίσκο, η χρήση άλλων μνημών, όπως της κύριας μνήμης θα βοηθούσε ως προς την ταχύτητα απόκρισης των απαντήσεων αλλά το μέγεθός της είναι μικρό και δεν θα βοηθούσε. Έτσι θα οδηγούμασταν γρήγορα στην απώλεια μέρους του ρεύματος δεδομένων εισόδου. Από όλα τα παραπάνω καταλήγουμε στο συμπέρασμα ότι η ανάπτυξη αλγορίθμων που με μικρό χρόνο επεξεργασίας θα μπορούν να εκμεταλλεύονται την κύρια μνήμη χωρίς την πρόσβαση στον σκληρό δίσκο είναι αναγκαία. Το πρόβλημα ανάπτυξης τέτοιων αλγορίθμων αποτελεί θέμα έρευνας μέχρι σήμερα και εστιάζει κυρίως στη μελέτη των παραθυρικών ερωτημάτων για προσεγγιστικές απαντήσεις, ώστε να αποφεύγεται η πρόσβαση στον δίσκο.

1.5.2 Προβληματικοί τελεστές

Η επέκταση των τελεστών έγινε βάσει των αντίστοιχων τελεστών στα συστήματα διαχείρισης βάσεων δεδομένων. Σε πολλές περιπτώσεις όμως η επέκταση αυτή παρουσίασε αδυναμίες, με κάποιες λειτουργίες τελεστών να συγκρούονται με τις βασικές αρχές του μοντέλου των βάσεων δεδομένων ([12]). Οι κυριότεροι τελεστές ήταν οι εξής:

- *Ανασχετικοί τελεστές* (*blocking operators*) : Χρησιμοποιούν μέρος του ρεύματος εισόδου ώστε να δώσουν κάποια απάντηση. Παράδειγμα τέτοιων τελεστών είναι οι τελεστές αθροίσματος (sum) και μέσου όρου (AVG). Το πρόβλημα γίνεται μεγαλύτερο εφόσον το ρεύμα έχει άπειρο μήκος και κάθε φορά αναμένουμε απάντηση από τον τελεστή και εφόσον ο τελεστής αναφέρεται σε μεγάλο όγκο δεδομένων. Δύο λύσεις που προτείνονται είναι η αντικατάσταση των προβληματικών τελεστών με άλλους που κάνουν την ίδια δουλειά και η υποδιαίρεση του ρεύματος δεδομένων που επεξεργάζονται σε μικρότερα τμήματα. Τα τμήματα αυτά φέρουν πληροφορίες για τα χαρακτηριστικά του, οι οποίες μπορούν να χρησιμοποιηθούν τελικά από τους τελεστές για προσεγγιστικές απαντήσεις.

- *Τελεστές διατήρησης κατάστασης* (*unbounded stateful operators*) : Οι τελεστές αυτοί διατηρούν τις τιμές των πλειάδων που επεξεργάζονται. Παράδειγμα τέτοιων τελεστών είναι η σύνδεση (join) και η τομή (intersect) μεταξύ ρευμάτων. Ο τρόπος αντιμετώπισης των προβλημάτων που δημιουργούν οι εν λόγω τελεστές είναι ο ίδιος με τον αντίστοιχο των ανασχετικών τελεστών, ώστε να αποτραπεί ή εξάντληση της μνήμης από ένα αποτέλεσμα που αυξάνεται ανεξέλεγκτα.

1.6 Αλγόριθμοι για ρεύματα δεδομένων

Οι αλγόριθμοι που χρησιμοποιούνται για την διαχείριση ρευμάτων δεδομένων και την επεξεργασία των ερωτημάτων διαρκείας θα πρέπει να εξασφαλίζουν την

γρήγορη και σωστή λειτουργία του συστήματος. Οφείλουν να επεξεργάζονται αποδοτικά τη μνήμη ώστε να μειώνουν το κόστος επεξεργασίας. Επιπροσθέτως, ο αλγόριθμος πρέπει να είναι σε θέση να παρέχει και να αποθηκεύει σημαντικά ενδιάμεσα αποτελέσματα, έτσι ώστε να μπορεί να αυξήσει τις επιδόσεις του. Τέλος πρέπει να παρέχεται η δυνατότητα να προβλεφθούν μελλοντικές πληροφορίες με την σωστή διαχείριση των δεδομένων. Υπάρχουν αλγόριθμοι που επιτρέπουν πολλαπλά περάσματα από το ρεύμα δεδομένων, ωστόσο οι πλέον αποδοτικοί είναι εκείνοι που επεξεργάζονται τα στοιχεία μόνο μια φορά (single pass). Τα στοιχεία σαρώνονται μόνο μία φορά και ο αλγόριθμος πρέπει να είναι σε θέση να υπολογίζει τόσο ενδιάμεσα αποτελέσματα τμημάτων των δεδομένων που έχει παρέλθει μέχρι εκείνη τη στιγμή, όπως το μέσο όρο των αριθμών, όσο και τελικά αποτελέσματα.

1.6.1 Μαζική επεξεργασία

Τα δεδομένα επεξεργάζονται μαζικά (batch processing) και όχι μεμονωμένα το καθένα όταν καταφθάνουν. Με αυτόν τον τρόπο η εκτέλεση των ερωτημάτων γίνεται γρηγορότερη, κυρίως σε περιπτώσεις όπου η άφιξη των στοιχείων γίνεται με μεγάλη συχνότητα και η επεξεργασία τους αργεί. Τα αποτελέσματα δίνονται προσεγγιστικά, αφού δεν λαμβάνονται έγκαιρα και αντιπροσωπεύουν την ακριβή απάντηση σε κάποια χρονική στιγμή στο πρόσφατο παρελθόν και όχι στο παρόν. Η τεχνική αυτή δίνει έγκυρες και σε πραγματικό χρονο απαντήσεις αφού η καθυστέρηση της απάντησης δεν βλάπτει την εγκυρότητα του αποτελέσματος. Ένας αλγόριθμος που παρουσιάζει μεγάλη συχνότητα λήψης πληροφοριών μπορεί να περιοριστεί σε μία μέση συχνότητα λήψης πληροφοριών αποθηκεύοντας προσωρινά τις πληροφορίες και στη συνέχεια να τις επεξεργαστεί ομαδικά, όταν η συχνότητα μειωθεί.

1.6.2 Δειγματοληψία

Με τη δειγματοληψία (sampling) ρευμάτων δεδομένων επεξεργαζόμαστε μόνο έναν περιορισμένο αριθμό δειγμάτων και όχι το σύνολο του ρεύματος δεδομένων. Η δειγματοληψία είναι δυνατόν να είναι [12] :

- *τυχαία* (*randomized sampling*), οπότε με τυχαίο τρόπο κάθε πλειάδα είτε αποθηκεύεται ως συστατικό του δείγματος είτε απορρίπτεται. Προϋποθέτει ότι το δείγμα του ρεύματος που λαμβάνεται είναι αντιπροσωπευτικό.
- *ομοιόμορφη* (*uniform sampling*), οπότε κάθε ένα συγκεκριμένο αριθμό δειγμάτων αποθηκεύεται μία πλειάδα. Οι υπόλοιπες απορρίπτονται.
- *διαστρωματωμένη* (*stratified sampling*), η οποία μειώνει τα σφάλματα λόγω διακύμανσης των δεδομένων και του σφάλματος στα ερωτήματα που έχουν συσταδοποιήσει δεδομένα.

Η τεχνική της δειγματοληψίας εφαρμόζεται κυρίως όταν η ενημέρωση του συστήματος με τις εισερχόμενες πλειάδες είναι χρονοβόρα, ενώ η τεχνική της μαζικής επεξεργασίας εφαρμόζεται όταν η επεξεργασία των διατηρούμενων πλειάδων είναι αργή. Η τεχνική της δειγματοληψίας μπορεί να εφαρμοσθεί ταυτόχρονα

με την τεχνική της μαζικής επεξεργασίας, βελτιώνοντας κατά πολύ την απόδοση του συστήματος.

1.6.3 Συνόψεις δεδομένων

Οι συνόψεις δεδομένων (summaries or data synopses) αποτελούν μια συνοπτική περίληψη της πληροφορίας με μειωμένη ακρίβεια. Το μέγεθός τους είναι σημαντικά μικρότερο, σε λογαριθμικό ή πολυλογαριθμικό βαθμό, σε σχέση με το σύνολο της πληροφορίας. Ο σχηματισμός των συνόψεων θα πρέπει να γίνεται με ένα πέρασμα των δεδομένων με τη σειρά που καταφτάνουν. Ο κεντρικός επεξεργαστής έχει τη δυνατότητα να λαμβάνει υπόψη του τις υπάρχουσες συνόψεις και να παράγει προσεγγιστικά αποτελέσματα για τα ερωτήματα που τίθενται. Η χρήση περιλήψεων συμβάλει στη δραστική μείωση του χώρου που καταλαμβάνουν τα δεδομένα σε κάθε αποθηκευτικό χώρο. Οι συνόψεις εφαρμόζονται συνήθως σε ερωτήματα σύνδεσης με παράλληλη χρήση τελεστών συνάθροισης. Οι κυριότερες τεχνικές συνόψεων είναι οι εξής:

Σκίτσα δεδομένων

Η τεχνική των σκίτσων δεδομένων εφαρμόζεται για την παραγωγή περίληψης με τυχαίο τρόπο (randomized sketching), χρησιμοποιώντας ένα μικρό μέρος της μνήμης ώστε να υπολογίσουμε μία απάντηση σε συγκεκριμένα ερωτήματα (συνήθως απόστασης) πάνω σε ένα σύνολο δεδομένων.

Κυματίδια

Τα κυματίδια (wavelets) αποτελούν μαθηματικούς μετασχηματισμούς που αναπαριστούν την πληροφορία με αριθμητικές συναρτήσεις, των οποίων οι συντελεστές είναι προβολές του σήματος σε ένα ορθοκανονικό σύνολο διανυσμάτων αναφοράς. Η επιλογή των κατάλληλων διανυσμάτων αναφοράς κρίνει και τον τύπο των wavelets. Η χρησιμότητα των κυματιδίων φαίνεται από το γεγονός ότι από μικρό πλήθος των σημαντικότερων συντελεστών των γραμμικών προβολών μπορούν να αναπαρασταθούν τα πρωτότυπα δεδομένα με κριτήριο την Ευκλείδεια νόρμα τους. Διατηρώντας μια αντιπροσωπευτική σύνοψη του ρεύματος με τη μορφή κυματιδίων, η οποία διαθέτει μικρό όγκο, είναι εύκολο να απαντηθούν προσεγγιστικά ερωτήματα συνάθροισης για μεμονωμένα χαρακτηριστικά των δεδομένων με αρκετά καλή ακρίβεια. Τα κυματίδια διευκολύνουν σημαντικά το χειρισμό των ρευμάτων δεδομένων.

Ιστογράμματα

Τα ιστογράμματα είναι συνηθισμένες δομές περίληψης που αναπαριστούν συνοπτικά την κατανομή των τιμών σε ένα σύνολο δεδομένων. Χρησιμοποιούνται κυρίως για τον υπολογισμό του μεγέθους ερωτημάτων, να δίνουν προσεγγιστικές απαντήσεις σε ερωτήματα και την εξόρυξη δεδομένων (σχήμα 1.5). Το αναμενόμενο εύρος του συνόλου τιμών χωρίζεται σε διαστήματα για καθένα από τα οποία

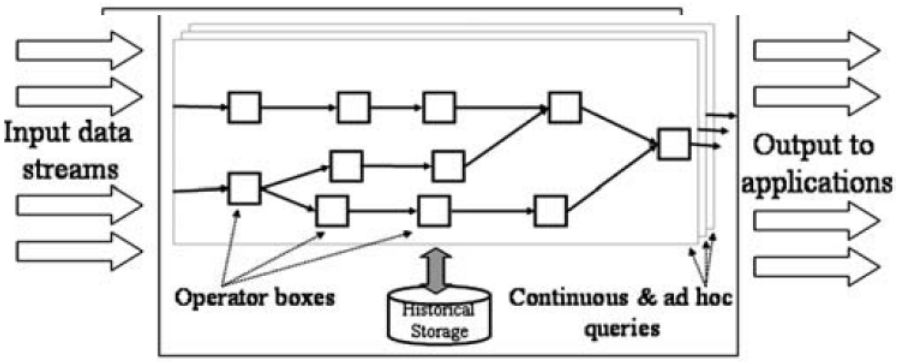
Σχήμα 1.5: Παράδειγμα ιστογράμματος.

διατηρείται η συχνότητα τιμών που καταφτάνει, σχηματίζοντας έτσι ένα ιστόγραμμα, το οποίο μπορεί να αναπαρασταθεί και με ένα διάγραμμα. Σαφώς, το πλήθος και το εύρος των διαστημάτων συνηγορεί στην ακρίβεια της αναπαράστασης. Ένα ρεύμα δεδομένων πρέπει να χωρίζεται σε τέτοια διαστήματα από συγκεκριμένους αλγόριθμους οι οποίοι θα επιτρέπουν τον δυναμικό προσδιορισμό των διαστημάτων. Κατόπιν, βάσει του εύρους των τιμών κάθε διαστήματος επιχειρείται να εντοπισθεί η αντιπροσωπευτικότερη τιμή, η οποία τελικά θα διατηρηθεί αποθηκευμένη. Η ανασύσταση του συνόλου των δεδομένων από το ιστόγραμμα γίνεται με βάση τις αντιπροσωπευτικές τιμές κάθε διαστήματος και τη συχνότητα των στοιχείων που αντιστοιχεί στο εν λόγω διάστημα. Ενώ έχουν προταθεί αρκετοί αλγόριθμοι ιστογραμμάτων, το ζήτημα εύρεσης δυναμικών αλγορίθμων προσδιορισμού των διαστημάτων με βάση τις πλειάδες που καταφτάνουν, επιδέχεται αρκετή έρευνα.

1.7 Γλώσσες ρευμάτων δεδομένων

Μία γλώσσα ρευμάτων δεδομένων μπορεί να έχει πολλά κοινά στοιχεία με γλώσσες που χρησιμοποιούνται από συστήματα διαχείρισης βάσεων δεδομένων όπως είναι η γλώσσα SQL. Αυτό συμβαίνει γιατί τόσο τα συστήματα διαχείρισης ρευμάτων δεδομένων όσο και τα αντίστοιχα βάσεων δεδομένων παρουσιάζουν αρκετά κοινά στοιχεία. Επιπροσθέτως, γλώσσες όπως η SQL χρησιμοποιούνται ευρέως και πιθανές επεκτάσεις οι οποίες είναι αναγκαίες να γίνουν, λ.χ. παράθυρα, δεν έρχονται σε αντίθεση με κανένα σημείο της δομής και της λειτουργίας τους. Έτσι η SQL είναι ιδανική ως γλώσσα από την οποία μπορεί να δανειστούν αρκετά στοιχεία οι αντίστοιχες των ρευμάτων δεδομένων. Μέχρι στιγμής τόσο για ερευνητικούς όσο και εμπορικούς σκοπούς έχουν αναπτυχθεί αρκετά τέτοια συστήματα. Τα κυριότερα από αυτά, τα οποία αναπτύχθηκαν σχεδόν ταυτόχρονα είναι τα εξής:

- Stream: Αναπτύχθηκε από το πανεπιστήμιο Stanford το 2001 και τα αρχικά του σημαίνουν STanford stREam datA Management. Η υλοποίησή του βασίστηκε στην γλώσσα βάσεων δεδομένων SQL, με την επέκτασή της σε μία καινούργια γλώσσα ρευμάτων δεδομένων με το όνομα CQL (Continuous



Σχήμα 1.6: Σύστημα Audora για διαχείριση ρευμάτων δεδομένων.

Query Language). Η CQL διατήρησε τα χαρακτηριστικά της SQL με επεκτάσεις για κυλιόμενα παράθυρα και δειγματοληψία. Το Stream παρέχει ένα ολοκληρωμένο περιβάλλον διασύνδεσης για την υποβολή ερωτημάτων που υποστηρίζει την επεξεργασία τους στην γλώσσα CQL με αρκετά εύχρηστο τρόπο ανάγνωσης και εγγραφής ρευμάτων δεδομένων.

- AUDORA: Αναπτύχθηκε από τα πανεπιστήμια MIT, Brown και Brandeis το 2001. Ο πρωταρχικός στόχος της AUDORA ήταν να μπορεί να υποστηρίξει αποτελεσματικά και απρόσκοπτα εφαρμογές παρακολούθησης αντικειμένων, διαχείρισης τηλεπικοινωνιακών δεδομένων κτλ που απαιτούν την ικανότητα χειρισμού τεράστιου όγκου δεδομένων συνεχών ρευμάτων που φθάνουν σε πραγματικό χρόνο. Η επίτευξη υψηλής κλιμακωσιμότητας σε κατανεμημένα συστήματα επεξεργασίας αποτέλεσε έναν ακόμα σημαντικό λόγο για την ανάπτυξή του. Το σύστημα υποβολής ερωτημάτων τα σχεδιάζει σε γραφικό περιβάλλον δημιουργώντας έτσι ένα διάγραμμα ροής αποτελούμενο από τετράγωνα και βέλη (σχήμα 1.6). Μεταξύ των χαρακτηριστικών του είναι η αρχιτεκτονική του συντονίζεται από έναν χρονοπρογραμματιστή (scheduler) και το γεγονός ότι υποστηρίζει και ένα σύστημα αποθήκευσης δεδομένων (storage manager).

- TelegraphCQ: Αναπτύχθηκε από το πανεπιστήμιο Berkeley. Η υλοποίησή του βασίστηκε στην γλώσσα προγραμματισμού C/C++ και στην αρχιτεκτονική της PostgreSQL. Χρησιμοποιεί τον μηχανισμό Eddy για να πετύχει καλύτερα και αποδοτικότερα αποτελέσματα στην εκτέλεση των ερωτημάτων διαρκείας. Μαζί με τα δεδομένα και τα ερωτήματα εμφανίζονται ως ρεύματα δεδομένων και αλλάζουν και αυτά βάσει του χρόνου. Τέλος, πρέπει να πούμε ότι ο πρωταρχικός στόχος του συστήματος αυτού ήταν να χρησιμοποιηθεί για δικτυακούς σκοπούς , όπως δίκτυα επικοινωνιών και αισθητήρων.

Άλλα συστήματα ρευμάτων δεδομένων είναι τα Nile (Purdue), Medusa (Brown). Όλα αυτά τα συστήματα έχουν επανεκδοθεί αρκετές φορές, παρέχοντας κάθε φορά επιπλέον δυνατότητες.

Κεφάλαιο 2

Διαχείριση κινούμενων αντικειμένων

2.1 Χωρικά και χωροχρονικά δεδομένα

Η πρόσφατη ραγδαία εξέλιξη των συστημάτων γεωγραφικού εντοπισμού αποτελεί τη βάση για την ανάπτυξη ποικίλων σχετικών εμπορικών και ερευνητικών εφαρμογών. Εφαρμογές πλοήγησης στο οδικό δίκτυο μιας πόλης σε πραγματικό χρόνο είναι ήδη πραγματικότητα. Ο χρήστης υποβάλει τη θέση προορισμού του και το σύστημα, του παρέχει σε πραγματικό χρόνο την συντομότερη διαδρομή. Ταυτόχρο-να, το σύστημα μπορεί να εντοπίζει τη θέση του και να τον να καθοδηγεί ανάλογα, ακόμα και να διορθώνει τη διαδρομή σε περίπτωση λάθους. Η ανάπτυξη τέτοιων εφαρμογών εισήγαγαν ένα νέο αντικείμενο έρευνας στο πεδίο των βάσεων δεδομένων, τις χωροχρονικές βάσεις δεδομένων.

Η ανάπτυξη κατάλληλων δομών, οι οποίες να αναπαριστούν τα χωρικά και τα χωροχρονικά δεδομένα και κατάλληλων πράξεων (συναρτήσεων) επί αυτών, που να περιγράφουν τα χωροχρονικά φαινόμενα είναι απαραίτητα στοιχεία για την υλοποίηση των εφαρμογών αυτών. Έχουν προταθεί διάφοροι τρόποι αναπαράστασης των πραγματικών χωρικών αντικειμένων στον υπολογιστή [9, 10] :

- *Το σημείο* (*point*) : που αναπαριστά τη θέση ενός αντικειμένου στο χάρτη. Επεκτείνεται για τις ανάγκες των χωροχρονικών βάσεων δεδομένων στο κινούμενο σημείο.

- *Η γραμμή* (*line*) : που αναπαριστά συνδέσεις μεταξύ αντικειμένων στο χάρτη, όπως π.χ. δρόμους. Αποτελείται από ένα ή περισσότερα ευθύγραμμα τμήματα. Επεκτείνεται για τις ανάγκες των χωροχρονικών βάσεων δεδομένων στην κινούμενη γραμμή.

- *Η περιοχή* (*region*) : που αναπαριστά μία γεωγραφική περιοχή στον χάρτη π.χ. πόλη, νομός ή χώρα. Ενδέχεται να περιέχει οπές ή να συνθέτεται από πολλά μη επικαλυπτόμενα μέρη. Επεκτείνεται για τις ανάγκες των

χωροχρονικών βάσεων δεδομένων στις κινούμενες περιοχές (π.χ. κινούμενα ερωτήματα περιοχής).

Για παράδειγμα σε ένα σύστημα παρακολούθησης της κυκλοφορίας αυτοκινήτων ένα ερώτημα θα ήταν «εντόπισε τα αυτοκίνητα που κινούνται στο εθνικό δίκτυο με προορισμό την Αθήνα και βρίσκονται εντός του νομού Αττικής». Τα ερωτήματα πρέπει να επεξεργάζονται αποδοτικά , ίσως και πολλά μαζί, και παράλληλα να παρακολουθείται η κίνηση και η μεταβολή των χωρικών αντικειμένων. Το πλήθος των δεδομένων στις περισσότερες περιπτώσεις χωρικών εφαρμογών είναι αρκετά μεγάλο. Η αποτίμηση των ερωτημάτων πρέπει να γίνεται με τρόπο αποδοτικό ώστε να εξασφαλίζεται ακρίβεια και καλές χρονικές επιδόσεις.Έτσι αναγκαία είναι η οργάνωση των δεδομένων σε κατάλληλες δομές για την εύκολη προσπέλασή τους από τα ερωτήματα, με χωρικές μεθόδους προσπέλασης (spatial access methods). Οι χωρικές μέθοδοι προσπέλασης έχουν να αντιμετωπίσουν το πρόβλημα της απουσίας ολικής διάταξης στα χωρικά δεδομένα, γεγονός που έχει οδηγήσει σε αρκετές προτάσεις για δομές και αλγορίθμους επεξεργασίας.

2.1.1 Μοντελοποίηση χωρικών δεδομένων

Όπως είδη αναφέρθηκε, τα χωρικά αντικείμενα που χειρίζονται οι χωρικές βάσεις δεδομένων είναι το σημείο, η γραμμή και η περιοχή με την επέκταση τους στις χωροχρονικές βάσεις δεδομένων σε κινούμενες χωρικές οντότητες. Η μοντελοποίηση των αντικειμένων γίνεται σε δύο επίπεδα:

- Στο αφηρημένο μοντέλο γίνεται η θεμελίωση που αποσκοπεί στη μελέτη των χωροχρονικών φαινομένων. Βασίζεται σε αφηρημένους τύπους δεδομένων (abstract data types) και αποτελεί επέκταση του αφηρημένου μοντέλου για στατικά χωρικά αντικείμενα. Οι τύποι δεδομένων και οι λειτουργίες του δημιουργούν μια άλγεβρα κινούμενων αντικειμένων, παρόλα αυτά το αφηρημένο μοντέλο δεν είναι κατάλληλο για την αναπαράσταση των δεδομένων στον υπολογιστή.

- Στο διακριτό μοντέλο όπου γίνεται η αναπαράσταση της κίνησης σε πεπερασμένα στιγμιότυπα. Μία κινούμενη περιοχή ορισμένη στο αφηρημένο μοντέλο θα μπορούσε να αντιστοιχεί στο διακριτό μοντέλο σε μια πολυγωνική περιοχή, η οποία προσεγγίζει την πραγματική.

2.1.2 Χωρικά ερωτήματα

Τα κυριότερα χωρικά ερωτήματα είναι οι εξής [12]:

- *Ισότητας* (*exact match query*) : Να βρεθούν όλα τα αντικείμενα που έχουν την ίδια γεωμετρία με ένα συγκεκριμένο αντικείμενο.

- *Σημείου* (*point query*) : Να βρεθούν όλα τα αντικείμενα που περιέχουν ένα σημείο.

- *Παραθύρου* (*window query*) : Να βρεθούν όλα τα αντικείμενα που έχουν ένα τουλάχιστον κοινό σημείο με ένα παράθυρο (π.χ. ορθογωνικό).

Σχήμα 2.1: Παραδείγματα χωρικών ερωτημάτων [12].

- *k-εγγύτερων γειτόνων* (*k-nearest neighbor*) : Να βρεθούν τα k κοντινότερα αντικείμενα σε ένα συγκεκριμένο αντικείμενο.
- *Χωρικής σύνδεσης* (*spatial query*) : Πραγματοποιεί τον συνδυασμό δύο ή περισσότερων δεδομένων βάσει ενός κοινού τους χαρακτηριστικού.
- *Τομής* (*intesection query*) : Να βρεθούν όλα τα αντικείμενα με τα οποία ένα συγκεκριμένο αντικείμενο έχει κοινά εσωτερικά σημεία.
- *Ενεργητικού εγκλεισμού* (*containment query*) : Να βρεθούν όλα τα αντικείμενα που περιέχονται από ένα συγκεκριμένο αντικείμενο.
- *Παθητικού εγκλεισμού* (*enclosure query*) : Να βρεθούν όλα τα αντικείμενα που περιέχουν ένα συγκεκριμένο αντικείμενο.
- *Γειτνίασης* (*adjacency query*) : Να βρεθούν αντικείμενα που συνορεύουν με το δοσμένο.

Όλα τα παραπάνω ερωτήματα εκφράζονται στην άλγεβρα και στο ΣΔΒΔ μέσω των λειτουργιών που έχουν αναπτυχθεί. Αντίστοιχα, μπορούν να επεκταθούν στο πεδίο του χρόνου για κινούμενα χωρικά αντικείμενα μέσω των αντίστοιχων λειτουργιών. Η χωρική διάσταση μπορεί να αναφέρεται είτε στο παρελθόν, είτε στο παρόν, είτε στο άμεσο μέλλον υποθέτοντας και κάποια τεχνική για την πρόβλεψη.

2.2 Κινούμενα αντικείμενα

Στην εποχή μας, ο τομέας της παρακολούθησης κινούμενων αντικειμένων παρουσιάζει ιδιαίτερη ανάπτυξη. Με χρήση συστημάτων όπως GPS και ασύρματων τηλεπικοινωνιών εντοπίζεται η θέση κάθε αντικειμένου και διοχετεύεται περιοδικά σε ένα κεντρικό επεξεργαστή (Server). Εκεί γίνεται η επεξεργασία των εκάστοτε

Σχήμα 2.2: Κίνηση αντικειμένου.

ερωτημάτων που τίθενται. Τα κινούμενα αντικειμενα σε αυτές τις εφαρμογές θεωρούνται σημειακά αφού ο χώρος που καλύπτουν είναι αμελητέος συγκριτικά με την έκταση του χώρου. Οι τύποι ερωτημάτων ποικίλουν ανάλογα με την εφαρμογή. Τα ερωτήματα που θα μας απασχολήσουν κυρίως στην παρούσα εργασία είναι ερωτήματα και έχουν να κάνουν θέσεις των αντικειμένων στο παρόν (now-related queries). Πρόκειται για ερωτήματα περιοχής (coordinate-based queries), τα οποία αναφέρονται μόνο στις τρέχουσες θέσεις των αντικειμένων. Έτσι ο κεντρικός επεξεργαστής δε χρειάζεται να καταγράφει την ιστορία των θέσεων των αντικειμένων (τροχιές) παρά μόνο τις συντεταγμένες της τελευταίας ενημέρωσης της θέσης τους. Οι θέσεις των κινούμενων αντικειμένων λαμβάνονται με τη μορφή ρευμάτων δεδομένων στον κεντρικό υπολογιστή και ακολουθούν τα χαρακτηριστικά τους (κίνηση αντικειμένου στο σχήμα 2.2). Επιπροσθέτως τα ερωτήματα που μελετώνται, αντιμετωπίζονται ως ερωτήματα διαρκείας (continuous queries). Παρέχονται και αυτά ως ρεύματα δεδομένων στον κεντρικό υπολογιστή και η αποτίμησή τους πρέπει να γίνεται έγκαιρα ώστε να παρέχονται άμεσα απαντήσεις στους χρήστες.

2.2.1 Θέσεις κινούμενων αντικειμένων

Η κίνηση αποτελεί χαρακτηριστικό για πολλά αντικείμενα. Στην πραγματικότητα η κίνηση των αντικειμένων διέπεται από περιορισμούς. Για παράδειγμα, ένας άνθρωπος δεν μπορει να περπατάει στον δρόμο με ταχύτητα που να ξεπερνάει τα 15 χμ/ώρα. Για να γίνουμε πιο συγκεκριμένοι παρουσιάζουμε τις εξής τρεις κατηγορίες κίνησης:

- Κίνηση χωρίς περιορισμούς, όπως για παράδειγμα, η κίνηση αεροπλάνων.
- Κίνηση με περιορισμούς, όπως για παράδειγμα η κίνηση πεζών όταν συναντούν φυσικά εμπόδια (λ.χ. βουνά, λίμνες) που επιβραδύνουν την κίνησή τους.
- Κίνηση σε ορισμένες τροχιές, όπως για παράδειγμα, η κίνηση των τρένων στο

σιδηροδρομικό δίκτυο. Τα τρένα μπορούν να βρίσκονται σε ορισμένα σημεία του δισδιάστατου χώρου, δηλαδή στις ράγες και δεν μπορούν να βγουν από εκεί.

Μελετώντας για κάθε περίπτωση τα χαρακτηριστικά της κίνησης των αντικειμένων που εξετάζονται μπορούμε να τα αξιοποιήσουμε ώστε να βρούμε κατάλληλες δομές δεικτοδότησης, διευκολύνοντας την επεξεργασία. Τα στίγματα των αντικειμένων δεν είναι σχεδόν ποτέ απολύτως ακριβή. Αυτό μπορεί να γίνεται είτε εσκεμμένα είτε να οφείλεται στον τρόπο λήψης και καταγραφής των θέσεών τους. Για λόγους προστασίας ιδιωτικότητας προσώπων, κανείς χρήστης μπορεί να μη θέλει να αποκαλύπτει επακριβώς το ακριβές στίγμα του. Επιπροσθέτως, το εγγενές σφάλμα λόγω μετρήσεων από GPS, RFID κτλ. και η αβεβαιότητα λόγω δειγματοληψίας για τις θέσεις του αντικειμένου στις χρονικές στιγμές μεταξύ δύο διαδοχικών λήψεων μειώνουν την ακρίβεια της θέσης του αντικειμένου. Για παράδειγμα, ένα κινούμενο αντικείμενο λόγω της ταχύτητάς του έχει απομακρυνθεί αρκετά από τη θέση που καταγράφηκε από την τελευταία που έστειλε το στίγμα του. Επιπλέον από τις συσκευές GPS, αλλά και από άλλες αιτίες όπως λ.χ. θόρυβος, παρεμβολές και διακοπές κατά τη διάρκεια της μετάδοσης, είναι αναπόφευκτο ότι στη διαδικασία καταγραφής των συντεταγμένων δημιουργούνται σφάλματα. Η πραγματική θέση του αντικειμένου δεν είναι δυνατόν να προσδιοριστεί με ακρίβεια, αλλά κυμαίνεται εντός κάποιων ορίων σφάλματος. Επιπρόσθετα, τα όρια σφάλματος των συσκευών εντοπισμού δεν είναι σταθερά, αλλά είναι δυνατόν να κυμαίνονται ανάλογα με άλλες συνθήκες όπως λ.χ. τις καιρικές συνθήκες και την ίδια τη γεωγραφική θέση του αντικειμένου λόγω πυκνής δόμησης.

Η κίνηση μπορεί να αναπαρασταθεί στο δισδιάστατο επίπεδο από δύο χωρικές (x, y) και μία χρονική συντεταγμένη (t). Πιο συγκεκριμένα η θέση ενός αντικειμένου δίνεται από μία πλειάδα της μορφής:

<*object id, timestamp, x coordinate, y coordinate*>, όπου :

- *object id* : Η ταυτότητα του κινούμενου αντικειμένου.
- *timestamp* : Το χρονόσημο στο οποίο γίνεται η καταγραφή.
- *x coordinate* : Συντεταγμένη ως προς τονxάξονα.
- *y coordinate* : Συντεταγμένη ως προς τονyάξονα.

Η συχνότητα λήψης και καταγραφής της θέσης ενός αντικειμένου είναι ένα σημαντικό ζήτημα για την απόδοση των συστημάτων παρακολούθησης κινούμενων αντικειμένων. Η συχνότητα αυτή δεν είναι απαραίτητα σταθερή και μπορεί να ποικίλει από αντικείμενο σε αντικείμενο. Η σταθερή και μεγάλη συχνότητα ενδεχομένως να παρουσιάζει αρκετά προβλήματα και να επιβαρύνει το σύστημα με πλεονάζουσα πληροφορία. Τόσο το κόστος αποστολής της θέσης όσο και το κόστος επεξεργασίας και διαχείρισης των πληροφοριών από τον κεντρικό υπολογιστή θα αυξανόταν. Για κινούμενα αντικείμενα, τα οποία μεταβάλλουν συχνά την κατεύθυνση της κίνησής τους, απαιτείται συχνότερη ανανέωση της θέσης τους για να αποτυπώνονται με μεγαλύτερη ακρίβεια οι λεπτομέρειες της τροχιάς τους.

Αντίθετα, για αντικείμενα που κινούνται με προκαθορισμένες τροχίες, απαιτείται μικρότερη συχνότητα δειγματοληψίας.

2.2.2 Ερωτήματα σε κινούμενα αντικείμενα

Δεν είναι όλα τα χωρικά ερωτήματα εφαρμόσιμα σε κινούμενα σημειακά αντικείμενα. Αυτό συμβαίνει κυρίως για τη θεώρησή τους ως αντικείμενα με αμελητέα έκταση. Τα ερωτήματα χωρίζονται σε δύο κύριες κατηγορίες : *τα ερωτήματα θέσης* (*coordinated-based* ή *location-based queries*) και τα *ερωτήματα τροχιάς* (*trajectory-based queries*)με τα τελευταία να διακρίνονται σε *τοπολογικά ερωτήματα* (*topological queries*) και ερωτήματα πλοήγησης (*navigational queries*).

Ερωτήματα θέσης

Τα ερωτήματα θέσης ασχολούνται με τη θέση των αντικειμένων σε κάποια χρονική στιγμή. Τα ερωτήματα είναι δυνατόν να αναφέρονται στο παρόν (ερωτήματα τρέχουσας θέσης), είτε να αναφέρονται στο βραχυπρόθεσμο μέλλον με τη χρήση κάποιας μεθόδου πρόβλεψης των μελλοντικών θέσεων. Για αναφορά στο παρελθόν ή στο μέλλον απαραίτητη θα ήταν η αποθήκευση των παρελθουσών θέσεων των αντικειμένων. Τα σημαντικότερα ερωτήματα θέσης είναι τα εξής [9, 10] :

- *Ερωτήματα περιοχής* (*range queries*). Για ένα χωρικό παράθυρο (λ.χ. ορθογώνια, τετραγωνική, κυκλική, πολυγωνική περιοχή κλπ) και ενός χρονικού παραθύρου , ζητούνται τα αντικείμενα, των οποίων η κίνηση περιέχεται εντός τους. Οι απαντήσεις επιστρέφονται με τη μορφή των ταυτοτήτων των αντικειμένων. Παράδειγμα: «ποιά αυτοκίνητα βρέθηκαν στο κέντρο της Αθήνας από τις 12:00 μέχρι τη 13:00·».

- *Ερωτήματα εγγύτερων γειτόνων* (*nearest neighbor queries*). Για ένα κινούμενο σημειακό αντικείμενο και ένα χρονικό παράθυρο, ζητούνται τα k α-ντικείμενα πλησίον του δοσμένου κατά τη διάρκεια του χρονικού παραθύρου. Ως απάντηση επιστρέφεται μία λίστα με τα αντικείμενα-γείτονες μαζί με τις σχετικές αποστάσεις από το σημείο αναφοράς σε αύξουσα ή φθίνουσα διάταξη. Παράδειγμα: «Επέστρεψε ως απάντηση τα δύο κοντινότερα οχήματα σε μένα.»

- *Ερωτήματα πυκνότητας* (*density queries*). Για μια δεδομένη τιμή κατωφλίου για την πυκνότητα των αντικειμένων και ενός χρονικού παραθύρου, ζητούνται οι περιοχές του χώρου, εντός των οποίων η πυκνότητα των κινούμενων αντικειμένων ξεπερνάει τη δοσμένη τιμή κατωφλίου κατά τη διάρκεια του χρονικού παραθύρου.

- *Ερωτήματα χρονικών τεμαχίων* (*time-sliced queries*). Αναζητούνται οι θέσεις του συνόλου των αντικειμένων κατά τη διάρκεια ενός χρονικού διαστήματος ή μιας χρονικής στιγμής.

Σχήμα 2.3: Κάνναβος 9×7 ως χωρικό ευρετήριο.

Τοπολογικά ερωτήματα

Τα τοπολογικά ερωτήματα είναι ερωτήματα τροχιάς, τα οποία εξετάζουν τις χωροχρονικές σχέσεις της τροχιάς των αντικειμένων με άλλα στατικά ή κινούμενα γειτονικά αντικείμενα ή με περιοχές του χώρου. Οι χωροχρονικές σχέσεις έχουν παρόμοια σημασιολογία με αυτές του τοπολογικού μοντέλου, οι οποίες υιοθετούνται από τα κλασικά συστήματα χωρικών βάσεων δεδομένων, επεκταμένες φυσικά κατά τη χρονική διάσταση. Για την εξακρίβωση μιας τοπολογικής σχέσης της τροχιάς ενός αντικειμένου με μια δοσμένη περιοχή του χώρου, ενδεχομένως να απαιτείται η εξέταση περισσότερων του ενός τμημάτων της τροχιάς. Παράδειγμα: «Επέστρεψε ως απάντηση όλα τα οχήματα που εισήλθαν στο κέντρο της Αθήνας από τις 12:00 μέχρι τη 13:00 και εξήλθαν από τις 14:00 μέχρι τις 15:00 .»

Ερωτήματα πλοήγησης

Τα ερωτήματα πλοήγησης είναι ερωτήματα τροχιάς με τη διαφορά ότι δίνονται απαντήσεις σε μεγέθη παραγόμενα από την τροχιά και όχι από τις αποθηκευμένες θέσεις. Χαρακτηριστικό παράδειγμα μεγέθους που μπορεί να υπολογιστεί από την τροχιά είναι η ταχύτητα ως πηλίκο της διανυθείσας απόστασης προς το αντίστοιχο χρονικό διάστημα. Συνήθως, οι αναζητούμενες τιμές των μεγεθών υπολογίζονται στα επιμέρους τμήματα της τροχιάς και στη συνέχεια υπολογίζεται κατά περίπτωση ο μέσος όρος ή η μέγιστη τιμή τους. Παράδειγμα: «Επέστρεψε ως απάντηση όλα τα οχήματα που κινούνται στο κέντρο της Αθήνας με ταχύτητα μεγαλύτερη από 80χμ/ώρα».

2.2.3 Δεικτοδότηση κινούμενων αντικειμένων

Η συγχονισμένη παρακολούθηση μεγάλου όγκου κινούμενων αντικειμένων δημιουργεί πολλά προβλήματα ως προς την προσπέλαση των πληροφοριών. Το γεγονός αυτό καθιστά τη δημιουργία κατάλληλων ευρετηρίων, για την γρήγορη προσπέλαση των πληροφοριών στο δίσκο, ώστε να αποφεύγονται οι καθυστερήσεις. Οι μέθοδοι προσπέλασης που έχουν προταθεί ακολουθούν κυρίως δύο τάσεις

Σχήμα 2.4: R-tree ως χωρικό ευρετήριο [12].

ανάλογα με τα χαρακτηριστικά της εφαρμογής, με γνώμονα τα δεδομένα (*data driven*) και με γνώμονα το χώρο (*space driven*). Οι μέθοδοι προσπέλασης της κατηγορίας με γνώμονα τα δεδομένα κρίνονται ακατάλληλες ως προς τις απαιτήσεις των ρευμάτων δεδομένων, αφού δεν εξασφαλίζουν γρήγορες και σταθερού χρόνου ενημερώσεις. Η κατηγορία περιλαμβάνει το σύνολο των ιεραρχικών δομών, οι οποίες διαμερίζουν το χώρο σε περιοχές ώστε κάθε μια να περιέχει ένα ανώτατο πλήθος αντικειμένων.

Η κατηγορία μεθόδων προσπέλασης με γνώμονα το χώρο χρησιμοποιείται συχνά προς τις απαιτήσεις του μοντέλου των ρευμάτων δεδομένων. Σύμφωνα με την τεχνική του κατακερματισμού (hashing) ο χώρος διαμερίζεται από ένα πλέγμα κελιών (κάνναβος). Κάθε σημειακό αντικείμενο βρίσκεται εντός των ορίων ενός κελιού. Κατά τη μετάβαση ενός αντικειμένου σε κάποιο γειτονικό κελί, διαγράφεται από το προηγούμενο και τοποθετείται στο καινούργιο. Σε κάθε κελί μπορεί να τοποθετηθεί απεριόριστο πλήθος αντικειμένων. Στο σχήμα 2.3 απεικονίζεται ένας δισδιάστατος τετραγωνικός χώρος, ο οποίος κατακερματίζεται από τον κάνναβο, τα κελιά του οποίου έχουν τετραγωνικό σχήμα.

Ο κάνναβος ουσιαστικά υλοποιείται από την αντίστοιχη συνάρτηση κατακερματισμού. Στην περίπτωση του σχήματος έχει εφαρμοσθεί γραμμικός κατακερματισμός δύο διαστάσεων με βαθμό κατάτμησης 9×7. Στη γενική περίπτωση επιλέγεται ο βαθμός κατάτμησης κάθε διάστασης χωριστά: έστω G_x ο βαθμός κατάτμησης της οριζόντιας διάστασης και G_y της κάθετης. Αν ο χώρος έχει μήκος width και ύψος height , τότε κάθε κελί θα έχει μήκος $dx = width/G_x$ και ύψος $dy = height/G_y$. Τότε, η συνάρτηση $f(x, y) = G_x\lfloor\frac{y}{dy}\rfloor + \lfloor\frac{x}{dx}\rfloor$ είναι συνάρτηση κατακερματισμού με πεδίο ορισμού το $[0, width) \times [0, height)$και πεδίο τιμών τους ακεραίους του διαστήματος $[0, G_xG_y - 1]$. Δηλαδή, η συνάρτηση μετασχηματίζει τις συντεταγμένες ενός σημείου του χώρου σε ένα ακέραιο αριθμό, ο οποίος είναι η ταυτότητα του κελιού στο οποίο ανήκει το σημείο. Στο σχήμα 2.3 φαίνονται οι ταυτότητες όλων των κελιών που προκύπτουν από την αντίστοιχη συνάρτηση κατακερματισμού για $G_x = 9, G_y = 7$.

Άλλες τεχνικές δεικτοδότησης κινούμενων αντικειμένων που βασίζονται κυρίως στο R-tree είναι το STR-tree (Spatio-temporal R-tree), το TB-tree (Trajectory-Bundle Tree), το TPR-tree (Time-Parameterized R-tree),το $R^{EXP} - tree$ και το STAR-tree. Η βασική ιδέα πίσω από το R-tree, που το συναντάμε γενικά στα χωρικά δεδομένα (βλέπε σχήμα 2.4), είναι η ομαδοποίηση αντικειμένων που βρίσκονται κοντά μεταξύ τους σε συστάδες. Στη συνέχεια οι συστάδες αυτές ομαδοποιούνται κατά τον ίδιο τρόπο κτλ. Οι παραλλαγές του R-tree προσπαθούν να καλύψουν την ανεπάρκεια του R-tree στην δεικτοδότηση κινούμενων αντικειμένων.

Κεφάλαιο 3

Διαχείριση δεδομένων με αβεβαιότητα

3.1 Εισαγωγή

Τα τελευταία χρόνια εμφανίστηκε ένα ευρύ φάσμα εφαρμογών που σχετίζονται με την *αβεβαιότητα (uncertainty)*. Τα αίτια της αβεβαιότητας στα δεδομένα ποικίλουν ανάλογα με την εφαρμογή. Για παράδειγμα, στην λειτουργία αισθητήρων η αβεβαιότητα οφείλεται στην ανακρίβεια λόγω σφαλμάτων κατά τη μέτρηση (π.χ. θερμοκρασία). Σε άλλες εφαρμογές, όπως είναι η προστασία της ιδιωτικότητας (privacy), υπάρχει η απαίτηση τα δεδομένα να είναι επίτηδες λιγότερο ακριβή. Η αβεβαιότητα συμβάλλει στην απόκρυψη προσωπικών δεδομένων προστατεύοντας ευαίσθητα χαρακτηριστικά των ατόμων, έτσι ώστε μικρότερο μέρος στοιχείων να μπορεί να δημοσιευτεί. Ένα άλλο παράδειγμα βρίσκεται στα συστήματα εντοπισμού στίγματος (GPS). Το στίγμα ενός αντικειμένου, π.χ. ενός αυτοκινήτου, δίνει την ακριβή θέση και ταχύτητά του κάθε χρονική στιγμή. Θα ήταν ασύμφορο για τον εντοπισμό του να στέλνει το στίγμα του πολύ συχνά, αφού με βάση την ταχύτητά του μπορούμε να ξέρουμε προσεγγιστικά πού βρίσκεται. Έτσι, η θέση του αντικειμένου όπως είναι γνωστή στο σύστημα, δεν ταυτίζεται πάντοτε με την τρέχουσα λόγω χρονικής υστέρησης κατά τη μετάδοση, οπότε θεωρείται αβέβαιη. Μέχρι πριν λίγα χρόνια, τα αβέβαια δεδομένα δεν είχαν καμία θέση στις παραδοσιακές, ακριβείς βάσεις δεδομένων, οι οποίες με την σειρά τους δεν ήταν προετοιμασμένες να τα αντιμετωπίσουν.

Σήμερα έχουν αναπτυχθεί πολλές μέθοδοι και αλγόριθμοι για καλύτερη επεξεργασία ερωτημάτων, πολλά από τα οποία τα συναντάμε γενικά στις β.δ. και εμπεριέχουν πλέον την αβεβαιότητα. Ο υπολογισμός των περισσότερων αλγορίθμων γίνεται με αριθμητικές μεθόδους, με αποτέλεσμα να προκύπτουν προσεγγιστικές λύσεις. Για την καλύτερη μοντελοποίηση των προβλημάτων αυτών επιλέγεται αντίστοιχα η κατάλληλη αναπαράσταση των δεδομένων. Σε άλλες εργασίες ακολουθείται ο συνεχής τρόπος αναπαράστασης υποθέτοντας διάφορες κατανομές (ομοιόμορφη, κανονική κτλ.) και σε άλλες ο διακριτός τρόπος με χρήση πεπερα-

	Όνομα	Τοποθεσία	Πιθανότητα P
t_1	Μάριος	κτίρια ΣΗΜΜΥ	$P_{11} = 0.3$
t_2	Μάριος	κτίρια ΣΕΜΦΕ	$P_{12} = 0.2$
t_3	Μάριος	παλιά κτίρια ΣΗΜΜΥ	$P_{13} = 0.5$
t_4	Γιάννης	παλιά κτίρια ΣΗΜΜΥ	$P_{21} = 0.7$
t_5	Γιάννης	κτίρια ΣΗΜΜΥ	$P_{22} = 0.3$

Πίνακας 3.1: Πιθανότητες ατόμων να βρίσκονται σε διάφορα μέρη του Πολυτεχνείου.

σμένου αριθμού διακριτών δειγμάτων. Τέλος, διάφορες ερευνητικές προσπάθειες εστιάζουν στην καλύτερη παρουσίαση των αποτελεσμάτων (π.χ. περιθώρια εμπιστοσύνης, εκτίμηση σφάλματος κτλ.), ώστε να παρέχεται στους χρήστες μία εποπτική εικόνα κατά πόσο μπορούν να βασιστούν στις απαντήσεις που δίνονται.

Στο κεφάλαιο αυτό γίνεται επισκόπηση της ευρύτερης έρευνας που έχει γίνει μέχρι στιγμής στον τομέα της αβεβαιότητας στις β.δ., εστιάζοντας κυρίως στους τομείς της αναπαράστασης, της επεξεργασίας και της τελικής παρουσίασης των στοιχείων.

3.2 Πιθανοτικά και αβέβαια δεδομένα

Για να μην δημιουργηθεί σύγχυση, αναφέρονται οι διαφορές ανάμεσα στις πιθανοτικές βάσεις δεδομένων και την αβεβαιότητα στις βάσεις δεδομένων. Η διάκριση αυτή γίνεται γιατί τα δύο μοντέλα παρουσιάζουν αρκετά κοινά στοιχεία, αφού και τα δύο εμπεριέχουν την έννοια της πιθανότητας, ωστόσο πρόκειται για δύο διαφορετικά πεδία έρευνας.

3.2.1 Πιθανοτικές βάσεις δεδομένων

Γενικά οι πιθανοτικές β.δ. ασχολούνται με την ύπαρξη και την μη-ύπαρξη αντικειμένων που είναι ακριβή. Για παράδειγμα, η πιθανοτική β.δ. του πίνακα 3.1 αντιπροσωπεύει την πιθανότητα καθενός από τα παραπάνω άτομα να βρίσκονται σε κάποιο μέρος του Πολυτεχνείου. Το άθροισμα των πιθανοτήτων κάθε ατόμου είναι ίσο με 1, αφού θεωρούμε ότι το άτομο υπάρχει και ξέρουμε ότι βρίσκεται κάπου μέσα στο Πολυτεχνείο.

Γνωρίζουμε ότι μία βάση δεδομένων που αναπαριστά πλήρη πληροφορία, μοντελοποιεί ένα μέρος του πραγματικού κόσμου. Αντίθετα, μία βάση δεδομένων που αναπαριστά μη πλήρη πληροφορία, στην ουσία αναπαριστά ένα σύνολο *πιθανών κόσμων (possible worlds)* [6]. Ένας πιθανός κόσμος είναι μια υποθετική αναπαράσταση του πραγματικού κόσμου και μπορεί να αναπαρασταθεί από μια βάση δεδομένων πλήρους πληροφορίας. Με άλλα λόγια, είναι η απεικόνιση όλων των στιγμιοτύπων που μπορούν να προκύψουν από μία πιθανοτική β.δ. Ειδικότερα, πιθανοτική βάση δεδομένων είναι ένα ζεύγος (W, P), όπου $W = W_1, W_2, \ldots, W_n$, είναι

Σχήμα 3.1: Περιοχές αντικειμένων που προσδιορίζονται από κάποια συνάρτηση πυκνότητας πιθανότητας.

ένα σύνολο από στιγμιότυπα στη βάση δεδομένων και P : W[0,1] είναι μία πιθανοτική κατανομή με $\sum P(W_j) = 1$. Κάθε στιγμιότυπο W_j,για το οποίο $P(W_j) > 0$, καλείται πιθανός κόσμος.

i	W_i	$P(W_i)$
1	0	$(1 - P_{11} - P_{12} - P_{13})(1 - P_{21} - P_{22}) = 0$
2	t_1	$P_{11}(1 - P_{21} - P_{22}) = 0$
3	t_2	$P_{12}(1 - P_{21} - P_{22}) = 0$
4	t_3	$P_{13}(1 - P_{21} - P_{22}) = 0$
5	t_4	$(1 - P_{11} - P_{12} - P_{13})P_{21} = 0$
6	t_1t_4	$P_{11}P_{21} = 0.21$
7	t_2t_4	$P_{12}P_{21} = 0.14$
8	t_3t_4	$P_{13}P_{21} = 0.35$
9	t_5	$(1 - P_{11} - P_{12} - P_{13})P_{22} = 0$
10	t_1t_5	$P_{11}P_{22} = 0.09$
11	t_2t_5	$P_{12}P_{22} = 0.6$
12	t_3t_5	$P_{13}P_{22} = 0.15$

Πίνακας 3.2: Πιθανοί κόσμοι

Για τα στοιχεία του πίνακα 3.1 οι πιθανοί κόσμοι φαίνονται στον πίνακα 3.2. Έτσι, στο i =7 βλέπουμε την πιθανότητα ο Μάριος να βρίσκεται στα κτίρια ΣΕΜΦΕ και ταυτόχρονα ο Γιάννης να βρίσκεται στα παλιά κτίρια ΣΗΜΜΥ, ενώ στο i =10 βλέπουμε την πιθανότητα ο Μάριος και ο Γιάννης να βρίσκονται ταυτόχρονα στα κτίρια ΣΗΜΜΥ.

Σχήμα 3.2: Συνάρτηση πυκνότητας-πιθανότητας αντικειμένου.

3.2.2 Βάσεις δεδομένων με αβεβαιότητα

Σε αντίθεση με τις πιθανοτικές β.δ., η αβεβαιότητα στις β.δ. ασχολείται με την ύπαρξη οντοτήτων, των οποίων η κατάσταση είναι μη ακριβής/σίγουρη. Για παράδειγμα, σε χωρικά δεδομένα η οντότητα αντιπροσωπεύεται από αντικείμενα και η κατάσταση ενός αντικειμένου από τη θέση του. Στον πίνακα 3.1 η διαφορά αυτή μεταξύ των δύο συγκρινόμενων μοντέλων φαίνεται από το γεγονός ότι σε μία πιθανοτική β.δ. ο Μάριος βρίσκεται στα κτίρια ΣΗΜΜΥ (ακριβής θέση) με πιθανότητα P_{11}=30/100. Η πιθανότητα αυτή αναφέρεται στην ύπαρξη ή την μή ύπαρξη του Μάριου στα κτίρια ΣΗΜΜΥ. Στις β.δ. με αβεβαιότητα, η ύπαρξη του Μάριου στην θέση αυτή θα ήταν δεδομένη και η ενασχόλησή μας θα επικεντρωνόταν στο ποια είναι η πιθανότητα ο Μάριος να βρισκόταν μέσα, κοντά ή και ενδεχομένως μακριά από τα κτίρια ΣΗΜΜΥ. Η θέση του γύρω από το κτίριο θα προσδιοριζόταν από μία κατανομή. Ο προσδιορισμός της κατάστασης γίνεται με πιθανοτικό τρόπο.

Για την καλύτερη κατανόηση των β.δ. με αβεβαιότητα δίνεται ένα παράδειγμα που βασίζεται σε χωρικά δεδομένα. Δοθέντος ενός σημείου X, θα θέλαμε να γνωρίζουμε ποιο από τα δύο κινούμενα αντικείμενα Y και Z είναι ο εγγύτερος γείτονας του X. Τα αντικείμενα Y, Z βρίσκονται σε δύο περιοχές (έστω κυκλικές) και η θέση τους μέσα στις περιοχές αυτές προσδιορίζεται από κάποια συνάρτηση πυκνότητας πιθανότητας (σχήμα 3.1). Ας υποθέσουμε ότι υπάρχουν εγγυήσεις ότι τη στιγμή που το ερώτημα αποτιμάται, τα Yκαι Z δεν μπορεί να είναι σε απόσταση dy και dz (ακτίνες κύκλων) αντίστοιχα από τις θέσεις που έχουν αποθηκευτεί στη βάση δεδομένων. Μπορούμε να πούμε με σιγουριά ότι ο Y είναι πλησιέστερα στον X, αν η ελάχιστη δυνατή απόσταση του Y από το X είναι λιγότερη από την αντίστοιχη απόσταση του Z από το X. Γενικά όμως ισχύει ότι η αβεβαιότητα των αντικειμένων μπορεί να μην μας επιτρέψει να καθορίσουμε με σιγουριά ποιο αντικείμενο είναι ο εγγύτερος γείτονας του X. Έτσι, κάθε αντικείμενο θα έχει διαφορετική πιθανότητα να είναι ο εγγύτερος γείτονας του X .

3.3 Αναπαράσταση Αβεβαιότητας

3.3.1 Μορφές αβεβαιότητας

Η ανάγκη για διαχείριση της αβεβαιότητας των δεδομένων στις β.δ. με τρόπο διαφανή προς τον χρήστη, οδήγησε στην ανάπτυξη ευέλικτων τρόπων αναπαράστασης των δεδομένων. Τα δεδομένα στις β.δ. με αβεβαιότητα, από το [17, 23],

Σχήμα 3.3: Διακριτά δείγματα.

διακρίνονται σε δύο κατηγορίες:

- στην *αβεβαιότητα πλειάδων (tuple uncertainty)*, η οποία χρησιμοποιείται κυρίως για την μοντελοποίηση πιθανοτικών σχεσιακών δεδομένων σε πιθανοτικές β.δ. Στην αβεβαιότητα πλειάδων, η παρουσία μιας πλειάδας σε μια σχέση είναι πιθανοτική και πολλαπλές πλειάδες μπορεί να έχουν περιορισμούς μεταξύ τους, όπως ο αμοιβαίος αποκλεισμός (mutually exclusive tuples).
- στην *αβεβαιότητα σε γνωρίσματα πλειάδων (attribute uncertainty)*, η οποία χρησιμοποιείται στις β.δ. με αβεβαιότητα. Αντίθετα με την αβεβαιότητα πλειάδων, τώρα κάθε πλειάδα θεωρείται ότι ανήκει στη β.δ., αλλά μία ή περισσότερες από τις ιδιότητές της δεν είναι γνωστές με βεβαιότητα. Ειδικότερα, κάθε αβέβαιο αντικείμενο μοντελοποιείται από μια περιοχή αβεβαιότητας, ε-ντός της οποίας η ύπαρξη του αντικειμένου περιγράφεται από κάποια πιθανοτική κατανομή. Η κατανομή του αντικειμένου μπορεί να περιγραφεί είτε από μια συνάρτηση πυκνότητας-πιθανότητας (*συνεχείς κατανομές*), είτε από *διακριτά δείγματα*. Η επιλογή της κατανομής εξαρτάται από το τί θέλουμε να μοντελοποιήσουμε και από τις δυνατότητες του συστήματος.

3.3.2 Μοντέλο συνεχούς κατανομής

Στις χωρικές β.δ. η αβεβαιότητα ενός αντικειμένου μπορεί να αναπαρασταθεί ως εξής: Η *περιοχή αβεβαιότητας (uncertain region)* ενός αντικειμένου O_i σε χρόνοt , συμβολίζεται με $U_i(t)$ και είναι μια κλειστή περιοχή, έτσι ώστε το O_i να βρίσκεται πάντα μέσα σε αυτή την περιοχή. Η συνάρτηση πυκνότητας-πιθανότητας ενός αντικειμένου O_i , συμβολίζεται με $f_i(x, y, t)$ και εκφράζει την πιθανότητα το αντικείμενο να βρίσκεται στην θέση (x, y) την χρονική στιγμή t , όπως φαίνεται στο σχήμα 3.2. Ενδεικτικά, κάποιες συνεχείς κατανομές που συναντάμε συχνά είναι :

- Η ομοιόμορφη κατανομή
- Η κανονική κατανομή με κατάλληλη διασπορά και μέση τιμή
- Κατανομές Zipf, Poisson για στοχαστικά μοντέλα που έχουν να κάνουν με την περιγραφή της συχνότητας εμφάνισης κάποιων συμβάντων.

Κάθε σημείο εκτός της περιοχής $U_i(t)$ έχει πιθανότητα 0 στη συνάρτηση πυκνότητας-πιθανότητας $f_i(x, y, t)$ για την οποία ισχύει η ιδιότητα $\int_{U_i(t)} f_i(x, y, z)dxdy = 1$.

Σχήμα 3.4: Τομή ορθογωνικής περιοχής με αβέβαιο αντικείμενο.

3.3.3 Μοντέλο διακριτών δειγμάτων

Σε αντιπαράθεση με το προηγούμενο παράδειγμα, πάλι στις χωρικές β.δ., η αβεβαιότητα ενός αντικειμένου X για μία χρονική στιγμή ι μπορεί να αναπαρασταθεί με την χρήση διακριτών δειγμάτων, έστω $x_1[i], x_2[i], \ldots, x_n[i]$ όπως στο σχήμα 3.3. Κάθε διακριτό δείγμα $x_j[i]$ βρίσκεται σε μία θέση και αντιπροσωπεύει την πιθανότητα το αντικείμενο X να βρίσκεται σε εκείνη τη θέση. Το αντικείμενο έχει πιθανότητα 0 να βρίσκεται σε άλλη θέση εκτός των δοσμένων διακριτών δειγμάτων του. Φυσικά, σημειώνουμε ότι το άθροισμα των πιθανοτήτων όλων των δειγμάτων είναι 1.

3.3.4 Προσομοίωση Monte Carlo

Ο αλγόριθμος Monte Carlo αποτελεί μια αριθμητική μέθοδο για εύκολο μεν, αλλά και γρήγορο προσεγγιστικό υπολογισμό ολοκληρωμάτων. Με αυτόν τον τρόπο δεν χρειάζεται να λύσουμε αναλυτικά ολοκληρώματα που μέσω της συνάρτησης πυκνότητας πιθανότητας δίνουν ακριβές αποτέλεσμα, κερδίζοντας σε απόδοση και χρόνο. Πιο συγκεκριμένα, έστω ότι έχουμε ένα σημειακό αντικείμενο o που βρίσκεται σε μία περιοχή αβεβαιότητας o.ur , με συνάρτηση πυκνότητας πιθανότητας o.pdf(x). Έστω ακόμα ότι έχουμε μία ορθογώνια περιοχή r_q με κατώφλι $p_q \in [0, 1]$, η οποία έχει κοινά σημεία με την $o.ur$. Θέλουμε να υπολογίσουμε ποιά είναι η πιθανότητα P_{app} το σημειακό αντικείμενο αυτό να βρίσκεται μέσα στο r_q(σχήμα 3.4). Η ύπαρξη του αντικειμένου μέσα στο χώρο αυτό προσδιορίζεται από το ολοκλήρωμα: $P_{app}(o, q) = \int_{o.ur \bigcap r_q} o.pdf(x)dx$.

Χρησιμοποιώντας τον αλγόριθμο Monte Carlo μπορούμε να διαλέξουμε τυχαία σημεία που βρίσκονται μέσα στην περιοχή $o.ur$, έστω $x_1, x_2, \ldots, x_{n1}$. Χωρίς βλάβη της γενικότητας, υποθέτουμε ότι n_2 από τα σημεία αυτά πέφτουν μέσα στο ορθογώνιο, έστω $x_1, x_2, \ldots, x_{n2}$. Για κάθε ένα από τα σημεία x_i υπολογίζουμε την συνάρτηση πυκνότητας πιθανότητάς. Τότε η πιθανότητα P_{app} το σημειακό αντικείμενο να βρίσκεται μέσα στην ορθογώνια περιοχή δίνεται από τον τύπο: $P_{app} \approx \frac{\sum_{i=1}^{n_2} o.pdf(x_i)}{\sum_{i=1}^{n_1} o.pdf(x_i)}$. Προφανώς, όσο περισσότερα δείγματα ληφθούν, τόσο ακριβέστερη γίνεται η εκτίμηση της πιθανότητας.

3.4 Επεξεργασία ερωτημάτων

Στην ενότητα αυτή παρουσιάζονται οι βασικότεροι τύποι ερωτημάτων καθώς και διάφορα ευρετήρια που συναντάμε συχνά στις β.δ. με αβεβαιότητα. Υπάρχουν τρία

Σχήμα 3.5: Παράδειγμα R-δέντρου.

θέματα στην επεξεργασία των δεδομένων που εξετάζονται πάνω σε βάσεις δεδομένων με αβεβαιότητα. Πρώτον, πρέπει να παρέχεται εγγύηση για την ακρίβεια των απαντήσεων. Δεύτερον, χρειάζεται σχεδιασμός αποτελεσματικών προσεγγίσεων για την διαχείριση μεγάλου όγκου αβέβαιων δεδομένων με μικρό χρόνο απόκρισης των απαντήσεων. Τρίτον, δεδομένου ότι το μέγεθος της διαθέσιμης μνήμης είναι συχνά περιορισμένο, προτιμώνται τεχνικές επεξεργασίας με μικρή κατανάλωση μνήμης. Πάνω σε αυτούς τους τρεις άξονες κινούνται όλες οι μεθοδολογίες, τεχνικές και αλγόριθμοι που ασχολούνται με επεξεργασία αβέβαιων δεδομένων. Οι κυριότερες από αυτές αναφέρονται στη συνέχεια με κατάλληλα παραδείγματα για την καλύτερη κατανόησή τους.

3.4.1 Ευρετήρια

Όπως και στις συμβατικές β.δ., για την επεξεργασία μεγάλου όγκου δεδομένων με αβεβαιότητα χρησιμοποιούμε ευρετήρια για την διευκόλυνση ανάκτησης υποψήφιων απαντήσεων. Τα ευρετήρια αποτελούν κυρίως μέρος της *φάσης φιλτραρίσματος (filtering phase)* όπου δίνεται προσεγγιστική λύση στο πρόβλημα, σε αντίθεση με την *φάση εκλέπτυνσης (refinement phase)*, στην οποία δίδεται η τελική λύση.

Στις β.δ. με αβεβαιότητα, όπως επίσης και στις πιθανοτικές β.δ., το ευρετήριο που χρησιμοποιείται περισσότερο κάθε άλλο (μαζί με παραλλαγές του) από είναι το πιθανοτικό *R-δέντρο* (*probabilistic R-tree*). Το πιθανοτικό *R-δέντρο* αποτελεί μία γενίκευση του *B-δέντρου* και είναι μία πολύ εύχρηστη δομή για οργάνωση κυρίως χωρικών δεδομένων. Ειδικότερα, στα χωρικά δεδομένα τα δείγματα των αντικειμένων που βρίσκονται κοντά το ένα με το άλλο οργανώνονται σε *συστάδες* (*clusters*) και εμπεριέχονται σε ορθογώνια που ονομάζονται ελάχιστα περιβάλλοντα ορθογώνια ή αλλιώς MBR (*minimum bounding rectangles*). Στη συνέχεια, τα MBR που βρίσκονται κοντά μεταξύ τους εμπεριέχονται μέσα σε άλλα MBR, με την διαδικασία αυτή να συνεχίζεται ωσότου όλα τα αντικείμενα να βρίσκονται μέσα σε ένα τελικό ορθογώνιο. *Το πιθανοτικό R-δέντρο* μοιάζει με το *R-δέντρο* που χρησιμοποιείται γενικά στα χωρικά δεδομένα με τη διαφορά ότι η πληροφορία για τις πιθανότητες των δειγμάτων τηρείται στο δέντρο. Πιθανότητες που υπολογίζονται με βάση τα στοιχεία που εμπεριέχουν, συνοδεύουν και όλα τα ενδιάμεσα MBR.

Για παράδειγμα, στο σχήμα 3.5 βλέπουμε πώς οργανώνονται τα δείγματα a_i, b_i, c_i, d_i, κάποιων αβέβαιων αντικειμένων A, B, C, D που αντιπροσωπεύουν την πιθανότητα τα αντικείμενα αυτά να βρίσκονται σε αυτές τις θέσεις. Το δείγμα b_4 εμπεριέχεται

στο ορθογώνιο E_3 μαζί με το a_8 . Το ορθογώνιοE_3 με τη σειρά του εμπεριέχεται στο ορθογώνιο E_1 το οποίο βρίσκεται μέσα στο ορθογώνιο R_3 , το οποίο μαζί με το R_2 και τα αντικείμενα a_1 και a_{14} αποτελούν το σύνολο των δειγμάτων (θεωρητικά ένα μεγάλο MBR, το οποίο δεν φαίνεται στο σχήμα, και τα περιέχει όλα).

Εκτός του πιθανοτικού R-δέντρου στην βιβλιογραφία συναντάμε και άλλα ευρετήρια ανάλογα με το πρόβλημα που αντιμετωπίζεται. Στην [25] παρουσιάζεται το U-δέντρο, μία πολυδιάστατη μέθοδος ανάκτησης αβέβαιων δεδομένων που ακολουθούν αυθαίρετες κατανομές. Η δομή αυτή ελαχιστοποιεί τους πιθανοτικούς υπολογισμούς σε ερωτήματα περιοχής. Διαισθητικά, αυτό επιτυγχάνεται με τον προϋπολογισμό κάποιας βοηθητικής πληροφορίας για κάθε αντικείμενο, η οποία μπορεί να χρησιμοποιηθεί για τον αποκλεισμό ή την επικύρωση υποψηφίων αντικειμέ-νων χωρίς να χρειάζεται να λάβει υπόψη την πιθανότητα εμφάνισής τους. Τέτοιες πληροφορίες διατηρούνται σε όλα τα επίπεδα του U-δέντρου ώστε να αποφεύγεται η πρόσβαση σε υποδέντρα που δεν περιέχουν κανένα αποτέλεσμα. Επιπλέον, τα U-δέντρα είναι πλήρως δυναμικά, δηλαδή τα αντικείμενα μπορούν να εισάγονται ή να διαγράφονται με οποιαδήποτε σειρά.

3.4.2 Βασικά ερωτήματα

Top-K

Τα Top-K ερωτήματα ανακτούν τα κορυφαία K στοιχεία (π.χ. K μεγαλύτερα ή K μικρότερα) από μια συλλογή, π.χ. για να βρούμε τους 10 πιο ψηλούς ανθρώπους ανάμεσα σε ένα σύνολο ανθρώπων. Ένας αλγόριθμος Top-K είναι επίσης γνωστός ως αλγόριθμος κατωφλίου, αφού τερματίζεται όταν ένα συγκεκριμένο όριο επιτυγχάνεται. Στο [21] περιγράφεται μία προσέγγιση για τον αποδοτικό υπολογισμό και κατάταξη των Top-K απαντήσεων ενός SQL ερωτήματος σε μία πιθανοτική β.δ. Ένας απλός υπολογισμός των Top-K απαντήσεων θα υπολόγιζε όλες τις πιθανότητες και στη συνέχεια θα επέλεγε τα Top-K. Αντ΄ αυτού, εδώ υπολογίζονται μόνο οι πιθανότητες που εξασφαλίζουν σε μεγάλο βαθμό ότι (α) οι μέχρι τότε επιστρεφόμενες Top-K απαντήσεις είναι οι σωστές και (β) η κατάταξη των Top-K απαντήσεων είναι σωστή. Ο αλγόριθμος διεξάγει παράλληλα πολλές προσομοιώσεις Monte Carlo, μία για κάθε υποψήφια λύση, κρατώντας μόνο τις λύσεις με τις καλύτερες πιθανότητες. Με την προσέγγιση αυτή αποφεύγονται άσκοποι υπολογισμοί, όπως π.χ. για πιθανότητες που σίγουρα δεν περιλαμβάνονται μέσα στις Top-K. Τέλος αποδεικνύεται, τόσο μαθηματικά όσο και πειραματικά, ότι η τεχνική αυτή είναι σχεδόν βέλτιστη (near optimal).

Σύνδεση βάσει ομοιότητας (Similarity Join)

Στις β.δ. με αβεβαιότητα υπάρχει ενδιαφέρον για τη σύνδεση στοιχείων μεταξύ δύο πινάκων βάσει της ομοιότητάς τους σε συγκεκριμένα γνωρίσματα, δηλαδή σε παρόμοια χαρακτηριστικά, χωρίς να είναι απαραίτητο τα στοιχεία να ταυτίζονται απολύτως. Ένα παράδειγμα βρίσκεται στην προστασία της δημόσιας ασφάλειας, όπου η τροχιά για κάθε ύποπτο εγκληματία παρακολουθείται από την αστυνομία συνεχώς σε πραγματικό χρόνο μέσω GPS. Ένα ερώτημα που μπορεί να τεθεί

Σχήμα 3.6: Παράδειγμα υπερσφαίρας.

και ζητάει απάντηση είναι π.χ. «εάν δύο κακοποιοί πρόσφατα πήγαν στον ίδιο τόπο μέσα σε σύντομο χρονικό διάστημα». Λόγω μειωμένης ακρίβειας του GPS ή καθυστερήσεων στη μετάδοση του στίγματος, οι πληροφορίες που δίνουν τη θέση μπορεί να μην αποκαλύπτουν ακριβώς την πραγματική θέση, καθιστώντας την αβέβαιη. Έτσι, στο συγκεκριμένο παράδειγμα, χρειάζεται να επεξεργαστούμε ένα θοιν το οποίο θα δίνει αποδοτικά και αποτελεσματικά τις λύσεις με την υψηλότερη εμπιστοσύνη (*high confidence*).

Όσον αφορά την αποτίμηση, στο [17] ορίζεται το πρόβλημα χρησιμοποιώντας διακριτά δείγματα για κάθε στοιχείο που εξετάζεται. Προτείνονται δύο τεχνικές κλαδέματος, ώστε να περιοριστούν οι πράξεις που εκτελούνται. Και στις δύο τεχνικές παρουσιάζεται η έννοια της υπερσφαίρας (*hypersphere*), η οποία είναι πρακτικά μία σφαίρα, με κέντρο και ακτίνα που ορίζονται ανάλογα με τα δείγματα, η οποία εμπεριέχει όλα τα διακριτά δείγματα. Στην πρώτη τεχνική (*Object-Level Pruning*) κλαδεύουμε τα ζεύγη των αντικείμενων των οποίων οι υπερφαίρες βρίσκονται σε απόσταση μεγαλύτερη από ένα κατώφλι απόστασης ε. Το ε (σχήμα 3.6 αριστερά) πρακτικά είναι η απόσταση ανάμεσα στα κέντρα των υπερσφαιρών χωρίς να υπολογίσουμε τις ακτίνες τους. Στη δεύτερη τεχνική (*Sample-Level Pruning*) παρουσιάζεται η έννοια της *(1-β)υπερφαίρας* (σχήμα 3.6 δεξιά) μέσα στην οποία το αντικείμενο περιέχεται με πιθανότητα *(1-β)*. Έτσι τώρα κλαδεύουμε όλα τα ζευγάρια αντικειμένων που οι *(1-β)*υπερσφαίρες τους βρίσκονται σε απόσταση μεγαλύτερη από ένα κατώφλι απόστασης ε και η πιθανότητα να βρίσκονται και τα δύο μέσα στις *(1-β)*υπερφαίρες τους είναι μεγαλύτερη από ένα πιθανοτικό κατώφλι *a*. Ακόμα, παρουσιάζεται μία στρατηγική, η οποία ακολουθεί ένα μοντέλο κόστους και έχει ως σκοπό την ελαχιστοποίηση του κόστους επεξεργασίας.

Χωρικά ερωτήματα

Σε αρκετές εργασίες, έχουν μελετηθεί αρκετά χωρικά ερωτήματα, τα οποία τα συναντάμε συχνά στην βιβλιογραφία. Μερικά από αυτά είναι τα εξής:

- *Πιθανοτικά ερωτήματα περιοχής*, όπου ζητείται να βρεθούν αντικείμενα που βρίσκονται εντός μιάς περιοχής με πιθανότητα μεγαλύτερη από ένα κατώφλι. Στο [4] γίνεται υπολογισμός τέτοιων ερωτήματων για κινούμενα αντικείμενα που ακολουθούν τυχαία κατανομή. Αρχικά με την χρήση του αθροίσματος Minkowski ως διευρυμένου ερωτήματος περιοχής επιτυγχάνεται κλάδεμα σε πιθανοτικά R-δέντρα, δομή στην οποία έχουν οργανωθεί τα κινούμενα

αντικείμενα. Στη συνέχεια, εναλλάσσοντας τους ρόλους ερωτημάτων και δεδομένων (*query-data duality*) απλοποιείται το κόστος υπολογισμού των πιθανοτήτων. Ο συνδυασμός με ένα σταθερό πιθανοτικό κατώφλι οδηγεί τελικά σε ακόμα καλύτερα αποτελέσματα κλαδέματος.

- *Πιθανοτικά ερωτήματα απόστασης*, όπου ζητείται να βρεθούν τα αντικείμενα που βρίσκονται σε απόσταση μικρότερη από δοσμένη απόσταση δ από ένα άλλο αντικείμενο, με πιθανότητα μεγαλύτερη από ένα κατώφλι.

- *Πιθανοτικά ερωτήματα χωρικών συνδέσεων (join).* Πιο συγκεκριμένα, δοσμένων δύο αβέβαιων πολυγώνων S_1, S_2 , μίας απόστασης d και ενός κατωφλίου $\gamma \in [0, 1]$, ένα πιθανοτικό ερώτημα χωρικών συνδέσεων επιστρέφει ένα σύνολο από τριάδες της μορφής (s_i, s_j, π_i), όπου $s_i \in S_1$ και $s_2 \in S_2$ απέχουν απόσταση μικρότερη από d με πιθανότητα $\pi_i \geq \gamma$.

- *Ερώτημα των k-εγγυτέρων γειτόνων (kNN).* Έστω σύνολο S αντικειμένων σε ένα χώρο. Θέλουμε να βρούμε τα k κοντινότερα σημεία $Q_i \in S$ που βρίσκονται πλησιέστερα στο σημείο Q. Οι περισσότερες εργασίες (λ.χ. [26, 20]) που ασχολούνται με το συγκεκριμένο ερώτημα αναφέρονται στη δημιουργία αριθμητικών προσεγγιστικών αλγορίθμων, χρησιμοποιώντας μεταξύ άλλων και τη μέθοδο Monte Carlo . Πιο συγκεκριμένα στο [26], αφού χρησιμοποιηθεί η μέθοδος Monte Carlo για να επιτύχουμε διακριτά δείγματα, χρησιμοποιείται ένα πιθανοτικό *R-δέντρο*. Για το δέντρο αυτό προσδιορίζονται οι κατάλληλες προσεγγίσεις για τα δείγματα των αντικειμένων μέσω της ομαδοποίησής τους σε συστάδες. Τα MBR που περικλείουν αυτές τις συστάδες χρησιμοποιούνται για τον εντοπισμό και παράλειψη περιττών υπολογισμών, βελτιώνοντας έτσι τις επιδόσεις της αποτίμησης του ερωτήματος.

- *Αντίστροφο ερώτημα των k-εγγυτέρων γειτόνων (RkNN)*, όπου ζητείται να βρεθούν ταk αντικείμενα που έχουν ένα αντικείμενο Q, ως κοντινότερο γείτονά τους με κάποια δοσμένη πιθανότητα.

Γενικά ο τρόπος που ακολουθείται για την επεξεργασία χωρικών ερωτημάτων είναι ο εξής: Αρχικά ορίζεται το μοντέλο αναπαράστασης που θα ακολουθηθεί. Τις περισσότερες φορές ακολουθείται η αναπαράσταση με διακριτά δείγματα όπως π.χ. έγινε στην [20] για την επίλυση πιθανοτικών ερωτημάτων χωρικών συνδέσεων. Στη συνέχεια ακολουθείται η τακτική filter and refinement. Κατά τη φάση φιλτραρίσματος χρησιμοποιείται ένα πιθανοτικό *R-δέντρο* ως ευρετήριο (Σχήμα 3.5) μέσω του οποίου δίνονται γρήγορα και προσεγγιστικά οι πρώτες λύσεις για το πρόβλημα. Αυτό γίνεται με τεχνικές κλαδέματος του δέντρου, με αποτέλεσμα μεγάλα κομμάτια του να μην εξετάζονται καθόλου. Κατά τη φάση εκλέπτυνσης αναπτύσσεται ένας αποδοτικός αλγόριθμος που υπολογίζει, αναλυτικά αυτή τη φορά, τα οριστικά αποτελέσματα.

Συνάθροιση (Aggregation)

Σε αυτά τα ερωτήματα η αλληλεπίδραση μεταξύ πολλών στοιχείων παίζει σημαντικό ρόλο στην αποτίμησή τους. Οι προκύπτουσες πιθανότητες επηρεάζονται πολύ από

Σχήμα 3.7: Παράδειγμα συνάρθροισης.

την αβεβαιότητα των χαρακτηριστικών των άλλων στοιχείων. Για παράδειγμα, στο σχήμα 3.7 η πιθανότητα ο x να έχει την ελάχιστη τιμή (θερμοκρασία) επηρεάζεται από την αντίστοιχη τιμή και τα όρια του y.

Ανάλογα με την φύση της απάντησης, τα συναθροιστικά ερωτήματα κατατάσσονται σε:

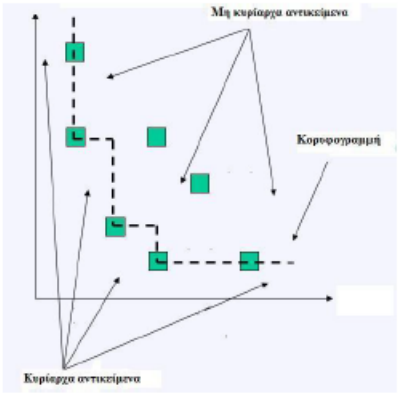
- *Συνάθροιση βάσει συνιστωσών (Entity-based Aggregate)*, η οποία επιστρέφει ένα σύνολο αντικειμένων που ικανοποιεί τις συνθήκες του ερωτήματος. Ένα χαρακτηριστικό παράδειγμα είναι το πιθανοτικό ερώτημα ελαχίστου (EMinQ), όπου επιστρέφεται ένα υποσύνολο πλειάδων (T_i, p_i) από τον πίνακα T , με το p_i να είναι η μη μηδενική πιθανότητα ότι η $T_i.a$ (τιμή γνωρίσματος a του T_i.) είναι η ελάχιστη μεταξύ όλων των στοιχείων του T.

- *Συνάθροιση βάσει τιμών (Value-based Aggregate)*, η οποία επιστρέφει μία μόνο τιμή, λ.χ. την μέση τιμή ενός υποσυνόλου των ενδείξεων του αισθητήρα. Χαρακτηριστικά παραδείγματα αποτελούν τα πιθανοτικό ερώτημα αθροίσματος (VSumQ) και μέσου όρου (VAvgQ) όπου επιστρέφεται μία συνάρτηση πυκνότητας πιθανότητας $p(x)$ της τυχαίας μεταβλητής T η οποία συμβολίζει το άθροισμα (αντίστοιχα το μέσο όρο), καθώς και ένα κλειστό διάστημα $[l, u]$ στο οποίο αυτή ισχύει, με $\int_l^u p(x)dx = 1$.

Στο [5] παρουσιάζονται μέθοδοι για τον υπολογισμό τέτοιων ερωτημάτων. Πιο συγκεκριμένα, ερωτήματα που ανήκουν στην κατηγορία συνάθροισης βάσει συνιστωσών (EMinQ, EMaxQ) ανάγονται στον υπολογισμό πιθανοτικών ερωτημάτων εγγυτέρων γειτόνων. Ακόμα, για ερωτήματα που ανήκουν στην κατηγορία συνάθροισης βάσει τιμών (VAvgQ, VsumQ) αποδεικνύεται ότι ο υπολογισμός τους είναι ουσιαστικά ο ίδιος. Για άλλα ερωτήματα της ίδιας κατηγορίας, όπως π.χ. για το VMinQ, μέρη του υπολογισμού τους ανάγονται στον υπολογισμό άλλων ερωτημάτων, όπως πχ του EMinQ, και άλλα μέρη τους υπολογίζονται ευριστικά με διάφορους αλγορίθμους.

Κορυφογραμμή

Τα ερωτήματα κορυφογραμμής (Skyline) έχουν αποδειχθεί χρήσιμο εργαλείο στη διαδικασία λήψης αποφάσεων. Λαμβάνοντας υπόψη ένα ορισμένο σύνολο δεδομένων D, για το οποίο δύο αντικείμενα s_1 και s_1 ανήκουν στο D, το αντικείμενο s_1 κυριαρχεί έναντι του άλλου s_2, εφόσον έχει τουλάχιστον ένα καλύτερο



Σχήμα 3.8: Κορυφογραμμή.

γνώρισμα και τα υπόλοιπα γνωρίσματά του δεν είναι χειρότερα εκείνων του s_2. Η κορυφογραμμή αποτελείται από όλα εκείνα τα αντικείμενα του συνόλου D έναντι των οποίων δεν μπορεί να κυριαρχήσει κανένα άλλο αντικείμενο που ανήκει στο D.

Για παράδειγμα, στις online παραγγελίες τα προϊόντα αξιολογούνται με βάση την τιμή τους, την κατάσταση και το όνομα της μάρκας τους. Έστω τα 4 προϊόντα της ίδιας μάρκας, θεωρώντας μόνο τις τρεις πρώτες στήλες του πίνακα 3.3. Τα προϊόντα X_2 και X_4 ανήκουν στην κορυφογραμμή επειδή το X_2 κυριαρχεί έναντι του X_1 και το X_4 κυριαρχεί έναντι του X_3 (λόγω τιμής). Τον υπολογισμό της κορυφογραμμής πάνω σε αβέβαια δεδομένα το συναντάμε σε πολλές εφαρμογές. Π.χ. έστω ότι στο προηγούμενο παράδειγμα θεωρήσουμε και μία επιπλέον στήλη που αντιπροσωπεύει την φερεγγυότητα του πωλητή που πουλάει το προϊόν. Η φερεγγυότητα του πωλητή μπορεί επίσης να χαρακτηριστεί ως η πιθανότητα (1=υψηλότερη τιμή, 0=χαμηλότερη τιμή) ότι το προϊόν εμφανίζεται ακριβώς, όπως περιγράφεται στη διαφήμιση, την παράδοση και την ποιότητά του.

Τώρα βλέπουμε ότι η φερεγγυότητα του X_1 είναι μεγαλύτερη από εκείνη του X_2. Το ίδιο ισχύει και για τα X_3 και X_4. Είναι ξεκάθαρο ότι πρέπει να υπάρχει ένας νέος υπολογισμός της κυριαρχίας των αντικειμένων έναντι των άλλων, λαμβάνοντας υπόψη μας τις πιθανότητες αυτές. Πολλές προσπάθειες έχουν γίνει για τον γρήγορο και αποτελεσματικό υπολογισμό skyline ερωτημάτων διαρκείας. Στο [31] εξετάζεται το πρόβλημα αποδοτικού υπολογισμού συνεχών skyline ερωτημάτων πάνω σε κυλιόμενα παράθυρα, με ρεύματα αβέβαιων δεδομένων, δοσμένης μιάς πιθανότητας κατωφλίου. Αρχικά, αφού γίνει η επιλογή των υποψήφιων στοιχείων που θα χρειαστούν στους υπολογισμούς, αναπτύσσονται αποδοτικές τεχνικές που βασίζονται στο R-δέντρο, αφού ο σκοπός είναι να γίνονται όσον το δυνατόν λιγότερες προσπελάσεις. Στη συνέχεια, επεκτείνονται οι τεχνικές αυτές σε εφαρμογές όπου δίδονται πολλαπλά πιθανοτικά κατώφλια (π.χ. ad-hoc ερώτημα κορυφογραμμής). Τέλος, δείχνεται ότι οι τεχνικές που αναπτύχθηκαν μπορούν επίσης να επεκταθούν για τον υπολογισμό της Top-K πιθανοτικής κορυφογραμμής σε κυλιόμενα παράθυρα.

3.5 Παρουσίαση αποτελεσμάτων

Στη βιβλιογραφία έχουν προταθεί διάφοροι τρόποι παρουσίασης των αποτελεσμάτων που προκύπτουν από την επεξεργασία των δεδομένων στις β.δ. με αβεβαιότητα. Οι τρόποι αυτοί ποικίλουν ανάλογα με τις τεχνικές επεξεργασίας που ακολουθούνται και με τους διατιθέμενους πόρους. Στην ενότητα αυτή δίδονται οι κυριότεροι τρόποι για παρουσίαση των απαντήσεων πέρα από τις απλές

Προϊόν	Τιμή	Κατάσταση	Φερεγγυότητα
X1	1000 euros	Εξαιρετική	0.9
X2	500 euros	Εξαιρετική	0.7
X3	300 euros	Καλή	1.00
X4	100 euros	Καλή	0.5

Πίνακας 3.3: Παράδειγμα Skyline

πιθανότητες και τα σφάλματα (errors), όπως είναι το *διάστημα εμπιστοσύνης*, η *ύπαρξη κατωφλίου* και *η κατάταξη των αποτελεσμάτων*.

3.5.1 Διαστήματα εμπιστοσύνης

Πολλές φορές, σε προβλήματα όπου οι υπολογισμοί οδηγούν σε τελικά αποτελέσματα που δεν μπορούν να δοθούν με μία απλή πιθανότητα, χρησιμοποιούμε διαστήματα εμπιστοσύνης. Για παράδειγμα, έστω ότι έχουμε 3 ανθρώπους x, y, z και 3 σημεία X, Y, Z (σημεία βάσει σχήματος 1) με τις αποστάσεις XY=2 και XZ=3. Έστω ότι ο x βρίσκεται στο X με πιθανότητα 1, ο y στο Y με πιθανότητα 0.1 (και σε διάφορα άλλα σημεία με άλλες πιθανότητες) σε και ο z στο Z με πιθανότητα 0.9 (και σε διάφορα άλλα σημεία με άλλες πιθανότητες). Επιθυμούμε να βρούμε ποιος από τους δύο αυτούς ανθρώπους (y, z) βρίσκεται πιθανότερα πλησιέστερα στον x. Τα διαστήματα εμπιστοσύνης μας δίνουν την δυνατότητα να μπορούμε να συγκρίνουμε τις δύο αυτές καταστάσεις και με αυτά εκφράζεται ο βαθμός της βεβαιότητας κατά πόσο κοντά βρίσκονται τα y, z στον x. Ανάλογα με την κατανομή που ακολουθείται, προσδιορίζεται η μέση τιμή και η διακύμανση του διαστήματος εμπιστοσύνης.

3.5.2 Χρήση κατωφλίων

Λαμβάνοντας υπόψη ένα όριο εμπιστοσύνης t, ένα ερώτημα με κατώφλι (Thresholding) επιστρέφει τα αντικείμενα (ή ζεύγη αντικειμένων, σε περίπτωση συνδέσεων) που μπορούν να επιλεγούν, επειδή συνοδεύονται από πιθανότητα τουλάχιστον t. Σε περιπτώσεις χωρικών ερωτημάτων (περιοχής, απόστασης και χωρικών συνδέσεων) τα τελικά αποτελέσματα που δίνονται πρέπει να ξεπερνούν ένα πιθανοτικό κατώφλι. Όσο μεγαλύτερη η τιμή του κατωφλίου τόσο λιγότερα αποτελέσματα θα παρουσιάζονται. Πιθανοτικά κατώφλια χρησιμοποιήθηκαν και σε άλλους τύπους ερωτημάτων, όπως στη σύνδεση βάσει ομοιότητας και στα συναθροιστικά ερωτήματα τιμής.

3.5.3 Κατάταξη

Για ένα ερώτημα Q σε μία πιθανοτική β.δ. με πιθανούς κόσμους W, θεωρητικά είναι εύκολο να δώσουμε όλα τα αποτελέσματα που υπάρχουν. Πρακτικά

όμως κάτι τέτοιο σε πολλές περιπτώσεις είναι αδύνατον ή και ασύμφορο, λόγω αποδοτικότητας. Η διαθέσιμη μνήμη και ο επεξεργαστής μπορεί να μην επιτρέπει κάτι τέτοιο. Ένα σημαντικό πρόβλημα στις πιθανοτικές β.δ. είναι καλύτερη παρουσίαση του συνόλου των πιθανών απαντήσεων του ερωτήματος στο χρήστη. Μία πρακτική προσέγγιση στο θέμα είναι η κατάταξη σε πλειάδες (ranking). Η παρουσίαση των αποτελεσμάτων με κατάταξη επιστρέφει ιεραρχημένα τα αντικείμενα (ή ζεύγη αντικειμένων, σε περίπτωση συνδέσεων) με βάση τα περιθώρια εμπιστοσύνης ή τις πιθανότητες που τα συνοδεύουν. Έτσι, το σύστημα επιστρέφει όλες τις δυνατές απαντήσεις και τις πιθανότητές τους $P_1, P_2, \ldots, P_n$) τις οποίες στη συνέχεια κατατάσσουμε και τέλος κρατάμε τις καλύτερες K (Top-K). Ένας τρόπος για να ταξινομήσουμε τις απαντήσεις-πλειάδες είναι κατά φθίνουσα σειρά με βάση τις πιθανότητές τους. Συχνά, βέβαια, μπορεί να υπάρξει η θέσπιση κάποιων κριτηρίων από τον χρήστη και έπειτα το σύστημα πρέπει να συνδυάσει κατάλληλα την κατάταξη του χρήστη, ώστε να δώσει τα αποτελέσματα εξόδου. Η ιεραρχημένη κατάταξη των αντικειμένων μπορεί επίσης να συνδυαστεί με χρήση κατωφλίου.

3.6 Τρέχοντα ερευνητικά ζητήματα

Τα τελευταία χρόνια η ερευνητική δουλειά στη διαχείριση δεδομένων με αβεβαιότητα, άρχισε να οδηγεί το χώρο σε μια σχετική ωρίμανση, δίνοντας την ευκαιρία για αξιοποίηση των δεδομένων με αβεβαιότητα σε πραγματικές εφαρμογές. Αυτή τη στιγμή υπάρχουν μια σειρά από επιστημονικά ζητήματα που απασχολούν την επιστημονική κοινότητα, όπως:

- η θεωρητική θεμελίωση πιο σύνθετων και χρήσιμων μοντέλων για διάφορες κατανομές.
- η σύνθεση αναλυτικών μοντέλων που θα υπολογίζουν με ακρίβεια το κόστος ερωτημάτων. Τέτοια μοντέλα μπορούν να χρησιμοποιηθούν για διευκόλυνση στην βελτιστοποίηση ερωτημάτων.
- η συσταδοποίηση και ταξινόμηση αβέβαιων δεδομένων, με ανάπτυξη αλγορίθμων εξόρυξης δεδομένων.
- η ανάπτυξη αποτελεσματικών τεχνικών σε προβλήματα ρευμάτων δεδομένων.
- η μελέτη αβέβαιων χωρικών δεδομένων, με ανάπτυξη ή και συνδυασμό αλγορίθμων για διάφορα προβλήματα, όπως τα kNN και $RkNN$.

Ιδιαίτερο ενδιαφέρον έχει η μελέτη ερωτημάτων σχετικά με το στίγμα μεγάλου πλήθους κινούμενων αντικειμένων, το οποίο λαμβάνει τη μορφή ρεύματος δεδομένων. Η ακρίβεια του στίγματος που γνωστοποιείται σε τρίτους δεν θα πρέπει να είναι βέβαιη για λόγους προστασίας. Έτσι, στο μέλλον, με την ανάπτυξη των κινητών συσκευών και των εφαρμογών κοινωνικής δικτύωσης η ύπαρξη τέτοιων εφαρμογών θεωρείται αναμενόμενη δίνοντας την δυνατότητα στους χρήστες να υποβάλλουν χωρικά *ερωτήματα διαρκείας (continuous queries)* για περιοχές ενδιαφέροντός τους, όπως π.χ. ποιός φίλος τους συχνάζει σε συγκεκριμένα μέρη με

συγκεκριμένη πιθανότητα. Τέτοιας φύσης είναι και η εφαρμογή που υλοποιείται στην παρούσα εργασία και ο αλγόριθμος της οποίας αναπτύσσεται στο επόμενο κεφάλαιο. Θεωρούμε ότι οι θέσεις των κινούμενων αντικειμένων που εξετάζονται είναι αβέβαιες και δεν γνωρίζουμε την ακριβή τους θέση. Πάνω σε τέτοια δεδομένα τίθενται πιθανοτικά ερωτήματα περιοχής με σκοπό την πιο έγκαιρη και όσο γίνεται έγκυρη αποτίμησή τους.

Κεφάλαιο 4

Επεξεργασία πιθανοτικών ερωτημάτων περιοχής

4.1 Εισαγωγή

Κύριο θέμα της παρούσας εργασίας αποτελεί η ανάπτυξη αλγορίθμου για αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων. Τα κινούμενα αντικείμενα διαθέτουν δυνατότητα γεωγραφικού εντοπισμού (GPS) όμως το στίγμα του κάθε αντικειμένου ποτέ δεν αποκαλύπτεται στον κεντρικό υπολογιστή. Ωστόσο, θεωρείται γνωστή η ευρύτερη περιοχή του. Η αβεβαιότητα της θέσης εκφράζεται με κάποια πιθανοτική κατανομή η οποία δεν θεωρείται ομοιόμορφη αλλά μπορεί να ποικίλλει. Οι χρήστες των κινούμενων αντικειμένων μπορούν να υποβάλλουν τα χωρικά ερωτήματα διαρκείας για περιοχές ενδιαφέροντός τους, οπότε ο επεξεργαστής οφείλει να συνεκτιμήσει τις σχετικές πιθανότητες των προσφάτως καταγεγραμμένων συμβάντων και να δίνει τακτικά ενημερωμένες προσεγγιστικές απαντήσεις με κυμαινόμενη ποιότητα.

Στο κεφάλαιο αυτό παρουσιάζεται ο αλγόριθμος που αναπτύχθηκε, με σκοπό την αποτίμηση ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων. Ο αλγόριθμος αυτός απαρτίζεται από δύο στάδια, ένα στάδιο προεπεξεργασίας και ένα στάδιο επεξεργασίας. Το στάδιο προεπεξεργασίας έχει ως σκοπό τα αποτελέσματα που παράγει να χρησιμοποιηθούν στο στάδιο επεξεργασίας. Το στάδιο επεξεργασίας αποτελείται από τα εξής μέρη:

- Ευρετήριο χωρικού πλέγματος (Grid Partitioning)
- Τρεις τεχνικές κλαδέματος (Pruning)

Και τα δύο αυτά μέρη ως σκοπό έχουν να μειώσουν το κόστος επεξεργασίας των δεδομένων. Η φιλοσοφία του αλγορίθμου βασίζεται σε δύο φάσεις, μία φάση φιλτραρίσματος (filtering phase) και μία φάση εκλέπτυνσης (refinement phase). Αναλυτικότερα τα αντικείμενα που εξετάζονται για κάθε ερώτημα φιλτράρονται από τις τρεις τεχνικές κλαδέματος που βασίζονται σε γεωμετρικά και πιθανοτικά

χαρακτηριστικά των αντικειμένων και των ερωτημάτων με αποτέλεσμα περιορισμένος αριθμός από αυτά να χρειάζεται να αποτιμηθεί περαιτέρω με μεγαλύτερη ακρίβεια. Η μεγαλύτερη αυτή ακρίβεια εισάγεται στην διαδικασία της εκλέπτυνσης όπου γίνεται και η αναλυτική και λεπτομερής αποτίμηση. Ο αλγόριθμος δεν παρέχει ακριβή αποτελέσματα αλλά είναι προσεγγιστικός. Ωστόσο, όπως προκύπτει κι από τα πειραματικά αποτελέσματα, προσεγγίζει με μεγάλη ακρίβεια τον εξαντλητικό αλγόριθμο (που παρέχει ακριβή πιθανοτικά αποτελέσματα) εξοικονομώντας σημαντικά σε χρόνο εκτέλεσης. Τέλος σημειώνεται ότι ο αλγόριθμος λειτουργεί online λαμβάνοντας περιοδικά είσοδο και εξάγοντας αποτελέσματα ανά κύκλους εκτέλεσης (κάθε timestamp).

4.2 Υποθέσεις

Οι υποθέσεις που γίνονται αφορούν τα γεωμετρικά και πιθανοτικά χαρακτηριστικά των αντικειμένων και των ερωτημάτων που τίθενται. Όλες αυτές οι υποθέσεις βασίστηκαν στην αρχική υπόθεση ότι τα αντικείμενα και τα ερωτήματα είναι ομοιόμορφα κατανεμημένα πάνω στο Ευκλείδειο επίπεδο.

4.2.1 Κινούμενα Αντικείμενα

Η θέση των αντικειμένων ανανεώνεται τακτικά, αλλά όχι με την ίδια περίοδο ανά αντικείμενο (timestamp). Γενικά, θεωρείται ότι ένα αντικείμενο , το οποίο το συμβολίζουμε με o_i, θα στείλει το στίγμα του όταν συμβεί μία σημαντική μεταβολή στην πορεία του λ.χ. μετακινηθεί 500μ. από την προηγούμενη θέση του. Τα αντικείμενα αναπαριστώνται με τη μορφή κύκλων που ακολουθούν την κανονική κατανομή δύο διαστάσεων (Circular Normal Distribution). Η επιλογή της κανονικής κατανομής δύο διαστάσεων έγινε λόγω του γεγονότος ότι μπορεί να μοντελοποιήσει με αποτελεσματικότητα το πρόβλημα που εξετάζουμε. Η θέση ενός κινούμενου αντικειμένου, διαισθητικά, μπορούμε να πούμε ότι αναπαρίσταται από ένα κέντρο και μία κυκλική περιοχή γύρω από αυτό που το μέγεθός της χαρακτηρίζεται από μία ακτίνα R . Η τιμή της ακτίνας R είναι σημαντική, όπως θα δούμε και στη συνέχεια, αφού καθορίζει το μέγεθος της περιοχής που η πραγματική θέση των αντικειμένων μπορεί να βρίσκεται. Η πιθανότητα της θέσης του αντικειμένου όσο περισσότερο κινούμαστε προς το κέντρο του κύκλου μεγαλώνει. Αντίθετα, όσο περισσότερο απομακρυνόμαστε μικραίνει. Η φυσική ερμηνεία αυτή την καθιστά ιδιαίτερα χρήσιμη για την αναπαράσταση της κατάστασης και της αβεβαιότητας των αντικειμένων έναντι άλλων κατανομών. Για παράδειγμα, η ομοιόμορφη κατανομή θα έδινε σε όλες τις θέσεις μιάς περιοχής ίσες πιθανότητες ύπαρξης του αντικειμένου χωρίς κάποιες θέσεις που βρίσκονται πιο κοντά στην πραγματική του θέση να αναπαριστώνται με μεγαλύτερα βάρη, πράγμα όχι επιθυμητό. Άλλες γνωστές κατανομές τόσο συνεχείς όσο και διακριτές όπως είναι η εκθετική, η γάμμα, η διωνυμική, η γεωμετρική, η Bernoulli ακόμα και τυχαία διακριτά δείγματα λόγω των ποιοτικών χαρακτηριστικών τους υπολείπονται έναντι της κυκλικής κανονικής κατανομής.

Σχήμα 4.1: Συνάρτηση πυκνότητας πιθανότητας κυκλικής κανονικής κατανομής

Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής δύο διαστάσεων δίνεται από τον τύπο:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2]}$$

όπου ρ ο συντελεστής συσχέτισης μεταξύ του x και του y. Στην συγκεκριμένη περίπτωση που έχουμε $\sigma_x = \sigma_y$ τότε το $\rho = 0$ και η παραπάνω σχέση γίνεται:

$$f(x,y) = \frac{1}{2\pi\sigma_x^2} e^{-\frac{1}{2}[(\frac{x-\mu_x}{\sigma_x})^2 + (\frac{y-\mu_y}{\sigma_y})^2]}$$

Αν υποθέσουμε ότι θέλουμε να απλοποιήσουμε περισσότερο την παραπάνω σχέση τότε μπορούμε να πάρουμε ως σημείο αναφοράς του (x,y)-συστήματος το $(\mu_x, \mu_y) = (0,0)$. Αυτό μπορεί πολύ εύκολα να επιτευχθεί αντικαθιστώντας τα x και τα y με το μετασχηματισμό $x^{'} = x - \mu_x$ και $y' = y - \mu_y$. Έτσι μπορούμε να ξαναγράψουμε τη συνάρτηση πυκνότητας πιθανότητας σε πολικές συντεταγμένες (r, θ) και με σημείο αναφοράς το $(x,y) = (0,0)$. Το σημείο (x,y) θα έχει πολικές συντεταγμένες με:

$$r = \sqrt{x^2 + y^2}$$

και

$$\theta = tan^{-1}(\frac{y}{x})$$

και η συνάρτηση πυκνότητας πιθανότητας γίνεται:

$$f(x,y) = \frac{1}{2\pi\sigma_x^2} e^{-\frac{1}{2}(\frac{r}{\sigma_x})^2}$$

Η γωνία θ δεν επηρεάζει τη συνάρτηση πυκνότητας πιθανότητας διότι η $f(x,y)$ εξαρτάται μόνο από την απόσταση του σημείου από το σημείο αναφοράς και όχι από την κατεύθυνσή του.

Στο σχήμα 4.1 παρουσιάζεται μια τρισδιάστατη αναπαράσταση της συνάρτησης πυκνότητας πιθανότητας $f(x,y)$ για τυχαία σ_x και σ_y . Το ύψος της επιφάνειας που σχηματίζεται δείχνει το μέγεθος της $f(x,y)$. Στο σχήμα 4.2 δίνεται ένα ακόμα

Σχήμα 4.2: Παράδειγμα συνάρτησης πυκνότητας πιθανότητας κυκλικής κανονικής κατανομής με $\sigma_x = \sigma_y = 1$

παράδειγμα συνάρτησης πυκνότητας πιθανότητας κυκλικής κανονικής κατανομής. Αυτή τη φορά η κατανομή έχει κέντρο στο $(\mu_x, \mu_y) = (0, 0)$ και οι τυπικές αποκλίσεις είναι $\sigma_x = \sigma_y = 1$.

Βασιζόμενοι σε όλα τα παραπάνω καταλήγουμε ότι τα αντικείμενα ακολουθούν την κανονική κατανομή δύο διαστάσεων (Circular Normal Distribution) με $\sigma_x = \sigma_y$ και $\mu_x = \mu_y$. Οι ακτίνες R_i των κύκλων των κανονικών κατανομών ποικίλουν και η τιμή τους λαμβάνεται από ένα σύνολο $\mathcal{R}$ που περιέχει n διαφορετικές προκαθορισμένες τιμές. Η επιλογή της ακτίνας δίνεται από τον χρήστη της εκάστοτε κινητής συσκευής και βασίζεται στο κατά πόσο θέλει να κρύψει την ακριβή του θέση. Εφόσον δεν τον ενδιαφέρει αρκετά να κρύψει τη θέση του επιλέγει μικρή ακτίνα, ενώ επιλέγει μεγαλύτερη ακτίνα στην περίπτωση που τον ενδιαφέρει να γίνεται αντιληπτός από μικρότερο αριθμό ερωτημάτων.

Σχήμα 4.3: Παράδειγμα ερωτημάτων με διάφορα κατώφλια και κατανομές αντικειμένων.

Σχήμα 4.4: Για $\lambda = 5$ ο πιθανότητες $p(b_k)$ των τετραγώνων b_k .

4.2.2 Κινούμενα Ερωτήματα

Η θέση των ερωτημάτων ανανεώνεται και αυτή κάθε κάποιο χρονόσημα (timestamp), την ίδια που ισχύει και για τα αντικείμενα. Τα ερωτήματα, τα οποία τα συμβολίζουμε με q_i, δίνονται με την μορφή ορθογωνίων. Οι διαστάσεις των ορθογωνίων αυτών δεν είναι απαραίτητα μεγαλύτερες από εκείνες των αντικειμένων. Η επιλογή του ορθογωνίου έγινε για λόγους ευκολίας. Ο υπολογισμός των πιθανοτικών αποτελεσμάτων βασίζεται στην τομή των ερωτημάτων με τα αντικείμενα. Ο υπολογισμός της τομής πολύπλοκων γεωμετρικών σχημάτων θα αύξανε κατά πολύ τόσο το κόστος επεξεργασίας όσο και δυσκολία υλοποίησης της εφαρμογής. Για το λόγο αυτό επιλέχτηκε ένα απλό γεωμετρικό σχήμα όπως το ορθογώνιο.

Κάθε κινούμενο ερώτημα το οποίο τίθέται συνοδεύεται και από ένα κατώφλι θ_i το οποίο εκφράζει την μικρότερη πιθανότητα που πρέπει να εμπεριέχει η τομή του με το κάθε αντικείμενο. Διαισθητικά, αποτελεί ένα κάτω όριο που η πιθανότητα της τομής των αντικειμένων με ένα ερώτημα πρέπει να υπερβαίνει ώστε να ληφθούν ως απάντηση. Τα κατώφλια θ_i των κύκλων των κανονικών κατανομών ποικίλουν και η τιμή τους παίρνεται από ένα σύνολο Θ που περιέχει m διαφορετικές προκαθορισμένες τιμές. Η επιλογή του κατωφλίου δίνεται από τον χρήστη της εκάστοτε κινητής συσκευής, που θέτει το ερώτημα και βασίζεται στο μέγεθος των απαντήσεων που θέλει να λάβει. Εφόσον τον ενδιαφέρει να συμπεριλάβει μόνο τα πολύ κοντινά στο ερώτημα αντικείμενα τότε επιλέγει μεγάλο κατώφλι. Στην περίπτωση που τον ενδιαφέρει να συμπεριλάβει και αντικείμενα των οποίων η πιθανότητα της τομής της κυκλικής κανονικής κατανομής τους με το ερώτημα θα είναι μικρή τότε επιλέγει μικρότερο κατώφλι. Ενδεικτικό παράδειγμα βλέπουμε στο σχήμα 4.3 όπου απεικονίζονται τέσσερα ερωτήματα με κατώφλια 0.6, 0.7, 0.8 και 0.95 αντίστοιχα. Διαισθητικά, βλέπουμε ότι για το ερώτημα με κατώφλι 0.95 δύο αντικείμενα στην περιοχή του δεν το τέμνουν καθόλου, ένα περιέχεται εξ ολοκλήρου μέσα του και για τα άλλα δύο γίνεται αναλυτικότερη αποτίμηση. Δεδομένου ότι η πιθανότητα κατωφλίου 0.95 είναι αρκετά μεγάλη ενδέχεται τα δύο αυτά αντικείμενα να μην δοθούν ως θετική απάντηση στην αποτίμηση του ερωτήματος.

4.3 Προεπεξεργασία Δεδομένων

4.3.1 Πρώτο Στάδιο: Διαίρεση κύκλων

Algorithm 1 Προεπεξεργασία

1: **Procedure** *SplitCircles*
2: **Είσοδος:** Υποθέτουμε σύνολο n προκαθορισμένων ακτίνων $\mathcal{R} = \{R_1, R_2, ..., R_n\}$.
3: **for** κάθε $R_i \in \mathcal{R}$ **do**
4: Κόψε κάθε κύκλο με ακτίνα R_i σε $\lambda \times \lambda$ τετράγωνα·
5: **for** κάθε κουτί b_k **do**
6: $p(b_k) \leftarrow$ πιθανότητα που περιέχει το b_k σύμφωνα με την κατανομή.·
7: **end for**
8: **end for**
9: **End Procedure**

Η προεπεξεργασία των δεδομένων χωρίζεται σε δύο στάδια. Ο σκοπός του πρώτου είναι να μετασχηματίσει την κυκλική κανονική κατανομή του κάθε αντικειμένου σε ένα τετράγωνο. Η διαδικασία αυτή έχει πολλά πλεονεκτήματα. Η τομή τετραγώνου με ορθογώνιο είναι υπολογίσιμη με πιο γρήγορο και εύκολο τρόπο απ΄ ό,τι η τομή κύκλου με ορθογώνιο. Με αυτόν τον τρόπο μειώνονται οι πράξεις που εκτελούνται. Το τετράγωνο που επιλέγεται να παραχθεί από τη θέση του κάθε κύκλου $((x, y), r)$ είναι το εγγεγραμμένο του τετράγωνο, το οποίο το συμβολίζουμε με $inbox(o_i)$. Ο μετασχηματισμός αυτός μπορεί να γίνει αφού έχει παρατηρηθεί ότι τα τέσσερα μέρη της κυκλικής κανονικής κατανομής που χάνονται (σχήμα 4.3 γραμμοσκιασμένα μέρη του κύκλου) δεν παίζουν ουσιαστικό ρόλο στους υπολογισμούς. Υπολογίστηκε ότι αποτελούν λιγότερο από το 5% της συνολικής κατανομής στην χειρότερη περίπτωση. Ακόμα και αν ένα ερώτημα τέμνει αυτή την περιοχή η πιθανότητα της τομής τους θα είναι πολύ μικρή, όχι σημαντική αν συλλογιστούμε ότι η μικρότερη πιθανότητα εμφάνισης άξια υπολογισμού είναι 50% ως μικρότερο κατώφλι που θέτουν τα ερωτήματα. Αρχικά γίνεται ο υπολογισμός των συντεταγμένων του τετραγώνου. Γεωμετρικά αυτό επιτυγχάνεται ως εξής: αν έχουμε (x, y) τις συντεταγμένες του κέντρου και r την ακτίνα, οι συντεταγμένες του τετραγώνου είναι $(x - \frac{r\sqrt{2}}{2}, y - \frac{r\sqrt{2}}{2})$ για κάτω αριστερά, $(x + \frac{r\sqrt{2}}{2}, y - \frac{r\sqrt{2}}{2})$ για κάτω δεξιά,$(x - \frac{r\sqrt{2}}{2}, y + \frac{r\sqrt{2}}{2})$ για πάνω αριστερά, $(x + \frac{r\sqrt{2}}{2}, y + \frac{r\sqrt{2}}{2})$ για πάνω δεξιά. Αφού υπολογιστούν οι συντεταγμένες, το εσωτερικό κάθε τέτοιου ορθογωνίου τεμαχίζεται σε $\lambda \times \lambda$ μικρότερα τετράγωνα, έστω b_k. Η μεταβλητή λ εκφράζει τον αριθμό των τετραγώνων που θα κοπεί το εγγεγραμμένο τετράγωνο. Όπως θα διαπιστωθεί και από τα πειραματικά αποτελέσματα, όσο μεγαλύτερη τιμή έχει τόσο ακριβέστερα τελικά αποτελέσματα μπορούμε να πάρουμε. Με το κόψιμο αυτό, γίνεται μία χοντρική εκτίμηση της αβεβαιότητας σε μικρά ¨κουτιά' (τετράγωνα). Κάθε κουτί εκφράζει την πιθανότητα το αντικείμενο να βρίσκεται μέσα σε αυτό. Με αυτόν τον τρόπο, η κυκλική κανονική κατανομή ¨διακριτοποιείται¨. Επειδή ο υπολογισμός αυτός είναι χρονοβόρος για να γίνεται κάθε φορά και λαμβάνοντας υπόψη ότι εξετάζουμε συγκεκριμένο εύρος από ακτίνες, προϋπολογίζουμε αρχικά τις πιθανότητες $p(b_k)$ που αντιστοιχούν στα εσωτερικά τετράγωνα b_k για όλα τα

Σχήμα 4.5: Για $\lambda = 7$ οι κανονικοποιημένες πιθανότητες $p(b_k)$ των εσωτερικών τετραγώνων b_k .

εύρη των ακτίνων(λόγω της συμμετρίας αφού $\sigma_x = \sigma_y$ τα νούμερα ταυτίζονται). Ο προϋπολογισμός αυτός γίνεται με τη προσομοίωση Monte Carlo και θα εξηγηθεί αναλυτικότερα στη συνέχεια. Στο σχήμα 4.4 για $\lambda = 5$ φαίνεται ο τεμαχισμός που έγινε. Αφού υπολογιστούν οι πιθανότητες αυτές, στην συνέχεια κανονικοποιούνται προκειμένου το άθροισμά τους να είναι μονάδα. Η κανονικοποίηση αυτή φαίνεται για $\lambda = 7$ στο σχήμα 4.5. Ο σκοπός του πρώτου σταδίου της προεπεξεργασίας είναι η διακριτοποίηση της κανονικής κατανομής για τις συγκεκριμένες ακτίνες έτσι ώστε να αποφεύγεται ο υπολογισμός πιθανοτήτων σε κάθε τομή ερωτήματος με αντικείμενο, ο οποίος θα είχε υπερβολικό κόστος.

4.3.2 Δεύτερο στάδιο: Υπολογισμός εμβαδών κατωφλίων

Algorithm 2 Προεπεξεργασία

1: **Procedure** *CalculateThresholdAreas*
2: **Είσοδος:** Υποθέτουμε σύνολο m προκαθορισμένων κατωφλίων $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$.
3: **for** κάθε $R_i \in \mathcal{R}$ **do**
4: **for** κάθε $\theta_j \in \Theta$ **do**
5: $A \leftarrow \{a_k, a_k \subset$ κύκλους με κέντρα $R_i, P(a_k) \leq \theta_j\}$
6: Επέλεξε $a \in A$, τέτοιο ώστε area$(a) \leq$ area (a_k), $P(a) \geq P(a_k), \forall a_k \in A$
7: $\epsilon_{ij} \leftarrow$ area(a); *//Εμβαδόν κύκλου γύρω από το κέντρο r*
8: **end for**
9: **end for**
10: **End Procedure**

Ο σκοπός του δεύτερου σταδίου της προεπεξεργασίας είναι ο υπολογισμός

Σχήμα 4.6: Υπολογισμός εμβαδού κύκλου για κατώφλι πιθανότητας 70% σε κύκλο ακτίνας 1χμ και σε κύκλο με ακτίνα 500μ.

του μικρότερου δυνατού εμβαδού που έχει ως πιθανότητα ένα κατώφλι θ_i. Η διαδικασία αυτή πραγματοποιείται ώστε τα κατώφλια αυτά να χρησιμοποιηθούν στην διαδικασία του υπολογισμού για κλάδεμα στο στάδιο επεξεργασίας (pruning3). Με αυτόν τον τρόπο μειώνονται οι πράξεις που εκτελούνται. Το εμβαδό ϵ_{ij} που επιλέγεται είναι εκείνο ενός κύκλου με κέντρο το κέντρο του κύκλου της κυκλικής κανονικής κατανομής. Ο προϋπολογισμός αυτός γίνεται προσεγγιστικά δοκιμάζοντας έναν μεγάλο αριθμό από ακτίνες με τη μέθοδο Monte Carlo και θα εξηγηθεί αναλυτικότερα στη συνέχεια. Π.χ. για να υπολογίσουμε τα κατώφλια του κύκλου με ακτίνα 1 χμ εξετάζουμε με Monte Carlo όλους τους κύκλους με ακτίνες από 0 μέχρι 1χμ με βήμα 10 μ. τι πιθανότητες περιέχουν. Από αυτούς κρατάμε μόνο εκείνους που περιέχουν πιθανότητες που είναι κοντά στα κατώφλια. Στο σχήμα 4.6 φαίνεται ο εσωτερικός κύκλος που περιέχει το 70% της πιθανότητας της κυκλικής κανονικής κατανομής ακτίνας 1 χμ. Το εμβαδόν του κύκλου, αφού βρεθεί η κατάλληλη ακτίνα, υπολογίζεται από τον τύπο υπολογισμού του εμβαδού του κύκλου $\pi\rho^2$. Αρχικά παρατηρήθηκε ότι για το ίδιο κατώφλι το εμβαδόν του κύκλου με διπλάσια ακτίνα που περικλείει πιθανότητα ίση με το κατώφλι είναι τετραπλάσιο. Συνεπώς στο σχήμα 4.6 το γραμμοσκιασμένο εμβαδόν του αριστερού σχήματος είναι τετραπλάσιο από το αντίστοιχο του δεξιού. Αυτό αποδεικνύεται μαθηματικά από το παρακάτω λήμμα.

Λήμμα

Η πιθανότητα του εμβαδού που περικλείεται σε κύκλο ακτίνας x είναι:

$$\int_{\theta=0}^{\theta=2\pi}\int_{r=0}^{r=x} f(r,\theta)drd\theta = \int_{\theta=0}^{\theta=2\pi}\int_{r=0}^{r=x} \frac{1}{2\pi\sigma_x^2}e^{-\frac{1}{2}(\frac{r}{\sigma_x})^2}drd\theta = \int_{r=0}^{r=x} \frac{\pi r}{2\pi\sigma_x^2}e^{-\frac{1}{2}(\frac{r}{\sigma_x})^2}dr$$

Με την αντικατάσταση $2r^{'} = r$ έχουμε:

$$\int_{r=0}^{r=2x^{'}} \frac{2r^{'}}{\sigma_x^2}e^{-\frac{1}{2}(\frac{4r^{'}}{\sigma_x})^2}dr^{'}$$

που με τον μετασχηματισμό $2\sigma_x^{'} = \sigma_x$ έχουμε:

$$\int_{r=0}^{r=2x^{'}} \frac{4r^{'}}{4\sigma_x^{'2}} e^{-\frac{1}{2}(\frac{2r^{'}}{2\sigma_x^{'}})^2} dr^{'}$$

, που είναι ο κύκλος με την διπλάσια ακτίνα (τετραπλάσιο εμβαδόν) από τον προηγούμενο σε κυκλική κανονική κατανομή διπλάσιας ακτίνας από πριν. Η ίδια διαδικασία μπορεί να γενικευτεί για ακτίνες n φορές μεγαλύτερες που σε αντίστοιχες καταστάσεις θα έχουν n^2 φορές μεγαλύτερο εμβαδόν.

4.3.3 Χρησιμότητα μεθόδου Monte Carlo

Για τα δύο στάδια προεπεξεργασίας και τον υπολογισμό των αποτελεσμάτων χρησιμοποιούμε την προσομοίωση Monte Carlo. Στο πρώτο χρησιμοποιείται για τον υπολογισμών των πιθανοτήτων των εσωτερικών τετραγώνων και στο δεύτερο τον υπολογισμό των κατωφλίων-εμβαδών ϵ_{ji}. Η χρησιμότητα της μεθόδου αυτής έχει ήδη περιγραφεί στο κεφάλαιο της αβεβαιότητας. Ο αριθμός των δειγμάτων που λαμβάνονται και στα δύο στάδια πρέπει να είναι αρκετά μεγάλος για κάθε επανάληψη ώστε τα αποτελέσματα αν και προσεγγιστικά να υπολογίζονται με ακρίβεια. Δεδομένου αυτού του γεγονότος, τα παραπάνω αποτελέσματα υπολογίζονται με μεγάλη ακρίβεια.

4.4 Επεξεργασία Δεδομένων

Η επεξεργασία των δεδομένων βασίζεται σε τεχνικές μείωσης των συνολικών πράξεων ώστε ο αλγόριθμος να είναι όσο το δυνατόν γρηγορότερος και αποτελεσματικότερος. Αρχικά ακολουθείται η τεχνική της ομοιόμορφης κατάτμησης σε πλέγμα (Grid Partitioning, κάνναβος) προκειμένου να τοποθετηθούν τα αντικείμενα στις αντίστοιχες θέσεις των κελιών του. Στη συνέχεια ακολουθείται η διαδικασία αποτίμησης των ερωτημάτων. Η αποτίμηση αυτή γίνεται χρησιμοποιώντας τεχνικές κλαδέματος. Οι τεχνικές αυτές χρησιμοποιούνται ώστε να μην γίνεται πλήρης αποτίμηση όλων των ερωτημάτων, παρά μόνο εκείνων που πληρούν τις προϋποθέσεις, π.χ. η συνολική πιθανότητα ένα αντικείμενο να βρίσκεται σε μία περιοχή υπερβαίνει κάποιο κατώφλι 95%.

4.4.1 Ευρετήριο πλέγματος

Η τεχνική του Grid Partitioning χρησιμοποιείται ευρέως ως ευρετήριο για κινούμενα αντικείμενα σε χωρικά ρεύματα δεδομένων. Σκοπός του είναι η ταξινόμηση των αντικειμένων σε κελιά με βάση την θέση τους ανά χρονόσημο (timestamp). Κάθε κελί κρατάει τα αντικείμενα εκείνα τα οποία το τέμνουν. Αυτό σημαίνει ότι ένα αντικείμενο μπορεί να περιέχεται σε περισσότερα από ένα κελιά. Η ανανέωση του πλέγματος γίνεται περιοδικά ανά κύκλο εκτέλεσης. Το πλεονέκτημά της είναι ότι συγκεντρώνει τα γειτονικά-κοντινά αντικείμενα, οπότε δεν χρειάζεται να εξετάζουμε όλα τα αντικείμενα που κινούνται στον χάρτη όταν αποτιμάται ένα ερώτημα παρά μόνο εκείνα που περιέχονται στα κελιά που τέμνει. Το κόστος

Σχήμα 4.7: Παράδειγμα μεθόδου πλέγματος για αβέβαια κινούμενα αντικείμενα

δημιουργίας του είναι μικρό, ενώ χωρίς αυτό η αποτίμηση ερωτημάτων θα ήταν χρονοβόρα. Επιπροσθέτως, θεωρούμε ότι τα αντικείμενα κινούνται με ταχύτητες μικρότερες από 200χμ την ώρα, υπόθεση αρκετά λογική. Το γεγονός αυτό οδηγεί στο αποτέλεσμα πολλά από τα αντικείμενα να παραμένουν στα κελιά του πλέγματος και δεν χρειάζεται να επανεξετάζεται συνεχώς η θέση τους. Το πλέγμα, το οποίο το συμβολίζουμε με G, έχει συγκεκριμένες διαστάσεις (width, height) και τεμαχίζεται σε $(c \times c)$ τετραγωνικές περιοχές. Τόσο οι διαστάσεις όσο και ο αριθμός των τετραγώνων που το χωρίζουν δίνονται ως παράμετροι και ο αριθμός τους επηρεάζει την απόδοση και το κόστος δημιουργίας και επεξεργασίας του αλγορίθμου. Π.χ. ένας μικρός αριθμός ευρετηρίου κελιών για μεγάλο αριθμό αντικειμένων στον χάρτη έχει μειωμένες αποδόσεις κάνοντάς το πιο αργό, αλλά επιφέρει μικρό κόστος δημιουργίας και επεξεργασίας. Αντιθέτως, ο μεγάλος αριθμός κελιών για τον ίδιο μεγάλο αριθμό αντικειμένων, βελτιώνει την απόδοση του αλγορίθμου κάνοντας τον πιο γρήγορο, αλλά τον επιβαρύνει με επιπλέον κόστος δημιουργίας και επεξεργασίας. Στο σχήμα 4.7 φαίνεται η χρησιμότητα του πλέγματος. Το ερώτημα που τίθεται εξετάζει μόνο τα αντικείμενα που περιέχονται στα κελιά που τέμνει. Π.χ. στο σχήμα 4.7 φαίνεται ότι το ερώτημα Q εξετάζει μόνο τα αντικείμενα που τέμνονται με τα κελιά του πλέγματος που τέμνει και το ερώτημα. Όλα τα υπόλοιπα αντικείμενα δεν εξετάζονται καθόλου.

4.4.2 Τεχνικές κλαδέματος

Όπως αναφέρθηκε και στην εισαγωγή του κεφαλαίου η φάση φιλτραρίσματος αποτελείται από τρεις τεχνικές κλαδέματος. Ο σκοπός τους είναι να μειώσουν το σύνολο των αντικειμένων για τα οποία πρέπει να γίνει αναλυτική αποτίμηση, άρα και τον χρόνο εκτέλεσης του αλγορίθμου. Η πρώτη τεχνική αναφέρεται στα στάσιμα αντικείμενα και ερωτήματα, η δεύτερη στα αντικείμενα που βρίσκονται εξ

ολοκλήρου μέσα στα ερωτήματα και η τρίτη στο εμβαδό της τομής του ερωτήματος με τα αντικείμενα.

Πρώτη τεχνική κλαδέματος

Η πρώτη συνθήκη κλαδέματος αφορά τα ερωτήματα που δεν έχουν κινηθεί ή τουλάχιστον έχουν κινηθεί , π.χ. 50μ. σε δύο συνεχόμενα χρονόσημα. Γι΄ αυτά τα ερωτήματα δεν χρειάζεται να επανεξετάσουμε αντικείμενα που δεν έχουν κινηθεί ή έχουν κινηθεί λίγο αντίστοιχα με τα ερωτήματα και δεν ανανεώνουμε τις απαντήσεις που είχαμε δώσει στο προηγούμενο χρονόσημα. Για όλα αυτά τα αντικείμενα δεν επιστρέφουμε απάντηση στον χρήστη της κινητής συσκευής και η αποτίμηση μένει ως έχει χωρίς να αλλάξει. Τα ερωτήματα που κινήθηκαν αρκετά εξετάζονται περαιτέρω. Επιπλέον, για τα ακίνητα ερωτήματα, τα υποψήφιά τους αντικείμενα που κινήθηκαν αρκετά εξετάζονται και αυτά περαιτέρω.

Δεύτερη τεχνική κλαδέματος

Η δεύτερη συνθήκη κλαδέματος αφορά τα αντικείμενα o_i που το $inbox(o_i)$ τους βρίσκεται εξ ολοκλήρου μέσα σε ένα ερώτημα. Επομένως έχουν πιθανότητα 100% κατά την αποτίμησή τους.. Η διαδικασία αυτή επιτυγχάνεται χωρίς μεγάλο κόστος υπολογισμών, αρκετά απλά με τέσσερις συγκρίσεις των συντεταγμένων των τετραγώνων με τα ορθογώνια. Για τα αντικείμενα αυτά επιστρέφεται απάντηση 100% στον χωρίς να χρειάζεται να γίνουν περαιτέρω αναλυτικοί υπολογισμοί. Τα αντικείμενα που δεν βρίσκονται 100% μέσα στα ερωτήματα εξετάζονται περαιτέρω.

Τρίτη τεχνική κλαδέματος

Η τρίτη συνθήκη κλαδέματος αφορά το εμβαδόν της τομής των ερωτημάτων με τα αντικείμενα και χρησιμοποιεί τα αποτελέσματα του δεύτερου σταδίου προϋπολογισμού. Βασίζεται πάνω τόσο στα γεωμετρικά όσο και στα πιθανοτικά χαρακτηριστικά της τομής των ερωτημάτων με των αντικειμένων. Εφόσον το εμβαδόν της τομής του ερωτήματος με κάποιο αντικείμενο, $area(\bigcap(q_i, inbox(o_j)))$, είναι μικρότερο από το εμβαδόν ϵ_{ji} που υπολογίστηκε στην προεπεξεργασία και αντιστοιχεί στο μικρότερο εμβαδό που αντιστοιχεί στο κατώφλι θ_i του ερωτήματος τότε το αντικείμενο δεν εξετάζεται αναλυτικά. Πιο συγκεκριμένα, θα πρέπει να ισχύει:

$$area(\bigcap(q_i, inbox(o_j))) < \epsilon_{ji}$$

για να απορριφθεί το αντικείμενο. Από τη συγκεκριμένη τεχνική κλαδέματος αναμένεται να απορρίπτονται αμέσως αντικείμενα που τομή τους με τα ερωτήματα είναι μικρή. Τα αντικείμενα που έχουν αρκετά μεγάλη τομή με ένα ερώτημα εξετάζονται περαιτέρω και οδηγούνται σε πιο λεπτομερής αποτίμηση κατά τη φάση εκλέπτυνσης (Refinement phase).

Σχήμα 4.8: Παράδειγμα τελικής αποτίμησης ερωτήματος για $\lambda = 5$

4.4.3 Λεπτομερής Αποτίμηση

Η τελική αποτίμηση αναφέρεται σε αντικείμενα που έχουν περάσει όλους τους περιορισμούς κλαδέματος. Η αποτίμηση αυτή γίνεται ως εξής: βρίσκουμε τις συντεταγμένες της τομής του ορθογωνίου ερωτήματος με το $inbox$ του αντικειμένου που αποτελεί ένα ορθογώνιο, έστω $((x_{min}, y_{min}), (x_{max}, y_{max}))$. Από τις συντεταγμένες αυτές υπολογίζουμε το άθροισμα θ_j^{min} των πιθανοτήτων των εσωτερικών τετραγώνων b_k που περιέχονται μόνο στο εσωτερικό του και το άθροισμα θ_j^{max} των πιθανοτήτων των εσωτερικών τετραγώνων b_k που τέμνονται από το κοινό ορθογώνιο. Αναλυτικότερα :

$$\theta_j^{min} = \sum(p(b_k)), \forall b_k \text{ εξ ολοκλήρου εντός του } q_i$$

και

$$\theta_j^{max} = \sum(p(b_k)), \forall b_k \text{ που τέμνει το } q_i$$

Από τα δύο παραπάνω μεγέθη σχηματίζουμε το διάστημα εμπιστοσύνης $(\theta_j^{min}, \theta_j^{max})$. Το διάστημα εμπιστοσύνης αποτελεί ένα σύνολο πιθανοτήτων με μία μικρότερη και μία μεγαλύτερη τιμή ανάμεσα στις οποίες θα βρίσκεται και η ακριβής τιμή της πιθανότητας τομής της κυκλικής κανονικής κατανομής του αντικειμένου με το ερώτημα. Όσο μικρότερο είναι το διάστημα εμπιστοσύνης τόσο πιο ακριβή αποτελέσματα θα δίνονταν τελικά. Η διαδικασία δημιουργίας του διαστήματος εμπιστοσύνης είναι απλή και αποδοτική αφού είναι απλώς το άθροισμα κάποιων εσωτερικών τετραγώνων, των οποίων οι τιμές έχουν προϋπολογιστεί στο πρώτο στάδιο της προεπεξεργασίας και δεν εμπλέκει τον υπολογισμό πιθανοτήτων με μεγάλο υπολογιστικό κόστος. Δύο παραδείγματα για διαφορετικές τιμές του λ δίνονται στα σχήματα 4.8 και 4.9. Το διάστημα εμπιστοσύνης που υπολογίστηκε μπορεί να δώσει ως απάντηση το αντικείμενο εφόσον το θ_j^{min} είναι μεγαλύτερο του κατωφλίου θ_i που

Σχήμα 4.9: Παράδειγμα τελικής αποτίμησης ερωτήματος για $\lambda = 7$

έχει τεθεί από το ερώτημα. Σε αντίθετη περίπτωση, το αντικείμενο δεν δίνεται ως απάντηση. Αναλυτικότερα θα επιστρέφεται θετική απάντηση, με το αντίστοιχο διάστημα εμπιστοσύνης, για όσα αντικείμενα ισχύει:

$$\theta_i < \theta_j^{min} < \theta_j^{max}$$

ή

$$\theta_j^{min} < \theta_i < \theta_j^{max}$$

και αρνητική για όσα ισχύει:

$$\theta_j^{min} < \theta_j^{max} < \theta_i$$

Η τελική απάντηση ανά κύκλο εκτέλεσης δίνεται στη μορφή $< o_i, \theta_j^{min}, \theta_j^{max} >$.

Στη συνέχεια ακολουθεί ο ψευδοκώδικας του κομματιού της επεξεργασίας.

Algorithm 3 Επεξεργασία

Procedure *ServerAlgorithm*
Είσοδος: Υποθέτουμε σύνολο M ερωτημάτων $q_i = (x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max}, t, \theta_i)$, N αντικειμένων $o_j = (c_j, R_j, t)$ και κανονική κατανομή .
Έξοδος: $\forall q_i : \{< o_j, \theta_j^{min}, \theta_j^{max} >: \bigcap(o_j, q_i)$ με διάστημα εμπιστοσύνης $((\theta_i < \theta_j^{min} < \theta_j^{max}) \vee (\theta_j^{min} < \theta_i < \theta_j^{max}))\}$
$G \leftarrow$ grid partitioning στα $c \times c$ τετραγωνικά κελιά. *//Χωρικό ευρετήριο*
Σε κάθε χρονόσημο t:
for κάθε $g \in G$ **do**
  Αντιστοίχησε στο g κάθε q_i και o_j που επικαλύπτει το g
  for κάθε q_i **do**
    $C_i = \{o_j$:που επικαλύπτεται με τα κελιά του $q_i\}$. *//Υποψήφιο σύνολο*
    for κάθε $o_j \in C_i$ **do**
      if (o_jείναι ακίνητο $\wedge$ q_i είναι ακίνητο) **then**
        συνέχισε ; *//Αποφεύγουμε την αποτίμηση ακίνητων οντοτήτων*
      else if $q_i \supset inbox(o_j)$ **then**
        Δώσε ως απάντηση $< o_j, 100, 100 >$
      else if $\bigcap(q_i,$ inbox$(o_j))= \emptyset$ **then**
        συνέχισε ; *//Pruning*
      else if εμβαδόν($\bigcap(q_i,$ inbox$(o_j))$) $< \epsilon_{ji}$ **then**
        συνέχισε ; *//Pruning*
      else
        Βρες την θέση όλων των τετραγώνων b_k των αντικειμένων o_j
        Κάτω όριο : $\theta_j^{min} = \sum(p(b_k)), \forall b_k$ Εξ ολοκλήρου εντός του q_i
        Πάνω όριο $\theta_j^{max} = \sum(p(b_k)), \forall b_k$ που τέμνει το q_i
        if $((\theta_i < \theta_j^{min} < \theta_j^{max}) \vee (\theta_j^{min} < \theta_i < \theta_j^{max}))$ **then**
          Δώσε ως απάντηση $< o_j, \theta_j^{min}, \theta_j^{max} >$ *//Αντικείμενα που επιστρέφονται ως απάντηση*
        end if
      end if
    end for
  end for
end for
End Procedure

Κεφάλαιο 5

Πειραματική Αξιολόγηση

Στο κεφάλαιο αυτό παρουσιάζονται ο τρόπος υλοποίησης του αλγορίθμου καθώς επίσης και η αξιολόγηση και ο έλεγχος της σωστής λειτουργίας του. Αρχικά περιγράφεται η αρχιτεκτονική του συστήματος, αναλύοντας τα μέρη από τα οποία αποτελείται η εφαρμογή. Στη συνέχεια, περιγράφονται οι επιλογές που είχαμε όσον αφορά τα αρχεία εισόδου για τη δοκιμή του αλγορίθμου και οι λόγοι που καθόρισαν την επιλογή μας. Παρουσιάζονται τα χαρακτηριστικά των αρχείων εισόδου και πληροφορίες για τα δεδομένα που χρησιμοποιήθηκαν στα πειράματα. Τέλος, παρατίθενται τα συγκριτικά πειράματα που εκτελέσαμε και γίνεται αξιολόγηση των επιδόσεων.

5.1 Αρχιτεκτονική συστήματος

Η αρχιτεκτονική του συστήματος βασίζεται σε τρία κομμάτια, εκείνο της εισόδου, της προεπεξεργασίας και επεξεργασίας των εισόδων και της εξόδου. Όλα τα κομμάτια του συστήματος απεικονίζονται από το σχήμα 5.1.

Το στάδιο της προεπεξεργασίας διακρίνεται σε δύο μέρη, εκείνο της δημιουργίας των εσωτερικών τετραγώνων b_k και εκείνο της δημιουργίας των κατωφλίων εμβαδού ϵ. Ο υπολογισμός των εσωτερικών κελιών b_k δεν δέχεται κάποια είσοδο, αφού είναι ανεξάρτητη της ακτίνας των κύκλων και γίνεται αναλυτικά με τη μέθοδο Monte Carlo όπως εξηγήθηκε στο κεφάλαιο 4. Ο υπολογισμός των κατωφλίων εμβαδού ϵ δέχεται ως είσοδο την ακτίνα R των κύκλων και το πιθανοτικό κατώφλι θ_i των ερωτημάτων και παράγει αναλυτικά με την μέθοδο Monte Carlo το μικρότερο εμβαδόν των κύκλων που περιέχει την πιθανότητα του κατωφλίου θ_i .

Το στάδιο της επεξεργασίας λαμβάνει ως εισόδους τα ρεύματα των κινούμενων αντικειμένων και ερωτημάτων και πράττει ως εξής: Αρχικά τοποθετεί τα κινούμενα αντικείμενα στα κελιά του πλέγματος. Στη συνέχεια για κάθε ερώτημα απομονώνονται τα κελιά τα οποία τέμνει. Από αυτά τα κελιά εξετάζονται όλα τα αντικείμενα τα τα τέμνουν. Για κάθε τέτοιο συνδυασμό ερωτήματος με αντικείμενο ακολουθεί το στάδιο φιλτραρίσματος (filter). Το στάδιο αυτό περιέχει τρεις τεχνικές κλαδέματος. Πιο αναλυτικά, η πρώτη τεχνική αναφέρεται στα στάσιμα αντικείμενα και ερωτήματα, η δεύτερη στα αντικείμενα που βρίσκονται εξ ολοκ-

Σχήμα 5.1: Γενική αρχιτεκτονική συστήματος.

λήρου μέσα στα ερωτήματα και η τρίτη στο εμβαδό της τομής του ερωτήματος με τα αντικείμενα. Για την υλοποίηση της τρίτης τεχνικής κλαδέματος το στάδιο φιλτραρίσματος δέχεται ως είσοδο αποτελέσματα που παρήχθησαν κατά την προεπεξεργασία και πιο συγκεκριμένα κατά τον υπολογισμό των κατωφλίων εμβαδών. Το στάδιο φιλτραρίσματος είτε δίνει κάποιες απαντήσεις είτε απορρίπτει κάποια αντικείμενα. Τα αντικείμενα που πληρούν τις προϋποθέσεις να περάσουν το στάδιο του φιλτραρίσματος οδηγούνται στο επόμενο, εκείνο της εκλέπτυνσης (refinement), όπου και γίνεται η λεπτομερής αποτίμηση. Για να το κάνει όμως αυτό δέχεται ως είσοδο τα υπολογισμένα τετράγωνα b_k, τα οποία υπολογίστηκαν κατά το στάδιο της προεπεξεργασίας. Η τελική αυτή αποτίμηση δίνει ως αποτελέσματα ένα διάστημα εμπιστοσύνης. Εφόσον το κατώφλι θ_iτου ερωτήματος είναι μικρότερο από αυτό το διάστημα εμπιστοσύνης τότε το αντικείμενο απορρίπτεται και δεν δίνεται ως απάντηση. Σε αντίθετη περίπτωση δίνεται ως έξοδος θετική απάντηση.

5.2 Πειραματικό πλαίσιο

Στην ενότητα αυτή αξιολογείται πειραματικά ο αλγόριθμος που δημιουργήθηκε στο κεφάλαιο 4. Όλε οι δομές υλοποιήθηκαν σε γλώσσα C++ και τα πειράματα εκτελέστηκαν σε λειτουργικό σύστημα Windows 7 σε προσωπικό υπολογιστή Intel(R) Core(TM) 2 Duo P8400, 2.53 GHz με μνήμη 3 GHz.

5.2.1 Παραγωγή συνθετικών δεδομένων

Τα πειραματικά δεδομένα των κινούμενων αντικειμένων παρήχθησαν βάσει ενός ψηφιακού χάρτη του οδικού δικτύου του πολεοδομικού συγκροτήματος Αθηνών. Το συγκεκριμένο ψηφιακό υπόβαθρο αφορά το βασικό οδικό δίκτυο της πρωτεύουσας,

Σχήμα 5.2: Τροχιές στην περιοχή της Αθήνας

προέρχεται από χάρτες κλίμακας 1:5000 και τηρείται σε διανυσματική (vector) μορφή, ενώ καλύπτει έκταση περίπου 300 τ.χλμ. Οι οδικοί άξονες διακρίνονται σε κατηγορίες (λεωφόροι ταχείας κυκλοφορίας, κύριες και δευτερεύουσες αρτηρίες, βοηθητικοί δρόμοι), και χαρακτηρίζονται από την μέση ταχύτητα κίνησης των οχημάτων στην διάρκεια της ημέρας, όπως έχει προκύψει από επιτόπιες μετρήσεις. Αυτό ακριβώς το στοιχείο μπορεί να αξιοποιηθεί για τον υπολογισμό του μέσου χρόνου διαδρομής ενός οχήματος κατά μήκος των συνδέσμων του δικτύου, καθιστώντας αυτήν την γεωγραφική βάση δεδομένων κατάλληλη για υπολογισμό της βέλτιστης διαδρομής (shortest path) μέσα στην πόλη. Με χρήση του λογισμικού ArcView GIS 3.2 και της επέκτασής του Network Analyst 1.0b , δημιουργήθηκαν συνολικά 100000 τροχιές ισάριθμων αντικειμένων, θέτοντας ως προέλευση και προορισμό κάθε διαδρομής τυχαία επιλεγμένα ζεύγη κόμβων του δικτύου. Οι κινήσεις διεξάγονται ως επί το πλείστον ακτινικά, θεωρώντας ότι τα περισσότερα αντικείμενα ξεκινούν από την περιφέρεια, διέρχονται από το κέντρο της πόλης και κατευθύνονται προς κάποιο προάστιο. Κατόπιν, από κάθε τροχιά έγινε δειγματοληψία σημειακών θέσεων, λαμβάνοντας συνολικά 200 στίγματα ανά τροχιά με ίση χρονική απόσταση μεταξύ τους ώστε να αντιπροσωπεύουν τακτικές ενημερώσεις της θέσης των αντικειμένων. Τελικά, προέκυψε ένα αρχείο τροχιών με εγγραφές που φέρουν την ταυτότητα (ID) του αντικειμένου, τις συντεταγμένες (x, y) και το χρονόσημο (t).

Επίσης, παρήχθησαν συνολικά 10000 θέσεις ερωτημάτων για την ίδια περιοχή. Για κάθε ερώτημα υπάρχει αρχική καταγραφή για t=0. Για κάθε επόμενο χρονόσημο, περί το 1% των τρεχουσών θέσεων επιλέγεται τυχαία και οι αντίστοιχες θέσεις μετατοπίζονται επίσης τυχαία. Λ.χ. υπάρχουν ερωτήματα που μετακινούνται έως και 9 φορές, ενώ άλλα καθόλου σε όλο το χρονικό διάστημα [0..200]. Επομένως, η κινητικότητα ($agility$) των ερωτημάτων έχει τεθεί στο 1%.

5.2.2 Πειραματικά δεδομένα

Για τον αλγόριθμο αποτίμησης πιθανοτικών ερωτημάτων σε αβέβαια κινούμενα αντικείμενα χρησιμοποιήθηκαν τέσσερα πειραματικά σύνολα δεδομένων αποτελούμενα από 10000, 20000, 50000 και 100000 κινούμενα αντικείμενα καθώς και 1000, 2000, 5000, 10000 κινούμενα ερωτήματα. Κάθε σύνολο περιέχει σημειακές θέσεις των αντικειμένων και ερωτημάτων του για 200 χρονόσημα. Κάθε εγγραφή του συνόλου των αντικειμένων έχει την μορφή $< t, id, x, y >$, όπου t είναι το χρονόσημο, id η ταυτότητα του αντικειμένου και x, y οι συντεταγμένες του κέντρου του κύκλου της κυκλικής κανονικής κατανομής του. Η ακτίνα των κύκλων δίνεται για όλα τα αντικείμενα ίδια, απλοποίηση που γίνεται για την ευκολία διαχείρισης των πειραματικών δεδομένων. Κάθε εγγραφή του συνόλου των ερωτημάτων έχει την μορφή $< t, id, x, y >$, όπου t είναι το χρονόσημο, id η ταυτότητα του αντικειμένου και x, y οι συντεταγμένες του κάτω αριστερά άκρου του ορθογωνίου του. Οι διαστάσεις των ορθογωνίων καθορίζονται τυχαία με βάση μία τυχαία διαδικασία καθορισμού που θα εξηγηθεί στη συνέχεια. Τα πειραματικά σύνολα των 10000, 20000, 50000 για τα αντικείμενα και των 1000, 2000, 5000 για τα ερωτήματα επιλέχθηκαν δειγματοληπτικά από τα αντίστοιχα σύνολα των 100000 αντικειμένων και 10000 ερωτημάτων. Έγιναν πειράματα για τις εξής παραμέτρους του αλγορίθμου:

- Πλήθος κινούμενων αντικειμένων N.
- Πλήθος κινούμενων ερωτημάτων M.
- Πλήθος κελιών $c \times c$ πλέγματος.
- Έκταση ερωτημάτων ως % της συνολικής περιοχής %.
- Ακτίνα κύκλων R.
- Πλήθος κουτιών $\lambda \times \lambda$ εγγεγραμμένου τετραγώνου.

Πιο συγκεκριμένα οι τιμές που δώθηκαν για κάθε παράμετρο ήταν:

Πλήθος αντικειμένων N	10k, 20k, 50k, **100k**
Πλήθος ερωτημάτων M	1k, 2k, 5k, **10k**
Πλήθος κελιών πλέγματος $c \times c$	5×5, **10×10**, 15×15, 20×20
Έκταση ερωτημάτων a	**1%**, 2%, 5%
Ακτίνα κύκλων R	200m, 600m, **1000m**, 2000m, 3000m
Πλήθος κουτιών $\lambda \times \lambda$	5×5, 7×7, **10×10**, 15×15, 20×20

Πίνακας 5.1: Παράμετροι πειραμάτων

Σχήμα 5.3: Χρόνος εκτέλεσης για διάφορες ακτίνες και μεγέθη καννάβου

5.3 Αξιολόγηση αποτελεσμάτων

Κατά τη διάρκεια των πειραμάτων μετρήθηκαν:

- Ο χρόνος τοποθέτησης των αντικειμένων στα κελιά του καννάβου.
- Ο χρόνος εκτέλεσης του αλγορίθμου ανά χρονόσημο.
- Το ποσοστό των αντικειμένων που πέφτουνε εξ ολοκλήρου μέσα στα ερωτήματα.
- Το ποσοστό των αντικειμένων που ανήκουν στα κελιά που τέμνουν τα ερωτήματα αλλά δεν έχουν κοινή τομή μαζί τους.
- Το ποσοστό των αντικειμένων που η πιθανότητα της τομής τους με τα ερωτήματα δεν ξεπερνάει το κατώφλι των ερωτημάτων και
- τέλος το ποσοστό των αντικειμένων για τα οποία γίνεται αναλυτική αποτίμηση.

Όλα τα παραπάνω ποσοστά αναφέρονται σε κάθε χρονική στιγμή. Επιπροσθέτως εξετάζεται δειγματοληπτικά η ακρίβεια του αλγορίθμου έναντι ενός εξαντλητικού αλγορίθμου που δίνει ακριβείς και όχι προσεγγιστικές απαντήσεις. Τέλος γίνεται η σύγκριση των δύο αλγορίθμων ως προς το κόστος εκτέλεσης τους ανά χρονόσημο.

5.3.1 Διαστασιολόγηση καννάβου

Η επιλογή του αριθμού των κελιών του καννάβου έγινε βάσει του παρακάτω πειράματος: Για κάθε μία από τις ακτίνες 200μ, 600μ, 1000μ, 2000μ και 3000μ μετρήθηκε ο χρόνος εκτέλεσης για πλήθος κελιών $c = 5, 10, 15, 20$. Από το σχήμα 5.3 βλέπουμε ότι οι γραφικές παραστάσεις του χρόνου ενημέρωσης του καννάβου

Σχήμα 5.4: Κλιμάκωση χρόνου εκτέλεσης

είναι αύξουσες για τις τέσσερις τιμές. Αντίθετα οι χρόνοι εκτέλεσης σχηματίζουν μία καμπύλη. Αρχικά μειώνονται μέχρι ενός σημείου (ελάχιστο). Από εκεί και πέρα αυξάνονται. Κάτι τέτοιο βλέπουμε ότι γίνεται πιο έντονα για μεγαλύτερες ακτίνες. Για μικρό πλήθος κελιών, ο κάνναβος είναι σχεδόν άχρηστος αφού αποτυγχάνει να διακρίνει επαρκώς την συνάφεια των αντικειμένων με ερωτήματα. Απ΄ την άλλη υπερβολικός κατακερματισμός σε κελιά οδηγεί σε αυξημένο διαχειριστικό κόστος, αφού η κυκλική κανονική κατανομή κάθε αντικειμένου θα αναφέρεται σε πολλά κελιά. Αυτό συμβαίνει σε μεγαλύτερο βαθμό για μεγαλύτερες ακτίνες αφού αντικείμενα με μεγαλύτερες ακτίνες αποθηκεύονται σε περισσότερα κελιά από αντικείμενα με μικρότερες ακτίνες. Έτσι μια μέση κατάτμηση c =10 αποδεικνύεται προτιμότερη για την εξεταζόμενη περίπτωση. Πράγματι, από τις μετρήσεις που έγιναν καταλήξαμε στο συμπέρασμα ότι η καλύτερη επιλογή κελιών καννάβου για το υπολογιστικό σύστημα στο οποίο έγιναν τα πειράματα είναι c =10. Η επιλογή αυτή δεν είναι η καλύτερη δυνατή για όλες τις ακτίνες π.χ. βλέπουμε ότι για ακτίνα 200μ. c =15 είναι πιο γρήγορη επιλογή. Παρόλα αυτά η επιλογή c =10 είναι πολύ καλή για το σύνολο των περιπτώσεων. Όλες οι υπόλοιπες μετρήσεις που παρουσιάζονται στη συνέχεια εκτελέστηκαν για κελιά 10×10.

5.3.2 Κλιμακωσιμότητα

Μετράμε την κλιμάκωση του χρόνου εκτέλεσης για δύο διαφορετικές ακτίνες αντικειμένων (1000μ και 2000μ αντίστοιχα), αρχικά για αυξανόμενο αριθμό αντικειμένων και σταθερό αριθμό ερωτημάτων και στη συνέχεια για αυξανόμενο αριθμό ερωτημάτων και σταθερό αριθμό αντικειμένων.

Αυξανόμενο πλήθος αντικειμένων για σταθερό αριθμό ερωτημάτων

Για το αρχείο με τα 10000 ερωτήματα έγιναν πειράματα με αυξανόμενο αριθμό αντικειμένων. Αρχικά για 10000 και στη συνέχεια για 20000, 50000 και 100000 αντικείμενα. Όλες οι άλλες παράμετροι παρέμειναν σταθερές. Συγκεκριμέ-

να το κατώφλι των ερωτημάτων ήταν $\theta_i = 0.95$ και οι διαστάσεις των ερωτημάτων αποτελούσαν το a =0.01% της συνολικής περιοχής. Από το σχήμα 5.4 (αριστερά) παρατηρούμε σε δύο διαφορετικές περιπτώσεις για ακτίνες κύκλων $R = 1000m$ και $R = 2000m$ ότι με την αύξηση του πλήθους των αντικειμένων αυξάνεται ο χρόνος εκτέλεσης. Φυσικά, ο χρόνος τοποθέτησης των αντικειμένων στον κάνναβο. Πιο συγκεκριμένα, βλέπουμε ότι ο χρόνος εκτέλεσης από τα 10000 στα 20000 αντικείμενα υπερδιπλασιάζεται. Το ίδιο γινεται και από τα 20000 στα 50000 αντικείμενα. Από τα 50000 στα 100000 ο χρόνος σχεδόν τριπλασιάζεται. Αυτό ερμηνεύεται ως εξής: Όταν αυξάνεται το πλήθος των αντικειμένων, υπάρχουν όλο και περισσότερες επικαλύψεις από τις κυκλικές κατανομές τους με τα ερωτήματα. Άρα κάθε ερώτημα οφείλει να εξετάζει πολλαπλάσιο αριθμό αντικειμένων. Το φαινόμενο επιτείνεται όταν τα αντικείμενα έχουν μεγαλύτερη ακτίνα.

Αυξανόμενο πλήθος ερωτημάτων για σταθερό αριθμό αντικειμένων

Για το αρχείο με τα 100000 αντικείμενα έγιναν πειράματα με αυξανόμενο αριθμό ερωτημάτων. Αρχικά για 1000 και στη συνέχεια για 2000, 5000 και 10000 ερωτήματα. Όλες οι άλλες παράμετροι παρέμειναν σταθερές. Συγκεκριμένα το κατώφλι των ερωτημάτων ήταν $\theta_i = 0.95$ και είχαμε ίδιες διαστάσεις ερωτημάτων ως προς % της συνολικής περιοχής a =0.01%. Από το σχήμα 5.4 (δεξιά)παρατηρούμε σε δύο διαφορετικές περιπτώσεις για ακτίνες κύκλων $R = 1000m$ και $R = 2000m$ ότι με την αύξηση του πλήθους των αντικειμένων αυξάνεται ο χρόνος εκτέλεσης. Φυσικά, ο χρόνος τοποθέτησης των αντικειμένων στον κάνναβο. Πιο συγκεκριμένα, βλέπουμε ότι ο χρόνος εκτέλεσης από τα 1000 στα 2000 αντικείμενα υπερδιπλασιάζεται. Το ίδιο γίνεται και από τα 2000 στα 5000 ερωτήματα και από τα 5000 στα 10000 ερωτήματα. Αυτό ερμηνεύεται ως εξής: ο συνολικός αριθμός των αντικειμένων αποτιμάται περισσοτερες φορές για μεγαλύτερο πλήθος ερωτημάτων.

5.3.3 Επίδραση της αβεβαιότητας

Βλέπουμε ότι ο χρόνος τοποθέτησης των αντικειμένων στον κάνναβο ανά χρονόσημο, για σταθερές άλλες παραμέτρους, αυξάνεται καθώς αυξάνεται η ακτίνα R των αντικειμένων. Πιο συγκεκριμένα στην εικόνα 5.5 διακρίνεται ο χρόνος εκτέλεση για μήκη ακτίνας 200μ, 600μ, 1000μ, 2000μ, 3000μ με σταθερό κατώφλι $\theta_i = 0.95$ και ίδιες διαστάσεις ερωτημάτων ως προς % της συνολικής περιοχής (1%). Η αύξηση στον χρόνο δεικτοδότησης στον κάνναβο οφείλεται στο γεγονός ότι αντικείμενα με μεγαλύτερο μήκος ακτίνων είναι πιο πιθανόν να τοποθετηθούν σε περισσότερα από ένα κελιά, σε αντίθεση με τα αντικείμενα μικρότερων ακτίνων τα οποία είναι πιο πιθανό να τοποθετηθούν σε λιγότερα κελιά.

Ακόμα βλέπουμε ότι ο χρόνος εκτέλεσης ανά χρονική στιγμή, για σταθερές άλλες παραμέτρους, αυξάνεται καθώς αυξάνεται το μήκος R της ακτίνας των αντικειμένων. Η αύξηση στον χρόνο αποτίμησης οφείλεται πάλι στο γεγονός ότι αντικείμενα με μεγαλύτερο μήκος ακτίνων είναι πιο πιθανόν να τοποθετηθούν σε πολλαπλά κελιά, σε αντίθεση με τα αντικείμενα μικρότερων ακτίνων τα οποία είναι πιο πιθανό να τοποθετηθούν σε λιγότερα κελιά. Έτσι κατά τη διάρκεια της επε-

Σχήμα 5.5: Επίδραση ακτίνας R στους μετρούμενους χρόνους.

ξεργασίας κάποια κελιά θα συγκεντρώνουν πολύ μεγάλο αριθμό αντικειμένων με αποτέλεσμα την αύξηση του χρόνου εκτέλεσης.

5.3.4 Επιδόσεις ανάλογα με τα χαρακτηριστικά των ερωτημάτων

Το κατώφλι των ερωτημάτων φαίνεται από την εικόνα 5.6 πως δεν επηρεάζει τον χρόνο τοποθέτησης αντικειμένων στον κάνναβο ανά χρονόσημο. Αυτό είναι αναμενόμενο αφού τα ερωτήματα δεν αποθηκεύονται στο πλέγμα.

Αντίθετα το κατώφλι των ερωτημάτων φαίνεται από την εικόνα 5.6 πως επηρεάζει σημαντικά τον χρόνο εκτέλεσης ανά χρονική στιγμή κάτι αναμενόμενο αφού όσο μεγαλύτερο κατώφλι έχουμε τόσο περισσότερο κλάδεμα αντικειμένων γίνεται, επομένως μειώνεται ο χρόνος εκτέλεσης. Αντιθέτως, όσο μικρότερο κατώφλι έχουμε τόσο περισσότερα αντικείμενα θα αποτιμώνται λεπτομερώς, αυξάνοντας τον χρόνο εκτέλεσης. Δοκιμάστηκαν τιμές κατωφλίων 0.6, 0.7, 0.8, 0.95 για ακτίνα 1000μ. με διαστάσεις ερωτημάτων ως προς % της συνολικής περιοχής 1%.

Οι διαστάσεις ερωτημάτων ως % της συνολικής περιοχής aγνωρίζουμε ότι δεν επηρεάζουν τον χρόνο τοποθέτησης αντικειμένων στο πλέγμα ανά χρονική στιγμή κάτι αναμενόμενο αφού τα ερωτήματα δεν αποθηκεύονται στο πλέγμα. Δοκιμάστηκαν τιμές 1%, 2%, 5%.

Αντίθετα οι διαστάσεις των ερωτημάτων ως % της συνολικής περιοχής % a φαίνεται από την εικόνα 5.6 πως επηρεάζουν και αυξάνουν τον χρόνο εκτέλεσης ανά χρονική στιγμή κάτι αναμενόμενο αφού μεγαλύτερα ερωτήματα εξετάζουν μεγαλύτερο αριθμό αντικειμένων. Δοκιμάστηκαν τιμές 1%, 2%, 5% για ακτίνα 1000μ. και τέσσερις διαφορετικές τιμές κατωφλίων ερωτημάτων 0.6, 0.7, 0.8, 0.95. Βλέπουμε ότι η διαφορά στον χρόνο ανάμεσα στα ερωτήματα με συντελεστής διαστάσεων ερωτημάτων 1% και 2% είναι τεράστια αφού τα ερωτήματα τετραπλασιάζουν τον εμβαδόν τους. Το ίδιο γίνεται αν συγκρίνουμε τα ερωτήματα με συντελεστή διαστάσεων ερωτημάτων 1% και 5% αφού πια θα έχουν 25-

Σχήμα 5.6: Χρόνος εκτέλεσης για κατώφλια ϑ=0.6, 0.7, 0.8, 0.95 και διάφορες τιμές διαστάσεων ερωτημάτων α=1%, 2%, 5%, για ακτίνα αντικειμένων 1000μ.

πλασιάσει το εμβαδόν τους. Μπορούμε να παρατηρήσουμε από το σχήμα ότι η σύγκριση των χρόνων θα είναι αυτής της τάξης μεγέθους παρόλο που η κατανομή των αντικειμένων στα κελιά του πλέγματος να μην είναι ομοιόμορφη κάθε χρονική στιγμή, με αποτέλεσμα μερικά κελιά να περιέχουν πολύ μεγαλύτερο αριθμό αντικειμένων από άλλα.

5.3.5 Επίδραση τεχνικών κλαδέματος

Εδώ εξετάζουμε πως επηρεάζονται συγκριμένα ποσοστά ανά χρονική στιγμή από την αύξηση της ακτίνας των κύκλων της κανονικής κατανομής των αντικειμένων. Η ακτίνα τον κύκλων της κυκλικής κατανομής εκφράζει την αβεβαιότητα των αντικειμένων. Όσο μεγαλύτερη είναι τόσο περισσότερο αβέβαια είναι η θέση του. Συγκεκριμένα, εξετάζονται:

- το ποσοστό των αντικειμένων που πέφτουνε εξ ολοκλήρου μέσα στα ερωτήματα,
- το ποσοστό των αντικειμένων που ανήκουν στα κελιά που τέμνουν τα ερωτήματα αλλά δεν έχουν κοινή τομή μαζί τους,
- το ποσοστό των αντικειμένων που η πιθανότητα της τομής τους με τα ερωτήματα δεν ξεπερνάει το κατώφλι των ερωτημάτων και τέλος
- το ποσοστό των αντικειμένων για τα οποία γίνεται αναλυτική αποτίμηση.

Η αύξηση της ακτίνας των κύκλων φαίνεται από την εικόνα 5.7 πως επηρεάζει και μειώνει το ποσοστό των αντικειμένων που πέφτουν εξ ολοκλήρου μέσα στα

Σχήμα 5.7: Χρόνος εκτέλεσης για διάφορες ακτίνες.

ερωτήματα. Αυτό είναι λογικό, αφού αυξάνοντας την ακτίνα, αυξάνουμε και το εμβαδό που καλύπτουν τα αντικείμενα με αποτέλεσμα όλο και λιγότερα να πέφτουν μέσα στα ερωτήματα, το εμβαδόν των οποίων παραμένει σταθερό.

Το ποσοστό των αντικειμένων που ανήκουν στα κελιά που τέμνουν τα ερωτήματα αλλά δεν έχουν κοινή τομή μαζί τους μειώνεται με την αύξηση της ακτίνας των κύκλων των αντικειμένων, όπως φαίνεται από την εικόνα 5.7. Πρακτικά, μετρήθηκε ότι γενικά το σύνολο των αντικειμένων που ανήκουν σε αυτήν την κατηγορία έμεινε σταθερό. Παρόλα αυτά λόγω των αυξομειώσεων των άλλων ποσοστών τελικά το ποσοστό του μειώθηκε. Όσο πιο μεγάλη ακτίνα έχουμε, για σταθερά χαρακτηριστικά πλέγματος, τόσο περισσότερα αντικείμενα θα χρειαστεί να εξετάσουμε που έχουν κοινή τομή με το ερωτήματα. Έτσι μειώνεται το ποσοστό αυτό. Επιπροσθέτως όλο και λιγότερα αντικείμενα θα βρίσκονται εξ ολοκλήρου μέσα στα ερωτήματα.

Η αύξηση της ακτίνας των κύκλων των αντικειμένων φαίνεται από την εικόνα 5.7 πως επηρεάζει και αυξάνει το ποσοστό των αντικειμένων που η πιθανότητα της τομής τους με τα ερωτήματα δεν ξεπερνάει το κατώφλι των ερωτημάτων. Όσο πιο μεγάλη ακτίνα έχουμε, για σταθερά χαρακτηριστικά πλέγματος, τόσο περισσότερα αντικείμενα θα χρειαστεί να εξετάσουμε που έχουν κοινή τομή με το ερωτήματα. Έτσι μειώνεται αυτό το ποσοστό.

Το ποσοστό των αντικειμένων για τα οποία γίνεται αναλυτική αποτίμηση μειώνεται με την αύξηση της ακτίνας των κύκλων των αντικειμένων. Αυτό γίνεται γιατί όσο πιο μεγάλη ακτίνα έχουμε, για σταθερά χαρακτηριστικά ερωτημάτων, τόσο περισσότερα αντικείμενα θα χρειαστεί να εξετάσουμε που έχουν κοινή τομή με το ερωτήματα τα οποία θα κοπούν στο κλάδεμα.

Σχήμα 5.8: Χρόνος εκτέλεσης για διάφορες ακτίνες και μεγέθη πλέγματος

Το κατώφλι των ερωτημάτων φαίνεται από την εικόνα 5.8 πως δεν επηρεάζει το ποσοστό των αντικειμένων που πέφτουν εξ ολοκλήρου μέσα στα ερωτήματα ανά χρονόσημο και το ποσοστό των αντικειμένων που ανήκουν στα κελιά που τέμνουν τα ερωτήματα αλλά δεν έχουν κοινή τομή με τα ερωτήματα.

Δοκιμάστηκαν τιμές κατωφλίων 0.6, 0.7, 0.8, 0.95 για ακτίνα 1000μ. και τιμές κατωφλίων 0.8, 0.95 για ακτίνα 2000μ. Το κατώφλι των ερωτημάτων φαίνεται από την εικόνα 5.8 πως επηρεάζει σημαντικά το ποσοστό των αντικειμένων που η πιθανότητα της τομής τους με τα ερωτήματα δεν ξεπερνάει το κατώφλι των ερωτημάτων ανά χρονόσημο. Αυτό είναι αναμενόμενο αφού όσο μεγαλύτερο κατώφλι έχουμε τόσο περισσότερο κλάδεμα αντικειμένων γίνεται, επομένως αυξάνεται το συγκεκριμένο ποσοστό. Αντιθέτως, όσο μικρότερο κατώφλι έχουμε τόσο περισσότερα αντικείμενα θα αποτιμώνται λεπτομερώς, μειώνοντας το ποσοστό αυτό.

Από την εικόνα 5.8, ακόμα φαίνεται ότι η μείωση του κατώφλι των ερωτημάτων αυξάνει σημαντικά το ποσοστό των αντικειμένων για τα οποία γίνεται αναλυτική αποτίμηση. Αυτό είναι αναμενόμενο αφού όσο μικρότερο κατώφλι έχουμε τόσο λιγότερο κλάδεμα αντικειμένων γίνεται, επομένως αυξάνεται το ποσοστό των αντικειμένων που αποτιμούνται λεπτομερώς. Αντιθέτως, όσο μεγαλύτερο κατώφλι έχουμε τόσο λιγότερα αντικείμενα θα αποτιμώνται λεπτομερώς, μειώνοντας το ποσοστό αυτό.

5.3.6 Σύγκριση χρόνων εκτέλεσης εξαντλητικού και προσεγγιστικού αλγορίθμου.

Συγκρίνουμε πειραματικά τον χρόνο εκτέλεσης του προσεγγιστικού αλγορίθμου που αναπτύχθηκε στο προηγούμενο κεφάλαιο με τον χρόνο εκτέλεσης ενός αλγορίθμου που δίνει λεπτομερείς απαντήσεις. Ο αλγόριθμος αυτός χρησιμοποιεί τον κάνναβο και τις δύο πρώτες συνθήκες κλαδέματος του προσεγγιστικού αλγορίθμου αλλά δεν χρησιμοποιεί καθόλου το κλάδεμα. Βασίζεται μόνο σε ένα στά-

Σχήμα 5.9: Σύγκριση εξαντλητικού με προσεγγιστικό αλγόριθμο για κατώφλια 95% και 80%.

διο επεξεργασίας και δεν περιέχει στάδια προεπεξεργασίας. Τέλος η λεπτομερής αποτίμηση γίνεται με τη μέθοδο Monte Carlo για 100 επαναλήψεις. Κάτι τέτοιο δεν θα δώσει πολύ καλά αποτελέσματα αφού θα χρειαζόταν αριθμός επαναλήψεων της τάξης των 10.000 ή και 100.000 δειγμάτων ώστε να θεωρείται ακριβής η κάθε αποτίμηση. Παρόλα αυτά βρέθηκε ότι ακόμα και με αριθμό επαναλήψεων 100 η εξαντλητική αποτίμηση είναι αρκετά πιο αργή όσον αφορά το χρόνο επεξεργασίας. Η δοκιμή έγινε για R=1000, c=10, a=0.01 και έδωσε μέσο χρόνο εκτέλεσης 213.207 sec. Για τα ίδια χαρακτηριστικά και κατώφλια ερωτημάτων 0.6, 0.7, 0.8, 0.95 είχαμε βρει μέσο χρόνο εκτέλεσης 22.5262, 21.2862, 19.6771, 18.1261 sec αντίστοιχα.

5.3.7 Ακρίβεια προσέγγισης

Για τα 10000 αντικείμενα διαλέγουμε ενδεικτικά 20 συγκεκριμένα ερωτήματα και για χρονικές στιγμές t=0, 40, 80, 120, 160, 200 βρίσκουμε όλα τα αντικείμενα που τα διαστήματα εμπιστοσύνης που σχηματίζουν με το ερώτημα ξεπερνούν τα κατώφλια των 80% και των 95%. Για αυτές τις συγκεκριμένες χρονικές στιγμές και αντικείμενα χρησιμοποιούμε τον εξαντλητικό αλγόριθμο (παρατίθεται πιο κάτω αναλυτικά) για να βρούμε την ακριβή πιθανότητα τομής των αντικειμένων με τα ερωτήματα. Την πιθανότητα αυτή την αντιπαραβάλλουμε με το αντίστοιχο διάστημα εμπιστοσύνης. Βρέθηκε ότι το ποσοστό των εξαντλητικών πιθανοτήτων που πέφτουν μέσα στα διαστήματα εμπιστοσύνης ξεπερνάει το 92% για κατώφλι 80% και το 94% για κατώφλι 95%. Πιο συγκεκριμένα, όπως φαίνεται στο σχήμα 5.9 αυτό έγινε όπως φαίνεται στο σχήμα για 262874 απο τα 281867 αντικείμενα για κατώφλι 80% και για 439045 απο τα 476614 αντικείμενα για κατώφλι 95%. Αυτό μας οδηγεί στο συμπέρασμα ότι κάποια αντικείμενα που η πραγματική πιθανότητα τομής τους με τα ερωτήματα είναι μικρότερη από το κατώφλι των ερωτημάτων δίνονται ως απάντηση. Έτσι έχουμε μια υπερεκτίμηση στις απαντήσεις. Επιπροσθέτως βρέθηκε ότι τα αντικείμενα που δίνει ο εξαντλητικός αλγόριθμος ως απάντηση και

δεν δίνει ο προσεγγιστικός (False negatives) δεν ξεπερνούν το 0,08% για κατώφλι 95% και το 0.12% για κατώφλι 80%. Τέλος για τα, συγκεκριμένα πειραματικά δεδομένα, για κατώφλι 95% βρέθηκε ότι ο μέσο όρος της μέσης τιμής του διαστήματος εμπιστοσύνης, δηλαδή το $avg(\frac{\theta_{max}+\theta_{min}}{2})$, διαφέρει κατά 1.43% από το μέσο όρο των αντίστοιχων εξαντλητικών κατωφλίων $\theta_{analytic}$ ενώ για κατώφλι 80% η αντίστοιχη διαφορά είναι 1.87%.

Κάποιοι τρόποι έτσι ώστε να αυξηθεί η ακρίβεια του αλγορίθμου είναι οι παρακάτω:

- Αύξηση του αριθμού λ των κουτιών του εγγεγραμμένου τετραγώνου για μεγαλύτερη ακρίβεια, αν και κάτι τέτοιο συνεπάγεται αύξηση στον χρόνο εκτέλεσης.
- Αύξηση αριθμού επαναλήψεων της μεθόδου Monte Carlo του εξαντλητικού αλγορίθμου για μεγαλύτερη ακρίβεια (π.χ. 100000 επαναλήψεις).

Algorithm 4 Αλγόριθμος χωρίς κλάδεμα με λεπτομερείς απαντήσεις

Procedure *ExhaustiveAlgorithm*
Είσοδος: Υποθέτουμε σύνολο M ερωτημάτων $q_i = (x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max}, t, \theta_i)$, N αντικειμένων $o_j = (c_j, R_j, t)$ και κανονική κατανομή .
Έξοδος: $\forall q_i : \{< o_j, \theta_j^{analytical} >: \bigcap(o_j, q_i)$ με$((\theta_i < \theta_j^{analytical})\vee)\}$
$G \leftarrow$ Κάνναβος με $c \times c$ τετραγωνικά κελιά. *//Χωρικό ευρετήριο*
Σε κάθε χρονόσημο t:
for κάθε $g \in G$ **do**
  Αντιστοίχησε στο g κάθε q_i και o_j που επικαλύπτει το g
  for κάθε q_i **do**
    $C_i = \{o_j$:που επικαλύπτεται με τα κελιά του $q_i\}$. *//Υποψήφιο σύνολο*
    for κάθε $o_j \in C_i$ **do**
      if (o_jείναι ακίνητο $\wedge$ q_i είναι ακίνητο) **then**
        συνέχισε ; *//Αποφεύγουμε την αποτίμηση ακίνητων οντοτήτων*
      else if $q_i \supset inbox(o_j)$ **then**
        Δώσε ως απάντηση $< o_j, 100 >$
      else if $\bigcap(q_i,$ inbox$(o_j))= \emptyset$ **then**
        συνέχισε ; *//Pruning*
      else
        Χρησιμοποιώντας τη μέθοδο Monte Carlo βρες την πιθανότητα $\theta_j^{analytic}$
        if $(\theta_i < \theta_j^{analytic})$ **then**
          Δώσε ως απάντηση $< o_j, \theta_j^{analytic} >$ *//Ακριβής απάντηση*
        end if
      end if
    end for
  end for
end for
End Procedure

5.4 Ανασκόπηση πειραμάτων

Γενικά, η κλιμάκωση των τιμών κάθε παραμέτρου επιβαρύνει τις επιδόσεις του συστήματος στις περισσότερες περιπτώσεις. Η επιλογή του βαθμού κατάτμησης του καννάβου ($c \times c$) επηρεάζει σημαντικά τις επιδόσεις, ιδίως για μικρό και μεγάλο πλήθος κελιών. Πιο συγκεκριμένα για μικρό (π.χ. c=5) και μεγάλο(π.χ. c=20) πλήθος κελιών ο χρόνος εκτέλεσης αυξάνεται. Έτσι επιλέγεται c=10 ώστε το πλέγμα να είναι όσο το δυνατόν αποτελεσματικότερο. Επιπλέον, παρατηρήθηκαν τα εξής:

- Όσο μεγαλύτερο κατώφλι τόσο μικρότερος ο χρόνος εκτέλεσης, αφού όλο και λιγότερα αντικείμενα αποτιμώνται λεπτομερώς.
- Όσο μεγαλύτερη ακτίνα τόσο αυξάνεται ο χρόνος εκτέλεσης αφού τα αντικείμενα αποθηκεύονται σε περισσότερα κελιά.
- Όσο μεγαλύτερες οι διαστάσεις των ερωτημάτων τόσο αυξάνεται ο χρόνος εκτέλεσης, αφού περισσότερα αντικείμενα εξετάζονται για κάθε ερώτημα.

Τέλος έγινε σύγκριση του αλγορίθμου με άλλον που κάνει εξαντλητικούς υπολογισμούς πιθανοτήτων τομής ερωτημάτων με αντικείμενα με τη χρήση της μεθόδου $MonteCarlo$ και βρέθηκε ότι ο προσεγγιστικός αλγόριθμος είναι πολύ γρηγορότερος (περίπου 10 φορές) και δίνει τις περισσότερες φορές (της τάξης 90-95%) διαστήματα εμπιστοσύνης που περιέχουν τις ακριβείς πιθανότητες.

Κεφάλαιο 6

Συμπεράσματα

Οι πρόσφατες ραγδαίες εξελίξεις στις τεχνολογίες εντοπισμού της γεωγραφικής θέσης αύξησαν το ενδιαφέρον για την ανάπτυξη εφαρμογών παρακολούθησης κινούμενων αντικειμένων, γνωστές με το όνομα Υπηρεσίες Εντοπισμού (Location Based Services). Οι χρήστες τέτοιων εφαρμογών αποστέλλουν τη θέση τους περιοδικά και είναι σε θέση να υποβάλλουν πολλαπλά ερωτήματα διαρκείας σε ένα κεντρικό επεξεργαστή, μεταξύ άλλων και ερωτήματα περιοχής. Τα ερωτήματα περιοχής εντοπίζουν τους χρήστες που βρίσκονται εντός μιας προσδιορισμένης γεωγραφικής έκτασης. Τα ερωτήματα είναι δυνατόν να είναι κινούμενα: η έκταση ή/και το κέντρο των περιοχών μπορεί να μεταβάλλεται, ενώ οι χρήστες μπορεί επίσης να κινούνται. Τα κινούμενα αντικείμενα, τα οποία διαθέτουν δυνατότητα γεωγ-ραφικού εντοπισμού (GPS), δεν θέλουν να αποκαλύπτουν το στίγμα τους στον κεντρικό υπολογιστή. Ωστόσο, θεωρείται γνωστή η ευρύτερη περιοχή του συμβάντος. Η πιθανότητα εκδήλωσης του γεγονότος (αβεβαιότητα) δεν θεωρείται ομοιόμορφη, αλλά μπορεί να ποικίλλει. Οι χρήστες μπορούν να υποβάλλουν τα χωρικά ερωτήματα διαρκείας για περιοχές ενδιαφέροντός τους, οπότε ο επεξεργαστής οφείλει να συνεκτιμήσει όλα τα δεδομένα και να παρέχει προσεγγιστικές απαντήσεις στους χρήστες αλλά με αρκετά καλή ακρίβεια. Τα πιθανοτικά ερωτήματα περιοχής υπάρχουν στη συντριπτική πλειοψηφία των εφαρμογών παρακολούθησης και απαιτείται να λειτουργούν σε πραγματικό χρόνο, εφαρμοζόμενοι διαρκώς πάνω στο ρεύμα ενημερώσεων των θέσεων των χρηστών. Το πλήρως δυναμικό μοντέλο τέτοιων συστημάτων διαφοροποιεί το χειρισμό τους από τις συμβατικές βάσεις δεδομένων, θέτοντας στόχους όπως:

- Η υποστήριξη ολοένα και μεγαλύτερου –ενδεχομένως κυμαινόμενου– πλήθους αντικειμένων και ερωτημάτων.
- Η συχνότερη καταγραφή των θέσεων των αντικειμένων με σκοπό τη μεγαλύτερη ακρίβεια στην τήρηση της τροχιάς τους.
- Η επεξεργασία ερωτημάτων σε πραγματικό χρόνο.

Σκοπός της διπλωματικής εργασίας ήταν η ανάπτυξη εφαρμογής για γρήγορη αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων α-

ντικειμένων. Ο αριθμός των ερωτημάτων και των αντικειμένων είναι μεγάλος. Ο αλγόριθμος στον οποίο στηρίζεται η υλοποίηση της εφαρμογής είναι προσεγγιστικός και σκοπός της ανάπτυξής του ήταν τα αποτελέσματα που δίνει να είναι όσο το δυνατόν πιο ακριβή. Από την μελέτη ζητημάτων σχεδίασης του συστήματος, προκύπτουν αξιοσημείωτες παρατηρήσεις:

- Η επιλογή του κατακερματισμού ως μεθόδου προσπέλασης, αποδείχτηκε ιδανική για την διαρκή παρακολούθηση της θέσης των αντικειμένων. Για μικρό πλήθος κελιών, ο κάνναβος είναι σχεδόν άχρηστος αφού αποτυγχάνει να διακρίνει επαρκώς την συνάφεια των αντικειμένων με ερωτήματα. Απ΄ την άλλη, υπερβολικός κατακερματισμός σε κελιά οδηγεί σε αυξημένο διαχειριστικό κόστος, αφού η κυκλική κανονική κατανομή κάθε αντικειμένου θα αναφέρεται σε πολλά κελιά. Αυτό συμβαίνει επειδή αντικείμενα με μεγάλες ακτίνες αποθηκεύονται σε περισσότερα κελιά από αντικείμενα με μικρότερες ακτίνες. Μια μέση κατάτμηση σε 100 κελιά αποδεικνύεται προτιμότερη για την εξεταζόμενη περίπτωση. Η επιλογή αυτή δεν είναι η καλύτερη δυνατή για όλες τις ακτίνες π.χ. βλέπουμε ότι για ακτίνα 200μ. c =15 είναι πιο γρήγορη επιλογή. Παρόλα αυτά η επιλογή c =10 είναι πολύ καλή για το σύνολο των περιπτώσεων.

- Οι τεχνικές κλαδέματος βάσει γεωμετρικών και πιθανοτικών χαρακτηριστικών επέφεραν σημαντικά πλεονεκτήματα:

  - Μικρός αριθμός αντικειμένων τελικά να απαιτούν αναλυτική αποτίμηση. Για μεγαλύτερα κατώφλια ερωτημάτων, ο συνολικός χρόνος αποτίμησης μειώνεται αφού περισσότερα αντικείμενα απορρίπτονται από τις τεχνικές κλαδέματος.
  - Ικανοποιητικές επιδόσεις για κλιμακούμενα πλήθη αντικειμένων και ερωτημάτων. Ο αλγόριθμος που υλοποιήθηκε, αποτιμά αποδοτικά κυμαινόμενο πλήθος ενεργών ερωτημάτων σε κυμαινόμενο πλήθος αντικειμένων.

- Αποδεκτή ακρίβεια για διαφορετικά μεγέθη περιοχών και βαθμών αβεβαιότητας. Ο προσεγγιστικός αλγόριθμος συγκρίθηκε με τον αντίστοιχο εξαντλητικό και βρέθηκαν ενθαρρυντικά αποτελέσματα ως προς την ακρίβεια. Τα αποτελέσματα του προσεγγιστικού αλγορίθμου είναι υπερεκτίμηση εκείνων του εξαντλητικού. Αρκετά από αυτά που δεν θα δίνονταν ως απάντηση από τον εξαντλητικό δίνονται από τον προσεγγιστικό. Επιπροσθέτως βρέθηκε ότι τα αντικείμενα που δίνει ο εξαντλητικός αλγόριθμος ως απάντηση και δεν δίνει ο προσεγγιστικός (False negatives) είναι λιγότερα από 1% των αποτελεσμάτων.

Από τη μελέτη του συγκεκριμένου προβλήματος, προκύπτουν ενθαρρυντικές προοπτικές επέκτασής του. Προφανώς, θα ήταν δυνατόν να μελετηθεί η μοντελοποίηση άλλων προβλημάτων (π.χ. καιρικά φαινόμενα) που στηρίζονται σε άλλες κατανομές, όπως την ομοιόμορφη ή κάποια άλλη κανονική κατανομή με διαφορετικές παραμέτρους. Είναι σαφές ότι ένα τέτοιο ολοκληρωμένο σύστημα

διαχείρισης ρευμάτων κινούμενων αντικειμένων θα μπορούσε να αξιοποιηθεί σε πρόσθετους πιθανοτικούς τύπους ερωτημάτων:

- *Ερωτήματα πυκνότητας* (*density queries*) αναζητούν συμπαγείς περιοχές με πυκνότητα σημειακών θέσεων πάνω από ένα προσδιορισμένο κατώφλι.
- *Ερωτήματα ανάστροφου εγγύτερου γείτονα* (*reverse nearest neighbors*), αναζητώντας το αντικείμενο του οποίου εγγύτερος γείτονας είναι το αντικείμενο ενδιαφέροντος.
- *Συναθροιστικά ερωτήματα εγγύτητας* (*aggregate NN queries*), τα οποία εντοπίζουν τα αντικείμενα που ελαχιστοποιούν τη συνολική τους απόσταση από ένα σύνολο σημειακών θέσεων ενδιαφέροντος.

Κεφάλαιο 7

Επίμετρο

7.1 Υλοποίηση Αλγορίθμου

Η υλοποίηση του αλγορίθμου πραγματοποιήθηκε σε περιβάλλον C++. Η συγκεκριμένη γλώσσα είναι πολύ εύχρηστη για την ανάπτυξη εφαρμογών σε αντικειμενοστρεφές περιβάλλον. Οι μέθοδοι που υλοποιήθηκαν περιγράφονται στη συνέχεια.

7.1.1 Κύριες δομές αλγορίθμου

- TObjChain: Λίστα που περιέχει τα αντικείμενα. Κάθε αντικείμενο περιέχει τα εξής πεδία:ts που συμβολίζει την χρονική στιγμή, id που συμβολίζει την ταυτότητά του, fresh για το αν έχει ανανεωθεί η θέση του, checked για το αν έχει εξεταστεί ήδη από ένα συγκεκριμένο ερώτημα μία δεδομένη χρονική στιγμή, $x_{min}, y_{min}, x_{max}, y_{max}$ για τις συντεταγμένες του και r για την ακτίνα του.

- TObjChain: Λίστα που περιέχει τα ερωτήματα. Κάθε ερώτημα περιέχει τα εξής πεδία:ts που συμβολίζει την χρονική στιγμή, id που συμβολίζει την ταυτότητά του, fresh για το αν έχει ανανεωθεί η θέση του, $x_{min}, y_{min}, x_{max}$, y_{max} για τις συντεταγμένες του και threshold για το κατώφλι που συμβολίζει την επιθυμητή πιθανότητα.

- GridCell: Δομή που αποτελείται από τη λίστα με τα αντικείμενα TObjChain και μία μεταβλητή processed: που δείχνει αν το αντικείμενο έχει τοποθετηθεί κάπου.

- ObjAssignments: Λίστα που δείχνει σε το κελί στο οποίο ένα αντικείμενο έχει προσωρινά αποθηκευτεί.

7.1.2 Κύριες μέθοδοι αλγορίθμου προεπεξεργασίας

SplitCircles

Η μέθοδος αυτή εκτελεί το πρώτο στάδιο της προεπεξεργασίας, δηλαδή προϋπολογίζει για κάθε αντικείμενο τις πιθανότητες $p(b_k)$ των εσωτερικών τετραγώνων b_k των εγγεγραμμένων τετραγώνων της κυκλικής κανονικής κατανομής κάθε αντικειμένου o_i . Δέχεται δύο παραμέτρους, τον αριθμό l των τετραγώνων στον οποίο θα διαιρεθεί το εγγεγραμμένο τετράγωνο και τον αριθμό των δειγμάτων (steps) που θα ληφθούν για την επανάληψη της μεθόδου Monte Carlo . Όσο μεγαλύτερος ο αριθμός της παραμέτρου steps τόσο μεγαλύτερη ακρίβεια θα επιτευχθεί. Η τάξη μεγέθους για το λ είναι 5-12 και για το steps είναι 10.000, ένα νούμερο που δίνει πολύ ακριβή αποτελέσματα. Η ακτίνα του κύκλου δεν δίδεται ως παράμετρος αφού τα αποτελέσματα θα είναι τα ίδια. Έτσι ως δοκιμαζόμενη ακτίνα επιλέχτηκε το 1χμ.

Normalization

Η μέθοδος αυτή βοηθάει και αυτή το πρώτο στάδιο της προεπεξεργασίας. Δέχεται ως παράμετρο μόνο τον αριθμό l των τετραγώνων στον οποίο θα διαιρεθεί το εγγεγραμμένο τετράγωνο και ο σκοπός της είναι τα αποτελέσματα των πιθανοτήτων που υπολογίστηκαν στην μέθοδο SplitCircles να κανονικοποιηθούν στη μονάδα. Η διαδικασία αυτή γίνεται διαιρώντας όλα αυτά τα αποτελέσματα με το άθροισμά του.

Calculate Threshold Areas

Η μέθοδος αυτή εκτελεί το δεύτερο στάδιο της προεπεξεργασίας, δηλαδή προϋπολογίζει για κάθε ακτίνα το μέγιστο εμβαδό που αντιστοιχεί σε συγκεκριμένες πιθανότητες – κατώφλια θ_i) που δίνουν τα ερωτήματα. Τα κατώφλια αυτά παίρνουν τιμές από 50% μέχρι 95% με βήμα 5% (10 κατώφλια σύνολο). Δεδομένου ότι εξετάζουμε κύκλους (των κυκλικών κανονικών κατανομών) με ακτίνες 200μ. μέχρι 3000χμ. με βήμα 200μ. (δηλαδή 15 ακτίνες) θα παραχθούν 150 συνολικά αριθμοί εμβαδών. Η μέθοδος δέχεται τρεις παραμέτρους, τον αριθμό threshold number ο οποίος κωδικοποιεί το κατώφλι το οποίο εξετάζεται (πχ. 1 για το 50%, 2 για το 55% κτλ), τον αριθμό των δειγμάτων (steps) που θα ληφθούν για την επανάληψη της μεθόδου Monte Carlo και την ακτίνα του κύκλου που θέλουμε να εξετάσουμε. Όσο μεγαλύτερος ο αριθμός της παραμέτρου steps τόσο μεγαλύτερη ακρίβεια θα επιτευχθεί. Ο υπολογισμός είναι προσεγγιστικός και γίνεται ως εξής: για κάθε ακτίνα που εξετάζεται, υπολογίζεται η πιθανότητα όλως των μικρότερων κύκλων που ο κύκλος περιέχει με ακτίνες από 0μ. μέχρι την ακτίνα του κύκλου με βήμα 10μ. Εφόσον ο κύκλος αυτός περιέχει πιθανότητα ίση με εκείνη του κατωφλίου τότε κρατάμε το εμβαδόν του, αλλιώς απορρίπτεται.

7.1.3 Κύριες μέθοδοι αλγορίθμου επεξεργασίας

RegularGrid (με τέσσερις παραμέτρους)

Η μέθοδος RegularGrid , με τέσσερις παραμέτρους, ουσιαστικά δημιουργεί το πλέγμα και υπολογίζει το μήκος των διαστάσεων, dx για το μήκος και dy για το πλάτος, του κάθε κελιού. Δέχεται τέσσερις παραμέτρους οι οποίες είναι:

- Το συνολικό μήκος του πλέγματος pHeight .
- Το συνολικό πλάτος του πλέγματος pWidth .
- Τον αριθμό των κελιών που χωρίζουμε το μήκος του πλέγματος pGrandX .
- Τον αριθμό των κελιών που χωρίζουμε το πλάτος του πλέγματος pGrandY .

RegularGrid

Η μέθοδος αυτή δε δέχεται παραμέτρους και ο σκοπός της είναι στο τέλος της επεξεργασίας των δεδομένων να αδειάσει όλα τα κελιά από τα αντικείμενα που περιέχουν εκείνη την χρονική στιγμή και τελικά να καταστρέψει το πλέγμα.

Allocate

Η μέθοδος Allocate δέχεται μόνο μία παράμετρο, το συνολικό αριθμό των αντικειμένων που μπορούν να αποθηκευτούν στο πλέγμα probj cnt. Εφόσον ο αριθμός των αντικειμένων υπερβαίνει αυτόν τον αριθμό τότε επιστρέφεται λάθος κατά τη διάρκεια της εκτέλεσης, αλλιώς αρχικοποιεί τα κελιά του πλέγματος.

Hash

Η Hash είναι η μέθοδος που χαρακτηρίζει το πλέγμα ως ευρετήριο. Δέχεται δύο παραμέτρους, τις συντεταγμένες ενός σημείου πάνω στον χάρτη (x, y) και επιστρέφει τον αριθμό του κελιού στον οποίο ανήκει. Η αντιστοιχία γίνεται με τον αλγόριθμο: αριθμός κελιού $= GranX * \lfloor \frac{y}{dy} \rfloor + GranY * \lfloor \frac{x}{dx} \rfloor$.

UpdateObject

Η UpdateObject δέχεται ως παράμετρο τα καινούργια χαρακτηριστικά ενός αντικειμένου και χρησιμοποιώντας τις συντεταγμένες (κάτω αριστερά (x_{min}, y_{min}) και πάνω δεξιά(x_{max}, y_{max})) του εγγεγραμμένου τετραγώνου του το τοποθετεί στα αντίστοιχα κελιά του πλέγματος με τα οποία είτε τέμνεται είτε τα εμπεριέχει στο εσωτερικό του.

UpdateQuery

Η UpdateQuery δέχεται ως παράμετρο τα καινούργια χαρακτηριστικά ενός ερωτήματος και χρησιμοποιώντας τις συντεταγμένες του(κάτω αριστερά (x_{min}, y_{min}) και πάνω δεξιά (x_{max}, y_{max}) εξετάζει ένα προς ένα τα αντικείμενα που βρίσκονται στα κελιά του πλέγματος με τα οποία είτε τέμνεται είτε τα εμπεριέχει στο εσωτερικό του. Η εξέταση κάθε αντικειμένου γίνεται μία φορά παρόλο που μπορεί να βρίσκεται σε περισσότερα από ένα κελιά χρησιμοποιώντας την μέθοδος Check Objects. Η εξέταση των αντικειμένων γίνεται με την βοήθεια της μεθόδου getCellObjects .

getCellObjects

Η μέθοδος getCellObjects δέχεται ως παράμετρο τον αριθμό ID του κελιού, τα αντικείμενα του οποίου θέλουμε να εξετάσουμε. Για το κελί αυτό εξετάζουμε όλα τα αντικείμενα που περιέχει. Στην μέθοδο αυτή εκτελούνται και οι τρεις τεχνικές κλαδέματος, οι πρώτες δύο με απλές συνθήκες και η τρίτη με την χρήση της μεθόδου Find Intersection Area. Τέλος εφόσον ένα αντικείμενο περάσει και τις τρεις αυτές συνθήκες γίνεται η τελική αποτίμησή του στην μέθοδο Find Probabilities .

Check Objects

Η χρήση της μεθόδου αυτής είναι η μη εξέταση των αντικειμένων που βρίσκονται σε παραπάνω από ένα κελί περισσότερες από μία φορές. Δέχεται μόνο μία παράμετρο, τον αριθμό ID του κελιού.

Find Intersection Area

Η μέθοδος αυτή βρίσκει το εμβαδό της τομής του ερωτήματος με το εγγεγραμμένο τετράγωνο της κυκλικής κανονικής κατανομής των αντικειμένων. Το αποτέλεσμα αυτό χρησιμοποιείται κατά τη διάρκεια του τρίτου κλαδέματος που αναφέρθηκε στον αλγόριθμο. Δέχεται έξι παραμέτρους. Τα χαρακτηριστικά του ερωτήματος, τα χαρακτηριστικά του αντικειμένου και τέσσερις επιπλέον αριθμός. Οι αριθμοί αυτοί είναι μηδενικά αρχικά και εφόσον υπάρχει κοινή τομή (ορθογώνιο) θα αποτελέσουν τις συντεταγμένες του κάτω αριστερά άκρου και τις συντεταγμένες του πάνω δεξιά άκρου. Εφόσον δεν υπάρχει τομή επιστρέφονται μηδενικά.

Find Probabilities

Η μέθοδος αυτή βρίσκει το τελικό περιθώριο εμπιστοσύνης της τομής του ερωτήματος με το εγγεγραμμένο τετράγωνο της κυκλικής κανονικής κατανομής των αντικειμένων και αποτελεί την τελική αποτίμηση. Δέχεται έξι παραμέτρους. Τα χαρακτηριστικά του αντικειμένου, τις κοινές του συντεταγμένες με το ερώτημα και τον αριθμό λ των τετραγώνων στον οποίο θα διαιρεθεί το εγγεγραμμένο τετράγωνο. Η χρησιμότητά της είναι να βρει πια εσωτερικά τετράγωνα b_k περιέχονται εξ ολοκλήρου και ποια τέμνονται από το ερώτημα. Με βάση αυτά και τα αποτελέσματα

του πρώτου σταδίου της προεπεξεργασίας επιστρέφει δύο αριθμούς-πιθανότητες, δηλαδή το προσεγγιστικό περιθώριο εμπιστοσύνης που θέλαμε να βρούμε.

Βιβλιογραφία

[1] A. Arasu, S. Babu, and J. Widom. The CQL Continuous Query Language: Semantic Foundations and Query Execution. VLDB Journal, 15(2): 121-142, 2006.

[2] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and Issues in Data Stream Systems. In Proceedings of the 21st ACM SIGACTQSIGMODQSIGART Symposium on Principles of Database Systems, Madison, Wisconsin, May 2002.

[3] S. Babu and J.Widom. ContinuousQueries over Data Streams. ACM SIGMOD Record, 30 (3):109-120, September 2001.

[4] J. Chen and R. Cheng. Efficient Evaluation of Imprecise Location-Dependent Queries. In Proceedings of the 23th IEEE International Conference on Data Engineering, pp. 586-595, Istanbul, Turkey, April 2007.

[5] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating Probabilistic Queries over Imprecise Data. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD 2003), pp. 551-562, San Diego, California, June 2003.

[6] N. Dalvi, C. Re, and D. Suciu. Probabilistic Databases: Diamonds in the Dirt. Communications of the ACM, 52(7):86-94, July 2009.

[7] R. Cheng, S. Prabhakar, and D. Kalashnikov. Querying Imprecise Data in Moving Object Environments. In Proceedings of the 19th International Conference on Data Engineering (ICDE'03), pp.723-725, Bangalore, India, 2003.

[8] X. Dai, M. L. Yiu, N. Mamoulis, Y. Tao, M. Vaitis. Probabilistic Spatial Queries on Existentially Uncertain Data. In Proceedings of the 9th International Symposium on Spatial and Temporal Databases (SSTD 2005), pp. 400-417, Angra dos Reis, Brazil, August 2005.

[9] M. Erwig, R.H. Guting, M. M. Schneider, and M.Vazirgiannis. Abstract and Discrete Modeling of Spatio-Temporal Data Types. In Proceedings of the 6th ACM Symposium on Geographic InformationSystems, Washington DC, pp.131-136, November 1998.

[10] R. H. Guting, M. H. Bohlen, M. Erwig, C. S. Jensen, N.A. Lorentzos, M. Schneider, and M. Vazirgiannis. A Foundation for Representing and Querying Moving Objects. ACM Transactions on Database Systems, 2000.

[11] B. Gedik and L. Liu. MobiEyes: Distributed Processing of Continuously Moving Queries on Moving Objects in a Mobile System. In Proceedings of the 9th International Conference on Extending Database Technology (EDBT'04), Heraklion (Crete), Greece, March 2004.

[12] V. Gaede, and O. Gunther. MultidimensionalAccess Methods. ACMComputing Surveys, 30 : 170-231, 1998.

[13] Lukasz Golab and M. Tamer Ozsu. Issues in data stream management. SIGMOD Record, 32(2):5-14, 2003.

[14] C. S. Jensen, D. Lin, B. Chin Ooi, and R. Zhang. Effective Density Queries on Continuously Moving Objects. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), pp. 71-81, Atlanta, Georgia, USA, April 2006.

[15] H.-P. Kriegel, P. Kunath, M. Pfeifle, and M. Renz. Probabilistic Similarity Join on Uncertain Data. In Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DASFAA 2006), pp. 295-309, Singapore, April 2006.

[16] H.-P. Kriegel, P. Kunath and M. Renz. Probabilistic Nearest-Neighbor Query on Uncertain Objects. In Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA 2007), pp. 337-348, Bangkok, Thailand, April 2007.

[17] X. Lian and L. Chen. Similarity Join Processing on Uncertain Data Streams. IEEE TKDE, 2011 .

[18] M.F. Mokbeland W.G. Aref. GPAC: Generic and Progressive Processing of Mobile Queries over Mobile Data.In Proceedings of the 6th international conference on Mobile data management (MDM'05), pp. 155-163, Ayia Napa, Cyprus, May 2005.

[19] R. Motwani , J. Widom, A. Arasu, B. Babcock, S. Babu,M. Datar, G. Manku, C. Olston, J. Rosenstein and R.Varma. QueryProcessing, Approximation, and Resource Management in a Data StreamManagement System. In Proceedings of the 2003 Conference onInnovative Data Systems Research (CIDR), January 2003.

[20] J. Ni, C. Ravishankar, and B. Bhanu. Probabilistic Spatial Database Operations. In Proceedings of the 8th International Symposium on Advances in Spatial and Temporal Databases (SSTD 2003), pp. 140-158, Santorini Island, Greece, July 2003.

[21] C. Re, N. Dalvi, and D. Suciu. Efficient Top-k Query Evaluation on Probabilistic Data. In Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007), pp. 886-895, Istanbul, Turkey, April 2007.

[22] M. Stonebraker, U. Cetintemel, and S. Zdonik. The 8 Requirements of Real-Time Stream Processing. ACM SIGMOD Record, 34(4):42-47, December 2005.

[23] S. Singh, C. Mayfield, R. Shah, S. Prabhakar, S. Hambrusch, J. Neville, and R. Cheng. Database Support for Probabilistic Attributes and Tuples. In Proceedings of the 24th IEEE International Conference on Data Engineering, pp. 1053-1061, Cancun , Mexico, April 2008.

[24] Z. Song and N. Roussopoulos. Hashing Moving Objects. In Proceedings of the 2nd International Conference on Mobile Data Management (MDM'01), pp. 161-172, Hong Kong, China, January 2001.

[25] Y. Tao, R. Cheng, X. Xiao, W. Ngai, B. Kao and S. Prabhakar. Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions. In Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005), pp. 922-933, Trondheim, Norway, September 2005.

[26] Y. Theodoridis. Ten Benchmark Database Queries for Location-based Services. Computer Journal, 46(6): 713-725, 2003.

[27] O. Wolfson, B. Xu, S. Chamberlain, L. Jiang. MovingObjectsDatabases: Issues and Solutions. In Proceedings of 10thInternationalConference on Scientific and Statistical DatabaseManagement (SSDB 1998),pp. 111-122, Capri, Italy, July 1998.

[28] T. Xia and D. Zhang. Continuous Reverse Nearest Neighbor Monitoring. In Proceedings of 22nd International Conference on Data Engineering (ICDE'06), Atlanta, Georgia, USA, April 2006.

[29] X. Yu and S. Mehrotra. Capturing Uncertainty in Spatial Queries over Imprecise Data. In Proceedings of the 14th International Conference on Database and Expert Systems Applications (DEXA 2003), pp. 192-201, Prague, Czech Republic, September 2003.

[30] J. Zhang, M. Zhu, D. Papadias, Y. Tao, and D. L. Lee. Location-based Spatial Queries. In Proceedings of the 22nd ACM SIGMOD International Conference on Management of Data,pp. 443-454, San Diego, California, June 2003.

[31] W. Zhang, X. Lin, Y. Zhang, W. Wang, and J. Xu Yu. Probabilistic Skyline Operator over Sliding Windows. In Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE'09), pp.1060-1071, Shanghai, China, March 2009.

Ορολογία

αβεβαιότητα	uncertainty
αποτίμηση ερωτημάτων	query evaluation
βελτιστοποίηση ερωτημάτων	query optimization
δεικτοδότηση	indexing
ελάχιστο περιβάλλον παραλληλόγραμμο	minimum bounding box
επεξεργασία από κοινού	shared execution
ερώτημα διαρκείας	continuous query
ερώτημα εγγύτερου γείτονα	nearest-neighbor query
ερώτημα θέσης	location-based query
ερώτημα κορυφογραμμής	skyline query
ερώτημα περιοχής	range query
εφαρμογές παρακολούθησης	monitoring applications
κάνναβος	grid
κατακερματισμός	hashing
κινούμενο αντικείμενο	moving object
κλάδεμα	pruning
κλιμάκωση	scalability
παράθυρο κυλιόμενο	sliding window
περιθώριο εμπιστοσύνης	confidence margin
περίληψη	summary
ρεύμα δεδομένων	data stream
σταδιακή αποτίμηση	incremental evaluation
συνάθροιση	aggregation
σύνδεση	join
σύνοψη	synopsis
τελεστής	operator
υπηρεσίες εντοπισμού	location-based services
χρονόσημο	timestamp
χωρικές μέθοδοι προσπέλασης	spatial access methods

Αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων

Μάριος Παπαμιχάλης
papamixmarios@gmail.com

Διπλωματική εργασία στο Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων
*Επιβλέπων: Καθηγητής **Τ. Σελλής***

1 Γενικό πλαίσιο

Η εργασία στοχεύει στη μελέτη και υλοποίηση μιας εφαρμογής που θα επιτρέπει online απαντήσεις σε *πιθανοτικά ερωτήματα διαρκείας (probabilistic continuous queries)* σχετικά με τη θέση μεγάλου αριθμού αβέβαιων θέσεων κινούμενων αντικειμένων. Τα πιθανοτικά ερωτήματα διαρκείας θα τίθενται από διάφορους κινούμενους χρήστες που εγγράφονται στην υπηρεσία και επιθυμούν να ενημερώνονται οποτεδήποτε στην περιοχή τους συμβαίνει κάποιο έκτακτο γεγονός (π.χ. επίσκεψη φίλου). Τέτοια γεγονότα καταγράφονται σε κεντρικό υπολογιστή (server) αλλά με αβεβαιότητα ως προς την ακριβή γεωγραφική τους θέση. Οι κινούμενες συσκευές διαθέτουν δυνατότητα γεωγραφικού εντοπισμού (GPS) όμως το στίγμα κάθε αντικειμένου ποτέ δεν αποκαλύπτεται στον κεντρικό υπολογιστή. Ωστόσο, θεωρείται γνωστή η ευρύτερη περιοχή του συμβάντος. Η πιθανότητα εκδήλωσης του γεγονότος δεν θεωρείται ομοιόμορφη, αλλά μπορεί να ποικίλλει. Οι χρήστες μπορούν να υποβάλλουν τα χωρικά ερωτήματα διαρκείας για περιοχές ενδιαφέροντός τους, οπότε ο επεξεργαστής οφείλει να συνεκτιμήσει τις σχετικές πιθανότητες των προσφάτως καταγεγραμμένων συμβάντων και να δίνει τακτικά ενημερωμένες προσεγγιστικές απαντήσεις με κυμαινόμενη ποιότητα. Τέτοια στοιχεία θα μπορούσαν να αξιοποιηθούν σε εφαρμογές κοινωνικής δικτύωσης με κινητά τηλέφωνα, εκτίμηση περιβαλλοντικού κινδύνου σε φυσικές καταστροφές (λ.χ. διαρροή πετρελαίου), πρόγνωση μετεωρολογικών φαινομένων (λ.χ. τυφώνες) κ.ά.

Ως βασική υπόθεση της εργασίας θεωρείται η παρακολούθηση πολλών κινούμενων αντικειμένων και ερωτημάτων τα οποία ανανεώνουν συχνά την περιοχή της θέσης τους και δημιουργούν *ρεύματα δεδομένων* (*data streams*):

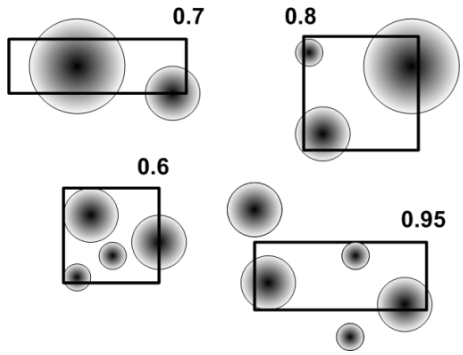
- Τα στοιχεία καταφθάνουν σε μεγάλο και ενδεχομένως μεταβλητό ρυθμό σε πραγματικό χρόνο (online).
- Τα ρεύματα έχουν μεγάλο μέγεθος, πιθανόν απεριόριστο.

Η φύση του μοντέλου συνηγορεί σε επιλογή προσεγγιστικών αλγορίθμων που εξοικονομούν κυρίως χρόνο, συνεκτιμώντας το σφάλμα που μπορεί να γίνει αποδεκτό από την εκάστοτε εφαρμογή.

2 Αβεβαιότητα δεδομένων

Τα τελευταία χρόνια εμφανίστηκε ένα ευρύ φάσμα εφαρμογών που σχετίζονται με την *αβεβαιότητα* (*uncertainty*). Τα αίτια της αβεβαιότητας στα δεδομένα ποικίλουν ανάλογα με την εφαρμογή. Για παράδειγμα, στην λειτουργία αισθητήρων η αβεβαιότητα οφείλεται στην ανακρίβεια λόγω σφαλμάτων κατά τη μέτρηση (π.χ. θερμοκρασία). Σε άλλες εφαρμογές, όπως είναι η προστασία της *ιδιωτικότητας* (*privacy*), υπάρχει η απαίτηση τα δεδομένα να είναι επίτηδες λιγότερο ακριβή. Η αβεβαιότητα συμβάλλει στην απόκρυψη προσωπικών δεδομένων, προστατεύοντας ευαίσθητα χαρακτηριστικά των ατόμων, έτσι ώστε μικρότερο μέρος στοιχείων να μπορεί να δημοσιευτεί. Ένα άλλο παράδειγμα βρίσκεται στα συστήματα εντοπισμού στίγματος (GPS). Το στίγμα ενός αντικειμένου (π.χ. αυτοκινήτου) δίνει την ακριβή θέση και ταχύτητά του κάθε χρονική στιγμή. Θα ήταν ασύμφορο για τον εντοπισμό του να στέλνει το στίγμα του πολύ συχνά, αφού με βάση την ταχύτητά του μπορούμε να ξέρουμε προσεγγιστικά πού βρίσκεται. Έτσι, η θέση του αντικειμένου όπως είναι γνωστή στο σύστημα, δεν ταυτίζεται πάντοτε με την τρέχουσα λόγω χρονικής υστέρησης κατά τη μετάδοση, οπότε θεωρείται αβέβαιη. Μέχρι πριν λίγα χρόνια, τα αβέβαια δεδομένα δεν είχαν καμία θέση στις παραδοσιακές, βάσεις δεδομένων, οι οποίες δεν ήταν προετοιμασμένες να τα αντιμετωπίσουν.

Σήμερα έχουν αναπτυχθεί πολλές μέθοδοι και αλγόριθμοι για την καλύτερη επεξεργασία ερωτημάτων, πολλά από τα οποία τα συναντάμε γενικά στις β.δ. και εμπεριέχουν πλέον την αβεβαιότητα. Ο υπολογισμός των περισσότερων αλγορίθμων γίνεται με αριθμητικές μεθόδους, με αποτέλεσμα τελικά να προκύπτουν προσεγγιστικές λύσεις. Για την καλύτερη μοντελοποίηση των προβλημάτων αυτών επιλέγεται αντίστοιχα η κατάλληλη αναπαράσταση των δεδομένων. Σε άλλες εργασίες ακο-



Σχήμα 1

λουθείται ο *συνεχής* τρόπος αναπαράστασης υποθέτοντας διάφορες κατανομές (ομοιόμορφη, Γκαουσιανή κτλ.) και σε άλλες ο *διακριτός* τρόπος με χρήση διακριτών δειγμάτων. Τέλος, διάφορες ερευνητικές προσπάθειες εστιάζουν στην καλύτερη παρουσίαση των αποτελεσμάτων (π.χ. περιθώρια εμπιστοσύνης, εκτίμηση σφάλματος κτλ.), ώστε να παρέχεται στους χρήστες μία εποπτική εικόνα κατά πόσο μπορούν να βασιστούν στις απαντήσεις που δίνονται.

Στο σημείο αυτό πρέπει να τονιστεί η διαφορά των πιθανοτικών β.δ. με τις β.δ. που ασχολούνται με την αβεβαιότητα (αντίστοιχα για ρεύματα δεδομένων). Οι πιθανοτικές β.δ. ασχολούνται με την ύπαρξη ή μη αντικειμένων που είναι ακριβή. Σε αντίθεση με τις πιθανοτικές β.δ., η αβεβαιότητα ασχολείται με την ύπαρξη οντοτήτων, των οποίων η κατάσταση είναι μη ακριβής. Στην εργασία αυτή ασχολούμαστε με ρεύματα δεδομένων που υπόκεινται σε αβεβαιότητα.

3 Η αβεβαιότητα στην κίνηση αντικειμένων

Οι εφαρμογές παρακολούθησης αβέβαιων θέσεων κινούμενων αντικειμένων σε πραγματικό χρόνο παρουσιάζουν ιδιαίτερο πρακτικό και εμπορικό ενδιαφέρον. Η έκταση των αντικειμένων θεωρείται μία κυκλική περιοχή με ακτίνα R γύρω από ένα κέντρο (x,y). Πιο συγκεκριμένα, θεωρούμε ότι ακολουθούν την κυκλική κανονική κατανομή (σχήμα 1). Τα πιθανοτικά ερωτήματα εκφράζονται ώς μία ορθογώνια περιοχή $((x_{min},y_{min}), (x_{max},y_{max}))$ που συνοδεύεται από ένα πιθανοτικό κατώφλι (θ_i). Το γενικό μοντέλο της εφαρμογής προβλέπει τον εντοπισμό *της περιοχής της θέσης κάθε αντικειμένου* (λ.χ. μέσω GPS) και την αποστολή της περιοδικά μαζί με ένα *χρονόσημο* (*timestamp*) t σε έναν κεντρικό επεξεργαστή. Επιπροσθέτως, προβλέπει τον εντοπισμό των περιοχών που περικλείουν τα ερωτήματα καθώς και των κατωφλίων τους.

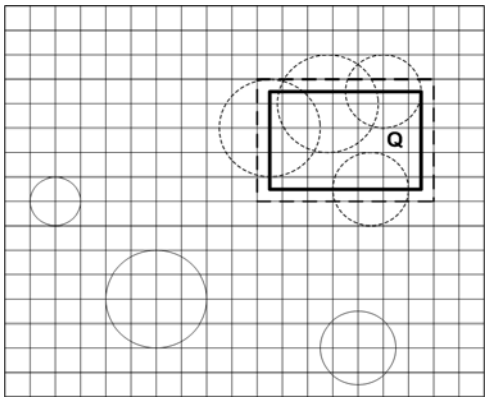
Τα ερωτήματα θέσης εξετάζουν τις περιοχές των θέσεων των αντικειμένων σε κάποια χρονική στιγμή. Πιο συγκεκριμένα για τα *πιθανοτικά ερωτήματα περιοχής (probabilistic range queries)* έχουμε: Δεδομένου ενός χωρικού παραθύρου (λ.χ. ορθογώνιας A), ενός κατωφλίου θ_i (λ.χ. 95%) και ενός χρονικού παραθύρου (χρονική στιγμή ή διάστημα), αναζητούνται τα αντικείμενα που κινούνται εντός της A κατά τη διάρκεια του διαστήματος με πιθανότητα ύπαρξης μεγαλύτερη του θ_i.

4 Επεξεργασία ερωτημάτων περιοχής

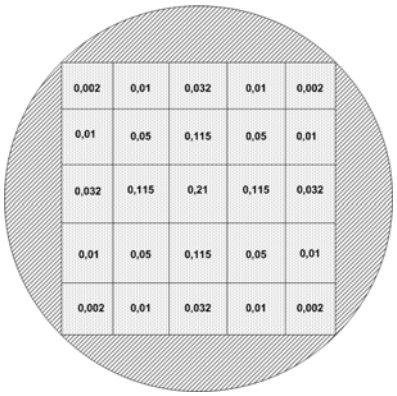
Ο αλγόριθμος που αναπτύχθηκε με σκοπό την αποτίμηση ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων απαρτίζεται από δύο στάδια, ένα στάδιο προεπεξεργασίας και ένα στάδιο επεξεργασίας. Το στάδιο προεπεξεργασίας έχει ως σκοπό τα αποτελέσματα που παράγει να χρησιμοποιηθούν στο στάδιο επεξεργασίας. Το στάδιο επεξεργασίας αποτελείται από τα εξής μέρη:

- Ευρετήριο χωρικού πλέγματος (*Grid Partitioning)*
- Τρεις τεχνικές κλαδέματος (*Pruning*)
- Λεπτομερής αποτίμηση αντικειμένων.

Και τα τρία αυτά μέρη ως σκοπό έχουν να μειώσουν το κόστος επεξεργασίας των δεδομένων. Η στρατηγική του αλγορίθμου διακρίνεται σε δύο φάσεις, μία φάση φιλτραρίσματος (*filtering phase*) και μία φάση εκλέπτυνσης (*refinement phase*). Αναλυτικότερα τα αντικείμενα που εξετάζονται για κάθε ερώτημα φιλτράρονται από τις τρεις τεχνικές κλαδέματος που βασίζονται σε γεωμετρικά και πιθανοτικά χαρακτηριστικά των αντικειμένων και των ερωτημάτων με αποτέλεσμα περιορισμένος αριθμός από αυτά να χρειάζεται να αποτιμηθεί περαιτέρω με μεγαλύτερη ακρίβεια. Η μεγαλύτερη ακρίβεια εισάγεται στην διαδικασία της εκλέπτυνσης όπου γίνεται και η αναλυτική και λεπτομερής αποτίμηση. Ο αλγόριθμος δεν παρέχει ακριβή αποτελέσματα, αλλά προσεγγίζει με μεγάλη ακρίβεια τον εξαντλητικό αλγόριθμο που παρέχει ακριβή πιθανοτικά αποτελέσματα εξοικονομώντας σημαντικά σε χρόνο εκτέλεσης. Τέλος σημειώνεται ότι ο αλγό-



Σχήμα 2



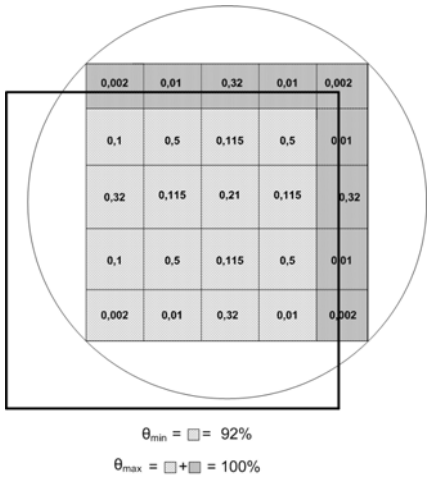
Σχήμα 3

ριθμος λειτουργεί online λαμβάνοντας περιοδικά είσοδο και εξάγοντας αποτελέσματα ανά κύκλους εκτέλεσης (κάθε χρονόσημο).

4.1 Προεπεξεργασία

Ο σκοπός του πρώτου σταδίου προεπεξεργασίας είναι να μετασχηματίσει την κυκλική κανονική κατανομή του κάθε αντικειμένου σε ένα τετράγωνο. Η διαδικασία αυτή έχει πολλά πλεονεκτήματα. Η τομή τετραγώνου με ορθογώνιο είναι υπολογίσιμη με πιο γρήγορο και εύκολο τρόπο απ' ό,τι η τομή κύκλου με ορθογώνιο. Το τετράγωνο που επιλέγεται να παραχθεί από τη θέση του κάθε κύκλου *((x,y), r)* είναι το εγγεγραμμένο τετράγωνο, το οποίο συμβολίζεται με $inbox(o_i)$. Ο μετασχηματισμός αυτός μπορεί να γίνει αφού έχει παρατηρηθεί ότι τα τέσσερα μέρη της κυκλικής κανονικής κατανομής που χάνονται (γραμμοσκιασμένα μέρη του κύκλου στο Σχήμα 3) δεν παίζουν ουσιαστικό ρόλο στους υπολογισμούς, αφού αποτελούν λιγότερο από το 5% της συνολικής κατανομής στην χειρότερη περίπτωση. Ακόμα και αν ένα ερώτημα τέμνει αυτή την περιοχή η πιθανότητα της τομής τους θα είναι πολύ μικρή, όχι σημαντική αν συλλογιστούμε ότι η μικρότερη πιθανότητα εμφάνισης άξια υπολογισμού είναι 50% ως μικρότερο πιθανοτικό κατώφλι που θέτουν τα ερωτήματα. Ακολούθως, το εγγεγραμμένο τετράγωνο υποδιαιρείται σε λxλ στοιχειώδη κουτιά, για το καθένα από τα οποία προϋπολογίζεται η πιθανότητα που περικλείει, ώστε να αποφεύγεται η ακριβής αποτίμηση πιθανοτήτων για υποπεριοχές αβεβαιότητας.

Ο σκοπός του δεύτερου σταδίου της προεπεξεργασίας είναι ο υπολογισμός του μικρότερου δυνατού εμβαδού ε_{ij} που αντιστοιχεί σε ένα πιθανοτικό κατώφλι ερωτήματος θ_i. Η διαδικασία πραγματοποιείται ώστε τα κατώφλια να χρησιμοποιηθούν για κλάδεμα στο στάδιο επεξεργασίας. Με αυτόν τον τρόπο μειώνονται οι πράξεις που εκτελούνται. Το εμβαδό ε_{ij} που επιλέγεται είναι εκείνο ενός κύκλου ομόκεντρου με την εκάστοτε κυκλική κανονική κατανομή.



Σχήμα 4

4.2 Επεξεργασία

Αρχικά, ο χώρος κατακερματίζεται ομοιόμορφα από ένα πλέγμα κελιών (*Grid Partitioning*) ανεξαρτήτως της κατανομής των αντικειμένων, με τη χρήση μιας συνάρτησης κατακερματισμού (*hash function*). Κάθε κινούμενο αντικείμενο εμπίπτει εντός των ορίων ενός ή περισσότερων κελιών με τα οποία τέμνεται. Κατά τη μετάβαση ενός αντικειμένου σε κάποιο νέο κελί, διαγράφεται από το προηγούμενο και εισάγεται στο νέο σε σταθερό χρόνο. Για κάθε κελί, διατηρείται μια λίστα κάδων στους οποίους αποθηκεύονται οι τρέχουσες θέσεις των αντικειμένων. Αυτός ο τρόπος αποτελεί ένα είδος ευρετηρίου (*index*) κατά τον οποίο τα κινούμενα ερωτήματα εξετάζουν μόνο τα αντικείμενα που περιέχονται στα κελιά με τα οποία επικαλύπτονται.

Αφού γίνει η τοποθέτηση των αντικειμένων στα κελιά ελέγχονται οι συνθήκες κλαδέματος:

- Η πρώτη συνθήκη αφορά τον συνδυασμό ερωτημάτων και αντικειμένων που δεν έχουν κινηθεί ή τουλάχιστον έχουν κινηθεί ελάχιστα, π.χ. 50μ. σε δύο συνεχόμενα χρονόσημα. Αυτά δεν χρειάζεται να επανεξεταστούν.
- Η δεύτερη συνθήκη αφορά τα αντικείμενα o_i που το $inbox(o_i)$ τους βρίσκεται εξ ολοκλήρου μέσα σε ένα ερώτημα.
- Η τρίτη συνθήκη αφορά το εμβαδόν της τομής των ερωτημάτων με τα αντικείμενα και χρησιμοποιεί τα αποτελέσματα του δεύτερου σταδίου προϋπολογισμού. Βασίζεται τόσο στα γεωμετρικά όσο και στα πιθανοτικά χαρακτηριστικά της τομής των ερωτημάτων και των αντικειμένων.

Η τελική αποτίμηση αναφέρεται σε αντικείμενα που έχουν περάσει όλους τους περιορισμούς κλαδέματος και σχηματίζει ως αποτέλεσμα ένα διάστημα εμπιστοσύνης

(θ_{min},θ_{max}). Το διάστημα εμπιστοσύνης αποτελεί ένα εύρος πιθανοτήτων με μία μικρότερη και μία μεγαλύτερη τιμή ανάμεσα στις οποίες θα βρίσκεται και η ακριβής τιμή της πιθανότητας τομής της κυκλικής κανονικής κατανομής του αντικειμένου με το αντίστοιχο ερώτημα.

'Οσο μικρότερο είναι το διάστημα εμπιστοσύνης τόσο πιο ακριβή αποτελέσματα θα δίνονταν τελικά. Παράδειγμα σχηματισμού του διαστήματος εμπιστοσύνης βλέπουμε στο σχήμα 4. Το άθροισμα των πιθανοτήτων των τετράγωνων που περιέχονται εξ ολοκλήρου μέσα στο ερώτημα αποτελούν το θ_{min}. Το άθροισμα των πιθανοτήτων των τετράγωνων που τέμνονται από το εμβαδόν του ερωτήματος αποτελούν το θ_{max}.

5 Αξιολόγηση επιδόσεων

Ο αλγόριθμος αποτίμησης πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων υλοποιήθηκε σε γλώσσα C++ και δοκιμάστηκε πειραματικά. Το πειραματικό σύνολο δεδομένων αποτελείται από 100000 κινούμενα αντικείμενα και 10000 κινούμενα ερωτήματα, τα οποία κινούνται για 200 χρονόσημα σε τροχιές που προσομοιώνουν την κυκλοφοριακή κίνηση στο οδικό δίκτυο της περιοχής της Αθήνας. Μελετήθηκαν οι επιδόσεις τους σε χρόνο και ακρίβεια, για τις εξής παραμέτρους:

- Πλήθος αντικειμένων (10k, 20k, 50k, 100k),
- Πλήθος ερωτημάτων (1k, 2k, 5k, 10k),
- Πλήθος κελιών cxc πλέγματος : 5x5, 10x10, 15x15, 20x20
- Διαστάσεις ερωτημάτων ως % της συνολικής περιοχής % : 1%, 2%, 5%
- Ακτίνα κύκλων R : 200, 600, 1000, 2000, 3000
- Κατώφλι ερωτημάτων θ_i : 60%, 70%, 80%, 95%
- Βαθμός κατάτμησης εγγεγραμμένου τετραγώνου λxλ : 5x5, 7x7, 10x10, 15x15, 20x20.

Γενικά, η κλιμάκωση των τιμών κάθε παραμέτρου επιβαρύνει τις επιδόσεις του συστήματος στις περισσότερες περιπτώσεις. Η επιλογή του βαθμού κατάτμησης c του πλέγματος επηρεάζει σημαντικά την απόδοση, ιδίως για μικρό και μεγάλο πλήθος κελιών. Πιο συγκεκριμένα για μικρό (c=5) και μεγάλο (c=20) πλήθος κελιών ο χρόνος εκτέλεσης αυξάνεται. Έτσι επιλέγεται c=10 ώστε το πλέγμα να είναι όσο το δυνατόν αποτελεσματικότερο. Επιπλέον, παρατηρήθηκαν τα εξής:

- Όσο μεγαλύτερο κατώφλι θ_i τόσο μικρότερος ο χρόνος εκτέλεσης, αφού όλο και λιγότερα αντικείμενα αποτιμώνται λεπτομερώς.
- Για μεγαλύτερη ακτίνα R τόσο αυξάνεται ο χρόνος εκτέλεσης αφού τα αντικείμενα αποθηκεύονται σε περισσότερα κελιά.
- Για ερωτήματα μεγαλύτερης έκτασης αυξάνεται ο χρόνος εκτέλεσης αφού περισσότερα αντικείμενα εξετάζονται για κάθε ερώτημα.

Τέλος έγινε σύγκριση του αλγορίθμου με άλλον που κάνει ακριβείς υπολογισμούς πιθανοτήτων τομής ερωτημάτων με αντικείμενα με τη χρήση της μεθόδου Monte Carlo και βρέθηκε ότι ο προσεγγιστικός αλγόριθμος είναι πολύ γρηγορότερος (περίπου 10 φορές) και δίνει πάντοτε διαστήματα εμπιστοσύνης που περιέχουν τις ακριβείς πιθανότητες.

6 Συμπεράσματα

Οι πρόσφατες ραγδαίες εξελίξεις στις τεχνολογίες εντοπισμού της γεωγραφικής θέσης αύξησαν το ενδιαφέρον για την ανάπτυξη εφαρμογών παρακολούθησης κινούμενων αντικειμένων. Ο σκοπός της διπλωματικής εργασίας ήταν η ανάπτυξη κατάλληλων δομών και αλγορίθμων για την αποτίμηση πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων. Από την μελέτη ζητημάτων σχεδίασης τέτοιων συστημάτων, προκύπτει ότι:

- Η επιλογή του κατακερματισμού ως μεθόδου προσπέλασης, αποδείχτηκε ιδανική για την διαρκή παρακολούθηση της θέσης των αντικειμένων.
- Οι τεχνικές κλαδέματος που υλοποιήθηκαν και βασίζονται σε γεωμετρικά και πιθανοτικά χαρακτηριστικά αποτιμούν αποδοτικά το πλήθος των κινούμενων ερωτημάτων για πλήθος κινούμενων αντικειμένων.
- Οι επιδόσεις, σε χρόνο και ακρίβεια, των αλγορίθμων μετρήθηκαν πειραματικά και ανέδειξαν την επάρκειά τους στον χειρισμό πολλαπλών κινούμενων ερωτημάτων διαρκείας.

Τέλος, ένα τέτοιο ολοκληρωμένο σύστημα διαχείρισης ρευμάτων κινούμενων πιθανοτικών ερωτημάτων περιοχής για αβέβαιες θέσεις κινούμενων αντικειμένων θα μπορούσε να επεκταθεί τόσο για επιπλέον κατανομές, τις οποίες θα επιβάλλει η μοντελοποίηση διαφορετικών προβλημάτων (π.χ. καιρικά φαινόμενα) όσο και για άλλου τύπου πιθανοτικά ερωτήματα (π.χ. πυκνότητας, εγγύτερου γείτονα, συναθροιστικά ερωτήματα εγγύτητας).