Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Αναζήτηση k-εγγύτερων γειτόνων μεταξύ περιοχών αβεβαιότητας με κανονική κατανομή

Διπλωματικη Εργασια

του

ΧΡΗΣΤΟΥ ΚΟΥΤΡΑ

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Βασεων Γνωσεων και Δεδομενων
Αθήνα, Ιούλιος 2015

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Αναζήτηση k-εγγύτερων γειτόνων μεταξύ περιοχών αβεβαιότητας με κανονική κατανομή

Διπλωματικη Εργασια

του

ΧΡΗΣΤΟΥ ΚΟΥΤΡΑ

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29η Ιουλίου 2015.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ιωάννης Βασιλείου	Νεκτάριος Κοζύρης	Ιωάννης Θεοδωρίδης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Παν. Πειραιώς

Αθήνα, Ιούλιος 2015

(Υπογραφή)

.....

Χρηστος Κουτρας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Ιωάννη Βασιλείου για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων. Επίσης ευχαριστώ ιδιαίτερα τον κ. Κώστα Πατρούμπα για την καθοδήγησή του και την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφερε όλα αυτά τα χρόνια.

Περίληψη

Αντικείμενο της διπλωματικής εργασίας είναι η ανάπτυξη και υλοποίηση ενός αλγορίθμου για την εύρεση πιθανότερων εγγύτερων γειτόνων από συγκεκριμένες σημειακές εστίες σε μία υποθετική υπηρεσία για κατόχους κινητών τηλεφώνων. Όποτε κάποιος χρήστης υποβάλλει ένα ερώτημα, θέτει τρία κριτήρια: (i) μία σημειακή εστία ενδιαφέροντος q, (ii) τον επιθυμητό αριθμό k των αναζητούμενων γειτόνων, καθώς και (iii) ένα κατώφλι πιθανότητας θ.

Για λόγους προστασίας του απορρήτου, κανένας χρήστης δεν αποκαλύπτει στους υπόλοιπους το ακριβές γωγραφικό στίγμα του, αλλά δηλώνει μία ευρύτερη *περιοχή αβεβαιότητας*. Στην προκειμένη περίπτωση, οι περιοχές αυτές μοντελοποιούνται σύμφωνα με την κανονική κατανομή. Φυσικά, η αβεβαιότητα μπορεί να εχει διαφορετικές παραμέτρους, εκφράζοντας διαφορετικές βαθμίδες ιδωτικότητας. Με τον όρο *΄πιθανότεροι εγγύτεροι γείτονες΄*, εννοούμε ότι σε μία συγεκριμένη περιοχή αναζήτησης γύρω από την εστία q, έχουν βρεθεί τουλάχιστον k κινούμενοι χρήστες με πιθανοτική κάλυψη μεγαλύτερη από το δεδομένο κατώφλι θ.

Η εργασία επικεντρώνεται κυρίως στην ανάπτυξη τεχνικών δεικτοδότησης, φιλτραρίσματος και κλαδέματος βάσει των οποίων θα μπορούμε να μειώσουμε το κόστος και τον χρόνο επεξεργασίας των δεδομένων. Ο αλγόριθμος που προτείνεται επιλέχτηκε να είναι προσεγγιστικός ως προς τον υπολογισμό της πιθανοτικής κάλυψης των περιοχών αβεβαιότητας και παρέχει μία λύση στο πρόβλημα της αποτίμησης πιθανοτικών ερωτημάτων εγγύτερων γειτόνων για αβέβαιες θέσεις κινούμενων αντικειμένων. Με εφαρμογή των παραπάνω τεχνικών, πραγματοποιήθηκαν πειράματα σε συνθετικά δεδομένα πάνω στον χάρτη της Αττικής, από τα οποία προέκυψαν θετικά αποτελέσματα. Επίσης, επιβεβαιώθηκαν οι αναμενόμενες επιδόσεις τους σχετικά με τους χρόνους εκτέλεσης και την ακρίβεια των απαντήσεων. Αυτό που μπορεί να εξαχθεί ως γενικό συμπέρασμα της εργασίας είναι ότι ο εν λόγω αλγόριθμος είναι κατάλληλος για προβλήματα πραγματικού χρόνου, θυσιάζοντας την ακρίβεια για χάρη της έγκαιρης απόκρισης.

Λέξεις Κλειδιά

Αβεβαιότητα, Πιθανοτικά ερωτήματα εγγύτερων γειτόνων, διδιάστατη κανονική κατανομή, κινούμενα αντικείμενα, ρεύματα δεδομένων.

Abstract

The purpose of this diploma thesis is to develop and implement an algorithm for most likely nearest neighbors monitoring from specific focal points in a hypothetical service for smartphone users. Whenever a user submits a most likely nearest neighbors query, sets three criteria: (i) a focal point of interest q, (ii) the desired number k of nearest neigbors, and (iii) a probability threshold θ.

Because of privacy protection reasons, no user compromises their geographical position to the rest, but declares a wider *uncertainty region*. In this case, these regions are modelled according to the bivariate Gaussian distribution. Of course, uncertainty can acquire different parameters, expressing different scales of privacy. By using the term "*most likely nearest neighbors*", we mean that in a certain search region arount point q, k moving users with probabilistic coverage above a certain threshold θ have been found.

This thesis mainly focuses on developing indexing, filtering and pruning techniques which will enable us to reduce the cost and processing time of data. The suggested algorithm is deliberately chosen to be approximate in the calculation of probabilistic coverage of uncertain regions and provides a solution to the problem of answering probabilistic nearest neighbor queries for uncertain positions of moving objects. By utilizing the above techniques, a experimental study was conducted against synthetic datasets generated using the map of Athens. In addition, the expected performance on the execution times and accuracy of answers was confirmed. The overall conclusion of this thesis is that the algorithm is suitable for real time problems, where some accuracy may be sacrificed for the benefit of timely response.

Keywords

Uncertainty, Probabilistic nearest neighbor queries, bivariate Gaussian distribution, moving objects, data streams.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η σύγχρονη τεχνολογία έχει δημιουργήσει την ανάγκη για εποπτεία και παρακολούθηση των κινήσεων διαφόρων αντικειμένων. Η ανάγκη αυτή έχει πολλαπλασιαστεί ιδιαίτερα από την ανάπτυξη που παρουσιάζει η βιομηχανία των "*έξυπνων*" κινητών τηλεφώνων (*smartphones*).

Πιο συγκεκριμένα, η πλειοψηφία των χρηστών κινητών τηλεφώνων χρησιμοποιεί καθημερινά εφαρμογές κοινωνικής δικτύωσης, όπως είναι το *Facebook*, το *Twitter* και το *Foursquare*. Η φύση των εφαρμογών αυτών κάνει υποχρεωτική την υποστήριξη υπηρεσιών εντοπισμού (*Location-based services*) των κοντινότερων στον χρήστη τοποθεσιών, φίλων και γενικότερα κινούμενων αντικειμένων. Για παράδειγμα, έστω ότι κάποιος χρήστης χρειάζεται κάποια συγκεκριμένη βοήθεια και είναι αναγκαίο αυτό να γίνει σε σύντομο χρονικό διάστημα. Είναι φανερό πως θα ήταν ιδιαίτερα χρήσιμο γι′ αυτόν, εάν μπορούσε μέσω μιας εφαρμογής κοινωνικής δικτύωσης να βρει γρήγορα και έγκυρα ποιοι φίλοι του βρίσκονται εγγύτερα στην τωρινή θέση του.

Η ευρεία χρήση των εφαρμογών κοινωνικής δικτύωσης καθώς και οι διάφορες υπηρεσίες εντοπισμού έχουν φέρει, ασφαλώς, στο προσκήνιο την ανάγκη για προστασία της ιδιωτικότητας (*privacy*) των χρηστών τους. Αυτή γίνεται εντονότερη, από τη στιγμή που όλο και περισσότερο έρχονται στο φως της δημοσιότητας πληροφορίες που καταδεικνύουν την εκμετάλλευση των προσωπικών δεδομένων στο Διαδίκτυο, από τέτοιου είδους υπηρεσίες για λόγους εμπορικούς και μή.

Εξαιτίας του ειδικού βάρους που έχει αποκτήσει η προστασία των προσωπικών δεδομένων είναι εύκολα κατανοητό ότι η υλοποίηση υπηρεσιών εντοπισμού κοντινότερων αντικειμένων γίνεται σαφώς πιο δύσκολη και απαιτητική. Τα δεδομένα τέτοιων εφαρμογών γίνονται πιο πολύπλοκα και η ανάλυσή τους για εξαγωγή απαντήσεων κοστίζει περισσότερο τόσο σε χρόνο όσο και σε επεξεργασία. Εκτός αυτού, πρέπει να συνυπολογίζεται ο τεράστιος όγκος των δεδομένων αυτών και η διαχείριση των αιτημάτων των χιλιάδων ή εκατομμυρίων συνδρομητών, οι οποίοι βέβαια επιθυμούν την εγκυρότητα και την ταχεία απόκριση.

Στην κατεύθυνση αυτή κρίνεται, λοιπόν, απαραίτητη η εξεύρεση λύσεων, οι οποίες θα επιτρέπουν σε εφαρμογές, όπως αυτές που αναφέραμε παραπάνω, να παρέχουν τη δυνατότητα στους χρήστες να ενημερώνονται για τους *εγγύτερους γείτονές* (*nearest neighbors*) τους, όποτε αυτοί το επιθυμούν, χωρίς καθυστέρηση και με σχετική ακρίβεια.

1.1 Αντικείμενο της διπλωματικής

Σκοπός της εργασίας είναι η μελέτη και η υλοποίηση ενός αλγορίθμου που θα επιτρέπει online απαντήσεις σε *πιθανοτικά ερωτήματα διαρκείας* (*probabilistic continuous queries*) σχετικά με την εύρεση *εγγύτερων γειτόνων* ανάμεσα σε κινούμενα αντικείμενα με αβέβαιες θέσεις. Τα εν λόγω πιθανοτικά ερωτήματα διαρκείας θα τίθενται από διάφορους κινούμενους χρήστες, για συγκεκριμένες σημειακές θέσεις ενδιαφέροντος (π.χ. ένα εστιατόριο ή ένας κινηματογράφος), οι οποίοι επιθυμούν να ενημερώνονται για τα αντικείμενα που είναι πιθανότερο να βρίσκονται κοντά σε αυτές. Οι θέσεις αυτές δεν δηλώνουν απαραίτητα την τωρινή θέση του εκάστοτε χρήστη και είναι γνωστές στο σύστημα, δηλαδή είναι δυνατός ο γεωγραφικός τους εντοπισμός (π.χ. μέσω GPS). Εξαιτίας της *αβεβαιότητας* (*uncertainty*) των κινούμενων αντικειμένων ως προς την ακριβή γεωγραφική τους θέση, ο επεξεργαστής του συστήματος οφείλει να δίνει εγκαίρως προσσεγιστικές απαντήσεις στους χρήστες, οι οποίες όμως θα πρέπει να είναι αρκετά αξιόπιστες. Επίσης, είναι επιθυμητό η ενημέρωση των απαντήσεων να γίνεται μόνο όταν υπάρχει κάποια αλλαγή σε σχέση με προηγούμενα αποτελέσματα. Τέτοια δεδομένα θα μπορούσαν κυρίως να αξιοποιηθούν σε εφαρμογές κοινωνικής δικτύωσης με κινητά τηλέφωνα (π.χ. *Facebook*), στις οποίες υπάρχει ανάγκη αξιοποίησης της γεωγραφικής θέσης των χρηστών. Είναι προφανές, πως βασική υπόθεση της εργασίας είναι η παρακολούθηση πολλών κινούμενων αντικειμένων και ερωτημάτων τα οποία θα ανανεώνουν συχνά την περιοχή της θέσης τους και δημιουργούν *ρεύματα δεδομένων* (*data streams*). Εδώ τονίζεται ότι:

- Τα στοιχεία καταφθάνουν στον server σε μεγάλο και ενδεχομένως μεταβλητό ρυθμό σε πραγματικό χρόνο.
- Τα δεδομένα αφορούν κινούμενους χρήστες και κινούμενα ερωτήματα.
- Κάθε κινούμενος χρήστης αποστέλλει τη θέση του ως μία περιοχή αβεβαιότητας, η οποία ανά χρήστη μπορεί να έχει μεταβλητό μέγεθος ανάλογα με την επιθυμία του για μεγαλύτερη προστασία της ιδωτικότητάς του ή όχι.
- Τα ρεύματα έχουν θεωρητικά απεριόριστο μέγεθος.

Η καινοτομία που προσφέρει η παρούσα εργασία , σε σχέση με προηγούμενες επιλύσεις, είναι πως γίνεται προσέγγιση του προβλήματος των k-εγγύτερων γειτόνων με τις εξής παραμέτρους:

- Γίνεται διαχείρηση μεγάλου πλήθους αντικειμένων και ερωτημάτων, των οποίων οι πληροφορίες ενημερώνονται συνεχώς, δηλαδή δεν είναι στατικά δεδομένα αλλά ρεύματα δεδομένων.
- Η περιοχή αβεβαιότητας κάθε αντικειμένου μοντελοποιείται με μία συνεχή κατανομή και όχι με διακριτά δείγματα.

Είναι λογικό πως με βάση τη φύση του μοντέλου, η καταλληλότερη επιλογή είναι η ανάπτυξη προσεγγιστικών αλγορίθμων που εξοικονομούν χρόνο και κόστος επεξεργασίας, συνεκτιμώντας πάντοτε το σφάλμα που μπορεί να γίνει αποδεκτό από την εκάστοτε εφαρμογή.

1.2 Οργάνωση του τόμου

Η παρούσα διπλωματική εργασία είναι οργανωμένη σε εννέα κεφάλαια. Στα κεφάλαια 2-5 παρέχονται το θεωρητικό υπόβαθρο και η βιβλιογραφική επισκόπηση βασικών εννοιών και σχετικών μεθόδων. Στα κεφάλαια 6-9 μελετάται το πρόβλημα των πιθανών k-εγγύτερων γειτόνων και περιγράφεται η προτεινόμενη μέθοδος επίλυσής του, η οποία αξιολογείται πειραματικά. Ειδικότερα:

Στα κεφάλαια 2, 3 περιγράφεται η έννοια των ρευμάτων κινούμενων αντικειμένων, όπου δίνεται έμφαση στην αδυναμία των συστημάτων διαχείρισης βάσεων δεδομένων όσον αφορά την αποδοτική επεξεργασία δυναμικά μεταβαλλόμενων δεδομένων. Παρουσιάζονται τα χαρακτηριστικά των ρευμάτων δεδομένων καθώς και οι επεκτάσεις που έγιναν στις συμβατικές βάσεις δεδομένων. Επίσης, γίνεται αναφορά στη διαχείριση των χωροχρονικών βάσεων δεδομένων.

Στο κεφάλαιο 4 περιγράφονται τα προβλήματα εγγύτερων γειτόνων όταν οι θέσεις των αντικειμένων είναι ακριβείς και παρουσιάζονται επιλεγμένες ερευνητικές προσεγγίσεις.

Στο κεφάλαιο 5 γίνεται ανάλυση των πιθανοτικών βάσεων δεδομένων, όπως και των πιθανοτικών χωρικών ερωτημάτων, τονίζοντας τη διαφοροποίηση σε σχέση με την περίπτωση όπου τα δεδομένα είναι ακριβή.

Στο κεφάλαιο 6 διατυπώνεται το πρόβλημα που στοχεύει να λύσει η εργασία, αναλύοντας τα μοντέλα των κινούμενων αντικειμένων και των ερωτημάτων και την μορφή των απαντήσεων.

Στο κεφάλαιο 7 παρουσιάζεται ο πρωτότυπος αλγόριθμος που επινοήθηκε για την αποτίμηση πιθανοτικών ερωτημάτων k-εγγύτερων γειτόνων για αβέβαιες θέσεις κινούμενων αντικειμένων. Ο αλγόριθμος δίνει προσεγγιστικές απαντήσεις, δέχεται ως εισόδους δύο ρεύματα δεδομένων, ένα για τα κινούμενα αντικείμενα και ένα για τα κινούμενα ερωτήματα και βασίζεται σε τεχνικές δεικτοδότησης, φιλτραρίσματος και κλαδέματος που σκοπεύουν στη μείωση του χρόνου εκτέλεσης και του κόστους επεξεργασίας.

Στο κεφάλαιο 8 αξιολογείται πειραματικά ο αλγόριθμος πάνω σε συνθετικά δεδομένα και σχολιάζονται διεξοδικά τα αποτελέσματα.

Τέλος, στο κεφάλαιο 9 εκτίθενται τα γενικά συμπεράσματα καθώς και πιθανές μελλοντικές προοπτικές της μεθόδου.

Κεφάλαιο 2

Ρεύματα δεδομένων

2.1 Εισαγωγή

Τα τελευταία χρόνια προέκυψε η ανάγκη για έρευνα πάνω σε θέματα που προέρχονται από ένα νέο μοντέλο στην επεξεργασία δεδομένων. Σύμφωνα με αυτό, τα δεδομένα δεν είναι πλέον στατικά όπως στις παραδοσιακές βάσεις δεδομένων, αλλά αλλάζουν δυναμικά με συνεχή , γρήγορο και χρονικά μεταβαλλόμενο τρόπο με τη μορφή ρευμάτων (*data streams*). Η νέα αυτή κατηγορία δεδομένων έχει γίνει ευρέως αναγνωρισμένη και υποστηρίζει πολλές εφαρμογές, δίκτυα αισθητήρων, κ.ά. Στις εφαρμογές αυτές, τα δεδομένα μεταβάλλονται με την πάροδο του χρόνου (π.χ. σε ένα δίκτυο αισθητήρων οι τιμές αλλάζουν όταν αλλάζει κάποια θερμοκρασία). Στο μοντέλο των ρευμάτων δεδομένων, αυτά μπορεί να παρουσιάζονται ως σχεσιακές πλειάδες (π.χ. μετρήσεις σε ένα δίκτυο, αρχεία κλήσεων, επισκεψιμότητα ιστοσελίδα κτλ). Παρόλα αυτά, η συνεχής άφιξή τους σε πολλαπλά, χρονικά μεταβαλλόμενα κι ενδεχομένων απρόβλεπτα και απεριόριστα ρεύματα γεννούν νέα προβλήματα προς επίλυση για την ερευνητική κοινότητα, διαφορετικά από αυτά των βάσεων δεδομένων.

Για όλες τις προαναφερθείσες εφαρμογές, δεν είναι δυνατό να χρησιμοποιήσουμε ένα παραδοσιακό σύστημα διαχείρισης βάσεων δεδομένων (ΔΒΜΣ) και να δουλέψουμε με αποτελεσματικότητα πάνω σε αυτό. Τα παραδοσιακά ΣΔΒΔ δεν είναι σχεδιασμένα για την ταχεία και συνεχή φόρτωση των συγκεκριμένων δεδομένων, με αποτέλεσμα να μην υποστηρίζουν άμεσα τα ερωτήματα διαρκείας (*continuous queries*). Επίσης, τα παραδοσιακά ΣΔΒΔ επικεντρώνονται σε μεγάλο βαθμό σε ακριβείς απαντήσεις οι οποίες υπολογίζονται από σταθερά πλάνα ερωτημάτων. Αντίθετα, χαρακτηριστικό των εφαρμογών αυτών αποτελούν οι προσεγγιστικές και οι προσαρμοστικές απαντήσεις στην εκτέλεση ερωτημάτων πάνω σε ρεύματα δεδομένων.

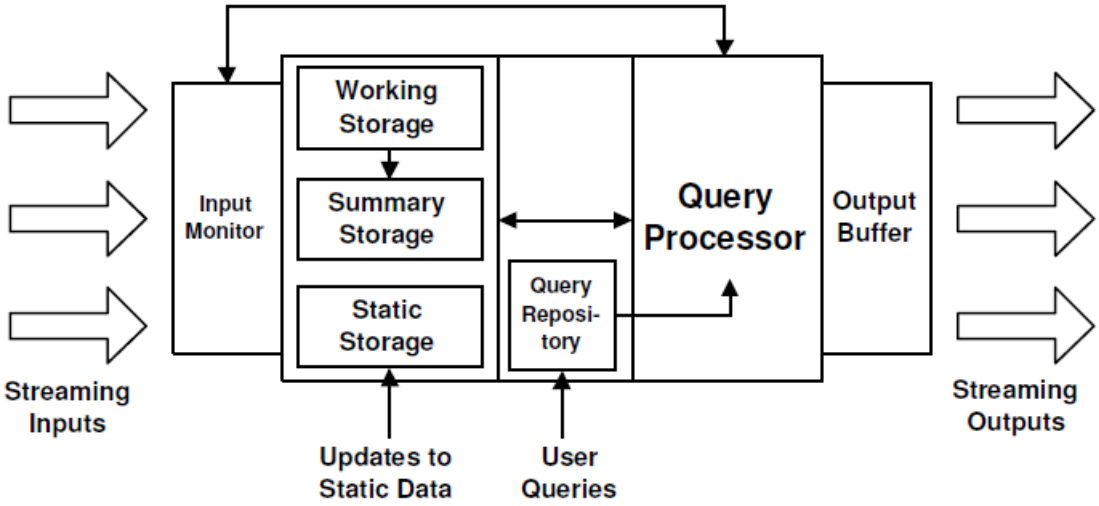
Στο κεφάλαιο αυτό, παρουσιάζεται το μοντέλο των ρευμάτων δεδομένων. Αρχικά αναπτύσσονται οι λόγοι ύπαρξης συστημάτων ρευμάτων δεδομένων και στη συνέχεια γίνεται μία επισκόπηση πάνων στα χαρακτηριστικά τους (μοντέλο, τελεστές), που έχουν παρουσιαστεί σε διάφορες ερευνητικές εργασίες. Στη συνέχεια, γίνεται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

2.2 Μοντέλο ρευμάτων δεδομένων

Στα συστήματα διαχείρισης βάσεων δεδομένων υπάρχει ενδιαφέρον 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, εξαιτίας κυρίως του τρόπου 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, γι΄ αυτό και παρουσιάστηκε η ανάγκη για ανάπτυξη ευέλικτων συστημάτων που να διαχειρίζονται με αποτελεσματικότητα τέτοια δεδομένα, το οποίο 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 η περισσότερα συνεχή ρεύματα. Αυτά αποτελούνται απο 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 [9]. 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 είνα πλέον ο χρήστης εκείνος που δρομολογεί τη ροή των δεδομένων, αλλά οι ίδιες οι πηγές, αποτελεί την ειδοποιό διαφορά με το πρότυπο των συμβατικών ΣΔΒΔ. Έτσι, ο χρήστης περιορίζεται στο ρόλο του παρατηρητή, έχοντας τη δυνατότητα μόνο να ρυθμίζει τις παραμέτρους του συστήματος και να υποβάλει ερωτήματα. Τα ρεύματα δεδομένων παρουσιάζουν τα εξής χαρακτηριστικά [9, 19]:

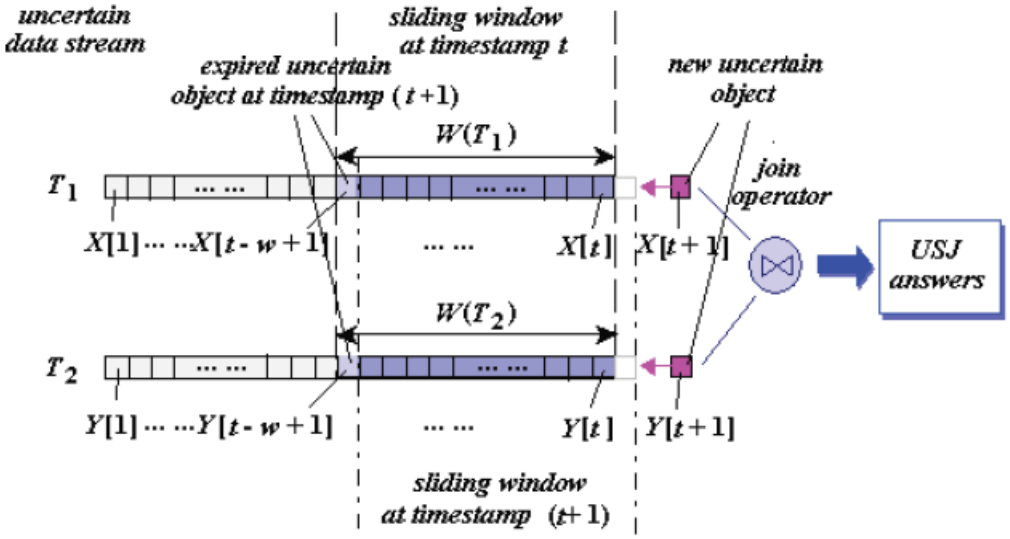
- Τα δεδομένα καταφθάνουν έγκαιρα, σε πραγματικό χρόνο (online) και τότε μπορούν να επεξεργαστούν.
- Τα ρεύματα δεδομένν μπορεί να είναι απεριόριστα σε μέγεθος.
- Το σύστημα οφείλει να δίνει έγκυρες και συνεπείς απαντήσεις σε πραγματικό χρόνο



Σχήμα 2.1: Σύστημα διαχείρισης ρευμάτων δεδομένων

ακόμη και αν το μέγεθος των ρευμάτων είναι απεριόριστο.

- Τα δεδομένα συνήθως δεν αποθηκεύονται και αντικαθίστανται από τα καινούρια.
- 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 κανένα έλεγχο πάνω στην εγκυρότητα καθώς επίσης και στη σειρά με την οποία τα δεδομένα φθάνουν ώστε να υποστούν επεξεργασία. Για τον λόγο αυτό πρέπει να εξασφαλίζεται η κανονική λειτουργία του συστήματος σε περίπτωση που αντιμετωπίζονται τέτοιες ατέλειες.
- 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 αποθηκεύονται ρητά στη μνήμη, κάτι το οποίο μπορεί να γίνει μόνο εφόσον έχουμε μικρό όγκο ρευμάτων δεδομένων.
- Η μη αποθήκευση του ρεύματος δεδομένων οδηγεί σε προσεγγιστικές (αλλά αξιόπιστες) απαντήσεις στην αποτίμηση ερωτημάτων που εμπλέκουν αποθήκευση μεγάλου όγκου δεδομένων (π.χ. δομές σύνοψης).



Σχήμα 2.2: Σύνδεση ρευμάτων δεδομένων

- 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 πως τα ρεύματα που αποτελούν είσοδο στο σύστημα μπορεί να παράγονται από την επεξεργασία άλλων ρευμάτων δεδομένων. Συγκεριμένα 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 το 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 στον παγκόσμιο ιστό).

- *Ασφάλεια δικτύων* : Καταγραφή επισκεπτών για εκτίμηση συμφόρησης σε διάφορους κόμβους του δικτύου.

Οι τελεστές που χρησιμοποιούνται στα συστήματα ρευμάτων δεδομένων [10] αποτελούν προεκτάσεις των παραδοσιακών ΣΔΒΔ ώστε να μπορούν να ανταπεξέλθουν στις απαιτήσεις τους. Σε αυτές έχει προστεθεί η έννοια του χρόνου (π.χ. σε ερωτήματα σε δεδομένα τα

τελευταία 10 λεπτά ή τις τελευταίες 10 χρονικές στιγμές). Οι κυριότεροι από αυτούς είναι οι εξής:

- *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 2.2). Για παράδειγμα, ποια λεωφορεία και τραμ από όλα τα κινούμενα αντικείμενα στην πόλη έχουν ταχύτητα μικρότερη από 10 χλμ.
- *Συνάθροιση* (*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 τη 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.

2.3 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 [9]:

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

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 δεοδμένα και πληροφορίες που καταγράφηκαν στο παρελθόν επιβαρύνουν ακόμα περισσότερο το εκάστοτε σύστημα. Για παράδειγμα, ο υπολογισμός του μέσου όρου των 1000 τελευταίων στοιχείων ενός ρεύματος δεδομένων θα απαιτούσε την καταγραφή των τελευταίων 1000 στοιχείνων του στη μνήμη και την αναθεώηρση της 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, τα ερωτήματα διαρκείας μπορούν να χωριστούν στις εξής δύο κατηγορίες:

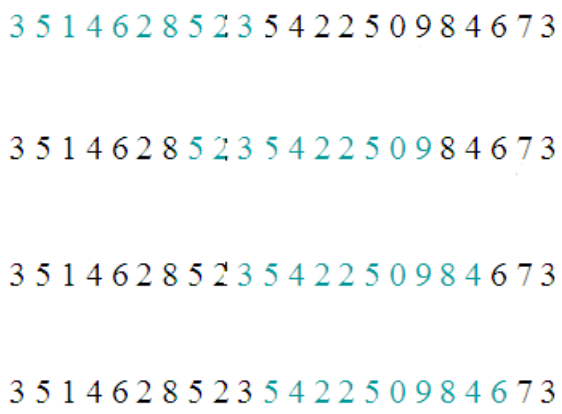
- Προκαθορισμένα ερωτήματα (predefined queries)
- Μη προκαθορισμένα ερωτήματα (ad-hoc queries)

Η διάκριση αυτή γίνεται για τον εξής λόγο: τα προκαθορισμένα ερωτήματα είναι εκ των προτέρων γνωστά στο σύστημα πριν καταφθάσουν τα δεδομένα, τα οποία θα είναι διαθέσιμα για επεξεργασία και το σύστημα μπορεί να κατανέμει καταλλήλως τους πόρους του (π.χ. τη μνήμη του για την επεξεργασία). Τα μη προκαθορισμένα ερωτήματα αναφέρονται σε δεδομένα που πιθανόν να μην είναι διαθέσιμα (π.χ. δεδομένα από το παρελθόν που δεν υπάρχουν πλέον), ενώ μπορεί να μην διαθέσιμοι και οι απαραίτητοι πόροι για την επεξεργασία τους.

2.4 Παράθυρα σε ερωτήματα διαρκείας

Ένα σύστημα διαχείρισης βάσεων δεδομένων εξετάζει ένα συγκεκριμένο τμήμα του συνόλου των πλειάδων που καταφθάνουν ως είσοδος. Όπως αναφέρθηκε και παραπάνω είναι ασύμφορο να αποθηκεύει όλα τα δεδομένα του ρεύματος εισόδου. Το εξεταζόμενο τμήμα του ρεύματος εισόδου αποτελεί ένα *παράθυρο* (*window*) επί των πιο πρόσφατων στοιχείων του ρεύματος. Συγκεριμένα, το παράθυρο περιλαμβάνει ένα τμήμα διαδοχικών πλειάδων στις οποίες τίθεται το ερώτημα. Για παράδειγμα, σε ένα δίκτυο αισθητήρων οι τιμές των θερμοκρασιών που καταγράφονται αφορούν τις τελευταίες μέρες ή το πολύ εβδομάδες, με τα παλαιότερα δεδομένα να διαγράφονται. Έτσι, τα παράθυρα απομονώνουν ένα πεπερασμένο πλήθος στοιχείων από ένα μεγάλο, πιθανώς απείρου μήκους ρεύμα δεδομένων. Τρία είδη των παραθύρων συναντώνται συνήθως στις περισσότερες εφαρμογές. Αυτά είναι (σύμφωνα με το [9]):

- *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.



Σχήμα 2.3: Παράθυρο ρεύματος δεδομένων.

- *Κυλιόμενα παράθυρα βάσει χρόνου* (*Time based sliding windows*) : Έχουν αφετηρία και πέρας που κινούνται ταυτόχρονα παρακολουθώντας την χρονική εξέλιξη των στοιχείων που συρρέουν στο σύστημα. Έτσι, παλαιότερα δεδομένα απορρίπτονται και καινούρια εισέρχονται με κυμαινόμενο ρυθμό. Το εύρος των παραθύρων παραμένει σταθερό, όμως ούτε το πλήθος των πλειάδων ούτε και τα περιεχόμενά τους διατηρούνται αμετάβλητα.

- *Κυλιόμενα παράθυρα βάσει πλειάδων* (*Tuple based sliding windows*) : Έχουν αφετηρία και πέρας που κινούνται ταυτόχρονα παρακολουθώντας τις πλειάδες που συρρέουν στο σύστημα (σχήμα 2.3).

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

2.5 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. Επιπρόσθετα, ο αλγόριθμος πρέπει να είναι σε θέση να παρέχει και να αποθηκεύει σημαντικά ενδιάμεσα αποτελέσματα, για να μπορεί να αυξήσει τις επιδόσεις του. Ακόμη, πρέπει να παρέχεται η δυνατότητα πρόβλεψης μελλοντικών πληροφοριών μέσω της σωστής διαχείρισης δεδομένων. Υπάρχουν αλγόριθμοι που επιτρέπουν πολλαπλά περάσματα από το ρεύμα δεδομένων, ωστόσο οι πλέον αποδοτικοί είναι εκείνοι που επεξεργάζονται τα στοιχεία μόνο μία 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 έχουν παρέλθει μέχρι εκείνη τη στιγμή όσο και τα τελικά αποτελέσματα.

2.5.1 Μαζική επεξεργασία

Τα δεδομένα επεξεργάζονται μαζικά (batch processing) και όχι μεμονωμένα το καθένα όταν καταφθάνουν. Με αυτόν τον τρόπο η εκτέλεση των ερωτημάτων γίνεται ταχύτερη, 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 στον παρόν. Η τεχνική αυτή δίνει έγκυρες και σε πραγματικό χρόνο απαντήσεις αφού η καθυστέρηση της απάντησης δεν βλάπτει την αξιοπιστία 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.

2.5.2 Δειγματοληψία

Με τη δειγματοληψία (*sampling*) ρευμάτων δεδομένων επεξεργάζεται μόνο ένας περιορισμένος αριθμός δειγμάτων και όχι το σύνολο του ρεύματος. Η δειγματοληψία μπορει να είναι [9]:

- *τυχαία* (*randomized sampling*) : 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*) : Κάθε ένα συγκεκριμένο αριθμό δειγμάτων αποθηκεύεται μία πλειάδα. Οι υπόλοιπες απορρίπτονται.

- *διαστρωμάτωμένη* (*stratified sampling*) : Μειώνονται 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.

2.5.3 Συνόψεις δεδομένων

Οι συνόψεις δεδομένων (summaries or data synopses) αποτελούν μία 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 καταφθάνουν. Ο κεντρικός επεξεργαστής έχει τη δυνατότητα να λαμβάνει υπόψη του τις υπάρχουσες συνόψεις και να παράγει προσεγγιστικά αποτελέσματα γαι τα ερωτήματα που τίθενται. Η χρήση περιλήψεων συμβάλλει 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* (*Sketches*)
- *Κυματίδια* (*Wavelets*)
- *Ιστογράμματα* (*Histograms*)

2.6 Γλώσσες ρευμάτων δεδομένων

Μία γλώσσα ρευμάτων δεδομένων μπορεί να έχει πολλά κοινά στοιχεία με γλώσσες που χρησιμποιούνται 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. Επίσης, γλώσσες όπως η SQL χρησιμοποιούνται ευρέως και πιθανές επεκτάσεις 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. Η υλοποίησή του βασίστηκε στη γλώσσα βάσεων δεδομένων SQL, με την επέκτασή της σε μία καινούρια γλώσσα ρευμάτων δεδομένων με το όνομα CQL (Continuous 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.

- AURORA : Αναπτύχθηκε από τα πανεπιστήμια MIT, Brown και Brandeis το 2001. Ο πρωταρχικός στόχος της AURORA 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 κλιμακωσημότητας σε κατανεμημένα συστήματα επεξεργασίας αποτέλεσε έναν ακόμα σημανιτκό λόγο για την ανάπτυξή του. Το σύστημα υποβολής ερωτημάτων τα σχεδιάζει σε γραφικό περιβάλλον δημιουργώντας έτσι ένα διάγραμμα ροής αποτελούμενο από τετράγωνα και βέλη. Μεταξύ των χαρακτηριστικών του είναι η αρχιτεκτονική του συντονίζεται από έναν χρονοπρογραμματιστή (scheduler) και το γεγονός ότι υποστηρίζει και ένα σύστημα αποθήκευσης δεδομένων.

- TelegraphCQ : Αναπτύχθηκε από το πανεπιστήμιο Berkeley. Η υλοποίηση 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, ο πρωταρχικός στόχος του συστήματος αυτού ήταν να χρησιμοποιηθεί για δικτυακούς σκοπούς, όπως δίκτυα επικοινωνιών και αισθητήρων.

Άλλα συστήματα ρευμάτων δεδομένων είναι τα AQuery, Tribeca, τα οποία κυρίως χρησιμοποιούνται για ανάλυση κυκλοφορίας και σε δίκτυα αισθητήρων.

Κεφάλαιο 3

Διαχείριση κινούμενων αντικειμένων

3.1 Χωρικά και χωροχρονικά δεδομένα

Με την ραγδαία εξέλιξη των συστημάτων γεωγραφικού εντοπισμού, δημιουργήθηκε η ανάγκη ανάπτυξης ποικίλων σχετικών εμπορικών και ερευνητικών εφαρμογών. Μία τέτοια εφαρμογή που πλέον χρησιμοποιείται ευρέως είναι και η πλοήγηση στο οδικό δίκτυο μιας πόλης σε πραγματικό χρόνο. Σε αυτήν ο εκάστοτε χρήστης υποβάλει τη θέση προορισμού του και το σύστημα του παρέχει την συντομότερη διαδρομή. Ακόμα, το σύστημα μπορεί να εντοπίζει τη θέση του και να τον καθοδηγεί ανάλογα, ακόμα και να επαναπροσαρμόζεται σε περίπτωση λάθους. Η ανάπτυξη τέτοιων εφαρμογών αποτέλεσαν την αφορμή για την εμφάνιση ενός νέου αντικειμένου έρευνας στο πεδίο των βάσεων δεδομένων, τις χωροχρονικές βάσεις δεδομένων.

Η ανάπτυξη κατάλληλων δομών, οι οποίες αναπριστούν τα χωροχρονικά δεδομένα καθώς και κατάλληλων πράξεων επί αυτών είναι απαραίτητα στοιχεία για την υλοποίηση των εφαρμογών αυτών. Έχουν προταθεί διάφοροι τρόποι αναπαράστασης των πραγματικών χωρικών αντικειμένων στον υπολογιστή [6, 8]:

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

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

- Η *περιοχή* (*region*), που αναπαριστά μία γεωγραφική περιοχή στον χάρτη. Είναι πιθανό να περιέχει οπές ή να αποτελείται από πολλά μη επικαλυπτόμενα μέρη. Για τις ανάγκες των χωροχρονικών βάσεων δεδομένων επεκτείνεται στις κινούμενες περιοχές.

Τα ερωτήματα που αφορούν κινούμενα αντικείμενα πρέπει να επεξεργάζονται αποδοτικά

και παράλληλα να παρακολουθούν την κίνηση και τη μεταβολή των χωρικών αντικειμένων. Το πλήθος των δεδομένων στις περισσότερες περιπτώσεις χωρικών εφαρμογών είναι αρκετά μεγάλο. Η αποτίμηση των ερωτημάτων οφείλει να γίνεται με τρόπο αποδοτικό και να εξασφαλίζει την αξιοπιστία και τις καλές χρονικές επιδόσεις. Οπότε, γίνεται κατανοητό πως είναι αναγκαία η οργάνωση των δεδομένων σε κατάλληλες δομές για την εύκολη προσπέλασή τους από τα ερωτήματα, με συγκεκριμένες μεθόδους. 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.

3.1.1 Μοντελοποίηση χωρικών δεδομένων

Όπως έχει ήδη αναφερθεί, τα χωρικά αντικείμενα που χειρίζονται οι χωρικές βάσεις δεδομένων είναι το σημείο, η γραμμή και η περιοχή με την επέκτασή 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 και αποτελεί επέκταση του αφηρημένου μοντέλου για στατικά χωρικά δεδομένα. Οι τύποι δεδομένων και οι λειτουργίες τους δημιουργούν μία άλγεβρα κινούμενων αντικειμένων, όμως αυτό 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 βέβαια προσεγγίζει την πραγματική.

3.1.2 Χωρικά ερωτήματα

Οι κυριότερες κατηγορίες χωρικών ερωτημάτων είναι οι εξής [9]:

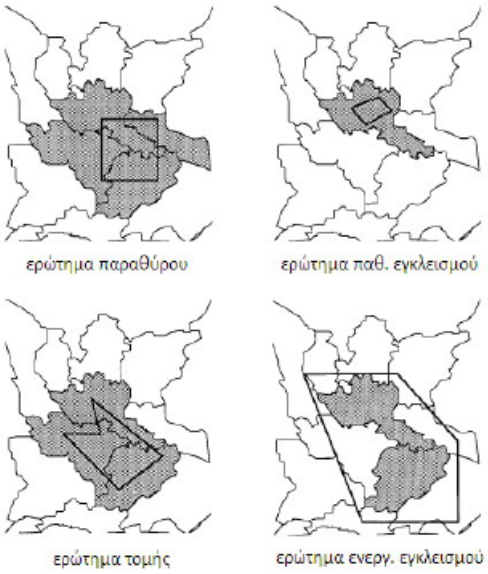
- *Ισότητας* (*Exact match query*) : Ζητείται να βρεθούν όλα τα αντικείμενα που έχουν την ίδια γεωμετρία με ένα συγκεκριμένο αντικείμενο.

- *Σημείου* (*Point query*) : Ζητείται να βρεθούν όλα τα αντικείμενα που περιέχουν ένα σημείο.

- *Παραθύρου* (*Window query*) : Ζητείται να βρεθούν όλα τα αντικείμενα που έχουν ένα τουλάχιστον κοινό σημείο με ένα παράθυρο.

- *k-εγγύτερων γειτόνων* (*k-nearest neighbors*) : Ζητείται να βρεθούν τα k κοντινότερα αντικείμενα σε ένα συγκεκριμένο σημείο ή αντικείμενο.

- *Χωρικής σύνδεσης* (*Spatial query*) : Πραγματοποιείται ο συνδυασμός δύο ή περισσοτέρων δεδομένων βάσει ενός κοινού τους χαρακτηριστικού.



Σχήμα 3.1: Παραδείγματα χωρικών ερωτημάτων [9]

- *Τομής* (*Intersection query*) : Ζητείται να βρεθούν όλα τα αντικείμενα με τα οποία ένα συγκεκριμένο αντικείμενο έχει κοινά εσωτερικά σημεία.
- *Ενεργητικού εγκλεισμού* (*Containment query*) : Ζητείται να βρεθούν όλα τα αντικείμενα που περιέχονται από ένα συγκεκριμένο αντικείμενο.
- *Παθητικού εγκλεισμού* (*Enclosure query*) : Ζητείται να βρεθούν όλα τα αντικείμενα που περιέχουν ένα συγκεκριμένο αντικείμενο.
- *Γειτνίασης* (*Adjacency query*) : Ζητείται να βρεθούν αντικείμενα που συνορεύουν με το δοθέν.

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.

3.2 Κινούμενα αντικείμενα

Ο τομέας της παρακολούθησης κινούμενων αντικειμένων παρουσιάζει ιδιαίτερη ανάπτυξη στις μέρες μας. Με την χρήση συστημάτων όπως το GPS και ασυρμάτων τηλεπικοινωνιών εντοπίζεται η θέση κάθε αντικειμένου και αποστέλλεται περιοδικά σε έναν κεντρικό επεξεργαστή (server). Εκεί γίνεται η επεξεργασία των εκάστοτε ερωτημάτων που τίθενται. Τα κινούμενα αντικείμενα σε αυτές τις εφαρμογές θεωρούνται σημειακά αφού η περιοχή που καλύπτουν είναι αμελητέος συγκριτικά με την έκταση του χώρου. Οι τύποι ερωτημάτων ποικίλουν ανάλογα με

την εφαρμογή. Τα ερωτήματα που θα μας αποσχολήσουν κυρίως στην παρούσα εργασία είναι ερωτήματα που έχουν να κάνουν με θέσεις των αντικειμένων στο παρόν. Πρόκειται για ερωτήματα k-εγγύτερων γειτόνων που 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 ρευμάτν δεδομένων στον κεντρικό υπολογιστή και ακολουθούν τα χαρακτηριστικά τους. Επίσης, τα ερωτήματα που μελετώντανι, 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.

3.2.1 Θέσεις κινούμενων αντικειμένων

Η κίνηση αποτελεί χαρακτηριστικό για πολλά αντικείμενα. Στην πραγματικότητα η κίνηση των αντικειμένων διέπεται από περιορισμούς. Συγκεκριμένα, υπάρχουν οι εξής τρεις κατηγορίες κίνησης:

- Κίνηση χωρίς περιορισμούς, όπως για παράδειγμα η κίνηση των αεροπλάνων.
- Κίνηση με περιορισμούς, όπως για παράδειγμα η κίνηση πεζών όταν συναντούν φυσικά εμπόδια.
- Κίνηση σε ορισμένες τροχιές, όπως για παράδειγμα η κίνηση τρένων στο σιδηροδρομικό δίκτυο.

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 συμβαίνει είτε εσκεμμένα είτε να οφείλεται στον τρόπο λήψης και καταγραφής των θέσεών τους. Για λόγους προστασίας ιδιωτικότητας, κάποιος χρήστης μπορεί να μην επιθυμεί να φανερώσει το ακριβές στίγμα του. Επίσης, 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. Ακόμα, από τις συσκευές γεωγραφικού εντοπισμού αλλά και από αιτίες όπως π.χ. θόρυβος, παρεμβολές και διακοπές κατά τη διάρκεια της μετάδοσης, είναι αναπόφευκτο ότι στη διαδικασία καταγραφής των συντεταγμένων δημιουργούνται σφάλματα. Η πραγματική θέση ενός αντικειμένου δεν είναι δυνατόν να προσδιοριστεί με ακρίβεια, αλλά κυμαίνεται εντός κάποιων ορίων σφάλματος, τα οποία από συσκευή σε συσκευή μπορεί να διαφέρουν.

Η κίηση μπορεί να αναπαρασταθεί στο διδιάστατο επίπεδο από δυό χωρικές συντεταγμένες (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* : Το αναγνωριστικό του κινούμενου αντικειμένου.

- *timestamp* : Το χρονόσημο στο οποίο γίνεται η καταγραφή.
- *x coordinate* : Συντεταγμένη ως προς τον x-άξονα.
- *y coordinate* : Συντεταγμένη ως προς τον y-άξονα.

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 διαφέρει από αντικείμενο σε αντικείμενο. Η σταθερή και μεγάλη συχνότητα ενδεχομένων να παρουσιάζει αρκετά προβλήματα και να επιβαρύνει το σύστημα με πλεονάζουσα πληροφορία. 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 τροχιές, απαιτείται μικρότερη συχνότητα δειγματοληψίας.

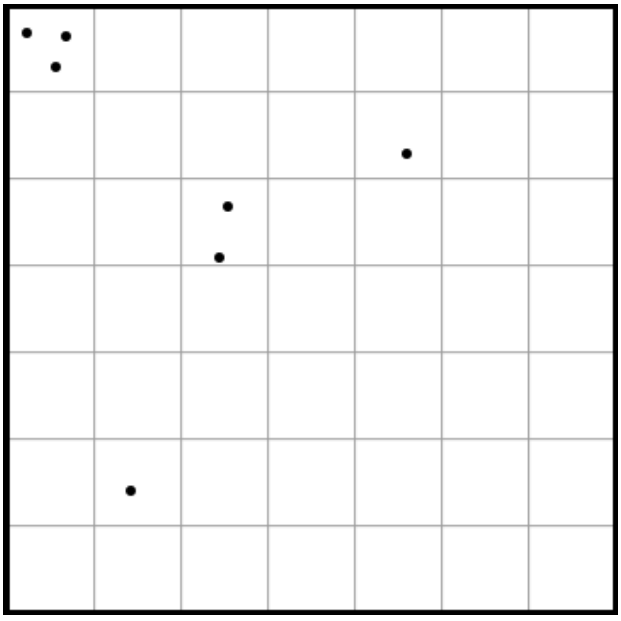
3.2.2 Ερωτήματα σε κινούμενα αντικείμενα

Τα χωρικά ερωτήματα που ανεφέρθηκαν προηγουμένως δεν είναι όλα εφαρμόσιμα σε κινούμενα σημειακά αντικείμενα. Αυτό συμβαίνει κυρίως για τη θεώρησή τους ως αντικείμενα με αμελητέα έκταση. Τα ερωτήματα σε κινούμενα αντικείμενα χωρίζονται στις εξής δύο κατηγορίες:

- *Ερωτήματα θέσης* (*location-based queries*)
- *Ερωτήματα τροχιάς* (*trajectory-based queries*)

Τα ερωτήματα θέσης ασχολούνται με τη θέση των αντικειμένων σε κάποια χρονική στιγμή. Τα ερωτήματα είναι δυνατόν να αναφέρονται στο παρόν, 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 είναι τα εξής [6, 8]:

- *Ερωτήματα περιοχής* (*range queries*) : Για ένα χωρικό παράθυρο και ένα χρονικό, ζητούνται τα αντικείμενα των οποίων η κίνηση περιέχεται εντός τους. Οι απαντήσεις επιστρέφονται με τη μορφή των ταυτοτήτων των αντικειμένων.
- *Ερωτήματα k-εγγύτερων γειτόνων* (*k-nearest neighbor queries*) : Για ένα κινούμενο σημειακό αντικείμενο και ένα χρονικό παράθυρο, ζητούνται τα k αντικείμενα 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.
- *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.



Σχήμα 3.2: Κάνναβος 7 × 7 ως χωρικό ευρετήριο

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

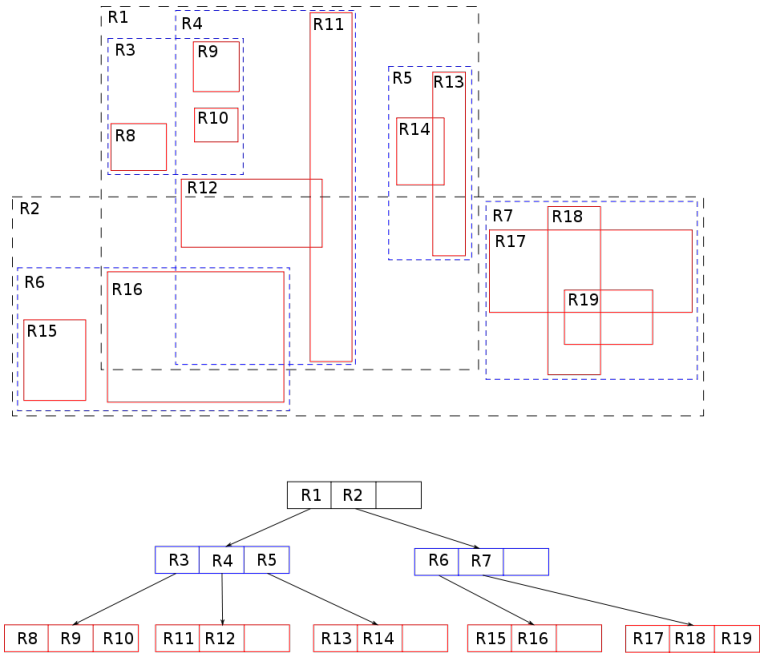
Τα ερωτήματα τροχιάς διακρίνονται στις εξής κατηγορίες:

- *Τοπολογικά ερωτήματα* (*topological queries*) : 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 κλασσικά 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.

- *Ερωτήματα πλοήγησης* (*navigational queries*) : 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 είανι 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 ή μέγιστη τιμή τους.

3.2.3 Δεικτοδότηση κινούμενων αντικειμένων

Η συγχρονισμένη παρακολούθηση μεγάλου όγκου κινούμενων αντικειμένων δημιουργεί πολλά προβλήματα ως προς την προσπέλαση των πληροφοριών. Το γεγονός αυτό καθιστά



Σχήμα 3.3: Παράδειγμα χρήσης R-tree ως χωρικό ευρετήριο

αναγκαία 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 ανάλογα με τα χαρακτηριστικά της εφαρμογής, με γνώμονα τα δεδομένα και με γνώμονα το χώρο. 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 καινούριο. Σε κάθε κελί μπορεί να τοποθετηθεί απεριόριστο πλήθος αντικειμένων. Στο σχήμα 3.2 απεικονίζεται ένας διδιάστατος τετραγωνικός χώρος, ο οποίος κατακερματίζεται από τον κάνναβο, τα κελιά του οποίου έχουν τετραγωνικό σχήμα.

Άλλες τεχνικές δεικτοδότησης κινούμενων αντικειμένων που βασίζονται κυρίως στο R-tree [7] είναι το STR-tree (Spatio-temporal R-tree), το TB-tree (Trajectory-Bundle Tree), το TPR-tree (Time-Parameterized R-tree) και το STAR-tree. Η βασική ιδέα πίσω από το R-tree (σχήμα 3.3) 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 ανεπάρκειά του στην δεικτοδότηση κινούμενων αντικειμένων.

Κεφάλαιο 4

Ερωτήματα k-εγγύτερων γειτόνων σε ακριβή στίγματα αντικειμένων

4.1 Εισαγωγή

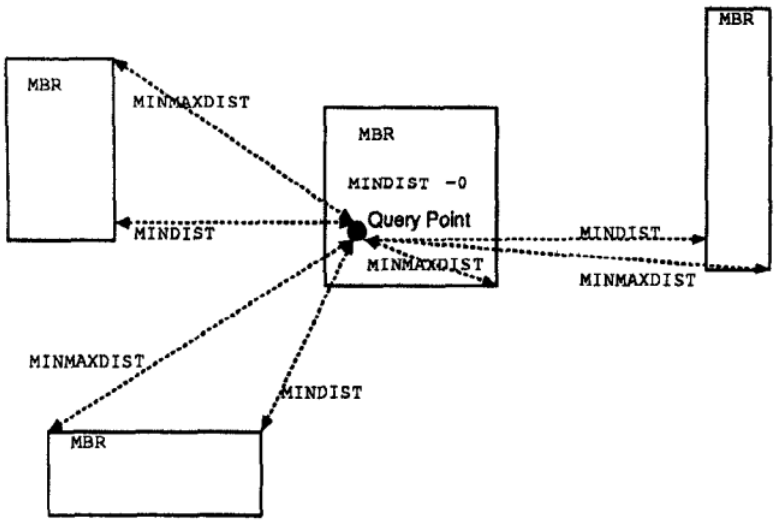
Ένας από τους σημαντικότερους τύπους ερωτημάτων στα συστήματα γεωγραφικού εντοπισμού είναι αυτός της εύρεσης των k-εγγύτερων γειτόνων (kNN) σε ένα δοσμένο σημείο στον χώρο. Η προσέγγιση του προβλήματος και ο τρόπος επίλυσής του μπορεί να διαφέρει ανάλογα με το εάν πρόκειται για στατικά δεδομένα, δηλαδή είναι γνωστά στο σύστημα όλα τα αντικείμενα με τις θέσεις τους και το σημείο του ερωτήματος και δεν υπάρχει ενημέρωση αυτών των πληροφοριών, ή για ρεύματα δεδομένων, όπου οι θέσεις των ερωτημάτων και των αντικειμένων αλλάζουν με το χρόνο και είναι αναγκαίο με κάθε τέτοια αλλαγή να γίνεται εκ νέου αποτίμηση.

Στο κεφάλαιο αυτό θα αναλύσουμε τις επικρατέστερες λύσεις του προβλήματος και για τις δύο παραπάνω κατηγορίες, στην περίπτωση όπου οι θέσεις των αντικειμένων και των ερωτημάτων είναι ακριβή στίγματα στο χώρο.

4.2 Αποτίμηση ερωτημάτων kNN σε στατικά δεδομένα

Οι πρώτες προσεγγίσεις του προβλήματος των k-εγγύτερων γειτόνων, αφορούσαν δεδομένα στατικά, των οποίων το μέγεθος δεν ήταν πολύ μεγάλο. Για τον λόγο αυτό, η χρησιμοποίηση δεντρικών δομών ως χωρικά ευρετήρια κρίθηκε η καταλληλότερη για την επίλυση του. Στην ενότητα αυτή θα γίνει περιγραφή της μεθόδου που αναπτύχθηκε στην εργασία [18], η οποία αντιμετωπίζει με βέλτιστο τρόπο το συγκεκριμένο πρόβλημα.

Ο αλγόριθμος χρησιμοποιεί ως χωρικό ευρετήριο ένα R-tree, μία δεντρική δομή στην οποία κάθε κόμβος αντιπροσωπεύει ένα ή περισσότερα Minimum Bounding Rectangles (MBRs). Ουσιαστικά, ένα MBR αντιστοιχεί στην ελάχιστη ορθογώνια περιοχή που περιλαμβάνει μία συστάδα από αντικείμενα στον χώρο. Οπότε, όσο κινούμαστε προς τα φύλλα του R-tree



Σχήμα 4.1: Παραδείγματα MINDIST και MINMAXDIST αποστάσεων στο διδιάστατο χώρο [18]

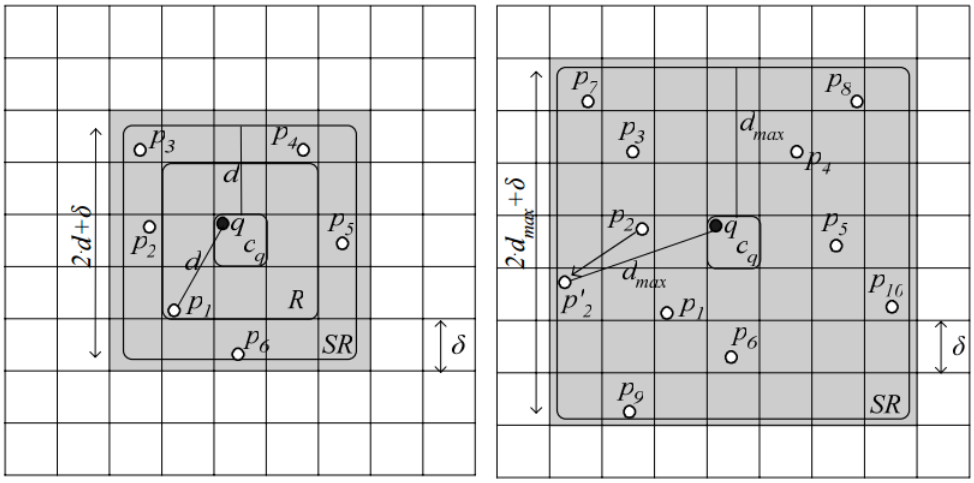
τόσο τα MBR είναι ολοένα και μικρότερα. Κάθε κόμβος γονιός έχει ως παιδιά κόμβους που περιέχουν τα ίδια αντικείμενα με αυτόν αλλά τοποθετημένα σε μικρότερα MBR. Επίσης, παρουσιάζονται δύο συγκεκριμένες μετρικές (σχήμα 4.1):

- Ελάχιστη απόσταση σημείου από ορθογώνια περιοχή (MINDIST), η οποία αποδεικνύεται πως αποτελεί ένα κάτω όριο για την απόσταση ενός σημείου από ένα αντικείμενο που εγκλείεται σε ένα MBR.

- Ελάχιστη μέγιστη απόσταση σημείου από ορθογώνια περιοχή (MINMAXDIST, η οποία αποδεικνύεται πως αποτελεί ένα άνω όριο για την απόσταση ενός σημείου από ένα αντικείμενο που εγκλείεται σε ένα MBR.

Με βάση αυτές τις μετρικές, ο αλγόριθμος κάνει χρήση κανόνων κλαδέματος (pruning) και η μέθοδος εύρεσης του εγγύτερου γείτονα είναι η εξής: Αρχικά, η απόσταση του εγγύτερου γείτονα τίθεται στο άπειρο. Ξεκινώντας από τη ρίζα του R-tree, το διασχίζουμε *κατά βάθος* και για κάθε κόμβο που δεν είναι φύλλο υπολογίζουμε τις MINDIST των MBRs του από το σημείο του ερωτήματος και με βάση αυτές ταξινομούμε τα εξεταζόμενα MBRs σε μία λίστα. Στη συνέχεια εφαρμόζουμε τις τεχνικές κλαδέματος στη λίστα αυτή και ο αλγόριθμος επαναλαμβάνεται για κάθε κόμβο που περιλαμβάνει το αντίστοιχο MBR έως ότου η λίστα να είναι κενή. Όταν φτάνουμε σε ένα φύλλο-MBR τότε υπολογίζουμε τις αντίστοιχες αποστάσεις των αντικειμένων του από τη θέση του ερωτήματος και ανανεώνουμε την εκτίμηση της ελάχιστης απόστασης του εγγύτερου γείτονα. Στο τέλος της διαδικασίας θα έχουμε ως αποτέλεσμα τον εγγύτερο γείτονα μαζί με την απόστασή του.

Η παραπάνω μέθοδος, εφαρμόζεται και για την εύρεση των k-εγγύτερων γειτόνων κάνοντας τις εξής παραλλαγές:



(a) NN search (b) Update handling

Σχήμα 4.2: Παραδείγματα εκτέλεσης αλγορίθμου [23]

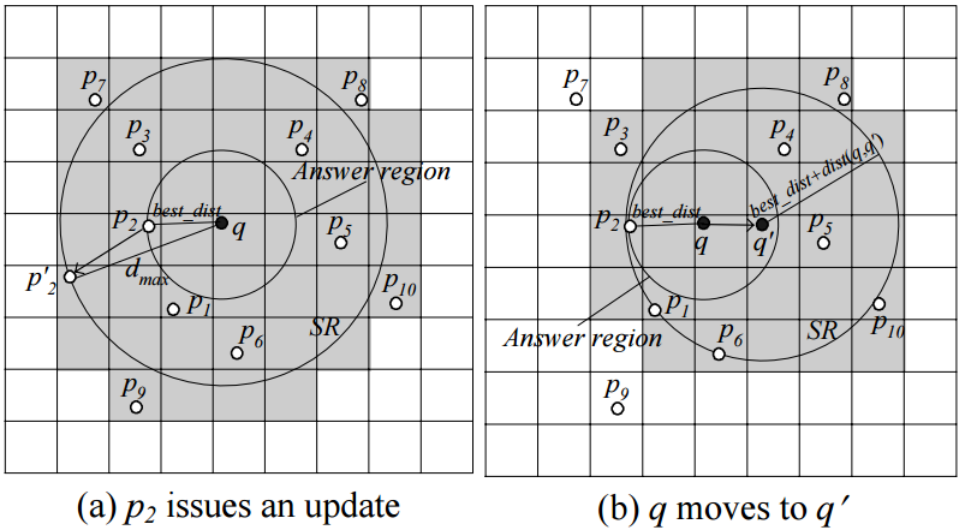
- Χρησιμοποιείται μία ταξινομημένη λίστα των μέχρι στιγμής k-εγγύτερων γειτόνων που έχουν βρεθεί.
- Οι τεχνικές κλαδέματος εφαρμόζονται με βάση την μεγαλύτερη του ερωτήματος από τον k-οστό γείτονα της λίστας αυτής.

Η πειραματική μελέτη έδειξε πως ο προτεινόμενος αλγόριθμος λειτουργεί σωστά και κλιμακώνεται καλά όσο ο αριθμός των εγγύτερων γειτόνων μεγαλώνει και το μέγεθος των δεδομένων αυξάνεται.

4.3 Αποτίμηση ερωτημάτων kNN σε ρεύματα δεδομένων

Η ανάπτυξη της τεχνολογίας, καθώς και η μεγάλη αύξηση των δεδομένων που κινούνται στο Διαδίκτυο, έκανε αναγκαία την προσέγγιση του προβλήματος των k-εγγύτερων γειτόνων για την περίπτωση όπου στην είσοδο έχουμε ρεύματα δεδομένων. Ασφαλώς, η χρήση δεντρικών δομών ως ευρετήρια για την επίλυση αυτού του προβλήματος απορρίφθηκε, αφού δεν είναι δυνατό να αναπτυχθεί αλγόριθμος που να μπορεί να κάνει αποτίμηση σε πραγματικό χρόνο. Επίσης, το μέγεθος των δεδομένων είναι αρκετά μεγάλο, γεγονός που καθιστά απαγορευτική την αποθήκευσή τους σε μία δεντρική δομή. Για τον λόγο αυτό η ερευνητική κοινότητα επέλεξε ως βασική μορφή ευρετηρίου, τη διαμέριση του χώρου σε πλέγματα κελιών (κάνναβος). Η δομή αυτή αποδείχθηκε κατάλληλη για την παρακολούθηση μεγάλου αριθμού αντικειμένων και την αποτίμηση συνεχών ερωτημάτων kNN. Παρακάτω, θα περιγραφούν οι τρεις επικρατέστερες μεθόδοι που αναπτύχθηκαν για την επίλυση του προβλήματος:

Στο [23] τα αντικείμενα δεικτοδοτούνται σε έναν κάνναβο με κελιά συγκεκριμένου μεγέθους. Κάθε ερώτημα kNN επαναϋπολογίζεται κάθε μία συγκεκριμένη περίοδο T και οι αλλαγές γίνονται άμεσα στον κάνναβο με κάθε ενημέρωση των δεδομένων. Όταν ένα ερώτημα αποτιμάται για πρώτη φορά, τότε ακολουθείται η εξής διαδικασία δύο βημάτων: Αρχικά,

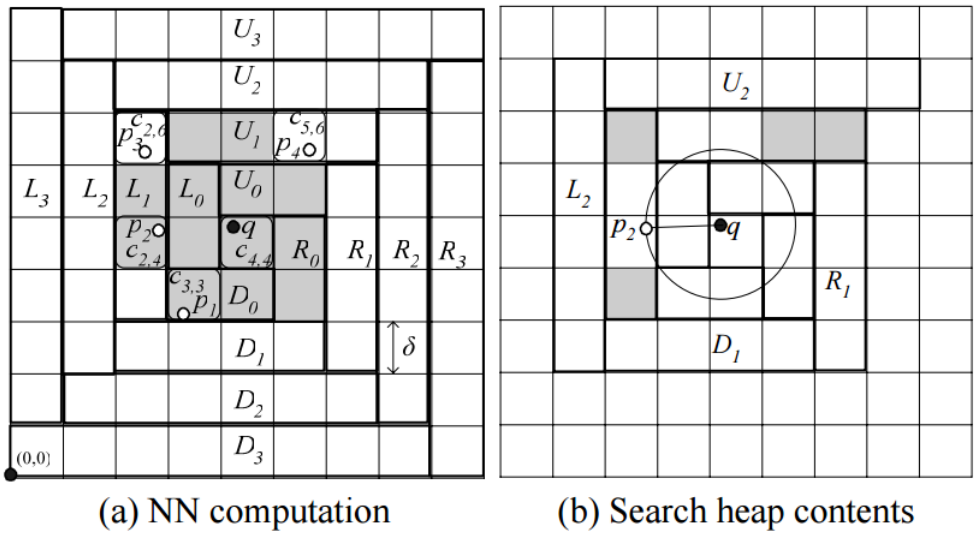


Σχήμα 4.3: Παραδείγματα χειρισμού ενημερώσεων [22]

εξετάζονται κελιά σε ένα τετράγωνο R γύρω από αυτό του ερωτήματος q έως ότου βρεθούν k. Στο σχήμα 4.2a παρατηρούμε ένα παράδειγμα ενός ερωτήματος εγγύτερου γείτονα όπου ο πρώτος υποψήφιος είναι το αντικείμενο *p1*, όμως δεν είναι αναγκαία και ο κοντινότερος, αφού μπορεί να υπάρχουν άλλα αντικείμενα έξω από το R που είναι σε μικρότερη απόσταση από το q. Για την ανάκτηση τέτοιων αντικειμένων επεικτεινόμαστε σε ένα μεγαλύτερο τετράγωνο γύρω από το κελί του ερωτήματος, στο οποίο αναζητούνται οι πραγματικοί k εγγύτεροι γείτονες, όπως φαίνεται και στο παράδειγμα του σχήματος 4.2a, όπου εξετάζονται τα αντικείμενα *p1-p6* και επιστρέφεται το *p2* ως απάντηση.

Ο επαναϋπολογισμός ενός ερωτήματος κάνει χρήση του προηγούμενου ευρεθέντος αποτελέσματος έτσι ώστε να μικρύνει τον χώρο αναζήτησης. Συγκεκριμένα, αποτιμά την μεγαλύτερη απόσταση από τα προηγούμενα αντικείμενα του συνόλου των kNN έχοντας ως δεδομένο τις τρέχουσες θέσεις τους και με βάση αυτή ερευνά τα αντικείμενα σε ένα νέο τετράγωνο γύρω από τη θέση του ερωτήματος q, όπως φαίνεται στο σχήμα 4.2b που αναφέρεται στο προηγούμενο παράδειγμα. Όταν ένα ερώτημα αλλάζει θέση τότε γίνεται αποτίμησή του από την αρχή. Ακόμη, ως επέκταση του αλγορίθμου γίνεται χρήση ενός ιεραρχικού καννάβου, που βελτιώνει την απόδοσή τους σε περίπτωση δεδομένων που είναι πολύ αραιά.

Σε αντίθεση με την παραπάνω μέθοδο, ο αλγόριθμος που προτείνεται στο [22] ασχολείται αποκλειστικά με την αποτίμηση των kNN ερωτημάτων όταν γίνεται ενημέρωση των δεδομένων, χωρίς να παρέχει κάποια τεχνική για τον υπολογισμό των εγγύτερων γειτόνων όταν αυτός γίνεται για πρώτη φορά. Και σε αυτή την προσέγγιση, τα αντικείμενα δεικτοδοτούνται σε έναν κάνναβο. Η περιοχή απάντησης (*answer region*) ορίζεται ως τα κελιά που τέμνονται από τον κύκλο με κέντρο το q και ακτίνα ίση με την απόσταση από τον μέχρι στιγμής k-οστό εγγύτερο γείτονα. Όταν έρχεται στο σύστημα μία ενημέρωση, ο αλγόριθμος προσδιορίζει την περιοχή αναζήτησης με βάση τα κελιά που επηρεάζουν οι αλλαγές και εάν αυτά έχουν επικάλυψη με την περιοχή απάντησης. Η ακτίνα αναζήτησης υπολογίζεται διαφορετικά για τις τρεις παρακάτω



Σχήμα 4.4: Παραδείγματα υπολογισμού εγγύτερου γείτονα [15]

περιπτώσεις:

- Εάν κάποιος από τους τρέχοντες εγγύτερους γειτόνους κινηθεί εντός της περιοχής απάντησης ή κάποια αντικείμενα εκτός αυτής εισέλθουν τότε οι εγγύτεροι γείτονες αναζητούνται σε αυτή την περιοχή.
- Εάν κάπιος από τους τρέχοντες εγγύτερους γειτόνους εξέλθει τις περιοχής απάντησης, τότε ο υπολογισμός των νέων kNN γίνεται παρόμοια με το [23].
- Εάν η θέση του ερωτήματος αλλάξει τότε η ακτίνα αναζήτησης τίθεται ίση με την ακτίνα της περιοχής απάντησης αυξημένης κατά τη μετατόπιση του q.

Στο σχήμα 4.3 φαίνονται παραδείγματα χειρισμού ενημερώσεων για την δεύτερη και τρίτη περίπτωση αντίστοιχα, όπου τα σκιαγραφημένα κελιά είναι αυτά στα οποία αναζητούνται κάθε φορά οι εγγύτεροι γείτονες.

Η αποδοτικότερη μέθοδος επίλυσης του προβλήματος των συνεχών ερωτημάτων kNN παρουσιάζεται στο [15], όπου η διαμέριση του χώρου γίνεται σε ορθογώνιες περιοχές γύρω από το κελί του ερωτήματος q έτσι ώστε να αποφευχθούν περιττοί υπολογισμοί που γίνονταν σε προηγούμενες ερευνητικές προσεγγίσεις. Ο αλγόριθμος επισκέπτεται κελιά γύρω από αυτό του ερωτήματος σε αύξουσα σειρά ελάχιστης απόστασης (MINDIST) και λειτουργεί ως εξής: Κάθε φορά ελέγχει εάν η οντότητα προς εξέταση είναι (i) κελί, οπότε και ελέγχει τα αντικείμενα που βρίσκονται σε αυτό και ανανεώνει τη λίστα με τους μέχρι στιγμής εγγύτερους γείτονες και την τιμή της απόστασης από τον k-οστό εγγύτερο γείτονα ($best\_dist$), ή είναι (ii) ορθογώνια περιοχή, οπότε τοποθετεί τα κελιά της σε ένα heap με τις προς εξέταση οντότητες (μαζί με τις ελάχιστες αποστάσεις τους από το ερώτημα), στο οποίο επίσης τοποθετεί την ορθογώνια περιοχή που βρίσκεται στο επόμενο επίπεδο προς την ίδια κατεύθυνση (επίσης μαζί με την ελάχιστη απόστασή της από το ερώτημα). Η διαδικασία τερματίζει όταν η επόμενη προς εξέταση οντότητα στο heap έχει ελάχιστη απόσταση από το ερώτημα q μεγαλύτερη από την

τρέχουσα *best_dist* και άρα αποκλείεται να υπάρξουν αλλαγές στη λίστα με τους εγγύτερους γείτονες. Στο σχήμα 4.4a φαίνεται ένα παράδειγμα υπολογισμού εγγύτερου γείτονα, όπου διακρίνεται η διαμέριση του χώρου σε ορθογώνιες περιοχές και με σκιά τα κελιά που εξετάζει ο αλγόριθμος. Στο σχήμα 4.4b παρουσιάζονται τα περιεχόμενα του heap κατά την εκτέλεση του αλγορίθμου για το συγκεκριμένο παράδειγμα.

Ο χειρισμός των ενημερώσεων για κάθε αντικείμενο που κινήθηκε λειτουργεί ως εξής: Αρχικά, ελέγχονται τα ερωτήματα τα οποία είχαν στην λίστα με τους εγγύτερους τους γείτονες το αντικείμενο που άλλαξε θέση. Για κάθε ένα από αυτά, εάν η νέα απόσταση του αντικειμένου είναι μικρότερη από την *best_dist* του ερωτήματος τότε η λίστα με τους εγγύτερους γείτονες δεν επηρεάζεται, παρά μόνο ίσως η κατάταξή τους, η οποία επαναπροσδιορίζεται. Εάν δεν ισχύει η υπόθεση αυτή, τότε το αποτέλεσμα του ερωτήματος πρέπει να επαναϋπολογιστεί. Στη συνέχεια, ελέγχονται, με βάση τη νέα θέση του αντικειμένου, τα ερωτήματα που πιθανώς να επηρεάζονται από την αλλαγή αυτή. Για κάθε ένα από αυτά εάν ισχύει πως η απόσταση του αντικειμένου είναι μικρότερη από την *best_dist*, τότε διαγράφεται το k-οστό αντικείμενο από τη λίστα με τους εγγύτερους γείτονες, στην οποία εισέρχεται το εξεταζόμενο αντικείμενο και προσδιορίζεται η καινούρια κατάταξη.

Στην εργασία αναφέρεται και μία βελτίωση σχετικά με τον χειρισμό πολλαπλών ενημερώσεων που καταφθάνουν στο σύστημα και αποδεικνύεται πειραματικά πως η μέθοδος που προτείνεται είναι γρηγορότερη και αποδοτικότερη συγκριτικά με τις αντίστοιχες των [23], [22].

Κεφάλαιο 5

Διαχείριση δεδομένων με αβεβαιότητα

5.1 Εισαγωγή

Τα τελευταία χρόνια εμφανίστηκε ένα ευρή φάσμα εφαρμογών που σχετίζονται με την *αβεβαιότητα* (*uncertainty*). Οι λόγοι για τους οποίους υπάρχει αβεβαιότητα στα δεδομένα ποικίλουν ανάλογα με την εφαρμογή. Για παράδειγμα, στην λειτουργία αισθητήρων η αβεβαιότητα οφείλεται στην ανακρίβεια λόγων σφαλμάτων κατά τη μέτρηση (π.χ. υγρασία). Σε άλλες εφαρμογές, όπως είναι η προστασία της ιδιωτικότητας, υπάρχει η απαίτηση τα δεδομένα να είναι επίτηδες λιγότερα ακριβή. Η αβεβαιότητα συμβάλλει στην απόκρυψη προσωπικών δεδομένων προστατεύοντας ευαίσθητα χαρακτηριστικά των ατόμων, έτσι ώστε μικρότερο μέρος στοιχείων να μπορεί να δημοσιευθεί. Ένα άλλο παράδειγμα έγκειται στα συστήματα εντοπισμού στίγματος (GPS). Το στίγμα ενός αντικειμένου 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 γνωρίζουμε προσεγγιστικά πού βρίσκεται. Έτσι, η θέση του αντικειμένου όπως είναι γνωστή στο σύστημα, δεν ταυτίζεται πάντοτε με την τρέχουσα λόγω χρονικής καθυστέρησης κατά τη μετάδοση, οπότε θεωρείται αβέβαιη. Μέχρι πρόσφατα, οι παραδοσιακές βάσεις δεδομένων δεν ήταν προετοιμασμένες για να αντιμετωπίσουν δεδομένα με αβεβαιότητα.

Σήμερα έχουν αναπτυχθεί πολλές μέθοδοι και αλγόριθμοι για καλύτερη επεξεργασία ερωτημάτων, πολλά από τα οποία τα συναντάμε γενικά στις βάσεις δεδομένων 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 δεδομένων. Σε άλλες εργασίες ακολουθείται ο συνεχής τρόπος αναπαράστασης, υποθέτοντας διάφορες κατανομές (ομοιόμορφη, κανονική κτλ.) και σε άλλες ο διακριτός τρόπος με χρήση πεπερασμένου 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 η δυνατότητα να διακρίνουν κατά πόσο

Researchers:

	Name	Affiliation	P	
t_1^1	Fred	U. Washington	$p_1^1 = 0.3$	$X_1 = 1$
t_1^2		U. Wisconsin	$p_1^2 = 0.2$	$X_1 = 2$
t_1^3		Y! Research	$p_1^3 = 0.5$	$X_1 = 3$
t_2^1	Sue	U. Washington	$p_2^1 = 1.0$	$X_2 = 1$
t_3^1	John	U. Wisconsin	$p_3^1 = 0.7$	$X_3 = 1$
t_3^2		U. Washington	$p_3^2 = 0.3$	$X_3 = 2$
t_4^1	Frank	Y! Research	$p_4^1 = 0.9$	$X_4 = 1$
t_4^2		M. Research	$p_4^2 = 0.1$	$X_4 = 2$

Σχήμα 5.1: Παράδειγμα πιθανοτικής βάσης δεδομένων [5]

οι απαντήσεις είναι αξιόπιστες.

5.2 Πιθανοτικά και αβέβαια δεδομένα

Στην ενότητα αυτή γίνεται αναφορά στις διαφορές που υπάρχουν ανάμεσα στις πιθανοτικές βάσεις δεδομένων και τις βάσεις δεδομένων με αβεβαιότητα, έννοιες που συχνά συγχέονται. 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.

5.2.1 Πιθανοτικές βάσεις δεδομένων

Οι πιθανοτικές βάσεις δεδομένων ασχολούνται με την ύπαρξη και την μη-ύπαρξη αντικειμένων που είναι ακριβή. Για παράδειγμα, η πιθανοτική βάση τους σχήματος 5.1 αντιπροσωπεύει την πιθανότητα καθενός από τα άτομα να εργάζεται σε κάποιο ερευνητικό κέντρο ή πανεπιστήμιο. Το άθροισμα των πιθανοτήτων κάθε ατόμου είναι ίσο με 1, αφού θεωρούμε ότι το άτομο υπάρχει και γνωρίζουμε ότι κατέχει κάποια ερευνητική θέση.

Είναι γνωστό πως 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*, [5]). Ένας πιθανός κόσμος είναι μία υποθετική αναπαράσταση του πραγματικού κόσμου και μπορεί να αναπαρασταθεί από μία βάση δεδομένων πλήρους πληροφορίας. Με άλλα λόγια, είναι η απεικόνιση όλων των στιγμιοτύπων που μπορούν να προκύψουν από μία πιθανοτική βάση δεδομένων.

5.2.2 Βάσεις δεδομένων με αβεβαιότητα

Σε αντίθεση με τις πιθανοτικές βάσεις δεδομένων, η αβεβαιότητα στις βάσεις δεδομένων 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. Ενώ, λοιπόν, σε μία πιθανοτική βάση δεδομένων θα γνωρίζαμε ότι ένα αντικείμενο βρίσκεται σε μία ακριβή θέση με μία συγκεκριμένη πιθανότητα, η οποία θα αναφερόταν στην ύπαρξη ή μη του αντικειμένου στη θέση αυτή, στις βάσεις δεδομένων με αβεβαιότητα η ύπαρξη του αντικειμένου σε μία περιοχή θα ήταν δεδομένη και το ενδιαφέρον θα επικεντρωνόταν στο ποιά είναι η πιθανότητα να βρίσκετα μέσα, κοντά ή και ενδεχομένω μακριά από αυτήν. Η θέση του γώρω από αυτή την περιοχή θα προσδιοριζόταν με μία κατανομή. Ο προσδιορισμός της κατάστασης γίνεται με πιθανοτικό τρόπο.

5.3 Αναπαράσταση αβεβαιότητας

5.3.1 Μορφές αβεβαιότητας

Η ανάγκη για διαχείριση της αβεβαιότητας των δεδομένων στις βάσεις δεδομένων με τρόπο διαφανή προς τον χρήστη οδήγησε στην ανάπτυξη ευέλικτων τρόπων αναπαράστασης των δεδομένων. Τα δεδομένα στις βάσεις δεδομένων με αβεβαιότητα, διακρίνονται σε δύο κατηγορίες:

- στην *αβεβαιότητα πλειάδων* (*tuple uncertainty*), η οποία χρησιμποιείται κυρίως για την μοντελοποίηση πιθανοτικών σχεσιακών δεδομένων σε πιθανοτικές βάσεις δεδομένων. Στην αβεβαιότητα πλειάδων, η παρουσία μία πλειάδας σε μια σχέση είναι πιθανοτική και πολλαπλές πλειάδες μπορεί να έχουν περιορισμούς μεταξύ τους, όπως ο αμοιβαίος αποκλεισμός.

- στην *αβεβαιότητα σε γνωρίσματα πλειάδων* (*attribute uncertainty*), η οποία χρησιμοποιείται στις βάσεις δεδομένων με αβεβαιότητα. Αντίθετα με την αβεβαιότητα πλειάδων, τώρα κάθε πλειάδα θεωρείται ότι ανήκει στη βάση δεδομένων, αλλά μία ή περισσότερες από τις ιδιότητές της δεν είναι γνωστές με βεβαιότητα. Ειδικότερα κάθε αβέβαιο αντικείμενο μοντελοποιείται από μία περιοχή αβεβαιότητας, εντός 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 επιθυμούμε να μοντελοποιήσουμε και από τις δυνατότητες του συστήματος.

5.3.2 Μοντέλο συνεχούς κατανομής

Στις χωροχρονικές βάσεις δεδομένων η αβεβαιότητα ενός αντικειμένου μπορεί να αναπαρασταθεί ως μία *περιοχή αβεβαιότητας* (*uncertainty region*) σε ένα συγκεκριμένο χρόνο t, η οποία είναι μία κλειστή περιοχή, έτσι ώστε το αντικείμενο να βρίσκεται πάντα μέσα σε αυτήν. Η συνάρτηση πυκνότητας πιθανότητας ενός αντικειμένου εκφράζει την πιθανότητα το αντικείμενο να βρίσκεται σε μία συγκεκριμένη θέση την χρονική στιγμή t. 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.

5.3.3 Μοντέλο διακριτών δειγμάτων

Σε αντιπαράθεση με την προηγούμενη μοντελοποίηση, στις χωροχρονικές βάσεις δεδομένων η αβεβαιότητα ενός αντικειμένου για μία χρονική στιγμή t μπορεί να αναπαρασταθεί με την χρήση διακριτών δειγμάτων. Κάθε διακριτό δείγμα βρίσκεται σε μία θέση και αντιπροσωπεύει την πιθανότητα το αντικείμενο να βρίσκεται σε αυτήν. Το αντικείμενο έχει πιθανότητα 0 να βρίσκεται σε άλλη θέση εκτός των δοσμένων διακριτών δειγμάτων του. Φυσικά, το άθροισμα των πιθανοτήτων όλων των δειγμάτων είναι 1.

5.3.4 Προσομοίωση Monte Carlo

Ο αλγόριθμος Monte Carlo αποτελεί μία αριθμητική μέθοδο για εύκολο 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. Συγκεκριμένα, επιλέγουμε τυχαία σημεία που βρίσκονται μέσα στην πειροχή αβεβαιότητας που ακολουθεί μία συνεχή κατανομή και για κάθε ένα από αυτά υπολογίζουμε την πιθανότητά τους, μέσω της τιμής της συνάρτησης πυκνότητας πιθανότητας. Προφανώς, όσο περισσότερα δείγματα ληφθούν, τόσο ακριβέστερη, αλλά και πιο χρονοβόρα, γίνεται η εκτίμηση της πιθανότητας.

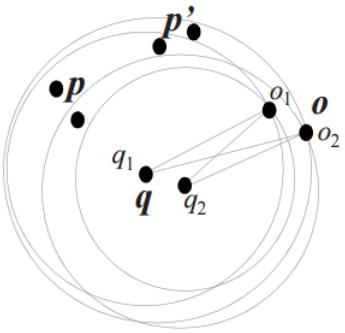
5.4 Επεξεργασία ερωτημάτων

Στην ενότητα αυτή παρουσιάζονται οι βασικότεροι τύποι ερωτημάτων καθώς και διάφορα ευρετήρια που συναντάμε συχνά στις βάσεις δεδομένων με αβεβαιότητα. Υπάρχουν τρία θέματα στην επεξεργασία των δεδομένων που εξετάζονται πάνω στις βάσεις δεδομένων με αβεβαιότητα: (ii) πρέπει να παρέχεται εγγύηση για την ακρίβεια των απαντήσεων, (ii) χρειάζεται σχεδιασμός αποτελεσματικών προσεγγίσεων για την διαχείριση μεγάλου όγκου αβέβαιων δεδομένων με μικρό χρόνο απόκρισης των απαντήσεων και (iii) 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.

5.4.1 Ευρετήρια

Όπως και στις συμβατικές βάσεις δεδομένων, για την επεξεργασία μεγάλου όγκου δεδομένων με αβεβαιότητα χρησιμποιούμε ευρετήρια για την διευκόλυνση ανάκτησης υποψήφιων

$$p_{nn}(q_1, o_1) = (1\text{-}1/2)\cdot(1\text{-}2/2) = 0/4$$
$$p_{nn}(q_1, o_2) = (1\text{-}2/2)\cdot(1\text{-}2/2) = 0/4$$
$$p_{nn}(q_2, o_1) = (1\text{-}0/2)\cdot(1\text{-}0/2) = 4/4$$
$$p_{nn}(q_2, o_2) = (1\text{-}0/2)\cdot(1\text{-}1/2) = 2/4$$
$$p_{nn}(q, o) = (4/4 + 2/4)/4 = 6/16 = 37{,}5\ \%$$



Σχήμα 5.2: Υπολογισμός πιθανοτήτων εγγύτερων γειτόνων στο [12]

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 δίνεται η τελική λύση.

Στις βάσεις δεδομένων με αβεβαιότητα, όπως και στις πιθανοτικές, το ευρετήριο που χρησιμοποιείται περισσότερο από κάθε άλλο (μαζί με παραλλαγές του) είναι το πιθανοτικό *R-δέντρο* (*probabilistic R-tree*). Το πιθανοτικό *R-δέντρο* αποτελεί 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 (τα οποία είδαμε και στο κεφάλαιο 4). Στη συνέχεια, τα MBR που βρίσκονται κοντά μεταξύ τους εμπεριέχονται μέσα σε άλλα MBR, με την διαδικασία αυτή να συνεχίζεται έως ότου 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 πληροφορίαγια τις πιθανότητες των δειγμάτων τηρείται στο δέντρο. Πιθανότητες που υπολογίζονται με βάση τα στοιχεία που εμπεριέχουν, συνοδεύουν και όλα τα ενδιάμεσα MBR. Εκτός του πιθανοτικού *R-δέντρου* στην βιβλιογραφία συναντάμε και άλλα ευρετήρια ανάλαογα με το πρόβλημα που αντιμετωπίζεται. Στο [21] παρουσιάζεται το U-tree, 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 αντικειμένων χωρίς να χρειάζεται να λάβει υπόψη την πιθανότητα εμφάνισής τους. Τέτοιες πληροφορίες διατηρούνται σε όλα τα επίπεδα του U-tree ώστε να αποφεύγεται η πρόσβαση σε υποδέντρα που δεν περιέχουν κανένα αποτέλεσμα. Επιπλέον τα U-trees είναι πλήρως δυναμικά, δηλαδή τα αντικείμενα μπορούν να εισάγονται ή να διαγράφονται με οποιαδήποτε σειρά.

5.4.2 Βασικά ερωτήματα

- **Top-k** : Τα top-k ερωτήματα ανακτούν τα κορυφαία k στοιχεία από μία συλλογή (π.χ. επιθυμούμε να βρούμε τους 10 πιο ακριβοπληρωμένους από ένα σύνονο ποδοσφαιρι-

στών). Ένας αλγόριθμος top-k είναι επίσης γνωστός ως αλγόριθμος κατωφλίου, αφού τερματζίεται όταν ένα συγκεκριμένο όριο επιτυγχάνεται.

- **Σύνδεση βάσει ομοιότητας (Similarity Join)** : Στις βάσεις δεδομένων με αβεβαιότητα υπάρχει ενδιαφέρον για τη σύνδεση στοιχείων μεταξύ δύο πινάκων βάσει της ομοιότητα 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 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 join το οποίο θα δίνει αποδοτικά και αποτελεσματικά τις λύσεις με την υψηλότερη εμπιστοσύνη.

- **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 [17] γίνεται υπολογισμός τέτοιων διαρκών ερωτημάτων για κινούμενα αντικείμενα τα οποία ακολουθούν τη διδιάστατη κανονική κατανομή.
  - *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.
  - *Ερώτημα των k-εγγύτερων γειτόνων*, όπου ζητείται να βρεθούν τα k κοντινότερα σημεία σε μία εστία q. Στο [20] προσεγγίζεται το πρόβλημα για διαρκή ερωτήματα πάνω σε αντικείμενα των οποίων η τροχιά είναι γνωστή, όμως η εκάστοτε θέση τους να συνοδεύεται από μία πιθανότητα εμφάνισης σε κάθε χρονική στιγμή. Επίσης, μία από τις πιο δημοφιλείς προσεγγίσεις του προβλήματος είναι αυτή του [12] όπου έχουμε στατικά δεδομένα και τα αντικείμενα μοντελοποιούνται με διακριτά δείγματα πιθανότητας. Όπως, φαίνεται στο σχήμα 5.2, τα αντικείμενα και το εκάστοτε ερώτημα μπορούν να αναπαρασταθούν με διακριτά δείγματα, κάθε ένα από τα οποία έχει μία πιθανότητα, και μπορεί να υπολογιστεί η πιθανότητα κάποιο αντικείμενο να είναι εγγύτερος γείτονας. Η μέθοδος που αναλύεται βασίζεται στις μετρικές mindist, minmaxdist καθώς και στη συσταδοποίηση σε MBR και αποδεικνύεται η αποτελεσμάτικότητά της στην πειραματική μελέτη. Η προσέγγιση του προβλήματος, όμως, δεν έχει γίνει για ρεύματα δεδομένων και διαρκή ερωτήματα όταν οι περιοχές αβεβαιότητας των αντικειμένων μοντελοποιούνται με συνεχή κατανομή και αυτό θα επιχειρεί να επιλύσει η παρούσα διπλωματική εργασία.

– *Αντίστροφο ερώτημα των k-εγγύτερων γειτόνων*, όπου ζητείται να βρεθούν τα k αντικείμενα που έχουν ένα άλλο ως κοντινότερο γείτονά τους με κάποια δοσμένη πιθανότητα. Το πρόβλημα έχει προσεγγιστεί για αβέβαια αντικείμενα στα [2],[4] με χρήση τεχνικών κλαδέματος, τόσο χωρικών όσο και πιθανοτικών.

- **Συνάθροιση (Aggregation)** : Σε αυτά τα ερωτήματα η αλληλεπίδραση μεταξύ πολλών στοιχείων παίζει σημαντικό ρόλο στην αποτίμησή τους. Οι προκύπτουσες πιθανότητες επηρεάζονται σε μεγάλο βαθμό από την αβεβαιότητα των χαρακτηριστικών των άλλων στοιχείων.

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

5.5 Παρουσίαση αποτελεσμάτων

Στη βιβλιογραφία έχουν προταθεί διάφοροι τρόποι παρουσίασης των αποτελεσμάτων που προκύπτουν από την επεξεργασία των δεδομένων στις βάσεις δεδομένων με αβεβαιότητα. Οι τρόποι αυτοί ποικίλουν ανάλογα με τις τεχνικές επεξεργασίας που ακολουθούνται και με τους διατιθέμενους πόρους. Στην ενότητα αυτή δίνονται οι κυριότεροι τρόποι για παρουσίαση των απαντήσεων πέρα από τις απλές πιθανότητες και τα σφάλματα, όπως είναι το διάστημα εμπιστοσύνης, η ύπαρξη κατωφλίου και η κατάταξη των αποτελεσμάτων.

5.5.1 Διαστήματα εμπιστοσύνης

Αρκετά συχνά, σε προβλήματα όπου οι υπολογισμοί οδηγούν σε τελικά αποτελέσματα που δεν μπορούν να δοθούν με μία απλή πιθανότητα, χρησιμοποιούμε διαστήματα εμπιστοσύνης. Τα διαστήματα εμπιστοσύνης μας δίνουν τη δυνατότητα να μπορούμε να συγκρίνουμε καταστάσεις και να εκφράζουμε το βαθμό βεβαιότητας της απάντησης. Ανάλογα με την κατανομή που ακολουθείται, προσδιορίζεται η μέση τιμή και η διακύμανση του διαστήματος εμπιστοσύνης.

5.5.2 Χρήση κατωφλίων

Λαμβάνοντας υπόψη ένα όριο εμπιστοσύνης θ, ένα ερώτημα με κατώφλι επιστρέφει τα αντικείμενα που μπορούν να επιλεγούν, επειδή συνοδεύονται από πιθανότητα μεγαλύτερη ή ίση του θ. Σε περιπτώσεις χωρικών ερωτημάτων, τα τελικά αποτελέσματα που δίνονται πρέπει να ξεπερνούν ένα πιθανοτικό κατώφλι. Όσο μεγαλύτερη η τιμή του κατωφλίου τόσο λιγότερα αποτελέσματα θα παρουσιάζονται συνήθως. Πιθανοτικά κατώφλια χρησιμοποιήθηκαν και σε

άλλους τύπους ερωτημάτων, όπως στη σύνδεση βάσει ομοιότητας και στα συναθροιστικά ερωτήματα τιμής.

5.5.3 Κατάταξη

Ένα σημαντικό πρόβλημα στις πιθανοτικές βάσεις δεδομένων είναι η 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 με βάση τα περιθώρια εμπιστοσύνης ή τις πιθανότητες που τα συνοδεύουν. Η ιεραρχημένη κατάταξη των αντικειμένων της απάντησης μπορεί επίσης να συνδυαστεί με τη χρήση κατωφλίου.

Κεφάλαιο 6

Μοντελοποίηση του προβλήματος

6.1 Εισαγωγή

Η παρούσα διπλωματική εργασία έχει ως κύριο θέμα την ανάπτυξη και υλοποίηση ενός αλγορίθμου για αποτίμηση πιθανοτικών ερωτημάτων k-εγγύτερων γειτόνων για αβέβαιες θέσεις κινούμενων αντικειμένων. Τα κινούμενα αυτά αντικείμενα μπορούν να εντοπιστούν γεωγραφικά (μέσω GPS), όμως το ακριβές στίγμα κάθε αντικειμένου δεν αποκαλύπτεται ποτέ στον κεντρικό υπολογιστή του συστήματος, αλλά γνωστοποιείται μία ευρύτερη περιοχή του. Συγκεκριμένα, η αβεβαιότητα της γεωγραφικής θέσης μοντελοποιείται με κάποια πιθανοτική κατανομή, η οποία δεν θεωρείται ομοιόμορφη, αλλά μπορεί να ποικίλλει. Οι χρήστες των κινούμενων αντικειμένων έχουν τη δυνατότητα να υποβάλλουν τα ερωτήματα διαρκείας εγγύτερων γειτόνων θέτοντας τις επιθυμητές παραμέτρους, οπότε ο επεξεργαστής οφείλει να δίνει τακτικά ενημερωμένες προσεγγιστικές απαντήσεις.

Στο κεφάλαιο αυτό γίνεται μία σαφής διατύπωση του μοντέλου του προβλήματος που μελετάται στην παρούσα διπλωματική εργασία. Για το λόγο αυτό περιγράφονται τα μοντέλα των κινούμενων αντικειμένων και των ερωτημάτων καθώς και η μορφή των απαντήσεων. Επίσης, γίνεται μία ανάλυση του πιθανοτικού μοντέλου που υιοθετήθηκε για τον καθορισμό των περιοχών αβεβαιότητας.

6.2 Μοντέλο συστήματος

Παρακάτω περιγράφονται οι υποθέσεις που αφορούν τα γεωμετρικά και πιθανοτικά χαρακτηριστικά των αντικειμένων και των ερωτημάτων που τίθενται. Οι υποθέσεις αυτές βασίζονται σε αυτή της ομοιόμορφης κατανομής των αντικειμένων και των ερωτημάτων πάνω στο Ευκλείδειο επίπεδο.

6.2.1 Κινούμενα Αντικείμενα

Η θέση των κινούμενων αντικειμένων θεωρείται πως ανανεώνεται με την ίδια περίοδο ανά αντικείμενο (timestamp). Κάθε κινούμενο αντικείμενο, το οποίο συμβολίζουμε με o_i, στέλνει

το γεωγραφικό του στίγμα σε κάθε τέτοια περίοδο. Η αναπαράσταση των κινούμενων αντικειμένων γίνεται με με τη μορφή κυκλικών περιοχών που ακολουθούν την κανονική κατανομή δύο διαστάσεων (Normal Distribution). Συγκεκριμένα, περιορίζουμε την κατανομή σε μία κυκλική περιοχή ώστε να είναι δυνατή η επεξεργασία των δεδομένων, αφού τυπικά η αβεβαιότητα της κανονικής κατανομής εκτείνεται στο άπειρο στον διδιάστατο χώρο. Ο λόγος για τον οποίο επιλέχθηκε η κανονική κατανομή για τις πιθανοτικές αυτές κυκλικές περιοχές, είναι επειδή παρέχει τη δυνατότητα να μοντελοποιήσουμε αποτελεσματικά το πρόβλημα προς εξέταση. Συγκεκριμένα, η θέση ενός κινούμενου αντικειμένου, διαισθητικά, αναπαρίσταται από το στίγμα ενός κέντρου και την κυκλική περιοχή γύρω από αυτό, που το μέγεθός της χαρακτηρίζεται από μία ακτίνα R. Όπως θα δούμε και στη συνέχεια, το μέγεθος αυτής της ακτίνας παίζει αρκετά σημαντικό ρόλο, διότι καθορίζει την έκταση της περιοχής που η πραγματική θέση του αντικειμένου μπορεί να βρίσκεται. Εξαιτίας της φύσης της κανονικής κατανομής, η πιθανότητα του αντικειμένου να βρίσκεται σε μία συγκεκριμένη θέση αυξάνεται όσο κινούμαστε προς το κέντρο του κύκλου, ενώ όσο απομακρυνόμαστε από αυτό μειώνεται. Αυτό το χαρακτηριστικό την καθιστά ιδιαιτέρως κατάλληλη για την αναπαράσταση της κατάστασης και της αβεβαιότητας των αντικειμένων σε σχέση με άλλες πιθανοτικές κατανομές. Προς υποστήριξη του παραπάνω επιχειρήματος, εάν λ.χ. υποθέταμε πως επιλέγαμε την ομοιόμορφη κατανομή, τότε όλες οι θέσεις μιας περιοχής θα έδιναν ίσες πιθανότητες ύπαρξης του αντικειμένου. Με αυτόν τον τρόπο δεν θα υπήρχε διαφοροποίηση στα βάρη ανάμεσα στις θέσεις που βρίσκονται στην πραγματικότητα πιο κοντά στην πραγματική του θέση και σε αυτές που θα ήταν πιο απίθανο να βρίσκεται, πράγμα μη επιθυμητό. Επιπρόσθετα, άλλες γνωστές κατανομές τόσο συνεχείς όσο και διακριτές,όπως είναι η γάμμα, η διωνυμική, η εκθετική, η γεωμετρική, η Bernoulli, αλλά ακόμα και τυχαία διακριτά δείγματα, υπολείπονται έναντι της κανονικής κατανομής λόγω των ποιοτικών χαρακτηριστικών τους.

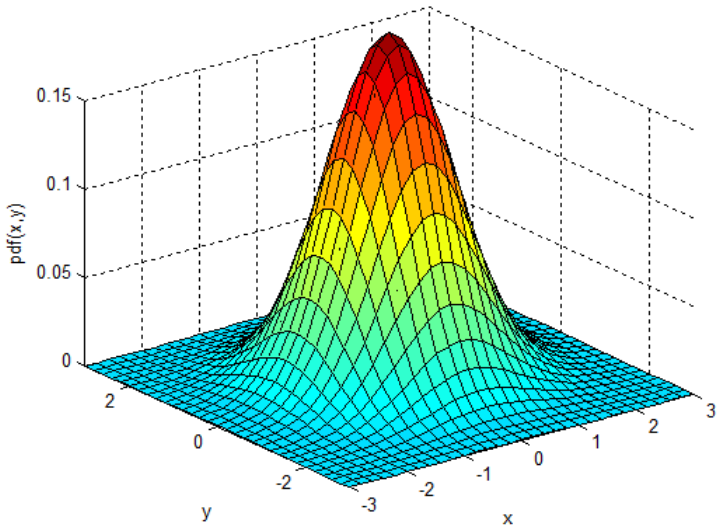
Αναλυτικά, η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής δύο διαστάσεων δίνεται από τον παρακάτω τύπο:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2]} \tag{6.1}$$

όπου με ρ συμβολίζουμε τον συντελεστή συσχέτισης (*correlation coefficient*) μεταξύ των x και y συντεταγμένων. Στην ειδική περίπτωση που έχουμε $\sigma_x = \sigma_y = \sigma$ τότε $\rho = 0$ και έτσι η παραπάνω σχέση γίνεται:

$$f(x,y) = \frac{1}{2\pi\sigma_x^2} e^{-\frac{1}{2}[(\frac{x-\mu_x}{\sigma_x})^2 + (\frac{y-\mu_y}{\sigma_y})^2]} \tag{6.2}$$

Με βάση την υπόθεση πως επιθυμούμε να απλοποιήσουμε ακόμα περισσότερο την παραπάνω σχέση, μπορούμε να πάρουμε ως σημείο αναφοράς του $(x,\ y)$ - συστήματος συντεταγμένων το $(\mu_x,\ \mu_y) = (0,\ 0)$. Για να το επιτύχουμε αυτό, πολύ εύκολα αντικαθιστούμε τα x και τα y με τις μετασχηματισμένες μεταβλητές $x' = x$ - μ_x και $y' = y$ - μ_y. Με τον τρόπο αυτό μπορούμε να αναδιατυπώσουμε τη συνάρτηση πυκνότητας πιθανότητας χρησιμοποιώντας πολικές συντεταγμένες $(r,\ \theta)$ και ως σημείο αναφοράς το $(x,\ y) = (0,\ 0)$. Στην περίπτωση αυτή, η θέση $(x,\ y)$ θα έχει πολικές συντεταγμένες με:



Σχήμα 6.1: Συνάρτηση πυκνότητας διδιάστατης πιθανότητας κανονικής κατανομής για $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 0.9$ και $\rho = 0$

$$r = \sqrt{x^2 + y^2}$$
$$\theta = \tan^{-1}(\frac{x}{y})$$

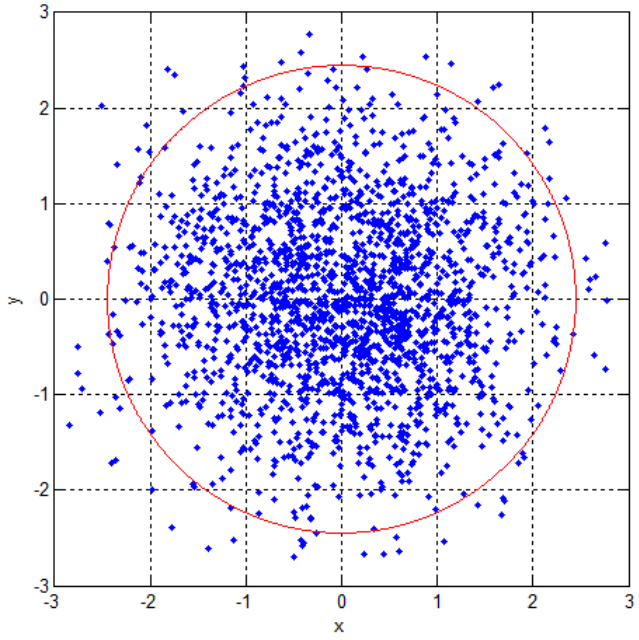
και η συνάρτηση πυκνότητας πιθανότητας γίνεται:

$$f(x,y) = \frac{1}{2\pi\sigma_x^2} e^{-\frac{1}{2}(\frac{r}{\sigma_x})^2} \tag{6.3}$$

Όπως φαίνεται και από την παραπάνω σχέση, η γωνία θ δεν επηρεάζει τη συνάρτηση πυκνότητας πιθανότητας καθώς η *f(x, y)* εξαρτάται μόνο από την απόσταση της θέσης από το σημείο αναφοράς και όχι από την κατεύθυνσή του.

Μια τρισδιάστατη αναπαράσταση της συνάρτησης πυκνότητας πιθανότητας *f(x, y)* για τυχαία σ_x και σ_y παρουσιάζεται στο σχήμα 6.1. Από το ύψος της επιφάνειας που σχηματίζεται μπορούμε να πάρουμε μία ένδειξη του μεγέθους της *f(x, y)*. Ακόμα, στο σχήμα 6.2 δίνεται ένα παράδειγμα διαγράμματος διασποράς διδιάστατης κανονικής κατανομής, την οποία περιορίζουμε σε μία συγκεκριμένη κυκλική περιοχή. Στο συγκεκριμένο παράδειγμα, η κατανομή έχει κέντρο στο $(\mu_x, \mu_y) = (0, 0)$ και οι τυπικές αποκλίσεις είναι $\sigma_x = \sigma_y = 1$.

Με βάση, λοιπόν, όλα τα παραπάνω καταλήγουμε στην υπόθεση πως οι κυκλικές περιοχές αβεβαιότητας των κινούμενων αντικειμένων ακολουθούν την κανονική κατανομή δύο διαστάσεων με $\sigma_x = \sigma_y = \sigma$ και $\mu_x = \mu_y$. Κάθε τέτοιος κύκλος κανονικής κατανομής μπορεί να λαμβάνει διαφορετική τιμή για την ακτίνα R_i. Η ακτίνα αυτή θα είναι ίση με το τριπλάσιο της τυπικής απόκλισης σ_i της κατανομής και έτσι σε κάθε περιοχή αβεβαιότητας θα καλύπτεται περίπου το 99.73% της συνολικής πιθανότητας. Η τιμή της τυπικής απόκλισης δεν είναι αυθαίρετη, αλλά λαμβάνεται από ένα προκαθορισμένο σύνολο που περιέχει n διαφορετικές τιμές.



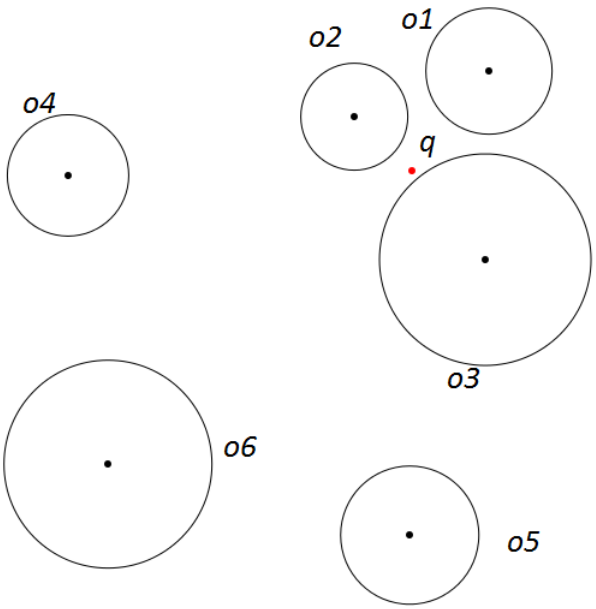
Σχήμα 6.2: Παράδειγμα διαγράμματος διασποράς διδιάστατης κανονικής κατανομής για $\sigma_x = \sigma_y = 1$

Ο εκάστοτε χρήστης κινητής συσκευής θα μπορεί να κάνει αυτή την επιλογή της τυπικής απόκλισης, και κατ'επέκταση της ακτίνας, η οποία θα βασίζεται στο κατά πόσο επιθυμεί να αποκρύψει την ακριβή γεωγραφική του θέση. Στην περίπτωση που δεν τον ενδιαφέρει ιδιαίτερα να αποκρύψει τη θέση του στο σύστημα τότε επιλέγει μικρή τυπική απόκλιση (άρα μικρότερη αβεβαιότητα), ενώ εάν επιθυμεί μεγαλύτερη προστασία της ιδιωτικότητάς του επιλέγει μεγαλύτερη τυπική απόκλιση (άρα μεγαλύτερη αβεβαιότητα).

6.2.2 Κινούμενα Ερωτήματα

Όπως και στην περίπτωση των αντικειμένων, έτσι και για τα ερωτήματα η θέση τους ανανεώνεται τακτικά, ανά κάποια χρονόσημα (*timestamp*). Τα ερωτήματα, τα οποία θα τα συμβολίζουμε με q_i, δίνονται με τη μορφή ακριβών γεωγραφικών στιγμάτων, που ορίζουν την εστία (*focal point*) ενδιαφέροντος. Ουσιαστικά, πρόκεινται για συγκεκριμένες περιοχές, όπως λ.χ. ένας κινηματογράφος ή ένα κατάστημα, των οποίων η γεωγραφική θέση είναι γνωστή.

Κάθε κινούμενο ερώτημα συνοδεύεται επίσης από έναν αριθμό k_i, ο οποίος θα εκφράζει τον αριθμό των εγγύτερων γειτόνων που είναι επιθυμητό να βρεθεί κοντά στη συγκεκριμένη εστία ενδιαφέροντος. Η τιμή του εκάστοτε k_i για κάθε ερώτημα θα λαμβάνεται από έναν συγκεκριμένο σύνολο, το οποίο θα έχει προκαθορίσει το σύστημα. Η επιλογή αυτή γίνεται από τον χρήστη, με βάση το μέγεθος της απάντησης που θέλει να λάβει. Επιπρόσθετα, κάθε κινούμενο ερώτημα δηλώνει ένα επιθυμητό κατώφλι θ_i, με το οποίο εκφράζεται η μικρότερη ανεκτή πιθανότητα για να θεωρηθεί κάποιο αντικείμενο υποψήφιο ως εγγύτερος γείτονας. Και στην περίπτωση αυτή η επιλογή γίνεται από τον χρήστη της εκάστοτε κινητής συσκευής



Σχήμα 6.3: Παράδειγμα πιθανοτικού ερωτήματος k-εγγύτερων γειτόνων, για εστία ενδιαφέροντος q

από ένα προκαθορισμένο σύνολο τιμών που του παρέχει το σύστημα. Ασφαλώς, η επιλογή κατωφλίου από τον χρήστη βασίζεται στο κατά πόσο επιθυμεί η απάντηση που θα πάρει να περιέχει αντικείμενα, τα οποία είναι αρκετά πιθανό να βρίσκονται κοντά του (και άρα επιλέγει μεγάλη τιμή για το θ_i) ή απλώς τον ενδιαφέρει να πάρει μια λιγότερο σαφή αλλά ποιοτική εικόνα (οπότε επιλέγει μικρότερες τιμές για το θ_i).

Στο σχήμα 6.3 παρατηρούμε ένα στιγμιότυπο πιθανοτικού ερωτήματος k-εγγύτερων γειτόνων, όπου έχουμε ορίσει ως εστία ενδιαφέροντος το q. Τα αντικείμενα 1-5 είναι κυκλικές περιοχές αβεβαιότητας, με μεταβλητό μέγεθος η κάθε μία. Ας υποθέσουμε, για παράδειγμα, ότι $k = 3$ και $\theta = 50\%$. Τότε, διαισθητικά, περιμένουμε πως ως αποτέλεσμα θα λάβουμε τα αντικείμενα 1,2 και 3, των οποίων οι περιοχές αβεβαιότητας βρίσκονται εγγύτερα στην εστία q.

6.3 Αποτελέσματα ερωτημάτων

Η αποτίμηση του κάθε ερωτήματος γίνεται κι αυτή ανά χρονόσημο. Η μορφή του αποτελέσματος που δίνεται ως απάντηση σε κάθε ερώτημα περιγράφεται ως εξής:

- Επιστρέφεται μία λίστα k αντικειμένων, όσος και ο επιθυμητός αριθμός εγγύτερων γειτόνων που θέτει το ερώτημα.
- Τα αντικείμενα αυτά είναι ταξινομημένα ως πρώτος, δεύτερος,...,k εγγύτερος γείτονας με βάση τη φθίνουσα τιμή πιθανότητας.
- Εννοείται πως κάθε τιμή πιθανότητας θα είναι μεγαλύτερη ή ίση σε σχέση με το κατώφλι

θ_i που έχει τεθεί από το ερώτημα.

- Για κάθε αντικείμενο αναφέρεται το αναγνωριστικό του, καθώς και η τιμή της πιθανότητάς που συγκεντρώνει.
- Η αποτίμηση ερωτημάτων γίνεται εκ νέου σε κάθε χρονόσημο. Τυχόν προηγούμενα αποτελέσματα δεν αξιοποιούνται σε επόμενους κύκλους εκτέλεσης.
- Ενημερώσεις αποτελεσμάτων μπορεί να στέλνονται μόνο όταν αλλάξει τουλάχιστον ένας από τους k ευρεθέντες γείτονες ή η κατάταξή τους.

Κεφάλαιο 7

Επεξεργασία πιθανοτικών ερωτημάτων k-εγγύτερων γειτόνων

7.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται ο αλγόριθμος που αναπτύχθηκε, με σκοπό την αποτίμηση ερωτημάτων εγγύτερων γειτόνων για αβέβαιες θέσεις κινούμενων αντικειμένων. Το στάδιο επεξεργασίας αυτών των ερωτημάτων χρησιμοποιεί:

- Ευρετήριο χωρικού καννάβου (Grid Partitioning) και έλεγχο ανά επάλληλες ζώνες (levels)
- Τεχνικές κλαδέματος (Pruning)

με στόχο την μείωση του κόστους επεξεργασίας των δεδομένων.

Ο αλγόριθμος αναλύεται στις εξής δύο φάσεις: μία φάση φιλτραρίσματος (filtering phase) και μία φάση εκλέπτυνσης (refinement phase). Πιο συγκεκριμένα, τα αντικείμενα που εξετάζονται για κάθε ερώτημα φιλτράρονται από τις τεχνικές κλαδέματος, οι οποίες βασίζονται σε γεωμετρικά και πιθανοτικά χαρακτηριστικά των κινούμενων αντικειμένων και των ερωτημάτων, έχοντας ως αποτέλεσμα πολύ μικρός αριθμός από τα αρχικά δεδομένα να χρειάζεται να ελεγχθούν περαιτέρω. Στην διαδικασία της εκλέπτυνσης γίνεται αναλυτική και λεπτομερής αποτίμηση για τα υποψήφια αντικείμενα που πέρασαν από την προηγούμενη φάση. Ο αλγόριθμος δεν παρέχει ακριβή αποτελέσματα αλλά είναι προσεγγιστικός. Ωστόσο, οι απαντήσεις που δίνει είναι αξιόπιστες ποιοτικά και αντιπροσωπευτικές της πραγματικότητας. Επίσης, όπως προκύπτει κι από τα πειραματικά αποτελέσματα ο χρόνος εκτέλεσης είναι αρκετά ικανοποιητικός και ανταποκρίνεται σε πραγματικές συνθήκες. Τέλος, επισημαίνεται πως ο αλγόριθμος λειτουργεί online, δεχόμενος είσοδο και εξάγοντας αποτελέσματα ανά κύκλους εκτέλεσης (ανά χρονόσημο).

7.2 Γενική ιδέα του αλγορίθμου

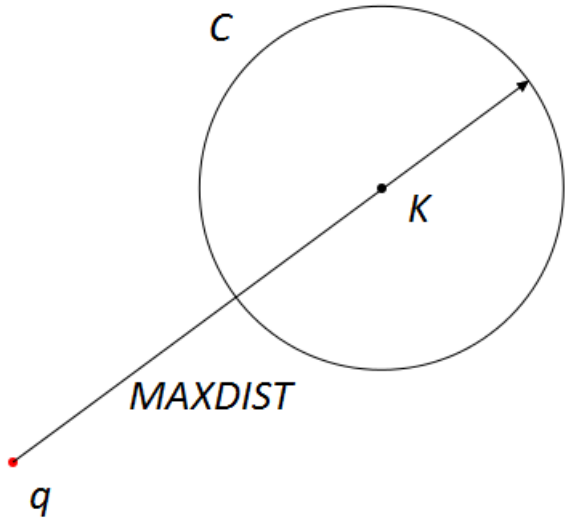
Είναι σαφές πως τα αποτελέσματα κάθε ερωτήματος διαρκείας πιθανότερων εγγύτερων γειτόνων (εν συντομία $k\theta NN$) πρέπει να γνωστοποιούνται με κάθε αλλαγή θέσεων των χρηστών. Όπως έχουμε ήδη αναφέρει, χρησιμοποιούμε ένα ευρετήριο χωρικού καννάβου (uniform grid partitioning), ώστε να τοποθετήσουμε στα κελιά τους τα κέντρα αβεβαιοτήτων των πιθανοτικών περιοχών των κινητών αλλά και τις ακριβείς εστίες των ερωτημάτων. Σε κάθε χρονόσημο, η αποτίμηση των ερωτημάτων ακολουθεί την, αρκετά διαδεδομένη (π.χ. χρησιμοποιείται στο [4]), στρατηγική *φιλτραρίσματος* και *εκλέπτυνσης* (*filter and refinement*):

- *Φιλτράρισμα*: Στο στάδιο αυτό για κάθε ερώτημα, ξεκινάμε από τη σημειακή εστία q και διερευνούμε επάλληλες ζώνες κελιών κυκλικά γύρω από το q. Σε κάθε τέτοιο κελί, ελέγχουμε τα κινούμενα αντικείμενα των οποίων τα κέντρα των περιοχών αβεβαιότητας είναι τοποθετημένα σε αυτό και αποφασίζουμε εάν κάποια από αυτά είναι υποψήφιοι εγγύτεροι γείτονες. Πρακτικά, η επιλογή ή μη των κινητών ως υποψηφίων βασίζεται κυρίως στα γεωμετρικά χαρακτηριστικά τους, αλλά και στο κατά πόσο καλύπτεται πιθανοτικά η περιοχή αβεβαιότητάς τους τη δεδομένη στιγμή. Η διαδικασία αυτή διακόπτεται όταν επιβεβαιωθεί ότι διερεύνηση σε περαιτέρω κελιά δε θα συνεισφέρει νέους υποψήφιους πλησιέστερους στο q γείτονες.

- *Εκλέπτυνση*: Έπειτα από το προηγούμενο στάδιο, επεκτεινόμαστε από τη σημειακή εστία q του ερωτήματος σε κυκλικές περιοχές αναζήτησης (*search regions*). Σε κάθε μία από αυτές υπολογίζουμε το *ποσοστό πιθανοτικής κάλυψης* Φ για κάθε υποψήφιο κινούμενο αντικείμενο που έχουμε βρει στη διαδικασία του φιλτραρίσματος. Προς αποφυγή του μεγάλου κόστους πράξεων που θα επέφερε ένας ακριβής και αναλυτικός υπολογισμός του Φ, η εκτίμηση της τιμής του γίνεται προσεγγιστικά. Αυτό επιτυγχάνεται, χρησιμοποιώντας την *αθροιστική συνάρτηση κατανομής* (*CDF, Cumulative Distribution Function*) κατά μήκος της ευθείας που ενώνει την εστία q και το κέντρο αβεβαιότητας του εκάστοτε υποψήφιου γείτονα. Ουσιαστικά, εξετάζουμε την κατανομή κατά μήκος της εγκάρσιας τομής της στο ύψος της διαμέτρου της κυκλικής περιοχής αβεβαιότητας. Εφαρμόζοντας αυτή την τεχνική υπολογισμού πιθανοτικής κάλυψης για τους υποψήφιους εγγύτερους γείτονες σε κάθε περιοχή αναζήτησης, διακόπτουμε τη διαδικασία όταν βρεθούν τουλάχιστον k αντικείμενα (ο αριθμός εγγύτερων γειτόνων που ζητεί το ερώτημα να βρεθούν) με πιθανοτική κάλυψη μεγαλύτερη ή ίση του κατωφλίου που θέτει το ερώτημα (δηλαδή $\Phi \geq \theta$). Τότε τυπώνουμε την κατάταξη των k-εγγύτερων γειτόνων για την συγκεκριμένη απόσταση από το q.

7.3 Βασικές έννοιες

Πριν την ανάλυση του αλγορίθμου, παρατίθενται μερικοί ορισμοί που είναι αναγκαίοι για την επεξήγησή του:



Σχήμα 7.1: Παράδειγμα μέγιστης απόστασης (MAXDIST) σημείου από περιφέρεια κύκλου

- *Ευκλείδεια απόσταση μεταξύ σημειακών θέσεων*

  Η Ευκλείδεια απόσταση μεταξύ δύο σημείων $o_1(x_1,\ y_1)$ και $o_2(x_2,\ y_2)$ στο διδιάστατο καρτεσιανό επίπεδο ορίζεται κλασικά ως εξής:

$$L_2(o_1, o_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{7.1}$$

- *Μέγιστη απόσταση σημείου από περιφέρεια κύκλου (MAXDIST)*

  Ορίζουμε ως μέγιστη απόσταση ενός σημείου q από την περιφέρεια ενός κύκλου C με κέντρο K, την Ευκλείδεια απόσταση μεταξύ του σημείου αυτού και του αντιδιαμετρικού σημείου της περιφέρειας του κύκλου κατά μήκος της ευθείας που διέρχεται από τα q και K. Αναλυτικά:

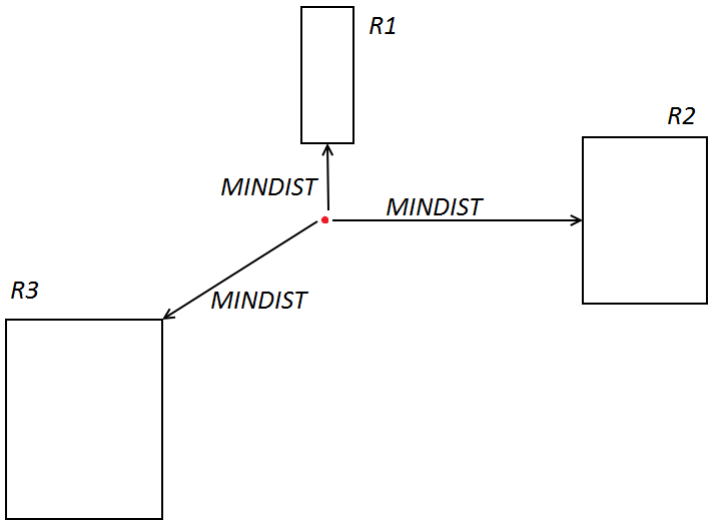
$$MAXDIST(q, C) = dist(q, K) + r(C) \tag{7.2}$$

  Όπου, r(C) η ακτίνα του κύκλου C. Στο σχήμα 7.1 παρατίθεται μία οπτικοποίηση της μέγιστης απόστασης μεταξύ ενός σημείου q και μίας περιφέρειας κύκλου C.

- *Ελάχιστη απόσταση σημείου από ορθογώνιο (MINDIST)*

  Ορίζουμε ως ελάχιστη απόσταση ενός σημείου $q(x,\ y)$ από ένα ορθογώνιο R με άκρα κυρίας διαγωνίου $(x_{min},\ y_{min})$ και $(x_{max},\ y_{max})$, την ελάχιστη απόσταση του σημείου από την περίμετρο του ορθογωνίου. Αναλυτικά:

$$MINDIST(q, R) = \sqrt{(x - r)^2 + (y - t)^2} \tag{7.3}$$



Σχήμα 7.2: Παράδειγμα ελάχιστης απόστασης (MINDIST) μεταξύ ενός σημείου και διαφόρων ορθογωνίων

Όπου σύμφωνα με το [18]

$$r = \begin{cases} x_{min} & \text{if } x < x_{min} \\ x_{max} & \text{if } x > x_{max} \\ x & \text{otherwise} \end{cases} \qquad t = \begin{cases} y_{min} & \text{if } y < y_{min} \\ y_{max} & \text{if } y > y_{max} \\ y & \text{otherwise} \end{cases}$$

Στο σχήμα 7.2 φαίνονται οι ελάχιστες αποστάσεις μεταξύ ενός σημείου και 3 διαφορετικών ορθογωνίων.

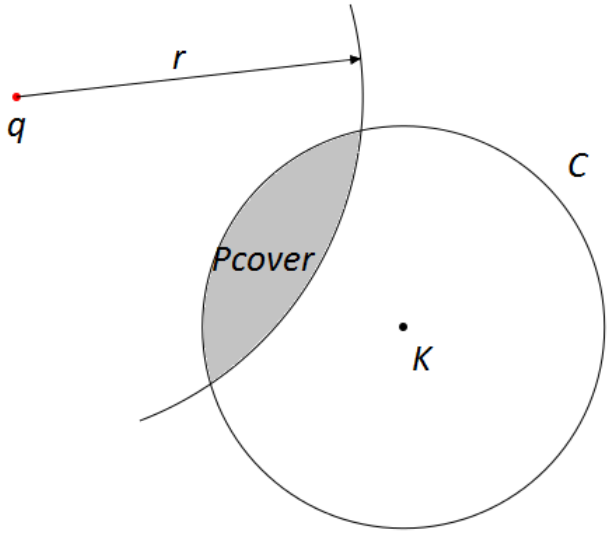
- *Πιθανοτική κάλυψη περιοχής* Έστω, σημείο q και κυκλική περιοχή αβεβαιότητας C που ακολουθεί μία οποιαδήποτε πιθανοτική κατανομή. Για έναν συγκεκριμένο κύκλο ακτίνας r με κέντρο το q (που συμβολίζουμε με $O(q,\ r)$), *η πιθανοτική κάλυψη της* C *ορίζεται:*

$$P_{cover}(O(q,r), C) = \int_{O(q,r)\cap C|y} \int_{O(q,r)\cap C|x} pdf(x,y)\, dxdy \tag{7.4}$$

όπου με $O(q,\ r) \cap C\,|x$ συμβολίζουμε το διάστημα κατά μήκος του x-άξονα στο οποίο οι περιοχές $O(q,\ r)$ και C επικαλύπτονται. Στο σχήμα 7.3 παρατηρείται η προς υπολογισμό πιθανοτική κάλυψη μίας περιοχής αβεβαιότητας.

7.4 Επεξεργασία Δεδομένων

Στην επεξεργασία των δεδομένων κύριο ρόλο έχουν οι τεχνικές μείωσης των συνολικών πράξεων ώστε ο αλγόριθμος να είναι όσο το δυνατόν αποδοτικότερος, τόσο από άποψη χρόνου εκτέλεσης όσο και από άποψη ποιότητας αποτελεσμάτων. Σε πρώτο στάδιο, ακολουθείται η τεχνική της ομοιόμορφης κατάτμησης σε κάνναβο, έτσι ώστε να δεικτοδοτηθούν τα αντικείμενα εντός των αντίστοιχων κελιών του. Αμέσως μετά, ακολουθείται η διαδικασία αποτίμησης



Σχήμα 7.3: Πιθανοτική κάλυψη περιοχής αβεβαιότητας σε ακτίνα αναζήτησης r από εστία q

των ερωτημάτων, στην οποία χρησιμοποιούνται τεχνικές κλαδέματος για την αποφυγή επεξεργασίας απίθανων υποψηφίων αντικειμένων.

7.4.1 Δομές δεδομένων

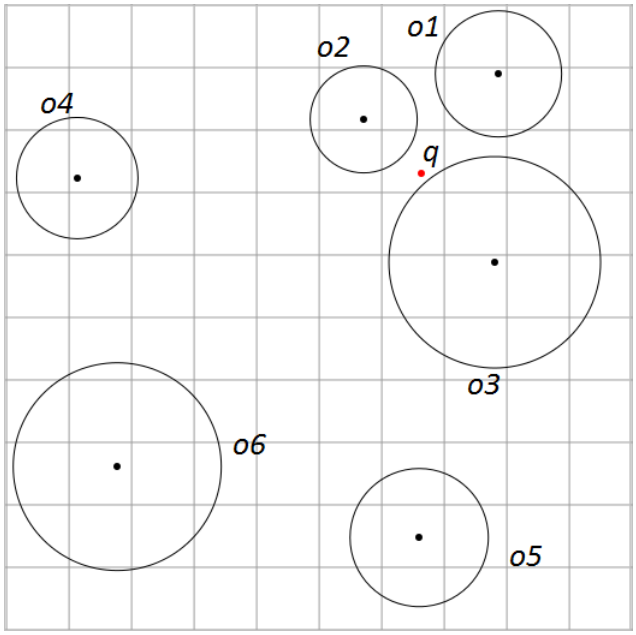
Στο σημείο αυτό γίνεται αναφορά των κύριων δομών δεδομένων που χρησιμοποιούνται στον αλγόριθμο:

- *ObjList*: Για κάθε κελί του καννάβου, αυτή η λίστα κρατάει τα αντικείμενα που έχουν αντιστοιχηθεί κατά το παρόν χρονόσημο.

- *cList*: Λίστα που περιέχει τα προς εξέταση κελία για το εκάστοτε ερώτημα. Τα αντικείμενα εντός κάθε κελιού ελέγχονται με τυχαία σειρά.

- Q: Ουρά προτεραιότητας μεγέθους k, η οποία αποθηκεύει τα υποψήφια αντικείμενα (*qualifying objects*) που στο παρόν χρονόσημο περιέχονται πλήρως στον κύκλο με κέντρο τη σημειακή εστία q του ερωτήματος και ακτίνα ίση με την μέχρι στιγμής k-οστή $MAXDIST$ (όπως αυτή έχει οριστεί προηγουμένως). Τα αντικείμενα τοποθετούνται ως προς φθίνουσα σειρά των $MAXDIST$ τους από την εστία q.

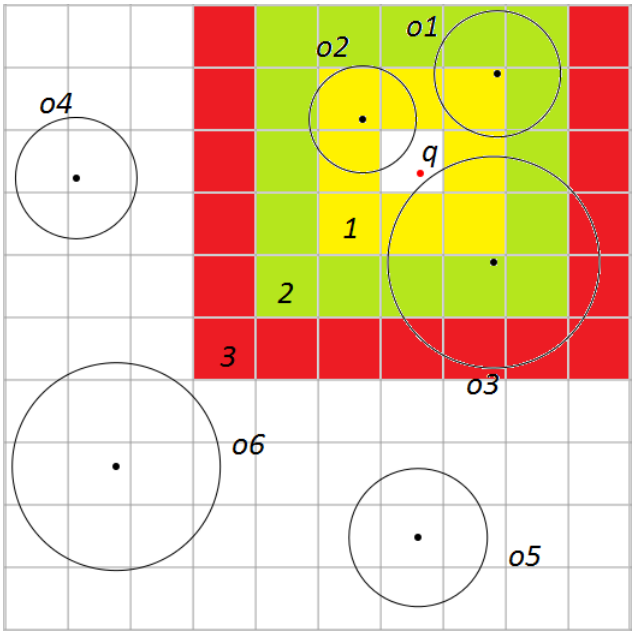
- P: Ουρά προτεραιότητας μεγέθους k, η οποία αποθηκεύει επιλαχόντα κινούμενα αντικείμενα (*auxiliary objects*), των οποίων τα κέντρα αβεβαιότητας βρίσκονται, στο εξεταζόμενο χρονόσημο, εγγύτερα στην εστία του ερωτήματος. Τα αντικείμενα τοποθετούνται ως προς φθίνουσα σειρά των αποστάσεων των κέντρων τους από την εστία q.



Σχήμα 7.4: Παράδειγμα μεθόδου καννάβου για αβέβαια κινούμενα αντικείμενα

7.4.2 Ευρετήριο καννάβου και επάλληλες ζώνες κελιών

Η τεχνική της κατάτμησης σε κάνναβο χρησιμοποιείται ευρέως ως ευρετήριο για κινούμενα αντικείμενα σε χωρικά ρεύματα δεδομένων (π.χ. στα [22], [23]). Σκοπός του είναι η διαμέριση των αντικειμένων σε κελιά με βάση την θέση τους ανά χρονόσημο. Κάθε κελί κρατάει τα αντικείμενα εκείνα των οποίων τα κέντρα της περιοχής αβεβαιότητας περιέχονται σε αυτό. Δηλαδή κάθε αντικείμενο δεν γίνεται να αντιστοιχηθεί σε περισσότερα από ένα κελιά. Ομοίως, γίνεται η αντιστοίχιση των θέσεων των ερωτημάτων σε κάθε κελί, των οποίων η εστία είναι ένα ακριβές γεωγραφικό στίγμα. Η ανανέωση του καννάβου γίνεται περιοδικά ανά κύκλο εκτέλεσης. Το βασικό πλεονέκτημα της τεχνικής αυτής είναι ότι τα αντικείμενα που βρίσκονται εγγύτερα μεταξύ τους συγκεντρώνονται σε γειτονικά κελιά. Επομένως δεν χρειάζεται να εξετάζουμε όλα τα αντικείμενα που κινούνται στον χάρτη όταν αποτιμάται ένα ερώτημα, παρά μόνο εκείνα που βρίσκονται πιο κοντά σε αυτό. Το κόστος δημιουργίας του καννάβου είναι αμελητέο, ειδικά εάν αναλογιστούμε πως χωρίς αυτό η αποτίμηση των ερωτημάτων θα ήταν αρκετά χρονοβόρα. Ο κάνναβος, το οποίο θα συμβολίζουμε με G, έχει συγκεκριμένες διαστάσεις (width, height) και τεμαχίζεται σε c $\times$ c τετραγωνικές περιοχές (κελιά). Οι διαστάσεις αυτές καθώς και ο αριθμός των τετραγώνων που το χωρίζουν δίνονται ως παράμετροι και ο αριθμός τους επηρεάζει την απόδοση και το κόστος επεξεργασίας του αλγορίθμου, όπως θα φανεί αργότερα στην πειραματική μελέτη. Για παράδειγμα, ένα αρκετά μικρό πλήθος κελιών για μεγάλο αριθμό αντικειμένων στον χάρτη έχει μειωμένες αποδόσεις κάνοντάς το πιο αργό, αλλά έχει μικρότερο κόστος κατασκευής. Αντίστοιχα, ένα υπερβολικά μεγάλο πλήθος κελιών για τον ίδιο (μεγάλο) αριθμό αντικειμένων, επιβαρύνει το κόστος επεξεργασίας δεδομένων, έχοντας και μεγαλύτερο κόστος κατασκευής. Οπότε, πρέπει να επιλεχθεί μία ενδιάμεση τιμή για τον αριθμό των κελιών στην οποία πρέπει να τεμαχιστεί ο κάνναβος. Στο σχήμα 7.4 φαίνεται



Σχήμα 7.5: Επάλληλες ζώνες κελιών γύρω από μία σημειακή εστία ερωτήματος

η χρησιμότητα του καννάβου, καθώς το ερώτημα που τίθεται εξετάζει μόνο τα αντικείμενα που βρίσκονται σε κελιά που βρίσκονται εγγύτερα στο κελί της εστίας. Όλα τα υπόλοιπα αντικείμενα δεν χρειάζεται να εξεταστούν.

Εκτός από την τεχνική του ευρετηρίου, μία ακόμη τεχνική αφορά την διάκριση των κελιών σε επάλληλες ζώνες γύρω από το κελί του ερωτήματος [15]. Ουσιαστικά, αυτή δίνει την σειρά με την οποία πρέπει να εξεταστούν τα αντικείμενα που βρίσκονται σε κελιά γύρω από την εστία. Η φιλοσοφία των επιπέδων είναι η εξής: Αρχικά, εξετάζονται τα αντικείμενα των οποίων τα κέντρα των περιοχών αβεβαιότητας βρίσκονται στο ίδιο κελί με αυτό του ερωτήματος (ζώνη 0, με κίτρινο χρώμα στο σχήμα 7.5). Στη συνέχεια εξετάζονται αντικείμενα που είναι τοποθετημένα σε κελιά γύρω από εκείνο του ερωτήματος (ζώνη 1, με πράσινο χρώμα στο σχήμα 7.5). Ακολουθεί η εξέταση των κελιών που βρίσκονται γύρω από το προηγούμενη ζώνη κ.ο.κ. . Έτσι, η σειρά που εξετάζονται αντικείμενα δεν είναι αυθαίρετη, αλλά συγκεκριμένη γίνεται με τέτοιο τρόπο ώστε να μην γίνονται περιττοί έλεγχοι. Στο σχήμα 7.5 φαίνεται πως γίνεται ο διαχωρισμός σε επάλληλες ζώνες γύρω από ένα συγκεκριμένο ερώτημα πάνω στον κάνναβο.

7.4.3 Φάση Φιλτραρίσματος

Στο στάδιο φιλτραρίσματος γίνεται επιλογή των υποψηφίων k-εγγύτερων γειτόνων ανά ερώτημα. Ουσιαστικά, μετά το πέρας αυτού του σταδίου έχουν οριστικοποιηθεί τα υποψήφια αντικείμενα που είναι δυνατό να είναι μέρος της απάντησης του ερωτήματος. Η διαδικασία του φιλτραρίσματος ξεκινάει αμέσως μετά την τοποθέτηση των κέντρων των περιοχών αβεβαιοτήτων των αντικειμένων και των σημειακών εστιών των ερωτημάτων στον κάνναβο στο τρέχον χρονόσημο. Αναλυτικά, για κάθε ερώτημα ισχύει:

Σε πρώτη φάση, αρχικοποιούνται οι ουρές Q (υποψηφίων αντικειμένων) και P (επιλαχόντων αντικειμένων), εντοπίζεται το κελί που είναι τοποθετημένο το ερώτημα και εισάγεται στη λίστα $cList$ των υπό εξέταση κελιών. Στη συνέχεια, γίνεται έλεγχος των αντικειμένων στο κελί αυτό, ανανεώνοντας τις ουρές Q και P, και γίνεται ενημέρωση της k-οστής μέγιστης απόστασης ($kMAXDIST$) για το ερώτημα, στο σύνολο των αντικειμένων που έχουν εξεταστεί μέχρι στιγμής. Ακολουθεί η εισαγωγή των προς εξέταση κελιών της επόμενης ζώνης στη $cList$. Αφού εξεταστούν τα αντικείμενα αυτών των κελιών, όπως προηγουμένως, η διερεύνηση συνεχίζεται στην επόμενη επάλληλη ζώνη κ.ο.κ, έως ότου η $kMAXDIST$ που έχει βρεθεί είναι μικρότερη ή ίση από την ελάχιστη απόσταση μεταξύ της εστίας του ερωτήματος και καθενός από τα κελιά της επόμενης ζώνης (*κριτήριο τερματισμού*). Ο λόγος για τον οποίο το κριτήριο τερματισμού της διερεύνησης ισχύει, είναι πως οποιοδήποτε αντικείμενο πέρα από από την τρέχουσα ζώνη θα έχει $MAXDIST$ μεγαλύτερη από την $kMAXDIST$ που έχει βρεθεί και άρα η εξέτασή του είναι περιττή.

Συγκεκριμένα, η εξέταση των αντικειμένων σε κάθε κελί και η πιθανή εισαγωγή τους στις ουρές γίνεται ως εξής: Για κάθε αντικείμενο που είναι τοποθετημένο στο κελί, αρχικά υπολογίζουμε την $MAXDIST$ του από τη σημειακή εστία του ερωτήματος. Τα πρώτα k εξεταζόμενα αντικείμενα τοποθετούνται στις ουρές Q και P χωρίς να εφαρμόζεται κάποιο κριτήριο για την εισαγωγή τουσ(γραμμές 4-8 στον Αλγόριθμο 1). Επομένως, οι δύο ουρές αρχικοποιούνται στα ίδια αντικείμενα. Στη συνέχεια, για κάθε επόμενο αντικείμενο αρχικά γίνεται έλεγχος εάν πρέπει να τοποθετηθεί στην ουρά Q. Αυτό συμβαίνει όταν η $MAXDIST$ του είναι μικρότερη από την τρέχουσα $kMAXDIST$. Τότε εξάγεται το αντικείμενο που βρίσκεται στην κορυφή της ουράς (αφού η περιφέρεια της περιοχής αβεβαιότητας βρίσκεται μακρύτερα από το q) και τοποθετείται το εξεταζόμενο. Επίσης, ανανεώνεται η τιμή της $kMAXDIST$ και γίνεται έλεγχος εαν το εξαγόμενο αντικείμενο πρέπει να τοποθετηθεί στην ουρά P των εγγύτερων, ως προς το κέντρο της περιοχής αβεβαιότητας, αντικειμένων(γραμμές 10-14 στον Αλγόριθμο 1). Ουσιαστικά, εισαγωγή ενός νέου αντικειμένου στην Q σημαίνει πως η $kMAXDIST$ πρέπει να μειωθεί και έτσι μικραίνει το εύρος των περιοχών αναζητήσεων εγγύτερων γειτόνων στην φάση της εκλέπτυνσης. Εάν το αντικείμενο απορριφθεί στον αρχικό έλεγχο για εισαγωγή στην Q, τότε εξετάζεται η ένταξή του στην P (γραμμές 15,16 στον Αλγόριθμο 1). Στην περίπτωση που δεν ικανοποιεί ούτε αυτά τα κριτήρια, σημαίνει πως δεν είναι υποψήφιος εγγύτερος γείτονας και ακολουθεί η εξέταση του επόμενου αντικειμένου. Ο ψευδοκώδικας του ελέγχου αντικειμένων μέσα σε συγκεκριμένο κελί, παρατίθεται παρακάτω ως αλγόριθμος 1.

Ο έλεγχος για την πιθανή εισαγωγή αντικειμένου στην P (γραμμή 2 στον Αλγόριθμο 2) βασίζεται σε 2 κριτήρια:

- Εξετάζεται η Ευκλείδεια απόσταση του κέντρου της περιοχής αβεβαιότητας περιοχής από την εστία του ερωτήματος, και αν αυτή είναι μικρότερη ή ίση από την αντίστοιχη απόσταση του αντικειμένου που βρίσκεται στην κορυφή της ουράς P. Το κριτήριο αυτό διασφαλίζει ότι στην ουρά P υπάρχουν αντικείμενα των οποίων τα κέντρα των περιοχών αβεβαιότητας είναι εγγύτερα στο ερώτημα και άρα πιο πιθανό να είναι μέρος της απάντησης.

Algorithm 1 Probabilistic $k\theta NN$ Monitoring

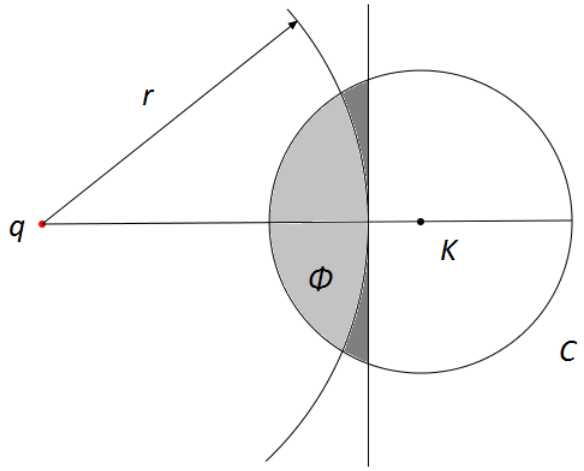
```
1: Procedure ProbeCellObjects (cell c, list of qualifying objects L, list of auxiliary objects P, focal query
   point q, integer k, threshold θ)
   //Q: list of k objects currently fully included in circle O(q, kMAXDIST)
   //P: list of k objects with their mean currently closest to q
2: for each object o ∈ c.ObjList do
3:    D ← MAXDIST(q, o);   //MAXDIST between q and uncertainty region of o
4:    if P.size() < k then
5:       P.push(o) ;     //List P contained less than k objects
6:    end if
7:    if  Q.size() < k then
8:       Q.push(o) ;                          //List Q contained less than k objects
9:       kMAXDIST ← MAXDIST(q, Q.top());      //Top item is farthest from q
10:   else if kMAXDIST > D then
11:      o' ← Q.pop();    //Remove the most extreme object o' currently in Q...
12:      Q.push(o);       //...and replace it with the current candidate object, and ...
13:      kMAXDIST ← MAXDIST(q, Q.top());      //...shrink kMAXDIST
         //Extreme object got evicted from Q, but could qualify for P
14:      VerifyCandidate(q, θ, o', P, kMAXDIST) ;
15:   else
16:      VerifyCandidate(q, θ, o, P, kMAXDIST) ; //Can it be an auxiliary object?
17:   end if
18: end for
19: End Procedure
```

- Εξετάζεται εάν η πιθανοτική κάλυψη της κυκλικής περιοχής του αντικειμένου είναι μεγαλύτερη ή ίση του κατωφλίου θ που έχει θέσει το ερώτημα. Ο υπολογισμός της πιθανοτικής κάλυψης δεν γίνεται αναλυτικά, όπως την ορίσαμε στην εξίσωση 7.4, αλλά μέσω μιάς προσεγγιστικής συνάρτησης Φ, η οποία χρησιμοποιεί την αθροιστική συνάρτηση κατανομής (*CDF, Cumulative Distribution Function*) κατά μήκος της ευθείας που ενώνει την εστία του ερωτήματος και το κέντρο αβεβαιότητας του αντικειμένου. Ουσιαστικά, η συνάρτηση αυτή είναι μία υπερεκτίμηση της ακριβούς τιμής της πιθανοτικής κάλυψης και αυτό φαίνεται στο σχήμα 7.6. Το σφάλμα αυτό στην εκτίμηση της πιθανοτικής κάλυψης, όμως, παρατηρήθηκε πειραματικά πως στην πλειονότητα των περιπτώσεων είναι ελάχιστο και δεν επηρρεάζει τα τελικά αποτελέσματα. Η συνάρτηση Φ υπολογίζεται σε απόσταση ίση με την τρέχουσα $kMAXDIST$ από την εστία του ερωτήματος. Αυτό το πιθανοτικό κριτήριο εξασφαλίζει πως στην ουρά P τοποθετούνται αντικείμενα που η πιθανοτική τους κάλυψη πιθανώς ικανοποιεί το κατώφλι θ για κάποιες περιοχές αναζήτησης (*search regions*) στην φάση της εκλέπτυνσης.

Ο ψευδοκώδικας για τον έλεγχο εισαγωγής αντικειμένων στην P φαίνεται παρακάτω και αναφέρεται ως αλγόριθμος 2.

Συμπερασματικά, ο τρόπος με τον οποίο γίνεται η εξέταση των αντικειμένων στα κελιά εξασφαλίζει πως μετά το πέρας της διαδικασίας του φιλτραρίσματος:

- Η ουρά Q θα περιέχει k αντικείμενα, των οποίων οι περιοχές αβεβαιότητας βρίσκονται



Σχήμα 7.6: Παράδειγμα υπολογισμού πιθανοτικής κάλυψης με τη συνάρτηση Φ

εξ ολοκλήρου μέσα σε ακτίνα ίση με $kMAXDIST$. Η συγκεκριμένη τιμή έχει οριστικοποιηθεί και ισούται με την k-οστή μεγαλύτερη Ευκλείδεια απόσταση του ερωτήματος από όλες τις περιφέρειες περιοχών αβεβαιότητας των αντικειμένων.

- Η ουρά P θα περιέχει k αντικείμενα των οποίων τα κέντρα αβεβαιότητας είναι πλησιέστερα στην εστία του ερωτήματος και πιθανώς ικανοποιούν το κατώφλι θ.

Algorithm 2 Probabilistic $k\theta NN$ Monitoring

1: **Procedure** *VerifyCandidate* (focal query point q, threshold θ, object o, list of auxiliary objects P, distance $kMAXDIST$)
2: **if** $\Phi(o, kMAXDIST) \geq \theta$ **and** $L_2(q, o) \leq L_2(q, P.\text{top}())$ **then**
3: P.pop(); *//Replace the most extreme element in P, since candidate o ...*
4: P.push(o); *//... has enough probability and has its mean closer to focal q*
5: **end if**
6: **End Procedure**

Η εισαγωγή κελιών στη λίστα $cList$ βασίζεται στην τεχνική των επάλληλων ζωνών που αναλύθηκε προηγουμένως. Έτσι, τοποθετούνται τα κελιά σε κάθε προσανατολισμό για την επόμενη ζώνη (γραμμές 3,4 στον Αλγόριθμο 3), αλλά μόνο αυτά για τα οποία ισχύει πως η $MINDIST$ του ερωτήματος από αυτά είναι μικρότερη ή ίση της τρέχουσας $kMAXDIST$ (γραμμή 6 στον Αλγόριθμο 3), έτσι ώστε να μην εξετάζονται αντικείμενα που αποκλείεται να είναι μέρος της τελικής απάντησης. Ο ψευδοκώδικας της παραπάνω διαδικασίας, παρατίθεται στη συνέχεια ως αλγόριθμος 3.

7.4.4 Φάση Εκλέπτυνσης

Στο στάδιο της εκλέπτυνσης (γραμμές 30-42 στον Αλγόριθμο 4) γίνονται οι λεπτομερείς υπολογισμοί για την εξαγωγή του τελικού αποτελέσματος. Λόγω του μικρού μεγέθους και των

δύο ουρών (k αντικείμενα η κάθε μία), γίνεται κατανοητό πως το κόστος της όλης διαδικασίας θα είναι σχετικά αμελητέο ως προς το συνολικό κόστος επεξεργασίας των δεδομένων ανά χρονόσημο. Αναλυτικά, για κάθε ερώτημα ακολουθείται η εξής διαδικασία:

Οι ακτίνες των περιοχών αναζήτησης είναι κάθε φορά ίσες με τις αντίστοιχες $MAXDIST$ των αντικειμένων της ουράς των υποψηφίων αντικειμένων Q, ξεκινώντας από την μικρότερη. Για κάθε τέτοια ακτίνα r, υπολογίζεται η τιμή της πιθανοτικής κάλυψής του για κάθε αντικείμενο των δύο ουρών (διότι μπορεί οι δύο ουρές να παρουσιάζουν κάποια επικάλυψη σε περιεχόμενα αντικείμενα). Ο υπολογισμός αυτός γίνεται και πάλι με βάση την προσεγγιστική συνάρτηση Φ. Αυτό επαναλαμβάνεται, έως ότου βρεθεί μία ακτίνα αναζήτησης για την οποία τουλάχιστον k αντικείμενα έχουν πιθανοτική κάλυψη μεγαλύτερη ή ίση του θ. Τότε σταματάει η διαδικασία της εκλέπτυνσης και εκδίδεται η απάντηση στο ερώτημα, περιλαμβάνοντας τα κορυφαία k αντικείμενα με τις μεγαλύτερες πιθανοτικές καλύψεις στην τελευταία περιοχή αναζήτησης.

Algorithm 3 Probabilistic $k\theta NN$ Monitoring

1: **Function** *FetchCells* (focal point q, query cell c, level lvl, distance $kMAXDIST$, integer $sizeQ$, integer k)
2: $cList \leftarrow \emptyset$; *//Container of cells that need be examined at the specified level*
//Check cells at the top (N), bottom (S), left (W), and right (E) of the query cell
3: **for each** orientation $h \in \{N, S, W, E\}$ **do**
4: $S \leftarrow \{s \in G$: grid cell s is one of the $2 \times lvl$ cells closest to q and is lvl cells away along direction h from query cell $c\}$;
5: **for each** cell $s \in S$ **do**
6: **if** $kMAXDIST \geq MINDIST(q, s)$ **or** $sizeQ < k$ **then**
7: $cList \leftarrow cList \cup \{s\}$; *//Include cell if list Q has less than k objects*
8: **end if**
9: **end for**
10: **end for**
11: return $cList$; *//List of cells that will be checked for qualifying objects*
12: **End Function**

Ο συνολικός ψευδοκώδικας των δύο φάσεων (φιλτραρίσματος και εκλέπτυνσης) παρατίθεται στην επόμενη σελίδα.

7.5 Αποτίμηση της μεθόδου

Η μέθοδος που αναλύθηκε παραπάνω για την αποτίμηση πιθανοτικών ερωτημάτων k-εγγύτερων γειτόνων παρουσιάζει τα εξής πλεονεκτήματα:

- Έλεγχος μόνο των αντικειμένων που βρίσκονται πλησιέστερα στις εκάστοτε εστίες των ερωτημάτων.
- Γρήγορος υπολογισμός της τιμής της πιθανοτικής κάλυψης, που επιτρέπει την ταχύτερη αποτίμηση μεγάλου πλήθους ερωτημάτων.
- Ποιοτική αξιοπιστία των αποτελεσμάτων.

Algorithm 4 Probabilistic $k\theta NN$ Monitoring

Input: Updates $\langle o_j, \mu_x^j, \mu_y^j, \sigma_j, \tau \rangle$ from $j = 1..N$ Bivariate Gaussian mobile users
Input: Specifications $\langle q_i, q_x^i, q_y^i, k_i, \theta_i, \tau \rangle$ from $i = 1..M$ continuous $k\theta$NN queries
Output: $R = \bigcup_i \{\langle q_i, R_i \rangle : R_i$ holds qualifying $k\theta$NN users per query $i = 1..M\}$
//Indexing objects, i.e., mobile users into the grid
for each cell $c \in G$ **do**
  $c.ObjList \leftarrow \emptyset$; *//Clear list of objects previously hashed into each cell*
end for
for each object o_j **do**
  $c \leftarrow \text{hash}(G, \mu_x^j, \mu_y^j)$;
  $c.ObjList \leftarrow c.ObjList \cup \{o_j\}$; *//Allocate object into a grid cell w.r.t. its mean*
end for
//Evaluation of queries
for each focal query point q_i **do**
  $Q \leftarrow \emptyset$; $P \leftarrow \emptyset$; *//Initialize lists of qualifying and auxiliary objects*
  //Filtering phase
  $c_i \leftarrow \text{hash}(G, q_i)$; *//The cell where this focal query point is currently located*
  $cList \leftarrow \{c_i\}$; *//Initialize list of cells that need be searched for q_i*
  $lvl \leftarrow 0$; *//Cell level is the one where the focal point is*
  repeat
    for each cell $c \in cList$ **do**
      ProbeCellObjects$(c, Q, P, q_i, k_i, \theta_i)$; *//Search for candidate $k\theta NN$ in cell*
    end for
    if $Q \neq \emptyset$ **then**
      $kMAXDIST \leftarrow Q.top()$; *//Top element of Q has the k-th MAXDIST*
      $lvl++$; *//Next level in the grid around focal point q*
      $cList \leftarrow \text{FetchCells}(c, lvl, q_i, kMAXDIST)$; *//All cells at next level*
      $stop \leftarrow$ **true**; *//If all cells at this level are farther than $kMAXDIST$*
      for each cell $c \in cList$ **do**
        $stop \leftarrow (stop$ **and** $(kMAXDIST \leq MINDIST(q_i, c)))$; *//Pruning*
      end for
    end if
  until $(cList = \emptyset$ **or** $stop)$; *//No more cells to check or they can be safely pruned*
  //Refinement phase
  for each $o_j \in Q$ **do**
    $r \leftarrow MAXDIST(q_i, o_j)$; *//Candidates in ascending search radius from q_i*
    $R_i \leftarrow \emptyset$; *//Container of final results for query q_i at execution cycle τ*
    for each $o' \in \{Q \cup P\}$ **do**
      $\phi \leftarrow \Phi(o', r)$; *//Estimate probability that o' is within radius r from q_i*
      if $\phi \geq \theta_i$ **then**
        $R_i \leftarrow R_i \cup \{\langle o', \phi \rangle\}$; *//$o'$ has enough probability to be a $k\theta NN$ of q_i*
      end if
    end for
    if $|R_i| \geq k$ **then**
      break; *//R_i may contain $>k$ results for query q_i; suppress excessive items*
    end if
  end for
  Report $\langle q_i, R_i \rangle$ with top-k objects in R_i sorted by ϕ;
end for
End Procedure

Ως μειονεκτήματα μπορούν να λογιστούν τα εξής:

- Το σφάλμα που παρουσιάζει η προσέγγιση Φ που χρησιμοποιείται για την πιθανοτική κάλυψη και το οποίο δεν είναι σταθερό. Για την καλύτερη ακρίβεια των αποτελεσμάτων θα μπορούσε να χρησιμοποιηθεί η μέθοδος Monte Carlo στην φάση της εκλέπτυνσης, με μεγαλύτερη, όμως, επιβάρυνση του συστήματος.
- Η μη αξιοποίηση προηγούμενων αποτελεσμάτων, αφού γίνεται επαναϋπολογισμός απαντήσεων εκ νέου ανά χρονόσημο.

Κεφάλαιο 8

Πειραματική Αξιολόγηση

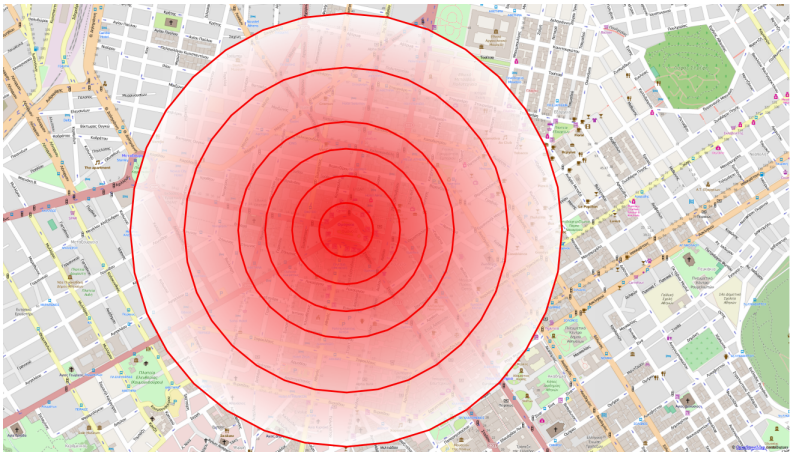
Στο κεφάλαιο αυτό παρουσιάζεται η πειραματική αξιολόγηση και ο έλεγχος σωστής λειτουργίας του αλγορίθμου. Γίνεται περιγραφή των χαρακτηριστικών των αρχείων εισόδου, των δεδομένων που χρησιμοποιήθηκαν και γίνεται παράθεση των συγκριτικών πειραμάτων που εκτελέστηκαν, με βάση τα οποία γίνεται αξιολόγηση των επιδόσεων.

8.1 Πειραματικό πλαίσιο

Στην ενότητα αυτή αξιολογείται πειραματικά ο αλγόριθμος που αναπτύχθηκε και υλοποιήθηκε στο κεφάλαιο 7. Όλες οι δομές υλοποιήθηκαν στη γλώσσα προγραμματισμού C++ και τα πειράματα εκτελέστηκαν σε λειτουργικό σύστημα Ubuntu Linux σε προσωπικό υπολογιστή Intel(R) Core(TM) i7, 2.8 GHz με μνήμη RAM 4 Gb.

8.1.1 Παραγωγή συνθετικών δεδομένων

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



Σχήμα 8.1: Ενδεικτικές περιοχές κλιμακούμενης αβεβαιότητας στον χάρτη της Αθήνας

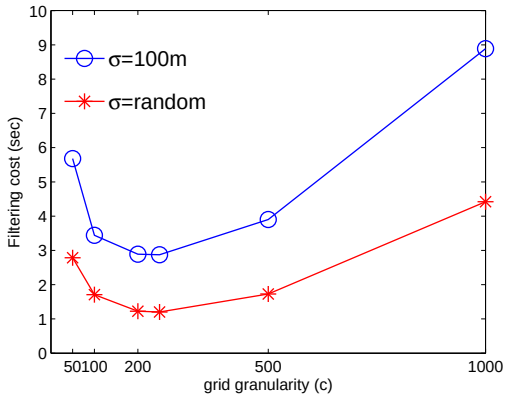
λαμβανοντας συνολικά 200 στίγματα ανά τροχιά με ίση χρονική απόσταση μεταξύ τους ώστε να αντιπροσωπεύουν τακτικές ενημερώσεις της θέσης των αντικειμένων. Τελικά, προέκυψε ένα αρχείο τροχιών με εγγραφές που φέρουν την ταυτότητα (ID) του αντικειμένου, τις συντεταγμένες (x, y) και το χρονόσημο (t).

Επιπρόσθετα, παρήχθησαν συνολικά 10000 εστίες ερωτημάτων για την ίδια περιοχή. Για κάθε ερώτημα υπάρχει αρχική καταγραφή για t=0. Για κάθε επόμενο χρονόσημο, περί το 1% των τρεχουσών εστιών επιλέγεται τυχαία και οι αντίστοιχες εστίες 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%.

8.1.2 Πειραματικά δεδομένα

Για τον αλγόριθμο αποτίμησης πιθανοτικών ερωτημάτων k-εγγύτερων γειτόνων σε αβέβαια κινούμενα αντικείμενα, $k\theta NN$, χρησιμοποιήθηκε ένα πειραματικό σύνολο δεδομένων αποτελούμενο από 100000 κινούμενα αντικείμενα και 10000 κινούμενα ερωτήματα. Το σύνολο αυτό περιέχει σημειακές θέσεις των αντικειμένων, δηλαδή των κέντρων των περιοχών αβεβαιοτήτων τους, και των εστιών των ερωτημάτων για 200 χρονόσημα. Κάθε εγγραφή του συνόλου των αντικειμένων έχει τη μορφή $< t, id, x, y >$, όπου t είναι το χρονόσημο, id η ταυτότητα του αντικειμένου και x, y οι συντεταγμένες του κέντρου της περιοχής αβεβαιότητας σύμφωνα με την διδιάστατη κανονική κατανομή. Η τυπική απόκλιση των κατανομών των κύκλων δίνεται είτε ίδια για όλα τα αντικείμενα, είτε επιλέγεται τυχαία για κάθε κινητό από ένα προκαθορισμένο σύνολο τιμών. Κάθε εγγραφή του συνόλου των ερωτηματων έχει την μορφή $< t, id, x, y >$, όπου t είναι το χρονόσημο, id η ταυτότητα του ερωτήματος και x, y οι συντεταγμένες της σημειακής εστίας του. Έγιναν πειράματα για τις εξής παραμέτρους του αλγορίθμου:

- Πλήθος κελιών $c \times c$ καννάβου.



Σχήμα 8.2: Κλιμάκωση χρόνου εκτέλεσης για διάφορες υποδιαιρέσεις του καννάβου

- Τυπική απόκλιση σ κανονικής κατανομής των κυκλικών περιοχών αβεβαιότητας.
- Αριθμός k-εγγύτερων γειτόνων που ζητείται.
- Πιθανοτικό κατώφλι θ για την εγκυρότητα απαντήσεων.

Πιο συγκεκριμένα, οι τιμές που δόθηκαν για κάθε παράμετρο ήταν (με έντονους χαρακτήρες (**bold**) φαίνονται οι τυπικές τιμές που χρησιμοποιήθηκαν στα πειράματα):

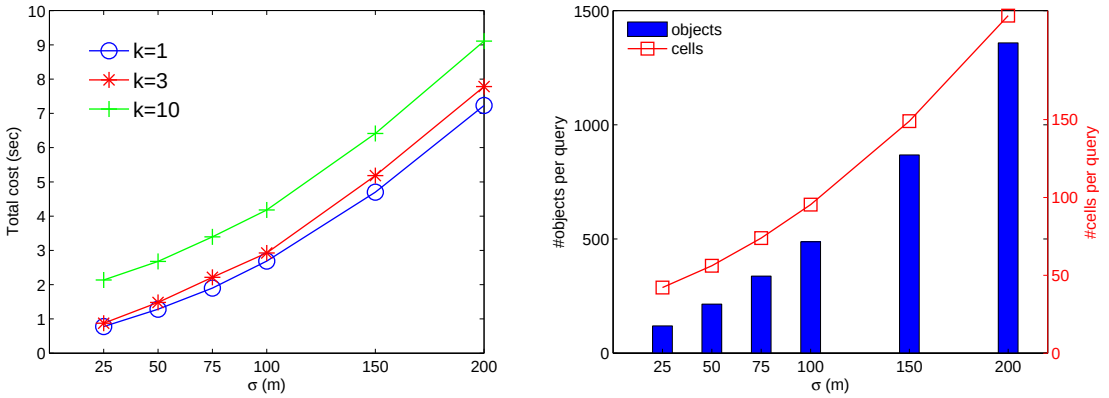
Πλήθος κελιών καννάβου $c \times c$	50×50, 100×100, 200×200, $\mathbf{250 \times 250}$, 500×500, 1000×1000
Τυπική απόκλιση σ	25m, 50m, 75m, **100m**, 150m, 200m
Αριθμός εγγύτερων γειτόνων k	1, 2, **3**, 4, 5, 10, 20
Πιθανοτικό κατώφλι θ	50%, 60%, 70%, **75%**, 80%, 90%, 99%

Πίνακας 8.1: Παράμετροι πειραμάτων

8.2 Αξιολόγηση αποτελεσμάτων

Κατά τη διάρκεια των πειραμάτων μετρήθηκαν τα εξής:

- Ο χρόνος εκτέλεσης του αλγορίθμου για όλα τα αντικείμενα και τα ερωτήματα, για τις φάσεις φιλτραρίσματος και εκλέπτυνσης ξεχωριστά, ανά χρονόσημο.
- Το πλήθος των κελιών και των αντικειμένων που εξετάστηκαν ανά ερώτημα και ανά χρονόσημο.
- Ο αριθμός των αντικειμένων που εισήχθησαν συνολικά στις ουρές Q και P ανά ερώτημα και ανά χρονόσημο.



(α′) Κλιμάκωση χρόνου εκτέλεσης (β′) Μέτρηση κελιών/αντικειμένων

Σχήμα 8.3: Επίδραση του βαθμού αβεβαιότητας των αντικειμένων

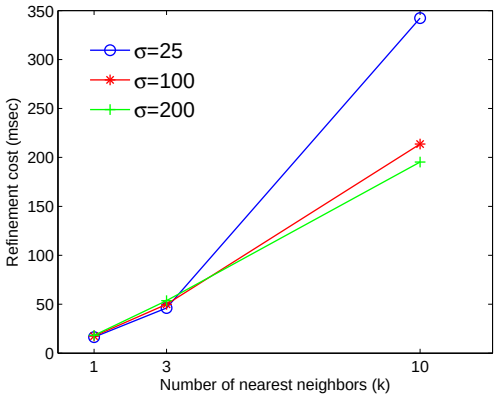
8.2.1 Διαστασιολόγηση καννάβου

Η επιλογή του αριθμού κελιών του καννάβου έγινε με βάση το παρακάτω πείραμα: Για σταθερή τυπική απόκλιση $\sigma = 100$μ, αλλά και για τυχαία ομοιόμορφη κατανομή της στα αντικείμενα από το σύνολο {25μ, 50μ, 75μ, 100μ, 150μ, 200μ} μετρήθηκε ο χρόνος εκτέλεσης για πλήθος κελιών ανά διάσταση $c = 50, 100, 200, 250, 500, 1000$. Στη γραφική παράσταση του σχήματος 8.2 απεικονίζονται οι χρόνοι εκτέλεσης της φάσης φιλτραρίσματος ανά ερώτημα, σε κάθε χρονόσημο και για τις δύο διαφορετικές περιπτώσεις. Οι χρόνοι αυτοί παρατηρείται πως σχηματίζουν μία καμπύλη. Αρχικά μειώνονται μέχρι ενός σημείου (ελάχιστο) και στη συνέχεια αυξάνονται. Αυτό συμβαίνει διότι για μικρό πλήθος κελιών ο κάνναβος δεν είναι χρήσιμος ως ευρετήριο, αφού μεγάλος αριθμός αντικειμένων τοποθετούνται στο ίδιο κελί και έτσι ο αλγόριθμος έχει μεγάλο επεξεργαστικό κόστος. Αλλά και ο υπερβολικός κατακερματισμός σε κελιά οδηγεί σε αυξημένο διαχειριστικό κόστος, επειδή οι περιοχές αβεβαιοτήτων καλύπτουν μεγάλο αριθμό κελιών η κάθε μία, με αποτέλεσμα να καθυστερεί ο τερματισμός του αλγορίθμου και να προσπελαύνονται αρκετά περισσότερα αντικείμενα. Για τους λόγους αυτούς, μία μέση κατάτμηση σε $c = 250$ κελιά ανά διάσταση αποδεικνύεται προτιμότερη για την εξεταζόμενη περίπτωση. Η επιλογή αυτή της τιμής επιβεβαιώθηκε και πειραματικά, στις δύο διαφορετικές εκτελέσεις, αποδεικνύοντας πως είναι η καταλληλότερη για το σύνολο των περιπτώσεων. Οι επόμενες μετρήσεις που θα παρουσιαστούν στη συνέχεια εκτελέστηκαν για 250×250 κελιά.

8.2.2 Επίδραση του βαθμού αβεβαιότητας (κυμαινόμενο σ)

Κρατώντας σταθερές τις υπόλοιπες παραμέτρους, έγινε εκτέλεση του αλγορίθμου για κάθε τιμή της τυπικής απόκλισης σ, για όλα τα κινούμενα αντικείμενα, στο σύνολο {25μ, 50μ, 75μ, 100μ, 150μ, 200μ}. Το πείραμα αυτό έγινε για τρεις διαφορετικές τιμές του αριθμού k των εγγύτερων γειτόνων (1, 3, 10).

Στην γραφική παράσταση του σχήματος 8.3α′ παρατηρείται η κλιμάκωση του συνολικού



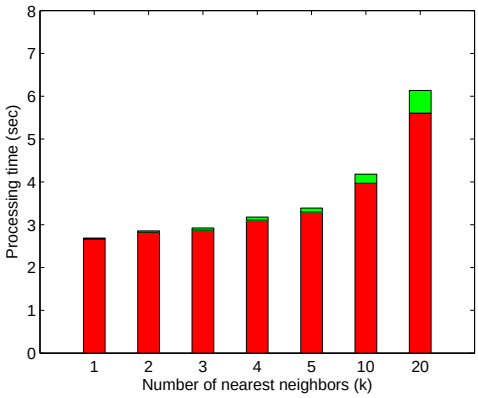
Σχήμα 8.4: Χρόνος εκτέλεσης φάσης εκλέπτυνσης για διαφορετικές τιμές σ/k

χρόνου εκτέλεσης του αλγορίθμου (φιλτράρισμα και εκλέπτυνση), ανα ερώτημα, για κάθε εκτέλεση του παραπάνω πειράματος. Όπως είναι φυσικό, ο χρόνος αυξάνεται όσο αυξάνει η τιμή της τυπικής απόκλισης. Το γεγονός αυτό οφείλεται στο μεγαλύτερο αριθμό κελιών και αντικειμένων που πρέπει να εξεταστούν ώστε να ικανοποιήσει το κριτήριο τερματισμού ο αλγόριθμος, το οποίο φαίνεται και στη γραφική παράσταση του σχήματος 8.3β′ (εκτέλεση μόνο για $k = 3$). Επίσης, οι χρόνοι είναι μεγαλύτεροι όσο ο αριθμός εγγύτερων γειτόνων αυξάνεται, χωρίς όμως να υπάρχει σημαντική επιβάρυνση του συστήματος. Η αιτιολόγηση της μικρής και όχι απότομης αύξησης για μεγαλύτερα k θα γίνει σε επόμενο πείραμα.

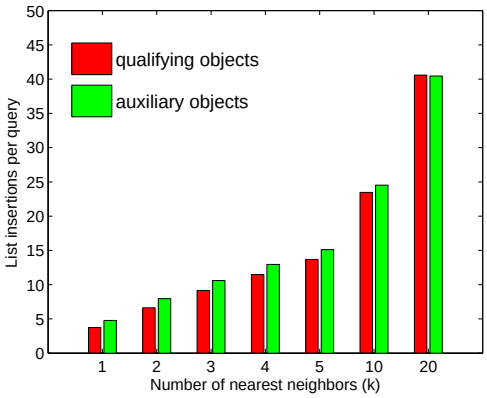
Ένα επιπλέον πείραμα που εκτελέστηκε ήταν το εξής: Για τρεις διαφορετικές τιμές του σ (25μ, 100μ, 200μ) και του k (1, 3, 10) και κρατώντας σταθερές τις υπόλοιπες παραμέτρους, εξετάστηκε ο χρόνος εκτέλεσης της φάσης εκλέπτυνσης. Στο σχήμα 8.4 παρατηρείται η γραφική παράσταση αυτού του χρόνου για τις διάφορες εκτελέσεις του πειράματος. Όσο αυξάνεται ο αναζητούμενος αριθμός των εγγύτερων γειτόνων, τόσο αυξάνεται ο χρόνος που διαρκεί η φάση εκλέπτυνσης, κάτι το οποίο είναι αναμενόμενο, αφού γίνεται αποτίμηση για μεγαλύτερο αριθμό αντικειμένων. Βέβαια, η επιβάρυνση αυτή είναι αμελητέα σε σχέση με το συνολικό χρόνο εκτέλεσης του αλγορίθμου (διάρκεια μερικών εκατοντάδων msec). Αυτό, όμως, που είναι άξιο σχολιασμού στο συγκεκριμένο πείραμα είναι πως, ενώ για $k = 1, 3$ ο χρόνος αυξάνεται (ελάχιστα) όσο μεγαλώνει η τιμή της τυπικής απόκλισης, για $k = 10$ παρατηρείται το αντίθετο. Η δικαιολόγηση αυτής της συμπεριφοράς βασίζεται στο γεγονός πως για μεγαλύτερα σ η επιθυμητή πιθανοτική κάλυψη θ μπορεί να επιτευχθεί για μικρότερες περιοχές αναζήτησης, καθώς οι περιοχές αβεβαιότητας είναι περισσότερο απλωμένες και καλύπτονται σε μεγαλύτερο βαθμό. Έτσι, όταν η τυπική απόκλιση λαμβάνει μικρές τιμές και το k είναι σχετικά μεγάλο θα χρειαστεί μεγαλύτερη ακτίνα αναζήτησης για να βρεθούν οι εγγύτεροι γείτονες με το επιθυμητό ποσοστό πιθανοτικής κάλυψης.

8.2.3 Επίδραση του αριθμού k των εγγύτερων γειτόνων

Στο πείραμα αυτό εξετάστηκε η απόδοση του αλγορίθμου όταν, με σταθερές τις υπόλοιπες παραμέτρους, έγινε εκτέλεση για όλες τις τιμές του k στο σύνολο $\{1, 2, 3, 4, 5, 10, 20\}$. Στο



(α′) Κλιμάκωση χρόνου εκτέλεσης



(β′) Αριθμός εισαγωγών αντικειμένων

Σχήμα 8.5: Μέγεθος ουρών για ποικίλες τιμές του k

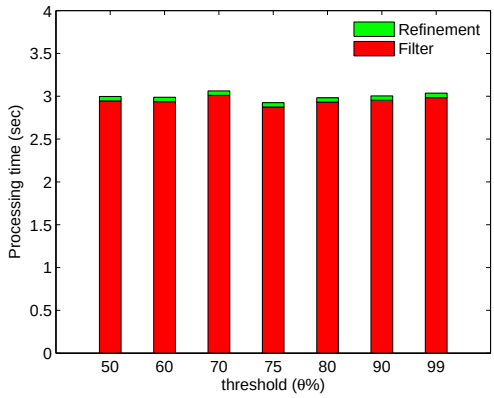
σχήμα 8.5α′ φαίνεται η διακύμανση του συνολικού χρόνου εκτέλεσης ανά ερώτημα, ανάλογα με τον επιθυμητό αριθμό των εγγύτερων γειτόνων. Ο χρόνος που διαρκεί η φάση εκλέπτυνσης είναι σχετικά αμελητέος σε αντιπαράθεση με τον χρόνο της φάσης φιλτραρίσματος, ο οποίος καλύπτει και το μεγαλύτερο κομμάτι της επεξεργασίας των δεδομένων. Αυτό είναι λογικό, καθώς στο φιλτράρισμα γίνεται έλεγχος αρκετά μεγαλύτερου αριθμού αντικειμένων, ενώ στην εκλέπτυνση εξετάζονται το πολύ $2k$. Επιπρόσθετα, η αύξηση του χρόνου εκτέλεσης δεν είναι απότομη, όταν ζητούνται περισσότεροι εγγύτεροι γείτονες, αφού δεν χρειάζεται να εξεταστεί αρκετά μεγαλύτερος αριθμός κινούμενων αντικειμένων στη φάση φιλτραρίσματος.

Στη γραφική παράσταση του σχήματος 8.5β′ παρατηρείται ο αριθμός εισαγωγών των αντικειμένων στις ουρές υποψηφίων και επιλαχόντων αντικειμένων (Q και P αντίστοιχα) , ανά ερώτημα, για κάθε τιμή του k. Το βασικό συμπέρασμα που εξάγεται, εκτός της φυσιολογικής αύξησης του αριθμού αυτού όσο αυξάνεται το k, είναι πως, τελικά, ο αριθμός των αντικειμένων που είναι σημαντικά για την εκτέλεση του αλγορίθμου είναι αρκετά μικρότερος από αυτόν των συνολικών αντικειμένων που εξετάζονται. Αυτό, βεβαίως, οφείλεται στο γεγονός πως κάθε φορά που εξετάζεται ένα κελί, ελέγχονται όλα τα αντικείμενά του ανεξαιρέτως. Ουσιαστικά, δεν υπάρχει κάποιο κριτήριο κλαδέματος για να μην ελέγχονται μη χρήσιμα αντικείμενα στο κελί και άρα ο αλγόριθμος τα προσπελαύνει αναγκαστικά.

8.2.4 Επίδραση πιθανοτικού κατωφλίου θ

Στο σημείο αυτό γίνεται σύγκριση των χρόνων εκτέλεσης του αλγορίθμου (φιλτράρισμα και εκλέπτυνση), ανά ερώτημα, όταν για σταθερές τιμές των άλλων παραμέτρων, αλλάζει το κατώφλι θ με πεδίο τιμών το σύνολο $\{50\%, 60\%, 70\%, 75\%, 80\%, 90\%, 99\%\}$. Στη γραφική παράσταση του σχήματος 8.6 παρατηρούνται οι εν λόγω χρόνοι εκτέλεσης.

Αντίθετα απ′ αυτό που θα ήταν αναμενόμενο, η αύξηση του κατωφλίου δεν επιφέρει επιβάρυνση στην εκτέλεση, αφού η τιμή του θ δεν έχει σημαντική επίδραση. Οι χρόνοι εκτέλεσης είναι παρόμοιοι και δεν παρουσιάζουν κάποια συγκεκριμένη συμπεριφορά. Η εξήγηση αυτού του φαινομένου έγκειται στο γεγονός πως οι έλεγχοι που γίνονται στον αλγόριθμο (ειδικά στη



Σχήμα 8.6: Κλιμάκωση χρόνου εκτέλεσης για διαφορετικές τιμές του θ

φάση φιλτραρίσματος καλύπτει και το μεγαλύτερο μέρος του χρόνου εκτέλεσης) με βάση το θ είναι ελάχιστοι, αφού κυρίως εξετάζονται γεωμετρικά χαρακτηριστικά των κινούμενων αντικειμένων σε σχέση με το εκάστοτε ερώτημα. Αυτό φάνηκε και από το προηγούμενο πείραμα όπου εξετάστηκαν οι εισαγωγές στις ουρές Q και P, των οποίων ο αριθμός ήταν ελάχιστος σε σχέση με το συνολικό εξεταζόμενο αριθμό αντικειμένων. Έτσι, το πιθανοτικό κατώφλι δεν φαίνεται να συνεισφέρει σημαντικά στην ελάττωση του κόστους επεξεργασίας των δεδομένων.

Κεφάλαιο 9

Συμπεράσματα και μελλοντικές επεκτάσεις

Οι πρόσφατες ραγδαίες εξελίξεις στις τεχνολογίες γεωγραφικού εντοπισμού αλλά και η μεγάλη δημοτικότητα των μέσων κοινωνικής δικτύωσης αύξησαν το ενδιαφέρον για την ανάπτυξη εφαρμογών παρακολούθησης κινούμενων αντικειμένων, γνωστές με το όνομα Υπηρεσίες Εντοπισμού (Location Services). Οι χρήστες τέτοιων εφαρμογών αποστέλλουν τη θέση τους περιοδικά και έχουν τη δυνατότητα να υποβάλλουν πολλαπλά ερωτήματα διαρκείας σε ένα κεντρικό επεξεργαστή, μεταξύ άλλων και ερωτήματα αναζήτησης k-εγγύτερων γειτόνων. Τα ερωτήματα αυτά εντοπίζουν τους k χρήστες που έχουν τη δυνατότητα να κινούνται και είναι πλησιέστερα σε κάποια τοποθεσία, η οποία μπορεί να ποικίλλει. Η τοποθεσία αυτή καθορίζεται από ένα γεωγραφικό στίγμα, ενώ τα κινούμενα αντικείμενα (δηλαδή οι χρήστες) διαθέτουν τη δυνατότητα γεωγραφικού εντοπισμού (GPS), όμως δεν επιθυμούν να αποκαλύπτουν την ακριβή θέση τους στον κεντρικό υπολογιστή. Ωστόσο, γνωστοποιείται μία ευρύτερη περιοχή, όπου η πιθανότητα να βρίσκεται το αντικείμενο σε κάθε θέση δεν είναι ομοιόμορφη, αλλά διαφοροποιείται. Οι χρήστες μπορούν να υποβάλλουν τα ερωτήματα διαρκείας για της τοποθεσίες ενδιαφέροντός τους, οπότε ο επεξεργαστής οφείλει να συνεκτιμήσει όλα τα δεδομένα και να παρέχει προσεγγιστικές απαντήσεις, οι οποίες όμως θα είναι αρκετά ικανοποιητικές ποιοτικά. Το δυναμικό αυτό μοντέλο τέτοιων συστημάτων επιχειρεί ένα διαφορετικό χειρισμό τους σε σχέση με τις συμβατικές βάσεις δεδομένων, θέτοντας στόχους όπως:

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

Σκοπός της παρούσας διπλωματικής εργασίας ήταν ο σχεδιασμός κατάλληλων δομών και η ανάπτυξη αλγορίθμου για την αποδοτική αποτίμηση πιθανοτικών ερωτημάτων εγγύτερων γειτόνων για αβέβαιες θέσεις κινούμενων αντικειμένων. Από τη μελέτη σχεδίασης του συστήματος, προκύπτουν τα εξής συμπεράσματα:

- Η επιλογή καννάβου ως χωρικού ευρετηρίου και η προσπέλαση των κελιών ανά επάλληλες ζώνες αποδείχτηκε ιδανική για την διαρκή παρακολούθηση των αβέβαιων αντικειμένων που είναι πλησιέστερα στα εκάστοτε ερωτήματα. Για μικρό πλήθος κελιών, ο κάνναβος χάνει τη χρησιμότητά του, αφού εξετάζεται μεγάλος αριθμός αντικειμένων σε κάθε κελί. Επίσης, ο υπερβολικός κατακερματισμός σε κελιά οδηγεί σε αυξημένο διαχειριστικό κόστος, διότι οι περιοχές αβεβαιοτήτων καλύπτουν μεγάλο αριθμό κελιών και έτσι καθυστερεί ο τερματισμός του αλγορίθμου. Μία μέση κατάτμηση $c = 250$ προκύπτει πως είναι κατάλληλη για την εξεταζόμενη περίπτωση.

- Η γνωστή στρατηγική *"φιλτράρισμα & εκλέπτυνση"* προσαρμόστηκε ώστε να αξιοποιεί κυρίως γεωμετρικά, αλλά και πιθανοτικά χαρακτηριστικά ων δεδομένων, προκειμένου να επιτυγχάνει αποδοτική αποτίμηση των ερωτημάτων. Με τον τρόπο αυτό, μικρός αριθμός αντικειμένων απαιτούν αναλυτική αποτίμηση.

- Οι χρονικές επιδόσεις του αλγόριθμου μετρήθηκαν πειραματικά και αποδείχθηκαν ιδιαιτέρως επαρκείς για τον χειρισμό πολλαπλών κινούμενων ερωτημάτων διαρκείας.

Από τη μελέτη του συγκεκριμένου προβλήματος, προκύπτουν ενθαρρυντικές προοπτικές επέκτασής του. Συγκεκριμένα, θα ήταν δυνατό να εξεταστεί η περίπτωση όπου οι περιοχές αβεβαιότητας των αντικειμένων ακολουθούν διαφορετικές πολυδιάστατες πιθανοτικές κατανομές, όπως οι χ^2, γ, Student κ.ά. . Τέλος, ένα τέτοιο ολοκληρωμένα σύστημα διαχείρισης ρευμάτων κινούμενων αντικειμένων $k\theta NN$ θα μπορούσε να αξιοποιηθεί σε πρόσθετους τύπους αναζήτησης. Για παράδειγμα, θα ήταν δυνατό να γίνει τροποποίησή του ώστε να αποτιμά ερωτήματα αντίστροφων πιθανοτικών k-εγγύτερων γειτόνων (*reverse k-NN*).

Βιβλιογραφία

[1] T. Bernecker, T. Emrich, H.-P. Kriegel, M. Renz, S. Zankl, and A. Z'ofle. Efficient Probabilistic Reverse Nearest Neighbor Query Processing on Uncertain Data. PVLDB, 4(10): 669-680, 2011.

[2] T. Bernecker, T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Z'ofle. A Novel Probabilistic Pruning Approach to Speed up Similarity Queries in Uncertain Databases. In Proceedings of the IEEE 27th International Conference on Data Engineering (ICDE), pp. 339-350, Hannover, Germany, 2011.

[3] C. Bohm, A. Pryakhin, and M. Schubert. Probabilistic Ranking Queries on Gaussians. In Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM'06), pp. 169-178, Vienna, Austria, July 2006.

[4] M.A. Cheema, X. Lin, W. Wang, W. Zhang, and J. Pei. Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data. IEEE Transactions on Knowledge and Data Engineering (TKDE), 22(4): 550-564, 2010.

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

[6] 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.

[7] A. Guttman. R-Trees: A Dynamic Index Structure for Spatial Searching. In Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data, pp. 47-57, Boston, Massachusetts, USA, June 1984.

[8] 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.

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

[10] L. Golab and M. Tamer Ozsu. Issues in Data Stream Management. ACM SIGMOD Record, 32(2):5-14, June 2003.

[11] G.R. Hjaltason and H. Samet. Distance Browsing in Spatial Databases. ACM Transactions on Database Systems, 24(2): 265-318, June 1999.

[12] 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.

[13] M. F. Mokbel, W. G. Aref, S. E. Hambrusch, and S.Prabhakar. Towards Scalable Location-aware Services: Requirements and Research Issues.In Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems (GIS'03), pp. 110-117,New Orleans,Louisiana, USA, November 2003.

[14] M.F. Mokbel, C. Chow, and W.G. Aref. The new Casper: Query Processing for Location Services without Compromising Privacy. In Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB), pp. 763-774, Seoul, Korea, September 2006.

[15] K. Mouratidis, M. Hadjieleftheriou, and D. Papadias. Conceptual Partitioning: An Efficient Method for Continuous Nearest Neighbor Monitoring. In Proceedings of the 24th ACM SIGMOD International Conference on Management of Data, pp. 634-645, Baltimore, Maryland, USA, June 2005.

[16] S. Papadopoulos, S. Bakiras, and D. Papadias. Nearest Neighbor Search with Strong Location Privacy. PVLDB, 3(1): 619-629, 2010.

[17] K. Patroumpas, M. Papamichalis, and T. Sellis. Probabilistic Range Monitoring of Streaming Uncertain Positions in GeoSocial Networks. In Proceedings of the 24th International Conference on Scientific and Statistical Data (SSDBM), pp. 20-37, Chania, Crete, Greece, June 2012.

[18] N. Roussopoulos, S. Kelley, and F. Vincent. Nearest Neighbor Queries. In Proceedings of the 1995 ACM International Conference on Management of Data (SIGMOD), pp. 71-79, San Jose, California, USA, June 1995.

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

[20] A.P. Sistla, O. Wolfson, and B. Xu. Continuous Nearest-Neighbor Queries with Location Uncertainty. VLDB Journal, 24(1):25-50, 2015.

[21] 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.

[22] X. Xiong, M.F. Mokbel, and W.G. Aref. SEA-CNN: Scalable Processing of Continuous K-Nearest Neighbor Queries in Spatio-temporal Databases. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05),pp. 643-654, Tokyo, Japan, April 2005.

[23] X. Yu, K. Q. Pu, and N. Koudas. Monitoring k-Nearest Neighbor Queries Over Moving Objects. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05),pp. 631-642, Tokyo, Japan, April 2005.

[24] P. Zhang, R. Cheng, N. Mamoulis, M. Renz, A. Zufle, Y. Tangs, and T. Emrich. Voronoi-based Nearest Neighbor Search for Multi-dimensional Uncertain Databases. In Proceedings of the 29th International Conference on Data Engineering (ICDE), pp. 158-169, Brisbane, Australia, April 2013.

[25] K. Zheng, P.C. Fung, and X. Zhou. k-Nearest Neighbor Search for Fuzzy Objects. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD), pp. 699-710, Indianapolis, Indiana, June 2010.

Γλωσσάριο

Ελληνικός όρος	Αγγλικός όρος
αβεβαιότητα	uncertainty
αθροιστική συνάρτηση κατανομής	cumulative distribution function
αποτίμηση ερωτημάτων	query evaluation
δειγματοληψία	sampling
δεικτοδότηση	indexing
εγγύτεροι γείτονες	nearest neighbors
εκλέπτυνση	refinement
επιλαχόντα αντικείμενα	auxiliary objects
επιλογή	selection
ζώνη	level
ερώτημα διαρκείας	continuous query
ερώτημα εγγύτερου γείτονα	nearest-neighbor query
ευρετήριο χωρικού καννάβου	uniform grid partitioning
ιδιωτικότητα	privacy
κάνναβος	grid
κανονική κατανομή	normal distribution
κινούμενο αντικείμενο	moving object
κλάδεμα	pruning
παράθυρο	window
περιοχή αναζήτησης	search region
πιθανοί κόσμοι	possible worlds
πολυπλεξία	multiplexing
ρεύμα δεδομένων	data stream
σημειακή εστία	focal point
συνάθροιση	aggregation
σύνδεση	join
συντελεστής συσχέτισης	correlation coefficient
υπηρεσίες εντοπισμού	location-based services
υποψήφια αντικείμενα	qualifying objects
φιλτράρισμα	filtering
χρονόσημο	timestamp

Αναζήτηση *k*-εγγύτερων γειτόνων μεταξύ περιοχών αβεβαιότητας με κανονική κατανομή

Χρήστος Κούτρας
koutras21@gmail.com

Διπλωματική εργασία στο Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων
*Επιβλέπων: Καθηγητής **Ι. Βασιλείου***

1 Γενικό πλαίσιο

Σκοπός της εργασίας είναι η μελέτη και η υλοποίηση ενός αλγορίθμου που θα επιτρέπει online απαντήσεις σε *πιθανοτικά ερωτήματα διαρκείας* (*probabilistic continuous queries*) σχετικά με την εύρεση *k-εγγύτερων γειτόνων* ανάμεσα σε κινούμενα αντικείμενα με αβέβαιες θέσεις. Τα εν λόγω πιθανοτικά ερωτήματα διαρκείας θα τίθενται από διάφορους κινούμενους χρήστες και θα αφορούν σημειακές εστίες με ακριβές στίγμα. Αυτές οι εστίες δεν συμπίπτουν απαραίτητα με την θέση του κινούμενου χρήστη, αλλά μπορεί να δηλώνουν ένα σημείο συνάντησης (λ.χ. καφέ, κινηματογράφος) ή κάποιο γνωστό τοπόσημο (landmark). Ο χρήστης επιθυμεί να ενημερώνεται για όσα αντικείμενα (δηλ. άλλους χρήστες των οποίων η θέση δεν είναι ακριβώς γνωστή) που είναι πιθανότερο να είναι πλησίον αυτής της εστίας ενδιαφέροντός του.

Εξαιτίας της *αβεβαιότητας* (*uncertainty*) ως προς την καταγραφή των ακριβών γεωγραφικών θέσεων των κινούμενων αντικειμένων, ο κεντρικός επεξεργαστής του συστήματος οφείλει να δίνει εγκαίρως προσεγγιστικές, αλλά σχετικά αξιόπιστες απαντήσεις στα εκάστοτε ερωτήματα. Τέτοια δεδομένα θα μπορούσαν κυρίως να αξιοποιηθούν σε εφαρμογές κοινωνικής δικτύωσης (π.χ. *Facebook*). Είναι προφανές, πως βασική υπόθεση της εργασίας είναι η παρακολούθηση πολλών κινούμενων αντικειμένων και ερωτημάτων τα οποία θα ανανεώνουν πολύ συχνά την περιοχή της θέσης τους και έτσι σχηματίζουν *ρεύματα δεδομένων* (*data streams*). Θα πρέπει να τονιστεί ότι:

- Όλα τα στοιχεία κίνησης καταφθάνουν στον κεντρικό επεξεργαστή σε μεγάλο και ενδεχομένως μεταβλητό ρυθμό σε πραγματικό χρόνο (*online*).
- Τα ρεύματα έχουν θεωρητικά απεριόριστο μέγεθος.
- Η περιοχή αβεβαιότητας κάθε χρήστη μπορεί να κυμαίνεται όσον αφορά το εύρος της, επιτρέποντας διαβαθμίσεις στο επίπεδο προστασίας της *ιδιωτικότητάς* του (*privacy protection*).

Είναι λογικό πως με βάση την πιθανοτική φύση του μοντέλου, καταλληλότερη επιλογή είναι η ανάπτυξη *προσεγγιστικών αλγορίθμων* που εξοικονομούν χρόνο και κόστος επεξεργασίας, συνεκτιμώντας πάντοτε το σφάλμα που μπορεί να γίνει αποδεκτό από την εκάστοτε εφαρμογή.

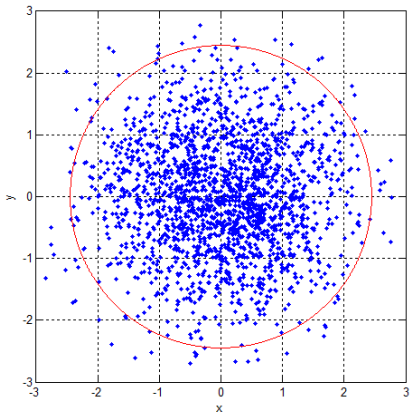
2 Αβεβαιότητα δεδομένων

Τα τελευταία χρόνια έχει εμφανιστεί ένα ευρύ φάσμα εφαρμογών που σχετίζονται με την *αβεβαιότητα*. Οι λόγοι για τους οποίους υπάρχει αβεβαιότητα στα δεδομένα διαφέρουν ανάλογα με την εφαρμογή. Για παράδειγμα, σε δίκτυα αισθητήρων η αβεβαιότητα οφείλεται σε σφάλματα κατά τις μετρήσεις (π.χ. υγρασία) που οδηγούν σε ανακρίβεια. Σε άλλες εφαρμογές (λ.χ. κοινωνικά δίκτυα), στις οποίες κυρίαρχο ρόλο έχει η προστασία της *ιδιωτικότητας*, τα δεδομένα είναι μη ακριβή διότι υπάρχει αυτή η απαίτηση εκ μέρους των ίδιων των χρηστών. Ουσιαστικά, η αβεβαιότητα συμβάλλει στην απόκρυψη προσωπικών δεδομένων, παρέχοντας προστασία σε ευαίσθητα χαρακτηριστικά των ατόμων (όπως φύλο, ηλικία, θρησκευτικές ή πολιτικές πεποιθήσεις κλπ.), με στόχο να γίνονται γνωστό λιγότερα στοιχεία. Ένα άλλο παράδειγμα που είναι άξιο να αναφερθεί είναι τα συστήματα γεωγραφικού εντοπισμού (GPS). Στην περίπτωση αυτή, το στίγμα ενός αντικειμένου (π.χ. κινητού τηλεφώνου) δίνει την ακριβή θέση του κατόχου και πιθανώς την ταχύτητα κίνησής του κάθε χρονική στιγμή. Είναι σαφές πως θα ήταν ασύμφορο για τον εντοπισμό του να παρέχει το στίγμα του αρκετά συχνά, αφού προσεγγιστικά θα μπορούσε να εντοπιστεί μέσω της ταχύτητάς του. Επομένως, η θέση του αντικειμένου όπως είναι γνωστή στο σύστημα, δεν ταυτίζεται πάντοτε με την τρέχουσα εξαιτίας της χρονικής υστέρησης κατά τη μετάδοση και άρα θεωρείται αβέβαιη.

Μέχρι πριν λίγα χρόνια, η αποδοτική επεξεργασία αβέβαιων δεδομένων δεν ήταν εφικτή, διότι οι συμβατικές βάσεις δεδομένων δεν ήταν προετοιμασμένες για ένα τέτοιο ενδεχόμενο. Σήμερα όμως, έχει αναπτυχθεί διάφορα μοντέλα και μία πληθώρα αλγορίθμων για την καλύτερη επεξεργασία ερωτημάτων που εμπεριέχουν πλέον την έννοια της αβεβαιότητας, μεγάλο αριθμό των οποίων συναντάμε αρκετά συχνά στις β.δ. Λόγω εγγενών δυσκολιών στην επεξεργασία τέτοιων ερωτημάτων, ο υπολογισμός των περισσότερων αλγορίθμων γίνεται με αριθμητικές

μεθόδους, με αποτέλεσμα οι λύσεις που προκύπτουν να είναι προσεγγιστικές.

Ανάλογα με το πρόβλημα ή την εφαρμογή, επιλέγεται η καταλληλότερη αναπαράσταση των αβέβαιων δεδομένων που συμβάλλει στην αποδοτική επεξεργασία τους. Στις σχετικές ερευνητικές εργασίες που



Σχήμα 1: *Κανονική κατανομή περιοχής αβεβαιότητας*

υπάρχουν στην πρόσφατη βιβλιογραφία ακολουθούνται τα εξής μοντέλα αναπαράστασης:

- Το *συνεχές μοντέλο αναπαράστασης*, όπου τα δεδομένα ακολουθούν κάποια συνεχή στατιστική κατανομή (ομοιόμορφη, Gaussian, κ.ά.).
- Το *διακριτό μοντέλο*, κάνοντας χρήση διακριτών δειγμάτων για τον υπολογισμό των πιθανοτήτων.

Επίσης, διάφορες ερευνητικές προσεγγίσεις επικεντρώνονται στην καλύτερη παρουσίαση των αποτελεσμάτων (π.χ. περιθώρια εμπιστοσύνης, εκτίμηση σφάλματος κτλ.) με σκοπό να παρέχεται στους χρήστες μία εποπτική εικόνα των απαντήσεων, δηλ. εάν και κατά πόσο μπορούν να βασιστούν σ' αυτές ή όχι.

Είναι αρκετά σημαντικό να τονιστεί η διαφορά μεταξύ των *πιθανοτικών* βάσεων δεδομένων και των βάσεων *δεδομένων με αβεβαιότητα* (αντίστοιχα για ρεύματα δεδομένων). Συγκεκριμένα, οι πιθανοτικές β.δ. ασχολούνται με την *ύπαρξη* ή μη αντικειμένων που είναι ακριβή. Απεναντίας, η αβεβαιότητα αναφέρεται στην ύπαρξη οντοτήτων, των οποίων η *κατάσταση* είναι μη ακριβής. Στην παρούσα διπλωματική εργασία ασχολούμαστε με ρεύματα δεδομένων που υπόκεινται σε αβεβαιότητα όσον αφορά την χωρική τους θέση.

3 Μοντέλο συστήματος

Για την ευκολότερη επίλυση του εξειδικευμένου προβλήματος εγγύτερων γειτόνων που καλείται να αντιμετωπίσει η εργασία, ακολουθείται μία συγκεκριμένη μοντελοποίηση των αντικειμένων και των ερωτημάτων.

Κάθε κινούμενο αντικείμενο αναπαρίσταται με μία *περιοχή αβεβαιότητας* (*uncertainty region*), η οποία ακολουθεί την κανονική κατανομή δύο μεταβλητών (*Bivariate Gaussian distribution*). Όπως φαίνεται στο Σχήμα 1, περίπου το 99.73% της συνολικής πιθανότητας περικλείεται σε έναν *κύκλο* με ακτίνα ίση με το τριπλάσιο της τυπικής απόκλισης *σ* γύρω από το κέντρο της κατανομής. Συνεπώς, για τους σκοπούς της επεξεργασίας, η περιοχή αβεβαιότητας κάθε αντικειμένου πρακτικά ισοδυναμεί με τον κύκλο αυτόν.

Κάθε *κινούμενο ερώτημα* προσδιορίζει έναν αριθμό *k* (1, 2, 3 ή παραπάνω) εγγύτερων γειτόνων που είναι επιθυμητό να βρεθούν γύρω από την (ενδεχομένως κινούμενη) σημειακή *εστία q*. Επιπλέον, το ερώτημα θέτει ένα *κατώφλι θ* (π.χ. 75%), το οποίο εκφράζει την μικρότερη ανεκτή πιθανότητα για να θεωρηθεί κάποιο αντικείμενο υποψήφιος εγγύτερος γείτονας.

Το γενικό μοντέλο της εφαρμογής προβλέπει την περιοδική μετάδοση των περιοχών αβεβαιότητας των αντικειμένων, καθώς και των εστιών ενδιαφέροντος των ερωτημάτων σε έναν κεντρικό επεξεργαστή. Κάθε τέτοια ανανέωση των στοιχείων αντιστοιχεί σε ένα κοινό χρονόσημο (*timestamp*) *t*. Σε κάθε τέτοιο χρονόσημο και για κάθε ερώτημα γίνεται εξαγωγή του αποτελέσματος το οποίο περιέχει τους *k* πιθανότερους εγγύτερους γείτονες με φθίνουσα σειρά πιθανότητας. Επομένως, τα ερωτήματα διαρκείας που τίθενται είναι ερωτήματα ορισμένου αριθμού εγγύτερων γειτόνων με βάση ένα πιθανοτικό κατώφλι (*kθNN*). Σημειώνεται ότι ο αλγόριθμος δεν παρέχει ακριβή, αλλά προσεγγιστικά αποτελέσματα. Ωστόσο, οι απαντήσεις θα είναι ποιοτικά αξιόπιστες και αντιπροσωπευτικές της πραγματικότητας.

4 Επεξεργασία ερωτημάτων *kθNN*

Με στόχο την μείωση του κόστους επεξεργασίας των δεδομένων, ο αλγόριθμος που αναπτύχθηκε για την αποτίμηση ερωτημάτων εγγύτερων γειτόνων με αβέβαιες θέσεις κινούμενων αντικειμένων χρησιμοποιεί:

- Χωρικό ευρετήριο βασισμένο σε κάνναβο (*grid partitioning*) για έλεγχο κελιών σε επάλληλες ζώνες.
- Τεχνικές κλαδέματος (*pruning*) για αποφυγή εξέτασης αντικειμένων άσχετων με το ερώτημα.

Σε κάθε χρονόσημο *t*, ο αλγόριθμος αντιστοιχεί τα κέντρα των περιοχών αβεβαιότητας των αντικειμένων και τις σημειακές εστίες των ερωτημάτων πάνω στα κελιά του καννάβου (Σχήμα 2). Ακολουθούν δύο φάσεις:

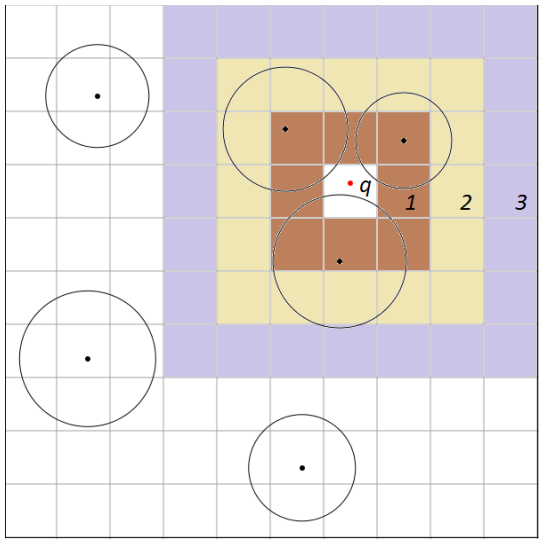
i. *Φιλτράρισμα υποψηφίων* (*filtering*): Τα αντικείμενα που εξετάζονται ανά ερώτημα φιλτράρονται βάσει

γεωμετρικών και πιθανοτικών χαρακτηριστικών των κινούμενων αντικειμένων και των ερωτημάτων.

ii. *Εκλέπτυνση τελικών απαντήσεων* (*refinement*): Αφού πολύ μικρός αριθμός από τα αρχικά δεδομένα χρειάζεται να επεξεργαστούν περαιτέρω, επιλέγονται τελικά οι *k*-εγγύτεροι γείτονες βάσει υπολογισμών πιθανοτικής κάλυψης γύρω από την εστία *q*.

4.1 Φιλτράρισμα υποψηφίων

Στο στάδιο αυτό και για κάθε ερώτημα, ξεκινάμε από την σημειακή εστία *q* και διερευνούμε επάλληλες ζώνες

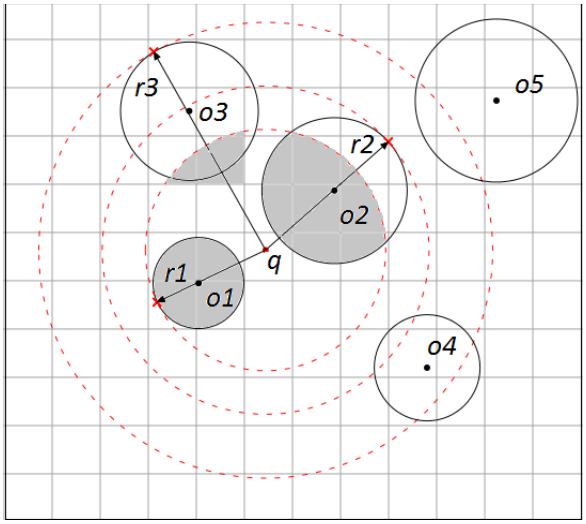


Σχήμα 2: *Έλεγχος υποψηφίων αντικειμένων με κάνναβο*

κελιών κυκλικά γύρω από το *q* (Σχήμα 2). Ουσιαστικά, με τον τρόπο αυτό εξετάζουμε τα αντικείμενα έτσι ώστε η σειρά να μην είναι αυθαίρετη αλλά να συμβάλλει σε αποφυγή ελέγχου περιττών αντικειμένων. Σε κάθε τέτοιο κελί, ελέγχονται λοιπόν όσα αντικείμενα έχουν κέντρα αβεβαιότητας εντός του κελιού και αποφασίζεται εάν κάποια από αυτά είναι *υποψήφιοι* εγγύτεροι γείτονες. Πρακτικά, η επιλογή υποψηφίων κατά το τρέχον χρονόσημο *t* βασίζεται σε γεωμετρικά χαρακτηριστικά τους (μέγιστη απόσταση περιοχής αβεβαιότητας, ελάχιστη απόσταση κελιού, Ευκλείδεια απόσταση κέντρου αβεβαιότητας) ως προς την εστία *q*. Η διαδικασία αυτή διακόπτεται όταν επιβεβαιωθεί ότι διερεύνηση σε περαιτέρω κελιά δεν είναι δυνατό να συνεισφέρει νέους υποψήφιους γείτονες πλησιέστερα προς την εστία *q*.

4.2 Εκλέπτυνση απαντήσεων

Στην φάση αυτή, επεκτεινόμαστε από την σημειακή εστία *q* του ερωτήματος σε κυκλικές περιοχές αναζήτησης (*search regions*), όπως φαίνεται με τις διακεκομμένες κόκκινες περιφέρειες στο Σχήμα 3. Για κάθε μία από αυτές, υπολογίζουμε το *ποσοστό πιθανοτικής κάλυψης* Φ για κάθε υποψήφιο κινούμενο αντικείμενο *o* (οι σκιασμένες περιοχές στο Σχήμα 3), από αυτά που έχουμε βρει στη διαδικασία του φιλτραρίσματος. Προς αποφυγή του μεγάλου κόστους πράξεων που θα επέφερε ένας ακριβής και αναλυτικός υπολογισμός του Φ, η εκτίμηση της τιμής του γίνεται προσεγγιστικά. Αυτό επιτυγχάνεται, χρησιμοποιώντας την *αθροιστική συνάρτηση κατανομής* (*CDF, cumulative distribution function*) κατά μήκος της ευθείας που ενώνει



Σχήμα 4: *Πιθανοτική κάλυψη υποψηφίων αντικειμένων σε διάφορες ακτίνες*

την εστία *q* με το κέντρο της εκάστοτε περιοχής αβεβαιότητας. Η κατανομή αυτή αφορά την *επιρροή* του αντικειμένου *o* στο *q*, αλλά μόνο κατά μήκος της εγκάρσιας τομής της στο ύψος της διαμέτρου της κυκλικής περιοχής αβεβαιότητας γύρω από το *o*. Αυτή η εκτιμώμενη επιρροή ακολουθεί την μονοδιάστατη κανονική κατανομή, έχοντας μέση τιμή που αντιστοιχεί στην απόσταση του *q* από το κέντρο της περιοχής αβεβαιότητας του *o* και με την ίδια τυπική απόκλιση *σ*. Εφαρμόζοντας αυτήν την προσεγγιστική τεχνική υπολογισμού πιθανοτικής κάλυψης για τους υποψήφιους εγγύτερους γείτονες σε κάθε περιοχή αναζήτησης, διακόπτουμε την διαδικασία όταν βρεθούν τουλάχιστον *k* αντικείμενα (ο ζητούμενος αριθμός εγγύτερων γειτόνων) με πιθανοτική κάλυψη μεγαλύτερη ή ίση του αντίστοιχου

κατωφλίου (δηλαδή $\Phi \geq \theta$). Τότε τυπώνουμε την κατάταξη των k-εγγύτερων γειτόνων για την συγκεκριμένη απόσταση από το q, μαζί με την αντίστοιχη εκτίμηση του Φ ανά γείτονα.

5 Πειραματική αξιολόγηση

Ο αλγόριθμος αποτίμησης πιθανοτικών ερωτημάτων k-εγγύτερων γειτόνων για αβέβαιες θέσεις κινούμενων αντικειμένων ($k\theta NN$), υλοποιήθηκε στη γλώσσα προγραμματισμού C++ και δοκιμάστηκε πειραματικά. Το συνθετικό 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 ευρύτερης περιοχής Αθηνών.

Μελετήθηκαν οι επιδόσεις σε χρόνο επεξεργασίας για τις εξής παραμέτρους:

- Πλήθος κελιών cxc στον κάνναβο: 50x50, 100x100, 200x200, 250x250, 500x500, 1000x1000.
- Τυπική απόκλιση σ κανονικής κατανομής της αβεβαιότητας: 25m, 50m, 75m, 100m, 150m, 200m. Επομένως, οι κυκλικές περιοχές αβεβαιότητας έχουν τριπλάσια ακτίνα (3σ).
- Αριθμός εγγύτερων γειτόνων k: 1, 2, 3, 4, 5, 10, 20.
- Κατώφλι θ επιλογής πιθανών απαντήσεων: 50%, 60%, 70%, 75%, 80%, 90%, 99%.

Γενικά παρατηρήθηκε είναι πως η κλιμάκωση των τιμών κάθε παραμέτρου επιβαρύνει τις επιδόσεις του συστήματος στις περισσότερες περιπτώσεις. Σημαντική επίδραση στις επιδόσεις έχει ο βαθμός κατάτμησης κατάτμηση c του καννάβου. Για μικρό (50x50) αλλά και για μεγάλο (1000x1000) πλήθος κελιών, ο χρόνος εκτέλεσης αυξάνεται. Για τον λόγο αυτό επιλέγεται c=250 ώστε το ευρετήριο να είναι όσο το δυνατόν αποτελεσματικότερο και να έχει μικρότερο κόστος συντήρησης. Επιπρόσθετα, παρατηρήθηκαν τα εξής:

- Η φάση φιλτραρίσματος επωμίζεται το μεγαλύτερο μέρος του κόστους ανά χρονόσημο. Αυτό συμβαίνει επειδή το φιλτράρισμα εφαρμόζεται επί όλων ανεξαιρέτως των αντικειμένων που έχουν κέντρα αβεβαιότητας εντός κάποιου κελιού. Απεναντίας, η φάση εκλέπτυνσης στοιχίζει πολύ λιγότερο, αφού χρειάζεται να ελέγξει το πολύ $2k$ αντικείμενα.
- Όσο μεγαλύτερο το πλήθος k των αναζητούμενων εγγύτερων γειτόνων, τόσο μεγαλύτερος προκύπτει ο χρόνος εκτέλεσης, αφού χρειάζεται να εξεταστούν περισσότερα κελιά και αντικείμενα.
- Για μεγαλύτερες τιμές στην τυπική απόκλιση σ αυξάνεται ο χρόνος εκτέλεσης, αφού η αβεβαιότητα των αντικειμένων εκτείνεται σε μεγαλύτερη έκταση και άρα ο αλγόριθμος τερματίζει αργότερα.
- Η τιμή του κατωφλίου θ φαίνεται να μην επηρεάζει σημαντικά τον χρόνο εκτέλεσης, αφού το σχετικό κριτήριο ελέγχου εφαρμόζεται λίγες φορές στην φάση φιλτραρίσματος που είναι δυσανάλογα βεβαρημένη συγκριτικά με την εκλέπτυνση.
- Σε όλες τις περιπτώσεις, η συνολική αποτίμηση όλων των ερωτημάτων ανά χρονόσημο ολοκληρώθηκε σε μερικά (<10) δευτερόλεπτα, επιβεβαιώνοντας την ανθεκτικότητα του προτεινόμενου αλγορίθμου.

6 Συμπεράσματα – Προοπτικές

Οι πρόσφατες ραγδαίες εξελίξεις στις τεχνολογίες γεωγραφικού εντοπισμού αλλά και η μεγάλη δημοτικότητα των μέσων κοινωνικής δικτύωσης, αύξησαν το ενδιαφέρον για την ανάπτυξη εφαρμογών παρακολούθησης κινούμενων αντικειμένων. Σκοπός της παρούσας διπλωματικής εργασίας ήταν ο σχεδιασμός κατάλληλων δομών και η ανάπτυξη αλγορίθμου για την αποτίμηση πιθανοτικών ερωτημάτων εγγύτερων γειτόνων για αβέβαιες θέσεις κινούμενων αντικειμένων ($k\theta NN$). Από την μελέτη ζητημάτων σχεδίασης τέτοιων συστημάτων, προκύπτει ότι:

- Η επιλογή καννάβου ως χωρικού ευρετηρίου και η προσπέλαση των κελιών ανά επάλληλες ζώνες αποδείχτηκε ιδανική για την διαρκή παρακολούθηση των αβέβαιων αντικειμένων που είναι πλησιέστερα στα εκάστοτε ερωτήματα.
- Η γνωστή στρατηγική *«φιλτράρισμα & εκλέπτυνση»* προσαρμόστηκε ώστε να αξιοποιεί κυρίως γεωμετρικά, αλλά και πιθανοτικά χαρακτηριστικά των δεδομένων, προκειμένου να επιτυγχάνει αποδοτική αποτίμηση των ερωτημάτων.
- Οι χρονικές επιδόσεις του αλγορίθμου μετρήθηκαν πειραματικά και αποδείχθηκαν ιδιαιτέρως επαρκείς για τον χειρισμό πολλαπλών κινούμενων ερωτημάτων διαρκείας.

Τέλος, ένα τέτοιο ολοκληρωμένο σύστημα διαχείρισης ρευμάτων κινούμενων ερωτημάτων $k\theta NN$, θα ήταν ενδιαφέρον αν μπορούσε να επεκταθεί για διαφορετικές πολυδιάστατες πιθανοτικές κατανομές (λ.χ. χ^2, γ, Fréchet, Student κ.ά.), όπως τυχόν επιβάλλει η μοντελοποίηση διαφορετικών προβλημάτων, αλλά επίσης και για αποτίμηση άλλου τύπου αναζητήσεων, π.χ. αντίστροφων πιθανοτικών k-εγγύτερων γειτόνων (*reverse k-NN*).