ΤΜΗΜΑ ΗΛΕΚΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη δεδομένων για υπηρεσίες με βάση τη θέση

Στύλος Σταύρος

Επιβλέπων Καθηγητής:
Βασιλακόπουλος Μιχαήλ
Αναπληρωτής Καθηγητής Π.Θ.

Συνεπιβλέπουσα Καθηγήτρια:
Δασκαλοπούλου Ασπασία
Επίκουρος Καθηγήτρια Π.Θ.

Βόλος,
Μάρτιος 2017

DATA MINING FOR LOCATION BASED SERVICES

Stylos Stavros

Στην οικογένειά μου

ΕΥΧΑΡΙΣΤΙΕΣ

Με την παρούσα Διπλωματική Εργασία ολοκληρώνεται ο προπτυχιακός κύκλος των σπουδών μου στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Θεσσαλίας.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων, αναπληρωτή Καθηγητή κ. Βασιλακόπουλο Μιχαήλ τόσο για την εμπιστοσύνη του με την ανάθεση της εν λόγω εργασίας, όσο και για την πάντοτε πολύτιμη και έγκαιρη καθοδήγησή του, όπως και την συνεπιβλέπουσα επίκουρο Καθηγήτρια κα. Δασκαλοπούλου Ασπασία.

Επίσης, ένα μεγάλο ευχαριστώ στους γονείς μου Ιωάννα, Δημήτρη, στον αδερφό μου Θανάση και στον θείο μου Δήμο. Δίχως τη στήριξη, τη κατανόηση και την αγάπη τους, αυτό το κείμενο πιθανόν να μην γραφόταν ποτέ.

Ένα ακόμη μεγάλο ευχαριστώ σε όλους τους ανθρώπους που γνώρισα, αγάπησα και αγαπήθηκα όλα αυτά τα χρόνια. Τόσο τους φίλους που ήρθαν για να μείνουν, όσο και εκείνους που απλά ήταν περαστικοί.

Τέλος, ένα μεγάλο ευχαριστώ για όλα και μία μεγάλη συγγνώμη που δεν πρόλαβα. Για τον παππού Σταύρο.

Περιληψη

Με τη διαρκή εξέλιξη των συστημάτων γεωγραφικών πληροφοριών (**GIS**) σε συνδυασμό με την πλέον διαδεδομένη χρήση υπηρεσιών με βάση τη θέση (**LBS**), δημιουργείται συνεχώς η ανάγκη για αυτόματη εξαγωγή νέας γνώσης από μεγάλες βάσεις χωρικών δεδομένων. Δοθέντος ενός συνόλου χωρικών αντικειμένων, η διαδικασία εξόρυξης συντοποθετημένων προτύπων ανακαλύπτει τα υποσύνολα των αντικειμένων αυτών, των οποίων η συνύπαρξη σε μία συγκεκριμένη γεωγραφική εμβέλεια παρουσιάζει μεγάλη συχνότητα. Το πρόβλημα εξόρυξης κανόνων χωρικής συντοποθεσίας (**spatial co-location rules mining**),οι οποίοι μας παραθέτουν τις εξαρτήσεις μεταξύ των αντικειμένων και προκύπτουν μέσω των προαναφερθέντων προτύπων, είναι διαφορετικό από το κλασσικό πρόβλημα εξόρυξης κανόνων συσχέτισης, καθώς δεν υπάρχουν προκαθορισμένες συναλλαγές. Έτσι χρησιμοποιείται η έννοια της γειτονιάς αντικειμένων με βάση κάποιο όριο απόστασης χρήστη. Μια γειτονιά υφίσταται όταν η απόσταση μεταξύ δύο αντικειμένων βρίσκεται εντός της εκάστοτε επιθυμητής εμβέλειας. Μέσω των γειτονιών, ανάλογα με το μοντέλο εξόρυξης που θα ακολουθήσουμε, δημιουργούμε και τις συναλλαγές μέσω των οποίων θα καταλήξουμε στους ζητούμενους κανόνες. Στην μεγάλη πλειοψηφία των μεθόδων της υπάρχουσας βιβλιογραφίας οι γειτονιές υπολογίζονται με βάση την Ευκλείδεια απόσταση των αντικειμένων, λαμβάνοντας το χώρο ως ομοιογενή και ισοτροπικό. Στα πλαίσια, λοιπόν, της εν λόγω διπλωματικής εργασίας υιοθε-τήθηκε ένα βασικό μοντέλο εξόρυξης κανόνων στοντοποθεσίας [1] και χρησιμοποιήθηκαν για λόγους πειραματικής αξιολόγησης τρείς διαφορετικοί τρόποι εύρεσης γειτονιών, οι δύο εκ των οποίων [2][3] λαμβάνουν υπόψη τους την περιοριστικότατα του δικτύου του δρόμου, ενώ ο τρίτος χρησιμοποιεί την προσέγγιση Ευκλείδειας απόστασης.

ABSTRACT

The continuous evolution of geographic information systems (GIS) combined with the widespread utilization of location based services (LBS), increases the need to extract automatically new knowledge from large spatial databases. Given a set of spatial features, the **co-location pattern mining procedure** discovers the subsets of these features whose co-existence at a specific geographic proximity presents high frequency. The problem of spatial co-location rules mining, which rules show us the dependencies between features and they are result of the above mentioned patterns, is different than the classic association rules mining problem, as there isn't existence of predefined transactions. So, we use the notion of user-specified neighborhood. A neighborhood is created if the distance between two features is under a specified geographic proximity. By the neighborhoods, depending the data mining model we will follow, we create the transactions with whom we will take the requested rules. The majority of methods to the existence literature, calculate the neighborhoods by the Euclidean distance, taking the space as homogeneous and isotropic. Within this thesis, was adopted a basic model of co-location rule mining [1] and they were used three different ways for neighborhoods finding for experimental evaluation reason, the two of them [2][3] take into account the road network constraints, as the third uses the Euclidean distance approach.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Σκοπος και Διαρθρωση Εργασιας

Αν κάποιος ανατρέξει στο διαδίκτυο να βρει εγχώριες πηγές που αφορούν την εξόρυξη προτύπων και κανόνων χωρικών δεδομένων **με βάση τη συχνότητα συνύπαρξής τους σε μία συγκεκριμένη γεωγραφική εμβέλεια,** θα καταλάβει πολύ γρήγορα πως πρόκειται για έναν τομέα ο οποίος δεν είναι ιδιαίτερα τετριμμένος, παρότι πολύ χρήσιμος και σημαντικός σε παγκόσμιο επίπεδο, καθώς όπως ο τομέας της κλασσικής εξόρυξης δεδομένων, βρίσκει πολλές εφαρμογές σε διάφορες πτυχές της καθημερινότητάς μας. Ο βασικός λόγος, λοιπόν, που περιλαμβάνω αρκετά συχνά τον όρο **"συντοποθεσία" -** ο οποίος δεν υφίσταται σε κάποιο λεξικό - στο κείμενο που ακολουθεί, είναι στη προσπάθεια μου να κάνω μία όσο πιο πιστή μετάφραση του όρου co-location.

Συνεπώς, ένας από τους βασικούς σκοπούς της εν λόγω εργασίας είναι να αποτελέσει μία όσο το δυνατότερο αξιόπιστη και βοηθητική πηγή, για άτομα τουλάχιστον σαν και εμένα που όταν ξεκίνησα να ασχολούμαι με το συγκεκριμένο θέμα, συνάντησα ένα χάος πληροφοριών και διαβάζοντας αρκετές μη καλά οργανωμένες ή πέρα του θέματος εν τέλει πηγές, έχασα αρκετό χρόνο. Ένας ακόμη βασικός σκοπός της εργασίας, είναι η εύρεση των χρονικών και σημασιολογικών διαφορών μεταξύ δύο διαφορετικών προσεγγίσεων υπολογισμού απόστασης αντικειμένων με σκοπό την εύρεση του ακρογωνιαίου ουσιαστικά λίθου της διαδικασίας, των γειτονιών, μέσω α) της Ευκλείδειας απόστασης και β)της απόστασης δικτύου.

Τέλος, θα ήθελα να αναφέρω πως λόγω του ότι δεν θεωρώ πως η έτοιμη "τροφή" βοηθάει ουσιαστικά, στις περισσότερες των περιπτώσεων τουλάχιστον, αποφάσισα να μην παραθέσω πολλά κομμάτια κώδικα, αλλά θα προσπαθήσω να εξηγήσω όσο το δυνατόν καλύτερα μπορώ τα σημαντικότερα σημεία της διαδικασίας που ακολουθήθηκε.

Η διάρθρωση της εργασίας έχει ως εξής:

Στο **κεφάλαιο 1** παρατίθενται κάποιες εισαγωγικές έννοιες που αφορούν τα **Συστήματα Γεωγραφικών Πληροφοριών**, τα **μοντέλα αναπαράστασης των γεωγραφικών δεδομένων**, τους **τρόπους συλλογής αυτών** και τις **Υπηρεσίες με βάση τη θέση**.

Στο **κεφάλαιο 2** γίνεται λόγος για τη **διαδικασία εξόρυξης γνώσης μέσα από βάσεις δεδομένων** και για το βασικό βήμα αυτής, την **εξόρυξη δεδομένων-προτύπων**. Το εν λόγω κεφάλαιο, θα μπορούσε να πει κανείς, πως λειτουργεί ως μία γέφυρα η οποία φέρνει σε σύνδεση, ουσιαστικά, τις έννοιες που παρουσιάζονται στο κεφάλαιο 1 με τη διαδικασία της εξόρυξης χωρικών δεδομένων που παρουσιάζεται στο κεφάλαιο 3, η οποία δεν αποτελεί κάτι άλλο πέρα από μια προσαρμογή της διαδικασίας εξόρυξης δεδομένων στις έννοιες των χωρικών δεδομένων.

Στο **κεφάλαιο 3** παρουσιάζονται η **διαδικασία εξόρυξης χωρικών δεδομένων και προτύπων** και οι τεχνικές που χρησιμοποιούνται προς αυτή τη κατεύθυνση, ενώ γίνεται μία εκτενής ανάλυση για τους **κανόνες χωρικής συντοποθεσίας** με την παρουσίαση δύο αλγοριθμικών προσεγγίσεων.

Στο **κεφάλαιο 4** παρουσιάζονται τα βήματα της διαδικασίας που ακολουθήθηκε, προκειμένου να φτάσουμε στη σύγκριση μεταξύ των διαφορετικών προσεγγίσεων εύρεσης χωρικών γειτονιών, **της χρήσης Ευκλείδειας απόστασης** και **της χρήσης απόστασης δικτύου**.

Τέλος, στο **κεφάλαιο 5**, παρουσιάζονται η **πειραματική αξιολόγηση** και τα **συμπεράσματα** που "εξορύχτηκαν" από τη διαδικασία που ακολουθήθηκε.

"Θα 'χουν τα κύματα ξεβράσει όλα τα μηνύματα

και τα πιο όμορφά μας ποιήματα-συνθήματα σε χείλη ξένα."

Σ.Σ.

1.ΕΙΣΑΓΩΓΙΚΕΣ ΈΝΝΟΙΕΣ

1.1 ΣΥΣΤΗΜΑ ΓΕΩΓΡΑΦΙΚΩΝ ΠΛΗΡΟΦΟΡΙΩΝ

Η γνώση συγκεκριμένης και στοχευόμενης γεωγραφικής πληροφορίας εκτός του ότι κάνει την καθημερινότητά μας πιο εύκολη, θεωρείται πως παίζει κομβικό ρόλο (άμεσα ή έμμεσα) και στη συντριπτική πλειοψηφία των οικονομικών, πολιτικών και άλλων σημαντικών αποφάσεων παγκοσμίως.

Την ανάγκη για **συλλογή, διαχείριση, επεξεργασία, ανάλυση, μοντελοποίηση** και **απεικόνιση** των **γεωγραφικών δεδομένων** – λειτουργίες που πλέον γίνονται κατά βάση μέσω υπολογιστικών συστημάτων – προκειμένου να καταλήξουμε σε οποιαδήποτε ζητούμενη γνώση (γεωγραφική ή μη), την καλύπτουν τα **Συστήματα Γεωγραφικών Πληροφοριών** (ΣΓΠ), ευρέως γνωστά ως **G.I.S.** – Geographic Information Systems.

Προσεγγίζοντας αναλυτικότερα την έννοια του όρου, σύμφωνα με τον διεθνή προμηθευτή λογισμικού του εν λόγω συστήματος, **ESRI** (Environmental Systems Research Institute), ένα ΣΓΠ είναι ένα υπολογιστικό εργαλείο το οποίο χρησιμοποιείται για τη χαρτογράφηση και την ανάλυση πραγμάτων που υπάρχουν και γεγονότων που συμβαίνουν στη Γη. Η τεχνολογία ΣΓΠ συνδυάζει λειτουργίες βάσεων δεδομένων, όπως τα ερωτήματα-queries και η στατιστική ανάλυση, με τα πλεονεκτήματα οπτικοποίησης και γεωγραφικής ανάλυσης που προσφέρουν οι χάρτες.

1.1.1 ΧΡΗΣΙΜΟΤΗΤΑ ΤΩΝ ΣΓΠ

Η επίδραση των ΣΓΠ, όπως προαναφέρθηκε, παρότι "αθόρυβη" για το ευρύ κοινό, είναι αρκετά μεγάλη σε διάφορες πτυχές της καθημερινότητάς μας. Για παράδειγμα, συμβουλευόμενοι μια υπολογιστική εφαρμογή χαρτογράφησης προκειμένου να βρούμε ποιο δρόμο θα πρέπει να ακολουθήσουμε ώστε να βρεθούμε σε κάποιο συγκεκριμένο πολυκατάστημα, τότε ουσιαστικά κάνουμε χρήση ΣΓΠ. Χρήση εργαλείων ΣΓΠ, όμως, πολύ πιθανό να έχουν κάνει και οι ιδιοκτήτες του πολυκαταστήματος αυτού, προτού καν αρχίσουν οι εργασίες κατασκευής του, προκειμένου βάσει στατιστικής ανάλυσης γεωγραφικών σε συνδυασμό με άλλα (περιγραφικά) δεδομένα, να αποφασίσουν ένα όσο το δυνατότερο καλύτερο σημείο, στο οποίο υπάρχει μεγαλύτερη πιθανότητα από κάποιο άλλο, να καλύψουν τις αγοραστικές ανάγκες των πελατών και κατ' επέκταση να αποκομίσουν

μεγαλύτερα κέρδη. Προς αποφυγή αποπροσανατολισμού, την επεξεργασία διάφορων χωρικών ερωτημάτων μέσω κάποιας κινητής συσκευής σε πραγματικό χρόνο, ενώ βρισκόμαστε σε κίνηση, την καλύπτουν οι Υπηρεσίες με βάση τη θέση (LBS),οι οποίες ουσιαστικά είναι ένα προϊόν το οποίο έχει προκύψει από τον συνδυασμό των GIS, GPS, GeoWeb και Mobile Computing και οι οποίες θα συζητηθούν αργότερα. Όπως, επίσης, η επεξεργασία πιο περίπλοκων, τουλάχιστον, ερωτημάτων, αποτελεί αποτέλεσμα ανάλυσης των ΣΓΠ βάσεων από διάφορες μεθόδους της υπολογιστικής διαδικασίας της **εξόρυξης δεδομένων**, η οποία και αποτελεί τη βάση αυτής της διπλωματικής εργασίας. Πέρα του παραδείγματος που παρατέθηκε, τα ΣΓΠ χρησιμοποιούνται για ακόμη πιο σημαντικές και κρίσιμες λειτουργίες, όπως η δημιουργία σχεδιαγράμματος κίνησης μιας ομάδας διάσωσης σε ένα μέρος που έχει υποστεί μία μεγάλη φυσική καταστροφή.

Η χαρακτηριστική δυνατότητα που παρέχουν, λοιπόν, τα ΣΓΠ είναι αυτή της σύνδεσης χωρικής πληροφορίας με άλλες μη χωρικές (περιγραφικές) ιδιότητες-πληροφορίες.

1.1.2 ΜΟΝΤΕΛΑ ΑΠΟΘΗΚΕΥΣΗΣ,ΔΙΑΜΟΡΦΩΣΗΣ ΚΑΙ ΑΝΑΚΤΗΣΗΣ ΣΓΠ

Όλες αυτές οι λειτουργικότητες που αναφέρθηκαν όπως και αρκετές άλλες, προσφέρονται μέσω εργαλείων και επεκτάσεων λογισμικών ΣΓΠ, ενώ η αποθήκευση, η διαμόρφωση και η ανάκτηση των δεδομένων μπορεί να γίνει μέσω:

- Σχεσιακών βάσεων δεδομένων, οι οποίες υποστηρίζουν τους βασικούς τύπους δεδομένων, με τους πίνακες των οποίων δημιουργούνται συνδέσεις με τα εκάστοτε χωρικά δεδομένα των οποίων η υποστήριξη γίνεται απλοϊκά (π.χ. ένα σημείο ως δύο αριθμοί που περιγράφουν το γεωγραφικό πλάτος και μήκος του).

- Αντικειμενοστραφών βάσεων δεδομένων, οι οποίες υποστηρίζουν γενικούς τύπους δεδομένων (abstract data types-ADT's) που ορίζονται από τον χρήστη, οπότε και η προσθήκη-δημιουργία χωρικών τύπων δεδομένων (π.χ. polygon) καθίσταται εφικτή.

Το αντικειμενοστραφές μοντέλο τείνει να χρησιμοποιείται όλο και περισσότερο σε εφαρμογές GIS εξαιτίας των αυξημένων δυνατοτήτων του σε σχέση με το σχεσιακό μοντέλο.

1.2 ΑΝΑΠΑΡΑΣΤΑΣΗ ΓΕΩΓΡΑΦΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Κάθε γεωγραφικό αντικείμενο ή συμβάν, την έννοια της χωρικής υπόστασης του οποίου τη δίνει η ίδια του η ύπαρξη σε ένα συγκεκριμένο σημείο πάνω στο χάρτη – συνήθως προσδιορισμός με βάση συντεταγμένες γεωγραφικού μήκους και πλάτους – , όπως εύκολα μπορεί να γίνει αντιληπτό, συνδέεται άρρηκτα και με άλλα μη-γεωγραφικά χαρακτηριστικά-ιδιότητες. Για παράδειγμα, το κτήριο το οποίο στεγάζει τα αμφιθέατρα μίας σχολής, μπορεί να λογιστεί χωρικά ως ένα **σημείο με μοναδικές τιμές συντεταγμένων,** σε συνδυασμό με κάποιες άλλου είδους πληροφοριακές ιδιότητες, όπως η οδός στην οποία αυτό βρίσκεται, ο ταχυδρομικός του κώδικας, αλλά και η ονομασία του τμήματος από το οποίο χρησιμοποιείται.

Με τον όρο **γεωγραφικά δεδομένα,** λοιπόν, αναφερόμαστε σε κάθε σύνολο δεδομένων που περιέχει κατ' ελάχιστο ένα στοιχείο τοποθεσίας το οποίο περιγράφει ένα σύνολο από σημεία, γραμμές, πολύγωνα ή επίπεδα, σε αντιστοίχιση με μη-γεωγραφικά χαρακτηριστικά-ιδιότητες.

Στην κατηγορία αυτή συναντάμε πολλά σύνολα δεδομένων, από δρόμους και σημεία ενδιαφέροντος μίας πόλης ή μίας ολόκληρης χώρας, μέχρι δεδομένα οικολογικού, βιολογικού και γενικότερα επιστημονικού ενδιαφέροντος. Θεμελιωδώς, τα γεωγραφικά δεδομένα είναι είτε **διανυσματικά** (**vector**) είτε **ψηφιδωτά** (**raster**) – απαρτισμένα από γεωμετρία ή από pixels αντίστοιχα. Οι δύο αυτοί τύποι, συχνά μπορεί να συνδυαστούν, όταν για παράδειγμα διανυσματικά δεδομένα δρόμου επικαλύπτονται από ψηφιδωτά δεδομένα δορυφόρου.

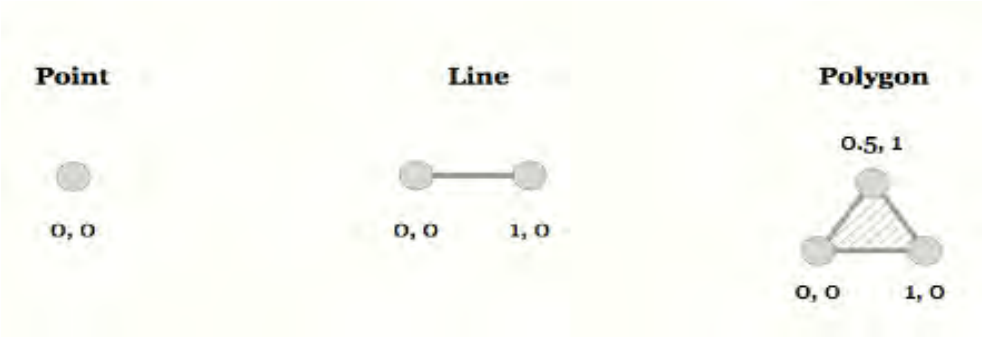
1.2.1 ΔΙΑΝΥΣΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

Τα **διανυσματικά (vector)** δεδομένα αναπαριστούν τα διάφορα γεωγραφικά χαρακτηριστικά-αντικείμενα ως **σημεία, γραμμές** ή **πολύγωνα,** ανάλογα τόσο με την εκάστοτε κλίμακα η οποία χρησιμοποιείται όσο και με το είδος των αντικειμένων προς αναπαράσταση.

- Τα διανυσματικά σημεία (Points) είναι απλώς ένα ζεύγος συντεταγμένων γεωγραφικού μήκους και πλάτους και αναπαριστούν χαρακτηριστικά τα οποία καταλαμβάνουν αρκετά μικρό μέρος του χώρου, όπως τα σημεία ενδιαφέροντος(POIs) μίας συγκεκριμένης περιοχής.

- Οι διανυσματικές γραμμές (LineStrings/MultilineStrings) είναι μια διατεταγμένη λίστα σημείων τα οποία συνδέονται μέσω μονοπατιών – παρουσιάζουν, δηλαδή, μια γραμμική συνέχεια (π.χ. δρόμοι και ποτάμια) –, όπως επίσης και άλλων νοητών σημείων, όπως είναι τα σύνορα μεταξύ δύο κρατών.

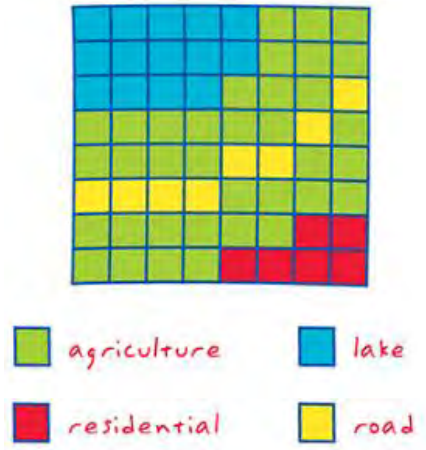
- Τέλος, τα διανυσματικά πολύγωνα (Polygons), ουσιαστικά, είναι διανυσματικές γραμμές οι οποίες παρουσιάζουν κλειστότητα – το αρχικό και τελικό σημείο τους, δηλαδή, είναι το ίδιο. Με την μορφή πολυγώνων συνήθως συναντάμε χαρακτηριστικά που καταλαμβάνουν αρκετά μεγάλο χώρο σε ένα χάρτη, όπως λίμνες, δάση, μεγάλα οικοδομικά συγκροτήματα κ.ά.



Εικόνα 1.1 – Διανυσματικοί τύποι δεδομένων

1.2.2 ΨΗΦΙΔΩΤΑ ΔΕΔΟΜΕΝΑ

Η έννοια των **ψηφιδωτών (raster)** δεδομένων μπορεί να γίνει πιο εύκολα κατανοητή αν λάβουμε υπόψη μας μία φωτογραφία ψηφιακής κάμερας όπου στο χαμηλότερο επίπεδο αφαίρεσής της, δεν είναι τίποτα περισσότερο από μία λίστα απαρτισμένη από pixels με τιμές. Τα διανυσματικά δεδομένα λαμβάνουν, δηλαδή, το χώρο ως ένα **επίπεδο** που απαρτίζεται από ένα πίνακα από κελιά ίδιου μεγέθους, το καθένα από τα οποία περιέχει μία τιμή ιδιότητας και bottom-left συντεταγμένες κελιού (σε **αντίθεση** με τα διανυσματικά δεδομένα που μας δίνουν τις ακριβείς συντεταγμένες ενός οποιοδήποτε σημείου πάνω σε ένα χάρτη). Τα κελιά που μοιράζονται τις ίδιες τιμές ιδιοτήτων, αντιπροσωπεύουν τον ίδιο τύπο ενός γεωγραφικού χαρακτηριστικού.



Εικόνα 1.2 - Ψηφιδωτά χωρικά δεδομένα

1.2.3 ΕΠΙΛΟΓΗ ΚΑΤΑΛΛΗΛΩΝ ΔΕΔΟΜΕΝΩΝ

Όσον αφορά την επιλογή ανάμεσα σε αυτούς τους δύο τύπους δεδομένων, αυτή εξαρτάται από το είδος της εργασίας που έχουμε να φέρουμε εις πέρας. Επί παραδείγματι, τα **pixels** στα **ψηφιδωτά** δεδομένα έχουν ιδιότητες όπως χρώμα, αδιαφάνεια ή ύψος και οι οποίες, θεωρητικά, έχουν μεγαλύτερη χρησιμότητα σε επιστημονικού επιπέδου έρευνες και αναλύσεις. Από την άλλη μεριά, τα **διανυσματικά** δεδομένα, συνήθως, μας παρέχουν πιο γενικές πληροφορίες που αφορούν ένα γεωγραφικό χαρακτηριστικό(εκτός του σχήματος και των συντεταγμένων του), όπως στην περίπτωση ενός δρόμου, όπου συνήθως αυτός συνοδεύεται από την κατηγορία (πρωτεύον, κορμού κλπ.),το αν είναι μονής ή διπλής κυκλοφορίας, αλλά και την ονομασία του.

1.3 ΣΥΛΛΟΓΗ ΓΕΩΓΡΑΦΙΚΩΝ-ΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Η διαδικασία της συλλογής και επεξεργασίας των γεωγραφικών δεδομένων, προκειμένου αυτά να υποστούν τις οποιεσδήποτε αναλύσεις ή και να αποδοθούν βάσει οπτικοποίσης, είναι μία από τις πιο δαπανηρές χρονικά και υπολογιστικά, αλλά ταυτόχρονα και πιο σημαντικές διαδικασίες στον κύκλο εργασιών ενός ΣΓΠ. Τα δεδομένα, όπως θα μπορούσε κανείς να πει, αποτελούν μαζί με το ανθρώπινο δυναμικό, τον πυρήνα ενός ΣΓΠ, καθώς χωρίς αυτά δεν υφίσταται η έννοια της πληροφορίας.

1.3.1 ΠΡΩΤΟΓΕΝΗ ΚΑΙ ΔΕΥΤΕΡΟΓΕΝΗ ΔΕΔΟΜΕΝΑ

Τα γεωγραφικά δεδομένα, πέρα των δύο κατηγοριών που αναφέρθηκαν στην προηγούμενη ενότητα, χωρίζονται και σε δύο ευρύτερες

κατηγορίες, με βάση την πηγή από την οποία συλλέγονται : σε **πρωτογενή** και **δευτερογενή** δεδομένα.

Με τον όρο **πρωτογενή δεδομένα** αναφερόμαστε σε δεδομένα που συλλέγονται απευθείας σε ψηφιακή μορφή, ενώ με τον όρο **δευτερογενή δεδομένα** αναφερόμαστε συνήθως σε δεδομένα τα οποία προέρχονται από **αρχειακές πηγές** (π.χ. αρχεία απογραφών πληθυσμού) και τα οποία χρίζουν **ψηφιοποίησης.**

1.3.2 ΤΡΟΠΟΙ ΣΥΛΛΟΓΗΣ ΠΡΩΤΟΓΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

Τις τελευταίες δεκαετίες, όπως είναι λογικό, οι πηγές από τις οποίες, κατά βάση, γίνεται η συλλογή δεδομένων, είναι άρρηκτα συνδεδεμένες με τις τεχνολογικές εξελίξεις. Συνεπώς, αυτά συλλέγονται απευθείας σε ψηφιακή μορφή. Τα βασικά μέσα, λοιπόν, που χρησιμοποιούνται προς αυτή την κατεύθυνση, είναι:

- **Παρατηρητικοί δορυφόροι κινητής τροχιάς και αεροσκάφη**, όπου μέσω των ειδικά τοποθετημένων αισθητήρων τους, μας παρέχουν δεδομένα τηλεπισκόπισης, δηλαδή ψηφιακές εικόνες οι οποίες ουσιαστικά περιλαμβάνουν ιδιότητες από μετρήσεις φυσικών, χημικών και βιολογικών ιδιοτήτων αντικειμένων – χωρίς άμεση επαφή, και οι οποίες δεν είναι τίποτε άλλο από **ψηφιδωτά δεδομένα.**

- **Τα Παγκόσμια Δορυφορικά Συστήματα Πλοήγησης (Global Navigation Satellite Systems)**, ένα εκ των οποίων είναι το **Παγκόσμιο Σύστημα Στιγματοθέτησης**, ευρέως γνωστό ως **GPS**, το οποίο βασίζεται σε ένα πλέγμα εικοσιτεσσάρων δορυφόρων της Γης, των οποίων οι ειδικές συσκευές εντοπισμού (πομπο-δέκτες) παρέχουν ακριβείς πληροφορίες για τη θέση ενός σημείου, το υψόμετρό του, την ταχύτητα και την κατεύθυνση της κίνησής του. Η μεγάλη πλειοψηφία των **διανυσματικών** δεδομένων που συναντάμε σήμερα σε ψηφιακούς χάρτες, είναι αποτέλεσμα συλλογής δεδομένων από αυτούς τους δορυφόρους. Παρόμοια συστήματα είναι το υπό ανάπτυξη ευρωπαϊκό σύστημα GALILEO , αλλά και το ρώσικο σύστημα GLONASS.

Οι μορφές αρχείων που συναντάμε τα ψηφιδωτά δεδομένα είναι η **GeoTIFF**, όπως επίσης και η **JPEG2000**, ενώ η βασική μορφή που συναντάμε τα περισσότερα διανυσματικά δεδομένα, είναι η **Shapefile.**

1.4 ΥΠΗΡΕΣΙΕΣ ΜΕ ΒΑΣΗ ΤΗ ΘΕΣΗ

Τόσο η πανταχού παρουσία του ασύρματου δικτύου σε συνδυασμό με τη δυνατότητα πρόσβασης σε εφαρμογές πλοήγησης GPS που προσφέρουν πλέον όλα τα τερματικά κινητών συσκευών, όσο και η διαρκής εξέλιξη των Γεωγραφικών Πληροφοριακών Συστημάτων, οδήγησαν στις αρχές της προηγούμενης δεκαετίας σε μία νέα γενιά λογισμικού για κινητές υπηρεσίες, γνωστές ως **Υπηρεσίες Με Βάση Τη Θέση-Location Based Services (LBS).** Οι υπηρεσίες αυτές μέσω των διάφορων εφαρμογών τους, συνδυάζοντας τις πληροφορίες που τους παρέχει το σύστημα GPS με τις πληροφορίες της βάσης δεδομένων του GIS, λαμβάνουν κάθε στιγμή την ακριβής τοποθεσία του εκάστοτε χρήστη και του προσφέρουν πληροφορίες σε πραγματικό χρόνο, απατώντας του για παράδειγμα σε queries όπως "Ποιό είναι το πιο κοντινό ATM με βάση την κατεύθυνσή μου και την κίνηση στο δρόμο αυτή τη στιγμή?". Συνεπώς, όπως προαναφέρθηκε, οι Υπηρεσίες με βάση τη θέση, μπορούν πολύ σωστά να θεωρηθούν ως ένα επιπέδου-λογισμικού προϊόν που έχει προκύψει από των συνδυασμό των υπηρεσιών **GIS**, **GPS**, **GeoWeb** και **Mobile Computing** σε μία προσπάθεια αύξησης των πλεονεκτημάτων που πρόσκεινται στις υπάρχουσες τηλεπικοινωνιακές υπηρεσίες. Τέλος, όπως γίνεται εύκολα αντιληπτό και σε συνδυασμό με τα όσα λέχθηκαν και στην ενότητα 1.1, οι εν λόγω υπηρεσίες, χρησιμοποιούν τεχνικές εξόρυξης δεδομένων και βρίσκουν μεγάλη χρησιμότητα και απήχηση σε διάφορες πτυχές της καθημερινότητας, απαντώντας μας από πιο απλά έως πιο περίπλοκα ερωτήματα.

2.ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

2.1 ΕΙΣΑΓΩΓΗ

Στο σημείο αυτό και στα πλαίσια της ανάγκης τόσο των υπηρεσιών που παρουσιάστηκαν στις ενότητες 1.1 και 1.4, **GIS** και **LBS** αντίστοιχα, όσο και άλλων υπηρεσιών οι οποίες δεν σχετίζονται απαραίτητα με χωρικά δεδομένα, για διαρκή εξέλιξη και κατ' επέκταση τη δυνατότητα παροχής όλο και ακριβέστερων-πολυπλοκότερων πληροφοριών ως προς τους χρήστες τους, μπορούμε να έρθουμε σε μία πρώτη επαφή και με τη διαδικασία της εξόρυξης δεδομένων-προτύπων, προκειμένου να θέσουμε κάποιες βασικές έννοιες, έτσι ώστε να περάσουμε αργότερα στη συζήτηση της επέκτασης, ουσιαστικά, των λειτουργιών της, με σκοπό την εξαγωγή χωρικής γνώσης.

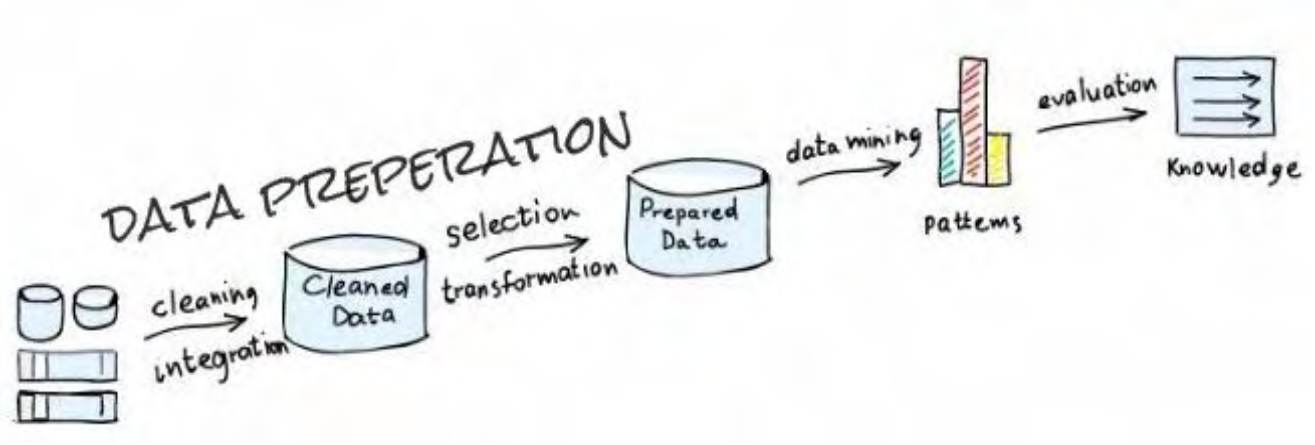
Η **εξόρυξη δεδομένων (data mining)**, αποτελεί μία υπολογιστική διαδικασία, η οποία χρησιμοποιώντας διάφορες κατηγορίες αλγορίθμων έχει ως τελικό σκοπό ,την εξαγωγή **ενδιαφέρουσας, κατανοητής** και μέχρι πρότινος **μη-γνωστής πληροφορίας ή προτύπων** μέσα από μεγάλες βάσεις δεδομένων.

Πρότυπο, μπορεί να θεωρηθεί κάθε σύνολο πραγμάτων-δεδομένων τα οποία σχετίζονται μεταξύ τους κάτω υπό κάποιες γενικές ή ειδικές συνθήκες και η συσχέτιση των οποίων χρίζει άξιας λόγου και προσοχής.

Εκτός απο μία υπολογιστική διαδικασία, η εξόρυξη δεδομένων, θεωρείται και ως ένα διαπανεπιστημιακό υποπεδίο της επιστήμης των υπολογιστών. Ο όρος που έχει επικρατήσει, αυτός καθ' αυτός, όμως, μπορεί να θεωρηθεί αν όχι λανθασμένος, τότε μη-ακριβής, καθώς ο βασικός στόχος των τεχνικών που χρησιμοποιούνται, όπως μόλις προαναφέρθηκε, είναι η -διαμέσου μεγάλου όγκου δεδομένων- εξόρυξη προτύπων και γνώσης και όχι απλά η εξαγωγή κάποιων δεδομένων, όπως αυτά μπορούν να ληφθούν υπόψη με την αυστηρή έννοια του δικού τους όρου. Γι' αυτόν τον λόγο, λοιπόν, σε αρκετές βιβλιογραφικές πηγές συναντάμε αντί για τον όρο **"εξόρυξη δεδομένων"**, τον όρο **"εξόρυξη προτύπων"**, χωρίς να αλλάζει κάτι, ουσιαστικά, ως προς τη σημασιολογία και τις μεθόδους που ακολουθούνται.

2.1.1 ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΜΕΣΑ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Την εξόρυξη δεδομένων, την συναντάμε πριν το βήμα ανάλυσης μίας ευρύτερης διαδικασίας, της "Ανακάλυψης γνώσης μέσα από βάσεις δεδομένων" (**Knowledge discovery in databases-KDD**)[4][5], η οποία όπως φαίνεται και στην εικόνα που ακολουθεί, ουσιαστικά, διαθέτει τρία βασικά βήματα: **α)**την **προ-επεξεργασία**, η οποία περιλαμβάνει κάποια "υποβήματα" όπως η επιλογή δεδομένων, ο καθαρισμός τους, η ενοποίησή τους και ο μετασχηματισμός τους, **β)**την **εξόρυξη προτύπων** μέσω των προεπεξεργασμένων δεδομένων και τέλος **γ)**την **ερμηνεία/αξιολόγηση** και **οπτικοποίηση** των αποτελεσμάτων που προέκυψαν από το βήμα της εξόρυξης.



Εικόνα 2.1 - Εξόρυξη γνώσης μέσα από βάσεις δεδομένων

Λόγω του τεράστιου φόρτου δεδομένων, λοιπόν, που κατακλύζουν την καθημερινότητά μας σε συνδυασμό με τη συνεχή εξέλιξη των τεχνολογιών διαδικτύου, όπως επίσης και με βάση τις όλο και περισσότερο πιο περίπλοκες απαιτήσεις των χρηστών, όπως κανείς μπορεί να καταλάβει, η ανακάλυψη προτύπων και κατ' επέκταση γνώσης, δεν θα μπορούσε να μην αποτελεί αναπόσπαστο κομμάτι πάρα πολλών λειτουργιών που συναντάμε σε μία πληθώρα πραγμάτων που συμβαίνουν γύρω μας, τα οποία ακολουθώντας έναν πιο αφαιρετικό τρόπο σκέψης, στο επίπεδο της επιστήμης υπολογιστών, θωρούνται ως δεδομένα.

2.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Προτού συνεχίσουμε τη συζήτησή μας ειδικότερα περί της εξόρυξης δεδομένων, καλό θα ήταν να πούμε κάποια πράγματα για το βήμα που προηγείται αυτής, στην ευρύτερη διαδικασία της "Ανακάλυψης

γνώσης μέσα από βάσεις δεδομένων", την προεπεξεργασία των δεδομένων.

Η **προεπεξεργασία**-προετοιμασία των δεδομένων, προκειμένου να τα φέρουμε σε μία όσο το δυνατότερο καταλληλότερη μορφή - έτσι ώστε να έχουμε εν τέλει όσο πιο ακριβή και ποιοτικά γίνεται αποτελέσματα - αποτελεί περίπου το 80% του χρόνου μίας οποιασδήποτε εργασίας με σκοπό την ανακάλυψη γνώσης. Αυτή η διαπίστωση, τουλάχιστον, ισχύει για την περίπτωση που η εν λόγω διαδικασία πραγματοποιείται μέσω κάποιου εργαλείου εξόρυξης δεδομένων και δεν υπάρχει ανθρωπίνου δυναμικού "εμπλοκή" σε προγραμματιστικό επίπεδο. Όπως, εύκολα μπορεί να γίνει αντιληπτό, χωρίς το βήμα της προεπεξεργασίας δεν μπορούμε να κάνουμε λόγο για το βήμα της εξόρυξης δεδομένων, πόσο μάλλον για την ανακάλυψη οποιασδήποτε πιο περίπλοκης, τουλάχιστον, γνώσης.

2.2.1 ΒΗΜΑΤΑ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ

Όπως συμβαίνει πάντα σε ένα οποιοδήποτε πρόβλημα, το πρώτο βασικό βήμα που πρέπει να κάνουμε, προκειμένου, βάσει κάποιας ανάλογης του μεγέθους του προσπάθειας, να καταλήξουμε στην επίλυσή του, δεν είναι άλλο από την πλήρη κατανόηση των απαιτήσεων αυτού, πράγμα που προφανώς δεν αλλάζει όσον αφορά τη διαδικασία ανακάλυψης γνώσης.

Το πρώτο βασικό βήμα, λοιπόν, που πρέπει να ακολουθηθεί πριν καν αρχίσει η διαδικασία της προεπεξεργασίας, είναι η **κατανόηση των απαιτήσεων του εκάστοτε προβλήματος προς ανάλυση** που μας δίνεται και κατ' επέκταση η επιλογή των σωστών δεδομένων.

Τα τρία βασικά βήματα της προεπεξεργασίας, που ακολουθούν μετά την επιλογή των κατάλληλων δεδομένων, όπως μπορούμε να δούμε στη σελίδα που ακολουθεί, είναι τα εξής:

Καθαρισμός Δεδομένων (Data Cleaning)

Τα δεδομένα στον πραγματικό κόσμο είναι "βρώμικα" :

- **Ελλειπή - incomplete** : μπορεί να τους λείπουν κάποιες τιμές γνωρισμάτων ή ακόμη και κάποια ενδιαφέροντα γνωρίσματα.
- **Με θόρυβο - noisy** : περιέχουν λάθη ή περιθωριακές τιμές(outliers).
- **Ασυνεπή - inconsistent** : περιέχουν μη λογικές συσχετίσεις ή διπλότυπες εγγραφές.

Η διόρθωση των τριών, αυτών, κατηγοριών "βρώμικων" δεδομένων, αποτελεί τη **διαδικασία του καθαρίσματος** που επιβάλλεται να ακο-

λουθηθεί, προκειμένου να αυξηθεί η ποιότητα των δεδομένων που θα χρησιμοποιηθούν στην περαιτέρω διαδικασία.

Ενοποίηση Δεδομένων (Data Integration)

Η **ενοποίηση των δεδομένων,** είναι η διαδικασία κατά την οποία **συγχωνεύουμε** και **ομαδοποιούμε,** πιθανώς, δεδομένα που βρίσκουμε μέσω πολλαπλών και ετερογενών πηγών, με βάση τις ομοιότητές τους ως προς κάποια συγκεκριμένα χαρακτηριστικά.

Μετασχηματισμός Δεδομένων (Data Transformation)

Ο **μετασχηματισμός των δεδομένων,** είναι η διαδικασία μετατροπής δεδομένων ή πληροφοριών από μία μορφή σε μία άλλη, συνήθως, όπως και στη περίπτωση για την οποία συζητάμε, από τη μορφή του πηγαίου συστήματος – στη ζητούμενη μορφή του συστήματος προορισμού.

2.3 Η ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΠΡΑΓΜΑΤΙΚΟ ΚΟΣΜΟ

Πρωτού περάσουμε στα πιο "βαθιά" ζητήματα της εξόρυξης δεδομένων, καλό θα ήταν να δούμε το πως αυτή επιδρά πρακτικά στον πραγματικό κόσμο. Όπως, πιθανόν, λοιπόν, να έχει γίνει έως τώρα αντιληπτό, η εξόρυξη δεδομένων χρησιμοποιείται ευρέως και ποικιλοτρόπως και παρότι που οι εταιρίες που ασχολούνται με αυτή όλο και πληθαίνουν, υπάρχουν ακόμη αρκετές προκλήσεις προς μελέτη και επίλυση. Οι προκλήσεις αυτές, αφορούν κυρίως αλγοριθμικές προσεγγίσεις, οι οποίες δεν μπορούν να έχουν άλλους βασικούς στόχους πέρα από την όλο και μεγαλύτερη ακρίβεια των αποτελεσμάτων, όπως επίσης και από την ελαχιστοποίηση του χρονικού και υπολογιστικού κόστους των τεχνικών που χρησιμοποιούνται. Παρόλα αυτά, προφανώς και για να βρίσκει σημαντικό χώρο τόσο στην οικονομία όσο και σε μία μεγάλη κλίμακα επιστημών, είναι σε σημείο που προσφέρει αρκετά υψηλού επιπέδου υπηρεσίες.

Ενδεικτικά και επιγραμματικά κάποια από τα πεδία όπου η εν λόγω διαδικασία είναι αρκετά δημοφιλής, είναι τα εξής:

- **Ανάλυση χρηματοοικονομικών δεδομένων**
- **Βιομηχανία Λιανεμπορίου**
- **Βιομηχανία Τηλεπικοινωνιών**
- **Ανάλυση βιολογικών δεδομένων (Βιοπληροφορική)**
- **Ανίχνευση απάτης**

2.3.1 ΧΡΗΣΙΜΟΤΗΤΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

Επειδή, όμως, μπορεί να μην έχει γίνει ακόμη πλήρως κατανοητό πως μπορούν να χρησιμοποιηθούν οι υπηρεσίες που προσφέρει η εξόρυξη δεδομένων, όπως επίσης και τι ακριβώς μπορούν να προσφέρουν στα πεδία που μόλις αναφέρθηκαν, ας "ρίξουμε" μία πιο αναλυτική ματιά σε κάποια από αυτά.

Πεδίο ανάλυσης χρηματοοικονομικών δεδομένων

Τα οικονομικά δεδομένα τόσο των τραπεζικών εργασιών, όσο και γενικότερα όλων των κλάδων που αποτελούν αναπόσπαστο κομμάτι της οικονομίας, είναι συνήθως σχετικά και μεγάλης ποιότητας, πράγμα που διευκολύνει τόσο την εξόρυξη προτύπων, όσο και τη στατιστική ανάλυση. Κάποιες απο τις προσφορές της εξόρυξης προτύπων στο συγκεκριμένο πεδίο, είναι οι εξής :

- Πρόβλεψη πληρωμής δανείων και ανάλυση των κινήσεων της πιστωτικής κάρτας των πελατών.
- Ταξινόμηση και κατηγοριοποίηση των πελατών με σκοπό συγκεκριμένες στοχεύσεις του κλάδου του marketing.
- Εντοπισμός ξεπλύματος χρήματος και άλλων οικονομικών εγκλημάτων.

Πεδίο Βιομηχανίας και Εμπορίου

Όσον αφορά το εν λόγω αυτό γενικότερο πεδίο, η εξόρυξη προτύπων έρχεται να του προσφέρει, ουσιαστικά, τη γνώση που αφορά τις συνήθειες των καταναλωτών ανά περίοδο, με σκοπό τόσο την ελαχιστοποίηση κινήσεων που περιέχουν ρίσκο, όσο και κατ' επέκταση την αποκόμιση όλο και μεγαλύτερων κερδών. Με τη βοήθεια της πληθώρας δεδομένων, λοιπόν, που διατίθενται μέσω διάφορων οργανισμών παγκοσμίως, η εξόρυξη προτύπων έρχεται να απαντήσει σε ερωτήσεις όπως :

- Ποιο προϊόν να προσφέρουμε στους πελάτες?
- Ποιες κατηγορίες πελατών θα επηρεαστούν και θα ανταποκριθούν σε μία συγκεκριμένη διαφημιστική εκστρατεία?
- Πως θα αναγνωρίσουμε τους πελάτες οι οποίοι δεν ακολουθούν πλέον τις υπηρεσίες μας και ποιος ο λόγος που συνέβη αυτό?
- Ποια κατηγορία πελατών είναι περισσότερη σημαντική και πως να την κρατάμε ευχαριστημένη?

Πεδίο Ανάλυσης Βιολογικών Δεδομένων (Βιοπληροφορική)

Οι προσεγγίσεις της εξόρυξης προτύπων είναι ιδανικές και για των ευρύτερο χώρο των Φυσικών και Θετικών Επιστημών, οι οποίες δεν θα μπορούσαν να μην διαθέτουν αρκετά "πλούσια" δεδομένα. Μία από αυτές τις επιστήμες είναι η Βιοπληροφορική, στη διαχείριση και την ανάλυση των αποτελεσμάτων της οποίας, διαδραματίζει σημαντικό ρόλο η εξόρυξη δεδομένων. Τα βιολογικά εξαγόμενα πρότυπα, όπως θα μπορούσε κανείς να τα χαρακτηρίσει, συμβάλουν σε αρκετές πτυχές της εν λόγω επιστήμης, κάποιες απο τις οποίες είναι :

- Η σημασιολογική ενοποίηση ετερογενών, κατανεμημένων γονιδιωμάτων και πρωτεομικών βάσεων δεδομένων.
- Η στοίχιση, ευρετηριοποίηση, αναζήτηση ομοιότητας και συγκριτική ανάλυση πολλαπλών αλληλουχιών νουκλεοτιδίων.
- Η ανακάλυψη δομημένων προτύπων και η ανάλυση γενετικών δικτύων και μονοπατιών πρωτεΐνης.

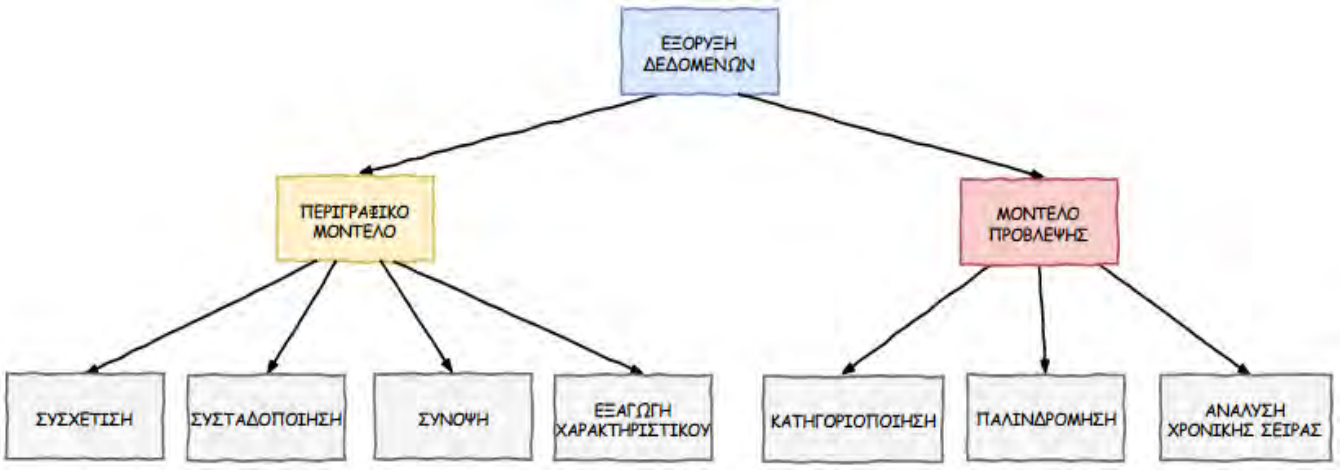
2.4 ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Έχοντας κάνει λόγο για τη διαδικασία ανακάλυψης γνώσης μέσα από βάσεις, το αρχικό της βήμα, την προεπεξεργασία των δεδομένων, και έχοντας παρουσιάσει κάποιες από τις πολλές υπηρεσίες που προσφέρει η εξόρυξη δεδομένων στον πραγματικό κόσμο, ήρθε η στιγμή να δούμε από μία πιο κοντινή οπτική γωνία το ποιες είναι οι μέθοδοι που χρησιμοποιεί η εν λόγω διαδικασία.

Οι μέθοδοι αυτοί δεν είναι κάτι άλλο από αλγοριθμικές προσεγγίσεις-μοντέλα που όπως προαναφέρθηκε ναι μεν παρουσιάζουν υψηλού επιπέδου λειτουργικότητα, αλλά στα πλαίσια της εξέλιξης, εξακολουθούν να αντιμετωπίζουν διάφορες προκλήσεις που έχουν να κάνουν με την εύρεση όλο και ακριβέστερων αποτελεσμάτων σε συνδυασμό πάντα με υψηλή ταχύτητα και χαμηλό υπολογιστικό κόστος. Συνεπώς, όπως εύκολα μπορεί να γίνει κατανοητό, στις διάφορες μεθόδους της εξόρυξης δεδομένων, παίζουν αρκετά σημαντικό ρόλο οι δομές δεδομένων.

Οι μέθοδοι που χρησιμοποιεί η εξόρυξη δεδομένων, προκειμένου να φτάσει στον τελικό της στόχο, μπορούν να χωριστούν σε δύο ευρύτερες κατηγορίες - μοντέλα προσέγγισης προβλήματος, με βάση το τι ακριβώς πληροφορία θέλουμε να εξάγουμε κάθε φορά : στο **περιγραφικό (descriptive) μοντέλο** και στο **μοντέλο πρόβλεψης (predictive)**. Οι μέθοδοι του περιγραφικού μοντέλου, ασχολούνται με τις γενικές ιδιότητες της παρουσίας των δεδομένων σε μία βάση δεδομένων, ενώ αντίθετα, οι μέθοδοι του μοντέλου πρόβλεψης,

χρησιμοποιούν τιμές ενός διαθέσιμου συνόλου, προκειμένου να προβλέψουν μη γνωστές ή μελλοντικές τιμές ενός άλλου, προφανώς σχετικού, συνόλου δεδομένων ενδιαφέροντος. Όπως φαίνεται και στην εικόνα που ακολουθεί, το **περιγραφικό μοντέλο** εξάγει πρότυπα βάσει των μεθόδων : **α) συσχέτισης, β) συσταδοποίησης, γ) σύνοψης και δ) εξαγωγής χαρακτηριστικού**, ενώ το **μοντέλο πρόβλεψης**, περιλαμβάνει τις μεθόδους : **α) κατηγοριοποίησης, β) παλινδρόμησης και γ) ανάλυσης χρονικής σειράς.**



Εικόνα 2.2 – Κατηγοριοποίηση μεθόδων εξόρυξης δεδομένων

Στην προσπάθεια, να μην ξεφύγουμε από το βασικό σκοπό του εν λόγω συγγράμματος, οι έννοιες που ακολουθούν πέρα της μεθόδου **συσχέτισης**, θα δοθούν όσο πιο συνοπτικά, αλλά συνάμα και κατανοητά, γίνεται.

2.4.1 ΜΟΝΤΕΛΟ ΠΡΟΒΛΕΨΗΣ

Ο βασικός σκοπός του μοντέλου πρόβλεψης, όπως αναφέρει και η ονομασία του, δεν είναι άλλος από την πρόβλεψη μελλοντικών τιμών-αποτελεσμάτων με βάση κάποια τρέχουσα συμπεριφορά γνωστών δεδομένων. Το μοντέλο πρόβλεψης, εκτός απο τις υπηρεσίες που προσφέρει στον κλάδο της μετεωρολογίας και την πρόγνωση του καιρού, χρησιμοποιείται ευρέως και σε μία πληθώρα άλλων λειτουργιών, όπως η ανίχνευση spamming ή απάτης και το στοχευόμενο marketing.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ (Classification)

Σε ένα πρόβλημα κατηγοριοποίησης, τυπικά, έχουμε ιστορικά-γνωστά δεδομένα που καλούνται παραδείγματα ετικέτας, όπως επίσης και κάποια νέα δεδομένα. Κάθε παράδειγμα ετικέτας περιέχει πολλαπλές ιδιότητες που μπορούν να χρησιμοποιηθούν για πρόβλεψη και μία

στοχευόμενη ιδιότητα η οποία καλείται εξαρτημένη μεταβλητή. Η τιμή της στοχευόμενης ιδιότητας καλείται ως ετικέτα κλάσης, όπου τα νέα δεδομένα εμπεριέχουν όλες τις άλλες ιδιότητες του παραδείγματος ετικέτας, εκτός, όμως, αυτής. Ο στόχος, λοιπόν, των διάφορων αλγοριθμικών μεθόδων της κατηγοριοποίησης είναι να κατασκευάσει ένα μοντέλο χρησιμοποιώντας γνωστά δεδομένα, όπου θα προβλέπει την κλάση ετικέτας των νέων δεδομένων. Γνωστοί αλγόριθμοι κατηγοριοποίησης, είναι ο αλγόριθμος του Naive Bayes και ο προσαρμοσμένος αλγόριθμος δικτύου Bayes.

ΠΑΛΙΝΔΡΟΜΗΣΗ (Regression)

Η παλινδρόμηση, ουσιαστικά, κάνει την ίδια δουλειά με τη κατηγοριοποίηση με μόνη βασική διαφορά ότι η παλινδρόμηση ασχολείται με αριθμητικές/συνεχόμενες στοχευόμενες ιδιότητες, τη στιγμή που η κατηγοριοποίηση ασχολείται με διακριτές/κατηγο-ρηματικές στο-χευόμενες ιδιότητες. Με άλλα λόγια, εάν η στοχευόμενη ιδιότητα περιλαμβάνει floating-point τιμές τότε απαιτείται παλινδρόμηση, ενώ αν περιλαμβάνει string ή διακριτές integer τιμές τότε απαιτείται κατηγοριοποίηση.

ΑΝΑΛΥΣΗ ΧΡΟΝΙΚΗΣ ΣΕΙΡΑΣ (Time-Series Analysis)

Οι σειρές χρόνου (time series), είναι σειρές από δεδομένα γεγονότων (events), καταχωρημένα σε μία δομή ευρετηρίου και τοποθετημένα με χρονική σειρά, όπου το κάθε επόμενο γεγονός προσδιορίζεται από ένα ή περισσότερα προηγούμενα γεγονότα. Όπως, λοιπόν εύκολα μπορεί να γίνει αντιληπτό, η ανάλυση των σειρών χρόνου περιλαμβάνει μεθόδους για την ανάλυση των εν λόγω δεδομένων, προκειμένου να εξάγει χρήσιμα πρότυπα, τάσεις, κανόνες και στατιστικές.

2.4.2 ΠΕΡΙΓΡΑΦΙΚΟ ΜΟΝΤΕΛΟ

Η δεύτερη προσέγγιση εξόρυξης δεδομένων από μεγάλα σύνολα δεδομένων, είναι γνωστή ως περιγραφική εξόρυξη δεδομένων όπου σε αντίθεση με τις μεθόδους πρόβλεψης, εστιάζει περισσότερο σε εσωτερικές δομές, σχέσεις και διασυνδεσιμότητες των δεδομένων. Το περιγραφικό μοντέλο, με τη μαθηματική έννοια του όρου, είναι η διαδικασία η οποία περιγράφει γεγονότα του πραγματικού κόσμου και τις σχέσεις μεταξύ των παραγόντων που είναι υπεύθυνοι γι' αυτά, λειτουργικότητα που προφανώς δεν διαφέρει ως προς τις μεθόδους εξόρυξης που θα συζητηθούν ακολούθως.

ΣΥΣΤΑΔΟΠΟΙΗΣΗ (Clustering)

Η **συσταδοποίηση**, με πολύ απλά λόγια, είναι η διαδικασία κατά την οποία βρίσκουμε διάφορες λογικές ομάδες(συστάδες-**clusters**) σε μία βάση δεδομένων. Με τον όρο "λογικές ομάδες" εννοούμε τα σύνολα των αντικειμένων τα οποία είναι όμοια ή σχετίζονται μεταξύ τους με βάση κάποια συγκεκριμένα χαρακτηριστικά, ενώ ταυτόχρονα είναι διαφορετικά ή μη συσχετιζόμενα από αντικείμενα άλλων ομάδων.

Εικόνα 2.3 - Συσταδοποίηση

Ενδεικτικά, σε επίπεδο εφαρμογής στον πραγματικό κόσμο, μπορούμε να κάνουμε λόγο για συσταδοποίηση :

- γονιδίων και πρωτεϊνών που έχουν την ίδια λειτουργία
- εικόνων επιπέδου βροχής
- μετοχών με παρόμοια διακύμανση τιμών
- πελατών με παρόμοια συμπεριφορά

Λόγω, λοιπόν, του ότι μπορούμε να κάνουμε λόγο για πολλών και διαφόρων ειδών συστάδες, οι ερευνητές έχουν αναπτύξει διάφορους αλγόριθμους κατά καιρούς, οι οποίοι με τη σειρά τους πρόσκεινται σε διαφορετικά μοντέλα προσέγγισης. Κάποια εξ αυτών, σε συνδυασμό με δημοφιλείς μεθόδους τους, είναι τα εξής:

- Μοντέλα κεντρικού σημείου : όπου ο αλγόριθμος k-means διαμοιράζει τα διάφορα σημεία σε συστάδες βάσει του κοντινότερου κεντρικού σημείου τους.
- Μοντέλα συνδεσιμότητας : όπου η ιεραρχική συσταδοποίηση χτίζει συστάδες βάσει απόστασης σύνδεσης.
- Κατανεμημένα μοντέλα : όπου οι συστάδες μοντελοποιούνται χρησιμοποιώντας στατιστικές κατανομές, όπως η πολυμεταβλητή κανονική κατανομή που χρησιμοποιείται από τον στατιστικό-υπολογιστικό αλγόριθμο expectation-maximization.

ΣΥΝΟΨΗ (Summarization)

Λόγω του ότι τα δεδομένα, στις αποθήκες δεδομένων (data warehouse) είναι πολύ μεγάλου όγκου, τίθενται ανάγκες ύπαρξης ενός μηχανισμού ο οποίος θα παίρνει μόνο τη σχετική και μεγάλης σημασίας πληροφορία και θα την αποδίδει σε μία καλύτερα οργανωμένη μορφή. Η διαδικασία της **σύνοψης** δεδομένων, λοιπόν, βοηθάει ακριβώς ως προς αυτή τη κατεύθυνση. Με άλλα λόγια, η σύνοψη χωρίζει τα δεδομένα σε υποσύνολα με απλές περιγραφές, δίνοντας έτσι μία πιο σφαιρική εικόνα για τις πληροφορίες που αυτά περιέχουν. Τέλος, λόγω του ότι η εν λόγω διαδικασία μπορεί να κλιμακωθεί σε διαφορετικά επίπεδα αφαίρεσης και να προσεγγιστεί από διάφορες οπτικές γωνίες, χρησιμοποιούνται διάφορες τεχνικές οι οποίες εφαρμόζονται σε αναλύσεις δεδομένων, οπτικοποιήσεις δεδομένων και αυτόματες αναφορές.

ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ (Feature Extraction)

Η διαδικασία εξαγωγής χαρακτηριστικών δημιουργεί νέα σύνολα χαρακτηριστικών βάσει των ιδιοτήτων ενός πρωτότυπου συνόλου δεδομένων. Αυτό χρησιμεύει στην περιγραφή δεδομένων μέσω ενός αριθμού από χαρακτηριστικά, πολύ μικρότερου απο τον αριθμό των πρωτότυπων ιδιοτήτων. Για περισσότερη κατανόηση, ας θεωρήσουμε ότι μία ταινία περιλαμβάνει "βία", "ηρωισμό" και "γρήγορα αυτοκίνητα" ως χαρακτηριστικές της ιδιότητες, τη στιγμή που το είδος-θέμα-κύριο εννοιολογικό χαρακτηριστικό της μπορεί να είναι η λέξη "περιπέτεια". Αυτή η ιδέα, λοιπόν, μπορεί να χρησιμοποιηθεί για την εξαγωγή του θέματος ενός αρχείου βάσει της συχνότητας συγκεκριμένων λέξεων-κλειδιών, όπως και σε άλλες περιπτώσεις, όπως η συμπίεση δεδομένων και η αναγνώριση προτύπων.

ΣΥΣΧΕΤΙΣΗ (Association)

Η συσχέτιση, αποτελεί τη μέθοδο της εξόρυξης δεδομένων η οποία ανακαλύπτει τη πιθανότητα συνύπαρξης δύο αντικειμένων μέσα σε ένα σύνολο συναλλαγών. Οι σχέσεις μεταξύ των συνυπαρχόντων αντικειμένων είναι γνωστές ως **κανόνες συσχέτισης.** Οι κανόνες συσχέτισης συνδέονται σε μεγάλο βαθμό με την ανάλυση συναλλαγών πωλήσεων και παίζουν σημαντικό ρόλο για λειτουργίες όπως η οργάνωση των ραφιών ή των φυλλαδίων ενός super market. Ένα κλασσικό παράδειγμα ανάλυσης, άλλωστε, που συναντάμε στη βιβλιογραφία περί κανόνων συσχέτισης είναι το εξής : "Την Παρασκευή το απόγευμα, οι νέοι άντρες Αμερικάνοι που αγοράζουν πάνες για βρέφη, έχουν τη προδιάθεση να αγοράζουν και μπύρα", το οποίο εν τη ρύμη του λόγου μπορεί να έχει μία λογική υπόσταση, αλλά κατά πάσα πιθανότητα δεν είναι πραγματικότητα.

Λόγου του ότι οι κανόνες συσχέτισης αποτελούν ένα από τα κεντρικά σημεία του εν λόγω συγγράμματος, περισσότερες λεπτομέρειες γι' αυτούς στην ενότητα που ακολουθεί.

2.5 ΑΝΑΛΥΣΗ ΣΥΣΧΕΤΙΣΗΣ

Σε καθημερινή βάση κάθε επιχείρηση μαζεύει τεράστιες ποσότητες δεδομένων, όπως για παράδειγμα τις συναλλαγές που πραγματοποιήθηκαν κατά τη διάρκεια της μέρας. Η ανάλυση των δεδομένων αυτών και η εξαγωγή γνώσης συσχετίσεων που, όπως μόλις ειπώθηκε, αφορά σε μεγάλο βαθμό τις αγοραστικές συνήθειες του κοινού, αποτελεί επιτακτική ανάγκη και συνδέεται άμεσα με το οργανόγραμμα των επιχειρήσεων αυτών.

Ένα συμπέρασμα όπως αυτό που αναφέρθηκε στη προηγούμενη υποενότητα, το ότι, δηλαδή, μία συγκεκριμένη μέρα και ώρα, μία συγκεκριμένη ομάδα ατόμων(με βάση το φύλλο και την ηλικία) έχει τη τάση να αγοράζει ένα συγκεκριμένο συνδυασμό προϊόντων και δη όσον αφορά το τελευταίο σκέλος αυτής της πρότασης, αποτελεί αντικείμενο ανάλυσης κανόνων συσχέτισης των οποίων η μορφή έχει ως εξής :

{Πάνες} ⟶ **{Μπύρα}** **[0,8]**

που σημασιολογικά μπορεί να αποδοθεί ως:

"Στο 80% των συναλλαγών κατά τις οποίες αγοράστηκαν πάνες, αγοράστηκε και μπύρα."

Ο κανόνας αυτός, λοιπόν, έρχεται να συνδυαστεί με μία πληθώρα άλλων κανόνων, διαμορφώνοντας σε μεγάλο βαθμό κινήσεις που αφορούν τόσο την πώληση όσο και την προώθηση. Εκτός, βέβαια, από λειτουργίες που αφορούν την οικονομία γενικότερα, όπως γίνεται αντιληπτό, η ανάλυση συσχετίσεων βρίσκει χώρο και σε αρκετά σημεία τόσο των φυσικών όσο και των θετικών επιστημών.

2.5.1 ΑΝΑΛΥΣΗ ΚΑΛΑΘΙΟΥ SUPER-MARKET

Για να εντρυφήσουμε περισσότερο στο θέμα της εξαγωγής κανόνων συσχέτισης, καλό θα ήταν να συνεχίσουμε τη συζήτησή μας με ένα αρκετά δημοφιλές παράδειγμα, αυτό των συναλλαγών του καλαθιού του super market (**market-basket**). Ας θεωρήσουμε, λοιπόν, πως ο πίνακας (2.1) που ακολουθεί, αποτελεί ένα δείγμα συναλλαγών-αποδείξεων με τα αντίστοιχα αναγνωριστικά τους.

ID	ΣΥΝΑΛΛΑΓΕΣ
1	{Γάλα, Καφές , Ψωμί, Τυρί}
2	{Δημητριακά, Γάλα, Καφές, Τυρί}
3	{Δημητριακά, Καφές, Ψωμί, Αυγά}
4	{Γάλα, Δημητριακά}
5	{Δημητριακά, Γάλα, Καφές, Ψωμί}

Πίνακας 2.1 Δείγμα συναλλαγών super-market

ΠΡΟΚΛΗΣΕΙΣ

Οι δύο βασικές προκλήσεις που έρχεται να αντιμετωπίσει η ανάλυση συσχετίσεων, όπως εν μέρει κάθε μέθοδος εξόρυξης δεδομένων, είναι :

- Η όσο το δυνατόν **ελαχιστοποίηση υπολογιστικού κόστους**, λόγω του ότι η εξόρυξη προτύπων μέσα από πολύ μεγάλες μεγάλα σύνολα δεδομένων παρουσιάζει μεγάλη υπολογιστική ακρίβεια, ειδικότερα με **brute force** (απλοποιημένες) προσεγγίσεις.
- Η **απομάκρυνση προτύπων τα οποία δεν θεωρούνται σημαντικά** και τα οποία μπορεί να προκύπτουν από ένα συνδυασμό τυχαίων παραγόντων.

ΟΡΙΣΜΟΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Ένας κανόνας συσχέτισης προκύπτει μέσω προτύπων-συνόλων δεδομένων τα οποία συναντάμε συχνά μέσα σε μία βάση συναλλαγών. Ο όρος "συχνά" πρόσκειται σε δύο τιμές κατωφλίου, οι οποίες θα αναφερθούν στη συνέχεια.

Σε αυτό το σημείο, λοιπόν, "ρίχνοντας" συχνές ματιές στον πίνακα 2.1, θα θέσουμε τις βασικές έννοιες του προβλήματος κανόνων συσχέτισης με βάση την προσέγγιση των R.Agrawal, T.Imielinski και A.Swami [6].

- Έστω $\mathbf{I=\{i_1,i_2,...,i_d\}}$ το σύνολο με όλα τα διακριτά στοιχεία **(item**s) των δεδομένων μας. Στο παράδειγμά μας το σύνολο αυτό είναι το εξής: I={Γάλα, Καφές, Ψωμί, Τυρί, Δημητριακά, Αυγά}

- Έστω **$T=\{t_1,t_2,...,t_N\}$** ένα σύνολο από **συναλλαγές (transactions)**, όπου κάθε συναλλαγή **t_i** αποτελεί ένα **k-στοιχειοσύνολο (k-itemset)** και είναι υποσύνολο του **I**. Το k αντιπροσωπεύει το πλάτος/αριθμό στοιχείων του εκάστοτε συνόλου. π.χ. το {Γάλα, Καφές} είναι ένα 2-στοιχειοσύνολο.

- **Μετρητής υποστήριξης (support count(σ)) στοιχειοσυνόλου** : Αντιπροσωπεύει τον αριθμό εμφάνισης ενός στοιχειοσυνόλου στο σύνολο των συναλλαγών. π.χ. σ({Γάλα, Καφές, Τυρί}) = 2.

- **Συχνό στοιχειοσύνολο (frequent/large itemset)** : Θεωρείται κάθε στοιχειοσύνολο του οποίου η υποστήριξη είναι μεγαλύτερη ή ίση από μία τιμή κατωφλίου(minsup).

- **Κανόνες συσχέτισης (association rules)** : Προκύπτουν μέσω των συχνών στοιχειοσυνόλων και είναι της μορφής **X → Y**, όπου το X και το Y είναι ασύνδετα(disjoint) στοιχειοσύνολα, δηλαδή, **$X \cap Y = \emptyset$**. Το πόσο **ισχυρός** είναι ένας κανόνας, αυτό εξαρτάται με το κατά πόσο η δική του **υποστήριξη** και η **εμπιστοσύνη (confidence)** του έχουν μεγαλύτερη ή ίση τιμή από τις αντίστοιχες τιμές κατωφλίου (minsup και minconf).

- **Υποστήριξη (support(s)) κανόνα** : Αντιπροσωπεύει το ποσοστό των συναλλαγών που περιέχουν την ένωση του X και του Y, δηλαδή:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

- **Εμπιστοσύνη (confidence(c)) κανόνα** : Αντιπροσωπεύει το ποσοστό των συναλλαγών που περιέχουν το X, περιέχουν και το Y, δηλαδή:

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Χάριν καλύτερης κατανόησης των δύο τελευταίων εννοιών, ας υποθέσουμε πως έχουμε τον κανόνα {Γάλα, Καφές} → {Δημητριακά}. Η υποστήριξη(s) και η εμπιστοσύνη(c) του, έχουν ως εξής :

$$s(\{\text{Γάλα},\text{Καφές}\} \rightarrow \text{Δημητριακά}\}) = \frac{\sigma(\{\text{Γάλα},\text{Καφές},\text{Δημητριακά}\})}{N} = \frac{2}{5}$$

$$c(\{\text{Γάλα},\text{Καφές}\} \rightarrow \text{Δημητριακά}\}) = \frac{\sigma(\{\text{Γάλα},\text{Καφές},\text{Δημητριακά}\})}{\sigma(\{\text{Γάλα},\text{Καφές}\})} = \frac{2}{3}$$

Οι μετρήσεις της **υποστήριξης** και της **εμπιστοσύνης**, όπως εύκολα μπορεί να καταλάβει κανείς, παίζουν σημαντικό ρόλο καθ' όλη τη διαδικασία εξόρυξης κανόνων συσχέτισης. Ένας κανόνας με μικρή υποστήριξη μπορεί να έχει προκύψει απλά κατά τύχη και έτσι δεν μπορεί να θεωρηθεί σημαντικός. Όπως, επίσης, και λόγω του ότι ουσιαστικά οι κανόνες συσχέτισης ψάχνουν για εξαρτήσεις μεταξύ των διάφορων προτύπων, όσο υψηλότερη εμπιστοσύνη έχει ένας κανόνας τόσο μεγαλύτερης σημασίας μπορεί να θεωρηθεί. Οι εν λόγω μετρήσεις, επίσης, όπως θα δούμε και στη συνέχεια βοηθούν και στην ελάττωση του υπολογιστικού κόστους της διαδικασίας.

Το πρόβλημα της ανακάλυψης κανόνων συσχέτισης, λοιπόν, μπορεί να χωριστεί σε δύο υποπροβλήματα :

1. Βρες όλα τα **συχνά στοιχειοσύνολα** που εμπεριέχονται στις συναλλαγές και
2. με βάση αυτά, δημιούργησε τους επιθυμητούς **ισχυρούς κανόνες**.

2.5.2 ΔΗΜΙΟΥΡΓΙΑ ΣΥΧΝΩΝ ΣΤΟΙΧΕΙΟΣΥΝΟΛΩΝ

Η δημιουργία συχνών στοιχειοσυνόλων, όπως γίνεται αντιληπτό σε σχέση με όσα ειπώθηκαν μόλις προηγουμένως, προϋποθέτει τον υπολογισμό των μετρητών υποστήριξής τους, διαδικασία που όταν έχουμε να κάνουμε με έναν μεγάλο όγκο δεδομένων είναι υπολογιστικά ακριβή. Αυτό είναι κάτι που δεν ισχύει άμεσα για τους κανόνες συσχέτισης καθώς όταν αυτοί ουσιαστικά δημιουργούνται οι μετρητές υποστήριξης των στοιχειοσυνόλων είναι ήδη γνωστοί.

Ειδικότερα, ένα σύνολο δεδομένων που περιέχει k items μπορεί να παράξει 2^k-1 συχνά στοιχειοσύνολα, εκτός του κενού συνόλου. Λόγω του ότι αυτό το k μπορεί να είναι πολυ μεγάλο, λοιπόν, η μείωση του υπολογιστικού κόστους της διαδικασίας δημιουργίας συχνών στοιχειοσυνόλων αποτελεί ουσιαστικά τη "νούμερο ένα" πρόκληση που καλούνται να αντιμετωπίσουν οι διάφοροι ερευνητές στη βιβλιογραφία, με πρωτεύοντα εργαλεία τη χρησιμοποίηση : α)τεχνικών κλαδέματος όπως γίνεται στον αλγόριθμο **Apriori**[6] με σκοπό τη μείωση των υποψήφιων στοιχειοσυνόλων και β)αποτελεσματικών δομών δεδομένων όπως δέντρα κατακερματισμού[7], tries[8] ή όπως η FP-tree δομή[35] με σκοπό τη μείωση των συγκρίσεων.

2.5.3 ΠΡΟΣΕΓΓΙΣΗ APRIORI

ΑΡΧΗ APRIORI

Ο **Apriori**[6], όπως θα μπορούσε να πει κανείς, αποτελεί τον ακρογωνιαίο λίθο των αλγόριθμων εξόρυξης κανόνων συσχέτισης, όντας ο πρώτος που χρησιμοποίησε τη **τεχνική κλαδέματος βάσει υποστήριξης** εκμεταλλευόμενος την εξής αρχή : Αν ένα στοιχειοσύνολο είναι συχνό, τότε όλα τα υποσύνολά του είναι συχνά ή αντίστροφα, αν ένα στοιχειοσύνολο δεν είναι συχνό, τότε κανένα του υπερσύνολό δεν είναι συχνό.

Βάσει αυτής της αρχής προκύπτει και η εξής **αντιμονότονη ιδιότητα**

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

που σημαίνει ότι η υποστήριξη ενός στοιχειοσυνόλου είναι μεγαλύτερη ή ίση από την υποστήριξη οποιουδήποτε υπερσυνόλου του. Έτσι, λοιπόν, κάθε στοιχειοσύνολο το οποίο έχει υποστήριξη ≤ minsup, κλαδεύεται και δεν λαμβάνεται υπόψη στην υπόλοιπη διαδικασία.

ΑΛΓΟΡΙΘΜΟΣ APRIORI

Έχοντας θέσει τις βασικές έννοιες γύρω από την εξόρυξη κανόνων συσχέτισης, μπορούμε να συνεχίσουμε τη συζήτησή μας με την παρουσίαση του αλγόριθμου Apriori[7], ο οποίος εστιάζει στη **δημιουργία συχνών στοιχειοσυνόλων** και είναι ο εξής :

1) $L_1 = \{\text{large 1-itemsets}\}$;
2) **for** ($k = 2$; $L_{k-1} \neq \emptyset$; $k{+}{+}$} **do begin**
3) $C_k = \text{apriori-gen}(L_{k-1})$; //New candidates
4) **forall** transactions $t \in D$ **do begin**
5) $C_t = \text{subset}(C_k, t)$; //Candidates contained in t
6) **forall** candidates $c \in C_t$ **do**
7) c.count++;
8) **end**
9) $L_k = \{c \in C_k | c.\text{count} \geq \text{minsup}\}$
10) **end**
11) Answer = $\bigcup_k L_k$;

Αλγόριθμος 2.1 Apriori

Με μία πρόχειρη ματιά γίνεται αντιληπτό πως με βάση τον εν λόγω αλγόριθμο, τα συχνά (frequent) ή αλλιώς μεγάλα (large) k-στοιχειοσύνολα που χρειαζόμαστε προκειμένου να δημιουργήσουμε τους κανόνες συσχέτισης, προκύπτουν μέσω **υποψήφιων** (**candidate**) k-στοιχειοσυνόλων.

Αναλυτικότερα, τα βήματα του Apriori (θεωρώντας ως γνωστή μία τιμή κατωφλίου minsup), είναι τα εξής :

- Αρχικά και ακριβώς πρωτού ξεκινήσει η επαναληπτική διαδικασία, γίνεται ένα μοναδικό πέρασμα στα δεδομένα προκειμένου να αποσαφηνιστεί το ποια από αυτά είναι συχνά και να δημιουργηθεί το σύνολο L_1.
- Ακολούθως, μέσω της συνάρτησης **apriori-gen**, δημιουργούνται τα νέα υποψήφια k-στοιχειοσύνολα (C_k) βάσει των συχνών (k-1)-στοιχειοσυνόλων.
- Στα βήματα 4 έως 8, υπολογίζεται η υποστήριξη των υποψήφιων στοιχειοσυνόλων με τη βοήθεια της συνάρτησης **subset** η οποία χρησιμοποιείται για τον "εντοπισμό" αυτών μέσα στη κάθε συναλλαγή.
- Τέλος, κλαδεύονται τα υποψήφια στοιχειοσύνολα των οποίων η υποστήριξη είναι μικρότερη από τη τιμή κατωφλίου minsup και δημιουργούνται τα εκάστοτε νέα συχνά στοιχειοσύνολα.
- Η διαδικασία επαναλαμβάνεται όσο υπάρχουν νέα συχνά στοιχειοσύνολα, δηλαδή όσο $L_{k-1} \neq \emptyset$.

ΔΗΜΙΟΥΡΓΙΑ ΥΠΟΨΗΦΙΩΝ ΣΤΟΙΧΕΙΟΣΥΝΟΛΩΝ

Ως επί το πλείστον, η συνάρτηση **apriori-gen** δημιουργεί υποψήφια στοιχειοσύνολα χρησιμοποιώντας την μέθοδο $\mathbf{F_{k-1}}$ **x** $\mathbf{F_{k-1}}$ κατά την οποία τα συχνά (κ-1)-στοιχειοσύνολα συγχωνεύονται ανά ζεύγη, με την προϋπόθεση ότι τα πρώτα κ-2 στοιχεία τους είναι ίδια. Σε SQL-like συντακτικό, η εν λόγω συνάρτηση έχει ως εξής :

insert into C_k
select $p.\text{item}_1, p.\text{item}_2, \ldots, p.\text{item}_{k-1}, q.\text{item}_{k-1}$
from $L_{k-1}\ p, L_{k-1}\ q$
where $p.\text{item}_1 = q.\text{item}_1, \ldots, p.\text{item}_{k-2} = q.\text{item}_{k-2}, p.\text{item}_{k-1} < q.\text{item}_{k-1}$;

ΠΑΡΑΤΗΡΗΣΕΙΣ:

◦Προς αποφυγή δημιουργίας διπλότυπων εγγραφών, κρατάμε κάθε στοιχειοσύνολο λεξικογραφικά ταξινομημένο.

◦Μετά το πέρας της βασικής διαδικασίας, η συνάρτηση εκτελεί

κλάδεμα στα στοιχειοσύνολα των οποίων τα (k-1)-υποσύνολα δεν βρίσκονται στα ήδη γνωστά συχνά (k-1)-στοιχειοσύνολα.

ΥΠΟΛΟΓΙΣΜΟΣ ΥΠΟΣΤΗΡΙΞΗΣ

Τα υποψήφια k-στοιχειοσύνολα αποθηκεύονται σε ένα **δέντρο κατακερματισμού (hash-tree)** του οποίου οι εσωτερικοί κόμβοι αποτελούν πίνακες κατακερματισμού βάθους d και δείχνουν σε κόμβους βάθους d+1. Οι κόμβοι αυτοί, με τη σειρά τους, αποτελούν είτε άλλους εσωτερικούς κόμβους (πανομοιότυπης δομής) ή φύλλα, τα οποία φύλλα αποτελούν λίστες όπου και αποθηκεύονται τα υποψήφια στοιχειοσύνολα. Κάθε στοιχειοσύνολο που εισάγεται διασχίζει το δέντρο προς τα κάτω, ξεκινώντας από τη ρίζα (βάθους 1) και ακολουθώντας τους εκάστοτε κόμβους βάθους j+1 με βάση τη συνάρτηση κατακερματισμού στο j^{th} στοιχείο τους, μέχρι να φτάσει σε κάποιο φύλλο. Όταν ένα φύλλο υπερβαίνει ένα καθορισμένο όριο, τότε μετατρέπεται σε εσωτερικό κόμβο και τα στοιχειοσύνολα που περιέχει διανέμονται σε νέα φύλλα βάσει τις j^{th} τιμές κατακερματισμού.

Μέσω της συνάρτησης **subset**, λοιπόν, τα στοιχεία των συναλλαγών κατακερματίζονται στους αντίστοιχους κάδους και αν και εφόσον υπάρξει πλήρης αντιστοίχιση με όλα τα στοιχεία ενός υποψήφιου στοιχειοσυνόλου, τότε αυτό αποθηκεύεται στο ενδιάμεσο σύνολο C_t. Στην επανάληψη που ακολουθεί, κάθε υποψήφιο στοιχειοσύνολο που βρίσκεται στο C_t, αυξάνει την υποστήριξή του κατά 1. Τέλος, να τονιστεί ότι για να βρούμε αν υπάρχει πλήρης αντιστοίχιση μεταξύ κάποιου στοιχειοσυνόλου και κάποιας συναλλαγής, χρησιμοποιούνται κατάλληλοι ενδιάμεσοι μετρητές στα φύλλα του δέντρου.

Με αυτό το τρόπο, ουσιαστικά, αποφεύγεται η σύγκριση κάθε στοιχείου κάθε συναλλαγής με κάθε στοιχείο κάθε υποψήφιου στοιχειοσυνόλου, πράγμα που βοηθάει σε πολύ σημαντικό βαθμό στη μείωση του κόστους της διαδικασίας.

ΔΗΜΙΟΥΡΓΙΑ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ

Όσον αφορά τους κανόνες συσχέτισης, σε αντίθεση με τη μέτρηση της υποστήριξης, η μέτρηση της εμπιστοσύνης δεν έγκειται σε κάποια αντιμονότονη ιδιότητα εκτός και αν μιλάμε για σύγκριση κανόνων που δημιουργήθηκαν από το ίδιο στοιχειοσύνολο, όπου ισχύει το εξής :

- Εάν ο κανόνας $X \rightarrow Y-X$ δεν ικανοποιεί τη τιμή κατωφλίου minconf, τότε κάθε κανόνας $X' \rightarrow Y-X'$, όπου $X' \subseteq X$, δεν πρέπει να ικανοποιεί το όριο της εμπιστοσύνης, επίσης.

Για τη δημιουργία των κανόνων συσχέτισης, λοιπόν, οι δημιουργοί του Apriori χρησιμοποιούν μία προσέγγιση ενημέρωσης επιπέδου, όπου κάθε επίπεδο (**level**) αντιστοιχεί στον αριθμό των items που ανήκουν στη δεξιά ακολουθία (**RHS**) του εκάστοτε κανόνα.

Αρχικά εξάγουμε όλους τους κανόνες υψηλής εμπιστοσύνης του πρώτου επιπέδου και εν συνεχεία, σε πρώτη φάση με βάση αυτούς, όσο ισχύει **k > level-1** για τη δημιουργία level-RHS συγχωνεύουμε ανά ζεύγη τα (level-1)-RHS (αποθήκευση σε ένα set προς αποφυγή διπλότυπων εγγραφών) , ενώ τα level-LHS είναι αποτέλεσμα ένωσης των αντίστοιχων (level-1)-LHS. Όλη η διαδικασία αφορά μόνο κανόνες υψηλής εμπιστοσύνης, όπως επίσης και κανόνες οι οποίοι δεν περιέχουν σε κάποιο από τα δύο μέλη τους το κενό σύνολο. Τέλος, όπως προαναφέρθηκε στην υποενότητα 2.5.2 για τον υπολογισμό της εμπιστοσύνης δεν χρειάζεται να κάνουμε επιπλέον περάσματα στη βάση των συναλλαγών καθώς οι απαραίτητες μετρήσεις των υποστηρίξεων είναι ήδη γνωστές μέσω της δομής αποθήκευσης του εκάστοτε k-στοιχειοσυνόλου.

3.ΕΞΟΡΥΞΗ ΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

3.1 ΕΙΣΑΓΩΓΗ

Με τη μεγάλη τεχνολογική ανάπτυξη και την ευρεία χρήση των υπηρεσιών GIS και LBS σε μία πληθώρα εφαρμογών, δημιουργούνται ζητήματα τόσο καλύτερης κατανόησης του γεωγραφικού πλαισίου όσο και ανάπτυξης ειδικών τεχνικών διαχείρισης μεγάλου όγκου χωρικών δεδομένων με σκοπό την παροχή ανάλυσης και αυτοματοποιημένης γνώσης υψηλής ποιότητας. Τα ζητήματα αυτά δημιουργούνται καθώς η εξόρυξη χρήσιμων και ενδιαφερόντων προτύπων από βάσεις χωρικών δεδομένων δεν μπορεί να θεωρηθεί ίδια διαδικασία με την εξόρυξη προτύπων από βάσεις παραδοσιακών αριθμητικών και κατηγορηματικών δεδομένων. Αυτό συμβαίνει καθώς η πρώτη παρουσιάζει περισσότερες δυσκολίες διαχείρισης λόγω της πολυπλοκότητας των χωρικών τύπων, σχέσεων και συσχετίσεων. Έτσι, λοιπόν, στα τέλη της δεκαετίας του 90' έγινε ένας πρώτος διαχωρισμός των κλασσικών προσεγγί-σεων της εξόρυξης δεδομένων με αυτές της εξόρυξης χωρικών δεδομένων [9],[10].

Η **εξόρυξη χωρικών δεδομένων (Spatial Data Mining)** αποτελεί την επέκταση-προσαρμογή των μεθόδων της υπολογιστικής διαδικασίας της **εξόρυξης δεδομένων (Data Mining)** στις έννοιες των χωρικών δεδομένων.

Με την πιο αυστηρή έννοια και υιοθέτηση, ουσιαστικά, του παραδοσιακού όρου, η εξόρυξη χωρικών δεδομένων είναι η διαδικασία ανακάλυψης ενδιαφερόντων και μέχρι πρότινος μη γνωστών, αλλά άξιων αναφοράς και προσοχής χωρικών προτύπων.

Όπως γίνεται αντιληπτό, το εν λόγω κεφάλαιο αποτελεί έναν συνδυασμό των όσων συζητήθηκαν έως τώρα. Για συντομία και εξοικονόμηση χώρου, από εδώ και στο εξής, σε πολλά σημεία, θα αναφερόμαστε στις έννοιες της εξόρυξης χωρικών δεδομένων και της κλασσικής εξόρυξης δεδομένων με τα αγγλικά τους ακρωνύμια, **SDM** και **DM** αντίστοιχα.

3.1.1 ΜΟΝΑΔΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΕΞΩΡΥΞΗΣ ΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Κάποια μοναδικά χαρακτηριστικά που κάνουν την εξόρυξη χωρικών δεδομένων να ξεχωρίζει από τις κλασσικές προσεγγίσεις των μεθόδων της εξόρυξης δεδομένων, είναι τα εξής :

- Τύπος δεδομένων : Όπως προαναφέρθηκε, η πρώτη βασική διαφορά μεταξύ των SDM και DM έγκειται στο γεγονός ότι τα δεδομένα που έχει να διαχειριστεί μία διαδικασία SDM είναι πιο περίπλοκα από εκείνα που διαχειρίζεται μία κλασσική διαδικασία DM, καθώς τα πρώτα περιέχουν επεκταμένα αντικείμενα όπως σημεία, γραμμές και πολύγωνα. Αυτό, προφανώς, δεν θα μπορούσε να μην έχει αντίκτυπο και στις σχέσεις μεταξύ των αντικειμένων, καθώς σε αντίθεση με τις σχέσεις μεταξύ αριθμητικών και αλφαριθμητικών αντικειμένων οι οποίες είναι συνήθως άμεσες και σαφείς, στην περίπτωση των χωρικών αντικειμένων υπάρχουν πιο έμμεσες σχέσεις καθώς προκύπτουν έννοιες όπως η απόσταση, η υπερκάλυψη, η τομή κ.α. Τις έμμεσες αυτές σχέσεις μπορούμε να τις αντιμετωπίσουμε ως εισόδους αριθμητικών ή κατηγορηματικών δεδομένων και να εφαρμόσουμε κλασσικές τεχνικές DM. Ωστόσο, με αυτόν το τρόπο μπορεί να χάσουμε πληροφορία. Έτσι, λοιπόν, η ανάπτυξη διάφορων τεχνικών για την ενσωμάτωση της γεωγραφικής πληροφορίας στη διαδικασία SDM, έρχεται για να δώσει τη λύση στο εν λόγω πρόβλημα.

- Στατιστική θεώρηση : Τα στατιστικά μοντέλα συχνά χρησιμοποιούνται για την αναπαράσταση παρατηρήσεων βάσει κάποιων τυχαίων μεταβλητών. Αυτά τα μοντέλα μπορούν να χρησιμοποιηθούν για εκτιμήσεις, περιγραφές και προβλέψεις ακολουθώντας τη θεωρία των πιθανοτήτων. Μία από τις θεμελιώδεις υποθέσεις της στατιστικής ανάλυσης είναι ότι τα δείγματα δεδομένων δημιουργούνται ανεξάρτητα. Παρ' όλα αυτά, στην ανάλυση χωρικών δεδομένων, η υπόθεση αυτή δεν μπορεί να ισχύει, καθώς αυτά παρουσιάζουν υψηλές συσχετίσεις. Τα χωρικά δεδομένα μπορούν να ληφθούν υπόψη ως ένα αποτέλεσμα παρατηρήσεων της στοχαστικής διεργασίας Z(s) : s 2 D, όπου s μία χωρική τοποθεσία και D ένα τυχαίο σύνολο ενός χωρικού πλαισίου. Τρία βασικά χωρικά στατιστικά προβλήματα είναι : η διαδικασία σημείου, το πλέγμα και η γεωστατιστική.

- Υπολογιστική διαδικασία : Για την εξόρυξη χωρικών δεδομένων έχουν υιοθετηθεί πολλές αλγοριθμικές στρατηγικές, όπως : διαίρει και βασίλευε, φιλτράρισμα και εκκαθάριση, διάταξη, ιεραρχική δομή και εκτίμηση παραμέτρου. Η χωρική αυτοσυσχέτιση και η χαμηλή διάσταση στο χώρο μας δίνουν περισσότερες ευκαιρίες για τη βελτίωση της υπολογιστικής αποδοτικότητας σε συνδυασμό με τις διάφορες δομές χωρικών ευρετηρίων που έχουν δημιουργηθεί, όπως η gist στην PostgreSQL .

3.1.2 ΧΡΗΣΙΜΟΤΗΤΑ SDM

Πολλά αποδοτικά εργαλεία εξαγωγής πληροφοριών από τεράστιους όγκους χωρικών δεδομένων παίζουν πολύ σημαντικό ρόλο σε λειτουργίες υπηρεσιών οι οποίες παίρνουν αποφάσεις με βάση αυτά, συμπεριλαμβανομένων της **NASA**, του Εθνικού Οργανισμού Απεικόνισης και Χαρτογράφησης (**NIMA**) και του Τμήματος Μεταφορών Ηνωμένων Πολιτειών (**USDOT**). Αυτοί οι οργανισμοί εξαπλώνονται σε πολλά πεδία εφαρμογών, όπως η οικολογία και η περιβαλλοντολογική διαχείριση, η δημόσια ασφάλεια, οι μεταφορές, η επιδημιολογία, η κλιματολογία, η διαχείριση γεγονότων έκτακτης ανάγκης κ.α. Επίσης, η εξόρυξη χωρικών δεδομένων χρησιμοποιείται και από άλλες υπηρεσίες οι οποίες ασχολούνται τόσο με το εμπόριο, όσο και με την οικονομία γενικότερα.

3.2 ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Όσον αφορά τις τεχνικές της SDM δεν υπάρχει κάποιος ενιαίος τρόπος βάσει τον οποίον αυτές κατηγοριοποιούνται, καθώς μέσα από βάσεις χωρικών δεδομένων μπορούν να ανακαλυφθούν πολλά είδη προτύπων, τα οποία μπορούν να αναπαρασταθούν σε διάφορες μορφές. Η κατηγοριοποίηση, συνήθως, εξαρτάται από το τι γνώση ακριβώς θέλουμε να εξάγουμε ανάλογα με το εκάστοτε ερώτημα-πρόβλημα που μας τίθεται, πράγμα που όπως είδαμε, εν μέρει, συμβαίνει και στις κλασσικές DM προσεγγίσεις. Ο Ester[11] διαχωρίζει τις SDM τεχνικές σε τέσσερις γενικές ομάδες : τους κανόνες συσχέτισης, τη χωρική συσταδοποίηση, την ανίχνευση χωρικής τάσης και τη χωρική κατηγοριοποίηση, ενώ σύμφωνα με τους Shekhar και Chawla [10] υπάρχουν τρείς μη-αμφιλεγόμενες τεχνικές, οι οποίες είναι : η **κατηγοριοποίηση**, η **συσταδοποίηση** και οι **κανόνες συσχέτισης**.

3.2.1 ΧΩΡΙΚΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Όπως είδαμε στο 2.4.1, ο στόχος των διάφορων αλγοριθμικών μεθόδων της κατηγοριοποίησης είναι η κατασκευή ενός μοντέλου,

μέσω της χρησιμοποίησης γνωστών δεδομένων, το οποίο θα προβλέπει την κλάση ετικέτας κάποιων νέων δεδομένων. Η βασική διαφορά που παρουσιάζει η χωρική κατηγοριοποίηση έναντι της κλασσικής προσέγγισης, είναι ότι στον υπολογισμό συμπεριλαμβάνει τον παράγοντα **γειτονιά**, καθώς, όπως είναι λογικό, στην ανίχνευση παρόμοιων κλάσεων πρέπει να λαμβάνεται υπόψη ο βαθμός σχετικότητας των τιμών ιδιοτήτων των γειτονικών αντικειμένων. Η διαδικασία της κατηγοριοποίησης μπορεί να εκτελεστεί με πολλούς τρόπους, ένας εκ των οποίων βασίζεται στη γραμμική παλινδρόμηση (Linear Regression-LR) [10].

3.2.2 ΑΝΙΧΝΕΥΣΗ ΧΩΡΙΚΗΣ ΤΑΣΗΣ

Μία χωρική τάση ορίζεται ως μία τακτική αλλαγή ενός ή περισσότερων μη χωρικών ιδιοτήτων όταν γίνεται μία χωρική απομάκρυνση από ένα αρχικό αντικείμενο. Ως εκ τούτου, η ανίχνευση χωρικής τάσης είναι μία τεχνική η οποία χρησιμοποιείται για την εύρεση προτύπων των οποίων οι ιδιότητες αλλάζουν σε σχέση με τη γειτονιά κάποιου χωρικού αντικειμένου [11]. Για καλύτερη κατανόηση, θα μπορούσαμε να θεωρήσουμε το εξής παράδειγμα : "Όταν απομακρυνόμαστε από μία μεγάλη πόλη, οι τιμές ακίνητης περιουσίας είναι χαμηλότερες", όπου η τάση χαρακτηρίζεται από την ανίχνευση μίας τακτικής αλλαγής της ιδιότητας της τιμής των ακίνητων περιουσιών σε αναλογία με την απόσταση αυτών από μία μεγάλη πόλη (αρχικό αντικείμενο).

3.2.3 ΧΩΡΙΚΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Όπως και η κλασσική συσταδοποίηση, η χωρική συσταδοποίηση είναι η διαδικασία ομαδοποίησης ενός συνόλου από χωρικά αντικείμενα. Τα αντικείμενα μέσα σε μία συστάδα έχουν μεγάλο βαθμό ομοιότητας μεταξύ τους, τη στιγμή που παρουσιάζουν διαφορές σε σχέση με τα αντικείμενα άλλων συστάδων. Με τη χρησιμοποίηση της εν λόγω διαδικασίας μπορούμε να βρούμε από συστάδες πόλεων με παρόμοια επίπεδα ανεργίας, μέχρι συστάδες από pixels παρόμοιων κλάσεων βάσει φασματικών χαρακτηριστικών. Μία από τις εφαρμογές της χωρικής συσταδοποίησης, άλλωστε, είναι η ανακάλυψη σεισμικών ρηγμάτων με την ομαδοποίηση των καταχωρήσεων ενός σεισμικού καταλόγου ή με τη δημιουργία θεματικών χαρτών οι οποίοι προκύπτουν από την ομαδοποίηση διανυσματικών χαρακτηριστικών γνωρισμάτων.

Η τεχνική των χωρικών συσχετίσεων δεν συμπεριλήφθηκε στη τρέχουσα ενότητα σκόπιμα, καθώς αποτελεί τον πυρήνα του εν λόγω συγγράμματος και όλο το κείμενο που ακολουθεί αφορά την ανάλυσή τους.

3.3 ΧΩΡΙΚΕΣ ΣΥΣΧΕΤΙΣΕΙΣ

Μία χωρική συσχέτιση ή διαφορετικά (και σύμφωνα με την πλειοψηφία των βιβλιογραφικών πηγών), μία **χωρική συντοποθεσία (spatial co-location)** αντιπροσωπεύει το υποσύνολο των χωρικών χαρακτηριστικών των οποίων οι περιπτώσεις είναι συχνά τοποθετημένες μαζί σε μία χωρική εμβέλεια (γειτονιά).

Το πρόβλημα της εξόρυξης χωρικά συσχετιζόμενων προτύπων και κατ' επέκταση κανόνων χωρικής συσχέτισης (συντοποθεσίας)[12] είναι διαφορετικό από το κλασσικό πρόβλημα εύρεσης προτύπων και κανόνων συσχέτισης μεταξύ αριθμητικών ή κατηγορηματικών δεδομένων. Αυτό συμβαίνει καθότι στο δεύτερο υπάρχει μία προκαθορισμένη βάση συναλλαγών από την οποία και προκύπτουν και οι εκάστοτε συσχετίσεις, ενώ στο πρώτο, περί του οποίου και γίνεται ο λόγος, δεν υπάρχει κάποια φυσική έννοια συναλλαγών. Αυτό δημιουργεί δυσκολίες τόσο στη χρησιμοποίηση των παραδοσιακών μετρήσεων, όσο και κατ' επέκταση στην εφαρμογή αλγορίθμων εξόρυξης συσχέτισης κατά τους οποίους τα μη συχνά στοιχειοσύνολα κλαδεύονται βάσει υποστήριξης και οι μη συχνοί κανόνες βάσει εμπιστοσύνης. Έτσι, λοιπόν, στη θέση των συναλλαγών συναντάμε την έννοια των **γειτονιών** βάσει ενός ορίου χωρικής εμβέλειας-χρήστη και στη θέση των παραδοσιακών μετρήσεων συναντάμε τις έννοιες : της επικράτησης (**prevalence**) - σχέση διαχωρισμού (**partition ratio**), διαχωρισμός περιεχομένου (**partition index**) - και της δεσμευμένης πιθανότητας (**condition probability**) [13],[14].

Ένας κανόνας συντοποθεσίας, σημασιολογικά, είναι της μορφής "στάσιμη πηγή νερού → ιός του Δυτικού Νείλου", όπου το εν λόγω παράδειγμα προσδίδει την παρουσία της επιδημίας του Δυτικού Νείλου σε περιοχές με στάσιμα νερά.

Για ευκολία της περαιτέρω συζήτησης, να τονίσουμε πως με τον όρο **χαρακτηριστικό** αναφερόμαστε σε μία χωρική οντότητα ή σε κάποιο χωρικό συμβάν, ενώ με τον όρο **αντικείμενο** αναφερόμαστε σε μία από τις περιπτώσεις του εκάστοτε χαρακτηριστικού. Για παράδειγμα, ο όρος supermarket (έστω Α) μπορεί να θεωρηθεί ως ένα χωρικό χαρακτηριστικό το οποίο έχει ως αντικείμενα κάθε υποκατάστημα (Α.1,Α.2,...,Α.Κ) σε μία καθορισμένη περιοχή.

3.3.1 ΙΣΤΟΡΙΚΟ ΥΠΟΒΑΘΡΟ

Το πρόβλημα εξόρυξης κανόνων συσχέτισης βάσει χωρικών σχέσεων για πρώτη φορά παρουσιάστηκε από τους Koperski και Han [12], οι οποίοι έθεσαν το ζήτημα ανακάλυψης υποσυνόλων χωρικών χαρακτηριστικών που συσχετίζονται συχνά με ένα συγκεκριμένο-αναφορικό χαρακτηριστικό (π.χ. καρκίνος). Σύμφωνα με τους εν λόγω ερευνητές, κάθε σύνολο από γειτονικά αντικείμενα ως προς κάθε αναφορικό χαρακτηριστικό μετατρέπεται σε μία συναλλαγή. Έτσι, οι κανόνες συσχέτισης δημιουργούνται από τις συναλλαγές που είναι σχετικές με το εκάστοτε αναφορικό χαρακτηριστικό. Όπως γίνεται αντιληπτό, όμως, η ακριβής εφαρμογή αυτής της προσέγγισης στη διαδικασία εξόρυξης συντοποθεσιών δεν μπορεί να συλλάβει την έννοια της συντοποθεσίας χωρίς την ύπαρξη ενός αναφορικού χαρακτηριστικού.

Ο Y. Morimoto[16], ανακάλυψε συχνά σύνολα γειτονικών κλάσεων(γειτονικών συντοποθετημένων χαρακτηριστικών) κάνοντας χρήση της μέτρησης υποστήριξης. Αυτή η προσέγγιση χρησιμοποιεί ένα διαχωρισμό χώρου και ένα σχήμα μη επικαλυπτόμενης ομαδοποίησης για την αναγνώριση των γειτονικών αντικειμένων. Ωστόσο, με αυτόν το τρόπο μπορούν να χαθούν περιπτώσεις συντοποθεσίας.

Οι Shekhar και Huang [13] πρότειναν τις μετρήσεις επικράτησης και δέσμευσης και έναν αλγόριθμο εξόρυξης συντοποθεσιών βάσει σύνδεσης. Ο χειρισμός των περιπτώσεων σύνδεσης για τη δημιουργία περιπτώσεων συντοποθεσίας είναι παρόμοιος με τη συνάρτηση apriori_gen. Έχοντας βρει όλα τα γειτονικά ζευγάρια (μεγέθους 2-περιπτώσεις συντοποθεσίας), η προσέγγιση αυτή, βρίσκει τις μεγέθους k>2 περιπτώσεις συντοποθεσίας συνδέοντας τις περιπτώσεις μεγέθους k-1 όπου τα πρώτα k-2 αντικείμενα είναι κοινά. Εν συνεχεία, γίνεται έλεγχος της γειτονικής σχέσης μεταξύ των (k-1)-στών αντικειμένων. Παρόλο, όμως, της ορθότητας και της πληρότητας του εν λόγω αλγόριθμου, η διαδικασία σύνδεσης παρουσιάζει αύξηση υπολογιστικής ακρίβειας με την αύξηση των περιπτώσεων συντοποθεσίας.

Όσον αφορά τη σύνδεση των περιπτώσεων συντοποθεσίας, έχουν προταθεί και άλλες τεχνικές όπως η χρησιμοποίηση R-Trees για πολλαπλά χωρικά χαρακτηριστικά [22][23], που όμως δεν λαμβάνουν υπόψη τους την ύπαρξη χωρικών ευρετηρίων.

Οι Shekhar και Yoo [24] πρότειναν έναν αλγόριθμο μερικής σύνδεσης για τη μείωση της ακριβής διαδικασίας σύνδεσης στην εύρεση περιπτώσεων συντοποθεσίας. Αυτή η προσέγγιση μειώνει σημαντικά τον αριθμό των διαδικασιών σύνδεσης, ωστόσο η απόδοσή της εξαρτάται από τον αριθμό των κομμένων περιπτώσεων από έναν κατηγορηματικό διαχωρισμό. Για την αποφυγή της διαδικασίας σύνδεσης περιπτώσεων, αργότερα [1] πρότειναν έναν νέο αλγόριθμο ο οποίος και υιοθετήθηκε στην εν λόγω διπλωματική εργασία.

Όσον αφορά τον ορισμό της γειτονιάς, στη βιβλιογραφία εκτός από τις μελέτες που χρησιμοποιούν μία σταθερή απόσταση και διαχωρισμό χώρου, υπάρχουν πολλές συλλήψεις των χωρικών σχέσεων που έχουν χρησιμοποιηθεί για την εξόρυξη χωρικών συσχετιζόμενων προτύπων. Αυτές περιλαμβάνουν buffer zones[25], τοπολογικές σχέσεις[26], k κοντινότερους γείτονες[27], διαγράμματα Delaunay [28] κ.α.

Τέλος, κάποιες πρόσφατες μελέτες εστιάζουν στην αποτελεσματικότητα της απόστασης δικτύου έναντι των μετρήσεων Ευκλείδειας απόστασης στην ανάλυση περιορισμένου δικτύου αντικειμένων και φαινομένων[29]. Οι Yamada και Thill[20] πρότειναν τη χρήση της μεθόδου K function-based με απόσταση δικτύου αντί της παραδοσιακής μεθόδου Ευκλείδειας απόστασης. Για τον υπολογισμό της εμβέλειας μεταξύ σημείων αστικών εγκαταστάσεων, επίσης, προτάθηκαν [29][30] μέθοδοι δημιουργίας Voronoi διαγράμματος που συνδυάζουν αποτελεσματικά τους περιορισμούς της μέτρησης απόστασης δικτύου, ενώ βρέθηκε πως η μέτρηση απόστασης δύο τοποθεσιών μέσω μίας ευθείας γραμμής μπορεί να υπερεκτιμήσει το σύμπλεγμα τάσης ενός φαινομένου δικτύου [31][32]. Τέλος, ο Wenhao Yu[2], πρότεινε έναν νέο αλγόριθμο εύρεσης απόστασης δικτύου, ο οποίος στηρίζεται στην επέκταση ακμής.

3.3? ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Δοθέντος ενός συνόλου απο χωρικά χαρακτηριστικά **F**, ενός συνόλου των αντικειμένων-περιπτώσεων τους **S** και μίας σχέσης γείτονα **R** πάνω στο S, μία **συντοποθεσία C (co-location C)** είναι ένα υποσύνολο των χωρικών χαρακτηριστικών, **C ⊆ F**, των οποίων οι περιπτώσεις σχηματίζουν μία κλίκα χρησιμοποιώντας τη σχέση γείτονα R. Αν d ένα όριο χρήστη και A.1,B.1 δύο χωρικά αντικείμενα, τότε αυτά είναι γείτονες αν η απόστασή τους είναι μικρότερη ή ίση του d.

Μία **περίπτωση συντοποθεσίας I (co-location instance I)** είναι ένα σύνολο από αντικείμενα, I ⊆ S, το οποίο περιλαμβάνει τα αντικείμενα όλων των χαρακτηριστικών σε μία συντοποθεσία και σχηματίζει μία σχέση **κλίκας**.

Ένας **κανόνας συντοποθεσίας (co-location rule)** είναι της μορφής : **C1→C2**, όπου **C1 ⊆ F,C2 ⊆ F** και **C1 ∩ C2 = ∅**. Για την εύρεση του αν ένας κανόνας είναι σημαντικός χρησιμοποιούνται οι μετρήσεις επικράτησης(p) και δεσμευμένη πιθανότητα(cp).

Η **σχέση διαχωρισμού (participation ratio) Pr(C,f_i)** ενός χαρακτηριστικού **f_i** σε μία συντοποθεσία **C={f_1, f_2,...., f_k}**,1 ≤ i ≤ k, είναι ένα κλάσμα από αντικείμενα του χαρακτηριστικού f_i

σε μία γειτονιά περιπτώσεων της συντοποθεσίας C-{f_i} και ορίζεται ως:

$$\Pr(C, fi) = \frac{\text{Αριθμός των ξεχωριστών αντικειμένων του fi στις περιπτώσεις της C}}{\text{Αριθμός των αντικειμένων του fi}}$$

Ο **διαχωρισμός περιεχομένου (participation index) Pi(C)** της συντοποθεσίας **C={f_1, f_2,...., f_k}** ορίζεται ως **Pi(C)=$\min_{fi \in C}${Pr(C,f_i)}.** Όσο πιο υψηλός διαχωρισμού περιεχομένου, τόσο και μεγαλύτερη η πιθανότητα τα χωρικά χαρακτηριστικά μίας συντοποθεσίας, να είναι μαζί.

Η **δεσμευμένη πιθανότητα (condition probability) P(C1|C2)** ενός κανόνα συντοποθεσίας C1→C2, αποτελεί το κλάσμα των περιπτώσεων του C2 στη γειτονιά των περιπτώσεων του C1.

$$P(C1|C2) = \frac{\text{Αριθμός των ξεχωριστών περιπτώσεων της C1 στις περιπτώσεις της ένωσης C1} \cup \text{C2}}{\text{Αριθμός περιπτώσεων του C1}}$$

Στο σημείο αυτό, να τονίσουμε πως τόσο η σχέση διαχωρισμού όσο και ο διαχωρισμός περιεχομένου δεν αυξάνονται μονοτονικά με την αύξηση του μεγέθους των συντοποθεσιών[13] και συνεπώς μπορούμε να χρησιμοποιήσουμε την **αντιμονότονη ιδιότητα** όπως στον Apriori αλγόριθμο για το κλάδεμα μη συχνών υποψήφιων συντοποθεσιών.

3.3.3 ΟΡΙΣΜΟΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Ο σκοπός, προφανώς, του εκάστοτε αλγόριθμου εξόρυξης κανόνων χωρικής συντοποθεσίας, δεν είναι άλλος από την υψηλή ποιότητα αποτελεσμάτων σε συνδυασμό με όσο το δυνατότερο χαμηλό υπολογιστικό κόστος. Το πρόβλημα, λοιπόν, έχει ως εξής :

Δοθέντος:

1. Ενός συνόλου από τύπους χωρικών χαρακτηριστικών F={f_1,.....,f_k}.
2. Ενός συνόλου από χωρικά αντικείμενα S = S_1U....US_n ,όπου S_i(1 ≤ i ≤ n) είναι ένα σύνολο από περιπτώσεις τύπου χαρακτηριστικού f_i. Κάθε αντικείμενο o ∈ S_i έχει μία διανυσματική πληροφορία [τύπος χαρακτηριστικού f_i,περίπτωση id j,τοποθεσία x,y] όπου 1 ≤ j ≤ |S_i|.
3. Μίας συμμετρικής σχέσης γείτονα R.
4. Ελάχιστων ορίων επικράτησης (min_prev) και δεσμευμένης πιθανότητας (min_cond_prob).

Βρες:
Ένα σύνολο από κανόνες συντοποθεσίας με διαχωρισμό περιεχομένου ≥ min_prev και δεσμευμένη πιθανότητα ≥ min_cond_prob.

3.3.4 ΠΡΟΚΛΗΣΕΙΣ

Λόγω του ότι η έννοια της γειτονιάς αποτελεί ακρογωνιαίο λίθο, όπως θα μπορούσε να πει κανείς, σε έναν αλγόριθμο εξόρυξης χωρικών συντοποθεσιών, η κατασκευή του γράφου αυτής αποτελεί τη βασική πρόκληση των εν λόγω αλγορίθμων. Οι περισσότερες έως τώρα μελέτες, όπως προαναφέρθηκε, βασίζονται στην Ευκλείδεια (ή επίπεδη) θεώρηση του χώρου. Οι ερευνητές θεωρούν ότι τα πρότυπα ενδιαφέροντος συμβαίνουν σε έναν απείρως ομοιογενή και ισοτροπικό χώρο, και η απόσταση μεταξύ δύο αντικειμένων μετριέται με μία ευθεία γραμμή Ευκλείδειας απόστασης. Ωστόσο, λόγου του ότι ο χώρος δεν είναι ούτε απείρως ομοιογενής ούτε ισοτροπικός και οι ανθρώπινες δραστηριότητες περιορίζονται μόνο από τα τμήματα δικτύου του επίπεδου χώρου, η προαναφερθείσα θεώρηση, όπως τόνισε και ο Miller[17], μπορεί να είναι απρόσφορη. Έτσι, λοιπόν, η απόσταση δικτύου μπορεί να θεωρηθεί πιο ουσιαστική και αξιόπιστη μέτρηση απόστασης για οικονομικές και κοινωνικές αναλύσεις [18][19][20].

Επίσης, όπως και στους κλασσικούς αλγόριθμους εξόρυξης κανόνων συσχέτισης, έτσι και στους αλγόριθμους εξόρυξης χωρικών συντοποθεσιών, ο υπολογισμός των μετρήσεων επικράτησης και δεσμευμένης πιθανότητας των υποψήφιων συντοποθεσιών απαιτεί αποτελεσματικές στρατηγικές, καθώς λόγω του μεγάλου όγκου των δεδομένων απλοϊκές (brute force) προσεγγίσεις είναι αρκετά δαπανηρές.

3.3.5 ΠΡΟΣΕΓΓΙΣΗ YOO & SHEKHAR (2006)

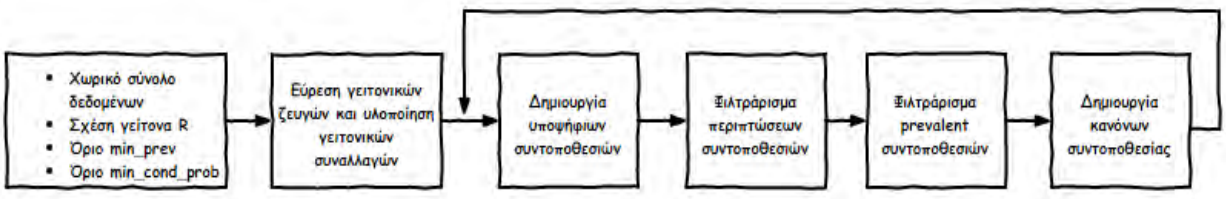
Οι Yoo και Shekhar συνεχίζοντας τη δουλειά τους [33], έκαναν μεταξύ άλλων τις εξής συνεισφορές [1] :

- Πρότειναν την υλοποίηση γειτονικών σχέσεων μεταξύ των χωρικών δεδομένων με σκοπό την αποτελεσματική εξόρυξη προτύπων συντοποθεσίας, παρουσιάζοντας δύο νέα μοντέλα διαχωρισμού γειτονιάς, το star neighborhood partitioning και το clique neighborhood partitioning, εκ των οποίων θα ασχοληθούμε με το πρώτο.
- Ανέπτυξαν έναν αλγόριθμο βασισμένο στο μοντέλο star neighborhood partition, ο οποίος είναι αποτελεσματικός καθώς χρησιμοποιώντας ένα σχήμα ψαξίματος-περίπτωσης (instance-lookup) αντί της ακριβής υπολογιστικά διαδικασίας

σύνδεσης (χωρικής ή περιπτώσεων) για το φιλτράρισμα των περιπτώσεων συντοποθεσίας. Επίσης, ένα από τα βήματά του περιλαμβάνει τραχύ κλάδεμα (coarse pruning), όπου οι υποψήφιες συντοποθεσίες φιλτράρονται χωρίς την εύρεση ακριβώς των περιπτώσεων συντοποθεσίας. Ο εν λόγω αλγόριθμος είναι σωστός και πλήρης καθώς δεν υπάρχουν ούτε ψευδή περιττώματα ούτε ψευδείς παραδοχές στη διαδικασία εύρεσης κανόνων συντοποθεσίας.

ΒΑΣΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ

Δοθέντος ενός συνόλου χωρικών δεδομένων, μίας σχέσης γείτονα και όρια min_prev, min_cond_prob, η βασική διαδικασία εξόρυξης προτύπων συντοποθεσίας περιλαμβάνει τέσσερα βήματα, όπως φαίνεται και στην εικόνα (3.1) που ακολουθεί. Αρχικά, δημιουργούνται τα υποψήφια σύνολα συντοποθεσιών των οποίων οι περιπτώσεις συλλέγονται από το δοθέν σύνολο χωρικών δεδομένων. Έπειτα, φιλτράρονται τα επικρατούντα σύνολα συντοποθεσιών που ικανοποιούν το όριο min_prev. Τέλος, δημιουργούνται οι κανόνες συντοποθεσίας βάσει του ορίου min_cond_prob. Λόγω του ότι ο περισσότερος υπολογιστικός χρόνος της εξόρυξης προτύπων χωρικής συντοποθεσίας αφιερώνεται στην εύρεση περιπτώσεων συντοποθεσίας με σχέσεις κλίκας γείτονα, προτείνεται ένα επιπλέον βήμα για την υλοποίηση των σχέσεων χωρικού γείτονα.



Εικόνα 3.1 Βήματα εξόρυξης προτύπων και κανόνων συντοποθεσίας

ΥΛΟΠΟΙΗΣΗ ΓΕΙΤΟΝΙΑΣ

Δοθέντος μίας σχέσης γείτονα, ένα σύνολο χωρικών δεδομένων μπορεί να αναπαρασταθεί ως ένας γράφος γειτόνων, όπου κάθε κόμβος αντιπροσωπεύει ένα χωρικό αντικείμενο και μία ακμή μεταξύ δύο κόμβων αντιπροσωπεύει μία γειτονική σχέση. Η υλοποίηση των γειτονικών σχέσεων με σκοπό την εξόρυξη συντοποθεσιών, πρέπει να ικανοποιεί δύο κριτήρια : α)να μη χαθεί καμία γειτονική σχέση κατα τη διαδικασία υλοποίησης και β)το υπολογιστικό κόστος πρέπει να είναι χαμηλό. Ένας επιπλέον στόχος μίας τέτοιας

υλοποίησης πρέπει να θεωρείται και η ελαχιστοποίηση των διπλότυπων εγγραφών.

Το **μοντέλο star neighborhood partition** διαχωρίζει τις σχέσεις γείτονα χρησιμοποιώντας μία σχέση **γείτονα αστεριού** (**star neighbor)** και μπορεί να χωριστεί σε δύο τύπους : α)**overlap star partitioning** και β)**disjoint star partitioning**. Από τους προαναφερθέντες τύπους, ο δεύτερος είναι προτιμότερος καθότι δεν μας δίνει διπλότυπες τιμές, χωρίς προφανώς να χάνει σχέσεις γείτονα. Αν A.1 και B.1 δύο χωρικά γειτονικά αντικείμενα τότε το disjoint star partitioning θα λάβει, σε αντίθεση με τον overlap star partitioning, μόνο το ζεύγος γειτονιάς {A.1,B.1} και όχι και το {B.1,A.1}.

Δοθέντος ενός χωρικού αντικειμένου $o_i \in S$ του οποίου ο τύπος χαρακτηριστικού είναι $f_i \in F$, **η γειτονιά αστεριού (star neighborhood)** του o_i καθορίζεται ως ένα σύνολο από χωρικά αντικείμενα $SN=\{o_j \in S \mid o_j = o_i \vee (f_j > f_i) \wedge R(o_j, o_i))\}$,όπου $f_j \in F$ είναι ο τύπος χαρακτηριστικού του o_j και R η σχέση γείτονα. Με άλλα λόγια, η γειτονιά αστεριού ενός αντικειμένου είναι ένα σύνολο από το κεντρικό αντικείμενο και των αντικειμένων της γειτονιάς του, των οποίων οι τύποι χαρακτηριστικού είναι λεξικογραφικώς μεγαλύτεροι από τον τύπο χαρακτηριστικού του.

ΦΙΛΤΡΑΡΙΣΜΑ ΠΕΡΙΠΤΩΣΕΩΝ ΣΥΝΤΟΠΟΘΕΣΙΑΣ

Αν $I=\{o_1,....,o_k\} \subseteq S$ ένα σύνολο χωρικών αντικειμένων των οποίων οι τύποι χαρακτηριστικού $\{f_1,....,f_k\}$ είναι διαφορετικοί. Εάν όλα τα αντικείμενα του I είναι γείτονες ως προς το πρώτο αντικείμενο o_1 ,τότε το σύνολο I καλείται **περίπτωση αστεριού** (**star instance)** της συντοποθεσίας $C=\{f_1,....,f_k\}$.

Το **σχήμα ψαξίματος-περίπτωσης (instance-lookup scheme**), αποτελεί την πρόταση των ερευνητών για το φιλτράρισμα των περιπτώσεων συντοποθεσίας από τις περιπτώσεις αστεριού. Μία περίπτωση αστεριού $\{o_1,....,o_k\}$ μπορεί να συλλεχθεί από τις γειτονιών αστεριού των οποίων ο τύπος του κεντρικού αντικειμένου είναι ο f_1. Η περίπτωση αστεριού δεν είναι περίπτωση συντοποθεσίας αν κι εφόσον δεν είναι βέβαιο ότι έχει μία σχέση κλίκας. Ωστόσο, αν όλα τα αντικείμενα σε αυτή εκτός του πρώτου (κεντρικού) σχηματίζουν μία κλίκα, τότε η περίπτωση αστεριού είναι περίπτωση συντοποθεσίας. Ένας τρόπος να το βρούμε αυτό είναι να ανατρέξουμε στις ήδη γνωστές k-1 περιπτώσεις συντοποθεσιών και να δούμε αν υπάρχει το υποσύνολο $\{f_2,....,f_k\}$.

ΦΙΛΤΡΑΡΙΣΜΑ ΣΥΝΤΟΠΟΘΕΣΙΑΣ

Για την εύρεση των επικρατουσών συντοποθεσιών ο αλγόριθμος χρησιμοποιεί τρία βήματα φιλτραρίσματος :

- Το **φιλτράρισμα επιπέδου-χαρακτηριστικού** ισχύει λόγω της αντιμονότονης ιδιότητας του διαχωρισμού περιεχομένου. Έτσι, μία υποψήφια συντοποθεσία μπορεί να κλαδευτεί χωρίς να εξεταστούν οι περιπτώσεις της αν ένα οποιοδήποτε υποσύνολό της δεν είναι επικρατών.

- Το **τραχύ φιλτράρισμα** γίνεται μετά την εύρεση όλων των περιπτώσεων αστεριού των συντοποθεσιών και χρησιμοποιείται για τη μείωση της ακριβής διαδικασίας φιλτραρίσματος των περιπτώσεων συντοποθεσίας. Εάν η τιμή επικράτησης των περιπτώσεων αστεριού μίας υποψήφιας συντοποθεσίας δεν ικανοποιεί το δοθέν όριο, η υποψήφια συντοποθεσία κλαδεύεται χωρίς να γίνει έλεγχος κλίκας των περιπτώσεων αστεριού της.

- Στο τελευταίο βήμα του αλγόριθμου, συναντάμε το **φιλτράρισμα εκκαθάρισης** (**refinement filtering)** το οποίο αποφασίζει τα επικρατούντα σύνολα συντοποθεσιών μετά το φιλτράρισμα όλων των περιπτώσεων συντοποθεσιών τους.

ΑΛΓΟΡΙΘΜΟΣ JOIN-LESS CO-LOCATION MINING

Ο αλγόριθμος Join-less co-location mining των Yoo και Shekhar [1], λοιπόν, έχει τρείς φάσεις :

- Μετατροπή της εισόδου του συνόλου χωρικών δεδομένων σε ένα σύνολο από ασύνδετες γειτονιές αστεριού.
- Συλλογή των περιπτώσεων αστεριού των υποψήφιων συντο-ποθεσιών από το υλοποιημένο σύνολο γειτονιών και τραχύ φιλτράρισμα των υποψήφιων με χρήση των τιμών επικράτησης των περιπτώσεων αστεριού τους.
- Φιλτράρισμα των περιπτώσεων συντοποθεσιών από τις περιπτώσεις αστεριού, εύρεση των επικρατουσών συντοποθεσιών και δημιουργία κανόνων συντοποθεσίας.

  Τόσο η δεύτερη, όσο και η τρίτη φάση επαναλαμβάνονται όσο αυξάνεται το μέγεθος των συντοποθεσιών.

3.3.6 ΠΡΟΣΕΓΓΙΣΗ WENHAO YU (2016)

Ο Wenhao Yu[2], κατά τον προηγούμενο, μόλις, χρόνο, δημοσίευσε μία έρευνα η οποία εστιάζει, ουσιαστικά, στο πρώτο κομμάτι ενός αλγόριθμου εξόρυξης κανόνων συντοποθεσίας, αυτό της δημιουργίας του γράφου γειτονιών. Ειδικότερα, ορίζεται ένα μοντέλο δημιουργίας γράφου γειτονιών βάσει απόστασης δικτύου σε συνδυασμό με έναν νέο αλγόριθμο ο οποίος αποτελεί μία επέκταση του αλγόριθμου Dijkstra[3], προσπαθώντας να τονιστεί τόσο η σημασιολογική όσο και η υπολογιστική διαφορά μεταξύ αυτών των προσεγγίσεων με τις προσεγγίσεις εύρεσης γειτονιών μέσω Ευκλείδειας απόστασης.

ΓΡΑΦΟΣ ΔΙΚΤΥΟΥ

Οι αστικές εγκαταστάσεις μπορούν να θεωρηθούν ως τυπικά σύνολα από σημεία δικτύου των οποίων οι τοποθεσίες και οι υπηρεσίες είναι "δεμένες" στα δίκτυα δρόμου. Για την ποσοτική αξιολόγηση της χωρικής συνέχειας και της αλληλεξάρτησης ποικίλων δραστηριοτήτων μέσα σε μία αστική περιοχή, στην εν λόγω έρευνα, παρατίθενται οι εξής όροι:

- Ένα **δίκτυο** είναι ένας βεβαρημένος γράφος **G=(N,E)**, όπου N ένα σύνολο από κόμβους και E σύνολο από ακμές.

- Μία **ακμή** είναι ένα τμήμα δρόμου και εκφράζεται ως **$e = \{n_i, n_j, w, O_e\}$**, όπου n_i και n_j αντιστοιχούν σε έναν αρχικό και ένα τελικό κόμβο αντίστοιχα **($n_i < n_j$)**. Το w(>0) είναι το βάρος κάθε ακμής, ενώ το O_e είναι ένα σύνολο των αντίστοιχων "δεμένων" αντικειμένων, τα οποία έχον μικρότερη Ευκλείδεια απόσταση από τη ζητούμενη ακμή σε σχέση με όλες τις υπόλοιπες του εκάστοτε γράφου.

  Τα αντικείμενα δικτύου προβάλλονται στις ακμές με **γραμμική αναφορά.** Αυτό σημαίνει ότι ένα αντικείμενο τοποθετείται πάνω σε μία ακμή βάσει την απόστασή του από τον αρχικό της κόμβο.

- Ένα **αντικείμενο δικτύου** περιγράφεται ως **$o_i = \{e, pos\}$** ,όπου το e αντιπροσωπεύει την κοντινότερη ακμή από το o_i και το pos $\in [0,w]$ εκφράζει την απόσταση μονοπατιού του "δεμένου" σημείου από τον αρχικό κόμβο. Αν, δηλαδή, $e = \{n_i, n_j, w, O_e\}$, τότε η απόσταση μεταξύ των o_i και n_i είναι το pos.

- Δοθέντος δύο αντικειμένων o_i και o_j, η **απόσταση δικτύου ND(o_i,o_j)** ορίζεται ως η απόσταση του μικρότερου μονοπατιού από το o_i στο o_j. Εάν τα δύο αντικείμενα βρίσκονται στην ίδια ακμή τότε η απόσταση τους υπολογίζεται ως $dis = |pos_i - pos_j|$,

διαφορετικά μπορεί να υπολογιστεί με την εφαρμογή ενός αλγόριθμου εύρεσης μονοπατιού(π.χ. Dijkstra[3]).

- Ένα σύνολο από δύο αντικείμενα δικτύου καλείται **γειτονιά** εάν η μεταξύ τους απόσταση δικτύου είναι μικρότερη ή ίση από ένα όριο χρήστη.

- Η **απόσταση ακμής ED(e_i,e_j)** μεταξύ δύο διαφορετικών ακμών e_i = {n_a,n_b,w_i,O_e} και e_j = {n_c,n_d,w_j,O_e} ορίζεται ως ND(n_x,n_y), όπου x ∈ {a,b}, y ∈ {c,d}. Η απόσταση ακμής μπορεί να χωριστεί σε τέσσερις τύπους (ET) : SS, SE, ES, EE, ανάλογα με τους τύπους των κόμβων με τους αυτή οποίους υπολογίζεται. Εάν και οι δύο κόμβοι είναι αρχικοί(start) τότε ο τύπος απόστασης ακμής (ET) είναι SS. Αν ο ένας κόμβος είναι αρχικός(start) και ο άλλος τελικός(end) τότε ο τύπος είναι SE κ.ο.κ.

ΕΥΡΕΣΗ ΓΕΙΤΟΝΙΚΩΝ ΖΕΥΓΩΝ

Για την εύρεση όλων των περιπτώσεων συντοποθεσίας δικτύου, πρέπει πρώτα να μετρήσουμε την απόσταση δικτύου μεταξύ όλων των αντικειμένων και να χτίσουμε τον αντίστοιχο γράφο γειτονιάς. Αν δύο αντικείμενα βρίσκονται στην ίδια ακμή, τότε ο υπολογισμός απόστασής τους, όπως προαναφέρθηκε, μπορεί να γίνει άμεσα, ειδάλλως πρέπει να υπολογιστεί το μικρότερο μονοπάτι μεταξύ αυτών. Η διαδικασία εύρεσης του μικρότερου μονοπατιού κατα μήκος πολλών κόμβων, όμως, είναι μία αρκετά ακριβή υπολογιστικά διαδικασία.

Ο αλγόριθμος του Dijkstra είναι μία από τις πιο δημοφιλείς μεθόδους στην έρευνα υπολογισμού μονοπατιού [3][34] και μπορεί να επεκταθεί στοιχειωδώς προκειμένου να δημιουργηθούν μία προς μία γειτονιές. Αρχικά λαμβάνει τις τομές και τα αντικείμενα ως δύο ξεχωριστούς τύπους κόμβων του γράφου και έπειτα, επαναληπτικώς, εκτελεί την επέκταση δικτύου σε κάθε αρχικό κόμβο για να βρει, μέσω της απόστασης δικτύου, γειτονικά αντικείμενα. Από εδώ και στο εξής, όπως και στην έρευνα η οποία παρουσιάζεται στην εν λόγω υποενότητα, θα αναφερόμαστε στην παραλλαγή αυτή του αλγόριθμου Dijkstra ως NNS (Node-based Neighbors Searching).

Ο NNS μπορεί να φτάσει σε μία βέλτιστη λύση, αλλά λόγω του ότι απαιτεί ακριβή επέκταση δικτύου για όλα τα ζητούμενα αντικείμενα, η απόδοσή του δεν είναι ικανοποιητική, ειδικά τη στιγμή που κάνουμε λόγο για πολύ μεγάλα σύνολα δεδομένων. Έτσι, λοιπόν, προτάθηκε ένας νέος αλγόριθμος, ο ENS (Edge-based Neighbors Searching) ο οποίος βασίζεται στην επέκταση ακμής και στην εκκαθάριση γειτονικών αντικειμένων. Ειδικά, η επέκταση

ακμής αποτελεί το πρώτο βήμα ως προς την εκτίμηση των αντικειμένων δικτύου που είναι τοποθετημένα σε διαφορετικές ακμές δικτύου και υποβοηθιέται από την απόσταση ακμής ED.

Εάν οι αποστάσεις μεταξύ δύο ακμών υπολογίζονται μέσω της απόστασης ακμής, τότε μπορεί να υπολογιστεί αποτελεσματικά και η χωρική εμβέλεια μεταξύ των αντικειμένων που είναι "δεμένα" σε αυτές. Με αυτόν το τρόπο, οι εκτιμώμενες γειτονιές σχηματίζουν μία υποψήφια ομάδα, η οποία εν συνεχεία μπορεί να εκκαθαριστεί προκειμένου να μας δώσει τις ακριβείς γειτονιές των εκάστοτε αντικειμένων βάσει γραμμικής αναφοράς. Ο αλγόριθμος, που προτείνεται, χρειάζεται μόνο εκτέλεση επέκτασης δικτύου σε κάθε ακμή, αντί της εκτέλεσης επέκτασης σε κάθε αντικείμενο και λόγω του ότι ο αριθμός ακμών στα πραγματικά δίκτυα, συνήθως, είναι πολύ μικρότερος από τον αριθμό των σημείων αντικειμένων, η απόδοση του ENS φαίνεται να υπερέχει αυτής του NNS.

ΠΟΛΥΠΛΟΚΟΤΗΤΑ ENS

Η βασική ιδέα του αλγόριθμου ENS, λοιπόν, είναι η δημιουργία ενός συνόλου από υποψήφιες γειτονιές με σταδιακές επεκτάσεις δικτύου γύρω από τις ζητούμενες ακμές. Στη χειρότερη περίπτωση η υπολογιστική πολυπλοκότητα του NNS είναι $O(n_g$[πολυπλοκότητα του αλγόριθμου Dijkstra]) ,ενώ του ENS είναι $O(n_e$[πολυπλοκότητα του αλγόριθμου Dijkstra]+ $n_e n_g \log n_g)$, όπου n_g ο αριθμός των αντικειμένων και n_e ο αριθμός των ακμών.

4.ΒΗΜΑΤΑ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Στα πλαίσια της εν λόγω διπλωματικής εργασίας, αρχικά έγινε η συλλογή αρχείων διανυσματικών χωρικών δεδομένων μορφής Shapefile μέσω του download server της σελίδας geofabrik.de, τα οποία προεπεξεργάστηκαν μέσω της αντικειμενο-σχεσιακής βάσης PostgreSQL η οποία διαθέτει επεκτάση GIS (postgist extension) και κατάλληλη δομή ευρετηρίου (gist) για χωρικά δεδομένα. Ακολούθως, τα προεπεξεργασμένα δεδομένα περάστηκαν σε τρία διαφορετικά Java projects όπου όσον αφορά τη διαδικασία εξόρυξης χωρικών συντοποθεσιών υιοθετήθηκε ο αλγόριθμος Join-less Colocation Mining των Yoo & Shekhar [1], ενώ όσον αφορά τη δημιουργία γράφου γειτονιών, χρησιμοποιήθηκαν τρεις διαφορετικές προσεγγίσεις : α) ο αλγόριθμος ENS [2], β) μία επέκταση του αλγορίθμου Dijkstra και γ) λόγω στενών χρονικών περιθωρίων, ένας απλοϊκός αλγόριθμος βάσει Ευκλείδειας απόστασης.

- Στη συζήτηση που αφορά τον κώδικα Java, τα αντικείμενα κλάσεων, για συντομία, θα αναφέρονται με τη κλήση της εκάστοτε κλάσης στο γ' πληθυντικό.
- Για καλύτερη παρακολούθηση και κατανόηση της συζήτησης, όσον αύτη αφορά, τουλάχιστον, τους γράφους και τις δομές, καλό θα ήταν παράλληλα με την ανάγνωση, να χρησιμοποιείται "χαρτί" και "μολύβι".

4.1 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Η προεπεξεργασία των δεδομένων θα μπορούσε να χαρακτηριστεί ως η πιο δαπανηρή χρονικά διαδικασία που ακολουθήθηκε κατα τη διάρκεια εκπόνησης της εργασίας. Το βασικό σημείο στασιμότητας ήταν η εύρεση του κατάλληλου κώδικα βάσης τόσο για το διαχωρισμό των δρόμων σε τμήματα ακμών, όσο και για τη γραμμική αναφορά των αντικειμένων δικτύου σε σχέση με την απόστασή τους από τον αρχικό κόμβο της πιο κοντινής τους ακμής. Οι δυσκολίες αυτές, όπως γίνεται αντιληπτό, αφορούσαν τις ανάγκες που πρέπει να ικανοποιηθούν προκειμένου να δημιουργηθεί ο γράφος δικτύου στις προσεγγίσεις εύρεσης γειτονιών βάσει απόστασης δικτύου.

Αρχικά, λοιπόν, συλλέχθηκαν Shapefiles δρόμων και σημείων ενδιαφέροντος (points of interest-POIs) κάποιων περιοχών, τα οποία περάστηκαν μέσω του εργαλείου **PostGis2.0 Shapefile and DBF Loader Exporter** στο βασικό εργαλείο διαχείρισης βάσης **PostgreSQL**, το **pgAdmin**. Εν συνεχεία, ακολούθησαν οι κατάλληλες διαμορφώσεις, για τον έλεγχο και την οπτικοποίηση των οποίων χρησιμοποιήθηκε η εφαρμογή **QGIS**.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΓΙΑ ΤΗΝ ΕΥΚΛΕΙΔΕΙΑ ΠΡΟΣΕΓΓΙΣΗ

Για την Ευκλείδεια προσέγγιση δημιουργίας γειτονιών το μόνο που χρειάστηκε από πλευράς προεπεξεργασίας ήταν ο διαχωρισμός των διάφορων σημείων ενδιαφέροντος και η δημιουργία αντίστοιχων πινάκων με τέσσερις στήλες : id, type και συντεταγμένες x, y. Τα στοιχεία αυτών των πινάκων συγχωνεύτηκαν σε έναν ενιαίο πίνακα με αλφαριθμητική ταξινόμηση.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΓΙΑ ΤΙΣ ΠΡΟΣΕΓΓΙΣΕΙΣ ΑΠΟΣΤΑΣΗΣ ΔΙΚΤΥΟΥ

Για τις δύο προσεγγίσεις δημιουργίας γειτονιών βάσει απόστασης δικτύου χρειάστηκε προεπεξεργασία τόσο των σημείων ενδιαφέροντος, όσο και των δρόμων.

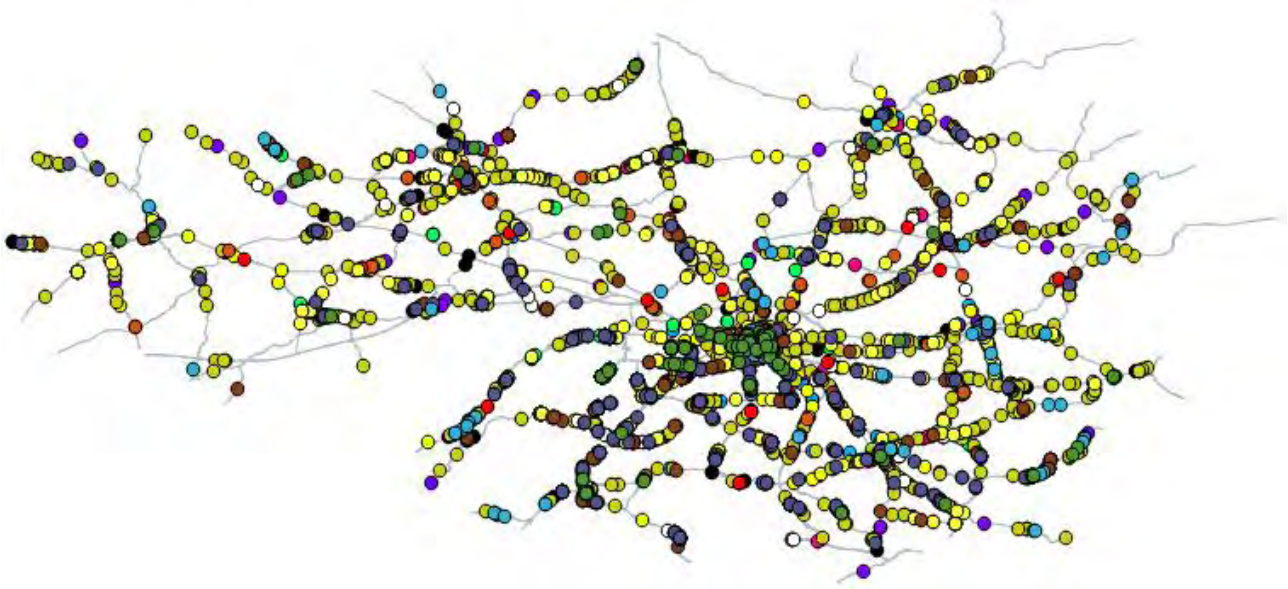
- Πίνακας δρόμων: Ο γεωμετρικός τύπος (geom) των δρόμων στην πλειοψηφία των όσων περιπτώσεων συναντήθηκαν ήταν multilinestring, πράγμα που περιπλέκει λίγο τη διαδικασία επεξεργασίας. Έτσι, λοιπόν, με τη βοήθεια των συναρτήσεων ST_NumGeometries και ST_GeometryN έγινε η μετατροπή αυτού σε linestring. Στη συνέχεια, επιλέχθηκαν κάποιοι συγκεκριμένοι τύποι δρόμων(π.χ. κορμού, πρωτεύον, δευτερεύον) προς εξοικονόμηση χρόνου, οι οποίοι και πέρασαν σε έναν αρχικό πίνακα **roads** : id, type, geom. Με βάση τον roads και με τη βοήθεια της συνάρτησης ST_DumpPoints πάρθηκαν όλοι οι κόμβοι κάθε δρόμου και δημιουργήθηκε ο πίνακας **nodes** : id, type, index, geom, όπου οι δύο πρώτες στήλες αντιστοιχούν στον εκάστοτε δρόμο που ανήκει ο κάθε κόμβος, ενώ το index είναι ένα μοναδικό, ουσιαστικά, αναγνωριστικό σε σχέση με τους υπόλοιπους κόμβους του ίδιου δρόμου. Ακολούθως, κάνοντας ένα LEFT OUTER JOIN στον nodes και με τη βοήθεια των συναρτήσεων ST_MakeLine, ST_DistanceSphere, ST_StartPoint και ST_EndPoint, καταλήξαμε στον ζητούμενο πίνακα ακμών **edges** : id, st_p, en_p, w , όπου st_p και en_p ο αρχικός και τελικός κόμβος κάθε ακμής, αντίστοιχα, και w το βάρος της. Όσες null τιμές προέκυψαν, διαγράφηκαν.

- Πίνακας αντικειμένων: Έχοντας καταλήξει στον τελικό πίνακα αντικειμένων-σημείων ενδιαφέροντος **objects** : id, type, geom, δημιουργήθηκε ένας νέος πίνακας temp για την εύρεση των αποστάσεων όλων των αντικειμένων από κάθε ακμή. Μέσω του **temp** και με τη βοήθεια της συνάρτησης min, βρέθηκε το ποια ακμή είναι η κοντινότερη σε κάθε αντικείμενο και έτσι δημιουργήθηκε ένας νέος πίνακας **lr_objects** : id, type, edge, όπου edge το αντίστοιχο id της κοντινότερης ακμής. Σε αυτό το σημείο να τονιστεί πως καθ' όλη τη διάρκεια επεξεργασίας των πινάκων αντικειμένων χρησιμοποιήθηκε η δομή ευρετηρίου gist. Έχοντας

τον lr_objects, λοιπόν, και με τη βοήθεια της συνάρτησης ST_ClosestPoint βρέθηκε το κοντινότερο σημείο της κοντινότερης ακμής στο οποίο και έγινε η προβολή του εκάστοτε αντικειμένου. Τέλος, με τη βοήθεια της συνάρτησης ST_DistanceSphere, έγινε και ο υπολογισμός της απόστασης των προβαλλόμενων αντικειμένων σε σχέση με τους αρχικούς κόμβους της πιο κοντινής τους ακμής.

Όπως εύκολα μπορεί να γίνει αντιληπτό, η διαδικασία που ακολουθήθηκε για την προβολή των αντικειμένων στις εκάστοτε κοντινότερες ακμές τους βάσει γραμμικής αναφοράς, ήταν αρκετά ακριβή υπολογιστικά και δεν προσφέρεται για πολύ μεγάλους όγκους δεδομένων, παρ' όλα αυτά δεν κατέστη δυνατή η εύρεση κάποιου άλλου πλήρη και πιο αποτελεσματικού τρόπου.

Η εικόνα που ακολουθεί είναι αποτέλεσμα χρήσης της εφαρμογής **QGIS**, στην οποία έγινε σύνδεση με τα προεπεξεργασμένα δεδομένα της μητροπολιτικής κοσμητείας του Great Manchester. Με διαφορετικό χρώμα τα διάφορα σημεία ενδιαφέροντος που χρησιμοποιήθηκαν.



Εικόνα 4.1 - Στιγμιότυπο χρήσης QGIS

4.2 ΣΥΝΔΕΣΗ ΒΑΣΗΣ

Έχοντας ολοκληρωθεί η διαδικασία της προεπεξεργασίας, το επόμενο βήμα που ακολουθήθηκε ήταν η διαμέσου τοπικού server σύνδεση των διαμορφωμένων πινάκων με τρία διαφορετικά Java projects ανάλογα με την προσέγγιση εύρεσης των γειτονικών αντικειμένων. Οι βιβλιοθήκες που προστέθηκαν προκειμένου να καταστεί εφικτή η εν λόγω σύνδεση ήταν η mysql-connector-java και η postgresql.

ΣΥΝΔΕΣΗ ΒΑΣΗΣ ΓΙΑ ΤΗΝ ΕΥΚΛΕΙΔΕΙΑ ΠΡΟΣΕΓΓΙΣΗ

Ο μόνος πίνακας που συνδέθηκε στο project εύρεσης γειτόνων μέσω Ευκλείδειας απόστασης ήταν αυτός των objects. Έτσι, η μόνη κλάση που δημιουργήθηκε ήταν η **NetworkObject**.java, τα αντικείμενα της οποίας περάστηκαν σε ένα Map με key τον τύπο-χαρακτηριστικό τους.

ΣΥΝΔΕΣΗ ΒΑΣΗΣ ΓΙΑ ΤΙΣ ΠΡΟΣΕΓΓΙΣΕΙΣ ΑΠΟΣΤΑΣΗΣ ΔΙΚΤΥΟΥ

Η σύνδεση των πινάκων nodes, edges και lr_objects στα projects εύρεσης γειτόνων μέσω απόστασης δικτύου ήταν, προφανώς, πιο περίπλοκη, καθότι έπρεπε να δημιουργηθεί ο γράφος δικτύου G=(N,E,W). Αυτό κατέστη δυνατό με τη δημιουργία των κλάσεων **Node**.java, **Edge**.java, **NetworkObject**.java και τη χρήση κατάλληλων μεταβλητών στιγμιοτύπου και δομών δεδομένων αποθήκευσης των αντικειμένων τους. Οι διασυνδέσεις που υλοποιήθηκαν αφορούσαν τόσο τους κόμβους οι οποίοι συνδέονταν μέσω μίας ίδιας ακμής, όσο και τις σχέσεις μεταξύ ακμών και NetworkObjects στα πλαίσια της γραμμικής αναφοράς.

4.3 ΕΥΡΕΣΗ ΓΕΙΤΟΝΙΚΩΝ ΑΝΤΙΚΕΙΜΕΝΩΝ

Έχοντας φύγει, πλέον, από τις λειτουργίες σχετικά με τη βάση και έχοντας δημιουργήσει τις κατάλληλες κλάσεις, δομές και διασυνδέσεις, το επόμενο βήμα που ακολουθήθηκε ήταν η υλοποίηση των τριών διαφορετικών προσεγγίσεων εύρεσης γειτονικών αντικει-μένων.

ΕΥΚΛΕΙΔΕΙΑ ΠΡΟΣΕΓΓΙΣΗ

Όπως προαναφέρθηκε, για την εύρεση των γειτονικών αντικειμένων μέσω Ευκλείδειας απόστασης, λόγω στενών χρονικών περιθωρίων, χρησιμοποιήθηκε μία brute force προσέγγιση αντί κάποιας

προσέγγισης σάρωσης χώρου, πράγμα που δεν προτείνεται όταν έχουμε να κάνουμε με πολύ μεγάλο αριθμό από objects, λόγω υπολογιστικής ακρίβειας.

Κατά την υλοποίηση αυτή, λοιπόν, χρησιμοποιήθηκαν δύο επαναληπτικοί βρόχοι για τον έλεγχο του αν κάθε ξεχωριστό και ένα προς ένα ζεύγος από NetworkObjects ικανοποιούσε το όριο απόστασης χρήστη. Προς αποφυγή διπλότυπων εγγραφών, τα ζεύγη που ικανοποιούσαν αυτό το όριο δημιουργούσαν ένα αντικείμενο NeighborPair η κλάση του οποίου λάμβανε τα NetworkObjects με αλφαβητική σειρά μέσω κατάλληλου ελέγχου στα μεταξύ τους χαρακτηριστικά. Εν συνεχεία, κάθε νέο NeighborPair περνούσε σε ένα hashSet. Λόγω, τέλος, του ότι σύμφωνα με το μοντέλο disjoint star partitioning που χρησιμοποιήθηκε για τη δημιουργία των συναλλαγών δεν ενδιαφερόμαστε για πρότυπα ίδιου χαρακτηριστικού, ο έλεγχος απόστασης και η καταχώριση γειτονικών αντικειμένων ίδιου τύπου αποφεύχθηκε.

Κώδικας 4.1 - Μέθοδος υπολογισμού Ευκλείδειας απόστασης

```
public double calcDist(NetworkObject o1,NetworkObject o2){

    final int R = 6370986 ; // Radius of the earth
    //x longitude-y latitude
    double lon1=o1.getX();
    double lon2=o2.getX();
    double lat1=o1.getY();
    double lat2=o2.getY();

    Double lonDistance = Math.toRadians(lon2 - lon1);
    Double latDistance = Math.toRadians(lat2 - lat1);
    Double a = Math.sin(latDistance / 2) * Math.sin(latDistance / 2)
        + Math.cos(Math.toRadians(lat1)) * Math.cos(Math.toRadians(lat2))
        * Math.sin(lonDistance / 2) * Math.sin(lonDistance / 2);
    Double c = 2 * Math.atan2(Math.sqrt(a), Math.sqrt(1 - a));
    double distance = (float) (R * c);

    return distance;

}
```

ΠΡΟΣΕΓΓΙΣΗ ΕΠΕΚΤΑΣΗΣ ΑΚΜΩΝ ΓΡΑΦΟΥ ΔΙΚΤΥΟΥ (ENS)

Για την εύρεση των γειτονικών αντικειμένων μέσω της προσέγγισης επέκτασης ακμών ακολουθήθηκαν τα βήματα του αλγόριθμου ENS [2]. Η κλάση **ENS**.java που υλοποιήθηκε, λοιπόν, λαμβάνει ως ορίσματα τα σύνολα nodes ,edges και ένα όριο χρήστη, ενώ οι μεταβλητές στιγμιοτύπου της είναι οι εξής :

- **EDC** : ένα σύνολο από πλειάδες (**Tuple**.java), κάθε μία εκ των οποίων έχει μεταβλητές στιγμιοτύπου δύο ακμές (**Edges**), την απόσταση μεταξύ αυτών και τον τύπο υπολογισμού της εν λόγω απόστασης (SS,SE κ.ο.κ.)

- **STC :** ένα ενδιάμεσο σύνολο από επικροτούμενα ζευγάρια Network-Objects (NeighborPair.java) το οποίο χρησιμοποιείται για το πέρασμα αυτών των ζευγαριών στο τελικό σύνολο **ST.**

Η αρχική μέθοδος **neigborsSearching** (μέθοδος 4.1), που μπορεί να χαρακτηριστεί ως ο διαχειριστής της διαδικασίας, διατρέχει όλες τις ακμές του γράφου δικτύου και σε κάθε ξεχωριστή επανάληψη καθαρίζει το σύνολο **EDC** προκειμένου να καταχωρηθούν σε αυτό οι εκάστοτε νέες πλειάδες μέσω της μεθόδου **edgeExpansion.** Έπειτα, κάθε νέα πλειάδα περνάει ως όρισμα στη μέθοδο **neighborObjectsRefining** κατα την οποία ελέγχονται οι αποστάσεις δικτύου μεταξύ όλων των αντίστοιχων προβαλλόμενων NetworkObjects (βάσει του τύπου απόστασης των εκάστοτε ακμών) και δημιουργούνται, όπως περιγράφηκε και στην προηγούμενη προσέγγιση, τα αντικείμενα **NeighborPair**, τα οποία προστίθενται στο σύνολο **STC.** Κάθε νέο στοιχείο του **STC**, τέλος, καταχωρείται στο ζητούμενο σύνολο **ST.**

Κώδικας 4.2 - Μέθοδος neigborsSearching

```
public void neigborsSearching(){

    for(Edge e:edges){

        EDC.clear();
        edgeExpansion(e);
        for(Tuple t:EDC){
            STC.clear();
            neighborObjectsRefining(t);

            for(NeighborPair trnsc:STC){

                ST.add(trnsc);

            }
        }

    }

}
```

Όσον αφορά τη πιο σημαντική μέθοδο της εν λόγω προσέγγισης, την **edge-Expansion**, κάθε φορά λαμβάνει ως όρισμα μία ακμή της οποίας και διατρέχει τον αρχικό και τελικό κόμβο. Ο εκάστοτε κόμβος που είναι προς **εξέταση**, αρχικοποιεί την απόσταση δικτύου του ως προς τον εαυτό του με 0, ενώ όλοι οι υπόλοιποι αρχικοποιούν την απόστασή δικτύου τους ως προς αυτόν με μία πολύ μεγάλη τιμή (τυπικά ∞) και την επισκέψιμότητα τους ως unvisited. Έπειτα, δημιουργείται μία δομή σωρού ελαχίστου (minHeap), στην οποία και αρχικά εισάγεται ο κόμβος προς εξέταση.

```
//minHeap Initialization
Queue<Node> minHeap = new PriorityQueue<>(node.getEdges().size(),new nodesDistComparator());
minHeap.add(node);
```

Κώδικας 4.3 - Αρχικοποίηση Σωρού

Όσο η εν λόγω δομή έχει περιεχόμενο, γίνεται poll του κόμβου με την μικρότερη απόσταση δικτύου ως προς τον εξεταζόμενο κόμβο, εκ των οποίων ο πρώτος (αν όχι ο ίδιος ο εξεταζόμενος) αλλάζει την αντίστοιχη μεταβλητή στιγμιοτύπου του σε visited. Στη συνέχεια, διατρέχονται όλες οι ακμές στις οποίες αυτός πρόσκειται και ακολουθούν οι εξής διαδικασίες:

- Λήψη της απόστασης του εκάστοτε visited κόμβου ως απόσταση ακμής και αν είναι μικρότερη ή ίση από το όριο χρήστη, τότε δημιουργούνται οι εκάστοτε πλειάδες, οι οποίες και καταχωρούνται στο σύνολο **EDC**.

- Αν ο άλλος κόμβος της ακμής που διατρέχεται, είναι unvisited, τότε υπολογίζεται η απόσταση δικτύου του σε σχέση με τον εξεταζόμενο κόμβο. Αν αυτή είναι μικρότερη ή ίση του ορίου χρήστη, τότε ο εν λόγω κόμβος εισέρχεται στη δομή σωρού.

ΠΡΟΣΕΓΓΙΣΗ ΕΠΕΚΤΑΣΗΣ ΚΟΜΒΩΝ ΓΡΑΦΟΥ ΔΙΚΤΥΟΥ (NNS)

Όσον αφορά την εν λόγω προσέγγιση, δυστυχώς, δεν κατέστη δυνατό να υλοποιηθεί με τέτοιο τρόπο που οι εκτυπώσεις της να είναι παρόμοιες με αυτές της προηγούμενης, καθώς στο [2] δεν γίνεται σαφής περιγραφή της διαδικασίας που ακολουθήθηκε για την εξαγωγή κάποιων συμπερασμάτων σε σχέση με τον ENS.

Η βασική επανάληψη του κώδικα που γράφηκε, περί της εν λόγω προσέγγισης, διατρέχει όλα τα NetworkObjects, από τα οποία παίρνει τον αρχικό κόμβο της εκάστοτε ακμής στην οποία πρόσκειται το καθένα από αυτά. Εν συνεχεία, κάνει τις κατάλληλες αρχικοποιήσεις βάσει αλγόριθμου Dijkstra, όπου σε σχέση με τη προηγούμενη περιγραφή, έχουμε ένα επιπλέον πεδίο για κάθε κόμβο, τον προκάτοχο (predecessor). Ακολούθως **(βήμα 2)**, όποια **NetworkObjects** πρόσκεινται στην ίδια ακμή με το τρέχον, δημιουργούν **NeighborPairs**, τα οποία και προστίθενται στη βασική δομή αποθήκευσης αυτών, την **ST**. Μετά από αυτή τη διαδικασία αρχικοποιείται η δομή σωρού και καταχωρείται σε αυτή ο κόμβος του NetworkObject. Όσο ο σωρός έχει περιεχόμενο, ο κόμβος **nMin** που εξάγεται από αυτή γίνεται visited και αν η απόστασή του είναι μικρότερη ή ίση από το όριο χρήστη, διατρέχονται όλοι οι κόμβοι με τους οποίους αυτός συσχετίζεται μέσω κάποιας κοινής

ακμής. Όποιος κόμβος από αυτούς δεν είναι visited, τότε υπολογίζεται η απόστασή του σε σχέση με τον αρχικό κόμβο, η οποία αν είναι μικρότερη από κάποια πιθανή προηγούμενή του ή από το όριο χρήστη, καταχωρείται καταλλήλως, μαζί με τον προκάτοχο του, nMin. Έπειτα, ελέγχεται και καταχωρείται η αρχική ακμή (**preEdge**) του μονοπατιού με βάση το οποίο έγινε ο υπολογισμός αυτός και ακολουθεί μία επανάληψη στις ακμές στις οποίες πρόσκειται ο κόμβος, προκειμένου να βρεθεί ποια ήταν η τελευταία ακμή του μονοπατιού και να καταχωρηθεί ως πεδίο σε αυτόν. Ακολούθως, αποφεύγοντας τον έλεγχο της ίδιας ακμής με αυτή που πρόσκειται το εκάστοτε τρέχον NetworkObject, λόγω του 2ου βήματος, προστίθενται νέες πλειάδες στο σύνολο **NDC** (το οποίο καθαρίζεται με τη κάθε αλλαγή του ελεγχόμενου NetworkObject), οι οποίες έχουν ως ορίσματα το τρέχον NetworkObject, τη τελευταία ακμή του μονοπατιού, την απόσταση που βρέθηκε προηγουμένως και ένα String το οποίο αντικατοπτρίζει την "ομάδα" με βάση την οποία θα γίνει ο υπολογισμός μεταξύ των NetworkObjects, αργότερα. Αν η preEdge εμπεριέχει το τρέχον NetworkObject, τότε η πλειάδα είναι "ομάδας A", διαφορετικά, "ομάδας B". Κάθε φορά που αδειάζει ο σωρός, διατρέχεται το εκάστοτε σύνολο NDC, όπου μέσω αυτού καλείται η μέθοδος υπολογισμού της απόστασης μεταξύ των NetworkObjects, όπου δημιουργούνται τα εκάστοτε Neighbor-Pairs βάσει του ορίου χρήστη.

Κώδικας 4.4- Υπολογισμός Απόστασης Δικτύου βάσει "ομάδων".

```
if(t.getNT().equals("A")){

    ND=t.getND()-nObj.getPos()-nObj2.getPos();

}
if(t.getNT().equals("B")){

    ND=t.getND()+nObj.getPos()-nObj2.getPos();

}
```

Η εν λόγω διαδικασία σε συνδυασμό με τη μετέπειτα διαδικασία της εξόρυξης των δεδομένων, παρήγε κατα βάση λιγότερα πρότυπα σε σχέση με την εξόρυξη δεδομένων κατα την οποία χρησιμοποιήθηκε ο αλγόριθμος ENS. Επίσης, τα πρότυπα αυτά ήταν στη μεγάλη τους πλειοψηφία μικρότερου ενδιαφέροντος. Ένας πιθανός λόγος που μπορεί να συνέβη αυτό, είναι ότι ο αλγόριθμος [2], λαμβάνει τέσσερις διαφορετικές περιπτώσεις υπολογισμού απόστασης μεταξύ των αντικειμένων, σε σχέση με τη μέθοδο βάσει Dijkstra που μόλις περιγράφηκε, η οποία χρησιμοποιεί δύο.

4.4 ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ ΣΥΝΤΟΠΟΘΕΣΙΑΣ

Έχοντας ολοκληρώσει την προεπεξεργασία των δεδομένων, έχοντας συνδέσει τους κατάλληλους πίνακες της βάσης με τα τρία Java projects μας - χρησιμοποιώντας τις κατάλληλες δομές δεδομένων αποθήκευσης - και έχοντας βρει όλα τα γειτονικά ζεύγη των NetworkObjects, μπορούμε να περάσουμε στη συζήτηση για τη διαδικασία εξόρυξης, η οποία είναι ίδια και για τις τρεις υλοποιήσεις.

Για την εξόρυξη προτύπων και κατ' επέκταση κανόνων συντοποθεσίας, λοιπόν, όπως προαναφέρθηκε, ακολουθήθηκαν τα βήματα του αλγόριθμου Join-less co-location mining[1]. Προκειμένου να περιγραφεί όσο το δυνατόν καλύτερα η διαδικασία, καλό θα ήταν να πάρουμε τις συναρτήσεις μία προς μία. Λόγω, όμως, του ότι τα περισσότερα σημεία είναι αρκετά μεγάλα από πλευράς κώδικα και είναι αδύνατο να αναλυθούν γραμμή προς γραμμή, θα γίνει μία προσπάθεια για όσο το δυνατότερο καλύτερη περιγραφή.

1. **Δημιουργία γειτονικών συναλλαγών βάσει μοντέλου disjoint star partitioning** : Για τη διεκπεραίωση αυτού του βήματος χρησιμοποιήθηκαν οι κλάσεις **StarNeighborhoods**.java και **Transaction**.java, όπου η πρώτη εξ αυτών δέχεται ως ορίσματα το σύνολο **ST** (4.3) και τον Map των NetworkObjects (4.2), έχοντας ως βασική μεταβλητή στιγμιοτύπου έναν Map (**starNeighborhoods**) με κλειδιά τα κεντρικά χαρακτηριστικά των γειτονιών αστεριού, το καθένα εκ των οποίων δείχνει σε μία λίστα από συναλλαγές (Transactions). Η βασική υλοποίηση γίνεται μέσω της μεθόδου **getTransactions()** της προαναφερθείσας κλάσης. Εκεί, αρχικά διατρέχονται τα κλειδιά του Map των (Network)Objects και σε κάθε αλλαγή χαρακτηριστικού δημιουργείται μία νέα λίστα από Transactions (**transactionsList**). Εν συνεχεία, διατρέχονται τα εκάστοτε NetworkObjects όπου σε πρώτη φάση τους δίνεται μία false flag και δημιουργούνται νέες Transactions, στις οποίες και προστίθεται το καθένα από αυτά. Σε αυτό το σημείο ας θωρήσουμε ένα τρέχον NetworkObject ως A.1. Στη τελευταία εμφωλευμένη επανάληψη που ακολουθεί, διατρέχεται το σύνολο ST όπου αν το πρώτο NetworkObject (έστω A.1) ενός NeighborPair (έστω A.1, B.1) είναι ίδιο με το A.1, τότε έχουμε αλλαγή του flag σε true και προσθήκη του B.1 στη Transaction του A.1. Κατα τη τελευταία επανάληψη, αν το flag είναι true, τότε η συναλλαγή προστίθεται στη τρέχουσα transactionsList, η οποία με τη σειρά της, στο τέλος της επόμενης επανάληψης καταχωρείται στον τρέχον starNeighborhoods.

Σε αυτό το σημείο, να τονιστεί πως η εν λόγω διαδικασία υλοποιήθηκε εκτός της βασικής κλάσης εξόρυξης, η οποία όμως, προφανώς και παίρνει ως όρισμα τον Map που δημιουργήθηκε.

Η βασική κλάση εξόρυξης προτύπων και κανόνων συντοποθεσίας, λοιπόν, ήταν η **ColocationMining**.java, η οποία πέρα του Map (**transactionsMap**) των συναλλαγών, πήρε ως ορίσματα :

- έναν ακόμη Map (**featuresSum**) με κλειδιά τα χαρακτηριστικά NetworkObjects, τα οποία δείχνουν σε έναν ακέραιο που αντιπροσωπεύει τον αριθμό των εκάστοτε προσκείμενων σε αυτά NetworkObjects και
- τα όρια χρήστη **minPrev** και **minCondProb**.

ενώ ως μεταβλητές στιγμιοτύπου χρησιμοποιήθηκαν σύνολα και Maps για τις υποψήφιες και επικρατούσες συντοποθεσίες, για τις περιπτώσεις αστεριού και κλίκας, αλλά και για τους κανόνες συντοποθεσίας.

2. **Δημιουργία υποψήφιων συντοποθεσιών** : Αρχικά, λαμβάνονται ως επικρατούσες συντοποθεσίες όλα τα διακριτά χαρακτηριστικά των NetworkObjects, καθώς από ορισμού οι 1-επικρατούσες συντοποθεσίες έχουν διαχωρισμό περιεχομένου ίσο με 1. Οι μεγέθους k > 1 υποψήφιες συντοποθεσίες δημιουργούνται από τις (k-1)-επικρατούσες συντοποθεσίες μέσω μίας μεθόδου παρόμοιας με αυτή της apriori_gen. Αν μία υποψήφια συντοποθεσία περιέχει κάποιο μη επικρατών (k-1)-υποσύνολο, τότε κλαδεύεται βάσει του επιπέδου-χαρακτηριστικού φιλτράρισμα.

Κώδικας 4.5- Φιλτράρισμα επιπέδου-χαρακτηριστικού

```
if(prevK>1){

    Iterator<CandidateColocation> iter=cCset.iterator();
    while(iter.hasNext()){

        boolean prevalent=false;
        int count=0;
        CandidateColocation cCol=iter.next();

        for(PrevalentColocation set:prevPc){

            if(cCol.getCandidate().containsAll(set.getPrevCol())){

                count++;

                if(count==cCol.getCandidate().size()){

                    prevalent=true;
                    break;
                }


            }

        }
        if(!prevalent){

            iter.remove();

        }
    }
}
```

3. **Φιλτράρισμα των περιπτώσεων αστεριών από τις υποψήφιες συντοποθεσίες μέσω της δομής συναλλαγών (transactionsMap) :** Για το εν λόγω βήμα, στη μέθοδο-διαχειριστή της Colocation-Mining, αρχικά διατρέχονται τα κλειδιά του transactionsMap και εν συνεχεία οι τρέχουσες k-υποψήφιες συντοποθεσίες. Αν ο τύπος χαρακτηριστικού του πρώτου μέλους της εκάστοτε υποψήφιας συντοποθεσίας είναι ίδιος με το εκάστοτε κλειδί, τότε καλείται η μέθοδος **filter_star_instances.** Η εν λόγω μέθοδος παίρνει ως ορίσματα την υποψήφια συντοποθεσία, τις συναλλαγές του κλειδιού και το τρέχον k. Με τις κατάλληλες μεταβλητές, δομές, επαναλήψεις και ελέγχους, κάθε φορά που τα χαρακτηριστικά της υποψήφιας συντοποθεσίας εμπεριέχονται στα χαρακτηριστικά των NetworkObjects μίας συναλλαγής, τότε δημιουργείται και μία νέα περίπτωση αστεριού για την εν λόγω συντοποθεσία. Αν το k είναι 2, τότε όπως εύκολα μπορεί να γίνει κατανοητό, οι περιπτώσεις αστεριού είναι και περιπτώσεις κλίκας.

```
for(String feature:transactionsMap.keySet()){

    for(CandidateColocation cC:ccMap.get(k)){

        if(cC.getCandidate().first().equals(feature)){

            filter_star_instances(cC,transactionsMap.get(feature),k);

        }
    }
}
```

Κώδικας 4.6 – Διαδικασία καλέσματος μεθόδου φιλτραρίσματος περιπτώσεων αστεριού

4. **Τραχεία επιλογή επικρατουσών συντοποθεσιών :** Αν το k είναι μεγαλύτερο του 2 τότε, πρωτού βρούμε τις περιπτώσεις κλίκας, γίνεται ένα πέρασμα από τη μέθοδο select_coarse_prevalent_co-locations, η οποία παίρνει ως μοναδικό όρισμα το k. Σε αυτή τη μέθοδο, λοιπόν, διατρέχονται οι k-υποψήφιες συντοποθεσίες, όπου για τις περιπτώσεις αστεριών της κάθε μίας καλείται η μέθοδος double **getPI**() της κλάσης **StarInstance**.java, με όρισμα τον featuresSum. Στην εν λόγω μέθοδο, ουσιαστικά, οι περιπτώσεις αστεριών διατρέχονται ανά χαρακτηριστικό των NetworkObjects τους και με τη βοήθεια των αντίστοιχων τιμών των κλειδιών του featuresSum, γίνεται γνωστό ο εκάστοτε διαχωρισμός περιεχομένου, με βάση τον οποίο η αρχική μέθοδος κλαδεύει τις μη συχνές συντοποθεσίες. Ακολουθεί ο υπολογισμός της σχέσης διαχωρισμού μέσα στη getPI().

Κώδικας 4.7- Υπολογισμός σχέσης διαχωρισμού

```
for(String elemType:cC.getCandidate()){
    tempSet=new HashSet<>();

    for(TreeSet<NetworkObject> nObSet:nObjList){

        for(NetworkObject nObj:nObSet ){

            if(nObj.getType().equals(elemType)){

                tempSet.add(nObj);
                break;
            }
        }
    }

    pr=tempSet.size()/(double)featuresSum.get(elemType);
    prMap.put(elemType,pr);
}
```

5. **Φιλτράρισμα των περιπτώσεων συντοποθεσίας** : Κατά τη διάρκεια αυτού του βήματος, ξεχωρίζουμε τις περιπτώσεις αστεριού (για k>2) που είναι και περιπτώσεις κλίκας-συντοποθεσίας. Αυτό γίνεται μέσω της μεθόδου **filter_clique_instances**, όπου με τις κατάλληλες δομές, μεταβλητές και επαναλήψεις ελέγχεται αν υπάρχει κάποια (k-1)-περίπτωση συντοποθεσίας, της οποίας τα NetworkObjects αντιστοιχίζονται με την εκάστοτε ουρά των NetworkObjects των περιπτώσεων αστεριού της εκάστοτε υποψήφιας συντοποθεσίας που διατρέχεται. Κάθε φορά που ο προαναφερων έλεγχος ικανοποιείται τότε δημιουργείτε και μία νέα περίπτωση συντοποθεσίας.

6. **Επιλογή επικρατούντων συνόλων συντοποθεσίας** : Σε αυτό το σημείο, μέσω της συνάρτησης **select_prevlent_colocations**, γίνεται το φιλτράρισμα εκκαθάρισης, κατα το οποίο υπολογίζεται ο διαχωρισμός περιεχομένου των υποψήφιων συντοποθεσιών βάσει των περιπτώσεων συντοποθεσίας. Οι υποψήφιες συντοποθεσίες που έχουν διαχωρισμό περιεχομένου μεγαλύτερο από το όριο χρήστη, περνάνε στο σύνολο των επικρατουσών συντοποθεσιών. Το εν λόγω βήμα, έχει παρόμοια λειτουργικότητα με το βήμα 4, με κύρια διαφορά, ότι ο διαχωρισμός περιεχομένου δεν υπολογίζεται μέσω των περιπτώσεων αστεριού.

7. **Δημιουργία κανόνων συντοποθεσίας** : Εν τέλει, με τη βοήθεια των μεθόδων **gen_colocation_rules** και **get_condition_probabilty** και μέσω των συνόλων επικρατουσών συντοποθεσιών, δημιουργούνται

και αποθηκεύονται οι κανόνες συντοποθεσίας με δεσμευμένη πιθανότητα μεγαλύτερη του αντίστοιχου ορίου χρήστη. Όσον αφορά τη διαδικασία δημιουργίας των κανόνων, υιοθετήθηκαν τα βήματα που περιγράφηκαν στον αλγόριθμο 2.2, ενώ όσον αφορά την εύρεση της δεσμευμένης πιθνότητας, η μέθοδος υπολογισμού αυτής :

- δέχεται ως ορίσματα τα σύνολα LHS ∪ RHS και LHS,
- διατρέχει τις αντίστοιχου μεγέθους του πρώτου ορίσματος περιπτώσεις συντοποθεσίας, όπου αν υπάρχει πλήρης αντιστοίχηση χαρακτηριστικών, με τους κατάλληλους εν συνεχεία ελέγχους γίνεται γνωστός ο αριθμός των ξεχωριστών περιπτώσεων του LHS στις περιπτώσεις LHS ∪ RHS και
- τέλος, διατρέχει τις αντίστοιχου μεγέθους με το μέγεθος του δεύτερου ορίσματος επικρατούσες συντοποθεσίες, όπου όποια αντιστοιχίζεται σε αυτό, απλά μας γνωστοποιεί τον αριθμό των περιπτώσεών της και κατ' επέκταση τον ζητούμενο παρανομαστή.

Μετά το πέρας της εν λόγω διαδικασίας, αυξάνεται κάθε φορά το k κατά 1, ενώ όσο υπάρχουν k-1 επικρατούσες συντοποθεσίες οι διαδικασίες 2 έως 7 επαναλαμβάνονται.

Τέλος, να τονιστεί πως κατα τη διάρκεια όλης της διαδικασίας σε κάθε κλάση χρησιμοποιήθηκαν οι κατάλληλοι μέθοδοι και "συγκριτές" (comparators) για τις συγκρίσεις μεταξύ των αντικειμένων της, τόσο για λόγους ταξινόμησης όσο και προς αποφυγή διπλότυπων εγγραφών.

Έχοντας ξεκινήσει, λοιπόν, από τη συλλογή των δεδομένων και έχοντας προεπεξεργαστεί αυτά, έχοντας δημιουργήσει τους κατάλληλους γράφους δικτύου για τις περιπτώσεις των προσεγγίσεων εύρεσης γειτόνων μέσω απόστασης δικτύου, έχοντας βρει τα γειτονικά ζεύγη αντικειμένων με τρεις διαφορετικούς τρόπους και έχοντας ακολουθήσει τα βήματα ενός αλγόριθμου εξόρυξης προτύπων και κανόνων συντοποθεσίας, το αποτέλεσμα που προκύπτει είναι της μορφής που φαίνεται στην επόμενη σελίδα.

Εικόνα 4.2- Αποτελέσματα ενός τρεξίματος κώδικα

```
================================================================
 Prevalent colocations-1:
================================================================
[sports_centre] (prev:1)
[restaurant]    (prev:1)
[post_office]   (prev:1)
[clothes]       (prev:1)
[supermarket]   (prev:1)
[attraction]    (prev:1)
[cafe]          (prev:1)
[library]       (prev:1)
[school]        (prev:1)
[pharmacy]      (prev:1)
[hotel]         (prev:1)
[playground]    (prev:1)
[atm]           (prev:1)
[pub]           (prev:1)
[fast_food]     (prev:1)
================================================================
================================================================
 Prevalent colocations-2:
================================================================
[atm, post_office]        (prev:0,3529)
[atm, pharmacy]           (prev:0,4194)
[pub, restaurant]         (prev:0,3349)
[fast_food, restaurant]   (prev:0,4)
[pharmacy, supermarket]   (prev:0,3226)
[fast_food, post_office]  (prev:0,3529)
[fast_food, pharmacy]     (prev:0,4252)
================================================================
================================================================
 Colocation rules-2
================================================================
[fast_food]--->[restaurant]     (Cond_Prob:0,4)
[atm]--->[pharmacy]             (Cond_Prob:0,4301)
[fast_food]--->[pharmacy]       (Cond_Prob:0,5161)
[post_office]--->[atm]          (Cond_Prob:0,5081)
[pharmacy]--->[atm]             (Cond_Prob:0,4194)
[restaurant]--->[fast_food]     (Cond_Prob:0,5714)
[post_office]--->[fast_food]    (Cond_Prob:0,4354)
[pharmacy]--->[fast_food]       (Cond_Prob:0,4252)
```

5.ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Για τη διεκπεραίωση των τρεξιμάτων του κώδικα (Java) και κατ' επέκταση για τη πειραματική αξιολόγηση της εργασίας που ακολουθήθηκε, χρησιμοποιήθηκαν τρεις διαφορετικές βάσεις δεδομένων, οι οποίες μετά τη προεπεξεργασία τους μέσω PostgreSQL, είχαν τα εξής στοιχεία :

	Ακμές	Κόμβοι	Σημεία Ενδιαφέροντος
Monaco	3695	3543	306
Leicestershire	16869	16345	1885
G.Manchester	30819	29519	4088

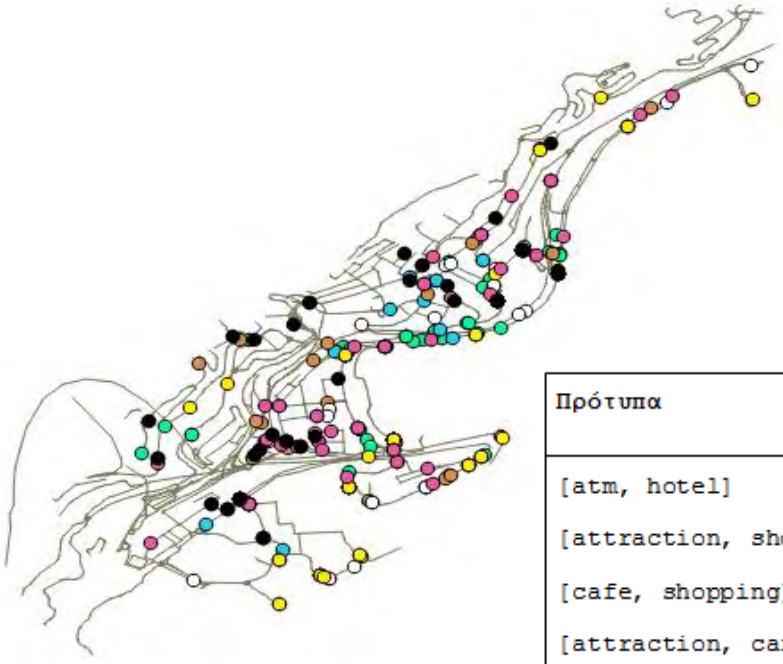
Τα σημεία ενδιαφέροντος των εν λόγω βάσεων, μέσω των οποίων έγιναν οι συγκρίσεις των διαφορετικών προσεγγίσεων εύρεσης γειτονικών αντικειμένων, βρίσκονται στις εικόνες 5.1-5.3.

Λόγω του γεγονότος ότι η προεπεξεργασία, ουσιαστικά, δεν αποτελούσε το κεντρικό θέμα της εργασίας και λόγω του ότι όσο πιο μεγάλο όγκο δεδομένων έχουμε να αντιμετωπίσουμε, τόσο πιο χρονοβόρα μπορεί να γίνει, σε συνδυασμό με τις υπόλοιπες απαιτήσεις που έπρεπε να ικανοποιηθούν, δεν κατέστη δυνατή η δημιουργία-διαμόρφωση μεγαλύτερων βάσεων, οι οποίες πιθανώς, να μας έδιναν ασφαλέστερα αποτελέσματα. Παρ' όλα αυτά τα δείγματα που "εξορύχτηκαν" μας οδηγούν σε παρόμοια συμπεράσματα με διάφορες μελέτες που σχετίζονται με τις προσεγγίσεις που χρησιμοποιήθηκαν. Επίσης, να τονιστεί ότι για τις συγκρίσεις που καταγράφονται ακολούθως, όσον αφορά τη προσέγγιση απόστασης δικτύου, χρησιμοποιήθηκε η μέθοδος [2] η οποία σε σχέση με τη μέθοδο βάσει Dijkstra που ακολουθήθηκε, όπως προαναφέρθηκε και στο προηγούμενο κεφάλαιο, μας παρείχε κατα πλειοψηφία πρότυπα λίγο υψηλότερου ενδιαφέροντος.

ΑΠΟΣΤΑΣΗ ΔΙΚΤΥΟΥ ΕΝΑΝΤΙ ΕΥΛΕΙΔΕΙΑΣ ΑΠΟΣΤΑΣΗΣ

Η πειραματική αξιολόγηση που ακολουθήθηκε, λοιπόν, μεταξύ των προσεγγίσεων εύρεσης γειτονικών αντικειμένων μέσω της Ευκλείδειας θεώρησης και μέσω της απόστασης δικτύου [2] έδειξε πως χρησιμοποιώντας το ίδιο όριο απόστασης, όπως επίσης και τα ίδια όρια επικράτησης και δεσμευμένης πιθανότητας, η τάση σχηματισμού προτύπων και κατ' επέκταση κανόνων συντοποθεσίας, μπορεί να αλλάξει σημαντικά. Αυτό, εύκολα, μπορεί να γίνει κατανοητό αν

ανατρέξει κανείς στις επόμενες τρείς εικόνες, οι οποίες εκτός των στιγμιότυπων μέσω QGIS και των πινάκων των σημείων ενδιαφέροντος κάθε περιοχής που χρησιμοποιήθηκε για την εξόρυξη προτύπων και κανόνων συντοποθεσίας, περιέχουν και από έναν πίνακα με πρότυπα συντοποθεσίας μεγέθους 2.



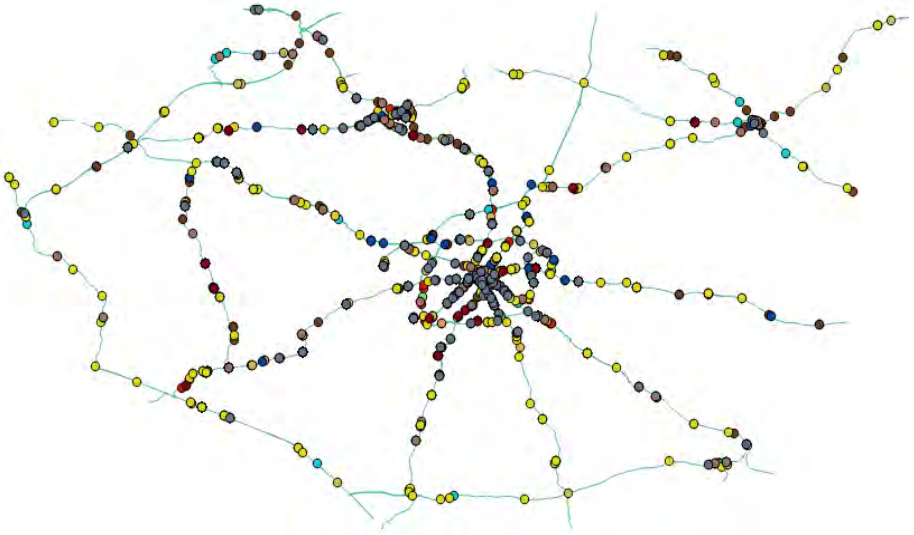
type character varying	count bigint
restaurant	56
shopping	37
atm	28
cafe	39
nightlife	39
attractions	79
hotel	28

Πρότυπα	Διαχωρισμός Περιεχομένου (PI)	
	Ευκλ.Απ.	Απ.Δικτύου
[atm, hotel]	(0,5)	(0,5714)
[attraction, shopping]	(0,7838)	(0,5946)
[cafe, shopping]	(0,6486)	(0,5676)
[attraction, cafe]	(0,7436)	(0,5641)
[hotel, nightlife]	(0,4872)	(0,4103)
[attraction, restaurant]	(0,7321)	(0,4821)
[atm, restaurant]	(0,3571)	(0,3036)
[cafe, hotel]	(0,6786)	(0,5714)
[cafe, nightlife]	(0,641)	(0,5385)
[cafe, restaurant]	(0,4643)	(0,3929)
[nightlife, shopping]	(0,6216)	(0,5135)
[restaurant, shopping]	(0,7568)	(0,4865)
[atm, cafe]	(0,4103)	(0,4103)
[attraction, hotel]	(0,8214)	(0,6582)
[hotel, shopping]	(0,5405)	(0,4054)
[atm, shopping]	(0,4324)	(0,3784)
[attraction, nightlife]	(0,641)	(0,5897
[nightlife, restaurant]	(0,5)	(0,3393)
[atm, nightlife]	(0,4103)	(0,3333)

Εικόνα 5.1 -

Πρότυπα συντοποθεσιών μεγέθους 2,

στη χώρα του Μονακό με όριο απόστασης 500m



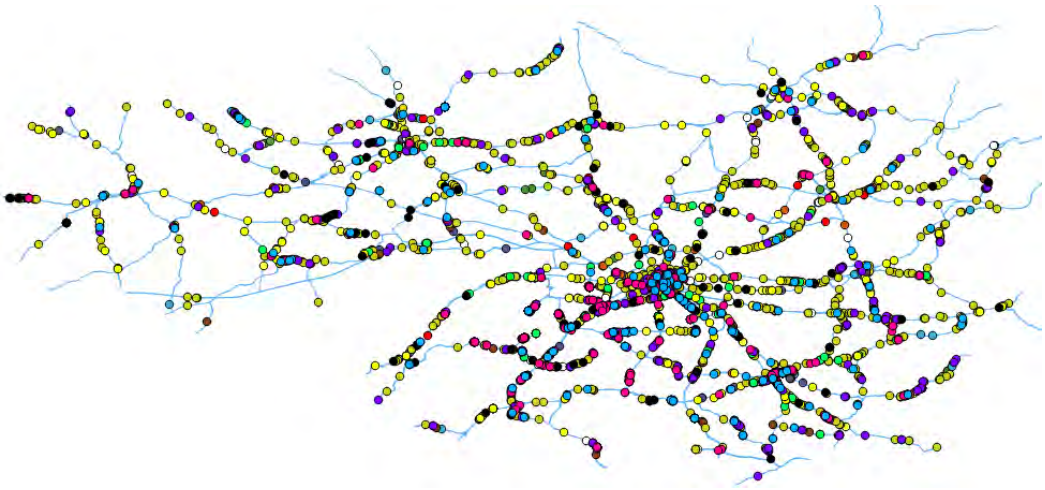
type character varying	count bigint
atm	124
attraction	28
cafe	200
clothes	94
fast_food	294
hotel	41
library	36
pharmacy	93
playground	59
post_office	136
pub	433
restaurant	225
school	35
sports_centre	22
supermarket	65

Πρότυπα	Διαχωρισμός Περιεχομένου (PI)	
	Ευκλ.Απ.	Απ.Δικτύου
[atm, post_office]	(0,4044)	(0,3824)
[atm, pharmacy]	(0,5269)	(0,5269)
[restaurant, school]	**(0,3429)**	**(0,3822)**
[cafe, restaurant]	(0,4311)	(0,36)
[pharmacy, supermarket]	(0,4731)	(0,3871)
[fast_food, hotel]	(0,449)	(0,3129)
[cafe, post_office]	(0,3162)	(0,3015)
[pub, restaurant]	**(0,4411)**	**(0,4688)**
[fast_food, restaurant]	(0,5644)	(0,5244)
[cafe, pharmacy]	(0,4194)	(0,4194)
[fast_food, post_office]	**(0,4412)**	**(0,4779)**
[cafe, clothes]	(0,575)	(0,505)
[pharmacy, post_office]	(0,3456)	(0,3162)
[restaurant, supermarket]	(0,5077)	(0,3244)
[atm, playground]	(0,3559)	(0,3306)
[atm, supermarket]	**(0,3692)**	**(0,379)**
[fast_food, pharmacy]	(0,5914)	(0,5612)

Εικόνα 5.2 -

Πρότυπα συντοποθεσιών μεγέθους 2,

στη περιοχή Leicestershire με όριο απόστασης 700m



type character varying	count bigint
fire_station	22
atm	376
attraction	58
parking	11
pharmacy	139
arts_centre	92
fast_food	533
restaurant	413
shopping	399
supermarket	163
cafe	298
nightlife	1149
hotel	82
library	62
school	103
post_office	188

Πρότυπα	Διαχωρισμός Περιεχομένου (PI)	
	Ευκλ.Απ.	Απ.Δικτύου
[restaurant, supermarket]	(0,4724)	(0,4601)
[fast_food, library]	(0,4503)	(0,3302)
[nightlife, school]	(0,3377)	(0,3107)
[fast_food, shopping]	(0,5063)	(0,4511)
[fast_food, supermarket]	(0,5347)	(0,5291)
[atm, hotel]	(0,3777)	(0,3005)
[cafe, shopping]	(0,401)	(0,3935)
[arts_centre, hotel]	(0,378)	(0,3293)
[cafe, pharmacy]	**(0,4532)**	**(0,4604)**
[cafe, restaurants]	(0,3947)	(0,3753)
[fast_food, post_office]	**(0,5053)**	**(0,516)**
[cafe, fast_food]	(0,3114)	(0,3058)
[pharmacy, post_office]	**(0,3191)**	**(0,3298)**
[atm, supermarket]	**(0,6196)**	**(0,6258)**
[fast_food, restaurants]	(0,46)	(0,4431)
[nightlife, pharmacy]	(0,4769)	(0,4473)
[atm, shopping]	(0,3935)	(0,3835)
[atm, library]	(0,4282)	(0,3165)
[fast_food, pharmacy]	(0,6473)	(0,6173)
[atm, attractions]	(0,3564)	(0,3112)
.....	.....	.....

Εικόνα 5.3 -

Στιγμιότυπο προτύπων συντοποθεσιών μεγέθους 2, στη περιοχή Great Manchester με όριο απόστασης 800m

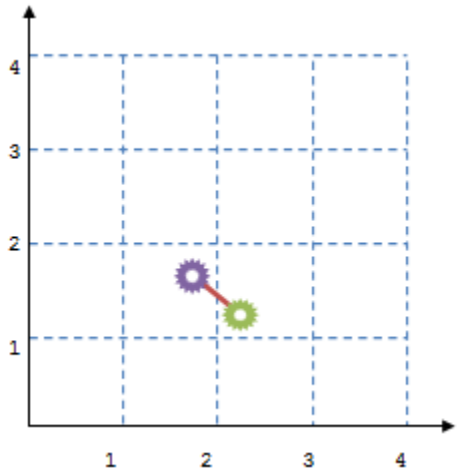
Στις σελίδες που προηγήθηκαν, με μία προσεκτική ματιά, μπορούμε να δούμε πως ο διαχωρισμός περιεχομένου μεταξύ των δύο διαφορετικών προσεγγίσεων, στη μεγάλη πλειοψηφία των περιπτώσεων, είναι μεγαλύτερος όταν χρησιμοποιείται η Ευκλείδεια θεώρηση του χώρου. Αυτό είναι μία λογική απόρροια, καθώς αρκετά γειτονικά ζεύγη τα οποία προκύπτουν μέσω αυτής, δεν λαμβάνονται υπόψη στις προσεγγίσεις που χρησιμοποιείται η έννοια της απόστασης δικτύου. Ο λόγος που συμβαίνει αυτό, δεν είναι άλλος από το γεγονός ότι σε έναν γράφο δικτύου, ο χώρος, όπως προαναφέρθηκε και στο κεφάλαιο 3, δεν λαμβάνεται υπόψη ως ομοιογενής και ισοτροπικός, λόγω των περιορισμών που υπάρχουν ή προκύπτουν σε αυτόν. Ως περιορισμούς μπορούμε να θεωρήσουμε, τους δρόμους ενός οδικού δικτύου, όπως αυτοί είναι διαμορφωμένοι στην εκάστοτε χωρική εμβέλεια, τα κτίρια και τα διάφορα άλλα σημεία ενδιαφέροντος που βρίσκονται κατά μήκος τους, αλλά και κάποιο πιθανό ατύχημα που έχει προκληθεί και έχει προκαλέσει μποτιλιάρισμα. Όπως εύκολα γίνεται κατανοητό, λοιπόν, η λήψη λιγότερων αντικειμένων, έχει επίπτωση και στις μετρήσεις που χρησιμοποιούνται τόσο για την εξόρυξη προτύπων όσο και κατ' επέκταση, για την εξόρυξη κανόνων συντοποθεσίας. Έτσι, όπως θα μπορούσε να πει κανείς, οι προσεγγίσεις εύρεσης γειτονικών αντικειμένων μέσω Ευκλείδειας απόστασης, δημιουργούν κατα βάση πρότυπα υπερεκτιμημένου ενδιαφέροντος, διαταράσσοντας την έννοια των ελκυστικών προτύπων. Βέβαια, στις συγκρίσεις που ακολουθήθηκαν, υπήρξαν και κάποιες περιπτώσεις (όπως τα πρότυπα των πινάκων με έντονα γράμματα), κατα τις οποίες μπορεί να υποτιμηθεί η απόσταση μεταξύ αντικειμένων, όπως όταν αυτά πρόσκεινται συχνά στην ίδια ακμή.

Έστω, λοιπόν, το ακόλουθο παράδειγμα κατά το οποίο ας θεωρήσουμε πως το μωβ αντικείμενο (Εικόνα 5.4) βρίσκεται στη θέση {x = 1.7, y = 1.8} και το πράσινο αντικείμενο στη θέση {x = 1.2, y = 1.4}. Όσον αφορά τον υπολογισμό της απόστασής τους μέσω της Ευκλείδειας θεώρησης, αυτός θα έχει ως εξής :

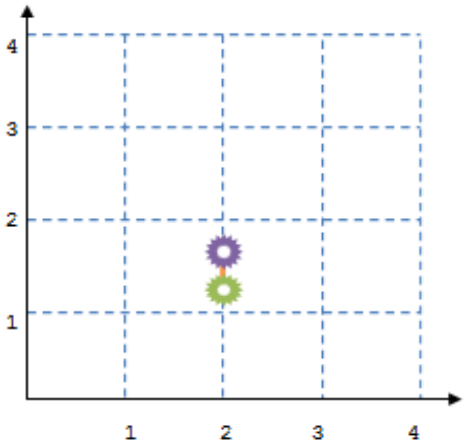
$$d1 = \sqrt{(1.2 - 1.7)^2 + (1.4 - 1.8)^2} = \sqrt{0.41} \approx 0.6403$$

Από την άλλη μεριά, όσον αφορά την προσέγγιση απόστασης δικτύου (Εικόνα 5.5) βάσει συστήματος γραμμικής αναφοράς, τα αντικείμενα προβάλλονται στο κοντινότερο σημείο της κοντινότερης ακμής τους και η θέση τους λαμβάνεται σε σχέση με τη θέση του εκάστοτε αρχικού κόμβου της ακμής αυτής. Συνεπώς, το μωβ αντικείμενο, πλέον, θα προβληθεί στην ακμή {x_1 = 2, y_1 = 1 , x_2 = 2, y_2 = 2} και θα έχει θέση pos = 1.8 - 1 = 0.8, ενώ το πράσινο που θα προβληθεί και αυτό στην ίδια ακμή θα έχει θέση pos = 1.4 - 1 = 0.4. Οπότε η απόσταση δικτύου μεταξύ των δύο αυτών αντικειμένων, θα έχει ως εξής :

$$d2 = |\, 0.4 - 0.8 \,| = 0.4 < d1$$



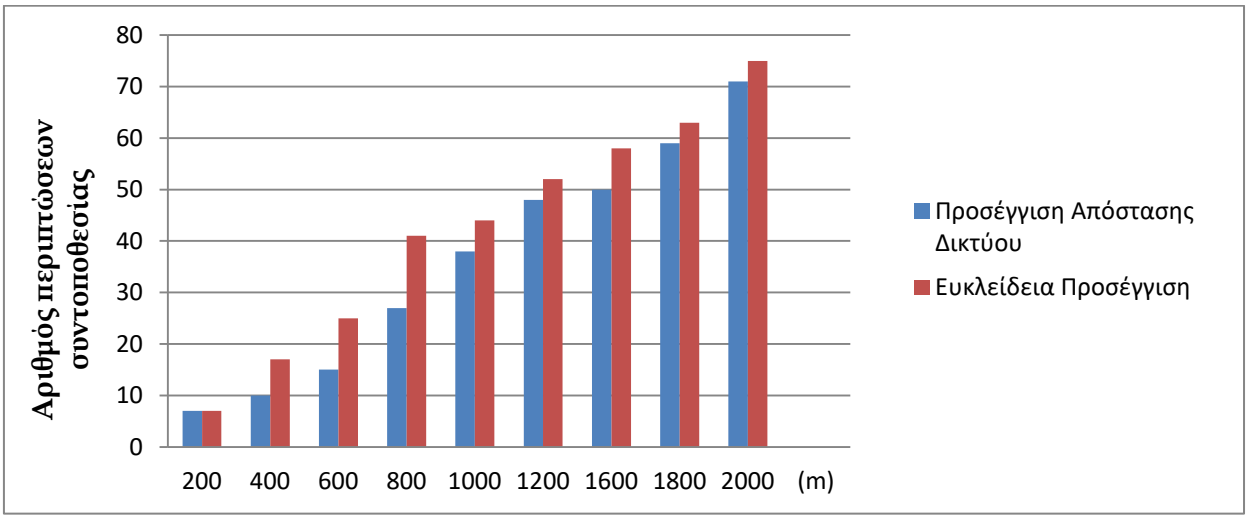
Εικόνα 5.4 – Ευκλείδεια απόσταση



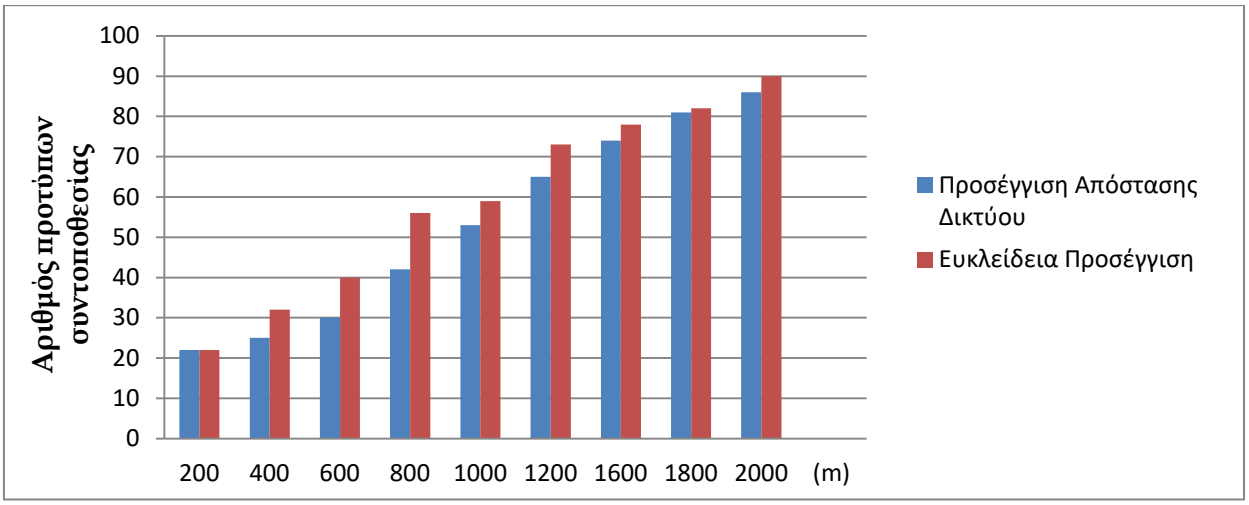
Εικόνα 5.5 – Απόσταση Δικτύου βάσει γραμμικής αναφοράς

Σε αυτό το σημείο να τονιστεί πως στο μεγαλύτερο ποσοστό των περιπτώσεων κατά το οποίο βρέθηκε πιο υψηλός διαχωρισμός περιεχομένου μέσω της προσέγγισης εύρεσης γειτονικών αντικειμένων με απόσταση δικτύου, αυτός ήταν πολύ κοντά με τον διαχωρισμό περιεχομένου της άλλης προσέγγισης.

Ο διαχωρισμός περιεχομένου, όμως, είναι η βασική μέτρηση με την οποία καθορίζονται-δημιουργούνται τα εκάστοτε πρότυπα. Όσο πιο μεγάλο διαχωρισμό περιεχομένου έχει μία υποψήφια συντοποθεσία, τόσο και περισσότερες οι πιθανότητες να είναι επικρατούσα βάσει το εκάστοτε όριο χρήστη. Οι γράφοι που ακολουθούν (5.1 και 5.2), είναι αποτέλεσμα εκτυπώσεων των αριθμών περιπτώσεων και προτύπων συντοποθεσίας με ίδιο όριο διαχωρισμού περιεχομένου στη βάση του Leicestershire και μας δείχνουν αυτό ακριβώς που ειπώθηκε προηγουμένως, ότι δηλαδή, κατά τη προσέγγιση εύρεσης γειτονικών αντικειμένων μέσω Ευκλείδειας θεώρησης, στις περισσότερες των περιπτώσεων συναντάμε πρότυπα υπερτιμημένου ενδιαφέροντος. Αυτό, θεωρητικά, σημαίνει πως πρότυπα τα οποία συναντάμε στις εν λόγω προσεγγίσεις, μπορεί να μην προκύπτουν καν μέσω των προσεγγίσεων που χρησιμοποιούν την έννοια της απόστασης δικτύου. Κατά τη πειραματική αξιολόγηση συναντήθηκε και η αντίθετη περίπτωση, αλλά μειοψηφικά.



Γράφος 5.1 -
Σύγκριση αριθμού περιπτώσεων συντοποθεσίας μέσω των δύο διαφορετικών προσεγγίσεων εύρεσης απόστασης



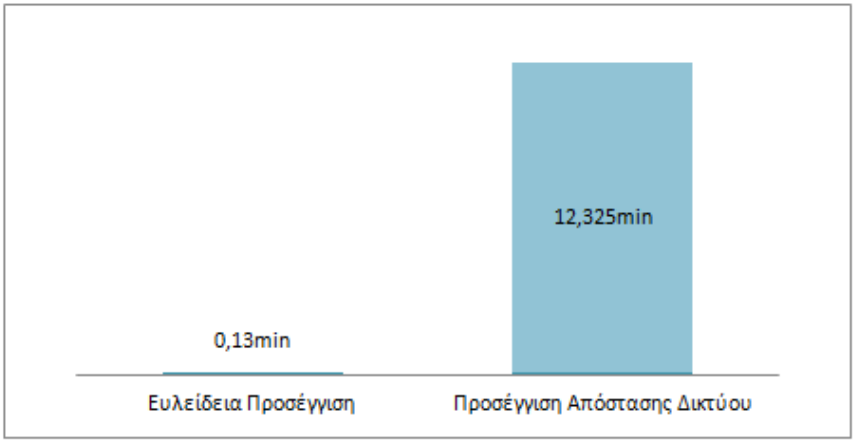
Γράφος 5.2 -
Σύγκριση αριθμού προτύπων συντοποθεσίας μέσω των δύο διαφορετικών προσεγγίσεων εύρεσης απόστασης

ΕΥΛΕΙΔΕΙΑ ΑΠΟΣΤΑΣΗ ΑΠΟΣΤΑΣΗΣ ΔΙΚΤΥΟΥ ΕΝΑΝΤΙ

Μέχρι αυτό το σημείο, ουσιαστικά, είδαμε τα πλεονεκτήματα των προσεγγίσεων απόστασης δικτύου έναντι αυτών που χρησιμοποιούν την Ευκλείδεια θεώρση του χώρου. Παρ' όλα αυτά, οι πρώτες έχουν να αντιμετωπίσουν προκλήσεις που αφορούν το υπολογιστικό κόστος και τον υπολογιστικό χρόνο των λειτουργιών τους.

Ενδεικτικά στο γράφο (5.3) που ακολουθεί, φαίνεται ο χρόνος προεπεξεργασίας που χρειάστηκε προκειμένου να διαμορφωθούν τα δεδομένα του Leicestershire κατάλληλα, ώστε να χρησιμοποιηθούν οι αντίστοιχες προσεγγίσεις. Τα 13 δευτερόλεπτα είναι κοινά και

για τις δύο περιπτώσεις και δεν είναι άλλα, από τα δευτερόλεπτα κατα τα οποία έγινε ο διαχωρισμός των σημείων ενδιαφέροντος και η προσθήκη αυτών σε έναν κοινό πίνακα. Τα επιπλέον 12.35 λεπτά που χρειάστηκαν για τη χρησιμοποίηση της προσέγγισης απόστασης δικτύου, είχαν να κάνουν τόσο με το διαχωρισμό των δρόμων σε κόμβους και ακμές, όσο και με τη προβολή των σημείων ενδιαφέροντος στις πιο κοντινές τους ακμές βάσει γραμμικής αναφοράς. Σίγουρα θα υπάρχουν και πιο γρήγοροι τρόποι από αυτόν που ακολουθήθηκε, αλλά θεωρητικά είναι αρκετά δύσκολο αυτοί να φτάνουν τους χρόνους που απαιτούνται για τη χρησιμοποίηση της Ευκλείδειας προσέγγισης. Να τονιστεί, επίσης, πως έχοντας ουσιαστικά έτοιμες τις εντολές λόγω προηγούμενων προεπεξεργασιών, εκτός του χρόνου που φαίνεται στο κατωτέρω διάγραμμα, χρειάστηκαν συνολικά γύρω στις δυόμιση εργατοώρες.



Γράφος 5.3 -
Χρόνος προεπεξεργασίας δεδομένων για τις δύο διαφορετικές προσεγγίσεις εύρεσης απόστασης

Όσον αφορά την υπόλοιπη διαδικασία, υπάρχει τόσο το "χτίσιμο" του εκάστοτε γράφου μέσω των προεπεξεργασμένων δεδομένων, όσο και εν τέλει η διαδικασία εύρεσης απόστασης δικτύου, η οποία μέσω των περιορισμών που λαμβάνει υπόψη της είναι πιο χρονοβόρα. Λόγω του ότι οι βάσεις που χρησιμοποιήθηκαν είχαν σχετικά μικρό αριθμό δεδομένων, με εξαίρεση τη βάση του Greater Manchester, στις άλλες δύο, οι χρόνοι ήταν μικροί και για τις δύο περιπτώσεις. Ενδεικτικά, όμως, για τη βάση του Greater Manchester, με όριο απόστασης τα 400 μέτρα, ο χρόνος που απαιτήθηκε για την εύρεση αποστάσεων δικτύου μεταξύ των σημείων ενδιαφέροντος ήταν 81.87 δευτερόλεπτα, εν αντιθέσει με τη προσέγγιση Ευκλείδειας θεώρησης που ήταν 11.28 δευτερόλεπτα.

ΕΝ ΚΑΤΑΚΛΕΙΔΙ

Με την όλο και αυξανομένη χρησιμοποίηση υπηρεσιών οι οποίες χειρίζονται χωρικά δεδομένα, όπως οι εφαρμογές Συστημάτων Γεωγραφικών Πληροφοριών και οι εφαρμογές Υπηρεσιών με βάση τη θέση, γίνεται όλο και μεγαλύτερη η ανάγκη για εξαγωγή αυτόματης, ακριβούς και περίπλοκης – ειδικότερα σε επίπεδο επιστημονικής έρευνας – γνώσης. Η ζητούμενη ακρίβεια αφορά σε μεγάλο βαθμό τη κατανόηση του ευρύτερου γεωγραφικού πλαισίου και εξαρτάται από το συνδυασμό δύο βασικών παραγόντων, οι οποίοι δεν είναι άλλοι από τη προεπεξεργασία και την εξόρυξη των εκάστοτε δεδομένων. Η προεπεξεργασία, όπως είδαμε στο εν λόγω σύγγραμμα, εξαρτάται από τη προσέγγιση που θα ακολουθηθεί με κατεύθυνση την εύρεση των ζευγών γειτονικών αντικειμένων σε μία χωρική εμβέλεια. Όσον αφορά τις προσεγγίσεις που χρησιμοποιούν την Ευκλείδεια θεώρηση, τα πράγματα είναι πιο απλά, καθώς ο χώρος λαμβάνεται ως ομοιογενής και ισοτροπικός και οι αποστάσεις είναι αποτέλεσμα μέτρησης του μεγέθους μίας ευθείας νοητής γραμμής μεταξύ του κάθε ζεύγους. Έτσι, το μόνο που απαιτείται είναι οι θέσεις αυτών πάνω στις συντεταγμένες της Γης. Αυτό σημαίνει, προφανώς πολύ λιγότερο χρονοβόρα προεπεξεργασία, τόσο από άποψη υπολογιστικού χρόνου και κόστους, όσο και από άποψη εργατοωρών. Παρ' όλα αυτά, τόσο οι ανθρώπινες δραστηριότητες, όσο και οι δραστηριότητες διάφορων γεγονότων που προκύπτουν πάνω στο χώρο, υπόκεινται, στις περισσότερες των περιπτώσεων, στους διάφορους περιορισμούς που υπάρχουν σε αυτόν, όπως ο τρόπος διαμόρφωσης του οδικού δικτύου ή ο τρόπος εξάπλωσης των σημείων ενδιαφέροντος. Έτσι, ουσιαστικά, μπορούμε να πούμε πως οι προσεγγίσεις που χρησιμοποιούν την έννοια της απόστασης δικτύου, έρχονται να ακυρώσουν τις προαναφερθείσες, προσφέροντας πιο αξιόπιστα πρότυπα σε συνδυασμό με τη διαδικασία εξόρυξης. Αυτά τα πρότυπα, μπορούν να διαδραματίσουν σημαντικό ρόλο σε λειτουργίες που αφορούν από τον επαναπροσδιορισμό του σχεδιαγράμματος ενός ταξιδιού μέχρι και τον επαναπροσδιορισμό μίας επιστημονικής μελέτης. Ωστόσο, με τη χρησιμοποίηση των εν λόγω προσεγγίσεων, προκύπτουν σημαντικές προκλήσεις που έχουν να κάνουν με το υπολογιστικό κόστος και χρόνο, τόσο στη διαδικασία της προεπεξεργασίας, όσο στη διαδικασία χτισίματος του εκάστοτε γράφου δικτύου, όσο και στη διαδικασία εύρεσης των ζευγών γειτονικών αντικειμένων, καθώς σε αυτά τα σημεία, οι προσεγγίσεις που χρησιμοποιούν την Ευκλείδεια θεώρηση του χώρου, υπερτερούν.

ΒΙΒΛΙΟΓΡΑΦΙΑ - ΑΝΑΦΟΡΕΣ

[1] Yoo J.S. & Shekhar S.(2006).A joinless approach for mining spatial colocation patterns. IEEE Transactions on Knowledge & Data Engineering,18(10),1323-1337.

[2] Wenhao Yu (2016). Spatial co-location pattern mining for location-based services in road networks. Expert Systems With Applications 46 (2016) 324–335

[3] Dijkstra E.W.(1959). A note on two problems in connexion with graphs. Numerische Mathematik, 1(1), 269-271.

[4] Fayyad U. M. 1997 "Knowledge Discovery in Databases: An Overview", ILP 1997, pp. 3-16.

[5] Fayyad U. M., Piatetsky-Shapiro G., and Smyth P. 1996 "Knowledge Discovery and Data Mining : Towards a Unifying Framework", Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, Oregon, AAAI Press, Menlo Park, California, pp. 82 - 88.

[6] R. Agrawal, T. Imielinski, and A. Swami (1993). Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993.

[7] Agrawal R. & Srikant R.(1994). Fast algorithms for mining association rules. In Proceedings of the 20th international conference on very large data bases (VLDB'94) (pp. 487–499)

[8] F. Bodon, L. Rónyai. Trie: An alternative data structure for data mining algorithms. Mathematical and Computer Modelling : An International Journal archive. Volume 38 Issue 7-9, October, 2003 Pages 739-751

[9] Roddick J.-F. and Spiliopoulou M. 1999. A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. SIGKDD Explorations 1(1): 34-38 (1999).

[10] Shekhar S. and Chawla S. 2002. A Tour of Spatial Databases. Prentice Hall (ISBN 0-7484-0064-6).

[11] Ester M., Kriegel H.-P. and Sander J., Spatial Data Mining: A Database Approach, proceedings of 5th International Symposium on Advances in Spatial Databases, pages 47-66, 1997

[12] Koperski K. & Han J.(1995).Discovery of spatial association rules in geographic information databases. In Proceedings of 4th international symposium on large spatial databases(pp.47–66). Portland, MaineSpringer.

[13] Shekhar S. & Huang Y.(2001).Discovering spatial co-location patterns : a summary of results. In Proceedings of Advances in Spatial and Temporal Databases(pp.236–256).Berlin Springer.

[14] Huang Y., Shekhar S. & Xiong H.(2004).Discovering colocation patterns from spatial datasets : a general approach. IEEE Transactions on Knowledge & Data Engineering,16(12),1472–1485.

[15] C. Berge. Graphs and Hypergraphs. American Elsevier, 1976.

[16] Y. Morimoto. Mining Frequent Neighboring Class Sets in Spatial Databases. In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001.

[17] Miller H.J.(1994).Market area delimitation within networks using geographic information systems.GeographicalSystems,1,157–173.

[18] Okabe A., Okunuki K.I. & Shiode S.(2006).The SANET toolbox : new methods for network spatial analysis. Transactions in GIS,10(4),535–550.

[19] Okabe A., Boots B., Sugihara K. & Chiu S.N.(2009).Spatial tessellations : concepts and applications of Voronoi diagrams. Chiche-ster, UK : Wiley.

[20] Yamada I. & Thill J.C.(2007).Local indicators of network-constrained clusters in spatial point patterns. Geographical Analy-sis,39(3),268–292.

[21] L. Arge, O. Procopiuc, S. Ramaswamy, T. Suel, and J.Vitter, “Scalable Sweeping-Based Spatial Join,” in Proc. of the Int’l Confe-rence on Very Large Databases, 1998.

[22] D. Papadias, N. Mamoulis, and Y. Theodoridis, “Processing and Optimization of Multiway Spatial Joins Using R-Trees,” in POPS, 1999.

[23] H. Park, G. Cha, and C. Chung, “Multi-way Spatial Joins Using R- Trees: Methodology and Performance Evaluation,” in SSD, 1999.

[24] J. Yoo and S. Shekhar, “A Partial Join Approach for Mining Co-location Patterns,” in Proc. of ACM International Symposium on Advances in Geographic Information Systems(ACM-GIS), 2004

[25] Appice A.,Ceci M.,Lanza A.,Lisi F.A. & Malerba D.(2003).Discovery of spatial association rules in georeferenced census data : a relational mining approach. In Proceedings of the intelligent data analysis(pp.541–566).

[26] Santos M.Y. & Amaral L.(2005). Geospatial data mining in the analysis of a demographic database. Soft Computing,9(5),374–384.

[27] Bembenik R. & Rybinski H.(2009). FARICS : a method of mining spatial association rules and collocations using clustering and delaunay diagrams. Journal of Intelligent Information Systems,33(1),41-64.

[28] Clementini E., Felice P.D. & Koperski K.(2000).Mining multiple-levels patial association rules for objects with a broad boundary. Data & KnowledgeEngineering,34(3),251-270.

[29] Yu W., Ai T. & Shao S.(2015).The analysis and delimitation of Central Business District using network kernel density estimation. Journal of Transport Geography,45,32-47.

[30] Okabe A., Satoh T., Furuta T., Suzuki A. & Okano K.(2008).Generalized network Voronoi diagrams : concepts, computational methods, and applications. International Journal of Geographical Information Science,22(9),965-994.

[31] Xie Z. & Yan J.(2008).Kernel density estimation of traffic accidents in a network space. Computers, Environment and Urban Systems,35(5),396-406.

[32] Shiode S.(2011).Street- level spatial scan statistic and STAC for analysing street crime concentrations. TransactionsinGIS,15(3),365-383.

[33] J. Yoo and S. Shekhar, "A Join-less Approach for Co-location Pattern Mining: A Summary of Results," in Proc. of IEEE International Conference on Data Mining(ICDM), 2005.

[34] Papadias D., Zhang J.,Mamoulis N. & Tao Y.(2003). Query processing in spatial network databases. In Proceedings of the 29th conference on very large databases(VLDB)(pp.790-801).

[35] J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000.

ΕΙΚΟΝΕΣ-ΑΛΓΟΡΙΘΜΟΙ

Εικόνα 1.1 και 1.2 - http://mapschool.io/

Εικόνα 2.1 - https://onthe.io/

Εικόνα 2.3 - https://line.do/

Εικόνα ενότητας 3.1 - http://www.esri.com/

Αλγόριθμος 2.1 - [7]

Κομμάτι κώδικα σύνταξης SQL (υποενότητα 2.5.3) - [14]