



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΠΡΟΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

**Εύρεση των πιο ποικιλόμορφων προϊόντων με ανάλυση
αποτελεσμάτων επερωτήσεων σε χωρο-κειμενικών
δεδομένων**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της/του

Τσίπου Ιωάννη

Επιβλέπων : Αν. Καθηγήτρια Ακριβή Βλάχου, Πανεπιστήμιο Αιγαίου

Μέλη εξεταστικής επιτροπής: Αν. Καθηγητής Θεόδωρος Κωστούλας, Πανεπιστήμιο Αιγαίου,
Αν. Καθηγητής Παναγιώτης Συμεωνίδης, Πανεπιστήμιο Αιγαίου

Σάμος, [Μάρτιος 2022]

Πρόλογος και ευχαριστίες

Ευχαριστώ την επιβλέπουσα καθηγήτρια κυρία Βλάχου Ακριβή για την βοήθεια και την καθοδήγηση που είχα στην περίοδο της διπλωματικής μου. Επίσης θα ήθελα επίσης να ευχαριστήσω και την οικογένεια μου και τους φίλους μου παρείχαν στήριξη σε όλη την διάρκεια των σπουδών μου και με βοήθησαν σε κάθε δυσκολία.

© 2022

του/της

Τσίπου Ιωάννη

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Περιεχόμενα

1	Εισαγωγή	1
1.1	Ανάκτηση δεδομένων με βάση την τοποθεσία	1
1.2	Βασικές έννοιες.....	2
1.2.1	Χωρικά δεδομένα(Spatial Dataset)	2
1.2.2	Geographic Information System(GIS)	3
1.2.3	Χωρικές βάσεις δεδομένων(Spatial Databases).....	4
1.2.4	Χωρικά Ερωτήματα(Spatial queries)	5
1.3	Αντικείμενο διπλωματικής.....	6
1.4	Δομή της διπλωματικής	6
2	Ορισμός του Προβλήματος.....	8
2.1	Ερωτήσεις κοντινότερου γείτονα με λέξεις κλειδιά.....	8
2.2	Συνάρτηση ομοιότητας και ποικιλομορφία	13
2.3	Παράδειγμα συνάρτησης ομοιότητας	15
2.4	Εξαιρέσεις.....	17
3	Παρόμοιες Εργασίες.....	18
3.1	Keyword Search on Spatial Databases.....	18
3.2	Keyword Search in Spatial Databases: Towards Searching by Document	19
3.3	Scalable Top-K Spatial Keyword Search.....	20
3.4	Dynamically Ranked Top-K Spatial Keyword Search	20
3.5	Παρούσα διπλωματική.....	21
4	Πειραματική αποτίμηση της λύσης - Αποτελέσματα	22
	Για να δούμε την απόδοση της μεθόδου μας, καθώς και την ικανότητα να επιλέγει ποικιλόμορφα σημεία ενδιαφέροντος αποτιμήσαμε πειραματικά την προτεινόμενη μέθοδο.	22
4.1	Δεδομένα διπλωματικής.....	22
4.1.1	Αρχείο με όλα τα εστιατόρια.....	22
4.1.2	Αρχείο log	23
4.1.3	Πρόσθετες πληροφορίες υλοποίησης	24
4.2	Αναλυτική υλοποίηση.....	25
4.2.1	Δοκιμές με διαφορετικά αρχεία ερωτήσεων.....	25
5	Ανάλυση αποτελεσμάτων και συμπεράσματα.....	39
5.1	Αξιοπιστία αποτελεσμάτων	39
5.1.1	Περιγραφή δεδομένων.....	39
5.1.2	Ανάλυση αποτελεσμάτων	40

5.2	Προβλήματα που προέκυψαν.....	42
5.3	Εξέταση δυνατότητας πρακτικής εφαρμογής	42
6	Συμπεράσματα.....	43
	Βιβλιογραφία.....	44

Λίστα Εικόνων

Εικόνα 1: Βασικά συστατικά ενός LBS.....	2
Εικόνα 2: Τα χωρικά δεδομένα με τη μορφή σημείων, γραμμών, πολυγώνων κ.λπ.....	4
Εικόνα 3: Παράδειγμα μία βάσης χωρικών δεδομένων(Spatial Database).....	5
Εικόνα 4: Παράδειγμα μορφής του αρχείου με τα εστιατόρια και τα χαρακτηριστικά τους.....	6
Εικόνα 5: Παράδειγμα μορφής του αρχείου με τα εστιατόρια και τα χαρακτηριστικά τους.....	6
Εικόνα 6: Παράδειγμα αρχείου εστιατορίων.....	9
Εικόνα 7: Εξίσωση εύρεσης απόστασης μεταξύ δύο σημείων.....	9
Εικόνα 8: Λεπτομερές πίνακας εστιατορίων και αποτελέσματα ερωτήσεων για το καθένα.....	12
Εικόνα 9: Τύπος εξίσωσης Jaccard.....	13
Εικόνα 10: Σχήμα τομής δύο σημείων.....	14
Εικόνα 11: Σχήμα ένωσης μεταξύ δύο σημείων.....	14
Εικόνα 12: Εξίσωση ευκλείδειας απόστασης δύο σημείων.....	16
Εικόνα 13: Παράδειγμα R-Tree για ένα σύνολο δεδομένων.....	19
Εικόνα 14: Ερώτημα mCK για 3 keywords.....	19
Εικόνα 15: Δομή του I3.....	20
Εικόνα 16: Δομή STARI.....	21
Εικόνα 17: Τυχαία εστιατόρια από το αρχείο.....	23
Εικόνα 18: Παράδειγμα ερωτήσεων στο αρχείο log.....	24
Εικόνα 19: Αποτελέσματα top-k.....	26
Εικόνα 20: Αποτελέσματα top-k δεύτερου αρχείου.....	31
Εικόνα 21: Αποτελέσματα top-k τρίτου αρχείου.....	34
Εικόνα 22: Διάγραμμα χρόνων εκτέλεσης.....	40
Εικόνα 23: Πίτα διαφορετικών ειδών εστιατορίων.....	41

Λίστα Πινάκων

Πίνακας 1: Αποτελέσματα πρώτης ερώτησης	10
Πίνακας 2: Αποτελέσματα δεύτερης ερώτησης.....	10
Πίνακας 3: Αποτελέσματα τρίτης ερώτησης	11
Πίνακας 4: Αποτελέσματα τέταρτης ερώτησης.....	11
Πίνακας 5: Αποτελέσματα πέμπτης ερώτησης.....	12
Πίνακας 6: Ερωτήσεις πρώτου αρχείου.....	26
Πίνακας 7: Τριάδα πρώτης δοκιμής.....	28
Πίνακας 8: Τετράδα πρώτης δοκιμής.....	29
Πίνακας 9: Ερωτήσεις δεύτερου αρχείου.....	30
Πίνακας 10: Τριάδα δεύτερης δοκιμής.....	32
Πίνακας 11: Τετράδα δεύτερης δοκιμής.....	33
Πίνακας 12: Ερωτήσεις τρίτου αρχείου.....	34
Πίνακας 13: Ερωτήσεις τέταρτου αρχείου.....	36
Πίνακας 14: Τριάδα τέταρτης δοκιμής.....	37
Πίνακας 15: Τετράδα τέταρτης δοκιμής.....	38
Πίνακας 16: Μέγιστες και ελάχιστες τιμές συντεταγμένων.....	39

Περίληψη

Τα τελευταία χρόνια έχουν δημιουργηθεί περισσότερες εφαρμογές που βασίζονται στην τοποθεσία ενός χρήστη με αποτέλεσμα να εμφανίζει τις κατάλληλες πληροφορίες με βάση αυτήν την τοποθεσία. Υπάρχουν πολλά συστήματα που το κάνουν αυτό χρησιμοποιώντας χώρο – κειμενικά δεδομένα και εξελίσσονται ολοένα και περισσότερο για να ικανοποιήσουν τις ανάγκες όλο και περισσότερων χρηστών. Πλέον πολλές από αυτές δίνουν την δυνατότητα σε χρήστες να αναζητεί σημεία ενδιαφέροντος με βάση τις προτιμήσεις του και την πιο κοντινή απόσταση από αυτούς. Αυτά τα σημεία μπορεί να είναι επιχειρήσεις που παράγουν προϊόντα ή προσφέρουν διάφορες υπηρεσίες, διάφορα μαγαζιά όπως εστιατόρια, σούπερ – μάρκετ, ρούχων, τεχνολογίας κ.α. Στην Διπλωματική αυτή θα ασχοληθούμε με την κατάταξη κάποιων σημείων ενδιαφέροντος και συγκεκριμένα εστιατορίων με βάση την απόσταση τους από τον χρήστη και κάποιων λέξεων – κλειδιών που τον ενδιαφέρει να αναζητήσει και από τα αποτελέσματα διαλέγονται τα πιο ανάμοιο μεταξύ τους έτσι ώστε να υπάρχει αυτό που λέμε ποικιλομορφία στα αποτελέσματα μας δηλαδή πολλά και διαφορετικά είδη εστιατορίων. Για όλα τα παραπάνω χρησιμοποιήθηκαν αλγόριθμοι τόσο για την εύρεση, την κατάταξη και την εμφάνιση των αποτελεσμάτων.

Λέξεις Κλειδιά: Χώρο – κειμενικά δεδομένα, σημεία ενδιαφέροντος, κατάταξη, απόσταση, αναζήτηση, λέξεις – κλειδιά, ανάμοια, ποικιλομορφία.

Abstract

In recent years, more applications have been created based on a user's location, resulting in the display of relevant information based on the location of the user. There are many systems that do this using spatial-textual data and now are evolving more and more to meet the needs of more users. Many of them allow users to search for points of interest based on his preferences and nearest distance from him. These points can be companies that produce products or offer various services, various shops such as restaurants, supermarkets, clothes, technology, etc. In this Thesis we will deal with the ranking of some points of interest and specifically restaurants based on their distance from the user and some keywords that he is interested in searching. The results that will be selected will be the most dissimilar to each other so that we could have diversity in our results, ie many different types of restaurants. For all the above, algorithms were used to find, rank and display the results.

Keywords: *spatial-textual data, points of interest, ranking, distance, search, keywords, dissimilar, diversity.*

1

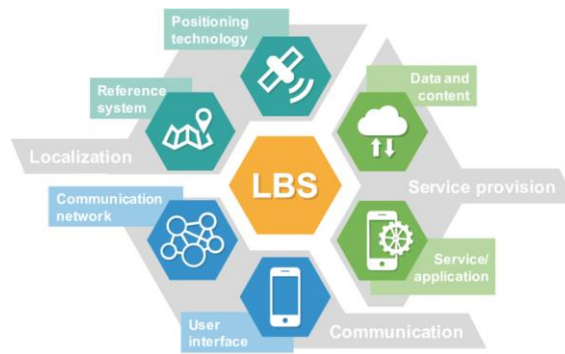
Εισαγωγή

1.1 Ανάκτηση δεδομένων με βάση την τοποθεσία

Πολλές εφαρμογές απαιτούν την εύρεση αντικειμένων – χώρων που βρίσκονται πιο κοντά στην θέση ενός χρήστη ή σε μια καθορισμένη τοποθεσία που περιέχει ένα σύνολο λέξεων-κλειδιών. Ένα παράδειγμα είναι οι γνωστές “yellow pages” όπου επιτρέπουν στους χρήστες να καθορίζουν μια διεύθυνση και ένα σύνολο λέξεων-κλειδιών. Έτσι σαν αποτέλεσμα ο χρήστης λαμβάνει μια λίστα επιχειρήσεων ή αντικειμένων των οποίων η περιγραφή περιέχει αυτές τις λέξεις-κλειδιά τις οποίες έχει χρησιμοποιήσει ο ίδιος για την αναζήτηση, ταξινομημένες με βάση την απόσταση που τον ενδιαφέρει.

Το LBS αναφέρεται στις υπηρεσίες λογισμικού που χρησιμοποιούν γεωγραφικά δεδομένα και πληροφορίες για την παροχή υπηρεσιών ή πληροφοριών στους χρήστες. Έτσι ένα LBS μπορεί να εντοπίσει την θέση ενός χρήστη ο οποίος φυσικά έχει ενεργοποιήσει αυτήν την λειτουργία και να βρει τι υπάρχει γύρω του πολύ εύκολα. Υπάρχουν πολλές εφαρμογές φυσικά κάποιες εκ των οποίων είναι εφαρμογές για τον εντοπισμό φίλων (Find your Friends), τον εντοπισμό πλησιέστερων μαγαζιών, εστιατορίων κ.α., εφαρμογές για παρακολούθηση ασθενών για απόκριση έκτακτης ανάγκης κ.α.¹

¹ Jiang, Bin, and Xiaobai Yao. "Location-based services and GIS in perspective." Computers, Environment and Urban Systems 30.6 (2006): 712-725.



Εικόνα 1. Βασικά συστατικά ενός LBS²

Σε αυτήν την διπλωματική, δεδομένων κάποιων αρχείων καταγραφής ερωτημάτων(query logs), κάποιων ερωτημάτων που έχουν γίνει σε ένα LBS(Location-based Service) και ενός σύνολο σημείων ενδιαφέροντος διάφορων χρηστών(π.χ. εστιατόρια) αντιμετωπίζουμε το πρόβλημα της ανακάλυψης ενός περιορισμένου συνόλου διαφορετικών σημείων ενδιαφέροντος που προσελκύουν τις προτιμήσεις διαφορετικών πελατών οπότε βασική προϋπόθεση είναι να έχουμε μια ποικιλομορφία στα αποτελέσματα των αναζητήσεων μας.

Έχοντας τα αρχεία καταγραφής ερωτήσεων(query logs) γνωρίζουμε τις ερωτήσεις που έχουν κάνει οι χρήστες οι οποίες περιλαμβάνουν τις θέσεις των χρηστών και κάποιες προτιμήσεις με λέξεις-κλειδιά. Έτσι έχουμε για παράδειγμα σαν ερώτημα “τα εστιατόρια σε απόσταση 5km που να έχουν “pizza””. Οπότε θέλουμε να αναλύσουμε την ποιότητα των σημείων ενδιαφέροντας με βάση πόσο αρέσουν στους χρήστες του LBS.

1.2 Βασικές έννοιες

1.2.1 Χωρικά δεδομένα(Spatial Dataset)

Για την λύση του προβλήματος μας στην συγκεκριμένη διπλωματική θα χρησιμοποιήσουμε χωρικά δεδομένα(Spatial data). Τι είναι όμως τα χωρικά δεδομένα και οι χωρικές βάσεις δεδομένων(Spatial database) αλλά και τα χωρικά ερωτήματα(spatial queries) και ποια η διαφορά τους με τα υπόλοιπα που ξέρουμε;

Τα χωρικά δεδομένα είναι πληροφορίες για ένα φυσικό αντικείμενο που μπορεί να αναπαρασταθεί με αριθμητικές τιμές σε ένα σύστημα γεωγραφικών συντεταγμένων. Σε γενικές γραμμές, τα χωρικά δεδομένα αντιπροσωπεύουν τη θέση, το μέγεθος και το σχήμα ενός αντικειμένου στον πλανήτη, όπως ένα κτίριο, λίμνη, βουνό κ.α.. Τα χωρικά δεδομένα μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες: τα διάνυσμα και τα ράστερ(Heywood and Cornelius, 2010). Τα διανυσματικά μοντέλα χωρικών δεδομένων χαρακτηρίζονται με σημεία, γραμμές και πολύγωνα και τις τοπολογίες μεταξύ τους. Για παράδειγμα, ένα ποτάμι μπορεί να θεωρηθεί ως γραμμή, και μια λίμνη μπορεί να

² Falkowski et al., 2016

αντιμετωπιστεί ως πολύγωνο. Τα δεδομένα ράστερ αντιπροσωπεύουν γεωγραφικά φαινόμενα με ένα πλέγμα πολυδιάστατων διακριτών τιμών, όπως εικόνες τηλεπισκόπησης(remote sensing images), σαρωμένους τοπογραφικούς χάρτες κ.α.³ Τα χωρικά δεδομένα μπορεί επίσης να περιλαμβάνουν και πρόσθετα χαρακτηριστικά που παρέχουν περισσότερες πληροφορίες σχετικά με την οντότητα που αντιπροσωπεύεται. Τα Γεωγραφικά Συστήματα Πληροφοριών(GIS)⁴ ή άλλες εξειδικευμένες εφαρμογές λογισμικού μπορούν να χρησιμοποιηθούν για πρόσβαση, οπτικοποίηση, χειρισμό και ανάλυση γεωχωρικών δεδομένων. Οπότε μπορούμε σε μία σειρά να πούμε ότι τα χωρικά δεδομένα είναι τα δεδομένα που σχετίζονται με ή περιέχουν πληροφορίες σχετικά με μια συγκεκριμένη τοποθεσία στην επιφάνεια της Γης. Ένα παράδειγμα χωρικών δεδομένων μπορεί έχοντας έναν χάρτη που να δείχνει κτίρια(π.χ. εστιατόρια) στην περιοχή της πόλης. Κάθε ένα από τα κτίρια αυτά, εκτός από την τοποθεσία του, μπορεί να έχει πρόσθετα χαρακτηριστικά, όπως τον τύπο χρήσης (στέγαση, επιχείρηση, κυβέρνηση κ.λπ.), το έτος κατασκευής κ.α.

Τα χωρικά δεδομένα χρησιμοποιούνται σε πολλούς τομείς κάθε ένα με διαφορετικό τρόπο. Ένας τομέας είναι για χρήση γραφικών π.χ. για την δημιουργία χαρτών. Για παράδειγμα κατά τη δημιουργία οπτικών, γραφικών ή χαρτών με χωρικά δεδομένα, υπάρχει μια ποικιλία γεωγραφικών στοιχείων που πρέπει να λαμβάνεται υπόψη. Ένα από τα πιο σημαντικά προβλήματα είναι η προβολή. Η προβολή ενός χάρτη περιγράφει τον τρόπο με τον οποίο ένα τρισδιάστατο σχήμα παρουσιάζεται σε μια δισδιάστατη επιφάνεια.

Ένας άλλος τομέας χρήσης χωρικών δεδομένων είναι στην στατιστική. Όπως συμβαίνει με οποιαδήποτε άλλα δεδομένα έτσι και τα χωρικά δεδομένα για να κατανοηθούν θα πρέπει γίνει μία στατιστική ανάλυση. Η κύρια διαφορά μεταξύ χωρικών δεδομένων και όλων των άλλων τύπων δεδομένων όσον αφορά τη στατιστική ανάλυση είναι η ανάγκη να ληφθούν υπόψη παράγοντες όπως το υψόμετρο, η απόσταση και η περιοχή στην διαδικασία της ανάλυσης. Για παράδειγμα, απλά δεδομένα μπορούν να βοηθήσουν στην εκτίμηση ή στην πρόβλεψη της τιμής ενός δείγματος και τα χωρικά δεδομένα μπορούν να βοηθήσουν στην εκτίμηση ή την πρόβλεψη της τιμής μιας μεταβλητής σε μια τοποθεσία δείγματος.⁵

1.2.2 Geographic Information System(GIS)

Ο πιο συνηθισμένος τρόπος επεξεργασίας και ανάλυσης χωρικών δεδομένων είναι η χρήση GIS ή σύστημα γεωγραφικών πληροφοριών. Αυτά μπορεί να είναι προγράμματα ή ένας συνδυασμός προγραμμάτων που συνεργάζονται για να βοηθήσουν τους χρήστες να κατανοήσουν τα δεδομένα τους. Ένας χρήστης χρησιμοποιεί συνήθως ένα σύνολο χωρικών δεδομένων και τα συγκρίνει ή τα συνδυάζει μεταξύ τους. Κάθε σύνολο χωρικών δεδομένων μπορεί να αναφέρεται ως “layer”. Το πεδίο και η μελέτη του GIS εκτείνεται πολύ περισσότερο από την ψηφιακή χαρτογράφηση και τη χαρτογραφία. Αποτελείται από μια ποικιλία κατηγοριών, όπως χωρική ανάλυση, τηλεπισκόπηση

³ Peng Yue, Zhenyu Tan, in Comprehensive Geographic Information Systems, 2018

⁴ Malczewski, Jacek. "GIS-based land-use suitability analysis: a critical overview." Progress in planning 62.1 (2004): 3-65.

⁵ www.safe.com

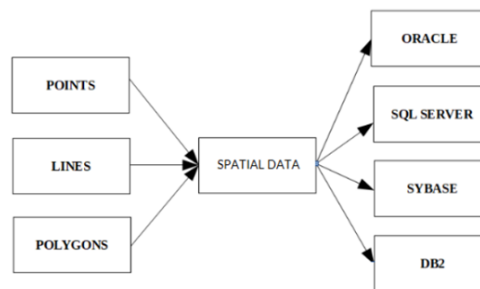
και γεω-οπτικοποίηση. Σε αυτά τα πεδία τα χωρικά δεδομένα γίνονται πολύ πιο περίπλοκα και δύσκολα στη χρήση.⁶

1.2.3 Χωρικές βάσεις δεδομένων(Spatial Databases)

Γνωρίζοντας πλέον για τα χωρικά δεδομένα οι χωρικές βάσεις δεδομένων, όπως υποδηλώνει το όνομα, είναι βάσεις δεδομένων που έχουν βελτιστοποιηθεί για την αποθήκευση και την αναζήτηση αυτών των χωρικών δεδομένων⁷. Συγκεκριμένα οι χωρικές βάσεις δεδομένων μπορούν να περιέχουν μεγάλες ποσότητες δεδομένων σε πολλούς πίνακες με μηχανισμούς σύνδεσης που διατηρούν την ακεραιότητα των δεδομένων αυτών. Είναι σημαντικό να γίνει διάκριση μεταξύ τυπικών βάσεων δεδομένων και χωρικών βάσεων δεδομένων καθώς οι χωρικές βάσεις δεδομένων είναι τυπικές βάσεις δεδομένων που έχουν επεκταθεί για να δέχονται χωρικούς τύπους δεδομένων(spatial datasets) και ερωτήματα(spatial queries). Προηγούμενες μελέτες χωρικών βάσεων δεδομένων επικεντρωνόταν κυρίως στον τρόπο τοποθέτησης διανυσματικών δεδομένων σε βάσεις δεδομένων, ενώ τα δεδομένα ράστερ εξακολουθούν να αποθηκεύονται ως αρχεία.

Τα βασικά χαρακτηριστικά των χωρικών βάσεων δεδομένων είναι τα εξής⁸:

- Είναι συστήματα βάσεων δεδομένων.
- Προσφέρουν διάφορους τύπους χωρικών δεδομένων στο μοντέλο δεδομένων που χρησιμοποιούν (data model) και στη γλώσσα ερωτημάτων (query language).
- Υποστηρίζει τύπους χωρικών δεδομένων στην εφαρμογή του, παρέχοντας χωρικά ευρετήρια και αποτελεσματικούς αλγόριθμους για χωρική ένωση(spatial join).



Εικόνα 2. Τα χωρικά δεδομένα με τη μορφή σημείων, γραμμών, πολυγώνων κ.λ.π. χρησιμοποιούνται από πολλές διαφορετικές βάσεις δεδομένων⁹

⁶ www.safe.com

⁷ Shaun Fontanella, Ningchuan Xiao, in Comprehensive Geographic Information Systems, 2018

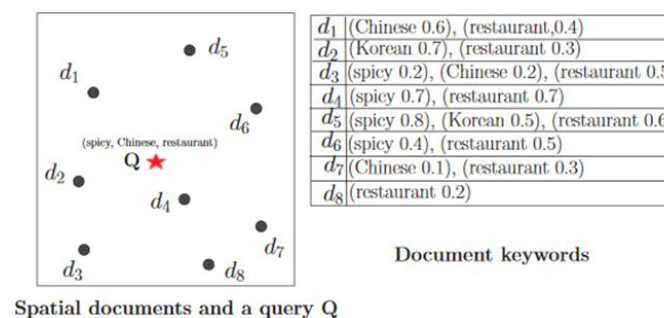
⁸ Güting, Ralf Hartmut. "An introduction to spatial database systems." the VLDB Journal 3.4 (1994): 357-399.

⁹ www.tutorialspoint.com

Με βάση αυτά τα χαρακτηριστικά ο σχεδιασμός και η εφαρμογή μιας χωρικής βάσης δεδομένων πρέπει να πληροί τις ακόλουθες απαιτήσεις:

- Να μπορεί η βάση δεδομένων να χρησιμοποιηθεί για αποθήκευση και διαχείριση δεδομένων.
- Να υποστηρίζει διάφορους τύπους χωρικών δεδομένων στο μοντέλο δεδομένων της.
- Να προσφέρει μια γλώσσα ερωτήματος (Query language) για την εκτέλεση χωρικών ερωτημάτων (spatial queries).
- Να παρέχει χωρικά ευρετήρια (spatial indices) για την επιτάχυνση χωρικών ερωτημάτων.

Συμπερασματικά τα μοντέλα βάσεων δεδομένων, τα χωρικά ερωτήματα και τα ευρετήρια είναι τρία σημαντικά ζητήματα που πρέπει να ληφθούν υπόψη για την παροχή μιας χωρικής βάσης δεδομένων. Χωρικά ερωτήματα που εκτελούνται σε χωρικά αντικείμενα περιλαμβάνουν κυρίως ερωτήματα βάσει τοποθεσίας, ερωτήματα βάσει χωρικής σχέσης και ερωτήματα βάσει χαρακτηριστικών. Τα δύο πρώτα είναι τυπικά χωρικά ερωτήματα.¹⁰



Εικόνα 3. Παράδειγμα μία βάσης χωρικών δεδομένων(Spatial Database)¹¹

1.2.4 Χωρικά Ερωτήματα(Spatial queries)

Τα χωρικά ερωτήματα είναι ερωτήματα σε μια χωρική βάση δεδομένων που μπορούν να απαντηθούν μόνο με βάση τις γεωμετρικές πληροφορίες, δηλαδή τη χωρική θέση και την έκταση των σχετικών αντικειμένων. Ένα χωρικό ερώτημα ορίζεται από ένα χώρο ερωτήματος S , δηλαδή είτε ολόκληρη την χωρική βάση δεδομένων, είτε ένα τμήμα της που λαμβάνεται μέσω κατάλληλων φίλτρων από ένα αντικείμενο ερωτήματος q που μπορεί είτε να ανήκει είτε να μην ανήκει στη βάση δεδομένων και με χωρική σχέση $\mathcal{R}$ ¹². Ένα παράδειγμα ερωτήματος ορίζεται έτσι ως εξής: **return all objects $s \in S$ that are in relation $\mathcal{R}$ with q .**

Εμείς θα ασχοληθούμε με ερωτήματα που βασίζονται σε μια τοποθεσία βρίσκοντας έτσι τα κοντινότερα σημεία στον χρήστη όπως θα δούμε παρακάτω.

¹⁰ Peng Yue, Zhenyu Tan, in Comprehensive Geographic Information Systems, 2018

¹¹ Zhang, Dongxiang, Kian-Lee Tan, and Anthony KH Tung. "Scalable top-k spatial keyword search." Proceedings of the 16th international conference on extending database technology. 2013

¹² Leila de Floriani, ... Enrico Puppo, in Handbook of Computational Geometry, 2000

1.3 Αντικείμενο διπλωματικής

Στην παρούσα διπλωματική θα ασχοληθούμε με χωρο-κειμενικά δεδομένα δηλαδή με δεδομένα τα οποία περιγράφονται από γεωγραφικές συντεταγμένες π.χ. X,Y καθώς και από κάποιες λέξεις κλειδιά το καθένα. Εμείς συγκεκριμένα έχουμε ένα αρχείο με κάποια εστιατόρια και τα χαρακτηριστικά τους όπως το όνομα, μια βαθμολογία αξιολόγησης, την τοποθεσία στο χώρο μας η οποία περιγράφεται από ένα γεωγραφικό μήκος και πλάτος, και ένα σύνολο από λέξεις κλειδιά που το περιγράφουν π.χ. Αμερικάνικο, Ιταλικό κ.α. Επίσης διαθέτουμε και ένα αρχείο ερωτήσεων διαφορετικών χρηστών που περιλαμβάνει την θέση του χρήστη πάλι σε μορφή γεωγραφικού πλάτους και μήκους καθώς επίσης και ένα σύνολο από λέξεις-κλειδιά με βάση την προτίμηση του δηλαδή τι τον ενδιαφέρει να βρει σε μία κοντινή απόσταση. Εμάς μας ενδιαφέρουν τα πιο κοντινά εστιατόρια οπότε θα χρειαστεί ένα top-k query για να βρούμε π.χ. τα top-2 κοντινότερα εστιατόρια με βάση τις προτιμήσεις δηλαδή τα keywords που έχει δώσει.

```
|1|Daphne's Greek Cafe|2.5|37.755557|-122.250360846519|Greek, Mediterranean
2|Laguna Pizza|4.5|38.423436|-121.41361|Pizza
3|Starbucks Coffee|4.5|34.095731|-117.717899|American, Coffee, Tea
4|Starbucks Coffee|5|36.341065|-119.347593|Coffee
5|A G Ferrari Foods|4|37.858281|-122.253116|Italian, Deli, Sandwiches
6|Palms|4|34.397338|-119.520776|Californian, Seafood, American, Chicken, Steak
7|A&W Restaurant|3|36.302863|-119.135775|American, Burgers, Hot Dogs, Chicken, Fast Food, Pub Food, Sandwiches
```

Εικόνα 4. Παράδειγμα μορφής του αρχείου με τα εστιατόρια και τα χαρακτηριστικά τους

```
Q1 = {American} και (33.672452|-118.004725)
Q2 = {Steak} και (35.563482|-119.556997)
Q3 = {Italian, Deli} και (32.78014|-120.230014)
Q4 = {American, Steak, Seafood} και (33.02365|-122.4930112)
Q5 = {American, Mexican} και (33.2554391|-117.9830225)
```

Εικόνα 5. Παράδειγμα μορφής του αρχείου με τα εστιατόρια και τα χαρακτηριστικά τους

Τέλος ο σκοπός μας είναι να έχουμε μια ποικιλομορφία δηλαδή θέλουμε το τελικό αποτέλεσμα να είναι πιο ποικιλόμορφο(diverse). Αυτό σημαίνει ότι θέλουμε να καλύψουμε τις προτιμήσεις από όσους περισσότερους χρήστες, να υπάρχει δηλαδή όσο το δυνατό περισσότερες επιλογές μεταξύ για διαφορετικούς χρήστες οπότε σκοπός μας είναι από τα αποτελέσματα των προηγούμενων μέσω κάποιων συναρτήσεων να βρούμε τα m εστιατόρια με την ελάχιστη ομοιότητα.

1.4 Δομή της διπλωματικής

Η διπλωματική μας αποτελείται από 6 κεφάλαια. Στο κεφάλαιο 2 γίνεται μια ανάλυση του προβλήματος που μελετάμε καθώς και ένα παράδειγμα για την κατανόηση του. Στο Κεφάλαιο 3 παρουσιάζονται άλλες προσεγγίσεις που μοιάζουν με την δική μας που έχουν γίνει στο χώρο και πως βοήθησαν. Στο Κεφάλαιο 4 αναλύουμε τον τρόπο που χρησιμοποιήσαμε για να πετύχουμε το επιθυμητό αποτέλεσμα, με διάφορα τεχνικά χαρακτηριστικά που χρησιμοποιήσαμε. Στο Κεφάλαιο 5 αναλύουμε τυχόν προβλήματα που προέκυψαν, την αξιοπιστία των αποτελεσμάτων μας, την

σημασία τους κ.α. Τέλος στο Κεφάλαιο 6 αναφέρουμε τα συμπεράσματα, τον βαθμό ικανοποίησης που είχαμε και τυχόν μελλοντικές απαιτήσεις, προτάσεις ή βελτιώσεις.

2

Ορισμός του Προβλήματος

2.1 Ερωτήσεις κοντινότερου γείτονα με λέξεις κλειδιά

Σε αυτήν την διπλωματική θα χρησιμοποιήσουμε το ερώτημα κατάταξης με λέξεις κλειδιά (top-k keyword search), το οποίο ψάχνει μέσα στον πολυδιάστατο χώρο να βρει τα k κοντινότερα αντικείμενα στο καθορισμένο σημείο p . Τα χωρικά δεδομένα ταξινομούνται με βάση την απόσταση και έτσι ένα αντικείμενο που είναι πιο κοντά στο σημείο p θα έχει υψηλότερη κατάταξη. Το p εδώ θα είναι το σημείο που βρίσκεται κάθε χρήστης ο οποίος θα περιέχει κάποιες λέξεις κλειδιά με τα οποία θα γίνεται και η αναζήτηση. Συγκεκριμένα $score(T) = distance(Tp, p)$.

Η ερώτηση λαμβάνει υπόψη και ένα σύνολο λέξεων κλειδιών και εφαρμόζουμε boolean keyword search που επιστρέφει το σύνολο όλων των αντικειμένων των οποίων η περιγραφή τους περιέχει όλα τα $w_1, \dots, w_m$.

Για παράδειγμα ας πούμε ότι έχουμε 25 τυχαίες πλειάδες οι οποίες επιλέχθηκαν από ένα μεγάλο σύνολο δεδομένων και θέλουμε με βάση την θέση του χρήστη και τις λέξεις κλειδιά που χρησιμοποίησε να βρούμε τα 2 πιο κοντινά σημεία που εξυπηρετούν τις προτιμήσεις του.

1|Daphne's Greek Cafe|2.5|37.755557|-122.250360846519|Greek, Mediterranean
 2|Laguna Pizza|4.5|38.423436|-121.41361|Pizza
 3|Starbucks Coffee|4.5|34.095731|-117.717899|American, Coffee, Tea
 4|Starbucks Coffee|5|36.341065|-119.347593|Coffee
 5|A G Ferrari Foods|4|37.858281|-122.253116|Italian, Deli, Sandwiches
 6|Palms|4|34.397338|-119.520776|Californian, Seafood, American, Chicken, Steak
 7|A&W Restaurant|3|36.302863|-119.135775|American, Burgers, Hot Dogs, Chicken, Fast Food, Pub Food, Sandwiches
 8|A&W Restaurant|4.5|34.272035|-118.50015|American, Barbecue, Dim Sum, Chinese, Seafood
 9|Basque Hotel|1.5|36.73146|-119.796036|American, Mexican
 10|El Cholo|4|33.938846|-117.937280033333|Mexican
 11|The Original Pantry Cafe|3.5|34.04633|-118.262558|American, Steak, Coffee, Cafe, Diner
 12|Whiskey Creek At Bishop|4|37.364352|-118.395234|American
 13|Alioto's Restaurant|3|37.80916|-122.41646|Seafood, Italian, American, European, Southern
 14|Genova Delicatessen|4.5|37.8966828417367|-122.064129710197|Deli, Italian
 15|Hong Kong Restaurant|4|37.332027|-120.469065|Burgers, Chinese
 16|Shepherd's Inn|4.5|36.738|-119.78269|Seafood, American, Mexican, Steak
 17|Mexicali Rose Restaurant|3.5|37.80045|-122.27659|Mexican, Fast Food
 18|Ole's Waffle Shop|4|37.765562|-122.24176|American, Diner, Traditional
 19|Santa Fe Basque|4|36.778652|-119.79952|American, Southwestern
 20|La Golondrina Mexican Cafe|4|34.05709|-118.23834|Mexican, American, Deli
 21|The Brighton Coffeeshop|4|34.0686|-118.403999|Coffee, American, Diner, Californian, Deli, Tea
 22|Village Inn|3.5|33.605509|-117.890027|Seafood, American, Steak
 23|Zals Museum Cafe|5|34.423128|-119.704073|Cafe, Coffee, French, Sandwiches
 24|Club Pheasant|3.5|38.551988|-121.53776|Italian, American
 25|Dairy Queen|3|33.558037|-117.724289|Burgers, Fast Food, Ice Cream, American

Εικόνα 6. Παράδειγμα αρχείου εστιατορίων

Στην Εικόνα 6 έχουμε και τα 25 εστιατόρια που το καθένα περιέχει:

- ένα αναγνωριστικό αριθμό(π.χ. 1, 2, 3 κτλ.)
- το όνομα του
- μια βαθμολογία
- το γεωγραφικό μήκος του(longitude)
- το γεωγραφικό πλάτος του(latitude)
- μια σειρά από λέξεις κλειδιά που χαρακτηρίζουν το εστιατόριο στο είδος του

Έστω ότι έχουμε έναν χρήστη ο οποίος βρίσκεται σε μία περιοχή με το αντίστοιχο γεωγραφικό πλάτος και μήκος και ψάχνει τα 2 πιο κοντινά εστιατόρια που έχουν αμερικάνικη κουζίνα. Οπότε το ερώτημα θα έχει την εξής μορφή:

Q1 = {American} και (33.672452 | -118.004725)

Ξεκινώντας θα πρέπει να μέσω του πλάτους και του μήκους να βρούμε την θέση του χρήστη και του κάθε εστιατορίου έτσι ώστε να μπορούμε να βρούμε την απόσταση μεταξύ τους. Αυτό για να γίνει θα πρέπει να χρησιμοποιήσουμε τον τύπο για τον υπολογισμό της απόστασης μεταξύ 2 σημείων.

Ο τύπος για την απόσταση δύο σημείων είναι όπως φαίνεται παρακάτω:

$$\sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

Εικόνα 7. Εξίσωση εύρεσης απόστασης μεταξύ δύο σημείων

Στον τύπο αυτόν οι μεταβλητές x_A και y_A αναπαριστούν συντεταγμένες του ενός σημείου και οι μεταβλητές x_B και y_B τις συντεταγμένες του άλλου.

Name	Q1 Distance
Village Inn	0.1328
Dairy Queen	0.3029

Πίνακας 1. Αποτελέσματα πρώτης ερώτησης

Έτσι στον παραπάνω πίνακα έχουμε τα 2 πιο κοντινά σημεία στον χρήστη με βάση την θέση του στον χάρτη. Το ίδιο μπορεί να γίνει φυσικά και με άλλα ερωτήματα με διαφορετική θέση και keywords.

Μπορούμε να χρησιμοποιήσουμε άλλο keyword και τοποθεσία χρήστη για να δούμε διαφορετικά αποτελέσματα. Για παράδειγμα αν αλλάξουμε τις συντεταγμένες και για λέξη χρησιμοποιήσουμε το Steak το ερώτημα μας διαμορφώνεται έτσι:

Q2 = {Steak} και (35.563482 | -119.556997)

Έτσι με βάση τα παραπάνω τα top 2 που προκύπτουν από την αναζήτηση με βάση τα keyword και την τοποθεσία του χρήστη είναι:

Name	Q2 Distance
Palms	1.1667
Shepherd's Inn	1.196

Πίνακας 2. Αποτελέσματα δεύτερης ερώτησης

Και τα 2 εστιατόρια περιέχουν το συγκεκριμένο φαγητό καθώς ο χρήστης αυτό ήθελε αλλά το πιο κοντινό σημείο με βάση την θέση του είναι το Palms σε απόσταση 0,1667.

Φυσικά μπορεί ένας χρήστης να θέλει 2 είδη φαγητού καθώς μπορεί αν είναι παραπάνω από ένας στο τραπέζι και να υπάρχουν διαφορετικές προτιμήσεις. Σε αυτήν την περίπτωση χρησιμοποιούμε 2 keywords στην ίδια ερώτηση και θέλουμε στα αποτελέσματα να περιέχουν και τα 2 είδη και όχι έστω ένα από αυτά.

Έτσι χρησιμοποιώντας διαφορετικές συντεταγμένες αλλά αυτήν την φορά 2 keywords η ερώτηση μας διαμορφώνεται ως εξής:

Q3 = {Italian, Deli} και (32.78014 | -120.230014)

Έτσι ο χρήστης που κάνει την αναζήτηση ψάχνει για εστιατόρια που έχουν και ιταλική κουζίνα και deli μαζί.

Οπότε με βάση τα παραπάνω τα τοπ 2 που προκύπτουν από την αναζήτηση με βάση τα keyword και την τοποθεσία του χρήστη είναι:

Name	Q3 Distance
Genova Delicatessen	5.4353
A G Ferrari Foods	5.4663

Πίνακας 3. Αποτελέσματα τρίτης ερώτησης

Και τα 2 εστιατόρια περιέχουν ιταλικό και deli καθώς ο χρήστης αυτό ήθελε αλλά το πιο κοντινό σημείο με βάση την θέση του είναι το Genova Delicatessen σε απόσταση 5,4353. Υπήρχαν και κάποια σημεία τα οποία είχα ιταλικό και deli αλλά περιείχε το ένα από τα 2 και οπότε δεν μας κάνουν.

Δοκιμάζοντας τώρα 3 keywords και διαφορετικές συντεταγμένες έχουμε το ερώτημα:

Q4 = {American, Steak, Seafood} και (33.02365 | -122.4930112)

Φυσικά όπως και στο προηγούμενο ψάχνουμε εστιατόρια που περιέχουν και τις 3 λέξεις και όχι έστω μία από αυτές.

Τα αποτελέσματα είναι τα εξής:

Name	Q4 Distance
Palms	3.2743
Shepherd's Inn	4.5981

Πίνακας 4. Αποτελέσματα τέταρτης ερώτησης

Και τα 2 εστιατόρια περιέχουν και τα 3 που επέλεξε ο χρήστης αλλά το πιο κοντινό σημείο με βάση την θέση του είναι το Palms σε απόσταση 3.2743. Και εδώ υπήρχαν άλλα εστιατόρια που να περιείχαν κάποια από αυτά αλλά όχι όλα οπότε και δεν τα επέστρεψε.

Και τελευταίο παράδειγμα δοκιμάζουμε άλλες 2 λέξεις οπότε έχουμε το ερώτημα:

Q5 = {American, Mexican} και (33.2554391 | -117.9830225)

Σε αυτό θέλουμε εστιατόρια και με αμερικάνικο και με μεξικάνικο φαγητό μαζί.

Με βάση τα παραπάνω τα τοπ 2 που προκύπτουν από την αναζήτηση με βάση τα keyword και την τοποθεσία του χρήστη είναι:

Name	Q5 Distance
La Golondrina Mexican Cafe	0.8413
Shepherd's Inn	3.9201

Πίνακας 5. Αποτελέσματα πέμπτης ερώτησης

Και τα 2 εστιατόρια περιέχουν και τα 2 που επέλεξε ο χρήστης αλλά το πιο κοντινό σημείο με βάση την θέση του είναι το La Golondrina Mexican Cafe σε απόσταση 0.8413. Και εδώ υπήρχαν άλλα εστιατόρια που να περιείχαν κάποια από αυτά αλλά όχι όλα οπότε και δεν τα επέστρεψε.

Οπότε έχουμε συνοπτικά όλα τα εστιατόρια και την απόσταση του καθενός από τον χρήστη για κάθε ερώτηση. Έχουμε συνολικά 7 στήλες, μία με το όνομα των εστιατορίων, τις 5 παραπάνω ερωτήσεις που χρησιμοποιήσαμε και μία για τα keywords του κάθε εστιατορίου ξεχωριστά.

Name	Q1 Distance	Q2 Distance	Q3 Distance	Q4 Distance	Q5 Distance	Keywords
Daphne's Greek Cafe						Greek, Mediterranean
Laguna Pizza						Pizza
Starbucks Coffee	0,5113					American, Coffee, Tea
Starbucks Coffee						Coffee
A G Ferrari Foods			5,4663			Italian, Deli, Sandwiches
Palms	1,6804	1,1667		3,2743		Californian, Seafood, American, Chicken, Steak
A&W Restaurant	2,8633					American, Burgers, Hot Dogs, Chicken, Fast Food, Pub Food, Sandwiches
A&W Restaurant	0,7778					American, Barbecue, Dim Sum, Chinese, Seafood
Basque Hotel	3,5449				3,9204	American, Mexican
El Cholo						Mexican
The Original Pantry Cafe	0,4542	1,9943				American, Steak, Coffee, Cafe, Diner
Whiskey Creek At Bishop	3,7125					American
Alioto's Restaurant	6,0478					Seafood, Italian, American, European, Southern
Genova Delicatessen			5,4353			Deli, Italian
Hong Kong Restaurant						Burgers, Chinese
Shepherd's Inn	3,5438	1,196		4,5981	3,9201	Seafood, American, Mexican, Steak
Mexicali Rose Restaurant						Mexican, Fast Food
Ole's Waffle Shop	5,8912					American, Diner, Traditional
Santa Fe Basque	3,5874					American, Southwestern
La Golondrina Mexican Cafe	0,45				0,8413	Mexican, American, Deli
The Brighton Coffeshop	0,5625					Coffee, American, Diner, Californian, Deli, Tea
Village Inn	0,1328	2,5715		4,6396		Seafood, American, Steak
Zals Museum Cafe						Cafe, Coffee, French, Sandwiches
Club Pheasant	6,0243		5,9181			Italian, American, Deli
Dairy Queen	0,3029					Burgers, Fast Food, Ice Cream, American

Εικόνα 8. Λεπτομερές πίνακας εστιατορίων και αποτελέσματα ερωτήσεων για το καθένα

Το παράδειγμα αυτό είναι για ένα ερώτημα που έγινε από ένα χρήστη αλλά εμείς θέλουμε να δοκιμάσουμε πολλά ερωτήματα από διαφορετικούς χρήστες έτσι ώστε στο τέλος να έχουμε ένα σύνολο από m εστιατόρια τα οποία καλύπτουν τις προτιμήσεις από όσο περισσότερους χρήστες, αρά είναι όσο πιο ποικιλόμορφο(diverse) γίνεται. Για να βρούμε αυτό το σύνολο από εστιατόρια έχουμε την ερώτηση που είναι ένα σημείο στο χώρο, μια ακτίνα και ένα set από keywords. Θεωρούμε ότι το k (πόσα γυρνάει μια ερώτηση) είναι δεδομένο και σταθερό για όλες τις ερωτήσεις. Άρα έχουμε ένα dataset τα δεδομένα μας (π.χ. εστιατόρια) και ένα set από ερωτήσεις.

2.2 Συνάρτηση ομοιότητας και ποικιλομορφία

Για να μιλήσουμε καλύτερα για ποικιλομορφία όμως θα πρέπει να βρούμε τα πιο ανόμοια από αυτά τα σημεία που έχουμε, οπότε πρέπει πρώτα να ορίσουμε τι είναι η ομοιότητα.

Για να βρούμε πόσο όμοια είναι 2 σημεία μεταξύ τους σε σχέση πάντα με τις ερωτήσεις και προτιμήσεις των χρηστών, όπως για παράδειγμα το Palms με το Village Inn που χρησιμοποιήθηκαν πιο πάνω θα πρέπει να κοιτάξουμε την τοποθεσία και τις λέξεις κλειδιά. Αν είναι πολύ κοντά στο χώρο πιθανόν να ενδιαφέρουν τους ίδιους χρήστες ή χαρακτηρίζονται από τις ίδιες λέξεις κλειδιά.

Για να το δούμε αυτό στην πράξη θα χρησιμοποιήσουμε μια συνάρτηση ομοιότητας που συνδυάζει αυτά τα 2 σημεία(p_1, p_2) που θέλουμε. Η συνάρτηση που θα χρησιμοποιήσουμε είναι η εξής:

$$\text{Sim}(p_1, p_2) = 0.5 * (\text{maxdist} - \text{dist}(p_1.q, p_2.q)) / \text{maxdist} + 0.5 * \text{Jaccard}(p_1.qw, p_2.qw)$$

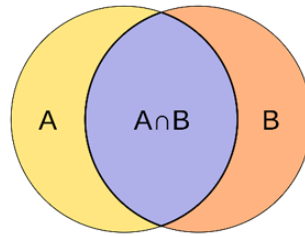
Αυτή η συνάρτηση που ονομάσαμε sim θα υπολογίζει την ομοιότητα 2 σημείων που μας ενδιαφέρουν και αποτελείται από 3 μεταβλητές:

- 1) Η πρώτη μεταβλητή είναι το maxdist που είναι η διαγώνιος στο χώρο ή η μέγιστη απόσταση και υπολογίζεται βρίσκοντας την ελάχιστη τιμή σε latitude και longitude δημιουργώντας το ένα σημείο, αλλά και την μέγιστη αυτών δημιουργώντας ένα άλλο. Οπότε η απόσταση μεταξύ αυτών των δυο σημείων με τις ελάχιστες και μέγιστες τιμές σε latitude και longitude είναι η διαγώνιος στο χώρο ή η μέγιστη απόσταση.
- 2) Η δεύτερη μεταβλητή είναι το dist και είναι η απόσταση μεταξύ των δύο σημείων που εξετάζουμε. Είναι δηλαδή μία συνάρτηση απόστασης μεταξύ των 2 σημείων, για εμάς είναι η ευκλείδεια όπως χρησιμοποιήθηκε πιο πάνω.
- 3) Η τελευταία μεταβλητή είναι η Jaccard και είναι μία ακόμα συνάρτηση για τον υπολογισμό της ομοιότητας και της ποικιλομορφίας των συνόλων δειγμάτων. Πιο συγκεκριμένα ο τύπος είναι:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

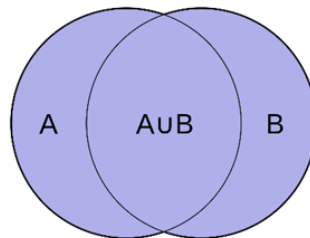
Εικόνα 9. Τύπος εξίσωσης Jaccard

Αυτή η συνάρτηση επιστρέφει τιμές στο διάστημα από 0 έως 1, όπου 0 σημαίνει ότι δεν υπάρχει καμία ομοιότητα στις λέξεις κλειδιά ενώ το 1 ότι είναι όμοια στο 100% στις λέξεις κλειδιά.



Εικόνα 10. Σχήμα τομής δύο σημείων

Παραπάνω βλέπουμε ένα σχήμα για την τομή μεταξύ του A και του B ή αλλιώς του πρώτου σημείου και του δεύτερου στο παράδειγμα μας. Η τομή θα είναι τα κοινά keywords μεταξύ των δύο σημείων από το σύνολο.



Εικόνα 11. Σχήμα ένωσης μεταξύ δύο σημείων

Ενώ εδώ βλέπουμε το σχήμα που δείχνει την ένωση μεταξύ του A και του B ή αλλιώς των δύο σημείων δηλαδή όλα τα keywords που εμφανίζονται για το A και όλα τα keywords που εμφανίζονται για το B.¹³

Επίσης θα πρέπει να αναφερθεί ότι το πρώτο κομμάτι της συνάρτησης δηλαδή το **maxdist -dist (p1.q, p2.q)/maxdist** δίνει και αυτό τιμές από το 0 έως το 1, οπότε όταν έχουμε 0 σημαίνει ότι τα δύο σημεία είναι πολύ μακριά μεταξύ τους ενώ όσο πιο κοντά στο 1 τόσο πιο κοντά είναι μεταξύ τους.

Όσον αφορά το maxdist που είναι η διαγώνιος του χώρου μας υπολογίζεται βρίσκοντας πρώτα το την ελάχιστη απόσταση τιμή για το latitude και την ελάχιστη τιμή για το longitude διαμορφώνεται το ένα σημείο.

Υπολογίζοντας στην συνέχεια την μέγιστη τιμή για το latitude και την μέγιστη για το longitude διαμορφώνεται το άλλο σημείο. Η μεταξύ τους απόσταση θα είναι και η διαγώνιος μας, το maxdist.

¹³ https://en.wikipedia.org/wiki/Jaccard_index

Οπότε έχουμε:

- Min long = 33.558037
- Min lat = -122.41646
- Max long = 38.551988
- Max lat = -117.717899

Οπότε αφού έτσι διαμορφώνονται τα σημεία η απόσταση μεταξύ τους και άρα η διαγώνιος μας ή το maxdist θα είναι **6.8568**.

Άρα συνολικά η συνάρτηση sim θα μας δώσει τιμές από το 0 έως το 1 όπου αν είναι 0 σημαίνει ότι τα δύο σημεία είναι ανόμοια μεταξύ τους ενώ 1 ότι είναι όμοια μεταξύ τους.

Για να καταλάβουμε καλύτερα την ομοιότητα μεταξύ δύο σημείων θα κάνουμε κάποια παραδείγματα.

2.3 Παράδειγμα συνάρτησης ομοιότητας

Έστω για παράδειγμα ότι παίρνουμε 2 τυχαία εστιατόρια από την λίστα μας και θέλουμε να βρούμε την ομοιότητα μεταξύ τους δηλαδή πόσο μοιάζουν μεταξύ τους αυτά τα 2 εστιατόρια. Έστω ότι επιλέγουμε το **La Golondrina Mexican Cafe** και το **Palms**.

Το La Golondrina Mexican Café έχει συντεταγμένες: 34.05709 και -118.23834 και keywords: Mexican, American, Deli.

Το Palms έχει συντεταγμένες: 34.397338 και -119.520776 και keywords: Californian, Seafood, American, Chicken, Steak.

❖ Jaccard Similarity

Από αυτά που έχουμε μέχρι στιγμής μπορούμε με το μάτι να βγάλουμε κάποια συμπεράσματα για να βρούμε το Jaccard similarity. Όπως εξηγήσαμε πιο πάνω για να βρούμε το Jaccard χρειαζόμαστε την τομή και την ένωση των keywords των 2 εστιατορίων που συγκρίνουμε τα οποία εμφανίστηκαν στις ερωτήσεις που έχουν γίνει. Για παράδειγμα αν το ένα εμφανίστηκε σε μία ερώτηση και το άλλο σε 2 η ένωση θα είναι τα keywords της ερώτησης στην οποία εμφανίστηκε το ένα και τα keywords των ερωτήσεων που εμφανίστηκε το άλλο.

Όπως κάναμε πιο πάνω το La Golondrina Mexican Café εμφανίστηκε στην 5η ερώτηση που κάναμε με keywords τα {American, Mexican}.

Το Palms εμφανίστηκε στην 2η και 4η ερώτηση με keywords { Steak } και { American, Steak, Seafood } αντίστοιχα.

Οπότε η ένωση αυτών των 2 θα είναι:

$$U = \{\text{American, Mexican, Steak, Seafood}\}.$$

Για την τομή ξέρουμε ότι χρειαζόμαστε τα κοινά keywords αυτών των 2. Στην συγκεκριμένη περίπτωση είναι μόνο το American.

$$T = \{\text{American}\}.$$

Έτσι στον αριθμητή θα έχουμε το {American} και στον παρανομαστή το {American, Mexican, Steak, Seafood} ή αλλιώς 1 / 4.

Έτσι βρίσκουμε ότι το Jaccard Similarity είναι 0,25.

❖ Ευκλείδεια απόσταση

Στην συνέχεια θα πρέπει να βρούμε την ευκλείδεια απόσταση μεταξύ αυτών των 2 εστιατορίων που επιλέξαμε για το παράδειγμα. Όπως και με την Jaccard έτσι και εδώ θα πρέπει να δούμε σε ποια ή σε ποιες ερωτήσεις εμφανίζεται το κάθε ένα εστιατόριο από αυτά τα 2. Πιο πριν είδαμε ότι το La Golondrina Mexican Café εμφανίζεται στην 5η και το Palms στην 2η και 4η. Ξέροντας αυτό μπορούμε να βρούμε τις τελικές συντεταγμένες για να βρούμε την μεταξύ τους απόσταση.

Έχουμε για το πρώτο ότι εμφανίζεται στην 5η ερώτηση με συντεταγμένες: 33.2554391 και 117.9830225.

Για το δεύτερο είναι λίγο διαφορετική η μέτρηση καθώς εμφανίζεται σε πάνω από μία ερωτήσεις οπότε θα πρέπει να προσθέσουμε τις συντεταγμένες της κάθε ερώτησης για το longitude και το latitude αντίστοιχα.

Οι ερωτήσεις μας είναι η 2η και η 4η με την 2η να έχει συντεταγμένες: 35.563482 και -119.556997 και την 4η: 33.02365 και -122.4930112 αντίστοιχα. Έτσι προσθέτουμε όλα τα longitude μεταξύ τους και τα διαιρούμε για να βρούμε τον μέσο όρο σε longitude οπότε στον παρανομαστή θα έχουμε τον αριθμό των ερωτήσεων στις οποίες εμφανίζεται το συγκεκριμένο

εστιατόριο το οποίο στην περίπτωση μας είναι 2 γιατί εμφανίζεται σε 2 ερωτήσεις(2η,4η). Το ίδιο κάνουμε και για το latitude.

Οπότε τώρα έχουμε 2 καινούργιες συντεταγμένες τις οποίες θα χρησιμοποιήσουμε για να βρούμε την απόσταση μεταξύ των 2 εστιατορίων. Συνολικά θα έχουμε για να υπολογίσουμε { 33.2554391, -117.9830225, 34.293566, -121.02500409999999 }. Με τον ίδιο τρόπο που υπολογίζουμε την απόσταση μεταξύ 2 σημείων κάνουμε και εδώ με τον τύπο της ευκλείδειας που ξέρουμε.

$$\sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

Εικόνα 12. Εξίσωση ευκλείδειας απόστασης δύο σημείων

Κάνοντας τους υπολογισμούς καταλήγουμε στο αποτέλεσμα που είναι: **3.2142432258996965**.

❖ Maxdist

Το maxdist το ξέρουμε ήδη γιατί το υπολογίσαμε στην αρχή πριν το παράδειγμα και παραμένει ίδιο για όλη την διάρκεια εκτός και άμα είχαμε κάποιες καινούργιες συντεταγμένες από κάποιο καινούργιο εστιατόριο στην περιοχή και θα έπρεπε έτσι να ξανακάνουμε τον υπολογισμό της. Οπότε το maxdist παραμένει 6.8568.

Με βάση τον τύπο μας έχουμε ότι:

- $\text{Maxdist} = 6.8568$
- $\text{dist}(p1.q, p2.q) = 3.2142432258996965$
- $\text{Jaccard}(p1.qw, p2.qw) = 0,25$

Κάνοντας τους υπολογισμούς προκύπτει το αποτέλεσμα:

$\text{Sim}(p1, p2) = 0.265616378930427$

Ο αριθμός αυτός μας λέει την ομοιότητα μεταξύ των 2 εστιατορίων η οποία είναι κοντά στο 0.3 το οποίο είναι μακριά από το 1 που είναι το μέγιστο και που είναι η τέλεια ομοιότητα, οπότε θα λέγαμε ότι αυτά τα 2 εστιατόρια είναι περισσότερο ανόμοια μεταξύ τους.

2.4 Εξαιρέσεις

Περίπτωση 1η

Υπάρχει φυσικά και η περίπτωση δύο εστιατόρια να έχουνε Jaccard similarity ίση με 0 και αυτό γίνεται φυσικά έχοντας 0 στον αριθμητή δηλαδή η τομή τους να είναι 0, να μην έχουν δηλαδή ούτε ένα κοινό keyword. Σε αυτήν την περίπτωση θα έχουμε απλά $\text{Jaccard} = 0$ και λογικά και η τελική τιμή ομοιότητας τους θα είναι πολύ χαμηλή.

Περίπτωση 2η

Μία άλλη περίπτωση εξαίρεσης είναι να μην συγκρίνουμε ποτέ 2 εστιατόρια που βρίσκονται στις ίδιες ερωτήσεις δηλαδή άμα το ένα εστιατόριο έχει εμφανιστεί στις ερωτήσεις 2 και 4 και ένα άλλο έχει εμφανιστεί και αυτό στις ίδιες και σε καμία άλλη τότε δεν μπορούμε να τα συγκρίνουμε καθώς το αποτέλεσμα της συνάρτησης ομοιότητας sim θα είναι ίσο με 1 το οποίο δεν θέλουμε.

3

Παρόμοιες Εργασίες

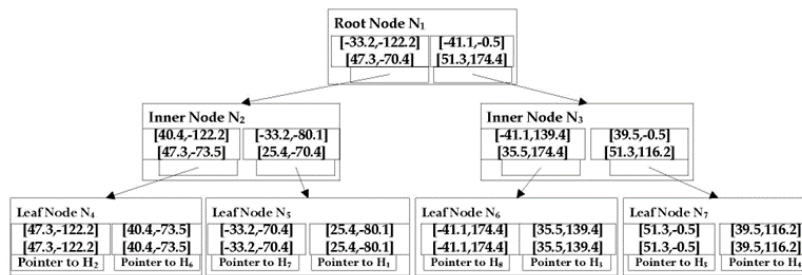
3.1 Keyword Search on Spatial Databases

Έχουν μελετηθεί σε πολλές δημοσιεύσεις διάφορα χωρικά ερωτήματα που χρησιμοποιούν διάφορους τρόπους όπως τον R-tree ¹⁴ και τον R*-tree ¹⁵. Πολλά κείμενα με βάση αυτούς του τρόπους χρησιμοποιούν καινούργια ευρετήρια και μεθόδους για αναζήτηση χωρικών keywords σε χωρικές βάσεις δεδομένων. Κάθε απάντηση στα ερωτήματα είναι συνήθως μία λίστα αντικειμένων που ταξινομούνται με βάση την απόσταση τους από την περιοχή του ερωτήματος. Μια μέθοδος που χρησιμοποιήθηκε είναι η Information Retrieval R-Tree ¹⁶ το οποίο βασίζεται στο R-tree αλλά χρησιμοποιήθηκαν και αλγόριθμοι όπως ο MIR2-Tree. Σε διάφορα πειράματα που έγιναν στην συνέχεια επικεντρώθηκαν στην απόσταση μεταξύ τους περισσότερο καθώς τα αποτελέσματα ήταν πιο εύκολα για ανάλυση. Τα αποτελέσματα ήταν ότι σε χρόνο ο αλγόριθμος IR2-Tree ήταν καλύτερος από τον R-Tree αλλά όχι από τον MIR2-Tree. Στα συμπεράσματα προέκυψε ότι υπάρχει μια πιο ταχύτερη προσέγγιση η οποία βασίζεται στις τεχνικές των R-Tree και των αρχείων υπογραφών.

¹⁴ A. Guttman. R-trees: A dynamic index structure for spatial searching. Proc. SIGMOD, pages 47–57, 1984

¹⁵ N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. Proc. SIGMOD, pages 322–331, 1990

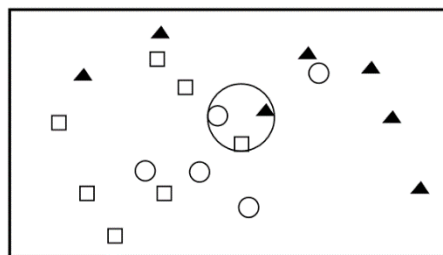
¹⁶ De Felipe, Ian, Vagelis Hristidis, and Naphtali Rishe. "Keyword search on spatial databases." 2008 IEEE 24th International Conference on Data Engineering. IEEE, 2008



Εικόνα 13. Παράδειγμα R-Tree για ένα σύνολο δεδομένων¹⁷

3.2 Keyword Search in Spatial Databases: Towards Searching by Document

Σε άλλες παρόμοιες δημοσιεύσεις εξετάστηκε ένα νέο ευρετήριο με όνομα bR * -Tree¹⁸ το οποίο αποτελεί επέκταση του R * -Tree όπου κατάφεραν να μειώσουν τον χώρο αναζήτησης. Μαζί εξετάστηκε και ένα νέο ερώτημα spatial keyword με όνομα mCK, το οποίο δεδομένου μιας βάσης δεδομένων όπου κάθε πλειάδα περιέχει ένα κείμενο με μορφή keyword στοχεύει στην εύρεση των πλησιέστερων πλειάδων που ταιριάζουν με τα keywords που καθορίζονται από το χρήστη. Έτσι δεδομένου ότι μία πλειάδα μπορεί να συσχετιστεί με ένα ή περισσότερα keywords, ο αριθμός των πλειάδων απόκρισης για το ερώτημα mCK είναι το πολύ m . Για τα πειράματα που έγιναν χρησιμοποιήθηκαν 4 παράμετροι όπως ο αριθμός των keywords στην βάση δεδομένων, το μέγεθος των δεδομένων, την διάσταση και τον αριθμό των keywords που σχετίζονται με τα δεδομένα. Το αποτέλεσμα ήταν ότι τόσο σε συνθετικά όσο και σε πραγματικά σύνολα δεδομένων αποδεικνύει ότι το bR * -tree απαντά αποτελεσματικά στα ερωτήματα mCK που έγιναν σε σχετικά σύντομο χρόνο απόκρισης.



Εικόνα 14. Ερώτημα mCK για 3 keywords¹⁹

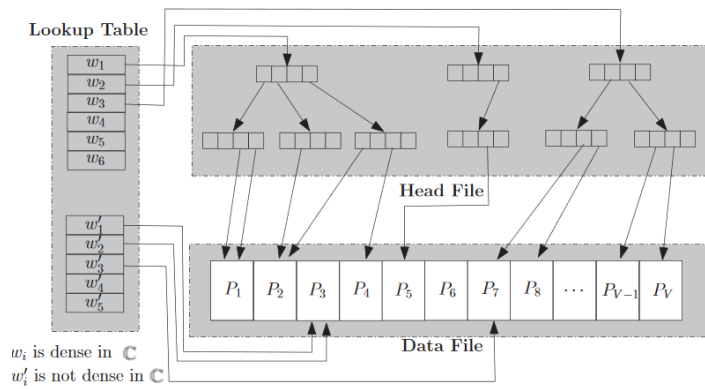
¹⁷ De Felipe, Ian, Vagelis Hristidis, and Naphtali Rishe. "Keyword search on spatial databases." 2008 IEEE 24th International Conference on Data Engineering. IEEE, 2008.

¹⁸ Zhang, Dongxiang, et al. "Keyword search in spatial databases: Towards searching by document." 2009 IEEE 25th International Conference on Data Engineering. IEEE, 2009

¹⁹ Zhang, Dongxiang, et al. "Keyword search in spatial databases: Towards searching by document." 2009 IEEE 25th International Conference on Data Engineering. IEEE, 2009

3.3 Scalable Top-K Spatial Keyword Search

Σε άλλη μελέτη γίνεται αναφορά για έναν επεκτάσιμο δείκτη(index) που ονομάζεται I3, ο οποίος υιοθετεί τη δομή Quadtree²⁰ που βοηθά στην αποσύνθεση χώρου για ιεραρχική κατανομή του χώρου δεδομένων σε κελιά²¹. Το βασικό τμήμα του I3 είναι το keyword cell, το οποίο καταγράφει τη χωρική τοποθεσία ενός keyword όπως επίσης εξετάζεται ένας νέος μηχανισμός αποθήκευσης για την αποτελεσματική ανάκτηση του. Μαζί με αυτά μελετάται τόσο η σημασιολογία AND όσο και η σημασιολογία OR για spatial keyword search. Χρησιμοποιείται ένα χαρακτηριστικό κείμενο(textual attribute) το οποίο είναι μια λίστα από keywords, καθεμία με ένα βάρος που μπορεί να προσαρμοστεί σε διαφορετικές εφαρμογές. Σε πειράματα που πραγματοποιήθηκαν σε πραγματικά χωρικά σύνολα δεδομένων (Twitter και Wikipedia) αποδεικνύουν την υπεροχή του I3 έναντι των υπάρχοντων σχεδίων όπως το IR-tree και το S2I. Μαζί με αυτό συνεπάγεται μικρότερο χρόνο κατασκευής για την κατασκευή του ευρετηρίου, όπως επίσης χαμηλότερο κόστος αποθήκευσης του.



Εικόνα 15. Δομή του I3²²

3.4 Dynamically Ranked Top-K Spatial Keyword Search

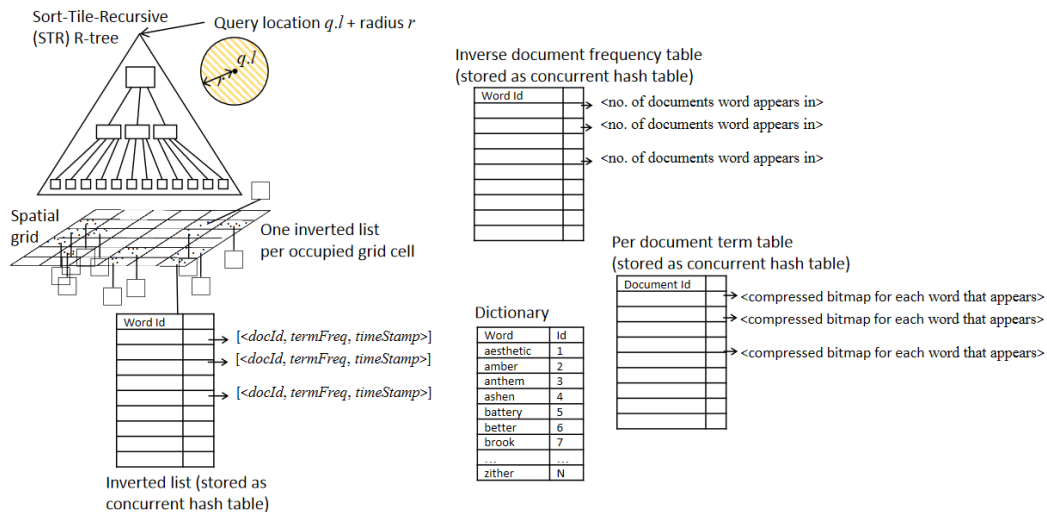
Ένα άλλο νέο ευρετήριο που μελετήθηκε είναι το STARI(Spatio-Textual Adaptive Ranking based Index), το οποίο χρησιμοποιεί αυτήν τη μέθοδο της δυναμικής κατάταξης των χωρο-κειμενικών αντικειμένων για να ανακτήσει τα πιο πρόσφατα σχετικά αντικείμενα(top-k). Στα πειράματα που έγιναν χρησιμοποιήσαν το Twitter ως σύνολο δεδομένων όπου τα tweets επισημάνθηκαν γεωγραφικά από τους συντάκτες του χρησιμοποιώντας ένα πραγματικό σύνολο δεδομένων οδικού δικτύου. Συνολικά σε σύγκριση με το IR-tree, το STARI μπορεί να υποστηρίξει ενημερώσεις σε

²⁰ Samet, Hanan. "The quadtree and related hierarchical data structures." ACM Computing Surveys (CSUR) 16.2 (1984): 187-260.

²¹ Zhang, Dongxiang, Kian-Lee Tan, and Anthony KH Tung. "Scalable top-k spatial keyword search." Proceedings of the 16th international conference on extending database technology. 2013

²² Zhang, Dongxiang, Kian-Lee Tan, and Anthony KH Tung. "Scalable top-k spatial keyword search." Proceedings of the 16th international conference on extending database technology. 2013

πραγματικό χρόνο, ενώ το IR-tree δεν το κάνει όπως επίσης μπορεί να υποστηρίξει υψηλά ποσοστά ενημερώσεων εγγράφων και να υποστηρίξει την εκτέλεση ερωτήματος χαμηλού latency. Οπότε συμπερασματικά το STARI ανακτά τα πιο πρόσφατα έγγραφα χρησιμοποιώντας ένα δυναμικό σχήμα κατάταξης για τον υπολογισμό της σχετικότητας κειμένου.²³



Εικόνα 16. Δομή STARI

3.5 Παρούσα διπλωματική

Στην παρούσα διπλωματική θα δοκιμαστεί κάτι διαφορετικό καθώς θα χρησιμοποιήσουμε ότι ξέρουμε μέχρι τώρα για χωρο – κειμενικά δεδομένα και αναζητήσεις top – k και θα δημιουργήσουμε διαφορετικά δεδομένα που χρειαζόμαστε στην παρούσα διπλωματική. Συγκεκριμένα θα χρησιμοποιήσουμε ότι δεδομένα έχουμε και με βάση κάποια ερωτήματα θα προσπαθήσουμε να συγκρίνουμε τα αποτελέσματα μεταξύ τους για να καταφέρουμε να έχουμε μια ποικιλομορφία στο τελικό μας αποτέλεσμα δηλαδή αποτελέσματα τα οποία μοιάζουν λιγότερο μεταξύ τους. Με αυτό τον τρόπο καταφέρνουμε να έχουμε πολλές και διαφορετικές επιλογές που ικανοποιούν όλο και περισσότερους χρήστες.

²³ Ray, Suprio, and Bradford G. Nickerson. "Dynamically ranked top-k spatial keyword search." Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data. 2016.

4

Πειραματική αποτίμηση της λύσης - Αποτελέσματα

Για να δούμε την απόδοση της μεθόδου μας, καθώς και την ικανότητα να επιλέγει ποικιλόμορφα σημεία ενδιαφέροντος αποτιμήσαμε πειραματικά την προτεινόμενη μέθοδο.

4.1 Δεδομένα διπλωματικής

Τα δεδομένα που χρησιμοποιήσαμε είναι τα εξής:

- I. Ένα αρχείο .txt με όλα τα εστιατόρια μιας περιοχής.
- II. Ένα αρχείο log που περιέχει ερωτήσεις χρηστών για να την αναζήτηση.

4.1.1 Αρχείο με όλα τα εστιατόρια

Το αρχείο αυτό είναι ένα αρχείο κειμένου(.txt) το οποίο περιέχει όλα τα εστιατόρια που θα αναζητήσουμε. Περιέχει συγκεκριμένα 78.981 εγγραφές κάθε μία με διάφορα χαρακτηριστικά που χαρακτηρίζουν το κάθε εστιατόριο.

Τα χαρακτηριστικά αυτά συγκεκριμένα είναι 6:

- ✓ Ένα id σε μορφή ενός ακεραίου
- ✓ Το όνομα του εστιατορίου
- ✓ Μια βαθμολογία με άριστα το 5
- ✓ Το γεωγραφικό μήκος του
- ✓ Το γεωγραφικό πλάτος του
- ✓ Ένα σύνολο από λέξεις – κλειδιά που χαρακτηρίζουν το εστιατόριο

Εμάς από αυτά, τα χαρακτηριστικά που ενδιαφέρουν και θα δώσουμε έμφαση είναι τα 4 από αυτά, το όνομα του, το γεωγραφικό πλάτος και μήκος του και τις λέξεις – κλειδιά που έχει. Όσον αφορά τις λέξεις – κλειδιά είναι λέξεις που χαρακτηρίζουν το εστιατόριο ως προς το είδος του δηλαδή αυτές μπορεί να είναι π.χ. Italian που σημαίνει ότι το συγκεκριμένο εστιατόριο έχει ιταλική κουζίνα. Μπορεί φυσικά να έχει περισσότερα από ένα οπότε εκτός από Italian να έχει και Mexican οπότε να περιέχει και μεξικάνικη μαζί με την ιταλική.

Οπότε με αυτά τα 4 χαρακτηριστικά θα ασχοληθούμε στην εργασία για να μπορέσουμε να βγάλουμε τα συμπεράσματα στο τέλος. Το όνομα το χρειαζόμαστε για να ξέρουμε το όνομα του εστιατορίου, το γεωγραφικό πλάτος και μήκος για να μπορέσουμε να βρούμε την θέση του πάνω στον χάρτη και φυσικά να μετρήσουμε αποστάσεις και τις λέξεις για να δούμε πόσο θα διαφέρουν τα αποτελέσματα μεταξύ τους. Οπότε έχοντας αυτά εμείς θα διατρέχουμε το αρχείο και θα κάνουμε τις απαραίτητες ενέργειες. Ένα παράδειγμα για το πως είναι μια εγγραφή στο αρχείο είναι η εξής:

```
1|Daphne's Greek Cafe|2.5|37.755557|-122.250360846519|Greek, Mediterranean
2|Laguna Pizza|4.5|38.423436|-121.41361|Pizza
3|Starbucks Coffee|4.5|34.095731|-117.717899|American, Coffee, Tea
4|Starbucks Coffee|5|36.341065|-119.347593|Coffee
5|A G Ferrari Foods|4|37.858281|-122.253116|Italian, Deli, Sandwiches
6|Palms|4|34.397338|-119.520776|Californian, Seafood, American, Chicken, Steak
7|A&W Restaurant|3|36.302863|-119.135775|American, Burgers, Hot Dogs, Chicken, Fast Food, Pub Food, Sandwiches
8|A&W Restaurant|4.5|34.272035|-118.50015|American, Barbecue, Dim Sum, Chinese, Seafood
```

Εικόνα 17. Τυχαία εστιατόρια από το αρχείο

Όπως φαίνεται το πρώτο νούμερο είναι ένα id στην περίπτωση μας ο αριθμός του εστιατορίου στο αρχείο. Στην συνέχεια είναι το όνομα του εστιατορίου, μετά η βαθμολογία, το γεωγραφικό πλάτος και μήκος και τέλος οι λέξεις – κλειδιά που μπορεί να είναι μία ή και περισσότερες και όλα αυτά χωρίζονται με μια κάθετη γραμμή.

4.1.2 Αρχείο log

Το αρχείο αυτό είναι ένα αρχείο κειμένου που εμπεριέχει ερωτήσεις χρηστών οι οποίοι ψάχνουν κάποιο εστιατόριο στην περιοχή τους. Στην πειραματική μελέτη χρησιμοποιήσαμε συνθετικά δεδομένα. Όπως και το αρχείο με τα εστιατόρια έτσι και αυτό περιέχει κάποια χαρακτηριστικά που χρειάζονται για την εύρεση των αποτελεσμάτων μέσω των ερωτήσεων. Τα χαρακτηριστικά αυτά είναι 5 και συγκεκριμένα:

- ✓ Ένα id
- ✓ Ένα k που περιγράφει πόσα εστιατόρια να επιστρέψει με βάση την απόσταση.
- ✓ Το γεωγραφικό πλάτος
- ✓ Το γεωγραφικό μήκος
- ✓ Ένα σύνολο από λέξεις – κλειδιά

Μπορεί να είναι τα ίδια με κάποια από αυτά του αρχείου των εστιατορίων αλλά είναι διαφορετικά από εκείνα.

Πιο αναλυτικά το id είναι ένας αριθμός σε μορφή ακεραίου. Πέρα από αυτό τα άλλα έχουν κάποιο διαφορετικό σκοπό. Το k δεν υπάρχει στο άλλο αρχείο αλλά σε αυτό μας χρειάζεται διότι εκπροσωπεί πόσα εστιατόρια ψάχνει ο χρήστης τα οποία είναι πιο κοντά του, δηλαδή αν το k είναι 2 τότε ψάχνουμε τα 2 πιο κοντινά εστιατόρια με βάσει τα χαρακτηριστικά που ψάχνει.

Το γεωγραφικό μήκος και πλάτος μιας ερώτησης εκπροσωπεί την θέση του χρήστη στον χάρτη οπότε με βάση αυτήν μπορούμε να μετρήσουμε την απόσταση από κάποιο εστιατόριο.

Τέλος οι λέξεις – κλειδιά είναι το είδος που ψάχνει ο χρήστης και τον ενδιαφέρει να βρει σε ένα εστιατόριο, οπότε αν έχει γράψει Italian και Mexican σημαίνει ότι ψάχνει εστιατόρια με ιταλική και μεξικάνικη κουζίνα. Ένα παράδειγμα για το πως είναι μια ερώτηση στο αρχείο αυτό είναι η εξής:

```
1|2|33.672452|-118.004725|American
2|2|35.563482|-119.556997|Steak
3|2|32.78014|-120.230014|Italian,Deli
4|2|33.02365|-122.4930112|American,Steak,Seafood
5|2|33.2554391|-117.9830225|American,Mexican
```

Εικόνα 18. Παράδειγμα ερωτήσεων στο αρχείο log

Όπως φαίνεται πάλι το πρώτο νούμερο είναι ο αριθμός της ερώτησης στο αρχείο, στην συνέχεια είναι το k, το γεωγραφικό πλάτος και μήκος ή αλλιώς η θέση του χρήστη στον χάρτη και τέλος οι λέξεις – κλειδιά που ψάχνει ο χρήστης για να βρει.

Όλα τα εστιατόρια βρίσκονται σε μία γεωγραφική περιοχή η οποία καθορίζεται από την διαγώνιο, οπότε στις ερωτήσεις των χρηστών ο χρήστης θα πρέπει να βρίσκεται εντός αυτής της περιοχής. Η διαγώνιος όπως αναφέραμε στα παραδείγματα του Κεφαλαίου 2 υπολογίζεται από το μικρότερο longitude και latitude που σχηματίζουν το ένα σημείο και από το μεγαλύτερο αντίστοιχα για το 2^ο σημείο, υπολογίζοντας την μεταξύ τους απόσταση εύρεση της διαγώνιου.

Εμείς για τις δοκιμές που έγιναν χρησιμοποιήθηκαν 4 διαφορετικά αρχεία με ερωτήσεις και αυτό για να μπορέσουμε να βγάλουμε διαφορετικούς χρόνους σε κάθε μία περίπτωση. Οπότε στην ουσία κάθε φορά για κάθε περίπτωση έχουμε από ένα αρχείο και συγκεκριμένα κάθε αρχείο περιέχει 5 παραπάνω ερωτήσεις από το προηγούμενο. Το πρώτο περιέχει 5 ερωτήσεις χρηστών, το δεύτερο περιέχει τις 5 προηγούμενες δηλαδή του πρώτου αρχείου μαζί με 5 ακόμα καινούργιες, το τρίτο τις 10 προηγούμενες μαζί με 5 καινούργιες και το τέταρτο περιέχει συνολικά 20 ερωτήσεις με τις 15 από τα άλλα αρχεία. Αυτά τα αρχεία δημιουργήθηκαν με τελείως τυχαίες εγγραφές και με ένα k το οποίο είναι διαφορετικό για κάθε χρήστη οπότε μπορεί ο ένας να θέλει το k να είναι 2, ένας άλλος 3, 4 κ.ο.κ. οπότε θα επιστρέφονται διαφορετικός αριθμός σημείων.

4.1.3 Πρόσθετες πληροφορίες υλοποίησης

Στην υλοποίηση μας εκτός από τα αποτελέσματα, χρησιμοποιήθηκαν και ειδικές μεταβλητές για την μέτρηση του χρόνου. Δηλαδή στην υλοποίηση υπάρχουν και τρεις μεταβλητές οι οποίες αποτελούν την έναρξη του χρόνου, την λήξη του και το τελικό αποτέλεσμα που προκύπτει από αυτές τις δύο. Οπότε έτσι θα είμαστε σε θέση να μετρήσουμε τους χρόνους απόδοσης χρησιμοποιώντας διαφορετικές παραμέτρους και βγάζοντας κάποια στατιστικά στοιχεία αυτών των απαντήσεων.

4.2 Αναλυτική υλοποίηση

Το πρόγραμμα μας ξεκινάει διαβάζοντας το αρχείο με τα εστιατόρια και στην συνέχεια τις ερωτήσεις των χρηστών. Τα εστιατόρια αποθηκεύονται σε μία λίστα καθώς και οι ερωτήσεις. Κάθε ερώτηση περιέχει διαφορετικό γεωγραφικό πλάτος και μήκος και διαφορετικές λέξεις-κλειδιά. Εκτός από αυτά υπάρχει και ένα k το οποίο μπορεί να είναι από 2 μέχρι 5 και όπως αναφέρθηκε παρουσιάζει τον αριθμό των εστιατορίων που επιστρέφει με βάση την απόσταση δηλ. τα πιο κοντινά του, π.χ. άμα το k είναι 2 επιστρέφει τα δύο πιο κοντινά εστιατόρια με βάση τις λέξεις-κλειδιά που έχει δηλώσει ενώ άμα το k είναι 5 θα επιστρέψει τα πέντε πιο κοντινά εστιατόρια με βάση τις λέξεις-κλειδιά.

Στο πρόγραμμα δεν χρησιμοποιήθηκαν διαφορετικές κλάσεις για να είναι πιο απλό όπως επίσης χρησιμοποιήθηκε και μόνο ένα αντικείμενο για την έναρξη του. Η γλώσσα προγραμματισμού η οποία χρησιμοποιήθηκε είναι η Java και υλοποιήθηκε στο NetBeans IDE για την εμφάνιση των αποτελεσμάτων.

Οι δοκιμές που έγιναν αναπαριστούν διαφορετικές περιπτώσεις στην εκτέλεση του προγράμματος δηλαδή σε κάθε δοκιμή χρησιμοποιήθηκαν διαφορετικές τιμές σε συγκεκριμένες μεταβλητές που μας ενδιαφέρουν και μετρήθηκαν οι χρόνοι και αξιολογήθηκαν τα αποτελέσματα που μας επιστράφηκαν. Πιο συγκεκριμένα έγιναν οι παρακάτω δοκιμές στο πρόγραμμα:

- ✓ Δοκιμή με 5 ερωτήσεις
- ✓ Δοκιμή με 10 ερωτήσεις
- ✓ Δοκιμή με 15 ερωτήσεις
- ✓ Δοκιμή με 20 ερωτήσεις

Εκτός από αυτές τις περιπτώσεις χρησιμοποιήθηκαν και διαφορετικές τιμές στο εύρος των αποτελεσμάτων που θέλουμε, για παράδειγμα πόσα εστιατόρια θα επιστρέφει το πρόγραμμα στο τέλος για την σύγκριση μεταξύ τους, οπότε μπορεί να θέλουμε να επιστραφούν μόνο 3 εστιατόρια στο τελικό αποτέλεσμα και να συγκρίνουμε την ποικιλομορφία ανάμεσα σε αυτά, ή να θέλουμε να επιστραφούν 4, 5, κ.ο.κ.

4.2.1 Δοκιμές με διαφορετικά αρχεία ερωτήσεων

ΔΟΚΙΜΗ 1

Ξεκινώντας τις δοκιμές του προγράμματος πρώτα θα το τρέξουμε με ανάμεικτα k στο διάστημα από 2 έως 5 και 5 ερωτήσεις .

Οι ερωτήσεις είναι οι εξής:

ID	k	Latitude	Longitude	Keywords
1	5	33,672452	-118,004725	American
2	3	35,563482	-119,556997	Steak
3	3	32,78014	-120,230014	Italian,Deli
4	2	33,02365	-122,4930112	Steak,Seafood
5	3	33,2554391	-117,9830225	American,Mexican

Πίνακας 6. Ερωτήσεις πρώτου αρχείου

Έχουμε το πρώτο αρχείο με τις ερωτήσεις χρηστών όπου ο καθένας έχει διαφορετικές συντεταγμένες και οπότε βρίσκεται σε διαφορετικό σημείο και έχουν διαφορετικές λέξεις – κλειδιά οπότε ψάχνουν διαφορετικά είδη εστιατορίων. Παρακάτω θα αναλυθούν βήμα - βήμα τα αποτελέσματα που προκύπτουν από το πρόγραμμα.

- 1) Για αρχή υπολογίζονται τα top-k οπότε τα πρώτα αποτελέσματα είναι τα k πιο κοντινά εστιατόρια στον κάθε χρήστη με βάση τα keywords που έχει δηλώσει.

```
Top-5 Results:
[Redz Juke Joint, Starbucks Coffee, Baja Fresh, Pizzarito, Sango Sushi Japanese Restaurant]
Top-3 Results:
[IHOP, Chuy's Mesquite Broiler, Applebee's]
Top-3 Results:
[Paesano's Pizzeria, CA' Dario Restaurant, Via Maestra 42]
Top-2 Results:
[Chef Rick's Ultimately Fine, F McIntocks Saloon and Dining]
Top-3 Results:
[Harbor House Cafe No 2, Cabo Cantina, Gemmell's]
```

Εικόνα 19. Αποτελέσματα top-k

Οπότε για κάθε ερώτηση έχουμε από k εστιατόρια τα οποία είναι με βάση την απόσταση τους από τον χρήστη και οπότε είναι τα k πιο κοντινά.

Μπορούμε να παρατηρήσουμε ότι κανένα εστιατόριο από τα παραπάνω δεν εμφανίζεται σε κάποια άλλη ερώτηση οπότε έχουμε 16 διαφορετικά εστιατόρια.

- 2) Το επόμενο βήμα που θα γίνει είναι ελεγχθούν όλες οι δυνάδες μεταξύ τους για να μπορέσει να βγει μέσω της συνάρτησης ομοιότητας ένας αριθμός που θα δηλώνει την ομοιότητα μεταξύ των εστιατορίων.

Αφού βρούμε το ποσοστό ομοιότητας μεταξύ των πρώτων δυάδων το πρόγραμμα διαλέγει το μικρότερο από αυτά δηλαδή την δυάδα με το μικρότερο ποσοστό οπότε και την μικρότερη ομοιότητα μεταξύ τους. Τα εστιατόρια που προκύπτουν στην πρώτη δυάδα είναι:

- F McIntocks Saloon and Dining
- Starbucks Coffee

Η ποσοστό αυτών των δύο βγήκε 0.4883395833872756 που σημαίνει ότι όλα τα άλλα μοιάζουν περισσότερο μεταξύ τους. Αυτό το νούμερο είναι κοντά στο 50% δηλαδή στην μέση περίπου. Όσο πιο κοντά είναι στο 1 τόσο περισσότερο μοιάζουν μεταξύ τους ενώ όσο πιο κοντά στο 0 τόσο περισσότερο δεν μοιάζουν μεταξύ τους.

- 3) Το τρίτο βήμα είναι πολύ σημαντικό καθώς η δυάδα που προκύπτει θα χρησιμοποιηθεί για να φτιαχτεί ένα καινούργιο σημείο το οποίο θα ελεγχθεί με τα υπόλοιπα εστιατόρια που μας έχουν μείνει δηλαδή τα υπόλοιπα 14. Αυτό το καινούργιο σημείο όπως κάναμε στο παράδειγμα του 2^ο κεφαλαίου έχει καινούργιες συντεταγμένες και νέες λέξεις – κλειδιά. Οπότε τώρα μπορεί να ελεγχθεί με τα υπόλοιπα 14 και να βγει ένα καινούργιο σημείο που θα αποτελείται από 3 εστιατόρια που θέλουμε να είναι και το αποτέλεσμα μας. Μετά τους υπολογισμούς το καινούργιο σημείο θα αποτελείται από 3 πλέον εστιατόρια τα οποία είναι:

- F McIntocks Saloon and Dining
- Starbucks Coffee
- CA' Dario Restaurant

Το ποσοστό ανέβηκε στο 0.4985389594945777 που είναι λογικό καθώς προστέθηκε ακόμη ένα στην τελική λίστα αλλά πάλι δεν είναι πάνω από το μισό. Οι λέξεις - κλειδιά που χαρακτηρίζουν αυτά τα 3 σημεία είναι:

- American
- Steak
- Seafood
- Italian
- Deli

Οπότε υπάρχει γενικά μία ποικιλομορφία η οποία δεν είναι η πιο ιδανική παρόλα αυτά έχουμε 5 είδη κουζινών που μπορεί να εξυπηρετήσει διαφορετικούς χρήστες.

Ο χρόνος που καταγράφηκε από την εκτέλεση του προγράμματος μέχρι τον τερματισμό του είναι στα 871 ms ο οποίος εξαρτάται και από το μηχάνημα το οποίο τρέχει το πρόγραμμα.

Εκτός από αυτά αυτήν την τριάδα υπάρχουν και άλλες τριάδες με εστιατόρια τα οποία έχουν τα ίδιες λέξεις – κλειδιά δηλαδή έχουν εμφανιστεί στην ίδια ερώτηση και οπότε το ίδιο ποσοστό.

Συγκεκριμένα άλλες 2 τριάδες οι οποίες είναι:

- F McIntocks Saloon and Dining
 - Starbucks Coffee
 - Via Maestra 42

Και η άλλη:

- F McIntocks Saloon and Dining
 - Starbucks Coffee
 - Paesano's Pizzeria

Οπότε μπορεί να γίνει η επιλογή από μία από αυτές τις τριάδες.

ID	Name	Keywords	# of question
1	F McIntocks Saloon and Dining	Steak,Seafood	4
2	Starbucks Coffee	American	1
3	CA' Dario Restaurant / Via Maestra 42 / Paesano's Pizzeria	Italian,Deli	3

Πίνακας 7. Τριάδα πρώτης δοκιμής

Φυσικά μπορούμε να έχουμε παραπάνω αποτελέσματα καθώς αυτό το καθορίζουμε εμείς. Για παράδειγμα μπορούσαμε να επιλέξουμε να εμφανιστούν 4 εστιατόρια αντί για 3 οπότε θα είχαμε ακόμη ένα στην λίστα καθώς και κάποιο πρόσθετο keyword στα αποτελέσματα μας.

Επιλέγοντας 4 εστιατόρια αντί για 3 το αποτέλεσμα θα είναι το εξής:

- F McIntocks Saloon and Dining
 - Starbucks Coffee
- CA' Dario Restaurant / Via Maestra 42 / Paesano's Pizzeria
 - Gemmell's

Έτσι παρατηρείται ότι προστέθηκε άλλο ένα στην τελική λίστα με το ποσοστό να είναι στο 0.5775181408323171 δηλαδή πάνω από την μέση αυτήν την φορά που σημαίνει ότι όλες οι

υπόλοιπες δυάδες είναι πάνω από αυτό το ποσοστό και οπότε κάποια από αυτά να είναι σχεδόν ίδια όσων αφορά την κουζίνα του καθενός. Οι λέξεις – κλειδιά ήταν:

- American
 - Steak
 - Seafood
 - Italian
 - Deli
 - Mexican

Οπότε έχουμε έτσι 6 διαφορετικές κουζίνες που μπορούν να εξυπηρετήσουν διαφορετικούς χρήστες ανάλογα με τις προτιμήσεις τους.

Και εδώ φυσικά έχουμε άλλες 2 τετράδες με διαφορετικά εστιατόρια οι οποίες μπορούν να επιλεγούν ανάλογα το εστιατόριο. Οι άλλες τετράδες είναι:

- F McIntocks Saloon and Dining
 - Starbucks Coffee
- CA' Dario Restaurant / Via Maestra 42 / Paesano's Pizzeria
 - Harbor House Cafe No 2

Και η άλλη:

- F McIntocks Saloon and Dining
 - Starbucks Coffee
- CA' Dario Restaurant / Via Maestra 42 / Paesano's Pizzeria
 - Cabo Cantina

ID	Name	Keywords	# of question
1	F McIntocks Saloon and Dining	Steak,Seafood	4
2	Starbucks Coffee	American	1
3	CA' Dario Restaurant / Via Maestra 42 / Paesano's Pizzeria	Italian,Deli	3
4	Harbor House Cafe No 2 / Cabo Cantina	American,Mexican	5

Πίνακας 8. Τετράδα πρώτης δοκιμής

Ο χρόνος σε αυτήν την περίπτωση ήταν 1093 ms ή αλλιώς 1 δευτερόλεπτο, προφανώς παραπάνω από τον προηγούμενο καθώς χρειάστηκε ένα παραπάνω υπολογισμό.

ΔΟΚΙΜΗ 2

Σε αυτήν την δοκιμή θα χρησιμοποιήσουμε διαφορετικό αρχείο ερωτήσεων που περιέχει 10 διαφορετικές ερωτήσεις χρηστών με 5 από αυτές να είναι από το προηγούμενο αρχείο. Οι ερωτήσεις είναι οι εξής:

ID	k	Latitude	Longitude	Keywords
1	5	33,672452	-118,004725	American
2	3	35,563482	-119,556997	Steak
3	3	32,78014	-120,230014	Italian,Deli
4	2	33,02365	-122,4930112	Steak,Seafood
5	3	33,2554391	-117,9830225	American,Mexican
6	3	34,2689	-121,785344	Burgers
7	5	33,289254	-119,2543105	American,Fast Food
8	2	35,269013	-118,2301756	Coffee
9	2	32,586501	-120,230158	Deli,Mexican
10	4	34,1089657	-117,125963	Healthy

Πίνακας 9. Ερωτήσεις δεύτερου αρχείου

Και εδώ όλα τα k είναι ανάμεικτα οπότε θέλουμε τα k πιο κοντινά. Εκτελώντας το πρόγραμμα έχουμε:

- 1) Παρακάτω φαίνονται τα top – k αποτελέσματα με βάση την πιο κοντινή απόσταση από τον χρήστη.


```

Top-5 Results:
[Redz Juke Joint, Starbucks Coffee, Baja Fresh, Pizzarito, Sango Sushi Japanese Restaurant]
Top-3 Results:
[IHOP, Chuy's Mesquite Broiler, Applebee's]
Top-3 Results:
[Paesano's Pizzeria, CA' Dario Restaurant, Via Maestra 42]
Top-2 Results:
[Chef Rick's Ultimately Fine, F McIntocks Saloon and Dining]
Top-3 Results:
[Harbor House Cafe No 2, Cabo Cantina, Gemmell's]
Top-3 Results:
[F McIntocks Saloon and Dining, Cheles Food and Spirits, Cracked Crab]
Top-5 Results:
[McDonald's, El Pollo Loco, Del Taco, McDonald's, Burger King]
Top-2 Results:
[Belladonna Tea Room, Panera Bread]
Top-2 Results:
[Catch 21, Casa Pulido]
Top-4 Results:
[A Healthier Las Brasas, Chili's, Souplantation, Chili's]

```

Εικόνα 20. Αποτελέσματα top-k δεύτερου αρχείου

Μπορούμε να παρατηρήσουμε ότι ένα από αυτά εμφανίζεται 2 φορές δηλαδή εμφανίστηκε σε 2 ερωτήσεις και συγκεκριμένα το F McIntocks Saloon and Dining εμφανίζεται στην 4^η και στην 6^η ερώτηση. Επίσης έχουμε 2 McDonald's και 2 Chill's στην ίδια ερώτηση αλλά πρόκειται για 2 διαφορετικά εστιατόρια καθώς στο αρχείο με όλα τα εστιατόρια υπάρχουν πολλά τα οποία είναι η ίδια εταιρία αλλά έχει εστιατόρια σε πολλά και διαφορετικά σημεία.

- 2) Στο 2^ο βήμα εξετάζαμε όλες τις δυάδες για να βρούμε τα 2 με την μικρότερη ομοιότητα με βάση την συνάρτηση sim. Η δυάδα που προκύπτει είναι η εξής:

- A Healthier Las Brasas
- Chef Rick's Ultimately Fine

Το ποσοστό αυτών των δύο ήταν 0.4859206968735555, πάλι κοντά στο 50%.

- 3) Στο τελευταίο βήμα αφού δημιουργηθεί το νέο σημείο μας και ελεγχτεί με τις υπόλοιπες δυάδες προκύπτει η τελική λίστα με τα εστιατόρια. Για 3 εστιατόρια το αποτέλεσμα είναι να έχουμε 2 τριάδες αυτήν την φορά όπου η μία είναι:

- Panera Bread
- A Healthier Las Brasas
- Chef Rick's Ultimately Fine

Και η άλλη είναι:

- Belladonna Tea Room
- A Healthier Las Brasas

- Chef Rick's Ultimately Fine

Η ομοιότητα είναι 0.49402861212809757 και τα keywords που τα χαρακτηρίζουν είναι:

- Healthy
- Steak
- Seafood
- Coffee

ID	Name	Keywords	# of question
1	A Healthier Las Brasas	Healthy	10
2	Chef Rick's Ultimately Fine	Steak,Seafood	4
3	Panera Bread / Belladonna Tea Room	Coffee	8

Πίνακας 10. Τριάδα δεύτερης δοκιμής

Παρατηρείται ότι εδώ έχουμε 4 τελείως άσχετα το ένα με το άλλο και οπότε μπορούμε να πούμε ότι έχουμε αρκετά καλή ποικιλομορφία όσων αφορά τα εστιατόρια μεταξύ τους καθώς η παραπάνω λίστα περιέχει από καφέ μέχρι αμερικάνικη κουζίνα που ενδεχομένως να περιέχει κρέας, όπως επίσης και θαλασσινά για κάποιον που δεν προτιμάει ίσως το κρέας αλλά και υγιεινό για κάποιον δεν προτιμάει τίποτα από τα παραπάνω.

Ο χρόνος εκτέλεσης του προγράμματος ήταν 1994 ms, προφανώς παραπάνω απ' ότι με 5 ερωτήσεις που είχαμε στην πρώτη δοκιμή.

Για 4 εστιατόρια έχουμε 2 τετράδες όπου η μία είναι:

- A Healthier Las Brasas
- Chef Rick's Ultimately Fine
- Panera Bread / Belladonna Tea Room
 - Cheles Food and Spirits

Και η άλλη είναι:

- A Healthier Las Brasas
- Chef Rick's Ultimately Fine
- Panera Bread / Belladonna Tea Room
 - Cracked Crab

ID	Name	Keywords	# of question
1	A Healthier Las Brasas	Healthy	10
2	Chef Rick's Ultimately Fine	Steak,Seafood	4
3	Panera Bread / Belladonna Tea Room	Coffee	8
4	Cheles Food and Spirits / Cracked Crab	Burgers	6

Πίνακας 11. Τετράδα δεύτερης δοκιμής

Ομοιότητα: 0.49355663900495766

Keywords: Healthy, Steak, Seafood, Coffee, Burgers

Χρόνος: 2245 ms

Η λίστα με τις κουζίνες είναι η ίδια αλλά αυτήν την φορά προτέθηκε και άλλη μία που είναι τα Burgers.

Δοκιμή 3

Σε αυτήν την δοκιμή οι ερωτήσεις γίνονται 15 και τα k παραμένουν ανάμεικτα. Οι ερωτήσεις είναι οι εξής:

ID	k	Latitude	Longitude	Keywords
1	5	33,672452	-118,004725	American
2	3	35,563482	-119,556997	Steak
3	3	32,78014	-120,230014	Italian,Deli
4	2	33,02365	-122,4930112	Steak,Seafood
5	3	33,2554391	-117,9830225	American,Mexican
6	3	34,2689	-121,785344	Burgers
7	5	33,289254	-119,2543105	American,Fast Food
8	2	35,269013	-118,2301756	Coffee
9	2	32,586501	-120,230158	Deli,Mexican
10	4	34,1089657	-117,125963	Healthy
11	3	33,250369	-119,2003699	Ice Cream,Bakery
12	3	35,2565899	-118,2300477	Asian
13	2	33,526589	-118,3300125	Coffee,Tea

14	4	33,589314	-119,7856254	Mediterranean
15	2	34,0012548	-118,845231	Salad,Sandwiches

Πίνακας 12. Ερωτήσεις τρίτου αρχείου

1) Το top-k εμφάνισε τα εξής εστιατόρια:

```

Top-5 Results:
[Redz Juke Joint, Starbucks Coffee, Baja Fresh, Pizzarito, Sango Sushi Japanese Restaurant]
Top-3 Results:
[IHOP, Chuy's Mesquite Broiler, Applebee's]
Top-3 Results:
[Paesano's Pizzeria, CA' Dario Restaurant, Via Maestra 42]
Top-2 Results:
[Chef Rick's Ultimately Fine, F McIntocks Saloon and Dining]
Top-3 Results:
[Harbor House Cafe No 2, Cabo Cantina, Gemmell's]
Top-3 Results:
[F McIntocks Saloon and Dining, Cheles Food and Spirits, Cracked Crab]
Top-5 Results:
[McDonald's, El Pollo Loco, Del Taco, McDonald's, Burger King]
Top-2 Results:
[Belladonna Tea Room, Panera Bread]
Top-2 Results:
[Catch 21, Casa Pulido]
Top-4 Results:
[A Healthier Las Brasas, Chili's, Souplantation, Chili's]
Top-3 Results:
[Baskin-Robbins, Baskin-Robbins, Infuzion Cafe]
Top-3 Results:
[Quiznos, Gold Fish Japanese Restaurant, Shandra's Thai-Chinese Cuisine]
Top-2 Results:
[Coffee Bean and Tea Leaf, It's A Grind Coffee House]
Top-4 Results:
[Pasucci, Wine Cask, CA' Dario Restaurant, Jonathans At Peiranos]
Top-2 Results:
[Chili's, California Pizza Kitchen]

```

Εικόνα 21. Αποτελέσματα top-k τρίτου αρχείου

2) Η πρώτη δυάδα που προκύπτει είναι η εξής:

- A Healthier Las Brasas
- Chef Rick's Ultimately Fine

Ομοιότητα: 0.4859206968735555

Τα νούμερά αλλά και τα εστιατόρια είναι τα ίδια με την προηγούμενη δοκιμή στο 2^ο βήμα που σημαίνει ότι παραμένει αυτό με το λιγότερο ποσοστό.

3) Στην τελική λίστα των εστιατορίων υπάρχουν 2 τριάδες. Η μία είναι:

- A Healthier Las Brasas

- Chef Rick's Ultimately Fine
- Panera Bread

Και η άλλη λίστα είναι:

- A Healthier Las Brasas
- Chef Rick's Ultimately Fine
- Belladonna Tea Room

Οι δύο λίστες εδώ μπορούμε να παρατηρήσουμε ότι έχουν τα ίδια εστιατόρια με της δοκιμής 2 οπότε έχουν και τα ίδιες λέξεις – κλειδιά. Αυτό κανονικά είναι δύσκολο να γίνει καθώς το πρόγραμμα παίρνει υπόψιν και την απόσταση οπότε θα έπρεπε να υπήρχε έστω μία ελάχιστη διαφορά καθώς προστέθηκαν καινούργιες ερωτήσεις στο τρίτο αρχείο αλλά στην συγκεκριμένη περίπτωση τα σημεία που βρέθηκαν στην δοκιμή 2 είναι και εδώ τα βέλτιστα καθώς είναι σε απόσταση πιο κοντινά από οποιαδήποτε καινούργιο σημείο που βρέθηκε και περιέχει τις λέξεις – κλειδιά που αναζητήθηκαν και οπότε θεωρούνται τα ίδια με της δοκιμής 2 με το ίδιο ποσοστό στο 0.49402861212809757 και keywords:

- Healthy
- Steak
- Seafood
- Coffee

Οπότε έτσι μπορεί να γίνει επιλογή για μία από τις 2 τριάδες για τελική λίστα.

Ο χρόνος εκτέλεσης ήταν στα 4186 ms ακόμα περισσότερο φυσικά καθώς έχουμε πιο πολλές ερωτήσεις. Και για την τετράδα έχουμε τα ίδια αποτελέσματα οπότε έχουμε τα ίδια 4 σημεία να παραμένουν ίδια με της δοκιμής 2.

Δοκιμή 4

Σε αυτήν την δοκιμή οι ερωτήσεις γίνονται 20 και τα k παραμένουν ανάμεικτα. Οι ερωτήσεις είναι οι εξής:

1) Το top-k εμφάνισε τα εξής εστιατόρια:

ID	k	Latitude	Longitude	Keywords
1	5	33,672452	-118,004725	American
2	3	35,563482	-119,556997	Steak

3	3	32,78014	-120,230014	Italian,Deli
4	2	33,02365	-122,4930112	Steak,Seafood
5	3	33,2554391	-117,9830225	American,Mexican
6	3	34,2689	-121,785344	Burgers
7	5	33,289254	-119,2543105	American,Fast Food
8	2	35,269013	-118,2301756	Coffee
9	2	32,586501	-120,230158	Deli,Mexican
10	4	34,1089657	-117,125963	Healthy
11	3	33,250369	-119,2003699	Ice Cream,Bakery
12	3	35,2565899	-118,2300477	Asian
13	2	33,526589	-118,3300125	Coffee,Tea
14	4	33,589314	-119,7856254	Mediterranean
15	2	34,0012548	-118,845231	Salad,Sandwiches
16	2	34,217774	-119,978231	Seafood,Steak
17	3	32,589645	-121,271002	Mexican,Tex Mex
18	3	35,545885	-119,565457	Diner
19	2	36,100256	-120,254698	Pizza
20	2	34,788965	-118,710251	Vegetarian

Πίνακας 13. Ερωτήσεις τέταρτου αρχείου

2) Η πρώτη δυάδα που προκύπτει είναι η εξής:

- Souplantation
- Chef Rick's Ultimately Fine

Τα αποτελέσματα εδώ έδειξαν μια διαφορετική δυάδα με τις προηγούμενες ερωτήσεις.

3) Και εδώ στην τελική λίστα των εστιατορίων υπάρχουν 2 τριάδες εδώ καθώς έχουν το ίδιο ποσοστό. Η μία λίστα είναι:

- Souplantation
- Chef Rick's Ultimately Fine
- Dominos Pizza

Και η άλλη είναι:

- Souplantation

- Chef Rick's Ultimately Fine
- Central Coast Pizza

ID	Name	Keywords	No. of question
1	Souplantation	Healthy	10
2	Chef Rick's Ultimately Fine	Steak, Seafood	4
3	Dominos Pizza / Central Coast Pizza	Pizza	19

Πίνακας 14. Τριάδα τέταρτης δοκιμής

Ποσοστό και των 2 εκδοχών είναι το ίδιο και είναι 0.49338480820877145.

Τα keywords εδώ είναι λίγο διαφορετικά σε σχέση με την προηγούμενη δοκιμή και τώρα έχουμε:

- Healthy
- Steak
- Seafood
- Pizza

Έχουμε 4 τελείως διαφορετικές κουζίνες και εδώ από υγιεινά μέχρι κρέας, θαλασσινά αλλά και πίτσα. Οπότε μπορεί να επιλεγθεί μία από τις 2 εκδοχές καθώς η ποικιλομορφία είναι η ίδια.

Χρόνος εκτέλεσης είναι: 6984 ms.

Για 4 εστιατόρια έχουμε 3 τετράδες, η μία είναι:

- Souplantation
- Chef Rick's Ultimately Fine
- Central Coast Pizza / Dominos Pizza
 - Gemmell's

Η άλλη είναι:

- Souplantation
- Chef Rick's Ultimately Fine
- Central Coast Pizza / Dominos Pizza
 - Cabo Cantina

Και η τελευταία είναι:

- Souplantation
- Chef Rick's Ultimately Fine
- Central Coast Pizza / Dominos Pizza
- Harbor House Cafe No 2

ID	Name	Keywords	No. of question
1	Souplantation	Healthy	10
2	Chef Rick's Ultimately Fine	Steak,Seafood	4
3	Dominos Pizza / Central Coast Pizza	Pizza	19
4	Gemmell's / Cabo Cantina / Harbor House Cafe No 2	American,Mexican	5

Πίνακας 15. Τετράδα τέταρτης δοκιμής

Ο χρόνος εκτέλεσης είναι 8199 ms.

5

Ανάλυση αποτελεσμάτων και συμπεράσματα

5.1 Αξιοπιστία αποτελεσμάτων

5.1.1 Περιγραφή δεδομένων

Για να φτάσουμε στα αποτελέσματα χρησιμοποιήσαμε 2 είδη αρχείων σε μορφή κειμένου: το αρχείο Restaurants που περιείχε 78981 εστιατόρια το καθένα με κάποια χαρακτηριστικά και το αρχείο Questions το οποίο το σπάσαμε σε 4 διαφορετικά αρχεία με ερωτήσεις για τα διαφορετικά αποτελέσματα μας. Στο πρόγραμμα χρησιμοποιήθηκαν όλα αυτά για την υλοποίηση και την εξακρίβωση των αποτελεσμάτων.

Εκτός από αυτά έχουμε και τον χώρο στον οποίο βρίσκονται τα εστιατόρια ή αλλιώς τα όρια τα οποία ορίζονται από τον υπολογισμό της διαγωνίου. Η διαγώνιος προκύπτει από 4 τιμές: το μέγιστο γεωγραφικό μήκος και πλάτος, το ελάχιστο γεωγραφικό μήκος και πλάτος. Οι τιμές τους φαίνονται στον παρακάτω πίνακα.

Μέγιστο γεωγραφικό μήκος	48.847931
Μέγιστο γεωγραφικό πλάτος	47.669728
Ελάχιστο γεωγραφικό μήκος	-42.118331
Ελάχιστο γεωγραφικό πλάτος	-124.20028

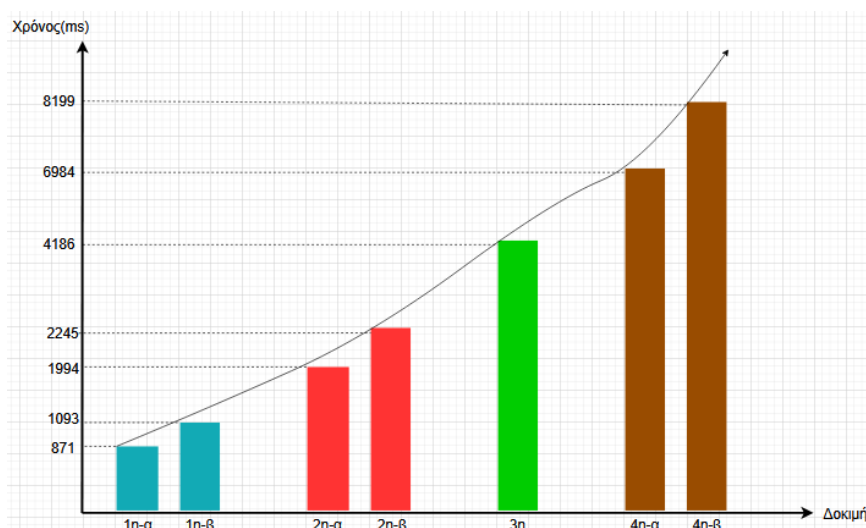
Πίνακας 16. Μέγιστες και ελάχιστες τιμές συντεταγμένων

Με βάση αυτές τις συντεταγμένες υπολογίστηκε η ευκλείδεια απόσταση με αποτέλεσμα η διαγώνιος μας να είναι 194.458634.

5.1.2 Ανάλυση αποτελεσμάτων

Στην ανάλυση αποτελεσμάτων θα αναλύσουμε τους χρόνους εκτέλεσης σε κάθε περίπτωση, τα διαφορετικά είδη εστιάτοριών που μας επιστράφηκαν καθώς και πόσο ικανοποιητικά ήταν με βάση αυτά που περιμέναμε.

5.1.2.1 Χρόνοι εκτέλεσης

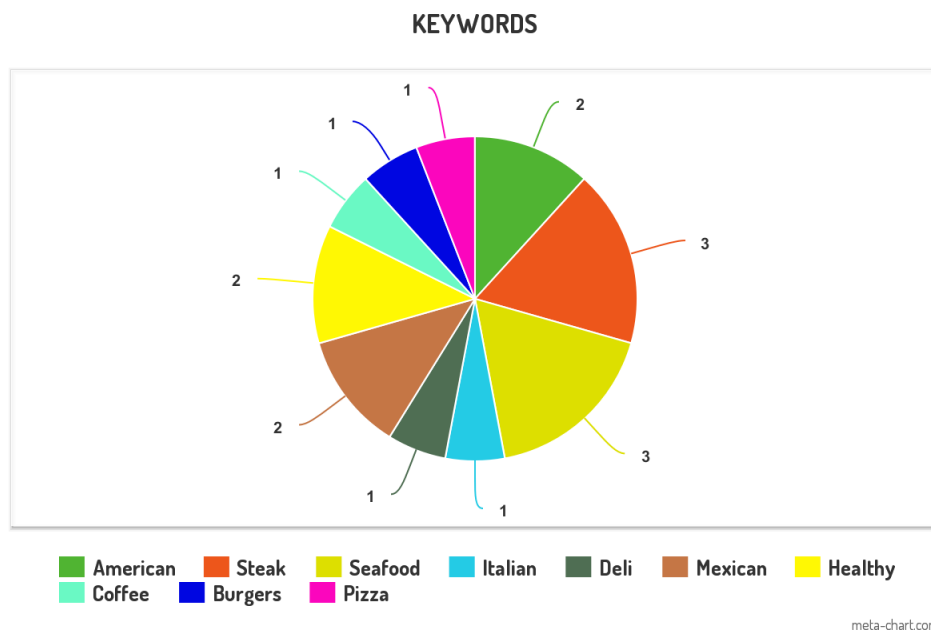


Εικόνα 22. Διάγραμμα χρόνων εκτέλεσης

Στο παραπάνω διάγραμμα το οποίο φτιάχτηκε με την βοήθεια έτοιμου εργαλείου²⁴ απεικονίζονται όλοι οι χρόνοι που σημειώθηκαν κατά της εκτέλεση του προγράμματος για διάφορες δοκιμές όπως αναφέρθηκαν στο Κεφάλαιο 4. Μπορούμε να παρατηρήσουμε ότι όσο περισσότερες ήταν οι ερωτήσεις στο αρχείο τόσο μεγαλύτερο χρόνο χρειαζόταν για τον υπολογισμό των αποτελεσμάτων πράγμα το οποίο είναι λογικό καθώς επεξεργάζεται περισσότερα δεδομένα στην περίπτωση μας περισσότερα ερωτήματα. Αν ξανατρέχαμε κάποια καινούργια δοκιμή οι χρόνοι πιθανώς να ήταν διαφορετικοί καθώς μιλάμε για κλάσματα του δευτερολέπτου(miliseconds), αλλά δεν θα είχαν μεγάλη διαφορά με τους προηγούμενους. Όπως παρατηρείται οι χρόνοι ξεκινάνε από 871 ms δηλαδή σχεδόν 1 δευτερόλεπτο και φτάνουν μέχρι και τα 8190 ms ή αλλιώς 8 δευτερόλεπτα. Να σημειωθεί ότι αυτοί οι χρόνοι εξαρτιόνται και από την υπολογιστική ισχύ υπάρχει, οπότε ένα πιο δυνατό μηχάνημα προφανώς θα χρειαζόταν λιγότερο χρόνο για να τρέξει τις παραπάνω δοκιμές.

²⁴ <https://app.diagrams.net>

5.1.2.2 Ανάλυση διαφορετικών ειδών εστιατορίων



Εικόνα 23. Πίτα διαφορετικών ειδών εστιατορίων

Όπως φαίνεται στο παραπάνω διάγραμμα πίτας το οποίο και αυτό δημιουργήθηκε με έτοιμο εργαλείο²⁵ έχουμε τα διαφορετικά είδη εστιατορίων που εμφανίστηκαν στα αποτελέσματα μας. Εκτός από το όνομα έχουμε και έναν αριθμό που αντιπροσωπεύει πόσες φορές εμφανίστηκε στα αποτελέσματα μας ή διαφορετικά σε πόσες δοκιμές. Συνολικά έχουμε 10 διαφορετικά είδη και για τις 4 δοκιμές που έγιναν. Οι περισσότερες από αυτές τις λέξεις εμφανίστηκαν μόνο μία φορά καθώς μετά προσθέτονταν και άλλες ερωτήσεις με διαφορετικές λέξεις – κλειδιά οπότε ήταν πιο κοντά ή πιο ανόμοια σε σχέση με τα άλλα. Δύο από αυτές τις λέξεις – κλειδιά που εμφανίστηκαν περισσότερες φορές είναι το Steak και το Seafood, δύο τελείως διαφορετικές κουζίνες μεταξύ τους που εμφανίστηκαν σε 3 από τις 4 δοκιμές. Σε 2 από τις 4 δοκιμές εμφανίστηκαν 3 είδη τα οποία είναι το American, Healthy και Mexican. Τα 3 αυτά είδη είναι και αυτά διαφορετικά μεταξύ τους οπότε μπορούν να ικανοποιήσουν διαφορετικούς χρήστες και εδώ. Όλα τα υπόλοιπα εμφανίστηκαν από 1 φορά σε 4 δοκιμές που έγιναν. Συνολικά οι λέξεις – κλειδιά και από τις 20 ερωτήσεις που έγιναν ήταν 30 με μοναδικές από αυτές να ήταν 21, οπότε είχαμε 21 διαφορετικές λέξεις για αναζήτηση. Υπενθυμίζεται ότι η αναζήτηση γινόταν όταν ένα εστιατόριο περιείχε όλα τις λέξεις – κλειδιά και όχι μία από αυτές, σε διαφορετική περίπτωση τα αποτελέσματα θα ήταν πολύ περισσότερα καθώς κάποια σημεία περιείχαν έστω μία από τις λέξεις – κλειδιά που αναζητούσε ένας χρήστης αλλά όχι τις υπόλοιπες.

²⁵ <https://www.meta-chart.com>

5.1.2.3 Βαθμός ικανοποίησης

Η αρχικός στόχος αυτής της διπλωματικής ήταν να βρούμε τα πιο ποικιλόμορφα σημεία έτσι ώστε να εξυπηρετούνται περισσότεροι διαφορετικοί χρήστες άρα και διαφορετικές προτιμήσεις. Με βάση τα αποτελέσματα μας μπορούμε να πούμε ότι ο στόχος επιτεύχθηκε καθώς βρέθηκαν διάφορα σημεία τα οποία είναι διαφορετικά μεταξύ τους όσον αφορά το είδος τους και την απόσταση από τον χρήστη καθώς θέλουμε να είναι και αυτή η βέλτιστη και συγκεκριμένα η μικρότερη. Μπορούσαν να γίνουν και άλλες δοκιμές μόνο με διαφορετικές λέξεις – κλειδιά όπου εκεί πέρα θα έπαιζε μεγαλύτερο ρόλο η απόσταση καθώς η ποικιλομορφία θα υπήρχε διότι όλες οι λέξεις θα ήταν διαφορετικές και θα υπήρχαν στις ερωτήσεις μόνο μία φορά. Στις δοκιμές τις δικές μας χρησιμοποιήθηκαν κάποιες λέξεις – κλειδιά πάνω από μία φορά για να διατηρήσουμε ότι μπορεί να υπάρχουν χρήστες που ψάχνουν το ίδιο με άλλους και ότι όχι όλοι ψάχνουν κάτι διαφορετικό. Τα αποτελέσματα ήταν ικανοποιητικά καθώς υπήρχε ποικιλομορφία στα εστιατόρια με κάποιες εξαιρέσεις καθώς υπήρχε και η μεταβλητή της απόστασης που επηρέαζε το τελικό αποτέλεσμα και όχι μόνο οι λέξεις – κλειδιά.

5.2 Προβλήματα που προέκυψαν

Κατά την διάρκεια της υλοποίησης δεν υπήρξαν κάποια προβλήματα με τον κώδικα ή με τα δεδομένα μας. Υπάρχει πάντα το πρόβλημα ότι μπορεί να προκύψουν κάποια καθώς έχουμε να διαχειριστούμε πολλά δεδομένα σε αριθμό και μας ενδιαφέρει τα αποτελέσματα να είναι ικανοποιητικά στο 100% και να μην υπάρχουν παραλήψεις και άλυτα θέματα. Παρόλα αυτά το πρόγραμμα αν και μεγάλο σε μέγεθος και ίσως όχι το πιο βέλτιστο από άποψη μνήμης και κώδικα καθώς δεν μας ενδιέφερε αυτό το κομμάτι παρά μόνο να τρέχει σωστά, κάνει την δουλειά που πρέπει εκτυπώνοντας τα αποτελέσματα που θέλουμε.

5.3 Εξέταση δυνατότητας πρακτικής εφαρμογής

Όπως αναφέρθηκε η διπλωματική αυτή εκτός από την ανάλυση των δεδομένων και των αποτελεσμάτων χρησιμοποιήθηκε κάποιο πρόγραμμα για να λυθούν αυτά και να έχουμε κάποια συμπεράσματα. Το πρόγραμμα αυτό γράφτηκε χρησιμοποιώντας μία γλώσσα προγραμματισμού και μία εφαρμογή για την εκτέλεση των εντολών. Παρόλα αυτά η υλοποίηση για την συγκεκριμένη διπλωματική είναι σχετικά απλή καθώς δεν χρειάστηκε να αναπτυχθεί κάποια εφαρμογή για υπολογιστή ή κινητό παρά μόνο να γραφτεί ο κώδικας για την επεξεργασία των δεδομένων και την εκτύπωση τους χωρίς κάποιο γραφικό περιβάλλον που θα αλληλοεπιδρά ο χρήστης. Φυσικά αυτή η υλοποίηση μπορεί να πάει ένα βήμα παραπάνω και δημιουργηθεί κάποιου είδους εφαρμογή είτε για κινητά είτε για υπολογιστή ή tablet με γραφικό περιβάλλον που θα μπορεί ένας χρήστης να επιλέγει με την μορφή επιλογών να του εμφανίσει αυτό που θέλει.

6

Συμπεράσματα

Η διπλωματική αυτή είχε ως στόχο μια πρωτότυπη ιδέα η οποία ήταν η εύρεση των πιο ποικιλόμορφων σημείων με βάση κάποιες λέξεις – κλειδιά και την απόσταση τους από τον χρήστη σε μία γεωγραφική περιοχή. Για να γίνει αυτό βοήθησαν τα δεδομένα που είχαμε στην διάθεση μας που ήταν τα εστιατόρια με κάποια χαρακτηριστικά που τα περιέγραφαν όπως τοποθεσία στον χάρτη, λέξει – κλειδιά και οι ερωτήσεις των χρηστών με βάση τις προτιμήσεις τους και την θέση τους στο χάρτη. Με βάση αυτά καταφέραμε να βρούμε διάφορα εστιατόρια διαφορετικά μεταξύ τους έτσι ώστε να υπάρχει αυτή η ποικιλομορφία(diversity) που θέλαμε και φυσικά επιλέγοντας πρώτα τα πιο κοντινά σε έναν χρήστη. Δημιουργήθηκε και πρόγραμμα το οποίο θα μπορούσε να μετατραπεί και σε εφαρμογή για την καλύτερη εξυπηρέτηση των χρηστών. Τα συγκεκριμένα αποτελέσματα φυσικά δεν μπορούν να ανταποκριθούν στην πραγματικότητα καθώς το αρχείο με τις ερωτήσεις των χρηστών το οποίο ήταν χωρισμένο σε 4 νέα αρχεία για τις δοκιμές που έγιναν έγινε με τυχαίο τρόπο δεν ήταν πραγματικά δεδομένα χρηστών. Με βάση αυτά υπάρχουν προοπτικές βελτίωσης για το μέλλον με πραγματικά δεδομένα για την καλύτερη εξυπηρέτηση πραγματικών χρηστών.

Βιβλιογραφία

Jiang, Bin, and Xiaobai Yao. "Location-based services and GIS in perspective." *Computers, Environment and Urban Systems* 30.6 (2006): 712-725.

A. Guttman. R-trees: A dynamic index structure for spatial searching. *Proc. SIGMOD*, pages 47–57, 1984.

N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. *Proc. SIGMOD*, pages 322–331, 1990.

De Felipe, Ian, Vagelis Hristidis, and Naphtali Rishe. "Keyword search on spatial databases." 2008 IEEE 24th International Conference on Data Engineering. IEEE, 2008.

Zhang, Dongxiang, et al. "Keyword search in spatial databases: Towards searching by document." 2009 IEEE 25th International Conference on Data Engineering. IEEE, 2009.

Samet, Hanan. "The quadtree and related hierarchical data structures." *ACM Computing Surveys (CSUR)* 16.2 (1984): 187-260.

Zhang, Dongxiang, Kian-Lee Tan, and Anthony KH Tung. "Scalable top-k spatial keyword search." *Proceedings of the 16th international conference on extending database technology*. 2013

Ray, Suprio, and Bradford G. Nickerson. "Dynamically ranked top-k spatial keyword search." *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data*. 2016.

Malczewski, Jacek. "GIS-based land-use suitability analysis: a critical overview." *Progress in planning* 62.1 (2004): 3-65.

Peng Yue, Zhenyu Tan, in *Comprehensive Geographic Information Systems*, 2018

Shaun Fontanella, Ningchuan Xiao, in *Comprehensive Geographic Information Systems*, 2018

Güting, Ralf Hartmut. "An introduction to spatial database systems." the VLDB Journal 3.4 (1994): 357-399.

Leila de Floriani, ... Enrico Puppo, in Handbook of Computational Geometry, 2000