ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Κατάταξη χωρο-κειμενικών δεδομένων μεγάλης κλίμακας με βάση καινοτόμους τρόπους ταξινόμησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Πάναλη Κυριάκου

Επιβλέπων : Αν. Καθηγήτρια Ακριβή Βλάχου, Πανεπιστήμιο Αιγαίου

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

Σάμος, Μάρτιος 2021

Πρόλογος και ευχαριστίες

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

ΠΑΝΑΛΗ ΚΥΡΙΑΚΟΥ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πίνακας περιεχομένων

Λίστα Εικόνων

Λίστα Πινάκων

Περίληψη

Με τις νεότερες τεχνολογίες που απευθύνονται στην ασύρματη σύνδεση στο διαδίκτυο μέσω των έξυπνων τηλεφώνων και στην πραγματοποίηση ολοένα και παραπάνω ενεργειών μέσω αυτών, προκύπτει η ανάγκη για την ανάπτυξη καινοτόμων εργαλείων που ανταποκρίνονται στις ανάγκες των χρηστών. Ένα ακόμη σημαντικό κομμάτι της σύγχρονης τεχνολογίας, είναι η δυνατότητα εύρεσης της τοποθεσίας του εκάστοτε χρήστη, όπως και η εξατομίκευση που προσφέρεται από αυτή. Οι έξυπνες κινητές συσκευές, προσφέρουν στους χρήστες την δυνατότητα να ψάχνουν για σημεία ενδιαφέροντος, τα οποία βρίσκονται σε κοντινή απόσταση και πληρούν κάποιες προϋποθέσεις, σύμφωνα με κριτήρια που καθορίζονται από τον χρήστη. Τα σημεία αυτά μπορούν να είναι επιχειρήσεις οι οποίες προσφέρουν προϊόντα ή υπηρεσίες. Σκοπός της παρούσας Διπλωματικής εργασίας, είναι η κατάταξη των χωρο-κειμενικών δεδομένων που αναπαριστούν κάποια σημεία ενδιαφέροντος, σύμφωνα με τις αναζητήσεις που πραγματοποιούν για αυτά οι χρήστες από μία συγκεκριμένη τοποθεσία. Για την κατάταξη των σημείων ενδιαφέροντος, δημιουργείτε ένας γράφος, όπου σε συνδυασμό με αλγορίθμους κατάταξης προκύπτουν χρήσιμα συμπεράσματα για αυτά.

Λέξεις Κλειδιά: *Υπηρεσίες τοποθεσίας, χωρο-κειμενικά δεδομένα, σημεία ενδιαφέροντος, αναζήτηση, γράφος, κατάταξη, λέξεις κλειδιά, χωρικές ερωτήσεις*

Abstract

With the latest technologies aimed at wireless internet connection through smart phones and the implementation of more and more actions through them, the need arises for the development of innovative tools that meet the needs of users. Another important piece of modern technology is the ability to find the location of each user, as well as the personalization offered by it. Smartphones offer users the ability to search for points of interest, which are within a certain distance and meet certain conditions, according to criteria set by the user. These points can be companies that offer products or services. The purpose of this Thesis is the classification of spatial-textual data that represent some points of interest, according to the searches performed for them by users from a specific location. To rank the points of interest, we create a graph, where in combination with rank algorithms, provide useful conclusions.

Keywords: *Location based services, spatial-textual data, points of interest, search, graph, rank, keywords, spatial queries*

1 Εισαγωγή

1.1 Ανάκτηση δεδομένων με βάση την τοποθεσία

Είναι γνωστό ότι βρισκόμαστε στην εποχή της πληροφορίας και της αυτοματοποίησης, με την τεχνολογία να αποτελεί πλέον αναπόσπαστο κομμάτι της καθημερινότητας του μέσου ανθρώπου. Κάνοντας χρήση των διαφόρων τεχνολογιών, έχουμε καταφέρει να κάνουμε την ζωή μας πολύ πιο εύκολη, εξοικονομώντας ταυτόχρονα πολύτιμο χρόνο, αφού έχουμε απαλλαγεί από διεργασίες που πλέον γίνονται αυτοματοποιημένα, όπως για παράδειγμα την αποπληρωμή λογαριασμών και την κατάθεση φορολογικών δηλώσεων. Κάθε σύγχρονος άνθρωπος κατέχει πλέον τουλάχιστον μία ηλεκτρονική συσκευή, η οποία συνήθως είναι φορητή και παρέχει την δυνατότητα για σύνδεση στο διαδίκτυο. Η πιο δημοφιλής συσκευή που χρησιμοποιείτε πλέον ευρέως είναι τα έξυπνα τηλέφωνα ή smartphones, με περισσότερους από 3,5 δισεκατομμύρια χρήστες παγκοσμίως[1] . Κάθε σύγχρονη υπολογιστική συσκευή, εκτός από την δυνατότητα σύνδεσης στο διαδίκτυο, παρέχει και την τεχνολογία εύρεσης της γεωγραφικής θέσης της συσκευής, και κατ' επέκταση του χρήστη της.

Η εξέλιξη των ασύρματων επικοινωνιών κατέστησε αναγκαία την ανάπτυξη εφαρμογών και υπηρεσιών, οι οποίες καλούνται να αξιοποιήσουν τις υπάρχουσες τεχνολογίες και να προσφέρουν στους τελικούς χρήστες την δυνατότητα να εκμεταλλευτούν την εύρεση της τοποθεσίας τους. Έτσι λοιπόν προέκυψε ο όρος Υπηρεσίες Τοποθεσίας (Location Based Services - LBS), ο οποίος αναφέρεται σε λογισμικό, το οποίο χρησιμοποιεί κατά βάση την τοποθεσία του χρήστη, και κατ' επέκταση προσφέρει υπηρεσίες με βάση αυτή. Στην συγκεκριμένη κατηγορία λογισμικού μπορούν να ενταχθούν συστήματα πλοήγησης, υπηρεσίες έκτακτης ανάγκης, εφαρμογές εντοπισμού κατοικίδιου κ.ο.κ. (Schiller et al, 2004).

Η αρχή για την εγκαθίδρυση των Υπηρεσιών Τοποθεσίας (Location Based Services – LBS) έγινε στην αρχή της δεκαετίας του 1970, όταν το υπουργείο εθνικής άμυνας των Ηνωμένων Πολιτειών της Αμερικής ξεκίνησε να λειτουργεί το σύστημα global positioning system γνωστό ως GPS. Το συγκεκριμένο σύστημα αναπτύχθηκε αρχικά για στρατιωτικούς σκοπούς, αλλά την δεκαετία του 1980 η κυβέρνηση των Ηνωμένων Πολιτειών αποφάσισε να το κάνει διαθέσιμο παγκοσμίως για

1. www.statista.com/statistics/330695/number-of-smartphone-users-worldwide

διάφορες άλλες χρήσεις. Σκοπός του συστήματος GPS είναι η ανεύρεση της τοποθεσίας ενός χρήστη, του οποίου η συσκευή διαθέτει την κατάλληλη κεραία. Το συγκεκριμένο σύστημα βασίζεται σε ένα πλέγμα εικοσιτεσσάρων δορυφόρων, εφοδιασμένων με ειδικές συσκευές εντοπισμού, οι οποίες ονομάζονται πομποδέκτες GPS. Οι συσκευές αυτές αναλαμβάνουν την παροχή πληροφοριών, προσδιορίζοντας τη θέση, το υψόμετρο και την ταχύτητα με την οποία κινείται στον χώρο μία συσκευή που διαθέτει την κατάλληλη κεραία. Επίσης γίνεται εφικτή η γραφική απεικόνιση των παραπάνω πληροφοριών, κάνοντας χρήση ειδικού λογισμικού χαρτογράφησης[2].

Για τον ακριβή εντοπισμό της θέσης, γίνεται κάθε φόρα χρήση τεσσάρων διαφορετικών δορυφόρων, ενώ χρησιμοποιείται ο μαθηματικός τριγωνισμός. Κάθε ένας απ' τους δορυφόρους συνδράμει με διαφορετικό τρόπο για τον ακριβή εντοπισμό της εκάστοτε συσκευής. Αντίστοιχα συστήματα με αυτό των Ηνωμένων Πολιτειών ανέπτυξαν αργότερα η Ρωσία (GLONASS), η Κίνα (BeiDou), καθώς και η Ευρωπαϊκή Ένωση (Galileo).

Με την ανάπτυξη τεχνολογιών που καθιστούν εφικτή την γνωστοποίηση της θέσης του εκάστοτε χρήστη, προκύπτει η ανάπτυξη νέων τεχνολογιών λογισμικού, οι οποίες καλούνται να εκμεταλλευτούν την δυνατότητα αυτή. Έτσι, προκύπτουν εφαρμογές όπως συστήματα πλοήγησης οχημάτων, οι οποίες λειτουργούν με βάση τον χάρτη της περιοχής που βρίσκεται ο εκάστοτε χρήστης. Το συγκεκριμένο πλεονέκτημα γίνεται εκμεταλλεύσιμο και από άλλους τομείς, όπως την διαφήμιση και την ανάπτυξη ηλεκτρονικών παιχνιδιών. Ειδικότερα στον τομέα της διαφήμισης, συχνά συνδυάζονται η τοποθεσία και οι προτιμήσεις του χρήστη, προκειμένου να προβληθούν τα κατάλληλα εξατομικευμένα αποτελέσματα.

Εικόνα 1: Υπηρεσίες τοποθεσίας

2 wikipedia.org/wiki/Global_Positioning_System

Οι έξυπνες κινητές συσκευές όπως τα έξυπνα τηλέφωνα, προσφέρουν στους χρήστες την δυνατότητα να ψάχνουν για σημεία ενδιαφέροντος, τα οποία βρίσκονται σε κοντινή απόσταση και πληρούν κάποιες προϋποθέσεις, με βάση τα κριτήρια που επιλέγει ο χρήστης. Τα σημεία αυτά μπορούν να είναι επιχειρήσεις οι οποίες προσφέρουν προϊόντα ή υπηρεσίες. Τέτοιου είδους σημεία ενδιαφέροντος είναι τα εστιατόρια, τα κομμωτήρια και τα καταστήματα με είδη ένδυσης. Για παράδειγμα, ένα εστιατόριο Α, μπορεί να γνωστοποιεί την τοποθεσία του και τον κατάλογό με τα φαγητά που προσφέρει. Έτσι, όταν κάποιος χρήστης βρεθεί κοντά στο συγκεκριμένο εστιατόριο και πραγματοποιήσει αναζήτηση για πίτσα, το εστιατόριο Α θα εμφανιστεί σαν αποτέλεσμα, αν στον κατάλογό του εμπεριέχεται η λέξη πίτσα. Αντίστοιχα σε αυτήν την περίπτωση θα εμφανιστούν και άλλα αποτελέσματα, τα οποία αναπαριστούν εστιατόρια τα οποία προσφέρουν πίτσα.

1.2 Αντικείμενο διπλωματικής

Η παρούσα διπλωματική εργασία έχει ξεκάθαρο τεχνολογικό πλαίσιο με σκοπό την κατάταξη χωρο-κειμενικών δεδομένων μεγάλης κλίμακας. Με τον όρο χωρο-κειμενικά δεδομένα αναφερόμαστε σε δεδομένα τα οποία περιγράφονται από γεωγραφικές συντεταγμένες (X,Y), καθώς και από ένα σύνολο από λέξεις-κλειδιά. Πιο συγκεκριμένα, έχουμε μια βάση δεδομένων η οποία περιέχει τις τοποθεσίες κάποιων σημείων ενδιαφέροντος, δηλαδή το γεωγραφικό πλάτος και μήκος ή αλλιώς x και y. Επίσης, για κάθε σημείο ενδιαφέροντος βρίσκονται αποθηκευμένες και κάποιες λέξεις-κλειδιά, οι οποίες περιγράφουν την ιδιότητα της κάθε εγγραφής. Στον πίνακα 1 ακολουθούν ενδεικτικά παραδείγματα που προσομοιώνουν τρεις εγγραφές οι οποίες αναφέρονται σε εστιατόρια:

ID	***Όνομα***	***Βαθμολογία***	***Γεωγραφικό πλάτος***	***Γεωγραφικό μήκος***	***Λέξεις κλειδιά***
1456	Carl's Jr	1	36.807414	-119.884527	American, Burgers, Fast Food
1457	Picnic Garden	4	40.765149	-73.818978	Barbecue, Japanese, Korean
1458	Star of India	3.5	42.460714	-83.136283	Indian, Pakistani

Πίνακας 1: Εγγραφές εστιατορίων

Εκτός από το σύνολο των σημείων ενδιαφέροντος, θα γίνεται καταγραφή των ερωτήσεων που πραγματοποιούν οι χρήστες σε ένα ξεχωριστό αρχείο που ονομάζεται αρχείο καταγραφής ερωτήσεων (query log). Έτσι, κάθε φορά που κάποιος χρήστης θα κάνει χρήση του LBS συστήματος, θα καταγράφεται η θέση του, καθώς και οι λέξεις κλειδιά που χρησιμοποιήθηκαν για την αναζήτηση.

Απώτερος σκοπός της διπλωματικής, είναι η κατάταξη των σημείων ενδιαφέροντος με βάση τις προτιμήσεις των χρηστών, οι οποίες εκφράζονται με τις παραμέτρους που έδωσαν στις ερωτήσεις. Για πρακτικούς λόγους, τα αρχεία καταγραφής ερωτήσεων θα αποτελούνται από συνθετικά δεδομένα, τα οποία θα δημιουργηθούν με τυχαίο τρόπο.

1.3 Δομή της διπλωματικής

Η παρούσα διπλωματική εργασία αποτελείτε από 7 κεφάλαια. Στο κεφάλαιο 2 γίνεται μία παρουσίαση του προβλήματος που μελετάμε, ενώ εμπεριέχεται ένα αναλυτικό παράδειγμα για την κατανόησή του. Στο κεφάλαιο 3 γίνεται μία επεξήγηση για τα εργαλεία που επιλέχθηκαν για την λύση του προβλήματος. Στο κεφάλαιο 4 παρουσιάζονται κάποιες από τις σημαντικότερες προσεγγίσεις που πραγματεύονται παρόμοια προβλήματα. Έπειτα ακολουθεί το κεφάλαιο 5, στο οποίο αναλύεται ο τρόπος με τον οποίο προσεγγίστηκε το πρόβλημα κατά την εκπόνηση του πρακτικού κομματιού της παρούσας εργασίας. Στο κεφάλαιο 6 παρουσιάζονται τα αποτελέσματα της εκτέλεσης του προγράμματος που δημιουργήθηκε, με χρήση ενδεικτικών παραδειγμάτων. Ακολουθεί το κεφάλαιο 7 στο οποίο αναφέρονται τα σημαντικότερα συμπεράσματα, ενώ στο τέλος αναγράφεται η βιβλιογραφία που χρησιμοποιήθηκε κατά την εκπόνηση της παρούσας διπλωματικής εργασίας.

2 Ορισμός Προβλήματος

Στην συγκεκριμένη ενότητα θα αναλυθεί περαιτέρω το πρόβλημα που καλείτε να λύσει η παρούσα διπλωματική. Στην εποχή μας πολλές εφαρμογές απαιτούν την εύρεση σημείων ενδιαφέροντος, ή αντικειμένων, τα οποία βρίσκονται σε μία συγκεκριμένη περιοχή. Τέτοιες εφαρμογές μπορούν να θεωρηθούν οι εφαρμογές online τηλεφωνικών καταλόγων, εφαρμογές για delivery φαγητού και τα λοιπά. Οι εφαρμογές αυτές, λειτουργούν με βάση το πρόβλημα της εύρεσης του κοντινότερου γείτονα σε χωρο-κειμενικά δεδομένα, το οποίο είναι γνωστό στην επιστήμη των υπολογιστών (Felipe et al, 2008).

2.1 Αναζήτηση χωρο-κειμενικών δεδομένων

Με την πρόοδο στις τεχνολογίες γεω-εντοπισμού, υπάρχει ραγδαία αυξανόμενη ποσότητα χωρο-κειμενικών αντικειμένων που συλλέγονται από πολλές εφαρμογές, όπως υπηρεσίες βασισμένες σε τοποθεσίες (LBS) και κοινωνικά δίκτυα, στις οποίες ένα αντικείμενο περιγράφεται από την θέση του στο χώρο και από ένα σύνολο λέξεων κλειδιών. Για παράδειγμα, ένας διαδικτυακός κατάλογος επιχειρήσεων, παρέχει πληροφορίες τοποθεσίας για κάθε επιχείρηση, καθώς και σύντομες περιγραφές αυτών (π.χ. ξενοδοχεία, εστιατόρια). Όπως είναι φυσικό λοιπόν, προκύπτει η ανάγκη της ανεύρεσης συγκεκριμένων αντικειμένων, που εντοπίζονται σε βάσεις δεδομένων όπως αυτές που μόλις περιγράφηκαν.

Εφαρμογές όπως αυτές που αναφέρθηκαν παραπάνω, απαιτούν τουλάχιστον δύο ειδών πληροφορίες, οι οποίες ουσιαστικά τίθενται ως ερώτηση από τον χρήστη. Για παράδειγμα, οι εφαρμογές online τηλεφωνικών καταλόγων, συνήθως απαιτούν μία διεύθυνση η οποία καθορίζει την περιοχή ή τον χώρο στον οποίο θα πραγματοποιηθεί η αναζήτηση, καθώς και μία ή παραπάνω λέξεις, οι οποίες περιγράφουν την ιδιότητα του αντικειμένου της αναζήτησης. Έπειτα, ο χρήστης λαμβάνει σαν απάντηση αποτελέσματα τα οποία είναι ταξινομημένα με βάση την απόστασή τους από την αρχική διεύθυνση και περιγράφονται από τις λέξεις-κλειδιά (Felipe et al, 2008). Ως άλλο παράδειγμα, οι ιστότοποι ακινήτων επιτρέπουν στους χρήστες να αναζητούν ακίνητα με συγκεκριμένες λέξεις-κλειδιά στην περιγραφή τους και να τα ταξινομούν ανάλογα με την

απόσταση τους από μια καθορισμένη τοποθεσία. Τέτοιου είδους ερωτήματα καλούνται χωρικές ερωτήσεις με λέξεις κλειδιά (spatial keyword queries).

Με τον όρο αναζήτηση χωρο-κειμενικών δεδομένων ή αλλιώς spatial keyword search (χωρικές ερωτήσεις με λέξεις κλειδιά), αναφερόμαστε στην ανάκτηση δεδομένων με βάση την τοποθεσία κάποιων χρηστών, σε συνδιασμό με κάποιες λέξεις. Τα δεδομένα χαρακτηρίζονται από την τοποθεσία τους και από μία λίστα λέξεων τα οποία χρησιμοποιούνται κατά την ανάκτηση. Πιο συγκεκριμένα, η τοποθεσία συνήθως αναπαρίσταται με συντεταγμένες, οι οποίες δηλώνουν την ακριβή θέση κάποιου σημείου ενδιαφέροντος ή την τοποθεσία κάποιου χρήστη της εφαρμογής. Πολλές φορές βέβαια, ο χρήστης μπορεί να ορίσει την τοποθεσία γύρω από την οποία θέλει να πραγματοποιήσει αναζήτηση. Για παράδειγμα, εάν κάποιος κάνει χρήση της εφαρμογής google maps και πραγματοποιήσει αναζήτηση για την λέξη « καφές », μπορεί να πλοηγηθεί στον χάρτη και να δει τα αποτελέσματα σε οποιαδήποτε περιοχή επιλέξει. Μία τέτοια εφαρμογή είναι το tripadvisor, η οποία λειτουργεί με αντίστοιχο τρόπο, παρέχοντας επιπλέον λειτουργίες, όπως αξιολογήσεις και σχόλια χρηστών οι οποίες βοηθούν τον χρήστη να επιλέξει το κατάλληλο σημείο ενδιαφέροντος.

Πιο συγκεκριμένα, στο πλαίσιο της παρούσας διπλωματικής εργασίας, καλούμαστε να απαντήσουμε στο εξής ερώτημα: « Ένα ερώτημα Q (query) αναζητάει στον χώρο για τα k κοντινότερα αντικείμενα στο σημείο p (σημείο που βρίσκεται ο χρήστης). Θέτοντας το παραπάνω ερώτημα, αναζητούμε τα k καλύτερα αντικείμενα με βάση την κοντινότερη απόστασή τους από την τοποθεσία του χρήστη, και θα αναφερόμαστε σε αυτά σαν top-k. Τα αντικείμενα που πληρούν τις προϋποθέσεις, ταξινομούνται με βάση την απόσταση, έτσι ώστε αυτά που βρίσκονται πιο κοντά στο σημείο p, να κατατάσσονται υψηλότερα στην ταξινομημένη λίστα. Έτσι έχουμε: score(T) = distance(Tp, p). ». Στην περίπτωσή μας, οι ταξινομημένες απαντήσεις θα πρέπει να περιγράφονται από τις λέξεις κλειδιά (wl, ...wm) που ενδιαφέρουν τον χρήστη.

2.2 Παράδειγμα ανάκτησης δεδομένων

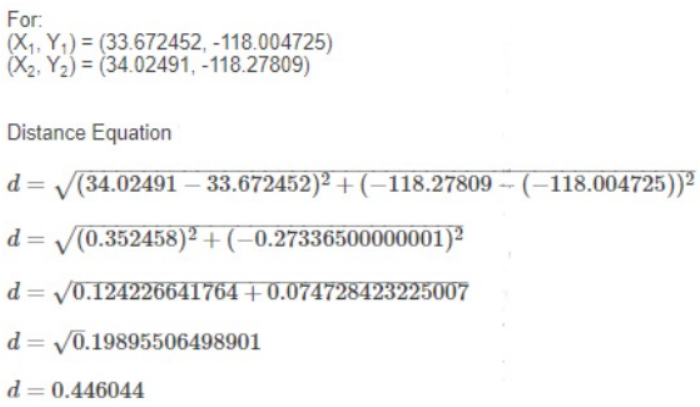
Προκειμένου να γίνει καλύτερα αντιληπτό το πρόβλημα που καλούμαστε να λύσουμε, κρίνεται σκόπιμη η παράθεση συγκεκριμένων παραδειγμάτων, τα οποία καλούνται να ορίσουν το πρόβλημα αλλά και την επιθυμητή λύση. Για τον λόγο αυτό δημιουργήθηκε ο πίνακας 2, ο οποίος εμπεριέχει πραγματικά δεδομένα (factual.com). Στον συγκεκριμένο πίνακα καταγράφονται δεδομένα, τα οποία αφορούν 25 εστιατόρια που βρίσκονται σε μία περιοχή. Η πρώτη στήλη (ID) αναπαριστά το αναγνωριστικό της κάθε εγγραφής, και δημιουργήθηκε για να διασφαλίσει την μοναδικότητα της. Η δεύτερη στήλη αφορά το όνομα του κάθε εστιατορίου, το οποίο θα μπορούσαν να μοιράζονται διαφορετικές επιχειρήσεις (πχ. McDonald's), ενώ η τρίτη και η τέταρτη στήλη αναπαριστούν τις συντεταγμένες στις οποίες βρίσκετε και λειτουργεί η κάθε επιχείρηση. Τέλος η πέμπτη στήλη περιέχει τις λέξεις-κλειδιά, οι οποίες στην περίπτωση που μελετάμε περιγράφουν τον τύπο του φαγητού που σερβίρεται σε κάθε εστιατόριο.

ID	*Όνομα*	*Γεωγραφικό πλάτος*	*Γεωγραφικό μήκος*	*Λέξεις κλειδια*
E1	El Adobe	33.499466	-117.6625	Mexican, American, Caribbean, Latin American
E2	Jack's Restaurant	34.864594	-120.44726	Bakery, American, Burgers, Chicken, Fast Food, Salad, Steak
E3	Yoshino River Park	36.849474	-119.784166666	Sushi, Japanese
E4	Comfort Cafe	34.015477	-118.49252646	American, Healthy, Sandwiches, Burgers, Ice Cream, Juices, Seafood, Smoothies, Soup, Steak
E5	Hinano Café	33.979293	-118.466451	Burgers, Pub Food, Sandwiches
E6	Palms	34.397338	-119.520776	Californian, Seafood, American, Chicken, Steak
E7	Fuzhou Super Buffet	38.09562	-122.562614	Chinese, Buffet
E8	Chiaramonte's Deli and Sausages	37.353099	-121.886424	Deli, Catering, Sandwiches
E9	Philippe The Original	34.059767	-118.23675	Sandwiches, American, Salad, Deli, French, Soup
E10	Barstow Station Liquor	34.891606	-117.000213	Fast Food
E11	Luigi's	35.374537	-118.993195	Italian, Deli, American
E12	McDonald's	37.781337	-122.420093	American, Burgers, Fast Food
E13	Swan Oyster Depot	37.79072	-122.42071	Seafood, American, Traditional
E14	Clearman's Galley	34.128218	-118.073139	American, Seafood
E15	Balboa Cafe	37.798894	-122.43589	American, Burgers, Californian, Mediterranean
E16	Sam's for Play Cafe and Catering On Cleveland Ave	38.462831	-122.728381	Catering, Diner

E17	Farmer Boys Restaurant	33.832059	-117.985020875	American, Burgers, Fast Food
E18	The Farm Of Beverly Hills	34.072022	-118.35754	American, Californian, Coffee, Sandwiches, Tea
E19	St. Francis Fountain	37.75285167	-122.40827	American, Diner, Coffee, Deli, Ice Cream, Frozen Yogurt, Sandwiches
E20	A G Ferrari Foods	37.368496	-122.03496	Deli, Catering, Italian, Sandwiches
E21	A&W Restaurant	36.302863	-119.135775	American, Burgers, Hot Dogs, Chicken, Fast Food, Pub Food, Sandwiches
E22	Camilo's Fine Cuisine and Catering	34.139198	-118.213025	Californian, French, Contemporary, Latin
E23	Musso and Frank Grill	34.101572	-118.335352	American, Steak, Continental, Seafood, Barbecue, Deli, European, Steakhouse
E24	Barney's Beanery	34.0905	-118.374115	American, Traditional, Diner, Bagels, Barbecue, Burgers, Deli, Donuts, Pizza,Sandwiches
E25	Bistango Cafe	34.02491	-118.27809	Italian, American, Pizza, Sandwiches, Pasta, Salad

Πίνακας 2: Δείγμα βάσης δεδομένων για 25 εγγραφές

Έχοντας υπόψιν τα παραπάνω, θέλουμε να θέσουμε μία ερώτηση, και με βάση αυτήν να λάβουμε ως αποτέλεσμα τα 3 βέλτιστα από τα 25 συνολικά εστιατόρια. Η ερώτηση που θα θέσουμε ως πρώτο παράδειγμα (παράδειγμα 1) είναι η εξής: « Ένα top-3 χωρικό ερώτημα, αναζητά στον χώρο τα 3 κοντινότερα αντικείμενα με βάση το σημείο p1:(33.672452 , -118.004725) ». Επίσης, θα πρέπει να θέσουμε και τουλάχιστον μία λέξη ως λέξη κλειδί. Στην περίπτωση μας υποθέτουμε ότι η λέξη κλειδί είναι η λέξη: « American ».



Εικόνα 2: Ευκλείδεια απόσταση

Λαμβάνοντας υπόψιν την λέξη κλειδί, πρέπει πλέον να κατατάξουμε τα αποτελέσματα με βάση το πόσο κοντά βρίσκονται στο σημείο p. Η αναζήτησή μας θα γίνει σε 17 από τα συνολικά 25 εστιατόρια, καθώς τα 8 από αυτά δεν περιέχουν την λέξη « American » στην περιγραφή τους. Για να καταλήξουμε στα 3 βέλτιστα αποτελέσματα πρέπει να υπολογίσουμε την απόσταση του σημείου d από κάθε ένα από τα 17 εναπομείναντα σημεία ενδιαφέροντος. Ο συγκεκριμένος υπολογισμός γίνεται με βάση τον τύπο της ευκλείδειας απόστασης[3], ο οποίος εφαρμόζεται ενδεικτικά για το σημείο E1 στην εικόνα 2. Καταλήγουμε λοιπόν σε 3 βέλτιστα αποτελέσματα τα οποία είναι κατά σειρά τα E17, E1 και E25. Ο πίνακας 3 περιλαμβάνει τις top-3 εγγραφές οι οποίες βρίσκονται κοντά στο στο σημείο p1, με σειρά από το πιο κοντινό στο πιο μακρινό σημείο. Να σημειωθεί ότι κατά την υλοποίηση, για τον υπολογισμό της απόστασης μεταξύ των σημείων, εκτός από τον τύπο της ευκλείδειας απόστασης, λαμβάνεται υπ όψιν και η καμπυλότητα της γης, κάτι που παραλείπεται από το συγκεκριμένο παράδειγμα για λόγους απλότητας, ενώ το αποτέλεσμα της εφαρμογής του τύπου μετατρέπεται σε μέτρα.

ID	***Όνομα***	***Γεωγραφικό πλάτος***	***Γεωγραφικό μήκος***	***Λέξεις κλειδιά***	***Απόσταση από p1***
E17	Farmer Boys Restaurant	33.832059	-117.985020875	American, Burgers, Fast Food	17.8km
E1	El Adobe	33.499466	-117.6625	Mexican, American, Caribbean, Latin, American	37km
E25	Bistango Cafe	34.02491	-118.27809	Italian, American, Pizza, Sandwiches, Pasta, Salad	46.6km

Πίνακας 3: Top-3 αποτελέσματα για p1 και «American»

Ας υποθέσουμε τώρα μία άλλη περίπτωση, (παράδειγμα 2) κατά την οποία ο χρήστης βρίσκεται και πάλι στο σημείο p1: (33.672452 , -118.004725), και πραγματοποιεί μια αναζήτηση στο ίδιο

3 el.wikipedia.org

σύνολο δεδομένων. Η δομή της παρούσας ερώτησης λοιπόν θα έχει την εξής δομή: « Ένα top-3 ερώτημα, αναζητά τα 3 κοντινότερα αντικείμενα με βάση το σημείο p1: (33.672452 , -118.004725) » . Αυτή τη φορά η αναζήτηση θα πραγματοποιείται με βάση την λέξη κλειδί : « Deli » , οπότε από τις 25 εγγραφές, θα χρειαστεί να γίνει υπολογισμός της απόστασης μόνο για τις 7, καθώς οι υπόλοιπες 18 δεν περιέχουν την συγκεκριμένη λέξη-κλειδί. Η απάντηση στην παραπάνω ερώτηση παρουσιάζεται στον πίνακα 4.

ID	***Όνομα***	***Γεωγραφικό πλάτος***	***Γεωγραφικό μήκος***	***Λέξεις κλειδιά***	***Απόσταση από p1***
E9	Philippe The Original	34.059767	-118.23675	Sandwiches, American, Salad, Deli, French, Soup	48.1km
E23	Musso and Frank Grill	34.101572	-118.335352	American, Steak, Continental, Seafood, Barbecue, Deli, European, Steakhouse	56.6km
E24	Barney's Beanery	34.0905	-118.374115	American, Traditional, Diner, Bagels, Barbecue, Burgers, Deli, Donuts, Pizza,Sandwiches	57.6km

Πίνακας 4: Top-3 αποτελέσματα για p1 και «Deli»

Συνεχίζοντας με τον ορισμό του προβλήματος, ορίζουμε ένα νέο σημείο στο οποίο θα βρίσκετε ο χρήστης. Το σημείο αυτό το ονομάζουμε p2 και του δίνουμε μία τυχαία τιμή : (32.870138,-120.084675). Αυτή τη φορά (παράδειγμα 3) υποθέτουμε ότι η λέξη-κλειδί είναι η λέξη « Pizza » . Παρατηρούμε ότι οι εγγραφές που έχουν ως λέξη κλειδί την συγκεκριμένη λέξη είναι 2 από τις 25. Επομένως, παρόλο που η ερώτηση ζητάει τρεις εγγραφές, οι απαντήσεις θα είναι 2. Στον πίνακα 5 παρουσιάζονται τα αποτελέσματα στην ερώτηση. Όπως φαίνεται, η απόσταση του βέλτιστου αποτελέσματος είναι αρκετά μεγαλύτερη σε σύγκριση με τα προηγούμενα παραδείγματα. Αυτό συμβαίνει γιατί τα υπόλοιπα σημεία ενδιαφέροντος που βρίσκονται πιο κοντά στο p2, δεν περιγράφονται από την λέξη κλειδί με βάση την οποία γίνεται η αναζήτηση.

ID	***Όνομα***	***Γεωγραφικό πλάτος***	***Γεωγραφικό μήκος***	***Λέξεις κλειδιά***	***Απόσταση από p2***
E24	Barney's Beanery	34.0905	-118.374115	American, Traditional, Diner, Bagels, Barbecue, Burgers, Deli, Donuts, Pizza,Sandwiches	208.7km
E25	Bistango Cafe	34.02491	-118.27809	Italian, American, Pizza, Sandwiches, Pasta, Salad	211.1km

Πίνακας 5: Top-2 αποτελέσματα για p2 και «Pizza»

Μέχρι τώρα, έχουμε υποθέσει ότι κάθε φορά η αναζήτηση πραγματοποιείται με βάση μία λέξη κλειδί. Στην πραγματικότητα οι λέξεις κλειδιά που μπορούμε να χρησιμοποιήσουμε, είναι περισσότερες. Για παράδειγμα, μπορούμε να πραγματοποιήσουμε μία αναζήτηση με 2 λέξεις-κλειδιά. Έτσι, υποθέτουμε (παράδειγμα 4) ότι από το σημείο p2 γίνεται μία ερώτηση, η οποία αφορά τις λέξεις : « American » και « Burgers ». Από τα 25 σημεία ενδιαφέροντος, τα 7 περιέχουν και τις 2 λέξεις, οπότε τα υπόλοιπα 18 περιέχουν είτε μία από αυτές είτε καμία. Τα αποτελέσματα αναγράφονται στον πίνακα 6.

ID	***Όνομα***	***Γεωγραφικό πλάτος***	***Γεωγραφικό μήκος***	***Λέξεις κλειδια***	***Απόσταση από p2***
E4	Comfort Cafe	34.015477	-118.4925264	American, Healthy, Sandwiches, Burgers, Ice Cream, Juices, Seafood, Smoothies, Soup, Steak	195km
E24	Barney's Beanery	34.0905	-118.374115	American, Traditional, Diner, Bagels, Barbecue, Burgers, Deli, Donuts, Pizza,Sandwiches	208.7km
E17	Farmer Boys Restaurant	33.832059	-117.985020875	American, Burgers, Fast Food	222.4km

Πίνακας 6: Top-3 αποτελέσματα για p2 και «American ,Burgers»

Στην τελευταία περίπτωση που θα μελετήσουμε, θα βρούμε με τυχαίο τρόπο ένα νέο σημείο p3 το οποίο θα αναπαριστά την τοποθεσία που βρίσκεται ο χρήστης. Έστω λοιπόν (παράδειγμα 6) p3 : (30.862141, -117.501512) και λέξεις « Smoothies, Seafood, Juices, Soup ». Στην συγκεκριμένη περίπτωση, μόνο 1 από τα 25 σημεία εμπεριέχει και τις 4 λέξεις. Οπότε η αναζήτηση θα επιστρέφει μόνο μία απάντηση, η οποία φαίνεται στον πίνακα 7. Παρατηρούμε ότι το αποτέλεσμα που εμφανίζεται, βρίσκεται αρκετά μακριά από το σημείο p3.

ID	***Όνομα***	***Γεωγραφικό πλάτος***	***Γεωγραφικό μήκος***	***Λέξεις κλειδιά***	***Απόσταση από p2***
E4	Comfort Cafe	34.015477	-118.4925264	American, Healthy, Sandwiches, Burgers, Ice Cream, Juices, Seafood, Smoothies, Soup, Steak	362.7km

Πίνακας 7: Top-1 αποτέλεσμα για p3 και «Smoothies, Seafood, Juices, Soup»

Μελετώντας τα παραπάνω παραδείγματα, παρατηρούμε ότι τα αποτελέσματα βρίσκονται αρκετά μακριά από τον χρήστη. Αυτό συμβαίνει διότι η βάση δεδομένων που χρησιμοποιείτε για τον σκοπό του παραδείγματος είναι πολύ μικρή σε σύγκριση με την πραγματική, με αποτέλεσμα τα

εστιατόρια να απέχουν αρκετά μεταξύ τους. Επίσης όπως έχει ήδη αναφερθεί, για κάθε ερώτηση που πραγματοποιεί ο χρήστης ορίζει και μία ακτίνα γύρω από την τοποθεσία του, μέσα στην οποία θα γίνεται η αναζήτηση. Έτσι, κάθε φορά που κάποιο σημείο ενδιαφέροντος βρίσκεται πιο μακριά από την ακτίνα αυτή, δεν επιστρέφεται σαν αποτέλεσμα από τον αλγόριθμο, ακόμη και αν δεν υπάρχουν σημεία εντός της συγκεκριμένης ακτίνας. Στο συγκεκριμένο παράδειγμα, θεωρούμε ότι η ακτίνα αυτή έχει οριστεί στα 400 χιλιόμετρα για όλα τα ερωτήματα, ενώ κατά την υλοποίηση η ακτίνα ορίζεται ξεχωριστά για κάθε ερώτημα.

3 Βασικές έννοιες

3.1 Γράφοι και PageRank

3.1.1 Τι είναι Γράφος

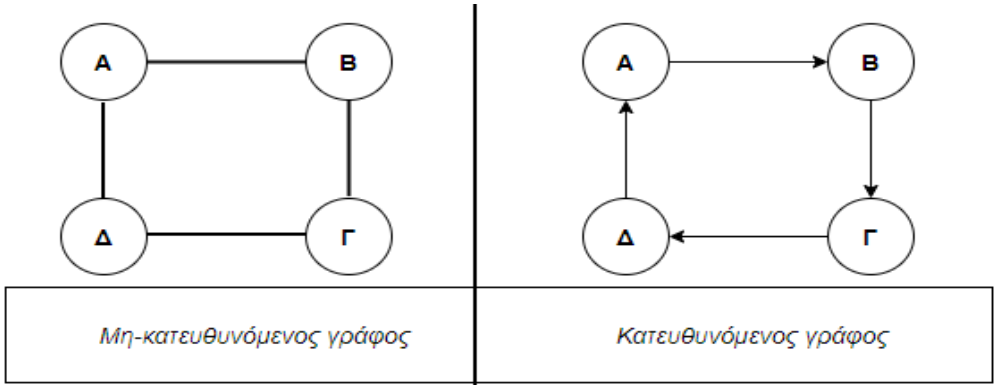
Στη θεωρία γραφημάτων, ένας γράφος ή γράφημα είναι μια δομή η οποία αντιστοιχεί σε ένα σύνολο αντικειμένων, στο οποίο ορισμένα ζεύγη συσχετίζονται. Τα αντικείμενα αναπαριστώνται συνήθως από σημεία που ονομάζονται κορυφές ή κόμβοι, ενώ οι γραμμές που εκφράζουν τις συσχετίσεις μεταξύ κόμβων ονομάζονται ακμές ή σύνδεσμοι. Με τον όρο γράφημα λοιπόν, αναφερόμαστε σε ένα σύνολο αποτελούμενο από ακμές και κόμβους. Συνήθως, κάθε κόμβος ενός γράφου αναπαριστά ένα μοναδικό αντικείμενο, και για λόγους ευκρίνειας του δίνουμε κάποιο όνομα, για παράδειγμα « κόμβος Α ». Αντίστοιχα οι ακμές λαμβάνουν συνήθως κάποιες ετικέτες οι οποίες αποτελούνται από λέξεις που εκφράζουν κάτι για την συσχέτιση, ενώ σε ορισμένες περιπτώσεις λαμβάνουν και έναν αριθμό ο οποίος εκφράζει κάποιο «κόστος». Τα γραφήματα αποτελούν ένα από τα αντικείμενα μελέτης των διακριτών μαθηματικών[4].

Η μελέτη του Leonhard Euler που δημοσιεύθηκε το 1736 θεωρείται το πρώτο έγγραφο στην ιστορία της θεωρίας γραφημάτων[5]. Η λέξη γράφημα ή γράφος χρησιμοποιήθηκε για πρώτη φορά με την έννοια που μελετάμε από τον James Joseph Sylveste το 1878. Στην επιστήμη των υπολογιστών, τα γραφήματα χρησιμοποιούνται για την αναπαράσταση δικτύων επικοινωνίας, για την οργάνωση δεδομένων κ.λπ. Για παράδειγμα, η δομή των συνδέσμων ενός ιστότοπου μπορεί να αναπαρασταθεί από ένα κατευθυνόμενο γράφημα, στο οποίο οι κόμβοι αντιπροσωπεύουν τις ιστοσελίδες και οι ακμές αντιπροσωπεύουν συνδέσμους από τη μία σελίδα στην άλλη. Η ανάπτυξη αλγορίθμων για τη διαχείριση γραφημάτων είναι μείζονος ενδιαφέροντος για την επιστήμη των υπολογιστών. Ο μετασχηματισμός των γραφημάτων συχνά τυποποιείται και αντιπροσωπεύεται από συστήματα επανεγγραφής γραφημάτων.

4 en.wikipedia.org/wiki/Graph_(discrete_mathematics)#cite_note-1

5 en.wikipedia.org/wiki/Graph_theory

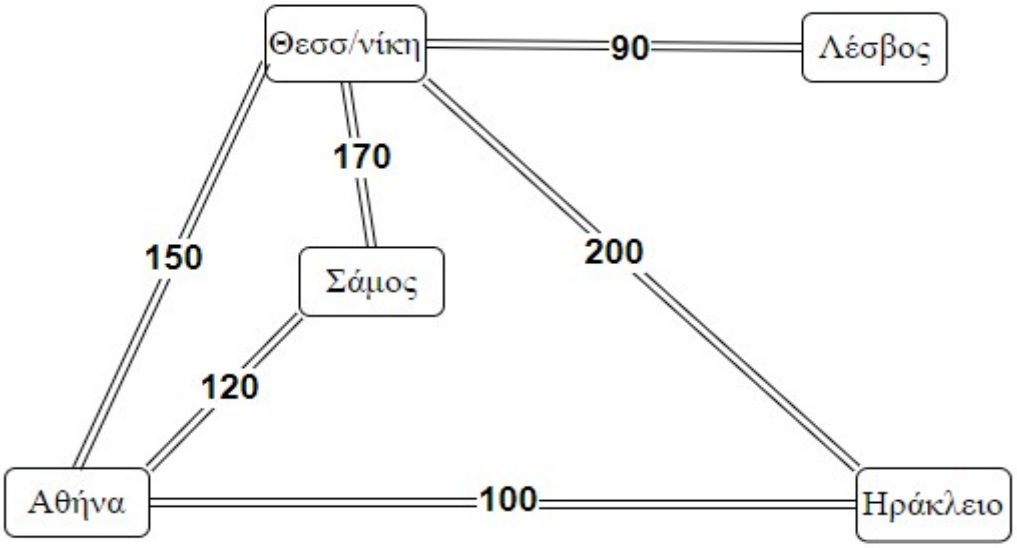
Οι γράφοι, χωρίζονται σε δύο μεγάλες κατηγορίες τους κατευθυνόμενους και τους μη κατευθυνόμενους. Ένας μη κατευθυνόμενος γράφος αποτελείτε από ένα σύνολο αντικειμένων - κόμβων που συνδέονται μεταξύ τους, μέσω αμφίδρομων ακμών, ενώ ένας γράφος όπου οι ακμές του δείχνουν προς μια κατεύθυνση ονομάζεται κατευθυνόμενος γράφος. Σε γενικές γραμμές, χρησιμοποιούμε ένα κατευθυνόμενο γράφημα, όταν οι κορυφές οι οποίες ενώνονται με μία ακμή συσχετίζονται άμεσα μεταξύ τους. Η συσχέτιση αυτή συνήθως εκφράζεται από την κορυφή που ξεκινάει η ακμή, προς την κορυφή στην οποία καταλήγει. Σε αντίθεση με τα κατευθυνόμενα γραφήματα, τα μη κατευθυνόμενα αναπαριστούν συνήθως περισσότερο αφηρημένα συστήματα. Στην εικόνα 3 αναπαρίστανται ένα κατευθυνόμενο και ένα μη κατευθυνόμενα γράφημα.



Εικόνα 3: Κατευθυνόμενος / Μη-κατευθυνόμενος γράφος

Συχνά θέλουμε να εκφράσουμε και κάποιο κόστος ή βάρος που πιθανόν να προκύψει για την μετάβαση από έναν κόμβο σε έναν άλλο. Σε αυτές τις περιπτώσεις χαρακτηρίζουμε τον γράφο ως βεβαρημένο, ενώ στην αντίθετη περίπτωση ως μη-βεβαρημένο. Έστω λοιπόν ότι θέλουμε να εκφράσουμε σε έναν γράφο τα δρομολόγια ενός αεροπλάνου, ενώ ταυτόχρονα θέλουμε να αποτυπώσουμε και το κόστος των καυσίμων που θα χρειαστούν. Για την δημιουργία ενός τέτοιου γράφου, πρέπει αρχικά να ορίσουμε του κόμβους, οι οποίοι θα αναπαριστούν κάποιες περιοχές της Ελλάδας: Αθήνα, Θεσσαλονίκη, Σάμος, Ηράκλειο, Λέσβος. Έπειτα, προχωράμε με την δημιουργία των ακμών, η οποία θα εκφράζει την σύνδεση μεταξύ των αεροδρομίων της περιοχής. Στην περίπτωση που μελετάμε, ο γράφος δεν χρειάζεται να είναι κατευθυνόμενος, ενώ κατά την δημιουργία κάθε ακμής προσθέτουμε και το κόστος των καυσίμων που απαιτούνται για κάθε δρομολόγιο. Έτσι, προκύπτει ο γράφος της εικόνας 4, ο οποίος είναι ένας Μη-κατευθυνόμενος βεβαρημένος γράφος .

Σύμφωνα με τον γράφο της εικόνας 4, εάν θέλουμε να εκτελέσουμε ένα δρομολόγια από Αθήνα προς Θεσσαλονίκη ή από Θεσσαλονίκη προς Αθήνα, θα έχουμε βάρος 200. Αντίστοιχα, εάν θέλουμε να μεταβούμε από Σάμο προς Ηράκλειο, θα πρέπει πρώτα να μεταβούμε στην Αθήνα με βάρος 120, και έπειτα από Αθήνα προς Ηράκλειο με βάρος 100. Επομένως, το συνολικό βάρος της μετάβασης από Σάμο προς Ηράκλειο είναι 220. Εναλλακτικά, θα μπορούσαμε να μεταβούμε από Σάμο προς Θεσσαλονίκη και έπειτα προς Ηράκλειο, με συνολικό κόστος 370. Συνήθως αναζητούμε το μονοπάτι με το μικρότερο συνολικά βάρος, και το ονομάζουμε βέλτιστο.



Εικόνα 4 : Μη-κατευθυνόμενος βεβαρημένος γράφος

3.1.2 Τι είναι ο Pagerank

Οι Larry Page και Sergey Brin ανέπτυξαν τον συγκεκριμένο αλγόριθμο στο Πανεπιστήμιο του Στάνφορντ το 1996, ως μέρος ενός ερευνητικού έργου σχετικά με ένα νέο είδος μηχανής αναζήτησης. Η βασική ιδέα για την κατασκευή του αλγορίθμου, είναι ότι οι πληροφορίες στο διαδίκτυο θα μπορούσαν να ταξινομηθούν ιεραρχικά με βάση την δημοτικότητα κάθε συνδέσμου. Θα έπρεπε λοιπόν μια σελίδα να κατατάσσεται υψηλότερα εφόσον υπάρχουν περισσότεροι σύνδεσμοι προς αυτήν. Το σύστημα αναπτύχθηκε με τη βοήθεια των Scott Hassan και Alan Steremberg, οι οποίοι έπαιξαν καθοριστικό ρόλο στην ανάπτυξη της μηχανής αναζήτησης: « Google ». Οι Rajeev Motwani και Terry Winograd συνέγραψαν με τους Page και Brin το πρώτο άρθρο σχετικά με τον αλγόριθμο Pagerank, που δημοσιεύθηκε το 1998. Λίγο μετά, οι Page και Brin ίδρυσαν την Google Inc., την εταιρεία πίσω από το Μηχανή αναζήτησης Google. Αν και πλέον υπάρχουν πολλοί παράγοντες που καθορίζουν την κατάταξη των αποτελεσμάτων αναζήτησης, το Pagerank συνεχίζει να παρέχει τη βάση για όλα τα εργαλεία αναζήτησης ιστού της Google[6].

Με τον όρο Pagerank αναφερόμαστε σε έναν αλγόριθμο ο οποίος αφορά μια κατανομή πιθανοτήτων, που χρησιμοποιείται για να εκπροσωπήσει την πιθανότητα ότι ένα άτομο κάνοντας

6 en.wikipedia.org/wiki/PageRank#History

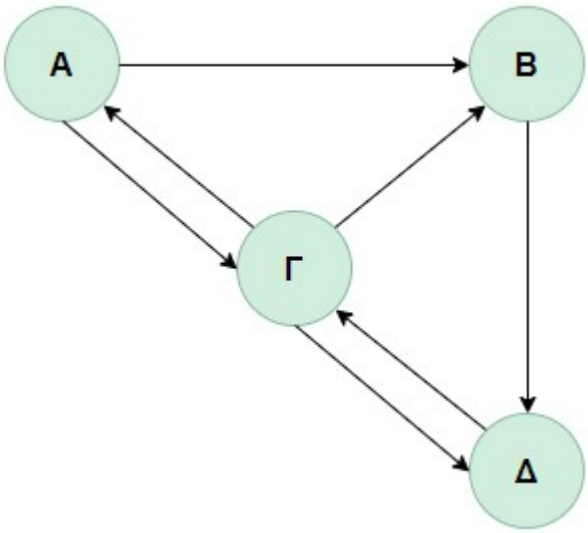
τυχαία κλικ σε συνδέσμους, θα καταλήξει σε κάποια συγκεκριμένη σελίδα. Ο υπολογισμός του PageRank απαιτεί πολλές προσπελάσεις της συλλογής, ώστε να υπολογίζεται με μεγαλύτερη ακρίβεια η βαθμολογία. Η πιθανότητα εκφράζεται ως αριθμητική τιμή μεταξύ 0 και 1. Για παράδειγμα, η πιθανότητα 0,3 εκφράζεται πιθανότητα 30% να συμβεί ένα ενδεχόμενο. Ως εκ τούτου, για ένα έγγραφο Α με Pagerank 0,3 σημαίνει ότι υπάρχει πιθανότητα 30% ένα άτομο που κάνει κλικ σε έναν τυχαίο σύνδεσμο να κατευθυνθεί προς το εν λόγω έγγραφο Α.

Προκειμένου να γίνει κατανοητή η λειτουργία του Pagerank, θα μελετηθεί αρχικά μία ελαφρώς απλοποιημένη εκδοχή R, η οποία λειτουργεί με βάση τον τύπο 1 (Page et al,1999):

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Τύπος 1

Ας υποθέσουμε λοιπόν ότι έχουμε μία ιστοσελίδα u. Στη συνέχεια, ορίζουμε ότι Fu είναι το σύνολο των σελίδων στις οποίες δείχνει η u, και Bu είναι το σύνολο των σελίδων που δείχνουν στην u. Επίσης ορίζουμε ότι Nu = |Fu|, δηλαδή ο αριθμός των συνδέσμων που εμπεριέχονται στην u, ενώ c είναι ένας παράγοντας που χρησιμοποιείται για την ομαλοποίηση, έτσι ώστε η συνολική κατάταξη όλων των ιστοσελίδων να είναι σταθερή. Συνοψίζοντας λοιπόν έχουμε ότι η τιμή Pagerank για μια σελίδα u εξαρτάται από τις τιμές Pagerank για κάθε σελίδα υ που περιέχεται στο σύνολο Bu διαιρούμενη με τον αριθμό των συνδέσμων που εμπεριέχονται στην u.



Εικόνα 5: Απλοποιημένο δίκτυο ιστοσελίδων

Στην εικόνα 5 απεικονίζεται ένα απλοποιημένο δίκτυο ιστοσελίδων. Κάθε κόμβος απεικονίζει και μία ιστοσελίδα, ενώ οι ακμές που συνδέουν τους κόμβους μεταξύ τους απεικονίζουν τους συνδέσμους που εμπεριέχονται από μία ιστοσελίδα προς μία άλλη. Για παράδειγμα, η ιστοσελίδα Β περιέχει έναν σύνδεσμο προς την ιστοσελίδα Δ, ενώ η ιστοσελίδα Α περιέχει έναν σύνδεσμο

προς την ιστοσελίδα Β και τα λοιπά. Θέλουμε λοιπόν να κατατάξουμε τις ιστοσελίδες Α, Β, Γ, και Δ , σύμφωνα με το Pagerank τους. Αρχικά θεωρούμε ότι κάποιος χρήστης έχει ίσες πιθανότητες να βρεθεί σε οποιαδήποτε από τις τέσσερις ιστοσελίδες, οπότε η αρχική πιθανότητα για κάθε ιστοσελίδα είναι 25%, ή αλλιώς 1/4.

Ιστοσελίδα	***Επανάληψη 0***	***Επανάληψη 1***	***Επανάληψη 2***	***PageRank***
Α	1/4	1/12	1.5/12	1
Β	1/4	2.5/12	2/12	2
Γ	1/4	4.5/12	4.5/12	4
Δ	1/4	4/12	4/12	3

Πίνακας 8: Διαδικασία Pagerank για απλοποιημένο δίκτυο

Για να υπολογίσουμε την πιθανότητα να βρεθεί κάποιος στην ιστοσελίδα Α έπειτα από την επιλογή ενός συνδέσμου ή αλλιώς από την πραγματοποίηση ενός «κλικ», πρέπει να παρατηρήσουμε και πάλι την εικόνα 5. Βλέπουμε ότι υπάρχει μόνο μία ιστοσελίδα η οποία περιέχει σύνδεσμο που οδηγεί στην Α, η οποία είναι η Γ. Επίσης, η ιστοσελίδα Γ περιέχει σύνδεσμο προς τρεις συνολικά ιστοσελίδες. Επομένως, για να βρούμε το Pagerank της ιστοσελίδας Α για την πρώτη επανάληψη πρέπει να διαιρέσουμε την πιθανότητα να βρεθεί κάποιος στην ιστοσελίδα Γ στην επανάληψη 0, (δηλαδή 1/4) με τον αριθμό των ιστοσελίδων που « δείχνει » η ιστοσελίδα Γ. Οπότε για την ιστοσελίδα Α στην Επανάληψη 1, έχουμε : PR= (1/4) / 3 = 1/12. Αντίστοιχα, για να βρούμε το Pagerank της Επανάληψης 1 για την ιστοσελίδα Β πρέπει και πάλι να παρατηρήσουμε το σχήμα της εικόνας 5. Βλέπουμε, ότι αυτή τη φορά υπάρχουν δύο ιστοσελίδες οι οποίες δείχνουν στην ιστοσελίδα Β, οι οποίες είναι οι Α και η Γ. Επομένως για την Β έχουμε: PR= ((1/4)/2 + (1/4)/3) = 2.5/12. Με αντίστοιχο τρόπο υπολογίζουμε το Pagerank για την Επανάληψη 1 και για τις υπόλοιπες δύο ιστοσελίδες (Πίνακας 8).

Για τον υπολογισμό, των Pagerank στην δεύτερη Επανάληψη, λειτουργούμε με τον ίδιο ακριβώς τρόπο όπως στην Επανάληψη 1. Έτσι, για την ιστοσελίδα Α, πρέπει να διαιρέσουμε την πιθανότητα να βρεθεί κάποιος στην ιστοσελίδα Γ στην επανάληψη 1, (δηλαδή 4.5/12) με τον αριθμό των ιστοσελίδων που « δείχνει » η ιστοσελίδα Γ, δηλαδή 3, άρα έχουμε: PR= (4.5/12) / 3 =1.5/12. Με την ίδια τακτική λοιπόν συμπληρώνουμε και την υπόλοιπη στήλη του πίνακα 7 που αφορά την Επανάληψη 2. Στην τελευταία στήλη του πίνακα 8 έχουμε την τελική βαθμολογία των ιστοσελίδων. Η ιστοσελίδα με την υψηλότερη τελική βαθμολογία είναι αυτή που θα εμφανιστεί πρώτη στον χρήστη, στην περίπτωσή μας η ιστοσελίδα Γ, ενώ ακολουθούν με την σειρά οι Δ, Β και Α.

Με βάση την παραπάνω ανάλυση συμπεραίνουμε ότι προκειμένου να έχει μία ιστοσελίδα Χ υψηλή βαθμολογία Pagerank, πρέπει να προτείνεται από πολλές ιστοσελίδες. Ωστόσο η ιστοσελίδα η οποία προτείνει την Χ θα πρέπει να μην προτείνει μεγάλο αριθμό ιστοσελίδων, καθώς και να έχει και η ίδια υψηλή βαθμολογία. Το παραπάνω γεγονός είναι αρκετά αποτρεπτικό

στην περίπτωση που κάποιος θέλει να « παραπλανήσει » τον αλγόριθμο, δημιουργώντας πολλές εικονικές ιστοσελίδες, προκειμένου να ανεβάσει την Pagerank βαθμολογία μίας πραγματικής.

3.1.3 Τι είναι ο Weighted Pagerank ή Personalized Pagerank

Με την ευρύτερη χρήση του διαδικτύου στο εμπόριο, την διδασκαλία και την ειδησεογραφία, η ανεύρεση των αναγκών των χρηστών και η παροχή χρήσιμων πληροφοριών είναι οι πρωταρχικοί στόχοι των κατόχων ιστοσελίδων. Έτσι, η ανάλυση της συμπεριφοράς των χρηστών γίνεται όλο και πιο σημαντική. Η τεχνική που χρησιμοποιείται για την ανακάλυψη του περιεχόμενου του διαδικτύου αλλά και τη συμπεριφορά των χρηστών σε αυτό, ονομάζεται Web Mining. Το Web Mining αποτελείται από το Web Content Mining (WCM), το Web Structure Mining (WSM), και το Web Usage Mining (WUM). Το Web Content Mining ασχολείται με την ανακάλυψη χρήσιμων πληροφοριών από το περιεχόμενο των ιστοσελίδων, το Web Structure Mining ανακαλύπτει σχέσεις μεταξύ ιστοσελίδων αναλύοντας δομές ιστού, ενώ το Web Usage Mining ελέγχει τα προφίλ των χρηστών και τη συμπεριφορά τους, που καταγράφονται στο αρχείο καταγραφής ιστού.

Με βάση την τοπολογία των υπερσυνδέσμων, το Web Structure Mining, κατηγοριοποιεί τις ιστοσελίδες, δημιουργώντας σχετικά μοτίβα βασισμένα στην ομοιότητα και τις σχέσεις μεταξύ διαφορετικών ιστοσελίδων. Τεχνικά, το WSM προσπαθεί να ανακαλύψει τη δομή των συνδέσεων των υπερσυνδέσμων μεταξύ ιστοσελίδων. Οι αριθμοί των συνδέσεων (σύνδεσμοι προς μια σελίδα) και των αποσυνδέσεων (σύνδεσμοι από μια σελίδα) είναι πολύτιμες πληροφορίες στο Web Mining.

Ο αλγόριθμος Pagerank, ένας από τους ευρύτερα χρησιμοποιούμενους αλγόριθμους κατάταξης δηλώνει ότι εάν μια σελίδα έχει σημαντικούς συνδέσμους προς αυτήν, οι σύνδεσμοι της προς άλλες σελίδες γίνονται επίσης σημαντικοί. Επομένως, μια σελίδα έχει υψηλή κατάταξη εάν το άθροισμα των βαθμών των πίσω συνδέσμων της είναι υψηλό. Στο Pagerank, η βαθμολογία που μπορεί να δώσει μια σελίδα p, σε άλλες σελίδες, κατανέμεται ομοιόμορφα μεταξύ των εξερχόμενων συνδέσμων της. Οι βαθμολογίες ενός ιστότοπου, θα μπορούσαν να υπολογιστούν επαναληπτικά ξεκινώντας από οποιαδήποτε ιστοσελίδα. Σε έναν ιστότοπο, δύο ή περισσότερες σελίδες ενδέχεται να συνδεθούν μεταξύ τους και να σχηματίσουν βρόχο. Εάν αυτές οι σελίδες δεν αναφέρουν, αλλά αναφέρονται από άλλες ιστοσελίδες εκτός βρόχου, θα συγκεντρώνουν βαθμό, αλλά δεν θα διανέμουν ποτέ βαθμό. Το σενάριο αυτό, ονομάζεται rank sink.

Για την επίλυση του προβλήματος του rank sink, ελήφθησαν υπόψιν οι δραστηριότητες των χρηστών. Παρατηρήθηκε λοιπόν ένα φαινόμενο, κατά το οποίο οι χρήστες δεν ακολουθούν πάντα τους υπάρχοντες συνδέσμους, αλλά για παράδειγμα, μετά την προβολή της σελίδας α, ορισμένοι χρήστες μεταβαίνουν απευθείας στη σελίδα β, η οποία δεν συνδέεται άμεσα με τη σελίδα α. Για το σκοπό αυτό, οι χρήστες απλώς πληκτρολογούν τη διεύθυνση URL της σελίδας β στο πεδίο

κειμένου URL και μεταβαίνουν απευθείας σε αυτήν. Στην συγκεκριμένη περίπτωση, η κατάταξη της σελίδας β θα πρέπει να επηρεάζεται από τη σελίδα α, παρόλο που οι δυο τους δεν είναι άμεσα συνδεδεμένες. Λαμβάνοντας υπόψιν το παρόν φαινόμενο ο τύπος του Pagerank διαμορφώνεται όπως ο τύπος 2.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v}$$

Τύπος 2

Σύμφωνα με τον τύπο 2, έχουμε όπου d ένας παράγοντας απόσβεσης που ορίζεται συνήθως σε 0,85. Θα μπορούσαμε επίσης να θεωρήσουμε το d ως την πιθανότητα των χρηστών να ακολουθούν τους συνδέσμους και το (1 - d) ως τη διανομή του Pagerank από μη απευθείας συνδεδεμένες σελίδες. Για να διαπιστώσει τη χρησιμότητα του αλγορίθμου Pagerank, η Google το εφάρμοσε στην αντίστοιχη μηχανή αναζήτησης. Στα πειράματα, διαπιστώθηκε ότι ο αλγόριθμος Pagerank λειτουργεί αποδοτικά, αφού η τιμή κατάταξης συγκλίνει σε μια λογική ανοχή σε περίπου λογαριθμικό χρόνο log n (Xing et al, 2004). Παρόλο που ο αλγόριθμος Pagerank χρησιμοποιείται επιτυχώς από την Google, εξακολουθεί να τίθεται ένα πρόβλημα, καθώς στο διαδίκτυο ορισμένοι σύνδεσμοι που βρίσκονται σε μια ιστοσελίδα, ενδέχεται να είναι πιο σημαντικοί από τους άλλους.

Όσο πιο δημοφιλής είναι μία ιστοσελίδα, τόσο περισσότεροι είναι οι σύνδεσμοι που περιέχουν άλλες ιστοσελίδες προς αυτήν. Με τον όρο Weighted Pagerank, αναφερόμαστε σε μία επέκταση του αλγορίθμου Pagerank, κατά την οποία αποδίδονται μεγαλύτερες τιμές κατάταξης σε δημοφιλέστερες σελίδες, αντί να πραγματοποιείται ομοιόμορφη κατανομή μεταξύ των συνδέσμων. Έτσι, κάθε ιστοσελίδα παίρνει μία τιμή η οποία ορίζει την σημαντικότητά της, με βάση το πόσο δημοφιλής είναι.. Για τον υπολογισμό της παρούσας τιμής, πρέπει να ληφθούν υπόψιν ο αριθμός των συνδέσμων από (inlinks) και προς (outlinks) την ιστοσελίδα οι οποίοι καταγράφονται ως Win (υ, u) και Wout (υ, u), αντίστοιχα.

Το Win(υ, u) είναι το βάρος της σύνδεσης (υ, u), που υπολογίζεται με βάση τον αριθμό των συνδέσμων της σελίδας u και τον αριθμό των συνδέσμων όλων των σελίδων αναφοράς της σελίδας υ, και αποδίδεται σύμφωνα με τον τύπο 3. Για τον τύπο 3 λοιπόν έχουμε Iu και Ip να αντιπροσωπεύουν τον αριθμό των συνδέσμων της σελίδας u και της σελίδας p, αντίστοιχα. Ενώ, το R(υ) δηλώνει τη λίστα σελίδων αναφοράς της σελίδας υ.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p}$$

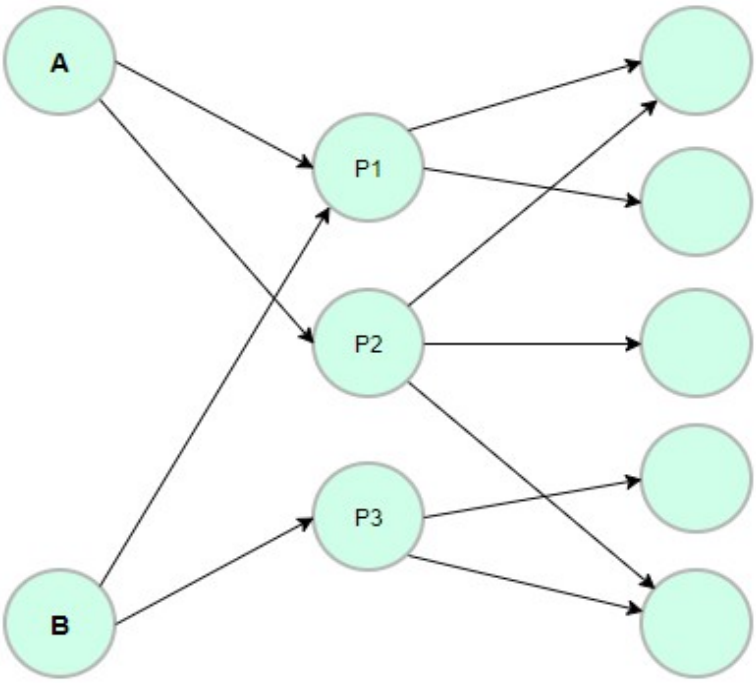
Τύπος 3

Αντίστοιχα, το Wout(υ, u) είναι το βάρος του συνδέσμου (υ, u), το οποίο υπολογίζεται με βάση τον αριθμό των εξωτερικών συνδέσμων της σελίδας u και τον αριθμό των συνδέσμων όλων των σελίδων αναφοράς της σελίδας υ. Στον τύπο 4, τα Ou και Op αντιπροσωπεύουν τον αριθμό των συνδέσμων της σελίδας u και της σελίδας p, ενώ το R(υ) δηλώνει τη λίστα σελίδων αναφοράς της σελίδας υ.

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

Τύπος 4

Στην εικόνα 6 απεικονίζεται ένα παράδειγμα κάποιων συνδέσμων σε μία ιστοσελίδα. Η ιστοσελίδα A, έχει δύο σελίδες αναφοράς την P1 και την P2. Ο αριθμός των συνδέσμων των δύο αυτών σελίδων (inlinks) είναι $L_{P1}=2$ και $L_{P2}=1$. Αντίστοιχα, ο αριθμός των συνδέσμων προς άλλες ιστοσελίδες (outlinks) είναι $O_{P1}=2$ και $O_{P2}=3$. Με βάση λοιπόν τα παραπάνω δεδομένα, έχουμε Win(A, P1)=Ip1/(Ip1 + Ip2) = 2/3 και Wout(A,p1) = Op1/(Op1 + Op2) =2/5.



Εικόνα 6: Δίκτυο Ιστοσελίδων

Προκειμένου λοιπόν να ληφθεί υπόψιν ο υπολογισμός των βαρών στην τελική Pagerank ταξινόμηση χρειάζεται να τροποποιηθεί ο τύπος δύο σύμφωνα με τον παρακάτω τύπο (Τύπος 5) :

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)}$$

Τύπος 5

Το Web Mining χρησιμοποιείται για την εξαγωγή πληροφοριών από την προηγούμενη συμπεριφορά των χρηστών. Η εξόρυξη δομών ιστού παίζει σημαντικό ρόλο σε αυτήν την προσέγγιση. Ο πιο γνωστός αλγόριθμος για την εξόρυξη δομών ιστού είναι ο Pagerank, ο οποίος χρησιμοποιείται για την κατάταξη σχετικών σελίδων. Το Pagerank αντιμετωπίζει όλους τους συνδέσμους με τον ίδιο τρόπο κατά την κατάταξη. Ο αλγόριθμος Weighted Pagerank, αποτελεί μια επέκταση στον αλγόριθμο Pagerank, η οποία λαμβάνει υπόψη τη σημασία τόσο των inlinks όσο και των outlinks των σελίδων, ενώ διανέμει βαθμολογίες με βάση τη δημοτικότητά τους. Μελέτες προσομοίωσης που χρησιμοποιούν τον ιστότοπο του Πανεπιστημίου Saint Thomas δείχνουν ότι το WPR είναι σε θέση να εντοπίσει μεγαλύτερο αριθμό σχετικών σελίδων σε ένα δεδομένο ερώτημα σε σύγκριση με το τυπικό Pagerank (Xing et al, 2004)

4 Ανασκόπηση Υπάρχουσας Βιβλιογραφίας

4.1 Υπάρχουσες σχετικές μελέτες

Όπως έχει ήδη αναφερθεί, το πρόβλημα της αναζήτησης χωρο-κειμενικών δεδομένων έχει απασχολήσει αρκετά την επιστημονική κοινότητα τα τελευταία χρόνια μετά την έλευση των τεχνολογιών ανεύρεσης τοποθεσίας και την ευρεία διάδοση των smartphones. Έτσι, έχουν προκύψει διάφορες μελέτες, οι οποίες καλούνται να προτείνουν αποδοτικές λύσεις, καθώς και διάφορες εφαρμογές οι οποίες τις εφαρμόζουν. Στην συγκεκριμένη ενότητα θα αναλυθούν σύντομα κάποιες από τις μελέτες αυτές.

4.1.1 Αναζήτηση σε χωρό-κειμενικά δεδομένα

Η συγκεκριμένη μελέτη (Felipe et al, 2008) είναι ίσως η δημοφιλέστερη που μελετά το συγκεκριμένο αντικείμενο. Σκοπός της είναι η παρουσίαση ενός νέου τρόπου με βάση τον οποίο θα πραγματοποιείτε αποδοτικότερα η αναζήτηση χωρο-κειμενικών δεδομένων. Πιο συγκεκριμένα οι συνεισφορές της είναι οι εξής:

- Καθορισμός του προβλήματος της top-k χωρο-κειμενικής αναζήτησης.
- Η πρόταση μίας αποδοτικής δομής για την αποθήκευση κειμενικών και χωρικών δεδομένων.
- Η παρουσίαση ενός αλγορίθμου ο οποίος χρησιμοποιεί την παραπάνω δομή, προκειμένου να απαντήσει αποδοτικά σε μία top-k χωρο-κειμενική ερώτηση.

Η μέθοδος η οποία προτείνεται, βασίζεται στη δομή R-Tree, ενώ οι συγγραφείς την ονομάζουν Information Retrieval R-Tree, και αποτελείτε ουσιαστικά από μία δομή R-Tree, στην οποία προστίθεται μία υπογραφή σε κάθε κόμβο του δέντρου. Σκοπός της προστιθέμενης υπογραφής είναι η δήλωση των λέξεων-κλειδιών με τρόπο τέτοιο, ώστε η αναζήτησή τους να είναι αποδοτική. Πιο τυπικά, ένα Information Retrieval R-Tree, είναι μια δενδρική δομή δεδομένων με

ισορροπημένο ύψος, όπου κάθε κόμβος έχει καταχωρήσεις της μορφής (ObjPtr, A, S). Τα ObjPtr και A ορίζονται όπως στην δομή R-Tree, ενώ το S είναι η υπογραφή του αντικειμένου που αναφέρεται από το ObjPtr. Οι διεργασίες που λαμβάνουν χώρα στην συγκεκριμένη δομή, είναι αυτές της εισαγωγής νέου κόμβου ή της διαγραφής του, κάτι που διασφαλίζει την μικρή πολυπλοκότητα του αλγορίθμου. Στο τέλος της μελέτης παρατίθενται πειραματικά αποτελέσματα τα οποία επιβεβαιώνουν την αποδοτικότητα του αλγορίθμου

4.1.2 Collective Spatial Keyword Querying

Οι συγγραφείς της παρούσας μελέτης (Cao et al, 2011) αναφέρουν, ότι τα ερωτήματα που μελετήθηκαν μέχρι τότε επικεντρώνονταν γενικότερα στην εύρεση μεμονωμένων αντικειμένων που το καθένα απαντούσε σε ένα ερώτημα, και όχι την εύρεση ομάδων αντικειμένων όπου ικανοποιούν συλλογικά ένα ερώτημα. Ορίζεται λοιπόν το πρόβλημα της ανάκτησης μιας ομάδας χωρικών αντικειμένων, έτσι ώστε οι λέξεις-κλειδιά της ομάδας να ικανοποιούν τις λέξεις-κλειδιά του ερωτήματος με σκοπό τα αντικείμενα να βρίσκονται πλησιέστερα στην τοποθεσία του χρήστη, έχοντας τις χαμηλότερες αποστάσεις μεταξύ αντικειμένων. Πιο συγκεκριμένα, ορίζεται ένα σύνολο χωρικών αντικειμένων D, και ένα ερώτημα $q = (\lambda, \psi)$, όπου λ μια θέση και ψ ένα σύνολο λέξεων-κλειδιών. Η μελέτη εξετάζει δύο περιπτώσεις του ερωτήματος.

- Στην πρώτη περίπτωση, στόχος είναι η εύρεση μίας ομάδας αντικειμένων χ που ικανοποιεί τις λέξεις-κλειδιά του ερωτήματος q, έτσι ώστε το άθροισμα των χωρικών αποστάσεων να ελαχιστοποιείται.
- Στην δεύτερη περίπτωση, στόχος είναι να βρεθεί μια ομάδα αντικειμένων χ που να ικανοποιεί τις λέξεις-κλειδιά του q, έτσι ώστε να ελαχιστοποιείται το άθροισμα της μέγιστης απόστασης μεταξύ των χ και του ερωτήματος q, αλλά και η μέγιστη απόσταση μεταξύ δύο αντικειμένων στην ομάδα αντικειμένων χ.

Η δομή που προτείνουν οι συγγραφείς για την κατάταξη των δεδομένων, είναι το IR-Tree, το οποίο στην ουσία είναι ένα R-Tree που χρησιμοποιεί ανεστραμμένα αρχεία. Κάθε κόμβος ενός IR-Tree, περιέχει καταχωρήσεις της μορφής (o, o.λ, o.di), όπου το o είναι ένα αντικείμενο στο σύνολο δεδομένων D. Το o.λ είναι το ορθογώνιο οριοθέτησης του o και o.di είναι ένα αναγνωριστικό της περιγραφής του o. Επίσης, κάθε κόμβος περιέχει ένα δείκτη σε ένα ανεστραμμένο αρχείο με τις λέξεις-κλειδιά των αντικειμένων που είναι αποθηκευμένα στον κόμβο. Ένας ανεστραμμένος δείκτης αρχείων έχει δύο κύρια στοιχεία:

- Το λεξιλόγιο όλων των διακριτών λέξεων που εμφανίζονται στην περιγραφή ενός αντικειμένου.
- Μια λίστα για κάθε λέξη t, που αποτελείτε από μια ακολουθία αναγνωριστικών των αντικειμένων των οποίων οι περιγραφές περιέχουν την λέξη t.

Άλλη μία προσφορά της παρούσας μελέτης είναι η παρουσίαση δύο συναρτήσεων οι οποίες αφορούν τον υπολογισμό του κόστους της αναζήτησης στο IR-Tree. Οι ερευνητές εφαρμόζουν

τους αντίστοιχους αλγορίθμους σε διαφορετικά σύνολα δεδομένων, και παρουσιάζουν τα αποτελέσματα.

4.1.3 Keyword Search in Spatial Databases: Towards Searching by Document

Η μελέτη (Zhang et al, 2009) αναφέρεται στην αναζήτηση χωρο-κειμενικών δεδομένων, ενώ σκοπός της δημοσίευσης, είναι η παρουσίαση του *m*-closest keywords (*m*CK) query. Έχοντας μια βάση δεδομένων με χωρικά αντικείμενα, κάθε πλειάδα συσχετίζεται με ορισμένες περιγραφικές πληροφορίες που εκπροσωπούνται από λέξεις-κλειδιά. Το ερώτημα mCK στοχεύει στην εύρεση των πλησιέστερων χωρικών αντικειμένων που ταιριάζουν με τις λέξεις-κλειδιά που καθορίζει ο χρήστης. Δεδομένου ενός συνόλου λέξεων-κλειδιών από ένα έγγραφο, το ερώτημα mCK μπορεί να είναι πολύ χρήσιμο για τη γεωγραφική προσθήκη ετικετών στο έγγραφο, συγκρίνοντας τις λέξεις-κλειδιά με άλλα έγγραφα τα οποία έχουν γεωγραφική ετικέτα σε μια βάση δεδομένων. Για να απαντηθεί αποτελεσματικά ένα ερώτημα mCK, παρουσιάζεται μία νέα δομή που ονομάζεται bR*-Tree, η οποία αποτελεί επέκταση του R*-Tree.

Οι κύριες συνεισφορές της συγκεκριμένης μελέτης είναι οι εξής:

- Προτείνεται ένα νέο spatial keyword query, που ονομάζεται mCK query, το οποίο έχει μεγάλο αριθμό εφαρμογών σε χωρικές βάσεις δεδομένων.
- Προτείνεται επίσης ένας νέος τρόπος αναζήτησης, που ονομάζεται bR*-Tree, το οποίο επεκτείνει το R*-Tree προκειμένου να συνοψίζονται αποτελεσματικά οι λέξεις-κλειδιά και οι χωρικές τους πληροφορίες.
- Ενσωματώνονται αποτελεσματικές στρατηγικές αναζήτησης, οι οποίες μειώνουν σημαντικά τον χώρο αναζήτησης.
- Ορίζονται δύο περιορισμοί μονοτονικού, συγκεκριμένα το distance mutex και το keyword mutex, ως ιδιότητες για διαγραφή των κόμβων του δέντρου. Παρέχονται επίσης αποδοτικές εφαρμογές για την εξέταση αυτών των περιορισμών.
- Τέλος πραγματοποιούνται εκτενή πειράματα για να δειχθεί ότι ο αλγόριθμός είναι αποτελεσματικός στη μείωση του χρόνου απάντησης ερωτημάτων mCK, ενώ εμφανίζει καλή επεκτασιμότητα όσον αφορά τον αριθμό των λέξεων-κλειδιών του ερωτήματος.

4.1.4 Efficient Processing of Top-k Spatial Keyword Queries

Σκοπός αυτού του άρθρου (Joao et al, 2011), είναι η πρόταση ενός νέου τρόπου ανάθεσης δεικτών, για τη βελτίωση της απόδοσης των top-k χωρο-κειμενικών ερωτημάτων. Ο νέος αυτός τρόπος ονομάζεται Spatial Inverted Index (S2I). Ο τρόπος που προτείνουν οι συγγραφείς, χαρτογραφεί κάθε ξεχωριστό όρο σε ένα σύνολο αντικειμένων που περιέχουν τον όρο. Τα αντικείμενα αποθηκεύονται ανάλογα με τη συχνότητα που βρίσκονται στο κείμενο, και μπορούν

να ανακτηθούν αποτελεσματικά με φθίνουσα σειρά της συνάφειας των λέξεων-κλειδιών και της χωρικής εγγύτητας. Επιπλέον, παρουσιάζονται αλγόριθμοι που εκμεταλλεύονται το S2I για την αποτελεσματική επεξεργασία χωρο-κειμενικών ερωτημάτων.

Το S2I κατατάσσει κάθε όρο t σε ένα συγκεντρωτικό R-Tree (aR-Tree) ή σε ένα μπλοκ που αποθηκεύει τα χωρο-κειμενικά αντικείμενα p που περιέχουν τον όρο t. Οι πιο συνηθισμένοι όροι αποθηκεύονται σε aR-tree, ενώ οι λιγότερο συχνοί όροι αποθηκεύονται σε ένα αρχείο. Ομοίως με έναν παραδοσιακό ανεστραμμένο δείκτη, το S2I αποθηκεύει τους όρους σε αντικείμενα που περιέχουν τον όρο. Επίσης, χρησιμοποιούνται δύο διαφορετικές δομές δεδομένων, μία για λιγότερο συχνούς όρους και μια άλλη για πιο συχνούς, στους οποίους παρέχεται πρόσβαση με φθίνουσα σειρά, ανάλογα με την συνάφεια των λέξεων-κλειδιών και της χωρικής απόστασης .

Οι υπόλοιπες συνεισφορές αυτού του άρθρου είναι:

- Η παρουσίαση αποτελεσματικών αλγόριθμων που εκμεταλλεύονται το S2I προκειμένου να επεξεργάζονται αποτελεσματικά ερωτήματα χωρικών λέξεων-κλειδιών.
- Τέλος, οι συγγραφείς δείχνουν μέσα από μια εκτεταμένη πειραματική αξιολόγηση ότι η προσέγγισή τους ξεπερνά τους « state-of-the-art » αλγορίθμους από την άποψη του χρόνου ενημέρωσης, το κόστος και τον χρόνο απόκρισης.

4.1.5 Spatial Keyword Query Processing: An Experimental Evaluation

Η μελέτη (Chen et al, 2013) δημοσιεύθηκε με σκοπό να αξιολογήσει τις μέχρι τότε υπάρχουσες μεθόδους, για την πραγματοποίηση αναζήτησης σε χωρο-κειμενικά δεδομένα. Στο κείμενο αναφέρονται τα αποτελέσματα που προέκυψαν κατά την εφαρμογή της παρούσας σύγκρισης, αποκαλύπτοντας έτσι νέες πληροφορίες που μπορεί να καθοδηγήσουν την επιλογή τρόπου πραγματοποίησης μίας τέτοιας αναζήτησης. Αρχικά, οι συγγραφείς κατατάσσουν τους τρόπους με τους οποίους τίθεται κάθε φορά η ερώτηση (query), με τον παρακάτω τρόπο:

- Boolean kNN Query: Ανακτήστε τα k αντικείμενα που βρίσκονται πλησιέστερα στην τρέχουσα τοποθεσία του χρήστη (που αντιπροσωπεύεται από ένα σημείο) έτσι ώστε η κειμενική περιγραφή κάθε αντικειμένου να περιέχει τις λέξεις-κλειδιά: « πίτσα και καπουτσίνο ».
- Top-k kNN Query: Ανάκτηση των αντικειμένων k με την υψηλότερη βαθμολογία, μετρούμενη ως συνδυασμός της απόστασής τους από την τοποθεσία του ερωτήματος (ένα σημείο) και της συνάφειας της κειμενικής περιγραφής τους με τις λέξεις-κλειδιά : « πίτσα και καπουτσίνο ».
- Boolean Range Query: Ανακτήστε όλα τα αντικείμενα των οποίων η κειμενική περιγραφή περιέχει τις λέξεις-κλειδιά: « πίτσα και καπουτσίνο » και των οποίων η τοποθεσία βρίσκεται σε απόσταση 10 χλμ από την τοποθεσία του ερωτήματος.

Εκτός από την κατηγοριοποίηση των ερωτημάτων που χρησιμοποιεί η κάθε τακτική για πραγματοποίηση αναζήτησης, οι συγγραφείς κατατάσσουν τις τακτικές με βάση τον τρόπο της αναζήτησης και τον τρόπο που πραγματοποιείται η δεικτοδότηση των λέξεων-κλειδιών. Η σύγκριση πραγματοποιείται έχοντας ως κριτήρια τον χρόνο που χρειάζεται για να πραγματοποιηθεί η αναζήτηση , καθώς και το κόστος σε χώρο. Επίσης τα τελικά πειράματα πραγματοποιούνται για διαφορετικούς όγκους δεδομένων. Με την έρευνα και την σύγκριση συνολικά 12 τεχνικών για πραγματοποίηση χωρο-κειμενικής αναζήτησης, η παρούσα μελέτη προσφέρει μία αντικειμενική αξιολόγηση η οποία σκοπεύει να βοηθήσει στην επιλογή της κατάλληλης.

5 Μέθοδος για την επίλυση του προβλήματος

5.1 Δημιουργία του αρχείου log

Για την εκπόνηση της παρούσας διπλωματικής εργασίας, ήταν απαραίτητη η χρήση εγγραφών, οι οποίες αναπαριστούν τα ερωτήματα των χρηστών, και χρησιμοποιούνται για την δημιουργία του τελικού γράφου. Το πλήθος των εγγραφών αυτών έπρεπε να είναι αρκετά μεγάλο, προκειμένου να παρέχονται αρκετά δεδομένα για την δημιουργία του γράφου και την κατάληξη σε χρήσιμα συμπεράσματα. Προκειμένου λοιπόν να διασφαλιστούν τα παραπάνω, δημιουργήθηκαν με τυχαίο τρόπο οι εγγραφές από τις οποίες αποτελείτε το το συγκεκριμένο αρχείο. Μία εγγραφή που υπάρχει στο αρχείο log αποτελείτε από τα εξής:

- ID: Αποτελεί το αναγνωριστικό το οποίο αντιστοιχεί σε ένα ερώτημα, και πρόκειται για έναν μοναδικό ακέραιο αριθμό που στην περίπτωσή μας είναι αύξοντας. Δηλαδή το ερώτημα που τέθηκε πρώτο θα έχει ID=1, το δεύτερο ID=2 και ούτω καθεξής.
- X: Πεδίο που αντιπροσωπεύει το γεωγραφικό μήκος του σημείου που βρίσκεται ο χρήστης.
- Y: Πεδίο που αντιπροσωπεύει το γεωγραφικό πλάτος του σημείου που βρίσκεται ο χρήστης.
- Radius: Πεδίο το οποίο εκφράζεται σε μέτρα και αναπαριστά την ακτίνα στην οποία είναι επιθυμητό να πραγματοποιηθεί η αναζήτηση. Η τιμή του πεδίου αποδίδεται τυχαία, και κυμαίνεται από 10 έως 100 χιλιόμετρα.
- k: Ο αριθμός των αποτελεσμάτων που θέλει να λάβει ο χρήστης ως απάντηση.
- keywords: Πεδίο το οποίο περιλαμβάνει τις λέξεις κλειδιά που χρησιμοποιούνται κατά την αναζήτηση

Ενδεικτικά ο πίνακας 9 αναπαριστά τρεις τυχαίες εγγραφές που βρίσκονται καταγεγραμμένες στο αρχείο log.

ID	*X*	*Y*	*Radius(m)*	*k*	*keywords*
35	36.703247	-118.438793	32451	1	Fish And Chips, Vegan
36	36.353357	-120.126421	29557	5	French, Burgers
37	37.923467	-117.449588	9622	4	Middle Eastern, Ice Cream, Smoothies

Πίνακας 9: Εγγραφές log file

Πιο συγκεκριμένα, για την δημιουργία ενός αρχείου log πρέπει αρχικά να καθοριστεί ο επιθυμητός αριθμός των εγγραφών που θα το αποτελούν. Έπειτα υπολογίζονται με τυχαίο τρόπο οι συντεταγμένες που αναπαριστούν το σημείο που βρίσκεται ο χρήστης. Για τα πεδία x και y δημιουργούνται αριθμοί τύπου double, το εύρος των οποίων μπορεί να μεταβληθεί ανάλογα με τις απαιτήσεις του εκάστοτε συστήματος και την περιοχή στην οποία βρίσκονται τα αντίστοιχα σημεία ενδιαφέροντος. Επίσης, για την αναπαράσταση της ακτίνας (Radius) και τον αριθμό των επιθυμητών αποτελεσμάτων δημιουργούνται με τυχαίο τρόπο τυχαίοι ακέραιοι αριθμοί τύπου int. Για τον αριθμό που αναπαριστά το ID, υπολογίζεται αυτοματοποιημένα και αυξάνεται κατά έναν κάθε φορά που πρόκειται να δημιουργηθεί μία εγγραφή.

Για την δημιουργία του πεδίου με τις λέξεις κλειδιά, χρειαζόμαστε αρχικά μία δομή τύπου πίνακα στον οποίο έχουν εκχωρηθεί όλες οι πιθανές λέξεις που θα μπορούσε να χρησιμοποιήσει ο χρήστης αναφερόμενος στα εστιατόρια που υπάρχουν στην βάση δεδομένων. Έπειτα, χρειαζόμαστε έναν τυχαίο ακέραιο, ο οποίος θα εκφράζει το πλήθος των λέξεων που πρόκειται να χρησιμοποιηθούν από τον χρήστη. Τέλος, αφού έχει καθοριστεί το πλήθος των λέξεων υπολογίζονται τυχαίοι ακέραιοι οι οποίοι εκφράζουν την θέση του πίνακα από την οποία πρόκειται να αντληθεί η κάθε λέξη κλειδί. Βέβαια, εκτός από τον τυχαίο τρόπο δημιουργίας εγγραφών, παρέχεται και η δυνατότητα για χειροκίνητη εισαγωγή των πεδίων μέσω της εφαρμογής. Ο χρήστης εισάγει την τοποθεσία του, την ακτίνα, το k και τις λέξεις κλειδιά, και η εγγραφή καταχωρείτε στο αρχείο παίρνοντας το αντίστοιχο ID.

5.2 Δημιουργία του γράφου

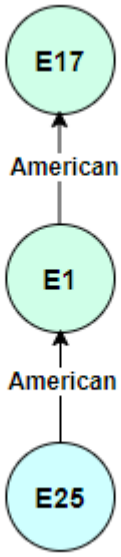
Σκοπός της παρούσας εργασίας, είναι η δημιουργία ενός γράφου, ο οποίος δημιουργείται με βάση τα ερωτήματα που θέτουν οι χρήστες προς μία βάση δεδομένων. Στην βάση αυτή εμπεριέχονται εγγραφές που αντιπροσωπεύουν εστιατόρια που βρίσκονται σε μία γεωγραφική τοποθεσία. Όπως έχει αναφερθεί, κάθε εγγραφή η οποία αφορά ένα εστιατόριο, αποτελείτε από έναν μοναδικό ακέραιο αριθμό (id), από τιμές που αναπαριστούν το γεωγραφικό μήκος και πλάτος x και y αντίστοιχα, καθώς και από τις λέξεις κλειδιά οι οποίες αναπαριστούν τους τύπους φαγητών που προσφέρει το κάθε εστιατόριο. Αντίστοιχα, η ερώτηση που πραγματοποιεί ο κάθε χρήστης αποτελείτε από ένα μοναδικό αναγνωριστικό, από την τοποθεσία του (x και y), από μία ακτίνα

γύρω από αυτήν στην οποία επιθυμεί να πραγματοποιηθεί η αναζήτηση, από τον αριθμό των αποτελεσμάτων που επιθυμεί να λάβει ως απάντηση, καθώς και από τις αντίστοιχες λέξεις κλειδιά. Κάθε ερώτηση που πραγματοποιείται, καταγράφεται στο αρχείο καταγραφής (log file).

Αρχικά κρίνεται σκόπιμο να ορίσουμε τον γράφο που καλούμαστε να υλοποιήσουμε, με βάση τις απαιτήσεις του προβλήματος που μελετάμε. Έτσι ορίζουμε τον γράφο ως εξής:

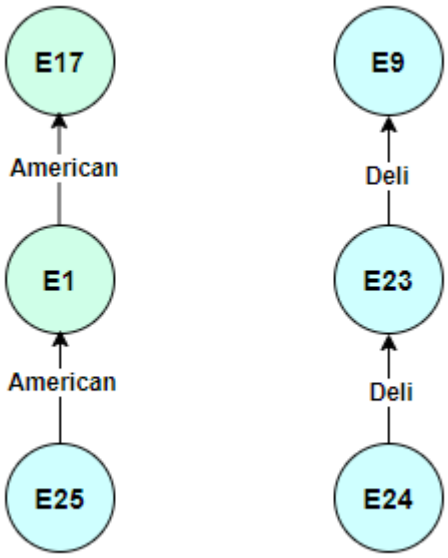
- Οι κορυφές του γράφου αναπαριστούν τα εστιατόρια επομένως κάθε κορυφή είναι ένα αντικείμενο της βάσης δεδομένων.
- Μια ακμή από τον κόμβο Α στον κόμβο Β, σημαίνει ότι υπάρχει στο log file μια top-k ερώτηση όπου ο κόμβος Β προηγείται από τον κόμβο Α.
- Κάθε ακμή έχει και κάποια tags τα οποία αντιστοιχούν στα keywords που είχε η ερώτηση με την οποία δημιουργήθηκε η ακμή.
- Κάθε μια κορυφή του γράφου αντιστοιχεί σε κάποιο εστιατόριο το οποίο έχει επιστραφεί τουλάχιστον μία φορά ως ένα από τα top-k αποτελέσματα για κάποια ερώτηση.

Στην ενότητα 2.3 μελετήσαμε κάποιες περιπτώσεις, κατά τις οποίες ένας χρήστης πραγματοποιεί αναζήτηση για κάποιο εστιατόριο από μία συγκεκριμένη τοποθεσία. Στο παράδειγμα 1 αναλύθηκε η περίπτωση κατά την οποία η ερώτηση πραγματοποιείτε από το σημείο p1 με βάση την λέξη-κλειδί: «American», ενώ έπρεπε να επιστραφούν τρία βέλτιστα αποτελέσματα. Έπειτα, η απάντηση που θα λάμβανε ο χρήστης αναπαραστάθηκε από τον πίνακα 3, όπου τα εστιατόρια με id: E17, E1 και E25, ήταν τα τρία κοντινότερα. Έτσι, ο γράφος που προκύπτει για το αντίστοιχο ερώτημα, αναπαρίσταται από το σχήμα της εικόνας 7. Παρατηρούμε ότι ο γράφος αποτελείτε από τρεις κορυφές, οι οποίες συνδέονται μεταξύ τους με δύο ακμές, οι οποίες χαρακτηρίζονται από την λέξη με βάση την οποία πραγματοποιήθηκε η αναζήτηση.



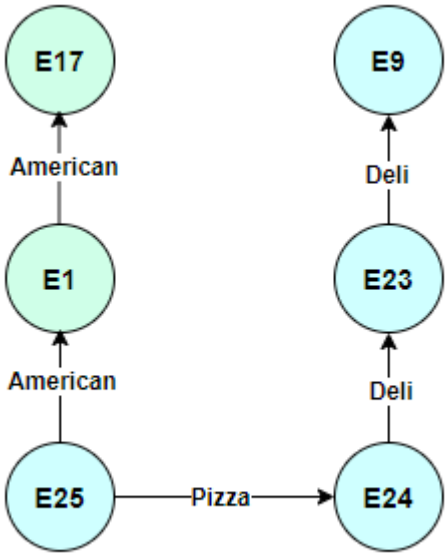
Εικόνα 7: Δημιουργία γράφου
- βήμα 1

Συνεχίζοντας, με την κατασκευή του γράφου, θα πρέπει να λάβουμε υπόψιν μας την επόμενη ερώτηση που τέθηκε προς την βάση δεδομένων. Έτσι, παίρνουμε την ερώτηση του παραδείγματος 2 από την ενότητα 2.3. Τα αποτελέσματα της παρούσας ερώτησης, αναπαραστάθηκαν στον πίνακα 4, και οι αντίστοιχες κορυφές που επιστράφηκαν ως απάντηση ήταν οι E9, E23 και E24. Ο γράφος λοιπόν, διαμορφώνεται σύμφωνα με το σχήμα της εικόνας 8. Ο ανανεωμένος γράφος αποτελείτε από τρεις επιπλέον κορυφές οι οποίες ενώνονται με δύο ακμές.



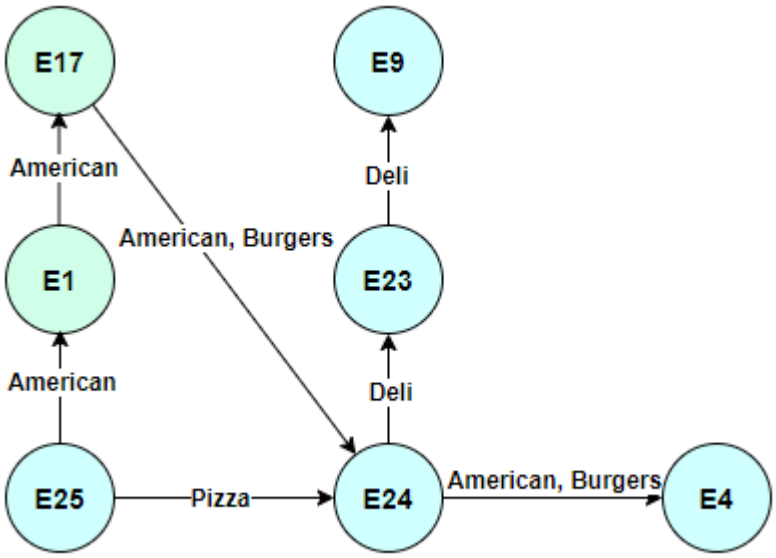
Εικόνα 8: Δημιουργία γράφου - βήμα 2

Σε αυτό το σημείο παρατηρούμε ότι οι δύο ερωτήσεις λαμβάνουν απαντήσεις οι οποίες δεν συνδέονται μεταξύ τους, καθώς καμία κορυφή που ικανοποιεί την μία ερώτηση δεν ικανοποιεί και την άλλη. Ως επόμενο βήμα, λαμβάνουμε υπόψιν το επόμενο παράδειγμα της ενότητας 2.2. Το αποτέλεσμα του παραδείγματος 3 αναπαρίσταται από τον πίνακα 5 και αποτελείται από τα εστιατόρια με id E24 και E25.



Εικόνα 9: Δημιουργία γράφου - βήμα 3

Στο σχήμα της εικόνας 9 παρατηρούμε την επίπτωση που είχε το ερώτημα στο υπό κατασκευή γράφημα. Οι κόμβοι E25 και E24 προ-υπήρχαν, οπότε η μόνη αλλαγή που προέκυψε από την ενσωμάτωση του νέου ερωτήματος είναι ακμή με την λέξη κλειδί που ενώνει τις δύο κορυφές. Στη συνέχεια προχωράμε στο επόμενο ερώτημα το οποίο διαμορφώνει τον γράφο. Τα αποτελέσματα του επόμενου ερωτήματος αναγράφονται στον πίνακα 6 και περιλαμβάνουν τις κορυφές E4, E24 και E17. Από τις παραπάνω κορυφές η μία πρέπει να δημιουργηθεί, ενώ οι υπόλοιπες υπάρχουν ήδη στο γράφημα. Η επίπτωση της νέας ερώτησης στον γράφο παρουσιάζεται στην εικόνα 10 . Το τελευταίο ερώτημα της ενότητας 2.3, λαμβάνει ένα μόνο αποτέλεσμα ως απάντηση, το E4. Στην περίπτωση αυτή ο κόμβος προϋπάρχει, επομένως δεν πραγματοποιείτε κάποια αλλαγή στο γράφημα. Στην περίπτωση που ο συγκεκριμένος κόμβος δεν υπήρχε ήδη στον γράφο, τότε απλά θα δημιουργούνταν χωρίς να συνδέεται με κάποια άλλη κορυφή μέσω ακμής.



Εικόνα 10: Δημιουργία γράφου - βήμα 4

Κατά την δημιουργία ενός αντίστοιχου γραφήματος, ενδέχεται να προκύψουν διάφορες περιπτώσεις. Για παράδειγμα, σε μια εφαρμογή θα μπορούσε μία ακμή να εμφανιστεί σε περισσότερα από ένα ερωτήματα. Στην περίπτωση αυτή, δεν δημιουργούμε την ακμή εκ νέου, αλλά ενημερώνουμε τις λέξεις τις οποίες την χαρακτηρίζουν, προσθέτοντας την λέξη κλειδί που οδήγησε στην επανεμφάνισή της. Σε κάθε αντίστοιχη περίπτωση θα πρέπει να γνωρίζουμε σε πόσα ερωτήματα έχει εμφανιστεί οποιαδήποτε ακμή.

5.3 Αποθήκευση του γράφου

Έπειτα από την δημιουργία του αρχείου καταγραφής ερωτημάτων και την δημιουργία του τελικού γραφήματος, κρίθηκε απαραίτητη η εύρεση ενός τρόπου κατά τον οποίο θα αποθηκεύεται ο γράφος. Η αποθήκευση του γράφου είναι ιδιαίτερα σημαντικό κομμάτι, καθώς εξοικονομεί πολύτιμο χρόνο κατά την εκτέλεση μίας εφαρμογής τέτοιου τύπου. Η δημιουργία ενός γραφήματος ή οποιαδήποτε άλλη διεργασία που απαιτεί την προσπέλαση και σύγκριση μεγάλων αρχείων δεδομένων, είναι αρκετά χρονοβόρα διαδικασία, και δεν θα ήταν συνετό να εκτελείτε από την αρχή κάθε φορά που προστίθενται νέα δεδομένα. Για τον λόγο αυτό, δημιουργήθηκε ένα νέο αρχείο, στο οποίο πραγματοποιείτε η αποθήκευση του γραφήματος (export), ενώ παρέχεται και η δυνατότητα της ανάγνωσής του (import), χωρίς να απαιτείτε εκ νέου προσπέλαση του αρχείου καταγραφής και του αρχείου με τις εγγραφές των εστιατορίων.

Η αποθήκευση του γραφήματος πραγματοποιήθηκε σε ένα αρχείο σε μορφή DOT (graph description language). Η συγκεκριμένη μορφή είναι μία από τις καταλληλότερες για την αποθήκευση και την σχεδίαση κατευθυνόμενων γραφημάτων. Τα χαρακτηριστικά του περιλαμβάνουν καλά συντονισμένους αλγόριθμους διάταξης για την τοποθέτηση κόμβων και ακμών, καθώς και ετικέτες, οι οποίες μπορούν να χρησιμεύσουν στην αποθήκευση των διάφορων γνωρισμάτων που κρίνονται απαραίτητα για την αναδημιουργία του γραφήματος. Στην περίπτωσή μας, για την αποθήκευση του γράφου χρειάστηκε η παροχή κάποιον επιπλέον πληροφοριών, πέρα από τις συσχετίσεις μεταξύ κορυφών.

Για την δημιουργία ενός αρχείου DOT απαιτείτε η παροχή ενός μοναδικού αναγνωριστικού για κάθε κόμβο, το οποίο χρησιμοποιείτε για την δημιουργία των συσχετίσεων μεταξύ των κόμβων. Στην περίπτωσή μας ένα τέτοιο αναγνωριστικό αποτελεί το «id», το οποίο έχει οριστεί ήδη για κάθε κόμβο που αναπαριστά ένα εστιατόριο, ενώ εάν οι κόμβοι του γραφήματος δεν διέθεταν ένα αντίστοιχο χαρακτηριστικό, θα προστίθενται με αυτόματο τρόπο, κατά την δημιουργία του τελικού αρχείου. Εκτός από το «id», ήταν αναγκαία και η αποθήκευση των υπολοίπων γνωρισμάτων που αποτελούν μία εγγραφή εστιατορίου. Για την ανάγκη αυτή παρέχεται η δυνατότητα της αποθήκευσης επιπλέον χαρακτηριστικών για κόμβους και ακμές.

Για την επίτευξη των παραπάνω, απαιτείτε η απόδοση μίας ετικέτας για κάθε αναγνωριστικό, έτσι, η αποθήκευση μίας εγγραφής εστιατορίου με την μορφή κόμβου πραγματοποιείται με την εξής μορφή: « 25 [id="25" name="Bistango Cafe" x="34.02491" y="-118.27809" keywords="Italian , American , Pizza , Sandwiches , Pasta , Salad "]; » Αντίστοιχα η αποθήκευση μίας ακμής που συσχετίζει για παράδειγμα τον κόμβο 25 με τον κόμβο 24 εκφράζεται ως εξής: « 25 --→ 24 [label="Pizza ,"]; » . Η ετικέτα με το όνομα «label», περιλαμβάνει τις λέξεις κλειδιά που χαρακτηρίζουν την ακμή.

5.4 Κατάταξη των δεδομένων

Εξίσου σημαντικό με την δημιουργία του γράφου, είναι η κατάταξη των δεδομένων που τον αποτελούν, προκειμένου να καταλήξουμε στα τελικά χρήσιμα συμπεράσματα. Για την κατάταξη των δεδομένων, επιλέχθηκαν οι αλγόριθμοι Pagerank και Weighted Pagerank. Έτσι, ο αλγόριθμος Pagerank εφαρμόζεται στον τελικό γράφο και λειτουργεί σύμφωνα με τον τρόπο που αναλύθηκε στο κεφάλαιο 3.1.2, με την διαφορά ότι πλέον δεν αναφερόμαστε σε ιστοσελίδες, αλλά σε εστιατόρια τα οποία αναπαρίστανται από τους κόμβους. Ενδεικτικά λοιπόν, για τον γράφο της εικόνας 10 του κεφαλαίου 5.2, ο κόμβος με την μεγαλύτερη βαθμολογία Pagerank, είναι ο E24 (≈0,209) και ακολουθούν οι κόμβοι E9 και E23 (≈0,192 και ≈0,152 αντίστοιχα).

Για την εφαρμογή του αλγορίθμου Weighted Pagerank, χρειάστηκε η ανάθεση κάποιου βάρους στις ακμές του γράφου. Έτσι αποφασίστηκε ότι κατά την δημιουργία του γράφου θα υπολογίζεται το πλήθος των φορών που θα εμφανίζεται μία ακμή. Στην περίπτωση λοιπόν που μία ακμή

εμφανίζεται για ένα ή παραπάνω ερωτήματα, τότε το βάρος της ακμής θα αυξάνεται κατά μισή μονάδα για κάθε ερώτημα. Αν λοιπόν έχουμε μία ακμή που εμφανίζεται σε 3 ερωτήματα, τότε το βάρος της συγκεκριμένης ακμής θα είναι 3 × 0,5=1,5. Επίσης, κρίθηκε σκόπιμη η μεταβολή του βάρους των ακμών, με βάση μία λέξη την οποία ορίζει ο χρήστης. Έτσι, το βάρος της κάθε ακμής η οποία περιλαμβάνει την λέξη που θέτει ο χρήστης πολλαπλασιάζεται με 10. Οπότε, αν έχουμε μία ακμή που εμφανίζεται τρεις σε τρεις ερωτήσεις και περιλαμβάνει την λέξη που ορίζεται από τον χρήστη, το συνολικό βάρος θα γίνει (3 × 0,5) × 10 =15. Ενδεικτικά λοιπόν, για τον γράφο της εικόνας 10 του κεφαλαίου 5.2, ο κόμβος με την μεγαλύτερη βαθμολογία Weighted Pagerank, είναι ο E9 (≈0,239) και ακολουθούν οι κόμβοι E23 και E24 (≈0,211 και ≈0,197 αντίστοιχα).

6 Πειραματική αποτίμηση

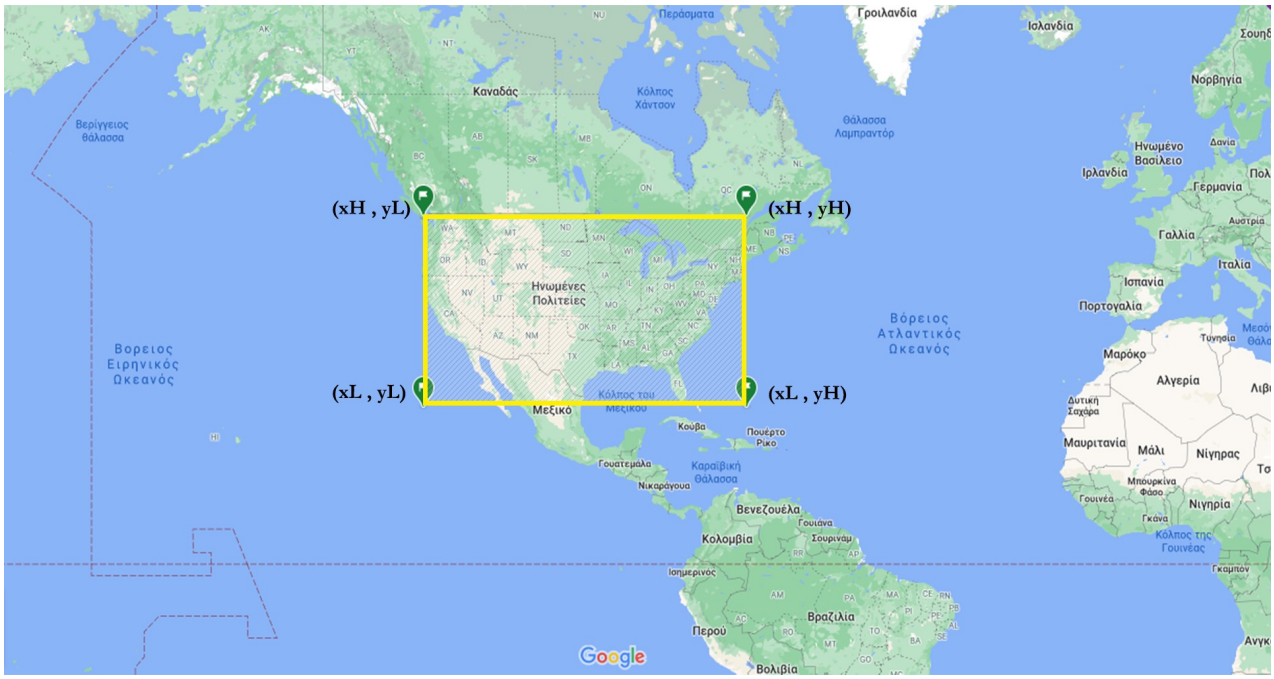
6.1 Περιγραφή των δεδομένων

Για την δημιουργία του πρακτικού κομματιού της παρούσας διπλωματικής εργασίας έγινε χρήση ενός αρχείου από το factual.com, μέσα στο οποίο εμπεριέχονται οι εγγραφές των εστιατορίων. Το συγκεκριμένο αρχείο είναι της μορφής txt, ενώ κάθε εγγραφή καταλαμβάνει μία γραμμή . Στο αρχείο αναπαρίστανται συνολικά 78970 εστιατόρια. Για κάθε εστιατόριο, καταγράφονται όλες οι πληροφορίες που μας ενδιαφέρουν, δηλαδή δεν υπάρχουν ελλιπής εγγραφές. Οι εγγραφές του αρχείου αναπαριστούν πραγματικά δεδομένα και αφορούν επιχειρήσεις που βρίσκονται στις ΗΠΑ. Επίσης κάνοντας χρήση κώδικα, βρέθηκαν οι μέγιστες και ελάχιστες τιμές του γεωγραφικού μήκους και πλάτους όλων των τοποθεσιών, και παρουσιάζονται στον πίνακα 10. Σε αυτό το σημείο πρέπει να αναφερθεί, ότι κατά την ανίχνευση των τιμών του πίνακα 10, παρατηρήθηκε ότι στο αρχικό αρχείο εμπεριέχονταν περίπου δέκα εγγραφές, των οποίων οι γεωγραφικές θέσεις δεν είχαν αποδοθεί σωστά. Οι συγκεκριμένες εγγραφές εντοπίστηκαν και απομονώθηκαν.

Μέγιστο γεωγραφικό μήκος (xH):	24.547224
Ελάχιστο γεωγραφικό μήκος (xL):	-124.20028
Μέγιστο γεωγραφικό πλάτος (yH):	-69.95395
Ελάχιστο γεωγραφικό πλάτος (yL):	-124.20028

Πίνακας 10: Μέγιστες και ελάχιστες τιμές συντεταγμένων

Η εύρεση των ελάχιστων και μέγιστων στις γεωγραφικές συντεταγμένες, παρέχει την δυνατότητα εύρεσης της γεωγραφικής έκτασης στην οποία βρίσκονται τα εστιατόρια. Έτσι, λαμβάνοντας υπόψιν τα τεχνητά σημεία (xH , yL), (xH , yH), (xL , yH) και (xL , yL), βρίσκουμε ότι οι εγγραφές μας αναφέρονται στο γεωγραφικό τετράγωνο που φαίνεται στην εικόνα 11.



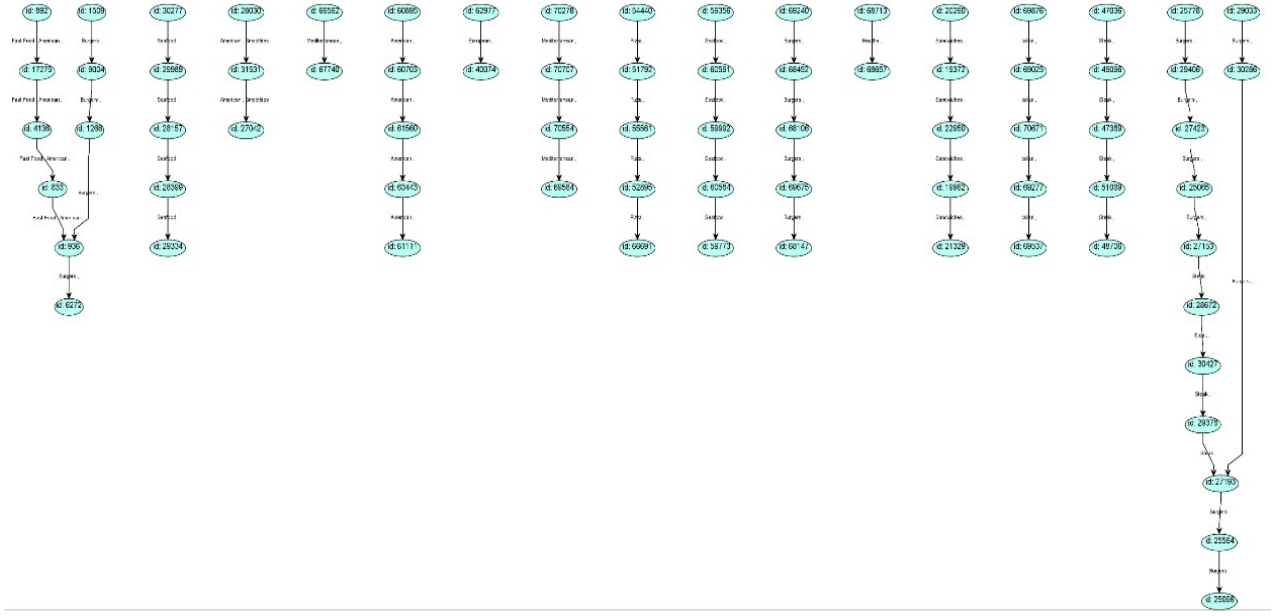
Εικόνα 11: Τοποθεσία εστιατορίων

6.2 Λειτουργία του προγράμματος

Για τις ανάγκες της εκπόνησης της παρούσας διπλωματικής εργασίας, επιλέχθηκε η παρουσίαση των αποτελεσμάτων με δύο ανεξάρτητα αρχεία καταγραφής (log files). Τα αποτελέσματα του πρώτου αρχείου καταγραφής θα παρουσιαστούν στο κεφάλαια 6.2.1. και αντίστοιχα του δεύτερου στο 6.2.2 . Για την παρακάτω αναπαράσταση, έγινε χρήση του εργαλείου jGraphX, μέσω του οποίου, δημιουργήθηκαν οι τελικές εικόνες.

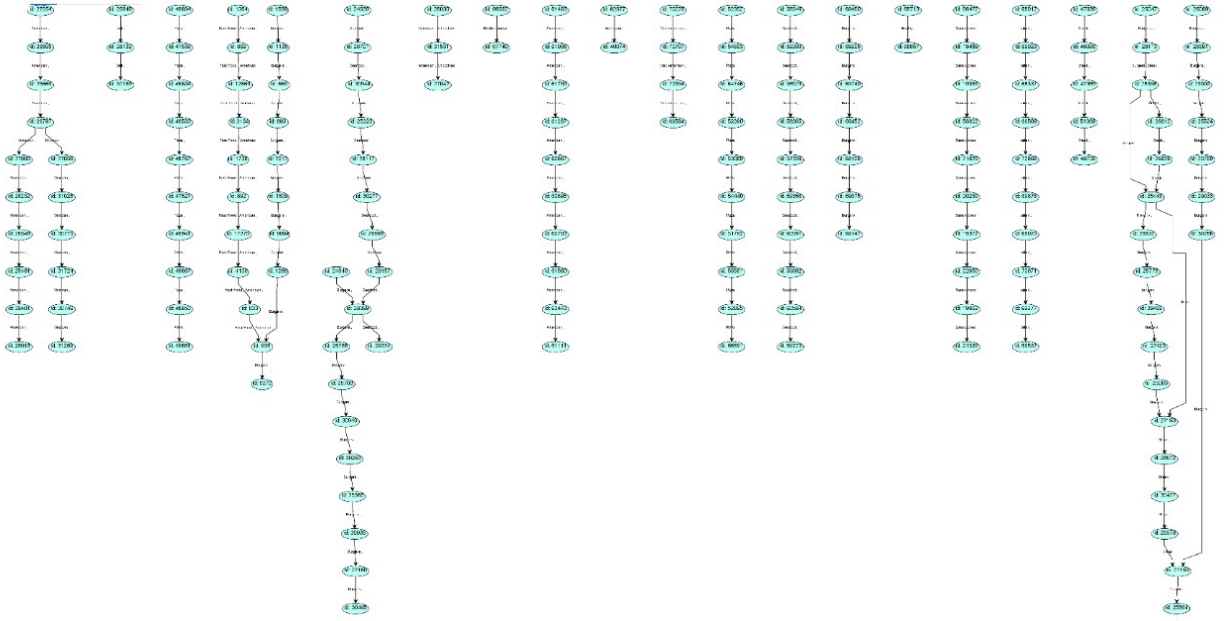
6.2.1 Αποτελέσματα με μεγάλο αρχείο καταγραφής

Το μεγάλο αρχείο καταγραφής, αποτελείτε από 2000 εγγραφές, οι οποίες δημιουργήθηκαν με τυχαίο τρόπο, όπως αναφέρθηκε στο κεφάλαιο 5.1. Για την αναπαράσταση του γεωγραφικού μήκους σε κάθε εγγραφή χρησιμοποιήθηκαν τυχαίες τιμές οι οποίες κυμαίνονταν από xL=24.000000, έως xH=48.000000 , ενώ αντίστοιχα για το γεωγραφικό πλάτος yL=-124.000000, έως yH=-69.000000. Οι παραπάνω ελάχιστες και μέγιστες τιμές, επιλέχθηκαν με σκοπό τα ερωτήματα να καλύψουν όλη την περιοχή στην οποία βρίσκονται τα εστιατόρια, για αυτό και συμπίπτουν με αυτές του πίνακα 9.



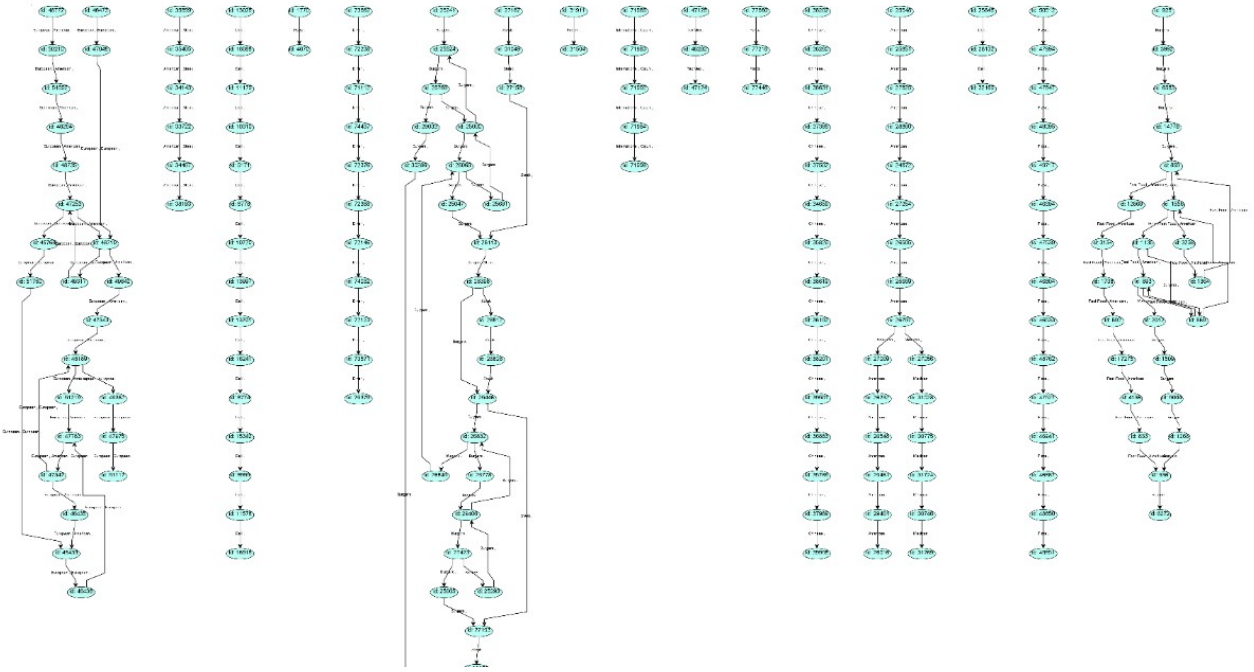
Εικόνα 12: Μερική αναπαράσταση γράφου για k=5

Στην εικόνα 12, παρουσιάζεται μέρος του γραφήματος, το οποίο δημιουργήθηκε, όταν το πλήθος των απαντήσεων που λάμβαναν οι χρήστες έπαιρνε την τιμή 5, δηλαδή είχαμε k=5. Στην πραγματικότητα, το ολοκληρωμένο γράφημα αποτελείται από περίπου 122 υπογραφήματα, ενώ τα μεγαλύτερα από αυτά διακρίνονται στην εικόνα. Επίσης, για την περίπτωση όπου k=5 δημιουργήθηκαν συνολικά 463 κόμβοι και 341 ακμές. Δυστυχώς δεν είναι δυνατή η παράθεση του υπόλοιπου γράφου στο παρόν κείμενο, λόγω του πλήθους των υπογραφημάτων. Παρακάτω ακολουθεί η εικόνα 13, στην οποία παρατίθεται μέρος του παραγόμενου γράφου αλλά αυτήν την φορά για k=10, δηλαδή επιστρέφονται μέχρι 10 αποτελέσματα στον χρήστη, εφόσον αυτά πληρούν τις υπόλοιπες προϋποθέσεις.



Εικόνα 13: Μερική αναπαράσταση γράφου για k=10

Παρατηρούμε ότι το μέγεθος των υπογραφημάτων αυξάνεται σημαντικά, ενώ κάποια υπογραφήματα ενώνονται μεταξύ τους λόγω των επιπλέων ακμών που προστίθενται. Βέβαια και πάλι δεν είναι δυνατή η παράθεση ολόκληρου του γράφου λόγω των περιορισμών που αναφέρθηκαν παραπάνω. Στην περίπτωση που k=10, έχουμε συνολικά 751 κορυφές και 637 ακμές, ενώ το πλήθος των υπογραφημάτων παραμένει υψηλό, με περίπου 120 υπογραφήματα. Έπειτα, ακολουθεί η εικόνα 14 στην οποία παρατίθεται μέρος του γράφου για k=15.



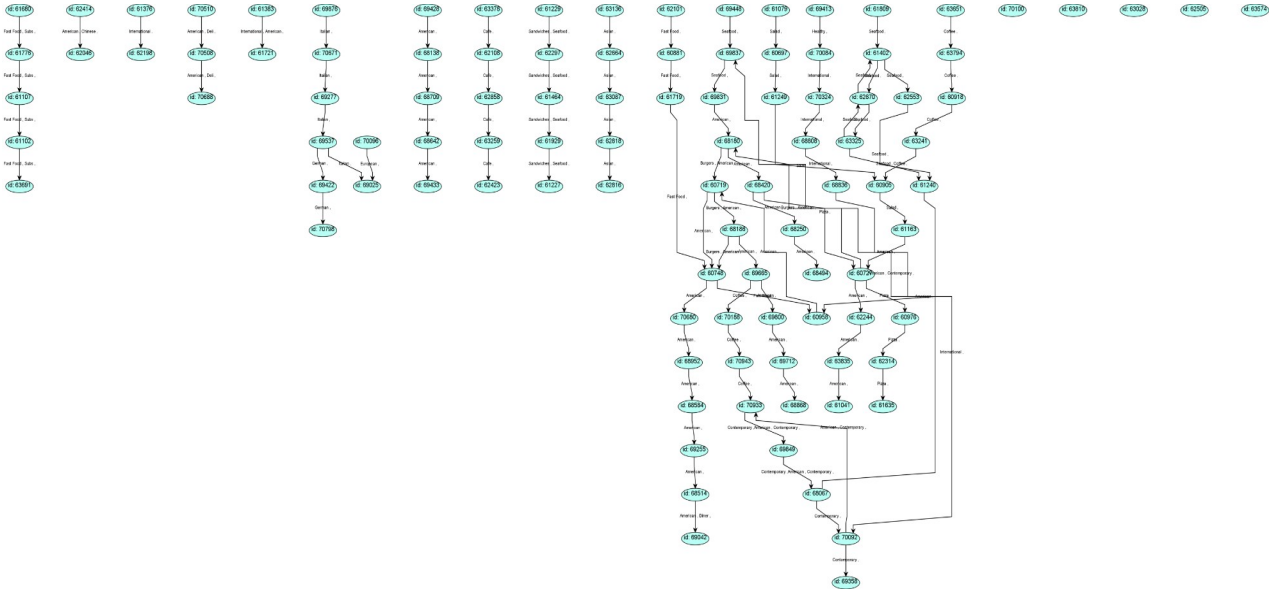
Εικόνα 14: Μερική αναπαράσταση γράφου για k=15

Στην περίπτωση κατά την οποία έχουμε ορίσει k=15, παρατηρούμε ότι τα υπογραφήματα μεγαλώνουν ακόμη περισσότερο. Επίσης σημειώνεται ότι ο τελικός γράφος αποτελείτε από 975 κορυφές και 875 ακμές, ενώ τα υπογραφήματα συνεχίζουν να είναι πολλά σε πλήθος (περίπου 118).

6.2.2 Αποτελέσματα με μικρότερο αρχείο καταγραφής

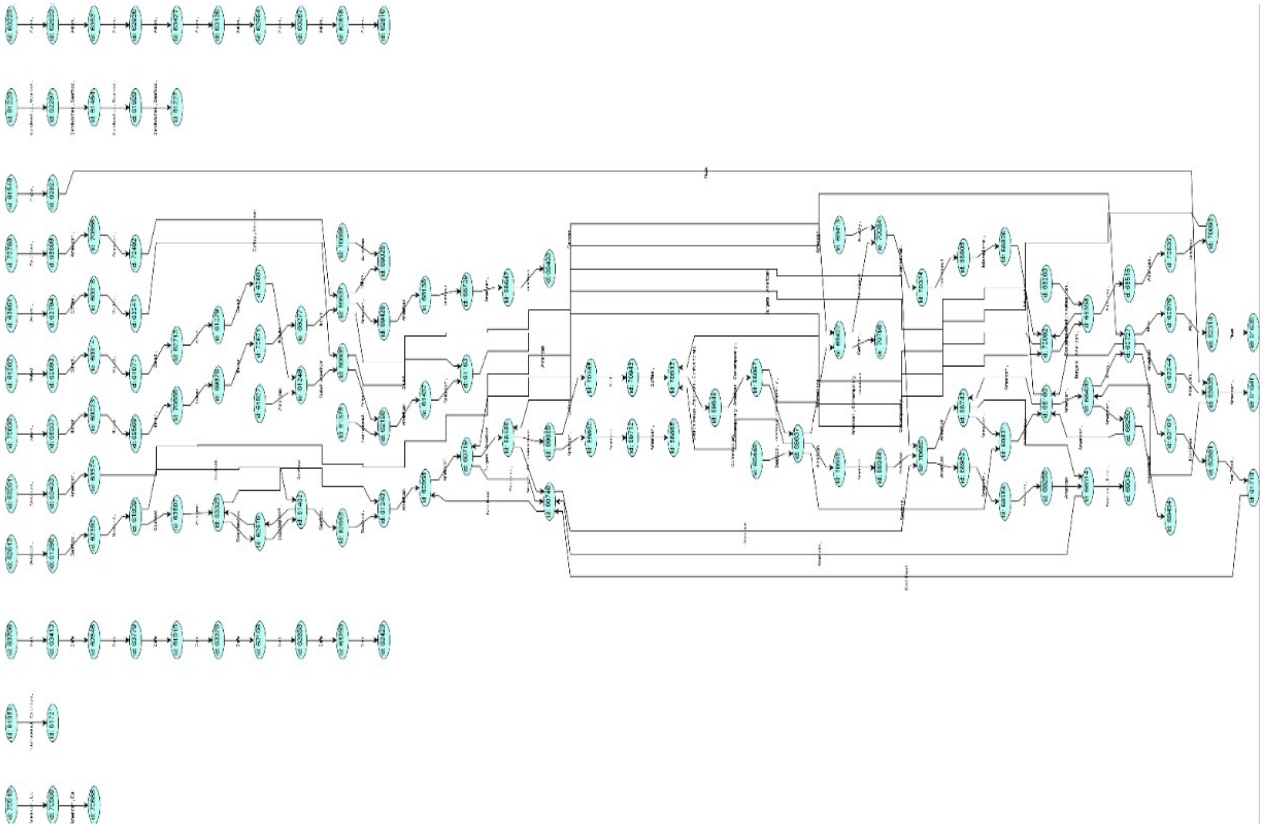
Τα αποτελέσματα που παρουσιάστηκαν παραπάνω δεν ήταν αρκετά για να καταλήξουμε σε ιδιαίτερα χρήσιμα συμπεράσματα. Για τον λόγο αυτό δημιουργήθηκε ένα επιπλέον αρχείο καταγραφής (log file), το οποίο σε αντίθεση με το προηγούμενο θα περιείχε λιγότερες εγγραφές (συνολικά 200), οι οποίες όμως θα προσομοιώνονταν σε ένα σημείο μικρότερης ακτίνας. Έτσι, για την δημιουργία του αρχείου χρησιμοποιήθηκαν τυχαίες τιμές οι οποίες κυμαίνονταν από xL=84.000000, έως xH=85.000000, ενώ αντίστοιχα για το γεωγραφικό πλάτος yL= -83.000000,

έως yH= -82.000000. Όπως και για το προηγούμενο log file, έτσι και τώρα ξεκινάμε παρουσιάζοντας τον γράφο έχοντας ορίσει k=5.



Εικόνα 15: Αναπαράσταση γράφου για k=5

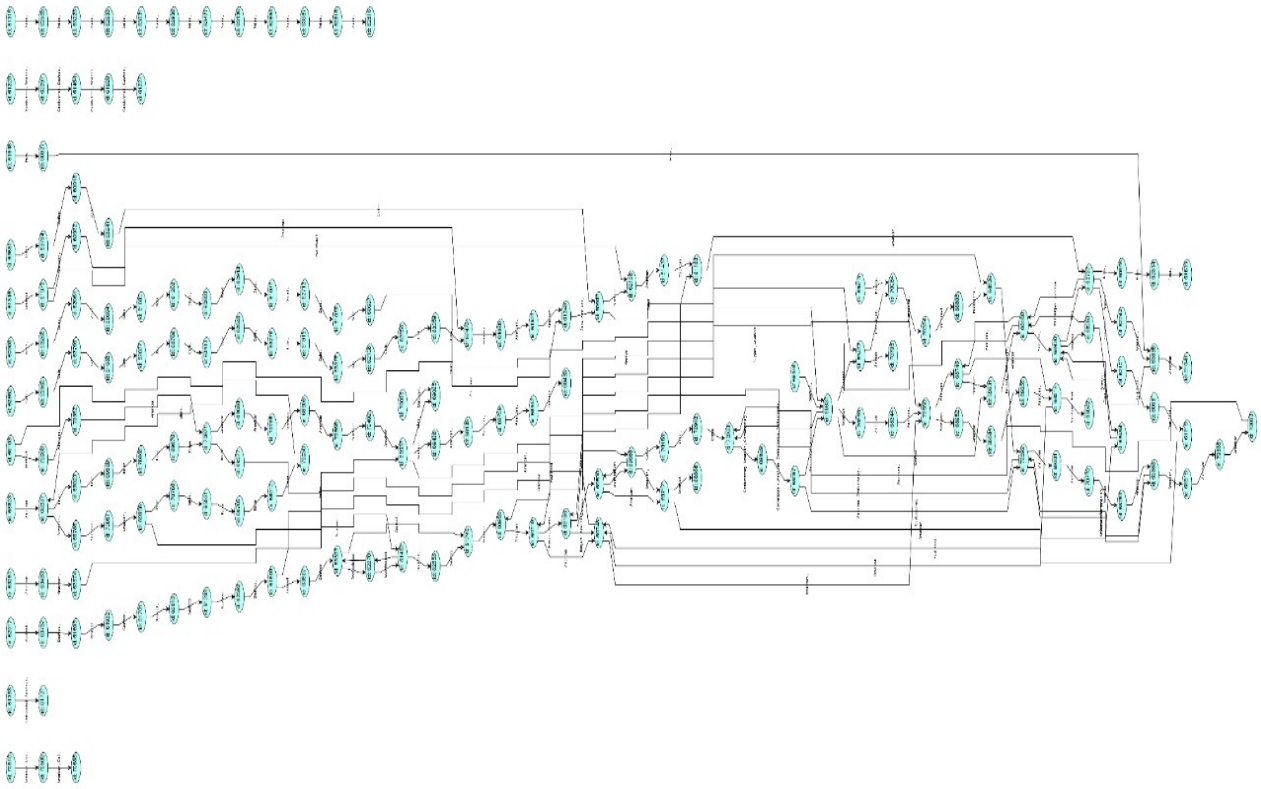
Στην εικόνα 15 παρουσιάζεται ο γράφος που προέκυψε εφαρμόζοντας τον αλγόριθμο. Σε αυτήν την περίπτωση έχουμε συνολικά 105 κόμβους και 97 ακμές, ενώ παρουσιάζεται ολόκληρος ο γράφος, λόγο του μικρού του μεγέθους. Στην εικόνα 16 παρουσιάζεται κομμάτι του γράφου που προκύπτει όταν ο αλγόριθμος εφαρμόζεται για k=10. Αυτή την φορά, ο γράφος παρουσιάζεται σε οριζόντια διάταξη, προκειμένου να γίνουν ορατές οι επιπτώσεις, της αύξησης των επιθυμητών αποτελεσμάτων k.



Εικόνα 16: Αναπαράσταση γράφου για k=10

Ο συνολικός γράφος, το μεγαλύτερο μέρος του οποίου παρουσιάζεται στην εικόνα 16, αποτελείτε από 147 κόμβους και 157 ακμές. Παρατηρούμε ότι υπάρχει σημαντικά μεγαλύτερη συνοχή στον γράφο σε σχέση με τα αποτελέσματα που προέκυπταν κατά την οπτικοποίηση των αποτελεσμάτων του μεγαλύτερου αρχείου καταγραφής ερωτήσεων για k=10.

Συνεχίζοντας με την παρουσίαση του γράφου, εφαρμόζουμε για άλλη μία φορά τον αλγόριθμο και αυτή την φορά ορίζουμε k=15. Η οπτικοποίηση του μεγαλύτερου μέρους του γράφου, παρουσιάζεται παρακάτω στην εικόνα 17, η οποία είναι και πάλι σε οριζόντια διάταξη. Αυτήν τη φορά ο γράφος αποτελείταιαπό181 κόμβους και 206 ακμές. Παρατηρούμε ότι παρόλο που προστίθενται αρκετοί νέοι κόμβοι, η συνοχή του γράφου αυξάνεται, καθώς οι περισσότεροι συμμετέχουν στην δημιουργία του κύριου υπογραφήματος.



Εικόνα 17: Αναπαράσταση γράφου για k=15

6.3 Κατάταξη των δεδομένων

6.3.3 Κατάταξη με Pagerank

Για την κατάταξη των δεδομένων του γράφου έγινε χρήση του αλγόριθμου Pagerank, ο οποίος παρουσιάστηκε στο κεφάλαιο 3.1.2. Η εφαρμογή του αλγόριθμου έγινε ενδεικτικά για τον γράφο της Εικόνας 13, ο οποίος αποτελείτε από 751 κορυφές και 637 ακμές. Στον Πίνακα 11 παρουσιάζονται τα 10 καλύτερα εστιατόρια με βάση τον αλγόριθμο Pagerank.

ID	*Όνομα εστιατορίου*	*Pagerank score*
27193	Applebee's	0.004308496404151227
47783	Aurelio's Pizza	0.004153424362815145
25564	McDonald's	0.004030846494367417
47342	Tinley Park The Original Tinley	0.003903825827107113
936	Burger King	0.0037986663909017828
25096	Jack in the Box	0.003788905340747061
6272	The Central Texan BBQ	0.003619682221094393
48189	Kerry Piper Irish Pub	0.003130713487838656
28672	Texas Roadhouse	0.0027856504473967878
27153	Applebee's	0.0027718405359870245

Πίνακας 11: Κατάταξη με Pagerank

6.3.4 Κατάταξη με Weighted Pagerank ή Personalized Pagerank

Εκτός από την κατάταξη των δεδομένων του γράφου με τον αλγόριθμο Pagerank, κρίθηκε χρήσιμη και η εφαρμογή του αλγορίθμου Weighted Pagerank ο οποίος παρουσιάστηκε στο κεφάλαιο 3.1.3. Ο συγκεκριμένος αλγόριθμος εφαρμόστηκε και πάλι για τον γράφο της Εικόνας 13. Επίσης η λέξη κλειδί που χρησιμοποιήθηκε για την ανάθεση των βαρών στις ακμές είναι η λέξη: «American», ενώ τα 10 καλύτερα εστιατόρια με βάση τον Weighted Pagerank παρουσιάζονται στον πίνακα 12.

ID	*Όνομα εστιατορίου*	*Pagerank score*
47783	Aurelio's Pizza	0.005720793323544162
47342	Tinley Park The Original Tinley	0.0051888254092051014
46433	Aurelio's is Pizza	0.0047263464759744675
46438	Aurelio's Pizza	0.004425825242089829
46435	Aurelio's Pizza	0.004415749742355861
27193	Applebee's	0.004261565932372681
25564	McDonald's	0.004025603045385557
25096	Jack in the Box	0.003823174787429326
936	Burger King	0.0038031546069740127
6272	The Central Texan BBQ	0.003635503836118046

Πίνακας 12: Κατάξη με Weighted Pagerank

7 Συμπεράσματα

Η παρούσα διπλωματική εργασία είχε ως σκοπό την παρουσίαση μίας πρωτότυπης ιδέας, η οποία αφορά την δημιουργία ενός γράφου αποτελούμενου από χωρο-κειμενικά δεδομένα, με βάση τα ερωτήματα των χρηστών. Με την δημιουργία ενός τέτοιου γράφου προκύπτουν νέες δυνατότητες για την ανάλυση αντίστοιχων δεδομένων. Ενδεικτικά εφαρμόστηκαν οι αλγόριθμοι κατάταξης Pagerank και Weighted Pagerank, ενώ σε μεταγενέστερες εφαρμογές θα μπορούσε να εφαρμοστεί οποιοσδήποτε αλγόριθμος σχετίζεται με γράφους. Ένας ιδιαίτερα σημαντικός παράγοντας για την κατασκευή του γράφου, είναι τα δεδομένα τα οποία αναπαριστούν τις ερωτήσεις των χρηστών. Στην περίπτωσή μας, το αρχείο καταγραφής ερωτήσεων δημιουργήθηκε με τυχαίο τρόπο, με αποτέλεσμα ο τελικός γράφος να μην ανταποκρίνεται στην πραγματικότητα.

Βιβλιογραφία

Sciller J., and Voisard A. (2004), Location-Based Services, 10-26.

Muhammed Miah, Gautam Das, Vagelis Hristidis, Heikki Mannila: Standing Out in a Crowd: Selecting Attributes for Maximum Visibility. ICDE 2008: 356-365

Akrivi Vlachou, Christos Doulkeridis, Kjetil Nørvåg, Yannis Kotidis: Identifying the Most Influential Data Objects with Reverse Top-k Queries. Proc. VLDB Endow. 3(1): 364-372 (2010)

Felipe, Ian & Hristidis, Vagelis & Rishe, N.. (2008). Keyword Search on Spatial Databases. Proceedings - International Conference on Data Engineering. 656 - 665. 10.1109/ICDE.2008.4497474.

Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

Xing, W., & Ghorbani, A. (2004). Weighted PageRank algorithm. Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. doi:10.1109/dnsr.2004.1344743

Cao, X., Cong, G., Jensen, C. S., & Ooi, B. C. (2011). Collective spatial keyword querying. Proceedings of the 2011 International Conference on Management of Data - SIGMOD '11. doi:10.1145/1989323.1989363

Zhang, D., Chee, Y. M., Mondal, A., Tung, A. K. H., & Kitsuregawa, M. (2009). Keyword Search in Spatial Databases: Towards Searching by Document. 2009 IEEE 25th International Conference on Data Engineering. doi:10.1109/icde.2009.77

Rocha-Junior, J. B., Gkorgkas, O., Jonassen, S., & Nørvåg, K. (2011). Efficient Processing of Top-k Spatial Keyword Queries. Lecture Notes in Computer Science, 205–222. doi:10.1007/978-3-642-22922-0_13

Chen, L., Cong, G., Jensen, C. S., & Wu, D. (2013). Spatial keyword query processing. Proceedings of the VLDB Endowment, 6(3), 217–228. doi:10.14778/2535569.2448955