ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ σε ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ και στα ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ»

Αξιολόγηση Μαρκοβιανών μοντέλων για την ακρίβεια πρόβλεψης ανθρώπινης κινητικότητας μέσα από επισημασμένες τοποθεσίες σε κοινωνικά δίκτυα

ΨΑΛΤΟΠΟΥΛΟΣ ΒΑΣΙΛΕΙΟΣ
ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: 09316050

ΕΠΙΒΛΕΠΩΝ: Διευθυντής Έρευνας ΕΠΙΣΕΥ, Δρ. Άγγελος Αμδίτης

ΑΘΗΝΑ, 11 / 02 / 2019

Περιεχόμενα

Λίστα εικόνων

Λίστα πινάκων

1. Περίληψη

Η κοινοποίηση παρουσίας στα κοινωνικά δίκτυα έχει γίνει εξαιρετικά δημοφιλής τα τελευταία χρόνια. Σε γενικές γραμμές μέσω της κοινοποίησης παρουσίας σε κάποια συγκεκριμένη τοποθεσία παρέχονται στους χρήστες διάφορες σχετικές υπηρεσίες επιτρέποντας τους την αλληλεπίδραση με τις επαφές τους αλλά και το διαμοιρασμό σχετικών πληροφοριών με την τοποθεσία αναφοράς.

Σκοπός της παρούσας εργασίας είναι η αξιολόγηση των διάφορων συνθέσεων του Μαρκοβιανού μοντέλου (Πρωτοτάξιο, Δευτεροτάξιο και Κρυφό) ως προς την ικανότητα πρόβλεψης της επόμενης επισκεπτόμενης τοποθεσίας κάποιου χρήστη σε ένα κοινωνικό δίκτυο - την τοποθεσία δηλαδή στην οποία ο χρήστης θα πραγματοποιήσει την επόμενη κοινοποίηση παρουσίας. Γενικά, τα μοντέλα Markov χρησιμοποιούνται για τη μοντελοποίηση πολλών προβλημάτων. Ανάμεσα στις χρήσεις ξεχωρίζουν για τον αποτελεσματικότητα τους στην μοντελοποίηση προβλημάτων πρόβλεψης.

Ο λόγος της παρούσας διερεύνησης είναι διότι μετά από μελέτες έχει προκύψει ότι γενικώς οι ανθρώπινες συμπεριφορές μπορούν να περιγραφούν με ακρίβεια ως ένα σύνολο δυναμικών μοντέλων που μπορούν να αλληλουχιστούν μέσα από αλυσίδες Markov. Η ανάλυση της απόδοσης και της ικανότητας των μοντέλων θα γίνει στα πλαίσια μιας γενικής αξιολόγησης αποτελεσματικότητας μέσω προσομοιώσεων με πραγματικά δεδομένα. Τα κοινωνικά δίκτυα με χρήση τοποθεσίας μέσω των δεδομένων που συλλέγουν στα αστικά κέντρα, μπορούν να μας βοηθήσουν να κατανοήσουμε τους παράγοντες που διέπουν την ανθρώπινη κινητικότητα. Έτσι, τα πειραματικά μας δεδομένα ελήφθησαν από την βάση δεδομένων του κοινωνικού δικτύου Foursquare από εξωτερικό website και η χρήση τους είναι καθαρά και μόνο ερευνητική χωρίς καμία αλλοίωση και αναδιανομή.

Ο βασικός στόχος μας είναι αφού αξιολογήσουμε τις συνθέσεις των μοντέλων και κρίνουμε ότι τα αποτελέσματα είναι ικανοποιητικά, να προχωρήσουμε στην ανάπτυξη ενός καθολικού μοντέλου που να μπορεί να προβλέψει την επόμενη τοποθεσία που κάποιος χρήστης θα επιλέξει να παρευρεθεί. Αυτό το μοντέλο θα είναι εξαιρετικής σημασίας γιατί μπορεί να χρησιμοποιηθεί σε συστήματα πρότασης τοποθεσιών, διαφημιστικές υπηρεσίες, ταξιδιωτικά πακέτα και σε πληθώρα σχετικών εφαρμογών.

Η πρόβλεψη και η μοντελοποίηση της ανθρώπινης κινητικότητας μπορεί επίσης να βελτιώσει τις υπηρεσίες βασιζόμενες στην τοποθεσία, τα ευφυή συστήματα μεταφορών, τα αστικά συστήματα πληροφορικής κλπ.

2. Abstract

The check-in functionality offered by most social networks has become extremely popular in recent years. In general, by sharing their particular location through a check-in, users are provided with various related services allowing them to interact with their contacts and to share relevant information with reference to their visited location.

The purpose of this assignment is to evaluate the variations and the different synthesis of the Markov model (First, Second, and Hidden) in terms of their ability to predict the next visited location of a user in a social network - the location where the user will check-in next. In general, Markov models are used to model many problems. Among the uses they stand out for their efficiency in modeling forecasting problems.

The reason for this research is because studies have shown that human behaviors can be described as a set of dynamic models that can be sequenced through Markov chains. Analysis of the performance and capability of the models will be performed in the context of a general efficacy assessment with the use of real data. Social networks that offer the check-in functionality collect data in urban centers that can help us understand the factors that affect human mobility. Thus, our experimental dataset was obtained from the social network Foursquare database, taken from an external website and its use is purely about research without any alteration or redistribution.

Our main goal is to evaluate the different model compositions and if we consider the results to be satisfactory, to develop a global model that can predict a certain user's next check-in location. This model will be of the utmost importance because it can be used in site propositions, advertising, travel packages and a variety of related applications.

The prediction and modeling of human mobility can also improve location based services, intelligent transport systems, urban IT systems, etc

3. Εισαγωγή

Η διάσταση της τοποθεσίας και η χρήση της στα κοινωνικά δίκτυα συγκεντρώνει τεράστιο ερευνητικό ενδιαφέρον αφού γεφυρώνει το χάσμα μεταξύ του φυσικού κόσμου και των υπηρεσιών κοινωνικής δικτύωσης στο διαδίκτυο.

Η κοινοποίηση παρουσίας με χρήση τοποθεσίας στα κοινωνικά μέσα έχει γίνει εξαιρετικά δημοφιλής τα τελευταία χρόνια. Σε γενικές γραμμές μέσα απ' αυτή τη δράση παρέχονται στους χρήστες διάφορες υπηρεσίες (σχετικές με την τοποθεσία) επιτρέποντας τους την αλληλεπίδραση με τις άλλους παρευρισκόμενους και το διαμοιρασμό πληροφοριών σχετικών με την τοποθεσία αυτή.

Η τρέχουσα θέση κάποιου μπορεί να προσδιοριστεί από τον ενσωματωμένο δέκτη GPS στο τηλέφωνο, το τάμπλετ, το λάπτοπ ή άλλες λοιπές έξυπνες συσκευές που έχει στην κατοχή του. Ο προσδιορισμός θέσης είναι εφικτός και μέσω των δικτύων κινητής τηλεφωνίας από τη συσκευή με βάση τον κωδικό της συνδεδεμένης κυψέλης και την ισχύ του σήματος.

Οι υπηρεσίες με χρήση τοποθεσίας στα κοινωνικά δίκτυα παράλληλα παρέχουν νέες ευκαιρίες για εφαρμογές έξυπνων τηλεφώνων και διαφημίσεις και συμβάλλουν κατά κόρον στον επιστημονικό κλάδο, για τη μελέτη ανθρωπίνων συμπεριφορών στην εκμάθηση της ανθρώπινης κινητικότητας. Το ιστορικό τοποθεσίας ενός χρήστη στον πραγματικό κόσμο υποδηλώνει, σε κάποιο βαθμό, τα ενδιαφέροντα και τις συμπεριφορές του. Κατά συνέπεια, τα άτομα που παρουσιάζουν παρόμοια ιστορικά επισκέψεων ορισμένων τοποθεσιών είναι πιθανό να έχουν κοινά ενδιαφέροντα και συμπεριφορά.

Η κινητικότητα που παρουσιάζουν οι άνθρωποι φαίνεται να μην είναι τυχαία και γενικά οι δείκτες της ακολουθούν μη-Γκαουσιανές κατανομές. Ένας παράγοντας που καθιστά την κινητικότητα μη τυχαία εντελώς είναι ότι οι άνθρωποι εκτελούν διαφορετικές δραστηριότητες σε διαφορετικές στιγμές της ημέρας, ανάλογα με το αν είναι καθημερινές ή Σαββατοκύριακα και αργίες.

Η πρόβλεψη και η μοντελοποίηση αυτής της κινητικότητας μπορεί να βελτιώσει τις βασιζόμενες στην τοποθεσία υπηρεσίες, τα ευφυή συστήματα μεταφορών, τα αστικά συστήματα πληροφορικής κλπ.

Στην παρούσα εργασία θα προσπαθήσουμε να αξιολογήσουμε κάποια Μαρκοβιανά μοντέλα στην ικανότητα τους να προβλέψουν την επόμενη τοποθεσία που θα βρεθεί κάποιος χρήστης με βάση το ιστορικό των κοινοποιήσεων παρουσίας που έχει κάνει στο παρελθόν σε κάποιο κοινωνικό δίκτυο.

Για να προβλέψουμε σωστά και με σχετικά μεγάλη ακρίβεια την επόμενη τοποθεσία του κάθε χρήστη από τις παρελθοντικές κοινοποιήσεις του επικεντρωθήκαμε στο διαχωρισμό κυρίων κατηγοριών τοποθεσιών. Έτσι σε πρώτο βήμα, προσπαθούμε να εντοπίσουμε την επόμενη κατηγορία στην οποία ο χρήστης θα επιλέξει να κοινοποιήσει την παρουσία του και μετά ψάχνουμε μέσα σε αυτή την κατηγορία για την τοποθεσία που θα γίνει η πράξη.

Εικόνα 1 Σύγχρονές έξυπνες συσκτεύες με χρήση εντοπισμού θέσης GPS

Σκοπός αυτής της μελέτης είναι να παραχθεί ένα γενικό μοντέλο το οποίο θα προβλέπει την επόμενη τοποθεσία ο χρήστης είναι πιο πιθανό να επισκεφτεί ώστε να δημιουργηθεί ένα ευφυές σύστημα προτάσεων τοποθεσιών για επίσκεψη.

Ένα τέτοιο σύστημα, στο οποίο οι ενδιαφερόμενοι θα ήξεραν με αρκετά μεγάλη ακρίβεια τον τρόπο με τον οποίο κάποιος πρόκειται να μετακινηθεί θα έδινε μεγάλες δυνατότητες σε εμπορικά καταστήματα, εστιατόρια, καφετέριες, κέντρα διασκέδασης τη δυνατότητα να προσελκύσουν περισσότερους πελάτες μέσω αυξημένης προβολής, στοχευμένων διαφημίσεων ακόμα και ειδικών προσφορών. Φυσικά οι εφαρμογές ενός τέτοιου συστήματος δεν θα μπορούσαν να είναι μόνο εμπορικές. Άλλες πιθανές εφαρμογές θα μπορούσαν να περιλαμβάνουν την καταγραφή της δυνητικής εξάπλωσης ασθενειών, τον καθορισμό σημείων συμφόρησης τις ώρες αιχμής και να βοηθήσουν στο γενικότερο σχεδιασμό και τη διαχείριση των μεταφορών.

4. Ιστορικό

Σε αυτή την εισαγωγική ενότητα θα προσφέρουμε μια σύντομη παρουσίαση των μελετών που έχουν γίνει με επίκεντρο την ανθρώπινη κινητικότητα με κατάλληλη αξιοποίηση δεδομένων.

Αρχίζοντας θα εξετάσουμε τις κυρίαρχες τάσεις στην μοντελοποίηση της ανθρώπινης κινητικότητας με αναφορά στα μοντέλα βαρύτητας και τα μοντέλα παρεμβατικών ευκαιριών και θα προχωρήσουμε σε μια απόδειξη ότι αυτά τα δύο διαφορετικά στην όψη μοντέλα εν τέλει συγκλίνουν.

Έπειτα θα δούμε τις πιο σύγχρονες προσεγγίσεις στο ερευνητικό αυτό πεδίο πριν τη ευρεία χρήση υπάρχοντων δεδομένων κάνοντας αναφορά στην προσέγγιση της μελέτης ανθρώπινης κινητικότητας με την ανίχνευση σημειωμένων χαρτονομισμάτων.

Συνεχίζοντας θα δούμε την τεράστια σημασία της χρήσης γεωγραφικών δεδομένων χρηστών κοινωνικών δικτύων όπως η κοινοποίηση παρουσίας αλλά και η ύπαρξη μιας πολύ μεγάλης βάσης δεδομένων τοποθεσιών.

Τέλος θα ασχοληθούμε με την παρουσίαση των παραγόντων που επηρεάζουν την κινητικότητα των χρηστών και θα διακρίνουμε συγκεκριμένα χαρακτηριστικά που αφορούν τους χρήστες όπως οι ιστορικές επισκέψεις, οι κατηγοριακές προτιμήσεις και οι γεωγραφικές αποστάσεις και θα δούμε ότι υπάρχουν και χαρακτηριστικά που επηρεάζουν γενικότερα την κινητικότητα και είναι κατά βάση χρονοεξαρτόμενα, διακρίνοντας ωρολογιακή κατηγορία χαρακτηριστικών και ημερολογιακή κατηγορία.

4.1. Κυρίαρχες τάσεις στην μοντελοποίηση της ανθρώπινης κινητικότητας

Πριν προχωρήσουμε σε περιγραφή των μοντέλων ανθρώπινης κινητικότητας, θα κάνουμε μια εισαγωγή στο ευρύτερο πρόβλημα.

Σε γενικές γραμμές, οι μελετητές κινητικότητας προσπαθούν να καταγράψουν τις στατιστικές ιδιότητες των μετακινήσεων, με συγκεκριμένο δεδομένο χώρο στον οποίο παρουσιάζεται κινητικότητα. Σε αυτό το πλαίσιο κύριος στόχος είναι η πρόβλεψη του αριθμού (ή του κλάσματος) των κινήσεων μεταξύ μιας αφετηρίας και ενός προορισμού.

Ευρύτερα, δεδομένου ενός συνόλου σημείων αφετηριών O και ενός συνόλου σημείων προορισμών D, ο στόχος είναι να κατασκευαστεί ένα μοντέλο που να απαριθμεί με ακρίβεια τον αριθμό των μετακινήσεων (ισοδύναμες μεταβάσεις) μεταξύ ενός σημείου i του O και ενός σημείου j του D.

Εφαρμογές αυτών των θεωρητικών μοντέλων συναντώνται στον πολεοδομικό σχεδιασμό, όπου οι αφετηρίες και οι προορισμοί συνήθως αντιστοιχούν στον τόπο κατοικίας και τους χώρους εργασίας που βρίσκονται στις γεωγραφικές ζώνες i και j αντίστοιχα. Συμπληρωματικά στη θεωρία της μετανάστευσης, κυρίαρχος στόχος είναι να μοντελοποιηθεί ο αριθμός των ατόμων που μεταναστεύουν από τη μια χώρα στην άλλη ή ισοδύναμα από μια πόλη στην άλλη ανάλογα με τη γεωγραφική κλίμακα στην οποία μελετάται η κίνηση.

4.1.1. Μοντέλα βαρύτητας

Όπως συμβαίνει και με τον νόμο της παγκόσμιας έλξης της Νευτώνιας Φυσικής, τα μοντέλα βαρύτητας προσπαθούν να μοντελοποιήσουν τη ροή κινητικότητας μεταξύ μιας αφετηρίας και ενός προορισμού ανάλογα με τις μάζες τους και αντιστρόφως ανάλογα με την απόστασή τους.

Γενικεύοντας, ορίζουμε δύο αντικείμενα i και j με αντίστοιχες μάζες m_i και m_j και γεωγραφική απόσταση d_{ij}, τότε η δύναμη έλξης μεταξύ i και j, που δίνεται από το F_{ij} και ισούται με:

$$F_{ij} = \gamma \frac{m_i m_j}{d_{ij}^2}$$

Όπου γ είναι μια σταθερά εξαρτώμενη από τα δεδομένα. Η ανάλογη διατύπωση στο πλαίσιο μοντελοποίησης μεταφορών θα ήταν

$$T_{ij} = k \frac{O_i O_j}{d_{ij}^2}$$

για ένα σύνολο αφετηριών O και ένα σύνολο προορισμών D και όπου k είναι και πάλι μια σταθερά.

4.1.2. Μοντέλα παρεμβατικών ευκαιριών

Παρά την κομψή μαθηματική μορφή των μοντέλων βαρύτητας, υπάρχει ένα στοιχείο αμφισβήτησης τους κατά τη διατύπωση της θεωρίας. Θα μπορούσαν οι άνθρωποι να κινούνται σαν μικρά σωματίδια των οποίων η συμπεριφορά απλώς διέπεται από τους Νευτώνειους νόμους της βαρυτικής έλξης;

Τον Δεκέμβριο του 1940, ο Samuel Stouffer δημοσίευσε ένα έργο στο οποίο προσπάθησε να εξηγήσει τη σχέση μεταξύ της ανθρώπινης κινητικότητας και της απόστασης με όρους που έθεσαν την ανθρώπινη λήψη αποφάσεων, τους κοινωνικούς παράγοντες και τη γνώση στο επίκεντρο της κινητικής διαδικασίας.

Η θεωρία που είναι γνωστή ως θεωρία παρεμβάσεων αναφέρει ότι:
Ο αριθμός των ατόμων που μετακινούνται σε μια δεδομένη απόσταση είναι ευθέως ανάλογος με τον αριθμό των ευκαιριών σε αυτή την απόσταση και αντιστρόφως ανάλογος με τον αριθμό των παρεμβατικών ευκαιριών.

Η θεωρία των παρεμβατικών ευκαιριών δοκιμάστηκε εμπειρικά σε ένα σύνολο δεδομένων απογραφής που περιγράφει το μεταναστευτικό κίνημα των οικογενειών του Cleveland, Ohio των Ηνωμένων Πολιτειών. Στη μελέτη, η χωρική κατανομή των ευκαιριών απασχόλησης ήταν γνωστή και ο Stouter έδειξε πώς η κινητικότητα των οικογενειών καθοδηγήθηκε από αυτή τη κατανομή.
Στη γενική ιδέα, δεδομένης της γεωγραφικής θέσης της κατοικίας μιας οικογένειας στο Cleveland, σχηματίζεται γύρω της μια ομάδα ομόκεντρων ζωνών και καταγράφεται η τοποθεσία των θέσεων εργασίας σε αυτές τις ζώνες. Η πιθανότητα μετανάστευσης σε μια ζώνη που περιέχει μιας τοποθεσία επιθυμητής εργασίας είναι αντιστρόφως ανάλογη με το άθροισμα των διαθέσιμων τοποθεσιών θέσεων εργασίας στις ζώνες μεταξύ της περιοχής αφετηρίας και της ζώνης προορισμού.

4.1.3. Μοντέλα βαρύτητας και παρεμβατικών ευκαιριών: Η απόδειξη σύγκλισης

Η θεμελιώδης διαφορά μεταξύ των δύο μοντέλων είναι μοντέλο παρεμβατικών ευκαιριών λαμβάνει ρητά υπόψη την ύπαρξη ευκαιριών μεταξύ της αφετηρίας και του προορισμού σε μια δυνητική μετακίνηση ενώ στα μοντέλα βαρύτητας υποθέτουμε ότι σημασία έχει μόνο η αλληλεπίδραση μεταξύ των δύο σημείων. Σε γενικές γραμμές όμως, δεχόμαστε ότι η ύπαρξη τρίτων δεν μπορεί να επηρεάσει τον όγκο μετακινήσεων T_{ij} μεταξύ μιας προέλευσης O και ενός προορισμού D.
Ο Alan G. Wilson, στο βιβλίο του του 1967, αναγνώρισε τις δυσκολίες του απλού μοντέλου βαρύτητας και την αναδιατύπωσε σε μια πιο γενικευμένη μορφή.

Πρώτον, ανέφερε ότι εάν διπλασιάσουμε τον όγκο των μετακινήσεων με αφετηρία O_i και επίσης διπλασιάσουμε τον αριθμό των προορισμών εργασίας D_i, ο αριθμός των ταξιδιών T_{ij} θα πρέπει να τετραπλασιάσει, ενώ προηγουμένως προσδοκούσαμε να διπλασιαστεί. Για να εκφράσει αυτό το επιχείρημα επίσημα, ο Wilson πρόσθεσε τους ακόλουθους περιορισμούς:

$$\sum_j T_{ij} = O_i$$

$$\sum_i T_{ij} = D_j$$

πράγμα που εγγυάται ότι ο αριθμός των μετακινήσεων που ξεκινούν από τη ζώνη αφετηρίας O_j και ο αριθμός των μετακινήσεων που προσελκύονται από τη ζώνη προορισμού D_j συνοψίζονται σωστά όταν δύο σταθερές A_i και B_j συνδεθούν με τις περιοχές εκκίνησης και προσέλκυσης μετακίνησης αντίστοιχα. Επιπλέον, ο Wilson παρείχε μια γενικότερη απεικόνιση της επίδρασης της απόστασης στην μοντελοποίηση των μεταφορών θεωρώντας μια συνάρτηση $f(d_{ij})$ η οποία προσαρμόζεται ανα περιοχή και χαλαρώνει την υπόθεση ότι το T_{ij} είναι αντιστρόφως ανάλογο προς το d_{ij} που με τη σειρά του υψώνεται στο τετράγωνο. Η νεοεισαχθείσα μαθηματική διατύπωση του μοντέλου βαρύτητας γίνεται:

$$T_{ij} = A_i B_j O_i D_j f(d_{ij})$$

Με

$$A_i = [\sum_j B_j D_j f(d_{ij})]^{-1}$$

Και

$$B_j = [\sum_i A_i O_i f(d_{ij})]^{-1}$$

Αυτό όχι μόνο προσέφερε μια γενικότερη ενσωμάτωση της θεωρίας της βαρύτητας στα υπάρχοντα μοντέλα μεταφοράς, αλλά επίσης απέδειξε ότι τα δύο μοντέλα ήταν στατιστικά ισοδύναμα.

4.1.4. Σύγχρονες προσεγγίσεις στη μοντελοποίηση ανθρώπινης κινητικότητας

Στο πρόσφατο παρελθόν λόγω της έλλειψης κινητών συσκευών για την μελέτη της ανθρώπινης κινητικότητας οι επιστήμονες χρησιμοποίησαν έναν πολύ έξυπνο τρόπο για τη μελέτη της ανθρώπινης κινητικότητας. Η δημιουργική ιδέα περιλάμβανε την ανίχνευση των σημειωμένων χαρτονομισμάτων που αντάλλασσαν οι άνθρωποι. Ειδικότερα, ακολούθησαν τις διαδρομές 464 χιλιάδων δολαρίων, με περίπου 1 εκατομμύριο αναφορές των παρατηρήσεών να έχουν υποβληθεί σε μια ειδικά κατασκευασμένη ιστοσελίδα. Η κύρια παραδοχή των συγγραφέων ήταν ότι η κίνηση του δολαρίου αντιπροσωπεύει μια μετεξέλιξη των ανθρώπινων κινήσεων (οι άνθρωποι χρησιμοποιούν τα χαρτονομίσματα στο χώρο) και επομένως αναμένεται να διέπονται από παρόμοιους στατιστικούς νόμους.

Για κάθε ζεύγος διαδοχικών αναφορών σε θέσεις x_i και x_j ενός συγκεκριμένου χαρτονομίσματος δολαρίου, μετρήθηκε η αντίστοιχη γεωγραφική απόσταση $\Delta r = |x_j - x_i|$. Ακολούθως, οι συγγραφείς μέτρησαν την πυκνότητα

πιθανότητας του Δr, $P(\Delta r)$ και διαπίστωσαν ότι η ακόλουθη σχέση παρουσιάζει μια καλή προσέγγιση για τη κατανομή των δεδομένων θέσης:

$$P(\Delta r) \propto \Delta r^{-\beta}$$

με $\beta = 1.59 \pm 0.02$. Υποστηρίχθηκε λοιπόν ότι η κατανομή των ανθρώπινων μετακινήσεων ακολουθεί επίσης μια κατανομή συγκεκριμένης δύναμης. Παρατηρούμε ότι το $P(\Delta r)$ ακολουθεί το νόμο της βαρύτητας, με την απόσταση έχει μια αποσυνθετική επίδραση, και τις μάζες των μετακινήσεων μεταξύ αφετηρίας και προορισμού να θεωρούνται ενιαίες.
Παρουσιάζει λοιπόν μια λογική υπόθεση, δεδομένου ότι δεν υπάρχουν σταθερά σύνολα αφετηριών O και προορισμών D που αντιστοιχούν σε διαφορετικές γεωγραφικές ζώνες όπως αναφέρεται στην αρχική θεωρία της βαρύτητας. Αντ' αυτού, λόγω της χωρικής ακρίβειας των παρατηρήσεων των δολαρίων που καταγράφηκαν με συντεταγμένες γεωγραφικού πλάτους και μήκους, η μάζα μιας δυνητικής αφετηρίας (ή του προορισμού) μπορεί να θεωρηθεί ότι είναι ίση με 1.0 και έτσι, στη συνέχεια καταλήγουμε στο ότι μόνο η απόσταση έχει σημασία στην κινητικότητα.
Μετά την πρώτη μελέτη μεγάλης κλίμακας για τα μοντέλα κινητικότητας με χρήση δεδομένων από κινητά τηλέφωνα επιβεβαιώθηκε η προαναφερθείσα εξίσωση με παρόμοιο εκθέτη που αυτή τη φορά προέκυψε $\beta = 1.75 \pm 0.15$. Στη μελέτη, η οποία περιλάμβανε δείγμα περίπου 16 εκατομμυρίων κινήσεων, προτάθηκε επίσης μια εκθετική αποκοπή και η εξίσωση τροποποιήθηκε ελαφρώς:

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} e^{-\frac{\Delta r}{\kappa}}$$

Όπου το Δr_0 και κ είναι συγκεκριμένες παράμετροι δεδομένων, στην περίπτωση αυτή $\Delta r_0 = 1{,}5$ km και $\kappa = 400$ km.
Το πιο σημαντικό επίτευγμα είναι ότι με αυτόν τον τρόπο έχουμε καταφέρει να μοντελοποιήσουμε την κινητικότητα σε μεγάλες αποστάσεις.
Στη συνέχεια θα δούμε πώς η νέα γενιά συνόλων δεδομένων τοποθεσίας από κοινωνικά δίκτυα θα μας βοηθήσει να εξετάσουμε την ανθρώπινη κινητικότητα όχι μόνο στην πόλη, αλλά και πώς μπορεί να αποτελέσει την πηγή δεδομένων για την ανάπτυξη νέων εφαρμογών κινητών τηλεφώνων.

4.2. Η σημασία των γεωγραφικών δεδομένων χρηστών κοινωνικών δικτύων για τη μελέτη της ανθρώπινης κινητικότητας

Οι λειτουργίες που προσφέρονται στους χρήστες σχετικά με την επισήμανση τοποθεσίας στην οποία βρίσκονται σε δεδομένη χρονική στιγμή δεν είναι παρά μια αναμενόμενη εξέλιξη των παλαιότερων διαδικτυακών κοινωνικών δικτύων. Συγκεκριμένα υπάρχουν κοινωνικά δίκτυα που βασίζουν την ύπαρξη τους στην κοινοποίηση παρουσίας και την επισήμανση τοποθεσίας.
Σε αυτή την περίπτωση οι τοποθεσίες είναι το κύριο σημείο εστίασής τους, γύρω από το οποίο δημιουργείται ένα σύστημα που αναπαριστά το κοινωνικό δίκτυο των χρηστών. Χαρακτηριστικό παράδειγμα τέτοιου δικτύου από το οποίο αντλήθηκαν και δεδομένα για να γίνει η παρούσα μελέτη είναι το Foursquare. Εκτός από τις παραδοσιακές επιλογές σύνδεσης δύο χρηστών που υπάρχουν στα κοινωνικά δίκτυα, όπως «μηχανισμοί σύστασης», «προτεινόμενοι φίλοι», «αναζήτηση χρήστη με βάση το όνομα ή κριτήρια», το Foursquare δίνει τη

δυνατότητα στους χρήστες να συνδεθούν ως φίλοι σε μέρη στα οποία κοινοποιούν την παρουσία τους.

Εικόνα 2 Λογότυπο πλατφόρμας κοινωνικής δικτύωσης Foursquare

Πιο συγκεκριμένα, ένας χρήστης μπορεί να αναζητήσει άτομα που έχουν κοινοποιήσει την παρουσία τους σε ένα μια συγκεκριμένη τοποθεσία και στη συνέχεια να τους στείλει αίτημα φιλίας με σκοπό να συνδεθεί μαζί τους.
Αυτή η διαφορετική δυνατότητα που προσφέρει στους χρήστες του το Foursquare το καθιστά ως ένα ελκυστικό μέσο για τους ερευνητές που ενδιαφέρονται για την αλληλεπίδραση της ανθρώπινης κινητικότητας και της κοινωνικής αλληλεπίδρασης.
Το Foursquare αυτή τη στιγμή είναι η πιο δημοφιλής βασιζόμενη στην τοποθεσία υπηρεσία και αριθμεί περισσότερους από 35 εκατομμύρια ενεργούς χρήστες.
Πρόσφατα έχουν εμφανιστεί πολλά μοντέλα στη βιβλιογραφία που προσπαθούν να προβλέψουν την ανθρώπινη κινητικότητα εκμεταλλευόμενα τους προερχόμενους από την πλατφόρμα γράφους κοινωνικών συνδέσεων και αντιστρόφως, μοντέλα που χρησιμοποιούν τις δεδομένες προτιμήσεις κίνησης και γεννούν μηχανισμούς προτάσεων για τη δημιουργία κοινωνικού δικτύου.
Θα δούμε τα κύρια χαρακτηριστικά λειτουργίας τους Foursquare που το καθιστούν μια αξιοσημείωτη πλατφόρμα για εξέταση.

- **Κοινοποίηση παρουσίας**. Αυτή είναι η κεντρική έννοια. Όταν ένας χρήστης βρίσκεται σε ένα μέρος, μπορεί να επικοινωνήσει με κάποιον άλλον μέσω της δυνατότητας κοινοποίησης παρουσίας. Συνήθως χρησιμοποιείται ο αισθητήρας GPS του κινητού τηλεφώνου για να παρέχει στην υπηρεσία πληροφορίες σχετικά με τη γεωγραφική του θέση ο οποία κωδικοποιείται μέσω συντεταγμένων γεωγραφικού πλάτους και γεωγραφικού μήκους. Η εφαρμογή στη συνέχεια αναζητά μέσα από μια βάση δεδομένων με εκατομμύρια εγγεγραμμένους χώρους από όλο τον κόσμο και επιστρέφει στον χρήστη μια λίστα με κοντινά μέρη. Στη συνέχεια, ο χρήστης ελέγχει στο χώρο όπου βρίσκεται και τον πραγματοποιεί την κοινοποίηση παρουσίας.
- **Βάση δεδομένων τοποθεσιών**. Αυτό που κάνει τις εγγραφές τοποθεσιών του Foursquare τόσο ξεχωριστές είναι τα επίπεδα πληροφοριών που σχετίζονται με αυτές.
  Σε κάθε τοποθεσία οι χρήστες μοιράζονται εκτός από την κοινοποίηση τους αρκετές σημασιολογικές πληροφορίες π.χ. Κινέζικο Εστιατόριο ή θερινός κινηματογράφος, αλλά και γλωσσικό περιεχόμενο όπως συμβουλές, σχόλια ή ετικέτες καθώς και περιεχόμενα πολυμέσων που περιλαμβάνουν φωτογραφίες και βίντεο. Επίσης διατίθενται και πληροφορίες σε πραγματικό χρόνο, όπως ο αριθμός των ατόμων που

επισκέφθηκαν κάποια τοποθεσία κατά τη διάρκεια της τελευταίας ώρας ή ακόμα και ειδικές προσφορές για τακτικούς πελάτες.
Συνολικά, η υπηρεσία απαριθμεί πάνω από 50 εκατομμύρια επισημασμένες τοποθεσίες παγκοσμίως, οι οποίες καλύπτουν γεωγραφικά τις περισσότερες χώρες και ηπείρους. Αυτό το σετ είναι αποτέλεσμα μιας βάσης δεδομένων που ξεκίνησε περιλαμβάνοντας κάποιες εκατοντάδες σημεία ενδιαφέροντος (POI) όταν τέθηκε πρώτη φορά σε λειτουργία το Foursquare και έκτοτε έχει αυξηθεί μέσω του crowdsourcing με τους χρήστες να προσθέτουν νέους χώρους κάθε μέρα.

- **Παγκόσμια προσβασιμότητα**. Στην εφαρμογή μπορεί να έχει πρόσβαση οποιοσδήποτε σε οποιοδήποτε μέρος του κόσμου με μοναδικό προαπαιτούμενο την παρουσία σύνδεσης στο Διαδίκτυο. Με αυτές τις πληροφορίες διαθέσιμες σε παγκόσμιο επίπεδο για πρώτη φορά δίνεται η ευκαιρία να μελετηθεί η κινητικότητα κίνηση σε μεγάλες κλίμακες (για παράδειγμα σε όλες τις χώρες).
- **Δημόσια διαθεσιμότητα**. Μέσω των API του Foursquare μπορούμε να συλλέξουμε δεδομένα σχετικά με την κινητικότητα των χρηστών όμως εδώ είναι σημαντικό να σημειωθεί ότι λαμβάνουμε μόνο δεδομένα τα οποία οι χρήστες έχουν προσφέρει ως δημόσια, ενώ δεν υπάρχει καμία πρόσβαση σε δεδομένα που οι χρήστες έχουν ορίσει ως ιδιωτικά. Έτσι, τα δεδομένα είναι διαθέσιμα στους ακαδημαϊκούς ερευνητές και, ως εκ τούτου, είναι πιθανό να προκύψουν νέες αναλύσεις, τεχνικές και μοντέλα (ερευνητική καινοτομία) και, η έρευνα μπορεί να εξεταστεί από άλλους ερευνητές κατά τη δημοσίευση (αναπαραγωγικότητα).

4.3. Παράγοντες που επηρεάζουν την κινητικότητα των χρηστών

Εξετάζοντας το πρόβλημα της ανθρώπινης κινητικότητας με τη χρήση δεδομένων χρηστών κοινωνικών δικτύων και τις κοινοποιημένες τοποθεσίες τους παρατηρούμε ότι η δραστηριότητα τους επηρεάζεται από κάποιο αριθμό παραγόντων. Πριν ξεκινήσουμε την μοντελοποίηση της κινητικότητας των χρηστών όμως, θα πρέπει να μιλήσουμε για αυτούς τους παράγοντες, να τους αναλύσουμε και να τους κατανοήσουμε.

4.3.1. Χαρακτηριστικά κινητικότητας χρηστών

Όπως είναι γενικά γνωστό, υπάρχουν ορισμένοι παράγοντες που επηρεάζουν το χρήστη θετικά ή αρνητικά ώστε να αποφασίσει σχετικά με το αν θα επισκεφτεί ξανά κάποια συγκεκριμένη τοποθεσία που έχει επισκεφτεί στο παρελθόν.

Σε αυτή την κατηγορία αναφερόμαστε σε χαρακτηριστικά σχετικά με τις κοινοποιήσεις παρουσίας που παράγονται από δεδομένο χρήστη του οποίου θέλουμε να προβλέψουμε τη συμπεριφορά ή από το ίδιο το κοινωνικό του δίκτυο.

Ο απώτερος στόχος μας μέσα από αυτή τη μελέτη είναι να καταγράψουμε τις πιθανότητες με τις οποίες κάποιος μπορεί να επιστρέψει σε ένα μέρος που έχει επισκεφτεί στο παρελθόν, αλλά και τις προτιμήσεις του όσον αφορά τους τύπους τοποθεσιών που του αρέσει να παρευρίσκεται.

4.3.1.1. Ιστορικές επισκέψεις

Προηγούμενες μελέτες σχετικές με την πρόβλεψη και τη μοντελοποίηση της ανθρώπινης κινητικότητας έδειξαν ότι οι μελλοντικές κινήσεις των ανθρώπων επηρεάζονται έντονα από την παρελθοντική κινητική τους συμπεριφορά.

Μετρώντας τον αριθμό επισκέψεων του χρήστη u στο σε κάποια συγκεκριμένη τοποθεσία k μέσα από τις κοινοποιήσεις παρουσίας του, θα προσπαθήσουμε να υπολογίσουμε με την παρακάτω σχέση σε ποιο βαθμό είναι πιθανό να πραγματοποιήσεις ξανά κοινοποίηση παρουσίας την αμέσως επόμενη φορά στο ίδιο μέρος ή σε κάποιο άλλο που έχει επισκεφθεί στο παρελθόν.

$$\widehat{r_k}(u) = |\{(l,t) \in C_u : t < t' \wedge l = k\}|$$

Με (l,t) να δείχνει μια κοινοποίηση παρουσίας, l την τοποθεσία, t το χρόνο, C_u το σετ των κοινοποιήσεων του χρήστη u και t' τον παρόντα χρόνο πρόβλεψης.

Σε γενικές γραμμές περιμένουμε να δούμε ότι υπάρχει μεγάλη πιθανότητα οι χρήστες να επισκεφτούν νέες τοποθεσίες ωστόσο η πιθανότητα επανεπίσκεψης παρελθοντικών τοποθεσιών όπως προκύπτει παραμένει αρκετά υψηλή (περίπου 30%) κάτι που αποδίδει αρκετά μεγάλη συμμετοχή στο δείγμα των πιθανών προορισμών του χρήστη ειδικά αν εξετάσουμε κάποιον ενεργό.

4.3.1.2. Κατηγοριακές προτιμήσεις

Μια άλλη πηγή πληροφοριών που βασίζεται στην ιστορικά καταγεγραμμένη κινητική συμπεριφορά είναι ο αριθμός των κοινοποιήσεων παρουσίας ενός χρήστη u σε μια συγκεκριμένη τοποθεσία k που ανήκει σε κάποια κατηγορία z, z_k. Με αυτόν τον τρόπο, εξάγουμε κάποια συμπεράσματα ώστε να προσδιορίζουμε τη σημασία των διαφόρων κατηγοριών τοποθεσιών (καφετέρια, θέατρο, γήπεδο, εμπορικό κέντρο, κινηματογράφος κ.λπ.) για ένα συγκεκριμένο χρήστη και τα κατατάσσουμε ανάλογα:

$$\widehat{r_k}(u) = |\{(l,t) \in C_u : t < t' \wedge z_l = z_k\}|$$

Στη συνέχεια ταξινομούμε χώρους που ανήκουν στην ίδια κατηγορία ανάλογα με τη δημοτικότητά, δηλαδή με τον συνολικό αριθμό κοινοποιήσεων παρουσίας. Έτσι, ανάμεσα σε όλες τις καφετέριες για παράδειγμα, εκείνες με τις περισσότερες κοινοποιήσεις παρουσίας κατατάσσονται υψηλότερα.

4.3.1.3. Κοινωνικό φιλτράρισμα

Στη συνέχεια, παρουσιάζουμε πώς μπορούμε να εκμεταλλευτούμε τις γενικότερες πληροφορίες που αφορούν μοτίβα κοινοποιήσεων παρουσίας χρηστών του Foursquare παρατηρώντας κάποιο συγκεκριμένο χρήστη και το κοινωνικό του δίκτυο. Σε αυτήν την κατηγορία θα συμπεριλάβουμε τη δημοτικότητα και τα γεωγραφικά χαρακτηριστικά μαζί με άλλα γενικά χαρακτηριστικά που σχετίζονται με τις μεταβάσεις μεταξύ τοποθεσιών

$$\widehat{r_k}(u) = \sum_{v \in \Gamma_u} |\{(l,t) \in C_v : t < t' \wedge l = k\}|$$

$\widehat{r_k}(u)$ είναι το χαρακτηριστικό τοποθεσίας k του χρήστη u
Γ_u είναι οι φίλοι/δίκτυο του u χρήστη
C_v το σετ των κοινοποιήσεων παρουσίας για την τοποθεσία v

4.3.2. Γενικά Χαρακτηριστικά Κινητικότητας

Η ανθρώπινη δραστηριότητα στα κοινωνικά δίκτυα, δεν βασίζεται μόνο στην ιστορία ή τις πληροφορίες του παρελθόντος. Σε γενικές γραμμές λαμβάνονται υπόψιν και άλλες παράμετροι όπως η δημοτικότητα, η γεωγραφική απόσταση και διάφορα άλλα γενικά χαρακτηριστικά που χρησιμοποιούνται για τη μετάβαση ανάμεσα σε τοποθεσίες.

4.3.2.1. Δημοτικότητα

Ορίζουμε αυτό το χαρακτηριστικό υπολογίζοντας τον συνολικό αριθμό των κοινοποιήσεων παρουσίας που πραγματοποιούνται από κάποιο σύνολο χρηστών U στο γενικό σύνολο δεδομένων μας για μια συγκεκριμένη τοποθεσία k:

$$\widehat{r_k}(U) = \sum_{u \in U} |\{(l,t) \in C_u : t < t' \wedge l = k\}|$$

4.3.2.2. Γεωγραφική Απόσταση

Για να μελετήσουμε την επίδραση της γεωγραφικής απόστασης στις κοινωνικές υπηρεσίες με βάση την τοποθεσία, παίρνουμε την τρέχουσα θέση l' του χρήστη u και μετράμε την απόσταση $dist(l', k)$ από όλες τις υπόλοιπες πιθανές τοποθεσίες μετάβασης με βάση τις γεωγραφικές συντεταγμένες τους. Οι τοποθεσίες κατατάσσονται στη συνέχεια με αύξουσα σειρά.

$$\widehat{r_k}(l') = dist(l', k)$$

4.3.2.3. Απόσταση κατάταξης

Παρόμοια με τη γεωγραφική απόσταση, ορίζουμε απόσταση κατάταξης που μετρά τη σχετική πυκνότητα μεταξύ της τρέχουσας θέσης του χρήστη, l' και όλων των άλλων πιθανών τοποθεσιών. Τυπικά, λαμβάνοντας υπόψη όλα τα μέρη ($l \in L$) ορίζουμε:

$$\widehat{r_k}(l) = |\{(l \in L) : dist(l', w) < dist(l', k)\}|$$

Με απλά λόγια αυτή η σχέση περιγράφει τον αριθμό των τοποθεσιών που βρίσκονται πιο κοντά στο l' από τον προορισμό k. Υποθέτουμε εδώ ότι η κίνηση των ανθρώπων δεν βασίζεται απόλυτα σε αριθμητικές τιμές απόστασης, αλλά στην πυκνότητα των εναλλακτικών επιλογών που βρίσκονται κοντά τους.

4.3.2.4. Μεταβάσεις σχετικές με τη δραστηριότητα.

Υποθέτοντας ότι η διαδοχή των ανθρώπινων δραστηριοτήτων δεν είναι τυχαία, για παράδειγμα μπορούμε να επισκεφτούμε το σούπερ μάρκετ μετά από τη δουλειά ή να επισκεφθούμε ένα ξενοδοχείο μετά από την προσγείωση μας σε κάποιο αεροδρόμιο, γι' αυτό ορίζουμε το αντίστοιχο χαρακτηριστικό που μας επιτρέπει να συλλάβουμε αυτό το σήμα στα δεδομένα των κοινοποιήσεων παρουσίας του Foursquare.

Τυπικά, γράφοντας μια πλειάδα (m, n), τα σημεία $m \in L$ και $n \in L$ που εμφανίζονται σε δύο διαδοχικές κοινοποιήσεις παρουσίας, με τα z_m και z_n να είναι οι αντίστοιχες κατηγορίες τους, έχουμε

$$\widehat{r_k}(l') = |\{(m,n) \in L_c : z_m = z_{l'} \wedge z_n = z_k\}|$$

Με το L_c να υποδηλώνει το σύνολο πλειάδων για τοποθεσίες που εμφανίζονται σε διαδοχικές μεταβάσεις πριν από τον τρέχοντα χρόνο πρόβλεψης t'.

4.3.2.5. Μεταβάσεις τοποθεσιών

Εξ 'ορισμού του προβλήματος πρόβλεψης της επόμενης κοινοποίησης παρουσίας, επιδιώκουμε να προβλέψουμε διαδοχικές μεταβάσεις χρηστών σε

όλες τις τοποθεσίες. Έτσι, δημιουργούμε ένα χαρακτηριστικό που εκμεταλλεύεται άμεσα αυτές τις πληροφορίες, μετρώντας τις άμεσες μεταβάσεις μεταξύ όλων των ζευγών τοποθεσιών στην πόλη. Συνεπώς, το σκορ μιας τοποθεσίας k λαμβάνεται με την απαρίθμηση των παρελθουσών μεταβάσεων που παρατηρούνται από οποιονδήποτε χρήστη από την τρέχουσα θέση του l' στην θέση k, την οποία τυπικά ορίζουμε ως

$$\widehat{r_k}(l') = |\{(m,n) \in L_c : m = l' \wedge n = k\}|$$

4.3.3. Χρονοεξαρτόμενα Χαρακτηριστικά

Στα συστήματα που μελετούν την ανθρώπινη συμπεριφορά ο χρόνος είναι ίσως η σημαντικότερη διάσταση. Οι χρήστες κοινωνικών δικτύων τείνουν να συμπεριφέρονται διαφορετικά ανάλογα με την ημέρα της εβδομάδας και ανάλογα με την ώρα της ημέρας.

Σε αυτό το εδάφιο, θα ορίσουμε χρονικά χαρακτηριστικά που μελετούν πληροφορίες τόσο για τη δραστηριότητα των χρηστών όσον αφορά την επίσκεψη συγκεκριμένων κατηγοριών τοποθεσιών, όσο και για τα χρονικά μοτίβα επισκέψεων σε συγκεκριμένα μέρη.

4.3.3.1. Ωρολογιακή κατηγορία

Συγκεκριμένα, με δεδομένο ότι ο όρος z_k υποδηλώνει τον τύπο της τοποθεσίας προορισμού k, ορίζουμε την δημοτικότητα της ωρολογιακής κατηγορίας ως το άθροισμα των παρελθοντικών κοινοποιήσεων παρουσίας σε μια τοποθεσία τύπου z_k για μια δεδομένη ώρα h της ημέρας.

$$\widehat{r_k}(t') = |\{(l,t) \in C : z_l = z_k \wedge tod(t) = tod(t')\}|$$

με το $tod(t) \in [0, 1, \dots, 24]$ να δίνει μια τιμή που αντιστοιχεί σε μια ώρα της ημέρας στον χρόνο t.

4.3.3.2. Ημερολογιακή κατηγορία

Έπειτα ορίζουμε και την δημοτικότητα ημερολογιακής κατηγορίας ως το άθροισμα των παρελθοντικών κοινοποιήσεων παρουσίας σε μια τοποθεσία τύπου z για μια δεδομένη ώρα h της εβδομάδας.

$$\widehat{r_k}(t') = |\{(l,t) \in C : z_l = z_k \wedge tow(t) = tow(t')\}|$$

με το $tow(t) \in [0, 1, \dots, 167]$ να δίνει μια τιμή που αντιστοιχεί σε μια ώρα της εβδομάδας στο χρόνο t.

4.3.3.3. Τοποθεσία ημέρας

Τέλος, ορίζουμε επίσης τη χρονική δραστηριότητα κοινοποιήσεων παρουσίας σε συγκεκριμένες τοποθεσίες. Μετράμε τον αριθμό των θέσεων των κοινοποιήσεων παρουσίας που έχει γίνει στη διάρκεια μιας ημέρας της εβδομάδας (τοποθεσία ημέρας) ως:

$$\widehat{r_k}(t') = |\{(l,t) \in C : l = k \wedge dow(t) = dow(t')\}|$$

όπου το $dow(t)$ επιστρέφει την ημέρα της εβδομάδας στο χρόνο t.

4.3.3.4. Τοποθεσία ώρας

Παρόμοια ορίζουμε για και τη σχέση που δίνει τον αριθμό των κοινοποιήσεων που έχουν γίνει σε μια τοποθεσία k σε μια δεδομένη ώρα μιας ημέρας (τοποθεσία ώρας), με στόχο να πιάσουμε τα εβδομαδιαία και καθημερινά μοτίβα, αντίστοιχα:

$$\widehat{r_k}(t') = |\{(l,t) \in C : l = k \wedge tod(t) = tod(t')\}|$$

4.3.4. Ανάλυση μοτίβων ανθρώπινης κινητικότητας με βάση χρονικά χαρακτηριστικά

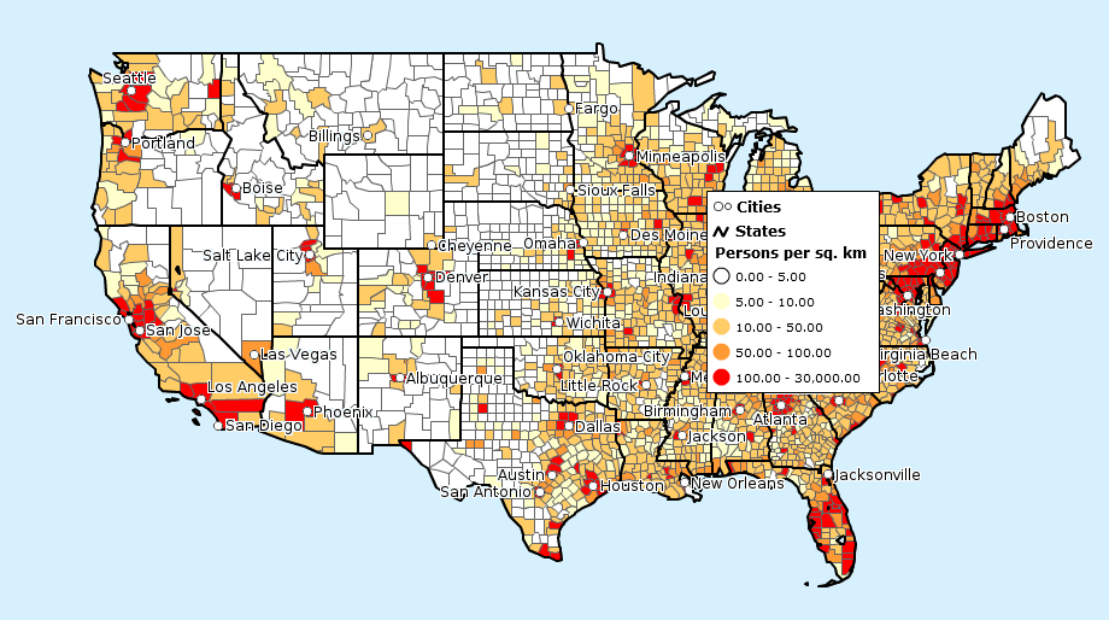
Στο σύνολο δεδομένων που χρησιμοποίησαν οι Dingqi Yang και Daqing Zhang στην μελέτη τους "Detecting Overlapping Communities in Location-Based Social Networks" παρατηρήθηκαν και εξάχθηκαν ορισμένα δυνατά πρότυπα και κανονικότητες σχετικά με την κινητικότητα που παρουσιάζουν οι άνθρωποι με βάση χρονικά χαρακτηριστικά.
Αρχικά μελετήθηκε η συμπεριφορά που παρουσιάζουν οι χρήστες όσον αφορά τις μετακινήσεις τους σε συνάρτηση με τις ημέρες της εβδομάδας.
Η μελέτη αυτή όπως αναμενόταν διαχώρισε δύο κύριες ημερολογιακές κατηγορίες, τις «Καθημερινές» και τα «Σαββατοκύριακα».
Σε αυτές τις μελέτες φάνηκε πως η συμπεριφορά των ανθρώπων είναι διαφορετική όχι μόνο ανάλογα με την ημέρα αλλά και ανάλογα με την ώρα τις καθημερινές και τα Σαββατοκύριακα.
Με τα δεδομένα που πήραμε από το Fourquare, μπορούμε να εξετάσουμε τη συμπεριφορά και την ανθρώπινη δραστηριότητα κατά τη διάρκεια των επτά ημερών της εβδομάδας.
Από το σύνολο δεδομένων, παρατηρήσαμε ότι η πιο δημοφιλής κατηγορία κοινοποιήσεων παρουσίας είναι η κατηγορία "Bar" με 15978 κοινοποιήσεις, με δεύτερη και πολύ κοντά σε αριθμητικές τιμές την κατηγορία "Home (private)" με 15382 κοινοποιήσεις και τρίτη την κατηγορία "Office" με 12740 κοινοποιήσεις.
Όσον αφορά τις ημέρες, βλέπουμε ότι μεγαλύτερη πυκνότητα κοινοποιήσεων παρατηρείται την Παρασκευή και το Σάββατο με 33824 και 34401 κοινοποιήσεις παρουσίας αντίστοιχα. Ακολουθούν η Κυριακή και η Δευτέρα με 32636 και 32486 κοινοποιήσεις αντίστοιχα ενώ ακολουθείται πτωτική πορεία από την Τρίτη ως και την Πέμπτη.
Στην μελέτη που κάναμε για την εξάρτηση του χρόνου με τις κοινοποιήσεις είδαμε ότι οι περισσότερες αριθμητικά κοινοποιήσεις γίνονται στις 23:00, στις 22:00 στις 13:00 και στις 00:00 με τη σειρά και αριθμητικά 15340, 14915, 14209 και 13874 αντίστοιχα. Παρατηρείται δηλαδή ότι οι περισσότερες κοινοποιήσεις παρουσίας γίνονται τις ώρες της νυχτερινής εξόδου, και του μεσημεριανού φαγητού.
Όπως είπαμε και παραπάνω, μια πιο προσεκτική παρατήρηση της χρονικής ανάλυσης των δεδομένων αποκαλύπτει ότι οι άνθρωποι τείνουν να συμπεριφέρονται διαφορετικά και σε διαφορετικές ώρες της ημέρας. Έτσι, με σταθερά την ημέρα, είναι σημαντικό να καταλάβουμε ποιες κατηγορίες γίνονται πιο δημοφιλείς και σε ποια χρονική στιγμή.
Τα αποτελέσματα αυτά ήταν και αναμενόμενα με την γενικότερη εμπειρία και βιβλιογραφία που έχουμε για την ανθρώπινη συμπεριφορά στα κοινωνικά δίκτυα.

4.3.5. Σχετικές εργασίες

Σε αυτό το εδάφιο θα παρουσιάσουμε τις σχετικές εργασίες που έχουν γίνει και αφορούν την πρόβλεψη ανθρώπινης κίνησης. Πρόκειται σε γενικές γραμμές για ένα πολύ καυτό θέμα που απασχολεί την επιστημονική κοινότητα για αρκετά χρόνια και έχει μελετηθεί αρκετές φορές.

Στο άρθρο “Friendship and Mobility: user Movement in Location Based Social Networks (LBSN)” των Jure Laskoveç, Seth A.Myers που δημοσιεύθηκε στις ΗΠΑ διερευνήθηκαν οι επιπτώσεις των κοινωνικών δικτύων και της περιοδικής κίνησης στην ανθρώπινη κινητικότητα. Πιο συγκεκριμένα στο άρθρο χρησιμοποιούνται τρία διαφορετικά σύνολα δεδομένων για τη μοντελοποίηση της κινητικότητας των χρηστών. Αυτά τα σύνολα δεδομένων λήφθηκαν από το Gowalla με 6,4 εκατομμύρια κοινοποιήσεις παρουσίας. Το άλλο σύνολο δεδομένων είναι το Brightike που περιλαμβάνει 4,5 εκατομμύρια κοινοποιήσεις παρουσίας σε δημόσιους χώρους και 900 εκατομμύρια κοινοποιήσεις παρουσίας από ευρωπαϊκούς παρόχους γραμμών.
Στο άρθρο προτείνονται δύο μοντέλα: Το μοντέλο περιοδικής κινητικότητας χρηστών που κατασκευάζεται με βάση το ιστορικό του χρήστη και το μοντέλο κοινωνικής περιοδικής κινητικότητας που βασίζεται στις σχέσεις που έχει ο κάθε χρήστης με τους άλλους χρήστες στο δίκτυο του. Το άρθρο αυτό συνέκρινε την απόδοση αυτών των δύο μοντέλων. Ως εξαγόμενο, πάνω στη μελέτη της κινητικότητα των χρηστών βρέθηκε πως όχι μόνο η περιοδική κινητικότητα αλλά και οι κοινωνικές σχέσεις παίζουν μεγάλο ρόλο.
Η αυξανόμενη δημοτικότητα των υπηρεσιών βασισμένων σε τοποθεσίες, π.χ. Facebook, Instagram, Foursquare, Gowalla κ.λπ., οδηγεί σε έναν νέο χώρο έρευνας και μόνο λίγες μελέτες έχουν γίνει μέχρι τώρα.

Στο άρθρο “Exploring Millions of Footprints in Location Sharing Services” των Cheng, Z., Caverlee, J., Lee, K., Sui D. οι ερευνητές μοντελοποίησαν την ανθρώπινη κινητικότητα μέσα από την ανάλυση χρονικών, χωρικών, κοινωνικών και κειμενικών πληροφοριών των χρηστών κοινωνικών δικτύων. Το σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτή την έρευνα λήφθηκε από το twitter και το Gowalla και αποτελούταν από 225.098 χρήστες και 22 εκατομμύρια κοινοποιήσεις παρουσίας. Με τη διερεύνηση του συνόλου δεδομένων αποδείχθηκε ότι οι χρήστες κατά βάση κοινοποιούν την παρουσία τους σε εστιατόρια, καφετέριες, εμπορικά κέντρα και αεροδρόμια.
Ένα από τα κύρια εξαγόμενα της ανάλυσης είναι ότι οι άνθρωποι που ζουν σε κατοικημένες περιοχές τείνουν να ταξιδεύουν περισσότερο από όσους ζουν σε αραιοκατοικημένες περιοχές. Ωστόσο, οι άνθρωποι που ζουν σε αραιοκατοικημένες περιοχές θεωρείται ότι ταξιδεύουν σε μεγαλύτερες αποστάσεις. Από την άλλη πλευρά, λαμβάνοντας υπόψη το εισόδημα των ατόμων, τα άτομα με υψηλότερο εισόδημα τείνουν να ταξιδεύουν περισσότερο σε απομακρυσμένες περιοχές. Το συμπέρασμα είναι ότι η κοινωνική δημοτικότητα των χρηστών φαίνεται να επηρεάζει την κινητικότητά τους. Ο λόγος του αριθμού των ατόμων που ακολουθούν τον χρήστη και του αριθμού των ατόμων που ακολουθεί ο χρήστης, δείχνει την κοινωνική δημοτικότητα του χρήστη. Σύμφωνα με το αποτέλεσμα της αξιολόγησης της κοινωνικής δημοτικότητας, το μέτρο της δημοτικότητας είναι αναλογικά παράλληλο με το ποσοστό των ταξιδιών.



Εικόνα 3 Πυκνότητα κατοίκων στις ΗΠΑ

Στο άρθρο “Mobile Location Prediction of Spatial-Temporal Context" των Huiji G., Jiliang T. και Huan L., σκοπός της μελέτης είναι να μοντελοποιηθεί η ανθρώπινη κινητικότητα με το χρονικό-ιστορικό πρότυπο και το χρονικό-περιοδικό πρότυπο των χρηστών. Το άρθρο περιγράφει την κινητικότητα των χρηστών με γκαουσιανή κατανομή. Το σύνολο δεδομένων που χρησιμοποιήθηκε δόθηκε από το διαγωνισμό δεδομένων Nokia Mobile και περιλάμβανε ένα έτος δεδομένων 80 χρηστών. Για το σύνολο δοκιμαστικών δεδομένων, χρησιμοποιήθηκαν δεδομένα τοποθεσιών των τελευταίων 50 ημερών και για το σύνολο εκπαίδευσης χρησιμοποιήθηκαν πληροφορίες θέσης 50 ημερών των 80 προαναφερθέντων χρηστών. Εξετάζοντας το σύνολο δεδομένων, διαπιστώνεται ότι σε διαφορετική χρονική περίοδο η κινητικότητα των χρηστών είναι διαφορετική. Για αυτό το λόγο, ο ερευνητής στο άρθρο περιέγραψε την έννοια της ημερήσιας κατανομής και πρότεινε το "Ιεραρχικό Μοντέλο Pitman Yordan Prior Hour-Day (HPHD)".

Γενικά, σε μελέτες προσομοίωσης της ανθρώπινης κινητικότητας σε κοινωνικά δίκτυα, οι τοποθεσίες που έχουν τις λιγότερες κοινοποιήσεις παρουσίας απομακρύνονται από το σύνολο δεδομένων.

Ωστόσο, στο άρθρο “Modelling geo-social correlations for new check-ins on location-based social networks” των Huiji G., Jiliang T. και Huan L., οι ερευνητές προσπάθησαν να αποδείξουν ότι οι κοινοποιήσεις παρουσίας κάποιου χρήστη σε μια τοποθεσία που δεν έχει επισκεφθεί νωρίτερα επηρεάζονται σημαντικά από τις κοινοποιήσεις άλλων χρηστών (στην ίδια τοποθεσία) του δικτύου του χρήστη αναφοράς.

Επιπλέον, για την καλύτερη κατανόηση της κοινωνικής συσχέτισης του χρήστη με την κοινοποίηση παρουσίας σε μια νέα τοποθεσία, ο κοινωνικός συσχετισμός χωρίζεται σε 4 υπο-συσχετισμούς.

Δηλαδή, να αναζητούνται σχέσεις μεταξύ εκείνων που κατοικούν κοντά στο χρήστη και βρίσκονται στη λίστα φίλων του (τοπικοί φίλοι), αυτών που ζουν κοντά στη χρήση, αλλά δεν βρίσκονται στη λίστα φίλων του (τοπικοί, μη φίλοι), αυτών που ζουν μακριά αλλά βρίσκονται στη λίστα φίλων του χρήστη (σε

απόσταση, φίλοι) και σε αυτούς που ζουν μακριά και δεν βρίσκονται στη λίστα φίλων του χρήστη , (σε απόσταση, μη φίλος).
Σχετικά με το σύνολο δεδομένων, χρησιμοποιήθηκαν δεδομένα κοινοποιήσεων από το Foursquare που συλλέχθηκαν μέσω Twitter. Αυτό το σύνολο περιλαμβάνει 1.385.223 κοινοποιήσεις από 11.326 χρήστες. Έτσι το προτεινόμενο μοντέλο γεω-κοινωνικής συσχέτισης του άρθρου, είναι σε θέση να εκτιμά κάποια επόμενη κοινοποίηση παρουσίας ενός χρήστη.

Υπάρχουν γενικά πολλές μελέτες που έχουν θίξει το συγκεκριμένο πρόβλημα. Στο άρθρο " Content-Aware Point of Interest Recommendation on Location-Based Social Networks" των Huiji G., Jiliang T., Xia H. και Huan L, οι συγγραφείς αναφέρονται στις χρονικές επιδράσεις στα κοινωνικά δίκτυα για να προτείνουν κάποια τοποθεσία. Κατασκευάζοντας πρότυπα ημερήσιας και εβδομαδιαίας βάσης ανέπτυξαν ένα πρωτοποριακό μοντέλο πρότασης νέων τοποθεσιών.

Οι ερευνητές του Facebook Backstrom, Sun, και Marlow στο άρθρο τους "Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity" ανέλυσαν την απόσταση των κοινωνικών σχέσεων μεταξύ χρηστών του Facebook και χρησιμοποιούσαν τοποθεσίες των φίλων ενός χρήστη αναφοράς για να προβλέψουν την επόμενη γεωγραφική θέση του. Στο άρθρο " You are where you Tweet: A content-based approach to geo-locating Twitter users" των Cheng, Caverlee, και Lee οι συγγραφείς μοντελοποίησαν τη χωρική κατανομή των λέξεων στο κείμενο που δημιούργησε κάποιος χρήστης στο Twitter για να προβλέψουν την τοποθεσία του.
Στο άρθρο του Hecht, "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles" μελετήθηκε η χωρική συμπεριφορά των χρηστών μέσω των προφίλ τους στο Twitter.
Σε ένα άλλο άρθρο των Gonzalez, Hidalgo, και Barabasi, "Understanding individual human mobility patterns" αναλύθηκαν 100.000 τροχιές χρηστών έξυπνων τηλεφώνων (2008) και αποδείχτηκε ότι η ανθρώπινη κινητικότητα εμφανίζει απλά αναπαραγώγιμα πρότυπα.
Στο άρθρο "Human mobility prediction based on individual and collective geographical preferences" των Calabrese F., Di Lorenzo G. και Ratti C., μελετήθηκε η κινητικότητα των χρηστών με βάση τις μεταβάσεις και με βάση τη δραστηριότητα σε ατομικό επίπεδο για το το σχεδιασμό και τη διαχείριση μεταφορών.
Άλλη μια εφαρμογή πρόβλεψης της θέσης έχει παρουσιαστεί στο άρθρο του Bellotti "Activity-Based Serendipitous Recommendations with the Magitti Mobile Leisure Guide" όπου αναπτύχθηκε το σύστημα συστάσεων Magitti
Τις τελευταίες δύο δεκαετίες έχουν γίνει εκτενείς ερεύνες για τα Μαρκοβιανά μοντέλα. Το άρθρο του L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" παρείχε μια ολοκληρωμένη εικόνα της εφαρμογής του κρυφού μοντέλου Markov και της απόδοσης του στην αναγνώρισης ομιλίας.
Πιο πρόσφατα το άρθρο του R.M. Altman, "Mixed Hidden Markov Models: An Extension of the Hidden Markov Model to the Longitudinal Data Setting" με δυναμική διαρρύθμιση δεδομένων εξέτασε την επέκταση του βασικού κρυφού μοντέλου Markov σε μικτό. Το μικτό HMM κάτω από κάποια τυποποιημένη κατανομή δεδομένων, όπως η γκαουσιανή κατανομή, αποδίδει καλά. Ωστόσο,

αυτά τα μοντέλα, απαιτούν εξαιρετικά εξειδικευμένες λειτουργίες κανονικών συνδέσεων σύμφωνα με τις συγκεκριμένες απαιτήσεις της εκαστοτε κατανομής, κάτι που είναι πολύ σπάνιο συμβεί στον τομέα κοινωνικής δικτύωσης σύμφωνα με το άρθο των Jihang Ye, Zhe Zhu και Hong Cheng., " What's Your Next Move: User Activity Prediction in Location-based Social Networks".
Σε μια άλλη δημοσίευση των Tang J., και Liu H. "Exploring Social-Historical Ties on Location-Based Social Networks" διερευνήθηκε το μοντέλο των εισερχόμενων χρηστών σε κοινωνικά δίκτυα σε σχέση με τους κοινωνικο-ιστορικούς δεσμούς. Πρότειναν ένα κοινωνικο-ιστορικό μοντέλο για να διερευνηθεί η συμπεριφορά των χρηστών στις κοινοποιήσεις παρουσίας εντός του κοινωνικού δικτύου. Το μοντέλο ενσωματώνει τις κοινωνικές και ιστορικές επιπτώσεις και αξιολογεί το ρόλο της κοινωνικής συσχέτισης στη συμπεριφορά του χρήστη για κοινοποίηση. Για να μοντελοποιηθεί η κινητικότητα των χρηστών, χρησιμοποιήθηκε ένα σύνολο δεδομένων από το Gowalla, που περιλαμβάνει 6,4 εκατομμύρια κοινοποιήσεις παρουσίας. Χρησιμοποιήθηκε η πρόβλεψη θέσης ως εφαρμογή για την αξιολόγηση του προτεινόμενου κοινωνικο-ιστορικού μοντέλου και άλλων μεθόδων αναφοράς. Κατέληξαν στο συμπέρασμα ότι το κοινωνικο-ιστορικό συνεχώς υπερέχει των βασικών μεθόδων γιατί λαμβάνει υπόψιν ιστορικούς και κοινωνικούς δεσμούς. Τα πειραματικά αποτελέσματά σχετικά με την πρόβλεψη θέσης δείχνουν ότι η προτεινόμενη προσέγγιση τους καταγράφει καταλλήλως τις ιδιότητες κοινοποίησης παρουσίας του χρήστη και ξεπερνά τα παρόντα μοντέλα.
Στην παρούσα εργασία, τα σκοπούμενα μοντέλα μας λαμβάνουν υπόψιν τις ιστορικές και χρονικές πληροφορίες για τη διερεύνηση της συμπεριφοράς των χρηστών στο κοινωνικό δίκτυο χρησιμοποιώντας μοντέλα Markov (πρωτοτάξιο, δευτεροτάξιο και κρυφό).

Στα άρθρα των Jia-Dong Z. και Chi-Yin C., "GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations", Huiji G., Xia H. και Huan L., Enriching short text representation in microblog for clustering front), Salvatore S. και Anastasios N., "An Empirical Study of Geographic User Activity Patterns in Foursquare" και Samiul H., Xianyuan Z. και Satish V.U., "Understanding Social Influence in Activity Location Choice and Lifestyle Patterns Using Geolocation Data from Social Media" οι ερευνητές χρησιμοποιούν διαφορετικά σύνολα δεδομένων και διαφορετικά πλαίσια (χωρικό, χρονικό και κοινωνικό) και αναπτύσσουν μοντέλα για την ανθρώπινη κινητικότητα. Οι προαναφερθείσες μελέτες μας παρέχουν αρκετές πληροφορίες σχετικά με την εκτίμηση των επιμέρους προτύπων κινητικότητας.

5. Ορισμός του προβλήματος

Λόγω της πρόσφατης εκθετικής ανάπτυξης υπηρεσιών με βάση τη γεωγραφική θέση σε κινητές συσκευές, όπως τα έξυπνα τηλέφωνα, τα έξυπνα ρολόγια και τα tablet, το πρόβλημα της πρόβλεψης της επόμενης θέσης του χρήστη γίνεται ένα σημαντικό ερευνητικό θέμα τόσο στον ακαδημαϊκό χώρο όσο και στον επιχειρησιακό κλάδο.
Το πρόβλημα επικεντρώνεται στην πρόβλεψη επίσκεψης μιας τοποθεσίας ενός χρήστη προτού αυτός φτάσει, με βάση τις παρελθούσες συμπεριφορές του που συνάγονται μέσω της χρήσης αισθητήρων όπως το GPS και το WiFi, που διαθέτουν οι σύγχρονες κινητές συσκευές.
Αν λάβουμε υπόψη και το επίπεδο λεπτομέρειας της γεωγραφικής τοποθεσίας όσον αφορά την πρόβλεψη, καταλαβαίνουμε ότι όσο πιο ακριβές είναι, τόσο περισσότερες είναι οι δυνατότητες ανάπτυξης περαιτέρω εξελιγμένων υπηρεσιών.

Γενικά, το πρόβλημα εκτίμησης θέσης συνοψίζεται ως εξής. Λαμβάνοντας υπόψη τις τροχιές του χρήστη $T = l_0 \rightarrow l_1 \rightarrow l_2 \rightarrow l_3 \dots l_t$, όπου L είναι το σύνολο θέσεων $L = \{L_1, L_2, \dots L_n\}$ και l_i είναι ο χρήστης στην i-τοποθεσία που επισκέφθηκε, θέλουμε να προβλέψουμε το $l_t + 1$ δηλαδή την επόμενη θέση του χρήστη.

Τα μοντέλα Markov φαίνονται μια καλή προσέγγιση για την πρόβλεψη της επόμενης θέσης με βάση το ιστορικό τοποθεσιών. Τυπικά ένα μοντέλο Markov θεωρεί ένα μοτίβο τελευταίων τοποθεσιών επισκέψεων ενός χρήστη για να προβλέψει την επόμενη τοποθεσία. Το μήκος του θεωρούμενου μοτίβου ονομάζεται τάξη. Έτσι, ένα μοντέλο Markov τάξης 3 χρησιμοποιεί τις τρεις τελευταίες επισκεπτόμενες τοποθεσίες.
Για όλα τα μοτίβα, το μοντέλο αποθηκεύει τις πιθανότητες των επόμενων τοποθεσιών που υπολογίζονται από ολόκληρη την ακολουθία των τοποθεσιών που επισκέφθηκε ο χρήστης. Ένα απλό μοντέλο Markov είναι το ονομαζόμενο Markov predictor. Το Markov predictor αποθηκεύει για κάθε μοτίβο τις συχνότητες των επόμενων τοποθεσιών.

Σύμφωνα με την υπάρχουσα βιβλιογραφία που μελετήθηκε, τα μοντέλα Markov υπερτερούν έναντι άλλων εναλλακτικών για την αντιμετώπιση και επίλυση αυτού του προβλήματος.

	Bayesian network	multi-layer perceptron	Elman net	Markov predictor	state predictor	Markov predictor with counter	state predictor with counter
accuracy (%) (quantity (%))	78.82 (89.89)	76.45 ($\approx$ 100)	79.68 (100)	76.53 (90.47)	70.89 (90.47)	81.14 (78.40)	81.88 (74.38)
stability (%)	29.67	32.59	71.57	24.67	29.97	24.95	23.99
learning	fast	slow	slow	fast	fast	fast	fast
relearning	slow	slow	slow	slow	fast	slow	fast
memory (bit)	6,500	3,880	7,215	36,960	2,730	37,380	3,150
computing costs	inefficient chain rule	training until $E < t$, otherwise one propagation	training over many learning cycles, otherwise one propagation	table look-up	table look-up	table look-up	table look-up
modelling costs	medium	high	high	low	low	low	low
expandability	yes	no	no	yes	yes	yes	yes
time prediction	integrated	parallel	parallel	parallel	parallel	parallel	parallel

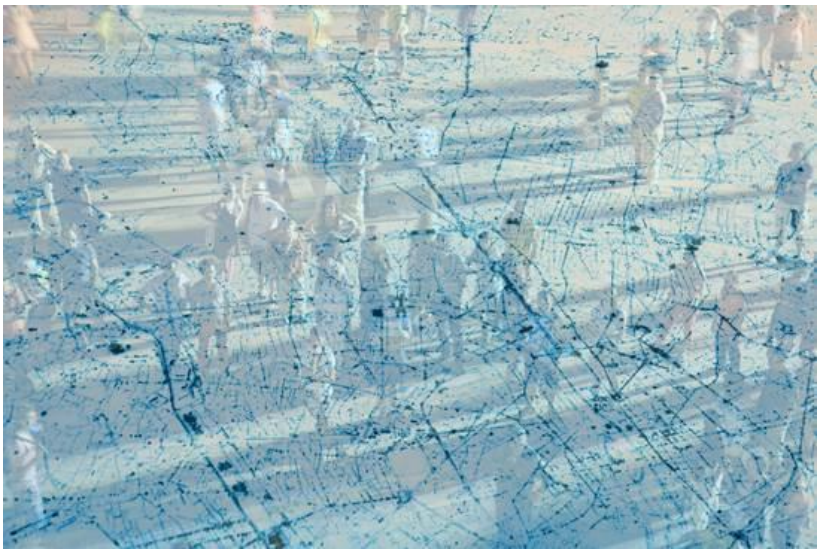
Εικόνα 4 Συγκριτικός πίνακας μεθόδων πρόβλεψης

Γενικά, τα μοντέλα Markov είναι από τα στατιστικά μοντέλα που χρησιμοποιούνται κατά κόρον στις προβλέψεις. Συγκεκριμένα στην περίπτωση μας τα μοντέλα Markov απλοποιεί τη διαδικασία επίλυσης παρέχοντας μας την θεμελιώδη ιδιότητα Markov, ή αλλιώς:
Η πιθανότητα ότι ένα γεγονός θα συμβεί, δεδομένων n παρελθοντικών γεγονότων, είναι περίπου ίσο με την πιθανότητα ότι ίδιο γεγονός θα συμβεί με δεδομένο μόνο το τελευταίο παρελθοντικό γεγονός.
Η χρήση της ιδιότητας Markov σημαίνει ότι δεν χρειάζεται να πάμε πολύ πίσω στο παρελθόν για προβλέψουμε μελλοντικά αποτελέσματα. Είναι εφικτό χρησιμοποιώντας μόνο τα πιο πρόσφατα παρελθοντικά γεγονότα. Το μοντέλο Markov εκτός από καλή αποτελεί και μια πολύ απλή προσέγγιση για την επίλυση του προβλήματος που εξετάζουμε.

5.1. Σκοπός

Σκοπός της παρούσας εργασίας είναι να αξιολογήσει διάφορες συνθέσεις του μοντέλου Markov ως προς την ικανότητα τους να προβλέψουν την επόμενη επισκεπτόμενη τοποθεσία κάποιου χρήστη σε ένα κοινωνικό δίκτυο.
Ο λόγος διερεύνησης των διαφόρων συνθέσεων είναι διότι στην υπάρχουσα βιβλιογραφία έχει προκύψει μετά από μελέτες ότι γενικώς οι ανθρώπινες συμπεριφορές μπορούν να περιγραφούν με ακρίβεια ως ένα σύνολο δυναμικών μοντέλων που μπορούν να αλληλουχιστούν μέσα από αλυσίδες Markov.
Υπάρχουν αρκετά επιστημονικά ερευνητικά άρθρα διαθέσιμα, τα οποία προσεγγίζουν το πρόβλημα της πρόβλεψης της επόμενης θέσης ενός ατόμου χρησιμοποιώντας διαφορετικές συνθέσεις του μοντέλου Markov στηριζόμενα στις παρατηρήσεις συμπεριφοράς του σχετικά με την κινητικότητα που

παρουσιάζει για δεδομένο χρονικό διάστημα και τις τοποθεσίες που έχει επισκεφθεί
Ελλιπής και ανεπαρκής έρευνα υπάρχει όμως για την αξιολόγηση της απόδοσης των προτεινόμενων μεθόδων υπό τις διαφορετικές συνθέσεις ώστε να καθοριστεί αν τα μοντέλα πρόβλεψης είναι ακριβή και παρέχουν σημαντικές και πολύτιμες πληροφορίες και αν πετυχαίνουν και κατά πόσο το στόχο τους.



Εικόνα 5 Ανθρώπινα ίχνη στο διαδίκτυο μέσα σε κοινωνικά δίκτυα

Η αφομοίωση των ψηφιακών ιχνών, που οι χρήστες αφήνουν στα κοινωνικά δίκτυα μέσω των κοινοποιήσεων παρουσίας, σε μοντέλα που βασίζονται και διαχειρίζονται δεδομένα μπορεί να αναπτύξει γενικά πρότυπα πρόβλεψης χωρικής κινητικότητας σε διαφορετικές κλίμακες ανάλυσης με υψηλή απόδοση και με σχετικά χαμηλές απαιτήσεις δεδομένων. Σε γενικές γραμμές, προσπαθούμε να βρούμε ένα τρόπο να ενσωματώσουμε αυτά τα ίχνη σε μοντέλα πρόβλεψης ανθρώπινης κινητικότητας που μπορούν να χρησιμοποιηθούν περαιτέρω για να περιγράψουν με ακρίβεια τις επιδημικές εκρήξεις, να διαχειριστούν την αντιμετώπιση καταστροφών, να σχεδιάσουν αστικές υπηρεσίες σε κατάλληλες χωρικές κλίμακες αλλά και να βοηθήσουν στην ανάπτυξη εφαρμογών κοινωνικής δικτύωσης, συστήματα προτάσεων και συστήματα υπενθυμίσεων.

5.2. Πεδίο εφαρμογής της εργασίας

Πρωταρχικός στόχος της εργασίας είναι να αναλύσουμε την απόδοση των μοντέλων Markov υπό διαφορετικές συνθέσεις στην προσέγγιση του προβλήματος την πρόβλεψης ανθρώπινης κινητικότητας.
Η ανάλυση της απόδοσης των μοντέλων θα γίνει στα πλαίσια της γενικής αξιολόγησης αποτελεσματικότητας μέσω προσομοιώσεων με πραγματικά δεδομένα.
Θα εξετάσουμε το πρόβλημα υπό το πρίσμα του ερευνητή και του προγραμματιστή σε εφαρμογές δεδομένων μεγάλης κλίμακας στα πλαίσια εκπόνησης μεταπτυχιακής εργασίας.
Η χρήση πραγματικών δεδομένων μεγάλης κλίμακας για την εκπόνηση της εργασίας κρίνεται απαραίτητη. Τα δεδομένα ελήφθησαν από την βάση

δεδομένων του κοινωνικού δικτύου Foursquare από εξωτερικό website και η χρήση τους θα είναι καθαρά και μόνο ερευνητική χωρίς καμία επεξεργασία και αναδιανομή.

5.3. Υπόθεση

Τα κοινωνικά δίκτυα με χρήση τοποθεσίας μέσω των δεδομένων που συλλέγουν στα αστικά κέντρα, μπορούν να μας βοηθήσουν να κατανοήσουμε τους παράγοντες που διέπουν την ανθρώπινη κινητικότητα. Μέσα από την ανάλυση των δεδομένων μπορούμε μοιραία να οδηγηθούμε στην ανάπτυξη εντυπωσιακών εφαρμογών και γενικότερων υπηρεσιών για τους χρήστες κινητών συσκευών.
Σε γενικές γραμμές οι άνθρωποι ακολουθούν κάποιες συνήθειες και υιοθετούν κάποιες συγκεκριμένες συμπεριφορές, οπότε είναι εν γένει εύκολο να προβλεφθεί η κίνηση τους. Υπάρχουν όμως και φορές που άλλοτε διακόπτουν τις συνήθειες τους ακανόνιστα ή τις τροποποιούν σε μεγάλο βαθμό.
Στην παρούσα μελέτη θα προσπαθήσουμε να μελετήσουμε την κίνηση μέσα από το φακό των διαφορετικών σημάτων πληροφοριών που είναι διαθέσιμα από αυτού του είδους τις υπηρεσίες. Εκτός από το να προσπαθήσουμε να ερμηνεύσουμε την πρόβλεψη της ανθρώπινης κινητικότητας, ο στόχος μας είναι να επιτύχουμε υψηλές αποδόσεις πρόβλεψης.
Ο βασικός στόχος μας είναι να αναπτύξουμε μοντέλο που να μπορεί να προβλέψει την επόμενη τοποθεσία που ενδεχομένως να επιλέξει ένας χρήστης να παρευρεθεί. Αυτό το μοντέλο μπορεί να χρησιμοποιηθεί σε συστήματα πρότασης τοποθεσιών, διαφημιστικές υπηρεσίες, ταξιδιωτικά πακέτα και γενικά προκύπτει πληθώρα σχετικών χρήσεων.
Πολλά εμπορικά καταστήματα, ξενοδοχεία και εστιατόρια μέσω τέτοιου είδους συστημάτων μπορούν να γνωρίζουν πότε ένας πελάτης μπορεί να παρουσιαστεί και μέσω στοχευμένων προσφορών, ή εξειδικευμένων υπηρεσιών να προσελκύσουν το ενδιαφέρον περισσότερων καταναλωτών στην επιχείρηση τους.
Είναι όμως εξαιρετικά μεγάλης σημασίας να οριστεί σαφώς τι πρόκειται να μελετηθεί, εξεταστεί και αξιολογηθεί σε αυτή τη μελέτη. Έτσι, θα προχωρήσουμε στη διατύπωση της υπόθεσης.
Κατ' επέκταση η βασική υπόθεση αυτής της μελέτης είναι ότι η εκτίμηση της επόμενης τοποθεσίας θα γίνει μέσω ακολουθιακής πρόβλεψης. Έτσι, μέσω της ακολουθιακής πρόβλεψης θα μπορέσουμε να προβλέψουμε και τις μελλοντικές τοποθεσίες των χρηστών.
Κατά συνέπεια, οι υποθέσεις της μελέτης σχετιζόμενες με τις αποδόσεις των μοντέλων Markov περιμένουμε να μας υποδείξουν ότι υπάρχει διαφορά μεταξύ των διαφόρων συνθέσεων μοντέλων Markov (πρώτης τάξης, δεύτερης τάξης και κρυφό μοντέλο).

5.4. Στόχοι

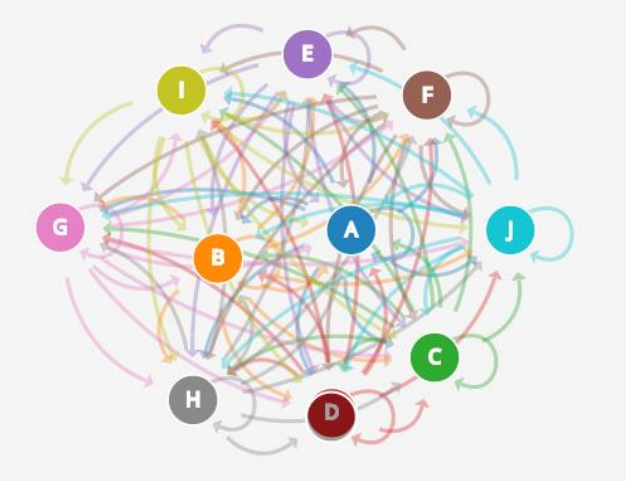
Η βασική παραδοχή στα πλαίσια αυτής της μελέτης είναι ότι όλες οι κινήσεις ενός χρήστη μέσα σε ένα κοινωνικό δίκτυο μπορούν να είναι και είναι προβλέψιμες. Δηλαδή, μέσω της στατιστικής παρατήρησης των κινήσεων ενός τυχαίου χρήστη, μπορούμε να έχουμε ένα πολύ καλό σημάδι για την πρόβλεψη των μελλοντικών τοποθεσιών που αυτός ο χρήστης αναφοράς θα επισκεφτεί.

Με δεδομένη αυτή την παραδοχή, ένας από τους στόχους αυτής της εργασίας είναι να παράσχει μια μέθοδο για την εκτίμηση των μελλοντικών τοποθεσιών, όπως το πού μπορεί κάποιος να βρεθεί χωροταξικά στο μέλλον, λαμβάνοντας υπόψη τις προηγούμενες κατηγορίες κοινοποιήσεων παρουσίας αλλά και το τις χρονικές στιγμές που έχουν γίνει αυτές οι κοινοποιήσεις από τον χρήστη αναφοράς.

Ο κύριος στόχος αυτής της εργασίας είναι η διεξαγωγή ανάλυσης σύγκρισης μεταξύ των διάφορων διαμορφώσεων του μοντέλου Markov.
Η διαδικασία για την επίτευξη του στόχου της μελέτης περιγράφεται εν γένει ως εξής:
1. Μεθοδολογική στρατηγική
2. Συλλογή και Ανάλυση Δεδομένων
3. Προγραμματισμός πειραμάτων και σχεδιασμός
4. Εργαλεία που χρησιμοποιούνται για την υλοποίηση
5. Αξιολόγηση εγκυρότητας
6. Αξιολόγηση και παρουσίαση του αποτελέσματος

5.5. Κίνητρο

Τα μοντέλα Markov (πρωτοτάξιο μοντέλο Markov, δευτεροτάξιο μοντέλο Markov και κρυφό μοντέλο Markov) χρησιμοποιούνται για τη μοντελοποίηση πολλών προβλημάτων. Ανάμεσα στις χρήσεις ξεχωρίζουν για τον αποτελεσματικότητα τους για την μοντελοποίηση προβλημάτων πρόβλεψης. Αυτό που δεν έχει μελετηθεί αρκετά είναι η αξιολόγηση επιδόσεων των μοντέλων Markov όσον αφορά το πόσο καλά μπορούν να χρησιμοποιηθούν με εφαρμογή στο πρόβλημα πρόβλεψης επόμενης θέσης κάποιου χρήστη με βάση το ιστορικό των κοινοποιήσεων παρουσίας, των κατηγοριών τους και με χρήση διάφορων χρονικών πληροφοριών σε ένα κοινωνικό δίκτυο. Για τη μελέτη αυτή θα χρησιμοποιηθεί όπως προείπαμε, ένα σύνολο δεδομένων από το δημοφιλές κοινωνικό δίκτυο Foursquare.



Εικόνα 6 Οπτικοποίηση μιας μικρής Μαρκοβιανής αλυσίδας 10 καταστάσεων

Ως εκ τούτου αυτό αποτελεί και το βασικό κίνητρο για τη διεξαγωγή αυτής της εργασίας. Τα μοντέλα που μελετήθηκαν, αποτελούν πρακτικές εφαρμογές και θα πρέπει να λειτουργούν και να παρουσιάζουν αποτελέσματα σε πραγματικές

συνθήκες και πραγματικό χρόνο σε διάφορες καταστάσεις ανά τον κόσμο. Ελπίζουμε αυτή η έρευνα να αποτελέσει δομικό λίθο και να χρησιμοποιηθεί περαιτέρω για την ανάπτυξη χρήσιμων εφαρμογών στον πραγματικό κόσμο.
Η επιλογή πεδίου μελέτης αφορά την περιοχή των κοινωνικών δικτύων με χρήση τοποθεσίας η οποία είναι μια εξαιρετικά εξελισσόμενη περιοχή έρευνας και καθιστά αυτή την εργασία πολύ ενδιαφέρουσα.
Μέσω αυτής της εργασίας ελπίζουμε να μπορέσουμε να κατανοήσουμε τις διαφορές απόδοσης των μοντέλων Markov στο συγκεκριμένο πεδίο μελέτης.
Η χρήση των μοντέλων Markov και των διαμορφώσεων τους είναι ευρέως γνωστή σε προγραμματιστές και ερευνητές. Άλλος ένας από τους στόχους και τα κίνητρα για τη σύνθεση αυτής της εργασίας είναι να μπορέσουμε μέσω της αξιολόγησης αυτών των μοντέλων να παράσχουμε υποστήριξη ώστε να γνωρίζουμε από πριν ποια διαμόρφωση Markov είναι η καλύτερη δυνατή για το μοντέλο που προσπαθούμε να κατασκευάσουμε.
Είναι σημαντικό να καταλάβουμε ποια είναι η διαφορά της απόδοσης και τι να αναμένουμε από τα μοντέλα Markov στα πρότυπα πρόβλεψης και αν είναι δυνατόν να μπορέσουμε να την βελτιώσουμε.

6. Μέθοδος

6.1. Ερευνητικός σχεδιασμός

Σύμφωνα με τις δημοσιεύσεις των Sekaran U., "Research methods for business: A skill building approach." και Burn R. B., "Introduction to research methods" η έρευνα μπορεί να περιγραφεί ως μια συστηματική και οργανωμένη προσπάθεια διερεύνησης ενός συγκεκριμένου προβλήματος για την παροχή μιας λύσης. Συνεπώς, το προϊόν της έρευνας είναι η παροχή νέων γνώσεων, η ανάπτυξη θεωριών καθώς και η συγκέντρωση στοιχείων για να αποδείξει τις γενικεύσεις στις οποίες θα καταλήξει. Στην δημοσίευση του "Foundations of Behavioral Research", ο Kerlinger F. N., δηλώνει ότι η επιστημονική έρευνα είναι μια συστηματική, ελεγχόμενη, εμπειρική και κριτική εξέταση προτάσεων σε συνάρτηση με υποτιθέμενες σχέσεις μεταξύ διαφόρων φαινομένων. Η έρευνα ταξινομείται σε τρεις βασικές κατηγορίες: ποσοτική, ποιοτική και μεικτή σύμφωνα με τον Creswell J. W. στη δημοσίευση του "Educational research: Planning, conducting, and evaluating quantitative and qualitative research"
Όσον αφορά τη φύση του σκοπού της παρούσας μελέτης προκειμένου να επιλεγεί μια χρησιμοποιήσιμη προσέγγιση για αξιολόγηση και να προσδιοριστεί το κίνητρο για το οποίο έχει επιλεγεί, διεξήχθη μια ποιοτική ερευνητική προσέγγιση. Αναζητήθηκε και μελετήθηκε η κατάλληλη σχετική βιβλιογραφία δεδομένου ότι υπάρχει. Έχει αποδειχτεί ότι η συλλογή δεδομένων με ποιοτικά χαρακτηριστικά είναι εξαιρετικά σημαντική. Επιπλέον επιλέξαμε να πραγματοποιήσουμε περαιτέρω πειραματική προσέγγιση πάνω σε πραγματικά δεδομένα μαζί με την σχετική ανάλυση τους ώστε να πετύχουμε να εξάγουμε τα αναμενόμενα αποτελέσματα και να είμαστε σε θέση να διεξάγουμε την συγκριτική ανάλυση για τις τεχνικές που επιλέξαμε να μελετήσουμε. Εναλλακτικά θα μπορούσαμε να διεξάγουμε μελέτη περίπτωσης ώστε να εφαρμοστεί ως στρατηγική συγκριτικής έρευνας. Να συγκρίνουμε δηλαδή το αποτέλεσμα χρήσης της μεθόδου που έχουμε διαλέξει με το αποτέλεσμα της χρήσης μιας άλλης διαφορετικής προσέγγισης.
Εντούτοις, επιλέξαμε την πειραματική προσέγγιση, αφού οι μελέτες περιπτώσεων που επιλέχτηκαν σε σχέση με τις μεταβλητές που χρησιμοποιούνται για τις προσομοιώσεις αντιπροσωπεύουν τυπικές καταστάσεις και εκτελέσαμε τις αντίστοιχες προσομοιώσεις (πειράματα) ώστε να έχουμε τον έλεγχο της κατάστασης και αν χρειαστεί να είμαστε σε θέση να χειραγωγήσουμε τη κατάσταση άμεσα, επακριβώς και συστηματικά. Προκειμένου να συγκρίνουμε και να επαληθεύσουμε τα πειραματικά μας αποτελέσματα διαλέξαμε περισσότερες από μία μεθόδους αξιολόγησης. Αναλύσαμε τις προαναφερθείσες μεθόδους αξιολόγησης με βάση τη σκοπιμότητα και τα αποτελέσματα. Είναι σημαντικό η επιλεγμένη μέθοδος αξιολόγησης να μπορεί να παράγει χρήσιμο αποτέλεσμα για την επαλήθευση της υπόθεσης.

6.2. Συλλογή και ανάλυση δεδομένων

Αρχικά κατασκευάσαμε τις τεχνικές που αναφέραμε νωρίτερα (πρωτοτάξιο, δευτεροτάξιο και κρυφό μοντέλο Markov) ώστε να είναι σε θέση να λύσουν το πρόβλημα που προηγουμένως έχουμε ορίσει, της πρόβλεψης την ανθρώπινης κίνησης μέσα από αλγορίθμους που σχετίζονται με τα Μαρκοβιανά μοντέλα.

Προκειμένου να επικυρώσουμε την εγκυρότητα και την ορθότητα των τεχνικών αυτών, χρησιμοποιήσαμε τους αλγορίθμους μας σε πραγματικά δεδομένα που προέρχονται από τη δημοφιλή πλατφόρμα κοινωνικής δικτύωσης Foursquare.
Το σύνολο δεδομένων που χρησιμοποιήσαμε το κατεβάσαμε από https://sites.google.com/site/yangdingqi/home/foursquare-dataset και έχει χρησιμοποιηθεί στο παρελθόν στη μελέτη των Dingqi Yang και Daqing Zhang, " Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs" όπου και δημοσίευσαν τα αποτελέσματα.
Το σύνολο δεδομένων, περιλαμβάνει μακροπρόθεσμα (περίπου 10 μήνες) δεδομένα κοινοποιήσεων παρουσίας από πραγματικούς χρήστες (ανωνυμοποιημένα) στη Νέα Υόρκη που συλλέχθηκαν από την πλατφόρμα από τις 3 Απριλίου 2012 έως τις 16 Σεπτεμβρίου 2013. Συνολικά υπάρχουν 227428 κοινοποιήσεις παρουσίας από 1084 χρήστες στην περιοχή της Νέας Υόρκης σε συνολικά 252 κύριες κατηγορίες και σε 38334 μοναδικές τοποθεσίες.
Για να εκτελέσουμε επαρκή ανάλυση των δεδομένων, αφιερώθηκε πάρα πολύς χρόνος ώστε να μπορέσουμε να τα κατανοήσουμε επαρκώς, να εξοικειωθούμε με το δείγμα, να κατανοήσουμε το πως θα τα χειριστούμε, τι εξαγόμενο ζητάμε από τα δεδομένα αναφοράς και να πάρουμε χρήσιμες γνώσεις από αυτά. Μετά τη συλλογή του συνόλου των δεδομένων και την αρχική επεξεργασία όπου αφαιρέθηκαν οι μη έγκυρες εγγραφές, εκτελέστηκε διεξοδική ανάλυση δεδομένων.
Αυτή η διαδικασία ανάλυσης των συλλεχθέντων δεδομένων θα μπορούσαμε να την αναφέρουμε και ως προσπάθεια κατανόησης τους, αναγνώρισης των μέσων και της συνοχής τους. Αυτό αναφέρεται ως ανάλυση περιεχομένου στη βιβλιογραφία στην δημοσίευση του Patton M., "A Guide to Using Qualitative Research Methodology".
Εξετάσαμε προσεκτικά όλο το σύνολο των δεδομένων και εντοπίσαμε τα βασικά τους χαρακτηριστικά. Σε γενικές γραμμές αναλύσαμε όλα τα χαρακτηριστικά των δεδομένων προκειμένου να κατανοηθεί επακριβώς η διαδικασία και η μεθοδολογία εργασίας που θα υιοθετηθεί για να έρθει εις πέρας αυτή η έρευνα και για να διαπιστώσουμε αν υπάρχουν αρκετά δεδομένα και είναι επαρκούς ποιότητας ώστε να μπορέσουν να βοηθήσουν στο μοντέλο πρόβλεψης που κατασκευάσαμε.

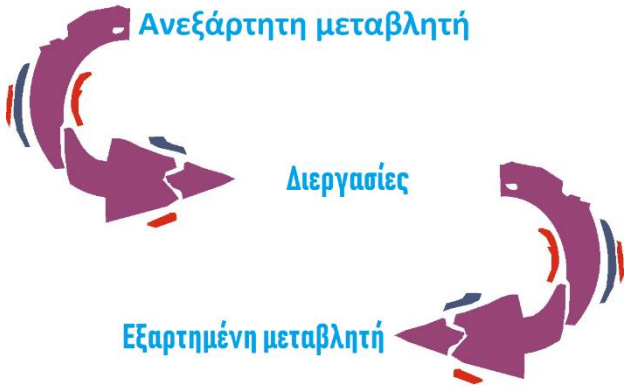
6.3. Πειραματικός σχεδιασμός

Σε αυτή την ενότητα θα περιγράψουμε τον σχεδιασμό των προσομοιώσεων που εκτελέσαμε.. Η διαδικασία του σχεδιασμού αποτελεί δομικό λίθο ώστε να μπορούν να εκτελεστούν σωστά και ορθολογικά τα πειράματα – προσομοιώσεις μας. Σε γενικές γραμμές ο σχεδιασμός περιλαμβάνει τον ορισμό των εξαρτημένων και ανεξάρτητων μεταβλητών, τον σχεδιασμό της ραχοκοκαλιάς των προσομοιώσεων και την ανάλυση των ρίσκων σχετικά με την εγκυρότητα τους.

Παρακάτω παρουσιάζεται το τυπικό διάγραμμα ροής πειραματικού σχεδιασμού.

Εικόνα 7 Διάγραμμα ροής πειραματικού σχεδιασμού

Αρχικά ξεκινάμε με τον προσδιορισμό των μεταβλητών. Στις προσομοιώσεις μας οι ανεξάρτητες μεταβλητές καθορίζουν τις περιπτώσεις για τις οποίες χρησιμοποιούμε δεδομένα των εξαρτώμενων μεταβλητών. Στις προσομοιώσεις μας οι ανεξάρτητες μεταβλητές είναι το σύνολο δεδομένων που περιλαμβάνει το ιστορικό των κοινοποιήσεων παρουσίας ενός χρήστη και η εξαρτημένη μεταβλητή είναι το αποτέλεσμα της απόδοσης των μοντέλων πάνω στα οποία εκτελούμε τις προσομοιώσεις μας.



Εικόνα 8 Σχέση ανεξάρτητης - εξαρτημένης μεταβλητής

Αφού το πρόβλημα έχει προσδιοριστεί σωστά και έχουμε επιλέξει τις ανεξάρτητες και της εξαρτημένες μεταβλητές καθώς και οι μετρικές αξιολόγησης για τις μεταβλητές μας, προχωράμε στο σχεδιασμό των προσομοιώσεων. Οι προσομοιώσεις μας αποτελούνται από μια ακολουθία δοκιμών. Για να αποδώσει στο έπακρο η προσομοίωση, πρέπει η ακολουθία δοκιμών να σχεδιαστεί προσεκτικά.

Οι γενικές αρχές σχεδιασμού περιλαμβάνουν τον προσδιορισμό της τυχαιοποίησης, της παρεμπόδιση και της εξισορρόπησης. Όσον αφορά την τυχαιοποίηση στον σχεδιασμό των πειραμάτων, ο απλούστερος τρόπος για τη σύγκριση των μεθόδων είναι ο "τυχαίος σχεδιασμός", η τυχαιοποίηση περιλαμβάνει την τυχαία κατανομή των πειραματικών μονάδων σε όλες τις μεθόδους. Οι προσομοιώσεις μας συγκρίνουν διαφορετικές διαμορφώσεις του μοντέλου Markov μεταξύ τους αφού έχει δοθεί ως είσοδος στο μοντέλο το ιστορικό κοινοποιήσεων παρουσίας του κάθε χρήστη για κάθε διαμόρφωση. Η ανάθεση σε κάθε μέθοδο επιλέγεται τυχαία. Σχετικά με την παρεμπόδιση, δεν εφαρμόζεται κάποια συστηματική προσέγγιση. Τέλος όσον αφορά την εξισορρόπηση, φροντίσαμε ώστε κάθε μέθοδος να έχει ίσο αριθμό πειραματικών μονάδων, ώστε να πετύχουμε έναν ισορροπημένο σχεδιασμό για να απλοποιήσουμε και ενισχύσουμε τη στατιστική ανάλυση των δεδομένων.
Ο ορισμός του προβλήματος, η καταγραφή της υπόθεσης και το μέτρο αξιολόγησης σημαίνουν ότι ο σχεδιασμός περιλαμβάνει: ένα παράγοντα με τρεις μεθόδους.
Η εξαρτώμενη μεταβλητή μετράται σε κλίμακα αναλογιών και έτσι μπορούν να γίνουν παραμετρικές δοκιμές. Η καταλληλότερη είναι η ανάλυση διακύμανσης.

6.4. Εργαλεία που χρησιμοποιήθηκαν για την εκτέλεση των προσομοιώσεων

Ένα από τα σημαντικότερα μέρη για την εκτέλεση των πειραμάτων και των προσομοιώσεων για την μελέτη του φαινομένου είναι ότι θα πρέπει να καθίσταται δυνατή και άμεσα μετρήσιμη η αξιολόγηση των αποτελεσμάτων με τα επιλεγμένα μοντέλα που θα εφαρμόσουμε. Μετά από αρκετή βιβλιογραφική έρευνα σχετική με το θέμα, επιλέξαμε το μοντέλο Markov και διαμορφώσεις του για να προσομοιώσουμε την πρόβλεψη της επόμενης θέσης του χρήστη.
Σε αυτή την ενότητα θα περιγράψουμε τα εργαλεία και τα προγράμματα που χρησιμοποιήθηκαν για την εκτέλεση των προσομοιώσεων για να προτείνουμε επιλογές υλικού και λογισμικού που είναι σε θέση να μας βοηθήσουν σε μελλοντική έρευνα.
Οι προτεινόμενες τεχνικές εφαρμόστηκαν γραμμένες στη γλώσσα προγραμματισμού. Το κύριο πλεονέκτημα της συγγραφής κώδικα είναι ότι είμαστε σε θέση να έχουμε τον πλήρη έλεγχο της εξέλιξης και να μπορούμε εύκολα να επέμβουμε σε περίπτωση που υπάρχει κάποιο σφάλμα. Το πρόγραμμα γράφτηκε προσαρμοσμένο σύμφωνα με το σύνολο δεδομένων που αποκτήσαμε. Έτσι η επαναχρησιμοποίηση του για διαφορετικά δεδομένα καθίσταται αρκετά δύσκολη. Όμως στις επαναλήψεις των πειραμάτων με τα ίδια δεδομένα είναι εύκολο να καταλάβουμε και να εντοπίσουμε (αν υπάρχουν) τυχόν δυσλειτουργίες ή περίεργη συμπεριφορά του προγράμματος και τους λόγους ώστε να μπορέσουμε να εκτελέσουμε σωστά αποσφαλμάτωση.
Όπως προαναφέρθηκε, η υλοποίηση του μοντέλου έγινε προγραμματιστικά. Οι αλγόριθμοι γράφτηκαν σε Matlab R2017a, ενώ για την ανάλυση των δεδομένων χρησιμοποιήθηκε το Microsoft Excel Professional Plus 2016. Ενώ μπορούσε να χρησιμοποιηθεί και άλλη γλώσσα προγραμματισμού, επιλέχθηκε η Matlab γιατί προσφέρει πολλά πακέτα μαθηματικών βιβλιοθηκών, και είναι επαρκώς γρήγορη για την εκτέλεση του κώδικα.

Ο υπολογιστής που χρησιμοποιήθηκε ήταν ένας τυπικός υπολογιστής γραφείου. Το πρόγραμμα είναι σε θέση να εκτελείται άμεσα και δεν είναι «βαρύ», έτσι δεν υπάρχει πραγματική ανάγκη για ισχυρό υπολογιστή.
Παρακάτω παραθέτω ενδεικτικά τη σύνθεση του υπολογιστή ο οποίο εκτέλεσε τον κώδικα και παρήγαγε τα αποτελέσματα που παρουσιάζουμε στα πλαίσια αυτής της εργασίας.

- Επεξεργαστής : Intel® Core™2 Quad Processor Q9450 (12M Cache, 2.66 GHz, 1333 MHz FSB)
- Μνήμες RAM: Corsair Vengeance 8GB DDR3-1600MHz (CMZ8GX3M2A1600C9)
- Σκληρός Δίσκος: Kingston SSDNow V300 240GB
- Λειτουργικό Σύστημα: Microsoft Windows 10

6.5. Αξιολόγηση εγκυρότητας

Ανάλογα με το στόχο του πειράματος και των προσομοιώσεων που εκτελούνται, διακρίνονται διάφορα επίπεδα που συμβάλλουν στην καθολική αξιολόγηση της εγκυρότητας. Η στατιστική εγκυρότητα του συμπεράσματος, η εσωτερική εγκυρότητα, η κατασκευαστική εγκυρότητα και η εξωτερική εγκυρότητα.
Η **στατιστική εγκυρότητα του συμπεράσματος** (SCV) ή πιο απλά εγκυρότητα του συμπεράσματος, είναι ένα μέτρο που μας δείχνει κατά πόσο λογικό είναι ένα ερευνητικό ή πειραματικό συμπέρασμα. Σε γενικές γραμμές χρησιμοποιείται τόσο για ποιοτική έρευνα όσο και για ποσοτική. Στις προσομοιώσεις μας εφαρμόζονται αρκετά γνωστές στατιστικές τεχνικές οι οποίες είναι εύρωστες για παραβίαση των υποθέσεων τους. Μια γενική απειλή για την εγκυρότητα του συμπεράσματος είναι, ωστόσο, ο χαμηλός αριθμός των εγγραφών που έχει το δείγμα μας, τα οποία ενδέχεται να μειώσουν την ικανότητα εξεύρεσης μοτίβων στα δεδομένα.
Η **εγκυρότητα του συμπεράσματος** αφορά μόνο το ερώτημα: Με βάση τα δεδομένα που έχουμε, το συμπέρασμα έχει κάποια σχέση ή όχι; Είναι σημαντικό να συνειδητοποιήσουμε ότι δεν υπάρχει τέλεια εγκυρότητα του συμπεράσματος και στην ουσία δεν μπορούμε να είμαστε ποτέ 100% βέβαιοι ότι τα συμπεράσματά μας είναι σωστά. Ωστόσο, η μετρική αυτή αναφέρεται σε εύλογα συμπεράσματα με βάση τα δεδομένα μας.
Η **εσωτερική εγκυρότητα** επικεντρώνεται κυρίως στην εγκυρότητα της μελέτης και αφορά τον τρόπο μέτρησης της ακρίβειας της. Με άλλα λόγια, είναι ένας τρόπος μέτρησης της ισχυρότητας των μεθόδων έρευνας που χρησιμοποιήθηκαν. Αφορά θέματα που μπορεί να επηρεάσουν την ανεξάρτητη μεταβλητή σε σχέση με την αιτιότητα, που δεν έχει λάβει υπόψιν ο ερευνητής. Γενικά, σχετίζεται με το πόσες μεταβλητές συγχέονται στο εκάστοτε πείραμά. Κατά την εκτέλεση ενός πειράματος όσο λιγότερες είναι οι μεταβλητές που συγχέονται, τόσο υψηλότερη θα είναι η εσωτερική του εγκυρότητα και αντιστρόφως. Σε ιδανικές συνθήκες, το πείραμά πρέπει να έχει υψηλή εσωτερική εγκυρότητα.
Μία από τις απειλές για την εσωτερική εγκυρότητα σε αυτή την εργασία είναι η επιλογή. Η επιλογή του δείγματος που εξετάστηκε ήταν τυχαία, πράγμα που συνεπάγεται απειλή για την εγκυρότητα της μελέτης.
Η **κατασκευαστική εγκυρότητα** είναι ένας τρόπος να ελέγξουμε την εγκυρότητα των πειραμάτων. Επιδεικνύει ότι το πείραμα μετράει στην πραγματικότητα

αυτό το οποίο κατασκευάστηκε για να μετράει. Επίσης αφορά τη γενίκευση των αποτελεσμάτων του πειράματος πάνω στη θεωρία στην οποία στηρίχθηκε. Δεν είναι πάντα εύκολο να μετρηθεί η κατασκευαστική εγκυρότητα. Συνήθως απαιτούνται διάφορα μέτρα για να αποδειχθεί. Ένας από τους λόγους για τους οποίους είναι τόσο δύσκολο να μετρηθεί είναι η ύπαρξη υποκειμενικότητας Μία σημαντική απειλή για την κατασκευαστική εγκυρότητα στη μελέτη μας είναι ότι οι τεχνικές για αξιολόγηση μπορεί να μην είναι αντιπροσωπευτικές. Αυτό περιορίζει το πεδίο των συμπερασμάτων που εξάγονται από αυτές τις συγκεκριμένες τεχνικές.
Η **εξωτερική εγκυρότητα** βοηθά στο να απαντηθεί η ερώτηση: μπορεί η συγκεκριμένη έρευνα να εφαρμοστεί στον «πραγματικό κόσμο»; Αφορά δηλαδή τη γενίκευση του αποτελέσματος του πειράματος σε περιβάλλον διαφορετικό από αυτό στο οποίο διεξήχθη. Εάν η έρευνά είναι εφαρμόσιμη και σε άλλα πειράματα, με άλλες παραμέτρους, δείγμα κλπ, τότε η εξωτερική εγκυρότητα του είναι υψηλή και αντίστροφα. Είναι σημαντικό η έρευνά να είναι αρχικά αποτελεσματική και να μπορεί να γενικευτεί και σε άλλες περιπτώσεις. Γενικά είναι αρκετά δύσκολο να γενικευθούν τα αποτελέσματα σε άλλα σενάρια.

Κάποιοι από τους παράγοντες που επηρεάζουν την εγκυρότητα της έρευνας μας είναι αρχικά ότι δεν είναι εντελώς αξιόπιστη η ανεξάρτητη μεταβλητή που χρησιμοποιούμε λόγω υψηλής μεταβλητότητας. Στο πείραμα μας η ανεξάρτητη μεταβλητή είναι το σύνολο δεδομένων των κοινοποιήσεων παρουσίας των χρηστών. Οι κοινοποιήσεις αυτές είναι εντελώς τυχαίες όπως ελήφθησαν από την πλατφόρμα. Παρ ‘όλη την επεξεργασία που κάναμε στο δείγμα, και αφού επιλέξαμε να χρησιμοποιήσουμε τις δέκα πιο δημοφιλής κατηγορίες για την διεξαγωγή των προσομοιώσεων, μπορούμε να πούμε ότι δεν εμφανίζονται σχεδόν καθόλου μοτίβα για τις τοποθεσίες. Άλλος ένας παράγοντας είναι η έλλειψη αντιπροσωπευτικότητας της ανεξάρτητης μεταβλητής. Οι χρήστες προέρχονται από όλες τις κοινωνικές ομάδες και όλα τα ηλικιακά γκρουπ, κάτι που κάνει τη δουλειά μας αρκετά πιο περίπλοκη και δύσκολη. Είναι ευκόλως κατανοητό ότι διαφορετικές επισκέψεις σε τοποθεσίες θα κάνει ένα ανήλικο άτομο από κάποιο ενήλικο και ένα πιο εύπορο άτομο από κάποιο λιγότερο. Δυστυχώς δεν μπορούσαμε στο δείγμα που λάβαμε να κάνουμε τέτοιου είδους διαχωρισμούς κάτι που μας φέρνει στον παράγοντα της μη αντιπροσωπευτικής δειγματοληψίας.
Τα προαναφερθέντα μας φέρνουν αντιμέτωπους με το πρόβλημα της έλλειψης επιρροής της ανεξάρτητης μεταβλητής. Τα αποτελέσματα του πειράματος πρέπει να είναι σε θέση να παράγουν ρεαλιστικές επιρροές στο δείγμα. Αυτό δεν είναι εφικτό σε μεγάλο βαθμό στις προσομοιώσεις μας.
Η εξαρτημένη μας μεταβλητή είναι το αποτέλεσμα της απόδοσης των μοντέλων, κάτι που μας δίνει άλλον έναν παράγοντα επιρροής της εγκυρότητας της μελέτης μας, μιας και υπάρχει μερική έλλειψη αξιοπιστίας και αντιπροσωπευτικότητας της εξαρτημένη μεταβλητής. Αυτό γιατί δεν μπορούμε να είμαστε εντελώς βέβαιοι ότι οι υλοποιήσεις των μοντέλων μας είναι κατασκευασμένα στο μαθηματικό επίπεδο πολυπλοκότητας που απαιτείται και τα δείγματα με τα οποία εκπαιδεύουμε τα μοντέλα μας ώστε να αποφανθούμε για την απόδοση τους είναι εντελώς τυχαία χωρίς τους διαχωρισμούς που αναφέραμε. Επίσης η εξαρτώμενη μεταβλητή είναι μια ποσοτική μονάδα που μετράει την απόδοση με πολύ μικρή ευαισθησία. Το εξαγόμενο του πειράματος

δεν είναι μια σειρά από αποτελέσματα τα οποία θα μπορούσαμε να εκτιμήσουμε και να μελετήσουμε καλύτερα αλλά τελικά ένας λόγος των σωστά εκτιμημένων επόμενων κοινοποιήσεων προς το σύνολο αυτών. Η μικρή ευαισθησία είναι αποτέλεσμα του ότι δεδομένου του δείγματος, όσες φορές και να εκτελεστούν οι προσομοιώσεις το αποτέλεσμα θα είναι το ίδιο. Ενδέχεται να διαφοροποιείται με άλλο επιλεγόμενο δείγμα μόνο όμως η κατασκευή των μοντέλων μας έχει γίνει πάνω στα πειραματικά δεδομένα που είχαμε προηγουμένως λάβει και ενδέχεται να υπάρχουν επιπτώσεις στην μεταφορά τους πάνω σε ένα πειραματικά δεδομένα. Παρ ’όλο που στην βιβλιογραφία έχει μελετηθεί και έχει προσεγγιστεί ξανά η μελέτη της ανθρώπινης κινητικότητας με χρήση μαρκοβιανών μοντέλων και δείγμα κοινοποιήσεις παρουσίας ληφθείσες από κοινωνικά δίκτυα ενδέχεται η χρήση αυτών των στατιστικών τεχνικών να είναι ακατάλληλη και το πρόβλημα να προσεγγίζεται πολύ καλύτερα με χρήση άλλων μαθηματικών-στατιστικών τεχνικών.

6.5.1. Ελαχιστοποίηση πιθανών απειλών

Για να ελαχιστοποιήσουμε τις πιθανές απειλές που αφορούν στην εγκυρότητα των πειραμάτων μας θα ακολουθήσουμε μια σειρά από βήματα.
Για την κατασκευαστική εγκυρότητα μπορούμε να πούμε ότι η τυποποίηση των όρων και των εννοιών με βάση τους οποίους διεξάγεται η έρευνα συμβάλει στην ελαχιστοποίηση των απειλών. Σε αυτή τη μελέτη εφαρμόζονται καλά γνωστές στατιστικές τεχνικές οι οποίες είναι εύρωστες στην υπόθεση μπορεί να μην είναι αντιπροσωπευτικές του σεναρίου.
Η απόκτηση όσο το δυνατόν περισσότερων πληροφοριών σχετικά με το πειραματικό δείγμα στην μελέτη βοηθά στην ελαχιστοποίηση των απειλών για την εσωτερική εγκυρότητα.
Η απόκτηση όσο το δυνατόν περισσότερων πληροφοριών σχετικά με τις διαδικαστικές λεπτομέρειες και το σενάριο της ερευνητικής μελέτης, για παράδειγμα, το πλαίσιο της μελέτης, το περιβάλλον, τα καθορισμένα χαρακτηριστικά, ελαχιστοποιούν την απειλή για εξωτερική εγκυρότητα.
Σαν γενικός κανόνας, η επιλογή κατάλληλου ερευνητικού σχεδιασμού βοηθάει στον έλεγχο των περισσότερων απειλών για την εγκυρότητα.
Η αύξηση του μεγέθους του δείγματος, η τυχαία επιλογή και η τυχαία ανάθεση των υποκειμένων ελαχιστοποιούν την επιλογή ως απειλή για την εσωτερική εγκυρότητα καθώς και ως απειλή για την εγκυρότητα του συμπεράσματος.
Ένα πείραμα που διεξάγεται εντός ενός συγκεκριμένου πλαισίου σχεδιάζεται για να απαντήσει αποκλειστικά σε ερωτήσεις που αφορούν το συγκεκριμένο πλαίσιο και θεωρείται επαρκές μόνο εάν τα αποτελέσματα είναι έγκυρα εντός αυτού του συγκεκριμένου πλαισίου. Η γενίκευση του αποτελέσματος του πειράματος σε άλλο πλαίσιο εκτός της μελετώμενης περίπτωσης, με άλλα λόγια, αν υποθέσουμε ότι το συμπέρασμα είναι καθολικό, τότε η εγκυρότητα πρέπει να καλύπτει το ευρύτερο πλαίσιο για το οποίο ισχυριζόμαστε ότι το αποτέλεσμα γενικεύεται κάτι που είναι εξαιρετικά δύσκολο έως και αδύνατο. Είναι σημαντικό να υπάρχουν επαρκείς γνώσεις σχετικά με την αντιπροσωπευτικότητα των ευρημάτων.
Προκειμένου να διασφαλίσουμε την γενικότερη εγκυρότητα μια μέθοδος είναι να εκτελέσουμε αρκετά μεγάλο αριθμό προσομοιώσεων πάνω στο δείγμα μας. Έτσι μπορούμε να πούμε ότι βρισκόμαστε σε ένα καλό επίπεδο.

Τέλος συμπεραίνουμε ότι οι απειλές για την εγκυρότητα του αποτελέσματος του πειράματος και την αξιολόγηση του είναι σημαντικό να προσδιοριστούν κατά τη φάση σχεδιασμού του, αφού ότι οι απειλές που αφορούν το πείραμα συνδέονται στενά με την πρακτική σημασία των αποτελεσμάτων.

6.6. Αξιολόγηση και παρουσίαση των αποτελεσμάτων

Ένα βασικό στοιχείο της επιστημονικής έρευνας παρουσιάζει τα αποτελέσματα της έρευνας στην επιστημονική κοινότητα.
Η τυπική δομή παρουσίασης αποτελεσμάτων περιλαμβάνει την εισαγωγή ο οποία παρέχει πληροφορίες σχετικά με το πρόβλημα που διερευνάται, την τρέχουσα γνώση σχετικά με το πρόβλημα και πρέπει να θέτει σαφώς την υπόθεση για μελέτη καθώς και τις πειραματικές προσδοκίες. Ακολουθεί η παρουσίαση των μεθόδων, όπου περιγράφονται οι διαδικασίες με τις οποίες θα γίνουν οι κατάλληλες δοκιμές για τη μελέτη της υπόθεσης, με αρκετή λεπτομέρεια ώστε να μπορεί το κάθε πείραμα να επαναληφθεί από άλλους ενδιαφερόμενους, καθώς και τα υλικά, ο εξοπλισμός, οι αναλυτικές και οι στατιστικές διαδικασίες που ακολουθούνται. Έπειτα, ακολουθεί το τμήμα των αποτελεσμάτων που δίνει μια σύνοψη του πειραματικού αποτελέσματος της μελέτης, περιλαμβάνει λεκτική περιγραφή τους, και προσφέρει πίνακες και μετρικές, όπου αναφέρονται τα στατιστικά αποτελέσματα καθώς και το πειραματικό σφάλμα. Φυσικά οι πίνακες και οι μετρικές πρέπει να περιλαμβάνουν τις σωστές λεζάντες που εξηγούν τι συνοψίζουν. Τέλος έρχεται το κομμάτι της συζήτησης όπου ερμηνεύονται τα αποτελέσματα και εξάγονται συμπεράσματα. Συνήθως σε αυτό το τμήμα, συγκρίνονται τα αποτελέσματα της μελέτης με αποτελέσματα άλλων παρεμφερών μελετών ώστε να δοθεί μια ευρύτερη σημασία των ευρημάτων και παρουσιάζονται οι περιορισμοί της μελέτης, πηγές σφαλμάτων αν υπάρχουν και σχέδια για μελλοντική εργασία.
Σε αυτό το σημείο του κειμένου, έχουμε ήδη παρουσιάσει εκτενώς το τμήμα της εισαγωγής. Παρακάτω θα δούμε την παρουσίαση των μεθόδων και τεχνικών σε θεωρητικό υπόβαθρο αναφορικά με τα μοντέλα Markov και θα εξηγήσουμε την υλοποίηση τους, καθώς και την πειραματική διαδικασία που περιλαμβάνει τις αντίστοιχες προσομοιώσεις.
Αφού ολοκληρωθούν οι υλοποιήσεις των προτεινόμενων μοντέλων, θα προσαρμόσουμε τα αποτελέσματα καταλλήλως ώστε να μπορούν να χρησιμοποιηθούν για αξιολογήσεις. Θα πραγματοποιηθεί επίσης συγκριτική ανάλυση των μοντέλων αναφορικά με την ακρίβεια τους.
Με χρήση του συνόλου δεδομένων και τα εξαγόμενα αποτελέσματα, θα κατασκευάσουμε τα αντίστοιχα γραφήματα, γραφικές παραστάσεις και πίνακες που παρουσιάζουν διαφορετικές καταστάσεις και παραμετροποιήσεις.
Τέλος, θα πάμε στο κομμάτι της συζήτησης στο τμήμα του επιλόγου, όπου θα δούμε τις συνεισφορές της μελέτης μας, θα σχολιάσουμε τα αποτελέσματα και θα δούμε σκέψεις και ιδέες για μελλοντική εργασία.

7. Ανάλυση Συνόλου Δεδομένων

Είναι εύκολα κατανοητό ότι σε μια τέτοιου είδους ανάλυση το σύνολο των δεδομένων παίζει κρίσιμο ρόλο.
Έτσι σε αυτό το εδάφιο κρίνεται απαραίτητο να παρουσιάσουμε το σύνολο των δεδομένων που χρησιμοποιήθηκαν, από πού ελήφθησαν και τι παρουσιάζουν. Αρχικά, θα παρουσιάσουμε το σύνολο των δεδομένων και έπειτα θα ασχοληθούμε με μια μικρή ανάλυση γύρω από αυτό.

7.1. Σύνολο δεδομένων

Προκειμένου να προσεγγίσουμε το πρόβλημα και να μοντελοποιήσουμε την ανθρώπινη κινητικότητα, χρησιμοποιήσαμε ένα πακέτο δεδομένων προερχόμενο από ένα από τα πιο γνωστά και δημοφιλή κοινωνικά δίκτυα "Foursquare".
Το σύνολο δεδομένων που χρησιμοποιήσαμε το κατεβάσαμε από https://sites.google.com/site/yangdingqi/home/foursquare-dataset και είναι το ίδιο που χρησιμοποιήθηκε στη διατριβή των Dingqi Yang και Daqing Zhang, “ Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs”.
Οι κοινοποιήσεις παρουσίας καταγράφηκαν από την περισυλλογή tweet με ετικέτες Foursquare μέσω του "Twitter Public Stream API".
Το σύνολο δεδομένων, περιλαμβάνει μακροπρόθεσμα (περίπου 10 μήνες) δεδομένα κοινοποιήσεων παρουσίας στη Νέα Υόρκη που συλλέχθηκαν από την πλατφόρμα από τις 3 Απριλίου 2012 έως τις 16 Σεπτεμβρίου 2013. Συνολικά υπάρχουν 227428 κοινοποιήσεις παρουσίας από 1084 χρήστες στην περιοχή της Νέας Υόρκης.
Κατά μέσο όρο ο κάθε χρήστης του dataset που επεξεργαστήκαμε έχει πραγματοποιήσει περίπου 210 κοινοποιήσεις παρουσίας.

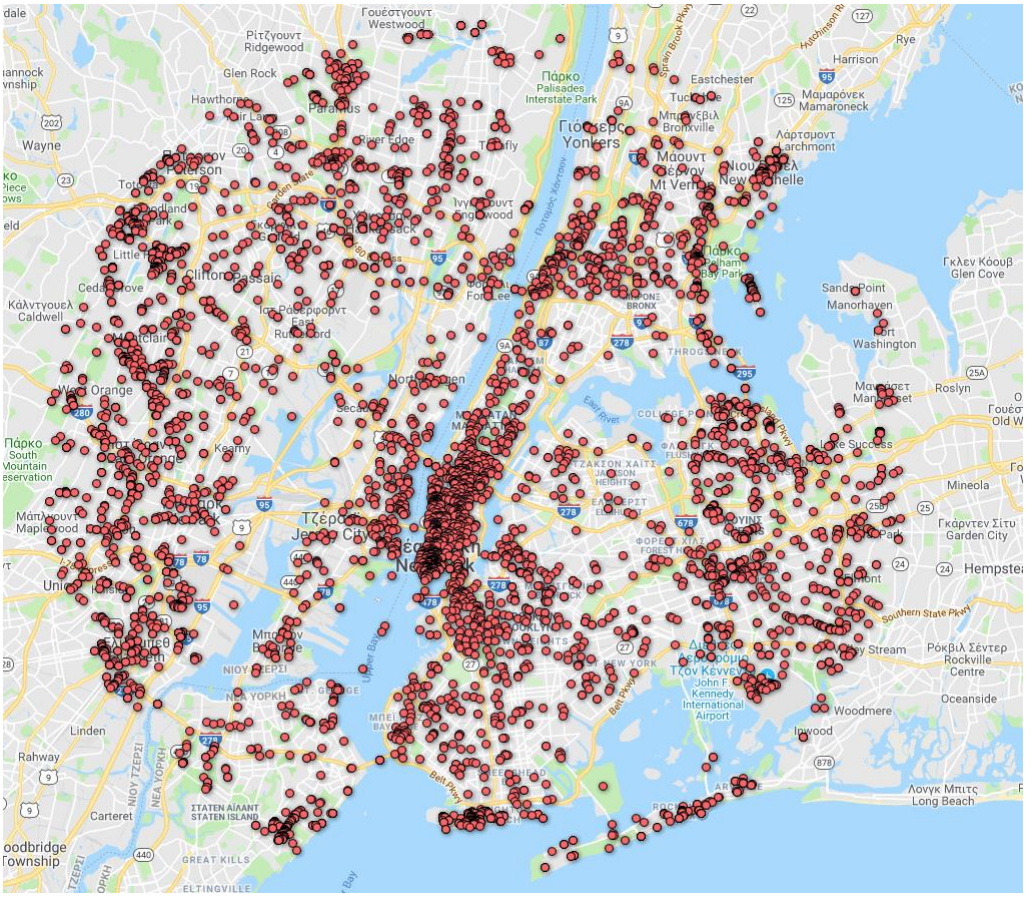
Το αρχείο έχει σαν δομή 8 στήλες:

1. Αναγνωριστικό χρήστη (ανώνυμο)
2. Αναγνωριστικό τοποθεσίας (Foursquare)
3. Αναγνωριστικό κατηγορίας τοποθεσίας (Foursquare)
4. Όνομα κατηγορίας τοποθεσίας (Foursquare)
5. Γεωγραφικό πλάτος
6. Γεωγραφικό μήκος
7. Μετατόπιση ζώνης ώρας σε λεπτά (Η χρονική μετατόπιση σε λεπτά μεταξύ της ίδιας της κοινοποίησης και της ώρας στο UTC)
8. Ώρα UTC

Στο σχήμα παρακάτω βλέπουμε μια οπτικοποίηση όλου του dataset πάνω στο χάρτη της Νέας Υόρκης. Η οπτικοποίηση έγινε με χρήση google fusion tables.

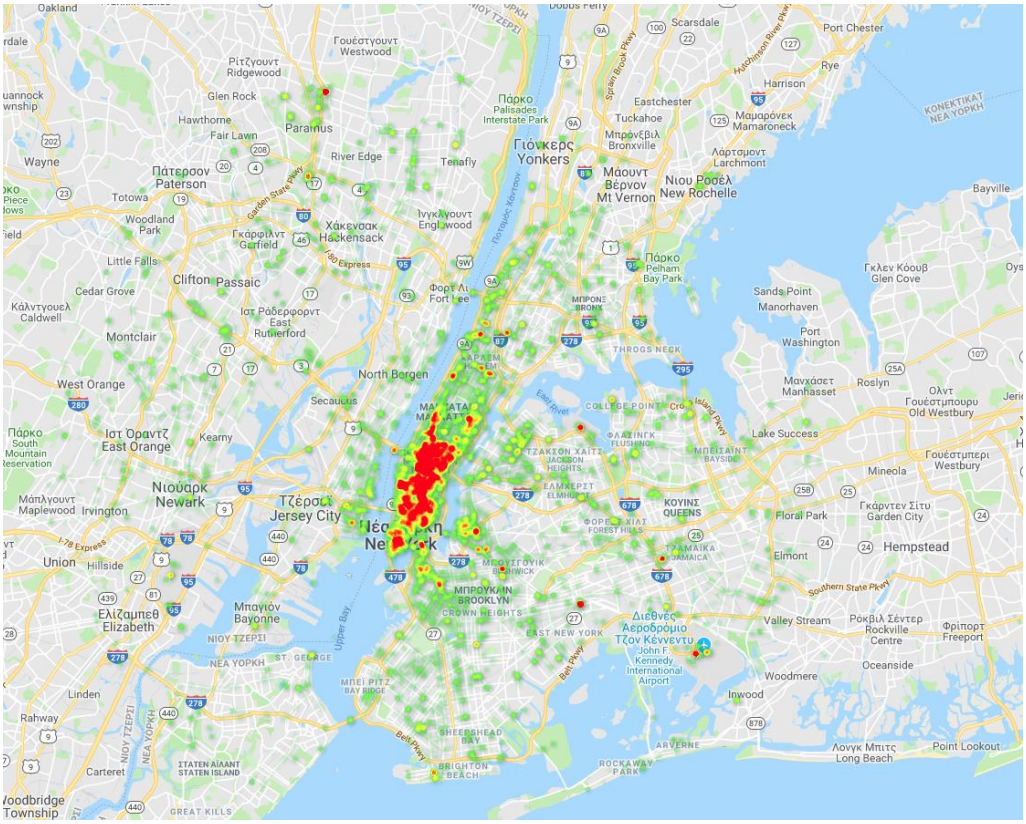
Για να κάνουμε τη μελέτη μας πιο σωστή έπρεπε να «καθαρίσουμε» το σύνολο δεδομένων. Έτσι, κάναμε ένα φιλτράρισμα στις λανθασμένες τιμές, σε τιμές θορύβου και προσπαθήσαμε να διακρίνουμε τις μη έγκυρες εγγραφές ώστε να τις απαλείψουμε. Επίσης ουσιώδες ήταν να επιλέξουμε να μελετήσουμε τους ενεργούς χρήστες της πλατφόρμας και κάναμε επιπλέον ελέγχους ώστε να

καταλήξουμε σε αυτές τις επιλογές. Επιλέξαμε να μελετήσουμε τους χρήστες που έχουν τουλάχιστον μια κοινοποίηση παρουσίας την εβδομάδα.



Εικόνα 9 Οπτικοποίηση του συνόλου δεδομένων στο χάρτη της Νέας Υόρκης

Όταν ο χρήστης κοινοποιεί την παρουσία του, το Foursquare, χρησιμοποιώντας το σύστημα επαλήθευσης που διαθέτει και μέσα από μια σειρά ελέγχων, αποφαίνεται για το αν ο χρήστης είναι στην πραγματικότητα κοντά σε αυτό το μέρος ή όχι. Μέσα στο σύνολο δεδομένων εντοπίστηκαν πολλές ψεύτικες κοινοποιήσεις τις οποίες εξαλείψαμε. Κάτι άλλο που παρατηρήθηκε και στο σύνολο δεδομένων ήταν οι λεγόμενες "ξαφνικές κινήσεις". Έτσι κάναμε έλεγχο για τον εντοπισμό διαδοχικών κοινοποιήσεων που είναι ταχύτερα από την ταχύτητα του αεροπλάνου θέτοντας ως ανώφλι τα 1300km/h για να αποδεχτούμε την αλληλουχία των κοινοποιήσεων ή να την απορρίψουμε. Τον όρο “ξαφνική κίνηση” ή “άμεση κοινοποίηση” τον εισήγαγαν οι Cheng Z., Caverlee J. και Lee K. στην εργασία τους “ Exploring Millions of Footprints in Location Sharing Services” Προσπαθήσαμε να αφαιρέσουμε όλες τις “ξαφνικές κινήσεις” υπήρχαν στο σύνολο δεδομένων μας.



Εικόνα 10 Πυκνότητα κοινοποιήσεων του dataset πάνω στο χάρτη της Νέας Υόρκης

Από τα δύο γραφήματα που παρουσιάζονται είναι ξεκάθαρο και ειδικότερα από το δεύτερο ότι οι περισσότερες κοινοποιήσεις παρουσίας γίνονται στο κέντρο της πόλης και σημαντικά λιγότερες γίνονται στα περίχωρα.

7.2. Ανάλυση του συνόλου δεδομένων

Όταν ο χρήστης κοινοποιεί την παρουσία του στην πλατφόρμα, το Foursquare, χρησιμοποιώντας ένα αλγοριθμικό σύστημα επαλήθευσης, ελέγχει αν ο χρήστης είναι πραγματικά κοντά σε αυτό το μέρος ή όχι.
Πραγματοποιήσαμε παρ'όλα αυτά φιλτράρισμα σε όλο το σύνολο δεδομένων για ψεύτικες κοινοποιήσεις.
Ένας δευτερογενής έλεγχος που πραγματοποιήθηκε ήταν για "ξαφνικές κινήσεις". Διαδοχικές κοινοποιήσεις παρουσίας που είναι ταχύτερες από την ταχύτητα του αεροπλάνου έπρεπε να αφαιρεθούν γιατί ενδεχομένως να περιείχαν σφάλματα και να μας αλλοίωναν τα αποτελέσματα της ανάλυσης.

Σε γενικές γραμμές διακρίνονται 252 κατηγορίες κοινοποιήσεων:
Arts & Crafts Store, Bridge , Home (private), Medical Center, Food Truck, Food & Drink Shop, Coffee Shop, Bus Station, Bank, Gastropub, Electronics Store, Mobile Phone Shop, Cafı, Automotive Shop, Restaurant, American Restaurant, Government Building, Airport, Ferry, Office, Other Great Outdoors, Building, Mexican Restaurant, Music Venue, Subway, Student Center, Park, Road, Burger Joint, Sporting Goods Shop, Pizza Place, Jewelry Store, Sandwich Place, Clothing Store, Neighborhood, Ice Cream Shop, Soup Place, College Academic Building, Department Store, Playground, Tattoo Parlor, Mall, Deli / Bodega, University,

Diner, Music Store, Light Rail, Salon / Barbershop, General College & University, Animal Shelter, Laundry Service, Residential Building (Apartment / Condo), Drugstore / Pharmacy, Cuban Restaurant, BBQ Joint, Other Nightlife, Gym / Fitness Center, Italian Restaurant, Stadium, Church, Train Station, Tanning Salon, Hotel, Miscellaneous Shop, Bar, Spanish Restaurant, Moving Target, Asian Restaurant, Factory, School, General Travel, Burrito Place, Fast Food Restaurant, Dumpling Restaurant, Cupcake Shop, Wings Joint, Caribbean Restaurant, Hardware Store, Performing Arts Venue, Convenience Store, French Restaurant, Bookstore, Bike Shop, Campground, Gas Station / Garage, Parking, Salad Place, Art Gallery, Video Game Store, Toy / Game Store, Chinese Restaurant, Event Space, Vegetarian / Vegan Restaurant, Sushi Restaurant, Convention Center, Latin American Restaurant, Spa / Massage, Paper / Office Supplies Store, Candy Store, Camera Store, Breakfast Spot, Southern / Soul Food Restaurant, Cosmetics Shop, Community College, Fried Chicken Joint, Plaza, Dessert Shop, Cemetery, Museum, Bagel Shop, Arcade, Concert Hall, Athletic & Sport, Middle Eastern Restaurant, Theater, Medical School, Tea Room, Movie Theater, Comedy Club, Post Office, Seafood Restaurant, Scenic Lookout, Housing Development, Synagogue, Donut Shop, General Entertainment, Pool, Japanese Restaurant, Arts & Entertainment, Pet Store, German Restaurant, Indian Restaurant, Garden, Hot Dog Joint, Steakhouse, Bowling Alley, Smoke Shop, Pool Hall, Harbor / Marina, Thai Restaurant, Bakery, Food, Ramen / Noodle House, College Theater, Mediterranean Restaurant, Beer Garden, African Restaurant, Outdoors & Recreation, River, Sorority House, Beach, Casino, Malaysian Restaurant, High School, Snack Place, Taxi, College & University, Record Shop, Temple, Historic Site, Arts & Crafts Store, Rest Area, Furniture / Home Store, History Museum, Recycling Facility, Bridal Shop, Library, Nail Salon, Professional & Other Places, Nursery School, Sculpture Garden, Antique Shop, Taco Place, South American Restaurant, Law School, Thrift / Vintage Store, Brazilian Restaurant, Winery, Greek Restaurant, Falafel Restaurant, Tapas Restaurant, City, Eastern European Restaurant, Korean Restaurant, Ski Area, Rental Car Location, Spiritual Center, Science Museum, Car Dealership, Flea Market, Art Museum, Gift Shop, Portuguese Restaurant, Flower Shop, Hobby Shop, Car Wash, Board Shop, Brewery, Cajun / Creole Restaurant, Mac & Cheese Joint, Shop & Service, Vietnamese Restaurant, Video Store, Travel & Transport, Dim Sum Restaurant, Racetrack, Elementary School, Zoo, Design Studio, Gaming Cafe, Swiss Restaurant, Travel Lounge, Trade School, Australian Restaurant, Funeral Home, Shrine, Peruvian Restaurant, College Stadium, Fraternity House, Bike Rental / Bike Share, Filipino Restaurant, Arepa Restaurant, Turkish Restaurant, Embassy / Consulate, Aquarium, Scandinavian Restaurant, Middle School, Financial or Legal Service, Fish & Chips Shop, Mosque, Afghan Restaurant, Motorcycle Shop, Fair, Ethiopian Restaurant, Distillery, Gluten-free Restaurant, Argentinian Restaurant, Moroccan Restaurant, Nightlife Spot, Planetarium, Storage Facility, Molecular Gastronomy Restaurant, Internet Cafe, Military Base, Newsstand, Public Art, Market, Photography Lab, Garden Center, Music School, Castle, Pet Service.

Για την ανάλυση μας, από αυτές τις κατηγορίες θα επιλέξουμε αυτές που εμφανίζονται περισσότερες φορές.

Στο πρώτο στάδιο της επεξεργασίας του συνόλου δεδομένων μας αναζητήσαμε τις 10 πιο δημοφιλείς κατηγορίες κοινοποίησης παρουσίας ώστε να τις χρησιμοποιήσουμε σαν βάση για τις αναλύσεις μας.
Όπως προέκυψε, οι 10 δημοφιλέστερες κατηγορίες που διαθέτουμε στα δεδομένα μας είναι οι εξής.

- **Bar** με 15978 κοινοποιήσεις παρουσίας
- **Home** (private) με 15382 κοινοποιήσεις παρουσίας
- **Office** με 12740 κοινοποιήσεις παρουσίας
- **Subway** με 9348 κοινοποιήσεις παρουσίας
- **Gym / Fitness Center** με 9171 κοινοποιήσεις παρουσίας
- **Coffee Shop** με 7510 κοινοποιήσεις παρουσίας
- **Food & Drink Shop** με 6596 κοινοποιήσεις παρουσίας
- **Train Station** με 6408 κοινοποιήσεις παρουσίας
- **Park** με 4804 κοινοποιήσεις παρουσίας
- **Neighborhood** με 4604 κοινοποιήσεις παρουσίας

Στον πίνακα παρακάτω φαίνεται ότι οι κατηγορίες που επιλέξαμε να χρησιμοποιήσουμε καλύπτουν λίγο παραπάνω από το 40% του αρχικού συνόλου δεδομένων μας. Στην δεύτερη στήλη παρουσιάζουμε επίσης τα ποσοστά με τα οποία εμφανίζεται ξεχωριστά η κάθε κατηγορία μέσα σε ολόκληρο το σύνολο ενώ στην τρίτη στήλη βλέπουμε πόσο καταλαμβάνει ποσοστιαία η κάθε κατηγορία μέσα στο υποσύνολο των 10 που έχουμε ξεχωρίσει.

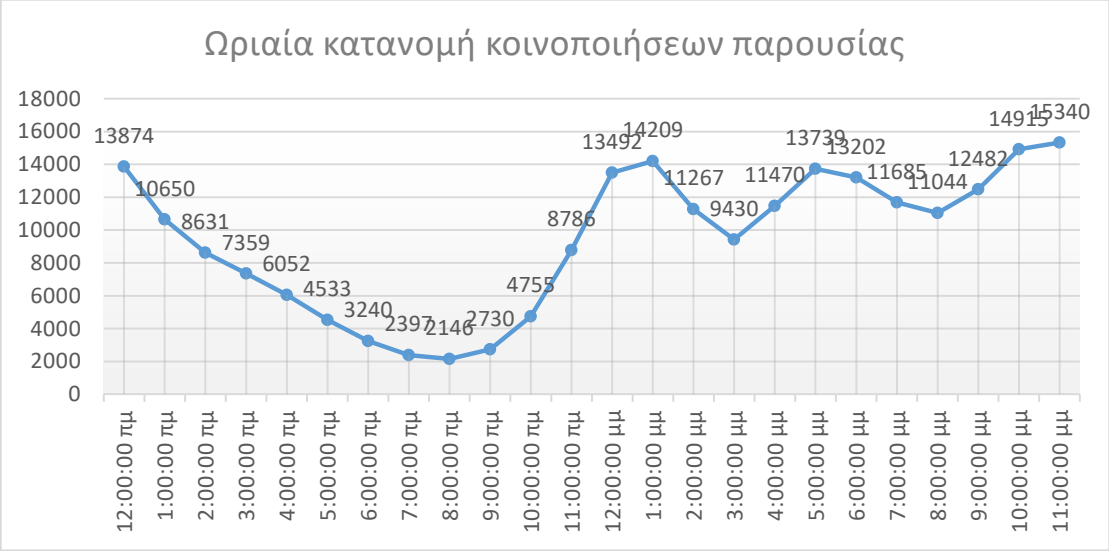
Κατηγορία	Ποσοστό (%) επί του συνόλου των δεδομένων	Ποσοστό (%) επί του υποσυνόλου των επιλεγμένων 10 κατηγοριών
Bar	7,02552	17,26586
Home	6,763459	16,62182
Office	5,601773	13,76687
Subway	4,110312	10,10147
Gym / Fitness Center	4,032485	9,910202
Coffee Shop	3,302144	8,115322
Food & Drink Shop	2,900259	7,127652
Train Station	2,817595	6,924498
Park	2,112317	5,191213
Neighborhood	2,024377	4,975092
Σύνολο	40,690241	100

Πίνακας 1 Οι δέκα δημοφιλέστερες κατηγορίες

Στην εικόνα 11 βλέπουμε τις κατηγορίες που έχουμε ξεχωρίσει και τις αριθμητικές τιμές των κοινοποιήσεων παρουσίας που υπάρχουν σε αυτές στο σύνολο δεδομένων μας.

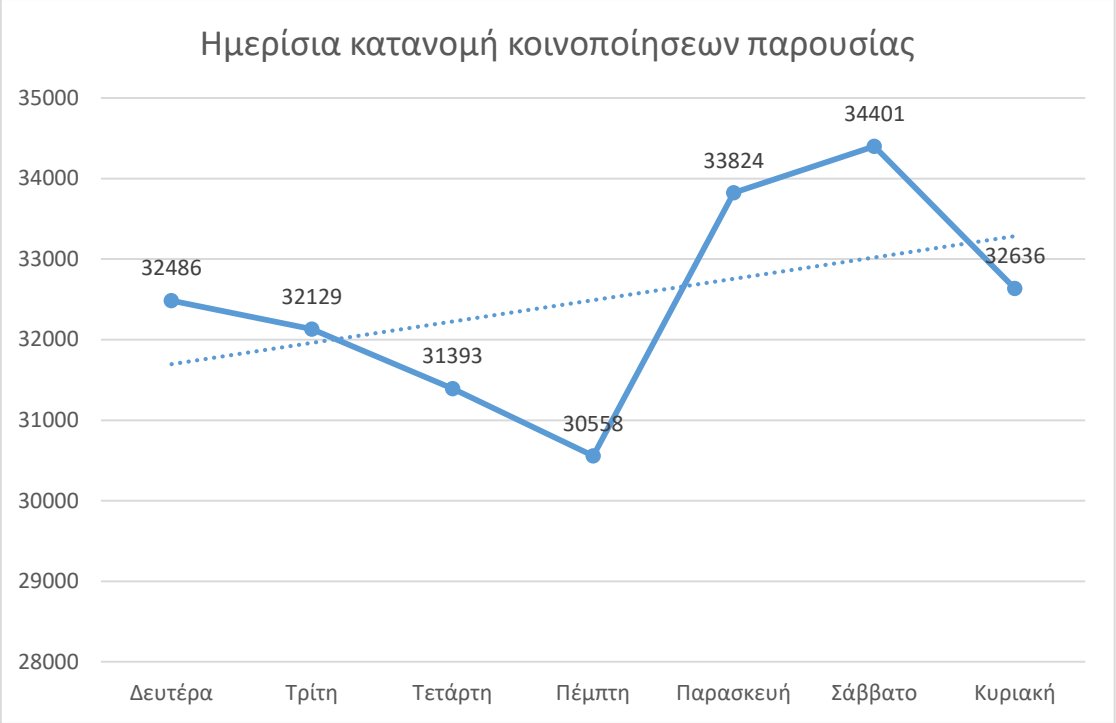
Εικόνα 11 Οι δέκα δημοφιλέστερες κατηγορίες με τις αριθμητικές τιμές

Στην εικόνα 12, κάναμε μια αρχική ανάλυση και παρουσίαση του πλήθους των κοινοποιήσεων παρουσίας ανεξαρτήτως τοποθεσίας ώστε να χρησιμοποιηθεί αργότερα για την εξέταση των τάσεων των ανθρώπων ανάλογα με την ώρα της ημέρας εξάγοντας κάποια πρώτα συμπεράσματα για την κινητικότητα των χρηστών με ωρολογιακή βάση.



Εικόνα 12 Ωριαία κατανομή κοινοποιήσεων παρουσίας

Στην εικόνα 13, εφαρμόσαμε την ίδια ανάλυση και οπτικοποιήσαμε τα αποτελέσματα για να εξετάσουμε τις τάσεις για κοινοποίηση παρουσίας ανάλογα με την ημέρα εξάγοντας κάποια αρχικά συμπεράσματα με ημερολογιακή βάση.



Εικόνα 13 Ημερήσια κατανομή κοινοποιήσεων παρουσίας

8. Τεχνική

8.1. Μοντελα Markov

Στη συνέχεια θα μελετήσουμε μαθηματικά μοντέλα τα οποία ανήκουν σε μια μεγάλη οικογένεια στοχαστικών-πιθανοθεωρητικών μοντέλων, τα οποία ονομάζονται μοντέλα εξάρτησης του Markov ή αλλιώς Μαρκοβιανά μοντέλα.
Θα εισαγάγουμε αρχικά την έννοια της αλυσίδας Markov (Markov Chain).
Η θεώρηση μιας ακολουθίας ενδεχομένων ως αλυσίδα Markov στηρίζεται, πολύ απλά, στην ιδέα ότι κάθε ένα από τα ενδεχόμενα εξαρτάται μόνο από το αμέσως προηγούμενό του ή αλλιώς το κάθε ενδεχόμενο καθορίζει με κάποια πιθανότητα το αμέσως επόμενό του. Αν αυτή η εξάρτηση επεκταθεί και σε 2,3,...,k προηγούμενα ενδεχόμενα τότε μιλάμε για αλυσίδες Markov 2ης, 3ης, ..., kης τάξης.
Ήδη από τη δεκαετία του 1970 τα μοντέλα αυτά χρησιμοποιούνταν και χρησιμοποιούνται ακόμα με σκοπό την αναγνώριση και επεξεργασία εικόνας, ήχου κ.α. και υπάρχει πλούσια βιβλιογραφία πάνω στα θέματα αυτά. Η πιο απλή εξήγηση για τα παραπάνω είναι το γεγονός ότι σε οποιοδήποτε κωδικοποιημένο σύστημα επικοινωνίας όπως στις φυσικές γλώσσες, υπάρχει μια εσωτερική δομή που καθορίζει κάποιο είδος εξάρτησης των συμβόλων. Για παράδειγμα, στην αγγλική γλώσσα το γράμμα Q ακολουθείται σχεδόν πάντοτε από το U, άρα η πιθανότητα να εμφανιστεί το U σε μια θέση δεν είναι πάντα ίδια αλλά εξαρτάται από το αν προηγήθηκε το Q. Για την ακρίβεια, ο ίδιος ο Ρώσος Μαθηματικός Andrey Markov (1856-1922) οδηγήθηκε στην σύλληψη της έννοιας των ομώνυμων αλυσίδων, μελετώντας τις εναλλαγές φωνηέντων και συμφώνων σε κάποιο ποίημα του Pushkin όπως αναφέρεται στο σύγγραμμα “The Life and Work of A. A. Markov” των Basharin G., Langville A., Naumov V.

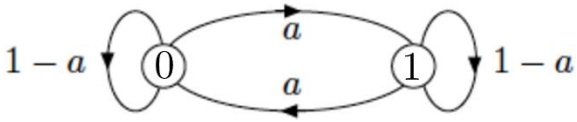
8.2. Στοχαστική διαδικασία Markov

Οι Μαρκοβιανές αλυσίδες είναι η σημαντικότερη κατηγορία των στοχαστικών διαδικασιών όσον αφορά στις εφαρμογές. Η μελέτη των Μαρκοβιανών αλυσίδων ξεκίνησε από τον ομώνυμο Ρώσσο μαθηματικό καθηγητή του Πανεπιστημίου της Μόσχας το 1907. Το αρχικό του κίνητρο ήταν να γενικεύσει το πρόβλημα σε μία σειρά από ανεξάρτητες δοκιμές Bernoulli τοποθετώντας μία εξάρτηση για την κάθε δοκιμή σε σχέση με την προηγούμενη. Σήμερα, οι εφαρμογές των Μαρκοβιανών αλυσίδων είναι σημαντικές και πολύ εκτεταμένες με εξαιρετικά μεγάλη δυναμική για νέες εφαρμογές.
Σε γενικές γραμμές μια διαδικασία Markov με διακριτό χώρο καταστάσεων ονομάζεται αλυσίδα Markov. Ένα σύνολο από τυχαίες μεταβλητές $\{X_t\}$ αποτελούν μία αλυσίδα Markov όταν η πιθανότητα η επόμενη τιμή (κατάσταση) να είναι ίση με x_{t+1} εξαρτάται μονάχα από την παρούσα κατάσταση x_t και όχι από οποιαδήποτε άλλη τιμή του παρελθόντος. Η ιδιότητα αυτή είναι γνωστή σαν έλλειψη μνήμης (memoryless property) και περιορίζει τη γενικότητα των διαδικασιών Markov.

8.3. Πρωτοτάξιο μοντέλο Markov

Μια αλυσίδα Markov πρώτης τάξης ορίζεται ως μια στοχαστική ανέλιξη διακριτών καταστάσεων σε διακριτό χρόνο με τη Μαρκοβιανή Ιδιότητα, δηλαδή με δεδομένη την παρούσα κατάσταση, οι παλαιότερες και οι μελλοντικές

καταστάσεις είναι ανεξάρτητες. Η ανέλιξη αποτελείται από την ακολουθία των τυχαίων μεταβλητών $x_1, x_2, x_3 \dots$, η οποία παίρνει τιμές από ένα αριθμήσιμο σύνολο S που ονομάζουμε χώρο καταστάσεων της αλυσίδας, με S= {1,2, … ,N}. Η τιμή x_t συμβολίζει την κατάσταση στην οποία βρίσκεται το σύστημα την χρονική στιγμή t.



Εικόνα 14 Απλή Μαρκοβιανή αλυσίδα δύο καταστάσεων

Το παραπάνω σχήμα απεικονίζει την απλούστερη μορφή μιας αλυσίδας Markov. Περιγράφει μια τυχαία κίνηση στο χώρο {0,1}. Η θέση που βρισκόμαστε τη στιγμή $t = 0,1,2,\dots$ είναι $x_t \in \{0,1\}$.

Η Μαρκοβιανή ιδιότητα ορίζει ότι η δεσμευμένη κατανομή των «μελλοντικών» παρατηρήσεων $x_{t+1}, x_{t+2}, x_{t+3}$ δεδομένου του «παρελθόντος» $x_1, x_2, \dots x_{t-1}, x_t$, εξαρτάται από το παρελθόν μόνο μέσω του x_t. Με άλλα λόγια, η γνώση της πιο πρόσφατης κατάστασης του συστήματος καθιστά τη λιγότερο πρόσφατη ιστορία άχρηστη.
Ορίζουμε:

$$P(X_t = x_t | X_1 = x_1, X_2 = x_2, \dots, X_{t-1} = x_{t-1}) = P(X_t = x_t | X_t = x_{t-1}) \ (1)$$

Οι Μαρκοβιανές αλυσίδες συχνά περιγράφονται από ένα κατευθυνόμενο γράφημα που οι άκρες του επιγράφουν τις πιθανότητες μετάβασης από τη μια κατάσταση στις άλλες.

Μια αλυσίδα Markov χαρακτηρίζεται από τον πίνακα των πιθανοτήτων μετάβασης, ή αλλιώς τον πίνακα μεταβάσεων.
Συνοπτικά ο πίνακας μεταβάσεων γράφεται ως εξής:

$$A = [a_{i,j}] = \begin{bmatrix} a_{1,1} & \cdots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{N,1} & \cdots & a_{N,N} \end{bmatrix} \ (2)$$

Ο πίνακας A είναι N x N διαστάσεων και το άθροισμα των σειρών του είναι 1.
Η πιθανότητα μετάβασης στο πρωτοτάξιο μοντέλο Markov ή αλλιώς τα στοιχεία του πίνακα A , δίνονται από την σχέση:

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \ (3)$$

η οποία δηλώνει, την πιθανότητα μετάβασης τη στιγμή t+1 στην κατάσταση j δεδομένου ότι η ακριβώς προηγούμενη κατάσταση τη στιγμή t είναι i.
Σε οποιαδήποτε στιγμή t, μπορούμε να βρούμε την ακολουθία παρατηρήσεων όταν γνωρίζουμε την κατάσταση Q_t και όταν το σύστημα μεταβαίνει από μια κατάσταση σε μια άλλη. Αν θεωρήσουμε ότι η ακολουθία παρατηρήσεων είναι O, με $O = Q = \{q_1, q_2, \dots, q_T\}$, τότε η πιθανότητα της υπολογίζεται ως

$$P(O = Q | A, \Pi) = P(q_1) \Pi_{t=2}^{T} P(q_t | q_{t-1}) = \pi_{q_1} a_{q_1, q_2}, \dots, a_{q_{T-1}, q_T} \ (4)$$

Ο όρος π_{q_1} εκφράζει την πιθανότητα να βρισκόμαστε στην αρχική κατάσταση και ο όρος a_{q_1,q_2} εκφράζει την πιθανότητα μετάβασης από τον κατάσταση q_1 στην κατάσταση q_2.
Στα κοινωνικά δίκτυα η κινητικότητα μπορεί να αναπαρασταθεί από μια στοχαστική διαδικασία. Επομένως, για να μοντελοποιθεί η ανθρώπινη κινητικότητα πρέπει να υπολογιστεί και να εξαχθεί ο πίνακας μεταβάσεων του πρωτοτάξιου μοντέλου Markov.
Στο παρόν σύνολο δεδομένων έχουμε διακρίνει 10 κύριες (δημοφιλέστερες) κατηγορίες. Έτσι προκειμένου να εκτιμηθεί σωστά η επόμενη μετάβαση (κοινοποίηση παρουσίας) του κάθε χρήστη πρέπει να υπολογιστεί και δημιουργηθεί ξεχωριστά για τον καθένα η πιθανοτική μήτρα μετάβασης (κοινοποιήσεων) μέσα σε αυτές τις 10 κατηγορίες.

8.4. Δευτεροτάξιο μοντέλο Markov

Μια $n^{ης}$ τάξεως αλυσίδα Markov, μπορεί να προκύψει αυτόματα από γενίκευση της Μαρκοβιανής ιδιότητας. Συγκεκριμένα, η σχέση αυτή τροποποιείται έτσι ώστε να συμπεριλάβει εξάρτηση στις n προηγούμενες παρατηρήσεις.
Σε αντίθεση με το πρωτοτάξιο Μαρκοβιανό μοντέλο, το δευτεροτάξιο Μαρκοβιανό μοντέλο λαμβάνει υπόψιν την παρούσα και την χρονικά αμέσως προηγούμενη κατάσταση της αλυσίδας προκειμένου να προβλέψει την χρονικά αμέσως επόμενη κατάσταση.
Η γενικότερη έκφραση του $n^{τάξιου}$ μοντέλου γράφεται ως:

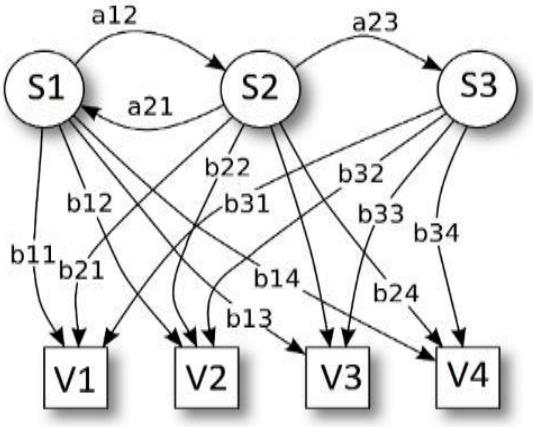
$$P(X_t = x_t|X_1 = x_1, X_2 = x_2, \dots, X_{t-1} = x_{t-1}) = P(X_t = x_t|X_{t-1} = x_{t-1}\, X_{t-2} = x_{t-2} \dots\, X_{t-n} = x_{t-n}) \quad (5)$$

Για το δευτεροτάξιο μοντέλο βάζουμε όπου n=2 και η σχέση γίνεται:

$$P(X_t = x_t|X_1 = x_1, X_2 = x_2, \dots, X_{t-1} = x_{t-1}) = P(X_t = x_t|X_{t-2} = x_{t-2}\, X_{t-1} = x_{t-1}) \quad (6)$$

8.5. Κρυφό μοντέλο Markov

Το κρυφό μοντέλο Markov (HMM) είναι ένα στατιστικό μοντέλο Markov κατά το οποίο το σύστημα που μοντελοποιείται υποτίθεται ότι είναι μια διαδικασία Markov με μη παρατηρημένες/κρυφές καταστάσεις. Το κρυφό μοντέλο Markov έχει τη μορφή του απλούστερου δυναμικού Bayesian δικτύου. Τα μαθηματικά πίσω από το HMM αναπτύχθηκαν από τον L. E. Baum και τους συνεργάτες του. Το κρυφό μοντέλο Markov, συνιστά ένα τύπο στοχαστικής μοντελοποίησης, κατάλληλο για μη στάσιμες στοχαστικές ακολουθίες, των οποίων οι στατιστικές ιδιότητες διέπονται από τυχαίες μεταβάσεις μεταξύ n διαφορετικών στάσιμων διεργασιών. Χρησιμοποιούνται δηλαδή για την μοντελοποίηση διεργασιών που είναι κατά τμήματα στάσιμες, δηλαδή όταν οι στατιστικές τους ιδιότητες στα στάσιμα τμήματα δεν μεταβάλλονται κατά το πέρασμα του χρόνου.



Εικόνα 15 Απλό κρυφό μοντέλο Markov

Στο παραπάνω σχήμα βλέπουμε ένα απλό παράδειγμα ενός κρυφού Μαρκοβιανού μοντέλου. Σύμφωνα με αυτό, τα V1, V2, V3, και V4 αποτελούν ένα διαφορετικό σύνολο παρατηρήσεων ενώ τα S1, S2 και S3 αποτελούν το σύνολο των κρυφών καταστάσεων.
Από το σχήμα βλέπουμε ότι η πιθανότητα να παράξουμε το V1 από την κατάσταση S1 είναι b11.
Υποθέτοντας $S = \{S_1, S_2, \ldots, S_N\}$ ένα σύνολο κρυφών καταστάσεων και $V=\{v_1, v_2,\ldots, v_M\}$ ένα σύνολο παρατηρήσεων, τότε, το HMM έχει τα εξής στοιχεία.

1) N: ο αριθμός των κρυφών καταστάσεων στο μοντέλο. $S = \{S_1, S_2, \ldots, S_N\}$
2) M: ο αριθμός των μοναδικών παρατηρήσεων του μοντέλου. $V=\{v_1, v_2,\ldots, v_M\}$
3) Πιθανότητα μετάβασης κατάστασης
$$A = [a_{ij}], a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$$
4) Πιθανότητα μιας παρατήρησης
$$B = [b_j(m)], b_j(m) = P(O_t = v_m | q_t = S_j)$$
5) Πιθανότητες αρχικών καταστάσεων
$$\Pi = [\pi_i], \pi_i = P(q_1 = S_i)$$

9. Πειραματική διαδικασία

9.1. Μετρικές αξιολόγησης

Προκειμένου να αξιολογήσουμε την ακρίβεια των αλγορίθμων που χρησιμοποιήθηκαν και τις προσομοιώσεις που εκτελέσαμε πρέπει να εκτελέσουμε και τις μετρικές αξιολόγησης.
Θα ορίσουμε λοιπόν την ικανότητα των αλγορίθμων μας ως προς την ακρίβεια της πρόβλεψης ως εξής:

$$A_u = \frac{N_u^T}{N_u}$$

και

$$A = \frac{\sum_u A_u}{|u|}$$

Ο πρώτος λόγος μας δίνει την ακρίβεια της πρόβλεψης του αλγορίθμου για έναν συγκεκριμένο χρήστη u μέσα από το σύνολο δεδομένων που έχουμε. Στον αριθμητή έχουμε τον όρο N_u^T που μας δείχνει τις σωστά εκτιμώμενες κοινοποιήσεις παρουσίας του αλγορίθμου μας και στον παρονομαστή έχουμε τον όρο N_u που εκφράζει τον συνολικό αριθμό κοινοποιήσεων παρουσίας του χρήστη u εντός του συνόλου δεδομένων μας.
Ο δεύτερος λόγος μας δίνει την καθολική απόδοση του αλγορίθμου για το σύνολο των χρηστών. Στον αριθμητή έχουμε το άθροισμα πάνω σε u χρήστες των επιμέρους αποδόσεων του αλγορίθμου μας και στον παρονομαστή είναι ο συνολικός αριθμός χρηστών που χρησιμοποιήσαμε για μελέτη από το σύνολο δεδομένων. Αυτός ο λόγος μας δίνει φυσικά και το ποσοστό επιτυχίας του μοντέλου μας.

Για να αξιολογήσουμε το κατά πόσο και πώς το αποτέλεσμα της στατιστικής ανάλυσης μπορεί να γενικευθεί στο σύνολο δεδομένων χρησιμοποιήσαμε την τεχνική επικύρωσης K-fold cross-validation στο μοντέλο μας.

Η γενική διαδικασία έχει ως εξής:

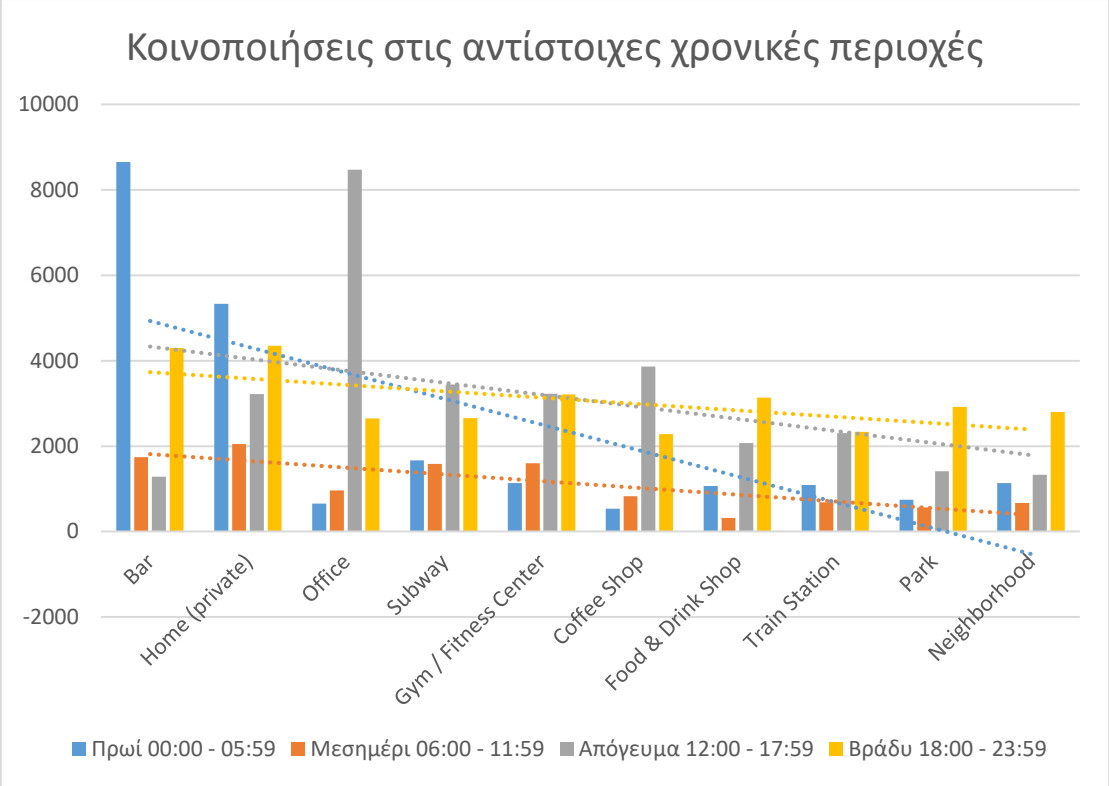
1) Ανακατεύουμε το σύνολο δεδομένων τυχαία.
2) Διαχωρίζουμε το σύνολο δεδομένων σε ομάδες k
3) Για κάθε μοναδική ομάδα:
    1) Παίρνουμε την ομάδα ως ομάδα αναφοράς (δοκιμαστικό σύνολο δεδομένων)
    2) Χρησιμοποιούμε τις υπόλοιπες ομάδες ως σύνολο δεδομένων εκπαίδευσης
    3) Ταιριάζουμε ένα μοντέλο πάνω στο σύνολο δεδομένων εκπαίδευσης και το αξιολογούμε πάνω στην ομάδα αναφοράς
    4) Κρατάμε σε ένα πίνακα τη βαθμολογία αξιολόγησης
4) Υπολογίζουμε την ικανότητα του μοντέλου μας το δείγμα των βαθμολογιών αξιολόγησης του μοντέλου

Τέλος, για τον υπολογισμό της ακρίβειας των μετρήσεων του χρησιμοποιήσαμε την τεχνική “The area under the ROC curve” υπολογίζοντας αρχικά την καμπύλη ROC.

Η δημιουργία της καμπύλης ROC μας δίνει μια πλήρη αναφορά ευαισθησίας / εξειδίκευσης και αποτελεί ένα θεμελιώδες εργαλείο για την αξιολόγηση διαγνωστικών δοκιμών.
Σε μια καμπύλη ROC ο πραγματικός θετικός ρυθμός ($TPR = \frac{TP}{TP+FN}$) σχεδιάζεται σε συνάρτηση με το ψευδοθετικό ρυθμό ($FPR = \frac{FP}{FP+TN}$) για διαφορετικά σημεία αποκοπής κάποιας παραμέτρου. Κάθε σημείο της καμπύλης ROC αντιπροσωπεύει ένα ζεύγος (TPR, FPR) που αντιστοιχεί σε ένα συγκεκριμένο όριο (κατώφλι) απόφασης. Η περιοχή κάτω από την καμπύλη ROC (AUC) είναι ένα μέτρο για το πόσο καλά μια παράμετρος μπορεί να κάνει διάκριση μεταξύ δύο ομάδων σε όλα τα πιθανά όρια απόφασης και μετρά ολόκληρη την δισδιάστατη περιοχή κάτω από ολόκληρη την καμπύλη ROC.

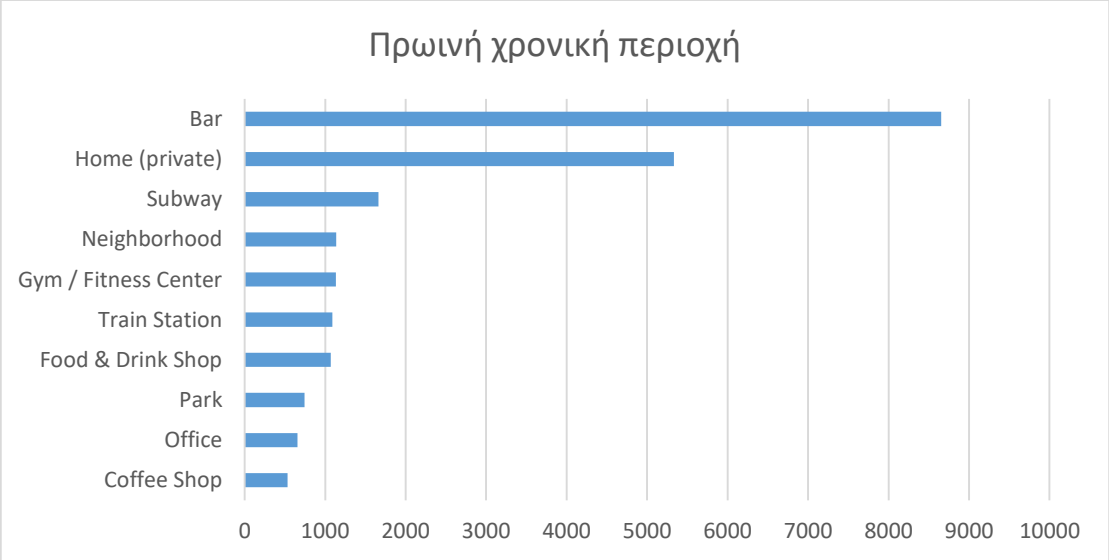
9.2. Κατηγοριακή πρόβλεψη

Σε αυτό το εδάφιο θα μιλήσουμε για τον τρόπο με τον οποίο υπολογίζουμε την πρόβλεψη της επόμενης κοινοποίησης παρουσίας κατηγοριακά.
Στα μοντέλα που αναπτύξαμε, θεωρήσαμε τις κοινοποιήσεις παρουσίας που έχουν κάνει οι χρήστες σε κάποια συγκεκριμένη κατηγορία τοποθεσιών όπως έχουμε διαχωρίσει προηγουμένως, ως χρονικά ακολουθιακούς πίνακες.
Έπειτα, υπολογίσαμε τους πίνακες μετάβασης για το πρωτοτάξιο και δευτεροτάξιο μοντέλο Markov. Μετά την εξαγωγή των πινάκων μετάβασης, τα μοντέλα μας είναι σε θέση να εκτιμήσουν σε ποια κατηγορία τοποθεσιών ο χρήστης u θα πραγματοποιήσει την επόμενη κοινοποίηση παρουσίας με δεδομένη πάντα την παρούσα (προηγούμενη) θέση του.
Παρομοίως ενεργήσαμε και για τον υπολογισμό της μήτρας μεταβάσεων για το κρυφό μοντέλο Markov. Ως κατάσταση παρατήρησης θεωρήσαμε τον χρόνο που έχουν πραγματοποιηθεί οι κοινοποιήσεις και για να απλουστεύσουμε τη διαδικασία χωρίσαμε το εικοσιτετράωρο σε τέσσερις εξάωρες χρονικές περιόδους.
Τα τέσσερα αυτά χρονικά διαστήματα δημιουργούν παρατηρήσιμες καταστάσεις. Το σύμπλεγμα κρυφών καταστάσεων αποτελείται από δέκα κατηγορίες. Λαμβάνουμε ως βάση τις συνολικές κοινοποιήσεις παρουσίας που έχει πραγματοποιήσει το σύνολο των χρηστών σε κάποια συγκεκριμένη κατηγορία και φυσικά την ώρα. Στη συνέχεια εφαρμόζονται αλγόριθμοι Baum-Welch στην ακολουθία που δημιουργήσαμε και είμαστε σε θέση να υπολογίσουμε και να προβλέψουμε σε ποια κατηγορία θα πραγματοποιηθεί η επόμενη κοινοποίηση παρουσίας κάποιου χρήστη.



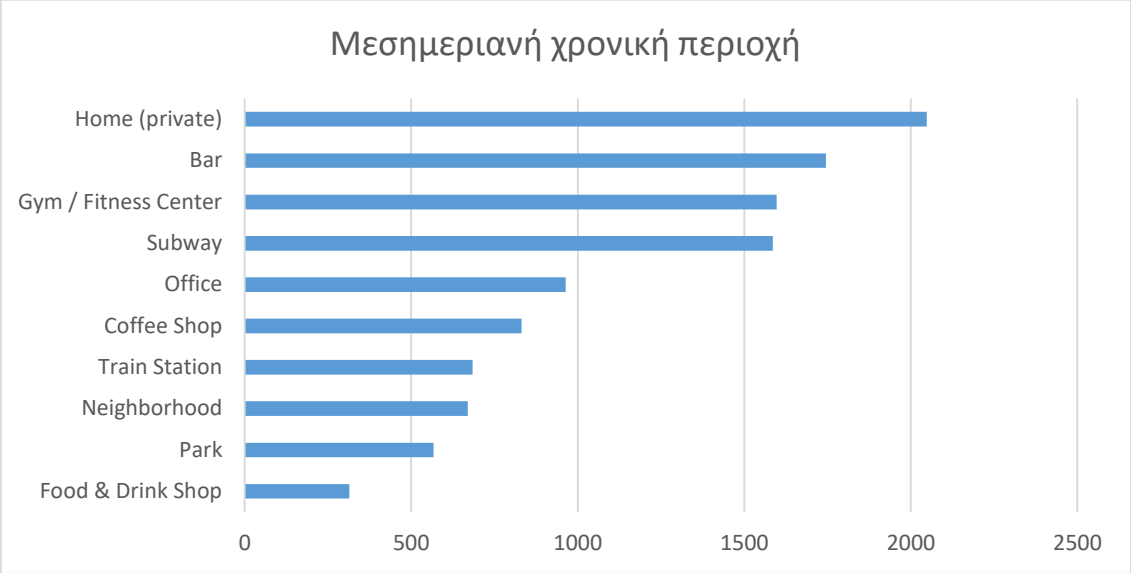
Εικόνα 16 Κατανομή κοινοποιήσεων παρουσίας σε τέσσερις κύριες χρονικές κατηγορίες

Η γραφική παράσταση της εικόνας 16 μας δείχνει το πλήθος των κοινοποιήσεων παρουσίας που έχει το σύνολο δεδομένων μας για τις 10 πιο δημοφιλείς κατηγορίες που έχουμε ξεχωρίσει. Παρατηρούμε ότι στο δίαστημα 00:00 με 6:00 οι περισσότερες κοινοποιήσεις έχουν πραγματοποιηθεί στην κατηγορία Bar. Από αυτό μπορούμε να εξάγουμε πληροφορίες για τις συνήθειες των χρηστών μας και να πούμε ότι μια μεγάλη πλειοψηφία εκείνο το διάστημα επιλέγει να διασκεδάσει, ενώ ακολουθούν οι χρήστες οι οποίοι περνούν το βράδυ στο σπίτι τους. Σαν τριτη πιο δημοφιλής κατηγορία είναι του «Subway». Αυτό μας δείχνει ότι υπάρχει μεταμεσονύκτια χρήση του υπόγειου σιδηροδρόμου (MMM), πιθανώς για την επιστροφή ή για την κατεύθυνση προς διασκέδαση.



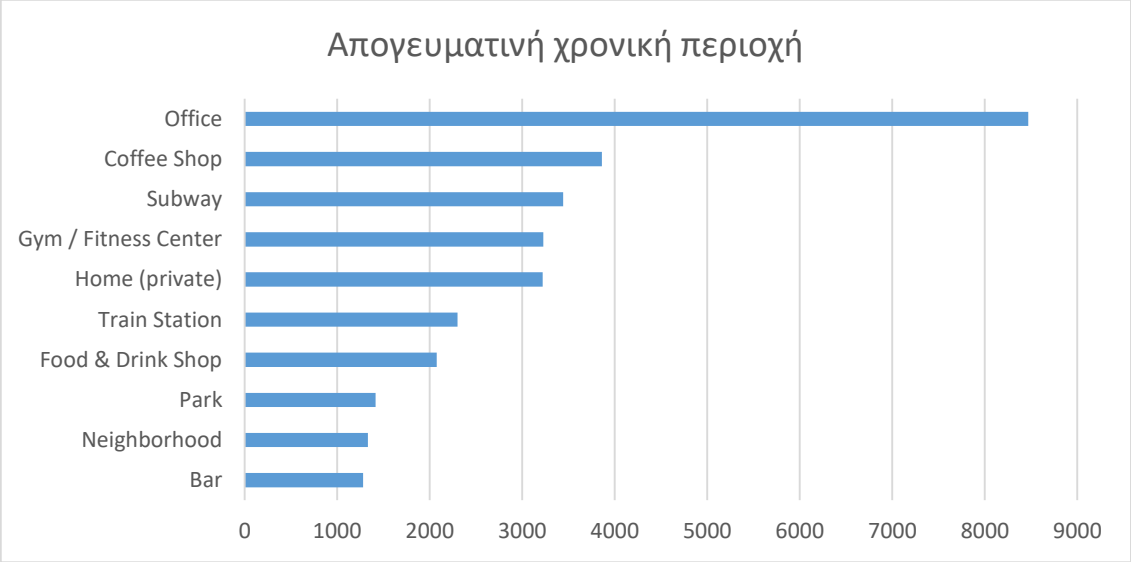
Εικόνα 17 Πρωινή χρονική περιοχή

Αρα, αν επιλέξουμε κάποιο χρήστη ώστε να εκτιμήσουμε την επόμενη του τοποθεσία μέσα σε αυτό το διάστημα, περιμένουμε το μοντέλο μας να μας προβλέψει είτε «Bar» είτε «Home» σαν τις επικρατέστερες επιλογές.



Εικόνα 18 Μεσημεριανή χρονική περιοχή

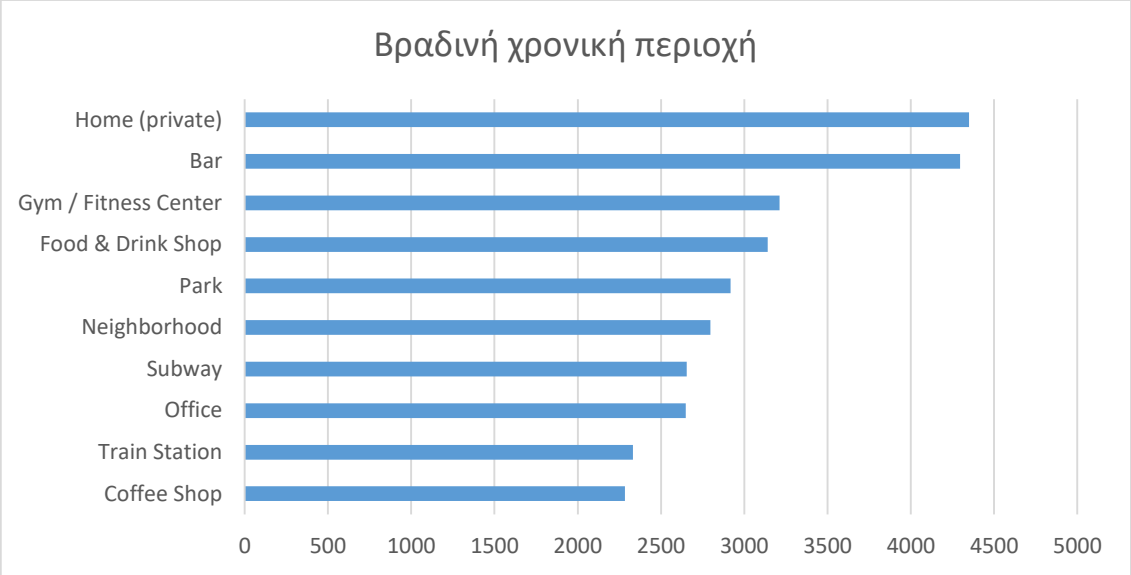
Στην δεύτερη χρονική περιοχή, τις «μεσημεριανές» ώρες όπως τι έχουμε ορίσει, δηλαδή από τις 6:00 ως τις 12:00, τα πράγματα είναι πιο δύσκολα για τις εκτιμήσεις μας, καθώς βλέπουμε ότι οι επικρατέστερες κατηγορίες όπως παρουσιάζονται στο σύνολο δεδομένων μας είναι 4, και είναι αρκετά «κοντά» μεταξύ τους. Η τοποθεσία «Home», έχει λίγο παραπάνω από 2000 κοινοποιήσεις, η τοποθεσία «Bar» περίπου 1750, ενώ οι «Gym/Fitness Center» και «Subway» έχουν λίγο πάνω από 1500. Περιμένουμε λοιπόν από τις προσομοιώσεις μας για τυχαίο χρήστη, να έχουμε κάποια πρόταση ανάμεσα σε αυτές τις 4.



Εικόνα 19 Απογευματινή χρονική περιοχή

Στην τρίτη κατηγορία, της απογευματινής περιοχής, από 12:00 μέχρι 18:00 βλέπουμε με μεγάλη διαφορά από τις υπόλοιπες την κατηγορία office. Οι ακόλουθες είναι πάρα πολύ κοντά σε πλήθος αναφορών. Οπότε για κάποιο

τυχαίο χρήστη, η πιο πιθανή πρόταση για την επόμενη κοινοποίηση μέσα σε αυτό το διάστημα θα περιμέναμε να είναι στην κατηγορία «Office».



Εικόνα 20 Βραδινή χρονική περιοχή

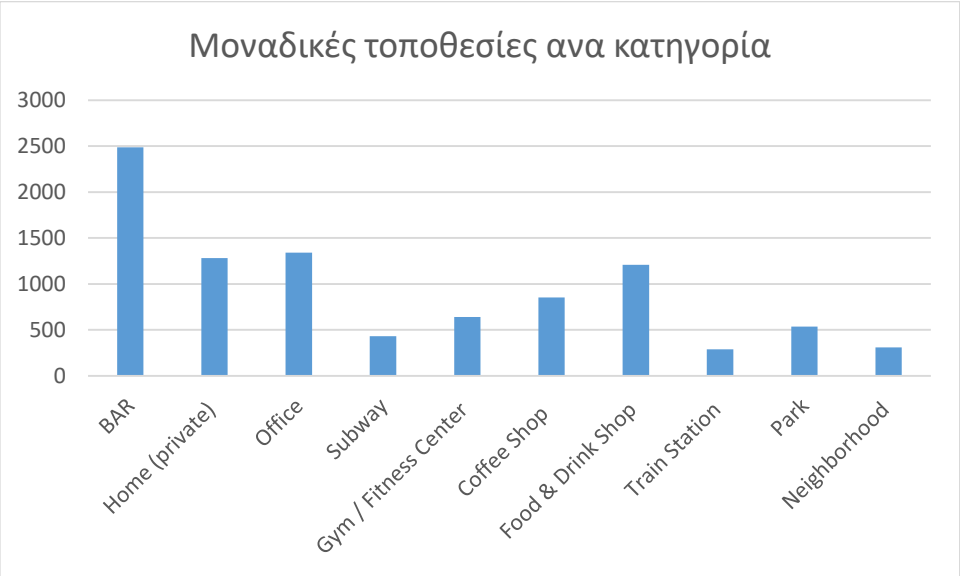
Στην βραδινή περιοχή, από 18:00 ως 00:00 ξεχωρίζουν πάλι οι δυο κατηγορίες της πρωινής περιοχής σαν επικρατέστερες, ωστόσο αυτή τη φορά δεν απέχουν πολύ σε πλήθος κοινοποιήσεων η μία από την άλλη. Ακολουθούν οι «Gym/Fitness Center» και η «Food & Drink Shop». Σε αυτή την κατηγορία για κάποιο τυχαίο χρήστη θα περιμέναμε μια πρόβλεψη κοινοποίησης από τα μοντέλα μας είτε στην κατηγορία «Home» είτε στην κατηγορία «Bar».

9.3. Προσομοιώσεις και αποτελέσματα

Κατά την πειραματική διαδικασία, στο στάδιο των προσομοιώσεων που εκτελέσαμε, αρχικά κάναμε εκτίμηση της επόμενης κατηγορίας μέσα στην οποία θα κάνει κοινοποίηση παρουσίας κάποιος χρήστης. Η κατηγορία, βέβαια, περιλαμβάνει αρκετές τοποθεσίες. Η καθεμία από τις κατηγορίες που επιλέξαμε να εισάγουμε στο δείγμα μας για μελέτη, περιείχε ένα συγκεκριμένο αριθμό από μοναδικές τοποθεσίες.
Η κατηγορία Bar είχε 2488 μοναδικές τοποθεσίες
Η κατηγορία Home (private) είχε 1282 μοναδικές τοποθεσίες
Η κατηγορία Office είχε 1341 μοναδικές τοποθεσίες
Η κατηγορία Subway είχε 430 μοναδικές τοποθεσίες
Η κατηγορία Gym / Fitness Center είχε 640 μοναδικές τοποθεσίες
Η κατηγορία Coffee Shop είχε 854 μοναδικές τοποθεσίες
Η κατηγορία Food & Drink Shop είχε 1209 μοναδικές τοποθεσίες
Η κατηγορία Train Station είχε 288 μοναδικές τοποθεσίες
Η κατηγορία Park είχε 536 μοναδικές τοποθεσίες
Η κατηγορία Neighborhood είχε 309 μοναδικές τοποθεσίες
Αυτό είναι πιο ευδιάκριτο στην εικόνα 21 από όπου έχουμε αφαιρέσει τα διπλότυπα. Οι αριθμοί κοινοποιήσεων μας δείχνουν σε ποιες κατηγορίες κάθε χρήστης έχει πραγματοποιήσει τουλάχιστον μία κοινοποίηση (αγνοώντας αν έχουν ξανακοινοποιήσει την παρουσία τους εκεί). Αυτό μας δίνει μια ιδέα για τις τοποθεσίες όπου οι χρήστες έχουν τη συνήθεια περισσότερο να κοινοποιούν την

παρουσία τους. Φαίνεται ότι η υψηλότερη προτίμηση στο δείγμα μας είναι της κατηγορίας «Bar».

Ο κώδικας γράφτηκε σε Matlab. Κομμάτια του κώδικα βρέθηκαν στο διαδίκτυο και αρκετά τμήματα που υπήρχαν στο stackoverflow προσαρμόστηκαν κατάλληλα και χρησιμοποιήθηκαν.



Εικόνα 21 Μοναδικές τοποθεσίες ανά κατηγορία

Στην πρώτη προσομοίωση, επιλέξαμε να εκτιμήσουμε μόνο μια κατηγορία η οποία είχε εξ αρχής την μεγαλύτερη πιθανότητα να είναι η επόμενη κατηγορία μέσα στην οποία ο χρήστης θα πραγματοποιήσει την επόμενη κοινοποίηση παρουσίας.
Αφού εκτιμήσουμε σωστά την κατηγορία, οι τοποθεσίες οι οποίες μας ενδιαφέρουν στη συνέχεια, είναι προφανές ότι υπάγονται στη συγκεκριμένη κατηγορία την οποία εκτιμήσαμε. Και η τοποθεσία στην οποία ο χρήστης θα κάνει την επόμενη κοινοποίηση παρουσίας, είναι αυτή που παρουσιάζει τη μεγαλύτερη πιθανότητα.
Και στις δύο περιπτώσεις, ακολουθήσαμε την έννοια της ταξινόμησης με βάση τη δημοτικότητα τόσο των κατηγοριών όσο και των τοποθεσιών. Οι εκτιμήσεις δημοτικότητας έγιναν αλγοριθμικά.
Έτσι λοιπόν, αναλόγως με το ποια θα προβλεφθεί σωστά ως η επόμενη κατηγορία, προβλέπουμε και την επόμενη τοποθεσία, η οποία δεν θα μπορούσε να είναι άλλη από αυτήν που παρουσιάζει το μεγαλύτερο αριθμό κοινοποιήσεων, δηλαδή τη μεγαλύτερη δημοτικότητα αφού έχουμε διατάξει το σύνολο σε φθίνουσα σειρά.
Τέλος, επιλέγουμε τις k δημοφιλέστερες τοποθεσίες (με $k = 5$, $k = 10$ και $k = 15$) ως τις επόμενες τοποθεσίες στις οποίες θα γίνει κοινοποίηση παρουσίας από το χρήστη.
Η συνολική ακρίβεια εκτίμησης του μοντέλου που μελετάται, εκφράζεται από το λόγο

$$\frac{\textit{σωστά εκτιμημένες τοποθεσίες}}{\textit{σωστά εκτιμημένες κατηγορίες}}$$

Παρακάτω παραθέτουμε τα αποτελέσματα της πρώτης πειραματικής διαδικασίας.

Το κρυφό μοντέλο Markov δείχνει να έχει την καλύτερη απόδοση από τα τρία εξεταζόμενα. Πιο συγκεκριμένα παρουσιάζει ακρίβεια πρόβλεψης περίπου 39,13% που είναι η καλύτερη των τριών με το πρωτοτάξιο μοντέλο Markov να ακολουθεί σχετικά κοντά (σε απόδοση) και το δευτεροτάξιο μοντέλο Markov να απέχει αρκετά από τα άλλα δύο και να παρουσιάζει τη χειρότερη απόδοση.

Μελετώμενο μοντέλο	*Απόδοση (%)*
Πρωτοτάξιο μοντέλο Markov	37,58
Δευτεροτάξιο μοντέλο Markov	25,94
Κρυφό μοντέλο Markov	39,13

Πίνακας 2 Απόδοση μελετώμενων μοντέλων με μία κατηγορία κοινοποιήσεων

Μελετώμενο μοντέλο	*Απόδοση (%) k=5*	*Απόδοση (%) k=10*	*Απόδοση (%) k=15*
Πρωτοτάξιο μοντέλο Markov	41,71	50,86	59,73
Δευτεροτάξιο μοντέλο Markov	35,87	46,03	51,24
Κρυφό μοντέλο Markov	46,34	52,35	61,94

Πίνακας 3 Απόδοση μελετώμενων μοντέλων με μία κατηγορία κοινοποιήσεων για διαφορετικό πλήθος υποψήφιων τοποθεσιών

Στην δεύτερη προσομοίωση, επιλέξαμε να εκτιμήσουμε τις τρείς πιο πιθανές κατηγορίες στις οποίες ο χρήστης θα μπορούσε να βρεθεί.
Η διαφορά αυτής της μεθόδου από την προηγούμενη είναι ότι σε αυτή την περίπτωση, αν η τοποθεσία, δεν μπορούσε να βρεθεί στην πρώτη κατηγορία, τότε ψάχναμε με τη σειρά δημοτικότητας στην δεύτερη και την τρίτη κατηγορία τοποθεσιών.
Αρχικά δηλαδή, ξεκινάμε εκτιμώντας τις τρεις πιο πιθανές κατηγορίες επόμενης κοινοποίησης παρουσίας κάποιου χρήστη. Έπειτα, για να εκτιμήσουμε και την τοποθεσία, αναζητούμε μέσα στις κατηγορίες που έχουμε ξεχωρίσει με σειρά δημοτικότητας σειριακά την σωστή τοποθεσία.
Παρατηρήθηκε ότι όταν αυξήθηκαν οι κατηγορίες τοποθεσιών, και άρα και οι λίστες με τις ακριβείς (μοναδικές) κατηγορίες, τα αποτελέσματα των αλγορίθμων ήταν περισσότερο ικανοποιητικά. Συγκεκριμένα, και πάλι ο αλγόριθμος πρόβλεψης του κρυφού μοντέλου Markov ήταν ο καλύτερος όλων με τον αλγόριθμο του πρωτοτάξιου να ακολουθεί και τη χειρότερη απόδοση να παρουσιάζει ο αλγόριθμος του δευτεροτάξιου μοντέλου.

Πιο αναλυτικά, στους παρακάτω πίνακες βλέπουμε τα αποτελέσματα των προσομοιώσεων

Μελετώμενο μοντέλο	*Απόδοση (%)*
Πρωτοτάξιο μοντέλο Markov	52,11
Δευτεροτάξιο μοντέλο Markov	42,49
Κρυφό μοντέλο Markov	53,76

Πίνακας 4 Απόδοση μελετώμενων μοντέλων με τρεις κατηγορίες κοινοποιήσεων

Μελετώμενο μοντέλο	*Απόδοση (%) k=5*	*Απόδοση (%) k=10*	*Απόδοση (%) k=15*
Πρωτοτάξιο μοντέλο Markov	40,12	48,91	58,33
Δευτεροτάξιο μοντέλο Markov	32,74	43,38	49,73
Κρυφό μοντέλο Markov	45,63	49,34	60,22

Πίνακας 5 Απόδοση μελετώμενων μοντέλων με τρεις κατηγορίες κοινοποιήσεων για διαφορετικό πλήθος υποψήφιων τοποθεσιών

Στην τρίτη φάση προσομοιώσεων, επιλέξαμε να εκτιμήσουμε ανάμεσα από τις δέκα κατηγορίες που είχαμε ξεχωρίσει νωρίτερα, την επόμενη κατηγορία τοποθεσιών που θα βρεθεί ο χρήστης.
Ομοίως με την προηγούμενη προσομοίωση, αφού εκτιμηθεί σωστά η επόμενη κατηγορία, αναζητούμε την ακριβή τοποθεσία ανάμεσα στις δέκα κατηγορίες που έχουμε ξεχωρίσει με σειρά δημοτικότητας σειριακά.
Τα αποτελέσματα στις εκτιμήσεις των αλγορίθμων όσον αφορά την κατηγορία, ήταν πάρα πολύ ικανοποιητικά. Για την εκτίμηση της ακριβής θέσης όμως είδαμε ότι οι αλγόριθμοι παρουσίασαν πολύ χαμηλές αποδόσεις λόγο της πληθώρας των μοναδικών τοποθεσιών ανάμεσα στις δέκα αυτές κατηγορίες. Ωστόσο, και πάλι ο αλγόριθμος πρόβλεψης του κρυφού μοντέλου Markov ήταν ο καλύτερος όλων με τον αλγόριθμο του πρωτοτάξιου να ακολουθεί και τη χειρότερη απόδοση να παρουσιάζει ο αλγόριθμος του δευτεροτάξιου μοντέλου.

Συνοπτικά παρουσιάζουμε στους παρακάτω πίνακες τα αποτελέσματα των προσομοιώσεων

Μελετώμενο μοντέλο	*Απόδοση (%)*
Πρωτοτάξιο μοντέλο Markov	73,48
Δευτεροτάξιο μοντέλο Markov	64,21
Κρυφό μοντέλο Markov	75,83

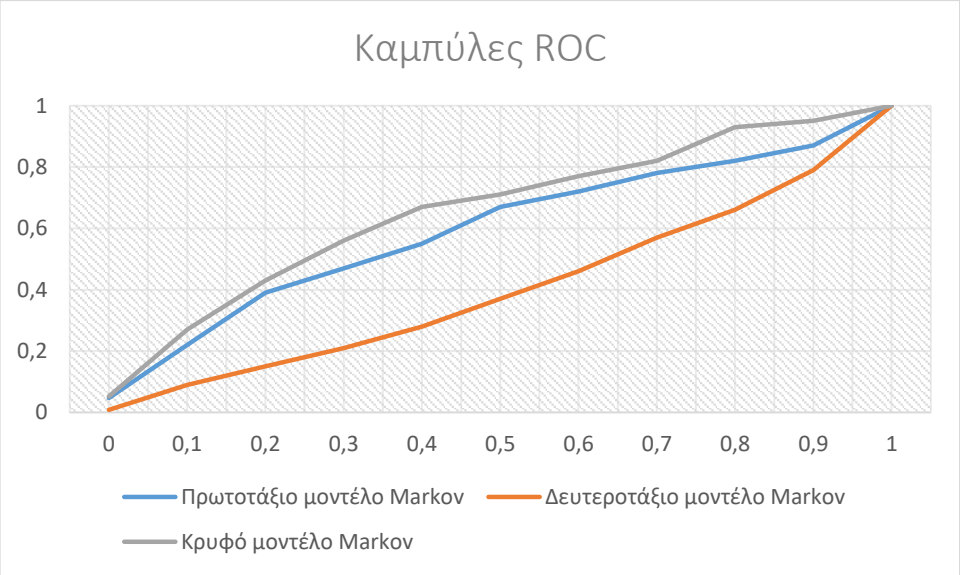
Πίνακας 6 Απόδοση μελετώμενων μοντέλων με δέκα κατηγορίες κοινοποιήσεων

Μελετώμενο μοντέλο	*Απόδοση (%) k=5*	*Απόδοση (%) k=10*	*Απόδοση (%) k=15*
Πρωτοτάξιο μοντέλο Markov	33,18	46,89	53,93
Δευτεροτάξιο μοντέλο Markov	26,72	39,67	45,16

Κρυφό μοντέλο Markov	37,70	48,08	55,34

Πίνακας 7 Απόδοση μελετώμενων μοντέλων με δέκα κατηγορίες κοινοποιήσεων για διαφορετικό πλήθος υποψήφιων τοποθεσιών

Αφού ολοκληρώσαμε τις προσομοιώσεις μας και την πειραματική διαδικασία, προχωρήσαμε σε ένα μέτρηση της ακρίβειας των μοντέλων μας μέσα από την εκτίμηση της περιοχής κάτω από την καμπύλη ROC (AUC).
Όπως προ είπαμε, ένας τρόπος για να αποφανθούμε για την ποιότητα των αλγορίθμων μας είναι η καμπύλη Receiver Operating Characteristic (ROC). Γενικά, πρόκειται για μια εξαιρετική μέθοδο για την ταυτόχρονη σύγκριση πολλαπλών αλγορίθμων.
Για να υπολογιστεί και να σχεδιαστεί η καμπύλη ROC, υπολογίζουμε τον ψευδοθετικό ρυθμό (λανθασμένες προβλέψεις) και τον πραγματικό θετικό ρυθμός (σωστές προβλέψεις).
Για να σχεδιάσουμε μια καμπύλη παράγουμε ένα ζευγάρι αριθμών (FPR, TPR) και τις τοποθετούμε σε άξονες Χ και Υ ως συντεταγμένες, αντίστοιχα.
Σε ιδανικές περιπτώσεις, ένα μοντέλο, το οποίο είναι τέλειο, έχει TPR =1 και FPR = 0 και θα παρουσίαζε την υψηλότερη δυνατή καμπύλη ROC. Το μέτρο που μας δίνει την απόδοση του αλγορίθμου είναι η κλίση της καμπύλης.
Όταν συγκρίνουμε τους αλγορίθμους με αυτή τη μέθοδο, αυτός με την υψηλότερη καμπύλη ROC είναι προφανώς ο καλύτερος. Το ίδιο ισχύει και για την περιοχή κάτω από την καμπύλη ROC, την AUC την οποία έχουμε περιγράψει σε προηγούμενο εδάφιο. Ο αλγόριθμος με την υψηλότερη AUC είναι αυτός που παρουσιάζει την καλύτερη απόδοση.
Στο σχήμα παρακάτω βλέπουμε ότι ο αλγόριθμος κρυφού μοντέλου Markov είναι ελαφρώς καλύτερος από τον αλγόριθμος πρωτοτάξιου μοντέλου Markov.
Όπως μπορούμε να δούμε, η καμπύλη ROC, και η περιοχή AUC είναι καλύτερες και παρουσιάζουν παρόμοια αποτελέσματα με τις αποδόσεις που έχουμε υπολογίσει παραπάνω. Το κρυφό μοντέλο Markov υπερέχει πάρα πολύ του δευτεροτάξιου μοντέλου Markov το οποίο φαίνεται να είναι το λιγότερο χρήσιμο.



Εικόνα 22 Καμπύλες ROC και περιοχή AUC των μελετώμενων μοντέλων

AUC για το πρωτοτάξιο μοντέλο Markov = 0,601
AUC για το δευτεροτάξιο μοντέλο Markov = 0,408
AUC για το κρυφό μοντέλο Markov = 0,664

Οι υπηρεσίες με βάση τοποθεσίες παρουσιάζουν νέες προκλήσεις, καθώς αποκαλύπτουν όχι μόνο την τοποθεσία ενός χρήστη αλλά και ένα επιπλέον επίπεδο πληροφοριών σχετικά με τους φυσικούς χώρους που επισκέπτεται. Έτσι, οι πάροχοι υπηρεσιών μπορούν πλέον να έχουν πρόσβαση στα δεδομένα σχετικά με πλήθος παραγόντων που μπορούν να επηρεάσουν τους χρήστες όταν αποφασίζουν ποιο μέρος να επισκεφθούν, που περιλαμβάνουν προσωπικά ενδιαφέροντα, κοινωνική επιρροή, χωρική εγγύτητα και χρονικό πλαίσιο. Έτσι μπορούν οι παρούσες τεχνικές να επεκταθούν πέρα από την πρόβλεψη των χωρικών τροχιών, ώστε να υπολογίσουν ακριβέστερα την επόμενη θέση που θα επισκέπτεται ο χρήστης. Ωστόσο, μαζί με τις νέες ευκαιρίες που προσφέρονται από τα πρόσθετα επίπεδα πληροφόρησης που περιλαμβάνονται σε αυτά τα δεδομένα, έρχονται και οι προκλήσεις.

Αρχικά, η προβλεψιμότητα καθίσταται πιο δύσκολη, αφού οι αλγόριθμοι πρόβλεψης γίνονται πολλαπλώς πολυπλοκότεροι για τον υπολογισμό της ακριβούς θέσης, μεταξύ χιλιάδων, που θα βρίσκεται ένας χρήστης. Η πειραματική μας προσέγγισή έχει διερευνήσει αυτό το συμβιβασμό ο οποίος κατέληξε να είναι μεγαλύτερη τροχοπέδη από αυτή που υπολογίζαμε. Όμως είδαμε πως είναι εφικτό, χάρη στα εκτενή διαθέσιμα δεδομένα κινητικότητας των χρηστών, να προσδιοριστούν και να αξιοποιηθούν πολλά διαφορετικά χαρακτηριστικά πρόβλεψης για τον υπολογισμό της πιθανότητας να επισκεφτεί κάποιος μια συγκεκριμένη τοποθεσία.

Ωστόσο, κάθε μεμονωμένο χαρακτηριστικό προσφέρει ελάχιστη πληροφόρηση για τη συμπεριφορά του χρήστη και, συνεπώς, (μεμονωμένα) δεν είναι σε θέση να παρέχει μια καθολική απάντηση στην συνολική διαδικασία πρόβλεψης. Ένας αποτελεσματικός τρόπος αντιμετώπισης αυτού του προβλήματος είναι η εκπαίδευση μοντέλων που μπορούν να αξιοποιήσουν τη συνδυασμένη ισχύ πολλαπλών χαρακτηριστικών. Επιπλέον, στην άλλη διάσταση, η εξαιρετική ακρίβεια των διαθέσιμων δεδομένων, καθιστά την κάθε εγγραφή σχεδόν μοναδική, ή με άλλα λόγια, «σπάνια» υπό την έννοια ότι δύσκολα θα βρεθούν πολλές ίδιες εγγραφές για διαφορετικούς χρήστες, και άρα, το σετ που θα χρησιμοποιούμε για την εκπαίδευση των μοντέλων απαιτεί συνδυασμό δεδομένων που συλλέγονται από πολλούς χρήστες στην ίδια γεωγραφική περιοχή.

Συνεχίζοντας, όπως είδαμε, οι περισσότεροι χρήστες στα κοινωνικά δίκτυα έχουν λίγες κοινοποιήσεις παρουσίας (σπάνια μοιράζονται δημοσίως την κοινοποίηση). Χρειάζονται έξυπνοι τρόποι ή ακόμη και συστήματα ανταμοιβής, τεχνικές παιχνιδοποίησης, ώστε να δίνουμε κίνητρα στους χρήστες να μας προσφέρουν τα δεδομένα τους και έχουν ενεργή συμμετοχή μιας και πρόκειται για ένα πάρα πολύ βασικό ζήτημα σε αυτές τις υπηρεσίες. Πολλά από τα χαρακτηριστικά που εξάγαμε στην εργασία, δεν περιλαμβάνουν ειδικές πληροφορίες χρηστών, και έτσι ο μόνος τρόπος που θα μπορούσαν να

αξιοποιηθούν είναι για τη βελτίωση συστάσεων σε μια συγκεκριμένη κατηγορία χρηστών, όχι προσωποποιημένα.

Επιπλέον, οι παρατηρήσεις μας υπογραμμίζουν ότι όχι μόνο η προβλεψιμότητα των χρηστών αλλάζει με την πάροδο του χρόνου, αλλά και ο τρόπος με τον οποίο διάφοροι παράγοντες καθορίζουν την κινητικότητα των χρηστών μπορεί να παρουσιάσει διαχρονικές διακυμάνσεις.
Για παράδειγμα, είδαμε ότι οι χρήστες τείνουν να κινούνται προς γεωγραφικά πλησιέστερους χώρους κατά το βράδυ και είναι λιγότερο πιθανό να επισκεφθούν τις ιστορικά παρατηρημένες τοποθεσίες τους κατά τη διάρκεια του Σαββατοκύριακου. Αυτό έχει δύο σημαντικές συνέπειες.

Πρώτον, τα νέα μοντέλα που θα καταγράφουν και αναπαράγουν η συμπεριφορά των χρηστών μέσω κινητών πρέπει να υπολογίζουν ρητά και να εκμεταλλεύονται αυτές τις χρονικές παραμέτρους. Δεύτερον, οι πάροχοι υπηρεσιών και οι προγραμματιστές εφαρμογών που έχουν ως στόχο να προσφέρουν σύσταση τοποθεσίας ή οποιοδήποτε άλλο παρόμοιο σύστημα πρέπει να λάβουν υπόψη ότι οι διαφορετικές πλευρές της συμπεριφοράς των χρηστών δυναμικά και ετερογενώς επηρεάζουν τις κινήσεις τους. Τα μαρκοβιανά μοντέλα που προσομοιώθηκαν σε αυτή την εργασία συνδυάζουν τα χαρακτηριστικά αυτά σε μια στατική χρονική αναπαράσταση, δηλαδή, όλα τα χαρακτηριστικά ενσωματώνονται με την ίδια στάθμιση για όλο τον εξεταζόμενο χρόνο.

Αν σκεφτόμασταν να ενσωματώσουμε και το χρόνο σε αυτό το πλαίσιο, για παράδειγμα με την κατασκευή διαφορετικών μοντέλων για διαφορετικές ώρες της εβδομάδας, τότε θα απαιτούνται σημαντικά μεγαλύτερες ποσότητες δεδομένων για την εκμάθηση. Αυτό δεν είναι εφικτό όμως με τη χρήση του παρόντος πακέτου δεδομένων του Foursquare. Ωστόσο δεν πρέπει να αποκλεισθούν μελλοντικές προσπάθειες για τη δημιουργία πιο δυναμικών μοντέλων υπό το πρίσμα καλύτερων δεδομένων.

Τέλος, να σημειώσουμε ότι στην παρούσα ανάλυση εκπαιδεύσαμε τα μοντέλα μηχανικής μάθησης με ένα σύνολο δεδομένων που λάβαμε αφού το είχαμε κατεβάσει από το διαδίκτυο. Η εκπαίδευση τους με δυναμικό τρόπο, με λήψη δεδομένων σε πραγματικό χρόνο (σε μια αρκετά πιο μεγάλη κλίμακα φυσικά) και στη συνέχεια η πρόταση θέσεων άμεσα για κάθε ένα χρήστη λαμβάνοντας υπόψιν όλους τους του διάφορους ελέγχους επιλύει το πρόβλημα που έχουμε ορίσει εξ αρχής. Αυτό όμως, αντιπροσωπεύει ένα σύστημα, που εγγυάται γρήγορες υπολογιστικές απαντήσεις αφού επεξεργαστεί σαν είσοδο μόνο ένα σετ χαρακτηριστικών που κωδικοποιούν πληροφορίες σχετικά με τα παρελθούσες κοινοποιήσεις παρουσίας του χρήστη και την τρέχουσα θέση του.

Σε αυτό το σενάριο πρόβλεψης απαιτείται γρήγορη και δυναμική υπολογιστική απάντηση όσον αφορά την κατάταξη της τοποθεσίας σε πραγματικό χρόνο. Οι προσεγγίσεις όπως το μοντέλο παραγοντοποίησης μήτρας και οι τυχαίοι περίπατοι δεν είναι ειδικά προσαρμοσμένες για σενάρια πραγματικού χρόνου και σε αυτό το πλαίσιο θα μπορούσαν να βλάψουν την ποιότητα των υπηρεσιών που παρέχονται στο χρήστη.

10. Επίλογος

10.1. Συνεισφορές

Η κύρια συνεισφορά αυτής της εργασίας στον επιστημονικό κλάδο έρχεται από την σύγκριση μεταξύ διαφορετικών μοντέλων Markov. Σε αυτή τη διατριβή αξιολογήθηκαν οι τεχνικές πρωτοτάξιο μοντέλο Markov, δευτεροτάξιο μοντέλο Markov και κρυφό μοντέλο Markov και που εφαρμόστηκαν σε ένα πραγματικό σύνολο δεδομένων προεχόμενο από τη δημοφιλή πλατφόρμα Foursquare για την πρόβλεψη της επόμενης θέσης (κοινοποίησης παρουσίας) ενός χρήστη. Η σύγκριση επιδόσεων των μοντέλων δείχνει ποιο μοντέλο αποδίδει καλύτερα από τα υπόλοιπα μοντέλα στην ικανότητα πρόβλεψης με την κατάλληλη παραμετροποίηση.
Στη διεθνή βιβλιογραφία είναι ένα θέμα το οποίο έχει προσεγγιστεί ξανά αρκετές φορές και υπάρχουν δημοσιευμένα ερευνητικά έργα που εφαρμόζουν αυτά τα μοντέλα ξεχωριστά για να προβλέπουν τη θέση του χρήστη.

10.2. Μελλοντική εργασία

Συμπερασματικά μπορούμε να πούμε ότι τα αποτελέσματα της παρούσας μελέτης που αφορά την πρόβλεψη της ανθρώπινης κίνησης εντός πόλεων χρησιμοποιώντας το ιστορικό των χρηστών που περιλαμβάνει παρελθοντικές κοινοποιήσεις παρουσίας από κοινωνικά δίκτυα, είναι αρκετά ενδιαφέροντα. Σε αυτό το εδάφιο θα μιλήσουμε για νέες ιδέες για μελλοντική εργασία αλλά και ενδεχόμενες επεκτάσεις της παρούσας μελέτης ώστε να γίνει πιο ισχυρή και πιο αξιόπιστη στο σκοπό για τον οποίο σχεδιάστηκε.
Στην παρούσα εργασία, ασχοληθήκαμε και πειραματιστήκαμε κατά βάση με τα τυπικά μοντέλα Markov, το πρωτοτάξιο, το δευτεροτάξιο και το κρυφό μοντέλο. Μια επέκταση της παρούσας μελέτης θα μπορούσε να περιλαμβάνει και να έχει ως βάση πιο εξελιγμένα μοντέλα, όπως το άπειρο κρυφό Μαρκοβιανό μοντέλο που παρουσιάστηκε στην μελέτη των Beal M. J., Ghahramani. Z., "The Infinite Hidden Markov Model" όπου οι ερευνητές έδειξαν ότι τα κρυφά μοντέλα Markov μπορούν να επεκταθούν ώστε να έχουν ένα μετρήσιμα άπειρο αριθμό κρυφών καταστάσεων, χρησιμοποιώντας τη θεωρία των διεργασιών του Dirichlet και ενσωματώνοντας θεωρητικά άπειρες μεταβατικές παραμέτρους, αφήνοντας μόνο τρεις υπερπαραμέτρους για εκμάθηση από τα δεδομένα. Ένα άλλο Μαρκοβιανό μοντέλο που παρουσιάστηκε πρόσφατα και θα μπορούσε να έχει εφαρμογή στη μελέτη μας είναι το υβριδικό μοντέλο Markov, το οποίο έχει χρησιμοποιηθεί μάλιστα στην πρόβλεψη ανθρώπινης κίνησης, στην εργασία των Qiao Y., Si Z. και Zhang Y., "A hybrid Markov-based model for human mobility prediction" όπου, υπολογίζεται η σειρά προβλέψεων της αλυσίδας Markov προσαρμοσμένης στο μήκος των μοτίβων ατομικής κινητικότητας, αντί για σταθερή, εξετάζεται η χρονική κατανομή των μοτίβων κινητικότητας κατά τον υπολογισμό της πιθανότητας μετάβασης στην επόμενη τοποθεσία και χρησιμοποιούνται τα αποτελέσματα πρόβλεψης χρηστών με παρόμοιες τροχιές εάν το πρόσφατο πλαίσιο δεν έχει εμφανιστεί προηγουμένως.

Ενδεχόμενη βελτίωση και ιδέα για μελλοντική εργασία που θα μπορούσε να γίνει στην παρούσα μελέτη είναι η a-posteriori κανονικοποίηση σε κατανομές a-posteriori πιθανολογικών μοντέλων (με περιορισμούς) λανθανουσών

μεταβλητών καθώς επίσης και η ανάπτυξη μοντέλων που κατασκευάζουν χωρικές κατανομές τοποθεσιών στις πόλεις και τις χρησιμοποιούν ως πηγή διακυμάνσεων για τις μετακινήσεις. Για παράδειγμα, δύο (ή και παραπάνω) διαδοχικές τοποθεσίες που βρίσκονται σχετικά κοντά θα πρέπει να παρουσιάζουν μεγαλύτερες πιθανότητες μετάβασης από την μια στην άλλη και το αντίστροφο.
Για μελλοντική έρευνα επίσης, θα μπορούσαμε εισάγουμε διακριτά μοντέλα για την ταξινόμηση του ιστορικού θέσης με τη χρήση σημασιολογικών κατηγοριών (SVM) και μοντέλα δομημένης πρόβλεψης βάσει αναζητήσεων (SEARN) και Ensemble για να προσεγγίσουμε μεθόδους ακολουθιακής ανάλυσης όπως έχει γίνει στο έργο των L. Liao, D. Fox και H. Kautz., "Learning and Inferring Transportation Routines". Πολλές προηγούμενες εργασίες έχουν επίσης προσεγγίσει το πρόβλημα της ανάλυσης και ταξινόμησης ακολουθιακών δεδομένων που συλλέχθηκαν από διάφορα μέρη. Μια αξιόλογη μελέτη έχει γίνει από τον Dietterich T. G., με μια αξιόλογη έρευνα στη δημοσίευση του "Machine Learning for Sequential Data: A Review".
Επιπλέον, η ανάπτυξη μοντέλων που λαμβάνουν υπόψιν τον τρόπο με τον οποίο διαφορετικοί παράγοντες στο πέρασμα του χρόνου επηρεάζουν την ανθρώπινη κίνηση αναμένεται να παράγουν ακριβέστερες προβλέψεις για τον εντοπισμό των προθέσεων μετακίνησης. Εκτός του χρόνου όμως, πρέπει να ληφθεί υπόψιν ότι δεν παρουσιάζουν όλοι οι άνθρωποι ίδιες δυνάμεις έλξης ή απώθησης από συγκεκριμένες τοποθεσίες. Συνεπώς σαν μελλοντική μελέτη θα έπρεπε να ενταχθούν στα μοντέλα μέθοδοι εξατομίκευσης που εκτιμούν πως διάφορες ομάδες χρηστών αντιδρούν και επηρεάζονται από διάφορες τοποθεσίες με τις οποίες συνδέονται.
Μια από τις κοινωνικού είδους παρατηρήσεις που κάναμε είναι ότι οι χρήστες των κοινωνικών δικτύων φαίνεται να προτιμούν να κάνουν κοινοποίηση παρουσίας σε νέους χώρους κάτι που δείχνει ότι δίνουν σημασία στην αστική εξερεύνηση και την ανακάλυψη νέων δραστηριοτήτων στην πόλη. Ως εκ τούτου, τα αλγοριθμικά μοντέλα που αξιοποιούν πολυδιάστατα σήματα πληροφοριών προερχόμενα από κοινοποιήσεις χρηστών αναμένεται να παραμείνουν στην πρώτη γραμμή της ακαδημαϊκής έρευνας προκειμένου σε εφαρμογές όπως συστάσεις δραστηριοτήτων και τοπικές αναζητήσεις. Πρόκειται για μια νέα πηγή ανταγωνισμού ανάμεσα στους τεχνολογικούς κολοσσούς όπως η Microsoft, η Google και το Facebook, αλλά και μεταξύ πολλών νεοσύστατων επιχειρήσεων, οι οποίες σήμερα επενδύουν πολύ σε τομείς που σχετίζονται με το γεωγραφικό εμπόριο.

10.3. Συζήτηση

Στην παρούσα μελέτη που είχε στόχο τη σύγκριση μεταξύ διαφορετικών μοντέλων Markov για την μοντελοποίηση και πρόβλεψη της ανθρώπινης κίνησης, επιλέξαμε την προσέγγιση πειραματισμού και την εκτέλεση προσομοιώσεων. Ενώ πειραματικά πήραμε σχετικά καλά αποτελέσματα για την απόδοση του πρωτοτάξιου, του δευτεροτάξιου και του κρυφού μοντέλου Markov δεδομένου του συνόλου δεδομένων που είχαμε και της επεξεργασίας που κάναμε πάνω σε αυτό, θα ήταν καλό να έχουμε κάνει και μια θεωρητική μελέτη για άλλες διατάξεις Μαρκοβιανών μοντέλων καθώς επίσης και για το ιεραρχικό μοντέλο Markov. Τα ιεραρχικά μοντέλα Markov έχουν χρησιμοποιηθεί

για κατηγοριοποίηση ανθρώπινης συμπεριφοράς στο παρελθόν από τον Theocharous. G., στο έργο του “Hierarchical learning and planning in partially observable markov decision processes” όπου παρουσίασε το μοντέλο ιεραρχικά μερικώς παρατηρήσιμων διαδικασιών απόφασης Markov για την κλιμάκωση της εκμάθησης και του προγραμματισμού σε μεγάλη κλίμακα μερικώς παρατηρήσιμων περιβαλλόντων. Ωστόσο, εμείς, ασχοληθήκαμε με τις πιο γνωστές και ευρέως χρησιμοποιούμενες τακτικές διαμόρφωσης του μοντέλου Markov και αφήσαμε τις άλλες επιλογές για μελλοντικές εργασίες.
Παρόλο που το σύνολο δεδομένων που λάβαμε από το Foursquare είχε πάρα πολύ μεγάλο αριθμό από κοινοποιήσεις παρουσίας, μετά τις διαδικασίες επεξεργασίας και απόκλισης αρκετών από τις κοινοποιήσεις φαίνεται πως αν είχαμε αρκετά μεγαλύτερο αριθμό κοινοποιήσεων τα πειραματικά μας αποτελέσματα θα ήταν αρκετά καλύτερα.
Το ζητούμενο μας εξ αρχής ήταν να αξιολογηθεί η απόδοση των τριών διαμορφώσεων του μοντέλου Markov ως μοντέλο πρόβλεψης τοποθεσιών σε κοινωνικά δίκτυα. Η απόδοση των διαμορφώσεων του μοντέλου Markov ποικίλει λόγω διαφορών συνόλων δεδομένων αλλά και της ίδιας της διαμόρφωσης.
Από τα συμπεράσματα μας προκύπτει ότι το ROC και το AUC αποδίδουν πάρα πολύ καλά και παρουσιάζουν εξαιρετικά αποτελέσματα αναφορικά με την ακρίβεια. Το κρυφό μοντέλο Markov υπερτερεί του του πρωτοτάξιου και δευτεροτάξιου μοντέλου Markov. Το δευτεροτάξιο μοντέλο φαίνεται να δίνει τα χειρότερα αποτελέσματα στην απόδοση του όσον αφορά τη ικανότητα πρόβλεψης. Τα αποτελέσματα της μελέτης είναι αρκετά απλά, πράγμα που σημαίνει ότι η υπόθεση που σχηματίστηκε στην αρχή της έρευνας αποδείχθηκε σωστή.
Στην εργασία αξιολογήθηκαν οι απειλές εγκυρότητας. Πρόκειται για μια διαδικαστική εργασία που είναι σημαντικό να γίνει εκ των προτέρων για να διασφαλίσουμε ότι οι απειλές θα ελαχιστοποιηθούν. Είναι απολύτως αδύνατο να εξαλειφθούν πλήρως απειλές εγκυρότητας. Υπάρχουν και θα υπάρχουν πάντα όπως αναφέρει και ο Wohlin στη δημοσίευση του “An experimental comparison of Usage-Based and Checklist-Based Reading”. Δεν είναι εφικτό να εξαλείψουμε πλήρως τις απειλές εγκυρότητας μπορούμε όμως να τις ελαχιστοποιήσουμε. Έτσι αφού εντοπίσαμε όλες τις απειλές, τις θέσαμε υπό έλεγχο καταφέραμε να τις μετριάσουμε στο βαθμό που ήταν εφικτό. Είναι σημαντικό να αναγνωρίζονται οι απειλές εγκυρότητας πριν από τις πειραματικές διαδικασίες. Στην εργασία μας, η πιθανότητα εμφάνισης αυτών των απειλών μετά τις διαδικασίες για την ελαχιστοποίηση τους είναι τόσο μικρή, που δεν επηρεάζει τα πειραματικά αποτελέσματα. Σύμφωνα με την Angie Schottmuller (https://conversionxl.com/blog/ab-test-validity-threats/), "Η ελαχιστοποίηση των δεδομένων «ρύπων» για τη βελτιστοποίηση της ακεραιότητας είναι το δύσκολο κομμάτι. Η κατανόηση και η ανασκόπηση ενός συνόλου τεχνικών και περιβαλλοντικών παραγόντων/μεταβλητών που θα μπορούσαν να καταστρέψουν την εγκυρότητα της δοκιμής πρέπει να γίνει σε πρώτη φάση". Τέλος όπως προ είπαμε, οι στόχοι της μελέτης επιτεύχθηκαν σε ικανοποιητικό βαθμό, κάτι που σημαίνει ότι η μελέτη απαντά στην ερώτηση για την οποία ξεκίνησε, διατηρώντας παράλληλα την εγκυρότητα σε ένα καλό επίπεδο.

Αναφορές

Altman, R. M. (2007). *Mixed Hidden Markov Models: An Extension of the Hidden Markov Model to the Longitudinal Data Setting.* Journal of the American Statistical Association.

Backstrom, L., Sun, E., & Marlow, C. (2010). *Find me if you can: Improving geographical prediction with social and spatial proximity.* WWW '10.

Basharin, G. P., Langville, A. N., & Naumov, V. A. (2004). *The Life and Work of A. A. Markov.* Elsevier.

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2010). *The Infinite Hidden Markov Model.* Gatsby Computational Neuroscience Unit University College London.

Burns, R. B. (2013). *Introduction to Research Methods 4th Edition.* ISBN-13: 978-0761965923.

Calabrese, F., Lorenzo, D. G., & Ratti, C. (2010). *Human mobility prediction based on individual and collective geographical preferences.* Funchal, Portugal: IEEE.

Cheng, Z., Caverlee, J., & Lee, K. (2010). *You are where you Tweet: A content-based approach to geo-locating Twitter users.* CIKM '10.

Cheng, Z., Lee, K., & Caverlee, J. (2011). *Exploring Millions of Footprints in Location Sharing Services.* Fifth International AAAI Conference on Weblogs and Social Media.

Cho, E., Myers, S. A., & Leskovec, J. (2011). *Friendship and Mobility: User Movement In Location-Based Social Networks.* Stanford University, Stanford, CA, USA.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* ISBN-13: 978-0131367395.

Dietterich, T. G. (2002). *Machine Learning for Sequential Data: A Review.* Oregon State University.

Ducasse, S. G. (χ.χ.). *StatisticsHowTo.com.* Ανάκτηση από http://www.statisticshowto.com/

Eubank, S., Guclu, H., Kumar, S. A., Marathe, V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). *Modelling disease outbreaks in realistic urban social networks.* Nature.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2003). *How to Design and Evaluate Research in Education.* McGraw-Hil.

Gao, H., Tang, J., & Liu, H. (2012). *Exploring Social-Historical Ties on Location-Based Social Networks.* Sixth International AAAI Conference on Weblogs and Social Media.

Gao, H., Tang, J., & Liu, H. (2012). *Mobile Location Prediction of Spatio-Temporal Context.* Nokia mobile data challenge workshop.

Gao, H., Tang, J., & Liu, H. (2012). *Modelling geo-social correlations for new check-ins on location-based social networks.* CIKM '12.

Gao, H., Tang, J., Hu, X., & Liu, H. (2015). *Content-Aware Point of Interest Recommendation on Location-Based Social Networks.* Arizona State University.

Gomes, J. B., Phua, C., & Krishnaswamy, S. (2013). *Where will you go? Mobile Data Mining for Next Place Prediction.* Institute for Infocomm Research (I2R).
González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). *Understanding individual human mobility patterns.* Nature.
Hecht, B., Suh, L. H., & Chi, E. H. (2011). *Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles.* CHI '11.
Kerlinger, F. N. (1964). *Foundations of Behavioral Research.* ISBN-13: 978-0155078970.
Kim, M., Kotz, D., & Kim, S. (2006). *Extracting a mobility model from real user traces.* INFOCOM '06.
Lai, Y.-C., Yan, X.-Y., & Zi-You-Gao. (2017). New human-mobility prediction model offers scalability, requires less data. *Arizona State University*.
Liao, L., Fox, D., & Kautz, H. (2004). *Learning and Inferring Transportation Routines.* ScienceDirect.
Noulas, A. (2013). *Human Urban Mobility in Location-based Social Networks: Analysis, Models and Applications.* phD Thesis.
Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). *Mining User Mobility Features for Next Place Prediction in Location-based Services.* IEEE.
Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). *An Empirical Study of Geographic User Activity Patterns in Foursquare.* Fifth International AAAI Conference on Weblogs and Social Media.
Patton, M., & Cocharn, M. (2002). *A Guide to Using Qualitative Research Methodology.* Paris: Médecins Sans Frontières.
Pérez, L. C. (2003). *Hidden Markov Models and the Baum-Welch Algorithm.* IEEE.
Qiao, Y., Si, Z., Zhang, Y., Abdesslem, F. B., Zhang, X., & Yang, J. (2018). *A hybrid Markov-based model for human mobility prediction.* Neurocomputing.
Rabiner, L. (1989). *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* IEEE.
Samiul Hasan, S. V. (2016). *Understanding Social Influence in Activity Location Choice and Lifestyle Patterns Using Geolocation Data from Social Media.* Front. ICT.
Schottmuller, A. (χ.χ.). *How to Minimize A/B Test Validity Threats.* Ανάκτηση από ConversionXL: https://conversionxl.com/blog/ab-test-validity-threats/
Stouffer, S. A. (1940). *Intervening Opportunities: A Theory Relating Mobility and Distance.* American Sociological Association.
Strigos, T. (2015). *Alternative Positioning Techniques, Diploma Thesis.* Athens: National Technical University of Athens.
TANG, J., WANG, X., GAO, H., HU, X., & LIU, H. (2012). *Enriching short text representation in microblog for clustering front.* Front. Comput. Sci., 2012.
Theocharous, G. (2002). *Hierarchical learning and planning in partially observable markov decision processes.* Michigan State University East Lansing.
Uma Sekaran, R. B. (2016). *Research Methods For Business: A Skill Building Approach, 7th Edition.* ISBN: 978-1-119-26684-6.
Victoria Bellotti, B. B. (2008). *Activity-Based Serendipitous Recommendations with the Magitti Mobile Leisure Guide.* Florence, Italy: CHI 2008.

Wang, Z., Zhang, D., Yang, D., Yu, Z., & Zhou, X. (2012). Detecting Overlapping Communities in Location-Based Social Networks. *International Conference on Social Informatics*, 110-123.

Wilson, A. G. (1966). Gravity models in the physical and social sciences, I: some comparisons. *Ministry of Transport, London*.

Wohlin, C., Runeson, P., & Thelin, T. (2003). *An experimental comparison of Usage-Based and Checklist-Based Reading.* IEEE.

Yang, D., & Zhang, D. (2014). *Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs.* IEEE.

Yasheen, S. (2016). *Evaluation of Markov Models in Location Based Social Networks in Terms of Prediction Accuracy.* University of Skovde.

Ye, J., Zhu, Z., & Cheng, H. (2013). *What's Your Next Move: User Activity Prediction in Location-based Social Networks.* SIAM International Conference on Data Mining.

Zhang, J.-D., & Chow, C.-Y. (2015). *GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations.* SIGIR '15.

Zheng, W., Zheng, Y., Xie, X., & Yang, Q. (2010). *Collaborative location and activity recommendations with GPS history data.* WWW '10.

Zheng, Y., Zhang, L., Xie, X., & Ma, W. (2009). *Mining interesting locations and travel sequences from gps trajectories.* WWW '09.