Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Συστήματα Προσωποποιημένων Συστάσεων για Προβλήματα Προσανατολισμού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευστράτιος Τσιρτσής

Επιβλέπων: Δημήτριος Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Συστήματα Προσωποποιημένων Συστάσεων για Προβλήματα Προσανατολισμού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευστράτιος Τσιρτσής

Επιβλέπων: Δημήτριος Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Μαρτίου 2019.

.....	.....	.....
Δημήτριος Φωτάκης	Αριστείδης Παγουρτζής	Νικόλαος Παπασπύρου
Αν. Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2019

.....

Ευστράτιος Τσιρτσής

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

Περίληψη

Ένας τομέας με ενδιαφέρουσες αλγοριθμικές προκλήσεις είναι αυτός του σχεδιασμού τουριστικών διαδρομών. Η ύπαρξη πολλαπλών αξιοθέατων σε μια πόλη σε συνδυασμό με τον περιορισμένο χρόνο που διαθέτει ένας τουρίστας δημιουργεί την ανάγκη για υπολογισμό μιας διαδρομής που να ικανοποιεί τις προσωπικές του προτιμήσεις ενώ την ίδια στιγμή τηρεί κάποιο αυστηρό χρονικό περιορισμό. Το πρόβλημα αυτό, γνωστό ως Πρόβλημα του Προσανατολισμού, αντιμετωπίζεται συνήθως με ευριστικές μεθόδους και προσεγγιστικούς αλγορίθμους.

Ενώ η υπάρχουσα βιβλιογραφία εστιάζεται στην σχεδίαση τέτοιων αλγορίθμων θεωρώντας το προφίλ προτιμήσεων του εκάστοτε τουρίστα γνωστό με βάση την προηγούμενη συμπεριφορά του, κάτι τέτοιο δεν ανταποκρίνεται πλήρως στην πραγματικότητα. Λόγω των διαφορών των διαθέσιμων αξιοθέατων στις διάφορες πόλεις του κόσμου, η προηγούμενη συμπεριφορά του τουρίστα δεν αντικατοπτρίζει πλήρως τις μελλοντικές του προτιμήσεις. Με άλλα λόγια, γνωρίζοντας ότι ένας τουρίστας επισκέφθηκε μουσεία στο Παρίσι, δεν σημαίνει υποχρεωτικά ότι τον ενδιαφέρει να επισκεφθεί μουσεία και στη Νέα Υόρκη. Στην παρούσα διπλωματική εργασία γίνεται μια πρώτη προσπάθεια αντιμετώπισης του συγκεκριμένου προβλήματος σχεδιάζοντας αλγόριθμους εμπνευσμένους από κλασικές μεθόδους προσωποποιημένων συστάσεων. Με πειραματική αξιολόγηση σε πραγματικά δεδομένα τουριστών, αποδεικνύεται ότι οι αλγόριθμοι αυτοί είναι αποτελεσματικότεροι από κλασικές μεθόδους σύστασης των διασημότερων αξιοθέατων σε κάθε πόλη. Από πλευρά διαχείρισης πληροφορίας, παρατηρείται ότι η χρήση αλγορίθμων συσταδοποίησης ως μέσο συμπίεσης των προφίλ προτιμήσεων των τουριστών μπορεί να επιταχύνει αισθητά την εκτέλεση των προαναφερθέντων αλγορίθμων συστάσεων, προκαλώντας ελάχιστες απώλειες στην προσωπική ικανοποίηση των τουριστών.

Λέξεις κλειδιά

Πρόβλημα προσανατολισμού, Συστήματα συστάσεων, Ευριστικές τεχνικές, Συσταδοποίηση, Πρόβλημα σχεδιασμού τουριστικών διαδρομών

Abstract

A field with interesting algorithmic challenges is the one concerning tourist trip design. The excess of touristic attractions in a city combined with the limited available time of a tourist creates the need to compute a path that satisfies the tourist's personal preferences while maintaining a given time budget. This problem, known as the Orienteering Problem, is usually faced with heuristics and approximation algorithms.

While existing literature focuses on the design of such algorithms considering the preference profile of each tourist to be known based on their previous behavior, that assumption is not entirely realistic. Because of the differences between the available attractions in each city around the world, the previous tourist behavior does not reflect their future preferences. In other words, knowing that a tourist visited museums in Paris, does not necessarily mean that they are interested in visiting museums in New York. In this thesis, a first attempt of facing that problem is presented, by designing algorithms inspired by classic recommendation techniques. By experimental evaluation on real tourists' data, such algorithms are proved to outperform classic approaches of recommending attractions based on popularity. From a data management perspective, it is observed that the use of clustering algorithms as a means of compression of the tourists' preference profiles can extremely speed up execution of the aforementioned recommendation algorithms, inflicting minimal losses in individual tourist satisfaction.

Keywords

Orienteering problem, Recommender systems, Heuristics, Clustering, Tourist trip design problem

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας, Δημήτρη Φωτάκη, όχι μόνο για την καθοδήγησή του και τις ιδέες του στο πλαίσιο αυτής αλλά επειδή ήταν πάντα πρόθυμος να με βοηθήσει στις μελλοντικές μου επιλογές και με ενθάρρυνε να μαθαίνω συνεχώς νέα πράγματα καθώς έψαχνα να βρω την προσωπική μου ισορροπία ανάμεσα στη θεωρητική και την πρακτική πλευρά των αλγορίθμων. Ακόμα, μέσω του ενθουσιώδους τρόπου διδασκαλίας του με έκανε να αγαπήσω το συγκεκριμένο αντικείμενο και να ασχοληθώ με αυτό στη μετέπειτα ακαδημαϊκή και επαγγελματική μου πορεία.

Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές Νίκο Παπασπύρου και Άρη Παγουρτζή, πρωτίστως για τις γνώσεις που μου προσέφεραν μέσω της διδασκαλίας τους στον τομέα της πληροφορικής και φυσικά για τη συμμετοχή τους στην επιτροπή εξέτασης αυτής της διπλωματικής εργασίας.

Ένα πολύ μεγάλο ευχαριστώ αξίζει στους γονείς μου, Γιώργο και Λένα, που η συμβολή τους στην ολοκλήρωση των σπουδών μου ήταν καθοριστική. Τους ευχαριστώ που ήταν πάντα δίπλα μου και με στήριξαν σε όλες τις επιλογές μου μέχρι το τέλος. Δεν θα είχα φτάσει μέχρι εδώ χωρίς εκείνους.

Τέλος, θέλω να ευχαριστήσω όλους τους φίλους και τις φίλες που γνώρισα στα χρόνια του πολυτεχνείου, αυτούς που συνάντησα τις πρώτες μέρες και είμαστε μαζί μέχρι σήμερα αλλά κι εκείνους με τους οποίους μοιραστήκαμε λιγότερες αλλά πολύ όμορφες στιγμές που θα θυμάμαι για πάντα.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

1 Εισαγωγή

1.1 Αντικείμενο

Η μεγάλη τεχνολογική εξέλιξη των τελευταίων δεκαετιών και η ανάπτυξη του διαδικτύου έχει οδηγήσει στη δημιουργία πλειάδας ηλεκτρονικών εφαρμογών που υποβοηθούν τους ανθρώπους στις καθημερινές τους δραστηριότητες. Οι εφαρμογές αυτές πολλαπλασιάζονται μέρα με τη μέρα και πλέον έχουν επεκταθεί σε μεγάλο αριθμό τομέων και αγορών.

Ένας τέτοιος τομέας με ισχυρή παρουσία σε πολλές χώρες του κόσμου είναι και αυτός του τουρισμού. Ένα από τα βασικότερα προβλήματα που αντιμετωπίζουν οι τουρίστες είναι το πώς θα επιλέξουν τις δραστηριότητες και τα μέρη που θα επισκεφθούν κατά τη διάρκεια των διακοπών τους. Συνήθως, οι επιλογές αυτές γίνονται λαμβάνοντας υπόψη πολλαπλά κριτήρια όπως η προσωπική προτίμηση, η φήμη των αξιοθέατων, ο διαθέσιμος χρόνος κ.λ.π. Τέτοιες πληροφορίες γίνονται συχνά γνωστές μέσω ενημερωτικών τουριστικών βιβλίων ενώ πολλές φορές οι άνθρωποι εμπιστεύονται για αυτή τη δουλειά εξειδικευμένα ταξιδιωτικά πρακτορεία.

Αυτές οι κλασικές μέθοδοι επιλογής τουριστικών δραστηριοτήτων χρησιμοποιούνται εδώ και πάρα πολλά χρόνια, παρ' όλα αυτά έχουν σημαντικά μειονεκτήματα. Το κυριότερο από αυτά είναι ότι συνήθως δεν απευθύνονται ικανοποιητικά στα προσωπικά ενδιαφέροντα του κάθε τουρίστα. Είναι πρακτικά αδύνατο για έναν διαχειριστή ταξιδιωτικού γραφείου ή για ένα συγγραφέα ταξιδιωτικών οδηγών να λάβει υπόψη του όλες τις δυνατές προτιμήσεις των ανθρώπων στους οποίους απευθύνεται και να τις συνδυάσει κατάλληλα ώστε να προσφέρει την καλύτερη δυνατή υπηρεσία στον κάθε ένα ξεχωριστά. Έτσι, τα προγράμματα που προτείνονται στους τουρίστες είναι συνήθως πολύ γενικά, αγνοώντας πλήρως τις ατομικές προτιμήσεις.

Όμως, τα τελευταία χρόνια, οι προσεγγίσεις επίλυσης του συγκεκριμένου προβλήματος έχουν στραφεί κυρίως στην περιοχή της επιστήμης υπολογιστών. Η εμφάνιση πολλών ιστοσελίδων με χωρικές πληροφορίες για σημεία των πόλεων (π.χ. Tripadvisor, Google Maps) έχει οδηγήσει στην παραγωγή τεράστιου όγκου πληροφορίας για τα αντικείμενα ενδιαφέροντος σε κάθε πόλη και η ραγδαία αύξηση των χρηστών στα ηλεκτρονικά μέσα κοινωνικής δικτύωσης (π.χ. Twitter, Flickr) έχει δώσει τη δυνατότητα αναλυτικής μελέτης της συμπεριφοράς τους στο πλαίσιο τουριστικών επισκέψεων.

Έτσι, έχουν αναπτυχθεί προηγμένα υπολογιστικά εργαλεία τα οποία αναλαμβάνουν την αυτοματοποιημένη δημιουργία ταξιδιωτικών προγραμμάτων για τουρίστες κάνοντας χρήση πληροφοριών από ιστοσελίδες όπως αυτές που αναφέρθηκαν προηγουμένως τα οποία συνήθως προσφέρουν αρκετά μεγαλύτερη ικανοποίηση στους χρήστες τους από τις κλασικές μεθόδους. Οι σχεδιαστές αυτών των εργαλείων όμως καλούνται να αντιμετωπίσουν δύσκολα υπολογιστικά προβλήματα και η πλήρης υιοθέτηση τέτοιων συστημάτων είναι ακόμη σε εξέλιξη όσο η μελέτη των τε-

χνικών δυσκολιών τους προχωράει. Η εργασία αυτή αποτελεί ένα μικρό βήμα στην κατανόηση κάποιων προβλημάτων που θέτουν τέτοια τεχνολογικά εργαλεία ώστε να συμβάλλει στη βελτίωση της αποτελεσματικότητάς τους στο μέλλον.

1.2 Κίνητρο

Τα ηλεκτρονικά συστήματα που προτείνουν προϊόντα ή υπηρεσίες στους χρήστες τους βρίσκονται στο επίκεντρο του ενδιαφέροντος των επιστημόνων πληροφορικής καθώς απαιτούν προηγμένες τεχνικές αντιμετώπισης ενώ είναι απολύτως χρήσιμα σε πολλές ηλεκτρονικές πλατφόρμες όπως οι ιστοσελίδες ηλεκτρονικού εμπορίου. Στο πλαίσιο σύστασης ταξιδιωτικών πλάνων σε τουρίστες, τα ηλεκτρονικά συστήματα που αναπτύσσονται μπορούν να εκτιμήσουν τα ενδιαφέροντα του κάθε χρήστη και να του προτείνουν μια σειρά από αξιοθέατα που ταιριάζουν απολύτως στα ενδιαφέροντά του. Ωστόσο, η σύσταση ταξιδιωτικών πλάνων σε τουρίστες παρουσιάζει έναν επιπλέον περιορισμό ο οποίος παίζει καθοριστικό ρόλο στο σχεδιασμό και στην πολυπλοκότητα τέτοιων συστημάτων.

Συνήθως, ο χρόνος που έχουν διαθέσιμο οι τουρίστες για την επίσκεψη αξιοθέατων σε μια πόλη είναι πολύ μικρός μπροστά στο χρόνο που απαιτείται για την επίσκεψη όλων των αξιοθέατων που ταιριάζουν στις προτιμήσεις τους. Για παράδειγμα, οι ταξιδιώτες ενός κρουαζιερόπλοιου πραγματοποιούν πολλαπλές στάσεις σε διαφορετικές πόλεις και έχουν στη διάθεσή τους μόλις λίγες ώρες ώστε να επισκεφθούν τα μέρη που τους ενδιαφέρουν σε κάθε στάση. Έτσι, συστήματα που αναλαμβάνουν να τους προτείνουν ένα σύντομο ταξιδιωτικό πλάνο πρέπει να πραγματοποιήσουν ένα συμβιβασμό ανάμεσα στις ατομικές προτιμήσεις και στους χρόνους μετακίνησης από σημείο σε σημείο δεδομένου του χρονικού περιορισμού. Αυτό οδηγεί σε ενδιαφέροντα και πολύπλοκα υπολογιστικά προβλήματα που απαιτούν προχωρημένες τεχνικές για την επίλυσή τους.

Επίσης, μια άλλη ενδιαφέρουσα παράμετρος τέτοιων συστημάτων είναι ο τρόπος που εκμεταλλεύονται πληροφορία από ηλεκτρονικά μέσα κοινωνικής δικτύωσης για να εξάγουν τα προφίλ προτιμήσεων των τουριστών. Συνήθως, το ζητούμενο στο σχεδιασμό τέτοιων υπηρεσιών είναι η μικρότερη δυνατή συμμετοχή των χρηστών. Καταστάσεις όπου ο χρήστης είναι απασχολημένος συνεχώς με την εφαρμογή ώστε να τη βοηθήσει να κατανοήσει τα ενδιαφέροντά του είναι χρονοβόρες και μη αποδοτικές. Συνεπώς, η αυτοματοποίηση της εξόρυξης γνώσης από πλατφόρμες κοινωνικής δικτύωσης είναι απαραίτητη για το σχεδιασμό ενός "έξυπνου" συστήματος δημιουργίας τουριστικών πλάνων.

Μάλιστα, όσο μεγαλύτερος είναι ο αριθμός των χρηστών τόσο η πολυπλοκότητα διαχείρισης της πληροφορίας τους γιγαντώνεται. Σε περιπτώσεις όπου το ηλεκτρονικό σύστημα συστάσεων χρησιμοποιεί στοιχεία της παρελθοντικής συμπεριφοράς των χρηστών για να αποφασίσει τι κρίνεται ως ενδιαφέρον και τι όχι για τον καθένα, οι απαραίτητοι υπολογισμοί μπορεί να φτάσουν σε τέτοιο μέγεθος που να

κάνουν το σύστημα "μη βιώσιμο" σε πραγματικές εφαρμογές. Επομένως, απαιτείται η χρήση μεθόδων που θα πραγματοποιήσουν ικανοποιητικές προτάσεις στους χρήστες χρησιμοποιώντας ταυτόχρονα πληροφορίες επιλεκτικά.

1.3 Συνεισφορά

Βασικό αντικείμενο της παρούσας εργασίας είναι η σύσταση διαδρομών σε πόλεις, οι οποίες ανταποκρίνονται στις εκάστοτε προτιμήσεις των χρηστών και υποβάλλονται σε κάποιο χρονικό περιορισμό. Συγκεκριμένα, το σημείο διαφοροποίησης από την προϋπάρχουσα βιβλιογραφία είναι στον τρόπο με τον οποίο το σύστημα καθορίζει το προφίλ προτιμήσεων του κάθε χρήστη. Ενώ οι περισσότερες ερευνητικές εργασίες επικεντρώνονται στις μεθόδους υπολογισμού των διαδρομών παίρνοντας ως δεδομένη τη γνώση για τις ατομικές προτιμήσεις, η αρχική διαδικασία εκτίμησης των προτιμήσεων δεν είναι καθόλου προφανής.

Οι μέθοδοι που εξετάζονται σε αυτή την εργασία για τον υπολογισμό αυτών των προτιμήσεων βασίζονται στο συνδυασμό της πληροφορίας για τις προηγούμενες διαδρομές που έχει ακολουθήσει ο κάθε χρήστης κατά τις τουριστικές του επισκέψεις καθώς και των συλλογικών τάσεων των τουριστών όπως εμφανίζονται σε κάθε πόλη. Στη συνέχεια, τα προφίλ προτιμήσεων που δημιουργούνται χρησιμοποιούνται για το σχεδιασμό διαδρομών για τους τουρίστες και τα αποτελέσματα συγκρίνονται με πραγματικά δεδομένα για να επαληθευθεί η ακρίβεια των εκτιμήσεων. Παρατηρείται ότι ο συνδυασμός αυτών των πληροφοριών για τη δημιουργία διαδρομών δίνει καλύτερα αποτελέσματα από τις κλασικές μεθόδους που συνήθως βασίζονται στη "διασημότητα" των αξιοθέατων και όχι στις προσωπικές προτιμήσεις των χρηστών για αυτά.

Ακόμη, ένα τμήμα αυτής της εργασίας έχει να κάνει με τον τρόπο διαχείρισης της συλλογικής πληροφορίας των τουριστών. Συγκεκριμένα, μελετάται η περίπτωση όπου η διαδρομή δεν αφορά ένα μεμονωμένο χρήστη αλλά μία ομάδα χρηστών καθώς και πώς το μέγεθος του group επηρεάζει τη ατομική ικανοποίηση. Η σημαντικότερη παρατήρηση που προκύπτει είναι ότι η συλλογή πληροφοριών για τις προτιμήσεις ενός μικρού αριθμού από groups είναι αρκετή για να αναπαράγει μεγάλο μέγεθος της ικανοποίησης των μεμονωμένων χρηστών, κάνοντας την αποθήκευση και τη διαχείριση των πληροφοριών σημαντικά ευκολότερη και αποδοτικότερη.

1.4 Περίγραμμα κεφαλαίων

Η δομή της παρούσας εργασίας έχει ως εξής. Στο Κεφάλαιο 2 παρουσιάζονται τα βασικά στοιχεία της βιβλιογραφίας σε σχετικές ερευνητικές περιοχές όπως ο σχεδιασμός τουριστικών διαδρομών, οι τεχνικές συσταδοποίησης και τα συστήματα προσωποποιημένων συστάσεων. Στο Κεφάλαιο 3 γίνεται πειραματική μελέτη της συμπεριφοράς αλγορίθμων συσταδοποίησης εφαρμοζόμενων στο πρόβλημα της

σύστασης διαδρομών σε groups. Στο Κεφάλαιο 4 παρουσιάζονται οι αλγόριθμοι εκτίμησης των ατομικών προτιμήσεων και σύστασης διαδρομών που εκμεταλλεύονται προηγούμενη πληροφορία για τους χρήστες. Τέλος, στο Κεφάλαιο 5 γίνεται μία σύνοψη των αποτελεσμάτων και δίνονται κάποιες ενδιαφέρουσες μελλοντικές ερευνητικές κατευθύνσεις.

2 Σχετικές ερευνητικές κατευθύνσεις

2.1 Σχεδιασμός τουριστικών διαδρομών

Το *Πρόβλημα του Σχεδιασμού Τουριστικών Διαδρομών (Tourist Trip Design Problem – TTDP)* [1] συγκεντρώνει μεγάλο ενδιαφέρον από ερευνητές στους τομείς των πληροφοριακών συστημάτων και της επιχειρησιακής έρευνας λόγω των τεράστιων δυνατοτήτων που μπορεί να προσφέρει η επίλυσή του στην ανάπτυξη του τουρισμού. Η μελέτη του έχει οδηγήσει στη σχεδίαση *Προσωποποιημένων Ηλεκτρονικών Τουριστικών οδηγών (Personalized Electronic Tourist guides – PETs)* [2][3][4] που εμφανίζονται σε πλειάδα διαδικτυακών ή κινητών εφαρμογών [5][6].

Κάθε τέτοιο σύστημα έχει συνήθως τρεις κεντρικούς στόχους [7], τη διαδικασία της αναζήτησης αξιοθέατων που σχετίζονται με το προφίλ προτιμήσεων του εκάστοτε χρήστη, τη δημιουργία της διαδρομής που αυτός πρέπει να ακολουθήσει προκειμένου να τα επισκεφθεί μέσα στο χρονικό πλαίσιο που διαθέτει καθώς και την παροχή δυνατοτήτων τροποποίησής της. Ο συνδυασμός των δύο πρώτων λειτουργιών αποτελεί ένα δύσκολο αλγοριθμικό πρόβλημα, το *Πρόβλημα του Προσανατολισμού (Orienteering Problem – OP)*.

Σε αυτή την ενότητα παρουσιάζεται μία τυπική μοντελοποίηση του προβλήματος στα πρότυπα της διακριτής βελτιστοποίησης, εξετάζονται οι τρόποι αντιμετώπισής του και αναφέρονται διάφορες ενδιαφέρουσες παραλλαγές του που αντλούν έμπνευση από πραγματικές εφαρμογές.

2.1.1 Το Πρόβλημα του Προσανατολισμού και η ιστορία του

Το Πρόβλημα του Προσανατολισμού ανήκει σε μια μεγάλη οικογένεια προβλημάτων που έχουν τις ρίζες τους σε ένα από τα κλασικότερα προβλήματα της επιστήμης υπολογιστών, το *Πρόβλημα του Περιοδεύοντος Πωλητή (Travelling Salesman Problem – TSP)* [8]. Σύμφωνα με τον ορισμό του, δεδομένου ενός συνόλου κόμβων και ακμών με βάρη μεταξύ τους, απαιτείται η εύρεση ενός μονοπατιού το οποίο ξεκινάει και τερματίζει σε ένα συγκεκριμένο κόμβο έχοντας περάσει από όλους τους υπόλοιπους χρησιμοποιώντας ακμές που ελαχιστοποιούν το συνολικό βάρος του μονοπατιού.

Κατά τη δεκαετία του 1980, αντλώντας έμπνευση από πρακτικές εφαρμογές, οι επιστήμονες πληροφορικής άρχισαν να μελετούν πιο περίπλοκες παραλλαγές του TSP όπως το *Πρόβλημα του Περιοδεύοντος Πωλητή με Κέρδη (Travelling Salesman Problem with Profits – TSPP)* [9]. Σύμφωνα με αυτή τη γενίκευση του TSP, η επίσκεψη σε κάθε κόμβο συνοδεύεται από κάποιο κέρδος και στόχος του προβλήματος είναι η εύρεση κάποιου υποσυνόλου των κόμβων καθώς και του μονοπατιού μεταξύ τους, ικανοποιώντας δύο αντικρουόμενους στόχους, τη μεγιστοποίηση του συλλεγόμενου κέρδους και την ελαχιστοποίηση του μήκους του διανυόμενου μο-

νοπατιού.

Έτσι, εμφανίστηκε το Πρόβλημα του Προσανατολισμού [10] ως μια μονοκριτηριακή παραλλαγή του TSPP όπου η αναζήτηση αφορά ένα μονοπάτι που μεγιστοποιεί το κέρδος κρατώντας το συνολικό μήκος κάτω από μία προκαθορισμένη τιμή (*budget*). Η ονομασία του, όπως καθιερώθηκε στη συνέχεια, δόθηκε για πρώτη φορά το 1984 [11]. Στο πρόβλημα αυτό βασίζεται η πιο απλή περίπτωση του σχεδιασμού τουριστικών διαδρομών όπου κάθε κόμβος αντιστοιχεί σε κάποιο αξιοθέατο συνοδευόμενο από κάποιο κέρδος (ικανοποίηση ενός τουρίστα), το μήκος ενός μονοπατιού είναι ο συνολικός χρόνος που απαιτείται για την επίσκεψη όλων των αξιοθέατων που αυτό περιλαμβάνει, ενώ ο χρόνος αυτός διατηρείται κάτω από κάποιο όριο (διαθέσιμος χρόνος του τουρίστα).

Ακολουθεί η μοντελοποίηση της απλούστερης περίπτωσης σχεδίασης τουριστικών διαδρομών ως μια μορφή του Προβλήματος του Προσανατολισμού και δίνονται χρήσιμες συμβάσεις και συμβολισμοί για τη συνέχεια. Έστω ένας κατευθυνόμενος γράφος $G = (V, E)$ με βάρη, όπου V είναι ένα σύνολο από n κόμβους που αντιστοιχούν σε *σημεία ενδιαφέροντος (Points Of Interest – POIs)* μέσα σε μια πόλη και E ένα σύνολο m ακμών που αντιπροσωπεύει τους δρόμους που συνδέουν τα σημεία αυτά. Σε κάθε κόμβο αντιστοιχεί ένας χρόνος επίσκεψης $d_V(u)$ και σε κάθε ζευγάρι κόμβων μία απόσταση $d_E(u, v)$ ίση με το μήκος του συντομότερου μονοπατιού που συνδέει τους δύο αυτούς κόμβους στον γράφο G. Είναι εύκολο να δούμε ότι μας δίνεται η δυνατότητα να συμπεριλάβουμε τους χρόνους αναμονής των αξιοθέατων στα βάρη των ακμών. Ως βάρος ορίζουμε το $w(u, v) = d_V(u) + d_E(u, v)$. Επιπλέον, κάθε POI αντιστοιχεί σε μία *αξία* $p : V \mapsto \mathbb{R}_+$ που αντιπροσωπεύει την ικανοποίηση του τουρίστα (ή χρήστη) από την επίσκεψη στο συγκεκριμένο POI.

Ένα *μονοπάτι* είναι μία ακολουθία από αξιοθέατα. Μας ενδιαφέρουν μονοπάτια T που έχουν ως σημεία εκκίνησης και τερματισμού δύο (όχι αναγκαία διαφορετικά) αξιοθέατα s και t. Δεδομένου ενός μονοπατιού $T = (s, v_1, ..., v_l, t)$ με $v_i \neq v_j$ για $i \neq j$, ορίζουμε την αξία του ως το άθροισμα των αξιών των κόμβων του $\sum_{i=1}^{l} p(v_i)$. Παρόλο που το αρχικό και το τελικό σημείο του μονοπατιού μπορεί να ταυτίζονται, αξίζει να αναφερθεί ότι τα ενδιάμεσα σημεία πρέπει να είναι όλα διαφορετικά μεταξύ τους και μόνο αυτά συνεισφέρουν στην αξία του μονοπατιού. Τέλος, δεδομένου ενός μονοπατιού $T = (s, v_1, ..., v_l, t)$, συμβολίζουμε ως $len(T) = w(s, v_1) + w(v_1, v_2) + ... + w(v_l, t)$ το συνολικό του μήκος και ως $T \oplus v$ το μονοπάτι $T' = (s, v_1, ..., v_l, v, t)$.

Ο σκοπός του προβλήματος είναι, δεδομένου ενός χρονικού περιορισμού $B \in \mathbb{R}_+$ και σημείων εκκίνησης και τερματισμού s και t, να βρεθεί ένα μονοπάτι $T = (s, v_1, ..., v_l, t)$ που να ικανοποιεί τον περιορισμό, δηλαδή $len(T) \leq B$ και η αξία του μονοπατιού να είναι η μέγιστη δυνατή.

Όπως έχει αναφερθεί και προηγουμένως, το πρόβλημα αυτό αποτελεί μία εξελιγμένη παραλλαγή του προβλήματος του περιοδεύοντος πωλητή. Έτσι, γεννιούνται

ερωτήματα για την υπολογιστική πολυπλοκότητα που το χαρακτηρίζει. Αν και οι πραγματικές εφαρμογές σχεδιασμού τουριστικών διαδρομών απαιτούν πιο πολύπλοκες παραλλαγές, ακόμα και αυτή η φαινομενικά απλή περίπτωση αποδεικνύεται ότι είναι εξαιρετικά δύσκολη. Το Πρόβλημα του Προσανατολισμού ανήκει στην κλάση πολυπλοκότητας NP-Hard [12] δηλαδή η ακριβής επίλυσή του σε πολυωνυμικό χρόνο είναι μάλλον αδύνατη.

Συνεπώς, ακριβείς λύσεις στο πρόβλημα μπορούν να δοθούν μόνο για στιγμιότυπα με μικρό αριθμό κόμβων. Αλγόριθμοι που δίνουν ακριβείς λύσεις χρησιμοποιούν τεχνικές περιορισμού του χώρου των δυνατών λύσεων όπως branch and bound [13][14] ή branch and cut [15][16] ώστε να επιταχύνουν τη διαδικασία επίλυσης. Προκειμένου να επιλυθούν μεγαλύτερα στιγμιότυπα του προβλήματος έχουν σχεδιαστεί προσεγγιστικοί αλγόριθμοι [17][18][19][20][21] για μικρές παραλλαγές του προβλήματος που αφορούν το αν ο γράφος είναι κατευθυνόμενος ή μη και αν τα σημεία εκκίνησης και τερματισμού s,t είναι δεδομένα. Οι αλγόριθμοι αυτοί συνοδεύονται από εγγυήσεις για την ακρίβεια της λύσης που πετυχαίνουν αλλά είναι αρκετά περίπλοκοι ώστε να χρησιμοποιηθούν σε πρακτικές εφαρμογές.

Η ανάγκη επίλυσης όλο και μεγαλύτερων στιγμιοτύπων σε όλο και μικρότερο χρόνο οδήγησε με την πάροδο του χρόνου στην προτίμηση αλγορίθμων οι οποίοι αντιμετωπίζουν το Πρόβλημα του Προσανατολισμού κάνοντας χρήση ευριστικών τεχνικών. Διάφορες τεχνικές που χρησιμοποιήθηκαν περιλαμβάνουν τη μέθοδο Monte Carlo [11], υπολογισμό κέντρου βαρύτητας (center of gravity) [12][22], four-phase heuristics [23], χρήση νευρωνικών δικτύων [24], άπληστα κριτήρια [25] και tabu search [26]. Οι αλγόριθμοι αυτοί δεν προσφέρουν εγγυήσεις για την ποιότητα της λύσης που δίνουν αλλά φαίνεται να είναι αρκετά αποδοτικοί στην πράξη.

2.1.2 Παραλλαγές του προβλήματος

Προκειμένου να προσαρμοστεί το Πρόβλημα του Προσανατολισμού σε περιορισμούς που θέτουν πρακτικές εφαρμογές όπως η σχεδίαση τουριστικών διαδρομών, έχουν μοντελοποιηθεί πολλαπλές εκδοχές του και στη συνέχεια παρουσιάζονται συνοπτικά μερικές από αυτές.

Δύο επεκτάσεις του OP με επιπλέον περιορισμούς αποτελούν το *Πρόβλημα του Προσανατολισμού με Χρονικά Παράθυρα (Orienteering Problem with Time Windows – OPTW)* και το *Πρόβλημα του Χρονικά Εξαρτώμενου Προσανατολισμού (Time Dependent Orienteering Problem – TDOP)*. Σύμφωνα με το *OPTW* [27], ο τουρίστας μπορεί να επισκεφθεί κάποιο κόμβο του γράφου G μόνο σε προκαθορισμένα χρονικά διαστήματα που μπορεί να είναι διαφορετικά για κάθε κόμβο. Με αυτό τον τρόπο, μοντελοποιούνται περιπτώσεις όπου τα σημεία ενδιαφέροντος είναι ανοιχτά μόνο για συγκεκριμένες ώρες (π.χ. μουσεία, μαγαζιά). Το *TDOP* [28] αφορά γράφους όπου το κόστος κάθε ακμής δεν είναι σταθερό αλλά αποτελεί συνάρτηση του χρόνου. Έτσι, λαμβάνονται υπόψη περιπτώσεις όπου ο χρόνος

μετάβασης εξαρτάται από την κίνηση στο δρόμο, από δρομολόγια μέσων μαζικής μεταφοράς κ.λ.π.

Στη βιβλιογραφία εμφανίζονται και παραλλαγές του αρχικού προβλήματος που αντί για επιπλέον περιορισμούς ή χρονική μεταβλητότητα παρουσιάζουν μεγαλύτερη πολυπλοκότητα στη συνάρτηση που βελτιστοποιούν ή περιλαμβάνουν στοχαστικά φαινόμενα. Δύο τέτοια προβλήματα είναι το *Γενικευμένο Πρόβλημα Προσανατολισμού (Generalized Orienteering Problem – GOP)* και το *Στοχαστικό Πρόβλημα Προσανατολισμού (Stochastic Orienteering Problem – SOP)*. Σύμφωνα με το *GOP* [24], ο κάθε κόμβος του γράφου αντιστοιχεί σε ένα σύνολο από τιμές κέρδους (π.χ. φυσική ομορφιά, ιστορική αξία) και το μέγεθος προς μεγιστοποίηση είναι κάποιος συνδυασμός αυτών των τιμών. Το *SOP* [29] αφορά περιπτώσεις όπου ο κάθε κόμβος αντιστοιχεί σε τυχαίο χρόνο επίσκεψης. Κάτι τέτοιο είναι πολύ χρήσιμο καθώς ο χρόνος που αφιερώνει κάθε τουρίστας σε κάποιο αξιοθέατο μπορεί να είναι αρκετά διαφορετικός και να μην μπορεί να χαρακτηριστεί από μία ντετερμινιστική τιμή.

Επίσης, σχετικά πρόσφατα, έχουν παρουσιαστεί και άλλες πιο περίπλοκες παραλλαγές μαζί με τους αλγορίθμους επίλυσής τους για να προσεγγίσουν πραγματικά προβλήματα σχεδίασης διαδρομών. Για παράδειγμα, το πρόβλημα μπορεί να επεκταθεί στη σχεδίαση πολλαπλών διαδρομών [30] με το σκεπτικό ότι οι τουρίστες μένουν σε μία πόλη για πάνω από μία μέρα και η επιλογή αξιοθέατων από μέρα σε μέρα εξαρτάται από το τι θα προταθεί τις υπόλοιπες. Άλλες προσεγγίσεις λαμβάνουν υπόψη τη συμμετοχή του χρήστη στο σχεδιασμό με κάποια μορφή βαθμολογίας [31] ώστε να ανακαλύψουν τα πραγματικά του ενδιαφέροντα. Ακόμα, υπάρχουν παραλλαγές που λαμβάνουν υπόψη στοιχεία θεωρίας παιγνίων [32]. Σε αυτό το πλαίσιο, μεγάλος αριθμός επισκέψεων σε κάποιο σημείο ενδιαφέροντος δημιουργεί καθυστερήσεις που οι χρήστες πρέπει να λάβουν υπόψη τους. Έτσι, καλούνται να επιλέξουν τη διαδρομή τους ανταγωνιστικά, σκεπτόμενοι τις πιθανές επιλογές των υπολοίπων.

Οι παραλλαγές του αρχικού Προβλήματος του Προσανατολισμού και οι αλγοριθμικές τους προσεγγίσεις είναι πολυάριθμες και με μεγάλη ποικιλομορφία. Η αναλυτική περιγραφή τους υπερβαίνει το αντικείμενο της παρούσας εργασίας και για μια πιο πλήρη παρουσίασή τους ο αναγνώστης παραπέμπεται στη βιβλιογραφία [33].

2.1.3 Σύσταση διαδρομών σε ομάδες

Μία πολύ ενδιαφέρουσα επέκταση του απλού Προβλήματος Σχεδιασμού Τουριστικών Διαδρομών είναι η Σύσταση Διαδρομών σε Ομάδες [34]. Η διαφορά από το κλασικό Πρόβλημα του Προσανατολισμού είναι ότι πλέον ενδιαφερόμαστε για τη βελτιστοποίηση της ικανοποίησης μιας ομάδας (group) τουριστών αντί για ένα μεμονωμένο τουρίστα. Για την επίλυση αυτού του προβλήματος πρέπει να ληφθούν υπόψη πιθανές διαφορές στις προτιμήσεις των μελών της ομάδας και έτσι η ποσότητα προς βελτιστοποίηση δεν είναι αρχικά πλήρως φανερή. Στη συνέχεια δίνεται

ένας ολοκληρωμένος ορισμός του προβλήματος και μερικές επιλογές αντικειμενικής συνάρτησης για την περιγραφή του.

Θεωρούμε ένα group k τουριστών $\{P_1, ..., P_k\}$. Παρομοίως με το Πρόβλημα του Προσανατολισμού, δίνεται ένας γράφος $G = (V, E)$ με βάρη ακμών $w(\cdot, \cdot)$ όπως προηγουμένως αλλά κάθε PoI v_i έχει διαφορετική αξία $p_j(v_i)$ για κάθε τουρίστα P_j. Δηλαδή, κάθε κόμβος v_i αντιστοιχεί σε ένα διάνυσμα τιμών $\mathbf{p}(v_i) : V \mapsto \mathbb{R}_+^k$ του οποίου η j-οστή τιμή είναι η $p_j(v_i)$.

Και σε αυτή την περίπτωση, ενδιαφερόμαστε για μονοπάτια $T = (s, v_1, ..., v_l, t)$ με σημεία εκκίνησης και τερματισμού $s, t \in V$ και αναζητούμε αυτά που το μήκος τους δεν υπερβαίνει ένα δοσμένο χρονικό περιορισμό B. Κάθε μονοπάτι έχει κάποια αξία (ή ικανοποίηση) για κάθε τουρίστα P_j που ορίζουμε ως $p_j(T) = \sum_{i=1}^{l} p_j(v_i)$.

Πρόβλημα 1. ***(TourGroup)*** *Δεδομένου ενός κατευθυνόμενου γραφήματος με βάρη $G = (V, E, w)$, δύο κόμβων $s, t \in V$, μίας τιμής $B \in \mathbb{R}_+$, ενός ακεραίου k, διανυσμάτων $\boldsymbol{p}(v) \in \mathbb{R}_+^k \ \forall v \in V$ και μιας συνάρτησης $\Phi : \mathbb{R}^k \mapsto \mathbb{R}_+$, ζητείται ένα μονοπάτι $T = (s, v_1, ..., v_l, t)$ ώστε $len(T) \leq B$ και η $\Phi(p_1(T), p_2(T), ..., p_k(T))$ να μεγιστοποιείται.*

Ο ορισμός αυτός καλύπτει μία οικογένεια διαφορετικών προβλημάτων που διαφοροποιούνται αρκετά ανάλογα με την επιλογή της συνάρτησης Φ. Ανάλογα με αυτή, μεταβάλλεται το μέγεθος του συμβιβασμού μεταξύ της συνολικής ικανοποίησης του group και του σεβασμού των ατομικών προτιμήσεων.

Υποπρόβλημα 1.1. ***(TourGroupSum)*** *Σε αυτή την περίπτωση, μεγιστοποιείται η $\Phi(p_1(T), p_2(T), ..., p_k(T)) = \sum_{j=1}^{k} p_j(T)$.*

Είναι εύκολο να δούμε ότι αυτή η εκδοχή μεγιστοποιεί τη συνολική ικανοποίηση του group χωρίς να λαμβάνει υπόψη περιπτώσεις τουριστών που ενδεχομένως να διαφωνούν πλήρως με την επίσκεψη σε κάποιο PoI.

Υποπρόβλημα 1.2. ***(TourGroupMin)*** *Σε αυτή την περίπτωση, μεγιστοποιείται η $\Phi(p_1(T), p_2(T), ..., p_k(T)) = min_{j \in [k]}\{p_j(T)\}$.*

Η συγκεκριμένη συνάρτηση είναι η πιο "δίκαιη" καθώς φροντίζει ώστε να κάνει μετριοπαθείς επιλογές που δεν θα δυσαρεστήσουν σε μεγάλο βαθμό κανένα μέλος του group. Το μειονέκτημά της είναι ότι δίνει μεγάλη σημασία στα άτομα αγνοώντας πλήρως τη συνολική ικανοποίηση.

Υποπρόβλημα 1.3. ***(TourGroupFair)*** *Σε αυτή την περίπτωση, μεγιστοποιείται η $\Phi(p_1(T), p_2(T), ..., p_k(T)) = avg_{j \in [k]}(p_j(T)) - \alpha \cdot std_{j \in [k]}(p_j(T))$, όπου $\alpha \in \mathbb{R}_+$ μία σταθερή παράμετρος που εκφράζει τη σχετική σημασία της δικαιοσύνης.*

Τέλος, η συνάρτηση αυτή είναι μία μέση λύση μεταξύ των δύο προηγούμενων. Ο πρώτος όρος (avg) έχει ως σκοπό να μεγιστοποιήσει τη μέση αξία του μονοπατιού άρα και το άθροισμα των επιμέρους ικανοποιήσεων των χρηστών. Ο δεύτερος όρος (std) ουσιαστικά επιβάλει ένα φραγμό στη βελτιστοποίηση της μέσης ικανοποίησης προσπαθώντας ταυτόχρονα να διατηρήσει χαμηλά την τυπική απόκλιση. Έτσι, φροντίζει ώστε η ελάχιστη ικανοποίηση να μην απέχει πολύ από τη μέση, διατηρώντας ένα βαθμό δικαιοσύνης του οποίου η σημασία είναι ανάλογη της σταθεράς α.

Οι τρεις αυτές συναρτήσεις και οι επιρροή τους στην εύρεση ομαδικών τουριστικών διαδρομών έχουν μελετηθεί πλήρως. Στα πλαίσια αυτής της εργασίας, χρησιμοποιείται μόνο η πρώτη, η οποία είναι και η απλούστερη από τις τρεις. Είναι εύκολο να δούμε ότι το Πρόβλημα του Προσανατολισμού ανάγεται στο TourGroupSum αφού το πρώτο αποτελεί ειδική περίπτωση του δεύτερου στην περίπτωση που το group αποτελείται μόνο από ένα χρήστη. Όμως, αυτά τα δύο προβλήματα είναι ισοδύναμα σύμφωνα και με το παρακάτω θεώρημα.

Θεώρημα 1. *Το πρόβλημα TourGroupSum ανάγεται στο απλό Πρόβλημα του Προσανατολισμού.*

Απόδειξη. Γνωρίζοντας τα διανύσματα αξίας $\mathbf{p}(v_i)$ για όλους τους κόμβους $v_i \in V$, μπορούμε να κατασκευάσουμε ένα χρήστη P_m όπου η αξία των κόμβων για αυτό το χρήστη είναι $p_m(v_i) = \frac{1}{k}\sum_{j=1}^{k} p_j(v_i)\ \forall v_i \in V$.

Αν $\mathbf{T}$ είναι το σύνολο όλων των εφικτών μονοπατιών μεταξύ s και t, λύνοντας το απλό Πρόβλημα του Προσανατολισμού για το χρήστη P_m έχουμε ένα μονοπάτι T^* το οποίο είναι βέλτιστο και για το TourGroupSum αφού:

$$\begin{aligned}
T^* &= \arg\max_{T\in\mathbf{T}} p_m(T) \quad (T = (s, v_1, v_2, ..., v_l, t)) \\
&= \arg\max_{T\in\mathbf{T}} \sum_{i=1}^{l} p_m(v_i) \\
&= \arg\max_{T\in\mathbf{T}} \sum_{i=1}^{l} \frac{1}{k} \sum_{j=1}^{k} p_j(v_i) \\
&= \arg\max_{T\in\mathbf{T}} \sum_{j=1}^{k} \sum_{i=1}^{l} p_j(v_i) \\
&= \arg\max_{T\in\mathbf{T}} \sum_{j=1}^{k} p_j(T) \quad (\Phi : TourGroupSum) \\
&= \arg\max_{T\in\mathbf{T}} \Phi(p_1(T), p_2(T), ..., p_k(T))
\end{aligned}$$

□

Όπως θα εξηγηθεί και στο επόμενο κεφάλαιο, θεωρούμε ότι ο κάθε τουρίστας και το κάθε POI απεικονίζονται ως διανύσματα πραγματικών αριθμών σε ένα κοινό διανυσματικό χώρο d διαστάσεων. Συνεπώς, ως αξία ενός POI για ένα χρήστη ορίζουμε το εσωτερικό γινόμενο των δύο αυτών διανυσμάτων. Δηλαδή, έστω ένας χρήστης P_j με διάνυσμα $\mathbf{u}_j$ και ένα POI v_i με διάνυσμα $\mathbf{m}_i$. Τότε, $p_j(v_i) = \mathbf{u}_j \cdot \mathbf{m}_i$. Ως συνέπεια αυτής της παραδοχής και του προηγούμενου θεωρήματος προκύπτει το παρακάτω λήμμα.

Λήμμα 1. *Η εύρεση μονοπατιού για την ομάδα στο TourGroupSum είναι ισοδύναμη με την εύρεση μονοπατιού για το κεντροειδές της (μέση τιμή διανυσμάτων ομάδας) στο απλό Πρόβλημα του Προσανατολισμού.*

Απόδειξη. Σύμφωνα με το προηγούμενο θεώρημα, αρκεί να αποδείξουμε ότι αν P_m είναι ο χρήστης τον οποίο προσομοιώνει το κεντροειδές τότε ισχύει $p_m(v_i) = \frac{1}{k}\sum_{j=1}^{k} p_j(v_i)\ \forall v_i \in V$.

Αφού ο P_m είναι το κεντροειδές της ομάδας τότε θα ισχύει $\mathbf{u}_m = \frac{1}{k}\sum_{j=1}^{k} \mathbf{u}_j$. Έστω ένα οποιοδήποτε POI $v_i \in V$ που απεικονίζεται με το διάνυσμα $\mathbf{m}_i$. Τότε, η αξία του συγκεκριμένου POI για το χρήστη P_m θα είναι:

$$\begin{aligned} p_m(v_i) &= \mathbf{u}_m \cdot \mathbf{m}_i \\ &= (\frac{1}{k}\sum_{j=1}^{k} \mathbf{u}_j) \cdot \mathbf{m}_i \\ &= \frac{1}{k}\sum_{j=1}^{k} \mathbf{u}_j \cdot \mathbf{m}_i \\ &= \frac{1}{k}\sum_{j=1}^{k} p_j(v_i) \end{aligned}$$

□

Έχοντας ορίσει πλήρως το πρόβλημα της Σύστασης Διαδρομών σε Ομάδες, παρακάτω παρουσιάζεται η αλγοριθμική του αντιμετώπιση. Καθώς η πλήρης ανάλυση και σύγκριση των αλγορίθμων δεν αποτελεί αντικείμενο αυτής της εργασίας, αναφέρονται μόνο κάποιοι αλγόριθμοι βασισμένοι σε άπληστες ευριστικές τεχνικές και παρουσιάζεται αναλυτικά ο καλύτερος από αυτούς που χρησιμοποιείται και στη συνέχεια. Αξίζει να αναφερθεί ότι οι αλγόριθμοι αυτοί μπορούν να δουλέψουν και με τις τρεις αντικειμενικές συναρτήσεις.

Οι αλγόριθμοι είναι οι εξής:

- *BestValue* – Το μονοπάτι κατασκευάζεται αυξητικά ξεκινώντας αρχικά με το (s, t). Σε κάθε επανάληψη, έχοντας ένα μονοπάτι T, ελέγχονται όλοι οι κόμβοι $v \notin T$ και επιλέγεται το μονοπάτι $T' = T \oplus v$ που διατηρεί μήκος μικρότερο από B και ταυτόχρονα μεγιστοποιεί την αντικειμενική συνάρτηση. Ο αλγόριθμος τερματίζει μόλις δεν μπορεί να κατασκευαστεί μονοπάτι που να διατηρεί το χρονικό περιορισμό. Έχει το μειονέκτημα ότι επιλέγει κόμβους με υψηλή αξία αγνοώντας πλήρως το χρονικό κόστος που επιφέρει η επίσκεψη σε αυτούς με αποτέλεσμα να δημιουργούνται μονοπάτια με πολύ μικρό αριθμό κόμβων.

- *BestDistance* – Η διαδικασία είναι όμοια με προηγουμένως αλλά σε κάθε βήμα επιλέγεται το εφικτό μονοπάτι που ελαχιστοποιεί το συνολικό μήκος. Ο συγκεκριμένος αλγόριθμος τείνει να επιστρέφει μονοπάτια αποτελούμενα από μεγάλο αριθμό κόμβων αλλά δεν εξετάζεται καθόλου η αξία που προσφέρουν αυτοί οι κόμβοι. Το αποτέλεσμα είναι τα μονοπάτια αυτά να απέχουν αρκετά από τα βέλτιστα.

- *BestRatio* – Η διαδικασία είναι όμοια με προηγουμένως αλλά σε κάθε βήμα επιλέγεται το εφικτό μονοπάτι που μεγιστοποιεί το λόγο της αξίας προς το μήκος. Ο αλγόριθμος αυτός δίνει καλύτερα αποτελέσματα από τους δύο προηγούμενους αφού εκμεταλλεύεται τη φύση του προβλήματος που είναι όμοια με αυτή του Προβλήματος του Σακιδίου. Παρ' όλα αυτά, αν ο αλγόριθμος φτάσει σε τοπικό βέλτιστο συναντώντας το άνω φράγμα B, δεν δίνεται η δυνατότητα να ξεφύγει από αυτό.

- *BestRatio+* – Ο αλγόριθμος αυτός αποτελεί επέκταση του *BestRatio* με τη διαφορά ότι αν σε κάποιο βήμα το μονοπάτι δεν μπορεί να επεκταθεί, γίνεται αλλαγή κάποιου από τους ήδη τοποθετημένους κόμβους. Έτσι, γίνεται αποφυγή τοπικών βέλτιστων. Οι λύσεις που δίνει ο αλγόριθμος είναι κοντά στις βέλτιστες και γι αυτό χρησιμοποιείται και στη συνέχεια της εργασίας. Η αναλυτική περιγραφή του αλγορίθμου δίνεται στην επόμενη σελίδα.

Algorithm 1: BestRatio+

Input: POIs, Users, Source (s), Destination (t), Graph weights (w)
Output: path
path $\leftarrow \{s, t\}$;
do
  bestPath $\leftarrow$ path;
  bestProfit $\leftarrow$ 0;
  bestCost $\leftarrow$ `len(`*path,w*`)`;
  change $\leftarrow$ False;
  for $v \in POIs \wedge v \notin path$ **do**
    newPath $\leftarrow$ path $\oplus$ v;
    if $\frac{\texttt{satisfaction(}newPath,Users\texttt{)}}{\texttt{len(}newPath,w\texttt{)}} > \frac{bestProfit}{bestCost} \wedge \texttt{len(}newPath,w\texttt{)} \leq B$ **then**
      bestPath $\leftarrow$ newPath;
      bestProfit $\leftarrow$ `satisfaction(`*newPath,Users*`)`;
      bestCost $\leftarrow$ `len(`*newPath,w*`)`;
      change $\leftarrow$ True;
    end
  end
  if $change = False$ **then**
    bestProfit $\leftarrow$ `satisfaction(`*bestPath,Users*`)`;
    bestCost $\leftarrow$ `len(`*bestPath,w*`)`;
    for $v_m \in POIs \wedge v_m \notin path$ **do**
      for $v_k \in path \setminus \{s, t\}$ **do**
        newPath $\leftarrow$ `replace(`*path,*v_m*,*v_k`)`;
        if $\frac{\texttt{satisfaction(}newPath,Users\texttt{)}}{\texttt{len(}newPath,w\texttt{)}} > \frac{bestProfit}{bestCost} \wedge \texttt{len(}newPath,w\texttt{)} \leq B$ **then**
          bestPath $\leftarrow$ newPath;
          bestProfit $\leftarrow$ `satisfaction(`*newPath,Users*`)`;
          bestCost $\leftarrow$ `len(`*newPath,w*`)`;
          change $\leftarrow$ True;
        end
      end
    end
  end
  if $change = False$ **then**
    return path;
  end
  path $\leftarrow$ bestPath;
while $change = True$;

Στον αλγόριθμο αυτό, η συνάρτηση $satisfaction$ πραγματοποιεί τον υπολογισμό της ικανοποίησης ολόκληρου του group όπως έχει οριστεί από την αντίστοιχη αντικειμενική συνάρτηση. Δηλαδή, υπολογίζει την ικανοποίηση $p_j(T)$ του κάθε χρήστη P_j για όλο το μονοπάτι T και δεδομένων αυτών των τιμών επιστρέφει την τιμή $\Phi(p_1(T), p_2(T), ..., p_k(T))$.

Μπορούμε να δούμε ότι αυτός ο υπολογισμός επαναλαμβάνεται πάρα πολλές φορές σε κάθε βήμα του αλγορίθμου και ότι η πολυπλοκότητα του είναι γραμμική στο μέγεθος του group. Συνεπώς, όσο το μέγεθος του group μεγαλώνει, ο υπολογισμός αυτός γίνεται όλο και πιο χρονοβόρος.

Παρ'όλα αυτά, στη συνέχεια της εργασίας, για τη μελέτη του προβλήματος χρησιμοποιείται μόνο η αντικειμενική συνάρτηση TourGroupSum. Συνεπώς, λαμβάνοντας υπόψη το Λήμμα 1, στη συνέχεια της εργασίας, όπου απαιτείται εύρεση μονοπατιού, για τη μείωση του χρόνου εκτέλεσης, υπολογίζεται ισοδύναμα το μονοπάτι για group ενός ατόμου που αποτελείται μόνο από το κεντροειδές του αρχικού group.

2.2 Τεχνικές συσταδοποίησης

Η συσταδοποίηση είναι μια διαδικασία κατάταξης ενός συνόλου οντοτήτων σε ομάδες έτσι ώστε τα μέλη κάθε ομάδας να μοιράζονται κάποια κοινά χαρακτηριστικά. Εμφανίζεται σε διάφορες μορφές και παρατηρείται συχνά σε πολλαπλές πρακτικές εφαρμογές. Για παράδειγμα, οι βιολόγοι προσπαθούν να διαχωρίσουν τα διάφορα είδη των ζώων σε πιο ομοιογενείς κατηγορίες με βάση τα χαρακτηριστικά τους, τα τμήματα marketing των εταιριών κατηγοριοποιούν τους πελάτες τους ώστε να προσφέρουν πιο ενδιαφέροντα προϊόντα ανάλογα με τις προτιμήσεις τους ενώ οι εγκληματολόγοι διαχωρίζουν τις περιοχές των πόλεων σε διαφορετικά επίπεδα κινδύνου με βάση τις ιδιαιτερότητες κάθε περιοχής.

Όσον αφορά την επιστήμη υπολογιστών, οι τεχνικές συσταδοποίησης χρησιμοποιούνται συχνά σε συστήματα προσωποποιημένων συστάσεων στο διαδίκτυο και συνήθως εμφανίζονται στα πλαίσια της *Μη-Επιβλεπόμενης Μάθησης*, όπου ένας αλγόριθμος λαμβάνει ένα σύνολο δεδομένων που αντιπροσωπεύουν κάποιες οντότητες και προσπαθεί να τους προσθέσει ετικέτες έτσι ώστε να παρουσιάζουν κάποια ενδιαφέρουσα δομή.

Στη συνέχεια αυτής της ενότητας αναλύονται οι βασικότερες κατηγορίες συσταδοποίησης και αναφέρονται κάποιοι αλγόριθμοι για τον αντίστοιχο υπολογισμό των συστάδων.

2.2.1 Συσταδοποίηση K-Μέσων

Η πρώτη κατηγορία είναι η *Συσταδοποίηση K-Μέσων (K-Means Clustering)* ή *Συσταδοποίηση βασισμένη σε Κεντροειδή (Centroid-based Clustering)*. Η βασική

ιδέα της είναι ότι οι οντότητες που ανήκουν σε μία συστάδα μπορούν να εκπροσωπηθούν πλήρως από ένα κοινό στοιχείο που ονομάζεται το κεντροειδές της συστάδας. Πρόκειται για μια παλιά ιδέα η οποία εμφανίστηκε για πρώτη φορά με αυτόν τον όρο το 1967 [35]. Σύμφωνα με τη συσταδοποίηση K-Μέσων, ουσιαστικά κάθε οντότητα εκπροσωπείται από το κοντινότερο κεντροειδές στο χώρο και αναζητούνται τα κατάλληλα κεντροειδή που ελαχιστοποιούν την αθροιστική διασπορά των συστάδων. Ένας πιο αυστηρός ορισμός του προβλήματος δίνεται παρακάτω.

Πρόβλημα 2. ***(K-Means)*** *Δεδομένων παρατηρήσεων* $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$*, όπου κάθε παρατήρηση είναι ένα διάνυσμα πραγματικών αριθμών διάστασης* d*, ζητείται μία διαμέριση των* n *παρατηρήσεων σε* k $(\leq n)$ *σύνολα* $\boldsymbol{S} = \{S_1, S_2, ..., S_k\}$ *ώστε να ελαχιστοποιείται η συνάρτηση* $\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} \| \boldsymbol{x} - \boldsymbol{\mu}_i \|^2$ *όπου* $\boldsymbol{\mu}_i$ *η μέση τιμή των διανυσμάτων στο* S_i*.*

Αυτή η μορφή συσταδοποίησης έχει το πλεονέκτημα ότι συμπιέζει σε μεγάλο βαθμό την υπάρχουσα πληροφορία εκπροσωπώντας κάθε παρατήρηση από το κοντινότερο κεντροειδές. Αυτή η απλούστευση είναι πολλές φορές αρκετή αν και έχει το μειονέκτημα ότι περιορίζει τη μορφή των συστάδων καθώς υποθέτει ότι έχουν σφαιρική μορφή. Το πρόβλημα αυτό είναι NP-Hard και για την επίλυσή του χρησιμοποιούνται ευριστικές τεχνικές που δίνουν υποβέλτιστες λύσεις. Ο πιο διαδεδομένος αλγόριθμος επίλυσης του προβλήματος παρουσιάστηκε το 1982 [36] ως τεχνική συμπίεσης σημάτων.

Ο αλγόριθμος αυτός δέχεται ως είσοδο τον αριθμό των συστάδων και λειτουργεί επαναληπτικά όπου σε κάθε βήμα ανανεώνει τα κεντροειδή. Ξεκινώντας από ένα αρχικό σύνολο κεντροειδών, σε κάθε βήμα, η κάθε παρατήρηση ανατίθεται στη συστάδα με το πλησιέστερο κεντροειδές και στη συνέχεια το κάθε κεντροειδές ανανεώνεται ως η μέση τιμή των παρατηρήσεων που του έχουν ανατεθεί. Ο αλγόριθμος τερματίζεται όταν οι συστάδες σταματήσουν να αλλάζουν.

Μία αναλυτική περιγραφή του αλγορίθμου (χωρίς την επιλογή των αρχικών κεντροειδών) δίνεται παρακάτω. Η συνάρτηση *mean* υπολογίζει τη μέση τιμή ενός συνόλου διανυσμάτων ενώ η *distance* υπολογίζει την ευκλείδεια απόσταση δύο διανυσμάτων.

Algorithm 2: Lloyd's algorithm

```
Input: Number of clusters (k)
Output: Clusters
Initialize centroids;
newClusters ← ∅;
centroids ← ∅;
do
    oldClusters ← newClusters;
    newClusters ← ∅ ;
    for vector ∈ Observations do
        bestDist ← +∞;
        for i ← 1 to k do
            dist ← distance(centroid[i],vector);
            if dist<bestDist then
                bestDist ← dist;
                bestCluster ← i;
            end
            newClusters[i] ← newClusters[i] ∪ vector;
        end
    end
    for i ← 1 to k do
        centroid[i] ← mean(newClusters[i]);
    end
while oldClusters ≠ newClusters;
```

Σε κάθε επανάληψη, ο αλγόριθμος διατρέχει όλες τις n παρατηρήσεις και ελέγχει πιο είναι το κοντινότερο από τα k κεντροειδή διάστασης d. Άρα, είναι εύκολο να δούμε ότι η χρονική πολυπλοκότητα κάθε επανάληψης είναι $\mathcal{O}(nkd)$. Αν και στη χειρότερη περίπτωση ο αλγόριθμος μπορεί να κάνει πολύ μεγάλο αριθμό βημάτων [37], στην πράξη παρατηρείται ότι συγκλίνει πολύ γρήγορα σε συστάδες οι οποίες δεν διαφοροποιούνται. Έτσι, για μικρά k,d, ο χρόνος εκτέλεσης θεωρείται σχεδόν γραμμικός στον αριθμό των παρατηρήσεων καθιστώντας τον αλγόριθμο αυτό πολύ αποδοτικό.

Όσον αφορά την επιλογή των αρχικών κεντροειδών, ένας απλός τρόπος είναι να επιλεγούν k τυχαία διανύσματα από το σύνολο των παρατηρήσεων. Όπως είναι λογικό, η επιλογή αυτή καθορίζει σε μεγάλο βαθμό την επιτυχία του αλγορίθμου. Μπορεί μία κακή επιλογή τους να οδηγήσει σε συσταδοποίηση που είναι εμφανώς μη βέλτιστη. Για το λόγο αυτό, έχουν αναπτυχθεί πιο εξελιγμένες τεχνικές για την επιλογή των κεντροειδών και μία τέτοια επέκταση είναι ο αλγόριθμος των K-

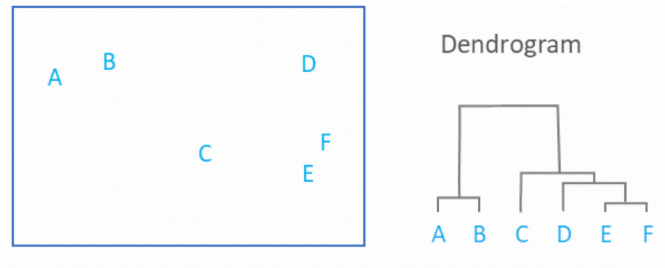
Μέσων++ [38].

Σύμφωνα με αυτή την εκδοχή του αλγορίθμου, αρχικά επιλέγεται ως πρώτο κεντροειδές ένα διάνυσμα από το σύνολο των παρατηρήσεων. Στη συνέχεια, τα επόμενα κεντροειδή επιλέγονται επαναληπτικά. Σε κάθε επανάληψη, για κάθε παρατήρηση υπολογίζεται η απόστασή της από το κοντινότερο ήδη υπάρχον κεντροειδές. Στη συνέχεια, το νέο κεντροειδές επιλέγεται τυχαία με βάση μία κατανομή πιθανότητας όπου η πιθανότητα μίας παρατήρησης να πάρει το ρόλο του κεντροειδούς είναι ανάλογη του τετραγώνου της προϋπολογισμένης απόστασής της. Διαισθητικά, σε κάθε επανάληψη, οι παρατηρήσεις που βρίσκονται μακριά από τα ήδη υπάρχοντα κεντροειδή έχουν μεγαλύτερη πιθανότητα να επιλεγούν. Με αυτό τον τρόπο, γίνεται προσπάθεια τα κεντροειδή να είναι ομοιόμορφα κατανεμημένα στο χώρο.

Από το Κεφάλαιο 3 και έπειτα, όπου αναφέρεται η συσταδοποίηση, θα υπονοείται χρήση του αλγορίθμου Κ-Μέσων++ όπως αναλύθηκε προηγουμένως. Τα υπόλοιπα ήδη συσταδοποίησης που περιγράφονται παρακάτω δεν έχουν χρησιμοποιηθεί στο πλαίσιο του σχεδιασμού τουριστικών διαδρομών για λόγους που θα αναφερθούν στη συνέχεια.

2.2.2 Ιεραρχική Συσταδοποίηση

Οι τεχνικές που αναφέρονται σε αυτή την ενότητα εμπίπτουν στην κατηγορία της *Ιεραρχικής Συσταδοποίησης (Hierarchical Clustering)* η οποία συχνά χαρακτηρίζεται και ως *Συσταδοποίηση βασισμένη στη Συνδεσιμότητα (Connectivity-based clustering)*. Το βασικό χαρακτηριστικό των τεχνικών αυτής της κατηγορίας είναι η δημιουργία μίας ιεραρχίας συστάδων [39]. Η κάθε συστάδα αποτελείται είτε από δύο μικρότερες συστάδες, είτε από δύο απλές παρατηρήσεις, είτε από μία συστάδα και μία παρατήρηση. Η ιεραρχία αυτή παρουσιάζεται με τη μορφή *δενδρογράμματος* και ένα παράδειγμα φαίνεται στο παρακάτω σχήμα.



Σχήμα 1: Ιεραρχική Συσταδοποίηση

Για τη σύνδεση δύο παρατηρήσεων μεταξύ τους, το κριτήριο που χρησιμοποιείται είναι κάποια μετρική απόστασης. Στον παρακάτω πίνακα δίνονται κάποιες τέτοιες μετρικές.

Μετρική	**Ορισμός**
Ευκλείδια απόσταση	$\| \mathbf{a} - \mathbf{b} \|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Απόσταση Manhattan	$\| \mathbf{a} - \mathbf{b} \|_1 = \sum_i \|a_i - b_i\|$
Μέγιστη απόσταση	$\| \mathbf{a} - \mathbf{b} \|_\infty = max_i \|a_i - b_i\|$

Πίνακας 1: Μετρικές απόστασης

Για να μπορέσει να πραγματοποιηθεί και σύνδεση μεταξύ συστάδων, απαιτείται ο συνδυασμός των παραπάνω ώστε να δημιουργηθούν μετρικές απόστασης μεταξύ συστάδων που οδηγούν σε διαφορετικούς τύπους συσταδοποίησης. Κάποιοι από τους πιο ευρέως διαδεδομένους δίνονται στον παρακάτω πίνακα. Έστω A,B τα δύο σύνολα (συστάδες) και d η μετρική απόστασης μεταξύ παρατηρήσεων.

Τύπος συσταδοποίησης	**Μετρική**
Πλήρους σύνδεσης	$max\{d(a,b) : a \in A, b \in B\}$
Μονής σύνδεσης	$min\{d(a,b) : a \in A, b \in B\}$
Μέσης σύνδεσης	$\frac{1}{\|A\| \cdot \|B\|} \sum_{a \in A} \sum_{b \in B} d(a,b)$

Πίνακας 2: Παραλλαγές ιεραρχικής συσταδοποίησης

Μία σημαντική διαφορά της Ιεραρχικής Συσταδοποίησης από τη Συσταδοποίηση Κ-Μέσων είναι ότι το αποτέλεσμα που δίνει δεν είναι ένας συγκεκριμένος αριθμός από συστάδες αλλά το πλήρες δενδρόγραμμα των παρατηρήσεων. Έτσι, ενώ στη Συσταδοποίηση Κ-Μέσων ο αριθμός k των συστάδων πρέπει να είναι καθορισμένος εκ των προτέρων, αυτό δεν είναι απαραίτητο στην Ιεραρχική Συσταδοποίηση. Εδώ, αφού κατασκευαστεί το δενδρόγραμμα, μπορεί να επιλεγεί εκ των υστέρων το επίπεδο στο οποίο είναι επιθυμητό να "κοπεί" ώστε να προκύψει ο επιθυμητός αριθμός συστάδων.

Ένα άλλο χαρακτηριστικό πλεονέκτημα τεχνικών αυτής της κατηγορίας είναι ότι μπορούν να λειτουργήσουν με οποιαδήποτε μετρική απόστασης, σε αντίθεση με τη Συσταδοποίηση Κ-Μέσων όπου η σύγκλιση των αλγορίθμων δεν είναι σίγουρη για μετρικές διαφορετικές της Ευκλείδιας απόστασης.

Παρ' όλα αυτά, η Ιεραρχική Συσταδοποίηση είναι πιο περίπλοκη ως διαδικασία και παρέχει πολλή πληροφορία που ενδεχομένως να μην είναι απαραίτητη. Πολλές φορές, σε απλές εφαρμογές, η απλή Συσταδοποίηση Κ-Μέσων είναι αρκετή για να δώσει τα επιθυμητά αποτελέσματα, συμπιέζοντας ταυτόχρονα την πληροφορία των παρατηρήσεων σε ένα μικρό σύνολο κεντροειδών.

Η Ιεραρχική Συσταδοποίηση βασίζεται σε δύο αλγοριθμικές προσεγγίσεις, τη *Συσσωρευτική* και τη *Διαιρετική*. Στην πρώτη περίπτωση, αρχικά η κάθε παρατήρηση αποτελεί μία ξεχωριστή συστάδα. Σε κάθε βήμα, ενώνονται δύο συστάδες ανάλογα με τη μετρική απόστασης που χρησιμοποιείται και η διαδικασία επαναλαμβάνεται μέχρι να καταλήξουμε σε μία συστάδα. Αλγόριθμοι αυτού του τύπου

παρουσιάστηκαν για πρώτη φορά τη δεκαετία του 1970 [40][41] και για κάποιες περιπτώσεις μετρικών απόστασης τα αποτελέσματα που δίνουν είναι βέλτιστα. Αν και υπάρχουν γρηγορότερες ειδικές περιπτώσεις, στη γενική περίπτωση η χρονική πολυπλοκότητα της Συσσωρευτικής Ιεραρχικής Συσταδοποίησης είναι $\mathcal{O}(n^3)$. Στη χειρότερη περίπτωση, σε κάθε βήμα θα ενώνεται μία παρατήρηση με μία υπάρχουσα συστάδα άρα θα απαιτούνται n επαναλήψεις ώστε ο αλγόριθμος να τερματιστεί. Επίσης, σε κάθε βήμα, λόγω της ένωσης συστάδων, η απόσταση κάθε παρατήρησης από τις υπάρχουσες συστάδες πρέπει να επαναϋπολογίζεται, κάτι που στη χειρότερη περίπτωση απαιτεί n^2 υπολογισμούς. Έτσι, προκύπτει η πολυπλοκότητα που αναφέρθηκε προηγουμένως.

Στη Διαιρετική Ιεραρχική Συσταδοποίηση, αρχικά θεωρούμε όλες τις παρατηρήσεις ως μία συστάδα και σε κάθε βήμα μία από τις υπάρχουσες συστάδες διαιρείται στα δύο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι κάθε παρατήρηση να αποτελεί μία ξεχωριστή συστάδα. Προφανώς, οι τρόποι να διαιρεθεί μία συστάδα σε δύο μικρότερες είναι εκθετικοί στο πλήθος τους ($\mathcal{O}(2^n)$) γι αυτό στην πράξη το πρόβλημα λύνεται με χρήση διαφόρων ευριστικών τεχνικών. Για παράδειγμα, ο αλγόριθμος DIANA (DIvisive ANAlysis Clustering) σε κάθε βήμα βρίσκει την παρατήρηση που απέχει περισσότερο από τις υπόλοιπες και την προσθέτει σε μια νέα συστάδα. Στη συνέχεια, προσθέτει στη νέα συστάδα της παρατηρήσεις που είναι πλησιέστερα σε αυτή παρά στη συστάδα που απέμεινε από τη διαίρεση.

Όπως αναφέρθηκε και προηγουμένως, η Ιεραρχική Συσταδοποίηση παρέχει πολύ μεγάλη πληροφορία για τη δομή των παρατηρήσεων και οι αλγόριθμοι που την υλοποιούν έχουν πολυπλοκότητα τέτοια που ο χρόνος εκτέλεσης γίνεται υπερβολικά μεγάλος ακόμα και για μετρίου μεγέθους σύνολα παρατηρήσεων. Συνεπώς, αν η φύση του προβλήματος προς μελέτη δεν απαιτεί την ιεραρχική δομή που προσφέρουν αυτοί οι αλγόριθμοι, είναι προτιμητέα η χρήση απλούστερων μεθόδων συσταδοποίησης όπως αυτή των K-Μέσων.

2.2.3 Συσταδοποίηση βασισμένη σε κατανομή

Αυτή η μορφή συσταδοποίησης είναι αρκετά διαφορετική από τις άλλες καθώς δεν βασίζεται σε αποστάσεις μεταξύ των παρατηρήσεων αλλά έχει μια πιο αυστηρή στατιστική φύση. Η *Συσταδοποίηση βασισμένη σε κατανομή (Distribution-based Clustering)*, όπως υπονοείται και από το όνομα, έχει ως στόχο να εκφράσει τις παρατηρήσεις ως δείγματα ενός συνόλου κατανομών πιθανότητας. Με άλλα λόγια, ο σκοπός είναι η εύρεση κάποιων κατανομών (και των παραμέτρων τους) εκ των οποίων είναι αναμενόμενο να προκύψουν οι υπάρχουσες παρατηρήσεις ως δείγματα. Στη συνέχεια, ως συστάδες προκύπτουν σύνολα παρατηρήσεων τα οποία πιθανότατα έχουν προκύψει από την ίδια κατανομή.

Καθώς η μορφή των παρατηρήσεων μπορεί να είναι αρκετά περίπλοκη ενώ οι κατανομές που επιλέγονται αρκετά απλές (π.χ. κανονικές), η ελαχιστοποίηση των

σφαλμάτων σε ικανοποιητικό βαθμό ενδεχομένως να απαιτεί μεγάλο πλήθος κατανομών [42]. Παρ' όλα αυτά, όσο αυξάνεται ο αριθμός των συνιστωσών κατανομών, η συγκεκριμένη μέθοδος εμφανίζει ένα συγκεκριμένο μειονέκτημα γνωστό ως *υπερπροσαρμογή (overfitting)*. Σε αυτή την περίπτωση, τα μοντέλα που προκύπτουν είναι κατάλληλα ώστε να περιγράψουν με πλήρη ακρίβεια τις παρατηρήσεις αλλά δεν έχουν τη δυνατότητα γενίκευσης, δηλαδή παρουσιάζουν μεγάλα σφάλματα σε μελλοντικές παρατηρήσεις. Γι αυτό το λόγο επιβάλλονται περιορισμοί στο πλήθος των συνιστωσών κατανομών ώστε να επιτυγχάνεται ένας συμβιβασμός μεταξύ των σφαλμάτων και της δυνατότητας γενίκευσης.

Αυτό το είδος συσταδοποίησης έχει το ιδιαίτερο χαρακτηριστικό της ισχυρής στατιστικής φύσης. Όμως, η υπόθεση ότι οι παρατηρήσεις είναι δείγματα συγκεκριμένων τύπων πιθανοτικών κατανομών αποτελεί ταυτόχρονα πλεονέκτημα και μειονέκτημα. Η υπόθεση αυτή είναι πολύ ισχυρή καθώς δίνει τη δυνατότητα όχι απλώς να ομαδοποιηθούν οι παρατηρήσεις αλλά να προσεγγισθεί και το ακριβές μοντέλο από το οποίο αυτές παρήχθησαν, δίνοντάς τους μια πιο αυστηρή περιγραφική δομή. Από την άλλη, αυτή η υπόθεση σπάνια μπορεί να εφαρμοστεί σε πραγματικά προβλήματα καθώς οι πραγματικές παρατηρήσεις περιέχουν μεγάλο ποσοστό θορύβου και η μορφή τους δεν είναι τέτοια ώστε να μπορούν να περιγραφούν εύκολα από ένα σύνολο απλών κατανομών πιθανότητας.

Από αλγοριθμική σκοπιά, η συσταδοποίηση με βάση την κατανομή βασίζεται σε προσεγγιστικούς επαναληπτικούς αλγόριθμους *Εκτίμησης Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation)*, με βασικότερο τον αλγόριθμο *Expectation-Maximization* [43] ο οποίος έχει δύο βασικούς στόχους. Ο πρώτος είναι να υπολογίσει από ποια από τις συνιστώσες κατανομές έχει προκύψει η κάθε παρατήρηση και αυτό συμβολίζεται ως ένα διάνυσμα πιθανοτήτων όπου κάθε στοιχείο αντιστοιχεί στην πιθανότητα η παρατήρηση να προέκυψε από την αντίστοιχη κατανομή. Ο δεύτερος στόχος είναι η εκτίμηση των παραμέτρων αυτών των κατανομών ώστε να ταιριάζουν πάνω στις παρατηρήσεις που τους αναλογούν.

Ο αλγόριθμος αυτός ξεκινάει από κάποιες αρχικές τιμές (ενδεχομένως τυχαίες) για τις παραπάνω παραμέτρους και εκτελεί επαναληπτικά δύο βήματα. Στο πρώτο βήμα, δεδομένων των κατανομών, υπολογίζεται η πιθανότητα μια παρατήρηση να προέκυψε από μια συγκεκριμένη κατανομή για κάθε δυνατό ζεύγος. Στο δεύτερο βήμα, δεδομένων αυτών των πιθανοτήτων, αναπροσαρμόζονται οι παράμετροι των συνιστωσών κατανομών ώστε να μεγιστοποιείται η πιθανότητα να έχουν εμφανιστεί οι παρατηρήσεις που αντιστοιχούν στην κάθε μία. Αυτή η διαδικασία επαναλαμβάνεται και τελικά συγκλίνει σε ένα σημείο που δεν πραγματοποιούνται περαιτέρω αλλαγές. Η λογική του αλγορίθμου μοιάζει αρκετά με τη Συσταδοποίηση Κ-Μέσων με τη διαφορά ότι οι παρατηρήσεις δεν ανήκουν αυστηρά σε μία κατανομή αλλά ταυτόχρονα σε όλες με διαφορετική πιθανότητα στην κάθε μία. Όπως και στη Συσταδοποίηση Κ-Μέσων, ενδεχομένως ο αλγόριθμος να μη συγκλίνει στο ολικό

βέλτιστο οπότε η πολλαπλή εκτέλεσή του με διαφορετικές αρχικές επιλογές παραμέτρων κρίνεται αναγκαία.

Όπως έγινε εμφανές και προηγουμένως, η Συσταδοποίηση με βάση την κατανομή είναι μια διαδικασία η οποία παρέχει ισχυρή πληροφορία για τα δεδομένα που εξετάζονται υπό την προϋπόθεση ότι υπάρχει κάποια προϋπάρχουσα γνώση για τη μορφή της κατανομής που ακολουθούν ώστε η μέθοδος να έχει τη θεωρητική βάση για να εφαρμοστεί.

2.3 Συστήματα προσωποποιημένων συστάσεων

Η ραγδαία εξάπλωση του διαδικτύου έχει ως φυσικό αποτέλεσμα την γιγάντωση του όγκου των διαθέσιμων πληροφοριών. Η διαχείρισή τους είναι μία από τις μεγαλύτερες προκλήσεις της σύγχρονης πληροφορικής. Έτσι, το πρόβλημα που αντιμετωπίζουν οι χρήστες του διαδικτύου είναι ότι έχουν πρόσβαση σε τόσο πολλές πηγές που αδυνατούν να τις διαχειριστούν αποδοτικά και να εντοπίσουν αυτό που πραγματικά αναζητούν. Αντιστοίχως, μεγάλες διαδικτυακές επιχειρήσεις (π.χ. Amazon, Spotify, Netflix) παρέχουν τόσο μεγάλο αριθμό προϊόντων που δυσκολεύονται να αναγνωρίσουν τις προτιμήσεις του εκάστοτε χρήστη ώστε να μεγιστοποιήσουν το κέρδος τους.

Για την αντιμετώπιση τέτοιων προβλημάτων έχουν αναπτυχθεί τα *Συστήματα Προσωποποιημένων Συστάσεων (Recommender Systems)*. Η μελέτη τους έχει ξεκινήσει από τη δεκαετία του 1990 [44] και μέσα στα χρόνια έχουν βρει εφαρμογή σε πολλούς τομείς όπως η σύσταση ταινιών [45][46] ή μουσικής [47]. Τα συστήματα αυτά συνήθως επιτελούν τρεις λειτουργίες. Η πρώτη είναι η συλλογή πληροφοριών για τα προφίλ των χρηστών όπως η ηλικία, οι συνήθεις προτιμήσεις, ο χρόνος που αφιερώνει ο χρήστης στο σύστημα, πληροφορίες από μέσα κοινωνικής δικτύωσης κ.λ.π. Η δεύτερη είναι η συλλογή των απαραίτητων χαρακτηριστικών των αντικειμένων που προτείνονται. Για παράδειγμα, σε ένα σύστημα συστάσεων για ταινίες τέτοια χαρακτηριστικά είναι το είδος της ταινίας, το έτος, οι πρωταγωνιστές, ο σκηνοθέτης κ.λ.π. Η τρίτη και πιο σημαντική λειτουργία είναι ο αποτελεσματικός συνδυασμός των παραπάνω πληροφοριών ώστε να γίνει εκτίμηση του ενδιαφέροντος του κάθε χρήστη για το κάθε αντικείμενο.

Υπάρχουν πολλοί τρόποι [48] να πραγματοποιηθεί αυτό το τελευταίο βήμα και αυτό είναι το ενδιαφέρον κομμάτι ενός συστήματος συστάσεων από αλγοριθμική πλευρά. Στη συνέχεια της ενότητας αυτής αναφέρονται δύο από τις βασικότερες προσεγγίσεις και παρουσιάζονται τα πλεονεκτήματα και τα μειονεκτήματά τους.

2.3.1 Σύσταση με βάση το περιεχόμενο

Πολλές διαδικτυακές υπηρεσίες βασίζουν τη λειτουργία τους σε *Συστήματα Σύστασης με βάση το περιεχόμενο (Content-based Recommender Systems)* [49]. Η

βασική αρχή που ακολουθούν αυτά τα συστήματα είναι η σύγκριση των χαρακτηριστικών των διαθέσιμων αντικειμένων με τις προτιμήσεις του χρήστη όπως αυτές έχουν διαμορφωθεί στο παρελθόν. Δηλαδή, για κάθε χρήστη διατηρείται ένα διάνυσμα όπου οι τιμές του αντιστοιχούν στην προτίμηση του χρήστη σε διάφορα χαρακτηριστικά των αντικειμένων. Με αυτή τη λογική, τα αντικείμενα θα πρέπει να είναι απεικονισμένα με τέτοιο τρόπο ώστε με χρήση κάποιας κατάλληλης συνάρτησης να μπορεί να γίνει υπολογισμός του ενδιαφέροντος του χρήστη γι αυτά με βάση τα χαρακτηριστικά του διανύσματός του.

Επίσης, όταν ο χρήστης βαθμολογεί διάφορα αντικείμενα είτε άμεσα (μέσω συστήματος βαθμολόγησης) είτε έμμεσα (μέσω αριθμού κλικ, αγορών κ.λ.π.) [50][51] το προφίλ του που είναι αποθηκευμένο στο σύστημα προσαρμόζεται έτσι ώστε να ταιριάζει στα αντικείμενα που έχει βαθμολογήσει υψηλά. Με αυτό τον τρόπο, το σύστημα συστάσεων καταφέρνει να κάνει συσχετίσεις μεταξύ χρηστών και αντικειμένων και να κάνει προτάσεις κοντινές στις προηγούμενες προτιμήσεις των χρηστών.

Οι συσχετίσεις αυτές πραγματοποιούνται κάνοντας χρήση κάποιας απλής συνάρτησης. Έστω ένα διάνυσμα $\mathbf{u}$ που περιλαμβάνει τις προτιμήσεις του χρήστη σε διάφορα χαρακτηριστικά και $\mathbf{p}$ ένα διάνυσμα που αντιστοιχεί στο βαθμό που ένα αντικείμενο περιέχει τα αντίστοιχα χαρακτηριστικά. Δύο πολύ διαδεδομένες συναρτήσεις που δίνουν το ενδιαφέρον του χρήστη για το συγκεκριμένο αντικείμενο δίνονται στον παρακάτω πίνακα.

Μετρική	**Ορισμός**
Ομοιότητα συνημιτόνου	$sim(\mathbf{u}, \mathbf{p}) = \frac{\mathbf{u} \cdot \mathbf{p}}{\lvert\mathbf{u}\rvert \lvert\mathbf{p}\rvert}$
Ευκλείδια ομοιότητα	$sim(\mathbf{u}, \mathbf{p}) = \frac{1}{1+\lvert\mathbf{u}-\mathbf{p}\rvert}$

Πίνακας 3: Μετρικές ομοιότητας

Όπως είναι φανερό, για να έχει νόημα η συγκεκριμένη μέθοδος συστάσεων, τα στοιχεία των διανυσμάτων $\mathbf{u}$ και $\mathbf{p}$ θα πρέπει να αντιστοιχούν σε συγκεκριμένα χαρακτηριστικά των διαθέσιμων αντικειμένων που καθορίζουν τις προτιμήσεις των χρηστών. Όταν τα αντικείμενα είναι αποθηκευμένα με τέτοιο τρόπο ώστε να παρέχονται σαφώς τα χαρακτηριστικά τους, η διαδικασία συστάσεων γίνεται αρκετά απλή. Παρ' όλα αυτά, πάρα πολύ συχνά, τα δεδομένα σχετικά με τα διαθέσιμα αντικείμενα δεν παρέχονται σε τόσο βολική μορφή δηλαδή ενδεχομένως να υπάρχει μόνο μια περιγραφή σε μορφή κειμένου από την οποία τα χαρακτηριστικά πρέπει να εξαχθούν αυτοματοποιημένα με κάποιο αλγόριθμο ανάλυσης κειμένου. Σε τέτοιες περιπτώσεις, επειδή οι αλγόριθμοι αυτοί δεν καταφέρνουν πάντα να εξάγουν τα κατάλληλα χαρακτηριστικά για τα αντικείμενα, τα συστήματα συστάσεων με βάση το περιεχόμενο παρουσιάζουν χαμηλότερη απόδοση.

Αυτή η μέθοδος εφαρμόζεται σε πολλά πραγματικά συστήματα συστάσεων διότι παρουσιάζει αρκετά πλεονεκτήματα. Δεδομένου ενός προφίλ για ένα χρήστη, μπορεί απευθείας να υπολογιστεί η προτίμησή του για όλα τα διαθέσιμα αντικείμενα. Δηλαδή, το σύστημα μπορεί να λειτουργήσει σωστά ακόμα και με ελάχιστους χρήστες. Επίσης, η συσχέτιση της προτίμησης με συγκεκριμένα χαρακτηριστικά των αντικειμένων κάνει τη μέθοδο συστάσεων πιο αιτιοκρατική. Αυτό είναι πολύ σημαντικό σε περιπτώσεις που αναζητείται ο λόγος για τον οποίο το σύστημα πρότεινε κάποιο αντικείμενο σε κάποιο χρήστη.

Από την άλλη, η σύσταση με βάση το περιεχόμενο παρουσιάζει και συγκεκριμένα μειονεκτήματα. Όπως αναφέρθηκε και προηγουμένως, το σημαντικότερο πρόβλημα είναι η δυσκολία στην εξαγωγή των απαραίτητων χαρακτηριστικών για τα αντικείμενα και τα προφίλ των χρηστών καθώς αυτά ίσως πρέπει να προκύψουν από ακατέργαστα κείμενα ή ακόμα πιο δύσκολες μορφές όπως ηχητικά σήματα και βίντεο σε περιπτώσεις συστάσεων τραγουδιών ή ταινιών. Ακόμα, καθώς η δημιουργία των προφίλ των χρηστών βασίζεται σε προηγούμενες επιλογές αντικειμένων, το σύστημα αδυνατεί να πραγματοποιήσει αποτελεσματικές συστάσεις όταν κάποιος χρήστης δεν έχει χρησιμοποιήσει το σύστημα πολλές φορές, κάτι που είναι γνωστό ως το *Πρόβλημα της Ψυχρής Εκκίνησης (Cold-Start Problem)* [52]. Τέλος, ένα άλλο σημαντικό πρόβλημα είναι ότι το σύστημα βασίζεται πλήρως στις προηγούμενες επιλογές των χρηστών άρα δεν έχει τη δυνατότητα να τους προτείνει αντικείμενα που ενδεχομένως τους αρέσουν αλλά έχουν χαρακτηριστικά που απέχουν από αυτά των προηγούμενων αντικειμένων για τα οποία είχαν υψηλή προτίμηση. Έτσι, τα συστήματα συστάσεων με βάση το περιεχόμενο τείνουν να εξειδικεύονται πολύ σε συγκεκριμένες κατηγορίες αντικειμένων.

Προβλήματα όπως αυτά που αναφέρθηκαν, γίνεται προσπάθεια να επιλυθούν με χρήση μεθόδων διαφορετικής λογικής όπως η σύσταση με συνεργατικό φιλτράρισμα.

2.3.2 Σύσταση με συνεργατικό φιλτράρισμα

Η σύσταση με *Συνεργατικό Φιλτράρισμα (Collaborative Filtering)* [53] είναι η πιο διαδεδομένη μέθοδος που χρησιμοποιείται στα σύγχρονα συστήματα συστάσεων. Το χαρακτηριστικό της που τη διαφοροποιεί από τη σύσταση με βάση το περιεχόμενο είναι ότι η λειτουργία της είναι πλήρως ανεξάρτητη από το περιεχόμενο και τα χαρακτηριστικά των αντικειμένων που περιέχονται στο σύστημα. Το μόνο στοιχείο που έχει σημασία είναι η προτίμηση που έχουν οι χρήστες για τα διάφορα αντικείμενα. Το σύστημα, γνωρίζοντας τις προτιμήσεις των χρηστών για κάποια αντικείμενα, μπορεί να εντοπίσει ζεύγη χρηστών που πιθανότατα έχουν παρόμοια ενδιαφέροντα παρατηρώντας όμοιο βαθμό προτίμησης σε μεγάλο αριθμό κοινών αντικειμένων. Για παράδειγμα, σε ένα συστάσεων ταινιών, αν δύο χρήστες U_1, U_2 έχουν βαθμολογήσει τις ταινίες M_1, M_2 με υψηλή βαθμολογία και ο χρήστης U_1 έχει βαθμολογήσει υψηλά την ταινία M_3, τότε είναι πολύ πιθανό η προτίμηση του

χρήστη U_2 για την ταινία M_3 να είναι κι αυτή υψηλή. Είναι προφανές από αυτό το απλό παράδειγμα ότι η εύρεση της ομοιότητας μεταξύ δύο χρηστών εξαρτάται αποκλειστικά από τις βαθμολογίες και καθόλου από το περιεχόμενο, επομένως η μέθοδος είναι απολύτως ίδια είτε πρόκειται για σύσταση τραγουδιών, είτε ταινιών, είτε ξενοδοχείων.

Δύο συνηθισμένες μετρικές [53] για την εύρεση της ομοιότητας δύο χρηστών είναι η ομοιότητα συνημιτόνου που αναφέρθηκε και στην περίπτωση της σύστασης με βάση το περιεχόμενο καθώς και η *Συσχέτιση Pearson (Pearson correlation)*. Για τον υπολογισμό αυτών των μετρικών απαιτούνται δύο διανύσματα που θα περιλαμβάνουν τις βαθμολογίες ή προτιμήσεις των δύο χρηστών για το σύνολο των αντικειμένων που έχουν βαθμολογήσει και οι δύο. Έστω $r_{a,i}$ ο συμβολισμός για τη βαθμολογία του αντικειμένου i από τον χρήστη a και $\bar{r}_a$ η μέση τιμή των βαθμολογιών που έχει δώσει ο χρήστης a στα διάφορα αντικείμενα. Τότε, η ομοιότητα συνημιτόνου μεταξύ δύο χρηστών ορίζεται ως εξής:

$$sim(a,b) = \frac{\sum_i r_{a,i} r_{b,i}}{\sqrt{\sum_i r_{a,i}^2}\sqrt{\sum_i r_{b,i}^2}}$$

Αντίστοιχα, η Συσχέτιση Pearson δίνεται από τον παρακάτω τύπο:

$$sim(a,b) = \frac{\sum_i (r_{a,i} - \bar{r_a})(r_{b,i} - \bar{r_b})}{\sqrt{\sum_i (r_{a,i} - \bar{r_a})^2}\sqrt{\sum_i (r_{b,i} - \bar{r_b})^2}}$$

Το πλεονέκτημα που παρουσιάζει η Συσχέτιση Pearson είναι ότι πραγματοποιεί ενός είδους κανονικοποίηση. Μέσω της αφαίρεσης της μέσης τιμής των βαθμολογιών του κάθε χρήστη, αποφεύγονται σφάλματα σε περιπτώσεις που οι χρήστες έχουν πολύ διαφορετικού τρόπους βαθμολόγησης. Για παράδειγμα, σε ένα σύστημα σύστασης ταινιών, αν ένα χρήστης έχει μέσο όρο βαθμολογιών 4.5 αστεριών, η βαθμολόγηση μια ταινίας με 5 αστέρια δεν έχει ιδιαίτερο ενδιαφέρον. Αντιθέτως, σε ένα χρήστη με μέσο όρο 2 αστεριών, η βαθμολόγηση μιας ταινίας με 5 αστέρια σημαίνει ότι του άρεσε πάρα πολύ και είναι ιδιαίτερα σημαντική.

Η άλλη λειτουργία που καλείται να εκτελέσει ένα σύστημα συστάσεων συνεργατικού φιλτραρίσματος είναι η πρόβλεψη της βαθμολογίας ή του ενδιαφέροντος ενός χρήστη για ένα αντικείμενο. Εφόσον έχει βρεθεί η ομοιότητα ενός χρήστη με τον καθένα από τους υπόλοιπους, τότε η εκτίμηση για τη βαθμολογία του σε ένα συγκεκριμένο αντικείμενο ορίζεται ως μια αθροιστική συνάρτηση των βαθμολογιών των παρόμοιων με αυτόν χρηστών. Προφανώς, η συνάρτηση αυτή πρέπει να είναι τέτοια ώστε να δίνεται μεγαλύτερη έμφαση στις βαθμολογίες των παρόμοιων χρηστών και λιγότερη σε χρήστες με διαφορετικά ενδιαφέροντα. Αν $p_{a,i}$ είναι η εκτίμηση της βαθμολογίας του χρήστη a για το αντικείμενο i, ένας συνηθισμένος τρόπος για τον υπολογισμό της δίνεται από την παρακάτω σχέση:

$$p_{a,i} = \bar{r}_a + \frac{\sum_u ((r_{u,i} - \bar{r}_u) \cdot sim(a,u))}{\sum_u sim(a,u)}$$

Με αυτή την τεχνική, ένα σύστημα συστάσεων συνεργατικού φιλτραρίσματος μπορεί να εκτιμήσει το ενδιαφέρον του κάθε χρήστη για τα αντικείμενα που δεν έχει αξιολογήσει και να του προτείνει αυτά που είναι πιο πιθανό να τον ενδιαφέρουν.

Αυτή η μέθοδος συστάσεων είναι ευρέως διαδεδομένη διότι παρουσιάζει σημαντικά πλεονεκτήματα σχετικά με τη σύσταση με βάση το περιεχόμενο. Καταρχάς, καθιστά τη σχεδίαση του συστήματος αρκετά απλούστερη αφού δεν απαιτεί την εξαγωγή χαρακτηριστικών από τα αντικείμενα ούτε κάποια διαδικασία δημιουργίας προφίλ για τους χρήστες. Οι βαθμολογίες των χρηστών είναι αρκετές για να περιγράψουν τις προτιμήσεις τους. Επίσης, το συνεργατικό φιλτράρισμα δίνει τη δυνατότητα γενίκευσης, δηλαδή της σύστασης στο χρήστη αντικειμένων που είναι αρκετά διαφορετικά από αυτά που έχει ήδη βαθμολογήσει αλλά ενδεχομένως να τον ενδιαφέρουν αφού ενδιαφέρουν και άλλους παρόμοιους χρήστες με αυτόν.

Η μέθοδος όμως ακολουθείται και από μια σειρά μειονεκτημάτων. Το σημαντικότερο από αυτά συναντάται συχνά με την ονομασία *Αραιότητα Δεδομένων (Data Sparsity)* [54] και αφορά την έλλειψη αρκετής πληροφορίας για να μπορεί να υπολογιστεί η ομοιότητα μεταξύ δύο χρηστών. Όπως αναφέρθηκε προηγουμένως, η χρήση της ομοιότητας συνημιτόνου και της συσχέτισης Pearson απαιτεί τη γνώση βαθμολογιών για αντικείμενα που έχουν βαθμολογήσει και οι δύο χρήστες. Στην περίπτωση που οι χρήστες έχουν βαθμολογήσει ελάχιστα αντικείμενα ο καθένας σε σχέση με το συνολικό πλήθος των υπαρχόντων, είναι αρκετά πιθανό τα διάφορα ζευγάρια χρηστών να μην έχουν βαθμολογήσει κοινά αντικείμενα ή αυτά να είναι πολύ λίγα ώστε να μπορεί να γίνει ακριβής εκτίμηση νέων βαθμολογιών.

Ο χώρος των συστημάτων προσωποποιημένων συστάσεων αποτελεί ένα πολύ ενεργό ερευνητικό πεδίο [55] και το συνεργατικό φιλτράρισμα είναι η καθιερωμένη μέθοδο συστάσεων. Για την εξάλειψη των αρνητικών της στοιχείων, έχουν αναπτυχθεί αρκετές τροποποιήσεις όπως τα *Υβριδικά Συστήματα Συστάσεων (Hybrid Recommender Systems)* [56] τα οποία βασίζονται στο συνδυασμό πληροφορίας για τις βαθμολογίες των αντικειμένων αλλά και σε ορισμένα στοιχεία του περιεχομένου προκειμένου να εκμεταλλευτούν τα πλεονεκτήματα και των δύο προαναφερθέντων τεχνικών.

3 Χρήση συσταδοποίησης στο σχεδιασμό ομαδικών διαδρομών

3.1 Εισαγωγή

Όπως αναφέρθηκε και στην Ενότητα 2.1.3. το πρόβλημα της Σύστασης Διαδρομών σε Ομάδες αφορά τη σχεδίαση μιας διαδρομής που να μεγιστοποιεί την ικανοποίηση ενός group τουριστών το οποίο μπορεί να αποτελείται από άτομα με διαφορετικές προτιμήσεις. Αυτό έχει ως αποτέλεσμα να υπάρχουν μέλη του group που να μην μένουν πολύ ικανοποιημένα από την κοινή διαδρομή ενώ μια εναλλακτική διαδρομή θα μεγιστοποιούσε την ικανοποίησή τους, μειώνοντας ταυτόχρονα την ικανοποίηση κάποιων άλλων. Αυτό το στοιχείο του αμοιβαίου συμβιβασμού είναι έντονο στο συγκεκριμένο πρόβλημα και η περαιτέρω μελέτη του δίνει νέες δυνατότητες στην κατανόηση και την επέκταση των συστημάτων σχεδιασμού τουριστικών διαδρομών.

Ένας τρόπος χαλάρωσης αυτών των συμβιβασμών ώστε να αυξηθεί η ικανοποίηση του κάθε χρήστη είναι ο διαχωρισμός ενός group σε μικρότερα, ανάλογα με τα επιμέρους ενδιαφέροντα των χρηστών. Με αυτό τον τρόπο οι χρήστες θα μπορούσαν να ακολουθήσουν κοινές διαδρομές με άτομα παρόμοιων προτιμήσεων με αποτέλεσμα να πραγματοποιήσουν μικρότερους προσωπικούς συμβιβασμούς.

Όπως περιγράφεται και στη συνέχεια του κεφαλαίου, πληροφορίες από ηλεκτρονικά μέσα κοινωνικής δικτύωσης δίνουν τη δυνατότητα απεικόνισης των τουριστών (και των προτιμήσεών τους) καθώς και των αξιοθέατων σε κάποιο διανυσματικό χώρο. Επομένως, είναι λογικό ο διαχωρισμός σε μικρότερα groups να γίνει κάνοντας χρήση των τεχνικών συσταδοποίησης που αναφέρθηκαν στην Ενότητα 2.2. Η φύση του προβλήματος είναι τέτοια που καθιστά χρήσιμη κυρίως τη μέθοδο συσταδοποίησης K-Μέσων. Η Ιεραρχική Συσταδοποίηση, πέραν του ότι είναι πιο χρονοβόρα από άποψη αλγοριθμικής πολυπλοκότητας, δίνει μεγάλη πληροφορία για τη δομή του group, κάτι που δεν είναι απαραίτητο καθώς το ενδιαφέρον περιορίζεται στο διαχωρισμό σε μικρότερα groups και όχι στις ιεραρχικές σχέσεις απόστασης π.χ. μεταξύ ζευγών ή τριάδων χρηστών. Από την άλλη, η χρήση Συσταδοποίησης με βάση την Κατανομή δεν έχει θεωρητική βάση ώστε να εφαρμοστεί καθώς θα ήταν εξιδανικευμένο το να θεωρηθεί ότι οι προτιμήσεις των χρηστών προκύπτουν από κάποια απλή κατανομή πιθανότητας.

Έτσι, το ενδιαφέρον του συγκεκριμένου κεφαλαίου επικεντρώνεται στη χρήση συσταδοποίησης K-Μέσων στο πρόβλημα της Σύστασης Διαδρομών σε Ομάδες. Χρησιμοποιώντας συσταδοποίηση, είναι χρήσιμη η ποσοτικοποίση του κέρδους στην ικανοποίηση των τουριστών όσο μεταβάλλεται ο αριθμός των συστάδων. Για παράδειγμα, οι χρήστες σε δύο groups των 10 ατόμων που θα ακολουθήσουν διαφορετικές διαδρομές είναι λογικό να μείνουν πιο ικανοποιημένοι από το να ακολουθούσαν

μια κοινή διαδρομή και οι 20. Με βάση αυτή τη συλλογιστική, τα ερωτήματα που προκύπτουν είναι τα εξής:

- Ποιος είναι ένας ικανοποιητικός τρόπος διαχωρισμού των τουριστών;
- Τι βελτίωση στη συνολική ικανοποίηση επιτυγχάνεται χωρίζοντας ένα group σε μικρότερες συστάδες;
- Από τι παράγοντες εξαρτάται η βελτίωση αυτή;

Στη συνέχεια του κεφαλαίου γίνεται προσπάθεια να δοθούν απαντήσεις στα παραπάνω ερωτήματα. Ο διαχωρισμός της ομάδας τουριστών σε μικρότερα groups με διαφορετικές διαδρομές οδηγεί τελικά στην επόμενη βασική υπόθεση.

Υπόθεση 1. *Η συνολική ικανοποίηση ενός συνόλου N τουριστών είναι αύξουσα συνάρτηση του αριθμού k των συστάδων στις οποίες θα χωριστεί αν κάθε συστάδα ακολουθήσει το δικό της μονοπάτι και παρουσιάζει άνω όριο στην περίπτωση δημιουργίας N ατομικών διαδρομών.*

Μία ενδιαφέρουσα ιδιότητα της συσταδοποίησης K-Μέσων που αναφέρθηκε και προηγουμένως είναι η δυνατότητα συμπίεσης της πληροφορίας. Η ιδιότητα αυτή είναι πάρα πολύ χρήσιμη αν το κάθε κεντροειδές μπορεί να "εκφράσει" σε ικανοποιητικό βαθμό τα διανύσματα της αντίστοιχης συστάδας χωρίς σημαντικές απώλειες. Στα πλαίσια της Σύστασης Διαδρομών σε Ομάδες, έχοντας ένα σύνολο από συστάδες, είναι λογικό οι χρήστες να διατηρούν ένα υψηλό ποσοστό της ικανοποίησής τους ακολουθώντας τη διαδρομή της συστάδας αντί για μία πλήρως προσωποποιημένη διαδρομή. Αυτό το στοιχείο οδηγεί στην παρακάτω υπόθεση, η οποία αποτελεί μία εναλλακτική διατύπωση της προηγούμενης.

Υπόθεση 2. *Έστω ένα σύνολο N διανυσμάτων τουριστών και k κεντροειδή που προκύπτουν από συσταδοποίηση K-Μέσων. Υπάρχει $k < N$ τέτοιο ώστε η μέση ικανοποίηση ενός χρήστη από τη διαδρομή της συστάδας του να αποτελεί πολύ μεγάλο ποσοστό της ικανοποίησής του από μία πλήρως προσωποποιημένη διαδρομή.*

Η επιβεβαίωση αυτής της υπόθεσης οδηγεί στο συμπέρασμα ότι για ένα τυχαίο χρήστη, με μεγάλη πιθανότητα, υπάρχει συστάδα στην οποία μπορεί να ενταχθεί προσφέροντάς του μεγάλο βαθμό ικανοποίησης.

Στη συνέχεια του κεφαλαίου αυτού, παρουσιάζονται οι μέθοδοι διαχωρισμού των τουριστών σε συστάδες, η προσαρμογή του προβλήματος σε πραγματικά δεδομένα χρηστών και εξετάζονται η ισχύς και οι παράμετροι των παραπάνω υποθέσεων με βάση πειραματική αξιολόγηση.

3.2 Ο αλγόριθμος K-Μέσων για χρήστες & POIs

Στη συνέχεια μελετάται η διαφοροποίηση στη συνολική ικανοποίηση των χρηστών καθώς χωρίζονται σε μικρότερες πιο ομοιογενείς ομάδες που θα ακολουθήσουν διαφορετικές διαδρομές. Ο κάθε χρήστης και το κάθε αξιοθέατο απεικονίζονται ως διανύσματα πραγματικών αριθμών όπως εξηγείται στην επόμενη ενότητα. Η απεικόνιση αυτή μας επιτρέπει να χρησιμοποιήσουμε τη συσταδοποίηση K-Μέσων για το διαχωρισμό των χρηστών σε μικρότερες ομάδες.

Η διαδικασία αυτή γίνεται κάνοντας χρήση δύο αλγορίθμων. Στην πρώτη περίπτωση, γίνεται απευθείας συσταδοποίηση των χρηστών, κάνοντας χρήση των διανυσμάτων τους, υπολογίζονται οι επιμέρους διαδρομές και οι συνολικές τους αξίες προστίθενται. Παρακάτω δίνεται μια τυπική περιγραφή του αλγορίθμου. Η συνάρτηση *kmeans* υλοποιεί τον αλγόριθμο K-Μέσων++ όπως περιγράφηκε στην Ενότητα 2.2.1. και η *findPath* βρίσκει ένα μονοπάτι μεγιστοποιώντας την αθροιστική ικανοποίηση των μελών του group με χρήση του αλγορίθμου BestRatio+ που αναφέρθηκε στην Ενότητα 2.1.3. Τέλος, η *satisfaction* υπολογίζει τη συνολική ικανοποίηση ενός group από τη διαδρομή που του προτείνεται.

Algorithm 3: Users' K-Means Clustering

```
Input: Number of clusters (k)
Output: Overall satisfaction (totalSat)
totalSat ← 0;
clusters ← kmeans(k,Users);
for i ← 1 to k do
    groupPath ← findPath(clusters[i]);
    totalSat ← totalSat + satisfaction(groupPath,clusters[i]);
end
```

Στη δεύτερη περίπτωση, γίνεται συσταδοποίηση των POIs και ο κάθε χρήστης αντιστοιχίζεται με την κοντινότερη συστάδα από POIs. Στη συνέχεια, χρήστες οι οποίοι έχουν αντιστοιχηθεί στην ίδια συστάδα ομαδοποιούνται και με αυτό τον τρόπο υλοποιείται μια έμμεση συσταδοποίηση των χρηστών. Διαισθητικά, υποθέτουμε ότι χρήστες οι οποίοι ενδιαφέρονται πολύ για μια ομάδα από αξιοθέατα, ενδεχομένως να έχουν γενικά παρόμοια ενδιαφέροντα. Τέλος, υπολογίζονται διαφορετικά μονοπάτια για όλες τις συστάδες χρηστών και υπολογίζεται η συνολική τους αξία όπως και στην προηγούμενη περίπτωση. Οι συναρτήσεις *mean* και *distance* έχουν την ίδια λειτουργικότητα με αυτή στην Ενότητα 2.2.1.

Algorithm 4: POIs' K-Means Clustering

```
Input: Number of clusters (k)
Output: Overall satisfaction (totalSat)
totalSat ← 0;
clusters ← kmeans(k,POIs);
for i ← 1 to k do
    groups[i] ← ∅
end
for user ∈ Users do
    closestDistance ← +∞;
    for i ← 1 to k do
        centroid ← mean(clusters[i]);
        if distance(centroid,user) < closestDistance then
            closestCentroid ← i;
            closestDistance ← distance(centroid,user);
        end
    end
    groups[closestCentroid]← groups[closestCentroid] ∪ user
end
for i ← 1 to k do
    if groups[i] ≠ ∅ then
        groupPath ← findPath(groups[i]);
        totalSat ← totalSat + satisfaction(groupPath,groups[i]);
    end
end
```

Έχοντας περιγράψει πλήρως τους τρόπους με τους οποίους γίνεται διαμέριση των χρηστών για τη μεγιστοποίηση της ικανοποίησής τους, στις επόμενες ενότητες αναλύεται η συμπεριφορά τους μέσω πειραματικής αξιολόγησης.

3.3 Ανάλυση του συνόλου δεδομένων

Όπως αναφέρθηκε στην Ενότητα 2.1.3 το πρόβλημα του σχεδιασμού ομαδικών τουριστικών διαδρομών αποτελεί ένα δύσκολο υπολογιστικό πρόβλημα και η αλγοριθμική του επίλυση βασίζεται σε ευριστικές τεχνικές. Γι αυτό το λόγο, κρίνεται απαραίτητη η πειραματική αξιολόγηση των αλγορίθμων με χρήση πραγματικών δεδομένων.

Για τη διαδικασία της αξιολόγησης είναι απαραίτητη η ύπαρξη ενός συνόλου δεδομένων που να περιέχει πληροφορίες για τους χρόνους μετάβασης μεταξύ των

αξιοθέατων μιας πόλης καθώς και το βαθμό ικανοποίησης ενός συνόλου τουριστών για κάθε αξιοθέατο της πόλης. Η πρώτη πληροφορία είναι εύκολο να βρεθεί μέσω κάποιας ηλεκτρονικής υπηρεσίας παροχής χαρτών. Παρ'όλα αυτά, μια διαδικασία για τη συλλογή της δεύτερης πληροφορίας θα απαιτούσε πλήρη συμμετοχή των τουριστών σε κουραστικό βαθμό και ενδεχομένως να ήταν ακριβή και χρονοβόρα. Γι αυτό το λόγο, χρησιμοποιούνται εκτενώς στη βιβλιογραφία αυτοματοποιημένοι τρόποι συλλογής δεδομένων από ιστοσελίδες κοινωνικής δικτύωσης [57] ώστε να ληφθούν έμμεσα πληροφορίες για τις διαδρομές που ακολουθούν οι τουρίστες σε κάθε πόλη και την ικανοποίησή τους από τα αξιοθέατα αυτών των διαδρομών.

Το σύνολο δεδομένων που χρησιμοποιείται στο κεφάλαιο αυτό είναι ανοικτά διαθέσιμο από προηγούμενη δημοσίευση των Anagnostopoulos et al. [34] σχετικά με τη Σύσταση Διαδρομών σε Ομάδες που αναφέρθηκε στην Ενότητα 2.3. Αρχικά, χρησιμοποιείται ένα σύνολο διαδρομών των Baraglia et al. [58] που έχει προκύψει από συνδυασμό φωτογραφιών χρηστών του Flickr και σελίδες της Wikipedia με χωρικό προσδιορισμό και περιέχει πληροφορίες για διαδρομές που ακολούθησαν οι χρήστες στις τρεις Ιταλικές πόλεις Ρώμη, Φλωρεντία, Πίζα. Για κάθε σημείο που επισκέφθηκαν δίνεται ο αριθμός των φωτογραφιών που τράβηξαν, καθώς και ο χρόνος που έμειναν σε αυτό. Το κάθε POI αναπαρίσταται από ένα διάνυσμα πραγματικών αριθμών που προκύπτει από λεξικογραφική ανάλυση της αντίστοιχης σελίδας της Wikipedia με χρήση της μεθόδου *Λανθάνουσας Σημασιολογικής Δεικτοδότησης (Latent Semantic Indexing – LSI)* [59] η οποία αναλύεται περισσότερο στην Ενότητα 4.3. Αντίστοιχα, ο κάθε χρήστης απεικονίζεται στον ίδιο διανυσματικό χώρο ως το σταθμισμένο άθροισμα των διανυσμάτων των αξιοθέατων που επισκέφθηκε με βάρη τον αριθμό των φωτογραφιών που έβγαλε στο καθένα από αυτά. Έτσι, ως μετρική για την ικανοποίηση ενός χρήστη από την επίσκεψη σε ένα POI, θεωρείται η τιμή του εσωτερικού γινομένου των δύο αυτών διανυσμάτων. Τέλος, ως χρόνος επίσκεψης για ένα σημείο θεωρείται ο μέσος όρος των χρόνων που έκαναν οι διάφοροι χρήστες που το επισκέφθηκαν.

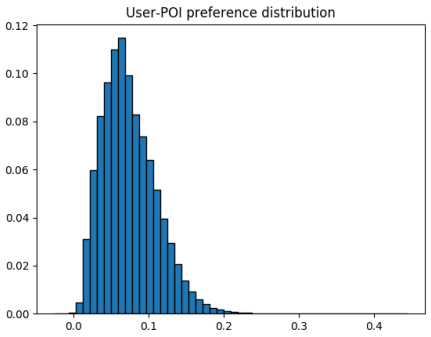
Αξίζει να σημειωθεί ότι από τους χρήστες που εμφανίζονται στις αρχικές διαδρομές, το σύνολο των δεδομένων περιέχει μόνο αυτούς που επισκέφθηκαν τουλάχιστον 10 POIs για να υπάρχει περισσότερη ακρίβεια στα προφίλ των χρηστών. Στον παρακάτω πίνακα φαίνονται κάποια συνολικά χαρακτηριστικά για το σύνολο δεδομένων.

Πόλη	**Αξιοθέατα**	**Χρήστες**
Ρώμη	671	1872
Φλωρεντία	1022	905
Πίζα	124	134

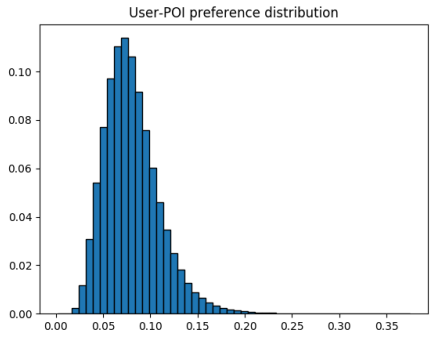
Πίνακας 4: Χαρακτηριστικά συνόλου δεδομένων

Στο Σχήμα 2 φαίνεται η κατανομή της ικανοποίησης ανά ζεύγος POI-χρήστη

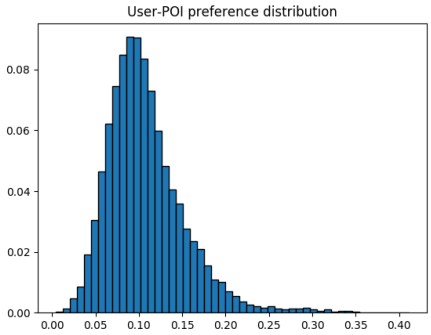
στις τρεις πόλεις. Παρατηρούμε ότι τα περισσότερα ζευγάρα παρουσιάζουν μία μέση προς χαμηλή ικανοποίηση ενώ υπάρχουν λίγα ζευγάρια στα οποία βλέπουμε σχετικά υψηλή ικανοποίηση σχηματίζοντας μια ουρά στην κατανομή. Αυτή η μορφή είναι αναμενόμενη καθώς ο κάθε χρήστης μπορεί να μείνει πλήρως ικανοποιημένος από μικρό αριθμό αξιοθέατων που ταιριάζουν στο προφίλ του ενώ τα υπόλοιπα του προσφέρουν μια μέση ικανοποίηση.



(α) Ρώμη



(β) Φλωρεντία

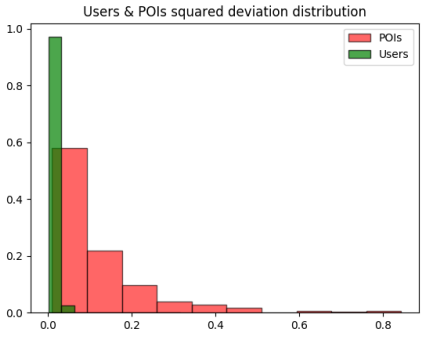


(γ) Πίζα

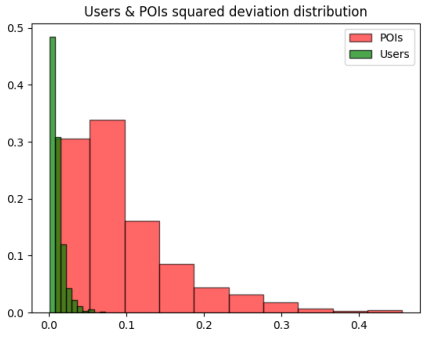
Σχήμα 2: Κατανομή της ικανοποίησης ανά ζεύγος POI-χρήστη

Κατά τη χρήση αλγορίθμων συσταδοποίησης μας ενδιαφέρει το κατά πόσο τα δεδομένα που επεξεργαζόμαστε παρουσιάζουν ποικιλομορφία ή ομοιογένεια. Στο Σχήμα 3 παρουσιάζεται η κατανομή της τετραγωνικής απόστασης των διανυσμάτων των χρηστών και των POIs από την αντίστοιχη μέση τιμή του συνόλου δεδομένων. Βλέπουμε ότι τα αξιοθέατα είναι συγκεντρωμένα γύρω από τη μέση τιμή αλλά με αρκετά μεγάλη απόκλιση, δηλαδή ότι έχουν σημαντικές διαφορές μεταξύ τους. Αντιθέτως, οι χρήστες είναι πολύ στενά συγκεντρωμένοι γύρω από τη μέση τιμή τους, δηλαδή τα προφίλ των προτιμήσεών τους δεν είναι και τόσο διαφορετικά.

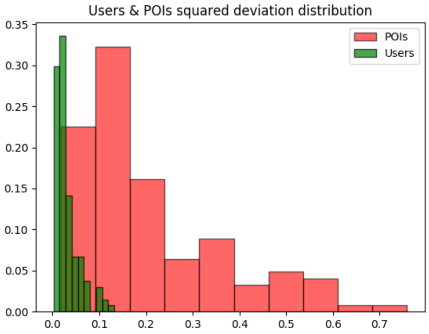
Προκειμένου να επιβεβαιώσουμε κάτι τέτοιο, επιπρόσθετα μπορούμε να κάνουμε



(α) Ρώμη



(β) Φλωρεντία



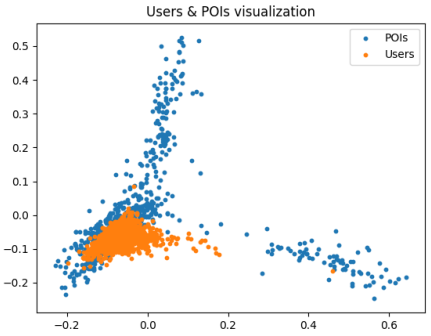
(γ) Πίζα

Σχήμα 3: Κατανομή της τετραγωνικής απόκλισης για POIs και χρήστες

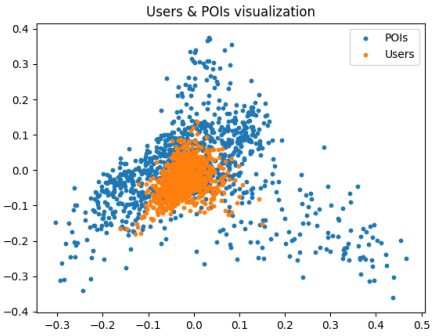
μία οπτικοποίηση των δεδομένων. Εφόσον τα δεδομένα είναι πολυδιάστατα δεν είναι δυνατή η άμεση απεικόνισή τους σε κάποιο δισδιάστατο σχήμα. Γι αυτό το λόγο, δημιουργούμε μία προβολή των δεδομένων σε ένα επίπεδο μέσω της μεθόδου *Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis – PCA)*. Η μέθοδος αυτή βρίσκει το επίπεδο στο οποίο η προβολή διατηρεί τη μέγιστη δυνατή πληροφορία για τα διανύσματα των POIs και χρησιμοποιείται για όλες τις δισδιάστατες απεικονίσεις στην παρούσα εργασία δίνοντας ικανοποιητικά αποτελέσματα. Η γραφική απεικόνιση χρηστών και POIs φαίνεται στο Σχήμα 4.

Μπορούμε να παρατηρήσουμε ότι στη Ρώμη και τη Φλωρεντία οι χρήστες είναι όντως πολύ συγκεντρωμένοι γύρω από τη μέση τιμή τους σε αντίθεση με τα POIs που είναι πιο ανοιγμένα στο χώρο. Στην Πίζα παρατηρούμε μεγαλύτερη ποικιλομορφία στην κατανομή των χρηστών.

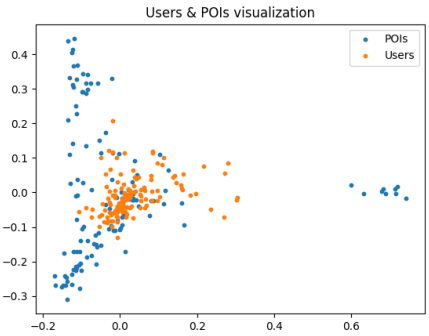
Καθώς οι χρήστες δεν έχουν σημαντική διαφοροποίηση, αναμένουμε το κέρδος που προσφέρει η συσταδοποίηση να είναι σχετικά μικρό για τη Ρώμη και τη Φλωρεντία ενώ να είναι μεγαλύτερο για την Πίζα.



(α) Ρώμη



(β) Φλωρεντία



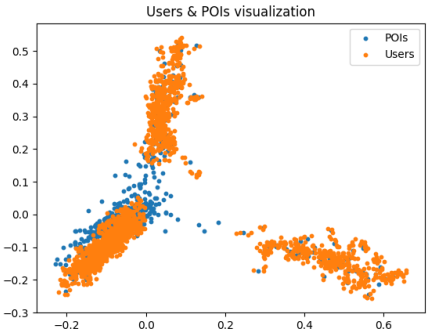
(γ) Πίζα

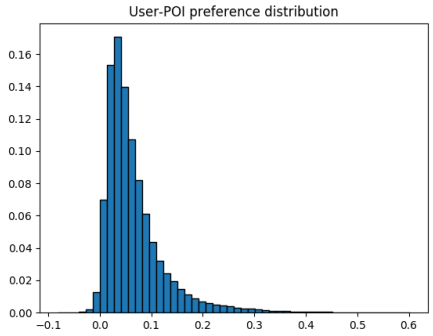
Σχήμα 4: Οπτικοποίηση

Προκειμένου να παρουσιαστεί σε πλήρη βαθμό η λειτουργικότητα των αλγορίθμων που εξετάζονται, απαιτείται ένα πιο ποικιλόμορφο σύνολο δεδομένων για τα προφίλ των χρηστών. Ένας τρόπος ώστε αυτό να γίνει εύκολα χωρίς επεξεργασία των αρχικών δεδομένων, είναι να δημιουργηθούν συνθετικά δεδομένα με βάση τα αρχικά ακολουθώντας την παρακάτω διαδικασία. Η δημιουργία των συνθετικών δεδομένων και η σύγκριση των αποτελεσμάτων γίνεται μόνο για την πόλη της Ρώμης.

Αρχικά, εκτελείται ο αλγόριθμος K-Μέσων για τα POIs της Ρώμης και αυτά χωρίζονται σε 5 εμφανείς κατηγορίες. Στη συνέχεια, επιλέγονται 3 από αυτές οι οποίες βρίσκονται μακριά η μία από την άλλη. Τέλος, για κάθε χρήστη της Ρώμης αντικαθίσταται το αρχικό του διάνυσμα με ένα τυχαίο διάνυσμα κάποιου POI από τις 3 κατηγορίες στο οποίο έχει προστεθεί μικρής κλίμακας θόρυβος. Τα χαρακτηριστικά αυτού του συνθετικού συνόλου δεδομένων φαίνονται στο Σχήμα 5.

Αξίζει να αναφερθεί ότι αυτή είναι μία ακραία περίπτωση όπου οι χρήστες είναι υπερβολικά διαχωρισμένοι, δηλαδή υπάρχουν χρήστες που ενδιαφέρονται αποκλειστικά για εκκλησίες, άλλοι που ενδιαφέρονται αποκλειστικά για μουσεία κ.λ.π.





(α) Οπτικοποίηση (Συνθετικά δεδομένα) (β) Οπτικοποίηση (Συνθετικά δεδομένα)

Τέτοιες περιπτώσεις δύσκολα παρατηρούνται στην πράξη και το συγκεκριμένο σύνολο δεδομένων χρησιμοποιείται μόνο για να υποστηρίξει καλύτερα τα επιχειρήματα γύρω από τις αρχικές υποθέσεις και δεν είναι αρκετό για να τις επιβεβαιώσει από μόνο του.

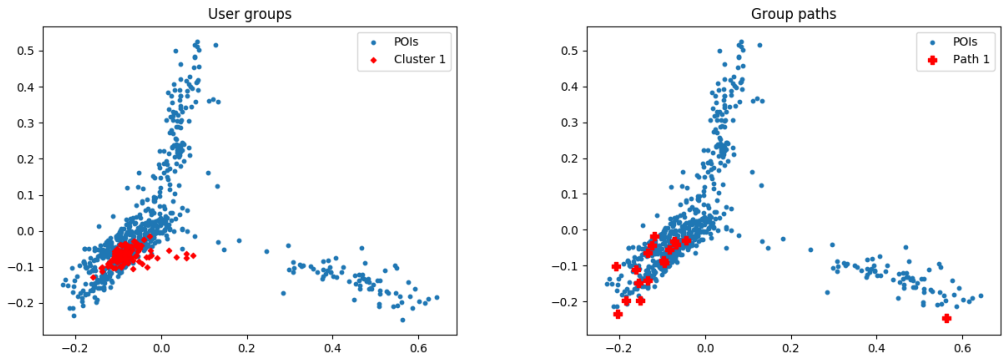
3.4 Πειραματική αξιολόγηση σε δεδομένα χρηστών

Σε αυτή την ενότητα μελετάται η απόδοση των αλγορίθμων που περιγράφηκαν στην Ενότητα 3.1. πάνω στο διαθέσιμο σύνολο δεδομένων. Η βασική ιδιότητα που εκμεταλλεύονται οι αλγόριθμοι είναι ότι μία ομάδα τουριστών τείνει να προτιμάει τα αξιοθέατα που βρίσκονται κοντά της στο διανυσματικό χώρο. Στο Σχήμα 6 δίνεται μία απεικόνιση ενός group 100 ατόμων καθώς και τα POIs της διαδρομής που ακολουθούν. Είναι εμφανές ότι τόσο οι τουρίστες όσο και το μονοπάτι βρίσκονται στην ίδια περιοχή του χώρου.

Αντίστοιχα, στο Σχήμα 7, μπορούμε να δούμε τη μεταβολή στα μονοπάτια αν οι ίδιοι 100 χρήστες χωριστούν σε 2 συστάδες μέσω του αλγορίθμου K-Μέσων. Με κόκκινο χρώμα απεικονίζεται η πρώτη συστάδα και με πράσινο η δεύτερη.

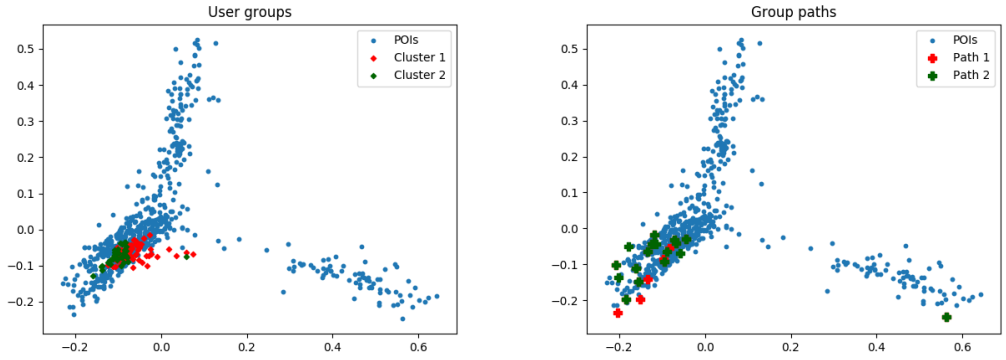
Από τα σχήματα αυτά μπορούμε να εξάγουμε δύο βασικά συμπεράσματα. Πρώτον, η συσταδοποίηση οδηγεί στη δημιουργία διαφορετικών μονοπατιών που ανταποκρίνονται καλύτερα στις προτιμήσεις των επιμέρους ομάδων χρηστών. Στο Σχήμα 6 βλέπουμε ότι τα POIs του μονοπατιού καλύπτουν ολόκληρη την κάτω αριστερή περιοχή του επιπέδου προκειμένου να καλύψουν τα ενδιαφέροντα όλου του group. Στο Σχήμα 7, το πράσινο group βρίσκεται πιο αριστερά στο επίπεδο από ό,τι το κόκκινο. Αντιστοίχως, το πράσινο μονοπάτι αποτελείται από POIs που βρίσκονται πιο αριστερά από αυτά του κόκκινου.

Η δεύτερη παρατήρηση είναι ότι τα δύο μονοπάτια έχουν μεγάλο αριθμό κοινών αξιοθέατων. Αυτό συμβαίνει διότι, παρ' ότι οι χρήστες έχουν χωριστεί σε δύο διαφορετικά groups, οι διαφορές τους είναι μικρές λόγω της μεγάλης ομοιογένειας των



(α) Χρήστες (β) Μονοπάτι

Σχήμα 6: Τουρίστες και μονοπάτια ως μία συστάδα

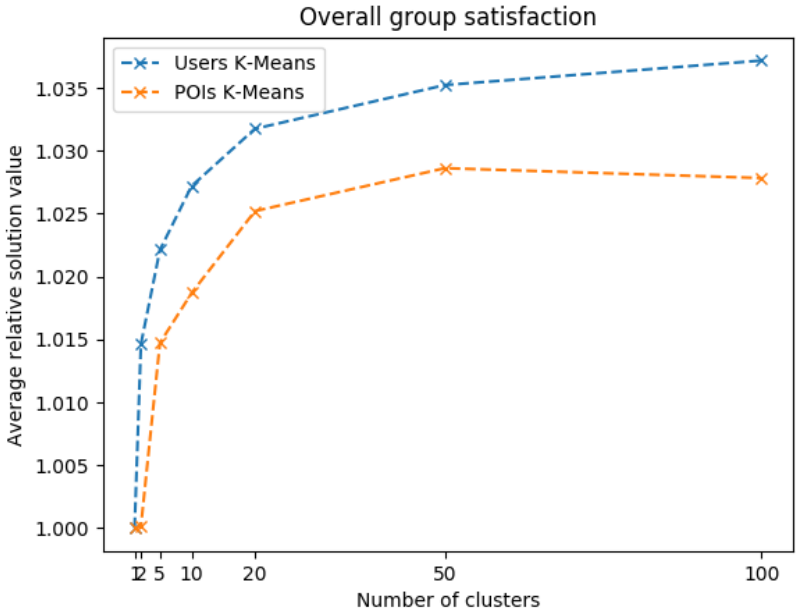


(α) Χρήστες (β) Μονοπάτια

Σχήμα 7: Τουρίστες και μονοπάτια ως δύο συστάδες

τουριστών. Έτσι, ο συμβιβασμός που κάνει το ένα group επισκεπτόμενο αξιοθέατα του άλλου είναι πολύ μικρός και δεν επηρεάζει σημαντικά τη λύση.

Για να γίνει αξιολόγηση των δύο αλγορίθμων που περιγράφηκαν στην Ενότητα 3.1. χρησιμοποιείται group 100 τουριστών το οποίο χωρίζεται σε αριθμό συστάδων ίσο με 1,2,5,10,20,50,100, για την κάθε συστάδα υπολογίζεται μία αποκλειστική διαδρομή και μελετάται η συνολική ικανοποίηση του group όσο αυξάνεται ο αριθμός των συστάδων. Γίνονται 500 επαναλήψεις του πειράματος όπου σε κάθε επανάληψη επιλέγονται τυχαία οι 100 τουρίστες που θα αποτελούν την ομάδα καθώς και τα σημεία s, t από όπου θα αρχίζουν και θα καταλήγουν αντίστοιχα τα μονοπάτια. Στο παρακάτω σχήμα φαίνεται η βελτίωση επί της αρχικής ικανοποίησης (μία μεγάλη συστάδα) καθώς μεταβάλλεται ο αριθμός των συστάδων.



Σχήμα 8: Συγκριτική απόδοση των αλγορίθμων συσταδοποίησης

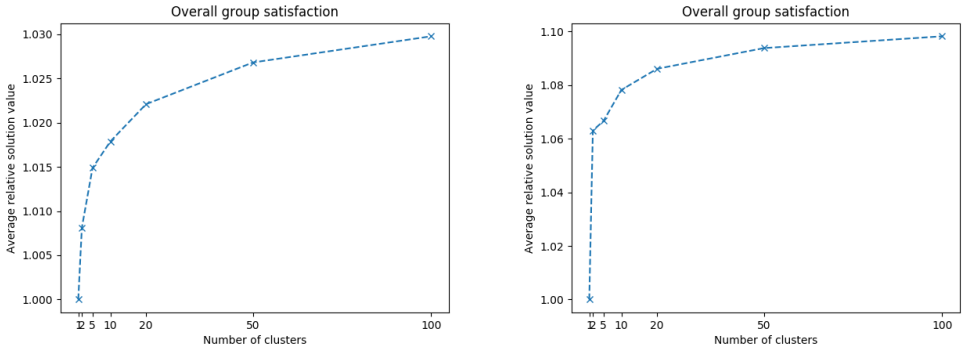
Όπως φαίνεται, η συνάρτηση ικανοποίησης-αριθμού συστάδων που προκύπτει και από τους δύο αλγορίθμους είναι αύξουσα και κοίλη. Με αυτό τον τρόπο επιβεβαιώνεται η υπόθεση ότι ο χωρισμός των χρηστών σε μικρότερα group αυξάνει τη συνολική ικανοποίηση. Από την άλλη, ο ρυθμός αυτής της αύξησης είναι πολύ μεγάλος για μικρό αριθμό συστάδων αλλά γρήγορα φθίνει καθώς οι συστάδες αυξάνονται αισθητά και πλησιάζουν το ολικό μέγεθος του group. Αφού το group αποτελείται από 100 άτομα στο πείραμα αυτό, ο διαχωρισμός τους σε 100 συστάδες αποτελεί το άνω φράγμα που μπορεί να επιτευχθεί καθώς αντιστοιχεί στη βέλτιστη περίπτωση που ο κάθε χρήστης ακολουθεί ένα προσωπικό μονοπάτι μόνος του χωρίς να υποβάλλεται σε συμβιβασμούς άλλων χρηστών.

Επίσης, μία άλλη παρατήρηση είναι ότι η διαφορά της ικανοποίησης μεταξύ των δύο ακραίων περιπτώσεων (1 συστάδα & 100 συστάδες) είναι της τάξης του 4%, δηλαδή πολύ μικρή. Όπως προαναφέρθηκε, αυτό ήταν αναμενόμενο καθώς τα προφίλ των χρηστών δεν έχουν μεγάλη ποικιλομορφία, οπότε οι συμβιβασμοί που επιβάλλονται λόγω της συνύπαρξης σε ένα μεγάλο group δεν είναι τόσο ισχυροί.

Όσον αφορά τη σύγκριση των δύο αλγορίθμων μπορούμε να παρατηρήσουμε ότι ο απλός διαχωρισμός των χρηστών με βάση τον αλγόριθμο K-Μέσων είναι αρκετά πιο αποδοτικός αφού η καμπύλη που παράγει είναι ψηλότερα στο σχήμα. Ο βασικός λόγος για αυτή τη διαφορά είναι ότι ο αλγόριθμος K-Μέσων για τα POIs λαμβάνει περισσότερο υπόψη τη σχέση του κάθε χρήστη με αυτά και όχι

τις σχέσεις μεταξύ των χρηστών κάνοντας έτσι μία έμμεση συσταδοποίηση που δεν φαίνεται να δίνει τα καλύτερα αποτελέσματα. Στη συνέχεια της εργασίας, όπου απαιτείται συσταδοποίηση θα χρησιμοποιείται αποκλειστικά ο αλγόριθμος K-Μέσων για τους χρήστες.

Αφού ο συγκεκριμένος αλγόριθμος φαίνεται να αποδίδει καλύτερα, στο Σχήμα 9 φαίνονται και τα αποτελέσματά του για τη Φλωρεντία και την Πίζα. Όπως είχε αναφερθεί και στην προηγούμενη ενότητα, το σύνολο δεδομένων της Φλωρεντίας παρουσιάζει αρκετή ομοιότητα με αυτό της Ρώμης, δηλαδή οι χρήστες είναι πολύ συγκεντρωμένοι γύρω από τη μέση τιμή τους ενώ οι χρήστες στην Πίζα έχουν λίγο μεγαλύτερη ποικιλομορφία. Αυτή η διαφοροποίηση μπορεί να παρατηρηθεί και στη συνάρτηση ικανοποίησης - αριθμού συστάδων για τις δύο πόλεις.



(α) Φλωρεντία (β) Πίζα

Σχήμα 9: Συνάρτηση ικανοποίησης - αριθμού συστάδων

Παρατηρούμε ότι η βελτίωση που προσφέρει η συσταδοποίηση στους χρήστες της Πίζας είναι ελαφρώς μεγαλύτερη (10%) σχετικά με αυτή που εμφανίζεται στη Ρώμη και στη Φλωρεντία (3%), αποτέλεσμα της ποικιλομορφίας των χρηστών για την πόλη αυτή. Η διαφοροποίηση αυτή γίνεται πιο εμφανής και αναλύεται περισσότερο με βάση το συνθετικό σύνολο δεδομένων στην επόμενη ενότητα.

Σε περίπτωση που το ενδιαφέρον περιορίζεται στην ποσοτικοποίηση της ικανοποίησης των τουριστών ανάλογα με τον αριθμό συστάδων, η παραπάνω ανάλυση επαρκεί. Παρ' όλα αυτά, όπως αναφέρθηκε και στην αρχή του κεφαλαίου, τα ίδια διαγράμματα μπορούν να αναγνωστούν και με εναλλακτικό τρόπο που προσφέρει χρήσιμη πληροφορία για τη διαχείριση της πληροφορίας στο συγκεκριμένο πρόβλημα. Αυτό μπορεί να γίνει εύκολα κατανοητό με το παρακάτω παράδειγμα.

Στην περίπτωση της Ρώμης, κάνοντας χρήση των δεδομένων αυτών, ένας διαχειριστής ταξιδιωτικού γραφείου μπορεί να συμπεράνει ότι χωρίζοντας 100 τουρίστες σε 10 διαφορετικά groups μπορεί να τους προσφέρει περίπου 3% περισσότερη ικανοποίηση. Από την άλλη, ο διαχειριστής ενός τέτοιου ηλεκτρονικού συστήματος

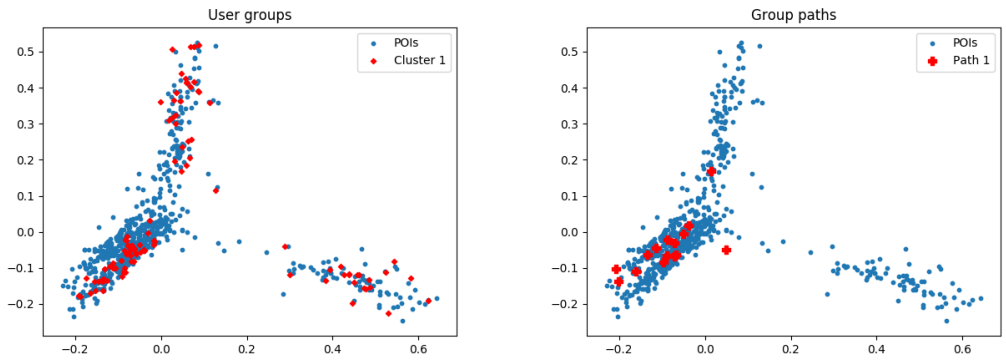
κατανοεί ότι χρησιμοποιώντας τα 10 κεντροειδή μπορεί να δημιουργήσει διαδρομές που θα προσφέρουν στους χρήστες 99% της ικανοποίησης που θα είχαν αν ακολουθούσαν διαδρομές σχεδιασμένες με βάση τα 100 προσωπικά τους διανύσματα. Έτσι, η μικρή απώλεια σε ικανοποίηση αντισταθμίζεται από τη μείωση του απαιτούμενου αποθηκευτικού χώρου και των αναγκαίων υπολογιστικών πόρων.

3.5 Πειραματική αξιολόγηση σε συνθετικά δεδομένα

Το συνθετικό σύνολο δεδομένων έχει κατασκευαστεί με τέτοιο τρόπο ώστε να υπάρχει υπερβολική διαφοροποίηση μεταξύ των χρηστών. Προκειμένου να πραγματοποιηθεί περαιτέρω μελέτη, αρχικά απαιτείται η επιβεβαίωση της βασικής υπόθεσης, ότι τουρίστες με διαφορετικά προφίλ προτιμήσεων ακολουθούν διαφορετικές διαδρομές που ταιριάζουν περισσότερο στα ενδιαφέροντά τους.

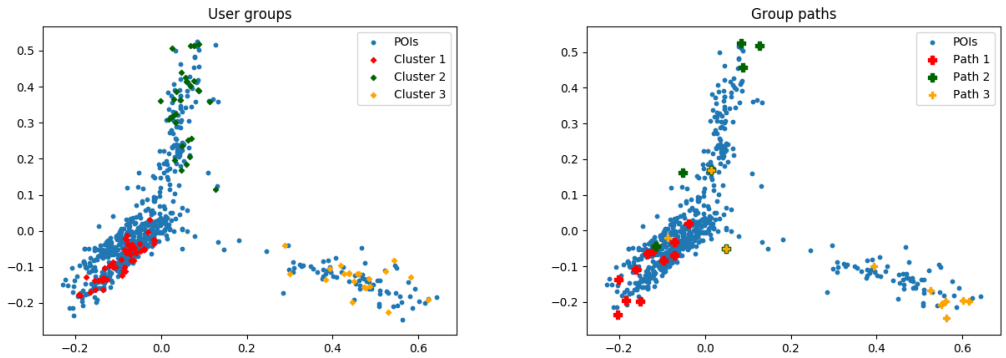
Έτσι, επιλέγεται μία τυχαία ομάδα 100 χρηστών και υπολογίζεται ένα μονοπάτι για αυτήν το οποίο φαίνεται στο Σχήμα 10. Καθώς τα μέλη της ομάδας βρίσκονται σε τρία διαφορετικά σημεία του επιπέδου, η μέση τιμή βρίσκεται στο κέντρο και παρατηρούμε ότι το μονοπάτι που επιλέγεται αποτελείται από POIs που βρίσκονται κοντά στο κέντρο.

Στη συνέχεια, οι ίδιοι 100 χρήστες διαχωρίζονται σε 3 συστάδες και η κάθε μία ακολουθεί τη διαδρομή που της ταιριάζει. Έτσι, στο Σχήμα 11 βλέπουμε ότι η κόκκινη ομάδα βρίσκεται κάτω αριστερά στο επίπεδο και ακολουθεί μία διαδρομή της οποίας τα POIs βρίσκονται και αυτά κάτω αριστερά. Αντιστοίχως, η πράσινη ομάδα και το μονοπάτι της καταλαμβάνουν το πάνω μέρος του επιπέδου ενώ η κίτρινη ομάδα το κάτω δεξιά. Η διαφορά που παρατηρούμε σε σχέση με το πραγματικό σύνολο δεδομένων είναι ότι εδώ η διαφοροποίηση μεταξύ των διαδρομών είναι πολύ μεγάλη.



(α) Χρήστες (β) Μονοπάτι

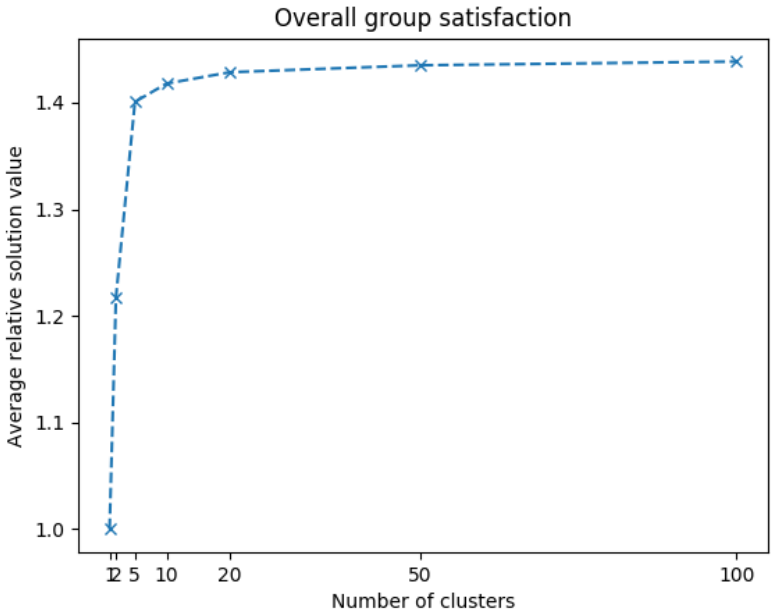
Σχήμα 10: Τουρίστες και μονοπάτια ως μία συστάδα



(α) Χρήστες (β) Μονοπάτια

Σχήμα 11: Τουρίστες και μονοπάτια ως τρεις συστάδες

Στη συνέχεια, εκτελείται ο αλγόριθμος K-Μέσων για χρήστες όπως περιγράφηκε στην Ενότητα 3.1. πάνω στο συνθετικό σύνολο δεδομένων. Οι διάφορες παράμετροι (π.χ. αριθμός χρηστών) και ο τρόπος εκτέλεσης (π.χ. αριθμός επαναλήψεων) είναι ακριβώς ίδια με τα αντίστοιχα που χρησιμοποιήθηκαν και στην προηγούμενη ενότητα. Βλέπουμε τα αποτελέσματά τους στο παρακάτω σχήμα.



Σχήμα 12: Συνάρτηση ικανοποίησης - αριθμού συστάδων (Συνθετικά δεδομένα)

Με βάση το διάγραμμα αυτό, οι παρατηρήσεις είναι παρόμοιες με αυτές της προηγούμενης ενότητας και έτσι ενισχύονται τα συμπεράσματα που αναφέρθηκαν. Είναι εμφανές ότι το άνω όριο της συνάρτησης είναι πολύ υψηλότερο από αυτό που εμφανιζόταν στην πόλη της Ρώμης κάνοντας χρήση του πραγματικού συνόλου δεδομένων. Παρατηρείται ότι το μέγιστο κέρδος στην ικανοποίηση που προκύπτει από το διαχωρισμό των 100 τουριστών του group σε συστάδες είναι πάνω από 40% ενώ στην προηγούμενη ενότητα το αντίστοιχο ποσοστό οριακά έφτανε το 4%. Η διαφοροποίηση αυτή οφείλεται στην ποικιλομορφία που εμφανίζουν οι συνθετικοί χρήστες ως προς τα προφίλ ενδιαφερόντων τους. Η συνύπαρξη πολύ διαφορετικών χρηστών στο ίδιο group οδηγεί σε ισχυρούς συμβιβασμούς στη σχεδίαση διαδρομής που έχουν μεγάλη επίπτωση στη συνολική ικανοποίηση.

Μία άλλη παρατήρηση που μπορεί να γίνει είναι ότι η συνάρτηση δεν είναι ιδιαίτερα ομαλή, δηλαδή φτάνει πολύ κοντά στη μέγιστη τιμή της ακόμα και για μικρό αριθμό συστάδων. Αυτό οφείλεται στο ότι οι συνθετικοί χρήστες απέχουν από τη μέση τιμή τους και αποτελούν καλώς διαχωρισμένες συστάδες, συγκεκριμένα 3 στο πλήθος. Έτσι, ο διαχωρισμός του group 100 ατόμων σε 3 ή παραπάνω συστάδες είναι αρκετός ώστε η συνολική ικανοποίηση να φτάσει πολύ κοντά στο άνω όριο. Με την εναλλακτική ερμηνεία που δόθηκε και προηγουμένως, αυτό σημαίνει ότι τα διανύσματα 3 κεντροειδών είναι αρκετά για το σχεδιασμό διαδρομών που ικανοποιούν σε μεγάλο βαθμό όλους τους χρήστες του συνθετικού συνόλου δεδομένων.

Εφόσον οι αρχικές υποθέσεις αυτού του κεφαλαίου έχουν επιβεβαιωθεί, οι βασικές ιδέες που αναπτύχθηκαν χρησιμοποιούνται στο Κεφάλαιο 4 ώστε να γίνει μια πλήρης περιγραφή ενός συστήματος σύστασης τουριστικών διαδρομών από πόλη σε πόλη.

4 Σύσταση διαδρομών από πόλη σε πόλη

4.1 Εισαγωγή

Το πρόβλημα του σχεδιασμού τουριστικών διαδρομών μελετάται από πολλούς ερευνητές λόγω της ενδιαφέρουσας αλγοριθμικής φύσης του. Στις περισσότερες προσεγγίσεις της βιβλιογραφίας, η μελέτη εστιάζεται στο κομμάτι της εύρεσης κατάλληλων ευριστικών τεχνικών για την επίλυση τέτοιων προβλημάτων σε αποδοτικό χρόνο λαμβάνοντας υπόψη τα αξιοθέατα μίας συγκεκριμένες πόλης και τις προτιμήσεις των χρηστών σε αυτά. Παρ' όλα αυτά, η δημιουργία αυτών των προφίλ προτιμήσεων δεν είναι καθόλου απλή υπόθεση. Όπως περιγράφηκε στο Κεφάλαιο 3, η κατασκευή ενός προφίλ εκ των υστέρων είναι μία σχετικά απλή διαδικασία και μας δίνει τη δυνατότητα να εξετάσουμε τις αλγοριθμικές πλευρές της σχεδίασης διαδρομών. Όμως, σε πραγματικές εφαρμογές, η δημιουργία προφίλ για έναν χρήστη αφού έχει ολοκληρώσει την επίσκεψή του στην εκάστοτε πόλη δεν έχει σχεδόν καμία ουσία καθώς δεν δίνει τη δυνατότητα να γίνει σύσταση διαδρομής που ο χρήστης μπορεί να ακολουθήσει. Έτσι, το αντικείμενο του κεφαλαίου αυτού εστιάζεται στον τρόπο εύρεσης του προφίλ του χρήστη. Συγκεκριμένα, γνωρίζοντας τα αξιοθέατα που επισκέπτεται σε μια πόλη Α, γίνεται μία εκτίμηση του προφίλ που θα έχει όταν επισκεφθεί μία πόλη Β.

Το βασικό εμπόδιο που δημιουργεί την αναγκαιότητα μελέτης του συγκεκριμένου προβλήματος είναι ότι το προφίλ του χρήστη όπως συντίθεται στην πόλη Α δεν είναι επαρκές προκειμένου να περιγράψει και τις προτιμήσεις του στην πόλη Β. Ο λόγος γι αυτή τη διαφοροποίηση είναι ότι οι πόλεις διαφέρουν μεταξύ τους ως προς την αναλογία, την ποιότητα και τις κατηγορίες των αξιοθέατων που περιλαμβάνει η κάθε μία. Για παράδειγμα, αν ένας χρήστης επισκεφθεί πολλά μουσεία στην πόλη Α, το σύστημα συστάσεων διαμορφώνει ένα προφίλ που έχει έντονη προτίμηση στα μουσεία. Αν ο χρήστης επισκεφθεί μία πόλη Β η οποία δεν περιλαμβάνει μουσεία, η παραγωγή συστάσεων με ακριβώς το ίδιο προφίλ ενδεχομένως να μην είναι αποτελεσματική.

Για τους λόγους αυτούς, προκειμένου να γίνει σύσταση διαδρομής σε ένα μεμονωμένο χρήστη στην πόλη Β, το σύστημα συστάσεων πρέπει να εκμεταλλευτεί πληροφορία και για τις παρελθοντικές προτιμήσεις του χρήστη στην πόλη Α αλλά και για τις συλλογικές προτιμήσεις των υπόλοιπων χρηστών στην πόλη Β. Αυτός ο συνδυασμός πληροφορίας μπορεί να γίνει με πολλούς τρόπους αλλά οι βασικές ιδέες στις οποίες στηρίζονται είναι δύο. Η παραδοχή που γίνεται και στις δύο περιπτώσεις είναι ότι τα αξιοθέατα και των δύο πόλεων είναι απεικονισμένα στον ίδιο διανυσματικό χώρο και αξιοθέατα με παρόμοια χαρακτηριστικά έχουν παραπλήσια διανύσματα ανεξαρτήτως της πόλης στην οποία βρίσκονται.

Η πρώτη ιδέα είναι η χρήση πληροφορίας για τα αξιοθέατα των διαδρομών που έχουν ήδη ακολουθήσει οι χρήστες στην πόλη Β. Δηλαδή, αν ένας χρήστης ακο-

λουθήσει μια διαδρομή στην πόλη Α και είναι γνωστό ένα σύνολο διαδρομών που έχουν ακολουθήσει χρήστες στην πόλη Β, μέσω κάποιας κατάλληλης διαδικασίας σύγκρισης μπορεί να επιλεγεί μία διαδρομή της οποίας τα αξιοθέατα "μοιάζουν" με αυτά που επισκέφθηκε ο χρήστης στην πόλη Α. Έτσι, το σύστημα συστάσεων παράγει εμμέσως ένα προφίλ για το χρήστη στην πόλη Β, βασιζόμενο σε πληροφορία για τα αξιοθέατα.

Η δεύτερη ιδέα είναι η άμεση δημιουργία νέου προφίλ για το χρήστη στην πόλη Β συνδυάζοντας το παρελθοντικό του προφίλ στην πόλη Α και τα προφίλ των χρηστών που έχουν ήδη επισκεφθεί την πόλη Β. Έτσι, το σύστημα συστάσεων παράγει ένα νέο προφίλ για το χρήστη χωρίς να λαμβάνει υπόψη του τις διαδρομές που ακολούθησαν οι χρήστες στην πόλη Β αλλά εκμεταλλευόμενο αποκλειστικά τα προφίλ τους.

Οι παραπάνω βασικές ιδέες χρησιμοποιούνται προκειμένου να σχεδιαστούν και να εξεταστούν κάποιοι αλγόριθμοι για την επίλυση του προβλήματος σύστασης διαδρομών από πόλη σε πόλη. Αν και η αρχική οπτική μελέτης του συγκεκριμένου προβλήματος είναι από την πλευρά των συστημάτων προσωποποιημένων συστάσεων, στην πορεία προκύπτουν και κάποια άλλα πιο ενδιαφέροντα ερωτήματα τα οποία γίνεται προσπάθεια να απαντηθούν στη συνέχεια του κεφαλαίου.

Ένας άλλος τρόπος να δει κανείς το πρόβλημα είναι από την πλευρά της μηχανικής μάθησης. Γνωρίζοντας τη συμπεριφορά ενός συνόλου χρηστών στην πόλη Β και την παρελθοντική συμπεριφορά ενός χρήστη στην πόλη Α, καλούμαστε να "προβλέψουμε" τη συμπεριφορά αυτού όταν επισκεφτεί την πόλη Β. Οι αλγόριθμοι που χρησιμοποιούνται στη συνέχεια του κεφαλαίου για τη σύσταση διαδρομών έχουν δυϊκή ερμηνεία καθώς ταυτόχρονα αποτελούν και μέσο πρόβλεψης για το προφίλ του χρήστη όταν αυτός επισκεφθεί την πόλη Β. Αυτή η οπτική του προβλήματος γεννά κάποια πολύ ενδιαφέροντα ερωτήματα:

- Ο αλγόριθμος σύστασης που δίνει τη διαδρομή με τη μεγαλύτερη ικανοποίηση για το χρήστη δίνει και την καλύτερη εκτίμηση για το πραγματικό του προφίλ;
- Από τι εξαρτάται η αποδοτικότητα των αλγορίθμων ως προς την ικανοποίηση που προσφέρει η διαδρομή που συστήνουν;
- Από τι εξαρτάται η αποδοτικότητα των αλγορίθμων ως προς την εκτίμηση του προφίλ του χρήστη;

Όπως αναφέρθηκε και προηγουμένως, ο κύριος λόγος για τον οποίο μελετάται το συγκεκριμένο πρόβλημα είναι ότι το προφίλ του χρήστη αλλάζει από πόλη σε πόλη ανάλογα με τα αξιοθέατα που η κάθε μία περιέχει. Σε μια προσπάθεια γενίκευσης, ένα τελευταίο ερώτημα που προκύπτει είναι το εάν υπάρχει κάποια συνάρτηση που δεδομένου του προφίλ ενός χρήστη στην πόλη Α και των αξιοθέατων των πόλεων

Α και Β δίνει το νέο προφίλ του ίδιου χρήστη στην πόλη Β. Με άλλα λόγια, τα θέματα που συνοδεύουν το παραπάνω ερώτημα είναι τα εξής:

- Ποιοι είναι οι παράγοντες που οδηγούν έναν χρήστη να αλλάξει τις προτιμήσεις του από πόλη σε πόλη;
- Η συνάρτηση μετασχηματισμού του προφίλ από πόλη σε πόλη έχει κάποια μορφή που ερμηνεύεται διαισθητικά ή ο μετασχηματισμός είναι ουσιαστικά τυχαίος;
- Υπάρχει τρόπος να μάθουμε αυτή τη συνάρτηση;

Κάποια από τα παραπάνω ερωτήματα απαντώνται στις επόμενες ενότητες ενώ για τα υπόλοιπα τίθενται οι απαραίτητες βάσεις για την περαιτέρω μελέτη τους στο μέλλον.

4.2 Αλγόριθμοι σύστασης διαδρομών

Σε αυτή την ενότητα περιγράφονται οι αλγόριθμοι που χρησιμοποιούνται για τη σχεδίαση διαδρομής για ένα χρήστη όταν αυτός επισκεφθεί μια πόλη Β γνωρίζοντας τη διαδρομή που ακολούθησε στην πόλη Α καθώς και τις διαδρομές που έχουν ακολουθήσει άλλοι χρήστες στην πόλη Β. Καθώς αυτοί οι αλγόριθμοι βασίζονται σε ευριστικές τεχνικές και η λειτουργία τους δεν μπορεί να αξιολογηθεί αυστηρά, παρατίθενται και κάποιοι απλοϊκοί αλγόριθμοι που χρησιμοποιούνται ως μέτρο σύγκρισης (baseline).

Το σύνολο δεδομένων που χρησιμοποιείται είναι το ίδιο με αυτό που περιγράφηκε στο Κεφάλαιο 3. Μία σειρά τροποποιήσεων που γίνονται ώστε αυτό να είναι κατάλληλα διαμορφωμένο για το συγκεκριμένο πρόβλημα παρουσιάζονται στην επόμενη ενότητα. Σε αυτό το σημείο, για την καλύτερη κατανόηση των αλγορίθμων, παρατίθεται συγκεκριμένη ορολογία για την περιγραφή των χρηστών και των πόλεων που θα ακολουθηθεί πλήρως και στη συνέχεια.

Για κάθε πόλη γνωρίζουμε ένα σύνολο χρηστών που την έχουν επισκεφθεί καθώς και τα αξιοθέατα που επέλεξαν σε αυτή. Καθώς το ενδιαφέρον επικεντρώνεται στην περίπτωση που ένας χρήστης επισκέπτεται μια πόλη και με βάση τη συμπεριφορά του εκεί γίνεται σύσταση διαδρομής σε μια δεύτερη, η πρώτη πόλη στο εξής θα ονομάζεται *SourceCity* και η δεύτερη *TargetCity*. Οι χρήστες για τους οποίους είναι διαθέσιμα δεδομένα επίσκεψης σε δύο πόλεις μας δίνουν τη δυνατότητα να τους χρησιμοποιήσουμε για την αξιολόγηση των αλγορίθμων και θα αναφέρονται ως *TestUsers*. Αντιθέτως, οι χρήστες που έχουν επισκεφθεί μόνο μία πόλη μπορούν να χρησιμοποιηθούν μόνο ως προϋπάρχουσα γνώση και συνεπώς θα ονομάζονται *TrainUsers*. Η πληροφορία που είναι γνωστή είναι τα προφίλ των TrainUsers στην TargetCity καθώς και τα προφίλ των TestUsers στην SourceCity. Η πληροφορία

που ουσιαστικά είναι ζητούμενη και γι αυτό το λόγο αποκρύπτεται είναι τα προφίλ των TestUsers στην TargetCity.

Οι αλγόριθμοι που μελετώνται περιγράφονται παρακάτω χωρίς να δίνονται πολλές λεπτομέρειες καθώς αποτελούν παραλλαγές του αρχικού αλγορίθμου *BestRatio+*. Η μόνη διαφοροποίησή τους είναι στον τρόπο με τον οποίο καθορίζεται η αξία κάθε αξιοθέατου για τον εκάστοτε χρήστη.

- ***PopularPath*** Η βασική του ιδέα είναι αυτή που χρησιμοποιείται πολλές φορές στην πράξη είτε από μεμονωμένους τουρίστες είτε από ταξιδιωτικά γραφεία. Δεν λαμβάνονται καθόλου υπόψη οι προσωπικές προτιμήσεις του κάθε χρήστη και το κάθε αξιοθέατο της πόλης αξιολογείται μόνο από το πόσο διάσημο είναι. Χρησιμοποιώντας τα διαθέσιμα δεδομένα, για κάθε αξιοθέατο της TargetCity, ως διασημότητα (popularity) ορίζεται ο αριθμός των τουριστών από το σύνολο των TrainUsers που το έχουν επισκεφθεί.

  Έτσι, όταν απαιτείται η σχεδίαση μίας διαδρομής για έναν χρήστη κάνοντας χρήση του συγκεκριμένου αλγορίθμου, ουσιαστικά εκτελείται ο *BestRatio+* με τη διαφορά ότι στον υπολογισμό του λόγου ικανοποίησης προς κόστος, αντί για βαθμό ικανοποίησης από κάθε αξιοθέατο χρησιμοποιείται ο βαθμός του αντίστοιχου popularity.

Ο προηγούμενος αλγόριθμος δεν χρησιμοποιεί καθόλου πληροφορία σχετικά με τα προφίλ των χρηστών οπότε είναι λογική η υπόθεση ότι πρόκειται για το πιο απλοϊκό μέτρο σύγκρισης. Οι επόμενοι 3 αλγόριθμοι χρησιμοποιούν στοιχεία από τις διαδρομές που έχουν πραγματοποιήσει προηγουμένως οι TrainUsers στην TargetCity καθώς και πληροφορία για τα αξιοθέατα που έχουν επισκεφθεί οι TestUsers στην SourceCity.

Σε αυτό το σημείο γίνεται χρήση των αποτελεσμάτων που παρουσιάστηκαν στο Κεφάλαιο 3. Όπως είναι λογικό, οι TrainUsers μπορεί να είναι υπερβολικά πολλοί σε πλήθος με αποτέλεσμα οι αλγόριθμοι που εκμεταλλεύονται στοιχεία για τις διαδρομές τους στην TargetCity να απαιτούν πολλούς υπολογιστικούς πόρους για την εκτέλεσή τους. Όμως, διαπιστώθηκε ότι οι προτιμήσεις των χρηστών παρουσιάζουν μεγάλη ομοιογένεια και ότι ένας μικρός αριθμός κεντροειδών αρκεί για το σχεδιασμό διαδρομών που ικανοποιούν όλους τους χρήστες σε πολύ μεγάλο βαθμό.

Έτσι, οι παρακάτω αλγόριθμοι δεν χρησιμοποιούν άμεσα πληροφορία για τις διαδρομές των TrainUsers αλλά για τις διαδρομές που θα ακολουθούσαν αν ήταν χωρισμένοι σε μικρό αριθμό από groups. Δηλαδή, πρωταρχικά, στην TargetCity γίνεται συσταδοποίηση K-Μέσων των TrainUsers σε μικρό αριθμό συστάδων και τα μόνα προφίλ που χρειάζεται να είναι γνωστά είναι αυτά των αντίστοιχων κεντροειδών.

Οι αλγόριθμοι σύστασης διαδρομών που βασίζονται σε αυτές τις ιδέες είναι οι εξής:

- ***RandomCluster*** Πρόκειται για άλλο ένα baseline αλγόριθμο που δεν χρησιμοποιεί πληροφορία για το προφίλ του TestUser στην SourceCity. Το πλεονέκτημά του σε σχέση με τον PopularPath είναι ότι χρησιμοποιεί τα κεντροειδή τα οποία προκύπτουν από τη συσταδοποίηση των προφίλ των TrainUsers στην TargetCity. Συγκεκριμένα, επιλέγεται τυχαία ένα από τα διαθέσιμα κεντροειδή και γίνεται η υπόθεση ότι το διάνυσμα αυτό αποτελεί το προφίλ του TestUser στην TargetCity. Στη συνέχεια, υπολογίζεται μία διαδρομή με βάση το συγκεκριμένο διάνυσμα. Ο αλγόριθμος αυτός είναι πλήρως τυχαιοκρατικός χωρίς να έχει προοπτικές επιτυχίας και γι αυτό χρησιμοποιείται ως baseline.

- ***MostSimilarCluster*** Η ιδέα που χρησιμοποιεί ο αλγόριθμος αυτός είναι η συσχέτιση διαδρομών. Έστω ότι για κάθε group (συστάδα) των TrainUsers στην TargetCity υπάρχει μία διαδρομή υπολογισμένη με βάση το κεντροειδές της. Τότε, γνωρίζοντας τη διαδρομή που ακολούθησε ένας TestUser στην SourceCity δίνεται η δυνατότητα να βρεθεί κάποιου είδους ομοιότητα μεταξύ αυτής και των προϋπολογισμένων διαδρομών στην TargetCity. Είναι λογικό να γίνει η υπόθεση ότι ο TestUser στην TargetCity θα έχει προφίλ παρόμοιο με τους TrainUsers που ακολουθούν παρόμοια διαδρομή με αυτή που ακολούθησε ο ίδιος στη SourceCity.

  Προκειμένου να γίνει αυτό, απαιτείται η επιλογή μιας μετρικής η οποία θα εκφράζει την ομοιότητα μεταξύ ενός συνόλου αξιοθέατων στην SourceCity με ένα σύνολο αξιοθέατων στην TargetCity. Έστω S, T τα σύνολα των αξιοθέατων δύο διαδρομών στις δύο πόλεις αντίστοιχα. Έστω ότι ένα POI $v_i \in S$ αναπαρίσταται με το διάνυσμα $\mathbf{s}_i$ και ένα POI $v_j \in T$ αναπαρίσταται με το διάνυσμα $\mathbf{t}_j$. Τότε, ορίζουμε ως ομοιότητα των δύο διαδρομών την εξής ποσότητα:

$$pathSim(S, T) = \sum_{\mathbf{t}_j \in T} \max_{\mathbf{s}_i \in S} \{\mathbf{s}_i \cdot \mathbf{t}_j\}$$

  Διαισθητικά, αυτή η μετρική αντιστοιχεί κάθε POI του συνόλου T σε ένα "παρόμοιο" POI του συνόλου S. Δύο POIs θεωρούνται παρόμοια όταν το εσωτερικό τους γινόμενο έχει μεγάλη τιμή. Η υπόθεση που γίνεται είναι ότι ο χρήστης που επιλέγει το σύνολο αξιοθέατων S στην SourceCity μένει ικανοποιημένος από αυτά και τα αξιοθέατα στην TargetCity που θα μεγιστοποιούν την ικανοποίησή του θα έχουν κάποιο αξιοθέατο που θα τους μοιάζει στο σύνολο S.

  Έτσι, ο αλγόριθμος MostSimilarCluster υπολογίζει διαδρομές για τις συστάδες της TargetCity και συστήνει στον TestUser αυτή που μεγιστοποιεί

την παραπάνω μετρική. Αντιστοίχως, γίνεται η εκτίμηση ότι το προφίλ προτιμήσεων του χρήστη αυτού ταυτίζεται με το διάνυσμα του κεντροειδούς της αντίστοιχης συστάδας. Αναλυτικά, αν T_k είναι το σύνολο POIs της διαδρομής της συστάδας k και $\mathbf{c}_k$ είναι το αντίστοιχο κεντροειδές, η εκτίμηση για το προφίλ του TestUser είναι:

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{c}_k} pathSim(S, T_k)$$

- ***WeightedCluster*** Ο αλγόριθμος αυτός μοιάζει σε μεγάλο βαθμό με τον προηγούμενο. Σχεδιάζει διαδρομές για όλες τις συστάδες των TrainUsers στην TargetCity και υπολογίζει τη μετρική pathSim μεταξύ κάθε μίας από αυτές τις διαδρομές και τη διαδρομή που ακολούθησε ο TestUser στην SourceCity. Στη συνέχεια, κάνει την εκτίμηση ότι το προφίλ του χρήστη στην TargetCity είναι ένα σταθμισμένο άθροισμα των κεντροειδών των συστάδων των TrainUsers με βάρη τις αντίστοιχες τιμές της μετρικής pathSim. Τέλος, συστήνει στον TestUser ένα μονοπάτι σχεδιασμένο με βάση το διάνυσμα που προέκυψε από την παραπάνω διαδικασία. Αναλυτικά, έστω S το σύνολο POIs που επισκέφθηκε ο TestUser στη SourceCity και T_k το σύνολο των POIs της διαδρομής της συστάδας k στην TargetCity. Αν $\mathbf{c}_k$ είναι το διάνυσμα του κεντροειδούς της συστάδας k, το εκτιμώμενο προφίλ του TestUser είναι:

$$\hat{\mathbf{u}} = \frac{\sum_k \mathbf{c}_k \cdot pathSim(S, T_k)}{\sum_k pathSim(S, T_k)}$$

  Η βασική διαφοροποίηση αυτού του αλγορίθμου από τον MostSimilarCluster είναι ότι το προφίλ του χρήστη δεν αντιστοιχίζεται πλήρως σε μία και μοναδική συστάδα αλλά λαμβάνονται υπόψη όλες οι συστάδες ανάλογα με το βαθμό ομοιότητας των διαδρομών τους με αυτή του TestUser στην SourceCity.

Στη συνέχεια, παρουσιάζονται δύο αλγόριθμοι οι οποίοι δεν χρησιμοποιούν διαδρομές για τη διαδικασία της εκτίμησης αλλά λαμβάνουν υπόψη τους απευθείας τα προφίλ των χρηστών.

- ***GlobalCentroid*** Ο αλγόριθμος αυτός αγνοεί πλήρως τις προσωπικές προτιμήσεις του TestUser και του συστήνει μια διαδρομή της TargetCity που ακολουθεί ο μέσος χρήστης που την επισκέπτεται. Προκειμένου να γίνει αυτός ο υπολογισμός, γίνεται η εκτίμηση ότι το προφίλ του TestUser ταυτίζεται με τη μέση τιμή των διανυσμάτων των TrainUsers στην TargetCity, ουσιαστικά το κεντροειδές που θα προέκυπτε αν οι χρήστες αυτοί συσταδοποιούνταν σε μία και μοναδική συστάδα. Αναλυτικά, αν οι TrainUsers στην TargetCity έχουν διανύσματα $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_l$, η εκτίμηση για το διάνυσμα του TestUser είναι:

$$\hat{\mathbf{u}} = \frac{1}{l}\sum_{i=1}^{l}\mathbf{u}_i$$

Ο αλγόριθμος αυτός είναι αρκετά απλοϊκός αφού πρακτικά θεωρεί ότι όλοι οι TestUsers είναι μεταξύ τους ίδιοι. Η λογική του είναι παρόμοια με αυτή του PopularPath με τη διαφορά ότι αντί για διασημότητα χαρακτηρίζει το κάθε αξιοθέατο με τη μέση προτίμηση που έχουν οι χρήστες για αυτό. Πρόκειται για ειδική περίπτωση του αλγόριθμου MixedProfile που περιγράφεται στη συνέχεια.

- ***MixedProfile*** Ο αλγόριθμος αυτός λαμβάνει υπόψη του δύο στοιχεία. Το προφίλ του TestUser στην SourceCity όπως αυτό δημιουργήθηκε από την προηγούμενη επίσκεψή του σε αυτή καθώς και το μέσο προφίλ των TrainUsers στην TargetCity, το ίδιο που χρησιμοποιείται και στον GlobalCentroid. Η πρώτη υπόθεση που γίνεται είναι ότι ο TestUser δεν θα αποκλίνει πολύ από το μέσο προφίλ των υπολοίπων χρηστών στην TargetCity, κάτι που έχει βάση λόγω της μεγάλης ομοιογένειας που παρατηρήθηκε στα προφίλ των χρηστών στο Κεφάλαιο 3. Η δεύτερη υπόθεση είναι ότι η όποια διαφοροποίηση από το μέσο προφίλ θα είναι στην ίδια κατεύθυνση με το προφίλ του χρήστη στην SourceCity. Διαισθητικά, αυτό σημαίνει ότι ο TestUser συμπεριφέρεται στην TargetCity με παρόμοιο τρόπο με τους υπόλοιπους έχοντας μια επιπλέον προτίμηση σε αξιοθέατα παρόμοια με αυτά που προτίμησε και στη SourceCity. Αναλυτικά, αν $\mathbf{u}_0$ είναι το προφίλ του TestUser στην SourceCity και $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_l$ τα προφίλ των TrainUsers στην TargetCity, η εκτίμηση για το προφίλ του TestUser στην TargetCity είναι:

  $$\hat{\mathbf{u}} = (1-\alpha)\frac{1}{l}\sum_{i=1}^{l}\mathbf{u}_i + \alpha\mathbf{u}_0$$

  Η παράμετρος α παίρνει τιμές στο διάστημα $[0, 1]$ και αντιστοιχεί στη σημασία των προηγούμενων προσωπικών προτιμήσεων έναντι των γενικών προτιμήσεων των υπόλοιπων τουριστών. Είναι εύκολο να δούμε ότι για $\alpha = 1$ η εκτίμηση για το προφίλ του χρήστη είναι ακριβώς το ίδιο προφίλ που είχε στην SourceCity δηλαδή δεν λαμβάνεται καθόλου υπόψη η διαφοροποίηση των ενδιαφερόντων από πόλη σε πόλη. Από την άλλη, στην περίπτωση που ισχύει $\alpha = 0$ ο αλγόριθμος είναι ίδιος με τον GlobalCentroid.

Προκειμένου να αξιολογηθούν αυτοί οι αλγόριθμοι καθώς και να απαντηθούν τα ερωτήματα που τέθηκαν στην αρχή του κεφαλαίου, στη συνέχεια γίνεται παρουσίαση του συνόλου δεδομένων που χρησιμοποιήθηκε κατά τη διαδικασία των πειραμάτων.

4.3 Μέθοδος δημιουργίας συνόλου δεδομένων

Προκειμένου να υπάρχει ένα μέτρο σύγκρισης, το σύνολο δεδομένων που χρησιμοποιήθηκε για την πειραματική αξιολόγηση των αλγορίθμων είναι κατά βάση το ίδιο με αυτό που παρουσιάστηκε στο Κεφάλαιο 3. Παρ' όλα αυτά, στην αρχική του μορφή δεν ήταν κατάλληλο για τη μελέτη του προβλήματος σύστασης διαδρομών από πόλη σε πόλη και γι αυτό το λόγο ανακατασκευάστηκε με παρόμοια διαδικασία με αυτή που προτείνουν οι Anagnostopoulos et al. [34] για την κατασκευή του αρχικού συνόλου δεδομένων όπως χρησιμοποιήθηκε στο Κεφάλαιο 3.

Το βασικό πρόβλημα που παρουσίαζε στην αρχική του μορφή ήταν ότι η διαδικασία παραγωγής διανυσμάτων για τα αξιοθέατα της κάθε πόλης με χρήση των κειμένων των αντίστοιχων σελίδων της Wikipedia είχε πραγματοποιηθεί ανεξάρτητα για κάθε πόλη. Έτσι, δεν υπήρχε τρόπος τα POIs και των τριών πόλεων να απεικονιστούν σε έναν κοινό διανυσματικό χώρο. Βέβαια, αυτό είναι πλήρως αναγκαίο κατά τη σύσταση διαδρομών από πόλη σε πόλη καθώς τόσο η μετρική ομοιότητας διαδρομών pathSim όσο και η συσχέτιση προφίλ των χρηστών όπως στην περίπτωση του αλγορίθμου MixedProfile δεν θα είχε κανένα απολύτως νόημα.

Ένα άλλο πρόβλημα που παρουσίαζε το προηγούμενο σύνολο δεδομένων είναι ότι οι χρήστες σε κάθε πόλη δεν συνοδεύονταν από κάποιο αναγνωριστικό (ID) και έτσι δεν ήταν δυνατό να βρεθούν ποιοι χρήστες έχουν επισκεφθεί μόνο μία πόλη ώστε να χρησιμοποιηθούν ως TrainUsers και ποιοι έχουν επισκεφθεί δύο ώστε να πάρουν το ρόλο των TestUsers. Ο συνδυασμός αυτών των δύο θεμάτων έκανε το προηγούμενο σύνολο δεδομένων προβληματικό κρίνοντας αναγκαία την ανακατασκευή του με τη διαδικασία που περιγράφεται παρακάτω.

Για τη δημιουργία προφίλ για τους χρήστες χρησιμοποιούνται οι διαδρομές τουριστών των Baraglia et al. [58] που αναφέρθηκαν και στο Κεφάλαιο 3. Για κάθε πόλη από τις Ρώμη, Φλωρεντία, Πίζα δίνεται ένα σύνολο χρηστών καθώς και οι διαδρομές που ακολούθησαν. Για κάθε POI σε αυτές τις διαδρομές δίνεται ο τίτλος της αντίστοιχης σελίδας στη Wikipedia και αντλείται το κείμενο της σελίδας μέσω του διαθέσιμου API.

Έχοντας το σύνολο των κειμένων για τα αξιοθέατα των τριών πόλεων, τα διανύσματα για το καθένα δημιουργούνται μέσω της μεθόδου Λανθάνουσας Σημασιολογικής Δεικτοδότησης (LSI). Ο σκοπός της μεθόδου είναι να αναλύσει τα κείμενα, να εξάγει ένα σύνολο σημασιολογικών εννοιών και να απεικονίσει το κάθε κείμενο με ένα διάνυσμα. Το κάθε στοιχείο του διανύσματος αντιστοιχεί στο βαθμό που το κάθε κείμενο αφορά τη συγκεκριμένη σημασιολογική έννοια. Στη συνέχεια, αναλύονται τα διάφορα στάδια της διαδικασίας αυτής.

Το πρώτο βήμα είναι η προεπεξεργασία των κειμένων. Από κάθε κείμενο αφαιρούνται οι συχνές συνδετικές λέξεις (π.χ. που, και) καθώς αυτές δεν προσφέρουν κανένα σημασιολογικό νόημα. Επίσης, αφαιρούνται όλες οι λέξεις με μήκος μικρότερο από 3 καθώς και αυτές πιθανότατα αποτελούν λέξεις μικρής σημασίας. Τέλος,

γίνεται λημματοποίηση και αποκοπή καταλήξεων στις λέξεις που περιλαμβάνει το κάθε κείμενο. Προκειμένου να εντοπιστεί μία έννοια είναι απαραίτητο λέξεις της ίδιας οικογένειας να αναχθούν στην κοινή τους ρίζα. Π.χ. οι λέξεις "εκκλησία" και "εκκλησιαστικός" είναι επιθυμητό να μετατραπούν στην ρίζα "εκκλησ" καθώς αυτό το τμήμα των λέξεων εκφράζει το κοινό τους νόημα. Στο τέλος αυτού του βήματος, το κάθε POI είναι ένα επεξεργασμένο κείμενο, δηλαδή μία λίστα από αποκομμένες λέξεις.

Το επόμενο βήμα είναι η δημιουργία ενός λεξικού από τις λέξεις που περιλαμβάνονται στα επεξεργασμένα κείμενα. Δηλαδή, για κάθε λέξη που υπάρχει μέσα στα κείμενα δημιουργείται ένας αριθμός-κλειδί και σε κάθε κείμενο μετράται πόσες φορές εμφανίζεται η συγκεκριμένη λέξη. Από το λεξικό αφαιρούνται πολύ συχνές λέξεις που εμφανίζονται στο 90% των κειμένων και πολύ σπάνιες που εμφανίζονται σε λιγότερο από 10 κείμενα. Έτσι, στο τέλος αυτού του βήματος, το κάθε POI είναι μία λίστα από ζεύγη (κλειδί, αριθμός εμφανίσεων) που αντιπροσωπεύουν τον αριθμό των εμφανίσεων του κάθε όρου μέσα στο εκάστοτε κείμενο.

Το τρίτο βήμα της διαδικασίας είναι η κατασκευή ενός δισδιάστατου πίνακα που να περιλαμβάνει τιμές που εκφράζουν τη σημασία του κάθε όρου για κάθε κείμενο. Αυτή η τιμή θα μπορούσε να είναι απλά το πλήθος των αντίστοιχων εμφανίσεων αλλά μία άλλη τιμή που χρησιμοποιείται ευρέως είναι η *Συχνότητα Όρου - Αντίστροφη Συχνότητα Εγγράφου (Term Frequency - Inverse Document Frequency – TF-IDF)*. Έστω ότι τα διαθέσιμα κείμενα είναι τα $D = \{d_1, d_2, ..., d_N\}$ και οι όροι που περιλαμβάνονται σε αυτά είναι οι $T = \{t_1, t_2, ..., t_M\}$. Ως $TF(t_i, d_j)$ συμβολίζεται η συχνότητα με την οποία εμφανίζεται ο όρος t_i στο κείμενο d_j δηλαδή ο λόγος του πλήθους εμφανίσεών του στο συγκεκριμένο κείμενο προς το συνολικό αριθμό των όρων σε αυτό. Επίσης, ως n_i συμβολίζεται το πλήθος των κειμένων στα οποία εμφανίζεται τουλάχιστον μία φορά ο όρος t_i. Σύμφωνα με τα παραπάνω, ο τύπος με τον οποίο ορίζεται η τιμή TF-IDF είναι ο εξής:

$$TF - IDF(t_i, d_j) = TF(t_k, d_j) \cdot log\frac{N}{n_i}$$

Διαισθητικά, μπορούμε να δούμε ότι μία τιμή TF-IDF είναι μεγάλη όταν ένας όρος εμφανίζεται πολλές φορές μέσα σε ένα κείμενο και σχετικά λίγες φορές στο σύνολο των κειμένων. Άρα αυτό το μέγεθος είναι κατάλληλο για να εκφράσει τη σημασία ενός όρου για ένα συγκεκριμένο POI. Έτσι, δημιουργείται ένας πίνακας με τις αντίστοιχες TF-IDF τιμές όπου η κάθε γραμμή του αντιστοιχίζεται σε έναν από τους όρους που εμφανίζονται στα κείμενα των POIs ενώ κάθε στήλη του αντιστοιχίζεται σε ένα από τα κείμενα αυτά. Ο πίνακας αυτός θα συμβολίζεται ως A.

Το τελευταίο βήμα της διαδικασίας LSI είναι η εξαγωγή των βασικότερων εννοιών που υπάρχουν στα κείμενα των POIs. Δεδομένου του πίνακα με τα ζεύγη τιμών TF-IDF (A), παρουσιάζουν ιδιαίτερο ενδιαφέρον γραμμές του πίνακα οι οποίες "μοιάζουν" μεταξύ τους. Αυτές οι γραμμές αντιστοιχούν σε λέξεις οι οποίες εμφα-

νίζονται συχνά στα ίδια κείμενα οπότε πιθανότατα αφορούν την ίδια σημασιολογική έννοια. Έτσι, λέξεις όπως "καθεδρικός", "εκκλησία", "βασιλική" μπορούν να συγχωνευθούν σε μία και τα κείμενα που τις περιέχουν να περιγραφούν ικανοποιητικά καλά με τη χρήση μίας κοινής έννοιας αντί για πολλές διαφορετικές λέξεις.

Αφού ο πίνακας A περιέχει τιμές TF-IDF για τα N κείμενα και τους $M > N$ όρους, είναι πάρα πολύ πιθανό ότι ο *βαθμός (rank)* του είναι ίσος με N. Άρα, από μαθηματική σκοπιά, το ζητούμενο της μεθόδου LSI είναι η εύρεση ενός πίνακα που θα προσεγγίζει τον A αλλά με αρκετά μικρότερο βαθμό. Διαισθητικά, με αυτό τον τρόπο, αναπαράγεται μεγάλο μέρος της υπάρχουσας πληροφορίας αλλά πολλά συστατικά στοιχεία των κειμένων (λέξεις) συγχωνεύονται σε ευρύτερες έννοιες.

Η μέθοδος που πραγματοποιεί αυτή τη διαδικασία ονομάζεται *Παραγοντοποίηση Ιδιόμορφων Τιμών (Singular Value Decomposition – SVD)*. Σύμφωνα με αυτή, ένας οποιοσδήποτε πίνακας A μεγέθους $M \times N$ με $M > N$ μπορεί να παραγοντοποιηθεί ως $A = S\Sigma U^T$. Η διαδικασία αυτή βασίζεται στην εύρεση των ιδιοτιμών των πινάκων $A^T A$ και AA^T που αποδεικνύεται ότι είναι πραγματικοί μη αρνητικοί αριθμοί και άρα μπορούν να γραφούν ως τετράγωνα μη αρνητικών πραγματικών ${\sigma_1}^2 \geq ... \geq {\sigma_n}^2$. Ο πίνακας $\Sigma_{n \times n}$ είναι διαγώνιος πίνακας με $\Sigma_{ii} = \sigma_i$.

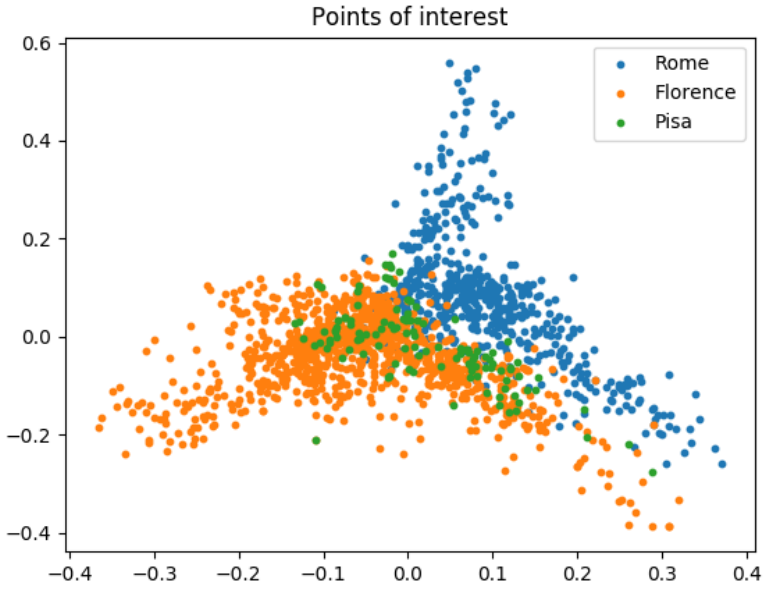
Στα πλαίσια της μεθόδου LSI, μπορούμε να δούμε ότι ο πίνακας $B = A^T A$ είναι ένας πίνακας $N \times N$ όπου κάθε στοιχείο περιέχει μια τιμή ανάλογη της συσχέτισης δύο κειμένων ενώ $C = AA^T$ ένας πίνακας $M \times M$ που περιέχει αντίστοιχες τιμές για τη συσχέτιση ζευγαριών λέξεων. Κάνοντας χρήση SVD παίρνουμε ότι $A = S\Sigma U^T$ όπου ο S $(M \times N)$ περιέχει τα ιδιοδιανύσματα του C, ο U $(N \times N)$ τα ιδιοδιανύσματα του B και ο $\Sigma_{N \times N}$ είναι ο διαγώνιος πίνακας που περιέχει τις ρίζες των ιδιοτιμών του B ή αλλιώς τις *ιδιόμορφες τιμές (singular values)*. Προκειμένου να "ρίξουμε" το βαθμό του πίνακα A χωρίς να χάσουμε πολύ πληροφορία, διατηρούνται οι r μεγαλύτερες ιδιόμορφες τιμές και αποκόπτονται οι υπόλοιπες μαζί με τα αντίστοιχα ιδιοδιανύσματα στους πίνακες S, U. Έτσι, σχηματίζονται οι νέοι πίνακες S_r $(M \times r)$, Σ_r $(r \times r)$, U_r $(N \times r)$. Παρ' όλα αυτά, είναι εύκολο να δούμε ότι το γινόμενό τους εξακολουθεί να έχει διάσταση $M \times N$ οπότε ισχύει $A \simeq S_r \Sigma_r {U_r}^T$.

Διαισθητικά, τα r ιδιοδιανύσματα που απομένουν αντιστοιχούν σε r κρυφές (latent) έννοιες των κειμένων. Έτσι, τα κείμενα των POIs μπορούν να αναπαρασταθούν ως διανύσματα στον ελαττωμένο διανυσματικό χώρο που δημιουργήθηκε. Σύμφωνα με τα παραπάνω, η κάθε στήλη του $r \times N$ πίνακα $\Sigma_r {U_r}^T$ αποτελεί το διάνυσμα απεικόνισης του αντίστοιχου κειμένου από τα N POIs.

Στα πλαίσια της εργασίας, πραγματοποιήθηκε η παραπάνω διαδικασία για το σύνολο των POIs Ρώμης, Φλωρεντίας και Πίζας ταυτόχρονα. Η διάσταση των διανυσμάτων επιλέχθηκε εμπειρικά ως $r = 10$ καθώς περισσότερες διαστάσεις δεν έδιναν αρκετή πληροφορία ώστε να υπάρχει μεγαλύτερη διαφοροποίηση των διαφόρων κατηγοριών αξιοθέατων.

Αν και τα βήματα της μεθόδου LSI είναι τα παραπάνω, απαιτείται ένα ακόμα βήμα

προκειμένου τα διανύσματα που προκύπτουν να είναι σε μορφή τέτοια που να επιτρέπει τη σύσταση διαδρομών από πόλη σε πόλη. Όπως αναφέρθηκε και προηγουμένως, τα διανύσματα αξιοθέατων με παρόμοια χαρακτηριστικά πρέπει να βρίσκονται κοντά στο χώρο ανεξαρτήτως της πόλης στην οποία βρίσκονται. Μία πρώτη απεικόνιση των POIs και των τριών πόλεων όπως προκύπτουν από τη μέθοδο LSI δίνεται στο Σχήμα 13.



Σχήμα 13: Αξιοθέατα διαφορετικών πόλεων

Είναι πολύ εύκολο να δούμε ότι η μέθοδος έχει καταφέρει να διαχωρίσει σε μεγάλο βαθμό τις πόλεις μεταξύ τους. Για παράδειγμα, ένα ζευγάρι POIs στη Ρώμη και στη Φλωρεντία, ακόμα κι αν είναι σχεδόν ίδια, δεν θα βρίσκονται κοντά στο διανυσματικό χώρο. Αυτό συμβαίνει διότι τα αξιοθέατα μιας πόλης περιέχουν πολλές φορές το όνομα της πόλης μέσα στην αντίστοιχη σελίδα της Wikipedia. Συνεπώς, χρησιμοποιώντας τα ονόματα των πόλεων και ίσως κάποιες σχετικές λέξεις ακόμα, η LSI εξάγει ως βασικό διαχωριστικό χαρακτηριστικό την πόλη στην οποία ανήκει το εκάστοτε POI.

Για να διορθωθεί αυτό το πρόβλημα, απαιτείται κάποιου είδους κανονικοποίηση των διανυσμάτων των αξιοθέατων που να αφαιρεί τις ιδιότητες των πόλεων. Η κανονικοποίηση που επιλέγεται μπορεί να εξηγηθεί διαισθητικά μέσω ενός παραδείγματος. Στην περίπτωση του Πύργου της Πίζας, το πραγματικό ενδιαφέρον δεν είναι στην απεικόνιση του συγκεκριμένου POI με ένα διάνυσμα. Ο στόχος είναι να απεικονιστεί ένας "γενικευμένος" Πύργος της Πίζας, δηλαδή ένα αξιοθέατο ανε-

ξαρτήτως πόλης που θα περιλαμβάνει όμως τα ίδια χαρακτηριστικά. Αφήνοντας τη μαθηματική αυστηρότητα και θεωρώντας ότι μπορούν να πραγματοποιηθούν πράξεις μεταξύ εννοιών, τα παραπάνω συνοψίζονται ως εξής:

$$\begin{aligned} GeneralizedTowerOfPisa &= (TowerOfPisa) - (Pisa) \\ &= (TowerOfPisa) - \{(POIOfPisa) - (POI)\} \\ &= (TowerOfPisa) - (POIOfPisa) + (POI) \end{aligned}$$

Κάνοντας την παραπάνω παραδοχή, μπορούμε να περιγράψουμε όλα τα αξιοθέατα ως διανύσματα με τη γενικευμένη τους μορφή αρκεί να ποσοτικοποιήσουμε τις έννοιες $(POIOfRome)$, $(POIOfFlorence)$, $(POIOfPisa)$ και (POI). Αν $M_C = \{\mathbf{m}_{C1}, \mathbf{m}_{C2}, ..., \mathbf{m}_{Cl}\}$ είναι το σύνολο διανυσμάτων των POIs της πόλης C, για τις τρεις πόλεις του συνόλου δεδομένων θα έχουμε τα σύνολα M_R, M_F, M_P. Συνεπώς, ορίζουμε την έννοια (POI) ως το μέσο όρο όλων των διανυσμάτων αξιοθέατων όλων των πόλεων:

$$(POI) = \frac{\sum_{C\in\{R,F,P\}} \sum_i \mathbf{m}_{Ci}}{\sum_{C\in\{R,F,P\}} |M_C|}$$

Αντιστοίχως, ορίζεται η έννοια του αξιοθέατου μιας συγκεκριμένης πόλης ως ο μέσος όρος των διανυσμάτων αξιοθέατων της αντίστοιχης πόλης:

$$(POIOfRome) = \frac{\sum_i \mathbf{m}_{Ri}}{|M_R|}$$

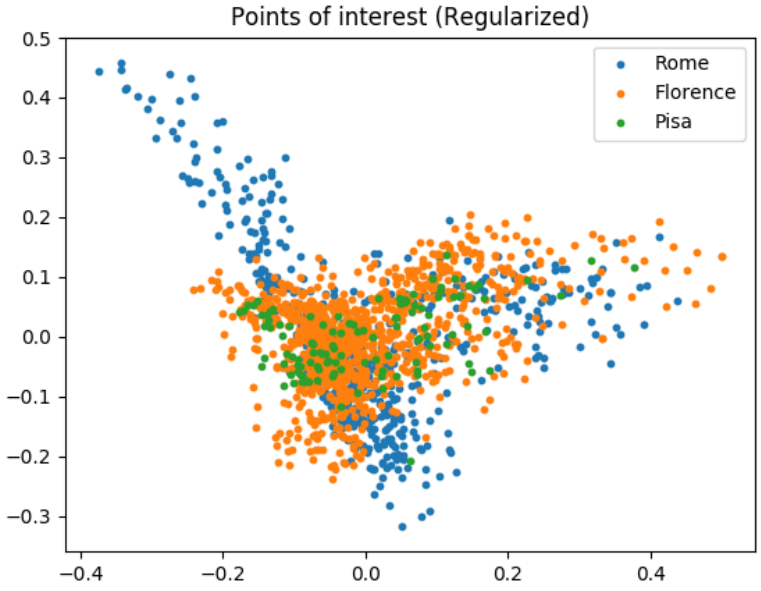
$$(POIOfFlorence) = \frac{\sum_i \mathbf{m}_{Fi}}{|M_F|}$$

$$(POIOfPisa) = \frac{\sum_i \mathbf{m}_{Pi}}{|M_P|}$$

Έχοντας ποσοτικοποιήσει τις απαραίτητες έννοιες, είναι πλέον εύκολο να γίνει κανονικοποίηση των διανυσμάτων των POIs έτσι ώστε να περιέχουν μόνο τα χαρακτηριστικά κάθε αξιοθέατου ανεξαρτήτως της πόλης στην οποία ανήκει. Για ένα αξιοθέατο μιας πόλης K με αρχικό διάνυσμα $\mathbf{m}_{Ki}$, το νέο του κανονικοποιημένο διάνυσμα θα δίνεται από την παρακάτω σχέση:

$$\mathbf{m}'_{Ki} = \mathbf{m}_{Ki} - \frac{\sum_j \mathbf{m}_{Kj}}{|M_K|} + \frac{\sum_{C\in\{R,F,P\}} \sum_j \mathbf{m}_{Cj}}{\sum_{C\in\{R,F,P\}} |M_C|}$$

Αφού εφαρμοστεί η παραπάνω διαδικασία στα διανύσματα των POIs των τριών πόλεων, αυτά μετασχηματίζονται και η νέα τους απεικόνιση δίνεται στο Σχήμα 14.



Σχήμα 14: Αξιοθέατα διαφορετικών πόλεων (Με κανονικοποίηση)

Μέθοδοι τέτοιου τύπου, αν και φαίνονται αρκετά διαισθητικές χωρίς θεωρητική βάση, επιβεβαιώνονται εμπειρικά τόσο στη βιβλιογραφία [60] όσο και στο διαθέσιμο σύνολο δεδομένων αυτής της εργασίας. Με σύντομη εξέταση των διανυσμάτων, μπορεί να παρατηρηθεί ότι οι τρεις πόλεις είναι πλέον αναμεμιγμένες και μάλιστα αξιοθέατα που συσχετίζονται βρίσκονται στην ίδια περιοχή του διανυσματικού χώρου. Για παράδειγμα, η περιοχή στα δεξιά του σχήματος που περιλαμβάνει διανύσματα και των τριών πόλεων αφορά αξιοθέατα με χαρακτηριστικά εκκλησιών, κατηγορία με μεγάλη συχνότητα και στις τρεις Ιταλικές πόλεις. Αντιθέτως, η περιοχή πάνω αριστερά όπου εμφανίζονται μόνο POIs της Ρώμης πρόκειται για απλές συνοικίες. Είναι λογικό μία μεγάλη πόλη ή πρωτεύουσα να περιέχει σελίδες της Wikipedia για τις συνοικίες της, κάτι τέτοιο δεν συμβαίνει όμως για μικρότερες πόλεις όπως η Φλωρεντία και η Πίζα.

Καθώς η διαδικασία απεικόνισης των αξιοθέατων σε διανύσματα έχει ολοκληρωθεί, αναφέρονται απλά κάποια τελικά βήματα που αφορούν άλλες πλευρές του συνόλου δεδομένων. Από τις διαθέσιμες τουριστικές διαδρομές που δίνονται, αφαιρούνται χρόνοι επίσκεψης σε αξιοθέατα μικρότεροι από 300 δευτερόλεπτα (5 λεπτά) καθώς πιθανότατα αποτελούν θόρυβο και επηρεάζουν σε μεγάλο βαθμό τους μέσους χρόνους επίσκεψης των αντίστοιχων αξιοθέατων. Επίσης, στο σύνολο δεδομένων του Κεφαλαίου 3 είχε αναφερθεί ότι αφαιρέθηκαν όσοι χρήστες είχαν ακολουθήσει διαδρομές αποτελούμενες από λιγότερα από 10 αξιοθέατα προκειμένου να είναι

πιο περιγραφικά τα προφίλ των χρηστών. Για να αυξηθεί ο αριθμός των διαθέσιμων χρηστών και να προκύψει λίγο μεγαλύτερη ποικιλομορφία στα διανύσματά τους, σε αυτό το κεφάλαιο, χαλαρώνεται αυτός ο περιορισμός και επιλέγονται οι χρήστες οι οποίοι έχουν επισκεφθεί τουλάχιστον 5 αξιοθέατα. Έτσι, το διάνυσμα-προφίλ του κάθε χρήστη ορίζεται ως το σταθμισμένο άθροισμα των γενικευμένων διανυσμάτων των POIs που επισκέφθηκε με βάρη ίσα με τον αριθμό των φωτογραφιών που τράβηξε στο καθένα. Τέλος, ως χρόνος επίσκεψης σε κάθε POI θεωρείται ο μέσος όρος των χρόνων που αφιέρωσαν όσοι χρήστες το επισκέφθηκαν.

Παρακάτω, παρατίθεται μία σύντομη ανάλυση του νέου συνόλου δεδομένων, όμοια με αυτή που έγινε στο Κεφάλαιο 3. Στον επόμενο πίνακα δίνονται κάποια βασικά χαρακτηριστικά του.

Πόλη	**Αξιοθέατα**	**Χρήστες**
Ρώμη	578	3749
Φλωρεντία	998	1935
Πίζα	116	288

Πίνακας 5: Χαρακτηριστικά νέου συνόλου δεδομένων

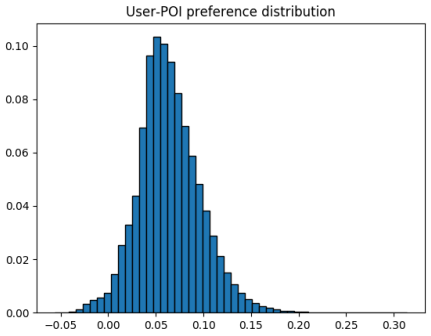
Επίσης, όπως αναφέρθηκε στην αρχή του κεφαλαίου, για το πρόβλημα της σύστασης διαδρομών από πόλη σε πόλη, ένα σύνολο χρηστών χρησιμοποιείται για την αξιολόγηση των συστάσεων και των εκτιμήσεων που κάνουν οι εξεταζόμενοι αλγόριθμοι. Το σύνολο αυτό περιλαμβάνει τους χρήστες οι οποίοι έχουν επισκεφθεί τουλάχιστον 2 από τις 3 πόλεις. Επομένως, για κάθε ζευγάρι πόλεων, δίνεται ο αριθμός των κοινών επισκεπτών τους.

Ζεύγος πόλεων	**Κοινοί χρήστες**
Ρώμη & Φλωρεντία	615
Φλωρεντία & Πίζα	135
Πίζα & Ρώμη	109

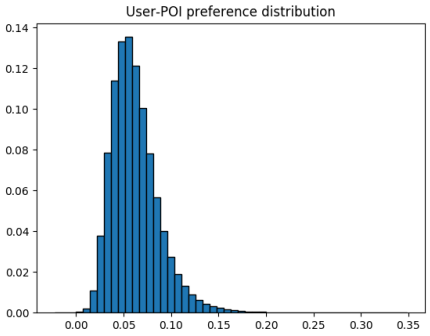
Πίνακας 6: Κοινοί επισκέπτες ανά ζεύγος πόλεων

Όπως αναφέρθηκε και προηγουμένως, για το πρόβλημα της σχεδίασης διαδρομών, έχει πολύ μεγάλη σημασία η κατανομή της προτίμησης ανά ζεύγος POI-χρήστη. Έτσι, για κάθε πόλη υπολογίζεται το εσωτερικό γινόμενο μεταξύ των ζευγαριών διανυσμάτων POI-χρηστών. Η κατανομή αυτής της προτίμησης για τις τρεις πόλεις του νέου συνόλου δεδομένων παρουσιάζεται στο Σχήμα 15.

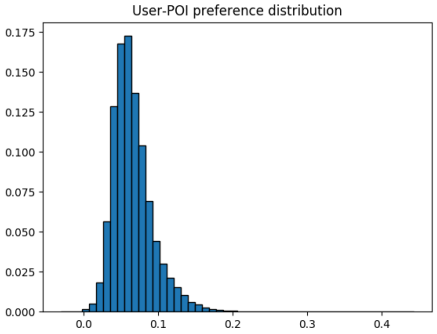
Η παρατήρηση που μπορεί να γίνει είναι ότι οι κατανομές έχουν παρόμοια μορφή με αυτή του Κεφαλαίου 3 οπότε η κοινή επεξεργασία του συνόλου δεδομένων για τις 3 πόλεις δεν επηρέασε σημαντικά τις προτιμήσεις των χρηστών. Οι κατανομές έχουν πάλι μία υψηλή συγκέντρωση γύρω από κάποια μέση τιμή προτίμησης ενώ



(α) Ρώμη



(β) Φλωρεντία



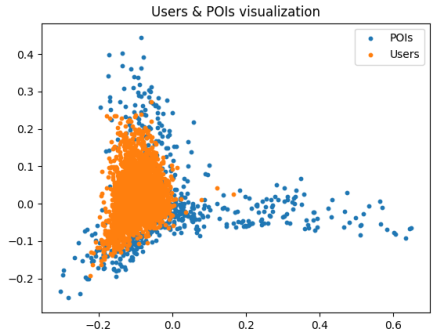
(γ) Πίζα

Σχήμα 15: Νέα κατανομή της ικανοποίησης ανά ζεύγος POI-χρήστη

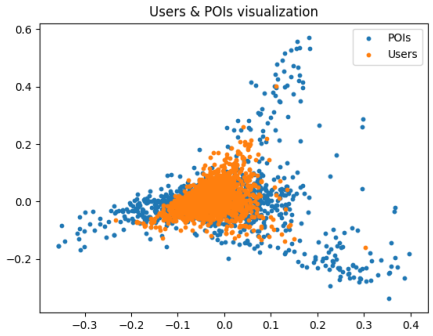
παρουσιάζουν μια ελαφριά ουρά στις υψηλότερες προτιμήσεις. Άρα, η αλγοριθμική βάση του προβλήματος παραμένει η ίδια.

Μία μικρή διαφοροποίηση που μπορεί να παρατηρηθεί είναι ότι στην περίπτωση της Ρώμης εμφανίζεται και μια μικρή ουρά στις χαμηλές προτιμήσεις κοντά στο μηδέν. Η ερμηνεία αυτής της διαφοράς είναι ότι πλέον υπάρχουν χρήστες οι οποίοι δεν μένουν καθόλου ικανοποιημένοι (ίσως είναι και ενοχλημένοι) από την επίσκεψη σε κάποια POIs. Αυτό πιθανότατα οφείλεται στην χαλάρωση του περιορισμού σχετικά με τον ελάχιστο αριθμό αξιοθέατων που πρέπει να έχουν επισκεφθεί οι χρήστες. Η ελάττωση του ορίου από 10 σε 5 ενδεχομένως να επιτρέπει τη δημιουργία προφίλ για χρήστες οι οποίοι δεν έχουν επισκεφθεί καθόλου κάποια κατηγορία αξιοθέατων επομένως το αντίστοιχο εσωτερικό γινόμενο θα δίνει μηδενική ή αρνητική τιμή. Παρ' όλα αυτά, η μορφή της κατανομής δεν αλλάζει αισθητά οπότε το αποτέλεσμα αυτού του συμβιβασμού μπορεί να θεωρηθεί αμελητέο.

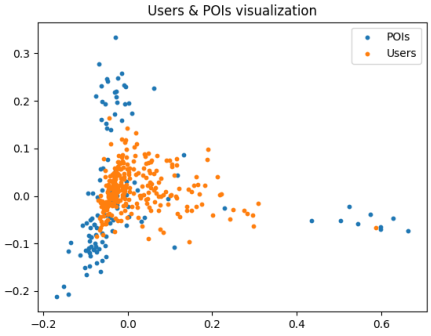
Κάτι τέτοιο μπορεί να φανεί και με μία πρόχειρη οπτικοποίηση των POIs και των χρηστών στις αντίστοιχες πόλεις. Τα αποτελέσματα φαίνονται στο Σχήμα 16.



(α) Ρώμη



(β) Φλωρεντία



(γ) Πίζα

Σχήμα 16: Οπτικοποίηση (Νέο σύνολο δεδομένων)

Μπορούμε να δούμε ότι σε ποιοτικό επίπεδο τα δεδομένα έχουν ίδια μορφή με το αρχικό σύνολο δεδομένων του Κεφαλαίου 3. Η μόνη αισθητή διαφορά φαίνεται στην περίπτωση της Ρώμης, όπου οι χρήστες είναι αρκετά πιο ανοιγμένοι στο επίπεδο συγκριτικά με το αρχικό σύνολο δεδομένων. Μάλιστα, φαίνεται πως καταλαμβάνουν σε μεγάλο βαθμό την περιοχή στα αριστερά του επιπέδου, κάτι που δεν συνέβαινε προηγουμένως. Αυτό έχει ως αποτέλεσμα οι χρήστες να έχουν μεγαλύτερη ποικιλομορφία, πράγμα χρήσιμο στη μελέτη του συγκεκριμένου προβλήματος. Αυτή η τάση των χρηστών προς τα αριστερά του επιπέδου μπορεί να εξηγήσει και τη μικρή ουρά που παρατηρείται στην κατανομή ικανοποίησης ανά ζεύγος POI-χρήστη καθώς μια μεγάλη μάζα χρηστών βρίσκεται στα αριστερά ενώ μια μερίδα από POIs βρίσκονται πολύ δεξιά στο επίπεδο. Όπως είναι λογικό, το εσωτερικό τους γινόμενο είναι μηδενικό, ίσως και αρνητικό.

Έχοντας περιγράψει πλήρως το τροποποιημένο σύνολο δεδομένων και τους αλγόριθμους που εξετάζονται μπορεί πλέον να μελετηθεί η συμπεριφορά των αλγορίθμων που περιγράφηκαν στην αρχή του κεφαλαίου.

4.4 Πειραματική αξιολόγηση σε δεδομένα χρηστών

Σε αυτή την ενότητα, εξετάζεται η αποτελεσματικότητα των αλγορίθμων που αναφέρθηκαν νωρίτερα, τόσο από την πλευρά της σύστασης διαδρομών όσο και από την πλευρά της εκτίμησης για το προφίλ των χρηστών.

Σε αυτό το σημείο υπενθυμίζεται ο τρόπος με τον οποίο εξετάζουμε το πρόβλημα. Δεδομένων δύο πόλεων SourceCity, TargetCity θεωρούμε ότι ένας χρήστης TestUser έχει ήδη επισκεφθεί την SourceCity και έχει δημιουργηθεί ένα προφίλ γι αυτόν το οποίο συμβολίζεται με το διάνυσμα $\mathbf{u}_0$ καθώς και μία διαδρομή η οποία του προσφέρει τη μέγιστη ικανοποίηση για αυτό το προφίλ, η T_0. Την TargetCity έχουν επισκεφθεί ήδη ένα σύνολο χρηστών TrainUsers για τους οποίους έχουν δημιουργηθεί προφίλ $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_l$ και τα προφίλ αυτά έχουν ομαδοποιηθεί με συσταδοποίηση K-Μέσων. Το πραγματικό προφίλ του TestUser στην TargetCity συμβολίζεται ως $\mathbf{u}^*$ και θεωρείται κρυφό από τους αλγόριθμους ενώ χρησιμοποιείται μόνο για να αξιολογηθεί η αποτελεσματικότητα της εκάστοτε εκτίμησης. Το εκτιμώμενο προφίλ που επιστρέφει ένας αλγόριθμος συμβολίζεται ως $\hat{\mathbf{u}}$. Επίσης, ως $\hat{T}$ συμβολίζεται η διαδρομή που συστήνεται στο χρήστη και ως T^* η διαδρομή που υπολογίζεται σύμφωνα με το πραγματικό προφίλ $\mathbf{u}^*$. Τέλος, ως $p_{\mathbf{u}}(T)$ συμβολίζεται η αξία μιας διαδρομής T με βάση ένα προφίλ $\mathbf{u}$.

Προκειμένου να αξιολογήσουμε την αποδοτικότητα των αλγορίθμων όσον αφορά το κομμάτι των συστάσεων, είναι απαραίτητη μία ποσότητα που θα χρησιμοποιηθεί ως μετρική. Είναι λογικό να υποτεθεί ότι μία σύσταση διαδρομής είναι πολύ καλή αν προσφέρει μεγάλη ικανοποίηση στον TestUser με βάση το πραγματικό του προφίλ. Έτσι, ορίζεται η μετρική recScore ως ο λόγος της πραγματικής αξίας της διαδρομής που συστήνεται προς την πραγματική αξία της βέλτιστης διαδρομής. Αναλυτικά, δίνεται από τον τύπο:

$$recScore = \frac{P_{\mathbf{u}^*}(\hat{T})}{P_{\mathbf{u}^*}(T^*)}$$

Είναι εύκολο να δούμε ότι θεωρητικά η μετρική $recScore$ μπορεί να πάρει μέγιστη τιμή 1 αφού δεν υπάρχει μονοπάτι T με $P_{\mathbf{u}^*}(T) > P_{\mathbf{u}^*}(T^*)$ αλλιώς το T^* δεν θα ήταν βέλτιστο. Αυτό δεν συμβαίνει πάντα λόγω του ότι οι διαδρομές υπολογίζονται με χρήση ευριστικών τεχνικών οπότε δεν υπάρχει αυστηρή εγγύηση ότι η T^* η οποία υπολογίζεται με βάση το προφίλ $\mathbf{u}^*$ είναι όντως η βέλτιστη διαδρομή για το προφίλ $\mathbf{u}^*$. Συνεπώς, ο ορισμός της $recScore$ μετατρέπεται ως εξής:

$$recScore = min\{1, \frac{P_{\mathbf{u}^*}(\hat{T})}{P_{\mathbf{u}^*}(T^*)}\}$$

Βλέποντας το πρόβλημα από την πλευρά της εκτίμησης του πραγματικού προφίλ του χρήστη, οι μετρικές που μπορούν να σχεδιαστούν είναι πολλές. Η πιο εμφανής είναι αυτή που συγκρίνει τα δύο προφίλ, το πραγματικό και το εκτιμώμενο.

Συγκεκριμένα, πρόκειται για την ευκλείδεια απόσταση των δύο προφίλ. Ο τυπικός ορισμός της είναι:

$$profDist = \| \hat{\mathbf{u}} - \mathbf{u}^* \|_2$$

Ένας άλλος τρόπος να μετρηθεί η αποδοτικότητα ενός αλγορίθμου είναι να υπολογιστεί κάποια μορφή απόστασης μεταξύ της διαδρομής που συστήνει και της βέλτιστης διαδρομής T^*. Έστω ότι η διαδρομή $\hat{T}$ περιλαμβάνει τα διανύσματα $\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_l$ ενώ η T^* περιλαμβάνει τα $\mathbf{m}_1^*, \mathbf{m}_2^*, ..., \mathbf{m}_L^*$. Τότε, ως μέτρο απόστασης ανάμεσα στα δύο σύνολα διανυσμάτων θεωρούμε την παρακάτω ποσότητα:

$$pathDist = max\{\frac{1}{L}\sum_{i=1}^{L}\min_{1\leq j\leq l} \| \mathbf{m}_i^* - \mathbf{m}_j \|_2, \frac{1}{l}\sum_{j=1}^{l}\min_{1\leq i\leq L} \| \mathbf{m}_i^* - \mathbf{m}_j \|_2\}$$

Η μετρική αυτή έχει περίπλοκη μορφή αλλά είναι πολύ απλή διαισθητικά. Έχοντας δύο σύνολα, για κάθε στοιχείο του ενός, βρίσκεται το πλησιέστερο στοιχείο του άλλου και υπολογίζεται η μεταξύ τους απόσταση. Από αυτές τις αποστάσεις υπολογίζεται η μέση τιμή. Από τις δύο μέσες τιμές επιστρέφεται τελικά η μεγαλύτερη προκειμένου η μετρική απόστασης να γίνει συμμετρική ως προς τη σειρά των διαδρομών.

Τέλος, εισάγεται μία ακόμα μετρική η οποία χρησιμοποιείται για τη μελέτη της δυνατότητας γενίκευσης των αλγορίθμων που παρουσιάζονται. Για την περιγραφή της είναι απαραίτητο να αναφερθούν κάποια στοιχεία από τον τομέα της μηχανικής μάθησης και συγκεκριμένα από τα προβλήματα ταξινόμησης (classification). Στα προβλήματα ταξινόμησης δίνεται ένα σύνολο δεδομένων τα οποία συνοδεύονται από μία ετικέτα 1 ή 0 και ένα αλγόριθμος προσπαθεί να αναθέσει ετικέτες στα δεδομένα χωρίς να γνωρίζει τις πραγματικές τους. Ένα τέτοιο παράδειγμα θα μπορούσε να είναι η αναγνώριση μιας εικόνας ενός όγκου ως καλοήθους ή κακοήθους από ένα ιατρικό σύστημα. Σε τέτοια προβλήματα, η ταξινόμηση μπορεί να επιφέρει τέσσερα δυνατά αποτελέσματα, τα True Positive (Πραγματική ετικέτα: 1, Εκτιμώμενη ετικέτα: 1), True Negative (Πραγματική ετικέτα: 0, Εκτιμώμενη ετικέτα: 0), False Positive (Πραγματική ετικέτα: 0, Εκτιμώμενη ετικέτα: 1), False Negative(Πραγματική ετικέτα: 1, Εκτιμώμενη ετικέτα: 0).

Πολλές φορές οι αλγόριθμοι ταξινόμησης δεν επιστρέφουν αυστηρά μία ετικέτα για κάθε δείγμα από τα δεδομένα αλλά ένα score ίσο με την πιθανότητα αυτή η ετικέτα να είναι 1. Μία μετρική η οποία χρησιμοποιείται για να περιγράψει την αποδοτικότητα τέτοιων αλγορίθμων ταξινόμησης σε προβλήματα είναι η *AUC (Area Under the Curve)* [61]. Χωρίς να μπούμε σε πολλές λεπτομέρειες, η μετρική αυτή αντιστοιχεί διαισθητικά στην εξής πιθανότητα. Αν επιλεγούν τυχαία δύο δείγματα, ένα με ετικέτα 1 και ένα με ετικέτα 0, η AUC είναι η πιθανότητα ο αλγόριθμος ταξινόμησης να δώσει υψηλότερο score στο πρώτο από ό,τι στο δεύτερο.

Για τον υπολογισμό αυτής της μετρικής γίνεται η εξής διαδικασία. Επιλέγεται τυχαία ένα σύνολο από δείγματα και ταξινομούνται σε φθίνουσα σειρά ως προς το score που τους δίνει ο αλγόριθμος. Για παράδειγμα, οι πραγματικές ετικέτες 10 δειγμάτων ταξινομημένων ως προς το score μπορεί να έχουν τη μορφή 1110100000. Είναι εμφανές ότι σε αυτό το παράδειγμα ο αλγόριθμος έχει κάνει ένα λάθος αφού υπάρχει ένα δείγμα με ετικέτα 1 με score χαμηλότερο από ένα δείγμα με ετικέτα 0. Προκειμένου η λίστα με τις ετικέτες να ταξινομηθεί σύμφωνα με την πραγματική τιμή των ετικετών απαιτείται ένας αριθμός από swaps, δηλαδή εναλλαγών διπλανών στοιχείων. Έτσι, αν έχουν επιλεγεί n δείγματα, ως AUC ορίζεται η ποσότητα:

$$AUC = 1 - \frac{\#swaps}{\frac{n(n-1)}{2}}$$

Στο παραπάνω απλό παράδειγμα απαιτείται μόνο ένα swap οπότε η AUC είναι 1-1/(10*9/2)=44/45=97.8%. Είναι εύκολο να δούμε πώς με αυτό τον τρόπο εξετάζεται η ικανότητα του αλγορίθμου να κατασκευάσει μία *σειρά κατάταξης (ranking)* για τα δείγματα κοντά στην πραγματική.

Για την αξιολόγηση των αλγορίθμων σύστασης διαδρομών από πόλη σε πόλη εισάγεται μία μετρική η οποία αποτελεί προσπάθεια γενίκευσης της προηγούμενης για προβλήματα που δεν εμπίπτουν στα στενά όρια του classification. Η νέα αυτή μετρική ονομάζεται *G-AUC (Generalized AUC)* και υπολογίζεται ως εξής: Έστω ότι δίνονται κάποια τυχαία δείγματα από το σύνολο δεδομένων. Ο αλγόριθμος που εξετάζεται δίνει κάποιο εκτιμώμενο score για το καθένα και τα δείγματα ταξινομούνται σε φθίνουσα σειρά σύμφωνα με το score αυτό. Μέχρι αυτό το σημείο η διαδικασία είναι ίδια με τον υπολογισμό της κλασικής AUC. Η διαφοροποίηση συνίσταται στο ότι τα δείγματα δεν συνοδεύονται από κάποια δυαδική ετικέτα 0 ή 1 αλλά από κάποιο πραγματικό score που είναι πραγματικός αριθμός. Δηλαδή, υπάρχει μια κατάταξη των δειγμάτων σύμφωνα με το εκτιμώμενο score που έδωσε ο υπό εξέταση αλγόριθμος και μία κατάταξή τους σύμφωνα με το πραγματικό τους score. Ο τύπος για τον υπολογισμό της G-AUC είναι ίδιος με αυτόν για την AUC όπου $\#swaps$ είναι ο αριθμός των εναλλαγών διπλανών στοιχείων που πρέπει να γίνουν στην εκτιμώμενη κατάταξη ώστε αυτή να ταυτιστεί με την πραγματική.

Στο συγκεκριμένο πρόβλημα που μελετάμε, τα δείγματα που επιλέγονται είναι 50 τυχαία κατασκευασμένες διαδρομές. Η πραγματική τους κατάταξη υπολογίζεται έτσι ώστε να είναι σε φθίνουσα σειρά ως προς την αξία $p_{\mathbf{u}^*}(T)$ ενώ η εκτιμώμενη κατάταξη είναι σε φθίνουσα σειρά ως προς την αξία $p_{\hat{\mathbf{u}}}(T)$. Διαισθητικά, η G-AUC αντιστοιχεί στην πιθανότητα ένας από τους αλγόριθμους που μελετώνται να χαρακτηρίσει μια "καλή" διαδρομή περισσότερο ικανοποιητική από μία "κακή".

Στη συνέχεια, παρουσιάζονται οι τιμές των προηγούμενων μετρικών για τους αλγόριθμους που περιγράφηκαν στην Ενότητα 4.2. Κάθε φορά επιλέγεται ένα ζεύγος SourceCity, TargetCity, ως TestUsers χρησιμοποιούνται αυτοί για τους οποίους

υπάρχουν δεδομένα διαδρομής και για τις δύο πόλεις ενώ αυτοί που έχουν επισκεφθεί μόνο την TargetCity χρησιμοποιούνται ως TrainUsers. Λόγω του ότι τα προφίλ των χρηστών έχουν πολύ μικρή διακύμανση οι διαφορές στις τιμές που παίρνουν οι μετρικές είναι μικρές και αποτελούν μια ένδειξη για την αποδοτικότητα των αλγορίθμων. Για την ακριβέστερη εξέτασή τους είναι απαραίτητη η μελλοντική μελέτη τους σε καταλληλότερα σύνολα δεδομένων. Λόγω των δυσκολιών που επιβάλλει το παρόν σύνολο δεδομένων, δεν παρουσιάζονται τα αποτελέσματα για όλους τους πιθανούς συνδυασμούς SourceCity, TargetCity αλλά μόνο για κάποιους που οι διαφορές στις τιμές των μετρικών είναι ιδιαίτερα εμφανείς.

Στον παρακάτω πίνακα παρουσιάζονται οι τιμές των μετρικών για SourceCity τη Φλωρεντία και TargetCity τη Ρώμη.

Αλγόριθμος/Μετρική	**recScore**	**pathDist**	**profDist**	**G-AUC**
PopularPath	0.5938	-	-	-
RandomCluster	0.9289	0.0741	0.1072	0.9411
MostSimilarCluster	0.9506	0.0782	0.1120	0.9324
WeightedCluster	0.9349	0.0645	0.0862	0.9559
GlobalCentroid	0.9382	0.0643	0.0843	0.9573

Πίνακας 7: Αποτελέσματα αλγορίθμων (Φλωρεντία → Ρώμη)

Η πρώτη παρατήρηση είναι ότι η τιμή της μετρικής recScore για τη διαδρομή που επιστρέφει ο αλγόριθμος PopularPath είναι πολύ χαμηλή σχετικά με τους υπόλοιπους αλγόριθμους. Δηλαδή, η σύσταση διαδρομών στους χρήστες με βάση τη διασημότητα των αξιοθέατων κάθε πόλης δίνει καθαρά μη βέλτιστα αποτελέσματα από πλευράς ικανοποίησης. Αυτό συμβαίνει διότι οι διαδρομές αυτές σχεδόν πάντα περιέχουν τα διασημότερα POIs της αντίστοιχης πόλης δίνοντας μικρή σημασία στο χρόνο που θα αφιερώσει ένας τουρίστας σε αυτά. Το αποτέλεσμα τέτοιων επιλογών είναι η δημιουργία διαδρομών που αποτελούνται από λίγα αξιοθέατα που είναι πολύ διάσημα ενώ θα μπορούσαν να δημιουργηθούν πιο πλούσιες διαδρομές με λιγότερο διάσημα αξιοθέατα που συνολικά θα αύξαναν την ικανοποίηση των τουριστών. Επομένως, οι αλγόριθμοι που έχουν σχεδιαστεί είναι σαφώς αποτελεσματικότεροι από τον πρώτο baseline αλγόριθμο.

Επίσης, μπορεί να παρατηρηθεί ότι και ο RandomCluster, αν και πετυχαίνει σαφώς καλύτερο recScore από τον PopularPath, είναι κι αυτός χειρότερος από τους επόμενους καθώς δεν βασίζεται σε κάποιο αυστηρό κριτήριο αλλά πραγματοποιεί την ανάθεση του TestUser σε ένα από τα υπάρχοντα κεντροειδή των TrainUsers εντελώς τυχαία.

Οι αλγόριθμοι που έχουν πραγματικό ενδιαφέρον και αξίζει να μελετηθούν πιο προσεκτικά είναι οι MostSimilarCluster, WeightedCluster, GlobalCentroid. Όσον αφορά την προσφερόμενη ικανοποίηση στο χρήστη, παρατηρούμε ότι ο αλγόριθμος

MostSimilarCluster πετυχαίνει το μεγαλύτερο recScore ενώ οι άλλοι δύο πετυχαίνουν χαμηλότερες τιμές. Αυτό είναι αναμενόμενο καθώς ο MostSimilarCluster επιλέγει για τον TestUser τη διαδρομή της συστάδας των TrainUsers που θα του προσφέρει τη μεγαλύτερη αναμενόμενη ικανοποίηση. Συνεπώς, ο MostSimilarCluster έχει ως αποκλειστικό στόχο τη μεγιστοποίηση της ικανοποίησης. Αντιθέτως, ο WeightedCluster θυσιάζει μερικώς το στόχο αυτό προκειμένου να κάνει ακριβέστερη εκτίμηση για το διάνυσμα του TestUser. Παρ' όλα αυτά, επειδή η διακύμανση των προφίλ των χρηστών είναι πολύ μικρή, ο WeightedCluster κάνει εκτίμηση για το διάνυσμα η οποία είναι παρόμοια με τον GlobalCentroid, δηλαδή στη μέση τιμή όλων των διανυσμάτων των TrainUsers. Όπως φαίνεται και από τον Πίνακα 7, οι WeightedCluster και GlobalCentroid έχουν παρόμοιες τιμές για όλες τις μετρικές.

Από την άλλη, είναι εύκολο να δούμε ότι στις μετρικές που έχουν να κάνουν με εκτίμηση και όχι με ικανοποίηση, δηλαδή στις pathDist, profDist, G-AUC, επικρατούν οι αλγόριθμοι WeightedCluster, GlobalCentroid. Το χαρακτηριστικό αυτό βασίζεται στο ότι οι αλγόριθμοι αυτοί είναι πιο ανεκτικοί στα λάθη. Ο MostSimilarCluster μπορεί να υπολογίσει ότι η αναμενόμενη ικανοποίηση μεγιστοποιείται αντιστοιχώντας τον TestUser με μία συγκεκριμένη συστάδα των TrainUsers χωρίς αυτό να σημαίνει απαραίτητα ότι η συστάδα αυτή είναι η πλησιέστερη στο πραγματικό (κρυφό) προφίλ του TestUser. Σε αυτή την περίπτωση, θα έχει ικανοποιήσει σε μεγάλο βαθμό το χρήστη αλλά θα έχει κάνει μεγάλο λάθος στην εκτίμηση του προφίλ του. Αντιθέτως, οι WeightedCluster και GlobalCentroid κάνουν πιο μετριοπαθή εκτίμηση και γι αυτό δεν πραγματοποιούν σημαντικά λάθη, κάτι που φαίνεται ιδιαίτερα από τη διαφορά που παρουσιάζουν στη μετρική G-AUC, η οποία δίνει τιμή κρίνοντας ένα σύνολο 50 διαδρομών και όχι μία μεμονωμένη διαδρομή.

Στη συνέχεια, δίνονται τα αντίστοιχα αποτελέσματα για SourceCity την Πίζα και TargetCity τη Ρώμη.

Αλγόριθμος/Μετρική	**recScore**	**pathDist**	**profDist**	**G-AUC**
PopularPath	0.5836	-	-	-
RandomCluster	0.9327	0.0769	0.0981	0.9420
MostSimilarCluster	0.9752	0.0727	0.0972	0.9457
WeightedCluster	0.9350	0.0657	0.0726	0.9638
GlobalCentroid	0.9332	0.0678	0.0710	0.9646

Πίνακας 8: Αποτελέσματα αλγορίθμων (Πίζα → Ρώμη)

Και σε αυτή την περίπτωση μπορούν να γίνουν παρόμοιες παρατηρήσεις με προηγουμένως. Ο αλγόριθμος MostSimilarCluster πετυχαίνει ξανά υψηλότερο recScore από τους υπόλοιπους, δηλαδή καταφέρνει να πραγματοποιήσει πολύ καλές συστάσεις στους χρήστες ώστε να μεγιστοποιήσει την ικανοποίησή τους. Αντιθέτως, βλέπουμε πάλι ότι οι αλγόριθμοι GlobalCentroid και WeightedCluster παρουσιά-

ζουν καλύτερη τιμή G-AUC δηλαδή κάνουν μικρότερα σφάλματα στην εκτίμηση του πραγματικού προφίλ του εκάστοτε TestUser.

Στους πίνακες 9 και 10 παρουσιάζονται τα αντίστοιχα αποτελέσματα για TargetCity την Πίζα και SourceCity τις Φλωρεντία, Ρώμη αντίστοιχα.

Αλγόριθμος/Μετρική	**recScore**	**pathDist**	**profDist**	**G-AUC**
PopularPath	0.8162	-	-	-
RandomCluster	0.8937	0.0804	0.2546	0.7630
MostSimilarCluster	0.9693	0.0680	0.1262	0.8588
WeightedCluster	0.9302	0.1042	0.1862	0.7324
GlobalCentroid	0.9616	0.0556	0.1011	0.8834

Πίνακας 9: Αποτελέσματα αλγορίθμων (Φλωρεντία → Πίζα)

Αλγόριθμος/Μετρική	**recScore**	**pathDist**	**profDist**	**G-AUC**
PopularPath	0.8366	-	-	-
RandomCluster	0.8930	0.0736	0.2406	0.7818
MostSimilarCluster	0.9539	0.0551	0.1494	0.8604
WeightedCluster	0.9353	0.0848	0.1796	0.7424
GlobalCentroid	0.9624	0.0470	0.1000	0.8906

Πίνακας 10: Αποτελέσματα αλγορίθμων (Ρώμη → Πίζα)

Στην περίπτωση που η πόλη που χρησιμοποιείται ως TargetCity είναι η Πίζα, τα αποτελέσματα που παρατηρούνται είναι διαφορετικά. Η Πίζα έχει το ιδιαίτερο χαρακτηριστικό ότι έχει πολύ λιγότερους TrainUsers. Αυτό ενδεχομένως να επηρεάζει την αποδοτικότητα των αλγορίθμων που εξετάζονται.

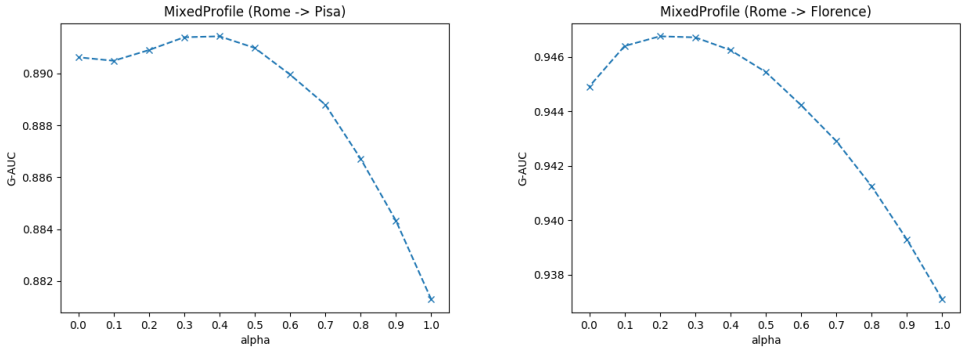
Η πρώτη παρατήρηση που μπορεί να γίνει είναι ότι, όπως και προηγουμένως, οι baseline αλγόριθμοι PopularPath και RandomCluster πετυχαίνουν χαμηλότερο recScore σχετικά με τους υπόλοιπους. Η άλλη παρατήρηση είναι ότι ο αλγόριθμος MostSimilarCluster δυσκολεύεται να σχεδιάσει διαδρομή που να μεγιστοποιεί την ικανοποίηση των τουριστών καθώς το recScore που δίνει είναι σχεδόν ίδιο με αυτό που δίνει ο GlobalCentroid. Η αδυναμία αυτή οφείλεται στην ανεπάρκεια σε TrainUsers ώστε να "εκπαιδευτεί" κατάλληλα από την προηγούμενη συμπεριφορά άλλων χρηστών πέραν του TestUser. Δηλαδή, η διαδικασία προσωποποίησης των διαδρομών είναι δυσκολότερη σε αυτή την περίπτωση και έτσι η ασφαλέστερη επιλογή είναι η χρήση ενός πιο απλού αλγορίθμου όπως ο GlobalCentroid.

Σε αυτή την περίπτωση ο GlobalCentroid έχει πολύ καλή συμπεριφορά καθώς το recScore του δεν διαφοροποιείται σημαντικά από αυτό του MostSimilarCluster αλλά ταυτόχρονα παρουσιάζει και υψηλότερη G-AUC δηλαδή δεν πραγματοποιεί σημαντικά λάθη κατά την εκτίμηση του προφίλ του χρήστη. Συνολικά, βλέπουμε

ότι όλοι οι αλγόριθμοι παρουσιάζουν χαμηλή G-AUC σχετικά με τις περιπτώσεις των Πινάκων 7,8 δηλαδή δεν καταφέρνουν επιτυχώς να κάνουν εκτίμηση για το προφίλ η οποία να δίνει σωστή κατάταξη για το σύνολο των προκατασκευασμένων διαδρομών, δείχνοντας την αδυναμία της εκτίμησης όταν το σύνολο των διαθέσιμων χρηστών δεν είναι επαρκές.

Όπως έχει αναφερθεί και νωρίτερα, από τις τρεις πόλεις, η Φλωρεντία είναι αυτή στην οποία τα προφίλ των χρηστών παρουσιάζουν τη μικρότερη διασπορά. Συνεπώς, οι αλγόριθμοι που εξετάζονται δεν έχουν σημαντικές διαφοροποιήσεις ως προς τις μετρικές και γι αυτό τα σχετικά αποτελέσματα δεν περιλαμβάνονται.

Τέλος, μελετάται ο αλγόριθμος MixedProfile ο οποίος προτείνει ως προφίλ για τον TestUser στην TargetCity ένα γραμμικό συνδυασμό του προφίλ του ίδιου χρήστη στη SourceCity και του μέσου προφίλ των TrainUsers στην TargetCity. Καθώς ο συγκεκριμένος αλγόριθμος λειτουργεί με μία υπερπαράμετρο α που ρυθμίζει το βάρος σε αυτό το γραμμικό συνδυασμό, λαμβάνονται αποτελέσματα της εκτέλεσής του για μια σειρά τιμών της υπερπαραμέτρου στο διάστημα $[0, 1]$. Τα αποτελέσματα που παρουσιάζονται στο Σχήμα 17 είναι αυτά που έχουν ως SourceCity τη Ρώμη, καθώς σε αυτές τις περιπτώσεις η συσχέτιση μεταξύ του προφίλ στην SourceCity και του προφίλ στην TargetCity φαίνεται να είναι πιο έντονη.



(α) Ρώμη → Πίζα (β) Ρώμη → Φλωρεντία

Σχήμα 17: Μεταβολή G-AUC στον αλγόριθμο MixedProfile

Από τα παραπάνω διαγράμματα μπορεί να γίνει μία σειρά παρατηρήσεων. Πρώτον, φαίνεται καθαρά ότι το προφίλ του κάθε χρήστη μεταβάλλεται από πόλη σε πόλη. Ο αλγόριθμος MixedProfile για $\alpha = 1$ υποθέτει ότι το προφίλ του TestUser στην TargetCity ταυτίζεται με αυτό που είχε στη SourceCity. Είναι εύκολο να δούμε ότι η G-AUC που πετυχαίνει ο αλγόριθμος γι αυτή την τιμή του α είναι η χειρότερη δυνατή, άρα η συγκεκριμένη εκτίμηση είναι σε μεγάλο βαθμό λάθος.

Η δεύτερη και πιο ενδιαφέρουσα παρατήρηση είναι ότι η G-AUC ως συνάρτηση του α είναι μία κοίλη συνάρτηση που παρουσιάζει μέγιστο σε κάποιο εσωτερικό ση-

μείο του διαστήματος $[0, 1]$. Αυτό σημαίνει ότι η καλύτερη εκτίμηση για το προφίλ του TestUser στην TargetCity γίνεται λαμβάνοντας κυρίως υπόψη πώς συμπεριφέρεται ένας μέσος χρήστης στην πόλη αυτή αλλά διατηρώντας ένα τμήμα της πληροφορίας για την προηγούμενη συμπεριφορά του χρήστη στη SourceCity.

Η παρουσία της κορυφής της συνάρτησης σε εσωτερικό σημείο του $[0, 1]$ δείχνει ότι το προφίλ ενός χρήστη στη SourceCity είναι συνδεδεμένο με το προφίλ του στην TargetCity. Με αυτό τον τρόπο, δίνεται μια σαφής ένδειξη για ένα από τα βασικότερα ερωτήματα που τέθηκαν στην αρχή του κεφαλαίου. Το ερώτημα αφορούσε το κατά πόσο παρατηρείται κάποιο μοτίβο στο μετασχηματισμό του προφίλ του χρήστη όταν αυτός μετακινείται από πόλη σε πόλη. Η κορυφή δείχνει ότι ο μετασχηματισμός αυτός δεν είναι τυχαίος αλλά βασίζεται σε ένα βαθμό στις προτιμήσεις του χρήστη όπως αυτός τις εκφράζει στην SourceCity καθώς και στα διαθέσιμα POIs που υπάρχουν στην TargetCity τα οποία αποτελούν τον κύριο λόγο της μεταβολής του προφίλ.

5 Συμπεράσματα

5.1 Σύνοψη αποτελεσμάτων

Ανακεφαλαιώνοντας, το πρόβλημα της σύστασης διαδρομών από πόλη σε πόλη είναι ένα πρόβλημα με πολλαπλές διαστάσεις που απαιτεί εξελιγμένες αλγοριθμικές προσεγγίσεις για την επίλυσή του. Οι τεχνικές που παρουσιάστηκαν στα προηγούμενα κεφάλαια αποτελούν μια πρώτη διερεύνησή του και δίνουν σαφείς ενδείξεις για τις δυνατότητες αλλά και τα εμπόδια στη σχεδίαση πραγματικών συστημάτων συστάσεων τουριστικών διαδρομών.

Αρχικά, μελετήθηκε η συμπεριφορά αλγορίθμων συσταδοποίησης στο πρόβλημα της σύστασης διαδρομών σε ομάδες. Κατά την πειραματική διαδικασία, ένα μεγάλο group τουριστών χωρίστηκε σε μικρότερες υποομάδες με χρήση συσταδοποίησης K-Μέσων και η κάθε ομάδα ακολούθησε διαφορετική διαδρομή, βελτιστοποιημένη με βάση τα προφίλ των μελών της. Παρατηρήθηκε ότι ο διαχωρισμός σε μεγαλύτερο πλήθος από υποομάδες με χρήση συσταδοποίησης συνεισφέρει στη μεγαλύτερη συνολική αύξηση της ικανοποίησης των τουριστών. Συγκεκριμένα, για τουρίστες με μικρή διακύμανση των προφίλ τους, ο διαχωρισμός σε 4-5 ομάδες ήταν αρκετός για να προσφέρει στον εκάστοτε τουρίστα ένα πολύ μεγάλο ποσοστό της ικανοποίησης που θα είχε αν ακολουθούσε μία διαδρομή βελτιστοποιημένη αποκλειστικά για το προσωπικό του προφίλ.

Σε δεύτερο στάδιο, ορίστηκε πλήρως το πρόβλημα της σύστασης διαδρομών από πόλη σε πόλη και σχεδιάστηκαν κάποιοι βασικοί αλγόριθμοι για την επίλυσή του. Επίσης, αναλύθηκε μία πλήρης διαδικασία για την κατασκευή συνόλου δεδομένων κατάλληλου για το συγκεκριμένο πρόβλημα η οποία απαιτεί πιο περίπλοκη επεξεργασία από περιπτώσεις συνόλων δεδομένων που αφορούν το κλασικό πρόβλημα του προσανατολισμού. Οι αλγόριθμοι που προτάθηκαν εξετάστηκαν πειραματικά στα διαθέσιμα δεδομένα με χρήση μιας πλειάδας μετρικών αξιολόγησης και μελετήθηκαν ενδιαφέροντα στοιχεία στη συμπεριφορά τους. Συγκεκριμένα, παρατηρήθηκε ότι οι αλγόριθμοι είναι καθαρά πιο αποτελεσματικοί από κλασικές μεθόδους σύστασης διαδρομών με βάση τη διασημότητα ορισμένων αξιοθέατων. Επίσης, ένα άλλο αποτέλεσμα είναι ότι η μεγιστοποίηση της ικανοποίησης του χρήστη κατά τη σύσταση διαδρομών δεν είναι ισοδύναμη με την ακριβή εκτίμηση του πραγματικού του προφίλ προτιμήσεων και ότι αυτοί οι δύο στόχοι είναι αντικρουόμενοι όταν οι αλγόριθμοι υπολογισμού των διαδρομών βασίζονται σε ευριστικές τεχνικές που δεν επιστρέφουν εγγυημένα βέλτιστα αποτελέσματα.

Γενικότερα, επιβεβαιώθηκε μέσα από τη διαδικασία των πειραμάτων ότι το προφίλ προτιμήσεων του κάθε χρήστη αλλάζει από πόλη σε πόλη ανάλογα με τα διαθέσιμα αξιοθέατα στην κάθε μία. Ακόμα, εντοπίστηκε μία συσχέτιση μεταξύ του προφίλ ενός χρήστη σε μια πόλη με το προφίλ του σε μια άλλη, δίνοντας κίνητρο για την περαιτέρω μελέτη αυτής της συσχέτισης στο μέλλον.

5.2 Μελλοντικές ερευνητικές κατευθύνσεις

Αν και όλα τα αποτελέσματα της παρούσας εργασίας έχουν εξηγηθεί σε μεγάλο βαθμό, κάποια από αυτά αποτελούν κυρίως ενδείξεις και όχι σαφείς εξηγήσεις για τις παραμέτρους του προβλήματος υπό μελέτη. Ο κύριος λόγος γι αυτό είναι οι περιορισμοί που επέβαλε το διαθέσιμο σύνολο δεδομένων σε συνδυασμό με κάποιες παραδοχές που έχουν πραγματοποιηθεί από προηγούμενες εργασίες πάνω στα προβλήματα σύστασης διαδρομών. Μία κατεύθυνση με αρκετό ενδιαφέρον είναι αυτή της εύρεσης της κατάλληλης μεθόδου για την αυτοματοποιημένη εξαγωγή προφίλ τουριστών με χρήση ψηφιακών αποτυπωμάτων από ηλεκτρονικά μέσα κοινωνικής δικτύωσης. Για παράδειγμα, χρήση δεδομένων από πλατφόρμες όπως το Twitter ή το Trip Advisor ενδεχομένως να είναι αποδοτικότερη από το Flickr το οποίο χρησιμοποιήθηκε στην παρούσα εργασία. Επίσης, η απεικόνιση των αξιοθέατων σε διανύσματα μέσω των περιγραφών τους ίσως μπορεί να πραγματοποιηθεί καλύτερα με χρήση κάποιας πιο προηγμένης μεθόδου από την LSI όπως η Word2vec [62].

Ένα άλλο ενδιαφέρον πρόβλημα που πρέπει να μελετηθεί προκειμένου η σύσταση διαδρομών από πόλη σε πόλη να πραγματοποιηθεί μέσω κάποιου ηλεκτρονικού συστήματος με δυνατότητες κλιμάκωσης (scalability) είναι η αποδοτική διαχείριση των πληροφοριών σχετικά με τα προφίλ των χρηστών όπως αποθηκεύονται αφού επισκεφθούν κάποια πόλη. Όπως εξηγήθηκε στο Κεφάλαιο 3, είναι προτιμότερο να μην αποθηκεύονται όλα τα διαθέσιμα προφίλ που έχουν παραχθεί αλλά ένα πολύ μικρότερο σύνολο τα οποία ουσιαστικά θα αποτελούν τα κεντροειδή κάποιων συστάδων των πρωταρχικών προφίλ. Τα ερωτήματα που τίθενται είναι πώς αυτό μπορεί να γίνει με ακρίβεια σε συνεχή χρόνο (online) και με ποιο τρόπο θα καθορίζεται ο αριθμός των απαραίτητων προφίλ χωρίς να χάνεται σημαντική ποσότητα πληροφορίας.

Τέλος, μία κατεύθυνση που μπορεί να προσφέρει χρήσιμη γνώση πάνω στο πρόβλημα της σύστασης διαδρομών από πόλη σε πόλη είναι η πιο λεπτομερής μελέτη της μεταβολής του προφίλ ενός τουρίστα από τη μία πόλη στην άλλη. Συγκεκριμένα, έχει ενδιαφέρον να βρεθεί από τι παράγοντες εξαρτάται αυτή η μεταβολή και να διαπιστωθεί κατά πόσο είναι αιτιοκρατική ή περιλαμβάνει μεγάλο βαθμό τυχαιότητας. Επίσης, ένα άλλο ερώτημα είναι αν αυτή η συνάρτηση απεικόνισης προφίλ από πόλη σε πόλη μπορεί να βρεθεί με χρήση αλγορίθμων μηχανικής μάθησης, υπό την προϋπόθεση της παρουσίας κατάλληλων δεδομένων. Η κατανόηση των ιδιοτήτων της συγκεκριμένης συνάρτησης μπορεί να δώσει τη δυνατότητα για ακριβέστερη εκτίμηση του προφίλ ενός χρήστη και συνεπώς για σύσταση διαδρομών που θα του προσφέρουν μεγαλύτερη ικανοποίηση.

Αναφορές

[1] Pieter Vansteenwegen and Dirk Van Oudheusden. The Mobile Tourist Guide: An OR Opportunity. *OR Insight*, 20(3):21–27, July 2007.

[2] Keith Cheverst, Keith Mitchell, and Nigel Davies. The Role of Adaptive Hypermedia in a Context-aware Tourist GUIDE. *Commun. ACM*, 45(5):47–51, May 2002.

[3] Ander Garcia, Maria Teresa Linaza, Olatz Arbelaitz, and Pieter Vansteenwegen. Intelligent Routing System for a Personalised Electronic Tourist Guide. In Wolfram Höpken, Ulrike Gretzel, and Rob Law, editors, *Information and Communication Technologies in Tourism 2009*, pages 185–197. Springer Vienna, 2009.

[4] Michael Kenteris, Damianos Gavalas, and Daphne Economou. Electronic mobile guides: a survey. *Personal and Ubiquitous Computing*, 15(1):97–111, January 2011.

[5] Pieter Vansteenwegen, Wouter Souffriau, Greet Vanden Berghe, and Dirk Van Oudheusden. The City Trip Planner. *Expert Syst. Appl.*, 38(6):6540–6546, June 2011.

[6] D. Gavalas, M. Kenteris, C. Konstantopoulos, and G. Pantziou. Web application for recommending personalised mobile tourist routes. *IET Software*, 6(4):313–322, August 2012.

[7] Ander Garcia, Olatz Arbelaitz, Maria Teresa Linaza, Pieter Vansteenwegen, and Wouter Souffriau. Personalized Tourist Route Generation. In *Proceedings of the 10th International Conference on Current Trends in Web Engineering*, ICWE'10, pages 486–497, Berlin, Heidelberg, 2010. Springer-Verlag. event-place: Vienna, Austria.

[8] R. Johnson and M. G. Pilcher. The traveling salesman problem, edited by E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B Shmoys, John Wiley & Sons, Chichester, 1985, 463 pp. *Networks*, 18(3):253–254, 1988.

[9] C P Keller and M F Goodchild. The Multiobjective Vending Problem: A Generalization of the Travelling Salesman Problem. *Environment and Planning B: Planning and Design*, 15(4):447–460, December 1988.

[10] Bruce Golden, Larry Levy, and Roy Dahl. Two generalizations of the traveling salesman problem. *Omega*, 9(4):439–441, 1981.

[11] T. Tsiligirides. Heuristic Methods Applied to Orienteering. *Journal of the Operational Research Society*, 35(9):797–809, September 1984.

[12] Bruce L. Golden, Larry Levy, and Rakesh Vohra. The orienteering problem. *Naval Research Logistics (NRL)*, 34(3):307–318, 1987.

[13] Gilbert Laporte and Silvano Martello. The selective travelling salesman problem. *Discrete Applied Mathematics*, 26(2):193–207, March 1990.

[14] R. Ramesh, Yong-Seok Yoon, and Mark H. Karwan. An Optimal Algorithm for the Orienteering Tour Problem. *ORSA Journal on Computing*, 4(2):155–165, May 1992.

[15] Michel Gendreau, Gilbert Laporte, and Frédéric Semet. A branch-and-cut algorithm for the undirected selective traveling salesman problem. *Networks*, 32(4):263–273, 1998.

[16] Matteo Fischetti, Juan José Salazar González, and Paolo Toth. Solving the Orienteering Problem through Branch-and-Cut. *INFORMS Journal on Computing*, 10(2):133–148, May 1998.

[17] Avrim Blum, Shuchi Chawla, David R. Karger, Terran Lane, Adam Meyerson, and Maria Minkoff. Approximation Algorithms for Orienteering and Discounted-Reward TSP. *SIAM J. Comput.*, 37(2):653–670, May 2007.

[18] Nikhil Bansal, Avrim Blum, Shuchi Chawla, and Adam Meyerson. Approximation Algorithms for deadline-TSP and Vehicle Routing with Time-windows. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 166–174, New York, NY, USA, 2004. ACM. event-place: Chicago, IL, USA.

[19] Chandra Chekuri, Nitish Korula, and Martin Pál. Improved Algorithms for Orienteering and Related Problems. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 661–670, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics. event-place: San Francisco, California.

[20] Chandra Chekuri and Martin Pal. A Recursive Greedy Algorithm for Walks in Directed Graphs. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '05, pages 245–253, Washington, DC, USA, 2005. IEEE Computer Society.

[21] Viswanath Nagarajan and R. Ravi. The Directed Orienteering Problem. *Algorithmica*, 60(4):1017–1030, August 2011.

[22] B. L. Golden, Qiwen Wang, and Li Liu. A multifaceted heuristic for the orienteering problem. *Naval Research Logistics (NRL)*, 35(3):359–366, 1988.

[23] R. Ramesh and Kathleen M. Brown. An efficient four-phase heuristic for the generalized orienteering problem. *Computers & Operations Research*, 18(2):151–165, January 1991.

[24] Qiwen Wang, Xiaoyun Sun, Bruce L. Golden, and Jiyou Jia. Using artificial neural networks to solve the orienteering problem. *Annals of Operations Research*, 61(1):111–120, December 1995.

[25] I-Ming Chao, Bruce L. Golden, and Edward A. Wasil. A fast and effective heuristic for the orienteering problem. *European Journal of Operational Research*, 88(3):475–489, February 1996.

[26] Michel Gendreau, Gilbert Laporte, and Frédéric Semet. A tabu search heuristic for the undirected selective travelling salesman problem. *European Journal of Operational Research*, 106(2):539–545, April 1998.

[27] Marisa G. Kantor and Moshe B. Rosenwein. The Orienteering Problem with Time Windows. *The Journal of the Operational Research Society*, 43(6):629–635, 1992.

[28] Fedor V. Fomin and Andrzej Lingas. Approximation algorithms for time-dependent orienteering. *Information Processing Letters*, 83(2):57–62, July 2002.

[29] Anupam Gupta, Ravishankar Krishnaswamy, Viswanath Nagarajan, and R. Ravi. Approximation Algorithms for Stochastic Orienteering. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 1522–1538, Philadelphia, PA, USA, 2012. Society for Industrial and Applied Mathematics. event-place: Kyoto, Japan.

[30] Zachary Friggstad, Sreenivas Gollapudi, Kostas Kollias, Tamas Sarlos, Chaitanya Swamy, and Andrew Tomkins. Orienteering Algorithms for Generating Travel Itineraries. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 180–188, New York, NY, USA, 2018. ACM. event-place: Marina Del Rey, CA, USA.

[31] Senjuti Basu Roy, Gautam Das, Sihem Amer-Yahia, and Cong Yu. Interactive Itinerary Planning. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ICDE '11, pages 15–26, Washington, DC, USA, 2011. IEEE Computer Society.

[32] Pradeep Varakantham, Hala Mostafa, Na Fu, and Hoong Chuin Lau. DIRECT: A Scalable Approach for Route Guidance in Selfish Orienteering Problems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 483–491, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems. event-place: Istanbul, Turkey.

[33] Damianos Gavalas, Charalampos Konstantopoulos, Konstantinos Mastakas, and Grammati Pantziou. A Survey on Algorithmic Approaches for Solving Tourist Trip Design Problems. *Journal of Heuristics*, 20(3):291–328, June 2014.

[34] Aris Anagnostopoulos, Reem Atassi, Luca Becchetti, Adriano Fazzone, and Fabrizio Silvestri. Tour Recommendation for Groups. *Data Min. Knowl. Discov.*, 31(5):1157–1188, September 2017.

[35] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[36] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28:129–136, 1982.

[37] David Arthur and Sergei Vassilvitskii. How Slow is the K-means Method? In *Proceedings of the Twenty-second Annual Symposium on Computational Geometry*, SCG '06, pages 144–153, New York, NY, USA, 2006. ACM. event-place: Sedona, Arizona, USA.

[38] David Arthur and Sergei Vassilvitskii. K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. event-place: New Orleans, Louisiana.

[39] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, September 1967.

[40] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, January 1973.

[41] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, January 1977.

[42] Carl Edward Rasmussen. The Infinite Gaussian Mixture Model. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 554–560, Cambridge, MA, USA, 1999. MIT Press. event-place: Denver, CO.

[43] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, September 1977.

[44] Ken Lang. NewsWeeder: Learning to Filter Netnews. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 331–339. Morgan Kaufmann, San Francisco (CA), January 1995.

[45] Walter Carrer-Neto, María Luisa Hernández-Alcaraz, Rafael Valencia-García, and Francisco García-Sánchez. Social knowledge-based recommender system. Application to the movies domain. *Expert Systems with Applications*, 39(12):10990–11000, September 2012.

[46] Carlos A. Gomez-Uribe and Neil Hunt. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4):13:1–13:19, December 2015.

[47] Shulong Tan, Jiajun Bu, Chun Chen, Bin Xu, Can Wang, and Xiaofei He. Using Rich Social Media Information for Music Recommendation via Hypergraph Model. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7S(1):22:1–22:22, November 2011.

[48] Michael J. Pazzani. A Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial Intelligence Review*, 13(5):393–408, December 1999.

[49] Robin van Meteren. Using Content-Based Filtering for Recommendation. 2000.

[50] Seok Kee Lee, Yoon Ho Cho, and Soung Hie Kim. Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations. *Information Sciences*, 180(11):2142–2155, June 2010.

[51] Keunho Choi, Donghee Yoo, Gunwoo Kim, and Yongmoo Suh. A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. *Electronic Commerce Research and Applications*, 11(4):309–317, July 2012.

[52] Al Mamunur Rashid, George Karypis, and John Riedl. Learning Preferences of New Users in Recommender Systems: An Information Theoretic Approach. *SIGKDD Explor. Newsl.*, 10(2):90–100, December 2008.

[53] Laurent Candillier, Frank Meyer, and Marc Boullé. Comparing State-of-the-Art Collaborative Filtering Systems. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, Lecture Notes in Computer Science, pages 548–562. Springer Berlin Heidelberg, 2007.

[54] Jesus Bobadilla and Francisco Serradilla. The Effect of Sparsity on Collaborative Filtering Metrics. In *Proceedings of the Twentieth Australasian Conference on Australasian Database - Volume 92*, ADC '09, pages 9–18, Darlinghurst, Australia, Australia, 2009. Australian Computer Society, Inc. event-place: Wellington, New Zealand.

[55] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, June 2005.

[56] Robin Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[57] Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu. Automatic Construction of Travel Itineraries Using Social Breadcrumbs. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 35–44, New York, NY, USA, 2010. ACM. event-place: Toronto, Ontario, Canada.

[58] Cristina Ioana Muntean, Franco Maria Nardini, Fabrizio Silvestri, and Ranieri Baraglia. On Learning Prediction Models for Tourists Paths. *ACM Trans. Intell. Syst. Technol.*, 7(1):8:1–8:34, October 2015.

[59] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, pages 465–480, New York, NY, USA, 1988. ACM. event-place: Grenoble, France.

[60] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[61] Andrew P. Bradley. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recogn.*, 30(7):1145–1159, July 1997.

[62] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org, 2014. event-place: Beijing, China.