Πολυτεχνείο Κρήτης
Σχολή Μηχανικών Παραγωγής και Διοίκησης

Ανάπτυξη ενός συστήματος εξατομικευμένων συστάσεων για ηλεκτρονικές κρατήσεις ξενοδοχείων βασισμένο στη πολυκριτήρια ανάλυση αποφάσεων

Διπλωματική Εργασία

Πρατικάκης Εμμανουήλ

Επιβλέπων : Ματσατσίνης Νικόλαος,
Καθηγητής

Χανιά 2017

Πρόλογος

Με αφορμή την ολοκλήρωση της διπλωματικής μου εργασίας και των προπτυχιακών μου σπουδών, θα ήθελα να ευχαριστήσω αρχικά τους γονείς μου και τον αδερφό μου για την έμπρακτη στήριξη που μου παρείχαν καθ' όλη τη διάρκεια των σπουδών μου. Επίσης θα ήθελα να ευχαριστήσω τους φίλους μου για όλες τις όμορφες στιγμές που περάσαμε καθώς και για τις εποικοδομητικές συζητήσεις που είχαμε. Τέλος, θα ήθελα να ευχαριστήσω το καθηγητή και επιβλέπων της παρούσας διπλωματικής εργασίας, κύριο Ματσατσίνη Νικόλαο για τη καθοδήγησή του κατά τη διάρκεια εκπόνησης της εργασίας.

Πρατικάκης Μάνος
Χανιά, Νοέμβριος 2017

Περίληψη

Η ανάπτυξη του διαδικτύου έχει αυξήσει σημαντικά τον όγκο και τη πολυπλοκότητα της διαθέσιμης πληροφορίας, γεγονός που καθιστά το πρόβλημα της λήψης αποφάσεων ακόμη δυσκολότερο. Τα συστήματα συστάσεων βοηθούν στην επίλυση αυτού του προβλήματος εντοπίζοντας προϊόντα ή υπηρεσίες ενδιαφέροντα για κάθε χρήστη του συστήματος μέσα από μια πληθώρα διαθέσιμων. Στη παρούσα εργασία αναπτύσσεται ένα τέτοιο σύστημα που στόχο έχει την εξαγωγή εξατομικευμένων συστάσεων για κάθε χρήστη, λαμβάνοντας υπόψη τις προτιμήσεις του. Αρχικά αναλύεται το θεωρητικό υπόβαθρο που αφορά το πρόβλημα συστάσεων και περιγράφονται οι υφιστάμενες μεθοδολογίες επίλυσής του. Στη συνέχεια αναπτύσσεται μια μεθοδολογία επίλυσης η οποία χρησιμοποιεί τεχνικές της πολυκριτήριας ανάλυσης αποφάσεων και συγκεκριμένα της αναλυτικής-συνθετικής προσέγγισης για τη μοντελοποίηση και δημιουργία διαφορετικών προφίλ χρηστών. Με χρήση τεχνικών από το τομέα των συστημάτων συστάσεων εξάγονται έπειτα εξατομικευμένες συστάσεις για κάθε χρήστη. Τέλος, προκειμένου να αξιολογήσουμε την απόδοσή του, το σύστημα εφαρμόζεται στο τομέα των συστάσεων για ηλεκτρονικές κρατήσεις ξενοδοχείων χρησιμοποιώντας ένα πραγματικό σετ δεδομένων με αξιολογήσεις ξενοδοχείων από την ιστοσελίδα trip advisor.

Abstract

The explosive growth of the Internet has increased the volume and complexity of available information, fact that can make decision-making inefficient. Recommendation or Recommender systems are software tools and techniques intended to help users in decision-making processes by suggesting products or services of potential interest. In this thesis a personalized recommender system is developed considering user preferences. Initially the theoretical framework of the recommendation problem is analyzed and current methodologies and approaches are reviewed. Then a methodological framework based on Multiple Criteria Decision Analysis is proposed. Specifically an aggregation-disaggregation approach is used for user modeling prior to the application of a collaborative filtering algorithm based on multi-criteria ratings. Finally in order to evaluate the system's performance it was applied in the online hotel booking recommendations domain using a data set of hotel reviews from trip-advisor website.

Πίνακας περιεχομένων

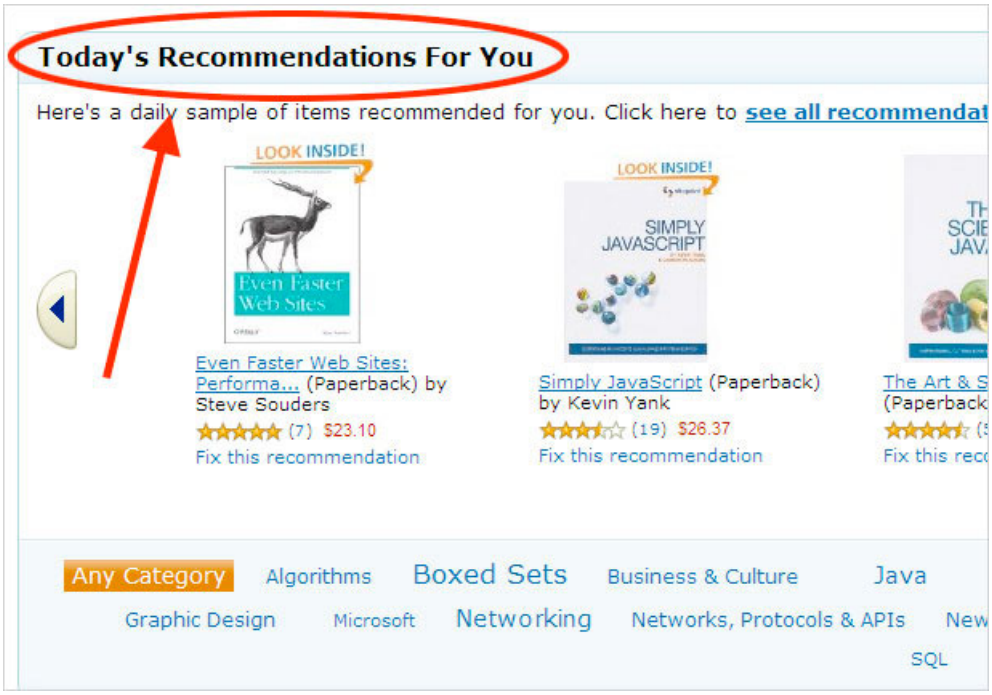
Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1 : Εισαγωγή

1.1 Συστήματα συστάσεων

Η ανάπτυξη των συστημάτων συστάσεων ξεκίνησε από την απλή παρατήρηση ότι κάθε άνθρωπος συχνά βασίζει μεγάλο μέρος της καθημερινότητάς του και διάφορες αποφάσεις που παίρνει, σε προτάσεις που λαμβάνει από το περίγυρό του. Για παράδειγμα συχνά λαμβάνουμε υπόψη προτάσεις από το περίγυρό μας για το πιο βιβλίο να διαβάσουμε ή πιο τουριστικό προορισμό να επισκεφθούμε για να καταλήξουμε σε μια τελική απόφαση. Μεγάλα ηλεκτρονικά καταστήματα επίσης αντιλήφθηκαν την ανάγκη για ανάπτυξη συστημάτων που θα υποβοηθούσαν τους πελάτες στη λήψη μιας απόφασης σχετικά με το πιο προϊόν να αγοράσουν. Ιδιαίτερα τα τελευταία χρόνια, η ανάπτυξη του διαδικτύου έχει αυξήσει σημαντικά τον όγκο και τη πολυπλοκότητα της διαθέσιμης πληροφορίας. Οι χρήστες του διαδικτύου λαμβάνουν ποικίλες πληροφορίες προτού καταφέρουν να απομονώσουν μόνο τις απαραίτητες, γεγονός που καθιστά το πρόβλημα της λήψης αποφάσεων ακόμη δυσκολότερο. Τα συστήματα συστάσεων βοηθούν στην επίλυση αυτού του προβλήματος εντοπίζοντας προϊόντα ή υπηρεσίες ενδιαφέροντα για κάθε χρήστη του συστήματος μέσα από μια πληθώρα διαθέσιμων. Πλέον τα συστήματα συστάσεων παίζουν σημαντικό ρόλο σε μεγάλες ιστοσελίδες όπως η Amazon, YouTube, Yahoo, Tripadvisor, Netflix κ.α. ενώ αποδεδειγμένα πλέον προσφέρουν αύξηση των επισκεπτών των ηλεκτρονικών καταστημάτων και μεγαλύτερη ικανοποίηση στους πελάτες. Τα συστήματα συστάσεων έχουν ως κύριο στόχο να βοηθήσουν χρήστες που δεν έχουν αρκετή προσωπική εμπειρία ή πληροφόρηση για να αξιολογήσουν τη πληθώρα των εναλλακτικών αντικειμένων που βρίσκεται στο διαδίκτυο. Αντικείμενο είναι ο όρος που χρησιμοποιείται για να εκφράσουμε το αποτέλεσμα ενός συστήματος συστάσεων. Συνήθως ένα σύστημα συστάσεων προτείνει ένα είδος αντικειμένων, το οποίο μπορεί να είναι ένα βιβλίο,μια ταινία,ένα τραγούδι, ένα ξενοδοχείο ή και ένας τουριστικός προορισμός. Οι συστάσεις που εξάγονται από το σύστημα είναι συνήθως εξατομικευμένες λίστες αντικειμένων, οπότε κάθε χρήστης ή κάθε ομάδα χρηστών λαμβάνει διαφορετικές συστάσεις. Υπάρχουν βέβαια και μη-εξατομικευμένες συστάσεις οι οποίες είναι ευκολότερο να εξαχθούν , αλλά δεν λαμβάνουν υπόψη τα ιδιαίτερα χαρακτηριστικά του κάθε χρήστη. Για τη δημιουργία εξατομικευμένων συστάσεων δίδεται ιδιαίτερη σημασία στη δημιουργία προφίλ χρηστών καθώς αυτά αποτελούν αναπαραστάσεις των ιδιαίτερων χαρακτηριστικών και αναγκών των χρηστών.



Εικόνα 1.1-1: Εμφάνιση του συστήματος συστάσεων της ιστοσελίδας amazon

Πηγή: www.webdesignerdepot.com/2009/10/an-analysis-of-the-amazon-shopping-experience

1.2 Σκοπός της εργασίας

Στη παρούσα εργασία γίνεται ανάπτυξη ενός συστήματος συστάσεων ξενοδοχείων για ηλεκτρονικές κρατήσεις με δεδομένα αξιολογήσεων από την ιστοσελίδα trip advisor. Το σύστημα αναπτύσσεται χρησιμοποιώντας γνωστές τεχνικές των συστημάτων συστάσεων και της πολυκριτήριας ανάλυσης αποφάσεων. Πιο συγκεκριμένα, αναπτύσσεται μια μεθοδολογία η οποία λαμβάνει υπόψη αξιολογήσεις χρηστών πάνω σε διαφορετικά κριτήρια για τη δημιουργία διαφορετικών προφίλ, με χρήση αλγορίθμων της πολυκριτήριας ανάλυσης αποφάσεων. Η πολυκριτήρια ανάλυση αποφάσεων έχει ως στόχο την αντιμετώπιση ενός προβλήματος που παρουσιάζεται κατά την εξέταση όλων των πιθανών παραμέτρων και κριτηρίων που επηρεάζουν τη λήψη μιας τελικής απόφασης. Αρχικά γίνεται μια επισκόπηση των βασικών μεθοδολογιών και προσεγγίσεων που χρησιμοποιούνται για την επίλυση του προβλήματος συστάσεων καθώς και των προβλημάτων που αντιμετωπίζουν οι υπάρχουσες μεθοδολογίες. Έπειτα αναλύεται το θεωρητικό υπόβαθρο που αφορά το πρόβλημα και γίνεται επεξήγηση των βασικών εννοιών που θα χρησιμοποιηθούν για την ανάπτυξη του συστήματος. Στο κεφάλαιο 4 αναλύεται η προτεινόμενη μεθοδολογία για την επίλυση του προβλήματος και έπειτα γίνεται αξιολόγηση του συστήματος με χρήση ενός συνόλου πραγματικών δεδομένων. Τέλος γίνεται μια επισκόπηση των μελλοντικών ενεργειών και προεκτάσεων που μπορούν να εφαρμοσθούν για τη βελτίωση του συστήματος.

Κεφάλαιο 2 : Υφιστάμενη κατάσταση

2.1 Εισαγωγή

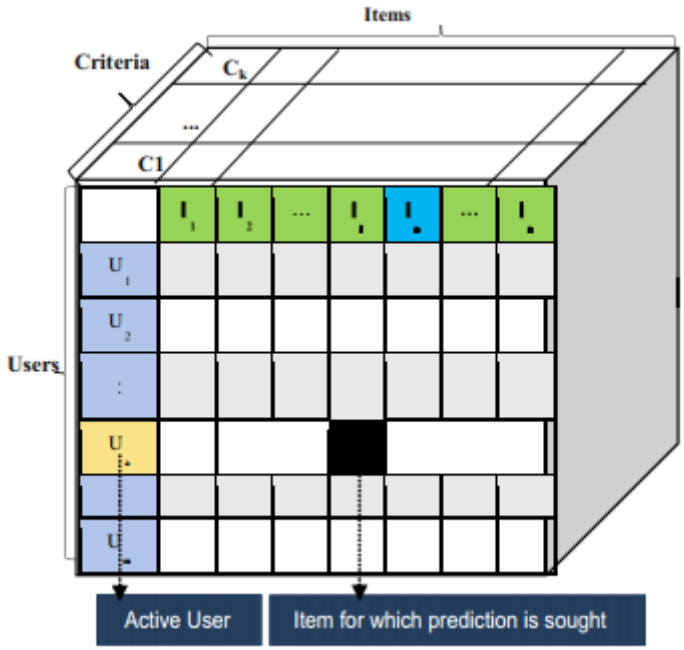
Η μελέτη των συστημάτων συστάσεων σαν ξεχωριστό τομέα ξεκίνησε γύρω στα μέσα του 1990, αρκετά αργότερα σε σχέση με μελέτες γύρω από άλλα πληροφοριακά συστήματα. Τα τελευταία 20 χρόνια βέβαια τα συστήματα συστάσεων απασχολούν πολλούς ερευνητές και έχουν αναπτυχθεί αρκετές μεθοδολογίες για την επίλυση του προβλήματος συστάσεων. Υπάρχουν όμως ακόμη αρκετά ανοικτά ζητήματα και προοπτικές βελτίωσης και ανάπτυξης νέων μεθοδολογιών. Η πλειονότητα των συστημάτων συστάσεων που χρησιμοποιούνται, βασίζεται σε μια αριθμητική αξιολόγηση η οποία αντιπροσωπεύει εξ' ολοκλήρου την άποψη του χρήστη (πελάτη) για ένα αντικείμενο (προϊόν / υπηρεσία) . Ως εκ' τούτου το πρόβλημα στη πιο απλή του μορφή αποκτά δύο πιθανές διαστάσεις: Χρήστες και Αντικείμενα. Το πρόβλημα αυτό συνήθως μοντελοποιείται με ένα πίνακα m×n διαστάσεων, όπου m οι χρήστες του συστήματος και n τα αντικείμενα και αφορά το προσδιορισμό μιας βαθμολογίας R(u,i) για κάθε αντικείμενο i που δεν έχει αξιολογήσει ο χρήστης u (Πίνακας 2.1-1). Πολλές μεθοδολογίες που χρησιμοποιούν μία μονή βαθμολογία ,έχουν καταφέρει να λύσουν επιτυχημένα το πρόβλημα συστάσεων σε αρκετές εφαρμογές.

	Item 1	Item 2	Item 3	...	Item n
User 1	2	3	?	...	5
User 2	?	4	3	...	?
User 3	3	2	?	...	3
...	...	...	...	...	...
User m	1	?	5	...	4

Πίνακας 2.1-1: Πίνακας χρηστών × αντικειμένων στο πρόβλημα σύστασης με αξιολόγηση ενός χαρακτηριστικού

Πηγή: www.statr.me/2016/07/recommender-system-using-parallel-matrix-factorization

Παρόλα αυτά , περιγράφοντας ένα αντικείμενο με πολλαπλά χαρακτηριστικά και λαμβάνοντας υπ' όψη τις προτιμήσεις των χρηστών πάνω σε αυτά, μπορεί να βοηθήσει στη δημιουργία πιο αποδοτικών συστημάτων. Πλέον όλο και περισσότερα ηλεκτρονικά καταστήματα ζητούν από τους πελάτες τους την αξιολόγηση των αντικειμένων με βάση διάφορα κριτήρια,οπότε το πρόβλημα γίνεται περισσότερο πολύπλοκο (Πίνακας 2.1-2). Ο αναλυτής, πρέπει να λάβει υπόψη του, τη σημασία που έχουν για κάθε χρήστη τα κριτήρια αυτά, αλλά και να βρει τρόπους για να συμπεριλάβει τη πληροφορία που δέχεται απ' τις αξιολογήσεις των κριτηρίων στη διαδικασία ανάπτυξης ενός συστήματος συστάσεων.



Πίνακας 2.1-2: Πίνακας χρηστών × αντικειμένων × κριτηρίων στο πρόβλημα σύστασης με αξιολόγηση k κριτηρίων

Πηγή: Mehrbakhsh N., Othman I. (2015). A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS, *Electronic Commerce Research and Applications.*

2.2 Υφιστάμενες μεθοδολογίες επίλυσης

Πολλοί ερευνητές επικεντρώνονται στο να αναπτύξουν μεθοδολογίες οι οποίες εκμεταλλεύονται με το πιο αποδοτικό τρόπο τη πληροφορία που λαμβάνεται μέσω των πολυκριτήριων βαθμολογιών για τη δημιουργία καλύτερων συστάσεων. Οι G. Adomavicius και YoungOk Kwon (“New Recommendation Techniques for Multi-Criteria Rating Systems”,2007) επιλύουν το πρόβλημα συστάσεων για πολυκριτήρια συστήματα βαθμολόγησης ,προτείνοντας δύο μεθοδολογίες. Σύμφωνα με τη πρώτη στοχεύουν στον υπολογισμό ενός μέτρου ομοιότητας μεταξύ χρηστών το οποίο υπολογίζεται μέσω της απόστασης (π.χ. απόλυτης διαφοράς) που έχουν οι βαθμολογήσεις των χρηστών στα k κριτήρια. Μέσω της μεθοδολογίας αυτής προσεγγίζεται με μεγαλύτερη ακρίβεια η έννοια της ομοιότητας μεταξύ χρηστών η οποία θα επεξηγηθεί αναλυτικά σε επόμενα κεφάλαια. Στη δεύτερη μεθοδολογία που εξετάζουν, αντιμετωπίζουν το πρόβλημα των k κριτηρίων ως k διαφορετικά προβλήματα , στόχος των οποίων είναι να υπολογίσουν σε πρώτο στάδιο την ομοιότητα μεταξύ χρηστών λαμβάνοντας υπ’ όψη μόνο το συγκεκριμένο κριτήριο. Έχοντας ως δεδομένο τις ομοιότητες μεταξύ χρηστών με βάση τα k κριτήρια, υπολογίζουν μέσω τυπικών αλγορίθμων των συστημάτων συστάσεων, πιθανές βαθμολογίες των χρηστών σε κάθε κριτήριο. Το τελικό αποτέλεσμα βασίζεται στη λογική ότι οι k πολυκριτήριες βαθμολογίες που

υπολογίσθηκαν, συνδέονται με τη συνολική προτίμηση του κάθε χρήστη μέσω μιας συνθετικής συνάρτησης. Οι δύο μεθοδολογίες εξετάσθηκαν στη περίπτωση ενός συστήματος σύστασης ταινιών, με δεδομένα από την Yahoo! Movies τα οποία περιέχουν αξιολογήσεις ταινιών με μία συνολική βαθμολογία αλλά και βαθμολογία σε 5 επιπλέον κριτήρια: σενάριο, ηθοποιοί, σκηνοθεσία και οπτικά εφέ. Και οι δυο μεθοδολογίες προσφέρουν μεγαλύτερη ακρίβεια στις συστάσεις τους, σε σχέση με τις συστάσεις που θα έκανε ένα σύστημα το οποίο θα χρησιμοποιούσε μόνο τη συνολική βαθμολογία των χρηστών.

Ενώ οι μεθοδολογίες των Adomavicius & Kwon προσφέρουν συστάσεις λαμβάνοντας υπόψη τη προτίμηση των χρηστών σε διαφορετικά κριτήρια, γίνεται επιπλέον ανάλυση από πολλούς ερευνητές για το πως ένα σύστημα συστάσεων που χρησιμοποιεί πολυκριτήριες αξιολογήσεις μπορεί να προσφέρει περισσότερο εξατομικευμένες συστάσεις. Στις περισσότερες υφιστάμενες μεθοδολογίες, οι αναλυτές στοχεύουν στη δημιουργία διαφορετικών ομάδων χρηστών με παρόμοιες προτιμήσεις και ανάγκες ή ομάδων παρόμοιων αντικειμένων . Η λογική πίσω αυτό βασίζεται στην έννοια της τμηματοποίησης της αγοράς. Η υποδιαίρεση της συνολικής αγοράς σε ομοιογενή τμήματα πελατών, που το καθένα μπορεί εύκολα να επιλεγεί και να αντιμετωπιστεί ως μια μικρότερη εξειδικευμένη αγορά, μπορεί να οδηγήσει στη καλύτερη κατανόηση των αναγκών και προτιμήσεων των πελατών. Δεδομένου λοιπόν ότι οι χρήστες δίδουν διαφορετική σημασία στα διάφορα κριτήρια όταν τους ζητείται να αξιολογήσουν ένα αντικείμενο, η δημιουργία ομάδων χρηστών με παρόμοιες προτιμήσεις μπορεί να οδηγήσει στη δημιουργία περισσότερο εξατομικευμένων συστημάτων. Ένα σύστημα το οποίο επιλύει το πρόβλημα συστάσεων με αυτή τη λογική είναι αυτό των K.Lakiotaki, N.Matsatinis & A. Tsoukias (“Multi-Criteria User Modeling in Recommender Systems”,2011).Στο σύστημα αυτό δημιουργούνται διαφορετικές ομάδες από προφίλ χρηστών και έπειτα εφαρμόζονται αλγόριθμοι συνεργατικού φιλτραρίσματος για την εξαγωγή συστάσεων. Τα προφίλ που δημιουργούνται είναι αποτέλεσμα μιας διαδικασίας μοντελοποίησης των χρηστών η οποία βασίζεται στη πολυκριτήρια ανάλυση αποφάσεων. Το σύστημα εξετάσθηκε στο σετ δεδομένων με αξιολογήσεις ταινιών της yahoo που περιγράφηκε παραπάνω , ενώ επέφερε καλύτερα αποτελέσματα σε σχέση με τα συστήματα αξιολόγησης που χρησιμοποιούν αξιολογήσεις σε ένα χαρακτηριστικό , αλλά και σε σχέση με άλλες μεθοδολογίες που εκμεταλλεύονται τις πολυκριτήριες αξιολογήσεις. Με βάση αυτό το σύστημα θα αναπτυχθεί η μεθοδολογία του κεφαλαίου 4, για τη δημιουργία συστήματος συστάσεων για ηλεκτρονικές κρατήσεις ξενοδοχείων. Παρόμοιο σύστημα το οποίο αναπτύχθηκε και εφαρμόζει τεχνικές ομαδοποίησης των χρηστών με στόχο την εξατομίκευση, είναι αυτό των Liwei Liu, Nikolay Mehandjiev & Dong-Ling Xu (“Multi-Criteria Service Recommendation Based on User Criteria Preferences”,2011). Σε αυτό το σύστημα εφαρμόζεται συσταδοποίηση (clustering) των χρηστών, με βάση τα κριτήρια που θεωρούν πιο σημαντικά (significant criteria) για να καταλήξουν στη τελική αξιολόγηση. Για την εύρεση των σημαντικών κριτηρίων εφαρμόζεται ένα μοντέλο γραμμικής παλινδρόμησης (linear regression model) , που στόχο έχει το προσδιορισμό μιας συνθετικής συνάρτησης της μορφής $r_o= f(r_1,r_2,\ldots,r_k)$, όπου r_o η συνολική βαθμολογία και $r_1,\ldots,r_k$ οι βαθμολογίες στα k κριτήρια, για κάθε χρήστη. Μέσω αυτής της συνάρτησης προκύπτουν συντελεστές βαρών οι οποίοι προσδιορίζουν τη σημασία που δίδει ο κάθε χρήστης σε κάθε κριτήριο. Το συγκεκριμένο σύστημα εφαρμόστηκε στο τομέα ηλεκτρονικών κρατήσεων ξενοδοχείων με χρήση ενός σετ δεδομένων με αξιολογήσεις ξενοδοχείων από την ιστοσελίδα trip

advisor. Στο συγκεκριμένο σετ οι χρήστες έχουν βαθμολογήσει τα ξενοδοχεία συνολικά αλλά και σε 5 επιμέρους κριτήρια : Value, Rooms, Location, Cleanliness και Service. Και σε αυτή τη περίπτωση το πολυκριτήριο σύστημα συστάσεων που αναπτύχθηκε είχε καλύτερη επίδοση σε σχέση με το σύστημα που εφαρμόσθηκε στα ίδια δεδομένα αλλά αγνοούσε τις αξιολογήσεις στα επιμέρους κριτήρια.

Ιδιαίτερα τα τελευταία χρόνια , όπου η χρησιμότητα των συστημάτων συστάσεων έχει αποδειχθεί έμπρακτα στο τομέα των επιχειρήσεων (e-commerce sites), δίδεται μεγάλη έμφαση στην ανάπτυξη αλγορίθμων και μεθοδολογιών που είναι σε θέση όχι μόνο να προσφέρουν αξιόπιστες συστάσεις αλλά και να ελαχιστοποιούν το χρόνο και το υπολογιστικό κόστος. Πολλοί αλγόριθμοι αποδεδειγμένα προσφέρουν αξιόπιστες προβλέψεις, αλλά στη πράξη, αποδίδουν πολύ αργά για να εφαρμοστούν σε ιστοσελίδες ηλεκτρονικών καταστημάτων που λειτουργούν σε πραγματικό χρόνο. Ένας από τους αλγόριθμους που έχει αναπτυχθεί λαμβάνοντας υπόψη την ελαχιστοποίηση του υπολογιστικού κόστους και χρόνου, είναι αυτός των Sarwar B., Karypis G., Konstan J. & Riedl J. (“Item-based collaborative filtering recommendation algorithms” ,2001). Στον αλγόριθμο αυτό υπολογίζονται ομοιότητες και εξετάζονται συσχετίσεις μεταξύ αντικειμένων. Οι συστάσεις που εξάγονται είναι σετ αντικειμένων τα οποία έχουν τη μεγαλύτερη ομοιότητα με τα αντικείμενα τα οποία έχει αξιολογήσει θετικά ο χρήστης στο παρελθόν. Τα μεγάλα ηλεκτρονικά καταστήματα αλλά και οι ιστοσελίδες που παρέχουν υπηρεσίες (Amazon , E-bay, Booking , Trip - Advisor), έχουν πολλούς περισσότερους χρήστες απ’ ότι προϊόντα. Σε αυτές τις περιπτώσεις ένας item-based αλγόριθμος έχει πιο γρήγορη ανταπόκριση στην εξαγωγή συστάσεων, ειδικά όταν οι σχέσεις μεταξύ των αντικειμένων έχουν υπολογισθεί εκ των προτέρων. Άλλοι αλγόριθμοι με πρακτικό ενδιαφέρον για online εφαρμογή σε ηλεκτρονικά καταστήματα χρησιμοποιούν τεχνικές μείωσης των διαστάσεων (π.χ. principal component analysis,singular value decomposition) για να εξάγουν ακριβείς συστάσεις (Billsus & Pazzani 1998, Sarwar 2002). Αυτές οι τεχνικές βασίζονται στη λογική ότι ο πίνακας χρηστών × αντικειμένων , μπορεί να έχει επιπλέον, ανεξάρτητες από τα υπόλοιπα δεδομένα διαστάσεις οι οποίες “κρύβουν” πληροφορίες για τις προτιμήσεις των χρηστών. Με τη χρήση αυτών των τεχνικών, μειώνονται οι διαστάσεις των δεδομένων, ώστε να περιέχουν μόνο την απαραίτητη πληροφορία σχετικά με τις σχέσεις μεταξύ χρηστών ή αντικειμένων. Η διαδικασία μείωσης των διαστάσεων θεωρείται σχετικά χρονοβόρα, εφόσον όμως πραγματοποιηθεί όλες οι παραδοσιακές τεχνικές και αλγόριθμοι συστάσεων εξάγουν αποτελέσματα γρηγορότερα.

2.3 Ανοιχτά ζητήματα στα συστήματα συστάσεων

Σε ένα σχετικά νέο τομέα έρευνας όπως τα συστήματα συστάσεων, υπάρχουν πολλές προεκτάσεις και ζητήματα τα οποία δεν έχουν ακόμη αναλυθεί επαρκώς. Πέραν από τα πολυκριτήρια συστήματα συστάσεων, υπάρχουν ακόμη πολλές άλλες κατευθύνσεις ως προς τις οποίες γίνεται έρευνα. Σε αυτή τη παράγραφο γίνεται μια επισκόπηση των θεμάτων αυτών.

2.3.1 Δημιουργία συστάσεων σε ομάδες χρηστών (Group recommendations)

Οι άνθρωποι πολλές φορές αναζητούν αντικείμενα τα οποία αρέσουν ή είναι χρήσιμα σε μία ομάδα στην οποία ανήκουν. Για παράδειγμα, οι τουρίστες κατά πλειοψηφία ταξιδεύουν σε ομάδες,

αλλά οι περισσότερες υπάρχουσες μεθοδολογίες στοχεύουν στη σύσταση προορισμών ή ξενοδοχείων κ.τ.λ. σε ένα μόνο άτομο. Θα πρέπει λοιπόν να αναπτυχθούν συστήματα τα οποία να λαμβάνουν υπόψη περισσότερο τη προτίμηση του συνόλου της ομάδας παρά των ξεχωριστών ατόμων που ανήκουν σε αυτή. Η δυσκολία στην ανάπτυξη τέτοιων συστημάτων βρίσκεται στην εύρεση ενός μοντέλου που μετρά τις προτιμήσεις μιας ομάδας και συστήνει αντικείμενα τα οποία θα είναι χρήσιμα για τη πλειοψηφία των ατόμων της ομάδας. Η διαδικασία λήψης αποφάσεων μέσα σε μία ομάδα είναι αρκετά δυσκολότερη και πολυπλοκότερη απ' ότι η λήψη ατομικών αποφάσεων. Πολλές φορές τα άτομα που ανήκουν σε μια ομάδα έχουν διαφορετικά ή και αντικρουόμενα ενδιαφέροντα.

2.3.2 Κινητά συστήματα συστάσεων (Mobile recommendation systems)

Τα κινητά συστήματα συστάσεων είναι συστήματα που βοηθούν ένα χρήστη ή μια ομάδα χρηστών στο να αποφασίσει για διάφορα ζητήματα τα οποία αντιμετωπίζει όταν βρίσκεται εν' κινήσει. Τέτοια ζητήματα μπορεί να είναι οι διάφορες αποφάσεις που χρειάζεται να πάρουν οι καταναλωτές όταν ψωνίζουν σε καταστήματα ή οι αποφάσεις που παίρνουν οι τουρίστες όταν επισκέπτονται ένα νέο μέρος (επιλογή εστιατορίου, μουσείου, ξενοδοχείου). Υπάρχει λοιπόν ανάγκη ανάπτυξης συστημάτων που να λαμβάνουν υπόψη πληροφορίες όπως τοποθεσία,ώρα ή και το καιρό για χρήστες που βρίσκονται εν' κινήσει. Ένα τέτοιο σύστημα είναι το Google Now το οποίο λαμβάνει πληροφορίες από κινητές συσκευές για να προσφέρει στους χρήστες εξατομικευμένες συστάσεις.

2.3.3 Συστήματα συστάσεων για εμπορική χρήση (Recommendation systems for commercial use)

Τα συστήματα συστάσεων αναπτύσσονται για να βοηθήσουν τους χρήστες στις επιλογές που παίρνουν καθημερινά με σύσταση αντικειμένων τα οποία θα είναι χρήσιμα για αυτούς. Στη περίπτωση όμως των επιχειρήσεων, κύριος σκοπός των συστημάτων συστάσεων είναι η επίτευξη οικονομικών στόχων και η επιτυχία ενός συστήματος εξαρτάται από το πόσο κέρδος επέφερε στην επιχείρηση η υλοποίησή του. Ο ρόλος λοιπόν που παίζουν τα συστήματα συστάσεων που αναπτύσσονται για εμπορική χρήση και για υποβοήθηση των χρηστών είναι διαφορετικός. Για παράδειγμα ένα σύστημα συστάσεων προϊόντων που αναπτύσσεται για μια επιχείρηση μπορεί να έχει ως στόχο την αύξηση των πωλήσεων προϊόντων μιας συγκεκριμένης κατηγορίας, την αύξηση του αριθμού πελατών ή καλύτερη διαφήμιση προϊόντων που δεν αγοράζονται συχνά. Ιδιαίτερα τα τελευταία χρόνια αναπτύσσονται τεχνικές και αλγόριθμοι συστάσεων για επιχειρήσεις οι οποίοι έχουν ως πρωταρχικούς στόχους:

- Αύξηση του αριθμού των πωλήσεων.
- Ανάπτυξη σχέσης εμπιστοσύνης μεταξύ επιχείρησης και πελατών.
- Αύξηση της ικανοποίησης των πελατών.
- Καλύτερη κατανόηση των αναγκών των πελατών.

Κεφάλαιο 3 : Θεωρητικό υπόβαθρο

3.1 Βασικές προσεγγίσεις στα συστήματα συστάσεων

Οι συνηθέστερες τεχνικές που χρησιμοποιούν τα συστήματα συστάσεων μέχρι σήμερα ανήκουν σε δύο κύριες κατηγορίες, το συνεργατικό φιλτράρισμα (collaborative filtering) και το φιλτράρισμα με βάση το περιεχόμενο (content-based filtering).Η διαφοροποίηση των δύο τεχνικών έγκειται στη λογική σύμφωνα με την οποία θα συσταθούν αντικείμενα στο κάθε χρήστη. Πέραν από αυτές τις δυο περιπτώσεις, τα συστήματα συστάσεων μπορούν να κατηγοριοποιηθούν και σε επιπλέον κατηγορίες λόγω κυρίως της διαφορετικής φύσης των προβλημάτων που επιλύουν. Στη παράγραφο αυτή θα περιγραφούν έξι συνολικά κατηγορίες με έμφαση στις δύο πρώτες που αναφέρθηκαν παραπάνω.

3.1.1 Συνεργατικό φιλτράρισμα (Collaborative Filtering)

Το συνεργατικό φιλτράρισμα είναι μια δημοφιλής τεχνική σύμφωνα με την οποία οι συστάσεις για κάθε χρήστη προκύπτουν με βάσει τις προτιμήσεις άλλων χρηστών με παρόμοια συμπεριφορά. Συνήθως, αυτές οι μέθοδοι δεν χρησιμοποιούν πληροφορίες που έχουν να κάνουν με το περιεχόμενο των αντικειμένων προς σύσταση αυτό καθ' αυτό, αλλά βασίζονται στις γνώμες των χρηστών (συνήθως αξιολογήσεις). Σε ένα τυπικό σύστημα συστάσεων που βασίζεται στο συνεργατικό φιλτράρισμα τα δεδομένα σχετικά με τις προτιμήσεις των χρηστών καταγράφονται σε ένα πίνακα χρηστών-αντικειμένων του οποίου οι γραμμές συνήθως αναφέρονται σε διαφορετικούς χρήστες και κάθε εγγραφή υποδεικνύει τη προτίμηση του χρήστη u για το αντικείμενο i. Οι τεχνικές που ανήκουν σε αυτή τη κατηγορία λοιπόν χρησιμοποιούν τις βαθμολογήσεις των χρηστών για να καθορίσουν τη σχέση μεταξύ τους και να εξάγουν προβλέψεις σε βαθμολογήσεις που θα έδινε κάθε χρήστης σε κάποιο νέο αντικείμενο. Οι τεχνικές συνεργατικού φιλτραρίσματος χρησιμοποιούνται σε πολλές εφαρμογές συστημάτων συστάσεων. Δημοφιλείς ιστοσελίδες που χρησιμοποιούν τέτοιες τεχνικές είναι οι LinkedIn, Facebook, Twitter και Amazon.

Συνήθως τα συστήματα που χρησιμοποιούν το συνεργατικό φιλτράρισμα διαχωρίζονται με βάση τους αλγορίθμους που χρησιμοποιούν σε βασιζόμενα στη μνήμη (memory-based) και βασιζόμενα στο μοντέλο (model-based) συστήματα. Τα συστήματα που ανήκουν στη πρώτη κατηγορία εμπεριέχουν αλγορίθμους που χρησιμοποιούν το σύνολο των αντικειμένων που έχει βαθμολογήσει κάθε χρήστης στο παρελθόν για να εξάγουν για αυτόν συστάσεις. Η δεύτερη κατηγορία model-based συστημάτων χρησιμοποιεί αλγόριθμους οι οποίοι εκμεταλλεύονται μόνο ένα μέρος των αντικειμένων που έχει βαθμολογήσει ο κάθε χρήστης για να δημιουργήσουν ένα μοντέλο το οποίο μετά χρησιμοποιείται για να γίνουν προβλέψεις βαθμολογιών του χρήστη σε νέα αντικείμενα.

Οι τεχνικές συνεργατικού φιλτραρίσματος μπορεί να κατηγοριοποιηθούν επιπλέον, σε βασιζόμενες στο χρήστη (user-based) και βασιζόμενες στο αντικείμενο (item-based). Οι βασιζόμενες στο χρήστη τεχνικές υπολογίζουν ομοιότητα μεταξύ χρηστών, ενώ οι βασιζόμενες στο

αντικείμενο τεχνικές υπολογίζουν ομοιότητα μεταξύ αντικειμένων, χρησιμοποιώντας τις βαθμολογήσεις των χρηστών.

Η έννοια της ομοιότητας μπορεί να καθοριστεί με πολλούς τρόπους. Τα δύο πιο διαδεδομένα μέτρα ομοιότητας είναι το μέτρο cosine similarity και ο συντελεστής συσχέτισης Pearson ,τα οποία δίδονται από τις εξισώσεις 3.1.1-1 και 3.1.1-2 αντίστοιχα. Όπως φαίνεται στην εξίσωση 3.1.1-1 η τιμή της ομοιότητας μεταξύ δύο χρηστών u,u' εξαρτάται από τις βαθμολογήσεις R(u,i) και R(u',i) που έχουν δώσει οι χρήστες στο παρελθόν,για κάθε αντικείμενο i που ανήκει στο σύνολο των κοινών αντικειμένων I(u,u').Το σύνολο I(u,u') αναφέρεται σε όλα τα αντικείμενα που έχουν βαθμολογήσει και οι δύο χρήστες.

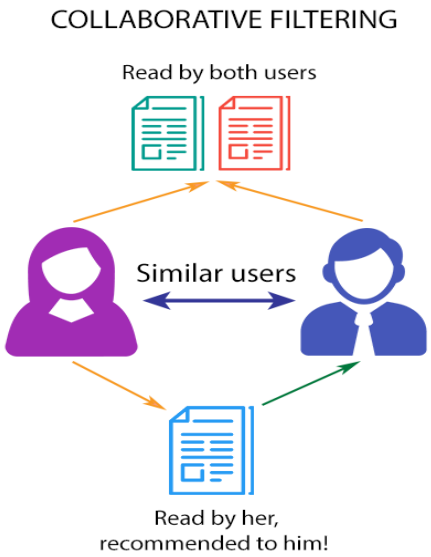
$$\mathrm{simil}(u,u')=\frac{\sum_{i\in I(u,u')} R(u,i)\cdot R(u',i)}{\sqrt{\sum_{i\in I(u,u')} R(u,i)^2}\cdot\sqrt{\sum_{i\in I(u,u')} R(u',i)^2}} \qquad \textbf{3.1.1-1}$$

Ομοίως χρησιμοποιώντας το συντελεστή Pearson υπολογίζουμε το simil(u,u') όπως φαίνεται στην εξίσωση 3.1.1-2

$$\mathrm{simil}(u,u')=\frac{\sum_{i\in I(u,u')} (R(u,i)-\bar{R}(u))\cdot(R(u',i)-\bar{R}(u'))}{\sqrt{\sum_{i\in I(u,u')} (R(u,i)-\bar{R}(u))^2}\cdot\sqrt{\sum_{i\in I(u,u')} (R(u',i)-\bar{R}(u'))^2}} \qquad \textbf{3.1.1-2}$$

Το μέτρο ομοιότητας με βάση το συντελεστή του Pearson κυμαίνεται από 1 (θετική συσχέτιση), για χρήστες με απόλυτη συμφωνία, έως -1 (αρνητική συσχέτιση) για χρήστες με απόλυτη διαφωνία.

Με βάση τις τεχνικές του συνεργατικού φιλτραρίσματος και σε συνδυασμό με μεθόδους εξόρυξης γνώσης από δεδομένα, μηχανικής μάθησης (data mining, machine learning) και στατιστικής, έχουν αναπτυχθεί πολλοί αλγόριθμοι συστάσεων.



Σχήμα 3.1-1: Αναπαράσταση ενός συστήματος συνεργατικού φιλτραρίσματος βασιζόμενο στο χρήστη

Πηγή: www.marutitech.com/recommendation-engine-benefits

Τα συστήματα συνεργατικού φιλτραρίσματος έχουν τα παρακάτω πλεονεκτήματα :

- Δεν απαιτείται γνώση του περιεχομένου των αντικειμένων π.χ. στη περίπτωση των ξενοδοχείων η τοποθεσία,το όνομα του ξενοδοχείου,ο ιδιοκτήτης κ.τ.λ. Απαιτείται μόνο η αξιολόγηση των αντικειμένων από τους χρήστες του συστήματος. Συνεπώς το συνεργατικό φιλτράρισμα μπορεί να εφαρμοστεί εν' γένει σε κάθε είδος αντικειμένου , π.χ. ξενοδοχεία, βιβλία, ταινίες, τραγούδια κ.τ.λ.
- Οι τεχνικές collaborative filtering μπορούν να εφαρμοστούν σε μεγάλο όγκο δεδομένων καθώς δεν απαιτείται παρέμβαση από τον άνθρωπο για το προσδιορισμό των χαρακτηριστικών των αντικειμένων.
- Μπορούν να πραγματοποιηθούν συστάσεις αντικειμένων τελείως διαφορετικών από αυτά που έχει βαθμολογήσει ο χρήστης στο παρελθόν.
- Δεν απαιτείται συγκεκριμένη γνώση πάνω στο τομέα που απασχολούν τα αντικείμενα που μελετώνται,για την επεξήγηση και ανάλυση των χαρακτηριστικών τους.

3.1.2 Φιλτράρισμα με βάση το περιεχόμενο (Content-based Filtering)

Τα συστήματα που χρησιμοποιούν φιλτράρισμα με βάση το περιεχόμενο αναλύουν ένα σύνολο από πληροφορίες, για τα χαρακτηριστικά των αντικειμένων που έχουν βαθμολογηθεί προηγουμένως από τους χρήστες. Τέτοιες πληροφορίες μπορεί να είναι π.χ. ο τίτλος ή ο

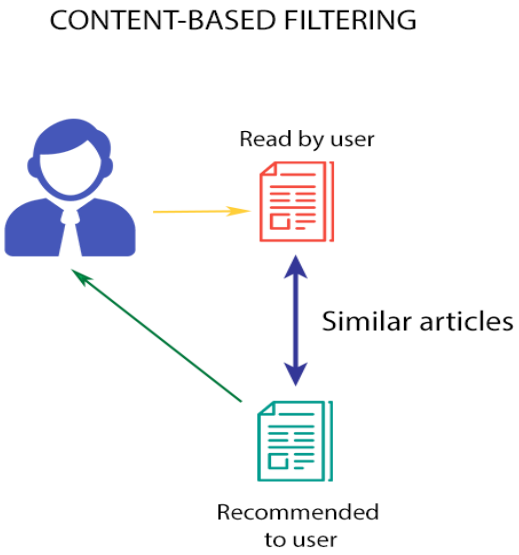
συγγραφέας σε ένα βιβλίο. Στη συνέχεια χτίζουν ένα μοντέλο ή προφίλ των ενδιαφερόντων του χρήστη που βασίζεται στα χαρακτηριστικά αυτά. Σε πολλές περιπτώσεις το προφίλ αυτό δίδεται άμεσα από το χρήστη. Το προφίλ αποτελεί μία δομημένη παρουσίαση των προτιμήσεων του χρήστη και προσαρμόζεται ανάλογα με τα νέα αντικείμενα για τα οποία δείχνει ενδιαφέρον. Η διαδικασία σύστασης συνίσταται ουσιαστικά στην αντιστοίχιση των χαρακτηριστικών του προφίλ των χρηστών έναντι των χαρακτηριστικών ενός αντικειμένου.

Ανάλογα με το τομέα που εξετάζεται, τα δεδομένα που περιγράφουν τα αντικείμενα μπορεί να έχουν μια συγκεκριμένη δομή, π.χ. για το τομέα των βιβλίων μια βάση δεδομένων μπορεί να έχει χαρακτηριστικά όπως συγγραφέας,είδος,τίτλος, είτε να βρίσκονται σε μια αδόμητη μορφή. Στη περίπτωση της αδόμητης μορφής των δεδομένων χρησιμοποιούνται τεχνικές επεξεργασίας φυσικής γλώσσας και επεξεργασίας κειμένων (natural language processing , text mining) για την εξαγωγή της απαραίτητης πληροφορίας.

Ο πιο διαδεδομένος τρόπος για να αντιστοιχηθεί μια αριθμητική τιμή σε κάθε αντικείμενο, είναι η τεχνική συχνότητας όρου – αντίστροφης συχνότητας όρου ή tf-idf αλγόριθμος. Ως όρος (term) συνήθως θεωρείται μία λέξη, λέξη-κλειδί ή και μια ολόκληρη πρόταση ανάλογα με την εφαρμογή. Ένα αρχείο για παράδειγμα, μπορεί να αναπαρασταθεί ως διάνυσμα σταθμισμένων όρων. Ο αλγόριθμος tf-idf αποδίδει σε κάθε όρο ενός εγγράφου ένα συντελεστή βάρους ανάλογα με το πόσες φορές εμφανίζεται αυτός στο έγγραφο. Ο συντελεστής βάρους είναι βασικά ένα στατιστικό μέτρο που χρησιμοποιείται για να μετρηθεί πόσο σημαντική είναι μια λέξη μέσα σε ένα έγγραφο.

Τα πλεονεκτήματα των συστημάτων που χρησιμοποιούν φιλτράρισμα με βάση το περιεχόμενο είναι:

- Δεν απαιτούνται πληροφορίες για τις σχέσεις μεταξύ χρηστών,καθώς δεν χρησιμοποιούνται μέτρα ομοιότητας μεταξύ χρηστών στους υπολογισμούς. Τα συστήματα φιλτραρίσματος με βάση το περιεχόμενο εκμεταλλεύονται μόνο τις αξιολογήσεις που παρέχει ο ενεργός χρήστης για να δημιουργήσουν το προφίλ του.
- Μπορούν να παρέχουν επεξηγήσεις για το ποιοι λόγοι οδήγησαν κάθε αντικείμενο στη λίστα των συστάσεων μέσω της αναγραφής των χαρακτηριστικών ή περιγραφών ενός αντικειμένου. Αυτά τα χαρακτηριστικά είναι δείκτες που μπορεί να συμβουλευτεί κάποιος προκειμένου να αποφασίσει εάν πρέπει να εμπιστευθεί μια σύσταση.
- Τα με βάση το περιεχόμενο συστήματα είναι σε θέση να συστήσουν νέα, μη δημοφιλή αντικείμενα που δεν έχουν βαθμολογηθεί από κανένα χρήστη.



Σχήμα 3.1-2: Αναπαράσταση ενός συστήματος βασιζόμενο στο φιλτράρισμα με βάση το περιεχόμενο

Πηγή: www.marutitech.com/recommendation-engine-benefits

3.1.3 Δημογραφικό φιλτράρισμα (Demographic Filtering)

Οι δημογραφικές προσεγγίσεις στα συστήματα συστάσεων χρησιμοποιούν τα δημογραφικά χαρακτηριστικά των χρηστών όπως π.χ. την ηλικία,το φύλο,το επάγγελμα ως κύριους παράγοντες πάνω στους οποίους βασίζονται οι εξαγόμενες συστάσεις. Το δημογραφικό φιλτράρισμα ακολουθεί τη λογική του συνεργατικού φιλτραρίσματος ότι παρόμοιοι χρήστες είναι πιθανόν να προτιμούν τα ίδια αντικείμενα .Η διαφορά τους έγκειται στο ότι η δημογραφική προσέγγιση δεν χρησιμοποιεί αξιολογήσεις αντικειμένων από χρήστες για τη δημιουργία προφίλ, αλλά δημογραφικές πληροφορίες. Τα κύρια μειονεκτήματα των συστημάτων που χρησιμοποιούν δημογραφική προσέγγιση, οφείλονται στο ότι χρησιμοποιούν μια γενικευμένη λογική για τη συλλογή των ενδιαφερόντων των χρηστών. Επιπλέον σε τέτοιου είδους συστήματα, υπάρχει αδυναμία προσαρμογής στις αλλαγές των προτιμήσεων των χρηστών στο πέρασμα του χρόνου. Παρόλα αυτά,τα δημογραφικά χαρακτηριστικά μπορούν να αποτελέσουν χρήσιμη πληροφορία εάν συμπεριληφθούν σε άλλες προσεγγίσεις.

3.1.4 Φιλτράρισμα με βάση τη γνώση (Knowledge-based Filtering)

Τα συστήματα που βασίζονται στη γνώση, χρησιμοποιούν τη γνώση γύρω από το πως τα διάφορα αντικείμενα ανταποκρίνονται στις ανάγκες των χρηστών (π.χ. όταν αγοράζουμε έναν ηλεκτρονικό υπολογιστή,ένα τέτοιο σύστημα μπορεί να μας προτείνει να αγοράσουμε και μία οθόνη ή ένα πληκτρολόγιο).Σε ένα τέτοιο σύστημα απαιτείται γνώση για τις ανάγκες του κάθε

χρήστη και γνώση γύρω από το πως το κάθε αντικείμενο μπορεί να τις καλύψει,ενώ η διαδικασία που ακολουθείται είναι ο προσδιορισμός της σχέσης μεταξύ των αναγκών που καλύπτει το αντικείμενο i και των αναγκών του χρήστη. Η χρήση των προσεγγίσεων με βάση τη γνώση περιορίζεται κυρίως μόνο όταν τα παραδοσιακά συστήματα σύστασης (με βάση το περιεχόμενο και συνεργατικά) δεν μπορούν να εξάγουν ικανοποιητικές συστάσεις.

3.1.5 Πολυκριτήρια συστήματα συστάσεων (Multi-criteria Recommendation Systems)

Τα πολυκριτήρια συστήματα συστάσεων ορίζονται ως συστήματα συστάσεων στα οποία οι προτιμήσεις των χρηστών καθορίζονται πάνω σε πολλαπλά κριτήρια. Οι προ αναφερόμενες προσεγγίσεις συνήθως χρησιμοποιούν βαθμολογήσεις πάνω σε ένα χαρακτηριστικό για να προσδιορίσουν τη συνολική ικανοποίηση του χρήστη από ένα αντικείμενο και ως εκ τούτου να εξάγουν συστάσεις. Παρόλα αυτά, σε πολλές περιπτώσεις,η αξιολόγηση ενός χαρακτηριστικού δεν συλλαμβάνει πλήρως τη συνολική προτίμηση των χρηστών στα αντικείμενα .Σε αντίθεση λοιπόν με τις προηγούμενες τεχνικές,η πολυκριτήρια προσέγγιση στοχεύει στο προσδιορισμό μιας βαθμολογίας για κάθε νέο ως προς το χρήστη αντικείμενο, εκμεταλλευόμενη τις προτιμήσεις του πάνω σε πολλαπλά κριτήρια. Η περίπτωση τέτοιων συστημάτων αναφέρεται ξεχωριστά, καθώς έχει αναπτυχθεί πληθώρα αλγορίθμων και τεχνικών (κυρίως προερχόμενων από το τομέα της Επιστήμης των Αποφάσεων)για να εκμεταλλευτούν πλήρως τη πληροφορία που λαμβάνεται από τις βαθμολογήσεις σε παραπάνω από ένα χαρακτηριστικά. Στις περισσότερες περιπτώσεις τα πολυκριτήρια συστήματα συστάσεων συνδυάζονται με τις παραδοσιακές τεχνικές που αναφέρθηκαν παραπάνω.

3.1.6 Υβριδικά συστήματα συστάσεων (Hybrid Recommendation Systems)

Τα υβριδικά συστήματα συνδυάζουν δύο ή περισσότερες προσεγγίσεις συστημάτων συστάσεων με σκοπό να παραχθούν βελτιωμένες συστάσεις. Επιδιώκουν να μεγιστοποιήσουν τα οφέλη από τις μεθόδους που συνενώνουν, ελαχιστοποιώντας παράλληλα τα όποια μειονεκτήματα τους.

Σύμφωνα με τον Burke (2007) υπάρχουν επτά τρόποι σύμφωνα με τους οποίους μπορούν να συνδυαστούν συστήματα συστάσεων σε ένα υβριδικό πλαίσιο:

- Εφαρμόζοντας ξεχωριστά τα διαφορετικά συστήματα και παρουσιάζοντας τα αποτελέσματά τους είτε μαζί, είτε σε ξεχωριστές λίστες.
- Σταθμίζοντας τα αποτελέσματα των επιμέρους συστημάτων για να εξάγουμε ένα ενιαίο αποτέλεσμα.
- Επιλέγοντας με βάση κάποιο κριτήριο να χρησιμοποιήσουμε τα αποτελέσματα από ένα μόνο σύστημα.
- Βελτιστοποιώντας τις εξαγόμενες συστάσεις του ενός συστήματος με χρήση των άλλων.

- Συνδυάζοντας δεδομένα από διαφορετικές πηγές και αναλύοντάς τα από ένα σύστημα συστάσεων.
- Χρησιμοποιώντας σαν είσοδο του ενός συστήματος τις εξαγόμενες συστάσεις του άλλου.
- Δημιουργώντας ένα μοντέλο βασισμένο σε ένα σύστημα , το οποίο μετά χρησιμοποιείται σαν είσοδο στο δεύτερο.

Εξαιτίας της ιδιότητας που έχουν οι υβριδικές μεθοδολογίες να ελαχιστοποιούν τα μειονεκτήματα των επιμέρους συστατικών τους ,αναπτύσσονται όλο και περισσότερα υβριδικά συστήματα συστάσεων.

3.2 Περιορισμοί υφιστάμενων προσεγγίσεων

Οι αλγόριθμοι που χρησιμοποιούν τα συστήματα συστάσεων συχνά αντιμετωπίζουν προβλήματα τα οποία χειροτερεύουν την απόδοσή τους. Διαφορετικές προσεγγίσεις συχνά αντιμετωπίζουν διαφορετικά προβλήματα. Σε αυτή την ενότητα θα αναλυθούν τα βασικά προβλήματα των υπάρχων προσεγγίσεων.

3.2.1 Περιορισμοί των προσεγγίσεων βασισμένων στο συνεργατικό φιλτράρισμα

- **Πρόβλημα νέου αντικειμένου** (cold start / new item problem)

  Τα συστήματα συνεργατικού φιλτραρίσματος (CF–Collaborative Filtering) στηρίζονται μόνο στις αξιολογήσεις των χρηστών για να παράγουν συστάσεις, και δεν κάνουν χρήση των πληροφοριών από το περιεχόμενο των υπάρχων αντικειμένων. Επομένως, για κάθε νέο αντικείμενο που εισάγεται στη βάση δεδομένων, το σύστημα δεν έχει επαρκή πληροφορία για αυτό και άρα δεν μπορεί να το προτείνει σε κανένα χρήστη. Ένα νέο αντικείμενο θα μπορεί να προταθεί, όταν έχει πλέον αξιολογηθεί από αρκετούς χρήστες ή όταν βρεθεί μια σχέση που να το συνδέει με τα υπόλοιπα αντικείμενα της βάσης δεδομένων.

- **Πρόβλημα νέου χρήστη** (new user problem)

  Το πρόβλημα που αναφέρεται παραπάνω για τη περίπτωση των αντικειμένων ισχύει και για τη περίπτωση των νέων χρηστών. Όταν ένας νέος χρήστης εισέρχεται στο σύστημα , δεν υπάρχει αρκετή πληροφορία γύρω από τις προτιμήσεις του (δεν έχει κάνει κάποια αξιολόγηση).Οι προσεγγίσεις συνεργατικού φιλτραρίσματος βασίζονται εξ' ολοκλήρου στις αξιολογήσεις των χρηστών για να εξάγουν συστάσεις και δεν χρησιμοποιούν το περιεχόμενο των αντικειμένων για να εισάγουν ομαλά τους νέους χρήστες.

- **Σποραδικότητα των δεδομένων** (data sparsity)

  Η έννοια της σποραδικότητας για τη περίπτωση των δεδομένων με αξιολογήσεις αναφέρεται στην ύπαρξη πολλών χρηστών, αλλά λίγων αξιολογήσεων. Η πλειονότητα των χρηστών σε ιστοσελίδες ηλεκτρονικών καταστημάτων βαθμολογεί μόνο ένα πολύ μικρό ποσοστό των διαθέσιμων αντικειμένων. Σε ένα πίνακα χρηστών / αντικειμένων , όπου κάθε

κελί περιέχει την αξιολόγηση του χρήστη για το αντικείμενο , η σποραδικότητα μπορεί να μετρηθεί ως ένα ποσοστό:

$$\text{Σποραδικότητα} = 1 - \frac{\text{Συνολικός αριθμός αξιολογήσεων}}{\text{Συνολικός αριθμός χρηστών} \times \text{Συνολικός αριθμός αντικειμένων}} \quad \textbf{3.2.1-1}$$

Τα περισσότερα μέτρα ομοιότητας που χρησιμοποιούνται στις τεχνικές συνεργατικού φιλτραρίσματος , λειτουργούν σωστά μόνο όταν υπάρχει επαρκής αριθμός αξιολογήσεων από όλους τους χρήστες.

3.2.2 Περιορισμοί των προσεγγίσεων βασισμένων στο φιλτράρισμα με βάση το περιεχόμενο

- **Περιορισμένη ανάλυση περιεχομένου**

  Τα συστήματα βάσει περιεχομένου, περιορίζονται από τα χαρακτηριστικά γνωρίσματα των αντικείμενων που πρόκειται να συσταθούν. Για να μπορέσει να υπάρξει ένα επαρκές σύνολο χαρακτηριστικών γνωρισμάτων θα πρέπει το περιεχόμενο να είναι σε τέτοια μορφή, ώστε να μπορεί να αναλυθεί αυτόματα από τον υπολογιστή. Για παράδειγμα η ανάλυση του περιεχομένου άρθρων ή ειδήσεων μπορεί να γίνει αυτόματα από τον υπολογιστή με μικρή παρέμβαση από τον άνθρωπο. Στη περίπτωση όμως των πολυμέσων (τραγούδια,ταινίες κ.τ.λ.) είναι δυσκολότερο να εφαρμοσθούν αυτόματες τεχνικές εξαγωγής των χαρακτηριστικών και πολλοί αλγόριθμοι αδυνατούν να αναλύσουν το περιεχόμενό τους.

- **Πρόβλημα νέου χρήστη**

  Και σε αυτή τη περίπτωση συστημάτων , υπάρχει το πρόβλημα του νέου χρήστη. Τα συστήματα με βάση το περιεχόμενο προτείνουν αντικείμενα τα οποία έχουν παρόμοιο περιεχόμενο με τα αντικείμενα που ο χρήστης επέλεξε στο παρελθόν. Όταν ο χρήστης δεν έχει δηλώσει τη προτίμησή του σε κάποια αντικείμενα , το σύστημα δεν μπορεί να εξάγει το περιεχόμενο τους και να προτείνει παρόμοια.

- **Το πρόβλημα της υπερειδίκευσης**

  Το πρόβλημα της υπερειδίκευσης αναφέρεται στο γεγονός ότι τα συστήματα συστάσεων με βάση το περιεχόμενο προτείνουν μόνο αντικείμενα παρόμοια (σύμφωνα με το περιεχόμενο) με αυτά που άρεσαν παλαιότερα στο χρήστη. Ως εκ τούτου το σύστημα δε μπορεί να προτείνει διαφορετικά αντικείμενα παρόμοιου περιεχομένου π.χ. διαφορετικά άρθρα που περιγράφουν το ίδιο γεγονός.

3.3 Αξιολόγηση συστημάτων συστάσεων

Το πρόβλημα συστάσεων μπορεί να επιλυθεί είτε με προσδιορισμό μιας βαθμολογίας R(u,i) για κάθε αντικείμενο i που δεν έχει αξιολογήσει ο χρήστης u, είτε με τη κατηγοριοποίηση των αντικειμένων σε ομάδες «προτείνονται» και «δεν προτείνονται».Για την αξιολόγηση λοιπόν των συστημάτων συστάσεων, έχουν προταθεί διάφοροι τρόποι (Herlocker, Konstan, 2004), καθώς πλέον τα συστήματα συστάσεων αξιολογούνται με διαφορετικούς τρόπους, ανάλογα με το τομέα

και το λόγο για τον οποίο πρόκειται να εφαρμοσθούν. Σε αυτή την ενότητα θα αναλυθούν μέτρα ακρίβειας για την αξιολόγηση των συστημάτων συστάσεων.

3.3.1 Στατιστικά μέτρα ακρίβειας (Statistical accuracy metrics)

Τα μέτρα που ανήκουν σε αυτή τη κατηγορία υπολογίζουν πόσο κοντά είναι η εκτιμώμενη από το σύστημα βαθμολογία R'(u,i) σε σχέση με τη πραγματική βαθμολογία R(u,i). Το πιο διαδεδομένο μέτρο αυτής της κατηγορίας είναι το μέσο απόλυτο σφάλμα (Mean Absolute Error- MAE) , το οποίο υπολογίζει για κάθε αντικείμενο i την απόλυτη διαφορά της εκτιμώμενης βαθμολογίας από την πραγματική βαθμολογία που έχει εισάγει ο χρήστης και στη συνέχεια τις σταθμίζει ως εξής :

$$MAE_u = \frac{1}{n} \sum_{i=1}^{n} |r_{ui} - r'_{ui}| \qquad \textbf{3.3.1-1}$$

Όπου n το πλήθος των αντικειμένων που έχει βαθμολογήσει ο χρήστης u. Το μέσο απόλυτο σφάλμα για όλο το σύστημα υπολογίζεται βρίσκοντας το μέσο όρο των MAE_u όλων των χρηστών.

Άλλο ένα διαδεδομένο μέτρο αυτής της κατηγορίας είναι η ρίζα του μέσου τετραγωνικού σφάλματος(Root Mean Squared Error- RMSE). Η διαφορά του RMSE και MAE είναι ότι στο MAE όλες οι επιμέρους διαφορές είναι εξίσου σταθμισμένες. Στη περίπτωση του RMSE οι διαφορές υψώνονται στο τετράγωνο προτού βρεθεί ο μέσος όρος, οπότε δίδεται σχετικά μεγαλύτερο βάρος σε μεγάλες διαφορές.

$$RMSE_u = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (r_{ui} - r'_{ui})^2} \qquad \textbf{3.3.1-2}$$

3.3.2 Μέτρα ακρίβειας κατηγοριοποίησης (Classification accuracy metrics)

Τα μέτρα αυτά υπολογίζουν την επιτυχία ενός αλγορίθμου στο να καθορίσει σωστά τη κατηγορία που ανήκει το κάθε αντικείμενο, στη περίπτωσή μας «προτείνονται» και «δεν προτείνονται».Σύμφωνα με τους Herlocker και Konstan (2004), πρέπει τα διαθέσιμα αντικείμενα να διαχωριστούν ως εξής:

- **Σχετικά αντικείμενα (Relevant)**, τα οποία είναι αυτά που ενδιαφέρουν το χρήστη.
- **Μη σχετικά αντικείμενα (Irrelevant)** , τα οποία δεν ενδιαφέρουν το χρήστη.
- **Επιλεγμένα αντικείμενα (Selected)** είναι αυτά τα οποία ο αλγόριθμος θεωρεί ότι ενδιαφέρουν το χρήστη.
- **Μη επιλεγμένα (Not Selected)**, τα αντικείμενα τα οποία δεν επέλεξε ο αλγόριθμος καθώς θεωρεί ότι δεν ενδιαφέρουν το χρήστη.

Με βάση τη κατηγοριοποίηση αυτή ορίζεται ο δείκτης ακρίβειας (Precision),ένα από τα πιο διαδεδομένα μέτρα ως:

$$\text{Δείκτης ακρίβειας} = \frac{\text{Σχετικά αντικείμενα} \cap \text{Επιλεγμένα αντικείμενα}}{\text{Επιλεγμένα αντικείμενα}} \quad \textbf{3.3.2-1}$$

Ο δείκτης ακρίβειας (precision) συμβολίζει την πιθανότητα επιλογής ενός αντικειμένου το οποίο είναι ταυτόχρονα σχετικό και επιλεγμένο.

Επίσης ορίζεται ο δείκτης ανάκλησης (Recall), ο οποίος χρησιμοποιείται συμπληρωματικά με τον δείκτη ακρίβειας ως:

$$\text{Δείκτης ανάκλησης} = \frac{\text{Σχετικά αντικείμενα} \cap \text{Επιλεγμένα αντικείμενα}}{\text{Σχετικά αντικείμενα}} \quad \textbf{3.3.2-2}$$

Τέλος ορίζεται το μέτρο F (F-score / F1 measure) το οποίο αποτελεί τον αρμονικό μέσο όρο του δείκτη ακρίβειας και ανάκλησης.

$$F = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \textbf{3.3.2-3}$$

Τα μέτρα ακρίβειας κατηγοριοποίησης εφαρμόζονται στη λίστα των αντικειμένων που προτείνει το σύστημα για κάθε χρήστη και η συνολική ικανότητα του συστήματος να κατηγοριοποιεί σωστά τα αντικείμενα υπολογίζεται λαμβάνοντας τις μέσες τιμές των παραπάνω μέτρων.

3.4 Μοντελοποίηση των χρηστών με βάση τη πολυκριτήρια ανάλυση αποφάσεων

Τα συστήματα συστάσεων χαρακτηρίζονται από την έννοια της εξατομίκευσης σε αντίθεση με τις μηχανές αναζήτησης που συνήθως εστιάζουν στην αντιστοίχηση. Για να αποκτήσουν αυτή την ιδιότητα τα συστήματα συστάσεων πρέπει να εξάγουν συστάσεις βασιζόμενα σε πληροφορίες σχετικά με το κάθε χρήστη όπως το προφίλ του , τις προτιμήσεις του ή τις προηγούμενες επιλογές του. Η δημιουργία συστάσεων και η εξατομίκευση λοιπόν είναι έννοιες που συνδέονται μεταξύ τους. Μια σύσταση μπορεί να θεωρηθεί ως «μη εξατομικευμένη» εάν δεν εξαρτάται από το προφίλ κάποιου χρήστη. Στις περιπτώσεις αυτές, τα συστήματα συστάσεων δεν ξεχωρίζουν τους χρήστες μεταξύ τους, και προσφέρουν τις ίδιες συστάσεις σε χρήστες με διαφορετικά χαρακτηριστικά. Αντίθετα, οι εξατομικευμένες συστάσεις είναι αυτές που βασίζονται σε δεδομένα σχετικά με τους χρήστες, τα οποία συλλέγονται και αναπαρίστανται στα προφίλ των χρηστών.

3.4.1 Εισαγωγή

Για το σχεδιασμό ενός επιτυχούς συστήματος συστάσεων πρέπει να δοθεί ιδιαίτερη σημασία στη δημιουργία του προφίλ των χρηστών,καθώς συμβάλλει στην εξατομίκευση των συστάσεων για κάθε χρήστη. Αρχικά πρέπει να διευκρινιστούν οι έννοιες της μοντελοποίησης και του προφίλ των χρηστών. Στη βιβλιογραφία υπάρχουν αρκετοί ορισμοί των όρων προφίλ και μοντέλο χρήστη (user model, user profile), ενώ πρέπει να σημειωθεί ότι αυτές είναι δύο διαφορετικές έννοιες. Οι Yehya Mohamad και Christos Kouroupetroglou (2014) ορίζουν τα μοντέλα χρηστών ως ξεκάθαρες

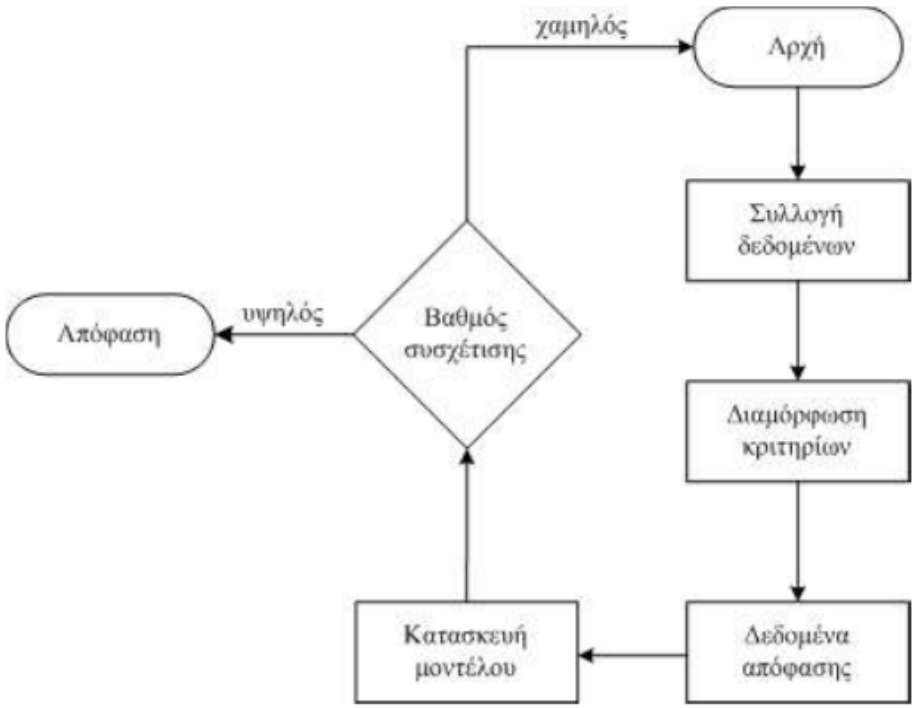
αναπαραστάσεις των ιδιοτήτων κάθε χρήστη, συμπεριλαμβανομένων των αναγκών, των προτιμήσεων καθώς και των φυσικών και γνωστικών του χαρακτηριστικών. Το μοντέλο κάθε χρήστη αναπαρίσταται έπειτα στη μορφή ενός προφίλ. Σε αυτή τη παράγραφό θα αναλυθούν τεχνικές της πολυκριτήριας ανάλυσης αποφάσεων που μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση των χρηστών και τη δημιουργία προφίλ.

3.4.2 Αναλυτική-Συνθετική προσέγγιση

Η πολυκριτήρια λήψη αποφάσεων είναι μια διαδικασία μοντελοποίησης και επίλυσης των προβλημάτων απόφασης που περιέχουν πολλαπλά κριτήρια. Στόχος της είναι να βοηθήσει τον αποφασίζοντα στη διαδικασία της λήψης μιας τελικής απόφασης. Ο Bernard Roy (1985) προτείνει ένα γενικό μεθοδολογικό πλαίσιο μοντελοποίησης των προβλημάτων απόφασης, το οποίο αποτελείται από τέσσερα διαδοχικά στάδια.

1. Καθορισμός του αντικειμένου της απόφασης. Το αντικείμενο της απόφασης οφείλει να αναλυθεί σε ένα διακριτό ή συνεχές σύνολο δράσεων (εναλλακτικών)πάνω στις οποίες θα παρθεί η απόφαση,το οποίο ονομάζουμε σύνολο Α. Ο ορισμός μιας προβληματικής πάνω στο σύνολο αυτό των δράσεων αποσκοπεί στο να δώσει επιχειρησιακό ρόλο στο έργο του αναλυτή. Μια προβληματική σχετίζεται άμεσα με το ερώτημα: «πώς θα διαχειριστούμε τις δράσεις;» ή πιο συγκεκριμένα «τι θέλουμε να επιτύχουμε;»Μπορούμε να διακρίνουμε 4 προβληματικές αναφοράς:
    - Προβληματική α: επιλογή (choice) μιας και μονής δράσης από το σύνολο Α.
    - Προβληματική β: ταξινόμηση (sorting) των δράσεων σε ομογενείς προκαθορισμένες κατηγορίες, οι οποίες είναι διατεταγμένες ως προς τις προτιμήσεις του αποφασίζοντος.
    - Προβληματική γ: κατάταξη (ranking) των δράσεων του συνόλου Α από τη καλύτερη μέχρι τη χειρότερη.
    - Προβληματική δ : περιγραφή (description) των δράσεων και των συνεπειών τους στη γλώσσα των εμπλεκόμενων στη διαδικασία της απόφασης.
2. Καθορισμός μιας συνεπούς οικογένειας κριτηρίων. Καθορισμός δηλαδή ενός συνόλου συναρτήσεων που δηλώνουν τις προτιμήσεις των χρηστών πάνω στις διάφορες εναλλακτικές. Η συνεπής οικογένεια κριτηρίων οφείλει να πληρεί τρεις θεμελιώδεις συνθήκες. Τη συνέπεια ή μονοτονία, την επάρκεια και το μη πλεονασμό.
3. Δημιουργία μοντέλου ολικής προτίμησης .Ένα μοντέλο ολικής προτίμησης αποτελεί λίγο πολύ το κανόνα σύνθεσης των κριτηρίων,δηλαδή των μοντέλων μερικής προτίμησης.
4. Υποστήριξη της απόφασης. Στο στάδιο αυτό, ο αναλυτής του προβλήματος αναζητά και οργανώνει τα στοιχεία απάντησης σε συγκεκριμένα ερωτηματικά που θέτουν ή ενδέχεται να θέσουν κάποιοι εμπλεκόμενοι στη διαδικασία της απόφασης και κυρίως ο αποφασίζων.

Σύμφωνα με την αναλυτική-συνθετική προσέγγιση(aggregation – disaggregation approach) η τελική απόφαση είναι γνωστή εκ των προτέρων και αναλύεται ώστε να προκύψουν τα χαρακτηριστικά που οδήγησαν τον αποφασίζοντα στη τελική του απόφαση. Με βάση τη πληροφορία αυτή δημιουργείται ένα μοντέλο προτιμήσεων του χρήστη που χρησιμοποιείται για να τον υποβοηθήσει σε μελλοντικές αποφάσεις. Ουσιαστικά, στις μεθόδους της συγκεκριμένης προσέγγισης, εκτιμώνται ή συμπεραίνονται οι παράμετροι εκείνες ενός μοντέλου απόφασης οι οποίες επιτρέπουν τη βέλτιστη ανασύσταση μιας απόφασης. Η αρχή της αναλυτικής-συνθετικής προσέγγισης παρουσιάζεται στο σχήμα 3.4-1, όπου πρέπει να σημειωθεί ότι, σε περίπτωση που διαπιστωθεί ασυνέπεια ανάμεσα στον αποφασίζοντα και το εκτιμώμενο μοντέλο απόφασης, αναθεωρείται είτε η συνεπής οικογένεια κριτηρίων είτε η αξιοπιστία των δεδομένων της απόφασης.



Σχήμα 3.4-1: Αρχή της αναλυτικής-συνθετικής προσέγγισης

Πηγή: Σίσκος Ι. (2008). *ΜΟΝΤΕΛΑ ΑΠΟΦΑΣΕΩΝ, Μεθοδολογία Επιχειρησιακής Έρευνας, Θεωρία Πολυκριτήριας Ανάλυσης, Εφαρμογές σε Επιχειρήσεις & Οργανισμούς*, Εκδόσεις Νέων Τεχνολογιών, Αθήνα, σελ. 307.

Για την αποσαφήνιση της ολικής προτίμησης ενός αποφασίζοντος, οι Jacquet-Lagrèze & Siskos τονίζουν την αναγκαιότητα ύπαρξης ενός συνόλου δράσεων αναφοράς A_R (reference actions), το οποίο μπορεί να είναι:

- ένα σύνολο προγενέστερων δράσεων (past decisions)

- ένα υποσύνολο των πραγματικών δράσεων του προβλήματος, ιδιαίτερα όταν το Α είναι αρκετά μεγάλο ($A_R \subset A$),
- ένα σύνολο εικονικών δράσεων, το οποίο μπορεί να αξιολογηθεί με ευκολία από τον αποφασίζοντα, ώστε αυτός να εκφράσει τις ολικές του προτιμήσεις.

Σε καθεμία από τις παραπάνω περιπτώσεις, ζητείται από τον αποφασίζοντα να εξωτερικεύσει ή/και επιβεβαιώσει τις ολικές προτιμήσεις του στο σύνολο A_R, λαμβάνοντας υπ' όψη τις επιδόσεις των δράσεων αναφοράς σε όλα τα κριτήρια.

3.4.3 Η μέθοδος UTA

Σύμφωνα με τους Siskos, Grigoroudis, Matsatsinis (2016), η μέθοδος UTA (UTility Additives), η οποία προτάθηκε από τους Jacquet-Lagrèze & Siskos (1982) είναι από τις πιο γνωστές πολυκριτήριες μεθοδολογίες λήψης αποφάσεων και θέτει τις βάσεις της σύγχρονης Αναλυτικής-Συνθετικής προσέγγισης. Έχει ως στόχο την εκτίμηση μιας ή περισσότερων προσθετικών συναρτήσεων αξίας από μία προδιάταξη ενός συνόλου αναφοράς A_R. Η μέθοδος χρησιμοποιεί ειδικές τεχνικές γραμμικού προγραμματισμού για να καθορίσει τις συγκεκριμένες συναρτήσεις, έτσι ώστε η κατάταξη που αποκτάται μέσω αυτών των συναρτήσεων στο A_R να είναι όσο το δυνατόν πιο συμβατή με την αρχική προ διάταξη.

Το μοντέλο σύνθεσης των κριτηρίων (μοντέλο απόφασης) στη μέθοδο UTA είναι μία προσθετική συνάρτηση αξίας της ακόλουθης μορφής:

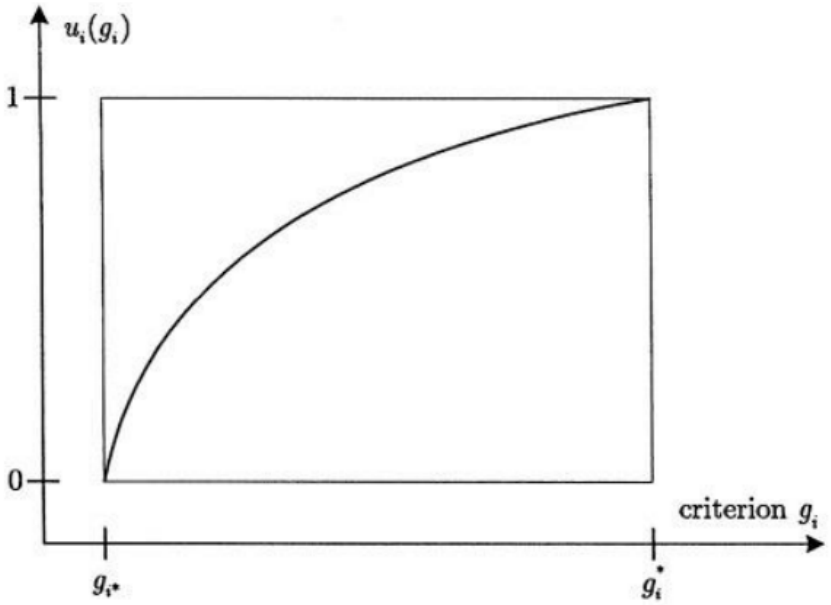
$$u(g)=\sum_{i=1}^{n} p_i u_i(g_i) \qquad \textbf{3.4.3-1}$$

υπό τους περιορισμούς κανονικοποίησης:

$$\begin{cases} \sum_{i=1}^{n} p_i = 1 \\ u_i(g_{i*}) = 0, u_i(g_i^*) = 1, \forall i = 1,2,\ldots n \end{cases} \qquad \textbf{3.4.3-2}$$

όπου u_i =1,2,...n είναι αύξουσες συναρτήσεις των g_i που καλούνται περιθώριες ή μερικές συναρτήσεις αξίας (marginal value functions) , κανονικοποιημένες στο διάστημα [0,1] και p_i τα βάρη των u_i .Και οι μερικές και οι ολικές συναρτήσεις αξίας έχουν την ιδιότητα της μονοτονίας του πραγματικού κριτηρίου. Για παράδειγμα, στην περίπτωση της συνάρτησης ολικής αξίας ισχύουν οι παρακάτω ιδιότητες :

$$\begin{cases} u[g(a)] > u[g(b)] \Leftrightarrow a > b \,(\text{προτίμηση}) \\ u[g(a)] = u[g(b)] \Leftrightarrow a \sim b \,(\text{αδιαφορία}) \end{cases} \qquad \textbf{3.4.3-3}$$



Σχήμα 3.4-2: Η κανονικοποιημένη μερική συνάρτηση αξίας

Πηγή: Σίσκος Ι. (2008). *ΜΟΝΤΕΛΑ ΑΠΟΦΑΣΕΩΝ, Μεθοδολογία Επιχειρησιακής Έρευνας, Θεωρία Πολυκριτήριας Ανάλυσης, Εφαρμογές σε Επιχειρήσεις & Οργανισμούς*, Εκδόσεις Νέων Τεχνολογιών, Αθήνα, σελ. 309.

Η μέθοδος UTA θεωρεί μια μορφή της προσθετικής συνάρτησης αξίας χωρίς βάρη,ισοδύναμη της μορφής 3.4.3-1 , 3.4.3-2 :

$$u(g)=\sum_{i=1}^{n} u_i(g_i) \qquad \textbf{3.4.3-4}$$

υπό τους περιορισμούς κανονικοποίησης:

$$\begin{cases} \sum_{i=1}^{n} u_i(g_i^*)=1 \\ u_i(g_{i*})=0, \forall i=1,2,\ldots n \end{cases} \qquad \textbf{3.4.3-5}$$

Η ύπαρξη ενός τέτοιου μοντέλου προϋποθέτει φυσικά τη προτιμησιακή ανεξαρτησία των κριτηρίων (preferential independence) για τον αποφασίζοντα. Χρησιμοποιώντας το προσθετικό μοντέλο 3.4.3-4 και 3.4.3-5 και λαμβάνοντας υπ' όψη τις σχέσεις προτίμησης 3.4.3-3 , η αξία κάθε εναλλακτικής $a \in A_R$ μπορεί να γραφεί ως εξής:

$$u'[g(a)]=\sum_{i=1}^{n} u_i[g_i(a)]+\sigma(a), \forall a \in A_R \qquad \textbf{3.4.3-6}$$

όπου σ(α) είναι το ενδεχόμενο σφάλμα σε σχέση με το u'[g(a)].Για την εκτίμηση των αντίστοιχων μερικών συναρτήσεων αξίας σε μια γραμμική κατά τμήματα μορφή, οι Jacquet-Lagrèze & Siskos προτείνουν τη χρήση της γραμμικής παρεμβολής. Έτσι για κάθε κριτήριο,το διάστημα $[g_{i*}\ ,\ g_i^*]$ χωρίζεται σε α_i-1 ίσα διαστήματα και τα τελικά σημεία g_i^j δίνονται από τη σχέση:

$$g_i^j = g_{i*} + \frac{j-1}{a_i - 1}(g_i^* - g_{i*}),\ \forall\, j = 1,2,\ldots \alpha_i \qquad \textbf{3.4.3-7}$$

Η μερική αξία μιας εναλλακτικής a υπολογίζεται με χρήση γραμμικής παρεμβολής :

$$u_i[g_i(a)] = u_j(g_i^j) + \frac{g_i(a) - g_i^j}{g^{j+1} - g_i^j}[u_i(g_i^{j+1}) - u_i(g_i^j)],\ \text{για}\ g_i(a) \in [g_i^j, g_i^{j+1}] \qquad \textbf{3.4.3-8}$$

Επίσης, το σύνολο αναφοράς $A_R = \{a_1, a_2, \ldots, a_m\}$ «ανακατατάσσεται» με τέτοιο τρόπο, ώστε οι δράσεις να είναι διατεταγμένες σε μια σειρά προτίμησης, δηλαδή η a_1 αποτελεί την κεφαλή και η a_m την ουρά της κατάταξης. Δεδομένου ότι η συγκεκριμένη κατάταξη έχει τη μορφή μιας προ διάταξης R, για κάθε ζεύγος διαδοχικών δράσεων $(a_k\ ,\ a_{k+1})$ ισχύει, είτε $a_k > a_{k+1}$ (προτίμηση) είτε $a_k \sim a_{k+1}$ (αδιαφορία). Έτσι, αν τεθεί

$$\Delta(a_k, a_{k+1}) = u'[g(a_k)] - u'[g(a_{k+1})] \qquad \textbf{3.4.3-9}$$

τότε ισχύει μια από τις ακόλουθες περιπτώσεις:

$$\begin{cases} \Delta(a_k, a_{k+1}) \geq \delta, \text{αν}\ a_k > a_{k+1} \\ \Delta(a_k, a_{k+1}) = 0, \text{αν}\ a_k \sim a_{k+1} \end{cases} \qquad \textbf{3.4.3-10}$$

όπου δ είναι ένας μικρός θετικός αριθμός που διαχωρίζει σημαντικά δύο διαδοχικές κλάσεις ισοδυναμίας της R. Λαμβάνοντας υπόψη την υπόθεση σχετικά με τη μονοτονία των προτιμήσεων, οι μερικές αξίες $u_i(g_i)$ πρέπει να ικανοποιούν το σύνολο των ακόλουθων περιορισμών:

$$u_i(g_i^{j+1}) - u_i(g_i^j) \geqslant s_i\ \forall\, j = 1,2,\ldots,\alpha_{i-1}, i = 1,2,\ldots,n \qquad \textbf{3.4.3-11}$$

όπου $s_i \geq 0$ είναι τα κατώφλια αδιαφορίας που ορίζονται για κάθε κριτήριο g_i .Τα συγκεκριμένα κατώφλια δεν είναι απαραίτητο να χρησιμοποιούνται σε κάθε περίπτωση εφαρμογής της μεθόδου UTA, αλλά είναι ιδιαίτερα χρήσιμα για την αποφυγή φαινομένων, όπου : $u_i(g_i^{j+1}) = u_i(g_i^j)$ όταν $g_i^{j+1} > g_i^j$.

Οι μερικές συναρτήσεις αξίας υπολογίζονται τελικά μέσω του ακόλουθου γραμμικού προγράμματος, όπου ως περιορισμοί χρησιμοποιούνται οι σχέσεις 3.4.3-4, 3.4.3-5, 3.4.3-10, 3.4.3-11.

$$\begin{cases} [\min] F = \sum_{a \in A_R} \sigma(\alpha) \\ \text{υπό τους περιορισμούς :} \\ \left.\begin{array}{l} \Delta(a_k, a_{k+1}) \geq \delta, \alpha\nu\, a_k > a_{k+1} \\ \Delta(a_k, a_{k+1}) = 0, \alpha\nu\, a_k \sim a_{k+1} \end{array}\right\} \forall\, k \\ u_i(g_i^{j+1}) - u_i(g_i^j) \geq 0 \;\forall\, i \,\kappa\alpha\iota\, j \\ \sum_{i=1}^{n} u_i(g_i^*) = 1 \\ u_i(g_{i*}) = 0,\; u_i(g_i^j) \geqslant 0, \sigma(\alpha) \geqslant 0 \;\forall\, a \in A_R, \forall\, i \,\kappa\alpha\iota\, j \end{cases} \qquad \textbf{3.4.3-12}$$

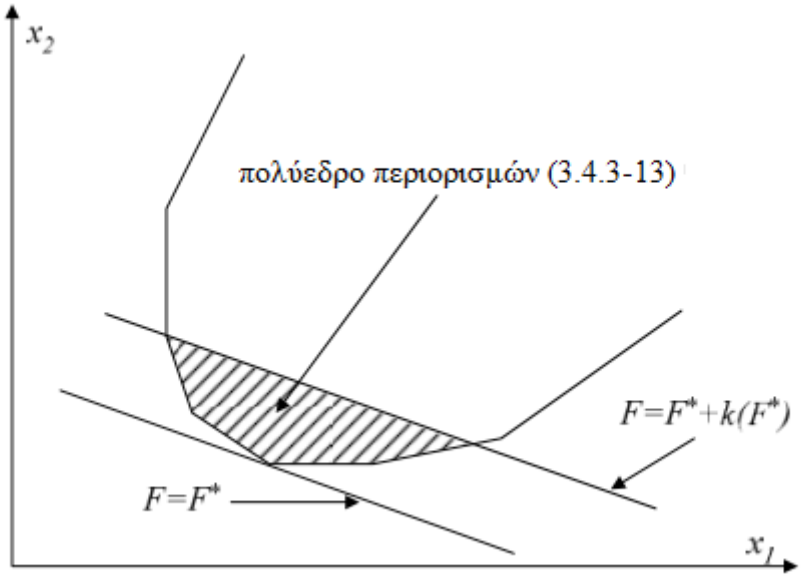
Η ανάλυση ευστάθειας των αποτελεσμάτων του γραμμικού προγράμματος (γ. π.) 3.4.3-12 αντιμετωπίζεται ως ένα πρόβλημα ανάλυσης μεταβελτιστοποίησης . Πράγματι, αν η βέλτιστη λύση δώσει F*=0, τότε το υπερπολύεδρο των αποδεκτών λύσεων για τα u_i (g_i) δεν είναι κενό, αλλά υπάρχουν πολλαπλές συναρτήσεις αξίας που είναι απόλυτα συνεπείς με τη προδιάταξη R. Ακόμη και στη περίπτωση που η βέλτιστη τιμή της αντικειμενικής συνάρτησης είναι μη μηδενική, υπάρχουν άλλες λύσεις, λιγότερο καλές για την F, που είναι σε θέση να βελτιώσουν άλλα εναλλακτικά κριτήρια βελτιστοποίησης (π.χ. το συντελεστή συσχέτισης τ του Kendall).

Όπως φαίνεται στο σχήμα 3.4-3, ο χώρος των μεταβέλτιστων λύσεων καθορίζεται από το υπερπολύεδρο:

$$\begin{cases} F \leq F^* + k(F^*) \\ \text{όλοι οι περιορισμοί του } \gamma.\pi.(3.4.3-12) \end{cases} \qquad \textbf{3.4.3-13}$$

όπου $k(F^*)$ είναι ένα θετικό (ή μηδέν) κατώφλι, το οποίο καθορίζεται ως ένα μικρό ποσοστό του σφάλματος F^*.Υπάρχει ένας σημαντικός αριθμός αλγορίθμων που είναι σε θέση να εξετάσουν τις λύσεις-κορυφές του υπερπολυέδρου,όπως μέθοδοι κλάδου και φράγματος ή η μέθοδος αντίστροφης simplex. Οι Jacquet-Lagrèze & Siskos, στην αρχική μορφή της μεθόδου UTA, προτείνουν τη διερεύνηση του πολυέδρου, μέσω μιας ευρετικής μεθόδου αναζήτησης (ημι)βέλτιστων λύσεων,επιλύοντας τα ακόλουθα γ.π. :

$$\begin{cases} [\min] u_i(g_i^*) \\ \text{στο πολύεδρο} \end{cases} \quad \kappa\alpha\iota \quad \begin{cases} [\max] u_i(g_i^*) \\ \text{στο πολύεδρο} \end{cases} \quad \forall\, i=1,2,...n \qquad \textbf{3.4.3-14}$$



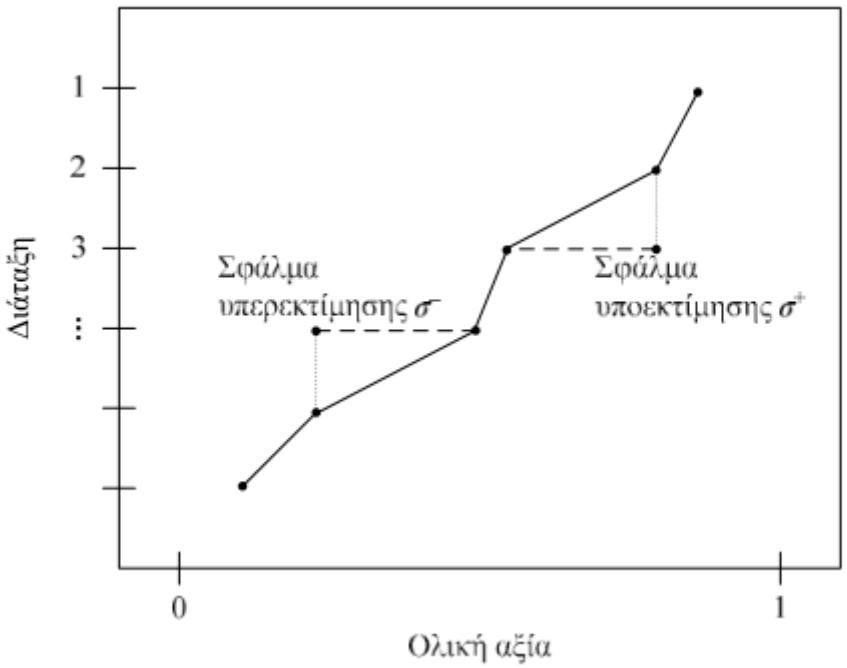
Σχήμα 3.4-3: Ανάλυση ευστάθειας στη μέθοδο UTA

Πηγή: Σίσκος Ι. (2008). *ΜΟΝΤΕΛΑ ΑΠΟΦΑΣΕΩΝ, Μεθοδολογία Επιχειρησιακής Έρευνας, Θεωρία Πολυκριτήριας Ανάλυσης, Εφαρμογές σε Επιχειρήσεις & Οργανισμούς*, Εκδόσεις Νέων Τεχνολογιών, Αθήνα, σελ. 312.

Ως τελική λύση του προβλήματος, υπολογίζεται η μέση τιμή των λύσεων τω προηγούμενων γ.π., που είναι και αυτή (ημι)βέλτιστη, λόγω της κυρτότητας του υπερπολυέδρου. Σε περίπτωση αστάθειας, οι λύσεις των γ.π. (3.4.3-14) εμφανίζουν μεγάλη απόκλιση μεταξύ τους και η εκτιμώμενη μέση λύση είναι λιγότερο αντιπροσωπευτική. Σε κάθε περίπτωση, οι επιμέρους αυτές λύσεις υποδεικνύουν τη διακύμανση των βαρών των κριτηρίων g_i και συνεπώς δίνουν μια ιδέα της σημαντικότητας αυτών των κριτηρίων στο σύστημα προτιμήσεων του αποφασίζοντος.

3.4.4 Η μέθοδος UTASTAR

Η μέθοδος UTASTAR προτάθηκε από τους Siskos-Yannacopoulos (1985) και αποτελεί μια βελτιωμένη έκδοση της πρωτότυπης μεθόδου UTA. Στην αρχική έκδοση της μεθόδου UTA, για καθεμία δράση $a \in A_R$ ορίζεται ένα μοναδικό σφάλμα σ(α).Αυτή η συνάρτηση σφάλματος δεν είναι επαρκής για την ελαχιστοποίηση της ολικής διασποράς των σημείων στη μονότονη καμπύλη του σχήματος (3.4-4). Το πρόβλημα αφορά τα σημεία που βρίσκονται δεξιά της καμπύλης, από τα οποία θα ήταν προτιμότερο να αφαιρεθεί μια ποσότητα αξίας χωρίς να αυξηθούν οι αξίες των άλλων (παράδειγμα της ποιοτικής ή μονότονης παλινδρόμησης, ordinal regression paradigm, σχήμα 3.4-4).



Σχήμα 3.4-4: Καμπύλη μονότονης παλινδρόμησης

Πηγή: Σίσκος Ι. (2008). *ΜΟΝΤΕΛΑ ΑΠΟΦΑΣΕΩΝ, Μεθοδολογία Επιχειρησιακής Έρευνας, Θεωρία Πολυκριτήριας Ανάλυσης, Εφαρμογές σε Επιχειρήσεις & Οργανισμούς*, Εκδόσεις Νέων Τεχνολογιών, Αθήνα, σελ. 313.

Στη μέθοδο UTASTAR, οι Siskos & Yannacopoulos εισάγουν μια διπλή θετική συνάρτηση σφάλματος και έτσι ο τύπος (3.4.3-6**)** γίνεται:

$$u'[g(a)]=\sum_{i=1}^{n} u_i[g_i(a)]-\sigma^{+}(a)+\sigma^{-}(a), \forall a \in A_R \qquad \textbf{3.4.4-1}$$

όπου όπου σ^{+} και σ^{-} είναι τα σφάλματα υποεκτίμησης και υπερεκτίμησης, αντίστοιχα. Επιπρόσθετα, μια άλλη τροποποίηση αφορά τους περιορισμούς μονοτονίας των κριτηρίων, οι οποίοι μοντελοποιούνται με τη βοήθεια των ακόλουθων μετασχηματισμών των μεταβλητών:

$$w_{ij}=u_i(g_i^{j+1})-u_i(g_i^{j}) \forall i=1,2,\ldots n \text{ και } j=1,2,\ldots,\alpha_i-1 \qquad \textbf{3.4.4-2}$$

Με αυτό τον τρόπο, οι συνθήκες μονοτονίας (3.4.3-11) μπορούν να αντικατασταθούν από περιορισμούς μη αρνητικότητας των μεταβλητών w_{ij}. Συνεπώς ο αλγόριθμος UTASTAR συνοψίζεται στα ακόλουθα βήματα:

Βήμα 1: Η ολική αξία των δράσεων $u[g(a_k)]$, k=1,2,…,m, εκφράζεται αρχικά ως συνάρτηση των μερικών αξιών $u_i(g_i)$ και στη συνέχεια των μεταβλητών w_{ij}, σύμφωνα με την εξίσωση (3.4.4-2), μέσω των ακόλουθων σχέσεων:

$$\begin{cases} u_i(g_i^1)=0 \ \forall \, i=1,2,\ldots n \\ u_i(g_i^j)=\sum_{i=1}^{j-1} w_{ij}, \forall \, i=1,2,\ldots n \text{ και } j=2,3,\ldots \alpha_i-1 \end{cases} \qquad \textbf{3.4.4-3}$$

Βήμα 2: Εισάγονται δύο συναρτήσεις σφάλματος σ^+ και σ^- στο A_R, γράφοντας για κάθε ζεύγος διαδοχικών δράσεων στη προδιάταξη τις αναλυτικές εκφράσεις:

$$\begin{gathered} (a_k, a_{k+1}) = u[g(a_k)] - \sigma^+(a_k) + \sigma^-(a_k) \\ -u[g(a_{k+1})] + \sigma^+(a_{k+1}) - \sigma^-(a_{k+1}) \end{gathered} \qquad \textbf{3.4.4-4}$$

Βήμα 3: Επιλύεται το ακόλουθο γραμμικό πρόγραμμα:

$$\begin{cases} [\min] z = \sum_{k=1}^{m} [\sigma^+(a_k) + \sigma^-(a_k)] \\ \text{υπό τους περιορισμούς:} \\ \left.\begin{array}{l} \Delta(a_k, a_{k+1}) \geq \delta, \text{αν } a_k > a_{k+1} \\ \Delta(a_k, a_{k+1}) = 0, \text{αν } a_k \sim a_{k+1} \end{array}\right\} \forall \, k \\ \sum_{i=1}^{n} \sum_{j=1}^{\alpha_i - 1} w_{ij} = 1 \\ w_{ij} \geqslant 0, \sigma^+(a_k) \geqslant 0, \sigma^-(a_k) \geqslant 0 \ \forall \, i, j \text{ και } k \end{cases} \qquad \textbf{3.4.4-5}$$

Βήμα 4: Ελέγχεται η ύπαρξη πολλαπλών βέλτιστων ή ημιβέλτιστων λύσεων στο γ.π. (3.4.4-5), υπολογίζοντας το βαρύκεντρο των προσθετικών συναρτήσεων αξίας που μεγιστοποιούν τις ακόλουθες αντικειμενικές συναρτήσεις :

$$u_i(g_i^*) = \sum_{j=1}^{\alpha_i - 1} w_{ij} \, \forall \, i = 1,2,\ldots,n \qquad \textbf{3.4.4-6}$$

στο υπερπολύεδρο των περιορισμών του γ.π. 3.4.4-5 που περιορίζεται από τον επόμενο νέο περιορισμό:

$$\sum_{k=1}^{m} [\sigma^+(a_k) + \sigma^-(a_k)] \leq z^* + \varepsilon \qquad \textbf{3.4.4-7}$$

όπου z^* είναι η βέλτιστη τιμή (σφάλμα) του γ.π. και ε είναι ένας πολύ μικρός θετικός αριθμός ή μηδέν.

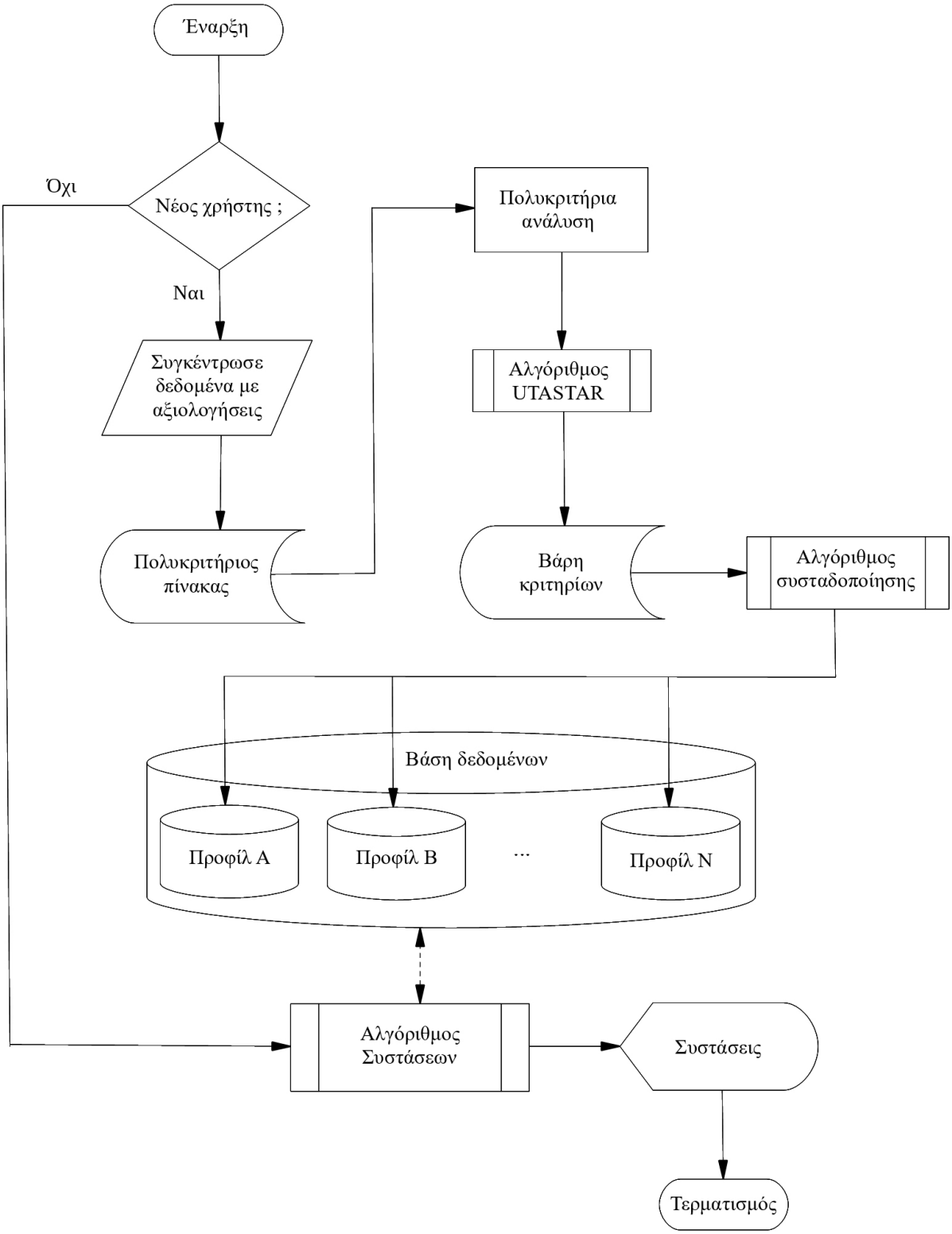
Οι Siskos & Yannacopoulos απέδειξαν, σε ένα σύνολο πειραματικών δεδομένων, ότι η UTASTAR δίνει καλύτερα αποτελέσματα από το πρωτότυπο αλγόριθμο UTA.

Οι πολυκριτηρίες μέθοδοι που αναλύθηκαν παραπάνω μπορούν να χρησιμοποιηθούν για τη δημιουργία προφίλ χρηστών μέσω της μοντελοποίησης των προτιμήσεών τους. Για την αναπαράσταση των προτιμήσεων των χρηστών μπορούν να χρησιμοποιηθούν π.χ. οι συναρτήσεις αξίας που προκύπτουν απ' την εφαρμογή των αλγορίθμων για κάθε χρήστη ή τα βάρη των κριτηρίων.

Κεφάλαιο 4 : Προτεινόμενη μεθοδολογία

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα αναλυθεί το μεθοδολογικό πλαίσιο στο οποίο βασίζεται το σύστημα σύστασης ξενοδοχείων για ηλεκτρονικές κρατήσεις που αναπτύχθηκε. Μια επισκόπηση της προτεινόμενης μεθοδολογίας φαίνεται στο σχήμα 4.1-1. Η μεθοδολογία αυτή ακολουθεί τη λογική του συνεργατικού φιλτραρίσματος βασιζόμενο στο χρήστη (user based collaborative filtering) και χρησιμοποιεί τη πολυκριτήρια ανάλυση για τη δημιουργία προφίλ χρηστών με στόχο την εξατομίκευση των συστάσεων. Χωρίζεται σε τέσσερεις φάσεις, όπου στη πρώτη αποκτώνται και επεξεργάζονται τα δεδομένα εισόδου για κάθε νέο χρήστη, στη δεύτερη κάθε χρήστης μοντελοποιείται με βάση τη πολυκριτήρια ανάλυση, στη τρίτη πραγματοποιείται η διαδικασία της συσταδοποίησης για τη δημιουργία προφίλ χρηστών και στη τέταρτη εφαρμόζεται ένας αλγόριθμος συστάσεων που χρησιμοποιεί τις πολυκριτήριες αξιολογήσεις για τη δημιουργία συστάσεων. Στις επόμενες παραγράφους θα αναλυθεί η διαδικασία και οι αλγόριθμοι που εφαρμόζονται σε κάθε φάση.



Σχήμα 4.1-1: Αρχιτεκτονική του μεθοδολογικού πλαισίου

4.2 Μεθοδολογικό πλαίσιο

4.2.1 Απόκτηση δεδομένων

Για τη δημιουργία συστάσεων σε κάθε χρήστη είναι απαραίτητο το σύστημα να έχει για αυτόν αρκετή πληροφορία. Τα δεδομένα που χρησιμοποιούνται ως είσοδος στους αλγορίθμους και στις διαδικασίες του συστήματος σε πρώτη φάση είναι σε μορφή αξιολογήσεων. Κάθε χρήστης είναι απαραίτητο να έχει αξιολογήσει μια ομάδα αντικειμένων μέσα από ένα σύνολο εναλλακτικών επιλογών. Το σύστημα λοιπόν δέχεται για κάθε χρήστη αριθμητικά δεδομένα τα οποία δηλώνουν τη προτίμησή του σε διάφορα αντικείμενα με βάση διαφορετικά κριτήρια. Θεωρώντας ως A το σύνολο των αντικειμένων, συγκεντρώνονται τα δεδομένα για κάθε χρήστη σε μορφή πίνακα, του οποίου οι γραμμές αντιστοιχούν στα αντικείμενα του συνόλου A, οι στήλες σε k κριτήρια και σε μια συνολική προτίμηση, ενώ τα κελιά περιέχουν βαθμολογίες οι οποίες ακολουθούν ορισμένη μονοτονία και κλίμακα μέτρησης π. χ. βαθμολογία από το 1 έως το 5. Η συνολική προτίμηση μετατρέπεται σε μια κατάταξη των εναλλακτικών με φθίνουσα σειρά ώστε να δημιουργηθεί ένας πολυκριτήριος πίνακας ως είσοδος για τη δεύτερη φάση, όπως φαίνεται παρακάτω:

Initial Data Form									Final Data Form								
AuthorID	HotelID	Service	Cleanliness	Value	Sleep Quality	Location	Rooms	Overall	AuthorID	HotelID	Service	Cleanlines	Value	Sleep Qua	Location	Rooms	Ranking
1	1010527	5	5	5	5	5	5	5	1	1010527	5	5	5	5	5	5	1
	121998	5	5	5	5	5	5	5		121998	5	5	5	5	5	5	1
	230405	4	5	4	5	4	4	4		230405	4	5	4	5	4	4	2
	2515626	5	5	4	5	5	4	5		2515626	5	5	4	5	5	4	1
	1022792	4	4	4	2	4	3	3		1022792	4	4	4	2	4	3	3
	2515527	5	5	5	4	3	3	5		2515527	5	5	5	4	3	3	1
	2072559	4	4	4	4	4	4	4		2072559	4	4	4	4	4	4	2
	268660	5	5	3	5	3	4	5		268660	5	5	3	5	3	4	1
	2515818	5	5	5	5	5	5	5		2515818	5	5	5	5	5	5	1
	190072	1	1	1	1	5	1	1		190072	1	1	1	1	5	1	5
2	2515499	5	5	3	5	5	5	5	2	2515499	5	5	3	5	5	5	1
	189076	2	1	2	3	3	2	2		189076	2	1	2	3	3	2	4

Πίνακας 4.2-1: Υπόδειγμα πολυκριτήριου πίνακα

4.2.2 Μοντελοποίηση χρηστών με βάση τη πολυκριτήρια ανάλυση

Στη φάση της πολυκριτήριας ανάλυσης για τη μοντελοποίηση των χρηστών, χρησιμοποιείται ο πολυκριτήριος πίνακας για κάθε χρήστη που αποκτήθηκε από τη πρώτη φάση ως είσοδος του αλγορίθμου utastar που αναλύθηκε στο κεφάλαιο 3. Σε πρώτο στάδιο της πολυκριτήριας ανάλυσης σύμφωνα με την αναλυτική συνθετική προσέγγιση πρέπει να ορισθεί το αντικείμενο της απόφασης . Η πολυκριτήρια μέθοδος utastar επιλύει το πρόβλημα της κατάταξης

εναλλακτικών επιλογών (γ προβληματική) , στη περίπτωση όμως της συγκεκριμένης μεθοδολογίας, η utastar χρησιμοποιείται αποκλειστικά για τη μοντελοποίηση των χρηστών, ενώ τελικός στόχος είναι η εκτίμηση βαθμολογιών για άγνωστα αντικείμενα. Έπειτα είναι απαραίτητο να ελεγχθεί η συνεπής οικογένεια κριτηρίων. Κάθε κριτήριο πρέπει να έχει χειρότερη και καλύτερη τιμή (g_{j*} , g_j^*), συγκεκριμένη μονοτονία (αύξον/φθίνον) και είδος (ποσοτικό/ποιοτικό). Τέλος για να εφαρμοσθεί ο αλγόριθμος utastar είναι απαραίτητο κάθε χρήστης να έχει αξιολογήσει κατ' ελάχιστο 2 εναλλακτικές επιλογές και να υπάρχει για αυτές μια ασθενής προτίμηση (weak preference).Τα αποτελέσματα του αλγορίθμου περιλαμβάνουν τις συναρτήσεις αξίας (value functions) για κάθε κριτήριο καθώς και τα βάρη σημαντικότητας (significance weights) κάθε κριτηρίου για κάθε χρήστη. Τα βάρη των κριτηρίων μέσω των οποίων μοντελοποιείται ο χρήστης, αποθηκεύονται σε ένα πίνακα και εκφράζουν το σύστημα αξιών του.

4.2.3 Συσταδοποίηση

Ένας αλγόριθμος συσταδοποίησης διαχωρίζει ένα σύνολο δεδομένων σε διακριτές ομάδες/συστάδες (clusters). Σκοπός είναι να βρεθούν οι φυσικές (ή ουσιαστικές) ομάδες που υπάρχουν στα δεδομένα. Πρόκειται για μια μη εποπτευόμενη τεχνική (unsupervised process) στόχος της οποίας είναι τα δεδομένα που ανήκουν στην ίδια ομάδα να μοιάζουν αρκετά ενώ τα δεδομένα διαφορετικών ομάδων να διαφέρουν το περισσότερο δυνατό. Υπάρχουν δύο κύριες κατηγορίες αλγορίθμων συσταδοποίησης: ιεραρχική και διαχωριστική. Οι αλγόριθμοι διαχωριστικής συσταδοποίησης χωρίζουν τα στοιχεία των δεδομένων σε μη επικαλυπτόμενες ομάδες έτσι ώστε κάθε στοιχείο να ανήκει ακριβώς σε μία ομάδα. Οι αλγόριθμοι ιεραρχικής συσταδοποίησης ομαδοποιούν διαδοχικά τα αντικείμενα μέσα σε συστάδες, παράγοντας ένα σύνολο από εμφωλευμένες συστάδες που είναι οργανωμένες σε ένα ιεραρχικό δέντρο. Σε αυτή τη φάση χρησιμοποιείται ο αλγόριθμος k-means ένας από τους πιο δημοφιλείς διαχωριστικούς αλγόριθμους συσταδοποίησης στον οποίο κάθε συστάδα σχετίζεται με ένα κεντρικό σημείο.

Υποθέτοντας ένα σύνολο δεδομένων $\{x_1,x_2,\ldots,x_n\}$, $x_n \in R^d$, ο k-means διαιρεί τα δεδομένα σε k διακριτές ομάδες C_1, C_2, …,C_k βελτιστοποιώντας κάθε φορά ένα συγκεκριμένο κριτήριο. Το πιο διαδεδομένο κριτήριο που εφαρμόζεται είναι το άθροισμα του τετραγωνικού σφάλματος (Sum of Squared Error-SSE, 4.2.3-1) μεταξύ κάθε σημείου x_i (i=1,2,…,n) και του κεντρικού σημείου m_j (j=1,2,…,k) ενός υποσυνόλου C_j, το οποίο περιλαμβάνει το x_i. Το κριτήριο συσταδοποίησης εξαρτάται από τα κέντρα των συστάδων $m_1,m_2,\ldots,m_k$.

$$\mathrm{SSE}(m_1, m_2, \ldots m_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} I(x_i \in C_j) |x_i - m_j|^2 \qquad \textbf{4.2.3-1}$$

Στη συγκεκριμένη εφαρμογή ο αλγόριθμος δέχεται ως είσοδο το πίνακα βαρών όλων των χρηστών που προέκυψε από τη φάση της μοντελοποίησης, ενώ τα αρχικά κεντρικά σημεία επιλέγονται τυχαία.

4.2.4 Δημιουργία συστάσεων

Έχοντας δημιουργήσει τις διαφορετικές ομάδες χρηστών, εφαρμόζεται ένας αλγόριθμος συστάσεων που χρησιμοποιεί τις πολυκριτήριες αξιολογήσεις και βασίζεται στη λογική του συνεργατικού φιλτραρίσματος σε κάθε διαφορετική ομάδα χρηστών.

Αρχικά υπολογίζεται η απόσταση μεταξύ του χρήστη αναφοράς u (χρήστης για τον οποίο ψάχνω σύσταση) και του χρήστη u' για το ίδιο αντικείμενο μέσω του τύπου ευκλείδειας απόστασης.

$$d_{uu'}=\sqrt{\sum_{n=1}^{k}(r_{un}-r_{u'n})^2} \qquad \textbf{4.2.4-1}$$

Όπου r_u και $r_{u'}$ είναι τα διανύσματα που περιέχουν τις βαθμολογίες των χρηστών u και u' για ένα αντικείμενο i αντίστοιχα. Λόγω των πολυκριτήριων δεδομένων τα r_u και $r_{u'}$ αναφέρονται σε διανύσματα k+1 μεγέθους όπου k τα κριτήρια συν τη συνολική βαθμολογία.

Στη συνέχεια υπολογίζεται η συνολική απόσταση μεταξύ δύο χρηστών σύμφωνα με το τύπο 4.2.4-2.

$$\mathrm{dist}(u,u')=\frac{1}{|U(u,u')|}\sum_{i\in U(u,u')} d_{uu'} \qquad \textbf{4.2.4-2}$$

Όπου U(u,u') ο συνολικός αριθμός των κοινών αντικειμένων που έχουν βαθμολογήσει και οι δύο χρήστες. Η συνολική απόσταση μεταξύ δύο χρηστών είναι ουσιαστικά η μέση τιμή των αποστάσεων των βαθμολογιών τους για όλα τα κοινά αντικείμενα και υφίσταται μόνο όταν το σύνολο αυτό είναι διάφορο του κενού.

Χρησιμοποιώντας τη συνολική απόσταση μεταξύ δύο χρηστών υπολογίζουμε την ομοιότητά τους με βάση το μέτρο:

$$\mathrm{simil}(u,u')=\frac{1}{1+\mathrm{dist}(u,u')} \qquad \textbf{4.2.4-3}$$

Το μέτρο ομοιότητας προσεγγίζει τη τιμή 0 όταν η απόσταση μεταξύ των χρηστών είναι πολύ μεγάλη, ενώ παίρνει τη τιμή 1 όταν οι βαθμολογίες των χρηστών στα κοινά τους αντικείμενα είναι ίσες.

Τέλος υπολογίζεται μια εκτιμώμενη συνολική βαθμολογία για κάθε αντικείμενο i που δεν έχει βαθμολογήσει ο χρήστης αναφοράς u σύμφωνα με το τύπο 4.2.4-4.

$$R(u,i)=\frac{1}{\sum_{u'\in C(u)} simil(u,u')}\cdot\sum_{u'\in C(u)}(simil(u,u')\cdot R(u',i)) \qquad \textbf{4.2.4-4}$$

Τα R(u,i) και R(u',i) αναφέρονται στη συνολική βαθμολογία. Το σύνολο C(u) περιέχει χρήστες που ανήκουν στη συστάδα του u και έχουν τουλάχιστον ένα κοινό αντικείμενο με αυτόν .Το ένα κοινό αντικείμενο μεταξύ των δύο χρηστών εξασφαλίζει ότι υφίσταται απόσταση άρα και το μέτρο της ομοιότητας μεταξύ τους. Στη περίπτωση όπου δεν έχει βαθμολογήσει κανείς το αντικείμενο i, τότε χρησιμοποιείται ένα νέο σύνολο C(u). Το νέο σύνολο αυτό περιέχει και χρήστες που ανήκουν στη $2^{η}$ κοντινότερη συστάδα του u (κοντινότερη με βάση την απόσταση των κέντρων). Ο αλγόριθμος συστάσεων περιγράφεται παρακάτω σε μορφή ψευδοκώδικα.

Αλγόριθμος συστάσεων

1. *Βρες όλους τους χρήστες u' που έχουν βαθμολογήσει το αντικείμενο i και ανήκουν στη συστάδα του u*
2. *Έλεγξε αν υπάρχει sim(u',u).Αν υπάρχει πρόσθεσε το χρήστη u' στο σύνολο C(u)*
3. *Αν C(u) =Ø επανάλαβε :*
4. *Χρησιμοποίησε τη κοντινότερη συστάδα του u με βάση την απόσταση των κέντρων*
5. *Πήγαινε στο βήμα 1*
6. *Μέχρι C(u) ≠Ø*
7. *Υπολόγισε R(u,i)*

Στη περίπτωση όπου χρειαστεί να χρησιμοποιήσουμε διαφορετική συστάδα παραπάνω από 2 φορές, σταματάει ο αλγόριθμος και σαν R(u,i) επιστρέφεται η μέση βαθμολογία των χρηστών της συστάδας του u.

Κεφάλαιο 5 : Αποτελέσματα

5.1 Περιγραφή δεδομένων

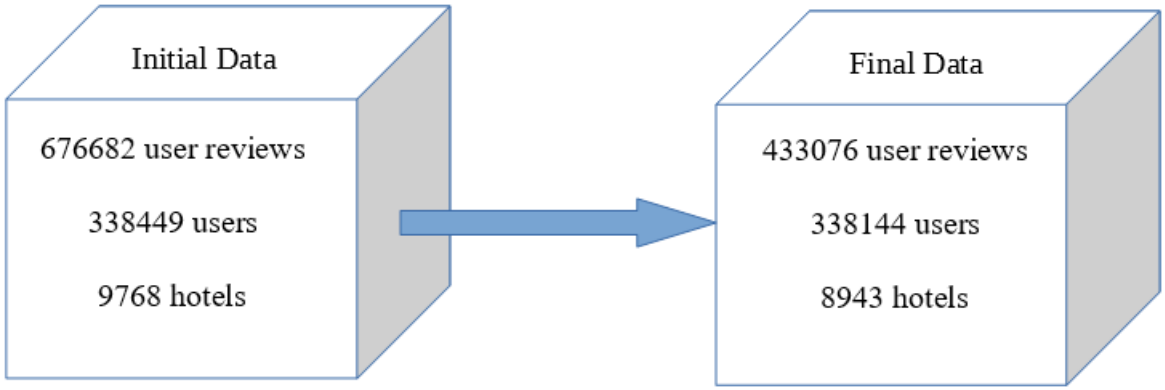
Για την αξιολόγηση του συστήματος συστάσεων που βασίζεται στη μεθοδολογία του κεφαλαίου 4 , έγινε εφαρμογή στο τομέα σύστασης ξενοδοχείων για ηλεκτρονική κράτηση με δεδομένα από την ιστοσελίδα tripadvisor. Στην ιστοσελίδα trip advisor οι χρήστες μπορούν να αξιολογήσουν τα ξενοδοχεία που έχουν επισκεφθεί στα παρακάτω χαρακτηριστικά : Service / Cleanliness / Value / Business Service (e.g Internet access / Check – in, front desk / Sleep Quality / Location / Rooms καθώς και να δώσουν τη συνολική τους βαθμολογία για το ξενοδοχείο στο χαρακτηριστικό Overall. Όλα τα χαρακτηριστικά μετρώνται σε 5-βάθμια κλίμακα όπου το 1 αντιστοιχεί σε Terrible , το 2 σε Poor, το 3 σε Average, το 4 σε Very good και το 5 σε Excellent. Κάθε χρήστης βέβαια μπορεί να βαθμολογήσει το ξενοδοχείο σε μερικά από τα παραπάνω χαρακτηριστικά. Το σετ δεδομένων που χρησιμοποιήθηκε περιέχει 7 κριτήρια καθώς για τα υπόλοιπα δεν υπήρχαν αρκετές αξιολογήσεις από τους χρήστες (οι περισσότεροι επέλεξαν να μην τα βαθμολογήσουν).Τα κριτήρια που χρησιμοποιούνται είναι :

- Service, που δηλώνει τη γενικότερη εξυπηρέτηση που έλαβαν οι πελάτες του ξενοδοχείου από το προσωπικό.
- Cleanliness, η καθαριότητα για το σύνολο του ξενοδοχείου.
- Value, η σχέση ποιότητας/τιμής του ξενοδοχείου.
- Sleep Quality, η ποιότητα του ξενοδοχείου σχετικά με τη διαμονή (άνεση, ηχομόνωση κ.τ.λ.).
- Location, η τοποθεσία του ξενοδοχείου.
- Rooms, κατά πόσο άρεσαν στους πελάτες τα δωμάτια του ξενοδοχείου (μέγεθος, διακόσμηση, θέα κ.τ.λ.).
- Overall, η συνολική βαθμολογία.

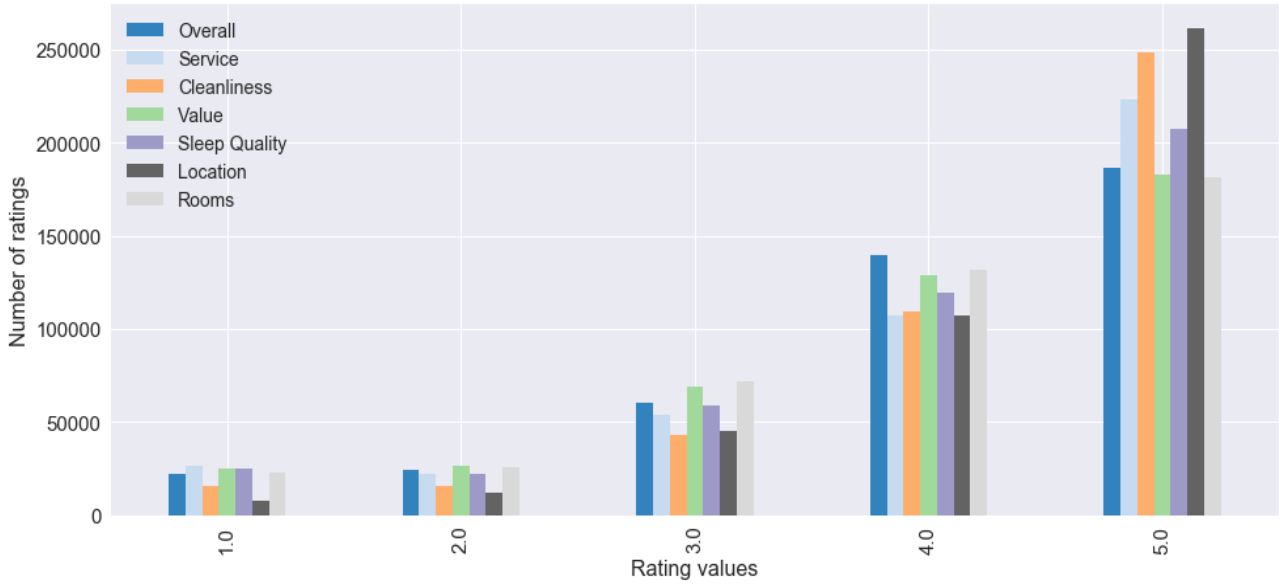
Πέραν από τις αξιολογήσεις, στο σετ δεδομένων έχουμε το όνομα χρήστη, τη τοποθεσία από την οποία πραγματοποίησε ο χρήστης την αξιολόγησή του καθώς και ένα μοναδικό id για κάθε διαφορετικό ξενοδοχείο. Τέλος σε κάθε διαφορετική αξιολόγηση που πραγματοποιείται, αποδίδεται ένα μοναδικό “review id”. Το σύνολο των αξιολογήσεων λοιπόν που περιέχει το σετ δεδομένων είναι ο συνολικός αριθμός των μοναδικών review-id. Τα δεδομένα που χρησιμοποιήθηκαν, αποκτήθηκαν από την ιστοσελίδα του εργαστηρίου Database and Information Systems του πανεπιστημίου του Illinois (**http://times.cs.uiuc.edu/~wang296/Data/**) και έχουν συλλεχθεί σε διαφορετικές χρονικές στιγμές από την ιστοσελίδα trip advisor. Τα δεδομένα που υπάρχουν στην ιστοσελίδα του εργαστηρίου βρίσκονται σε μορφή αρχείων JSON (JavaScript Object Notation), και συγκεκριμένα για κάθε ξενοδοχείο υπάρχει ένα αρχείο json που περιέχει τις πληροφορίες που αναφέρθηκαν παραπάνω. Για την εφαρμογή της μεθοδολογίας που αναπτύχθηκε στο κεφάλαιο 4, τα δεδομένα αξιολογήσεων κάθε ξενοδοχείου συγκεντρώθηκαν σε ένα ενιαίο αρχείο csv (comma-

separated values) και αναπαριστώνται σε μορφή πίνακα, όπου κάθε γραμμή περιέχει ξεχωριστές αξιολογήσεις (review id), και στήλες το όνομα χρήστη, το id του ξενοδοχείου και τις βαθμολογίες στα 7 χαρακτηριστικά.

Αρχικά πραγματοποιήθηκε καθαρισμός των δεδομένων που επέφερε μείωση κατά 36%, για να αφαιρεθούν πανομοιότυπες εγγραφές, αλλά και εγγραφές που έχουν έστω και ένα κενό στοιχείο. Ως πανομοιότυπες εγγραφές θεωρούνται αυτές οι οποίες έχουν κοινό "review id". Οι εγγραφές με κοινό review-id περιέχουν τις ίδιες αξιολογήσεις για ένα ξενοδοχείο από ένα χρήστη και εμφανίζονται στα δεδομένα περισσότερο από μία φορά. Τελικά το σετ δεδομένων που χρησιμοποιείται όπως φαίνεται στο σχήμα 5.1-1 περιέχει 338144 χρήστες, 8943 διαφορετικά ξενοδοχεία και συνολικά 433076 αξιολογήσεις στις οποίες όλοι οι χρήστες έχουν δώσει βαθμολογία σε όλα τα κριτήρια.



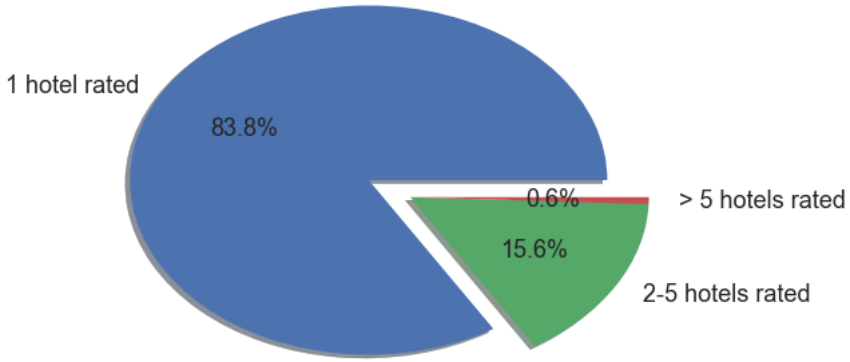
Σχήμα 5.1-1: Καθαρισμός δεδομένων



Σχήμα 5.1-2: Κατανομή των βαθμολογιών στο σύνολο των δεδομένων

Όπως φαίνεται στο σχήμα 5.1-2 οι περισσότεροι χρήστες έχουν δώσει υψηλές βαθμολογίες σε όλα τα χαρακτηριστικά. Οι περισσότερες αξιολογήσεις έχουν βαθμολογία 5, ενώ οι λιγότερες 1.

Στη συνέχεια οι χρήστες κατηγοριοποιήθηκαν σε ομάδες ανάλογα με τον αριθμό των ξενοδοχείων που είχαν βαθμολογήσει. Όπως φαίνεται στο σχήμα 5.1-3, οι περισσότεροι χρήστες έχουν βαθμολογήσει 1 ξενοδοχείο,ενώ το ποσοστό των χρηστών που έχουν βαθμολογήσει περισσότερα από 5 ξενοδοχεία είναι πολύ μικρό.



Σχήμα 5.1-3: Ποσοστά χρηστών ανάλογα με τον αριθμό των ξενοδοχείων που έχουν βαθμολογήσει

	Service	Cleanliness	Overall	Value	Sleep Quality	Location	Rooms
Service	1.000000	0.688347	0.808412	0.712501	0.652258	0.463143	0.686292
Cleanliness	0.688347	1.000000	0.768932	0.655920	0.689958	0.442063	0.757015
Overall	0.808412	0.768932	1.000000	0.787778	0.764244	0.521679	**0.819391**
Value	0.712501	0.655920	0.787778	1.000000	0.671384	0.475834	0.712505
Sleep Quality	0.652258	0.689958	0.764244	0.671384	1.000000	0.431569	0.757443
Location	0.463143	0.442063	0.521679	0.475834	**0.431569**	1.000000	0.447383
Rooms	0.686292	0.757015	0.819391	0.712505	0.757443	0.447383	1.000000

Πίνακας 5.1-1: Συντελεστής συσχέτισης Pearson μεταξύ των μεταβλητών των δεδομένων

Στο πίνακα 5.1-1 παρατηρούμε τις συσχετίσεις των μεταβλητών με βάση το συντελεστή Pearson(-1 αρνητική συσχέτιση, 1 θετική συσχέτιση). Η μεγαλύτερη τιμή παρατηρείται μεταξύ των μεταβλητών Overall και Rooms , συντελεστής Pearson : 0.819391, ενώ η μικρότερη τιμή παρατηρείται μεταξύ των μεταβλητών Location και Sleep Quality : 0.431569.

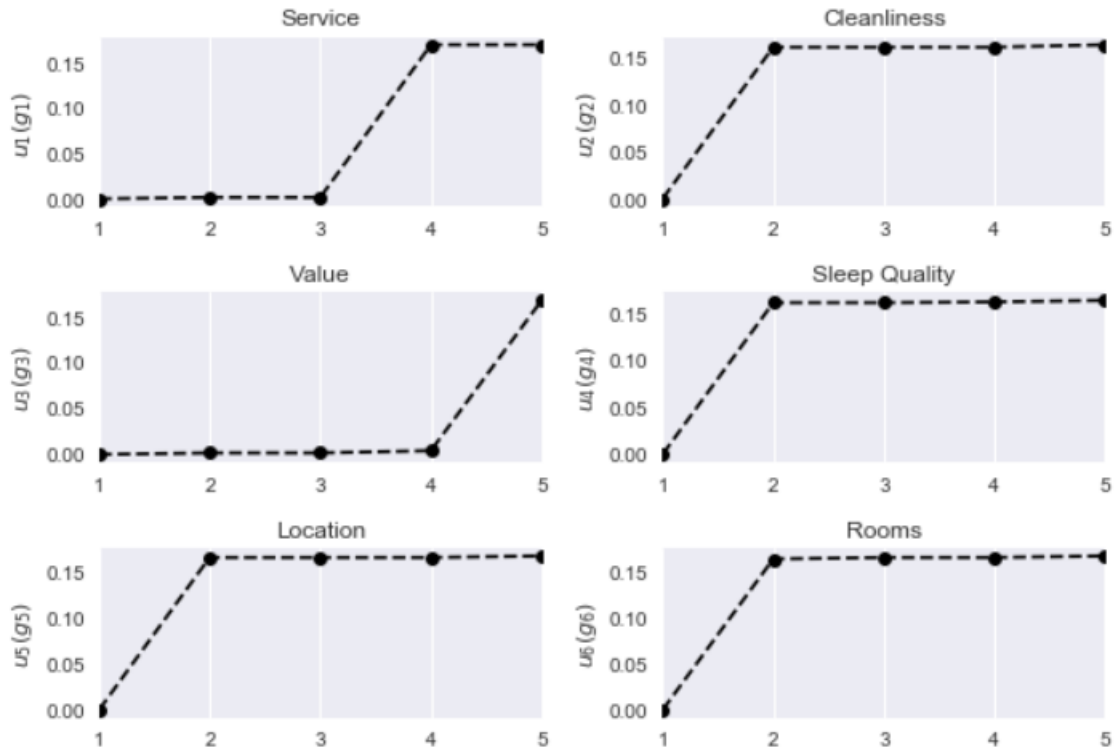
5.2 Αποτελέσματα UTASTAR

Για την εφαρμογή του αλγορίθμου UTASTAR , είναι απαραίτητο ο κάθε χρήστης να έχει βαθμολογήσει περισσότερα από δύο ξενοδοχεία, και να υπάρχει για αυτά μια συνολική διάταξη από το καλύτερο στο χειρότερο. Πραγματοποιήθηκε λοιπόν μετασχηματισμός στα δεδομένα όπως αναφέρεται στη παράγραφο 4.2.1, στον οποίο η βαθμολογία στο χαρακτηριστικό Overall χρησιμοποιείται για τη δημιουργία της αρχικής διάταξης των εναλλακτικών ξενοδοχείων.

Initial Data Form									Final Data Form							
AuthorID	HotelID	Service	Cleanliness	Value	Sleep Quality	Location	Rooms	Overall	AuthorID	HotelID	Service	Cleanlines	Value	Sleep Qua	Location	Rooms	Ranking
1	1010527	5	5	5	5	5	5	5	1	1010527	5	5	5	5	5	5	1
	121998	5	5	5	5	5	5	5		121998	5	5	5	5	5	5	1
	230405	4	5	4	5	4	4	4		230405	4	5	4	5	4	4	2
	2515626	5	5	4	5	5	4	5		2515626	5	5	4	5	5	4	1
	1022792	4	4	4	2	4	3	3		1022792	4	4	4	2	4	3	3
	2515527	5	5	5	4	3	3	5		2515527	5	5	5	4	3	3	1
	2072559	4	4	4	4	4	4	4		2072559	4	4	4	4	4	4	2
	268660	5	5	3	5	3	4	5		268660	5	5	3	5	3	4	1
	2515818	5	5	5	5	5	5	5		2515818	5	5	5	5	5	5	1
	190072	1	1	1	1	5	1	1		190072	1	1	1	1	5	1	5
2	2515499	5	5	3	5	5	5	5	2	2515499	5	5	3	5	5	5	1
	189076	2	1	2	3	3	2	2		189076	2	1	2	3	3	2	4

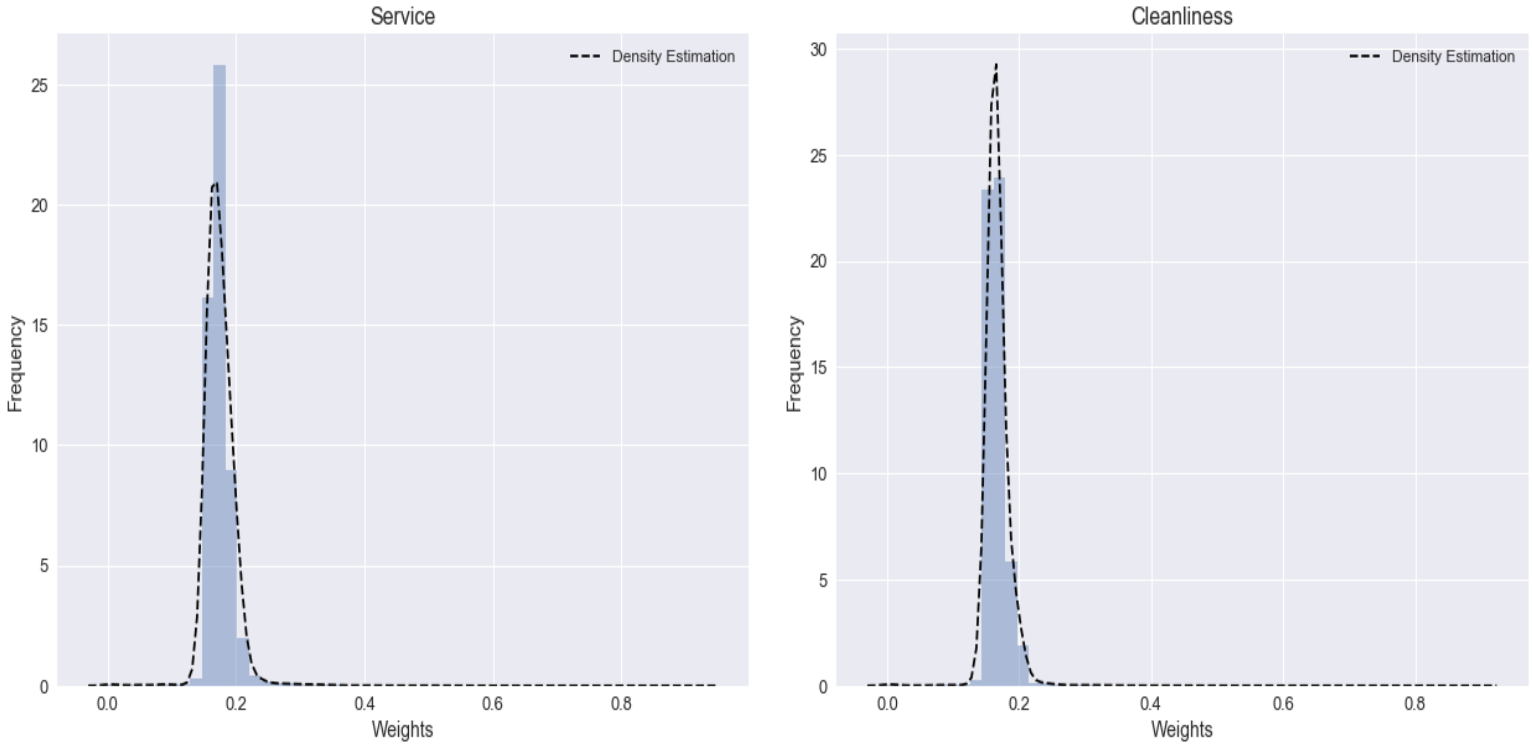
Για την εφαρμογή του αλγορίθμου είναι επιπλέον απαραίτητο να ορισθούν τα χαρακτηριστικά των κριτηρίων. Όλα τα κριτήρια είναι αύξοντα και βαθμολογούνται σε μια ποιοτική κλίμακα από το 1 έως το 5. Επιλέχθηκε λοιπόν ένα σετ χρηστών από το σύνολο των δεδομένων οι οποίοι έχουν βαθμολογήσει 2 ή περισσότερα ξενοδοχεία και μέσα από τις βαθμολογίες τους στο χαρακτηριστικό Overall να μπορεί να προκύψει μια αρχική προδιάταξη. Το σετ αυτό αποτελείται από 15892 χρήστες, 6318 διαφορετικά ξενοδοχεία και συνολικά 51923 αξιολογήσεις. Σαν είσοδο του αλγορίθμου λαμβάνεται ο πολυκριτήριος πίνακας για κάθε χρήστη, ενώ σαν αποτέλεσμα αποθηκεύουμε τα βάρη των κριτηρίων για κάθε χρήστη σε ένα πίνακα 15892×6 που θα χρησιμοποιηθεί έπειτα για τη δημιουργία προφίλ χρηστών.

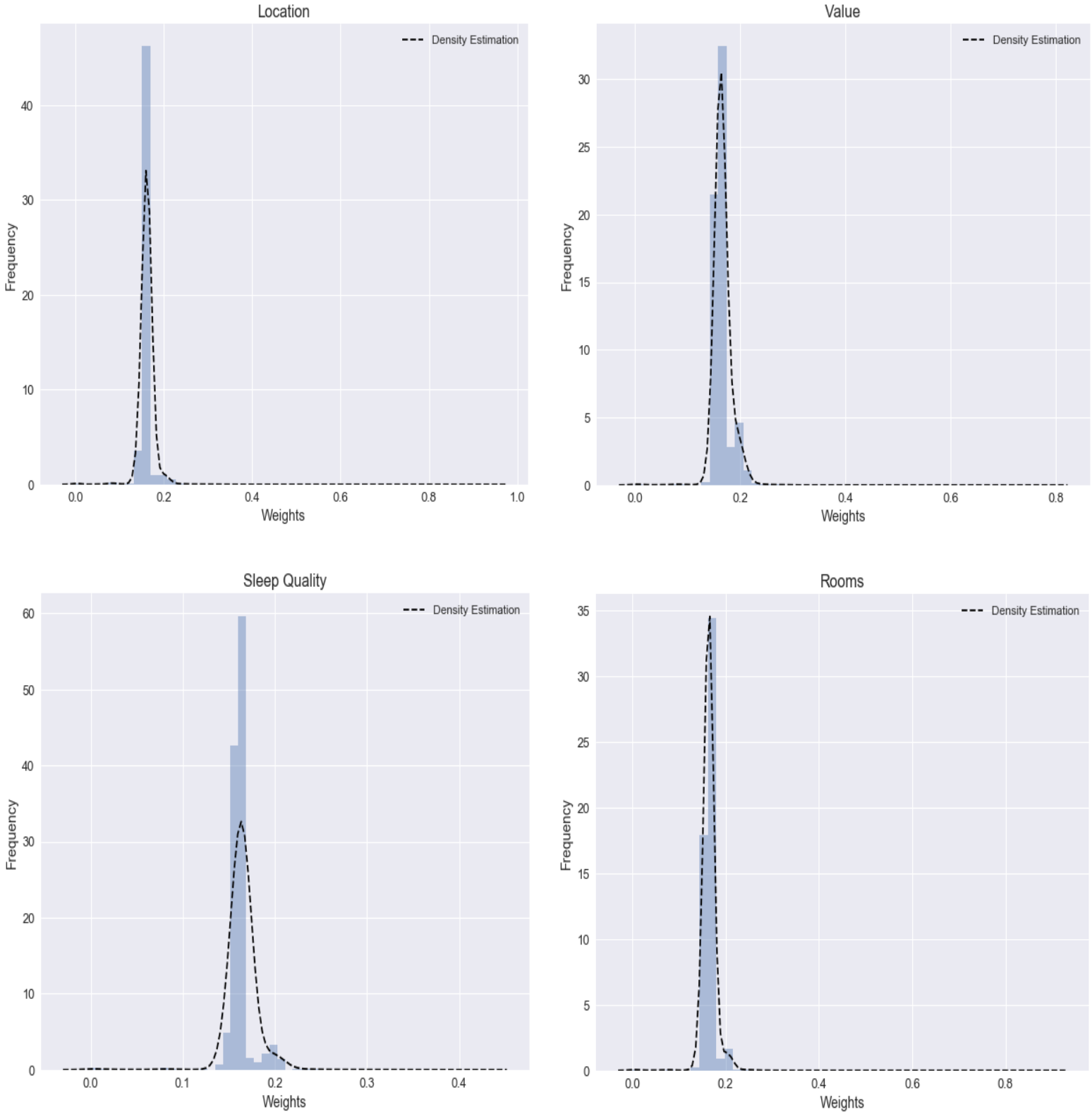
Στο σχήμα 5.2-1 παρατηρούμε τις μερικές (περιθώριες) συναρτήσεις αξίας που προέκυψαν από την εφαρμογή της UTASTAR για ένα τυχαίο χρήση.



Σχήμα 5.2-1: Μερικές συναρτήσεις αξίας

Στο σχήμα 5.2-2 επίσης παρατηρούμε τις κατανομές των βαρών για κάθε κριτήριο για το σύνολο των χρηστών.





Σχήμα 5.2-2: Κατανομή βαρών των κριτηρίων

Στο πίνακα 5.2-1 παρατηρούμε τη μέγιστη, ελάχιστη και μέση τιμή, καθώς και τη τυπική απόκλιση των βαρών του κάθε κριτηρίου.

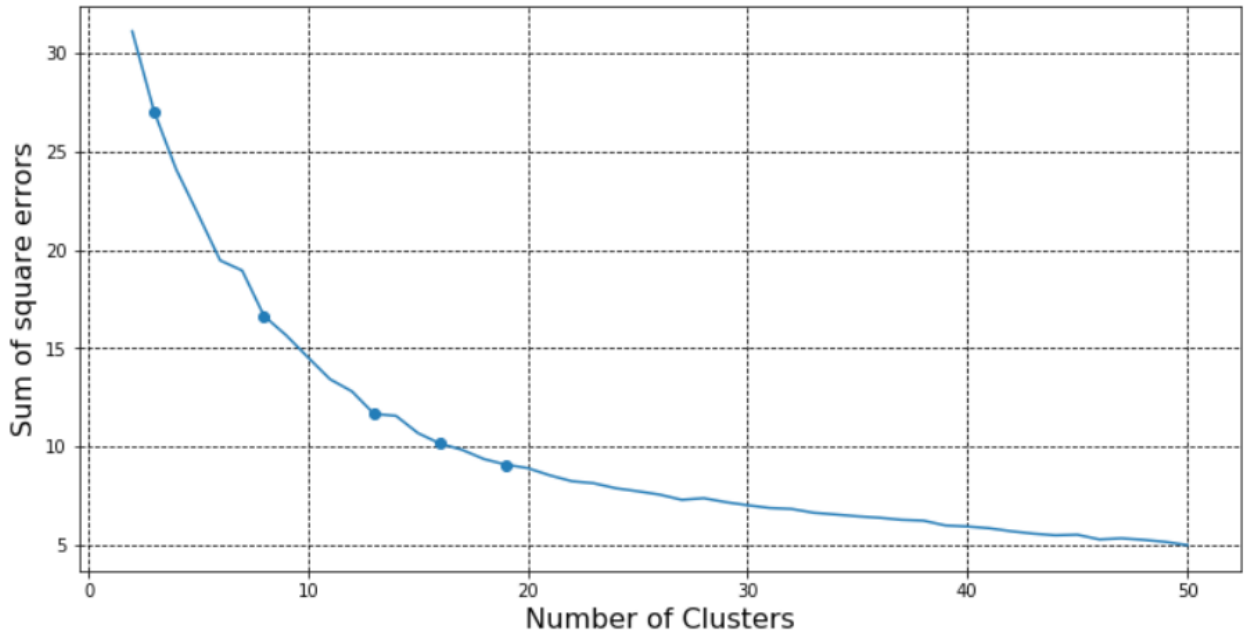
	Service	Cleanliness	Value	Sleep Quality	Location	Rooms
count	15892	15892	15892	15892	15892	15892
mean	0.174781	0.166948	0.166730	0.164412	0.161965	0.165164
std	0.024785	0.019804	0.017685	0.015761	0.020075	0.014928
min	0	0	0	0	0	0
max	0.918333	0.893333	0.791667	0.421667	0.943333	0.900000

Πίνακας 5.2-1: Στατιστικά των βαρών των κριτηρίων

Όπως φαίνεται στο πίνακα 5.2-1 η μέγιστη τιμή δίδεται στο κριτήριο Location 0.943, ενώ η ελάχιστη στο κριτήριο Sleep Quality, 0.421. Κατά μέσο όρο το κριτήριο Service παίρνει μεγαλύτερες τιμές βάρους απ' ότι τα υπόλοιπα κριτήρια, ενώ έχει και τη μεγαλύτερη τυπική απόκλιση.

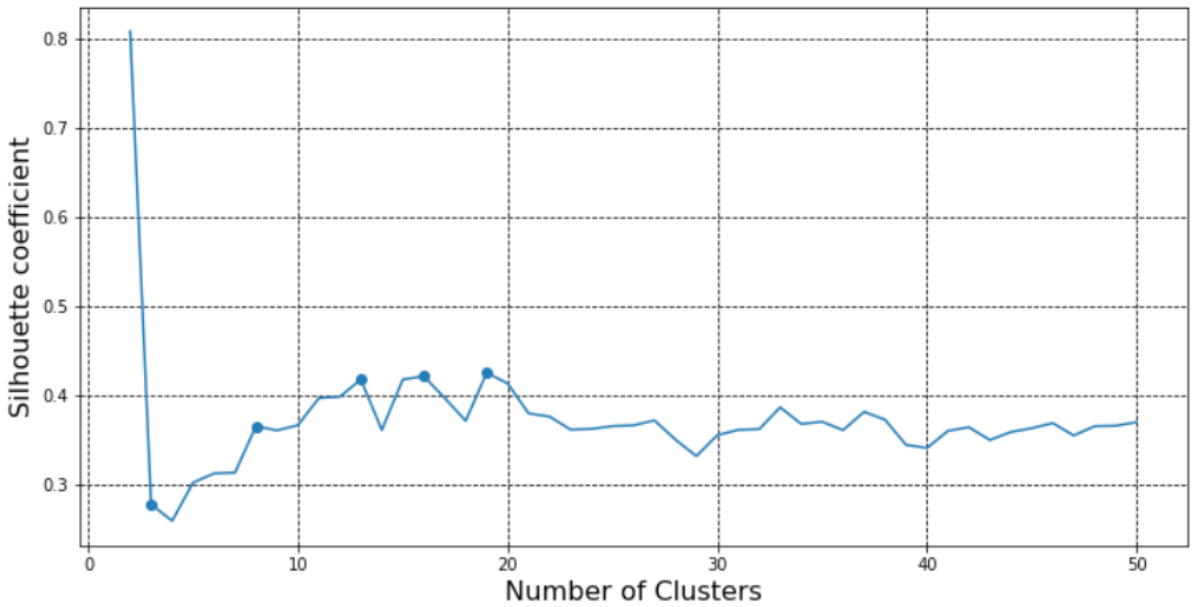
5.3 Αποτελέσματα συσταδοποίησης

Η συσταδοποίηση πραγματοποιήθηκε μέσω του αλγορίθμου k-means. Ως είσοδος του αλγορίθμου λαμβάνεται ο πίνακας βαρών των κριτηρίων (15892×6) , ενώ πρέπει να ορισθούν παράμετροι όπως ο αριθμός των συστάδων που θα δημιουργηθούν, καθώς και τα αρχικά κέντρα των συστάδων. Για την επιλογή του αριθμού των συστάδων, ο αλγόριθμος k means υλοποιήθηκε για 2 έως και 50 συστάδες, με τυχαία αρχικοποίηση των κέντρων. Σε κάθε περίπτωση μετρήθηκε ο μέσος συντελεστής Silhouette καθώς και το άθροισμα των αποστάσεων της κάθε εγγραφής από το κοντινότερο κέντρο(sum of square errors-sse).



Σχήμα 5.3-1: Άθροισμα αποστάσεων κάθε εγγραφής από το κοντινότερο κέντρο σε συνάρτηση του αριθμού των συστάδων

Στο σχήμα 5.3-1 παρατηρούμε μείωση του sse με αύξηση του αριθμού των συστάδων, καθώς σε κάθε στάδιο του αλγορίθμου , επιδιώκεται η μείωση των αποστάσεων των εγγραφών από τα κοντινότερα κέντρα. Στα επιλεγμένα σημεία πάνω στη καμπύλη φαίνονται οι αντίστοιχες τιμές για 3, 8, 13, 16 και 19 συστάδες. Με συνεχή αύξηση του αριθμού των συστάδων, θα έχουμε και μείωση του sse.



Σχήμα 5.3-2: Μέσος συντελεστής Silhouette σε συνάρτηση του αριθμού των συστάδων

Ο συντελεστής silhouette υπολογίζεται ως εξής :

$$s(i)=\frac{b(i)-a(i)}{\max[a(i),b(i)]} \qquad \textbf{5.3-1}$$

Όπου a(i) είναι η μέση απόσταση του στοιχείου i από όλα τα υπόλοιπα στοιχεία που ανήκουν στη συστάδα του και b(i) η μέση απόσταση του στοιχείου i από όλα τα στοιχεία της κοντινότερης συστάδας του. Ο μέσος συντελεστής silhouette για όλα τα στοιχεία i ,δείχνει πόσο "στενά συνδεδεμένα" είναι τα στοιχεία που ανήκουν σε κάθε συστάδα. Στο σχήμα 5.3-2 παρατηρούμε ότι ο μέσος συντελεστής silhouette παραμένει περίπου ίδιος με μερικά τοπικά ελάχιστα και μέγιστα μετά τις 10 συστάδες.

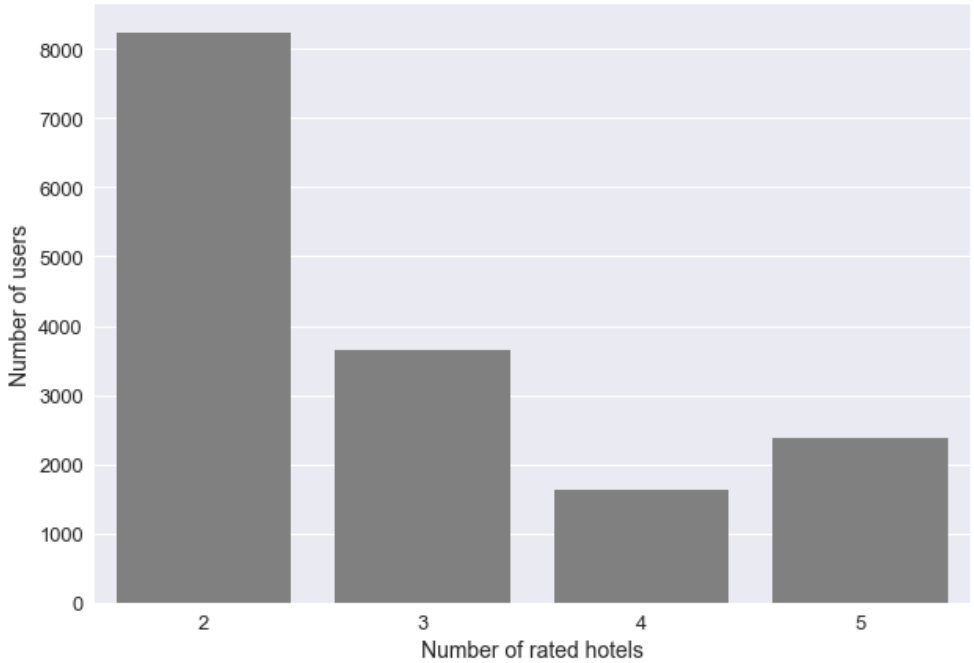
Η συσταδοποίηση των χρηστών με βάση τα βάρη που έχουν αποδοθεί στα κριτήρια μέσω της UTASTAR, πραγματοποιήθηκε με 19 συστάδες ,όπου ο μέσος συντελεστής silhouette σε αυτή τη περίπτωση είναι 0.4253 και το άθροισμα των αποστάσεων κάθε στοιχείου από το κοντινότερο κέντρο του είναι 9.1091.Τα κέντρα των 19 συστάδων φαίνονται στο πίνακα 5.3-1 και για κάθε συστάδα έχει σημειωθεί η μέγιστη τιμή. Παρατηρούμε λοιπόν ότι οι χρήστες που ανήκουν στη 1η συστάδα θεωρούν πιο σημαντικό κριτήριο το Sleep Quality, στη 2η το Service κ.τ.λ.

	Service	**Cleanliness**	**Value**	**Sleep Quality**	**Location**	**Rooms**
0	0.161177	0.158333	0.157786	0.201808	0.158980	0.161916
1	0.199066	0.160214	0.161649	0.159860	0.158339	0.160872
2	0.173201	0.165909	0.165631	0.165697	0.163517	0.166046
3	0.164170	0.157489	0.157608	0.157096	0.204006	0.159630
4	0.110556	0.065972	0.199306	0.093333	0.464722	0.066111
5	0.348111	0.136429	0.096159	0.150452	0.139857	0.128992
6	0.030333	0.671667	0.131667	0.035000	0.109667	0.021667
7	0.073376	0.328746	0.178184	0.140605	0.128440	0.150648
8	0.160616	0.158750	0.197314	0.161247	0.159554	0.162518
9	0.024167	0.036389	0.031111	0.042500	0.819306	0.046528
10	0.759667	0.011667	0.093167	0.044000	0.047500	0.044000
11	0.169826	0.054010	0.348281	0.140330	0.161615	0.125938
12	0.158898	0.191489	0.162367	0.162844	0.160064	0.164338
13	0.284528	0.177230	0.157628	0.038478	0.173269	0.168867
14	0.048333	0.006667	0.119167	0.019167	0.001667	0.805000
15	0.183248	0.213236	0.149818	0.151630	0.145044	0.157024
16	0.161158	0.159304	0.157945	0.158079	0.156616	0.206898
17	0.001667	0.098333	0.729167	0.084167	0.002500	0.084167
18	0.273094	0.165436	0.165792	0.171275	0.047135	0.177269

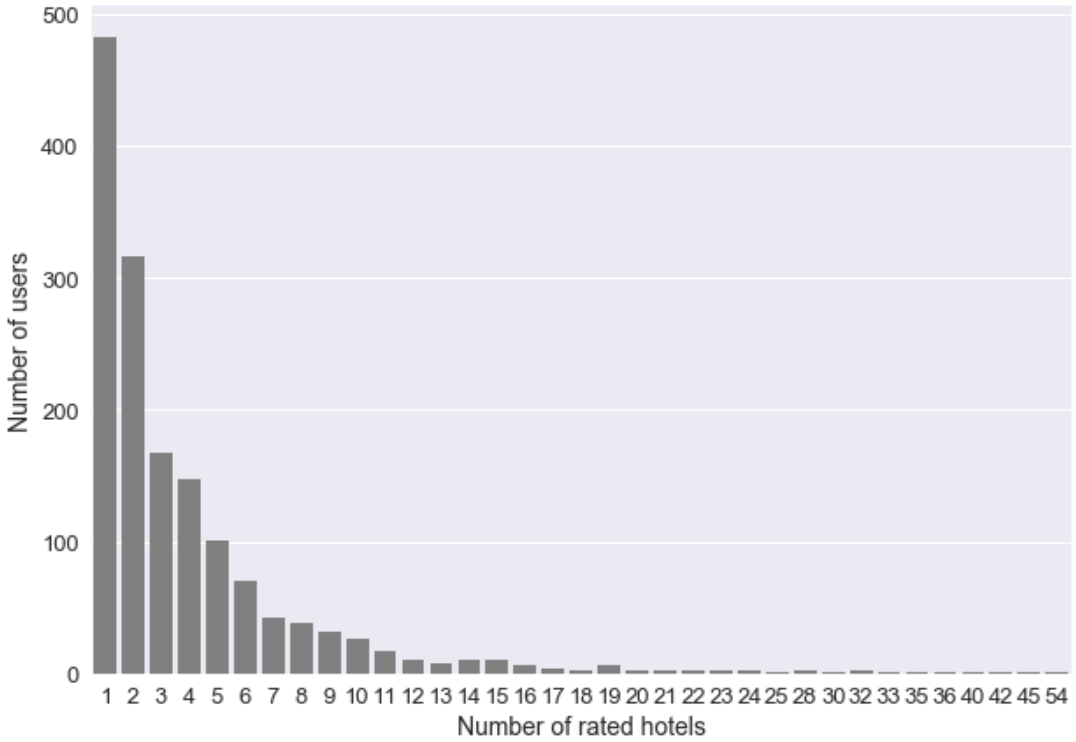
Πίνακας 5.3-1: Κέντρα των 19 συστάδων

5.4 Αποτελέσματα της φάσης συστάσεων

Σύμφωνα με τη μεθοδολογία που αναπτύχθηκε στο κεφάλαιο 4, στη φάση συστάσεων υπολογίζεται από το σύστημα μία βαθμολογία R(u,i) για κάθε αντικείμενο i που δεν έχει βαθμολογήσει ο χρήστης u. Για την αξιολόγηση των αποτελεσμάτων οι βαθμολογίες R(u,i) υπολογίσθηκαν σε ένα μέρος των δεδομένων το οποίο ονομάζουμε test set. Το test set αποτελεί ένα κομμάτι των δεδομένων το οποίο αφαιρείται από το σύνολο. Το υπόλοιπο μέρος των δεδομένων, train set χρησιμοποιείται από τον αλγόριθμο για τον υπολογισμό των R(u,i). Ουσιαστικά το train set αποτελεί τη βάση δεδομένων του συστήματος, ενώ το test set περιέχει αξιολογήσεις βάσει των οποίων θα μετρηθεί η ακρίβεια των συστάσεων. Από το σύνολο των δεδομένων που πήραμε από τη φάση της συσταδοποίησης, όλοι οι χρήστες που έχουν βαθμολογήσει από 2-5 ξενοδοχεία συμπεριλαμβάνονται στο train set. Το test set περιλαμβάνει όλους τους χρήστες που ανήκουν στο train set και έχουν αξιολογήσει περισσότερα από 5 ξενοδοχεία. Με το διαχωρισμό αυτό, κάθε χρήστης που ανήκει στο test set, ανήκει υποχρεωτικά και στο train set,με διαφορετικές αξιολογήσεις, οπότε το σύστημα έχει επαρκή πληροφορία για να εξάγει για αυτόν συστάσεις. Το train set λοιπόν περιέχει 15892 χρήστες, 5931 ξενοδοχεία και συνολικά 45837 αξιολογήσεις, ενώ το test set περιέχει 1523 χρήστες ,2597 ξενοδοχεία και 6086 αξιολογήσεις.

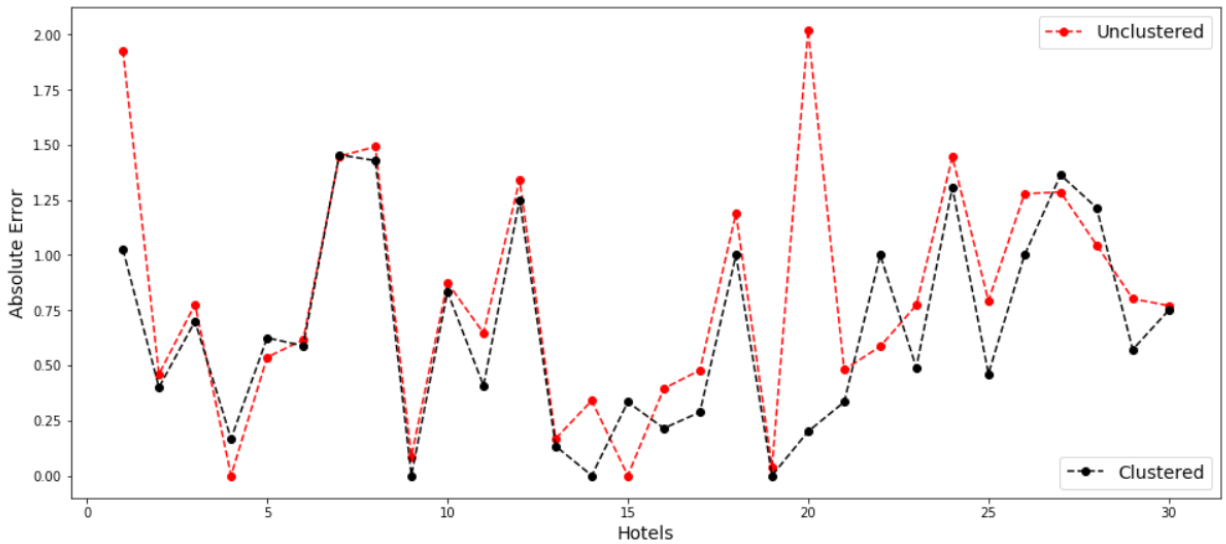


Σχήμα 5.4-1: Αριθμός χρηστών του train set ανάλογα με τον αριθμό των ξενοδοχείων που έχουν βαθμολογήσει



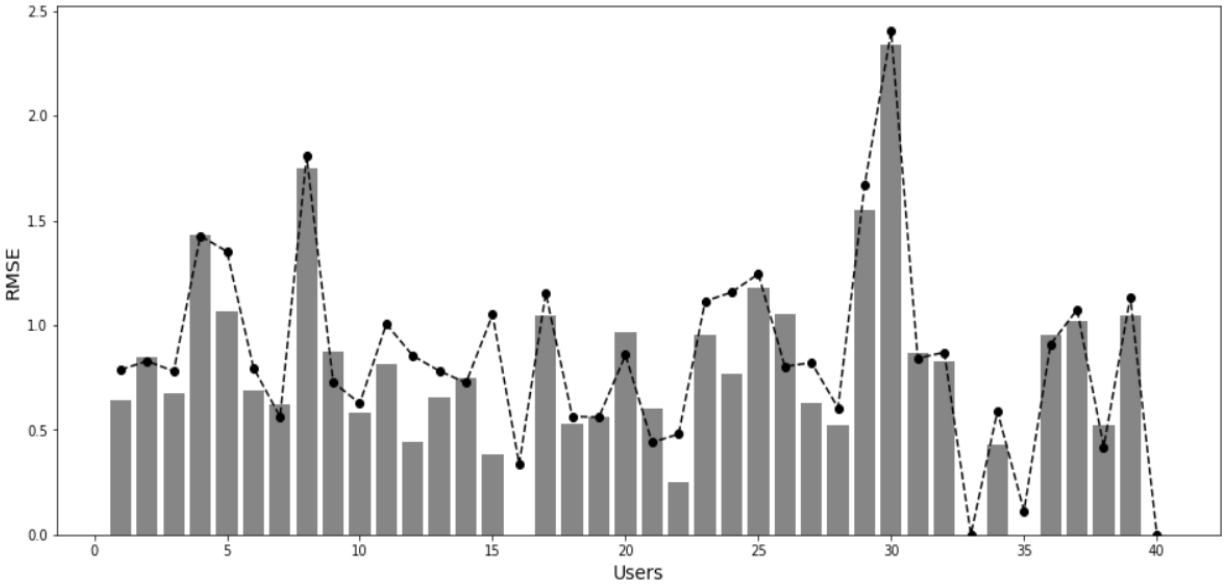
Σχήμα 5.4-2: Αριθμός χρηστών του test set ανάλογα με τον αριθμό των ξενοδοχείων που έχουν βαθμολογήσει

Αρχικά υπολογίζονται για κάθε χρήστη του test set στατιστικά μέτρα ακρίβειας, όπως το μέσο απόλυτο σφάλμα και η ρίζα του μέσου τετραγωνικού σφάλματος (MAE , RMSE), όπως αυτά περιγράφηκαν στο κεφάλαιο 3.



Σχήμα 5.4-3: Απόλυτο σφάλμα μεταξύ εκτιμώμενης και πραγματικής βαθμολογίας για ένα τυχαίο χρήστη

Στο σχήμα 5.4-3 παρατηρούμε την απόλυτη διαφορά, μεταξύ των εκτιμώμενων από το σύστημα βαθμολογιών και των πραγματικών, στη περίπτωση συσταδοποίησης των χρηστών αλλά και στη περίπτωση όπου δεν έχουμε πραγματοποιήσει συσταδοποίηση, για ένα τυχαίο χρήστη ο οποίος έχει αξιολογήσει 30 διαφορετικά ξενοδοχεία. Παρατηρούμε ότι το σύστημα στις περισσότερες περιπτώσεις δίδει εκτιμώμενες βαθμολογίες με μικρότερο σφάλμα όταν έχει πραγματοποιηθεί συσταδοποίηση.



Σχήμα 5.4-4: Ρίζα μέσου τετραγωνικού σφάλματος για 40 τυχαίους χρήστες

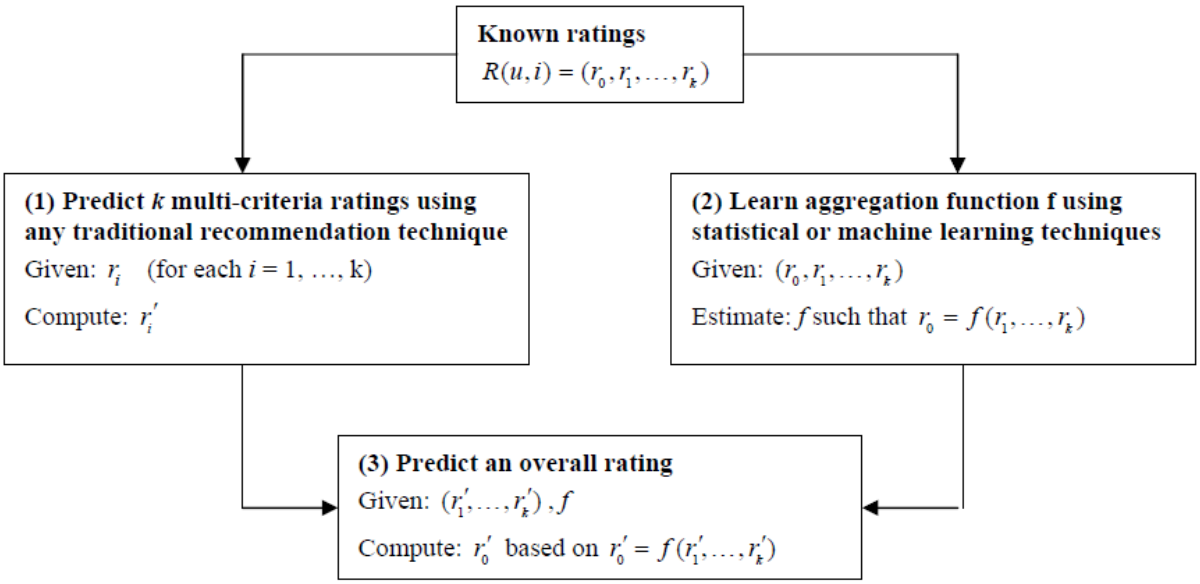
Στο σχήμα 5.4-4 παρατηρούμε τη τιμή του $RMSE_u$ για 40 τυχαίους χρήστες. Οι ράβδοι αντιστοιχούν σε τιμές σφάλματος στη περίπτωση όπου έχουν δημιουργηθεί 19 συστάδες, ενώ οι τελείες αντιστοιχούν σε σφάλματα για τη περίπτωση όπου δεν πραγματοποιείται συσταδοποίηση. Το σύστημα πετυχαίνει μικρότερο σφάλμα κατά μέσο όρο όταν πραγματοποιούμε συσταδοποίηση, παρατηρούμε όμως ότι για αρκετούς χρήστες η συσταδοποίηση επιφέρει μεγαλύτερο σφάλμα.

5.5 Σύγκριση βάσει αποτελεσμάτων με άλλες μεθοδολογίες

Για τη καλύτερη κατανόηση της απόδοσης του συστήματος ,συγκρίνονται τα αποτελέσματα του με τα αποτελέσματα δύο επιπλέον μεθοδολογιών που εφαρμόσθηκαν στα ίδια δεδομένα. Σε κάθε περίπτωση μετρώνται στατιστικά μέτρα ακρίβειας, αλλά και μέτρα ακρίβειας κατηγοριοποίησης που αναλύθηκαν στο κεφάλαιο 3, για το σύνολο των αξιολογήσεων που περιέχει το test set.

5.5.1 Συνθετική συνάρτηση (Aggregation Function)

Η μεθοδολογία αυτή βασίζεται στη μεθοδολογία που αναπτύχθηκε από τους G. Adomavicius και YoungOk Kwon (“New Recommendation Techniques for Multi-Criteria Rating Systems”,2007) και χρησιμοποιεί μια συνθετική συνάρτηση της μορφής $r_o= f(r_1,r_2,\ldots,r_k)$, για το προσδιορισμό της συνολικής βαθμολογίας,λαμβάνοντας υπόψη τις βαθμολογίες στα k κριτήρια.



Σχήμα 5.5-1: Επισκόπηση της προσέγγισης με συνθετική συνάρτηση

Πηγή: Adomavicius G., Kwon Y. (2007). New Recommendation Techniques for Multi-Criteria Rating Systems, *IEEE Intelligent Systems,* Volume 22, Issue 3.

Στη συγκεκριμένη εφαρμογή χρησιμοποιείται ένα μοντέλο γραμμικής παλινδρόμησης (5.5.1-1) για το προσδιορισμό της συνολικής βαθμολογίας, ενώ ο προσδιορισμός των βαθμολογιών στα k κριτήρια αναλύεται σε k προβλήματα σύστασης με συνεργατικό φιλτράρισμα μονής βαθμολογίας (single rating collaborative filtering).

$$r_0=\sum_{i=1}^{k} w_i r_i+\varepsilon \qquad \textbf{5.5.1-1}$$

Αναλυτικότερα χρησιμοποιείται μια σταθμισμένη προσέγγιση του μέτρου cosine similarity (5.5.1-2) για τον υπολογισμό ομοιότητας μεταξύ χρηστών με βάση τη βαθμολογία τους σε κάθε ένα κριτήριο ξεχωριστά. Το μοντέλο γραμμικής παλινδρόμησης χρησιμοποιεί έπειτα τις εκτιμώμενες από το σύστημα βαθμολογίες στα k κριτήρια για τον υπολογισμό της τελικής βαθμολογίας, βάσει της οποίας προτείνονται αντικείμενα στο χρήστη.

$$\text{simil}(u,u') = A \frac{\sum_{i \in I(u,u')} R(u,i) \cdot R(u',i)}{\sqrt{\sum_{i \in I(u,u')} R(u,i)^2} \cdot \sqrt{\sum_{i \in I(u,u')} R(u',i)^2}} \quad \textbf{5.5.1-2}$$

$$A = \frac{U(u,u')}{U(u)} \quad \textbf{5.5.1-3}$$

Ο συντελεστής A ορίζεται ως ο λόγος των κοινών αντικειμένων μεταξύ των χρηστών u και u' δια το σύνολο των αντικειμένων που έχει βαθμολογήσει ο χρήστης u. Χρησιμοποιείται επειδή το παραδοσιακό μέτρο cosine similarity δίδει πάντα τιμές ίσες με 1 (απόλυτη ομοιότητα) όταν οι χρήστες έχουν βαθμολογήσει 1 μόνο κοινό αντικείμενο, ανεξάρτητα από τη τιμή της βαθμολογίας.

5.5.2 Προσέγγιση με βάση το συνεργατικό φιλτράρισμα και χρήση μονής βαθμολογίας (Single rating collaborative filtering approach)

Στη προσέγγιση αυτή χρησιμοποιείται μόνο η συνολική βαθμολογία για τον υπολογισμό της ομοιότητας μεταξύ χρηστών και δεν λαμβάνονται υπόψη στους υπολογισμούς οι βαθμολογίες στα υπόλοιπα κριτήρια. Για τον υπολογισμό της ομοιότητας μεταξύ χρηστών χρησιμοποιήθηκαν οι τύποι 5.5.1-2 , 5.5.1-3.

Στο πίνακα 5.5-1 φαίνονται τα αποτελέσματα των 3 μεθόδων με βάση στατιστικά μέτρα ακρίβειας και μέτρα ακρίβειας κατηγοριοποίησης. Η κατηγοριοποίηση των ξενοδοχείων σε "Recommended" και "Not Recommended" για κάθε χρήστη έγινε με βάση την εκτιμώμενη συνολική βαθμολογία. Συγκεκριμένα, όλα τα ξενοδοχεία για τα οποία εκτιμήθηκε συνολική βαθμολογία 4 και 5 ,ανήκουν στη κατηγορία "Recommended", ενώ όλα τα υπόλοιπα στη κατηγορία "Not Recommended".Τέλος και οι 3 μεθοδολογίες εφαρμόσθηκαν για τη περίπτωση δημιουργίας 19 συστάδων και για τη περίπτωση όπου δεν πραγματοποιείται συσταδοποίηση.

	MAE	**RMSE**	**Precision**	**Recall**	**F1**	
1)	**0,6936**	**0,7802**	**0,8268**	**0,8547**	**0,8405**	**19 clusters**
2)	0,7307	0,8184	0,8165	0,8536	0,8346	
3)	0,9025	1,0135	0,7271	0,7553	0,7409	
1)	**0,7206**	**0,8286**	**0,8158**	**0,838**	**0,8267**	**unclustered**
2)	0,7445	0,8352	0,8163	0,8323	0,8242	
3)	0,9236	1,0983	0,7023	0,7307	0,7162	

Πίνακας 5.5-1: Αποτελέσματα αξιολόγησης μεθοδολογιών

Με 1) συμβολίζεται η μεθοδολογία που αναλύθηκε στο κεφάλαιο 4, με 2) η προσέγγιση που βασίζεται στη συνθετική συνάρτηση και 3) η προσέγγιση με χρήση μόνο της συνολικής βαθμολογίας. Με έντονα γράμματα φαίνονται τα καλύτερα αποτελέσματα από τις 3 προσεγγίσεις

για κάθε μέτρο. Όπως φαίνεται στο πίνακα 5.5-1 η μεθοδολογία του κεφαλαίου 4 πετυχαίνει καλύτερα αποτελέσματα υπό την έννοια της ακρίβειας των εξαγόμενων συστάσεων σε σχέση με τις άλλες 2 μεθοδολογίες. Χειρότερα αποτελέσματα έχει η 3η μέθοδος όπου στους υπολογισμούς λαμβάνεται υπόψη μόνο η συνολική βαθμολογία. Τέλος παρατηρούμε ότι με τη χρήση 19 συστάδων και οι 3 προσεγγίσεις επιφέρουν καλύτερα αποτελέσματα άλλα με μικρή διαφορά.

5.6 Συμπεράσματα

Στο κεφάλαιο αυτό εξετάσθηκαν τα αποτελέσματα από την εφαρμογή της προτεινόμενης μεθοδολογίας του κεφαλαίου 4 για την επίλυση του προβλήματος συστάσεων. Το σύστημα εφαρμόσθηκε σε πραγματικά δεδομένα αξιολογήσεων ξενοδοχείων από την ιστοσελίδα trip advisor και όλες οι επιμέρους διαδικασίες και αλγόριθμοι υλοποιήθηκαν σε γλώσσα προγραμματισμού Python. Αρχικά έγινε προεπεξεργασία των δεδομένων ώστε να εφαρμοσθεί το σύστημα. Στη συνέχεια παρουσιάστηκαν τα διάφορα στάδια για τη δημιουργία διαφορετικών ομάδων χρηστών με βάση τα προφίλ που προέκυψαν από τη μοντελοποίησή τους με βάση τη πολυκριτήρια μέθοδο UTASTAR και αναλύθηκαν σημαντικά ζητήματα όπως η επιλογή του αριθμού των συστάδων για τη δημιουργία ομάδων χρηστών. Τέλος έγινε σύγκριση με βάση τα αποτελέσματα, με άλλες 2 μεθοδολογίες επίλυσης του προβλήματος συστάσεων με πολυκριτήριες αξιολογήσεις. Ειδικότερα, με βάση τα μέτρα ακρίβειας των εξαγόμενων συστάσεων συμπεραίνουμε ότι η ενσωμάτωση των πολυκριτήριων βαθμολογιών επιφέρει καλύτερα αποτελέσματα. Επίσης μέσω της εφαρμογής των 3 προσεγγίσεων σε συσταδοποιημένα και μη δεδομένα, παρατηρούμε ότι η δημιουργία διαφορετικών ομάδων χρηστών με βάση τα προφίλ που προκύπτουν από τη πολυκριτήρια ανάλυση συνεισφέρει στη μεγαλύτερη ακρίβεια των συστάσεων.

Η μικρή διαφοροποίηση στην ακρίβεια μεταξύ της εφαρμογής σε συσταδοποιημένα και μη δεδομένα, οφείλεται στο πρόβλημα των νέων αντικειμένων(cold start). Ο αλγόριθμος συστάσεων που εφαρμόζεται για την εκτίμηση των βαθμολογιών R(u,i), "απαιτεί" κάθε αντικείμενο να έχει αξιολογηθεί από τουλάχιστον ένα χρήστη της συστάδας που ανήκει ο u, για να είναι δυνατός ο υπολογισμός ομοιότητας μεταξύ χρηστών. Στο σετ δεδομένων όμως υπάρχουν αρκετά ξενοδοχεία τα οποία έχουν αξιολογηθεί από ένα μόνο χρήστη, οπότε δεν υπάρχουν για αυτά αρκετές αξιολογήσεις στη βάση δεδομένων (train set). Η συσταδοποίηση λοιπόν δεν βελτιώνει πάρα πολύ τα αποτελέσματα γιατί είναι δύσκολο να βρεθεί ξενοδοχείο που να έχουν βαθμολογήσει δύο χρήστες της ίδιας συστάδας. Η δημιουργία βέβαια διαφορετικών ομάδων χρηστών μειώνει σημαντικά τον υπολογιστικό χρόνο, καθώς τα μέτρα ομοιότητας μεταξύ χρηστών υπολογίζονται μόνο για χρήστες ίδιας συστάδας, σε αντίθεση με τις μεθόδους που δεν χρησιμοποιούν συσταδοποίηση , όπου εκεί γίνονται υπολογισμοί για όλους τους χρήστες των δεδομένων. Τέλος η μοντελοποίηση των χρηστών με βάση τη πολυκριτήρια ανάλυση και η δημιουργία προφίλ χρηστών, αντιμετωπίζει το πρόβλημα νέων χρηστών. Σε αντίθεση με άλλες μεθοδολογίες επίλυσης όπου απαιτούν για κάθε χρήστη πολλές αξιολογήσεις , η προτεινόμενη μεθοδολογία χρησιμοποιεί κατ' ελάχιστο 2 αξιολογήσεις από κάθε χρήστη, για να υπολογίσει για αυτόν βάρη των κριτηρίων και να τον εντάξει σε μία συστάδα.

Κεφάλαιο 6 : Επίλογος

6.1 Συμπεράσματα

Σε αυτή την εργασία αναπτύχθηκε ένα σύστημα σύστασης ξενοδοχείων για ηλεκτρονικές κρατήσεις με χρήση τεχνικών της πολυκριτήριας ανάλυσης αποφάσεων και του τομέα των συστημάτων συστάσεων. Συγκεκριμένα μελετήθηκε η χρήση της αναλυτικής συνθετικής προσέγγισης στη μοντελοποίηση των χρηστών με βάση τις προτιμήσεις τους και η δημιουργία προφίλ χρηστών. Στη συγκεκριμένη μεθοδολογία περιορίζονται γνωστά προβλήματα των συστημάτων συστάσεων, όπως αυτό των νέων χρηστών, καθώς οι χρήστες μοντελοποιούνται με βάση τη πολυκριτήρια ανάλυση και ομαδοποιούνται πριν την εφαρμογή του αλγόριθμου συστάσεων συνεργατικού φιλτραρίσματος, προσδίδοντας έτσι σε κάθε νέο χρήστη τις ιδιότητες της ομάδας του και μειώνοντας τον υπολογιστικό χρόνο . Επίσης μέσω του αλγορίθμου συστάσεων που αναπτύχθηκε, αξιοποιείται με αποδοτικό τρόπο η πληροφορία σχετικά με τις πολυκριτήριες βαθμολογίες, ενώ αυξάνεται και η ακρίβεια των εξαγόμενων συστάσεων. Συμπερασματικά η πολυκριτήρια ανάλυση αποφάσεων και τα συστήματα συστάσεων, δυο διαφορετικοί τομείς έρευνας, μπορούν να συνδυαστούν για την αποδοτικότερη επίλυση προβλημάτων απόφασης. Η εφαρμογή της προτεινόμενης μεθοδολογίας επιλύει το πρόβλημα επιλογής ξενοδοχείου για ηλεκτρονική κράτηση, σημαντικό πρόβλημα που αντιμετωπίζουν πλέον αρκετοί χρήστες του διαδικτύου. Ο τουριστικός τομέας περιλαμβάνει μεγάλο όγκο πληροφοριών που υπάρχει στο διαδίκτυο και δεν χρησιμοποιείται στη μέγιστη δυναμική του. Τα συστήματα συστάσεων έχουν μεγάλη δυναμική στο να αυξήσουν την ικανοποίηση των τουριστών και να βοηθήσουν στις διαδικασίες λήψης των αποφάσεών τους.

6.2 Μελλοντικές προεκτάσεις

Στην εργασία αυτή εξετάσθηκε ένας τρόπος επίλυσης του προβλήματος συστάσεων. Λόγω του ότι τα συστήματα συστάσεων αποτελούν ένα νέο σχετικά τομέα έρευνας, υπάρχουν αρκετές προεκτάσεις και κατευθύνσεις ως προς τι οποίες μπορεί να γίνει έρευνα για τη βελτίωση των υπάρχων προσεγγίσεων. Ορισμένες από αυτές τις κατευθύνσεις αναφέρονται παρακάτω:

- Στη περίπτωση των πολυκριτήριων συστημάτων συστάσεων μπορούν να εξετασθούν νέοι τρόποι αξιολόγησης και ενσωμάτωσης της πληροφορίας στους αλγορίθμους συστάσεων. Πέρα από τη χρήση των μέτρων ομοιότητας που χρησιμοποιούνται σε αλγόριθμους συνεργατικού φιλτραρίσματος, μπορούν να αναπτυχθούν αλγόριθμοι συστάσεων με χρήση τεχνικών από τη τεχνητή νοημοσύνη, τη μηχανική μάθηση και την εξόρυξη γνώσης από δεδομένα για το προσδιορισμό μιας τελικής βαθμολογίας.
- Επιπλέον η χρήση πολυκριτήριων αξιολογήσεων σε μορφή αριθμητικών δεδομένων σε συνδυασμό με τη χρήση πληροφοριών για το περιεχόμενο των αντικειμένων μπορεί να

οδηγήσει στη δημιουργία υβριδικών συστημάτων με αυξημένη ακρίβεια εξαγόμενων συστάσεων. Ιδιαίτερα στη περίπτωση του τουριστικού τομέα η ανάλυση του περιεχομένου των αντικειμένων (ξενοδοχείων, προορισμών κ.τ.λ.) μπορεί να οδηγήσει στη καλύτερη κατανόηση των αναγκών των χρηστών.

- Για τη καλύτερη κατανόηση της νοοτροπίας των χρηστών σχετικά με τις αξιολογήσεις που δίνουν, μπορούν να ερευνηθούν νέες τεχνικές για τη μοντελοποίηση και αναπαράσταση των διαφορετικών ομάδων χρηστών (π.χ. ιεραρχική συσταδοποίηση , συσταδοποίηση με χρήση τεχνικών μείωσης διαστάσεων).
- Εφαρμογή της προτεινόμενης μεθοδολογίας σε διαφορετικούς τομείς, αλλά και σε διαφορετικά σετ δεδομένων. Η μεθοδολογία του κεφαλαίου 4 εφαρμόσθηκε σε ένα σετ δεδομένων με αξιολογήσεις ξενοδοχείων με βαθμολογίες από το 1 έως το 5. Η υλοποίηση σε διαφορετικές εφαρμογές αλλά και σε διαφορετικά δεδομένα (π.χ. βαθμολογίες με διαφορετική κλίμακα,διαφορετικά είδη κριτηρίων) μπορεί να οδηγήσει σε νέους τρόπους βελτίωσης της μεθοδολογίας και αντιμετώπισης περιορισμών της.

Βιβλιογραφία

1. An integrated Recommender System based on Multi-Criteria Decision Analysis and Data Analysis methods: Methodology, implementation and evaluation-Kleanthi Lakiotaki – 2010

2. New Recommendation Techniques for Multi-Criteria Rating Systems-G. Adomavicius, YoungOk Kwon – 2007

3. Multi-Criteria User Modelling in Recommender Systems-K.Lakiotaki, N.Matsatinis & A. Tsoukias-2011

4. Multi-Criteria Service Recommendation Based on User Criteria Preferences-Liwei Liu, Nikolay Mehandjiev & Dong-Ling Xu-2011

5. Item-based collaborative filtering recommendation algorithms-Sarwar B., Karypis G., Konstan J. & Riedl J-2001

6. Learning Collaborative Information Filters- Billsus D., Pazzani M.J-1998

7. Incremental singular value decomposition algorithms for highly scaleable recommender systems-Sarwar B., Karypis G., Konstan J.A, Riedl J.-2002

8. Contextual Recommendation-B. Mobasher and S. S. Anand-2007

9. Hybrid Systems for Personalized Recommendations. Intelligent Techniques for Web Personalization-R. Burke-2005

10. Evaluating collaborative filtering recommender systems-J. L. Herlocker, J. A. Konstan, L.G. Terveen & J. T. Riedl-2004

11. Methodologie Multicritère d'Aide a la Decision-B. Roy-1985

12. Assessing a set of additive utility functions for multicriteria decision-making, the UTA method-E. Jacquet-Lagreze & J. Siskos -1982

13. Utastar-an ordinal regression method for building additive value functions-Y. Siskos and D. Yannacopoulos -1985

14. Siskos, Y., E. Grigoroudis, N.F. Matsatsinis (2016), UTA Methods, in: S. Greco, M. Ehrgott, J.R. Figueira (Eds.), Multiple Criteria Decision Analysis: State of the Art Surveys, Springer

15. Research Report on User Modeling for Accessibility-Yehya Mohamad,Christos Kouroupetroglou-2014

16. Recommender Systems in Commercial Use- Susan E. Aldrich- 2011

17. Multicriteria Recommender System for Life Insurance Plans based on Utility Theory-2017

18. Multicriteria Predictors using Aggregation Functions based on Item Views-Fabián P. Lousame, Eduardo Sánchez-2010

19. Data Mining Methods for Recommender Systems, Chapter 2-Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol-2010

20. Neighborhood-Based Collaborative Filtering, Chapter 2- Charu C. Aggarwal-2016

21. Introduction to Recommender Systems Handbook-Francesco Ricci, Lior Rokach and Bracha Shapira-2011

22. New Hybrid Recommender Approaches : An Application to Equity Funds Selection-Nikolaos F. Matsatsinis, Eleftherios A. Manarolis-2009

23. Challenges in Recommender Systems for Tourism-Manoj Reddy Dareddy-2016

24. Hotel recommendation based on user preference analysis- Keqiang Wang, Xiaoling Wang, Cheqing Jin, Aoying Zhou-2015

25. Multi-Criteria Recommender Systems- Gediminas Adomavicius, Nikos Manouselis, YoungOk Kwon-2010

26. http://times.cs.uiuc.edu/~wang296/Data/, TripAdvisor Data Set- University of Illinois, Database and Information Systems Laboratory

27. ΜΟΝΤΕΛΑ ΑΠΟΦΑΣΕΩΝ, Μεθοδολογία Επιχειρησιακής Έρευνας, Θεωρία Πολυκριτήριας Ανάλυσης, Εφαρμογές σε Επιχειρήσεις & Οργανισμούς, Εκδόσεις Νέων Τεχνολογιών-Σίσκος Ι. -2008