ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Προηγμένα Πληροφοριακά Συστήματα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ομαδικές συστάσεις βάσει περίπτωσης για διαμορφώσιμα προϊόντα με χρήση πολυδιάστατης ομαδοποίησης

Χαρίκλεια Α. Παπαδημητρίου

***Επιβλέποντες*:** **Ευστάθιος Χατζηευθυμιάδης,** Αναπληρωτής Καθηγητής ΕΚΠΑ
Κωνσταντίνος Κολομβάτσος, Διδάκτωρ ΕΚΠΑ

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ομαδικές συστάσεις βάσει περίπτωσης για διαμορφώσιμα προϊόντα με χρήση πολυδιάστατης ομαδοποίησης

Χαρίκλεια Α. Παπαδημητρίου

Α.Μ.: Μ1218

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ευστάθιος Χατζηευθυμιάδης,** Αναπληρωτής Καθηγητής
Κωνσταντίνος Κολομβάτσος, Διδάκτωρ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ **Όνομα Επώνυμο,** Τίτλος (π.χ Αναπληρωτής Καθηγητής) (εάν υπάρχει)**:**

Νοέμβριος 2016

ΠΕΡΙΛΗΨΗ

Τα συστήματα συστάσεων παρέχουν εξατομικευμένες προτάσεις στους χρήστες σχετικά με αντικείμενα ή θέματα που εκτιμάται ότι θα τους ενδιαφέρουν. Τα μοντέρνα συστήματα συστάσεων υποστηρίζουν τους χρήστες στην επιλογή αντικειμένων ενός συγκεκριμένου είδους (για παράδειγμα, ταινίες, βιβλία και τραγούδια). Η παρούσα εργασία επικεντρώνεται σε ένα σχετικά νέο τομέα συστημάτων συστάσεων που αφορά τα διαμορφώσιμα προϊόντα (configurable products) τα οποία αποτελούνται από επιμέρους αντικείμενα και προτείνονται στο χρήστη ως σύνολο, όπως είναι για παράδειγμα ένας Η/Υ. Συνήθως, τα συστήματα συστάσεων επωφελούνται των τεχνικών του συνεργατικού φιλτραρίσματος (ΣΦ) που προβλέπουν αντικείμενα για το νέο χρήστη με βάση τις προτιμήσεις άλλων όμοιων χρηστών. Εκτός από το συνεργατικό φιλτράρισμα, τα συστήματα συστάσεων χρησιμοποιούν επίσης άλλες τεχνικές μηχανικής μάθησης όπως ομαδοποίηση (clustering) και κατηγοριοποίηση (classification) δεδομένων. Η παρούσα διπλωματική εργασία στοχεύει στην πρόταση μιας νέας αποδοτικής τεχνικής συστάσεων ανασχηματιζόμενων προϊόντων για ομάδες χρηστών. Προτείνεται η δημιουργία ενός υβριδικού συστήματος συστάσεων ΣΦ και συστάσεων βάσει περίπτωσης (case-based) το οποίο θα προτείνει διαμορφώσιμα προϊόντα σε ομάδες χρηστών μέσω της υιοθέτησης τεχνικών πολυδιάστατης ομαδοποίησης και κατηγοριοποίησης. Ειδικότερα, χρησιμοποιούμε τα δημογραφικά δεδομένα και τις προτιμήσεις των χρηστών για να τους ομαδοποιήσουμε σε πολλαπλές κατηγορίες και στη συνέχεια δημιουργούμε ένα μοντέλο που εντάσσει το νέο χρήστη σε μία από αυτές. Οι νέοι χρήστες ομαδοποιούνται βάσει κατηγορίας και οι συστάσεις παρέχονται στην ομάδα βάσει των διαμορφώσεων εγγεγραμμένων χρηστών που οι προτιμήσεις τους μοιάζουν περισσότερο με της εκάστοτε ομάδας. Η πειραματική αξιολόγηση αποδεικνύει ότι η ενσωμάτωση της πολυδιάστατης ομαδοποίησης βελτιώνει την ακρίβεια των συστάσεων. Παράλληλα, αντιμετωπίζει τα κυριότερα προβλήματα των τεχνικών ΣΦ που είναι η αραιότητα των αξιολογήσεων και το πρόβλημα της ψυχρής εκκίνησης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Συστήματα Συστάσεων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ομαδοποίηση υποχώρων, κατηγοριοποίηση πολλαπλών κλάσεων, συνεργατικό φιλτράρισμα, συστάσεις βάσει περίπτωσης, ομαδικές συστάσεις για ανασχηματιζόμενα προϊόντα

ABSTRACT

Recommender systems provide personalized suggestions to end users regarding items or concepts that they will probably find interesting. Modern recommenders help users to select items of a specific kind, for instance films, books or songs. This thesis focuses on a relatively new field of recommender systems concerning configurable products which consist of individual attributes or parts. These parts are recommended to the user as a whole, for example a customizable PC. Usually, recommenders leverage collaborative filtering methods that predict items for new users based on the preferences of other similar users. Apart from collaborative filtering, recommenders are likely to use other techniques common in data mining such as clustering and classification of data. The aim of this thesis is to propose an effective approach for recommendation of configurable products for groups of users. We suggest and describe the creation of a hybrid collaborative filtering and case-based recommender system, which will propose configurations to groups by applying multidimensional clustering and classification algorithms. Specifically, we use demographic data and users' preferences to cluster them in multiple classes and then we create the model which classifies the new user into one of these classes. New users are grouped by class and recommendations are provided to each group based on the configurations of registered users whose preferences are more similar to the group's aggregated preferences. Experimental evaluation of the aforementioned system proves that the integration of multidimensional clustering improves the precision of the recommendations. At the same time, it deals with the major problems of collaborative filtering approaches, which are the sparseness of rankings (for new items) and the cold start problem (for new users).

SUBJECT AREA: Recommender Systems

KEYWORDS: Subspace clustering, multiclass classification, collaborative filtering, case-based recommendation, group recommendations for configurable products

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της παρούσας Διπλωματικής Εργασίας, θα ήθελα να ευχαριστήσω θερμά τους επιβλέποντες, καθ. Ευστάθιο Χατζηευθυμιάδη και διδάκτορα Κωνσταντίνο Κολομβάτσο, για τη συνεργασία, την εμπιστοσύνη και την πολύτιμη συμβολή τους στην ολοκλήρωση της.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, κατεύθυνση «Προηγμένα Πληροφοριακά Συστήματα». Ο στόχος της παρούσας διπλωματικής εργασίας ήταν η εμβάθυνση στον τομέα των συστημάτων συστάσεων και η μελέτη της εξέλιξης τους με σκοπό τη διερεύνηση κάποιου από τα ανοικτά θέματα έρευνας στο συγκεκριμένο τομέα. Από την αρχική συζήτηση του θέματος προσανατολιστήκαμε στον σχετικά νέο τομέα των συστάσεων για ανασχηματιζόμενα προϊόντα καθώς και στις συστάσεις για ομάδες χρηστών. Η πρακτική εφαρμογή και χρησιμότητα των συστάσεων σε καθημερινές εφαρμογές όπως π.χ. ταξιδιωτικά πακέτα για ομάδες χρηστών, με ώθησε στο να ασχοληθώ με το συγκεκριμένο θέμα.

1. ΕΙΣΑΓΩΓΗ

Τα συστήματα συστάσεων (recommender systems) αποτελούν εργαλεία λογισμικού, σκοπός των οποίων είναι η προσφορά φιλτραρισμένης και εξατομικευμένης πληροφορίας μέσα από ένα χώρο πληροφορίας (information space) χαρακτηριστικό του οποίου είναι το μεγάλο μέγεθος ή/και η πολυπλοκότητα. Η παρεχόμενη πληροφορία αφορά προϊόντα ή θέματα που ενδιαφέρουν το χρήστη και παρουσιάζεται δίνοντας προτεραιότητα στα συγκεκριμένα αντικείμενα που (είναι πιθανό να) τον ενδιαφέρουν περισσότερο. Το πεδίο των συστημάτων συστάσεων, που εμφανίστηκε για πρώτη φορά το 1995, γνώρισε ευρεία ανάπτυξη μέχρι σήμερα λόγω της πληθώρας προβλημάτων που επιλύει και των διαφορετικών τεχνικών και τεχνολογιών που χρησιμοποιεί καθώς και των διαφόρων πρακτικών εφαρμογών του. Τα συστήματα συστάσεων θεωρούνται πλέον ένα σημαντικό εργαλείο στο ηλεκτρονικό εμπόριο (e-commerce) καθώς έχει αποδειχθεί ότι η ενσωμάτωση τους σε ιστοσελίδες αυξάνει τις πωλήσεις.

Το πρωτεύον χαρακτηριστικό ενός συστήματος συστάσεων είναι τα δεδομένα, τα οποία προέρχονται από πληθώρα μέσων όπως αξιολογήσεις χρηστών, ανατροφοδότηση (feedback) ή κριτικές αγοραστών. Μετά τη συλλογή δεδομένων, τα συστήματα συστάσεων χρησιμοποιούν αλγορίθμους μηχανικής μάθησης (machine learning), για να εντοπίσουν ομοιότητες και σχέσεις μεταξύ χρηστών και προϊόντων.

Η έρευνα στο χώρο των συστημάτων συστάσεων χρησιμοποιεί αρκετές τεχνικές από το επιστημονικό πεδίο της Τεχνητής Νοημοσύνης, όπως η μηχανική μάθηση, η εξόρυξη δεδομένων (data mining) και η ικανοποίηση περιορισμών (constraint satisfaction). Συνήθως, τα συστήματα συστάσεων επωφελούνται των τεχνικών του συνεργατικού φιλτραρίσματος (collaborative filtering) που περιλαμβάνουν μοντέλα πρόβλεψης (predictive models), ευριστική αναζήτηση (heuristic search), συλλογή δεδομένων (data collection) και αλληλεπίδραση με το χρήστη (user interaction). Η αλληλεπίδραση με το χρήστη έχει ως στόχο τη δημιουργία του προφίλ προτιμήσεων του χρήστη που βασίζεται στις αξιολογήσεις προϊόντων. Τα διαφορετικά προφίλ συγκρίνονται το ένα με το άλλο και σε συνδυασμό με συγκεκριμένους αλγορίθμους χρησιμοποιούνται για τον καθορισμό και την πρόβλεψη των αντικειμένων που είναι πιθανότερο να αρέσουν περισσότερο ή να ανταποκρίνονται καλύτερα στις ανάγκες του χρήστη.

Εκτός από το συνεργατικό φιλτράρισμα, τα συστήματα συστάσεων χρησιμοποιούν επίσης άλλες τεχνικές μηχανικής μάθησης όπως ομαδοποίηση (clustering) και κατηγοριοποίηση (classification) δεδομένων. Η ομαδοποίηση είναι μια τεχνική που χρησιμοποιείται για να συνδυάσει μεγάλες ποσότητες δεδομένων σε παρόμοιες κατηγορίες. Επίσης, χρησιμοποιείται για να εντοπίσει πρότυπα δεδομένων (data patterns) και να καταστήσει απλούστερη τη διαχείριση τεράστιου όγκου δεδομένων. Η κατηγοριοποίηση χρησιμοποιείται για να αποφανθεί αν νέα δεδομένα ή κάποιος όρος αναζήτησης ταιριάζει με ένα μοτίβο που έχει παρατηρηθεί σε προηγούμενο χρόνο.

Η πολυδιάστατη ομαδοποίηση βελτιώνει την απόδοση και την ποικιλομορφία των συστάσεων και παράλληλα αντιμετωπίζει τα κυριότερα προβλήματα των τεχνικών

συνεργατικού φιλτραρίσματος (ΣΦ) που είναι η αραιότητα των αξιολογήσεων και το πρόβλημα της ψυχρής εκκίνησης που παρουσιάζεται όταν ένας νέος χρήστης γραφτεί στο σύστημα και δεν υπάρχει καμία αποθηκευμένη πληροφορία για αυτόν. Τα μοντέρνα συστήματα συστάσεων υποστηρίζουν τους χρήστες στην επιλογή αντικειμένων ενός συγκεκριμένου είδους (για παράδειγμα, ταινίες, βιβλία και τραγούδια). Η παρούσα εργασία επικεντρώνεται σε ένα νεότερο τομέα συστημάτων συστάσεων που αφορά τα διαμορφώσιμα προϊόντα (configurable products) που αποτελούνται από επιμέρους αντικείμενα και προτείνονται στο χρήστη ως σύνολο, όπως είναι για παράδειγμα ένας Η/Υ ή ένα ταξιδιωτικό πακέτο. Ανοικτό θέμα έρευνας στο πεδίο των συστάσεων για ανασχηματιζόμενα προϊόντα αποτελεί η βασισμένη σε ομάδες διαμόρφωση (group-based configuration), δηλαδή η εύρεση αποδοτικών τεχνικών συστάσεων ανασχηματιζόμενων προϊόντων για ομάδες χρηστών. Για αυτό το λόγο η τεχνική που προτείνουμε απευθύνεται σε ομάδες χρηστών αντί για μεμονωμένους χρήστες.

Στην προτεινόμενη τεχνική ενσωματώνουμε τις τεχνικές πολυδιάστατης ομαδοποίησης και κατηγοριοποίησης σε ένα υβριδικό σύστημα συστάσεων ΣΦ και συστάσεων βάσει περίπτωσης (case-based). Συγκεκριμένα χρησιμοποιούμε τα δημογραφικά δεδομένα των χρηστών καθώς και τις προτιμήσεις τους για να ομαδοποιήσουμε τους χρήστες σε πολλαπλές κατηγορίες και στη συνέχεια δημιουργούμε ένα μοντέλο που εντάσσει το νέο χρήστη σε μία από τις κατηγορίες. Οι νέοι χρήστες ομαδοποιούνται και οι συστάσεις παρέχονται στην ομάδα με βάση τις ολοκληρωμένες διαμορφώσεις (configurations) εγγεγραμμένων χρηστών που οι προτιμήσεις τους και το προφίλ τους μοιάζει περισσότερο με την εκάστοτε ομάδα. Η χρήση παλαιότερων συνόδων με διαμορφώσεις που έχουν επιλέξει οι εγγεγραμμένοι χρήστες είναι στοιχείο που παραπέμπει σε συστάσεις βάσει περίπτωσης, οι οποίες είναι μια μορφή συστάσεων βάσει γνώσης που τυπικά λαμβάνει υπόψη ιδιότητες των αντικειμένων από προηγούμενες επιλογές των χρηστών και χρησιμοποιεί συναρτήσεις ομοιότητας. Η πρωτοπορία της παρούσας εργασίας έγκειται στο γεγονός ότι εστιάζει στην παροχή συστάσεων για ομάδες χρηστών. Επιπροσθέτως, οι ομάδες αυτές δεν είναι γνωστές εκ των προτέρων αλλά προκύπτουν από τη διαδικασία της ομαδοποίησης (clustering).

Η δομή της παρούσας εργασίας είναι η ακόλουθη: Στο δεύτερο κεφάλαιο αναλύουμε συνοπτικά τις διάφορες τεχνικές συστάσεων καθώς και τα προβλήματα που αντιμετωπίζουν. Στο τρίτο κεφάλαιο αναλύουμε ειδικά το πρόβλημα των συστάσεων για ανασχηματιζόμενα προϊόντα και παρουσιάζουμε τη σχετική ορολογία. Στη συνέχεια του κεφαλαίου αναφέρουμε εκτενέστερα τους αλγορίθμους ομαδοποίησης για προβολές του υποχώρου (subspace clustering) και τις διάφορες προσεγγίσεις τους μια εκ των οποίων εφαρμόσαμε στην προτεινόμενη τεχνική. Στο τέταρτο κεφάλαιο, που αποτελεί το κυριότερο μέρος της παρούσας εργασίας, παρουσιάζουμε αναλυτικά την προτεινόμενη τεχνική, τη δομή του συστήματος καθώς και τους εμπλεκόμενους αλγόριθμους (ομαδοποίηση, κατηγοριοποίηση, συνεργατικές συστάσεις κ.λπ.). Στο πέμπτο κεφάλαιο παρουσιάζουμε την πειραματική αξιολόγηση του συστήματος βάσει συγκεκριμένων μετρικών και σεναρίων για να αξιολογήσουμε την αποδοτικότητα της τεχνικής. Τέλος, στο έκτο κεφάλαιο

παραθέτουμε τα σημαντικότερα συμπεράσματα καθώς και προτάσεις για μελλοντικές επεκτάσεις της προτεινόμενης τεχνικής.

2. ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ

Ο πρώτος ορισμός που δόθηκε για τα συστήματα συστάσεων (ΣΣ) στο άρθρο των Resnick και Varian το 1997 [1], εστιάζει στην υποστήριξη που παρέχουν τα συστήματα αυτά στη συνεργασία μεταξύ χρηστών. Ο ρόλος τους περιορίζεται στη συγκέντρωση συστάσεων (δοσμένων ως είσοδοι στο σύστημα) και την προώθηση τους στους κατάλληλους χρήστες. Με την πάροδο του χρόνου ο ορισμός επεκτείνεται ώστε να περιλαμβάνει «κάθε σύστημα που παράγει εξατομικευμένες συστάσεις ως έξοδο ή καθοδηγεί το χρήστη με εξατομικευμένο τρόπο σε χρήσιμα ή ενδιαφέροντα αντικείμενα μέσα από ένα μεγάλο χώρο δυνατών επιλογών» [2]. Ένας πιο τυπικός ορισμός δόθηκε το 2005 από τους Adomavicius και Tuzhilin:

"[…] το πρόβλημα συστάσεων (recommendation problem) μπορεί να διατυπωθεί ως εξής: έστω *C* το σύνολο όλων των χρηστών και S το σύνολο όλων των πιθανών αντικειμένων που μπορούν να προταθούν. Έστω *u* μια συνάρτηση χρησιμότητας (utility function) που μετράει τη χρησιμότητα του αντικειμένου *s* για τον χρήστη *c*, η οποία είναι: u: $C \times S \Rightarrow R$, όπου το *R* είναι πλήρως διατεταγμένο σύνολο (π.χ. οι πραγματικοί αριθμοί σε ένα συγκεκριμένο διάστημα). Ακολούθως, για κάθε χρήστη $c \in C$ πρέπει να επιλεγεί αντικείμενο $s' \in S$ τέτοιο ώστε να μεγιστοποιεί τη χρησιμότητα για το συγκεκριμένο *c*" [3].

Ο παραπάνω ορισμός φυσικά δεν είναι τόσο γενικός ώστε να περιγράφει επίσης και τα ΣΣ που λειτουργούν με σύνολα διαμορφώσεων (configurations) και όχι σύνολα αντικειμένων ή για τα συστήματα που παρέχουν συστάσεις σε ομάδες χρηστών αντί για ένα μεμονωμένο χρήστη. Από τους ορισμούς που παρατέθηκαν είναι εμφανές ότι τα ΣΣ διαφέρουν από τις μηχανές αναζήτησης και τις σχετικές εφαρμογές ανάκτησης πληροφορίας καθώς σε αυτά το σύνολο αποτελεσμάτων που ανακτάται για συγκεκριμένο ερώτημα θα είναι το ίδιο ανεξάρτητα του ποιος το θέτει.

2.1 Τεχνικές συστάσεων

Στο συγκεκριμένο κεφάλαιο αναλύουμε τις κατηγορίες των ΣΣ ανάλογα με την πληροφορία που χρησιμοποιούν ως είσοδο.

2.1.1 Συνεργατικό φιλτράρισμα (Collaborative Filtering)

Η πιο επιφανής τεχνική συστάσεων, είναι οι «συνεργατικές συστάσεις» (collaborative recommendation), κεντρική ιδέα της οποίας είναι ότι αν για δύο χρήστες τα αντικείμενα 1 ως k έχουν την ίδια χρησιμότητα (utility) τότε είναι αρκετά πιθανό να έχουν επίσης την ίδια για το αντικείμενο k+1 [4]. Το κυριότερο πλεονέκτημα της τεχνικής αυτής είναι η απλότητά της καθώς το πρόβλημα του υπολογισμού της χρησιμότητας μετατρέπεται σε πρόβλημα συμπλήρωσης των τιμών που λείπουν σε πίνακες αξιολογήσεων, τους αραιούς πίνακες όπου κάθε χρήστης είναι μια γραμμή, κάθε αντικείμενο στήλη και οι τιμές είναι οι γνωστές αξιολογήσεις. Για την εύρεση γειτόνων ομότιμων χρηστών με παρόμοιες προτιμήσεις έχουν

χρησιμοποιηθεί τεχνικές εύρεσης κοντινότερου γείτονα (nearest-neighbor), παραγοντοποίηση πίνακα και άλλες τεχνικές μείωσης διαστάσεων.

2.1.2 Φιλτράρισμα βασισμένο στο περιεχόμενο (Content – Based Filtering)

Πριν η έρευνα επικεντρωθεί στις συστάσεις με χρήση συνεργατικού φιλτραρίσματος (ΣΦ), το ερευνητικό ενδιαφέρον εστιαζόταν στο συνδυασμό γνώσης για τα αντικείμενα με πληροφορίες για τις προτιμήσεις των χρηστών προκειμένου να προταθούν τα πλέον κατάλληλα αντικείμενα. Εξ' αιτίας της συσχέτισης της προσέγγισης αυτής με το περιεχόμενο της πηγής γνώσης και συγκεκριμένα με τα χαρακτηριστικά των αντικειμένων, η τεχνική αυτή έγινε γνωστή ως «συστάσεις βάσει περιεχομένου» (content-based recommendation). Σε αυτές τις μεθόδους, η χρησιμότητα $u(c,s)$ ενός αντικειμένου s για τον χρήστη c εκτιμάται βάσει των χρησιμοτήτων $u(c,s_i)$ των αντικειμένων $s_i \in S$ που είναι παρόμοια με το αντικείμενο s, όπως τέθηκαν αυτές από τον χρήστη c. Η προσέγγιση σχετίζεται στενά με την ανάκτηση και το φιλτράρισμα γνώσης (information retrieval and filtering) και με την επιβλεπόμενη μηχανική μάθηση καθώς μπορεί να μετασχηματιστεί σε πρόβλημα εκμάθησης ενός συνόλου ταξινομητών όπου οι κατηγορίες είναι: «χρήσιμο για τον χρήστη Χ» και «όχι χρήσιμο για τον ίδιο χρήστη Χ».

2.1.3 Χρήση δημογραφικών δεδομένων (Demographic – Based Filtering)

Επιπλέον των παραδοσιακών χαρακτηριστικών του προφίλ των χρηστών, όπως οι λέξεις-κλειδιά και τα δημογραφικά στοιχεία, πιο προηγμένες τεχνικές ανάπτυξης προφίλ που βασίζονται σε κανόνες εξόρυξης δεδομένων, ακολουθίες και υπογραφές (signatures) που περιγράφουν τα ενδιαφέροντα του χρήστη μπορούν να χρησιμοποιηθούν για τη δημιουργία του προφίλ ενός χρήστη. Επίσης, παρόμοιες προηγμένες τεχνικές δημιουργίας προφίλ μπορούν να χρησιμοποιηθούν για την ανάπτυξη περιεκτικών (comprehensive) προφίλ αντικειμένων. Τέτοιες τεχνικές που βασίζονται στην εξόρυξη δεδομένων έχουν κυρίως χρησιμοποιηθεί στο πλαίσιο ανάλυσης χρήσης του Ιστού, π.χ. για την ανακάλυψη μοτίβων χρήσης του Ιστού, δηλαδή τις αλληλουχίες σελίδων στις οποίες πλοηγήθηκαν οι χρήστες, ώστε τα ΣΣ να παρέχουν καλύτερες προτάσεις στους χρήστες σχετικά με ιστοχώρους που τους ενδιαφέρουν.

Μετά την ανάπτυξη των προφίλ χρηστών και αντικειμένων, μια γενικότερη συνάρτηση εκτίμησης αξιολογήσεων μπορεί να οριστεί με βάση τα προφίλ αυτά και τις προηγουμένως καθορισμένες αξιολογήσεις ως εξής: έστω ότι το προφίλ του χρήστη i ορίζεται ως διάνυσμα p χαρακτηριστικών:

$$\vec{c_i} = (a_{i1}, \dots, a_{ip}) \quad (1)$$

Επίσης, έστω ότι το προφίλ του αντικειμένου j ορίζεται ως διάνυσμα r χαρακτηριστικών:

$$\vec{s_j} = (b_{j1}, \dots, b_{jr}) \quad (2)$$

Επίσης, ορίζουμε το $\vec{c} = (\overrightarrow{c_1}, \dots, \overrightarrow{c_n})$ ως διάνυσμα όλων των προφίλ χρήστη και $\vec{s} = (\overrightarrow{s_1}, \dots, \overrightarrow{s_m})$ ως διάνυσμα όλων των προφίλ αντικειμένων. Τότε η πιο γενική διαδικασία εκτίμησης αξιολόγησης μπορεί να οριστεί:

$$r'_{ij} = \begin{cases} r_{ij}, \alpha\nu\ r_{ij} \neq \emptyset \\ u_{ij}\,(R, \vec{c}, \vec{s}), \alpha\nu\ r_{ij} = \emptyset \end{cases} \quad (3)$$

Η εξίσωση (3) εκτιμά κάθε άγνωστη αξιολόγηση r'_{ij} με βάση τις γνωστές αξιολογήσεις $R = \{ r_{ij} \neq \emptyset \}$, τα προφίλ χρήστη $\vec{c}$ και τα προφίλ αντικειμένων $\vec{s}$. Η συνάρτηση χρησιμότητας u_{ij} μπορεί να εκτιμηθεί με διάφορους τρόπους που περιλαμβάνουν ευριστικές συναρτήσεις, ταξινομητές κοντινότερου γείτονα, δέντρα απόφασης, συναρτήσεις ακτινικής βάσης (radial basis functions) και νευρωνικά δίκτυα. Η παραπάνω εξίσωση παρουσιάζει ένα γενικό μοντέλο που εξαρτάται από ολόκληρο το φάσμα εισόδων που περιλαμβάνει τα χαρακτηριστικά του χρήστη *i* ($\overrightarrow{c_i}$) και, πιθανώς, άλλων χρηστών ($\vec{c}$). Επίσης, περιλαμβάνει τα χαρακτηριστικά του αντικειμένου *j* ($\vec{s}_j$) και άλλων αντικειμένων ($\vec{s}$), προτιμήσεις (αξιολογήσεις) R_i όπως εκφράστηκαν από τον *i* χρήστη καθώς και προτιμήσεις όλων των υπόλοιπων χρηστών $R = \{ r_{ij} \neq \emptyset \}$. Συνεπώς η συνάρτηση u_{ij} προϋποθέτει τόσο συνεργατικές (collaborative) όσο και μεθόδους συστάσεων βασισμένες στο περιεχόμενο (content-based). Ένα ενδιαφέρον ερευνητικό πρόβλημα αποτελεί η επέκταση των βασισμένων σε χαρακτηριστικά προφίλ, όπως ορίστηκαν από τα $\vec{c}$ και $\vec{s}$ ώστε να χρησιμοποιούν τις πιο προηγμένες τεχνικές κατασκευής προφίλ που αναφέρθηκαν νωρίτερα.

2.1.4 Υβριδικές τεχνικές συστάσεων (Hybrid Recommendation Methods)

Οι υβριδικές τεχνικές συστάσεων συνδυάζουν συνεργατικές και βασισμένες στο περιεχόμενο μεθόδους προκειμένου να αποφύγουν τους περιορισμούς και τα προβλήματα που παρουσιάζει κάθε μια από τις παραπάνω τεχνικές και τα οποία αναλύονται σε επόμενη ενότητα. Για το συνδυασμό συνεργατικών προσεγγίσεων με τεχνικές συστάσεων βάσει περιεχομένου υπάρχουν διάφορες τεχνικές που κατηγοριοποιούνται ως εξής:

- Υλοποίηση συνεργατικών και βάσει περιεχομένου τεχνικών ξεχωριστά και συνδυασμός των προβλέψεων τους (παράλληλη υβριδοποίηση),
- Ενσωμάτωση χαρακτηριστικών βασισμένων σε περιεχόμενο τεχνικών σε συνεργατική προσέγγιση,
- Ενσωμάτωση χαρακτηριστικών συνεργατικών τεχνικών σε βασισμένες σε περιεχόμενο μεθόδους,
- Κατασκευή γενικού ενοποιημένου μοντέλου που ενσωματώνει χαρακτηριστικά και των δύο τεχνικών.

Για τα συστήματα παράλληλων υβριδικών τεχνικών χρησιμοποιούνται οι ακόλουθες στρατηγικές [2]:

Σταθμισμένη (Weighted): η στρατηγική αυτή συνδυάζει τις συστάσεις δύο ή περισσότερων συστημάτων υπολογίζοντας σταθμισμένα αθροίσματα των αποτελεσμάτων τους.

Μεταβατική (Switching): με τη στρατηγική αυτή αποφασίζεται ποιο σύστημα συστάσεων πρέπει να χρησιμοποιηθεί σε μια συγκεκριμένη περίπτωση αναλόγως της ποιότητας των παραγόμενων αποτελεσμάτων και του προφίλ χρήστη καθώς και συναφών παραμέτρων (contextual parameters) όπως είναι οι προθέσεις και οι προσδοκίες του χρήστη.

Ανεξάρτητος συνδυασμός (Mixed): τα αποτελέσματα ανεξάρτητων συστημάτων συστάσεων συνδυάζονται στο επίπεδο της διεπαφής χρήστη ώστε να παρουσιαστούν μαζί τα αποτελέσματα διαφορετικών τεχνικών.

Συνδυασμός χαρακτηριστικών (Feature combination): με τη στρατηγική αυτή συνεργατικές τεχνικές συγχωνεύονται με τεχνικές βάσει περιεχομένου (ΒΠ). Οι πληροφορίες ΣΦ χρησιμοποιούνται ως δεδομένα - χαρακτηριστικά και οι τεχνικές ΒΠ συνδυάζονται σε αυτό το επαυξημένο σύνολο δεδομένων για την υλοποίηση του αλγορίθμου συστάσεων. Με τον τρόπο αυτό μειώνεται η ευαισθησία του συστήματος στον αριθμό των χρηστών που έχουν βαθμολογήσει ένα αντικείμενο και επιπλέον το σύστημα διαθέτει πληροφορία για την ομοιότητα των αντικειμένων σε αντίθεση με τα συστήματα ΣΦ.

Μέθοδος καταρράκτης (Cascade): Σε αντίθεση με τις προηγούμενες μεθόδους υβριδισμού, η μέθοδος καταρράκτη περιλαμβάνει μια σταδιακή διαδικασία: μία τεχνική συστάσεων χρησιμοποιείται πρώτα για να παράγει ένα σύνολο υποψηφίων προς σύσταση αντικειμένων και μια δεύτερη τεχνική τελειοποιεί τη σύσταση μειώνοντας τον αριθμό των υποψηφίων.

Επαύξηση χαρακτηριστικών (Feature augmentation): Μια τεχνική που χρησιμοποιείται για την παραγωγή βαθμολογίας ή κατηγοριοποίησης ενός στοιχείου και η παραγόμενη πληροφορία ενσωματώνεται στη συνέχεια στην επεξεργασία της επόμενης τεχνικής συστάσεων.

Μετα-επίπεδο (Meta-level): Το μοντέλο που παράγεται από μία τεχνική χρησιμοποιείται ως είσοδος σε μία άλλη τεχνική. Διαφέρει από τη μέθοδο επαύξησης χαρακτηριστικών που χρησιμοποιεί μόνο τα χαρακτηριστικά αφού η συγκεκριμένη κατηγορία χρησιμοποιεί ολόκληρο το μοντέλο.

Στο **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.** συγκρίνεται όλο το εύρος υβριδικού σχεδιασμού χρησιμοποιώντας τα δεδομένα προφίλ χρηστών του γνωστού συστήματος συστάσεων εστιατορίων Entree. Οι μέθοδοι συστάσεων που αναλύονται περιλαμβάνουν *φιλτράρισμα βάσει περιεχομένου* (ΒΠ), ανάκτηση βάσει γνώσης και δύο συνεργατικούς αλγορίθμους. Για τη μέτρηση της απόδοσης των εξεταζόμενων υβριδικών συνδυασμών υπολογίστηκε η μέση βαθμολόγηση των σωστών συστάσεων (average rank of the correct recommendations – ARC). Τα υβρίδια επαύξησης χαρακτηριστικών παρουσίασαν την καλύτερη απόδοση στην περίπτωση που μια τεχνική συστάσεων ΒΠ συνέβαλε σε συνεργατική τεχνική.

2.1.5 Ομαδοποίηση (Clustering) σε συστάσεις συνεργατικού φιλτραρίσματος

Οι παραδοσιακές τεχνικές ομαδοποίησης (clustering) στοχεύουν στην εύρεση ομάδων αντικειμένων χρησιμοποιώντας όλα τα χαρακτηριστικά τους στον πλήρη χώρο δεδομένων. Κάθε αντικείμενο μπορεί να συμμετέχει σε διάφορες ομάδες που σχηματίζονται σε διαφορετικά υποσύνολα των χαρακτηριστικών σε ένα πολυδιάστατο χώρο δεδομένων. Για παράδειγμα οι χρήστες μπορούν να περιγραφούν από τα πολλαπλά χαρακτηριστικά που προσδιορίζουν το προφίλ τους. Σε ένα ΣΣ για αξιολόγηση ταινιών, για κάθε χρήστη μπορούν να παρατηρηθούν πολλά πιθανά ενδιαφέροντα τα οποία πρέπει να εντοπιστούν ως ομάδες (clusters). Εύκολα γίνεται αντιληπτό ότι οι ομάδες μπορούν να επικαλύπτονται με την έννοια ότι ένας χρήστης μπορεί να ανήκει σε πολλές ομάδες και επιπροσθέτως επειδή κάθε ενδιαφέρον ή συμπεριφορά του χρήστη περιγράφεται από συγκεκριμένα χαρακτηριστικά, οι ομάδες εντοπίζονται στις συγκεκριμένες προβολές στον υποχώρο των δεδομένων.

Ανεξαρτήτως του αλγορίθμου ομαδοποίησης που χρησιμοποιείται, οι προσεγγίσεις ομαδοποίησης σε ολόκληρο το χώρο δεδομένων δεν είναι εξίσου αποδοτικές για πολυδιάστατους χώρους που καλύπτουν πολλαπλά χαρακτηριστικά. Για παράδειγμα, για τη χρήση ομαδοποίησης K-means σε συστάσεις ΣΦ τα πειραματικά αποτελέσματα έδειξαν ότι η τεχνική αυτή δεν είναι δυνατόν να εντοπίσει πολλαπλές ομάδες σε πολυδιάστατα δεδομένα [19]. Η έρευνα στον τομέα της ομαδοποίησης σε χώρους δεδομένων πολλών διαστάσεων (high dimensional) ανέδειξε πληθώρα διαφορετικών μεθόδων ομαδοποίησης οι οποίες συνοψίζονται στα [20] και [21]. Ο κοινός τους στόχος είναι η εύρεση ομάδων σε αυθαίρετες προβολές σε υποχώρους των δεδομένων με τους οποίους σχετίζεται ένα υποσύνολο διαστάσεων. Στο [22] οι συγγραφείς προτείνουν μια υβριδική τεχνική που χρησιμοποιεί πολυδιάστατη ομαδοποίηση σε συστάσεις ΣΦ προκειμένου να βελτιώσουν την απόδοση και την ποικιλομορφία των συστάσεων και ταυτόχρονα να αντιμετωπίσουν τα κυριότερα προβλήματα των συνεργατικών τεχνικών συστάσεων όπως είναι η ψυχρή εκκίνηση και η αραιότητα των αξιολογήσεων. Επιπροσθέτως, στο [23] προτείνεται ένα εκτενές μοντέλο για ομαδικές συστάσεις (group recommendation) που εκμεταλλεύεται συστάσεις για αντικείμενα που έχουν δώσει στο παρελθόν χρήστες παρόμοιοι με αυτούς που ανήκουν στην ίδια ομάδα (group). Βάσει αυτών των ιδεών παρουσιάζουμε σε επόμενο κεφάλαιο έναν υβριδικό αλγόριθμό συστάσεων ανασχηματιζόμενων προϊόντων για ομάδες χρηστών με χρήση πολυδιάστατης ομαδοποίησης και κατηγοριοποίησης των χρηστών.

2.2 Τα προβλήματα των τεχνικών συστάσεων

Τα προβλήματα της τεχνικής του συνεργατικού φιλτραρίσματος είναι γνωστά: προκειμένου να γίνουν συστάσεις για νέα αντικείμενα πρέπει να χρησιμοποιηθεί και κάποια επιπλέον πηγή γνώσης και όταν ο πίνακας αξιολογήσεων είναι πολύ αραιός ή ένας χρήστης έχει κάνει πολύ λίγες αξιολογήσεις, το σύστημα δεν μπορεί να βασιστεί σε αυτές απόλυτα για να κάνει προβλέψεις. Το πρόβλημα των νέων αντικειμένων χωρίς αξιολόγηση και νέων χρηστών για τους οποίους δεν είναι γνωστές οι προτιμήσεις (αφού δεν έχουν αξιολογήσει

επαρκή αριθμό αντικειμένων) είναι γνωστό ως πρόβλημα της ψυχρής εκκίνησης (Cold Start problem) [5].

Εκτός από το «πρόβλημα ψυχρής εκκίνησης», μια άλλη πρόκληση που καλούνται να αντιμετωπίσουν τα συστήματα συνεργατικών συστάσεων είναι να παρέχουν συστάσεις σε ένα χρήστη που δεν είναι μέλος κάποιας συγκεκριμένης ομάδας. Οι χρήστες αυτοί ονομάζονται “gray sheep” και το πρόβλημα παροχής συστάσεων σε νέους ή άγνωστους χρήστες ονομάζεται “gray sheep problem” [16]. Το συγκεκριμένο πρόβλημα προκύπτει κυρίως σε περιπτώσεις Διαδικτυακών εφαρμογών με μεγάλες και ανομοιογενείς βάσεις χρηστών. Μια πιθανή λύση σε αυτό το πρόβλημα είναι η αναζήτηση παρόμοιων χρηστών με βάση παρόμοια δημογραφικά στοιχεία ή χαρακτηριστικά. Προκειμένου να προσδιοριστεί η ομοιότητα, μπορεί να χρησιμοποιηθεί ένας μηχανισμός κατηγοριοποίησης μηχανικής μάθησης όπου οι όμοιοι χρήστες (που ονομάζονται «γείτονες» – neighbors) βρίσκονται είτε με την υλοποίηση διαφορετικών αλγορίθμων ομοιότητας (π.χ. Naïve Bayes classifier) ή ελέγχοντας, ως προς την ομοιότητα τους, διάφορες ιδιότητες των χρηστών.

Τα συστήματα συστάσεων βάσει περιεχομένου έχει παρατηρηθεί ότι υπόκεινται στους παρακάτω περιορισμούς:

- **Περιορισμένη ανάλυση περιεχομένου**: Οι βασισμένες στο περιεχόμενο συστάσεις περιορίζονται από τα χαρακτηριστικά που είναι ρητά συνδεδεμένα με τα αντικείμενα τα οποία προτείνουν αυτά τα συστήματα. Το περιεχόμενο πρέπει να είναι σε μορφή που μπορεί αυτόματα να αναλυθεί από υπολογιστή ή τα χαρακτηριστικά θα πρέπει να ανατεθούν χειροκίνητα στα αντικείμενα προκειμένου να υπάρχουν επαρκή σύνολα χαρακτηριστικών. Ο περιορισμός αυτός έχει το μειονέκτημα ότι σε ορισμένους τύπους δεδομένων, π.χ. δεδομένα πολυμέσων δεν μπορεί να εφαρμοστεί η μέθοδος αυτόματης εξαγωγής χαρακτηριστικών και η ανάθεση γνωρισμάτων από χρήστες συχνά δεν είναι πρακτική λόγω των περιορισμένων ανθρωπίνων πόρων. Επίσης δυο διαφορετικά αντικείμενα με ίδιο σύνολο χαρακτηριστικών δεν μπορούν να αναγνωριστούν ως διακριτά από το σύστημα [18].
- **Πρόβλημα νέου χρήστη**: Ο χρήστης πρέπει να βαθμολογήσει επαρκή αριθμό αντικειμένων πριν το σύστημα συστάσεων να μπορεί πραγματικά να κατανοήσει τις προτιμήσεις του και να παρουσιάσει αξιόπιστες προτάσεις. Ως εκ τούτου, ένας νέος χρήστης, έχοντας πολύ λίγες αξιολογήσεις, δεν θα είναι σε θέση να πάρει ακριβείς συστάσεις.
- **Υπερ-εξειδίκευση (over specialization):** Αυτό το πρόβλημα παρουσιάζεται όταν το σύστημα προτείνει μόνο αντικείμενα που συγκεντρώνουν υψηλή βαθμολογία σύμφωνα με το προφίλ του χρήστη οπότε περιορίζεται στη σύσταση αντικειμένων παρόμοιων με όσα έχουν ήδη βαθμολογηθεί.

Οι περισσότερες μέθοδοι συστάσεων δεν είναι ευέλικτες από την άποψη ότι προσδιορίζουν μόνο ένα προκαθορισμένο και αμετάβλητο σύνολο συστάσεων. Συνεπώς, ο χρήστης δε μπορεί να προσαρμόσει τις συστάσεις ανάλογα με τις εκάστοτε ανάγκες του. Το πρόβλημα αυτό αναγνωρίστηκε στο [17] και η *Γλώσσα Επερωτήσεων Συστάσεων* (Recommendation

Query Language – RQL) προτάθηκε για την αντιμετώπιση του. Η RQL είναι γλώσσα παρόμοια με την SQL που εξυπηρετεί στην έκφραση ευέλικτων επερωτήσεων συστάσεων που ορίζονται από το χρήστη. Επίσης, τα περισσότερα από τα ΣΣ προτείνουν ένα συγκεκριμένο αντικείμενο σε μεμονωμένους χρήστες και δεν ασχολούνται με αθροιστικές προτάσεις, δηλαδή συστάσεις που να αφορούν ένα πλήθος εφαρμογών, για παράδειγμα, να προτείνουν μάρκες ή κατηγορίες αντικειμένων σε συγκεκριμένα σύνολα χρηστών (π.χ. ταξιδιωτικά πακέτα [κατηγορία αντικειμένων] για προπτυχιακούς φοιτητές [σύνολο χρηστών]). Ένας τρόπος υποστήριξης των αθροιστικών (aggregated) συστάσεων είναι η χρήση προσεγγίσεων βασισμένων σε OLAP (Online Analytical Processing) για πολυδιάστατες συστάσεις (multidimensional recommendations).

3. ΣΥΣΤΑΣΕΙΣ ΓΙΑ ΑΝΑΣΧΗΜΑΤΙΖΟΜΕΝΑ ΠΡΟΙΟΝΤΑ

3.1 Το πρόβλημα των συστάσεων για ανασχηματιζόμενα προϊόντα (configurable products)

Όπως αναφέρθηκε σε προηγούμενη ενότητα, τα ΣΣ δεν αφορούν μόνο σε μεμονωμένα αντικείμενα, αλλά είναι δυνατόν να αφορούν και σε διαμορφώσιμα, ανασχηματιζόμενα προϊόντα (configurable products) που αποτελούνται από επιμέρους αντικείμενα και προτείνονται στο χρήστη ως σύνολο, όπως είναι για παράδειγμα ένας Η/Υ, ένα αυτοκίνητο ή ένα ταξιδιωτικό πακέτο. Για τα διαμορφώσιμα προϊόντα κάθε χαρακτηριστικό αναπαριστά μια επιλογή για το χρήστη που εκτελεί την παραγγελία. Κάθε χαρακτηριστικό έχει πολλές διαφορετικές εναλλακτικές τιμές, οπότε η παραγγελία στην πραγματικότητα θα είναι μια λίστα επιλογών που καθορίζουν το τελικό προϊόν. Η βάση γνώσης που ορίζει τα συγκεκριμένα διαμορφώσιμα προϊόντα περιγράφει τις ιδιότητες των επιτρεπόμενων στιγμιότυπων. Αν και η αναπαράσταση γνώσης είναι διαφορετική σε σχέση με αυτή των μη ανασχηματιζόμενων προϊόντων, ο στόχος της εύρεσης της λύσης που ανταποκρίνεται καλύτερα στις ανάγκες του πελάτη είναι και σε αυτή την περίπτωση ο ίδιος.

Κατά τη διαδικασία δημιουργίας μιας διαμόρφωσης (configuration), οι χρήστες προσδιορίζουν και συχνά τροποποιούν τις απαιτήσεις τους και το σύστημα τους παρέχει ανάδραση (feedback) π.χ. λύσεις που ικανοποιούν το σύνολο περιορισμών που έθεσαν ή προτάσεις για χαλάρωση των περιορισμών όταν δεν είναι δυνατό να βρεθεί λύση. Οι απαιτήσεις μπορούν να είναι ορισμοί χαρακτηριστικών τιμών ή ακόμα και επερωτήσεις κειμένου εκφρασμένες χωρίς αυστηρή μαθηματική περιγραφή.

Τα σύγχρονα ΣΣ παρέχουν συστάσεις βασισμένες μόνο σε πληροφορίες σχετικά με τους χρήστες και τα αντικείμενα. Αυτό σημαίνει ότι δεν λαμβάνονται υπόψη επιπλέον πληροφορίες σχετικά με τα συμφραζόμενα που πιθανόν να είναι πολύ σημαντικά. Σε πολλές περιπτώσεις η χρησιμότητα ενός αντικειμένου σε κάποιον χρήστη μπορεί να εξαρτάται σημαντικά από το χρόνο (π.χ. μήνας ή ημέρα της εβδομάδας) καθώς και τα άτομα για τα οποία προορίζεται ή υπό ποιες περιστάσεις. Για τις περιπτώσεις αυτές δεν είναι αρκετό το σύστημα να προτείνει απλώς αντικείμενα στους χρήστες αλλά πρέπει να δέχεται επιπλέον συναφείς πληροφορίες όπως για παράδειγμα ο χρόνος ή η τοποθεσία του χρήστη. Προκειμένου να προταθεί ένα ταξιδιωτικό πακέτο, η διαδικασία συστάσεων θα πρέπει να περιλαμβάνει πληροφορία για το χρόνο του έτους, τα άτομα με τα οποία θα ταξιδέψει ο χρήστης, τυχόν περιορισμούς και τις συνθήκες του ταξιδιού π.χ. είδος καταλύματος ή μέσο μεταφοράς.

Για να ληφθεί υπόψη η συναφής πληροφορία, προτάθηκε ο προσδιορισμός της συνάρτησης χρησιμότητας σε πολυδιάστατο χώρο D_1 x … x $D_n \rightarrow R$, σε αντίθεση με τον παραδοσιακά δισδιάστατο χώρο χρήστη – αντικειμένου (*User x Item* space). Επομένως, η συνάρτηση χρησιμότητας *u* ορίζεται:

$$u: D_1 \times \ldots \times D_n \rightarrow R \qquad (4)$$

Έτσι, ένα πρόβλημα συστάσεων ορίζεται επιλέγοντας συγκεκριμένες διαστάσεις σχετικά με το αντικείμενο D_{i1} , … , D_{ik} (όπου $k < n$), αλλά και σχετικά με το χρήστη D_{j1} , … , D_{jl} (όπου $l < n$), οι οποίες δεν επικαλύπτονται. Έτσι για κάθε πλειάδα (d_{j1}, …, d_{jl}) ∈ D_{j1} x … x D_{jl} προτείνεται η πλειάδα (d_{i1}, …, d_{ik}) ∈ D_{i1} x … x D_{ik} που μεγιστοποιεί τη χρησιμότητα u (d_1, …, d_n).

Ανοικτό θέμα έρευνας στο πεδίο των συστάσεων για ανασχηματιζόμενα προϊόντα αποτελεί η βασισμένη σε ομάδες διαμόρφωση (group – based configuration) και συγκεκριμένα η βέλτιστη υποστήριξη των διαδικασιών διαμόρφωσης για ομάδες χρηστών που αποτελούν το αντικείμενο της παρούσας εργασίας. Ένα τέτοιο παράδειγμα είναι η διαμόρφωση ταξιδιωτικών πακέτων για ομάδες. Σε αυτά τα σενάρια η διαμόρφωση δημιουργείται συνεργατικά από την ενδιαφερόμενη ομάδα και όχι από μεμονωμένους χρήστες. Η τυπική λειτουργικότητα που υποστηρίζεται στις διαμορφώσεις για ένα χρήστη πρέπει να υποστηρίζεται και στα σενάρια κατανεμημένων διαμορφώσεων. Η λειτουργικότητα περιλαμβάνει πρόταση των χαρακτηριστικών και των τιμών για το εκάστοτε χαρακτηριστικό καθώς και πρόταση επεξηγήσεων.

Οι ανοιχτές ερωτήσεις που αποτελούν αντικείμενο έρευνας στο πλαίσιο αυτό είναι:

- Σε περίπτωση ασυνέπειας των απαιτήσεων πως θα επιτευχθεί ομοφωνία μεταξύ των διαφορετικών χρηστών;
- Ποιοι είναι οι καλύτεροι αλγόριθμοι διαμορφώσεων για το σκοπό αυτό;
- Σε ποιο βαθμό θα πρέπει να είναι ορατές οι προτιμήσεις των χρηστών σε άλλους χρήστες;
- Πώς θα μπορέσουμε να αξιοποιήσουμε τις θεωρίες της λήψης αποφάσεων της ομάδας (group decision making) για τη βελτίωση της ποιότητας των διαδικασιών λήψης αποφάσεων;

3.1.1 Βασισμένες σε γνώση διαμορφώσεις (Knowledge - Based Configurations)

Για τους παρακάτω ορισμούς ισχύει η παραδοχή ότι μια εργασία διαμόρφωσης (configuration task) εκφράζεται ως ένα βασικό πρόβλημα ικανοποίησης περιορισμών (Constraint Satisfaction Problem – CSP)

- Configuration task

Μια εργασία διαμόρφωσης μπορεί να οριστεί ως ένα πρόβλημα ικανοποίησης περιορισμών (V,D,C) όπου το σύνολο $V = \{v_1, v_2, \ldots, v_n\}$ αναπαριστά το σύνολο μεταβλητών πεπερασμένου πεδίου (finite domain variables) και $D = \{ \text{dom}(v_1), \text{dom}(v_2), \ldots, \text{dom}(v_n)\}$ είναι το σύνολο των αντίστοιχων πεδίων των μεταβλητών. Το $C = P_{KB} \cup C_R$ αποτελεί το σύνολο των περιορισμών όπου το $P_{KB} = \{c_1, c_2, \ldots, c_m\}$ εκφράζει τη γνώση για το αντικείμενο και το $C_R = \{c_{m+1}, c_{m+2}, \ldots, c_u\}$ εκφράζει ένα σύνολο απαιτήσεων. Η τριπλέτα (V, D, P_{KB}) συμβολίζει τη βάση γνώσης για τη διαμόρφωση (configuration knowledge base).

- Διαμόρφωση (λύση)

Μια διαμόρφωση (αποτέλεσμα διαμόρφωσης ή λύση – configuration result, solution) αποτελείται από μια λίστα των συστατικών και συνιστωσών που είναι μέρη του τελικού διαμορφώσιμου προϊόντος. Συγκεκριμένα, μια διαμόρφωση είναι ένα σύνολο $I = \{v_1 = i_1, v_2 = i_2, \ldots, v_n = i_n\}$ που περιλαμβάνει τις συγκεκριμένες τιμές – στοιχεία i_j από το πεδίο τιμών $dom(v_j)$ που αποδίδονται σε κάθε μεταβλητή v_j. Το ενδιαφέρον των χρηστών επικεντρώνεται στις έγκυρες διαμορφώσεις – λύσεις, δηλαδή διαμορφώσεις που είναι ***πλήρεις*** (σε κάθε μεταβλητή έχει ανατεθεί τιμή) και ***συνεπείς*** (οι αναθέσεις είναι συνεπείς με τους περιορισμούς στο σύνολο *C*.

3.1.2 Πρόταση χαρακτηριστικών (Feature Recommendation)

Οι τεχνολογίες συστάσεων μπορούν να βοηθήσουν το χρήστη κατευθύνοντας τον στο να προσδιορίσει τα επιθυμητά χαρακτηριστικά (δηλαδή τις μεταβλητές) του ανασχηματιζόμενου προϊόντος και με αυτό τον τρόπο μειώνεται ο χρόνος επιλογής χαρακτηριστικών. Η μείωση επιτυγχάνεται με δύο τρόπους: αφενός κάποια χαρακτηριστικά μπορούν να αποκλειστούν ρητά αν δεν είναι απαραίτητα σε συγκεκριμένο πλαίσιο, για παράδειγμα, αν η τιμή κάποιου χαρακτηριστικού μπορεί να προκύψει από τις απαιτήσεις του χρήστη C_R σε συνδυασμό με τη βάση γνώσης για το αντικείμενο P_{KB} (δηλαδή από το σύνολο των περιορισμών μέχρι τη δεδομένη στιγμή). Αφετέρου μπορούν να ταξινομηθούν ανάλογα με τη σημασία τους για το χρήστη, ώστε να προτείνονται πρώτα τα πιο σχετικά με τις ανάγκες και τις επιθυμίες του χαρακτηριστικά.

Εισαγωγή και Εξαγωγή χαρακτηριστικών: Η εξαγωγή των χαρακτηριστικών που δεν είναι σχετικές με το τρέχον πλαίσιο διαμόρφωσης μπορεί να υλοποιηθεί βάσει ροών διαδικασιών (process flows) που καθορίζουν τη σειρά με την οποία οι ερωτήσεις θα τεθούν στο χρήστη [11].

Αυτή η επιλογή σχετικών ερωτήσεων που βασίζεται σε διαδικασίες μπορεί να ερμηνευθεί ως ένας απλός τύπος συστάσεων βασισμένων σε γνώση όπου οι περιορισμοί εκφράζουν τις προϋποθέσεις για την επιλογή ερωτήσεων.

Αξιολόγηση χαρακτηριστικών: Εκτός από την προσέγγιση ρητής εισαγωγής ή εξαγωγής χαρακτηριστικών, τα χαρακτηριστικά μπορούν επίσης να ταξινομηθούν ως προς τη σημασία τους για το χρήστη. Διάφορες προσεγγίσεις για την πρόταση χαρακτηριστικών έχουν ήδη αναπτυχθεί: για παράδειγμα η βασισμένη στη δημοτικότητα (popularity-based), βασισμένη στη χρηστικότητα (utility-based) πρόταση χαρακτηριστικών [14], συνεργατική (collaborative) και τέλος βασισμένη στην εντροπία (entropy-based) πρόταση χαρακτηριστικών [15].

- Επιλογή χαρακτηριστικών βασισμένη στη δημοτικότητα

Στην προσέγγιση αυτή, τα χαρακτηριστικά (v_i) αξιολογούνται ανάλογα με τη δημοτικότητά τους που ορίζεται ως το μερίδιο των επιλογών των χρηστών για την v_i σε σχέση με τον συνολικό αριθμό επιλογών μεταβλητών. Το πλεονέκτημα της προσέγγισης αυτής είναι ότι δεν απαιτεί πολύπλοκους υπολογισμούς και πρωτίστως δίνει σημασία στα χαρακτηριστικά που θέλουν να προσδιορίσουν οι χρήστες.

$$popularity(v_i) = \frac{\#selected(\{v_i\})}{\#selected(\{v_1 \ldots v_n\})} \quad (5)$$

- Επιλογή χαρακτηριστικών βάσει εντροπίας

Ο όρος εντροπία χρησιμοποιείται για τον καθορισμό του μικρότερου πλήθους bits που χρειάζονται για τη μετάδοση πληροφορίας με συγκεκριμένη κατανομή εμφάνισης. Όσο υψηλότερη είναι η εντροπία τόσο μεγαλύτερος και ο βαθμός πληροφορίας στα μεταφερόμενα δεδομένα. Στην επιλογή χαρακτηριστικών η εντροπία μπορεί να χρησιμοποιηθεί για επιλογή χαρακτηριστικών με υψηλή εντροπία (άρα και παρεχόμενη πληροφορία) ώστε να ελαχιστοποιηθεί ο συνολικός αριθμός ερωτήσεων που απαιτούνται και να καθοριστεί επιτυχώς μια διαμόρφωση. Η εντροπία ενός χαρακτηριστικού (της μεταβλητής v_i) καθορίζεται από την εξίσωση (2) όπου p_{aj} είναι η πιθανότητα εμφάνισης της τιμής $a_j \in \text{dom}(v_i)$.

$$entropy(v_i) = -\sum_{j=1}^{|dom(v_i)|} p_{aj} * \log_2(p_{aj}) \quad (6)$$

Για τον υπολογισμό της εντροπίας του χαρακτηριστικού εκμεταλλευόμαστε την πληροφορία που περιλαμβάνεται σε αρχεία καταγραφής επιτυχημένων και ολοκληρωμένων συνόδων διαμόρφωσης. Η τεχνική μέτρησης της εντροπίας δεν υπολογίζει τη σημασία των χαρακτηριστικών για κάθε χρήστη και για αυτό θα πρέπει να εφαρμόζεται μόνο σε περιπτώσεις όπου δεν ενδιαφέρουν οι προτιμήσεις του χρήστη όσον αφορά την προδιαγραφή των χαρακτηριστικών. Ειδικά στο πλαίσιο των διεργασιών διαμόρφωσης, η εντροπία πρέπει να συνδυάζεται με άλλες μεθόδους που λαμβάνουν υπόψη τη συνάφεια των χαρακτηριστικών (όπως η συνεργατική και η βασισμένη στη δημοτικότητα κατάταξη).

- Επιλογή χαρακτηριστικών βασισμένη στη χρησιμότητα

Η τεχνική αυτή συνδυάζει τα πλεονεκτήματα της επιλογής χαρακτηριστικών βάσει δημοτικότητας και βάσει εντροπίας. Η χρησιμότητα για μεταβλητή v_i δίνεται από την ακόλουθη εξίσωση:

$$utility(v_i) = entropy(v_i) * popularity(v_i) \quad (7)$$

3.1.3 Πρόταση επεξηγήσεων (Explanation Recommendation)

Μερικές φορές οι απαιτήσεις C_R που καθορίστηκαν από τους χρήστες είναι ασυνεπείς με τη βάση γνώσης για το αντικείμενο P_{KB}. Οι τεχνολογίες συστάσεων (recommendation technologies) μπορούν να κατευθύνουν το χρήστη στην τροποποίηση των τιμών των χαρακτηριστικών (δηλαδή μεταβλητών) του διαμορφώσιμου προϊόντος σε περίπτωση που δεν μπορεί να βρεθεί λύση με τις τρέχουσες τιμές. Προκειμένου να μην περιοριστεί ο χρήστης αναφορικά με το πόσες και ποιες απαιτήσεις θα μπορεί να θέτει, η τεχνολογία συστάσεων παρέχει τη δυνατότητα να εμφανίζονται στο χρήστη *ελάχιστες επεξηγήσεις* (minimal explanations) και *διαγνώσεις* (diagnoses) [6], [7] ή *μέγιστες χαλαρώσεις τιμών* (maximal relaxations) [12], [13] στις περιπτώσεις που μια λύση δεν είναι δυνατόν να βρεθεί με τους δοσμένους περιορισμούς. Οι ελάχιστες επεξηγήσεις είναι το σύνολο των ελάχιστων

απαιτήσεων που πρέπει να τροποποιηθούν οι τιμές τους ή να μη δοθεί καμία τιμή για αυτές, ώστε να προσδιοριστεί κάποια λύση. Η μέγιστη χαλάρωση τιμών είναι το συμπλήρωμα της ελάχιστης επεξήγησης.

- Διάγνωση βασισμένη σε μοντέλο (Model based diagnosis)

Οι επεξηγήσεις (διαγνώσεις) βασίζονται στην επίλυση των ελαχίστων συνόλων σύγκρουσης (conflict set). Μια ελάχιστη επεξήγηση είναι το ελάχιστο σύνολο περιορισμών που πρέπει να διαγραφούν ή να τροποποιηθούν ώστε το C_R να γίνει συνεπές με το P_{KB}. Επίλυση μιας ελάχιστης σύγκρουσης επιτυγχάνεται με τη διαγραφή τουλάχιστον ενός περιορισμού από το αντίστοιχο σύνολο των συγκρούσεων.

3.1.4 Πρόταση χαρακτηριστικών τιμών (Feature Value Recommendation)

Με την πρόταση χαρακτηριστικών τιμών παρουσιάζονται στο χρήστη δημοφιλείς επιλογές χαρακτηριστικών με βάση παλιότερες επιλογές χρηστών με παρόμοιες απαιτήσεις ώστε να ολοκληρωθούν οι μη πλήρεις διαμορφώσεις του χρήστη. Οι προτάσεις χαρακτηριστικών τιμών βοηθούν το χρήστη να καταλάβει τις εξαρτήσεις μεταξύ των απαιτήσεων που έχει θέσει και τις πιθανές αναθέσεις τιμών για τις υπόλοιπες μεταβλητές. Για παράδειγμα χρησιμοποιώντας ένα σύστημα γνώσης βασισμένο σε κανόνες, το σύστημα συστάσεων υπολογίζει και υποδεικνύει τα χαρακτηριστικά που είναι συμβατά με τις προηγούμενες επιλογές του χρήστη. Όταν μια επιλογή δεν είναι συμβατή, το σύστημα θα πληροφορήσει το χρήστη για τις αναγκαίες αλλαγές, ώστε να διατηρήσει την πιο πρόσφατη αλλαγή του. Επίσης, το σύστημα μπορεί να επιλέξει χαρακτηριστικά που βελτιστοποιούν μια μεταβλητή (π.χ. την τιμή). Με αυτό τον τρόπο, ο χρήστης επιλέγει μόνο τα πιο σημαντικά για αυτόν χαρακτηριστικά θέτοντας τιμές μόνο σε αυτά και το σύστημα επιλέγει και θέτει τιμές στα υπόλοιπα ώστε να προκύψουν πλήρεις διαμορφώσεις [8].

Εκτός από τις βασισμένες σε κανόνες και τις στατικές συστάσεις (όπου δεν λαμβάνεται υπόψη το πόσο σημαντικό είναι το εκάστοτε χαρακτηριστικό για το συγκεκριμένο χρήστη, αλλά πόσο δημοφιλές είναι αυτό γενικά και προτείνεται η τιμή που επέλεξαν οι περισσότεροι χρήστες) έχουν προταθεί: η προσέγγιση του καθορισμού διαμορφώσεων *k-κοντινότερων γειτόνων* (k-nearest neighbors configurations) που είναι παρόμοιες με το τρέχον σύνολο απαιτήσεων και η προσέγγιση καθορισμού συστάσεων *με βάση την πλειοψηφία* (majority voting).

Ο χρήστης μπορεί να βοηθηθεί στον προσδιορισμό εναλλακτικών διαμορφώσεων με την παροχή προεπιλογών (defaults). Στο πλαίσιο των διαδραστικών διαλόγων διαμόρφωσης οι παρέχονται προεπιλεγμένες απαντήσεις στις ερωτήσεις του χρήστη. Για παράδειγμα, χρησιμοποιούνται για να εκφράσουν εξατομικευμένες προτάσεις χαρακτηριστικών τιμών. Στην έρευνα του Felfernig [10], μελετήθηκε ο αντίκτυπος των προσωποποιημένων προτάσεων χαρακτηριστικών, στην ικανοποίηση των χρηστών σε μια διαδικασία συστάσεων βασισμένη σε γνώση. Για τον υπολογισμό των προεπιλεγμένων τιμών χρησιμοποιήθηκαν αλγόριθμοι εύρεσης κοντινότερου γείτονα και Naïve Bayes

κατηγοριοποιητές. Τα αποτελέσματα της έρευνας υποδηλώνουν ότι η ικανοποίηση των χρηστών αυξάνεται με προσωποποιημένες προεπιλογές.

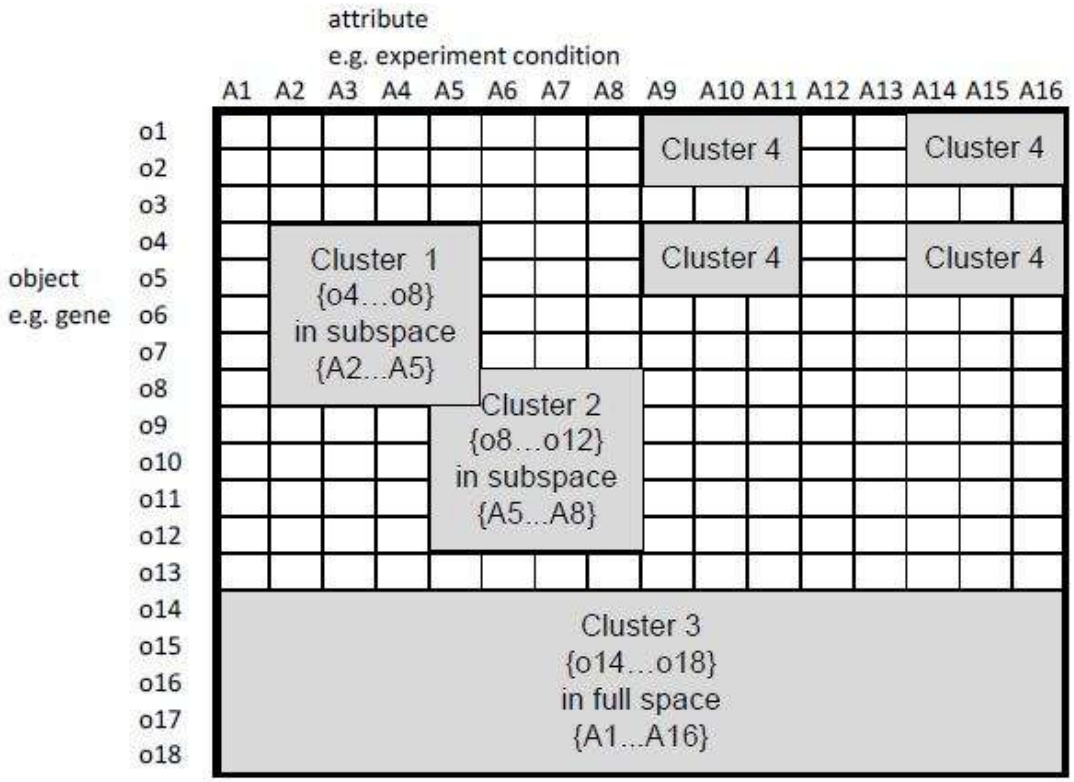
Ο παρακάτω μαθηματικός τύπος εκφράζει τη βασική προσέγγιση για τον καθορισμό συστάσεων χαρακτηριστικών τιμών για τη μεταβλητή v_i βάσει πλειοψηφίας όπου a_j είναι η j-ιοστή τιμή στο dom(v_i) και v_{ik} δηλώνει την τιμή της v_i της διαμόρφωσης για την σύνοδο $s_k \in S$.

$$majority(v_i) = \text{argmax}_{(j=1\ldots|dom(v_i)|)} \left(\sum_{k=1}^{|S|} v_{ik} = a_j \right) \quad (8)$$

Με εξαίρεση τις βασισμένες σε κανόνες προσεγγίσεις, οι υπόλοιπες τεχνικές για πρόταση χαρακτηριστικών τιμών δεν εγγυώνται τη συνέπεια των προτεινόμενων τιμών με τις δοσμένες απαιτήσεις C_R και τη δεδομένη γνώση για το αντικείμενο P_{KB}. Πριν την πρόταση κάποιας τιμής χαρακτηριστικού είναι απαραίτητος ο έλεγχος συνέπειας των προτάσεων με τους περιορισμούς στο σύνολο C_R. Εναλλακτικά, αντί να καθοριστούν τιμές χαρακτηριστικών συνεπείς με τους περιορισμούς για τη διαμόρφωση, μπορεί να γίνει από το σύστημα εκκίνηση χειρισμού ασυνέπειας με την παρουσίαση ελαχίστων επεξηγήσεων των ασυνεπειών μεταξύ των απαιτήσεων και των υπολογισμένων προτάσεων τιμών χαρακτηριστικών [9].

3.2 Αλγόριθμοι Ομαδοποίησης σε προβολές του υποχώρου (Clustering in Subspace Projections)

Η ομαδοποίηση σε προβολές του υποχώρου δεδομένων στοχεύει στην ανίχνευση ομάδων παρόμοιων αντικειμένων και ενός συνόλου σχετικών διαστάσεων για κάθε ομάδα αντικειμένων. Στη βιβλιογραφία συναντώνται κυρίως δύο διαφορετικά ονόματα, *ομαδοποίηση υποχώρων* (subspace clustering) και *προβολική ομαδοποίηση* (projective clustering) αλλά μπορούμε να διακρίνουμε τρεις κύριες κατηγορίες πολυδιάστατης ομαδοποίησης με βάση τον υποκείμενο ορισμό της ομαδοποίησης καθώς και την παραμετροποίηση της προκύπτουσας ομαδοποίησης [24]. Στις επόμενες ενότητες περιγράφουμε τις εν λόγω κατηγορίες καθώς και τους αλγόριθμους που τις καθιέρωσαν θεωρώντας μόνο την περίπτωση ομαδοποίησης υποχώρων σε γνωρίσματα (attributes) με συνεχείς (αριθμητικές) τιμές. Θεωρούμε μια αφηρημένη, πολλαπλών διαστάσεων βάση δεδομένων με αντικείμενα που περιγράφονται από διάφορα γνωρίσματα. Όπως εικονίζεται για παράδειγμα στην Εικόνα 1 μια προβολή υποχώρου είναι ένα αυθαίρετο υποσύνολο γνωρισμάτων. Κάθε ομάδα (cluster) περιγράφεται από ένα υποσύνολο αντικειμένων (γραμμές) και ένα υποσύνολο γνωρισμάτων (στήλες). Όπως έχει προαναφερθεί από προηγούμενο κεφάλαιο σε ορισμένες προσεγγίσεις οι ομάδες σε προβολές υποχώρου μπορεί να επικαλύπτονται και στα αντικείμενα και στις διαστάσεις αφού η ομοιότητα μεταξύ αντικειμένων εκτιμάται υπολογίζοντας μόνο τις σχετικές διαστάσεις.



Εικόνα 1: Παράδειγμα ομαδοποίησης υποχώρων

3.2.1 Προσέγγιση βασισμένη σε κελιά (Cell-based Approach)

Η βασισμένη σε κελιά προσέγγιση βασίζεται σε μια προσέγγιση του χώρου δεδομένων με κελιά. Για όλες τις προσεγγίσεις αυτού του τύπου, το αποτέλεσμα ομάδας R αποτελείται από σύνολο κελιών που το καθένα περιέχει περισσότερα από «τ» αντικείμενα όπου το «τ» έχει οριστεί ως κατώφλι. Αυτά τα κελιά περιγράφουν τα αντικείμενα των ομάδων είτε με υπερκύβο μεταβλητού πλάτους *w* ή με σταθερό πλέγμα *ξ* κελιών σε κάθε διάσταση.

Ομάδα υποχώρου βασισμένη σε κελιά: Μια ομάδα υποχώρου βασισμένη σε κελιά (O, S) ορίζεται αναφορικά με τον ελάχιστο αριθμό αντικειμένων *τ* σε κελιά CS πλάτους w που προσδιορίζονται από διαστήματα I_i ανά διάσταση $\forall\ i \in S$. Κάθε διάστημα είναι μέρος του κοινού χώρου $I_i = [l_i, \dots, u_i] \subseteq [0 \dots u]$ με κάτω και άνω όριο l_i και u_i. Για όλες τις μη σχετικές διαστάσεις $\forall\ j \in D \backslash S$ το διάστημα είναι ο πλήρης χώρος I_j = [0 … u]. Τα ομαδοποιημένα αντικείμενα O = { o | o ∈ $DB \cap CS$ } και ικανοποιούν το | O | ≥ *τ*.

Η πρώτη προσέγγιση για ομαδοποίηση βασισμένη σε κελιά εισήχθη με τον αλγόριθμο CLIQUE [20]. Ο αλγόριθμος CLIQUE βρίσκει ομάδες πρώτα χωρίζοντας κάθε διάσταση σε ξ διαστήματα ίσου πλάτους και στη συνέχεια αποθηκεύει ως ομάδες τα διαστήματα εκείνα στα οποία η πυκνότητα είναι μεγαλύτερη από το κατώφλι *τ*. Συνεπώς οι είσοδοι του αλγορίθμου είναι εκτός από τα δεδομένα και ο αριθμός των διαστημάτων ξ καθώς και το κατώφλι πυκνότητας *τ*. Στη συνέχεια κάθε σύνολο δύο διαστάσεων εξετάζεται: εάν υπάρχουν δύο τεμνόμενα διαστήματα σε αυτές τις δύο διαστάσεις και η πυκνότητα στην

τομή των διαστημάτων αυτών είναι μεγαλύτερη από *τ*, τότε η τομή των διαστημάτων αποθηκεύεται ως ομάδα. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα σύνολα τριών, τεσσάρων, πέντε και ούτω καθεξής διαστάσεων. Μετά από κάθε βήμα, οι γειτονικές (adjacent) ομάδες αντικαθίστανται από μια κοινή (joint) ομάδα. Πιο πρόσφατες προσεγγίσεις αυτού του τύπου όπως ο DOC χρησιμοποιούν ευέλικτους υπερκύβους πλάτους *w* [33]. Στον MINECLUS οι υπερκύβοι υποστηρίζονται από FP δέντρα που είναι γνωστό από την εξόρυξη συνολοστοιχείων ότι πετυχαίνουν καλύτερους χρόνους εκτέλεσης [34]. Τέλος ο SCHISM βελτιώνει τον ορισμό των ομάδων με μεταβλητά κατώφλια που προσαρμόζονται στο πλήθος διαστάσεων του υποχώρου [35].

Όσον αφορά την αποδοτικότητα της ομαδοποίησης, καθώς η ομαδοποίηση υποχώρων ψάχνει για ομάδες σε αυθαίρετους υποχώρους η αφελής αναζήτηση είναι εκθετική στον αριθμό των διαστάσεων. Ο CLIQUE προτείνει ένα κριτήριο κλαδέματος (pruning) για αποδοτική ομαδοποίηση υποχώρων με βάση μια ιδιότητα μονοτονίας. Μια παρόμοια ιδιότητα μονοτονίας εισήχθη στον αλγόριθμο apriori [25] για αποδοτική εξόρυξη συχνών συνολοστοιχείων και υιοθετήθηκε από την ομαδοποίηση υποχώρων.

3.2.2 Προσέγγιση βασισμένη στην πυκνότητα (Density-based Approach)

Οι προσεγγίσεις βάσει πυκνότητας βασίζονται στην ομαδοποίηση που προτάθηκε με τον αλγόριθμο DBSCAN [26]. Στις προσεγγίσεις αυτές η πυκνότητα κάθε αντικειμένου υπολογίζεται μετρώντας το πλήθος αντικειμένων στην ε-γειτονιά του χωρίς να έχει γίνει προηγουμένως καμία διακριτοποίηση (discretization) του χώρου. Μια ομάδα ορίζεται ως σύνολο πυκνών αντικειμένων που έχουν περισσότερα από έναν ελάχιστο αριθμό αντικειμένων στην ε-γειτονιά τους. Ομάδες αυθαίρετου σχήματος σχηματίζονται από αλυσίδα αντικειμένων που απέχουν το πολύ απόσταση ε το ένα από το άλλο. Για τον καθορισμό της γειτονιάς των αντικειμένων χρησιμοποιείται κάποια συνάρτηση απόστασης όπως για παράδειγμα η Ευκλείδεια απόσταση. Η παραμετροποίηση της ομοιότητας μέσω της παραμέτρου ε και της επιλογής διαφορετικής συνάρτησης απόστασης καθιστά την προσέγγιση αυτή ευέλικτη αλλά απαιτεί πρότερη γνώση για τα δεδομένα που συνήθως δεν είναι εφικτή στην χωρίς επίβλεψη μάθηση.

Η πρώτη βασισμένη στην πυκνότητα τεχνική ομαδοποίησης ήταν ο SUBCLU [27] προέκταση του αλγορίθμου DBSCAN για ομαδοποίηση υποχώρων με τον περιορισμό του υπολογισμού πυκνότητας μόνο στις σχετικές διαστάσεις. Ο αλγόριθμος ακολουθεί προσέγγιση από τη βάση και πάνω ξεκινώντας με τη δημιουργία μονοδιάστατων ομάδων με DBSCAN και στην συνέχεια κάθε ομάδα επεκτείνεται με μια διάσταση τη φορά σε μια διάσταση που είναι γνωστό ότι έχει ομάδα. Όλες οι ομάδες σε έναν υποχώρο με περισσότερες διαστάσεις θα είναι υποσύνολα των ομάδων που βρέθηκαν στην πρώτη ομαδοποίηση. Ο SUBCLU παράγει αναδρομικά κ+1 – διάστατους υποψήφιους υποχώρους συνδυάζοντας κ-διάστατους υποχώρους με ομάδες που μοιράζονται κ–1 γνωρίσματα. Μετά το κλάδεμα περιττών υποψηφίων, ο DBSCAN εφαρμόζεται στον υποψήφιο υποχώρο για να βρεθεί αν ακόμα περιέχει ομάδες. Αν ναι, ο υποψήφιος υποχώρος χρησιμοποιείται

για τον επόμενο συνδυασμό υποχώρων. Για να βελτιωθεί η ταχύτητα εκτέλεσης του DBSCAN λαμβάνονται υπόψη μόνο τα σημεία που είναι γνωστό ότι ανήκουν σε ομάδες στον κ-διάστατο υποχώρο.

3.2.3 Προσέγγιση προσανατολισμένη στην ομαδοποίηση (Clustering-oriented Approach)

Σε αντίθεση με τις προηγούμενες προσεγγίσεις, οι τεχνικές που προσανατολίζονται στην ομαδοποίηση εστιάζουν στο αποτέλεσμα ομαδοποίησης R καθορίζοντας άμεσα αντικειμενικές συναρτήσεις όπως ο αριθμός των ομάδων που θα ανιχνευθούν ή ο μέσος αριθμός διαστάσεων των ομάδων όπως στον πρωτοπόρο στον τομέα αυτό αλγόριθμο PROCLUS [28]. Ο συγκεκριμένος αλγόριθμος επεκτείνει τον K-means χωρίζοντας τα δεδομένα σε *k* ομάδες με μέσο αριθμό διαστάσεων *l.* Οι προσεγγίσεις που προσανατολίζονται στην ομαδοποίηση αντί να βασίζονται σε κάποιο ορισμό ομάδας καθορίζουν τις ιδιότητες του συνόλου των παραγόμενων ομάδων. Κάθε αντικείμενο ανατίθεται στην ομάδα στην οποία ταιριάζει καλύτερα. Στην ίδια κατηγορία αλγορίθμων ανήκουν ο P3C [29] και ο STATPC [30]. Ο P3C βασίζεται στη χρήση στατιστικών κατανομών για την εύρεση ομάδων. Κάθε διάσταση πρώτα χωρίζεται σε $1 + \log_2 nrow(data)$ κάδους που στη βιβλιογραφία ονομάζονται *πυρήνες* ομάδων (cluster cores) και ο έλεγχος x – τετραγώνου (chi-square, x^2) χρησιμοποιείται για να υπολογίσει την πιθανότητα οι πλευρές των κάδων αυτών να είναι ομοιόμορφα κατανεμημένες. Αν αυτή η πιθανότητα είναι μεγαλύτερη από 1 – chi-square alpha τότε τίποτα δε συμβαίνει. Διαφορετικά οι μεγαλύτεροι κάδοι θα αφαιρεθούν μέχρι να φτάσει η πιθανότητα στο όριο του 1 - chi-square alpha. Οι κάδοι που αφαιρέθηκαν με αυτό τον τρόπο χρησιμοποιούνται για την εύρεση ομάδων. Οι γειτονικοί κάδοι συγχωνεύονται και κατόπιν οι ομάδες σχηματίζονται παίρνοντας ένα κάδο από μια διάσταση και καθορίζοντας την πιθανότητα να μοιραστούν τόσα σημεία όσα με κάθε άλλο κάδο από άλλη διάσταση. Τέλος, ο κάδος τέμνεται με τον κάδο από την άλλη διάσταση όπου αυτή η πιθανότητα είναι η μικρότερη δεδομένου ότι είναι τουλάχιστον μικρότερη από $1 \times 10^{-Poisson\ Threshold}$ και η διαδικασία επαναλαμβάνεται μέχρι να μη βρίσκεται τέτοιος κάδος.

Ο STATPC προσδιορίζει μια στατιστικά σημαντική πυκνότητα για να επιλέξει την βέλτιστη μη πλεοναστική ομαδοποίηση. Παρόλο που προσδιορίζει ιδιότητες των ομάδων επιχειρεί μια συνολική βελτιστοποίηση του αποτελέσματος ομαδοποίησης *R*.

Και οι τεχνικές που βασίζονται σε κελιά αλλά και αυτές που βασίζονται σε πυκνότητα παρέχουν ορισμό των ομάδων και κάθε σύνολο αντικειμένων *O* και σύνολο διαστάσεων *S* που ικανοποιούν αυτό τον ορισμό ανιχνεύονται ως ομάδα υποχώρου (*O, S*). Δεν υπάρχει διαδικασία βελτιστοποίησης στην επιλογή ομάδων. Από την άλλη πλευρά οι προσανατολισμένες σε ομαδοποίηση τεχνικές δεν επηρεάζουν τις ομάδες που ανιχνεύονται και με αυτό τον τρόπο στις ομάδες εισάγεται θόρυβος. Επειδή αυτές οι προσεγγίσεις προσπαθούν να βελτιστοποιήσουν συνολικά την ομαδοποίηση, προσπαθούν να αναθέσουν κάθε αντικείμενο σε ομάδα με αποτέλεσμα ομάδες που περιέχουν πολύ

ανόμοια αντικείμενα και συνεπώς θόρυβο. Παρόλο που χρησιμοποιούνται μηχανισμοί για την απομάκρυνση του θορύβου στις προσεγγίσεις αυτές η ποιότητα της ομαδοποίησης επηρεάζεται ακόμα από την ανομοιότητα των αντικειμένων.

4. ΠΡΟΤΕΙΝΟΜΕΝΗ ΤΕΧΝΙΚΗ

4.1 Περιγραφή τεχνικής

Στο συγκεκριμένο κεφάλαιο, θα περιγράψουμε την τεχνική που προτείνουμε για την επίλυση του προβλήματος συστάσεων για ανασχηματιζόμενα προϊόντα σε ομάδες χρηστών χρησιμοποιώντας μια υβριδική μέθοδο ΣΦ με χρήση ομαδοποίησης υποχώρων. Σε αντίθεση με τις παραδοσιακές τεχνικές ομαδοποίησης, η ομαδοποίηση σε υποχώρους επιτρέπει τα αντικείμενα να ανήκουν σε πολλαπλές ομάδες σε διαφορετικούς υποχώρους.

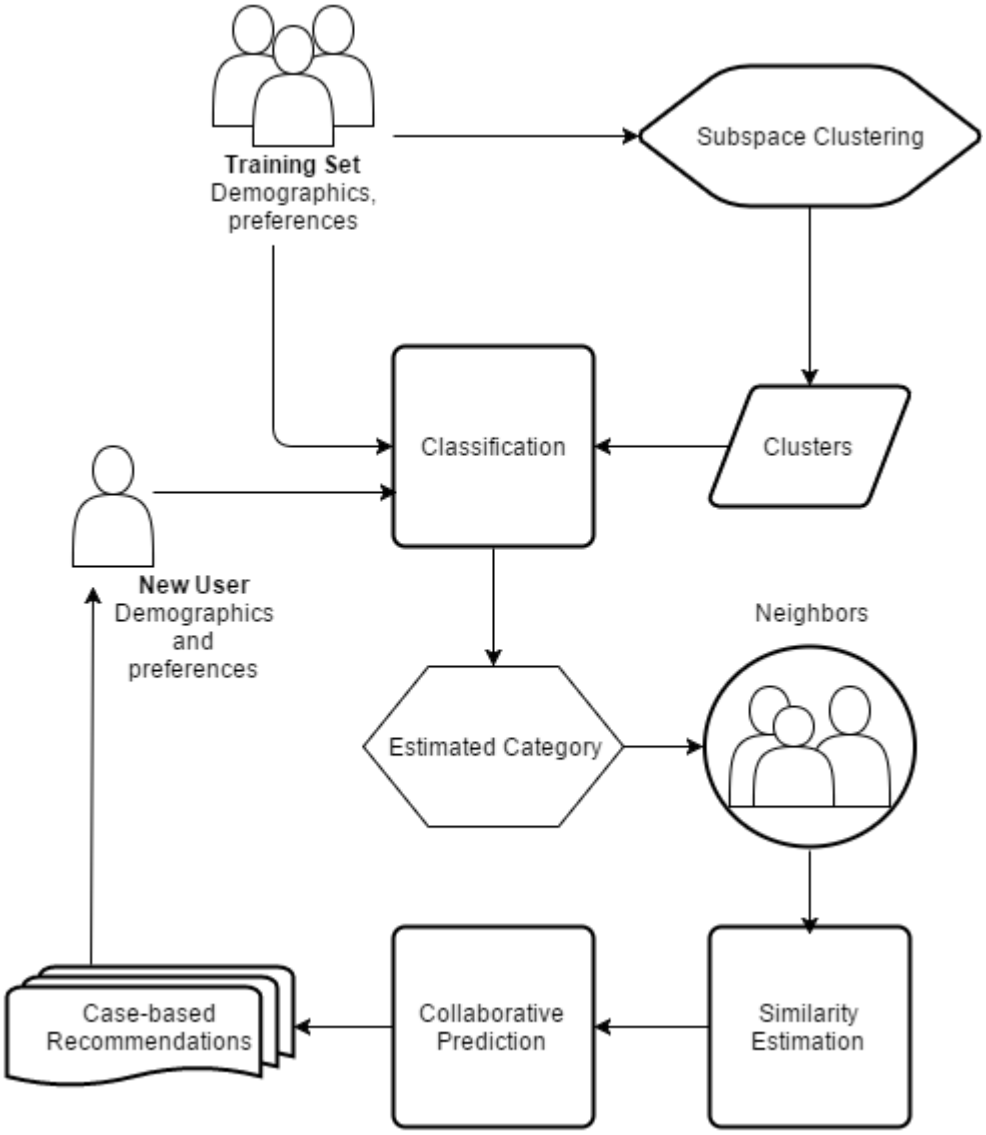
Προσεγγίζουμε το πρόβλημα από την πλευρά ενός συνόλου νέων χρηστών για τους οποίους δεν υπάρχει ιστορικό προηγούμενων ανασχηματιζόμενων προϊόντων που επέλεξαν στο σύστημα αλλά είναι διαθέσιμα τα δημογραφικά τους στοιχεία και οι προτιμήσεις τους για κάθε χαρακτηριστικό του αντικειμένου (δηλαδή κατά πόσο τους ενδιαφέρει και είναι σημαντικό για αυτούς). Η διαδικασία πρόβλεψης χωρίζεται στα ακόλουθα στάδια: 1) ομαδοποίηση όλων των χρηστών με πολυδιάστατη ομαδοποίηση υποχώρων (multidimensional subspace clustering) την οποία χρησιμοποιούμε ως κατηγοριοποίηση χωρίς εκμάθηση προκειμένου να γίνει μια αρχική ομαδοποίηση των χρηστών βάσει δημογραφικών στοιχείων και προτιμήσεων, 2) κατηγοριοποίηση των υπαρχόντων χρηστών με βάση τις ομάδες του πρώτου βήματος και δημιουργία ενός μοντέλου που βασίζεται στα ανωτέρω δεδομένα και στη συνέχεια εφαρμογή του μοντέλου αυτού για την εύρεση γειτόνων κάθε νέου χρήστη, 3) υπολογισμός ομοιότητας του νέου χρήστη και των γειτόνων βάσει σταθμισμένου μέσου όρου που λαμβάνει υπόψη τα δημογραφικά δεδομένα των χρηστών, και 4) συνεργατική πρόβλεψη για ομάδες νέων χρηστών που συνδυάζει το ιστορικό επιλογών αντικειμένων των γειτόνων με τις συνολικές ομοιότητες που έχουν οι ομάδες με καθέναν από τους γείτονες τους.

Τα σημαντικότερα στοιχεία της προτεινόμενης τεχνικής που απεικονίζεται στο διάγραμμα της Εικόνα 2: Διάγραμμα ροής για την προτεινόμενη τεχνική αναλύονται παρακάτω:

Ομαδοποίηση υποχώρων (Subspace clustering): Με την ομαδοποίηση υποχώρων που σχηματίζουν οι προτιμήσεις και τα χαρακτηριστικά των χρηστών μπορούμε να ανιχνεύσουμε πολλαπλές έννοιες (δηλαδή ομάδες) που αντιπροσωπεύουν ενδιαφέροντα χρηστών σε σχέση με τα δημογραφικά τους δεδομένα που δεν θα μπορούσαν να βρεθούν με τεχνικές παραδοσιακής ομαδοποίησης.

Κατηγοριοποίηση (Classification): Με την τεχνική αυτή παράγεται το μοντέλο του κατηγοριοποιητή. Αρχικά καθορίζονται τα γνωρίσματα που χρησιμοποιούνται ως κριτήρια κατά τη διαδικασία της εκπαίδευσης του συνόλου δεδομένων για την παραγωγή του μοντέλου. Στη δική μας περίπτωση το σύνολο εκπαίδευσης θα είναι οι προτιμήσεις των χρηστών για τα διάφορα μέρη του ανασχηματιζόμενου προϊόντος καθώς και τα δημογραφικά στοιχεία των χρηστών που είναι ήδη εγγεγραμμένοι βάσει των οποίων δημιουργείται το μοντέλο. Στο σύνολο αυτό έχει εφαρμοστεί αλγόριθμος ομαδοποίησης οπότε οι κλάσεις θα προέρχονται από τις ομάδες που προέκυψαν. Κατά την

κατηγοριοποίηση νέου χρήστη, η είσοδος του μοντέλου είναι τα δημογραφικά στοιχεία και οι προτιμήσεις του νέου χρήστη ενώ η έξοδος είναι η εκτιμώμενη κατηγορία.



Εικόνα 2: Διάγραμμα ροής για την προτεινόμενη τεχνική

Εκτιμώμενη κατηγορία (Estimated Category): Είναι η έξοδος του μοντέλου δηλαδή η κατηγορία στην οποία ανήκει ο νέος χρήστης. Οι κατηγορίες δεν είναι εκ των προτέρων γνωστές και επειδή εκφράζουν έννοιες στον πολυδιάστατο χώρο που ορίζεται από τα γνωρίσματα των χρηστών και τα γνωρίσματα/μέρη των ανασχηματιζόμενων προϊόντων προσδιορίζονται στο αρχικό στάδιο της ομαδοποίησης.

Γείτονες (Neighbors): Είναι οι χρήστες που ανήκουν στην ίδια κατηγορία με το νέο χρήστη. Οι γείτονες παίζουν σημαντικό ρόλο στη διαδικασία σύστασης αφού το αντικείμενο προτείνεται τελικά βάσει του ιστορικού του γείτονα του οποίου οι προτιμήσεις είναι οι πιο κοντινές (όμοιες) με την αθροιστική (aggregated) προτίμηση για την ομάδα νέων χρηστών που ανήκει στην κατηγορία αυτή. Ο αλγόριθμος για την εύρεση των γειτόνων έχει ως είσοδο την κατηγορία στην οποία ανήκει ο νέος χρήστης καθώς και όλους τους

εγγεγραμμένους χρήστες τους συστήματος. Ο αλγόριθμος δίνει ως έξοδο τους γείτονες του νέου χρήστη.

Υπολογισμός Ομοιότητας (Similarity Estimation): Με τη διαδικασία αυτή υπολογίζεται η ομοιότητα κάθε νέου χρήστη με τους γείτονες του ως άθροισμα σταθμισμένων ομοιοτήτων για τα δημογραφικά στοιχεία τους. Θεωρούμε ότι όσο περισσότερο μοιάζει το προφίλ των χρηστών τόσο θα μοιάζουν και οι προτιμήσεις τους. Για κάθε δημογραφικό στοιχείο χρησιμοποιείται η κατάλληλη τεχνική υπολογισμού ομοιότητας. Για την ηλικία εφαρμόζεται μια εκθετική συνάρτηση, για το επάγγελμα μια μετρική σημασιολογικής ομοιότητας και για το φύλο μια δυαδική μεταβλητή. Σε επόμενη ενότητα εξηγείται αναλυτικά ο τρόπος υπολογισμού. Οι επιμέρους ομοιότητες συνδυάζονται με κάποιους συντελεστές βαρύτητας και προκύπτει η συνολική σταθμισμένη ομοιότητα που αποτελεί και την έξοδο της διαδικασίας αυτής.

Πρόβλεψη βάσει συνεργατικού φιλτραρίσματος (Collaborative Prediction) και Συστάσεις βάσει Περιεχομένου (Case-based Recommendation): Το συγκεκριμένο κομμάτι του συστήματος είναι θεμελιώδες αφού παρέχει την τελική σύσταση για το κάθε μέρος του ανασχηματιζόμενου προϊόντος σε μια ομάδα χρηστών. Η σύσταση βασίζεται στις σταθμισμένες ομοιότητες που υπολογίστηκαν στο προηγούμενο στάδιο αλλά και στις προτιμήσεις των νέων χρηστών για το συγκεκριμένο μέρος του αντικειμένου και επίσης στο ιστορικό των γειτόνων. Πρώτα υπολογίζουμε την εκτιμώμενη προτίμηση ενός νέου χρήστη σε κάποιο γνώρισμα υπολογίζοντας το σταθμισμένο μέσο όρο των προτιμήσεων για το σύνολο των γειτόνων και στη συνέχεια αθροίζουμε τις εκτιμώμενες προτιμήσεις για τους νέους χρήστες που ο κατηγοριοποιητής ανέθεσε στην ίδια κατηγορία. Ο σταθμισμένος μέσος όρος προτιμήσεων για τους γείτονες προκύπτει από το άθροισμα του γινομένου των ομοιοτήτων μεταξύ νέου χρήστη και καθενός από τους γείτονές του sim(n,u) πολλαπλασιασμένο με την προτίμηση του νέου χρήστη για το συγκεκριμένο γνώρισμα του προϊόντος. Στη συνέχεια, αυτό το άθροισμα διαιρείται με το άθροισμα των sim(n, u). Τέλος βρίσκοντας με ποιον γείτονα μοιάζει περισσότερο η συνολική εκτιμώμενη προτίμηση για την ομάδα ενώ το ιστορικό από προηγούμενες επιτυχημένες συνόδους (sessions) αξιοποιείται ώστε να προταθεί στην ομάδα χρηστών το μέρος του προϊόντος που αυτός ο γείτονας είχε επιλέξει.

4.2 Προτεινόμενος Αλγόριθμος

Ορίζουμε ένα σύνολο νέων χρηστών $N = \{n_1, n_2 \ldots\ldots n_n\}$, ένα σύνολο χρηστών, $U = \{u_1, u_2 \ldots\ldots u_n\}$ που υπάρχουν ήδη στο σύστημα και έχουν εκτός από αποθηκευμένες προτιμήσεις επίσης το ιστορικό των επιλογών που έκαναν τελικά για τα μέρη του ανασχηματιζόμενου προϊόντος, ένα σύνολο δημογραφικών δεδομένων D = {ηλικία, επάγγελμα, φύλο, ταχυδρομικός κωδικός} για κάθε χρήστη και ένα σύνολο γνωρισμάτων που σχηματίζουν το διαμορφώσιμο αντικείμενο $P = \{p_1, p_2, \ldots, \ldots p_n\}$. Προβλέπουμε την προτίμηση για κάθε νέο χρήστη $n_i \in N$ για το μέρος του αντικειμένου p βρίσκοντας αρχικά τους γειτονικούς χρήστες του n_i, ένα υποσύνολο του U. Η γειτονιά του n_i ορίζεται με τη

βοήθεια ενός ταξινομητή (classifier), ο οποίος ταξινομεί τον n_i σε μία ομάδα (γειτονιά) με βάση τα δημογραφικά του δεδομένα (ηλικία, επάγγελμα, φύλο και Τ.Κ. - αυτή η λίστα μπορεί εύκολα να επεκταθεί) και τις προτιμήσεις του για κάθε μέρος του συνόλου *P*. Για κάθε ζευγάρι (n_i, u_i) υπολογίζουμε ένα δείκτη ομοιότητας βάσει μιας κανονικοποιημένης συνάρτησης παλινδρόμησης με ανεξάρτητες μεταβλητές τους δείκτες ομοιότητας για την ηλικία, το επάγγελμα και το φύλο των δύο χρηστών. Η ομοιότητα λοιπόν μεταξύ των χρηστών παραμετροποιείται από επιμέρους βάρη που καθορίζουν σε ποιο δημογραφικό δεδομένο θα δοθεί μεγαλύτερη βαρύτητα. Η συνολική σταθμισμένη ομοιότητα λαμβάνεται υπόψη στη διαδικασία πρόβλεψης των συστάσεων σε ομάδες χρηστών για τα ανασχηματιζόμενα αντικείμενα. Στη συνέχεια του κεφαλαίου παρουσιάζονται τα στάδια του προτεινόμενου αλγορίθμου.

4.2.1 Πολυδιάστατη ομαδοποίηση χρηστών

Σε πρώτο στάδιο έχουμε την προ-επεξεργασία (preprocessing) και την πολυδιάστατη ομαδοποίηση των δεδομένων. Χρησιμοποιούμε την ομαδοποίηση προκειμένου να βρούμε τις ομάδες που θα χρησιμεύσουν ως κατηγορίες - κλάσεις για την εκπαίδευση του αλγορίθμου κατηγοριοποίησης σε ένα σύνολο χρηστών. Στο βήμα αυτό συλλέγουμε ή δημιουργούμε τα δεδομένα με τη μορφή προφίλ χρηστών και αντικειμένων (αφορά δημογραφικά δεδομένα και ιστορικό προτιμήσεων αλλά και τελικών επιλογών ανασχηματιζόμενων αντικειμένων).

Παρακάτω παρουσιάζεται σε ψευδογλώσσα ο αλγόριθμος πολυδιάστατης ομαδοποίησης.

Algorithm: Multidimensional clustering

Input: Instances $u_i \in \{u_1,u_2,\ldots\ldots,u_n\}$, input attributes set $I_A = \{A_1, A_2,\ldots, A_d\}$, the number of intervals ξ and the density limit τ

Output: Clusters with clustered instances $u'_i \in \{u'_1,u'_2,\ldots\ldots,u'_n\}$

Begin

S = $A_1 \; x \; A_2 \; x$ ... A_d is the d-dimensional numerical space

Define options: ξ and τ

Select attributes $A_i \in I_A$ from 1-dimensional subspace to d-dimensional subspace to generate the attribute space s.

for each attribute subspace s:

Detect all the clusters C = {c_1, c_2, … c_n} in each attribute subspace s: given the threshold τ if the number of users in unit is greater than τ, consider this unit as candidate

Join the connected candidate as cluster $c_i \subset C$

Collect full clusters in the current subspace

end for

End

4.2.2 Κατηγοριοποίηση νέων χρηστών

Από το προηγούμενο βήμα έχουν προκύψει οι ομάδες των χρηστών τις οποίες θα χρησιμοποιήσουμε ως κατηγορίες-κλάσεις για τον αλγόριθμο κατηγοριοποίησης τόσο για την εκπαίδευση του αλγορίθμου σε ένα σύνολο χρηστών και τη δημιουργία του μοντέλου όσο και στη συνέχεια και την κατηγοριοποίηση των νέων χρηστών. Το μοντέλο λαμβάνει υπόψη τα δημογραφικά δεδομένα και τις προτιμήσεις του κάθε χρήστη. Στο σημείο αυτό είναι απαραίτητη η προεπεξεργασία των δεδομένων αλφαριθμητικού τύπου όπως π.χ. το επάγγελμα προκειμένου να μπορεί να τα χειριστεί ο κατηγοριοποιητής πολλαπλών κλάσεων (multiclass classifier) ο οποίος θα χρησιμοποιηθεί για την ταξινόμηση των χρηστών. Στην πρώτη φάση του αλγορίθμου εκπαιδεύουμε τον κατηγοριοποιητή με ένα σύνολο χρηστών U για τη δημιουργία του μοντέλου και ακολούθως με το μοντέλο αυτό μπορεί να προβλεφτεί η κατηγορία στην οποία θα ανήκει κάθε νέος χρήστης.

Για τους αλγορίθμους χρησιμοποιήσαμε τις υλοποιήσεις του Weka, ένα εργαλείο το οποίο παρέχει αλγορίθμους για τους τομείς της μηχανική μάθησης και της εξόρυξης δεδομένων. Παρακάτω παρουσιάζουμε τον αλγόριθμο πολλαπλών κλάσεων του (MultiClassClassifier) στον οποίο καλούμε τον C4.5 ή τον Logistic από τους κατηγοριοποιητές που παρέχει το Weka. Ο Logistic δημιουργεί και χρησιμοποιεί ένα μοντέλο πολυωνυμικής λογιστικής παλινδρόμησης με εκτιμητή κορυφών. Ο C4.5 (J48 στο Weka) δημιουργεί ένα δέντρο απόφασης.

Algorithm: Multi-class Classification (Logistic base)

Input: Instances $u_i \in \{u_1,u_2,\ldots\ldots,u_n\}$, Unlabeled Instances $n_i \in \{n_1,n_2,\ldots\ldots,n_n\}$

Output: Labeled $n_i \in \{n_1,n_2,\ldots\ldots,n_n\}$

Begin

Define Options: one-against-all, Random width factor, Random number seed, base classifier to use and its options

Set class index

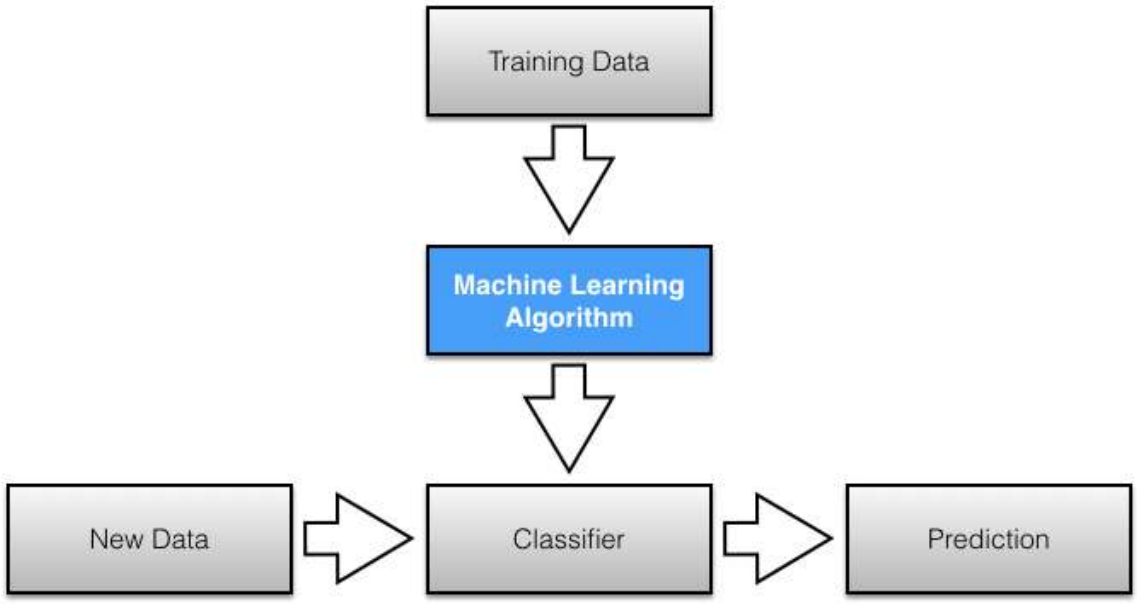
Build the Multi-class Classifier

Label unlabeled instances n_i

for each $n_i \in \{n_1,n_2,\ldots\ldots,n_n\}$

```
            Classify instance with trained model
        end for
End
```

Στην ακόλουθη εικόνα (Εικόνα 3) απεικονίζονται τα βήματα της διαδικασίας κατηγοριοποίησης.



Εικόνα 3: Διάγραμμα ροής για τη διαδικασία κατηγοριοποίησης

Δεδομένα εκπαίδευσης και αλγόριθμοι μηχανικής εκμάθησης (Training data and machine learning algorithm): Η φάση της εκπαίδευσης είναι η πιο σημαντική για τη δημιουργία του μοντέλου. Η είσοδος του αλγορίθμου είναι τα γνωρίσματα κλάσης (class attributes) και το σύνολο δεδομένων. Η ανάλυση του συνόλου δεδομένων παράγει το μοντέλο που χρησιμοποιείται για να προβλέψει την κατηγορία σε μέχρι πρότινος αταξινόμητα δεδομένα, δηλαδή μπορεί να ταξινομήσει δεδομένα χωρίς κλάση, όπως είναι οι νέοι χρήστες στην περίπτωση μας.

Μοντέλο (Model): Το μοντέλο προκύπτει από τον αλγόριθμο κατηγοριοποίησης και εξαρτάται από το σύνολο εκπαίδευσης καθώς και από τον εκάστοτε αλγόριθμο κατηγοριοποίησης.

4.2.3 Εύρεση της γειτονιάς των νέων χρηστών

Αφού έχει δημιουργηθεί το μοντέλο από τα δεδομένα εκπαίδευσης κι έχουμε βρει σε ποια κατηγορία ανήκει ο εκάστοτε n_i συγκεντρώνουμε όλους τους χρήστες u_i σε ομάδες ανάλογα

με την κατηγορία που ανήκει ο καθένας και η οποία προέκυψε από τον ταξινομητή πολλαπλών κλάσεων. Με αυτόν τον τρόπο ορίζουμε τους γείτονες του n_i.

Algorithm: Find neighbors of new users

Input: Labeled $n_i \in \{n_1, n_2, \ldots\ldots, n_n\}$, Instances $u_i \in \{u_1, u_2, \ldots\ldots, u_n\}$

Output: List of neighbors

Begin

for each labeled $n_i \in \{n_1, n_2, \ldots\ldots, n_n\}$

Find n_iCategory

for each u_i $\{u_1, u_2, \ldots\ldots, u_n\}$

Find u_iCategory

if n_iCategory.equals(u_iCategory) **then**

Add u_i in neighborsList

end if

end for

end for

End

4.2.4 Υπολογισμός της ομοιότητας χρηστών και αθροιστικής ομοιότητας ομάδων

Στη συνέχεια υπολογίζουμε την ομοιότητα των χρηστών με καθένα από τους γείτονες του. Η ομοιότητα ορίζεται από τον παρακάτω σταθμισμένο μέσο όρο:

$$sim(n, u) = \sum_{i=1}^{n} b_i w_i \quad (9)$$

Όπου w_i το βάρος της εκάστοτε συνάρτησης που υπολογίζει την ομοιότητα των χρηστών για καθένα από τα δημογραφικά δεδομένα π.χ. w_1 είναι το βάρος ομοιότητας του n_i με τον u_i για το επάγγελμα και b_i ο συντελεστής βαρύτητας για καθεμιά από τις επιμέρους ομοιότητες.

Στην πειραματική μας αποτίμηση εστιάζουμε σε τρία είδη δημογραφικών δεδομένων. Σε γενικότερη περίπτωση, η μεθοδολογία που περιγράφουμε πιο κάτω μπορεί εύκολα να επεκταθεί. Η παραπάνω εξίσωση μετατρέπεται ως εξής:

$$sim(n, u) = b_1 AgeSim + b_2 OccupationSim + b_3 GenderSim \quad (10)$$

Για τον υπολογισμό της ομοιότητας της ηλικίας AgeSim χρησιμοποιούμε την παρακάτω εκθετική συνάρτηση λαμβάνοντας υπόψη τη διαφορά της ηλικίας των δύο χρηστών καθώς και μια μέγιστη τιμή ηλικίας:

$$AgeSim(n,u) = \left(1 - \frac{|D|}{D_{max}}\right)^{a} \quad (11)$$

Όπου *D* η διαφορά ηλικίας, D_{max} η εκτιμώμενη μέγιστη τιμή της ηλικίας και α η παράμετρος της εκθετικής συνάρτησης με τιμές στο διάστημα (0, ∞). Όταν $a < 1$ τότε όσο μεγάλη να είναι η διαφορά στην ηλικία θεωρούμε ότι υπάρχει ομοιότητα.

Για να υπολογίσουμε την ομοιότητα του επαγγέλματος χρησιμοποιήσαμε μια από τις μετρικές σημασιολογικής ομοιότητας, την Wu and Palmer [31]. Η συγκεκριμένη μετρική συγκρίνει δύο έννοιες/λέξεις. Υπολογίζει την ομοιότητα βάσει του μήκους μονοπατιού από το least common subsumer (LCS) αλλά και από τα βάθη των δύο ομάδων συνωνύμων (synsets) στις ταξονομίες του Wordnet [32], μια λεξιλογική βάση δεδομένων της Αγγλικής γλώσσας. Δεδομένου δύο λέξεων, το LCS ορίζεται ο πιο κοινός κόμβος-έννοια από τον οποίο προέρχονται οι συγκεκριμένες έννοιες που εξετάζουμε. Το εννοιολογικό δέντρο ορίζεται από σχέσεις is-a. Μια έννοια θεωρείται ότι είναι πρόγονος μια άλλης όπως ακριβώς ορίζεται σε ένα γενεαλογικό δέντρο. Για παράδειγμα το LCS για τις έννοιες «αυτοκίνητο» και «βάρκα» μπορεί να οριστεί ότι είναι το «όχημα». Η εξίσωση που υπολογίζει την μετρική για δύο έννοιες c1, c2 είναι η εξής:

$$OccupationSim(n.u) = sim_{wup}\ (c1, c2) = \frac{2\ x\ depth\ (LCS\ (c1,c2))}{depth(c1) + depth(c2)} \quad (12)$$

Στην υλοποίηση χρησιμοποιήθηκε το WS4J[1] (WordNet Similarity for Java) το οποίο παρέχει ένα Java API για διάφορους δημοσιευμένους αλγορίθμους συσχέτισης και ομοιότητας και αποτελεί υλοποίηση σε Java του WordNet::Similarity[2] που είναι η πρωτότυπη υλοποίηση σε Perl από την ομάδα του καθηγητή Ted Pedersen στο Πανεπιστήμιο της Μινεσότα.

Τέλος, το GenderSim προκύπτει απλά από δύο δυαδικές τιμές, 0 εάν το φύλο του *ni* δεν είναι ίδιο με το φύλο του *ui* και 1 εφόσον τα δύο φύλα είναι ίδια.

Ακολουθεί η συνοπτική παρουσίαση του αλγορίθμου για εύρεση της ομοιότητας χρηστών:

Algorithm: Weighted similarity Calculation

Input: New users $n_i \in \{n_1, n_2, \ldots\ldots, n_n\}$, Neighbors $u_i \in \{u_1, u_2, \ldots\ldots, u_n\}$

Output: Average weight $\overline{w}$

Begin

for each $n_i \in \{n_1, n_2, \ldots\ldots, n_n\}$

Find weight for age

Find weight for occupation

Find weight for gender

[1] https://code.google.com/archive/p/ws4j/
[2] http://wn-similarity.sourceforge.net/

Calculate average weight

end for

End

4.2.5 Συνεργατικές συστάσεις ανασχηματιζόμενων προϊόντων

Για να προβλέψουμε την προτίμηση (preference) ενός νέου χρήστη σε ένα χαρακτηριστικό a υπολογίζουμε ένα σταθμισμένο μέσο όρο των προτιμήσεων που έχει το σύνολο U των γειτόνων του (όπου γείτονες οι χρήστες που ανήκουν στην κατηγορία στην οποία ταξινομήθηκε από τον classifier ο νέος χρήστης). Η προτίμηση για το νέο χρήστη δίνεται από τον ακόλουθο τύπο:

$$P_{n,a} = \frac{\sum_{u \in U} sim(n,u) * p_{u,a}}{\sum_{u \in U} sim(n,u)} \quad (13)$$

Όπου sim(n,u) είναι η ομοιότητα του νέου χρήστη *n* με τον γείτονα *u* και $p_{u,a}$ είναι η προτίμηση του χρήστη *u* για το χαρακτηριστικό του προϊόντος a.

Στη συνέχεια για τους νέους χρήστες που ανήκουν στην ίδια κατηγορία θα πρέπει να προκύψουν οι μέσες αθροιστικές προτιμήσεις (ως περιορισμοί που καθορίζουν την ομάδα) ώστε να δώσουμε προτάσεις που απευθύνονται σε ομάδες νέων χρηστών και όχι για μεμονωμένους χρήστες. Σε αυτή τη φάση χρησιμοποιούμε και τις αποθηκευμένες συνόδους από το ιστορικό των (παλιών) χρηστών με τις επιλογές που έκαναν για τα ανασχηματιζόμενα προϊόντα. Δηλαδή, βρίσκουμε με ποιον χρήστη από την ομάδα μοιάζει περισσότερο για κάθε γνώρισμα η αθροιστική προτίμηση που προέκυψε (ελάχιστη απόσταση) και τέλος προτείνουμε το αντίστοιχο μέρος όπως το επέλεξε αυτός βάσει των δεδομένων από το ιστορικό του. Στόχος είναι η τελική σύσταση να είναι η πλέον αποδεκτή και βέλτιστη για την ομάδα.

Algorithm: Preference Prediction and recommendation

Input: New users $n_i \in \{n_1, n_2, \ldots\ldots, n_n\}$, Neighbors $u_i \in \{u_1, u_2, \ldots\ldots, u_n\}$, Preferences for each attribute a, stored sessions

Output: recommendation for attribute a

Begin

for each $n_i \{n_1, n_2, \ldots\ldots, n_n\}$

Calculate weighted similarities

end for

for each $u_i \{u_1, u_2, \ldots\ldots, u_n\}$

Find the preference in attribute a

end for

Calculate the predicted preference $P_{n,a}$ and aggregate predicted preference for new users in the same group

Find the minimum distance between group neighbor users' preferences and the group's aggregated preference

Recommend part selected on stored session of the user that is most similar to the group

End

5. ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ

Στο παρόν κεφάλαιο παρουσιάζεται ο τρόπος με τον οποίο έγινε η πειραματική αξιολόγηση του συστήματος. Αρχικά παρουσιάζονται οι μετρικές απόδοσης που καθορίζουν την αποδοτικότητα του συστήματος και αξιολογούν την ικανότητα πρόβλεψης. Επιπλέον αναφέρονται εκτενώς τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική αποτίμηση. Τέλος, αναφερόμαστε στα διαφορετικά σενάρια βάσει των οποίων ορίζονται τα πειράματα και σχολιάζουμε τα αντίστοιχα αποτελέσματα. Στόχος της αξιολόγησης είναι η μέτρηση της απόδοσης του συστήματος καθώς και η επιβεβαίωση της ορθής λειτουργίας του.

5.1 Μετρικές απόδοσης

Οι μετρικές δείχνουν το πόσο διαφέρουν οι προβλέψεις από τις πραγματικές αξιολογήσεις. Χρησιμοποιούμε δύο μετρικές ευρέως αποδεκτές για την αξιολόγηση συστημάτων συστάσεων οι οποίες έχουν χρησιμοποιηθεί σε αρκετές ερευνητικές μελέτες στον τομέα των συστημάτων συστάσεων. Μία μετρική απόδοσης ενός συστήματος συστάσεων είναι το *μέσο απόλυτο σφάλμα* (Mean Absolute Error - MAE). Το MAE υπολογίζει το μέσο όρο της απόλυτης τιμής της διαφοράς ανάμεσα στις προβλεπόμενες και τις πραγματικές αξιολογήσεις. Η εξίσωση του MAE είναι η ακόλουθη:

$$MAE = \frac{1}{N} \sum_{ij} |p_{i,j} - r_{i,j}| \quad (14)$$

όπου *N* είναι το σύνολο των αντικειμένων για τα οποία γίνεται πρόβλεψη, p_{ij} είναι η τιμή πρόβλεψης για ένα χρήστη i σε ένα αντικείμενο j και r_{ij} είναι η πραγματική αξιολόγηση. Συγκεκριμένα, για τα πειράματα μας τα οποία αφορούν ανασχηματιζόμενα προϊόντα με συγκεκριμένο πλήθος χαρακτηριστικών η εξίσωση (14) γράφεται:

$$MAE = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} |p_{i,j} - r_{i,j}| \quad (15)$$

όπου *n* το σύνολο των χρηστών και *m* το σύνολο των χαρακτηριστικών, p_{ij} είναι η εκτιμώμενη προτίμηση για ένα χρήστη i σε ένα χαρακτηριστικό j και r_{ij} είναι η πραγματική του προτίμηση. Μία άλλη μετρική απόδοσης για την αξιολόγηση της αποδοτικότητας των συστημάτων συστάσεων είναι η *ρίζα του μέσου τετραγωνικού σφάλματος* (Root Mean Squared Error - RMSE). Το RMSE υπολογίζει τη τετραγωνική ρίζα της μέσης τιμής της διαφοράς υψωμένη στο τετράγωνο και ο μαθηματικός του τύπος δίνεται παρακάτω:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i,j} \left(p_{i,j} - r_{i,j}\right)^2} \quad (16)$$

όπου $N = n \times m$ είναι το σύνολο των αντικειμένων για τα οποία γίνεται πρόβλεψη, p_{ij} είναι η τιμή πρόβλεψης για ένα χρήστη i σε ένα χαρακτηριστικό j και r_{ij} είναι η πραγματική αξιολόγηση. Όσο μικρότερες είναι οι τιμές αυτές τόσο πιο ακριβείς είναι οι προβλέψεις και κατά συνέπεια και καλύτερη η απόδοση του αλγορίθμου.

5.2 Σύνολα δεδομένων

Στα πειράματά μας υιοθετούμε το σύνολο δεδομένων του MovieLens[3], μία εφαρμογή η οποία προσφέρει προσωποποιημένες συστάσεις για ταινίες. Οι ερευνητές του GroupLens έχουν συλλέξει και δημοσιοποιήσει σύνολα δεδομένων με βαθμολογίες χρηστών για ένα πλήθος ταινιών. Για την πειραματική αποτίμηση της τεχνικής χρησιμοποιήσαμε ένα σύνολο δεδομένων με 100.000 αξιολογήσεις 943 χρηστών για 1682 διαφορετικές ταινίες[4] όπου κάθε χρήστης έχει βαθμολογήσει τουλάχιστον 20 ταινίες. Επιλέξαμε τους πρώτους Χ χρήστες από το σύνολο των 943 χρηστών να αποτελούν τους εγγεγραμμένους χρήστες του συστήματος και τους τελευταίους 50 χρήστες ως νέους χρήστες. Τα πειράματα ξεκίνησαν με 300 εγγεγραμμένους χρήστες και στα πλαίσια των πειραμάτων μας αυξάνουμε τον αριθμό τους κατά 100 μέχρι το όριο των 671 χρηστών για να δούμε την επίδραση του αριθμού των εγγεγραμμένων χρηστών στα αποτελέσματα. Για την εκτέλεση των πειραμάτων επεξεργαστήκαμε κατάλληλα το σύνολο δεδομένων που αφορά τους χρήστες (u.user) καθώς και το σύνολο με τις βαθμολογίες των χρηστών για τις ταινίες (u.data) το οποίο τροποποιήθηκε κατάλληλα ώστε να εκφράζει αποθηκευμένες συνόδους. Τα δεδομένα τα οποία κρατήσαμε από το σύνολο u.user είναι το αναγνωριστικό (id) και τα δημογραφικά στοιχεία (ηλικία, επάγγελμα, φύλο και μέρος του Τ.Κ.). Σε αυτά τα δεδομένα για κάθε χρήστη προστέθηκαν οι προτιμήσεις του για κάποια χαρακτηριστικά του ταξιδιού (καθώς εξετάζουμε την περίπτωση που το ανασχηματιζόμενο προϊόν είναι ταξιδιωτικό πακέτο). Παράδειγμα τέτοιων χαρακτηριστικών αποτελούν το κόστος, ο προορισμός, ο τύπος καταλύματος για τη διαμονή των ταξιδιωτών, η επιλογή μεταφορικού μέσου για τον προορισμό (π.χ. αεροπορικώς, οδικώς, με πλοίο κ.λπ.) και η ενοικίαση οχήματος στον προορισμό και η προτίμηση του χρήστη για κάθε ένα από αυτά τα χαρακτηριστικά εκφράζεται ως δεκαδικός αριθμός στο διάστημα [0 , 1]. Οι αποθηκευμένες σύνοδοι των παλιών εγγεγραμμένων χρηστών σχηματίστηκαν από το σύνολο u.data κρατώντας για κάθε χρήστη τα 5 πρώτα αντικείμενα που βαθμολόγησε ως τιμές που επιλέχτηκαν για κάθε ένα από τα γνωρίσματα - μέρη του ανασχηματιζόμενου αντικείμενου σχηματίζοντας πλειάδες της μορφής: «user id | X1 | X2 | .. | Xn» όπου X1 η επιλογή για το πρώτο γνώρισμα κ.ο.κ.

5.3 Παράμετροι πειραμάτων

Ο προτεινόμενος αλγόριθμος, όπως περιγράφηκε σε προηγούμενη ενότητα, περιλαμβάνει το σημαντικό στάδιο της κατηγοριοποίησης για την αξιολόγηση του οποίου χρησιμοποιούμε δύο αλγόριθμους κατηγοριοποίησης. Στον κατηγοριοποιητή πολλαπλών κλάσεων εκτελέσαμε πειράματα χρησιμοποιώντας στη θέση του δυαδικού κατηγοριοποιητή για την One-against-All στρατηγική τον C4.5 (η υλοποίηση του στο Weka ονομάζεται J48) και τον αλγόριθμο Logistic. Για να συγκρίνουμε τις δύο διαφορετικές προσεγγίσεις χρησιμοποιούμε επίσης έναν αλγόριθμο αναφοράς (baseline) κατά τον οποίο δε δημιουργήθηκε μοντέλο για να προβλέψει την κατηγορία στην οποία ανήκει ο νέος χρήστης

[3] https://movielens.org/
[4] http://files.grouplens.org/datasets/movielens/ml-100k/

αλλά αυτή ανατέθηκε τυχαία χωρίς να έχει προηγηθεί το βήμα της πολυδιάστατης ομαδοποίησης. Τα αποτελέσματα του αλγορίθμου αναφοράς (που ονομάζουμε Random Classification Algorithm – RCA) συγκρίθηκαν με τα αποτελέσματα των υπόλοιπων μεθόδων. Δίνοντας διαφορετικές τιμές για τους συντελεστές βαρύτητας της εξίσωσης που δίνει την ομοιότητα των χρηστών με καθένα από τους γείτονες του καθώς και τις τιμές του εκθέτη για τον υπολογισμό της ομοιότητας της ηλικίας AgeSim προέκυψαν τα διαφορετικά σενάρια. Οι διάφορες τιμές για τα βάρη b_i επιλέχτηκαν αυθαίρετα αλλά πάντα λαμβάνοντας υπόψη και τηρώντας τον ακόλουθο τύπο:

$$\sum_{i=1}^{3} b_i = 1 \quad (1718)$$

Οι παράμετροι και οι αντίστοιχες τιμές βάσει των οποίων έγιναν τα πειράματα παρουσιάζονται συνοπτικά στον παρακάτω πίνακα.

Πίνακας 1: Παράμετροι πειραμάτων

Παράμετροι	Τιμές
Αλγόριθμοι κατηγοριοποίησης	C4.5, Logistic, Baseline Random Classification Algorithm
Συντελεστές βαρύτητας b_i	$b_i \in \{0 \dots 1\}$ με $\sum_{i=1}^{3} b_i = 1$
a	0.8, 4

5.4 Σενάρια και αξιολόγηση αποτελεσμάτων

Στην παρούσα ενότητα παρουσιάζουμε τα αποτελέσματα της προτεινόμενης τεχνικής βάσει κάποιων σεναρίων εκτέλεσης. Σε κάθε σενάριο υπολογίζονται οι επιμέρους ομοιότητες του νέου χρήστη και των γειτόνων για καθένα από τα δημογραφικά δεδομένα ηλικία, επάγγελμα και φύλο όπως ορίστηκε στην ενότητα **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**. Σε καθένα από αυτούς τους δείκτες ομοιότητας αναθέτουμε συγκεκριμένους συντελεστές βαρύτητας. Οι εν λόγω συντελεστές καθώς και η παράμετρος της εκθετικής συνάρτησης για το δείκτη ομοιότητας της ηλικίας των χρηστών αποτελούν τα διαφορετικά σενάρια που εκτελούνται για την αξιολόγηση του αλγορίθμου. Στις γραφικές παραστάσεις που απεικονίζουν τα αποτελέσματα των μετρικών MAE και RMSE συγκρίνονται οι διαφορετικοί κατηγοριοποιητές που εφαρμόστηκαν. Ο χρήστης μπορεί να ανήκει σε μία από αυτές πολλαπλές κατηγορίες που αντιστοιχούν στις κλάσεις που προέκυψαν από την ομαδοποίηση. Το Multi - C4.5 και Multi - Logistic αφορά τη χρήση των αλγορίθμων C4.5 και Logistic αντίστοιχα ως δυαδικούς κατηγοριοποιητές για την κατηγοριοποίηση πολλαπλών κλάσεων εφαρμόζοντας την One-Against-All στρατηγική.

Σενάριο 1^{ο}

Για το 1ο σενάριο θεωρούμε ότι οι επιμέρους ομοιότητες που σχετίζονται με καθένα από τα δημογραφικά δεδομένα συνεισφέρουν κατά το ίδιο ποσοστό στον υπολογισμό του συνολικού βάρους. Συνεπώς, $b_1 = b_2 = b_3 = 1/3$, όπου b_1 ο συντελεστής βαρύτητας για την

ηλικία, b_2 ο συντελεστής βαρύτητας για το επάγγελμα και b_3 ο συντελεστής βαρύτητας για το φύλο.

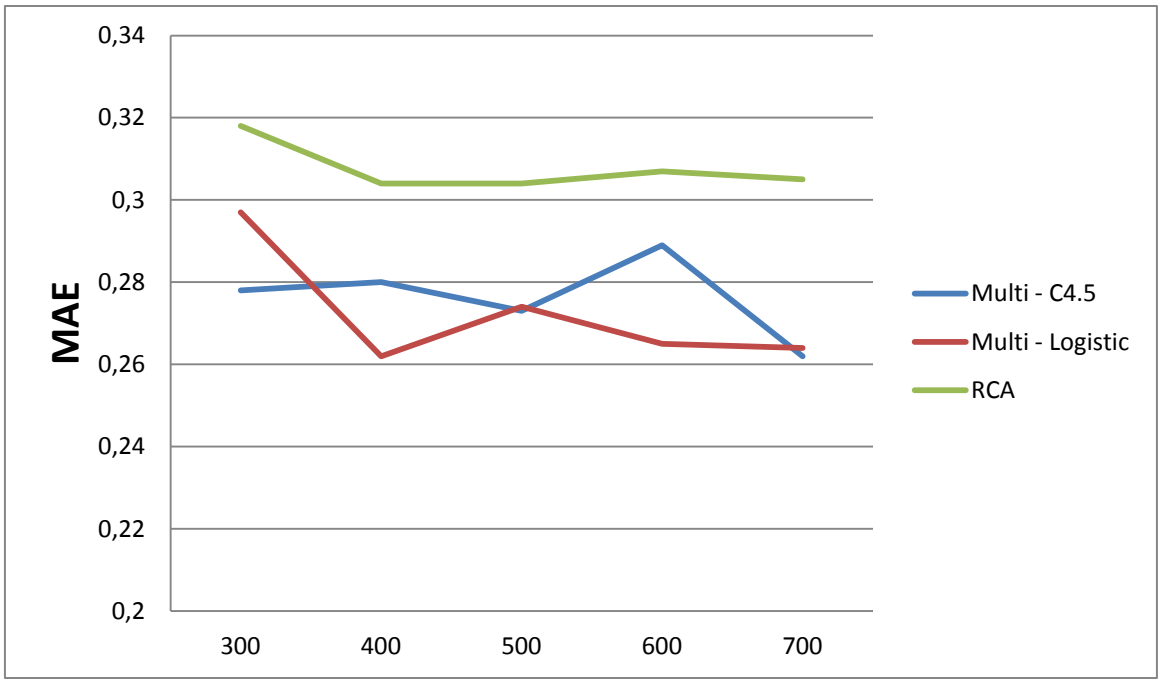
Στη συνέχεια, παρουσιάζουμε τα αποτελέσματα για το συγκεκριμένο σενάριο.

Πίνακας 2: Αποτελέσματα MAE για το 1ο σενάριο ($b_1 = b_2 = b_3 = 1/3$ και $\alpha = 0.8$)

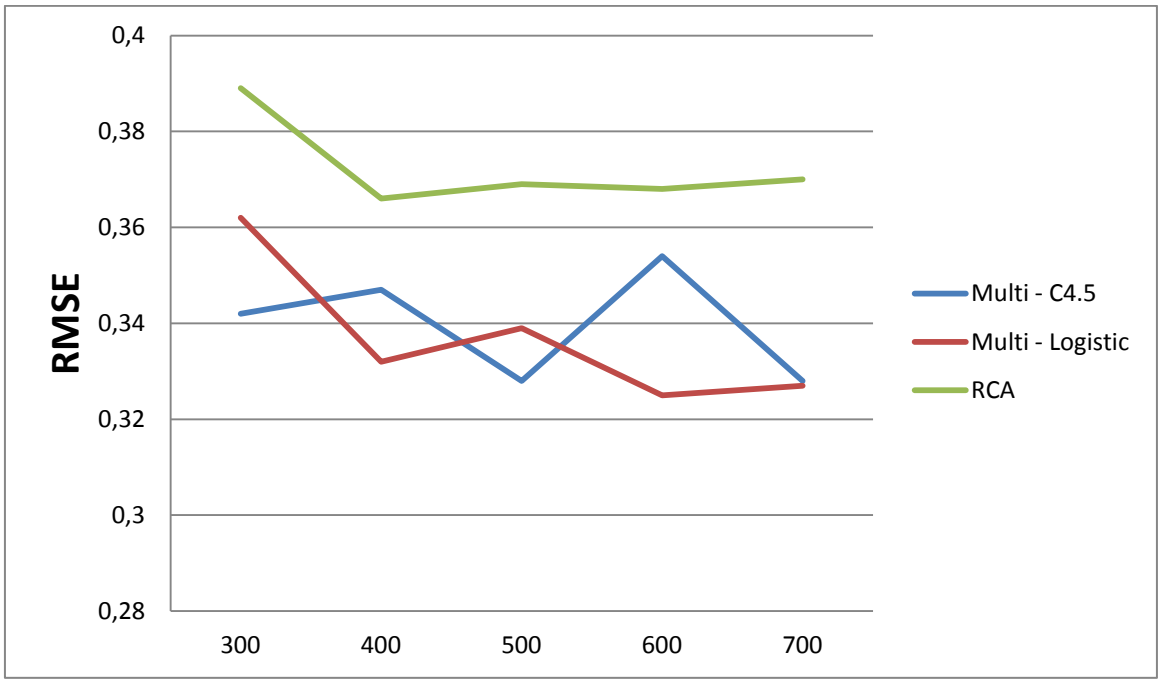
Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.278	0.297	0.318
400	0.280	0.262	0.304
500	0.273	0.274	0.304
600	0.289	0.265	0.307
700	0.262	0.264	0.305

Πίνακας 3: Αποτελέσματα RMSE για το 1ο σενάριο ($b_1 = b_2 = b_3 = 1/3$ και $a = 0.8$)

Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.342	0.362	0.389
400	0.347	0.332	0.366
500	0.328	0.339	0.369
600	0.354	0.325	0.368
700	0.328	0.327	0.370



Σχήμα 1: Αποτελέσματα MAE για το 1^{ο} σενάριο



Σχήμα 2: Αποτελέσματα RMSE για το 1^{ο} σενάριο

Από τα Σχήματα 1 και 2 παρατηρούμε ότι και οι 2 προσεγγίσεις που χρησιμοποιούμε για την κατηγοριοποίηση πολλαπλών κλάσεων έχουν καλύτερα αποτελέσματα από τον αλγόριθμο αναφοράς RCA όπου οι κλάσεις επιλέγονται τυχαία χωρίς το βήμα της κατηγοριοποίησης. Όπως φαίνεται και στον Πίνακα 2, οι τιμές του MAE είναι γενικά χαμηλές άρα και η ακρίβεια των τεχνικών που χρησιμοποιήθηκαν είναι αρκετά

ικανοποιητική και η μέση διαφορά μεταξύ εκτιμώμενης και πραγματικής τιμής για τις προτιμήσεις του χρήστη δεν ξεπερνά το 0,3. Παρατηρείται επίσης ότι η τιμή του MAE ελαττώνεται όσο αυξάνεται ο αριθμός των εγγεγραμμένων χρηστών ακόμα και για την περίπτωση του αλγόριθμου RCA αλλά ακόμα περισσότερο για τις προσεγγίσεις όπου χρησιμοποιούμε το μοντέλο για την κατηγοριοποίηση των νέων χρηστών. Οι προσεγγίσεις Multi – C4.5 και Multi - Logisic παρουσιάζουν συνολικά καλύτερη εικόνα παρά τις αυξομειώσεις που εμφανίζουν κατά τη διάρκεια εκτέλεσης των πειραμάτων με διαφορετικό πλήθος εγγεγραμμένων χρηστών. Η προσέγγιση με τον αλγόριθμο Logistic φαίνεται να έχει καλύτερη απόδοση από την κατηγοριοποίηση πολλαπλών κλάσεων με τον C4.5 ο οποίος παρουσιάζει αστάθεια όσο αυξάνεται το πλήθος των εγγεγραμμένων χρηστών. Εξ' ορισμού το RMSE δίνει σχετικά υψηλό βάρος σε μεγάλα σφάλματα (επειδή αθροίζουμε τα τετράγωνα των σφαλμάτων πριν να πάρουμε το μέσο όρο τους). Παρατηρώντας τον Πίνακα 3 με τα αποτελέσματα του RMSE για το 1^{ο} σενάριο είναι φανερό ότι τα μεγαλύτερα σφάλματα παρουσιάζονται με τον αλγόριθμο RCA όπως είναι αναμενόμενο και επιβεβαιώνουν την καλύτερη απόδοση του Multi – Logistic που εμφανίζεται πιο σταθερός από τον Multi – C4.5. Στα ανωτέρω πειραματικά αποτελέσματα η τιμή του α είναι 0.8. Τα πειράματα επαναλήφτηκαν για α = 4 και τα αποτελέσματα ήταν παρόμοια.

Σενάριο 2^{ο}

Στα επόμενα σενάρια τροποποιούμε την παραμετροποίηση με τέτοιο τρόπο ώστε να δώσουμε μεγαλύτερη βαρύτητα σε κάποιο από τα δημογραφικά δεδομένα κατά τον υπολογισμό της ομοιότητας. Στο συγκεκριμένο σενάριο δίνουμε μεγαλύτερη βαρύτητα στην μετρική OccuSim για το επάγγελμα του χρήστη. Οι τιμές των b_i διαμορφώνονται ως εξής:

$$b_1 = 0.3,\ b_2 = 0.6,\ b_3 = 0.1$$

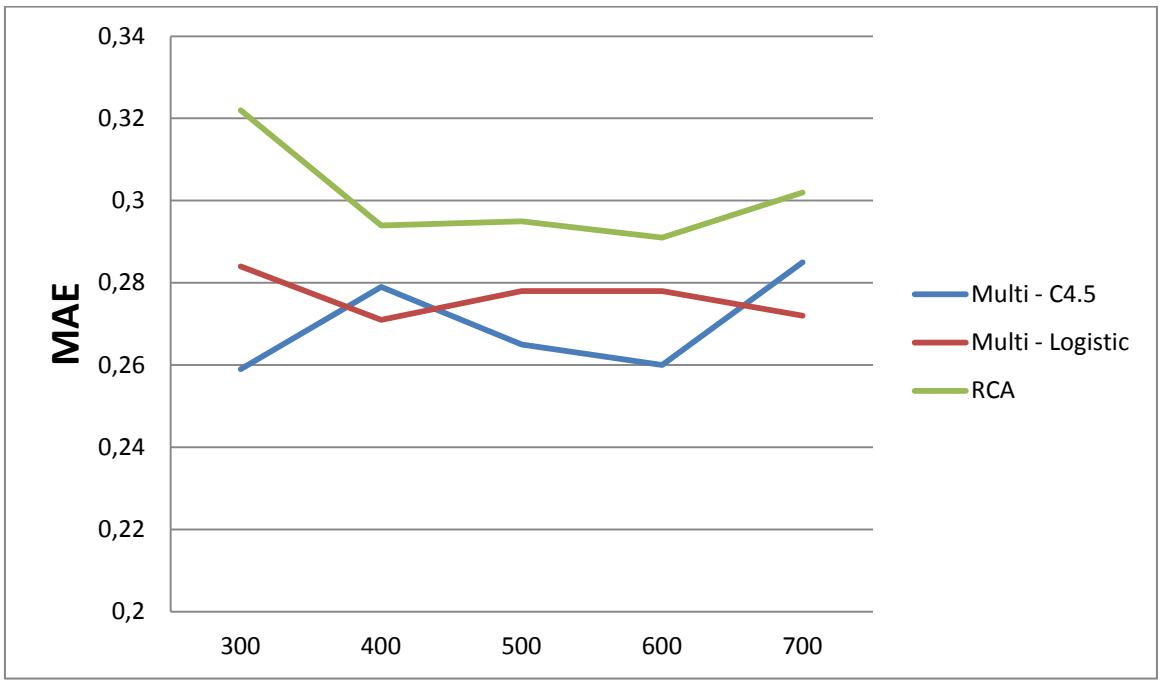
Στη συνέχεια παρουσιάζονται τα αποτελέσματα εκτέλεσης του 2^{ου} σεναρίου για α = 0.8 και τα πλήθη εγγεγραμμένων χρηστών του πρώτου σεναρίου.

Πίνακας 4: Αποτελέσματα MAE για το 2^{ο} σενάριο (α = 0.8)

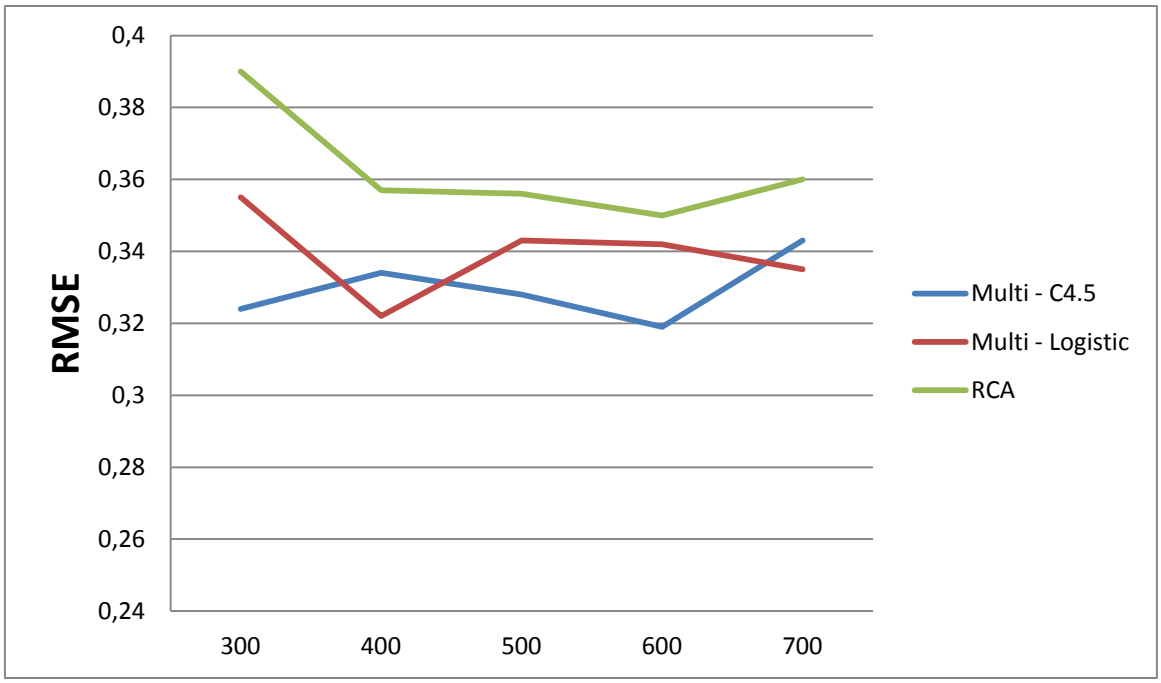
Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.259	0.284	0.322
400	0.279	0.271	0.294
500	0.265	0.278	0.295
600	0.260	0.278	0.291
700	0.285	0.272	0.302

Πίνακας 5: Αποτελέσματα RMSE για το 2ο σενάριο (α = 0.8)

Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.324	0.355	0.390
400	0.334	0.322	0.357
500	0.328	0.343	0.356
600	0.319	0.342	0.350
700	0.343	0.335	0.360



Σχήμα 3: Αποτελέσματα MAE για το 2ο σενάριο



Σχήμα 4: Αποτελέσματα RMSE για το 2ο σενάριο

Παρατηρούμε στα Σχήματα 3 και 4 ότι τα αποτελέσματα είναι παρόμοια με το προηγούμενο σενάριο. Αξίζει να τονίσουμε ότι και με την παραμετροποίηση που επιλέχτηκε στο σενάριο αυτό τα αποτελέσματα είναι ενθαρρυντικά και δείχνουν καλή απόδοση για το προτεινόμενο σύστημα. Συμπερασματικά, οι τιμές των πειραματικών αποτελεσμάτων δεν επηρεάστηκαν σημαντικά από το βάρος που δόθηκε στο επάγγελμα για τον υπολογισμό της ομοιότητας. Επίσης, στο συγκεκριμένο σενάριο ο Multi – Logistic εμφανίζεται πιο σταθερός από τον Multi – C4.5 αλλά ο C4.5 παρουσιάζει μικρότερα σφάλματα όπως φαίνεται στους Πίνακες 4 και 5.

Σενάριο 3ο

Στο συγκεκριμένο σενάριο δίνουμε μεγαλύτερη βαρύτητα στην μετρική GenSim που αφορά το φύλο του χρήστη. Οι τιμές των b_i διαμορφώνονται ως εξής:

$$b_1 = 0.3,\ b_2 = 0.1,\ b_3 = 0.6$$

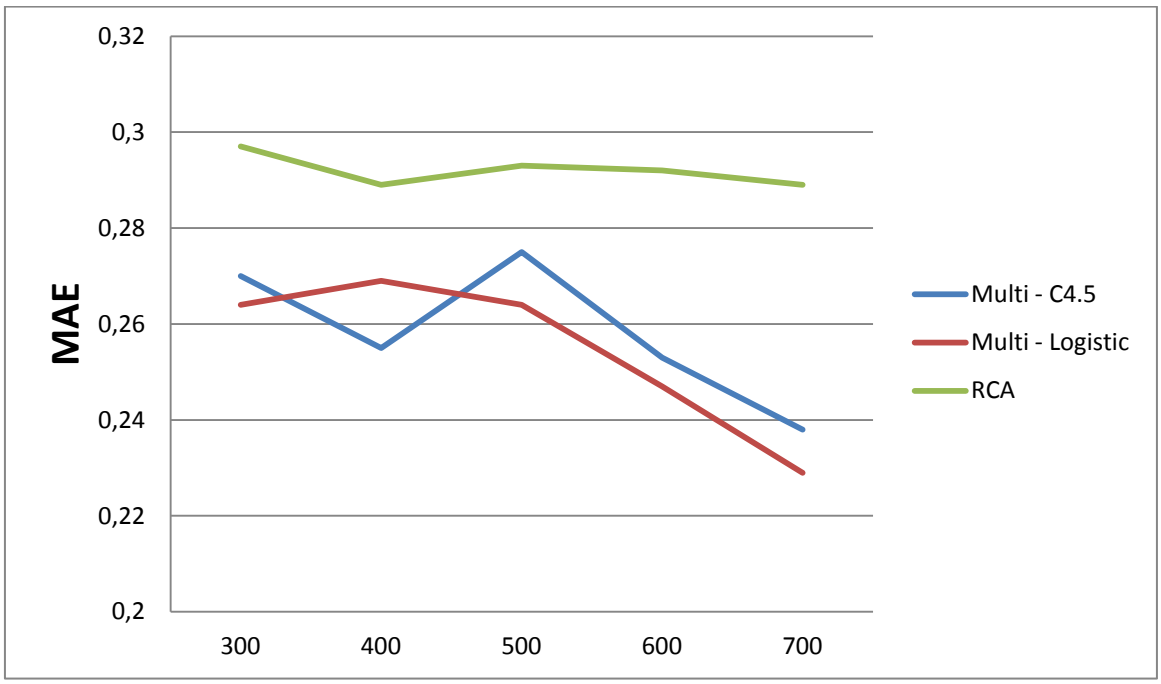
Παρακάτω παρουσιάζονται τα αποτελέσματα για το 3ο σενάριο.

Πίνακας 6: Αποτελέσματα MAE για το 3ο σενάριο (α = 0.8)

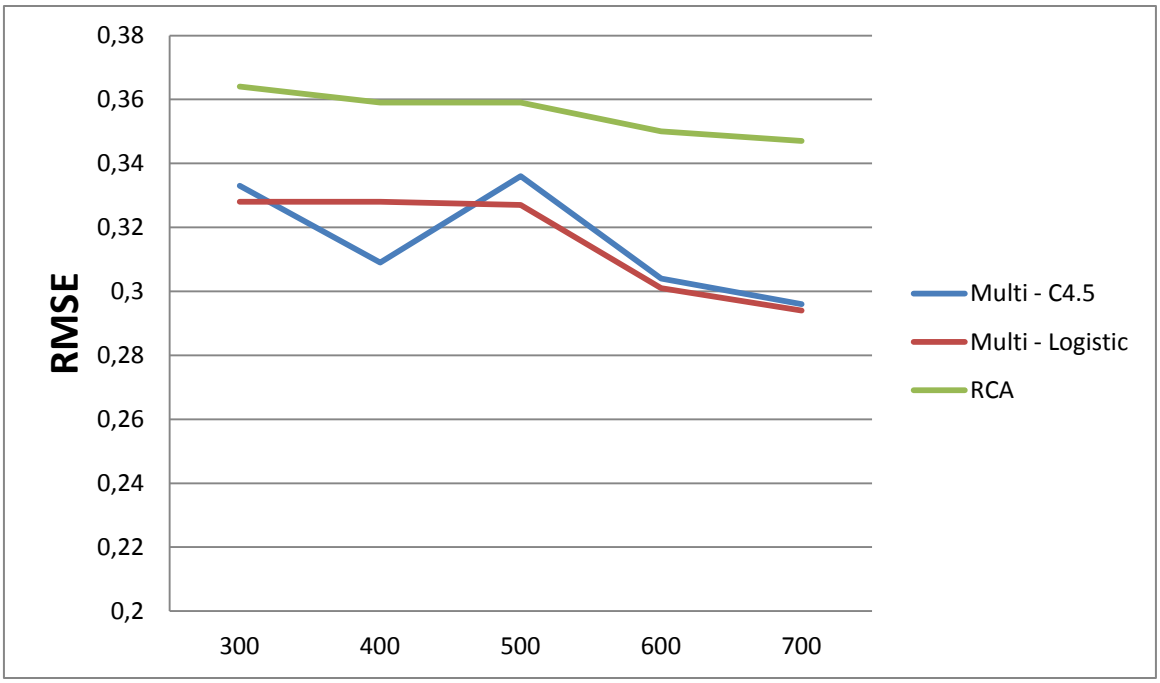
Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.270	0.264	0.297
400	0.255	0.269	0.289
500	0.275	0.264	0.293
600	0.253	0.247	0.292
700	0.238	0.229	0.289

Πίνακας 7: Αποτελέσματα RMSE για το 3^{ο} σενάριο (α = 0.8)

Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.333	0.328	0.364
400	0.309	0.328	0.359
500	0.336	0.327	0.359
600	0.304	0.301	0.350
700	0.296	0.294	0.347



Σχήμα 5: Αποτελέσματα MAE για το 3^{ο} σενάριο



Σχήμα 6: Αποτελέσματα RMSE για το 3ο σενάριο

Από τους πίνακες 6 και 7 παρατηρούμε ότι το συγκεκριμένο σενάριο παρουσιάζει καλύτερα αποτελέσματα για τις μετρικές MAE και RMSE. Αξίζει να σημειωθεί ότι δίνοντας μεγαλύτερη βαρύτητα στο φύλο των χρηστών για τον υπολογισμό της ομοιότητας φαίνεται να έχουμε καλύτερη απόδοση. Συγκεκριμένα εμφανίζεται μικρότερο σφάλμα σε σχέση με τα προηγούμενα αποτελέσματα και για την περίπτωση του Multi – Logistic έχουμε σταθερή μείωσή του σφάλματος αυξάνοντας τον αριθμό των εγγεγραμμένων χρηστών.

Σενάριο 4ο

Στο συγκεκριμένο σενάριο δίνουμε μεγαλύτερη έμφαση στην μετρική AgeSim για να εξετάσουμε την επίδραση που έχει η ηλικιακή ομοιότητα των χρηστών στα πειραματικά αποτελέσματα. Οι τιμές των b_i διαμορφώνονται ως εξής:

$$b_1 = 0.6,\ b_2 = 0.3,\ b_3 = 0.1$$

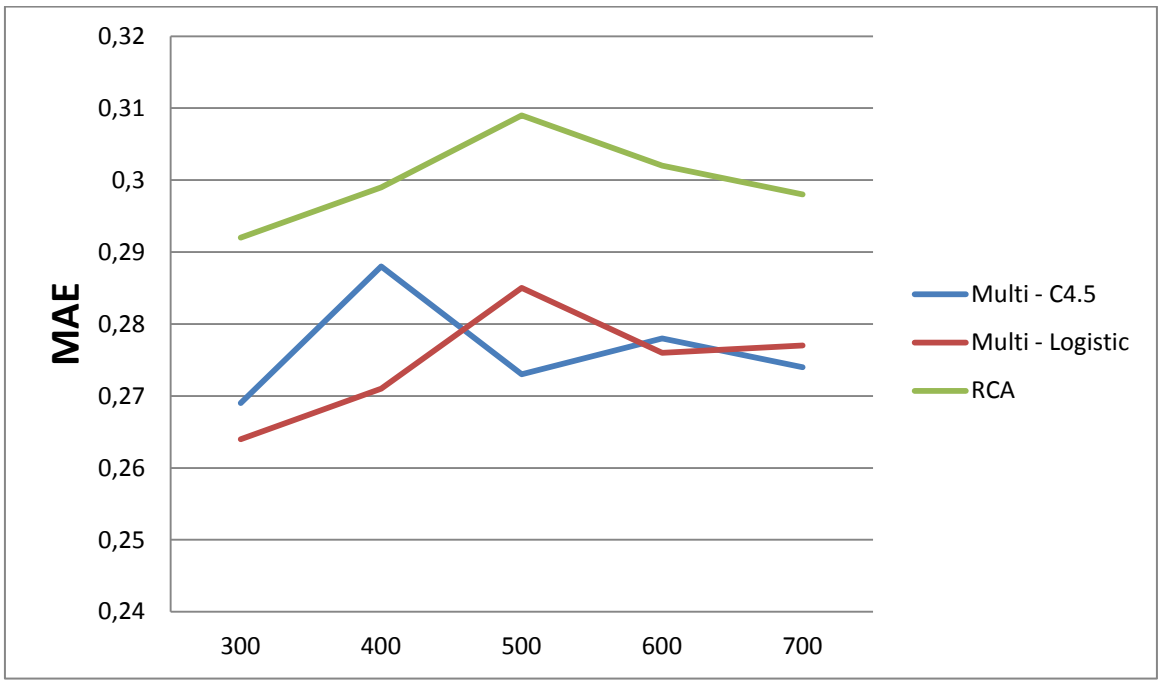
Στη συνέχεια παρουσιάζονται τα αποτελέσματα εκτέλεσης του 4ου σεναρίου για α = 0.8.

Πίνακας 8: Αποτελέσματα MAE για το 4ο σενάριο (α = 0.8)

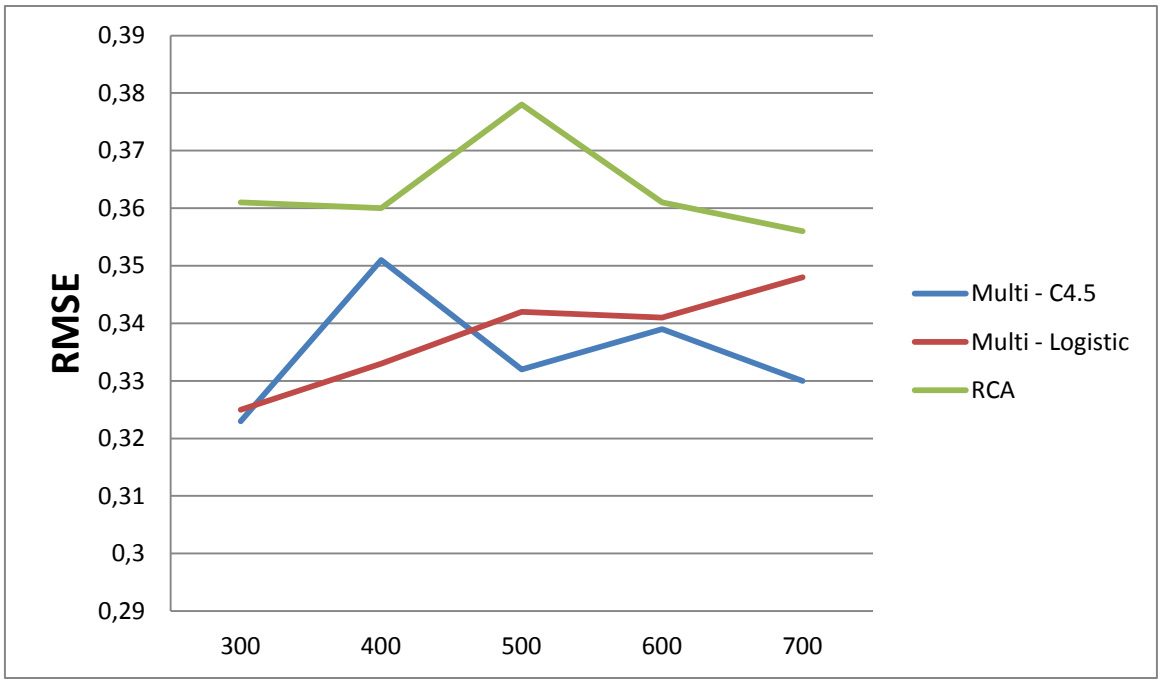
Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.269	0.264	0.292
400	0.288	0.271	0.299
500	0.273	0.285	0.309
600	0.278	0.276	0.302
700	0.274	0.277	0.298

Πίνακας 9: Αποτελέσματα RMSE για το 4^{ο} σενάριο (α = 0.8)

Εγγεγραμμένοι χρήστες #	Multi – C4.5	Multi – Logistic	RCA
300	0.323	0.325	0.361
400	0.351	0.333	0.360
500	0.332	0.342	0.378
600	0.339	0.341	0.361
700	0.330	0.348	0.356



Σχήμα 7: Αποτελέσματα MAE για το 4^{ο} σενάριο (α = 0.8)



Σχήμα 8: Αποτελέσματα RMSE για το 4^{ο} σενάριο (α = 0.8)

Όπως απεικονίζεται στα Σχήματα 7 και 8 τα αποτελέσματα είναι παρόμοια με αυτά των δύο πρώτων σεναρίων. Ενώ παρατηρούνται αυξομειώσεις και με τις δύο τεχνικές Multi – C4.5 και Multi – Logistic η διαφορά του συγκεκριμένου σεναρίου είναι ότι ο Multi – C4.5 φαίνεται να έχει καλύτερη απόδοση όσο αυξάνεται ο αριθμός των χρηστών σε αντίθεση με προηγούμενα σενάρια που η προσέγγιση του Logistic παρουσιάζει καλύτερα αποτελέσματα.

Το βέλτιστο σενάριο είναι το 3^{ο} όπου για τον υπολογισμό της ομοιότητας δίνεται μεγαλύτερη έμφαση στο φύλο. Στο συγκεκριμένο σενάριο η κατηγοριοποίηση εμφανίζει τα χαμηλότερα ποσοστά λαθών και οι εκτιμώμενες προτιμήσεις των χρηστών είναι παραπλήσιες με τις αληθινές τους προτιμήσεις. Επίσης, παρατηρήθηκε ότι η τιμή του α δεν έχει ουσιαστική επίδραση στην απόδοση της προτεινόμενης τεχνικής.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία προτείνουμε μια νέα προσέγγιση για ένα υβριδικό σύστημα συστάσεων που συνδυάζει τεχνικές ομαδοποίησης, κατηγοριοποίησης και συνεργατικού φιλτραρίσματος προκειμένου να παρέχει συστάσεις ανασχηματιζόμενων προϊόντων σε ομάδες χρηστών. Η προτεινόμενη τεχνική αντιμετωπίζει τα προβλήματα των αραιών αξιολογήσεων και της ψυχρής εκκίνησης και παράλληλα βελτιώνει την απόδοση και την ποικιλομορφία των παρεχόμενων συστάσεων. Η προσέγγιση μας βασίζεται στην ιδέα ότι άτομα με τα ίδια χαρακτηριστικά είναι πιθανόν να έχουν ίδιες προτιμήσεις. Για αυτό το λόγο επιχειρούμε να ανακαλύψουμε συσχετίσεις στον πολυδιάστατο χώρο δεδομένων που σχηματίζουν τα δημογραφικά δεδομένα και οι προτιμήσεις των χρηστών με τη χρήση πολυδιάστατης ομαδοποίησης. Τις συγκεκριμένες ομάδες που προκύπτουν τις χρησιμοποιούμε ως κλάσεις για να δημιουργήσουμε το μοντέλο με το οποίο ο κατηγοριοποιητής θα αναθέσει κάθε νέο χρήστη σε μια κατηγορία/κλάση.

Στη συνέχεια, εκτιμούμε την προτίμηση του για κάθε μέρος του ανασχηματιζόμενου προϊόντος βάσει των βαθμολογιών των ήδη εγγεγραμμένων χρηστών που ανήκουν στην ίδια κατηγορία. Η συνεργατική πρόβλεψη γίνεται με βάση μια συνάρτηση ομοιότητας η οποία συνδυάζει τους δείκτες ομοιότητας για καθένα από τα δημογραφικά δεδομένα: ηλικία, επάγγελμα και φύλο. Τέλος αθροίζουμε τις προβλεπόμενες προτιμήσεις για κάθε νέο χρήστη που ανήκει στην ίδια κατηγορία και με βάση την αθροιστική προτίμηση που προκύπτει για κάθε μέρος του προϊόντος παρέχουμε συστάσεις χρησιμοποιώντας το ιστορικό επιλογών του εγγεγραμμένου χρήστη της ίδιας κατηγορίας του οποίου η προτίμηση πλησιάζει περισσότερο την αθροιστική προτίμηση της ομάδας.

Το στάδιο της κατηγοριοποίησης συμβάλλει στην εκτίμηση της προβλεπόμενης προτίμησης μειώνοντας τα σφάλματα και κατά συνέπεια συμβάλλει στην ακρίβεια της προτεινόμενης τεχνικής. Αυτό προκύπτει από την σύγκριση των αλγορίθμων κατηγοριοποίησης πολλαπλών κλάσεων που υλοποιήθηκαν σε σχέση με τον αλγόριθμο αναφοράς RCA ο οποίος επιλέγει τυχαία την κατηγορία που θα αναθέσει το νέο χρήστη. Σημειώνεται ότι για τις περισσότερες περιπτώσεις ο αλγόριθμος κατηγοριοποίησης Logistic είναι πιο αποδοτικός αλλά ο C 4.5 είναι καλύτερος όταν δίνεται περισσότερη βαρύτητα στην ηλικία για τον υπολογισμό ομοιότητας των χρηστών.

6.2 Μελλοντικές προεκτάσεις

Στην παρούσα εργασία παρουσιάσαμε έναν υβριδικό αλγόριθμο που συνδυάζει παραδοσιακές τεχνικές ΣΦ με ομαδοποίηση και κατηγοριοποίηση και επίσης τεχνικές συστάσεων βάσει περίπτωσης (case-based) για την παροχή συστάσεων ανασχηματιζόμενων προϊόντων σε ομάδες χρηστών το οποίο είναι ανοικτό θέμα έρευνας στα σύγχρονα ΣΣ. Μία μελλοντική επέκταση της προσέγγισης είναι η δοκιμή περισσότερων

αλγορίθμων ομαδοποίησης υποχώρου και κατηγοριοποιητών, για την εύρεση του βέλτιστου για την συγκεκριμένη μέθοδο.

Μεγάλο ενδιαφέρον θα είχε η επέκταση αυτών των δημογραφικών δεδομένων με δεδομένα σχετικά με το ανασχηματιζόμενο προϊόν και τους χρήστες του [35] (για παράδειγμα, για την περίπτωση του ταξιδιωτικού πακέτου τέτοια δεδομένα θα μπορούσε να είναι ο χρόνος του έτους που θα γίνει το ταξίδι, η σχέση του χρήστη με τους συνταξιδιώτες του σε περίπτωση που δεν ταξιδεύει μόνος π.χ. οικογένεια, φίλοι κλπ.) για να ερευνήσουμε την επίδρασή τους στη δημιουργία του μοντέλου καθώς και στο τελικό αποτέλεσμα.

Επίσης μια ανοιχτή ερώτηση που αποτελεί αντικείμενο έρευνας στο πλαίσιο των συστάσεων για ομάδες χρηστών και δεν διερευνήθηκε στην παρούσα εργασία είναι το πώς θα επιτευχθεί σύμπνοια στις προτιμήσεις μερών με αντικρουόμενα ενδιαφέροντα μέσα στην ομάδα για το ποίο έχουν προταθεί διάφορες τεχνικές [37].

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Recommender Systems	Συστήματα Συστάσεων
Information Space	Χώρος Πληροφορίας
e-commerce	Ηλεκτρονικό εμπόριο
Feedback	Ανατροφοδότηση
Machine Learning	Μηχανική Μάθηση
Constraint Satisfaction	Ικανοποίηση Περιορισμών
Collaborative Filtering	Συνεργατικό Φιλτράρισμα
Predictive Model	Μοντέλο πρόβλεψης
Heuristic Search	Ευριστική Αναζήτηση
Data Collection	Συλλογή Δεδομένων
User Interaction	Αλληλεπίδραση με το Χρήστη
Clustering	Ομαδοποίηση
Classification	Κατηγοριοποίηση
Data Pattern	Πρότυπα Δεδομένων
Configurable Products	Διαμορφώσιμα / Ανασχηματιζόμενα Προϊόντα
Group - based Configuration	Βασισμένη σε Ομάδες Διαμόρφωση
Subspace Clustering	Ομαδοποίηση για Προβολές του Υποχώρου
Case – based Recommendations	Συστάσεις βάσει περίπτωσης
Utility Function	Συνάρτηση Χρησιμότητας
Configuration	Διαμόρφωση
Collaborative Recommendation	Συνεργατικές Συστάσεις
Content - based Recommendation	Συστάσεις Βάσει Περιεχομένου
Information Retrieval and Filtering	Ανάκτηση και Φιλτράρισμα γνώσης
Demographic - Based Filtering	Φιλτράρισμα Βάσει Δημογραφικών Στοιχειών
Radial Basis Functions	Συναρτήσεις Ακτινικής Βάσης
Hybrid Recommendation Methods	Υβριδικές Μέθοδοι Συστάσεων
Weighted	Σταθμισμένη
Switching	Μεταβατική
Cold start problem	Πρόβλημα Ψυχρής Εκκίνησης
Feature Combination	Συνδυασμός Χαρακτηριστικών
Feature Augmentation	Επαύξηση Χαρακτηριστικών
Data Mining	Εξόρυξη Δεδομένων
Training	Εκπαίδευση
Classifier	Κατηγοριοποιητής
Supervised Learning	Μέθοδος Εποπτευόμενης Μάθησης
MultiClass	Πολλαπλές Κλάσεις
Neighbors	Γείτονες
Weights Estimation	Υπολογισμός Βαρών
Attribute	Γνώρισμα
Meta - level	Μετα-επίπεδο
High - dimensional	Πολυδιάστατος
Cold Start Problem	Πρόβλημα Ψυχρής Εκκίνησης
Over Specialization	Υπερ-εξειδίκευση
Group Decision Making	Ομαδική Λήψη Αποφάσεων

Knowledge-based Configuration	Βασισμένη σε Γνώση Διαμόρφωση
Configuration Result	Αποτέλεσμα Διαμόρφωσης
Feature Recommendation	Πρόταση Χαρακτηριστικών
Process Flow	Ροή Διαδικασίας
Explanation Recommendation	Πρόταση Επεξηγήσεων
Maximal Relaxations	Μέγιστες Χαλαρώσεις Τιμών
Model Based Diagnosis	Διάγνωση Βασισμένη σε Μοντέλο
Conflict Set	Σύνολο Σύγκρουσης
Feature Value Recommendation	Πρόταση Χαρακτηριστικών Τιμών
Subspace Projection	Προβολή του Υποχώρου
Projective Clustering	Προβολική Ομαδοποίηση
Attributes	Γνωρίσματα
Pruning	Κλάδεμα / Βελτιστοποίηση
Cluster Core	Πυρήνας Ομάδας
Similarity Estimation	Υπολογισμός Ομοιότητας
User Sessions	Σύνοδοι Χρηστών
Preprocessing	Προ-επεξεργασία
Multiclass Classifier	Κατηγοριοποιητής Πολλαπλών Κλάσεων
Synset	Ομάδα Συνωνύμων
Baseline	Αλγόριθμος Αναφοράς

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

ΣΦ	Συνεργατικό Φιλτράρισμα
ΦΒΠ	Φιλτράρισμα Βάσει Περιεχομένου
ΣΣ	Συστήματα Συστάσεων
OvA	One Against All
OLAP	Online Analytical Processing
LCS	Least Common Subsumer
OccuSim	Occupation Similarity
AgeSim	Age Similarity
GenSim	Gender Similarity
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
RCA	Random Classification Algorithm
CF	Collaborative Filtering
ARC	Average Rank of the Correct Recommendations
WS4J	Wordnet Similarity For Java

ΑΝΑΦΟΡΕΣ

[1] P. Resnick, and H. R. Varian, Recommender Systems. *Communications of the ACM,* vol. 40(3), 1997, pp. 56–58.

[2] R. Burke, Hybrid Recommender Systems: Survey and Experiments, *User Modeling and User-Adapted Interaction*, vol. 12(4), 2002, pp. 331–370.

[3] G. Adomavicius, and A. Tuzhilin, Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on Knowledge and Data Engineering, vol. 17(6), pp. 734–749.

[4] J. Schafer, D. Frankowski, J. Herlocker, and S. Sen, Collaborative Filtering Recommender Systems. In The Adaptive Web, ed. P. Brusilovsky, A. Kobsa, and W. Nejdl. Lecture Notes in Computer Science 4321. Berlin: Springer, 2007, pp. 291–324.

[5] R. Burke, A. Felfernig, and M. Goeker. Recommender Systems: An Overview, AI Magazine, AAAI, vol. 32(3), 2011, pp. 13-18.

[6] R. Reiter, A Theory of Diagnosis from First Principles, Artificial Intelligence, vol. 23(1), 1987, pp. 57–95.

[7] A. Felfernig, M. Schubert, G. Friedrich, M. Mandl, M. Mairitsch, and E. Teppan, Plausible Repairs for Inconsistent Requirements*, In Proc. of the 21st International Joint Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press, 2009, pp. 791–796.

[8] R. Coester, A. Gustavsson, R. Olsson, and A. Rudstroem, Enhancing Web-Based Configuration with Recommendations and Cluster-Based Help, 2002

[9] A. Haag, and S. Riemann, Product Configuration as Decision Support: The Declarative Paradigm in Practice. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 25(2), 2011, pp. 131–142.

[10] A. Felfernig, M. Mandl, J. Tiihonen, M. Schubert, G. Leitner, Personalized User Interfaces for Product Configuration. *Int'l Conf. on Intelligent User Interfaces (IUI'2010)*, 2010, pp. 317 – 320.

[11] A. Felfernig, R. and Burke, Constraint-Based Recommender Systems: Technologies and Research Issues, *In Proc. of the ACM International Conf. on Electronic Commerce*, New York: Association for Computing Machinery, 2008, pp. 17-26.

[12] T. Petit, C. Bessiere, and J. Regin, A General Conflict-Set Based Framework for Partial Constraint Satisfaction, *In Proc. of the 9th International Conf. on Principles and Practice of Constraint Programming*, Lecture Notes in Computer Science 2833, Berlin: Springer, 2003, pp. 1–14.

[13] B. O'Sullivan, A. Papadopoulos, B. Faltings, P. Pu, Representative Explanations for Over-Constrained Problems. In Proc. of the 22nd AAAI Conf. on Artificial Intelligence, Menlo Park, CA: AAAI Press, 2007, pp. 323–328.

[14] N. Mirzadeh, and F. Ricci, Cooperative Query Rewriting for Decision Making Support. Journal of Applied Artificial Intelligence, vol. 21(10), 2007, pp. 895–932.

[15] C. Thompson, M. Goker, and P. Langley, A Personalized System for Conversational Recommendations. Journal of Artificial Intelligence Research vol. 21, 2004, pp. 393–428.

[16] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, Combining Content-Based and Collaborative Filters in an Online Newspaper. In Proc. of ACM SIGIR Workshop on Recommender Systems, 1999

[17] G. Adomavicius and A. Tuzhilin, Multidimensional recommender systems: a data warehousing approach, *In Proc. of the 2nd Intl. Workshop on Electronic Commerce (WELCOM'01)*. Lecture Notes in Computer Science, vol. 2232, Springer, 2001b.

[18] U. Shardanand, and P. Maes. Social information filtering: Algorithms for automating 'word of mouth'. *In Proc. of the Conf. on Human Factors in Computing Systems*, 1995

[19] L. H. Ungar and D. P. Foster, Clustering Methods for Collaborative Filtering, Papers from 1998 Workshop, AAAI Technical Report WS-98-08, AAAI Press, 1998, pp. 114 – 129.

[20] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications, In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1998, pp.94–105.

[21] L. Parsons, E. Haque, and H. Liu, Subspace clustering for high dimensional data: A review, SIGKDD Explorations Newsletter, vol. 6, no. 1, 2004, pp.90–105.

[22] X. Li and T. Murata, Multidimensional clustering based collaborative filtering approach for diversified recommendation, *In 7th Intl. Conf. on Computer Science and Education (ICCSE 2012)*, July 14-17, 2012 Melbourne, Australia

[23] E. Ntoutsi, K. Stefanidis, K. Nørvåg and HP. Kriegel, Fast group recommendations by applying user clustering, *Intl. Conf. on Conceptual Modeling*, Springer, 2012, pp. 126 – 140.

[24] E. Muller, S. Gunnemann, I. Assent, T. Seidl, Evaluating clustering in subspace projections of high dimensional data, *In Proc. 35th International Conference on Very Large Data Bases* (VLDB 2009), Lyon, France. (2009)

[25] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, In VLDB, 1994, pp. 487-499.

[26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases, In KDD, 1996, pp. 226-231.

[27] K. Kailing, H.-P. Kriegel, and P. Kröger, Density-connected subspace clustering for high-dimensional data, In SDM, 2004, pp. 246-257.

[28] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park, Fast algorithms for projected clustering, In SIGMOD, 1999, pp 61-72.

[29] G. Moise, J. Sander, and M. Ester. P3C: A robust projected clustering algorith, In ICDM, 2006, pp 414-425.

[30] G. Moise and J. Sander. Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering, In KDD, 2008, pp 533-541.

[31] Z. Wu and M. Palmer, Verb semantics and lexical selection, *In ACL '94 Proc. of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp 133 – 138.

[32] G. A. Miller, WordNet: A large lexical database of English; http://wordnet.princeton.edu/ [Προσπελάστηκε στις 28/08/2016]

[33] C. E. A. Procopiuc, A Monte Carlo algorithm for fast projective clustering, In SIGMOD, 2002, pp. 418-427.

[34] M. L. Yiu and N. Mamoulis, Frequent-pattern based iterative projected clustering, In ICDM, 2003, pp. 689-692.

[35] K. Sequeira and M. Zaki, SCHISM: A new approach for interesting subspace mining, In ICDM, 2004, pp. 186-193.

[36] G. Adomavicius , R. Sankaranarayanan , S. Sen , A. Tuzhilin, Incorporating contextual information in recommender systems using a multidimensional approach, ACM Transactions on Information Systems (TOIS), v.23 n.1, January 2005, pp.103-145.

[37] K. McCarthy, L. McGinty, B. Smyth, Case-based group recommendation: compromising for success. In: Weber, R.O., Richter, M.M. (eds.) ICCBR 2007. LNCS (LNAI), vol. 4626, Springer, Heidelberg (2007), pp. 299–313.