ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΚΗ ΤΩΝ ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΔΙΚΤΥΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη Κατηγοριοποίηση Χρηστών ως Μέθοδος Επίλυσης του Προβλήματος Ψυχρής Εκκίνησης

Μπλερίνα Π. Λίκα (Blerina P. Lika)

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

ΑΘΗΝΑ

Ιούλιος 2013

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη Κατηγοριοποίηση Χρηστών ως Μέθοδος Επίλυσης του Προβλήματος Ψυχρής Εκκίνησης

Μπλερίνα Π. Λίκα (Blerina P. Lika)
Α.Μ.: ΜΟΠ10319

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ευστάθιος Χατζιευθυμιάδης,** Αναπληρωτής Καθηγητής ΕΚΠΑ
Κωνσταντίνος Κολομβάτσος, Διδάκτωρ ΕΚΠΑ

ΠΕΡΙΛΗΨΗ

Τα συστήματα συστάσεων παρέχουν εξατομικευμένες συστάσεις που ενδέχεται να ενδιαφέρουν τους χρήστες του Διαδικτύου. Μία δημοφιλής τεχνική που εφαρμόζουν τα περισσότερα συστήματα συστάσεων είναι το συνεργατικό φιλτράρισμα (ΣΦ) που προβλέπει νέα αντικείμενα για τον ενεργό χρήστη με βάση τις προτιμήσεις άλλων χρηστών. Τα ΣΦ συστήματα πραγματοποιούν προβλέψεις καθορίζοντας μία γειτονιά όμοιων χρηστών και συγκρίνουν τις βαθμολογίες του ενεργού χρήστη με τους όμοιούς του για ένα σύνολο αντικειμένων. Ωστόσο, σε πολλές περιπτώσεις οι ΣΦ τεχνικές υποφέρουν από ένα κοινό πρόβλημα που ονομάζεται πρόβλημα ψυχρής εκκίνησης και για αυτό δεν είναι σε θέση να παράγουν ακριβείς προβλέψεις για χρήστες για τους οποίους δεν έχουν καταχωρημένα στοιχεία. Στη παρούσα διπλωματική εργασία, προτείνουμε μία αποτελεσματική τεχνική, η οποία ξεπερνά το μειονέκτημα αυτό. Η προτεινόμενη τεχνική επιλύει τη ψυχρή εκκίνηση για νέους χρήστες εφαρμόζοντας μεθόδους κατηγοριοποίησης σε ένα παραδοσιακό ΣΦ σύστημα. Για τη φάση της κατηγοριοποίησης εκπαιδεύουμε έναν κατηγοριοποιητή για να δημιουργήσουμε ένα μοντέλο βάσει δημογραφικών δεδομένων. Στη συνέχεια, χρησιμοποιούμε το μοντέλο για την εύρεση της γειτονιάς του ενεργού χρήστη και προβλέπουμε τη βαθμολογία για ένα αντικείμενο βάσει των αξιολογήσεων των γειτόνων. Με αυτό το τρόπο βρίσκουμε χρήστες στο σύστημα που πιθανόν οι προτιμήσεις τους να ταιριάζουν με εκείνες του νέου χρήστη. Η αξιολόγηση και τα αποτελέσματα των πειραμάτων αποδεικνύουν την αποτελεσματικότητα και την αποδοτικότητα της μεθοδολογίας μας. Τέλος, κατά την αποτίμηση της τεχνικής παρουσιάζουμε την απόδοση του συστήματος με τη χρήση διαφορετικών κατηγοριοποιητών όπως δέντρα απόφασης και naïve bayes προσεγγίσεις.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Συστήματα Συστάσεων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Συνεργατικό φιλτράρισμα, πρόβλημα ψυχρής εκκίνησης, μέθοδοι κατηγοριοποίησης, δέντρα απόφασης, σημασιολογική ομοιότητα

ABSTRACT

Recommender systems (RS) provide personalized recommendation for items that are of of interest to the web users. Most RSs adopt collaborative filtering (CF) techniques in order to predict items for a new user. These approaches are based on predictions for finding the user's neighborhood and compare neighbors' ratings for a set of items. However, in many cases, CF techniques suffer from a common problem called the cold-start problem. Therefore, they are not able to make accurate predictions. In this thesis, we propose a novel and effective approach that solves the discussed problem. This method is based on the idea that people with the same features will also share the same interests. More specifically, the proposed approach deals with the cold-start problem for new users by incorporating classification methods in a pure CF system. At first, we train a classifier to create a model based on demographic data. We use this model to classify new users in a specific category according to their profile type. In this way we find a neighborhood for an active user and make predictions according to neighbors' ratings. We evaluate our algorithm and present experimental results that reveal the efficiency and the performance of our methodology. Finally, we make performance comparisons using different classifiers such as decision trees and naïve bayes approaches.

SUBJECT AREA: Recommender Systems

KEYWORDS: Collaborative filtering, cold - start problem, classification methods, decision trees, semantic similarity

Στην πολυαγαπημένη μου οικογένεια και στον Herald.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα σε αυτό το σημείο να εκφράσω τις θερμότατες ευχαριστίες μου στον επιβλέποντα της διπλωματικής εργασίας, Αναπληρωτή Καθηγητή κ. Ευστάθιο Χατζηευθυμιάδη για την ευκαιρία που μου έδωσε για την εκπόνηση της συγκεκριμένης εργασίας και για τη συνεργασία μας στην ερευνητική ομάδα Διάχυτου Υπολογισμού (p-comp). Ακόμη, θα ήθελα να ευχαριστήσω θερμά τον Κωνσταντίνο Κολομβάτσο, για τις χρήσιμες οδηγίες και συμβουλές του καθ' όλη τη διάρκεια της εκπόνησης αυτής της εργασίας. Η συνεχής παρακολούθηση της προόδου της διπλωματικής εργασίας, οι εύστοχες επισημάνσεις τους καθώς και η αμεσότητά τους για την επίλυση οποιουδήποτε προβλήματος προέκυπτε συνετέλεσαν στη διαμόρφωση του τελικού αποτελέσματος. Επίσης, θα ήθελα να ευχαριστήσω τους συνεργάτες μου Κάκια, Βαγγέλη, Βασίλη, Γιώργο και Χρήστο για τη στήριξή τους στα πλαίσια της συνεργασίας μας στο p-comp. Τέλος, ευχαριστώ πολύ την οικογένεια μου και ιδιαίτερα τον Herald για την αμέριστη συμπαράσταση και κατανόηση καθ' όλη τη διάρκεια των σπουδών μου.

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, ειδίκευση «Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών στη Διοίκηση και Οικονομική των Τηλεπικοινωνιακών Δικτύων». Στόχος μας, όταν συζητήσαμε το θέμα για πρώτη φορά, ήταν η εξερεύνηση μιας ερευνητικής περιοχής άμεσα συσχετιζόμενης με τα συστήματα συστάσεων και τα προβλήματα που αντιμετωπίζουν. Πιο συγκεκριμένα, σκοπός της παρούσας διπλωματικής εργασίας ήταν η εμβάθυνση στον τομέα των συστημάτων συστάσεων, η μελέτη της εξελιξής τους καθώς και η δημιουργία ενός συστήματος συστάσεων που ξεπερνάει το πρόβλημα της ψυχρής εκκίνησης. Η ενασχόληση με την προσπάθεια επίλυσης ενός προβλήματος το οποίο συναντάται σε καθημερινές εφαρμογές με βοήθησε ιδιαιτέρως στο να ασχοληθώ με ακόμα μεγαλύτερη διάθεση.

1. ΕΙΣΑΓΩΓΗ

Κατά τη διάρκεια των τελευταίων δύο δεκαετιών, το Διαδίκτυο έχει αναδειχθεί ως το κύριο μέσο για online αγορές, κοινωνική δίκτυωση, ηλεκτρονικό ταχυδρομείο κ.α. Η ταχύτητα με την οποία εξαπλώνεται ο όγκος της πληροφορίας μέσω του Διαδικτύου έχουν αναδείξει το φαινόμενο της υπερπληροφόρησης (information overload). Οι χρήστες έχουν μία δυσκολία να αναζητήσουν και να βρουν εύκολα και γρήγορα αυτό που επιθυμούν. Για αυτό έχουν αναπτυχθεί συστήματα τα οποία βοηθούν στην λύση του συγκεκριμένου πρόβληματος. Αυτά τα συστήματα ονομάζονται συστήματα συστάσεων (Recommender Systems) και θεωρούνται επέκταση των παραδοσιακών συστημάτων πληροφόρησης. Ένα σύστημα συστάσεων είναι ένα εργαλείο λογισμικού που επιτρέπει την εύρεση και τη σύσταση οντοτήτων (όπως προϊόντα, υπηρεσίες, πληροφορίες, πρόσωπα) που μπορεί να ενδιαφέρουν ένα χρήστη [1]. Η βασική τους λειτουργία είναι να προβλέπουν βαθμολογίες (ratings) για αντικείμενα (όπως βιβλία, ταινίες, μουσική) καθώς και προτιμήσεις που αφορούν κυρίως φίλους και ομάδες που δραστηριοποιούνται σε εφαρμογές κοινωνικών δικτύων (όπως Facebook, LinkedIn, MySpace, κ.τ.λ.).

Τα συστήματα αυτά είναι ιδιαίτερα δημοφιλή στο ηλεκτρονικό εμπόριο (E-Commerce) όπου έχουν υιοθετηθεί ευρέως για να βοηθήσουν τους πελάτες να εντοπίσουν προϊόντα που τους ταιριάζουν και πιθανόν θα ήθελαν να αγοράσουν. Ένα αποδοτικό σύστημα συστάσεων συμβάλλει στην ικανοποίηση του πελάτη καθώς βελτιώνει την εμπειρία του και αυξάνει την εμπιστοσύνη του προς το σύστημα. Αυτό έχει ως αποτελέσμα την αύξηση των πωλήσεων των ηλεκτρονικών καταστημάτων που τα παρέχουν. Συνεπώς, τα συστήματα συστάσεων έχουν τραβήξει το ενδιαφέρον πολλών εταιριών που διαθέτουν ηλεκτρονικά καταστήματα. Η Amazon.com χρησιμοποιεί τεχνικές βασισμένες στις προηγούμενες δραστηριότητες του χρήστη (πληροφορίες ιστορικού) για να παρέχει εξατομικευμένο περιεχόμενο. Το eBay.com συλλέγει την ανατροφοδότηση του χρήστη για τα προϊόντα που παρέχει και στη συνέχεια τη χρησιμοποιεί για να προτείνει προϊόντα σε χρήστες που έχουν παρόμοια συμπεριφορά. Τέλος, το Netflix.com και το IMDB.com είναι εφαρμογές ψυχαγωγίας που προτείνουν ταινίες. Τα συγκεριμένα

συστήματα χρησιμοποιούν παρόμοια μεθοδολογία με τη διαφορά ότι το πρώτο ζητάει από το χρήστη να βαθμολογήσει ταινίες και στη συνέχεια πραγματοποιεί τις συστάσεις ενώ το δεύτερο προτείνει κατευθείαν στο χρήστη ταινίες αναλόγως με τις τρέχουσες αναζητήσεις του. Κάποιες άλλες γνωστές εφαρμογές του διαδικτύου που χρησιμοποιούν συστήματα συστάσεων παρουσιάζονται στον Πίνακα 1. Πολλές είναι οι ερευνητικές ομάδες που μελετούν την περιοχή των συστημάτων συστάσεων ενσωματώνοντας τεχνικές από διάφορες περιοχές. Οι περιοχές αυτές είναι μεταξύ άλλων οι τεχνητή νοημοσύνη (artificial intelligence), εξόρυξη δεδομένων (data mining), μηχανική μάθηση (machine learning).

Συνήθως, τα συστήματα συστάσεων για να παρέχουν εξατομικευμένες προτάσεις, μαθαίνουν τη συμπεριφορά του χρήστη χρησιμοποιώντας κάποιες μεθόδους. Οι πιο δημοφιλείς είναι το συνεργατικό φιλτράρισμα(ΣΦ) και το φιλτράρισμα βάση περιεχομένου (ΦΒΠ). Οι συγκεκριμένες τεχνικές σύστασης χρησιμοποιούν πληροφορία που προέρχεται από το ιστορικό του ενεργού χρήστη(ΦΒΠ) ή από τις αξιολογήσεις εγγεγραμμένων χρηστών που μοιάζουν με τον ενεργό χρήστη(ΣΦ). Παρόλα αυτά, η ΦΒΠ τεχνική αντιμετωπίζει ένα σημαντικό πρόβλημα: το πρόβλημα της ψυχρής εκκίνησης. Το πρόβλημα αυτό εμφανίζεται όταν ένας νέος χρήστης γραφτεί στο σύστημα και δεν υπάρχει καμία πληροφορία για αυτόν. Το πρόβλημα διατυπώνεται ως εξής: Έστω $N = \{n_1, n_2 \ldots\ldots n_n\}$ ένα σύνολο νέων χρηστών, $U= \{u_1,u_2 \ldots\ldots u_n\}$ ένα σύνολο εγγεγραμμένων χρηστών στο σύστημα, και $I = \{i_1, i_2, \ldots\ldots i_n\}$ ένα σύνολο αντικειμένων. Τότε ορίζουμε μία συνάρτηση χρησιμότητας $f: N \times I \rightarrow R$ η οποία έχοντας ως είσοδο κάθε νέο χρήστη και ένα οποιοδήποτε αντικείμενο θα επιστρέφει μία αξιολόγηση R που δηλώνει το πόσο αρέσει το συγκεκριμένο αντικείμενο στο νέο χρήστη.

Για την επίλυση του παραπάνω προβλήματος προτείνουμε μία τεχνική η οποία ενσωματώνει τη μέθοδο της κατηγοριοποίησης με ένα παραδοσιακό συνεργατικό σύστημα συστάσεων. Πιο συγκεκριμένα, χρησιμοποιούμε τα δημογραφικά δεδομένα των χρηστών για να δημιουργήσουμε ένα μοντέλο που εντάσσει το νέο χρήστη σε μία κατηγορία. Με αυτό το τρόπο βρίσκουμε χρήστες στο σύστημα που πιθανόν οι προτιμήσεις τους να ταιριάζουν με εκείνες του νέου χρήστη. Οι συστάσεις γίνονται βάση

των αξιολογήσεων των όμοιων χρηστών. Αυτό που είναι ιδιαίτερα πρωτοποριακό είναι ότι δεν δίνουμε την ίδια έμφαση σε όλες τις αξιολογήσεις των όμοιων χρηστών που ανήκουν στην κατηγορία που προκύπτει από την κατηγοριοποίηση. Αντιθέτως, λαμβάνουμε υπόψη περισσότερο τις αξιολογήσεις των χρήστών που έχουν μεγαλύτερη ομοιότητα στα δημογραφικά δεδομένα με τον νέο χρήστη.

Η δομή της παρούσας εργασίας είναι ως εξής: Στο δεύτερο κεφάλαιο παρουσιάζουμε μία επισκόπηση των τεχνικών συστάσεων καθώς και τα προβλήματα που αντιμετωπίζουν. Στο τρίτο κεφάλαιο αναλύουμε την έννοια της «Κατηγοριοποίησης» και αναφέρουμε εκτενέστερα τους αλγορίθμους που εφαρμόσαμε στη προτεινόμενη τεχνική. Στο τέταρτο κεφάλαιο, που αποτελεί και πυρήνα της παρούσας εργασίας, παρουσιάζουμε αναλυτικά την προτεινόμενη τεχνική, την αρχιτεκτονική του συστήματος, και τον αλγόριθμο. Στο πέμπτο κεφάλαιο γίνεται μία πειραματική αποτίμηση του συστήματος με βάση συγκεκριμένες μετρικές και σενάρια για να αξιολογήσουμε την αποδοτικότητα του αλγορίθμου. Τέλος, στο έκτο κεφάλαιο παραθέτουμε τα σημαντικότερα συμπεράσματα καθώς και μελλοντικές επεκτάσεις της προτεινόμενης τεχνικής.

Πίνακας 1: Γνωστές εφαρμογές που χρησιμοποιούν τεχνικές συστάσεων

Εφαρμογή	**Κατηγορία**	**Δικτυακός Τόπος**
Youtube	Ψυχαγωγία	http://www.youtube.com/
Pandora Radio	Ραδιόφωνο	http://www.pandora.com
Movielens	Ταινίες	http://www.movielens.org
Last.fm	Μουσική	http://www.last.fm/
Google News	Ειδήσεις	https://news.google.com/
Facebook	Κοινωνικά Δίκτυα	www.facebook.com
Myspace	Κοινωνικά Δίκτυα	http://gr.myspace.com
LinkedIn	Κοινωνικά Δίκτυα	http://www.linkedin.com/
Twitter	Κοινωνικά Δίκτυα	https://twitter.com/
Rotten Tomatoes	Ταινίες	http://www.rottentomatoes.com/
Jinni	Ταινίες	http://www.jinni.com/
Rovi Corporation	Ψηφιακή Ψυχαγωγία	http://www.rovicorp.com/
Foursquare	Κοινωνικά Δίκτυα	https://foursquare.com/
Tripadvisor	Ταξίδια	http://www.tripadvisor.com/

2. ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ

2.1 Τεχνικές συστάσεων

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

2.1.1 Συνεργατικό φιλτράρισμα (Collaborative Filtering)

Η δημοφιλέστερη και πλέον κυρίαρχη μέθοδος συστάσεων είναι η σύσταση συνεργατικού φιλτραρίσματος. Τα συστήματα συνεργατικού φιλτραρίσματος (ΣΦ) στηρίζονται στις γνώμες και τις προτιμήσεις άλλων χρηστών [1, 44]. Οι τεχνικές ΣΦ επικεντρώνονται κυρίως στη συλλογή πληροφοριών του προφίλ και της συμπεριφοράς του χρήστη για την εύρεση ομοιότητας με άλλους χρήστες. Οι προβλέψεις για κάποιο αντικείμενο βασίζονται στις βαθμολογίες - προτιμήσεις όμοιων χρηστών. Δηλαδή, τα συστήματα ΣΦ, για να παράξουν συστάσεις για ένα χρήστη, χρησιμοποιούν πληροφορία από χρήστες που είναι όμοιοι με αυτόν. Τα συστήματα ΣΦ διακρίνονται σε δύο κύριες κατηγορίες ανάλογα με τον τρόπο εύρεσης της γειτονιάς του χρήστη. Η πρώτη βασίζεται στη μνήμη (memory - based) και η δεύτερη βασίζεται σε μοντέλο (model- based).

2.1.1.1 Memory - based προσέγγιση

Η memory - based προσέγγιση χρησιμοποιεί άμεσα τις βαθμολογίες των χρηστών οι οποίες έχουν αποθηκευτεί στο σύστημα. Η συγκεκριμένη τεχνική συνήθως προσεγγίζεται από μία δομή, ένα δισδιάστατο πίνακα, ο οποίος συσχετίζει τους χρήστες, τα αντικείμενα και τις βαθμολογίες. Παρακάτω ορίζουμε έναν User x Item πίνακα με σειρές τα ονόματα των χρηστών και στήλες ταινίες τις οποίες έχουν αξιολογήσει. Στον Πίνακα 2 η ελάχιστη βαθμολογία που μπορεί να δώσει ο χρήστης είναι 1 και η μέγιστη 5.

Πίνακας 2: Παράδειγμα ενός User x Item πίνακα

	The Notebook	Inception	Gladiator	Star Wars	Titanic
Mary	5	3	2	2	4
Roger	1	4	5	5	2
Kate	2	5	4	5	2
Helen	4	4	3	2	5

Οι memory - based αλγόριθμοι μπορούν με τη βοήθεια ενός πίνακα όπως ο Πίνακας 2 να βρουν ομοιότητα ανάμεσα στους χρήστες ή ανάμεσα στα αντικείμενα. Στη βιβλιογραφία συναντάμε δύο τεχνικές, μία που βασίζεται στον χρήστη (user-based) και μία που βασίζεται στο αντικείμενο (item-based). Η user - based προσέγγιση [2, 3] βρίσκει μια γειτονιά του χρήστη με βάση τις αξιολογήσεις των αντικειμένων. Έστω u ένας χρήστης του συστήματος, οι γείτονες του u είναι εκείνοι οι οποίοι έχουν αξιολογήσει με παρόμοιο τρόπο αντικείμενα που έχει αξιολογήσει ο u. Στη συνέχεια, για την πρόβλεψη της βαθμολογίας του u για ένα άγνωστο αντικείμενο, χρησιμοποιούνται οι βαθμολογίες των γειτόνων του για το συγκεκριμένο αντικείμενο. Ορίζουμε μία συνάρτηση ομοιότητας $sim(u,n)$, η οποία υπολογίζει την ομοιότητα του χρήστη u και του γείτονα του n. Η ομοιότητα προκύπτει από τις αξιολογήσεις που έχουν δώσει ο u και n σε ένα σύνολο αντικειμένων. Για την εκτίμηση της βαθμολογίας του χρήστη u στο αντικείμενο i χρησιμοποιούμε τον σταθμισμένο μέσο όρο που ορίζεται ως εξής:

$$r_{u,i} = \frac{\sum_{n \in N} sim(u,n) * r_{n,i}}{\sum_{n \in N} sim(u,n)} \quad (1)$$

όπου N είναι το σύνολο των γειτόνων του u και $r_{n,i}$ η βαθμολογία του n για το αντικείμενο i.

Αντίστοιχα, τα item - based συστήματα [4, 5, 6] βασίζονται στην ομοιότητα των αντικειμένων. Τα συγκεκριμένα συστήματα προβλέπουν την βαθμολογία του u για ένα

αντικείμενο i βάση των βαθμολογιών που είχε δώσει ο u σε παρόμοια αντικείμενα. Για αυτό το σκοπό ορίζουμε μία συνάρτηση ομοιότητας $sim(i, i_0)$, η οποία υπολογίζει την ομοιότητα του αντικειμένου i και του i_0. Δύο αντικείμενα είναι όμοια εφόσον διάφοροι χρήστες του συστήματος έχουν αξιολογήσει τα αντικείμενα αυτά με παρόμοιο τρόπο. Οπότε, για την εκτίμηση της βαθμολογίας του χρήστη u στο αντικείμενο i τροποποιούμε ανάλογα το σταθμισμένο μέσο όρο που ορίζεται ως εξής:

$$r_{u,i} = \frac{\sum_{i_0 \in I} sim(i,i_0) * r_{u,i_0}}{\sum_{i_0 \in I} sim(i,i_0)} \quad (2)$$

όπου I είναι το σύνολο των όμοιων αντικειμένων με το i και r_{u,i_0} η βαθμολογία του u στο αντικείμενο i_0.

Η πιο γνωστή item-based τεχνική είναι της Amazon [5]. Για κάθε αντικείμενο ο αλγόριθμος δημιουργεί έναν πίνακα με παρόμοια αντικείμενα που έχουν ήδη αξιολογήσει οι πελάτες. Η συνάρτηση ομοιότητας υπολογίζει ένα βάρος το οποίο επιτρέπει την επιλογή έμπιστων γειτόνων και παρέχει έναν τρόπο να δώσουμε μεγαλύτερη ή μικρότερη βαρύτητα στους γείτονες οι οποίοι επηρεάζουν την τελική εκτίμηση της βαθμολογίας. Για αυτό το σκοπό, στη βιβλιογραφία έχουν προταθεί διάφορες μετρικές. Παρακάτω περιγράφονται οι σημαντικότερες από αυτές.

Ευκλείδια Απόσταση (Euclidean Distance): Μία απλή μετρική είναι η ευκλείδια απόσταση που υπολογίζει τη τετραγωνική ρίζα του αθροίσματος της διαφοράς τετραγώνου δύο μεταβλητών. Ο μαθηματικός τύπος είναι ο εξής:

$$similarity(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (3)$$

όπου n είναι ο αριθμός συνιστωσών των διανυσμάτων ενώ X_i και Y_i είναι οι τιμές της i-οστής συνιστώσας.

Απόσταση Manhattan (Manhattan Distance): Η συγκεκριμένη μετρική υπολογίζει το άθροισμα της διαφοράς των συνιστωσών των διανυσμάτων. Αυτό υπολογίζεται με βάση την επόμενη εξίσωση:

$$similarity(X,Y) = \sum_{i=1}^{n}|X_i - Y_i| \quad (4)$$

Συντελεστής ομοιότητας Jaccard (Jaccard Similarity Coefficient): Ο συντελεστής ομοιότητας Jaccard βασίζεται στον αριθμό των αντικειμένων που έχουν αξιολογηθεί και από τους δύο χρήστες και όχι στις βαθμολογίες που έχουν δώσει στα αντικείμενα. Ορίζεται ως το πηλίκο της τομής δύο διανυσμάτων προς την ενωσή τους. Η τιμή του συντελεστή είναι 0 οταν οι χρήστες δεν εχουν αξιολογήσει κανένα ίδιο αντικείμενο και 1 όταν οι χρήστες έχουν αξιολογήσει τον ίδιο αριθμό αντικειμένων. Η εξίσωση για τον συντελεστή ομοιότητας Jaccard είναι:

$$similarity(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (5)$$

όπου X και Y είναι τα αντίστοιχα διανύσματα, χρήστες ή αντικείμενα, $X \cap Y$ η τομή τους και $X \cup Y$ η ένωσή τους.

Ομοιότητα Σημιτόνου (Cosine Similarity): Οι όροι χρήστης και αντικείμενο αναπαριστούνται ως διανύσματα σε ένα πολυδιάστατο χώρο. Η συγκεκριμένη μετρική υπολογίζει την ομοιότητά των χρηστών από το συνημίτονο της γωνίας που σχηματίζουν τα δύο διανύσματα. Η ελάχιστη τιμή της μετρικής είναι -1 υποδηλώνοντας την απόκλιση των διανυσμάτων και η μέγιστη 1, υποδηλώνοντας την απόλυτη ταύτιση. Όταν τα διανύσματα είναι κάθετα μεταξύ τους και σχηματίζουν γωνία 90^0 το συνημίτονο είναι 0 υποδηλώνοντας ότι τα διανύσματα είναι ανεξάρτητα. Η εξίσωση που εφαρμόζουμε για τον υπολογισμό της μετρικής αυτής είναι η ακόλουθη:

$$similarity(X,Y) = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}} \quad (6)$$

όπου το · συμβολίζει το εσωτερικό γινόμενο των X και Y και το $\| X \|$ είναι ο κανόνας του φορέα για το διάνυσμα X.

Προσαρμοσμένη Ομοιότητα Συνημιτόνου (Adjusted Cosine Similarity): Είναι παρόμοια μετρική με την ομοιότητα συνημιτόνου με τη διαφορά ότι περιλαμβάνει το μέσο όρο των τιμών της καθεμίας συνιστώσας του διανύσματος. Η εξίσωση που υπολογίζει την προσαρμοσμένη ομοιότητα συνημιτόνου είναι:

$$similarity(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{\imath})(Y_i - \bar{\imath})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{\imath})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{\imath})^2}} \quad (7)$$

όπου $\bar{\imath}$ είναι ο μέσος όρος των τιμών τις i-οστής συνιστώσας και X_i και Y_i είναι οι τιμές της i-οστής συνιστώσας των διανυσμάτων.

Συντελεστής Συσχέτισης Pearson (Pearson's Correlation Coefficient): Ο συντελεστής συσχέτισης Pearson είναι η δημοφιλέστερη μετρική που εφαρμόζεται στα συστήματα συστάσεων. Ο Pearson υπολογίζει τη γραμμική συσχέτιση δύο μεταβλητών-διανυσμάτων. Επιπλέον, προέρχεται από ένα γραμμικό μοντέλο παλινδρόμησης που στηρίζεται σε μία σειρά από παραδοχές όσον αφορά τα δεδομένα. Η σχέση πρέπει να είναι γραμμική, τα λάθη πρέπει να είναι ανεξάρτητα και να έχουν κατανομή πιθανότητας με μέση τιμή 0 και σταθερή διακύμανση. Η ελάχιστη τιμή είναι -1 όταν τα διανύσματα είναι αντίθετα(δεν είναι όμοια), και η μέγιστη 1 όταν τα διανύσματα είναι όμοια. Όταν ο συντελεστής είναι 0 δηλώνει ότι οι τιμές των συνιστωσών των διανυσμάτων είναι γραμμικώς ανεξάρτητες. Η εξίσωση για το Pearson είναι:

$$similarity(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (8)$$

όπου $\overline{X}$ και $\overline{Y}$ είναι οι μέσοι όροι των τιμών των συνιστωσών για τα διανύσματα X και Y. Εάν υποθέσουμε ότι τα X και Y είναι χρήστες τότε σε ένα user-based σύστημα η εξίσωση (9) μετατρέπεται ως εξής:

$$similarity(x, y) = \frac{\sum_{i=1}^{n}(r_{x,i} - \overline{r_x})(r_{y,i} - \overline{r_y})}{\sqrt{\sum_{i=1}^{n}(r_{x,i} - \overline{r_x})^2}\sqrt{\sum_{i=1}^{n}(r_{y,i} - \overline{r_y})^2}} \quad (9)$$

όπου $r_{x,i}$ είναι η βαθμολογία του χρήστη x στο αντικείμενο i, $r_{y,i}$ είναι η βαθμολογία του χρήστη y στο i, $\overline{r_x}$ και $\overline{r_y}$ η μέση τιμή των βαθμολογιών που έχουν δώσει αντίστοιχα οι χρήστες x και y. Το Pearson σε ένα item-based σύστημα τροποποιείται ως εξής:

$$similarity(i, i_o) = \frac{\sum_{u=1}^{U}(r_{u,i} - \overline{r_\imath})(r_{u,i_o} - \overline{r_{\imath_o}})}{\sqrt{\sum_{u=1}^{U}(r_{u,i} - \overline{r_\imath})^2}\sqrt{\sum_{i=1}^{n}(r_{u,i_o} - \overline{r_{\imath_o}})^2}} \quad (10)$$

όπου U είναι είναι ένα σύνολο χρηστών, $r_{u,i}$ είναι η βαθμολογία του χρήστη u στο αντικείμενο i , και r_{u,i_o} είναι η βαθμολογία του χρήστη u στο i_o .

Κάποιες ερευνητικές προσπάθειες παρουσιάζουν διάφορες παραλλαγές των παραπάνω μετρικών όπως ο συνδυασμός τους για τη δημιουργία μίας συνδυαστικής μετρικής. Στο [7] για τη δημιουργία ενός υβριδικού συστήματος συστάσεων συνδυάζονται γραμμικά ο συντελεστής Pearson και η προσαρμοσμένη ομοιότητα συνημιτόνου Αρχικά, χρησιμοποιείται η Pearson για το υπολογισμό της ομοιότητας αντικείμενων από ένα πίνακα με αντικείμενα και βαθμολογίες (item-ratings matrix). Έπειτα, εφαρμόζεται η προσαρμοσμένη ομοιότητα συνημιτόνου για τον υπολογισμό της ομοιότητας από ένα πίνακα με ομάδες και βαθμολογίες (group-ratings matrix). Στο [8] προτείνονται συνδυαστικές σταθμισμένες μετρικές (weighting approaches) για memory – based αλγορίθμους. Οι συγγραφείς συνδυάζουν την ομοιότητα Jaccard με άλλες μετρικές ομοιότητας όπως Pearson, Cosine και Manhattan. Στα πειραματικά τους αποτελέσματα αποδεικνύεται ότι ο συνδυασμός Jaccard - Manhattan είναι ο συνδυασμός που δίνει την καλύτερη εκτίμηση της βαθμολογίας. Τέλος, στο [9] αποδεικνύεται ότι Pearson δίνει καλύτερα αποτελέσματα μεταξύ άλλων μετρικών που περιγράφονται στη μελέτη.

2.1.1.2 Model – based προσέγγιση

Οι αλγόριθμοι που βασίζονται σε μια model-based προσέγγιση χρησιμοποιούν τις αξιολογήσεις των χρηστών ως σύνολο εκπαίδευσης αλγορίθμων μηχανικής μάθησης για τη δημιουργία ενός μοντέλου πρόβλεψης. Στη συνέχεια, το μοντέλο χρησιμοποιείται για την πρόβλεψη της βαθμολογίας για ένα αντικείμενο. Δηλαδή, υπολογίζουμε την αναμενόμενη βαθμολογία σε ένα άγνωστο αντικείμενο δεδομένου της πληροφορίας που έχουμε για το χρήστη. Στη βιβλιογραφία υπάρχουν πολλοί τρόποι για να δημιουργήσουμε ένα μοντέλο. Οι σημαντικότεροι είναι η συσταδοποίηση (clustering), οι πιθανοτικοί αλγόριθμοι (probabilistic algorithms). Η συσταδοποίηση δημιουργεί ομάδες από χρήστες (user-based) με παρόμοιες προτιμήσεις ή ομάδες αντικειμένων (item-based) με παρόμοιο περιεχόμενο. Η δημιουργία των συστάδων βασίζεται στις αξιολογήσεις που έχουν δώσει οι χρήστες στα αντικείμενα. Στα [10] και [11], οι συγγραφείς χρησιμοποιούν την προσέγγιση της σκληρής συσταδοποίησης (“hard-clustering”) για να κατατάξουν το χρήστη σε μία συστάδα και η πρόβλεψη γίνεται

σύμφωνα με τις βαθμολογίες των χρηστών στη συγκεκριμένη συστάδα. Υπάρχουν και πιο ασαφείς (fuzzy) προσεγγίσεις που υπολογίζουν τη πιθανότητα ένας χρήστης να ανήκει σε κάποια κλάση και ύστερα ύστερα εκτιμούν την αξιολόγησή του για ένα αντικείμενο. Οι πιθανοτικοί αλγόριθμοι [11, 12, 13], όπως τα Bayesian δίκτυα [11], δημιουργούν ένα πιθανοτικό μοντέλο για να λύσουν το πρόβλημα του συνεργατικού φιλτραρίσματος.

Οι ερευνητές προτιμούν τη memory - based προσέγγιση έναντι της model-based. Το σημαντικότερο πλεονέκτημα της memory - based είναι η δυνατότητα να υλοποιηθεί απλά προσφέροντας μία ακριβή πρόβλεψη. Επιπλέον, παρέχει μία συνοπτική και διαισθητική αιτιολόγηση για την προβλεπόμενη βαθμολογία. Για παράδειγμα, ο χρήστης μπορεί να δει τη λίστα με τα αντικείμενα των γειτόνων καθώς και τις βαθμολογίες για αυτά. Μία άλλη χρήσιμη ιδιότητα της μεθόδου είναι ότι προσφέρει σταθερότητα. Τα συστήματα που εφαρμόζουν τη συγκεκριμένη τεχνική επηρέαζονται ελάχιστα από την προσθήκη χρηστών, αντικειμένων και αξιολογήσεων. Αυτό είναι ιδιαίτερα θετικό για μεγάλες εμπορικές εφαρμογές. Από την άλλη πλευρά, η model - based προσέγγιση είναι λιγότερο αποτελεσματική αλλά μπορεί να λύσει το πρόβλημα της κλιμάκωσης το οποίο συναντάμε σε ένα παραδοσιακό memory - based σύστημα.

2.1.2 Φιλτράρισμα βασισμένο στο περιεχόμενο (Content - Based Filtering)

Μία άλλη τεχνική συστάσεων είναι το φιλτράρισμα βάση περιεχομένου (ΦΒΠ – Content Based Filtering) [1] στο οποίο το σύστημα συστήνει αντικείμενα παρόμοια με εκείνα που ο χρήστης προτίμησε στο παρελθόν. Τα ΦΒΠ συστήματα προέρχονται από τα επιστημονικά πεδία της Ανάκτηση Πληροφορίας (Information Retrieval) και της Τεχνητής Νοημοσύνης (Artificial Intelligence). Οι βασικές λειτουργίες των ΦΒΠ συστημάτων είναι η ανάλυση του περιεχομένου των αντικείμενων, τα οποία ο χρήστης προτίμησε και αξιολόγησε στο παρελθόν, και η δημιουργία του προφίλ του χρήστη βάσει των χαρακτηριστικών των αντικειμένων. Αρχικά, τα συστήματα αυτά πραγματοποιούν μία ανάλυση του περιεχομένου, δηλαδή αναλύουν τα χαρακτηριστικά από τα οποία αποτελείται ένα αντικείμενο. Στη συνέχεια, τα δεδομένα αυτής της ανάλυσης αποτελούν την είσοδο του αλγορίθμου για την εκμάθηση του προφίλ του

χρήστη. Το τελευταίο στάδιο είναι το ταίριασμα των αντικειμένων με το προφίλ του χρήστη για να αποφασιστεί αν το αντικείμενο είναι υποψήφιο προς σύσταση.

Το προφίλ αντιπροσωπεύει τα ενδιαφέροντα του χρήστη και υιοθετεί κάποια χαρακτηριστικά. Η κύρια τεχνική για τη δημιουργία προφίλ του χρήστη είναι η σχετική ανατροφοδότηση για ένα αντικείμενο. Η ανατροφοδότηση έχει διάφορες μορφές: α) βαθμολογίες (ratings), β) σχόλια υπό τη μορφή κειμένου (text comments), και γ) αρέσω/δεν αρέσω (like/dislike). Για το αντικείμενο ορίζεται εξίσου ένα προφίλ με χαρακτηριστικά. Για παράδειγμα, σε ένα σύστημα που προτείνει ταινίες, ορίζεται για κάθε ταινία ένα προφίλ που περιλαμβάνει τους ηθοποιούς, το σκηνοθέτη, το είδος ταινίας, κ.λπ. Με βάση αυτά τα χαρακτηριστικά δημιουργείται το προφίλ του χρήστη και του προτείνονται ταινίες οι οποίες ταιριάζουν περισσότερο με τα χαρακτηριστικά του. Για να ορίσουμε το προφίλ ενός χρήστη και ενός αντικειμένου χρησιμοποιούμε διανύσματα βαρών [14]. Για ένα χρήστη u ορίζουμε ένα διάνυσμα $w_u = (w_{u1}, \ldots\ldots, w_{un})$, όπου το κάθε βάρος του διανύσματος δηλώνει το βαθμό συσχέτισης, δηλαδή τη σημασία που έχει το αντίστοιχο χαρακτηριστικό για το προφίλ του χρήστη. Παρόμοια, για ένα αντικείμενο i ορίζουμε ένα διάνυσμα $w_i = (w_{i1,} \ldots\ldots, w_{in})$, όπου το κάθε βάρος δηλώνει τη βαρύτητα που έχει το χαρακτηριστικό για το προφίλ του αντικειμένου. Για την πρόβλεψη της βαθμολογίας υπολογίζουμε μία συνάρτηση χρησιμότητας $Utility(u, i)$ που δηλώνει το ενδιάφερον του u για το αντικείμενο i. Μία ευρεστική μέθοδος για αυτό το σκοπό, είναι η μετρική ομοιότητας συνημιτόνου (βλ. εξίσωση (7)). Ο αντίστοιχη εξίσωση για ένα ΦΒΠ είναι ο εξής:

$$Utility(u, i) = \frac{w_u \cdot w_i}{\|w_u\|_2 \times \|w_i\|_2} \tag{11}$$

όπου u είναι ο χρήστης i το αντικείμενο που πρόκειται να προταθεί και w_u είναι το διάνυσμα για το προφίλ του χρήστη και w_i το διάνυσμα για το προφίλ του αντικειμένου. Για παράδειγμα, εάν ο χρήστης u προτιμά περισσότερο ταινίες κωμωδίας τότε θα του προτείνονται ταινίες αυτού του είδους.

Εκτός από τις παραδοσιακές ευρεστικές μεθόδους, τα ΦΒΠ χρησιμοποιούν και άλλες τεχνικές οι οποίες βασίζονται στη δημιουργία μοντέλου όπως Bayesian classifiers[15] [16], δέντρα απόφασης , και τεχνητά νευρωνικά δίκτυα (Artificial Neural Network).

Συστήματα τα οποία εφαρμόζουν ΦΒΠ τεχνική είναι το Pandora [ρεφ], ένα δικτυακό ραδιόφωνο που παίζει τραγούδια ίδια με αυτά που ο χρήστης άκουσε στο παρελθόν. Το Pandora προτείνει τραγούδια ανάλογα με τις ιδιότητες της μουσικής όπως το tempo και η τονικότητα. Το LIBRA [17] συστήνει βιβλία από μία ψηφιακή βιβλιοθήκη εφαρμόζοντας τη μέθοδο κατηγοριοποίησης Naïve Bayes. Το σύστημα συστάσεων του LIBRA χρησιμοποιεί τις περιγραφές των βιβλίων από διάφορες ιστοσελίδες καθώς και από την on-line ψηφιακή βιβλιοθήκη της Amazon.com. Η CiteSeer [18] είναι επιστημονική ψηφιακή βιβλιοθήκη, η οποία υλοποιεί ένα σύστημα συστάσεων για να ανιχνεύσει επιστημονικά άρθρα και βιβλία που σχετίζονται με τα ερευνητικά ενδιαφέροντα του χρήστη. Με αυτό το τρόπο, η CiteSeer επιτρέπει στους χρήστες της να ενημερώνονται για τις νέες δημοσιεύσεις που είναι σχετικές με το προφίλ τους.

2.1.3 Χρήση δημογραφικών δεδομένων (Demographic - Based Filtering)

Σε αυτή την ενότητα μελετάμε τη τεχνική συστάσεων που βασίζεται στα δημογραφικά δεδομένα του προφίλ του χρήστη. Οι συστάσεις μπορεί να γίνουν σύμφωνα με την ηλικία, το φύλο και την χώρα διαμονής των χρηστών. Η ιδέα είναι ότι χρήστες με παρόμοια δημογραφικά δεδομένα είναι πιθανόν να έχουν και παρόμοιες προτιμήσεις. Οπότε, η εκτίμηση της βαθμολογίας του χρήστη u για το i προκύπτει από την βαθμολογία που έχουν δώσει οι χρήστες με ίδια δημογραφικά δεδομένα με τον u. Τα συστήματα συστάσεων που εφαρμόζουν τεχνική βασισμένη στο περιεχόμενο είναι λίγα. Ο λόγος είναι η δυσκολίας συλλογής των δημογραφικών δεδομένων αφού πολλοί χρήστες δεν επιθυμούν να αποκαλύψουν τα δημογραφικά τους δεδομένα και δηλώνουν ψευδή. Το INTRIGUE [19] είναι ένα σύστημα που παρέχει πληροφορίες στους επισκέπτες για την πόλη Τορίνο της Ιταλίας. Το INTRIGUE προτείνει προορισμούς για αξιοθέατα λαμβάνοντας υπόψιν τις προτιμήσεις ετερογενών ομάδων τουριστών (όπως οικογένειες με παιδιά ή ηλικιωμένους). Στο [28] παρουσιάζεται ένας έξυπνος πράκτορας (Intelligent Agent) ο οποίος αλληλεπιδρά με τους χρήστες του διαδικτύου και χρησιμοποιεί τα προφίλ τους για να προτείνει ιστοσελίδες. Η σύσταση των ιστοσελίδων βασίζεται στα δημογραφικά δεδομένα των χρηστών.

2.1.4 Υβριδικές τεχνικές συστάσεων (Hybrid Recommendation Methods)

Υβριδικά συστήματα συστάσεων είναι τα συστήματα που συνδυάζουν δύο ή περισσότερες τεχνικές συστάσεων. Σκοπός της δημιουργίας τους είναι ότι μπορούν να ξεπεράσουν τα προβλήματα που παρουσιάζουν οι υπάρχουσες τεχνικές συστάσεων. Τα προβλήματα της κάθε τεχνικής παρουσιάζονται παρακάτω. Υπάρχουν διάφορες κατηγορίες υβριδικών συστημάτων ανάλογα με τον τρόπο που συνδυάζουν τις τεχνικές συστάσεων [20]. Παρακάτω αναλύονται οι σημαντικότερες:

Σταθμισμένη (Weighted): Η κατηγορία αυτή υπολογίζει μία συνολική εκτίμηση για τη βαθμολογία ενός αντικειμένου συνδυάζοντας σταθμισμένα τα αποτελέσματα από διάφορες τεχνικές συστάσεων.

Μεταβατική (Switching): Εξάρτηση από την τρέχουσα κατάσταση και τις συνθήκες. Υλοποιεί μετάβαση από μία τεχνική στην άλλη ανάλογα με τις συνθήκες, το επίπεδο εμπιστοσύνης και από εξωτερικά κριτήρια. Για παράδειγμα, είναι πιθανό ένα σύστημα να χρησιμοποιεί αρχικά τη ΒΠΦ τεχνική. Αν οι συστάσεις δεν παρουσιάζουν τότε μεταβαίνουν στη ΣΦ τεχνική.

Ανεξάρτητος συνδυασμός (Mixed): Συστάσεις οι οποίες παράγονται από ανεξάρτητες τεχνικές συστάσεων παρουσιάζονται μαζί.

Συνδυασμός χαρακτηριστικών (Feature combination): Δεδομένα-χαρακτηριστικά από τις επιμέρους τεχνικές χρησιμοποιούνται για την υλοποίηση ενός ενιαίου αλγορίθμου συστάσεων. Πληροφορία από το συνεργατικό φιλτράρισμα χρησιμοποιείται επιπρόσθετα στο φιλτράρισμα βασισμένο στο περιεχόμενο.

Μέθοδος καταρράκτης (Cascade): Με τη χρήση μίας τεχνικής μπορεί να επιτευχθεί βελτίωση μίας άλλης. Σε πρώτο στάδιο μία τεχνική συστάσεων παράγει ένα σύνολο υποψήφιων προς σύσταση αντικειμένων, σε δεύτερο στάδιο η εφαρμογή μίας άλλης τεχνικής μειώνει τον αριθμό των στοιχείων της λίστας.

Επαύξηση χαρακτηριστικών (Feature augmentation): Η έξοδος μίας τεχνικής χρησιμοποιείται ως είσοδο σε άλλη τεχνική. Για παράδειγμα, το LIBRA [17], ένα σύστημα με ΦΒΠ τεχνική χρησιμοποιεί τα δεδομένα από την Amazon.com, ένα σύστημα με ΣΦ τεχνική, για τη σύσταση προϊόντων.

Μετα-επίπεδο (Meta-level): Το μοντέλο που παράγεται από μία τεχνική εφαρμόζεται σε μία άλλη τεχνική. Σε αντίθεση με την μέθοδο επαύξησης χαρακτηριστικών που χρησιμοποιεί μόνο τα χαρακτηριστικά, η συγκεκριμένη κατηγορία χρησιμοποιεί όλο το μοντέλο.

Ένα παράδειγμα υβριδικού συστήματος είναι το NewsDude [21], ένα ραδιόφωνο που προτείνει στον χρήστη ειδήσεις ενώ βρίσκεται στο αυτοκίνητό του. Το σύστημα συνδυάζει αποτελέσματα από δύο διαφορετικούς αλγορίθμους ταξινόμησης (Naïve Bayes και k-nearest neighbor algorithm (k-NN)) για τη δημιουργία μίας ενιαίας ΦΒΠ τεχνικής. Το [22] περιγράφει ένα σύστημα που συνδυάζει τεχνικές ΦΒΠ και ΣΦ για να προτείνει πακέτα ταξιδιών στους πελάτες του. Αρχικά, χρησιμοποιείται η ΦΒΠ τεχνική για να βελτιστοποιήσει τις επερωτήσεις του χρήστη και να φιλτράρει πακέτα τα οποία δεν τον ενδιαφέρουν. Εάν το αίτημα του χρήστη παράγει αρκετά ή καθόλου αποτελέσματα τότε το σύστημα προτείνει χρήσιμες αλλαγές των επερωτήσεων για να επιτευχθούν καλύτερα αποτελέσματα. Έπειτα, τα αποτελέσματα της πρώτης μεθόδου ταξινομούνται με βάση τη ΣΦ τεχνική. Με αυτό τον τρόπο, μειώνονται οι επερωτήσεις του χρήστη και η διαδικασία επιλογής του κατάλληλου πακέτου. Το [23] ενισχύει το συνεργατικό φιλτράρισμα χρησιμοποιώντας πληροφορία από ΦΒΠ τεχνική. Σύμφωνα με τα πειράματα, το προτεινόμενο υβριδικό μοντέλο έχει καλύτερα αποτελέσματα από ένα απλό σύστημα ΦΒΠ ή ΣΦ. Παρόμοια, το [7] προτείνει ένα τρόπο για να ενσωματωθεί η πληροφορία από το προφίλ των αντικειμένων (content - based) σε ένα item - based συνεργατικό φιλτράρισμα.

2.2 Τα προβλήματα των τεχνικών συστάσεων

Στην ενότητα αυτή παρουσιάζουμε τα σημαντικότερα προβλήματα που αντιμετωπίζουν οι τεχνικές συστάσεων.

Τα συστήματα συστάσεων αντιμετωπίζουν προβλήματα τα οποία διαφέρουν ανάλογα με την τεχνική που υλοποιούν. Στα ΣΦ συστήματα εμφανίζονται προβλήματα τα οποία οφείλονται στην εξάρτηση που έχουν μεταξύ τους οι χρήστες του συστήματος. Ενώ στα ΣΒΠ τα προβλήματα συνδέονται κυρίως με την ανάλυση του περιεχομένου και τη δημιουργία κατάλληλου προφίλ για τον ενδιαφερόμενο χρήστη. Ένας κοινός τρόπος

επίλυσης αυτών των προβλημάτων είναι ο συνδυασμός των δύο τεχνικών. Τα σημαντικόντερα προβλήματα των συστημάτων συστάσεων είναι τα εξής:

Το πρόβλημα Ψυχρής Εκκίνησης (Cold-Start Problem): Τα συστήματα συστάσεων αντιμετωπίζουν το πρόβλημα ψυχρής εκκίνησης όταν μία οντότητα (χρήστης ή αντικείμενο) είναι νέα, δηλαδή δεν έχει γίνει / υπάρχει κάποια αναφορά από / για αυτή στο σύστημα. Στη βιβλιογραφία αναφέρονται τρεις τύποι του προβλήματος

a) διαχείριση σύστασης για νέους χρήστες (New User Cold-Start Problem),
b) διαχείριση σύστασης για νέα αντικείμενα (New Item Cold-Start Problem)
c) διαχείρισης σύστασης για νέους χρήστες και νέα αντικείμενα (New Item and User Cold-Start Problem).

Το πρόβλημα ψυχρής εκκίνησης για νέο χρήστη δημιουργείται όταν ο χρήστης είναι καινούριος και το σύστημα δεν γνωρίζει τις προτιμήσεις του μιας και δεν έχει αξιολογήσει κάποιο αντικείμενο. Το πρόβλημα αυτό αντιμετωπίζουν εξίσου συστήματα που χρησιμοποιούν ΣΦ και ΦΒΠ τεχνικές. Στα ΦΒΠ συστήματα η απουσία αξιολογήσεων (προτιμήσεων) έχει ως συνέπεια το μοντέλο να αδυνατεί να δημιουργήσει το προφίλ του χρήστη που βασίζεται στα χαρακτηριστικά των αντικειμένων που έχει αξιολογήσει. Παρόμοια, τα ΣΦ συστήματα αδυνατούν να βρουν όμοιους χρήστες για να παράξουν προβλέψεις.

Αντίστοιχα, το πρόβλημα ψυχρής εκκίνησης για ένα νέο αντικείμενο δημιουργείται όταν το αντικείμενο θεωρείται ότι ‘εισέρχεται’ για πρώτη φορά στο σύστημα, δηλαδή δεν έχει αξιολογηθεί από κανένα χρήστη. Το πρόβλημα αυτό αντιμετωπίζουν κυρίως τα απλά ΣΦ συστήματα αφού για τη σύσταση σε έναν χρήστη βασίζονται στις βαθμολογίες που έχουν δοθεί στα αντικείμενα από άλλους χρήστες. Συνεπώς, κανένα νέο αντικείμενο δεν μπορεί να προταθεί εφόσον δεν έχει κάποια αξιολόγηση.

Τέλος, το πρόβλημα ψυχρής εκκίνησης για νέους χρήστες και ταυτόχρονα για αντικείμενα αποτελεί συνδυασμό των παραπάνω. Δηλαδή, παρουσιάζεται όταν στο σύστημα υπάρχουν χρήστες που δεν έχουν αξιολογήσει κάποιο αντικείμενο αλλά και αντικείμενα που δεν έχουν αξιολογηθεί από κάποιον χρήστη.

Αξίζει σε αυτό το σημείο να σημειωθεί ότι πολλές ερευνητικές εργασίες επικεντρώνονται στην επίλυση του παραπάνω κυρίως για τα ΣΦ συστήματα που αντιμετωπίζουν σε μεγάλο βαθμό το πρόβλημα της ψυχρής εκκίνησης. Στο [24], το σύστημα ζητάει από τους νέους χρήστες να βαθμολογήσουν άγνωστα αντικείμενα. Συνεπώς, απαιτείται ένα στάδιο 'συνέντευξης' ώστε το σύστημα να λειτουργήσει ικανοποιητικά. Με βάση αυτές τις αρχικές αξιολογήσεις η ΣΦ τεχνική συστάσεων βρίσκει όμοιους χρήστες και παράγει συστάσεις για το νέο χρήστη. Αυτός ίσως είναι και ο πιο άμεσος τρόπος επίλυσης του προβλήματος. Ωστόσο, η χρήση ενός σταδίου 'συνέντευξης' για νεοεισερχόμενους χρήστες μπορεί να αποδειχθεί μη αποδοτική αφού κάτι τέτοιο είναι πιθανό να μην γίνει αποδεκτό από τους ίδιους τους χρήστες. Παρόμοια, στο [25] χρησιμοποιείται επίσης μία αρχική 'συνέντευξη' για να δημιουργηθεί το προφίλ του χρήστη. Στη 'συνέντευξη' οι χρήστες εκτός από τα δημογραφικά τους στοιχεία δηλώνουν και τις προτιμήσεις τους σε επιλεγμένα αντικείμενα ή κατηγορίες. Στην εκμάθηση του προφίλ προσαρμόζονται δυναμικά οι ερωτήσεις της συνέντευξης σύμφωνα με τις απαντήσεις του χρήστη. Με αυτό τον τρόπο βελτιώνεται αποτελεσματικά και εξελίσσεται το προφίλ του χρήστη. Άλλα συστήματα επιλέγουν μια μέθοδο που στηρίζεται στην εμπιστοσύνη των νέων χρηστών για τους υπάρχοντες χρήστες του συστήματος [27] [28]. Η ιδέα βασίζεται στο ότι οι νέοι χρήστες επιλέγουν αυτούς που εμπιστεύονται και θεωρούν πως οι προτιμήσεις τους είναι αντιπροσωπευτικές των δικών τους. Με αυτόν τον τρόπο, το σύστημα συστάσεων βασίζεται στις αξιολογήσεις των έμπιστων χρηστών (trusted users) για να προτείνει αντικείμενα στους νεοεισερχόμενους.

Μία αποτελεσματική προσέγγιση για την επίλυση του προβλήματος ψυχρής εκκίνησης είναι ο συνδυασμός των ΣΦ και ΦΒΠ. Μία κοινή μέθοδος είναι η αξιοποίηση της πληροφορίας που προέρχεται από το φιλτράρισμα βασισμένο στο περιεχόμενο. Για την εκτίμηση της βαθμολογίας ενσωματώνουμε την πληροφορία αυτή στο ΣΦ. Στο [45] εφαρμόζονται τεχνικές συσταδοποίησης για την διαχείριση των νέων αντικειμένων. Η μέθοδος συνδυάζει υβριδικά τις ΦΒΠ και ΣΦ τεχνικές για να παράξει μία συνδυαστική συνάρτηση ομοιότητας. Αρχικά, το σύστημα χρησιμοποιεί την πληροφορία βάση περιεχομένου για να δημιουργήσει ομάδες(συστάδες) από παρόμοια αντικείμενα. Για κάθε ομάδα παράγεται και μία συνολική βαθμολογία. Η συνδυαστική συνάρτηση ομοιότητας λαμβάνει υπόψη την ομοιότητα που προκύπτει από έναν πίνακα που

περιέχει τις ομάδες και την αντίστοιχη βαθμολογία (group - rating matrix) καθώς και την ομοιότητα που προκύπτει από έναν πίνακα που περιέχει τα αντικείμενα και τις αξιολογήσεις (item - raiting matrix). Όπως αναφέρουν οι συγγραφείς η συγκεκριμένη τεχνική μπορεί να αντιμετωπίσει με επιτυχία και το πρόβλημα της ψυχρής εκκίνησης για νέα αντικείμενα. Στο [37] αναπτύσσεται έναν αμφίδρομο πιθανοτικό μοντέλο (aspect model) για item - based συστήματα. Ουσιαστικά, η τεχνική υλοποιεί ένα υβριδικό μοντέλο συνδυάζοντας το ΦΒΠ με το ΣΦ φιλτράρισμα για να επιλύσει το πρόβλημα ψυχρής εκκίνησης για νέα αντικείμενα. Το aspect model βασίζεται στην πληροφορία που σχετίζεται με τον χρήστη όπως δημογραφική πληροφορία. Αντίστοιχα, στο [28] αναπτύσσεται ένα παρόμοιο μοντέλο για την επίλυση του προβλήματος ψυχρής εκκίνησης για νέους χρήστες. Στο ΣΦ ενσωματώνεται πληροφορία από τα χαρακτηριστικά των χρηστών. Άλλες ερευνητικές εργασίες δημιουργούν μοντέλα παλινδρόμησης για να παρέχουν εξατομικευμένες συστάσεις στη περίπτωση του νέου χρήστη ή του νέου αντικειμένου. Στο [29] προτείνεται ένα μοντέλο παλινδρόμησης που προβλέπει τη βαθμολογία για το ζεύγος χρήστης/αντικείμενο. Το μοντέλο, εκμεταλλεύεται εξίσου τη διαθέσιμη πληροφορία για τους χρήστες και τα αντικείμενα. Οι χρήστες είναι διανύσματα που αποτελούνται από δημογραφικά δεδομένα και τα αντικείμενα είναι διανύσματα που αποτελούνται από χαρακτηριστικά του περιεχομένου. Παρόμοια, στο [30] χρησιμοποιούνται δημογραφικά δεδομένα από το προφίλ των χρηστών και χαρακτηριστικά των αντικειμένων για δημιουργία ενός διγραμμικού μοντέλου παλινδρόμησης. Στο [31] υιοθετούνται κανόνες συσχέτισης (association rules) για να επιλύσει το πρόβλημα της ψυχρής εκκίνησης και να βελτιώσει την απόδοση του συστήματος όταν αντιμετωπίζει το πρόβλημα αυτό. Η τεχνική αυτή αναπτύσσει κανόνες συσχέτισης που βελτιώνουν το προφίλ των χρηστών για να επιλύσει το πρόβλημα της ψυχρής εκκίνησης. Τέλος, στο [32] προτείνονται κοινωνικές επισημάνσεις (social tags) για την επίλυση του προβλήματος ψυχρής εκκίνησης.

Υπερ-Εξειδίκευση (Over-Specialization): Το πρόβλημα αυτό παρουσιάζεται όταν το σύστημα προτείνει στον χρήστη συνεχώς αναμενόμενα αντικείμενα. Αυτό είναι εμφανές στις ΦΒΠ τεχνικές που το σύστημα προτείνει αντικείμενα με βάση το ιστορικό του χρήστη. Το αποτέλεσμα είναι ότι πολλές φορές κάποια αντικείμενα να είναι ίδια με αυτά που έχει ήδη ο χρήστης επιλέξει. Το συγκεκριμένο μειονέκτημα αναδεικνύει την απουσία

νεωτερισμού-καινοτομίας των ΦΒΠ συστημάτων. Για παράδειγμα, εάν ο χρήστης δηλώσει ότι του αρέσει η ταινία με τίτλο “Star Wars”, το οποίο είναι επιστημονικής φαντασίας, τότε το σύστημα θα του προτείνει βιβλία από τη συγκεκριμένη κατηγορία. Αντιθέτως, τα ΣΦ μπορεί να προσφέρουν καινοτόμες επιλογές.

Εξάρτηση του χρήστη: Η εξάρτηση του χρήστη είναι ένα μειονέκτημα των ΣΦ συστημάτων που απαιτούν βαθμολογίες από άλλους χρήστες και τους γείτονες για να κάνουν συστάσεις. Δηλαδή τα συστήματα αυτά δεν στηρίζονται στις επιλογές του χρήστη με αποτέλεσμα ο χρήστης να μην θεωρείται ανεξάρτητος. Όπως γίνεται αντιληπτό, τα ΣΒΠ παρουσιάζουν ανεξαρτησία από το χρήστη αφού για την πρόβλεψη των συστάσεων η τεχνική στηρίζεται αποκλειστικά και μόνο στις αξιολογήσεις του.

Αδιαφάνεια: Το πρόβλημα αυτό παρουσιάζεται όταν το σύστημα δεν μπορεί να δώσει εξηγήσεις για τον τρόπο λειτουργίας του καθώς και αιτιολογία για τη σύσταση του εκάστοτε αντικειμένου. Αδιαφάνεια αντιμετωπίζουν κυρίως τα ΣΦ συστήματα, αφού η μόνη εξήγηση που μπορούν να δώσουν για τη σύσταση ενός αντικειμένου είναι ότι άγνωστοι χρήστες που έχουν παρόμοιες προτιμήσεις επέλεξαν το συγκεκριμένο αντικείμενο. Από την άλλη πλευρά τα ΦΒΠ είναι διαφανή διότι έχουν τη δυνατότητα να παρουσιάσουν περιγραφή ή λίστα χαρακτηριστικών τα οποία συνέλαβαν στη διαδικασία σύστασης.

Περιορισμένη Ανάλυση Περιεχομένου: Τα ΦΒΠ συστήματα έχουν έναν περιορισμό στον αριθμό και στον τύπο των χαρακτηριστικών των αντικειμένων. Για ακριβείς προβλέψεις, οι ΦΒΠ τεχνικές χρειάζονται αρκετή πληροφορία για να διακρίνουν τα αντικείμενα που αρέσουν στον εκάστοτε χρήστη από αυτά που δεν του αρέσουν. Για παράδειγμα, μερικές προσεγγίσεις μπορεί να λαμβάνουν υπόψη μόνο μερικά από τα χαρακτηριστικά του περιεχομένου ενώ απαιτούνται και τα υπόλοιπα για να προσφέρουν μία πιο ακριβή σύσταση.

3. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ(CLASSIFICATION)

3.1 Το πρόβλημα της Κατηγοριοποίησης

Η κατηγοριοποίηση είναι μία μέθοδος Εξόρυξης Δεδομένων (Data Mining) κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών (target category). Η μέθοδος πρέπει να παράξει μία ακριβή πρόβλεψη της κατηγορίας στην οποία ανήκει το κάθε στοιχείο. Για παράδειγμα, με τη βοήθεια της κατηγοριοποίησης μπορούμε να διακρίνουμε εάν ένα e-mail είναι spam ανάλογα με το περιεχόμενο ή την επικεφαλίδα του. Το πρόβλημα της κατηγοροποίησης απαντάται σε μία πληθώρα εφαρμογών. Μερικά παραδείγματα είναι η Μηχανική Όραση (Computer Vision), η Αναγνώριση Προτύπου (Pattern Recognition), η Βιολογία (Biological Classification), η Αναγνώριση Φωνής (Speech Recognition), και η Γεωστατική (Geostatistics). Επίσης, η κατηγοριοποίηση εφαρμόζεται στα αποτελέσματα των μηχανών αναζήτησης, στην ανακάλυψη φαρμάκων, στη στατιστική επεξεργασία της φυσικής γλώσσας καθώς και στην αναγνώριση πιστοληπτικής ικανότητας που αφορά τους πελάτες τραπεζών. Πιο συγκεκριμένα, η κατηγοριοποίηση μπορεί να περιγραφεί ως μία διαδικασία δύο βημάτων:

Α. Εκπαίδευση - Εκμάθηση (Training): Το πρώτο βήμα είναι η εκπαίδευση του μοντέλου μέσω ενός συνόλου δεδομένων (training set). Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο κατηγοριοποίησης προκειμένου να σχηματιστεί το μοντέλο το οποίο ονομάζεται κατηγοριοποιητής (classifier). Διαφορετικοί αλγόριθμοι κατηγοριοποίησης χρησιμοποιούν διαφορετικές τεχνικές για τη δημιουργία του μοντέλου. Στην κατηγοριοποίηση, τα δεδομένα της εκπαίδευσης ανήκουν σε μία κατηγορία γνωστή εκ των προτέρων. Για αυτό το λόγο, η κατηγοριοποίηση είναι μία μέθοδος εποπτευομένης μάθησης (supervised learning).

Β. Αποτίμηση του Μοντέλου: Για την αξιολόγηση του μοντέλου, χρησιμοποιούμε ένα σύνολο δοκιμαστικών δεδομένων (test data) το οποίο είναι διαφορετικό από το σύνολο εκπαίδευσης που χρησιμοποιήθηκε για τη δημιουργία του μοντέλου. Το μοντέλο κατηγοριοποιεί τα δεδομένα. Στη συνέχεια συγκρίνεται η τιμή της πρόβλεψης της κατηγορίας που σχηματίστηκε από τα δοκιμαστικά δεδομένα με την υπάρχουσα τιμή των δεδομένων εκπαίδευσης. Για την αξιολόγηση του μοντέλου χρησιμοποιούμε

διάφορες μετρικές οι οποίες κρίνουν εάν το μοντέλο είναι αποδεκτό για την συγκεκριμένη χρήση. Η απόδοση των κατηγοριοποιητών οφείλεται αποκλειστικά στα χαρακτηριστικά των δεδομένων. Δεν υπάρχει ένας μόνο κατηγοριοποιητής που δουλεύει καλά σε όλα τα προβλήματα κατηγοριοποίησης.

Στη βιβλιογραφία συναντάμε δύο τύπους του προβλήματος κατηγοριοποίησης οι οποίοι αναλύονται παρακάτω:

I. Δυαδική Κατηγοριοποίηση
II. Κατηγοροποίηση Πολλαπλών Κλάσεων

3.1.1 Δυαδική Κατηγοριοποίηση (Binary Classification)

Η πιο απλή μορφή κατηγοριοποίησης είναι η δυαδική κατηγοριοποίηση. Στη δυαδική κατηγοριοποίηση, η κατηγορία στόχος έχει δύο μόνο τιμές. Για παράδειγμα, εάν θέλουμε να προβλέψουμε την πραγματοποίηση ή όχι ενός αθλητικού αγώνα με βάση τα χαρακτηριστικά του καιρού (όπως θερμοκρασία, υγρασία, εάν είναι θυελλώδης ή όχι) θα πρέπει να οριστούν δύο κλάσεις. Οι κλάσεις μπορεί να οριστούν με τα σύμβολα “Θ” ή “Α” όπου το “Θ” δηλώνει ΘΕΤΙΚΟ, δηλαδή ότι ο αγώνας θα πραγματοποιηθεί, ενώ το “Α” δηλώνει ΑΡΝΗΤΙΚΟ δηλαδή ότι ο αγώνας δεν θα πραγματοποιηθεί.

3.1.2 Κατηγοριοποίηση Πολλαπλών Κλάσεων (Mutliclass Classification)

Η κατηγοριοποίηση πολλαπλών κλάσεων είναι πιο πολύπλοκη διαδικασία από την διαδική. Το πρόβλημα παρουσιάζεται όταν θέλουμε να κατατάξουμε δείγματα σε παραπάνω από δύο κλάσεις. Μία απλή ιδέα για την επίλυση του συγκεκριμένου προβλήματος είναι η εφαρμογή δυαδικών κατηγοριοποιητών. Παρακάτω αναλύονται μερικές στρατηγικές που βασίζονται στη συγκεκριμένη ιδέα:

Ένας - εναντίον - όλων (one - against - all (OvA)) [33]**:** Στη φάση εκπαίδευσης, για κάθε προκαθορισμένη κλάση, εκπαιδεύουμε ένα δυαδικό κατηγοριοποιητή. Για Ν προκαθορισμένες κλάσεις κατασκευάζουμε Ν δυαδικούς κατηγοριοποιητές. Για να κατηγοριοποιήσουμε ένα στοιχείο σε μία κλάση, εφαρμόζουμε όλους τους κατηγοριοποιητές και διαλέγουμε εκείνον με τη μεγαλύτερη ακρίβεια πρόβλεψης. Για

παράδειγμα στη περίπτωση του Naïve Bayes επιλέγουμε τον κατηγοριοποιητή με τη μεγαλύτερη πιθανότητα. Η πρόβλεψη υπολογίζεται από την παρακάτω εξίσωση:

$$f(x) = \operatorname{argmax}_i f_i(x) \quad (12)$$

όπου f_i είναι ο δυαδικός ταξινομητής της i-οστής κλάσης.

Όλοι - εναντίον - όλων (all - against - all (AvA)) [33]**:** Σε αντίθεση με την προηγούμενη στρατηγική, στη φάση εκπαίδευσης δημιουργούμε έναν δυαδικό κατηγοριοποιητή για ένα ζεύγος κλάσεων. Οπότε εάν έχουμε N προκαθορισμένες κλάσεις κατασκευάζουμε N·(N-1) δυαδικούς κατηγοριοποιητές. Για την πρόβλεψη λαμβάνουμε υπόψη τον κατηγοριοποιητή που έχει τη μεγαλύτερη πιθανότητα:

$$f(x) = \operatorname{argmax}_i \sum_j f_{ij}(x) \quad (13)$$

όπου f_{ij} είναι ο δυαδικός κατηγοριοποιητής για τις κλάσεις i, j. Η στρατηγική αυτή αναφέρεται στη βιβλιογραφία και ως ένας - εναντίον - έναν (one - against - one(OvO)).

Κώδικας διόρθωσης λαθών [33]**:** Η συγκεκριμένη προσέγγιση προσπαθεί να συνδυάσει κατηγοριοποιητές με τέτοιο τρόπο που να επιτρέπει τη διόρθωση λαθών.

Σε γενική ανάλυση δεν μπορούμε να δώσουμε μία ξεκάθαρη απάντηση για το ποια είναι η καλύτερη στρατηγική. Αυτό εξαρτάται από το εκάστοτε πρόβλημα. Παρόλα αυτά, οι AvA ή OvO είναι απλές στρατηγικές και δουλεύουν καλά με τους περισσότερους αλγορίθμους κατηγοριοποίησης. Η διαφορά των δύο στρατηγικών έγκειται στην υπολογιστική τους ικανότητα. Η AvA στρατηγική απαιτεί $O(N^2)$ κατηγοριοποιητές σε σχέση με το OvO που απαιτεί O(N). Παρόλα αυτά μερικές φορές η AvA έχει καλύτερη απόδοση επειδή κατά μέσο όρο ο κάθε κατηγοριοποιητής. Επιπλέον, η AvA είναι καλύτερη επιλογή εάν ο χρόνος για τη δημιουργία ενός ταξινομητή είναι μία γραμμική συνάρτηση των αντίστοιχων δειγμάτων.

3.2 Αλγόριθμοι Κατηγοριοποίησης

3.2.1 Δέντρα Απόφασης (Decision Trees)

Μία γνωστή μέθοδος κατηγοριοποίησης είναι τα δέντρα απόφασης [34]. Κάθε εσωτερικός κόμβος προσδιορίζει τον έλεγχο των γνωρισμάτων και κάθε κλαδί που

συνδέει τους εσωτερικούς κόμβους με τους απόγονους τους αντιστοιχεί σε μία πιθανή τιμή για το γνώρισμα. Το κάθε φύλλο αντιστοιχεί στη τιμή της κλάσης. Τα δέντρα απόφασης μας βοηθούν στη λήψη απόφασης καθώς και στο να αποφανθούμε εάν ένα δείγμα ανήκει ή όχι σε μία κλάση-κατηγορία (δυαδική κατηγοριοποίηση).

Θα εξήγησουμε τα δέντρα απόφασης βασισμένοι σε ένα απλό παράδειγμα. Έστω ότι θέλουμε να προβλέψουμε τη διεξαξωγή ενός αθλητικού αγώνα ανάλογα με τις καιρικές συνθήκες που επικρατούν σε μία συγκεκριμένη περιοχή. Υποθέτουμε ότι οι κατηγορίες στόχος έχουν δύο τιμές "Α" (ΑΡΝΗΤΙΚΟ) ή "Θ" (ΘΕΤΙΚΟ) όπως αναφέραμε παραπάνω. Τα γνωρίσματα (attributes) του καιρού που λαμβάνουμε υπόψη είναι:

i. Πρόγνωση με τιμές {Ηλιόλουστος, Συννεφιασμένος, Βροχερός}
ii. Υγρασία με τιμές {Υψηλή, Κανονική}
iii. Θερμοκρασία με τιμές {Δροσερός, Ηπιος, Ζεστός}
iv. Θυελλώδης {Αληθές, Ψευδές}

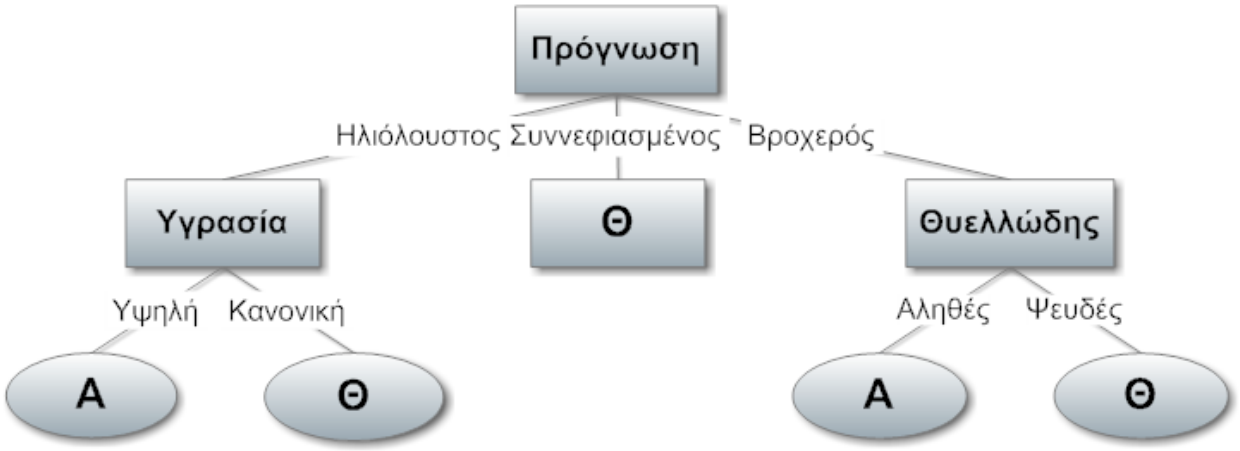
Στον Πίνακα 2 παρουσιάζουμε ένα μικρό σύνολο εκπαίδευσης για τη δημιουργία ενός δέντρου απόφασης:

Πίνακας 2: Παράδειγμα συνόλου εκπαίδευσης για τη διεξαγωγή αθλητικού αγώνα

Δείγμα	**Πρόγνωση**	**Θερμοκρασία**	**Υγρασία**	**Θυελλώδης**	**Κλάση**
1	Ηλιόλουστος	Ζεστός	Υψηλή	Ψευδές	Α
2	Ηλιόλουστος	Ζεστός	Υψηλή	Αληθές	Α
3	Συννεφιασμένος	Ζεστός	Υψηλή	Ψευδές	Θ
4	Βροχερός	Ήπιος	Υψηλή	Ψευδές	Θ
5	Βροχερός	Δροσερός	Κανονική	Ψευδές	Θ
6	Βροχερός	Δροσερός	Κανονική	Αληθές	Α
7	Συννεφιασμένος	Δροσερός	Κανονική	Αληθές	Θ

8	Ηλιόλουστος	Ήπιος	Υψηλή	Ψευδές	Α
9	Ηλιόλουστος	Δροσερός	Κανονική	Ψευδές	Θ
10	Βροχερός	Ήπιος	Κανονική	Ψευδές	Θ
11	Ηλιόλουστος	Ήπιος	Κανονική	Αληθές	Θ
12	Συννεφιασμένος	Ήπιος	Υψηλή	Αληθές	Θ
13	Συννεφιασμένος	Ζεστός	Κανονική	Ψευδές	Θ
14	Βροχερός	Ζεστός	Υψηλή	Αληθές	Α

Παρακάτω παρουσιάζουμε ένα απλό δέντρο απόφασης που παράγεται από το σύνολο δεδομένων του Πίνακα 2.



Εικόνα 1: Δέντρο Απόφασης για τη διεξαγωγή ενός αθλητικού αγώνα

Σύμφωνα με το δέντρο απόφασης της Εικόνας 1, για το κάθε δείγμα λαμβάνουμε υπόψη το γνώρισμα της πρόγνωσης. Όπως απεικονίζεται στην Εικόνα 1, όταν η πρόγνωση του καιρού είναι “Συννεφιασμένος” τότε δεν χρειάζεται να λάβουμε υπόψη κάποιο άλλο γνώρισμα. Το αποτέλεσμα όπως δείχνει το αντίστοιχο φύλλο του δέντρου είναι “Θ” δηλαδή ότι ο αγώνας θα διεξαχθεί. Αντίστοιχα, όταν ο καιρός είναι “Ηλιόλουστος” πρέπει να λάβουμε υπόψη και την υγρασία. Στην περίπτωση που η

υγρασία είναι υψηλή τότε η απόφαση είναι “Α” δηλαδή ότι ο αγώνας δεν θα πραγματοποιηθεί. Εάν η υγρασία είναι “Κανονική” τότε η απόφαση είναι “Θ” δηλαδή ο αγώνας θα πραγματοποιηθεί.

Σε περίπτωση που τα γνωρίσματα είναι επαρκή, είναι δυνατόν να κατασκευαστεί ένα ορθό δέντρο απόφασης που κατατάσσει κάθε δείγμα στη σωστή κλάση. Το νόημα της επαγωγής είναι να δημιουργηθεί ένα δέντρο απόφασης, που δεν κατατάσσει μόνο δείγματα από το σύνολο εκπαίδευσης, αλλά και άγνωστα δείγματα τα οποία πρέπει να κατηγοριοποιηθούν σε μία κλάση. Για να γίνει αυτό, ο εκάστοτε αλγορίθμος πρέπει να εντοπίσει μία σχέση ανάμεσα στα δείγματα και στα χαρακτηριστικά τους. Ο κάθε αλγόριθμος δημιουργεί ένα δέντρο απόφασης το οποίο εξαρτάται από την πολυπλοκότητά του. Στη περίπτωση που για το ίδιο σύνολο δεδομένων έχουν παραχθεί δύο διαφορετικά δέντρα απόφασης (από διαφορετικούς αλγορίθμους) θα επιλέξουμε το πιο απλό δέντρο απόφασης, δηλαδή αυτό με τα λιγότερα μονοπάτια.

Για τη δημιουργία ενός δέντρου απόφασης, συναντάμε στη βιβλιογραφία διάφορες οικογένειες αλγορίθμων. Η πρώτη οικογένεια αλγορίθμων είναι η Hunt's Concept Learning System (CLS) [35]. Οι αλγόριθμοι που ανήκουν στη CLS, δημιουργούν ένα δέντρο απόφασης το οποίο κατηγοριοποιεί ένα δείγμα σύμφωνα με το ελάχιστος κόστος. Οι απόγονοι της CLS είναι η ID3 οικογένεια [34], η οποία έχει κάποια βασικά στοιχεία της CLS οικογένειας. Ένας αλγόριθμος που θεωρείται εξέλιξη - επέκταση της ID3 ομάδας είναι και ο C4.5 ο οποίος θα περιγραφεί παρακάτω. Οι CLS και ID3 απαιτούν το κάθε δείγμα να έχει γνωρίσματα μόνο από ένα συγκεκριμένο σύνολο δεδομένων. Από την ID3 οικογένεια προήλθαν η ACLS και η ASSISTANT οι οποίες έχουν κάποιες σημαντικές διάφορες. Για παράδειγμα για τους ACLS αλγορίθμους δεν είναι ανάγκη για το κάθε δείγμα να υπάρχουν γνωρίσματα μόνο από ένα σύνολο δεδομένων (όπως στο ID3) αλλά για ένα γνώρισμα μπορεί να χρησιμοποιηθούν και ακέραιες τιμές που δεν έχουν προκαθοριστεί.

3.2.1.1 C4.5 Αλγόριθμος

Στην ενότητα αυτή θα μελετήσουμε τον C4.5 αλγόριθμο [36] που ανήκει στην οικογένεια ID3. Ο C4.5 προσφέρει αρκετές βελτιώσεις σε σχέση με τον ID3. Ο συγκεκριμένος αλγόριθμος μπορεί να εφαρμοστεί στην περίπτωση που τα γνωρίσματα ενός δείγματος

έχουν είτε διακριτές ή συνεχείς τιμές. Επιπλέον, ο C4.5 μπορεί να κατηγοριοποιήσει ένα δείγμα όταν κάποιο από τα γνωρίσματα δεν έχει τιμές. Τέλος, άλλη μία δυνατότητα που προσφέρει είναι “κλάδεμα” – βελτιστοποίηση (pruning). Μέσω του “κλαδέματος” μπορεί να ελαχιστοποιήσουμε το δέντρο απόφασης το οποίο παράγεται να αφαιρέσει κάποια κλαδιά τα οποία δεν συντελούν στην εκτίμηση της πρόβλεψης και να τα αντικαταστήσει με φύλλα. Με αυτό τον τρόπο, βρίσκει ένα μικρό δέντρο απόφασης. Παρακάτω παρουσιάζουμε σε ψευδοκώδικα τον C4.5:

Algorithm: C4.5

Input: Instances
Output: Decision for each instance

Begin
    Check for base cases
    for each attribute a
        Find the normalized information gain from splitting on a
    end for
    Let a_best be the attribute with the highest normalized information gain
    Create a decision node that splits on a_best
    Recurse on the sublists obtained by splitting on a_best
        Add those nodes as children of node
End

Για τη διαδικασία παραγωγής του δέντρου απόφασης χρησιμοποιούμε την εντροπία για να υπολογίσουμε το κέρδος του κάθε γνωρίσματος $\bar{y}$ για το σε μία συγκεκριμένη τιμή :

$$Entropy(\bar{y}) = -\sum_{j=1}^{n} \frac{|y_j|}{|\bar{y}|} \log \frac{|y_j|}{|\bar{y}|} \quad (14)$$

όπου $\bar{y}$ είναι το διάνυσμα ενός γνωρίσματος, $Entropy(\bar{y})$ είναι η εντροπία του αντίστοιχου γνωρίσματος. Επίσης, υπολογίζουμε την εντροπία υπό όρους σε περίπτωση που η τιμή του γνωρίσματος είναι η j.

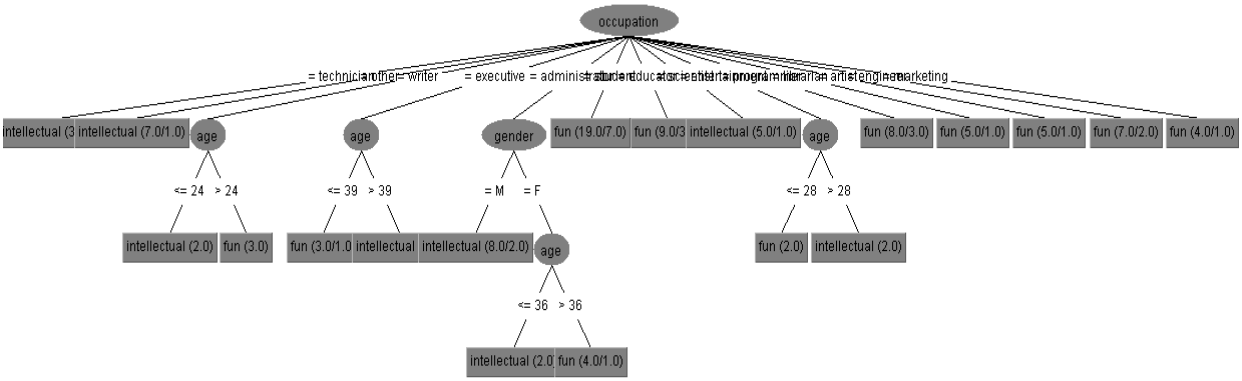
$$Entropy(j|\bar{y}) = \frac{|y_i|}{|\bar{y}|} log \frac{|y_i|}{|\bar{y}|} \quad (15)$$

Τέλος, υπολογίζουμε το κέρδος σύμφωνα με την παρακάτω εξίσωση:

$$Gain(\bar{y}\,, \mathrm{j}) = Entropy(\bar{y} - Entropy(j|\bar{y})\,) \tag{16}$$

όπου $Gain$ είναι το αντίστοιχο κέρδος, $\bar{y}$ είναι το διάνυσμα ενός γνωρίσματος, και $Entropy(\bar{y}\,)$ είναι η εντροπία του αντίστοιχου γνωρίσματος.

Στην Εικόνα 2 παρουσιάζεται ένα δέντρο απόφασης C4.5 το οποίο παράγεται από την τεχνική την οποία προτείνουμε και που θα παρουσιαστει αναλυτικά στο τέταρτο κεφάλαιο.



Εικόνα 2: Δέντρο απόφασης C4.5 δυαδικής κατηγοριοποίησης

3.2.2 Naïve Bayes

Ο Naïve Bayes [37] είναι ένας πιθανοτικός κατηγοριοποιητής ο οποίος βασίζεται στο θεώρημα του Bayes [38]. Ο αλγόριθμος αύτος προβλέπει τη πιθανότητα ένα δείγμα X να ανήκει σε μία κατηγορία.

Ο συγκεκριμένος αλγόριθμος θεωρεί ότι οι τιμές των γνωρισμάτων ενός δείγματος είναι ανεξάρτητες από τις τιμές των υπολοίπων γνωρισμάτων. Έστω ότι θέλουμε να

προβλέψουμε εάν κάποιο δείγμα είναι άνδρας ή γυναίκα δεδομένου κάποιων χαρακτηριστικών όπως το βάρος, το ύψος, και το μέγεθος του παπουτσιού. Ο Naïve Bayes υπολογίζει την εκ των υστέρων (a posteriori) πιθανότητα και για τις δύο κλάσεις. Έστω p(man | X) η εκ των υστέρων πιθανότητα το δείγμα X να ανήκει στη κλάση man και p(woman | X) το δείγμα X να ανήκει στη κλάση woman. Ο Naïve Bayes υπολογίζει και τις δύο πιθανότητες και κατατάσσει το χρήστη στην κλάση με τη μεγαλύτερη εκ των υστέρων πιθανότητα. Δηλαδή εάν p(man | X) μεγαλύτερη από p(woman | X) τότε το X θα είναι άντρας.

Έστω ένα δείγμα το οποίο ορίζεται ως ένα διάνυσμα $X = (x_1, x_2, \ldots, x_n)$ όπου x_i είναι η τιμή ενός χαρακτηριστικού και c μία προκαθορισμένη κλάση στην οποία υποθέτουμε ότι ανήκει το δείγμα. Τότε η πιθανότητα το δείγμα να ανήκει στη κλάση c υπολογίζεται από το θεώρημα του Bayes:

$$p(c|X) = \frac{p(X|c)p(c)}{p(X)} \tag{17}$$

όπου p(c | X) είναι η εκ των υστέρων πιθανότητα το δείγμα X να ανήκει στη κλάση c, p(c) η πιθανότητα της συγκεκριμένης κλάσης (class prior), p(X | c) είναι η κατανομή πιθανότητας των χαρακτηριστικών και p(X) είναι η πιθανότητα που εξαρτάται αποκλειστικά από τα χαρακτηριστικά.

Εάν υποθέσουμε ότι όλα τα χαρακτηριστικά είναι ανεξάρτητα από την τιμή της κλάσης τότε p(X | c) υπολογίζεται ως εξή:

$$p(X|c) = p(x_1, x_2, \ldots\ldots\ldots x_n|c) = \prod_{i=1}^{n} p(x_i|c) \tag{18}$$

Όπου n το σύνολο των χαρακτηριστικών και $p(x_i | c)$ η πιθανότητα του x_i δεδομένου ότι το στιγμιότυπο ανήκει στη κλάση c. Οπότε η εξίσωση (15) μετατρέπεται ως εξής:

$$p(x_1, x_2, \ldots\ldots\ldots x_n|c) = \frac{\prod_{i=1}^{n} p(x_i|c)p(c)}{p(x_1, x_2, \ldots\ldots\ldots x_n)} \tag{19}$$

Η εξίσωση (17) είναι το πιθανοτικό μοντέλο το οποίο δημιουργείται στη φάση εκπαίδευσης. Το δείγμα κατηγοριοποιείται στη κλάση που αναλογεί η μεγαλύτερη εκ των υστέρων πιθανότητα δηλαδή η συνάρτηση πρόβλεψης μπορεί να οριστεί ως εξής:

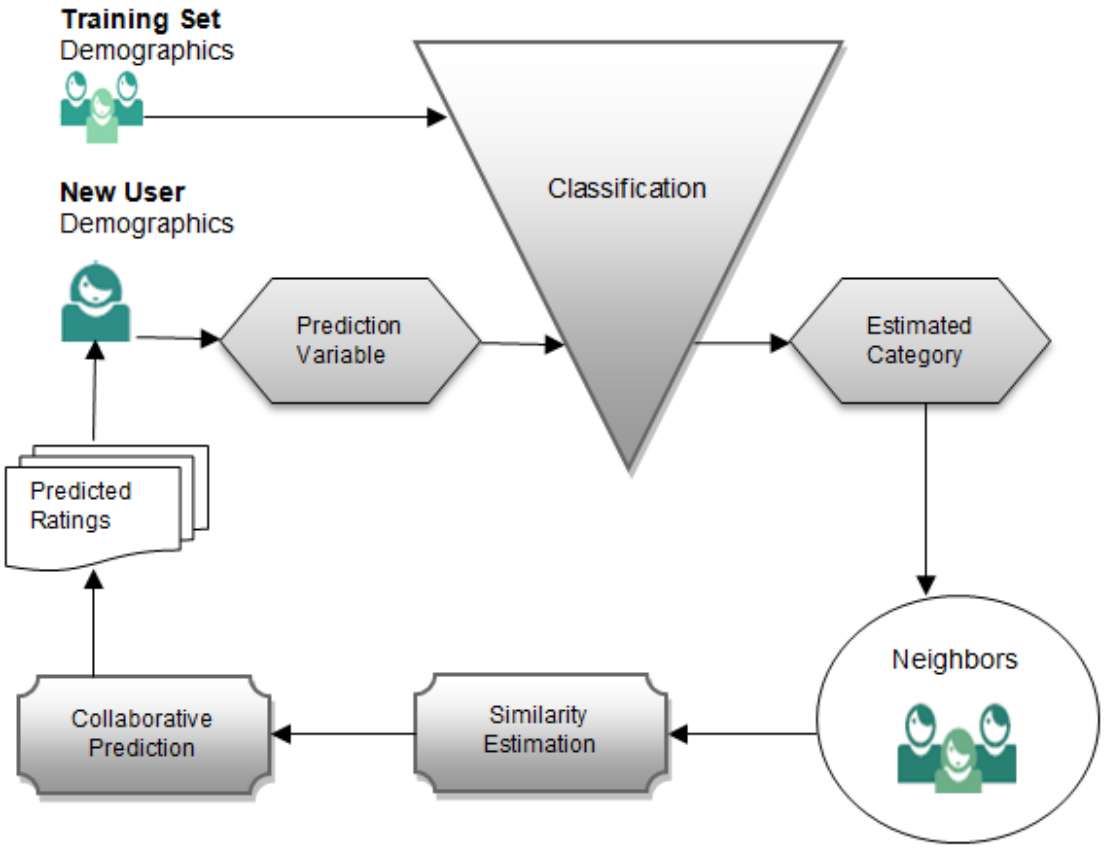
$$classify(X_1, X_2, \ldots\ldots\ldots, X_n) = \operatorname{argmax}_c p(C = c) \prod_{i=1}^{n} p(X_i|c) \tag{20}$$

Πριν περάσουμε στην ανάλυση της προτεινόμενης τεχνικής θα προσπαθήσουμε να εξηγήσουμε τους λόγους για τους οποίους επιλέξαμε τους παραπάνω αλγορίθμους για να την κατηγοριοποίηση του νέου χρήστη. Τα δέντρα απόφασης και συγκεκριμένα ο αλγόριθμος C4.5 παρουσιάζουν αρκετά πλεονεκτήματα. Ο C4.5 δίνει καλά αποτελέσματα δηλαδή κατηγοριοποιεί το δείγμα σε κάποια συγκεκριμένη κλάση ακόμη και όταν τα γνωρίσματα είναι λίγα και κάποια εξ' αυτών οι τιμές απουσιάζουν. Επίσης, δημιουργεί ένα απλό και κατανοητό δέντρο του οποίου η κάθε διαδρομή μπορεί εύκολα να ερμηνευτεί. Από την άλλη πλευρά, ο Naïve Bayes ανήκει σε μία άλλη κατηγορία αλγορίθμων κατηγοριοποίησης και λαμβάνει υπόψη ότι τα γνωρίσματα δεν είναι ανεξάρτητα. Οπότε είναι αρκετά ενδιαφέρον να δούμε τα αποτελέσματα που έχουν και οι δύο αλγόριθμοι όταν εφαρμόζονται στην προτεινόμενη τεχνική για την εύρεση της γειτονιάς του νέου χρήστη.

4. ΠΡΟΤΕΙΝΟΜΕΝΗ ΤΕΧΝΙΚΗ

4.1 Περιγραφή τεχνικής

Στο συγκεκριμένο κεφάλαιο θα περιγράψουμε την τεχνική που προτείνουμε για την επίλυση του προβλήματος ψυχρής εκκίνησης στα συστήματα που εφαρμόζουν τη μέθοδο ΣΦ. Ειδικότερα, προσεγγίζουμε το πρόβλημα από την πλευρά του νέου χρήστη για τον οποίο δεν υπάρχει προηγούμενη βαθμολογία για κανένα αντικείμενο του συστήματος. Η διαδικασία πρόβλεψης αποτελείται από τρία στάδια: 1) δημιουργία ενός μοντέλου το οποίο βασίζεται στα δημογραφικά δεδομένα των εγγεγραμμένων χρηστών του συστήματος. Στο συγκεκριμένο στάδιο εφαρμόζουμε το μοντέλο για την εύρεση γειτονιάς του νέου χρήστη 2) υπολογισμός δεικτών ομοιότητας του νέου χρήστη και των γειτόνων βάση σταθμισμένου μέσου όρου που λαμβάνει υπόψη τα δημογραφικά δεδομένα των χρηστών και 3) συνεργατική πρόβλεψη η οποία συνδυάζει τα βάρη ομοιότητας και τις αξιολογήσεις των γειτόνων. Στην Εικόνα 3 απεικονίζουμε τις σημαντικότερες πτυχές της προτεινόμενης τεχνικής οι οποίες αναλύονται παρακάτω:



Εικόνα 3: Διάγραμμα ροής της προτεινόμενης τεχνικής

Μεταβλητή Πρόβλεψης (Prediction Variable): Το κάθε δείγμα αποτελείται από διάφορα χαρακτηριστικά. Αρχικά είναι σημαντικό να εντοπίσουμε τα κριτήρια βάση των οποίων θα γίνει η ταξινόμηση των χρηστών. Ο εντοπισμός των κριτηρίων γίνεται με βάση τα γνωρίσματα που χρησιμοποίηθηκαν ως κριτήρια κατά τη διαδικασία της εκπαίδευσης του συνόλου δεδομένων για την παραγωγή του μοντέλου. Στη δική μας περίπτωση, τα κριτήρια είναι τα δημογραφικά δεδομένα του χρήστη. Στο συγκεκριμένο στάδιο αποφασίζουμε ποια θα είναι η προβλεπόμενη μεταβλητή δηλαδή βάση ποιου χαρακτηριστικού θα ταξινομήσουμε τον νέο χρήστη.

Για παράδειγμα εάν ένα σύνολο δεδομένων έχουν γνωρίσματα όπως ύψος, βάρος, αριθμός παπουτσιού, φύλο (άντρας ή γυναίκα), τότε έχουμε την κατάλληλη πληροφορία και την επιλογή για να ορισούμε ως μεταβλητή πρόβλεψης το φύλο.

Κατηγοριοποίηση (Classification): Το συγκεκριμένο μέρος παράγει το μοντέλο του κατηγοριοποιητή. Για τη δημιουργία του μοντέλου απαιτείται αρχικά να καθοριστεί η μεταβλητή πρόβλεψης και να πραγματοποιηθεί η φάση εκπαίδευσης του συνόλου δεδομένων. Η διαδικασία της εκπαίδευσης έχει ως αποτέλεσμα την παραγωγή του μοντέλου. Τα βήματα αυτά παρουσιάζονται στην Εικόνα 4 και αναφέρονται αναλυτικότερα παρακάτω. Στη δική μας περίπτωση το σύνολο εκπαίδευσης είναι τα δημογραφικά στοιχεία των χρηστών που είναι ήδη εγγεγραμμένοι στο σύστημα βάσει των οποίων δημιουργείται το μοντέλο. Κατά την κατηγοριοποίηση του νέου χρήστη η είσοδος του μοντέλου είναι τα δημογραφικά δεδομένα του νέου χρήστη και η έξοδος είναι η εκτιμώμενη κατηγορία.

Εκτιμώμενη κατηγορία (Estimated Category): Είναι η έξοδος του μοντέλου δηλαδή η κατηγορία στην οποία ανήκει ο νέος χρήστης. Σύμφωνα με το αμέσως προηγούμενο παράδειγμα αφού λάβαμε υπόψη το φύλο ως μεταβλητή πρόβλεψης τότε η εκτιμώμενη κατηγορία είναι άντρας ή γυναίκα.

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 βάσει των αξιολογήσεων που έχουν κάνει σε παρόμοια αντικείμενα. 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. Η έξοδος του αλγορίθμου είναι οι γείτονες του νέου χρήστη.

Υπολογισμός βαρών (Weights Estimation): Το συγκεκριμένο κομμάτι υπολογίζει τα βάρη ομοιότητας που αντιστοιχούν στο νέο χρήστη και στον εκάστοτε γείτονα. Είσοδος του υποσυστήματος είναι τα δημογραφικά δεδομένα του νέου χρήστη και των γειτόνων. Για το στάδιο αυτό έχουν υλοποιηθεί αλγόριθμοι που υπολογίζουν το βάρος ομοιότητας ξεχωριστά για καθένα από τα δημογραφικά δεδομένα. Στη δική μας περίπτωση για την ηλικία εφαρμόζεται μία εκθετική συνάρτηση, για το επάγγελμα μία μετρική σημασιολογικής ομοιότητας και για το φύλο μία δυαδική μεταβλητή. Όλα τα βάρη συνδυάζονται και προκύπτει ένα συνολικό βάρος ομοιότητας το οποίο ορίζει το πόσο ταιριάζει ο νέος χρήστης με τον εκάστοτε γείτονά του. Το συνολικό βάρος αποτελεί και έξοδο του υποσυστήματος. Με αυτό τον τρόπο θεωρούμε ότι εάν ο νέος χρήστης και ο

γείτονας του μοιάζουν στα δημογραφικά δεδομένα τότε είναι μεγαλύτερη η πιθανότητα να έχουν και τις ίδιες προτιμήσεις.

Πρόβλεψη βάσει συνεργατικού φιλτραρίσματος (Collaborative Prediction): Το συγκεκριμένο υποσύστημα είναι ζωτικής σημασία αφού είναι υπεύθυνο για τον υπολογισμό της βαθμολογίας. Στον υπολογισμό λαμβάνεται υπόψη το προηγούμενο βήμα, ο υπολογισμός των βαρών, έτσι ώστε η τελική εκτίμηση να στηρίζεται στους γείτονες που μοιάζουν περισσότερο με το νέο χρήστη. Η είσοδος του υποσυστήματος είναι τα βάρη και οι αξιολογήσεις των γειτόνων για το αντικείμενο που είναι προς σύσταση. Η έξοδος του υποσυστήματος όπως φαίνεται στην Εικόνα 3 είναι οι βαθμολογίες για τα αντικείμενα που προτείνονται στο νέο χρήστη.

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 βαθμολογίες για κάποια αντικείμενα, ένα σύνολο δημογραφικών δεδομένων D = {ηλικία, επάγγελμα, φύλο, κατηγορία ταινίας} για κάθε χρήστη και ένα σύνολο αντικειμένων $I = \{i_1, i_2, \ldots\ldots i_n\}$. Όπως θα δούμε, η λίστα των χαρακτηριστικών μπορεί εύκολα να επεκταθεί ώστε να καλύψει περισσότερα δημογραφικά δεδομένα και έτσι να αυξήσει την αποδοτικότητα του συστήματος. Για κάθε $n_i \in N$ προβλέπουμε τη βαθμολογία του σε κάποιο αντικείμενο i ορίζοντας αρχικά μια γειτονιά, ένα υποσύνολο του U, όπου ανήκει ο n_i . 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. Στη συνέχεια για κάθε ζευγάρι (n_i , u_i) υπολογίζουμε ένα δείκτη ομοιότητας βάσης μίας κανονικοποιημένης συνάρτησης παλινδρόμησης με ανεξάρτητες μεταβλητές το δείκτη ομοιότητας για την ηλικία, το επάγγελμα και το φύλο των δύο χρηστών. Δηλαδή το συνολικό βάρος που ορίζεται για κάθε κοινωνικό δεσμό παραμετροποιείται από επιμέρους βάρη και καθορίζουνε σε ποια ποιο δημογραφικό δεδομένο επιθυμούμε να δώσουμε μεγαλύτερη έμφαση. Το συνολικό βάρος για κάθε συνάρτηση λαμβάνεται υπόψη στη διαδικασία πρόβλεψης της βαθμολογίας του n_i για ένα αντικείμενο. Για την πρόβλεψη βασιζόμαστε σε ένα σταθμισμένο μέσο όρο των αξιολογήσεων που έχουν οι γείτονες του n_i στο αντικείμενο i. Παρακάτω παρουσιάζοντα τα στάδα του προτεινόμενου αλγορίθμου.

4.2.1 Κατηγοριοποίηση του νέου χρήστη

Όπως αναφέρθηκε παραπάνω, ταξινομούμε το νέο χρήστη n_i σε μία από τις κατηγορίες-κλάσεις, $C = \{C_u^1, C_u^2, \ldots\ldots, C_u^n\}$ οι οποίες προκύπτουν από την εκπαίδευση του αλγορίθμου σε ένα σύνολο χρηστών. Τα γνωρίσματα τα οποία λαμβάνει υπόψη το μοντέλο για την εκπαίδευση είναι τα δημογραφικά δεδομένα των χρηστών. Για να ταξινομήσουμε τους χρήστες σε παραπάνω από δυο κλάσεις χρησιμοποιούμε τη κατηγοριοποίηση πολλαπλών κλάσεων(Multiclass classification). Στο προτεινόμενο μοντέλο έχουμε συνδυάσει έναν δυαδικό ταξινομητή με τη στρατηγική ένας-εναντίον-όλων (one-against-all (OvA)) για να επιτύχουμε κατηγοριοποίηση πολλαπλών κλάσεων. Ο αλγόριθμος μας έχει δύο φάσεις, η πρώτη φάση αφορά την εκπαίδευση ενός συνόλου χρηστών U για τη δημιουργία του μοντέλου, η δεύτερη φάση αφορά τη πρόβλεψη του μοντέλου για την κατηγορία στην οποία ανήκει κάθε νέος χρήστης.

Για το OvA εκπαιδεύουμε ένα δυαδικό ταξινομητή για κάθε κλάση, με αυτό τον τρόπο διακρίνουμε τη συγκεκριμένη κλάση από τις υπόλοιπες κλάσεις. Οπότε εάν θέλουμε να δημιουργήσουμε K κλάσεις τότε θα δημιουργηθούν K διαφορετικοί ταξινομητές. Παρακάτω παρουσιάζεται σε ψευδογλώσσα ο αλγόριθμος OvA.

Algorithm: One-against-all

Input:

L, a training algorithm for binary classifiers
Instances X
Labels y where $y_i \in \{1, \ldots K\}$ is the label for the instance x_i

Output:

List of classifiers f_k for $k \in \{1, \ldots K\}$

Begin

for each k in {1 … *K*}:
Construct a new label vector $y_i' = 1$ where $y_i = k$, 0 (or -1) elsewhere
Apply L training algorithm to *X*, *y'* to obtain f_k
end for

End

Για να προβλέψουμε την κλάση - κατηγορία του νέου στιγμιότυπου εφαρμόζουμε όλους τους ταξινομητές που έχει δημιουργήσει το μοντέλο. Η κλάση του νέου στιγμιότυπου είναι εκείνη για την οποία ο εκάστοτε ταξινομητής έχει το μεγαλύτερο βαθμό και υπολογίζεται με τον εξής τύπο:

$$\hat{y} = \arg\max_{1 \le k \le K} f_k(\mathrm{x}) \tag{21}$$

Για το δυαδικό ταξινομητή δοκιμάσαμε τους αλγορίθμους που περιγράφηκαν παραπάνω (C4.5 και Naïve Bayes). Για τους συγκεκριμένους αλγορίθμους χρησιμοποιήσαμε τις υλοποιήσεις του Weka [39], ένα εργαλείο το οποίο παρέχει αλγορίθμους για τους τομείς της μηχανική μάθησης και της εξόρυξης δεδομένων. Παρακάτω παρουσιάζουμε τον αλγόριθμο πολλαπλών κλάσεων του Weka (MultiClassClassifier) στον οποίο καλούμε το C4.5 ή το NaiveBayesSimple.

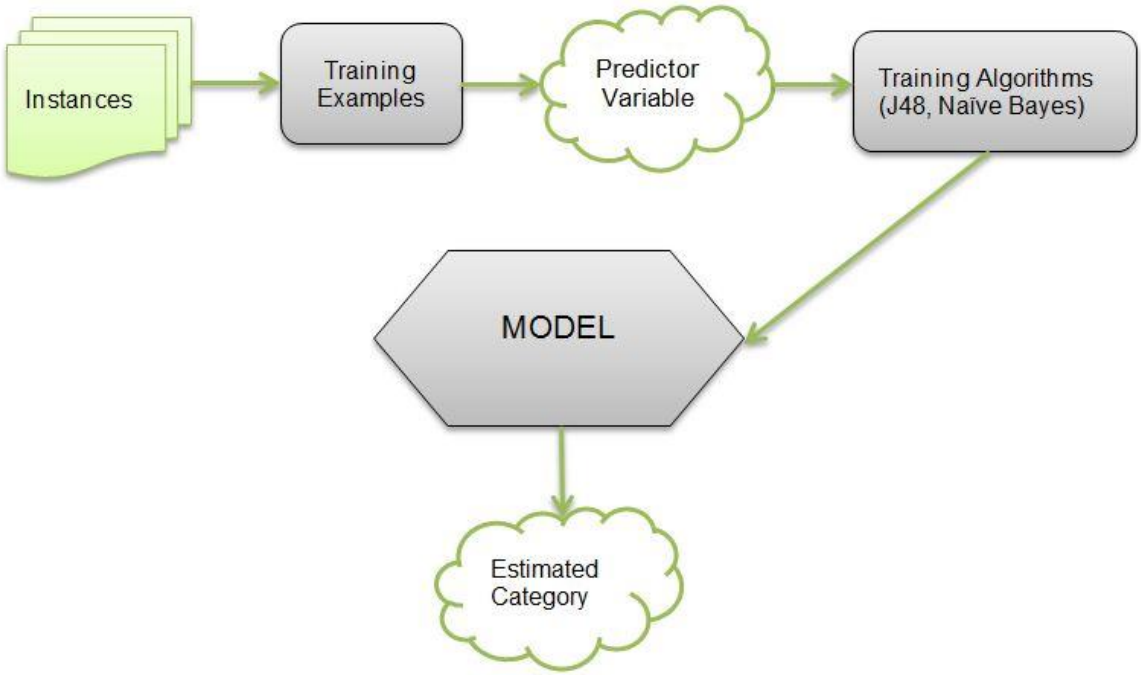
Algorithm: Multi-class Classification (C4.5 base)

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
Set class index
C4.5 tree
setUnpruned (C4.5)
Build the Mutli-class Classifier
Label unlabeled instances ni
for each $n_i \in \{n_1,n_2,\ldots\ldots,n_n\}$
Classify instance with trained model
end for

End

Στην Εικόνα 4 απεικονίζουμε τα βήματα που ακολουθούμε για τη δημιουργία του μοντέλου κατηγοριοποίησης.



Εικόνα 4: Διάγραμμα ροής για τη διαδικασία της κατηγοριοποίησης

Μεταβλητή Πρόβλεψης (Prediction variable): Όπως εξηγήσαμε και παραπάνω η μεταβλητή πρόβλεψης είναι ένα χαρακτηριστικό των δειγμάτων βάση του οποίου καθορίζεται η κατηγοριοποίηση. Η μεταβλητή πρόβλεψης της Εικόνας 3 πρέπει να ίδια με αυτή της Εικόνας 4 για να γίνει η πρόβλεψη της κατηγορίας.

Εκπαίδευση - Αλγόριθμοι εκπαίδευσης (Training algorithms): Η φάση της εκπαίδευσης είναι η πιο σημαντική για τη δημιουργία του μοντέλου. Η είσοδος του αλγορίθμου είναι η μεταβλητή πρόβλεψης και το σύνολο δεδομένων. Η ανάλυση του συνόλου δεδομένων παράγει το μοντέλο το οποίο αποτελεί έξοδο του αλγορίθμου κατηγοριοποίησης.

Μοντέλο (Model): Το μοντέλο είναι το αποτελέσμα του αλγορίθμου κατηγοριοποίησης και εξαρτάται από το σύνολο εκπαίδευσης καθώς και από τον εκάστοτε αλγόριθμο κατηγοριοποίησης.

4.2.2 Εύρεση της γειτονιάς του νέου χρήστη

Αφού έχουμε βρει σε ποια κατηγορία ανήκει ο εκάστοτε n_i συγκεντρώνουμε όλους τους χρήστες σε ομάδες ανάλογα με την κατηγορία που ανήκει ο καθένας και η οποία προέκυψε από τον ταξινομητή πολλαπλών κλάσεων. Με αυτόν τον τρόπο ορίζουμε τη γειτονιά του n_i.

```
Algorithm: Grouping

Input: Labeled n_i ∈ {n_1,n_2,……,n_n}, Instances u_i ∈ {u_1,u_2,……..,u_n}
Output: List of neighbors

Begin
        for each labeled n_i ∈ {n_1,n_2,……,n_n}
            Find n_iCategory
        end for
        for each u_i ∈ {u_1,u_2,……..,u_n}
            Find u_iCategory
            if  n_iCategory.Equals(u_iCategory) Then
                Add u_i in neighborsList
            end if
       end for
End
```

4.2.3 Υπολογισμός της ομοιότητας χρηστών

Στη συνέχεια υπολογίζουμε ένα βάρος ομοιότητας του n_i με καθένα από τους γείτονές του. Το βάρος ομοιότητας προκύπτει από τον παρακάτω σταθμισμένο μέσο όρο:

$$\mathrm{sim}(\mathrm{n},\mathrm{u}) = {}^{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{b_i w_i}} \tag{22}$$

Όπου w_i το απότελεσμα της εκάστοτε μετρικής 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 την ηλικία και b_i είναι ο συντελεστής βαρύτητας για καθένα από τα επιμέρους βάρη.

Για την προτεινόμενη τεχνική λαμβάνουμε υπόψη μόνο τρία είδη δημογραφικών δεδομένων. Οπότε στη δική μας περίπτωση η εξίσωση (22) μετατρέπεται ως εξής:

$$\mathrm{sim}(\mathrm{n},\mathrm{u}) = \mathrm{b_1}\,\mathrm{AgeSim} + \mathrm{b_2}\mathrm{OccuSim} + \mathrm{b_3}\mathrm{GenSim} \tag{23}$$

Για να υπολογίσουμε το βάρος ομοιότητας της ηλικίας AgeSim χρησιμοποιούμε την παρακάτω εκθετική συνάρτηση λαμβάνοντας υπόψη τη διαφορά της ηλικίας των δύο χρηστών καθώς και μία μέγιστη τιμή ηλικίας:

$$\mathrm{AgeSim(u,n)} = (1 - \frac{|D|}{D_{max}})^{a} \quad (24)$$

όπου D η διαφορά ηλικίας, D_{max} η εκτίμηση για τη μέγιστη τιμή της ηλικίας και α παράμετρος της εκθετικής συνάρτησης με τιμές στο διάστημα (0,∞). Όταν α < 1 τότε όσο μεγάλη να είναι η διάφορα στην ηλικία θεωρούμε ότι υπάρχει ομοιότητα. Ενώ για α > 1 τότε όσο μικρή είναι η διάφορα τόσο πιο δύσκολα θεωρούμε ότι υπάρχει ομοιότητα.

Για να υπολογίσουμε το βάρος ομοιότητας του επαγγέλματος χρησιμοποιήσαμε μία από τις μετρικές σημασιολογικής ομοιότητας, τη Wu and Palmer [40]. Η συγκεκριμένη μετρική συγκρίνει δύο έννοιες - λέξεις. Υπολογίζει την ομοιότητα συναρτήσει του μήκους του μονοπατιού από το least common subsumer (LCS). Δεδομένου δύο έννοιων το LCS ορίζεται ο πιο κοινός κόμβος από τον οποίο προέρχονται οι συγκεκριμένες έννοιες. Για παράδειγμα εάν θέλουμε να συγκρίνουμε τις έννοιες “αυτοκίνητο” και “σκάφος” τότε το LCS θα μπορούσε να είναι το “όχημα”. Η εξίσωση που υπολογίζει τη μετρική είναι η εξής:

$$\mathrm{OccuSim(u,n)} = \mathrm{sim_{wup}(c1,c2)} = \frac{2\times \mathrm{depth(LCS(c_1,c_2))}}{\mathrm{depth(c_1)} + \mathrm{depth(c_2)}} \quad (25)$$

όπου c1 και c2 είναι οι δύο εξεταζόμενες έννοιες δηλαδή τα επάγγελματα των χρηστών, $\mathrm{depth(c)}$ είναι η συνάρτηση που υπολογίζει το βάθος από την έννοια c στην ιεραρχία του WordNet [41], μία λεξιλογική βάση δεδομένων της αγγλικής γλώσσας.
Το GenSim προκύπτει απλά από δύο δυαδικές τιμές, 0 εάν το φύλο του n_i δεν είναι ίδιο με το φύλο του u_i και 1 εφόσον τα δύο φύλα είναι ίδια.

Παρακάτω παρουσιάζουμε συνοπτικά τον αλγόριθμο για τον υπολογισμό των βαρών.

Algorithm: Weights Calculation

Input: NewUsers $n_i \in \{n_1,n_2,\ldots\ldots,n_n\}$, Neighbors $u_i \in \{u_1,u_2,\ldots\ldots,u_n\}$
Output: Average weight $\bar{w}$

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
        Calculate average weight
    end for
End

4.2.4 Συνεργατική πρόβλεψη βαθμολογίας

Για την εκτίμηση της τιμής προτίμησης χρησιμοποιούμε συνεργατική πρόβλεψη προτιμήσεων βάσει της ομοιότητας των χρηστών. Με απλά λόγια, για να προβλέψουμε τη βαθμολογία $R_{n,i}$ ενός νέου χρήστη σε ένα αντικείμενο i υπολογίζουμε ένα σταθμισμένο μέσο όρο των βαθμολογιών που έχει το σύνολο των U των γειτόνων του.

$$R_{n,i} = \frac{\sum_{u \in U} sim(n,u) * r_{u,i}}{\sum_{u \in U} sim(n,u)} \quad (26)$$

όπου $sim(n,u)$ είναι η ομοιότητα του νέου χρήστη n με τον γείτονα u και $r_{u,i}$ είναι η βαθμολογία του χρήστη u για το αντικείμενο i. Το συγκεκριμένο μέρος αφορά το τελευταίο στάδιο της Εικόνας 3.

Algorithm: Rating Prediction

Input: Users $n_i \in \{n_1,n_2,\ldots\ldots,n_n\}$, Neighbors $u_i \in \{u_1,u_2,\ldots\ldots,u_n\}$, Ratings, Item i
Output: Predicted rating

Begin
    for each $n_i \in \{n_1,n_2,\ldots\ldots,n_n\}$
        Calculate weights
    end for
    for each $u_i \in \{u_1,u_2,\ldots\ldots,u_n\}$
        Find the rating in item i
    end for

Calculate the predicted rating $R_{n,i}$

End

5. ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ

Στη παρούσα ενότητα παρουσιάζεται ο τρόπος με τον οποίο έγινε η αξιολόγηση του συστήματος. Αρχικά, αναλύονται οι μετρικές απόδοσης που καθορίζουν την αποδοτικότητα του συστήματος και αξιολογούν την ικανότητα πρόβλεψης. Επιπλέον, γίνεται μία εκτενής αναφορά στα σύνολα δεδομένων που χρησιμοποιούνται για τις εκτελέσεις των πειραμάτων. Τέλος, παρουσιάζουμε τα διαφορετικά σενάρια βάσει των οποίων ορίζονται τα πειράματα και σχολιάζουμε τα αντίστοιχα αποτελέσματα. Στόχος της αξιολόγησης είναι τόσο η μέτρηση των επιδόσεων του συστήματος όσο και η επικύρωση της ορθής λειτουργίας του.

5.1 Μετρικές απόδοσης

Οι μετρικές μετρούν το πόσο διαφέρουν οι προβλέψεις από τις πραγματικές αξιολογήσεις. Μία μετρική απόδοσης ενός συστήματος συστάσεων είναι το μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE) [42]. 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:

$$\mathrm{MAE} = \frac{1}{N} \sum_{ij} \left| p_{i,j} - r_{i,j} \right| \tag{26}$$

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.

Μία άλλη μετρική απόδοσης για την αξιολόγηση της αποδοτικότητας των συστημάτων συστάσεων είναι η ρίζα του μέσου τετραγωνικού σφάλματος (Root Squared Mean Error - RMSE) [25].

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i,j}} \left(\mathrm{p_{i,j}} - \mathrm{r_{i,j}} \right)^2} \tag{27}$$

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. Το RMSE υπολογίζει τη τετραγωνική ρίζα της μέσης τιμής της διαφοράς υψωμένη στο τετράγωνο. Αξίζει να σημειωθεί ότι όσο πιο μικρές είναι οι τιμές των δύο μετρικών τόσο καλύτερες είναι οι προβλέψεις άρα τόσο μεγαλύτερη και η ακρίβεια του αλγορίθμου μας.

Οι συγκεκριμένες μετρικές είναι ευρέως αποδεκτές για την αξιολόγηση συστημάτων συστάσεων και έχουν κατά καιρούς χρησιμοποιηθεί από πολλές ερευνητικές μελέτες.

5.2 Σύνολα δεδομένων

Στα πειράματά μας έχουμε χρησιμοποιήσει σύνολα δεδομένων από το MovieLens[1], μία εφαρμογή η οποία προσφέρει συστάσεις για ταινίες. Η ερευνητική ομάδα του GroupLens [43] έχει συλλέξει κατά καιρούς σύνολα δεδομένων με αξιολογήσεις των χρηστών του MovieLens. Για τα πειράματά μας χρησιμοποιήσαμε ένα σύνολο δεδομένων που περιέχει 100.000 αξιολογήσεις 1000 χρηστών για 1700 διαφορετικές ταινίες. Από το σύνολο των 1000 χρηστών επιλέξαμε τους πρώτους 700 να αποτελούν τους εγγεγραμμένους χρήστες του συστήματος και τους τελευταίους 50 ως νέους χρήστες. Τα πειράματα ξεκίνησαν με 100 εγγεγραμμένους χρήστες που θεωρούμε ότι είναι ήδη στο σύστημα και έχουν κάνει αξιολογήσεις σε διάφορες ταινίες. Στα πλαίσια των πειραμάτων μας αυξάνουμε τον αριθμό τους κατά 100 για να δούμε την επίπτωση που έχει ο αριθμός των εγγεγραμμένων χρηστών στο τελικό αποτελέσμα. Το σύνολο των αξιολογήσεων έχει βαθμολογίες από 1 που είναι η ελάχιστη αξιολόγηση έως 5 που είναι η μέγιστη. Για την εκτέλεση των πειραμάτων επεξεργαστήκαμε κατάλληλα το σύνολο δεδομένων που αφορά τους χρήστες. Τα δεδομένα τα οποία κρατήσαμε από το συγκεκριμένο σύνολο είναι το αναγνωριστικό (id) και τα δημογραφικά στοιχεία (age, occupation, gender). Σε αυτά προσθέσαμε ένα χαρακτηριστικό (property) το οποίο συνδέεται με το χαρακτήρα του χρήστη με τις ταινίες που επιλέγει να δει. Οι τιμές του χαρακτηριστικού μπορεί να είναι πολλές, εμείς υιοθετούμε τις ‘διασκεδαστικός’, ‘διανοούμενος’, ‘περιπετειώδης’ και ‘ρομαντικός’ (fun, intellectual, adventurous, romantic). Η τιμή του χαρακτηριστικού (property) που προστέθηκε σε κάθε δείγμα λαμβάνει υπόψη το επάγγελμα του χρήστη. Για παράδειγμα εάν ο χρήστης έχει το επάγγελμα του ‘επιστήμονας’ υποθέτουμε ότι είναι πιο πιθανό να είναι ‘διανοούμενος’ σε σχέση με τα υπόλοιπα χαρακτηριστικά. Δηλαδή δημιουργούμε μία πλειάδα που, για το συγκεκριμένο παράδειγμα, καταχωρούμε τις εξής πιθανότητες (0.0, 0.8, 0.2, 0.0). Η πρώτη τιμή της πλειάδας είναι η πιθανότητα που αντιστοιχεί στον χρήστη για το χαρακτηριστικό διασκεδαστικός, η δεύτερη για το διανοούμενος, η τρίτη για το

[1] http://www.movielens.org

περιπετειώδης και η τέταρτη για το ρομαντικός. Ουσιαστικά, το χαρακτηριστικό (property) αποτελεί τη μεταβλητή πρόβλεψης η οποία αναφέρθηκε στο τέταρτο κεφάλαιο. Συγκεκριμένα, δημιουργήσαμε δύο είδη συνόλων ένα που είναι κατάλληλο για δυαδική κατηγοριοποίηση και ένα το οποίο είναι κατάλληλο για την κατηγοριοποίηση πολλαπλών κλάσεων. Στην περίπτωση της δυαδικής κατηγοριοποίησης (βλ. Εικόνα 5) χρησιμοποιήσαμε ένα σύνολο δεδομένων που έχει δύο τιμές για το property 'διανοούμενος' και 'διασκεδαστικός'. Η τιμή που εμφανίζεται σε κάθε διάνυσμα της Εικόνας 5 είναι αυτή που αντιστοιχεί στο 'διασκεδαστικός'. Δηλαδή παρατηρούμε ότι για το επάγγελμα του προγραμματιστή (programmer) η πιθανότητα να είναι 'διασκεδαστικός' είναι 0.5 οπότε θα είναι και 0.5 να είναι 'διανοούμενος'. Άρα οι προγραμματιστές (programmers) θα είναι μισοί "διασκεδαστικοί" (fun) και μισοί "διανοούμενοι" (intellectual).

```
private static final AttrWeight weights[] = {
  new AttrWeight("technician", 0),
  new AttrWeight("other", 0.5),
  new AttrWeight("writer", 0),
  new AttrWeight("executive", 0.3),
  new AttrWeight("administrator", 0.4),
  new AttrWeight("student", 1),
  new AttrWeight("lawyer", 0),
  new AttrWeight("educator", 0),
  new AttrWeight("scientist",0),
  new AttrWeight("entertainment", 1),
  new AttrWeight("programmer", 0.5),
  new AttrWeight("librarian", 0.3),
  new AttrWeight("homemaker", 0.7),
  new AttrWeight("artist", 0.7),
  new AttrWeight("engineer", 0.4),
  new AttrWeight("marketing", 0.8),
  new AttrWeight("healthcare", 0.9),
  new AttrWeight("retired", 0.5),
  new AttrWeight("salesman", 0.5),
  new AttrWeight("doctor", 0)
};
```

Εικόνα 5: Τιμές του property στη δυαδική κατηγοριοποίηση

```
private static final MultiAttrWeight weights[] = {
  new MultiAttrWeight("technician", 0, 0.7, 0.3, 0),
  new MultiAttrWeight("other", 0.5, 0, 0.2, 0.3),
  new MultiAttrWeight("writer", 0, 1, 0, 0),
  new MultiAttrWeight("executive", 0.1, 0.8, 0, 0.1),
  new MultiAttrWeight("administrator", 0.3, 0.7, 0, 0),
  new MultiAttrWeight("student", 1, 0, 0, 0),
  new MultiAttrWeight("lawyer", 0.1, 0.8, 0, 0.1),
  new MultiAttrWeight("educator", 0, 1, 0, 0),
  new MultiAttrWeight("scientist",0, 0.8, 0.2, 0),
  new MultiAttrWeight("entertainment", 0.9, 0, 0.1, 0),
  new MultiAttrWeight("programmer", 0.8, 0.1, 0.1, 0),
  new MultiAttrWeight("librarian", 0.3, 0, 0, 0.7),
  new MultiAttrWeight("homemaker", 0.7, 0, 0, 0.3),
  new MultiAttrWeight("artist", 0.2, 0, 0.3, 0.5),
  new MultiAttrWeight("engineer", 0.4, 0.6, 0, 0),
  new MultiAttrWeight("marketing", 0.6, 0, 0.2, 0.2),
  new MultiAttrWeight("healthcare", 0.8, 0, 0.1, 0.1),
  new MultiAttrWeight("retired", 0.3, 0.1, 0.3, 0.3),
  new MultiAttrWeight("salesman", 0.8, 0.1, 0, 0.1),
  new MultiAttrWeight("doctor", 0, 1, 0, 0)
```

Εικόνα 6: Τιμές του property στην κατηγοριοποίηση πολλαπλών κλάσεων

Για την κατηγοριοποίηση πολλαπλών κλάσεων το σύνολο δεδομένων περιέχει τις τέσσερις τιμές που ορίσαμε παραπάνω (fun, intellectual, adventurous, romantic). Οι τιμές τις οποίες ορίσαμε για κάθε πλειάδα παρουσιάζονται στην Εικόνα 6. Με αυτό τον τρόπο, το σύνολο δεδομένων μπορεί να επεκταθεί έτσι ώστε το χαρακτηριστικό (property) να έχει παραπάνω τιμές.

5.3 Παράμετροι πειραμάτων

Όπως αναφέρθηκε παραπάνω, ο προτεινόμενος αλγόριθμος περιλαμβάνει ένα σημαντικό στάδιο, που είναι αυτό της κατηγοριοποίησης. Για να αξιολογήσουμε τη τεχνική την οποία προτείνουμε, πραγματοποιήσαμε δοκιμές με δύο κατηγοριοποιητές (αλγόριθμοι κατηγοριοποίησης). Αρχικά, εκτελούμε πειράματα χρησιμοποιώντας τον δυαδικό C4.5 (στο Weka η αντίστοιχη υλοποίηση ονομάζεται J48) στη θέση του κατηγοριοποιητή. Ο C4.5 επιτυγχάνει δυαδική κατηγοριοποίηση προβλέποντας ότι ο νέος χρήστης θα ανήκει σε μία από τις δύο κατηγορίες που έχουμε ορίσει. Επίσης, πραγματοποιήσαμε πειράματα, επεκτείνοντας την προηγούμενη προσέγγιση, έτσι ώστε

οι κατηγορίες πρόβλεψης να είναι τέσσερις και να εντάξουμε το νέο χρήστη σε μία από αυτές. Ουσιαστικά, υλοποιήσαμε μία κατηγοριοποίηση πολλαπλών κλάσεων εφαρμόζοντας στη θέση του δυαδικού κατηγοριοποιητή για την OvA στρατηγική τον C4.5 και τον Naïve Bayes. Επίσης, συγκρίναμε όλες αυτές τις προσεγγίσεις με έναν αλγόριθμο αναφοράς (baseline) με τον οποίο δεν δημιουργούμε κάποιο μοντέλο για να προβλέψουμε την κατηγορία στην οποία ανήκει ο νέος χρήστης. Στη προκειμένη περίπτωση, αναθέτουμε τυχαία την κατηγορία στην οποία ανήκει ο νέος χρήστης χωρίς να λάβουμε υπόψη τα δημογραφικά δεδομένα. Η συγκεκριμένη μέθοδος ονομάστηκε Random Classification Algorithm (RCA) και τα αποτελέσματά της συγκρίθηκαν με τα αποτελέσματα των υπολοίπων προσεγγίσεων. Τα διαφορετικά σενάρια προέκυψαν με βάση τις διαφορετικές τιμές για τους συντελεστές βαρύτητας της εξίσωσης (23) καθώς και τις διαφορετικές τιμές του εκθέτη για την εξίσωση (24). Δοκιμάσαμε διάφορες τιμές για τα βάρη b_i λαμβάνοντας υπόψη το εξής:

$$\sum_{i=1}^{3} bi = 1 \qquad (28)$$

Στον Πίνακα 3, παρουσιάζονται συνοπτικά οι παράμετροι βάσει των οποίων έγιναν τα πειράματα.

Πίνακας 3: Παράμετροι πειραμάτων

Παράμετροι	**Τιμές**
Αλγόριθμοι	C4.5, Naïve Bayes, RCA
Συντελεστές βαρύτητας b_i	$b_i \in \{0....1\}$ με $\sum_{i=1}^{3} b_i = 1$
α	0.8 και 5

5.4 Σενάρια και αξιολόγηση αποτελεσμάτων

Στο σημείο αυτό παρουσιάζουμε τα αποτελέσματα του προτεινόμενου αλγορίθμου βάσει κάποιων σεναρίων εκτέλεσης. Σε κάθε σενάριο υπολογίζονται οι επιμέρους ομοιότητες του νέου χρήστη και των γειτόνων για καθένα από τα δημογραφικά

δεδομένα. Δηλαδή υπολογίζουμε το βάρος ομοιότητας των χρηστών στην ηλικία, στο επάγγελμα και στο φύλο. Σε καθένα από αυτούς τους δείκτες ομοιότητας αναθέτουμε και συγκεκριμένους συντελεστές βαρύτητας. Τα διαφορετικά σενάρια σχετίζονται με τις διαφορετικές τιμές που αναθέτουμε στους συντελεστές βαρύτητας b_i της εξίσωσης (23) και την παράμετρο α της εξίσωσης (24). Σε όλα σχεδόν τα σχήματα που απεικονίζουν τα αποτελέσματα των μετρικών MAE και RMSE συγκρίνουμε τους διαφορετικούς κατηγοριοποιητές που εφαρμόσαμε στον αλγόριθμό μας. Οι συγκρίσεις γίνονται στις περιπτώσεις όπου ο κατηγοριοποιητής εξάγει τέσσερις πιθανές κλάσεις. Δηλαδή, ο χρήστης μπορεί να ανήκει σε μία από αυτές τις τέσσερις κατηγορίες. Το Multi - C4.5 και Multi - Naïve Bayes αφορά τη χρήση των αλγορίθμων C4.5 και Naïve Bayes αντίστοιχα ως δυαδικούς κατηγοριοποιητές για την κατηγοριοποίηση πολλαπλών κλάσεων εφαρμόζοντας τη OvA στρατηγική.

Σενάριο 1^{ο}

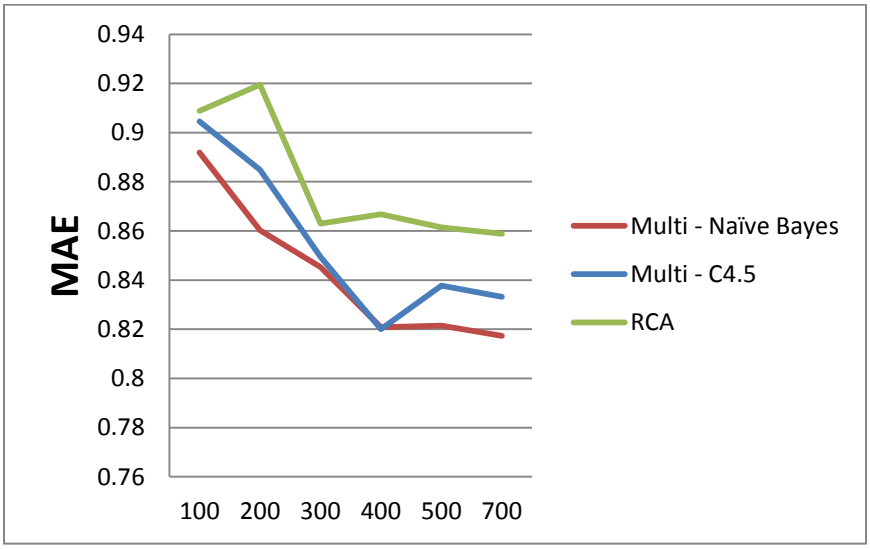
Στο 1^{ο} σενάριο θεωρούμε ότι οι επιμέρους ομοιότητες που σχετίζονται με καθένα από τα δημογραφικά δεδομένα συνεισφέρουν κατά το ίδιο ποσοστό στον υπολογισμό του συνολικού βάρους της εξίσωσης (23). Συνεπώς, $b_1 = b_2 = b_3 = \frac{1}{3}$, όπου b_1 ο συντελεστής βαρύτητας για την ηλικία, b_2 ο συντελεστής βαρύτητας για το επάγγελμα και b_3 ο συντελεστής βαρύτητας για το φύλο. Παρακάτω, παρουσιάζουμε τα αποτελέσματα για το συγκεκριμένο σενάριο.

Πίνακας 4: Αποτελέσματα MAE για το 1° σενάριο (α = 0.8)

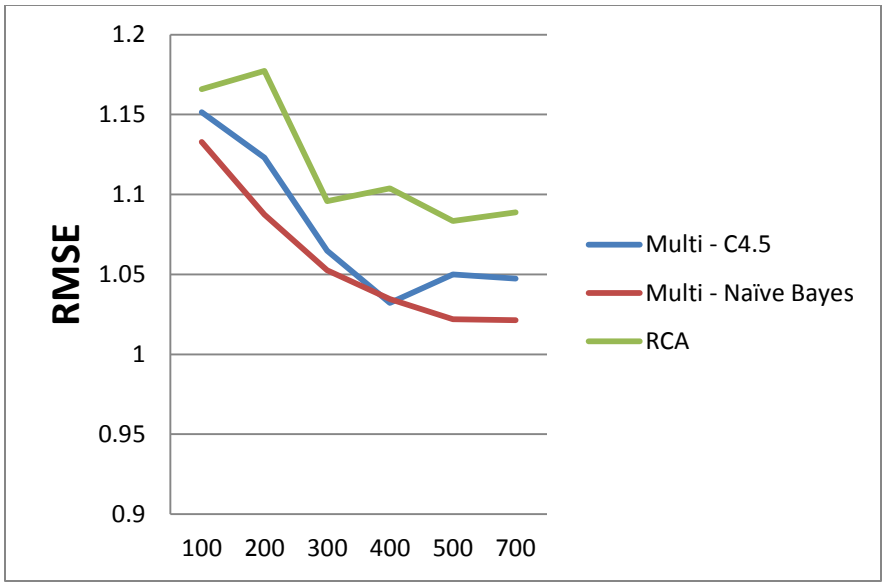
Users	Multi – C4.5	Multi – Naïve Bayes	RCA
100	0.9	0.89	0.91
200	0.88	0.86	0.92
300	0.85	0.85	0.86
400	0.82	0.82	0.87
500	0.84	0.82	0.86
700	0.83	0.82	0.86

Πίνακας 5: Αποτελέσματα RMSE για το 1° σενάριο (α = 0.8)

Users	Multi – C4.6	Multi – Naïve Bayes	RCA
100	1.15	1.13	1.17
200	1.12	1.09	1.18
300	1.06	1.05	1.1
400	1.03	1.03	1.1
500	1.05	1.02	1.08
700	1.05	1.02	1.09



Σχήμα 1: Αποτελέσματα MAE για το 1° σενάριο



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

Από τα Σχήματα 1 και 2 φαίνεται ότι και οι δύο προσεγγίσεις που χρησιμοποιούμε στην κατηγοριοποίηση πολλαπλών κλάσεων έχουν καλύτερα αποτελέσματα από εκείνη του RCA όπου οι κλάσεις επιλέγονται τυχαία. Στον Πίνακα 4 παρατηρούμε ότι όλες οι τιμές του MAE είναι μικρότερο της μονάδας το οποίο είναι αρκετά ενθαρρυντικό. Η μικρότερη τιμή παρατηρείται στην περίπτωση των 900 εγγεγραμμένων χρηστών. Όπως βλέπουμε στα σχήματα η καμπύλη του RCA παρουσιάζει αυξομειώσεις καθ'όλη τη διάρκεια εκτέλεσης των πειραμάτων. Από την άλλη πλευρά οι προσεγγίσεις Multi - C4.5 και Multi – Naïve Bayes παρουσιάζουν πολύ καλύτερη εικόνα. Είναι ενθαρρυντικό ότι οι τιμές του MAE είναι μικρότερες από τη μονάδα για όλες τις δοκιμές. Παρόλα αυτά ο Multi - Naïve Bayes έχει καλύτερη απόδοση από τον Multi - C4.5 με το δεύτερο να παρουσιάζει μία αστάθεια καθώς ο αριθμός των χρηστών αυξάνεται. Όταν ο αριθμός των εγγεγραμμένων χρηστών γίνει 400, εμφανίζεται μία μικρή άνοδος στις τιμές και των δύο μετρικών άρα και άνοδος στο ποσοστό των σφαλμάτων. Όπως θα δούμε παρακάτω, σε ένα άλλο σενάριο το οποίο συνδυάζει διαφορετικά τα βάρη, τα αποτελέσματα του Multi - C4.5 δεν παρουσιάζουν το συγκεκριμένο μειονέκτημα. Στα πειράματα ορίσαμε α = 0.8. Κάναμε αντίστοιχα πειράματα με α = 5 και τα αποτελέσματα είναι παρόμοια.

Σενάριο 2^{ο}

Στα επόμενα σενάρια τροποποιούμε την παραμετροποίηση με τέτοιο τρόπο ώστε να δώσουμε μεγαλύτερη έμφαση στον υπολογισμό της ομοιότητας σε κάποια από τα δημογραφικά δεδομένα του χρήστη.

Στο συγκεκριμένο σενάριο δίνουμε μεγαλύτερη έμφαση στο επάγγελμα και στην αντίστοιχη μετρική OccuSim. Οπότε οι τιμές των b_i διαμορφώνονται ως εξής:

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

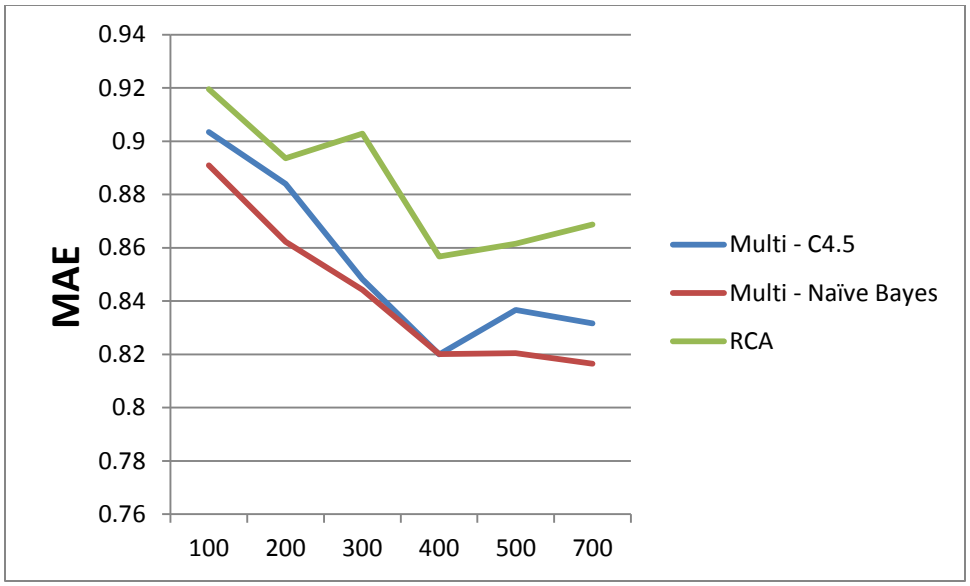
Παρακάτω παρουσιάζουμε τα αποτελέσματα του 2^{ο} σεναρίου.

Πίνακας 6: Αποτελέσματα MAE για το 2^{ο} σενάριο (α = 0.8)

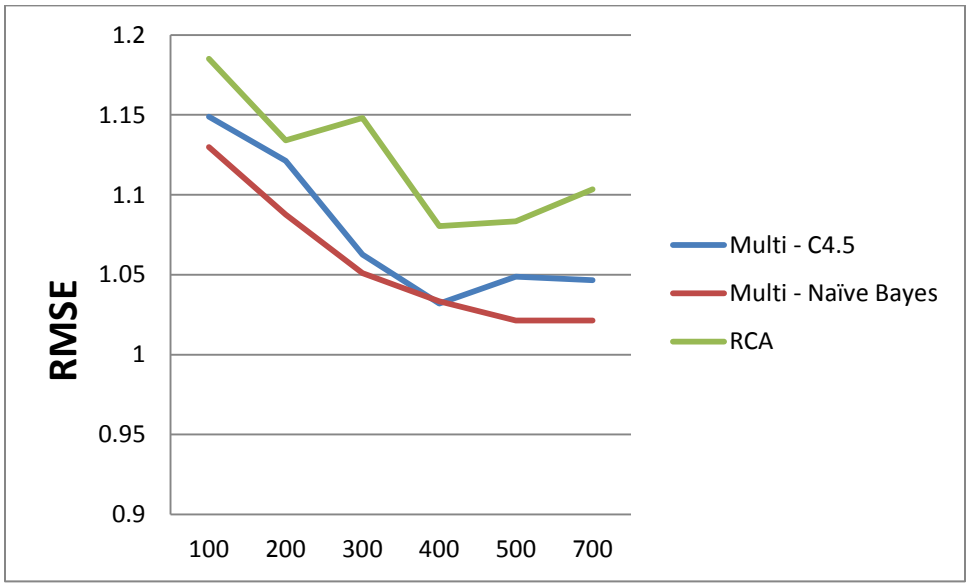
Users	Multi - C4.5	Multi - Naïve Bayes	RCA
100	0.9	0.89	0.92
200	0.88	0.86	0.89
300	0.85	0.84	0.9
400	0.82	0.82	0.86
500	0.84	0.82	0.86
700	0.83	0.82	0.87

Πίνακας 7: Αποτελέσματα RMSE για το 2^{ο} σενάριο (α = 0.8)

Users	Multi - C4.5	Multi - Naïve Bayes	RCA
100	1.15	1.13	1.19
200	1.12	1.09	1.13
300	1.06	1.05	1.15
400	1.03	1.03	1.08
500	1.05	1.02	1.08
700	1.05	1.02	1.1



Σχήμα 3: Αποτελέσματα MAE για το 2ο σενάριο



Σχήμα 4: Αποτελέσματα RMSE για το 2ο σενάριο

Παρατηρούμε στα Σχήματα 3 και 4 ότι η εικόνα είναι παρόμοια με το προηγούμενο σενάριο. Αξίζει να τονίσουμε ότι και εδώ οι τιμές των μετρικών είναι μικρές άρα και η απόδοση του συστήματος αρκετά καλή. Επιπλέον, όπως φαίνεται στους Πίνακες 6 και 7 μετά τους 400 χρήστες η Multi - C4.5 προσέγγιση παρουσιάζει μία ανωμαλία αφού οι μετρικές MAE και RMSE αυξάνονται. Οι τιμές των μετρικών δεν επηρεάζονται σε μεγάλο βαθμό από το το βάρος που δίνουμε στο επάγγελμα.

Σενάριο 3^{ο}

Στη συνέχεια δίνουμε μεγαλύτερη έμφαση στο βάρος για το φύλο και στην αντίστοιχη μετρική GenSim. Συνεπώς, οι τιμές των b_i διαμορφώνονται ως εξής:

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

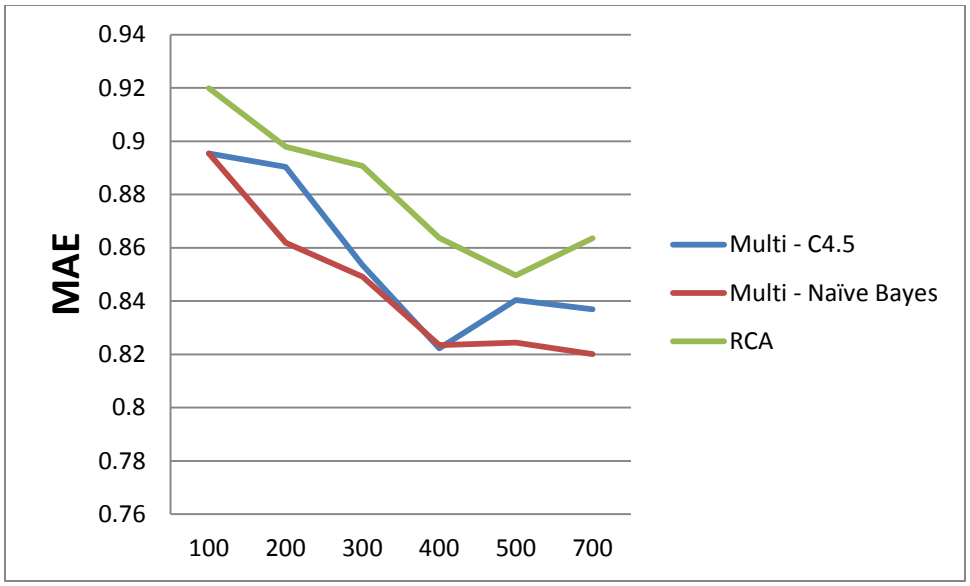
Παρακάτω, παρουσιάζονται τα αποτελέσματα για το 3^{ο} σενάριο.

Πίνακας 8: Αποτελέσματα MAE για το 3^{ο} σενάριο (α = 0.8)

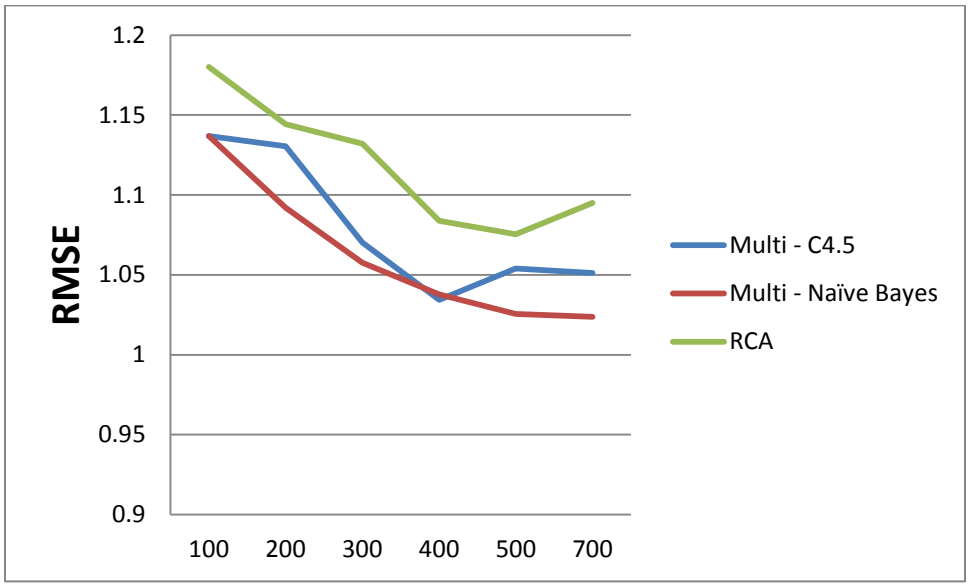
Users	Multi - C4.5	Multi - Naïve Bayes	RCA
100	0.9	0.9	0.92
200	0.89	0.86	0.9
300	0.85	0.85	0.89
400	0.82	0.82	0.86
500	0.84	0.82	0.85
700	0.84	0.82	0.86

Πίνακας 9: Αποτελέσματα RMSE για το 3^{ο} σενάριο (α = 0.8)

Users	Multi - C4.5	Multi - Naïve Bayes	RCA
100	1.14	1.14	1.18
200	1.13	1.09	1.14
300	1.07	1.06	1.13
400	1.03	1.04	1.08
500	1.05	1.03	1.08
700	1.05	1.02	1.1



Σχήμα 5: Αποτελέσματα MAE για το 3ο σενάριο



Σχήμα 6: Αποτελέσματα RMSE για το 3ο σενάριο

Όπως απεικονίζεται στα Σχήματα 5 και 6 τα αποτελέσματα είναι σχεδόν παρόμοια με τα προηγούμενα σενάρια. Η διαφορά σε αυτό το σενάριο είναι ότι παρατηρούνται αυξομειώσεις στη μετρική MAE όχι μόνο στην προσέγγιση του Multi - C4.5 αλλά και του Naïve Bayes. Παρόλα αυτά και σε αυτό το σενάριο όπως και στα προηγούμενα σενάρια μεγαλύτερη αποδοτικότητα παρουσιάζει ο Naïve Bayes.

Σενάριο 4^{ο}

Στο σενάριο αυτό εξετάζουμε την επιρροή που έχει η ομοιότητα των χρηστών που αφορά την ηλικία στον τελικό υπολογισμό της προβλεπόμενης τιμής. Δίνοντας μεγαλύτερη έμφαση στην ηλικία, AgeSim, τα βάρη της εξίσωσης (23) διαμορφώνονται ως εξής:

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

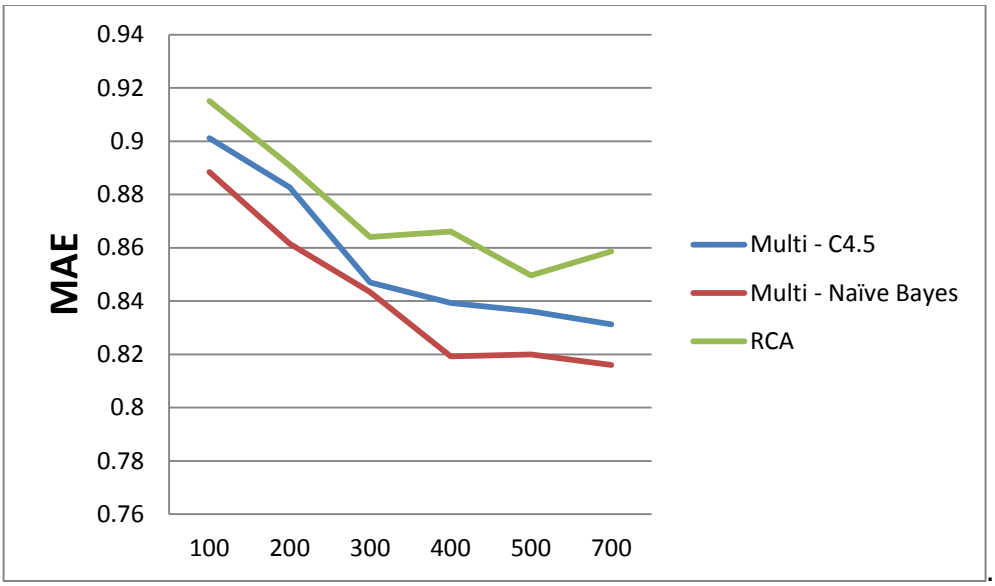
Τα αποτελέσματα που προκύπτουν από την εκτέλεση των πειραμάτων παρουσιάζονται παρακάτω:

Πίνακας 10: Αποτελέσματα MAE για το 4^{ο} σενάριο (α = 0.8)

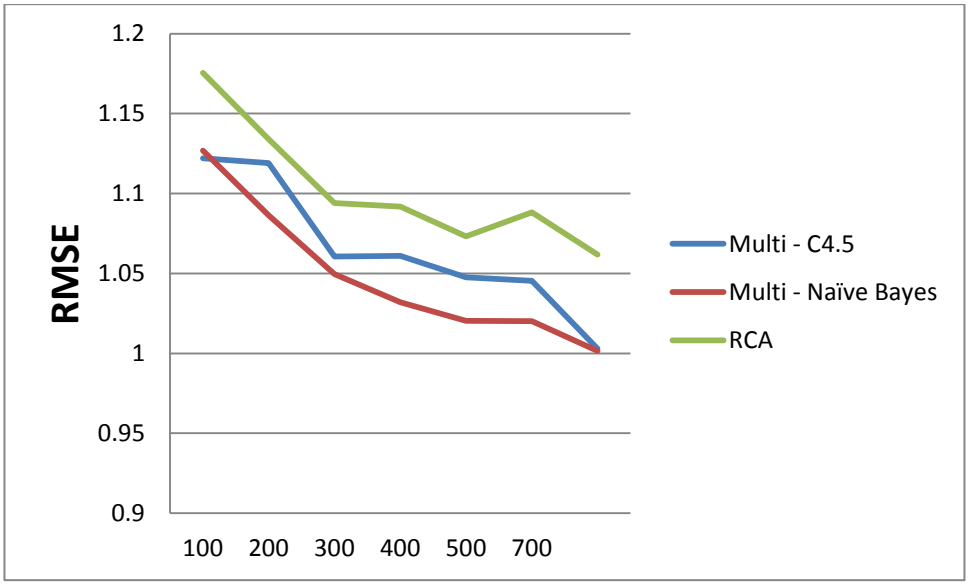
Users	Multi - C4.5	Multi - Naïve Bayes	RCA
100	0.9	0.89	0.92
200	0.88	0.86	0.89
300	0.85	0.84	0.86
400	0.84	0.82	0.87
500	0.84	0.82	0.85
700	0.83	0.82	0.86

Πίνακας 11: Αποτελέσματα MAE για το 4^{ο} σενάριο (α = 0.8)

Users	Multi – C4.5	Multi – Naïve Bayes	RCA
100	1.12	1.13	1.18
200	1.12	1.09	1.13
300	1.06	1.05	1.09
400	1.06	1.03	1.09
500	1.05	1.02	1.07
700	1.05	1.02	1.09



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



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

Το συγκεκριμένο σενάριο παρουσιάζει τα καλύτερα αποτελέσματα και για τις δύο μετρικές αφού οι τιμές τους είναι μικρότερες σε σχέση με τα προηγούμενα σενάρια. Όπως απεικονίζεται στα Σχήματα 7 και 8 η καμπύλη του Multi – C4.5 δεν παρουσιάζει την απότομη αυξομείωση που είδαμε στα προηγούμενα σενάρια. Αντιθέτως, και για τις δύο προσεγγίσεις Multi - C4.5 και Multi - Naïve Bayes καθώς αυξάνεται ο αριθμός των εγγεγραμμένων χρηστών στο σύστημα η μετρική MAE μειώνεται σταδιακά και δεν

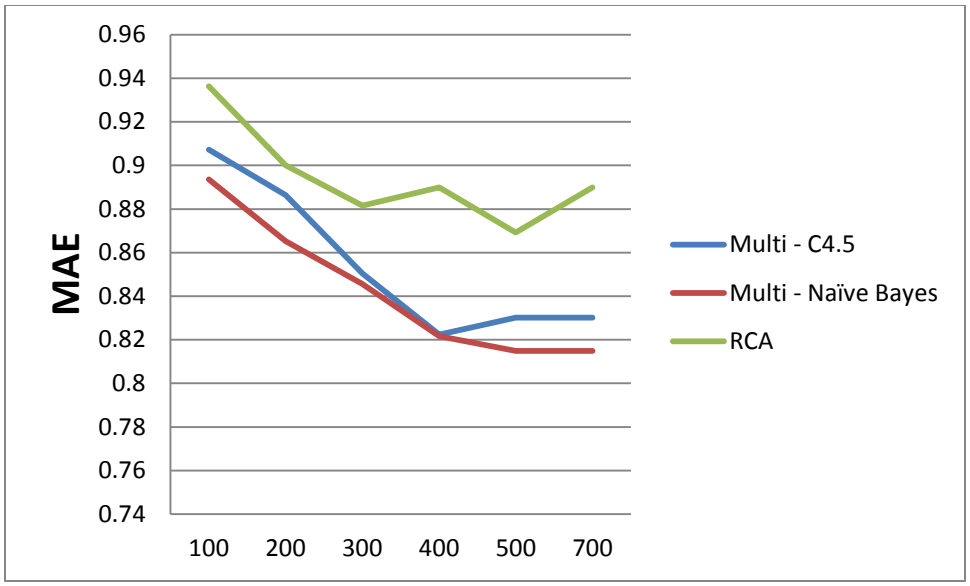
αυξάνεται απότομα όπως είδαμε και στα προηγούμενα σενάρια. Ο Multi - Naïve Bayes παρουσιάζει μικρότερα ποσοστά λαθών από το Multi - C4.5. Οπότε για την προτεινόμενη τεχνική ο Multi - Naïve Bayes είναι πιο αποδοτικός. Στα αποτελέσματα που απεικονίζονται στα Σχήματα 7 και 8 ορίσαμε α = 0.8. Παρακάτω παρουσιάζουμε τα αποτελέσματα της μετρικής MAE με α = 5.

Πίνακας 12: Αποτελέσματα MAE για το 4^{ο} σενάριο (α = 5)

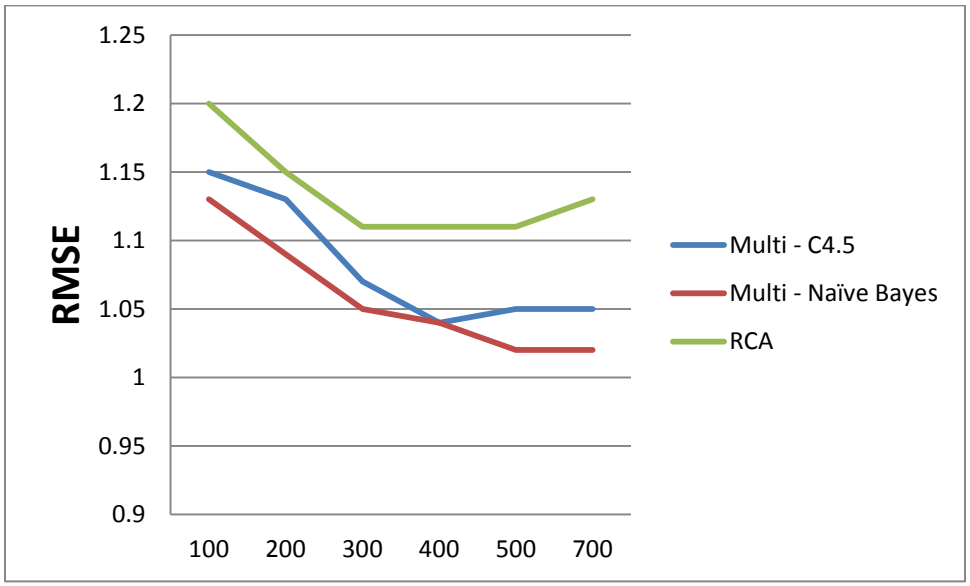
Users	Multi - C4.5	Multi - Naïve Bayes	RCA
100	0.91	0.89	0.94
200	0.89	0.87	0.9
300	0.85	0.85	0.88
400	0.82	0.82	0.89
500	0.83	0.81	0.87
700	0.81	0.81	0.84

Πίνακας 13: Αποτελέσματα RMSE για το 4^{ο} σενάριο (α = 5)

Users	Multi - C4.5	Multi - Naïve Bayes	RCA
100	1.15	1.13	1.2
200	1.13	1.09	1.15
300	1.07	1.05	1.11
400	0.04	1.04	1.11
500	1.05	1.02	1.11
700	1.05	1.02	1.13



Σχήμα 9: Αποτελέσματα MAE για το 4^{ο} σενάριο (α = 5)



Σχήμα 10: Αποτελέσματα RMSE για το 4^{ο} σενάριο (α = 5)

Όπως φαίνεται στο Σχήμα 9, για α = 5, ο Multi – C4.5 δεν παρουσιάζει τόσο καλή εικόνα όσο προηγουμένως, αφού παρατηρούμε πάλι μία αυξομείωση με την άνοδο των εγγεγραμμένων χρηστών. Άρα, για το συγκεκριμένο σενάριο όσο πιο μικρή είναι η τιμή του a τόσο ποιο αποδοτικός είναι ο προτεινόμενος αλγόριθμος.

Αφού συμπεράναμε από τα αντίστοιχα διαγράμματα ότι το 4^{ο} σενάριο είναι το καλύτερο για την προτεινόμενη τεχνική θα παρουσιάσουμε τα ποσοστά βελτίωσης του MAE όταν εφαρμόζουμε τους Multi - C4.5 και Multi - Naïve Bayes κατηγοριοποιητές σε σχέση με τον RCA.

Πίνακας 14: Ποσοστό βελτίωσης των C4.5 & Naïve Bayes σε σχέση με τον RCA

Users	C4.5 (2 cls)	Multi - C4.5 (4 cls)
100	1.52%	2.91%
200	0.92%	3.29%
300	1.98%	2.39%
400	3.09%	5.40%
500	1.59%	3.50%
600	3.19%	4.96%
700	5.39%	5.55%

Από τον Πίνακα 13 παρατηρούμε ότι η χρήση των κατηγοριοποιητών Multi - C4.5 και Multi - Naïve Bayes βελτιώνει την απόδοση του αλγορίθμου. Είναι αρκετά ενθαρρυντικό ότι το ποσοστό βελτίωσης σε σχέση με τον RCA φτάνει στο 5.55%.

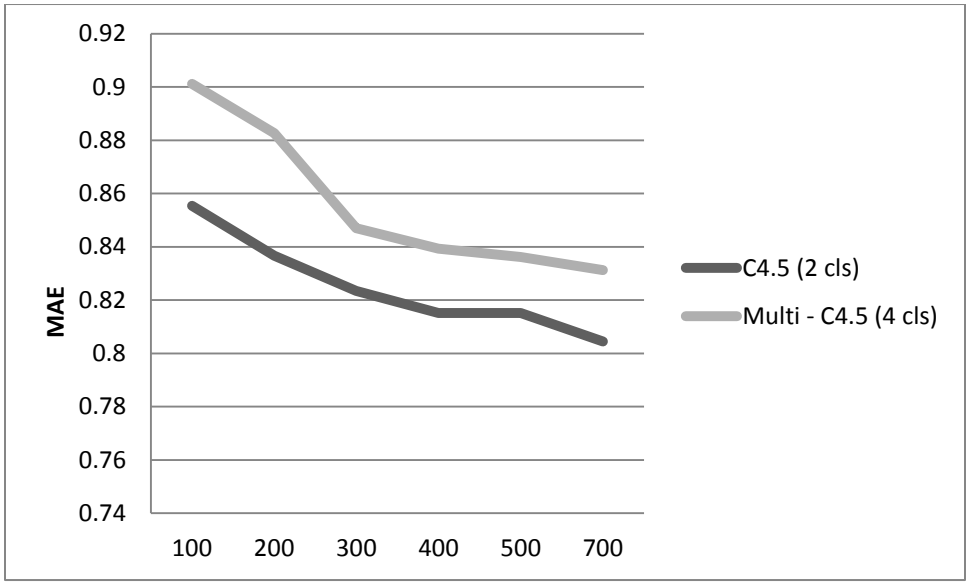
Στο σημείο αυτό, παρουσιάζουμε για το 4^{ο} σενάριο μία τελευταία σειρά αποτελεσμάτων η οποία συγκρίνει την περίπτωση όπου για τον κατηγοριοποιητή C4.5 πραγματοποιούμε δυαδική κατηγοριοποίηση (δύο κλάσεις) με την περίπτωση πολλαπλής κατηγοριοποίησης για τον ίδιο αλγόριθμο (βλ. Σχήμα 10).

Πίνακας 15: Αποτελέσματα MAE για το δυαδικό C4.5 και Multi - C4.5

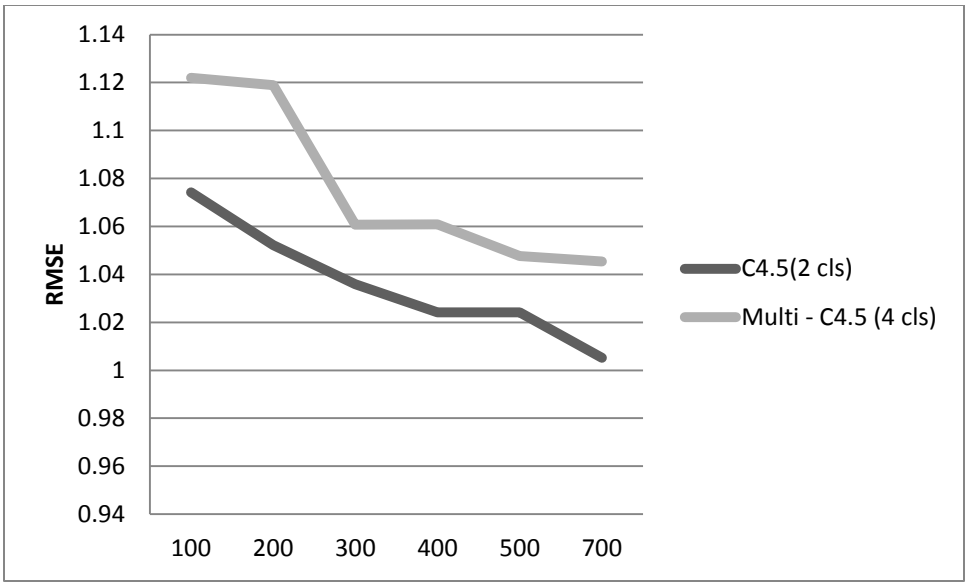
Users	C4.5 (2 cls)	Multi - C4.5
100	0.86	0.9
200	0.84	0.88
300	0.82	0.85
400	0.82	0.84
500	0.82	0.84
700	0.8	0.83

Πίνακας 16: Αποτελέσματα RMSE για το δυαδικό C4.5 και Multi - C4.5

Users	C4.5 (2 cls)	Multi - C4.5
100	1.07	1.12
200	1.05	1.12
300	1.04	1.06
400	1.02	1.06
500	1.02	1.05
700	1.01	1.05



Σχήμα 11: Αποτελέσματα MAE για το δυαδικό C4.5 και Multi - C4.5



Σχήμα 12: Αποτελέσματα RMSE για το δυαδικό C4.5 και Multi - C4.5

Στο Σχήμα 10, η καμπύλη C4.5 (2 cls) ή C4.5 (2 classes) εκφράζει την εφαρμογή του C4.5 κατηγοριοποιητή για την παραγωγή ενός μοντέλου με δύο πιθανές προβλεπόμενες κλάσεις. Το Multi - C4.5 (4 cls) ή C4.5 (4 classes) εκφράζει την εφαρμογή του ίδιου κατηγοριοποιητή για την παραγωγή ενός μοντέλου με τέσσερις πιθανές προβλεπόμενες κλάσεις. Παρατηρούμε ότι η C4.5 (4 cls) παρουσιάζει καλύτερα αποτελέσματα από την Multi – C4.5 (4 cls) πρώτη. Δηλαδή η περίπτωση το να έχουμε δύο κλάσεις είναι καλύτερη από το να έχουμε τέσσερις κλάσεις. Στην πρώτη περίπτωση οι κλάσεις είναι πιο πυκνές αφού είναι λίγες και οι γείτονες του νέου χρήστη που ανήκουν σε κάθε κλάση είναι πιο πολλοί με αποτέλεμα η πρόβλεψη να είναι και πιο αξιόπιστη. Σε αντίθεση με την δεύτερη περίπτωση που οι κλάσεις είναι πιο αραιές αφού το σύνολο των χρηστών κατανείμεται σε πολλές διαφορετικές κατηγορίες τις οποίες ορίζει το μοντέλο. Η συγκεκριμένη προσέγγιση μπορεί εύκολα να επεκταθεί και να εκτελεστούν πειράματα με περισσότερες από τέσσερις κλάσεις.

Το βέλτιστο σενάριο είναι το 4^{ο} στο οποίο για την πρόβλεψη δίνουμε μεγαλύτερη έμφαση στην ομοιότητα των χρηστών που προκύπτει εάν συγκρίνουμε τις ηλικίες τους. Η διαφορά αυτού του σεναρίου από τα υπόλοιπα είναι ότι σε αυτό οι κατηγοριοποιητές εμφανίζουν πολύ χαμηλά ποσοστά λαθών με αποτελέσμα οι προβλεπόμενες

αξιολογήσεις να είναι πολύ κοντά στις πραγματικές. Αξίζει επίσης να σημειώσουμε ότι η τιμή του α δεν έχει ουσιαστική επιρροή στην απόδοση του προτεινόμενου αλγορίθμου.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

6.1 Συμπεράσματα

Στη παρούσα διπλωματική εργασία αναπτύχθηκε μία καινοτόμα τεχνική η οποία επιλύει με ιδιαίτερη επιτυχία το πρόβλημα της ψυχρής εκκίνησης που παρουσιάζουν τα συστήματα συστάσεων. Η προτεινόμενη τεχνική ενσωματώνει τη μέθοδο της κατηγοριοποίησης σε ένα παραδοσιακό συνεργατικό φιλτράρισμα αξιοποιώντας τα δημογραφικά δεδομένα του νέου χρήστη. Η ιδέα στην οποία βασιζόμαστε είναι ότι άτομα με τα ίδια χαρακτηριστικά πιθανόν να έχουν και τις ίδιες προτιμήσεις. Αρχικά, με τη βοήθεια ενός κατηγοριοποιητή τοποθετούμε το νέο χρήστη σε μία κατηγορία και έπειτα κάνουμε εκτίμηση της βαθμολογίας του για ένα αντικείμενο βάσει των βαθμολογιών των χρηστών που ανήκουν στην συγκεκριμένη κατηγορία. Για τη συνεργατική πρόβλεψη χρησιμοποιούμε μία συνάρτηση ομοιότητας, η οποία συνδυάζει τους δείκτες ομοιότητας για καθένα από τα δημογραφικά δεδομένα. Μία σημαντική παρατήρηση είναι πως το στάδιο της κατηγοριοποίησης παίζει θετικό ρόλο στην εκτίμηση του τελικού αποτελέσματος, και ως αποτέλεσμα στην ακρίβεια του προτεινόμενου αλγορίθμου. Προτείνουμε, με τη συγκεκριμένη τεχνική, να χρησιμοποιείται πάντα ένας κατηγοριοποιητής ο οποίος θα παράγει ένα μοντέλο βάσει των δημογραφικών δεδομένων. Αυτό προκύπτει από τη σύγκριση των αλγορίθμων που υλοποιήσαμε με έναν αλγόριθμο αναφοράς, τον RCA, ο οποίος δεν χρησιμοποιεί κατηγοριοποιητή αλλά επιλέγει τυχαία την κατηγορία στην οποία ανήκει ο νέος χρήστης. Ο RCA δεν παρουσίασε ικανοποιητικά αποτελέσματα και για αυτό καταλήξαμε στο συμπέρασμα ότι οι κατηγοριοποιητές Multi – C 4.5 και Multi - Naïve Bayes είναι περισσότερο αποδοτικοί για τη προτεινόμενη τεχνική. Αξίζει να σημειώσουμε ότι ο αποδοτικότερος αλγόριθμος είναι ο Naïve Bayes. Επίσης, ιδιαίτερο ενδιαφέρον έχουν τα πειράματα τα οποία συγκρίνουν τις μετρικές απόδοσης με βάση τον αριθμό των κλάσεων που παράγει ο ίδιος αλγόριθμος, στην περίπτωση μας ο C4.5. Όπως αποδείξαμε, είναι προτιμότερο στην προτεινόμενη τεχνική να χρησιμοποιήσουμε τον C4.5 ως δυαδικό κατηγοριοποιητή.

6.2 Μελλοντικές Προεκτάσεις

Στα πλαίσια της παρούσας εργασίας, παρουσιάσαμε έναν αλγόριθμο, ο οποίος επιλύει το πρόβλημα της ψυχρής εκκίνησης για νέους χρήστες με τη χρήση κατηγοριοποιητών. Από αυτούς δοκιμάσαμε δύο κατηγοριοποιητές, τους C4.5 και Naïve Bayes. Μία μελλοντική επέκταση της προσέγγισης είναι η δοκιμή περισσότερων κατηγοριοποιητών (όπως k - nearest neighbor, Bayesian networks), για να αποφανθεί ποιος είναι ο βέλτιστος για την συγκεκριμένη μέθοδο. Για όλους αυτούς τους κατηγοριοποιητές η είσοδος θα είναι τα δημογραφικά δεδομένα. Για την παρούσα εργασία λάβαμε υπόψη μόνο τρία είδη δημογραφικών στοιχείων. Μεγάλο ενδιαφέρον θα είχε η επέκταση αυτών των δεδομένων για να ερευνήσουμε την επίδρασή τους στη δημιουργία του μοντέλου καθώς και στο τελικό αποτέλεσμα.

Άλλη μία ενδιαφέρουσα εξέλιξη είναι η σύγκριση της παρούσας τεχνικής με άλλες τεχνικές που επιλύουν με διαφορετικό τρόπσο το πρόβλημα της ψυχρής εκκίνησης για νέους χρήστες σε ένα σύστημα συστάσεων [25, 31]. Τέλος, θα μπορούσαμε να ερευνήσουμε το πως η προτεινόμενη τεχνική θα επιλύει το πρόβλημα της ψυχρής εκκίνησης όταν στο σύστημα εμφανίζονται νέα αντικείμενα που είναι συμμετρικό του προβλήματος των νέων χρηστών.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός όρος**
Information Overload	Υπερπληροφόρηση
Recommender Systems	Συστήματα Συστάσεων
Ratings	Βαθμολογίες
E - Commerce	Ηλεκτρονικό Εμπόριο
Collaborative Filtering	Συνεργατικό Φιλτράρισμα
Content Based Filtering	Φιλτράρισμα βασισμένο στο περιεχόμενο
Memory	Μνήμη
Model	Μοντέλο
User	Χρήστης
Distance	Απόσταση
Similarity Coefficient	Συντελεστής Ομοιότητας
Cosine Similarity	Ομοιότητα Συνημιτόνου
Adjusted	Προσαρμοσμένος
Similarity	Ομοιότητα
Item - ratings Matrix	Πίνακας με Αντικείμενα και Βαθμολογίες
Group - ratings Matrix	Πίνακας με Χρήστες και Βαθμολογίες
Probalistic Algorithms	Πιθανοτικός Αλγόριθμος
Clustering	Συσταδοποίηση
Fuzzy	Ασαφές
Information Retrieval	Ανάκτηση Πληροφορίας
Artificial Intelligence	Τεχνητή Νοημοσύνη
Text Comments	Σχόλια Κειμένου
Artificial Neural Network	Τεχνητά Νευρωνικά Δίκτυα
Intelligent Agent	Έξυπνος Πράκτορας
Hybrid Recommendation Methods	Υβριδικές Μέθοδοι Συστάσεων
Weighted	Σταθμισμένη
Switching	Μεταβατική
Feature Combination	Συνδυασμός Χαρακτηριστικών
Meta - level	Μετά - Επίπεδο
Cold Start problem	Προβλημα Ψυχρής Εκκίνησης
New User	Νέος Χρήστης
New Item	Νέο Αντικείμενο
Association Rules	Κανόνες Συσχέτισης
Over Specialization	Υπερ - Εξειδίκευση
Classification	Κατηγοριοποίηση
Data Mining	Εξόρυξη Δεδομένων
Target Category	Κατηγορία Στόχος
Computer Vision	Μηχανική Όραση
Pattern Recognition	Αναγνώριση Προτύπων
Speech Recognition	Αναγνώριση Φωνής
Geostatistics	Γεωστατική
Training	Εκπαίδευση

Classifier	Κατηγοριοποιητής
Supervised Learning	Μέθοδος Εποπτευόμενης Μάθησης
Binary	Δυαδικός
MultiClass	Πολλαπλές Κλάσεις
Decision Trees	Δέντρα Απόφασης
Pruning	Κλάδεμα - Βελτιστοποίηση
Prediction Variable	Μεταβλητή Πρόβλεψης
Estimated	Εκτιμώμενη
Neighbors	Γείτονες
Weights Estimation	Υπολογισμός Βαρών
Intellectual	Διανοούμενος
Adventurous	Περιπετειώδης
Fun	Διασκεδαστικός
Romantic	Ρομαντικός
Property	Χαρακτηριστικό

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

ΣΦ	Συνεργατικό Φιλτράρισμα
ΦΒΠ	Φιλτράρισμα Βάση Περιεχομένου
KNN	K Nearest Neighbors
OvA	One Against All
AvA	All Against All
OvO	One Against One
OvA	One Against All
CLS	Hunt’s Concept Learning Systems
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

ΑΝΑΦΟΡΕΣ

[1] F. Ricci, L. Rokach, and B. Shapira, *Recommender System Handbook*, Springer 2011.

[2] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Riedl, GroupLens: Applying collaborative filtering to usenet news*, Communications of the ACM*, vol. 40(3), 1997, pp. 77–87.

[3] U. Shardanand, P. Maes, "Social information filtering: Algorithms for automating "word of mouth"", *ACM Conference on Human Factors in Computing Systems*, 1995, pp. 210–217.

[4] B.M. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Item-based collaborative filtering recommendation algorithm", *10th International World Wide Web Conference*, 2001.

[5] Linden, G., Smith, B., York, J: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003, pp. 76–80.

[6] M. Deshpande, G. Karypis, "Item-based top-N recommendation algorithms", *ACM Transactions on Information Systems*, 2004, pp. 143–177.

[7] Q. Li, B.M. Kim, "Clustering approach for Hybrid Recommender System", *WI '03 Proceedings of the IEEE/WIC International Conference on Web Intelligence*, 2003

[8] L. Candillier, F. Meyer, F. Fessant," Designing Specific Weighted Similarity Measures to Improve Collaborative Filtering Systems", *IEEE International Conference on Data Mining (ICDM)*, 2008, pp. 242-255.

[9] J.L. Herlocker, J.A. Konstan, J. Riedl, "An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms", *Information Retrieval*, 2002, pp. 287-310.

[10] L. H. Ungar and D. P. Foster, "Clustering Methods for Collaborative Filtering", *In Proc. Workshop on Recommendation Systems at the 15th National Conf. On Artificial Intelligence*, 1998.

[11] J.S. Breese, D. Heckerman, C. Kadie, Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Technical Report, Microsoft Research, May 1998.

[12] D.M. Pennock, E.Horvitz, S.Lawrence, C.L.Giles, "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory and Model-Based Approach", *Uncertainty in Artificial Intelligence(UAI)*, 2000, pp. 473-480.

[13] B. Sarwar, G. Karypis, J. Konstan, and J.Riedl, "Item-based collaborative filtering recommendation algorithms", *In WWW '01: Proceedings of the 10th international conference on World Wide Web*, 2001 pp. 285–295.

[14] 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, 2005, pp. 734 -749

[15] M. Pazzani and D. Billsus, Learning and Revising User Profiles: The Identification of Interesting Web Sites, Machine Learning, vol. 27, 1997, pp. 313-331

[16] R.J. Mooney, P.N. Bennett, and L. Roy, Book Recommending Using Text Categorization with Extracted Information, Technical Report, 1998.

[17] R.J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization", *Proceedings of the 5th ACM Conference on Digital Libraries,* pp. 195–204.

[18] K.D Bollacker and C.L Giles, "CiteSeer: An AutonomousWeb Agent for Automatic Retrieval and Identification of Interesting Publications", *Proc. Of the 2nd International Conference of Autonomous Agents*, 1998, pp. 116-123.

[19] L. Adrissono, A. Goy, G. Petrone, M. Segnan, P. Torasso, "Intrigue: Personalized recommendation of tourist attractions for desktop and handheld devices", *Applied Artificial Intelligence*, 2003, pp. 687–714.

[20] R.D. Burke, Hybrid Recommender Systems: Survey and Experiments, *User Model. User-Adapt. Interact.* vol. 12, p. 4, pp. 331-370.

[21] D. Billsus and M. Pazzani, User Modeling for Adaptive News Access, **User Modeling and User-Adapted Interaction** (UMUAI), vol. 10 , p. 2, 2000, pp. 147-180

[22] F. Ricci, A. Venturini, D. Cavada, N. Mirzadeh, D. Blaas, and M. Nones, "Product Recommendation with Interactive Query Management and twofolSimilarity", *Proc. of the 5th international conference on Case-based reasoning: Research and Development,* pp. 479-493.

[23] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations", *Proc. of the 18th National Conference on Artificial Intelligence (AAAI)*, 2002, pp. 187-192.

[24] A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan, and J.Riedl, "Learning new user preferences in recommender systems", Proc. Of the 7th International Conference on Intelligent User Interface, pp. 127–134.

[25] N. Golbandi, Y. Koren, and R. Lempel," Adaptive bootstrapping of recommender systems using decision trees, Proc. of the fourth ACM international conference on Web search and data mining, pp. 595-604.

[26] Paolo Massa and Bobby Bhattacharjee, Using trust in recommender systems: An experimental analysis, *In Proc. of iTrust2004 International Conference,* 2004, pp. 221–235.

[27] Patricia Victor, Chris Cornelis, Martine De Cock, and Ankur M. Teredesai, "Key figure impact in trust-enhanced recommender systems", *AI Commun*, 2008, pp .127–143,

[28] X. N. Lam, T. Vu, T.D. LE, A. D. Duong ,"Addressing cold-start problem in recommendation systems*", In Proc. of the 2nd international conference on Ubiquitous information management and communication*, 2008, pp. 208-211.

[29] Seung-Taek Park and Wei Chu, "Pairwise preference regression for cold-start recommendation*". In RecSys '09: Proceedings of the third ACM conference on Recommender systems,* 2009, pp. 21–28.

[30] Seung-Taek Park and Wei Chu,"Personalized recommendation on dynamic content using predictive bilinear models", In Proc. of the 18th international Conference on World Wide Web, 2009, pp. 691-700.

[31] G. Shaw, Y. Xu, and S. Geva, Using association rules to solve the cold-start problem in recommender systems, PAKDD, 2010, vol. 1 pp. 340-347.

[32] Zi-Ke Zhang, Chuang Liu, Yi-Cheng Zhang, Tao Zhou: Solving the Cold-Start Problem in Recommender Systems with Social Tags**,** *CoRR abs*, 2010.

[33] R. Rifkin, Multiclass Classification, MIT lecture, Springer 2008; http://www.mit.edu/~9.520/spring09/Classes/multiclass.pdf [Προσπελάστηκε 10/05/2013].

[34] J. R. Quinlan, Induction of decision trees, "*Machine Learning*", 1986, vol. 1(1), pp. 81-106.

[35] Hunt, E. B., Marin, J., and Stone, "P. J. Experiments in induction", New York: Academic Press, 1966.

[36] J. R. Quinlan, C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers*, 1993.

[37] S. Theodoridis and K. Koutroumbas, *Pattern Recognition: Theory and Application*", 4rth edition, Academic Press, 2009.

[38] Papoulis A. Probability, "Bayes Theorem in Statistics Random variables and Stochastic Processes" 2nd New York: McGraw Hill, 1984, pp. 38–39**.**

[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, *The WEKA data mining software: an update*, ACM SIGKDD Explorations Newsletter, 2009, vol. 11(1), pp. 10-18.

[40] C. Corley and R. Mihalcea, "Measuring the Semantic Similarity of Texts", *In Proc. of the ACM Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, June 2005, pp. 13-18.

[41] George A. Miller, WordNet: A large lexical database of English; http://wordnet.princeton.edu/ [Προσπελάστηκε 14/05/2013].

[42] J. L. Herlocker, J. A. Konstant, L. G. Terveen, J. T. Riedl, Evaluating collaborative filtering recommender Systems, ACM Transactions on Information Systems (TOIS), vol. 22(1), 2004, pp. 5-53.

[43] GroupLens Research Group; http://www.grouplens.org/node/73, [Προσπελάστηκε 17/05/2013].

[44] G. Xue, C. Lin, Q. Yang, W. Xi, H. Zeng, Y. Yu, and Z. Chen, "Scalable collaborative filtering using cluster - based smoothing", *In the Proc. of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 114 -121.

[45] Q. Li and B. M. Kim, "Clustering Approach for Hybrid Recommender Systems", *In the Proc. Of the 2003 IEEE/WIC International Conference on Web Intelligance*, pp. 33