ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία Συστήματος Συστάσεων Βασισμένο σε Χωροχρονικές Πληροφορίες

Γεώργιος Κ. Λιάπατας

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

ΑΘΗΝΑ

Νοέμβριος 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία Συστήματος Συστάσεων Βασισμένο σε Χωροχρονικές Πληροφορίες

Γεώργιος Κ. Λιάπατας
Α.Μ.: M1271

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

Νοέμβριος 2016

ΠΕΡΙΛΗΨΗ

Ο ρόλος ενός συστήματος συστάσεων είναι η παροχή στους χρήστες εξατομικευμένων προτάσεων. Σκοπός της παρούσας εργασίας είναι η δημιουργία ενός συστήματος το οποίο θα παράγει ως έξοδο συστάσεις, με τη βοήθεια χρονικής πληροφορίας (διαφορά της χρονικής στιγμής βαθμολόγησης από αυτή της σύστασης) και της γεωγραφικής θέσης του χρήστη. Αρχικά το σύστημα κατηγοριοποιεί το χρήστη, σε ένα πεπερασμένο αριθμό κλάσεων, με βάση δημογραφικά δεδομένα. Στη συνέχεια, υπολογίζει ένα βάρος για κάθε βαθμολογία που έχουν πραγματοποιήσει οι χρήστες οι οποίοι ανήκουν στην ίδια κατηγορία με τον πελάτη που θα δεχτεί τις συστάσεις. Στο επόμενο βήμα (συσταδοποίηση), δημιουργεί μία συστάδα με προϊόντα τα οποία πιθανόν θα ενδιέφεραν το χρήστη και μία δεύτερη με προϊόντα που ενδεχομένως θα του ήταν αδιάφορα. Χρησιμοποιώντας την πρώτη συστάδα, δημιουργεί κάποιες προβλέψεις βαθμολογίας, βάσει των οποίων εξάγονται οι τελικές συστάσεις προς το χρήστη, ενώ δε λαμβάνει υπόψη του τη δεύτερη. Για την υλοποίηση του συστήματος χρησιμοποιήθηκε το εργαλείο εξόρυξης δεδομένων WEKA, και πιο συγκεκριμένα οι αλγόριθμοι κατηγοριοποίησης Naïve Bayes, C4.5 και ο αλγόριθμος συσταδοποίησης Expectation Maximization. Η αξιολόγηση και τα αποτελέσματα των πειραμάτων αποδεικνύουν την ορθότητα της μεθοδολογίας καθώς και την αποδοτικότητα του προτεινόμενου συστήματος. Τέλος, γίνεται αποτίμηση του αλγορίθμου με βάση διαφορετικά σενάρια υπολογισμού των βαρών και παράθεση μετρικών σφάλματος για κάθε σενάριο ξεχωριστά.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Συστήματα Συστάσεων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Συστάσεις, Κατηγοριοποίηση, Ομαδοποίηση, Υπολογισμός βαρών, Πρόβλεψη βαθμολογίας

ABSTRACT

The role of a recommendation system is to provide to users personalized recommendations. The purpose of this thesis is the creation of a recommender system which produces as output recommendations on top of the geographical location of each user as well as temporal information. Initially, the system classifies the user, based on demographical characteristics, to a finite number of classes. Then calculates a weight for each rating, that users, of the same class, have made. In the next step, it creates a cluster (clustering) with products which may interest the user and a second cluster with products that, probably, they are not of interest of the user. Based on the first cluster, it creates some rating predictions based on which the final recommendations are exported. We adopt the WEKA data mining tool for the implementation of the system, and more specifically the classification algorithms Naïve Bayes and C4.5. The adopted clustering algorithm is the Expectation Maximization. We provide specific simulations and the evaluation of the proposed system. Our experimental results reveal the advantages and the efficiency of the proposed algorithm.

SUBJECT AREA: Recommender Systems

KEYWORDS: Recommendations, Classification, Clustering, Weight estimation, Rating prediction

Στην οικογένεια μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα της διπλωματικής εργασίας αναπληρωτή καθηγητή κ. Ευστάθιο Χατζιευθυμιάδη που μου έδωσε την ευκαιρία να ασχοληθώ με την εκπόνηση της παρούσας εργασίας. Επίσης είμαι ευγνώμων στον Κωνσταντίνο Κολομβάτσο για την πολύτιμη καθοδήγηση και βοήθεια του κατά τη διάρκεια της δουλειάς μου. Θα ήθελα, επιπλέον, να τον ευχαριστήσω για την άμεση ανταπόκριση και υποστήριξη του, όσες φορές και αν χρειάστηκε. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου, για την αμέριστη συμπαράσταση και στήριξη.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα εργασία εκπονήθηκε στα πλαίσια της απόκτησης του μεταπτυχιακού τίτλου σπουδών με τίτλο «Συστήματα Επικοινωνιών και Δίκτυα» του Τμήματος πληροφορικής και Τηλεπικοινωνιών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών. Η εργασία εκπονήθηκε στην Αθήνα κατά το έτος 2016, ενώ το αποτέλεσμα αυτής είναι δημιούργημα έρευνας αρκετών μηνών που είχε ως αντικείμενο τη δημιουργία ενός συστήματος συστάσεων που παράγει αποτελέσματα με τη βοήθεια χωροχρονικής πληροφορίας.

1. ΕΙΣΑΓΩΓΗ

1.1 Ιστορικό

Τις τελευταίες δεκαετίες το διαδίκτυο έχει εισέλθει στη ζωή μας και τείνει να την αλλάξει ολοκληρωτικά. Με την εξέλιξη της τεχνολογίας και τη δημιουργία νέων συσκευών, ταχύτερων, μικρότερων με μεγαλύτερη αυτονομία και οικονομικά πιο προσιτών, η χρήση του Διαδικτύου αγγίζει κάθε πτυχή της καθημερινότητας μας. Επηρεάζει, πλέον, τον τρόπο που σκεφτόμαστε, που ενεργούμε, βελτιώνει και απλουστεύει διαδικασίες και καταστάσεις. Αποτελεί, χωρίς υπερβολή, το πολυτιμότερο εργαλείο κάθε ανθρώπου και σίγουρα μία μόνιμη και αστείρευτη πηγή πληροφορίας. Ο όγκος της υπάρχουσας πληροφορίας αυξάνεται συνεχώς και με πολύ γρήγορο ρυθμό, με αποτέλεσμα να πολλαπλασιάζεται ο αριθμός των προσφερόμενων επιλογών. Παράλληλα, αυξάνει και ο φόρτος που απαιτείται από το χρήστη για την επιλογή της επιθυμητής πληροφορίας, γεγονός που δυσκολεύει την αναζήτηση και απόκτηση δεδομένων. Για το λόγο αυτό, κρίνεται απαραίτητη η υλοποίηση τεχνικών που θα προσφέρουν την επιθυμητή πληροφορία στο χρήστη, χωρίς αυτός να σπαταλά την ώρα του ψάχνοντας μέσα σε έναν αχανές όγκο δεδομένων. Οι τεχνικές αυτές προσφέρουν εξατομικευμένες υπηρεσίες προσμετρώντας τις επιθυμίες και τα «θέλω» του χρήστη ενώ εμπλέκοντας μία μορφή τεχνητής νοημοσύνης, φιλτράρουν τις πληροφορίες και παρουσιάζουν τα τελικά αποτελέσματα.

Σήμερα, οι τεχνικές αυτές χρησιμοποιούνται σε μία σειρά από εφαρμογές, ενώ πολλές φορές λειτουργούν διαφανώς, χωρίς, δηλαδή, ο χρήστης να καταλαβαίνει τη λειτουργία τους. Τέτοιες τεχνικές συναντάμε σε μία πληθώρα τομέων, κυριότερους από τους οποίους είναι τα ηλεκτρονικά καταστήματα (e-shop), οι μηχανές αναζήτησης ακόμα και η ηλεκτρονική αλληλογραφία. Στο κυρίως μέρος της εργασίας γίνεται αναφορά στην υλοποίηση ενός συστήματος που αφορά ένα ηλεκτρονικό κατάστημα.

1.2 Τι είναι Σύστημα Συστάσεων

Τα συστήματα που έχουν τα παραπάνω χαρακτηριστικά ονομάζονται **συστήματα συστάσεων** (Recommender System, RS) και λειτουργούν εφαρμόζοντας εξατομικευμένες τεχνικές σε μία προσπάθεια να προσφέρουν στους πελάτες προϊόντα, υπηρεσίες ή πληροφορίες που ικανοποιούν τις ανάγκες τους. Ουσιαστικά, ένα σύστημα συστάσεων είναι ένα σύστημα που προτείνει προϊόντα στους πελάτες και τους παρέχει επαρκή ενημέρωση ώστε να είναι σε θέση να επιλέξουν ποια προϊόντα επιθυμούν να αγοράσουν. Ένα αποδοτικό σύστημα συστάσεων συμβάλλει στην ικανοποίηση του πελάτη καθώς βελτιώνει την εμπειρία του και αυξάνει την εμπιστοσύνη του προς το σύστημα.

Κάθε σύστημα συστάσεων δημιουργείται και ακολουθεί ένα καθορισμένο αλγόριθμο βάσει του οποίου συστήνει προϊόντα. Έτσι τα προϊόντα που προτείνει μπορεί να βασίζονται στις συνολικές πωλήσεις του καταστήματος ή ακόμα και στα δημογραφικά χαρακτηριστικά κάθε πελάτη. Η απόφαση μπορεί, όμως, να προκύπτει και με πιο σύνθετο τρόπο, απορρέοντας δηλαδή από την ανάλυση της προηγούμενης αγοραστικής συμπεριφοράς του πελάτη ως πρόβλεψη για τη μελλοντική του συμπεριφορά ή με τη βοήθεια στοιχείων από κοινότητες πελατών που έχουν παρόμοια χαρακτηριστικά και κοινή αγοραστική συμπεριφορά.

Εκτός από τον τρόπο που δημιουργούνται οι συστάσεις, παρουσιάζονται διαφορές και στη μορφή των συστάσεων αυτών από σύστημα σε σύστημα. Πιο συγκεκριμένα, άλλα συστήματα περιορίζονται σε απλή πρόταση προϊόντων προς τους πελάτες, ενώ, άλλα προσφέρουν εξατομικευμένες πληροφορίες για τα προϊόντα, γενικές απόψεις μίας κοινότητας πελατών με κοινά χαρακτηριστικά ή ακόμα και εξατομικευμένες κριτικές από πελάτες ή κοινότητες πελατών. Έστω για παράδειγμα ένα ηλεκτρονικό κατάστημα (e-

shop) το οποίο πουλά βιβλία και ανάμεσα σε άλλους έχει πελάτες τον Α και τον Β. Ο Α έχει δηλώσει πως προτιμά συγκεκριμένο βιβλίο επιστημονικής φαντασίας, έστω το Χ, και αγοράζει ένα άλλο βιβλίο, έστω το Υ, επίσης επιστημονικής φαντασίας. Αν ο Β αναζητήσει το Χ, μία πιθανή ενέργεια του συστήματος συστάσεων είναι να προτείνει στον Β το Υ (που αγόρασε ο πελάτης Α) με το σκεπτικό ότι θα ήταν αρκετά πιθανό ο Β να έχει τις ίδιες παραπλήσιες προτιμήσεις με τον Α.

1.3 Λόγος χρησιμοποίησης των Συστημάτων Συστάσεων

Ο βασικός λόγος χρησιμοποίησης ενός συστήματος συστάσεων σε ένα ηλεκτρονικό κατάστημα είναι η ενίσχυση των πωλήσεων του. Έρευνες έχουν δείξει ότι ένα ηλεκτρονικό κατάστημα το οποίο χρησιμοποιεί RS έχει περισσότερες πιθανότητες να μετατρέψει ένα περιηγητή σε αγοραστή, σε αντίθεση με κάποιο άλλο που δε χρησιμοποιεί. Είναι γεγονός πως οι περισσότεροι χρήστες που εισέρχονται σε ένα e-shop συχνά απλώς περιηγούνται σε αυτό χωρίς να αγοράζουν κάποιο προϊόν του καταστήματος. Εδώ ξεκινά η λειτουργία του RS το οποίο με βάση όσα έχει καταφέρει να συλλέξει για το χρήστη, προτείνει προϊόντα τα οποία είναι πιθανό να τον ενδιαφέρουν και έτσι αυξάνεται η πιθανότητα αγοράς. Η απόδοση του RS εξαρτάται από την επιτυχία αυτών των προτάσεων.

Επιπρόσθετα, ένας ακόμα λόγος χρησιμοποίησης του RS είναι η αύξηση του cross-selling στο κατάστημα. Ο όρος cross-selling ή πρόσθετη πώληση υποδηλώνει όλες τις παράλληλες πωλήσεις που σχετίζονται με μία δεδομένη αγορά. Για παράδειγμα, αν κάποιος πάει σε ένα κατάστημα ρούχων και αγοράσει ένα πουκάμισο, cross-selling θεωρείται το να πειστεί από την πωλήτρια να αγοράσει μανικετόκουμπα και γραβάτα που ταιριάζουν με το πουκάμισο αυτό. Στη συνέχεια, η χρήση του RS οδηγεί σε αύξηση και του up-selling. Μία αγορά χαρακτηρίζεται με τον όρο up-selling αν ο αγοραστής έχει στο μυαλό του ένα συγκεκριμένο προϊόν και πειστεί να αγοράσει κάποιο παρόμοιο ανώτερης ποιότητας και τιμής.

Ένας άλλος λόγος χρησιμοποίησης είναι η σταδιακή βελτίωση της εμπιστοσύνης των πελατών προς το κατάστημα. Καθώς ο αριθμός των e-shops αυξάνεται πολύ γρήγορα, είναι πολύ σημαντικό ένα κατάστημα να έχει σταθερούς πελάτες που να το εμπιστεύονται και να αποκτά μία καλή φήμη. Η εμπιστοσύνη αυτή που χτίζεται αργά αλλά σταθερά, με τη βοήθεια του συστήματος συστάσεων, εξαργυρώνεται καθώς είναι βέβαιο πως ένας πελάτης γυρνά ξανά σε ένα ηλεκτρονικό κατάστημα που τον ικανοποιεί. Έτσι όσο περισσότερες φορές επιστρέψει ο πελάτης και όσο περισσότερο έρχεται σε επαφή με το σύστημα που φέρει χαρακτηριστικά αυτό-εκμάθησης, τόσο καλύτερες θα είναι οι υπηρεσίες (συστάσεις που θα λαμβάνει) και ταυτόχρονα τόσο θα αυξάνεται η εμπιστοσύνη του προς αυτό.

1.4 Τρόπος λειτουργίας

Ένα σύστημα συστάσεων λειτουργεί όπως ακριβώς ορίζει η έννοια του συστήματος. Πιο συγκεκριμένα, ένα σύστημα καθορίζεται από τις εισόδους, τις εξόδους και τις διαδικασίες επεξεργασίας των δεδομένων εισόδου βάσει των οποίων προκύπτουν οι έξοδοι. Ακριβώς με αυτόν τον τρόπο λειτουργεί και ένα σύστημα συστάσεων. Δέχεται μία συλλογή από δεδομένα εισόδου που παίρνουν τη μορφή προτιμήσεων των πελατών και χαρακτηριστικών των προϊόντων μαζί με μία σειρά συσχετίσεων και στη συνέχεια το σύστημα χρησιμοποιεί αυτά τα δεδομένα για να παράγει τις προτάσεις προϊόντων. Τα δεδομένα αυτά μπορούμε να τα χωρίσουμε σε δύο μεγάλες κατηγορίες ανάλογα με την προέλευση τους. Η πρώτη είναι τα δεδομένα που είναι σχετικά με τον πελάτη ο οποίος δέχεται τις συστάσεις, ενώ στη δεύτερη κατηγορία εντάσσονται τα δεδομένα εκείνα που είναι σχετικά με την κοινότητα άλλων πελατών.

Η τάση που επικρατεί σήμερα στα συστήματα συστάσεων είναι τα δεδομένα που απορροφά το σύστημα να προέρχονται από την περιήγηση του χρήστη στην ιστοσελίδα, έτσι ώστε το σύστημα να ανταποκρίνεται στην παρούσα κατάσταση και στις επιθυμίες του. Έτσι ο τρόπος που συμπεριφέρεται ο χρήστης μέσα στην ιστοσελίδα του ηλεκτρονικού καταστήματος μετατρέπεται σε είσοδο του συστήματος. Η είσοδος των πληροφοριών, λοιπόν, θα μπορούσε να γίνεται με σχετική διαφάνεια ως προς το χρήστη, χωρίς δηλαδή εκείνος να το γνωρίζει. Αυτό θα μπορούσε να συμβεί καταγράφοντας τις κινήσεις και την πλοήγηση του χρήστη στα διάφορα προϊόντα. Για παράδειγμα, θα μπορούσε να καταγραφεί ως είσοδος το γεγονός ότι ο επισκέπτης ενός καταστήματος ενδιαφέρθηκε για ένα συγκεκριμένο βιβλίο επιστημονικής φαντασίας και επισκέφτηκε τη σελίδα της περιγραφής του. Αντίθετα, η είσοδος του συστήματος μπορεί να είναι «προκαλούμενη» από το χρήστη. Δηλαδή ο ίδιος ο επισκέπτης του ηλεκτρονικού καταστήματος θα μπορούσε να συμπληρώσει κάποια φόρμα ή ερωτηματολόγιο ή ακόμα και να βαθμολογήσει κάποια προϊόντα. Η παραπάνω κατάσταση κατά την οποία ο ίδιος ο χρήστης διαμορφώνει την είσοδο του συστήματος συνηθίζεται να αποκαλείται «σαφής πλοήγηση».

Υπάρχουν μία σειρά από μεθοδολογίες που χρησιμοποιούνται στα σύστημα συστάσεων και «προδίδουν» τις προτιμήσεις των χρηστών. Πέραν των ερωτηματολογίων και της βαθμολόγησης που αναφέρθηκαν πιο πάνω, ένας ακόμα τρόπος είναι οι λέξεις κλειδιά. Πιθανότατα μέσω κάποιας μηχανής αναζήτησης, ο χρήστης πληκτρολογεί τις προτιμήσεις του ή τα χαρακτηριστικά κάποιου επιθυμητού προϊόντος με τη μορφή μεμονωμένων λέξεων κλειδιών. Στη συνέχεια, το σύστημα χρησιμοποιεί τα δεδομένα αυτά ως είσοδο για να παράγει τις δικές του συστάσεις. Επιπρόσθετα, το ρόλο της εισόδου του συστήματος θα μπορούσε να παίξει το ιστορικό των αγορών του χρήστη. Λαμβάνοντας υπόψη την εμπιστοσύνη του χρήστη σε ένα προϊόν το ιστορικό υπό προϋποθέσεις αποτελεί μία μορφή βαθμολόγησης.

Από την άλλη μεριά, υπάρχουν και τα δεδομένα που είναι προσανατολισμένα σε κοινότητες πελατών. Τέτοιου είδους δεδομένα, δίνουν συλλογικές πληροφορίες για το πως αντιλαμβάνεται μία ομάδα πελατών συγκεκριμένα προϊόντα. Ένας τρόπος λήψης τέτοιων πληροφοριών θα μπορούσε να είναι τα σχόλια που κάνει κάθε πελάτης κάτω από ένα προϊόν. Αν και οι πληροφορίες που μπορεί να εξαχθούν από τέτοιου είδους δεδομένα είναι πολύ σημαντικές, ο χρόνος επεξεργασίας τους είναι απαγορευτικός και τα καθιστούν δύσχρηστα.

Ολοκληρώνοντας την περιγραφή του συστήματος ως οντότητα, το τελευταίο του κομμάτι είναι η έξοδος, η οποία εξαρτάται από την ποσότητα και την ποιότητα των εισόδων που έχει λάβει. Η έξοδος ενός συστήματος συστάσεων μπορεί να πάρει τη μορφή προβλέψεων, συστάσεων ακόμα και βαθμολογιών. Ο πιο διαδεδομένος τύπος είναι οι προτάσεις που έχουν συχνά τη μορφή προτροπής προς το χρήστη ώστε να δοκιμάσει ένα συγκεκριμένο προϊόν. Τέλος, υπάρχουν αλγόριθμοι που παρουσιάζουν τη βαθμολογία που θα έδινε κάποιος χρήστης σε προϊόντα (προβλέψεις).

1.5 Σύστημα συστάσεων βασισμένο σε χωροχρονικές πληροφορίες

Η παρούσα εργασία πραγματεύεται τη δημιουργία ενός συστήματος με παρόμοια χαρακτηριστικά με αυτά των συστημάτων που περιγράφηκαν παραπάνω με τη μόνη διαφορά ότι εμπλέκει τη μεταβλητή της γεωγραφικής θέσης σε συνδυασμό με χρονικές πληροφορίες ώστε να φτάσει στο αποτέλεσμα των συστάσεων. Για την πραγματοποίηση του όλου εγχειρήματος, το σύστημα είναι δυνατόν να διαιρεθεί σε υποσυστήματα το καθένα από τα οποία εκτελεί μία ξεχωριστή λειτουργικότητα, ενώ όλα μαζί σε συνδυασμό αποφέρουν το τελικό αποτέλεσμα.

Πιο συγκεκριμένα, ένας χρήστης που εισέρχεται στο ηλεκτρονικό κατάστημα και δηλώνει κάποια στοιχεία, όπως φύλο και ηλικία, κατατάσσεται σε μία κατηγορία με βάση

αυτά τα χαρακτηριστικά. Η διαδικασία αυτή ονομάζεται κατηγοριοποίηση (classification βλ. κεφάλαιο 4). Για κάθε χρήστη που κάνει εγγραφή το σύστημα κρατά τόσο την ημερομηνία εγγραφής όσο και το γεωγραφικό της στίγμα. Επιπρόσθετα, το γεωγραφικό στίγμα του χρήστη ανανεώνεται και μετά την εγγραφή ούτως ώστε να λαμβάνει τις ανάλογες συστάσεις. Εκτός από την εγγραφή γεωγραφική θέση αποθηκεύεται και για κάθε βαθμολόγηση. Το σύστημα, όταν εισέρχεται ένας καινούργιος χρήστης για τον οποίο πρέπει να παράξει συστάσεις, συγκρίνει τη θέση του με τις θέσεις των βαθμολογιών και άλλα χαρακτηριστικά και βγάζει ένα βάρος (βλ. κεφάλαιο 5). Αυτό το βάρος δείχνει ουσιαστικά πόσο κοντά είναι ένα προϊόν σε ένα χρήστη. Όσο πιο μεγάλο είναι το βάρος τόσο μεγαλύτερη είναι και η πιθανότητα ο χρήστης να ενδιαφερθεί για το προϊόν. Με βάση το υπολογιζόμενο βάρος το σύστημα επιτελεί, στη συνέχεια, τη διαδικασία της συσταδοποίησης (Clustering βλ. κεφάλαιο 6). Κατά τη διάρκεια της συσταδοποίησης δημιουργούνται μέσα στην κατηγορία που έχει κατανεμηθεί ο χρήστης δύο ομάδες. Η πρώτη ομάδα (Cluster) περιέχει τα χαμηλότερα βάρη, δηλαδή τα προϊόντα τα οποία με μεγάλη πιθανότητα δεν ενδιαφέρουν το χρήστη ενώ αντίθετα η δεύτερη ομάδα περιέχει τα υψηλότερα βάρη. Ουσιαστικά η δεύτερη συστάδα είναι αυτή που περιέχει τα προϊόντα οι βαθμολογίες των οποίων θα χρησιμοποιηθούν ώστε να προκύψουν οι τελικές συστάσεις προς το νέο επισκέπτη.

Η τελική μορφή της εργασίας διαμορφώνεται με βάση τη λειτουργία των υποσυστημάτων αυτών. Πιο συγκεκριμένα, μετά το δεύτερο κεφάλαιο, στο οποίο γίνεται μία πιο αναλυτική παρουσίαση των συστημάτων συστάσεων σε συνάρτηση με τις λειτουργίες του προτεινόμενου αλγορίθμου, ξεκινά η παρουσίαση των παραπάνω υποσυστημάτων, των τεχνικών που χρησιμοποιήθηκαν και των αποτελεσμάτων που έδωσαν. Στο τέλος της εργασίας γίνεται μία πειραματική αποτίμηση του συστήματος, λαμβάνοντας υπόψιν διαφορετικά σενάρια και μετρικές.

2. ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ

2.1 Τεχνικές Συστάσεων

Πριν ξεκινήσει η λεπτομερής περιγραφή των τμημάτων του προτεινόμενου συστήματος συστάσεων κρίνεται απαραίτητη η αναφορά στις ήδη υπάρχουσες κατηγορίες συστημάτων. Πιο συγκεκριμένα, υπάρχουν αρκετοί τύποι συστημάτων αφενός λόγω των διαφορετικών τομέων στους οποίους απευθύνονται, αφετέρου, λόγω των δεδομένων τα οποία χρησιμοποιούν ή τον τρόπο δημιουργίας τους αλλά και παρουσίασης των συστάσεων στο χρήστη ή ακόμα και λόγω των διαφορετικών αλγορίθμων που χρησιμοποιούν για να φτάσουν στο τελικό αποτέλεσμα. Τα πιο σύγχρονα συστήματα συστάσεων βασίζονται σε δύο, κυρίως, τρόπους λειτουργίας. Ο πρώτος από τους δύο τρόπους βασίζεται στο **περιεχόμενο** (βάσει περιεχομένου, content based) και δεύτερος **βάσει φίλτρου συνεργασίας** (collaborative filtering). Ακόμα, υπάρχουν συστήματα με **βάση τη γνώση** (knowledge-based), **δημογραφικά** και **υβριδικά**.

2.2 Συστήματα με βάση το περιεχόμενο

Αυτός ο τύπος συστήματος δημιουργεί συστάσεις προϊόντων βάσει των προτιμήσεων του χρήστη που έχει αντλήσει σε προηγούμενες αλληλεπιδράσεις του με το σύστημα. Προτείνονται, δηλαδή, αντικείμενα όμοια με αυτά που είχε προτιμήσει ο χρήστης στο παρελθόν. Για κάθε ξεχωριστό χρήστη «χτίζεται» σταδιακά και διατηρείται ένα προφίλ, σύμφωνα με την ανάλυση που έχει γίνει στα προϊόντα που έχει κρίνει ή έχει αξιολογήσει. Ο υπολογισμός της ομοιότητας των προτεινόμενων προϊόντων γίνεται σε σχέση με τα χαρακτηριστικά. Για παράδειγμα, αν κάποιος είχε επιλέξει στο παρελθόν ένα βιβλίο που ανήκει στην κατηγορία της επιστημονικής φαντασίας, τότε το σύστημα παραγωγής συστάσεων μαθαίνει να συστήνει και άλλα βιβλία που ανήκουν στο συγκεκριμένο είδος. Ένα τέτοιο σύστημα έχει την απαίτηση της αξιολόγησης από πλευράς χρήστη καθώς και την ύπαρξη ενός μηχανισμού ανάκτησης των χαρακτηριστικών όρων που περιγράφουν ένα προϊόν.

Συμπερασματικά, το σύστημα για να λειτουργήσει επιτελεί ένα κύκλο εργασιών για να παραγάγει τις τελικές συστάσεις. Αρχικά, αναλύει τα χαρακτηριστικά των προϊόντων και αντλεί όλες τις απαραίτητες πληροφορίες ώστε να καταφέρει στο τέλος να τα κατατάξει σε σχετικά ή μη σχετικά με το χρήστη. Το κύριο σημείο εδώ είναι το ίδιο το αντικείμενο. Δηλαδή τα χαρακτηριστικά που το εκφράζουν και το πώς το ένα αντικείμενο διαφέρει από τα υπόλοιπα. Κάθε αντικείμενο είναι διαφορετικό έχοντας διαφορετικά γνωρίσματα που το περιγράφουν. Το αποτέλεσμα αυτής της ανάλυσης αποτελεί την είσοδο του αλγορίθμου που είναι υπεύθυνος για τη δημιουργία προφίλ του χρήστη. Κατά το τελευταίο στάδιο, γίνεται σύγκριση των χαρακτηριστικών των προϊόντων με το προφίλ του χρήστη και λαμβάνεται η τελική απόφαση για το συγκεκριμένο προϊόν θα μπορούσε να σταθεί ως σύσταση προς το χρήστη.

Όπως προκύπτει από τα παραπάνω, ένα από τα σημαντικότερα συστατικά της προσέγγισης αυτής θεωρείται το προφίλ του χρήστη. Στο προφίλ περιέχονται όλες οι πληροφορίες για τα ενδιαφέροντα και τις προτιμήσεις του χρήστη. Αποδεικνύεται ότι οι πληροφορίες του προφίλ των χρηστών πηγάζουν ρητά από τον ίδιο το χρήστη, μέσω ερωτηματολογίων που υποβλήθηκαν ή από την συναλλακτική συμπεριφορά τους στην πάροδο του χρόνου. Για κάθε αντικείμενο, ορίζεται επίσης ένα παρόμοιο προφίλ. Χαρακτηριστικό παράδειγμα μπορεί να θεωρηθεί ένα σύστημα συστάσεων ταινιών. Για κάθε ταινία δημιουργείται ένα προφίλ στο οποίο περιλαμβάνεται το είδος της ταινίας, ο σκηνοθέτης, ηθοποιοί, έτος κυκλοφορίας και άλλες πληροφορίες. Με βάση αυτό το προφίλ και πιο συγκεκριμένα τα χαρακτηριστικά των ταινιών κατασκευάζεται ο «σκελετός» του προφίλ του χρήστη, στον οποίο προτείνονται ταινίες με χαρακτηριστικά

που παρομοιάζουν με αυτά που βρίσκονται στο προφίλ του. Για να κατασκευάσουμε το προφίλ του χρήστη, ορίζουμε αρχικά ένα διάνυσμα βαρών $W_u=(W_{u1}, W_{u2}, \ldots, W_{un})$, κάθε ένα από τα οποία αντιπροσωπεύει τη σημασία του συγκεκριμένου χαρακτηριστικού για το χρήστη. Ομοίως, για το προφίλ του αντικειμένου ορίζουμε ένα δεύτερο διάνυσμα $W_i=(W_{i1}, W_{i2}, \ldots, W_{in})$, όπου, αντίστοιχα με το προφίλ του χρήστη, το κάθε βάρος δηλώνει τη βαρύτητα του χαρακτηριστικού για το αντικείμενο. Για την πρόβλεψη της σύστασης ή μη του αντικειμένου s στο χρήστη c, υπολογίζουμε μία συνάρτηση χρησιμότητας U(c,s). Με αυτόν τον τρόπο, η πρόταση μιας ταινίας σε ένα χρήστη, σχετίζεται με την ανάδειξη των ομοιοτήτων μεταξύ των ταινιών, όπου ο χρήστης έχει βαθμολογήσει υψηλά (συγκεκριμένους ηθοποιούς, σκηνοθέτες, είδος). Μόνο για όσα αντικείμενα η τιμή της συνάρτησης ομοιότητας U είναι υψηλή μπορεί να προκύψει σύσταση.

2.3 Συνεργατικό φιλτράρισμα

Η συνεργατική προσέγγιση παρουσιάζει αρκετές διαφοροποιήσεις. Αντί να προταθούν αντικείμενα με βάση την ομοιότητα τους με τα στοιχεία που ένας χρήστης επιδοκίμασε στο παρελθόν, προτείνονται προϊόντα που άλλοι «παρόμοιοι» χρήστες έχουν συμπαθήσει. Αντίθετα με τα συστήματα βασισμένα στο περιεχόμενο που υπολογίζουν την ομοιότητα των προϊόντων, υπολογίζεται η ομοιότητα των χρηστών. Πιο συγκεκριμένα, για κάθε χρήστη βρίσκεται ένα σύνολο πλησιέστερων χρηστών, που ονομάζονται «γείτονες», με τους οποίους βάσει υπολογισμών υπάρχει ισχυρότερος συσχετισμός. Έτσι υπάρχει δυνατότητα πρόβλεψης αποτελεσμάτων για άγνωστα στοιχεία χρησιμοποιώντας συνδυασμό αποτελεσμάτων που είναι γνωστά από τους πλησιέστερους «γείτονες».

Για παράδειγμα, ας υποθέσουμε πως έχουμε δύο χρήστες, έστω Α και Β. Με βάση το ιστορικό των αγορών αλλά και το πλήθος των βαθμολογιών που ο Α αλλά και ο Β υποθέτουμε ότι κατατάσσονται στην ίδια γειτονιά. Δηλαδή, ο Α και ο Β αποτελούν ένα ζευγάρι γειτόνων που συνδέονται με ισχυρό συσχετισμό λόγω πιθανότατα κοινών προτιμήσεων. Υποθέτουμε πως ο παρακάτω πίνακας είναι ο πίνακας βαθμολογιών των δύο χρηστών για πέντε προϊόντα του ηλεκτρονικού καταστήματος (προϊόν 1, προϊόν 2, προϊόν 3, προϊόν 4, προϊόν 5). Οι βαθμολογίες είναι από το ένα μέχρι το πέντε, ενώ, όπως γίνεται αντιληπτό δεν είναι υποχρεωτικό όλοι οι χρήστες να βαθμολογούν όλα τα προϊόντα.

Πίνακας 1: Βαθμολογίες χρηστών για διαφορετικά προϊόντα

	Προϊόν 1	Προϊόν 2	Προϊόν 3	Προϊόν 4	Προϊόν 5
Χρήστης Α	4	5	2		3
Χρήστης Β	4	**5**	**2**		**3**

Παρατηρώντας τον πίνακα 1, οι χρήστες Α και Β έχουν κοινή βαθμολογία για το προϊόν 1, όμως, μόνο ο χρήστης Α έχει βαθμολογήσει τα προϊόντα δύο, τρία και πέντε. Οι βαθμολογίες που εμφανίζονται στον πίνακα με πιο έντονη γραμματοσειρά αποτελούν προβλέψεις του συστήματος και όχι πραγματικές βαθμολογίες του χρήστη. Έχοντας ως δεδομένο ότι οι χρήστες είναι γείτονες, το σύστημα υποθέτει πως έχουν κοινές προτιμήσεις και με αποτέλεσμα να προβλέπει τις βαθμολογίες του χρήστη Β με βάση το

χρήστη Α, για όσα προϊόντα αυτός δεν έχει ακόμα βαθμολογήσει. Στη συνέχεια, και με τη βοήθεια των παραπάνω προβλέψεων γίνονται οι συστάσεις προς το χρήστη Β.

Όπως είναι φυσικό, στην πραγματικότητα η λειτουργία του συστήματος είναι πολύ πιο σύνθετη, ενώ δημιουργούνται πίνακες με τεράστιο όγκο δεδομένων και πιο σύνθετες σχέσεις μεταξύ των χρηστών. Το παραπάνω παράδειγμα, αποτελεί μόνο μία απλοποιημένη μορφή του αληθινού συστήματος για λόγους συντομίας και ευκολότερης κατανόησης.

Το συνεργατικό φιλτράρισμα υπερβαίνει τους περιορισμούς του φιλτραρίσματος περιεχομένου (content-based). Το σύστημα μπορεί να προτείνει αντικείμενα στους χρήστες βασισμένο στις εκτιμήσεις των στοιχείων, αντί του περιεχομένου των στοιχείων, γεγονός που μπορεί να βελτιώσει την ποιότητα των συστάσεων. Αυτό συμβαίνει γιατί η εκτίμηση που δίνει κάποιος χρήστης είναι πιθανό να σε στοιχεία που είναι αδύνατο να περιγραφούν, όπως για παράδειγμα η ποιότητα ηθοποιίας και κινηματογραφικής λήψης σε μια ταινία ή το φιλόξενο κλίμα σε ένα εστιατόριο.

Όπως και στην περίπτωση των content-based συστημάτων, θα ήταν χρήσιμος ο καθορισμός μίας καθαρής έκδοσης του συνεργατικού φιλτραρίσματος. Ένα καθαρό συνεργατικό σύστημα συστάσεων είναι ένα σύστημα το οποίο δεν πραγματοποιεί καμία ανάλυση των στοιχείων. Στην πραγματικότητα, το μόνο γνωστό χαρακτηριστικό για ένα στοιχείο είναι η εκτίμηση. Συνεπώς, συστάσεις για έναν χρήστη υποβάλλονται απλώς βάσει των ομοιοτήτων με άλλους χρήστες. Η καθαρή συνεργατική σύσταση λύνει όλες τις ανεπάρκειες που απαντώνται στα καθαρά συστήματα φιλτραρίσματος περιεχομένου (content-based συστήματα). Με τη βοήθεια των συστάσεων άλλων χρηστών, μπορούμε να εξετάσουμε οποιοδήποτε είδος περιεχομένου και να λάβουμε στοιχεία με διαφορετικό περιεχόμενο από εκείνο που έχει εκτιμήσει ο χρήστης στο παρελθόν. Δεδομένου ότι η ανατροφοδότηση των άλλων χρηστών επηρεάζει τη συστάσεις του συστήματος, δίνεται η δυνατότητα να διατηρηθεί η αποτελεσματικότητα των συστάσεων αυτών και η αποδοτικότητα του συστήματος ακόμα και με λιγότερες εκτιμήσεις από οποιοδήποτε μεμονωμένο χρήστη.

Εντούτοις, αυτή η προσέγγιση εισάγει ορισμένα καινούργια προβλήματα. Το πιο σημαντικό από αυτά, αποτελεί το **«πρόβλημα της ψυχρής εκκίνησης»** (cold-start). Εάν ένα νέο στοιχείο εμφανιστεί στη βάση δεδομένων, δεν υπάρχει κανένας τρόπος να συστηθεί σε έναν χρήστη μέχρις ότου να ληφθούν περισσότερες πληροφορίες για αυτό. Κάτι τέτοιο θα μπορούσε να συμβεί μέσω μιας νέας εκτίμησης χρήστη είτε διευκρινίζοντας παρόμοια με αυτό αντικείμενα. Το πρόβλημα αναφέρεται στη βιβλιογραφία με την ονομασία «ψυχρή εκκίνηση» (cold-start) ή αλλιώς πρόβλημα της «πρώτης εκτίμησης» (first-rater). Η εναλλακτική ονομασία προκύπτει από το γεγονός ότι απαιτείται αξιολόγηση/βαθμολόγηση για τα νέο-εισερχόμενα αντικείμενα που κανένας χρήστης δεν έχει εκτιμήσει ακόμα.

Εάν ο αριθμός των χρηστών είναι σημαντικά μικρότερος σε σχέση με τον όγκο των πληροφοριών που είναι αποθηκευμένα στο σύστημα, όπως είναι φυσικό, υπάρχει ο κίνδυνος να η πυκνότητα των εκτιμήσεων να είναι ιδιαίτερα αραιή, καθιστώντας έτσι το σύνολο των αξιοσύστατων στοιχείων σαφώς μικρότερο. Το παραπάνω συμβαίνει καθώς δημιουργείται μια εξαιρετικά μεγάλη ή γρήγορα μεταβαλλόμενη βάση δεδομένων.

Ένα ακόμα πρόβλημα προκύπτει από το γεγονός ότι ένας χρήστης είναι πιθανό να έχει ασυνήθιστες προτιμήσεις σε σχέση με τον υπόλοιπο πληθυσμό. Σε αυτή την περίπτωση υπάρχει ο κίνδυνος να λάβουμε ως αποτέλεσμα μη επαρκείς συστάσεις, καθώς θα είναι ιδιαίτερα μικρός ο αριθμός των χρηστών που παρουσιάζουν ομοιότητες.

Τα παραπάνω προβλήματα εξαρτώνται αυστηρά από τον αριθμό και την ομοιογένεια των χρηστών, τα οποία επηρεάζουν επίσης την ομάδα των «γειτόνων» του χρήστη. Σε

μια κατάσταση όπου η ανατροφοδότηση αποτυγχάνει να αναγκάσει αυτήν την ομάδα πλησιέστερων γειτόνων να αλλάξει, η αποδοκιμασία για ένα στοιχείο δεν θα αποτρέψει απαραιτήτως το χρήστη από να λάβει παρόμοια στοιχεία στο μέλλον.

Ένα ακόμα πρόβλημα στην παραδοσιακή συνεργατική προσέγγιση φιλτραρίσματος αποτελεί το γεγονός ότι δεν ασχολείται καθόλου με οποιεσδήποτε πληροφορίες μπορούν να εξαχθούν από το περιεχόμενο. Η έλλειψη πρόσβασης στο περιεχόμενο των στοιχείων αποτρέπει σε παρόμοιους χρήστες να βρεθούν στην ίδια γειτονιά εκτός και αν έχουν βαθμολογήσει ακριβώς ίδια στοιχεία. Για παράδειγμα, εάν ένας χρήστης βαθμολόγησε με υψηλό βαθμό τη σελίδα προτάσεων εξόδου μίας ιστοσελίδας Α και ένας άλλος τις αντίστοιχες προτάσεις της σελίδας Β, οι δύο δεν θα κατέληγαν να θεωρούνται απαραίτητα γείτονες.

Είναι προφανές ότι στο φιλτράρισμα περιεχομένου δεν υφίσταται τα ανωτέρω προβλήματα. Έτσι συνηθίζεται να χρησιμοποιείται ο συνδυασμός τους προκειμένου να επιτευχθεί καλύτερη απόδοση φιλτραρίσματος, να αμβλυνθούν οι αρνητικές συνέπειες των προβλημάτων που χαρακτηρίζουν κάθε μέθοδο και να αξιοποιηθούν τα προτερήματα τους.

2.4 Συστήματα Βασισμένα στη Γνώση

Τα Βασισμένα στη Γνώση συστήματα συστάσεων (knowledge-based recommenders) είναι μία ειδική περίπτωση συστημάτων συστάσεων που βασίζονται στη ρητή γνώση σχετικά με την ποικιλία στοιχείων, προτιμήσεων χρηστών και κριτήριων συστάσεων (ποια στοιχεία θα πρέπει να συνίστανται και σε ποιο πλαίσιο). Εξαρτώνται, δηλαδή, από τη γνώση και προτείνουν αντικείμενα που εντάσσονται σε ένα συγκεκριμένο πεδίο γνώσης. Αναλυτικότερα, ασχολούνται με το πώς ορισμένα χαρακτηριστικά στοιχείου ανταποκρίνονται στις ανάγκες και τις προτιμήσεις των χρηστών και, τελικά, πως τα αντικείμενα είναι χρήσιμα για το χρήστη.

Τα βασισμένα στη γνώση συστήματα είναι χρήσιμα σε σενάρια κατά τα οποία παραδοσιακές προσεγγίσεις όπως το συνεργατικό φιλτράρισμα ή το φιλτράρισμα με βάση το περιεχόμενο δεν μπορούν να εφαρμοστούν. Τα παραδοσιακά συστήματα σύστασης (φιλτράρισμα με βάση το περιεχόμενο και συνεργατικά) είναι κατάλληλα για τη σύσταση προϊόντων όπως βιβλία, ταινίες, ή ειδήσεις. Ωστόσο, ειδικά όσον αφορά προϊόντα όπως αυτοκίνητα, υπολογιστές, διαμερίσματα, ή οι χρηματοπιστωτικές υπηρεσίες, αυτές οι προσεγγίσεις δεν είναι η κατάλληλη επιλογή. Για παράδειγμα, θα πρέπει να συνιστώνται μόνο χρηματοπιστωτικές υπηρεσίες που υποστηρίζουν την επενδυτική περίοδο που καθορίζεται από τον πελάτη.

Αντικείμενα όπως τα διαμερίσματα και τα αυτοκίνητα δεν αγοράζονται πολύ συχνά, ως εκ τούτου, τα συστήματα βασισμένα στη βαθμολόγηση συχνά δεν αποδίδουν καλά λόγω του χαμηλού αριθμού των διαθέσιμων αξιολογήσεων για ένα συγκεκριμένο αντικείμενο (όπως ακριβώς οι βαθμολογίες που απαιτούνται από τους αλγορίθμους συνεργατικού φιλτραρίσματος). Επιπλέον, οι χρήστες των εφαρμογών δε θα μπορούσαν να είναι ικανοποιημένοι από βαθμολογίες προηγούμενων ετών και τις αντίστοιχες συστάσεις αντικειμένων (όπως θα ήταν, για παράδειγμα, οι συστάσεις ενός αλγόριθμου με βάση το περιεχόμενο). Ακόμα, στα περίπλοκα πεδία στοιχείων, οι πελάτες επιθυμούν να καθορίζουν τις προτιμήσεις τους ρητά (για παράδειγμα «η μέγιστη τιμή του αυτοκινήτου είναι Χ"). Σε αυτό το πλαίσιο, οι περιορισμοί πρέπει να ληφθούν υπόψη από το σύστημα συστάσεων.

Οι δύο τελευταίες πτυχές δεν υποστηρίζονται από προσεγγίσεις όπως το συνεργατικό φιλτράρισμα και το φιλτράρισμα με βάση το περιεχόμενο. Περαιτέρω παραδείγματα τομέων αντικειμένων που ενδείκνυνται για συστήματα συστάσεων που βασίζονται στη

γνώση είναι οι χρηματοπιστωτικές υπηρεσίες, οι ψηφιακές φωτογραφικές μηχανές, και οι τουριστικοί προορισμοί.

Οι μέθοδοι συστάσεων που βασίζονται στη γνώση βοηθούν στην αντιμετώπιση των προκλήσεων με την αξιοποίηση ρητών απαιτήσεων των χρηστών και βαθιάς γνώσης για τον τομέα των προϊόντων, από όπου και προέρχεται ο υπολογισμός των συστάσεων. Τα συστήματα αυτά σε μεγάλο βαθμό επικεντρώνονται σε πηγές γνώσης που δεν αξιοποιούνται από το συνεργατικό και το φιλτράρισμα με βάση το περιεχόμενο. Το δυνατό σημείο των knowledge-based συστημάτων συστάσεων είναι η απουσία του προβλήματος της ψυχρής εκκίνησης (cold-start). Ένα μειονέκτημα είναι τα πιθανά σημεία συμφόρησης στην απόκτηση γνώσεων που προκαλούνται από την ανάγκη καθορισμού των γνώσεων σύστασης με σαφή τρόπο. Σε σύγκριση, λοιπόν, με το συνεργατικό φιλτράρισμα και το φιλτράρισμα με βάση το περιεχόμενο, τα συστήματα αυτά μπορεί να μην αντιμετωπίζουν το πρόβλημα της «ψυχρής εκκίνησης», η υλοποίησή τους, όμως, είναι σαφώς δυσκολότερη.

Στα συστήματα συστάσεων που βασίζονται στη γνώση, οι απαιτήσεις των χρηστών και οι προτιμήσεις προκαλούνται εντός του πεδίου εφαρμογής ενός βρόχου ανάδρασης. Ένας σημαντικός λόγος για το χαρακτηριστικό, αυτό, των συστημάτων συστάσεων (που είναι βασιζόμενα στη γνώση) είναι η πολυπλοκότητα του τομέα του στοιχείου ο οποίος είναι συχνά αδύνατο να φανερώσει όλες τις προτιμήσεις του χρήστη από την πρώτη στιγμή. Επιπλέον, οι προτιμήσεις του χρήστη δεν είναι συνήθως γνωστές από την αρχή, αλλά φανερώνονται ή κατασκευάζονται εντός του πεδίου εφαρμογής μιας συνόδου σύστασης.

Σε ένα σύστημα συστάσεων που βασίζεται στη γνώση, η ανατροφοδότηση των χρηστών δίνεται μέσω των απαντήσεων τους σε ερωτήσεις που περιορίζουν το σύνολο των σχετικών στοιχείων. Ένα παράδειγμα μιας τέτοιας ερώτηση είναι «Ποιο είδος συστήματος φακών προτιμάτε: σταθερό ή ανανεώσιμους φακούς». Σε τεχνικό επίπεδο, η αναζήτηση με βάση σενάρια σύστασης μπορεί να εφαρμοστεί στη βάση των συστημάτων συστάσεων που βασίζονται σε περιορισμούς. Τα συστήματα συστάσεων που βασίζονται σε περιορισμούς, εφαρμόζονται στη βάση του περιορισμού αναζήτησης ή διαφορετικού τύπου προσεγγίσεων με βάση τα ερωτήματα.

Σε ένα σύστημα συστάσεων με βάση την πλοήγηση, η ανατροφοδότηση των χρηστών δίνεται μέσα από τις "κριτικές", οι οποίες προσδιορίζουν τις αιτήσεις αλλαγής όσον αφορά το αντικείμενο που συστήνεται στο χρήστη. Οι κριτικές, στη συνέχεια, χρησιμοποιούνται για τη σύσταση του επόμενου «υποψήφιου» στοιχείου. Ένα παράδειγμα μιας κριτικής στο πλαίσιο ενός σεναρίου σύστασης μίας ψηφιακής φωτογραφικής μηχανής είναι το σχόλιο: "Θα ήθελα να έχουμε μια φωτογραφική μηχανή σαν αυτή, αλλά με χαμηλότερη τιμή". Αυτό είναι ένα παράδειγμα μιας «κριτικής», η οποία αντιπροσωπεύει μια αίτηση αλλαγής σε ένα χαρακτηριστικό του στοιχείου. Ένα σύνθετο σώμα κριτικών επιτρέπει τον προσδιορισμό περισσότερων από μία αιτήσεων αλλαγής σε μια στιγμή. Η «Δυναμική κριτική» λαμβάνει επίσης υπόψη τις προηγούμενες κριτικές του χρήστη (το ιστορικό κριτικής). Πιο πρόσφατες προσεγγίσεις εκμεταλλεύονται, επιπροσθέτως, πληροφορίες που είναι αποθηκευμένες στα αρχεία καταγραφής της αλληλεπίδρασης του χρήστη με το σύστημα για να μειωθεί περαιτέρω η προσπάθεια αλληλεπίδρασης από την άποψη του αριθμού των απαιτούμενων κύκλων κριτικής.

Αξιοσημείωτα συστήματα που βασίζονται στη γνώση είναι τα βασισμένα στην περίπτωση (case based). Σε αυτά τα συστήματα, η συνάρτηση ομοιότητας υπολογίζει κατα πόσο οι ανάγκες των χρηστών (περιγραφή του προβλήματος) ταιριάζουν με τις συστάσεις (λύσεις του προβλήματος). Εδώ ο βαθμός ομοιότητας μπορεί να ερμηνεύεται άμεσα ως η χρησιμότητα της σύστασης για τον χρήστη.

Τα συστήματα που βασίζονται σε περιορισμούς (constraint based) είναι ένα άλλο είδος των συστημάτων σύστασης που βασίζονται στη γνώση. Όσον αφορά τη γνώση που χρησιμοποιούν, τα δύο συστήματα είναι παρόμοια. Αρχικά, συλλέγονται οι απαιτήσεις των χρηστών. Σε περιπτώσεις ασυνεπών απαιτήσεων που δεν μπορούν να βρεθούν λύσεις προτείνονται αυτομάτως επιδιορθώσεις και τα αποτελέσματα συστάσεων μπορούν να εξηγηθούν. Η σημαντική διαφορά έγκειται στον τρόπο που υπολογίζονται οι λύσεις. Τα συστήματα με βάση την περίπτωση καθορίζουν τις συστάσεις τους βάσει μετρικών ομοιότητας ενώ τα συστήματα με βάση τους περιορισμούς εκμεταλλεύονται κυρίως προκαθορισμένες βάσεις γνώσης που περιέχουν σαφείς κανόνες για το πώς να συσχετίζουν τις απαιτήσεις του πελάτη με τα χαρακτηριστικά στοιχείου.

Τα συστήματα που βασίζονται στη γνώση τείνουν να λειτουργούν καλύτερα από τα άλλα κατά την έναρξη της εφαρμογής τους, αλλά αν δεν είναι εξοπλισμένα με εργαλεία μάθησης δεν υπερτερούν σε σχέση με άλλες μεθόδους που μπορούν να εκμεταλλευτούν τα αρχεία καταγραφής της αλληλεπίδρασης ανθρώπου / υπολογιστή. Τα συστήματα που βασίζονται στους περιορισμούς τυπικά ορίζονται από δύο σύνολα μεταβλητών (VC, $VPROD$) και τρία διαφορετικά σύνολα περιορισμών (CR, CF, $CPROD$). Αυτές οι μεταβλητές και οι περιορισμοί είναι τα κύρια συστατικά ενός προβλήματος ικανοποίησης περιορισμών. Μια λύση για ένα πρόβλημα ικανοποίησης περιορισμού αποτελείται από ένα συνδυασμό των μεταβλητών, έτσι ώστε όλοι οι περιορισμοί που ορίζονται να πληρούνται. Οι απαιτήσεις του χρήστη/πελάτη VC περιγράφουν τις πιθανές απαιτήσεις των πελατών. Οι ιδιότητες του προϊόντος $VPROD$ περιγράφουν τις ιδιότητες ενός δεδομένου είδους προϊόντων. Για Παράδειγμα οι ιδιότητες του προϊόντος είναι το είδος του προϊόντος ή ακόμα και το όνομα του προϊόντος. Οι περιορισμοί CR είναι οι περιορισμοί που συστηματικά περιορίζουν τις απαιτήσεις του πελάτη. Το φίλτρο προϋποθέσεων CF καθορίζει τη σχέση μεταξύ των απαιτήσεων των πελατών και το συγκεκριμένο είδος προϊόντων. Για παράδειγμα, οι πελάτες χωρίς εμπειρία στον τομέα των χρηματοοικονομικών υπηρεσιών δε θα πρέπει να έχουν προτάσεις που περιλαμβάνουν προϊόντα υψηλού κινδύνου. Τέλος, η μεταβλητή $CPROD$ αντιπροσωπεύει τους βασικούς περιορισμούς σχετικά με τις πιθανές μεταβλητές $VPROD$. Με δεδομένο ένα σύστημα με βάση τη γνώση με τις παραπάνω μεταβλητές και ένα σύνολο απαιτήσεων υπάρχει δυνατότητα να υπολογιστούν οι συστάσεις. Ο σκοπός τώρα του συστήματος που είναι η ανάδειξη ενός συνόλου προϊόντων που ταιριάζουν στις επιθυμίες και τις ανάγκες του πελάτη μπορεί να οριστεί ως ένα πρόβλημα ικανοποίησης περιορισμών (VC,$VPROD$, $CC \cup CF \cup CR \cup CPROD$) .

Η μέθοδος σύστασης που βασίζεται στην υπόθεση (case based) είναι μία από τις πιο επιτυχημένες μεθόδους μηχανικής μάθησης. Η τεχνική αυτή είναι μία ολοκληρωμένη διαδικασία επίλυσης προβλημάτων που στηρίζεται στη μάθηση από την εμπειρία και έχει τέσσερα κύρια στάδια: την ανάκτηση, την επαναχρησιμοποίηση, την προσαρμογή και τη διατήρηση. Με την εμφάνιση ενός νέου προβλήματος αναζητά ένα παρόμοιο πρόβλημα του παρελθόντος (που έχει ήδη επιλύσει μία παρόμοια υπόθεση), και στη συνέχεια επαναχρησιμοποιεί την υπόθεση εκείνη για την επίλυση του σημερινού προβλήματος. Σε αυτές τις προσεγγίσεις μια υπόθεση και ένα προϊόν ουσιαστικά θεωρούνται πανομοιότυπα αντικείμενα. Το πρόβλημα της σύστασης συνήθως αντιπροσωπεύεται από ένα σύνολο χαρακτηριστικών του προϊόντος, που καθορίζονται από το χρήστη, ενώ η λύση της υπόθεσης είναι το ίδιο το προϊόν. Στο βασικό σενάριο χρήσης, ο πελάτης ψάχνει να αγοράσει κάποιο προϊόν και καθιστά σαφείς ορισμένες απαιτήσεις σχετικά με το προϊόν αυτό. Το σύστημα αναζητά την βασική υπόθεση για τα προϊόντα που ταιριάζουν με τις απαιτήσεις του χρήστη. Η διαδικασία ανάκτησης οδηγείται από ένα μετρητή ομοιότητας που υπολογίζει την ομοιότητα της περιγραφής του προβλήματος, δηλαδή, τις σημερινές απαιτήσεις των χρήστη με τα προϊόντα στη

βάση δεδομένων της υπόθεσης. Μια σειρά από προϊόντα στη συνέχεια ανακτώνται από τη βασική υπόθεση και τα προϊόντα αυτά μετατρέπονται σε συστάσεις προς το χρήστη.

2.5 Συστήματα Βασισμένα σε Δημογραφικά Δεδομένα

Ο όρος δημογραφικά δεδομένα αναφέρεται στα στοιχεία που υπάρχουν στο προφίλ κάθε χρήστη και έχουν να κάνουν με στοιχεία όπως την ηλικία του, το φύλο του, τα ενδιαφέροντά του και άλλα τα οποία βοηθούν να γίνει καλύτερη συσχέτιση των εναλλακτικών με τον χρήστη. Τα δημογραφικά δεδομένα χρησιμοποιούνται βασιζόμενα στην ιδέα πως άτομα που έχουν τα ίδια ενδιαφέροντα, την ίδια ηλικία και τις ίδιες ασχολίες είναι πολύ πιθανόν να αναζητούν το ίδιο αντικείμενο. Τα συστήματα που χρησιμοποιούν αλγόριθμούς που βασίζονται στα δημογραφικά δεδομένα εκτιμούν πως μια εναλλακτική για ένα χρήστη θα του είναι τόσο ενδιαφέρουσα όσο ενδιαφέρουσα είναι η ίδια εναλλακτική σε χρήστες με παρόμοια δημογραφικά δεδομένα.

Οι δημογραφικές προσεγγίσεις φιλτραρίσματος χρησιμοποιούν τις περιγραφές των ανθρώπων για να κατανοήσουν τη σχέση μεταξύ ενός στοιχείου και του τύπου χρηστών στους οποίους ταιριάζει. Τα προφίλ χρηστών δημιουργούνται με την ταξινόμηση των χρηστών σε στερεοτυπικές περιγραφές, που αντιπροσωπεύουν τα χαρακτηριστικά γνωρίσματα των κατηγοριών των χρηστών. Τα προσωπικά στοιχεία του χρήστη είναι απαραίτητα και χρησιμοποιούνται για την ταξινόμηση. Οι ταξινομήσεις χρησιμοποιούνται ως γενικοί χαρακτηρισμοί για τους χρήστες και τα ενδιαφέροντά τους. Συνήθως, τα προσωπικά στοιχεία του χρήστη λαμβάνονται κατά την αίτηση εγγραφής στο σύστημα. Τα προφίλ που προκύπτουν επεκτείνουν το εύρος των πληροφοριών που περιλαμβάνεται στη δημογραφική βάση δεδομένων.

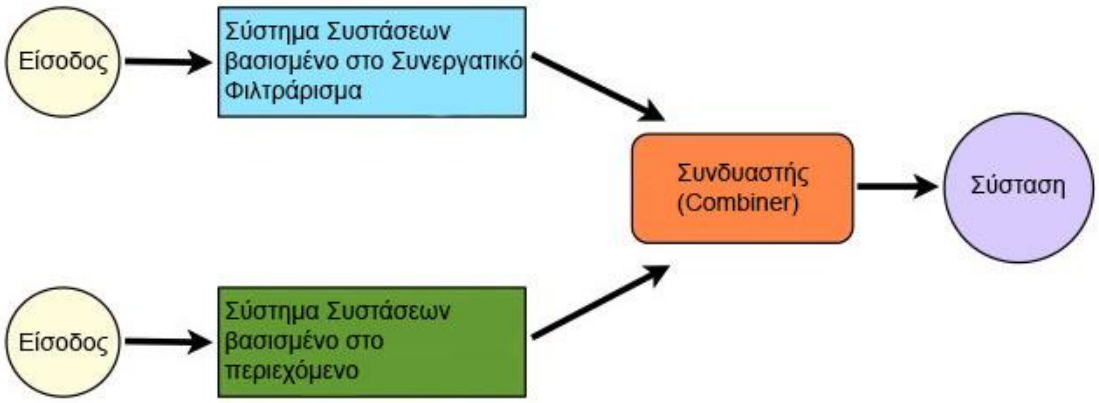
Για παράδειγμα, η μέθοδος που εφαρμόζεται στο LifeStyle Finder (Krulwich 1997) χρησιμοποιεί ένα δημογραφικό σύστημα που καλείται PRIZM και διαιρεί τον πληθυσμό των Ηνωμένων Πολιτειών σε 62 δημογραφικές συστάδες σύμφωνα με το ιστορικό των αγορών τους, τα χαρακτηριστικά του τρόπου ζωής τους και τις απαντήσεις τους σε διάφορες έρευνες.

Ένα δημογραφικό σύστημα φιλτραρίσματος έχει δύο βασικά μειονεκτήματα: οι συστάσεις αποδεικνύονται πάρα πολύ γενικές και δεν προσαρμόζονται στις αλλαγές ενδιαφέροντος. Το μεγαλύτερο, όμως, πρόβλημα που υπάρχει με αυτή τη μέθοδο εύρεσης εναλλακτικών προτάσεων είναι πως πολλοί χρήστες, ιδιαίτερα αυτοί που είναι μεγάλης ηλικίας, δεν θέλουν να εισάγουν τα προσωπικά τους στοιχεία μέσα στο διαδίκτυο και έτσι τα στοιχεία που χρησιμοποιούν οι αλγόριθμοι είναι ελλιπή και δεν μπορούν να δώσουν αξιόπιστα αποτελέσματα. Ένας ακόμα λόγος που δεν υπάρχουν αρκετά δημογραφικά στοιχεία είναι πως οι χρήστες δεν θέλουν να συμπληρώνουν μεγάλα ερωτηματολόγια σχετικά με αυτούς θεωρώντας πως είναι άσκοπα και πως δεν θα ωφελήσει σε τίποτα αν το συμπληρώσουν. Για να αντιμετωπιστεί αυτό το πρόβλημα έχουν δημιουργηθεί αλγόριθμοι στους οποίους ο χρήστης δίνει μια προσωπική σελίδα και από εκεί μπορούν να χρησιμοποιήσουν πολλά στοιχεία από την γλώσσα που χρησιμοποιεί, τις εκφράσεις του, τα θέματα με τα οποία ασχολείται και άλλα. Παρόλα τα προβλήματα που παρουσιάζουν, οι δημογραφικές πληροφορίες μπορούν να αποτελέσουν μια χρήσιμη τεχνική εάν συνδυαστούν με κάποια από τις προηγούμενες προσεγγίσεις.

2.6 Υβριδικά Συστήματα

Πρόσφατες έρευνες απέδειξαν ότι μία υβριδική προσέγγιση, που συνδυάζει το συνεργατικό φιλτράρισμα και το φιλτράρισμα βάσει περιεχομένου μπορεί να είναι πιο αποτελεσματική σε ορισμένες περιπτώσεις. Οι υβριδικές προσεγγίσεις μπορούν να εφαρμοστούν με διάφορους τρόπους. Πιο συγκεκριμένα, κάνοντας προβλέψεις βασισμένες στο περιεχόμενο και προβλέψεις βασισμένες στη συνεργασία χωριστά και

στη συνέχεια συνδυάζοντας τες. Ακόμα, με την προσθήκη δυνατοτήτων βασισμένων στο περιεχόμενο σε μια συνεργατική προσέγγιση (και αντίστροφα) ή με την ενοποίηση των προσεγγίσεων σε ένα μοντέλο. Αρκετές μελέτες συγκρίνουν εμπειρικά την απόδοση του υβριδικού με τις καθαρές μεθόδους συνεργασίας και τις μεθόδους με βάση το περιεχόμενο και αποδεικνύουν ότι οι υβριδικές μέθοδοι μπορούν να παρέχουν πιο ακριβείς συστάσεις από τις καθαρές προσεγγίσεις.



Εικόνα 1: Παράδειγμα Υβριδικού Συστήματος Συστάσεων

Στην εικόνα 1 παρατίθεται η σχηματική αναπαράσταση ενός υβριδικού συστήματος συστάσεων το οποίο συνδυάζει τεχνικές βασισμένες στο συνεργατικό φιλτράρισμα και στα περιεχόμενα. Κάθε ένα λαμβάνει μία είσοδο κάνει την απαραίτητη επεξεργασία και αποδίδει την έξοδο του σε ένα συνδυαστή (Combiner). Αυτό το στοιχείο του συστήματος είναι επιφορτισμένο με την εκτέλεση μίας μορφής συγχώνευσης των αποτελεσμάτων ώστε να προκύψει η τελική έξοδος του συστήματος. Το Netflix αποτελεί ένα χαρακτηριστικό παράδειγμα χρήσης υβριδικού συστήματος. Κάνει συστάσεις συγκρίνοντας τις συνήθειες παρακολούθησης παρόμοιων χρηστών (παράδειγμα συνεργατικού φιλτραρίσματος) και επίσης προσφέροντας ταινίες που μοιράζονται χαρακτηριστικά με άλλες ταινίες τις οποίες ένας χρήστης έχει βαθμολογήσει με υψηλό βαθμό (φιλτράρισμα με βάση το περιεχόμενο).

Μια ποικιλία από τεχνικές έχουν προταθεί ως βάση για συστήματα συστάσεων η συνεργατική, οι βασισμένες στο περιεχόμενο, αυτές που βασίζονται στη γνώση, και οι τεχνικές που χρησιμοποιούν δημογραφικά δεδομένα. Κάθε μία από αυτές τις τεχνικές έχουν γνωστές αδυναμίες, όπως είναι το γνωστό πρόβλημα της ψυχρής εκκίνησης για τη συνεργατικά συστήματα και τα συστήματα με βάση το περιεχόμενο (το πρόβλημα που προκύπτει από την έλλειψη αξιολογήσεων από τους καινούργιους χρήστες ή για τα καινούργια προϊόντα). Επίσης το πρόβλημα της μηχανικής συμφόρησης γνώσης (knowledge engineering bottleneck) σε προσεγγίσεις που βασίζονται στη γνώση. Ένα υβριδικό σύστημα συστάσεων είναι αυτό το οποίο συνδυάζει πολλαπλές τεχνικές μαζί για να επιτύχει κάποια συνέργεια μεταξύ τους.

Συνεργατική προσέγγιση: Το σύστημα παράγει προτάσεις χρησιμοποιώντας μόνο τις πληροφορίες σχετικά με τα προφίλ αξιολόγησης για διαφορετικούς χρήστες. Στα συνεργατικά συστήματα εντοπίζονται ομότιμοι χρήστες με τη βοήθεια του ιστορικού βαθμολογίας. Για κάθε παρόμοιο με τον τρέχοντα χρήστη, δημιουργούνται προτάσεις οι οποίες πηγάζουν από την παρατήρηση της «γειτονιάς».

Προσεγγίσεις με βάση το περιεχόμενο: Το σύστημα παράγει προτάσεις από δύο πηγές. Πρώτα από όλα, από τα χαρακτηριστικά που σχετίζονται με τα προϊόντα και στη συνέχεια από τις βαθμολογίες που ένας χρήστης τους έχει δώσει. Τα συστήματα που είναι βασισμένα στο περιεχόμενο μεταχειρίζονται τη σύσταση ως ένα πρόβλημα

ταξινόμησης και «μαθαίνουν» ένα ταξινομητή για τις προτιμήσεις και αντιπάθειες του χρήστη με βάση τα χαρακτηριστικά του προϊόντος.
Δημογραφική προσέγγιση: Ένα δημογραφικό σύστημα συστάσεων παρέχει συστάσεις που βασίζονται σε ένα δημογραφικό προφίλ του χρήστη. Τα προτεινόμενα προϊόντα μπορούν να παραχθούν για διαφορετικές δημογραφικές ομάδες, συνδυάζοντας τις αξιολογήσεις των χρηστών στις ομάδες αυτές.

Προσεγγίσεις που βασίζονται στη γνώση: Ένα σύστημα συστάσεων με βάση τη γνώση προτείνει προϊόντα που βασίζονται σε συμπεράσματα σχετικά με τις ανάγκες και τις προτιμήσεις του χρήστη. Αυτή η γνώση μερικές φορές περιέχει ρητή λειτουργική γνώση για το πώς ορισμένα χαρακτηριστικά του προϊόντος ανταποκρίνονται στις ανάγκες των χρηστών.

Ο όρος υβριδικό σύστημα χρησιμοποιείται εδώ για να περιγράψει οποιοδήποτε σύστημα συστάσεων συνδυάζει πολλαπλές τεχνικές σύστασης μαζί, έτσι ώστε να παράξει την έξοδο του. Δεν υπάρχει λόγος για τον οποίο δεν θα μπορούσαν πολλές διαφορετικές τεχνικές του ίδιου τύπου να γίνουν υβριδικές. Για παράδειγμα, δύο διαφορετικά συστήματα συστάσεων με βάση το περιεχόμενο θα μπορούσαν να εργαστούν από κοινού, ενώ στη βιβλιογραφία υπάρχουν πολλά παρόμοια παραδείγματα.

Γενικά υπάρχουν πολλές υβριδικές τεχνικές που δύναται να χρησιμοποιηθούν. Κάποιες από αυτές είναι:

- **Με βάρη:** Η βαθμολογία διαφορετικών συστατικών σύστασης συνδυάζεται αριθμητικά.
- **Με εναλλαγή:** Το σύστημα επιλέγει ανάμεσα στα συστατικά συστάσεων και εφαρμόζει τα επιλεγμένα.
- **Ανάμικτα:** Συστάσεις από διαφορετικά συστατικά σύστασης παρουσιάζονται μαζί.
- **Συνδυασμός χαρακτηριστικού:** Τα χαρακτηριστικά που προέρχονται από διαφορετικές πηγές γνώσης συνδυάζονται μαζί και δίνονται σε ένα και μόνο αλγόριθμο σύστασης.
- **Αύξηση χαρακτηριστικού:** Μία τεχνική σύστασης χρησιμοποιείται για να υπολογίσει ένα χαρακτηριστικό ή ένα σετ χαρακτηριστικών, το οποίο γίνεται στη συνέχεια μέρος της εισόδου της επόμενης τεχνικής.
- **Αλληλουχία:** Δίνεται στα συστήματα συστάσεων αυστηρή προτεραιότητα.
- **Μετα-επίπεδο:** Εφαρμόζεται μία τεχνική σύστασης και παράγει μία μορφή μοντέλου. Το μοντέλο αυτό χρησιμοποιείται ως είσοδος της επόμενης τεχνικής.

Πολλά συστήματα σύστασης χρησιμοποιούν μια υβριδική προσέγγιση, συνδυάζοντας την συνεργατική μέθοδο και τη μέθοδο με βάση το περιεχόμενο, η οποία βοηθά στην αποφυγή ορισμένων περιορισμών των με βάση το περιεχόμενο και των συνεργατικών συστημάτων. Οι διαφορετικοί τρόποι για να συνδυαστούν οι προαναφερθέντες μέθοδοι σε ένα υβριδικό σύστημα συστάσεων μπορούν να ταξινομηθούν ως εξής: (1) εφαρμογή συνεργατικής και με βάση το περιεχόμενο μεθόδου ξεχωριστά και συνδυάζοντας τις προβλέψεις τους, (2) ενσωμάτωση κάποιων χαρακτηριστικών με βάση το περιεχόμενο σε μια συνεργατική προσέγγιση, (3) ενσωμάτωση ορισμένων χαρακτηριστικών συνεργατικής προσέγγισης σε ένα σύστημα με βάση το περιεχόμενο και (4) κατασκευής ενός γενικού ενοποιητικό μοντέλο που ενσωματώνει τα χαρακτηριστικά και των δύο. Όλες οι παραπάνω προσεγγίσεις έχουν χρησιμοποιηθεί από ερευνητές συστημάτων συστάσεων, όπως περιγράφεται παρακάτω.

Συνδυάζοντας ξεχωριστά τα συστήματα: Ένας τρόπος για να οικοδομηθεί ένα υβριδικό σύστημα συστάσεων είναι να εφαρμόσει ξεχωριστά συστήματα που βασίζονται σε περιεχόμενο και συνεργασία. Στη συνέχεια, υπάρχουν δύο διαφορετικά σενάρια.

Πρώτον, υπάρχει η δυνατότητα να συνδυάσουμε τις εξόδους (αξιολογήσεις) που λαμβάνονται από κάθε σύστημα συστάσεων ξεχωριστά σε μια τελική σύσταση χρησιμοποιώντας είτε ένα γραμμικό συνδυασμό των αξιολογήσεων ή ένα σύστημα ψηφοφορίας. Εναλλακτικά, μπορεί να χρησιμοποιηθεί ένα από τα επιμέρους συστήματα, σε οποιαδήποτε χρονική στιγμή δύναται να επιλεγεί ώστε να χρησιμοποιηθεί αυτό που είναι «καλύτερο» βασισμένο σε κάποιο μέτρο ποιότητας σύστασης. Προσθέτοντας χαρακτηριστικά τεχνικών που βασίζονται στο περιεχόμενο σε συνεργατικά μοντέλα. Αρκετά υβριδικά εισηγητικά συστήματα βασίζονται σε παραδοσιακές τεχνικές συνεργασίας, αλλά διατηρούν και προφίλ με βάση το περιεχόμενο για κάθε χρήστη. Αυτά τα προφίλ χρησιμοποιούνται για τον υπολογισμό της ομοιότητας μεταξύ δύο χρηστών, και όχι εκείνα των κοινών στοιχείων που έχουν αξιολογηθεί. Αυτό επιτρέπει να ξεπεραστούν τα προβλήματα των ελαχίστων αναφορών που σχετίζονται με αμιγώς συνεργατικές προσεγγίσεις. Ένα άλλο όφελος της αυτής της προσέγγισης είναι ότι στους χρήστες μπορεί να συστηθεί ένα στοιχείο όχι μόνο όταν το στοιχείο αυτό κατέχει υψηλή θέση από τους χρήστες με παρόμοιο προφίλ, αλλά και άμεσα, δηλαδή, όταν βαθμολογείται υψηλά συγκριτικά με το προφίλ του χρήστη.

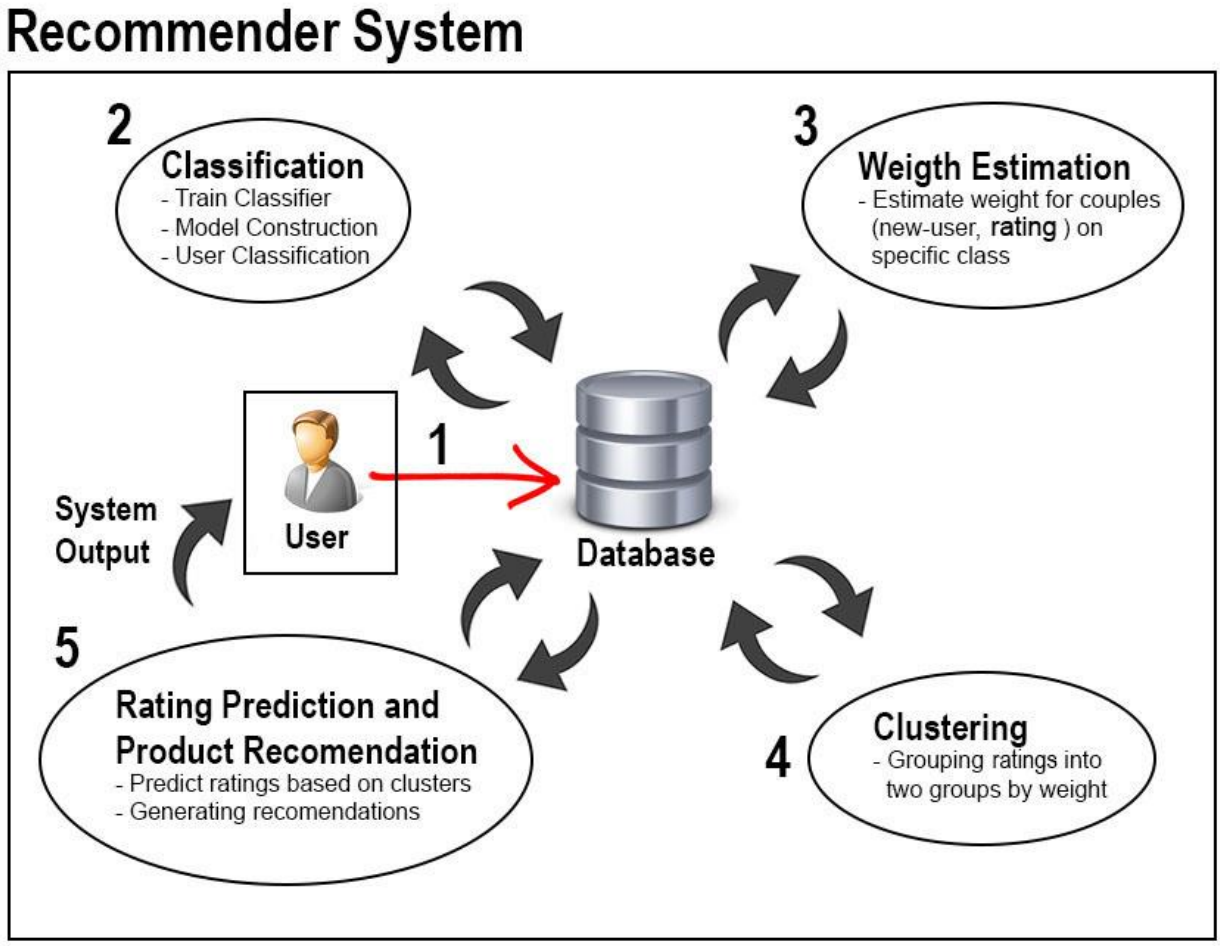
Προσθέτοντας συνεργατικά χαρακτηριστικά σε μοντέλα που βασίζονται στο περιεχόμενο: Η πιο δημοφιλής προσέγγιση σε αυτή την κατηγορία είναι η τεχνική μείωσης διάστασης σε μια ομάδα προφίλ με βάση το περιεχόμενο.

Αναπτύσσοντας ένα ενιαίο ενοποιητικό μοντέλο σύστασης: Πολλοί ερευνητές έχουν ακολουθήσει τη προσέγγιση αυτή κατά τα τελευταία χρόνια. Για παράδειγμα, στο μπορεί να προταθεί η χρήση χαρακτηριστικών τεχνικών με βάση το περιεχόμενο καθώς και συνεργατικών (π.χ., η ηλικία ή το φύλο των χρηστών ή το είδος των ταινιών) σε έναν ενιαίο ταξινομητή που διέπεται από κάποιους κανόνες. Ακόμα θα μπορούσε να προταθεί μία ενιαία πιθανολογική μέθοδος για το συνδυασμό συνεργατικών και με βάση το περιεχόμενο συστάσεων, η οποία βασίζεται στην πιθανολογική λανθάνουσα σημασιολογική ανάλυση. Τελικά, πολλές έρευνες, όπως, αυτές που εμπειρικά συγκρίνουν την απόδοση της υβριδικής τεχνικής με τις τεχνικές συνεργασίας και με βάση το περιεχόμενο αποδεικνύουν ότι οι υβριδικές μέθοδοι μπορούν να παρέχουν πιο ακριβείς συστάσεις από τις καθαρές προσεγγίσεις.

3. ΠΕΡΙΓΡΑΦΗ ΣΥΣΤΗΜΑΤΟΣ ΣΥΣΤΑΣΕΩΝ ΒΑΣΙΣΜΕΝΟΥ ΣΕ ΧΩΡΟΧΡΟΝΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ

3.1 Περιγραφή Συστήματος

Στο παρόν κεφάλαιο, γίνεται μία αποσπασματική περιγραφή του προτεινόμενου συστήματος προτού γίνει αναλυτική παρουσίαση όλων των συστατικών του. Πιο συγκεκριμένα, αποτελείται από τέσσερα βασικά μέρη (υποσυστήματα). Αρχικά συναντάμε το υποσύστημα κατηγοριοποίησης (classification), όπου κάθε πελάτης του ηλεκτρονικού καταστήματος κατατάσσεται σε μία κατηγορία με βάση κάποια δημογραφικά χαρακτηριστικά. Στη συνέχεια, για κάθε βαθμολογία υπολογίζεται ένα βάρος (υποσύστημα υπολογισμού βαρών), στο οποίο συνεισφέρουν τόσο η ηλικία του βαθμολογητή, όσο και χωροχρονικές πληροφορίες. Επιπρόσθετα, με βάση τα βάρη πραγματοποιείται μία συσταδοποίηση (clustering) και προκύπτουν δύο ομάδες βαθμολογιών. Μία που περιέχει τις μικρότερες τιμές βαρών και μία δεύτερη με τις μεγαλύτερες. Τελικά, το σύστημα αγνοεί την πρώτη συστάδα και λαμβάνοντας υπόψη τις βαθμολογίες της δεύτερης, πραγματοποιεί μία διαδικασία πρόβλεψης βαθμολογιών για λογαριασμό του χρήστη. Οι βαθμολογικά καλύτερες από τις παραπάνω προβλέψεις δημιουργούν τη λίστα συστάσεων, η οποία αποτελεί την έξοδο του συστήματος (υποσύστημα πρόβλεψης βαθμολογιών και εξαγωγής συστάσεων).



Εικόνα 2: Διάγραμμα υποσυστημάτων του προτεινόμενου συστήματος συστάσεων

Η διαδικασία που περιεγράφηκε απεικονίζεται σχηματικά στην εικόνα 2, όπου ξεχωρίζουν τόσο τα υποσυστήματα όσο και η σειρά με την οποία ενεργοποιούνται. Αξίζει να σημειωθεί πως, αν και τα όρια κάθε διαδικασίας είναι ευδιάκριτα, τα

υποσυστήματα αλληλεξαρτώνται αφού η έξοδος του ενός χρησιμοποιείται ως είσοδος του επόμενου.

3.2 Το Προτεινόμενο Σχήμα Δεδομένων

Οι λειτουργίες του προτεινόμενου αλγορίθμου συστάσεων απαιτούν ένα σύστημα αποθήκευσης το οποίο θα αλληλοεπιδρά με κάθε υποσύστημα ξεχωριστά, θα λαμβάνει, θα αποθηκεύει και θα μοιράζει δεδομένα. Οι πληροφορίες οι οποίες αποθηκεύονται περιλαμβάνουν τα χαρακτηριστικά των χρηστών που είναι εγγεγραμμένοι στο ηλεκτρονικό κατάστημα. Αυτά ποικίλουν από δημογραφικές πληροφορίες μέχρι ιστορικό αγορών και γεωγραφική θέση. Επιπρόσθετα, αποθηκεύονται προτιμήσεις, είτε αυτές πρόκειται για προβλέψεις του συστήματος, είτε για επιθυμίες που έχει εκφράσει ο χρήστης κατά τη διάρκεια της περιήγησης του στο ηλεκτρονικό κατάστημα. Στη συνέχεια, κρίνεται απαραίτητη η αποθήκευση πληροφοριών που έχουν να κάνουν με τα προϊόντα που διαθέτει προς πώληση το κατάστημα και τα χαρακτηριστικά τους. Τέτοιες πληροφορίες θα μπορούσαν να είναι μία περιγραφή, η τιμή ή η κατηγορία στην οποία ανήκει το προϊόν. Τέλος, τα σημαντικότερα, ίσως, από τα δεδομένα που κρατά το κατάστημα και συνδέονται άμεσα με την εξαγωγή των συστάσεων είναι οι βαθμολογίες των χρηστών για τα προς πώληση προϊόντα. Μαζί με τις βαθμολογίες κρατούνται οι προβλέψεις των βαθμολογιών που εξάγει το σύστημα καθώς και το ιστορικό των συστάσεων.

3.3 Κατηγοριοποίηση

Ξεκινώντας την περιγραφή του συστήματος θα εξεταστεί το γεγονός της εισόδου ενός νέου χρήστη στο σύστημα. Η κατηγοριοποίηση του χρήστη γίνεται κατά τη διάρκεια της εγγραφής του στο ηλεκτρονικό κατάστημα. Μόλις, λοιπόν, ολοκληρωθεί η εισαγωγή, το πρώτο υποσύστημα που «ενεργοποιείται» είναι αυτό της κατηγοριοποίησης. Πιο συγκεκριμένα, ο χρήστης δηλώνει κάποια δημογραφικά του στοιχεία, όπως είναι η ηλικία και το φύλο του. Με βάση αυτά τα στοιχεία το σύστημα τον κατατάσσει σε μία κατηγορία η οποία τον συνοδεύει καθ' όλη την πορεία του μέσα στο σύστημα.

Για λόγους που θα αναλυθούν στη συνέχεια, οι χρήστες του ηλεκτρονικού καταστήματος κατατάσσονται αρχικά με βάση το φύλο, αν δηλαδή είναι άντρες ή γυναίκες, και στη συνέχεια με βάση τη δεκαετία της ηλικίας τους. Αν για παράδειγμα, κάποιος χρήστης είναι 42 ετών και γυναίκα τότε θα του αποδοθεί η κατηγορία «γυναίκα-σαράντα» (woman_forty). Η λογική της κατηγοριοποίησης είναι πως γυναίκες και άντρες έχουν διαφορετικές προτιμήσεις. Άρα είναι πολύ πιο πιθανό οι προτιμήσεις δύο γυναικών να μοιάζουν περισσότερο από ότι μίας γυναίκας και ενός άντρα. Με το ίδιο σκεπτικό κοντινές ηλικίες είναι πιο πιθανό να έχουν κοινές προτιμήσεις από ότι πιο μακρινές. Για παράδειγμα, δύο γυναίκες είκοσι χρονών είναι πιο πιθανό να έχουν όμοιες προτιμήσεις από ότι μία γυναίκα είκοσι και μία ογδόντα χρονών. Ολοκληρώνοντας την κατηγοριοποίηση των χρηστών, το σύστημα δίνει τη σκυτάλη στο επόμενο υποσύστημα του υπολογισμού των βαρών.

3.4 Υπολογισμός Βαρών

Το υποσύστημα του υπολογισμού των βαρών λειτουργεί για κάθε πελάτη ξεχωριστά. Είναι ουσιαστικά ένας αριθμός που συνδέει τον πελάτη για τον οποίο προορίζεται η έξοδος του συστήματος, με μία βαθμολογία την οποία έχει πραγματοποιήσει κάποιος άλλος χρήστης του καταστήματος. Το βάρος είναι, ουσιαστικά, ο αριθμός εκείνος, ο οποίος δηλώνει την ομοιότητα των δύο χρηστών. Του χρήστη ο οποίος δέχεται την έξοδο του συστήματος και αυτού που βαθμολόγησε. Όσο μεγαλύτερο βάρος προκύψει, τόσο μεγαλύτερη είναι η ομοιότητα και κατά συνέπεια τόσο πιο πιθανό να συσταθεί το προϊόν το οποίο έχει βαθμολογήσει ο όμοιος χρήστης.

Προχωρώντας στη διαδικασία του υπολογισμού, κάθε καινούργιος χρήστης ο οποίος εγγράφεται στο ηλεκτρονικό κατάστημα κατατάσσεται σε μία κατηγορία με τη βοήθεια της διαδικασίας κατηγοριοποίησης. Στη συνέχεια, όταν το σύστημα ξεκινήσει τη διαδικασία παραγωγής συστάσεων αρχίζει να συγκρίνει κατά μία έννοια τον πελάτη με κάθε μία από τις καταχωρημένες βαθμολογίες. Το αποτέλεσμα αυτής της σύγκρισης είναι το βάρος και προκύπτει με βάση δημογραφικά χαρακτηριστικά του χρήστη που βαθμολογεί όπως η ηλικία αλλά και πιο εξεζητημένες πληροφορίες, όπως είναι η γεωγραφική θέση του ή χρονικές πληροφορίες που αφορούν τη βαθμολόγηση.

Το σύνολο των παραπάνω πληροφοριών αποθηκεύονται ώστε να είναι προσβάσιμο από τα υπόλοιπα υποσυστήματα. Κατά αυτόν τον τρόπο, η έξοδος του συστήματος χρησιμοποιείται ως είσοδος για το επόμενο υποσύστημα, δηλαδή αυτό της συσταδοποίησης (clustering).

3.5 Συσταδοποίηση

Έχοντας πλέον αποθηκευμένες τις πληροφορίες που παραπέμπουν στην ομοιότητα του πελάτη με τις βαθμολογίες και κατ' επέκταση τα προϊόντα, το σύστημα χρησιμοποιεί ένα τρόπο ώστε να διαχωρίσει τις βαθμολογίες με τη μεγαλύτερη ομοιότητα, από αυτές με τη μικρότερη. Αυτό επιτυγχάνεται με τη βοήθεια του υποσυστήματος συσταδοποίησησης. Ουσιαστικά, το υποσύστημα της συσταδοποίησης χωρίζει τους πελάτες σε δύο ομάδες με βάση το βάρος του ζευγαριού χρήστη-βαθμολογίας. Η πρώτη ομάδα, ή ομάδα υψηλού ενδιαφέροντος, συμβολίζεται με το 0 και περιέχει τις βαθμολογίες με τα μεγαλύτερα βάρη. Η δεύτερη, η οποία χαρακτηρίζεται αλλιώς και ως ομάδα χαμηλού ενδιαφέροντος, συμβολίζεται με το 1 και περιέχει τις υπόλοιπες βαθμολογίες.

Το σκεπτικό σύμφωνα με το οποίο επιλέχθηκε η δημιουργία δύο ομάδων από τον αλγόριθμο, κρύβεται πίσω από την απαίτηση για διαχωρισμό των βαθμολογιών και κατ' επέκταση των προϊόντων σε πιθανά για σύσταση στο χρήστη και μη συμβατά με αυτόν. Οι προαναφερόμενες δύο ομάδες εμπεριέχουν ακριβώς τα παραπάνω στοιχεία. Η ομάδα υψηλού ενδιαφέροντος περιλαμβάνει αυτά που είναι πιθανό να συσταθούν, αντίθετα, στην ομάδα χαμηλού ενδιαφέροντος ανήκουν οι βαθμολογίες, το προφίλ των οποίων δεν ταίριαξε με αυτό το χρήστη. Αξίζει να σημειωθεί πως η επιλογή των δύο ομάδων έγινε χάριν απλότητας, αφού θέση τους θα μπορούσαν να βρίσκονται ακόμα περισσότερες. Σε μία τέτοια περίπτωση το σύστημα θα έπρεπε να αποφασίσει ποιες από αυτές τις ομάδες θα συνεισέφεραν τις βαθμολογίες τους για την κατάρτιση της λίστας συστάσεων.

Ομοίως με τα προηγούμενα αποτελέσματα, έτσι και με αυτά της συσταδοποίση, με την ολοκλήρωση των εργασιών του υποσυστήματος η έξοδος του αποθηκεύεται και χρησιμοποιείται ως είσοδος για το επόμενο στάδιο, αυτό της διαδικασίας εξαγωγής συστάσεων.

3.6 Πρόβλεψη Βαθμολογίας και Εξαγωγή Συστάσεων

Με δεδομένο πως υπάρχουν αποθηκευμένες πληροφορίες για την συστάδα που ανήκει μία βαθμολογία σε σχέση με κάποιον χρήστη και έχοντας συμπληρώσει τη λεγόμενη ομάδα υψηλού ενδιαφέροντος το σύστημα προχωρά στη πρόβλεψη βαθμολογίας. Η έννοια της πρόβλεψης βαθμολογίας περιλαμβάνει τη μελέτη των ήδη υπαρχουσών βαθμολογιών από τους πελάτες μέσα στην ομάδα υψηλού ενδιαφέροντος και τον υπολογισμό μίας βαθμολογίας με βάση τα παραπάνω δεδομένα για προϊόντα που ο πελάτης δεν έχει βαθμολογήσει ο ίδιος.

Έχοντας συμπληρώσει έναν αριθμό προβλέψεων για ένα χρήστη, το σύστημα επιλέγει με βάση κάποια κριτήρια ποιες προβλέψεις θα μετατρέψει σε συστάσεις. Με την πρόβλεψη της βαθμολογίας και την εξαγωγή των συστάσεων το σύστημα ολοκληρώνει

τη λειτουργία του. Δεδομένου, όμως, πως πρόκειται για ένα προτεινόμενο σύστημα κρίνεται απαραίτητος ο έλεγχος της λειτουργίας και η πειραματική αποτίμηση των αποτελεσμάτων του.

4. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

4.1 Το πρόβλημα της Κατηγοριοποίησης

Κατηγοριοποίηση ονομάζεται η μέθοδος Εξόρυξης Δεδομένων (Data Mining) κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών (target category). Είναι αλλιώς γνωστή ως κατηγοριοποίηση δέντρων ή δέντρα αποφάσεων καθώς η έξοδος της μπορεί να μοντελοποιηθεί ως ένα δέντρο του οποίου κάθε κόμβος αναπαριστά ένα σημείο όπου μία απόφαση θα πρέπει να ληφθεί βασισμένη στην είσοδο. Η ροή του αλγορίθμου μεταφέρεται από κόμβο σε κόμβο και όταν φτάσει σε ένα κόμβο φύλλο προκύπτει η προβλεπόμενη έξοδος. Η μέθοδος παράγει μία ακριβή πρόβλεψη της κατηγορίας στην οποία ανήκει το κάθε στοιχείο. Το πρόβλημα της κατηγοριοποίησης απαντάται σε μία πληθώρα εφαρμογών, όπως η Μηχανική Όραση (Computer Vision), η Αναγνώριση Προτύπου (Pattern Recognition), η Βιολογία (Biological Classification), η Αναγνώριση Φωνής (Speech Recognition), και η Γεωστατιστική (Geostatistics). Πιο συγκεκριμένα, η κατηγοριοποίηση μπορεί να περιγραφεί ως μία διαδικασία δύο βημάτων, της κατασκευής του μοντέλου και της εφαρμογής του μοντέλου.

4.1.1 Κατασκευή και Εφαρμογή του μοντέλου

Αρχικά, το πιο σημαντικό βήμα είναι η εκπαίδευση του μοντέλου, η οποία πραγματοποιείται με τη βοήθεια ενός συνόλου δεδομένων που βρίσκονται ήδη αποθηκευμένα. Τα δεδομένα αυτά ονομάζονται σετ εκπαίδευσης και χρησιμοποιούνται ως είσοδος του αλγορίθμου κατηγοριοποίησης. Στη συνέχεια ο κατηγοριοποιητής, ή αλλιώς classifier, ξεκινά να τα αναλύει μέχρι ότου να φτάσει σε ένα συμπέρασμα για τον τρόπο κατηγοριοποίησης. Το συμπέρασμα αυτό είναι ουσιαστικά η διαδικασία που θα πρέπει να ακολουθήσει το μοντέλο ώστε να φτάσει σε μία σωστή έξοδο σύμφωνα με τα δεδομένα που έχουν δοθεί. Η διαδικασία του βήματος της εκπαίδευσης παράγει ένα μοντέλο το οποίο περιέχει τον τρόπο με τον οποίο κατηγοριοποιούνται τα δεδομένα. Αξίζει να σημειωθεί πως η διαδικασία αυτή πραγματοποιείται αυτόματα, με τον αλγόριθμο να «μαντεύει» τη λογική με την οποία αποφασίστηκε μία μεταβλητή να έχει μία συγκεκριμένη τιμή. Το μόνο που χρειάζεται να γίνει για την έναρξη της εκπαίδευσης είναι η παροχή ενός σετ ήδη κατηγοριοποιημένων δεδομένων τα οποία ο αλγόριθμος θα χρησιμοποιήσει ώστε να εκπαιδευτεί και να παράγει το δικό του μοντέλο. Οι τιμές, δηλαδή, των μεταβλητών είναι συμπληρωμένες και για αυτό το λόγο η κατηγοριοποίηση μπορεί να θεωρηθεί μέθοδος εποπτευόμενης μάθησης (supervised learning). Έχοντας, πλέον, δημιουργήσει το μοντέλο κατηγοριοποίησης ο αλγόριθμος είναι σε θέση να μαντέψει με σχετική ακρίβεια, ανάλογα και με τη φύση των δεδομένων εισόδου, την τιμή της μεταβλητής για στοιχεία που δεν την έχουν αποτιμήσει εκ των προτέρων. Διαφορετικοί αλγόριθμοι κατηγοριοποίησης είναι πιθανό να χρησιμοποιούν διαφορετικές τεχνικές για τη δημιουργία του μοντέλου.

4.1.2 Αποτίμηση του Μοντέλου

Για την αποτίμηση της απόδοσης του μοντέλου χρησιμοποιείται ένα σύνολο δεδομένων (test data) που, όμως, θα πρέπει να είναι διαφορετικό από το σύνολο των δεδομένων εκπαίδευσης (training data) που χρησιμοποιήθηκε στο παραπάνω βήμα. Η διαδικασία αποτίμησης της απόδοσης του μοντέλου, ουσιαστικά, ποσοτικοποιεί την απόδοση του, δίνοντας τη δυνατότητα αξιολόγησης της λειτουργίας του. Για να καταλήξει στα, εν λόγο, αποτελέσματα η μέθοδος αποτίμησης συγκρίνει την τιμή πρόβλεψης της κατηγορίας, από τα δοκιμαστικά δεδομένα, με την υπάρχουσα τιμή των δεδομένων εκπαίδευσης. Είναι σημαντικό να τονιστεί πως είναι πιθανό να υπάρχουν περισσότεροι από ένας κατηγοριοποιητές οι οποίοι είναι ικανοί να ανταπεξέλθουν σε ένα πρόβλημα

κατηγοριοποίησης και, ως εκ τούτου, η απόφαση της χρησιμοποίησης κάποιου εξ αυτών εναπόκειται στο δημιουργό του συστήματος.

Στη βιβλιογραφία συναντάμε δύο τύπους κατηγοριοποιητών. Η πρώτη κατηγορία είναι η **δυαδική κατηγοριοποίηση** (Binary classification). Χαρακτηριστικό των δυαδικών κατηγοριοποιητών είναι πως η έξοδος τους λαμβάνει μόνο δύο τιμές. Η δεύτερη κατηγορία είναι η **κατηγοριοποίηση πολλαπλών κλάσεων**. Σε αντίθεση με τη δυαδική, η κατηγοριοποίηση πολλαπλών κλάσεων (Multiclass Classification), έχει τη δυνατότητα να αποδίδει έξοδο με πάνω από δύο τιμές. Λόγω των παραπάνω τιμών που αποδίδει, η κατηγορία αυτή έχει πιο πολύπλοκη υλοποίηση σε σχέση με τους δυαδικούς.

4.2 Αλγόριθμος Κατηγοριοποίησης

Η υλοποίηση του υποσυστήματος έγινε με τη βοήθεια του εργαλείου εξόρυξης δεδομένων WEKA. Το εργαλείο αυτό, δίνει τη δυνατότητα υλοποίησης της κατηγοριοποίησης, προσφέροντας μία πληθώρα αλγορίθμων. Στα πλαίσια της παρούσας εργασίας χρησιμοποιούνται οι αλγόριθμοι Naïve Bayes καθώς και ο αλγόριθμος C4.5. Η χρήση των προαναφερθέντων αλγορίθμων δε γίνεται παράλληλα αλλά ένας ανά εκτέλεση. Αυτό συμβαίνει για τη λήψη μίας πλήρους εικόνας της συμπεριφοράς του συστήματος κατά τη χρήση διαφορετικών αλγορίθμων και την τροποποίηση που επιφέρει αυτή στα αποτελέσματα του. Πιο συγκεκριμένα, γίνεται αναφορά στη διαδικασία που ακολουθούν για τον προσδιορισμό της κατηγορίας, παρουσίαση των αποτελεσμάτων τους και σύγκριση της απόδοσης τους.

4.2.1 Naïve Bayes

Ο αλγόριθμος Naive Bayes [1] είναι μία απλή τεχνική για την κατασκευή κατηγοριοποιητών. Οι κατηγοριοποιητές είναι, ουσιαστικά, μοντέλα τα οποία αναθέτουν «ταμπέλες» κλάσεων στα στοιχεία ενός προβλήματος, όπου οι κλάσεις προέρχονται από ένα πεπερασμένο σετ. Φυσικά δεν είναι ο μοναδικός αλγόριθμος για την εκπαίδευση τέτοιων κατηγοριοποιητών αλλά μία οικογένεια αλγορίθμων βασισμένη σε μία βασική αρχή. Η αρχή αυτή είναι πως όλοι οι Naive Bayes κατηγοριοποιητές υποθέτουν πως η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη της τιμής οποιουδήποτε άλλου χαρακτηριστικού, δοσμένης της κλάσης. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί ότι είναι μήλο αν είναι κόκκινο στρογγυλό και έχει διάμετρο περίπου δέκα εκατοστά. Ένας Naive Bayes classifier θεωρεί ότι καθένα από αυτά τα χαρακτηριστικά συμβάλλει ανεξάρτητα στην πιθανότητα το φρούτο να είναι μήλο, ανεξάρτητα από τυχόν συσχετίσεις μεταξύ των χαρακτηριστικών χρώματος, στρογγυλάδας, και διαμέτρου.

Για ορισμένους τύπους μοντέλων πιθανοτήτων, σε ελεγχόμενο περιβάλλον μάθησης, οι ταξινομητές Naive Bayes μπορούν να εκπαιδευτούν πολύ αποτελεσματικά. Σε πολλές πρακτικές εφαρμογές, η εκτίμηση παραμέτρων για μοντέλα Naive Bayes χρησιμοποιεί τη μέθοδο της μέγιστης πιθανοφάνειας. Με άλλα λόγια, μπορεί κανείς να συνεργαστεί με ένα μοντέλο Naïve Bayes, χωρίς την αποδοχή της Bayesian πιθανότητα ή τη χρήση οποιασδήποτε Bayesian μεθόδου.

Παρά τον αφελή σχεδιασμό τους και προφανώς τις υπεραπλουστευμένες υποθέσεις, οι Naïve Bayes ταξινομητές συμπεριφέρονται αρκετά καλά σε πολλές περίπλοκες πραγματικές καταστάσεις. Ένα πλεονέκτημα των Naive Bayes αλγορίθμων είναι ότι απαιτούν μόνο ένα μικρό αριθμό δεδομένων εκπαίδευσης για την εκτίμηση των παραμέτρων που είναι αναγκαίες για την ταξινόμηση.

Αφηρημένα, ένας αλγόριθμος Naive Bayes [2] είναι ένα υπό όρους μοντέλο πιθανοτήτων. Ένα στιγμιότυπο του προβλήματος που πρόκειται να ταξινομηθεί,

αντιπροσωπεύεται από ένα διάνυσμα $x=(x_1, \ldots, x_n)$ το οποίο υποδηλώνει n χαρακτηριστικά (ανεξάρτητες μεταβλητές) και αναθέτει πιθανότητες $p(C_k \mid x_1, \ldots, x_n)$ για καθεμία από τις K πιθανές κλάσεις C_k.

Το πρόβλημα με την παραπάνω εξίσωση είναι ότι αν ο αριθμός των χαρακτηριστικών n είναι μεγάλος ή εάν ένα χαρακτηριστικό μπορεί να πάρει ένα μεγάλο αριθμό τιμών, τότε είναι ανέφικτο το να βασιστεί ένα τέτοιο μοντέλο σε πίνακες πιθανοτήτων. Ως εκ τούτου, το μοντέλο μπορεί να αναδιατυπωθεί για να γίνει πιο προσιτό. Χρησιμοποιώντας το θεώρημα του Bayes, η υπό όρους πιθανότητα μπορεί να αναλυθεί όπως δείχνει στην εξίσωση 1.

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

Εξίσωση 1: Η υπό όρους πιθανότητα

Στη συνέχεια, το μοντέλο ανασυντάσσεται ούτως ώστε να γίνει πιο προσιτό. Με απλά λόγια, χρησιμοποιώντας τη Bayesian ορολογία πιθανότητας προκύπτει η εξίσωση 2.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Εξίσωση 2: Η υπό όρους πιθανότητα σε Bayesian ορολογία

Στην πράξη, υπάρχει ενδιαφέρον μόνο στον αριθμητή του κλάσματος (Εξίσωση 3) ο οποίος είναι ισοδύναμος με το κοινό μοντέλο πιθανότητας

$$p(C_k, x_1, \ldots, x_n)$$

Εξίσωση 3: Αριθμητής υπό όρους πιθανότητας

το οποίο μπορεί να επαναδιατυπωθεί χρησιμοποιώντας τον κανόνα της αλυσίδας (Εξίσωση 4) για επαναλαμβανόμενες εφαρμογές του ορισμού της υπό όρους πιθανότητας:

$$\begin{aligned} p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\ &= p(x_1|x_2, \ldots, x_n, C_k)p(x_2, \ldots, x_n, C_k) \\ &= p(x_1|x_2, \ldots, x_n, C_k)p(x_2|x_3, \ldots, x_n, C_k)p(x_3, \ldots, x_n, C_k) \\ &= \ldots \\ &= p(x_1|x_2, \ldots, x_n, C_k)p(x_2|x_3, \ldots, x_n, C_k) \ldots p(x_{n-1}|x_n, C_k)p(x_n|C_k)p(C_k) \end{aligned}$$

Εξίσωση 4: Επαναδιατύπωση με τη βοήθεια του κανόνα της αλυσίδας

Στη συνέχεια, εστιάζουμε στις "αφελείς" υπό όρους υποθέσεις ανεξαρτησίας. Υποθέτουμε ότι κάθε χαρακτηριστικό είναι ανεξάρτητο κάθε άλλου, δεδομένης της κλάσης (Εξίσωση 5). Αυτό σημαίνει ότι

$$p(x_i|x_{i+1}, \ldots, x_n, C_k) = p(x_i|C_k)$$

Εξίσωση 5: Ανεξαρτησία χαρακτηριστικών

Έτσι το μοντέλο θα μπορούσε να εκφραστεί όπως παρουσιάζεται στην εξίσωση 6.

$$
\begin{aligned}
p(C_k|x_1,\dots,x_n) &\propto p(C_k,x_1,\dots,x_n)\\
&\propto p(C_k)\,p(x_1|C_k)\,p(x_2|C_k)\,p(x_3|C_k)\,\cdots\\
&\propto p(C_k)\prod_{i=1}^{n}p(x_i|C_k)\,.
\end{aligned}
$$

Εξίσωση 6: Διαφορετική έκφραση του παραπάνω μοντέλου

Αυτό σημαίνει ότι, σύμφωνα με τις παραπάνω παραδοχές για την ανεξαρτησία, η υπό όρους κατανομή (Εξίσωση 7) πάνω από τη μεταβλητή της κατηγορίας είναι

$$
p(C_k|x_1,\dots,x_n)=\frac{1}{Z}p(C_k)\prod_{i=1}^{n}p(x_i|C_k)
$$

Εξίσωση 7: Η υπό όρους κατανομή

όπου τα στοιχεία Z=p(x) είναι ένας παράγοντας κλιμάκωσης που εξαρτάται μόνο από τα $x_1, \dots, x_n$.

Η κατασκευή ενός κατηγοριοποιητή από το μοντέλο πιθανοτήτων έγκειται στο γεγονός ότι ο naive Bayes συνδυάζει το παραπάνω μοντέλο με ένα κανόνα απόφασης. Ένας κοινός κανόνας είναι η επιλογή της υπόθεσης που είναι πιο πιθανή, το οποίο είναι γνωστό και ως maximum a posteriori ή κανόνας MAP.

Ο αντίστοιχος Bayes κατηγοριοποιητής, είναι η συνάρτηση που αντιστοιχίζει μια ετικέτα κατηγορίας για κάποιο από τα k όπως προκύπτει από την εξίσωση 8.

$$
\hat{y}=\operatorname*{argmax}_{k\in\{1,\dots,K\}}p(C_k)\prod_{i=1}^{n}p(x_i|C_k).
$$

Εξίσωση 8: Η συνάρτηση που αντιστοιχίζει μία ετικέτα κατηγορίας

4.2.2 Αλγόριθμος C4.5

Ο C4.5 [4] είναι ένας αλγόριθμος που χρησιμοποιείται για τη δημιουργία ενός δέντρου απόφασης. Τα δέντρα απόφασης που παράγονται από τον C4.5 μπορούν να χρησιμοποιηθούν για κατηγοριοποίηση, και για το λόγο αυτό, ο C4.5 συχνά αναφέρεται ως ένα στατιστικός κατηγοριοποιητής. Χτίζει δέντρα απόφασης από ένα σύνολο δεδομένων εκπαίδευσης, χρησιμοποιώντας την έννοια των πληροφοριών εντροπίας. Τα δεδομένα εκπαίδευσης είναι ένα σύνολο S = S1, S2, ... των ήδη ταξινομημένων δειγμάτων. Κάθε S_i δείγμα αποτελείται από ένα p-διάστατο διάνυσμα ($x_{1,i}$, $x_{2,i}$, ..., $x_{p,\theta}$), όπου τα X_J αντιπροσωπεύουν τις αξίες του χαρακτηριστικού ή των χαρακτηριστικών του δείγματος, καθώς και την κατηγορία στην οποία ανήκει το s_i.

Σε κάθε κόμβο του δέντρου, το C4.5 επιλέγει το χαρακτηριστικό των δεδομένων που χωρίζει το πλέον αποτελεσματικό σύνολο των δειγμάτων σε υποσύνολα εμπλουτισμένο στη μία κλάση ή στην άλλη. Το κριτήριο διαχωρισμού είναι το κανονικοποιημένο κέρδος πληροφοριών (διαφορά εντροπίας). Το χαρακτηριστικό με το υψηλότερο κανονικοποιημένο όφελος των πληροφοριών θα επιλεγεί έτσι ώστε να πάρει την απόφαση. Ο αλγόριθμος C4.5, στη συνέχεια, επαναλαμβάνεται σε μικρότερους υποκαταλόγους.

Αυτός ο αλγόριθμος έχει μερικές βασικές περιπτώσεις:

- Όλα τα δείγματα στον πίνακα ανήκουν στην ίδια κατηγορία. Όταν συμβαίνει αυτό, απλά δημιουργεί έναν κόμβο φύλλο για το δέντρο απόφασης λέγοντας να γίνει επιλογή αυτής της κατηγορίας.

- Κανένα από τα χαρακτηριστικά δεν παρέχουν κάποιο κέρδος πληροφοριών. Σε αυτή την περίπτωση, C4.5 δημιουργεί έναν κόμβο απόφασης υψηλότερα από το δέντρο χρησιμοποιώντας την αναμενόμενη τιμή της κατηγορίας.
- Στοιχείο της προηγούμενης αθέατης κατηγορίας που αντιμετωπίζουν. Και πάλι, ο C4.5 δημιουργεί έναν κόμβο απόφασης υψηλότερα από το δέντρο χρησιμοποιώντας την αναμενόμενη τιμή.

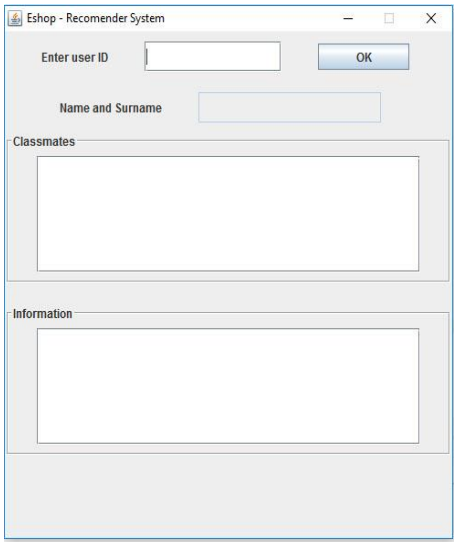
Σε ψευδοκώδικα, ο γενικός αλγόριθμος για την οικοδόμηση δέντρων απόφασης είναι:

1. Ελέγξτε για τις παραπάνω περιπτώσεις.
2. Για κάθε χαρακτηριστικό α, βρες την κανονικοποιημένη αναλογία κέρδους πληροφορίας διαχωρίζοντας στο α.
3. Έστω a_best το χαρακτηριστικό με την υψηλότερο κανονικοποιημένο κέρδος πληροφορίας.
4. Δημιουργήστε έναν κόμβο απόφασης που διαχωρίζεται στο a_best.

Επανέλαβε στους υποκαταλόγους που λαμβάνονται από τη διάσπαση στο a_best, και πρόσθεσε αυτούς τους κόμβους ως παιδιά του κόμβου.

4.3 Υλοποίηση Κατηγοριοποίησης

Το εργαλείο εξόρυξης δεδομένων WEKA επιτρέπει δύο τρόπους χρήσης του κατηγοριοποιητή. Ο πρώτος τρόπος είναι μέσω της εφαρμογής του WEKA και του γραφικού περιβάλλοντος που προσφέρει δίνοντας τη δυνατότητα αναπαράστασης των αποτελεσμάτων με γραφικές παραστάσεις και πρόσθετες πληροφορίες. Ο δεύτερος τρόπος είναι ο τρόπος που ακολουθήθηκε στα πλαίσια της παρούσας εργασίας, δηλαδή, ο προγραμματιστικός. Κατά την υλοποίηση του δεύτερου τρόπου κρίθηκε αναγκαία η δημιουργία ενός γραφιστικού περιβάλλοντος φιλικό προς το χρήστη για την παρουσίαση των αποτελεσμάτων του συστήματος με δομημένο τρόπο, όπως παρουσιάζεται στην παρακάτω εικόνα.



Εικόνα 3: Η διεπαφή χρήστη

Συμπληρώνοντας στη διεπαφή (Εικόνα 3) το μοναδικό αριθμό του χρήστη ο οποίος επιθυμεί τη λήψη συστάσεων, το σύστημα βάζει σε λειτουργία τα επιμέρους υποσυστήματα προσφέροντας τελικά τις συστάσεις και εμφανίζοντας τα αποτελέσματα κάθε υποσυστήματος στην αντίστοιχη περιοχή. Εκτός του παραδοσιακού προγραμματιστικού τρόπου θα γίνει στη συνέχεια παρουσίαση των αποτελεσμάτων από την εφαρμογή του WEKA.

4.3.1 Εκπαίδευση και χρήση του Μοντέλου

Όπως προαναφέρθηκε η αρχική διαδικασία για τη δημιουργία ενός κατηγοριοποιητή περιλαμβάνει την εκπαίδευση ενός μοντέλου. Με απλά λόγια, παρέχεται στον εκάστοτε αλγόριθμο, τον οποίο υποστηρίζει το WEKA, ένα πλήθος δεδομένων τα οποία είναι κατανεμημένα σε ένα πεπερασμένο αριθμό κλάσεων. Στη συνέχεια, δόθηκε στους πελάτες μία κλάση που σχετίζεται με την ηλικία αλλά και το φύλο του πελάτη. Η λογική της απόδοσης μίας τέτοιας κλάσης σε κάθε πελάτη έγκειται στο γεγονός πως υπάρχει μεγάλη πιθανότητα πελάτης με κοντινή ηλικία και ίδιο φύλο να έχουν παρόμοιες προτιμήσεις. Παραδείγματα, κλάσεων που αποδόθηκαν από την εφαρμογή παρουσιάζονται στην παρακάτω εικόνα (Εικόνα 4).

birth_date	class	sex
1992-09-01	man_twenty	man
1963-10-26	woman_fifty	woman
1993-07-13	man_twenty	man
1992-11-05	man_twenty	man
1983-02-06	woman_thirty	woman
1974-07-25	man_fourty	man
1959-09-11	man_fifty	man
1980-03-26	man_thirty	man

Εικόνα 4: Στιγμιότυπο της βάσης δεδομένων του πίνακα users

Στα πλαίσια της παρούσας εργασίας, τα ονόματα στο πεπερασμένο σετ κλάσεων ξεκινούν από το φύλο στα αγγλικά και στη συνέχεια ακολουθεί η δεκαετία στην οποία ανήκει η ηλικία του πελάτη, χωρισμένα με κάτω παύλα. Όταν ολοκληρωθεί η διαδικασία της εκπαίδευσης η έξοδος του συστήματος είναι ένα μοντέλο που έχει τη δυνατότητα δεδομένης της ημερομηνίας γέννησης ενός χρήστη και του φύλου του, να τον κατατάσσει σε μία από της προαναφερθείσες κατηγορίες. Τα αποτελέσματα που παράγει το συγκεκριμένο μοντέλο, αποθηκεύονται στη βάση και συγκεκριμένα στον πίνακα users στο χαρακτηριστικό class.

Αξίζει να σημειωθεί πως λόγους ευκολίας λαμβάνεται ως δεδομένο πως κάθε πελάτης κατά την εγγραφή του στο σύστημα συμπληρώνει μία φόρμα όπου δηλώνει το φύλο και την ηλικία του.

5. ΥΠΟΛΟΓΙΣΜΟΣ ΒΑΡΩΝ

5.1 Διαδικασία Υπολογισμού

Ο επόμενος σταθμός για την τελική επίτευξη των επιθυμητών συστάσεων προς το χρήστη είναι ο υπολογισμός των βαρών. Ξεκινώντας από εκεί που σταμάτησε το προηγούμενο υποσύστημα, υπολογίζονται τα βάρη για κάθε ζευγάρι (χρήστη, βαθμολογίας χρηστών στην ίδια κλάση). Πιο συγκεκριμένα, μόλις το πρώτο υποσύστημα αποθηκεύσει την έξοδο του, που είναι η τιμή της κλάσης που ανατέθηκε στο χρήστη, το παρόν υποσύστημα λαμβάνει αυτή την τιμή ως είσοδο και βρίσκει όλους τους υπόλοιπους πελάτες που έχουν βαθμολογήσει οποιοδήποτε προϊόν και τους έχει αποδοθεί η συγκεκριμένη τιμή της κλάσης. Στη συνέχεια, με τρόπο που θα αναλυθεί στην επόμενη παράγραφο υπολογίζει επαναληπτικά για κάθε ζευγάρι ένα βάρος το οποίο ουσιαστικά αποτιμά την ομοιότητα των χαρακτηριστικών των βαθμολογιών και του χρήστη με βάση κάποια κριτήρια. Ως ζευγάρι λαμβάνεται πάντα ένας χρήστης ο οποίος θα δεχθεί τελικά τις συστάσεις και κάθε βαθμολογία που έχει αποδοθεί από κάθε χρήστη που βρίσκεται στην ίδια κλάση με τον προαναφερόμενο. Στη συνέχεια, παρατίθεται ο τρόπος με τον οποίο γίνεται ο υπολογισμός του συνολικού βάρους κάθε βαθμολογίας.

5.2 Συναρτήσεις Υπολογισμού

Όπως προαναφέρθηκε, για κάθε ζευγάρι πελάτη-βαθμολογίας υπολογίζεται ένας αριθμός ο οποίος παίρνει τιμές στο διάστημα [0,1]. Ο αριθμός αυτός δείχνει τη σχέση που έχουν οι βαθμολογίες με το χρήστη όσον αφορά την ηλικία του βαθμολογητή, το χρόνο αλλά και τη γεωγραφική θέση στην οποία πραγματοποιήθηκαν. Πιο συγκεκριμένα, όσο μεγαλύτερος είναι αυτός ο αριθμός, όσο δηλαδή πλησιάζει στο ένα, τόσο μεγαλύτερη είναι και η σχέση μεταξύ πελάτη-βαθμολογίας. Για τον υπολογισμό του παραπάνω αριθμού έχουν οριστεί κάποιες συναρτήσεις οι οποίες προσδίδουν στην τελική διαμόρφωση του αριθμού με διαφορετικό βάρος η κάθε μία. Οι συναρτήσεις αυτές εμπλέκουν τα δημογραφικά χαρακτηριστικά του βαθμολογητή, όπως είναι η ηλικία, με χωροχρονικές πληροφορίες για τη βαθμολογία, όπως είναι η γεωγραφική θέση και η χρονική στιγμή.

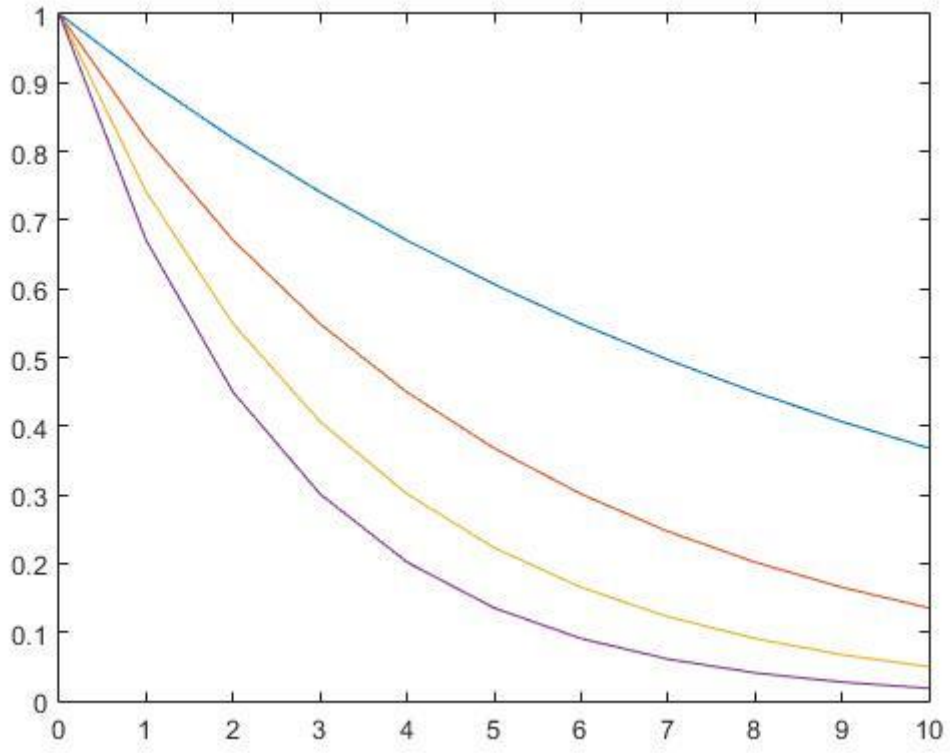
5.2.1 Συνάρτηση βάρους Ηλικίας

Ένα χαρακτηριστικό που συνεισφέρει στον καθορισμό του τελικού βάρους είναι η ηλικία του πελάτη. Ο τρόπος με τον οποίο γίνεται η σύγκριση των ηλικιών και προκύπτει ο αριθμός σύμφωνα με τον οποίο γίνεται η συνεισφορά της ηλικίας στο τελικό βάρος καθορίζεται από μία συνάρτηση με όνομα ageSim. Το συνολικό βάρος που προκύπτει θα πρέπει να είναι ένας αριθμός στο [0,1], κατ' επέκταση, τα επι μέρους χαρακτηριστικά επιβάλλεται να συνεισφέρουν με έναν ανάλογο αριθμό στο σύνολο. Για το λόγο αυτό επιλέχθηκε η ageSim, η μορφή της οποίας παρουσιάζεται στην εξίσωση 9.

$$ageSim = \frac{1}{e^{(\alpha * \Delta Years)}}$$

Εξίσωση 9: Συνάρτηση υπολογισμού βάρους ηλικίας

Όπου ΔYears είναι η διαφορά ηλικίας μεταξύ του βαθμολογητή και του χρήστη για τον οποίο προορίζονται οι συστάσεις. Η γραφική παράσταση της συνάρτησης παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 5).



Εικόνα 5: Γραφική παράσταση της συνάρτησης ageSim για διαφορετικές τιμές του α

Στο σχήμα της εικόνας 5 εμφανίζονται οι γραφικές παραστάσεις της συνάρτησης για διαφορετικές τιμές του α και, πιο αναλυτικά, από πάνω προς τα κάτω για α=0,1, α=0,2, α=0,3 και α=0,4. Συμπερασματικά, όσο αυξάνεται το α η γραφική παράσταση της συνάρτησης γίνεται πιο απότομη. Αυτό πρακτικά σημαίνει πως όσο πιο μεγάλο είναι το α, αυξανομένης της διαφοράς ηλικίας το βάρος μειώνεται πιο γρήγορα. Για παράδειγμα, αν η διαφορά ηλικίας μεταξύ του πελάτη και του βαθμολογητή είναι 7 χρόνια, τότε, για α=0,1 το βάρος (w) είναι περίπου 0,5, για α=0,2 w=0,25, για α=0,3 w=0,11 και για α=0,4 το w=0,7.

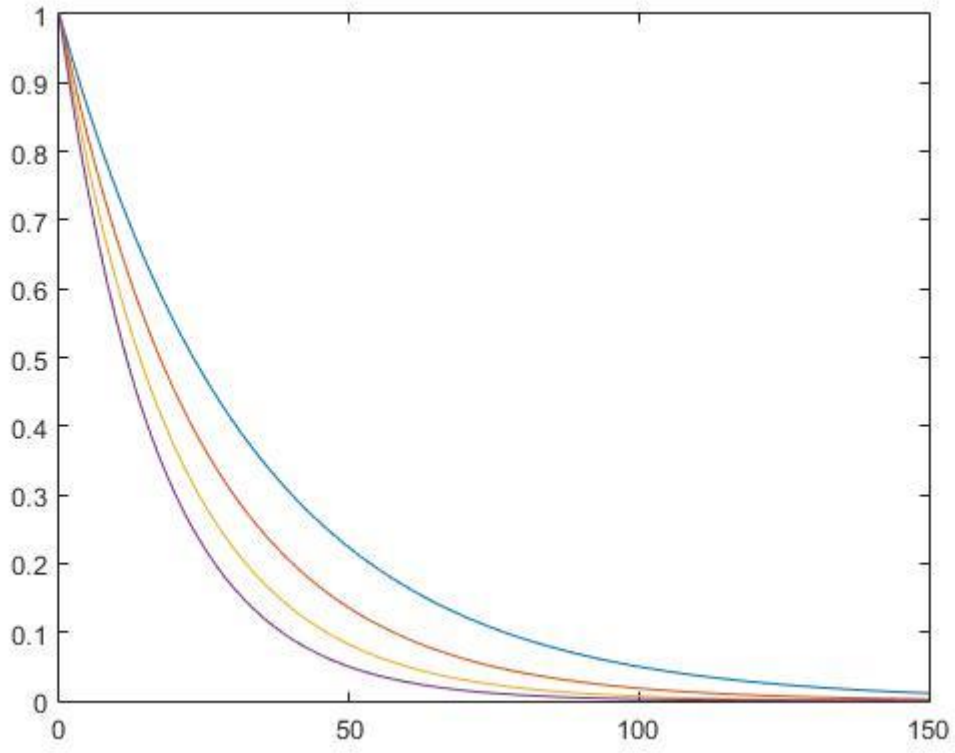
5.2.2 Συνάρτηση βάρους Τοποθεσίας

Ομοίως με την ηλικία στον τελικό αριθμό του βάρους συνεισφέρει και η θέση του χρήστη με ένα πιο αναβαθμισμένο ρόλο. Για τη συγκεκριμένη περίπτωση, ο τρόπος με τον οποίο γίνεται η σύγκριση των γεωγραφικών τοποθεσιών του χρήστη και της βαθμολογίας καθώς και το αποτέλεσμα που προκύπτει για διαφορετικούς αριθμούς της παραπάνω σύγκρισης ορίζεται από τη συνάρτηση logSim (Εξίσωση 10), για την οποία ισχύει:

$$locSim = \frac{1}{e^{(\beta * \Delta Location)}}$$

Εξίσωση 10: Συνάρτηση υπολογισμού του βάρους γεωγραφικής θέσης

Η συνάρτηση είναι φθίνουσα εκθετική με ΔLocation τη διαφορά της γεωγραφικής θέσης του χρήστη από τη βαθμολογία σε χιλιόμετρα. Όσο αυξάνονται τα χιλιόμετρα διαφοράς των δύο χρηστών, τόσο η τιμή που δίνει η συνάρτηση μειώνεται.



Εικόνα 6: Γραφική παράσταση της συνάρτησης locSim για διαφορετικές τιμές του β

Στην πιο πάνω εικόνα (Εικόνα 6) παρουσιάζεται η γραφική παράσταση της εξίσωσης για διαφορετικές τιμές του β. Πιο συγκεκριμένα, (από πάνω προς τα κάτω) για τις τιμές β=0,03, β=0,04, β=0,05 και β=0,06. Είναι φανερό πως όσο η τιμή του β μειώνεται τόσο πιο απότομα μειώνεται η τιμή της συνάρτησης αυξανομένης της απόστασης. Σκοπός της locSim είναι μετά από ένα κατώφλι να μηδενίζεται η τιμή της κάτι που θα σήμαινε εξαγωγή συστάσεων προϊόντων με βαθμολογίες που έχουν πραγματοποιηθεί σε χιλιομετρική απόσταση μικρότερη του προαναφερθέντος ορίου. Για παράδειγμα, για β=0,06 η συνάρτηση λαμβάνει μία εξαιρετικά μικρή τιμή μετά τα 100 χιλιόμετρα απόστασης, πράγμα που σημαίνει πως η τιμή του βάρους θέσης θα είναι 0 και η βαθμολογία ανάλογα και με τις τιμές των βαρών των άλλων χαρακτηριστικών θα έχει μειωμένες πιθανότητες να εισέλθει στη συστάδα από την οποία θα προταθούν προϊόντα (συσταδοποίηση βλ Κεφάλαιο 6). Ανεξάρτητα από τον αριθμό β για διαφορά απόστασης 0 χιλιομέτρων η συνάρτηση έχει τιμή 1 και όσο αυξάνεται η απόσταση χρήστη-βαθμολογίας τόσο η τιμή της μειώνεται.

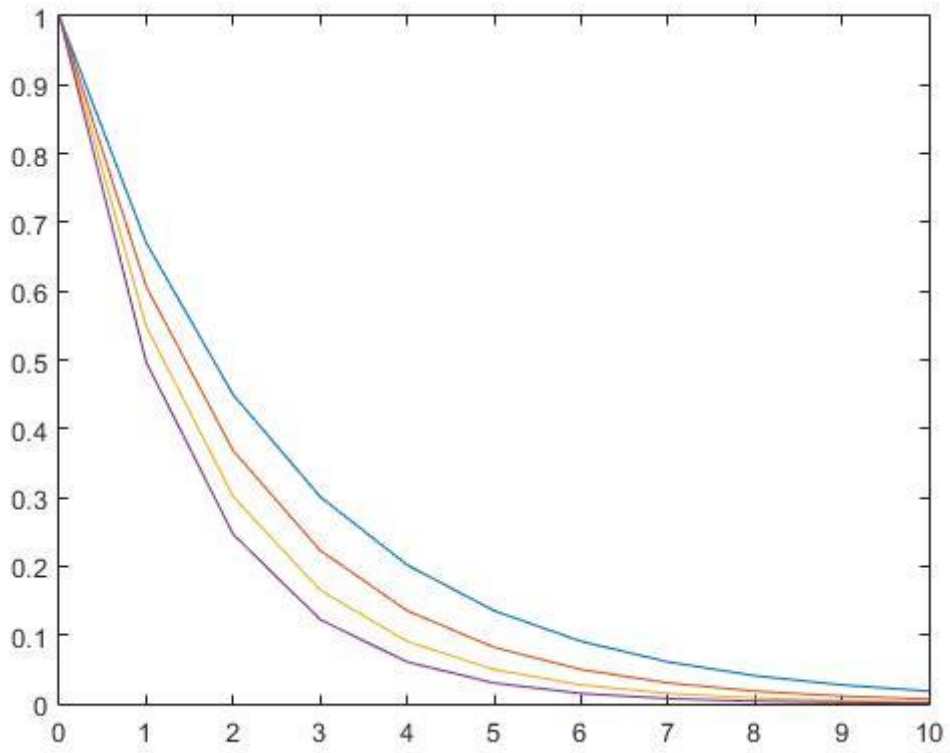
5.2.3 Συνάρτηση βάρους Χρόνου

Το τρίτο και τελευταίο χαρακτηριστικό που συνεισφέρει στον τελικό υπολογισμό του συνολικού βάρους ενός ζευγαριού χρήστη-βαθμολογίας είναι ο χρόνος. Όπως και με τα παραπάνω χαρακτηριστικά έτσι και με το χρόνο απαιτείται η παρουσία μίας φθίνουσα συνάρτηση. Όσο μεγαλύτερη είναι η διαφορά χρόνου (Δtime) τόσο μικρότερη θα πρέπει να είναι η τιμή της. Η συνάρτηση που περιγράφεται πιο πάνω ονομάζεται timeSim και περιγράφεται από την παρακάτω εξίσωση:

$$timeSim = \frac{1}{e^{(\gamma * \Delta time)}}$$

Εξίσωση 11: Συνάρτηση του βάρους Χρόνου

Η γραφική παράσταση της εξίσωσης 11 παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 7) για διαφορετικές τιμές του γ. Πιο συγκεκριμένα, έχει σχεδιαστεί η γραφική παράσταση της συνάρτησης (από πάνω προς τα κάτω) για γ=0,4, γ=0,5, γ=0,6 και γ=0,7. Όπως είναι αντιληπτό, όσο μικρότερη είναι η τιμή του γ τόσο πιο απότομα μειώνεται η τιμή της συνάρτησης. Στον άξονα των y παρουσιάζεται η τιμή της timeSim, ενώ, στον άξονα των x η τιμή της μεταβλητής Δtime.



Εικόνα 7: Γραφική παράσταση της Συνάρτησης timeSim για διαφορετικές τιμές του γ

Η τιμή της πιο πάνω μεταβλητής αντιπροσωπεύει τη διαφορά, σε χρόνια, της χρονικής στιγμής της βαθμολογίας από τη χρονική στιγμή κατά την οποία γίνεται η σύσταση. Αφού η συνάρτηση είναι φθίνουσα όσο μεγαλύτερος είναι ο αριθμός της διαφοράς χρόνου τόσο μικρότερη είναι η τιμή της συνάρτησης. Αυτό πρακτικά σημαίνει πως όσο πιο μακρινές είναι οι δύο ημερομηνίες τόσο μικρότερη είναι η τιμή της συνάρτησης (η οποία συνεισφέρει στο συνολικό βάρος) και άρα λιγοστεύουν οι πιθανότητες για τη βαθμολογία να εισέλθει στη συστάδα των προϊόντων που είναι πιθανόν να συσταθούν (συσταδοποίηση βλ Κεφάλαιο 6).

5.2.4 Υπολογισμός συνολικού βάρους

Αφού υπολογιστούν οι τιμές των παραπάνω συναρτήσεων συνδυάζονται όλες μαζί για να προκύψει το συνολικό βάρος για κάθε ζευγάρι χρηστών. Για να γίνει αυτό, θα πρέπει να δοθεί σε κάθε συνάρτηση ένα ξεχωριστό βάρος που αντικατοπτρίζει τη σημασία που έχει κάθε χαρακτηριστικό στην τελική αποτίμηση του ολικού βάρους. Για το σύστημα που παρουσιάζεται στην παρούσα εργασία το συνολικό βάρος προκύπτει από την αποτίμηση της τιμής της παρακάτω εξίσωσης.

$$Weight = \kappa * ageSim + \lambda * timeSim + \mu * locSim$$
$$= \kappa * \frac{1}{e^{(\alpha * \Delta Years)}} + \lambda * \frac{1}{e^{(\gamma * \Delta time)}} + \mu * \frac{1}{e^{(\beta * \Delta Location)}}$$

Εξίσωση 12: Εξίσωση της συνάρτησης του συνολικού βάρους

Στην παραπάνω εξίσωση (Εξίσωση 12) τη μεγαλύτερη βαρύτητα στον υπολογισμό του τελικού αριθμού βάρους θα πρέπει να την έχει η τοποθεσία. Μετά θα πρέπει να ακολουθεί ο χρόνος ενώ τελευταίο χαρακτηριστικό με το μικρότερο υποβάρος θα πρέπει να είναι η ηλικία. Η επιλογή των αριθμών κ, λ και μ αυτών έγινε βάση της επιθυμητής στρατηγικής που ακολουθήθηκε κατά τη δημιουργία του συστήματος η οποία ήθελε τη γεωγραφική θέση του χρήστη να έχει σημαντικό ρόλο στην επιλογή των συστάσεων. Εφόσον και οι τρεις συναρτήσεις έχουν μέγιστη τιμή 1, το συνολικό άθροισμα των υποβαρών τους θα πρέπει και αυτό να είναι 1, ούτως ώστε να προκύψει αριθμός με μέγιστη τιμή τη μέγιστη τιμή των 3 συναρτήσεων. Τα υποβάρη, λοιπόν, που θα επιλεγούν θα πρέπει να έχουν άθροισμα το 1, δηλαδή κ+λ+μ=1. Αξίζει να σημειωθεί πως για να επιτευχθεί ο αριθμός 1 στο συνολικό βάρος θα πρέπει τόσο η ηλικία όσο και η θέση και ο χρόνος να ταυτίζονται πράγμα εξαιρετικά σπάνιο.

6. ΣΥΣΤΑΔΟΠΟΙΗΣΗ

6.1 Το πρόβλημα της συσταδοποίησης

Η συσταδοποίηση είναι η διαδικασία της ομαδοποίησης ενός συνόλου αντικειμένων, με τέτοιο τρόπο ώστε όλα τα αντικείμενα μέσα στην ίδια ομάδα (που ονομάζεται cluster) να είναι περισσότερο όμοια μεταξύ τους, σε σχέση με αυτά των άλλων ομάδων. Η ομαδοποίηση αποτελεί μία βασική λειτουργία της διερευνητικής εξόρυξης δεδομένων, καθώς και μια κοινή τεχνική για την ανάλυση των στατιστικών στοιχείων, που χρησιμοποιούνται σε πολλούς τομείς, συμπεριλαμβανομένης της μηχανικής μάθησης, της αναγνώρισης προτύπων, της ανάλυσης εικόνας, της ανάκτησης πληροφοριών, της βιοπληροφορικής, της συμπίεσης δεδομένων καθώς επίσης και των γραφικών των υπολογιστών.

Η λύση μπορεί να επιτευχθεί με διάφορους αλγόριθμους που διαφέρουν σημαντικά στην αντίληψη τους για το τι συνιστά μία συστάδα και πώς αυτό μπορεί να βρεθεί αποτελεσματικά. Δημοφιλής έννοιες των συστάδων περιλαμβάνουν ομάδες με μικρές αποστάσεις μεταξύ των μελών του συμπλέγματος, πυκνές περιοχές του χώρου δεδομένων, διαστήματα ή ειδικές στατιστικές κατανομές. Η ομαδοποίηση μπορεί επομένως να διατυπωθεί ως ένας στόχος ενός προβλήματος βελτιστοποίησης. Οι κατάλληλες ρυθμίσεις του αλγορίθμου και των παραμέτρων συσταδοποίησης (συμπεριλαμβανομένων των τιμών, όπως η συνάρτηση απόστασης που θα χρησιμοποιηθεί, το όριο πυκνότητας ή τον αριθμό των αναμενόμενων συστάδων) εξαρτώνται από το μοναδικό σύνολο δεδομένων και τη σκοπούμενη χρήση των αποτελεσμάτων. Η συσταδοποίηση, ως εκ τούτου, δεν είναι μια αυτοματοποιημένη εργασία, αλλά μια επαναληπτική διαδικασία ανακάλυψης γνώσης ή ένας διαδραστικός στόχος βελτιστοποίησης που περιλαμβάνει τη μελέτη και την αποτυχία. Συχνά είναι απαραίτητη η τροποποίηση των προεπεξεργασμένων δεδομένων και των παραμέτρων του μοντέλου μέχρι το αποτέλεσμα που θα επιτευχθεί να έχει τις επιθυμητές ιδιότητες.

Εκτός από τον όρο της συσταδοποίησης, υπάρχουν μια σειρά από όρους με παρόμοιες έννοιες, συμπεριλαμβανομένης της αυτόματης ταξινόμησης, της αριθμητικής ταξινόμησης, του όρου «botryology» (από την ελληνική βότρυς "σταφυλιών") και της τυπολογικής ανάλυσης. Οι λεπτές διαφορές είναι συχνά στη χρήση των αποτελεσμάτων. Στην εξόρυξη δεδομένων, το θέμα του ενδιαφέροντος είναι οι προκύπτουσες ομάδες, από την άλλη, όμως στην αυτόματη ομαδοποίηση το ενδιαφέρον μετατοπίζεται στην προκύπτουσα διακριτική ισχύ.

6.2 Αλγόριθμος Συσταδοποίησης

Όσον αφορά τις στατιστικές, ένας αλγόριθμος expectation-maximization (EM) [3] είναι μια επαναληπτική μέθοδος για την εύρεση των, μέγιστης πιθανοφάνειας ή των maximum a posteriori (MAP) εκτιμήσεων των παραμέτρων σε στατιστικά μοντέλα, όπου το μοντέλο εξαρτάται από απαρατήρητες λανθάνουσες μεταβλητές. Η επανάληψη της ΕΜ εναλλάσσεται μεταξύ της εκτέλεσης ενός βήματος μέσης τιμής (Ε), το οποίο δημιουργεί μια συνάρτηση για τη μέση τιμή της λογαριθμικής πιθανότητας χρησιμοποιώντας την τρέχουσα εκτίμηση για τις παραμέτρους, και ενός βήματος μεγιστοποίησης (Μ), το οποίο υπολογίζει τις παραμέτρους, μεγιστοποιώντας την αναμενόμενη λογαριθμική πιθανοφάνεια που βρέθηκε στο στάδιο Ε. Αυτές οι παράμετροι-εκτιμήσεις, στη συνέχεια, χρησιμοποιούνται για να προσδιοριστεί η κατανομή των λανθανουσών μεταβλητών στο επόμενο στάδιο Ε.

Ο αλγόριθμος ΕΜ χρησιμοποιείται για την εύρεση (τοπικά) παραμέτρων μέγιστης πιθανοφάνειας ενός στατιστικού μοντέλου σε περιπτώσεις όπου οι εξισώσεις δεν μπορούν να επιλυθούν άμεσα. Συνήθως αυτά τα μοντέλα περιλαμβάνουν λανθάνουσες μεταβλητές εκτός από άγνωστες παραμέτρους και γνωστές παρατηρήσεις δεδομένων.

Δηλαδή, είτε υπάρχουν ελλείπουσες τιμές μεταξύ των δεδομένων, ή το μοντέλο μπορεί να διατυπωθεί πιο απλά υποθέτοντας την ύπαρξη πρόσθετων απαρατήρητων σημείων δεδομένων. Για παράδειγμα, ένα μίγμα μοντέλου μπορεί να περιγραφεί πιο απλά με την παραδοχή ότι κάθε σημείο δεδομένων που έχει ήδη παρατηρηθεί έχει ένα αντίστοιχο απαρατήρητο σημείο δεδομένων, ή λανθάνουσα μεταβλητή, προσδιορίζοντας το συστατικό μίγματος στο οποίο ανήκει το κάθε σημείο δεδομένων.

Η εύρεση μίας λύσης μέγιστης πιθανότητας, συνήθως, απαιτεί τη λήψη των παραγώγων της συνάρτησης πιθανότητας σε σχέση με όλες τις άγνωστες τιμές - τις παραμέτρους και τις λανθάνουσες μεταβλητές - και ταυτόχρονα την επίλυση των εξισώσεων που προκύπτουν. Σε στατιστικά μοντέλα με λανθάνουσες μεταβλητές, αυτό συνήθως δεν είναι δυνατό. Αντί αυτού, το αποτέλεσμα είναι συνήθως μια σειρά αλληλένδετων εξισώσεων στην οποία η λύση των παραμέτρων απαιτεί τις τιμές των λανθανουσών μεταβλητών και αντιστρόφως, αλλά αντικαθιστώντας το ένα σύνολο εξισώσεων στο άλλο παράγεται μία άλυτη εξίσωση.

Ο αλγόριθμος ΕΜ προχωρά από την παρατήρηση ότι τα ακόλουθα είναι ένας τρόπος, για να λύσει αριθμητικά αυτά τα δύο σύνολα εξισώσεων. Κάποιος μπορεί απλά να πάρει αυθαίρετες τιμές για ένα από τα δύο σύνολα αγνώστων και να τις χρησιμοποιήσει για να εκτιμήσει το δεύτερο σετ, στη συνέχεια, χρησιμοποιώντας αυτές τις νέες τιμές μπορεί να βρει μια καλύτερη εκτίμηση του πρώτου σετ, και μετά να συνεχίσει να εναλλάσσεται μεταξύ των δύο, μέχρι και οι δύο προκύπτουσες τιμές να συγκλίνουν σε σταθερά σημεία. Δεν είναι προφανές ότι αυτό θα λειτουργήσει, αλλά στην πραγματικότητα μπορεί να αποδειχθεί ότι για αυτό το συγκεκριμένο πλαίσιο λειτουργεί, και ότι το παράγωγο της πιθανότητας είναι (αυθαίρετα κοντά στο) μηδέν, το οποίο με τη σειρά του σημαίνει ότι το σημείο είναι ένα μέγιστο σημείο. Σε γενικές γραμμές μπορεί να υπάρχουν πολλαπλά μέγιστα, και δεν υπάρχει καμία εγγύηση ότι θα βρεθεί το καθολικό μέγιστο. Επίσης, μερικές πιθανοτήτων έχουν κάποιες ανωμαλίες, όπως, παράλογα μέγιστα. Για παράδειγμα, μία από τις «λύσεις» που μπορεί να βρεθεί από τον αλγόριθμο ΕΜ σε ένα μίγμα μοντέλου περιλαμβάνει τον καθορισμό ενός από τα συστατικά να έχει μηδενική διακύμανση και η μέση παράμετρος για την ίδια συνιστώσα να είναι ίση με ένα από τα σημεία δεδομένων.

Λαμβάνοντας υπόψη το στατιστικό μοντέλο, το οποίο δημιουργεί ένα σύνολο X των παρατηρούμενων δεδομένων, ένα σύνολο απαρατήρητος λανθανουσών δεδομένων ή τις τιμές Z που λείπουν, και ένα διάνυσμα άγνωστων παραμέτρων θ, μαζί με μια συνάρτηση πιθανοφάνειας L (θ; X, Z) = P (X, Z | θ), η εκτίμηση της μέγιστης πιθανοφάνειας (MLE) των αγνώστων παραμέτρων καθορίζεται από την οριακή πιθανότητα των παρατηρούμενων δεδομένων (Εξίσωση 13).

$$L = (\theta; X) = p(X|\theta) = \sum_{Z} p(X, Z|\theta)$$

Εξίσωση 13: Οριακή πιθανότητα των παρατηρούμενων δεδομένων

Ωστόσο, αυτή η ποσότητα είναι συχνά δυσεπίλυτη. Για παράδειγμα εάν η Z είναι μια ακολουθία γεγονότων στην οποία ο αριθμός των τιμών αυξάνεται εκθετικά με το μήκος της αλληλουχίας, ο ακριβής υπολογισμός του αθροίσματος καθίσταται εξαιρετικά δύσκολος.

Ο αλγόριθμος ΕΜ προσπαθεί να βρει το MLE της οριακής πιθανότητας από την επαναληπτική εφαρμογή με τα ακόλουθα δύο βήματα:

Expectation (βήμα Ε): Υπολογισμός της αναμενόμενης τιμής της συνάρτησης της λογαριθμικής πιθανότητας (Εξίσωση 14), σε σχέση με την υπό συνθήκη κατανομή του **Z** δεδομένου X κάτω από την τρέχουσα εκτίμηση των παραμέτρων $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = E_{Z|X,y^{(t)}}[logL(\theta; X, Z)]$$

Εξίσωση 14: Αναμενόμενη τιμή της συνάρτησης της λογαριθμικής πιθανότητας

Maximization (βήμα Μ): Εύρεση της παραμέτρου που μεγιστοποιεί την ποσότητα (Εξίσωση 15).

$$\theta^{(t+1)} = arg\max_{\theta} Q(\theta|\theta^{(t)})$$

Εξίσωση 15: Μεγιστοποίηση της ποσότητας

Στα τυπικά μοντέλα στα οποία εφαρμόζεται ο αλγόριθμος ΕΜ:

1. Τα σημεία δεδομένων X μπορεί να είναι διακριτά (λαμβάνοντας τιμές σε ένα πεπερασμένο ή αριθμήσιμα άπειρο σύνολο) ή συνεχή (λαμβάνοντας τιμές σε άπειρο σύνολο). Μπορεί στην πραγματικότητα να υπάρχει ένα διάνυσμα των παρατηρήσεων που σχετίζεται με κάθε σημείο δεδομένων.
2. Οι τιμές που λείπουν (γνωστές και ως λανθάνουσες μεταβλητές) Z είναι διακριτές, που προέρχονται από ένα σταθερό αριθμό τιμών, και υπάρχει μια λανθάνουσα μεταβλητή ανά σημείο δεδομένων.
3. Οι παράμετροι είναι συνεχής, και είναι δύο ειδών: οι παράμετροι που σχετίζονται με όλα τα σημεία δεδομένων, και οι παράμετροι που σχετίζονται με μια συγκεκριμένη αξία μιας λανθάνουσας μεταβλητής (δηλαδή συνδέονται με όλα τα σημεία δεδομένων των οποίων οι αντίστοιχες λανθάνουσες μεταβλητές έχουν μια συγκεκριμένη τιμή).

Ωστόσο, είναι δυνατή η εφαρμογή του αλγορίθμου ΕΜ και για άλλα είδη των μοντέλων.

Το κίνητρο είναι το εξής. Εάν είναι γνωστή η αξία των θ παραμέτρων, είναι δυνατό να βρεθεί η τιμή των λανθανουσών μεταβλητών Z, μεγιστοποιώντας τη λογαριθμική πιθανότητα πάνω από όλες τις πιθανές τιμές του Z, είτε απλά με την επανάληψη του Z ή μέσω ενός αλγορίθμου όπως ο Viterbi αλγόριθμος για κρυμμένα μοντέλα Markov. Αντίθετα, αν γνωρίζουμε την αξία των λανθάνουσες μεταβλητές Z, μπορούμε να βρούμε μια εκτίμηση των θ παραμέτρων αρκετά εύκολα, συνήθως με μία απλή ομαδοποίηση των σημείων δεδομένων, σύμφωνα με την τιμή των συνδεδεμένων λανθανουσών μεταβλητών και βρίσκοντας το μέσο όρο των τιμών, ή κάποια συνάρτηση των τιμών, των σημείων σε κάθε ομάδα. Αυτό υποδηλώνει έναν επαναληπτικό αλγόριθμο, στην περίπτωση όπου τόσο το θ όσο και το Z είναι άγνωστα:

1. Πρώτα γίνεται αρχικοποίηση των παραμέτρων θ με κάποιες τυχαίες τιμές.
2. Υπολογισμός της καλύτερης τιμής του Z δοσμένων των παραπάνω τιμών των παραμέτρων.
3. Στη συνέχεια, χρησιμοποίηση των μόλις υπολογισμένων τιμών του Z για τον υπολογισμό μίας καλύτερης εκτίμησης για τις παραμέτρους θ. Οι παράμετροι που συνδέονται με μία συγκεκριμένη τιμή του Z θα χρησιμοποιήσουν μόνο εκείνα τα σημεία δεδομένων των οποίων οι συσχετιζόμενες λανθάνουσες μεταβλητές έχουν αυτή την τιμή.
4. Επανάληψη των βημάτων 2 και 3 μέχρι να υπάρξει η τελική σύγκλιση.

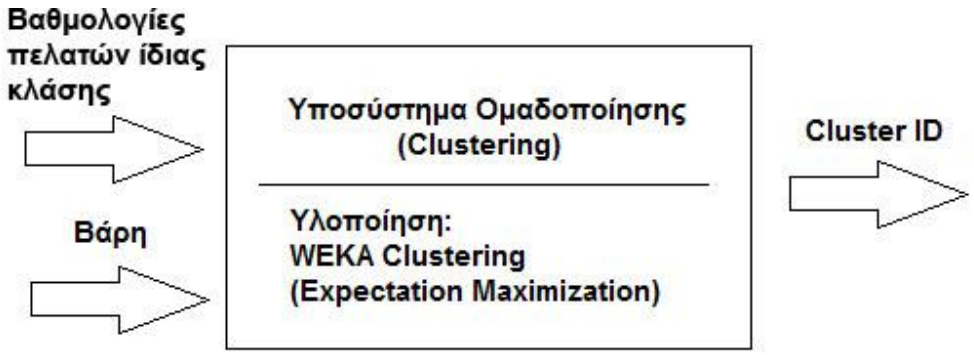
Ο εν λόγω αλγόριθμος μονοτονικά προσεγγίζει ένα τοπικό ελάχιστο της συνάρτησης κόστους, και είναι γνωστός με το όνομα «hard EM». Ο αλγόριθμος k-means είναι ένα παράδειγμα αυτής της κατηγορίας αλγορίθμων.

Ωστόσο, ο αλγόριθμος επιδέχεται βελτιώσεις. Θα μπορούσε, δηλαδή, αντί να κάνει μια σκληρή επιλογή για τα Z, να καθορίζει την πιθανότητα κάθε πιθανής τιμής του Z για κάθε σημείο δεδομένων, και στη συνέχεια να χρησιμοποιεί τις πιθανότητες που συνδέονται με μια συγκεκριμένη τιμή του Z για να υπολογίσει ένα σταθμισμένο μέσο όρο για το σύνολο των σημείων δεδομένων. Ο προκύπτον αλγόριθμος συνήθως ονομάζεται «soft EM», και είναι το είδος του αλγόριθμου που συνήθως συνδέεται με τον αλγόριθμο EM. Οι μετρήσεις που γίνονται για τον υπολογισμό αυτών των σταθμισμένων μέσων όρων ονομάζονται «soft counts» (σε αντίθεση με τις μετρήσεις που χρησιμοποιούνται σε έναν αλγόριθμο τύπου hard-EM, όπως ο K-means). Οι πιθανότητες που υπολογίζονται για το Z είναι posterior πιθανότητες και είναι αυτές που υπολογίζονται στο βήμα E. Τα «soft counts» χρησιμοποιούνται για τον υπολογισμό των νέων τιμών των παραμέτρων και είναι αυτές που υπολογίζονται στο στάδιο M.

6.3 Η Προτεινόμενη Συσταδοποίηση

Η διαδικασία που ακολουθήθηκε για την υλοποίηση της συσταδοποίησης (Clustering) είναι παρόμοια με αυτή που χρησιμοποιήθηκε για την υλοποίηση της κατηγοριοποίησης (Classification). Πιο συγκεκριμένα, χρησιμοποιήθηκε και πάλι το εργαλείο εξόρυξης δεδομένων WEKA και για την ακρίβεια ο αλγόριθμος ομαδοποίησης expectation-maximization (EM), ο οποίος παρουσιάστηκε στην προηγούμενη παράγραφο. Το χαρακτηριστικό πάνω στο οποίο στηρίχθηκε η συσταδοποίηση είναι το βάρος των ζευγαριών χρήστη-βαθμολογίας, για το οποίο η διαδικασία υπολογισμού του παρουσιάστηκε στο προηγούμενο κεφάλαιο.

Ο βασικός λόγος για την ύπαρξη της συσταδοποίησης είναι η ανάγκη να γίνει ένας διαχωρισμός των βαθμολογιών που ανήκουν σε χρήστες με την ίδια τιμή κλάσης (που προκύπτει από τη διαδικασία classification) σε σχέση με το χρήστη ο οποίος θα λάβει τις συστάσεις του συστήματος. Ο διαχωρισμός αυτός θα πρέπει να γίνει με βάση το ποιες βαθμολογίες καλύπτουν τα κριτήρια που θέτει ο υπολογισμός του βάρους. Με απλά λόγια, αν δοθεί ένας χρήστης και η κλάση στην οποία ανήκει, είναι απαραίτητος ο αλγόριθμος ο οποίος θα διαχωρίζει τις βαθμολογίες που ταιριάζουν στο χρήστη με βάση το βάρος, και θα αγνοεί αυτές που θεωρεί πως αφορούν προϊόντα που δεν ενδιαφέρουν τον πελάτη.



Εικόνα 8: Διάγραμμα εισόδων-εξόδου του συστήματος ομαδοποίησης

Αν γίνει μία πιο προσεκτική παρατήρηση των εισόδων του υποσυστήματος, όπως προκύπτουν από το σχήμα της εικόνας 8, μπορεί εύκολα να διαπιστωθεί πως οι δύο είσοδοι (τιμή κλάσης και βάρη) αποτελούν τις εξόδους των προηγούμενων υποσυστημάτων.

Επιστρέφοντας στο διαχωρισμό των βαθμολογιών, λαμβάνουμε ως δεδομένο πως ο αλγόριθμος συσταδοποίησης έχει αρχικοποιηθεί κατά τέτοιο τρόπο ώστε να διαχωρίζει τις βαθμολογίες σε δύο συστάδες. Στη μία συστάδα περιέχονται οι βαθμολογίες με τα λεγόμενα μεγάλα βάρη, τα βάρη δηλαδή που είναι πιο κοντά στην τιμή του 1. Η δεύτερη ομάδα είναι η ομάδα βαθμολογιών με τα μικρότερα βάρη, τα βάρη δηλαδή που είναι πιο κοντά στην τιμή μηδέν. Κάνοντας μία υπενθύμιση της έννοιας του βάρους που αναλύθηκε στην προηγούμενη ενότητα, τα μεγάλα βάρη σημαίνουν ότι τα χαρακτηριστικά των βαθμολογιών έχουν μεγάλη ομοιότητα (όσον αφορά την ηλικία του βαθμολογητή, τη γεωγραφική θέση της βαθμολογίας και το χρόνο που πραγματοποιήθηκε) με το χρήστη σε αντίθεση με τα μικρά βάρη που μεταφράζονται σε πολύ μικρή ομοιότητα. Δημιουργούνται, λοιπόν, οι δύο ομάδες, η μία μικρής ομοιότητας η οποία, ουσιαστικά, αγνοείται από το χρήστη μη έχοντας κάποια χρησιμότητα και η δεύτερη ομάδα, η ομάδα μεγάλης ομοιότητας ή ομάδα υψηλού ενδιαφέροντος στην οποία περιέχονται οι βαθμολογίες με τα υψηλά βάρη, η οποία αποτελεί τη δεξαμενή από την οποία θα προκύψουν οι προβλέψεις βαθμολογιών και οι τελικές συστάσεις προς το χρήστη.

Όπως προκύπτει από το σχήμα της παραπάνω εικόνας ο αλγόριθμος παρουσιάζει ως έξοδο μία τιμή (Cluster ID), ανάλογα με την ομάδα στην οποία τοποθετεί το χρήστη. Η έξοδος του αλγορίθμου για να μπορεί, μετέπειτα, να επαναχρησιμοποιηθεί από άλλο υποσύστημα αποθηκεύεται. Αξίζει να σημειωθεί πως οι δύο συστάδες (Clusters) αποθηκεύονται με βάση ένα μοναδικό αριθμό. Κατά σύμβαση, ο αριθμός μηδέν χρησιμοποιείται ως id για τη συστάδα των μεγάλων βαρών, ενώ ο αριθμός 1, χρησιμοποιείται για τη συστάδα χαμηλού ενδιαφέροντος.

7. ΠΡΟΒΛΕΨΗ ΒΑΘΜΟΛΟΓΙΑΣ ΚΑΙ ΕΞΑΓΩΓΗ ΣΥΣΤΑΣΕΩΝ

7.1 Το πρόβλημα της πρόβλεψης βαθμολογίας

Έχοντας αναφερθεί σε όλες τις υπόλοιπες λειτουργίες μέχρι και την συσταδοποίηση (clustering), μένει μόνο ένα κομμάτι για να ολοκληρωθεί το σύστημα συστάσεων. Αυτό το κομμάτι είναι η πρόβλεψη της βαθμολογίας και κατ' επέκταση η εξαγωγή των συστάσεων. Μέχρι στιγμής, το σύστημα έχει τη δυνατότητα να δημιουργεί κάποιες συστάδες βαθμολογιών κάποιων προϊόντων, με τέτοιο τρόπο ώστε να αυξάνεται η πιθανότητα τα προϊόντα μέσα στην ίδια συστάδα να προταθούν στο χρήστη. Το πρόβλημα το οποίο επιλύεται στην παρούσα ενότητα είναι ο τρόπος με τον οποίο το σύστημα εξάγει τα δεδομένα από την συστάδα υψηλού ενδιαφέροντος, τα μετατρέπει σε προβλέψεις βαθμολογίας και στη συνέχεια πως αυτές οι προβλέψεις μετατρέπονται σε συστάσεις και με ποια κριτήρια.

7.2 Μέθοδος Πρόβλεψης Βαθμολογίας

Σκοπός της μεθόδου είναι η λήψη των μοναδικών αριθμών που χαρακτηρίζουν τις βαθμολογίες που βρίσκονται στη συστάδα υψηλού ενδιαφέροντος, η μετατροπή τους σε είσοδο του υποσυστήματος και η επεξεργασία τους με τέτοιο τρόπο ώστε στο τέλος να προκύψουν οι προβλέψεις των βαθμολογιών ανάμεσα στις οποίες βρίσκονται οι συστάσεις. Κατά το τελικό στάδιο της εξαγωγής των συστάσεων το υποσύστημα προσφέρει ως έξοδο προς τον πελάτη μία λίστα συστάσεων.

Η διαδικασία που περιγράφηκε παραπάνω είναι σαφώς πιο περίπλοκη. Αρχικά το σύστημα ανασύρει από το χώρο αποθήκευσης την έξοδο του συστήματος συσταδοποίησης (Clustering). Αδιαφορεί για τις βαθμολογίες που έχουν επισημανθεί με το μοναδικό αριθμό της συστάδας χαμηλού ενδιαφέροντος. Αντίθετα, κάθε βαθμολογία που βρίσκεται στη συστάδα υψηλού ενδιαφέροντος ομαδοποιείται ανάλογα με το προϊόν στο οποίο αναφέρεται. Στο επόμενο βήμα, υπολογίζεται ο μέσος όρος βαθμολογίας για κάθε μία από τις παραπάνω ομάδες. Ουσιαστικά προκύπτει ένας μέσος όρος βαθμολογίας για κάθε προϊόν που εμφανίζεται στη συστάδα υψηλού ενδιαφέροντος.

Η τελική εξαγωγή των συστάσεων προς το χρήστη περνά πρώτα από την παραπάνω διαδικασία η οποία ονομάζεται **διαδικασία πρόβλεψης βαθμολογίας**. Δηλαδή ο αλγόριθμος προβλέπει τη βαθμολογία για μία σειρά προϊόντων και στο τέλος διαλέγει κάποια προϊόντα από αυτά ως συστάσεις. Στην παρούσα εργασία η πρόβλεψη της βαθμολογίας είναι η πρόσθεση των βαθμολογιών για κάθε προϊόν που έχει βαθμολογήσει κάθε χρήστης στη συστάδα υψηλού ενδιαφέροντος προς τον αριθμό των χρηστών μέσα σε αυτή την ομάδα που βαθμολόγησαν το προϊόν αυτό. Όπως αναφέρθηκε και στην προηγούμενη παράγραφο, στο τέλος της διαδικασίας θα βγει ουσιαστικά ένας μέσος όρος των βαθμολογιών κάθε προϊόντος με τη διαφορά πως αν ο μέσος όρος αυτός προκύπτει από λιγότερες από 10 βαθμολογίες τότε αγνοείται. Το όριο των 10 βαθμολογιών τέθηκε έτσι ώστε να μεγιστοποιείται η αξιοπιστία των συστάσεων. Σε αντίθετη περίπτωση το σύστημα θα μπορούσε να εξάγει ως σύσταση μία πρόβλεψη βαθμολογίας βαθμού 5 (άριστα) που, όμως, προέκυψε από τη βαθμολογία ενός και μόνο βαθμολογητή. Μία τέτοια πρόβλεψη δε θα μπορούσε να αντικατοπτρίζει την εικόνα που έχουν οι χρήστες για το προϊόν αυτό, συνεπώς δε θα μπορούσε να σταθεί ως αξιόπιστη σύσταση. Η πρόβλεψη βαθμολογίας που προκύπτει από την παραπάνω διαδικασία αποθηκεύεται στο σύστημα όπως ακριβώς θα αποθηκευόταν μία βαθμολογία ενός χρήστη, με την ένδειξη πως αποτελεί πρόβλεψη.

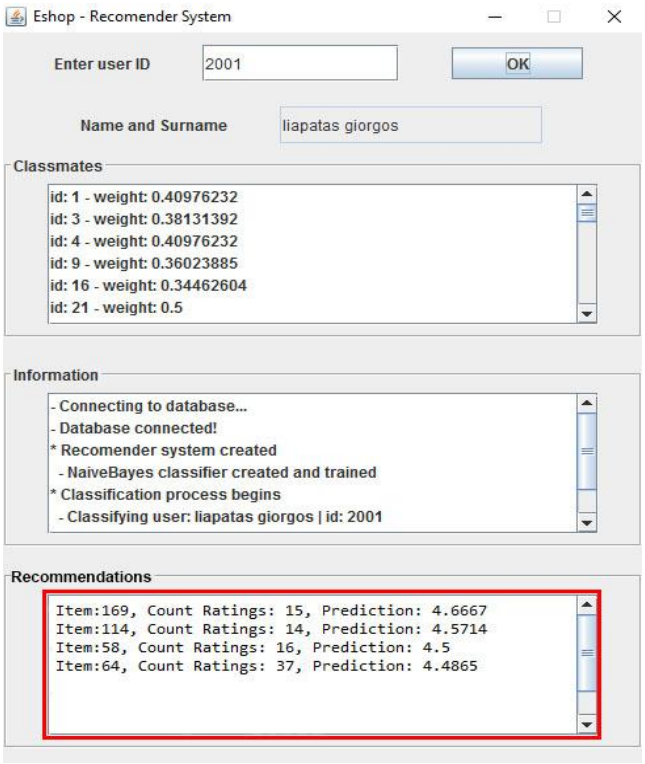
Έχοντας παρουσιάσει την υλοποίηση της πρόβλεψης βαθμολογίας η εξαγωγή συστάσεων δεν απέχει πολύ. Οι προβλέψεις των βαθμολογιών μας δίνουν μία εικόνα της πρόβλεψης των προτιμήσεων του χρήστη. Για παράδειγμα, μέσα στις παραπάνω

προβλέψεις υπάρχουν αντικείμενα που έχουν μέσο όρο βαθμολογίας ο οποίος είναι αρκετά χαμηλός. Θεωρούμε πως τα παραπάνω, με μεγάλη πιθανότητα δε είναι επιθυμητά από το χρήστη. Αντίθετα υπάρχουν στοιχεία που ο μέσος όρος των βαθμολογιών τους είναι πολύ υψηλός (κοντά στο 5). Τα αντικείμενα αυτά, συγκαταλέγονται μέσα στις προτιμήσεις του χρήστη με αρκετά μεγάλη πιθανότητα. Όπως είναι λογικό, οι συστάσεις που εξάγονται προέρχονται από αυτές τις προβλέψεις. Δε θα μπορούσαν, όμως, σε καμία περίπτωση να συμπεριληφθούν ανάμεσα σε αυτές αντικείμενα που έχουν λάβει ως πρόβλεψη χαμηλή βαθμολογία.

Η στρατηγική που χρησιμοποιείται για την υλοποίηση της εξαγωγής συστάσεων βασίζεται σε πρώτη φάση στις μεγαλύτερες βαθμολογίες και στη συνέχεια στον αριθμό των ατόμων που έχουν βαθμολογήσει την ίδια ταινία. Για παράδειγμα. έστω ότι έχουμε 2 ταινίες. Την πρώτη ταινία την έχει βαθμολογήσει ένα άτομο με άριστα (πέντε) ενώ τη δεύτερη την έχουν βαθμολογήσει σαράντα (40) άτομα και ο μέσος όρος των βαθμολογιών τους είναι 4.62. Σε καμία περίπτωση δε θα μπορούσε να συσταθεί η πρώτη ταινία ως πιο πιθανή να ανήκει στις προτιμήσεις του χρήστη. Στην πράξη όσες ταινίες έχουν λιγότερες από δέκα (10) βαθμολογίες αγνοούνται. Το σύστημα χρησιμοποιεί αυτές που έχουν πάνω από δέκα βαθμολογίες, βρίσκει το μέσο όρο τις ταξινομεί με βάση αυτόν το μέσο όρο και εν τέλει κρατάει της πρώτες τέσσερις δημιουργώντας μία λίστα συστάσεων προς το χρήστη. Η λίστα αυτή εμφανίζεται στο χρήστη αλλά και αποθηκεύεται ως ιστορικό. Ουσιαστικά η διαδικασία της πρόβλεψης βαθμολογιών και της εξαγωγής συστάσεων είναι η ίδια ακριβώς διαδικασία. Η διαφορά είναι πως κατά την εξαγωγή της εξόδου του συστήματος εφαρμόζεται ένα είδος ταξινόμησης με βάση τη βαθμολογία και τον αριθμό βαθμολογιών πάνω στις προβλέψεις και εν τέλει επιλέγονται οι 4 πρώτες προβλέψεις στη λίστα ως συστάσεις προς το χρήστη. Ο αριθμός 4 επιλέχθηκε, αφενός, επειδή είναι αρκετά μικρός ώστε να συγκρατηθεί η αξιοπιστία του συστήματος (μεγάλος αριθμός στη λίστα θα μπορούσε να σημαίνει είσοδος σε αυτή προϊόντων με υψηλή βαθμολογία αλλά λιγότερους βαθμολογητές), αφετέρου, επειδή είναι αρκετά μεγάλος ώστε να περιλαμβάνει περισσότερες επιλογές σε περίπτωση αστοχίας κάποιας από τις υπόλοιπες συστάσεις.

7.3 Εξαγωγή Συστάσεων

Όταν ολοκληρωθεί η αποθήκευση των αποτελεσμάτων, οι συστάσεις εμφανίζονται στο κάτω μέρος της διεπαφής χρήστη και ουσιαστικά ολοκληρώνεται η λειτουργία του συστήματος. Στην πιο κάτω εικόνα (Εικόνα 9) φαίνεται ένα παράδειγμα χρήσης του συστήματος και τα αποτελέσματα τα οποία προκύπτουν. Οι συστάσεις εμφανίζονται στο κάτω μέρος της οθόνης σημειωμένες με κόκκινο χρώμα.



Εικόνα 9: Παρουσίαση συστάσεων στο UI του χρήστη

Για την κατανόηση της λειτουργίας του συστήματος δεν αρκεί μόνο ένα απλό παράδειγμα. Για το λόγο αυτό, στην επόμενη ενότητα γίνεται μία πιο ολοκληρωμένη εκτέλεση του αλγορίθμου που περιλαμβάνει την αλλαγή θέσης του χρήστη ώστε να αποτυπωθεί με τον πλέον κατάλληλο τρόπο η διαφοροποίηση των συστάσεων σε σχέση με τη γεωγραφική θέση. Ουσιαστικά, γίνεται η προσομοίωση του αλγορίθμου με πραγματικά δεδομένα. Αφού ολοκληρωθεί και αυτό το βήμα, η εργασία ολοκληρώνεται με την αποτίμηση της ορθότητας των αποτελεσμάτων.

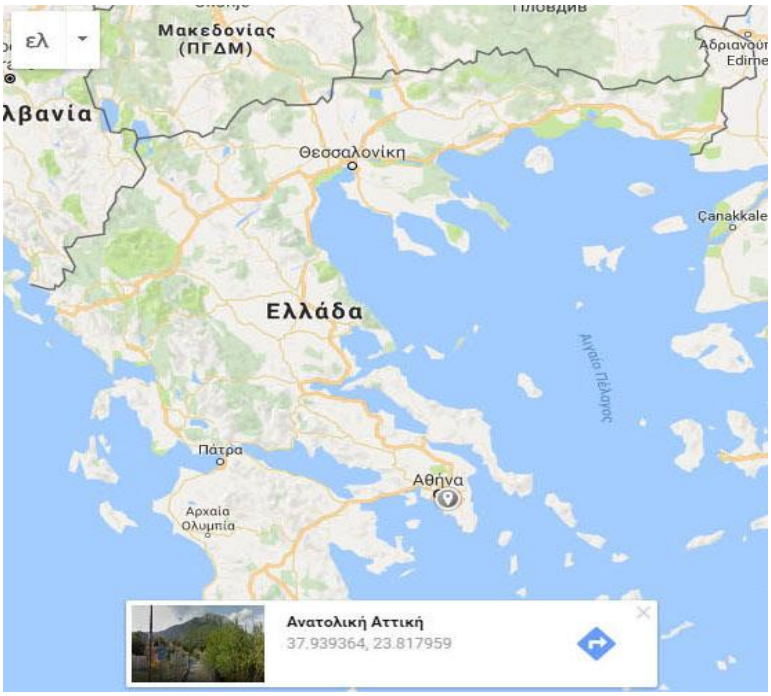
8. ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ

Πριν παρουσιαστεί η πειραματική αποτίμηση του προτεινόμενου αλγορίθμου, παρατίθενται κάποιες πληροφορίες για τον τρόπο με τον οποίο πραγματοποιήθηκε το προγραμματιστικό κομμάτι της εργασίας. Αρχικά, το σύνολο των υποσυστημάτων υλοποιήθηκαν σε περιβάλλον Windows και γλώσσα προγραμματισμού Java με τη βοήθεια του λογισμικού Eclipse. Για τα υποσυστήματα της κατηγοριοποίησης και της συσταδοποίησης χρησιμοποιήθηκε η Java διεπαφή (Application programming interface, API) του WEKA. Πιο συγκεκριμένα, με τη βοήθεια της υλοποιήθηκαν προγραμματιστικά οι αλγόριθμοι κατηγοριοποίησης Naïve Bayes και C4.5 (ή J48 για το WEKA) και ο αλγόριθμος συσταδοποίησης Expectation Maximization.

Όσον αφορά, τα απαραίτητα για τη λειτουργία της εφαρμογής του συστήματος συστάσεων, δεδομένα τα οποία εισήχθησαν για την πραγματοποίηση της πειραματικής αποτίμησης, προέρχονται από την ιστοσελίδα MovieLens. Στα δεδομένα περιλαμβάνονται μία λίστα 950 χρηστών περιλαμβανομένων των δημογραφικών τους χαρακτηριστικών, μία λίστα περισσότερων από 1500 ταινιών μαζί με περίπου 100.000 αξιολογήσεις που σχετίζονται τόσο με τους χρήστες όσο και με τις ταινίες. Οι παραπάνω λίστες έχουν δημιουργηθεί από δεδομένα που κατά καιρούς συλλέγει η ερευνητική ομάδα του GroupLens.

Ολοκληρώνοντας το κομμάτι της παρουσίασης της υλοποίησης του συστήματος συστάσεων, κρίνεται σκόπιμη η πραγματοποίηση μίας πειραματικής αποτίμησης για την αρτιότερη κατανόηση της λειτουργίας του. Ξεκινώντας, επιλέχθηκαν πέντε γεωγραφικά σημεία στα οποία εκτελείται ο αλγόριθμος. Τα σημεία αυτά έχουν μεταξύ τους απόσταση ανώτερη των εκατό χιλιομέτρων έτσι ώστε να είναι ορατή η επίδραση της γεωγραφικής θέσης στα αποτελέσματα. Θεωρούμε πως ο χρήστης περνά από αυτά τα σημεία και εισέρχεται στο σύστημα με σκοπό να δεχθεί τις συστάσεις. Το σύστημα με βάση τη γεωγραφική του θέση και άλλα στοιχεία παρέχει τις ανάλογες εξόδους.

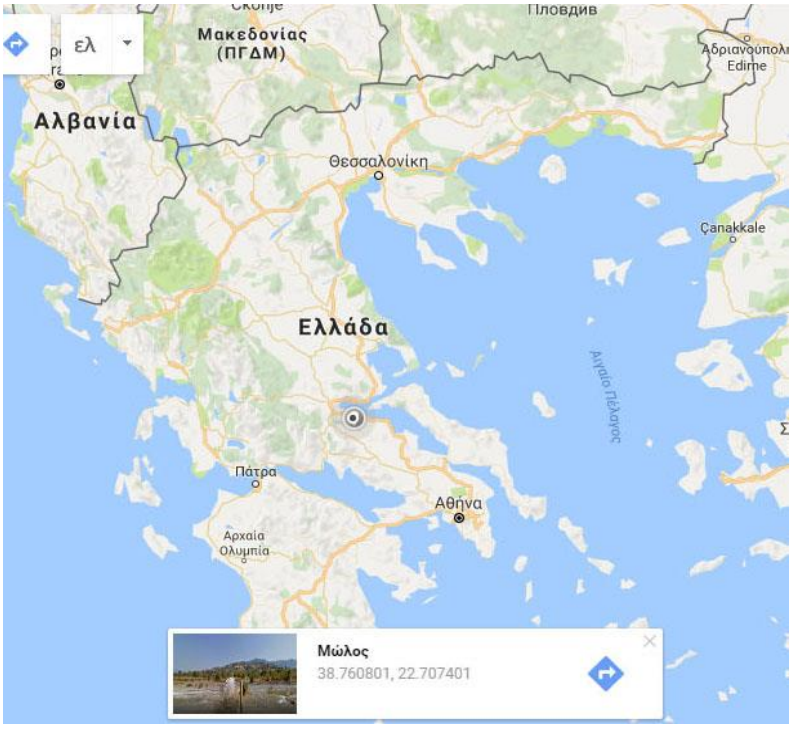
Για την διευκόλυνση της ανάλυσης της πειραματικής αποτίμησης παρουσιάζονται τα πέντε σημεία στα οποία ο χρήστη θα δεχτεί συστάσεις πάνω στο χάρτη. Ξεκινώντας, το πρώτο σημείο από το οποίο περνά ο πελάτης, έστω θέση Α, βρίσκεται στην περιοχή της ανατολικής Αττικής.



Εικόνα 10: Αφετηρία πελάτη (θέση Α)

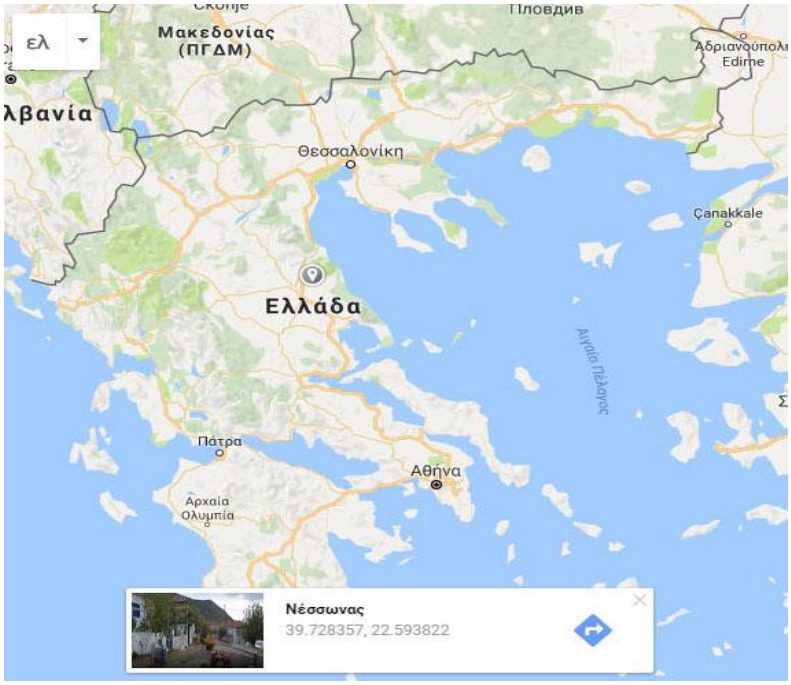
Το γεωγραφικό στίγμα είναι (37.939364, 23.817959) και παρουσιάζεται στην πιο πάνω εικόνα (Εικόνα 10).

Στη συνέχεια, ο πελάτης μετακινείται από την ανατολική Αττική περίπου 130 χιλιόμετρα βορειοδυτικά στη θέση Β (38.760801, 22.707401) κοντά στο χωριό Μώλος στο νομό Φθιώτιδας. Η ακριβής θέση του χρήστη παρουσιάζεται στην εικόνα 11.



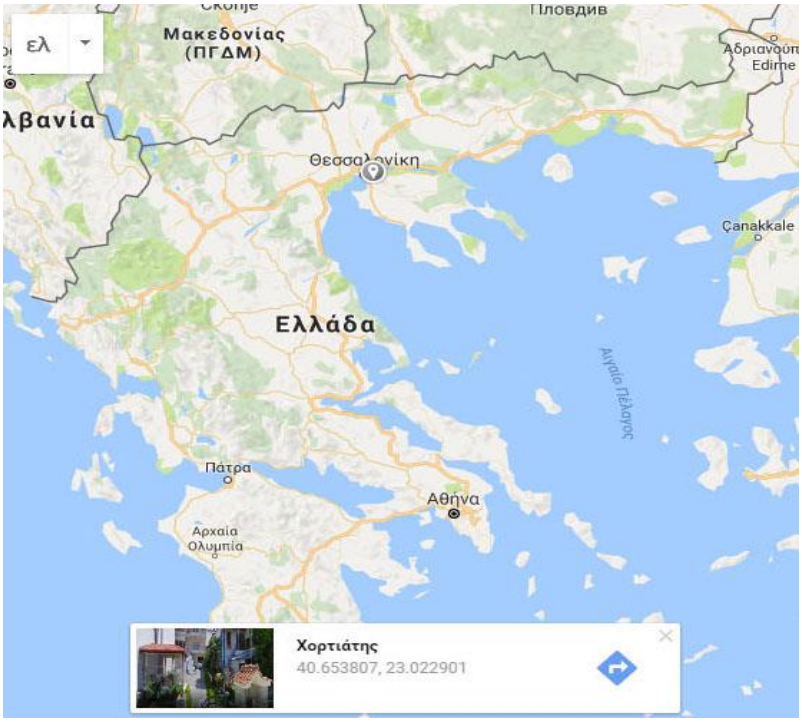
Εικόνα 11: Θέση Β, Μώλος

Ανεβαίνοντας 108 χιλιόμετρα βορειότερα φτάνει στο σημείο Γ (39.728357, 22.593822) έξω από το χωριό Νέσσωνας του Νομού Λάρισας (Εικόνα 12).



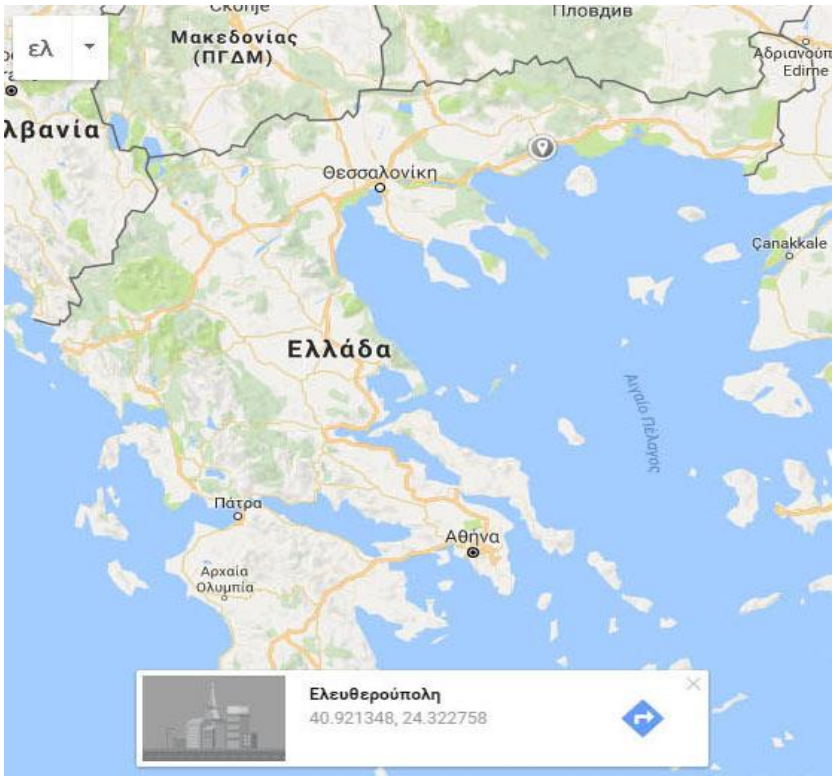
Εικόνα 12: Θέση Γ του πελάτη

Έπειτα πραγματοποιεί την τελευταία στάση στην κωμόπολη του Χορτιάτη στο σημείο Δ (40.653807, 23.022901), το οποίο φαίνεται στην εικόνα 13. Η απόσταση που χωρίζει το σημείο Γ από το σημείο Δ είναι 109 χιλιόμετρα.



Εικόνα 13: Θέση Δ, Χορτιάτης

Τέλος καταλήγει στην Ελευθερούπολη του Νομού Καβάλας στο σημείο Ε (40.921348, 24.322758), το οποίο φαίνεται στην εικόνα 14. Το σημείο Δ και Ε απέχουν 114 χιλιόμετρα.



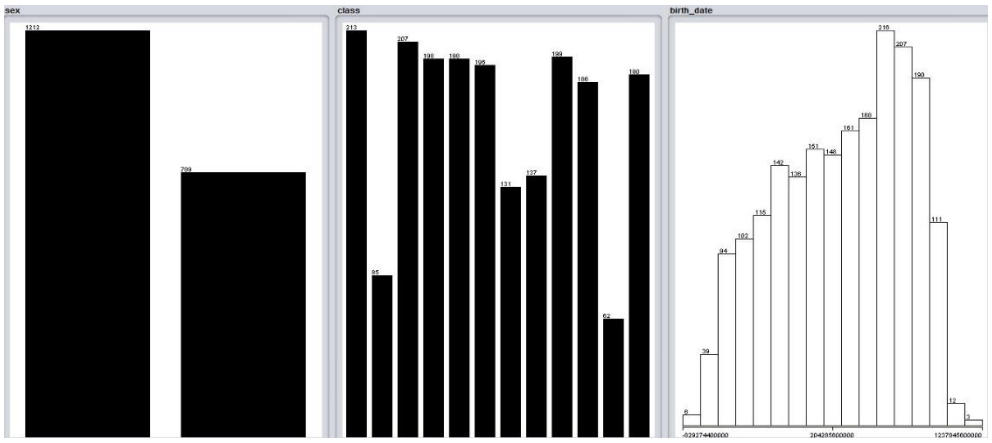
Εικόνα 14: Θέση Ε, Ελευθερούπολη

Για λόγους συντομίας για το υπόλοιπο της εργασίας οι τρείς θέσεις θα αναφέρονται με τα πέντε γράμματα του ελληνικού αλφάβητου, όπως ορίστηκαν πιο πάνω. Για την έναρξη της πειραματικής αποτίμησης θεωρούμε πως ο χρήστης είναι καινούργιος στο σύστημα και πραγματοποιεί την εγγραφή του στη αφετηρία της διαδρομής στη θέση Α. Η εγγραφή του καινούργιου χρήστη θέτει σε λειτουργία το υποσύστημα της κατηγοριοποίησης τα αποτελέσματα του οποίου παρουσιάζονται στην παράγραφο 8.1.

8.1 Παρουσίαση Αποτελεσμάτων Κατηγοριοποίησης

Η εκτέλεση του αλγορίθμου της κατηγοριοποίησης γίνεται με τη βοήθεια της εφαρμογής του WEKA. Για την υλοποίηση του υποσυστήματος όπως αναφέρθηκε στο κεφάλαιο 5 της εργασίας χρησιμοποιήθηκαν δύο αλγόριθμοι: ο Naïve Bayes και ο C4.5, τα αποτελέσματα των οποίων παρουσιάζονται παρακάτω.

Ξεκινώντας τη διαδικασία μέσω της εφαρμογής του WEKA και κατά την εισαγωγή των δεδομένων, το σύστημα κάνει τα αναλύει και προσφέρει μία γραφική απεικόνιση της φύσης τους. Δίνει δηλαδή μία άποψη των δεδομένων από πλευράς στατιστικής. Ο τρόπος με τον οποίο παρουσιάζονται τα στοιχεία φαίνεται στην εικόνα 15.



Εικόνα 15: Δημογραφικά δεδομένα κατά την εκτέλεση της εφαρμογής του WEKA

Στο αριστερό μέρος της εικόνας διακρίνεται το φύλο των πελατών. Για το δείγμα πελατών που χρησιμοποιήθηκε ο αριθμός των αντρών υπερτερεί έναντι αυτού των γυναικών. Πιο συγκεκριμένα ο αριθμός των αντρών ανέρχεται σε 1212 ενώ των γυναικών σε 789. Στη συνέχεια, στο κέντρο της οθόνης παρατηρούμε την κατανομή των χρηστών σε κλάσεις, ενώ, τέλος στο δεξί άκρο της εικόνας παρατηρούμε την αντίστοιχη κατανομή σε διαστήματα ηλικιών. Αξίζει να σημειωθεί πως για την εκπαίδευση του μοντέλου δε χρησιμοποιήθηκαν όλα τα χαρακτηριστικά παρά μόνο όσα ήταν χρήσιμα (ηλικία, κλάση και φύλο).

8.1.1 Παρουσίαση Αποτελεσμάτων αλγορίθμου Naïve Bayes

Προχωρώντας, στην εκπαίδευση του μοντέλου γίνεται αρχικά επιλογή του αλγορίθμου Naïve Bayes και ακολουθεί η έξοδος της εφαρμογής.

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     QueryResult
Instances:    2001
Attributes:   3
              sex
              class
              decate
Test mode:    split 20.0% train, remainder test
```

Εικόνα 16: Πληροφορίες εκτέλεσης

Το πρώτο βήμα είναι εξαγωγή κάποιων πληροφοριών εκτέλεσης (Εικόνα 16). Πιο συγκεκριμένα, το σχήμα του αλγορίθμου που χρησιμοποιήθηκε (Naïve Bayes στη συγκεκριμένη περίπτωση), τον αριθμό των εγγραφών που δέχθηκε ο αλγόριθμος ως είσοδο (2001 στο συγκεκριμένο παράδειγμα) τα πεδία της βάσης που χρησιμοποιήθηκαν (φύλο, κλάση και ηλικία) και τέλος το ποσοστό επί των πελατών που χρησιμοποιήθηκε ως σετ εκπαίδευσης. Στο συγκεκριμένο παράδειγμα χρησιμοποιήθηκε 20% του δείγματος, ποσοστό που θεωρείται ικανοποιητικό για τη δημιουργία ενός αξιόπιστου μοντέλου.

Παρακάτω, δίνεται μία απεικόνιση του μοντέλου, μετά το πέρας της εκπαίδευσης, σε μορφή κειμένου (Εικόνα 17).

```
Naive Bayes Classifier

                Class
Attribute           2       5       3       4       0       1       6       7
                 (0.27)  (0.16)  (0.21)  (0.19)     (0)  (0.08)  (0.09)     (0)
=============================================================================
sex
  man            351.0   183.0   270.0   218.0     2.0    96.0    96.0     4.0
  woman          183.0   136.0   157.0   165.0     1.0    70.0    84.0     1.0
  [total]        534.0   319.0   427.0   383.0     3.0   166.0   180.0     5.0

class
  man_twenty     351.0     1.0     1.0     1.0     1.0     1.0     1.0     1.0
  woman_fifty      1.0   136.0     1.0     1.0     1.0     1.0     1.0     1.0
  woman_thirty     1.0     1.0   157.0     1.0     1.0     1.0     1.0     1.0
  man_fourty       1.0     1.0     1.0   218.0     1.0     1.0     1.0     1.0
  man_fifty        1.0   183.0     1.0     1.0     1.0     1.0     1.0     1.0
  man_thirty       1.0     1.0   270.0     1.0     1.0     1.0     1.0     1.0
  woman_twenty   183.0     1.0     1.0     1.0     1.0     1.0     1.0     1.0
  woman_fourty     1.0     1.0     1.0   165.0     1.0     1.0     1.0     1.0
  man_zero         1.0     1.0     1.0     1.0     2.0     1.0     1.0     1.0
  woman_ten        1.0     1.0     1.0     1.0     1.0    70.0     1.0     1.0
  man_ten          1.0     1.0     1.0     1.0     1.0    96.0     1.0     1.0
  man_sixty        1.0     1.0     1.0     1.0     1.0     1.0    96.0     1.0
  woman_sixty      1.0     1.0     1.0     1.0     1.0     1.0    84.0     1.0
  man_seventy      1.0     1.0     1.0     1.0     1.0     1.0     1.0     4.0
  [total]        546.0   331.0   439.0   395.0    15.0   178.0   192.0    17.0
```

Εικόνα 17: Απεικόνιση του μοντέλου σε μορφή κειμένου

Σύμφωνα με το μοντέλο το 27% του πληθυσμού (πελάτες που δόθηκαν ως είσοδος) διανύει τη δεύτερη δεκαετία της ζωής του. Το 16% την πέμπτη, το 21% την Τρίτη, το 19% την τέταρτη το 0,08% την πρώτη και το 0.09% την έκτη. Επιπρόσθετα δίνονται κάποιες πληροφορίες για το πλήθος των αντρών και των γυναικών που βρίσκονται σε κάθε μία από τις παραπάνω δεκαετίες της ζωής τους.

Μετά την απεικόνιση του μοντέλου, εμφανίζεται η ώρα που χρειάστηκε η εφαρμογή για την αποτίμηση του με τη βοήθεια του δείγματος δοκιμής.

```
=== Evaluation on test split ===

Time taken to test model on training split: 0.01 seconds
```

Εικόνα 18: Ο χρόνος εκτέλεσης του αλγορίθμου

Για τη συγκεκριμένη εκτέλεση χρειάστηκαν, μόλις, 0,01 δευτερόλεπτα, ώστε να δοκιμαστεί το μοντέλο (Εικόνα 18).

```
=== Summary ===

Correctly Classified Instances        1599               99.8751 %
Incorrectly Classified Instances         2                0.1249 %
Kappa statistic                          0.9985
Mean absolute error                      0.0348
Root mean squared error                  0.0616
Relative absolute error                 17.1698 %
Root relative squared error             19.3581 %
Total Number of Instances             1601
```

Εικόνα 19: Πληροφορίες ακρίβειας του μοντέλου

Εν συνεχεία δίνονται πληροφορίες για την ακρίβεια του αλγορίθμου (Εικόνα 19). Πιο συγκεκριμένα, η εφαρμογή παρουσιάζει μία λίστα με στατιστικά που ενημερώνουν για το πόσο ακριβές ήταν το μοντέλο στην πρόβλεψη της σωστής κλάσης. Στο παράδειγμα, αφού η εφαρμογή διαχώρισε το 20% του δείγματος (400 πελάτες) από το δείγμα, κατηγοριοποίησε τους υπόλοιπους χίλιους εξακόσιους έναν πελάτες (1601) με ακρίβεια 99,8751%. Δηλαδή μόλις σε δύο πελάτες στους 1601 αποδόθηκε λάθος κλάση. Το αποτέλεσμα του αλγορίθμου θα μπορούσε να χαρακτηριστεί εξαιρετικό, για να εξαχθεί, όμως, οποιοδήποτε συμπέρασμα θα πρέπει να συγκριθεί και με το αποτέλεσμα του δεύτερο αλγορίθμου (C4.5 βλ. παράγραφο 8.1.2).

Τέλος, παρατίθεται ένας πίνακας σύγχυσης ή αλλιώς confusion matrix στον οποίο παρουσιάζονται αναλυτικά τα αποτελέσματα κάθε κλάσης και στον οποίο φαίνεται ποια κλάση θα έπρεπε να αποδοθεί και ποια αποδόθηκε τελικά. Στον παρακάτω πίνακα της εικόνας 20 στην τελευταία σειρά προκύπτει πως τα δύο λάθη του αλγορίθμου προέρχονται από την έβδομη κλάση στα στοιχεία της οποίας ο αλγόριθμος απέδωσε λανθασμένα την τρίτη κλάση.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 428   0   0   0   0   0   0   0 |   a = 2
   0 259   0   0   0   0   0   0 |   b = 5
   0   0 335   0   0   0   0   0 |   c = 3
   0   0   0 303   0   0   0   0 |   d = 4
   0   0   0   0   0   0   0   0 |   e = 0
   0   0   0   0   0 125   0   0 |   f = 1
   0   0   0   0   0   0 149   0 |   g = 6
   0   0   2   0   0   0   0   0 |   h = 7
```

Εικόνα 20: Confusion matrix

8.1.2 Παρουσίαση Αποτελεσμάτων του Αλγορίθμου C4.5

Το WEKA δημιουργεί με παρόμοιο τρόπο τα αποτελέσματα για τον αλγόριθμο C4.5. Η διαφορά των δύο αλγορίθμων είναι πως το μοντέλο που δημιουργείται κατά το πέρας της διαδικασίας από τον C4.5 έχει τη μορφή ενός δέντρου τα φύλλα του οποίου αποτελούν τις κλάσεις των πελατών. Τα δεδομένα που εισήχθησαν για τον αλγόριθμο Naïve Bayes είναι κοινά με αυτά του παρόντος αλγορίθμου και το δέντρο που δημιουργείται παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 21).

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

class = man_twenty: 2 (350.0)
class = woman_fifty: 5 (135.0)
class = woman_thirty: 3 (156.0)
class = man_fourty: 4 (217.0)
class = man_fifty: 5 (182.0)
class = man_thirty: 3 (269.0)
class = woman_twenty: 2 (182.0)
class = woman_fourty: 4 (164.0)
class = man_zero: 0 (1.0)
class = woman_ten: 1 (69.0)
class = man_ten: 1 (95.0)
class = man_sixty: 6 (95.0)
class = woman_sixty: 6 (83.0)
class = man_seventy: 7 (3.0)

Number of Leaves  :     14

Size of the tree :      15
```

Εικόνα 21: Το μοντέλου του αλγορίθμου C4.5 σε μορφή κειμένου

Το δέντρο διαθέτει 14 φύλλα όσες είναι και οι κλάσεις στις οποίες διαμοιράστηκε το δείγμα των πελατών. Ο χρόνος που χρειάστηκε για τη δημιουργία του μοντέλου δε διαφέρει σε σχέση με τον αλγόριθμο Naïve Bayes. Προχωρώντας στην ανάλυση της απόδοσης του αλγορίθμου διαπιστώνουμε πως ο C4.5 ξεπερνά τα, ήδη εξαιρετικά, αποτελέσματα του Naïve Bayes.

```
=== Summary ===

Correctly Classified Instances        1601               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             1601
```

Εικόνα 22: Ανάλυση της απόδοσης του αλγορίθμου C4.5

Πιο συγκεκριμένα, καταφέρνει μέσα στον ίδιο χρόνο να επιτύχει το απόλυτο, δηλαδή να κατηγοριοποιήσει τους πελάτες στις σωστές κλάσεις με 100% επιτυχία (Εικόνα 22). Τα παραπάνω αποτελέσματα επιβεβαιώνονται και από τον confusion matrix που παρατίθεται στην από κάτω εικόνα (Εικόνα 23).

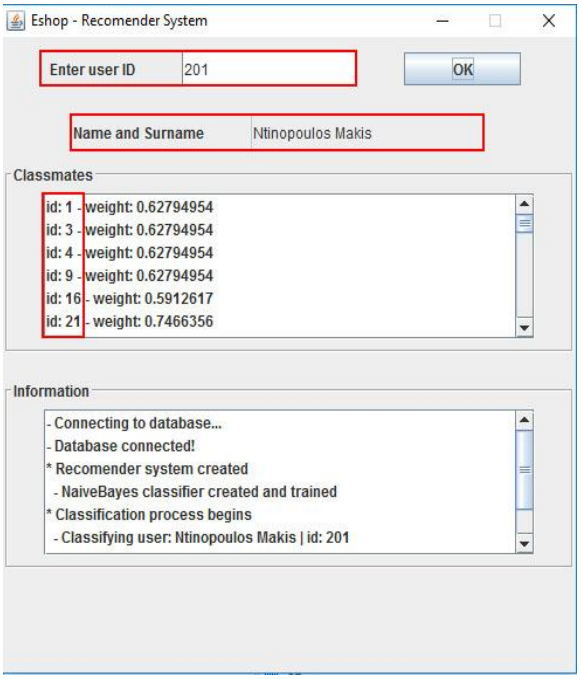
```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 428   0   0   0   0   0   0   0 |   a = 2
   0 259   0   0   0   0   0   0 |   b = 5
   0   0 335   0   0   0   0   0 |   c = 3
   0   0   0 303   0   0   0   0 |   d = 4
   0   0   0   0   0   0   0   0 |   e = 0
   0   0   0   0   0 125   0   0 |   f = 1
   0   0   0   0   0   0 149   0 |   g = 6
   0   0   0   0   0   0   0   2 |   h = 7
```

Εικόνα 23: Confusion matrix του αλγορίθμου C4.5

Σύμφωνα με τον παραπάνω πίνακα και σε σύγκριση με τον αντίστοιχο πίνακα του αλγορίθμου Naïve Bayes, δεν υπήρξε καμία απώλεια (λάθος απόδοση κλάσης) σε καμία από τις κλάσεις των πελατών.

8.1.3 Αποτελέσματα των αλγορίθμων σε προγραμματιστικό περιβάλλον

Τα παραπάνω αποτελέσματα παρουσιάζονται στη διεπαφή χρήστη που χρησιμοποιείται για την προβολή των εξόδων του συνόλου του συστήματος. Τα αποτελέσματα που προκύπτουν δεν είναι συγκεντρωτικά, αλλά αναφέρονται σε ένα και μοναδικό πελάτη. Η διαδικασία της εκτέλεσης του αλγορίθμου ξεκινά με την επιλογή του μοναδικού αριθμού ID του πελάτη του ηλεκτρονικού καταστήματος και στη συνέχεια το σύστημα ξεκινά να την παρουσίαση όπως προκύπτει από την παρακάτω εικόνα (Εικόνα 24).

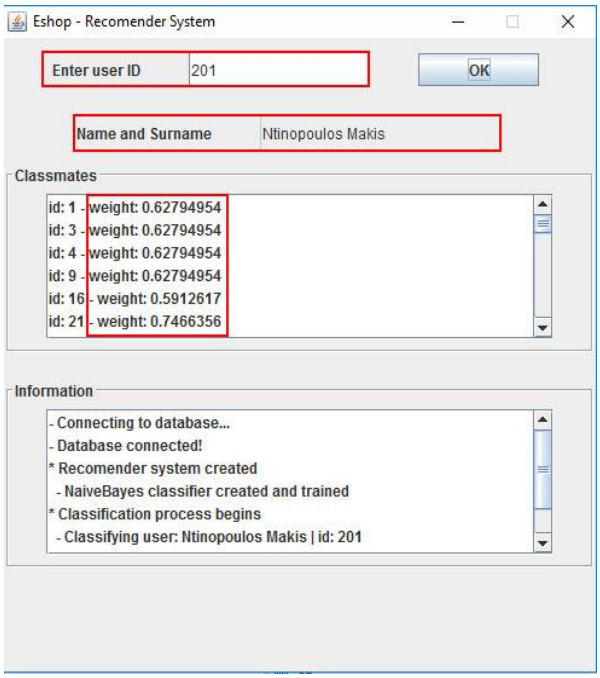


Εικόνα 24: Εμφάνιση του id, του ονόματος και των χρηστών με ίδια τιμή κλάσης από τη διεπαφή χρήστη

Αρχικά, με βάση το ID προσδιορίζεται το όνομα του πελάτη, ενώ στη συνέχεια με έναν από τους παραπάνω αλγορίθμους του αποδίδεται η κλάση. Στη συνέχεια εμφανίζονται όλοι οι χρήστες που μοιράζονται την ίδια τιμή κλάσης με αυτόν. Με αυτόν το τρόπο ολοκληρώνεται η διαδικασία κατηγοριοποίησης του πελάτη και το σύστημα προχωρά στον υπολογισμό των βαρών.

8.2 Αποτελέσματα Υπολογισμού Βαρών

Για να προκύψει η το βάρος του ζευγαριού {χρήστη-βαθμολογίας}, το υποσύστημα λαμβάνει ως είσοδο την έξοδο του υποσυστήματος της κατηγοριοποίησης. Για ένα συγκεκριμένο χρήστη βρίσκει τους γείτονες του (τους χρήστες με τους οποίους έχουν την ίδια τιμή κλάσης) και υπολογίζει επαναληπτικά το βάρος των ζευγαριών του συγκεκριμένου χρήστη και κάθε βαθμολογίας που πραγματοποίησαν οι γείτονες. Για κάθε ζευγάρι ξεχωριστά, εφαρμόζονται και οι τρεις συναρτήσεις που παρουσιάστηκαν στο κεφάλαιο 5. Τα αποτελέσματα του υποσυστήματος εμφανίζονται κατά την εκτέλεση του αλγορίθμου δίπλα στο μοναδικό αριθμό κάθε βαθμολογίας. Ο τρόπος εμφάνισης των αποτελεσμάτων στη διεπαφή παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 25). Με κόκκινο χρώμα είναι τονισμένα το id του πελάτη, ο οποίος θα είναι ο τελικός δέκτης των συστάσεων, το όνομα και το επίθετο του καθώς και η λίστα με τις βαθμολογίες των πελατών του ηλεκτρονικού καταστήματος που έχουν την ίδια τιμή κλάσης με το χρήστη αυτό.



Εικόνα 25: Εμφάνιση των βαρών κάθε ζευγαριού χρήστη-βαθμολογίας στη διεπαφή χρήστη

Η έξοδος της διαδικασίας υπολογισμού των βαρών, χρησιμοποιείται ως είσοδος του υποσυστήματος της συσταδοποίησης.

8.3 Παρουσίαση Αποτελεσμάτων Συσταδοποίησης

Όπως προκύπτει από την παρακάτω εικόνα (Εικόνα 26), η εφαρμογή, αρχικά παρέχει κάποιες πληροφορίες που αφορούν την εκτέλεση του αλγορίθμου. Αρχικά, εμφανίζεται το όνομα και τα στοιχεία αρχικοποίησης, το πιο σημαντικό από τα οποία είναι ο αριθμός των συστάδων που θα προκύψουν κατά την εκτέλεση του (-N 2, στην πρώτη σειρά της εικόνας 26). Αναγράφεται, ακόμα, ο αριθμό των στοιχείων που εμπλέκονται στη συσταδοποίηση και τέλος αναφέρονται τα χαρακτηριστικά των βαθμολογιών που συνετέλεσαν στη διαμόρφωση του αποτελέσματος. Στην πιο κάτω εκτέλεση, τα στοιχεία που χρησιμοποιήθηκαν ήταν 198 σε αριθμό, ενώ, τα χαρακτηριστικά που χρησιμοποιήθηκαν ήταν ο μοναδικός αριθμός της βαθμολογίας των χρηστών και το βάρος του (weight) με βάση το οποίο πραγματοποιήθηκε η συσταδοποίηση.

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N 2 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     QueryResult
Instances:    198
Attributes:   2
              classmate_id
              weight
Test mode:    evaluate on training data
```

Εικόνα 26: Πληροφορίες εκτέλεσης του αλγορίθμου EM

Στη συνέχεια, η εφαρμογή εκτυπώνει σε μορφή κειμένου το μοντέλο που δημιουργήθηκε (Εικόνα 27), από το οποίο μπορούν να εξαχθούν πληροφορίες όπως ο μέσος όρος των βαρών σε κάθε ομάδα ή το ποσοστό των στοιχείων που ανήκουν σε κάθε μία από τις συστάδες αυτές.

```
=== Clustering model (full training set) ===


EM
==

Number of clusters: 2
Number of iterations performed: 100


                Cluster
Attribute            0        1
                  (0.42)   (0.58)
===================================
classmate_id
  mean          1563.5441  547.3629
  std. dev.      250.0456  368.2308

weight
  mean             0.6363   0.6471
  std. dev.         0.086    0.055


Time taken to build model (full training data) : 0.02 seconds
```

Εικόνα 27: Το μοντέλο και ο χρόνος δημιουργίας του απο τον αλγόριθμο Expectation Maximization

Τέλος, προσφέρονται περισσότερες πληροφορίες σχετικά με τον αριθμό των στοιχείων που περιέχονται σε κάθε συστάδα μαζί με τη λογαριθμική πιθανοφάνεια (log likelihood) που λειτουργεί ως κατώφλι για την επιλογή της συστάδας.

```
=== Model and evaluation on training set ===

Clustered Instances

0       83 ( 42%)
1      115 ( 58%)


Log likelihood: -6.40677
```

Εικόνα 28: Αποτελέσματα αποτίμησης μοντέλου συσταδοποίσης

Στην παραπάνω εικόνα (Εικόνα 28), ο αριθμός του log likelihood είναι αρνητικός. Η πιο πάνω παρατήρηση είναι απόλυτα λογική καθώς αναφέρεται στο λογάριθμο της πιθανότητας και όχι στην κοινή πιθανότητα που παίρνει τιμές στο [0,1].

8.4 Παρουσίαση Εξόδων Συστήματος

Για τη μελέτη της συμπεριφοράς του συστήματος δημιουργήθηκαν 3 διαφορετικά σενάρια, τα οποία διαφέρουν μεταξύ τους στον τρόπο υπολογισμού των βαρών. Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι εξισώσεις 9, 10 και 11 που παρουσιάστηκαν στο Κεφάλαιο 5 για τον υπολογισμό του βάρους της ηλικίας, της γεωγραφικής θέσης και του χρόνου με α=0,3, β=0,6, γ=0,05. Το σημείο, όμως, στο οποίο διαφέρει κάθε σενάριο είναι η συνεισφορά κάθε ενός από τα παραπάνω βάρη στον υπολογισμού του συνολικού βάρους και για την ακρίβεια οι αριθμοί κ, λ και μ της εξίσωσης 12 του Κεφαλαίου 5. Ο διαφορετικός υπολογισμός του συνολικού βάρους σημαίνει διαφορετικά βάρη για ίδια ζευγάρια βαθμολογίας-χρήστη (από σενάριο σε σενάριο), που οδηγεί σε διαφορετική συσταδοποίηση και τελικά σε διαφοροποιημένη λίστα συστάσεων.

Το πρώτο σενάριο λαμβάνει υπόψη του μόνο την τοποθεσία για τη διαμόρφωση του βάρους. Έτσι, οι βαθμολογίες που πραγματοποιήθηκαν μακρύτερα από το κατώφλι που θέτει η συνάρτηση locSim (εξίσωση 10) αυτομάτως λαμβάνουν βάρος μηδέν, κατατάσσονται στο cluster 1 και αγνοούνται. Από την άλλη πλευρά, οι βαθμολογίες οι οποίοι βρίσκονται εντός της ακτίνας που ορίζει το κατώφλι, διαγωνίζονται και ανάλογα το ποιος είναι πιο κοντά στον πελάτη (συνεπάγεται μεγαλύτερο βάρος) κερδίζουν μία θέση στη συστάδα υψηλού ενδιαφέροντος.

Στο δεύτερο σενάριο, το μίγμα των χαρακτηριστικών αλλάζει, καθώς η γεωγραφική θέση διαμορφώνει το 80% του βάρους, η ηλικία το 10% και ο χρόνος το υπόλοιπο 10%. Εδώ παρατηρείται μία αύξηση του αριθμού των μακρινών (σε σχέση με το χρήστη που δέχεται συστάσεις) βαθμολογήσεων που εισέρχονται στη συστάδα υψηλού ενδιαφέροντος. Οι βαθμολογήσεις που πραγματοποιήθηκαν πολύ κοντά στη θέση του χρήστη, βέβαια, συνεχίζουν να έχουν τον πρώτο λόγο στη διαμόρφωση των προβλέψεων. Χαρακτηριστικό αυτού του σεναρίου είναι η διεύρυνση της ακτίνας στην οποία έχουν πραγματοποιηθεί οι βαθμολογήσεις που ανήκουν στη συστάδα υψηλού ενδιαφέροντος.

Στο τρίτο σενάριο, το βάρος υπολογίζεται κατά 50% από τη γεωγραφική θέση του πελάτη, σε ποσοστό 20% από την ηλικία του και 30% από το χρόνο που πραγματοποιήθηκαν οι βαθμολογήσεις. Αυτό σημαίνει, πρακτικά, πως μπορούν να επιλεγούν βαθμολογίες πελατών οι οποίες δεν πραγματοποιήθηκαν τόσο κοντά στο χρήστη, όμως, ο βαθμολογητής, πιθανότατα, έχει παρόμοια ηλικία ή ο χρόνος της βαθμολογίας είναι σχετικά κοντινός. Με πιο απλά λόγια, το έδαφος που χάνει ο βαθμολογητής ώστε να εισέλθει η βαθμολογία του στη συστάδα υψηλού ενδιαφέροντος, αναπληρώνεται είτε από την ηλικία του (κοντινή με αυτή του χρήστη) είτε απο τη χρονική στιγμή που βαθμολόγησε, είτε και από τα δύο μαζί.

8.4.1 Πρώτο σενάριο (μ=1, λ=0, κ=0)

Για το πρώτο σενάριο, αντικαθίστανται, στην εξίσωση 12 του κεφαλαίου 5, οι αριθμοί 1, 0, 0 στη θέση των μ, λ και κ αντίστοιχα. Αυτό, πρακτικά, σημαίνει ότι για τον υπολογισμό του συνολικού βάρους, αγνοείται τόσο η τιμή του βάρους της ηλικίας, όσο και αυτή του χρόνου. Σύμφωνα, λοιπόν, με τη συνάρτηση locSim (Εξίσωση 10), η οποία είναι η μοναδική που συνεισφέρει, και με το β που επιλέχθηκε, η τιμή της τοποθεσίας μηδενίζεται ή τείνει προς το μηδέν για διαφορά μεγαλύτερη των εκατό χιλιομέτρων. Δηλαδή, κάθε βαθμολογία που βρίσκεται εντός της ακτίνας των εκατό χιλιομέτρων από το χρήστη, που θα γίνει δέκτης των συστάσεων, μπορεί να βρεθεί με σχετικά μεγάλη πιθανότητα στη συστάδα υψηλού ενδιαφέροντος και να προσφέρει στον υπολογισμό των προβλέψεων. Αντίστροφα, αν κάποια βαθμολογία βρίσκεται έξω από την ακτίνα των εκατό χιλιομέτρων τότε αναπόφευκτα το βάρος του ζευγαριού της με τον πελάτη θα είναι μηδέν (συστάδα χαμηλού ενδιαφέροντος), με αποτέλεσμα να αγνοηθεί.

Τα αποτελέσματα του σεναρίου 3 για τις πέντε τοποθεσίες παρουσιάζονται στην παρακάτω εικόνα (Εικόνα 29).

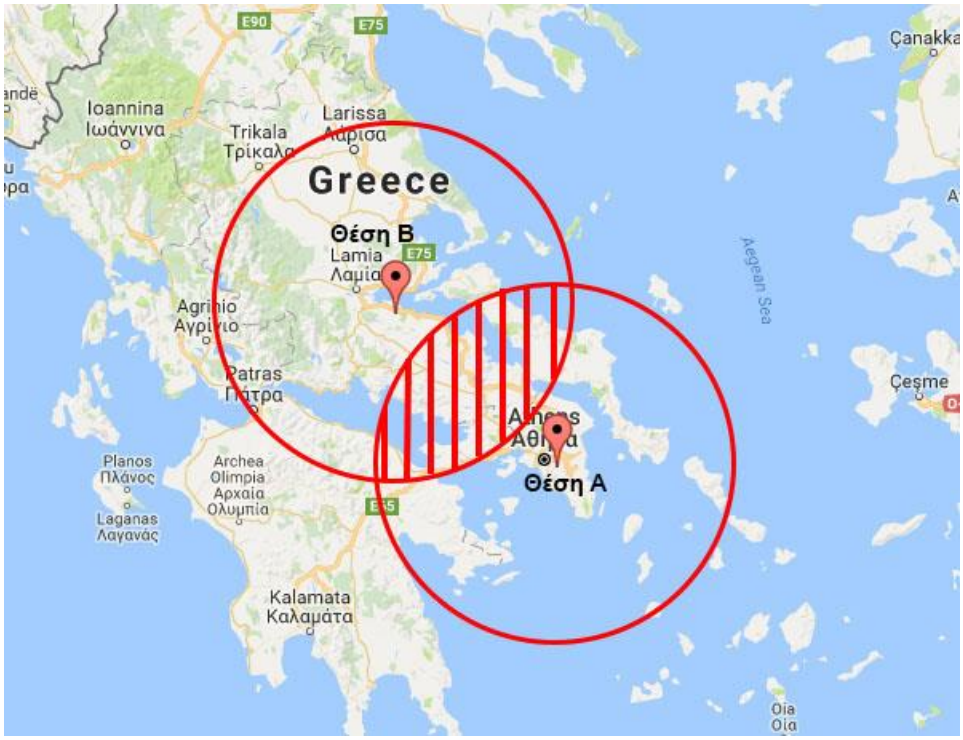
```
Θέση Α
Rating Prediction => Item:50, Count Ratings: 17, Prediction: 4.8235
Rating Prediction => Item:98, Count Ratings: 12, Prediction: 4.5833
Rating Prediction => Item:79, Count Ratings: 11, Prediction: 4.5455
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.4286

Θέση Β
Rating Prediction => Item:127, Count Ratings: 13, Prediction: 4.6154
Rating Prediction => Item:174, Count Ratings: 15, Prediction: 4.6
Rating Prediction => Item:50, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:79, Count Ratings: 12, Prediction: 4.25

Θέση Γ
Rating Prediction => Item:98, Count Ratings: 12, Prediction: 4.3333
Rating Prediction => Item:174, Count Ratings: 18, Prediction: 4.2778
Rating Prediction => Item:50, Count Ratings: 13, Prediction: 4.1538
Rating Prediction => Item:172, Count Ratings: 16, Prediction: 4.125


Θέση Δ
Rating Prediction => Item:150, Count Ratings: 13, Prediction: 4.4615
Rating Prediction => Item:96, Count Ratings: 15, Prediction: 4.4
Rating Prediction => Item:50, Count Ratings: 17, Prediction: 4.2941
Rating Prediction => Item:181, Count Ratings: 12, Prediction: 4.25

Θέση Ε
Rating Prediction => Item:313, Count Ratings: 11, Prediction: 4.5455
Rating Prediction => Item:50, Count Ratings: 13, Prediction: 4.3846
Rating Prediction => Item:176, Count Ratings: 12, Prediction: 4.25
Rating Prediction => Item:237, Count Ratings: 12, Prediction: 4.1667
```

Εικόνα 29: Αποτελέσματα αλγορίθμου για το πρώτο σενάριο εκτέλεσης

Η έξοδος του συστήματος περιλαμβάνει, εκτός από το μέσο όρο των βαθμολογιών που λαμβάνονται υπόψη στην πρόβλεψη (Prediction), το πλήθος τους (count ratings) καθώς και το ID του προϊόντος (Item). Εδώ αξίζει να σημειωθεί πως οι προβλέψεις ταξινομούνται, αρχικά, με βάση την τιμή της βαθμολογίας και στη συνέχεια με βάση το πλήθος των βαθμολογιών (ratings) που συνέβαλαν στη διαμόρφωση της τιμής τους. Αυτό σημαίνει πως αν υπάρχουν πέντε προϊόντα προς σύσταση και έχουν όλα την ίδια βαθμολογία, προτεραιότητα για το σύστημα έχουν τα προϊόντα με το μεγαλύτερο πλήθος συμβαλλόμενων βαθμολογιών (count rating).

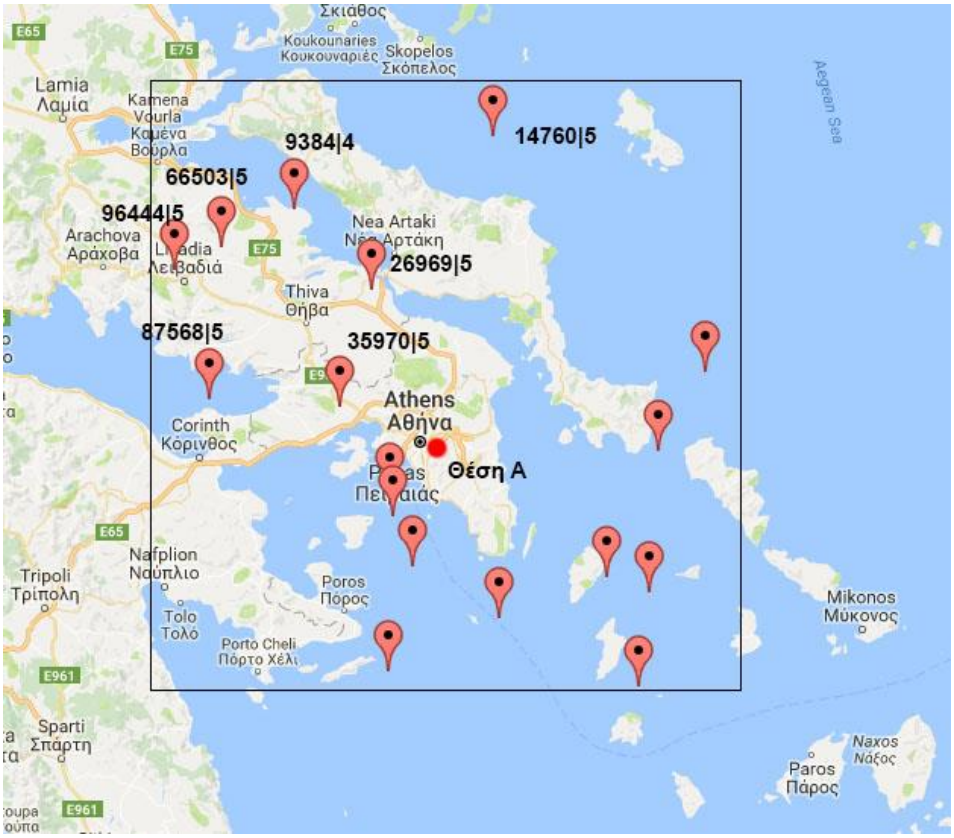
Παρατηρώντας πιο προσεκτικά τα αποτελέσματα της εικόνας 29 για το πρώτο σενάριο μπορεί εύκολα να διαπιστωθεί πως η σύσταση κάποιων προϊόντων επαναλαμβάνεται. Για παράδειγμα, στη θέση Α τρία από τα τέσσερα προϊόντα που συστήνονται, είναι κοινά με αυτά της λίστας συστάσεων της θέσης Β (ID 50, 79 και 174). Συνεχίζοντας, από τη θέση Β στη θέση Γ υπάρχει επανάληψη των συστάσεων για τα προϊόντα με ID 174 και 50. Το ίδιο συμβαίνει και για τα ζευγάρια των επόμενων θέσεων.

Ο λόγος για τον οποίο συμβαίνει αυτή η επανάληψη συστάσεων σε δύο κοντινές θέσεις είναι η αλληλοκάλυψη του χώρου (με ακτίνα περίπου 100 χιλιομέτρων) που δημιουργεί το κατώφλι της συνάρτησης βάρους της γεωγραφικής θέσης. Η αλληλοκάλυψη φαίνεται γραμμοσκιασμένη στην παρακάτω εικόνα (Εικόνα 30). Μέσα στο χώρο αυτό έχουν πραγματοποιηθεί βαθμολογήσεις οι οποίες συμβάλλουν στη διαμόρφωση των τιμών των προβλέψεων τόσο για τη θέση Α όσο και για τη θέση Β.



Εικόνα 30: Ο χώρος που ορίζει το κατώφλι της συνάρτησης locSim

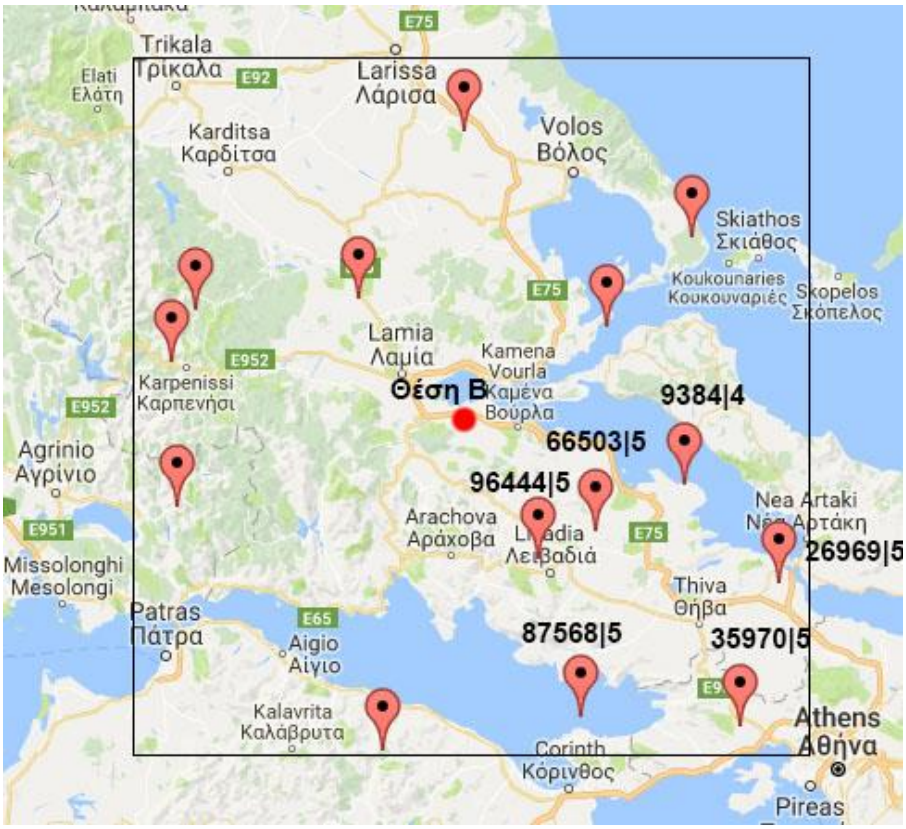
Ο πιο πάνω ισχυρισμός, αποδεικνύεται από το χάρτη των θέσεων των βαθμολογιών για το προϊόν με ID 50 για τη θέση Α, ο οποίος παρουσιάζεται στην εικόνα 31 σε αντιστοιχία με τον χάρτη με το χάρτη για τη θέση Β της εικόνας 32.



Εικόνα 31: Χάρτης βαθμολογιών για το προϊόν με id 50 στη θέση Α

Όπως φαίνεται από της εικόνα (Εικόνα 31) οι συνολικά 17 βαθμολογίες (με μέσο όρο 4,8236) περικλείονται σε ένα νοητό παραλληλόγραμμο το οποίο έχει διαστάσεις

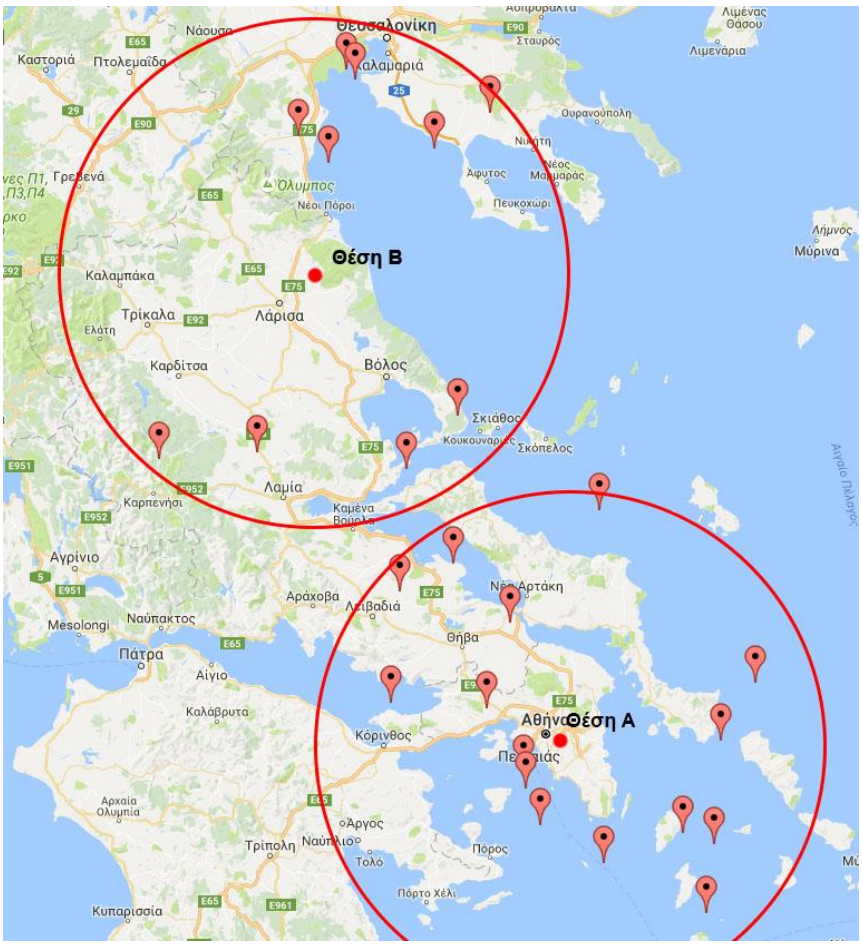
(174,28*167,97)km. Οι θέσεις των βαθμολογιών αναπαρίστανται με ένα κόκκινο δείκτη, ενώ η θέση του χρήστη παρουσιάζεται με μία κόκκινη τελεία.



Εικόνα 32: Χάρτης βαθμολογιών για το προϊόν με id 50 στη θέση Β

Προχωρώντας στη θέση Β, για την πρόβλεψη συνεισέφεραν 14 βαθμολογίες, οι θέσεις των οποίων παρουσιάζονται στο χάρτη της εικόνας 32. Περικλείονται σε ένα νοητό παραλληλόγραμμο με διαστάσεις (157,59*158,3)km ενώ η θέση του χρήστη παρουσιάζεται και πάλι με την κόκκινη τελεία περίπου στο κέντρο της εικόνας. Στους δύο χάρτες, έχουν σημειωθεί με το ID και την τιμή τους οι βαθμολογήσεις που είναι κοινές για τις δύο θέσεις. Πιο συγκεκριμένα, οι βαθμολογίες αυτές είναι οι 9384, 66503, 96444, 87568, 35970 και η 26969. Η γεωγραφική θέση στην οποία πραγματοποιήθηκαν βρίσκεται μέσα στα όρια της περιοχής αλληλοκάλυψης, που παρουσιάστηκε στην εικόνα 30. Η παραπάνω παρατήρηση, εξηγεί το λόγο για τον οποίο ο χρήστης λαμβάνει λίστα συστάσεων με κοινά προϊόντα σε δύο κοντινές θέσεις.

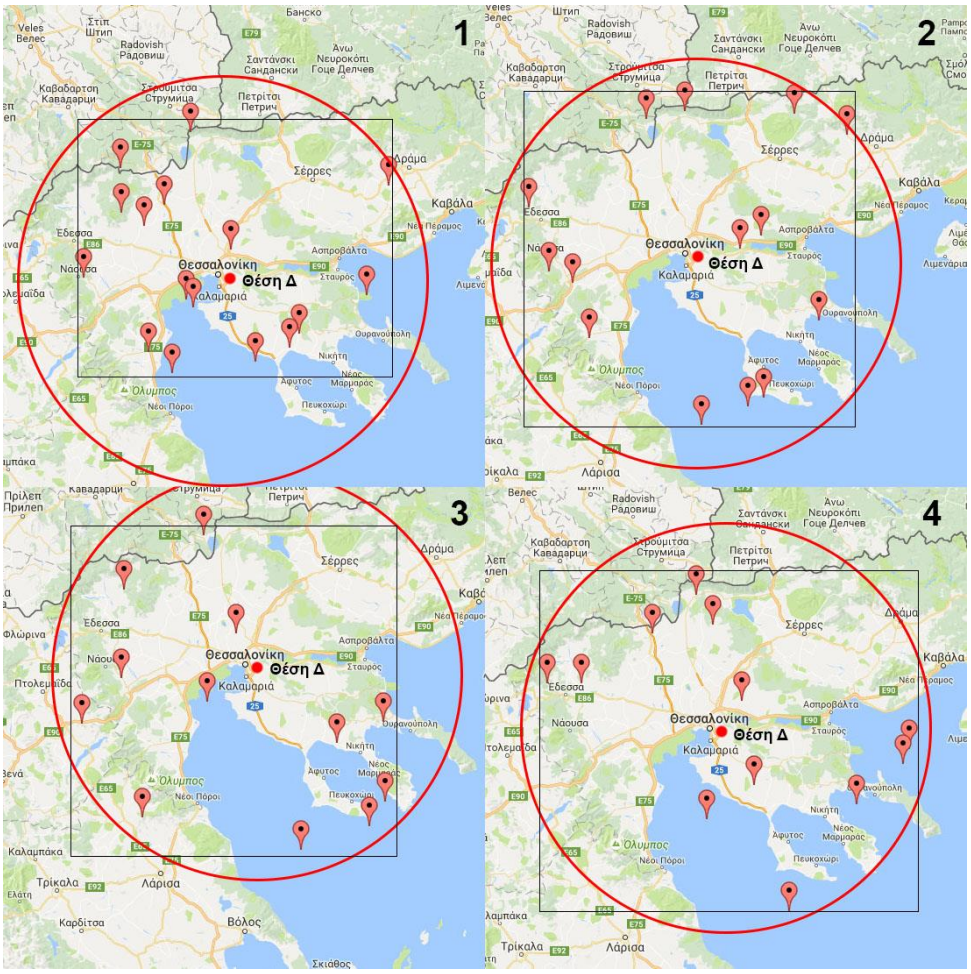
Όπως, όμως, προκύπτει από τις συστάσεις του συστήματος (Εικόνα 29), το προϊόν με ID 50 επανεμφανίζεται στην έξοδο του για τη θέση Γ. Με σκοπό να αποδειχθεί ο τοπικός χαρακτήρας των συστάσεων παρατίθενται στην εικόνα 33 ο χάρτης των βαθμολογιών της συστάδας υψηλού ενδιαφέροντος για τις θέσεις Α και Γ του εν λόγω προϊόντος.



Εικόνα 33: Χάρτης βαθμολογιών που συμβάλλουν στην πρόβλεψη του προϊόντος (ID 50) για τις θέσεις Α και Γ

Όπως προκύπτει από το χάρτη της εικόνας (Εικόνα 33) δεν υπάρχουν κοινές βαθμολογίες στις 2 περιοχές. Αυτό σημαίνει πως η επανεμφάνιση του προϊόντος στις συστάσεις οφείλεται αποκλειστικά και μόνο στις πολύ καλές βαθμολογήσεις που έχει αποσπάσει κοντά στις 2 περιοχές που μελετήθηκαν. Κοινά αποτελέσματα λαμβάνουμε και για την θέση Δ, της οποίας, ο χάρτης των βαθμολογιών των προβλέψεων της λίστας συστάσεων παρουσιάζεται στην εικόνα 34.

Στο πρώτο κομμάτι της εικόνας (πάνω αριστερά) εμφανίζονται οι βαθμολογήσεις του προϊόντος με ID 50. Αυτές εκτείνονται σε μία περιοχή εύρους 167x127 χιλιομέτρων και βρίσκονται όλες εντός της ακτίνας που δημιουργεί το κατώφλι της συνάρτησης locSim. Ομοίως, στο δεύτερο κομμάτι (πάνω δεξιά) οι βαθμολογίες του προϊόντος με ID 96, στο τρίτο (κάτω αριστερά) του προϊόντος με ID 150 και στο τέταρτο (κάτω δεξιά) του προϊόντος με ID 174.



Εικόνα 34: Χάρτης βαθμολογιών των συστάσεων για τη θέση Δ

Τα στοιχεία της διασποράς των βαθμολογιών γύρω από το χρήστη συμπίπτουν απόλυτα με το αναμενόμενο αποτέλεσμα, που διαμορφώνει η συνάρτηση του βάρους γεωγραφικής θέσης. Συμπερασματικά, λόγω της συνεισφοράς μόνο του παραπάνω χαρακτηριστικού στον υπολογισμό του συνολικού βάρους, οι βαθμολογίες που διαμορφώνουν τις προβλέψεις (και μετέπειτα τις συστάσεις) περικλείονται σε μία σχετικά μικρή ακτίνα γύρω από το χρήστη. Τα προϊόντα που συστήνονται, λοιπόν, αντιπροσωπεύουν τις ανά τόπους προτιμήσεις των βαθμολογητών (τοπικός χαρακτήρας της σύστασης).

8.4.2 Δεύτερο σενάριο (μ=0.8, λ=0.1, κ=0.1)

Για την εκτέλεση του δεύτερου σεναρίου θεωρούμε τις ίδιες τιμές των μεταβλητών α, β και γ, των εξισώσεων 9,10 και 11, όπως αυτές ορίστηκαν από το προηγούμενο σενάριο της παραγράφου 8.4.1. Αυτό που αλλάζει είναι το μίγμα της συνεισφοράς των βαρών, όπως αυτό ορίζεται από την εξίσωση 12. Πιο συγκεκριμένα, οι μεταβλητές κ, λ και μ παίρνουν τις τιμές 0.1, 0.1 και 0.8 αντίστοιχα. Το παραπάνω σημαίνει πως μειώνεται η εισφορά της γεωγραφικής θέσεις στο 80% του συνολικού βάρους, ενώ ταυτόχρονα αυξάνεται το ποσοστό του χρόνου και της ηλικίας του βαθμολογητή. Είναι άξιο αναφοράς πως, ήδη, από την κατηγοριοποίηση του χρήστη το εύρος των ηλικιών έχει μειωθεί σε περίοδο μίας δεκαετίας πράγμα που οδηγεί με μεγάλη πιθανότητα σε μη μηδενικό βάρος ηλικίας (όπως προκύπτει από την τιμή της συνάρτησης ageSim μετά τον προσδιορισμό της μεταβλητής α). Για παράδειγμα, έστω ότι συγκρίνονται οι ηλικίες ενός ζευγαριού πελάτη-βαθμολογητή. Το κατώφλι το οποίο ορίζει η συνάρτηση ageSim με α=0,3 είναι περίπου τα 10 δέκα χρόνια διαφοράς. Εφόσον από την κατηγοριοποίηση η μέγιστη διαφορά μεταξύ των ηλικιών του ζευγαριού είναι τα 10 χρόνια, με μεγάλη πιθανότητα η τιμή της συνάρτησης θα είναι μη μηδενική. Αν και η τιμή του βάρους που θα προκύψει συνεισφέρει μόνο κατά 10% στο σύνολο, βαθμολογήσεις που

πραγματοποιήθηκαν σε μακρινές αποστάσεις ή σε παλιότερη ημερομηνία θα μπορούσαν να επωφεληθούν και να εισέλθουν στη συστάδα υψηλού ενδιαφέροντος. Το ίδιο θα μπορούσε να συμβεί σε βαθμολογήσεις που έγιναν πολύ πρόσφατα, αλλά, σε πολύ μακρινή απόσταση από τον πελάτη, λόγω της αυξημένης τιμής της συνάρτησης timeSim.

Η παραπάνω παρατήρηση αποτυπώνεται και στα αποτελέσματα του αλγορίθμου. Δηλαδή, στο συγκεκριμένο σενάριο παρατηρούνται στη συστάδα υψηλού ενδιαφέροντος βαθμολογίες οι οποίες ξεφεύγουν από τα στενά όρια που θέτει η συνάρτηση locSim. Λόγω, όμως, του μικρού ποσοστού με το οποίο συνεισφέρουν ηλικία και χρόνος (συνολικά 20%), η πλειοψηφία των βαθμολογήσεων εντοπίζεται, όπως και στο προηγούμενο σενάριο σε μικρή χιλιομετρική απόσταση από το χρήστη που δέχεται τις συστάσεις. Τα αποτελέσματα του αλγορίθμου παρουσιάζονται στην από κάτω εικόνα (βλ. Εικόνα 35).

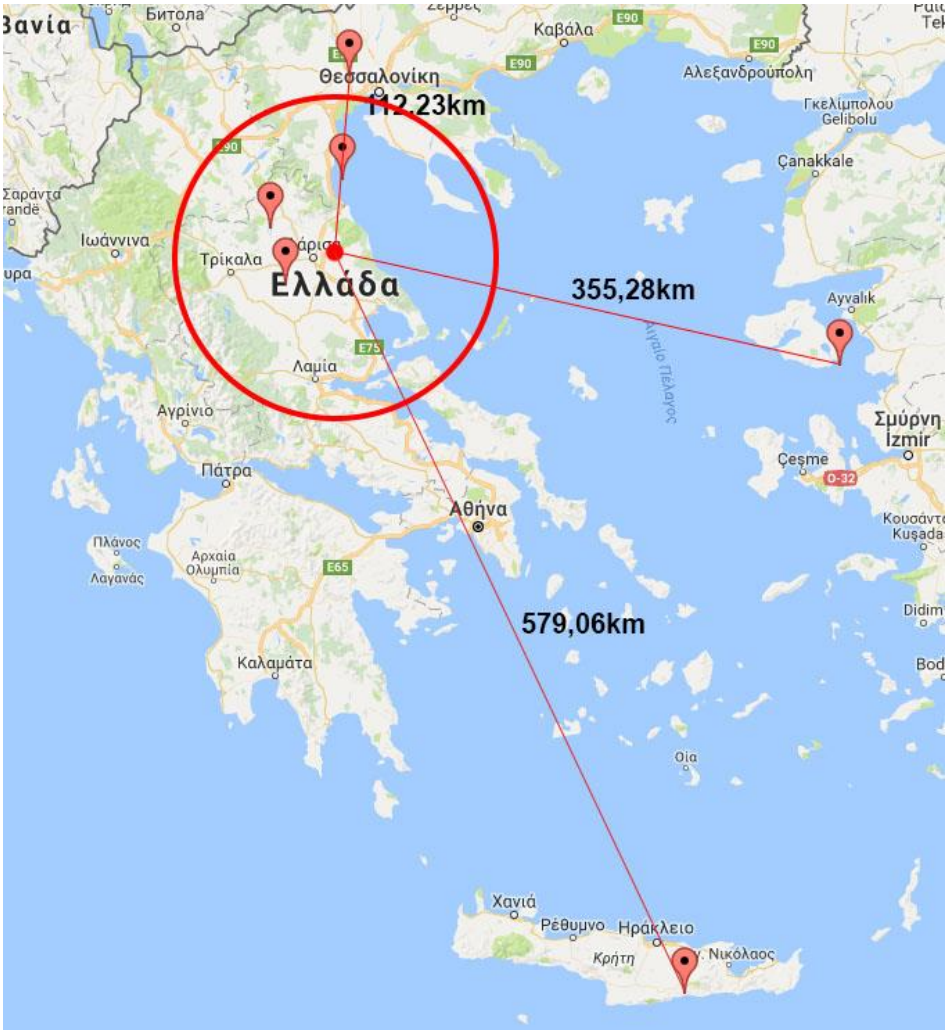
```
Θέση Α
Rating Prediction => Item:50, Count Ratings: 7, Prediction: 4.8571
Rating Prediction => Item:258, Count Ratings: 6, Prediction: 4.1667
Rating Prediction => Item:7, Count Ratings: 6, Prediction: 3.5
Rating Prediction => Item:117, Count Ratings: 7, Prediction: 3.4286

Θέση Β
Rating Prediction => Item:50, Count Ratings: 7, Prediction: 4.5714
Rating Prediction => Item:100, Count Ratings: 9, Prediction: 4.2222
Rating Prediction => Item:288, Count Ratings: 8, Prediction: 4.0
Rating Prediction => Item:268, Count Ratings: 6, Prediction: 4.0

Θέση Γ
Rating Prediction => Item:174, Count Ratings: 6, Prediction: 4.8333
Rating Prediction => Item:96, Count Ratings: 6, Prediction: 4.3333
Rating Prediction => Item:313, Count Ratings: 6, Prediction: 4.1667
Rating Prediction => Item:288, Count Ratings: 6, Prediction: 4.1667

Θέση Δ
Rating Prediction => Item:50, Count Ratings: 7, Prediction: 4.1429
Rating Prediction => Item:313, Count Ratings: 7, Prediction: 4.1429
Rating Prediction => Item:181, Count Ratings: 8, Prediction: 4.125
Rating Prediction => Item:315, Count Ratings: 6, Prediction: 4.0

Θέση Ε
Rating Prediction => Item:176, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:313, Count Ratings: 10, Prediction: 4.4
Rating Prediction => Item:1, Count Ratings: 6, Prediction: 3.8333
Rating Prediction => Item:100, Count Ratings: 8, Prediction: 3.125
```

Εικόνα 35: Αποτελέσματα αλγορίθμου για το δεύτερο σενάριο εκτέλεσης

Παρατηρώντας τις συστάσεις του αλγορίθμου διαπιστώνεται η ύπαρξη μίας συσχέτισης μεταξύ τους. Στη θέση Α και Β ο χρήστης λαμβάνει ως σύσταση το προϊόν με ID 50 το οποίο εμφανίζεται και στη θέση Δ. Αυτή η επανάληψη θα μπορούσε να οφείλεται σε τρείς λόγους. Ο πρώτος είναι η ύπαρξη πολλών υψηλών βαθμολογιών του προϊόντος γύρω από το χρήστη οι οποίες εισέρχονται στη συστάδα υψηλού ενδιαφέροντος λόγω της υψηλής τιμής του βάρους γεωγραφικής θέσεις. Ο δεύτερος είναι πως σε διπλανές θέσεις, όπως είναι η θέση Α και Β οι δύο χρήστες θα μπορούσαν να μοιράζονται ίδιες βαθμολογίες λόγω της αλληλοκάλυψης των δύο χώρων που ορίζει η συνάρτηση locSim. Ο τρίτος είναι η ύπαρξη στη συστάδα υψηλού ενδιαφέροντος βαθμολογήσεις που

εισήλθαν σε αυτή λόγω της τιμής είτε του βάρους χρόνου είτε του βάρους ηλικίας. Οι βαθμολογήσεις αυτές θα μπορούσαν να επηρεάζουν τις προβλέψεις, ανάλογα με τον ανταγωνισμό που δέχονται από άλλες, για οποιαδήποτε γεωγραφική θέση.

Ένα χαρακτηριστικό παράδειγμα σύστασης του δεύτερου σεναρίου είναι αυτό της σύστασης του προϊόντος με ID 288 στη θέση Γ. Οι βαθμολογήσεις που τη διαμορφώνουν (την πρόβλεψη βαθμολογίας) παρουσιάζονται στο χάρτη της εικόνας 36. Ο πελάτης που δέχεται τις συστάσεις αναπαρίσταται με την κόκκινη τελεία και γύρω από αυτόν σημειώνεται με ένας κόκκινος κύκλος, εκτός του οποίου, η τιμή του βάρους της συνάρτησης γεωγραφικής θέσης είναι μηδενική.

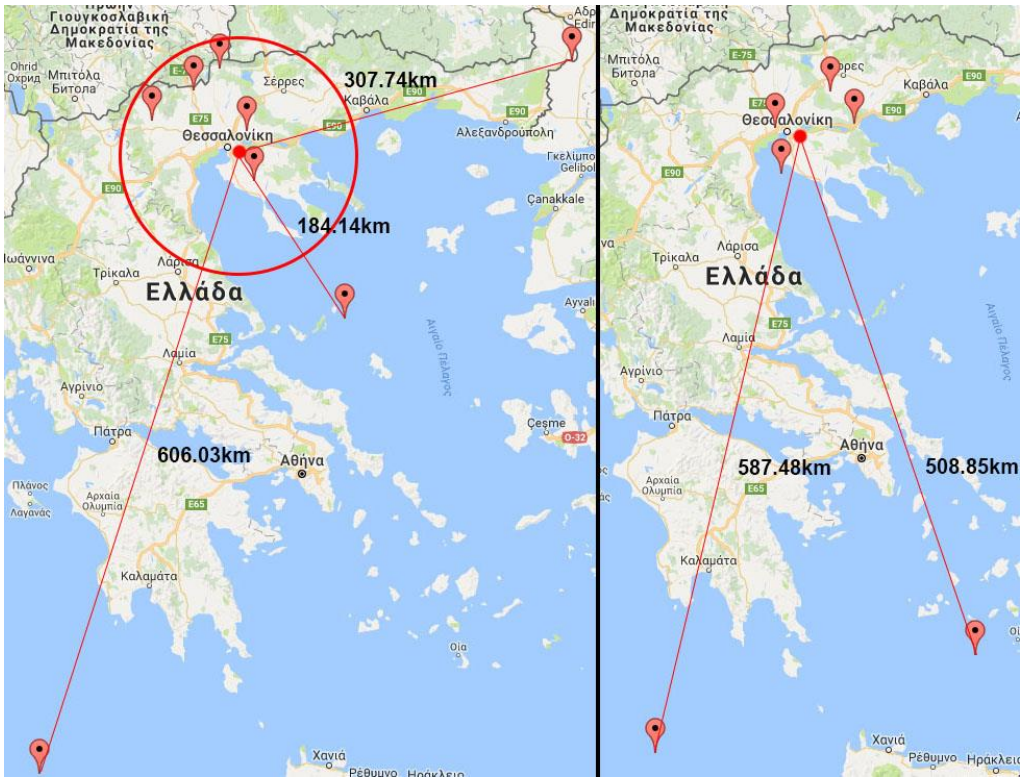


Εικόνα 36: Οι βαθμολογίες που διαμορφώνουν την πρόβλεψη βαθμολογίας του προϊόντος 288

Η πρόβλεψη της βαθμολογίας διαμορφώνεται από 6 συνολικά βαθμολογήσεις, 3 εκ των οποίων βρίσκονται εντός του κύκλου, ενώ οι υπόλοιπες βρίσκονται σε απόσταση άνω των 100 χιλιομέτρων. Όπως είναι λογικό, οι βαθμολογήσεις εντός του κύκλου εισήλθαν στη συστάδα υψηλού ενδιαφέροντος λόγω της υψηλής τιμής του βάρους γεωγραφικής θέσης. Για τις υπόλοιπες 3, θα μπορούσαμε να υποθέσουμε πως η είσοδος τους οφείλεται είτε στην ηλικία του βαθμολογητή (μικρή διαφορά ηλικίας με τον πελάτη), είτε στην πρόσφατη πραγματοποίηση τους, χωρίς να αποκλείονται φυσικά και οι δύο περιπτώσεις. Πράγματι, ερευνώντας τα αίτια της εισόδου τους στη συστάδα υψηλού ενδιαφέροντος, διαπιστώνεται πως ο βαθμολογητής που βρίσκεται 579 χιλιόμετρα μακριά (Νομός Ηρακλείου Κρήτης), έχει διαφορά ηλικίας, μόλις, 2 μήνες ενώ ο βαθμολογητής που βρίσκεται στα 355 χιλιόμετρα, 7 μήνες. Επιπρόσθετα, οι δύο βαθμολογίες πραγματοποιήθηκαν 1 μήνα πριν την εξαγωγή των συστάσεων. Έτσι, ενώ

το βάρος της γεωγραφικής του θέσης είναι σίγουρα μηδενικό, καταφέρνουν να επηρεάσουν την πρόβλεψη μέσω των βαρών χρόνου και ηλικίας.

Τα παραπάνω χαρακτηριστικά απαντώνται σε όλες τις βαθμολογίες που διαμορφώνουν τις προβλέψεις του δεύτερου σεναρίου. Στην εικόνα 37, παρουσιάζονται άλλα δύο παραδείγματα για τα προϊόντα με ID 181 και 335. Πιο συγκεκριμένα, η πλειοψηφία των βαθμολογιών που συμβάλλουν στις προβλέψεις εντοπίζεται πολύ κοντά στο χρήστη, ενώ αντίθετα ένα μικρότερο ποσοστό βρίσκεται διασπαρμένο σε μακρινά σημεία του χάρτη. Αυτές είναι οι βαθμολογίες οι οποίες εισήλθαν στη συστάδα υψηλού ενδιαφέροντος με μη γεωγραφικά κριτήρια.



Εικόνα 37: Χάρτης βαθμολογιών που διαμορφώνουν τις προβλέψεις των προϊόντων με ID 181 (αριστερά) και 335 (δεξιά) για τη θέση Δ.

Με βάση τα παραπάνω αποτελέσματα, η πιο ουσιαστική παρατήρηση για το μίγμα υπολογισμού του βάρους του δεύτερου σεναρίου είναι η απελευθέρωση του συστήματος πρόβλεψης από την ακτίνα των 100 χιλιομέτρων. Αυτή είχε παρατηρηθεί στο πρώτο σενάριο, λόγω της μη χρησιμοποίησης του χρόνου των βαθμολογήσεων και της ηλικίας του βαθμολογητή στον καθορισμό του συνολικού βάρους. Τα γεωγραφικά όρια των συστάσεων του σεναρίου δεν είναι, τόσο αυστηρά, με την έννοια ότι βαθμολογήσεις από οποιοδήποτε μέρος του χάρτη θα μπορούσαν να επηρεάσουν τις προβλέψεις. Αν και το ποσοστό των πιο μακρινών βαθμολογιών είναι μικρότερο σε σχέση με αυτό των κοντινότερων, θα μπορούσαν να διαμορφώσουν την έξοδο του συστήματος είτε αποκλείοντας ένα προϊόν είτε πριμοδοτώντας το.

8.4.3 Τρίτο σενάριο (μ=0.5, λ=0.3,κ=0.2)

Σύμφωνα με το τρίτο σενάριο, το μίγμα του υπολογισμού του βάρους των βαθμολογιών διαμορφώνεται ως εξής: το 50% οφείλεται στην τιμή της συνάρτησης locSim και κατ' επέκταση στη γεωγραφική θέση που έλαβε χώρα η βαθμολόγηση, το 30% στο χρόνο κατά τον οποίο έγινε η βαθμολόγηση και κατά 20% στην ηλικία του χρήστη που βαθμολόγησε (διαφορά ηλικίας με τον πελάτη που θα δεχτεί τις συστάσεις). Το κοινό με

τα αποτελέσματα του προηγούμενου σεναρίου είναι πως οι βαθμολογίες δεν είναι περιορισμένες σε κάποια γεωγραφική περιοχή. Για την ακρίβεια, όσο το ποσοστό με το οποίο συμβάλλουν η ηλικία και ο χρόνος μεγαλώνει, τόσο η διασπορά των βαθμολογιών στο χάρτη επεκτείνεται. Οι βαθμολογήσεις που διαμορφώνουν τις προβλέψεις και κατ' επέκταση οι συστάσεις, παύουν να έχουν τοπικό χαρακτήρα. Αυτό οφείλεται στην απουσίας από τη συστάδα υψηλού ενδιαφέροντος ικανοποιητικού αριθμού βαθμολογήσεων κοντινών στο χρήστη που θα λειτουργούσαν ως διαμορφωτές των προβλέψεων.

Τα αποτελέσματα του συστήματος για το παρόν σενάριο εκτέλεσης παρουσιάζονται στην εικόνα 38.

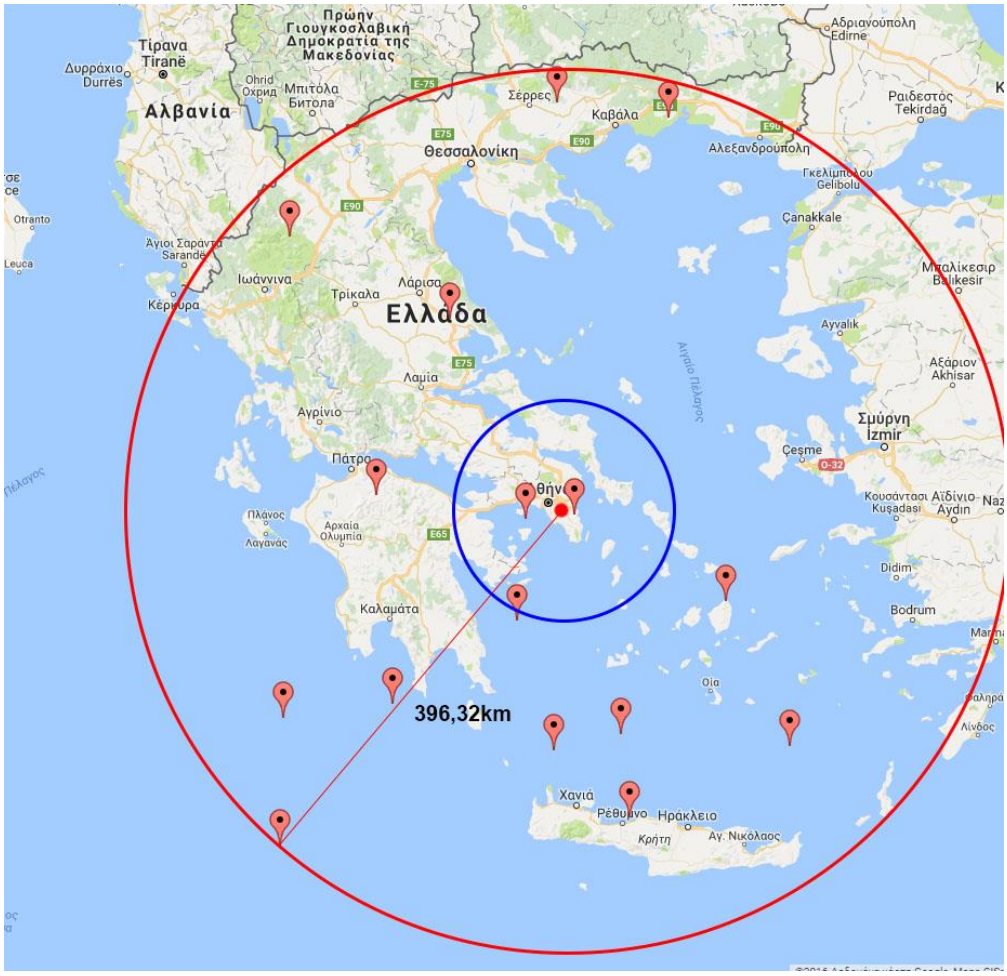
```
Θέση Α
Rating Prediction => Item:174, Count Ratings: 16, Prediction: 4.5625
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:192, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:211, Count Ratings: 6, Prediction: 4.5

Θέση Β
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:518, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:211, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:318, Count Ratings: 9, Prediction: 4.4444

Θέση Γ
Rating Prediction => Item:174, Count Ratings: 15, Prediction: 4.6
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:22, Count Ratings: 10, Prediction: 4.4
Rating Prediction => Item:12, Count Ratings: 8, Prediction: 4.375

Θέση Δ
Rating Prediction => Item:135, Count Ratings: 6, Prediction: 4.6667
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:242, Count Ratings: 6, Prediction: 4.5

Θέση Ε
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:22, Count Ratings: 11, Prediction: 4.3636
Rating Prediction => Item:12, Count Ratings: 11, Prediction: 4.3636
```

Εικόνα 38: Αποτελέσματα αλγορίθμου για το τρίτο σενάριο εκτέλεσης

Παρατηρείται πως η έξοδος του συστήματος παρουσιάζει αρκετές ομοιότητες από περιοχή σε περιοχή πάνω στο χάρτη. Χαρακτηριστικό είναι το παράδειγμα του προϊόντος με ID 174 που εμφανίζεται ως σύσταση και στις 5 τοποθεσίες του χάρτη. Το σημαντικότερο σημείο της πιο πάνω παρατήρησης είναι πως οι βαθμολογίες που αναδεικνύουν το προϊόν σε σύσταση είναι σε όλες τις θέσεις σχεδόν αμετάβλητες. Πιο συγκεκριμένα, για τις θέσεις Β, Δ και Ε, ενεπλάκησαν στην πρόβλεψη 14 βαθμολογήσεις οι οποίες απέφεραν τιμή πρόβλεψης 4,5714 (ίδιες βαθμολογίες). Στις θέσεις Α και Β, 16 και 15 βαθμολογήσεις, οι οποίες άλλαξαν το αποτέλεσμα της τιμής πρόβλεψης ελάχιστα (4,5625 και 4,6 αντίστοιχα). Δηλαδή, για τις παραπάνω θέσεις προστέθηκαν συνολικά 3 επιπλέον βαθμολογίες, η ύπαρξη των οποίων στη συστάδα

υψηλού ενδιαφέροντος αποδίδεται στην κοντινή τους θέση σε σχέση με το χρήστη που δέχεται τις συστάσεις. Παρόμοια συμπεράσματα θα μπορούσαν να εξαχθούν για το προϊόν με ID 318 που εμφανίζεται, ομοίως με το 174, σε όλες τις θέσεις στις οποίες ο πελάτης δέχεται συστάσεις.

Η παραπάνω παρατηρήσεις τεκμηριώνονται από το χάρτη των βαθμολογήσεων, της εικόνας 39, για το προϊόν με ID 174 στη θέση Α.



Εικόνα 39: Βαθμολογίες που διαμορφώνουν την πρόβλεψη του προϊόντος 174 για τη θέση Α

Στην εικόνα 39 φαίνονται οι 14 βαθμολογήσεις που παραμένουν σταθερές καθ' όλη τη διάρκεια του ταξιδιού του χρήστη (για όλες τις θέσεις). Οι βαθμολογήσεις αυτές, βρίσκονται έξω από το μπλε κύκλο που διαμορφώνει η συνάρτηση locSim και μέσα στον κόκκινο κύκλο με ακτίνα 396,32 χιλιομέτρων. Επιπρόσθετα, μέσα στο μπλε κύκλο και σε πολύ κοντινή απόσταση από τον πελάτη διακρίνονται οι δύο επιπλέον βαθμολογίες που διακρίνουμε στα αποτελέσματα του συστήματος της εικόνας 39. Οι βαθμολογίες αυτές δεν έχουν γίνει πρόσφατα (μικρό βάρος χρόνου) και ο βαθμολογητής δεν είναι ηλικιακά τόσο κοντά στο πελάτη (μικρό βάρος ηλικίας). Αυτός είναι και ο λόγος για τον οποίο εμφανίζονται μόνο στη θέση Α.

Αξίζει ακόμα να σημειωθεί, πως όπως προκύπτει από τα από τα αποτελέσματα του πρώτου σεναρίου μέσα στον μπλε κύκλο υπάρχουν συνολικά 14 βαθμολογίες που αφορούν το προϊόν με ID 174. Από αυτές τις βαθμολογίες μόνο οι δύο πιο κοντινές στο χρήστη κατάφεραν να μπουν στη συστάδα υψηλού ενδιαφέροντος ενώ οι υπόλοιπες 12 απέτυχαν να ξεπεράσουν το κατώφλι του βάρους ώστε να εμπλακούν στη διαμόρφωση της πρόβλεψης. Αυτό συνέβη, καθώς εντοπίστηκαν 14 βαθμολογίες (εκτός του μπλε κύκλου) οι οποίες έλαβαν μεγαλύτερο βάρος και κατάφεραν να εισέλθουν στη συστάδα υψηλού ενδιαφέροντος κατά τη διάρκεια της συσταδοποίησης.

Από την παραπάνω ανάλυση των αποτελεσμάτων του σεναρίου 3, προκύπτει πως όσο μειώνεται το ποσοστό της συμβολής της γεωγραφικής θέσης στο συνολικό βάρος, τόσο η συσχέτιση μεταξύ των προβλέψεων σε διαφορετικές θέσεις αυξάνεται. Οι συστάσεις που προσφέρονται στο χρήστη χάνουν τον τοπικό χαρακτήρα τους και η σημασία της χωρικής πληροφορίας βαίνει μειούμενη. Η παραπάνω παρατήρηση αποδεικνύεται από τα αποτελέσματα των δύο τελευταίων σεναρίων. Με βάση το συγκεκριμένο μίγμα χαρακτηριστικών οι συστάσεις θα μπορούσαν να χαρακτηριστούν, γενικότερα, καθολικές, καθώς ο ίδιος χρήστης παίρνει σχεδόν τις ίδιες συστάσεις για διαφορετικές θέσεις.

8.5 Αριθμητική αποτίμηση

Ένα από τα πιο ουσιαστικά κομμάτια της εργασίας είναι η πειραματική αποτίμηση της προτεινόμενης μεθόδου. Μετά την εξαγωγή των συστάσεων, οι παραπάνω προβλέψεις βαθμολογίας συγκρίνονται με πραγματικές βαθμολογήσεις χρηστών, με τη βοήθεια μετρικών απόκλισης και εξάγονται συμπεράσματα. Η πειραματική αποτίμηση συμβάλλει στην αποτύπωση μίας εικόνας της αποτελεσματικότητας του αλγορίθμου με βάση τα διαφορετικά σενάρια που μελετήθηκαν στο παρόν κεφάλαιο (Κεφάλαιο 8). Πιο συγκεκριμένα σκοπός της είναι η εξακρίβωση των επιδόσεων του προτεινόμενου συστήματος αλλά και η απόδειξη της ορθής του λειτουργίας. Σε πρώτη φάση παρουσιάζονται οι μετρικές απόδοσης, σύμφωνα με τις οποίες καθορίζεται η αποδοτικότητα του συστήματος και αξιολογείται η έξοδος του. Τέλος, γίνεται παράθεση των αποτελεσμάτων για τα διάφορα σενάρια, βάσει των οποίων δημιουργήθηκαν οι τεχνικές αξιολόγησης.

8.5.1 Μετρικές Απόδοσης

Οι μετρικές απόδοσης [5] ουσιαστικά μετρούν το πόσο διαφορετικές είναι οι προβλέψεις του συστήματος σε σχέση με τα πραγματικά δεδομένα των βαθμολογήσεων που έχουν πραγματοποιήσει οι πελάτες. Η πιο διαδεδομένη μετρική απόδοσης ενός συστήματος συστάσεων είναι το απόλυτο σφάλμα ή αλλιώς Mean Absolute Error (MAE). Το MAE υπολογίζει το μέσο όρο της απόλυτης τιμής της διαφοράς ανάμεσα στις προβλεπόμενες και τις πραγματικές αξιολογήσεις. Η εξίσωση του απόλυτου σφάλματος είναι η:

$$MAE = \frac{1}{N}\sum_{ij} |prediction_{ij} - real_{ij}|$$

Εξίσωση 16: Η εξίσωση του απόλυτου σφάλματος

Για την εξίσωση 16, όπου N είναι το σύνολο των βαθμολογήσεων για τις οποίες έχει γίνει πρόβλεψη, το $prediction_{ij}$ είναι η τιμή πρόβλεψης για ένα χρήστη i σε ένα αντικείμενο j ενώ το $real_{ij}$ είναι η πραγματική αξιολόγηση που έχει δώσει ο χρήστης i για το προϊόν j.

Μία δεύτερη τεχνική απόδοσης που χρησιμοποιείται στα πλαίσια της παρούσας εργασίας, αλλά και γενικότερα για την αξιολόγηση της αποδοτικότητας των συστημάτων συστάσεων είναι η ρίζα του μέσου τετραγωνικού σφάλματος (Root Squared Mean Error - RMSE). Η εξίσωση του RMSE είναι η:

$$RMSE = \sqrt{\frac{1}{N}\sum_{ij} (prediction_{ij} - real_{ij})^2}$$

Εξίσωση 17: Η εξίσωση του μέσου τετραγωνικού σφάλματος

Για την εξίσωση 17, όπου N είναι το σύνολο των αντικειμένων για τα οποία γίνεται πρόβλεψη, $prediction_{ij}$ είναι η τιμή πρόβλεψης για ένα χρήστη i σε ένα αντικείμενο j και

$real_{ij}$ είναι η πραγματική βαθμολόγηση. Το RMSE υπολογίζει την τετραγωνική ρίζα της μέσης τιμής της διαφοράς υψωμένη στο τετράγωνο. Αξίζει να σημειωθεί πως όσο πιο μικρές είναι οι τιμές των δύο μετρικών τόσο καλύτερες είναι οι προβλέψεις άρα τόσο μεγαλύτερη η ακρίβεια του αλγορίθμου που προτείνεται. Τόσο η μία τεχνική όσο και η άλλη είναι καθολικά αποδεκτές και συναντώνται πολύ συχνά στη βιβλιογραφία που αναφέρεται σε μετρικές απόδοσης συστημάτων συστάσεων.

8.6 Σενάρια και αξιολόγηση αποτελεσμάτων

Η αξιολόγηση των αποτελεσμάτων πραγματοποιήθηκε με βάση τις μετρικές που παρουσιάστηκαν στην παράγραφο 8.5. Οι μετρικές εφαρμόστηκαν και στα 3 σενάρια για κάθε μία από τις θέσεις στις οποίες ο πελάτης έλαβε τις συστάσεις του συστήματος. Στη θέση του prediction χρησιμοποιήθηκαν οι προβλέψεις του προτεινόμενου συστήματος, ενώ για πραγματικά δεδομένα (real) χρησιμοποιήθηκαν βαθμολογίες που πραγματοποίησαν πελάτες του καταστήματος σε μία κοντινή απόσταση από την εκάστοτε θέση του πελάτη.

8.6.1 Απόλυτο Σφάλμα

Αρχικά, ο προτεινόμενος αλγόριθμος μετρήθηκε με βάση το απόλυτο σφάλμα και τα αποτελέσματα για τα 3 σενάρια εκτέλεσης του, παρουσιάζονται στον πίνακα 2. Υπενθυμίζεται πως όσον αφορά την απόδοση του, όσο μικρότερος είναι ο αριθμός του απόλυτου σφάλματος, τόσο ακριβέστερα είναι τα αποτελέσματα που αυτός αποδίδει.

Πίνακας 2: Το απόλυτο σφάλμα σε κάθε θέση των 3 σεναρίων

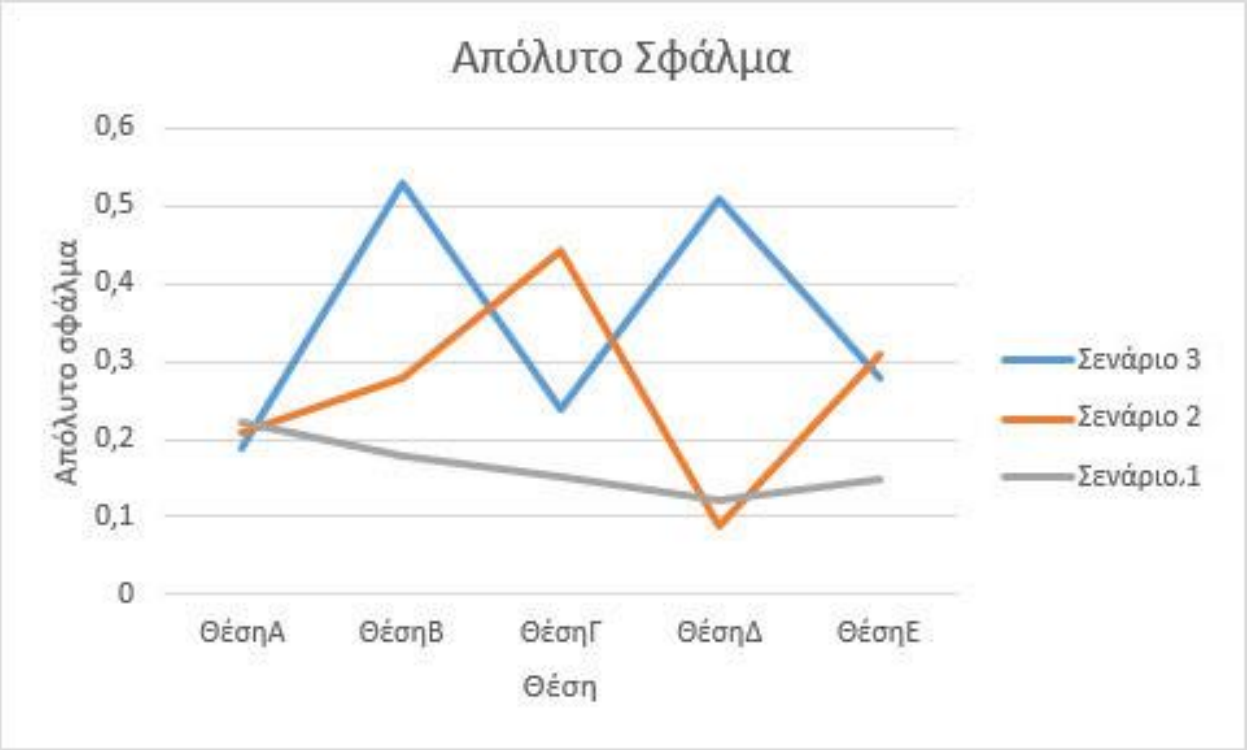
	Θέση Α	**Θέση Β**	**Θέση Γ**	**Θέση Δ**	**Θέση Ε**
Σενάριο 1	0.2226	0.1797	0.15	0.1231	0.14675
Σενάριο 2	0.2091	0.28	0.4421	0.087	0.3079
Σενάριο 3	0.188	0.53	0.24	0.51	0.28

Κάθε σειρά του πίνακα 2 αντιπροσωπεύει τα αποτελέσματα ενός εκ των 3 σεναρίων εκτέλεσης, ενώ κάθε στήλη τις τιμές του απολύτου σφάλματος για μία συγκεκριμένη θέση. Παρατηρώντας τις τιμές της μετρικής διαπιστώνεται πως είναι ιδιαίτερα χαμηλές στο πρώτο σενάριο και παρουσιάζουν μία αυξητική τάση στα επόμενα 2, χωρίς όμως αυτή να επηρεάζει τη λειτουργία του προτεινόμενου συστήματος. Ο μικρός αριθμός σφάλματος που προκύπτει, συνολικά, σε όλες τις θέσεις των 3 σεναρίων προδίδει τη μεγάλη ακρίβεια των συστάσεων του αλγορίθμου.

Επιστρέφοντας στα αποτελέσματα ανά εκτέλεση, το απόλυτο σφάλμα για τα πρώτο σενάριο είναι εξαιρετικά χαμηλό, δείγμα του ότι ο αλγόριθμος αποδίδει ακριβείς συστάσεις. Οι τιμές δεν έχουν μεγάλες αποκλίσεις μεταξύ τους. Η μεγαλύτερη τιμή εντοπίζεται στην θέση Α (0,2226) ενώ η μικρότερη στη θέση Δ (0,1231). Για το δεύτερο σενάριο παρατηρείται μία γενικότερη αύξηση του σφάλματος. Η μικρότερη τιμή μετρήθηκε στη θέση Δ (0,087) ενώ η μεγαλύτερη στη θέση Γ (0,4421). Η αύξηση αυτή οφείλεται στο γεγονός ότι κατά το δεύτερο σενάριο οι βαθμολογήσεις ξεφεύγουν από τα όρια που δημιουργεί η συνάρτηση βάρους γεωγραφικής θέσης. Έτσι, λαμβάνοντας υπόψη στη μέτρηση του σφάλματος μόνο τις πραγματικές βαθμολογίες που είναι κοντά στο χρήστη η απόκλιση μεγαλώνει. Για το τρίτο σενάριο, παρατηρείται μία συνολική αύξηση του σφάλματος σε σχέση με το πρώτο αλλά και με το δεύτερο σενάριο. Ο λόγος για τον οποίο συμβαίνει είναι κοινός με αυτόν του σεναρίου 2. Όπως παρουσιάστηκε στην παράγραφο 8.5, στο τρίτο σενάριο το ποσοστό των βαθμολογιών που

συμβάλλουν στη διαμόρφωση της σύστασης και είναι κοντά στο χρήστη είναι πολύ μικρότερο από των μακρινών βαθμολογιών. Έτσι η απόκλιση σε σχέση με τις πραγματικές βαθμολογίες, που επιλέχθηκαν να είναι κοντά στο χρήστη, μεγαλώνει.

Οι γραφικές παραστάσεις του απολύτου σφάλματος για κάθε ένα από τα 3 σενάρια εκτέλεσης του αλγορίθμου παρουσιάζονται στην εικόνα 40.



Εικόνα 40: Γραφική παράσταση του απόλυτου σφάλματος και για τα 3 σενάρια

Παρατηρώντας πιο προσεκτικά τη γραφική παράσταση του απολύτου σφάλματος του δεύτερου σεναρίου (εμφανίζεται στην εικόνα 40 με πορτοκαλί χρώμα), διαπιστώνεται η ύπαρξη μεγάλων διαφορών από θέση σε θέση. Για παράδειγμα, στη θέση Γ παίρνει τιμή κοντά στο 0,45 και στην επόμενη θέση πέφτει κατακόρυφα στο 0,08. Αυτό πιθανότατα συμβαίνει για 2 λόγους. Αρχικά στη θέση Γ, η πλειοψηφία των βαθμολογιών που διαμορφώνουν τις συστάσεις εντοπίζονται μακριά από το χρήστη και έχουν διαφορετικές τιμές από αυτές που είναι πιο κοντά. Έτσι η απόκλιση της εξόδου του προτεινόμενου συστήματος από τις πραγματικές βαθμολογίες αυξάνεται. Στην επόμενη θέση (θέση Δ), είτε οι βαθμολογίες που συμβάλλουν στην πρόβλεψη τυχαίνει να βρίσκονται (χιλιομετρικά) κοντά στο χρήστη, είτε βρίσκονται μακριά από αυτόν όμως οι τιμές τους δεν έχουν μεγάλες διαφορές από τις κοντινές. Υπάρχει δηλαδή μία ομοιομορφία στις τιμές των βαθμολογήσεων του προϊόντος. Το ίδιο φαινόμενο παρατηρείται και για το τρίτο σενάριο και πιο συγκεκριμένα για τις θέσεις Β, Δ και Γ.

Τέλος, συμπεραίνουμε πως η τιμή του απολύτου σφάλματος για τα παραπάνω σενάρια εκτέλεσης, εξαρτάται από τη γεωγραφική θέση των βαθμολογιών (απόσταση από το χρήστη) σε συνδυασμό με τις διαφοροποιήσεις των πιο μακρινών βαθμολογιών σε σχέση με αυτές που βρίσκονται πιο κοντά στον πελάτη. Δηλαδή, όσο μακρύτερα βρίσκονται οι βαθμολογίες της πρόβλεψης και όσο πιο πολύ διαφοροποιούνται από τις πιο κοντινές τόσο μεγαλύτερη είναι η τιμή του απόλυτου σφάλματος.

8.6.2 Μέσο Τετραγωνικό σφάλμα

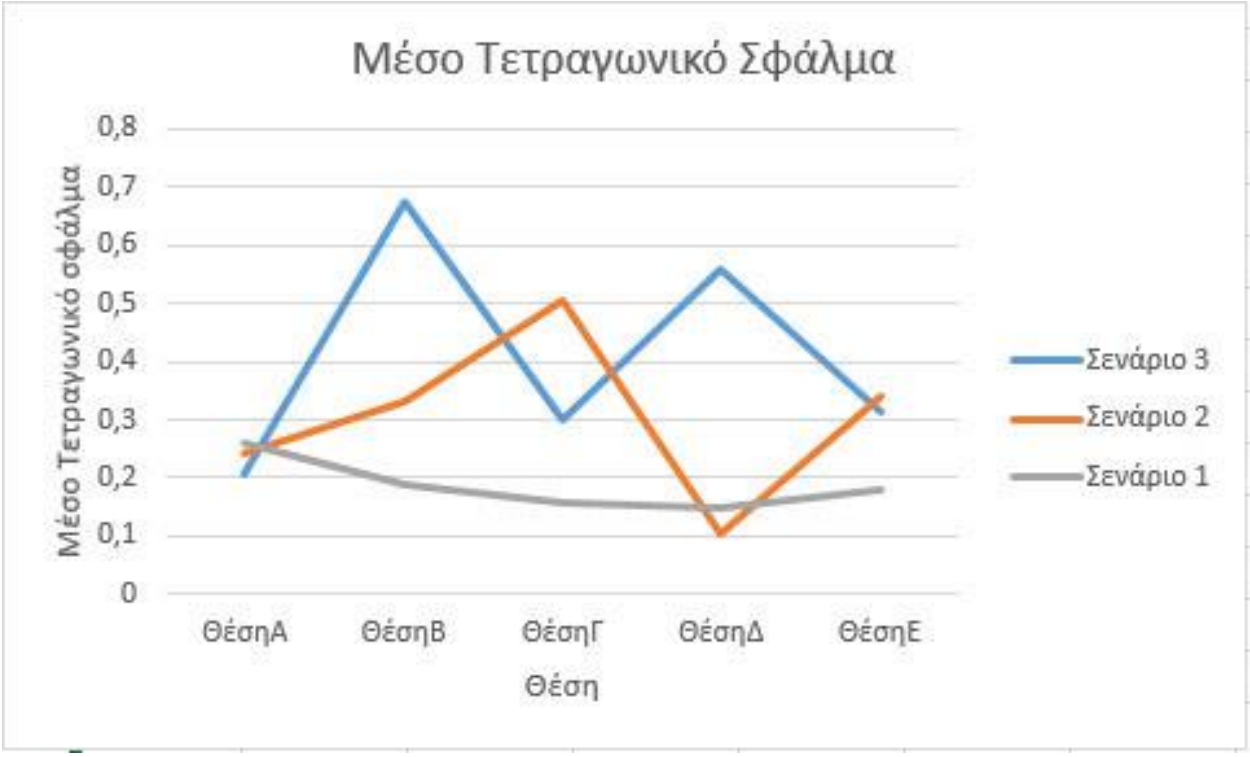
Τα αποτελέσματα του υπολογισμού του μέσου τετραγωνικού σφάλματος για κάθε σενάριο και για κάθε θέση στην οποία ο χρήστης δέχτηκε συστάσεις παρουσιάζονται στον πίνακα 3.

Πίνακας 3: Το μέσο τετραγωνικό σφάλμα σε κάθε θέση των τριών σεναρίων

	Θέση Α	**Θέση Β**	**Θέση Γ**	**Θέση Δ**	**Θέση Ε**
Σενάριο 1	0.26	0.1887	0,1576	0.1465	0.1787
Σενάριο 2	0.2401	0.33	0.5066	0.1055	0.34
Σενάριο 3	0.205	0.6753	0.3	0.56	0.312

Ξεκινώντας από το πρώτο σενάριο η τιμή του σφάλματος εμφανίζεται ιδιαίτερα χαμηλή από 0,1465 (θέση Δ) και φτάνει μέχρι και 0,26 (Θέση Α). Στο δεύτερο σενάριο, το μέσο τετραγωνικό σφάλμα είναι ελαφρώς αυξημένο (συνολικά). Η μικρότερη τιμή που παίρνει είναι 0,1055 για τη θέση Δ ενώ η μεγαλύτερη 0,5066 (θέση Γ). Το σενάριο 3 εμφανίζει αύξηση του σφάλματος, παρουσιάζοντας τιμές από 0,205 (θέση Α) έως και 0,6753 για τη θέση Β.

Οι γραφικές παραστάσεις του μέσου τετραγωνικού σφάλματος για τα 3 σενάρια εμφανίζονται στην εικόνα 41.



Εικόνα 41: Γραφική παράσταση του μέσου τετραγωνικού σφάλματος και για τα 3 σενάρια

Παρατηρείται πως κατά το πρώτο σενάριο το σφάλμα παραμένει γενικά σταθερό, χωρίς ιδιαίτερες διακυμάνσεις. Αντίθετα, το σενάριο 2 (πορτοκαλί γραφική παράσταση, βλ Εικόνα 41) έχει μεγάλες διαφοροποιήσεις ακόμα και σε κοντινές θέσεις (θέση Γ και Δ). Το φαινόμενο αυτό εμφανίζεται λόγω του χαρακτηριστικού του δεύτερου σεναρίου, που φέρει ένα ποσοστό των βαθμολογιών που εμπλέκονται στον υπολογισμό της πρόβλεψης (βαθμολογίας της σύστασης) να βρίσκεται σε μακρινή απόσταση από το

χρήστη. Δεδομένου ότι έχουν ληφθεί ως πραγματικές βαθμολογίες (real) οι βαθμολογίες που βρίσκονται σε μικρή ακτίνα από το χρήστη, όταν οι μακρύτερες βαθμολογίες διαφέρουν σημαντικά από τις πραγματικές καταλήγουμε σε μεγάλο σφάλμα (θέση Γ). Αντίθετα, όταν οι βαθμολογίες δεν έχουν μεγάλη διαφορά το σφάλμα εμφανίζεται μειωμένο (θέση Δ).

Τέλος, στο σενάριο 3 παρατηρείται κατά μέσο όρο το μεγαλύτερο σφάλμα. Εκτός αυτού, είναι το σενάριο για το οποίο μετρήθηκε η μεγαλύτερη τιμή του μέσου τετραγωνικού σφάλματος (θέση Β) η οποία πλησίασε το 0,7. Ο λόγος που το σφάλμα εμφανίζεται αυξημένο είναι πως, ομοίως με το σενάριο 2, ένα ποσοστό των βαθμολογιών που διαμορφώνουν την πρόβλεψη βρίσκεται μακριά από τον πελάτη που δέχεται τις συστάσεις. Η διαφορά των δύο σεναρίων, στην οποία οφείλεται η αύξηση του σφάλματος, είναι πως κατά το σενάριο 3 το ποσοστό αυτό είναι σημαντικά υψηλότερο λόγω του ποσοστού επιρροής της ηλικίας και του χρόνου σε βάρος της γεωγραφικής θέσεις.

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

Συμπεράσματα

Στην παρούσα διπλωματική εργασία δημιουργήθηκε ένα καινοτόμο σύστημα συστάσεων το οποίο προσφέρει ως έξοδο προτάσεις που βασίζονται σε χρονική και χωρική πληροφορία. Με πιο απλά λόγια οι συστάσεις που προκύπτουν εξαρτώνται από τη γεωγραφική θέση του πελάτη αλλά και τη χρονική στιγμή την οποία γίνονται. Για την καλύτερη παρουσίαση του, το σύστημα χωρίστηκε σε 4 υποσυστήματα. Το πρώτο υποσύστημα, είναι αυτό της κατηγοριοποίησης των χρηστών, κατά τη διάρκεια της οποίας ένας χρήστης εντάσσεται σε μία κατηγορία ανάλογα με το φύλο και την ηλικία του. Το δεύτερο υποσύστημα είναι αυτό του υπολογισμού των βαρών. Ανάλογα με την κατηγορία στην οποία ανήκει ο πελάτης, το σύστημα συγκρίνει τα χαρακτηριστικά όλων των βαθμολογιών των πελατών που βρίσκονται στην ίδια κατηγορία με χαρακτηριστικά του ίδιου του πελάτη, όπως είναι η ηλικία και η γεωγραφική θέση. Η έξοδος του συστήματος είναι ένας πίνακας βαρών για κάθε βαθμολογία, της οποίας ο βαθμολογητής βρίσκεται στην ίδια κατηγορία με τον πελάτη που θα δεχτεί τις συστάσεις. Το βάρος αυτό αντικατοπτρίζει κατά πόσο η βαθμολογία αυτή και κατ' επέκταση το προϊόν στο οποίο αναφέρεται ταιριάζει στις προτιμήσεις του χρήστη. Στο επόμενο υποσύστημα, αυτό της συσταδοποίησης, γίνεται ένα ξεσκαρτάρισμα των βαθμολογιών και παραμένουν μόνο οι βαθμολογίες με τα υψηλότερα βάρη. Το τελευταίο υποσύστημα επεξεργάζεται τα δεδομένα του προηγούμενου και εξάγει κάποιες προβλέψεις από τις οποίες προκύπτουν οι τελικές συστάσεις προς τον πελάτη.

Προχωρώντας σε περισσότερες λεπτομέρειες υλοποίησης, για το υποσύστημα της κατηγοριοποίησης και αυτό της συσταδοποίησης χρησιμοποιήθηκαν δύο έτοιμοι αλγόριθμοι με τη βοήθεια του εργαλείου εξόρυξης δεδομένων WEKA. Για το πρώτο υποσύστημα χρησιμοποιήθηκαν οι αλγόριθμοι Naïve Bayes και C4.5 ενώ για το δεύτερο ο αλγόριθμος expectation maximization (EM). Αρχικά, τα αποτελέσματα που προέκυψαν από την ανάλυση των δύο αλγορίθμων κατηγοριοποίησης με τη βοήθεια της εφαρμογής του WEKA ήταν εξαιρετικά. Για τον Naïve Bayes το ποσοστό επιτυχίας άγγιζε το 99,9% και ο C4.5 κατάφερε να επιτύχει το απόλυτο στις προβλέψεις του. Τα αποτελέσματα του αλγορίθμου συσταδοποίησης μπορούν να γίνουν αντιληπτά από την αποτίμηση του απόλυτου και μέσου τετραγωνικού σφάλματος στα οποία σημειώθηκαν εξαιρετικά μικρές τιμές.

Όσον αφορά τα αποτελέσματα του αλγορίθμου διαπιστώθηκε διαφορετική συμπεριφορά του συστήματος ανάλογα με το μίγμα χαρακτηριστικών που χρησιμοποιήθηκε για τον υπολογισμού των βαρών κάθε βαθμολογίας. Πιο συγκεκριμένα, η φύση των συστάσεων αλλάζει σε σχέση με το ποσοστό με το οποίο συμβάλει η γεωγραφική θέση στον υπολογισμό του βάρους. Για μεγάλα ποσοστά οι συστάσεις έχουν μία τοπική μορφή και περιλαμβάνουν βαθμολογίες πολύ κοντά στο χρήστη. Όσο τα ποσοστά μειώνονται ο χαρακτήρας των συστάσεων μετατρέπεται ολοένα και περισσότερο σε καθολικό, δηλαδή, οι ίδιες βαθμολογίες παίζουν σημαντικό ρόλο στην αποτίμηση της πρόβλεψης βαθμολογίας ανεξάρτητα από τη γεωγραφική θέση. Αυτό συμβαίνει, διότι, στο παιχνίδι του υπολογισμού των βαρών μπαίνουν η ηλικία του χρήστη και η χρονική στιγμή που έγινε η βαθμολόγηση , δεδομένα που δε διαφοροποιούνται ανάλογα με τη θέση.

Τέλος, ο διαφορετικός αυτός χαρακτήρας των συστάσεων αποτυπώνεται και στην αποτίμηση του σφάλματος τόσο με τη μετρική του απόλυτου όσο και με αυτή του μέσου τετραγωνικού σφάλματος με βάση, πάντα, την τεχνική που χρησιμοποιείται. Λόγω της διεύρυνσης των χωρικών ορίων των βαθμολογιών που συμβάλλουν σε μία πρόβλεψη, παρατηρείται μία αύξηση των τιμών του σφάλματος που οφείλεται στον τρόπο που χρησιμοποιήθηκε για τον υπολογισμό του.

Μελλοντικές Προεκτάσεις

Το σύστημα που παρουσιάστηκε στην παρούσα εργασία αποτελεί ένα ολοκληρωμένο σύστημα συστάσεων το οποίο εξάγει αξιόπιστα αποτελέσματα σύμφωνα με την αποτίμηση του σφάλματος που διενεργήθηκε στα τελευταία κεφάλαια. Παρά τις εξαιρετικές συστάσεις θα μπορούσε να δεχτεί στο μέλλον ορισμένες προεκτάσεις ή βελτιώσεις οι οποίες θα επέκτειναν τον τρόπο λειτουργίας του. Πιο συγκεκριμένα, θα μπορούσε να δοθεί η δυνατότητα στο χρήστη να αποδέχεται κάποια από αυτές τις συστάσεις είτε άμεσα, αξιολογώντας κάθε μία ξεχωριστά, είτε έμμεσα παρατηρώντας ποιες από αυτές τις προτάσεις μετατρέπονται σε αγορές. Στη συνέχεια, το σύστημα θα μπορούσε να εντάξει αυτό το χαρακτηριστικό στον υπολογισμό των μετέπειτα βαρών και την εξαγωγή των συστάσεων προς το συγκεκριμένο χρήστη.

Μία ακόμα ενδιαφέρουσα προέκταση θα μπορούσε να είναι εμπλοκή μεγαλύτερου αριθμού δημογραφικών χαρακτηριστικών όπως είναι το επάγγελμα στην κατηγοριοποίηση των χρηστών. Ένα τέτοιο ενδεχόμενο θα αύξανε κατά πολύ τον αριθμό των κατηγοριών με αποτέλεσμα οι συστάσεις που θα προέκυπταν θα ήταν πιθανότατα πιο στοχευμένες και ακριβείς. Τέλος, θα μπορούσε να γίνει μία αναλυτική μελέτη των αντοχών του παραπάνω συστήματος σε γνωστά προβλήματα των συστημάτων συστάσεων όπως είναι το πρόβλημα της ψυχρής εκκίνησης.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Recommender System	Σύστημα συστάσεων
E-shop	Ηλεκτρονικό κατάστημα
Email	Ηλεκτρονική αλληλογραφία
Cross-selling	Πρόσθετη πώληση
Up-selling	Υπεραξία πώλησης
Classification	Κατηγοριοποίηση
Clustering	Ομαδοποίηση
Content-based	Βάσει περιεχομένου
Collaborative filtering	Φιλτράρισμα συνεργασίας
Knowledge-based	Βάσει της γνώσης
Cold start	Ψυχρή εκκίνηση
Case based	Βασισμένα στην περίπτωση
Constraint based	Βασισμένα στους περιορισμούς
Engineering bottleneck	Μηχανική συμφόρηση
Weighted	Με βάρος
Switching	Μεταγωγή
Mixed	Ανακατεμένος
Feature Combination	Συνδυασμός χαρακτηριστικών
Feature Augmentation	Αύξηση χαρακτηριστικών
Cascade	Αλληλουχία
Meta-level	Μετα-επίπεδο
localhost	Τοπικός server
Pattern Recognition	Αναγνώριση μοτίβου
Biological Classification	Βιολογική κατηγοριοποίηση
Speech Recognition	Αναγνώριση ομιλίας
server	Εξυπηρετητής
Data Mining	Εξόρυξη δεδομένων
Computer Vision	Υπολογιστικό όραμα
Geostatistics	Γεωστατιστική
Training set	Συλλογή δεδομένων για εκπαίδευση
classifier	Κατηγοριοποιητής
supervised learning	Εποπτευόμενη μάθηση
Binary Classification	Δυαδική κατηγοριοποίηση
queries	Επερώτηση
Driver	Οδηγός
Database	Βάση δεδομένων
maximum a posteriori	Μέγιστη εκ των υστέρων
Rule	Κανόνας
User interface	Διεπαφή χρήστη
confusion matrix	Μήτρα σύγχυσης
soft counts	Μαλακό μέτρημα
hard-EM	Σκληρός ΕΜ
posterior	Μεταγενέστερο
threshold	Κατώφλι
Prediction	Πρόβλεψη
Count Ratings	Αριθμός βαθμολογιών
Item	Στοιχείο

Trained	Εκπαιδεύτηκε
Mean Absolute Error	Μέσο απόλυτο σφάλμα
Root Squared Mean Error	Μέση τετραγωνική ρίζα απόλυτου σφάλματος

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

RS	Recommender System
id	Identification number
JDBC	Java Database Connectivity
UI	User Interface
WEKA	Waikato Environment for Knowledge Analysis
MLE	Maximum Likelihood Expectation
EM	Expectation Maximisation
MAE	Mean Absolute Error
RS	Recommender System
RSME	Root Squared Mean Error

ΑΝΑΦΟΡΕΣ

[1] Daniel Lawd, Pedro Domingos, “Naïve Bayes Models for Probability Estimation”, Dept. Computer Science and Engineering, Univ. Washington.
[2] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[3] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
[4] https://en.wikipedia.org/wiki/C4.5_algorithm
[5] Christian Desrosiers, George Karypis, “A comprehensive survey of neighborhood-based recommendation methods”
[6] Robin Burke, Alexander Felfernig, Mehmet H. Goker, “Recommender System An Overview”
[7] Guido Jan de Nooij, “Recommender systems: An overview”, Vrije Universiteit Amsterdam
[8] Yehuda Koren and Robert Bell, “Advances in Collaborative Filtering, *Recommender Systems Handbook*,”
[9] Gediminas Adomavicius and Alexander Tuzhilin, “Context-Aware Recommender Systems, *Recommender Systems Handbook*”