

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία γενικευμένου συστήματος προτάσεων με εφαρμογή σε ένα σύνολο αλληλεξαρτώμενων δεδομένων

Εκπόνηση:                                          Επιβλέποντες:

Ιωάννης-Παναγιώτης Μπουντουρίδης                        Καθ. Περικλής Α.Μήτκας

giannis.[boudou@gmail.com](mailto:boudou@gmail.com)                                       Δρ. Αντώνης Χρυσόπουλος

ΑΕΜ: 8872                                          Ερευν. Σωτήρης Τσαρούχης

ΘΕΣΣΑΛΟΝΙΚΗ 2020

# Ευχαριστίες

Καταρχάς θα ήθελα να ευχαριστήσω τον καθηγητή κ. Περικλή Μήτκα την εμπιστοσύνη που μου έδειξε αναθέτοντάς αυτήν την διπλωματική εργασία.

Επίσης θα ήθελα να ευχαριστήσω τον μεταδιδακτορικό ερευνητή του εργαστηρίου Δρ. Αντώνη Χρυσόπουλο για την εξαιρετική συνεργασία κατά τη διάρκεια εκπόνησης της διπλωματικής εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω τον ερευνητή του εργαστηρίου Σωτήρη Τσαρούχη για τις συμβουλές του, την καθοδήγηση και τη συνεχή επικοινωνία που είχαμε.

Τέλος θα ήθελα να ευχαριστήσω τους ανθρώπους που με έχουν στηρίξει όσο κανένας άλλος δηλαδή τους γονείς μου Μπάμπη και Στέλλα και τον αδερφό μου Τάσο που αποτελούν τα σημαντικότερα πρόσωπα στην πορεία της ζωής μου.

# Περίληψη

Στην σημερινή εποχή μεγάλοι κολοσσοί λογισμικού και τεχνολογίας οφείλουν ένα σημαντικό κομμάτι της επιτυχίας τους στον χρόνο που επένδυσαν με στόχο να δημιουργήσουν σύνθετα και αξιόπιστα συστήματα προτάσεων. Ο ρόλος των συστημάτων προτάσεων εξυπηρετεί την ερμηνεία της δραστηριότητας των χρηστών και προτείνει προϊόντα ή υπηρεσίες που είναι όσο κατά το δυνατόν πιο πιθανά να είναι κοντά στα ενδιαφέροντα των χρηστών. H βιομηχανία της ένδυσης αποτελεί ένα πεδίο που αναμφίβολα βασίζεται στα παραπάνω συστήματα. Είναι ένας κλάδος συνεχώς αναπτυσσόμενος με την μόδα και τις τάσεις να κερδίζουν τις παραστάσεις ως τα πιο βασικά συστατικά της επιτυχίας.

Ο σχεδιασμός και η παραγωγή ρούχων φέρνει την ανάγκη έμπνευσης και καθοδήγησης των σχεδιαστών από υψηλής νοημοσύνης τεχνολογικά συστήματα προτάσεων. Τα δεδομένα που συλλέγονται τροφοδοτούν τα συστήματα προτάσεων τα οποία με την σειρά τους εξάγουν κάποια ισχυρά συμπεράσματα για τις προτιμήσεις και την οπτική που έχει ο σχεδιαστής δημιουργώντας αποτελεσματικές προτάσεις που αποτελούν ερεθίσματα στους σχεδιαστές για νέες κολεξιόν.

Σκοπός της παρούσας διπλωματικής είναι η δημιουργία ενός γενικευμένου συστήματος προτάσεων που θα εφαρμόζει τεχνικές προεπεξεργασίας δεδομένων και πρόβλεψης προτιμήσεων. Το πεδίο εφαρμογής στο παραπάνω γενικευμένο σύστημα θα είναι η βιομηχανία της ένδυσης και πιο συγκεκριμένα η δημιουργία ενός συστήματος το οποίο θα υποβοηθά τους σχεδιαστές παρέχοντας τις απαραίτητες προτάσεις για τον σχεδιασμό ρούχων.

Το τελικό σύστημα προτάσεων περιλαμβάνει τρία μοντέλα συνεργατικού φιλτραρίσματος και δύο μοντέλα επεξεργασίας περιεχομένου. Ο συνδυασμός αυτών των μοντέλων επιτυγχάνεται με την χρήση ενός υβριδικού συστήματος που συμβάλει στην αξιοποίηση όλων των θετικών στοιχείων κάθε υλοποίησης.

Τα πειράματα που διεξάχθηκαν αξιολογούνται χρησιμοποιώντας μετρικές που δίνουν άλλη διάσταση στην ερμηνεία του συστήματος καθώς αποτελούν οι βέλτιστες επιλογές που μπορεί να διαλέξει κανείς για να ερμηνεύσει συστήματα τόσο σύνθετου περιεχομένου. Η διαδικασία αυτή είναι απαραίτητη καθώς αποτελεί την επιστημονική τεκμηρίωση της ορθής λειτουργίας του τελικού συστήματος.

# Abstract

Να γίνει η μετάφραση της περίληψης μετά την ολοκλήρωση της

# Λίστα εικόνων

[Εικόνα 1 - Διάγραμμα Ρίζας Μέσου Τετραγωνικού Σφάλματος (RMSE) 25](#_Toc54541165)

[Εικόνα 2 - Διάγραμμα Μέσου Απόλυτου Σφάλματος (MAE) 26](#_Toc54541166)

# Λίστα πινάκων

[Πίνακας 1 - Αναπαράσταση δεδομένων και εύρεση ομοιότητας 16](#_Toc54270775)

[Πίνακας 2- Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων 19](#_Toc54270776)

[Πίνακας 3 - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων - Υπολογισμός ομοιότητας 19](#_Toc54270777)

[Πίνακας 4 - Αναπαράσταση πίνακα με κεντραρισμένη μέση τιμή αξιολογήσεων 20](#_Toc54270778)

Πίνακας περιεχομένων

[Ευχαριστίες 2](#_Toc58837920)

[Περίληψη 3](#_Toc58837921)

[Abstract 4](#_Toc58837922)

[Λίστα εικόνων 5](#_Toc58837923)

[Λίστα πινάκων 6](#_Toc58837924)

[1 Εισαγωγή 9](#_Toc58837925)

[1.1 Γενικά 10](#_Toc58837926)

[1.2 Ορισμός προβλήματος 10](#_Toc58837927)

[1.3 Στόχοι διπλωματικής 11](#_Toc58837928)

[1.4 Μεθοδολογία διπλωματικής 11](#_Toc58837929)

[1.5 Οργάνωση Κεφαλαίων 11](#_Toc58837930)

[2 Θεωρητικό Υπόβαθρο 13](#_Toc58837931)

[2.1 Κατηγορίες συστημάτων προτάσεων 13](#_Toc58837932)

[2.1.1 Συστήματα Προτάσεων Συνεργατικού Φιλτραρίσματος (Collaborative filtering) 13](#_Toc58837933)

[2.1.2 Συστήματα Προτάσεων φιλτραρίσματος περιεχομένου (Content-based Filtering) 18](#_Toc58837934)

[2.1.3 Υβριδικά Συστήματα Προτάσεων 19](#_Toc58837935)

[2.2 Μετρικές αξιολόγησης (Χρήσιμες μετρικές) 19](#_Toc58837936)

[2.2.1 Ομοιότητα δεδομένων 19](#_Toc58837937)

[2.2.2 Ρίζα μέσου τετραγωνικού σφάλματος (RMSE) 20](#_Toc58837938)

[2.2.3 Μέσο απόλυτο σφάλμα (MAE) 21](#_Toc58837939)

[2.2.4 Κάλυψη (Coverage) 21](#_Toc58837940)

[2.2.5 Ποικιλομορφία (diversity) 22](#_Toc58837941)

[2.2.6 Καινοτομία (novelty) 22](#_Toc58837942)

[3 Υλοποίηση συστημάτων προτάσεων 24](#_Toc58837943)

[3.1 Συνεργατικό φιλτράρισμα με επίκεντρο τον χρήστη 24](#_Toc58837944)

[3.1.1 Προεπεξεργασία δεδομένων (pre-processing) 24](#_Toc58837945)

[3.1.2 Εύρεση όμοιων χρηστών 25](#_Toc58837946)

[3.1.3 Στάθμιση σημασίας ομοιότητας (significance weighting) 25](#_Toc58837947)

[3.1.4 Κεντράρισμα μέση τιμής δεδομένων (mean-center data) 26](#_Toc58837948)

[3.1.5 Συνάρτηση πρόβλεψης αξιολόγησης 27](#_Toc58837949)

[3.2 Συνεργατικό φιλτράρισμα με επίκεντρο τον προϊόν 27](#_Toc58837950)

[3.2.1 Προεπεξεργασία δεδομένων (pre-processing) 27](#_Toc58837951)

[3.2.2 Κεντράρισμα μέση τιμής δεδομένων (mean-center data) 28](#_Toc58837952)

[3.2.3 Εύρεση όμοιων προϊόντων (ρούχων) 28](#_Toc58837953)

[3.2.4 Συνάρτηση πρόβλεψης αξιολόγησης 29](#_Toc58837954)

[3.3 Συνεργατικό φιλτράρισμα με ανάλυση πίνακα σε ιδιάζουσες τιμές 29](#_Toc58837955)

[3.3.1 Προεπεξεργασία δεδομένων (pre-processing) 29](#_Toc58837956)

[3.3.2 Διαχωρισμός δεδομένων (train – test split) 30](#_Toc58837957)

[3.3.3 Μετατροπή δεδομένων εκπαίδευσης σε νέο πίνακα χρηστών-αξιολογήσεων 31](#_Toc58837958)

[3.3.4 Ανάλυση δεδομένων εκπαίδευσης σε πίνακα ιδιαζουσών τιμών 32](#_Toc58837959)

[4 Πειράματα και αποτελέσματα 33](#_Toc58837960)

[4.1 Υλοποίηση συνεργατικού φιλτραρίσματος με επίκεντρο τον χρήστη 33](#_Toc58837961)

[4.1.1 Διάγραμμα Ρίζας Μέσου Τετραγωνικού Σφάλματος (RMSE) ανά γείτονα 34](#_Toc58837962)

[4.1.2 Διάγραμμα Μέσου Απόλυτου Σφάλματος (MAE) ανά γείτονα 34](#_Toc58837963)

[4.1.3 Πίνακας επιδόσεων εκτίμησης αξιολογήσεων 35](#_Toc58837964)

[4.2 Υλοποίηση συνεργατικού φιλτραρίσματος με ανάλυση πίνακα ιδιαζουσών τιμών 37](#_Toc58837965)

[4.2.1 Διάγραμμα Ρίζας Μέσου Τετραγωνικού Σφάλματος (RMSE) ανά feature 37](#_Toc58837966)

[4.2.2 Διάγραμμα Ρίζας Μέσου Απόλυτου Σφάλματος (MAE) ανά feature 38](#_Toc58837967)

[4.2.3 Πίνακας επιδόσεων εκτίμησης αξιολογήσεων 39](#_Toc58837968)

[Βιβλιογραφία 40](#_Toc58837969)

Κεφάλαιο

# Εισαγωγή

## Γενικά

Στην σημερινή εποχή ένας ευρέως αναπτυσσόμενος κλάδος είναι αυτός της ένδυσης και της μόδας που αποτελεί μία κερδοφόρα και αρκετά εξελισσόμενη βιομηχανία. Σε πολλές χώρες του κόσμου η έννοια της ένδυσης παύει να αποτελεί απλά μία βασική ανάγκη. Πολλοί άνθρωποι οδηγούνται στην αγορά προϊόντων ένδυσης για την ικανοποίηση ευρύτερων αναγκών όπως κοινωνικών, επαγγελματικών ή και προσωπικών .

Σε αυτό καταναλωτικό κλίμα μεγάλοι οίκοι μόδας καλούνται να λύσουν προβλήματα που θα μεγιστοποιήσουν την απήχηση και το κέρδος που έχουν τα ρούχα τους. Σκοπός για την επίτευξη αυτού του στόχου είναι η δημιουργία κολεξιόν που θα ακολουθούν τις συνεχώς εξελισσόμενες τάσεις της μόδας που θα κάνουν τους καταναλωτές να ενδώσουν. Η έλλειψη πωλήσεων και η δημιουργία υψηλών αποθεμάτων ρούχων παρωχημένων είναι προβλήματα που πάση θυσία μεγάλοι κολοσσοί προσπαθούν να αποφύγουν.

Συνεπώς μεγάλοι οίκοι μόδας σε συνεργασία με άλλες εταιρίες προϊόντων παραγωγής ένδυσης οφείλουν σε κάθε καμπάνια τους να δημιουργούν τις προϋποθέσεις ώστε να κερδίζουν τις εντυπώσεις, να μειώνουν την αβεβαιότητα των πωλήσεων τους και να ακολουθούν πιστά τις τάσεις ώστε τελικά οδηγηθούν στην παραπάνω επίτευξη του στόχου τους.

## Ορισμός προβλήματος

Μεγάλοι οίκοι μόδας για να κατανοήσουν και να προσαρμοστούν στις ανάγκες και τις προτιμήσεις των καταναλωτών ακολουθούν και προσανατολίζονται με τεχνολογικά εργαλεία που είναι εύκολο να αποκτήσει πρόσβαση κανείς, στην σημερινή κοινωνία.

Σκοπός αυτών των τεχνολογικών εργαλείων δεν είναι να αντικαταστήσουν τους σχεδιαστές που εργάζονται σε αυτούς τους οίκους. Άλλωστε μία έννοια τόσο αφηρημένη όσο η μόδα χρειάζεται και την βοήθεια της ανθρώπινης αισθητικής και δεν θα μπορούσε ποτέ να καλυφθεί μονομερώς μέσα από μια αυτοματοποιημένη διαδικασία της τεχνολογίας. Βασικός στόχος των εργαλείων είναι να σταθούν υποβοηθοί του σχεδιαστή και να τον καθοδηγήσουν σε μία νέα πετυχημένη κολεξιόν προσαρμοσμένη στις ανάγκες των καταναλωτών.

## Στόχοι διπλωματικής

Στόχος της παρούσας διπλωματικής είναι η δημιουργία ενός γενικευμένου συστήματος προτάσεων που θα δέχεται ως είσοδο ένα σύνολο δεδομένων. Το σύστημα αυτό θα περιλαμβάνει τεχνικές προεπεξεργασίας δεδομένων και παροχής προτάσεων. Το σύνολο δεδομένων που δίνεται ως είσοδο στην παρούσα εργασία είναι αξιολογήσεις χρηστών σε ρούχα αλλά και τα χαρακτηριστικά τα οποία διέπουν. Μετά την είσοδο των δεδομένων, βάζει σε λειτουργία υλοποιημένες τεχνικές και προτείνει στο σχεδιαστή προϊόντα ώστε να μπορεί να σχεδιάζει πιο γρήγορα και αποτελεσματικά τα ρούχα που επιθυμεί. Σκοπός λοιπόν είναι η οπτική παροχή ρούχων που θα αποτελέσουν πηγή έμπνευσης για τη δημιουργία μιας πετυχημένης κολεξιόν. Οι σχεδιαστές θα χρησιμοποιούν ένα σύστημα στο οποίο οι καταγεγραμμένες προτιμήσεις τους θα αναλύονται από διάφορα μοντέλα και θα προτείνονται ρούχα ένδυσης τα οποία θα είναι όσο το δυνατόν γίνεται κοντά στις προτιμήσεις τους.

## Μεθοδολογία διπλωματικής

Για την πραγματοποίηση της διπλωματικής εργασίας χρησιμοποιήθηκε σετ δεδομένων που περιλαμβάνει:

1. Τα χαρακτηριστικά κάθε ρούχου
2. Αξιολογήσεις από χρήστες

Τα στάδια για την δημιουργία του τελικού συστήματος ήταν:

1. Η προεπεξεργασία του συνόλου δεδομένων των ρούχων
2. Η δημιουργία αξιολογήσεων που προσομοιώνει ανθρώπινη συμπεριφορά
3. Η δημιουργία των υλοποιήσεων του γενικευμένο συστήματος προτάσεων
4. Η χρήστη μετρικών για την αξιολόγηση και επιστημονική τεκμηρίωση της ορθής λειτουργίας του συστήματος

## Οργάνωση Κεφαλαίων

Κεφάλαιο

# Θεωρητικό Υπόβαθρο

## Κατηγορίες συστημάτων προτάσεων

Μετά από μια εισαγωγική ανασκόπηση για την έννοια, τον ρόλο και την συμβολή στα συστήματα προτάσεων είναι αναγκαίο να αναλυθούν κάποιες συγκεκριμένες κατηγορίες συστημάτων που όπως θα δούμε στα επόμενα κεφάλαια θα αποτελέσουν θεμέλια για την δημιουργία του τελικού συστήματος υποβοήθησης των σχεδιαστών.

### Συστήματα Προτάσεων Συνεργατικού Φιλτραρίσματος (Collaborative filtering)

Τα συστήματα προτάσεων που ανήκουν στην κατηγορία του συνεργατικού φιλτραρίσματος (collaborative filtering), σε αντίθεση με άλλα συστήματα, δε χρησιμοποιούν πληροφορίες περιεχομένου των προϊόντων, παρά μόνο τον τις βαθμολογίες χρηστών δηλαδή τον πίνακα χρήστη-αντικειμένου βαθμολογιών (user-item, rating matrix). H βασική ιδέα σε αυτή την κατηγορία αλγόριθμων είναι η εύρεση χρηστών με αρκετά παρόμοιες αξιολογήσεις. Η αποτελεσματικότητα των αλγορίθμων στηρίζεται στην παρατήρηση ότι οι χρήστες που έχουν παρόμοιες προτιμήσεις, αξιολογούν με παρόμοιο τρόπο. Οι τεχνικές αυτές συνήθως λαμβάνουν ένα σύνολο από τις βαθμολογίες των χρηστών του συστήματος και παράγουν προβλέψεις σχετικά με το τι χρειάζεται ένας χρήστης, βασιζόμενες στους πιο κοντινούς (ως προς τις προτιμήσεις) σε αυτόν χρήστες. Tα μοντέλα συνεργατικού φιλτραρίσματος χρησιμοποιούν αξιολογήσεις παρεχόμενες από όμοιους χρήστες με σκοπό τη δημιουργία προτάσεων. Η κυρία πρόκληση στην σχεδίαση ενός μοντέλου συνεργατικό φιλτραρίσματος είναι η σωστή διαχείριση των πινάκων αξιολογήσεων (user-item rating matrix). Οι πίνακες αυτοί, καθώς πολλές θέσεις αξιολογήσεων παραμένουν κενές έχουν πυκνότητα που είναι αρκετά αραιή (sparse). Πιο συγκεκριμένα σε μία βάση δεδομένων με χιλιάδες ρούχα οι περισσότεροι χρήστες αξιολογούν ένα μικρό κομμάτι του συνόλου όλων των ρούχων ως αποτέλεσμα οι περισσότερες αξιολογήσεις να παραμένουν απροσδιόριστες. Για κάθε χρήστη η κυρία ιδέα του συνεργατικού φιλτραρίσματος είναι να καλύψει αυτές στις απροσδιόριστες αξιολογήσεις λαμβάνοντας υπ’ όψη τους υψηλά συσχετιζόμενους χρήστες που έχουν παρόμοιες προτιμήσεις. Υπάρχουν δύο είδη μεθόδων που χρησιμοποιούνται κατά κόρον στις υλοποιήσεις συνεργατικού φιλτραρίσματος γνωστοί και ως μέθοδοι βασισμένοι στη μνήμη (memory-based methods) και μέθοδοι βασισμένοι στα μοντέλα (model-based methods).

#### Συστήματα βασισμένα στη μνήμη (memory-based methods)

Oι μέθοδοι βασισμένοι στη μνήμη (memory-based methods) αποτελούν τα θεμέλια του συνεργατικό φιλτραρίσματος καθώς αποτέλεσαν τις πρώτες επιδιώξεις δημιουργίας προτάσεων που βασίστηκαν σε μια παρατήρηση ότι όμοιες αξιολογήσεις υποδηλώνουν και όμοιες προτιμήσεις. Για να εξερευνήσει κανείς την προτίμηση των χρηστών με αυτή την κατηγορία μεθόδων μπορεί να το κάνει με δύο τεχνικές: η μια τεχνική έχει επίκεντρο τις βαθμολογίες του χρήστη ενώ η άλλη τεχνική έχει επίκεντρο τις βαθμολογίες των προϊόντων.

##### Συστήματα μνήμης με επίκεντρο τον χρήστη (user-based)

Σε αυτή την περίπτωση αν υποθέσουμε ότι ο χρήστης Α είναι το επίκεντρο που θέλουμε να κάνουμε προτάσεις, τότε χρήστες όμοιοι με τις αξιολογήσεις του A δηλαδή χρήστες με όμοιες προτιμήσεις αξιοποιούνται για την εκτίμηση των αξιολογήσεων του χρήστη Α σε προϊόντα που δεν έχει εξερευνήσει. Το σύνολο των αξιολογήσεων του χρήστη Α συγκρίνεται με τα σύνολα αξιολογήσεων των υπόλοιπων χρηστών ώστε να βρεθούν ποιοι χρήστες είναι πιο όμοιοι. Η ομοιότητα κάθε χρήστη, με του χρήστη Α, βασίζεται μόνο σε αξιολογήσεις που έχουν κάνει οι χρήστες σε κοινά προϊόντα. Συναρτήσεις ομοιότητας με είσοδο τις κοινές αξιολογήσεις των συγκρινόμενων χρηστών επιστρέφουν μια τιμή που κυμμένεται από το -1 έως 1. Όσο πιο μεγάλη τιμή επιστρέφει η συνάρτηση τόσο πιο όμοιες θεωρούνται οι αξιολογήσεις των δύο συγκρινόμενων χρηστών. Ένα πρόβλημα που προκύπτει είναι ότι κάποιοι χρήστες τείνουν να βαθμολογούν αρκετά ψηλά ενώ άλλοι χαμηλά. Η λύση σε αυτό το πρόβλημα είναι η αφαίρεση της μέσης τιμής των αξιολογήσεων του χρήστη από κάθε αξιολόγηση προϊόντος ξεχωριστά ώστε να δημιουργηθούν mean-centered δεδομένα.

##### Συστήματα μνήμης με επίκεντρο το προϊόν (item-based)

Σε αυτή την περίπτωση αν πάλι υποθέσουμε ότι ο χρήστης Α είναι το επίκεντρο που θέλουμε να κάνουμε προτάσεις, τότε για ένα προϊόν Β ψάχνουμε ένα σύνολο προϊόντων S όμοιο με το προϊόν Β και χρησιμοποιούμε τις αξιολογήσεις του συνόλου S για την πρόβλεψη της αξιολόγησης του Β. Είναι προφανές λοιπόν πως σε αυτή την μέθοδο για την πρόβλεψη των απροσδιόριστων αξιολογήσεων του χρήστη Α γίνεται αναζήτηση των πιο όμοιων προϊόντων και όχι των πιο όμοιων χρηστών. Συνεπώς τα Κ πιο όμοια προϊόντα του συνόλου S έχουν αξιολογήσεις που θα καθορίσουν το αποτέλεσμα της πρόβλεψης για την τιμή αξιολόγησης που θα πάρει το προϊόν Β.

#### Συστήματα βασισμένα στα μοντέλα (model-based methods)

Σε αντίθεση με τους αλγόριθμους που βασίζονται στη μνήμη, οι αλγόριθμοι βασιζόμενοι στα μοντέλα χρησιμοποιούν τις βαθμολογίες των χρηστών προκειμένου να εκπαιδεύσουν ένα μοντέλο. Οι μέθοδοι βασισμένοι στα μοντέλα προσαρμόζουν ένα παραμετροποιημένο σύστημα στον πίνακα αξιολογήσεων το οποίο στη συνέχεια χρησιμοποιείται για την πρόβλεψη απροσδιόριστων βαθμολογιών και κατ’ επέκταση για τη δημιουργία προτάσεων προς τους χρήστες. Σε αυτή την υποκατηγορία συστημάτων ανήκουν: οι ταξινομητές κοντινότερου γείτονα (knn classifiers), οι Bayesian ταξινομητές (Bayesian classifiers) αλλά και οι μέθοδοι βασιζόμενοι σε παλινδρόμηση (regression-based methods). Το ευρύτερο φάσμα αυτών των τεχνικών είναι: η μηχανική μάθηση (machine learning) και η εξόρυξη δεδομένων (data mining). Μία αποτελεσματική κατηγορία μοντέλων που βασίζεται στην χαμηλού βαθμού παραγοντοποίηση μήτρας (low-rank matrix factorization) με τεχνικές όπως το Singular Value Decomposition (SVD), το regularized SVD, το Non-negative Matrix Factorization (NMF), το Probabilistic Matrix Factorization (PMF), το Bayesian PMF, το Non-linear PMF, το Maximum Margin Matrix Factorization (MMMF) και Nonlinear Principal Component Analysis (NPCA) . Τα μοντέλα που αναφέρθηκαν τελευταία αποκαλούνται λανθάνοντα παράγοντα (Latent factor models) [1]. Αυτές οι μέθοδοι έχουν γίνει δημοφιλείς τα τελευταία χρόνια επειδή συνδυάζουν καλό βαθμό κλιμάκωσης και ευστοχία προβλέψεων. Επιπρόσθετα, προσφέρουν ευελιξία στην μοντελοποίηση διαφόρων καταστάσεων από την πραγματική ζωή. Ωστόσο, επισημαίνεται ότι όλων των ειδών οι model-based προσεγγίσεις είναι συχνά χρονοβόρες στην υλοποίηση και την ενημέρωσή τους. [2]

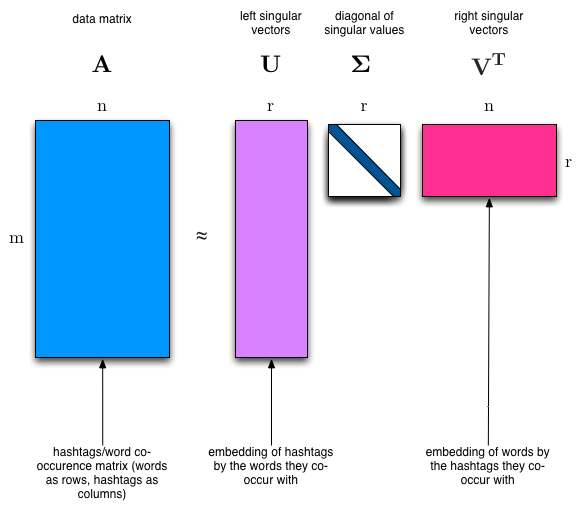
##### Συστήματα ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD)

Η μέθοδος ανάλυσης πίνακα σε ιδιάζουσες τιμές (Singular Value Decomposition) αποτελεί μια τεχνική παραγοντοποίησης πινάκων με βάση την οποία ένας πίνακας R διαστάσεων παραγοντοποιείται στο γινόμενο πινάκων

όπου:

* ένας ορθογώνιος πίνακας διαστάσεων
* ο ανάστροφος πίνακας διαστάσεων
* ένας διαγώνιος πίνακας διαστάσεων

Τα στοιχεία του πίνακα είναι οι ιδιάζουσες τιμές του πίνακα και είναι ταξινομημένα σε σειρά φθίνουσα. Με την βοήθεια των πινάκων και οι χρήστες και τα προϊόντα (ρούχα) μπορούν να αναπαρασταθούν σε έναν χώρο διαστάσεων. Η χρησιμότητα της ανάλυσης πίνακα σε ιδιάζουσες τιμές έγκειται στη δυνατότητα διατήρησης των k μεγαλύτερων ιδιαζουσών τιμών του διαγώνιου πίνακα και η αποβολή των υπόλοιπων τιμών όπως αντίστοιχα η διατήρηση των πρώτων στηλών των πινάκων και και η αποβολή των υπόλοιπων στηλών. Τα πλεονεκτήματα που προκύπτουν από τη χρήση της ανάλυσης πίνακα σε ιδιάζουσες τιμές συμβάλλουν σημαντικά στην απόδοση και ακρίβεια του συστήματος συστάσεων.



Εικόνα - Αναπαράσταση ανάλυσης πίνακα σε ιδιάζουσες τιμές (SVD) [3]

Πρώτον, σημαντική είναι η ανάλυση των λανθανουσών σχέσεων που υπάρχουν ανάμεσα στους χρήστες και τα προϊόντα (ρούχα). Κάθε ιδιάζουσα τιμή αναφέρεται σε έναν άγνωστο παράγοντα, που χαρακτηρίζει σημαντικά τη σχέση ανάμεσα στους χρήστες και τα προϊόντα. Διατηρώντας τις μεγαλύτερες ιδιάζουσες τιμές, εξασφαλίζεται η διατήρηση των πιο σημαντικών παραγόντων, άσχετα εάν η ερμηνεία τους είναι κατά πάσα πιθανότητα αδύνατη. Επιπλέον, η τεχνική αυτή εξασφαλίζει μείωση διαστάσεων (dimensionality reduction) του αρχικού προβλήματος. Εξαλείφοντας τις μικρότερες ιδιάζουσες τιμές αρκεί να αποθηκεύσουμε μόνο τους τρεις παραγόμενους πίνακες μειώνοντας το σύνολο των αποθηκευμένων στοιχείων από σε .Συνεπώς η επιλογή της τιμής έχει άμεσες συνέπειες στην ακρίβεια συστάσεων αλλά και στη απόδοσή του συστήματος ως προς τις απαιτήσεις σε αποθηκευτικό χώρο. Τα μειονεκτήματα της προσέγγισης αφορούν τον χρόνο εκτέλεσής της, ο οποίος είναι συνήθως μεγαλύτερος από τη προσέγγιση συνεργατικού φιλτραρίσματος με βάση τη μνήμη. Ωστόσο τα αποτελέσματα της ανάλυσης του αρχικού πίνακα μπορούν να αποθηκευτούν και να επαναχρησιμοποιηθούν για την παραγωγή προβλέψεων. Μάλιστα, εξαιτίας της μείωσης των διαστάσεων, παρουσιάζονται πιο σύντομοι χρόνοι εκτέλεσης κατά το στάδιο δημιουργίας συστάσεων σε σύγκριση με τα αντίστοιχα στάδια των υπόλοιπων προσεγγίσεων. Δεύτερο μειονέκτημα αποτελεί η ανάγκη για εκ νέου εκτέλεση της διαδικασίας ανάλυσης εξαιτίας αλλαγών στον αρχικό πίνακα δεδομένων (προσθήκη νέων χρηστών ή αντικειμένων). Στη βιβλιογραφία όμως παρατηρούνται προσπάθειες για τη δημιουργία τεχνικών ανανέωσης και επέκτασης των τριών παραγόμενων πινάκων (SVD updating & folding up). Τέλος η προσέγγιση απαιτεί συνήθως απαλοιφή τον μηδενικών τιμών του αρχικού πίνακα μέσω εξειδικευμένης προεπεξεργασίας. [4]

### Συστήματα Προτάσεων Φιλτραρίσματος Περιεχομένου (Content-based Filtering)

Οι μέθοδοι φιλτραρίσματος βασισμένοι στο περιεχόμενο έχουν κύριο στόχο την ανάλυση των βασικών χαρακτηριστικών και ιδιοτήτων των προϊόντων. Αναγκαία προϋπόθεση είναι η ύπαρξη ενός συνόλου δεδομένων που θα περιλαμβάνει αναλυτικά τα χαρακτηριστικά κάθε προϊόντος. Τα δεδομένα αυτά υφίστανται απαραίτητη επεξεργασία ώστε να εξαχθούν τα πιο καίρια και σημαντικά χαρακτηριστικά και να αποβληθεί η περιττή πληροφορία. Τα χαρακτηριστικά κάθε ρούχου ορίζονται από την βιομηχανία παραγωγής του. Το μέγεθος, η μάρκα, το μάκρος των μανικιών και άλλα χαρακτηριστικά είναι λίγο πολύ τετριμμένα δηλαδή αποτελούν έννοιες που χρησιμοποιούνται ευρέως σε όλο τον κόσμο με αποτέλεσμα η προεπεξεργασία των δεδομένων να αποτελεί εύκολη διαδικασία για τον διαχειριστή. Μετά την ολοκλήρωση αυτής της διαδικασίας δύναται η δυνατότητα αναπαράστασης ενός προϊόντος ως ένα διάνυσμα σε έναν πεπερασμένο χώρο διαστάσεων. Με αυτή την αναπαράσταση γίνεται εφικτό να αξιοποιηθούν διάφορα μοντέλα, να εκπαιδευτούν ώστε τελικά να προτείνουν νέα προϊόντα της προτίμησης αντίστοιχα με της προηγούμενης αρεσκείας. Δυστυχώς, τα συστήματα βασισμένα στο περιεχόμενο υποφέρουν από σημαντικούς περιορισμούς που τα καθιστούν δύσχρηστα και συνεπώς τις περισσότερες φορές συνδυάζονται με άλλες τεχνικές και μεθόδους. Βασικό πρόβλημα αποτελεί η δυσκολία ορισμού των μεταβλητών που περιγράφουν τα προφίλ χρηστών και προϊόντων. Η συλλογή δεδομένων για κάθε προϊόν ξεχωριστά καθώς και η προεπεξεργασία των δεδομένων αυτών είναι χρονοβόρα διαδικασία. Επιπλέον, δύσκολο εγχείρημα αποτελεί και η επιλογή των κατάλληλων μεταβλητών (feature selection), η οποία συνήθως προϋποθέτει ιδιαίτερη γνώση και εμπειρία για τη φύση του προβλήματος. Εξίσου σημαντικό πρόβλημα είναι το πρόβλημα της εξειδίκευσης. Τα προϊόντα που προτείνονται από το σύστημα περιορίζονται κυρίως στα χαρακτηριστικά των προϊόντων που παρουσίασε αρχικά ο χρήστης. Με αυτό τον τρόπο, είναι πρακτικά αδύνατο να παρουσιαστούν νέα προϊόντα στον χρήστη που να καλύπτουν διαφορετικές ανάγκες του. Τέλος ιδιαίτερο πρόβλημα αποτελεί η αδυναμία δημιουργίας προφίλ και παροχής συστάσεων στους νέους χρήστες του συστήματος εξαιτίας της έλλειψης σε δεδομένα που περιγράφουν προηγούμενες αγορές ή προτιμήσεις τους.

### Υβριδικά Συστήματα Προτάσεων (Hybrid)

Τα Υβριδικά Συστήματα Προτάσεων συνδυάζουν δύο ή περισσότερες τεχνικές για να επιτύχουν καλύτερη απόδοση και να περιορίσουν όσο περισσότερο γίνεται τα μειονεκτήματα οποιουδήποτε μεμονωμένου συστήματος πρότασης. Φυσικά ο τρόπος με τον οποίο συνδυάζονται τα συστήματα ποικίλει, παρακάτω παρουσιάζονται οι επικρατέστερες κατηγορίες υβριδικών συστημάτων σύμφωνα πάντα βιβλιογραφία.

#### Σταθμισμένα Υβριδικά Συστήματα

Στα σταθμισμένα υβριδικά συστήματα, βαθμολογίες διαφορετικών συστημάτων προτάσεων συνδυάζονται για τον υπολογισμό της εξόδου του συστήματος. Οι αξιολογήσεις προϊόντων από χρήστες περιέχονται στον πίνακα αξιολογήσεων *διαστάσεων* . Ένα υβριδικό σύστημα που περιλαμβάνει συστήματα προτάσεων παράγει νέους πίνακες αξιολογήσεων . Οι έξοδοι των συστημάτων συνδυάζονται με την χρήση ενός συνόλου βαρών για την εκτίμηση του τελικού αποτελέσματος. Για ένα σύνολο βαρών η συνάρτηση εξόδου του σταθμισμένου υβριδικού συστήματος ορίζεται ώς:

Στην πιο απλή περίπτωση οι τιμές για τα βάρη παίρνουν την τιμή της ίσης ψήφου, στην συγκεκριμένη περίπτωση αυτή είναι . Ωστόσο αυτή η περίπτωση είναι αρκετά ιδανική καθώς τα συστήματα προτάσεων ποικίλουν. Για να επιλεχθούν τα βέλτιστα βάρη είναι απαραίτητο να αξιολογηθεί η επίδοση της αποτελεσματικότητας του συνδυασμού των βαρών

σύστημα υπολογίζει την βαθμολογία ενός αντικειμένου συναρτήσει του αποτελέσματος όλων των διαθέσιμων τεχνικών που υπάρχουν στο σύστημα. Έστω R = [] οι βαθμολογίες ενός πίνακα. Στα σταθμισμένα υβριδικά συστήματα η έξοδος συνδυάζεται με ένα σύνολο από βάρη ώστε αν οι τελικές εκτιμήσεις των αξιολογήσεων για χρήστες σε προϊόντα τότε ένα σύνολο από βάρη ορίζεται και η συνάρτηση εκτίμησης των αξιολογήσεων ορίζεται ως:

Στην απλούστερη περίπτωση γίνεται η επιλογή

## Μετρικές αξιολόγησης (Χρήσιμες μετρικές)

### Ομοιότητα δεδομένων

Υπάρχουν διαφορές μετρικές για τον υπολογισμό της ομοιότητας των δεδομένων, οι επικρατέστερες είναι η μετρική Pearson (Pearson correlation similarity) και η μετρική του συνημίτονου (Cosine similarity). Κάθε μέθοδος επιστρέφει έναν αριθμό από το -1 ως το 1. Όσο μεγαλύτερος είναι ο αριθμός αυτός τόσο μεγαλύτερη είναι θεωρείται πως είναι η ομοιότητα των δεδομένων. Στο παράδειγμα που ακολουθεί έχουμε ένα πίνακα αξιολογήσεων χρηστών v και u και αντίστοιχα τους τύπους για τον υπολογισμό της ομοιότητας των δύο χρηστών.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *cloth 1* | *cloth 2* | *cloth 3* | *cloth 4* | *cloth 5* |
| *User: v* | 7 |  | 3 | 8 | 2 |
| *User: u* | 9 | 6 | 2 |  | 2 |

Πίνακας - Αναπαράσταση δεδομένων και εύρεση ομοιότητας

Εξίσωση - Μετρική ομοιότητας pearson

Εξίσωση - Μετρική ομοιότητας συνημίτονου

Όπου οι θέσεις που βαθμολόγησε ο χρήστης *u*, οι θέσεις που βαθμολόγησε ο χρήστης *v* και οι θέσεις που βαθμολόγησαν και οι δύο χρήστες *u* και *v.*

Αντίστοιχα και είναι οι βαθμολογίες των χρηστών u και v στο προϊόν k ενώ και είναι η μέσες βαθμολογίες των χρηστών u και v.

Από το πιο πάνω παράδειγμα προκύπτει ότι

Συνεπώς οι ομοιότητες των χρηστών θα είναι ως προς pearson και για την ομοιότητα συνημίτονου.

### Ρίζα μέσου τετραγωνικού σφάλματος (RMSE)

Η ρίζα μέσου τετραγωνικού σφάλματος είναι μια μετρική αξιολόγησης της εκτίμησης ενός μοντέλου πρόβλεψης. Δεδομένου ότι το σφάλμα πρόβλεψης τετραγωνίζεται η μετρική RMSE δίνει βάρος περισσότερο στα σφάλματα με μεγάλη απόκλιση και λιγότερο σε αυτά με μικρή. [5] Αυτό σημαίνει ότι για μοντέλα που τα μεγάλα σφάλματα είναι ανεπιθύμητα ή ακόμη και απαγορευτικά, η μετρική αυτή θα πρέπει να λαμβάνεται σοβαρά υπ’ όψη. Πιο αναλυτικά η μετρική δίνεται από τον τύπο:

όπου είναι οι τιμές εκτίμησης, είναι οι πραγματικές τιμές και είναι ο αριθμός των πραγματικών τιμών.

### Μέσο απόλυτο σφάλμα (MAE)

To μέσο απόλυτο σφάλμα είναι μια μετρική αξιολόγησης που μετρά πόσο κοντά εκτιμήθηκαν οι προβλέψεις του συστήματος σε σχέση με τις πραγματικές βαθμολογίες. Ο τύπος περιγράφεται από την σχέση:

όπου είναι οι τιμές εκτίμησης, είναι οι πραγματικές τιμές και είναι ο αριθμός των πραγματικών τιμών.

Το μέσο απόλυτο σφάλμα σε αντίθεση με την ρίζα του μέσου τετραγωνικού σφάλματος λαμβάνει υπ’ όψη τα μεγάλα και μικρά σφάλματα εκτίμησης με το ίδιο βάρος [5] ενώ αγνοεί την κατεύθυνση του σφάλματος αφού περιλαμβάνει την μέση τιμή στον υπολογισμό του τύπου.

### Κάλυψη (Coverage)

Σύμφωνα με την βιβλιογραφία [6] η κάλυψη (coverage) ενός συστήματος προτάσεων είναι το ποσοστό των στοιχείων από το σύνολο του συστήματος για τα οποία το ίδιο το σύστημα μπορεί να κάνει προτάσεις. Ο όρος κάλυψη συσχετίστηκε κυρίως με δύο έννοιες:

* Με το ποσοστό των προϊόντων για τα οποία το σύστημα είναι σε θέση να δημιουργήσει μια πρόταση
* Με το ποσοστό των διαθέσιμων προϊόντων που πράγματι συνιστώνται ποτέ σε χρήστες

Παρόλο που διαφορετικοί συγγραφείς διαφέρουν ως προς την ορολογία, στην παρούσα διπλωματική εργασία υιοθετείται ο πρώτος ορισμός δηλαδή ως κάλυψη προβλέψεων από το σύνολο του καταλόγου. [7]

### Ποικιλομορφία (diversity)

Υπάρχουν πολλοί ορισμοί που περιγράφουν την έννοια της ποικιλομορφίας ως μετρική, η απλούστερη είναι αυτή του Bradley και Smyth που ορίζουν την ποικιλομορφία ως το αντίθετο της ομοιότητας η οποία και υιοθετείται στην παρούσα διπλωματική. [8] Πιο συγκεκριμένα θεωρούμε την ποικιλομορφία ώς τον μέσο όρο της ανομοιότητας μεταξύ όλων των πιθανών ζευγών που μπορούν να προκύψουν από μια πρόταση. Ο τύπος περιγράφεται από την σχέση:

Είναι σαφές ότι η ποιότητα της μέτρησης εξαρτάται σε μεγάλο βαθμό από την συνάρτηση ομοιότητας που θα εφαρμοστεί στο σύνολο των δεδομένων. [9]

### Καινοτομία (novelty)

Η έννοια της καινοτομίας (novelty) στα συστήματα προτάσεων σύμφωνα με τη βιβλιογραφία για ένα δεδομένο χρήστη, καθορίζεται από την αναλογία των άγνωστων αντικειμένων στη λίστα προτάσεων προς το σύνολο. Δηλαδή είναι ένα μέγεθος που εκφράζει το ποσοστό των προτάσεων που τα προϊόντα δεν είχαν στο παρελθόν κάποια αλληλεπίδραση απ’ τον δεδομένο χρήστη δηλαδή του ήταν άγνωστα [10]. Αναλυτικά ο τύπος περιγράφεται από την σχέση:

Όπου το R είναι το σύνολο των προτεινόμενων προϊόντων και είναι μια δυαδική συνάρτηση που επιστρέφει 1 εάν ο δεδομένος χρήστης γνωρίζει ήδη το προϊόν διαφορετικά επιστρέφει 0. Καθώς τα συστήματα προτάσεων αδυνατούν με μεγάλη σιγουριά να κρίνουν εάν ο χρήστης δεν αλληλοεπίδρασε ποτέ με προϊόν, υπάρχει μια προσέγγιση θεωρεί τα λιγότερο δημοφιλή προϊόντα ως καινοτόμες προτάσεις στους χρήστες [11].

Κεφάλαιο

# Υλοποίηση συστημάτων προτάσεων

## Συνεργατικό φιλτράρισμα με επίκεντρο τον χρήστη

### Προεπεξεργασία δεδομένων (pre-processing)

Στην παρούσα υλοποίηση σκοπός της προεπεξεργασίας δεδομένων είναι η δημιουργία ενός πίνακα χρηστών-αξιολογήσεων που θα περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων. Παρακάτω ακολουθεί η αναπαράσταση του πίνακα χρηστών-αξιολογήσεων

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ρούχο 1 | ρούχο 2 | … | ρούχο m-1 | ρούχο m |
| χρήστης 1 | 5 | 5 | … | 6 |  |
| χρήστης 2 | 4 | 9 | … | 3 | 2 |
| …. | ….. | … | … | … | … |
| χρήστης n | 7 | 9 | … | 8 |  |

Πίνακας - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων

### Εύρεση όμοιων χρηστών

Το αμέσως επόμενο βήμα αυτής της υλοποίησης είναι η εύρεση των πιο όμοιων χρηστών ως προς τον χρήστη που είναι το επίκεντρο. Για την συγκεκριμένη εφαρμογή θεωρούμε επίκεντρο τον χρήστη 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ρούχο 1 | … | ρούχο m-1 | ρούχο m | ομοιότητα |
| χρήστης 1 | 5 | … | 6 |  | 1 |
| χρήστης 2 | 4 | … | 3 | 2 | 0.67 |
| …. | ….. | … | … | … |  |
| χρήστης n | 7 | … | 8 |  | 0.789 |

Πίνακας - Αναπαράσταση του πίνακα χρηστών-αξιολογήσεων - Υπολογισμός ομοιότητας

### Στάθμιση σημασίας ομοιότητας (significance weighting)

Η αξιοπιστία του αποτελέσματος της συνάρτησης ομοιότητας πολλές φορές εξαρτάται από τον αριθμό των κοινών αξιολογήσεων για τους συγκρινόμενους χρήστες. Όταν δύο χρήστες έχουν μικρό αριθμό κοινών αξιολογήσεων τότε το αποτέλεσμα της ομοιότητας θα πρέπει να μειώνεται από έναν παράγοντα που θα τονίζει την σημασία των κοινών αξιολογήσεων. Αυτή η μέθοδος γνωστή και ως στάθμιση της ομοιότητας (significance weighting) [8] δεδομένου των κοινών αξιολογήσεων ορίζεται ένα threshold b τέτοιο ώστε:

Εξίσωση - Αναπαράσταση τύπου στάθμισης σημασίας ομοιότητας

### Κεντράρισμα μέση τιμής δεδομένων (mean-center data)

Αμέσως μετά την εύρεση των πιο όμοιων χρηστών μέσα από την στάθμιση της ομοιότητας είναι αναγκαίο το κεντράρισμα της μέσης τιμής των δεδομένων κάθε όμοιου χρήστη για την προσαρμογή του bias των αξιολογήσεων.

To bias των αξιολογήσεων προκαλείται όταν οι χρήστες τείνουν να δίνουν πάντα υψηλές ή πάντα χαμηλές βαθμολογίες σε όλα τα στοιχεία. [9]

Η νέα αξιολόγηση κάθε χρήστη προκύπτει από τον τύπο

Όπου η αξιολόγηση του χρήστη u στο ρούχο j και η μέση τιμή των αξιολογήσεων.

Παρακάτω ακολουθεί σχηματικά η αναπαράσταση του νεου πίνακα χρηστών-αξιολογήσεων με κεντραρισμένη μέση τιμή

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Ρούχο 1 | … | Ρούχο m | Μέση τιμή | Σταθμισμένη ομοιότητα |
| Χρήστης 1 |  | … |  |  | …. |
| …. | …. | … | …. | … | …. |
| Χρήστης n |  |  |  |  | …. |

Πίνακας - Αναπαράσταση πίνακα με κεντραρισμένη μέση τιμή αξιολογήσεων

### Συνάρτηση πρόβλεψης αξιολόγησης

Οι προβλέψεις των απροσδιόριστων αξιολογήσεων υπολογίζονται ως ο σταθμισμένος μέσος όρος των αποκλίσεων από τον μέσο όρο του κάθε χρήστη προσθέτοντάς τον στη μέση βαθμολογία του χρήστη που είναι στο επίκεντρο.

Εξίσωση - Τύπος της πρόβλεψης των απροσδιόριστων αξιολογήσεων

Παραπάνω παρουσιάζεται ο τύπος της πρόβλεψης των απροσδιόριστων αξιολογήσεων.

Αν u είναι ο χρήστης που έχουμε στο επίκεντρο και j το ρούχο που θέλουμε να εκτιμήσουμε τότε   είναι η εκτιμώμενη αξιολόγηση. Όπου η μέση τιμή των αξιολογήσεων του χρήστη u, το σύνολο των k πιο όμοιων χρηστών, η ομοιότητα του χρήστη u με τον χρήστη v και

## Συνεργατικό φιλτράρισμα με επίκεντρο τον προϊόν

### Προεπεξεργασία δεδομένων (pre-processing)

Στην παρούσα υλοποίηση σκοπός της προεπεξεργασίας δεδομένων είναι παρόμοιος με τις προηγούμενης δηλαδή η δημιουργία ενός πίνακα χρηστών-αξιολογήσεων που θα περιλαμβάνει τις αξιολογήσεις κάθε χρήστη για όλα τα ρούχα που περιέχονται στη βάση δεδομένων.

### Κεντράρισμα μέση τιμής δεδομένων (mean-center data)

Αμέσως μετά την δημιουργία του πίνακα χρηστών-αξιολογήσεων είναι αναγκαίο το κεντράρισμα της μέσης τιμής των δεδομένων του κάθε χρήστη για την προσαρμογή του bias των αξιολογήσεων.

Αντίθετα με την προηγούμενη εφαρμογή το κεντράρισμα της μέσης τιμής των αξιολογήσεων γίνεται πριν την εύρεση των πιο όμοιων ρούχων.

### Εύρεση όμοιων προϊόντων (ρούχων)

Έχοντας κεντράρει την μέση τιμή των αξιολογήσεων κάθε χρήστη στο μηδέν το επόμενο βήμα αυτής της υλοποίησης είναι η εύρεση των πιο όμοιων ρούχων. Ο υπολογισμός της ομοιότητας με τη χρήση του βασικού μέτρου συνημίτονο (regular cosine similarity) σε περίπτωση που βασίζεται σε προϊόντα (item-based) έχει ένα σημαντικό μειονέκτημα - η διαφορά στην κλίμακα βαθμολογίας μεταξύ διαφορετικών χρηστών δεν λαμβάνεται υπόψη. Η προσαρμοσμένη ομοιότητα συνημίτονου αντισταθμίζει αυτό το μειονέκτημα αφαιρώντας τον αντίστοιχο μέσος όρο των χρηστών από κάθε προσωπική του βαθμολογία. [10]

Θεωρούμε τις και της κεντραρισμένες ως προς την μέση τιμή αξιολογήσεις ενώ αντίστοιχα και τις θέσεις που έχουν υποβάλει οι χρήστες και σε προϊόντα.

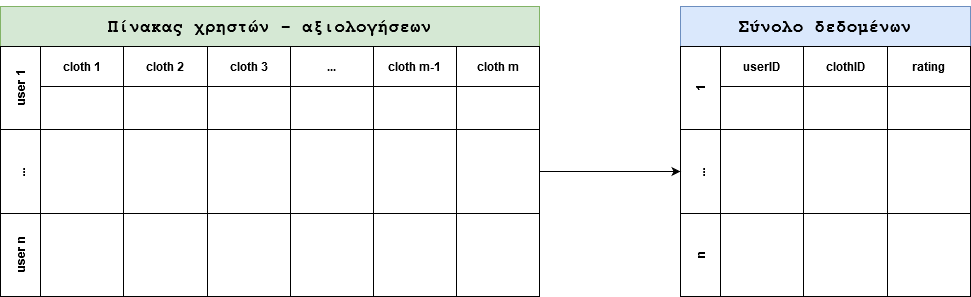
Εξίσωση - Αναπαράσταση τύπου προσαρμοσμένης ομοιότητας

### Συνάρτηση πρόβλεψης αξιολόγησης

## Συνεργατικό φιλτράρισμα με ανάλυση πίνακα σε ιδιάζουσες τιμές

### Προεπεξεργασία δεδομένων (pre-processing)

Στην παρούσα υλοποίηση σκοπός της προεπεξεργασίας είναι η μετατροπή του πίνακα αξιολογήσεων-χρηστών σε ένα σύνολο δεδομένων που για κάθε εγγραφή του, μια αξιολόγηση θα αντιστοιχεί σε ένα ρούχο και χρήστη. Η εικόνα που ακολουθεί αναπαριστά την μετατροπή του πίνακα χρηστών-αξιολογήσεων σε πίνακα συνόλου δεδομένων

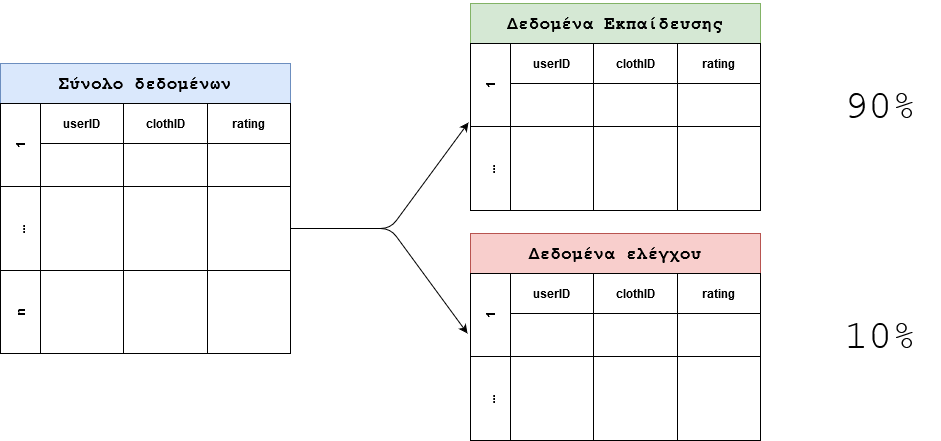


Εικόνα - Μετατροπή πίνακα χρηστών-αξιολογήσεων σε πίνακα συνόλου δεδομένων

Η μετατροπή αυτή έχει στόχο να κάνει ευκολότερο τον διαχωρισμό των δεδομένων στο επόμενο βήμα της υλοποίησης.

### Διαχωρισμός δεδομένων (train – test split)

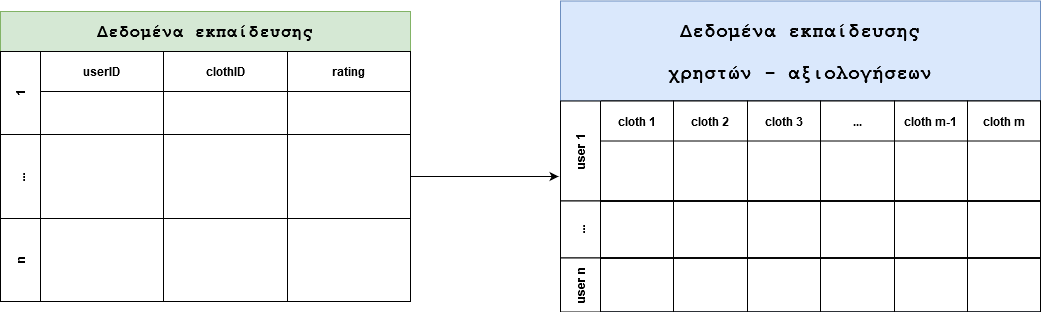
Για την εξακρίβωση των επιδόσεων πρόβλεψης της υλοποίησης οι γνωστές βαθμολογίες του χρήστη χωρίζονται σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Τα δεδομένα εκπαίδευσης αποτελούν το 90 % του συνόλου δεδομένων και τα δεδομένα ελέγχου το υπόλοιπο 10 %. Η εικόνα που ακολουθεί αναπαριστά των διαχωρισμό του συνόλου δεδομένων.



Εικόνα - Αναπαράσταση διαχωρισμού συνόλου δεδομένων

### Μετατροπή δεδομένων εκπαίδευσης σε νέο πίνακα χρηστών-αξιολογήσεων

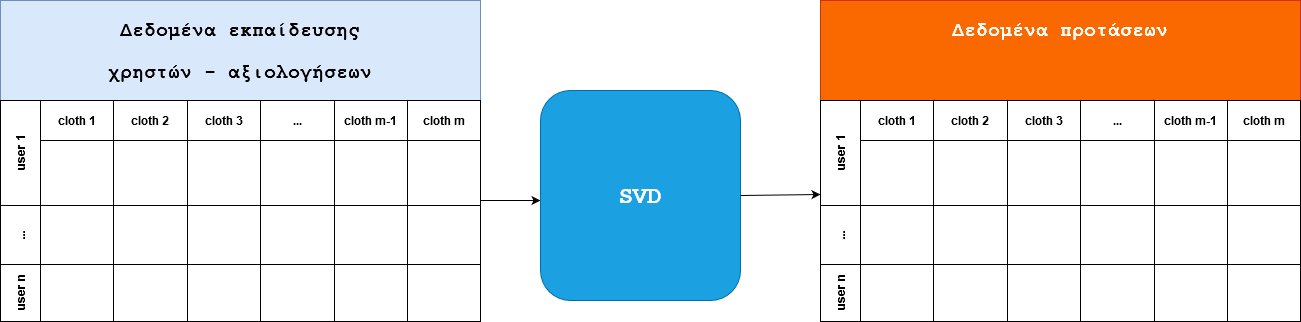
Μετά των διαχωρισμό των δεδομένων είναι αναγκαίο ο πίνακας των δεδομένων εκπαίδευσης να μετατραπεί στην οργάνωση δεδομένων που διέπει ο αρχικός πίνακας αξιολογήσεων ώστε στη συνέχεια να ακολουθήσει η ανάλυση πίνακα σε ιδιάζουσες τιμές.



Εικόνα - Μετατροπή δεδομένων εκπαίδευσης σε πίνακα χρηστών αξιολογήσεων

### Ανάλυση δεδομένων εκπαίδευσης σε πίνακα ιδιαζουσών τιμών

Μετά την δημιουργία του πίνακα δεδομένων εκπαίδευσης χρηστών-αξιολογήσεων η ανάλυση του πίνακα σε ιδιάζουσες τιμές δημιουργεί προτάσεις για τις απροσδιόριστες βαθμολογίες του συνόλου δεδομένων.



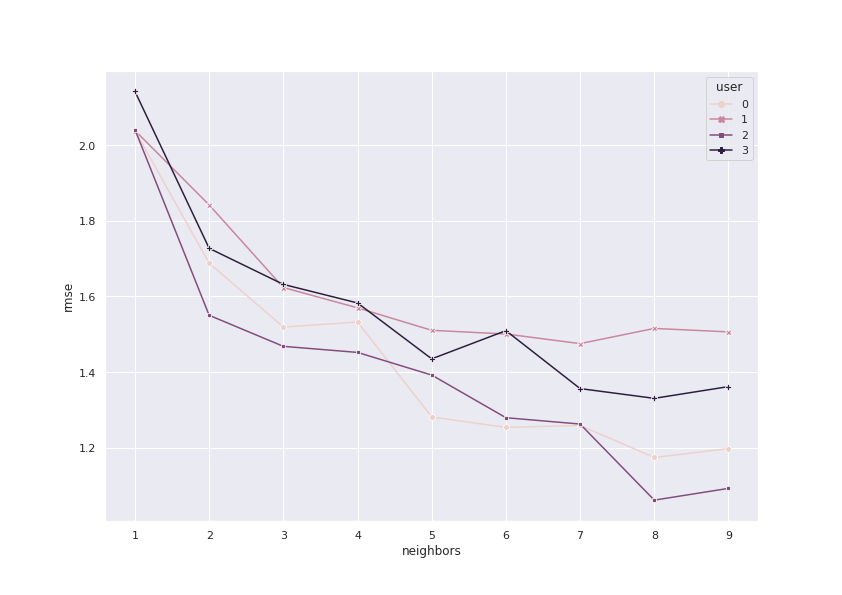
Κεφάλαιο

# Πειράματα και αποτελέσματα

## Υλοποίηση συνεργατικού φιλτραρίσματος με επίκεντρο τον χρήστη

Το σύνολο δεδομένων για την διεξαγωγή πειραμάτων αυτής της υλοποίησης αποτελείται από 9 χρήστες που έχουν αξιολογήσει ρούχα με αρνητικές, ουδέτερες ή θετικές βαθμολογίες. Ενδεικτικά για 4 από τους 9 συνολικούς χρήστες διεξάχθηκαν πειράματα για την εκτίμηση των βαθμολογιών τους. Καθώς δεν υπήρχε άλλος τρόπος αξιολόγησης των αποτελεσμάτων πέρα από την χρήση των ήδη γνωστών βαθμολογιών τα δεδομένα χωρίστηκαν σε δεδομένα εκπαίδευσης (train set) και δεδομένα δοκιμής (test set). Πιο συγκεκριμένα το 90 % των βαθμολογιών του χρήστη που ήταν στο επίκεντρο προτάσεων χρησιμοποιήθηκαν για την εύρεση των πιο όμοιων χρηστών ενώ το 10 % των ήδη γνωστών βαθμολογιών επανεκτιμήθηκε βάση την συνάρτηση πρόβλεψης ώστε να υπάρχει σύγκριση αποτελεσμάτων.

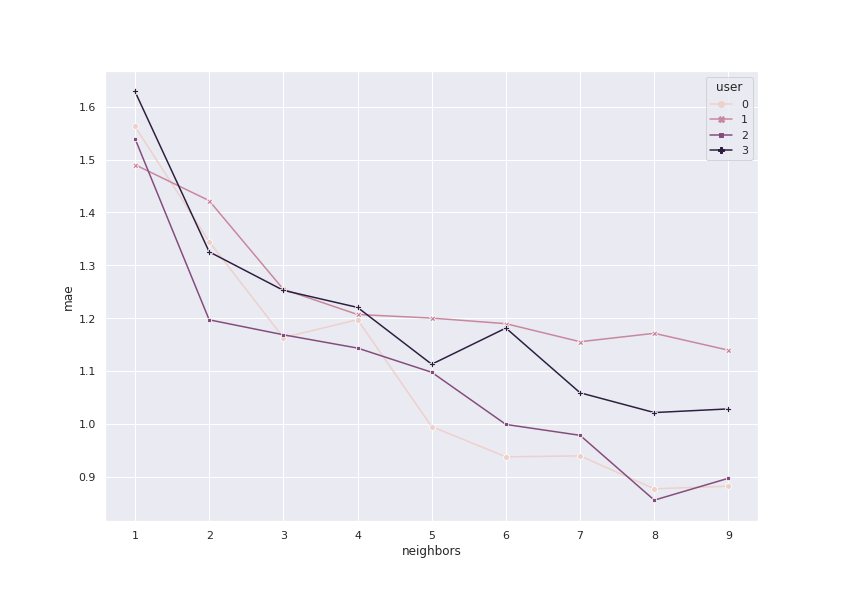
### Διάγραμμα Ρίζας Μέσου Τετραγωνικού Σφάλματος (RMSE) ανά γείτονα



Εικόνα - Διάγραμμα Ρίζας Μέσου Τετραγωνικού Σφάλματος (RMSE)

Είναι προφανές ότι η ρίζα μέσου τετραγωνικού σφάλματος παίρνει υψηλές τιμές για μικρό αριθμό γειτόνων ενώ μειώνεται σημαντικά όσο ο αριθμός των γειτόνων αυξάνει. Βέβαια εφόσον η υλοποίηση στηρίζεται στους όμοιους χρήστες είναι αναμενόμενο όσο για μεγάλους αριθμούς γειτόνων οι κοινές προτιμήσεις που δεν έχουν εξερευνηθεί ακόμη από τον χρήστη που είναι στο επίκεντρο να λιγοστεύουν.

### Διάγραμμα Μέσου Απόλυτου Σφάλματος (MAE) ανά γείτονα



Εικόνα - Διάγραμμα Μέσου Απόλυτου Σφάλματος (MAE)

Όμοια με την ρίζα μέσου τετραγωνικού σφάλματος, το μέσο απόλυτο σφάλμα παίρνει υψηλές τιμές για μικρό αριθμό γειτόνων και χαμηλές τιμές όσο ο αριθμός των γειτόνων αυξάνει. Βέβαια εφόσον η υλοποίηση στηρίζεται στους όμοιους χρήστες είναι αναμενόμενο όσο ο αριθμός των γειτόνων αυξάνει οι κοινές προτιμήσεις που δεν έχουν εξερευνηθεί ακόμη από τον χρήστη που είναι στο επίκεντρο να λιγοστεύουν.

### Πίνακας επιδόσεων εκτίμησης αξιολογήσεων

*Χρήστης 0*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Γείτονες* | *Accuracy* | *Recall* | *Precision* | *F1 Score* | *Ρούχα* |
| 1 | 0.63 | 0.73 | 0.72 | 0.73 | 243 |
| 2 | 0.71 | 0.9 | 0.75 | 0.82 | 160 |
| 3 | 0.74 | 0.9 | 0.79 | 0.84 | 119 |
| 4 | 0.77 | 0.96 | 0.78 | 0.86 | 99 |
| 5 | 0.86 | 0.97 | 0.88 | 0.92 | 79 |
| 6 | 0.88 | 0.97 | 0.9 | 0.93 | 73 |
| 7 | 0.88 | 1.0 | 0.88 | 0.94 | 69 |
| 8 | 0.94 | 1.0 | 0.93 | 0.97 | 62 |
| 9 | 0.93 | 1.0 | 0.93 | 0.96 | 57 |

*Χρήστης 1*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Γείτονες* | *Accuracy* | *Recall* | *Precision* | *F1 Score* | *Ρούχα* |
| 1 | 0.68 | 0.81 | 0.75 | 0.78 | 244 |
| 2 | 0.7 | 0.88 | 0.76 | 0.81 | 164 |
| 3 | 0.74 | 0.93 | 0.77 | 0.84 | 116 |
| 4 | 0.76 | 0.97 | 0.76 | 0.85 | 99 |
| 5 | 0.74 | 0.98 | 0.75 | 0.85 | 90 |
| 6 | 0.77 | 1.0 | 0.76 | 0.87 | 81 |
| 7 | 0.74 | 1.0 | 0.74 | 0.85 | 74 |
| 8 | 0.72 | 1.0 | 0.72 | 0.84 | 64 |
| 9 | 0.72 | 1.0 | 0.72 | 0.84 | 58 |

*Χρήστης 2*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Γείτονες* | *Accuracy* | *Recall* | *Precision* | *F1 Score* | *Ρούχα* |
| 1 | 0.66 | 0.75 | 0.78 | 0.76 | 253 |
| 2 | 0.75 | 0.85 | 0.84 | 0.84 | 159 |
| 3 | 0.78 | 0.9 | 0.84 | 0.87 | 122 |
| 4 | 0.82 | 0.93 | 0.86 | 0.89 | 100 |
| 5 | 0.81 | 0.93 | 0.85 | 0.89 | 88 |
| 6 | 0.87 | 0.97 | 0.89 | 0.93 | 76 |
| 7 | 0.87 | 0.95 | 0.9 | 0.93 | 68 |
| 8 | 0.89 | 0.96 | 0.91 | 0.94 | 61 |
| 9 | 0.89 | 0.98 | 0.91 | 0.94 | 55 |

*Χρήστης 4*

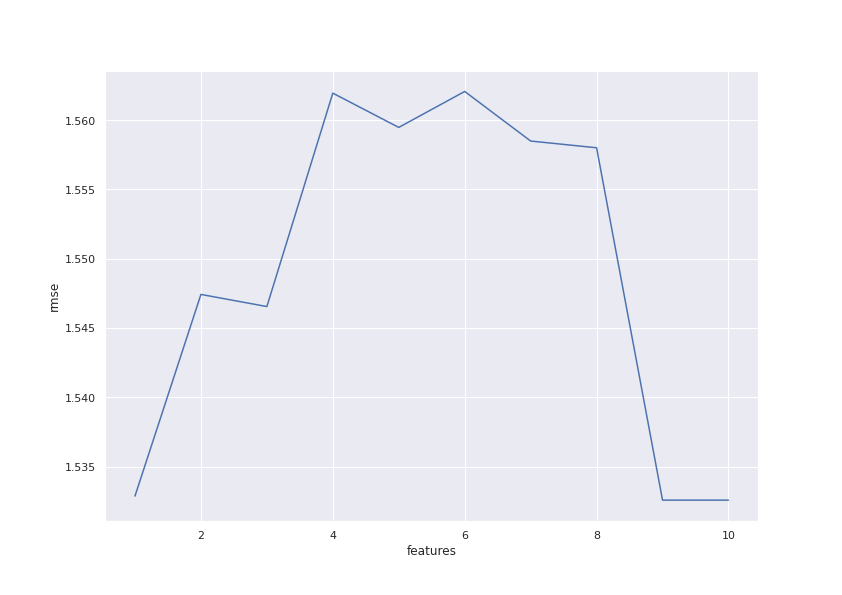
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Γείτονες* | *Accuracy* | *Recall* | *Precision* | *F1 Score* | *Ρούχα* |
| 1 | 0.61 | 0.7 | 0.75 | 0.72 | 243 |
| 2 | 0.71 | 0.8 | 0.81 | 0.81 | 164 |
| 3 | 0.75 | 0.86 | 0.82 | 0.84 | 132 |
| 4 | 0.79 | 0.94 | 0.81 | 0.87 | 109 |
| 5 | 0.84 | 0.96 | 0.87 | 0.91 | 90 |
| 6 | 0.84 | 0.98 | 0.84 | 0.91 | 79 |
| 7 | 0.88 | 1.0 | 0.88 | 0.94 | 69 |
| 8 | 0.89 | 1.0 | 0.89 | 0.94 | 65 |
| 9 | 0.88 | 1.0 | 0.88 | 0.93 | 58 |

## Υλοποίηση συνεργατικού φιλτραρίσματος με ανάλυση πίνακα ιδιαζουσών τιμών

Το σύνολο δεδομένων αποτελείται από 10 χρήστες ο καθένας από αυτούς έχει καταχωρημένες διαφορετικές αξιολογήσεις σε διαφορετικά προϊόντα (ρούχα). Για την αξιολόγηση των προβλέψεων έγινε διαχωρισμός του συνόλου δεδομένων σε δεδομένα εκπαίδευσης ( 90 % του συνόλου δεδομένων ) και δεδομένα ελέγχου ( υπόλοιπο 10 % ). Η παραγοντοποίηση των πινάκων εξυπηρετεί την μείωση διαστάσεων του πίνακα χρηστών και προϊόντων . Η μείωση της διάστασης γίνεται με όσο δυνατότερη λιγότερη απώλεια πληροφορίας. Αυτό έχει ως αποτέλεσμα οι νέες διαστάσεις παραγωντοποίησης πινάκων να αποτελούν τα σημαντικότερα features για κάθε χρήστη και κάθε προϊόν αντίστοιχα. Όσα περισσότερα features διατίθενται τόσο λιγότερη πληροφορία χάνεται βέβαια εις βάρος της επίδοσης και της ταχύτητας του αλγορίθμου.

### Διάγραμμα Ρίζας Μέσου Τετραγωνικού Σφάλματος (RMSE) ανά feature

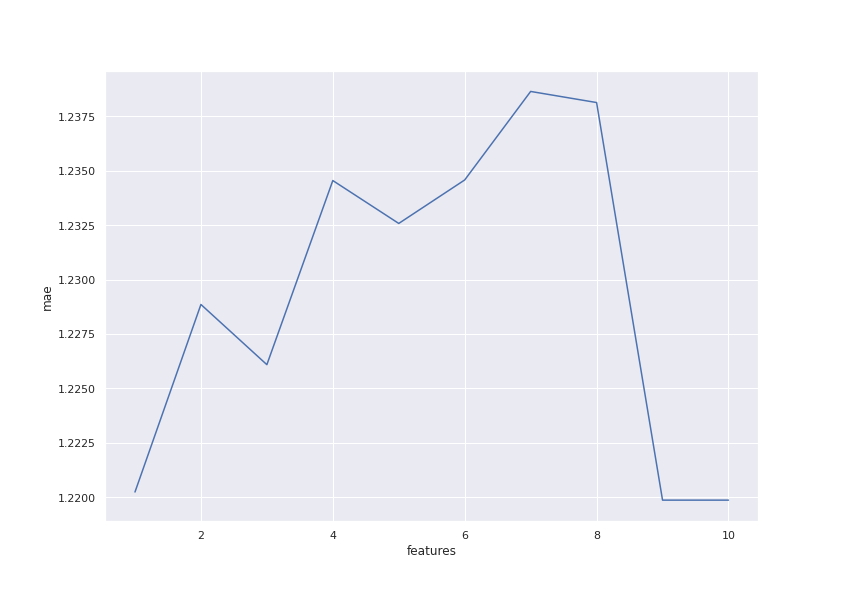
Βλέπουμε πως παρά την επιλογή πολλών ή λίγων features η τιμή του μέσου τετραγωνικού σφάλματος κυμαίνεται σε κλίματα απόκλισης πράγμα θετικό καθώς η μείωση των διαστάσεων για τα δεδομένα γίνεται χωρίς να υπάρχει ιδιαίτερη απώλεια πληροφορίας που αντικατοπτρίζεται μέσω της μικρής μεγέθυνσης της ρίζας τετραγωνικού σφάλματος.



Εικόνα - Ρίζα του Μέσου Τετραγωνικού Σφάλματος (RMSE) ανά feature

### Διάγραμμα Ρίζας Μέσου Απόλυτου Σφάλματος (MAE) ανά feature

Όμοια με την περίπτωση της ρίζας μέσου τετραγωνικού σφάλματος παρά την επιλογή πολλών ή λίγων features η τιμή του μέσου απόλυτου σφάλματος κυμαίνεται σε κλίματα απόκλισης πράγμα θετικό καθώς η μείωση των διαστάσεων για τα δεδομένα γίνεται χωρίς να υπάρχει αντίκτυπο στην απώλεια πληροφορίας η οποία αν συνέβαινε θα αντικατοπτρίζονταν μέσω της μεγέθυνσης του μέσου απόλυτου σφάλματος.



Εικόνα - Ρίζα του Μέσου Απόλυτου Σφάλματος (MAE) ανά feature

### Πίνακας επιδόσεων εκτίμησης αξιολογήσεων

Για την εκτίμηση των επιδόσεων στην ικανότητα εκτίμησης των αξιολογήσεων χρησιμοποιούμε το σύνολο δεδομένων ελέγχου. Για τυχαίους χρήστες εκπαιδεύουμε το σύστημα με το 90 % των αξιολογήσεων τους και ελέγχουμε την ικανότητα πρόβλεψης στο υπόλοιπο 10 % που αποτελεί το σύνολο δεδομένων ελέγχου. Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν από τα πειράματα στους χρήστες του συνόλου δεδομένων

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Χρήστες | Accuracy | Recall | Precision | F1 score | RMSE | MAE |
| 0 | 0.65 | 0.75 | 0.73 | 0.74 | 1.59 | 1.28 |
| 1 | 0.73 | 0.81 | 0.79 | 0.8 | 1.46 | 1.17 |
| 2 | 0.7 | 0.75 | 0.78 | 0.77 | 1.66 | 1.35 |
| 3 | 0.67 | 0.75 | 0.76 | 0.76 | 1.63 | 1.29 |
| 4 | 0.66 | 0.77 | 0.71 | 0.74 | 1.69 | 1.38 |
| 5 | 0.69 | 0.8 | 0.78 | 0.79 | 1.54 | 1.22 |
| 6 | 0.73 | 0.81 | 0.79 | 0.8 | 1.57 | 1.28 |
| 7 | 0.68 | 0.74 | 0.77 | 0.76 | 1.65 | 1.35 |

User based

Svd

Rf

hubrid

# Βιβλιογραφία

|  |  |
| --- | --- |
| [1] | Y. Koren, Factorization Meets the Neighborhood: a Multifaceted Collabora-tive Filtering Model, 2008. |
| [2] | Υ. Koren, R. Bell και C. Volinsky,, Matrix factorization Techniques forRecommender Systems, 2009. |
| [3] | A. Tulloch, «Facebook Research,» [Ηλεκτρονικό]. Available: https://research.fb.com/blog/2014/09/fast-randomized-svd/. |
| [4] | E. Jane και T. Bailey , Folding-up: A Hybrid Method for Updating the Partial Singular Value Decomposition in Latent Semantic Indexing, 2006. |
| [5] | T. Chai και . R. Draxler, «Root mean square error (RMSE) or mean absolute error (MAE)?,» p. 2, 2014. |
| [6] | J. Herlocker , . J. Konstan, L. Terveen και J. Riedl , Evaluating collaborative filtering recommender systems., ACM Transactions on Information Systems, 2004, pp. 5-53. |
| [7] | D. Jannach, M. Ge και C. Delgado-Battenfeld, Beyond accuracy: Evaluating recommender systems by coverage and serendipity, Barcelona, 2010. |
| [8] | G. Karypis και C. Desrosiers, A comprehensive survey of neighborhood-basedrecommendation methods, Springer, 2011, pp. 21-22. |
| [9] | X. Su και T. M. Khoshgoftaar, A Survey of Collaborative Filtering Techniques, USA: Hindawi Publishing Corporation, 2009, pp. 6-7. |
| [10] | B. Sarwar, G. Karypis, J. Konstan και J. Riedl, Item-based Collaborative Filtering RecommendationAlgorithms, Hong Kong, 2001, pp. 6-7. |
| [11] | C. C. Aggarwal, Recommender Systems - The Textbook, Springer, 2006. |