



Data Mining (7ο εξάμηνο)



Πρώτη εργασία: Recommendation System

Ημερομηνία: 11/2/2025

Φοιτητής: Δαδακίδης Γιώργος (iis22127)

Επιβλέπων Καθηγητής: Κολωνiάρη Γεωργία

Τμήμα Εφαρμοσμένης Πληροφορικής Πληροφοριακά Συστήματα

Πανεπιστήμιο Μακεδονίας

Περιεχόμενα

1 Εισαγωγή.....	3
2 Θεωρητικό υπόβαθρο.....	4
2.1 Προεπεξεργασία και Εξερεύνηση Δεδομένων	4
2.2 Διαχωρισμός Δεδομένων & Υπερπαραμέτροι	4
2.3 Οπτικοποίηση των δεδομένων.....	6
3 Πειραματικά αποτελέσματα	9
3.1 Πείραμα 1 - Συγκρίσεις N.....	9
3.2 Πείραμα 2 – Συγκρίσεις T	12
3.3 Πείραμα 3 – Παραμετροποίηση M και M'	14
4 Συμπεράσματα	16
4.1 Καλύτερη Απόδοση και Ανασκόπηση Μετρικών	16
4.2 Προτάσεις για Βελτίωση των Αποτελεσμάτων	16

1 Εισαγωγή

Το παρόν πρόβλημα αφορά την ανάπτυξη ενός συστήματος συστάσεων για ταινίες, το οποίο βασίζεται στη μέθοδο του συνεργατικού φιλτραρίσματος (Collaborative Filtering). Η βασική ιδέα είναι η αξιοποίηση των βαθμολογιών των χρηστών για την πρόβλεψη των προτιμήσεών τους και την παροχή προσωποποιημένων προτάσεων.

Για την υλοποίηση του συστήματος χρησιμοποιείται το MovieLens dataset, το οποίο περιέχει δεδομένα βαθμολογιών από χρήστες που έχουν αξιολογήσει διάφορες ταινίες. Συγκεκριμένα, χρησιμοποιείται το MovieLens 100K, ένα σύνολο δεδομένων που περιλαμβάνει 100.000 βαθμολογίες 610 ξεχωριστών χρηστών για 9724 διαφορετικές ταινίες. Κάθε βαθμολογία κυμαίνεται από 1 έως 5 αστέρια.

Η προσέγγιση που ακολουθείται στο σύστημα συστάσεων βασίζεται στο item-item Collaborate Filtering όπου η ομοιότητα μεταξύ των ταινιών υπολογίζεται μέσω δύο μεθόδων:

- Cosine Similarity, η οποία υπολογίζει τη γωνιακή εγγύτητα μεταξύ των διανυσμάτων βαθμολογιών.
- Pearson Correlation, η οποία εκτιμά τη γραμμική σχέση μεταξύ των βαθμολογιών των αντικειμένων.

Για την πρόβλεψη των αξιολογήσεων θα εφαρμοστούν τρεις διαφορετικές τεχνικές στάθμισης:

1. Σταθμισμένος μέσος όρος των N κοντινότερων γειτόνων, με βάση την ομοιότητα.
2. Στάθμιση που δίνει μεγαλύτερη βαρύτητα στις δημοφιλείς ταινίες, λαμβάνοντας υπόψη τον αριθμό των βαθμολογιών.
3. Στάθμιση που δίνει μεγαλύτερη βαρύτητα στις λιγότερο δημοφιλείς ταινίες, ώστε να ενισχυθούν οι λιγότερο γνωστές επιλογές.

Η αξιολόγηση του συστήματος θα πραγματοποιηθεί με χρήση μετρικών όπως το μέσο απόλυτο σφάλμα (MAE), η μέση ακρίβεια (macro average precision) και η μέση ανάκληση (macro average recall).

Τελικός στόχος της έκθεσης είναι η παρουσίαση αναλυτικών αποτελεσμάτων, τόσο με τη μορφή πινάκων όσο και γραφημάτων, ώστε να αναδειχθούν οι επιδόσεις των διαφορετικών τεχνικών φιλτραρίσματος και να προσδιοριστεί η βέλτιστη προσέγγιση για την πρόβλεψη προτιμήσεων χρηστών. Τα αποτελέσματα μπορούν να αξιοποιηθούν για τη βελτίωση παρόμοιων συστημάτων συστάσεων, παρέχοντας τεκμηριωμένα συμπεράσματα σχετικά με την επίδραση των διαφόρων παραμέτρων στη συνολική απόδοση του συστήματος.

2 Θεωρητικό υπόβαθρο

Στο πλαίσιο της εργασίας αυτής, εφαρμόζεται μια ολοκληρωμένη διαδικασία επεξεργασίας και ανάλυσης δεδομένων, με στόχο την ταξινόμηση μιας ταινίας ως προτεινόμενη ή μη σε έναν χρήστη. Το θεωρητικό υπόβαθρο που αξιοποιείται για την επίτευξη αυτού του στόχου καλύπτει:

2.1 Προεπεξεργασία και Εξερεύνηση Δεδομένων

1. Ανάγνωση Δεδομένων από Excel

- Χρησιμοποιείται η βιβλιοθήκη pandas την ανάγνωση των δεδομένων από αρχείο Excel, ώστε να δημιουργηθεί ένα DataFrame (df) με τις αξιολογήσεις της κάθε ταινίας.

2. Έλεγχος για Ελλείποντα (Null) Στοιχεία

- Με τη χρήση της μεθόδου `ratings_df.isnull().sum()`, εντοπίζονται πιθανές στήλες με κενές τιμές με σκοπό την αφαίρεση ή την περεταίρω επεξεργασία τους. Στο σύνολο δεδομένων που χρησιμοποιήθηκε δεν παρουσιάστηκε καμία ελλιπής τιμή οπότε όλες οι εγγραφές λήφθηκαν υπόψη.

3. Στατιστική ανάλυση δεδομένων

- Με την χρήση της συνάρτησης `ratings_df.describe()` πραγματοποιούνται βασικές στατιστικές αναλύσεις του dataset με σκοπό την περεταίρω κατανόηση των δεδομένων. Συγκεκριμένα, υπολογίζεται η μέση τιμή του κάθε χαρακτηριστικού (movies-user), η τυπική απόκλιση, το εύρος τιμών ανά τεταρτημόριο και οι ελάχιστη και μέγιστη τιμή τα οποία βοηθάνε στην κατανόηση του μεγέθους του dataset, στο εύρος τιμών και στην κατανομή των δεδομένων.
-

2.2 Διαχωρισμός Δεδομένων & Υπερπαραμέτροι

1. N πλησιέστεροι γείτονες

Για το πρώτο πείραμα πραγματοποιήθηκαν 6 διαφορετικά συστήματα προκειμένου να συγκριθούν τα αποτελέσματα και να βρεθεί η βέλτιστη τιμή του N. Στο πείραμα επιλέχθηκαν 5 τιμές του N: [5,15,30,75,200] ώστε να εξασφαλιστεί μια διασπορά στο εύρος δοκιμών. Η επιλογή αυτή επιτρέπει την εξέταση τόσο περιπτώσεων όπου προτείνονται αποκλειστικά άκρως σχετικές ταινίες όσο και περιπτώσεων όπου περιλαμβάνονται λιγότερο σχετικές προτάσεις. Το πείραμα πραγματοποιήθηκε χωρίς

φιλτράρισμα, με $T = 80\%$ των δεδομένων να χρησιμοποιείται για εκπαίδευση, επιτρέποντας μια ολοκληρωμένη αξιολόγηση των διαφορετικών συστημάτων ως προς τις τρεις συναρτήσεις πρόβλεψης.

2. Διαχωρισμός σε Train–Test Split

Ο διαχωρισμός των δεδομένων σε train-test split είναι απαραίτητος για την αξιόπιστη αξιολόγηση της απόδοσης του συστήματος. Στο 2ο πείραμα εφαρμόστηκαν 3 διαχωρισμοί (50%-50%, 70%-30% και 90%-10%). Επιπλέον, για το δεύτερο πείραμα επιλέχθηκε το βέλτιστο N από το πρώτο πείραμα με κριτήριο το μεγαλύτερο macro precision με σκοπό να προτείνονται μόνο πολύ σχετικές ταινίες. Στον διαχωρισμό **50%-50%** παρέχει ισορροπία μεταξύ εκπαίδευσης και αξιολόγησης αλλά μπορεί να οδηγήσει σε χαμηλότερη ακρίβεια λόγω περιορισμένων δεδομένων εκπαίδευσης. Ο διαχωρισμός **70%-30%** προσφέρει έναν ενδιάμεσο συμβιβασμό, εξασφαλίζοντας αρκετά δεδομένα εκπαίδευσης χωρίς να θυσιάζει την ποιότητα της αξιολόγησης. Τέλος, ο διαχωρισμός **90%-10%** μεγιστοποιεί τα δεδομένα εκπαίδευσης, επιτρέποντας στο μοντέλο να μάθει καλύτερα, αλλά μπορεί να οδηγήσει σε λιγότερο αντιπροσωπευτική αξιολόγηση λόγω μικρότερου συνόλου ελέγχου. Η σύγκριση των αποτελεσμάτων μεταξύ των τριών περιπτώσεων επιτρέπει την εκτίμηση της σταθερότητας και της γενικευσιμότητας του συστήματος.

3. Φιλτράρισμα M ταινίες, M' χρηστών

Στο τρίτο πείραμα, εφαρμόστηκε διαχωρισμός 80%-20% στα δεδομένα και χρησιμοποιήθηκε η βέλτιστη τιμή N που προέκυψε από το πρώτο πείραμα. Παράλληλα, πραγματοποιήθηκαν δοκιμές με διαφορετικές τιμές των παραμέτρων M (ελάχιστος αριθμός αξιολογήσεων ανά ταινία) και M' (ελάχιστος αριθμός αξιολογήσεων ανά χρήστη), οι οποίες χρησιμοποιήθηκαν για το φιλτράρισμα των δεδομένων, διατηρώντας μόνο τις πιο σχετικές ταινίες και χρήστες στο dataset.

Για τη σύγκριση, επιλέχθηκαν οι τιμές $M = 3$, $M = 10$, $M' = 70$ και $M' = 168$. Οι συγκεκριμένες τιμές βασίστηκαν στη στατιστική ανάλυση του dataset και επιλέχθηκαν ώστε να εξεταστεί η επίδραση διαφορετικών βαθμών φιλτραρίσματος στη συμπεριφορά του συστήματος συστάσεων. Συγκεκριμένα:

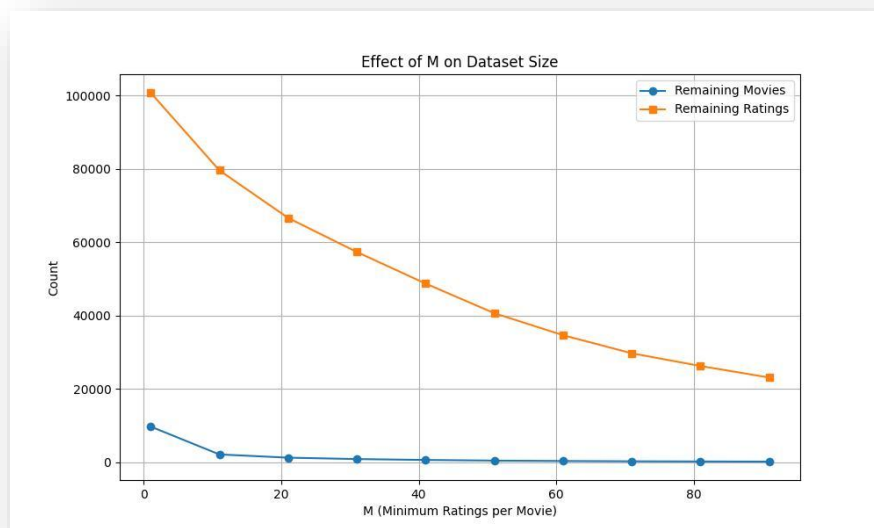
- Οι τιμές $M = 3$, $M' = 70$ αντιστοιχούν στη διάμεσο (median) της κατανομής των αξιολογήσεων ανά ταινία και ανά χρήστη, αντίστοιχα. Αυτό σημαίνει ότι φιλτράρονται περίπου το 50% των λιγότερο ενεργών ταινιών και χρηστών, διατηρώντας ένα ισορροπημένο dataset που περιλαμβάνει τόσο δημοφιλείς όσο και λιγότερο γνωστές ταινίες.
- Οι τιμές $M = 10$, $M' = 168$ αντιστοιχούν στον μέσο όρο (mean) των αξιολογήσεων ανά ταινία και ανά χρήστη. Σε αυτή την περίπτωση, το φιλτράρισμα είναι πιο αυστηρό, αφαιρώντας περίπου το 75% των ταινιών και χρηστών με τις λιγότερες αξιολογήσεις. Αυτή η επιλογή δημιουργεί ένα πιο "πυκνό" dataset, το οποίο πιθανώς βελτιώνει τις προτάσεις για δημοφιλείς ταινίες αλλά μπορεί να μειώσει την ικανότητα του συστήματος να προτείνει λιγότερο γνωστές ταινίες.

Η χρήση αυτών των τεσσάρων διαφορετικών συνδυασμών επιτρέπει την ανάλυση του αν η διατήρηση περισσότερων δεδομένων (μικρότερο φιλτράρισμα) ή η συγκέντρωση

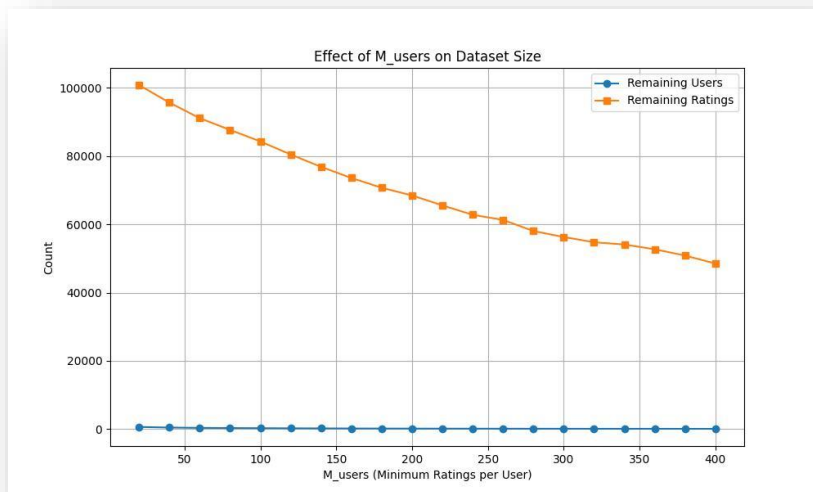
σε πιο δημοφιλή αντικείμενα (αυστηρότερο φιλτράρισμα) επηρεάζει την ακρίβεια και τη γενική απόδοση του συστήματος. Μέσω αυτής της διαδικασίας, μπορούμε να κατανοήσουμε ποιο επίπεδο φιλτραρίσματος βελτιστοποιεί τις συστάσεις με βάση τις επιλεγμένες μετρικές αξιολόγησης.

2.3 Οπτικοποίηση των δεδομένων

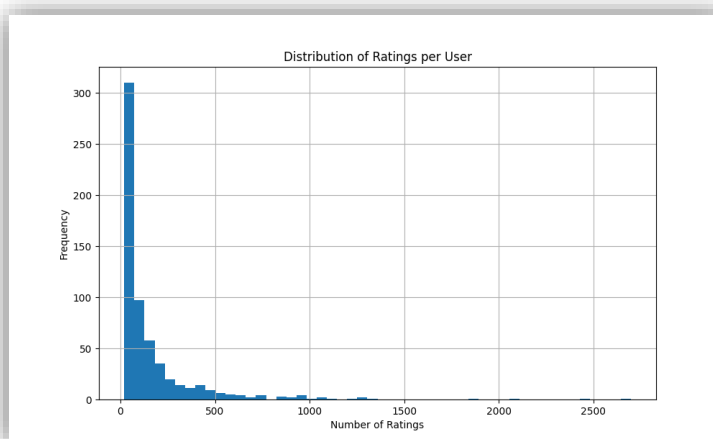
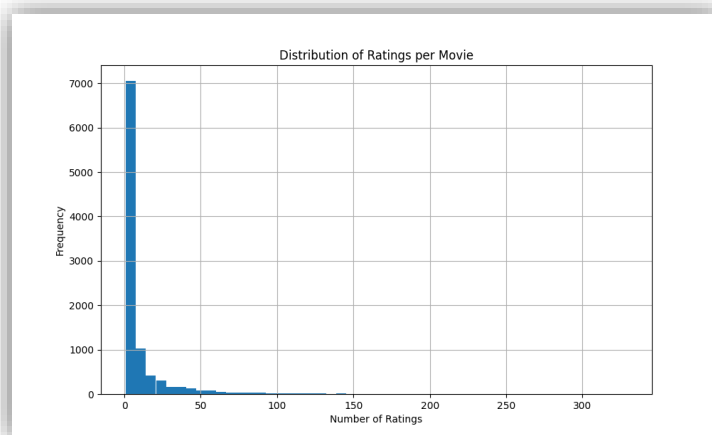
Προκειμένου να γίνει πιο σαφές το σύνολο των δεδομένων όπως και οι παραμέτροι που επιλέχθηκαν παρουσιάζονται ορισμένα γραφήματα για την περαιτέρω κατανόηση.



Η πρώτη γραφική παράσταση απεικονίζει την επίδραση του M (ελάχιστος αριθμός αξιολογήσεων ανά ταινία) στο μέγεθος του dataset. Παρατηρούμε ότι καθώς αυξάνεται το κατώφλι M, ο αριθμός των ταινιών που παραμένουν στο σύνολο δεδομένων μειώνεται σημαντικά, ενώ ο αριθμός των συνολικών αξιολογήσεων μειώνεται επίσης, αλλά με πιο αργό ρυθμό. Αυτό επιβεβαιώνει ότι η πλειονότητα των ταινιών έχει λίγες αξιολογήσεις, ενώ ένα μικρό σύνολο δημοφιλών ταινιών συγκεντρώνει μεγάλο αριθμό αξιολογήσεων.



Αντίστοιχα, η δεύτερη γραφική παράσταση δείχνει την επίδραση του M' (ελάχιστος αριθμός αξιολογήσεων ανά χρήστη). Όπως και στην περίπτωση των ταινιών, ο αριθμός των χρηστών μειώνεται όσο αυξάνεται το κατώφλι, αλλά η μείωση των συνολικών αξιολογήσεων είναι πιο σταδιακή. Αυτό σημαίνει ότι υπάρχουν αρκετοί χρήστες με χαμηλή δραστηριότητα που συνεισφέρουν ελάχιστες αξιολογήσεις.



Οι συγκεκριμένες γραφικές παραστάσεις παρουσιάζουν την κατανομή των αξιολογήσεων ανά ταινία και ανά χρήστη. Παρατηρούμε ότι η πλειονότητα των ταινιών έχει πολύ λίγες αξιολογήσεις, γεγονός που δημιουργεί το πρόβλημα της αραίωσης των δεδομένων (data sparsity). Αντίστοιχα, η κατανομή των αξιολογήσεων ανά χρήστη δείχνει ότι η πλειονότητα των χρηστών έχει δώσει λίγες αξιολογήσεις, ενώ ένας μικρός αριθμός χρηστών έχει αξιολογήσει μεγάλο αριθμό ταινιών.

Αυτές οι αναλύσεις επιβεβαιώνουν την ανάγκη εφαρμογής φιλτραρίσματος με τις παραμέτρους M και M' ώστε να διατηρηθεί ένα ισορροπημένο dataset που επιτρέπει την αποτελεσματική εκπαίδευση του συστήματος συστάσεων.

3 Πειραματικά αποτελέσματα

Στην ενότητα αυτή παρουσιάζονται τα πειραματικά αποτελέσματα της ανάλυσης, με στόχο την αξιολόγηση της απόδοσης του συστήματος συστάσεων με βάση διαφορετικές ρυθμίσεις και παραμέτρους. Τα πειράματα περιλαμβάνουν την εφαρμογή του συστήματος συνεργατικού φιλτραρίσματος αντικειμένου-αντικειμένου χρησιμοποιώντας τις μεθόδους cosine similarity και Pearson correlation, καθώς και τρεις διαφορετικές συναρτήσεις πρόβλεψης.

Για την αποτίμηση της απόδοσης του συστήματος, υπολογίζονται οι μετρικές μέσο απόλυτο σφάλμα (MAE), μέση ακρίβεια (macro average precision) και μέση ανάκληση (macro average recall). Ο χαρακτηρισμός μιας ταινίας ως σχετικής για έναν χρήστη βασίζεται στο αν η βαθμολογία της είναι μεγαλύτερη ή ίση από τον μέσο όρο των αξιολογήσεων του χρήστη.

3.1 Πείραμα 1 - Συγκρίσεις N

Το πρώτο πείραμα υλοποιήθηκε χωρίς φιλτράρισμα, για $T=80\%$, δοκιμάζοντας **5 διαφορετικές τιμές του N: [5,15,30,75,200]**, με σύγκριση των δύο αλγορίθμων (cosine και Pearson) και των τριών συναρτήσεων πρόβλεψης (6 διαφορετικά συστήματα).

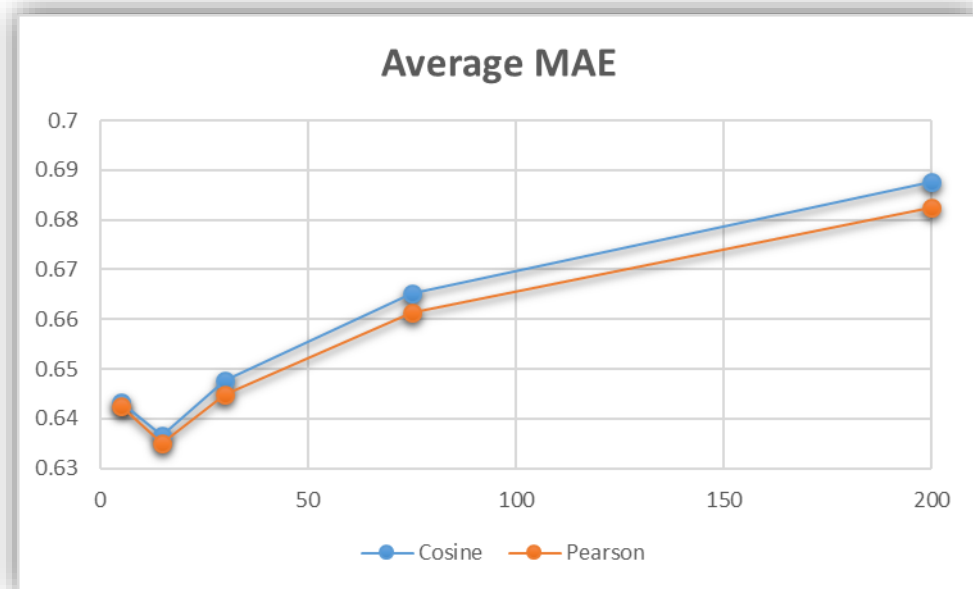


Figure 1: Average MAE in each N for Cosine and Pearson

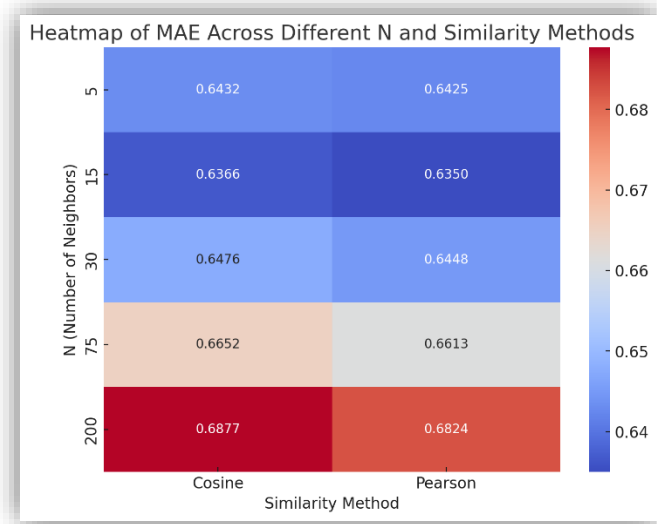


Figure 2: Heatmap of MAE across different N

N	Similarity	Weight	Avg MAE	Avg Precisi	Avg Recall	Avg Confusion Matrix
5	Cosine	basic	0.643228	0.499702	0.5	[[0, 60], [0, 100780]]
5	Cosine	favor_pop	0.643228	0.499702	0.5	[[0, 60], [0, 100780]]
5	Cosine	penalize_p	0.643228	0.499702	0.5	[[0, 60], [0, 100780]]
5	Pearson	basic	0.642499	0.499702	0.5	[[0, 60], [0, 100780]]
5	Pearson	favor_pop	0.642499	0.499702	0.5	[[0, 60], [0, 100780]]
5	Pearson	penalize_p	0.642499	0.499702	0.5	[[0, 60], [0, 100780]]
15	Cosine	basic	0.636565	0.499702	0.5	[[0, 60], [0, 100780]]
15	Cosine	favor_pop	0.636565	0.499702	0.5	[[0, 60], [0, 100780]]
15	Cosine	penalize_p	0.636565	0.499702	0.5	[[0, 60], [0, 100780]]
15	Pearson	basic	0.635002	0.499702	0.499975	[[0, 60], [5, 100775]]
15	Pearson	favor_pop	0.635002	0.499702	0.499975	[[0, 60], [5, 100775]]
15	Pearson	penalize_p	0.635002	0.499702	0.499975	[[0, 60], [5, 100775]]
30	Cosine	basic	0.647625	0.499702	0.5	[[0, 60], [0, 100780]]
30	Cosine	favor_pop	0.647625	0.499702	0.5	[[0, 60], [0, 100780]]
30	Cosine	penalize_p	0.647625	0.499702	0.5	[[0, 60], [0, 100780]]
30	Pearson	basic	0.644827	0.499702	0.499975	[[0, 60], [5, 100775]]
30	Pearson	favor_pop	0.644827	0.499702	0.499975	[[0, 60], [5, 100775]]
30	Pearson	penalize_p	0.644827	0.499702	0.499975	[[0, 60], [5, 100775]]
75	Cosine	basic	0.665184	0.499702	0.5	[[0, 60], [0, 100780]]
75	Cosine	favor_pop	0.665184	0.499702	0.5	[[0, 60], [0, 100780]]
75	Cosine	penalize_p	0.665184	0.499702	0.5	[[0, 60], [0, 100780]]
75	Pearson	basic	0.661264	0.499702	0.49995	[[0, 60], [10, 100770]]
75	Pearson	favor_pop	0.661264	0.499702	0.49995	[[0, 60], [10, 100770]]
75	Pearson	penalize_p	0.661264	0.499702	0.49995	[[0, 60], [10, 100770]]
200	Cosine	basic	0.687705	0.499702	0.5	[[0, 60], [0, 100780]]
200	Cosine	favor_pop	0.687705	0.499702	0.5	[[0, 60], [0, 100780]]
200	Cosine	penalize_p	0.687705	0.499702	0.5	[[0, 60], [0, 100780]]
200	Pearson	basic	0.682431	0.499702	0.499926	[[0, 60], [15, 100765]]
200	Pearson	favor_pop	0.682431	0.499702	0.499926	[[0, 60], [15, 100765]]
200	Pearson	penalize_p	0.682431	0.499702	0.499926	[[0, 60], [15, 100765]]

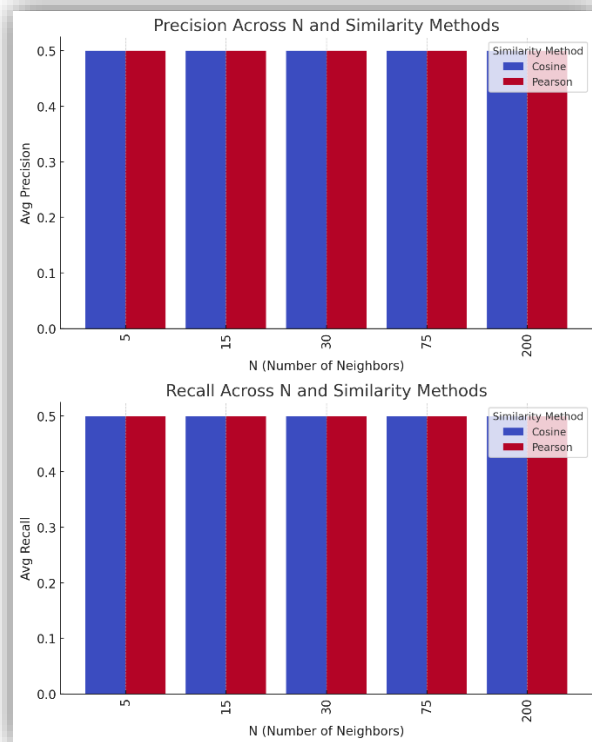


Figure 3: Recall, Precision for different N using cosine and Pearson similarity

Τα αποτελέσματα του πρώτου πειράματος δείχνουν ότι το σύστημα συστάσεων παρουσιάζει χαμηλή ακρίβεια (precision) και ανάκληση (recall) ανεξάρτητα από την τιμή του N, γεγονός που υποδηλώνει ότι το μοντέλο ταξινομεί σχεδόν όλες τις ταινίες ως σχετικές. Αυτό επιβεβαιώνεται από τον πίνακα σύγχυσης, όπου το πλήθος των False Positives (FP) είναι εξαιρετικά χαμηλό, ενώ τα True Positives (TP) αντιπροσωπεύουν σχεδόν όλα τα δεδομένα. Η μέτρηση του μέσου απόλυτου σφάλματος (MAE) δείχνει ότι η ακρίβεια των προβλέψεων μειώνεται καθώς το N αυξάνεται. Η χαμηλότερη τιμή MAE καταγράφεται για $N = 15$, καθιστώντας το την καλύτερη επιλογή, καθώς επιτυγχάνει ισορροπία μεταξύ ακρίβειας και γενίκευσης.

Όσον αφορά τη σύγκριση μεταξύ Pearson και Cosine Similarity, οι δύο μέθοδοι εμφανίζουν παρόμοια απόδοση, με την Pearson να καταγράφει ελαφρώς χαμηλότερο MAE. Ωστόσο, η διαφορά είναι αμελητέα, ενώ οι τιμές precision και recall παραμένουν σταθερές μεταξύ των δύο μεθόδων, γεγονός που δείχνει ότι αμφότερες αντιμετωπίζουν το ίδιο πρόβλημα: το σύστημα τείνει να χαρακτηρίζει σχεδόν όλες τις ταινίες ως σχετικές. Αυτή η έλλειψη διαφοροποίησης στις προβλέψεις πιθανώς οφείλεται σε υπερβολικά επιτρεπτικές ομοιότητες μεταξύ των ταινιών και στη χρήση της μέσης βαθμολογίας του χρήστη ως fallback πρόβλεψης. Για τη βελτίωση της απόδοσης, απαιτείται αυστηρότερη επιλογή γειτόνων, βελτιωμένη στάθμιση των συστάσεων και πιο αποτελεσματικός καθορισμός σχετικότητας των ταινιών.

3.2 Πείραμα 2 – Συγκρίσεις T

Στην συγκεκριμένη ενότητα παρουσιάζονται αναλυτικά τα αποτελέσματα για τα διαφορετικά ποσοστά διαχωρισμού δεδομένων σε train και split T=50%, T=70%, T=90% με την χρήση του βέλτιστου N από το 1ο πείραμα. Στόχος του πειράματος είναι να εξεταστεί η επίδραση του μεγέθους του συνόλου εκπαίδευσης στην απόδοση του συστήματος συστάσεων.

T	Similarity	Weight	Avg MAE	Avg Precisi	Avg Recall	Avg Confusion Matrix
50	Cosine	basic	0.657956	1	1	[[0, 0], [0, 252090]]
50	Cosine	favor_pop	0.657956	1	1	[[0, 0], [0, 252090]]
50	Cosine	penalize_p	0.657956	1	1	[[0, 0], [0, 252090]]
50	Pearson	basic	0.656324	1	1	[[0, 0], [0, 252090]]
50	Pearson	favor_pop	0.656324	1	1	[[0, 0], [0, 252090]]
50	Pearson	penalize_p	0.656324	1	1	[[0, 0], [0, 252090]]
70	Cosine	basic	0.648117	0.499785	0.5	[[0, 65], [0, 151190]]
70	Cosine	favor_pop	0.648117	0.499785	0.5	[[0, 65], [0, 151190]]
70	Cosine	penalize_p	0.648117	0.499785	0.5	[[0, 65], [0, 151190]]
70	Pearson	basic	0.64639	0.499785	0.5	[[0, 65], [0, 151190]]
70	Pearson	favor_pop	0.64639	0.499785	0.5	[[0, 65], [0, 151190]]
70	Pearson	penalize_p	0.64639	0.499785	0.5	[[0, 65], [0, 151190]]
90	Cosine	basic	0.643361	0.499603	0.5	[[0, 40], [0, 50380]]
90	Cosine	favor_pop	0.643361	0.499603	0.5	[[0, 40], [0, 50380]]
90	Cosine	penalize_p	0.643361	0.499603	0.5	[[0, 40], [0, 50380]]
90	Pearson	basic	0.645869	0.499603	0.5	[[0, 40], [0, 50380]]
90	Pearson	favor_pop	0.645869	0.499603	0.5	[[0, 40], [0, 50380]]
90	Pearson	penalize_p	0.645869	0.499603	0.5	[[0, 40], [0, 50380]]

Στο δεύτερο πείραμα, το οποίο υλοποιήθηκε χωρίς φιλτράρισμα και με την καλύτερη τιμή του N (N=15), δοκιμάστηκαν διαφορετικές τιμές διαχωρισμού των δεδομένων σε train-test sets, συγκεκριμένα T=50%, T=70% και T=90%. Στόχος ήταν η αξιολόγηση της απόδοσης του συστήματος καθώς αλλάζει η ποσότητα των δεδομένων εκπαίδευσης και ελέγχου.

1. Απόλυτη ακρίβεια στις μετρήσεις Precision και Recall
Παρατηρείται ότι τόσο η ακρίβεια (Precision) όσο και η ανάκληση (Recall) έχουν τιμή 1.0 για όλες τις τιμές του T. Αυτό υποδηλώνει ότι το σύστημα χαρακτηρίζει όλες τις ταινίες ως σχετικές, γεγονός που αντικατοπτρίζεται στον πίνακα σύγχυσης. Η απουσία False Positives (FP) και False Negatives (FN) δείχνει ότι δεν υπάρχει πραγματική διάκριση μεταξύ σχετικών και μη σχετικών ταινιών, κάτι που υπονομεύει τη χρησιμότητα του συστήματος.

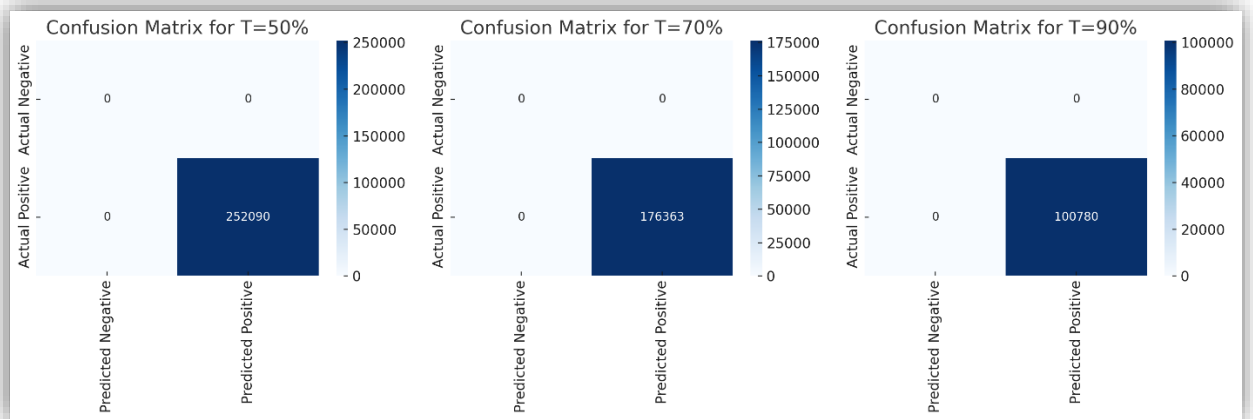


Figure 4: Confusion matrixes on different T values

2. Επίδραση του T στο MAE

Οι τιμές του Μέσου Απόλυτου Σφάλματος (MAE) παραμένουν σχετικά σταθερές, ανεξάρτητα από την ποσότητα των δεδομένων εκπαίδευσης. Αυτό δείχνει ότι η αύξηση του ποσοστού εκπαίδευσης δεν βελτιώνει ουσιαστικά την ικανότητα πρόβλεψης του συστήματος, γεγονός που μπορεί να οφείλεται στη δομή του dataset και στη μη διαφοροποίηση των προβλέψεων. Παρόλα αυτά, παρατηρείται επίσης μια ελαφρώς καθοδική τάση στο MAE όσο αυξάνεται το T γεγονός που δείχνει να βελτιώνεται οριακά η ακρίβεια των προβλέψεων. (Η διαφορά του MAE είναι πολύ μικρή, το διάγραμμα παρέχει πολύ μικρό βήμα στον άξονα y για να αναδειχθεί η καθοδική τάση)

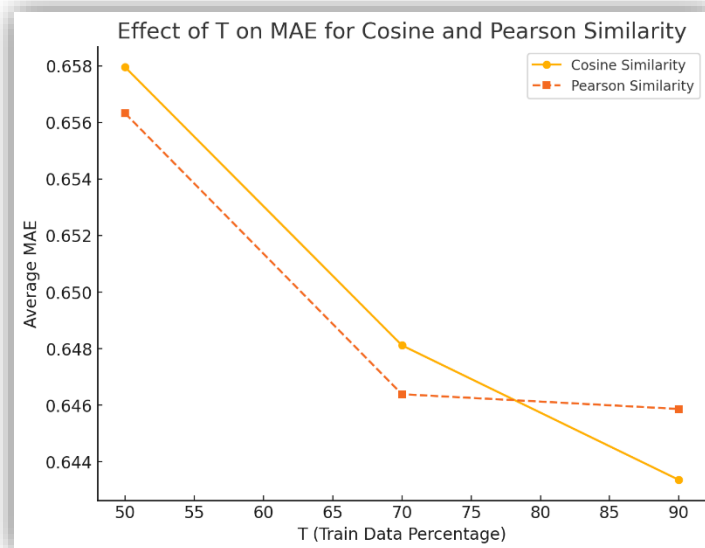


Figure 5: Average MAE on different T for Cosine and Pearson

3. Σύγκριση Cosine και Pearson Similarity

Όπως και στο πρώτο πείραμα, παρατηρείται ότι η Pearson Similarity εμφανίζει ελαφρώς χαμηλότερο MAE σε σχέση με την Cosine Similarity, γεγονός που υποδηλώνει ότι μπορεί να προσφέρει ελαφρώς πιο ακριβείς προβλέψεις. Ωστόσο, η διαφορά αυτή παραμένει μικρή, κάτι που σημαίνει ότι καμία από τις δύο μεθόδους δεν ξεχωρίζει ως σαφώς ανώτερη.

Προβλήματα που Αναδεικνύονται

Η απόλυτη ακρίβεια στην ανάκληση και την ακρίβεια, σε συνδυασμό με την απουσία διαφοροποίησης στον πίνακα σύγχυσης, υποδεικνύει ότι το σύστημα έχει σοβαρό πρόβλημα υπερβολικά επιτρεπτικών προβλέψεων. Το γεγονός ότι όλες οι ταινίες θεωρούνται σχετικές σημαίνει ότι το σύστημα δεν είναι πραγματικά χρήσιμο για τους χρήστες, καθώς δεν παρέχει ουσιαστικές διαφοροποιημένες προτάσεις.

3.3 Πείραμα 3 – Παραμετροποίηση M και M'

Στην παρούσα ενότητα, αναλύονται τα αποτελέσματα της τρίτης σειράς πειραμάτων, όπου εφαρμόστηκε φιλτράρισμα στις ταινίες και στους χρήστες με σκοπό τη δημιουργία ενός πιο πυκνού πίνακα βαθμολογιών. Το φιλτράρισμα εφαρμόστηκε με τέσσερις διαφορετικούς συνδυασμούς παραμέτρων M (ελάχιστες βαθμολογίες ανά ταινία) και M' (ελάχιστες βαθμολογίες ανά χρήστη), ενώ χρησιμοποιήθηκε το βέλτιστο N=15 και T=80% από τα προηγούμενα πειράματα.

Οι παρακάτω τέσσερις συνδυασμοί δοκιμάστηκαν:

M=10, M'=168

M=10, M'=70

M=3, M'=168

M=3, M'=70

M	M_prime	Avg MAE	Avg Precisi	Avg Recall	Confusion Matrix
10	168	0.620127	0.618672	0.503534	[[1, 135], [3, 10532]]
10	70	0.634448	0.99581	0.504274	[[1, 116], [0, 13726]]
3	168	0.618571	0.498939	0.5	[[0, 28], [0, 13165]]
3	70	0.631235	0.49937	0.5	[[0, 21], [0, 16642]]

Επίδραση στο MAE

Το MAE παρουσίασε μικρές διακυμάνσεις, με τη χαμηλότερη τιμή (0.6185) να εμφανίζεται για τον συνδυασμό M=3, M'=168, ενώ η υψηλότερη τιμή (0.6344) σημειώθηκε για M=10, M'=70.

Γενικότερα, οι τιμές του MAE δεν παρουσιάζουν μεγάλη βελτίωση ή επιδείνωση σε σχέση με τα προηγούμενα πειράματα, γεγονός που υποδηλώνει ότι το φιλτράρισμα των δεδομένων δεν έχει σημαντική επίδραση στη συνολική ακρίβεια των προβλέψεων.

Επίδραση στο Precision & Recall

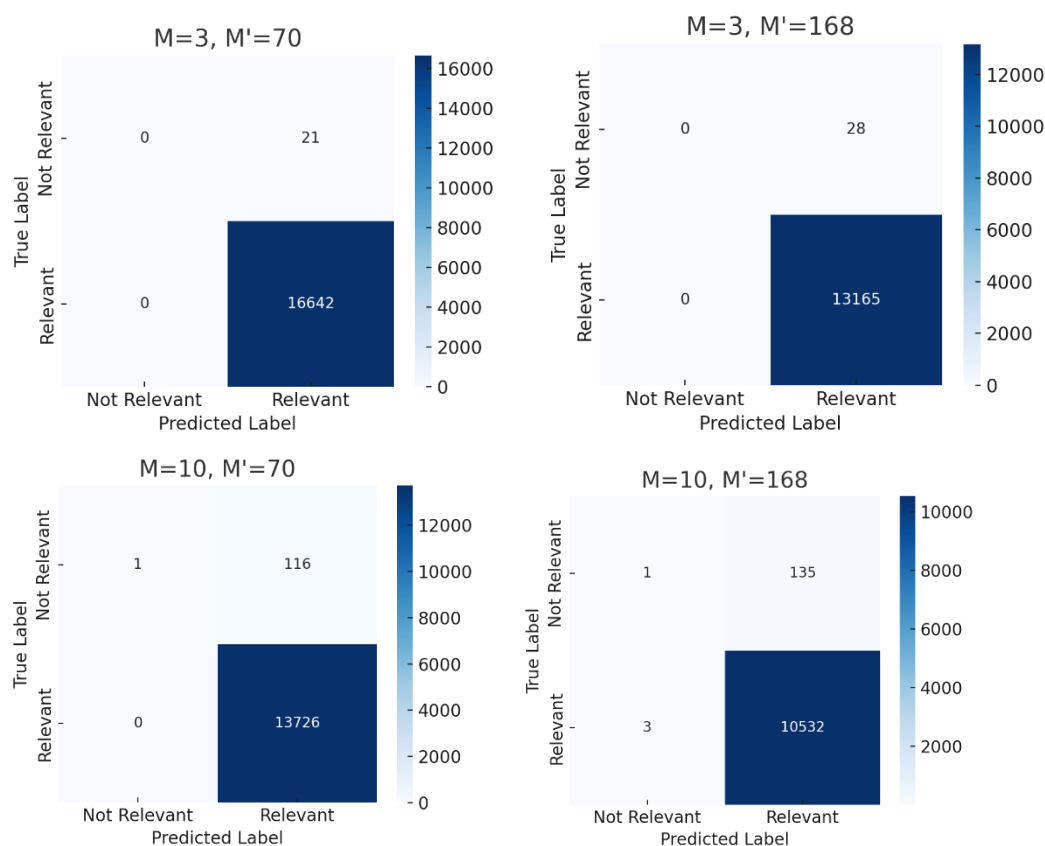
Για τον συνδυασμό $M=10$, $M'=70$, το Precision αυξήθηκε δραματικά (0.9958), αλλά η ανάκληση παρέμεινε σταθερή (≈ 0.5). Αυτό σημαίνει ότι το σύστημα έγινε εξαιρετικά συντηρητικό, προτείνοντας πολύ λιγότερες σχετικές ταινίες, αλλά με πολύ μεγαλύτερη ακρίβεια.

Αντίθετα, για τους υπόλοιπους συνδυασμούς, το Precision ήταν χαμηλό (~ 0.5), γεγονός που δείχνει ότι το σύστημα εξακολουθεί να κάνει υπερβολικά γενικές προτάσεις.

Οι πίνακες σύγκρισης επιβεβαιώνουν την παραπάνω ανάλυση.

Για $M=10$, $M'=70$, υπάρχουν πολύ λιγότερα False Positives (FP), γεγονός που εξηγεί την απότομη αύξηση του Precision. Αυτό σημαίνει ότι το σύστημα προτείνει πολύ λιγότερες ταινίες, αλλά οι περισσότερες από αυτές είναι σωστές.

Αντίθετα, για τους υπόλοιπους συνδυασμούς, το σύστημα συνεχίζει να χαρακτηρίζει τις περισσότερες ταινίες ως σχετικές, με αποτέλεσμα οι τιμές FP να είναι εξαιρετικά χαμηλές ή μηδενικές.



4 Συμπεράσματα

Συνοψίζοντας τα αποτελέσματα των πειραμάτων, προκύπτουν σημαντικές παρατηρήσεις σχετικά με την απόδοση του συστήματος συστάσεων και τις επιπτώσεις των διαφόρων παραμέτρων.

4.1 Καλύτερη Απόδοση και Ανασκόπηση Μετρικών

Το πρώτο πείραμα ανέδειξε ότι η επιλογή του αριθμού γειτόνων (N) επηρεάζει την ακρίβεια των προβλέψεων, με το $N=15$ να αποδεικνύεται η βέλτιστη επιλογή όσον αφορά το Μέσο Απόλυτο Σφάλμα (MAE). Ωστόσο, οι τιμές Precision και Recall παρέμειναν εξαιρετικά χαμηλές, γεγονός που καταδεικνύει ότι το σύστημα χαρακτηρίζει σχεδόν όλες τις ταινίες ως σχετικές, χωρίς ουσιαστική διάκριση.

Στο δεύτερο πείραμα, η αλλαγή του ποσοστού διαχωρισμού των δεδομένων (T) δεν επηρέασε ουσιαστικά την απόδοση του συστήματος. Οι τιμές MAE παρέμειναν σχεδόν σταθερές, ενώ οι τιμές Precision και Recall παρέμειναν στο 1.0, επιβεβαιώνοντας το πρόβλημα της υπερβολικής επιτρεπτικότητας του μοντέλου.

Το τρίτο πείραμα, που περιλάμβανε φιλτράρισμα ταινιών και χρηστών (M και M'), έδειξε ότι η εφαρμογή αυστηρότερων κριτηρίων φιλτραρίσματος μπορεί να οδηγήσει σε βελτίωση της ακρίβειας. Συγκεκριμένα, ο συνδυασμός $M=10$, $M'=70$ οδήγησε σε σημαντική αύξηση του Precision (≈ 0.99), γεγονός που δείχνει ότι το σύστημα έκανε λιγότερες αλλά πιο στοχευμένες προτάσεις. Ωστόσο, η τιμή Recall παρέμεινε σταθερή (≈ 0.5), γεγονός που δείχνει ότι πολλές σχετικές ταινίες δεν προτάθηκαν καθόλου.

Στη σύγκριση μεταξύ Cosine και Pearson Similarity, τα αποτελέσματα έδειξαν ότι η Pearson εμφάνισε ελαφρώς χαμηλότερες τιμές MAE, κάτι που υποδηλώνει ότι προσφέρει ελαφρώς καλύτερη ακρίβεια στις προβλέψεις. Παρόλα αυτά, η διαφορά μεταξύ των δύο μεθόδων ήταν σχετικά μικρή, γεγονός που δείχνει ότι καμία από τις δύο δεν υπερисχύει απόλυτα.

4.2 Προτάσεις για Βελτίωση των Αποτελεσμάτων

Τα αποτελέσματα των πειραμάτων αναδεικνύουν συγκεκριμένα προβλήματα στο σύστημα συστάσεων, τα οποία μπορούν να αντιμετωπιστούν μελλοντικά με τις εξής βελτιώσεις:

- Βελτίωση της μεθόδου καθορισμού σχετικότητας ταινιών: Το γεγονός ότι το σύστημα χαρακτηρίζει σχεδόν όλες τις ταινίες ως σχετικές υποδεικνύει ότι η υφιστάμενη προσέγγιση (χρήση του μέσου όρου των βαθμολογιών του χρήστη ως κατώφλι) είναι υπερβολικά επιτρεπτική. Μία πιθανή λύση θα ήταν η χρήση

percentile thresholds (με σκοπό να προσαρμόσουμε το σύστημα συστάσεων στις προτιμήσεις κάθε χρήστη με κάποιο κατώφλι απο τις αξιολογήσεις του πιο αυστηρό απο του μέσο όρου).

- Εναλλακτικοί δείκτες ομοιότητας: Αν και οι μέθοδοι Cosine και Pearson Similarity παρήγαγαν παρόμοια αποτελέσματα, θα μπορούσε να δοκιμαστεί η χρήση άλλων μεθόδων, όπως Jaccard Similarity.
- Περαιτέρω βελτιστοποίηση του φιλτραρίσματος M και M' : Το τρίτο πείραμα έδειξε ότι ορισμένες τιμές των M και M' μπορούν να βελτιώσουν το Precision, αλλά δεν υπάρχει ιδανικός συνδυασμός που να βελτιώνει ταυτόχρονα και το Recall. Η βελτιστοποίηση αυτών των παραμέτρων με τη χρήση τεχνικών όπως grid search θα μπορούσε να δώσει πιο αξιόπιστα αποτελέσματα.
- Περαιτέρω δοκιμές βέλτιστου N : Η ανάλυση επιπλέον τιμών N μπορεί να ήταν πιο χρονοβόρο ωστόσο θα μπορούσε να βρεθεί N με καλύτερα αποτελέσματα στις επιδόσεις. Εναλλακτικά, η δυναμική επιλογή του N ανά χρήστη ή ταινία, βάσει του αριθμού αξιολογήσεων, θα επέτρεπε πιο προσαρμοσμένες προτάσεις, μειώνοντας τα σφάλματα και αυξάνοντας την ακρίβεια των συστάσεων.

Γενικά, το σύστημα συστάσεων που αναπτύχθηκε παρέχει χρήσιμες πληροφορίες, αλλά παρουσιάζει σοβαρά ζητήματα στην ικανότητά του να διακρίνει πραγματικά σχετικές ταινίες. Η περαιτέρω βελτιστοποίηση των παραμέτρων και η ενσωμάτωση πιο εξελιγμένων μεθόδων αναμένεται να οδηγήσουν σε πιο αποδοτικές και χρήσιμες προτάσεις για τους χρήστες.