Πανεπιστήμιο Μακεδονίας

Τμήμα Εφαρμοσμένης Πληροφορικής

Μάθημα: Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Προγραμματιστική Εργασία 2

Θέμα: Σύστημα Συστάσεων

Παρασκευή Ξανθοπούλου (it1490)

Καθηγητές: Γεώργιος Ευαγγελίδης, Γεωργία Κολωνιάρη

Ακαδημαϊκό Έτος: 2023-2024

Θεσσαλονίκη, Φεβρουάριος 2024

Περιεχόμενα

1. Εισαγωγή

Στο πλαίσιο αυτής της εργασίας, αναπτύχθηκε ένα σύστημα συστάσεων συνεργατικού φιλτραρίσματος αντικειμένου-αντικειμένου για ταινίες, με χρήση του συντελεστή Pearson ως μέτρο ομοιότητας, χρησιμοποιώντας το σύνολο δεδομένων του MovieLens. Το σύστημα εκτελεί προβλέψεις για τις βαθμολογίες ταινιών για συγκεκριμένους χρήστες χρησιμοποιώντας διάφορες μεθόδους βαθμολόγησης.

2. Κώδικας

Ο κώδικας υλοποιήθηκε σε Python και περιλαμβάνει τις ακόλουθες λειτουργίες:

- Προεπεξεργασία του συνόλου δεδομένων, συγκεκριμένα δειγματοληψία και αφαίρεση χρηστών και ταινιών με λιγότερες από 12 βαθμολογίες.
- Υλοποίηση τεσσάρων διαφορετικών μεθόδων πρόβλεψης βαθμολογίας ταινιών.
- Εκτέλεση πειραμάτων εκτίμησης απόδοσης του συστήματος με χρήση 5-fold Cross Validation.

2.1. Προεπεξεργασία Δεδομένων

Για τη μείωση του μεγέθους του συνόλου δεδομένων, αρχικά πραγματοποιήθηκε δειγματοληψία, όπου αφαιρέθηκε το 60% των εγγραφών και στη συνέχεια απομακρύνθηκαν οι εγγραφές που αντιστοιχούσαν σε χρήστες ή ταινίες με λιγότερες από 12 βαθμολογίες. Πριν την προεπεξεργασία, το αρχείο είχε μέγεθος 2.426 KB, ενώ μετά την εφαρμογή της δειγματοληψίας και του φιλτραρίσματος, το μέγεθος του αρχείου μειώθηκε σημαντικά στα 523 KB. Συνολικά, το ποσοστό μείωσης είναι περίπου 78,51%.

2.2. Μέτρο ομοιότητας

Το μέτρο ομοιότητας που χρησιμοποιήθηκε ονομάζεται "Adjusted Cosine Similarity". Αυτό το μέτρο αντιπροσωπεύει έναν τρόπο για τη μέτρηση της ομοιότητας μεταξύ δύο ταινιών. Είναι παρόμοιο με τον συντελεστή Pearson, με τη διαφορά ότι εφαρμόζεται στις ταινίες (στήλες) ενός πίνακα οφέλους (user-item matrix), αντί για τους χρήστες (γραμμές).

Ο αλγόριθμος υπολογισμού περιλαμβάνει δύο βήματα:

- Αρχικά, από κάθε γραμμή (διάνυσμα με τις βαθμολογίες ενός χρήστη για κάθε ταινία, στις ταινίες που δεν υπάρχει βαθμολογία βάζουμε 0) αφαιρούμε τον μέσο όρο της γραμμής.
- Στη συνέχεια, υπολογίζουμε το cosine similarity για τη στήλη (διάνυσμα με τις βαθμολογίες ενός αντικειμένου από κάθε χρήστη).

2.3. Συναρτήσεις Πρόβλεψης Βαθμού

Οι τέσσερις συναρτήσεις πρόβλεψης βαθμού περιγράφονται ως εξής:

1. **Σταθμισμένος Μέσος Όρος**: Υπολογίζει τη βαθμολογία μιας ταινίας με βάση τον σταθμισμένο μέσο όρο των βαθμολογιών των γειτόνων της. Έστω N το σύνολο των ταινιών που είναι πιο όμοιες με την m και έχουν βαθμολογηθεί από τον χρήστη u. Τότε:

$$r_{mu} = \frac{\sum_{n \epsilon N} sim(m,n) * r_{nu}}{\sum_{n \epsilon N} sim(m,n)}$$

2. **Σταθμισμένος Μέσος Όρος με Προσαρμογή της Μέσης Βαθμολογίας του Χρήστη και Αφαίρεση του Bias των Γειτόνων**: Υπολογίζει τη βαθμολογία μιας ταινίας με βάση τον σταθμισμένο μέσο όρο των βαθμολογιών των γειτόνων της, προσαρμόζοντας την μέση βαθμολογία του χρήστη και αφαιρώντας το bias των γειτόνων. Έστω N το σύνολο των ταινιών που είναι πιο όμοιες με την m και έχουν βαθμολογηθεί από τον χρήστη u. Τότε:

$$r_{mu} = \overline{r_u} + \frac{\sum_{n \epsilon N} sim(m,n) * (r_{nu} - \overline{r_n})}{\sum_{n \epsilon N} sim(m,n)}$$

3. **Σταθμισμένος Μέσος Όρος με Βάση το Πλήθος των Κοινών Χρηστών**: Υπολογίζει τη βαθμολογία μιας ταινίας με βάση τον σταθμισμένο μέσο όρο των βαθμολογιών των γειτόνων της. Ωστόσο, το βάρος κάθε γείτονα υπολογίζεται ανάλογα με το πλήθος των κοινών χρηστών που έχουν βαθμολογήσει τις δύο ταινίες, δηλαδή υποθέτουμε ότι το πλήθος των κοινών χρηστών μεταξύ δύο ταινιών είναι κατα κάποιο τρόπο ένα μέτρο ομοιότητας. Με άλλα λόγια, όσο περισσότεροι οι κοινοί χρήστες μεταξύ δύο αντικειμένων, τόσο μεγαλύτερη η βαρύτητα του ενός στη βαθμολογία του άλλου. Αυτό βοηθάει στην αύξηση της σημασίας των περισσότερο "αξιόπιστων" γειτόνων στον υπολογισμό της τελικής βαθμολογίας. Έστω N το σύνολο των ταινιών που είναι πιο όμοιες με την m και έχουν βαθμολογηθεί από τον χρήστη u. Τότε:

$$r_{mu} = \frac{\sum_{n \epsilon N} sim(m,n)*W(m,n)*r_{nu}}{\sum_{n \epsilon N} sim(m,n)*W(m,n)}$$

$$W(m,n) = \frac{commonUsers(m,n) + 1}{maxCommonUsers(m) + 1}$$

Μετά τον υπολογισμό του βάρους W(m,n), γίνεται προσαρμογή του στο διάστημα [0.9,1], έτσι ώστε να διατηρηθεί μέσα σε μια κατάλληλη περιοχή.

4. **Σταθμισμένος Μέσος Όρος με Βάση τη Διακύμανση των Βαθμολογιών της Κάθε Ταινίας**: Υπολογίζει τη βαθμολογία μιας ταινίας με βάση τον σταθμισμένο μέσο όρο των βαθμολογιών των γειτόνων της. Ωστόσο, σε αυτήν την περίπτωση, η στάθμιση βασίζεται στη διακύμανση των βαθμολογιών της κάθε ταινίας. Άρα, όσο μεγαλύτερη είναι η διακύμανση που υπάρχει στη βαθμολογία μιας ταινίας, τόσο μεγαλύτερο το βάρος της ταινίας στον σταθμισμένο μέσο όρο. Αυτό σημαίνει ότι οι ταινίες με μεγαλύτερη διακύμανση στις βαθμολογίες τους θα έχουν μεγαλύτερη επίδραση στον υπολογισμό της τελικής βαθμολογίας. Έστω N το σύνολο των ταινιών που είναι πιο όμοιες με την m και έχουν βαθμολογηθεί από τον χρήστη u. Τότε:

$$r_{mu} = \frac{\sum_{n \epsilon N} sim(m,n)*W(n)*r_{nu}}{\sum_{n \epsilon N} sim(m,n)*W(n)}$$

$$W(n) = log(variance(n))$$

Μετά τον υπολογισμό του βάρους W(n), γίνεται προσαρμογή του στο διάστημα [0.8,1], έτσι ώστε να διατηρηθεί μέσα σε μια κατάλληλη περιοχή.

3. Πειράματα

3.1. Αποτελέσματα

Τα αποτελέσματα των πειραμάτων παρουσιάζονται στον πίνακα **3.1.1**. Οι πίνακες σύγχυσης που περιλαμβάνονται, είναι της παρακάτω μορφής:

TN	FP
FN	TP

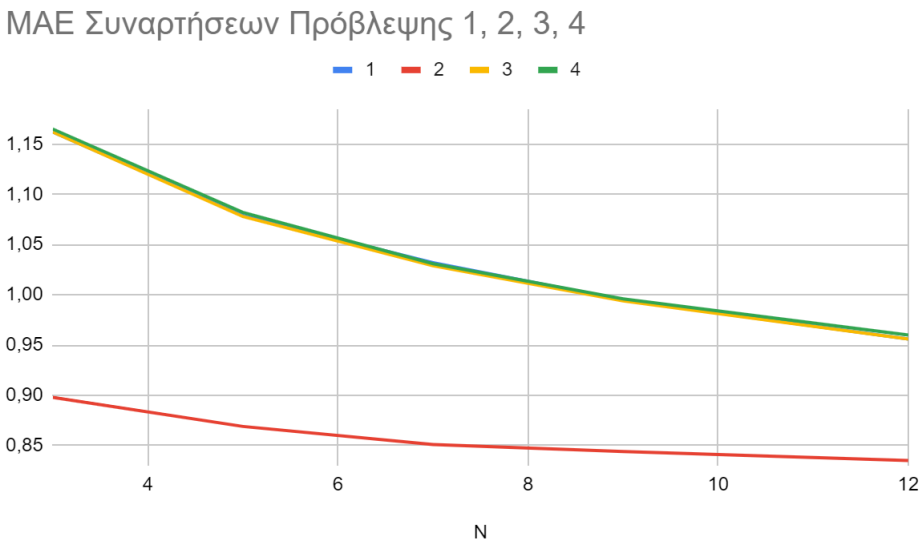
Prediction Method	N	MAE	Precision	Recall	Confusion Matrix	
1	3	1,162	0,350	0,425	2.463	722
					9.878	8.865
1	5	1,079	0,374	0,437	2.415	770
					9.252	9.491
1	7	1,032	0,393	0,446	2.388	797
					8.697	10.046
1	9	0,995	0,405	0,452	2.345	840
					8.483	10.260
1	12	0,956	0,421	0,460	2.316	869
					8.064	10.679
2	3	0,898	0,584	0,550	1.793	1.392
					5.040	13.703
2	5	0,869	0,593	0,558	1.795	1.390
					4.703	14.040
2	7	0,851	0,609	0,569	1.738	1.447
					4.435	14.308
2	9	0,844	0,608	0,570	1.747	1.438
					4.389	14.354
2	12	0,835	0,620	0,577	1.694	1.491
					4.281	14.462
3	3	1,162	0,349	0,425	2.475	710
					9.815	8.928
3	5	1,078	0,373	0,436	2.436	749
					9.164	9.579
3	7	1,029	0,384	0,442	2.400	785
					8.951	9.792
3	9	0,994	0,409	0,454	2.336	849
					8.414	10.329
3	12	0,956	0,430	0,463	2.294	891
					7.891	10.852
4	3	1,165	0,340	0,421	2.492	693
					10.069	8.674

4	5	1,082	0,365	0,433	2.448	737
					9.398	9.345
4	7	1,031	0,385	0,442	2.410	775
					8.857	9.886
4	9	0,996	0,405	0,452	2.348	837
					8.476	10.267
4	12	0,963	0,432	0,465	2.267	918
					7.934	10.809

3.1.1. Πίνακας αποτελεσμάτων

3.2. Γραφικές Παραστάσεις

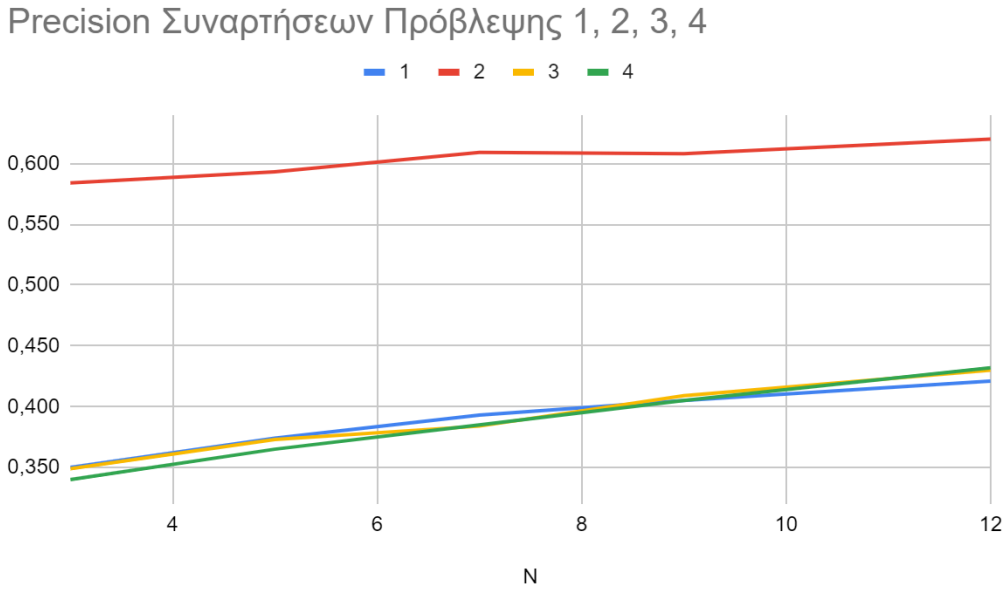
Στο Διάγραμμα **3.2.1** παρατηρούμε το MAE (Μέσο Απόλυτο Σφάλμα) των τεσσάρων συναρτήσεων πρόβλεψης βαθμολογιών.



3.2.1. Διάγραμμα Σύγκρισης MAE (Μέσο Απόλυτο Σφάλμα)

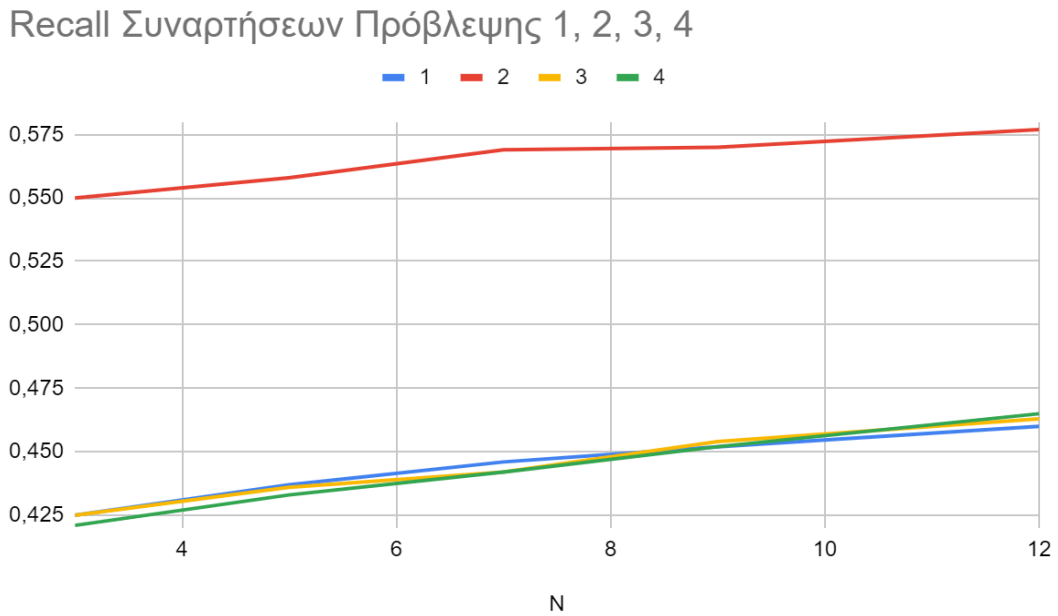
Όπως φαίνεται στο διάγραμμα, η συνάρτηση πρόβλεψης 2 ξεχωρίζει, με το μικρότερο MAE σε σύγκριση με τις άλλες τρεις συναρτήσεις. Αυτό υποδεικνύει ότι οι βαθμολογίες που προβλέπει είναι πιο κοντά στις πραγματικές βαθμολογίες, σε σύγκριση με τις προβλέψεις των υπολοίπων συναρτήσεων. Μια ενδιαφέρουσα παρατήρηση είναι ότι, όσο αυξάνεται το N (πλήθος γειτόνων), τόσο μειώνεται το MAE για κάθε συνάρτηση. Αυτό υποδεικνύει ότι η αύξηση του αριθμού των γειτόνων συμβάλλει στη βελτίωση των προβλέψεων. Επιπλέον, παρατηρούμε ότι οι συναρτήσεις πρόβλεψης 1, 3 και 4 φαίνεται να έχουν παρόμοιο MAE. Ωστόσο, στην πραγματικότητα, το MAE της συνάρτησης 3 είναι ελαφρώς μικρότερο από τις υπόλοιπες.

Στο Διάγραμμα **3.2.2** παρουσιάζονται οι ακρίβειες των τεσσάρων συναρτήσεων πρόβλεψης.



3.2.2. Διάγραμμα Σύγκρισης Precision (Ακρίβεια)

Παρατηρούμε ότι η συνάρτηση πρόβλεψης 2 διακρίνεται για τη μεγαλύτερη ακρίβεια. Επίσης, όπως είδαμε και στο διάγραμμα των Μέσων Σφαλμάτων, η αύξηση του N (πλήθος γειτόνων), έχει θετική επίδραση και στην ακρίβεια για όλες τις συναρτήσεις πρόβλεψης. Επιπλέον, οι συναρτήσεις πρόβλεψης 1, 3 και 4 παρουσιάζουν πολύ κοντά αποτελέσματα σε ακρίβεια. Ωστόσο, για μεγαλύτερες τιμές του N, οι συναρτήσεις 3 και 4 εμφανίζουν ελαφρώς μεγαλύτερη ακρίβεια.



3.2.3. Διάγραμμα Σύγκρισης Recall (Ανάκληση)

Στο Διάγραμμα **3.2.3**, μπορούμε να δούμε ότι η συνάρτηση 2 εμφανίζει την υψηλότερη ανάκληση. Επιπλέον, όπως παρατηρήσαμε και στα προηγούμενα δύο διαγράμματα, η αύξηση του N, αυξάνει και την ανάκληση για όλες τις συναρτήσεις. Τέλος, οι συναρτήσεις 1,3 και 4, παρουσιάζουν παρόμοιους βαθμούς ανάκλησης, όμως, όπως παρατηρήθηκε και στο διάγραμμα της ακρίβειας, για μεγαλύτερες τιμές N, οι βαθμοί ανάκλησης των συναρτήσεων 3 και 4 υπερβαίνουν τον βαθμό της 1.

3.3. Συμπεράσματα

Μετά από την ανάλυση των διαγραμμάτων σύγκρισης MAE, Precision και Recall, προκύπτουν τα εξής συμπεράσματα:

- Η συνάρτηση 2, η οποία παράγει την πρόβλεψη με υπολογισμό του σταθμισμένου μέσου όρου με προσαρμογή της μέσης βαθμολογίας του χρήστη και αφαίρεση του bias των γειτόνων, αποφέρει τα υψηλότερα μέτρα αποτίμησης, κάτι που την καθιστά την προτιμότερη επιλογή ανάμεσα στις τέσσερις συναρτήσεις πρόβλεψης.
- Η αύξηση του πλήθους των ομοιότερων γειτόνων που λαμβάνονται υπόψη για τον υπολογισμό της βαθμολογίας μιας ταινίας, αποτελεί σημαντικό παράγοντα που επηρεάζει όλα τα μέτρα αποτίμησης, με αποτέλεσμα να βελτιώνεται η ευστοχία της πρόβλεψης, ανεξαρτήτως της συνάρτησης που χρησιμοποιείται.
- Οι συναρτήσεις 3 και 4, αν και παρουσιάζουν παρόμοιες επιδόσεις με τη συνάρτηση 1, παρατηρήθηκε ότι σε ορισμένες περιπτώσεις υπερέχουν ελαφρώς της συνάρτησης 1 σε επίδοση. Αυτό καταδεικνύει ότι το πλήθος των κοινών χρηστών και η διακύμανση της βαθμολογίας μιας ταινίας μπορούν να παρέχουν επιπλέον πληροφόρηση για την πρόβλεψη μιας βαθμολογίας από ότι μόνο η ομοιότητα, και η χρήση τους για τη στάθμιση μπορεί να οδηγήσει σε πιο εύστοχες προβλέψεις. Καθοριστικό σημείο αποτελεί ο ορισμός κατάλληλων συναρτήσεων στάθμισης που βασίζονται στο πλήθος κοινών χρηστών και τη διακύμανση των βαθμολογιών, καθώς και η ρύθμισή τους με στόχο τη βελτιστοποίηση των επιδόσεων.

Τέλος, μέσω πειραματικών δοκιμών και αναλύσεων, είναι εφικτό να εντοπιστούν οι κατάλληλες συναρτήσεις στάθμισης που βασίζονται στο πλήθος κοινών χρηστών και τη διακύμανση των βαθμολογιών. Αυτή η διαδικασία μπορεί να συμβάλει στη βελτιστοποίηση της πρόβλεψης βαθμολογιών ταινιών και να ενισχύσει την ακρίβεια και την απόδοση του συστήματος προτάσεων.