ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Μια προσέγγιση μεταφοράς γνώσης σε συστήματα συστάσεων συνεργατικού φιλτραρίσματος με μη-επικαλυπτόμενες οντότητες

Από

Γεώργιο Σιράγα

ΜΕ2148

Υποβάλλεται
για την εκπλήρωση των προϋποθέσεων λήψης
Μεταπτυχιακού Διπλώματος
στην ειδίκευση «Προηγμένα Πληροφοριακά Συστήματα»
του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”
στο
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Ιούλιος 2024

Επιβλέπουσα: Χαλκίδη Μαρία
Ακαδημαϊκή Θέση: Αναπληρώτρια Καθηγήτρια

Συγγραφέας: Σιράγας Γεώργιος

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή: Σιράγας Γεώργιος
Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας: Μια προσέγγιση μεταφοράς γνώσης σε συστήματα συστάσεων συνεργατικού φιλτραρίσματος με μη-επικαλυπτόμενες οντότητες

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις …… [ημερομηνία έγκρισης] από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπων/ουσα(Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς)……[ονοματεπώνυμο, βαθμίδα]
Μέλος Εξεταστικής Επιτροπής: ……[ονοματεπώνυμο, βαθμίδα]
Μέλος Εξεταστικής Επιτροπής: ……[ονοματεπώνυμο, βαθμίδα]

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Ο Σιράγας Γεώργιος γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Μια προσέγγιση μεταφοράς γνώσης σε συστήματα συστάσεων συνεργατικού φιλτραρίσματος με μη-επικαλυπτόμενες οντότητες», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.
Επιπλέον *δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.*
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο ΔΗΛΩΝ
Ονοματεπώνυμο: Σιράγας Γεώργιος
Αριθμός Μητρώου: ΜΕ2148
Υπογραφή: Σιράγας Γεώργιος

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω πρωτίστως και από καρδιάς την επιβλέπουσα καθηγήτρια κα. Μαρία Χαλκίδη για την ανάθεση του θέματος, την άμεση και ουσιαστική επικοινωνία που διατηρήσαμε κατά την διάρκεια της συγγραφής της διπλωματικής εργασίας καθώς και για τον χρόνο τον οποίο διέθεσε για διορθώσεις και παρατηρήσεις.

Ευχαριστώ επίσης τα μέλη της εξεταστικής επιτροπής, καθώς και γενικότερα όλους τους καθηγητές του μεταπτυχιακού προγράμματος σπουδών για την εξαιρετική διδασκαλία τους, η οποία με εξόπλισε με σύγχρονες γνώσεις και δεξιότητες, τόσο σε ακαδημαϊκό, όσο και σε επαγγελματικό επίπεδο.

Τέλος, ευχαριστώ τους γονείς μου και τον αδερφό μου για την αμέριστη στήριξη που μου παρείχαν κατά την διάρκεια των σπουδών μου.

ΠΕΡΙΛΗΨΗ

Τα συστήματα συστάσεων χρησιμοποιούνται ευρέως στις πλατφόρμες ψυχαγωγίας και στις επιχειρήσεις ηλεκτρονικού εμπορίου για την εξατομίκευση της εμπειρίας των χρηστών και τη βελτίωση της ικανοποίησης των πελατών. Ωστόσο, αντιμετωπίζουν σημαντικές προκλήσεις, όπως το πρόβλημα της αραιότητας των δεδομένων, όπου τα δεδομένα αλληλεπίδρασης είναι αραιά, και το πρόβλημα της ψυχρής εκκίνησης (cold start), όπου νέοι χρήστες ή αντικείμενα δεν έχουν επαρκή δεδομένα αλληλεπίδρασης ώστε οι αλγόριθμοι που παράγουν συστάσεις να εκπαιδευθούν και να αποδώσουν ικανοποιητικά.

Η μεταφορά γνώσης αποτελεί μια προσέγγιση μηχανικής μάθησης, η οποία αξιοποιεί τη γνώση που αποκτάται από την επίλυση ενός προβλήματος για την ενίσχυση της λύσης ενός διαφορετικού αλλά σχετικού προβλήματος. Η εφαρμογή τεχνικών μεταφοράς γνώσης σε συστήματα συστάσεων αποσκοπεί στην αντιμετώπιση των ζητημάτων που αντιμετωπίζουν οι παραδοσιακές προσεγγίσεις επιτρέποντας την εκπαίδευση μοντέλων με λιγότερα δεδομένα και με βελτιωμένη απόδοση.

Στην παρούσα διπλωματική εργασία διερευνάται η εφαρμογή της μεταφοράς μάθησης σε συστήματα συστάσεων συνεργατικού φιλτραρίσματος με μη-επικαλυπτόμενες οντότητες. Πιο συγκεκριμένα, η μελέτη επικεντρώνεται σε σενάρια όπου δεν υπάρχουν κοινοί χρήστες ή αντικείμενα μεταξύ των τομέων, χρησιμοποιώντας μόνο τις αξιολογήσεις των χρηστών για τη μεταφορά γνώσης, χωρίς να γίνει χρήση μεταδεδομένων. Οι τομείς που χρησιμοποιούνται είναι Ταινίες και Βιβλία και τα πειραματικά αποτελέσματα αξιολογούνται με βάση την ακρίβεια της προτεινόμενης μεθόδου σε σύγκριση με παραδοσιακές τεχνικές συστάσεων που δεν κάνουν χρήση μεταφοράς γνώσης.

Αν και ο προτεινόμενος αλγόριθμος έδειξε ότι μπορεί να βελτιώσει τις συστάσεις και να μεταφέρει θετικά γνώση υπό συγκεκριμένες συνθήκες, δεν ήταν αρκετός για να ξεπεράσει σε απόδοση όλες τις παραδοσιακές μεθόδους με τις οποίες συγκρίθηκε και συγκεκριμένα την μέθοδο SVD (singular value decomposition).

ABSTRACT

Recommender systems are widely used in entertainment platforms and e-commerce businesses to personalize user experiences and enhance customer satisfaction. However, they face significant challenges, such as the sparsity problem, where interaction data is sparse, and the cold start problem, where new users or items lack sufficient interaction data for recommendation algorithms to train effectively.

Transfer learning is a machine learning approach that leverages knowledge gained from solving one problem to improve the solution of a different but related problem. Applying transfer learning techniques to recommender systems aims to address the challenges faced by traditional approaches, enabling the training of models with less data and improved performance.

This dissertation investigates the application of transfer learning in collaborative filtering recommender systems with non-overlapping entities. Specifically, the study focuses on scenarios where there are no common users or items between domains, using only user ratings to transfer knowledge without relying on metadata. The domains used are Movies and Books, and the experimental results are evaluated based on the accuracy of the proposed method compared to traditional recommendation techniques that do not utilize transfer learning.

Although the proposed algorithm demonstrated the potential to improve recommendations and positively transfer knowledge under certain conditions, it was not sufficient to outperform all traditional methods it was compared against, particularly the singular value decomposition (SVD) method.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

1.1 Σκοπός της διπλωματικής εργασίας

Ο στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η διερεύνηση της μεταφοράς μάθησης και των συστημάτων συστάσεων, καθώς και το πώς η μεταφορά γνώσης μπορεί να χρησιμοποιηθεί για την αντιμετώπιση των προκλήσεων που παρουσιάζουν τα συστήματα συστάσεων.

Την τελευταία εικοσαετία, τα συστήματα συστάσεων έχουν διεισδύσει ολοένα και περισσότερο στην καθημερινή ζωή λόγω της δημοφιλίας των ψηφιακών πλατφορμών ψυχαγωγίας και του ηλεκτρονικού εμπορίου. Ωστόσο, οι παραδοσιακές μέθοδοι μηχανικής μάθησης απαιτούν μεγάλο όγκο δεδομένων για να λειτουργήσουν αποτελεσματικά και βασίζονται σε ορισμένες υποθέσεις που συχνά δεν ισχύουν στην πραγματικότητα, όπως η σταθερότητα της κατανομής των δεδομένων με την πάροδο του χρόνου. Η μεταφορά γνώσης είναι ένα υποσχόμενο πεδίο της μηχανικής μάθησης που στοχεύει στην αξιοποίηση της γνώσης που έχει αποκτηθεί από την επίλυση ενός προβλήματος για την αντιμετώπιση ενός νέου προβλήματος.

Επιπλέον, στην παρούσα εργασία, επιδιώκεται η εφαρμογή μιας προσέγγισης μεταφοράς γνώσης μεταξύ τομέων σε συστήματα συστάσεων συνεργατικού φιλτραρίσματος με μη-επικαλυπτόμενες οντότητες. Πιο αναλυτικά, στο συγκεκριμένο σενάριο συστάσεων, γίνεται μεταφορά γνώσης μεταξύ δύο ή περισσότερων τομέων χωρίς να υπάρχουν κοινοί χρήστες και αντικείμενα, εκμεταλλευόμενοι μόνο τις αξιολογήσεις που έχουν κάνει οι χρήστες. Η συγκεκριμένη προσέγγιση είναι ιδιαίτερα ενδιαφέρουσα, καθώς δεν κάνει χρήση μεταδεδομένων. Τα αποτελέσματα παρουσιάζονται και κρίνονται βάσει της αποτελεσματικότητας και της ακρίβειας της προσέγγισης σε σχέση με παραδοσιακές μεθόδους συστάσεων. Τέλος, η προσέγγιση που έχει επιλεγεί, αντιμετωπίζεται με κριτική ματιά και αναλύονται τα θετικά και τα αρνητικά της στοιχεία.

1.2 Διάρθρωση της διπλωματικής εργασίας

Πέραν του εισαγωγικού πρώτου κεφαλαίου η υπόλοιπη εργασία διαρθρώνεται σε έξι κεφάλαια ως εξής:

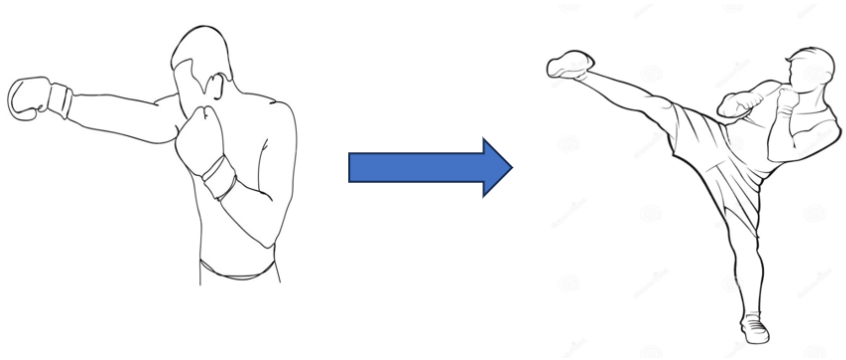
2. ΜΕΤΑΦΟΡΑ ΓΝΩΣΗΣ: Αναλύεται η μεταφορά γνώσης σε θεωρητικό επίπεδο, παρουσιάζονται τα κρίσιμα ερευνητικά ερωτήματα που την αφορούν και παρατίθενται ορισμένες τεχνικές που εφαρμόζονται στην βιβλιογραφία.
3. ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ: Γίνεται εκτενής παρουσίαση και ανάλυση του προβλήματος των συστάσεων, παρουσιάζονται ορισμένες χαρακτηριστικές τεχνικές, παρατίθενται τα ζητήματα που αντιμετωπίζουν τα παραδοσιακά συστήματα καθώς και οι διάφορές μετρικές αξιολόγησης που χρησιμοποιούνται. Τέλος, παρουσιάζονται βασικές προσεγγίσεις μεταφοράς γνώσης επί συστημάτων συστάσεων.
4. ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ ΚΑΙ ΕΡΓΑΛΕΙΑ ΕΠΙΛΥΣΗΣ: Παρουσιάζεται το πρόβλημα με το οποίο ασχολείται η διπλωματική εργασία, καθώς και τα εργαλεία και λογισμικά που χρησιμοποιήθηκαν. Τέλος, γίνεται ανάλυση των συνόλων δεδομένων του προβλήματος.
5. ΜΕΘΟΔΟΛΟΓΙΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΜΕΛΕΤΗΣ: Παρουσιάζεται ,σε θεωρητικό επίπεδο, η μέθοδος που έχει επιλεγεί για να επιλύσει το πρόβλημα.
6. ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΜΕΛΕΤΗΣ: Παρατίθενται τα αποτελέσματα των διαφόρων πειραμάτων που εκτελέστηκαν.
7. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ: Πραγματοποιείται κριτική ανασκόπηση της μεθόδου υλοποίησης με βάση τα πειραματικά αποτελέσματα και μια αναφορά σε μελλοντική ερασία.

2. ΜΕΤΑΦΟΡΑ ΓΝΩΣΗΣ

2.1 Εισαγωγή

Η μεταφορά γνώσης (transfer learning -TL) είναι μια μέθοδος μηχανικής μάθησης (machine learning–ML) που όπως και πολλές άλλες προσεγγίσεις και τεχνικές που εφαρμόζονται στην πληροφορική, οφείλει την έμπνευση της σε άλλο επιστημονικό πεδίο. Πιο συγκεκριμένα, στο πεδίο της γνωσιακής επιστήμης (cognitive science) η μεταφορά γνώσης ορίζεται ως η διαδικασία κατά την οποία η μάθηση σε μια κατάσταση ή σε ένα πλαίσιο συμβάλλει στην απόδοση ή την εκμάθηση σε ένα διαφορετικό, νέο πλαίσιο [1]. Η συμβολή στο νέο γνωστικό πλαίσιο δεν είναι κατά ανάγκη θετική (positive transfer), καθώς ενδέχεται η προγενέστερη εμπειρία να μην επηρεάσει καθόλου ή ακόμα και αρνητικά (negative transfer) ως προς την εκμάθηση.

Για παράδειγμα, η εξάσκηση σε μία πολεμική τέχνη μπορεί να βοηθήσει στην εκμάθηση μίας άλλης, καθώς πολλές πολεμικές τέχνες μοιράζονται βασικές αρχές και τεχνικές που μπορούν να μεταφερθούν και να ενισχύσουν την εκπαίδευση σε άλλες τέχνες. Αντίθετα, η εκμάθηση πιάνου δεν παρέχει κάποιο ουσιαστικό όφελος ως προς την εκμάθηση του ποδηλάτου, καθώς οι δεξιότητες που απαιτούνται για την καθεμία δραστηριότητα είναι εντελώς διαφορετικές και δεν υπάρχει κοινό έδαφος για μεταφορά γνώσεων ή δεξιοτήτων. Τέλος, η εκμάθηση γαλλικών ενδέχεται να επηρεάσει αρνητικά την σωστή εκμάθηση και προφορά των γερμανικών, καθώς οι δύο γλώσσες έχουν διαφορετικές φωνολογικές δομές και προφορά, γεγονός που μπορεί να προκαλέσει σύγχυση.

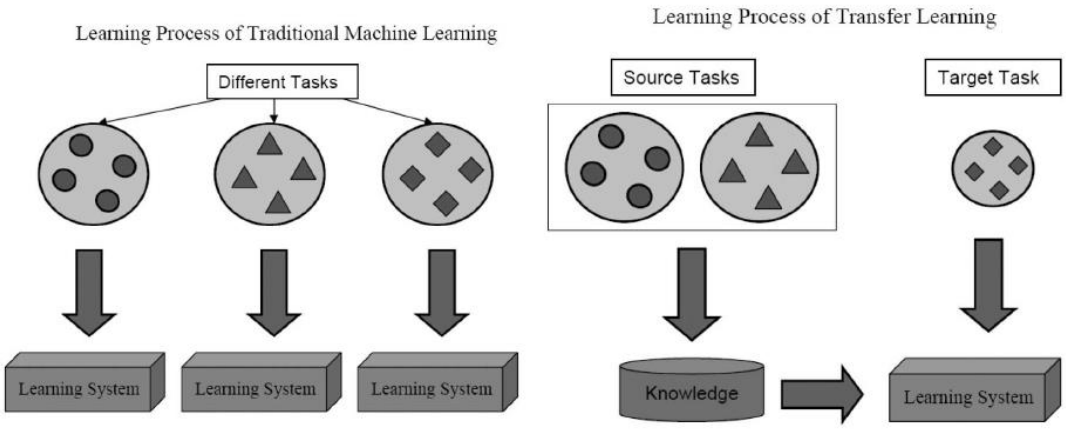


Εικόνα 1- Θετική Μεταφορά γνώσης: boxing to kick-boxing

Η μεταφορά γνώσης ως μέθοδος μηχανικής μάθησης επιχειρεί να αξιοποιήσει την παραπάνω συλλογιστική, χρησιμοποιώντας την γνώση που έχει αποκτηθεί από ένα ή περισσότερα ήδη εκπαιδευμένα μοντέλα με σκοπό να βελτιώσει ή να επιταχύνει την διαδικασία εκπαίδευσης ενός νέου μοντέλου. Όπως στην περίπτωση της γνωσιακής επιστήμης, έτσι και όσον αφορά την μηχανική μάθηση, θα πρέπει οι εργασίες (tasks) και οι τομείς (domains) που μοντελοποιούνται να σχετίζονται επαρκώς μεταξύ τους ώστε να επιτευχθεί θετική μεταφορά γνώσης

Το TF, πέρα από το ενδιαφέρον που παρουσιάζει σε εννοιολογικό επίπεδο έχει ιδιαίτερα χρήσιμες πρακτικές εφαρμογές, καθώς επιχειρεί να ξεπεράσει κάποιους από τους μεγαλύτερους

περιορισμούς της παραδοσιακής μηχανικής μάθησης, όσον αφορά τις πρακτικές εφαρμογές. Οι αλγόριθμοι μηχανικής μάθησης και τεχνητής νοημοσύνης (πχ. deep neural networks) χρειάζονται ένα μεγάλου όγκου σύνολο δεδομένων ώστε να είναι αποδοτικοί [2]. Η συλλογή και διαχείριση επισημασμένων (labeled) ή μη-επισημασμένων (unlabeled) δεδομένων για εποπτευόμενους (supervised) ή μη-εποπτευόμενους (unsupervised) αλγορίθμους αντίστοιχα, αποτελεί μια ιδιαίτερα κοστοβόρα, χρονοβόρα και σε κάποιες περιπτώσεις ακόμη και μη εφικτή διαδικασία. Επιπλέον, μια από τις βασικές υποθέσεις των παραδοσιακών αλγορίθμων μηχανικής μάθησης είναι ότι τα δεδομένα εκπαίδευσης (train) και δοκιμής (test) προέρχονται από την ίδια κατανομή και ότι αυτή η κατανομή δεν θα αλλάξει με την πάροδο του χρόνου. Σε περίπτωση που συμβεί το τελευταίο, τότε το μοντέλο θα πρέπει να εκπαιδευθεί εκ νέου, αυξάνοντας έτσι το συνολικό κόστος ανάπτυξης και εφαρμογής του πληροφοριακού συστήματος. Τέλος, η Εικόνα 2 παρουσιάζει την διαφορά μεταξύ ML και TL σε εννοιολογικό επίπεδο.



Εικόνα 2- Traditional ML vs TL [3]

2.2 Ορισμός της μεταφοράς γνώσης

Για να προχωρήσουμε στην παρουσίαση και ανάλυση των διάφορων τεχνικών, αλλά και στην κατηγοριοποίηση αυτών θα πρέπει πρώτα να δοθούν αυστηροί ορισμοί για τους όρους τομέας (domain), εργασία (task) και μεταφορά γνώσης (TF). Οι παρακάτω ορισμοί δίνονται σύμφωνα με το [4]:

Τομέας (Domain): Ένας τομέας D αποτελείται από δύο μέρη, τον χώρο χαρακτηριστικών $\mathcal{X}$ και την οριακή κατανομή $P(X)$, όπου $X = \{x_i | \, x_i \in \mathcal{X}, i = 1, \dots, n\}$, δηλαδή X ένα σύνολο παρατηρήσεων. Πιο συνοπτικά ο τομέας μπορεί να συμβολιστεί ως $D = \{\mathcal{X}, P(X)\}$. Η οριακή κατανομή P είναι γενικά άγνωστη και είναι δύσκολο να εκφραστεί σε κλειστή μορφή.

Εργασία (Task): Μια εργασία $\mathcal{T}$ αποτελείται από έναν χώρο ετικετών $\mathcal{Y}$ (labeled instances) και μία συνάρτηση απόφασης f, η οποία είναι άγνωστη και μαθαίνεται από τα δειγματικά δεδομένα. Πιο συνοπτικά η εργασία μπορεί να συμβολιστεί ως $\mathcal{T} = \{\mathcal{Y}, f\}$. Επιπλέον, σε πολλές εφαρμογές ML, η συνάρτηση απόφασης f αποτελεί ουσιαστικά την πρόβλεψη για την κατανομή της δεσμευμένης

πιθανότητας της ετικέτας-μεταβλητής στόχου δεδομένης της επεξηγηματικής μεταβλητής, ή πιο αυστηρά: $f(x_i) = \{P(y_k | x_i) \mid y_k \in \mathcal{Y}, k = 1, \dots, |\mathcal{Y}|\}$

Πριν δοθεί ο αυστηρός ορισμός τους TL θα πρέπει να αναφερθεί ότι οι διάφοροι τομείς που συμμετέχουν στο πρόβλημα παρουσιάζονται πρακτικά ως ζεύγη χαρακτηριστικών και ετικετών. Για παράδειγμα, ένας source τομέας αναμένεται να έχει την εξής δομή:
$D_S = \{(x, y) \mid x_i \in \mathcal{X}^S, y_i \in \mathcal{Y}^S, i = 1, \dots, n^S\}$. Σε αντίθεση με τους source τομείς, ένας target τομέας ενδέχεται να διαθέτει μόνο μη-επισημασμένες παρατηρήσεις και λίγες ή ακόμα και καθόλου επισημασμένες παρατηρήσεις.

Μεταφορά Γνώσης (TL): Δεδομένων κάποιον παρατηρήσεων που αντιστοιχούν σε $m^S \in \mathbb{N}^+$ το πλήθος source τομέων και εργασιών, ή αλλιώς $\{(D_{S_i}, \mathcal{T}_{S_i}) \mid i = 1, \dots, m^S\}$ και κάποιον παρατηρήσεων που αντιστοιχούν σε $m^T \in \mathbb{N}^+$ το πλήθος target τομέων, ή αλλιώς $\{(D_{T_j}, \mathcal{T}_{T_j}) \mid j = 1, \dots, m^T\}$, τότε με τον όρο μεταφορά γνώσης ορίζεται η αξιοποίηση της γνώσης που έχει εξαχθεί από τους source τομείς με σκοπό την βελτίωση της απόδοσης των συναρτήσεων απόφασης f^{T_j}, με $j = 1, \dots, m^T$, για τους τομείς στόχους.

2.3 Βασικά ερωτήματα για την μεταφορά γνώσης

Παρόλο που ο παραπάνω ορισμός κάνει λόγο για παραπάνω από ένα τομείς source και target, για τα παραδείγματα και τους ορισμούς που ακολουθούν θα χρησιμοποιηθεί το σενάριο με έναν τομέα εκατέρωθεν, χωρίς βλάβη της γενικότητας. Πριν προχωρήσουμε στην κατηγοριοποίηση και ανάλυση των διάφορων προσεγγίσεων TL παρουσιάζονται συνοπτικά τα 3 κύρια ερευνητικά ερωτήματα (what,how and when to transfer) που θα πρέπει να λαμβάνονται υπόψιν σε κάθε υλοποίηση TL:

- What to transfer: Σε ένα σενάριο TL με έναν source τομέα D_S και έναν target D_T με τις αντίστοιχες εργασίες $\mathcal{T}_S$ και $\mathcal{T}_T$, ενδέχεται κάποια από την γνώση που έχει εξαχθεί για τον D_S να έχει αξία μόνο για τον τομέα αυτόν. Γενικότερα, ορισμένες φορές η γνώση που εξάγεται έχει συγκεκριμένη χρήση για μεμονωμένους τομείς ή εργασίες. Συνεπώς, θα πρέπει να γίνει επιλογή για το ποιο κομμάτι της γνώσης που αποκτήθηκε για τον D_S έχει νόημα να μεταφερθεί στον D_T, έτσι ώστε να μπορεί να βοηθήσει στη βελτίωση των επιδόσεων της εργασίας $\mathcal{T}_T$, κάτι που αποτελεί και τον απώτερο σκοπό του συστήματος.
- How to transfer: Αφού ανακαλυφθεί ποια γνώση μπορεί και έχει νόημα να μεταφερθεί μεταξύ των τομέων στο πρώτο βήμα, σειρά έχει η ανάπτυξη αλγορίθμων μάθησης, οι οποίοι θα εκτελέσουν την μεταφορά.
- When to transfer: Ένα κρίσιμο ερώτημα, το οποίο πολύ συχνά παραμελείται στην βιβλιογραφία, είναι το αν θα πρέπει να εφαρμοστεί μια προσέγγιση μεταφοράς γνώσης για την επίλυση ενός προβλήματος ή όχι. Όπως επισημάνθηκε και στην εισαγωγή, στην περίπτωση όπου οι τομείς D_S και D_T δεν σχετίζονται επαρκώς μεταξύ τους, η μεταφορά γνώσης μπορεί να επιφέρει έως και αρνητικά αποτελέσματα. Επιπλέον, θα πρέπει να ληφθεί υπόψιν το γεγονός ότι καθώς το TL προσθέτει επιπλέον πολυπλοκότητα στις ήδη υπολογιστικά κοστοβόρες διαδικασίες ML που εφαρμόζονται σε έναν τομέα, ενδεχομένως

μια μικρή βελτίωση στην επίδοση του συστήματος δεν αρκεί για να δικαιολογήσει το πρόσθετο κόστος όσον αφορά την ταχύτητα εκπαίδευσης του μοντέλου και τον χρόνο που αφιερώνεται για την βελτιστοποίηση και συντήρηση του.

2.4 Ταξινομήσεις του προβλήματος μεταφοράς γνώσης

Στην βιβλιογραφία παρουσιάζονται διάφοροι τρόποι ταξινόμησης του προβλήματος της μεταφοράς γνώσης και των διάφορων προσεγγίσεων που εφαρμόζονται για την επίτευξη της. Στα παρακάτω χρησιμοποιούνται για χάριν παραδείγματος δύο τομείς D_S(source) και D_T(target) και τα αντίστοιχα στοιχεία τους-εργασίες, χώροι χαρακτηριστικών κλπ.

Μια πρώτη, απλή, κατηγοριοποίηση του TF γίνεται με βάση την συνοχή των χώρων χαρακτηριστικών $\mathcal{X}^{\mathrm{S}}$ και $\mathcal{X}^{T}$ ή/και των χώρων ετικετών $\mathcal{Y}^{\mathrm{S}}$, $\mathcal{Y}^{\mathrm{T}}$ των δύο τομέων. Πιο συγκεκριμένα, εάν ισχύει $\mathcal{X}^{\mathrm{S}} = \mathcal{X}^{T}$και $\mathcal{Y}^{\mathrm{S}} = \mathcal{Y}^{\mathrm{T}}$, τότε το σενάριο αυτό ονομάζεται ομογενής μεταφορά γνώσης (homogenous transfer learning) , ενώ αν ισχύει $\mathcal{X}^{\mathrm{S}} \neq \mathcal{X}^{T}$ ή/και $\mathcal{Y}^{\mathrm{S}} \neq \mathcal{Y}^{\mathrm{T}}$ τότε έχουμε ετερογενή μεταφορά γνώσης (heterogenous transfer learning)

Μια πιο σύνθετη και λεπτομερής ταξινόμηση του προβλήματος TF έχει δοθεί από τους Pan και Yang [3] και έχει δύο επίπεδα, τα οποία προκύπτουν ως προς:

- Την σύγκριση των D_S και D_T ως σχετικά με το αν υπάρχουν ή όχι διαθέσιμα δεδομένα με ετικέτα (labeled data) σε καθένα από τους δύο τομείς.
- Την επιλογή στο «τι θα μεταφερθεί» ως γνώση από τον D_S στον D_T. Πρόκειται ουσιαστικά για την μελέτη και κατηγοριοποίηση του TF με βάση το ερώτημα “what to transfer” που αναλύθηκε στην προηγούμενη παράγραφο.

Στην επόμενη παράγραφο παρουσιάζεται η ταξινόμηση του TF ως προς το πρώτο επίπεδο, δηλαδή την διαθεσιμότητα ή όχι δεδομένων με ετικέτα.

2.4.1 Επαγωγική, μεταγωγική και μη-εποπτευόμενη μεταφορά γνώσης

Επαγωγική μεταφορά

Αν στον τομέα D_T υπάρχουν δεδομένα με ετικέτα τότε η μεταφορά γνώσης χαρακτηρίζεται ως επαγωγική (inductive). Στην επαγωγική μεταφορά γνώσης ισχύει $\mathcal{T}_S \neq \mathcal{T}_T$,δηλαδή η εργασία στον τομέα στόχο D_T είναι διαφορετική από την εργασία στον source τομέα D_S, ανεξαρτήτως από το αν οι D_S και D_T είναι ίδιοι ή όχι. Επιπλέον, ανάλογα με το αν υπάρχουν δεδομένα με ετικέτα στον τομέα D_S τότε προκύπτουν δύο υποκατηγορίες, οι οποίες έχουν ως εξής:

- Υπάρχει πληθώρα διαθέσιμων δεδομένων με ετικέτα στον τομέα D_S. Σε αυτή την περίπτωση, η επαγωγική μεταφορά γνώσης είναι παρόμοια με τη ταυτόχρονη μάθηση (multitask learning). Η ουσιώδης διαφορά μεταξύ των δύο είναι ότι, στην επαγωγική μεταφορά γνώσης μοναδικός στόχος είναι η βελτίωση της εργασίας $\mathcal{T}_T$, σε αντίθεση με το

multitask-learning, όπου γίνεται προσπάθεια για βελτίωση της απόδοσης των εργασιών και των δύο τομέων, δηλαδή των $\mathcal{T}_S$, $\mathcal{T}_T$.

- Δεν υπάρχουν διαθέσιμα δεδομένα με ετικέτα στον τομέα D_S. Το συγκεκριμένο σενάριο είναι παρόμοιο με αυτό της αυτοδίδακτης μάθησης (self-taught learning), όπου ένα σύστημα μαθαίνει να αναγνωρίζει πρότυπα ή να κάνει προβλέψεις χωρίς να διαθέτει προκαθορισμένες ετικέτες για τα δεδομένα εκπαίδευσης. Αντίθετα, χρησιμοποιεί μη επισημασμένα δεδομένα για να κατασκευάσει χρήσιμες αναπαραστάσεις που μπορούν στη συνέχεια να βελτιώσουν την απόδοση της μάθησης με επισημασμένα δεδομένα [5].

Μεταγωγική μεταφορά

Αν στον τομέα D_S υπάρχουν πολλά δεδομένα με ετικέτα και ταυτόχρονα όχι στον D_T τότε η μεταφορά γνώσης χαρακτηρίζεται ως μεταγωγική (transductive). Επίσης, στην επαγωγική μεταφορά γνώσης ισχύει $D_S \neq D_T$, δηλαδή οι τομείς είναι διαφορετικοί μεταξύ τους. Όπως έχει αναφερθεί σε προηγούμενη παράγραφο, ένας τομέας ορίζεται μέσω του χώρου χαρακτηριστικών $\mathcal{X}$ και της οριακής κατανομής $P(X)$, συνεπώς προκύπτουν δύο υποκατηγορίες για την μεταγωγική ,μεταφορά γνώσης:

- Οι χώροι χαρακτηριστικών είναι διαφορετικοί, δηλαδή $\mathcal{X}^S \neq \mathcal{X}^T$ και ουσιαστικά προκύπτει το σενάριο της ετερογενούς μεταφοράς γνώσης.
- Οι χώροι χαρακτηριστικών είναι όμοιο, δηλαδή $\mathcal{X}^S = \mathcal{X}^T$ και η διαφορά έγκειται στις οριακές κατανομές, $P(X_S) \neq P(X_T)$. Στο συγκεκριμένο σενάριο χρησιμοποιούνται διάφορες τεχνικές που επιδιώκουν να γεφυρώσουν το χάσμα μεταξύ των διαφορετικών κατανομών δεδομένων, ώστε να μπορέσει να γίνει μεταφορά γνώσης και να βελτιωθεί η απόδοση του μοντέλου στον D_T.

Μη-εποπτευόμενη μεταφορά

Αν δεν υπάρχουν δεδομένα με ετικέτα σε κανέναν από τους δύο τομείς D_S και D_T τότε η μεταφορά γνώσης χαρακτηρίζεται ως μη-εποπτευόμενη (unsupervised). Στην μη-εποπτευόμενη μεταφορά γνώσης ισχύει $\mathcal{T}_S \neq \mathcal{T}_T$, όπως και στην περίπτωση της επαγωγικής. Παρόλου που οι εργασίες διαφέρουν μεταξύ τους, ταυτόχρονα σχετίζονται έως ένα βαθμό. Σε αυτή την περίπτωση χρησιμοποιούνται τεχνικές όπως συσταδοποίηση (clustering) και τεχνικές για μείωση των διαστάσεων, ώστε να βελτιωθεί η απόδοση στον D_T.

2.4.2 Ταξινόμηση βάση του τρόπου επίλυσης

Όπως αναφέρθηκε και παραπάνω το δεύτερο επίπεδο ταξινόμησης των προσεγγίσεων του προβλήματος της μεταφοράς γνώσης έχει να κάνει με το «τι μεταφέρεται» από τον D_S στον D_T. Πιο συγκεκριμένα, ορίζονται τέσσερις διαφορετικές προσεγγίσεις οι οποίες είναι οι εξής:

- Στην μεταφορά γνώσης βασισμένη σε στιγμιότυπα (instance-based) γίνεται η υπόθεση ότι ορισμένα από τα δεδομένα του τομέα D_S μπορούν να επαναχρησιμοποιηθούν για τη

μάθηση στον D_T με την μέθοδο της επαναστάθμισης (reweighting). Στα πλαίσια της επαγωγικής μεταφοράς γνώσης μια χαρακτηριστική τεχνική instance-based είναι ο αλγόριθμος TrAdaBoost [6]. Ο αλγόριθμος χρησιμοποιεί μια μικρή ποσότητα επισημασμένων δεδομένων στον D_T και ταυτόχρονα προσπαθεί να επανασταθμίσει τα δεδομένα του D_S που είναι περισσότερα, ώστε να κατασκευάσει ένα υψηλής απόδοσης μοντέλο ταξινόμησης για τον τομέα στόχο D_T. Κατά την εκπαίδευση το TrAdaBoost προσπαθεί να επαναπροσδιορίσει το βάρος των δεδομένων του D_S σε κάθε επανάληψη για να μειώσει την επίδραση των "αρνητικών" δεδομένων, ενώ ενθαρρύνει τα "καλά" δεδομένα πηγής να συνεισφέρουν περισσότερο για τον D_T.

- Η μεταφορά γνώσης βασισμένη σε χαρακτηριστικά (feature-based) μπορεί να εφαρμοστεί στο πλαίσιο και των τριών κατηγοριών που αναλύθηκαν στην προηγούμενη παράγραφο (inductive, transductive, un-supervised). Η βασική ιδέα είναι να μάθει μια "καλή" αναπαράσταση χαρακτηριστικών ο τομέα-στόχος D_T. Η συγκεκριμένη κατηγορία μεταφοράς μάθησης, χωρίζεται σε δύο υποκατηγορίες, την συμμετρική και μη-συμμετρική. Στην συμμετρική προσέγγιση μετασχηματίζονται τα χαρακτηριστικά του D_S ώστε να προσεγγίσουν τα χαρακτηριστικά του D_T. Στην μη-συμμετρική προσέγγιση γίνεται προσπάθεια για την εύρεση ενός κοινού κρυμμένου χώρου χαρακτηριστικών μεταξύ των D_S και D_T. Σκοπός είναι η αξιοποίηση των κοινών χαρακτηριστικών για τη βελτίωση της μάθησης στον D_T, παρόλο που τα πλήρη σύνολα χαρακτηριστικών των D_S και D_T δεν είναι πανομοιότυπα. Για να γίνει ο μετασχηματισμός των χαρακτηριστικών θα πρέπει να ελαχιστοποιηθεί η διαφορά της οριακής και της δεσμευμένης κατανομής μεταξύ των τομέων, να διατηρηθούν οι ιδιότητες των δεδομένων και να βρεθεί αντιστοιχία μεταξύ των χαρακτηριστικών για τους τομείς D_S και D_T. Μια μετρική που χρησιμοποιείται ευρέως στον χώρο του TL και είναι η Maximum Mean Discrepancy (MMD) [7]. Η συγκεκριμένη μετρική αποτελεί μέτρο ομοιότητας για τις κατανομές και ορίζεται ως: $\mathrm{MMD}(X^S, X^T) = || \frac{1}{n^S}\sum_{i=1}^{n^S}\Phi(x_i^S) - \frac{1}{n^T}\sum_{j=1}^{n^T}\Phi(x_j^T) ||_{\mathcal{H}}^2$, όπου το Φ είναι μια απεικόνιση (μετασχηματισμός) των δεδομένων σε έναν Reproducing Kernel Hilbert Space (RKHS).

- Η μεταφορά γνώσης βασισμένη σε παράμετρο (parameter-transfer) εφαρμόζεται κυρίως στο πλαίσιο της επαγωγικής μεταφοράς γνώσης και υποθέτει ότι οι εργασίες $\mathcal{T}_S$ και $\mathcal{T}_T$ μοιράζονται κάποιες παραμέτρους ή προγενέστερες κατανομές (prior distributions) των υπερπαραμέτρων (hyperparameters). Οι περισσότερες προσεγγίσεις που για αυτό το σενάριο έχουν σχεδιαστεί για να λειτουργούν υπό το πλαίσιο του multi-task learning. Ωστόσο, μπορούν εύκολα να τροποποιηθούν για να εξυπηρετήσουν την μεταφορά γνώσης αναθέτοντας μεγαλύτερο βάρος στη συνάρτηση απώλειας του τομέα στόχου, αντί για ίσα βάρη που χρησιμοποιούνται στην multi-task προσέγγιση , ώστε να βελτιστοποιηθεί μόνο η εργασία $\mathcal{T}_T$.

- Η προσέγγιση μεταφοράς γνώσης σε σχέσεις (relational knowledge transfer) διαφέρει κατά πολύ σε σχέση τις προηγούμενες προσεγγίσεις, καθώς στους σχεσιακούς τομείς τα δείγματα-δεδομένα δεν είναι ανεξάρτητα και ομοιόμορφα κατανεμημένα, καθώς

αναπαρίστανται μέσω γράφων και δεδομένων κοινωνικών δικτύων. Οι αλγόριθμοι που χρησιμοποιούνται κάνουν χρήση των δομικών εννοιών «οντότητα» και «σχέση» και ο απώτερος σκοπός είναι να βρεθούν τρόποι σύνδεσης των σχέσεων μεταξύ των τομέων D_S και D_T και αντίστοιχα των οντοτήτων μεταξύ των τομέων D_S και D_T.

3. ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ

3.1 Εισαγωγή

Τα συστήματα συστάσεων (recommender systems) είναι λογισμικά τα οποία, βασιζόμενα σε τεχνικές στατιστικής και μηχανικής μάθησης, καθορίζουν τα πιο κατάλληλα αντικείμενα που θα πρέπει να παρουσιαστούν σε έναν χρήστη [8]. Τα αντικείμενα αυτά μπορεί να περιλαμβάνουν μια ευρεία γκάμα από κατηγορίες, όπως ταινίες, βιβλία, μουσικά κομμάτια, άρθρα από ηλεκτρονικές εφημερίδες, καθώς και διάφορα άλλα προϊόντα, όπως ρούχα, ηλεκτρονικά είδη και υπηρεσίες.

Η ανάπτυξη αυτών των συστημάτων έχει παρουσιάσει σημαντική πρόοδο και έχει προσελκύσει μεγάλο ερευνητικό ενδιαφέρον τα τελευταία χρόνια, κυρίως λόγω της ραγδαίας αύξησης του ηλεκτρονικού εμπορίου. Η περιορισμένη επιφάνεια προβολής στις οθόνες υπολογιστών και κινητών τηλεφώνων, σε συνδυασμό με την πληθώρα των διαθέσιμων προϊόντων καθιστούν αναγκαία την εξατομίκευση των προτάσεων για τη βελτίωση της εμπειρίας του χρήστη. Ταυτόχρονα, σε ένα ιδιαίτερα ανταγωνιστικό περιβάλλον, όπως αυτό του ηλεκτρονικού εμπορίου, η ικανότητα των συστημάτων συστάσεων να προσαρμόζουν τις προτάσεις στις ατομικές προτιμήσεις των χρηστών με υψηλό βαθμό ακρίβειας αποτελεί σημαντικό ανταγωνιστικό πλεονέκτημα για τις επιχειρήσεις. Ενδεικτικά, το 35% [9] των πωλήσεων στην διαδικτυακή πλατφόρμα αγορών της Amazon, καθώς και περίπου το 80% [10] του περιεχομένου που καταναλώνεται στην πλατφόρμα ψηφιακής ψυχαγωγίας Netflix, προκύπτουν ως αποτέλεσμα των προτάσεων RS.

3.2 Κλασικός ορισμός του προβλήματος των συστάσεων

Στην πιο κοινή και απλή του μορφή, το πρόβλημα των συστάσεων (recommendation problem) ανάγεται στην συμπλήρωση των κενών στοιχείων ενός πίνακα R, ο οποίος ονομάζεται πίνακας αξιολογήσεων (rating matrix). Πιο συγκεκριμένα, υποθέτοντας m χρήστες και n αντικείμενα, τότε ως πίνακας αξιολογήσεων ορίζεται ο πίνακας R, διαστάσεων $m \times n$, όπου το στοιχείο R_{ij} του πίνακα αντιπροσωπεύει την αξιολόγηση που έχει δώσει ο χρήστης i για το αντικείμενο j (Εικόνα 3). Αν ο χρήστης i δεν έχει αξιολογήσει το αντικείμενο j, τότε το στοιχείο R_{ij} παραμένει κενό ή λαμβάνει την τιμή μηδέν. Σκοπός των διάφορων τεχνικών και αλγορίθμων που χρησιμοποιούνται στα RS είναι να προβλέψουν τις αξιολογήσεις που λείπουν, δηλαδή να συμπληρώσουν τα στοιχεία του πίνακα R με τις βέλτιστες δυνατές τιμές για κάθε συνδυασμό χρήστη-αντικειμένου που δεν έχει παρατηρηθεί ακόμη.

Αξίζει επίσης να αναφερθεί ότι τα στοιχεία του πίνακα R δύναται να αναπαριστούν είτε ρητές αξιολογήσεις (explicit ratings), οι οποίες συνήθως προκύπτουν από ενέργειες των χρηστών, είτε έμμεσες αξιολογήσεις (implicit ratings), οι οποίες προέρχονται από την παρατήρηση της συμπεριφοράς του χρήστη ως προς τα αντικείμενα, χωρίς να απαιτείται άμεση ενέργεια από τον ίδιο. Ενδεικτικό παράδειγμα ρητών αξιολογήσεων αποτελούν οι βαθμολογίες που δίνει ένας χρήστης στις ταινίες που έχει παρακολουθήσει σε μια κλίμακα, όπως από 1 έως 5 αστέρια, ενώ για τις έμμεσες αξιολογήσεις, ένα χαρακτηριστικό παράδειγμα είναι ο χρόνος που αφιερώνει ο χρήστης σε μια ιστοσελίδα.

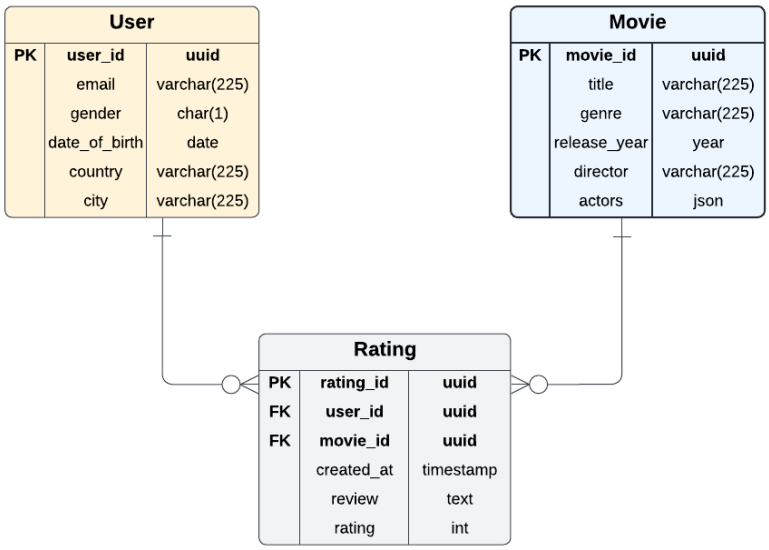
		Movies (n)			
		Star Wars	**The Godfather**	**Interstellar**	**Titanic**
Users (m)	**John**	3	Ø	1	5
	Helen	4	2	Ø	1
	Maria	Ø	4	5	3
	George	3	5	4	Ø

Εικόνα 3- Παράδειγμα rating matrix R, $R_{ij} \in \{1,2,3,4,5\}$

Ένας πιο αυστηρός ορισμός του προβλήματος των συστάσεων είναι ο εξής [11]:
Έστω U το σύνολο των χρηστών, V το σύνολο όλων των πιθανών αντικειμένων που μπορούν να προταθούν και η συνάρτηση χρησιμότητας (utility function) f, $f: U \times V \rightarrow R$, όπου το R είναι ένα ολικά διατεταγμένο σύνολο (θετικοί ακέραιοι ή πραγματικοί αριθμοί με συγκεκριμένο εύρος τιμών). Τότε, για κάθε $u \in U$ το RS προσπαθεί να επιλέξει αντικείμενο $v' \in V$, τέτοιο ώστε να μεγιστοποιεί την συνάρτηση χρησιμότητας του χρήστη, ή αλλιώς $\forall u \in U, v' = \underset{v' \in V}{arg\max} f(u, v)$.

Από το παράδειγμα της Εικόνας 3 γίνεται εμφανές ότι η συνάρτηση χρησιμότητας δεν ορίζεται σε όλο τον χώρο $U \times V$, αλλά σε ένα υποσύνολο αυτού, καθώς υπάρχουν κενά στοιχεία. Πιο συγκεκριμένα, ο πίνακας R είναι συνήθως αραιός (sparse), καθώς στην πραγματικότητα κάθε χρήστης αλληλοεπιδρά με πολύ λίγα αντικείμενα σε σχέση με το συνολικό πλήθος αυτών που προσφέρονται. Υπό αυτό το πρίσμα, σκοπός του RS είναι να ορίσει την συνάρτηση f με τέτοιο τρόπο ώστε να επιβεβαιώνονται εμπειρικά οι υπάρχουσες τιμές, καθώς και να προβλέψει τις άγνωστες τιμές μεγιστοποιώντας κάποιο κριτήριο απόδοσης, όπως το μέσο απόλυτο σφάλμα (mean absolute error).

Σε ένα παραγωγικό πληροφοριακό σύστημα, όπως αυτό ενός μεγάλου παρόχου ταινιών σε ψηφιακή μορφή, οι χώροι U (χρήστες) και V (αντικείμενα - ταινίες) είναι συνήθως πολυδιάστατοι. Για παράδειγμα, στη βάση δεδομένων ενός τέτοιου συστήματος αποθηκεύονται, συνήθως, εκτός από το μοναδικό προσδιοριστικό κάθε χρήστη (user_id), και μεταδεδομένα (metadata) που αφορούν και περιγράφουν τον χρήστη, όπως ονοματεπώνυμο, φύλο, ηλικία, τόπος διαμονής. Παρομοίως, για κάθε ταινία αποθηκεύονται επίσης μεταδεδομένα όπως τίτλος, είδος, έτος κυκλοφορίας, και σκηνοθέτης. Στην Εικόνα 4 παρουσιάζεται ένα απλουστευμένο, αλλά πιο ρεαλιστικό μοντέλο δεδομένων επί του οποίου εφαρμόζονται οι διάφορες τεχνικές των συστημάτων συστάσεων.



Εικόνα 4- Μοντέλο Οντοτήτων Συσχετίσεων για αξιολογήσεις ταινιών

Ανάλογα με τις διάφορες τεχνικές και τον τύπο δεδομένων που αξιοποιεί ένα σύστημα συστάσεων ώστε να πραγματοποιήσει τις προτάσεις προκύπτουν οι παρακάτω βασικές κατηγορίες και υποκατηγορίες [12, 13] , οι οποίες και θα αναλυθούν περεταίρω στην συνέχεια:

- Τεχνικές Φιλτραρίσματος βασισμένες στο περιεχόμενο (content-based filtering)
- Τεχνικές Συνεργατικού Φιλτραρίσματος (collaborative filtering)
    - Βασιζόμενες στην Μνήμη (memory-based)
    - Βασιζόμενες σε Μοντέλο (model-based)
- Υβριδικές Προσεγγίσεις (hybrid approaches)

3.3 Εναλλακτικές προσεγγίσεις του προβλήματος των συστάσεων

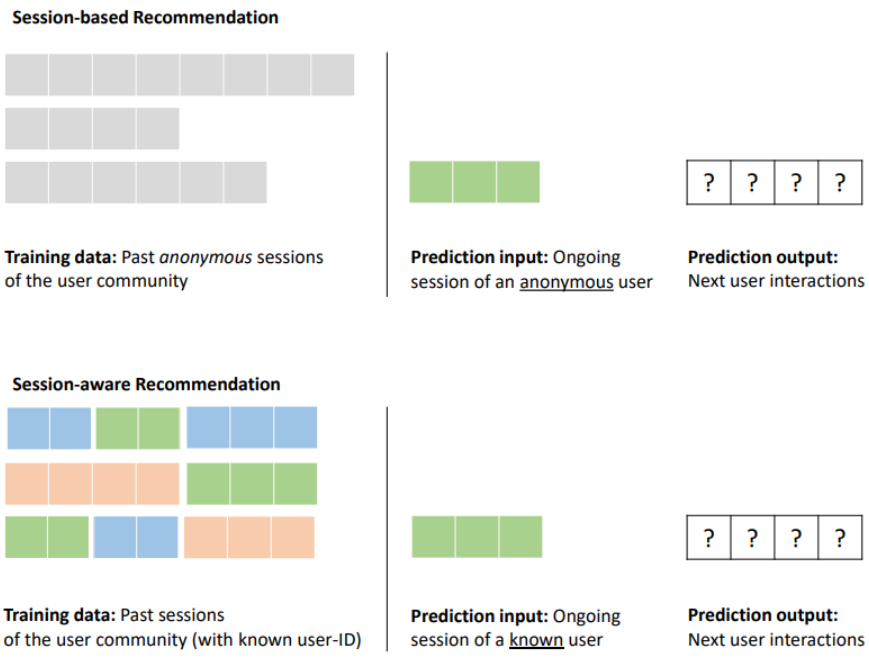
Οι προηγούμενοι ορισμοί μέσω του πίνακα αξιολογήσεων και της συνάρτησης χρησιμότητας δεν είναι οι μοναδικοί, καθώς καλύπτουν μια συγκεκριμένη έκφανση του προβλήματος των συστάσεων. Στην σύγχρονη έρευνα παρουσιάζονται και μελετώνται διαρκώς νέες προσεγγίσεις του προβλήματος. Χαρακτηριστικά παραδείγματα αποτελούν τα sequence-aware και τα multistakeholder recommender systems.

Τα sequence-aware RS χρησιμοποιούν τις διάφορες αλληλεπιδράσεις ενός χρήστη με μια εφαρμογή για ένα συγκεκριμένο χρονικό ορίζοντα με σκοπό να προβλέψουν την επόμενη ή τις επόμενες δράσεις του χρήστη. Οι αλληλεπιδράσεις αυτές μπορεί να είναι το γεγονός ότι ο χρήστης είδε ένα αντικείμενο, αν το έβαλε στο καλάθι αγορών κ.α. Σε ένα τέτοιο σενάριο γίνεται προφανές ότι η συνάρτηση χρησιμότητας και ο πίνακας αξιολογήσεων που ορίστηκε παραπάνω δεν επαρκούν για τον ορισμό και την αντιμετώπιση του προβλήματος. Μια κατηγοριοποίηση των

sequence aware RS ως προς τον χρονικό ορίζοντα που χρησιμοποιείται για τις προτάσεις είναι η εξής [14]:

- Last-N interactions-based recommendation: Λαμβάνονται υπόψιν μόνο οι τελευταίες N δράσεις του χρήστη και σκοπός είναι η πρόβλεψη της αμέσως επόμενης δράσης. Μια συνηθισμένη εφαρμογή του συγκεκριμένου τρόπου προτάσεων είναι η πρόβλεψη της επόμενης τοποθεσίας (check-in), όπου είτε δεν υπάρχουν πολλές προηγούμενες ενέργειες για ένα χρήστη, είτε οι παρελθοντικές παρατηρήσεις δεν προσφέρουν ιδιαίτερη αξία για την πρόβλεψη.
- Session-based recommendation: Λαμβάνονται υπόψιν μόνο οι δράσεις του χρήστη για το συγκεκριμένο session, καθώς ο χρήστης εισέρχεται για πρώτη φορά στην εφαρμογή ή περιηγείται ανώνυμα. Κάτι τέτοιο συμβαίνει πολύ συχνά σε πλατφόρμες ψηφιακού περιεχομένου καθώς και ηλεκτρονικού εμπορίου.
- Session-aware recommendation: Λαμβάνονται υπόψιν οι δράσεις του χρήστη για το συγκεκριμένο αλλά και για προγενέστερα sessions, καθώς ο χρήστης έχει χρησιμοποιήσει στο παρελθόν την εφαρμογή.
-

Στην Εικόνα 5 παρουσιάζεται γραφικά ο τρόπος με τον οποίο ορίζονται το session-based και το session-aware πρόβλημα, σύμφωνα με το [15]

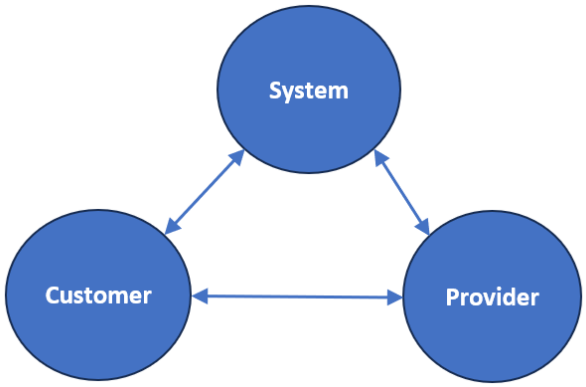


Εικόνα 5-Session-based vs Session-aware RS

Όσον αφορά το multistakeholder recommender πρόβλημα, σε αυτό το σενάριο η συνάρτηση χρησιμότητας f που ορίστηκε παραπάνω δεν λαμβάνει υπόψιν μόνο τους χρήστες, αλλά και τον οργανισμό που υλοποιεί το σύστημα συστάσεων καθώς και άλλα εμπλεκόμενα μέρη

(stakeholders). Σκοπός του συστήματος είναι να μεγιστοποιήσει την συνάρτηση f για όλα τα εμπλεκόμενα μέρη εξισορροπώντας τις ανάγκες τους [16]. Η συγκεκριμένη προσέγγιση μοιάζει εμφανώς πιο ρεαλιστική και αναπαριστά με μεγαλύτερη ακρίβεια το πως λειτουργεί η αγορά στον πραγματικό κόσμο, ιδίως όταν οι οργανισμοί λειτουργούν ως πλατφόρμες ηλεκτρονικών αγορών (marketplaces), όπως eBay, Amazon, Backmarket, Skroutz κ.α.

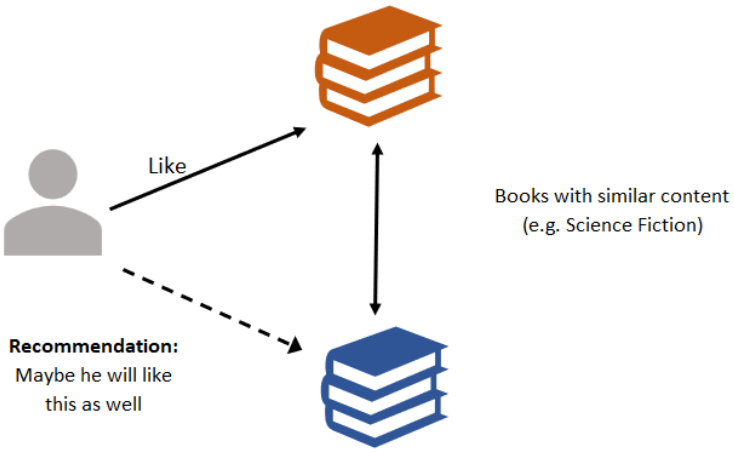
Ακολουθώντας το παράδειγμα των marketplaces, μπορούμε να θεωρήσουμε τις εξής τρεις βασικές οντότητες – stakeholders: σύστημα (system), πάροχος (provider) και πελάτης (customer) οι οποίες παρουσιάζονται στην Εικόνα 6. Οι πελάτες ταυτίζονται ουσιαστικά με τους χρήστες στο παραδοσιακό πρόβλημα των συστάσεων, καθώς συνδέονται στο σύστημα με σκοπό να αγοράσουν προϊόντα και αναμένουν να λάβουν προτάσεις που θα ικανοποιήσουν τις ανάγκες τους. Οι πάροχοι είναι εκείνοι που προσφέρουν το προϊόν μέσω της πλατφόρμας-συστήματος και λαμβάνουν αξία (utility) εφόσον οι πελάτες τους επιλέξουν για τις αγορές τους. Τέλος, το σύστημα είναι η πλατφόρμα που φέρνει σε επαφή πελάτες και παρόχους, έχει υλοποιήσει το σύστημα συστάσεων και λαμβάνει αξία συνήθως υπό την μορφή προμήθειας από τα υπόλοιπα εμπλεκόμενα μέρη (συνήθως από τους providers).



Εικόνα 6- Multistakeholder overview

3.4 Τεχνικές φιλτραρίσματος βασισμένες στο περιεχόμενο

Τα συστήματα προτάσεων βασισμένα στο περιεχόμενο (content-based recommender systems - CBF) αναζητούν αντικείμενα παρόμοια με εκείνα που προτιμά ένας χρήστης, βασιζόμενα στην σύγκριση μεταξύ του περιεχομένου των αντικειμένων [17], ή πιο απλά, το σύστημα προτείνει αντικείμενα παρόμοια με αυτά που ο χρήστης εκτίμησε θετικά στο παρελθόν (Εικόνα 7).



Εικόνα 7- Content-based overview

Στην συγκεκριμένη προσέγγιση αξιοποιούνται, εκτός από τις αξιολογήσεις των χρηστών (πίνακας R), και μεταδεδομένα που αφορούν τα αντικείμενα. Για παράδειγμα, εάν οι συστάσεις γίνονται για ταινίες, κάθε ταινία μπορεί να χαρακτηρίζεται από το είδος (π.χ. δράσης, ρομαντική, κλπ.), τους ηθοποιούς, τον σκηνοθέτη, τη διάρκεια, την ημερομηνία κυκλοφορίας και τη γλώσσα. Αυτά τα χαρακτηριστικά θεωρείται ότι περιγράφουν επαρκώς το αντικείμενο και κατά συνέπεια χρησιμοποιούνται για να δημιουργηθεί μια αναπαράσταση του τελευταίου. Έπειτα, οι αναπαραστάσεις των αντικειμένων χρησιμοποιούνται για να συνθέσουν το προφίλ ενός χρήστη με βάση τις προηγούμενες προτιμήσεις του. Τέλος, οι προτάσεις γίνονται με βάση το προφίλ του χρήστη, προσπαθώντας να βρεθούν αντικείμενα που ταιριάζουν σε αυτό.

Πιο συγκεκριμένα και σύμφωνα με το [18] τα CBF συστήματα μοιράζονται εν γένει την παρακάτω αρχιτεκτονική:

- Content Analyzer: Ο σκοπός του συγκεκριμένου βήματος είναι να αναπαραστήσει το περιεχόμενο των αντικειμένων (π.χ. έγγραφα, ιστοσελίδες, ειδήσεις, ταινίες) με τρόπο κατάλληλο για τα επόμενα βήματα επεξεργασίας. Χρησιμοποιούνται τεχνικές εξαγωγής χαρακτηριστικών (feature extraction) ώστε να μετατραπεί η πληροφορία από την αρχική της μορφή στην επιθυμητή, ώστε να λειτουργήσει ως μεταβλητή εσόδου μετέπειτα. Για παράδειγμα, μια ιστοσελίδα αναπαρίσταται ως ένα διάνυσμα λέξεων-κλειδιών.
- Profile Learner: Σε αυτό το βήμα αξιοποιούνται δεδομένα που αντιπροσωπεύουν τις προτιμήσεις του χρήστη, όπως ο πίνακας αξιολογήσεων, σε συνδυασμό με το περιεχόμενο των αντικειμένων που άρεσαν ή δεν άρεσαν στον χρήστη στο παρελθόν. Χρησιμοποιούνται τεχνικές μηχανικής μάθησης με σκοπό να γενικευτούν αυτές οι προτιμήσεις και να κατασκευαστεί το προφίλ χρήστη. Για παράδειγμα, σε ένα σύστημα συστάσεων ιστοσελίδων συνδυάζονται τα διανύσματα θετικών και αρνητικών παραδειγμάτων ένα πρωτότυπο διάνυσμα που αντιπροσωπεύει το προφίλ χρήστη.
- Filtering Component: Το προφίλ χρήστη που προέκυψε από το προηγούμενο βήμα συγκρίνεται με τις αναπαραστάσεις των αντικειμένων που προέκυψαν στο πρώτο βήμα, συνήθως μέσω κάποιον μετρικών ομοιότητας, όπως ,για παράδειγμα, με βάση το

συνημίτονο των διανυσμάτων (cosine similarity). Με βάση την τιμή ομοιότητας που προκύπτει για κάθε ζεύγος αντικειμένου και προφίλ χρήστη γίνονται και οι αντίστοιχες προτάσεις.

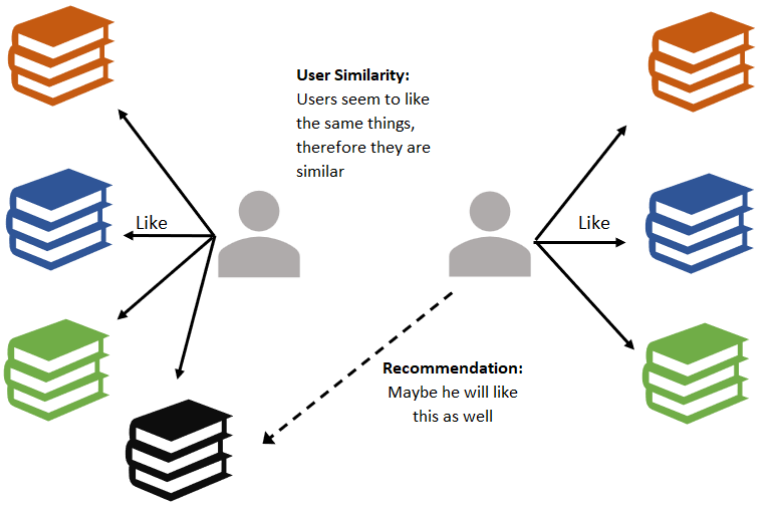
3.5 Συνεργατικό φιλτράρισμα

Ο όρος «συνεργατικό φιλτράρισμα» (collaborative filtering – CF) εμφανίστηκε πρώτη φορά στη βιβλιογραφία το 1992 στο πλαίσιο του συστήματος Tapestry, το οποίο αναπτύχθηκε από τους Goldberg, Nichols, Oki, και Terry. Ουσιαστικά, ο όρος CF περιγράφει τη διαδικασία κατά την οποία οι άνθρωποι συνεργάζονται για να φιλτράρουν και να βρουν χρήσιμη πληροφορία, καταγράφοντας τις αντιδράσεις τους είτε θετικές είτε αρνητικές σε διάφορα έγγραφα ή αντικείμενα. Οι αντιδράσεις αυτές, που μπορεί να είναι αξιολογήσεις ή σχολιασμοί, χρησιμοποιούνται στη συνέχεια από άλλους χρήστες για να φιλτράρουν το περιεχόμενο και να βρουν πληροφορίες που πιθανώς τους ενδιαφέρουν [19].

Η κεντρική ιδέα στην οποία βασίζεται το CF είναι ότι οι χρήστες που είχαν παρόμοιες προτιμήσεις στο παρελθόν θα συνεχίσουν να έχουν παρόμοιες προτιμήσεις και στο μέλλον. Σε αντίθεση με την content-based προσέγγιση που κάνει χρήση δεδομένων σχετικών με το περιεχόμενο των αντικειμένων, τα CF συστήματα στηρίζονται εξ ολοκλήρου στον πίνακα αξιολογήσεων R για να πραγματοποιήσουν τις προτάσεις. Οι προσεγγίσεις CF διαιρούνται σε δύο βασικές κατηγορίες: memory-based και model-based οι οποίες και αναλύονται στην συνέχεια.

3.5.1 Συνεργατικό φιλτράρισμα βασισμένο στην μνήμη

Στα συστήματα που βασίζονται στην μνήμη (memory-based), δεν γίνεται υπόθεση σχετικά με κάποιο μοντέλο που εξηγεί τον τρόπο με τον οποίο οι χρήστες αλληλοεπιδρούν με τα αντικείμενα. Αντίθετα, οι αλγόριθμοι αυτής της κατηγορίας ορίζουν συνήθως ένα μέτρο ομοιότητας μεταξύ των χρηστών ή των αντικειμένων και στη συνέχεια χρησιμοποιούν αναζήτηση των κοντινότερων γειτόνων για να πραγματοποιήσουν τις προτάσεις. Ανάλογα με την διάσταση στην οποία ορίζεται η μετρική ομοιότητας, τα συστήματα αυτά χαρακτηρίζονται ως User-based ή Item-based (εναλλακτικά ονομάζονται και User-User ή Item-Item). Στην Εικόνα 8 παρουσιάζεται γραφικά η λογική με την οποία λειτουργεί ένα user-user σύστημα.



Εικόνα 8- User-based Collaborative Filtering

Η ομοιότητα μεταξύ χρηστών ή αντικειμένων μπορεί να οριστεί με διάφορους τρόπους. Δύο από τους πιο συνηθισμένους στην βιβλιογραφία είναι ο συντελεστής συσχέτισης Pearson (Pearson correlation coefficient)και η ομοιότητα συνημίτονου (cosine similarity) και. Οι μετρικές αυτές χρησιμοποιούνται για να ορίσουν ομοιότητα επί διανυσμάτων σε πολυδιάστατους χώρους, στην συγκεκριμένη περίπτωση επί των σειρών (user-based) ή των στηλών (item-based) του πίνακα αξιολογήσεων R.

Ακολουθεί ο συντελεστής συσχέτισης Pearson μεταξύ δύο χρηστών u και v :

$$PCC(u,v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})(r_{vi} - \overline{r_v})}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \overline{r_v})^2}}$$

, όπου I_{uv} είναι το σύνολο των αντικειμένων που έχει βαθμολογηθεί και από τους δύο χρήστες, $\overline{r_u}$ και $\overline{r_v}$ είναι η μέση βαθμολογία των u και v αντίστοιχα για τα αντικείμενα που ανήκουν στο I_{uv} και r_{ui}, r_{vi} είναι οι βαθμολογίες για το ίδιο αντικείμενο i. Αντιστοίχως ορίζεται και ο τύπος για την ομοιότητα μεταξύ δύο αντικειμένων.

Ακολουθεί η ομοιότητα συνημίτονου μεταξύ δύο αντικειμένων i και j:

$$Cosine(i,j) = \frac{\sum_{u \in U_{ij}} r_{ui} r_{uj}}{\sqrt{\sum_{u \in U_i} {r_{ui}}^2} \sqrt{\sum_{u \in U_j} {r_{uj}}^2}}$$

, όπου U_i και U_j είναι τα σύνολα των χρηστών που έχουν βαθμολογήσει τα αντικείμενα i και j αντίστοιχα, U_{ij} το σύνολο των χρηστών που έχουν βαθμολογήσει και τα δύο αντικείμενα και r_{ui},

r_{uj} είναι οι βαθμολογίες του χρήστη u για τα αντικείμενα i και j αντίστοιχα . Αντιστοίχως ορίζεται και ο τύπος για την ομοιότητα μεταξύ δύο χρηστών.

Εφόσον οριστούν και υπολογιστούν τα μέτρα ομοιότητας, η πρόβλεψη των βαθμολογιών πραγματοποιείται συνήθως με κάποια μέθοδο κοντινότερων γειτόνων. Παραθέτουμε δύο από τους πιο συνηθισμένους τρόπους που συναντώνται στην βιβλιογραφία: weighted sum prediction [20] και mean-centered prediction [21] για την user-user προσέγγιση:

- weighted sum prediction : $\widetilde{r_{ui}} = \frac{\sum_{v \in N_u^i} Sim(u,v) * r_{vi}}{\sum_{v \in N_u^i} |Sim(u,v)|}$

- mean-centered prediction: $\widetilde{r_{ui}} = \overline{r_u} + \frac{\sum_{v \in N_u^i} Sim(u,v) * (r_{vi} - \overline{r_v})}{\sum_{v \in N_u^i} |Sim(u,v)|}$

, όπου N_u^i είναι το σύνολο των γειτόνων (κ-πλησιέστερων) με την μεγαλύτερη ομοιότητα με τον χρήστη u και ταυτόχρονα έχουν βαθμολογήσει το αντικείμενο i , v είναι ένας χρήστης που ανήκει στο N_u^i και $Sim(u, v)$ είναι η ομοιότητα (όπως για παράδειγμα ομοιότητα συνημίτονου) μεταξύ των χρηστών u και v. Για την mean-centered πρόβλεψη οι μεταβλητές παραμένουν ίδιες και τα $\overline{r_u}$, $\overline{r_v}$ είναι η μέση βαθμολογία για τους χρήστες u και v αντίστοιχα. Στην Εικόνα 9 παρουσιάζεται ένα διάγραμμα ροής που συνοψίζει την διαδικασία που ακολουθούν εν γένει τα memory-based συστήματα

Εικόνα 9 -Memory-based CF flowchart

Αξίζει να αναφερθεί το γεγονός ότι υπάρχουν διάφορες προσεγγίσεις οι οποίες συνδυάζουν τις user-based και item-based μεθόδους, όπως στο [22], όπου υπολογίζεται η ομοιότητα των χρηστών και των αντικειμένων χρησιμοποιώντας τον συντελεστή συσχέτισης Pearson. Η τελική πρόβλεψη προκύπτει ως σταθμικός μέσος (weighted average) των βαθμολογιών και των ομοιοτήτων που έχουν υπολογιστεί έναντι του αριθμού των δεδομένων που χρησιμοποιήθηκαν σε κάθε μέθοδο.

3.5.2 Συνεργατικό φιλτράρισμα βασισμένο σε μοντέλο

Το συνεργατικό φιλτράρισμα βασισμένο σε μοντέλο (model-based) χρησιμοποιεί τεχνικές μηχανικής μάθησης για την πρόβλεψη αλληλεπιδράσεων χρήστη-αντικειμένου. Σε αντίθεση με την memory-based προσέγγιση που στηρίζεται μόνο στα δεδομένα του πίνακα αξιολογήσεων R, η model-based προσέγγιση υποθέτει ένα μοντέλο που εξηγεί τις προτιμήσεων χρηστών. Το μοντέλο εκπαιδεύεται επί των δεδομένων του πίνακα R μέσω διάφορων τεχνικών, όπως είναι η παραγοντοποίηση πινάκων (matrix factorization), η συσταδοποίηση (clustering), οι τεχνικές

βαθιάς μάθησης (deep learning) κ.ά. Σε αυτή την κατηγορία ανήκουν κάποιες από τις πιο δημοφιλείς και επιτυχημένες τεχνικές των συστημάτων συστάσεων που χρησιμοποιούνται ευρέως στην βιομηχανία. Χαρακτηριστικό παράδειγμα αποτελεί η τεχνική παραγοντοποίησης πινάκων SVD (singular value decomposition), παραλλαγή της οποίας χρησιμοποιήθηκε από την νικήτρια ομάδα του διάσημου διαγωνισμού Netflix Prize που διοργάνωσε η πλατφόρμα Netflix το 2006 [23]. Ακολουθεί η παρουσίαση και ανάλυση ορισμένων τεχνικών model-based CF.

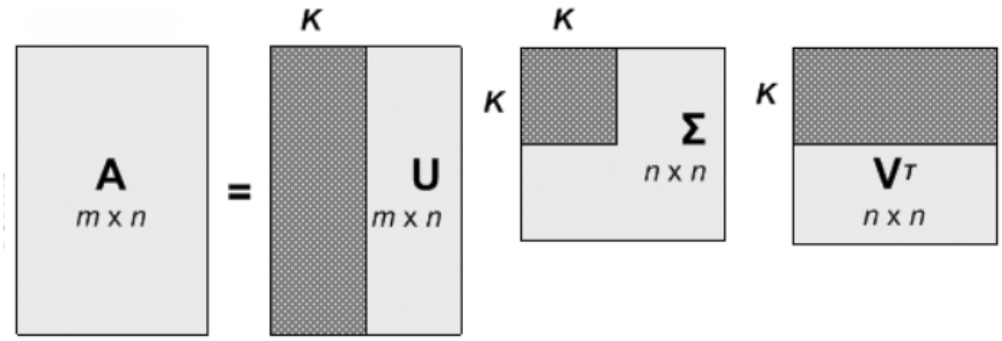
SVD (singular value decomposition)

Πρόκειται για μια μαθηματική μέθοδο που έχει τις ρίζες της στην γραμμική άλγεβρα και εφαρμόζεται ευρέως σε εφαρμογές μηχανικής μάθησης, καθώς χρησιμοποιείται συνήθως για την μείωση των διαστάσεων των επεξηγηματικών μεταβλητών. Η διαδικασία SVD αποσυνθέτει έναν πίνακα σε τρείς απλούστερους πίνακες, με συγκεκριμένες επιθυμητές μαθηματικές ιδιότητες. Πιο συγκεκριμένα, στα πλαίσια του CF , ένας πίνακας αξιολογήσεων $R^{m \times n}$ μπορεί να αναπαρασταθεί ως $R \approx U\Sigma V^T$ όπου:

- U είναι ένας ορθογώνιος πίνακας ($U^T U = UU^T = I$), διαστάσεων $m \times k$ και οι στήλες του αντιπροσωπεύουν τις κύριες συνιστώσες των χρηστών
- Σ είναι ένας διαγώνιος πίνακας , διαστάσεων $k \times k$, όπου τα στοιχεία της διαγώνιου είναι οι σιγματικές τιμές (singular values) διατεταγμένες σε φθίνουσα σειρά. Οι τιμές της στην διαγώνιο του πίνακα Σ αντιπροσωπεύουν την βαρύτητα ή σημασία των κύριων συνιστωσών των πινάκων U και V ως προς τον αρχικό πίνακα R
- V^T είναι ένας ορθογώνιος πίνακας, διαστάσεων $k \times n$ και οι γραμμές του αντιπροσωπεύουν τις κύριες συνιστώσες των αντικειμένων

Ο αλγόριθμος SVD λαμβάνει ως παράμετρο το k, δηλαδή το πλήθος των σιγματικών τιμών που θα χρησιμοποιηθούν. Όπως αναφέρεται παραπάνω, οι σιγματικές τιμές διατάσσονται σε φθίνουσα σειρά. Επιλέγοντας τις πρώτες k από αυτές ταυτόχρονα γίνεται επιλογή των k κύριων συνιστωσών με την μεγαλύτερη βαρύτητα, οι οποίες συνδυαστικά προσεγγίζουν σε ικανοποιητικό βαθμό τον αρχικό πίνακα R, όπως φαίνεται και στην Εικόνα 10. Με αυτόν τον τρόπο επιτυγχάνεται μείωση των διαστάσεων με τέτοιο τρόπο ώστε να διατηρείται η αρχική πληροφορία, ενώ παράλληλα οι νέοι πίνακες U, Σ και V γενικεύουν τις προτιμήσεις των χρηστών και μπορούν να χρησιμοποιηθούν για την πρόβλεψη των κενών στοιχείων του πίνακα αξιολογήσεων R.

Ένα μειονέκτημα της μεθόδου SVD είναι το γεγονός ότι δεν μπορεί να χειριστεί τις τιμές που λείπουν (missing values) στον πίνακα R. Μια λύση για το παραπάνω πρόβλημα είναι η συμπλήρωση (imputation) των τιμών που λείπουν, συνήθως με τον μέσο όρο ή με κάποια άλλη στατιστική τεχνική, όπως η γραμμική παλινδρόμηση (linear regression). Ο τρόπος αυτός είναι υπολογιστικά αργός και ενδεχομένως προβληματικός υπό την έννοια ότι μπορεί να μεταβάλλει την αρχική κατανομή των δεδομένων. Προτιμώνται συνεπώς, αριθμητικές μέθοδοι βελτιστοποίησης, όπως gradient descent ή alternating least squares οι οποίες εφαρμόζονται μόνο στις γνωστές βαθμολογίες, έτσι ώστε να προκύψει η καλύτερη δυνατή αποσύνθεση του πίνακα R.

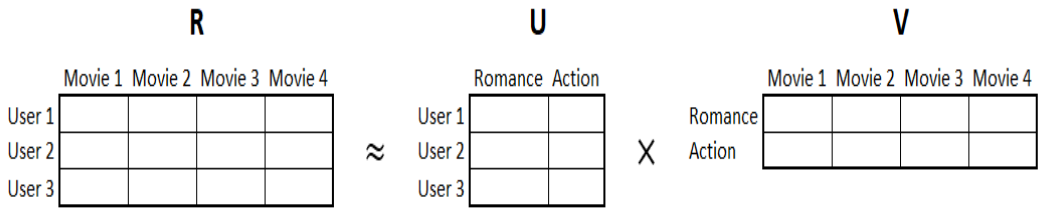


Εικόνα 10- SVD algorithm

NMF (non-negative matrix factorization)

Πρόκειται για μία μέθοδο παραγοντοποίησης πινάκων παρόμοια με την SVD, καθώς και σε αυτή την περίπτωση στόχος είναι το να εκφραστεί ο αρχικός πίνακας συστάσεων R σε μια πιο συνεπτυγμένη μορφή. Πιο συγκεκριμένα, ο πίνακας R αναλύεται σε δύο πίνακες μικρότερης διάστασης ως εξής: $R \approx UV$, όπου U και V είναι πίνακες με μη αρνητικά στοιχεία και διαστάσεων $m \times k$ και $k \times n$ αντίστοιχα. Όπως και για στην SVD προσέγγιση, έτσι και στον αλγόριθμο NMF χρησιμοποιούνται διάφορες επαναληπτικές διαδικασίες ώστε να βρεθούν οι κατάλληλοι πίνακες U και V. Μια από αυτές είναι ο αλγόριθμος gradient descent, ο οποίος όμως είναι αρκετά αργός όταν το μέγεθος του αρχικού πίνακα R είναι μεγάλο και επιπλέον η σύγκλιση εξαρτάται κατά μεγάλο βαθμό από την επιλογή του «βήματος» ως αρχική παράμετρος του αλγορίθμου. Για αυτό τον λόγο χρησιμοποιούνται πιο σύνθετες μέθοδοι, όπως πολλαπλασιαστικοί ή αθροιστικοί κανόνες ενημέρωσης [24]. Σε κάθε περίπτωση, η εύρεση των κατάλληλων πινάκων U και V ανάγεται στο πρόβλημα ελαχιστοποίησης της συνάρτησης κόστους $\|R - UV\|^2$, όπου $\|\cdot\|$ η ευκλείδεια νόρμα.

Διαισθητικά, θεωρούμε ότι ο πίνακας U εκφράζει τις προτιμήσεις των χρηστών και ο V τα χαρακτηριστικά των αντικειμένων σε μία μικρότερη διάσταση k. Για παράδειγμα, όπως παρουσιάζεται και στην Εικόνα 11, σε ένα σύστημα συστάσεων ταινιών οι γραμμές του πίνακα U αντιπροσωπεύουν τις προτιμήσεις κάθε χρήστη επί των k διαφορετικών ειδών (π.χ. δράσης, ρομαντική κ.λπ.) και αντίστοιχα οι στήλες του πίνακα V αντιπροσωπεύουν το κατά πόσο μια ταινία ανήκει σε ένα από τα k συγκεκριμένα είδη.



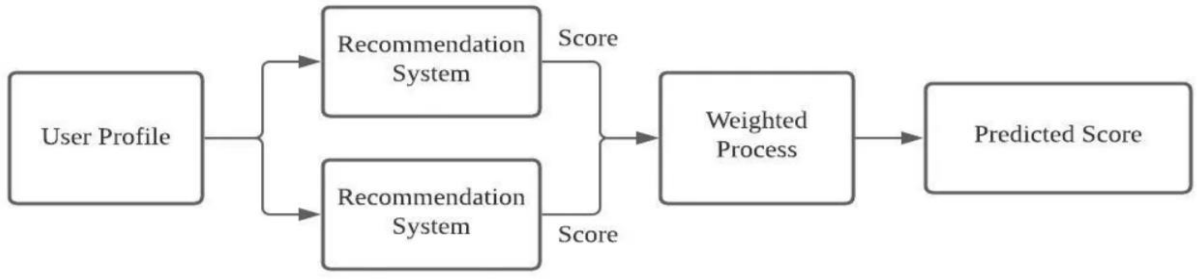
Εικόνα 11- NMF algorithm

3.6 Υβριδικές μέθοδοι

Οι προσεγγίσεις content-based και collaborative filtering που αναλύθηκαν στις προηγούμενες παραγράφους, διαθέτουν αμφότερες θετικά και αρνητικά στοιχεία, τα οποία και θα αναλυθούν στην συνέχεια με λεπτομέρεια. Γενικότερα, οποιαδήποτε προσέγγιση και αν ακολουθηθεί για την παραγωγή προτάσεων αποδίδει ικανοποιητικά κατά περίπτωση.Η σωστή προσέγγιση και μεθοδολογία εξαρτάται από πολλούς παράγοντες όπως η κατανομή των δεδομένων, η ύπαρξη ή μη μεταδεδομένων για την περιγραφή των αντικειμένων, η πυκνότητα ή αραιότητα του πίνακα αξιολογήσεων R κ.α. Το πρόβλημα των συστάσεων είναι πολύπλοκο και δεν υπάρχει μία "one size fits all" λύση, όπως και σε πολλά άλλα πεδία της μηχανικής μάθησης. Η επιλογή της κατάλληλης μεθόδου εξαρτάται από το είδος και την ποιότητα των διαθέσιμων δεδομένων, καθώς και από τις συγκεκριμένες ανάγκες και τους στόχους του συστήματος συστάσεων που αναπτύσσεται.

Ακολουθώντας την συλλογιστική της προηγούμενης παραγράφου, μια υβριδική προσέγγιση (hybrid approach) συνδυάζει διάφορες μεθόδους συστάσεων για να βελτιώσει την ακρίβεια και την ποιότητα των προτάσεων, καθώς και για να αντισταθμίσει τα μειονεκτήματά τους. Για παράδειγμα, ένας συνδυασμός CBF και CB μεθόδων μπορεί να εκμεταλλευτεί τις προσωπικές προτιμήσεις του χρήστη, καθώς και τις πληροφορίες που σχετίζονται με το περιεχόμενο των αντικειμένων. Ακολουθούν οι διάφορες τεχνικές υβριδοποίησης για RS όπως ορίζονται στο [25] καθώς και μια γραφική αναπαράσταση για την weighted προσέγγιση (Εικόνα 12)

- Weighted: Οι βαθμολογίες (ή ψήφοι) διαφόρων τεχνικών σύστασης συνδυάζονται για να παραχθεί μία μόνο σύσταση.
- Switching: Το σύστημα εναλλάσσεται μεταξύ των τεχνικών σύστασης ανάλογα με την τρέχουσα κατάσταση.
- Mixed: Παρουσιάζονται ταυτόχρονα συστάσεις από διάφορες διαφορετικές τεχνικές σύστασης.
- Feature Combination: Χαρακτηριστικά από διαφορετικές πηγές δεδομένων σύστασης συνδυάζονται σε έναν αλγόριθμο σύστασης.
- Cascade: Μία τεχνική σύστασης βελτιώνει τις συστάσεις που δίνονται από μία άλλη.
- Feature Augmentation: Η έξοδος από μία τεχνική χρησιμοποιείται ως χαρακτηριστικό εισόδου για μία άλλη.
- Meta-level: Το μοντέλο που μαθαίνεται από μία τεχνική σύστασης χρησιμοποιείται ως είσοδος για μία άλλη.



Εικόνα 12-Weighted Hybrid RS [26]

3.7 Μετρικές αξιολόγησης στα συστήματα συστάσεων

Όπως και σε οποιοδήποτε εφαρμογή της μηχανικής μάθησης, έτσι και στα συστήματα συστάσεων χρησιμοποιούνται διάφορες μετρικές για την αξιολόγηση τους (evaluation metrics) και πιο συγκεκριμένα για το κατά πόσο οι προτάσεις που παράγει το σύστημα είναι ικανοποιητικές. Οι μετρικές αξιολογήσεων εξυπηρετούν στην σύγκριση μεταξύ διαφορετικών προσεγγίσεων και τεχνικών, ενώ πολύ συχνά λειτουργούν ως baselines για ένα συγκεκριμένο dataset. Χαρακτηριστικά, στον διαγωνισμό Netflix Prize [23] ο στόχος που δόθηκε ήταν η μείωση της μετρικής RMSE κατά 10%. Ακολουθεί ο ορισμός και μια σύντομη ανάλυση για 7 από τα πιο συνηθισμένα μέτρα στον τομέα των RS. Στο σημείο αυτό θα πρέπει να σημειωθεί ότι τα RS δεν παράγουν προτάσεις μόνο υπό την μορφή αξιολογήσεων αλλά και ως διατεταγμένες λίστες, όπως για παράδειγμα τις κορυφαίες n-προτάσεις για έναν ενεργό χρήστη.

- Precision $= \frac{|\{\text{relevant items}\} \cap \{\text{recommended items}\}|}{|\{\text{recommended items}\}|}$, ή αλλιώς ακρίβεια, μετράει την αναλογία των σχετικών αντικειμένων σε σχέση με τον συνολικό αριθμό των προτεινόμενων αντικειμένων. Μεγαλύτερη τιμή ισοδυναμεί με καλύτερο αποτέλεσμα.

- Recall $= \frac{|\{\text{relevant items}\} \cap \{\text{recommended items}\}|}{|\{\text{relevant items}\}|}$, ή αλλιώς ανάκληση, μετράει την αναλογία των σχετικών αντικειμένων που προτείνονται επιτυχώς από το σύστημα σε σχέση με τον συνολικό αριθμό των σχετικών αντικειμένων.

- F1-Score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$, πρόκειται για τον αρμονικό μέσο των δύο παραπάνω μετρικών. Ακρίβεια και ανάκληση μεγιστοποιούνται η μια εις βάρος της άλλης, συνεπώς το F1-Score συνδυάζει τις δύο μετρικές με σκοπό την ταυτόχρονη μεγιστοποίηση τους.

- MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$, ή αλλιώς μέσο απόλυτο σφάλμα (mean absolute error). Πρόκειται για ένα κλασσικό μέτρο αξιολόγησης της απόδοσης αλγορίθμων και μετρά το μέσο μέγεθος των σφαλμάτων σε ένα σύνολο προβλέψεων.

- RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$, ή αλλιώς ρίζα μέσου τετραγωνικού σφάλματος (root mean square error), μοιάζει αρκετά με το MAE με την διαφορά ότι «τιμωρεί» περισσότερο τις μεγαλύτερες αποκλίσεις.

- MAPE $= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \times 100$,ή αλλιώς μέσο απόλυτο ποσοστιαίο σφάλμα (mean absolute percentage error). Εκφράζει το σφάλμα ως ποσοστό της πραγματική τιμής.

- Coverage $= \frac{|\{\text{recommended items}\}|}{|\{\text{total items}\}|}$, ή αλλιώς κάλυψη. Εκφράζει την αναλογία των αντικειμένων που το σύστημα μπορεί να προτείνει, έναντι του συνολικού αριθμού των αντικειμένων. Αποτελεί μια σημαντική μέτρηση για την αξιολόγηση της ικανότητας του συστήματος να εξυπηρετήσει μια ευρεία γκάμα προτιμήσεων.

3.8 Ζητήματα που αντιμετωπίζουν οι παραδοσιακές προσεγγίσεις

Ακολουθεί μια σύντομη παρουσίαση ορισμένων εκ των κύριων προβλημάτων που αντιμετωπίζουν οι προσεγγίσεις CB και CF. Όσον αφορά τα υβριδικά συστήματα και αυτά παρουσιάζουν εν πολλοίς τα κάτωθι ζητήματα, ως συνδυασμός των δύο μεθόδων, όχι όμως στον ίδιο βαθμό.

- Cold Start Problem: Το συγκεκριμένο πρόβλημα δυσκολία δημιουργίας ικανοποιητικών προτάσεων όταν υπάρχουν ανεπαρκή δεδομένα και έχει τρείς εκδοχές:
    - New User: Δυσκολία προτάσεων για χρήστες που μόλις έχουν εγγραφεί στην πλατφόρμα και έχουν λίγη ή καθόλου ιστορία αλληλεπιδράσεων.
    - New Item: Δυσκολία προτάσεων για νέα αντικείμενα που μόλις έχουν προστεθεί στην πλατφόρμα και δεν έχουν λάβει ακόμα αξιολογήσεις ή αλληλεπιδράσεις.
    - New System: Υπάρχουν ελάχιστες αλληλεπιδράσεις γενικότερα στο σύστημα καθώς είναι καινούργιο
- Data Sparsity Problem: Υπάρχει μεγάλο ποσοστό κενών ή ελλιπών δεδομένων σε ένα σύστημα συστάσεων, καθιστώντας δύσκολη την δημιουργία ικανοποιητικών προτάσεων. Το συγκεκριμένο πρόβλημα είναι ιδιαίτερα σύνηθες, καθώς το αντιμετωπίζουν σχεδόν όλα τα συστήματα μεγάλου μεγέθους. Για παράδειγμα, σε έναν τομέα ταινιών, είναι αναμενόμενο ότι ορισμένες, λίγες στο πλήθος ταινίες συγκριτικά με αυτές που διαθέτει το σύστημα, θα συγκεντρώνουν την πλειοψηφία των βαθμολογιών. Αυτό έχει ως αποτέλεσμα ο πίνακας αξιολογήσεων R να είναι συνήθως ένας πολύ αραιός (sparse) πίνακας.
- Gray Sheep Problem: Ο όρος gray sheep επινοήθηκε από τον Claypool [27] και αναφέρεται στους χρήστες που έχουν μοναδικά ή μη χαρακτηριστικά γούστα, τα οποία δεν ευθυγραμμίζονται εύκολα με τις προτιμήσεις της πλειοψηφίας των χρηστών.
- Stability vs Plasticity Problem: Αναφέρεται στο γεγονός ότι αν το σύστημα έχει μάθει τις προτιμήσεις ενός χρήστη και αυτές μεταβληθούν τότε θα δυσκολευτεί να αναγνωρίσει την αλλαγή αυτή και να πραγματοποιήσει τις αντίστοιχες προτάσεις. Συνεπώς, για να καλυφθεί αποφευχθεί το συγκεκριμένο σενάριο, θα πρέπει το σύστημα να διατηρεί τη γνώση που έχει ήδη μάθει (σταθερότητα), ενώ ταυτόχρονα να προσαρμόζεται σε νέα δεδομένα και (πλαστικότητα).

Οι προσεγγίσεις CF αντιμετωπίζουν και τα 4 παραπάνω προβλήματα. Αξίζει να σημειωθεί ότι οι model-based μέθοδοι αποδίδουν καλύτερα εν γένει σε σχέση με τις memory-based, ιδίως στο ζήτημα των αραιών δεδομένων στον πίνακα R- data sparsity problem. Οι CB προσεγγίσεις αντίθετα δεν αντιμετωπίζουν τα New-Item και Gray Sheep προβλήματα, αφενός γιατί στηρίζονται στην πληροφορία που περιγράφει τα αντικείμενα (άρα και ένα νέο αντικείμενο θα διαθέτει αυτή την πληροφορία) και αφετέρου διότι διαθέτουν την δυνατότητα προσαρμογής στο ιδιαίτερο προφίλ του κάθε χρήστη. Επιπλέον, οι CB μέθοδοι παράγουν προτάσεις οι οποίες είναι σχετικά εύκολο να ερμηνευθούν πως τον τρόπο με τον οποίον προέκυψαν, καθώς ο χρήστης λαμβάνει συστάσεις βάσει κατηγοριών που έχει εκτιμήσει θετικά στο παρελθόν. Ταυτόχρονα όμως, και για τον ίδιο λόγο, οι προτάσεις CB είναι αρκετά περιορισμένες σε σχέση με αυτές των CF συστημάτων, καθώς

«παγιδεύονται» σε προτάσεις με συγκεκριμένα χαρακτηριστικά, χάνοντας έτσι όσον αφορά την ποικιλία.

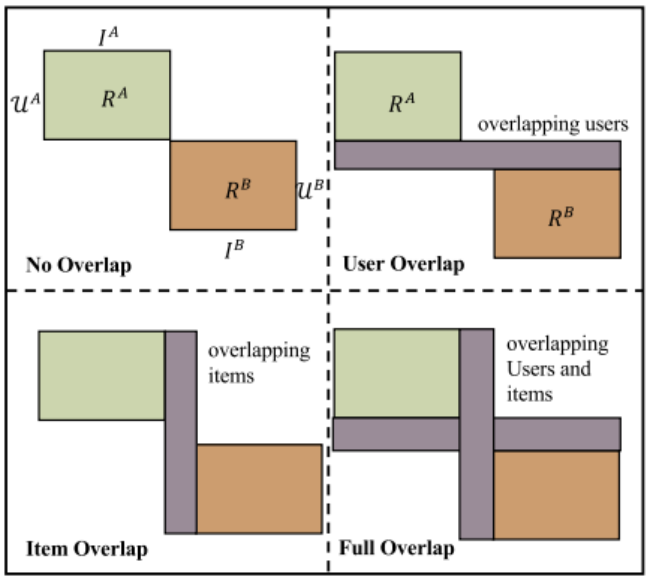
3.9 Μεταφορά γνώσης σε συστήματα συστάσεων

Τα cross-domain recommender systems (CDRS) αποτελούν ένα ταχέως αναπτυσσόμενο πεδίο των recommender systems, καθώς οι τεχνικές που εφαρμόζουν με την χρήση της μεταφοράς μάθησης βοηθούν στην αντιμετώπιση των προβλημάτων που αντιμετωπίζουν τα κλασικά RS και πιο συγκεκριμένα τα Data Sparsity και Cold Start προβλήματα. Πριν παρουσιαστούν οι διάφορες κατηγοριοποιήσεις των τεχνικών και οι εφαρμογές τους, ακολουθούν ενδεικτικά ορισμένοι συμβολισμοί που υιοθετούνται στα CDRS.

Υποθέτουμε δύο τομείς D^A και D^B, χωρίς βλάβη της γενικότητας, καθώς οι συμβολισμοί που ακολουθούν μπορούν εύκολα να επεκταθούν σε καταστάσεις με πολλαπλούς τομείς και τα αντίστοιχα σύνολα χρηστών $\mathcal{U}^A$,$\mathcal{U}^B$ και αντικειμένων $\mathcal{I}^A$,$\mathcal{I}^B$ για κάθε τομέα αντίστοιχα. Τότε οι πίνακες $R^A \in \mathbb{R}^{m^A \times n^A}$ και $R^B \in \mathbb{R}^{m^B \times n^B}$ αναπαριστούν τις αλληλεπιδράσεις μεταξύ χρηστών και αντικειμένων σε κάθε τομέα, όπου $m^A = |\mathcal{U}^A|$, $n^A = |\mathcal{I}^A|$,$m^B = |\mathcal{U}^B|$ και $n^B = |\mathcal{I}^B|$ δηλώνουν τον αριθμό των χρηστών και αντικειμένων στους αντίστοιχους τομείς. Κάθε στοιχείο των πινάκων αξιολόγησης $r_{mn} \in R^A$ ή $r_{mn} \in R^B$ μπορεί να είναι αριθμητικό, δηλώνοντας την ρητή αξιολόγηση ενός χρήστη για ένα αντικείμενο, ή δυαδικό, αντιπροσωπεύοντας μια έμμεση αλληλεπίδραση (π.χ., κλικ, αγορά, προσθήκη στο καλάθι) μεταξύ ενός ζεύγους χρήστη-αντικειμένου. Τέλος, σε κάθε τομέα ενδεχομένως να υπάρχουν δύο ζεύγη επιπλέον πινάκων $X^A \in R^{m^A \times p^A}$, $Y^A \in R^{n^A \times q^A}$ και $X^B \in R^{m^B \times p^B}$, $Y^B \in R^{n^B \times q^B}$που αναπαριστούν τα προφίλ χρηστών και τις ιδιότητες αντικειμένων, αντίστοιχα.

Τα CDRS παρουσιάζουν μεγαλύτερη πολυπλοκότητα σε σχέση με τα παραδοσιακά RS, όχι μόνο ως προς την υλοποίηση και τις διάφορες τεχνικές που χρησιμοποιούνται αλλά και ως προς τα σενάρια σύστασης. Μια από τις πρώτες κατηγοριοποιήσεις ως προς τα σενάρια σύστασης, τα οποία παρουσιάζονται στην Εικόνα 13, έχει δοθεί από τους Cremonesi et al. [28] και είναι η εξής:

- No Overlap: Δεν υπάρχουν κοινοί χρήστες αλλά ούτε και κοινά αντικείμενα μεταξύ των τομέων D^A και D^B , δηλαδή $\mathcal{U}^A \cap \mathcal{U}^B = \emptyset$ και $\mathcal{I}^A \cap \mathcal{I}^B = \emptyset$
- User Overlap: Υπάρχουν κάποιοι κοινοί χρήστες μεταξύ των τομέων αλλά όχι κοινά αντικείμενα, δηλαδή $\mathcal{U}^A \cap \mathcal{U}^B \neq \emptyset$ και $\mathcal{I}^A \cap \mathcal{I}^B = \emptyset$
- Item Overlap: Υπάρχουν ορισμένα κοινά αντικείμενα μεταξύ των τομέων αλλά όχι κοινοί χρήστες, δηλαδή $\mathcal{U}^A \cap \mathcal{U}^B = \emptyset$ και $\mathcal{I}^A \cap \mathcal{I}^B \neq \emptyset$
- User and Item Overlap: Υπάρχουν ορισμένα κοινά αντικείμενα και κάποιοι κοινοί χρήστες μεταξύ των τομέων, δηλαδή $\mathcal{U}^A \cap \mathcal{U}^B \neq \emptyset$ και $\mathcal{I}^A \cap \mathcal{I}^B \neq \emptyset$

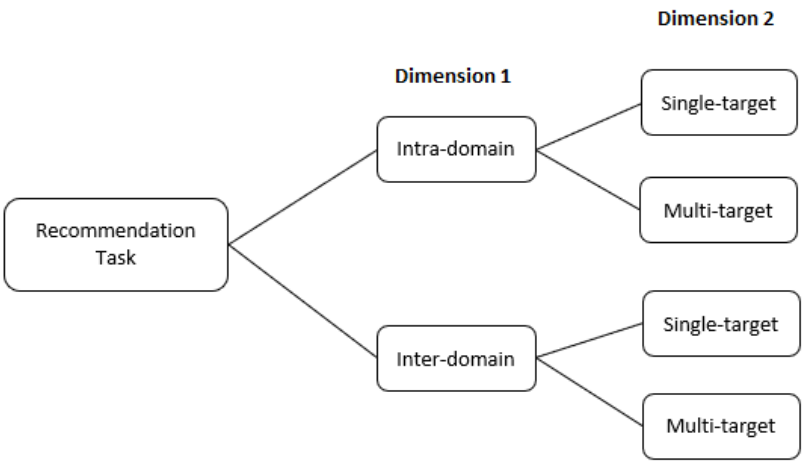


Εικόνα 13- Cross Domain Recommendation Scenarios, user & item overlap [29]

Οι Zang et al. [29] επέκτειναν την παραπάνω ταξινόμηση προσθέτοντας ένα δεύτερο επίπεδο το οποίο αφορά την κατηγοριοποίηση των εργασιών σύστασης. Τα σενάρια που σχετίζονται με τις εργασίες ταξινομούνται με την σειρά τους σε δύο διαστάσεις όπως παρουσιάζονται στην Εικόνα 14.

Πιο αναλυτικά, στην πρώτη διάσταση διαχωρίζονται οι εργασίες με βάση το αν τα προτεινόμενα αντικείμενα βρίσκονται στον ίδιο τομέα με τον χρήστη ή όχι. Στην σύσταση εντός τομέα (intra-domain) τα προτεινόμενα αντικείμενα βρίσκονται στον ίδιο τομέα με τον χρήστη. Αντίθετα, στην σύσταση εκτός τομέα ,τα προτεινόμενα αντικείμενα δεν ανήκουν στον ίδιο τομέα με τον χρήστη. Στην περίπτωση αυτή, ο χρήστης δεν έχει καμία αλληλεπίδραση με τα αντικείμενα για τα οποία λαμβάνει προτάσεις, συνεπώς τα συστήματα που παράγουν συστάσεις με αυτόν τον τρόπο επιχειρούν να επιλύσουν το cold-start πρόβλημα υπό το πρίσμα ενός νέου χρήστη.

Στην δεύτερη διάσταση γίνεται διαχωρισμός των εργασιών με βάση το αν παράγονται συστάσεις μόνο για έναν τομέα-στόχο ή για περισσότερους. Πιο συγκεκριμένα, στη σύσταση με έναν στόχο (single-target), συνήθως ο τομέας με τον πιο πυκνό πίνακα αλληλεπιδράσεων χρησιμοποιείται ως βοηθητικός τομέας (auxiliary domain) με σκοπό να βελτιωθεί η απόδοση των συστάσεων στον τομέα-στόχο, ο οποίος συνήθως είναι αρκετά αραιός. Αντίθετα, στη σύσταση σε πολλαπλούς στόχους (multi-target), δεν υπάρχει διάκριση πηγής-στόχου και ο απώτερος σκοπός είναι να βελτιωθούν οι συστάσεις ταυτόχρονα και στους δύο (ή περισσότεροι). Στην περίπτωση αυτή συνήθως και οι δύο τομείς έχουν αραιούς πίνακες αλληλεπιδράσεων και γίνεται η υπόθεση ότι οι δύο τομείς μπορούν να μάθουν ο ένας από τον άλλον.



Εικόνα 14- Cross Domain Recommendation Scenarios, recommendation task

Ακολουθεί η παρουσίαση κάποιον βασικών μεθοδολογιών που χρησιμοποιούνται στα CDRS ανά κατηγορία, δίνοντας έμφαση στην κατηγορία No-User/No-Item overlap καθώς σχετίζεται άμεσα με το πρόβλημα που πραγματεύεται η διπλωματική εργασία.

3.9.1 Μη αλληλοεπικάλυψη χρηστών και αντικειμένων

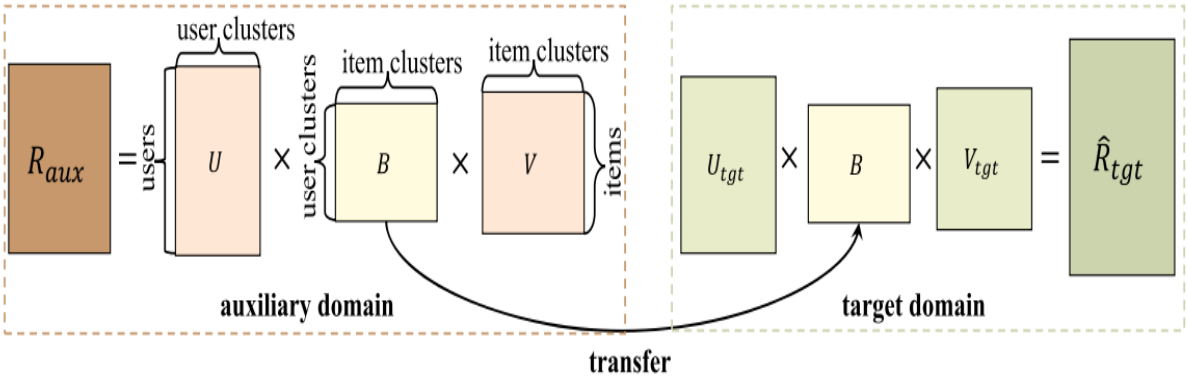
Στο συγκεκριμένο σενάριο δεν υπάρχουν κοινοί χρήστες μεταξύ των D^A και D^B ή ακόμα και αν υπάρχουν δεν είναι διαθέσιμη αυτή η πληροφορία σε εμάς. Το ίδιο ακριβώς ισχύει και όσον αφορά τα αντικείμενα. Σε αυτό το σενάριο εμφανίζονται τρείς κύριες κατηγορίες προσεγγίσεων οι οποίες είναι οι εξής: Εξαγωγή κοινών μοτίβων αξιολογήσεων σε επίπεδο συστάδων (cluster-level rating pattern), αποτύπωση συσχετίσεων ετικετών (capturing tag correlations), εφαρμογή ενεργητικής μάθησης (active learning).

Η αποτύπωση συσχετίσεων ετικετών βασίζεται στην υπόθεση ότι, παρόλο που οι χρήστες και τα αντικείμενα είναι διαφορετικά μεταξύ των τομέων, οι χρήστες μπορεί να χρησιμοποιούν τις ίδιες ή παρόμοιες ετικέτες για να περιγράψουν αντικείμενα ενδιαφέροντος και, παρομοίως, τα αντικείμενα σε διαφορετικούς τομείς ενδέχεται να χαρακτηρίζονται με τις ίδιες όσον αφορά τις ιδιότητες τους.

3.9.1.1 Εξαγωγή κοινών μοτίβων αξιολογήσεων σε επίπεδο συστάδων

Βασική υπόθεση της συγκεκριμένης προσέγγισης είναι ότι τα αντικείμενα που προσφέρονται στους τομείς D^A και D^B απευθύνονται στο γενικό πληθυσμό. Αυτό σημαίνει ότι οι χρήστες που αλληλοεπιδρούν με τα αντικείμενα αυτά ενδεχομένως να ανήκουν σε ομάδες που μοιράζονται παρόμοιες προτιμήσεις. Κάτι τέτοιο είναι ίσως αναμενόμενο σε ιδιαίτερα δημοφιλείς πλατφόρμες. Επιπλέον, παρόλο που τα αντικείμενα είναι διαφορετικά μεταξύ των τομέων, ενδεχομένως να μοιράζονται ορισμένες κοινές ιδιότητες. Για παράδειγμα, αν οι D^A και D^B αφορούν ταινίες και βιβλία αντίστοιχα, τότε ένα κοινό τους χαρακτηριστικό είναι το είδος (ρομαντικό, μυστηρίου κλπ.). Εν ολίγοις ,αν και δεν υπάρχουν αλληλεπικαλυπτόμενοι

χρήστες/αντικείμενα μεταξύ των τομέων, D^A και D^B δύο τομείς μπορεί να μοιράζονται μοτίβα αξιολογήσεων σε επίπεδο συστάδων (clusters). Επομένως, οι προσεγγίσεις αυτής της κατηγορίας στοχεύουν στην εξαγωγή αυτών των μοτίβων από έναν τομέα και στη μεταφορά τους στον άλλο τομέα, όπως φαίνεται και στην Εικόνα 15.



Εικόνα 15- Shared cluster-level rating pattern [29]

Η βασική μεθοδολογία για να επιτευχθεί η μεταφορά γνώσης μέσω κοινών μοτίβων αξιολογήσεων σε επίπεδο συστάδων ξεκινάει με την παραγοντοποίηση του πίνακα αξιολογήσεων του βοηθητικού τομέα R_{aux} σε τρείς πίνακες $U, V \; και \; S$ και συγκεκριμένα στους $U \in R_+^{m_A \times k}, V \in R_+^{n_A \times l}$, και $S \in R_+^{k \times l}$. Η παραγοντοποίηση αυτή γίνεται χρησιμοποιώντας τον αλγόριθμο ορθογώνιας μη αρνητικής τρι-παραγοντοποίησης (orthogonal nonnegative matrix tri-factorization -ONMTF), που είναι ισοδύναμος με τον αλγόριθμο k-means σε δύο διαστάσεις:

$$\min_{U \geq 0, V \geq 0, S \geq 0} || R_{aux} - USV^T ||_F^2$$
$$\text{όπου } U^T U = I, V^T V = I$$

Κάθε γραμμή των $U \; και \; V$ υποδηλώνει την συμμετοχή ενός χρήστη ή ενός αντικειμένου στις αντίστοιχες συστάδες. Το επόμενο βήμα είναι η κατασκευή ενός πίνακα B ο οποίος κωδικοποιεί τα μοτίβα αξιολογήσεων των k συστάδων για τους χρήστες και των l συστάδων για τα αντικείμενα τα οποία είναι κοινά και για τους δύο τομείς. Έπειτα, σκοπός είναι η εύρεση πινάκων U_{tgt} και V_{tgt} οι οποίοι υποδηλώνουν την συμμετοχή των χρηστών και των αντικειμένων στις συστάδες για τον target τομέα X_{tgt}. Αυτό γίνεται μεταφέροντας τον πίνακα B από το προηγούμενο βήμα στον τομέα στόχο ως εξής:

$$\min_{U_{tgt} \in \{0,1\}^{p \times k}, V_{tgt} \in \{0,1\}^{q \times l}} || \left(X_{tgt} - U_{tgt} B V_{tgt}^T \right) \circ W ||_F^2$$
$$\text{όπου } U_{tgt} 1 = 1, V_{tgt} 1 = 1$$

Στον παραπάνω τύπο, ο πίνακας W είναι ένας πίνακας 0-1 που χρησιμοποιείται για να καλύψει τις μη παρατηρούμενες καταχωρήσεις στον πίνακα αξιολογήσεων του X_{tgt}. Πιο συγκεκριμένα ισχύει $W_{ij} = 1$ αν το $\left[X_{tgt}\right]_{ij}$ έχει αξιολογηθεί και $W_{ij} = 0$ αν όχι. Επίσης, ο τελεστής $\circ$ συμβολίζει το στοιχειώδες γινόμενο (element-wise product) που σημαίνει ότι κάθε στοιχείο του πίνακα

πολλαπλασιάζεται ξεχωριστά με το αντίστοιχο στοιχείο ενός άλλου πίνακα. Τέλος οι προβλέψεις στον τομέα στόχο δημιουργούνται ως εξής:

$$\widetilde{X_{tgt}} = W \circ X_{tgt} + (1 - W) \circ \left(U_{tgt} B V_{tgt}^{T}\right)$$

Η παραπάνω μέθοδος προτάθηκε αρχικά από τους Li et.al [30] , όπου ο συμπαγής πίνακας B ονομάστηκε “codebook”. Ο πίνακας codebook κατασκευάζεται με τον μέσο όρο όλων των αξιολογήσεων σε κάθε co-cluster χρήστη-αντικειμένου ως εξής:
$B = [U_{aux}^{T} X_{aux} V_{aux}] \div [U_{aux}^{T} 1 1^{T} V_{aux}]$, όπου με ÷ συμβολίζεται η στοιχειώδης διαίρεση (entry-wise division).

Η μέθοδος “codebook” αποτέλεσε αφετηρία για την ανάπτυξη και άλλων τεχνικών που στηρίζονται στην παραπάνω προσέγγιση. Μια άλλη ιδιαίτερα επιδραστική εργασία ως προς τις συγκεκριμένες τεχνικές είναι το έργο των Si και Jin [31] , οι οποίοι υλοποιούν ένα ευέλικτο μοντέλο μίξης (Flexible Mixture Model - FMM) κάνοντας co-cluster χρήστες και αντικείμενα σε έναν τομέα χωρίς την υπόθεση ότι κάθε χρήστης και αντικείμενο ανήκουν μόνο σε ένα cluster και χρησιμοποιώντας τον αλγόριθμο EM (expectation maximization). Το μοντέλο FMM επεκτείνεται από τους Li et.al [32] ως Rating Matrix Generative Model (RMGM), καθώς εφαρμόζεται σε cross-domain σενάρια. Παρόμοια προσέγγιση ακολουθούν και οι Gao et.al [33] , οι οποίοι επεκτείνουν το “codebook” ώστε χειρίζεται περιπτώσεις όπου χρήστες και αντικείμενα ανήκουν σε παραπάνω από ένα cluster. Επιπλέον, χωρίζουν την γνώση σε domain specific και transferable και εφαρμόζουν περιορισμούς ως προς τι θα μεταφερθεί από τον έναν τομέα στον άλλο με στόχο να αποφύγουν την αρνητική μεταφορά γνώσης.

Μια άλλη πιο εξεζητημένη προσέγγιση προτείνεται από τους Zhang et.al [34] και ονομάζεται consistent knowledge transfer by subspace alignment (CKTSA). Στην συγκεκριμένη μέθοδο γίνεται co-cluster χρηστών και αντικειμένων σε κάθε τομέα ξεχωριστά και έπειτα για να αντιμετωπιστεί η απόκλιση των οριακών κατανομών μεταξύ των τομέων εφαρμόζεται subspace alignment. Ουσιαστικά, χρησιμοποιείται η τεχνική PCA (principal component analysis) ώστε να βρεθούν αναπαραστάσεις των χώρων των χρηστών και αντικειμένων σε χαμηλότερες διαστάσεις. Έπειτα οι υπόχωροι αυτοί χρησιμοποιούνται για την εύρεση πινάκων μετατροπής, τέτοιους ώστε οι κατανομές των υποχώρων να έλθουν πιο κοντά, τόσο για τους χρήστες όσο και για τα αντικείμενα. Στο επόμενο βήμα γίνεται εξαγωγή της κοινής γνώσης μέσω ελαχιστοποίησης συνάρτησης κόστους που υπολογίζει την διαφορά των πραγματικών πινάκων αξιολόγησης και των αντίστοιχων προσεγγίσεων. Το τελευταίο βήμα, πριν την παραγωγή συστάσεων, εφαρμόζεται feature representation regulation, όπου ο στόχος είναι η τροποποίηση των αναπαραστάσεων του πίνακα αξιολογήσεων στον τομέα στόχο ώστε το μοντέλο να ταιριάζει καλύτερα. Το συγκεκριμένο βήμα γίνεται διότι θεωρείται ότι ορισμένα χαρακτηριστικά που είναι ειδικά για τον source τομέα ενσωματώνονται στη μικρή ποσότητα διαθέσιμων δεδομένων στον πίνακα αξιολογήσεων του τομέα στόχου.

Παρόμοια λογική ακολουθήθηκε και στο έργο των ίδιων συγγραφέων [35] με την διαφορά ότι στο δεύτερο βήμα χρησιμοποιήθηκε η τεχνική geodesic flow kernel (GFK), η οποία είναι μια domain adaptation τεχνική, ώστε να έρθουν κοντά οι οριακές κατανομές χρηστών και αντικειμένων.

3.9.2 Αλληλοεπικάλυψη χρηστών ή/και αντικειμένων

Όταν υπάρχει αλληλοεπικάλυψη χρηστών ή/και αντικειμένων τότε, όπως είναι αναμενόμενο υπάρχουν περισσότερες επιλογές όσον αφορά τις μεθόδους που χρησιμοποιούνται για την μεταφορά γνώσης, όπως προσεγγίσεις βασισμένες σε γράφους, deep learning προσεγγίσεις, factorization machines κ.α.

Μια από τις πιο συνηθισμένες τεχνικές, όταν υπάρχουν κοινοί χρήστες και όχι κοινά αντικείμενα, είναι η collective matrix factorization, η οποία αποτελεί επέκταση των παραδοσιακών μεθόδων παραγοντοποίησης πινάκων αξιολογήσεων. Η συγκεκριμένη προσέγγιση προτάθηκε για πρώτη φορά από τους Singh και Gordon [36] όπου παραγοντοποιούνται συλλογικά οι πίνακες αξιολογήσεων R_A και R_B των δύο τομέων σε αναπαραστάσεις χρηστών U_A και U_B και αναπαραστάσεις αντικειμένων V_A και V_B με τον περιορισμό οι αναπαραστάσεις των χρηστών είναι κοινές στους διαφορετικούς τομείς, δηλαδή να ισχύει $U_A = U_B$.

4. ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ ΚΑΙ ΕΡΓΑΛΕΙΑ ΕΠΙΛΥΣΗΣ

4.1 Ορισμός προβλήματος

Αφού αναλύθηκαν η μεταφορά γνώσης και τα συστήματα συστάσεων στα προηγούμενα κεφάλαια, ο στόχος είναι να συνδυαστούν οι δύο αυτές έννοιες σε μια πρακτική εφαρμογή. Πιο συγκεκριμένα, το πρόβλημα που εξετάζεται είναι η προσπάθεια μεταφοράς γνώσης μεταξύ δύο διαφορετικών τομέων: ταινίες και βιβλία. Σκοπός είναι να ερευνηθεί κατά πόσο η μεταφορά γνώσης μπορεί να βελτιώσει τις συστάσεις σε αυτούς τους τομείς.

Τα σύνολα δεδομένων που χρησιμοποιούνται για τους τομείς ταινίες και βιβλία είναι τα MovieLens 25M [37] και Goodbooks-10k [38] αντίστοιχα. Ακολουθεί μια σύντομη περιγραφή των συνόλων δεδομένων:

- MovieLens 25M: Περιλαμβάνει 25 εκατομμύρια αξιολογήσεις ταινιών, από περίπου 162,000 χρήστες. Εκτός από τις αξιολογήσεις, το σύνολο δεδομένων περιέχει επίσης μεταδεδομένα για τις ταινίες, όπως είδος, τίτλο και έτος κυκλοφορίας. Είναι ευρέως χρησιμοποιούμενο στην έρευνα για συστήματα συστάσεων λόγω του μεγέθους και της ποιότητάς του.
- Goodbooks-10k: Περιέχει αξιολογήσεις βιβλίων για 10,000 τίτλους από περίπου 53,000 χρήστες. Συνολικά, περιλαμβάνει περίπου 6 εκατομμύρια αξιολογήσεις. Και το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει metadata για τα βιβλία, όπως τίτλο, συγγραφέα, είδος και έτος έκδοσης

Η υλοποίηση που ακολουθείται για την μεταφορά γνώσης είναι ανάλογη με αυτήν που προτείνεται στην εργασία Transfer Learning for Collaborative Filtering via a Rating-Matrix Generative Model (RMGM) [32]. Πρόκειται για μια cross-domain collaborative filtering μέθοδο η οποία μεταφέρει γνώση μεταξύ διαφορετικών, αλλά σχετικών μεταξύ τους τομέων, εντοπίζοντας κοινά μοτίβα αξιολογήσεων. Η συγκεκριμένη μέθοδος παράγει συστάσεις ταυτόχρονα και στους δύο τομείς χωρίς να προϋποθέτει την ύπαρξη κοινών χρηστών ή αντικειμένων. Επιπλέον, παρόλο που τα παραπάνω σύνολα δεδομένων περιέχουν metadata δεν θα γίνει χρήση αυτών, καθώς η μέθοδος δουλεύει μόνο με τους πίνακες αξιολογήσεων.

4.2 Εργαλεία επίλυσης

Ακολουθεί μια σύντομη αναφορά στα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση της διπλωματικής εργασίας.

4.2.1 Περιβάλλον υλοποίησης

Η εργασία υλοποιήθηκε στο εργαλείο Spyder (Scientific Python Development Environment), το οποίο είναι ένα περιβάλλον ανάπτυξης ανοιχτού κώδικα για τη γλώσσα προγραμματισμού Python. Το Spyder είναι ειδικά διαμορφωμένο για να διευκολύνει στην εκτέλεση επιστημονικών υπολογισμών και στην ανάλυση δεδομένων, καθώς παρέχει πολλαπλά παράθυρα για τον κώδικα, τα δεδομένα και τα αποτελέσματα. Τέλος, χρησιμοποιήθηκε η έκδοση 3.10 της γλώσσας προγραμματισμού Python διανεμημένη από την πλατφόρμα Anaconda.

4.2.2 Βιβλιοθήκες Python

Numpy: Πρόκειται για μια βιβλιοθήκη ειδικά σχεδιασμένη για την υποστήριξη επιστημονικών υπολογισμών. Πιο συγκεκριμένα η βιβλιοθήκη υποστηρίζει διανύσματα και πολυδιάστατους πίνακες, προσφέρει μαθηματικές συναρτήσεις και είναι ιδιαίτερα ισχυρή στην εκτέλεση πράξεων γραμμικής άλγεβρας. Αποτελεί, επίσης την βάση πάνω στην οποία χτίζονται πολλές άλλες βιβλιοθήκες επιστημονικού υπολογισμού.

Pandas: Πρόκειται για την πιο δημοφιλή βιβλιοθήκη όσον αφορά τη διαχείριση και ανάλυση δεδομένων σε Python, καθώς παρέχει ιδιαίτερα εύχρηστες δομές δεδομένων, όπως το DataFrame. Χρησιμοποιείται, μεταξύ άλλων, για την φόρτωση, τον καθαρισμο , τον μετασχηματισμό και την ανάλυση συνόλων δεδομένων. Τέλος, είναι βασισμένη στην βιβλιοθήκη Numpy, κάτι που την καθιστά ιδιαίτερα αποδοτική στην διαχείριση μεγάλου όγκου δεδομένων.

Matplotlib: Βιβλιοθήκη η οποία χρησιμοποιείται για την οπτικοποίηση και παρουσίαση δεδομένων. Παρέχει ιδιαίτερη ευελιξία στον χρήστη όσον αφορά την κατασκευή και παραμετροποίηση των διαγραμμάτων.

Surprise: Είναι μια εξειδικευμένη βιβλιοθήκη για την ανάπτυξη και αξιολόγηση συστημάτων συστάσεων συνεργατικού φιλτραρίσματος, η οποία παρέχει ορισμένα δημοφιλή σύνολα δεδομένων και διάφορους έτοιμους προς χρήση αλγορίθμους πρόβλεψης (SVD, NMF, κ.α.). Τέλος, διαθέτει εργαλεία για την αξιολόγηση, ανάλυση και σύγκριση της απόδοσης των διάφορων αλγορίθμων.

4.3 Ανάλυση των συνόλων δεδομένων

Πριν προχωρήσουμε στην παρουσίαση και εφαρμογή του συστήματος συστάσεων θα πρέπει να γίνει πρώτα ανάλυση των δεδομένων. Η Επεξηγηματική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA) είναι μια κρίσιμη διαδικασία στην επιστήμη δεδομένων και τη στατιστική ανάλυση, καθώς βοηθάει στην κατανόηση της δομής των δεδομένων και παρέχει χρήσιμες πληροφορίες που μπορούν να χρησιμοποιηθούν για τη λήψη αποφάσεων και την περαιτέρω ανάλυση.

Επίσης, πριν γίνει η ανάλυση, είναι σημαντικό είναι να ελεγχθεί η ποιότητα των δεδομένων για ελλείπουσες τιμές (missing values) και ακραίες παρατηρήσεις (outliers) οι οποίες ενδεχομένως να επηρεάζουν τις περιγραφικές στατιστικές και τις κατανομές των δεδομένων δίνοντας μια εσφαλμένη εικόνα. Χαρακτηριστικά, στο σύνολο MovieLens [37] βρέθηκε χρήστης με 33K βαθμολογίες, ο οποίος και αφαιρέθηκε από το σύνολο, καθώς θεωρήθηκε ότι είτε η τιμή αυτή είναι λανθασμένη, είτε ότι πρόκειται για μια μεμονωμένη περίπτωση. Ταυτόχρονα, θα πρέπει να διατηρηθεί μια ισορροπία μεταξύ του «καθαρισμού» των δεδομένων και του να μην επηρεαστεί η πραγματική τους κατανομή. Εν κατακλείδι, δεν βρέθηκαν missing values και εκτός την απόρριψη του συγκεκριμένου χρήστη που αναφέρθηκε προηγουμένως δεν τροποποιήθηκαν περεταίρω τα αρχικά σύνολα.

Statistic	Average of Ratings	Median of Ratings	STD of Ratings	Average Ratings per User	Median Ratings per User	Average Ratings per Item	Median Ratings per Item	# of Unique Users	# of Unique Items	Non-zero Ratings Ratio (%)
MovieLens	3.68	4	1.06	153.81	71	423.39	6	162541	59047	0.26
Goodbooks	3.92	4	0.99	111.87	111	597.65	248	53424	10000	1.12

Εικόνα 16-Basic Statistics for MovieLens and Gookdbooks

Στην Εικόνα 16 παρουσιάζονται κάποια βασικά στατιστικά στοιχεία των δύο συνόλων δεδομένων. Στο σημείο αυτό θα πρέπει να αναφερθεί ότι οι βαθμολογίες και στα δύο σύνολα είναι στην κλίμακα 1-5, με την διαφορά ότι στο σύνολο MovieLens επιτρέπονται και βαθμολογίες με δεκαδικό. Για τον λόγο αυτό έγινε κανονικοποίηση ώστε να υπάρχουν μόνο ακέραιες βαθμολογίες:

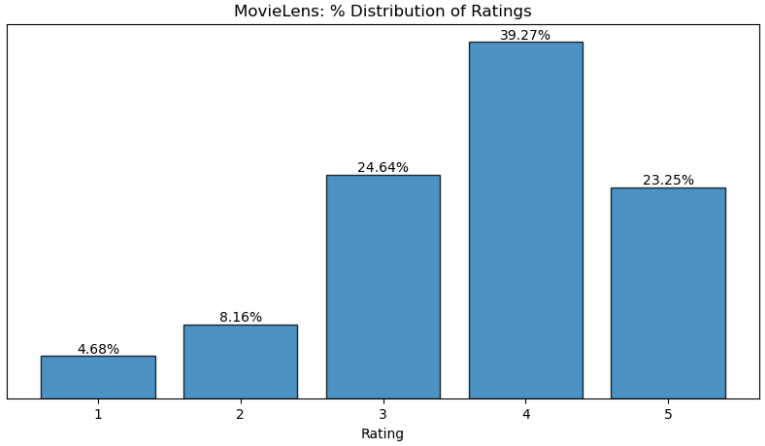
$$\text{normalized\_rating} = \left(\frac{\text{rating} - \text{min\_rating}}{\text{max\_rating} - \text{min\_rating}} \times 4 + 1 \right).round().astype(int)$$

Από την Εικόνα 16, προκύπτει ότι και στα δύο σύνολα η διάμεση βαθμολογία είναι μεγαλύτερη από την μέση τιμή, κάτι που υποδηλώνει αρνητική ασσυμετρία (left-skewed distribution) ως προς την κατανομή των βαθμολογιών. Η συγκεκριμένη παρατήρηση επιβεβαιώνεται και από τα διαγράμματα στις Εικόνες 17 και 18. Η εικόνα που παρουσιάζουν τα δεδομένα είναι συνηθισμένη στο πρόβλημα τον συστάσεων, καθώς γενικότερα οι χρήστες τείνουν να βαθμολογούν με υψηλό βαθμό τα αντικείμενα που προτιμούν, όπως είναι αναμενόμενο, αλλά αντίθετα, η συχνότητα των

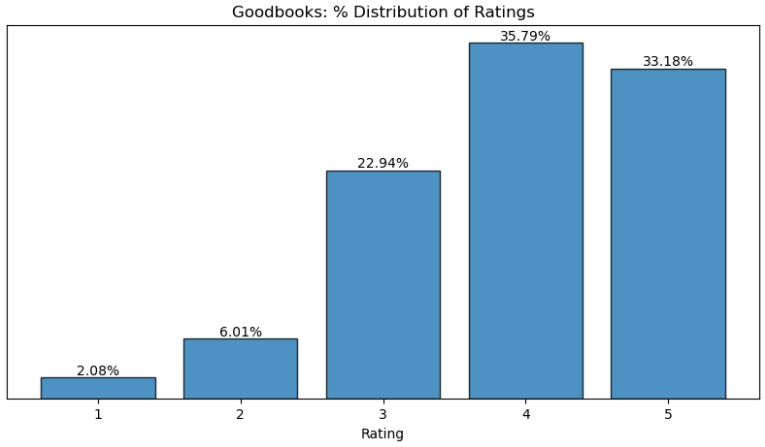
χαμηλών βαθμολογιών είναι μικρότερη. Αυτό συμβαίνει διότι οι χρήστες είναι πιο πιθανό να βαθμολογήσουν αντικείμενα που τους άρεσαν και να παραλείψουν την αξιολόγηση αντικειμένων που δεν τους άρεσαν ή δεν τους ενδιέφεραν αρκετά για να τα αξιολογήσουν. Τέλος, αξίζει να αναφερθεί ότι και τα δύο σύνολα δεδομένων είναι αραιά (sparse). Για παράδειγμα, στο σύνολο MovieLens, έχει βαθμολογία μόλις το 0,26% των δυνητικών αλληλεπιδράσεων χρηστών και ταινιών. Η έννοια της αραιότητας (sparsity) ορίζεται αυστηρά ως εξής:

sparsity= $(1 - \frac{\#\, ratings}{(\#\, users) \times (\#\, of\ items)}) \times 100\%$

και στην συγκεκριμένη περίπτωση τα δύο σύνολα δεδομένων έχουν 0,9974 (MovieLens) και 0,9888 (Goodbooks).



Εικόνα 17 - MovieLens Ratings Distribution

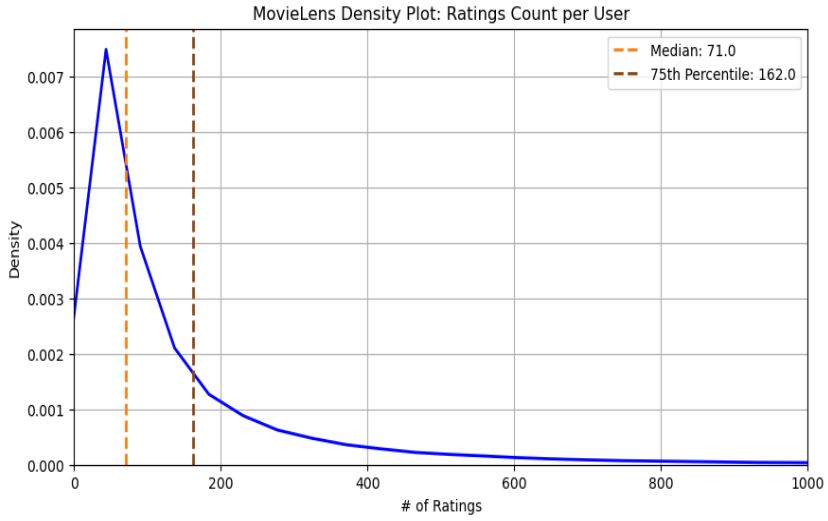


Εικόνα 18- Goodbooks Ratings Distribution

Συνεχίζοντας με την ανάλυση των δύο συνόλων, εξετάζουμε το πώς κατανέμεται ο αριθμός βαθμολογιών ανά χρήστη. Η συγκεκριμένη μεταβλητή είναι ιδιαίτερα σημαντική καθώς δείχνει τη δραστηριότητα των χρηστών και μπορεί να παρέχει πληροφορίες για την συμπεριφορά τους μέσα στην πλατφόρμα. Εξετάζουμε την συγκεκριμένη μεταβλητή μέσω των διαγραμμάτων πυκνότητας (density plots) που ακολουθούν (Εικόνες 19 και 20) . Ένα διάγραμμα πυκνότητας είναι μια γραφική απεικόνιση της κατανομής μιας συνεχούς μεταβλητής. Εμφανίζει την κατανομή των δεδομένων, δείχνοντας πού συγκεντρώνονται οι τιμές και πόσο συχνά εμφανίζονται διαφορετικές τιμές.

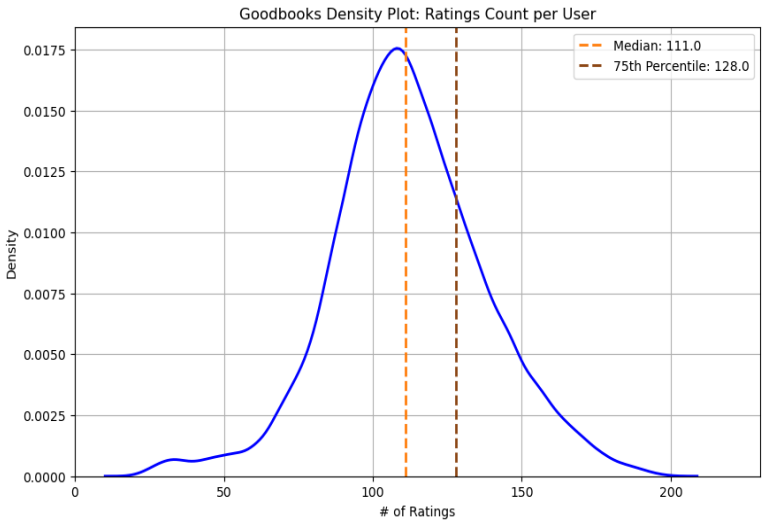
Το διάγραμμα πυκνότητας για το MovieLens δείχνει ότι η πλειοψηφία των χρηστών έχει αξιολογήσει λίγες ταινίες, με τη διάμεσο να είναι στις 71 αξιολογήσεις ανά χρήστη. Η πυκνότητα μειώνεται γρήγορα καθώς ο αριθμός των αξιολογήσεων αυξάνεται, δείχνοντας ότι υπάρχουν λίγοι χρήστες που έχουν αξιολογήσει πολλές ταινίες. Η ουρά του γραφήματος στα δεξιά υποδηλώνει την παρουσία λίγων χρηστών με υψηλό αριθμό αξιολογήσεων (δεξιά ασυμμετρία).

Στο Goodbooks, το διάγραμμα πυκνότητας παρουσιάζει μια πιο ισορροπημένη κατανομή, με τη διάμεσο να βρίσκεται στις 111 αξιολογήσεις ανά χρήστη. Η κατανομή είναι πιο συμμετρική και συγκεντρωμένη γύρω από τη διάμεσο, κάτι που υποδηλώνει ότι η πλειοψηφία των χρηστών αξιολογεί έναν αριθμό βιβλίων κοντά στη διάμεσο τιμή, με λιγότερες ακραίες τιμές.



.

Εικόνα 19- MovieLens: Ratings per User

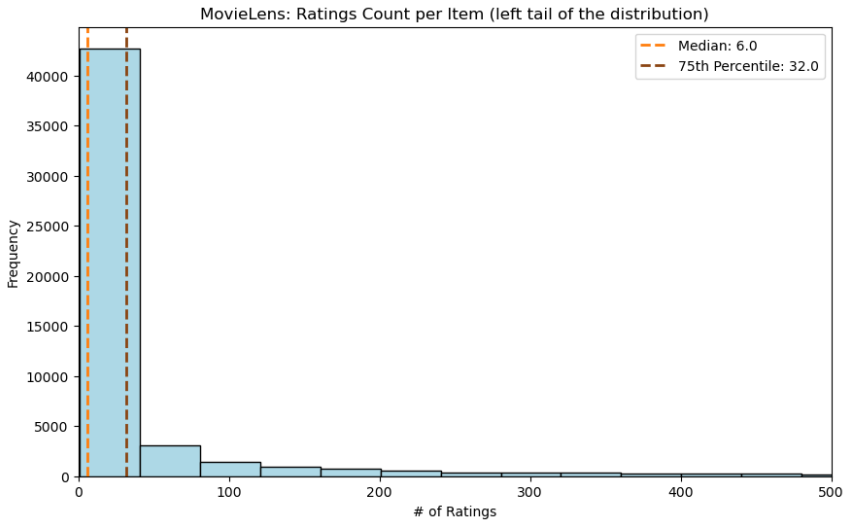


Εικόνα 20- Goodbooks Ratings per User

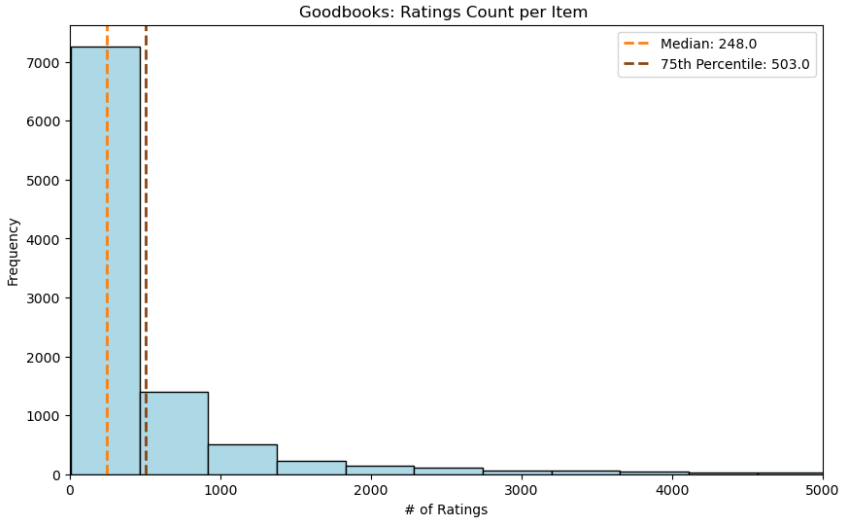
Τελευταίο βήμα στην ανάλυση αποτελεί την εξέταση της συμπεριφοράς των αντικειμένων. Πιο συγκεκριμένα, τα διαγράμματα συχνότητας που ακολουθούν (Εικόνα 21 και 22) απεικονίζουν το πως κατανέμονται οι βαθμολογίες ανά αντικείμενο για τα σύνολα δεδομένων MovieLens και Goodbooks αντίστοιχα.

Το διάγραμμα για το MovieLens δείχνει ότι η πλειοψηφία των ταινιών έχει πολύ λίγες βαθμολογίες, με τη διάμεσο στις 6 βαθμολογίες ανά ταινία. Αυτό υποδηλώνει ότι οι περισσότερες ταινίες αξιολογούνται από έναν μικρό αριθμό χρηστών, ενώ υπάρχουν λίγες ταινίες που λαμβάνουν πολλές βαθμολογίες. Η κατανομή είναι δεξιά ασύμμετρη, υποδεικνύοντας ότι λίγες ταινίες συγκεντρώνουν τη μεγάλη πλειοψηφία των βαθμολογιών.

Το διάγραμμα για το Goodbooks δείχνει ότι τα βιβλία λαμβάνουν πολύ περισσότερες βαθμολογίες κατά μέσο όρο σε σύγκριση με τις ταινίες στο MovieLens. Η διάμεσος είναι στις 248 βαθμολογίες ανά βιβλίο, και το 75ο ποσοστημόριο είναι στις 503 βαθμολογίες. Παρόλο που η κατανομή εξακολουθεί να είναι δεξιά ασύμμετρη, η ασυμμετρία είναι λιγότερο έντονη σε σχέση με το MovieLens, υποδεικνύοντας μια πιο ισορροπημένη κατανομή των βαθμολογιών μεταξύ των βιβλίων.



Εικόνα 21- MovieLens: Ratings per Item



Εικόνα 22-Goodbooks: Ratings per Item

5. ΜΕΘΟΔΟΛΟΓΙΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΜΕΛΕΤΗΣ

5.1 Λογική και υποθέσεις της μεθόδου

Η μέθοδος Rating-Matrix Generative Model (RMGM) [32] ανήκει στην κατηγορία των cross-domain collaborative filtering προσεγγίσεων και επιχειρεί να μεταφέρει γνώση μεταξύ δύο ή περισσότερων διαφορετικών μεταξύ τους τομέων για τους οποίους δεν υπάρχει αλληλοεπικάλυψη χρηστών και αντικειμένων. Ακόμα και αν υπάρχουν κοινοί χρήστες μεταξύ των τομέων, αυτή η γνώση δεν είναι διαθέσιμη σε εμάς. Επιπλέον, η συγκεκριμένη μέθοδος στηρίζεται αποκλειστικά στους πίνακες αξιολογήσεων και δεν κάνει χρήση μεταδεδομένων.

Η βασική υπόθεση της μεθόδου είναι ότι τομείς που έχουν κάποια σχέση μεταξύ τους, όπως τα βιβλία, η μουσική και οι ταινίες (αποτελούν έργα ψυχαγωγίας) ενδεχομένως να μοιράζονται παρόμοια κρυφά πρότυπα βαθμολόγησης. Πιο αναλυτικά, για δύο διαφορετικούς αλλά συγγενικούς τομείς:

- Τα αντικείμενα, παρόλο που είναι διαφορετικά, μπορεί να μοιράζονται κοινές ιδιότητες, όπως το είδος ή το θέμα. Για παράδειγμα, τόσο τα βιβλία όσο και οι ταινίες μπορεί να κατηγοριοποιούνται σε είδη όπως ρομαντικά, μυστήρια, δράματα κ.λπ.
- Οι χρήστες, παρόλο που είναι διαφορετικοί, αποτελούν δείγματα του γενικότερου πληθυσμού και άρα είναι πιθανό να μοιράζονται παρόμοια κοινωνικά χαρακτηριστικά και προτιμήσεις. Η υπόθεση αυτή έχει νόημα για δημοφιλείς πλατφόρμες.

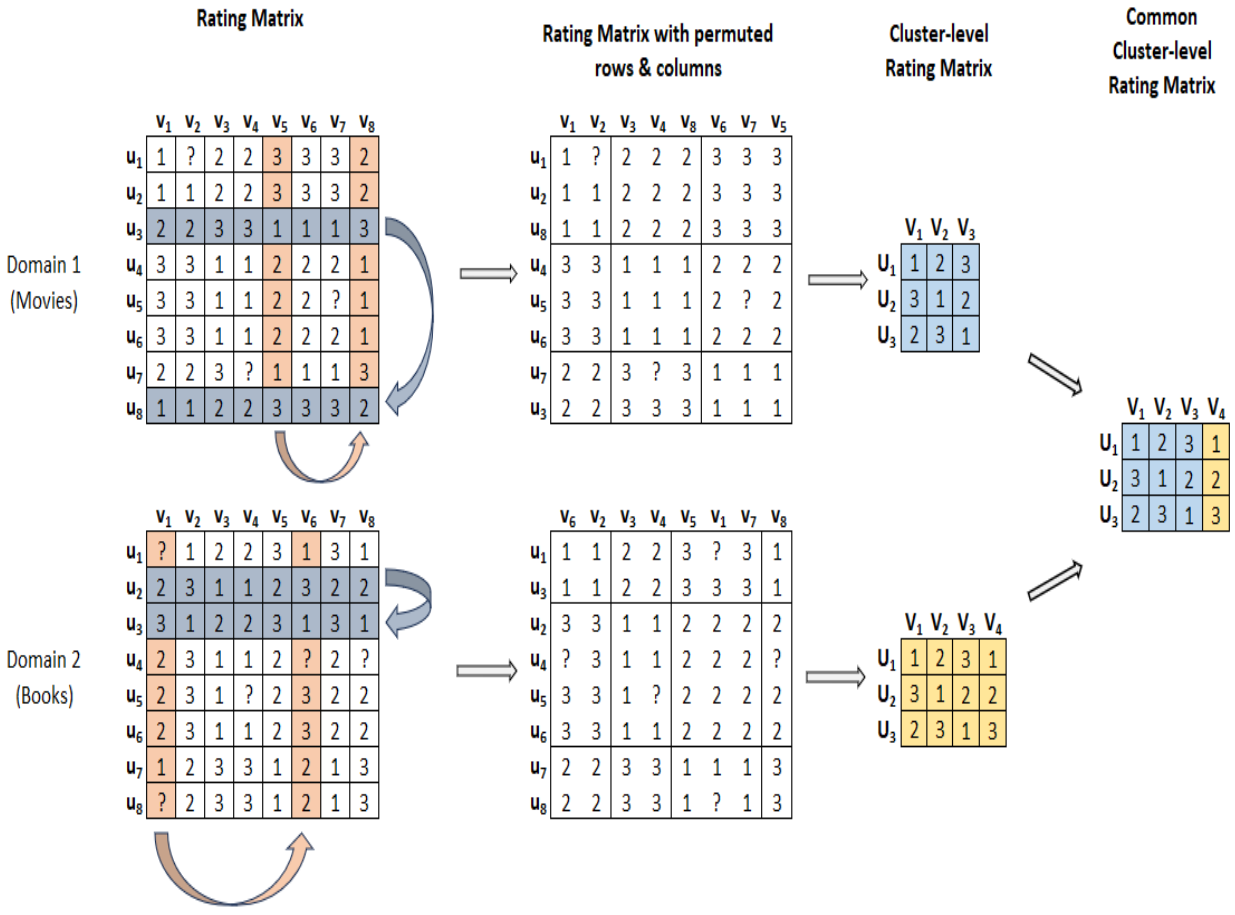
Ο τρόπος με τον οποίο, η μέθοδος RMGM επιχειρεί να αξιοποιήσει τα παραπάνω είναι μέσω της εύρεσης ενός κοινού προτύπου βαθμολόγησης σε επίπεδο συστάδων (clusters) μεταξύ των τομέων. Εφόσον το πρότυπο αυτό βρεθεί, τότε μπορεί να βοηθήσει στην παραγωγή προτάσεων και στην αντιμετώπιση των προβλημάτων cold start και sparsity που συνήθως αντιμετωπίζουν τα παραδοσιακά συστήματα συστάσεων.

Ακολουθεί ένα παράδειγμα με δύο τομείς μέσω του οποίου εξηγείται η λογική επί της οποίας στηρίζεται η προσέγγιση RMGM. Το παράδειγμα παρουσιάζεται γραφικά στην Εικόνα 23. Στο σημείο αυτό θα πρέπει να αναφερθεί ότι η Εικόνα 23 παρουσιάζει την γενική ιδέα πίσω από την προσέγγιση και δεν αποτελεί γραφική παράσταση του αλγορίθμου που εφαρμόζει το μοντέλο RMGM. Η ανάλυση του αλγορίθμου γίνεται σε επόμενη παράγραφο.

Το παράδειγμα της Εικόνας 23 ξεκινάει με δύο πίνακες αξιολογήσεων R_1 και R_2 που ανήκουν σε διαφορετικούς τομείς D_1 (ταινίες) και D_2 (βιβλία) αντίστοιχα. Τα στοιχεία των πινάκων αξιολογήσεων παριστάνουν τις βαθμολογίες (κλίμακα 1-5) που έχουν δώσει οι χρήστες u για τα αντικείμενα v . Πιο συγκεκριμένα, το στοιχείο R_{ij} ενός πίνακα αξιολογήσεων αντιπροσωπεύει την βαθμολογία που έχει δώσει ο χρήστης u_i για το αντικείμενο v_j. Σημαντικό είναι να αναφερθεί ότι οι χρήστες και τα αντικείμενα είναι διαφορετικά μεταξύ των δύο τομέων. Επίσης, ορισμένα κελιά στους πίνακες συμβολίζονται με ? , εννοώντας ότι το αντικείμενο δεν έχει βαθμολογηθεί από τον συγκεκριμένο χρήστη.

Αφού ορίστηκαν οι πίνακες που παρουσιάζονται στην Εικόνα 23 το επόμενο βήμα είναι η αναδιάταξη γραμμών και στηλών των πινάκων, ξεχωριστά σε κάθε τομέα. Η διαδικασία αυτή είναι ισοδύναμη με το co-clustering πινάκων, δηλαδή την ταυτόχρονη ομαδοποίηση των γραμμών και των στηλών του πίνακα. Αυτό σημαίνει ότι αντί να ομαδοποιούμε μόνο τις γραμμές (χρήστες) ή μόνο τις στήλες (αντικείμενα), το co-clustering προσπαθεί να βρει ομάδες γραμμών και ομάδες στηλών που έχουν παρόμοια πρότυπα. Καταλήγουμε έτσι σε δύο πίνακες οι οποίοι είναι πίνακες αξιολογήσεων σε επίπεδο συστάδων και πιο συγκεκριμένα για τον D_1 προκύπτουν 3 συστάδες χρηστών και αντικειμένων, ενώ για τον D_2 3 συστάδες χρηστών και 4 αντικειμένων.

Σε αυτό το σημείο μπορεί κανείς να παρατηρήσει ότι οι δύο ξεχωριστοί cluster-level πίνακες μοιάζουν αρκετά μεταξύ τους, κάτι το οποίο μας επιτρέπει να θεωρήσουμε ότι οι συστάσεις στους δύο τομείς ακολουθούν παρόμοιο μοτίβο, το οποίο και αποτυπώνεται συγκεντρωτικά ως κοινός πίνακας αξιολογήσεων σε επίπεδο συστάδων (common cluster-level Rating Matrix). Ο συγκεκριμένος πίνακας θεωρείται ότι αναπαριστά τα μοτίβα αξιολογήσεων και για τους δύο τομείς ταυτόχρονα.



Εικόνα 23- Shared Rating Pattern across domains

5.2 Μαθηματική μοντελοποίηση της μεθόδου

Στο παράδειγμα που προηγήθηκε παρουσιάστηκε ένα ιδεατό σενάριο όπου κάθε χρήστης ή αντικείμενο ανήκει σε μια μόνο συστάδα. Κάτι τέτοιο βεβαίως δεν συμβαίνει στην πραγματικότητα, καθώς ένας χρήστης ενδέχεται να έχει πολλές διαφορετικές προτιμήσεις ή να κατατάσσεται σε διαφορετικά κοινωνικά, δημογραφικά κλπ. σύνολα. Επιπλέον, ένα αντικείμενο μπορεί να ανήκει σε παραπάνω από μια κατηγορίες, όπως για παράδειγμα στις κατηγορίες "ρομαντική" και "οσκαρική". Για τον λόγο αυτό ο αλγόριθμος RMGM δεν δημιουργεί συστάδες με hard memberships, επιτρέποντας έτσι σε ένα χρήστη ή ένα αντικείμενο να ανήκουν σε παραπάνω από μια συστάδα με μια συγκεκριμένη πιθανότητα. Ακολουθεί ο αυστηρός μαθηματικός ορισμός της μεθόδου.

Υποθέτοντας Z τομείς σχετικούς μεταξύ τους, έχουμε Z διαφορετικούς πίνακες αξιολογήσεων. Για έναν τομέα z, έχουμε το σύνολο χρηστών $U_z = \{u_1^{(z)}, \dots, u_{n_z}^{(z)}\}$, το σύνολο αντικείμενών $V_z = \{v_1^{(z)}, \dots, v_{m_z}^{(z)}\}$ και τις βαθμολογίες r που έχουν δώσει οι χρήστες στα αντικείμενα. Συνεπώς ο τομέας μπορεί να αναπαρασταθεί ως εξής : $D_z = \{\left(u_1^{(z)}, v_1^{(z)}, r_1^{(z)}\right), \dots, \left(u_{s_z}^{(z)}, v_{s_z}^{(z)}, r_{s_z}^{(z)}\right)\}$. Οι βαθμολογίες για όλους τους τομείς θα πρέπει να είναι στην ίδια κλίμακα, για παράδειγμα $R = \{1, \dots, 5\}$

Υποθέτουμε ότι υπάρχουν K συστάδες για τους χρήστες $\{c_U^{(1)}, \dots, c_U^{(K)}\}$ και L συστάδες για τα αντικείμενα $\{c_V^{(1)}, \dots, c_V^{(L)}\}$. Οι συστάδες για τους χρήστες και τα αντικείμενα είναι οι ίδιες για όλους τους τομείς καθώς θεωρείται ότι οι Z τομείς έχουν παρόμοια μοτίβα αξιολογήσεων. Για παράδειγμα, το $c_U^{(1)}$ είναι η στοιβάδα 1 για τους χρήστες και αναφέρεται για τους χρήστες καθολικά, δηλαδή σε όλους τους τομείς.

Ορίζεται συνεπώς η πιθανότητα $P\left(c_U^{(k)}, c_V^{(l)} \middle| u, v\right)$ η οποία είναι η δεσμευμένη πιθανότητα του ζεύγους $c_U^{(k)}, c_V^{(l)}$, δηλαδή του k cluster για τους χρήστες και του l για τα αντικείμενα ,δεδομένου του ζεύγους χρήστη-αντικειμένου u, v. Διαισθητικά, η πιθανότητα $P\left(c_U^{(k)}, c_V^{(l)} \middle| u, v\right)$ συμβολίζει τον βαθμό βεβαιότητας με τον οποίο το ζεύγος χρήστη-αντικειμένου (u, v) ανήκει στο user-item co-cluster $(c_U^{(k)}, c_V^{(l)})$.

Επίσης, ένα co-cluster χρηστών και αντικειμένων μπορεί να έχει διαφορετικές βαθμολογίες με διαφορετικές πιθανότητα $P\left(r \middle| c_U^{(k)}, c_V^{(l)}\right)$, όπου εδώ r είναι η βαθμολογία (π.χ. από 1 έως 5)

Η βαθμολογία ενός χρήστη u για το αντικείμενο v, υποθέτοντας ότι οι τυχαίες μεταβλητές u και v είναι ανεξάρτητες μεταξύ τους, και σύμφωνα με το $[32]$, δίνεται μέσω της συνάρτησης :

$$f_R(u,v) = \sum_r r\, P(r|u,v)$$
$$= \sum_r r \sum_{k,l} P\left(r\middle|c_U^{(k)}, c_V^{(l)}\right) P\left(c_U^{(k)}, c_V^{(l)}\middle|u,v\right)$$
$$= \sum_r r \sum_{k,l} P\left(r\middle|c_U^{(k)}, c_V^{(l)}\right) P\left(c_U^{(k)}\middle|u\right) P\left(c_V^{(l)}\middle|v\right) \qquad (1)$$

Η εξίσωση (1) υπολογίζει την αναμενόμενη βαθμολογία λαμβάνοντας υπόψη όλες τις πιθανές βαθμολογίες r και τα πιθανά ζεύγη συστάδων χρηστών και αντικειμένων. Κάθε όρος στην εξίσωση συνεισφέρει στην τελική αναμενόμενη βαθμολογία με βάση τις πιθανότητες συμμετοχής του χρήστη και του αντικειμένου στις αντίστοιχες συστάδες και την πιθανότητα της βαθμολογίας r για το συγκεκριμένο ζεύγος συστάδων.

Η εξίσωση (1) μπορεί να εκφραστεί σύμφωνα με το [32] σε μορφή πινάκων ως :

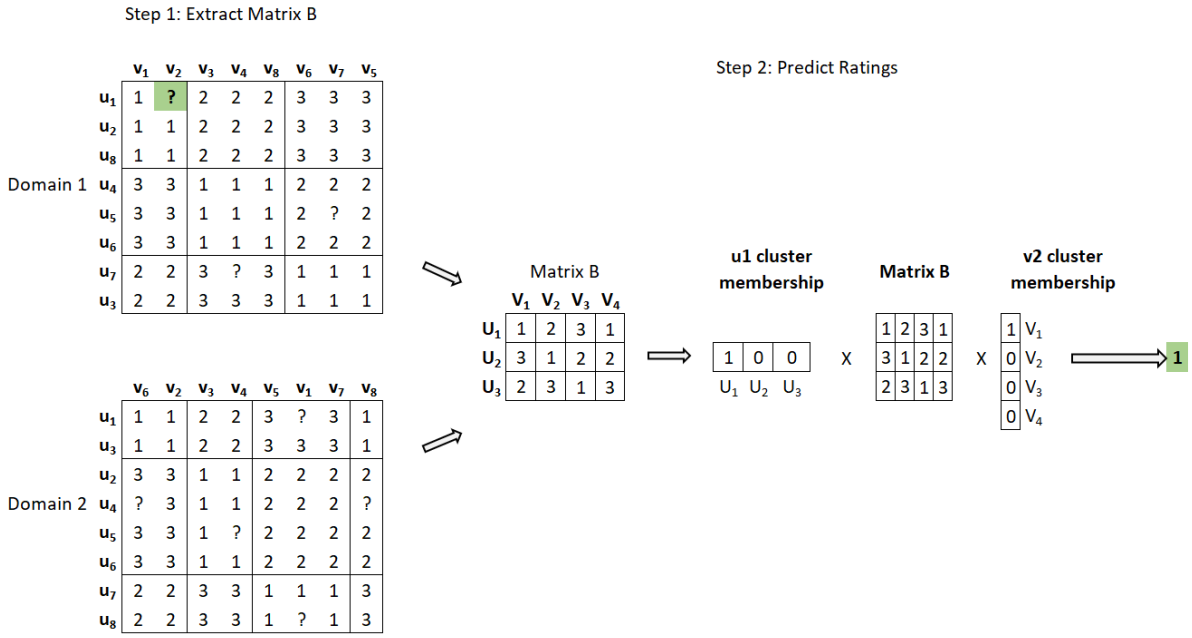
$$f_R(u,v) = p_u^\top B p_v, \quad , p_u^\top 1 = 1, \quad p_v^\top 1 = 1 \qquad (2)$$

, όπου $p_u \in \mathbb{R}^K$ και $p_v \in \mathbb{R}^L$ είναι διανύσματα των οποίων τα στοιχεία αθροίζουν την μονάδα και ουσιαστικά συμβολίζουν την πιθανότητα με την οποία ο χρήστης u ανήκει σε καθεμία από τις K συστάδες χρηστών και αντιστοίχως την πιθανότητα με την οποία το αντικείμενο v ανήκει σε καθεμία από τις L συστάδες αντικειμένων. Ο πίνακας B είναι ένας πίνακας K x L , όπου τα στοιχεία του είναι οι αναμενόμενες βαθμολογίες για κάθε co-cluster και μπορεί να γραφεί ως σύμφωνα με το [32] ως:

$$B_{kl} = \sum_r r\, P\left(r\middle|c_U^{(k)}, c_V^{(l)}\right) \qquad (3)$$

Παρατηρώντας τις εξισώσεις (2) και (3), ο πίνακας B κωδικοποιεί ουσιαστικά το κοινό μοτίβο βαθμολόγησης μεταξύ των τομέων που η μέθοδος υποθέτει ότι υπάρχει. Με άλλα λόγια, η πρόβλεψη για την βαθμολογία που δίνει ένας χρήστης u σε ένα αντικείμενο v είναι ένας γραμμικός συνδυασμός των:

- p_u που είναι το διάνυσμα πιθανοτήτων του χρήστη u να ανήκει στις διάφορες συστάδες χρηστών.
- B που είναι ο πίνακας βαθμολογιών σε επίπεδο συστάδων.
- p_v που είναι το διάνυσμα πιθανοτήτων του αντικειμένου v να ανήκει στις διάφορες συστάδες αντικειμένων.



Εικόνα 24- RMGM Prediction example

Στην Εικόνα 24 παρουσιάζεται γραφικά το πως η συγκεκριμένη μέθοδος κάνει προτάσεις, ως συνέχεια του παραδείγματος στην Εικόνα 23. Πιο συγκεκριμένα, θεωρούμε ότι έχουμε ήδη βρει τον πίνακα B που περιέχει την πληροφορία για τα κοινά μοτίβα αξιολόγησης των δύο τομέων. Συνεπώς, γίνεται πρόβλεψη της βαθμολογίας του χρήστη u1 για το αντικείμενο v1 χρησιμοποιώντας το διάνυσμα cluster-membership του u1 (τα πιθανά clusters είναι U1,U2,U3 , εδώ έχουμε hard membership συνεπώς ο u1 ανήκει με βεβαιότητα στο cluster U1) , τον πίνακα B και το διάνυσμα cluster-membership του v2 (και εδώ έχουμε hard membership και το v2 ανήκει με βεβαιότητα στο cluster V1).

5.3 Εκπαίδευση του μοντέλου

Η εκπαίδευση του μοντέλου γίνεται μέσω του αλγορίθμου EM (expectation maximization) ο οποίος προτάθηκε από τους Dempster et.al [39]. Στόχος είναι η εύρεση της συνάρτησης (1) από τα δεδομένα.

Δεδομένων K συστάδων χρηστών και L αντικειμένων το επόμενο βήμα σύμφωνα με το [32] είναι να εκφραστεί η οριακή κατανομή των χρηστών και των αντικειμένων ως μίξη κατανομών ως προς τις αντίστοιχες στοιβάδες, δηλαδή:

$$P(u) = \sum_{k} P\left(c_U^{(k)}\right) P\left(u \middle| c_U^{(k)}\right) \qquad (4)$$

$$P(v) = \sum_{l} P\left(c_V^{(l)}\right) P\left(v \middle| c_V^{(l)}\right) \qquad (5)$$

Θεωρώντας ότι οι τυχαίες μεταβλητές χρήστες u και αντικείμενα v είναι ανεξάρτητες μεταξύ τους, τότε η από κοινού οριακή τους κατανομή προκύπτει από τον συνδυασμό των δύο μίξεων. Πιο συγκεκριμένα, αξιοποιώντας τις σχέσεις (4) και (5), τότε ένα ζεύγος χρήστη-αντικειμένου μπορεί να θεωρηθεί ότι προκύπτει ως :

$$\left(u_i^{(z)}, v_i^{(z)}\right) \sim \sum_{k,l} P\left(c_U^{(k)}\right) P\left(c_V^{(l)}\right) P\left(u \middle| c_U^{(k)}\right) P\left(v \middle| c_V^{(l)}\right) \qquad (6)$$

Επιπλέον, σύμφωνα με το [32], μια βαθμολογία μπορεί να θεωρηθεί ότι προκύπτει ως η δεσμευμένη κατανομή δεδομένων των στοιβάδων, η πιο συγκεκριμένα ως εξής:

$$r_i^{(z)} \sim P\left(r \middle| c_U^{(k)}, c_V^{(l)}\right) \qquad (7)$$

Οι σχέσεις (6) και (7) περιέχουν πέντε μεταβλητές $P\left(c_U^{(k)}\right)$, $P\left(c_V^{(l)}\right)$, $P\left(u \middle| c_U^{(k)}\right)$, $P\left(v \middle| c_V^{(l)}\right)$ και $P\left(r \middle| c_U^{(k)}, c_V^{(l)}\right)$, οι οποίες εφόσον εκτιμηθούν μας δίνουν την δυνατότητα υπολογισμού της εξίσωσης (1) που είναι και το ζητούμενο. Πιο συγκεκριμένα, χρησιμοποιώντας τον κανόνα του Bayes έχουμε ότι:

$$P\left(c_U^{(k)} \middle| u\right) = \frac{P\left(u \middle| c_U^{(k)}\right) P\left(c_U^{(k)}\right)}{P(u)} = \frac{P\left(u \middle| c_U^{(k)}\right) P\left(c_U^{(k)}\right)}{\sum_k P\left(c_U^{(k)}\right) P\left(u \middle| c_U^{(k)}\right)} \qquad (8)$$

$$P\left(c_V^{(l)} \middle| v\right) = \frac{P\left(v \middle| c_V^{(l)}\right) P\left(c_V^{(l)}\right)}{P(v)} = \frac{P\left(v \middle| c_V^{(l)}\right) P\left(c_V^{(l)}\right)}{\sum_l P\left(c_V^{(l)}\right) P\left(v \middle| c_V^{(l)}\right)} \qquad (9)$$

, συνεπώς η εύρεση των 5 παραπάνω μεταβλητών αρκεί για να προσδιορίσει την ζητούμενη συνάρτηση (1).

Ακολουθεί η παρουσίαση του αλγορίθμου EM (expectation maximization) [31], ο οποίος εφαρμόζεται πολύ συχνά στα πλαίσια της μηχανικής μάθησης και της στατιστικής για την εκτίμηση παραμέτρων. Ένα κλασικό παράδειγμα χρήσης του EM είναι η εκτίμηση των παραμέτρων ενός μίγματος κανονικών κατανομών (Gaussian Mixture Models). Σε αυτό το σενάριο, τα δεδομένα προέρχονται από πολλές κανονικές κατανομές, αλλά δεν γνωρίζουμε από ποια κατανομή προέρχεται κάθε παρατήρηση. Σκοπός του αλγορίθμου είναι να εκτιμήσει τις παραμέτρους των κανονικών κατανομών βάση των παρατηρήσεων που είναι διαθέσιμες. Ουσιαστικά, ο αλγόριθμος προσπαθεί να βρει τις παραμέτρους εκείνες, οι οποίες εξηγούν με τον βέλτιστο δυνατό τρόπο τις τιμές που έχουν παρατηρηθεί (εκτίμηση μέγιστης πιθανοφάνειας).

Πρακτικά, και όσον αφορά την μέθοδο RMGM, ο αλγόριθμος αποτελείται από τρία βήματα:

- Αρχικοποίηση: Επιλογή αρχικών τιμών για τις 5 παραμέτρους. Η αρχικοποίηση μπορεί να γίνει με τυχαίο τρόπο ή και με πιο εκλεπτυσμένες μεθόδους, όπως για παράδειγμα μέσω co-clustering του πίνακα αξιολογήσεων του πιο πυκνού τομέα. Το συγκεκριμένο βήμα είναι ιδιαίτερα σημαντικό, καθώς η ποιότητα της συσταδοποίησης ενδέχεται να εξαρτάται σε μεγάλο βαθμό από τις αρχικές τιμές που δίνονται.
- E-Step (Expectation Step): Υπολογίζεται η εκ των υστέρων (posterior) πιθανότητα για κάθε συνδυασμό συστάδων χρηστών και συστάδων αντικειμένων, δεδομένων των παρατηρούμενων δεδομένων.
- M-Step (Maximization Step): Χρησιμοποιούνται οι υπολογισμένες εκ των υστέρων πιθανότητες από το E-step, ώστε να ενημερωθούν οι παράμετροι του μοντέλου, με τέτοιο τρόπο ώστε να μεγιστοποιείται η πιθανότητα τον παρατηρούμενων τιμών.

Ο αλγόριθμος ΕΜ εφαρμόζεται επαναληπτικά, εναλλάσσοντας τα βήματα E-step και M-step για προκαθορισμένο αριθμό επαναλήψεων ή μέχρι να ικανοποιηθεί κάποιο κριτήριο σύγκλισης, δηλαδή οι παράμετροι του μοντέλου να μην αλλάζουν σημαντικά μεταξύ των επαναλήψεων. Το E-step υπολογίζει τις εκ των υστέρων πιθανότητες, ενώ το M-step χρησιμοποιεί αυτές τις πιθανότητες για να ενημερώσει τις παραμέτρους του μοντέλου.

Ο συγκεκριμένος αλγόριθμος έχει ορισμένα αρνητικά στοιχεία. Ορισμένα από αυτά είναι το ότι ενδέχεται η σύγκλιση να είναι αργή, δηλαδή να χρειάζεται πολλές επαναλήψεις για να επιτευχθεί ένα καλό αποτέλεσμα. Επιπλέον, ο αλγόριθμος είναι υπολογιστικά ακριβός καθώς κάθε επανάληψη περιλαμβάνει την εκτίμηση παραμέτρων για όλα τα σημεία δεδομένων. Κάτι τέτοιο, μπορεί να δημιουργήσει πρακτικό πρόβλημα σε μεγάλα σύνολα δεδομένων. Το πιο βασικό όμως, είναι το γεγονός ότι είναι ένας non-convex αλγόριθμος, δηλαδή δεν είναι εγγυημένη η σύγκλιση σε ολικό μέγιστο. Αντιθέτως, ο αλγόριθμος συγκλίνει σε τοπικά μέγιστα, τα οποία δεν είναι απαραιτήτως τα καλύτερα δυνατά.

Στην Εικόνα 25 παρουσιάζεται ο αλγόριθμος ΕΜ και στην Εικόνα 26 ο τρόπος με τον οποίο γίνεται η πρόβλεψη των αξιολογήσεων αφότου ολοκληρωθεί η εκπαίδευση του μοντέλου.

Algorithm 1: EM Algorithm for RMGM

Input: Data $D_z = \{(u_1^{(z)}, v_1^{(z)}, r_1^{(z)}), \ldots, (u_{s_z}^{(z)}, v_{s_z}^{(z)}, r_{s_z}^{(z)})\}$,
# of user clusters K, # of item clusters L, # of iterations T

1. Initialize model parameters: $P(c_U^{(k)})$, $P(c_V^{(l)})$, $P(u|c_U^{(k)})$, $P(v|c_V^{(l)})$, $P(r|c_U^{(k)}, c_V^{(l)})$

for $t = 1$ to T do

// E-step
2. Compute the joint posterior probability:

$$P(c_U^{(k)}, c_V^{(l)}|u_i^{(z)}, v_i^{(z)}, r_i^{(z)}) = \frac{P(u_i^{(z)}, c_U^{(k)})P(v_i^{(z)}, c_V^{(l)})P(r_i^{(z)}|c_U^{(k)}, c_V^{(l)})}{\sum_{p,q} P(u_i^{(z)}, c_U^{(p)})P(v_i^{(z)}, c_V^{(q)})P(r_i^{(z)}|c_U^{(p)}, c_V^{(q)})}$$

where: $P(u_i^{(z)}, c_U^{(k)}) = P(c_U^{(k)})P(u_i^{(z)}|c_U^{(k)})$, $P(v_i^{(z)}, c_V^{(l)}) = P(c_V^{(l)})P(v_i^{(z)}|c_V^{(l)})$

// M-step
For simplicity we define $P(k, l|j^{(z)})$ as $P(c_U^{(k)}, c_V^{(l)}|u_j^{(z)}, v_j^{(z)}, r_j^{(z)})$

Update model parameters:

4. $P(c_U^{(k)}) = \frac{\sum_z \sum_l \sum_j P(k, l|j^{(z)})}{\sum_z s_z}$

5. $P(c_V^{(l)}) = \frac{\sum_z \sum_k \sum_j P(k, l|j^{(z)})}{\sum_z s_z}$

6. $P(u_i^{(z)}|c_U^{(k)}) = \frac{\sum_l \sum_{j:u_j^{(z)} = u_i^{(z)}} P(k, l|j^{(z)})}{P(c_U^{(k)}) \sum_z s_z}$

7. $P(v_i^{(z)}|c_V^{(l)}) = \frac{\sum_k \sum_{j:v_j^{(z)} = v_i^{(z)}} P(k, l|j^{(z)})}{P(c_V^{(l)}) \sum_z s_z}$

8. $P(r|c_U^{(k)}, c_V^{(l)}) = \frac{\sum_z \sum_{j:r_j^{(z)} = r} P(k, l|j^{(z)})}{\sum_z \sum_j P(k, l|j^{(z)})}$

end for

Εικόνα 25-RMGM EM Algorithm [32]

RMGM-Based Prediction

After training the RMGM, the missing values in the Z given rating matrices can be generated by:

$$f_R(u_i^{(z)}, v_i^{(z)}) = \sum_r r \sum_{k,l} P(r|c_U^{(k)}, c_V^{(l)}) P(c_U^{(k)}|u_i^{(z)}) P(c_V^{(l)}|v_i^{(z)}) \quad (1)$$

where $P(c_U^{(k)}|u_i^{(z)})$ and $P(c_V^{(l)}|v_i^{(z)})$ can be computed using the learned parameters based on the Bayes rule:

$$P(c_U^{(k)}|u_i^{(z)}) = \frac{P(u_i^{(z)}|c_U^{(k)})P(c_U^{(k)})}{\sum_k P(u_i^{(z)}|c_U^{(k)})P(c_U^{(k)})}, \quad P(c_V^{(l)}|v_i^{(z)}) = \frac{P(v_i^{(z)}|c_V^{(l)})P(c_V^{(l)})}{\sum_l P(v_i^{(z)}|c_V^{(l)})P(c_V^{(l)})}$$

If we set for simplicity $u_i^{(z)}=u$ and $v_i^{(z)}=v$ then

we can further rewrite (1) in the form of matrices: $f_R(u,v) = p_u^T B p_v$ (2) ,

where $p_u \in \mathbb{R}^K$ and $p_v \in \mathbb{R}^L$ are the user- and item-cluster membership vectors

($[p_u]_k = P(c_U^{(k)}|u)$ and $[p_v]_l = P(c_V^{(l)}|v)$) and B is a $K \times L$ relaxed cluster-level rating matrix

in which an entry can have multiple ratings with different probabilities: $B_{kl} = \sum_r r P(r|c_U^{(k)}, c_V^{(l)})$.

Equation (2) implies that the relaxed cluster-level rating matrix B is a cluster-level rating model.

Εικόνα 26-RMGM Prediction [32]

6. ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΜΕΛΕΤΗΣ

6.1 Πείραμα 1- Μεταφορά γνώσης από έναν πυκνό σε έναν αραιό τομέα

Για την υλοποίηση του πειράματος επιλέχθηκαν δύο υποσύνολα 1000 χρηστών και 500 αντικειμένων από το κάθε σύνολο δεδομένων.

Πιο συγκεκριμένα, για το σύνολο MovieLens [37] πρώτα αποκλείστηκαν οι ταινίες που έχουν λιγότερες από 50 αξιολογήσεις και έπειτα επιλέχθηκαν τυχαία 1000 χρήστες και 500 ταινίες, με κάθε χρήστη να έχει πάνω από 100 βαθμολογίες. Η συγκεκριμένη διαδικασία μας έδωσε ένα σύνολο δεδομένων με μέση βαθμολογία 3,40 και sparsity 85.15% (ή αντίστροφα rating ratio 14.85%).

Αντίστοιχα, για το σύνολο Goodbooks [38] πρώτα αποκλείστηκαν τα βιβλία που έχουν λιγότερες από 25 αξιολογήσεις και έπειτα επιλέχθηκαν τυχαία 1000 χρήστες και 500 βιβλία, με κάθε χρήστη να έχει πάνω από 15 βαθμολογίες. Η συγκεκριμένη διαδικασία μας έδωσε ένα σύνολο δεδομένων με μέση βαθμολογία 3,86 και sparsity 98.23% (ή αντίστροφα rating ratio 1.77%).

Ο πίνακας αξιολογήσεων που προέκυψε από τις ταινίες MovieLens χρησιμοποιήθηκε, ως ο πιο πυκνός από τους δύο, σαν βοηθητικός πίνακας (auxiliary matrix) και ο έλεγχος του αλγορίθμου έγινε επί του συνόλου Goodbooks. Πιο συγκεκριμένα, έγινε προσπάθεια για μεταφορά γνώσης από τον τομέα Ταινίες στον τομέα Βιβλία.

Για την αξιολόγηση του αλγορίθμου ακολουθήθηκε διαδικασία ανάλογη με αυτήν που προτείνουν σχετικά έργα και πιο συγκεκριμένα επιλέχθηκαν τυχαία 100 χρήστες από το σύνολο Goodbooks και για αυτούς τους χρήστες επιλέχθηκαν τυχαία 15 από τις παρατηρήσεις τους για εκπαίδευση και οι υπόλοιπες για test στο μοντέλο. Η διαδικασία αυτή επαναλήφθηκε 10 φορές.

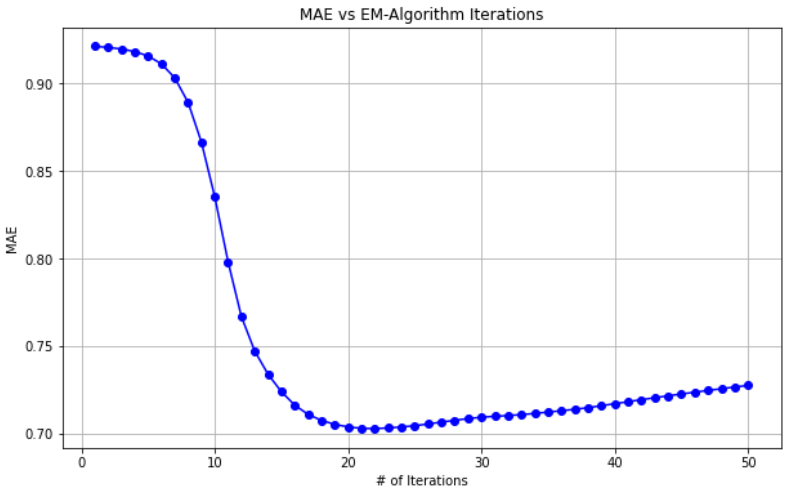
Η μετρική αξιολόγησης που χρησιμοποιήθηκε είναι το μέσο τετραγωνικό σφάλμα (mean absolute error) το οποίο ορίζεται ως εξής: MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

Για την αξιολόγηση της απόδοσης του αλγορίθμου τα αποτελέσματα συγκρίθηκαν με αυτά τριών μεθόδων που δεν κάνουν χρήση μεταφοράς γνώσης και συγκεκριμένα των SVD (singular value decomposition) ,NMF (non-negative matrix factorization) και KNN-pearson (k-nearest neighbors with pearson), όπως αυτές υλοποιούνται από την βιβλιοθήκη surprise.

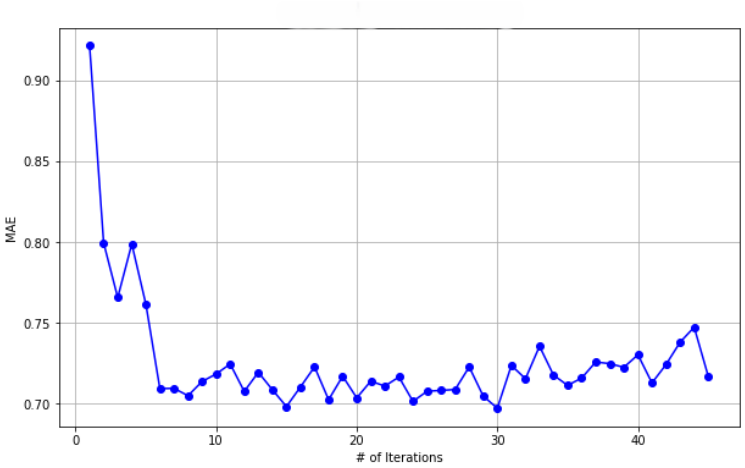
Οι SVD και NMF, όπως αναλύθηκαν και στο κεφάλαιο 2 είναι model-based προσεγγίσεις , αναλύουν τον πίνακα αξιολογήσεων σε γινόμενο πινάκων χαμηλότερων διαστάσεων και χρησιμοποιούν τις k πιο σημαντικές διαστάσεις ώστε να κάνουν προβλέψεις. Ο KNN-pearson , όπως παρουσιάστηκε στο κεφάλαιο 2, είναι μια memory-based προσέγγιση, υπολογίζει την ομοιότητα μεταξύ χρηστών χρησιμοποιώντας τον συντελεστή συσχέτισης Pearson και δημιουργεί προβλέψεις χρησιμοποιώντας τους k-κοντινότερους γείτονες για έναν χρήστη.

Για τον αλγόριθμο RMGM τέθηκαν ως παράμετροι K,L,T=20. K είναι ο αριθμός συστάδων των χρηστών, L ο αριθμός συστάδων των αντικειμένων και T είναι ο αριθμός των επαναλήψεων για τον αλγόριθμο expectation maximization. Στις εικόνες 27 και 28 παρουσιάζονται οι τιμές του MAE για ένα test set για τις διάφορες τιμές των παραμέτρων K,L και T.

Όσον αφορά την επιλογή του T, παρατηρήθηκε εμπειρικά, αλλά και από το διάγραμμα της Εικόνας 27, ότι η συγκεκριμένη επιλογή δίνει καλά αποτελέσματα. Στο σημείο αυτό θα πρέπει να αναφερθεί ότι ο αλγόριθμος EM (expectation maximization) μέσω του οποίου εκπαιδεύεται το μοντέλο είναι ένας αλγόριθμος non-convex, κάτι το οποίο σημαίνει ότι υπάρχουν τοπικά μέγιστα στα οποία ο αλγόριθμος μπορεί να παγιδευτεί και η σύγκλιση δεν είναι εγγυημένη. Όσον αφορά την επιλογή των K και L, παρομοίως απέδωσαν καλύτερα στα δεδομένα και επίσης προτείνονται από την σχετική βιβλιογραφία. Τέλος, για τους SVD και NMF επιλέχθηκαν αντίστοιχα 20 factors ως παράμετρος και για τον KNN οι 20 κοντινότεροι γείτονες.



Εικόνα 27 - MAE vs algorithm iterations



Εικόνα 28- MAE vs # of clusters

Ακολουθούν τα αποτελέσματα (MAE) των 10 δοκιμών:

Run	**RMGM**	**SVD**	**NMF**	**PCC**
1	0.71208	0.70213	0.72036	0.71812
2	0.70371	0.71239	0.79958	0.75097
3	0.69489	0.68682	0.72763	0.68659
4	0.62669	0.62704	0.73687	0.70198
5	0.72015	0.68986	0.73143	0.74371
6	0.70432	0.65813	0.79041	0.715
7	0.71723	0.69166	0.78059	0.72247
8	0.67189	0.65454	0.71365	0.69053
9	0.73552	0.68704	0.74821	0.7312
10	0.72295	0.68779	0.75392	0.6862
Average	0.70094	**0.67974**	0.75027	0.71467

Εικόνα 29- MAE comparison, Books Domain – Given 15

Παρατηρούμε ότι ο αλγόριθμος RMGM αποδίδει καλύτερα από τους NMF και PCC, αλλά όχι από τον SVD , καθώς μικρότερη τιμή σημαίνει καλύτερο αποτέλεσμα για το μέσο τετραγωνικό σφάλμα.

6.2 Πείραμα 2- Μεταφορά γνώσης μεταξύ δύο αραιών τομέων

Εφόσον η μέθοδος RMGM βρίσκει ένα κοινό μοτίβο αξιολογήσεων σε επίπεδο συστάδων μεταξύ διαφορετικών τομέων, ώστε να δημιουργήσει προβλέψεις, έχει ενδιαφέρον να εξεταστεί η απόδοση της μεθόδου και στους δύο τομείς ταυτόχρονα. Πιο συγκεκριμένα, η προσέγγιση αυτή ανήκει σε μια υποκατηγορία της μεταφοράς μάθησης, που ονομάζεται multi-task learning. Στο multi-task learning χρησιμοποιούνται και οι δύο (ή περισσότεροι) τομείς και εργασίες ταυτόχρονα ώστε να βελτιωθεί καθολικά η επίδοση των εργασιών.

Επιπλέον, μια ανάλογη προσέγγιση με το RMGM είναι το Flexible Mixture Model (FMM) [31], που ουσιαστικά πρόκειται για τον αντίστοιχο αλγόριθμο εφαρμοσμένο όμως σε έναν τομέα, χωρίς μεταφορά γνώσης. Ο αλγόριθμος που ορίστηκε στο προηγούμενο κεφάλαιο (κεφάλαιο 5) δεν έχει ως αυστηρό περιορισμό την ύπαρξη παραπάνω από ενός τομέα για να εφαρμοστεί. Συνεπώς, στην συνέχεια, χρησιμοποιείται ο αλγόριθμος RMGM όπως έχει οριστεί, σε κάθε τομέα ξεχωριστά και ονομάζεται FMM.

Η τεχνική FMM λειτουργεί με παρόμοιο τρόπο, όπως και το υπό εξέταση μοντέλο, δημιουργώντας co-clusters χρηστών και αντικειμένων και στοχεύοντας στην εύρεση ενός μοτίβου αξιολογήσεων που αναπαριστά την συμπεριφορά των χρηστών ως προς τα αντικείμενα καθολικά για τον κάθε τομέα ξεχωριστά.

Για την υλοποίηση του πειράματος επιλέχθηκαν δύο υποσύνολα 1000 χρηστών και 500 αντικειμένων από το κάθε σύνολο δεδομένων.

Πιο συγκεκριμένα, για το σύνολο MovieLens πρώτα αποκλείστηκαν οι ταινίες που έχουν λιγότερες από 25 αξιολογήσεις και έπειτα επιλέχθηκαν τυχαία 1000 χρήστες και 500 ταινίες, με κάθε χρήστη να έχει πάνω από 15 βαθμολογίες. Η συγκεκριμένη διαδικασία μας έδωσε ένα σύνολο δεδομένων με μέση βαθμολογία 3,68 και sparsity 96.57% (ή αντίστροφα rating ratio 3.43%).

Αντίστοιχα, για το σύνολο Goodbooks πρώτα αποκλείστηκαν τα βιβλία που έχουν λιγότερες από 25 αξιολογήσεις και έπειτα επιλέχθηκαν τυχαία 1000 χρήστες και 500 βιβλία, με κάθε χρήστη να έχει πάνω από 15 βαθμολογίες. Η συγκεκριμένη διαδικασία μας έδωσε ένα σύνολο δεδομένων με μέση βαθμολογία 3,96 και sparsity 98.79% (ή αντίστροφα rating ratio 1.21%).

Για την αξιολόγηση του αλγορίθμου ακολουθήθηκε διαδικασία train-test split με ποσοστά 80-20 σε κάθε τομέα ξεχωριστά. Πρακτικά, 20% από τις παρατηρούμενες βαθμολογίες των δύο πινάκων «κρύφτηκε» και τα διάφορα μοντέλα αξιολογήθηκαν ως προς τις προβλέψεις που παρήγαγαν για τα συγκεκριμένα στοιχεία των πινάκων. Η μετρική αξιολόγησης που χρησιμοποιείται είναι και πάλι το μέσο απόλυτο σφάλμα (MAE). Η διαδικασία του train-test split πραγματοποιήθηκε 10 φορές και η σύγκριση μεταξύ των διάφορων αλγορίθμων γίνεται με βάση την μέση τιμή του MAE για αυτές τις επαναλήψεις.

Τέλος, χρησιμοποιήθηκαν οι ίδιες παράμετροι για τις διάφορες μεθόδους, όπως και στο πρώτο πείραμα, δηλαδή K=L=T=20 για τα RMGM και FMM και 20 factors και γείτονες για τις μεθόδους SVD,NMF και PCC αντίστοιχα. Στο σημείο αυτό θα πρέπει να διευκρινιστεί το γεγονός ότι για την μεταφορά γνώσης Ταινίες -> Βιβλία και Βιβλία-> Ταινίες χρησιμοποιείται όλη η πληροφορία από τους πίνακες στον source τομέα, δηλαδή δεν αποκρύπτονται βαθμολογίες στους τομείς Ταινίες και Βιβλία αντίστοιχα.

Στις Εικόνες 30 και 31 παρατίθενται τα αποτελέσματα για τις προβλέψεις στους τομείς Βιβλία και Ταινίες αντίστοιχα, τα οποία και σχολιάζονται:

- Ταινίες -> Βιβλία: Η μέθοδος RMGM υπερτερεί των single-domain προσεγγίσεων NMF και PCC. Επιπλέον, υπερτερεί και της παρεμφερούς single-domain προσέγγισης FMM, κάτι που υποδεικνύει ότι έχουμε θετική μεταφορά γνώσης και συνεπώς η πρόβλεψη για τον τομέα Βιβλία βελτιώνεται χρησιμοποιώντας τον τομέα Ταινίες. Τέλος, η single-domain προσέγγιση SVD υπερτερεί όλων των υπολοίπων.
- Βιβλία -> Ταινίες: Η μέθοδος RMGM υπερτερεί των single-domain προσεγγίσεων NMF και PCC. Αντίθετα, η single-domain προσέγγιση SVD υπερτερεί της RMGM, αλλά και των υπολοίπων προσεγγίσεων. Ιδιαίτερα σημαντικό είναι το γεγονός ότι η single-domain

μέθοδος FMM εμφανίζεται οριακά καλύτερη σε σχέση με την RMGM, κάτι που υποδηλώνει ότι δεν έχουμε θετική μεταφορά γνώσης από τον τομέα Ταινίες στον τομέα Βιβλία. Το συγκεκριμένο εύρημα υποδηλώνει επίσης ότι η μεταφορά γνώσης δεν λειτουργεί αμφίδρομα με θετικό τρόπο, καθώς η πρόβλεψη βελτιώθηκε για Ταινίες -> Βιβλία, αλλά όχι και αντιστρόφως.

Run	**RMGM**	**FMM**	**SVD**	**NMF**	**PCC**
1	0.7264	0.7279	0.7236	0.7996	0.7644
2	0.7284	0.7409	0.7171	0.7805	0.7601
3	0.7292	0.7371	0.7215	0.7931	0.7683
4	0.7250	0.7382	0.6848	0.7683	0.7214
5	0.7401	0.7440	0.7135	0.7950	0.7468
6	0.7482	0.7573	0.7119	0.7914	0.7641
7	0.7369	0.7450	0.7312	0.7805	0.7800
8	0.7242	0.7376	0.7232	0.7979	0.7653
9	0.7464	0.7398	0.7388	0.8103	0.7721
10	0.7388	0.7334	0.7115	0.7713	0.7643
Average	0.7344	0.7401	**0.7177**	0.7888	0.7607

Εικόνα 30 – MAE Comparison, Books Domain – 80/20 split

Run	**RMGM**	**FMM**	**SVD**	**NMF**	**PCC**
1	0.7054	0.6996	0.6771	0.7968	0.7091
2	0.7018	0.6925	0.6983	0.7952	0.7104
3	0.6957	0.7039	0.6848	0.7659	0.7248
4	0.7089	0.7064	0.6945	0.7860	0.7128
5	0.7103	0.7046	0.6884	0.7909	0.7324
6	0.6926	0.6985	0.6847	0.7833	0.7210
7	0.6955	0.6963	0.6997	0.7857	0.7438
8	0.7003	0.6960	0.6949	0.7980	0.7200
9	0.7107	0.7040	0.6887	0.7986	0.7091
10	0.7064	0.7061	0.6919	0.7772	0.7146
Average	0.7028	0.7008	**0.6903**	0.7878	0.7198

Εικόνα 31- MAE Comparison, Movies Domain - 80/20 split

7. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

7.1 Συμπεράσματα

Στα πειράματα που εφαρμόστηκαν, έγινε προσπάθεια για μεταφορά γνώσης από τον τομέα Ταινίες στον τομέα Βιβλία και αντιστρόφως. Η προσέγγιση RMGM που χρησιμοποιήθηκε για την μεταφορά γνώσης απέδωσε καλύτερα από δύο single-domain προσεγγίσεις, αλλά όχι καλύτερα από την μέθοδο SVD. Στο σημείο αυτό θα πρέπει να αναφερθεί ότι ο αλγόριθμος δοκιμάστηκε σε πολλά διαφορετικά πειράματα, με διαφορετικά επίπεδα sparsity και διαφορετικό τρόπο test (90-10 train-test split, new users given 5 or 10) και τα αποτελέσματα ήταν παρόμοια. Επίσης δοκιμάστηκε ο αλγόριθμος με πολλές διαφορετικές τυχαίες αρχικοποιήσεις.

Ένας λόγος για τον οποίο ενδεχομένως η μέθοδος RMGM δεν κέρδισε όλες τις single-domain προσεγγίσεις με τις οποίες συγκρίθηκε είναι επειδή η γνώση που μεταφέρθηκε ως κοινά μοτίβα αξιολόγησης από τον ένα τομέα να περιέχει αρνητική πληροφορία για τον δεύτερο, δηλαδή να έχουμε την περίπτωση του negative transfer. Πιο συγκεκριμένα, ενδεχομένως κάθε τομέας να περιέχει πληροφορία που είναι σχετική μόνο για αυτόν, συνεπώς θα πρέπει να γίνει ένας διαχωρισμός της πληροφορίας σε domain-specific και non-domain specific. Μια πιθανή επίλυση στο συγκεκριμένο πρόβλημα θα ήταν να εξεταστεί ποια πληροφορία έχει νόημα να μεταφερθεί, από τον έναν τομέα στον άλλο, μέσω τεχνικών μεταφοράς γνώσης όπως instance weighting και μετά να τρέξει ο αλγόριθμος.

Ένας άλλος λόγος για τον οποίον ενδεχομένως η μέθοδος RMGM είχε χειρότερη απόδοση από το SDV είναι διότι οι κατανομές τον δεδομένων διαφέρουν από τομέα σε τομέα. Για παράδειγμα, η μέση τιμή των βαθμολογιών είναι πολύ σημαντικός παράγοντας όσον αφορά την συγκεκριμένη μέθοδο, καθώς αν τα δείγματα διαφέρουν σημαντικά μεταξύ τους ως προς αυτήν, δεν δημιουργούνται “καλές” συστάδες οι οποίες είναι αντιπροσωπεύτηκες για κάθε τομέα και συνεπώς τα ratting patterns που μεταφέρονται ως γνώση δεν είναι ιδανικά. Μια επίλυση στον συγκεκριμένο ζήτημα θα μπορούσε να είναι η εφαρμογή μεθόδων domain adaptation, ώστε οι κατανομές των δεδομένων να έρθουν πιο κοντά η μια στην άλλη και μετά να εφαρμοστεί ο αλγόριθμος RMGM.

Ταυτόχρονα, σημαντικό είναι το γεγονός ότι η μέθοδος RMGM έδειξε να αποδίδει καλύτερα από την single-domain παρεμφερή μέθοδο FMM, κάτι που υποδηλώνει ότι μπορεί να χρησιμοποιηθεί για να μεταφέρει γνώση μεταξύ τομέων. Παρόλα αυτά, η θετική μεταφορά γνώσης δεν έγινε με αξιόπιστο και αμφίδρομο τρόπο.

Γενικότερα, το cross domain collaborative filtering είναι μια ενδιαφέρουσα και ιδιαίτερα θελκτική προσέγγιση, υπό την έννοια ότι δεν κάνει χρήση του προφίλ χρηστών ή αντικειμένων, το οποίο σε ορισμένες περιπτώσεις είναι δύσκολο να αποκτηθεί. Ταυτόχρονα όμως, θα πρέπει να εφαρμόζονται τεχνικές που να διασφαλίζουν το ότι οι τομείς είναι σχετικοί μεταξύ τους και το ότι η μεταφορά γνώσης γίνεται με έναν αξιόπιστο τρόπο.

7.2 Μελλοντική Εργασία

Στην παρούσα διπλωματική εργασία μελετήθηκε μια συγκεκριμένη μέθοδος μεταφοράς γνώσης σε συστήματα συνεργατικού φιλτραρίσματος χωρίς κοινούς χρήστες και αντικείμενα μεταξύ των τομέων. Η συγκεκριμένη μέθοδος θα μπορούσε να επεκταθεί χρησιμοποιώντας κάποια από τις προσεγγίσεις που αναφέρθηκαν στις προηγούμενες παραγράφους, όπως για παράδειγμα με εφαρμογή domain adaptation τεχνικών. Μια τέτοια προέκταση ενδεχομένως να βελτίωνε την απόδοση της συγκεκριμένης μεθόδου, καθώς θα διασφάλιζε ότι οι κατανομές που ακολουθούν τα δεδομένα στους δύο τομείς είναι παρεμφερείς μεταξύ τους.

Μια άλλη προέκταση της συγκεκριμένης εργασίας έγκειται στην βελτίωση του αλγορίθμου EM (Expectation-Maximization). Μια προσέγγιση είναι η πιο εκλεπτυσμένη επιλογή της αρχικοποίησης των μεταβλητών, όπως για παράδειγμα κάνοντας co-cluster του χρήστες και τα αντικείμενα του πιο πυκνού τομέα και ξεκινώντας από εκεί τον αλγόριθμο. Μια άλλη προσέγγιση θα ήταν η χρήση της τεχνικής annealing , ώστε να αποφευχθούν τα «κακά» τοπικά μέγιστα στον αλγόριθμο EM. Αυτό επιτυγχάνεται με το διαχωρισμό των δεδομένων σε δύο υποσύνολα και την προσθήκη μιας παραμέτρου ρυθμού αλλαγής στο βήμα Μ. Η απόδοση του αλγορίθμου παρακολουθείται συγκρίνοντας την απόδοση έναντι των δύο υποσυνόλων. Όταν βρεθεί η καλύτερη παράμετρος μάθησης, ο αλγόριθμος εκτελείται για όλο το σύνολο δεδομένων.

Ιδιαίτερο ενδιαφέρον παρουσιάζει επίσης και η βελτιστοποίηση του αλγορίθμου ως προς την ταχύτητα εκπαίδευσης. Στην συγκεκριμένη εργασία χρησιμοποιήθηκε ένα μικρό υποσύνολο δεδομένων, συνεπώς μια πιθανή προέκταση είναι η εφαρμογή της μεθόδου σε μεγάλου όγκου δεδομένα ή ακόμα και σε real-time, με τις κατάλληλες τροποποιήσεις.

Τέλος, στις προσεγγίσεις μεταφοράς γνώσης σε συστήματα χωρίς κοινούς χρήστες και αντικείμενα, είναι πολύ σημαντικό να διερευνηθεί πρωτίστως κατά πόσο οι τομείς σχετίζονται μεταξύ τους. Γενικότερα, η μεταφορά γνώσης, εκτός από ότι αυξάνει την πολυπλοκότητα ενός συστήματος συστάσεων, ενδεχομένως και να επιφέρει αρνητικά αποτελέσματα ή απλώς να μην αποδίδει τόσο καλά όσο μια παραδοσιακή μέθοδος. Συνεπώς, μια μελλοντική εργασία η οποία θα παρουσίαζε ιδιαίτερο ενδιαφέρον είναι η δημιουργία ενός ενοποιημένου πλαισίου μέσω του οποίου θα εξετάζεται κατά πόσο οι τομείς παρουσιάζουν κοινά μοτίβα αξιολόγησης και κατά συνέπεια αν ενδείκνυται η μεταφορά γνώσης.

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] D. Perkins and G. Salomon, "Transfer of Learning," in *The International Encyclopedia of Education,Secon Edition*, Oxford, England, Pergamon Press, 1992, pp. 425-441.

[2] I. Elezi, Z. Yu, A. Anandkumar, L. Leal-Taixé and J. Alvarez, "Not All Labels Are Equal: Rationalizing the Labeling Costs for Training Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14492-14501.

[3] S. Pan and Q. Yang, "A Survey on Transfer Learning," *in IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2010.

[4] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE,* vol. 109, no. 1, pp. 43-76, 2020.

[5] R. Raina, A. Battle, H. Lee, B. Packer and A. Ng, "Self-Taught Learning: Transfer Learning from Unlabeled Data," *Proc. 24th Int'l Conf. Machine Learning,* pp. 759-766, 2007.

[6] W. Dai, Q. Yang, G. Xue and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 193-200.

[8] P. Castells and D. Jannach, Advanced Topics for Information Retrieval (Preprint), Madrid,Spain: ACM Press, 2023.

[9] I. MacKenzie, C. Meyer and S. Noble, "How retailers can keep up with consumers," McKinsey & Company, 2023. [Online]. Available: https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers. [Accessed 01/ 06/ 2024].

[10] S. Chhabra, "Netflix says 80 percent of watched content is based on algorithmic recommendations," MobileSyrup, 2017. [Online]. Available: https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/. [Accessed 01/ 06/ 2024].

[11] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering,* vol. 17, no. 6, pp. 734-749, 2005.

[12] Breese, John; Heckerman, David; Kadie, Carl;, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, 1998.

[13] M. Balabanovic and Y. Shoham, "Fab: Content-Based, Collaborative Recommendation," *Communications of the ACM,* vol. 40, no. 3, pp. 66-72, 1997.

[14] M. Quadrana, P. Cremonesi and J. Dietmar, "Sequence-Aware Recommender Systems," *ACM Computing Surveys (CSUR),* vol. 51, no. 4, pp. 1-36, 2018.

[15] S. Latifi, N. Mauro and D. Jannach, "Session-aware recommendation: A surprising quest for the state-of-the-art," *Information Sciences,* vol. 573, pp. 291-315, 2021.

[16] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski and L. Pizzato, "Beyond Personalization: Research Directions in Multistakeholder Recommendation," *ArXiv,* vol. abs/1905.01986, 2019.

[17] D. Mladenic, "Text-learning and related intelligent agents: a survey," *IEEE Intelligent Systems and their Applications,* vol. 14, no. 4, pp. 44-54, 1999.

[18] P. Lops, M. Gemmis and G. Semeraro, "Content-based Recommender Systems: State of the Art and Trends," in *Recommender Systems Handbook*, Springer, 2011, pp. 73-105.

[19] D. Goldberg, D. Nichols, M. Oki and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM,* vol. 35, no. 12, pp. 61-70, 1992.

[20] B. Sarwar, G. Karypis, J. Konstan and . J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, Hong Kong, Association for Computing Machinery, 2001, p. 285–295.

[21] C. Aggarwal, Recommender Systems, Springer, 2016.

[23] Y. Koren, R. Bell and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer,* vol. 42, no. 8, pp. 30-37, 2009.

[24] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems,* vol. 13, p. 535–541, 2000.

[25] R. Burje, "Hybrid Recommender Systems: Survey and Experiments," *User Modeling and User-Adapted Interaction,* vol. 12, pp. 331-370, 2002.

[26] J. Chiang, "Medium," 2021. [Online]. Available: https://medium.com/analytics-vidhya/7-types-of-hybrid-recommendation-system-3e4f78266ad8. [Accessed 01/ 06/ 2024].

[27] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes and M. Sartin, "Combining Content-Based and Collaborative Filters in an Online Newspaper," *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1999.

[28] P. Cremonesi, A. Tripodi and R. Turrin, "Cross-Domain Recommender Systems," *IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada,2011,* pp. 496-503.

[29] T. Zang, Y. Zhu, H. Liu, R. Zhang and J. Yu, "A Survey on Cross-domain Recommendation: Taxonomies, Methods, and Future Directions," *ACM Transactions on Information Systems,* vol. 41, no. 2, pp. 1-39, 2022.

[30] B. Li and Q. Yang, ". Can movies and books collaborate? Cross-domain collaborative filtering for sparsity reduction," *Proceedings of the 21st International Joint Conference on Artificial Intelligence,* pp. 2052-2057, 2009.

[31] L. Si and R. Jin, "Flexible Mixture Model for Collaborative FIltering," *Proceedings, Twentieth International Conference on Machine Learning,* vol. 2, pp. 704-711, 2003.

[32] B. LI, Q. Yang and X. Xue, "Transfer learning for collaborative filtering via a rating-matrix generative model," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, Association for Computing Machinery, 2009, pp. 617-624.

[33] S. Gao, H. Luo, D. Chen, S. Li,, P. Gallinari and J. Guo, "Cross-domain recommendation via cluster-level latent factor model.," in *In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*, Prague, Czech Republic, Springer Berlin Heidelberg, 2013, pp. 161-176.

[34] Q. Zhang, D. Wu, J. Lu, F. Liu and G. Zhang, "A cross-domain recommender system with consistent information transfer," *Decis. Support Syst.,* vol. 104, pp. 49-63, 2017.

[35] Q. Zhang, D. Wu, J. Lu, F. Liu and G. Zhang, "A cross-domain recommender system with consistent information transfer," *Decis. Supp. Syst.,* vol. 104, pp. 49-63, 2017.

[36] A. Singh and G. Gordon, "Relational learning via collective matrix factorization," in *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, AΨM, 2008, pp. 650-658.

[37] [Online]. Available: https://grouplens.org/datasets/movielens/25m/. [Accessed 01/ 06/ 2024].

[38] [Online]. Available: https://github.com/zygmuntz/goodbooks-10k/tree/master?tab=readme-ov-file. [Accessed 01/ 06/ 2024].

[39] P. Dempster, M. Laird and B. Rubin, "Maximum likelihood from incomplete data via the EM algortihm," *J. of the Royal Statistical Society,* vol. 39, no. 1, p. 1–38, 1977.

[40] [Online]. Available: https://github.com/AImpressionist/cross-domain-collaborative-filtering/tree/main. [Accessed 01/ 06/ 2024].

ΠΑΡΑΡΤΗΜΑ Ι. Python Code: Συνάρτηση RMGM_EM

```
import numpy as np
import time

# Function to run RMGM_EM
def RMGM_EM(Data, Size, K, L, T):
    D = Size.shape[0]  # number of domains
    M = int(np.sum(Size[:, 0]))  # total number of users
    N = int(np.sum(Size[:, 1]))  # total number of items
    P = int(np.sum(Data > 0))  # total number of observed ratings
    R = int(np.max(Data))  # rating scales (1 ... R)

    # Represent the data as a sparse matrix
    X = np.zeros((P, 3))
    i = 0
    for m in range(M):
        for n in range(N):
            if Data[m, n] > 0:
                X[i, 0] = m
                X[i, 1] = n
                X[i, 2] = Data[m, n]
                i += 1

    # Initialize model parameters
    ZU = np.ceil(np.random.rand(P) * K).astype(int)
    ZI = np.ceil(np.random.rand(P) * L).astype(int)

    Post = np.zeros((K, L, P))
    for p in range(P):
        Post[ZU[p]-1, ZI[p]-1, p] = 1

    PrrU = np.zeros(K)
    PrrI = np.zeros(L)
    CndU = np.zeros((M, K))
    CndI = np.zeros((N, L))
    CndR = np.zeros((K, L, R))

    start_time = time.time()
    # EM algorithm iterations
    for t in range(T):
        # M-step
        PrrU = np.sum(np.sum(Post, axis=2), axis=1) / P
        PrrI = np.sum(np.sum(Post, axis=2), axis=0) / P
        for r in range(R):
            Idx = (X[:, 2] == (r + 1))
            CndR[:, :, r] = np.sum(Post[:, :, Idx], axis=2) / np.sum(Post, axis=2)
        for m in range(M):
            Idx = (X[:, 0] == m)
            CndU[m, :] = np.sum(np.sum(Post[:, :, Idx], axis=2), axis=1) / (P * PrrU)
            CndU[m, :] = CndU[m, :] / np.sum(CndU[m, :])
        for n in range(N):
            Idx = (X[:, 1] == n)
            CndI[n, :] = np.sum(np.sum(Post[:, :, Idx], axis=2), axis=0) / (P * PrrI)
            CndI[n, :] = CndI[n, :] / np.sum(CndI[n, :])
        # E-step
        for p in range(P):
            m = int(X[p, 0])
            n = int(X[p, 1])
            r = int(X[p, 2]) - 1
            Prr = np.outer(PrrU, PrrI)
            Cnd = np.outer(CndU[m, :], CndI[n, :])
            Post[:, :, p] = Prr * Cnd * CndR[:, :, r]
            Post[:, :, p] = Post[:, :, p] / np.sum(Post[:, :, p])
    end_time = time.time()
    elapsed_time= end_time - start_time
    print(elapsed_time)
    # Compute core matrix and memberships
```

```
Core = np.zeros((K, L))
for r in range(R):
    Core += (r + 1) * CndR[:, :, r]
print('Shared group-level rating matrix across domains:')
print(np.round(Core))

UserMem = np.zeros((M, K))
for m in range(M):
    total = np.dot(CndU[m, :], PrrU)
    UserMem[m, :] = CndU[m, :] * PrrU / total
ItemMem = np.zeros((N, L))
for n in range(N):
    total = np.dot(CndI[n, :], PrrI)
    ItemMem[n, :] = CndI[n, :] * PrrI / total

# Fill in the missing entries
StartU = np.ones(D+1, dtype=int)
StartI = np.ones(D+1, dtype=int)
for d in range(1, D):
    StartU[d] = StartU[d-1] + Size[d-1, 0]
    StartI[d] = StartI[d-1] + Size[d-1, 1]
StartU[D] = M + 1
StartI[D] = N + 1

for d in range(D):
    for m in range(StartU[d]-1, StartU[d+1]-1):
        for n in range(StartI[d]-1, StartI[d+1]-1):
            if Data[m, n] == 0:
                Data[m, n] = np.dot(UserMem[m, :], np.dot(Core, ItemMem[n, :].T))

 # Compute the co-clustering result
IdxU = np.zeros(M, dtype=int)
IdxI = np.zeros(N, dtype=int)
for m in range(M):
     IdxU[m] = np.argmax(UserMem[m, :]) + 1
for n in range(N):
     IdxI[n] = np.argmax(ItemMem[n, :]) + 1

print('Co-clustering results and filled rating matrices:')
for d in range(D):
     SortU = np.argsort(IdxU[StartU[d]-1:StartU[d+1]-1])
     SortI = np.argsort(IdxI[StartI[d]-1:StartI[d+1]-1])
     sorted_data = Data[np.ix_(SortU + StartU[d] - 1, SortI + StartI[d] - 1)]
     print(np.round(sorted_data))

return Data, Core, CndU, CndI, PrrU, PrrI ,StartU, StartI,ItemMem , UserMem
```

Ο κώδικας που χρησιμοποιείται προέρχεται από την αντίστοιχη εργασία [32] και βρίσκεται διαθέσιμος [40]. Ο κώδικας μετατράπηκε από MATLAB στην γλώσσα Python με κάποιες μικρές αλλαγές.