ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

UNIVERSITY OF PATRAS

ΠΑΝΕΠΗΣΤΙΜΙΟ ΠΑΤΡΩΝ

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Διπλωματική Εργασία

Ανάλυση κοινωνικών δικτύων και Παροχή Συστάσεων σε χρήστες

ΜΠΑΣΤΟΥΛΗΣ ΓΕΩΡΓΙΟΣ

ΑΜ: 235593

ΕΠΙΒΛΕΠΩΝΤΕΣ

Χατζηλυγερούδης Ιωάννης, Καθηγητής

Μακρής Χρήστος, Αναπλ. Καθηγητής

ΣΥΝΕΠΙΒΛΕΠΩΝ

Περίκος Ισίδωρος, Διδάσκων ΑΑΔΕ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Μακρής Χρήστος, Αναπλ. Καθηγητής

Σιούτας Σπυρίδων, Καθηγητής

Ηλίας Αριστείδης, Μέλος ΕΔΙΠ

ΠΑΤΡΑ, ΣΕΠΤΕΜΒΡΙΟΣ 2021

Ευχαριστίες

Η παρούσα διπλωματική εργασία με τίτλο «Ανάλυση κοινωνικών δικτύων και Παροχή Συστάσεων σε χρήστες» ολοκληρώθηκε το διάστημα 07/2021. Θα ήθελα να ευχαριστήσω θερμά για την αποδοχή και τη στήριξη τον ουσιαστικό επιβλέποντα Καθηγητή κ. Ιωάννη Χατζηλυγερούδη, καθώς και για την καθοδήγηση τον συνεπιβλέποντα κ. Ισίδωρο Περίκο, διδάσκοντα ΑΑΔΕ. Επίσης, τον κ. Χρήστο Μακρή, Αναπληρωτή Καθηγητή, που δέχτηκε να αντικαταστήσει τον κ. Χατζηλυγερούδη, λόγω εκπαιδευτικής του άδειας. Η συνεργασία μας ήταν ιδιαίτερα εποικοδομητική και μέσα από αυτή κέρδισα πολύτιμες εμπειρίες που θα με συνοδεύσουν στην μελλοντική μου καριέρα ως μηχανικός ηλεκτρονικών υπολογιστών και πληροφορικής. Τέλος δε θα μπορούσα να μην αναφέρω τους Πατρινούς μου φίλους και τους γονείς μου, που υπήρξαν στήριγμα κατά την περίοδο εκπόνησης της εργασίας μου.

Πάτρα, Σεπτέμβριος 2021

Μπαστούλης Γιώργος

Περίληψη

Ένα αντικείμενο το οποίο έχει απασχολήσει σε πολύ μεγάλο βαθμό τον επιχειρησιακό αλλά και ερευνητικό κόσμο είναι τα Συστήματα Προτάσεων *(Recommender Systems)*. Πρόκειται για μία ευρύτατα διαδεδομένη τεχνολογία η οποία βασίζεται σε μεθόδους μηχανικής μάθησης και εξόρυξης πληροφοριών από δεδομένα. Τα Συστήματα Προτάσεων, με αφετηρία το έτος 1995, έχουν αναπτυχθεί ραγδαία όσον αφορά την ποικιλία των προβλημάτων που καλούνται να αντιμετωπίσουν, τις τεχνικές που χρησιμοποιούν καθώς και τις πρακτικές εφαρμογές τους. Υλοποιήσεις τους μπορούν να βρεθούν σε πολύ δημοφιλή διαδικτυακά συστήματα όπως το Netflix, η Amazon,η Pandora και πολλά άλλα. Η πλειονότητα των Συστημάτων Προτάσεων βασίζεται στην τεχνική του συνεργατικού φιλτραρίσματος, γνωστό και ως *'Collaborative Filtering'*. Η τεχνική του *Collaborative Filtering* είναι μία διαδικασία φιλτραρίσματος ή αξιολόγησης αντικειμένων χρησιμοποιώντας τις απόψεις άλλων χρηστών και βασίζεται στην υπόθεση ότι εάν ένα άτομο *A* έχει την ίδια γνώμη με ένα άτομο *B* για ένα ζήτημα, ο *A* είναι πιο πιθανό να έχει τη γνώμη του *B* για ένα διαφορετικό ζήτημα από αυτό ενός τυχαία επιλεγμένου ατόμου. Τέτοιες μέθοδοι έχουν απασχολήσει σε μεγάλο βαθμό τον ερευνητικό κόσμο και συνεπώς οι τεχνικές που έχουν αναπτυχθεί γύρω από αυτόν τον τομέα είναι πάρα πολλές.

Συγκεκριμένα στην παρούσα διπλωματική ασχολούμαστε με τα Συστήματα Προτάσεων σε κοινωνικά δίκτυα. Ένα κοινωνικό δίκτυο μπορεί να θεωρηθεί ως ένας Γράφος, όπου οι οντότητες παριστάνουν τους χρήστες και οι ακμές αναπαριστούν τις σχέσεις μεταξύ τους. Η μελέτη και τα πειράματά μας γίνονται σε τέτοια συστήματα όπου η πηγή από την οποία αντλούμε πληροφορίες για την παροχή προτάσεων, επεκτείνεται πέραν από τις αξιολογήσεις των χρηστών προς αντικείμενα, στις σχέσεις που έχουν αναπτύξει οι χρήστες μεταξύ τους. Επιπρόσθετα, χρησιμοποιούμε τεχνικές γνωστές ως *Link Prediction* οι οποίες είναι κατάλληλες για την αποτίμηση της ομοιότητας μεταξύ χρηστών μέσα σε ένα Γράφο, προκειμένου να εμπλουτίσουμε τα δεδομένα μας.

Abstract

Recommender Systems is a subject that has occupied the business and research world to a great extent. It is a widely used technology based on methods of machine learning and information retrieval. Recommender Systems, starting in 1995, have developed rapidly in terms of the variety of problems they face, the techniques they use and their practical applications. Such implementations can be found on very popular online systems such as Netflix, Amazon, Pandora and many more. The majority of Recommender Systems are based on the Collaborative Filtering technique. The Collaborative Filtering technique is a process of filtering or evaluating items using the opinions of other users and is based on the assumption that if a person *A* has the same opinion as a person *B* on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. Such methods have greatly occupied the research world and therefore the amount of techniques and algorithms that have been developed around this field is great.

Specifically in this dissertation we conduct our study on Recommender Systems in social networks. A social network is a set of users who, in addition to interacting with objects, also develop interactions with each other. Our study and experiments are carried out in such systems where the source from which we derive information for the provision of recommendations, extends beyond the user ratings to items, to the relationships that users have developed with each other. In addition, we use techniques known as Link Prediction which are suitable for evaluating similarity between users within a graph, in order to enrich our data.

Περιεχόμενα

Λίστα Πινάκων και Σχημάτων

1.0 Εισαγωγή

1.1 Μηχανική Μάθηση (Machine Learning)

Η μηχανική μάθηση είναι μια υπό-ενότητα της τεχνητής νοημοσύνης, όπου υπολογιστικοί αλγόριθμοι λειτουργούν αυτόνομα προκειμένου να 'μάθουν' από τα δεδομένα που χρησιμοποιούν χρήσιμες πληροφορίες που θα τους οδηγήσουν στην καλύτερη απόδοσή τους. Η μηχανική μάθηση, λοιπόν, είναι η προσπάθεια να επιτρέπουμε στους υπολογιστές να τροποποιήσουν ή να προσαρμόσουν τις ενέργειές τους έτσι ώστε αυτές οι ενέργειες να γίνουν περισσότερο ακριβείς, όπου η ακρίβεια μετράται από το πόσο καλά οι επιλεγμένες ενέργειες αντικατοπτρίζουν το σωστό αποτέλεσμα.

Ας φανταστούμε ότι παίζουμε Scrabble (ή κάποιο άλλο παιχνίδι) σε έναν υπολογιστή. Ξεκινώντας, μπορεί να νικάμε κάθε φορά στην αρχή, αλλά μετά από πολλά παιχνίδια ο υπολογιστής μαθαίνει να γίνεται ανταγωνιστικός, μέχρι που τελικά δεν θα κερδίσουμε ξανά ποτέ. Το συμπέρασμα θα είναι πως είτε χειροτερεύουμε εμείς, είτε ο υπολογιστής μαθαίνει πώς να κερδίζει στο Scrabble. Έχοντας μάθει να μας νικάει, ο υπολογιστής μπορεί να συνεχίσει να χρησιμοποιεί τις ίδιες στρατηγικές εναντίον άλλων παικτών, έτσι ώστε να μην ξεκινά από το μηδέν με κάθε νέο παίκτη και να χρειάζεται να αναπτύξει από την αρχή την στρατηγική του. Πρόκειται για ένα παράδειγμα που αποτελεί μία γενίκευση αυτού που αποκαλούμε μηχανική μάθηση. [1]

Στη μηχανική μάθηση, οι υπολογιστές δεν χρειάζεται να προγραμματίζονται ρητά, αλλά μπορούν να αλλάζουν και να βελτιώνουν τους αλγόριθμους τους μόνοι τους. Η βελτίωση των αλγορίθμων είναι συνήθως μία επαναληπτική διαδικασία καθώς ο αλγόριθμος χρειάζεται να μελετήσει τα δεδομένα πολλές φορές ώστε να πετύχει την επιθυμητή ακρίβεια. Σε κάθε επανάληψη θα υπάρχει ένα μικρό ποσοστό βελτίωσης και εκμάθησης, ώσπου μετά από ένα

όριο η συμπεριφορά του αλγορίθμου θα έχει φτάσει σε ένα ικανοποιητικό για τον άνθρωπο βαθμό.

Οι πρώτες συζητήσεις για την μηχανική μάθηση χρονολογούνται στην δεκαετία του 1950, ενώ τα τελευταία 15 χρόνια έχει λάβει χώρα η ραγδαία εξέλιξή της. Ιστορικά, το 1950, ο Alan Turing, δημιούργησε το 'Turing Test' προκειμένου να καθορίσει εάν ένας υπολογιστής έχει πραγματική νοημοσύνη. Η επιτυχία αυτού του τεστ απαιτούσε από έναν υπολογιστή να είναι σε θέση να ξεγελάσει έναν άνθρωπο ώστε να πιστεύει ότι είναι επίσης άνθρωπος. Αξίζει επίσης να αναφέρουμε πως ο πρώτος αλγόριθμος μηχανικής μάθησης δημιουργήθηκε το 1952 από τον 'Arthur Samuel' για το παιχνίδι της Ντάμας [2]. Ο υπολογιστής βελτίωνε το παιχνίδι του και ανέπτυσσε στρατηγικές όσο περισσότερο έπαιζε, μελετώντας ποιες κινήσεις τον οδηγούσαν στη νίκη.

Σήμερα η μηχανική μάθηση, είτε το αντιλαμβανόμαστε είτε όχι, έχει εισέλθει κατά κόρον στις ζωές μας σε πολλούς τομείς. Αναφορικά, ορισμένες συνήθεις εφαρμογές της βρίσκονται στην πληροφορική, στην αναγνώριση προσώπων και αντικειμένων, στην επεξεργασία φυσικής γλώσσας, στην ιατρική καθώς επίσης στην παροχή προτάσεων η οποία θα μας απασχολήσει στην συνέχεια της παρούσας διπλωματικής.

1.2 Είδη Μηχανικής Μάθησης

Με βάση το είδος του προβλήματος που χρήζουν λύσης, διάφορες μορφές μηχανικής μάθησης έχουν αναπτυχθεί με τις πιο διαδεδομένες να είναι η *Επιβλεπόμενη* και η *Μη-Επιβλεπόμενη* μηχανική μάθηση. Το κύριο κριτήριο που τις διαχωρίζει είναι το είδος πληροφοριών που χρησιμοποιούν οι αλγόριθμοι κατά τη διαδικασία της εκπαίδευσής τους.

- Supervised Learning (Επιβλεπόμενη μάθηση)

Στην επιβλεπόμενη μάθηση, το μηχάνημα εκπαιδεύεται χρησιμοποιώντας δεδομένα που είναι καλά "επισημασμένα", δεδομένα δηλαδή με ετικέτες (labels) για τα οποία γνωρίζουμε εξ αρχής το επιθυμητό αποτέλεσμα [1]. Μπορεί να συγκριθεί με τη μάθηση που πραγματοποιείται παρουσία ενός επόπτη ή ενός δασκάλου. Στην περίπτωση της επιβλεπόμενης μάθησης έχουμε στη διάθεσή μας ένα training set παραδειγμάτων για τα οποία έχουμε τις σωστές απαντήσεις. Ο αλγόριθμος σε αυτήν την περίπτωση, μαθαίνει από τα δεδομένα εκπαίδευσης με ετικέτα, και γενικεύει από τα δεδομένα αυτά προκειμένου να είναι σε θέση να απαντήσει σωστά σε νέα, άγνωστα δεδομένα. Μετά από την εκμάθηση από το training set, έπεται η διαδικασία της παραγωγής αποτελεσμάτων σε νέα, άγνωστα για τον υπολογιστή δεδομένα, για τα οποία όμως είναι γνωστή η σωστή απάντηση (test set). Χρησιμοποιώντας αυτά τα δεδομένα με ετικέτες, ο αλγόριθμος μπορεί να μετρήσει την απόδοσή του και να βελτιωθεί.

Η τεχνική της επιβλεπόμενης μάθησης είναι η πιο διαδεδομένη τεχνική μηχανικής μάθησης [1]. Οι πιο συνηθισμένες εφαρμογές της είναι η ταξινόμηση (Classification) και η οπισθοδρόμηση (Regression).
Τα προβλήματα ταξινόμησης (Classification) χρησιμοποιούν έναν αλγόριθμο για την ακριβή εκχώρηση δεδομένων δοκιμής σε συγκεκριμένες κατηγορίες, όπως ο διαχωρισμός των μήλων από τα πορτοκάλια. Ή, στον πραγματικό κόσμο, οι επιβλεπόμενοι αλγόριθμοι μάθησης μπορούν να χρησιμοποιηθούν για την ταξινόμηση των ανεπιθύμητων μηνυμάτων (spam) σε ξεχωριστό φάκελο από τα εισερχόμενά μας. Ένας αλγόριθμος αυτής της κατηγορίας που χρησιμοποιείται συχνά είναι ο *k-nearest-neighbours*, ο οποίος ταξινομεί κάθε δεδομένο σε γειτονιές που περιέχουν όμοιους προς αυτό γείτονες.

Η οπισθοδρόμηση (Regression) είναι ένας άλλος τύπος επιβλεπόμενης μάθησης που χρησιμοποιεί έναν αλγόριθμο για να κατανοήσει τη σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Τα μοντέλα οπισθοδρόμηση είναι χρήσιμα για την πρόβλεψη αριθμητικών τιμών βάσει διαφορετικών σημείων δεδομένων, όπως προβλέψεις εσόδων από πωλήσεις για μια δεδομένη επιχείρηση. Μερικοί δημοφιλείς αλγόριθμοι οπισθοδρόμησης είναι η γραμμική οπισθοδρόμηση (linear regression), η λογαριθμική οπισθοδρόμηση (logistic regression) και η πολυωνυμική οπισθοδρόμηση (polynomial regression).

- Unsupervised Learning (Μη επιβλεπόμενη μάθηση)

Η μη επιβλεπόμενη μάθηση είναι μια τεχνική μηχανικής μάθησης, όπου δεν χρειάζεται να επιβλέπουμε το μοντέλο [1]. Αντ' αυτού, πρέπει να επιτρέψουμε στο μοντέλο να λειτουργεί από μόνο του για να ανακαλύψει πληροφορίες. Η μη επιβλεπόμενη μάθηση χρησιμοποιεί αλγόριθμους μηχανικής μάθησης για την ανάλυση και τη συγκέντρωση συνόλων δεδομένων χωρίς ετικέτα, χωρίς δηλαδή να ξέρουμε εκ των προτέρων τη σωστή απάντηση για τα δεδομένα. Αυτοί οι αλγόριθμοι ανακαλύπτουν κρυμμένα μοτίβα στα δεδομένα χωρίς την ανάγκη ανθρώπινης παρέμβασης (ως εκ τούτου, είναι «χωρίς επίβλεψη»). Έτσι, για λειτουργίες όπως η παλινδρόμηση (regression) δεν είναι δυνατή η χρησιμοποίηση μη επιβλεπόμενης μάθησης, καθώς δεν έχουμε στη διάθεσή μας τις τιμές των μεταβλητών.

Οι πιο συνηθισμένες εφαρμογές της μη επιβλεπόμενης μηχανικής μάθησης είναι η ομαδοποίηση (Clustering) και η μείωση των διαστάσεων (Dimensionality Reduction) [3].

Η ομαδοποίηση (Clustering) [3] είναι μια τεχνική εξόρυξης δεδομένων για την ομαδοποίηση δεδομένων χωρίς ετικέτα με βάση τις ομοιότητες ή τις διαφορές τους. Για παράδειγμα, οι αλγόριθμοι ομαδοποίησης *K-means Clustering* αντιστοιχούν παρόμοια σημεία δεδομένων σε ομάδες, όπου η τιμή K αντιπροσωπεύει το μέγεθος των ομάδων που θα δημιουργηθούν και της ευκρίνειας. Αρχικά αναθέτει στις K ομάδες τυχαία δεδομένα και στη συνέχεια προσθέτει σε κάθε ομάδα τα πιο όμοια δεδομένα. Η διαδικασία ολοκληρώνεται όταν ανατεθούν όλα τα δεδομένα στις ομάδες.
Η μείωση διαστάσεων (Dimensionality Reduction) [3] είναι μια τεχνική μη επιβλεπόμενης μάθησης που χρησιμοποιείται όταν ο αριθμός των χαρακτηριστικών (features) σε ένα σύνολο δεδομένων είναι πολύ υψηλός. Μειώνει τον αριθμό των εισόδων δεδομένων σε διαχειρίσιμο μέγεθος, διατηρώντας παράλληλα την ακεραιότητα των δεδομένων. Συχνά, αυτή η τεχνική χρησιμοποιείται στο στάδιο της προ επεξεργασίας δεδομένων, όπως όταν οι αυτό-κωδικοποιητές (auto encoders) αφαιρούν θόρυβο από οπτικά δεδομένα για τη βελτίωση της ποιότητας της εικόνας.

2.0 Συστήματα Συστάσεων (Recommender Systems)

2.1 Εισαγωγή

Τα Συστήματα Προτάσεων (Recommender Systems) είναι ένα εργαλείο φιλτραρίσματος πληροφοριών που μπορούν να λειτουργήσουν σε μεγάλο και περίπλοκο όγκο δεδομένων [4]. Ιστορικά, εισήλθαν στο «προσκήνιο» το 1995 και έχουν αναπτυχθεί σε πολύ μεγάλο βαθμό, τόσο ως προς τα προβλήματα που καλούνται να λύσουν, όσο και τις πρακτικές τους εφαρμογές. Χρησιμοποιούνται ώστε να προβλέψουν την «βαθμολογία» ή «προτίμηση» που θα έδινε ένας χρήστης σε ένα στοιχείο, προτείνοντάς του τα αντικείμενα που είναι πιθανότερο να τον ενδιαφέρουν. Συναντιόνται συνήθως σε διαδικτυακές πλατφόρμες. Τα Συστήματα Προτάσεων χρησιμοποιούνται για την πρόταση διάφορων ειδών αντικειμένων. Ο όρος «αντικείμενα» είναι γενικός και μπορεί να είναι ταινίες, βιβλία, ταξιδιωτικοί προορισμοί, ειδήσεις, εστιατόρια, μουσική ή ακόμα άλλοι χρήστες και πολλά άλλα.

Χάρη στο συνεχώς μειούμενο κόστος αποθήκευσης και επεξεργασίας δεδομένων, τα Συστήματα Προτάσεων έχουν εξαπλωθεί σταδιακά στις περισσότερες πτυχές της ζωής μας. Οι πωλητές παρακολουθούν προσεκτικά την καταναλωτική μας συμπεριφορά καθώς σερφάρουμε στο διαδίκτυο προκειμένου να μας προτείνουν τα κατάλληλα προϊόντα που θα αυξήσουν τις πωλήσεις και τα κέρδη τους.
Οι ιστότοποι κοινωνικής δικτύωσης αναλύουν τις επαφές μας για να μας βοηθήσουν να συνδεθούμε με νέους φίλους και οι διαδικτυακοί ραδιοφωνικοί σταθμοί θυμούνται ποια τραγούδια παραλείπονται για να μας εξυπηρετήσουν καλύτερα στο μέλλον. Σε γενικές γραμμές, όποτε υπάρχουν πολλά διαφορετικά προϊόντα και οι πελάτες δεν μοιάζουν, η εξατομικευμένη πρόταση μπορεί να βοηθήσει ώστε να παραδοθεί το σωστό περιεχόμενο στο σωστό άτομο. Αυτό ισχύει ιδιαίτερα για εκείνες τις εταιρείες που βασίζονται στο Διαδίκτυο που προσπαθούν να κάνουν χρήση των λεγόμενων μακράς ουράς αγαθών (long-tail items) [5] που σπάνια αγοράζονται αλλά λόγω του πλήθους τους μπορούν να αποφέρουν σημαντικά κέρδη.

Η σωστή λειτουργία ενός Συστήματος Προτάσεων είναι ζωτικής σημασίας όταν χρησιμοποιείται σε εμπορικές εφαρμογές για διάφορους λόγους όπως:

- Αύξηση των πωλήσεων [6]. Αυτή είναι ίσως και η πιο σημαντική λειτουργία για ένα Σύστημα Προτάσεων, να είναι δηλαδή ικανό να πουλήσει ένα επιπλέον σύνολο αντικειμένων σε σύγκριση με αυτά που πουλήθηκαν χωρίς την παροχή προτάσεων. Αυτός ο στόχος επιτυγχάνεται επειδή τα προτεινόμενα είδη είναι πιθανό να ταιριάζουν στις ανάγκες και τις επιθυμίες του χρήστη. Πιθανώς ο χρήστης θα το αναγνωρίσει αφού δοκιμάσει αρκετές προτάσεις.
- Αύξηση πωλήσεων πιο διαφορετικών ειδών (diverse items) [6]. Μία άλλη πολύ σημαντική λειτουργία ενός Συστήματος Προτάσεων είναι ότι επιτρέπει στον χρήστη να διαλέξει αντικείμενα που πιθανώς να ήταν δύσκολο ή απίθανο να βρει χωρίς τη βοήθεια των προτάσεων. Για παράδειγμα, σε ένα Σύστημα Προτάσεων τουριστικών προορισμών, ο πάροχος υπηρεσιών ενδιαφέρεται να προωθήσει όλα τα μέρη ενδιαφέροντος μιας περιοχής και όχι μόνο τα πιο γνωστά. Αυτό θα ήταν δύσκολο χωρίς ένα Σύστημα Προτάσεων καθώς ο πάροχος δεν θα ήθελε να διακινδυνεύσει να διαφημίσει μέρη που δεν είναι πιθανό να ταιριάζουν με το γούστο ενός συγκεκριμένου χρήστη.
- Αύξηση της ικανοποίησης του χρήστη [6]. Ένα καλά σχεδιασμένο Σύστημα Προτάσεων μπορεί επίσης να βελτιώσει την εμπειρία του χρήστη με τον ιστότοπο ή την εφαρμογή. Ο χρήστης θα βρει τις προτάσεις ενδιαφέρουσες, σχετικές και, με μια σωστά σχεδιασμένη αλληλεπίδραση ανθρώπου-υπολογιστή, θα απολαύσει επίσης τη χρήση του συστήματος. Ο συνδυασμός αποτελεσματικών, ακριβών προτάσεων και μια χρήσιμη διεπαφή θα αυξήσει την υποκειμενική αξιολόγηση του συστήματος από τον χρήστη. Αυτό, με τη σειρά του, θα αυξήσει το σύστημα χρήση και την πιθανότητα να γίνουν αποδεκτές οι προτάσεις.
- Αύξηση της πιστότητας του χρήστη [6]. Ένας χρήστης πρέπει να είναι πιστός σε έναν ιστότοπο ο οποίος, όταν επισκέπτεται, αναγνωρίζει τον παλιό πελάτη και τον αντιμετωπίζει ως αξιόλογο επισκέπτη. Αυτό είναι ένα πρότυπο χαρακτηριστικό ενός Συστήματος Προτάσεων, δεδομένου ότι πολλά από αυτά υπολογίζουν προτάσεις, αξιοποιώντας έτσι τις πληροφορίες που αποκτήθηκαν από τον χρήστη κατά τη διάρκεια προηγούμενων αλληλεπιδράσεων όπως τις αξιολογήσεις του χρήστη για αντικείμενα. Κατά συνέπεια, όσο περισσότερο αλληλοεπιδρά ο χρήστης με τον

ιστότοπο, τόσο πιο εκλεπτυσμένο γίνεται το μοντέλο του χρήστη. Η αναπαράσταση των προτιμήσεων του χρήστη από το σύστημα αναπτύσσεται και η αποτελεσματικότητα των προτάσεων από το σύστημα αυξάνεται.

2.2 Σκοπός των Συστημάτων Προτάσεων

Σκοπός των Συστημάτων Προτάσεων είναι να εκτιμηθεί μια συνάρτηση χρησιμότητας (utility function) της οποίας η έξοδος είναι η πρόβλεψη του επιπέδου στο οποίο ένας χρήστης θα ήθελε να αλληλοεπιδράσει με ένα στοιχείο βάσει της προηγούμενης συμπεριφοράς του, της ομοιότητάς του με άλλους χρήστες ή άλλα στοιχεία.
Αν υποθέσουμε πως το "S" απεικονίζει το σύνολο όλων των χρηστών, "I" το σύνολο όλων των πιθανών προτεινόμενων στοιχείων, τότε το "u" θα είναι η συνάρτηση χρησιμότητας [7] (utility function) η οποία μετράει τη χρησιμότητα του στοιχείου "i" στο χρήστη "s", δηλ. u: S $\times$ I $\rightarrow$ R, όπου R είναι η βαθμολογία για το σύνολο των στοιχείων. Για κάθε χρήστη "s" που ανήκει στο σύνολο "S", η συνάρτηση χρησιμότητας "u" δείχνει τα αντικείμενα "i" που θα προταθούν από το σύνολο "I" έτσι ώστε η τιμή της να μεγιστοποιείται.

$$u = S \times I \rightarrow R$$

Η βιβλιογραφία σχετικά με τα Συστήματα Προτάσεων διακρίνεται τυπικά μεταξύ δύο ευρείας κατηγορίας μετρήσεων, της ακρίβειας των προτάσεων: πρόβλεψη βαθμολογίας ή αλλιώς rating prediction, συχνά ποσοτικοποιείται σε όρους του ριζικού μέσου τετραγωνικού σφάλματος (RMSE) και της κατάταξης ή αλλιώς ranking, μετρούμενη με όρους μετρήσεων όπως ακρίβεια (precision) και ανάκληση (recall) μεταξύ άλλων.
Όσον αφορά την πρόβλεψη βαθμολογίας (rating prediction), η πιο συνηθισμένη λειτουργία είναι η πρόβλεψη των τιμών αξιολόγησης για αυτά τα στοιχεία που ένας χρήστης έχει επιλέξει να αξιολογήσει. Αυτό το είδος δεδομένων μπορεί να συλλέγεται εύκολα και, ως εκ τούτου, είναι άμεσα διαθέσιμα για εκπαίδευση και αξιολόγηση του Συστήματος Προτάσεων εκτός

σύνδεσης. Επιπλέον, το root mean square error (RMSE), η πιο δημοφιλής μέτρηση ακρίβειας στη μέτρηση των προτάσεων, μπορεί εύκολα να αξιολογηθεί στα ζεύγη στοιχείων χρήστη που έχουν πραγματικά μια βαθμολογία στα δεδομένα.

2.3 Δεδομένα που χρησιμοποιούν τα Συστήματα Προτάσεων

Τα Συστήματα Προτάσεων λειτουργούν βάσει 2 ειδών πληροφοριών:

- Χαρακτηριστικές Πληροφορίες [8]. Αυτές οι πληροφορίες αφορούν τόσο τα αντικείμενα(π χ Κατηγορία αντικειμένου, keywords), όσο και τους χρήστες (π χ προτιμήσεις, δημογραφικά χαρακτηριστικά)
- Αλληλεπιδράσεις μεταξύ Χρηστών και Αντικειμένων [9]. Αυτές οι πληροφορίες περιγράφουν πώς ένας χρήστης αλληλοεπίδρασε με κάποιο αντικείμενο για παράδειγμα μέσω κάποιας βαθμολογίας που έβαλε στο αντικείμενο. Συνήθως απεικονίζονται με έναν πίνακα όπου στις γραμμές του βρίσκονται οι χρήστες, στις στήλες τα αντικείμενα και οι τιμές απεικονίζουν τις αξιολογήσεις των χρηστών προς τα αντικείμενα.

2.4 Κατηγορίες Συστημάτων Προτάσεων

Βασιζόμενοι στις πληροφορίες που χρησιμοποιεί ένα Σύστημα Προτάσεων, μπορούμε να τα χωρίσουμε σε κατηγορίες ως εξής:

1) Συστήματα Προτάσεων βάση φιλτραρίσματος περιεχομένου (Content-based filtering)
2) Συστήματα Προτάσεων βάση συνεργατικού φιλτραρίσματος (Collaborative filtering)
3) Υβριδικά Συστήματα Προτάσεων (Hybrid Systems)

Δύο επιπλέον κατηγορίες Συστημάτων Προτάσεων που μπορούμε να διακρίνουμε είναι τα δημογραφικά, τα οποία στοχεύουν στην κατηγοριοποίηση με βάση τα χαρακτηριστικά και την υποβολή προτάσεων βάσει δημογραφικών τάξεων και τέλος αυτά που βασίζονται στη γνώση τα οποία προσπαθούν να προτείνουν αντικείμενα βάσει συμπερασμάτων σχετικά με τις ανάγκες και τις προτιμήσεις του χρήστη.

2.4.1 Συστήματα Προτάσεων βάσει φιλτραρίσματος περιεχομένου (Content based filtering)

Το σύστημα μαθαίνει να κάνει προτάσεις στον χρήστη βασιζόμενο στα χαρακτηριστικά του ίδιου αλλά και των αντικειμένων [8]. Υποθέτει πως αν ένας χρήστης ενδιαφερόταν για κάποιο αντικείμενο στο παρελθόν, θα δείξει ξανά ενδιαφέρον για το ίδιο αντικείμενο στο μέλλον. Παρόμοια αντικείμενα συνήθως είναι ομαδοποιημένα με βάση τα χαρακτηριστικά τους και έτσι το σύστημα είναι σε θέση να προτείνει παρόμοια αντικείμενα στο χρήστη. Τα προφίλ των χρηστών δομούνται με βάση τις αλληλεπιδράσεις που έχουν με τα αντικείμενα. Στοχεύοντας στην αντιστοίχηση των χαρακτηριστικών του προφίλ του χρήστη έναντι των χαρακτηριστικών στοιχείων των αντικειμένων, παρέχει προτάσεις στους χρήστες. Στις περισσότερες περιπτώσεις, τα χαρακτηριστικά στοιχεία των αντικειμένων είναι απλά λέξεις-κλειδιά που εξάγονται από την περιγραφή τους.

Σε ένα σύστημα βάση φιλτραρίσματος περιεχομένου, κάθε χρήστης θεωρείται μοναδικός και κάθε αντικείμενο παρουσιάζεται ως το σύνολο των χαρακτηριστικών του. Για παράδειγμα για την σύσταση ταινιών σε χρήστες, το σύστημα επεξεργάζεται τα χαρακτηριστικά των ταινιών, όπως το γένος της ταινίας, οι ηθοποιοί, το πόσο δημοφιλείς είναι , ο σκηνοθέτης ή η γλώσσα της ταινίας. Τα δεδομένα αυτά χρησιμοποιούνται για να εκπαιδευτεί το σύστημα. Μία γενική απεικόνιση της αρχιτεκτονικής ενός τέτοιου συστήματος φαίνεται παρακάτω.

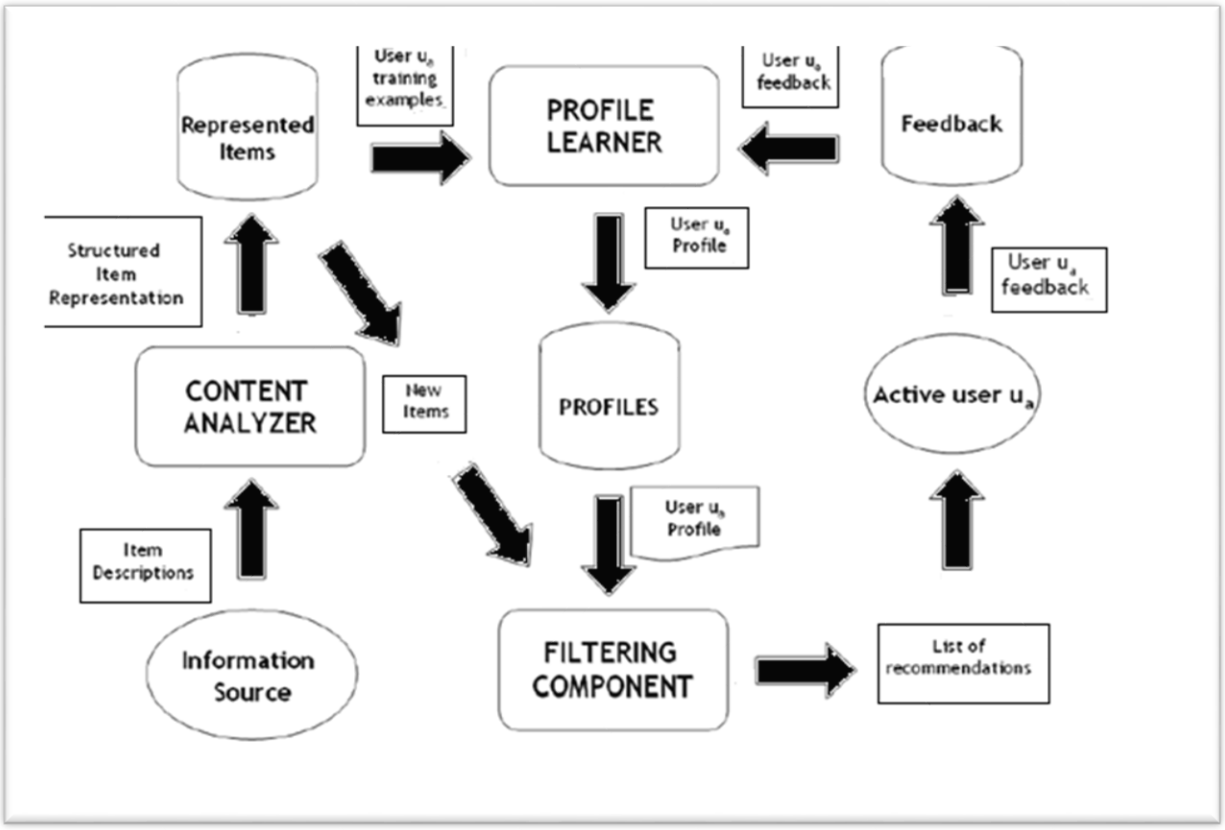

Figure 1. Φιλτράρισμα βάσει Περιεχομένου [40]

Content Analyzer

Ο αναλυτής περιεχομένου βοηθά στην εξαγωγή σχετικών πληροφοριών από την πηγή δεδομένων

όταν δεν έχουν καμία δομή. Ο αναλυτής περιεχομένου προ επεξεργάζεται τα δεδομένα και τα διαμορφώνει για άλλο επίπεδο λειτουργίας. Στην περίπτωση της σύστασης ταινιών, ο αναλυτής περιεχομένου είναι υπεύθυνος να εξάγει από τη λίστα όλων των ταινιών τις λέξεις-κλειδιά που προαναφέραμε, οι οποίες χαρακτηρίζουν κάθε ταινία. Αυτές είναι οι πληροφορίες που θα χρησιμοποιηθούν ύστερα από το σύστημα για την παραγωγή των προτάσεων.

Profile Learner

Ο εκπαιδευτής προφίλ συλλέγει τα δεδομένα που προέρχονται από τον αναλυτή περιεχομένου και γενικεύονται χρησιμοποιώντας διαφορετικές τεχνικές μηχανικής μάθησης, ταιριάζοντάς τα με τις προτιμήσεις του χρήστη και το προφίλ του. Για παράδειγμα, στην περίπτωση

ανάλυσης περιεχομένου ταινιών, όπου ο αναλυτής περιεχομένου παρήγαγε δεδομένα από μία λίστα ταινιών με βάση τους σκηνοθέτες, ο εκπαιδευτής προφίλ για έναν χρήστη μπορεί να δημιουργηθεί εξετάζοντας εάν ο χρήστης έχει παρακολουθήσει στο παρελθόν ταινίες του ίδιου σκηνοθέτη ή όχι.

Filtering Component
Αυτή η ενότητα χρησιμοποιεί το προφίλ χρήστη που δημιουργήθηκε από τον εκπαιδευτή προφίλ του
προηγούμενου βήματος για τη δημιουργία μιας λίστας προτάσεων στον χρήστη.

Τα πλεονεκτήματα των content-based Συστήματα Προτάσεων είναι ότι το μοντέλο δεν χρειάζεται δεδομένα για άλλους χρήστες του συστήματος για να παράγει προτάσεις σε έναν συγκεκριμένο χρήστη καθώς οι προτάσεις προορίζονται αποκλειστικά στον χρήστη αυτόν και δεν στηρίζεται σε υπολογισμούς μεταξύ διαφορετικών χρηστών. Έτσι η επέκτασή τους σε μεγάλο όγκο χρηστών καθίσταται εύκολη. Επίσης το μοντέλο μπορεί να συλλέγει πολύ συγκεκριμένα ενδιαφέροντα ενός χρήστη και μπορεί να του προτείνει πολύ συγκεκριμένα και εξειδικευμένα αντικείμενα για τα οποία ενδιαφέρονται πολύ λίγοι άλλοι χρήστες. Τέλος, σε περιπτώσεις που υπάρχουν πολλές πληροφορίες που περιγράφουν τα αντικείμενα, οι content-based τεχνικές είναι ιδανικές ώστε να αξιοποιήσουν αυτά τα δεδομένα για καλύτερες προτάσεις.

Τα μειονεκτήματα που έχουν παρατηρηθεί για τις content-based μεθόδους είναι τα εξής: υπάρχουν περιπτώσεις όπου το σύστημα δεν έχει πρόσβαση στα χαρακτηριστικά που περιγράφουν τα αντικείμενα ή τους χρήστες. Όπως αναφέραμε και παραπάνω, οι content-based μέθοδοι λειτουργούν αποκλειστικά με βάση αυτά τα χαρακτηριστικά. Επομένως, σε τέτοιες περιπτώσεις, η απόδοση του content-based Συστήματος Προτάσεων θα μειωθεί δραματικά. Αυτό το φαινόμενο ονομάζεται limited content analysis [10].

Ένα επιπλέον πρόβλημα που συναντάται σε τέτοιου είδους Συστήματα Προτάσεων είναι το overspecialization [10]. Ο όρος αυτός αναφέρεται στο φαινόμενο όπου οι προτάσεις που γίνονται σε έναν χρήστη μοιάζουν πολύ μεταξύ τους. Η φύση αυτών των μοντέλων είναι να προτείνουν αντικείμενα στον χρήστη με βάση τα υπάρχοντα ενδιαφέροντα του χρήστη.

Συνεπώς παρατηρείται πως το μοντέλο παρουσιάζει περιορισμένη δυνατότητα επέκτασης στα υπάρχοντα ενδιαφέροντα του χρήστη και περιορίζεται σε προτάσεις αντικειμένων με ίδια χαρακτηριστικά.

2.4.2 Συστήματα Προτάσεων βάσει συνεργατικού φιλτραρίσματος (Collaborative Filtering)

Η πιο συνήθης τεχνική που συναντάται στα Συστήματα Προτάσεων είναι αυτή του collaborative filtering (ή στα ελληνικά συνεργατικό φιλτράρισμα) και αυτό γιατί είναι ευκολότερη η υλοποίηση της αλλά συγχρόνως παρουσιάζει καλύτερα αποτελέσματα από αυτά των content-based τεχνικών [11]. Η λογική που ακολουθείται σε αυτή την προσέγγιση φιλτραρίσματος είναι πως εάν ένας χρήστης «Α» έχει την ίδια γνώμη με έναν χρήστη «Β» για ένα αντικείμενο, τότε ο χρήστης «Α» είναι πιθανότερο να έχει την γνώμη του «Β» για ένα διαφορετικό αντικείμενο σε σχέση με άλλους τυχαίους χρήστες. Από αυτό μπορεί κανείς να καταλάβει πως προκειμένου να παραχθούν προτάσεις για έναν συγκεκριμένο χρήστη, απαιτείται η γνώση και μελέτη πληροφοριών περισσότερων χρηστών. Ο παράγοντας που καθορίζει ποιοι χρήστες θα επηρεάσουν τις προτάσεις που θέλουμε να παρέχουμε σε έναν χρήστη είναι η ομοιότητα μεταξύ του ίδιου και των υπολοίπων.

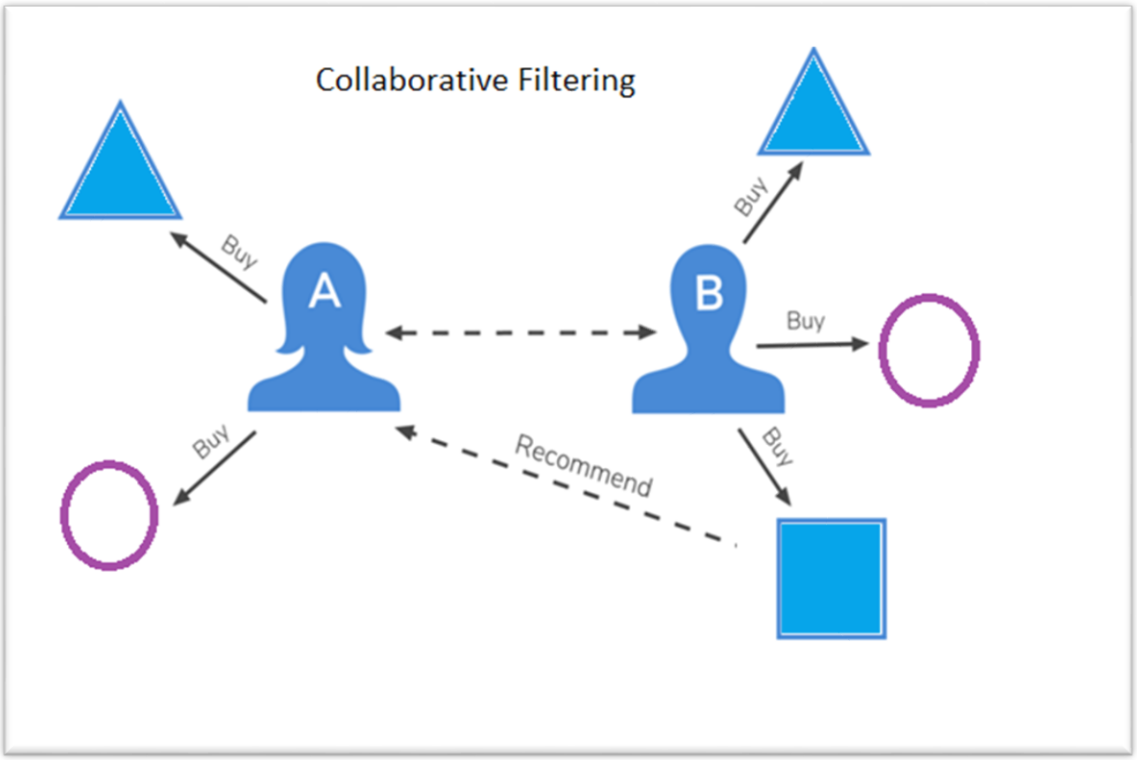

Figure 2. Συνεργατικό Φιλτράρισμα (Collaborative Filtering) [41]

2.4.2.1 Υποκατηγορίες Συστημάτων Προτάσεων βάσει συνεργατικού φιλτραρίσματος

Οι *collaborative filtering* προσεγγίσεις μπορούν να χωριστούν σε 2 υποκατηγορίες ως εξής :

- Memory-based προσεγγίσεις
- Model-based προσεγγίσεις

Οι Memory-based προσεγγίσεις (*Neighbourhood-based*) μπορούν με τη σειρά τους να χωριστούν σε 2 υποκατηγορίες ως εξής:

- User-based collaborative filtering
- Item-based collaborative filtering

Οι *User-based collaborative filtering* προσεγγίσεις χρησιμοποιούν στατιστικές τεχνικές προκειμένου να ομαδοποιήσουν όμοιους χρήστες σε γειτονιές. Έπειτα ένας σταθμισμένος συνδυασμός των βαθμολογιών που έχουν δώσει οι γείτονες του χρήστη χρησιμοποιείται για να παραχθούν προβλέψεις για αυτόν. Ο τρόπος με τον οποίον γίνεται η ομαδοποίηση των χρηστών σε γειτονιές είναι υπολογίζοντας την ομοιότητα μεταξύ τους με διάφορους τρόπους που θα αναφέρουμε παρακάτω. Μία γενικευμένη μεθοδολογία αυτών των προσεγγίσεων παρουσιάζεται παρακάτω:

1) Εκχωρήστε βάρος σε όλους τους χρήστες βάση της ομοιότητας με τον ενεργό χρήστη
2) Επιλέξτε τους "k" χρήστες που έχουν την υψηλότερη ομοιότητα με τον ενεργό χρήστη –συνήθως ονομάζεται γειτονιά
3) Υπολογίστε μια πρόβλεψη από έναν σταθμισμένο συνδυασμό των αξιολογήσεων των επιλεγμένων γειτόνων

Οι *Item-based collaborative filtering* προσεγγίσεις προτάθηκαν για να καταπολεμήσουν κάποια εμπόδια που αντιμετώπιζαν οι *User-based collaborative filtering* προσεγγίσεις. Τα εμπόδια αυτά αφορούσαν συστήματα όπου ο αριθμός των χρηστών είναι κατά πολύ μεγαλύτερος από αυτόν των αντικειμένων. Αυτό οδηγεί σε ανεπιθύμητα αποτελέσματα για τον χρήστη, που είναι να έχει πολλούς αλλά συγχρόνως αναξιόπιστους γείτονες. Ένα άλλο πρόβλημα που παρατηρήθηκε ήταν σε περιπτώσεις όπου μεγάλος αριθμός νέων χρηστών εισερχόταν στο σύστημα και απαιτούσε συνεχή υπολογισμό των παραμέτρων τους. Συνέπεια αυτού ήταν να επηρεάζεται αρνητικά το σύστημα τόσο σε ταχύτητα αλλά όσο και ακρίβεια στις προτάσεις.

Οι *Item-based collaborative filtering* προσεγγίσεις, προβλέπουν τις προτιμήσεις του χρήστη για ένα αντικείμενο βάσει των αξιολογήσεών του για παρόμοια αντικείμενα.

Αμφότερες οι memory-based collaborative filtering προσεγγίσεις συνήθως αντιμετωπίζουν προβλήματα με μεγάλους αραιούς πίνακες (data sparsity), καθώς ο αριθμός των αλληλεπιδράσεων μεταξύ χρηστών-αντικειμένων μπορεί να είναι πολύ χαμηλός για τη δημιουργία συμπλεγμάτων υψηλής ποιότητας.

Οι *Model-based collaborative filtering* προσεγγίσεις [12], σε αντίθεση με τις memory-based που βασίζονται στην ομαδοποίηση χρηστών/αντικειμένων και χρησιμοποιούν απευθείας τις αποθηκευμένες αξιολογήσεις για την παραγωγή προτάσεων, χρησιμοποιούν αυτές τις βαθμολογίες για να μάθουν ένα προγνωστικό μοντέλο. Τα χαρακτηριστικά των χρηστών και των αντικειμένων καταγράφονται από τις παραμέτρους του μοντέλου, οι οποίες μαθαίνονται από τα δεδομένα εκπαίδευσης (training data) και αργότερα χρησιμοποιούνται για την πρόβλεψη αξιολογήσεων. Ένα τέτοιο μοντέλο, δέχεται ως είσοδο τον πίνακα αλληλεπιδράσεων μεταξύ χρηστών-αντικειμένων και μέσω μεθόδων παραγοντοποίησης, μαθαίνει να προβλέπει τις αξιολογήσεις των χρηστών για νέα αντικείμενα.

Ένα πλεονέκτημα των *Model-based collaborative filtering* τεχνικών είναι ότι είναι σε θέση να προτείνουν μεγαλύτερο αριθμό αντικειμένων σε μεγαλύτερο αριθμό χρηστών, σε σύγκριση με τις memory-based προσεγγίσεις. Έχουν δηλαδή μεγάλη κάλυψη (*coverage*), ακόμη και όταν δουλεύουν με μεγάλους και αραιούς πίνακες.

Σε αυτές τις προσεγγίσεις χρησιμοποιούνται τεχνικές που ανήκουν στον τομέα της μηχανικής μάθησης και με σωστή υλοποίηση μπορούν να ξεπεράσουν κατά πολύ την επίδοση άλλων προσεγγίσεων.

Τα μοντέλα λανθάνοντος παράγοντα *(Latent factor model)* είναι τα πιο διαδεδομένα και περισσότερο χρησιμοποιούμενα σε αυτήν την κατηγορία τεχνικών.

Τα *Latent factor* μοντέλα, σε αντίθεση με τις memory-based τεχνικές που παράγουν προτάσεις βάσει των γειτόνων και των ομοιοτήτων μεταξύ χρηστών ή αντικειμένων, υποθέτουν ότι η ομοιότητα προκαλείται ταυτόχρονα από κάποια κρυμμένη και χαμηλότερης διάστασης δομή στα δεδομένα. Για παράδειγμα, η βαθμολογία που δίνει ένας χρήστης σε μια ταινία μπορεί να θεωρηθεί ότι εξαρτάται από κάποιους σιωπηρώς παράγοντες, όπως η προτίμηση του χρήστη για διάφορα είδη ταινιών.

Οι *Matrix Factorization* τεχνικές αποτελούν ένα ευρέως επιτυχημένο *Latent Factor* μοντέλο. Αυτή η οικογένεια μεθόδων έγινε ευρέως γνωστή κατά τη διάρκεια της απόσπασης του

βραβείου Netflix λόγω της αποτελεσματικότητάς της [13]. Πρόκειται για παραγοντοποίηση/διάσπαση του μητρώου user-item το οποίο περιέχει ratings των users στα items, σε χαμηλότερων διαστάσεων μητρώα user και item αντίστοιχα, και στη συνέχεια ανακατασκευή του με τις προβλέψεις για νέα αντικείμενα. Οι χρήστες και τα αντικείμενα αντιπροσωπεύονται ως άγνωστα διανύσματα λανθανουσών χαρακτηριστικών (latent feature vectors) "k" διαστάσεων. Αυτά τα διανύσματα χαρακτηριστικών μαθαίνονται έτσι ώστε τα εσωτερικά γινόμενα να προσεγγίζουν τις γνωστές αξιολογήσεις προτιμήσεων, σε σχέση με κάποιο μέτρο απώλειας (loss function). Πολύ κρίσιμο ρόλο σε αυτά τα Συστήματα Προτάσεων παίζει η ελαχιστοποίηση της απώλειας της αντικειμενικής συνάρτησης (minimizing objective function) και επιτυγχάνεται με διάφορους αλγορίθμους με πιο γνωστούς να είναι οι: Stochastic Gradient Descent (SGD) και Alternating Least Square(ALS).

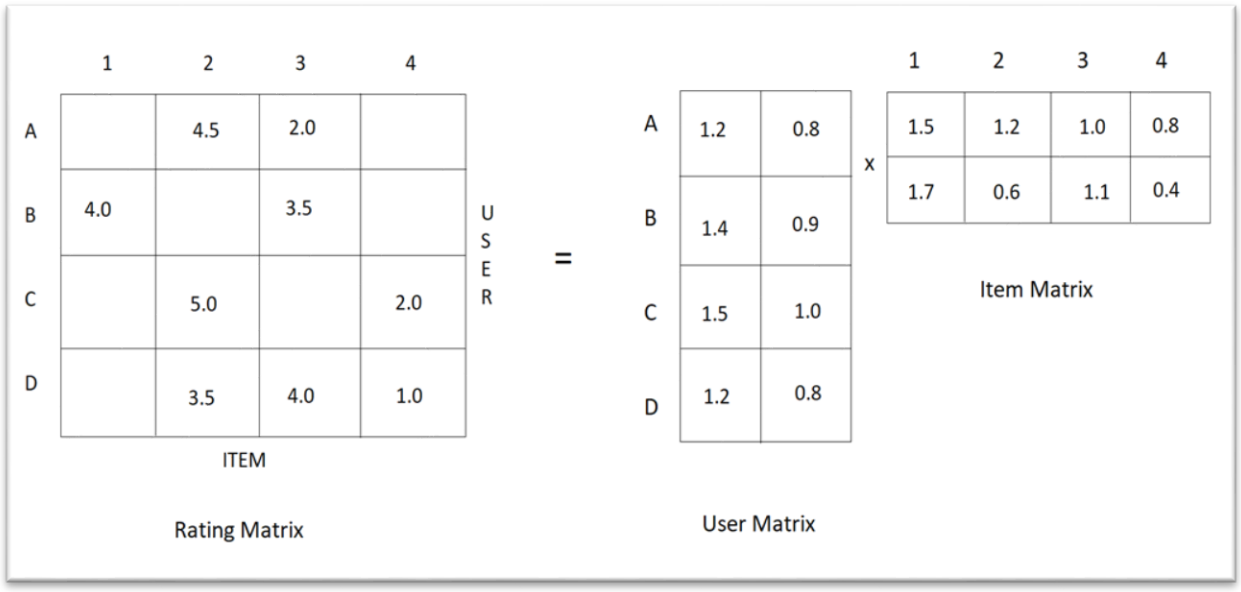

Figure 3. Παράδειγμα λειτουργίας Matrix Factorization [39]

Υπάρχουν διάφορες τεχνικές που χρησιμοποιούνται για την αποσύνθεση του πίνακα αλληλεπιδράσεων χρηστών-αντικειμένων σε μητρώα χρήστη και αντικειμένου ξεχωριστά και στη συνέχεια για την ανακατασκευή τους με τις προβλεπόμενες αξιολογήσεις όπως το κλασσικό Matrix Factorization, Observed Only Matrix Factorization, Weighted Matrix Factorization, Singular Value Decomposition και άλλα [14].

Ένα πολύ σημαντικό πλεονέκτημα των μεθόδων που βασίζονται σε *collaborative filtering* για την παραγωγή προτάσεων είναι ότι βοηθούν τους χρήστες να ανακαλύψουν νέα ενδιαφέροντα παράγοντας πρωτότυπες προτάσεις (*serendipity*).

Ωστόσο, παρά τις βελτιώσεις που έχουν επιφέρει στα Συστήματα Προτάσεων, οι *collaborative filtering* προσεγγίσεις έχουν παρουσιάσει ορισμένα προβλήματα, τα οποία θα αναφέρουμε σε επόμενη υπό-ενότητα, με το πιο σοβαρό να είναι το λεγόμενο *cold start problem* [15]. Με τον όρο αυτό αναφερόμαστε στην αδυναμία του Συστήματος Προτάσεων να λειτουργήσει αποδοτικά για χρήστες που μόλις έχουν εισέλθει στο σύστημα ή παρόμοια για νέα αντικείμενα για τα οποία δεν υπάρχουν αξιολογήσεις. Καθώς οι *collaborative filtering* τεχνικές βασίζονται στην αλληλεπίδραση μεταξύ των χρηστών και των αντικειμένων και απαιτούν την ύπαρξη δεδομένων για τη λειτουργία τους το πρόβλημα αυτό τείνει να γίνεται πολύ έντονο.

2.4.2.2 Προκλήσεις των Συστημάτων Προτάσεων βάσει συνεργατικού φιλτραρίσματος

Λόγω του μεγάλου αριθμού δεδομένων και πληροφοριών στο Διαδίκτυο, ειδικά στον τομέα του διαδικτυακού μάρκετινγκ όπως το Amazon, τα Συστήματα Προτάσεων όπως έχουμε ήδη αναφέρει έχουν γίνει πολύ σημαντικά. Τα Σύστημα Προτάσεων συχνά αυξάνουν την παραγωγή πωλήσεων των εταιρειών και βοηθά τους πελάτες να επιλέξουν το κατάλληλο προϊόν ενδιαφέροντός τους.
Ωστόσο, υπάρχουν κύριες προκλήσεις [16] που σχετίζονται με την απόδοση του συστήματος όπως η αραιότητα των δεδομένων, το cold start problem και η κλιμάκωση.

Cold start problem

Το cold start problem, ίσως το πιο σημαντικό εμπόδιο που συναντάται σε Συστήματα Προτάσεων συνεργατικού φιλτραρίσματος, είναι ένα είδος προβλήματος αδυναμίας που εμφανίζεται όταν ένας νέος χρήστης ή ένα αντικείμενο μόλις εισήλθε στο σύστημα. Στην περίπτωση προσέγγισης συνεργατικού φιλτραρίσματος, είναι προβληματικό να γίνει πρόταση σε έναν νέο χρήστη, για τον οποίο υπάρχουν πολύ περιορισμένες πληροφορίες, όπως επίσης για ένα νέο αντικείμενο, συνήθως δεν είναι αρκετές αξιολογήσεις
διαθέσιμες και ως εκ τούτου οδηγούν σε μια ασθενώς και μη χρήσιμη σύσταση. Έχουν αναπτυχθεί πολλές στρατηγικές για τον μετριασμό αυτού του προβλήματος. Η κύρια προσέγγιση είναι να βασιστούμε σε υβριδικά Συστήματα Προτάσεων, προκειμένου να μετριάσουμε τα μειονεκτήματα μιας κατηγορίας ή μοντέλου συνδυάζοντάς το με μια άλλη. Όταν ένα αντικείμενο ή χρήστης δεν παρέχουν αρκετή πληροφορία χαρακτηρίζεται ως “cold user/item” ενώ όταν έχουμε στη διάθεσή μας αρκετές πληροφορίες ονομάζεται “warm user/item”. Μια κοινή στρατηγική όταν ασχολούμαστε με νέα αντικείμενα είναι να συνδυάσουμε ένα συνεργατικού φιλτραρίσματος σύστημα, για ζεστά αντικείμενα, με ένα βάσει φιλτραρίσματος περιεχομένου, για κρύα αντικείμενα. Ενώ οι δύο αλγόριθμοι μπορούν να συνδυαστούν με διαφορετικούς τρόπους, το κύριο μειονέκτημα αυτής της μεθόδου σχετίζεται με την κακή ποιότητα των προτάσεων που συχνά παρουσιάζονται εξαιτίας του μέρους του συστήματος που λειτουργεί βάσει περιεχομένου σε σενάρια όπου είναι δύσκολο να παρέχεται μια ολοκληρωμένη περιγραφή των χαρακτηριστικών του αντικειμένου. Σε περίπτωση νέων χρηστών, εάν δεν υπάρχει δημογραφική περιγραφή ή η ποιότητά της είναι πολύ κακή, μια κοινή στρατηγική είναι να τους προσφέρεται μη εξατομικευμένες προτάσεις. Αυτό σημαίνει ότι θα μπορούσαν να προταθούν απλά τα πιο δημοφιλή αντικείμενα είτε παγκοσμίως είτε για τη συγκεκριμένη γεωγραφική περιοχή ή γλώσσα τους.

Data sparsity

Το πρόβλημα της αραιότητας των δεδομένων είναι ένα από τα μεγαλύτερα προβλήματα που αντιμετωπίζουν τα Σύστημα Προτάσεων και παρουσιάζει μεγάλο αντίκτυπο στην ποιότητα των προτάσεων. Ο αριθμός των αντικειμένων που είναι διαθέσιμα στο Διαδίκτυο έχει γίνει τεράστιο. Οι περισσότεροι χρήστες δεν έχουν βαθμολογήσει αρκετά μεγάλο αριθμό στοιχείων

που σημαίνει ότι το μητρώο χρήστη-στοιχείου (user-item matrix), το οποίο είναι υπεύθυνο για την απόδοση του συστήματος, μπορεί να έχει πολλές βαθμολογίες που λείπουν και να γίνει πολύ μεγάλο και αραιό.

Επομένως, η εύρεση ομοιότητας μεταξύ χρηστών ή στοιχείων γίνεται πολύ δύσκολη και μπορεί να οδηγήσει σε αδύναμες προτάσεις.

Προκειμένου να αντιμετωπιστεί αυτή η αραιότητα των δεδομένων που είναι επιβλαβής για την παροχή προτάσεων, έχουν προταθεί διάφορες τεχνικές με πιο διαδεδομένες να είναι αυτές της μείωσης των διαστάσεων των δεδομένων. Ως παράδειγμα μπορούμε να φέρουμε την τεχνική Singular Value Decomposition (SVD), η οποία αφαιρεί το ασήμαντο και μη αντιπροσωπευτικό μέρος των χρηστών και αντικειμένων, απευθείας από τη λίστα. Αυτό διευκολύνει τη χαρτογράφησης της ομοιότητας μεταξύ των χρηστών και του αντικειμένου αφού ο χώρος είναι μειωμένος και τα δεδομένα είναι λιγότερο αραιά. Μια άλλη μέθοδος, το Principal Component Analysis (PCA), μειώνει τη διάσταση του συνόλου δεδομένων εξασφαλίζοντας τη διατήρηση των σημαντικών συστατικών. Αυτό, ωστόσο, μπορεί να αφαιρέσει πολύ σημαντικούς χρήστες ή στοιχεία που περιέχουν πολύ σημαντικές πληροφορίες και συνεπώς να μειώσει την απόδοση του συστήματος.

Scalability

Η επεκτασιμότητα αναφέρεται στην αύξηση του αριθμού των αντικειμένων και των χρηστών. Τα σετ δεδομένων έχουν αναπτυχθεί δραματικά και συνεπώς έγινε πολύ δύσκολο να χρησιμοποιηθούν το παραδοσιακοί αλγόριθμοι συνεργατικού φιλτραρίσματος για την αντιμετώπισή του. Το κύριο εμπόδιο είναι ότι δεν θα υπάρχουν αρκετοί υπολογιστικοί πόροι που να πληρούν τις νέες απαιτήσεις του τεράστιου αυτού όγκου δεδομένων. Υπάρχουν μερικοί αλγόριθμοι που ασχολούνται με τον αυξανόμενο αριθμό χρηστών και αντικειμένων, αλλά συνοδεύονται από αύξηση των υπολογισμών, μεγαλύτερο κόστος και μερικές φορές οδηγούν σε ανακριβή αποτελέσματα. Για παράδειγμα, οι τεχνικές μείωσης διαστάσεων (PCA/SVD) μπορούν να χειριστούν αυτό το πρόβλημα επεκτασιμότητας, αλλά χρειάζεται κάποια διαδικασία όπως, κατασκευή παραγοντοποίησης μητρώου, η οποία είναι πολύπλοκη και δαπανηρή.

Synonymy

Η συνωνυμία αναφέρεται στην περίπτωση όπου όμοια αντικείμενα σε ένα σύνολο δεδομένων έχουν διαφορετική ονομασία.. Τα Συστήματα Προτάσεων συνήθως δεν είναι σε θέση να ανακαλύψουν αυτήν τη σχέση μεταξύ τους και τείνουν να συμπεριφέρονται προς αυτά ως διαφορετικά αντικείμενα.

Gray Sheep

Η πρόκληση αυτή σχετίζεται με την προσέγγιση συνεργατικού φιλτραρίσματος, ειδικά,
όταν χρησιμοποιείται η έννοια των ομαδοποιήσεων (clusters). Εμφανίζεται όταν ορισμένοι χρήστες δεν
συμφωνούν ή διαφωνούν με οποιοδήποτε σύμπλεγμα ανθρώπων. Αποτέλεσμα αυτού είναι να μην μπορούν να ωφεληθούν από τις τεχνικές του συνεργατικού φιλτραρίσματος καθώς δε συμμερίζονται ενδιαφέροντα με άλλους χρήστες.

2.4.3 Υβριδικά Συστήματα Προτάσεων

Στην προσπάθεια παροχής ακριβέστερων προτάσεων, έχουν προταθεί διάφορες υβριδικές προσεγγίσεις. Ένας υβριδικός αλγόριθμος συνδυάζει διάφορες προσεγγίσεις προκειμένου να παρέχει πιο ακριβή αποτελέσματα. Η δομή μιας τέτοιας εφαρμογής διαφέρει ανάλογα με τη στρατηγική υβριδικής προσέγγισης. Η πιο διαδεδομένη υβριδική προσέγγιση ενσωματώνει χαρακτηριστικά αξιολόγησης από την προσέγγιση βάσει περιεχομένου *(content-based filtering)* και τη συνεργατική προσέγγιση φιλτραρίσματος *(collaborative filtering)* [17].

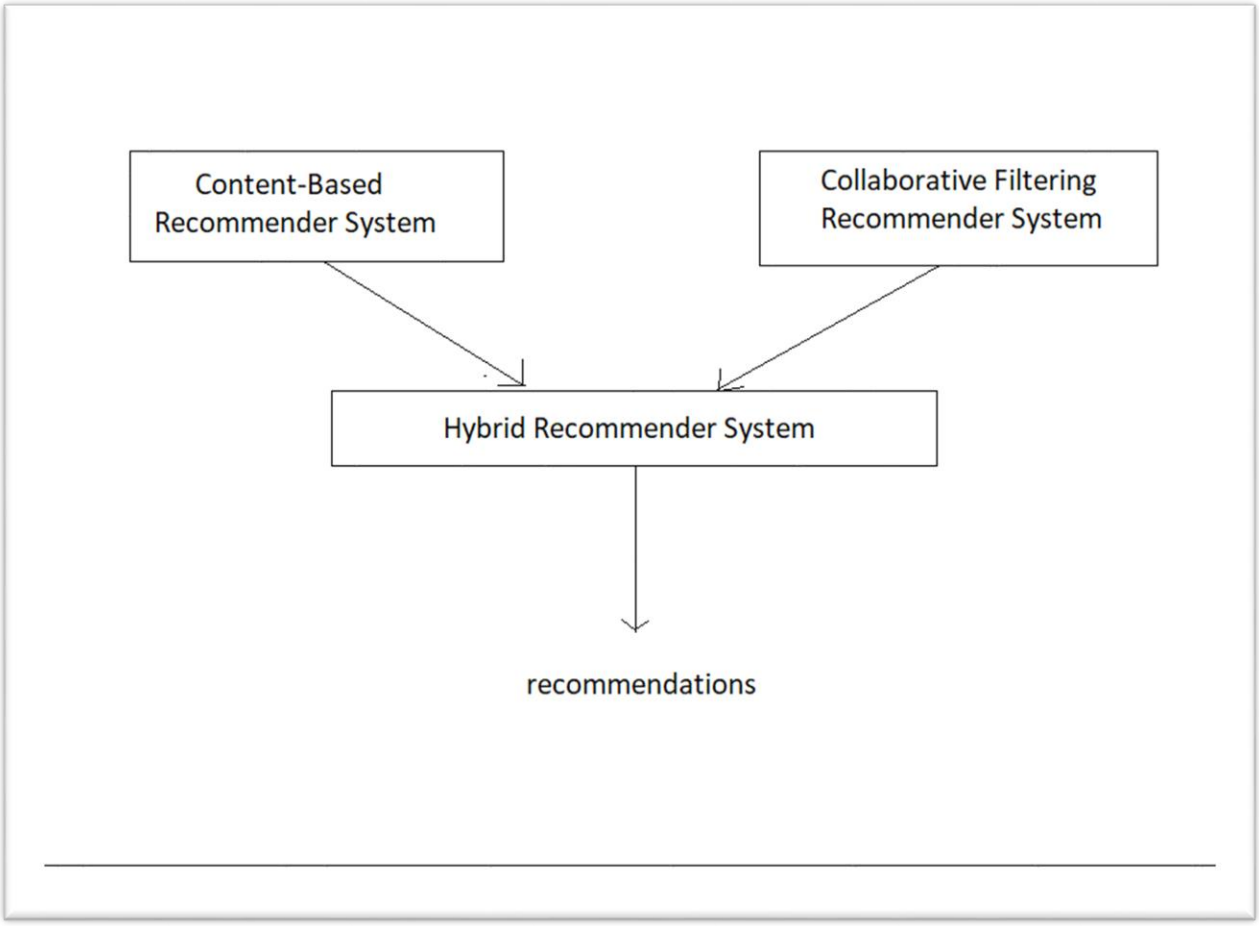

Figure 4. Υβριδικό μοντέλο Συστήματος Προτάσεων (Hybrid Recommender System)

2.4.4 Επιπλέον τύποι Συστημάτων Προτάσεων

Πέραν των Συστημάτων Προτάσεων που έχουμε αναφέρει ως τώρα στην παρούσα διπλωματική εργασία, υπάρχουν και άλλες προσεγγίσεις που διαφοροποιούν την παροχή προτάσεων.

Η συντριπτική πλειονότητα των υπαρχουσών προσεγγίσεων επικεντρώνεται στη σύσταση αντικειμένων στους χρήστες και δεν λαμβάνουν υπόψη τυχόν πρόσθετες πληροφορίες με βάση

ένα γενικότερο πλαίσιο, όπως ο χρόνος, ο τόπος, η εταιρεία άλλων ατόμων. Πρόκειται για τα Context-aware Συστήματα Προτάσεων τα οποία ασχολούνται με τη μοντελοποίηση και την πρόβλεψη των προτιμήσεων των χρηστών
ενσωματώνοντας τις διαθέσιμες πληροφορίες με βάση ένα γενικότερο πλαίσιο στη διαδικασία σύστασης ως πρόσθετες κατηγορίες δεδομένων. Αυτές οι μακροπρόθεσμες προτιμήσεις
και τα γούστα εκφράζονται συνήθως ως βαθμολογίες και μοντελοποιούνται ως συνάρτηση όχι μόνο μεταξύ αντικειμένων και χρηστών, αλλά και του γενικότερου πλαισίου. Με άλλα λόγια, οι βαθμολογίες ορίζονται με μία επιπλέον προσθήκη σε σχέση με τη γενικότερη συνάρτηση χρησιμότητας που αναφέραμε στο κεφάλαιο 2.1 ως εξής:

$$u: S \times I \times C \rightarrow R$$

Όπου πέραν του συνόλου των χρηστών “S” και αντικειμένων “I” προστίθεται ο παράγοντας “C” που είναι οι επιπλέον πληροφορίες με βάση το γενικότερο πλαίσιο προκειμένου να οδηγηθούμε στην πρόβλεψη της αξιολόγησης.

Ένας άλλος τύπος Συστημάτων Προτάσεων αυτά της δημογραφικής κατηγοριοποίησης (Demographic Recommender Systems). Σε σύγκριση με τις άλλες προσεγγίσεις, το πλεονέκτημα αυτού του τύπου συστήματος είναι ότι χρησιμοποιεί μόνο τα δημογραφικά δεδομένα χρηστών όπως το φύλο, ηλικία, εκπαίδευση κ.λπ. και ενδέχεται να μην χρειάζεται το ιστορικό των αξιολογήσεων των χρηστών, την περιγραφή κειμένου ή τη γνώση των αντικειμένων. Επομένως, νέοι χρήστες μπορούν να αποκτήσουν προτάσεις προτού αξιολογήσουν τυχόν στοιχεία.

2.5 Μέτρα Ομοιότητας

Τόσο το content-based filtering όσο και το collaborative filtering χαρτογραφούν κάθε στοιχείο σε έναν διάνυσμα ενσωμάτωσης (embedding vector) σε έναν κοινό χώρο ενσωμάτωσης (embedding space) [18]. Συνήθως, ο χώρος ενσωμάτωσης είναι χαμηλής διάστασης και καταγράφει κάποια λανθάνουσα δομή του αντικειμένου. Παρόμοια αντικείμενα, όπως για παράδειγμα βίντεο του YouTube που συνήθως παρακολουθούνται από τον ίδιο χρήστη, ή σύνολα βιβλίων που προτιμά ένας χρήστης, καταλήγουν κοντά στον χώρο ενσωμάτωσης. Η έννοια της «εγγύτητας» ορίζεται από ένα μέτρο ομοιότητας. Ένα μέτρο ομοιότητας είναι μια συνάρτηση που παίρνει ένα ζευγάρι διανυσμάτων ενσωμάτωσης και επιστρέφει μια βαθμίδα που μετρά την ομοιότητά τους. Για παράδειγμα, δεδομένου ότι ένα αντικείμενο αντιστοιχίζεται σε έναν διάνυσμα ενσωμάτωσης "y", το σύστημα αναζητά διανύσματα ενσωμάτωσης αντικειμένων "x" που είναι κοντά στο "y", δηλαδή ενσωματώσεις με υψηλή ομοιότητα. Το πιο σημαντικό και κρίσιμο βήμα στους αλγόριθμους συνεργατικού φιλτραρίσματος (collaborative filtering) είναι η εύρεση παρόμοιων στοιχείων και χρηστών. Αφού βρεθούν παρόμοιοι χρήστες και αντικείμενα, είναι εύκολο να αποφανθεί κανείς για την ομοιότητα μεταξύ αυτών των χρηστών και αντικειμένων και τελικά να επιλέξει μια ομάδα από χρήστες και αντικείμενα που μοιάζουν περισσότερο με τον χρήστη-στόχο.

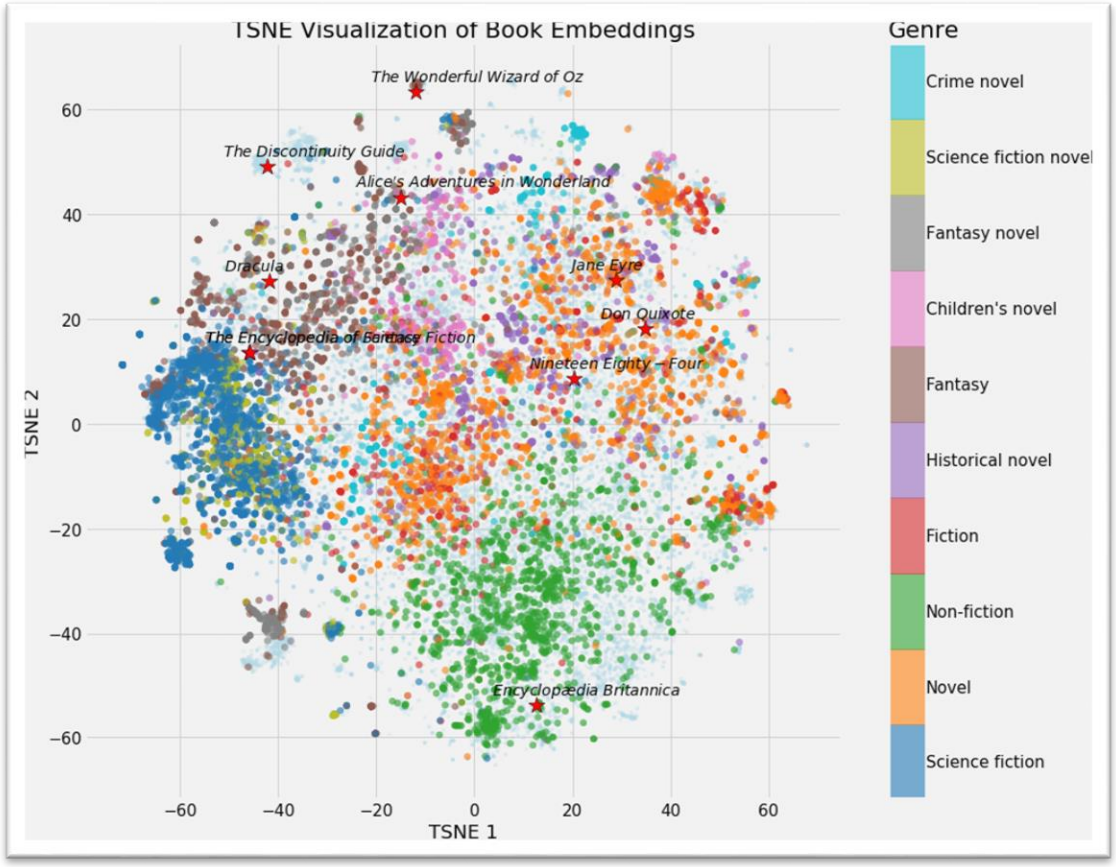

Figure 5. Παράδειγμα του Χώρου Ενσωμάτωσης με κατηγορίες βιβλίων (Embedding Space) [38]

Για να γίνει πιο σαφές με ένα απλοϊκό παράδειγμα από την ομάδα developers.google [18], ας σκεφτούμε πως έχουμε ένα Σύστημα Προτάσεων ταινιών βασισμένο στο Collaborative Filtering στο οποίο τα δεδομένα εκπαίδευσης αποτελούνται από έναν πίνακα σχολίων στο οποίο:

- Κάθε σειρά αντιπροσωπεύει έναν χρήστη.
- Κάθε στήλη αντιπροσωπεύει ένα στοιχείο (μια ταινία).

Τα σχόλια για τις ταινίες εμπίπτουν σε μία από τις δύο κατηγορίες:

- Άμεσα, Ρητή ανατροφοδότηση (implicit feedback) - οι χρήστες καθορίζουν πόσο τους άρεσε μια συγκεκριμένη ταινία παρέχοντας μια αριθμητική βαθμολογία.

- Σιωπηρή, Σιωπηρή ανατροφοδότηση - εάν ένας χρήστης παρακολουθεί μια ταινία, το σύστημα υπονοεί ότι τον ενδιαφέρει η ταινία αυτή.

Για απλοποίηση, θα υποθέσουμε ότι ο πίνακας ανατροφοδότησης είναι δυαδικός. Δηλαδή, η τιμή 1 υποδηλώνει ενδιαφέρον για την ταινία.

Όταν ένας χρήστης επισκέπτεται την αρχική σελίδα, το σύστημα θα πρέπει να προτείνει ταινίες με βάση και τα δύο:

- ομοιότητα με ταινίες που άρεσε ο χρήστης στο παρελθόν
- ταινίες που άρεσαν σε παρόμοιοι χρήστες

Ας υποθέσουμε ότι εκχωρούμε σε κάθε ταινία μια βαθμίδα που περιγράφει εάν η ταινία είναι για παιδιά (αρνητικές τιμές) ή ενήλικες (θετικές τιμές). Ας υποθέσουμε ότι εκχωρούμε επίσης μια κλίμακα σε κάθε χρήστη που περιγράφει το ενδιαφέρον του χρήστη για παιδικές ταινίες (κοντά στο -1) ή ταινίες ενηλίκων (πλησιέστερα στο +1). Το γινόμενο της ενσωμάτωσης (dot product) ταινιών και η ενσωμάτωση χρήστη πρέπει να είναι υψηλότερα (πλησιέστερα στο 1) για ταινίες που περιμένουμε να αρέσουν στο χρήστη.

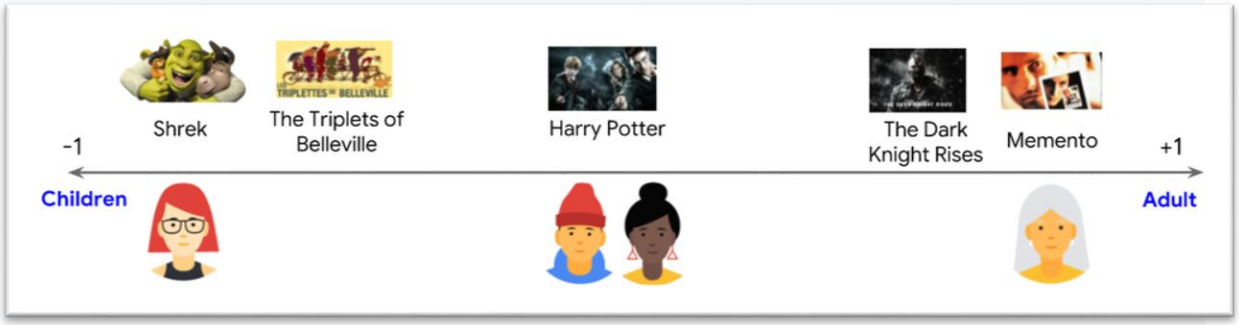

Figure 6. Μονοδιάστατος Χώρος Ενσωμάτωσης (1-D embedding space) [14]

Χρήστες οι οποίοι απολαμβάνουν και τις 2 από τις κατηγορίες βρίσκονται κοντά στο κέντρο του διανυσματικού χώρου. Φυσικά ένας χώρος ενσωμάτωσης δεν είναι ποτέ τόσο απλός. Έχει πολλά χαρακτηριστικά τα οποία καθορίζουν τη θέση του κάθε χρήστη μέσα σε αυτόν.

Στο παρακάτω διάγραμμα, κάθε σημάδι επιλογής προσδιορίζει μια ταινία που παρακολούθησε ένας συγκεκριμένος χρήστης. Ο τρίτος και ο τέταρτος χρήστης έχουν προτιμήσεις που εξηγούνται καλά από αυτήν τη λειτουργία - ο τρίτος χρήστης προτιμά ταινίες για παιδιά και ο τέταρτος χρήστης προτιμά ταινίες για ενήλικες. Ωστόσο, οι προτιμήσεις του πρώτου και του δεύτερου χρήστη δεν εξηγούνται καλά από αυτό το μοναδικό χαρακτηριστικό.

Figure 7. Μητρώο Ανατροφοδότησης με 1 feature (Feedback Matrix) [14]

Ένα χαρακτηριστικό δεν ήταν αρκετό για να εξηγήσει τις προτιμήσεις όλων των χρηστών. Για να ξεπεράσουμε αυτό το πρόβλημα, ας προσθέσουμε ένα δεύτερο χαρακτηριστικό: τον βαθμό στον οποίο κάθε ταινία είναι ένα blockbuster ή μια ταινία arthouse. . Όπου με τον όρο 'Blockbuster' εννοούμε πόσο εμπορική επιτυχία είναι η ταινία. Με ένα δεύτερο

χαρακτηριστικό, μπορούμε πλέον να αναπαραστήσουμε κάθε ταινία με τις ακόλουθες δισδιάστατες ενσωματώσεις:

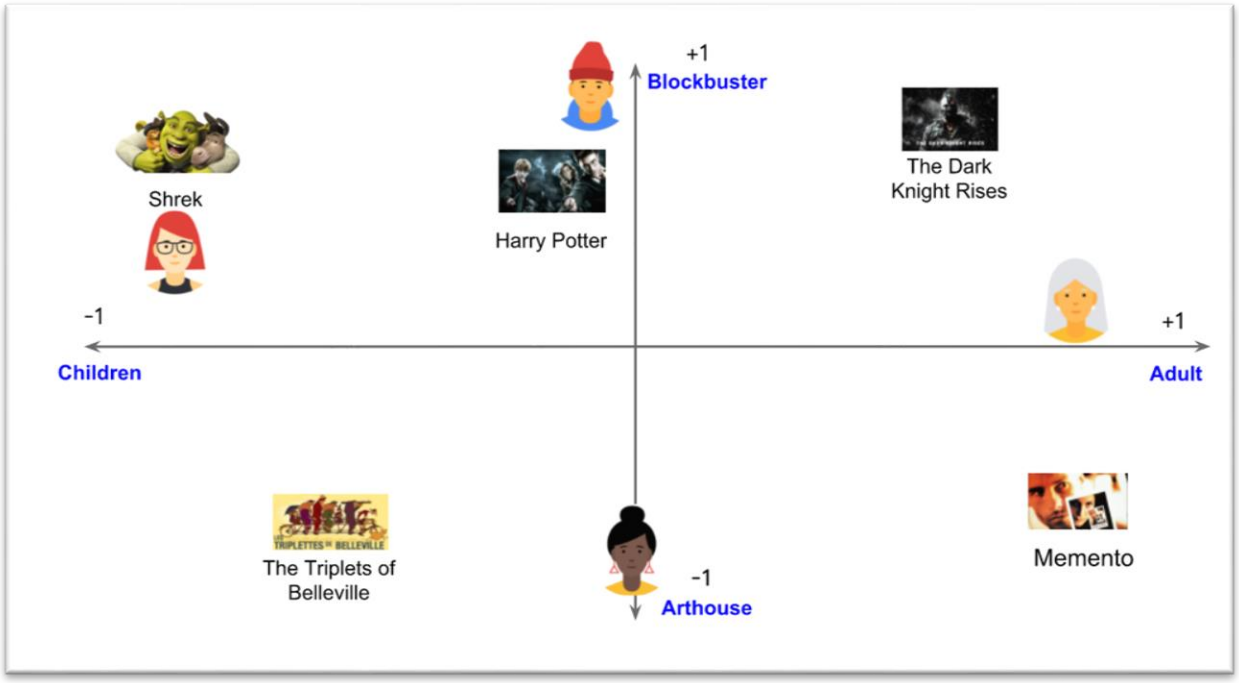

Figure 8. Δισδιάστατος Χώρος Ενσωμάτωσης 2-D (2-D Embedding Space) [14]

Τοποθετούμε και πάλι τους χρήστες μας στον ίδιο χώρο ενσωμάτωσης για να εξηγήσουμε καλύτερα τη συμπεριφορά τους: για κάθε ζεύγος (χρήστης, στοιχείο), θα θέλαμε το προϊόν (dot product) της ενσωμάτωσης χρήστη και της ενσωμάτωσης στοιχείου να είναι κοντά στο 1 όταν ο χρήστης παρακολούθησε την ταινία και κοντά στο 0 διαφορετικά. Φυσικά, όσα περισσότερα χαρακτηριστικά έχουμε για την περιγραφή του γούστου των χρηστών, τόσο πιο εξατομικευμένες και ακριβείς θα είναι οι προτάσεις που θα παραχθούν.

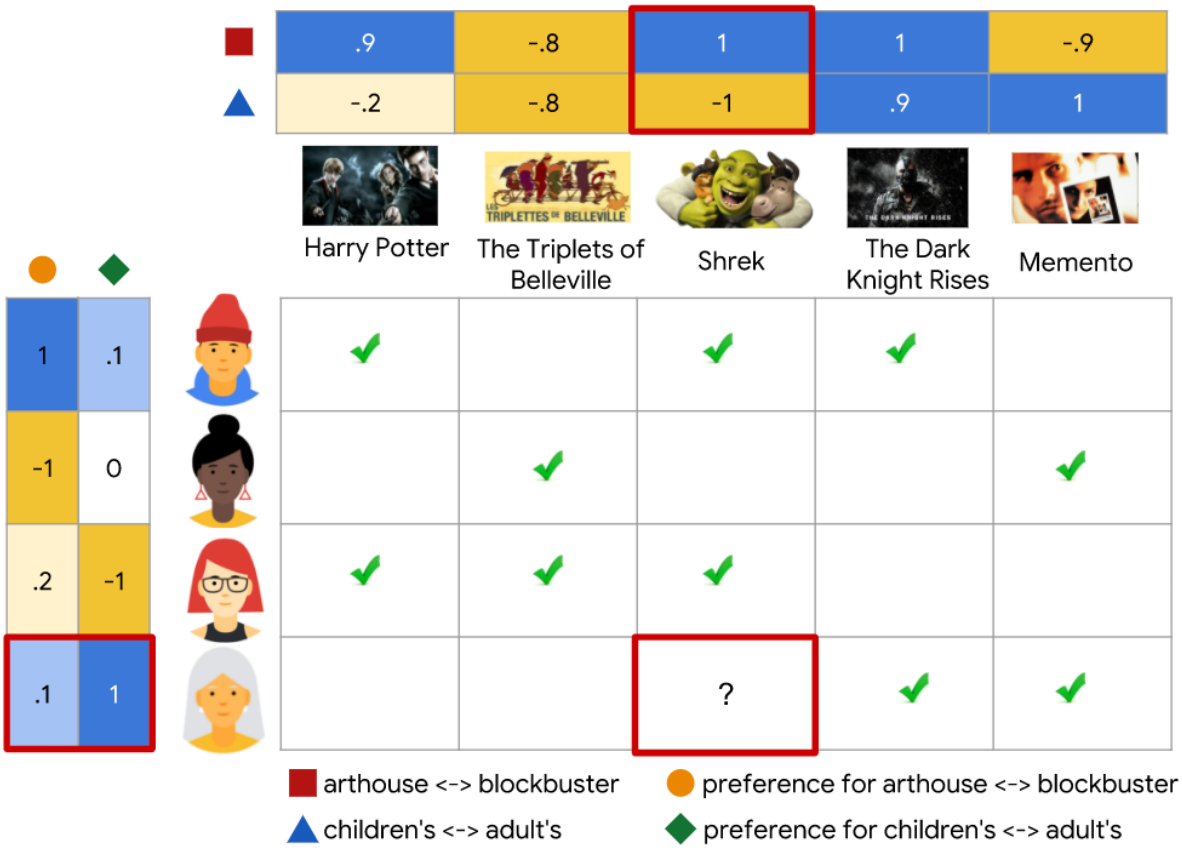

Figure 9. Μητρώο Ανατροφοδότησης με δύο features (Feedback Matrix) [14]

Το Matrix Factorization είναι ένα απλό μοντέλο ενσωμάτωσης. Δεδομένου του πίνακα ανατροφοδότησης A, το μοντέλο μαθαίνει:

- Έναν πίνακα ενσωμάτωσης χρήστη (user embedding)
- Έναν πίνακα ενσωμάτωσης στοιχείου (item embedding)

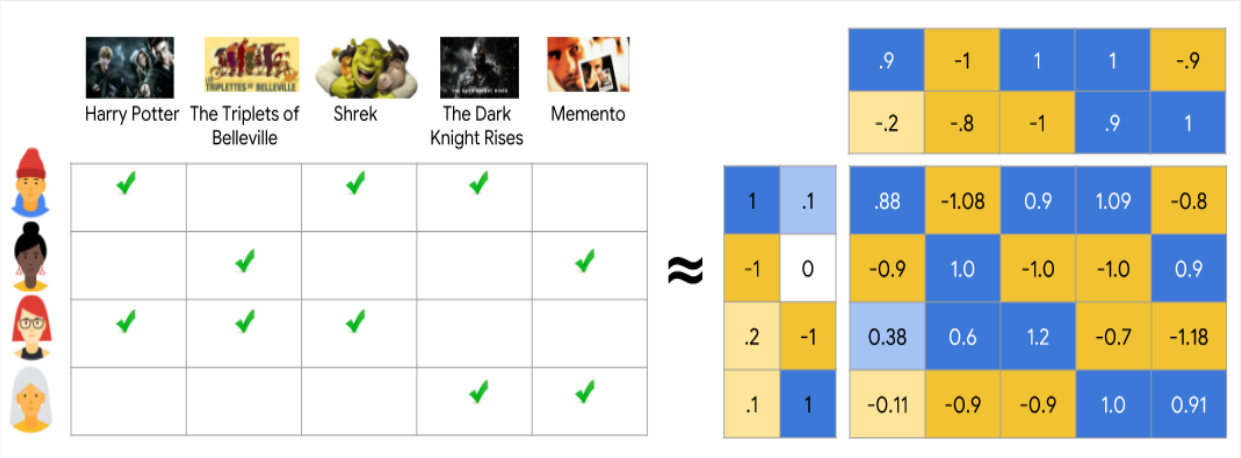

Figure 10. Matrix Factorization [User embedding - Item Embedding] [14]

Παρατηρούμε πως ο πίνακας Α περιέχει το dot product των τιμών της αρεσκείας του χρήστη και των χαρακτηριστικών της ταινίας. Οι ενσωματώσεις των χρηστών και των ταινιών μαθαίνονται με βάση τον πίνακα Α, με τέτοιον τρόπο ώστε το dot product τους να είναι μία καλή προσέγγιση του πίνακα Α. Ένα τεράστιο πλεονέκτημα που παρέχει η τεχνική του Matrix Factorization που εξηγήσαμε παραπάνω με ένα πολύ απλό παράδειγμα είναι πως από το feedback Matrix Α, το οποίο είναι πολύ αραιό καθώς οι χρήστες σπανίως αξιολογούν αντικείμενα, μας οδηγεί σε μία πολύ πιο συμπαγή αναπαράσταση.

Για να προσδιοριστεί ο βαθμός ομοιότητας, τα περισσότερα Συστήματα Προτάσεων βασίζονται σε ένα ή περισσότερα από τα ακόλουθα ανάλογα με τον τομέα που ερευνούν και το είδος των πληροφοριών που χρησιμοποιούν:

- *Cosine distance*

Πρόκειται για το συνημίτονο της γωνίας μεταξύ των διανυσμάτων ενσωμάτωσης [19]. Τα διανύσματα αυτά αναπαριστούν είτε τους χρήστες είτε τα αντικείμενα για τα οποία θέλουμε να εκτιμήσουμε την ομοιότητά τους. Συνήθως η μέτρηση ομοιότητας με βάση το συνημίτονο χρησιμοποιείται για την εκτίμηση της ομοιότητας μεταξύ δύο παρουσιών α και β στην

ανάκτηση πληροφοριών που το τα αντικείμενα έχουν μορφή διανύσματος και υπολογίζει την ομοιότητα του Cosine Vector (CV) (ή Vector Space) μεταξύ τους.
Αυτοί οι φορείς δείχνουν την απόσταση μεταξύ τους.

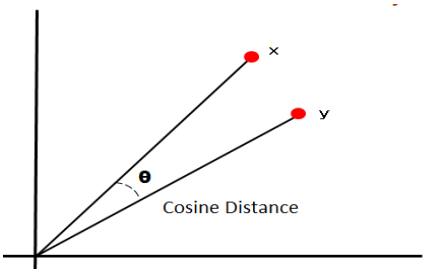

Figure 11. Γραφική Αναπαράσταση του Cosine Distance

$$cos\,(x,y) = \frac{(x.y)}{||x||\,||y||}$$

Equation 1. Εξίσωση του Cosine Distance

Όπου το (x.y) αναπαριστά το εσωτερικό γινόμενο των διανυσμάτων, ενα τα ||x|| και ||y|| τις νόρμες. Ένα πλεονέκτημα αυτής της μεθόδου είναι η χαμηλή πολυπλοκότητά του, ειδικά για αραιά διανύσματα καθώς πρέπει να ληφθούν υπόψη μόνο οι μη μηδενικές διαστάσεις.

- *Minkowski distance*

‘Όταν η διάσταση ενός σημείου δεδομένων είναι αριθμητική, η γενική φόρμα που χρησιμοποιείται ονομάζεται απόσταση Minkowski [19].

$$d(x,y) = (\sum_{i}^{n}(|x_i - y_i|)^q)^{\frac{1}{q}}$$

Equation 2. Εξίσωση της Minkowski Distance

Η μέτρηση Minkowski απόσταση ή Minkowski είναι μια μέτρηση σε ένα διαμορφωμένο φορέα διανυσμάτων που μπορεί να θεωρηθεί ως γενίκευση τόσο της Ευκλείδειας (q=2) απόστασης όσο και της απόστασης του Μανχάταν (q=1).

- *Manhattan distance*

Η Μανχάταν απόσταση είναι η απόσταση μεταξύ δύο σημείων που μετράται κατά μήκος των αξόνων σε ορθή γωνία [19].

$$d(x,y) = \sum_{i}^{n} |x_i - y_i|$$

Equation 3. Εξίσωση της Manhattan Distance

Πρόκειται ουσιαστικά για την Minkowski απόσταση με q=1.

- *Euclidean distance*

Στα μαθηματικά, η Ευκλείδεια απόσταση μεταξύ δύο σημείων στον Ευκλείδειο χώρο είναι το μήκος ενός τμήματος γραμμής μεταξύ των δύο σημείων [19]. Μπορεί να υπολογιστεί από τις Καρτεσιανές συντεταγμένες των σημείων χρησιμοποιώντας το Πυθαγόρειο θεώρημα, επομένως περιστασιακά ονομάζεται Πυθαγόρεια απόσταση.

Η ευκλείδεια απόσταση χρησιμοποιείται κατά κόρον στα Συστήματα Προτάσεων και είναι η τετραγωνική ρίζα του αθροίσματος των τετραγώνων της διαφοράς μεταξύ των συντεταγμένων και δίνεται από το Πυθαγόρειο θεώρημα.

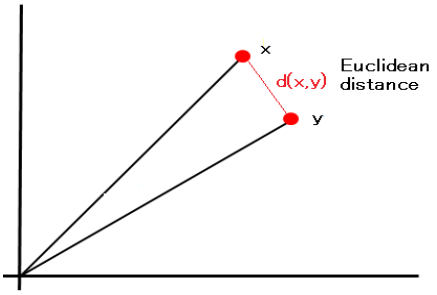

Figure 12. Γραφική Αναπαράσταση της Ευκλείδειας Απόστασης

$$d(x,\ y)\ =\ \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Equation 4. Εξίσωση της Ευκλείδειας Απόστασης

Άλλες ονομασίες για την ευκλείδεια απόσταση είναι επίσης L1-norm / L1-distance / ευθύγραμμη απόσταση.

- *Pearson Correlation Coefficient*

Πρόκειται για ένα μέτρο συσχέτισης μεταξύ δύο τυχαίων μεταβλητών και εύρους μεταξύ [-1, 1]. Για -1 η συσχέτιση μεταξύ των μεταβλητών είναι η μέγιστη θετική ενώ εάν είναι -1 η συσχέτιση είναι η μέγιστη αρνητική [19]. Στην στατιστική, ο συντελεστής συσχέτισης Pearson (PCC) αναφέρεται επίσης ως Pearson's r, ο συντελεστής συσχέτισης προϊόντος-ροής Pearson (PPMCC) ή η συσχέτιση διμερούς, είναι ένα μέτρο γραμμικής συσχέτισης μεταξύ δύο συνόλων δεδομένων. Είναι η συν διακύμανση δύο μεταβλητών, διαιρεμένη με το γινόμενο των τυπικών τους αποκλίσεων. Επομένως είναι ουσιαστικά μια κανονικοποιημένη μέτρηση της συν διακύμανσης, έτσι ώστε το αποτέλεσμα να έχει πάντα μια τιμή μεταξύ −1 και 1. Όπως με την ίδια τη συν διακύμανση, το μέτρο μπορεί να αντικατοπτρίζει μόνο μια γραμμική συσχέτιση των μεταβλητών και αγνοεί πολλούς άλλους τύπους σχέσης ή συσχέτισης . Ο συντελεστής συσχέτισης του Pearson, όταν εφαρμόζεται σε ένα δείγμα, αντιπροσωπεύεται συνήθως από το ‘r’ και μπορεί να αναφέρεται ως συντελεστής συσχέτισης δείγματος ή δείκτης συντελεστής συσχέτισης Pearson.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Equation 5. Εξίσωση του Pearson Correlation Coefficient

- *Jaccard Similarity*

Στις άλλες μετρήσεις ομοιότητας, συζητήσαμε μερικούς τρόπους για να βρούμε την ομοιότητα μεταξύ αντικειμένων, όπου τα αντικείμενα είναι σημεία ή διανύσματα. Ο δείκτης Jaccard, επίσης γνωστός ως συντελεστής ομοιότητας Jaccard, είναι μια στατιστική που χρησιμοποιείται για τον υπολογισμό της ομοιότητας και της ποικιλομορφίας των συνόλων δειγμάτων [19]. Ο συντελεστής Jaccard μετρά την ομοιότητα μεταξύ πεπερασμένων συνόλων δειγμάτων και ορίζεται ως το μέγεθος της τομής διαιρούμενο με το μέγεθος της ένωσης των συνόλων δειγμάτων

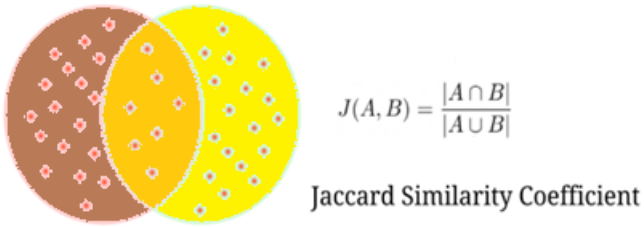

Figure 13. Γραφική Αναπαράσταση του Jaccard Similarity Coefficient

Η απόσταση Jaccard, η οποία μετρά την ομοιότητα μεταξύ των συνόλων δειγμάτων, είναι συμπληρωματική του συντελεστή Jaccard και επιτυγχάνεται αφαιρώντας τον συντελεστή Jaccard από 1, ή, ισοδύναμα, διαιρώντας τη διαφορά των μεγεθών της ένωσης και της τομής δύο συνόλων με το μέγεθος της ένωσης:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Equation 6. Εξίσωση του Jaccard Similarity Coefficient

- *Hamming distance*

Όλες οι ομοιότητες που αναφέραμε ήταν μετρήσεις απόστασης για συνεχείς μεταβλητές. Στην περίπτωση κατηγορηματικών μεταβλητών, το μέτρο ομοιότητας που συχνά χρησιμοποιείται είναι η απόσταση Hamming [20]. Στην περίπτωση που οι μεταβλητές που συγκρίνονται είναι ίδιες, τότε η απόστασή τους είναι 0 ενώ στην περίπτωση που διαφέρουν η απόσταση είναι 1. Η απόσταση Hamming μεταξύ δύο κατηγορηματικών μεταβλητών ίσου μήκους είναι ο αριθμός των θέσεων στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά. Με άλλα λόγια, μετρά τον ελάχιστο αριθμό αντικαταστάσεων που απαιτούνται για να αλλάξει η μία συμβολοσειρά στην άλλη ή τον ελάχιστο αριθμό σφαλμάτων που θα μπορούσαν να έχουν μετατρέψει τη μία συμβολοσειρά στην άλλη.

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

Equation 7. Εξίσωση της Hamming Distance

Οι κατηγορηματικές μεταβλητές που μετράμε μπορεί να είναι δυαδικές, για παράδειγμα, η μέτρηση Hamming δυο μεταβλητών φαίνεται στο παρακάτω σχήμα:

X	Y	Distance
red	red	0
red	green	1

Figure 14. Δυαδικός Πίνακας στο παράδειγμα του Hamming Distance

Τα σύμβολα των κατηγορηματικών μεταβλητών που συγκρίνονται μπορεί να είναι γράμματα, δυαδικά ψηφία ή δεκαδικά ψηφία, μεταξύ άλλων δυνατοτήτων.

Ως ένα παράδειγμα μέτρησης της απόστασης Hamming μεταξύ δυο κατηγορηματικών μεταβλητών πέραν από την διαδική του μορφη για την οποία μιλήσαμε είναι το παρακάτω:

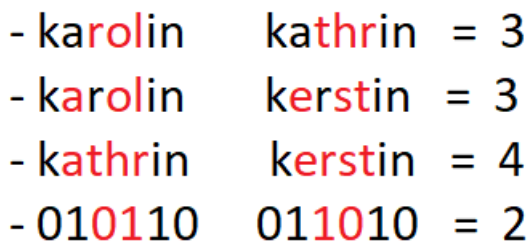

Figure 15. Αποτίμηση Ομοιότητας μέσω του Hamming Distance

2.6 Μετρήσεις Αξιολόγησης των Συστημάτων Προτάσεων

Η ποιότητα των προτάσεων βασίζεται στο πόσο συναφείς και ενδιαφέρουσες είναι για τους χρήστες. Όταν οι προτάσεις είναι πολύ προφανείς, δεν είναι χρήσιμες και το Σύστημα Προτάσεων δεν θεωρείται αποδοτικό. Για τη συνάφεια των προτάσεων, χρησιμοποιούμε μετρήσεις όπως ανάκληση(*recall*) και ακρίβεια(*precision*) [21]. Για την πρωτοτυπία, χρησιμοποιούνται μετρήσεις όπως η διαφορετικότητα (*diversity*), η κάλυψη (*coverage*), η έκπληξη (*serendipity*) και η καινοτομία (*novelty*) [22].

2.6.1 Μετρήσεις για τη συνάφεια

Σε ένα πρόβλημα ταξινόμησης, συνήθως χρησιμοποιούμε τις μετρήσεις αξιολόγησης ακρίβειας και ανάκλησης. Ομοίως, για Συστήματα Προτάσεων, χρησιμοποιούμε ένα μείγμα ακρίβειας και ανάκλησης -την μέτρηση μέσης ακρίβειας (Mean Average Precision-MAP), συγκεκριμένα MAP @ k, όπου παρέχονται k προτάσεις.

Η ακρίβεια *(precision)* υπολογίζεται από το πηλίκο της διαίρεσης των παρεχόμενων προτάσεων που είναι συναφείς για τον χρήστη με όλες τις προτάσεις που παρείχαμε στον χρήστη.

$$P = \frac{\#\text{ of our recommendations that are relevant}}{\#\text{ of items we recommended}}$$

Equation 8. Εξίσωση της μετρικής Precision

Η ανάκληση (*recall*) υπολογίζεται από το πηλίκο της διαίρεσης των παρεχόμενων προτάσεων που είναι συναφείς για τον χρήστη με τον αριθμό όλων των πιθανών προτάσεων που θα ήταν συναφείς για τον χρήστη ακόμα και αν δεν έγιναν.

$$r = \frac{\#\text{ of our recommendations that are relevant}}{\#\text{ of all the possible relevant items}}$$

Equation 9. Εξίσωση της μετρικής recall

Υπάρχει συνήθως μια αντίστροφη σχέση μεταξύ ανάκλησης και ακρίβειας. Η ακρίβεια ανησυχεί για το πόσες στάσεις είναι σχετικές μεταξύ των παρεχόμενων προτάσεων. Η ανάκληση ανησυχεί για το πόσες προτάσεις παρέχονται μεταξύ όλων των σχετικών προτάσεων.

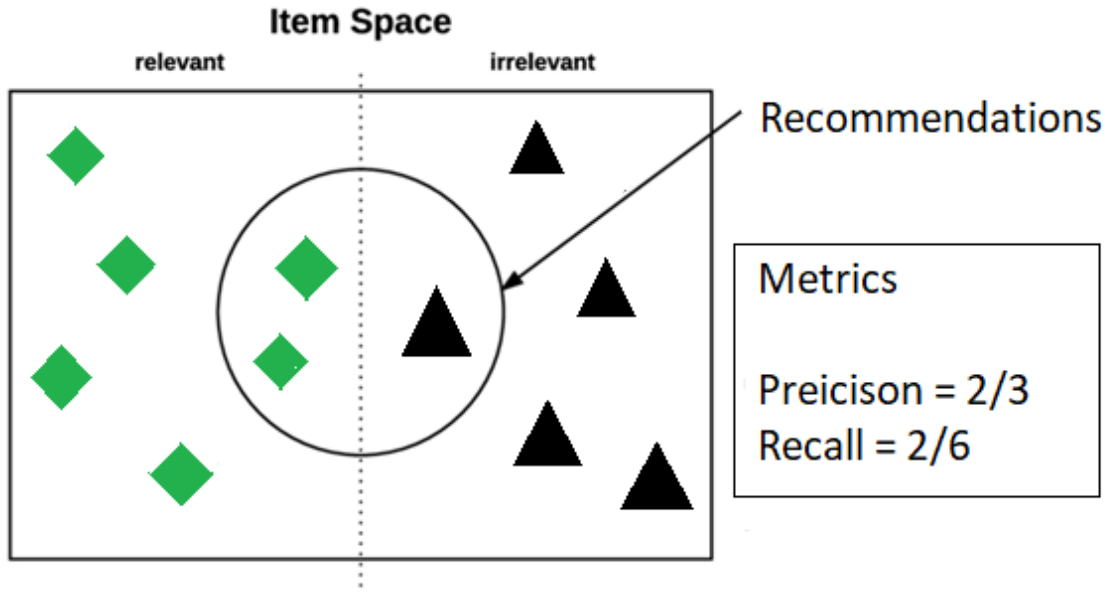

Figure 16. Γραφική Απεικόνιση των μετρικών Precision-Recall

Το μέσο απόλυτο σφάλμα (*Mean Absolute Error- MAE*) είναι μια μέτρηση που χρησιμοποιείται για τον υπολογισμό του μέσου όρου όλων των απόλυτων διαφορών τιμής μεταξύ της πραγματικής και της προβλεπόμενης βαθμολογίας. Όσο χαμηλότερο είναι το *MAE* τόσο καλύτερη είναι η ακρίβεια. Γενικά, το *MAE* μπορεί να κυμαίνεται από 0 έως άπειρο, όπου το άπειρο είναι το μέγιστο σφάλμα ανάλογα με την κλίμακα βαθμολογίας της μετρούμενης εφαρμογής.

$$MAE = \sum_{t=1}^{n} |\hat{y}_t - y|/n$$

Equation 10. Εξίσωση της μετρικής Mean Absolute Error (MAE)

Το *Root Mean Square Error(RMSE)* υπολογίζει τη μέση τιμή όλων των διαφορών των τετραγώνων μεταξύ της πραγματικής και της προβλεπόμενης βαθμολογίας και στη συνέχεια προχωρά στον υπολογισμό της τετραγωνικής ρίζας από το αποτέλεσμα. Κατά συνέπεια, τα μεγάλα σφάλματα ενδέχεται να επηρεάσουν δραματικά την αξιολόγηση *RMSE*, καθιστώντας τη μέτρηση *RMSE* πολύτιμη όταν είναι ανεπιθύμητα σημαντικά μεγάλα σφάλματα.

$$RMSE = \sqrt{\sum_{t=1}^{n}(\hat{y}_t - y)/n}$$

Equation 11. Εξίσωση της μετρικής Root Mean Square Error (RMSE)

2.6.2 Μετρήσεις πέραν της συνάφειας

Ωστόσο, η ποιότητα ενός Συστήματος Προτάσεων δεν βασίζεται μόνο στη συνάφεια των Προτάσεων. Για να δημιουργήσουμε ένα επιτυχημένο Σύστημα Προτάσεων, πρέπει να έχουμε κατά νου ότι η διαφορετικότητα παίζει σημαντικό αντίκτυπο στις προτάσεις. Υπάρχουν περισσότερες μετρήσεις που πρέπει να μετρηθούν και να αναλυθούν προκειμένου να επιτευχθεί μια ισχυρή απόδοση ενός Συστήματος Προτάσεων όπως η διαφορετικότητα (*diversity*), η κάλυψη (*coverage*), η έκπληξη (*serendipity*) και η καινοτομία (*novelty*).

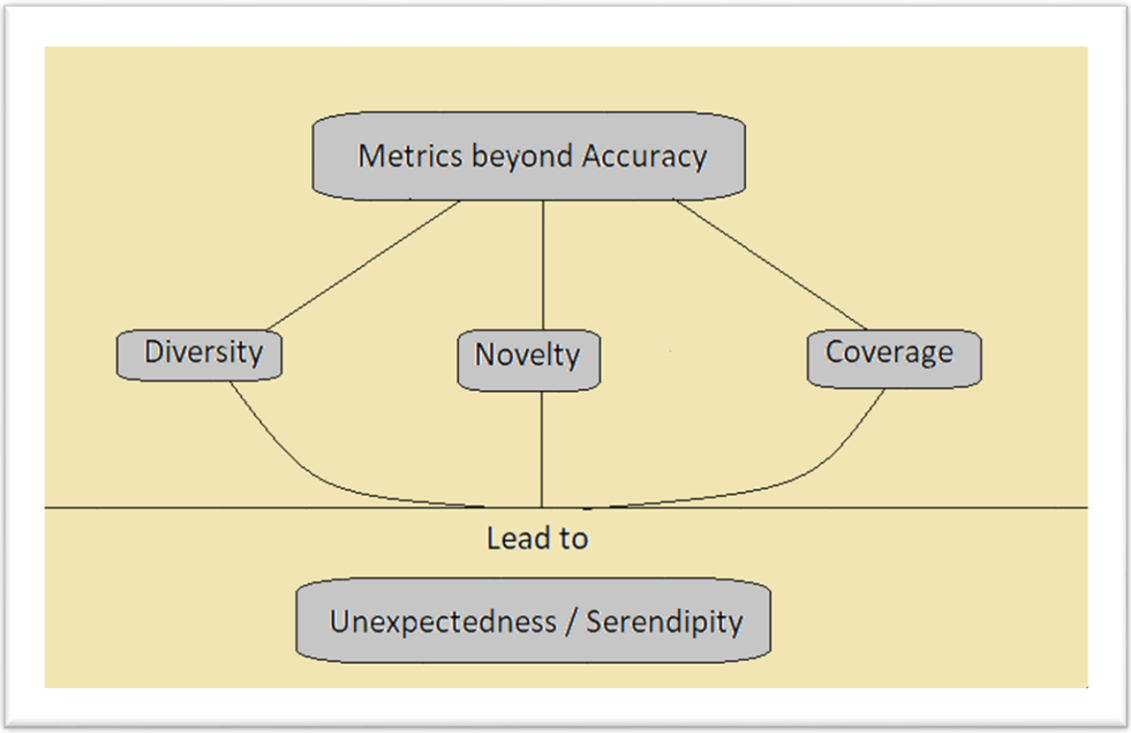

Figure 17. Απεικόνιση των μετρικών πέραν της ακρίβειας/συνάφειας

- *Ποικιλομορφία / Διαφορετικότητα (Diversity)*

Η ποικιλομορφία / διαφορετικότητα (*diversity*) μετρά πόσο διαφορετικά είναι τα προτεινόμενα στοιχεία για έναν χρήστη. Αυτή η ομοιότητα καθορίζεται συχνά χρησιμοποιώντας το περιεχόμενο του αντικειμένου (π.χ. είδη ταινιών), αλλά μπορεί επίσης να προσδιοριστεί χρησιμοποιώντας τον τρόπο με τον οποίο αξιολογούνται παρόμοια στοιχεία. Η περισσότερη έρευνα έχει επικεντρωθεί παραδοσιακά στην ακρίβεια, ιδίως στο πόσο κοντά είναι οι προβλεπόμενες αξιολογήσεις του Συστήματος Προτάσεων με τις πραγματικές αξιολογήσεις των χρηστών. Ωστόσο, έχει αναγνωριστεί ότι άλλες ιδιότητες προτάσεων, όπως το αν ο κατάλογος των προτάσεων είναι διαφορετικός και αν περιέχει νέα στοιχεία, μπορεί να έχει σημαντικό αντίκτυπο στη συνολική ποιότητα ενός συστήματος προτάσεων. Παρ 'όλα αυτά, η ποικιλομορφία γίνεται όλο και πιο σημαντικό θέμα, με την παραδοχή ότι οι χρήστες είναι πιο ικανοποιημένοι με ποικίλες λίστες προτάσεων, ακόμη και αν η ποικιλομορφία κοστίζει κάποια

απώλεια ακρίβειας. Κατά συνέπεια, τα τελευταία χρόνια, το επίκεντρο της έρευνας συστημάτων προτάσεων έχει αλλάξει για να συμπεριλάβει ένα ευρύτερο φάσμα στόχων «πέρα από την ακρίβεια».

- *Έκπληξη (Serendipity)*

Ο ορισμός της βασίζεται σε μεγάλο βαθμό στον ορισμό του βασικού συστατικού της – την έκπληξη. Στη επιστημονική βιβλιογραφία, η έκπληξη έχει συνδεθεί με γεγονότα που διαφέρουν από τις προσδοκίες κάποιου ή είναι δύσκολο να εξηγηθεί. Τέτοιοι ορισμοί είναι σημαντικοί για να λειτουργήσουν στον τομέα της ανάκτησης πληροφοριών ή των Συστημάτων Προτάσεων. Μια συνήθης πρόταση είναι εκείνη που βοηθά τον χρήστη να βρει ένα "εκπληκτικά ενδιαφέρον" αντικείμενο που μπορεί να μην είχε ανακαλύψει διαφορετικά.

- *Καινοτομία (Novelty)*

Η καινοτομία (novelty) μετρά πόσο νέες, πρωτότυπες ή ασυνήθιστες είναι οι προτάσεις για τον χρήστη. Σε γενικές γραμμές, οι προτάσεις θα αποτελούνται κυρίως από δημοφιλή αντικείμενα επειδή
(i) τα δημοφιλή αντικείμενα έχουν περισσότερα δεδομένα και
(ii) τα δημοφιλή αντικείμενα έχουν καλή απόδοση σε αξιολογήσεις εκτός σύνδεσης και στο διαδίκτυο. Ωστόσο, εάν ένα αντικείμενο είναι δημοφιλές ή με κορυφαίες πωλήσεις, ένας χρήστης θα είχε ήδη εκτεθεί σε αυτό. Επομένως, είναι λογικό να τροποποιήσουμε ένα Σύστημα Προτάσεων για να μειώσουμε τον αριθμό των δημοφιλών αντικειμένων που προτείνει. Ο κοινός τρόπος με τον οποίο μετράται η καινοτομία είναι να συγκρίνουμε τα προτεινόμενα στοιχεία ενός χρήστη με αυτά του υπόλοιπου πληθυσμού. Πόσο συχνά εμφανίζονται οι προτάσεις ενός χρήστη στις υπόλοιπες προτάσεις του πληθυσμού;

- *Κάλυψη (Coverage)*

Υπάρχουν δύο γενικές προσεγγίσεις για τη μέτρηση της κάλυψης προτάσεων - «κάλυψη χρηστών», η οποία μετρά το βαθμό στον οποίο το σύστημα καλύπτει τους χρήστες του (π.χ., την αναλογία χρηστών για τους οποίους ένας προτεινόμενος είναι σε θέση να παρέχει πρόταση) και «κάλυψη στοιχείων», που μετράει ο βαθμός στον οποίο οι προτάσεις καλύπτουν το σύνολο των διαθέσιμων αντικειμένων (δηλαδή, ο κατάλογος στοιχείων). Όπως και με άλλους στόχους πέραν της ακρίβειας, η ορολογία που χρησιμοποιείται για τον προσδιορισμό του στόχου κάλυψης ποικίλλει σε διαφορετικά έργα. Ο πιο διαδεδομένος ορισμός του είναι το κλάσμα των στοιχείων που εμφανίζονται στις λίστες προτάσεων των χρηστών.

2.7 Deep Learning στα Συστήματα Προτάσεων

Παρόλο που τα Συστήματα Προτάσεων παρέχουν αποτελεσματικούς τρόπους αντιμετώπισης του προβλήματος υπερφόρτωσης πληροφοριών, αντιμετωπίζουν επίσης πολλές διαφορετικές προκλήσεις, όπως η ακρίβεια των προτάσεων, η αραιότητα των δεδομένων (data sparsity) , το πρόβλημα εκκίνησης (cold start problem) και η κλιμάκωση (scalability). Οι τεχνικές βαθιάς μάθησης χρησιμοποιούνται όλο και περισσότερο τα τελευταία χρόνια στον τομέα προτάσεων ξεκινώντας την περίοδο 2013-2014. Δεδομένου ότι η ικανότητα επεξεργασίας δεδομένων των τεχνικών βαθιάς μάθησης αυξάνεται λόγω των εξελίξεων σε μεγάλες εγκαταστάσεις δεδομένων και ύπερ-υπολογιστών, οι ερευνητές έχουν ήδη αρχίσει να επωφελούνται από τεχνικές βαθιάς μάθησης σε Συστήματα Προτάσεων. Έχουν χρησιμοποιήσει τεχνικές βαθιάς μάθησης για να παράγουν πρακτικές λύσεις στις προκλήσεις των Συστημάτων Προτάσεων, όπως η επεκτασιμότητα και η αραιότητα των δεδομένων. Επιπλέον, έχουν χρησιμοποιήσει τη βαθιά μάθηση για την παραγωγή προτάσεων, τη μείωση διαστάσεων, την εξαγωγή χαρακτηριστικών από διαφορετικές πηγές δεδομένων και την ενσωμάτωσή τους στα συστήματα προτάσεων. Οι τεχνικές βαθιάς μάθησης χρησιμοποιούνται σε συστήματα

σύστασης για να μοντελοποιήσουν είτε τον πίνακα αλληλεπιδράσεων μεταξύ χρηστών-αντικειμένων είτε παράπλευρες πληροφορίες περιεχομένου που ενισχύουν τις προτάσεις.

2.7.2 Συνεισφορές του Deep Learning στα Συστήματα Προτάσεων

Μελετώντας και ψάχνοντας ερευνητικές δημοσιεύσεις όπως και συνέδρια (Deep Learning for Recommender Systems | Alexandros Karatzoglou) [23], θα μπορούσαμε να κατηγοριοποιήσουμε την συνεισφορά της βαθιάς μάθησης στα Συστήματα Προτάσεων ως εξής:

1) Εκμάθηση διανυσμάτων ενσωμάτωσης *(Learning the embeddings)*

   Για παράδειγμα, το Word2Vec είναι μια Neural Network τεχνική που μαθαίνει ενσωματώσεις για λέξεις με συγκεκριμένες ιδιότητες. Στην πράξη, μαθαίνει ένα διάνυσμα αναπαράστασης σε έναν χώρο ενσωμάτωσης που είναι κοντά σε άλλα διανύσματα που αντιπροσωπεύουν παρόμοιες λέξεις με το πρωτότυπο. Αυτός ο αλγόριθμος λειτουργεί με το "παράθυρο περιβάλλοντος" (“context window”) που αποτελείται από μια σειρά λέξεων, ας πούμε 5, και προσπαθεί να προβλέψει την 6η λέξη. Είναι ένα πρόβλημα ταξινόμησης που προσπαθεί να προβλέψει ποια λέξη θα είναι η επόμενη. Έτσι, εάν έχουμε μεγάλο όγκο κείμενο, μπορούμε να χρησιμοποιήσουμε αυτήν την Neural Network τεχνική για να μάθουμε ένα μοντέλο που μπορεί να προβλέψει ποια λέξη θα ακολουθήσει μετά από μια σειρά συγκεκριμένων λέξεων. Αυτές οι ενσωματώσεις που μαθαίνονται μπορούν να χρησιμοποιηθούν στο χώρο ενσωμάτωσης (Embedding Space) και έχουν χρήσιμες γεωμετρικές ιδιότητες, πράγμα που σημαίνει ότι παρόμοιες λέξεις είναι πολύ κοντά η μια στην άλλη.
   Έτσι, αυτός ο αλγόριθμος μπορεί πραγματικά να χρησιμοποιηθεί απευθείας στα Συστήματα Προτάσεων χρησιμοποιώντας αντικείμενα αντί για ενσωματώσεις λέξεων. Χρησιμοποιώντας ένα προφίλ χρήστη και τα στοιχεία που αγόρασε ο χρήστης, μπορούμε να μάθουμε το μοντέλο για να προβλέψουμε το επόμενο στοιχείο που θα

αγόραζε ο χρήστης. Ή θα μπορούσε ακόμη και να χρησιμοποιηθεί σε άλλους αλγόριθμους ως είσοδος, αν χρειαζόμαστε μια αναπαράσταση για τα αντικείμενα.

2) Βαθύ συνεργατικό Φιλτράρισμα (Deep Collaborative Filtering)

Οι Auto Encoders χρησιμοποιούνται για την παροχή προτάσεων βάση συνεργατικού φιλτραρίσματος (collaborative filtering). Οι Autoencoders είναι Deep Neural Networks που έχουν τον ίδιο χώρο εισόδου και εξόδου και προσπαθούν να ανακατασκευάσουν την είσοδο στην έξοδο. Για παράδειγμα, ένα Auto Encoder μοντέλο θα μπορούσε να χρησιμοποιηθεί για έναν χρήστη που έχει παρακολουθήσει 20 ταινίες. Εάν αφαιρέσουμε 1 ταινία από τις 20 και τις τροφοδοτήσουμε στον Auto Encoder, θα μπορούσαμε να προσπαθήσουμε να προβλέψουμε ολόκληρο το σύνολο των ταινιών. Με αυτόν τον τρόπο, το Auto Encoder μαθαίνει τις σχέσεις μεταξύ των ταινιών που έχει παρακολουθήσει ο χρήστης και επειδή πολλά άτομα έχουν τα ίδια ενδιαφέροντα, πολλά μοτίβα θα επαναληφθούν, ώστε το Auto Encoder να μάθει αυτά τα μοτίβα.
Ένας άλλος τρόπος με τον οποίο χρησιμοποιείται το Deep Collaborative Filtering είναι η λήψη χαρακτηριστικών χρήστη και στοιχείων και η ενσωμάτωσή τους σε χώρο χαμηλότερης διάστασης μέσω ξεχωριστών δικτύων deep feed forward και υπολογισμός της ομοιότητας μεταξύ τους.

3) Εξαγωγή χαρακτηριστικών από το περιεχόμενο (Feature extraction from content)

Για παράδειγμα, όταν ένας χρήστης αγοράζει ορισμένα είδη στο Amazon, δεν διαβάζει μόνο την περιγραφή ενός αντικειμένου, αλλά επίσης κοιτάζει και την εικόνα του. Είναι λογικό να υποθέσουμε ότι οι εικόνες περιέχουν επίσης σχετικές πληροφορίες σχετικά με τον χρήστη. Επομένως, θα ήταν χρήσιμο να εξάγονται χαρακτηριστικά από τις εικόνες, προκειμένου να παρέχονται καλύτερες προτάσεις. Αυτό θα μπορούσε εύκολα να γίνει από ένα προ-εκπαιδευμένο Συνελικτικό Νευρωνικό Δίκτυο (*CNNs*) που αποδεδειγμένα είναι πολύ αποτελεσματικό στην εξαγωγή πληροφοριών από εικόνες. Αυτές οι δυνατότητες μπορούν στη συνέχεια να

περιληφθούν σε ένα πρόβλημα Matrix Factorization ή σε οποιονδήποτε αλγόριθμο προτάσεων βαθιάς μάθησης.
Ένα άλλο παράδειγμα στην εξαγωγή χαρακτηριστικών από περιεχόμενο θα μπορούσε να είναι σε σύσταση ειδήσεων ή σε σύσταση ερευνητικών δημοσιεύσεων όπου ένα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (*RNNs*) μπορούν να χρησιμοποιηθούν για την εξαγωγή χαρακτηριστικών από το κείμενο. Μπορούμε να χαρτογραφήσουμε αυτό το κείμενο σε χώρο χαμηλότερης διάστασης και να χρησιμοποιήσουμε αυτές τις πληροφορίες για την παροχή προτάσεων σε χρήστες.

4) Προτάσεις βάσει Συνεδρίας (Session-based Recommendations)

Η συνεδρία χρήστη είναι στην πραγματικότητα μια σειρά συμβάντων που πραγματοποιούνται όταν ένας χρήστης πλοηγείται σε μια ιστοσελίδα. Για παράδειγμα, στο Amazon όταν ένας χρήστης αγοράζει κάτι, όλα τα κλικ και ο χρόνος που αφιερώνεται σε κάθε στοιχείο της σελίδας θεωρείται ως ολόκληρη περίοδος σύνδεσης. Έτσι, για να μοντελοποιήσουμε αυτήν την ακολουθία συμβάντων, τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs) έχουν εισέλθει στον τομέα των προτάσεων με βάση την περίοδο σύνδεσης. Ο λόγος για τον οποίο τα RNNs είναι πραγματικά ισχυρά στην αντιμετώπιση διαδοχικών δεδομένων είναι ότι έχουν μια εσωτερική κρυφή κατάσταση "s" που θα μπορούσε να θεωρηθεί ως σύνοψη των τρεχόντων και των προηγούμενων δεδομένων και η έξοδος αυτού του δικτύου είναι συνάρτηση αυτής της ίδιας κατάστασης.

2.8 Συστήματα Προτάσεων σε Κοινωνικά Δίκτυα

Έχοντας αναλύσει τα παραδοσιακά Συστήματα Προτάσεων με τις παραλλαγές που συναντώνται, τους τρόπους που μετράται η ομοιότητα μεταξύ χρηστών ή αντικειμένων, τα κριτήρια που πρέπει να λαμβάνουμε υπόψιν προκειμένου να τα αξιολογήσουμε, τις προκλήσεις που καλούνται να αντιμετωπίσουν, καθώς και τεχνικές που σχετίζονται με

διάφορες μεθόδους βαθιάς μάθησης θα παρουσιάσουμε την έννοια των κοινωνικών δικτύων και τα Συστήματα Προτάσεων σε αυτά.

Ως κοινωνικό δίκτυο εννοούμε ένα σύνολο χρηστών οι οποίοι συσχετίζονται μεταξύ τους και μοιράζονται κοινά ενδιαφέροντα. Οι συσχετίσεις μεταξύ των χρηστών μπορεί να απεικονίζεται ως η φιλία μεταξύ τους ή ως η εμπιστοσύνη και ο συμμερισμός των προτιμήσεών τους [24]. Τα διαδικτυακά κοινωνικά δίκτυα προσφέρουν νέες ευκαιρίες για περαιτέρω βελτίωση της ακρίβειας των Συστημάτων Προτάσεων καθώς πέραν των πληροφοριών που εξάγουμε από τις αξιολογήσεις και συμπεριφορές που φέρουν οι χρήστες για αντικείμενα, μπορούμε να εκμεταλλευτούμε τις συσχετίσεις με άλλους χρήστες. Ένα κοινωνικό δίκτυο μπορεί να θεωρηθεί ως ένα γράφημα που απεικονίζει τους χρήστες ως οντότητες και τις συσχετίσεις τους ως ακμές.

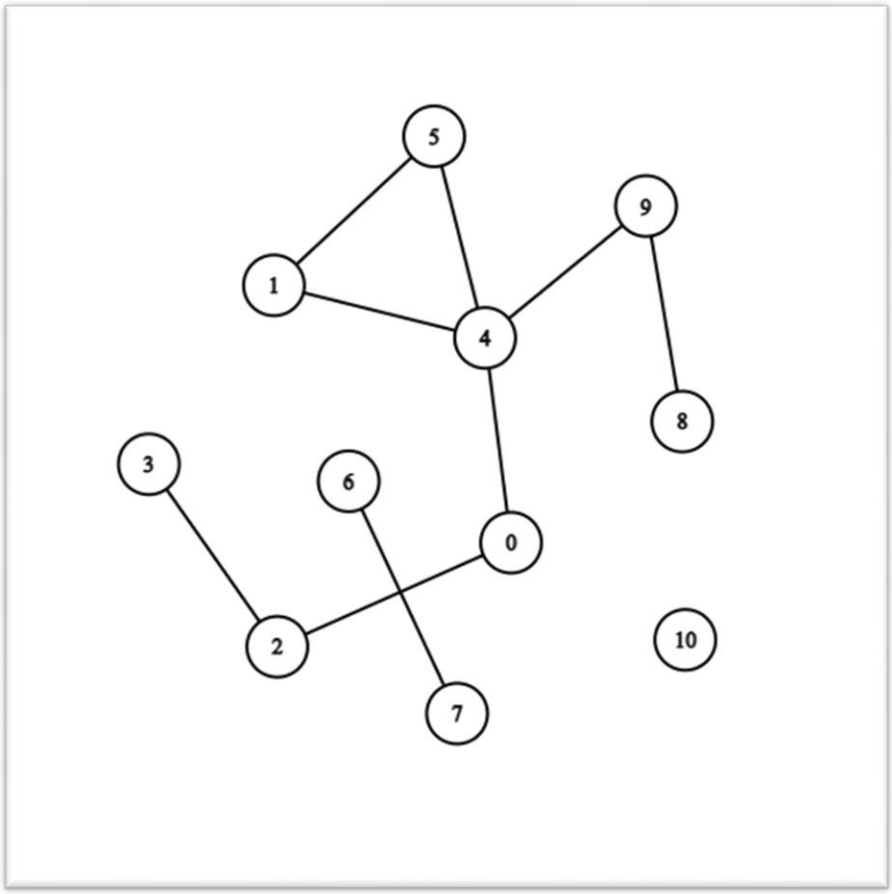

Figure 18. Γραφική Αναπαράσταση ενός κοινωνικού δικτύου

Στην πραγματική ζωή, οι άνθρωποι καταφεύγουν συχνά σε φίλους στα κοινωνικά τους δίκτυα για συμβουλές πριν αγοράσουν ένα προϊόν ή καταναλώσουν μια υπηρεσία [25]. Τα ευρήματα στους τομείς της κοινωνιολογίας και της ψυχολογίας δείχνουν ότι οι άνθρωποι τείνουν να συσχετίζονται και να συνδέονται με παρόμοιους άλλους χρήστες. Λόγω των σταθερών και μακροχρόνιων κοινωνικών δεσμών, οι άνθρωποι είναι πιο πρόθυμοι να μοιραστούν τις προσωπικές τους απόψεις με τους φίλους τους, και συνήθως εμπιστεύονται τις προτάσεις των φίλων τους περισσότερο από αυτές των ξένων και των πωλητών. Τα δημοφιλή διαδικτυακά κοινωνικά δίκτυα, όπως το Facebook το Twitter και το YouTube, παρέχουν νέους τρόπους στους ανθρώπους να επικοινωνούν και να δημιουργούν εικονικές κοινότητες. Κοινωνικά δίκτυα θεωρούνται επίσης ιστότοποι όπου χρήστες γράφουν αναθεωρήσεις και αξιολογήσεις για αντικείμενα τα οποία κατανάλωσαν και έχουν τη δυνατότητα να δηλώσουν βαθμό εμπιστοσύνης προς άλλους χρήστες βάση των αξιολογήσεων που έχουν κάνει. Ένα παράδειγμα τέτοιου ιστότοπου είναι το Epinions και το Ciao. Τα διαδικτυακά κοινωνικά δίκτυα όχι μόνο διευκολύνουν τους χρήστες να μοιράζονται τις απόψεις τους μεταξύ τους, αλλά και να χρησιμεύουν ως πλατφόρμα για την ανάπτυξη νέων αλγορίθμων για τα Συστήματα Προτάσεων για την αυτοματοποίηση των κατά τα άλλα χειροκίνητων κοινωνικών προτάσεων σε πραγματικά κοινωνικά δίκτυα.

Ένα Σύστημα Προτάσεων σε κοινωνικό δίκτυο βελτιώνει την ακρίβεια του παραδοσιακού Συστήματος Προτάσεων λαμβάνοντας υπόψη κοινωνικά ενδιαφέροντα και κοινωνική εμπιστοσύνη μεταξύ των χρηστών ως πρόσθετες εισόδους [26]. Για παράδειγμα, λόγω κοινωνικού ενδιαφέροντος, ένας χρήστης μπορεί να διαβάσει ένα συγκεκριμένο άρθρο ειδήσεων σχετικά με ένα συμβάν μόνο και μόνο επειδή το συμβάν συνέβη σε ένα μέρος όπου ζει η οικογένειά του. Λόγω της κοινωνικής εμπιστοσύνης, ένας χρήστης μπορεί να ακούσει ένα τραγούδι που προτείνουν οι στενοί φίλοι του στο Facebook. Η κοινωνική εμπιστοσύνη μεταξύ ενός ζεύγους φίλων μπορεί να δημιουργηθεί με βάση τη ρητή ανατροφοδότηση του χρήστη Α σχετικά με τον χρήστη Β (π.χ., με ψηφοφορία /*explicit feedback*), ή μπορεί να συναχθεί από τα σιωπηρά σχόλια (π.χ., τη συχνότητα και την ποσότητα αλληλεπίδρασης / επικοινωνίας / ανταλλαγής email μεταξύ Α και Β /*implicit feedback*). Διαφορετικοί

αλγόριθμοι Συστημάτων Προτάσεων που λειτουργούν σε κοινωνικά δίκτυα διερευνούν διαφορετικά τα κοινωνικά δίκτυα και τις ενσωματωμένες κοινωνικές πληροφορίες.

3 Ομοιότητα μέσα στον Γράφο

3.1 Γενικά για την ομοιότητα σε Γράφους

Το θέμα του προσδιορισμού της ομοιότητας των γραφημάτων θεωρήθηκε ως ιδιαίτερα σημαντικό ερευνητικό πεδίο στον παγκόσμιο ιστό, την τεχνητή νοημοσύνη και την έρευνα πληροφοριών. Πολλές εφαρμογές απαιτούν ένα μέτρο "ομοιότητας" μεταξύ αντικειμένων. Ένα προφανές παράδειγμα είναι η εύρεση παρόμοιου εγγράφου κατάλληλου εγγράφου δοσμένης μίας εντολής “query”, ή στον Παγκόσμιο Ιστό. Γενικότερα, ένα μέτρο ομοιότητας μπορεί να χρησιμοποιηθεί για ομαδοποίηση αντικειμένων, όπως για συλλογικό φιλτράρισμα σε ένα σύστημα σύστασης (collaborative filtering), στο οποίο οι "παρόμοιοι" χρήστες και αντικείμενα ομαδοποιούνται με βάση τις προτιμήσεις των χρηστών.

Στο πεδίο των Συστημάτων Προτάσεων σε ένα κοινωνικό σύνολο, όπου χρήστες και αντικείμενα μπορούν να αναπαρασταθούν σε έναν Γράφο ως οντότητες και ακμές, υπάρχουν διάφοροι τρόποι ώστε να αποφανθούμε για το ποσοστό που ένας χρήστης είναι όμοιος με άλλους χρήστες. Οι οντότητες στο Γράφο αναφέρονται στους χρήστες, ενώ οι ακμές στις μεταξύ τους σχέσεις. Οι ακμές αυτές μπορεί να είναι αριθμημένες ή όχι. Στην περίπτωση που είναι αριθμημένες, ο αριθμός επάνω στην ακμή αντιπροσωπεύει τη σημαντικότητα της σχέσης μεταξύ των συνδεδεμένων χρηστών, ενώ στην περίπτωση που οι ακμές δεν είναι αριθμημένες οι σχέσεις μεταξύ των χρηστών είναι απόλυτες και μη μετρήσιμες. Συνήθως σε

τέτοιου είδους γραφήματα οι ακμές δεν είναι αριθμημένες καθώς οι χρήστες επιλέγουν είτε να είναι συνδεδεμένοι με άλλους χρήστες είτε όχι.

Τα κοινωνικά δίκτυα έχουν αναπτυχθεί πολύ γρήγορα τα τελευταία χρόνια και είναι πολύ σημαντικά σε πολλούς τομείς εφαρμογών πληροφορικής. Μια σημαντική πτυχή αυτών των εφαρμογών βασίζεται σε μέτρα ομοιότητας μεταξύ κόμβων στο δίκτυο. Κάθε μέτρο ομοιότητας έχει τη δική του δυνατότητα εφαρμογής και ταιριάζει καλύτερα σε συγκεκριμένο τύπο τιμών και εάν αυτά τα μέτρα συνδυάζονται συλλογικά, μπορούν να βοηθήσουν στην κατασκευή ενός καλύτερου προφίλ για τους χρήστες. Όπως αναφέραμε, ένα κοινωνικό δίκτυο μπορεί να αναπαρασταθεί ως γράφημα, δηλαδή μια συλλογή συνδεδεμένων κόμβων. Η ομοιότητα μεταξύ κόμβων θα μπορούσε να βασίζεται σε χαρακτηριστικά κόμβων (κείμενο) ή / και άκρες / συνδέσμους (δομή). Ορισμένα μέτρα ομοιότητας λαμβάνουν υπόψιν τους γειτονικούς κόμβους και τη δομή του γραφήματος, ενώ άλλα επιτρέπουν στους κόμβους να είναι παρόμοιοι ακόμη και όταν δεν έχουν κοινούς γείτονες ή διασυνδέσεις.

3.2 Τύποι Ομοιότητας μέσα στον Γράφο

Σύμφωνα με τις πληροφορίες που χρησιμοποιούνται για τον προσδιορισμό της ομοιότητας των οντοτήτων μέσα στο Γράφο, μπορούμε να κατηγοριοποιήσουμε τα μέτρα ομοιότητας σε 3 ευρύτερες κατηγορίες [27] ως εξής:

- Δομική ομοιότητα (link – based). Σε αυτόν τον τύπο ομοιότητας, εξετάζονται οι σύνδεσμοι μεταξύ των κόμβων στο γράφημα, δηλαδή οι ακμές του. Οι σύνδεσμοι μπορούν να αντιπροσωπεύουν: σχέση εμπιστοσύνης, φιλία, πληρωμές κ.λπ. Έχει αποδειχθεί ότι αναλογικά με την ανθρώπινη κρίση είναι καλύτερη από άλλες μεθόδους ομοιότητας όπως η ομοιότητα περιεχομένου που θα δούμε παρακάτω.

- Ομοιότητα περιεχομένου (text - based). Σε αυτόν τον τύπο ομοιότητας, εξετάζονται τα χαρακτηριστικά των κόμβων στο γράφημα. Η ομοιότητα περιεχομένου ενός ιστότοπου φιλίας θα μπορούσε ενδεχομένως να βασίζεται στην ημερομηνία γέννησης, στα χόμπι, στο ενδιαφέρον των ταινιών και στην ηλικία των χρηστών. Ένας τρόπος για την καταγραφή περιεχομένου είναι με τη χρήση ετικετών καθορισμένων από τον χρήστη (π.χ. οι ετικέτες μπορούν να αντιπροσωπεύουν το περιεχόμενο μιας ταινίας που ενδιαφέρουν τον χρήστη). Με βάση την ομοιότητα των ετικετών, μπορεί να αναπτυχθεί ένας αλγόριθμος προτάσεων.

- Ομοιότητα λέξεων-κλειδιών (word - based). Όπως και στην ομοιότητα ετικετών παραπάνω, η ομοιότητα κόμβων μπορεί να οριστεί βάσει συλλογών λέξεων ή λέξεων-κλειδιών που αντιπροσωπεύουν τους χρήστες. Ένα παράδειγμα ομοιότητας λέξεων-κλειδιών είναι το δασικό μοντέλο (forest model), όπου οι λέξεις-κλειδιά διατάσσονται σε μια ιεραρχική δομή για να σχηματίσουν δέντρα διαφορετικών υψών.

3.3 Ο αλγόριθμος SimRank

Στη θεωρία δικτύου, η πρόβλεψη συνδέσμου είναι το πρόβλημα της πρόβλεψης της ύπαρξης σύνδεσης μεταξύ δύο οντοτήτων σε ένα δίκτυο. Παραδείγματα προβλέψεων συνδέσμων περιλαμβάνουν την πρόβλεψη συνδέσμων φιλίας μεταξύ χρηστών σε ένα κοινωνικό δίκτυο, την πρόβλεψη συνδέσμων συν-συγγραφέα σε ένα δίκτυο παραπομπών και την πρόβλεψη αλληλεπιδράσεων μεταξύ γονιδίων και πρωτεϊνών σε ένα βιολογικό δίκτυο. Η πρόβλεψη συνδέσμου μπορεί επίσης να έχει μια χρονική πτυχή, όπου, δεδομένου ενός στιγμιότυπου του συνόλου συνδέσμων τη στιγμή t, ο στόχος είναι να προβλεφθούν οι σύνδεσμοι τη στιγμή t + 1. Η πρόβλεψη συνδέσμου ισχύει ευρέως. Στο ηλεκτρονικό εμπόριο, η πρόβλεψη συνδέσμου είναι συχνά ένα δευτερεύον έργο για τη σύσταση στοιχείων στους χρήστες. Στην επιμέλεια των βάσεων δεδομένων παραπομπών, μπορεί να χρησιμοποιηθεί για την αντιγραφή δίσκων [28].

Ο αλγόριθμος SimRank [29] είναι ένα γενικό μέτρο ομοιότητας, βασισμένο σε ένα απλό και διαισθητικό μοντέλο γραφήματος. Ο SimRank εφαρμόζεται σε οποιονδήποτε τομέα με σχέσεις αντικειμένου-προς-αντικείμενο, που μετρά την ομοιότητα του δομικού πλαισίου στο οποίο συμβαίνουν αντικείμενα, με βάση τις σχέσεις τους με άλλα αντικείμενα. Συνεπώς πρόκειται για ένα μέτρο ομοιότητας σε Γράφους που βασίζεται στη δομή και αντιστοιχίζεται στην πρώτη κατηγορία που αναφέραμε στην προηγούμενη υπό-ενότητα. Η λογική πίσω από τον αλγόριθμο SimRank είναι ότι, σε πολλούς τομείς, παρόμοια αντικείμενα αναφέρονται από παρόμοια αντικείμενα. Πιο συγκεκριμένα, τα αντικείμενα ‘a’ και ‘b’ θεωρούνται όμοια εάν είναι στραμμένα από αντικείμενα ‘c’ και ‘d’ αντίστοιχα, και τα ‘c’ και ‘d’ είναι μεταξύ τους όμοια. Στην βασική του μορφή, ο SimRank αναθέτει τη μέγιστη τιμή ομοιότητας των αντικειμένων προς τους εαυτούς τους.

Για έναν κόμβο v σε έναν κατευθυνόμενο Γράφο, δηλώνουμε με $I\ (v)$και $O(v)$το σύνολο των γειτόνων που δείχνουν στον χρήστη και το σύνολο των γειτόνων στους οποίους δείχνει ο χρήστης αντίστοιχα. Οι χρήστες που ανήκουν στα δύο αυτά σύνολα αναπαρίστανται ως $I_i\ (v)$ και $O_i(v)$ αντίστοιχα για $1 \leq \mathrm{i} \leq |\ I\ (v)|$ και $1 \leq \mathrm{i} \leq |\ O(v)|$ αντίστοιχα. Ας θέσουμε την ομοιότητα μεταξύ των αντικειμένων ‘a’ και ‘b’ ως $s(a,b) \in [0,1]$ έχουμε:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

Όπου το C είναι μια σταθερά μεταξύ 0 και 1. Μία ειδική περίπτωση συναντάτι όταν ο κόμβος ‘α’ ή ‘β’ δεν έχει γείτονες. Σε αυτήν την περίπτωση η ομοιότητα μεταξύ των χρηστών αυτών δεν μπορεί να υπολογιστεί και τίθεται ως 0.

Τα βήματα που ακολουθεί ο αλγόριθμος SimRank είναι τα εξής:

- Αρχικοποίησε τα σκορ ομοιότητας για κάθε ζευγάρι κόμβων ως εξής:

  Αν κόμβος1 = κόμβος2

  SimRank(κόμβος1,κόμβος2) = 1

  Αλλιώς:

  SimRank(κόμβος1,κόμβος2) = 0
- Για κάθε επανάληψη (iteration) ενημέρωσε το σκορ ομοιότητας για κάθε ζευγάρι κόμβων στο Γράφο.
- Αν οι 2 κόμβοι είναι ίδιοι τότε: SimRank(κόμβος1,κόμβος2) = 1
- Εάν κάποιος κόμβος δεν έχει κανένα εισερχόμενο γείτονα τότε: SimRank(κόμβος1,κόμβος2) = 0
- Αλλιώς υπολόγισε το νέο SimRank σκορ με τη βοήθεια της εξίσωσης:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

Προκειμένου να καταλάβουμε καλύτερα τη λειτουργία του SimRank, ας δούμε τα αποτελέσματα που θα παραχθούν σε διαφορετικούς Γράφους μετά την υλοποίηση του αλγορίθμου σε αυτούς.
Πρέπει να σημειωθεί πως η σταθερά C είχε οριστεί στην τιμή 0.9 κατά τα πειράματα σε όλους του Γράφους που ακολουθούν.

Γράφος 1

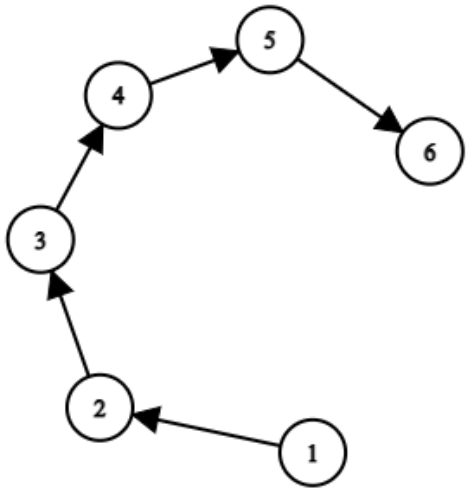

Figure 19. Γράφος με 6 οντότητές και 5 διασυνδέσεις

Ο πίνακας των αποτελεσμάτων διαμορφώνεται ως εξής:

1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1

Table 1. Αποτελέσματα του SimRank στο Γράφο 1

Τα αποτελέσματα ακολουθούν τη σειρά των κόμβων στο γράφημα, δηλαδή 1 – 2 – 3 – 4 – 5 – 6 τόσο στις στήλες όσο και τις γραμμές του πίνακα.

Η κύρια διαγώνιος του πίνακα είναι παντού 1 κάτι το οποίο είναι φυσιολογικό και ακολουθεί τον κανόνα που αναφέραμε, ότι το σκορ ομοιότητας ανάμεσα σε έναν κόμβο και τον εαυτό του ισούται με 1.
Επίσης μπορούμε να διακρίνουμε εύκολα από το γράφημα πως κανένα ζευγάρι κόμβων δεν έχει κοινούς γείτονες, κάτι το οποίο μας οδηγεί σε μηδενικό σκορ ομοιότητας.

Γράφος 2

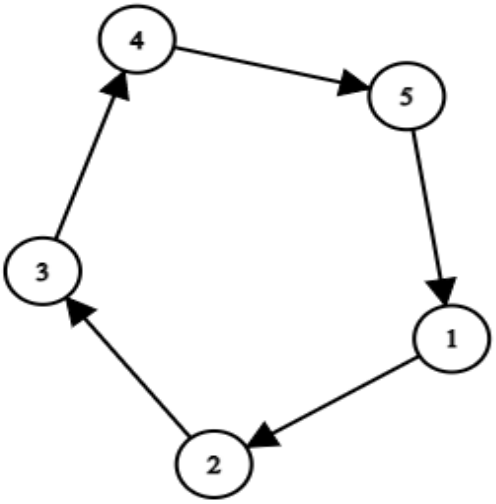

Figure 20. Γράφος με 5 οντότητες και 5 διασυνδέσεις

Ο πίνακας των αποτελεσμάτων διαμορφώνεται ως εξής:

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Table 2. Αποτελέσματα του SimRank στο Γράφο 2

Ομοίως, στο παραπάνω Γράφο δεν υπάρχουν κόμβοι που να έχουν κοινούς γείτονες. Οπότε το σκορ ομοιότητας έχει ακριβώς την ίδια συμπεριφορά με το Γράφο 1.

Γράφος 3

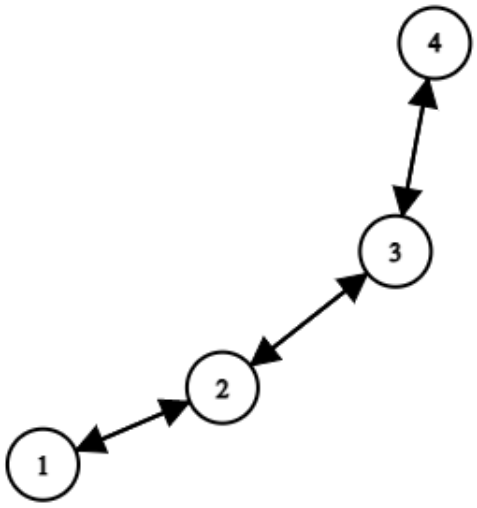

Figure 21. Γράφος με 4 οντότητες και 3 αμφίπλευρες διασυνδέσεις

Ο πίνακας των αποτελεσμάτων διαμορφώνεται ως εξής:

1	0	0.818	0
0	1	0	0.818
0.818	0	1	0
0	0.818	0	1

Table 3. Αποτελέσματα του SimRank στο Γράφο 3

Από τα αποτελέσματα, βλέπουμε πως οι κόμβοι (1-3) και (2-4) έχουν έναν κοινό γείτονα και αυτό οδηγεί σε σκορ ομοιότητας προφανώς μη μηδενικό.

Μία άλλη παρατήρηση που μπορούμε να κάνουμε είναι πως ο πίνακας αποτελεσμάτων παρουσιάζει συμμετρία ως προς την κύρια διαγώνιο.
Δηλαδή SimRank(κόμβος1,κόμβος3) = SimRank(κόμβος3,κόμβος1)

Γράφος 4

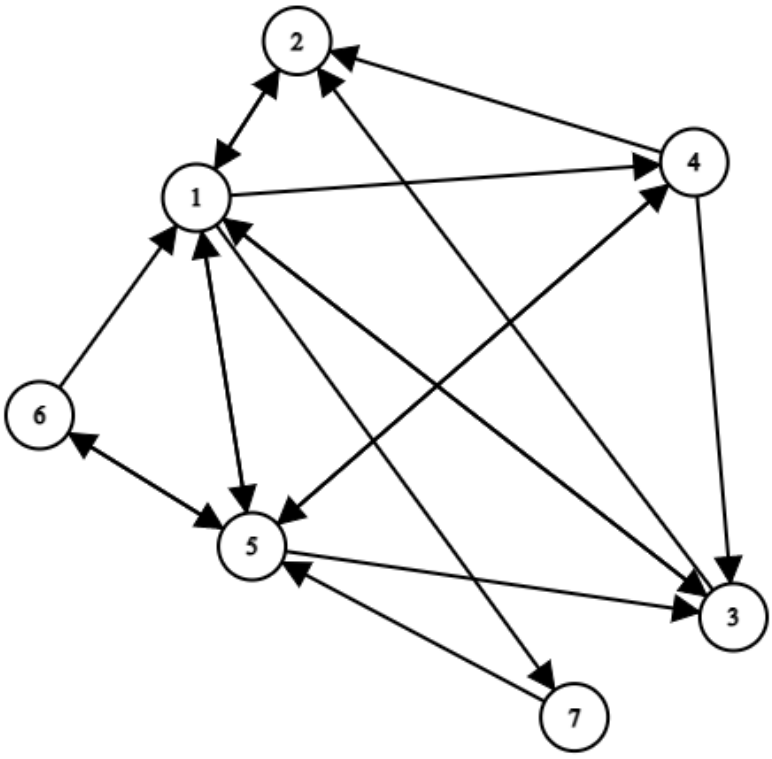

Figure 22. Γράφος με 7 οντότητες και πολλαπλές διασυνδέσεις

Ο πίνακας των αποτελεσμάτων διαμορφώνεται ως εξής:

1	0.563	0.553	0.556	0.544	0.603	0.509
0.563	1	0.597	0.568	0.604	0.502	0.633
0.553	0.597	1	0.628	0.585	0.627	0.63
0.556	0.568	0.628	1	0.545	0.695	0.695
0.544	0.604	0.585	0.545	1	0.491	0.6
0.603	0.502	0.627	0.695	0.491	1	0.49
0.509	0.633	0.63	0.695	0.6	0.49	1

Table 4. Αποτελέσματα του SimRank στο Γράφο 4

Αν παρατηρήσουμε τα ζευγάρια κόμβων (6-4) και (7-4) βλέπουμε πως ο κοινός γείτονας στους κόμβους 6 και 7 δεν είναι άλλος από τον κόμβο 4.

- Ο κόμβος 4 έχει δύο εισερχόμενους γείτονες, τον 1 και τον 5.
- Ο κόμβος 5 είναι ο μοναδικός εισερχόμενος γείτονας για τον κόμβο 6.

Αυτή η σχέση μεταξύ τους μας δίνει το ίδιο σκορ ομοιότητας 0.695.

4.0 Πείραμα

4.1 Πείραμα και Σύνολα δεδομένων

Τα Συστήματα Προτάσεων σε κοινωνικά δίκτυα όπως αναφέραμε μπορούν να κάνουν προβλέψεις / προτάσεις χρησιμοποιώντας περισσότερες πληροφορίες από τα παραδοσιακά

που λαμβάνουν υπόψιν μόνο αξιολογήσεις αντικειμένων ή τα χαρακτηριστικά αυτών και των χρηστών.

Αυτές οι πληροφορίες είναι ένας τύπος κοινωνικής σχέσης μεταξύ ενός συνόλου χρηστών. Η σχέση αυτή μπορεί μερικές φορές να είναι αν ένας χρήστης είναι «φίλος» με άλλους χρήστες ή ίσως εάν ένας χρήστης έχει επιλέξει ορισμένους χρήστες ως άτομα που εμπιστεύεται. Αυτό το είδος πληροφοριών είναι πολύ χρήσιμο για ένα Σύστημα Προτάσεων δεδομένου ότι εμπλουτίζει τους λανθάνοντες παράγοντες που πρέπει να μάθει ο αλγόριθμος και συνεπώς επιτυγχάνεται μεγαλύτερη ακρίβεια στις προτάσεις που παράγει.
Αρκετές μέθοδοι έχουν εφαρμοστεί σε τέτοια κοινωνικά δίκτυα, όπως οι παραδοσιακές τεχνικές Matrix Factorization που χρησιμοποιούν πληροφορίες βαθμολογίας και κοινωνικές σχέσεις ή τεχνικές Deep Learning. Από το 2008 έχει γίνει αρκετή ερευνητική προσπάθεια για να μελετηθεί η επίδραση που έχουν οι κοινωνικές σχέσεις μεταξύ των χρηστών στο έργο των προτάσεων. Πρόκειται για έναν τομέα για τον οποίο γίνεται μια συνεχής προσπάθεια βελτίωσης από πολλές ερευνητικές ομάδες και επιχειρήσεις και είναι προφανές ότι υπάρχουν πολλά περισσότερα που πρέπει και μπορούν να επιτευχθούν.

Στο παρόν πείραμα εφαρμόζουμε διάφορους αλγόριθμους που προσεγγίζουν την παροχή προτάσεων σε κοινωνικά δίκτυα με διαφορετικό τρόπο. Ο στόχος είναι να μάθουμε ποια είναι τα μέσα για να πετύχουμε ένα καλό σε ακρίβεια σύνολο προτάσεων σε έναν χρήστη, λαμβάνοντας υπόψη τόσο το προφίλ αξιολόγησής του που αποκτάται μέσω των προηγούμενων αξιολογήσεων του, αλλά και μέσω των κοινωνικών του συνδέσεων με άλλους χρήστες.

Το πείραμα διεξήχθη σε δύο διαφορετικά σύνολα δεδομένων για να αξιολογήσουμε και να συγκρίνουμε διάφορους αλγόριθμους προτάσεων που λαμβάνουν υπόψη τόσο τις βαθμολογίες όσο και τις κοινωνικές πληροφορίες μεταξύ των χρηστών. Ο τομέας στον οποίο γίνονται οι προτάσεις είναι οι ταινίες και τα σύνολα δεδομένων είναι το FilmTrust και το CiaoDVD. Και τα δύο σύνολα δεδομένων ήταν σε μορφή CSV και ήταν ήδη σχεδόν έτοιμα για διοχέτευση στους αλγόριθμους.

FilmTrust

Το FilmTrust [30] είναι ένα μικρό σύνολο δεδομένων που ανιχνεύτηκε από ολόκληρο τον ιστότοπο FilmTrust τον Ιούνιο του 2011. Το σύνολο δεδομένων FilmTrust δεν περιέχει πρόσθετες πληροφορίες σχετικά με κριτικές αξιολόγησης ή χαρακτηριστικά ταινιών. Αποτελείται από δύο αρχεία, rating.txt με μορφή: userID, movieID, movieRating και trust.txt. που περιέχει κατευθυνόμενες αξιολογήσεις αξιοπιστίας με μορφή: trustorID, trusteeID, trustRating. Η τιμή trustRating είναι δυαδική που σημαίνει ότι εάν ένας χρήστης εμπιστεύεται κάποιον, η τιμή αξιολόγησης είναι απόλυτη 1. Το FilmTrust περιέχει 1508 χρήστες, 2071 στοιχεία, 35497 αξιολογήσεις και 1853 κατευθυνόμενες συνδέσεις εμπιστοσύνης.

CiaoDVD

Το σύνολο δεδομένων CiaoDVD [30] είναι ένα σύνολο δεδομένων που ανιχνεύτηκε από ολόκληρη την κατηγορία DVD από τον ιστότοπο dvd.ciao.co.uk τον Δεκέμβριο του 2013. Σε αντίθεση με το FilmTrust, το CiaoDVD περιέχει πρόσθετες πληροφορίες εκτός από αξιολογήσεις και κοινωνικές πληροφορίες μεταξύ των χρηστών. Αποτελείται από τρία αρχεία, movie-rating.txt με μορφή: userID, movieID, movie-categoryID, reviewID, movieRating, reviewDate και trusts.txt που περιέχει κατευθυνόμενες αξιολογήσεις εμπιστοσύνης με μορφή: trustorID, trusteeID, trustRating και τέλος review-rating.txt με περιέχει αξιολογήσεις στις κριτικές ταινιών με μορφή userID, reviewID, reviewRating (χρησιμότητα). Οι στήλες πέραν των userID,movieID και movieRating στο αρχείο movie-rating.txt καταργήθηκαν για να ταιριάζουν στους αλγόριθμους. Το CiaoDVD περιέχει 6212 χρήστες, 15001 αξιολογήσεις και 40133 αξιόπιστες συνδέσεις.

Aspect	FilmTrust	CiaoDVD
Num of Users	1,508	7,375
Num of Items	2,071	99,746
Num of Ratings	35,497	280,391
Num of Trust	1,853	111,781
Rating density degree	1.14%	0.03%
Trust density degree	0.42%	0.23%

Table 5. Απεικόνιση των χαρακτηριστικών των dataset FilmTrust - CiaoDVD

4.2 Μετρικές Απόδοσης

Τα Συστήματα Προτάσεων λειτουργούν με σκοπό είτε να παράγουν λίστες αντικειμένων τα οποία θα αρέσουν περισσότερο στο χρήστη, είτε να προβλέψουν τις βαθμολογίες που θα έδιναν οι χρήστες σε ορισμένα αντικείμενα με σκοπό να προτείνουν φυσικά αυτά με την υψηλότερη βαθμολογία. Στην περίπτωσή μας, ασχολούμαστε με την πρόβλεψη βαθμολογίας και όχι με την παραγωγή λίστας προτάσεων, γνωστή και ως TOP-N Recommendation task. Εάν ασχολούμασταν με την παραγωγή λίστας προτάσεων, τότε οι μετρικές απόδοσης που θα επιλέγαμε θα ήταν η ακρίβεια (Precision) και η ανάκληση (Recall).

Στην περίπτωση μας ωστόσο που προσπαθούμε να προβλέψουμε τον βαθμό αρεσκείας των χρηστών προς αντικείμενα, οι καταλληλότερες μετρικές απόδοσης είναι το Round Mean Square Error (RMSE) και το Mean Absolute Error (MAE).

Μία παρατήρηση όσον αφορά τα μέτρα απόδοσης RMSE και MAE είναι πως το πρώτο ‘τιμωρεί’ περισσότερο τις ακραίες τιμές λόγω του τετραγώνου.

Η αξιολόγηση όλων των αλγορίθμων γίνεται με τη μέτρηση ακρίβειας Round Mean Square Error (RMSE) και Mean Absolute Error (MAE). Τα πειράματα διεξήχθησαν με διαφορετική αναλογία δοκιμής εκπαίδευσης του συνόλου δεδομένων για τμήματα εκπαίδευσης 60% και 80% (training / test). Αυτό είναι σημαντικό για να αναγνωρίσουμε το ποσοστό των δεδομένων που κάθε αλγόριθμος χρειάζεται για να μάθει καλύτερα. Η τεχνική K-fold cross-validation χρησιμοποιείται στη διαδικασία εκπαίδευσης, η οποία είναι μια διαδικασία δειγματοληψίας που χρησιμοποιείται για την αξιολόγηση των μοντέλων μηχανικής μάθησης σε ένα περιορισμένο δείγμα δεδομένων. Ο παράγοντας K ορίζεται ως 5. Το όφελος της διαδικασίας αυτής στο πείραμά μας είναι το αποτέλεσμα μιας λιγότερο προκατειλημμένης (less biased) εκτίμησης της δεξιότητας του μοντέλου.

4.3 Περιγραφή των Αλγορίθμων και Αποτελέσματα

Οι τεχνικές που χρησιμοποιήθηκαν καθώς και τα αποτελέσματά τους περιγράφονται παρακάτω.

SVD

Ο αλγόριθμος Singular Value Decomposition (SVD) [31] έγινε πολύ δημοφιλής στο πεδίο των συστημάτων προτάσεων από την διοργάνωση του Netflix για τη βελτίωση του Συστήματος Προτάσεων της πλατφόρμας. Βασιζόμενος στην τεχνική του Matrix Factorization, ο SVD χρησιμοποιείται ως θεμελιώδης αλγόριθμος μείωσης διαστάσεων. Ωστόσο, δεν μπορεί να εφαρμοστεί αποδοτικά στις ρητές αξιολογήσεις (explicit ratings) σε μεθόδους συνεργατικού φιλτραρίσματος, καθώς οι χρήστες συνήθως επιλέγουν να βαθμολογούν μόνο ένα μικρό μέρος των προϊόντων με αποτέλεσμα το μητρώο χρηστών-αντικειμένων να είναι πολύ αραιό.

Η πρόβλεψη της βαθμολογίας γίνεται ως εξής:

$$r_{ui} = \mu + b_u + b_i + q_i^T p_u$$

Όπου μ = μέσος όρος βαθμολογιών

b_u= προκατάληψη του χρήστη (user bias)

b_i = προκατάληψη του αντικειμένου (item bias)

$q_i^T p_u$ = αλληλεπίδραση μεταξύ χρήστη και αντικειμένου που δείχνει το ενδιαφέρον του χρήστη για τα χαρακτηριστικά του αντικειμένου

Στην περίπτωση που ο χρήστης είναι άγνωστος, το b_u και p_u θεωρούνται 0 και στην περίπτωση που το αντικείμενο είναι άγνωστο τα b_i και q_i^T θεωρούνται 0.

Στα πειράματά μας δεν χρησιμοποιήσαμε τον αλγόριθμο SVD, αλλά μία επέκτασή του, προκειμένου να αναλύσουμε την επίδραση των κοινωνικών σχέσεων στην παραγωγή προτάσεων.

SVD ++

Ο SVD ++ [32] είναι μια επέκταση της ήδη υπάρχουσας τεχνικής SVD που επιτρέπει στο υπάρχον SVD μοντέλο να λαμβάνει υπόψη περαιτέρω έμμεσα δεδομένα ανατροφοδότησης (implicit feedback). Τα έμμεσα δεδομένα αναφέρονται στα δεδομένα αξιολόγησης των διαφόρων προϊόντων των χρηστών που δεν εκφράζονται ρητά, για παράδειγμα, ακόμη και αν ένας χρήστης αξιολογεί ένα συγκεκριμένο προϊόν ανεξάρτητα από τη βαθμολογία αξιολόγησης του προϊόντος, υπάρχει πιθανότητα να υπάρχει ενδιαφέρον για το προϊόν. Για παράδειγμα, στην περίπτωση DVD ταινιών, ανεξάρτητα από την αξιολόγηση, μπορεί να κριθεί ότι υπάρχει προτίμηση για την ταινία με μόνο μία εγγραφή ενοικίασης. Αυτό δεν εξετάζεται στον SVD, επομένως ο SVD επεκτείνεται έτσι ώστε τέτοια δεδομένα να μπορούν να ληφθούν υπόψη στον SVD ++. Με άλλα λόγια, ο SVD ++ χρησιμοποιεί αυτά τα επιπλέον σιωπηρά δεδομένα για να

εμπλουτίσει τα διανύσματα χαρακτηριστικών των χρηστών και επομένως να ενισχύσει τις προτάσεις.

Η βαθμολογία του προϊόντος του χρήστη προβλέπεται μέσω της εξίσωσης:

$$r_{u,j} = b_u + b_j + \mu + q_j^T (p_u + |I|_u^{-1/2} \sum_{i \in I_u} y_i)$$

Όπου το μ δείχνει τη συνολική μέση βαθμολογία, b_j και b_u δείχνουν τις παρατηρούμενες αποκλίσεις από τη μέση τιμή για τον χρήστη u και το προϊόν j αντίστοιχα και το y_i υποδηλώνει την σιωπηρή επίδραση των στοιχείων που έχουν αξιολογηθεί από τον χρήστη 'u' στο παρελθόν στις αξιολογήσεις των αγνώστων στοιχείων στο μέλλον. Το $q_j^T p_u$ δείχνει την αλληλεπίδραση μεταξύ του χρήστη u και του αντικειμένου 'j' που δείχνει το γενικό ενδιαφέρον του για τα χαρακτηριστικά του αντικειμένου.

Θα πρέπει να σημειωθεί εδώ ότι ο SVD ++ δεν χρησιμοποιεί πληροφορίες περί κοινωνικών σχέσεων μεταξύ χρηστών για την εργασία προτάσεων και ο στόχος της δοκιμής του είναι να επιβεβαιωθεί ότι οι προτάσεις ενισχύονται κατά τη χρήση κοινωνικών πληροφοριών των χρηστών. Συγκεκριμένα, η απόφαση επιλογής του SVD ++ ως αλγορίθμου που χρησιμοποιεί μόνο πληροφορίες βαθμολογίας είναι ότι δείχνει ισχυρή απόδοση μεταξύ άλλων παρόμοιων τεχνικών.

Αποτελέσματα του SVD++

CiaoDVD (60%)	MAE	0.8794
	RMSE	1.1799
CiaoDVD (80%)	MAE	0.8889
	RMSE	1.1943
FilmTrust (60%)	MAE	0.8911
	RMSE	1.1815
FilmTrust (80%)	MAE	0.8922
	RMSE	1.1806

Table 6. Αποτελέσματα του SVD++

Social Matrix Factorization (SocialMF)

Το Social Matrix Factorization [33] είναι μια προσέγγιση βάσει μοντέλου (model-based) για σύσταση σε κοινωνικά δίκτυα χρησιμοποιώντας τεχνικές παραγοντοποίησης μητρώου χρηστών-αντικειμένων (Matrix Factorization). Η κύρια ιδέα πίσω από αυτό το μοντέλο είναι η διάδοση εμπιστοσύνης μεταξύ των χρηστών. Το διάνυσμα λανθάνουσας δυνατότητας ενός χρήστη εξαρτάται από τα διανύσματα κρυμμένων χαρακτηριστικών των άμεσων γειτόνων του. Συγκεκριμένα, σε αυτήν την εργασία, η εκτίμηση του διανύσματος κρυμμένων χαρακτηριστικών ενός χρήστη είναι ο σταθμισμένος μέσος όρος των διανυσμάτων κρυμμένων χαρακτηριστικών των άμεσων γειτόνων του.

Οι διατυπώσεις των παραπάνω φαίνονται στις παρακάτω εξισώσεις:

$$U_u = \sum_{v \in N_u} T_{u,v} U_v$$

Αυτή η εξίσωση αντιπροσωπεύει το εκτιμώμενο διάνυσμα κρυμμένων χαρακτηριστικών του χρήστη, δεδομένου του διανύσματος χαρακτηριστικών των άμεσων γειτόνων του. Το N_u δηλώνει τους γείτονες του χρήστη και το $T_{u,v}$ τη σχέση εμπιστοσύνης μεταξύ του χρήστη και των γειτόνων του και δεδομένου ότι η εκτίμηση του διανύσματος κρυμμένων χαρακτηριστικών ενός χρήστη είναι ο σταθμισμένος μέσος όρος των λανθάνουσων διανυσμάτων χαρακτηριστικών των γειτόνων του, το διάνυσμα κρυμμένων χαρακτηριστικών του χρήστη u μπορεί να συναχθεί ως:

$$\begin{pmatrix} U_{u,1} \\ U_{u,2} \end{pmatrix} \dots U_{u,K} = \begin{pmatrix} U_{1,1} & U_{2,1} & \dots & U_{N,1} \\ U_{1,2} & U_{2,2} & \dots & U_{N,2} \end{pmatrix} \dots U_{1,K} \quad U_{2,K} \quad \dots \quad U_{N,K} \begin{pmatrix} T_{u,1} \\ T_{u,2} \end{pmatrix} \dots T_{u,N}$$

Υπάρχουν δίκτυα όπου υπάρχει η δυνατότητα έκφρασης του βαθμού εμπιστοσύνης σε έναν χρήστη σε βαθμίδα και άλλα δίκτυα όπου ο βαθμός εμπιστοσύνης είναι δυαδικός. Στο πείραμά μας, ο βαθμός εμπιστοσύνης είναι δυαδικός, επομένως 0 σημαίνει καμία σχέση εμπιστοσύνης μεταξύ των χρηστών, ενώ 1 σημαίνει ότι υπάρχει σχέση εμπιστοσύνης.

Στο βασικό Matrix Factorization, τα χαρακτηριστικά μαθαίνονται με βάση μόνο τις παρατηρούμενες βαθμολογίες. Ωστόσο, στα δίκτυα κοινωνικής αξιολόγησης της πραγματικής ζωής, ένα μεγάλο μέρος των χρηστών δεν έχει εκφράσει βαθμολογίες ενώ αρέσκονται στο να συμμετέχουν μόνο στο κοινωνικό δίκτυο. Έτσι, τα χαρακτηριστικά τους δεν μπορούν να εξακριβωθούν με βάση τις παρατηρούμενες βαθμολογίες τους.

Ωστόσο, το μοντέλο SocialMF μπορεί να χειριστεί αυτούς τους χρήστες πολύ καλά. Το μοντέλο SocialMF μαθαίνει να συντονίζει τα λανθάνοντα χαρακτηριστικά αυτών των χρηστών κοντά στους γείτονές τους. Έτσι, παρά το γεγονός ότι δεν έχουν εκφραστεί αξιολογήσεις, τα διανύσματα χαρακτηριστικών αυτών των χρηστών θα μάθουν να είναι κοντά στους γείτονές τους. Ουσιαστικά, οι σχέσεις κοινωνικής εμπιστοσύνης μεταξύ των χρηστών είναι μια παρατηρούμενη εξάρτηση μεταξύ των διανυσμάτων χαρακτηριστικών των χρηστών.

Αποτελέσματα του SocialMF

CiaoDVD (60%)	MAE	0.8553
	RMSE	1.1146
CiaoDVD (80%)	MAE	0.8531
	RMSE	1.1101
FilmTrust (60%)	MAE	0.8586
	RMSE	1.1172
FilmTrust (80%)	MAE	0.8774
	RMSE	1.1323

Table 7. Αποτελέσματα του SocialMF

Social Recommendation Trust Ensemble (RSTE)

Ακολουθώντας το παράδειγμα του Social Matrix Factorization, το RSTE [34] είναι μια προσέγγιση που βασίζεται στην παραγοντοποίηση του μητρώου χρηστών-αντικειμένων για την παροχή κοινωνικών προτάσεων. Το σκεπτικό πίσω από αυτήν την προσέγγιση είναι ότι το γούστο μας συνδέεται με το γούστο των φίλων μας και οι προτιμήσεις μας μπορούν να επηρεαστούν από αυτούς.

Έτσι, η πρόβλεψη της βαθμολογίας που θα έδινε ένας χρήστης σε ένα στοιχείο μπορεί να γενικευθεί ως:

$$R_{i,k} = \sum_{j \in T(i)} R_{j,k} S_{i,j}$$

όπου $R_{j,k}$είναι το σκορ που ο χρήστης j έδωσε στο στοιχείο k, και $T(i)$ είναι ο αριθμός αξιόπιστων φίλων του χρήστη j. Ο όρος $S_{i,j} \in (0, 1]$ και μπορεί να ερμηνευθεί ως το πόσο εμπιστεύεται ή γνωρίζει ένας χρήστης i έναν χρήστη j σε ένα κοινωνικό δίκτυο.

Στη συνέχεια, η πρόβλεψη των αξιολογήσεων του χρήστη j σε όλα τα στοιχεία μπορεί να συναχθεί ως εξής:

$$\begin{pmatrix} R_{i,1} \\ R_{i,2} \end{pmatrix} \dots R_{i,n} = \begin{pmatrix} R_{1,1} & R_{2,1} & \dots & R_{m,1} \\ R_{1,2} & R_{2,2} & \dots & R_{m,2} \end{pmatrix} \dots R_{1,n} \quad R_{2,n} \quad \dots \quad R_{m,n} \begin{pmatrix} S_{i,1} \\ S_{i,2} \end{pmatrix} \dots S_{i,m}$$

Ωστόσο, κάθε χρήστης έχει το δικό του γούστο και ταυτόχρονα, μπορεί να επηρεαστεί από τους φίλους του / της που εμπιστεύεται. Ως εκ τούτου, προκειμένου να οριστεί το μοντέλο πιο ρεαλιστικά, κάθε παρατηρούμενη βαθμολογία στο μητρώο χρηστών-αντικειμένων πρέπει να αντικατοπτρίζει και τους δύο αυτούς παράγοντες. Με βάση αυτό το κίνητρο, η κατανομή υπό όρους πάνω από τις παρατηρούμενες αξιολογήσεις επηρεάζεται και από τους δύο παραπάνω παράγοντες, το δικό μας γούστο και τις προτιμήσεις των αξιόπιστων φίλων μας. Έτσι, προστίθεται μια παράμετρος ‘α’ προκειμένου να συντονίσει την επιρροή που ασκούν οι έμπιστοι φίλοι στις προτάσεις του χρήστη.

Αποτελέσματα του RSTE

CiaoDVD (60%)	MAE	0.8695
	RMSE	1.1470
CiaoDVD (80%)	MAE	0.8636
	RMSE	1.1416
FilmTrust (60%)	MAE	0.8735
	RMSE	1.1500
FilmTrust (80%)	MAE	0.8785
	RMSE	1.1429

Table 8. Αποτελέσματα του RSTE

Social Regularization (SoReg)

Το SoReg [35] βασίζεται στην παραγοντοποίηση μητρώου χρηστών-αντικειμένων (Matrix Factorization) μέσω κανονικοποίησης των κοινωνικών σχέσεων. Οι πληροφορίες κοινωνικού δικτύου χρησιμοποιούνται στο σχεδιασμό δύο όρων κοινωνικής κανονικοποίησης για τον περιορισμό της αντικειμενικής συνάρτησης (objective function) παραγοντοποίησης του μητρώου. Αυτός ο όρος κοινωνικής κανονικοποίησης υποθέτει ότι το γούστο κάθε χρήστη πλησιάζει το μέσο γούστο των φίλων του. Συγκεκριμένα, ένας όρος κοινωνικής κανονικοποίησης προστίθεται στην αντικειμενική συνάρτηση, προκειμένου να ελαχιστοποιηθούν οι αποστάσεις των προτιμήσεων μεταξύ του χρήστη και των φίλων του.
Τυπικά ο όρος κοινωνικής κανονικοποίησης παρουσιάζεται στην παρακάτω εξίσωση:

$$\frac{\alpha}{2}\sum_{i=1}^{m}|U_i - \frac{1}{F^+(i)}|\sum_{f\in F^+(i)} U_f||_F^2$$

Συγκεκριμένα, εάν η λίστα φίλων του χρήστη i είναι $F^+(i)$,τότε θα μπορούσαμε να υποθέσουμε ότι το γούστο του, U_i (διάνυσμα χαρακτηριστικών), πρέπει να είναι κοντά στις μέσες προτιμήσεις όλων των χρηστών στο σύνολο $F^+(i)$. Ο παράγοντας ‘α’ καθορίζει το ποσοστό της κοινωνικής επιρροής που θέλουμε να εισάγουμε κατά την παροχή προτάσεων. Ωστόσο, ορισμένοι φίλοι μπορεί να έχουν παρόμοια γούστα με αυτόν τον χρήστη, ενώ μερικοί άλλοι μπορεί να έχουν εντελώς διαφορετικ. Ως εκ τούτου, ένα πιο ρεαλιστικό μοντέλο πρέπει να αντιμετωπίζει με διαφορετικό τρόπο όλους τους φίλους ανάλογα με το πόσο παρόμοιοι είναι. Έτσι, ένας καλύτερος όρος κοινωνικής κανονικοποίησης είναι:

$$\frac{a}{2}\sum_{i=1}^{m}||U_i - \left(\frac{\sum_{f\in F^+(i)} Sim(i,f)\times U_f}{\sum_{f\in F^+(i)} Sim(i,f)}\right)||_F^2$$

Όπου το $Sim(i,f)$ είναι μια συνάρτηση ομοιότητας που επιτρέπει στον όρο κοινωνικής κανονικοποίησης να αντιμετωπίζει διαφορετικά τους φίλους των χρηστών. Η υψηλή βαθμολογία του $Sim(i,f)$ οδηγεί σε υψηλή συνεισφορά στο μέσο γούστο.

Το συγκεκριμένο μοντέλο επιβάλλει έναν όρο κοινωνικής κανονικοποίησης για να περιορίσει το γούστο του χρήστη στη μέση γεύση των φίλων του. Ωστόσο, αυτή η προσέγγιση δεν είναι ευαίσθητη στους χρήστες των οποίων οι φίλοι έχουν διαφορετικά γούστα. Αυτό θα προκαλέσει πρόβλημα απώλειας πληροφοριών, το οποίο θα οδηγήσει σε ανακριβή μοντελοποίηση του διανύσματος χαρακτηριστικών U_i. Ως εκ τούτου, για την αντιμετώπιση αυτού του προβλήματος, προτείνεται ένας άλλος όρος κοινωνικής κανονικοποίησης για την επιβολή περιορισμών μεταξύ ενός χρήστη και των φίλων του ξεχωριστά. Ο όρος κοινωνικής κανονικοποίησης είναι:

$$\frac{\beta}{2}\sum_{i=1}^{m}\sum_{f\in F^{+}(i)} Sim(i,f)\left|\left|U_i - U_f\right|\right|_F^2$$

Όπου το $\beta > 0$ καθορίζει το ποσοστό επηρροής των φίλων όπως προηγουμένως και το $Sim(i,f)$ είναι η συνάρτηση ομοιότητας μεταξύ των χρηστών.

Η *Pearson Correlation Coefficient* χρησιμοποιείται ως συνάρτηση ομοιότητας στην εξίσωση και βασίζεται στα αντικείμενα τα οποία οι συγκρινόμενοι χρήστες έχουν βαθμολογήσει από κοινού:

$$\mathrm{S}im(i,f) = \frac{\sum_{j\in I(i)\cap I(f)}(R_{ij}-R`_i)(R_{fj}-R'_f)}{\sqrt{\sum_{j\in I(i)\cap I(f)}(R_{ij}-R`_i)^2}\sqrt{\sum_{j\in I(i)\cap I(f)}(R_{fj}-R'_f)^2}}$$

Όπου το j ανήκει στο υποσύνολο των στοιχείων που ο χρήστης i και ο χρήστης f βαθμολόγησαν. Το R_{ij} αναφέρεται στην βαθμολογία που έδωσε ο χρήστης i στο αντικείμενο j. Το $R`_i$αντιπροσωπεύει τη μέση βαθμολογία που έχει δώσει ο χρήστης i. Από αυτόν τον ορισμό, η ομοιότητα χρήστη $\mathrm{S}im(i,f)$ κυμαίνεται από $[-1, 1]$ και μια μεγαλύτερη τιμή σημαίνει ότι οι χρήστες i και f είναι πιο παρόμοιοι.

Αποτελέσματα SoReg

CiaoDVD (60%)	MAE	0.8455
	RMSE	1.0964
CiaoDVD (80%)	MAE	0.8448
	RMSE	1.0996
FilmTrust (60%)	MAE	0.8486
	RMSE	1.1037
FilmTrust (80%)	MAE	0.8663
	RMSE	1.1202

Table 9. Αποτελέσματα του SoReg

TrustSVD

Ο TrustSVD [36] είναι μια τεχνική παραγοντοποίησης βασισμένης στην εμπιστοσύνη, η οποία επεκτείνει το υπάρχον μοντέλο SVD ++ που περιεγράφηκε παραπάνω, το οποίο εμπεριέχει την ρητή και σιωπηρή επίδραση των αξιολογήσεων (implicit – explicit feedback).

Όπως αναφέραμε και παραπάνω το σκεπτικό πίσω από τον SVD++ είναι ότι πέρα από τα χαρακτηριστικά διανύσματα των χρηστών και των αντικειμένων, λαμβάνονται υπόψη και οι προκαταλήψεις που υπάρχουν τόσο στους χρήστες όσο και στα αντικείμενα (user bias, item bias). Επίσης ο SVD++ λαμβάνει υπόψη περαιτέρω έμμεσα δεδομένα ανατροφοδότησης (implicit feedback) όσον αφορά τις αξιολογήσεις των αντικειμένων, τα οποία ενδέχεται να μην έχουν αναφερθεί ρητά από τους χρήστες.

Υπενθυμίζουμε τον τρόπο με τον οποίο γίνεται η πρόβλεψη των αξιολογήσεων με το μοντέλο SVD++:

$$r_{u,j} = b_u + b_j + \mu + q_j^T (p_u + |I|_u^{-1/2} \sum_{i \in I_u} y_i)$$

Όπου το μ δείχνει τη συνολική μέση βαθμολογία, b_j και b_u δείχνουν τις παρατηρούμενες αποκλίσεις από τη μέση τιμή για τον χρήστη u και το προϊόν j αντίστοιχα και το y_i υποδηλώνει την σιωπηρή επίδραση των στοιχείων που έχουν αξιολογηθεί από τον χρήστη 'u' στο παρελθόν στις αξιολογήσεις των αγνώστων στοιχείων στο μέλλον. Το $q_j^T p_u$ δείχνει την αλληλεπίδραση μεταξύ του χρήστη u και του αντικειμένου 'j' που δείχνει το γενικό ενδιαφέρον του για τα χαρακτηριστικά του αντικειμένου.

Ο TrustSVD ενσωματώνει τόσο τη ρητή (explicit) όσο και τη σιωπηρή (implicit) επίδραση των αξιολογήσεων στοιχείων καθώς και την εμπιστοσύνη των χρηστών. Επομένως, το χαρακτηριστικό διάνυσμα για ένα χρήστη 'u' διαμορφώνεται περαιτέρω στα δίκτυα κοινωνικής αξιολόγησης, λαμβάνοντας υπόψη την επίδραση τόσο των αξιολογημένων στοιχείων όσο και των αξιόπιστων χρηστών καθώς επίσης και τις σιωπηρές επιδράσεις αυτών. Η βελτίωση του μοντέλου SVD ++ γίνεται με την ενσωμάτωση της επιρροής εμπιστοσύνης. Συγκεκριμένα, η έμμεση επίδραση των αξιόπιστων χρηστών στις αξιολογήσεις αντικειμένων μπορεί να εξεταστεί με τον ίδιο τρόπο όπως τα βαθμολογημένα στοιχεία, που δίνονται από την εξίσωση:

$$r_{u,j} = b_u + b_j + \mu + q_j^T (p_u + |I|_u^{-1/2} \sum_{i \in I_u} y_i + |T|_u^{-1/2} \sum_{u \in T_u} w_u$$

Όπου w_u είναι τα χαρακτηριστικά διανύσματα των χρηστών που εμπιστεύεται ο χρήστης u, και έτσι το $q_j^T w_u$ μπορεί να εξηγηθεί ως οι αξιολογήσεις που προβλέπονται από τους αξιόπιστους χρήστες, δηλαδή από την επίδραση των αξιόπιστων χρηστών στην πρόβλεψη αξιολόγησης. Ο όρος $|T|_u^{-1/2}$ είναι το σύνολο των χρηστών που εμπιστεύεται ο χρήστης.

Αποτελέσματα TrustSVD

CiaoDVD (60%)	MAE	0.8199
	RMSE	1.0805
CiaoDVD (80%)	MAE	0.8203
	RMSE	1.0799
FilmTrust (60%)	MAE	0.8201
	RMSE	1.0687
FilmTrust (80%)	MAE	0.8196
	RMSE	1.0800

Table 10. Αποτελέσματα του TrustSVD

GraphRec

Το GraphRec [37] είναι ένα σύστημα νευρωνικού δικτύου γραφημάτων για κοινωνικές προτάσεις. Αυτή η προσέγγιση συλλαμβάνει από κοινού αλληλεπιδράσεις και απόψεις στο γράφημα χρήστη-αντικειμένων και συνεπώς μοντελοποιεί δύο γραφήματα και ετερογενείς δυνάμεις. Το μοντέλο αποτελείται από τρία στοιχεία: μοντελοποίηση χρήστη, μοντελοποίηση αντικειμένων και πρόβλεψη βαθμολογίας.

Το πρώτο συστατικό είναι η *μοντελοποίηση χρηστών*, η οποία είναι η εκμάθηση διανυσμάτων κρυμμένων χαρακτηριστικών των χρηστών.
Καθώς τα δεδομένα στα συστήματα κοινωνικών προτάσεων περιλαμβάνουν δύο διαφορετικά γραφήματα, ένα γράφημα κοινωνικής δικτύωσης και ένα γράφημα χρήστη-αντικειμένου, οι αναπαραστάσεις των χρηστών μαθαίνονται από διαφορετικές οπτικές γωνίες.
Επομένως, εισάγονται δύο συγκεντρώσεις για την επεξεργασία αυτών των δύο διαφορετικών γραφημάτων.

Η πρώτη είναι η συγκέντρωση στοιχείων, η οποία μπορεί να χρησιμοποιηθεί για την κατανόηση των χρηστών μέσω των αλληλεπιδράσεων τους με τα αντικειμένων στο γράφημα χρήστη-αντικειμένου. Η δεύτερη συγκέντρωση είναι η κοινωνική συσσωμάτωση, η σχέση μεταξύ των χρηστών στο κοινωνικό γράφημα, η οποία μπορεί να βοηθήσει τη μοντελοποίηση των χρηστών από την κοινωνική σκοπιά. Στη συνέχεια, συνδυάζοντας πληροφορίες τόσο από τον χώρο των αντικειμένων όσο και από τον κοινωνικό χώρο και έπειτα εξάγονται τα διανύσματα κρυμμένων χαρακτηριστικών για τους χρήστες.

Για τη μοντελοποίηση του χρήστη συνδυάζονται τα στοιχεία συγκέντρωσης αντικειμένων και κοινωνικής συγκέντρωσης:

$$h_i^I = \sigma(W \times Aggre_{items}(\{x_{ia}, \forall a \in C(i)\}) + b)$$

$$h_i^S = \sigma(W \times Aggre_{neighbors}(\{h_o^I, \forall o \in N(i)\}) + b)$$

Η πρώτη εξίσωση απεικονίζει τη συγκέντρωση των αντικειμένων για την περιγραφή του χρήστη. Όπου $C(i)$ είναι το σύνολο των στοιχείων με τα οποία αλληλεπίδρασαν οι χρήστες u_i , το x_{ia} είναι ένας φορέας αναπαράστασης για να υποδηλώσει την αλληλεπίδραση σε συνδυασμό με την αξιολόγηση μεταξύ του χρήστη u_i και ενός στοιχείου a και το $Aggre_{items}$ είναι η συνάρτηση συνάθροισης των αντικειμένων.

Η δεύτερη εξίσωση απεικονίζει την κοινωνική συγκέντρωση που συνδυάζει τα διανύσματα κρυμμένων χαρακτηριστικών των γειτόνων του χρήστη από το γράφημα χρήστη-αντικειμένου. Το $N(i)$ είναι το σύνολο των γειτόνων του u_i , το h_o^I είναι το διάνυσμα κρυμμένων χαρακτηριστικών των γειτόνων του χρήστη και το $Aggre_{neighbors}$ δηλώνει τη συνάθροιση των γειτόνων του χρήστη.

Για την καλύτερη εκμάθηση των διανυσμάτων χαρακτηριστικών των χρηστών, τόσο οι απεικονίσεις τους από την κοινωνική σκοπιά όσο και οι απεικονίσεις τους από το χώρο των αντικειμένων πρέπει να συνδυαστούν, καθώς το γράφημα κοινωνικής δικτύωσης και το γράφημα χρήστη-στοιχείου παρέχουν πληροφορίες για χρήστες από διαφορετικές οπτικές γωνίες. Έτσι, συνδυάζοντας τα δύο αυτά διανύσματα χαρακτηριστικών μέσω ενός Multi-

Layer-Perceptron (MLP), παράγεται μία ολοκληρωμένη και πιο πλούσια σε πληροφορίες αναπαράσταση των χρηστών.

$$h_i = h_i^I * h_i^S$$

Το δεύτερο συστατικό είναι η *μοντελοποίηση αντικειμένων*, η οποία είναι η εκμάθηση διανυσμάτων κρυμμένων χαρακτηριστικών των αντικειμένων. Προκειμένου να ληφθούν υπόψη τόσο οι αλληλεπιδράσεις όσο και οι απόψεις (σε μορφή αξιολογήσεων) στο γράφημα χρήστη-αντικειμένου, εισάγεται η συσσωμάτωση χρηστών, η οποία είναι η συγκέντρωση τόσο των αλληλεπιδράσεων όσο και των αξιολογήσεων των χρηστών προς τα αντικείμενα.

Για κάθε αντικείμενο, οι πληροφορίες από το σύνολο των χρηστών που έχουν αλληλεπιδράσει με αυτό συγκεντρώνονται ως εξής:

$$z_j = \sigma\left(W \times Aggre_{users}\left(\{f_{jt}, \forall t \in B(j)\}\right) + b\right)$$

όπου το f_{jt} είναι η συνένωση πληροφοριών για την αλληλεπίδραση και την αξιολόγηση των χρηστών για το αντικείμενο j, το $B(j)$ είναι το σύνολο των χρηστών που έχουν αλληλεπιδράσει με το στοιχείο j και το $Aggre_{users}$ είναι η συνάρτηση συνάθροισης των χρηστών.

Το τρίτο συστατικό είναι η πρόβλεψη βαθμολογίας των αντικειμένων από τους χρήστες. Έχοντας πρώτα αναλύσει τα χαρακτηριστικά των χρηστών και των αντικειμένων από την σκοπιά των αντικειμένων και των κοινωνικών σχέσεων, μένει να συνδυαστούν αυτές οι πληροφορίες για την ορθότερη πρόβλεψη της βαθμολογίας και εν συνεχεία την ακριβέστερη παροχή προτάσεων.
Τα διανύσματα κρυμμένων χαρακτηριστικών των χρηστών και στοιχείων (δηλαδή τα h_i και z_j), συνενώνονται πρώτα και στη συνέχεια τροφοδοτούνται σε ένα Multi-Layer Perceptron (MLP) για την πρόβλεψη βαθμολογίας ως:

$$g_1 = \left[h_i * z_j\right]$$

$$g_2 = \sigma(W_2 c_1 + b_2)$$

...

$$g_{l-1} = \sigma(W_l g_{l-1} + b_l)$$

$$r'_{ij} = w^T g_{l-1}$$

όπου l είναι ο δείκτης του κρυφού επιπέδου του MLP και το r'_{ij} είναι η προβλεπόμενη βαθμολογία από τον χρήστη i στο αντικείμενο j.

Αποτελέσματα GraphRec

CiaoDVD (60%)	MAE	0.8977
	RMSE	1.2167
CiaoDVD (80%)	MAE	0.8111
	RMSE	1.1009
FilmTrust (60%)	MAE	0.7294
	RMSE	0.9683
FilmTrust (80%)	MAE	0.6985
	RMSE	0.9129

Table 11. Αποτελέσματα του GraphRec

4.5 Εμπλουτισμός των Δεδομένων

Ένα μοντέλο μηχανικής μάθησης είναι τόσο αποδοτικό όσο πλούσιες είναι οι πληροφορίες που έχει στη διάθεσή του για επεξεργασία και εκμάθηση. Βασισμένοι σε αυτήν την ιδέα θα ήταν λογικό να προσπαθήσουμε να εμπλουτίσουμε τις πληροφορίες που παρέχουμε στα μοντέλα που χρησιμοποιούμε. Φυσικά η επέκταση των πληροφοριών που χρησιμοποιεί ένα μοντέλο πρέπει να είναι ακριβής και ποιοτική και όχι απλώς μία επέκταση ως προς το μέγεθος που δεν θα επιφέρει περισσότερη πληροφορία.

Έχοντας λοιπόν στη διάθεση μας τα προαναφερθέντα μοντέλα που λειτουργούν σε κοινωνικά δίκτυα, μπορούμε να χρησιμοποιήσουμε αλγορίθμους Link Prediction προκειμένου να επεκτείνουμε τα δεδομένα μας και να δώσουμε περισσότερη πληροφορία στους αλγορίθμους.

Στα πειράματά μας χρησιμοποιήσαμε τον αλγόριθμο που αναλύσαμε σε προηγούμενο κεφάλαιο, SimRank. Ο αλγόριθμος SimRank χρησιμοποιήθηκε στο τμήμα δεδομένων που περιέχει τις διασυνδέσεις των χρηστών μεταξύ τους. Αναλύοντας έτσι τη δομή του κοινωνικού Γράφου των δεδομένων μας, καταλήξαμε σε ένα νέο σύνολο δεδομένων, φυσικά βασισμένο στο προηγούμενο, το οποίο όμως περιέχει νέες διασυνδέσεις μεταξύ των χρηστών. Οι προσθήκες που έγιναν ήταν ουσιαστικά τρεις νέοι ‘έμπιστοι χρήστες’ για κάθε χρήστη του Γράφου. Με αυτόν τον τρόπο, οι αλγόριθμοι που εκ φύσεως αναλύουν τις προτιμήσεις των χρηστών με γνώμονα τόσο το προφίλ του ίδιου του χρήστη όσο και τις προτιμήσεις των έμπιστων προς αυτόν χρηστών, έχει στη διάθεση του περισσότερη πληροφορία προς επεξεργασία.

Ακολουθεί αναπαράσταση της δομής του Γράφου πριν και μετά την υλοποίηση του SimRank στο σύνολο δεδομένων FilmTrust. Ο λόγος που επιλέξαμε το FilmTrust αντί του CiaoDVD είναι η πολύ μεγαλύτερη πυκνότητα τόσο των βαθμολογιών όσο και των σχέσεων εμπιστοσύνης μεταξύ των χρηστών.

Ως κόμβος προς εξέταση επιλέχθηκε ο κόμβος με αριθμό 60, καθώς είναι ευδιάκριτος μέσα σε όλο το Γράφο και μπορούμε να παρατηρήσουμε ξεκάθαρα την εισαγωγή νέων συνδέσεων με άλλους κόμβους.

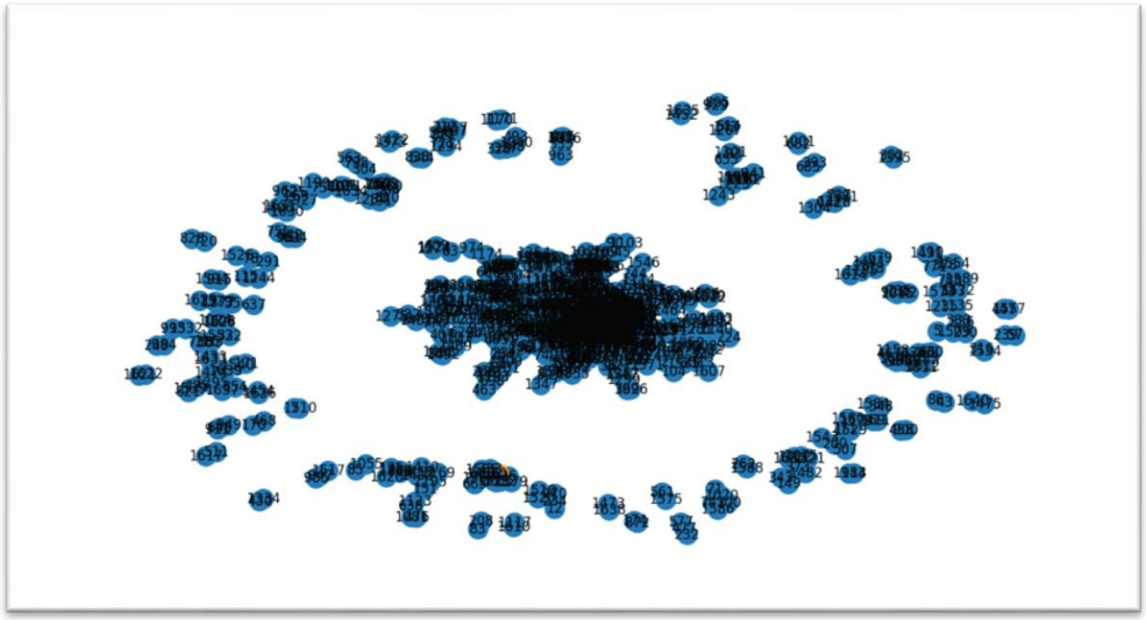

Figure 23. Αναπαράσταση Γράφου πριν την επέκτασή του 1

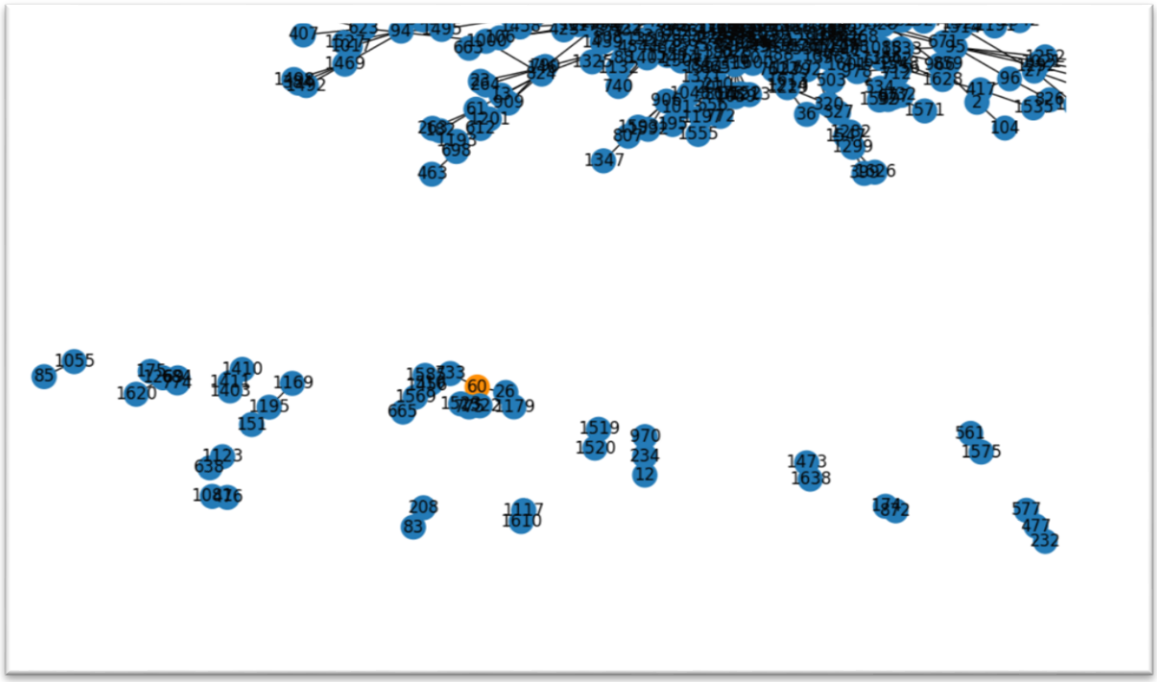

Figure 24. Αναπαράσταση Γράφου πριν την επέκτασή του 2

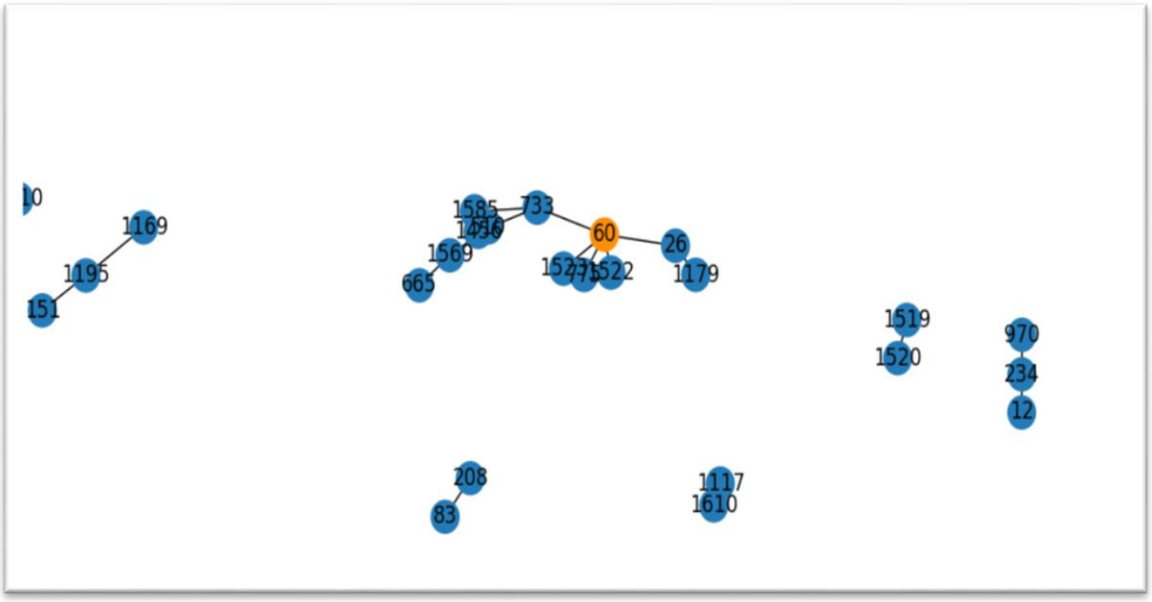

Figure 25. Αναπαράσταση Γράφου πριν την επέκτασή του 3

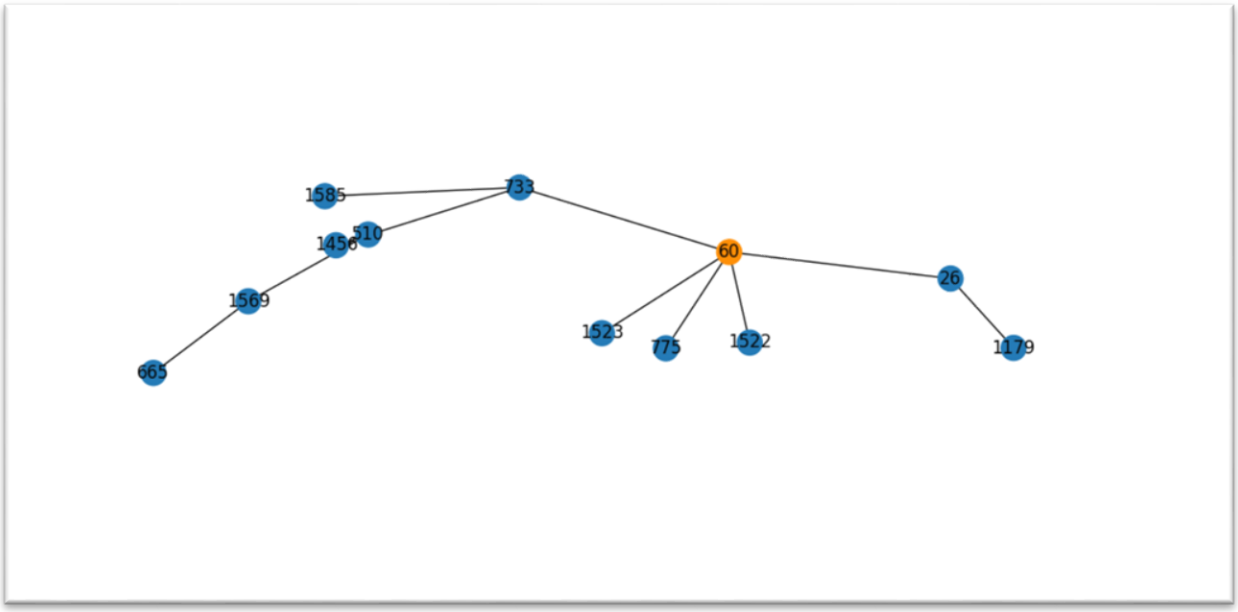

Figure 26. Αναπαράσταση Γράφου πριν την επέκτασή του 4

Στην συνέχεια ακολουθεί σειρά από απεικονίσεις του ίδιου Γράφου μετά την εισαγωγή των νέων χρηστών με τη χρήση του αλγορίθμου SimRank.

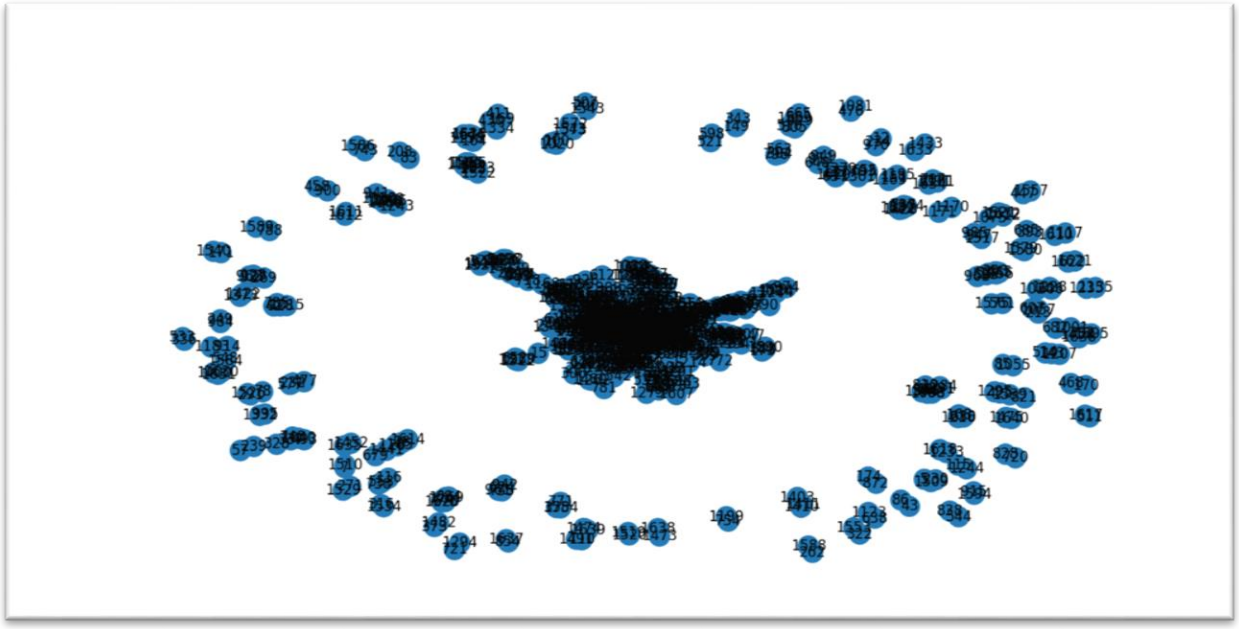

Figure 27. Αναπαράσταση Γράφου μετά την επέκτασή του 1

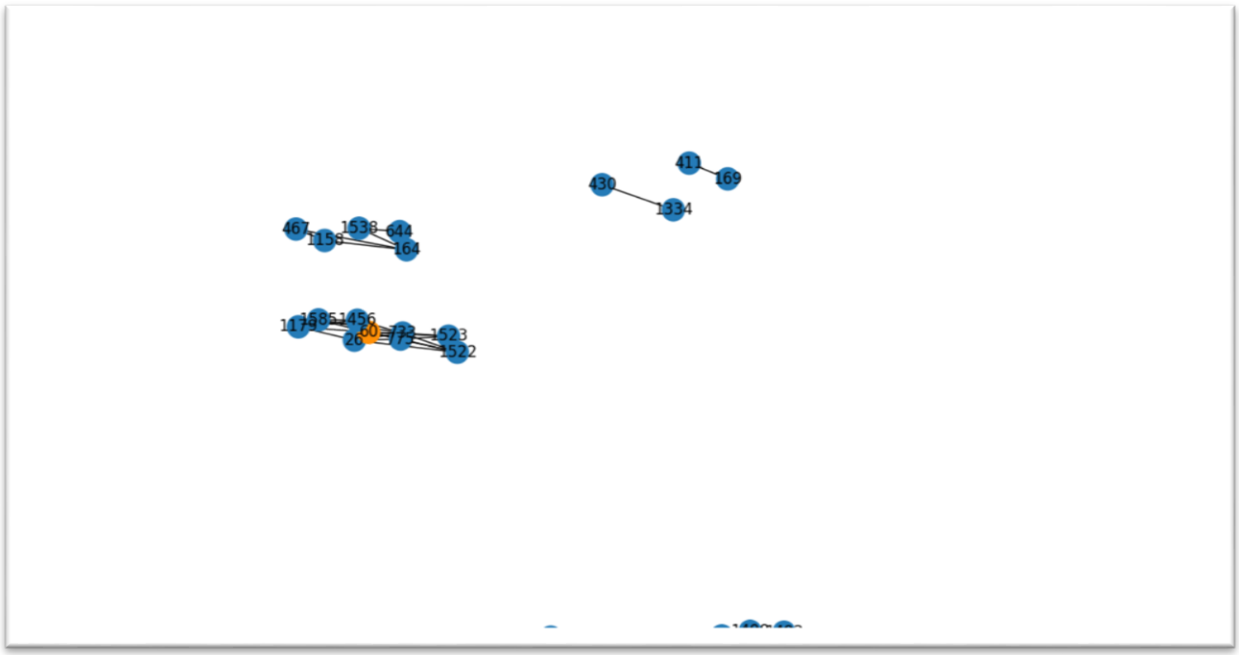

Figure 28. Αναπαράσταση Γράφου μετά την επέκτασή του 2

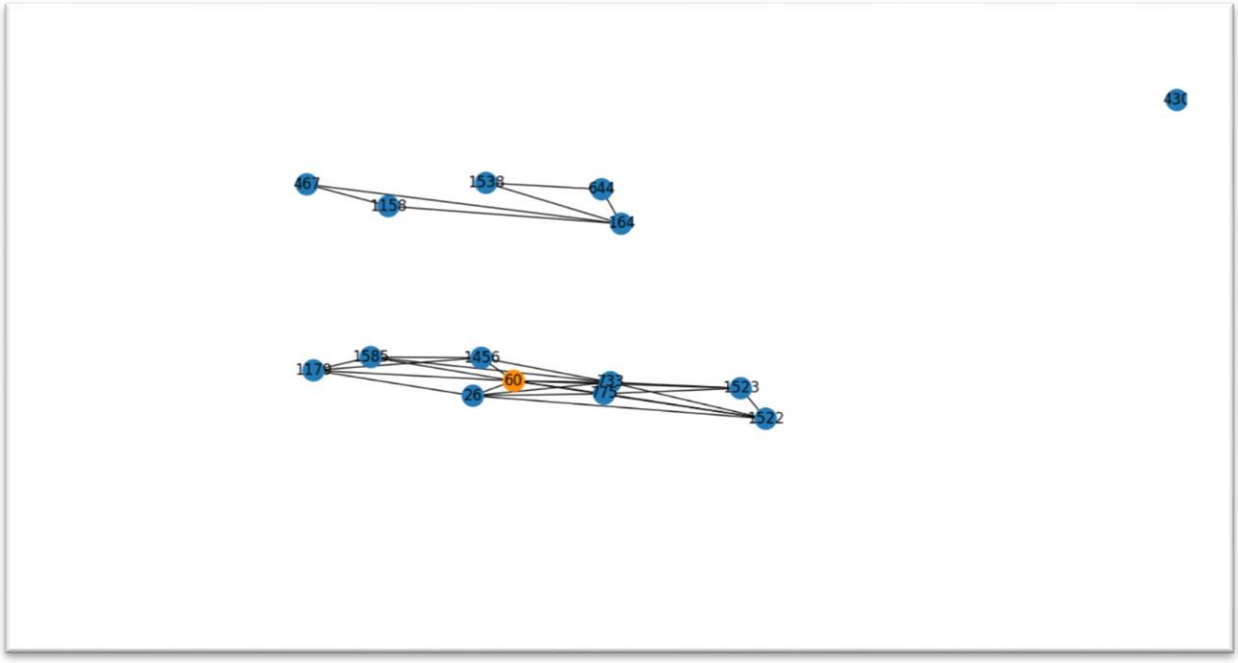

Figure 29. Αναπαράσταση Γράφου μετά την επέκτασή του 3

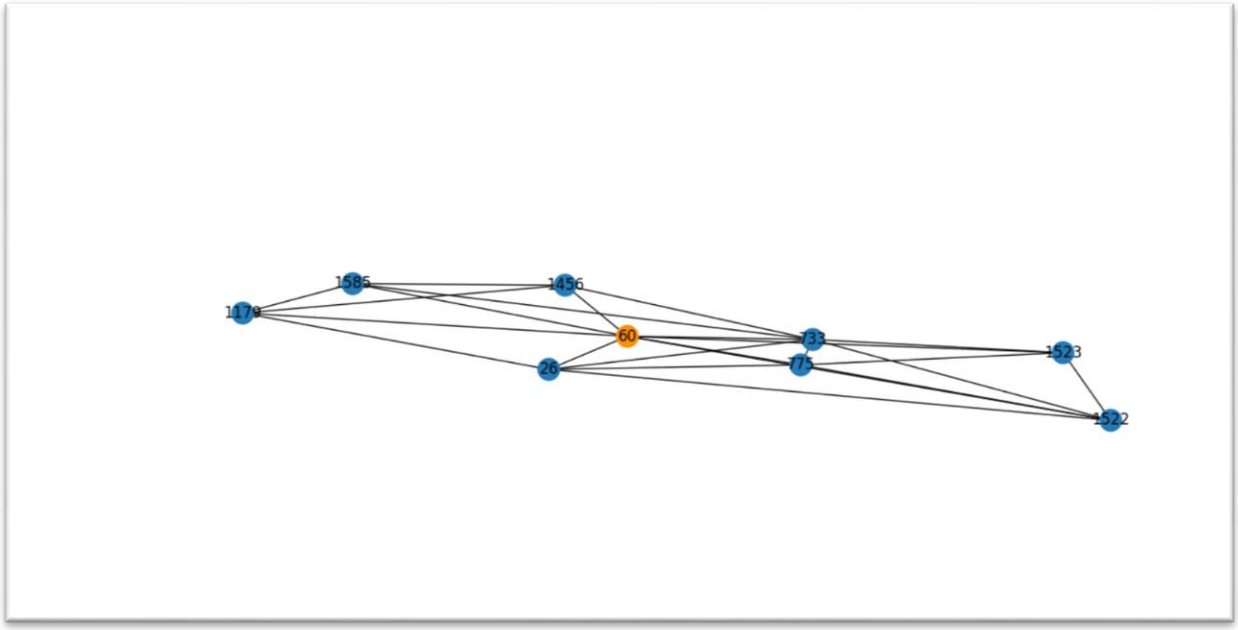

Figure 30. Αναπαράσταση Γράφου μετά την επέκτασή του 4

Είναι φανερό πως ο χρήστης 60 πλέον είναι συνδεδεμένος με περισσότερους χρήστες. Ομοίως φυσικά και οι υπόλοιποι χρήστες στην γειτονία που φαίνεται στην παραπάνω εικόνα έχουν πλέον αναπτύξει νέες σχέσεις εμπιστοσύνης με περισσότερους χρήστες. Συγκεκριμένα, επιλέχθηκε να εισάγουμε μέχρι τρεις νέους χρήστες για κάθε χρήστη του συνόλου δεδομένου

μας. Η επιλογή έγινε διαισθητικά αρχικά καθώς πρόκειται για πολύ μικρό και περιορισμένο dataset και έτσι η επιλογή μεγαλύτερου αριθμού από τρεις θα ήταν καταστροφικό για τα αποτελέσματα των αλγορίθμων μας. Σε περιπτώσεις όπου το σκορ ομοιότητας μεταξύ των χρηστών δεν ήταν ικανοποιητικό, τότε δεν συμπεριλάβαμε νέες συνδέσεις στα δεδομένα μας.

Μετά μάλιστα από δοκιμές παρατηρήσαμε πως εισάγοντας περισσότερους νέους χρήστες εμπιστοσύνης οδηγούμαστε σε χειρότερη απόδοση των αλγορίθμων καθώς παρά τα περισσότερα δεδομένα που παρέχουμε στους αλγορίθμους, η εγκυρότητά τους δεν είναι η επιθυμητή.

Εάν παρακολουθήσει κανείς τις πρώτες εικόνες στις οποίες φαίνεται ολόκληρη η δομή του Γράφου, θα μπορούσε να παρατηρήσει πως μετά τον εμπλουτισμό των δεδομένων, ο πυρήνας του Γράφου είναι πιο συνωστισμένος στο κέντρο, κάτι το οποίο είναι λογικό αφού οι κοντινοί χρήστες έχουν αναπτύξει πλέον σχέσεις μεταξύ τους.

Μετά τον εμπλουτισμό των δεδομένων, οι συνδέσεις μεταξύ των χρηστών ανήλθαν στις 3640 από τις αρχικά 1853 που περιείχε το FilmTrust dataset.

4.6 Περιγραφή του Framework που χρησιμοποιήθηκε

Όλοι οι αλγόριθμοι που χρησιμοποιήθηκαν στα πειράματα ήταν υλοποιημένοι στην γλώσσα προγραμματισμού Python 3.

Η Python είναι πλέον η πιο διαδεδομένη γλώσσα προγραμματισμού στους τομείς της Μηχανικής Μάθησης.

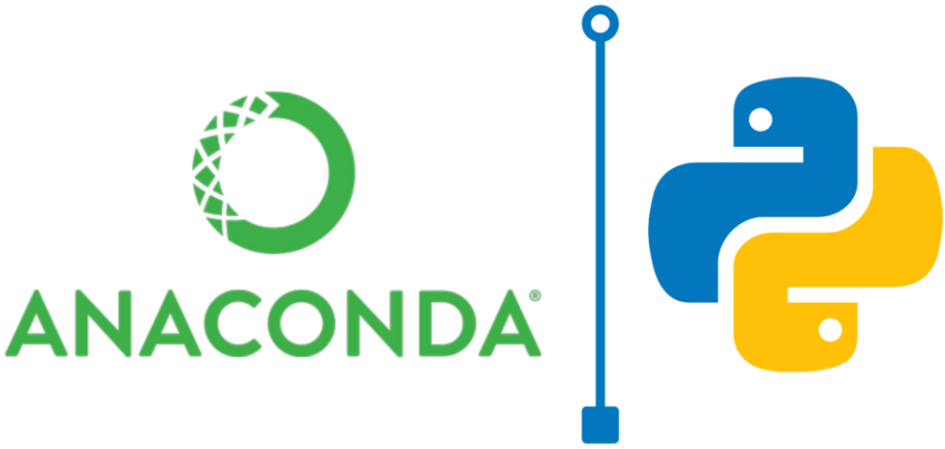

Figure 31. Anaconda - Python

Οι αλγόριθμοι γράφτηκαν επάνω στο anaconda framework , μία open source python distribution πλατφόρμα, ευνοϊκή στη χρήση πακέτων με σημαντικό ρόλο στην υλοποίηση μας

Πέραν από τους αλγορίθμους, η διαδικασία εκκαθάρισης, διαχείρισης και επεξεργασίας των δεδομένων έγιναν με τη βοήθεια βιβλιοθηκών της Python, όπως η Pandas και η Numpy.

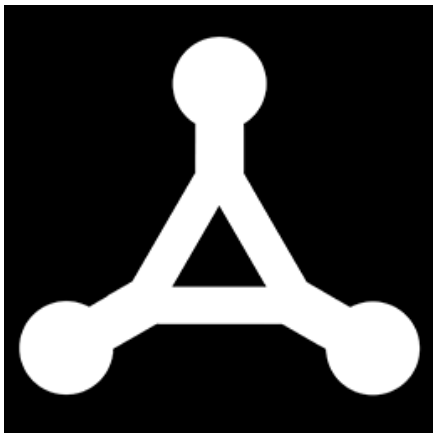

Figure 32. GraphEditor

Οι γράφοι πάνω στους οποίους υλοποιήθηκε ο αλγόριθμος SimRank σχεδιάστηκαν online με το δωρεάν εργαλείο GraphEditor από την csacademy

Figure 33. NetworkX

Η αναπαράσταση του συνόλου δεδομένων πριν και μετά την υλοποίηση του αλγορίθμου SimRank έγινε με τη βοήθεια του NetworkX, το οποίο είναι ευρέως γνωστό για τις δυνατότητες που προσφέρει για τη διαχείριση και μελέτη της δομής γράφων και δικτύων.

5.0 Αποτελέσματα

5.1 Συνολικά Αποτελέσματα και Σχολιασμός τους

Παρουσιάζουμε τα αποτελέσματα των αλγορίθμων με τα dataset FilmTrust, CiaoDVD σε ποσοστά train / test 60%-40% και 80% -20% αντίστοιχα.

Training	Metrics	Algorithms					
		SVD++	Social_MF	Social_RSTE	Social_Reg	TrustSVD	GraphRec
CiaoDVD (60%)	MAE	0.8794	0.8553	0.8695	0.8455	0.8199	0.8977
	RMSE	1.1799	1.1146	1.1470	1.0964	1.0805	1.2167
CiaoDVD (80%)	MAE	0.8889	0.8531	0.8636	0.8448	0.8203	0.8111
	RMSE	1.1943	1.1101	1.1416	1.0996	1.0799	1.1009
FilmTrust (60%)	MAE	0.8911	0.8586	0.8735	0.8486	0.8201	0.7294
	RMSE	1.1815	1.1172	1.1500	1.1037	1.0687	0.9683
FilmTrust (80%)	MAE	0.8922	0.8774	0.8785	0.8663	0.8196	0.6985
	RMSE	1.1806	1.1323	1.1429	1.1202	1.0800	0.9129

Table 12. Συνολικά Αποτελέσματα των Αλγορίθμων

Η αποτίμηση των επιδόσεων των αλγορίθμων σε όλα μας τα πειράματα έγινε με βάση τις μετρικές RMSE – MAE που περιγράψαμε σε παραπάνω κεφάλαιο. Παρατηρώντας τα αποτελέσματα μπορούμε να βγάλουμε σημαντικά συμπεράσματα.

Αρχικά μπορούμε να επιβεβαιώσουμε την βασική και λογική υπόθεση στην οποία βασίζεται η έρευνα καθώς και η πειραματική διαδικασία που παρουσιάζουμε, ότι οι προτάσεις που παράγει ένα Σύστημα Προτάσεων ενισχύεται σε μεγάλο βαθμό όταν λαμβάνονται υπόψιν οι σχέσεις μεταξύ των χρηστών. Οδηγηθήκαμε σε αυτό το συμπέρασμα παρατηρώντας τον Πίνακα [12], όπου βλέπουμε χαμηλότερη επίδοση του αλγορίθμου SVD++ ο οποίος λειτουργεί αποκλειστικά με βάση τις αλληλεπιδράσεις των χρηστών με τα αντικείμενα. Η απόδοση του SVD++ είναι σταθερά χειρότερη από αυτήν όλων των υπόλοιπων αλγορίθμων και στα 2 dataset. Η επιλογή του SVD++ έγινε καθώς παρουσιάζει πολύ καλή απόδοση σε σχέση με παρόμοιους αλγορίθμους. Είναι ασφαλές λοιπόν το συμπέρασμα μας.

Συνεχίζοντας στην εξέταση των υπόλοιπων τεχνικών μας, παρατηρούμε πως οι αλγόριθμοι SocialMF – Social_RSTE – Social_Reg παρουσιάζουν πολύ όμοια συμπεριφορά. Η συμπεριφορά αυτή είναι απολύτως λογική καθώς αυτές οι τεχνικές ακολουθούν πολύ παρόμοια λογική όσον αφορά τη χρησιμοποίηση του κοινωνικού παράγοντα των δεδομένων. Ελαφρώς καλύτερη απόδοση φαίνεται να έχει η τεχνική Social_Reg η οποία ακολουθεί το παράδειγμα των 2 προηγούμενων τεχνικών με τη διαφορά ότι εισάγει την έννοια της κοινωνικής κανονικοποίησης. Η κοινωνική κανονικοποίηση επιτρέπει στο μοντέλο να διαλέξει μόνον εκείνες τις κοινωνικές συνδέσεις του χρήστη οι οποίες θα επωφελήσουν τις προτάσεις και όχι άλλες που πιθανώς θα τις βλάψουν. Αυτή η κανονικοποίηση έχει ισχύ και στον πραγματικό κόσμο καθώς πιθανότατα δε θα συμβουλευτούμε τα ίδια άτομα για όλες μας τις αποφάσεις.

Ο αλγόριθμος TrustSVD φαίνεται να έχει καλύτερη απόδοση σε όλα τα στάδια της πειραματικής διαδικασίας από τους αλγορίθμους Social_MF - Social_RSTE - Social_Reg. Η καλύτερη απόδοση του TrustSVD μας δείχνει πόσο σημαντικό είναι να εμπλουτίζουμε στον μέγιστο δυνατό βαθμό τα χαρακτηριστικά διανύσματα των χρηστών (latent feature vectors). Η διαφορά του με τις προαναφερθέντες τεχνικές είναι πως λαμβάνει υπόψιν τόσο τη ρητή (explicit) όσο και τη σιωπηρή (implicit) επίδραση των αξιολογήσεων στοιχείων καθώς και την εμπιστοσύνη των χρηστών. Είναι ζωτικής σημασίας λοιπόν η κατασκευή των χαρακτηριστικών διανυσμάτων των χρηστών αλλά και των αντικειμένων να γίνεται με όσο το δυνατό περισσότερες πληροφορίες. Αυτές οι πληροφορίες δεν θα είναι πάντα ορατές και ρητά διατυπωμένες αλλά μερικές φορές χρειάζεται να αναλύσουμε διάφορες συμπεριφορές προκειμένου να τις ανακαλύψουμε.

Ιδιαίτερο ενδιαφέρον έχει η συμπεριφορά της τεχνικής GraphRec. Ο GraphRec είναι διαφορετικός από τους υπόλοιπους αλγορίθμους καθώς είναι ο μοναδικός που βασίζεται σε νευρωνικά δίκτυα για την παροχή προτάσεων. Ένα πρώτο χαρακτηριστικό για την συμπεριφορά του GrpahRec είναι η πολύ μεγάλη βελτίωση του αλγορίθμου όταν χρησιμοποιεί το 80% των δεδομένων ως train set από ότι όταν χρησιμοποιεί το 60%. Οι υπόλοιποι αλγόριθμοι που χρησιμοποιούμε έχουν παρόμοια συμπεριφορά και στις δύο περιπτώσεις, κάτι που μας οδηγεί στο συμπέρασμα πως αρχιτεκτονικές που βασίζονται σε νευρωνικά δίκτυα, βελτιώνουν την απόδοσή τους όταν εκπαιδεύονται με μεγαλύτερο όγκο δεδομένων σε σύγκριση με πιο συμβατικά μοντέλα μηχανικής μάθησης.

Ο αλγόριθμος GraphRec φαίνεται να έχει πάρα πολύ καλή απόδοση στο FilmTrust dataset σε σχέση με τους υπόλοιπους αλγορίθμους. Ωστόσο δεν ισχύει το ίδιο και για το CiaoDVD, όπου η απόδοσή του φαίνεται ικανοποιητική μόνο στο 80% train / test των δεδομένων. Αυτό αρχικά μας επιβεβαιώνει στο παραπάνω συμπέρασμα. Επιπλέον αν αναλογιστούμε τα χαρακτηριστικά των δύο συνόλων δεδομένων, όπου το FilmTrust είναι πιο πυκνό όσον αφορά τις βαθμολογίες και τις κοινωνικές σχέσεις ενώ το CiaoDVD είναι αισθητά πιο αραιό, οδηγούμαστε στο συμπέρασμα πως μία τεχνική νευρωνικού δικτύου όπως ο GraphRec μπορεί να εξάγει περισσότερη χρήσιμη πληροφορία σε σχέση με τους υπόλοιπους αλγορίθμους.

5.2 Αποτελέσματα των Αλγορίθμων στο εμπλουτισμένο dataset και Σχολιασμός τους

Παρουσιάζουμε τα αποτελέσματα των αλγορίθμων στο εμπλουτισμένο FilmTrust datasets με τη χρήση την τεχνικής SimRank. Τα ποσοστά train / test είναι 60%-40% και 80%-20% αντίστοιχα.

Training	Metrics	Algorithms				
		Social_MF	**Social_RSTE**	**Social_Reg**	**TrustSVD**	**GraphRec**
FilmTrust (60%)	MAE	0.8597	0.8736	0.8609	0.8211	1.3591
	RMSE	1.1189	1.1401	1.1185	1.0623	1.7332
FilmTrust (80%)	MAE	0.8736	0.8712	0.8821	0.8197	1.2645
	RMSE	1.1385	1.1391	1.1295	1.0796	1.5863

Table 13. Αποτελέσματα του Αλγορίθμων στα εμπλουτισμένα δεδομένα του FilmTrust dataset

Όσον αφορά τους αλγορίθμους Social_MF – Social_RSTE – Social_Reg – TrustSVD παρατηρούμε πως η επίδοσή τους παρουσιάζει παρόμοια ή βελτιωμένη συμπεριφορά. Τόσο το RMSE όσο και το MAE στις περισσότερες περιπτώσεις έχει μειωθεί που σημαίνει ότι οι προτάσεις που παράγονται μετά την εισαγωγή των νέων κοινωνικών σχέσεων από τον αλγόριθμο SimRank είναι πιο σχετικές για τους χρήστες μας.

Ξανά ιδιαίτερο ενδιαφέρον παρουσιάζει ο αλγόριθμος GraphRec. Η απόδοσή του έχει μειωθεί σημαντικά με τη μετρική RMSE να ανέρχεται από 0.9683 σε 1.732 και 0.9129 σε 1.5863, ενώ η μετρική MAE φτάνει από 0.8201 σε 1.3591 και από 0.6985 σε 1.2645. Αυτό δείχνει πως ο αλγόριθμος GraphRec βασίζεται σε μεγάλο βαθμό στις κοινωνικές σχέσεις των χρηστών για να φτάσει στη παροχή προτάσεων σε αυτούς. Είναι σημαντικό να σημειωθεί ότι ο αλγόριθμος SimRank είναι ένας γενικός αλγόριθμος που καθορίζει την ομοιότητα οντοτήτων βασιζόμενος μόνο στη δομή του Γράφου και είναι σημαντικό για την εποικοδομητική του λειτουργία οι διασυνδέσεις στο Γράφο να είναι πολλές. Η πυκνότητα ωστόσο του FilmTrust όσον αφορά τις κοινωνικές σχέσεις μεταξύ των χρηστών είναι πολύ μικρή με ποσοστό 0.42%.

6.0 Μελλοντική Έρευνα

Έχοντας ολοκληρώσει την περιγραφή της ερευνητικής και πειραματικής διαδικασίας που ακολουθήσαμε θα μπορούσαμε να καταλήξουμε σε έναν επίλογο και κάποιες ιδέες για

μελλοντική έρευνα. Μία γενική κατεύθυνση που θα προτείναμε ανεπιφύλακτα είναι η επιλογή αρχιτεκτονικών που βασίζονται σε νευρωνικά δίκτυα. Μπορούμε να διακρίνουμε από τα αποτελέσματά μας πως οι δυνατότητες τέτοιων αρχιτεκτονικών είναι πολύ μεγάλες σε σχέση με πιο συμβατικές μεθόδους μηχανικής μάθησης.

Τα κοινωνικά δίκτυα είναι εκ φύσεως απολύτως αντιπροσωπευτικά με τη βοήθεια γράφων. Τα νευρωνικά γραφήματα μπορούν να ενσωματώσουν πληροφορίες σχετικά με τους κόμβους και την τοπολογική δομή του Γράφου πολύ αποδοτικά. Οι σχέσεις μεταξύ των χρηστών αλλά και οι αλληλεπιδράσεις τους με αντικείμενα μπορούν να αναπαρασταθούν τέλεια με Γράφους. Αυτό καθιστά τα νευρωνικά γραφήματα ιδανικά για την παροχή προτάσεων σε κοινωνικά δίκτυα. Έτσι οι μελλοντικές έρευνες που θα διεξαχθούν στον τομέα των προτάσεων σε κοινωνικά δίκτυα θα πρέπει να επικεντρωθούν στη χρήση νευρωνικών γραφημάτων (Graph Neural Network).

Μία άλλη πρόταση που θα μπορούσαμε να κάνουμε για μελλοντική έρευνα είναι η χρησιμοποίηση πολλαπλών τεχνικών για την επέκταση των δεδομένων. Στην πειραματική μας διαδικασία χρησιμοποιήσαμε μόνο τον SimRank ο οποίος είναι ένας αλγόριθμος που βασίζεται αποκλειστικά στη δομή του Γράφου προκειμένου να αποτιμήσει την ομοιότητα μεταξύ χρηστών. Είναι βέβαιο πως αυτό δεν είναι αρκετό για να οδηγηθούμε σε ένα μεγαλύτερο και γεμάτο σημαντική πληροφορία σύνολο δεδομένων. Τα σύνολα δεδομένων που χρησιμοποιήσαμε στην παρούσα διπλωματική εργασία παρείχαν μόνο αξιολογήσεις αντικειμένων από τους χρήστες και σχέσεις εμπιστοσύνης μεταξύ τους. Στον πραγματικό κόσμο τα δεδομένα είναι σαφώς πιο πλούσια. Για παράδειγμα πέρα από τα παραπάνω μπορεί ένα σύνολο δεδομένων να περιέχει και σημασιολογικές πληροφορίες για τα αντικείμενα ή δημογραφικά χαρακτηριστικά των χρηστών. Σε ένα τέτοιο σύνολο δεδομένων ο αλγόριθμος SimRank μπορεί να συνδυαστεί με άλλους αλγορίθμους και να επιφέρει πολύ καλύτερα αποτελέσματα όσον αφορά την επέκταση των δεδομένων.

Μία ακόμη πρόταση για μελλοντική έρευνα είναι να αναλυθεί περισσότερο ο διαχωρισμός των δεδομένων σε train / test sets προκειμένου οι προτάσεις να είναι πιο αντιπροσωπευτικές για τον κάθε χρήστη ξεχωριστά. Είναι σαφές πως αν χρησιμοποιήσουμε λιγότερο ποσοστό των δεδομένων για εκπαίδευση του μοντέλου μας οι προτάσεις θα είναι λιγότερο εξατομικευμένες καθώς θα συμβαδίζει περισσότερο με τη μέση άποψη των υπόλοιπων χρηστών. Μία σκέψη πάνω σε αυτό θα ήταν ταινίες που έχουν λάβει λίγες αξιολογήσεις να μην αφήνονται εκτός του train set για χάρη ταινιών που έχουν πολύ περισσότερες αξιολογήσεις.

Ένα εμπόδιο που παρουσιάστηκε στην παρούσα διπλωματική ήταν πως τα σύνολα δεδομένων είχαν δυαδικές τιμές ως προς την εμπιστοσύνη μεταξύ των χρηστών. Μία πρόταση που θα μπορούσαμε να σκεφτούμε σε αυτό το σημείο είναι να δημιουργηθεί ένας μηχανισμός αποτίμησης αυτής της εμπιστοσύνης μεταξύ των χρηστών σε τιμές σε όλο το εύρος (0,1]. Αυτό θα οδηγήσει σίγουρα σε καλύτερα αποτελέσματα καθώς οι προτάσεις για τους χρήστες θα επηρεάζονται από τους γειτονικούς χρήστες με βάση το ποσοστό της εμπιστοσύνης που έχουν μεταξύ τους.

Επιπλέον, θα είχε ενδιαφέρον να χρησιμοποιηθούν οι αλγόριθμοι που αναλύσαμε σε περισσότερα σύνολα δεδομένων προκειμένου να γίνει καλύτερη αξιολόγησή τους. Ελπίζω το

κείμενο της διπλωματικής και η πειραματική διαδικασία να αποτελέσει πηγή πληροφοριών για μελλοντικές εργασίες και βάση για να εξελιχθούν και άλλες κατασκευές. Τα Συστήματα Προτάσεων ήταν και θα είναι σίγουρα ένας τομές που θα απασχολήσει για πολύ καιρό ακόμα τον ερευνητικό και επιχειρησιακό κόσμο, και ειδικά πάνω σε κοινωνικά δίκτυα τα οποία έχουν πλέον επεκταθεί σε τεράστιο βαθμό.

Βιβλιογραφία

[1] Marsland, S. (2015). Machine Learning: An Algorithmic Perspective, second edition. CRC Press.

[2] McCarthy, John, and Edward A. Feigenbaum. "In memoriam: Arthur samuel: Pioneer in machine learning." *AI Magazine* 11.3 (1990): 10-10.

[3] Cohen, M. B., Elder, S., Musco, C., Musco, C., & Persu, M. (2015, June). Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing* (pp. 163-172).

[4] Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K., & Zhou, T. (2012). Recommender systems. *Physics reports*, *519*(1), 1-49.

[5] Park, Yoon-Joo, and Alexander Tuzhilin. "The long tail of recommender systems and how to leverage it." *Proceedings of the 2008 ACM conference on Recommender systems*. 2008.

[6] Zhang, Xuirui, and Hengshan Wang. "Study on recommender systems for business-to-business electronic commerce." *Communications of the IIMA* 5.4 (2005): 8.

[7] Adomavicius, Gediminas, et al. "Incorporating contextual information in recommender systems using a multidimensional approach." *ACM Transactions on Information systems (TOIS)* 23.1 (2005): 103-145.

[8] Lops, Pasquale, Marco De Gemmis, and Giovanni Semeraro. "Content-based recommender systems: State of the art and trends." *Recommender systems handbook* (2011): 73-105.

[9] Pu, Pearl, Li Chen, and Rong Hu. "Evaluating recommender systems from the user's perspective: survey of the state of the art." *User Modeling and User-Adapted Interaction* 22.4-5 (2012): 317-355.

[10] De Gemmis, Marco, et al. "Semantics-aware content-based recommender systems." *Recommender systems handbook*. Springer, Boston, MA, 2015. 119-159.

[11] Schafer, J. Ben, et al. "Collaborative filtering recommender systems." *The adaptive web*. Springer, Berlin, Heidelberg, 2007. 291-324.

[12] Do, Minh-Phung Thi, D. V. Nguyen, and Loc Nguyen. "Model-based approach for collaborative filtering." *6th International Conference on Information Technology for Education*. 2010.

[13] KOREN, Yehuda; BELL, Robert; VOLINSKY, Chris. Matrix factorization techniques for recommender systems. *Computer*, 2009, 42.8: 30-37.

[14] https://developers.google.com/machine-learning/recommendation/collaborative/basics

[15] Gope, Jyotirmoy, and Sanjay Kumar Jain. "A survey on solving cold start problem in recommender systems." *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2017.

[16] Ricci, Francesco, Lior Rokach, and Bracha Shapira. "Recommender systems: introduction and challenges." *Recommender systems handbook*. Springer, Boston, MA, 2015. 1-34.

[17] Woerndl, Wolfgang, Christian Schueller, and Rolf Wojtech. "A hybrid recommender system for context-aware recommendations of mobile applications." *2007 IEEE 23rd International Conference on Data Engineering Workshop*. IEEE, 2007.

[18] https://developers.google.com/machine-learning/recommendation/overview/candidate-generation

[19] Jain, Gourav, Tripti Mahara, and Kuldeep Narayan Tripathi. "A survey of similarity measures for collaborative filtering-based recommender system." *Soft computing: theories and applications* (2020): 343-352.

[20] Norouzi, Mohammad, David J. Fleet, and Russ R. Salakhutdinov. "Hamming distance metric learning." *Advances in neural information processing systems*. 2012.

[21] Schröder, Gunnar, Maik Thiele, and Wolfgang Lehner. "Setting goals and choosing metrics for recommender system evaluations." *UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA*. Vol. 23. 2011.

[22] KAMINSKAS, Marius; BRIDGE, Derek. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2016, 7.1: 1-42.

[23] https://www.youtube.com/watch?v=hDjEd43R7Ik

[24] Sun, Z., Han, L., Huang, W., Wang, X., Zeng, X., Wang, M., & Yan, H. (2015). Recommender systems based on social networks. *Journal of Systems and Software*, *99*, 109-119.

[25] Palau, Jordi, et al. "Collaboration analysis in recommender systems using social networks." *International Workshop on Cooperative Information Agents*. Springer, Berlin, Heidelberg, 2004.

[26] Pitsilis, G., & Knapskog, S. J. (2012). Social Trust as a solution to address sparsity-inherent problems of Recommender systems. *arXiv preprint arXiv:1208.1004*.

[27] Rawashdeh, Ahmad, and Anca L. Ralescu. "Similarity Measure for Social Networks-A Brief Survey." *Maics*. 2015.

[28] https://en.wikipedia.org/wiki/Link_prediction

[29] Jeh, Glen, and Jennifer Widom. "Simrank: a measure of structural-context similarity." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002.

[30] https://guoguibing.github.io/librec/datasets.html

[31] Gower, Stephen. "Netflix prize and SVD." *University of Puget Sound* (2014).

[32] Kim, Min-Gun, and Kyoung-jae Kim. "Recommender Systems using SVD with Social Network Information." *Journal of Intelligence and Information Systems* 22.4 (2016): 1-18.

[33] Jamali, Mohsen, and Martin Ester. "A matrix factorization technique with trust propagation for recommendation in social networks." *Proceedings of the fourth ACM conference on Recommender systems*. 2010.

[34] Ma, Hao, Irwin King, and Michael R. Lyu. "Learning to recommend with social trust ensemble." *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009.

[35] Ma, Hao, et al. "Recommender systems with social regularization." *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011.

[36] Guo, Guibing, Jie Zhang, and Neil Yorke-Smith. "Trustsvd: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. No. 1. 2015.

[37] Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., & Yin, D. (2019, May). Graph neural networks for social recommendation. In *The World Wide Web Conference* (pp. 417-426).

[38]https://github.com/WillKoehrsen/wikipedia-data-science/blob/master/notebooks/Book%20Recommendation%20System.ipynb

[39]https://medium.com/analytics-vidhya/matrix-factorization-made-easy-recommender-systems-7e4f50504477

[40] https://medium.com/@cfpinela/content-based-recommender-systems-a68c2aee2235

[41] https://medium.com/analytics-vidhya/collaborative-filtering-simple-practice-pairwise-correlations-c87576a7c65