# MovieLens
# Προτασιακό Σύστημα

Στέλλα Ζαραφίδου - *2037*
stellanz@csd.auth.gr

Χριστίνα Ισάκογλου - *2056*
christci@csd.auth.gr

**Περίληψη. Με την εξέλιξη των web 2.0 τεχνολογιών το στατικής φύσεως πρόσωπο του διαδικτύου έδωσε τη θέση του σε εφαρμογές και υπηρεσίες δυναμικού περιεχομένου, όπου ο χρήστης αποτελεί το κέντρο της προσοχής καθώς πληροφορίες ρέουν προς αλλά και πλέον από αυτόν. Τα προτασιακά συστήματα αποτελούν ένα τέτοιο παράδειγμα άντλησης δεδομένων από το χρήστη με σκοπό την εξαγωγή πληροφορίας χρήσιμης για αυτόν και σχετιζόμενης πάντα με το προφίλ του. Εφαρμόζοντας τεχνικές εξόρυξης γνώσης καθιστούν εφικτή τη δημιουργία προσωποποιημένων προτάσεων και την κατηγοριοποίηση χρηστών και αντικειμένων σε ομάδες. Το παρόν έγγραφο, εστιαζόμενο στο προτασιακό σύστημα του MovieLens, έχει σκοπό να αναδείξει: τον τρόπο που δομείται το συγκεκριμένο προτασιακό σύστημα, τα δεδομένα τα οποία αξιοποιεί προκειμένου να καταλήξει σε συμπεράσματα, το περιβάλλον διάδρασης αυτού με το χρήστη, τα χαρακτηριστικά και τις λειτουργίες του και τέλος τους αλγορίθμους που καθιστούν εφικτή την ύπαρξη όλων των παραπάνω.**

Keywords: recommender systems | collaborative filtering | hybrid algorithms | item-based filtering | content-based filtering | user-generated data | cosine similarity

## I. Εισαγωγή

Το παρόν έγγραφο, εστιαζόμενο στο προτασιακό σύστημα του MovieLens, έχει σκοπό να αναδείξει: τον τρόπο που δομείται το συγκεκριμένο προτασιακό σύστημα, τα δεδομένα τα οποία αξιοποιεί προκειμένου να καταλήξει σε συμπεράσματα, το περιβάλλον διάδρασης αυτού με το χρήστη, τα χαρακτηριστικά και τις λειτουργίες του και τέλος τους αλγορίθμους που καθιστούν εφικτή την ύπαρξη όλων των παραπάνω. Στο πρώτο μέρος γίνεται μια αρχική περιγραφή του συγκεκριμένου συστήματος, που αποτελεί και την πρώτη επαφή που έχει ένας χρήστης με αυτό. Στη συνέχεια δίνεται έμφαση στην εσωτερική δομή και διεργασίες του με σκοπό την κατανόηση της λειτουργίας του.

## II. Ανάλυση

### Τομέας

Αντικείμενο προτάσεων στο MovieLens είναι οι ταινίες. Η στατική φύση των δεδομένων συνεπάγεται τη μόνιμη αποθήκευση τους σε βάση δεδομένων, η οποία μέχρι τις μέρες μας έχει φτάσει τις 22200 εγγραφές και εμπλουτίζεται καθημερινά.

Στο στατικό αυτό περιεχόμενο προστίθενται τα μεταδεδομένα που παράγονται από τους χρήστες. Αποτελούμενα από ετικέτες και βαθμολογήσεις, έχουν τον κυρίαρχο ρόλο στην εφαρμογή των αλγορίθμων που θα αναλυθούν στη συνέχεια.

### Σκοπός

Ο πρωταρχικός σκοπός του MovieLens είναι η εξαγωγή προσωποποιημένων προτάσεων με βάση το προφίλ που διαμορφώνει κάθε χρήστης εκτελώντας διάφορες δραστηριότητες στο σιτε. Το σύστημα δηλαδή είναι σε θέση, έπειτα από διεργασίες αναλυτικής φύσεως (analytics, classification algorithms) που τρέχουν στο πίσω μη-ορατό μέρος του, να εντοπίσει ποιες είναι οι ταινίες με μεγαλύτερη πιθανότητα να ικανοποιήσουν το χρήστη και ποιες αυτές που θα τον αφήσουν αδιάφορο και θα ήταν καλό να αποφύγει.

Πέρα από τον εμφανή, και συναφή με τα υπόλοιπα προτασιακά συστήματα, στόχο, το MovieLens επιδιώκει τη δημιουργία αίσθησης μιας ισχυρής κοινότητας χρηστών. Συμμετέχοντας στο forum(Q&A) οι κινηματογραφόφιλοι από όλα τα γεωγραφικά μήκη και πλάτη μπορούν να ανταλλάξουν απόψεις για τις ταινίες που τους συστήνονται.

Το περιεχόμενο μιας τέτοιας συμμετοχής από την πλευρά των χρηστών, αν και δεν εντάσσεται στα μεταδεδομένα που μπορούν να αξιοποιηθούν στην κατηγοριοποίηση που υλοποιεί το σύστημα, αυξάνουν τη διαδραστικότητα και ενισχύουν το κλίμα συνεργατικότητας.

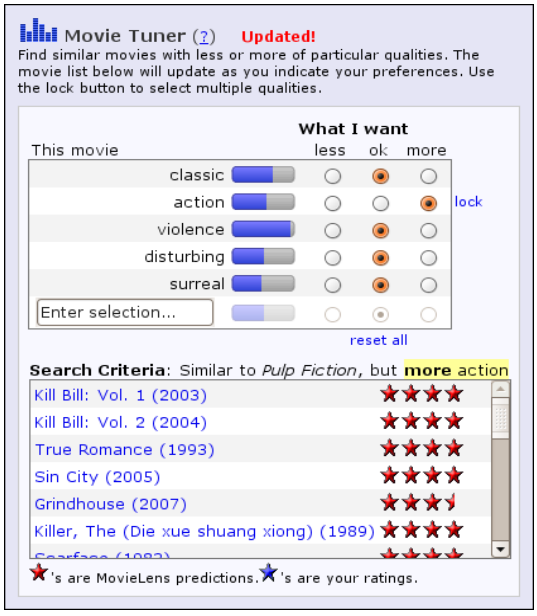
### Πλαίσιο Συστάσεων

Ιδιαίτερο γνώρισμα του πλαισίου συστάσεων του MovieLens αποτελεί η απουσία οποιασδήποτε αυθαίρετης παρουσίασης συστάσεων χωρίς την έγκριση και επιθυμία του χρήστη, όπως συμβαίνει σε παρεμφερή προτασιακά συστήματα, παραδείγματος χάριν στο YouTube.

Με αυτό τον τρόπο το σύστημα, έχοντας ως βασική του λειτουργία αυτήν της απλής αναζήτησης ταινιών στη βάση του, παροτρύνει την περαιτέρω ενεργητική συμμετοχή των χρηστών του (pull based model) περιμένοντας από αυτούς να επιλέξουν τον τρόπο και το είδος των συστάσεων που επιθυμούν να λάβουν. Οι επιλογές που έχουν για να δεχτούν προτάσεις αναλύονται στις εξής:

**Top Picks For You**. Τρόπος παρουσίασης των ταινιών που ταιριάζουν περισσότερο στον χαρακτήρα αυτών που έχει ήδη παρακολουθήσει και δείξει μεγαλύτερη προτίμηση ο εκάστοτε χρήστης, με βάση τη βαθμολογία (rating) που έχει αποδώσει σε κάθε μία απ' αυτές. Οι ταινίες αυτές παρουσιάζονται ταξινομημένες στον χρήστη με φθίνουσα σειρά και με βάση την πρόβλεψη (prediction) που έχει κάνει το σύστημα για αυτές, δηλαδή τη βαθμολογία που υπολογίζει πως θα της δώσει ο χρήστης αφότου την παρακολουθήσει.

**Movie Tuner**. Υπηρεσία στην οποία ο χρήστης έχει πρόσβαση μέσω της μεμονωμένης σελίδας κάθε ταινίας και τον εξυπηρετεί στο να βρει ταινίες που μοιάζουν αφενός με αυτήν, στη σελίδα της οποίας βρίσκεται, αλλά διαφέρει σε ορισμένα χαρακτηριστικά, με τρόπο που ρυθμίζει ο ίδιος. Παραδείγματα χαρακτηριστικών που ο χρήστης μπορεί να αυξομειώσει το βαθμό εμφάνισης τους, αποτελούν η βία, το δράμα, στοιχεία βασισμένα σε αληθινή ιστορία κλπ. Η υπερεσία αυτή βασίζεται εσωτερικά στην ύπαρξη ετικετών (tags), τις οποίες μπορεί ο χρήστης να αποδώσει σε ταινίες.




Σχήμα 1: Υπηρεσία Movie Tuner

**Use selected buddies**. Εναλλακτικός τρόπος αναζήτησης ταινιών που συνδυάζει τα αποτελέσματα της φιλτραρισμένης αναζήτησης του χρήστη με τις ταινίες που ταιριάζουν τόσο στους 'φίλους' που επέλεξε, όσο και στο προφίλ του ίδιου. Συμπεριλαμβάνοντας, δηλαδή, προφίλ άλλων χρηστών στην αναζήτηση, εμφανίζονται εκ νέου ταινίες, ταξινομημένες σε φθίνουσα σειρά προβλεπόμενης βαθμολόγησης, που ταιριάζουν στην τομή, αυτή τη φορά, των προφίλ των χρηστών που επιλέχτηκαν.

Καθοριστικό ρόλο παίζει η μορφή με την οποία οι συστάσεις παρουσιάζονται στον χρήστη, σε κάθε μία από τις περιπτώσεις που αναφέρθηκαν. Έτσι, προκύπτει το ερώτημα κατά πόσο ο τρόπος που εφαρμόζει το MovieLens είναι ικανοποιητικός και δεν αποσπά τους χρήστες από τον σκοπό του, αυτόν της αμερόληπτης εξαγωγής συμπερασμάτων σχετικά με τις προτιμήσεις τους και της μετέπειτα κατηγοριοποίησης τους με βάση αυτές.

Αφενός, γίνεται φανερό ότι ο εκάστοτε χρήστης έχει κάθε ελευθερία να παρέμβει και να διαμορφώσει διαφορετικά το προφίλ του αξιολογώντας περισσότερες ταινίες ή αλλάζοντας τις υπάρχουσες βαθμολογήσεις του (συμπεράσματα από αυτές του τις κινήσεις παρουσιάζονται από το MovieLens με συνοπτικό τρόπο κάνοντας χρήση διαγραμμάτων, προκειμένου ο καθένας να μπορεί να δει τη συνολική εικόνα που παρουσιάζει).

Αφετέρου όμως, ένα θέμα που είναι προς συζήτηση είναι το κατά πόσο η εμφάνιση των προβλεπόμενων βαθμολογιών για κάθε χρήστη επηρεάζει τη βαθμολόγηση που γίνεται στη συνέχεια, με αποτέλεσμα να αλλοιώνεται το αληθινό προφίλ του. Γι' αυτό το λόγο, το MovieLens έχει προνοήσει διαθέτοντας μία επιλογή που απενεργοποιεί την εμφάνιση των προβλέψεων αυτών, ούτως ώστε να βρίσκονται εκτός του πεδίου προσοχής του και να μην τον επηρεάζουν[1,4,9].

## Πηγές Ανάδρασης

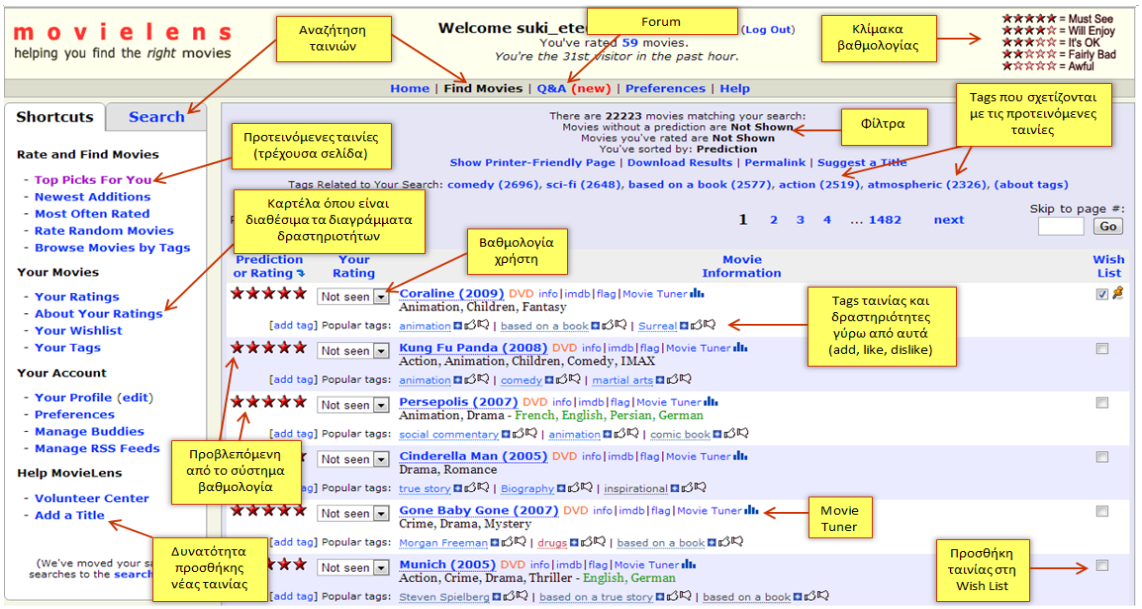
Το MovieLens απευθύνεται σε όλους τους χρήστες του Διαδικτύου, με κοινό γνώρισμα την αγάπη τους για τις ταινίες. Οι τεχνικές ή ειδικές γνώσεις σε κάποιον τομέα δεν κρίνονται απαραίτητες, ενώ επιδιώκεται οι χρήστες να είναι διαφορετικοί μεταξύ τους, ούτως ώστε όσο το δυνατόν περισσότερες και ποικίλες απόψεις και προτιμήσεις να λαμβάνονται υπόψη.

Το διαφορετικό υπόβαθρο των χρηστών, αφενός ενισχύει την αξιοπιστία του συστήματος και αφετέρου, καθώς όλο το πλαίσιο προτάσεων εντάσσεται σε αυτό της εξόρυξης πληροφορίας και γνώσης από ένα μεγάλο πλήθος (wisdom of the crowds), υποβοηθάει τις μεθόδους συλλογικής ευφυΐας (collective intelligence[1]) που εφαρμόζονται και τις καθιστά περισσότερο ακριβείς. Καθώς αυτό το οποίο ουσιαστικά συλλέγεται είναι απόψεις και προτιμήσεις, μπορεί εύκολα κάποιος να ισχυριστεί ότι όσο περισσότεροι άνθρωποι αξιολογούν μια ταινία ως καλή, καθένας με διαφορετικές προσλαμβάνουσες και ανεξάρτητος οποιασδήποτε επιρροής, τόσο μεγαλύτερη αξία και ισχύ έχει η άποψή τους.

## Επίπεδα Προσωποποίησης

Για να πετύχουν το στόχο τους, τα προτασιακά συστήματα λειτουργούν στα πλαίσια κάποιου ή κάποιων

[1]Με τον όρο συλλογική ευφυΐα αναφερόμαστε περισσότερο στις απόψεις του κοινού σε σχέση με ένα θέμα, παρά στην ύπαρξη μιας απάντησης σχετικής με αυτό που μπορεί να χαρακτηριστεί λανθασμένη ή σωστή.



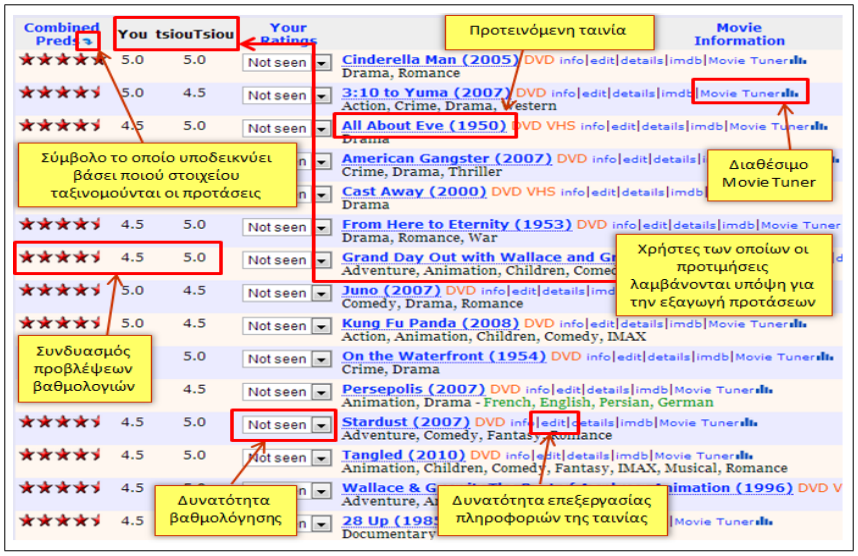

Σχήμα 2: Συστάσεις μέσω της επιλογής Top Picks For You

επιπέδων προσωποποίησης. Ο χαρακτήρας του MovieLens βασίζεται στη χρήση αλγορίθμων που κινούνται σε τρία επίπεδα προσωποποίησης.

Έτσι, στο σύστημα προωθούνται προτάσεις γενικές, που απευθύνονται σε όλους τους χρήστες, έχουν σκοπό την ενημέρωση κι έχουν στατικό χαρακτήρα (New Movies, New DVDs). Οι προτάσεις αυτές δεν είναι προσωποποιημένες μιας και όλοι οι χρήστες λαμβάνουν τις ίδιες. Ακόμη, παράγονται εφήμερες προτάσεις που βασίζονται σε επιτόπου δραστηριότητες του χρήστη, σύμφωνα με τις προτιμήσεις του γύρω από μια συγκεκριμένη ταινία (Movie Tuner). Οι προτάσεις αυτές παράγονται δυναμικά κι εμφανίζονται μόνο τη στιγμή της δραστηριότητας αυτής του χρήστη. Τέλος, το σύστημα παράγει προτάσεις που βασίζονται στα προσωπικά και μεγαλύτερης διάρκειας ενδιαφέροντα του χρήστη, λαμβάνοντας υπόψη το είδος των ταινιών που έχει βαθμολογήσει και την βαθμολογία που έχει αποδώσει σε αυτές (Top Picks For You).

## Ασφάλεια και Εμπιστοσύνη

Για να κάνει κανείς χρήση του προτασιακού συστήματος MovieLens, καλείται να δημιουργήσει πρώτα έναν λογαριασμό. Στο στάδιο αυτό το σύστημα δεν ζητάει από τον χρήστη να μοιραστεί προσωπικά του δεδομένα, αντιθέτως, τα τρία και μοναδικά στοιχεία που απαιτούνται είναι: μία έγκυρη διεύθυνση ηλεκτρονικού ταχυδρομείου, ένα ψευδώνυμο κι ένας κωδικός πρόσβασης. Έτσι διασφαλίζεται η ανωνυμία των χρηστών κατά την περιήγηση τους στο σύστημα, γεγονός που ενισχύει την εμπιστοσύνη τους προς αυτό. Επιπρόσθετα πεδία, κατά τη διάρκεια δημιουργίας λογαριασμού, μπορούν να συμ-




Σχήμα 3: Συνδυασμός προβλέψεων με την επιλογη Use selected buddies

πληρωθούν, αν και η παράλειψη αυτών δεν επηρεάζει τη λειτουργία του συστήματος και την ευστοχία των προτάσεων που θα εξάγονται για το συγκεκριμένο χρήστη.

Ακόμη ένα στοιχείο που ενισχύει την εμπιστοσύνη και την ασφάλεια που νιώθει ο χρήστης στην περιήγηση του στο σύστημα, είναι η δυνατότητα που του δίνεται να αποκρύψει από τους υπόλοιπους τις βαθμολογίες που έχει αποδώσει σε ταινίες, ούτως ώστε να μπορεί να διατηρήσει κρυφές τις προτιμήσεις του.

Πέρα όμως από τη σωστή αξιοποίηση των δεδομένων, για να κερδίσει την εμπιστοσύνη του χρήστη, το σύστημα θα πρέπει να δώσει την εντύπωση πως κάνει αυτό ακριβώς που ισχυρίζεται ότι κάνει.

Έτσι, από τη μία πλευρά θα πρέπει να γίνεται φανερό πως οι πληροφορίες που παρουσιάζονται εκπροσωπούν ισότιμα τις προτιμήσεις ενός μεγάλου ποσοστού των χρηστών και όχι αυτές των λίγων. Στο MovieLens, αυτό επιτυγχάνεται με την δυνατότητα που έχει κάθε χρήστης να επηρεάσει το μέσο όρο της βαθμολογίας (average rating) μιας ταινίας, να εμπλουτίσει το περιεχόμενο της βάσης των ταινιών με δικές του προτάσεις για ταινίες, σε περίπτωση που το κρίνει ελλιπές, να προσθέσει τις προσωπικές του ετικέτες (tags) σε ταινίες, οι οποίες μάλιστα δε χρειάζεται κατ' ανάγκη να υπόκεινται σε κάποιο επίσημο λεξιλόγιο (vocabulary), ενισχύοντας έτσι την ελευθερία που παρέχεται.

Ακόμη ένα στοιχείο αξιοπιστίας του συστήματος είναι το ότι σε κάθε ταινία παρουσιάζεται ο αριθμός των χρηστών που συμμετείχε στη διαμόρφωση του μέσου όρου βαθμολογίας της (rated by) και οι δραστηριότητες γύρω από τα ταγς της ταινίας (δημοφιλέστερα tags, likes και dislikes στα tags, tag cloud).

Από την άλλη πλευρά, δε θα πρέπει να δίνεται η εντύπωση στον χρήστη πως του γίνονται στοχευμένες προτάσεις που σκοπό έχουν όχι την πληροφόρηση αλλά τη χειραγώγησή του.

Πηγαίνοντας ένα βήμα παρά πέρα, συγκρίνοντας το MovieLens με κάποιο άλλο προτασιακό σύστημα, όπως για παράδειγμα το Amazon, παρατηρείται ότι δεν υπάρχει απόλυτη διαφάνεια αναφορικά με τις προτάσεις που γίνονται στο χρήστη. Για κάθε ταινία που αποτελεί σύσταση, δηλαδή, δε διευκρινίζεται ο ακριβής λόγος που οδήγησε στην ανάδειξη αυτής, ώστε να γίνεται σαφής ο τρόπος λειτουργίας του συστήματος. Αντισταθμίζοντας αυτό το χαρακτηριστικό, το MovieLens παρέχει τη δυνατότητα, όπως έχει ήδη αναφερθεί, ο χρήστης να μελετήσει τη γενική εικόνα που έχει διαμορφώσει, μέχρι τώρα, το σύστημα γι' αυτόν (καρτέλα About Your Ratings) και να εκτιμήσει την ποιότητα των προτάσεων που του γίνονται με βάση αυτήν.

## Περιβάλλοντα Διάδρασης

Η εκτέλεση ενός προτασιακού αλγορίθμου, προϋποθέτει την ύπαρξη δεδομένων που δέχεται σαν είσοδο και, ύστερα από την επεξεργασία τους, την παραγωγή άλλων δεδομένων στην έξοδο. Ένα προτασιακό σύστημα πρέπει να οπτικοποιήσει αυτά τα δεδομένα, ώστε ο χρήστης να μπορεί εύκολα να καταλάβει ποιες πληροφορίες θα πρέπει να δώσει στο σύστημα για να ενισχύσει την απόδοση του προτασιακού συστήματος και ποιες πληροφορίες αποτελούν τις εξόδους προς αυτόν.

Στο MovieLens, οι είσοδοι δίνονται με άμεση ανάδραση, μέσω των βαθμολογιών και των ετικετών (tags) που αποδίδουν οι χρήστες στις ταινίες που έχουν παρακολουθήσει. Η βαθμολογία οπτικοποιείται με τη βοήθεια μιας κλίμακας που αναπαρίσταται με 5 αστέρια (1-Awful, 2-Fairly Bad, 3-It's OK, 4-Will Enjoy, 5-Must See). Τα tags οπτικοποιούνται με τη βοήθεια ενός tag cloud στη σελίδα της κάθε ταινίας ή απλά αναφέρονται με απλό κείμενο τα δημοφιλέστερα tags της κάθε ταινίας όταν αυτές προτείνονται.

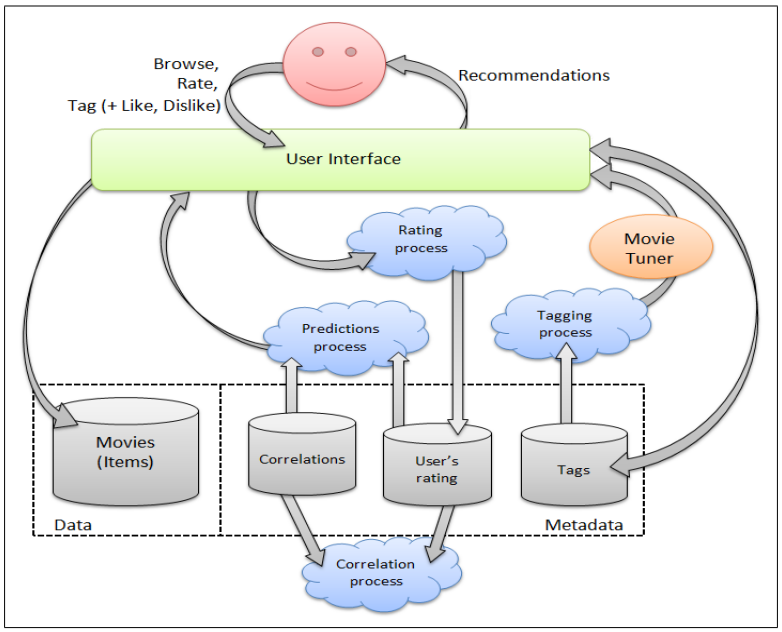
Λαμβάνοντας υπόψη και αυτά τα στοιχεία, το σύστημα παράγει, σαν εξόδους, προβλέψεις βαθμολόγησης και προτάσεις ταινιών. Στην πρώτη περίπτωση το σύστημα προσπαθεί να προβλέψει την βαθμολογία που θα έδινε ο χρήστης, δεδομένων των δραστηριοτήτων που έχει επιτελέσει στο παρελθόν. Στη δεύτερη περίπτωση, το σύστημα προτείνει στο χρήστη ταινίες που, κατά τους υπολογισμούς του, ταιριάζουν περισσότερο στο χαρακτήρα του και πιθανότατα να τον ενδιέφερε να τις παρακολουθήσει.

Μια ειδική περίπτωση προτάσεων, είναι αυτή που πραγματοποιείται λαμβάνοντας υπόψη τις σχέσεις μεταξύ συγκεκριμένων χρηστών (buddies). Αυτές οι προτάσεις παρουσιάζονται με προβλεπόμενη βαθμολογία, τον συνδυασμό των προβλέψεων που έχουν γίνει για τον κάθε χρήστη ξεχωριστά.

## Προτασιακοί Αλγόριθμοι

Το μοντέλο στο οποίο βασίζεται το αναφερόμενο σύστημα χαρακτηρίζεται υβριδικό, καθώς οι αλγόριθμοι που υλοποιούνται στο πίσω του μέρος προέρχονται τόσο από την πλευρά των αλγορίθμων συνεργατικής διήθησης (collaborative filtering) όσο και από την πλευρά των αλγορίθμων διήθησης με βάση το περιεχόμενο (content-based filtering).

Ο καθορισμός των αλγορίθμων που χρησιμοποιούνται υπάγεται στον καθορισμό των δεδομένων που το σύστημα αποθηκεύει, στον τρόπο αποθήκευσης τους (δομές δεδομένων) και στον τρόπο σύγκρισης αυτών, ώστε να εξαχθούν συμπεράσματα. Έτσι λοιπόν, έχουμε πως το MovieLens στηρίζεται κατά κύριο λόγο στη συνεργατική διήθηση (collaborative filtering), καθώς ο χρήστης για να λάβει συστάσεις και να αποφύγει το γνωστό πρόβλημα της 'ψυχρής εκκίνησης' (cold start problem), καλείται αρχικά να χτίσει το προφίλ του αξιολογώντας, στην κλίμακα 1 με 5, δεκαπέντε ταινίες. Με βάση τα προφίλ των χρηστών το σύστημα εξάγει ομάδες αντικειμένων (item-to-item model)[2], ανάλογα με τις αξιολογήσεις που τα συνδέουν. Από πρακτικής σκοπιάς η ουσία αυτού του μοντέλου υπάγεται στο ότι όταν ένα αντικείμενο ανήκει σε μια ομάδα και αξιολογηθεί θετικά από έναν χρήστη, οι συστάσεις που θα δρομολογηθούν για τον ίδιο θα ανήκουν στην ομάδα του αντικειμένου, για την οποία μόλις δήλωσε έμμεσα ενδιαφέρον.




Σχήμα 4: Εσωτερική δομή του εξυπηρέτη MovieLens

Ο ανωτέρω αλγόριθμος αναφέρεται στο πλαίσιο συστάσεων του "Top Picks For You" που παρέχει το MovieLens. Πέραν όμως αυτού, ο χρήστης έχει τη δυνατότητα, όπως ήδη αναφέρθηκε, να βρει ταινίες που θα τον ενδιαφέρουν αξιοποιώντας τις ετικέτες (tags) που έχουν αποδοθεί σε αυτές. Η λειτουργία του Movie Tuner, για παράδειγμα, στηρίζεται στους αλγορίθμους διήθησης βασιζόμενης στο περιεχόμενο (content-based filtering), καθώς οι ετικέτες αποτελούν τα διακριτά χαρακτηριστικά των αντικειμένων και η κατηγοριοποίηση τους επιτυγχάνεται με τη χρήση αυτών.

# III. Ερευνητική Προσέγγιση

## Αρχιτεκτονική συστήματος

Η λειτουργία του MovieLens στηρίζεται σε τεσσάρις βάσεις δεδομένων, από τις οποίες αντλεί τα δεδομένα πάνω στα οποία, στη συνέχεια, εκτελεί διεργασίες με σκοπό την εξαγωγή συστάσεων. Οι τέσσερις αυτές βάσεις αποτελούνται από μία κύρια, που περιέχει τα δεδομένα (data) του συστήματος, με το καθορισμένο περιεχόμενο που εμφανίζεται στο sitekαι όπως αυτά ορίζονται από τα αντικείμενα του τομέα του, και τρεις επιπλέον βάσεις για τη διαχείριση των μεταδεδομένων (metadata) που προκύπτουν απο τις ενέργειες του χρήστη.

Τα μεταδεδομένα περιλαμβάνουν τις ετικέτες και τις βαθμολογίες, που αποδίδει ο χρήστης στις ταινίες και τις συσχετίσεις που σχηματίζονται μεταξύ αυτών. Τα στοιχεία αυτά προκύπτουν από την αλληλεπίδραση του χρήστη με το περιβάλλον διάδρασης (user interface), με εξαίρεση τις συσχετίσεις, των οποίων ο υπολογισμός γίνεται σε μη-πραγματικό χρόνο με τρόπο που αναλύεται στη συνέχεια(offline).

Το στάδιο εκτέλεσης διεργασιών πάνω στο περιεχόμενο των παραπάνω βάσεων συνδέει τη φάση της εισόδου των μεταδεδομένων στο σύστημα με αυτό της δημιουργίας προβλέψεων και εξαγωγής συστάσεων από αυτό. Οι διεργασίες αναφέρονται αντίστοιχα ως: διεργασίες βαθμολογίας, κατά τις οποίες νέες βαθμολογίες καταγράφονται, διεργασίες ετικετών, όπου νέες ετικέτες προστίθενται, διεργασίες συσχετίσεων, όπου από τις βαθμολογίες που περιέχονται στην αντίστοιχη βάση εξάγονται συνδέσεις μεταξύ των αντικειμένων της κύριας βάσης και τέλος διεργασίες πρόβλεψης, που εκμεταλλεύονται την πληροφορία που περιέχουν οι προηγούμενες βάσεις, για να υπολογίσουν τις προβλέψεις του ενεργού χρήστη[2](active user), και ακολούθως τις συστάσεις που θα του γίνουν με βάση αυτές.

## Μοντέλο συνεργατικής διήθησης

Όπως τονίστηκε παραπάνω οι διεργασίες συσχετίσεων εξάγουν σχέσεις μεταξύ αντικειμένων κάνοντας χρήση των βαθμολογιών των χρηστών. Σ' αυτές λοιπόν τις συσχετίσεις, ή αλλιώς ομοιότητες, βασίζονται οι διεργασίες πρόβλεψης που παρέχουν ουσιαστικά τις προβλέψεις(και συστάσεις όταν ταξινομούνται αναλόγως) που μπορεί να λάβει ο χρήστης. Η μέθοδος αυτή αποτελεί υποκατηγορία της συνεργατικής διήθησης, αποκαλούμενη ως βασισμένη στο αντικείμενο[2],(item-based collaborative filtering ) και έρχεται σε αντιδιαστολή με αυτής της βασιζόμενης στο χρήστη(user-based collaborative filtering), διαφορετικής υποκατηγορίας του ίδιου μοντέλου.

Το κοινό σημείο των παραπάνω υποκατηγοριών είναι και το βασικό χαρακτηριστικό των αλγορίθμων συνεργατικής δίηθησης, ο στόχος δηλαδή της σύστησης νέων αντικείμενων ή πρόβλεψης της αρεσκείας αυτών από τον χρήστη, με βάση το ιστορικό των προτιμήσεων του αλλά και αυτό άλλων χρηστών, παρόμοιων από πλευράς

[2]Ενεργός χρήστης αποκαλειται ο χρήστης-στόχος για τον οποίο υπολογίζονται οι προβλέψεις

προτιμήσεων με αυτόν. Με αυτόν τον τρόπο, σε ένα σενάριο όπου υπάρχουν m χρήστες $U = u_1, u_2, ..., u_m$ και n αντικείμενα $I = i_1, i_2, ..., i_m$, ένας αλγόριθμος αυτού του μοντέλου θα επιστρέψει για κάθε χρήστη $u_i$ μία λίστα αντικειμένων $I_{u_i}$, για τα οποία ο χρήστης αυτό έχει αποδώσει κάποια βαθμολογία. Με αυτόν τον τρόπο δημιουργείται ένας πίνακας $m \times n$.

Σ' αυτό ακριβώς το σημείο ειπισέρχεται το στοιχείο των αλγορίθμων διήθησης βασιζόμενης σε αντικείμενο που τα διαφοροποιεί απο αυτούς της βασιζόμενης σε χρήστη. Στον πίνακα που σχηματίζεται, με βάση τα προηγούμενα, η προσέγγιση αυτών των αλγορίθμων είναι να κοιτάξουν στα αντικείμενα που ο χρήστης-στόχος έχει βαθμολογήσει, να βρουν πόσο όμοια είναι σε σχέση με το αντικείμενο-στόχος i και στη συνέχεια να επιλέξουν τα k περισσότερα όμοια αντικείμενα $i_1, i_2, ..., i_k$ αυτού. Ταυτόχρονα, υπολογίζονται οι ομοιότητες $s_{i1}, s_{i2}, ..., s_{ik}$ αυτών των αντικειμένων. Όταν οι υπολογισμοί αυτοί ολοκληρωθούν, η πρόβλεψη του αντικειμένου-στόχου είναι έτοιμη να παραχθεί ως ο ζυγισμένος μέσος όρος των βαθμολογιών του χρήστη-στόχου στα όμοια αντικείμενα αυτού.

Η βασικά ιδέα στο να μπορέσει να υπολογιστεί η ομοιότητα μεταξύ δύο αντικειμένων, έστω i και j, είναι να απομονωθούν οι χρήστες που έχουν βαθμολογήσει και τα δύο αυτά αντικείμενα και στη συνέχεια να εφαρμοστεί μια τεχνική υπολογισμού ομοιότητας προκειμένου να αποφασιστεί η ομοιότητα $s_{ij}$. Στο MovieLens η τεχνική αυτή είναι η λεγόμενη προσαρμοσμένη ομοιότητα συνημιτόνου(adjusted cosine similarity) και έχει το μειονέκτημα πως οι διαφορές στην κλίμακα βαθμολογίας που αποδίδουν οι χρήστες δε λαμβάνονται υπόψη. Παρ' όλα αυτά, ο αλγόριθμος αντισταθμίζει αυτό το μειόνεκτημα αφαιρόντας το μέσο όρο βαθμολογίας του αντίστοιχου χρήστη από το ζευγάρι υπολογισμού ομοιότητας. Έτσι, ο υπολογισμός της ομοιότητας εκφράζεται με τον εξής τύπο:

$$s_{ij} = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_u)^2}},$$

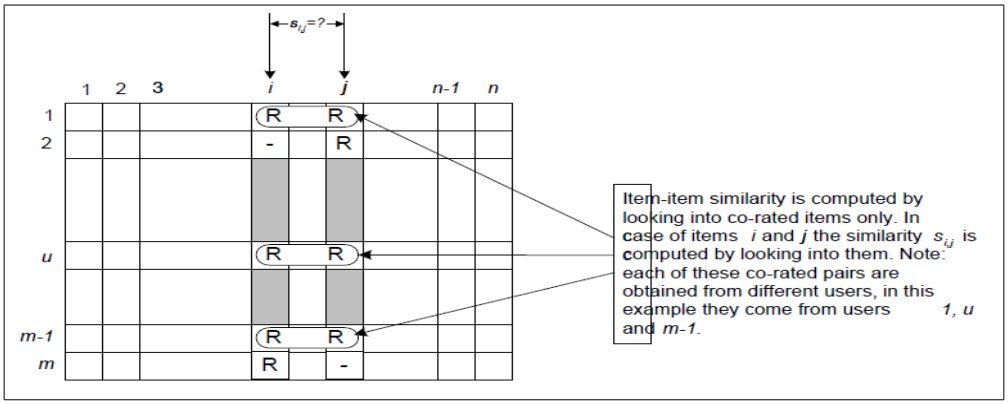
όπου $\bar{R}_u$ αντιστοιχεί στο μέσο όρο των βαθμολογιών του υοστού χρήστη.

## Ανεπάρκεια βαθμολογιών

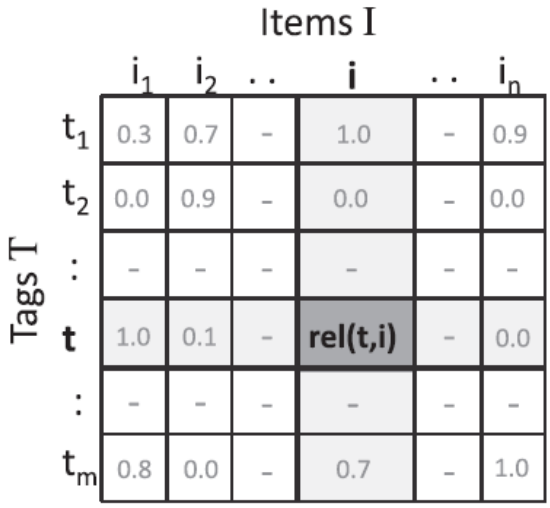
Με βάση τα όσα περιγράφηκαν, τίθεται το πρόβλημα που αντιμετωπίζουν συχνά οι αλγόριθμοι συνεργατικής διήθησης, αυτό της ανεπάρκειας βαθμολογιών. Λόγω του μεγάλου πλήθους δεδομένων που υπάρχει στη βάση δεδομένων του συστήματος, ένας χρήστης αλληλεπιδρά μόνο με ένα μικρό ποσοστό αυτού. Με αυτόν τον τρόπο γίνεται δυσκολότερη η εύρεση συσχετίσεων μεταξύ των αντικειμένων, μιας και η τομή των προτιμήσεων των χρηστών είναι μικρή. Κάτι τέτοιο πολλές φορές αντισταθμίζεται από το γεγονός ότι υπάρχουν ταινίες οι οποίες είναι πολύ δημοφιλείς και υπάρχει μεγάλη δραστηριότητα γύρω από αυτές με αποτέλεσμα να συντελεί στη δημιουργία συσχετίσεων μεταξύ ταινιών με λιγότερη απήχηση και των δημοφιλέστερων, ξεκινώντας μια αλυσίδα συσχετίσεων προς πολλές «κατευθύνσεις».

## Γονιδίωμα ετικετών και Movie Tuner

Το παραδοσιακό μοντέλο προσθήκης ετικετών (Traditional Tagging Model) χειρίζεται την προσθήκη αυτής της πληροφορίας ως δυαδική. Για παράδειγμα μπορεί κανείς να δηλώσει πως μια ταινία είναι βίαια, αλλά δεν μπορεί να ορίσει πόσο βίαια είναι. Επίσης, λόγο της δυαδικής απεικόνισης αυτής της πληροφορίας, δεν υπάρχει νόημα διαφορετικοί χρήστες να προσθέσουν πολλές φορές την ίδια ετικέτα σε μία ταινία. Ως εξέλιξη του μοντέλου αυτού, έρχεται το μοντέλο του «σακιδίου» (bag model) το οποίο επιτρέπει σε περισσότερους χρήστες πλέον να προσθέσουν την ίδια ετικέτα στην ίδια ταινία. Έτσι, από το πλήθος των ίδιων ετικετών που συλλέγει μια ταινία, θα μπορούσαν να εξάγονται συμπεράσματα για τη σχέση της ταινίας με αυτές[3]. Στην προηγούμενη περίπτωση όμως σε ένα πιθανό σενάριο όπου από τις τριάντα ετικέτες μιας ταινίας, οι οκτώ περιέχουν τη λέξη «βία» οι υπόλοιπες να μοιράζονται διαφορετικές ετικέτες, θα οδηγούσε στο συμπέρασμα πως η ταινία είναι βίαιη,χωρίς να γνωρίζουμε τον ακριβή βαθμό αυτού του χαρακτηρισμού. Το Γονιδίωμα Ετικετών (The Tag Genome)[7], αποτελεί ένα νέο μοντέλο που έρχεται να ενισχύσει τα παραδοσιακά μοντέλα που αναφέρθηκαν, ξεφεύγοντας από τη δυαδική αναπαράσταση της πληροφορίας των ετικετών και εισαγάγοντας μια συνεχή κλίμακα βαθμολόγησης, από το 0 έως το 1. Η κλιμακα




Σχήμα 5: Απομόνωση αντικειμένων που έχουν ψηφιστεί μαζί και υπολογισμός ομοιότητας




Σχήμα 6: Το γονιδίωμα ετικετών. Κάθε καταχώρηση του πίνακα δηλώνει τη συνάφεια της ετικέτας με την αντίστοιχη ταινία (στην κλίμακα από 0 μέχρι 1)

αυτή έχει ως σκοπό να καταγράψει πόσο ισχυρή είναι η συσχέτιση της ετικέτας με την ταινία στην οποία αποδίδεται. Η τιμή που παίρνει η ετικέτα σε αυτό το μοντέλο, καλείται συνάφεια (relevance) (0 = καθόλου συνάφεια, 1 = πλήρης συνάφεια). Ο δυναμικός χαρακτήρας των ετικετών ωθεί την κάθε τιμή της συνάφειας να μεταβάλλεται συνεχώς, με ανάλογο τρόπο. Το Movie Tuner του MovieLens λαμβάνει υπόψη τις τιμές της συνάφειας των ταινιών κι έτσι συγκρίνοντάς τις με αυτές άλλων ταινιών, με βάση τα κριτήρια που επέλεξε ο χρήστης, εντοπίζει ταινίες για να προτείνει εκ νέου και με τρόπο που παρουσιάζει μια δυναμικότητα σε σχέση με τους υπόλοιπους τρόπους προτάσεων που αναφέρθηκαν.

Όπως ένας οργανισμός ορίζεται από μια ακολουθία από γονίδια, έτσι κι ένα αντικείμενο, στο σύστημα που αναλύουμε, μπορεί να ορισθεί από τη σχέση του με μια σειρά από ετικέτες. Αν T το σύνολο των ετικετών και I το σύνολο των αντικειμένων, ποσοτικοποιούμε τη σχέση μεταξύ κάθε ετικέτας $t \in T$ και αντικειμένου $i \in I$, με τη συνάφεια (relevance) του t ως προς το i , και συμβολίζεται rel(t, i) .

Το γονιδίωμα ετικετών G είναι το σύνολο των τιμών των συναφειών για όλα τα ζευγάρια ετικέτας-ταινίας στο $T \times I$, που παριστάνεται σαν ένας πίνακας ετικετών-ταινιών όπως φαίνεται στο Σχήμα 6.

Ορίζουμε το G ως: $G_{t,i} = rel(t, i)$ Για κάθε $t \in T$ και $i \in I$. Ο όρος γονιδίωμα ετικετών χρησιμοποιείται, επίσης, για να περιγραφεί το διάνυσμα της συνάφειας μιας ετικέτας με ένα συγκεκριμένο αντικείμενο I, το οποίο δηλώνεται ως $G_i$ :

$$G_i = (rel(t_{1,i}), \ldots, rel(t_{m,i}))$$

για κάθε $t_j \in T$, όπου κάθε $G_i$ αντιπροσωπεύει μια στήλη του πίνακα G.

## Αναφορές


[1] Alexander Felfernig , Robin Burke, & Pearl Pu. 2011. *Preface to the special issue on user interfaces for recommender systems*. ACM 22, Issue 4-5 pp. 313-316 (2012).

[2] Badrul Sarwar, George Karypis, Joseph A. Konstan, & John Riedl. 2001. *Item-Based Collaborative Filtering Recommendation Algorithms*. ACM 1581133480/01/0005 (2001).

[3] Chen-Ting Huang, & Yin-Fu Huang. 2010. *Exploiting Social Tagging in a Web 2.0 Recommender System*. Internet Computing, IEEE (2010).

[4] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan,& John Riedl. 2003. *Is Seeing Believing? How Recommender Interfaces Affect Users" Opinions*. ACM 1-58113-630-7/03/0004 pp. 585-592 (2003).

[5] F. Maxwell Harper, Joseph A. Konstan, Xin Li, & Yan Chen. 2005. *User Motivations and Incentive Structures in an Online Recommender System*.

[6] Gawesh Jawaheer, Martin Szomszor, & Patty Kostkova. 2010. *Comparison of Implicit and Explicit Feedback from an Online Music Recommendation Service*. ACM 978-1-4503-0407-8/10/09 pp. 47-51 (2010).

[7] Jesse Vig, Shilad Sen, & John Riedl. 2012. *The tag genome: Encoding community knowledge to support novel interaction*. ACM Trans. Interact. Intell. Syst. 2, 3, Article 13 pp. 13-44 (2012).

[8] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, & John Riedl. 1999. *An Algorithmic Framework for Performing Collaborative Filtering*. ACM Inc. pp. 230-237 (1999).

[9] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, & John Riedl. 1997. *GroupLens: applying collaborative filtering to Usenet news*. Commun. ACM 40, 3 pp. 77-87 (1997).

[10] Michael Barnes. 2007. *User-Generated Metadata in Social Software: An Analysis of Findability in Content Tagging and Recommender Systems*. In partial fulfillment of the requirement for the degree of Master of Science.