ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΜΣ :«ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»
ΚΑΤΕΥΘΥΝΣΗ:«ΨΗΦΙΑΚΕΣ ΕΠΙΚΟΙΝΩΝΙΕΣ ΚΑΙ ΔΙΚΤΥΑ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ ΦΟΙΤΗΤΗ ΙΩΑΝΝΗ ΜΟΣΧΟΒΗ ΤΟΥ ΓΕΩΡΓΙΟΥ

ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ : ΜΕ1557

ΘΕΜΑ

ΠΡΟΣΕΓΓΙΣΕΙΣ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ ΓΙΑ ΣΥΣΤΑΣΕΙΣ ΒΑΣΕΙ ΤΕΧΝΙΚΩΝ ΣΥΝΕΡΓΑΤΙΚΟΥ ΦΙΛΤΡΑΡΙΣΜΑΤΟΣ

ΕΠΙΒΛΕΠΩΝ: ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ ΔΗΜΟΣΘΕΝΗΣ ΚΥΡΙΑΖΗΣ

ΠΕΙΡΑΙΑΣ 2018

Περιεχόμενα

Πίνακας Εικόνων

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου στον Επίκουρο Καθηγητή κ. Δρ. Δημοσθένη Κυριαζή για την δυνατότητα που μου έδωσε να πραγματοποιήσω την πτυχιακή μου εργασία.

Επιπλέον θα ήθελα να ευχαριστήσω όλους τους καθηγητές του ΠΜΣ «Ψηφιακές Επικοινωνίες και Δίκτυα» για τις πολύτιμες γνώσεις που μου προσέφεραν.

Ένα πρόσθετο ευχαριστώ στην Turi.com για την χορήγηση ακαδημαϊκής άδειας στα πλαίσια της εκπόνησης της εργασίας αυτής.

Τέλος, θέλω να εκφράσω ένα τεράστιο ευχαριστώ στην οικογένεια μου, για την στήριξη και την εμπιστοσύνη που μου έδειξαν όλα αυτά τα χρόνια των σπουδών μου.

Abstract

Με την άνθιση του διαδικτύου, όπου ο αριθμός των επιλογών βρίσκεται σε υπεραφθονία, προέκυψε η ανάγκη να φιλτράρουμε, να ιεραρχήσουμε και να παραδώσουμε αποδοτικά τις σχετικές πληροφορίες έτσι ώστε να ελαττώσουμε το πρόβλημα της υπερτροφοδότησης πληροφοριών, το οποίο αποτελεί πρόβλημα για αρκετούς χρήστες του Διαδικτύου. Οι μηχανές συστάσεων(recommendation engines) επιλύουν αυτό το πρόβλημα, αναζητώντας μέσα από ένα μεγάλο όγκο δυναμικά παραγόμενων πληροφοριών, και εφοδιάζοντας τους χρήστες με εξατομικευμένες υπηρεσίες και περιεχόμενο. Σκοπός αυτής της εργασίας είναι να μελετήσουμε τις διαφορετικές τεχνικές πρόβλεψης σε συστήματα συστάσεων στην πράξη προσομοιώνοντας την λειτουργία τους, τα προβλήματα και τις προκλήσεις που προκύπτουν κατά την λειτουργία τους, καθώς και να προταθούν τυχόν λύσεις σε αυτά.

Εισαγωγή

Με το ξεκίνημα της εποχής του Web 2.0 το διαδίκτυο γιγαντώθηκε και αναπτύχθηκε με εντυπωσιακή ταχύτητα. Αρκετές ευκαιρίες προέκυψαν, όπως η ανταλλαγή γνώσεων, πληροφοριών και απόψεων μεταξύ των χρηστών. Αυτές οι ευκαιρίες με την σειρά τους ευνόησαν την ανάπτυξη των μέσων κοινωνικής δικτύωσης(social networks) όπως το Facebook, το Twitter και το LinkedIn. Αποτέλεσμα αυτού, συντάκτες και συγγραφείς μοιράζονται τις δουλειές τους με εκατομμύρια αναγνώστες ανά την υφήλιο, ερασιτέχνες μουσικοί να γίνονται πιο γρήγορα από ποτέ διάσημοι, απλά επαναφορτώνοντας τα τραγούδια τους online.

Εικόνα 1: Το Web 2.0

Παράλληλα, ο επαγγελματικός κόσμος έχει επεκταθεί στο διαδίκτυο βρίσκοντας περισσότερους πελάτες και κατά συνέπεια αυξάνοντας τα κέρδη του. Η ποικιλία των καταστημάτων, των δημοπρατηρίων και των ανταλλακτηρίων επέκτεινε ακόμα περισσότερο το διαδίκτυο. Σήμερα, κάθε χρήστης μπορεί να προμηθευτεί σχεδόν οτιδήποτε, από οποιαδήποτε χώρα στον κόσμο. Στην πραγματικότητα πλέον το διαδίκτυο μοιάζει σαν ένας ατελείωτος κόσμος, με αποτέλεσμα ένα νέο ζήτημα να έρθει στο προσκήνιο. Ο όγκος των πληροφοριών και των αντικειμένων αυξήθηκε υπερβολικά οδηγώντας σε μια υπερφόρτωση πληροφοριών. Το τι ζητάει ο κάθε χρήστης έγινε ένα σημαντικό πρόβλημα. Αυτό επιλύθηκε εν μέρει με τις μηχανές αναζήτησης, ωστόσο αυτές απέτυχαν να δώσουν προσωποποιημένες πληροφορίες σαν αποτέλεσμα. Οι προγραμματιστές έδωσαν την λύση σε αυτό το πρόβλημα παρουσιάζοντας τα συστήματα συστάσεων(recommendation systems). Τα συστήματα συστάσεων αποτελούν εργαλεία για το φιλτράρισμα και την ταξινόμηση αντικειμένων και πληροφοριών. Ουσιαστικά, είναι απόψεις μιας κοινότητας χρηστών, με στόχο να βοηθήσουν μεμονωμένα άτομα αυτής της κοινότητας να αναγνωρίσουν πιο αποτελεσματικά το περιεχόμενο που τους ενδιαφέρει από ένα δυνητικά υπερβολικό αριθμό επιλογών. Υπάρχει τεράστια ποικιλία από αλγορίθμους και προσεγγίσεις που βοηθούν στην δημιουργία προσωποποιημένων συστάσεων. Δύο από αυτές έγιναν αρκετά γνωστές: το συνεργατικό φιλτράρισμα(collaborative filtering) και το φιλτράρισμα με βάση το περιεχόμενο(content-based filtering). Αυτές αποτελούν και τις βάσεις για τα περισσότερα μοντέρνα συστήματα συστάσεων.

Επιπλέον η εμφάνιση κινητών συσκευών με νέες τεχνολογίες, όπως τα GPS που κάνουν χρήση προτύπων 3G και 4G, δημιούργησαν νέες ευκαιρίες στην αγορά. Τα συστήματα συστάσεων αναμείχθηκαν και στην πρόοδο διαφόρων τομέων, όπως ο τουρισμός, η ασφάλεια και η διασκέδαση. Τα μοντέρνα συστήματα συστάσεων

βελτίωσαν την ακρίβεια των συστάσεων, χρησιμοποιώντας τεχνικές με βάση τα συμφραζόμενα(context-aware) ή με βάση την σημασιολογία(semantic). Σήμερα οι συστάσεις είναι πιο εξειδικευμένες και προσωποποιημένες. Όπως είναι λογικό όμως τα προβλήματα της συνδυασμένης χρήσης διαφορετικών τεχνολογιών και τεχνικών για την εξασφάλιση καλύτερων αποτελεσμάτων πάντα θα υπάρχουν και θα είναι ο λόγος για νέες προκλήσεις και έρευνες.

Problem Statement

Παρά την ύπαρξη και χρήση διαφορετικών τεχνικών προφανώς και κάποια προβλήματα είναι δυνατόν να προκύψουν. Αυτά είναι τα εξής:

Α) Το πρόβλημα της κρύας εκκίνησης(Cold-Start Problem)

Ο όρος «κρύα εκκίνηση» προέρχεται από τα αυτοκίνητα. Όταν η μηχανή είναι κρύα το αυτοκίνητο δεν δύναται να δουλέψει όπως προβλέπεται, αλλά μόλις η μηχανή φτάσει σε μια κατάλληλη θερμοκρασία η μηχανή δουλεύει άψογα. Για τις μηχανές συστάσεων, απλά σημαίνει ότι δεν έχουν ευδοκιμήσει οι κατάλληλες συνθήκες ώστε αυτές να λειτουργήσουν επαρκώς και να αποδώσουν τα βέλτιστα αποτελέσματα. Το πρόβλημα χωρίζεται σε δύο υποκατηγορίες: την κρύα εκκίνηση από την σκοπιά του προϊόντος και την κρύα εκκίνηση από τη σκοπιά του χρήστη. Στην πρώτη περίπτωση το προϊόν δεν έχει τον κατάλληλο αριθμό αξιολογήσεων ώστε να μπορεί να προταθεί αποτελεσματικά, ενώ στην δεύτερη το ιστορικό του χρήστη δεν επαρκεί ώστε οι συστάσεις να έχουν ακρίβεια.
Σε μερικά συστήματα συστάσεων η κρύα εκκίνηση από την σκοπιά του χρήστη επιλύεται με την χρήση ερευνών ή με την δημιουργία ενός προφίλ με κάποιες βασικές πληροφορίες.

Β) Το πρόβλημα της εμπιστοσύνης(Trust Issue)

Όπως είναι λογικό, η βαρύτητα των προφίλ ή των χρηστών με μικρό ιστορικό/ δραστηριότητα δεν μπορεί να είναι ίδια με αυτήν κάποιον άλλων με εμφανώς πλουσιότερο. Το πρόβλημα της εμπιστοσύνης εμπίπτει πάνω στις αξιολογήσεις ενός συγκεκριμένου χρήστη ή μιας ομάδας χρηστών.
Αυτό το πρόβλημα μπορεί να επιλυθεί με την διανομή προτεραιοτήτων μεταξύ των χρηστών.

Γ) Το πρόβλημα τις προσαρμοστικότητας(Scalability Issue)

Με την αύξηση του αριθμού των χρηστών και των αντικειμένων το σύστημα χρειάζεται ολοένα και περισσότερους πόρους για να επεξεργαστεί τις πληροφορίες και να διαμορφώσει τις ανάλογες συστάσεις. Η πλειονότητα των πόρων καταναλώνεται με στόχο να προσδιορίσει τους χρήστες με ανάλογες προτιμήσεις, καθώς και τα αγαθά με ανάλογες περιγραφές. Το πρόβλημα αυτό επιλύεται με τον συνδυασμό διαφορετικών τύπων φίλτρων καθώς και με την βελτίωση των συστημάτων σε φυσικό επίπεδο.

Δ) Το πρόβλημα της ισχνότητας(Sparsity Problem)

Στα περισσότερα διαδικτυακά καταστήματα που έχουν μεγάλο αριθμό από χρήστες και αντικείμενα, υπάρχουν σχεδόν πάντα χρήστες που έχουν αξιολογήσει μόνο ελάχιστα αντικείμενα. Με την χρήση συλλογικών(collaborative) και άλλων προσεγγίσεων τα συστήματα συστάσεων δημιουργούν εικονικές «γειτονιές» από χρήστες με βάση τα προφίλ αυτών. Εάν ένας χρήστης έχει αξιολογήσει ελάχιστα αντικείμενα είναι αρκετά δύσκολο να αποσαφηνιστεί το γούστο του και μπορεί αρκετά εύκολα να σχετιστεί με μια λάθος «γειτονιά». Η ισχνότητα ουσιαστικά αποτελεί το πρόβλημα της έλλειψης πληροφορίας.

Ε) Το πρόβλημα της ιδιωτικότητας(Privacy Problem)

Η ιδιωτικότητα αποτελεί το πιο σημαντικό πρόβλημα. Για να συγκεντρωθούν οι πιο ακριβείς και σωστές πληροφορίες, το σύστημα συστάσεων πρέπει να συλλέξει όσων των δυνατών περισσότερες πληροφορίες για τον χρήστη, συμπεριλαμβανομένων των δημογραφικών πληροφοριών του χρήστη καθώς και πληροφορίες σχετικές με την τοποθεσία του εκάστοτε χρήστη. Είναι τότε απολύτων φυσιολογικό να διεγείρονται ερωτήσεις σχετικά με την αξιοπιστία, την ασφάλεια και την εμπιστευτικότητα των εν λόγω πληροφοριών. Αρκετά διαδικτυακά καταστήματα προσφέρουν αποδοτική προστασία της ιδιωτικότητας των χρηστών εκμεταλλευόμενα εξειδικευμένους αλγορίθμους και προγράμματα.

Όλα τα προβλήματα που παρουσιάστηκαν, με εξαίρεση αυτό της ιδιωτικότητας, μπορούν να επιλυθούν με την χρήση των αλγορίθμων για την δημιουργία συστάσεων, είτε μεμονωμένα, είτε συνδυάζοντας τους. Η αυτοτελής παρουσίαση των αλγορίθμων και των προσεγγίσεων θα γίνει στο κεφάλαιο που ακολουθεί.

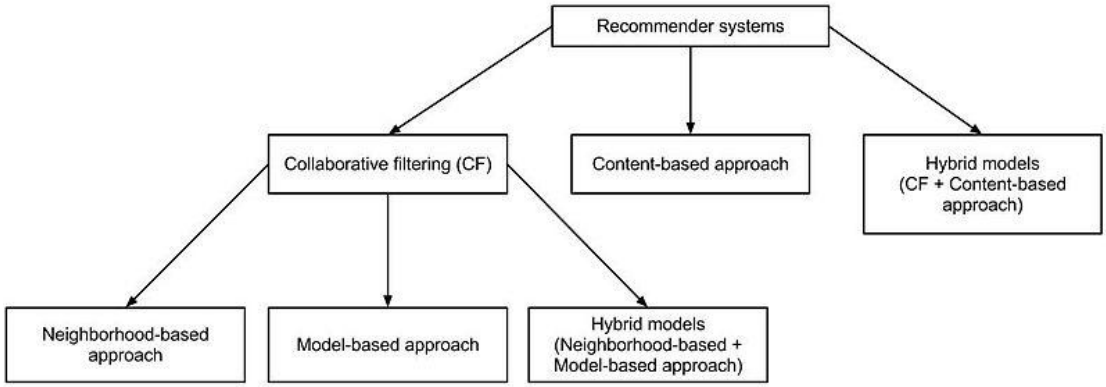
Όσων αφορά την ιδιωτικότητα, η ασφάλεια των δεδομένων έρχεται σε πρώτο λόγο. Επειδή η πλειονότητα των μηχανών συστάσεων χρησιμοποιούν τεχνολογία νέφους(cloud technology), έχουμε ένα συνδυασμό από κρυπτογράφηση, τεχνικών που εμποδίζουν την απώλεια δεδομένων και την προστασία της αξιοπιστίας, αυθεντικοποίηση καθώς και τεχνικών εξουσιοδότησης. Όταν οι επιχειρήσεις χρησιμοποιούν αλγόριθμους κρυπτογραφίας, είναι αρκετά σημαντικό αυτοί οι αλγόριθμοι αφενός να είναι γνωστοί και αφετέρου να είναι εγκεκριμένοι από το NIST. Είναι επίσης αρκετά χρήσιμο να υπάρχει μια επανεξέταση σε ετήσια βάση για τους αλγόριθμους και τα κλειδιά που χρησιμοποιούνται έτσι ώστε να επιβεβαιώνεται η αξιοπιστία της ασφάλειας.

Προσεγγίσεις

Υπάρχουν αρκετές προσεγγίσεις που να επιλύουν τα ζητήματα που αναφέρθηκαν στην προηγούμενη ενότητα. Αυτές οι προσεγγίσεις διαιρούνται σε δύο μεγάλες κατηγορίες: τις παραδοσιακές και τις μοντέρνες.

Οι παραδοσιακές προσεγγίσεις είναι οι πρώτες μέθοδοι που αναπτύχθηκαν για τις πρώτες μηχανές συστάσεων περί τα τέλη των 90s και τις αρχές των 00s. Με την εξάπλωση όμως του διαδικτύου και την πληθώρα από πληροφορίες σε αυτό, νέες μοντέρνες προσεγγίσεις βγήκαν στο προσκήνιο.

Παραδοσιακές Μέθοδοι



Εικόνα 2: Οι παραδοσιακές μέθοδοι συστάσεων

I) Φιλτράρισμα με βάση το περιεχόμενο(Content-based Filtering)

Τα συστήματα συστάσεων που είναι βασισμένα στο περιεχόμενο χρησιμοποιούν προφίλ χρηστών τα οποία δημιουργούνται κατά την εγγραφή τους. Το εκάστοτε προφίλ περιέχει πληροφορίες σχετικά με τον χρήστη και τα γούστα του. Τα γούστα του βασίζονται στις αξιολογήσεις που έχει δώσει ο χρήστης στο κάθε αντικείμενο. Ουσιαστικά, με την δημιουργία του προφίλ τα συστήματα συστάσεων διεξάγουν μια έρευνα, για να λάβουν κάποια αρχική γνώστη ώστε να αποφύγουν το πρόβλημα της ψυχρής εκκίνησης.

Τα συστήματα που φιλτράρουν με βάση το περιεχόμενο μοιράζονται από κοινού μέσα που χρησιμοποιούν τις περιγραφές των αντικειμένων που θα μοιραστούν, τα προφίλ με τις αρέσκειες του κάθε χρήστη και τις συγκρίσεις με τις αρέσκειες σε συνάρτηση με τα αντικείμενα ώστε να δοθούν οι τελικές συστάσεις. Τα αντικείμενα θα διαθέτουν περιγραφές, όπως το είδος μιας ταινίας ή η τοποθεσία ενός καταστήματος, βασιζόμενα στον τύπο του αντικειμένου που θα συσταθεί. Επιπλέον, τα αντικείμενα που έχουν μεγάλο βαθμό ομοιότητας με βάση τις δοθέντες προτιμήσεις του χρήστη θα συσταθούν.

Το προφίλ ενός χρήστη πρέπει να οικοδομηθεί έμμεσα από τις προτιμήσεις του χρήστη όσων αφορά τα αντικείμενα, αναζητώντας ομοιότητες μεταξύ

περιγραφών αντικειμένων που άρεσαν ή όχι στον χρήστη, βασισμένα στο ιστορικό ενεργειών του, ή άμεσα με την χρήση ερωτηματολογίων σχετικά με τις προτιμήσεις που αφορούν τις περιγραφές των αντικειμένων.

Ένα μοντέλο χρήστη πρέπει να δημιουργηθεί έμμεσα από μια αυτοματοποιημένη μέθοδο μάθησης, χρησιμοποιώντας τις περιγραφές των αντικειμένων σαν είσοδο ενός αλγόριθμου μάθησης υπό επίβλεψη και θα παράγει σαν αποτέλεσμα τις εντυπώσεις του χρήστη από το αντικείμενο αυτό.

Τα προφίλ συχνά απεικονίζονται σαν διανύσματα και οι περιγραφές ως βάρη. Οι επιλογές των χρηστών εξασφαλίζουν μια σχέση μεταξύ του χρήστη και των δοθέντων δεδομένων του. Κατά την έρευνα για συστήματα συστάσεων, μια επιλογή πρέπει να είναι κωδικοποιήσημη, από την πλευρά της μηχανής και να μεταφέρει χρήσιμες πληροφορίες, ώστε να δημιουργούνται οι κατάλληλες συστάσεις. Με αυτό τον τρόπο μπορούμε να δημιουργήσουμε μοναδικές επιλογές ή ακόμα και δυαδικές επιλογές(με την χρήση σύγκρισης). Αυτές οι επιλογές μπορούν να βοηθήσουν στην βελτίωση των συστάσεων όμως πάσχουν από συγκεκριμένα ελαττώματα, όπως αυτό της κάλυψης. Η κάλυψη μιας προτίμησης είναι άμεσα συσχετισμένη με την κάλυψη των ιδιοτήτων που εφαρμόστηκαν σε αυτή. Μια ιδιότητα που έχει υψηλή κάλυψη όταν εμφανίζεται σε πολλά αντικείμενα και περιορισμένη όταν εφαρμόζεται σε λίγα. Γι' αυτό και τα συστήματα συγκρίσεων που βασίζονται αποκλείστηκα με βάση το περιεχόμενο, συχνά απαιτούν μεγάλο όγκο από λεπτομέρειες με βάση τον χρήστη ώστε να αποδώσουν εύστοχες συστάσεις. Είναι ωστόσο δυνατό να επεκτείνουμε την κάλυψη μιας προτίμησης εισάγοντας την έννοια της ομοιότητας στις ιδιότητες. Μία επιλογή που αφορά μια συγκεκριμένη ιδιότητα μπορεί να εξαχθεί και σε όλες τις άλλες ιδιότητες που είναι παραπλήσιες. Αυτή η προσθήκη επεκτείνει την κάλυψη των επιλογών, αρκεί όμως να μην αντικρούει άλλες επιλογές.

Αρκετές προσεγγίσεις έχουν από κοινού ακολουθηθεί ώστε να κρίνουν την ομοιότητα μεταξύ των ιδιοτήτων. Παραδοσιακά το βάρος πέφτει στον αρμόδιο που σχεδιάζει την εκάστοτε εφαρμογή. Όσο και αν αυτή η προσέγγιση παραμένει δημοφιλής σε περιορισμένου εύρους εφαρμογές, οι μηχανές συστάσεων που δημιουργούνται για μεγάλου εύρους εφαρμογές αδυνατούν να την ακολουθήσουν διότι χρειάζονται πιο αυτοματοποιημένες διαδικασίες. Εναλλακτικά, μέτρα για την ομοιότητα μπορούν να χρησιμοποιηθούν, εκμεταλλευόμενα την πληθώρα πληροφοριών που υπάρχουν στο διαδίκτυο. Μια παρόμοια μετρική είναι το Normalised Google Distance (Cilibrasi & Vitanyi, 2007), το οποίο εξάγει ομοιότητες μεταξύ γραπτών όρων, χρησιμοποιόντας την συνύπαρξη αυτών σε ιστοσελίδες, όπως τις βρήκε η Google. Αυτή η μετρική ακόμα και αν είναι αποτελεσματική σε συγκεκριμένες περιπτώσεις δεν μπορεί να χρησιμοποιηθεί ασφαλώς διότι ουσιαστικά βασίζεται σε όλο το διαδίκτυο. Δυστυχώς είναι δύσκολο να βρει ολοκληρωτικές ομοιότητες σε μεγάλης κλίμακας βάσεις δεδομένων, λόγω των περιορισμών στην χρήση του API της Google. Για να αποφευχθούν τέτοιοι περιορισμοί, μετρικές ομοιότητας που αναλύουν άμεσα τις υπάρχουσες σύστασεις από την βάση δεδομένων του συστήματος χρησιμοποιούνται συχνά.

Άλλη μια προσέγγιση αφορά την χρήση ταξινομητή(classifier), σαν τον Naive Bayes, έχοντας σαν είσοδο την περιγραφή των αντικειμένων και σαν έξοδο τα γούστα ενός χρήστη για μία υποκατηγορία αντικειμένων. Ο ταξινομητής έχει ήδη χρησιμοποιήσει μια ομάδα από αντικείμενα που έχει επισκεφτεί ο χρήστης και στην συνέχεια είναι ικανός να προβλέψει αν ένα νέο αντικείμενο θα αρέσει ή όχι του χρήστη, με βάση την περιγραφή του αντικειμένου(Adomavicius & Tuzhilin, 2005). Με βάση αυτές τις μεθόδους μπορούν να συσταθούν και νέα αντικείμενα που ακόμα δεν έχουν καν αξιολογηθεί. Επιπλέον μπορούν να χειριστούν περιπτώσεις που οι

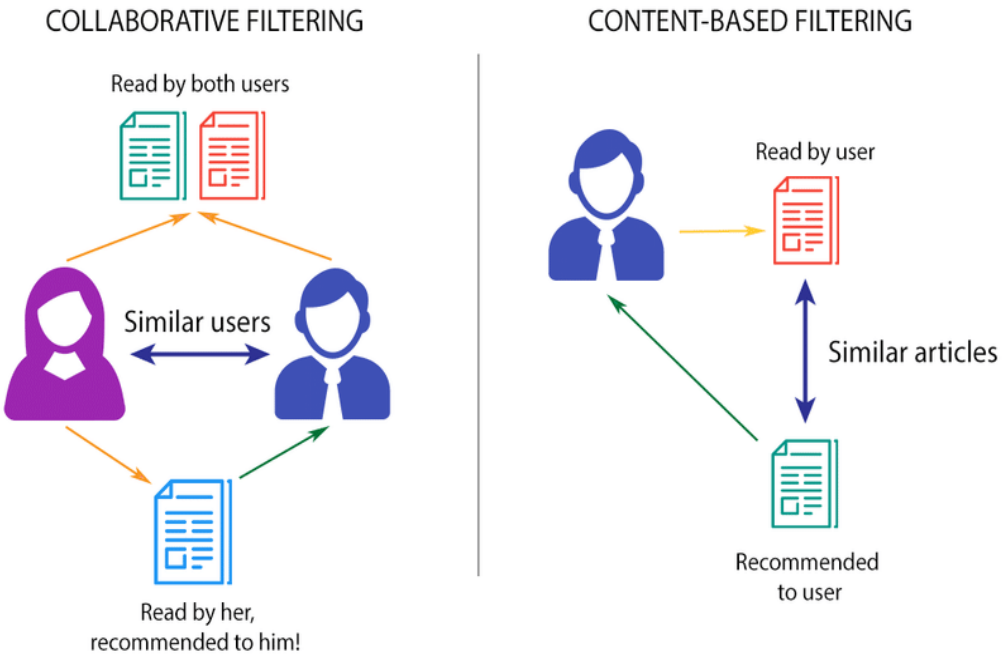
χρήστες δεν αναζήτησαν τα ίδια, αλλά παρόμοια αντικείμενα.

Ωστόσο, για να είσαι αποδοτικός όσων αφορά την προσέγγιση βασισμένη στην περιγραφή χρειάζεσαι πλούσιες περιγραφές και ολοκληρωμένες περιγραφές αντικειμένων καθώς και καλά στημένα προφίλ χρηστών. Αυτός είναι και ο κύριος περιορισμός τέτοιων συστημάτων. Μιας και είναι ελάχιστες οι σελίδες με καλογραμμένες περιγραφές, τέτοιες προσεγγίσεις έχουν εφαρμοστεί σε αντικείμενα που χαρακτηρίζονται από περιγραφή κειμένου και μπορούν να προσπελαστούν αυτόματα(Pazzani&Billsus,1997 και Mooney & Roy 1999).

Ένα άλλο πρόβλημα που μπορεί να προκύψει είναι από της υπερειδίκευσης(Zhang et al., 2002), και είναι ουσιαστικά η συνεχόμενη πρόταση αντικειμένων με παρόμοιο περιεχόμενο με αυτά που συστήνονται με αποτέλεσμα να έχουμε έλλειψη πρωτοτυπίας. Από την άλλη πλευρά όμως προβλήματα ιδιωτικότητας (Lam et al., 2006) όπως αυτό των χρηστών που δεν κοινοποιούν τις επιλογές τους, εμποδίζονται.

Η εκτίμηση ενός αντικειμένου από ένα χρήστη συχνά βασίζεται στις ολοένα και περισσότερες πληροφορίες που έχουν αποθηκευτεί στην περιγραφή του αντικειμένου. Ακόμα και υπερβολικά πλούσιες βάσεις δεδομένων μπορούν να παραλείψουν πληροφορίες που μπορεί να είναι κομβικές για τον χρήστη στην απόφασή του για το αν του αρέσει ή όχι το αντικείμενο.

Οι μέθοδοι συνεργατικού φιλτραρίσματος δεν απαιτούν από μεριάς τους καλοδομημένες περιγραφές. Αντιθέτως βασίζονται στις επιλογές του χρήστη σχετικά με τα αντικείμενα, οι οποίες μπορεί να αποδειχθούν πιο σημαντικές από μια απλή περιγραφή. Έχουν το πλεονέκτημα να αποδίδουν μια διαφορετική σκοπιά όσων αφορά το πόσο ενδιαφέροντα και ποιοτικά είναι τα αντικείμενα. Ο συνδυασμός αυτών των δύο σκοπιών ενθάρρυνε τον σχεδιασμό των υβριδικών συστημάτων.



Εικόνα 3 : Διαφορές συνεργατικού φιλτραρίσματος και φιλτραρίσματος με βάση το περιεχόμενο

ΙΙ) Συνεργατικό Φιλτράρισμα (Collaborative Filtering)

Το συνεργατικό φιλτράρισμα γρήγορα καθιερώθηκε ως μια από τις πιο ερευνημένες τεχνικές των συστημάτων συστάσεων, απ' όταν αυτή η μέθοδος πρωτοαναφέρθηκε και περιγράφηκε από τους Paul Resnick και Hal Varian το 1997. Η ιδέα του συνεργατικού φιλτραρίσματος αφορά στην εύρεση χρηστών με παρόμοια γούστα. Αν δύο χρήστες έχουν παρόμοιες ή ακόμα και ίδιες αξιολογήσεις σε ίδια αντικείμενα, τότε το γούστο τους είναι παρεμφερές. Τέτοιοι χρήστες ουσιαστικά σχηματίζουν μια εικονική ομάδα που την ονομάζουμε γειτονιά. Η αξιολόγηση ενός αντικειμένου για ένα δεδομένο χρήστη μπορούν να προβλεφτούν βασισμένες στις αξιολογήσεις που έχουν δοθεί από την γειτονιά του χρήστη ή από την γειτονιά του αντικειμένου. Μπορούμε να ξεχωρίσουμε τρεις προσεγγίσεις:

- Με βάση το χρήστη
- Με βάση το αντικείμενο
- Με βάση το μοντέλο

Έστω U μια ομάδα από N χρήστες, I μια ομάδα από M αντικείμενα και R μια ομάδα από αξιολογήσεις r από u χρήστες σε ι αντικείμενα($u \in U$ και $\iota \in I$) τότε το S είναι η ομάδα των αντικειμένων που έχουν αξιολογηθεί από ένα χρήστη από την ομάδα u. Ο στόχος του συνεργατικού φιλτραρίσματος είναι να μπορεί να προγνώσει την αξιολόγηση p ενός χρήστη a σε ένα αντικείμενο i. Ο χρήστης a υποθέτουμε ότι είναι ενεργός και έχει αξιολογήσει ήδη κάποια αντικείμενα άρα $S \neq \emptyset$. Το αντικείμενο που θα προγνωσθεί δεν είναι γνωστό στον χρήστη άρα $i \notin S$.

Προσεγγίσεις με βάση το χρήστη

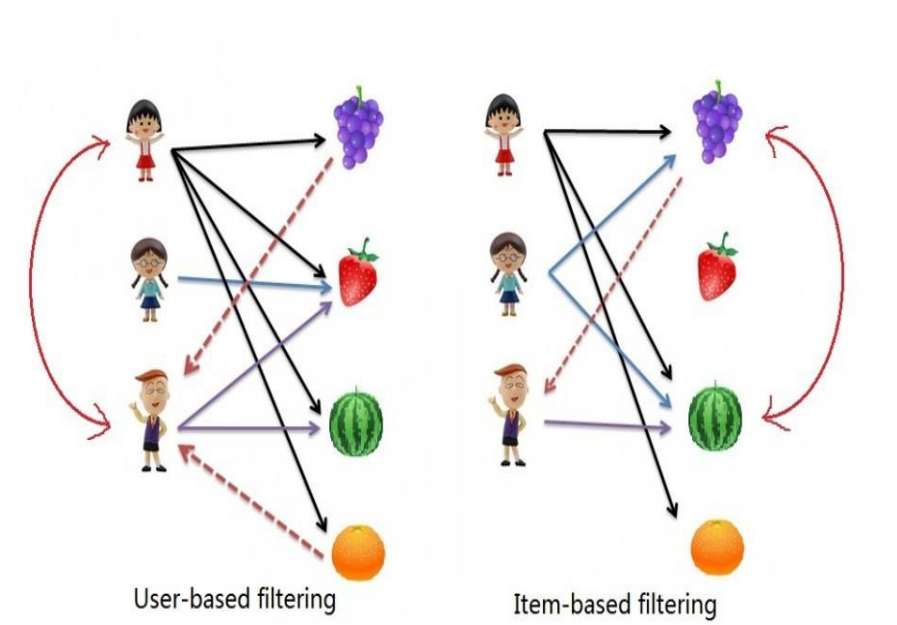
Για τις προσεγγίσεις με βάση τον χρήστη(Resnick et al., 1994, Shardanand & Maes,1995), η πρόβλεψη για την αξιολόγηση ενός αντικειμένου από ένα χρήστη, βασίζεται στις αξιολογήσεις των γειτονικών του αντικειμένων. Γι' αυτό το λόγο ένα μέτρο ομοιότητας μεταξύ των χρηστών πρέπει να οριστεί πριν γίνει η επιλογή των κοντινότερων γειτόνων του. Επιπλέον μια μέθοδος σύμπτυξης των αξιολογήσεων αυτών των γειτόνων ως προς το επίμαχο αντικείμενο πρέπει να επιλεχθεί.

Με αυτό τον τρόπο οι συστάσεις θα δίνονται βασισμένες στην αξιολόγηση των αντικειμένων από χρήστες που ανήκουν στην ίδια γειτονιά με τον χρήστη και με τους οποίους ο χρήστης έχει κοινά ενδιαφέρονται και αξιολογήσεις. Εάν το αντικείμενο έχει αξιολογηθεί θετικά από την γειτονιά του, τότε θα συσταθεί στον χρήστη. Βλέπουμε λοιπόν πως τα αντικείμενα που έχουν αξιολογηθεί πρότερα από των χρήστη παίζουν αρκετά σημαντικό ρόλο στην ενσωμάτωση του σε μία γειτονιά που θα χαρακτηρίζετε από κοινές αξιολογήσεις.

Προσεγγίσεις με βάση το αντικείμενο

Αυτή η προσέγγιση προτάθηκε από τους ερευνητές του πανεπιστημίου της Μινεσότα το 2001. Βασιζόμενοι στο γεγονός ότι το γούστο του χρήστη παραμένει σταθερό ή μεταβάλλεται ελάχιστα, παρόμοια αντικείμενα «χτίζουν γειτονιές»

βασιζόμενες στις αξιολογήσεις των χρηστών. Μετέπειτα το σύστημα δημιουργεί προτάσεις με βάση τα γειτονικά αντικείμενα των προτιμήσεων του κάθε χρήστη.



Εικόνα 4 : Διαφορές μεταξύ προσεγγίσεων με βάση το χρήστη ή το αντικείμενο

Προσεγγίσεις με βάση το μοντέλο

Η γενική ιδέα της προσέγγισης με βάση το μοντέλο είναι η παραγωγή ενός μοντέλου από δεδομένα εκτός σύνδεσης, έτσι ώστε να προβλέπει τις συνδεδεμένες αξιολογήσεις το ταχύτερο δυνατό. Τα πρώτα μοντέλα που προτάθηκαν αποτελούνται από ομαδοποιήσεις χρηστών και στην συνέχεια στην πρόβλεψη της αξιολόγησης ενός χρήστη σε ένα αντικείμενο με βάση την αξιολόγηση στο ίδιο αντικείμενο από τους χρήστες της ομάδας στην οποία ο χρήστης ανήκει.

Τα μοντέλα του Bayes είχαν επίσης προταθεί ώστε να ορίσουν της ανεξαρτησίες μεταξύ των αντικειμένων. Η ομαδοποίηση των αντικειμένων έχει μελετηθεί ενδελεχώς (παραδείγματος χάρη Ungar & Foster 1998, Conner & Herlocker 1999). Επιπλέον, μοντέλα με βάση εταιρικούς κανόνες έχουν μελετηθεί από (Sarwar et al., 2000) και (Lin et al., 2002).

Πιθανολογικοί αλγόριθμοι ομαδοποίησης έχουν επιπλέον χρησιμοποιηθεί, έτσι ώστε να δίνουν την δυνατότητα στους χρήστες να ανήκουν σε περισσότερες από μια ομάδες (Pennock et al. 2000, Kleinberg & Sandler 2004). Ακόμα και ιεράρχηση των ομάδων έχει προταθεί, έτσι ώστε αν μια ομάδα από χρήστες δεν έχει άποψη για ένα αντικείμενο, τότε το μοντέλο να προβαίνει στην δημιουργία μιας υπερομάδας(Kelleher & Bridge, 2003).

Σε τέτοιες προσεγγίσεις ο αριθμός των ομάδων έχει κομβική σημασία. Σε αρκετές περιπτώσεις διαφορετικός αριθμός ομάδων ελέγχεται, και εκείνη η οποία έχει τον μικρότερο ρυθμό σφαλμάτων, μετά από μια διασταυρωμένη επικύρωση, επιλέγεται. Οι ομάδες αντιπροσωπεύονται από έναν μέσο, και οι προβλεπόμενες αξιολογήσεις ενός χρήστη με βάση ένα αντικείμενο μπορούν να εξαχθούν κατευθείαν από τον κοντινότερο του μέσο. Εάν χρησιμοποιείται ομαδοποίηση τόσο στους χρήστες όσο και στα αντικείμενα, η προβλεπόμενη αξιολόγηση είναι η μέση αξιολόγηση. Τέτοιου είδους αλγόριθμοι χρειάζεται να εκτελεστούν αρκετές φορές με

τυχαίες αρχικές λύσεις, ώστε να αποφευχθούν τα τοπικά ελάχιστα.

Μετρικές Ομοιότητας

Η ομοιότητα που ορίζεται μεταξύ χρηστών ή αντικείμενων είναι κομβική στο συνεργατικό φιλτράρισμα. Η πρώτη που προτάθηκε στο (Resnick et al., 1994) είναι η συσχέτιση του Pearson και είναι ανάλογη με το συνημίτονο της απόκλισης από το μέσο. Μια εξίσου παραδοσιακή είναι και το απλό συνημίτονο.

Για εκείνες τις μετρικές ομοιότητας μόνο μια ομάδα από ιδιότητες που είναι όμοιες μεταξύ δύο διανυσμάτων χρησιμοποιούνται. Με αυτό τον τρόπο δυο διανύσματα αρκεί να έχουν ένα κοινό χαρακτηρισμό ή μια κοινή ιδιότητα για να χαρακτηριστούν όμοια. Τέτοιες βέβαια μετρικές έχουν και τα αρνητικά τους.

Το βασικό αρνητικό που μπορούμε να αναφέρουμε είναι το πρόβλημα της κρύας εκκίνησης, το οποίο προκύπτει όταν ένας νέος χρήστης δεν έχει αξιολογήσει κανένα αντικείμενο ή όταν ένα αντικείμενο δεν έχει αξιολογηθεί από κανένα χρήστη. Τότε το σύστημα απλά δεν μπορεί να δημιουργήσει συστάσεις λόγω της έλλειψης δεδομένων. Από την άλλη για το πρόβλημα του νέου χρήστη(Nguyen et al.,2007), εκμεταλλευόμαστε δημογραφικά στοιχεία του χρήστη σχετικά με την ηλικία, την τοποθεσία και την κατοικία του ώστε να βελτιώσουμε τις πρώτες συστάσεις που παρέχονται στο νέο χρήστη, χωρίς αυτός να έχει αξιολογήσει κάποιο αντικείμενο. Σημαντικό ρόλο στην επίλυση του νέου αντικειμένου και του νέου χρήστη μας παρέχουν οι υβριδικές προσεγγίσεις.

III) Υβριδικό Φιλτράρισμα

Όσων αφορά την περίπτωση του υβριδικού φιλτραρίσματος εκμεταλλευόμαστε τα θετικά τόσο του συνεργατικού φιλτραρίσματος όσο και αυτού με βάση το περιεχόμενο. Αυτές οι τεχνολογίες μπορούν να συνδυαστούν με ποικίλους τρόπους που χρησιμοποιούν τόσο τις εντυπώσεις των χρηστών από τα αντικείμενα όσο και τις επιλογές αυτών βασισμένες στις περιγραφές. Επιπλέον άλλες πηγές από δεδομένα όπως κοινωνικού και δημογραφικού περιεχομένου σχετικά με τους χρήστες μπορούν να χρησιμοποιηθούν.

Ο πιο άμεσος τρόπος να σχεδιάσεις ένα υβριδικό σύστημα συστάσεων είναι να εκτελέσεις ανεξάρτητα τόσο την συλλογική όσο και την με βάση το περιεχόμενο μέθοδο και στην συνέχεια να συνδυάσεις τα αποτελέσματα χρησιμοποιώντας ένα σύστημα ψήφων.

Στο (Balabanovic & Sholam,1997), ο συνδυασμός εκτελείται επιβάλλοντας στα αντικείμενα να βρίσκονται στον ίδιο χρόνο κοντά στο θεματικό προφίλ του χρήστη και στους υψηλά αξιολογημένους γείτονές τους. Στο (Pazzani, 1999), οι χρήστες συγκρίνονται με βάση το περιεχόμενο του προφίλ τους και στην συνέχεια χρησιμοποιούνται οι μετρικές που χρησιμοποιήθηκαν στο συνεργατικό φιλτράρισμα.

Στα (Polcicova et al., 2000, Melville et al., 2002) πίνακες με βάση τις αξιολογήσεις εμπλουτίζονται με προβλέψεις βασισμένες στο περιεχόμενο αφότου εκτελέσουμε και το συνεργατικό φιλτράρισμα. Στο (Vozalis & Margaritis, 2004), η ομοιότητα μεταξύ των αντικειμένων υπολογίζεται χρησιμοποιώντας το περιεχόμενο των περιγραφών και τα σχετικά διανύσματα αξιολογήσεων. Στην συνέχεια ένας αλγόριθμος συνεργατικού φιλτραρίσματος με βάση το αντικείμενο εκτελείται. Παράλληλα στην εργασία εκείνη οι συγγραφείς εξερεύνησαν πως αρκετοί αλγόριθμοι συνεργατικού φιλτραρίσματος μπορούν να ενισχυθούν με την χρήση των δημογραφικών στοιχείων των χρηστών. Δύο χρήστες μπορούν να θεωρηθούν παρόμοιοι όχι μόνο αν έχουν αξιολογήσει ανάλογα ίδια στοιχεία, αλλά και αν

ανήκουν και στην ίδια δημογραφική κατηγορία.

Στο (Han & Karypis, 2005) προτείνεται μια επέκταση τις λίστας των προτάσεων του συνεργατικού φιλτραρίσματος με αντικείμενα που το περιεχόμενο τους είναι παρόμοια με τα προτεινόμενα αντικείμενα. Βασιζόμενο στην ίδια ιδέα η ομοιότητα με βάση το περιεχόμενο που χρησιμοποιείται στο (Wang et al.,2006) ώστε να συγκρίνει τους χρήστες όχι μόνο με βάση τις εντυπώσεις που έχουν κοινοποιήσει για τα συγκεκριμένα αντικείμενα, αλλά και εξετάζοντας τις κοινές εντυπώσεις από αντικείμενα με παρόμοιο περιεχόμενο.

Ένα υβριδικό σύστημα μπορεί να σχεδιαστεί ακλουθώντας το φιλτράρισμα με βάση το περιεχόμενο και εμπλουτίζοντας τα δεδομένα που παράγονται με τη χρήση του συνεργατικού φιλτραρίσματος με στόχο να εμπλουτίσουν τις περιγραφές παρόμοιων αντικειμένων. Στον πυρήνα του, αυτό αποτελεί ένα σύστημα φιλτραρίσματος με βάση το περιεχόμενο που χρησιμοποιεί ομοιότητες στα χαρακτηριστικά, παρόμοιο με ένα σύστημα απορρόφησης προσωποποιημένων πληροφοριών, όπου τα αιτήματα είναι κενά. Τα αντικείμενα που προτείνονται με αυτό τον τρόπο λαμβάνουν υπόψη μόνο το τι αρέσει στον χρήστη, αγνοώντας όμως το τι δεν του αρέσει. Έτσι όσο πιο πολύ ένα σύστημα συνεργατικού φιλτραρίσματος θα προτείνει ένα αντικείμενο, τόσο πιο κοντά εκείνο το αντικείμενο θα είναι στο προφίλ του χρήστη. Αυτός ο τρόπος υβριδικού συστήματος αντιμετωπίζει τα κοινωνικά δεδομένα(που αντλήθηκαν από το συνεργατικό φιλτράρισμα) σαν μια ιδιότητα του αντικειμένου, όπως όλες οι άλλες. Ζυγίζοντας στην σημασία του κάθε χαρακτηριστικού, το σύστημα μπορεί να ποικίλει από το να είναι καθαρά βασισμένο στο περιεχόμενο με το να είναι καθαρά συνεργατικό.

Οι συνεργατικές τεχνικές φιλτραρίσματος εφαρμόζονται πιο συχνά από τις άλλες δυο μεθόδους και συχνά έχουν σαν αποτέλεσμα μια πιο καλύτερη απόδοση στις προβλέψεις. Από την άλλη το συνεργατικό φιλτράρισμα δείχνει καταλληλότερο όντας μια κομβική μέθοδος στα συστήματα πρόβλεψης, όταν το φιλτράρισμα με βάση το περιεχόμενο παρέχει λύσεις στους περιορισμούς του συνεργατικού φιλτραρίσματος, καθώς και ένα πιο φυσικό τρόπο διεπαφής με τους χρήστες. Μάλιστα οι χρήστες θα πρέπει να έχουν την δυνατότητα να ελέγξουν το σύστημα, χτίζοντας έτσι μια σχέση με νόημα και οδηγούμενοι στο ψυχολογικό κέρδος, με την ενίσχυση της εμπιστοσύνης στις προτάσεις. Αυτό όπως είναι φυσικό οδηγεί στην εξέταση της απόδοσης τους συστήματος συστάσεων.

Μοντέρνες Μέθοδοι

I) Προσεγγίσεις με βάση το context

Με την έννοια context αναφερόμαστε στις πληροφορίες για το περιβάλλον του χρήστη και για τις λεπτομέρειες της κατάστασης στην οποία εκείνος/η βρίσκεται. Τέτοιες λεπτομέρειες ενδέχεται να παίζουν σημαντικότερο ρόλο στις συστάσεις από τις ίδιες τις αξιολογήσεις που έχει δώσει ο χρήστης για τα αντικείμενα, μιας και αυτές οι αξιολογήσεις δεν περιέχουν λεπτομερείς πληροφορίες σχετικές με το κάτω υπό ποιες συνθήκες η αξιολόγηση αυτή δόθηκε από τον κάθε χρήστη. Μερικές συστάσεις μπορεί να είναι καλύτερο να δοθούν σε ένα χρήστη το βράδυ, κάποιες άλλες δεν ταιριάζουν με τις πρωινές του επιλογές καθώς και αυτός/ή έχει διαφορετικές επιλογές

όταν έξω έχει κρύο και τελείως διαφορετικές όταν έχει ζέστη. Τα συστήματα συστάσεων δίνουν προσοχή και εκμεταλλεύονται τέτοιες πληροφορίες στο να δώσουν συστάσεις ονομάζονται συστήματα με βάση το context(context-aware).

Την σημερινή εποχή που οι κινητές συσκευές αποτελούν αναπόσπαστο κομμάτι της καθημερινότητάς μας αυτού του είδους οι μηχανές συστάσεων είναι αναγκαίες. Με την χρήση του 4G και του GPS αλλά και άλλων τεχνολογιών, τα συστήματα συστάσεων μπορούν να αποσπάσουν ταχύτατα πληροφορίες για την τοποθεσία, αλλά και για τον ίδιο το χρήστη.

Σε αντίθεση με το είδος των πληροφοριών που αποθηκεύονται στα προφίλ, οι αλλαγές στο context γίνονται δυναμικά και σπάνια αποθηκεύονται μόνιμα, μιας και μπορεί να χάσουν την βαρύτητά τους με την πάροδο του χρόνου. Τα συστήματα με βάση το context, κυρίως λόγω αυτής της περιοδικής ανανέωσης των πληροφοριών, απέκτησαν αρκετή προβολή, καθώς αύξησαν εντυπωσιακά την ποιότητα των προτάσεών τους και αυτές οι προσεγγίσεις έγιναν ακόμα πιο εξειδικευμένες σε συγκεκριμένους τομείς.

Οι Gediminas Adomavicius και Alexander Tuzhilin παρουσίασαν τρία διαφορετικά παραδείγματα αλγορίθμων με την χρήση πληροφορίας βασισμένης στο context στην διαμόρφωση προτάσεων: το contextual pre-filtering, post-filtering και modeling. Στο contextual pre-filtering αντλείται context από επιλεγμένα δεδομένα, ώστε να τα χρησιμοποιήσουμε σε μελλοντικές προβλέψεις, ενώ στο post-filtering η άντληση λαμβάνει χώρα μετά από τις προβλέψεις. Τέλος στο τελικό παράδειγμα η πληροφορία από το context χρησιμοποιείται μόνο στην προσέγγιση της πρόβλεψης.

Από την άλλη οι Baltrunas και Ricci πρότειναν μια προσέγγιση με αξιολόγηση του context των αντικειμένων, βασισμένη στον διαχωρισμό των αντικειμένων. Στην αρχή αυτή η προσέγγιση αναζητά αντικείμενα που έχουν σαφής διαφορές στην αξιολόγηση. Κάθε αντικείμενο που μας επιστρέφεται λόγο ενός χαρακτηριστικού από το context, με βάση το οποίο οι αξιολογήσεις αλλάζουν. Για κάθε ένα τέτοιο δημιουργεί δύο νέα αντικείμενα αντί για το ένα παλιό που έχει, με σκοπό να ξεχωρίσει το χαρακτηριστικό του context από το αντικείμενο. Ουσιαστικά αυτή η προσέγγιση αποτελεί μια επέκταση του συνεργατικού φιλτραρίσματος.

Η ίδια αρχή χρησιμοποιείται στην τεχνική «μικρο-προφίλ» των Baltrunas και Amatriain. Αυτή η προσέγγιση που εξαρτάται από τον χρόνο, διαχωρίζει τα προφίλ των χρηστών βασισμένη στις αλλαγές των προτιμήσεων σε διαφορετικές χρονικές περιόδους.

Υπάρχουν αρκετοί τρόποι αναπαράστασης της πληροφορίες από context και της σχέσης μεταξύ αντικειμένων και χρηστών. Ένας από αυτούς τους τρόπου είναι ένας contextual γράφος. Εκτός από την αναπαράσταση, αν χρησιμοποιήσουμε αλγορίθμους βασισμένος σε γράφους, μπορούμε να βελτιώσουμε τις προβλέψεις μας. Σε contextual γράφους αντικείμενα, χρήστες και contextual αντικείμενα αποτελούν κόμβοι ενώ οι αξιολογήσεις καθώς και οι παράγοντες ομοιότητας αναπαρίστανται με την χρήση ακμών. Ο Toine Bogers χρησιμοποίησε contextual γράφους στον αλγόριθμο του, τον ContextWalk, που βασίζεται στην σύσταση ταινιών πάνω σε τυχαίες διαδρομές μιας αλυσίδας Markov. Παράλληλα οι τα συστήματα συστάσεων με βάση το context έγιναν με την σειρά τους αναπόσπαστο κομμάτι της τουριστικής βιομηχανίας.

Ένα από τα σημαντικότερα προβλήματα των συστημάτων συστάσεων που βασίζονται στο context είναι η άντληση context πληροφορίας. Η πληροφορία μπορεί να λαμβάνετε λεπτομερώς με τον χρήστη να συμπληρώνει κάποια φόρμα ή κάποια έρευνα. Βέβαια είναι καλύτερο να λαμβάνουμε context πληροφορίες χωρίς να δημιουργήσουμε ένα πολύπλοκο σύστημα αξιολογήσεων και κριτικών. Εναλλακτικός

τρόπος να αντλήσουμε πληροφορίες είναι η χρήση της τοποθεσίας μέσω GPS, ή η χρήση χρονοσφραγίδας σε μια συναλλαγή. Παράλληλα ένας άλλος εναλλακτικός τρόπος είναι η άντληση πληροφοριών αναλύοντας την συμπεριφορά των χρηστών ή χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων.

ΙΙ) Προσεγγίσεις με βάση τη σημασιολογία

Οι περισσότερες περιγραφές από αντικείμενα που παρουσιάζονται στους χρήστες είτε από τα συστήματα συστάσεων είτε από το υπόλοιπο διαδίκτυο βρίσκονται σε μορφή κειμένου. Με την χρήση ετικέτας ή λέξεων-κλειδιών τα οποία δεν έχουν κανένα νόημα με βάση τη σημασιολογία η ακρίβεια των συστάσεων δεν βελτιώνεται σε όλες τις περιπτώσεις, μιας και αρκετές λέξεις-κλειδιά μπορεί να αντικατασταθούν με ομώνυμές τους. Με αυτό τον τρόπο γίνεται σαφές το πόσο σημαντική είναι η κατανόηση και η δόμηση του κειμένου παίζουν αρκετά σημαντικό ρόλο στην αποτελεσματικότητα των συστάσεων. Οι παραδοσιακές τεχνικές εξόρυξης κειμένου που βασίζονται σε λεκτικές και συντακτικές αναλύσεις δείχνουν περιγραφές που μπορεί να τις κατανοήσει ένας χρήστης και όχι ένας υπολογιστής ή ένα σύστημα συστάσεων. Αυτός είναι ο λόγος που δημιουργήθηκαν νέες τεχνικές εξόρυξης κειμένου που βασίστηκαν στην σημασιολογική ανάλυση. Τα συστήματα συστάσεων που λειτουργούν έχοντας ως βάση αυτές τις τεχνικές ονομάζονται συστήματα συστάσεων με βάση τη σημασιολογία.

Η επίδοση τέτοιων συστημάτων βασίζεται στη γνωσιακή βάση που ορίζεται από ένα πρότυπο διάγραμμα(όπως η ταξινομία ή η οντολογία).

Ο ορισμός της Wikipedia μας αναφέρει ότι η Ταξινομία είναι η επιστήμη της ταξινόμησης. Η ταξινομία παίζει αρκετά σημαντικό ρόλο στην σημασιολογική ανάλυση. Η ταξινόμηση των αντικειμένων και των χρηστών που αφορά στον χώρο και στις ομάδες που εκείνοι/α ανήκουν προσδίδει περισσότερη αποδοτικότητα στα συστήματα συστάσεων.

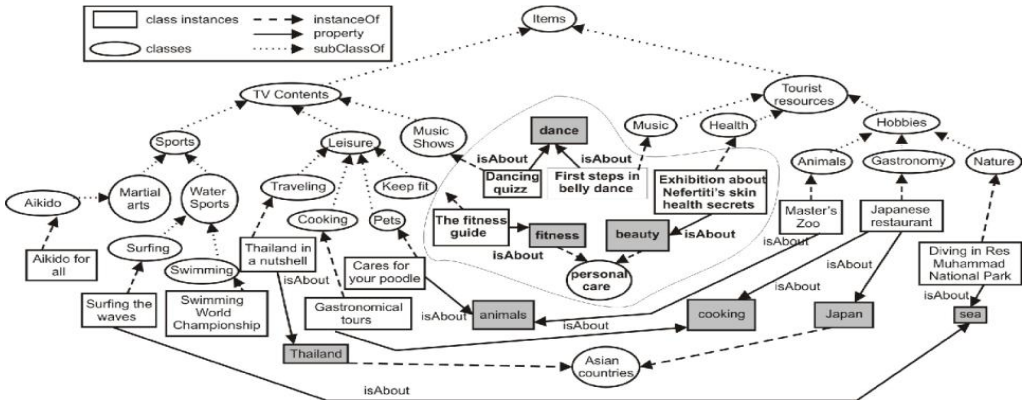
Ωστόσο, σε μερικές περιπτώσεις οι ταξινομίες δεν είναι ο καλύτερος τρόπος για να ταξινομήσουμε και να διαχωρίσουμε αντικείμενα, μιας και η εφαρμογή της ταξινομίας μπορεί να αποδειχθεί ιδιαίτερα πολύπλοκη και ακριβή, αν όχι ανέφικτη. Οι διαδικτυακές υπηρεσίες για κοινή χρήση φωτογραφιών είναι ένα καλό παράδειγμα δύσκολης δημιουργίας συστήματος με βάση την ταξινομία. Μια από τις καλύτερες λύσεις σε αυτό το πρόβλημα είναι οι folksonomies (πάντρεμα του όρου folk που σημαίνει συγγενής, με τον όρο ταξινομία). Η folksonomy είναι η προσέγγιση ταξινόμησης αντικειμένων, όπου οι χρήστες προσθέτουν ετικέτες στα αντικείμενα και με την σειρά τους οι πιο συχνές ή δημοφιλείς ετικέτες θα δημιουργήσουν κατηγορίες από αντικείμενα.

Οι ετικέτες στις folksonomies και τις ταξινομίες είναι συγγενικές μεταξύ τους χρησιμοποιώντας διάφορες οντολογίες. Ο Swartout όρισε την οντολογία σαν μία ιεραρχικά δομημένη ομάδα από όρους που περιγράφουν ένα χώρο και μπορούν να χρησιμοποιηθούν ως ο σκελετός για την δημιουργία μιας γνωσιακής βάσης. Ο Gruber είπε πως η οντολογία ορίζει του βασικούς όρους και τις σχέσεις που αποτελούν ένα λεξικό μιας θεματικής ενότητας, τους κανόνες για τον συνδυασμό των όρων αυτών μεταξύ τους και τις σχέσεις μεταξύ αυτών για να ορίσουν επεκτάσεις σε αυτό το λεξικό. Με την βοήθεια λοιπόν της οντολογίας τα συστήματα συστάσεων μπορούν να κατανοήσουν πως μερικοί όροι(αντικείμενα, χρήστες, ταμπέλες) σχετίζονται μεταξύ τους.

Ως γλώσσα περιγραφής των οντολογιών οι προγραμματιστές χρησιμοποιούν την Web Ontology Language(OWL). Συγκεκριμένα από το 2009 χρησιμοποιούν την OWL 2.

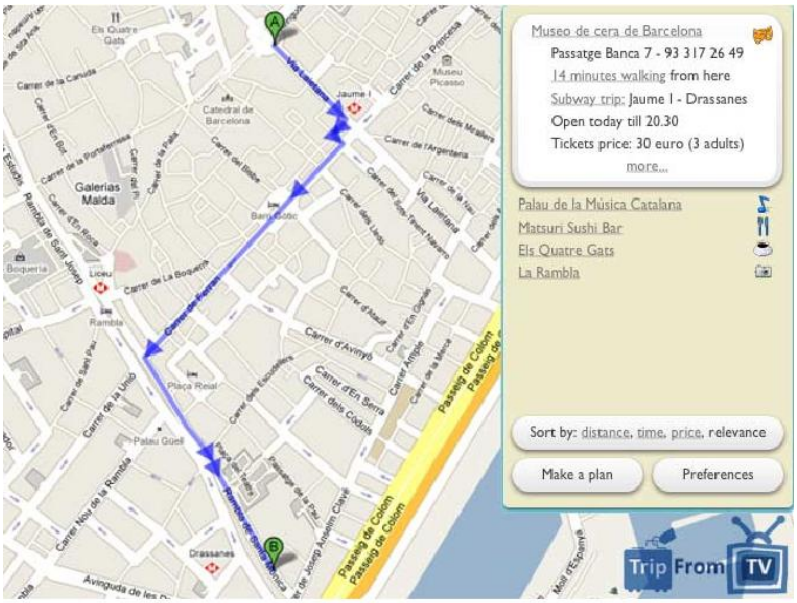
Παρά την ύπαρξη συχνά χρησιμοποιημένων οντολογιών, όπως το Word Net, το OpenCyc και το Gene Ontology, υπάρχουν ακόμα κάποιες προκλήσεις. Μία από αυτές είναι το συνάλλαγμα. Οι σχέσεις μεταξύ των αντικειμένων μπορεί να αλλάξουν και τότε οι οντολογίες δεν αντιστοιχούν στα δεδομένα που ισχύουν αυτή τη στιγμή.

Ο Ahmed Elgohary πρότεινε σε μια προσέγγισή του την χρήση της Wikipedia σαν μια οντολογία για την σημασιολογική ανάλυση ενός κειμένου. Τα άρθρα στην Wikipedia ανανεώνονται σε συχνή βάση και μερικά από αυτά υπάρχουν σε αρκετές γλώσσες. Επιπλέον είναι συνδεδεμένα μεταξύ τους με υπέρ-συνδέσμους. Με την χρήση ενός υπομνηματιστή βασισμένο στην Wikipedia το σύστημα συστάσεων δημιουργεί μια «αποθήκη» από υπομνήματα, η οποία θα χρησιμοποιηθεί για τις περεταίρω σημασιολογικές περιγραφές των όρων(αντικειμένων).
Η σημασιολογική ανάλυση μπορεί να χρησιμοποιηθεί σε διαφορετικού τύπου συστήματα συστάσεων για καλύτερη κατανόηση των σχέσεων μεταξύ αντικειμένων και χρηστών, όπως η δημιουργία μιας σημασιολογικής γειτονιάς μεταξύ διαφόρων προφίλ οντολογιών. Η ταξονομία από αντικείμενα είναι μια από τις πιο επιτυχημένες λύσεις για το πρόβλημα του «Νέου αντικειμένου».



Εικόνα 5 : Παράδειγμα OWL

Ένα παράδειγμα τέτοιας μηχανής συστάσεων είναι το TripFromTV που δουλεύει με ψηφιακές τηλεοράσεις και βασίζεται στην σημασιολογική πληροφορία. Η εφαρμογή συλλέγει πληροφορίες σχετικά με το τι είδους κανάλια, προγράμματα και ταινίες αρέσουν στον χρήστη που παρακολουθεί τηλεόραση και δημιουργεί κατηγοριοποιήσεις. Βασισμένη σε αυτές τις πληροφορίες ένα προφίλ χρήστη δημιουργείται από την εφαρμογή. Στην συνέχεια με βάση το ιστορικό και το προφίλ του χρήστη δημιουργούνται συστάσεις σχετικά με μέρη που θα μπορούσε να επισκεφθεί ή να ταξιδέψει. Ανάλογα με τις τηλεοπτικές προτιμήσεις του χρήστη είτε θα του προτείνει τον ταξιδιωτικό προορισμό της αρεσκείας του ή θα του προτείνει ένα εστιατόριο που θα σερβίρει την ανάλογη κουζίνα και βρίσκεται κοντά σε αυτόν. Παράλληλα με την βοήθεια του διαδικτύου η εφαρμογή επίσης να εμφανίσει το ωράριο λειτουργίας ή ακόμα και τις προσφορές του εν λόγω εστιατορίου.



Εικόνα 6 : Στιγμιότυπο από την χρήση TripFromTV

III) Cross-domain Προσεγγίσεις

Η εύρεση παρόμοιων χρηστών και η οικοδόμηση σωστών γειτονιών είναι ένα σημαντικό κομμάτι της διαδικασίας συστάσεων των συλλογικών συστημάτων συστάσεων. Ομοιότητες μεταξύ δύο χρηστών ανακαλύπτονται βασισμένες στις αξιολογήσεις των αντικειμένων. Παρόμοιες αξιολογήσεις σε ένα είδος δεν σημαίνει σίγουρα ότι οι αξιολογήσεις σε άλλα είδη είναι παρόμοιες.

Τα κλασσικά συλλογικά συστήματα συστάσεων συγκρίνουν τους χρήστες χωρίς να διαχωρίζουν τα αντικείμενα σε διαφορετικές κατηγορίες. Στα cross-domain συστήματα οι ομοιότητες μεταξύ των αντικειμένων για κάθε χρήστη λαμβάνουν υπόψη τους και το κάθε είδος. Μια μηχανή δημιουργεί τοπικές γειτονιές για κάθε χρήστη σχετικά με τα είδη του. Στην συνέχεια υπολογίζει τις ομοιότητες και οι ορισμένοι πλησιέστεροι γείτονες αποστέλλονται για συνεργατικό υπολογισμό ομοιοτήτων. Τα συστήματα συστάσεων αποφασίζουν την συνολική ομοιότητα και δημιουργούν ολοκληρωμένες γειτονιές ώστε να δημιουργήσουν προβλέψεις και συστάσεις.

IV) Peer-to-Peer Προσεγγίσεις

Τα συστήματα συστάσεων που λειτουργούν με Peer-to-Peer(P2P) προσεγγίσεις είναι αποκεντρωμένα. Κάθε χρήστης μπορεί να σχετιστεί με μια ομάδα από άλλους χρήστες με κοινά ενδιαφέροντα και να λάβει συστάσεις από τους χρήστες αυτής της ομάδας. Οι συστάσεις μπορούν επίσης να δοθούν με βάση το ιστορικό του κάθε χρήστη. Η αποκέντρωση του συστήματος σύστασης επιλύει το πρόβλημα της προσαρμοστικότητας(scalability problem).

V) Διαγλωσσικές (Cross-lingual) Προσεγγίσεις

Τα συστήματα συστάσεων βασισμένα στη διαγλωσσική προσέγγιση δίνουν την δυνατότητα στους χρήστες να λάβουν συστάσεις για αντικείμενα που έχουν περιγραφές σε γλώσσες που αυτοί ούτε ομιλούν ούτε καταλαβαίνουν.

Οι Yang, Chen και Wu πρότειναν μια προσέγγιση για διαγλωσσικές ομαδικές συστάσεις στις ειδήσεις. Η κύρια ιδέα είναι να χαρτογραφηθεί τόσο το κείμενο όσο και οι λέξεις κλειδιά από διαφορετικές γλώσσες σε ένα κοινό χαρακτηριστικό, λαμβάνοντας υπόψη την πιθανότητα διανομής λάθους θεμάτων. Από τις περιγραφές των αντικειμένων το σύστημα προσπελαύνει τις λέξεις κλειδιά μεταφράζοντάς τις σε μια από τις ορισμένες γλώσσες με την χρήση των λεξικών. Μετά από τη μετάφραση, χρησιμοποιείτε συνεργατικό ή άλλο φιλτράρισμα ώστε το σύστημα να δώσει τις συστάσεις στους χρήστες.

Με την χρήση της σημασιολογικής ανάλυσης είναι εφικτή η δημιουργή μιας γλωσσικά ανεξάρτητης αναπαράστασης σε κείμενο. Ο Pasquale Lops πρότεινε μια προσέγγιση με ένα σύστημα συστάσεων που ονομάζεται MARS(MultilAnguage Recommender System). Βασισμένο στον αλγόριθμο Word Sense Disambiguation, που εκμεταλλεύεται μια διαδικτυακή διαγλωσσική βάση δεδομένων για να εξασφαλίσει το ότι θα βγαίνει νόημα, το σύστημα δίνει το σωστό νόημα στις λέξεις αποφεύγοντας το πρόβλημα των συνώνυμων.

Παρόμοια προσέγγιση προτάθηκε από τους Murad Magableh και Antonio Cau, που χρησιμοποίησαν την λεξική οντότητα του WordNet με το πολύγλωσσο λεξικό EuroWordNet σαν πόρο ώστε να αποφύγουν το πρόβλημα των συνωνύμων και να δημιουργήσουν πιο ακριβής σχέσεις μεταξύ ετικετών σε folksonomies. Κάθε φορά που ο χρήστης προσθέτει μια ετικέτα σε ένα αντικείμενα, συνώνυμα της λέξης ετικέτα ανακαλύπτονται από το συνεργατικό σύστημα ετικετών και προσθέτονται στην ετικέτα με τη χρήση της οντολογίας του WordNet. Το EuroWordNet οικοδομεί σχέσεις μεταξύ ετικετών σε διαφορετικές ευρωπαϊκές γλώσσες, οι οποίες επιτρέπουν στα συστήματα συστάσεων να συγκρίνουν και να κατατάξουν αντικείμενα με ετικέτες σε διαφορετικές γλώσσες.

Τα διαγλωσσικά συστήματα συστάσεων σπάνε το γλωσσικό φράγμα και δίνουν δυνατότητες για εύρεση αντικειμένων, πληροφοριών, εργασιών ή βιβλίων σε άλλες γλώσσες.

Προτεινόμενες Λύσεις

Μιας και σκοπός της εργασίας είναι η απεικόνιση και η πειραματική λειτουργία μιας μηχανής σύστασης, πρέπει να παρουσιαστούν τα εργαλεία που χρησιμοποιήθηκαν κατά την υλοποίησή της καθώς και τα δεδομένα βάση των οποίων οι συστάσεις δόθηκαν.

Σε αυτή την ενότητα θα αναλυθεί τόσο το λογισμικό, όσο και οι βιβλιοθήκες που χρησιμοποιήθηκαν κατά την προσομοίωσή μας. Ταυτόχρονα θα παρουσιαστεί και η βάση δεδομένων που χρησιμοποιήθηκε.

Βάση Δεδομένων

Η βάση που χρησιμοποιήθηκε είναι η βάση MovieLens 1m. Η εν λόγω βάση διαθέτει 1.000.209 ανώνυμες αξιολογήσεις σε περίπου 3.900 ταινίες από 6.040 χρήστες του MovieLens το έτος 2000.

Η βάση δημιουργήθηκε από το GroupLens, μιας ερευνητικής ομάδας του τμήματος Επιστήμης των Υπολογιστών και Μηχανικών του Πανεπιστημίου της Μινεσότα, και της έρευνας ηγούνται οι καθηγητές John Riedl και Joseph Konstan.

Όλες οι αξιολογήσεις βρίσκονται στο αρχείο ratings.dat και έχουν την ακόλουθη μορφή:

UserID::MovieID::Rating::Timestamp

Όσων αφορά το κάθε πεδίο:

- Το πεδίο UserID, που αφορά τον αριθμό μητρώου του χρήστη έχει εύρος μεταξύ 1 και 6040.
- Το πεδίο MovieID, που αφορά τον αριθμό μητρώου της ταινίας έχει εύρος μεταξύ 1 και 3952.
- Οι αξιολογήσεις(Rating) γίνονται στην κλίμακα του 1 με 5 αστέρια, και είναι ακέραιες.
- Το timestamp είναι η χρονοσφραγίδα που έλαβε χώρο η αξιολόγηση
- Κάθε χρήστης έχει τουλάχιστον 20 αξιολογήσεις

Οι πληροφορίες για κάθε χρήστη βρίσκονται στο αρχείο users.dat και έχουν την ακόλουθη μορφή

UserID::Gender::Age::Occupation::Zip-code

Όσων αφορά το κάθε πεδίο:

- Το φύλο(gender) αναπαριστάται με «M» για αρσενικό και «F» για θηλυκό
- Οι ηλικίες(Age) έχουν την εξής διαφοροποίηση
    - 1: Για κάτω από 18
    - 18: Για 18-24
    - 25: Για 25-34
    - 35: Για 35-44
    - 45: Για 45-49
    - 50: Για 50-55
    - 56: Για άνω των 56
- Όσων αφορά την εργασία(occupation) η διάκριση γίνεται με τις ακόλουθες επιλογές:

- 0: Άλλη ή μη συμπληρωμένη
- 1: Ακαδημαϊκός/Δάσκαλος
- 2: Καλλιτέχνης
- 3: Ιδιωτικός υπάλληλος
- 4: Φοιτητής/Μεταπτυχιακός
- 5: Εξυπηρέτηση πελατών
- 6: Γιατρός/Υπηρεσίες υγείας
- 7: Προϊστάμενος
- 8: Αγρότης
- 9: Αρχιτέκτονας
- 10: Μαθητής δημοτικού
- 11: Δικηγόρος
- 12: Προγραμματιστής
- 13: Συνταξιούχος
- 14: Πωλητής
- 15: Επιστήμονας
- 16: Ελεύθερος επαγγελματίας
- 17: Τεχνικός/Μηχανικός
- 18: Έμπορος/Βιοτέχνης
- 19: Άνεργος
- 20: Συγγραφέας

Οι πληροφορίες για τις ταινίες βρίσκονται στο αρχείο movies.dat και έχουν την παρακάτω μορφή:

MovieID::Title::Genres

Όσων αφορά το κάθε πεδίο:

- Οι τίτλοι των ταινιών είναι οι ίδιοι με αυτούς του IMDB(Internet Movie DataBase) και έχουν σε παρένθεση το έτος που προβλήθηκαν πρώτη φορά σε κινηματογράφο
- Τα είδη(genres) διακρίνονται στα παρακάτω και διαχωρίζονται με κόμμα:
  - Action
  - Adventure
  - Animation
  - Children’s
  - Comedy
  - Crime
  - Documentary
  - Drama
  - Fantasy
  - Film-Noir
  - Horror
  - Musical
  - Mystery
  - Romance
  - Sci-Fi
  - Thriller
  - War
  - Western

Λογισμικό

Ο βέλτιστος τρόπος να μελετήσουμε τις μηχανές συστάσεων ήταν σε περιβάλλον Python. Η Python διαθέτει τις πλέον κατάλληλες βιβλιοθήκες, ώστε να μπορούμε να μελετήσουμε τα δεδομένα μιας βάσης, να προβούμε σε συστάσεις, αλλά και προβλέψεις.

Οι διερμηνευτές της Python είναι διαθέσιμοι για εγκατάσταση σε πολλά λειτουργικά συστήματα, επιτρέποντας στην Python την εκτέλεση κώδικα σε ευρεία γκάμα συστημάτων. Χρησιμοποιώντας εργαλεία τρίτων ο κώδικας της Python μπορεί να πακεταριστεί σε αυτόνομα εκτελέσιμα προγράμματα για μερικά από τα πιο δημοφιλή λειτουργικά συστήματα, επιτρέποντας τη διανομή του βασισμένου σε Python λογισμικού για χρήση σε αυτά τα περιβάλλοντα χωρίς να απαιτείται εγκατάσταση του διερμηνευτή της Python. Στην περίπτωση μας χρησιμοποιήσαμε το Anaconda και συγκεκριμένα την Version 5.2.

Βιβλιοθήκες

Στα πλαίσια της μελέτης χρησιμοποιήσαμε μια βασική βιβλιοθήκη:

- Το Turi, γνωστό και ως GraphLab

Σχετικά με το Turi

Το Turi είναι ένα βασισμένος σε γράφους, υψηλής απόδοσης λογισμικό κατανεμημένων υπολογισμών που είναι γραμμένο σε C++. Σαν GraphLab project ξεκίνησε από τον καθηγητή Carlos Guestrin του Carnegie Mellon University το 2009. Είναι ένα project ανοικτού λογισμικού που χρησιμοποιεί άδεια Apache. Παρότι το GraphLab αρχικά αναπτύχθηκε έχοντας ως αντικείμενο το Machine Learning, έχει πετύχει υψηλή επιτυχία σε αρκετό εύρος εργασιών εξόρυξης δεδομένων.

Το λογισμικό αποτελεί μια προσέγγιση σε παράλληλο προγραμματισμό με στόχο την αραίωση αλγορίθμων επαναληπτικών γράφων. Οι κύριοι πυλώνες στους οποίους βασίζεται ο σχεδιασμός του GraphLab είναι οι εξής:

- Αραίωση δεδομένων με τοπικές εξαρτήσεις
- Επαναληπτικοί αλγόριθμοι
- Πιθανή ασύγχρονη εκτέλεση

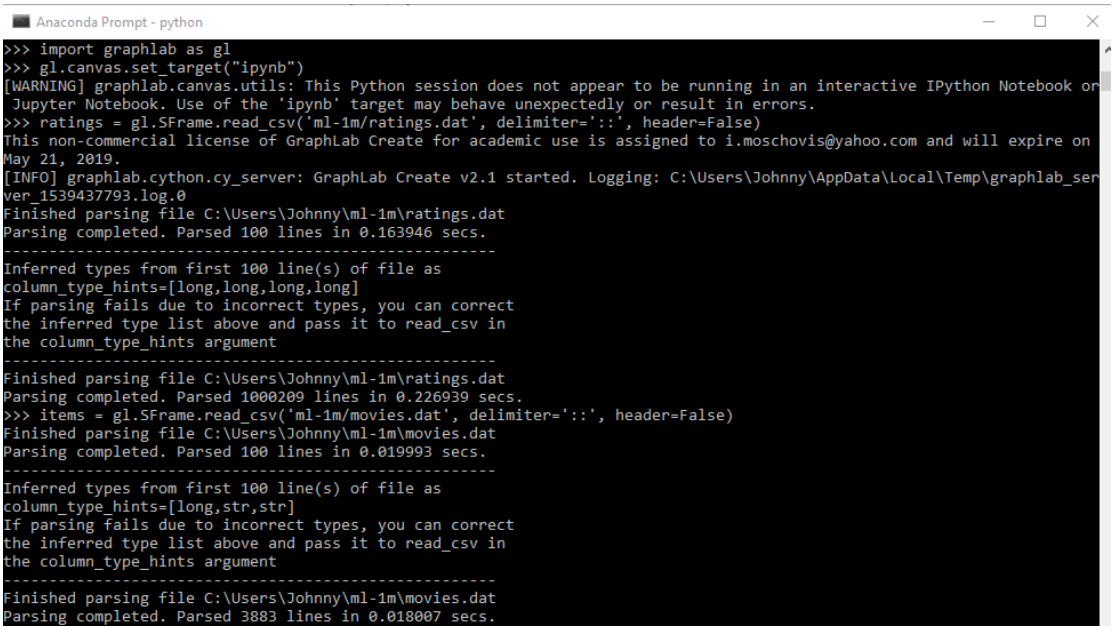
Δύο από τα βασικά εργαλεία του GraphLab που χρησιμοποιήθηκαν στα πλαίσια της εργασίας είναι:

- Η μοντελοποίηση, βάση της οποίας κατηγοριοποιήσαμε δεδομένα ανάλογα με τον τύπο τους
- Η χρήση αλγορίθμων συνεργατικού φιλτραρίσματος για την δημιουργία συστάσεων και προβλέψεων.

Στα πλαίσια της εργασίας χρησιμοποιήθηκε η έκδοση 2.2 του Turi η οποία χρησιμοποιεί την έκδοση 2.6 της Python. Επιπλέον για την χρήση του Turi μου χορηγήθηκε ακαδημαϊκό license.

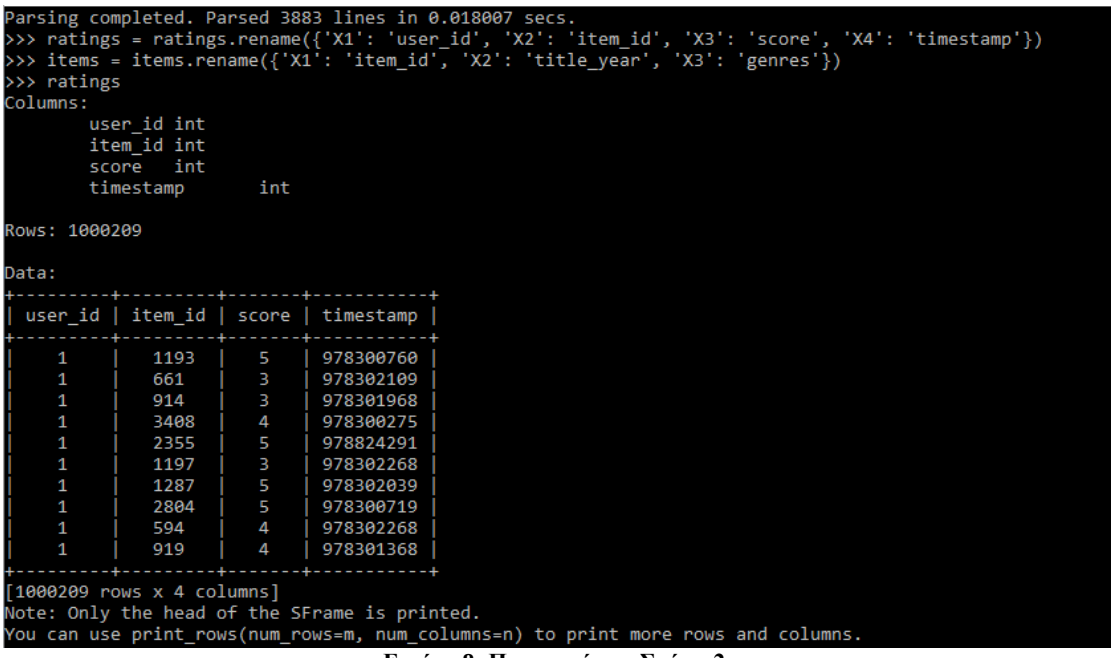
Προσομοίωση

Τρέχουμε το Anaconda command prompt και στην συνέχεια κάνουμε import την βιβλιοθήκη του graphlab(Turi)



Εικόνα 7: Προσομοίωση Σχήμα 1

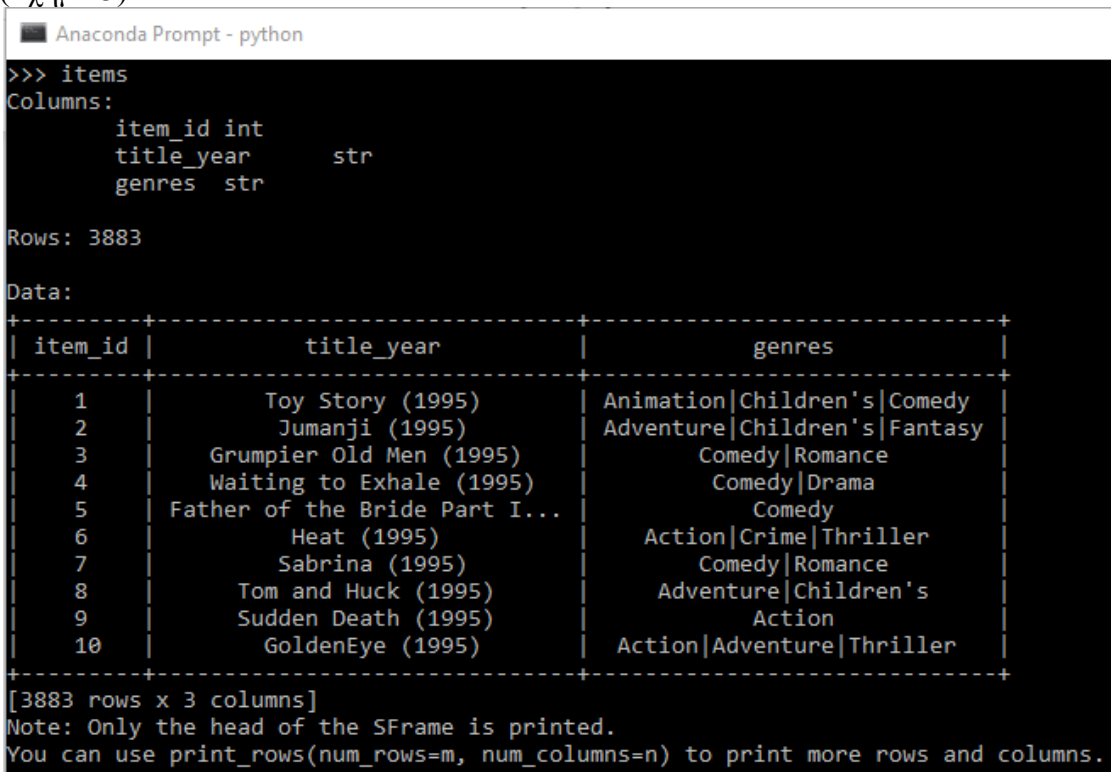
Στην συνέχεια καταχωρούμε σε Sframe τα αρχεία σε μορφή .dat διαβάζοντας τα csv των χρηστών, των ταινιών και των αξιολογήσεων(Σχήμα 1). Αφού έχει εξασφαλιστεί η προσπέλαση των δεδομένων, μετονομάζουμε τις στήλες στις αντίστοιχες, με βάση το αρχείο readme που διαθέτει η βάση(Σχήμα 2).



Εικόνα 8: Προσομοίωση Σχήμα 2

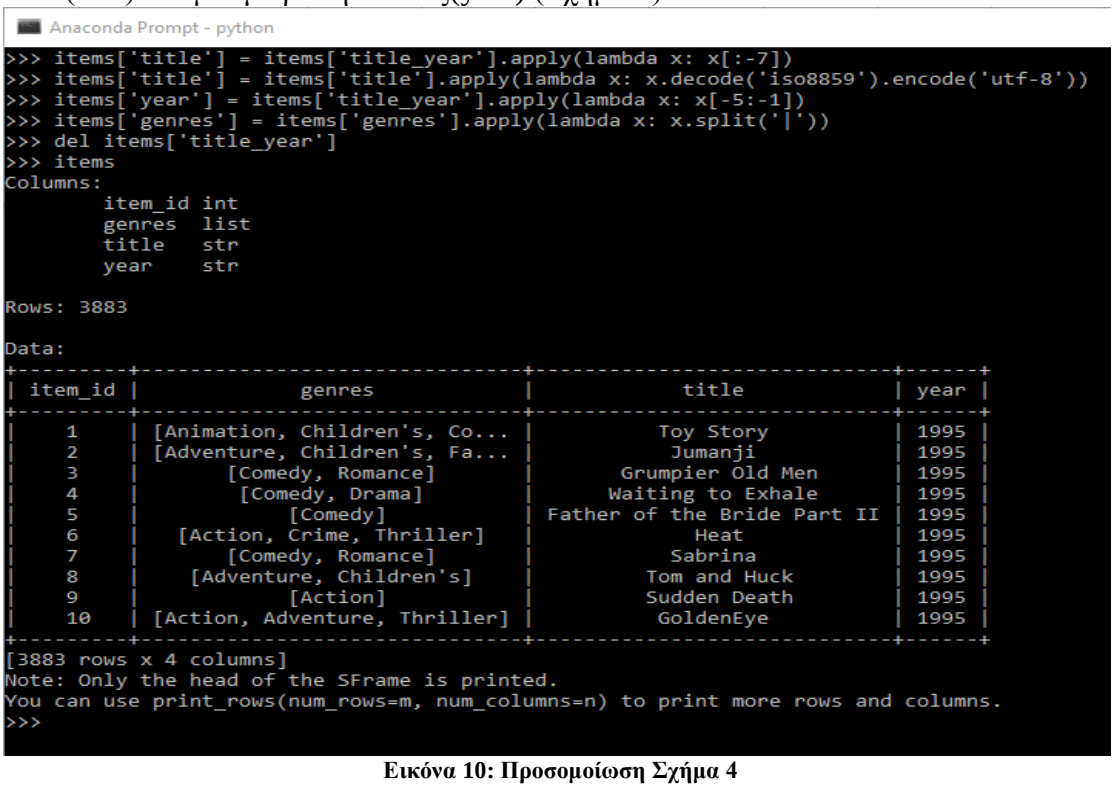
Εκτελούμε την ratings και μας δίνονται οι 10 πρώτες αξιολογήσεις του χρήστη με id 1.

Με την χρήση της εντολής items βλέπουμε τις 10 πρώτες ταινίες με βάση το id τους. (Σχήμα 3)



Εικόνα 9: Προσομοίωση Σχήμα 3

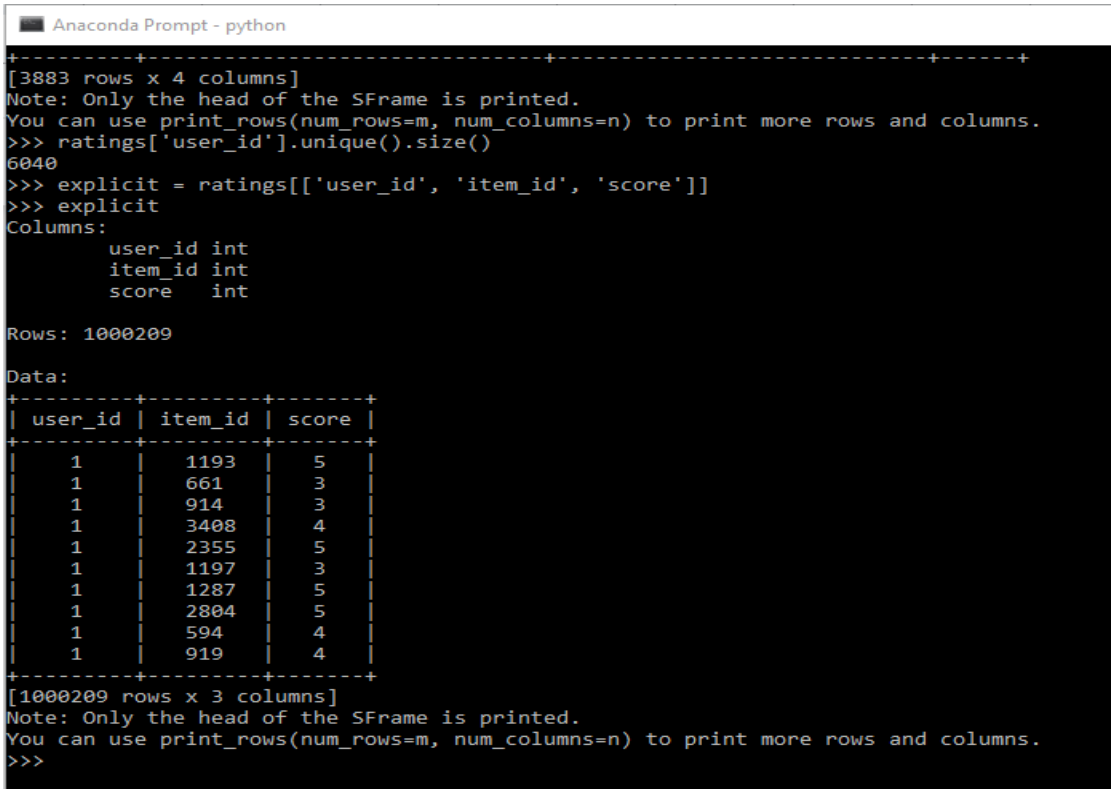
Στην συνέχεια διαχωρίζουμε τη στήλη title_year σε δύο νέες στήλες: μία με βάση τον τίτλο(title) και μια με βάση το έτος(year).(Σχήμα 4)



Εικόνα 10: Προσομοίωση Σχήμα 4

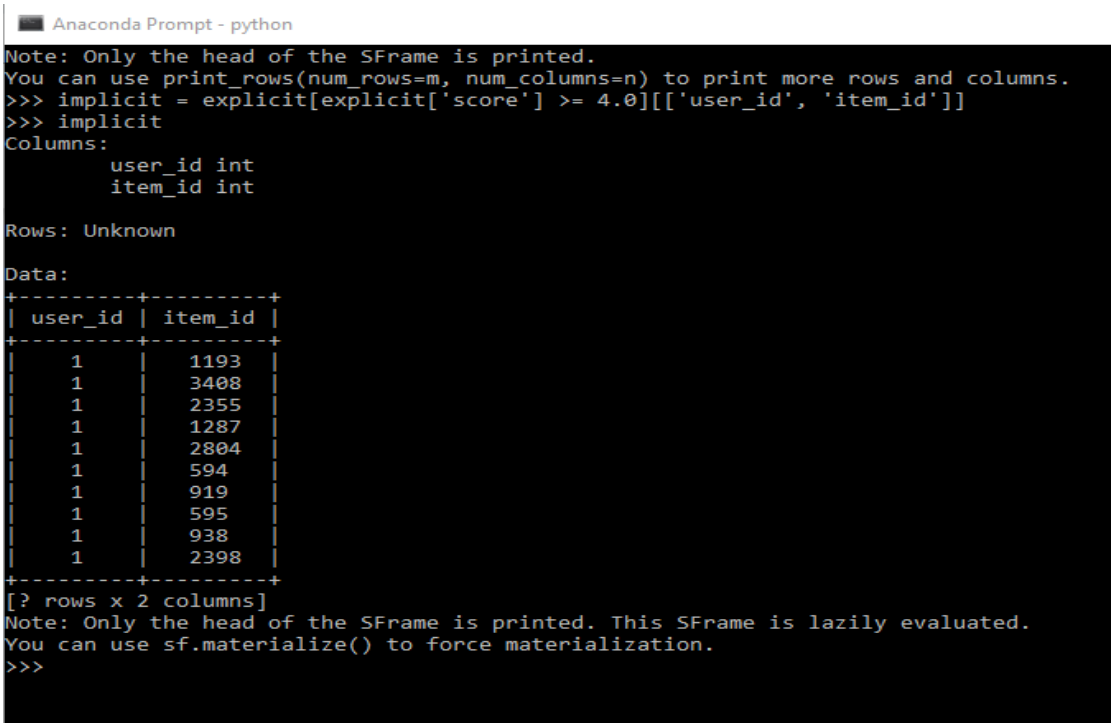
Αφού οι στήλες δημιουργηθούν διαγράφουμε τη στήλη title_year.

Στην συνέχεια δημιουργούμε μια ομάδα δεδομένων(dataset). Η πρώτη είναι η explicit που περιλαμβάνει όλες τις αξιολογήσεις(Σχήμα 5).



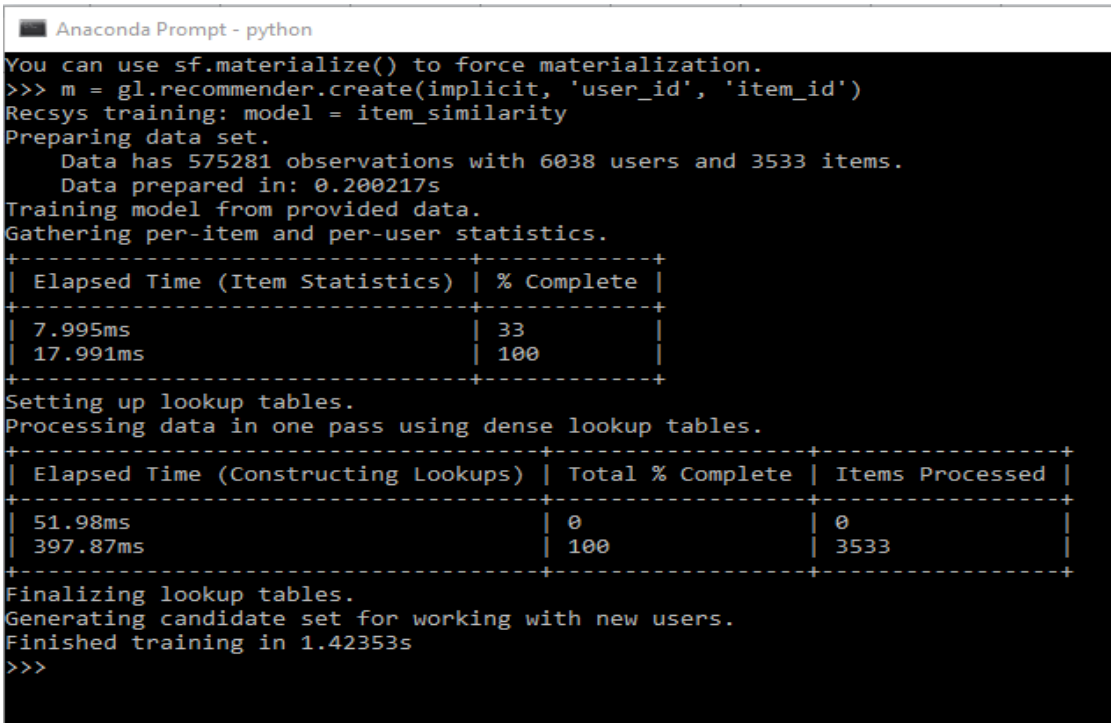
Εικόνα 11: Προσομοίωση Σχήμα 5

Η δεύτερη είναι η implicit και αφορά αξιολογήσεις με βαθμό 4 και 5(Σχήμα 6). Για την χρήση της δεύτερης ομάδας δεδομένων κρατάμε μόνο το id του χρήστη και της ταινίας.



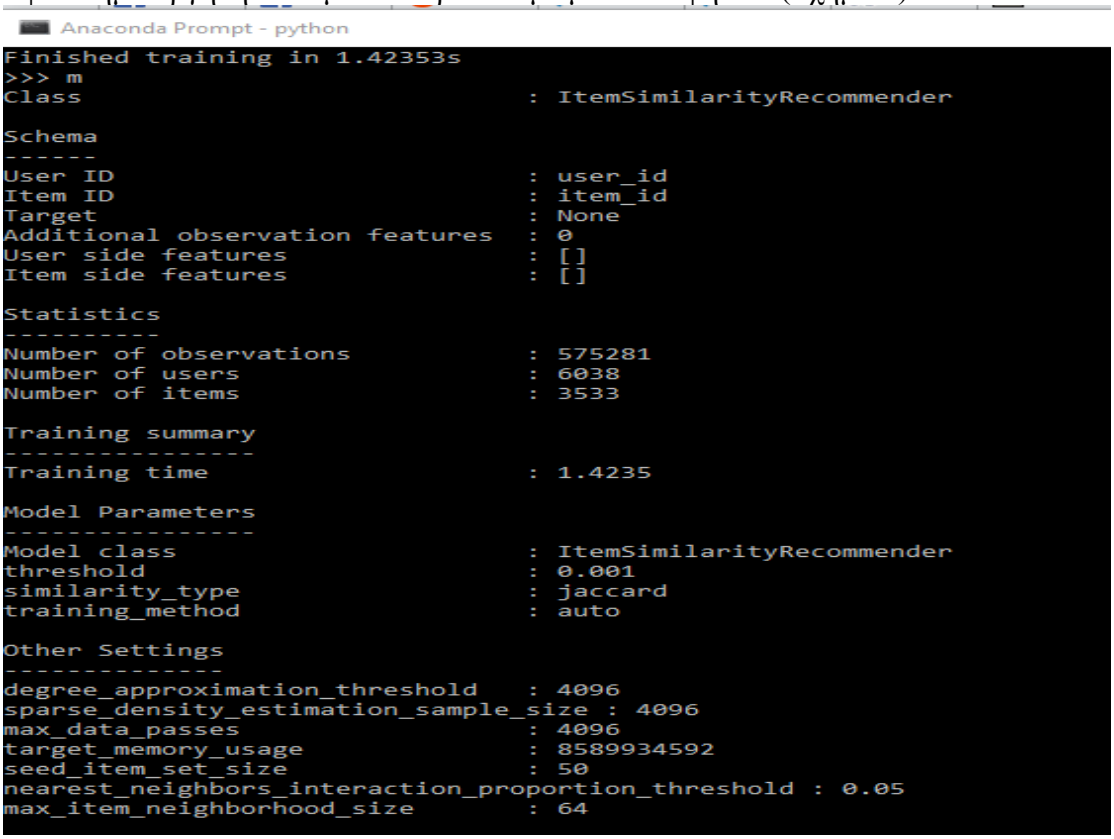
Εικόνα 12: Προσομοίωση Σχήμα 6

Έχοντας έτοιμες τις ομάδες δεδομένων, δημιουργούμε το πρώτο μας μοντέλο με βάση την ομάδα δεδομένων implicit, ώστε να προχωρήσουμε στις πρώτες μας συστάσεις.(Σχήμα 7)



Εικόνα 13: Προσομοίωση Σχήμα 7

Αφού δημιουργήθηκε το μοντέλο βλέπουμε μία σύνοψή του.(Σχήμα 8)



Εικόνα 14: Προσομοίωση Σχήμα 8

Στην συνέχεια διαλέγουμε μία τυχαία ταινία(Σχήμα 9).

```
>>> items[items['item_id'] == 1211]
Columns:
        item_id int
        genres  list
        title   str
        year    str

Rows: Unknown

Data:
+---------+--------------------------+-------------------------------+------+
| item_id |          genres          |             title             | year |
+---------+--------------------------+-------------------------------+------+
|   1211  | [Comedy, Drama, Romance] | Wings of Desire (Der Himme... | 1987 |
+---------+--------------------------+-------------------------------+------+
[? rows x 4 columns]
Note: Only the head of the SFrame is printed. This SFrame is lazily evaluated.
You can use sf.materialize() to force materialization.
```

Εικόνα 15: Προσομοίωση Σχήμα 9

Αφού την διαλέξαμε χρησιμοποιούμε φιλτράρισμα με βάση το αντικείμενο και δεχόμαστε τις πρώτες μας συστάσεις(Σχήμα 10).

```
Anaconda Prompt - python
>>> m.get_similar_items([1211]).join(items, on={'similar': 'item_id'}).sort('rank')
Columns:
        item_id int
        similar int
        score   float
        rank    int
        genres  list
        title   str
        year    str

Rows: 10

Data:
+---------+---------+----------------+------+---------------------------------+
| item_id | similar |     score      | rank |              genres             |
+---------+---------+----------------+------+---------------------------------+
|   1211  |   3089  | 0.159883737564 |  1   |            [Drama]              |
|   1211  |   3521  | 0.142348766327 |  2   |     [Comedy, Crime, Drama]      |
|   1211  |   1295  | 0.136518776417 |  3   |            [Drama]              |
|   1211  |   1300  | 0.134860038757 |  4   |            [Drama]              |
|   1211  |   247   | 0.134361207485 |  5   | [Drama, Fantasy, Romance, ...   |
|   1211  |   2065  | 0.133165836334 |  6   |    [Comedy, Drama, Romance]     |
|   1211  |   1172  | 0.132045090199 |  7   |    [Comedy, Drama, Romance]     |
|   1211  |   3067  | 0.130890071392 |  8   |        [Comedy, Drama]          |
|   1211  |   1289  | 0.129963874817 |  9   |       [Documentary, War]        |
|   1211  |   3742  | 0.129870116711 |  10  |          [Drama, War]           |
+---------+---------+----------------+------+---------------------------------+
+--------------------------------+------+
|             title              | year |
+--------------------------------+------+
| Bicycle Thief, The (Ladri ...  | 1948 |
|          Mystery Train         | 1989 |
| Unbearable Lightness of Be...  | 1988 |
| My Life as a Dog (Mitt liv...  | 1985 |
|       Heavenly Creatures       | 1994 |
|   Purple Rose of Cairo, The    | 1985 |
|         Cinema Paradiso        | 1988 |
| Women on the Verge of a Ne...  | 1988 |
|          Koyaanisqatsi         | 1983 |
| Battleship Potemkin, The (...  | 1925 |
+--------------------------------+------+
[10 rows x 7 columns]

>>>
```

Εικόνα 16: Προσομοίωση Σχήμα 10

Με αυτόν τον τρόπο έχουμε τις συστάσεις ταξινομημένες με βάση την συνάφεια των δύο ταινιών.

Στην συνέχεια δημιουργούμε ένα μοντέλο πρόβλεψης αξιολόγησης(Σχήμα 11).

```
>>> m2 = gl.recommender.create(explicit, 'user_id', 'item_id', target='score')
Recsys training: model = ranking_factorization_recommender
Preparing data set.
    Data has 1000209 observations with 6040 users and 3706 items.
    Data prepared in: 0.474369s
Training ranking_factorization_recommender for recommendations.
+--------------------------------+--------------------------------------------------+----------+
| Parameter                      | Description                                      | Value    |
+--------------------------------+--------------------------------------------------+----------+
| num_factors                    | Factor Dimension                                 | 32       |
| regularization                 | L2 Regularization on Factors                     | 1e-009   |
| solver                         | Solver used for training                         | sgd      |
| linear_regularization          | L2 Regularization on Linear Coefficients         | 1e-009   |
| ranking_regularization         | Rank-based Regularization Weight                 | 0.25     |
| max_iterations                 | Maximum Number of Iterations                     | 25       |
+--------------------------------+--------------------------------------------------+----------+
  Optimizing model using SGD; tuning step size.
  Using 125026 / 1000209 points for tuning the step size.
+---------+-------------------+----------------------------------------+
| Attempt | Initial Step Size | Estimated Objective Value              |
+---------+-------------------+----------------------------------------+
| 0       | 25                | Not Viable                             |
| 1       | 6.25              | Not Viable                             |
| 2       | 1.5625            | Not Viable                             |
| 3       | 0.390625          | Not Viable                             |
| 4       | 0.0976562         | 1.75494                                |
| 5       | 0.0488281         | 1.83749                                |
| 6       | 0.0244141         | 1.83546                                |
| 7       | 0.012207          | 1.85414                                |
+---------+-------------------+----------------------------------------+
| Final   | 0.0976562         | 1.75494                                |
+---------+-------------------+----------------------------------------+
Starting Optimization.
+---------+--------------+-------------------+-----------------------+-------------+
| Iter.   | Elapsed Time | Approx. Objective | Approx. Training RMSE | Step Size   |
+---------+--------------+-------------------+-----------------------+-------------+
| Initial | 2.999ms      | 2.44676           | 1.11711               |             |
+---------+--------------+-------------------+-----------------------+-------------+
| 1       | 445.981ms    | DIVERGED          | DIVERGED              | 0.0976562   |
| RESET   | 603.448ms    | 2.44672           | 1.1171                |             |
| 1       | 919.432ms    | 1.72028           | 1.0401                | 0.0488281   |
| 2       | 1.21s        | 1.53057           | 0.991039              | 0.0290334   |
| 3       | 1.51s        | 1.42306           | 0.948319              | 0.0214205   |
| 4       | 1.86s        | 1.34356           | 0.919278              | 0.0172633   |
| 5       | 2.16s        | 1.28586           | 0.896202              | 0.014603    |
| 6       | 2.46s        | 1.24795           | 0.881063              | 0.0127367   |
| 9       | 3.34s        | 1.18408           | 0.854229              | 0.00939698  |
| 11      | 3.95s        | 1.16123           | 0.844689              | 0.00808399  |
| 14      | 4.87s        | 1.1374            | 0.834637              | 0.00674643  |
| 19      | 6.46s        | 1.11263           | 0.823591              | 0.00536543  |
| 24      | 8.03s        | 1.09737           | 0.816951              | 0.0045031   |
+---------+--------------+-------------------+-----------------------+-------------+
Optimization Complete: Maximum number of passes through the data reached.
Computing final objective value and training RMSE.
       Final objective value: 1.09702
       Final training RMSE: 0.787861
```

Εικόνα 17: Προσομοίωση Σχήμα 11

Με αυτόν τον τρόπο υπολογίζουμε το αντικειμενικό βάρος καθώς και την ρίζα της μέσης αντικειμενικής απόκλισης. Ο χρόνος έρχεται σε συνάρτηση με την επεξεργαστική δυνατότητα του συστήματος που εκτελεί τις εντολές(όσων αφορά την ικανότητα του ενός πυρήνα) και παρατηρούμε όπως είναι λογικό ότι μετά από κάθε απόπειρα βελτιστοποίησης το βήμα ολοένα και ελαττώνεται.

Επιπλέον παρατηρούμε ότι για την βελτιστοποίηση του μοντέλου χρησιμοποιούμε την χρήση στοχαστικής γραμμικής παλινδρόμησης.

Ένας άλλος τρόπος δημιουργίας συστάσεων είναι οι ομαδοποιημένες συστάσεις. (Σχήμα 12).

```
>>> recs = m.recommend()
recommendations finished on 1000/6038 queries. users per second: 66688.9
recommendations finished on 2000/6038 queries. users per second: 71451.5
recommendations finished on 3000/6038 queries. users per second: 68203.5
recommendations finished on 4000/6038 queries. users per second: 65595.3
recommendations finished on 5000/6038 queries. users per second: 67589.5
recommendations finished on 6000/6038 queries. users per second: 65955.1
>>> recs
Columns:
        user_id int
        item_id int
        score   float
        rank    int

Rows: 60380

Data:
+---------+---------+----------------+------+
| user_id | item_id |     score      | rank |
+---------+---------+----------------+------+
|    1    |   318   | 0.119366906749 |  1   |
|    1    |   1307  | 0.118893305461 |  2   |
|    1    |   1259  | 0.117540442944 |  3   |
|    1    |   1198  | 0.115310272906 |  4   |
|    1    |   1265  | 0.112698629167 |  5   |
|    1    |   593   | 0.112643359767 |  6   |
|    1    |   296   | 0.112566664484 |  7   |
|    1    |    50   | 0.108486062951 |  8   |
|    1    |   1196  | 0.107369073232 |  9   |
|    1    |   1210  | 0.102146388425 |  10  |
+---------+---------+----------------+------+
[60380 rows x 4 columns]
Note: Only the head of the SFrame is printed.
You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.
```

Εικόνα 18: Προσομοίωση Σχήμα 12

Εξετάζουμε τις αξιολογήσεις ενός χρήστη (στο παράδειγμά μας εκείνου με το id 6, Σχήμα 13).

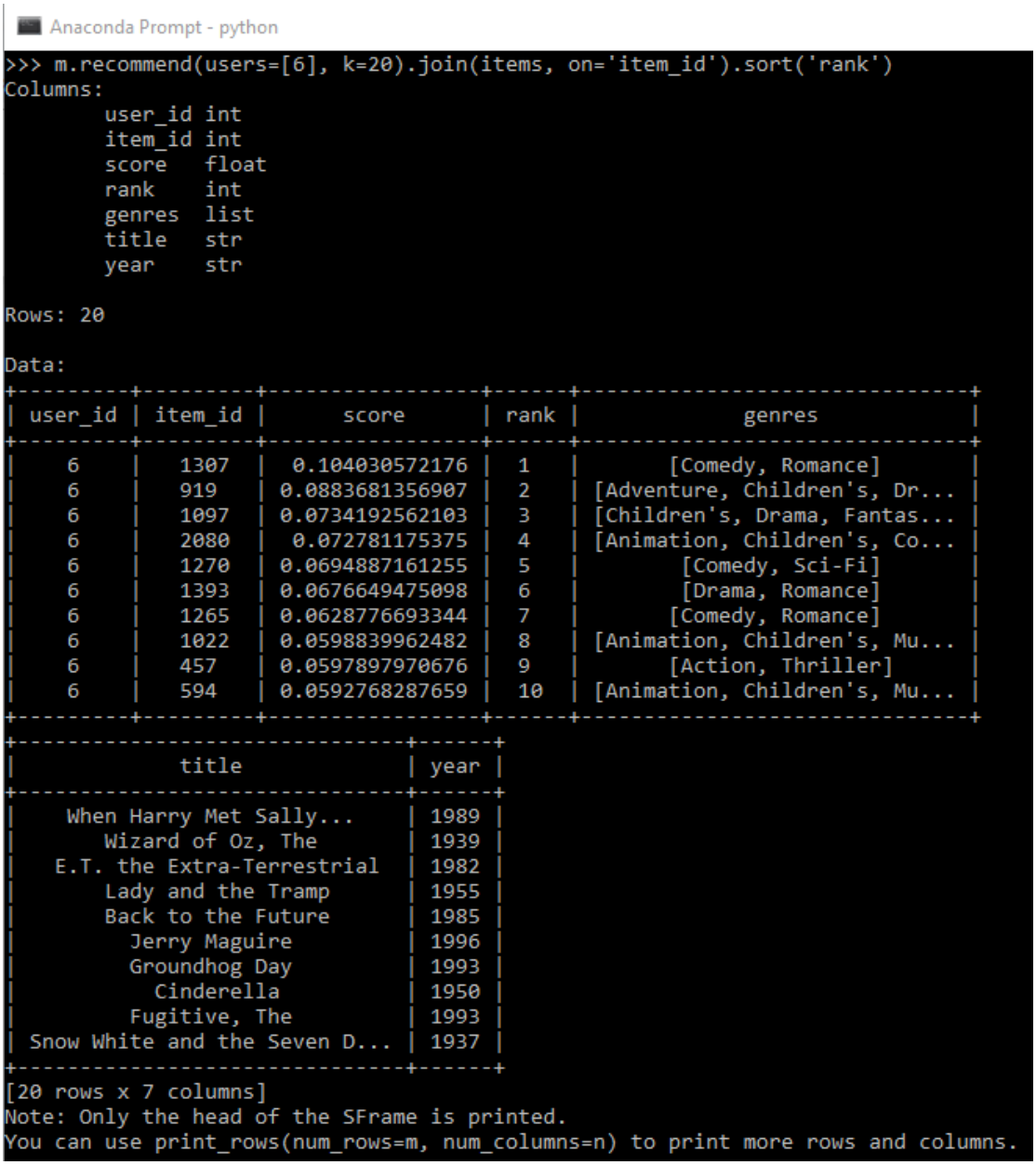
```
Anaconda Prompt - python
>>> ratings[ratings['user_id'] == 6].join(items, on='item_id')
Columns:
        user_id int
        item_id int
        score   int
        timestamp       int
        genres  list
        title   str
        year    str

Rows: 71

Data:
+---------+---------+-------+-----------+--------------------------------+
| user_id | item_id | score | timestamp |             genres             |
+---------+---------+-------+-----------+--------------------------------+
|    6    |    1    |   4   | 978237008 | [Animation, Children's, Co...  |
|    6    |    17   |   4   | 978236383 |        [Drama, Romance]        |
|    6    |    34   |   4   | 978237444 |  [Children's, Comedy, Drama]   |
|    6    |    48   |   5   | 978237570 | [Animation, Children's, Mu...  |
|    6    |   199   |   5   | 978237570 |        [Drama, Musical]        |
|    6    |   266   |   4   | 978237909 | [Drama, Romance, War, Western] |
|    6    |   296   |   2   | 978237379 |         [Crime, Drama]         |
|    6    |   364   |   4   | 978237570 | [Animation, Children's, Mu...  |
|    6    |   368   |   4   | 978237909 |  [Action, Comedy, Western]     |
|    6    |   377   |   3   | 978236383 |  [Action, Romance, Thriller]   |
+---------+---------+-------+-----------+--------------------------------+
+--------------------------------+------+
|             title              | year |
+--------------------------------+------+
|           Toy Story            | 1995 |
|     Sense and Sensibility      | 1995 |
|              Babe              | 1995 |
|           Pocahontas           | 1995 |
| Umbrellas of Cherbourg, Th...  | 1964 |
|      Legends of the Fall       | 1994 |
|         Pulp Fiction           | 1994 |
|        Lion King, The          | 1994 |
|           Maverick             | 1994 |
|            Speed               | 1994 |
+--------------------------------+------+
[71 rows x 7 columns]
Note: Only the head of the SFrame is printed.
You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.
>>>
```

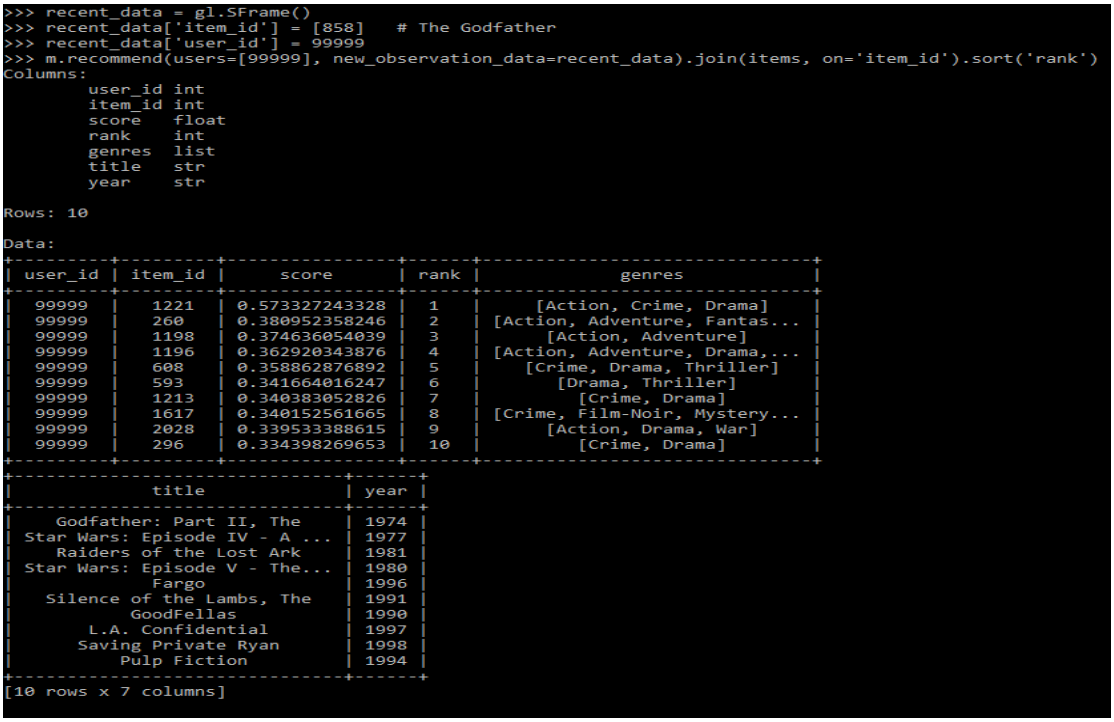
Εικόνα 19: Προσομοίωση Σχήμα 13

Προχωράμε στην σύσταση με βάση τον χρήστη και με του βρίσκουμε 20 συστάσεις(Σχήμα 14)



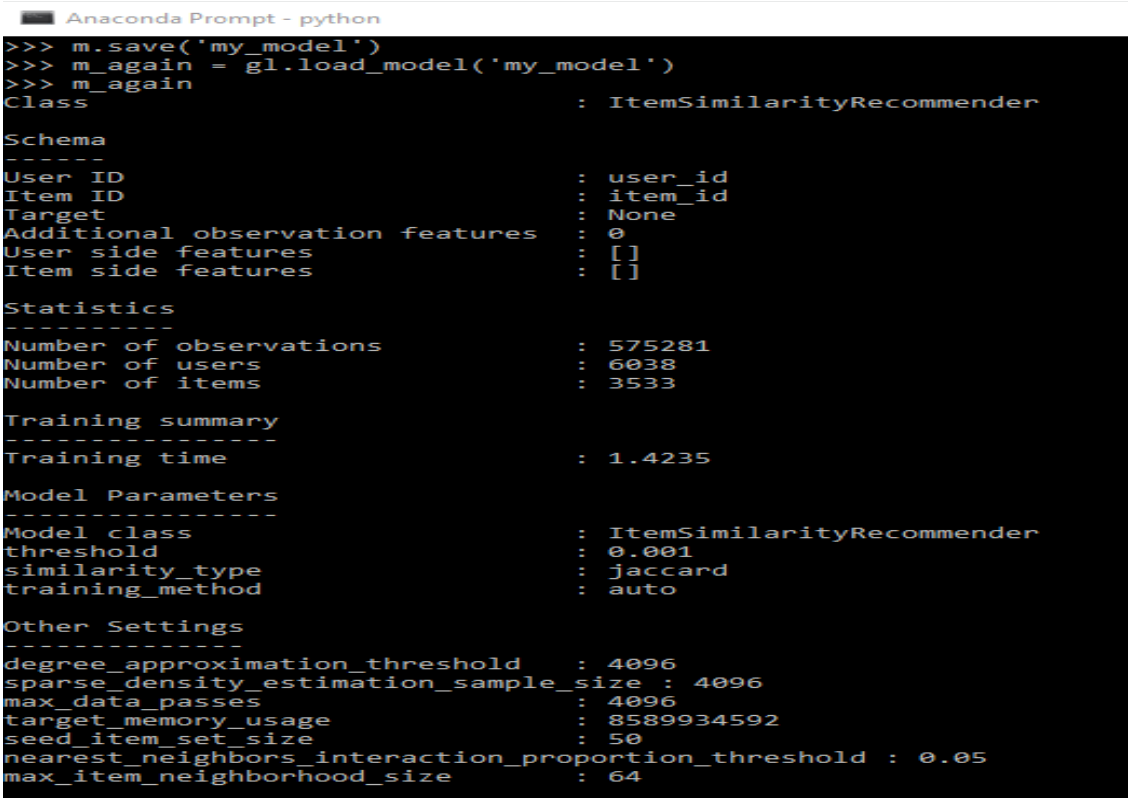
Εικόνα 20: Προσομοίωση Σχήμα 14

Με τον ίδιο τρόπο προσεγγίζουμε το πρόβλημα του νέου χρήστη. Εξετάζουμε ένα σενάριο στο οποίο ένας νέος χρήστης παρατήρησε την ταινία με id 858, η οποία είναι ο Νονός, και με τον ίδιο τρόπο παρατηρούμε τις ταινίες που θα του προταθούν (Σχήμα 15). Παρατηρούμε ότι οι συστάσεις γίνονται καθαρά με βάση την απόκλιση στην μέση αξιολόγηση της ταινίας και εξασφαλίζουμε ότι θα ταξινομηθούν με βάση αυτήν.



Εικόνα 21: Προσομοίωση Σχήμα 15

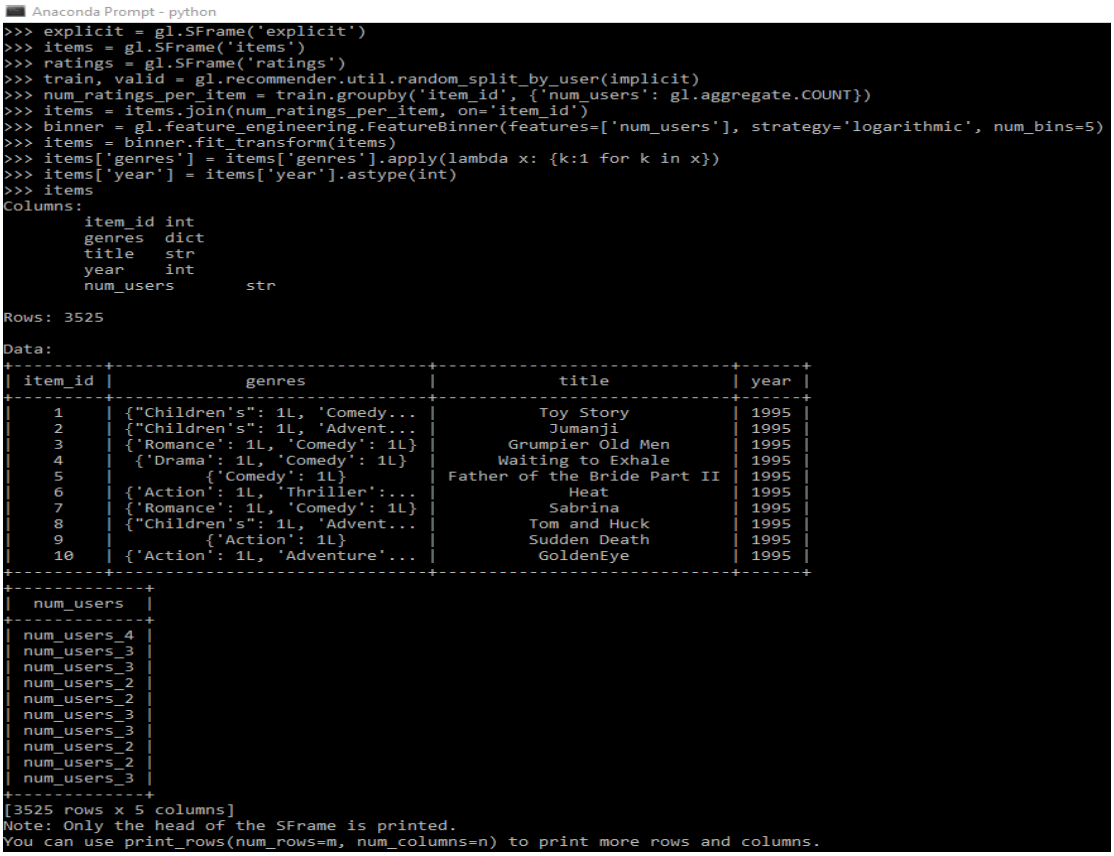
Αφού ολοκληρώσαμε τα σενάρια αποθηκεύουμε και επαναφορτώνουμε το μοντέλο μας(Σχήμα 16).



Εικόνα 22: Προσομοίωση Σχήμα 16

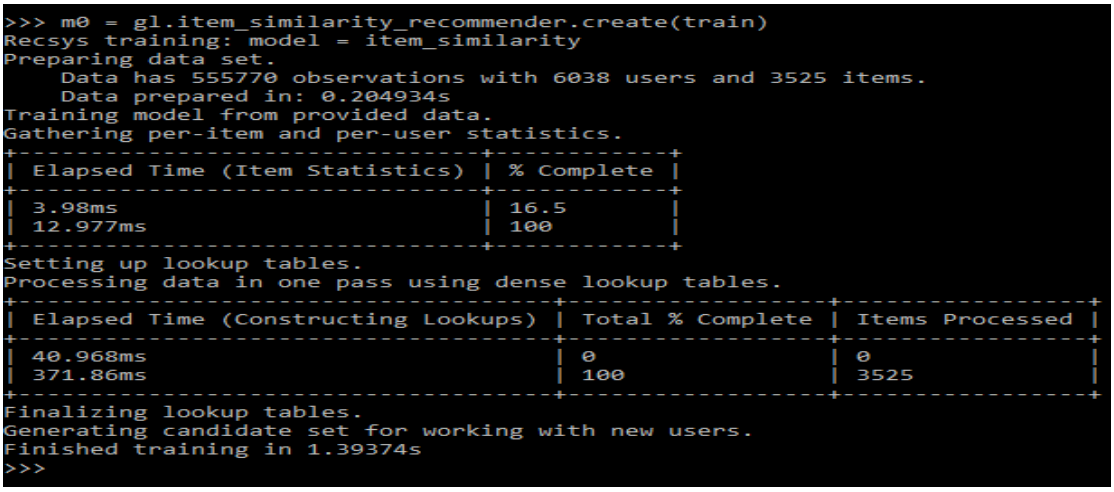
Ολοκληρώνοντας αυτή τη λειτουργία επανεκκινήθηκε η γραμμή εντολών, για να αδειάσει η προσωρινή μνήμη.

Διαβάζουμε πάλι τις ομάδες δεδομένων που αποθηκεύσαμε πριν. Διαιρούμε τα δεδομένα σε δύο ομάδες, ώστε να αποφύγουμε σφάλματα γενίκευσης, και υπολογίζουμε το πόσες φορές αξιολογήθηκε κάθε αντικείμενο. Με την χρήση του feature_engineering module μετατρέπουμε τον αριθμό των αξιολογήσεων σε κατηγορηματική μεταβλητή και μετατρέπουμε κάθε είδος ταινίας και κάθε έτος σε ακέραιο. (Σχήμα 17)



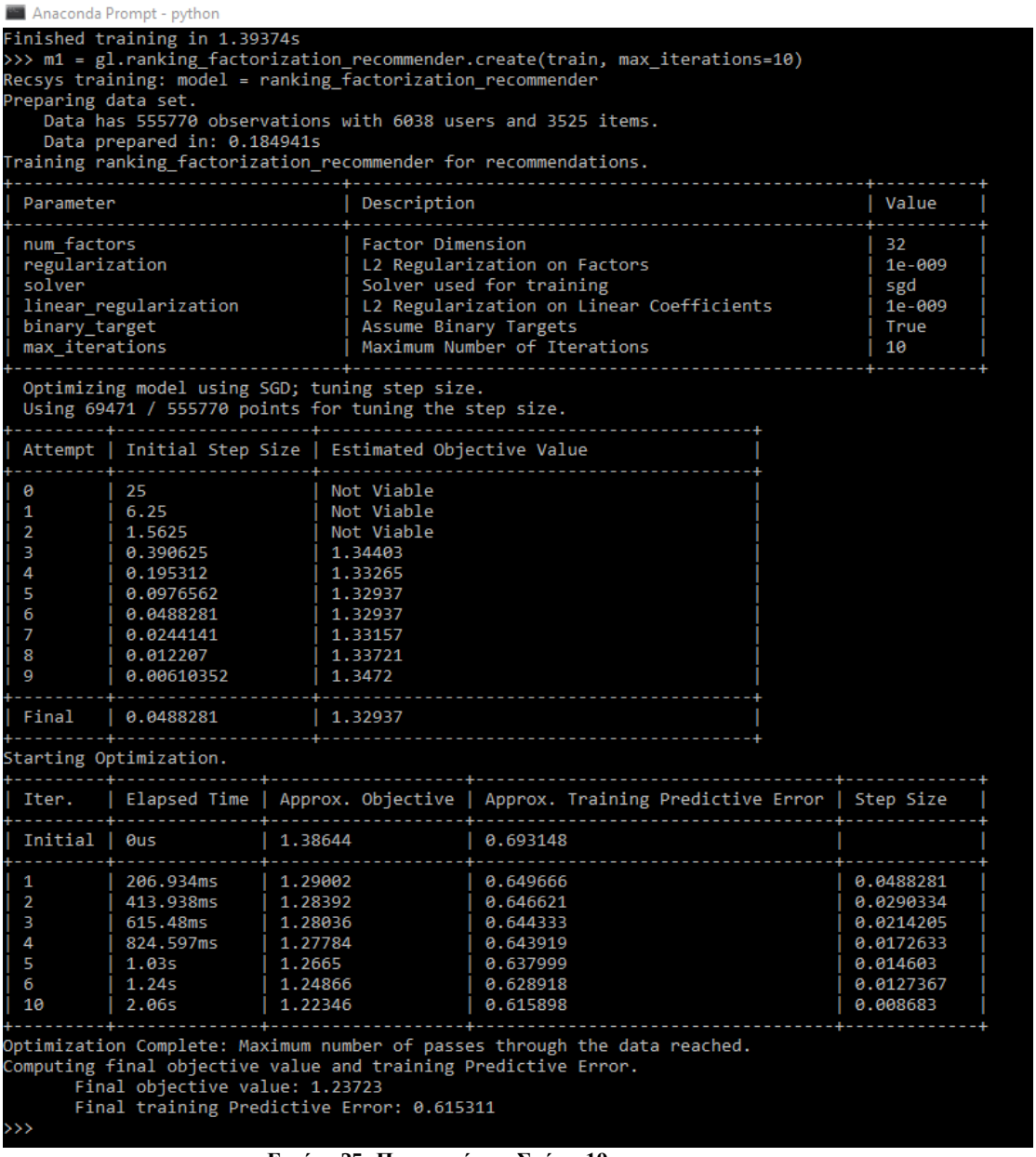
Εικόνα 23: Προσομοίωση Σχήμα 17

Στην συνέχεια προχωράμε στην δημιουργία μοντέλων.
Δημιουργούμε το μοντέλο m0, το οποίο αποτελεί μια προσέγγιση συνεργατικού φιλτραρίσματος που χρησιμοποιεί την ομοιότητα Jaccard για τις λίστες αντικειμένων δύο χρηστών(φιλτράρισμα με βάση τον χρήστη, Σχήμα 18).



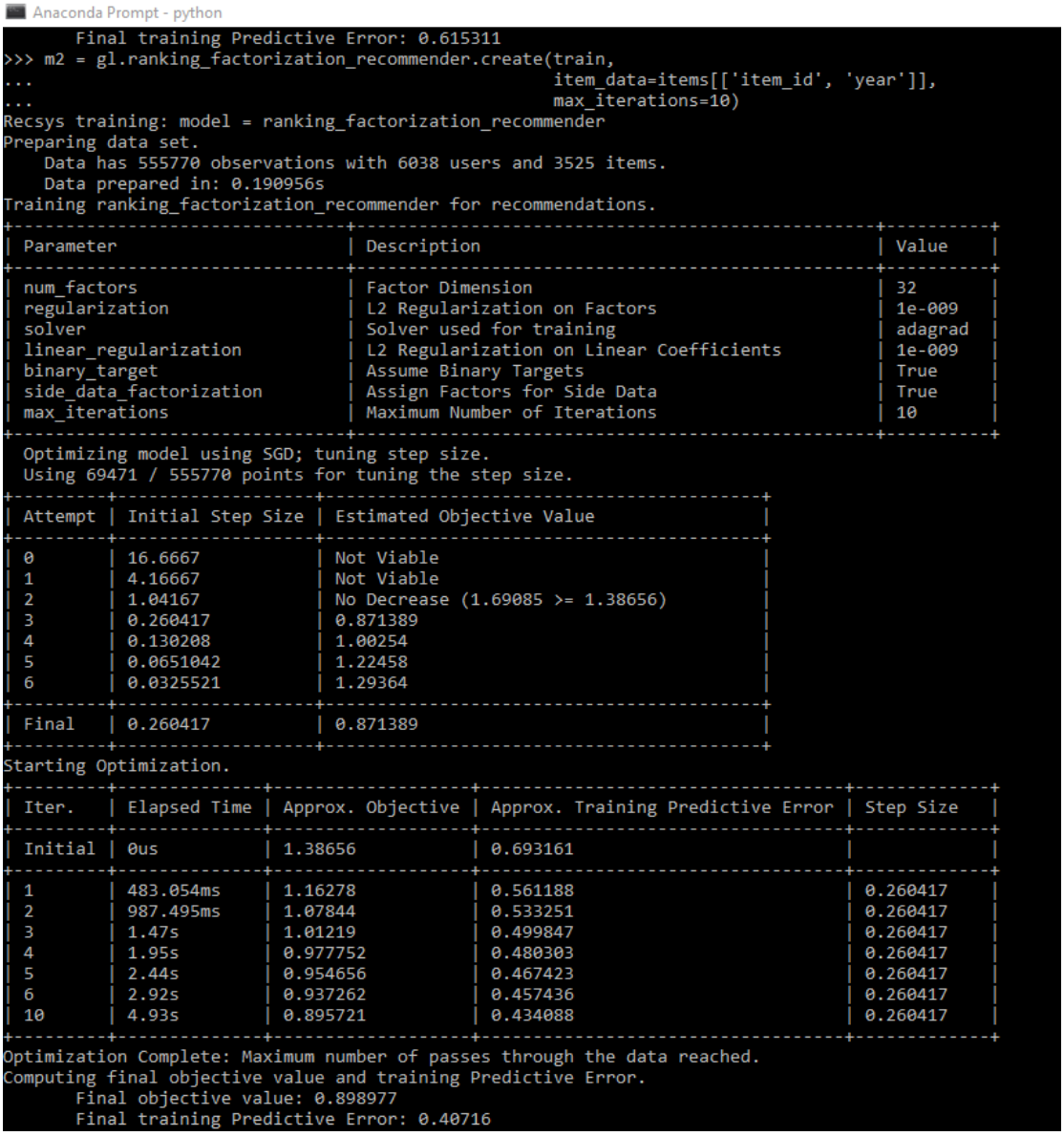
Εικόνα 24: Προσομοίωση Σχήμα 18

Το επόμενο μοντέλο είναι το m1, το οποίο αποτελεί μια προσέγγιση συνεργατικού φιλτραρίσματος με βάση τους λανθάνων παράγοντες για κάθε χρήστη και κάθε αντικείμενο(Σχήμα 19).



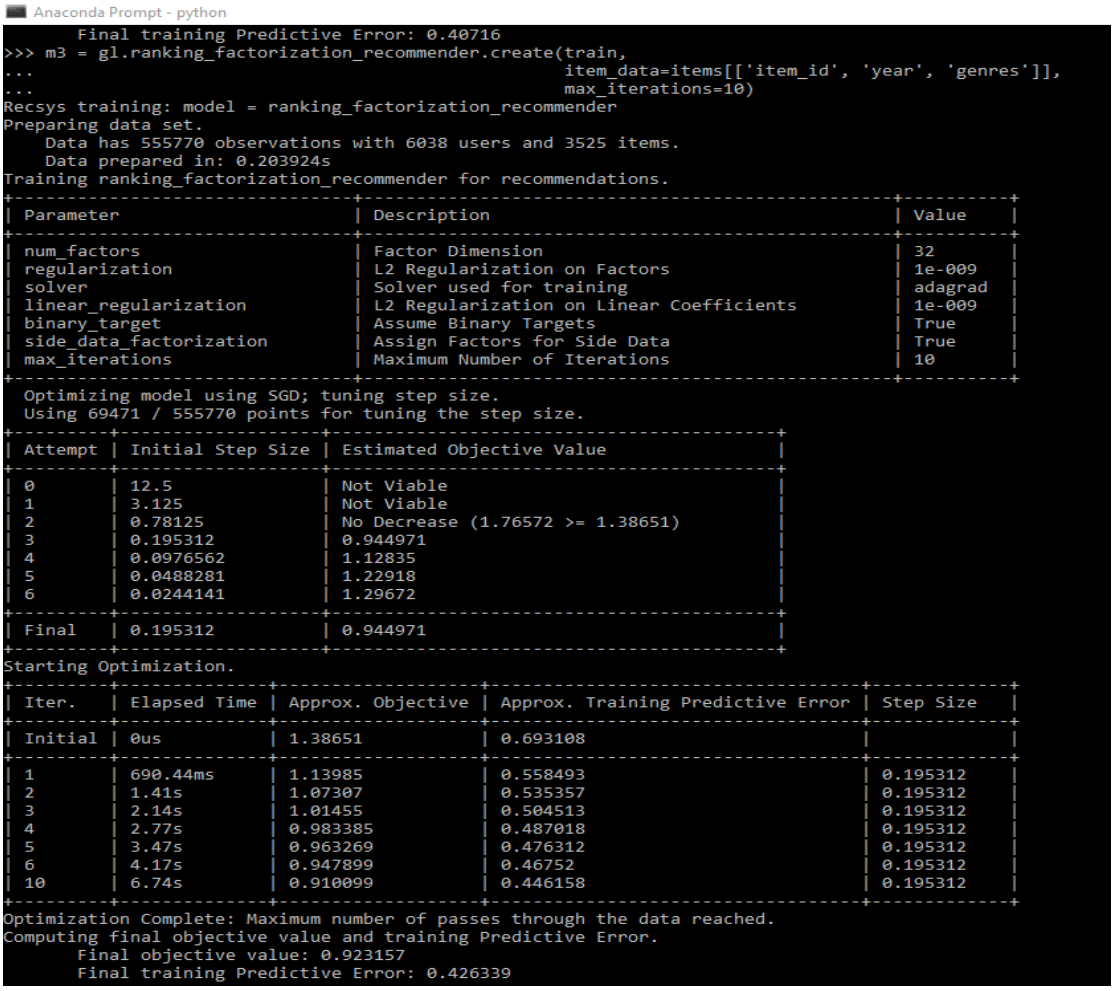
Εικόνα 25: Προσομοίωση Σχήμα 19

Παρατηρούμε μία ιδιαίτερα αυξημένη τιμή στην αντικειμενική αξία, αλλά παράλληλα και μία πιθανότητα σφάλματος πρόβλεψης της τάξης το 61.53%.
Με παρόμοιο τρόπο δημιουργούμε το m2 το οποίο είναι μια προσέγγιση συνεργατικού φιλτραρίσματος που υπολογίζει τους λανθάνων παράγοντες με βάση τους χρήστες, τα αντικείμενα και το έτος. (Σχήμα 20)



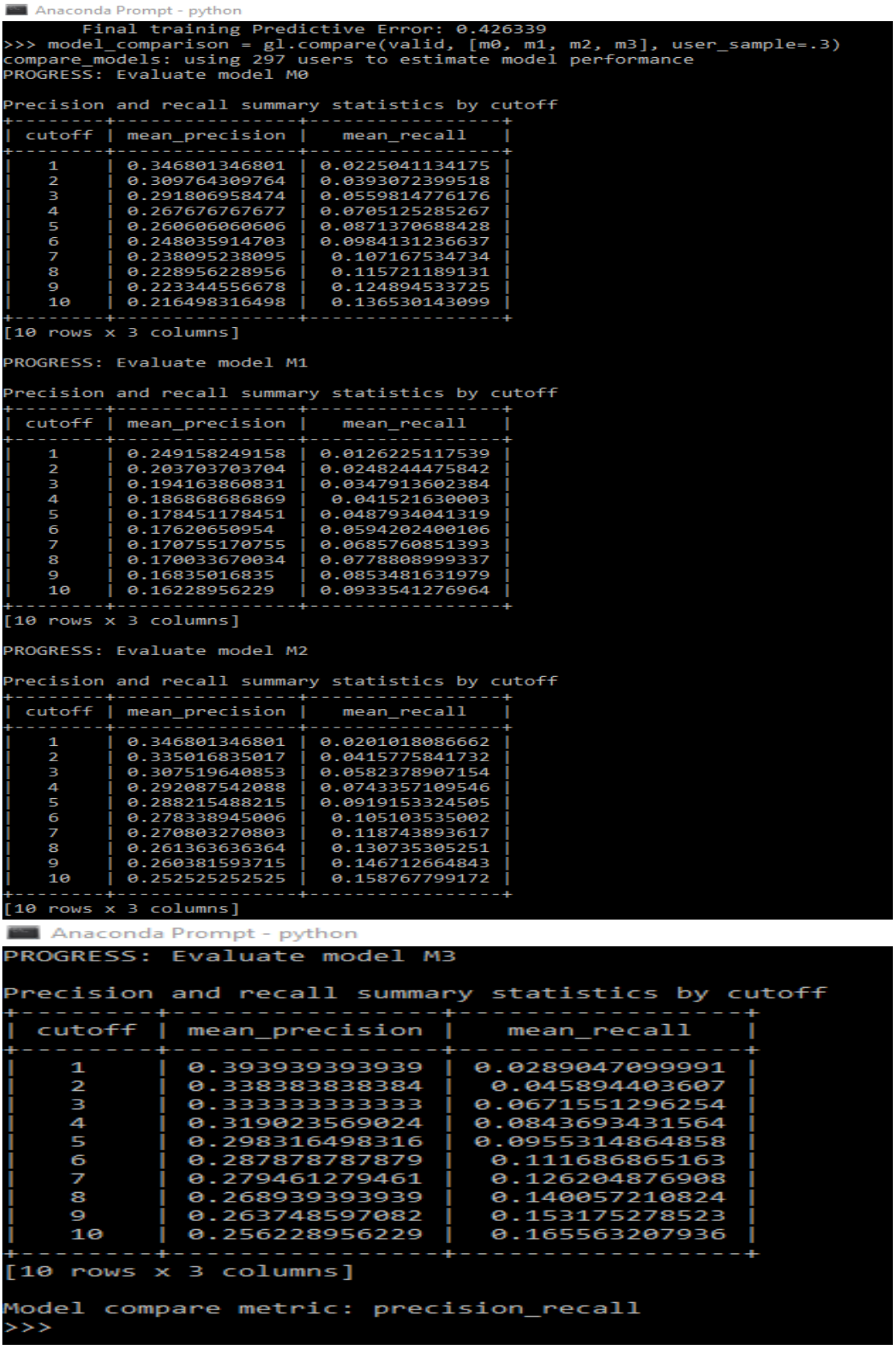
Εικόνα 26: Προσομοίωση Σχήμα 20

Βλέπουμε ελάττωση τόσο στην αντικειμενική αξία όσο και στην πιθανότητα σφάλματος πρόβλεψης(40.716%). Στο ίδιο μοτίβο δημιουργούμε και το μοντέλο m3 το οποίο λαμβάνει υπόψη του εκτός από τα προηγούμενα και το είδος. (Σχήμα 21)



Εικόνα 27: Προσομοίωση Σχήμα 21

Επειδή ο αριθμός των ειδών είναι είκοσι, έχουμε αύξηση τόσο της αντικειμενικής αξίας όσο και της πιθανότητας σφάλματος πρόβλεψης. Έχοντας δημιουργήσει τα τέσσερα αυτά μοντέλα(m0, m1, m2, m3), προχωράμε στην διαδικασία της αξιολόγησης. Με βάση τα ακριβή δεδομένα που λαμβάνουμε υπολογίζουμε την μέση ακρίβεια και την μέση ανάκληση με βάση τα στάδια αποκοπής(Σχήμα 22).



Εικόνα 28: Προσομοίωση Σχήμα 22

Σαν τελευταίο σενάριο προσομοίωσης θα δημιουργήσουμε ένα μοντέλο το οποίο με γνώμονα το είδος και τον χρόνο της ταινίας θα δημιουργήσει γειτονιές(Σχήμα 23).

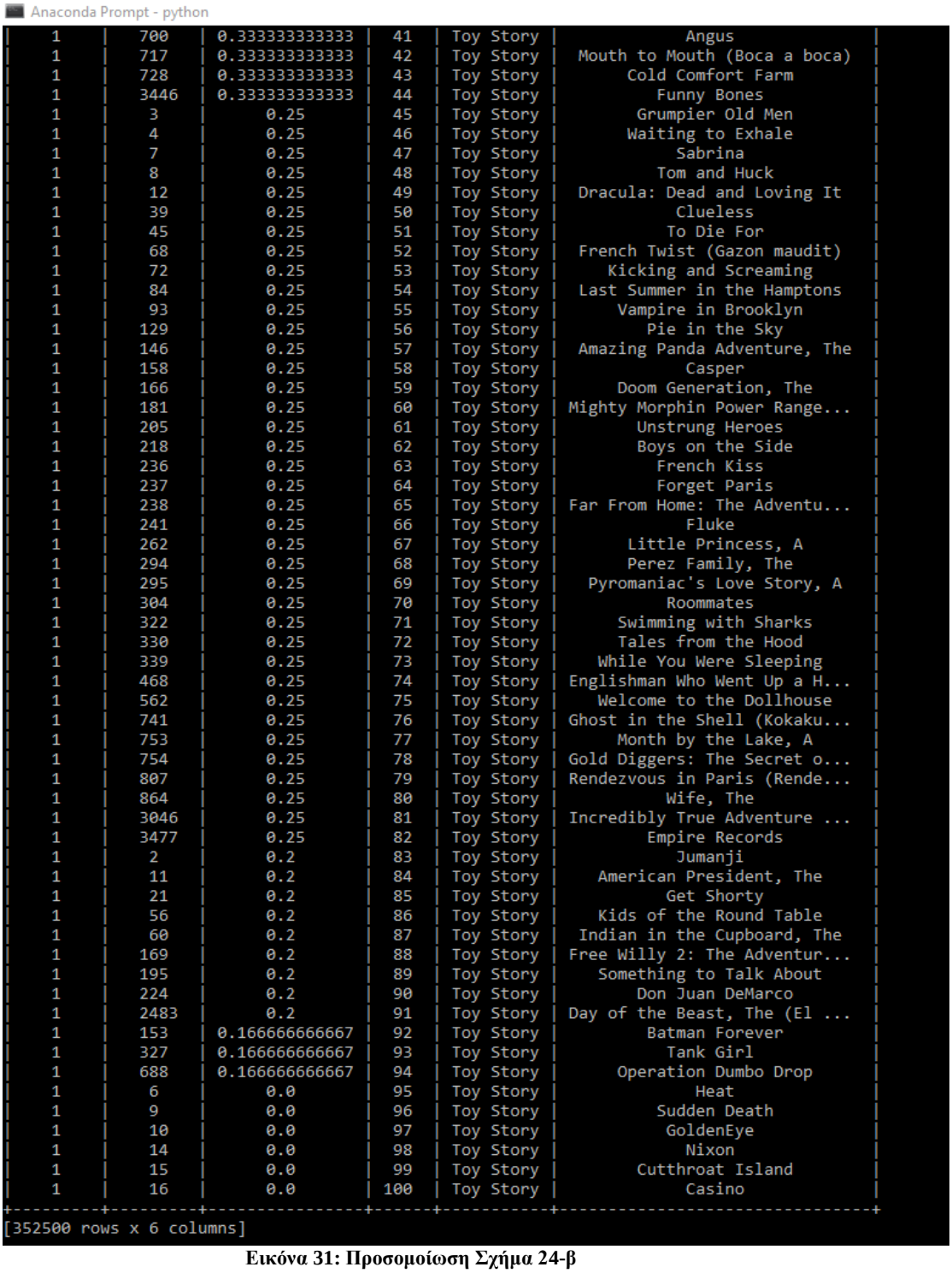
```
>>> dist = [[['genres'], 'jaccard', 1.0],
...         [['year'], 'euclidean', 1.0]]
>>> nn_model = gl.nearest_neighbors.create(items, 'item_id', distance=dist)
Defaulting to brute force instead of ball tree because there are multiple distance components.
Starting brute force nearest neighbors model training.
```

Εικόνα 29: Προσομοίωση Σχήμα 23

Χρησιμοποιούμε brute force διότι υπάρχουν αρκετοί συντελεστές ομοιότητας. Αφού το εξασφαλίσουμε αυτό δημιουργούμε την γειτονιά και αρχίζουμε την ομαδοποίηση. Στο σενάριο που ακολουθεί φτιάξαμε την γειτονιά που αντιστοιχεί στην ταινία με id 1, το Toy Story. Αποτέλεσμα αυτού είναι 3.525 ζεύγη και αποτυπώνεται στα Σχήματα 24-α και 24-β.

```
>>> gl.nearest_neighbors.create
<function create at 0x0000000009B6A4A8>
>>> similar = nn_model.query(items, 'item_id', k=100)\
...              .rename({'query_label': 'item_id', 'reference_label': 'similar', 'distance': 'score'})\
...              .join(items[['item_id', 'title']], on='item_id')\
...              .join(items[['item_id', 'title']], on={'similar': 'item_id'})
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 1            | 3525    | 0.0283688   | 2.987ms      |
| Done         |         | 100         | 207.95ms     |
+--------------+---------+-------------+--------------+
>>> similar['score'] = 1 - similar['score']
>>> similar.print_rows(100, max_row_width=200)
+---------+---------+----------------+------+-----------+-------------------------------+
| item_id | similar |     score      | rank |   title   |           title.1             |
+---------+---------+----------------+------+-----------+-------------------------------+
|    1    |    1    |      1.0       |  1   | Toy Story |           Toy Story           |
|    1    |   239   |      0.75      |  2   | Toy Story |         Goofy Movie, A        |
|    1    |    13   | 0.666666666667 |  3   | Toy Story |            Balto              |
|    1    |    54   | 0.666666666667 |  4   | Toy Story |        Big Green, The         |
|    1    |   888   | 0.666666666667 |  5   | Toy Story | Land Before Time III: The ... |
|    1    |    34   |      0.5       |  6   | Toy Story |              Babe             |
|    1    |   745   |      0.5       |  7   | Toy Story |        Close Shave, A         |
|    1    |    48   |      0.4       |  8   | Toy Story |          Pocahontas           |
|    1    |    5    | 0.333333333333 |  9   | Toy Story |  Father of the Bride Part II  |
|    1    |    19   | 0.333333333333 |  10  | Toy Story | Ace Ventura: When Nature Calls |
|    1    |    38   | 0.333333333333 |  11  | Toy Story |         It Takes Two          |
|    1    |    52   | 0.333333333333 |  12  | Toy Story |        Mighty Aphrodite       |
|    1    |    69   | 0.333333333333 |  13  | Toy Story |             Friday            |
|    1    |    96   | 0.333333333333 |  14  | Toy Story |    In the Bleak Midwinter     |
|    1    |   119   | 0.333333333333 |  15  | Toy Story |    Steal Big, Steal Little    |
|    1    |   144   | 0.333333333333 |  16  | Toy Story |    Brothers McMullen, The     |
|    1    |   156   | 0.333333333333 |  17  | Toy Story |       Blue in the Face        |
|    1    |   171   | 0.333333333333 |  18  | Toy Story |           Jeffrey             |
|    1    |   174   | 0.333333333333 |  19  | Toy Story |           Jury Duty           |
|    1    |   176   | 0.333333333333 |  20  | Toy Story |      Living in Oblivion       |
|    1    |   180   | 0.333333333333 |  21  | Toy Story |           Mallrats            |
|    1    |   186   | 0.333333333333 |  22  | Toy Story |          Nine Months          |
|    1    |   187   | 0.333333333333 |  23  | Toy Story |          Party Girl           |
|    1    |   189   | 0.333333333333 |  24  | Toy Story |           Reckless            |
|    1    |   203   | 0.333333333333 |  25  | Toy Story | To Wong Foo, Thanks for Ev... |
|    1    |   212   | 0.333333333333 |  26  | Toy Story |         Bushwhacked           |
|    1    |   216   | 0.333333333333 |  27  | Toy Story |        Billy Madison          |
|    1    |   228   | 0.333333333333 |  28  | Toy Story |  Destiny Turns on the Radio   |
|    1    |   258   | 0.333333333333 |  29  | Toy Story | Kid in King Arthur's Court, A |
|    1    |   274   | 0.333333333333 |  30  | Toy Story |       Man of the House        |
|    1    |   278   | 0.333333333333 |  31  | Toy Story |        Miami Rhapsody         |
|    1    |   310   | 0.333333333333 |  32  | Toy Story |          Rent-a-Kid           |
|    1    |   312   | 0.333333333333 |  33  | Toy Story |    Stuart Saves His Family    |
|    1    |   325   | 0.333333333333 |  34  | Toy Story | National Lampoon's Senior Trip |
|    1    |   333   | 0.333333333333 |  35  | Toy Story |          Tommy Boy            |
|    1    |   343   | 0.333333333333 |  36  | Toy Story |    Baby-Sitters Club, The     |
|    1    |   467   | 0.333333333333 |  37  | Toy Story |       Live Nude Girls         |
|    1    |   585   | 0.333333333333 |  38  | Toy Story |    Brady Bunch Movie, The     |
|    1    |   603   | 0.333333333333 |  39  | Toy Story |        Bye Bye, Love          |
|    1    |   633   | 0.333333333333 |  40  | Toy Story |       Denise Calls Up         |
|    1    |   700   | 0.333333333333 |  41  | Toy Story |            Angus              |
|    1    |   717   | 0.333333333333 |  42  | Toy Story |  Mouth to Mouth (Boca a boca) |
|    1    |   728   | 0.333333333333 |  43  | Toy Story |      Cold Comfort Farm        |
```

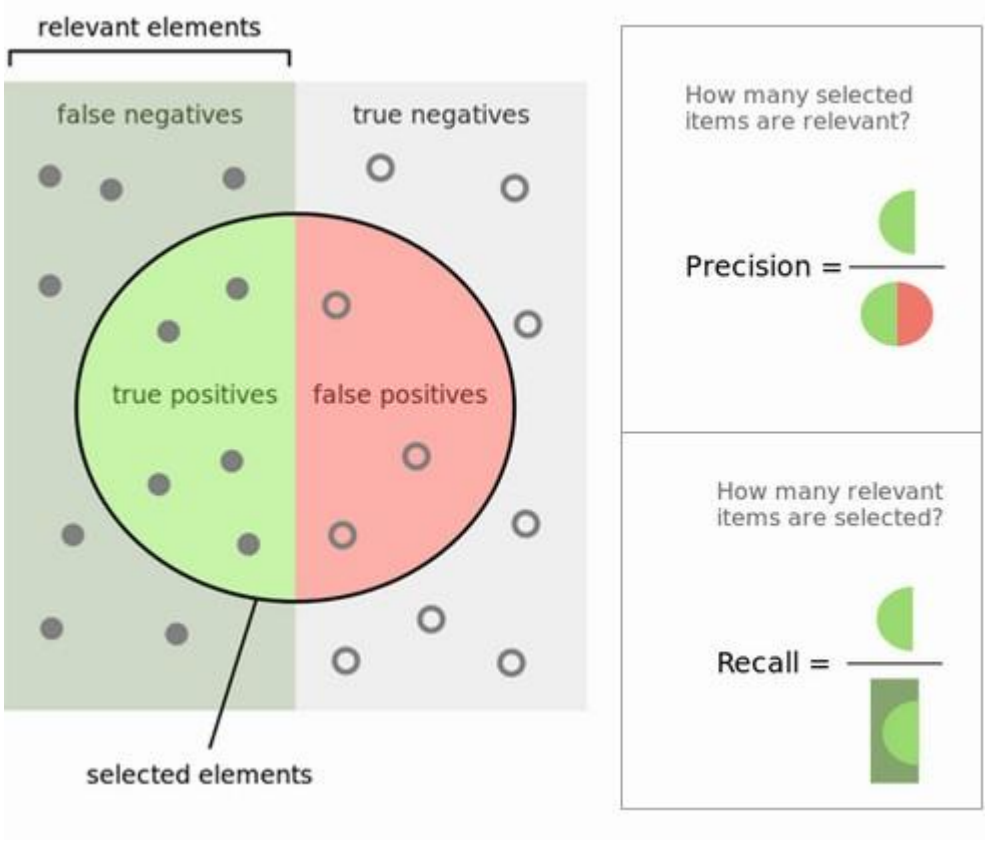
Εικόνα 30: Προσομοίωση Σχήμα 24-α



Εικόνα 31: Προσομοίωση Σχήμα 24-β

Συμπεράσματα

Στα πλαίσια της αναγνώρισης προτύπων γνωρίζουμε πως η μέση ακρίβεια υπολογίζει την σχετικότητα του αποτελέσματος, ενώ η μέση ανάκληση υπολογίζει το πιο ποσοστό από τα σχετικά αποτελέσματα αποτελεί ορθή σύσταση(Εικόνα 32).



Εικόνα 32: Επεξήγηση ακρίβειας και συνάφειας

Παρατηρούμε με βάση τη συγκριτική μελέτη(Σχήμα 22) που διεξήγαμε ανάμεσα στα τέσσερα μοντέλα ότι σε πρώτο βαθμό η μέση ακρίβεια μειώνεται και η μέση ανάκληση αυξάνεται, όσο περισσότερα στάδια αποκοπής έχουμε. Επιπλέον, παρατηρούμε ότι το m3, όντας το μοντέλο με την πιο ακριβής περιγραφή, είναι αυτό το οποίο έχει την μεγαλύτερη μέση ακρίβεια και μέση ανάκληση. Αυτό βασίζεται περισσότερο στο ότι όσο περισσότεροι παράγοντες λαμβάνονται υπόψη στο φιλτράρισμα τόσο περισσότερο αυξάνεται η μέση ακρίβεια και η μέση ανάκληση, το οποίο είναι και λογικό διότι όσο πιο αναλυτικό είναι το ερώτημα, τόσο πιο εξειδικευμένο είναι το αποτέλεσμα.

Μελλοντικές Κατευθύνσεις

Έχοντας παρουσιάσει αρκετά προβλήματα σχετικά με τις μηχανές συστάσεων, το σημαντικότερο συνεχίζει να αποτελεί η αποτελεσματική πρόβλεψη της αξιολόγησης των αντικειμένων από τους χρήστες. Οι τρέχουσες έρευνες για την βελτίωση αυτής της πρόβλεψης εστιάζει περισσότερο στον συνδυασμό διαφορετικών προσεγγίσεων. Σε αυτό το πεδίο αρκετές συνολικές μέθοδοι χρησιμοποιούνται(ensemble methods) έχουν προταθεί(Paterek, 2007 και Takacs et al., 2007). Συγκεκριμένοι συνδυασμοί διαφορετικών συστημάτων συστάσεων έχουν επίσης προταθεί, όπως και αλγόριθμοι που είναι ικανοί να κατανοήσουν νέες μετρικές ομοιότητας αντί να χρησιμοποιούν τις προκαθορισμένες(Bell et al.,2007). Λαμβάνοντας υπόψη την ρυθμική εξέλιξη των αξιολογήσεων σε ένα λογαριασμό, όπου οι πιο πρόσφατες αξιολογήσεις έχουν μεγαλύτερη βαρύτητα από τις νεότερες, κατορθώνουμε να βελτιώσουμε τις προβλέψεις.

Οι επιδόσεις των καλύτερων αλγορίθμων συστάσεων με γνώμονα την ακρίβεια σχεδόν ταυτίζονται και δεν μπορούμε να χαρακτηρίσουμε κάποια ως βέλτιστη. Στην πράξη αφού οι χρήστες δεν μπορούν να εντοπίσουν τις διαφορές είναι καλύτερο να εστιάσουμε σε άλλους παράγοντες, πλην της ακρίβειας, που προσπαθούν να εστιάσουν στην ποιότητα και την χρηστικότητα(για παράδειγμα η κάλυψη, η αλγοριθμική πολυπλοκότητα, η κλιμακωσιμότητα, η προσαρμοστικότητα, η εμπιστοσύνη και η ικανοποίηση του χρήστη, Herlocker et al.,2004).

Τα πειράματα με πραγματικούς χρήστες μας δείχνουν, ότι τα συστήματα πρέπει να οδηγούν και να ενθαρρύνουν τους χρήστες στην δημιουργία και οικοδόμηση των προφίλ τους. Το πρόβλημα αυτό ενθάρρυνε μια νέα έρευνα η οποία σχετίζεται με το επιστημονικό αντικείμενο της ενεργής μάθησης(Active Learning, Cohn et., 1996). Άλλα θέματα που αξίζει να ερευνηθούν είναι το ποια αντικείμενα ή πιο είδος αντικειμένων πρέπει να αξιολογηθεί για να βελτιώσει την ακρίβεια των συστημάτων συνεργατικού φιλτραρίσματος, καθώς και ποιες ιδιότητες των αντικειμένων είναι πιο κομβικά για τις συστάσεις με βάση το περιεχόμενο. Όπως είναι φυσικό με τον ίδιο τρόπο διεγείρεται το ερώτημα του πώς να σχεδιαστεί μια αποτελεσματική διεπαφή χρήστη για μία τέτοια αλληλεπίδραση.

Επίλογος

Στο πεδίο των συστημάτων συστάσεων, οι συνεργατικές μέθοδοι φιλτραρίσματος συχνά έχουν καλύτερη προγνωστική απόδοση απ'ότι αυτές με βάση το περιεχόμενο, τουλάχιστον όταν δεν υπάρχουν αρκετά δεδομένα αξιολόγησης. Η σύσταση αντικείμενων με βάση τις προτιμήσεις των χρηστών είναι πιο σημαντική από τις πληροφορίες που περιγράφουν το αντικείμενο.

Από την άλλη το φιλτράρισμα με βάση το περιεχόμενο προτείνει λύσεις στους περιορισμούς του συνεργατικού φιλτραρίσματος και παρέχει ένα πιο φυσικό τρόπο επαφής με τους χρήστες. Επιπλέον το θέμα της σχεδίασης φιλικών προς το χρήστη διεπαφών δεν πρέπει να υποτιμηθεί, καθώς οι χρήστες πρέπει να νιώθουν ότι ελέγχουν τις συστάσεις τους και να μπορούν να αποφύγουν άβολες καταστάσεις. Αλλιώς, ακόμα και αν το σύστημα είναι ακριβές στις προβλέψεις του, μπορεί να υποστεί το πρόβλημα της μιας επίσκεψης, όπου οι χρήστες το αποφεύγουν διότι το σύστημα τους εμποδίζει από το να εκφράσουν ορισμένες προτιμήσεις τους ή θεωρούν πως το σύστημα πάσχει ως προς την ευελιξία του. Μια διασκεδαστική επικοινωνία που αντέχει στο χρόνο είναι κομβική για να δοθούν καλές συστάσεις.

Όσων αφορά τις προσεγγίσεις συνεργατικού φιλτραρίσματος, αυτές που είναι βασισμένες στο αντικείμενο δείχνουν να αποδίδουν καλύτερα απέναντι σε αυτές που είναι βασισμένες στους χρήστες. Εκτός από τα καλά αποτελέσματα, οι συστάσεις βασισμένες στο αντικείμενο έχουν ταχύτερους χρόνους εκμάθησης και πρόβλεψης, τουλάχιστον για ομάδες δεδομένων που διαθέτουν περισσότερους χρήστες από αντικείμενα και είναι ικανές να παράγουν σχετικές προβλέψεις με την πρώτη αξιολόγηση ενός χρήστη. Επιπλέον τέτοια μοντέλα είναι κατάλληλα για πλοήγηση σε καταλόγους αντικειμένων ακόμα και αν δεν έχουμε διαθέσιμες πληροφορίες για τον τρέχον χρήστη, αφού μπορούν να παρουσιάσουν ένα χρήστη από μια κοντινή γειτονιά και από ένα αντικείμενο που ενδιαφέρεται εκείνος. Τέλος το γράφημα της γειτονιάς μπορεί να εξαχθεί σε συστήματα που εκμεταλλεύονται τις ομοιότητες των αντικειμένων χωρίς να εκθέτουν την ιδιωτικότητα του χρήστη.

Ένα σημαντικό θέμα που αφορά τα συστήματα συστάσεων είναι η εξερεύνηση κριτηρίων, που προσπαθούν να εξασφαλίσουν την ποιότητα και την χρησιμότητα των συστάσεων από την πλευρά της ικανοποίησης του χρήστη, όπως η κάλυψη, η αλγοριθμική πολυπλοκότητα, η καινοτομία, η προσαρμοστικότητα, η εμπιστοσύνη και η διεπαφή με τον χρήστη.

Τα συστήματα συστάσεων έχουν διευρύνει τις επιλογές για αναζήτηση και φιλτράρισμα πληροφοριών. Διαδικτυακά καταστήματα έχουν αυξημένα κέρδη από αυτά και με το τόσο ευρύ φάσμα επιλογών οι πελάτες έχουν ελαττώσει το χρόνο που ξόδευαν στις αγορές τους λόγω των συστάσεων. Παράλληλα όμως υπάρχουν προβλήματα, περιορισμοί και αποκλίσεις. Χρειάζονται αρκετές βελτιώσεις στο κομμάτι της ανάπτυξης του μοντέλου του χρήστη, με την βελτιστοποίηση της σημασιολογικής ανάλυσης της πληροφορίας καθώς και της επιτάχυνσης και της εκκαθάρισης της σύστασης. Τα συστήματα συστάσεων δεν περιορίζονται μόνο από τους υπολογιστές και τις κινητές συσκευές, αλλά μπορούν επιπλέον να ανοίξουν νέες δυνατότητες ασφαλείας κυρίως όταν ενσωματωθούν στην αυτοκινητοβιομηχανία ή σε άλλες συσκευές που χρησιμοποιούμε καθημερινά. Αυτό με τη σειρά του θα χρειαστεί ανάπτυξη πιο εξειδικευμένων συστημάτων σύστασης. Όλα αυτά τα γεγονότα μας δείχνουν πως οι μηχανές συστάσεων έχουν αρκετό δρόμο μπροστά τους, και εμείς απλά βρισκόμαστε στο αρχικό στάδιο της ανάπτυξής τους.

Βιβλιογραφία

- A. Elgohary, H. Nomir, I. Sabek, M. Samir, M. Badawy, and N. A. Yousri, "Wiki-rec: A semantic-based recommendation system using wikipedia as an ontology," in Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on, 29 2010.
- A. Gomez-Perez, O. Corcho-Garcia, and M. Fernandez-Lopez, Ontological Engineering. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2003.
- A. Sieg, B. Mobasher, and R. Burke, "Improving the effectiveness of collaborative recommendation with ontology-based user profiles," in Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, ser. HetRec '10. New York, NY, USA: ACM, 2010, pp. 39–46. [Online]. Available: http://doi.acm.org/10.1145/1869446.1869452
- Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734–749.
- B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web, ser. WWW '01. New York, NY, USA: ACM, 2001, pp. 285–295. [Online]. Available: http://doi.acm.org/10.1145/371920.372071
- Balabanovic, M., & Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. Communications of the ACM, 40(3), 66-72.
- Bell, R., Koren, Y., & Volinsky, C. (2007). Modeling relationships at multiple scales to improve accuracy of large recommender systems. In 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 95–104). New York, NY, USA. ACM.
- Breese, J., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In 14th Conference on Uncertainty in Artificial Intelligence (pp. 43–52). Morgan Kaufman.
- C.-Z. Yang, I.-X. Chen, and P.-J. Wu, "Cross-lingual news group recommendation using cluster-based cross-training." Taiwan: Computational Linguistics and Chinese Language Processing, 2008, pp. 41–60.
- Cilibrasi, R., & Vitanyi, P.M.B. (2007). The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 19(3), 370-383.
- Cohn, D. A., Ghahramani, Z., & Jordan M. I. (1996). Active Learning with Statistical Models. In Journal of Artificial Intelligence Research, 4, 129-145.
- Deshpande, M., & Karypis, G. (2004). Itembased top-N recommendation algorithms. In ACM Transactions on Information Systems, 22(1), 143–177.
- E. Peis, J. M. Morales-del Castillo, and J. A. Delgado-Lpez, "Semantic recommender systems. analysis of the state of the topic." Seoul, South Korea: Department of Computer Science, Yonsei Univerisity, 2008.
- G. Adomavicius and A. Tuzhilin, "Context-aware recommender systems," in Recommender Systems Handbook: A Complete Guide for Research Scientists and Practitioners. Springer, 2010.
- H. K. Farsani and M. Nematbakhsh, "A semantic recommendation procedure for electronic product catalog," 2006.

- Han, E.-H. S., & Karypis, G. (2005). Feature-based recommendation system. In 14th Conference of Information and Knowledge Management (pp. 446-452).
- Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. In ACM Transactions on Information Systems, 22(1), 5–53.
- J. Porter, "Folksonomies: A User-Driven Approach to Organizing Content," Apr. 2005. [Online]. Available: http://www.uie.com/articles/folksonomies/
- Jack, K., & Duclaye, F. (2007). Etude de la pertinence de critères de recherche en recherche d'informations sur des données structurées. In PeCUSI, INFORSID (pp. 285-297). PerrosGuirec, France.
- Jack, K., & Duclayee, F. (2008). Improving Explicit Preference Entry by Visualising Data Similarities. In Intelligent User Interfaces, International Workshop on Recommendation and Collaboration (ReColl). Spain.
- Karypis, G. (2001). Evaluation of item-based topN recommendation algorithms. In 10th International Conference on Information and Knowledge Management (pp. 247–254).
- Kelleher, J., & Bridge, D. (2003). Rectree centroid : An accurate, scalable collaborative recommender. In Cunningham, P., Fernando, T., & Vogel, C. (Ed.), 14th Irish Conference on Artificial Intelligence and Cognitive Science (pp. 89–94).
- Kleinberg, J., & Sandler, M. (2004). Using mixture models for collaborative filtering. In 36th ACM Symposium on Theory of Computing (pp. 569–578). ACM Press.
- L. Baltrunas and F. Ricci, "Context-dependent items generation in collaborative filtering," 2009.
- L. Baltrunas and X. Amatriain, "Towards time-dependant recommendation based on implicit feedback," 2009.
- Lam, S. K., Frankowski, D., & Riedl, J. (2006). Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In International Conference on Emerging Trends in Information and Communication Security.
- Lin, W., Alvarez, S., & Ruiz, C. (2002). Efficient adaptive-support association rule mining for recommender systems. In Data Mining and Knowledge Discovery, 6, 83–105.
- Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. In IEEE Internet Computing, 7(1), 76–80.
- M. Magableh, A. Cau, H. Zedan, and M. Ward, "Towards a multilingual semantic folksonomy," in Proceedings of the IADIS International Conferences Collaborative Technologies 2010 and Web Based Communities 2010, July 2010, pp. 178–182.
- M. van Setten, S. Pokraev, and J. Koolwaaij, "Context-aware recommendations in the mobile tourist application compass," in Adaptive Hypermedia and Adaptive Web-Based Systems, ser. Lecture Notes in Computer Science, P. De Bra and W. Nejdl, Eds. Springer Berlin / Heidelberg, 2004, vol. 3137, pp. 515–548, 10.1007/978 − 3 − 540 − 27780 − 427. [Online]. Available: http : //dx.doi.org/10.1007/978 − 3 − 540 − 27780 – 427
- Melville, P., Mooney, R., & Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. In 18th National Conference on Artificial Intelligence (pp. 187-192).

- Mooney, R., & Roy, L. (1999). Content-based book recommending using learning for text categorization. In ACM SIGIR'99, Workshop on Recommender Systems: Algorithms and Evaluation
- Nageswara Rao, K., & Talwar, V.G. (2008). Application domain and functional classification of recommender systems a survey. In Desidoc journal of library and information technology, vol 28, n°3, 17-36.
- Nguyen, A., Denos, N., & Berrut, C. (2007). Improving new user recommendations with rulebased induction on cold user data. In RecSys2007 (pp. 121-128).
- P. Lops, C. Musto, F. Narducci, M. De Gemmis, P. Basile, and G. Semeraro, "Mars: a multilanguage recommender system," in Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, ser. HetRec '10. New York, NY, USA: ACM, 2010, pp. 24–31. [Online]. Available: http://doi.acm.org/10.1145/1869446.1869450
- P. Resnick and H. R. Varian, "Recommender systems," Commun. ACM, vol. 40, pp. 56–58, March 1997. [Online]. Available: http://doi.acm.org/10.1145/245108.245121
- Paterek A. (2007). Improving regularized singular value decomposition for collaborative filtering. In KDD cup Workshop at SIGKDD.
- Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. In Artificial Intelligence Review, 13(5-6), 393–408.
- Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. In Machine Learning, 27, 313–331.
- Pazzani, M., & Billsus, D. (2007). Content-Based Recommendation Systems. In The Adaptive Web, 325-341.
- Pennock, D., Horvitz, E., Lawrence, S., & Giles, C. L. (2000). Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. In 16th Conference on Uncertainty in Artificial Intelligence (pp. 473–480).
- Polcicova, G., Slovak, R., & Navrat, P. (2000). Combining content-based and collaborative filtering. In ADBIS-DASFAA Symposium (pp. 118-127).
- R. Baraglia, M. Mordacchini, P. Dazzi, and L. Ricci, "A p2p recommender system based on gossip overlays (prego)," in Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, 2010, pp. 83 –90.
- R. Kanawati and H. Karoui, "A p2p collaborative bibliography recommender system," in Proceedings of the 2009 Fourth International Conference on Internet and Web Applications and Services. Washington, DC, USA: IEEE Computer Society, 2009, pp. 90–96. [Online]. Available: http://portal.acm.org/citation.cfm?id=1585689.1586218
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In Conference on Computer Supported Cooperative Work (pp. 175–186). ACM.
- S. Aciar, "Mining context information from consumers reviews," 2010
- S. Berkovsky, T. Kuflik, and F. Ricci, "Cross-domain mediation in collaborative filtering," in Proceedings of the 11th international conference on User Modeling, ser. UM '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 355–359. [Online]. Available: http : //dx.doi.org/10.1007/978 − 3 − 540 − 73078 – 144

- Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. In ACM Conference on Electronic Commerce (pp. 158–167).
- T. Bogers, "Movie recommendation using random walks over the contextual graph," 2010.
- T. Gruber. (2005) Ontology of Folksonomy: A Mash-up of Apples and Oranges. First on-Line conference on Metadata and Semantics Research (MTSR'05). [Online]. Available: http://tomgruber.org/writing/mtsr05-ontology-of-folksonomy.htm
- T. hun Kima, J. woo Song, and S. bong Yang, "L-prs: A location-based personalized recommender system." Seoul, South Korea: Department of Computer Science, Yonsei Univerisity
- Takacs, G., Pilaszy, I., Nemeth, B., & Tikk, D. (2007). On the gravity recommendation system. In KDD cup Workshop at SIGKDD.
- Ungar, L., & Foster, D. (1998). Clustering methods for collaborative filtering. In Workshop on Recommendation Systems. AAAI Press.
- Vozalis, M., & Margaritis, K. G. (2004). Enhancing collaborative filtering with demographic data: The case of item-based filtering. In 4th International Conference on Intelligent Systems Design and Applications (pp. 361–366).
- W. Woerndl and J. Schlichter, "Introducing context into recommender systems." Muenchen, Germany: Technische Universitaet Muenchen, pp. 138–140.
- Wang, J., de Vries, A. P., & Reinders, M. J. (2006). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In 29th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 501-508).
- Y. Blanco-Fernandez, M. Lopez-Nores, J. Pazos-Arias, A. Gil-Solla, and M. Ramos-Cabrer, "Exploiting digital tv users' preferences in a tourism recommender system based on semantic reasoning," in Consumer Electronics (ICCE), 2010 Digest of Technical Papers International Conference on, 2010, pp. 139 –140.
- Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In ACM SIGIR'02.

Σχετικά με τη βάση δεδομένων:

- F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI= http://dx.doi.org/10.1145/2827872

Παράρτημα

Οδηγίες Εγκατάστασης Turi/GraphLab

Βήμα 1: Κατεβάζουμε το Anaconda2(έκδοση 4.0.0)
Βήμα 2: Εγκαθιστούμε το Anaconda
Βήμα 3: Δημιουργούμε περιβάλλον conda:

```
# Δημιουργούμε περιβάλλον conda σε Python 2.7.x
conda create -n gl-env python=2.7 anaconda=4.0.0

#Ενεργοποιούμε το περιβάλλον που δημιουργήσαμε
activate gl-env
```

Βήμα 4: Εξασφαλίζουμε ότι το pip είναι τουλάχιστον στην έκδοση 7:

```
conda update pip
```

Βήμα 5:Εγκαθηστούμε το GraphLab Create
Για να γίνει το ακόλουθο βήμα πρέπει να εγγραφούμε συμπληρώνοντας την φόρμα (https://turi.com/download/academic.html). Στην συνέχεια από το mail εγγραφής μας θα μας σταλεί ένα product key. Στην συνέχεια στην κονσόλα του Anaconda πληκτρολογούμε το εξής:

```
pip install --upgrade --no-cache-dir https://get.graphlab.com/GraphLab-Create/2.1/your registered email address here/your product key here/GraphLab-Create-License.tar.gz
```

Βήμα 6: Εξασφαλίζουμε την εγκατάσταση του IPython(πλεον Jupyter)

```
conda install ipython-notebook
```

Απαιτήσεις Συστήματος για Turi/GraphLab

Οι απαιτήσεις είναι οι εξής:

- Windows 7 ή μεταγενέστερα ή Windows 2012 R2
- 64-bit αρχιτεκτονική επεξεργαστή
- Τουλάχιστον 4GB RAM
- Τουλάχιστον 2GB ελεύθερου χώρου στον σκληρό δίσκο

Σε περίπτωση που το σύστημα του χρήστη δεν καλύπτει τις απαιτήσεις, σε περίπτωση εξασφάλισης ακαδημαϊκής άδειας το GraphLab σου δίνει την δυνατότητα να δουλέψεις με χρήση της υπηρεσίας Amazon Web Services.