ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Υβριδική προσαρμοστική μέθοδος πρόβλεψης μεταβαλλόμενων προτιμήσεων χρηστών

Αλέξανδρος Μ. Μουζακίδης

Επιβλέποντες: Ιωάννης Ιωαννίδης, Καθηγητής ΕΚΠΑ
Γεώργιος Παλιούρας, Ερευνητής Β΄ ΕΚΕΦΕ “Δημόκριτος”

ΑΘΗΝΑ

ΜΑΪΟΣ 2010

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Υβριδική προσαρμοστική μέθοδος πρόβλεψης μεταβαλλόμενων προτιμήσεων χρηστών

Αλέξανδρος Μ. Μουζακίδης

Α.Μ.: Μ853

ΕΠΙΒΛΕΠΟΝΤΕΣ: Ιωάννης Ιωαννίδης, Καθηγητής ΕΚΠΑ
Γεώργιος Παλιούρας, Ερευνητής Β΄ ΕΚΕΦΕ "Δημόκριτος"

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: Ιωάννης Ιωαννίδης, Καθηγητής ΕΚΠΑ

ΜΑΪΟΣ 2010

ΠΕΡΙΛΗΨΗ

Καθημερινά επισκεπτόμαστε μεγάλες διαδικτυακές πύλες αλλά και μεγάλα πολυκαταστήματα που παρέχουν συνεχώς νέα και πολλά σε αριθμό προϊόντα και υπηρεσίες στους επισκέπτες τους. Απαραίτητο για την σωστή λειτουργία των ηλεκτρονικών ή μη χώρων είναι να μπορούν να ξεχωρίσουν τι από όσα μπορούν να παρέχουν θα ενδιαφέρει έναν πελάτη τους ώστε να τον εξυπηρετήσουν καλύτερα. Αυτή την ανάγκη έρχονται να καλύψουν τα συστήματα συστάσεων των οποίων ο στόχος είναι να εντοπίζουν τις ανάγκες των χρηστών και να προτείνουν προϊόντα που θα τους αρέσουν. Για να μπορεί ένα σύστημα να κάνει τέτοιες συστάσεις, πρέπει να είναι σε θέση να εκτιμήσει το πόσο θα αρέσει ένα προϊόν στον εκάστοτε πελάτη-χρήστη. Στην παρούσα εργασία έγινε μελέτη πρόσφατων επιστημονικών δημοσιεύσεων πάνω στο αντικείμενο των συστημάτων συστάσεων και πρόβλεψης προτιμήσεων. Έγινε συνδυασμός από τεχνικές για την δημιουργία νέου υβριδικού προσαρμοστικού αλγορίθμου πρόβλεψης προτίμησης βασισμένο σε τεχνικές συστάσεων βάσει περιεχομένου, συνεργατικές και δημογραφικές μαζί με τεχνικές αντιμετώπισης της εννοιολογικής απόκλισης.

Χρησιμοποιούμε το ιστορικό ενός χρήστη που περιέχει την αλληλεπίδραση του με το σύστημα μέσα στον χρόνο και μέσω αυτού προσπαθούμε να βγάλουμε συμπεράσματα για το ποιες είναι οι προτιμήσεις και τα ενδιαφέροντα του. Για να μπορέσουμε να προβλέψουμε με μεγαλύτερη επιτυχία τις προτιμήσεις των χρηστών εξάγουμε πληροφορία από το ιστορικό του κάθε χρήστη ξεχωριστά, το προφίλ όμοιων χρηστών, τα δημογραφικά τους χαρακτηριστικά άλλα και τα δομικά χαρακτηριστικά των αντικειμένων που υπάρχουν στο σύστημα. Η μέθοδος εκτιμά την προτίμηση του χρήστη για κάθε δομικό χαρακτηριστικό ενός αντικειμένου και μέσω αυτών για το ίδιο το αντικείμενο. Οι εκτιμήσεις για τις προτιμήσεις των χαρακτηριστικών γίνονται είτε βάσει των αντικειμένων που περιέχουν το χαρακτηριστικό και για τα οποία γνωρίζουμε την προτίμηση του χρήστη, είτε από την προτίμηση που έχουν εκφράσει όμοιοι χρήστες. Από το ιστορικό του χρήστη εντοπίζουμε τα χαρακτηριστικά που είναι πιο σημαντικά για κάθε πρόβλεψη καθώς και το πως αλλάζουν οι προτιμήσεις που έχουν δοθεί στο παρελθόν.

Η καλή λειτουργία του αλγορίθμου που παράχθηκε επιβεβαιώθηκε με την πραγματοποίηση πειραμάτων πάνω σε πραγματικά δεδομένα χρηστών. Τα δεδομένα που χρησιμοποιήσαμε τα πήραμε από τα αρχεία καταγραφής της διαδικτυακής υπηρεσίας movielens. Τα αρχεία αυτά

περιέχουν βαθμολογίες ταινιών που έχουν δώσει χρήστες του συστήματος σε διάρκεια δύο ετών. Χωρίσαμε τις βαθμολογίες σε δύο σύνολα και προσπαθήσαμε να προβλέψουμε τις βαθμολογίες του δεύτερου βάσει των προφίλ που φτιάξαμε από το πρώτο.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Εξόρυξη Γνώσης από δεδομένα και Πρόβλεψη Προτιμήσεων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ Μηχανική Μάθηση, Μέθοδοι Φιλτραρίσματος, Μοντελοποίηση Χρηστών, Απόκλιση Ενδιαφέροντος, Εννοιολογική Απόκλιση, Στατιστική Σημαντικότητα.

ABSTRACT

Every day we visit big web portal and stores that provide continuously many new products and services in their customers. To improve the serve of these portals and stores is essential to distinguish what are the products that customers will like most. For this reason recommender systems exists, these information filtering systems attempts to items that ale likely to be of interest to the users. Typically these systems try to predict the “rating” that a user would give to an item that had not been yet considered. In this dissertation we have studied the recently scientific publications on the topic of recommender systems and user rate prediction. We have combine different techniques of information filtering to create a new hybrid adaptive algorithm that predicts rates based on content based, collaborative filtering , demographic filtering and concept drift detection.

We use the logged interaction between users and the system over the time and we try to detect what users likes and what is interesting for them. We extract content based, collaborative based and demographic based user profiles and we combine these to make estimations about how interesting and significant are the features of the items to any user. The new algorithm uses the content-based, collaborative based and demographic based estimations about item features to predict “rates” for items that are not yet considered. From the log files we extract the statistical significance about item features and past user rates to adapt to the current rates that a user would give.

To evaluate the performance of the new method we run some experiments on real user data. I have used the log files of the movielens service. These files contains movie rates that users have given in a two years long time period. We have separated these rates in tow sets and we have tried to predict the rates that exists on the second set based on the profiles that we had create from the first one.

SUBJECT AREA: Data Mining and User Rate Prediction

KEYWORDS: Machine Learning, Filtering Methods, Interest Drift, Concept Drift, User Profiling, User Adaptivity, Statistical Significance

Περιεχόμενα

Πρόλογος

Αυτή η διπλωματική εργασία εκπονήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος ειδίκευσής στην πληροφορική του τμήματος “Πληροφορικής και Τηλεπικοινωνιών” του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών και την κατεύθυνση της “Υπολογιστικής Επιστήμης”. Το θέμα της εργασίας είναι μια πρόταση του εργαστηρίου “Τεχνολογίας Γνώσεων και Λογισμικού” του Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών «Δημόκριτος» και πραγματοποιήθηκε υπό την επίβλεψη του ερευνητή κ. Γιώργου Παλιούρα. Σε αυτό το σημείο οφείλω να ευχαριστήσω τον κ. Γιώργο Παλιούρα για την καθοδήγηση του και την ελευθερία κινήσεων που μου παρείχε κατά την διάρκεια της συνεργασίας μας. Επίσης οφείλω να ευχαριστήσω τον καθηγητή του τμήματος “Πληροφορικής και Τηλεπικοινωνιών” του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών κ. Ιωαννίδη Γιάννη που μου έδωσε την ευκαιρία να ασχοληθώ με την παρούσα εργασία και που με συμπεριέλαβε στην ομάδα ανάπτυξης του MADGIK μέσω της οποίας μου δόθηκε η ευκαιρία να ασχοληθώ με πολλά θέματα έρευνας και ανάπτυξης πάνω στον τομέα της εξατομίκευσης. Τέλος πρέπει να ευχαριστήσω τους συμφοιτητές και συμφοιτήτριες μου Βαγιανού Μαρία, Ιατροπούλου Κατερίνα, Μαριαλένα Κυριακίδη, Σταματογιαννάκη Λευτέρη, Μέι Λι Τριανταφυλίδη με τους οποίους συνεργάστηκα στην ομάδα του MADGIK και οι συζητήσεις που είχαμε βοήθησαν στο να γεννηθούν ιδέες που εφαρμόστηκαν σε αυτή την εργασία.

Το κείμενο της εργασίας έχει οργανωθεί σε πέντε κεφάλαια. Στο πρώτο κεφάλαιο γίνεται μια περιγραφή των συστημάτων συστάσεων και των υπηρεσιών που παρέχουν. Στο κεφάλαιο δύο γίνεται αναφορά σε αλγόριθμους και τεχνικές πρόβλεψης προτιμήσεων που έχουν δημοσιευτεί σε επιστημονικά περιοδικά και συνέδρια καθώς και αντιστοιχείς αναφορές σε αλγορίθμους και τεχνικές που έχουν δημοσιευτεί και ασχολούνται με το θέμα της εννοιολογικής απόκλιση. Στο κεφάλαιο τρία περιγράφεται η υβριδική προσαρμοστική μέθοδος προβλέψεων βαθμολογιών που προτείνουμε και τέλος στα κεφάλαια τέσσερα και πέντε τα πειραματικά αποτελέσματα που πετυχαίνει η εφαρμογή της στην πρόβλεψη προτιμήσεων ταινιών και τα τελικά συμπεράσματα.

1 Συστήματα Συστάσεων

1.1. Λόγος ύπαρξης

Η εξάπλωση του διαδικτύου και η συνεχής αύξηση της παρεχόμενης πληροφορίας μέσω των μεγάλων διαδικτυακών πυλών δημιουργεί το λεγόμενο πρόβλημα της υπερπληροφόρησης (information overload**).** Ο όρος περιγράφει την δυσκολία που αντιμετωπίζουν οι χρήστες να αντιληφθούν τι βλέπουν στην οθόνη τους αλλά και να αποφασίσουν τι να επιλέξουν. Στην προσπάθεια της δημιουργίας πιο φιλικών προς τον χρήστη εφαρμογών και αντιμετώπισής του προβλήματος δημιουργήθηκαν τα συστήματα συστάσεων (Recommender systems). Τα συστήματα αυτά επιλέγουν ποίο από το διαθέσιμο περιεχόμενο (όπως cd, βιβλία, ταινίες, άρθρα κτλ) είναι πιθανό να φανεί χρησιμότερο στον χρήστη και το προτείνουν με την μορφή συστάσεων. Αυτά τα συστήματα είναι ιδιαιτέρως σημαντικά στις εφαρμογές ηλεκτρονικού εμπορίου αφού μπορούν να εντοπίσουν αντικείμενα που είναι πιθανό να αγοράσει ο χρήστης. Η υλοποίηση πετυχημένων συστημάτων συστάσεων έχουν προσφέρει ανταγωνιστικότητα σε πολλά ηλεκτρονικά καταστήματα και διαδικτυακές εφαρμογές γενικότερα.

Μερικές γνωστές εφαρμογές που χρησιμοποιούν συστήματα συστάσεων για να παράγουν εξατομικευμένο περιεχόμενο είναι.

- blinkx (ταινίες και τηλεόραση)
- Facebook (κοινωνική δικτύωση)
- The Filter (διασκέδαση και πληροφόρηση)
- Introanalytics (κοινωνική δικτύωση, ταινίες και τηλεόραση, ηλεκτρονικό εμπόριο)
- IMDB (πληροφορίες ταινιών)
- Jinni search engine (ταινίες και τηλεόραση)
- Rotten Tomatoes (ταινίες)
- TV Genius (τηλεόραση)
- Last.Fm (ηλεκτρονικό ραδιόφωνο)
- Amazone (ηλεκτρονικό εμπόριο)
- Google AdSense (εξατομικευμένο διαφημιστικό υλικό)
- Gravity Technologies (ταινίες και τηλεόραση)
- iTunes (εφαρμογή αναπαραγωγής μουσικής)

1.2. Προδιαγραφές ενός συστήματος συστάσεων

Οι βασικές λειτουργίες που πρέπει να υλοποιεί κάθε σύστημα παραγωγής συστάσεων είναι η επιλογή των k καταλληλότερων αντικειμένων, ο υπολογισμός ομοιότητας ανάμεσα στα αντικείμενα και η πρόβλεψη της προτίμησης του κάθε χρήστη με το κάθε αντικείμενο.

Επιλογή των k καταλληλότερων αντικειμένων: Η βασική λειτουργία ενός συστήματος συστάσεων είναι να μπορέσει από μια μεγάλη λίστα αντικειμένων να επιλέξει τα k σημαντικότερα που πρέπει να προβάλλει. Συνήθως τα συστήματα πρέπει να υπολογίζουν πολλές τέτοιες λίστες αντικειμένων για

- κάθε χρήστη του συστήματος, χωρίς στοιχεία εξατομίκευσης δηλαδή χωρίς να υπολογίζει τι αρέσει προσωπικά στον χρήστη
- κάθε ομάδα χρηστών με κοινά δημογραφικά χαρακτηριστικά
- τον εκάστοτε χρήστη εξατομικευμένα και βάσει τις πρόσφατες επιλογές του
- τον εκάστοτε χρήστη εξατομικευμένα και βάσει τις μακροχρόνιες συνήθειές του

Υπολογισμός ομοιότητας αντικειμένων: Με την συνεχή εισαγωγή νέων αντικειμένων για να μπορεί ένα σύστημα να επιλέξει ποια από αυτά είναι τα καταλληλότερα προς σύσταση πρέπει να μπορεί να τα βρει πόσο μοιάζουν με τα αντικείμενα που αρέσουν στις χρήστες. Επίσης πρέπει να μπορεί να ομαδοποιεί αντικείμενα που παρουσιάζουν κοινό ενδιαφέρον (πχ πωλούνται συχνά μαζί). Για να μπορεί λοιπόν ένα σύστημα να κάνει καλές συστάσεις πρέπει με κάποιο τρόπο να συγκρίνει ομοιότητα είτε βάσει των χαρακτηριστικών των αντικειμένων είτε βάσει του ενδιαφέροντος που παρουσιάζουν προς τους χρήστες.

Πρόβλεψη προτιμήσεων: Για να είναι επιτυχημένη μία λίστα προτεινόμενων αντικειμένων θα πρέπει να απαρτίζεται από αντικείμενα που εκτιμούμε ότι θα αρέσουν στον χρήστη. Άρα πρέπει να εκτιμηθεί το πόσο θα του άρεσε ή πως θα βαθμολογούσε ένα αντικείμενο για το οποίο δεν έχει εκφραστεί ρητή/άμεση προτίμηση.

1.3. Τα δεδομένα των συστημάτων συστάσεων

Για να υποστηρίζει τις λειτουργίες που αναφέρονται στην παράγραφο 1.2 τα συστήματα συστάσεων δέχονται ως εισόδου δεδομένα που αφορούν: τα χαρακτηριστικά των αντικειμένων, και τα προφίλ των χρηστών.

Τα χαρακτηριστικά των αντικειμένων αποτελούν όλα τα δεδομένα που περιγράφουν ένα αντικείμενο στον χρήστη. Αν για παράδειγμα έχουμε μια εφαρμογή ηλεκτρονικού εμπορίου τα χαρακτηριστικά ενός εμπορεύματος είναι όλα όσα εμφανίζονται στον χρήστη όπως το είδος (cd, βιβλία, παπούτσια κτλ), η εταιρία κατασκευής, η τιμή και ότι άλλο θέλουμε να του γνωστοποιηθεί.

Τα προφίλ των χρηστών αποτελούνται από ότι πληροφορίες μπορούμε να συλλέξουμε για αυτόν. Δημογραφικά χαρακτηριστικά όπως ηλικία, επάγγελμα, οικογενειακή κατάσταση κτλ το ιστορικό του από την χρήση του συστήματος και αν είναι δυνατόν το ιστορικό του από την χρήση άλλων συστημάτων. Το ιστορικό του χρήστη περιέχει δεδομένα που δηλώνουν άμεσα ή έμμεσα τις προτιμήσεις και συνήθως είναι αποθηκευμένα σε αρχεία ημερολογίου (log files). Οι άμεσες προτιμήσεις δηλώνονται με αριθμητικές τιμές όπως οι βαθμολογίες, με αποτελέσματα συγκρίσεων ή με σημασιολογικά δεδομένα (semantic information) όπως οι ετικέτες.

Οι έμμεσες προτιμήσεις δηλώνονται μέσω τον “κινήσεων” που έκανε ο χρήστης στο παρελθόν. Όσο χρησιμοποιούσε το σύστημα καταγράφονται το αντικείμενα που επέλεξε, τι αγόρασε, πού μετακίνησε τον δείκτη του ποντικιού και γενικότερα ότι επιλογές μπορούσε να πραγματοποιήσει μέσω της παρεχόμενης διεπαφής που του προσφέρεται. Σκοπός είναι από τα έμμεσα δεδομένα να γίνει εξαγωγή κάποιας αριθμητικής τιμής

1.4. Η αυστηρότητα κρίσης των συστάσεων

Πολύ σημαντικό θέμα για την κρίση των συστάσεων που παράγουν τα συστήματα παίζει το πεδίο εφαρμογής που χρησιμοποιούνται (ηλεκτρονικό εμπόριο, πληροφορική ιατρική, gps κτλ). Μία εσφαλμένη πρόταση για την αγορά ενός βιβλίου δεν κρίνεται τόσο αυστηρά όσο η

προτροπή λήψης ενός λανθασμένου φαρμάκου. Για αυτό και οι τεχνικές φιλτραρίσματος διαφοροποιούνται αναλόγως της θεματικής περιοχής και του πόσο αυστηρά ορθές πρέπει να είναι οι συστάσεις.

Τα συστήματα συστάσεων δεν έχουν εφαρμογή μόνο στις εφαρμογές διαδικτύου αλλά σε οποιονδήποτε χώρο όπου η χρήση έξυπνων υπολογιστικών συστημάτων μπορεί να βοηθήσει στην λήψη αποφάσεων, όπως για χρηματιστηριακές επενδύσεις, ταξιδιωτικά γραφεία και λοιπές επιχειρήσεις όπου ο πελάτης επιλέγει ανάμεσα σε πολλές επιλογές.

1.5. Το αντικείμενο της διπλωματικής εργασίας

Στόχος της παρούσας διπλωματικής εργασίας ήταν η μελέτη διαφόρων τεχνικών πρόβλεψης προτιμήσεων ώστε να δημιουργηθεί ένας αλγόριθμος που συνδυάζοντάς διαφορετικές ιδέες να πετύχει καλύτερα αποτελέσματα. Αρχικά έγινε μία μελέτη της απόδοσης που μπορούν να επιτύχουν οι διάφορες μέθοδοι προβλέψεων ξεχωριστά και πώς βελτιώνονται όταν εμπλουτιστούν με τεχνικές αντιμετώπισης της εννοιολογικής απόκλισης και συνδιαστούν τα αποτελέσματα τους. Συγκεκριμένα ελέγχθηκαν τα αποτελέσματα που μπορούν να επιτύχουν το φιλτράρισμα βάσει περιεχομένου, το δημογραφικό και το συνεργατικό φιλτράρισμα από μόνα τους και εμπλουτισμένα με τεχνικές προσαρμόζονται στις αλλαγές προτιμήσεων των χρηστών. Από τα πειραματικά αποτελέσματα της πρώτης φάσης βγήκαν συμπεράσματα για το ποιες είναι οι πιο αποδοτικές μέθοδοι και πώς αυτές μπορούν να συνδιαστούν. Η υβριδική προσαρμοστική μέθοδος φιλτραρίσματος που προκύπτει παρουσιάζει μεγαλύτερη ακρίβεια στις προβλέψεις της.

2 Πρόβλεψη προτιμήσεων

2.1. Το πρόβλημα

Όπως αναφέρθηκε στην παράγραφο 1.2 μία από τις βασικές λειτουργίες ενός συστήματος συστάσεων είναι η εκτίμηση της προτίμησης οποιουδήποτε χρήστη και οποιοδήποτε αντικείμενο του συστήματος. Έστω $U\{1,2,\ldots\ldots,N\}$ το σύνολο των χρηστών του συστήματος και $I\{1,2,\ldots\ldots,M\}$ το σύνολο των αντικειμένων που μπορούν να προταθούν. Έστω μία συνάρτηση $F:U\,x\,I\rightarrow R$ που δέχεται ως είσοδο έναν χρήστη και ένα αντικείμενο και επιστρέφει το πόσο αρεστό είναι αυτό το αντικείμενο για τον χρήστη. Το R είναι ένα ολικά διατεταγμένο σύνολο*.

Ο ορισμός της συνάρτησης F διαφέρει από εφαρμογή σε εφαρμογή. Για τα γνωστά αντικείμενα, η τιμής της συνάρτηση F μπορεί να υπολογιστεί από τις βαθμολογίες των αντικειμένων από τους χρήστες, μπορεί να είναι η συχνότητα που έχει επιλέξει ο κάθε χρήστης ένα αντικείμενο καθώς και κάθε στατιστικό ή σημασιολογικό μέτρο σύγκρισης, που μπορεί να οριστεί στο σύστημα. Στην ιδανική περίπτωση, η F θέλουμε να μπορεί να επιστρέψει μία τιμή για κάθε στοιχείο το συνόλου $U\,x\,I$, για παράδειγμα να γνωρίζουμε όλες τις βαθμολογίες των χρηστών για όλα τα αντικείμενα, αλλά κάτι τέτοιο στην πράξη δεν είναι πάντα εφικτό και ο υπολογισμός της για τα άγνωστα αντικείμενα πρέπει να γίνει με κάποια υπολογιστική μέθοδο που βασίζεται σε παρέκταση (extrapolation). Επίσης το σύνολο UxI μπορεί να είναι πάρα πολύ μεγάλο και για λόγους κλιμάκωσης να είναι απαγορευτική η εκτίμηση της F σε ολόκληρο το χώρο. Σε αυτή την περίπτωση θα πρέπει να γίνει προσεκτική επιλογή βάσει μιας δεύτερης συνάρτησης έστω $G:UxI\rightarrow S\subset UxI$ που θα επιλέξει για ποιους συνδυασμούς χρήστη-αντικείμενο αξίζει να γίνει η εκτιμηθεί της F. Δηλαδή η συνάρτηση G θα περιορίσει τον χώρο UxI.

Η παρέκταση της συνάρτησης F αλλά και της συνάρτησης G συνήθως βασίζονται σε ευριστικές μεθόδους που είτε επαληθεύονται εμπειρικά μέσω της παρατήρησης, είτε που ικανοποιούν τα κριτήρια κάποιας συνάρτησης βελτιστοποίησης. Για την υλοποίηση τους

*Ο ορισμός των συμβόλων για τα σύνολα και τις συναρτήσεις ισχύουν και για τις παρακάτω παραγράφους του κεφαλαίου

χρησιμοποιούνται αλγόριθμοι και τεχνικές από τους χώρους της μηχανικής μάθησης, της υπολογιστικής θεωρίας αλλά και της ανάκτησης πληροφορίας.

Τα συστήματα συστάσεων τα κατηγοριοποιούμε σε τέσσερις τύπους αναλόγως του είδους της πληροφορίας που χρησιμοποιούν ως είσοδο. Αυτοί οι τύποι είναι:

Συστήματα με συστάσεις βάσει περιεχομένου (Content-based recommendation systems): Είναι τα συστήματα που για να παράξουν συστάσεις χρησιμοποιούν πληροφορία από το προφίλ του κάθε χρήστη ξεχωριστά και ότι επιλογές έχει κάνει στο ιστορικό του συνδυασμένες με τα χαρακτηριστικά των αντικειμένων.

Συστήματα με δημογραφικές συστάσεις (Demographic-based recommendation systems): Είναι τα συστήματα που για να παράξουν συστάσεις για έναν χρήστη χρησιμοποιούν πληροφορία από χρήστες με παρόμοια δημογραφικά χαρακτηριστικά πχ ηλικία, επάγγελμα.

Συστήματα με συνεργατικές συστάσεις (Collaborative recommendation systems): Είναι τα συστήματα που για να παράξουν συστάσεις για έναν χρήστη χρησιμοποιούν πληροφορία από χρήστες που του μοιάζουν στον τρόπο συμπεριφοράς (έχουν βαθμολογήσει παρόμοια αντικείμενα και με παρόμοιο τρόπο).

Συστήματα με υβριδικές συστάσεις (Hybrid recommendation systems): Είναι αυτά τα συστήματα που συνδυάζουν τεχνικές από τα τρία είδη που αναφέρθηκαν παραπάνω.

2.2. Συστάσεις βάσει περιεχομένου

Στις συστάσεις βάσει περιεχομένου προσπαθούμε να υπολογίσουμε την τιμή της συνάρτησής F για τα άγνωστα αντικείμενα βάσει των τιμών της πάνω στα γνωστά αντικείμενα με τα οποία υπάρχει ομοιότητα στην δομής τους. Οι συστάσεις βάσει περιεχομένου στηρίζονται στην ιδέα ότι οι χρήστες εκτιμούν περίπου το ίδιο όλα τα αντικείμενα που έχουν όμοια χαρακτηριστικά.

Για να υπολογίσουμε την ομοιότητα των αντικειμένων πρέπει να ορίσουμε κάποια μετρική

συνάρτησή. Για να είναι δυνατός ο ορισμός μιας τέτοιας μετρικής ομοιότητας είναι απαραίτητο να έχουν οριστεί χαρακτηριστικά πάνω στα αντικείμενα που θα αποτελούν τον χώρο στον οποίο γίνονται οι συγκρίσεις. Για παράδειγμα στα κείμενα μπορούμε να ορίσουμε ως χαρακτηριστικά τις λέξεις τους και να χρησιμοποιήσουμε μετρικές που υπολογίζουν την ομοιότητα δύο κειμένων βάσει των συχνοτήτων την κοινών λέξεων. Στις ταινίες μπορούμε να ορίσουμε ως χαρακτηριστικά τους συντελεστές (ηθοποιούς, σεναριογράφους κτλ) και να χρησιμοποιηθούν μετρικές που βασίζονται στο ποσοστό των κοινών συντελεστών. Γενικά ένα σύστημα συστάσεων βάσει περιεχομένου πρέπει σε ότι είδος αντικειμένων συστήνει να κάνει εξαγωγή χαρακτηριστικών και με βάση αυτά να υπολογίσει ομοιότητες. Τα συστήματα συστάσεων συνδυάζουν την ομοιότητα βάσει περιεχομένου με τις βαθμολογίες των γνωστών αντικειμένων ώστε να μπορέσουν να κάνουν συστάσεις.

Φορμαλιστικά κάθε αντικείμενο του συστήματος ορίζεται από ένα διάνυσμα έστω $I\{c_1,c_2,.....,c_k\}$ όπου $c_1,c_2,.....,c_k$ είναι τα χαρακτηριστικά του και k είναι το πλήθος τους. Επίσης μπορούμε να χρησιμοποιήσουμε και ένα επιπλέον διάνυσμα $W\{w_1,w_2,.....,w_k\}$ για να προσδιορίσουμε και το βάρος της συσχέτισης ανάμεσα στα αντικείμενα και τα χαρακτηριστικά τους. Για παράδειγμα σε ένα σύστημα συστάσεων για κείμενα το διάνυσμα I μπορεί να περιέχει τις διαφορετικές λέξεις που εμφανίζονται μέσα στο κείμενο και το διάνυσμα W τις συχνότητες τους. Επίσης ορίζεται ένα διάνυσμα $P\{p_1,p_2,.....,p_N\}$ με τις προτιμήσεις του χρήστη για τα αντικείμενα και που περιέχει την τιμή *Null για τα αντικείμενα που δεν γνωρίζουμε την προτίμηση του. Τείος ορίζεται μια μετρική συνάρτηση* $S(i_1,i_2,w_1,w_2),i_1,i_2\in I,w_1,w_2\in I$ *που υπολογίζει την ομοιότητα ανάμεσα στα αντικείμενα* i_1 *και* i_2 *Αρχικά η συνάρτηση* S *χρησιμοποιείται για να υπολογιστούν όλες οι ομοιότητες των αντικειμένων. Η τιμή της* F *για ένα άγνωστο αντικείμενο* c *υπολογίζεται συνήθως βάσει μιας συνάρτησης που λαμβάνει ως είσοδο τις γνωστές προτιμήσεις πάνω στα όμοια με το* c *αντικείμενα.* Για παράδειγμα μπορεί να υπολογιστεί χρησιμοποιώντας τον σταθμισμένο μέσο όρο της πιθανότητας να παρουσιάζει ίδια τιμή προτίμησης με κάποια από τα γνωστό αντικείμενο.

Τα προβλήματα της παραγωγής συστάσεων βάσει περιεχομένου

Η παραγωγή συστάσεων βάσει περιεχομένου αντιμετωπίζει τρία προβλήματα: τους περιορισμούς της ανάλυσης περιεχομένου, την υπερεξειδίκευση και την αδυναμία συστάσεων σε χρήστες με μικρό ιστορικό,

Περιορισμοί της ανάλυσης περιεχομένου: Τα συστήματα που βασίζονται σε τεχνικές ανάλυσης περιεχομένου πρέπει να είναι σε θέση να εξάγουν τα χαρακτηριστικά των αντικειμένων που συστήνουν. Όταν τα δεδομένα των χαρακτηριστικών μπορούν να ανακτηθούν αυτόματα η διαδικασία είναι απλή. Για παράδειγμα όταν τα χαρακτηριστικά μπορεί να υπάρχουν σε μια βάση δεδομένων ή να εξαχθούν με απλές στατιστικές μεθόδους. Όταν η διαδικασία εξαγωγής χαρακτηριστικών πρέπει να γίνει από τον άνθρωπο το πρόβλημα είναι πολύ δύσκολα διαχειρίσιμο και για μεγάλο όγκο αντικειμένων γίνεται απαγορευτικό.

Λόγω αυτού του περιορισμού οι συστάσεις βάσει περιεχομένου μπορούν να παραχθούν από αντικείμενα όπως τα κείμενα και οι ιστοσελίδες όπου η αυτόματη εξαγωγή χαρακτηριστικών είναι εφικτή. Είναι όμως δύσκολο να εφαρμοστούν σε συστήματα που συστήνουν πολυμεσικό υλικό (ήχος, εικόνες, βίντεο) όταν αυτά δεν συνοδεύονται από βάσεις δεδομένων με χαρακτηριστικά των αρχείων.

Ένας επιπλέον περιορισμός της ανάλυσης περιεχομένου είναι η δυσκολία ποιοτικής αξιολόγησης των χαρακτηριστικών. Δύο αντικείμενα που έχουν τα ίδια χαρακτηριστικά θα βαθμολογηθούν το ίδιο από το σύστημα ανεξάρτητα από καθοριστικές ποιοτικές τους διαφορές. Αν για παράδειγμα τα αντικείμενα μας είναι δύο ταινίες με τους ίδιους ηθοποιούς το σύστημα θα τις αντιμετωπίσει το ίδιο ανεξαρτήτως του πόσο καλά έπαιξαν τους ρόλους τους.

Υπερεξειδίκευση: Η ανάλυση περιεχομένου βασίζει τις συστάσεις της στα αντικείμενα για τα οποία υπάρχει γνωστή κρίση από τον χρήστη. Λόγω αυτού του παράγοντα είναι αδύνατο να συστήσουν αντικείμενα των οποίων τα χαρακτηριστικά δεν συσχετίζονται με τα αντικείμενα από το ιστορικό του χρήστη. Επομένως είναι και αδύνατο να συστήσουν αντικείμενα που εισήχθησαν στο σύστημα με νέα χαρακτηριστικά, όπως ταινίες με νέους ηθοποιούς.

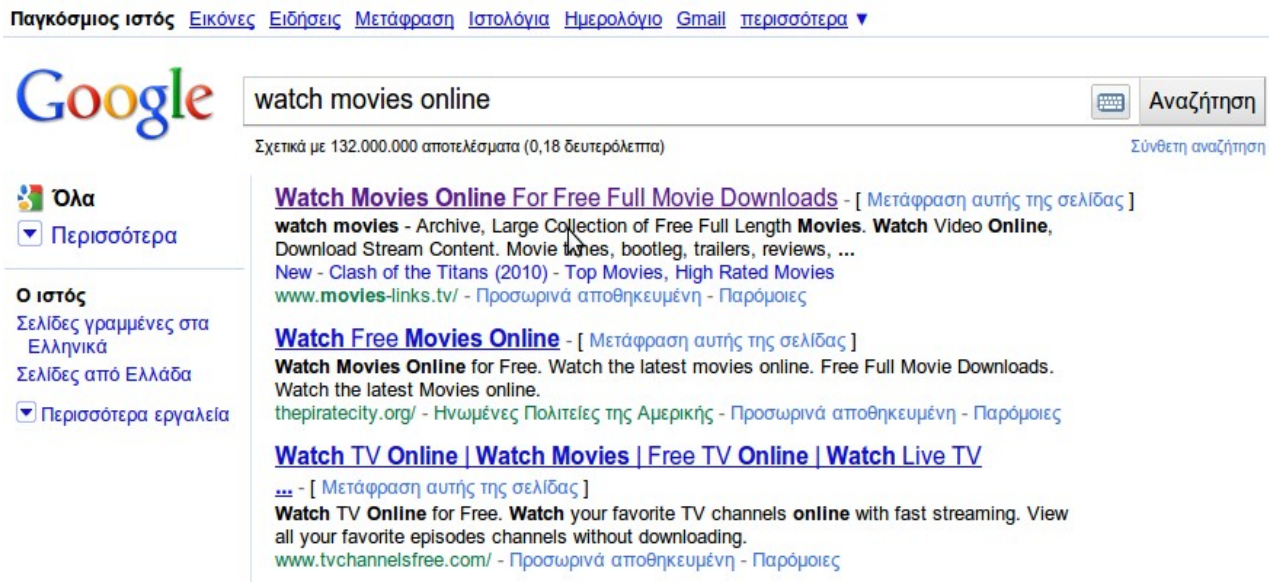
Αδυναμία συστάσεων σε χρήστες με μικρό ιστορικό: Οι χρήστες πού είναι είτε νέοι στο σύστημα είτε δεν το χρησιμοποιούν συχνά. δεν μπορούν να λάβουν καθόλου συστάσεις ή

όσες λαμβάνουν είναι ελάχιστα επιτυχημένες. Χωρίς αρκετή πληροφορία από τις κινήσεις του χρήστη το σύστημα δεν μπορεί να παράξει αξιόπιστες συστάσεις.

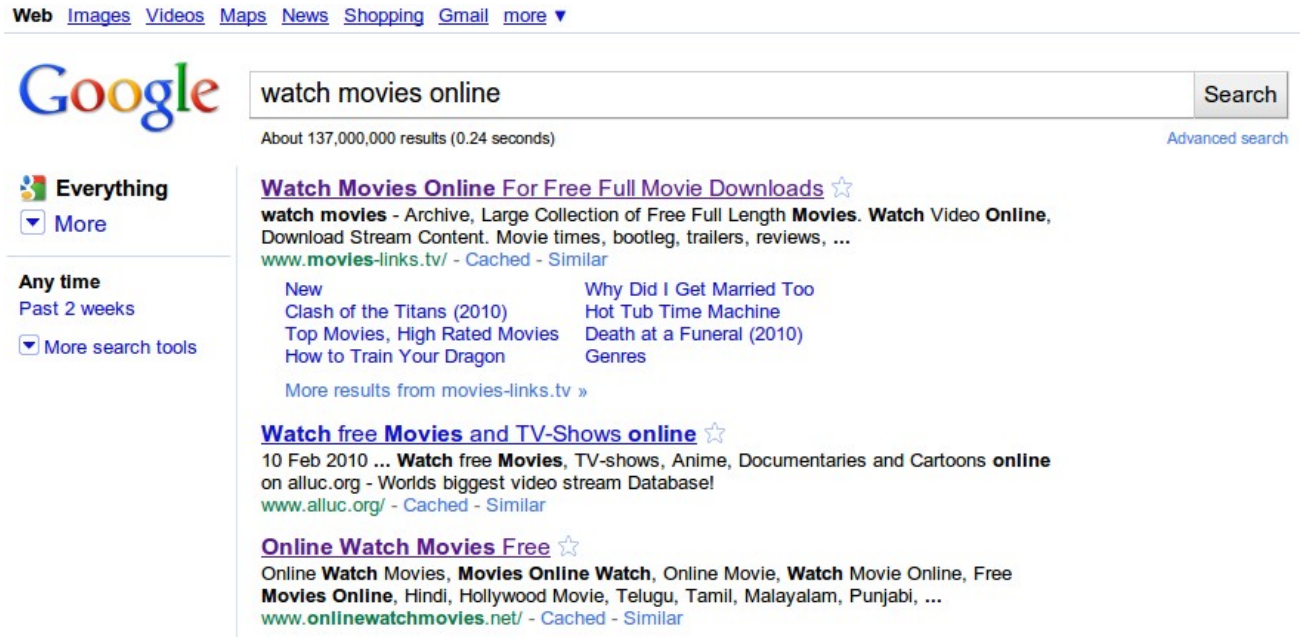
2.3. Δημογραφικές συστάσεις

Στο δημογραφικό φιλτράρισμα προσπαθούμε να υπολογίσουμε την τιμή της συνάρτησης F για τα αντικείμενα που είναι άγνωστα για κάποιον χρήστη χρησιμοποιώντας τις προτιμήσεις χρηστών με παρόμοια δημογραφικά χαρακτηριστικά. Η ιδέα πάνω στην οποία βασίζεται η παραγωγή δημογραφικών συστάσεων είναι ότι οι χρήστες με κοινά χαρακτηριστικά (επάγγελμα, τόπο διαμονής, ηλικία κτλ) είναι πιθανό να μοιράζονται και κοινά ενδιαφέροντα. Αυτή η τεχνική μοιάζει με το φιλτράρισμα περιεχομένου αλλά αντί να χρησιμοποιούμε τα προφίλ των χρηστών απομονωμένα χρησιμοποιούμε τύπους/στερεότυπα χρηστών και κατασκευάζουμε τα προφίλ για αυτές από τα δεδομένα όλων όσων αντιπροσωπεύουν. Τα δημογραφικά χαρακτηριστικά τα οποία ορίζουν τα στερεότυπα εξαρτώνται συνήθως από τα προσωπικά χαρακτηριστικά που θεωρούμε χρήσιμα και μπορούμε να συλλέξουμε από τους χρήστες. Ο ορισμός δε των στερεοτύπων γίνεται με κανόνες που κατασκευάζονται είτε από τους διαχειριστές των συστημάτων συστάσεων είτε δυναμικά χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων πάνω στις βαθμολογίες/προτιμήσεις των χρηστών.

Παράδειγμα στατικού ορισμού κλάσεων είναι η μηχανή αναζήτησης google που ορίζει κλάσεις που βασίζονται στον τόπο διαμονής του χρήστη. Ένας χρήστης από την Ελλάδα που χρησιμοποιεί το google θα δει την ιστοσελίδα στα ελληνικά και τα αποτελέσματα της αναζήτησης του φιλτράρονται ώστε να εμφανιστούν πρώτες οι σελίδες που είναι πιο δημοφιλείς στην Ελλάδα. Για επιβεβαίωση στις εικόνες 1 και 2 βλέπουμε τις τρεις πρώτες ιστοσελίδες από τα αποτελέσματα της αναζήτησης στο google για τις λέξεις “watch movies online”. Στην εικόνα 1 φαίνονται τα αποτελέσματα όταν η αναζήτηση πραγματοποιηθεί από χρήστες της Ελλάδος και στην εικόνα 2 από χρήστες της Αμερικής. Στις θέσεις δύο και τρία οι ιστοσελίδες διαφοροποιούνται γιατί το google κρίνει ότι από χώρα σε χώρα αλλάζει η δημοτικότητα των ιστοσελίδων.



Εικόνα 1. Αποτελέσματα αναζήτησης στο google για χρήστες από την Ελλάδα



Εικόνα 2. Αποτελέσματα αναζήτησης στο google για χρήστες από την Αμερική

Όταν θέλουμε να εξάγουμε τους κανόνες των στερεωτύπων από τα δεδομένα, χρησιμοποιούμε ως σύνολο εκπαίδευσης τα δημογραφικά χαρακτηριστικά των χρηστών, σε συνδυασμό με το ιστορικό των προτιμήσεων τους. Στον πίνακα 1 υπάρχει ένα απλό παράδειγμα για το πώς μπορεί να είναι ένα το τέτοιο σύνολο εκπαίδευσης. Από τα δεδομένα

μπορούν να εξαχθούν συστάδες χρηστών με όμοια ενδιαφέροντα και αυτές να οριστούν ως στερεότυπα. Εφαρμόζοντας αλγόριθμους επιβλεπόμενης μάθησης στα χαρακτηριστικά των χρηστών που ανήκουν στις συστάδες μπορούμε να ορίσουμε και τους κανόνες που πρέπει να ικανοποιούν τα δημογραφικά χαρακτηριστικά ενός νέου χρήστη για να ενταχθεί σε κάποια από αυτές. Στην δημοσίευση [6] παρουσιάζεται μια τεχνική στην οποία ορίζονται στατικά τα είδη των κλάσεων/στερεοτύπων (πχ σε όσους αρέσει ο ηθοποιός Al Pacino) και στην συνέχεια με την χρήση δέντρων απόφασης υπολογίζονται μόνο οι κανόνες.

Χρήστης	Ηλικία	Επάγγελμα	Βαθμός προτίμησης για τον ηθοποιό Al Pacino
Αλέξανδρος	27	Προγραμματιστής	10
Αντώνης	30	Προγραμματιστής	9,5
Κατερίνα	30	Φιλόλογος	4
Βασίλης	32	Φιλόλογος	3

Πίνακας 1. Βαθμολογίες χρηστών σε συνδυασμό με δημογραφικά χαρακτηριστικά

Λόγω της φύσης της ομαδοποίησης που κάνει, ένα σύστημα δημογραφικών συστάσεων τοποθετεί μεγάλο αριθμό χρηστών στο ίδιο γκρουπ και τα αποτελέσματα του είναι αρκετά επιρρεπή στα σφάλματα. Το γεγονός όμως της εύκολης υλοποίησης τέτοιον τεχνικών και του ότι μπορούν να παρέχουν συστάσεις για οποιονδήποτε νέο χρήστη για τον οποίο δεν υπάρχει ιστορικό αλλά μόνο πληροφορία για τα προσωπικά του στοιχεία τις καθιστούν ιδιαιτέρως χρήσιμες.

Τα προβλήματα του δημογραφικού φιλτραρίσματος

Το κυρίως πρόβλημα που αντιμετωπίζουν τα συστήματα δημογραφικών συστάσεων είναι η συλλογή των δημογραφικών χαρακτηριστικών. Πολλοί χρήστες δεν έχουν την διάθεση να παρέχουν τα προσωπικά τους στοιχεία, είτε γιατί δεν έχουν επιθυμούν να συμπληρώσουν μεγάλες φόρμες εισαγωγής είτε δεν θέλουν να αποκαλύψουν τα προσωπικά τους δεδομένα. Επίσης υπάρχουν και χρήστες που δηλώνουν ψευδή στοιχεία που παράγουν θόρυβο. Στην εργασία που παρουσιάζεται στην αναφορά [9] αντί να ζητηθεί από τους χρήστες να παρέχουν οι ίδιοι τα προσωπικά τους στοιχεία, οι χρήστες δηλώνουν απλώς την προσωπική τους

σελίδα. Από το κείμενο που υπάρχει στις σελίδες εξάγονται στοιχεία για τα δημογραφικά χαρακτηριστικά βάσει της γλώσσας και των λέξεων που εμφανίζονται.

2.4. Συνεργατικές συστάσεις

Τα συστήματα που παράγουν συνεργατικές συστάσεις βασίζεται στην ιδέα ότι χρήστες που συμφωνούσαν στο παρελθόν συνηθίζουν να συμφωνούν και στο μέλλον, βάσει αυτού προσπαθούμε να υπολογίσουμε την τιμή της συνάρτησής F για τα άγνωστα αντικείμενα ενός χρήστη, χρησιμοποιώντας τις προτιμήσεις άλλων χρηστών. Το συνεργατικό φιλτράρισμα αποτελεί την δημοφιλέστερη επιλογή των δημιουργών εφαρμογών που παράγουν συστάσεις.

Εμπορικές εφαρμογές που υλοποιούν τέτοιο φιλτράρισμα είναι

- Amazon ηλεκτρονικό πολυκατάστημα
- Amie Street ηλεκτρονικό πολυκατάστημα μουσικής
- Barilliance εφαρμογή υποστήριξης ηλεκτρονικοί εμπορίου
- Barnes and Noble ηλεκτρονικό κατάστημα βιβλίων
- Baynote μηχανή αναζήτησης
- ChoiceStream εταιρεία παροχής διαδικτυακό διαφημίσεων
- Collarity εταιρεία παροχής διαδικτυακών διαφημίσεων και μηχανής αναζήτησης
- Digg.com ιστοσελίδα κοινωνικής δικτύωσης
- Directed Edge εταιρεία που παρέχει συστάσεις πάνω σε ιστοσελίδες κοινωνικής δικτύωσης, ηλεκτρονικών καταστημάτων και ιστοσελίδων πληροφόρησης.
- eBay διαδικτυακή πύλη ηλεκτρονικού εμπορίου
- Google News υπηρεσία της Google που παρέχει εξατομικευμένα νέα
- Gravity R&D εταιρεία παροχής ψηφιακού περιεχομένου
- half.ebay.com
- Heeii Παρέχει συστάσεις για ιστοσελίδες
- Hollywood Video Διαδικτυακή πύλη παρακολούθησης
- Hulu τηλεόραση μέσω διαδικτύου
- Internet Movie Database παρέχει πληροφορίες για ταινίες
- iTunes διαδικτυακή πύλη με εμπορείο πολυμετοχικού υλικού
- Last.fm διαδικτυακό ραδιόφωνο
- LibraryThing πληροφορίες βιβλίων
- Loomia παροχή συστάσεων ιστοσελίδων
- Musicmatch Εφαρμογή αναπαραγωγής μουσικής
- MyStrands ιστοσελίδα κοινωνικής δικτύωσης
- Netflix εκπομπή βίντεο μέσω διαδικτύου

- StumbleUpon κοινωνική δικτύωση
- Threadless διαδικτυακή ιστοσελίδα για παραγγελίες t-shirt

Τους αλγόριθμους παραγωγής συνεργατικών συστάσεων μπορούμε να τους κατατάξουμε σε δύο κατηγορίες: στους αλγορίθμους βασισμένους σε μνήμη (memory-based) και στους αλγόριθμους βασισμένους σε μοντέλο (model-based).

2.4.1. Αλγόριθμοι βασισμένοι σε μνήμη

Αυτή η κατηγορία αλγορίθμων παραγωγής συνεργατικών συστάσεων διατηρεί στην μνήμη όλες τις γνωστές βαθμολογίες/προτιμήσεις και τις χρησιμοποιεί για να βρει ομοιότητες ανάμεσα σε χρήστες ή αντικείμενα. Δύο χρήστες μοιάζουν όταν ενδιαφέρονται για παρόμοια πράγματα ενώ δύο αντικείμενα μοιάζουν όταν ένα σύνολο χρηστών τα αντιμετωπίζουν με παρόμοια αρέσκεια.

Στην περίπτωση που θέλουμε να βασιστούμε στις ομοιότητες χρηστών (user-based collaborative filtering), αναπαριστούμε τους χρήστες ως διανύσματα στο χώρο των αντικειμένων και υπολογίζουμε την ομοιότητα τους με βάση την απόσταση αυτών των διανυσμάτων. Όταν θέλουμε να εκτιμήσουμε την προτίμηση ενός χρήστη για ένα άγνωστο αντικείμενο συγκεντρώνουμε τις προτιμήσεις των N κοντινότερων χρηστών που έχουν εκφράσει την προτίμηση τους για το αντικείμενο. Εκτιμούμε την προτίμηση του χρήστη εφαρμόζοντας μια συναθροιστική συνάρτηση, συνήθως τον σταθμισμένο μέσο όρο, πάνω στις τιμές που συγκεντρώσαμε.

Φορμαλιστικά, έστω $s(u_i,u_j)$ η συνάρτηση που υπολογίζει την ομοιότητα ανάμεσα στους χρήστες u_i και u_j , $S \subseteq A$ το σύνολό των N όμοιων χρηστών, $r_{u,I}$ η προτίμηση του χρήστη u για το αντικείμενο I . Η γενική μορφή της συνάρτησης για τον υπολογισμό της εκτιμώμενης προτίμησης θα είναι $E(u,I)=aggr(S)$ και αν η συναθροιστική συνάρτηση είναι ο σταθμισμένος μέσος όρος.

$$E(u,I)=\frac{\sum_{u_o \in S} s(u,u_o)*r_{u_o,I}}{\sum_{u_o \in S} s(u,u_o)}$$

Αντίστοιχα όταν βασιζόμαστε στις ομοιότητες αντικειμένων (item-based collaborative filtering) αναπαριστούμε τα αντικείμενα ως διανύσματα στον χώρο των χρηστών και υπολογίζουμε τις αποστάσεις των διανυσμάτων. Όταν θέλουμε να εκτιμήσουμε την προτίμηση ενός χρήστη για ένα άγνωστο αντικείμενο βρίσκουμε τα N κοντινότερα αντικείμενα τα οποία έχει βαθμολογήσει ο χρήστης και υπολογίζουμε μέσω μιας συνάρτησης συνήθως με σταθμισμένο μέσο όρο, την τιμή βάσει των άλλων προτιμήσεων του. Φορμαλιστικά έστω $s(I_i, I_j)$ η ομοιότητα ανάμεσα στα αντικείμενα I_i και I_j και $SI \subseteq B$ το σύνολο των N όμοιων αντικειμένων. Η εκτίμηση της προτίμησης είναι $E(u,I) = aggr(SI)$ και αν υποθέσουμε ότι χρησιμοποιούμε τον σταθμισμένο μέσο όρο

$$E(u,I) = \frac{\sum_{I_o \in SI} s(I,I_o) * r_{u,I_o}}{\sum_{I_o \in SI} s(I,I_o)}$$

Πολύ σημαντικός παράγοντας για την επιτυχία των αλγορίθμων είναι σωστή επιλογή της μετρικής συνάρτησης ομοιότητας. Στην βιβλιογραφία έχουν προταθεί πάρα πολλές τέτοιες μετρικές και ενδεικτικά ακολουθεί η περιγράφει των δημοφιλέστερων που χρησιμοποιήθηκαν και στην παρούσα εργασία.

Ομοιότητα Συνημιτόνου (Cosine Similarity): Είναι το συνημίτονο της γωνίας που σχηματίζουν τα δύο διανύσματα που συγκρίνονται. Επιστρέφει τιμές ανάμεσα στο -1 και το 1 με το 1 να δηλώνει την απόλυτη ταύτιση, το -1 την απόλυτη απόκλιση των διανυσμάτων. Όταν τα διανύσματα είναι κάθετα μεταξύ τους η επιστρεφόμενη τιμή είναι 0 και συνήθως αυτό λαμβάνεται ως ανεξαρτησία. Ο τύπος της μετρικής είναι ο ακόλουθος

$$similarity(A,B) = \cos(\theta) = \frac{\vec{A} * \vec{B}}{(\vec{A})(\vec{B})} = \frac{\sum_{i=1}^{N} A_i * B_i}{\sqrt{\sum_{i=1}^{N} A_i^2} * \sqrt{\sum_{i=1}^{N} B_i^2}}$$

Το σύμβολο N αντιστοιχεί στον αριθμό των διαστάσεων του χώρου των διανυσμάτων. Όταν η προτιμήσεις των χρηστών δηλώνονται μόνο με θετικές τιμές τότε το σύνολο των

επιστρεφόμενων τιμών περιορίζεται στο [0,1].

Προσαρμοσμένη Ομοιότητα Συνημιτόνου (Adjusted Cosine Similarity): Είναι μια παραλλαγή της απόστασης συνημιτόνου η οποία συμπεριλαμβάνει τον μέσο όρο των τιμών που υπάρχουν σε κάθε διάσταση. Σκοπός της παραλλαγής είναι να αντιμετωπίσει την διαφορετική αντίληψη που έχουν οι χρήστες για την κλίμακα των βαθμολογιών.

$$similarity(A,B) = \frac{\sum_{i=1}^{N} (A_i - \mu o(i)) * (B_i - \mu o(\iota))}{\sqrt{\sum_{i=1}^{N} (A_i - \mu o(i))^2} * \sqrt{\sum_{i=1}^{N} (B_i - \mu o(i))^2}}$$

Όπου $\mu o(i)$ είναι ο μέσος όρος των τιμών της *i*-οστής διάστασης των διανυσμάτων. Αν συγκρίνουμε χρήστες θα είναι ο μέσος όρων των τιμών των προτιμήσεων που έχουν δείξει όλοι οι χρήστες στο αντικείμενο *i* ενώ αν συγκρίνουμε αντικείμενα θα είναι ο μέσος των τιμών από τις προτιμήσεις που έχει εκφράσει ο *i* χρήστης.

Συντελεστής συσχέτισης του Pearson (Pearson's Correlation Coefficient): Υπολογίζει την γραμμική συσχέτιση δύο διανυσμάτων και επιστρέφει τιμές από -1 έως 1. Όσο η τιμή πλησιάζει το 1 τα διανύσματα μοιάζουν όλο και περισσότερο ενώ όσο η τιμή πλησιάζει στο -1 τα διανύσματα είναι όλο και πιο αντίθετα. Η τιμή 0 δηλώνει ότι οι τιμές των συνιστωσών των διανυσμάτων είναι γραμμικώς ανεξάρτητες. Ο τύπος της μετρικής είναι ο ακόλουθος

$$similarity(A,B) = \frac{\sum_{i=1}^{N} (A_i - \mu o(A)) * (B_i - \mu o(B))}{\sqrt{\sum_{i=1}^{N} (A_i - \mu o(A))^2} * \sqrt{\sum_{i=1}^{N} (B_i - \mu o(B))^2}}$$

Οι συμβολισμοί μο(A) και μο(B) είναι οι μέσοι όροι των τιμών των συνιστωσών των διανυσμάτων A και B αντίστοιχα. Αν συγκρίνουμε χρήστες θα είναι οι μέσοι όροι των τιμών των προτιμήσεων του κάθε χρήστη ενώ αν συγκρίνουμε αντικείμενα θα είναι οι μέσοι όροι των τιμών των προτιμήσεων που έχουν δείξει οι χρήστες για τα αντικείμενα.

Ομοιότητα Jaccard: Αυτή η μετρική δεν υπολογίζει αποστάσεις ανάμεσα σε διανύσματα άλλα το ποσοστό αλληλεπικάλυψης συνόλων. Εφαρμόζεται όταν ελέγχουμε μόνο το ποία είναι τα αντικείμενα για τα οποία έχουν εκφράσει ενδιαφέρον οι χρήστες ενώ δεν μας ενδιαφέρει το αν συμφωνούν στο πόσο τους άρεσαν. Ο τύπος της μετρικής είναι

$$similarity(A,B) = \frac{(A \cap B)}{(A \cup B)}$$

Αντίστοιχα για τον υπολογισμό της απόστασης Jaccard δηλαδή το πόσο διαφέρουν δύο σύνολα ισχύει ο τύπος

$$dissimilarity(A,B) = \frac{(A \cup B) - (A \cap B)}{(A \cup B)}$$

Υπάρχουν εργασίες που συγκρίνουν την απόδοση των μετρικών σε διάφορα προβλήματα συσταδοποίησης (clustering), συνεργατικού φιλτραρίσματος και γενικώς όπου έχουν εφαρμογή. Στις δημοσιεύσεις των αναφορών [10] και [11] τα πειράματα δείχνουν ότι ο συντελεστής συσχέτισης του Pearson είναι η καλύτερη επιλογή για τον υπολογισμό ομοιότητας χρηστών. Στην δημοσίευση της αναφοράς [12] υπάρχουν αποτέλεσμα πειραμάτων που δείχνουν ότι για τον υπολογισμό ομοιότητας αντικειμένων η καλύτερη επιλογή είναι η προσαρμοσμένη ομοιότητα συνημιτόνου. Εκτός από την ομοιότητα Jaccard οι μετρικές που αναφέρθηκαν παραπάνω εφαρμόζονται μόνο για τις διαστάσεις που υπάρχει τιμή και για τα δύο διανύσματα. Δηλαδή όταν συγκρίνουμε δύο χρήστες τα διανύσματα περιέχουν τιμές για τα αντικείμενα που έχουν βαθμολογήσει και οι δύο ενώ όταν συγκρίνουμε δύο αντικείμενα τα διανύσματα περιέχουν τιμές μόνο από τους χρήστες οι οποίοι έχουν βαθμολογήσει και τα δύο. Αυτό έχει ως αποτέλεσμα να αποκρύπτονται σοβαρές διαφορές των χρηστών και να κρίνονται όμοιοι χρήστες που απλά έχουν ένα, δυο κοινά αντικείμενα ενώ έχουν βαθμολογήσει πολύ περισσότερα. Στην εργασία της αναφοράς [11] υπάρχουν πειραματικά αποτελέσματα που δείχνουν ότι ο συνδυασμός της ομοιότητας Jaccard με τον συντελεστή συσχέτισης Pearson αλλά και την ομοιότητα συνημιτόνου βελτιώνει την απόδοση της πρόβλεψης βαθμολογιών.

$$wcosine(A,B) = JaccardSimilarity(A,B) * CosineSimilarity(A,B)$$

$$wpearson(A,B) = JaccardSimilarity(A,B) * PearsonCorrelationCoefficient(A,B)$$

2.4.2. Αλγόριθμοι βασισμένοι σε μοντέλο

Τα περισσότερα συστήματα συνεργατικού φιλτραρίσματος βασίζονται σε αλγορίθμους μνήμης καθώς υλοποιούνται εύκολα. Παρουσιάζουν όμως πρόβλημα κλιμάκωσης και είναι ευαίσθητα στην εισαγωγή ψεύτικων προφίλ (profile injection). Σε αυτού του τύπου τις επιθέσεις γίνεται προσπάθεια εισαγωγής ψεύτικων προφίλ χρηστών ώστε μέσω των προτιμήσεων να ευνοηθούν αντικείμενα που στην πραγματικότητα δεν θα ήταν το ίδιο δημοφιλή. Οι αλγόριθμοι που βασίζονται σε μοντέλο χρησιμοποιούν τα δεδομένα των προτιμήσεων/βαθμολογιών ως σύνολο εκπαίδευσης αλγόριθμων μηχανικής μάθησης για να παράξουν μοντέλα πρόβλεψης βαθμολογιών. Η επεξεργασία γίνεται σε μη πραγματικό χρόνο και έτσι μειώνεται το πρόβλημα κλιμάκωσης καθώς μπορεί να γίνει επεξεργασία μεγάλου αριθμό από προφίλ καθώς δεν απαιτείται απάντηση σε πραγματικό χρόνο. Η δυνατότητα της επεξεργασίας στο παρασκήνιο επιτρέπει την εκτέλεση πολύπλοκων αλγορίθμων αντιμετώπισης του θορύβου που προκαλούν οι επιθέσεις ψεύτικων προφίλ.

Τα μοντέλα πολλές φόρες τείνουν να έχουν λιγότερο ακριβή αποτελέσματα από τους αλγορίθμους μνήμης (O'Conner & Herlocker 1999). Για αυτό οι σχεδιαστές συστημάτων συστάσεων θα πρέπει να αναλογιστούν την σχέση κλιμάκωσης προς απόδοση πριν αποφασίσουν ποιες τεχνικές θα χρησιμοποιήσουν.

Στην κλασική μορφή αλγορίθμων που βασίζονται σε μοντέλα, χρησιμοποιείται το σύνολο βαθμολογιών και εφαρμόζεται ένας αλγόριθμος συσταδοποίησης (k-means[13], hierarchical clustering[14], maximal cliques[15] και δημιουργούνται συστάδες από κοντινούς χρήστες. Για κάθε συστάδα υπολογίζεται ένα προφίλ με τις βαθμολογίες από τον μέσω όρο όλων των χρηστών που ανήκουν σε αυτή.

Φορμαλιστικά, έστω u_k μία συστάδα στην οποία αντιστοιχούν ένα σύνολο από χρήστες U .

χρήστη τότε ισχύει

$$\vec{u}_k = \frac{\sum_{n \in U} \vec{u}_n}{(U)}$$

Για να γίνει εκτίμηση της προτίμησης ενός χρήστη για ένα άγνωστο αντικείμενο επιλέγεται η συστάδα στην οποία ανήκει ο χρήστης και επιστρέφεται η τιμή που έχει το προφίλ της.

Μια άλλη μεγάλη κατηγορία αλγορίθμων συνεργατικού φιλτραρίσματος βασισμένων σε μοντέλο είναι οι πιθανοτικοί (Bayesian CF Algorithms [10][13], Probabilistic Latent Semantic Analysis [14]). Στόχος των αλγορίθμων είναι να υπολογιστεί η πιθανότητα p(r|u,i) ένας χρήστης u να δώσει την βαθμολογία r σε ένα αντικείμενο i. Αφού υπολογιστούν οι πιθανότητες για όλες τις δυνατές τιμές η εκτίμηση γίνεται βάσει τους σταθμισμένου μέσου όρου τους

$$r_{ui} = \frac{\sum_{r \in R} p(r|u,i) * r}{\sum_{r \in R} p(r|u,i)}$$

όπου R είναι το σύνολο των τιμών που επιτρέπονται για την βαθμολογία ενός αντικειμένου.

Στην βιβλιογραφία υπάρχουν πολλά ακόμα είδη αλγορίθμων που βασίζονται σε προβλήματα βελτιστοποίησης περιορισμών, δέντρα αποφάσεων και γενικά άλλες μεθόδους της τεχνητής νοημοσύνης.

2.4.3. Τα προβλήματα των συνεργατικών συστάσεων

Τα προβλήματα που παρουσιάζονται στην παραγωγή συνεργατικών συστάσεων είναι η αδυναμία παροχής συστάσεων σε νέους χρήστες, αδυναμία παροχής συστάσεων που περιέχουν νέα αντικείμενα και η αντιμετώπιση αραιών δεδομένων.

Το πρόβλημα των νέων χρηστών: Όπως και στην παραγωγή συστάσεων βάσει περιεχομένου έτσι και εδώ για να παραχθούν συστάσεις χρειάζονται δεδομένα για τις προτιμήσεις του χρήστη. Για τους νέους χρήστες δεν είναι δυνατό να εντοπιστεί άλλος

χρήστης με τον οποίο να μοιάζουν. Το συνεργατικό φιλτράρισμα βασισμένο σε αντικείμενα καταφέρνει να ξεπεράσει γρηγορότερα το πρόβλημα του νέου χρήστη καθώς μόλις ο χρήστης εκδηλώσει τις πρώτες του προτιμήσεις είναι πιθανό να μπορούν να γίνουν συστάσεις αφού οι συσχετίσεις των αντικειμένων έχει υπολογιστώ βάσει των παλιών χρηστών.

Το πρόβλημα των νέων αντικειμένων: Για να γίνει σύσταση ενός αντικειμένου θα πρέπει πρώτα κάποιοι χρήστες να εκφράσουν την προτίμηση τους για αυτό. Επομένως για τα νέα αντικείμενα που εισέρχονται στο σύστημα στο σύστημα δεν υπάρχουν δεδομένα που να τα καθιστούν άξια σύστασης. Το πρόβλημα ξεπερνιέται γρηγορότερα στην περίπτωση που βασιζόμαστε σε συσχετίσεις χρηστών αφού μόλις κάποιοι χρήστες αξιολογήσουν το νέο αντικείμενο αυτό μπορεί να συσταθεί στους όμοιους του.

Το πρόβλημα της σποραδικότητας δεδομένων (sparsity problem): Στις περισσότερες εφαρμογές που παρέχουν περιεχόμενο τα αντικείμενα που παρέχονται είναι πολύ περισσότερα από όσα βαθμολογούν/αξιολογούν οι χρήστες. Αυτό έχει ως αποτέλεσμα τα συστήματα συστάσεων να πρέπει να εκτιμήσουν την αρέσκεια των χρηστών αντικείμενα από όσα είναι τα δεδομένα τους. Επίσης υπάρχουν χρήστες που οι προτιμήσεις τους διαφέρουν αρκετά από την πλειοψηφία των άλλων χρηστών με αποτέλεσμα να είναι δύσκολο να εντοπιστούν ομοιότητες και να τους προταθούν αντικείμενα.

2.4.4. Υβριδικές συστάσεις

Τα συστήματα που παράγουν υβριδικές συστάσεις χρησιμοποιούν αλγόριθμους που συνδυάζουν τεχνικές που ανήκουν σε δύο ή περισσότερες από τις κατηγορίες συστάσεων που περιγράφηκαν παραπάνω. Στόχος είναι να αντιμετωπιστούν τα προβλήματα που παρουσιάζει το κάθε είδος παραγωγής συστάσεων όταν χρησιμοποιείται μόνη της από τα πλεονεκτήματα των άλλων. Στην αναφορά [16] περιέχει μια έρευνα για διάφορα σημαντικά συστήματα και τεχνικές για συστήματα συστάσεων που είχαν δημοσιευτεί μέχρι το 2005. Στο κείμενο καταγράφεται ότι έχουν χρησιμοποιηθεί οι έξεις τρόποι υλοποίησης υβριδικού φιλτραρίσματος:

1. Παραγωγή εκτιμήσεων βάσει περιεχομένου και συνεργατικού ξεχωριστά και στην

συνέχεια να γίνει συνδυασμός τους.

2. Εμπλουτισμός του συνεργατικού φιλτραρίσματος με χαρακτηριστικά από το φιλτράρισμα βάσει περιεχομένου.

3. Εμπλουτισμός του φιλτραρίσματος βάσει περιεχομένου με χαρακτηριστικά από το συνεργατικό.

4. Κατασκευή ενός γενικού μοντέλου που ενοποιεί το συνεργατικό με το φιλτράρισμα περιεχομένου

Στην συγκεκριμένη έρευνα δεν γίνεται καθόλου αναφορά στο δημογραφικό φιλτράρισμα και απλώς σημειώνεται ότι τα δημογραφικά χαρακτηριστικά μπορούν να χρησιμοποιηθούν για να επεκταθεί το συνεργατικό φιλτράρισμα ώστε να αντιμετωπιστεί το πρόβλημα των νέων χρηστών. Στις παραπάνω υποδείξεις συνδυασμού τεχνικών φιλτραρίσματος μπορούν να προστεθούν

5. Εμπλουτισμός του συνεργατικού φιλτραρίσματος με δημογραφικά χαρακτηριστικά

6. Εξαγωγή προφίλ δημογραφικού φιλτραρίσματος από συνεργατικό φιλτράρισμα και επιλογή χρήσης αυτού για τους νέους χρήστες που δεν έχουν καταγεγραμμένο ιστορικό.

Ένα παράδειγμα παραγωγής υβριδικών συστάσεων υπάρχει και στην εργασία [33] που προσπαθεί να προβλέψει την βαθμολογία ταινιών βασισμένο στην κατασκευή ενός μοντέλου γράφου. Ως κόμβους στον γράφο τοποθετεί τις ταινίες, τα χαρακτηριστικά τους και τους χρήστες του συστήματος. Δημιουργεί ακμές ανάμεσα στους χρήστες και τα αντικείμενα και ανάμεσα στα αντικείμενα και τα χαρακτηριστικά τους. Στις ακμές των χρηστών με τα αντικείμενα τοποθετεί βάρη ανάλογα των βαθμολογιών που έχει δώσει ο χρήστης. Η ιδέα είναι ότι η εκτίμηση της προτίμησης για ένα άγνωστο αντικείμενο μπορεί να γίνει βάσει των βαρών της βέλτιστης διαδρομής που συνδέει το αντικείμενο με τον χρήστη.

2.5 Το πρόβλημα της εννοιολογικής απόκλισης

Τα συστήματα παραγωγής συστάσεων αλλά και γενικότερα κάθε σύστημα που παρέχει εξατομικευμένο υλικό επεξεργάζονται ροές δεδομένων που είτε άμεσα (με βαθμολογίες) είτε έμμεσα (με επιλογές) περιέχουν τις εκφρασμένες προτιμήσεις του κάθε χρήστη. Οι προτιμήσεις όμως δεν είναι αναγκάστηκα στατικές και μπορεί να αλλάζουν μέσα στον χρόνο (interest drift). Συνήθως δεν είναι σωστό να χρησιμοποιηθεί όλη η ροή στις προβλέψεις νέων προτιμήσεων και δεν πρέπει να αντιμετωπιστούν όλα τα δεδομένα με την ίδια αξία. Το πρόβλημα που παρουσιάζεται στις ροές δεδομένων, όπου ο ορισμός μιας ένοιας αλλάζει με άγνωστο τρόπο ονομάζεται εννοιολογική απόκλιση (concept drift).

Φορμαλιστικά, κάθε ροή αποτελείται από στιγμιότυπα $I_1.I_2.I_3.I_4....$ της μορφής $\iota=(x,y)$ όπου x είναι είναι οποιοδήποτε διάνυσμα του ν-διαστατού χώρου (πχ κάποιο αντικείμενο)) και y μία ετικέτα (πχ, μια βαθμολογία). Τα στιγμιότυπα φτάνουν σε (ισομεγέθεις) ομάδες $b_{t_1}.b_{t_2}.b_{t_3}.b_{t_4}$ τις διαδοχικές χρονικές στιγμές $t_1.t_2.t_3.t_4....$. Γνωρίζονταςτην κατανομή των στιγμιοτύπων για τις ομάδες μέχρι την χρονική στιγμή t_i θέλουμε να προβλέψουμε την κατανομή για την χρονική στιγμή t_{i+1} .

Η εννοιολογική απόκλιση εκφράζεται είτε σταδιακά (gradual concept drift) είτε απότομα (instant concept drift). Στην πρώτη περίπτωση, τα ενδιαφέροντα του χρήστη αλλάζουν σταδιακά με την πάροδο του χρόνου. Για παράδειγμα οι προτιμήσεις σε βιβλία που αλλάζουν όσο ο άνθρωπος μεγαλώνει. Στη δεύτερη περίπτωση, τα ενδιαφέροντα του χρήστη αλλάζουν απότομα. Για παράδειγμα ένας πελάτης που θέλει να αγοράσει έναν φορητό υπολογιστή και για κάποιο χρονικό διάστημα κάνει έρευνα αγοράς. Στο διάστημα της έρευνας θέλει να βλέπει προτάσεις υπολογιστών αν όμως τον αγοράσει δεν ενδιαφέρεται πλέον.

Στην εννοιολογική απόκλιση είναι γενικά αποδεκτό ότι οι όσο πιο πρόσφατα είναι τα δεδομένα μιας ροής με χρονικά δεδομένα τόσο πιο έγκυρα. Υπάρχει όμως και το φαινόμενο όπου αντικείμενα που παρουσιάζουν ενδιαφέρον για συγκεκριμένο χρήστη στο παρελθόν στην συνέχεια να το χάνουν και στο μέλλον να επανεμφανίζονται στα ενδιαφέροντα του.

2.6. Κατηγορίες αλγορίθμων

Οι αλγόριθμοι για την αντιμετώπιση της εννοιολογική απόκλισης κατηγοριοποιούνται στις ακόλουθες ομάδες [19][20].

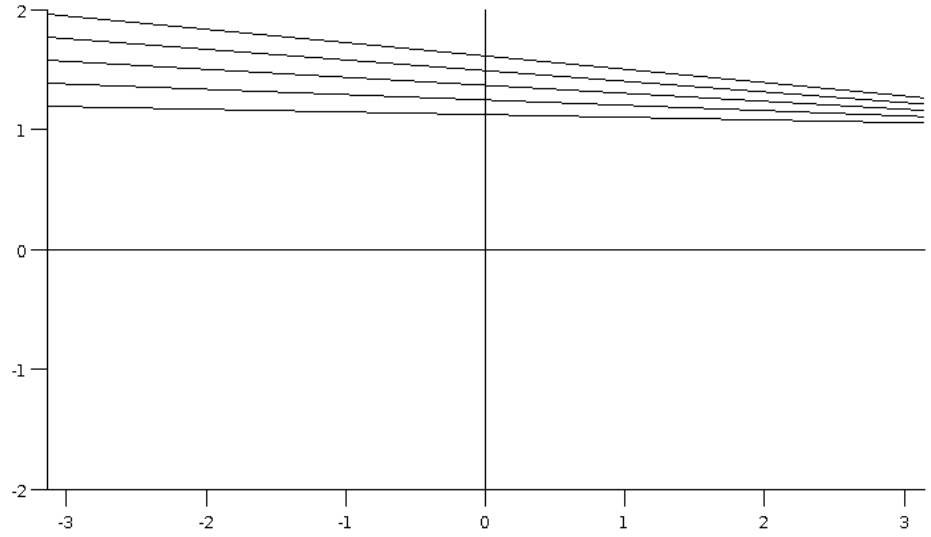
Επιλογή στιγμιότυπων (instance selection): Οι αλγόριθμοι αυτής της κατηγορίας κρατούν μόνο όσα από τα δεδομένα της ροής προτιμήσεων θεωρούν καταλληλότερα για τις προβλέψεις. Οι περισσότερο αλγόριθμοι προσπαθούν να βρουν ένα σταθερό ή προσαρμοζόμενο παράθυρο χρόνου (time-window) μέσα στο οποίο είναι τα χρήσιμα δεδομένα της ροής. Στόχος είναι να απορρίψουν γνώση που είναι πλέον άχρηστη, βασιζόμενοι στην παραδοχή ότι από μια χρονική στιγμή και πίσω τα δεδομένα είναι απλός θόρυβος. Στην κατηγορία αυτή ανήκουν οι εργασίες στις αναφορές [22][23][24][25].

Συχνά η εννοιολογική απόκλιση παρουσιάζεται μόνο σε συγκεκριμένα στιγμιότυπα της ροής δεδομένων και όχι γενικά στο σύνολο της (local concept drift). Οι αλγόριθμοι που χρησιμοποιούν παράθυρο αγνοούν τις συγκεκριμένες περιπτώσεις και θεωρούν πως η εννοιολογική απόκλιση είναι καθολική. Ακόμα και αν γίνει χρήση προσαρμοστικών βέλτιστων παραθύρων χρόνου αυτά εφαρμόζονται σε αλληλουχίες όμοιων στιγμιοτύπων και όχι σε μεμονωμένα στιγμιότυπα. Στην γενική περίπτωση, η εννοιολογική απόκλιση έχει κοινά χαρακτηριστικά στις μεγαλύτερες περιόδους των ροών δεδομένων και όποτε η τοπική εννοιολογική απόκλιση δεν θεωρείται σοβαρό πρόβλημα.

Τοποθέτηση βαρών στα στιγμιότυπα (Instance weighting): Οι αλγόριθμοι αυτής της κατηγορίας δεν ξεχνούν αλλά θεωρούν ότι τα πιο πρόσφατα δεδομένα είναι περισσότερο χρήσιμα από τα παλαιότερα και τους δίνουν επιπλέον βάρος. Για την υλοποίηση τέτοιων αλγορίθμων χρησιμοποιείται κάποια συνάρτηση εξασθένισης (decay function) $f(t)$ της οποίας η τιμή αντιστοιχεί στα βάρη των στιγμιοτύπων που υπάρχουν μέσα στην ροή. Η $f(t)$ συνήθως είναι μονότονη και έχει πεδίο ορισμού το (0,1], Στο κείμενο της αναφοράς [26] προτείνεται μία γραμμική συνάρτηση και αποδεικνύεται πειραματικά η καλή λειτουργίας της.

$$f(t_i) = 1 + k - \frac{2k(t_i - 1)}{n - 1}$$

Ο συντελεστής k είναι το ποσό που ελαττώνεται η συνεισφορά κάθε ομάδας στιγμιοτύπων όσο απομακρυνόμαστε από την χρονική στιγμή 0. Ο συντελεστής n είναι ο αριθμός των ομάδων στιγμιοτύπων. Η εικόνα 3 έχει την μορφή της συνάρτησης για k = {0.1,0.2,0.3,0.4,0.5} και n = 10. Όσο αυξάνεται η τιμή του k μεγαλώνει η κλήση της γραφικής παράστασης.



Εικόνα 3. Γραμμική συνάρτηση εξασθένησης

Αρκετά δημοφιλής συνάρτηση εξασθένησης είναι και η εκθετική. Για τον ορισμό αυτής της συνάρτησης πρέπει πρώτα να οριστεί ο χρόνος ημιζωής. Ο χρόνος ημιζωής T_0 είναι ο χρόνος που απαιτείται ώστε ένα στιγμιότυπο να χάσει την μισή του αξία. Δηλαδή να ισχύει

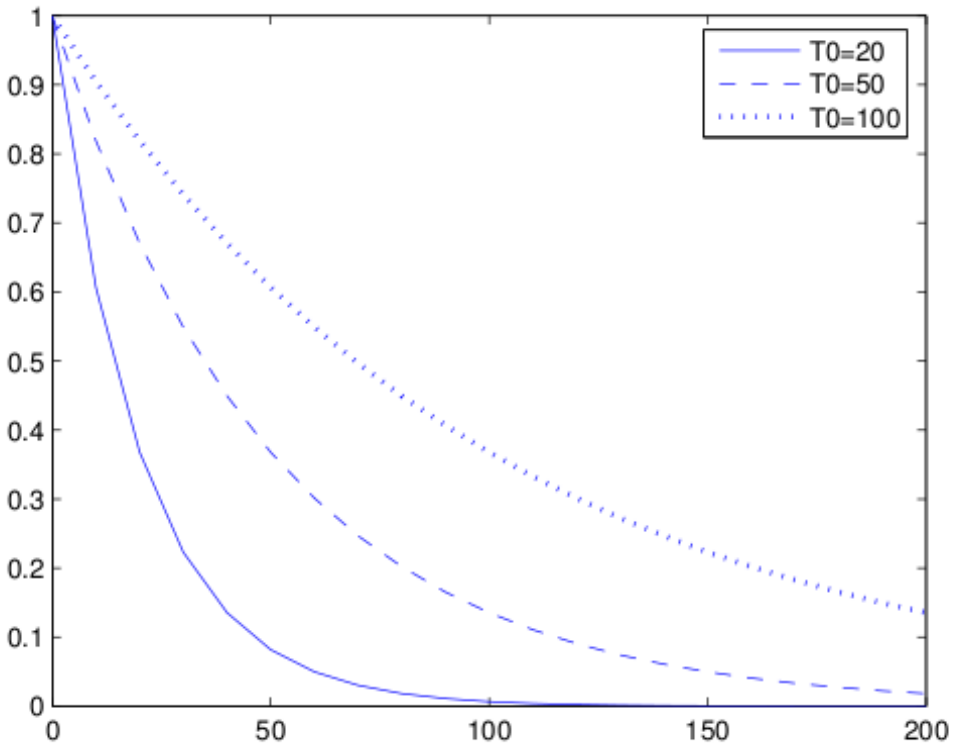
$$f(T_0)=\frac{f(0)}{2}$$

στην συνέχει ορίζεται ο συντελεστής εξασθένησης (decay factor), ως:

$$\lambda=\frac{1}{T_0}$$

και τέλος ο ορισμός της ίδιας της εκθετικής συνάρτησης είναι

$$f(t)=e^{-\lambda t}$$



Εικόνα 3. Εκθετική συνάρτηση για διαφορετικές τιμές του T_0 . αναγεγραμμένη από το [30]

Στην εργασία της αναφοράς [29] χρησιμοποιεί την εκθετική συνάρτηση εξασθένησης ώστε να παράξει βάρη για τις προτιμήσεις των χρηστών. Καταφέρνει να βελτιώσει τα αποτελέσματα προβλέψεων του κλασικού συνεργατικού φιλτραρίσματος με αντικείμενα (παράγραφος 2.4.1) εμπλουτίζοντας τις βαθμολογίες με τα βάρη της εκθετικής συνάρτησης.

$$E(u,i)=\frac{\sum_{i_o \in I} s(i,i_o)*r_{u,i}*f(t)}{\sum_{i_o \in S} s(i,i_o)*f(t)}$$

Εκτός από τους αλγορίθμους που υπολογίζουν τα βάρη των στιγμιοτύπων βάσει συναρτήσεων εξασθένησης υπάρχουν και εργασίες που υπολογίζουν τα βάρη βάσει του πόσο σχετικό είναι το κάθε στιγμιότυπο από τις παλιές ομάδες εκπαίδευσης με αυτά των νέων. Για παράδειγμα το σύστημα Nootropia [31] προσπαθεί να βρει τις αλλαγές των προτιμήσεων των αναγνωστών, με βάση τις λέξεις των κειμένων. Από κάθε κείμενο που έχει δει ο χρήστης εξάγει τις λέξεις και την συχνότητα που αυτές εμφανίζονται. Κατασκευάζει έναν γράφο αλληλεξάρτησης του οποίου οι κόμβοι είναι οι λέξεις των κειμένων και δημιουργεί ακμές

ανάμεσα στις λέξεις που συνεμφανίζονται μέσα στα κείμενα και σε ένα σταθερό παράθυρο (10) λέξεων. Υπολογίζει βάρη στις ακμές βάσει των συχνοτήτων εμφάνισης και συν εμφάνισης των κόμβων. Για να εκτιμήσει το πόσο θα αρέσει ένα νέο κείμενο στον χρήστη εξάγει τις λέξεις που υπάρχουν σε αυτό, τις ταξινομεί βάσει της συχνότητας εμφάνισης στο ιστορικό του και χρησιμοποιεί ένα μοντέλο διάχυσης ενέργειας για να υπολογίσει την στατιστική σημαντικότητα κάθε κόμβου. Η διάχυσή αρχίζει από τις λέξεις με την μικρότερη συχνότητα εμφανίσεις και εξαπλώνεται προς τους κόμβους με τους οποίους υπάρχει σύνδεση και έχουν μεγαλύτερη συχνότητα. Όσο ο χρήστης βλέπει νέα κείμενα το σύστημα ανανεώνει τα βάρη συσχέτισης των λέξεων και αφαιρεί και χαρακτηριστικά με μικρή στατιστική σημαντικότητα. Η χρήση του παραπάνω μοντέλου έχει σαν αποτέλεσμα το προφίλ του κάθε χρήστη να προσαρμόζεται γρήγορα σε σύνολα από λέξεις που είναι κοντά στο αντικείμενο που πραγματεύονται τα νέα κείμενα υποβαθμίζοντας μεμονωμένες λέξεις που εμφανίζονται συχνά στο ιστορικό του.

Μέθοδοι ομάδων ταξινομιτών (Ensemble Methods): Σε αυτή την κατηγορία αλγορίθμων χρησιμοποιούνται περισσότεροι από έναν ταξινομητές. Καθένας από τους οποίους είναι εκπαιδευμένος να εντοπίζει την εννοιολογική απόκλιση σε μία μόνο έννοια. Το ζητούμενο είναι να εντοπιστεί ποία έννοια (concept) είναι περισσότερο σχετική με τα νέα αντικείμενα, ώστε να το αναλάβει ο κατάλληλος ταξινομιτής. Παράδειγμα αυτής της κατηγορίας είναι ο αλγόριθμος LEARN^{++}.NSE που περιγράφεται στην αναφορά [32]. Ο αλγόριθμος δεν κρατάει καθόλου ιστορικό από τα δεδομένα που υπάρχουν στις ροές άλλα εκπαιδεύει με κάθε εισερχόμενη ομάδα στιγμιοτύπων έναν νέο ταξινομητή. Όταν πρέπει να κάνει προβλέψεις χρησιμοποιεί έναν αλγόριθμο ψηφοφορίας για να αποφασίσει ποιος από όλους τους ταξινομητές είναι ο πιο αντιπροσωπευτικός και τον επιλέγει.

3 Υβριδική προσαρμοστική μέθοδος πρόβλεψης μεταβαλλόμενων προτιμήσεων χρηστών

3.1. Βασική ιδέα

Σκοπός της εργασίας ήταν να δημιουργηθεί μια μέθοδος πρόβλεψης προτιμήσεων βασισμένη σε υβριδικό φιλτράρισμα η οποία να συνδυάζει την πληροφορία που μπορεί να παραχθεί από το φιλτράρισμα βάσει περιεχομένου με το συνεργατικό και το δημογραφικό φιλτράρισμα και να προσαρμόζεται στις αλλαγές των προτιμήσεων του κάθε χρήστη. Για την λειτουργία της μεθόδου όπως και με κάθε άλλη μέθοδο παραγωγής συστάσεων που περιγράφηκε, πρέπει να υπάρχει ένα σύνολο αντικειμένων I , ένα σύνολο από χαρακτηριστικά C , ένα σύνολο χρηστών U και ένα σύνολο δημογραφικών χαρακτηριστικών D . Είσοδος στην μέθοδο είναι το σύνολο από τις βαθμολογίες που έχουν εξαχθεί από το ιστορικό των χρηστών.

Η μέθοδος υλοποιεί αλγορίθμους φιλτραρίσματος βάσει περιεχομένου, δημογραφικού και συνεργατικού φιλτραρίσματος που υπολογίζουν την ομοιότητα χρηστών ξεχωριστά και συνδυάζει τα αποτελέσματα τους για να κάνει την τελική εκτίμηση. Κάθε μέθοδος παραγωγής συστάσεων που περιγράψαμε στο κεφάλαιο 2 εκτιμά τις πρωτιμήσεις για άγνωστων αντικειμένα βάσει γνωστών προτιμήσεων που υπήρχαν για όμοια αντικείμενα. Με τον όρο όμοια εννοούμε είτε ως προς τα δομικά τους χαρακτηριστικά είτε ως προς την απήχυση τους στους χρήστες. Στην πρωτινόμενη μέθοδο βασιζόμαστε στις εκτιμήσεις των προτιμήσεων των χρηστών για τα χαρακτηριστικά των αντικειμένων και από εκεί εκτιμούμε την προτίμηση για τα ίδια τα αντικείμενα. Κάθε μορφή φιλτραρίσματος είναι εμπλουτισμένη με μεθόδους αντιμετώπισης της εννοιολογικής απόκλισης βασισμένη σε τοποθέτηση βαρών σε στιγμιότυπα.

3.2. Φιλτράρισμα βάσει περιεχομένου με χαρακτηριστικά και βάρη προσαρμογής.

Η βασική ιδέα της μεθόδου είναι ότι η προτίμηση που έχει ο χρήστης u για το αντικείμενο i εκτιμάται από την προτίμηση του για το σύνολο C_i των χαρακτηριστικών του αντικειμένου.

Συγκεκριμένα η τιμή της προτίμησης $R_{u,i}$ υπολογίζεται από τον μέσο όρο των τιμών των προτιμήσεων $r_{u_o,c}$:

$$R_{u,i}=\frac{\sum_{c\in C_i} r_{u,c}}{\left(C_i\right)} \quad (1)$$

Έστω I_c το σύνολο των αντικειμένων που περιέχουν το χαρακτηριστικό c . Η προτίμηση $r_{u,c}$ του χρήστη u για ένα χαρακτηριστικό c εκτιμάται από τον μέσο όρο των προτιμήσεων που έχει εκφράσει για αντικείμενα που περιέχουν το συγκεκριμένο χαρακτηριστικό.

$$r_{u,c}=\frac{\sum_{i\in I_c} R_{u,i}}{\left(I_c\right)} \quad (2)$$

Πιο απλά χρησιμοποιούμε τις βαθμολογίες του χρήστη για να εκτιμήσουμε την προτίμηση του πάνω στα χαρακτηριστικά των αντικειμένων και χρησιμοποιούμε αυτές τις εκτιμήσεις για να συμπεράνουμε την προτίμηση προς τα άγνωστα αντικείμενα.

Για κάποια από τα χαρακτηριστικά C_i είναι πιθανό να μη μπορεί να γίνει εκτίμηση της προτίμησης $r_{u,c}$ γιατί απλά δεν υπάρχουν σε κανένα από τα αντικείμενα που βρίσκονται στο ιστορικό του χρήστη. Αυτά τα χαρακτηριστικά αγνοούνται και δεν τα συμπεριλαμβάνουμε στους υπολογισμούς. Προφανώς αν δεν μπορεί να εκτιμηθεί η προτίμηση για κανένα χαρακτηριστικό δεν μπορεί να εκτιμηθεί και η προτίμηση του αντικειμένου.

3.2.1 Η αξία των βαθμολογιών φθίνει στον χρόνο

Αν κάνουμε την παραδοχή πως όσο πιο πρόσφατες είναι οι βαθμολογίες που έχει δώσει ο χρήστης τόσο πιο αντιπροσωπευτικές είναι, τότε δεν πρέπει να έχουν όλες οι βαθμολογίες του ιστορικού το ίδιο βάρος. Έστω t_i η χρονική στιγμή που δόθηκε η βαθμολογία του αντικειμένου i , η σχέση (2) εμπλουτίζεται με ένα βάρος που υπολογίζεται από μια συνάρτηση εξασθένησης $f\left(t\right)$ η οποία φθίνει αναλόγως του t .

$$r_{u,c} = \frac{\sum_{i \in I_c} R_{u,i} * f(t_i)}{\sum_{i \in I_c} f(t_i)} \quad (3)$$

Στην μέθοδος μας κάνουμε την υπόθεση ότι η φθορά που παρουσιάζει η αξία κάθε βαθμολογίας είναι γραμμική ως προς τον χρόνο. Για αυτό η $f(t)$ που χρησιμοποιούμε είναι η γραμμική συνάρτηση εξασθένησης του χρησιμοποιείται και στην εργασία της αναφοράς [26].

$$f(t) = 1 + k - \frac{2k(t-1)}{n-1}$$

Για την εφαρμογή της χωρίζουμε το ιστορικό του χρήστη σε n ισομεγέθεις χρονικές περιόδους με την πρώτη χρονική περίοδο $t = 0$ να είναι η πιο πρόσφατη. Όσο αυξάνται το t η αξία των βαθμολογιών που ανήκουν στην αντίστοιχη χρονική περίοδο μειώνεται γραμμικά με το k .

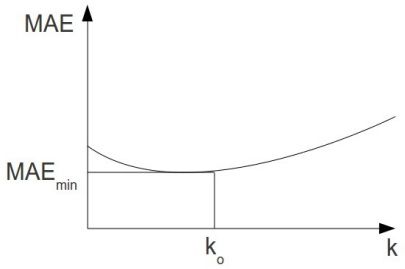
Επίσης χρειάζεται να υπολογίσουμε την σωστή τιμή του k η οποία θα ανταποκρίνεται κατά μέσο όρο στον ρυθμό με τον οποίο αλλάζουν τα ενδιαφέροντα των χρήστων. Κάνουμε την παραδοχή ότι το k είναι ίδιο για όλα τα χαρακτηριστικά των αντικειμένων. Αυτό δεν είναι πάντα αληθές υπολογιστικά είναι πιο ρεαλιστικό από το να υπολογισμό ένός k για κάθε χαρακτηριστικό ή κάθε ομάδα χαρακτηριστικών.

Για να υπολογίσουμε το k επιλέγουμε μέσα από το σύνολο δεδομένων, μία χρονική στιγμή t_s και χωρίζουμε τις χρονικές περιόδους βάσει αυτού. Σκοπός είναι να χρησιμοποιήσουμε όλα τα δεδομένα που υπάρχουν στις χρονικές περιόδους μέχρι και t_s ως σύνολο εκπαίδευσης και να υπολογίσουμε την βέλτιστη τιμή k με την οποία θα προβλέψουμε με μεγαλύτερη ακρίβεια τις βαθμολογίες που δόθηκαν μετά από αυτή τη χρονική στιγμή.

Για να ελέγξουμε την απόδοση του k χρησιμοποιούμε την μετρική του μέσου απόλυτου σφάλματος (Mean Average Error, MAE). Η συγκεκριμένη μετρική υπολογίζει το πόσο απέχουν οι προβλέψεις ενός αλγόριθμου πρόβλεψης από τις πραγματικές τιμές που έδωσε ο χρήστης.

$$MAE = \frac{\sum_{i=0}^{N} (r_i - R_i)}{N}$$

όπου N είναι ο αριθμός των προβλέψεων, r_i ή ι-οστή πρόβλεψη του αλγορίθμους και R_i ή ι-οστή βαθμολογία του χρήστη. Έστω $MAE(k)$ η συνάρτηση, που δοσμένου του k υπολογίζει όλες τις προβλέψεις και συγκρίνει τα αποτελέσματα με τις πραγματικές βαθμολογίες και επιστρέφει το μέσο απόλυτο σφάλμα. Αν δώσουμε στο k τιμές από ένα σύνολο [0,μ) η συνάρτηση της μετρικής $MAE(k)$ ως προς το k θα έχει μια μορφή σαν της γραφικής παράστασης που υπάρχει στην εικόνα 4. θα έχει δηλαδή ένα τοπικό ελάχιστο που θα είναι και το ολικό. Το να βρεθεί το τοπικό ελάχιστο είναι απλό, αρκεί να εφαρμόσουμε κάποιον από τους γνωστούς αλγόριθμους της αριθμητικής ανάλυσης που βρίσκουν ακρότατα.



Εικόνα 4. Η συνάρτηση της αλλαγής του μέσου σφάλματος σε σχέση με το k.

Το παραπάνω θα ισχύει γιατί εφόσον έχουμε κάνει την υπόθεση ότι η αξία των βαθμολογιών φθήνει γραμμικά θα υπάρχει μια συνάρτηση έστω $g(t) = at + b$ που θα είναι η βέλτιστη συνάρτηση εξασθένησης. Ο συντελεστής k της συνάρτησης $f(t)$ που χρησιμοποιούμε εμείς αποτελεί τον συνελεστή διεύθηνσης της ευθείας. Άρα θα υπάρχει ένα μόνο k για το οποίο η $f(t)$ και η $g(t)$ θα είναι παράλληλες και άρα η συνάρτηση $MAE(k)$ θα λαμβάνει την ελάχιστη τιμή της.

3.2.2 Εύρεση αντιπροσωπευτικών χαρακτηριστικών

Όταν εμφανίζεται ένα νέο αντικείμενο απαρτίζεται από χαρακτηριστικά που πιθανότατα

εμφανίζονται στο ιστορικό του χρήστη. Για να εκτιμήσουμε την αξία του αντικειμένου θα θέλαμε να ξέρουμε ποία απο τα χαρακτηριστικά του είναι πιο αντιπροσωπευτικά ως προς τον χρήστη που θα το βαθμολογήσεις. Δηλαδή ποια από τα χαρακτηριστικά του επηρεάζουν περισσότερο την κρίση του. Υπολογίζουμε την αξία που έχει ένα χαρακτηριστικό με μια από τις δύο παρακάτω ευριστικές:

α) Τα χαρακτηριστικά που έχουν την μεγαλύτερη συχνότητα εμφάνισης στο ιστορικό του χρήστη τον ενδιαφέρουν περισσότερο, άσχετο με το αν έχουν καλή βαθμολογία ή όχι. Ο τύπος του επιπλέον βάρους για ένα χαρακτηριστικό c και έναν χρήστη u είναι

$$w_{u,c}=freq(c) \quad (4)$$

β) Τα χαρακτηριστικά ενός αντικειμένου που εμφανίζονται πιο συχνά μαζί επηρεάζουν περισσότερο την κρίση του χρήστη, άσχετα από την βαθμολογία τους. Υπολογίζουμε την ομοιότητα των χαρακτηριστικών του αντικειμένου μεταξύ τους και πριμοδοτούμε τα χαρακτηριστικά που έχουν την μεγαλύτερη κατά μέσο όρο ομοιότητα. Για τον υπολογισμό της ομοιότητας χρησιμοποιούμε την μετρική conditional probability-based similarity. Ο τύπος της μετρικής ανάμεσα σε δύο χαρακτηριστικά c_1,c_2, είναι

$$w_{u,c}=\frac{freq(c_1 c_2)}{freq(c_2)} \quad (5)$$

όπου $freq(c_1 c_2)$ είναι η συχνότητα συνεμφάνισης των c_1,c_2 και $freq(c_2)$ η συχνότητα εμφάνισις του c_2 μέσα στο ιστορικό του χρήστη. Άρα ο τύπος της συνάρτησης υπολογισμού του έξτρα βάρους είναι

$$w_{u,c}=\frac{\sum_{c_0 \in C_i, c_o \neq C} P(c|c_0)}{(C_i)-1} \quad (6)$$

Η συνάρτηση (1) αλλάζει και γίνεται

$$R_{u,i} = \frac{\sum_{c \in C_i} r_{u,c} * w_{u,c}}{\sum_{c \in C_i} w_{u,c}} \quad (7)$$

3.3. Δημογραφικό φιλτράρισμα με χαρακτηριστικά και βάρη προσαρμογής.

Για το δημογραφικό φιλτράρισμα κατασκευάζεται ένα στερεότυπο για κάθε διαφορετική ομάδα τιμών που εμφανίζεται στα δημογραφικά χαρακτηριστικά των χρηστών. Στο προφίλ του στερεότυπου αποθηκεύεται ο μέσος όρος των βαθμολογιών όλων των χρηστών που ανήκουν στο στερεότυπο, ο μέσος όρος των συχνοτήτων εμφάνισης του κάθε χαρακτηριστικού κάθε ζευγαριού χαρακτηριστικών από τα αντικείμενα που βαθμολογήθηκαν, και τέλος ο μέσος όρος των τιμών των εκτιμήσεων που έγιναν για το πόσο αρέσει το κάθε χαρακτηριστικό στους χρήστες. Για την εκτίμηση της τιμής προτίμησης R_{si}^{u} ενός χρήστη u που ανήκει σε ένα στερεότυπο s για ένα αντικείμενο i χρησιμοποιείτε ο σταθμισμένος μέσος όρος όπως και στο φιλτράρισμα βάσει περιεχομένου.

$$R_{ui}^{s} = \frac{\sum_{c \in C_i} r_{s,c} * w_{s,c}}{\sum_{c \in C_i} w_{s,c}} \quad (8)$$

Τα βάρη $w_{s,c}$ υπολογίζονται και αυτά με μία από τις δύο ευριστικές που αναφέρθηκαν στην παράγραφο 4.2.

3.3. Συνεργατικό φιλτράρισμα με χαρακτηριστικά και βάρη προσαρμογής.

Το συνεργατικό φιλτράρισμα πραγματοποιείται βάσει των όμοιων χρηστών του συστήματος. Η ομοιότητα των χρηστών υπολογίζεται βάσει των εκτιμώμενων τιμών πάνω στα χαρακτηριστικά των αντικειμένων και όχι πάνω στα ίδια τα αντικείμενα. Η μετρική που χρησιμοποιείται για τον υπολογισμό της ομοιότητας είναι ο συνδυασμός της μετρικής ομοιότητας Jaccard με τον συντελεστή συσχέτισης Pearson όπως αναφέρθηκε στην παράγραφο 2.4.1.

Για να υπολογιστεί η προτίμηση ενός χρήστη u σε ένα αντικείμενο εκτιμούμε πρώτα την προτίμηση του πάνω στα χαρακτηριστικά του. Η εκτίμηση της προτίμησης κάθε χαρακτηριστικού c γίνεται από τον σταθμισμένο μέσο όρο της προτίμησης που έχει το σύνολο U_c των k κοντινότερων χρηστών.

$$r_{u,c}^{c}=\frac{\sum_{u_0 \in U_c} s\left(u,u_0\right)*r_{u_0,c}}{\sum_{u_0 \in U_i} s\left(u,u_0\right)} \quad (9)$$

Μέσω του ίδιου συνόλου και του σταθμισμένου μέσου όρου γίνεται και ο υπολογισμός των βαρών $w_{u,c}$.

$$w_{u,c}=freq\left(c\right)=\frac{\sum_{u_0 \in U_i} s\left(u,u_0\right)*freq_{u_0}\left(c\right)}{\sum_{u_0 \in U_i} s\left(u,u_0\right)} \quad (10)$$

ή

$$w_{u,c}=\frac{\sum_{c_0 \in C_i, c_o \neq C} \frac{\sum_{u_0 \in U_i} s\left(u,u_0\right)*P_{u_0}\left(c|c_0\right)}{\sum_{u_0 \in U_i} s\left(u,u_0\right)}}{\left(C_i\right)-1} \quad (11)$$

Όπου $freq_{u_0}\left(c\right)$ είναι η συχνότητα του χαρακτηριστικό c στο προφίλ του χρήστη u_0 και $P_{u_0}\left(c|c_0\right)$ είναι η ομοιότητα του χαρακτηριστικού c με το χαρακτηριστικό c_0 στο προφίλ του χρήστη u_0 . Η τελική εκτίμηση της τιμής γίνεται από τον μέσο όρο των εκτιμώμενων τιμών των χαρακτηριστικών

$$R_{ui}^{c}=\frac{\sum_{c \in C_i} r_{u,c}*w_{u,c}}{\sum_{c \in C_i} w_c} \quad (12)$$

3.4. Υβριδική προσαρμοστική μέθοδος πρόβλεψης μεταβαλλόμενων προτιμήσεων χρηστών

Το κάθε αντικείμενο του συστήματος μπορεί να αποτελείτε από μεγάλο αριθμό χαρακτηριστικών κάποια εκ των οποίων να μη υπάρχει σε κανένα από τα αντικείμενα που υπάρχουν στο ιστορικό του χρήστη. Όπως περιγράφηκε στις μεθόδους των παραγράφων 3.1, 3.2, 3.3 όταν δεν μπορεί να εκτιμηθεί η τιμή ενός χαρακτηριστικού αναγκάστηκα δεν το συμπεριλαμβάνουμε στους υπολογισμούς. Αποτέλεσμα είναι να πρέπει να εκτιμήσουμε την τιμή προτίμησης σε ένα αντικείμενο χωρίς να γνωρίζουμε την άποψη του χρήστη για όλα τα χαρακτηριστικά και να μειώνεται η απόδοση των προβλέψεων. Για να αντιμετωπίσουμε το πρόβλημα με τα αραιά δεδομένα συνδυάζουμε τις εκτιμήσεις από τις τρεις μεθόδους φιλτραρίσματος ώστε να μειώσουμε τον αριθμό των άγνωστων αντικειμένων.

Για να συνδυάσουμε τις εκτιμήσεις κάνουμε της υπόθεση ότι αφού το φιλτράρισμα βάσει περιεχομένου παρουσιάζει τον μεγαλύτερο βαθμό εξατομίκευσης θα εκφράζει και καλύτερα τις προτιμήσεις των χρηστών. Στην γενική περίπτωση δηλαδή θα κάνει τις καλύτερες εκτιμήσεις. Εν αναλογία το συνεργατικό φιλτράρισμα θα κάνει καλύτερες εκτιμήσεις από ότι το δημογραφικό.

Έστω $C_{ι,u}, C_{ι,c}, C_{ι,d} \in C_i$ τα υποσύνολο των χαρακτηριστικών του αντικειμένου i που μπορούν να εκτιμηθούν από το προφίλ του χρήστη, των ομοίων χρηστών και των στερεοτύπων αντίστοιχα. Για να εκτιμηθεί η προτίμηση που έχει ο χρήστης u για το αντικείμενο i από το σύνολο των χαρακτηριστικών C_i χρησιμοποιούμε την σχέση (13)

$$R_{ui} = \frac{\sum_{c_0 \in C_{ι,u}} r_{u,c} * w_{u,i} + \sum_{c_1 \in C_{i,c_0} - C_{i,c}} r^{c}_{u,c} * w_{c,c_1} + \sum_{c_2 \in C_{i,d} - C^{c}_{i,u} \cup C_{i,c}} r^{s}_{u,c} * w_{s,c_2}}{\left(c_0 \cup c_1 \cup c_2\right)} \quad (13)$$

όπου w_{c,c_i}, w_{s,c_i} είναι τα βάρη που υπολογίζονται από το συνεργατικό και το δημογραφικό φιλτράρισμα αντιστοίχως.

Η παραπάνω μέθοδος μπορεί να εκτελεστεί και με την απουσία του συνεργατικού ή του δημογραφικού φιλτραρίσματος και οι εκτιμήσεις να δοθούν από τους παρακάτω τύπους.

$$R_{ui}=\frac{\sum_{c_0\in C_{\iota,u}} r_{u,c}*w_{u,i}+\sum_{c_1\in C_{i,c_0}-C_{i,c}} r^{c}_{u,c}*w_{c,c_1}}{\left(c_0\cup c_1\right)} \quad (14)$$

$$R_{ui}=\frac{\sum_{c_0\in C_{\iota,u}} r_{u,c}*w_{u,i}+\sum_{c_2\in C_{i,d}-C_{i,u}} r^{s}_{u,c}*w_{s,c_2}}{\left(c_0\cup c_2\right)} \quad (15)$$

4 Αξιολόγηση αλγορίθμων

4.1. Σύνολο δεδομένωνs

Για τον έλεγχο των αλγόριθμων που δημιουργήθηκαν επιλέξαμε να χρησιμοποιήσουμε το σύνολο με τα δεδομένα από το GroupLens. Το GroupLens είναι ένα ερευνητικό εργαστήριο του τμήματος επιστήμης υπολογιστών του πανεπιστημίου της Μινεσότα και παρέχει δεδομένα από ένα ερευνητικό έργο που έχουν φτιάξει το MovieLens. Το MovieLens είναι μια διαδικτυακή υπηρεσία που οι χρήστες δέχονται προτάσεις για ταινίες και δίνουν βαθμολογίες από 0 έως 5. Τα δεδομένα λοιπόν του GroupLens περιέχουν τα δημογραφικά χαρακτηριστικά για ένα σύνολο από 6040 χρήστες, χαρακτηριστικά για 3900 ταινίες και 1.000.000 βαθμολογίες ταινιών που δόθηκαν σε διάρκεια δύο ετών. Το συγκεκριμένο σύνολο επιλέχτηκε γιατί είναι αρκετά δημοφιλές και συμπεριλαμβάνεται στην πειραματική διαδικασία των περισσοτέρων ερευνητικών εργασιών.

Στα δεδομένα των ταινιών το GroupLens παρέχει ελάχιστη πληροφορία, μόνο τον τίτλο και την κατηγορία στην οποία ανήκει. Εμείς για να ελέγξουμε τις μεθόδους μας χρειαζόμασταν τα χαρακτηριστικά τους τα οποία προμηθευτήκαμε από την βάση δεδομένων του IMDB. Η συγκεκριμένη ιστοσελίδα περιέχει πληροφορίες σχεδόν για όποια ταινία έχει γυριστεί και παρέχει την βάση δεδομένων της δωρεάν αρκεί να μη χρησιμοποιηθεί για εμπορικές εφαρμογές. Από τα δεδομένα των ταινιών επιλέξαμε να χρησιμοποιήσουμε μόνο τους ηθοποιούς. Ο περιορισμός στους ηθοποιούς έγινε γιατί θέλαμε όλα τα χαρακτηριστικά να ανήκουν στην ίδια κατηγορία ώστε να μπορούμε να βγάλουμε ασφαλή συμπεράσματα για την απόδοση των ευριστικών συναρτήσεων που αναφέρουμε στην παράγραφο 4.2.2. Αν τα χαρακτηριστικά μας ανήκαν σε περισσότερες από μία κατηγορίες ,για παράδειγμα ηθοποιοί και είδος ταινίας, πιθανότατα θα αντιμετωπίζαμε πρόβλημα με την χρήση των συχνοτήτων εμφανίσεις των χαρακτηριστικών. Χαρακτηριστικά που ανήκουν σε μικρό σύνολο όπως τα είδη θα κατέληγαν να έχουν υψηλή συχνότητα εμφάνισης και τελικά να επηρεάζουν περισσότερο χωρίς απαραίτητα να σημαίνει ότι έχουν και μεγαλύτερη αξία. Ως χαρακτηριστικά επιλέχτηκαν οι ηθοποιοί γιατί είναι πολυπληθείς και θεωρούμε πως σίγουρα παίζουν

σημαντικό ρόλο στο πώς αξιολογεί ο μέσος θεατής μια ταινία.

Δυστυχώς οι τίτλοι των ταινιών που παρέχει το GroupLens περιέχουν πολλά λάθη και είναι αδύνατο να εντοπιστούν όλες μέσα στην βάση του IMDB χωρίς να προηγηθεί επεξεργασία. Πραγματοποιήσαμε τις απαραίτητες διορθώσεις στους τίτλους χρησιμοποιώντας λογισμικό και όπου αυτό δεν ήταν δυνατό τις κάναμε χειροκίνητα. Το διορθωμένο σύνολο δεδομένων παρέχεται στο συνοδευτικό υλικό της εργασίας.

Για τους χρήστες τα δημογραφικά δεδομένα που παρέχονται είναι το φύλο τους, η ηλικιακή τους ομάδα το επάγγελμα τους και ο ταχυδρομικός κωδικός της περιοχής που διαμένουν. Επιλέξαμε να τα χρησιμοποιήσουμε όλα εκτός από τον ταχυδρομικό κωδικό. Για το δημογραφικό φιλτράρισμα φτιάχνουμε ένα στερεότυπο για κάθε συνδυασμό των χαρακτηριστικών που υπάρχει και η χρήση του ταχυδρομικού κωδικού θα αύξανε υπερβολικά των αριθμό των στερεοτύπων και θα μείωνε τον αριθμό των χρηστών που ανήκουν σε αυτές. Αυτό θα ήταν καθιστούσε άχρηστα τα στερεότυπα καθώς δεν θα υπήρχαν αρκετά στατιστικά δεδομένα για να διαφοροποιηθούν από το φιλτράρισμα βάσει περιεχομένου.

Επειδή στις μεθόδους μας χρησιμοποιούμε συναρτήσεις η απόδοση των οποίων έχουν σχέση με τον μέγεθος του χρονικού διαστήματος που ένας χρήστης εξέφρασε τις βαθμολογίες του βάλαμε ένα κατώφλι και θεωρήσαμε πως οι χρήστες που μας ήταν χρήσιμοι ήταν μόνοι όσοι χρησιμοποίησαν το σύστημα για περισσότερο από δύο εβδομάδες. Οπότε από τους 3900 χρήστες εμείς χρησιμοποιήσαμε 968 και από τις 1.000.000 βαθμολογίες τις 268.208. Ο μέσος όρος του χρονικού διαστήματος που χρησιμοποίησε ο κάθε χρήστης το σύστημα είναι 10 εβδομάδες.

4.2. Προετοιμασία πειραμάτων

Για την εκτέλεση των πειραμάτων χωρίσαμε το σύνολο των βαθμολογιών σε δύο υποσύνολα. Ένα που περιέχει το 75% των βαθμολογιών που κατέθεσε ο κάθε χρήστης και ένα με το άλλο 25%. Χρησιμοποιήσαμε το πρώτο ως σύνολο εκπαίδευσης για να κατασκευάσουμε τα προφίλ των χρηστών και το αλλά 25% για να ελέγξουμε την απόδοση.

Για την εκτέλεση των μεθόδων που χρησιμοποιούν την συνάρτηση εξασθένησης έπρεπε να υπολογιστεί ο συντελεστής εξασθένησης k. Αυτό πραγματοποιήθηκε σπάζοντας το 75% των βαθμολογιών που είναι στο σύνολο εκπαίδευσης σε δύο υποσύνολα. Το ένα που περιέχει το 90% των βαθμολογιών και χρησιμοποιείται για εκπαίδευση και το άλλο που περιέχει το 10% και χρησιμοποιήθηκε για τους ελέγχους της συνάρτησης αναζήτησης του βέλτιστου k.

Σε όλες τις μεθόδους συνεργατικού φιλτραρίσματος στον υπολογισμό τον κοντινότερων χρηστών κρατήσαμε μόνο όσους η απόσταση από τον χρήστη ήταν μεγαλύτερη από 0.7 χωρίς να διερευνηθεί ποια θα ήταν η βέλτιστη επιλογή.

4.3. Πειραματικά αποτελέσματα

Για να μετρήσουμε την απόδοση των αλγορίθμων μας χρησιμοποιήσαμε τα έξεις μέτρα απόδοσης

- Το μέσο απόλυτο σφάλμα

$$MAE = \frac{\sum_{i=0}^{N} \left(r_i - R_i \right)}{N}$$

- Τον αριθμό των προβλέψεων που μπόρεσαν να γίνουν
- Τον αριθμό των προβλέψεων που πέτυχαν ακριβώς την βαθμολογία του χρήστη.

Σε όλες τις μετρήσεις χρησιμοποιήσαμε τις στρογγυλοποιημένες εκτιμώμενες τιμές.

Τα πειράματα που εκτελέσαμε θέλαμε να ελέγξουμε ότι:

- Η εκτίμηση της προτίμησης βάσει των χαρακτηριστικών (3.2, 3.3, 3.4) έχει καλύτερα ή έστω τόσο καλά αποτελέσματα όσο η εκτίμηση βάσει των αντικειμένων. Δηλαδή όταν εκτιμούμε τις τιμές βάσει του σταθμισμένου μέσου όρου των τιμών στις προτιμήσεις των χαρακτηριστικών επιτυγχάνουμε καλύτερα αποτελέσματα από ότι όταν το κάνουμε μέσω των αντίστοιχων τιμών στα αντικείμενα.

- Ισχύει η υπόθεση ότι το φιλτράρισμα βάσει περιεχομένου παρέχει καλύτερες

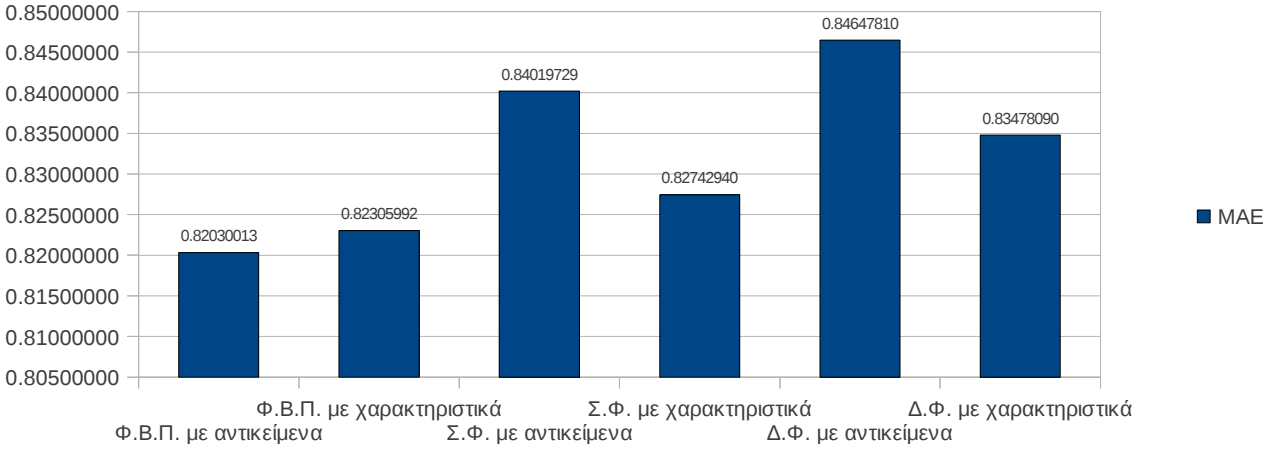
προβλέψεις από τα άλλα είδη φιλτραρίσματος και ότι το συνεργατικό φιλτράρισμα έχει καλύτερα αποτελέσματα από το δημογραφικό.

- Οι ευριστικές που προτείνουμε (παράγραφος 3.2) να χρησιμοποιηθούν βελτιώνουν την απόδοση των προβλέψεων.

- Ότι η τεχνική του υβριδικού φιλτραρίσματος που προτείνουμε έχει την μεγαλύτερη ακρίβεια και παράγει τις περισσότερες προβλέψεις.

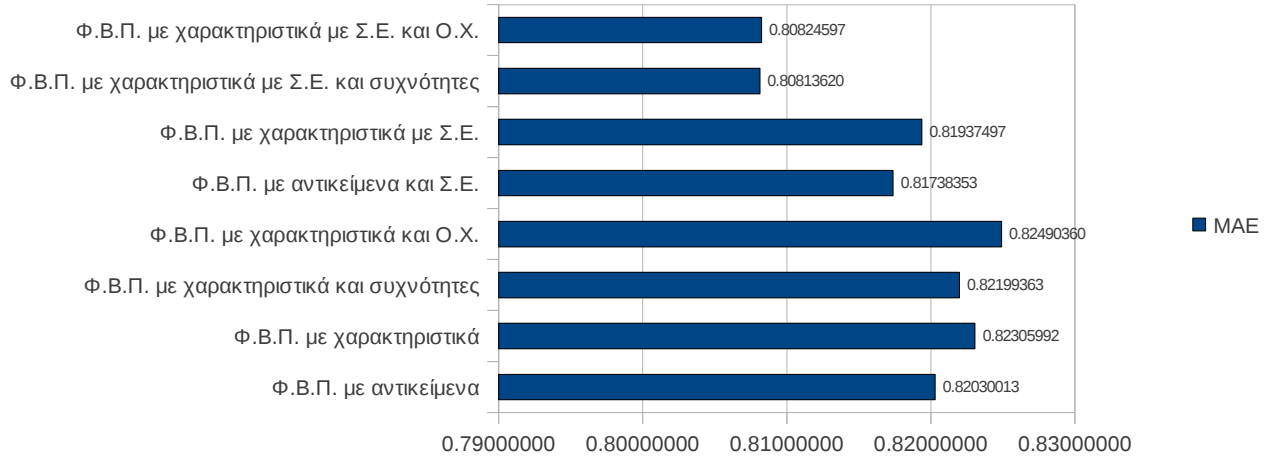
Για να είναι πιο ευανάγνωστα τα διαγράμματα χρησιμοποιήθηκαν κάποια ακρωνύμια, τα οποία είναι:

- Φ.Β.Π Φιλτράρισμα Βάσει Περιεχομένου
- Σ.Φ. Συνεργατικό Φιλτράρισμα
- Δ.Φ. Δημογραφικό Φιλτράρισμα
- Ο.Χ. Ομοιότητα Χαρακτηριστικών
- Σ.Ε. Συνάρτηση εξασθένισης
- Υ.Δ. Υβριδικό φιλτράρισμα



Διάγραμμα 1: Το μέσο απόλυτο σφάλμα για τα διαφορετικά είδη φιλτραρίσματος

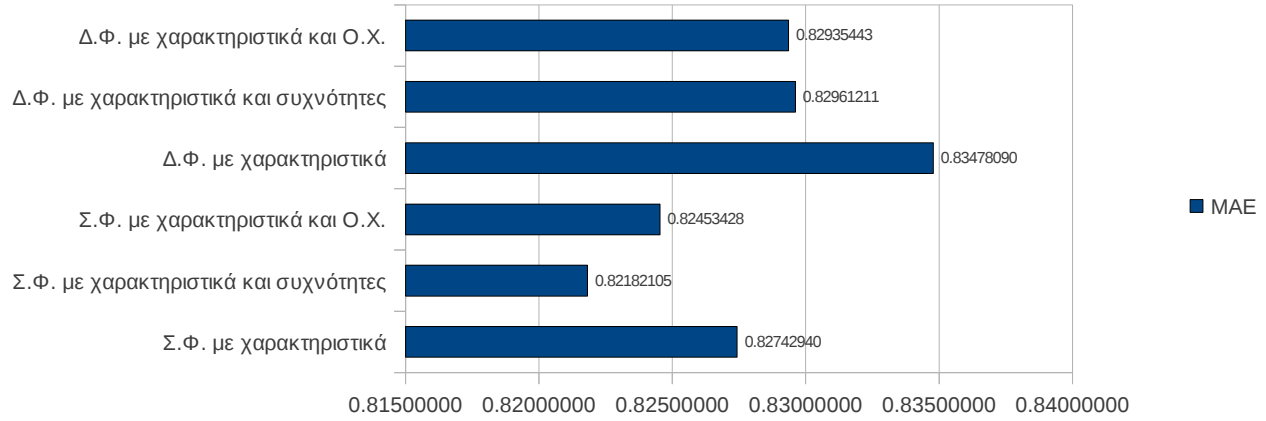
Από το διάγραμμα 1 φαίνεται ότι το φιλτράρισμα βάσει περιεχομένου και συγκεκριμένα χρησιμοποιώντας τα ίδια τα αντικείμενα έχει την μεγαλύτερη ακρίβεια. Στο συνεργατικό και το δημογραφικό φιλτράρισμα η χρήση των χαρακτηριστικών έχει πολύ καλύτερα αποτελέσματα.



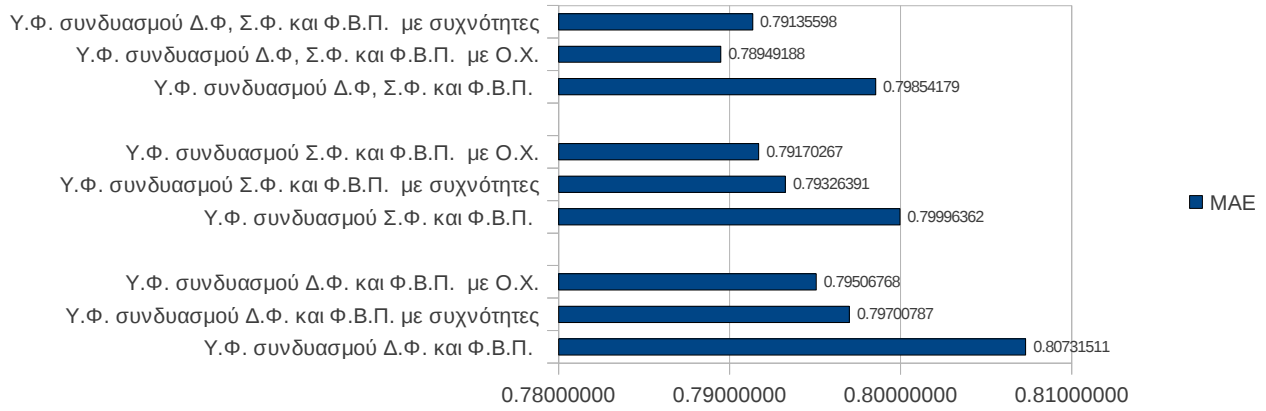
Διάγραμμα 2: Το μέσο απόλυτο σφάλμα για τις βελτιώσεις που μπορούν να γίνουν στο φιλτράρισμα βάσει περιεχομένου

Στο διάγραμμα 2 βλέπουμε ότι η εφαρμογή των ευριστικών συναρτήσεων που περιγράφονται στην παράγραφο 4.2 βελτιώνουν την απόδοση της φιλτραρίσματος βάσει περιεχομένου με

χαρακτηριστικά και το καθιστούν αποδοτικότερο από το αντίστοιχο με την χρήση αντικειμένων πάνω στα οποία δεν μπορούν να εφαρμοστούν οι ευριστικές αφού δεν υφίστανται συχνότητες. Αντίστοιχα η βελτίωση που μπορεί να πραγματοποιηθεί από την χρήστη των ευριστικών στο δημογραφικό και στο συνεργατικό φιλτράρισμα φαίνεται και στο διάγραμμα 3.



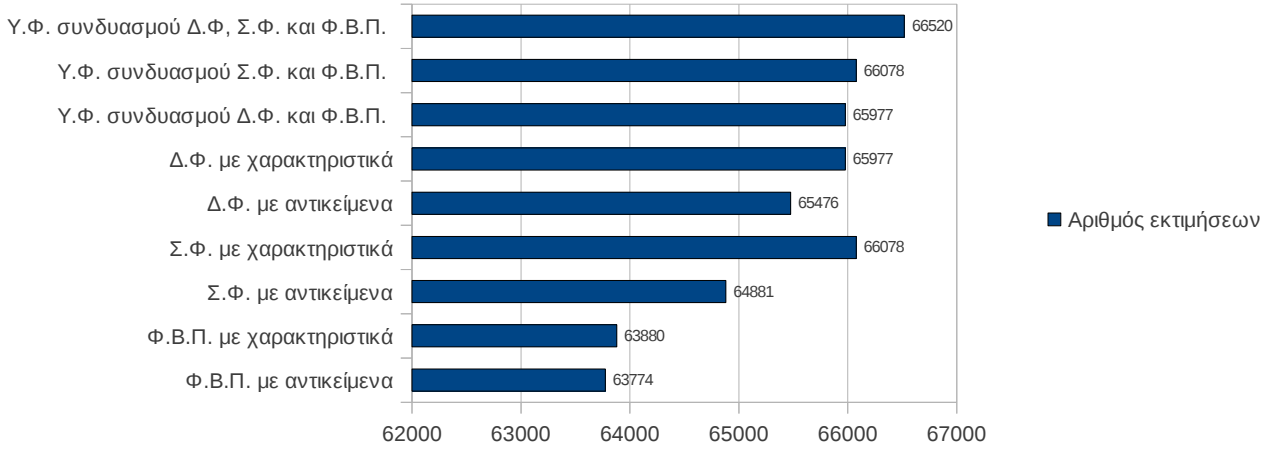
Διάγραμμα 3: Το μέσο απόλυτο σφάλμα για τις βελτιώσεις που μπορούν να γίνουν στο δημογραφικό και το συνεργατικό φιλτράρισμα αν χρησιμοποιηθούν οι ευριστικές συναρτήσεις



Διάγραμμα 4: Το μέσο απόλυτο σφάλμα για τις διαφορετικές εκδοχές του υβριδικού φιλτραρίσματος

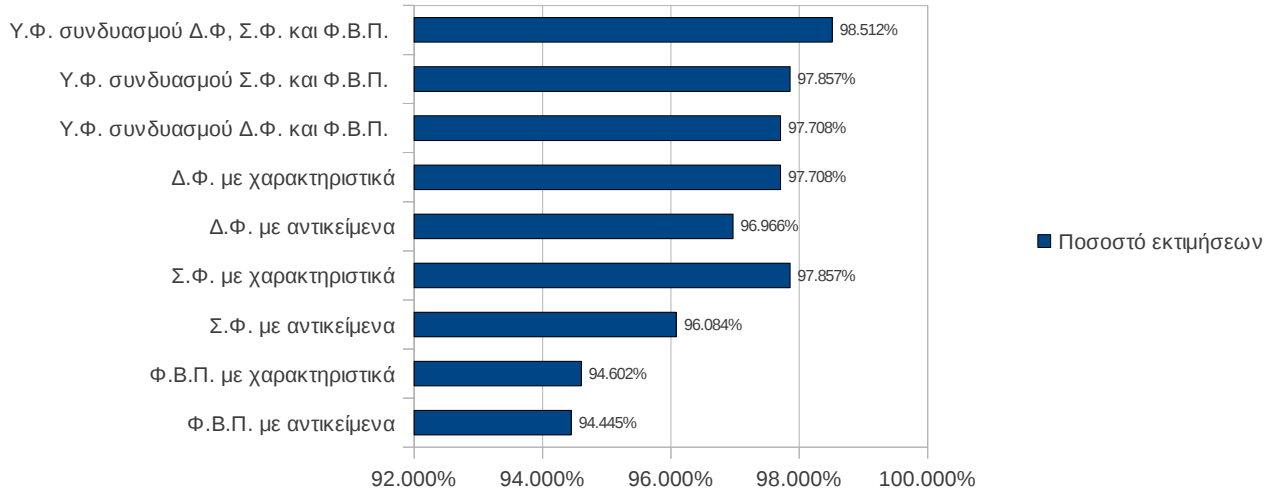
Στο διάγραμμα 4 φαίνεται η απόδοση του υβριδικού φιλτραρίσματος αναλόγως της ευριστικής συνάρτησης που θα χρησιμοποιηθεί και του ποιες μεθόδους θα συνδυαστούν. Βλέπουμε ότι

παρουσιάζει την μεγαλύτερη ακρίβεια πρόβλεψης βαθμολογίων από οποιαδήποτε άλλη μέθοδο της οποίας τα αποτελέσματα υπάρχουν στα διαγράμματα 1 έως 3.



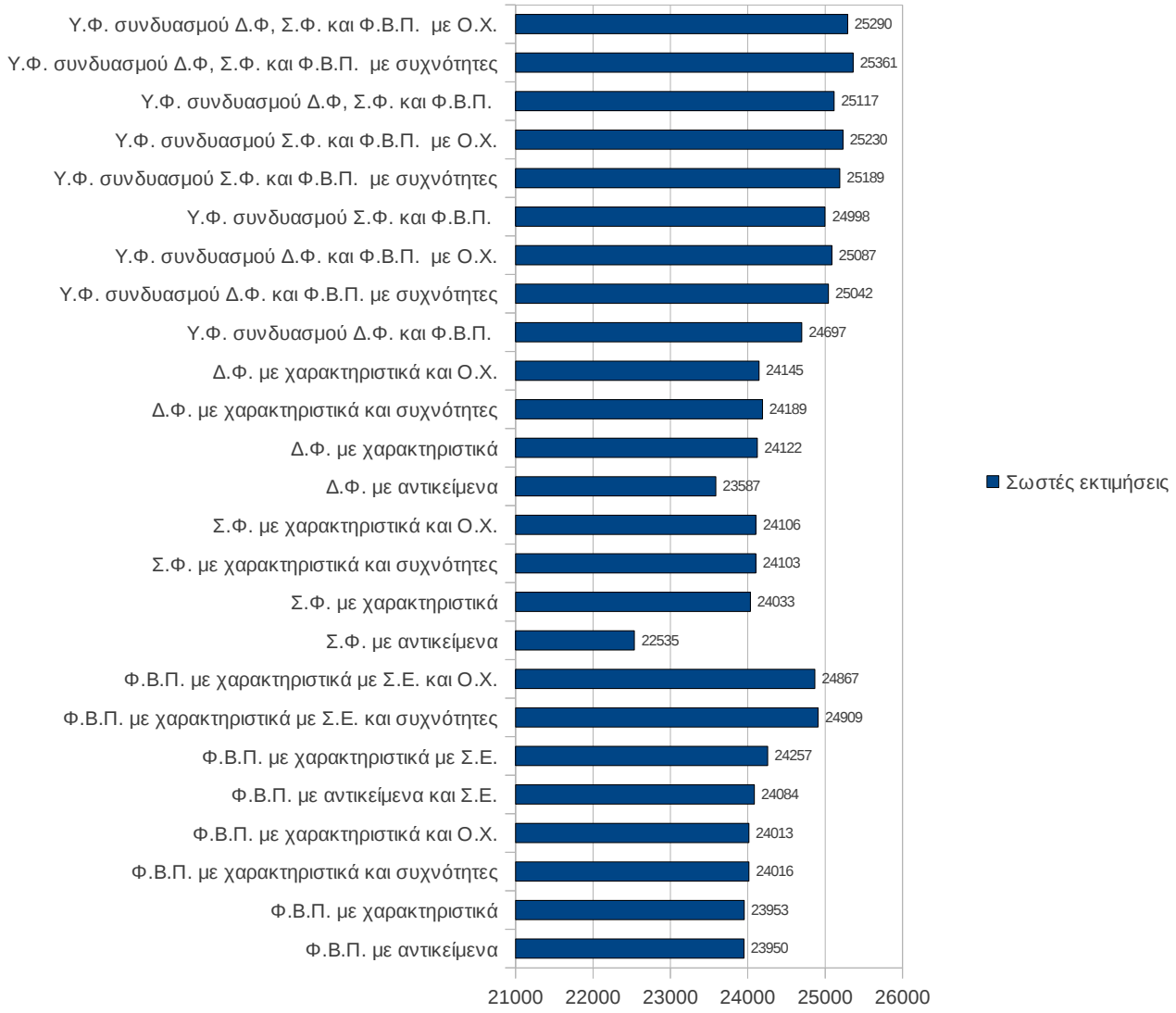
Διάγραμμα 5: Ο αριθμός των εκτιμήσεων που κατάφερε να κάνει η κάθε μέθοδος

Σε κάθε πείραμα ζητήθηκαν να πραγματοποιηθούν 67525 εκτιμήσεις βαθμολογιών. Η κάθε μέθοδος μπόρεσε να απαντήσει σε ένα ποσοστό από αυτές Στο διάγραμμα 5 υπάρχει ο ακριβής αριθμός των προβλέψεων και στο διάγραμμα 6 το ποσοστό επί του συνολικού αριθμού. Προφανώς η χρήση των ευριστικών δεν αλλάζει τον αριθμό των προβλέψεων που μπόρεσαν να πραγματοποιηθούν και για αυτό δεν συμπεριλαμβάνονται στα διαγράμματα.



Διάγραμμα 6: Το ποσοστό των εκτιμήσεων που κατάφερε να κάνει η κάθε μέθοδος

Τέλος στο διάγραμμα 7 μπορούμε να δούμε τον αριθμό των απόλυτα επιτυχημένων προβλέψεων που πέτυχε η κάθε μέθοδος.



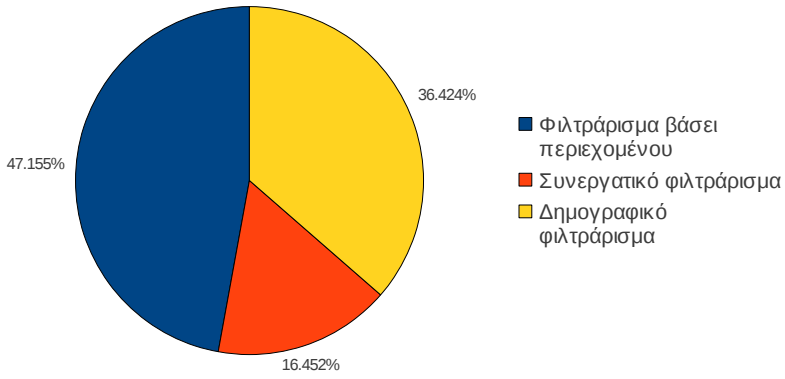
Διάγραμμα 7: Ο αριθμός των απόλυτα σωστών προβλέψεων

4.3.1 Απόπειρα συνδυασμού εκτιμήσεων

Από τα αποτελέσματα της προηγούμενης παραγράφου φαίνεται ότι οι εφαρμογή όσων υποθέσεων έγιναν στην νέα μέθοδος που προτείνουμε βελτιώνουν την ακρίβεια των προβλέψεων και άρα είναι χρήσιμες. Η μέθοδος μας βασίζεται στην ιδέα ότι το φιλτράρισμα βάσει περιεχομένου έχει πάντα καλύτερες εκτιμήσεις από το συνεργατικό φιλτράρισμα, το

οποίο έχει καλύτερες εκτιμήσεις από το δημογραφικό φιλτράρισμα. Η υπόθεση αυτή πιθανότατα δεν θα είναι πάντα αληθής και για αυτό θέλαμε να δούμε αν μπορεί να βελτιωθεί η μέθοδος.

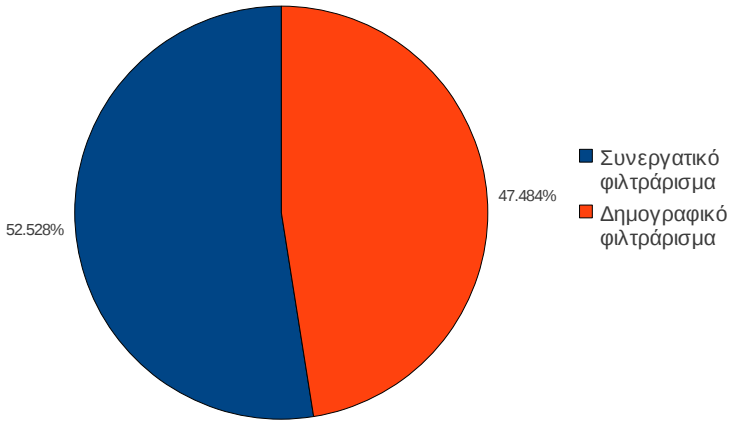
Τρέξαμε ένα πείραμα όπου χρησιμοποιήσαμε το 90% του σύνολου εκπαίδευσης για να εκτιμήσουμε τις τιμές του υπόλοιπου 10%. Σε αυτό το πείραμα κρατήσαμε στατιστικά για το ποιο είναι το ποσοστό όπου η εκτίμηση του φιλτραρίσματος βάσει περιεχομένου υπερτερεί του συνεργατικού και του δημογραφικού φιλτραρίσματος και του ποσοστού πού το συνεργατικό φιλτράρισμα υπερτερεί του δημογραφικού. Δηλαδή σε τι ποσοστό πράγματι το να επιλέξουμε πρώτα την εκτίμηση για ένα χαρακτηριστικό από το φιλτράρισμα βάσει περιεχομένου είναι η καλύτερη επιλογή.



Διάγραμμα 8: Τα ποσοστά που αναλογούν σε ποία μέθοδος έχει τις καλύτερες εκτιμήσεις χαρακτηριστικών όταν το χαρακτηριστικό υπάρχει στο προφίλ του χρήστη.

Στο διάγραμμα 8 μπορούμε να δούμε ότι στην περίπτωση που το προφίλ του χρήστη (φιλτράρισμα βάσει περιεχομένου) μπορεί να εκτιμήσει την προτίμηση για ένα χαρακτηριστικό τότε στο 47,155% των περιπτώσεων αυτή η εκτίμηση θα είναι πιο κοντά στην πραγματικότητα. Δηλαδή θα απέχει λιγότερο από την προτίμηση του συγκεκριμένου αντικειμένου για το οποίο θέλουμε να προβλέψουμε τη βαθμολογία που θα του έδινε ο χρήστης. Στο 16,452% των περιπτώσεων την βέλτιστη εκτίμηση μπορούμε να την πάρουμε από τα προφίλ των όμοιων χρηστών (συνεργατικό φιλτράρισμα) και το 36,424% από τα προφίλ των χρηστών με όμοια δημογραφικά χαρακτηριστικά.

Από το ίδιο πείραμα μετρήσαμε τα αντίστοιχα ποσοστά για τις περιπτώσεις που το προσωπικό προφίλ δεν μπορεί να εκτιμήσει την προτίμηση για ένα χαρακτηριστικό.



Διάγραμμα 9: Τα ποσοστά που αναλογούν σε ποία μέθοδος έχει τις καλύτερες εκτιμήσεις χαρακτηριστικών όταν το χαρακτηριστικό δεν υπάρχει στο προφίλ του χρήστη.

Στο διάγραμμα 9 βλέπουμε ότι σε αυτή την περίπτωση τα προφίλ των όμοιων χρηστών κατά 52,528% θα κάνει καλύτερη πρόβλεψη από τις εκτιμήσεις του προφίλ των χρηστών με όμοια δημογραφικά χαρακτηριστικά.

Προσπαθήσαμε να χρησιμοποιήσουμε αυτά τα ποσοστά ως πιθανότητες για να συνδυάσουμε τις εκτιμήσεις που μπορούν να γίνουν από το προσωπικό προφίλ του χρήστη, τα προφίλ των όμοιων χρηστών και τα προφίλ των χρηστών με κοινά δημογραφικά χαρακτηριστικά. Συγκεκριμένα η εκτίμηση ενός χρήστη u για ένα αντικείμενο i που έχει χαρακτηριστικά C υπολογίζεται από τον τύπο

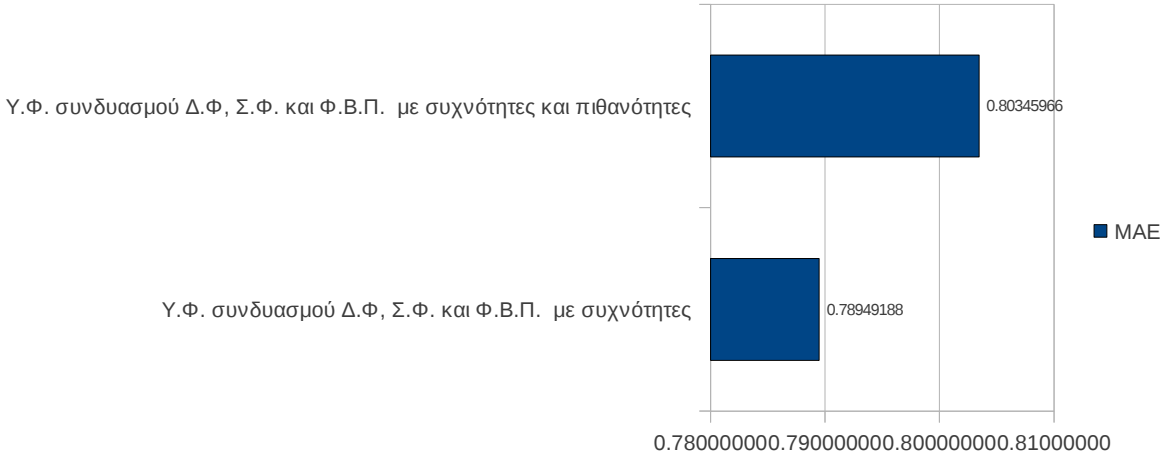
$$R_{ui}=\frac{\sum_{c_0\in C_{i,u}} r_{ucs,c}*w_{u,i}+\sum_{c_1\in C_{i,c_0}-C_{i,c}} r_{cs,c}*w_{c,c_1}+\sum_{c_2\in C_{i,d}-C_{i,u}\cup C_{i,c}} r^{s}_{u,c}*w_{d,c_2}}{\left(c_0\cup c_1\cup c_2\right)} \quad (16)$$

$$r_{ucs,c}=0{,}471*r_{u,c}+0{,}162*r^{c}_{u,c}+0{,}364*r^{s}_{u,c} \quad (17)$$

$$r_{cs,c} = 0{,}525 * r^{c}_{u,c} + 0{,}474 * r^{s}_{u,c} \quad (18)$$

Όπου $r^{c}_{u,c}$ είναι η εκτίμηση της τιμής της προτίμησης του χρήστη u για το χαρακτηριστικό c μέσω του συνόλου των πιο όμοιων χρηστών του. Και $r^{s}_{u,c}$ είναι η εκτίμηση της τιμής της προτίμησης του χρήστη u για το χαρακτηριστικό c μέσω του συνόλου των χρηστών που έχουν όμοια χαρακτηριστικά.

Εφαρμόσαμε τον τύπο 16 στα δεδομένα που έγιναν τα πειράματα της παραγράφου 5.3 και συγκρίναμε τα αποτελέσματα με την βέλτιστη εκδοχή της μεθόδου όπως προέκυψε από τα προηγούμενα πειράματα.



Διάγραμμα 10: Η σύγκριση του μέσου απόλυτου σφάλματος με την χρήση των πιθανοτήτων και χωρίς.

Τα αποτελέσματα με την χρήση των πιθανοτήτων είναι χειρότερα από την προηγούμενη εκδοχή και για αυτό η μέθοδος απορρίφθηκε.

5 Συμπεράσματα

5.1. Συμπεράσματα και μελλοντική διερεύνηση

Από τα αποτελέσματα των πειραμάτων του κεφαλαίου 4 συμπεραίνουμε ότι η υβριδική μέθοδος φιλτραρίσματος που σχεδιάσαμε και υλοποιήσαμε με τα συγκεκριμένα δεδομένα παράγει περισσότερες και ακριβέστερες προβλέψεις από τις κλασικές μεθόδους που χρησιμοποιούνται συνήθως. Οι μέθοδοι συνεργατικού και δημογραφικού φιλτραρίσματος αποδίδουν γενικά καλύτερα όταν βασίζονται στα χαρακτηριστικά των αντικειμένων και όχι στα ίδια τα αντικείμενα.

Το φιλτράρισμα βάσει περιεχομένου όταν χρησιμοποιεί τα χαρακτηριστικά των αντικειμένων για να βελτιώσει την απόδοση του χρειάζεται συναρτήσεις που να αναγνωρίζουν τα αντιπροσωπευτικά χαρακτηριστικά ενός αντικειμένου. Οι δύο ευριστικές συναρτήσεις προσδιορισμού των πιο αντιπροσωπευτικών χαρακτηριστικών που προτείνουμε λειτούργησαν θετικά και φαίνεται να αντιμετωπίζουν ως ένα βαθμό το θέμα της εννοιολογικής απόκλισης. Από τις δύο ευριστικές στην περίπτωση του συνεργατικού και του υβριδικού φιλτραρίσματος η ευριστική με την χρήση των συχνοτήτων δουλεύει καλύτερα από ότι η ευριστική της ομοιότητα των χαρακτηριστικών. Για να ερευνήσουμε όμως περισσότερο την απόδοση των συγκεκριμένων ευριστικών και να αποφανθεί ποια είναι προτιμότερα θα πρέπει να εφαρμόσουμε την μέθοδο και σε άλλα δεδομένα ίσως από την θεματική περιοχή των άρθρων όπου οι συνεμφανήσεις χαρακτηριστικών να είναι συχνότερες.

Από τα δεδομένα μας φαίνεται ότι η εφαρμογή της συνάρτησης εξασθένησης βελτιώνει λίγο τα αποτελέσματα αλλά μάλλον είναι κάλο να εκτελεστούν ξανά τα πειράματα για μικρότερο σύνολο χρηστών. Μάλλον θα πρέπει να επιλέξουμε όσους έχουν εκφράσει τις βαθμολογίες τους μέσα σε μεγαλύτερο χρονικό διάστημα για παράδειγμα για ένα χρόνο και να δούμε καλύτερα την επίδραση του χρόνου στις αλλαγές προτιμήσεων ταινιών.

Τέλος στην συγκεκριμένη υβριδική μέθοδο που προτείνουμε δεχόμαστε πάντα ότι το φιλτράρισμα βάσει περιεχομένου έχει καλύτερα αποτελέσματα από το συνεργατικό και αυτό

με την σειρά του καλύτερα από το δημογραφικό φιλτράρισμα. Αυτή η υπόθεση όπως αποδεικτικέ στα συγκεκριμένα δεδομένα βελτιώνει τα αποτελέσματα αλλά δεν είναι απολύτως αληθής. Θα πρέπει να ερευνήσουμε περισσότερο το πότε θα πρέπει να εμπιστευτούμε περισσότερο τα προφίλ των όμοιων sχρηστών ή των στερεοτύπων ή πότε θα πρέπει να συνδυάσουμε τις εκτιμήσεις με κάποια (γραμμική) συνάρτηση.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Clustering\	Συσταδοποίηση
Collaborative filtering recommendation systems	Συστήματα συστάσεων βασισμένα στο συνεργατικό φιλτράρισμα
Concept	Έννοια
Concept drift	Εννοιολογική απόκλιση
Content-based filtering recommendation systems	Συστήματα συστάσεων που χρησιμοποιούν φιλτράρισμα βάσει περιεχομένου
Data mining	Εξόρυξη δεδομένων
Decay function	Συνάρτηση εξασθένισης
Demographic-based filtering recommendation systems	Συστήματα συστάσεων βασισμένα στο δημογραφικό φιλτράρισμα
Dessimilarity	Διαφορά ομοιότητας
Extrapolation	Παρέκταση
Gradual concept drift	Σταδιακή εννοιολογική απόκλιση
Hybrid-based filtering recommendation systems	Συστήματα συστάσεων βασισμένα στο υβριδικό φιλτράρισμα
Instant concept drift	Απότομη εννοιολογική απόκλιση
Information retrieval	Ανάκτηση πληροφορίας
Interest drift	Αλλαγή των προτιμήσεων του χρήστη
local concept drift	Τοπική εννοιολογική απόκλιση
Log files	Αρχεία ημερολογίου
Mean Average Error	Μέσο απόλυτο σφάλμα
Profile injection	Εισαγωγής ψεύτικων προφίλ
Recommender systems	Συστημάτων συστάσεων
Semantic information	Σημασιολογική πληροφορία
Similarity	Ομοιότητα
Sparsity problem	Το πρόβλημα της σποραδικότητας

Αναφορές

[1] Ken Lang, NewsWeeder: Learning to Filter Netnews, 1995R.J. Mooney and L. Roy,: Content-Based Book Recommending Using Learning for Text Categorization, Proc. ACM SIGIR '99, Workshop Recommender Systems: Algorithms and Evaluation, 1999

[2] M. Pazzani and D. Billsus, Learning and Revising User Profiles: The Identification of Interesting Web Sites, Machine Learning, vol. 27, pages 313-331, 1997

[3] Meteren, R. V. & Someren, Using Content-Based Filtering for Recommendation, MLnet / ECML2000 Workshop, 2000

[4] Michael J. Pazzani & Daniel Billsus, Content-Based Recommendation Systems, In, The Adaptive Web, pages 325-341, 2007

[5] Yehuda Koren , Cohen, W., Fast Effective Rule Induction. In: Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, pages 115-123, 1995

[6] Exploiting learning techniques for the acquisition of user stereotypes and communities Source Proceedings of the seventh international conference on User modeling

[7] Shanle Ma, Xue Li, Yi Ding and Maria E. Orlowska, A Recommender System with Interest-Drifting , In: Web Information Systems Engineering – WISE 2007, pages 633-642, 2007

[8] Tsvi Kuflik and Bracha Shapira and Peretz Shoval, Stereotype-Based versus Personal-Based Filtering Rules in Information Filtering Systems, In: Journal of the American Society for Information Science and Technology, volume 54, pages 243-250, 2003

[9] Michael J. Pazzani, A Framework for Collaborative, Content-Based and Demographic Filtering, pages: 393-408, 1999

[10] Laurent Candillier Contact Information, Frank Meyer and Marc Boullé, Comparing State-of-the-Art Collaborative Filtering Systems, In: Machine Learning and Data Mining in Pattern Recognition, pages 548-562, 2007

[11] Laurent Candillier, Frank Meyer, Françoise Fessant, Designing Specific Weighted Similarity Measures to Improve Collaborative Filtering Systems, In: Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, pages 242-255, 2008

[12] Badrul Sarwar and George Karypis and Joseph Konstan and John Riedl, Item-Based Collaborative Filtering Recommendation Algorithms, 2001

[13] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In WWW '01: Proceedings of the 10th international conference on World Wide Web, pages 285–295, New York, NY, USA, 2001. ACM.

[14] Bamshad Mobasher, Robin Burke and JJ Sandvig , Model-Based Collaborative Filtering as a Defense Against Profile Injection Attacks, 2006,

[15] D Pierrakos, G Paliouras, C Papatheodorou, C Spyropoulos, KOINOTITES: A Web

Usage Mining Tool for Personalization, In: Proc. of the Panhellenic Conference on Human Computer Interaction, Patras

[16] Gediminas Adomavicius and Alexander Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, In: IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Volume 15, pages 734-749, 2005

[17] Michael Leben , Applying Item-based and User-based collaborative filtering on the Netflix data 2008

[18] Ralf Klinkenberg , Predicting Phases in Business Cycles Under Concept Drift, 2003

[19] Tsymbal, A., The problem of concept drift: definitions and related work, In: Technical Report TCD-CS-2004-15,2004.

[20] Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas , Incremental Clustering for the Classification of Concept-Drifting Data Streams

[21] Fan, W. Systematic data selection to mine concept-drifting data streams. In: Tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining. Seattle, WA, USA: ACM Press: pages 128-137, 2004

[22] Klinkenberg, R., Learning Drifting Concepts: Example Selection vs. Exam- ple Weighting Intelligent Data Analysis, In: Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, volume 8(3), pages 281-200. 2004

[23] Widmer, G. and Kubat, M., Learning in the Presense of Concept Drift and Hidden Contexts, In: Machine Learning, volume 23(1), pages 69-101, 1996.

[24] J. Schlimmer and R. Granger, Beyond incremental processing: Tracking concept drift. In: Proc. 5th National Conference on Artificial Intelligence, pages 502-507, 1986.

[25] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. In: Machine Learning, Volume 23, pages 69–101, 1996

[26] Koychev I. and Schwab I., Adaptation to Drifting User's Interests, In: Proc. of ECML2000 Workshop: Machine Learning in New Information Age, Barcelona, Spain, pages 39-45, 2000

[27] Dries Anton, Rückert Ulrich, Adaptive Concept Drift Detection, In: Statistical Analysis and Data Mining, volume 2, issue 5-6, pages 311-327, 2009

[28] Alexey Tsymbala, Mykola Pechenizkiy, Pádraig Cunningham, Seppo Puuronen, Dynamic Integration of Classifiers for Handling Concept Drift, In: Information Fusion, pages 56-68, 2008

[29] Yi Ding, Xue Li, Time Weight Collaborative Filtering, In: Proceedings of the 14th ACM international conference on Information and knowledge management, pages: 485-492, 2005

[30] Nikolaos Nanas, Victoria S. Uren, Anne N. De Roeck, Nootropia: A User Profiling Model Based on a Self-Organising Term Network, In: Proceedings of ICARIS 04, pages146-160

[31] Matthew Karnick, Metin Ahiskali, Michael D. Muhlbaier and Robi Polikar , Learning Concept Drift in Nonstationary Environments Using an Ensemble of Classifiers Based

Approach, World Congress on Computational Intelligence / IEEE International Joint Conference on Neural Networks 2008

[32] Sofus Macskassy, A Comparison of Two On-line Algorithms that Adapt to. Concept Drift

[33] Nguyen Duy Phuong , Le Quang Thang and Tu Minh Phuong،A Graph-Based Method for Combining Collaborative and Content-Based Filtering, In: PRICAI 2008: Trends in Artificial Intelligence, pages 859-869, 2008s