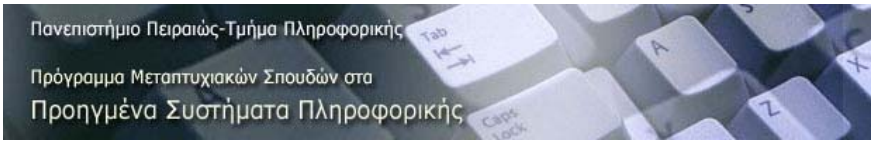
Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**ΥΛΟΠΟΙΗΣΗ ΚΑΙ ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΜΕΘΟΔΩΝ ΣΥΝΕΡΓΑΤΙΚΗΣ ΔΙΗΘΗΣΗΣ**
Ονοματεπώνυμο Φοιτητή	**ΗΡΑΚΛΕΙΔΗΣ ΓΕΩΡΓΙΟΣ ΤΟΥ ΒΑΣΙΛΕΙΟΥ**
Αριθμός Μητρώου	**ΜΠΣΠ09050**
Κατεύθυνση	**ΔΙΚΤΥΟΚΕΝΤΡΙΚΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**
Επιβλέπων	**ΤΣΙΧΡΙΝΤΖΗΣ ΓΕΩΡΓΙΟΣ, ΚΑΘΗΓΗΤΗΣ**



Ημερομηνία Παράδοσης **Ιούνιος 2011**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
ΤΖΙΧΡΙΝΤΖΗΣ ΓΕΩΡΓΙΟΣ ΚΑΘΗΓΗΤΗΣ	ΚΩΝΣΤΑΝΤΟΠΟΥΛΟΣ ΧΑΡΑΛΑΜΠΟΣ ΛΕΚΤΟΡΑΣ	ΠΙΚΡΑΚΗΣ ΑΓΓΕΛΟΣ ΛΕΚΤΟΡΑΣ

ΠΕΡΙΛΗΨΗ

Τα συστήματα σύστασης προσπαθούν να εξομοιώσουν τις προτιμήσεις των χρηστών, έχοντας σαν στόχο την εκτίμηση του πόσο ενδιαφέροντα θα είναι κάποια αντικείμενα, πληροφορίες, ή ακόμα και υπηρεσίες για το χρήστη και την υποβολή προτάσεων για στοιχεία που θα αγοράσει ένα άτομο ή θα εξετάσει. Αυτά τα συστήματα χρησιμοποιούνται συνεχώς στο διαδίκτυο και έχουν γίνει βασικό συστατικό πλέον του ηλεκτρονικού εμπορίου και της ανάκτησης πληροφοριών, παρέχοντας προτάσεις που αποτελεσματικά φιλτράρουν μεγάλους χώρους πληροφοριών ώστε κάθε χρήστης να μπορεί να κατευθύνεται προς τα στοιχεία που ανταποκρίνονται με τον καλύτερο τρόπο στις ανάγκες και τις προτιμήσεις του. Η τρομακτική αύξηση των πληροφοριών και αντικειμένων που είναι διαθέσιμα στο διαδίκτυο, καθώς και ο συνεχώς αυξανόμενος αριθμός επισκεπτών τα τελευταία χρόνια, θέτουν κάποιες προκλήσεις στα συστήματα σύστασης. Αυτές είναι, η παραγωγή σωστών προτάσεων, ο υπολογισμός μεγάλου αριθμού συστάσεων σε πολύ μικρό χρονικό διάστημα και η επιτυχής εκτέλεση ενός μεγάλου ποσοστού αιτημάτων σύστασης που δέχεται το σύστημα. Αρκετές τεχνικές όπως η συνεργατική διήθηση, η βασισμένη στο περιεχόμενο και η δημογραφική, όπως και συνδυασμοί τους, έχουν χρησιμοποιηθεί για την δημιουργία συστάσεων. Αυτή η μεταπτυχιακή διατριβή έρευνά ένα πλήθος από αλγόριθμους συνεργατικής διήθησης, αναλύοντας τα βήματα για την δημιουργία ενός τέτοιου συστήματος και συγκρίνει ποιοι αλγόριθμοι έχουν καλύτερα αποτελέσματα απέναντι στις προκλήσεις που αντιμετωπίζουν τα συστήματα σύστασης.

ABSTRACT

Recommender systems try to simulate people preferences for the purpose of estimate how much a user will be interested in some objects, information or services and to place a suggestion for items an individual should buy or at least examine. These systems is being used constantly in the web and have become basic components of electronic commerce and information retrieval, making suggestions that filter large information spaces so as to every user could be shown items that suits well his needs and interests. The huge growth of available information and items in the internet as well as the constantly increasing number of visitors of web sites in the recent years poses some challenges for recommender systems. These are the production of accurate recommendations, the calculation of large number of recommendations in a very small time period and the successful completion of a very big proportion of the requests that the system accepts. Various techniques like collaborative filtering, content based and demographic as well as some combinations of them have been used to create recommendations. This postgraduate thesis surveys a number of collaborative filtering algorithms analyzing the steps for creating such a system and compares which methods provide better results in the challenges that a recommender system encounters.

ΕΙΣΑΓΩΓΗ

Στην σημερινή κοινωνία, στην οποία έχει ζωτική σημασία η πληροφορία και η αναζήτηση της, ερχόμαστε καθημερινά αντιμέτωποι με το πρόβλημα της υπερφόρτωσης πληροφοριών. Ελάχιστος χρόνος και ταυτόχρονα τεράστιοι όγκοι δυναμικών και μη δεδομένων συνθέτουν αυτήν την κατάσταση, που μόνο σύγχυση μπορεί να προκαλέσει ακόμα και στον έμπειρο χρήστη. Αν αναλογιστούμε ότι ο μέσος χρήστης διαθέτει πολύ λίγο χρόνο για να πραγματοποιήσει κάποια αναζήτηση ,τότε κρίνοντας και από το πλήθος των δεδομένων που θα αντιμετωπίσει, θα πρέπει να βρίσκεται συνεχώς σε σύγχυση. Σε αυτήν την κατάσταση, εξειδικευμένο λογισμικό, θα μπορούσε να βοηθήσει τον χρήστη σε μεγάλο βαθμό να αποφύγει αυτό το πρόβλημα της υπερφόρτωσης. Επομένως η χρήση της μοντελοποίησης χρηστών, με στόχο την παροχή εξατομικευμένων πληροφοριών μέσω συστημάτων σύστασης, δείχνει να είναι μονόδρομος.

Ως σύστημα σύστασης, περιγράφεται κάθε σύστημα το οποίο είναι ικανό να παρέχει εξατομικευμένα αποτελέσματα σύστασης, ώστε να βοηθήσει τον χρήστη να γνωρίσει ενδιαφέροντα αντικείμενα ,μέσα σε ένα τεράστιο χώρο από πιθανές επιλογές. Τα συστήματα αυτά βασίζονται στην προσαρμογή στα χαρακτηριστικά και τη συμπεριφορά του χρήστη, μαθαίνοντας γι' αυτόν, είτε κάνοντας του σχετικές ερωτήσεις, είτε παρακολουθώντας την συμπεριφορά του σε κάθε αντικείμενο που έρχεται αντιμέτωπος κατά την αλληλεπίδραση του με αυτά. Σκοπός του συστήματος είναι να δημιουργήσει ένα προφίλ χρήστη ,το οποίο να το εκμεταλλευτεί κατάλληλα, ώστε να του δίνει κάθε φορά επιλογές που ταιριάζουν με αυτό.

Όταν κάνουμε μια επιλογή, και δεν έχουμε άμεση γνώση του αντικειμένου ,είναι καλή πρακτική, να επιλέξουμε αυτό που προτείνουν άλλοι χρήστες, που έχουν ίδιο, η παρόμοιο τρόπο σκέψης με μας ή να υπάρχουν κοινά στοιχεία που έχουν χαρακτηριστεί και από τις δυο πλευρές ως ενδιαφέροντα. Τα συστήματα που υλοποιούν τέτοιες μεθόδους, μπορούν να βρουν εφαρμογή σε πολλά διαφορετικά είδη πληροφοριών και δεδομένων όπως, μουσικά κομμάτια , ταινίες, βιβλία και πολλά άλλα.

Η δημιουργία προτάσεων είναι αποτέλεσμα τριών συστατικών στοιχείων : ενός προφίλ χρήστη , ενός μηχανισμού που θα ανανεώνει το προφίλ χρήστη αναλόγως με τα πιο πρόσφατα δεδομένα χρήσης και ενός ακόμα μηχανισμού, ο οποίος θα λαμβάνει υπόψη του το προφίλ και θα παράγει συστάσεις. Τα δεδομένα τα οποία αποθηκεύονται στο προφίλ ,ποικίλουν ανάλογα με τον τρόπο υλοποίησης του μηχανισμού. Διαφορετικές προσεγγίσεις απαιτούν διαφορετικά δεδομένα για τον χρήστη, κάνοντας τη δομή του προφίλ να διαφέρει από σύστημα σε σύστημα.

Ένα από τα μεγαλύτερα σύνολα δεδομένων που μπορούν να χρησιμοποιηθούν τα συστήματα σύστασης είναι ο χώρος του διαδικτύου όπου στην σύγχρονη κοινωνία της πληροφορίας, ο όγκος των επιλογών που μπορεί κάποιος να κάνει και των πληροφοριών που μπορεί κανείς να αντλήσει είναι χαώδης. Η μεγάλη ανάγκη για ελάττωση του χρόνου ενασχόλησης με την αναζήτηση έχει δημιουργήσει εδώ και αρκετά χρόνια, τις μηχανές αναζήτησης όπου προτείνουν στο χρήστη ιστοσελίδες με πληροφορίες, που είναι πιθανότατα κοντά σε αυτές που ψάχνει ο χρήστης. Εκτός όμως από τον πολύ μεγάλο όγκο των ιστοσελίδων υπάρχει επίσης πολύ μεγάλη ποικιλία πληροφοριών για κομμάτια μουσικής, κινηματογραφικές ταινίες, ειδήσεις και πολλά άλλα αντικείμενα που θα μπορούσε ενδεχομένως να ψάξει ο χρήστης. Σε αυτές τις περιπτώσεις μπορεί να βρει εφαρμογή ένα σύστημα σύστασης που θα προτείνει στον χρήστη ταινίες για παράδειγμα, σύμφωνα με τις κινηματογραφικές του προτιμήσεις, περιορίζοντας έτσι τον χρόνο και την πληθώρα των άχρηστων που θα χαρακτήριζαν αντιπαραγωγική και χρονοβόρα την αναζήτησή του. Ουσιαστικά τα συστήματα σύστασης λειτουργούν πρακτικά σαν φίλτρα, και παρουσιάζουν συστάσεις που πιστεύουν ότι είναι κοντά σε αυτό που θέλει ο χρήστης.

Σκοπός ενός συστήματός σύστασης είναι να μπορεί να αποφασίσει και να δώσει κάποια συμβουλή στον χρήστη για κάποιο αντικείμενο που ο ίδιος δεν έχει γνωρίσει ή δεν έχει αξιολογήσει προηγουμένως. Άρα το πρόβλημα της σύστασης αντικείμενων συνεπάγεται με την πραγματοποίηση εκτιμήσεων για όσα αντικείμενα δεν έχει αλληλεπιδράσει ακόμα κάποιος χρήστης, χρησιμοποιώντας πληροφορίες που διαφέρουν ανά περίπτωση. Κάποιες από αυτές είναι το ιστορικό των προτιμήσεων του χρήστη πάνω σε άλλα αντικείμενα και η ομοιότητα των αντικειμένων με αυτά που τον ενδιέφεραν στο παρελθόν. Οι πληροφορίες που χρειάζονται κάθε

φορά, εξαρτώνται από την τεχνική που θα χρησιμοποιηθεί για την υλοποίηση του κάθε συστήματος.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1
ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΗΣ

Το πρόβλημα της σύστασης

Το πρόβλημα της σύστασης μπορεί να αναλυθεί ως εξής: Έστω ότι το *U* είναι το σύνολο όλων των χρηστών και το *I* να είναι το σύνολο όλων των πιθανών αντικειμένων που μπορούν να προταθούν, όπως μουσικά κομμάτια, ταινίες, βιβλία, άρθρα, ηλεκτρονικά παιχνίδια και άλλα. Το σύνολο *I* των πιθανών στοιχείων μπορεί να είναι αχανές και να περιέχει εκατοντάδες χιλιάδες ή ακόμα και εκατομμύρια αντικείμενα σε μερικές εφαρμογές, όπως η σύσταση βιβλίων ή και μουσικών CD. Ομοίως, το σύνολο των χρηστών μπορεί επίσης να είναι πολύ μεγάλο και να προσεγγίζει τιμές εκατομμυρίων σε μερικές περιπτώσεις. Αν η *R* είναι μια συνάρτηση χρησιμότητας που μετρά το ποσό της χρησιμότητας του στοιχείου *i* για τον χρήστη u, θα ισχύει:

$$R: U \times I \rightarrow r$$

Εξίσωση 1

Όπου *r* είναι ένα πλήρως διατεταγμένο σύνολο, για παράδειγμα ένα σύνολο μη αρνητικών ακεραίων. Κατόπιν, για κάθε χρήστη $\mathbf{u} \in \mathbf{U}$, θέλουμε να επιλέξουμε τέτοιο στοιχείο $i' \in \mathbf{I}$ το οποίο να βρίσκεται στο σύνολο των αντικειμένων που δεν έχει γνωρίσει ο χρήστης και ταυτόχρονα να μεγιστοποιεί τη συνάρτηση χρησιμότητας του χρήστη **R**.

$$\forall u \in U, \qquad i'_u = \arg\max_{i \in I} R(u, i)$$

Εξίσωση 2

Στα συστήματα σύστασης, η ποσότητα της χρησιμότητας ενός στοιχείου συνήθως αντικατοπτρίζεται σε μια εκτίμηση ,που δείχνει με ποιο σημείο της κλίμακας βαθμολογίας, ένας χρήστης αξιολόγησε ένα στοιχείο. Για παράδειγμα ο χρήστης Κώστας αξιολόγησε την ταινία "Τιτανικός" με 4 (έχοντας σαν ανώτερο όριο το 5). Ωστόσο η συνάρτηση χρησιμότητας ,δεν είναι πάντα συγκεκριμένη. Ανάλογα με την εφαρμογή, η συνάρτηση R μπορεί είτε να καθορίζεται από τον χρήστη είτε να υπολογίζεται από την εφαρμογή.

Κάθε στοιχείο u του συνόλου U των χρηστών μπορεί να περιγραφτεί με ορισμένες πληροφορίες που περιέχονται στο προφίλ του, όπως η ηλικία, το φύλο, το εισόδημα, η οικογενειακή κατάσταση και άλλα. Στο πιο απλουστευμένο σενάριο, το προφίλ μπορεί να περιέχει ένα μόνο χαρακτηριστικό που να τον ξεχωρίζει από τους άλλους, όπως ο κωδικός-χρήστη (User-ID). Ομοίως, κάθε αντικείμενο i του συνόλου I των στοιχείων είναι καθορισμένο από έναν αριθμό χαρακτηριστικών. Ένα απλό παράδειγμα είναι το εξής: για ένα σύνολο I, στοιχείων που θα είναι κινηματογραφικές ταινίες, η κάθε μια από αυτές εκτός από τον κωδικό της μπορεί να έχει τον τίτλο της, τον σκηνοθέτη, τους πρωταγωνιστές, το είδος της και άλλα χαρακτηριστικά που θα μπορούσαν να βοηθήσουν σε μια πιθανή κατηγοριοποίηση της και επιλογή της.

Το κεντρικό πρόβλημα των συστημάτων σύστασης εντοπίζεται στη συνάρτηση χρησιμότητας R , αφού συνήθως δεν καθορίζεται σε ολόκληρο το διάστημα U x I, αλλά μόνο σε κάποιο υποσύνολό του. Αυτό συνεπάγεται ότι το R πρέπει να επεκταθεί σε ολόκληρο το διάστημα. Στα συστήματα σύστασης το ποσό της χρησιμότητας αναπαριστάται από τη βαθμολογία του στοιχείου και καθορίζονται αρχικά μόνο τα στοιχεία που έχουν αξιολογηθεί από τους χρήστες. Πρακτικά δηλαδή σε ένα σύστημα σύστασης βιβλίων, ο χρήστης καλείται να

αξιολογήσει (έστω στην κλίμακα του 10) κάποια από τα βιβλία που έχει διαβάσει. Αυτό σημαίνει πως αρκετά από τα στοιχεία (βιβλία) δεν θα έχουν αξιολόγηση από τον συγκεκριμένο χρήστη. Ωστόσο, ο στόχος του συστήματος σύστασης είναι να προβλέψει τις πιθανές αξιολογήσεις του χρήστη για τα υπόλοιπα στοιχεία και να παρουσιάσει συστάσεις κατάλληλες για τις προτιμήσεις του χρήστη, βασισμένες στις προβλέψεις αυτές .

Οι εκτιμήσεις των αγνώστων βαθμολογιών από τις ήδη υπάρχουσες, πραγματοποιούνται συνήθως από:

i. Μεθόδους εύρεσης που καθορίζουν τη συνάρτηση χρησιμότητας και εμπειρικά αξιολογούν την απόδοσή της
ii. Υπολογίζοντας την συνάρτηση χρησιμότητας που βελτιστοποιείται για ένα συγκεκριμένο κριτήριο απόδοσης.

Όταν υπολογιστούν οι άγνωστες βαθμολογίες, υποβάλλονται οι συστάσεις στοιχείων σε έναν χρήστη, οι οποίες είναι συνήθως τα στοιχεία με την μεγαλύτερη βαθμολογία και σε σχέση με τις προτιμήσεις του χρήστη.

Οι τρεις επικρατέστεροι τρόποι προσέγγισης που χρησιμοποιούνται στα σημερινά συστήματα σύστασης είναι [5]:

1. Βασισμένος στη συνεργατική διήθηση (Collaborative Filtering Recommender Systems)
2. Βασισμένος στο περιεχόμενο (Content Based Recommender Systems)
3. Υβριδικά συστήματα (hybrid Recommender Systems)

Ενώ έχουν υλοποιηθεί και κάποιοι άλλοι μηχανισμοί [15]:

4. Δημογραφικά συστήματα (Demographic Recommender Systems)
5. Βασισμένος σε γνώση (Knowledge based Recommender Systems)
6. Βασισμένος στην ωφέλεια (Utility based Recommender systems)

Σύστημα Σύστασης βασισμένο στη συνεργατική διήθηση (collaborative filtering recommender systems)

Στα συστήματα συνεργατικής διήθησης, οι συστάσεις βασίζονται σε αξιολογήσεις χρηστών, οι οποίοι έχουν κοινά ενδιαφέροντα. Το όλο εγχείρημα είναι θεμελιωμένο πάνω στην ιδέα ότι το σύνολο των χρηστών που έκαναν της ίδιες επιλογές, πιθανότατα να έχουν και τις ίδιες γενικότερες προτιμήσεις. Έτσι θα μπορούσαμε σε κάποιον χρήστη, μέσα από το σύνολο αυτό, να του προτείνουμε αντικείμενα που οι υπόλοιποι τα έχουν αξιολογήσει ως ικανοποιητικά στο παρελθόν. Οι γνώμες μπορούν να εκφραστούν σε βαθμολογίες από τους χρήστες με κλίμακες από κακό έως καλό ή επίσης και με καταγραφή από το σύστημα διαφόρων ενεργειών ή αντιδράσεων του χρήστη. Για παράδειγμα, αν ο χρήστης θα διάβαζε κάποιο βιβλίο ή θα έβλεπε κάποια κινηματογραφική ταινία θα είχε θετική αξιολόγηση, αλλιώς αν την αγνοούσε, αρνητική, όπως επίσης και αν πέρναγε αρκετή ώρα διαβάζοντας κάποιο άρθρο ή απλά του έριχνε μια ματιά και προχωρούσε σε επόμενο.

Αναλύοντας τις προτιμήσεις των χρηστών, μπορούμε να εξάγουμε κάποιο συμπέρασμα για τον βαθμό ομοιότητας των προτιμήσεών τους, χωρίς να δίνεται βαρύτητα στο είδος της πληροφορίας. Σαν αποτέλεσμα το μεγαλύτερο πλεονέκτημα αυτής της μεθόδου είναι η δυνατότητα να παράγει συστάσεις για πολλές κατηγορίες αντικειμένων όπως μουσική , ταινίες βιβλία καθώς και πολλά άλλα παραδείγματα. Θα ακολουθήσει εκτενής αναφορά σε επόμενο κεφάλαιο για αυτήν την κατηγορία συστημάτων σύστασης.

Σύστημα Σύστασης βασισμένο στο περιεχόμενο (Content based recommender systems)

Σε αντίθεση με τα συστήματα συνεργατικής διήθησης, όπου οι βαθμολογίες του χρήστη αποτελούν τον κυριότερο παράγοντα στην συσχέτιση τους με αυτές άλλων χρηστών για τη δημιουργία συστάσεων, στα συστήματα τα βασισμένα στο περιεχόμενο, δεν μας αφορούν οι γνώμες, άρα και το ιστορικό των άλλων ατόμων. Το κύριο συστατικό αυτής της προσέγγισης είναι η επεξεργασία των διαφόρων χαρακτηριστικών των αντικειμένων που είναι έτοιμα να προταθούν. Ένα μεγάλο πλήθος αντικειμένων μπορεί να αντιπροσωπευτεί από κάποιες τιμές

σε κάποια χαρακτηριστικά γνωρίσματα. Κάθε κατηγορία αντικειμένων μπορεί να περιγραφεί από τα ίδια χαρακτηριστικά και κάθε χαρακτηριστικό μπορεί να έχει ένα πεπερασμένο σύνολο τιμών.

Το σύστημα μαθαίνει τις προτιμήσεις του χρήστη, με βάση τα δεδομένα που έχει μαζέψει από τα αντικείμενα που έχει βαθμολογήσει στο παρελθόν. Συνδυάζοντας τις τιμές αυτών των διανυσμάτων χαρακτηριστικών των αντικειμένων με τα υπόλοιπα διανύσματα χαρακτηριστικών που διαθέτει το σύστημα, προσπαθεί να ανιχνεύσει τα ομοιότερα, σε σχέση με αυτά του ενεργού χρήστη. Σε αρκετές περιπτώσεις χρησιμοποιείται η συσχέτιση ημιτόνου σαν μέτρο ομοιότητας.

Στη μέθοδο σύστασης με βάση το περιεχόμενο , η συνάρτηση χρησιμότητας R(u,i) του στοιχείου i για το χρήστη u υπολογίζεται με βάση την αξιολόγηση που προσδίδει ο χρήστης στο στοιχείο $i^* \in I$ που είναι παρόμοιο με το i. Δηλαδή στην περίπτωση της εφαρμογής κινηματογραφικών ταινιών για παράδειγμα, το σύστημα προσπαθεί να βρει τη συσχέτιση μεταξύ των ταινιών που αξιολόγησε ο χρήστης με υψηλές βαθμολογίες, χρησιμοποιώντας τα χαρακτηριστικά (σκηνοθέτης, είδος, πρωταγωνιστές και άλλα). Για να προτείνει στη συνέχεια αυτές τις ταινίες που έχουν μεγάλο βαθμό ομοιότητας με τις προτιμήσεις του συγκεκριμένου χρήστη. Η συνάρτηση χρησιμότητας συνήθως ορίζεται ως εξής:

$$\mathbf{R(u,i) = score(ContentBasedProfile(u), Contents(i))}$$

Η συγκεκριμένη μέθοδος έχει άμεση σχέση με την ανάκτηση πληροφοριών, αλλά και την έρευνα φιλτραρίσματος πληροφοριών, αφού βασίζεται στο προφίλ χρήστη που περιέχει πληροφορίες οι οποίες προέρχονται από το χρήστη και περιγράφει κατά κάποιο τρόπο τις προτιμήσεις και τις ανάγκες του. Οι πληροφορίες αυτές αποκομίζονται συνήθως από τις προηγούμενες επιλογές και αξιολογήσεις του χρήστη.

Έστω λοιπόν ότι το διάνυσμα Content(i) είναι το σύνολο των χαρακτηριστικών του αντικείμενου i, δηλαδή, ένα σύνολο ιδιοτήτων που χαρακτηρίζει το στοιχείο i, το οποίο συνήθως είναι αποτέλεσμα της εξαγωγής ενός συνόλου χαρακτηριστικών γνωρισμάτων από το στοιχείο i. Αυτό χρησιμοποιείται για να καθορίσει την καταλληλότητα του στοιχείου για σύσταση. Η μέθοδος αυτή χρησιμοποιείται αρκετά συχνά σε κομμάτια κειμένου, τα οποία συνήθως αναλύονται εύκολα, λαμβάνοντας υπόψη σαν χαρακτηριστικά τους, λέξεις, των οποίων η σημαντικότητα μπορεί να μετρηθεί χρησιμοποιώντας κάποια μέτρα βαρύτητας.

Αυτή είναι η πρώτη μαθηματική προσέγγιση, που έχει της ρίζες της στην ανάκτηση πληροφοριών. Υλοποιείται κυρίως με το μέτρο ομοιότητας συνημίτονου (cosine similarity measurre). Στη συγκεκριμένη περίπτωση τα στοιχεία μπορεί να υπολογίζονται σαν διανύσματα των λέξεων-κλειδιών (keywords) όπου κάθε χαρακτηριστικό, υποδηλώνει την σημαντικότητα της λέξης-κλειδί. Στη συνέχεια χρησιμοποιούνται από το σύστημα κάποιες ευρετικές μέθοδοι (αλγόριθμος Winnon) για να υπολογιστεί η πιθανότητα, ένα στοιχείο να είναι στις προτιμήσεις του χρήστη. Ακολουθεί ένα παράδειγμα μιας τέτοιας συνημιτονοειδούς συνάρτησης (όπου w η σημαντικότητα) :

$$R(u,i) = \cos(\overrightarrow{w_u}, \overrightarrow{w_t}) = \frac{\overrightarrow{w_u}.\overrightarrow{w_t}}{\left\|\overrightarrow{w_u}\right\| \times \left\|\overrightarrow{w_t}\right\|}$$

Εξίσωση 3

Εκτός όμως από την προσέγγιση με τη χρήση του μέτρου ομοιότητας συνημίτονου και των ευρετικών μεθόδων, υπάρχει επίσης μια άλλη μαθηματική προσέγγιση, η οποία χρησιμοποιεί τεχνικές στατιστικής και μηχανικής μάθησης. Το βασικό χαρακτηριστικό της μεθόδου αυτής είναι ο υπολογισμός της πιθανότητας να προταθεί ένα στοιχείο , έχοντας ως δεδομένα τα χαρακτηριστικά του στοιχείου και ένα δείγμα επιλογών του χρήστη. Το σύστημα προσπαθεί να κατατάξει το αντικείμενο σε μια κλάση. Για παράδειγμα, αν θα το επέλεγε ο χρήστης ή όχι. Μια χαρακτηριστική τέτοια περίπτωση είναι η χρήση του αλγόριθμου ταξινόμησης του Bayes. Μια τέτοια συνάρτηση πρόβλεψης πιθανότητας είναι η ακόλουθη:

$$P(C_i)\prod_{x} P(k_{x,j}|C_i)$$

Εξίσωση 4

Μειονεκτήματα συστημάτων σύστασης βασισμένα στο περιεχόμενο

Τα συστήματα σύστασης με βάση το περιεχόμενο όμως έχουν κάποιους περιορισμούς που μειώνουν την αποδοτικότητα τους και περιγράφονται παρακάτω [2].

- Περιορισμένη Ανάλυση Περιεχομένου

Τα συστήματα που χρησιμοποιούν τεχνικές με βάση το περιεχόμενο, περιορίζονται από τα χαρακτηριστικά που είναι άμεσα συσχετισμένα με τα αντικείμενα που θα συστήσουν. Ως εκ τούτου, προκειμένου να έχουν ένα επαρκές σύνολο χαρακτηριστικών, η πληροφορία των αντικειμένων πρέπει να είναι είτε σε μορφή που να μπορεί να αναλύεται αυτόματα από ηλεκτρονικό υπολογιστή για παράδειγμα κείμενο ή τα χαρακτηριστικά θα πρέπει να αντιστοιχισθούν με τα στοιχεία, χειροκίνητα.

Ενώ τεχνικές ανάκτησης πληροφοριών λειτουργούν αρκετά καλά στην εξόρυξη χαρακτηριστικών από έγγραφα κειμένου, σε ορισμένα άλλα είδη αντικειμένων παρουσιάζονται προβλήματα με την αυτόματη εξαγωγή χαρακτηριστικών. Για παράδειγμα, οι αυτόματες μέθοδοι εξαγωγής χαρακτηριστικών είναι πολύ δυσκολότερο να εφαρμοσθούν σε αρχεία πολυμέσων, όπως φωτογραφίες, αρχεία ήχου και video. Επιπλέον, αρκετά συχνά η χειροκίνητη εκχώρηση ιδιοτήτων, χαρακτηρίζεται ως μη αποδοτική χρονοβόρα πρακτική, λόγω των περιορισμένων πόρων.

Ακόμα ένα πρόβλημα που παρουσιάζει η περιορισμένη ανάλυση περιεχομένου ,είναι ότι, εάν δύο διαφορετικά στοιχεία εκπροσωπούνται από τις ίδιες τιμές στα διαφορά χαρακτηριστικά, για το σύστημα δεν θα έχουν καμία διαφορά μεταξύ τους. Ως εκ τούτου, δεδομένου ότι τα έγγραφα κειμένου συνήθως εκπροσωπούνται από τις σημαντικότερες λέξεις κλειδιά, τα συστήματα που είναι βασισμένα στο περιεχόμενο δεν μπορούν να κάνουν κάποια διάκριση ανάμεσα σε ένα καλογραμμένο άρθρο και ένα κακογραμμένο, αν τυχαίνει να χρησιμοποιούν τις ίδιους όρους.

- Πρόβλημα Υπερειδίκευσης (Overspecialization Problem)

Όταν το σύστημα μπορεί να κάνει προτάσεις μόνο με αντικείμενα που ταιριάζουν με το προφίλ ενός χρήστη, ο χρήστης είναι καταδικασμένος στη σύσταση μόνο στοιχείων, που είναι παρόμοια με αυτά που έχει ήδη βαθμολογήσει. Για παράδειγμα, ένα άτομο που δεν έχει γνωρίσει ποτέ την Ελληνική κουζίνα ,ποτέ δεν θα λάβει μια σύσταση ακόμη και για το καλύτερο ελληνικό εστιατόριο.

Επιπλέον, το πρόβλημα της υπερειδίκευσης,, δεν περιορίζεται μόνο στο ότι τα συστήματα με βάση το περιεχόμενο δεν μπορούν να συστήσουν στοιχεία διαφορετικά, από οτιδήποτε ο χρήστης έχει γνωρίσει μέχρι τώρα. Σε ορισμένες περιπτώσεις μάλιστα κάποια στοιχεία δεν θα πρέπει να συνιστώνται, εάν έχουν πολύ μεγάλο βαθμό ομοιότητας με κάτι που ο χρήστης έχει ήδη γνωρίσει. Για παράδειγμα διαφορετικά άρθρα ειδήσεων που περιγράφουν το ίδιο γεγονός. Ως εκ τούτου, ορισμένα συστήματα συστάσεων βασισμένα στο περιεχόμενο, απορρίπτουν αντικείμενα, όχι μόνο εφόσον είναι πολύ διαφορετικά από αυτά των προτιμήσεων του χρήστη, αλλά ακόμα και εάν έχουν πολύ μεγάλο βαθμό ομοιότητας με κάποια που ο χρήστης έχει ξανασυναντήσει.

- Πρόβλημα νέου χρήστη

Ο χρήστης θα πρέπει να έχει βαθμολογήσει έναν επαρκή αριθμό στοιχείων, πριν κάποιο σύστημα συστάσεων με βάση το περιεχόμενο μπορέσει πραγματικά να κατανοήσει τις προτιμήσεις του χρήστη και να τον συμβουλέψει με αξιόπιστες συστάσεις. Σαν αποτέλεσμα, ίσως να χρειαστεί ένα μεγάλο χρονικό διάστημα και αρκετή υπομονή για έναν καινούριο χρήστη μέχρι να λάβει κάποια ικανοποιητική συμβουλή για κάποιο αντικείμενο.

Υβριδικά συστήματα (Hybrid recommender systems)

Πολλά συστήματα σύστασης χρησιμοποιούν υβριδικές προσεγγίσεις, συνδυάζοντας δυο ή περισσότερες τεχνικές, οι οποίες συμβάλλουν στην αποφυγή ορισμένων περιορισμών, που διέπουν τον καθένα από τους ανεξάρτητους μηχανισμούς. Οι διαφορετικοί τρόποι που μπορούν να συνδυαστούν διάφορες μέθοδοι μεταξύ τους μπορούν να κατηγοριοποιηθούν στους παρακάτω [15]:

- Σταθμισμένος (Weighted)

Ένα υβριδικό σύστημα βασισμένο στη στάθμιση, χρησιμοποιεί όλες τις τεχνικές που είναι διαθέσιμες στο σύστημα. Για παράδειγμα μπορεί να συνδυάσει βαθμολογίες που λαμβάνονται από τα δυο αυτόνομα συστήματα συστάσεων, σε μια τελική σύσταση κάνοντας χρήση μιας γραμμικής φόρμουλας

- Με εναλλαγή (Switching)

Στην περίπτωση αυτή το σύστημα χρησιμοποίει κάποια κριτήρια για να πραγματοποιήσει εναλλαγή από μια τεχνική σε μια άλλη. Για παράδειγμα όταν η τεχνική η βασισμένη στο περιεχόμενο δεν μπορεί να δώσει κάποια σίγουρη πρόβλεψη, τότε εκτελείται μηχανισμός συνεργατικής διήθησης. Αυτού του τύπου οι συνδυαστικοί μηχανισμοί επιφέρουν επιπρόσθετη πολυπλοκότητα στη διαδικασία, μιας και τα κριτήρια της εναλλαγής πρέπει να καθοριστούν, προσθέτοντας ακόμα ένα επίπεδο παραμετροποίησης.

- Μικτός (Mixed)

Στην κατηγορία αυτή, συστάσεις από περισσότερες από μια τεχνικές παρουσιάζονται μαζί. Για παράδειγμα σε ένα σενάριο σύστασης τηλεοπτικών εκπομπών θα ήταν δυνατό να γίνει χρήση των κειμένων περιγραφής, με τεχνική, βασισμένη στο περιεχόμενο και μεθόδων συνεργατικής διήθησης για τις προτιμήσεις των άλλων χρηστών. Οι συστάσεις και από τους δυο μηχανισμούς συνδυάζονται και γίνεται εκτίμηση της τελικής σύστασης.

- Συνδυασμός χαρακτηριστικών (Feature combination)

Άλλος ένας τρόπος συνδυασμού δυο διαφορετικών τεχνικών είναι η χρήση του ενός, ως απλό χαρακτηριστικό, με σκοπό να εμπλουτίσει το σύνολο δεδομένων με περισσότερες πληροφορίες και ύστερα να υλοποιηθεί ο δεύτερος αλγόριθμος με το διευρυμένο πλέον σύνολο.

- Διαδοχικός (Cascade)

Σε αντίθεση με τις προηγούμενες μεθόδους, η διαδοχική, συνεπάγεται μια σταδιακή διαδικασία. Σε αυτήν την τεχνική, μια τεχνική σύστασης, καλείται πρώτη για την παραγωγή μιας γενικής κατάταξης των υποψηφίων και μια δεύτερη τεχνική τελειοποιεί το αποτέλεσμα της αρχικής. Η διαδοχή επιτρέπει στο σύστημα να αποφύγει να χρησιμοποιήσει τη χαμηλότερης προτεραιότητας δεύτερη τεχνική, με αντικείμενα που έχουν ήδη διαφοροποιηθεί από το αρχικό ή έχουν εκτιμηθεί επαρκώς ως άχρηστα και ποτέ δεν θα προταθούν.

Επειδή το δεύτερο στάδιο διαδοχής επικεντρώνεται μόνο σε εκείνα τα σημεία για τα οποία η επιπλέον τελειοποίηση είναι αναγκαία, είναι πιο αποδοτική από ό, τι, για παράδειγμα, ένα βασισμένο στη βαρύτητα υβριδικό σύστημα, που εφαρμόζει όλες τις τεχνικές σε όλα τα στοιχεία.

Επιπλέον, η διαδοχή είναι ανεκτική στα παράσιτα της τεχνικής χαμηλής προτεραιότητας, δεδομένου ότι οι εκτιμήσεις των συστάσεων υψηλής προτεραιότητας μπορούν μόνο να τελειοποιηθούν και όχι να ανατραπούν.

- Διεύρυνση χαρακτηριστικών (Feature augmentation)

Μια τεχνική χρησιμοποιείται για την παραγωγή μιας αξιολόγησης ή ταξινόμησης για ένα αντικείμενο και στη συνέχεια αυτή η πληροφορία ενσωματώνεται στην επεξεργασία της επόμενης τεχνικής σύστασης.

- Μέτα-επιπέδου (Meta-level)

Στην προσέγγιση αυτή, δυο τεχνικές σύστασης μπορούν να συνδυαστούν, χρησιμοποιώντας το μοντέλο που δημιούργησε η μια ως είσοδο για την άλλη . Η διαφορά του με τη διεύρυνση χαρακτηριστικών, έγκειται στο ότι δημιουργείται το μοντέλο για να δώσει κάποια χαρακτηριστικά και όχι για να χρησιμοποιηθεί ως είσοδο στον δεύτερο μηχανισμό.

Το όφελος για αυτή τη μέθοδο, ιδιαίτερα για περιπτώσεις συνδυασμού βασισμένων στο περιεχόμενο τεχνικών με μεθόδους συνεργατικής διήθησης, είναι ότι το εξαγόμενο μοντέλο είναι ουσιαστικά μια συμπιεσμένη αναπαράσταση των προτιμήσεων του χρήστη και ο συνεργατικός μηχανισμός που ακολουθεί, μπορεί να χειριστεί ευκολότερα αυτή την πληροφορία παρά τα ακατέργαστα δεδομένα.

Τέλος, αρκετές συγκρίσεις που έχουν πραγματοποιηθεί μελετώντας τις επιδόσεις του υβριδικού σε σχέση με τα καθαρά συνεργατικά ή με βάση το περιεχόμενο συστήματα, έχουν δείξει ότι οι υβριδικές μέθοδοι μπορούν να παρέχουν συστάσεις με μεγαλύτερη ακρίβεια από τις ανεξάρτητες προσεγγίσεις.

Δημογραφικά συστήματα σύστασης (Demographic recommender systems)

Τα συστήματα αυτά επιχειρούν να κατατάξουν τον χρήστη σε κάποια δημογραφική κλάση, αντλώντας τα απαραίτητα στοιχεία για αυτόν, συνήθως μέσω κάποιων ερωτηματολογίων ή συνεντεύξεων. Αφού κάποιο άτομο κατηγοριοποιηθεί κάπου, τότε παίρνει συστάσεις από αντικείμενα που έχουν χαρακτηρίσει καλά, οι χρήστες της ίδιας κλάσης. Στην περίπτωση αυτή το ιστορικό των προηγουμένων βαθμολογιών του χρήστη δεν είναι απαραίτητο, όπως απαιτείται στις περιπτώσεις συνεργατικής διήθησης και βασισμένης στο περιεχόμενο, μιας και το σύστημα αρκείται μόνο σε δημογραφικά δεδομένα. Για πολλά χρόνια ήταν δύσκολο να υλοποιηθεί ένας τέτοιος μηχανισμός, μιας και οι άνθρωποι ήταν πολύ φειδωλοί στην αποκάλυψη προσωπικών στοιχείων. Πλέον όμως με την έξαρση του διαδικτύου και των ιστοτόπων κοινωνικής δικτύωσης το εμπόδιο αυτό φαίνεται πως ξεπερνιέται.

Βασισμένα σε γνώση συστήματα σύστασης (Knowledge - based recommender systems)

Σε αυτήν την περίπτωση τα συστήματα, συγκρατούν πληροφορίες για τη χρησιμότητα που έχει κάποιο αντικείμενο στον χρήστη, δίνοντας έτσι συστάσεις, βασισμένες στο πως, ένα αντικείμενο μπορεί να καλύψει τις ανάγκες του χρήστη. Ένα παράδειγμα μιας τυπικής πληροφορίας που αποθηκεύεται στο προφίλ χρήστη αυτής της κατηγορίας είναι ένα ερώτημα σε μια μηχανή αναζήτησης.

Βασισμένα στη ωφέλεια συστήματα σύστασης (Utility - based recommender systems)

Τα βασισμένα στην ωφέλεια συστήματα σύστασης πραγματοποιούν εκτιμήσεις, υπολογίζοντας το ποσό της ωφέλειας κάθε αντικειμένου για τον χρήστη. Το πρόβλημα που αντιμετωπίζει αυτή η κατηγορία συστημάτων, έγκειται στο πώς θα δημιουργηθεί η συνάρτηση ωφέλειας για κάθε χρήστη. Συνεπώς στα συστήματα αυτά, το προφίλ χρήστη αποτελείται από την συνάρτηση ωφέλειας στην οποία έχει καταλήξει ο μηχανισμός και υλοποιεί τεχνικές ικανοποίησης περιορισμών για να εντοπίσει την καλύτερη αντιστοιχία. Το πλεονέκτημα αυτού του είδους σύστασης είναι, ότι μπορεί να λάβει υπόψη παράγοντες που είναι ανεξάρτητοι με το αντικείμενο, όπως για παράδειγμα διαθεσιμότητα εμπόρων και αντικειμένων στη συνάρτηση ωφέλειας. Με αυτό τον τρόπο, γίνεται δυνατή η προσαρμογή στην τρέχουσα κατάσταση του χρήστη, όπως για παράδειγμα η επιλογή γρηγορότερης διανομής παρά καλύτερης τιμής για κάποιον που έχει άμεση ανάγκη.

ΚΕΦΑΛΑΙΟ 2
ΣΥΝΕΡΓΑΤΙΚΗ ΔΙΗΘΗΣΗ

Έστω **U** το σύνολο των χρηστών μιας βάσης και **I** το σύνολο των αντικειμένων που περιέχει τότε το σύστημα προσπαθεί να υπολογίσει τη συνάρτηση βαθμολογίας (εξίσωση 1) για κάθε ζεύγος τιμών (user, item) τα οποία δεν έχουν αξιολογηθεί ακόμα από τους χρήστες. Στην προσέγγιση της συνεργατικής διήθησης, ο μηχανισμός υπολογίζει την βαθμολογία ενός αντικειμένου, λαβαίνοντας υπόψη το πόσο πολύ άρεσε στους άλλους χρήστες το συγκεκριμένο στοιχείο. Δηλαδή γίνεται εκτίμηση της βαθμολογίας **R(u, i)** του χρήστη **u** για το αντικείμενο **i** βασισμένη στις βαθμολογίες **R(u', i)** άλλων χρηστών για το αντικείμενο **i** οι οποίοι χρήστες $u' \in U$ είναι παρόμοιοι με τον χρήστη **u**. Οι αλγόριθμοι συνεργατικής διήθησης μπορούν να διαχωριστούν σε δυο κύριες κλάσεις:

- Μνήμης (memory based)
- Μοντέλου (model based)

Συστήματα σύστασης Μνήμης (memory based)

Οι αλγόριθμοι μνήμης χρησιμοποιούν ευρετικές μεθόδους που εκμεταλεύονται ολόκληρη τη συλλογή των βαθμολογιών των χρηστών προκειμένου να φτάσουν σε κάποιο αποτέλεσμα. Η εκτιμώμενη τιμή της βαθμολογίας **R(u, i)** του χρήστη **u** για το αντικείμενο **i** εξάγεται από το συμψηφισμό των αξιολογήσεων για το **i** των υπολοίπων χρηστών η κάποιου υποσυνόλου τους $\hat{U}$ που για παράδειγμα περιέχει τους **N** περισσότερο παρόμοιους με τον **u** .

$$R(u,i) = aggr R(u',i) \text{ με } u' \in U$$

Εξίσωση 5

Κάποια παραδείγματα συναρτήσεων που υλοποιούν τέτοιες συναθροίσεις είναι:

$$R(u,i) = \frac{1}{N} \sum_{u' \in \hat{U}} R(u',i)$$

Εξίσωση 6

Εξίσωση 7

Εξίσωση 8

όπου k παράγοντας κανονικοποίησης και συνήθως ισούται με:

$$k = 1 / \sum_{u' \in U} |similarity(u, u')|$$

Το μέτρο ομοιότητας ανάμεσα στους u u': similarity (u ,u') ουσιαστικά εκφράζει την απόσταση των διανυσμάτων των αξιολογήσεων των δυο χρηστών, και μπορεί να χρησιμοποιηθεί ως συντελεστής βαρύτητας για το ποσοστό που θα λάβει κάθε βαθμολογία του χρήστη u' $R(u', i)$ στην τελική πρόβλεψη, όσο κοντινότεροι είναι δυο χρήστες τόσο μεγαλύτερη θα είναι και η βαρύτητα. Για τον υπολογισμό αυτού του μέτρου λαμβάνονται υπόψη οι αξιολογήσεις των αντικείμενων, τα οποία έχουν και οι δυο βαθμολογήσει. Υλοποιούνται συνήθως με δύο μεθόδους: τους συσχετισμούς (correlations) και την ομοιότητα συνημίτονου των u και u' (cosine based).

Στον τομέα των συσχετίσεων συχνότερα χρησιμοποιούνται οι συσχετίσεις Pearson (**Pearson correlation coefficient**) ή οι συσχετίσεις Spearman (**Spearman's rank correlation coefficient ή Spearman's rho**). Στην πρώτη περίπτωση υπολογίζεται η γραμμική εξάρτηση μεταξύ των διανυσμάτων των αξιολογήσεων των χρηστών u και u'. Στην δεύτερη προσέγγιση, αυτή του Spearman, έχουμε πάλι το ίδιο σύνολο τιμών όμως σε αυτήν την περίπτωση αλλάζει η εξάρτηση μεταξύ των μεταβλητών.

Στην προσέγγιση με βάση το συνημίτονο, ξανά οι χρήστες u και u' μεταφράζονται σε διανύσματα σε n-διάστατο χώρο, όπου **n** ο αριθμός στοιχείων, που έχουν αξιολογήσει και οι δυο. Το μέτρο ομοιότητας των χρηστών προκύπτει από το συνημίτονο της γωνίας που παρουσιάζουν τα δυο διανύσματα. Έστω x, y τα διανύσματα των χρηστών τότε η ζητούμενη τιμή αποδίδεται από τη σχέση $\cos(\vec{x}, \vec{y})$ (εξίσωση 3).

Οι αλγόριθμοι που βασίζονται στην ομοιότητα των χρηστών για να κάνουν προτάσεις, ανήκουν στην κατηγορία χρηστών (user based). Επιπλέον, καθώς οι παραπάνω τεχνικές κυρίως έχουν χρησιμοποιηθεί για τον υπολογισμό ομοιοτήτων μεταξύ των χρηστών, έχουν γίνει προσπάθειες, χρησιμοποιώντας τις ίδιες συσχετίσεις και τεχνικές ομοιότητας συνημίτονου, να υπολογιστούν οι ομοιότητες μεταξύ αντικειμένων αντί χρηστών και να ληφθούν οι εκτιμήσεις με βάση αυτές τις συσχετίσεις [7]. Αυτές οι προσεγγίσεις που βασίζονται στην ομοιότητα των αντικειμένων είναι οι λεγόμενες αντικειμένων (Item based)..

Model based συστήματα σύστασης

Σε αντίθεση με τα memory based συστήματα οι model based αλγόριθμοι χρησιμοποιούν την συλλογή των αξιολογήσεων για να δημιουργήσουν ένα μοντέλο, το οποίο στη συνέχεια θα εκμεταλλευτούν για να κάνουν προβλέψεις βαθμολογίας. Ως εκ τούτου, σε σύγκριση με τις model based μεθόδους, οι memory based μπορούν να θεωρηθούν ως "τεμπέληδες στη μάθηση", υπό την έννοια ότι δεν δημιουργούν ένα μοντέλο και εκτελούν τους ευρετικούς υπολογισμούς κατά το χρόνο που λαμβάνουν τα αιτήματα για να κάνουν συστάσεις.

Από πιθανοτική άποψη, η διαδικασία μπορεί να χαρακτηριστεί ως υπολογισμός της τιμής μιας βαθμολογίας ,έχοντας σαν δεδομένα, το ιστορικό των αξιολογήσεων του χρήστη. Για τον ενεργό χρήστη αν γίνει η υπόθεση ότι οι τιμές τις αξιολόγησης ποικίλλουν από 0 ως m τότε:

$$R_{(u,i)} = \sum_{x=0}^{m} \Pr\left(R_{(u,i)} = x \middle| R_{(u,i')}, i' \in S_u \right) \times x$$

Εξίσωση 9

όπου η πιθανοτική σχέση είναι η πιθανότητα ότι ο χρήστης u θα δώσει στο αντικείμενο i κάποια συγκεκριμένη βαθμολογία, δεδομένου του ιστορικού των αξιολογήσεων του S_u (εξίσωση 10). Για

τον υπολογισμό αυτής της πιθανότητας προτείνονται δυο πιθανοτικά μοντέλα : Τα πρότυπα ομαδοποίησης (cluster models) και τα Μπεϋζιανά δίκτυα [8].

Πρότυπα ομαδοποίησης (cluster models)

Ένα πιθανοτικό μοντέλο που θα μπορούσε να χρησιμοποιηθεί στα συστήματα συνεργατικής διήθησης είναι ο ταξινομητής Bayesian, όπου η πιθανότητα των βαθμολογιών είναι ανεξάρτητη δοθείσας της κλάσης C. Η κεντρική ιδέα είναι, ότι υπάρχουν συγκεκριμένες ομάδες ή τύποι χρηστών που έχουν ένα κοινό σύνολο προτιμήσεων. Δοθείσης της κλάσης οι επιλογές που εκφράζουν τα αντικείμενα, δηλαδή οι βαθμολογίες, είναι ανεξάρτητες. Το μοντέλο πιθανοτήτων που αφορούν την συνδυασμένη πιθανότητα της τάξης και των βαθμολογιών σε ένα σύνολο υποθετικών και οριακών κατανομών είναι η διατύπωση του Bayes.

$$\Pr(C = c, u_1, u_2, \dots, u_n) = \Pr(C = c) \prod_{i=1}^{n} \Pr(u_i | C = c)$$

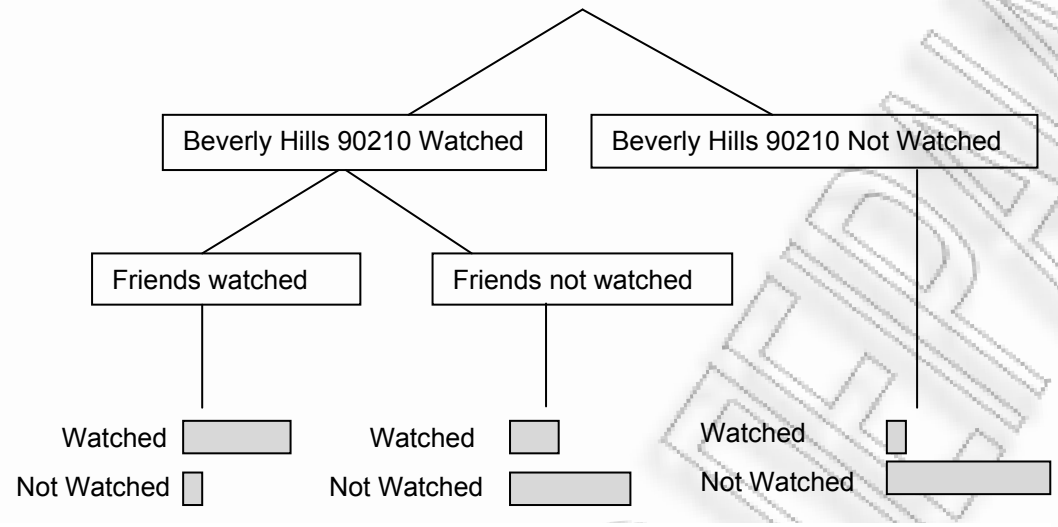
Εξίσωση 10

Η αριστερή μεριά της σχέσης είναι η πιθανότητα να παρατηρηθεί ένας χρήστης συγκεκριμένης κλάσης και ένα ολοκληρωμένο σύνολο από τιμές αξιολογήσεων. Οι παράμετροι του μοντέλου, η πιθανότητα της κλάσης να είναι μέλος $\Pr(C = c)$ και οι υποθετικές πιθανότητες των βαθμολογιών δοθείσας της κλάσης $\Pr(u_1 | C = c)$ εκτιμώνται από ένα εκπαιδευτικό υποσύνολο του πληθυσμού των χρηστών.

Μπεϋζιανά Δίκτυα

Ένα εναλλακτικό πιθανοτικό μοντέλο για συστήματα συνεργατικής διήθησης,είναι τα Μπεϋζιανά δίκτυα ,με κάθε κόμβο να αντιστοιχεί σε ένα αντικείμενο στη βάση. Οι καταστάσεις κάθε κόμβου αντιπροσωπεύουν τις πιθανές τιμές της κλίμακας βαθμολογίας, όπως επίσης υπάρχει και μία κατάσταση η οποία χαρακτηρίζει το αντικείμενο όταν δεν έχει ακόμα αξιολογηθεί. Ο αλγόριθμος εκμάθησης πραγματοποιεί αναζήτηση πάνω σε διάφορες δομές του μοντέλου για εξαρτήσεις μεταξύ αντικειμένων.

Στο τελικό δίκτυο, κάθε κόμβος – αντικείμενο θα έχει ένα σύνολο από μητρικούς κόμβους όπου θεωρούνται και οι καλύτεροι πάροχοι δεδομένων για την πρόβλεψη. Κάθε πίνακας υποθετικής πιθανότητας απεικονίζεται από ένα δένδρο απόφασης όπως φαίνεται και στο παρακάτω σχήμα. Στο συγκεκριμένο παράδειγμα φαίνεται ένα δένδρο απόφασης για την τηλεοπτική σειρά Melrose Place με μητρικούς κόμβους τα Friends και Beverly Hills 90210 και αναλόγως των καταστάσεων των γονέων υπολογίζεται η πιθανότητα να δει κάποιος το Melrose Place ή όχι.



Εικόνα 1 Δέντρο απόφασης για την τηλεοπτική σειρά Melrose place

Πλεονεκτήματα Μειονεκτήματα συστημάτων συνεργατικής διήθησης

Τα συστήματα σύστασης βασισμένα στη συνεργατική διήθηση, εφόσον αδιαφορούν για το περιεχόμενο του κάθε αντικειμένου, δεν διαθέτουν και τις ελλείψεις των συστημάτων που το λαμβάνουν υπόψη, δηλαδή την περιορισμένη ανάλυση περιεχομένου και την υπερειδίκευση. Για τον ίδιο λόγο δεν χρειάζεται να γίνει εξαγωγή δεδομένων από τα στοιχεία και να μετατραπεί η πληροφορία σε διάνυσμα χαρακτηριστικών, το οποίο αρκετές φορές γίνεται χειροκίνητα, προκειμένου να μπορεί να επεξεργασθεί από υπολογιστή για να προταθεί,. Επιπλέον, δεδομένου ότι τα συνεργατικά συστήματα δημιουργούν συστάσεις χρησιμοποιώντας τις αξιολογήσεις άλλων χρηστών, μπορούν να έρθουν αντιμέτωπα με οποιοδήποτε είδους περιεχομένου και να συστήσουν οποιαδήποτε είδη αντικειμένων, ακόμη και αυτά που διαφέρουν από εκείνα που έχουν προταθεί στο παρελθόν. Ουσιαστικά εφόσον υπάρχουν βαθμολογίες για διαφορετικά είδη στοιχείων, όπως ταινίες, μουσική και λοιπά, το σύστημα θα είναι πάντα ικανό να δώσει συστάσεις.

Παρόλα αυτά, τα συνεργατικά συστήματα έχουν τους δικούς τους περιορισμούς που περιγράφονται παρακάτω:

- Πρόβλημα νέου χρήστη

Όπως και στα συστήματα με βάση το περιεχόμενο ,προκειμένου να γίνουν ακριβείς συστάσεις, το σύστημα πρέπει να πρώτα να αποκτήσει εμπειρία στις προτιμήσεις του χρήστη παρατηρώντας τις αξιολογήσεις που αυτός δίνει. Διάφορες τεχνικές έχουν προταθεί για να αντιμετωπιστεί αυτό το πρόβλημα. Οι περισσότερες από αυτές χρησιμοποιούν υβριδικά συστήματα σύστασης, τα οποία συνδυάζουν τα βασισμένα στο περιεχόμενο συστήματα με τις συνεργατικές τεχνικές.

- Πρόβλημα νέου αντικειμένου

Τα καινούρια στοιχεία που προστίθενται τακτικά στο σύστημα προκειμένου να συσταθούν, έρχονται σε αντίθεση με τις τεχνικές των συνεργατικών συστημάτων, που βασίζονται αποκλειστικά στην επεξεργασία των ήδη διατυπωμένων προτιμήσεων των χρηστών. Συνεπώς, έως ότου το νέο στοιχείο που έχει εισέλθει στο σύστημα αξιολογηθεί από ένα ικανοποιητικό αριθμό χρηστών, το συνιστόν σύστημα δεν θα βρίσκεται σε θέση να το προτείνει. Αυτό το πρόβλημα μπορεί επίσης να αντιμετωπιστεί με χρήση υβριδικών συστημάτων.

- Πρόβλημα ανεπάρκειας δεδομένων (Sparsity problem)

Σε κάθε σύστημα συστάσεων, ο αριθμός των αξιολογήσεων που έχουν ήδη ληφθεί, είναι συνήθως πολύ μικρότερος σε σύγκριση με τον αριθμό των αξιολογήσεων που πρέπει να προβλεφθούν. Η αποτελεσματική πρόβλεψη βαθμολογιών από έναν αρκετά μικρό μέγεθος δείγματος είναι σημαντική. Επίσης, η επιτυχία των συστημάτων συνεργατικής διήθησης, εξαρτάται από τη διαθεσιμότητα ενός κρίσιμου αριθμού χρηστών. Για παράδειγμα, σε κάποιο σύστημα σύστασης κινηματογραφικών ταινιών, μπορεί να υπάρχουν αρκετές ταινίες που έχουν

βαθμολογηθεί από πολύ λίγα άτομα και αυτές οι ταινίες θα συστηθούν πολύ σπάνια, ακόμη και αν αυτοί οι λίγοι χρήστες τους έδωσαν υψηλές βαθμολογίες.

Επιπλέον , για τους χρήστες των οποίων τα γούστα είναι αρκετά ασυνήθιστα σε σχέση με αυτά του υπόλοιπου πληθυσμού, δύσκολα θα υπάρξουν ικανοποιητικά όμοιοι με άλλους χρήστες, με αποτέλεσμα οι προβλέψεις να είναι άστοχες. Οι χρήστες αυτοί λέγονται και “μαύρα πρόβατα” ενώ οι υπόλοιποι που δεν έχουν ασυνήθιστες βαθμολογίες με συνέπεια να έχουν ικανοποιητικά μέτρα συσχετίσεων ονομάζονται “λευκά πρόβατα”.

Ένας τρόπος για να ξεπεραστεί το πρόβλημα του πολύ μικρού αριθμού αξιολογήσεων, είναι η δυνατότητα κάποιου συστήματος να χρησιμοποιεί, δημογραφικές πληροφορίες από το προφίλ του χρήστη για τον υπολογισμό ομοιότητας με άλλους χρήστες [16]. Δηλαδή, δύο χρήστες μπορούν να θεωρηθούν παρόμοιοι όχι μόνο αν κατέτασσαν ταινίες με ίδιο τρόπο, αλλά ακόμα και αν ανήκουν στην ίδια δημογραφική κατηγορία. Κάποιες υλοποιήσεις τέτοιων συστημάτων κάνουν χρήση πληροφοριών όπως το φύλο, η ηλικία, ο κωδικός περιοχής, η εκπαίδευση και η απασχόληση των χρηστών για να πραγματοποιήσουν συστάσεις. Αυτού του είδους η επέκταση των παραδοσιακών τεχνικών συνεργατικής διήθησης καλείται μερικές φορές “δημογραφική διήθηση".

- Πρόβλημα Κλιμάκωσης (Scalability Problem)

Ένα σύστημα σύστασης θα πρέπει να είναι ικανό να παράγει προβλέψεις σε μικρό χρονικό διάστημα. Καθώς όμως αυξάνεται ο πληθυσμός των χρηστών και των αντικειμένων, τότε αυξάνονται και οι πόροι που καταναλώνει το σύστημα για να υπολογίσει τα μέτρα ομοιότητας ανάμεσα στους χρήστες. Σαν συνέπεια, πολλές φορές ο υπολογισμός της εκτίμησης, μπορεί να υπερβαίνει τα ανεκτά χρονικά περιθώρια, φαινόμενο ανεπίτρεπτο για ιστοτόπους με χιλιάδες επισκέψεις το δευτερόλεπτο.

Συνήθως για να αντιμετωπισθεί αυτού του είδους το πρόβλημα, γίνεται χρήση κάποιας βάσης δεδομένων, όπου έχουν προϋπολογιστεί και αποθηκευτεί τα μέτρα των ομοιοτήτων των χρηστών, για να μην είναι αναγκασμένο το σύστημα να τα υπολογίζει κάθε φορά παρά μόνο να εκτελεί ένα ερώτημα για ανάκτηση δεδομένων.

ΚΕΦΑΛΑΙΟ 3
ΜΕΘΟΔΟΛΟΓΙΑ

Σύνολα Δεδομένων (Data set)

Για τη σύγκριση των αποτελεσμάτων των διαφορετικών αλγορίθμων συνεργατικής διήθησης, εκτελέστηκαν μηχανισμοί πρόβλεψης, χρησιμοποιώντας δεδομένα που έχουν εξαχθεί από τον ιστότοπο της MovieLens, σκοπός του οποίου είναι η προτάσεις κινηματογραφικών ταινιών. Το πακέτο δεδομένων αποτελείται από 100.000 αξιολογήσεις 943 χρηστών, κλίμακας 1 ως 5, πάνω σε 1682 ταινίες με κάθε χρήστη να έχει βαθμολογήσει τουλάχιστον 20 αντικείμενα. Λαμβάνοντας υπόψη τον αριθμό των αξιολογήσεων προς τον συνολικό αριθμό τον παρατηρήσεων, αν όλοι οι χρήστες είχαν βαθμολογήσει όλες τις ταινίες μπορούμε να χαρακτηρίσουμε το συγκεκριμένο σύνολο δεδομένων ότι έχει επίπεδο αραιότητας (Sparsity level):

Επίσης για κάθε άτομο περιλαμβάνονται τα παρακάτω δημογραφικά στοιχεία, ταχυδρομικός κώδικας, ηλικία, φύλλο και επάγγελμα, μέσα από μια λίστα 21 δυνατών επιλογών.

Λαμβάνοντας υπόψη τον αριθμό των ταινιών N και των χρηστών M του συνόλου δεδομένων της MovieLens έχει δημιουργηθεί ένας **N x M** πίνακας **R**. Κάθε θέση του $R(i,u) = x$ δείχνει την αξιολόγηση του χρήστη **u** για το αντικείμενο **i**. Το $x \in \{1,2,3,4,5\}$ ενώ η προεπιλεγμένη τιμή του x είναι η μηδενική αν και μόνο αν, κάποιος χρήστης δεν έχει ακόμα βαθμολογήσει κάποια ταινία.

Κάθε κολώνα του πίνακα μπορεί να γραφεί με τη μορφή $u_m = R(:,m)$ $όπου$ $m \in M$ και απεικονίζει όλες τις παρατηρήσεις του κάθε χρήστη ενώ αντίστοιχα κάθε γραμμή του πίνακα $t_n = R(n,:)$ $όπου$ $n \in N$ μπορεί να υποδείξει τις βαθμολογίες όλων των χρηστών για μια ταινία. Στον παρακάτω πίνακα απεικονίζεται ένα δείγμα πίνακα αξιολογήσεων

	Νίκος	Κώστας	Μαρία	Γιώργος
Star wars	5	2	5	4
Titanic	2	4	0	2
The Shawshank Redemption	2	2	4	2
The Godfather	5	1	5	0

Πίνακας 1 Παράδειγμα πίνακα αξιολογήσεων

Για την εκτέλεση των μετρήσεων έχει επιλεγεί τυχαία ένα δείγμα 100 ατόμων περίπου όσο το ένα δέκατο του συνολικού πληθυσμού. Από κάθε χρήστη στο δοκιμαστικό σύνολο, δέκα αξιολογήσεις έχουν αποκρυφτεί ,για τις οποίες έχει ενεργοποιηθεί ο μηχανισμός και έχει εκτιμηθεί από το σύστημα η βαθμολογία.

Μετρήσεις

Προκειμένου να μετρηθεί η απόδοση κάθε συστήματος υπάρχουν ποσότητες που μπορούν να χαρακτηρίσουν την αποτελεσματικότητα ενός συστήματος σύστασης.

Η πρώτη είναι η ακρίβεια του συστήματος πρόβλεψης ,η οποία αξιολογεί πόσο σωστές είναι οι προβλέψεις και μπορεί να υπολογιστεί συγκρίνοντας τις εκτιμήσεις με τις πραγματικές βαθμολογίες, για κάποιο αντικείμενο το οποίο ο χρήστης έχει ήδη αξιολογήσει. Ένα μέτρο που να μπορεί να χαρακτηρίσει την διαφορά των εκτιμώμενων τιμών από τις πραγματικές είναι το μέσο απόλυτο σφάλμα (**Mean Absolute error**) και μπορεί να υπολογιστεί από την παρακάτω σχέση:

Εξίσωση 11

όπου **est** οι εκτιμώμενες τιμές και real οι βαθμολογίες των χρηστών για το αντικείμενο **i**.

Μια δεύτερη ποσότητα που είναι ικανή να αξιολογήσει ένα σύστημα σύστασης, είναι η κάλυψη (**Coverage**) δηλαδή το ποσοστό των αντικειμένων που το σύστημα καταφέρνει να δώσει κάποια πρόβλεψη προς το συνολικό αριθμό των αιτημάτων που δέχεται. Συνηθισμένοι παράγοντες που επηρεάζουν αρνητικά αυτό το μέτρο, είναι συνήθως τα πολύ μικρά σύνολα ομοίων χρηστών και τα καινούρια αντικείμενα που εισάγονται στο σύστημα και δεν έχουν αρκετές ως και καθόλου αξιολογήσεις να τα υποστηρίξουν.

Τέλος, σκοπός ενός συστήματος σύστασης, είναι η πρόταση αντικειμένων τα οποία ο χρήστης θα αξιολογούσε με σημεία από το πάνω όριο της κλίμακας. Επομένως είναι επιβεβλημένο ένα μέτρο που να εκφράζει την ευστοχία του αλγορίθμου στις υψηλές βαθμολογίες. Για παράδειγμα κάποιο άτομο ενδιαφέρεται να ασχοληθεί μόνο με ταινίες που έχει αξιολογήσει με τρία τέσσερα ή πέντε (D+), τα χαμηλότερα σημεία είναι περιττό να ληφθούν υπόψη ,μιας και θα αποτελούν το σύνολο των μη αρεστών ταινιών για το χρήστη (D-), κατά την μέτρηση της απόδοσης του συστήματος. Ουσιαστικά χρειάζεται κάποια ποσότητα, η οποία θα χαρακτηρίζει την ικανότητα υποστήριξης απόφασης του μηχανισμού. Η ποσότητα αυτή είναι η ευαισθησία λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristics Sensitivity) [13].

Όπως αναφέρθηκε και παραπάνω, η σύσταση αντικειμένων μπορεί να θεωρηθεί δυαδική διαδικασία. Υπάρχουν δηλαδή δυο ενδεχόμενα για τον χρήστη: είτε θα θέλει να δει την ταινία (positive P) είτε δεν θα θέλει (negative N). Επιλέγοντας ένα σημείο στη συνεχή κλίμακα (σημείο απόφασης), για παράδειγμα το σημείο 3 ή το 4, το σύστημα ταξινομεί ως μη άξιες για πρόταση τις ταινίες με εκτιμώμενη βαθμολογία κάτω του σημείου απόφασης (ένδειξη αρνητική) ενώ όλες τις ταινίες με εκτιμώμενη βαθμολογία πάνω από το σημείο απόφασης ως άξιες για να τις προτείνει (ένδειξη θετική). Λόγω της ατέλειας των συστημάτων σύστασης θα υπάρχουν ταινίες όπου ο χρήστης θα τις παρακολουθούσε άνετα αλλά ταξινομήθηκαν ως μη αρεστές όπως και το αντίθετο.

Ορίζεται ως ευαισθησία (sensitivity ή true positive rate TPR) του συστήματος σύστασης, το ποσοστό των θετικών ενδείξεων στον πληθυσμό των ταινιών που ο χρήστης θα παρακολουθούσε (και τη συμβολίζουμε με TP: true positive fraction) και ως ειδικότητα (specificity), το ποσοστό των αρνητικών ενδείξεων στον πληθυσμό των ταινιών που δεν θα ενδιέφεραν τον χρήστη (και τη συμβολίζουμε με TN: true negative fraction). Για κάθε σημείο απόφασης ορίζεται ένας 2x2 πίνακας όπως ο παρακάτω πίνακας:

	D-	D+
D-	TN	FP
D+	FN	TP

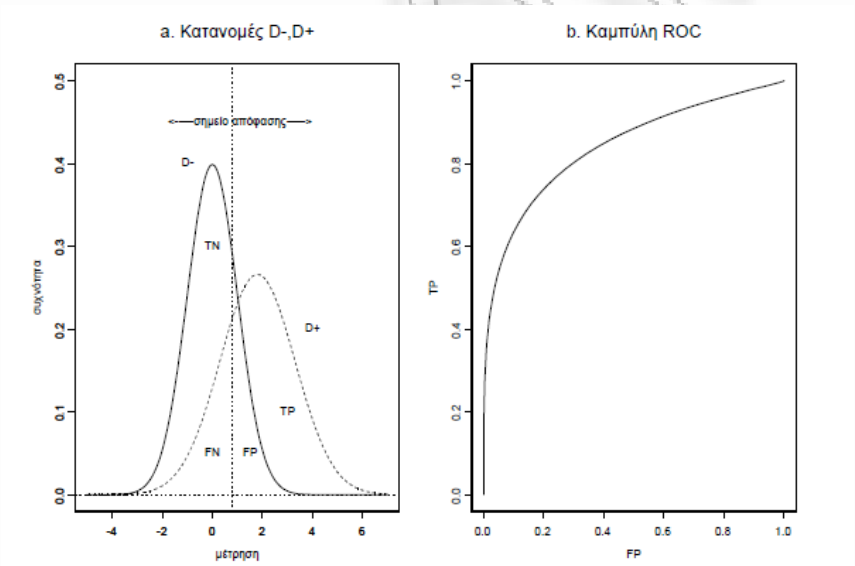
Πίνακας 2

όπου FP=1-TN (δηλαδή False Positive είναι το ποσοστό των θετικών ενδείξεων στις ταινίες όπου ο χρήστης δεν θα παρακολουθούσε) και FN=1-TP (δηλαδή False Negative είναι το ποσοστό των αρνητικών ενδείξεων στις μη ικανοποιητικές ταινίες).

Εξίσωση 12

Η καμπύλη ROC που αντιστοιχεί σε μια πρόβλεψη βαθμολογίας είναι το συνεχές γράφημα που ορίζουν τα σημεία (FP,TP) για όλα τα δυνατά σημεία απόφασης στο μοναδιαίο τετράγωνο [0,1] x [0,1] και ξεκινά από το σημείο (0,0) για να καταλήξει στο σημείο (1,1) (εικόνα 2). Πρακτικά, ορίζονται αντίστοιχοι πίνακες με τον πίνακα 2 και τα αντίστοιχα σημεία (FP, TP) τα οποία ενώνονται με ευθύγραμμα τμήματα και ορίζουν την καμπύλη ROC.

Το εμβαδόν κάτω από την καμπύλη ROC χρησιμοποιείται ως δείκτης διαχωρισμού των κατανομών αρεστών και μη αρεστών ταινιών και υπολογίζεται μη-παραμετρικά σύμφωνα με τον κανόνα του τραπεζίου με βάση τα σημεία (FP,TP) που έχουν υπολογιστεί. Η ελάχιστη τιμή που μπορεί να πάρει είναι 0.5 όταν οι δύο κατανομές συμπίπτουν απόλυτα και η μέγιστη 1.0 όταν οι δύο κατανομές δε συμπίπτουν πουθενά.



Εικόνα 2 a. Κατανομές αρεστών - μη αρεστών αντικειμένων και ποσοστά αποφάσεων b. Καμπύλη ROC

Για τις μετρήσεις πρέπει να θέσουμε ένα σημείο απόφασης το οποίο θεωρήσαμε το 4. Λαμβάνοντας αυτό υπόψη υπολογίστηκε η ευαισθησία (TPR) δηλαδή στο σύνολο των ικανοποιητικών για τον χρήστη ταινιών, σε τι ποσοστό είχε το σύστημα επιτυχία και εκτίμησε ότι θα του άρεσαν δηλαδή τις πρόβλεψε τη βαθμολογία τους στο 4 ή 5.

Υλοποίηση συστήματος σύστασης συνεργατικής διήθησης

Από την περισυλλογή των αξιολογήσεων των χρηστών μέχρι την παραγωγή πρόβλεψης, η υλοποίηση του συστήματος σύστασης συνεργατικής διήθησης είναι προϊόν τριών επιμέρους διαδικασιών :

- Υπολογισμός ομοιότητας κάθε χρήστη σε σχέση με τον ενεργό χρήστη
- Επιλογή υποσυνόλου χρηστών (γειτόνων) οι οποίοι θα χρησιμοποιηθούν για την πρόβλεψη της βαθμολογίας του χρήστη
- Εκτίμηση πρόβλεψης σύμφωνα με συμψηφισμό των αξιολογήσεων ενός συγκεκριμένου αντικειμένου, από τους γείτονες του ενεργού χρήστη.

Υπολογισμός ομοιότητας

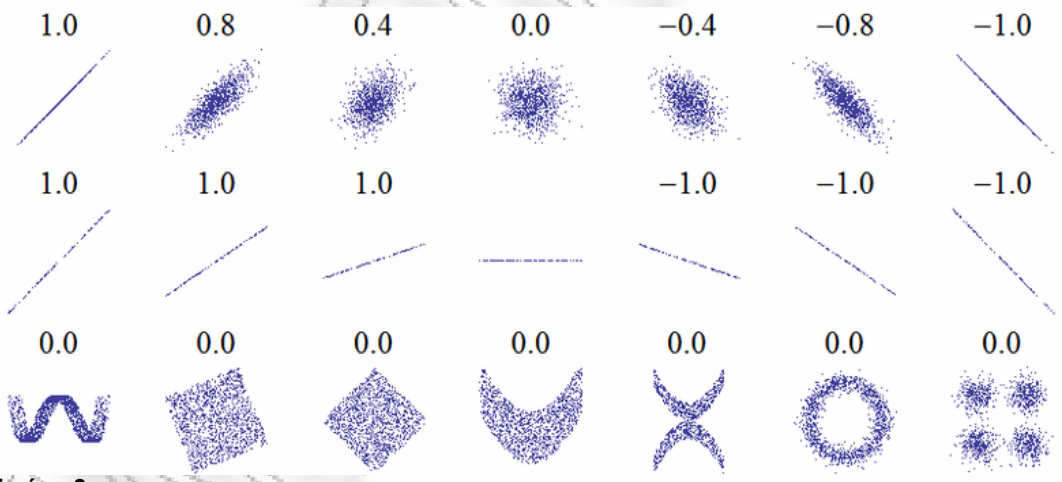
Όταν υπολογίζεται μια πρόβλεψη, λογικό είναι να τρέφει το σύστημα εμπιστοσύνη σε άτομα που θεωρούνται ότι έχουν παρόμοιες προτιμήσεις με τον ενεργό χρήστη και να βασίζεται σε αυτά. Το πρώτο βήμα λοιπόν στην υλοποίηση, αποτελείται από την προσπάθεια να βρεθεί μια ποσότητα που να περιγράφει πόσο όμοιες προτιμήσεις έχουν δυο άτομα μεταξύ τους. Οι πιο συνηθισμένες μέθοδοι υπολογισμού ομοιότητας είναι οι συσχετίσεις (correlations) με κυριότερο αντιπρόσωπο την συσχέτιση Pearson (Pearson correlation coefficient ή συχνά και **Pearson r**) και ομοιότητα συνημίτονου.

Η συσχέτιση Pearson πραγματοποιεί μια εκτίμηση για να χαρακτηρίσει τον βαθμό στον οποίο παρατηρείται μια γραμμική σχέση ανάμεσα σε δυο μεταβλητές **u** , **u'**. Οι τιμές του μέτρου συσχέτισης είναι μέσα στο κλειστό διάστημα [-1 , 1] και αν οι μεγαλύτερες τιμές τις μεταβλητής **u** αντιστοιχούν στις μεγαλύτερες τιμές τις μεταβλητής **u'** και αντίστοιχα οι μικρότερες τιμές της **u** στις μικρότερες της **u'** τότε το μέτρο είναι θετικό και τείνει στο +1, οπότε και έχουμε θετική συσχέτιση. Αντίθετα αν οι μεγαλύτερες τιμές της μιας μεταβλητής τείνουν στις μικρότερες της άλλης και αντίστροφα τότε, το μέτρο συσχέτισης θα πρέπει να έχει μία τιμή αρνητική, η οποία να είναι κοντά στην τιμή -1, οπότε και οι μεταβλητές είναι αρνητικά συσχετισμένες. Τέλος αν οι τιμές της μεταβλητής **u** φαίνεται να αντιστοιχούν με τυχαίο τρόπο στις τιμές τις **u'** τότε χαρακτηρίζονται ασυσχέτιστες και το μέτρο τείνει στην τιμή μηδέν. Αν λοιπόν **u** το διάνυσμα με τις αξιολογήσεις του ενεργού χρήστη και **u'** κάποιου άλλου χρήστη για να βρεθεί το μέτρο της ομοιότητας χρησιμοποιείτε η παρακάτω σχέση.

$$\mathbf{similarity}(u, u') = r = \frac{\sum_{i=1}^{m}(u_i - \bar{u})(u'_i - \bar{u}')}{\sigma_u \sigma_{u'}}$$

Εξίσωση 13

Στα παρακάτω γραφήματα διασποράς φαίνεται το μέτρο συσχέτισης σε σχέση με τις τιμές των δυο μεταβλητών.



Εικόνα 3

Ο συσχετισμός Pearson εξάγεται από ένα γραμμικό μοντέλο το οποίο είναι βασισμένο σε ένα σύνολο από παραδοχές που αφορούν τα δεδομένα και κυρίως ότι πρέπει να είναι γραμμική η σχέση που τα χαρακτηρίζει και τα σφάλματα να είναι ανεξάρτητα και να έχουν πιθανοτική κατανομή με μέση τιμή μηδέν και σταθερή διακύμανση. Αν δεν ικανοποιούνται αυτοί οι περιορισμοί τότε η συγκεκριμένη μέθοδος χαρακτηρίζεται ως αναξιόπιστη.

Εκτός του Pearson correlation coefficient έχει προταθεί και ο συσχετισμός του Spearman, (Spearman rank correlation coefficient ή Spearman rho) όπου δεν τον διέπουν οι περιορισμοί του Pearson, και ουσιαστικά δεν είναι κάτι άλλο από τον συντελεστή **Pearson r**

υπολογιζόμενο, όμως, με βάση την κατάταξη που έχει η κάθε παρατήρηση και όχι αυτές καθαυτές τις βαθμολογίες

$$\mathbf{similarity}(u, u') = \mathbf{rho} = \frac{\sum_{i=1}^{m}(\mathbf{rank}_{u,i} - \overline{\mathbf{rank}_u})(\mathbf{rank}_{u',i} - \overline{\mathbf{rank}'_u})}{\sigma_u \sigma'_u}$$

Εξίσωση 14

Περνώντας στη δεύτερη κατηγορία υπολογισμού παρόμοιων χρηστών, το μέτρο ομοιότητας συνημίτονου, ουσιαστικά αποτελεί τη γωνία των διανυσμάτων των παρατηρήσεων των χρηστών και υπολογίζεται από τη εξίσωση 3. Αν η γωνία τους είναι 0 τότε το συνημίτονο παίρνει τιμή 1 και θεωρείται ότι τα διανύσματα χαρακτηρίζονται από τέλεια συσχέτιση ενώ -1 αν η γωνία τους είναι 180 και θεωρούνται αντίθετα.

Σε αρκετές όμως περιπτώσεις έχει παρατηρηθεί ότι η τιμή που θα αποδοθεί από τον υπολογισμό κάποιας συσχέτισης μεταξύ δυο χρηστών, μπορεί να είναι παραπλανητική. Ένας πολύ μικρός αριθμός στοιχείων με ίδιες αξιολογήσεις ανάμεσα στους δυο χρήστες ενδεχομένως να έχει τη δυνατότητα να τους χαρακτηρίσει με ασυνήθιστα μεγάλα μέτρα συσχέτισης [3]. Για να αποφευχθεί η παγίδα αυτή, υπάρχει η δυνατότητα υποβάθμισης του αποτελέσματος της συσχέτισης. Όσο περισσότερες ίδιες αξιολογήσεις έχουν δυο χρήστες τόσο περισσότερη εμπιστοσύνη θα πρέπει να δείξουμε στον χρήστη αυτόν και στο μέτρο της συσχέτισης τους. Γι' αυτό το λόγο χρησιμοποιήθηκε ένα σύστημα βαρύτητας, σε συνδυασμό με τις συσχετίσεις, που αναλόγως του αριθμού των κοινών βαθμολογιών δυο χρηστών σε κάποια προϊόντα, μεγαλύτερο η μικρότερο θα είναι το ποσοστό εμπιστοσύνης που μπορεί να υπάρξει στον μεταξύ τους συσχετισμό. Δηλαδή εάν δυο χρήστες έχουν n κοινά σημεία με n μικρότερο ενός αριθμού N τον οποίο θεωρούμε ικανοποιητικό σαν αριθμό κοινών σημείων, τότε θα πολλαπλασιαστεί το μεταξύ τους μέτρο συσχέτισης με τον συντελεστή n/N ώστε να υποβαθμιστεί, ενώ αν το n είναι μεγαλύτερο η ίσο του N τότε το μέτρο θα παραμείνει ανέπαφο.

Επιλογή γειτόνων

Ύστερα από την απόδοση μέτρων συσχέτισης ανάμεσα στους χρήστες και την υποβάθμιση των αφερέγγυων χρηστών με χρήση βαρύτητας εμπιστοσύνης, πρέπει να αποφασιστεί ποιοι θα εξαιρεθούν από όλο τον πληθυσμό, δηλαδή, ποιοι χρήστες θα δυσκόλευαν τη διαδικασία χωρίς να μπορούσαν να προσφέρουν κάτι ουσιαστικό. Χρήστες με βαθμολογίες που έχουν μεγάλη πιθανότητα να συνεισφέρουν σε σωστές προβλέψεις, ενδεχομένως να χαθούν μέσα στο θόρυβο από το μεγάλο σύνολο των χρηστών που έχουν πολύ μικρή πιθανότητα, αν η επιλογή δεν πραγματοποιηθεί με σωστά κριτήρια. Εκτός από το όφελος της αξιοπιστίας υπάρχει και αυτό της επίδοσης, ειδικά όταν ένα σύστημα βρίσκεται αντιμέτωπο με έναν πολύ μεγάλο πληθυσμό χρηστών, όπου δυσχεραίνει κατά πολύ την όλη διαδικασία. Για τους λόγους αυτούς έχουν εφαρμοστεί κάποιοι τρόποι φιλτραρίσματος και επιλογής χρηστών που θα μπορούσαν να χαρακτηριστούν ως ικανοποιητικά όμοιοι με τον ενεργό χρήστη. Σε αυτήν την εργασία έχουν δοκιμαστεί δυο τρόποι στους οποίους γίνεται έλεγχος για το αν το μέτρο συσχέτισης του υποψήφιου γείτονα με τον ενεργό χρήστη: είναι είτε στα **N** μεγαλύτερα, είτε αν ξεπερνάει κάποιο όριο.

Όταν όμως ελαχιστοποιηθεί ή μειωθεί αρκετά ο αριθμός των γειτόνων το σύστημα αδυνατεί να δώσει προβλέψεις λόγω ελλιπών δεδομένων σε έναν σημαντικό αριθμό αντικείμενων. Ένα αποτέλεσμα είναι η μείωση κατά πολύ της κάλυψης του συνόλου των αντικειμένων και τελικά η παρουσία μεγάλου κίνδυνου να μην είναι δυνατό για κάποιον χρήστη να λάβει σύσταση για κάποιο προϊόν που είναι πιθανό να το αξιολογούσε με μεγάλη βαθμολογία.

Εκτίμηση πρόβλεψης

Εφόσον πλέον έχει δημιουργηθεί το σύνολο των γειτόνων, οι αξιολογήσεις τους πρέπει να συναθροιστούν για την τελική εκτίμηση της βαθμολογίας του ενεργού χρήστη πάνω σε κάποιο

αντικείμενο. Μια από τις πρώτες τεχνικές που εφαρμόστηκαν ήταν η μέθοδος της **GroupLens** [6] στην οποία θεωρούνται γείτονες όσοι έχουν βαθμολογήσει το ζητούμενο αντικείμενο και όχι όσοι ικανοποιούν μια από τις συνθήκες του προηγούμενου βήματος.

$$R(u,i) = \bar{u} + \frac{\sum_{u'=1}^{n} (R(u',i) - \overline{u'}) * sim(u,u')}{\sum_{u'=1}^{n} sim(u,u')}$$

Εξίσωση 15

Στην περίπτωση της **GroupLens** γίνεται η εκτίμηση της βαθμολογίας με την απόκλιση από τη μέση τιμή των αξιολογήσεων του ενεργού χρήστη $\bar{u}$, χρησιμοποιώντας τις τιμές της συσχέτισης Pearson σαν τις βαρύτητες για το ποσό της εμπιστοσύνης που θα ανατεθεί και άρα το ποσοστό της συμμετοχής του χρήστη στο τελικό αποτέλεσμα.

Κάποιοι άλλοι μηχανισμοί που έχουν χρησιμοποιηθεί είναι η μέση τιμή των αξιολογήσεων των γειτόνων για το αντικείμενο **i**

$$R(u,i) = \frac{\sum_{u'=1}^{n} R(u',i)}{n}$$

Εξίσωση 16

όπως επίσης και η απόκλιση από τη μέση τιμή που αντίθετα με την υλοποίηση της GroupLens δεν χρησιμοποιεί βαρύτητες για να προσδιορίσει το ποσοστό συμμετοχής του κάθε χρήστη ,ούτε εφαρμόζεται σε όσους χρήστες έχουν βαθμολογήσει το αντικείμενο i.

$$R(u,i) = \bar{u} + \frac{\sum_{u'=1}^{n} (R(u',i) - \overline{u'})}{n}$$

Εξίσωση 17

Συνοπτικά στον παρακάτω πίνακα φαίνονται οι τεχνικές που έχουν χρησιμοποιηθεί σε αυτήν την εργασία.

ΔΙΑΔΙΚΑΣΙΑ	ΜΕΘΟΔΟΣ
ΥΠΟΛΟΓΙΣΜΟΣ ΟΜΟΙΟΤΗΤΑΣ	ΣΥΣΧΕΤΙΣΕΙΣ PEARSON
	ΣΥΣΧΕΤΙΣΕΙΣ SPEARMAN
	ΟΜΟΙΟΤΗΤΑ ΣΥΝΗΜΙΤΟΝΟΥ
	ΥΠΟΒΑΘΜΙΣΗ ΜΕ ΧΡΗΣΗ ΣΥΝΤΕΛΕΣΤΗ ΒΑΡΥΤΗΤΑΣ
ΕΠΙΛΟΓΗ ΓΕΙΤΟΝΩΝ	ΕΠΙΛΟΓΗ **N** ΠΙΟ ΟΜΟΙΩΝ
	ΕΠΙΛΟΓΗ ΜΕ ΕΦΑΡΜΟΓΗ ΟΡΙΟΥ
ΥΠΟΛΟΓΙΣΜΟΣ ΠΡΟΒΛΕΨΗΣ	ΜΕΣΗ ΤΙΜΗ ΒΑΘΜΟΛΟΓΙΩΝ
	ΑΠΟΚΛΙΣΗ ΑΠΟ ΤΗ ΜΕΣΗ ΤΙΜΗ

Πίνακας 3 Μέθοδοι που έχουν χρησιμοποιηθεί σε αυτήν την εργασία

Επιπλέον για λόγους αναφοράς και μόνο, έχει γίνει υλοποίηση και ενός υβριδικού συστήματος συνδυασμού χαρακτηριστικών (Feature combination) λαμβάνοντας υπόψη τα δημογραφικά στοιχεία που περιλαμβάνει για τους χρήστες το σύνολο δεδομένων της MovieLens [14]. Πριν χρησιμοποιηθούν αυτά τα δεδομένα πρέπει να δημιουργηθούν διανύσματα δημογραφικών χαρακτηριστικών για κάθε χρήστη. Έχουμε δημιουργήσει για κάθε άτομο ένα διάνυσμα 27 χαρακτηριστικών, σύμφωνα με τον πίνακα 4 από τα οποία τρία είναι 1 ένα για ηλικία, ένα για το φύλλο και ένα για επάγγελμα, ενώ τα υπόλοιπα 0. Για παράδειγμα αν ο χρήστης είναι 25 ετών τότε θα έχει 1 στο 2 χαρακτηριστικό ενώ 0 σε όλα τα άλλα που αφορούν την ηλικία του. Εκτός από τον υπολογισμό ομοιότητας ανάμεσα στα διανύσματα αξιολογήσεων (sim_ratings) των χρηστών υπολογίζουμε και την ομοιότητα ανάμεσα στα διανύσματα δημογραφικών χαρακτηριστικών (sim_demographic) και χρησιμοποιούμε τις δυο ομοιότητες για να δημιουργήσουμε ένα συντελεστή βαρύτητας:

$$\mathbf{Weight} = \mathbf{sim}_{\mathbf{ratings}} + \mathbf{sim}_{\mathbf{demographic}} * \mathbf{sim}_{\mathbf{demographic}}$$

Εξίσωση 18

Στην συνέχεια έγινε επιλογή N ομοιότερων να λαμβάνουν τον χαρακτηρισμό του γείτονα. Με τη χρήση αυτού του συνόλου των ατόμων υπολογίστηκε με βάση την εξίσωση 19 η πρόβλεψη για κάθε αντικείμενο του δοκιμαστικού συνόλου δεδομένων.

$$\mathbf{R(u,l)} = \overline{\mathbf{u}} + \frac{\sum_{\mathbf{u'}=1}^{\mathbf{n}} (\mathbf{R(u',l)} - \overline{\mathbf{u'}}) * \mathbf{weight(u,u')}}{\sum_{\mathbf{u'}=1}^{\mathbf{n}} |\mathbf{weight(u,u')}|}$$

Εξίσωση 19

χαρακτηριστικό	Περιεχόμενα χαρακτηριστικών	
1	Ηλικία <=18	Σε κάθε χρήστη τοποθετείτε 1 μόνο στο χαρακτηριστικό που αντιστοιχεί στη ζώνη ηλικίας του, στο φύλλο του και στο επάγγελμα του ενώ σε όλα τα άλλα 0.
2	18<ηλικία<=29	
3	29< ηλικία<=49	
4	Ηλικία>49	
5	Άνδρας	
6	Γυναίκα	
7-27	Επάγγελμα	

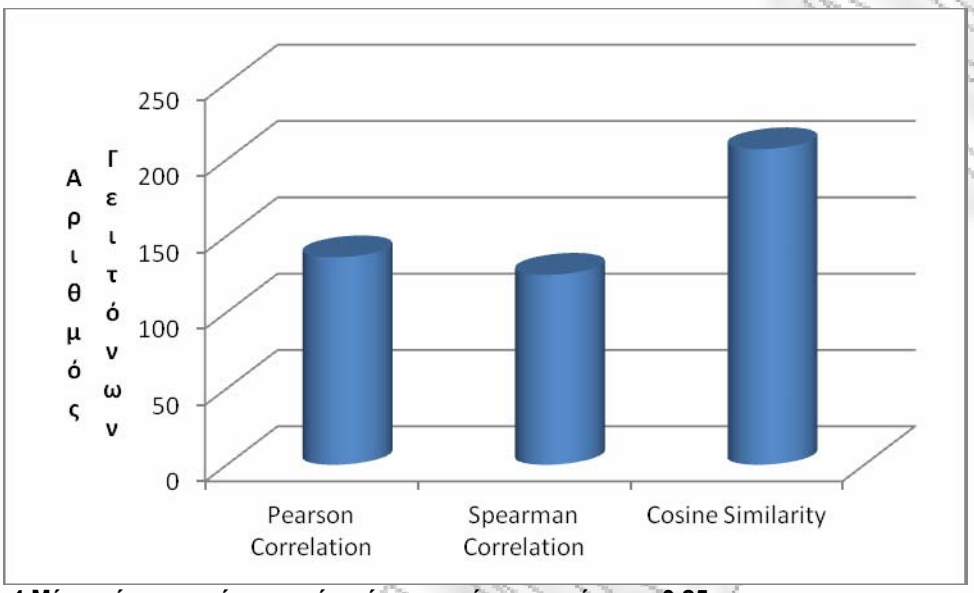
Πίνακας 4 Διάνυσμα δημογραφικών χαρακτηριστικών

Αποτελέσματα

Έχουν υλοποιηθεί διαφορετικά συστήματα, με σκοπό να δημιουργηθούν όλοι οι δυνατοί συνδυασμοί, των μεθόδων του πίνακα 3. Για να υπάρξει ένα καλό μέτρο, για την αποδοτικότητα των αλγορίθμων στον τομέα της κάλυψης, έχει γίνει μέτρηση για τον αριθμό των αιτημάτων ,που το σύστημα κατάφερε να φτάσει σε πρόβλεψη, σε σχέση με τον συνολικό αριθμό αιτημάτων που επεξεργάστηκε. Ενώ όσον αφορά την ακρίβεια των προβλέψεων, έχει μετρηθεί, για όλες τις επιτυχείς εκπληρώσεις αιτημάτων, η μέση απόκλιση από τις πραγματικές τιμές αξιολόγησης των χρηστών, όπως επίσης και για την ικανότητα υποστήριξης απόφασης, η ευαισθησία ROC.

Αρχικά για την πρώτη ομάδα του πίνακα 2, για τον υπολογισμό δηλαδή της ομοιότητας των ατόμων με τον ενεργό χρήστη, έχουν χρησιμοποιηθεί δυο βασικές τεχνικές συσχέτισης: η Pearson r και η Spearman rho, καθώς και η ομοιότητα συνημίτονου. Επιπλέον έγινε χρήση της μεθόδου υποβάθμισης των αναξιόπιστων γειτόνων, με χρήση βαρυτήτων στα ευρήματα των συσχετίσεων. Για να εκτιμηθεί ποια από τις τρεις συσχετίσεις παρέχει πολυτιμότερα αποτελέσματα, ένα πρώτο μέτρο είναι ο έλεγχος του αριθμού των γειτόνων που προσφέρει η κάθε συσχέτιση, στην περίπτωση βέβαια που η επιλογή γίνεται με εφαρμογή

κάποιου ορίου. Περισσότεροι γείτονες αναλογούν σε μεγαλύτερη πιθανότητα, το σύστημα να καταφέρει να δώσει μια εκτίμηση. Οπότε μελετώντας το μέγεθος της γειτονίας ,μπορεί να βγει ένα αρχικό συμπέρασμα για το ποσοστό κάλυψης.



Γράφημα 4 Μέσος όρος γειτόνων ανά χρήστη για όριο ομοιότητας 0.25

Στο γράφημα 4 φαίνεται ότι η ζυγαριά της κάλυψης τείνει προς τη μεριά της ομοιότητας συνημίτονου. Η συγκεκριμένη ομοιότητα είναι πιο ελαστική στις σχέσεις μεταξύ διανυσμάτων και τις βαθμολογεί με μεγαλύτερα σημεία, ενώ οι συσχετίσεις είναι περισσότερο αυστηρές στις βαθμολογίες τους.

Στην συνέχεια συγκρίνοντας ποια μέθοδος της πρώτης διαδικασίας αποδίδει καλύτερα στον τομέα της ακρίβειας πρόβλεψης, έχει χρησιμοποιηθεί σαν μέτρο το Mean Absolute Error που όσο πιο πολύ τείνει στο μηδέν, τόσο πιο ακριβείς είναι οι προβλέψεις του συστήματος. Όπως επίσης έχει χρησιμοποιηθεί και η ROC sensitivity για να εκφραστεί η υποστήριξη απόφασης του εκάστοτε αλγόριθμου. Σε αυτήν την περίπτωση, όσο πιο κοντά τείνει στη μονάδα η ποσότητα αυτή, τόσο πιο εύστοχος κρίνεται μηχανισμός στις προτάσεις του.

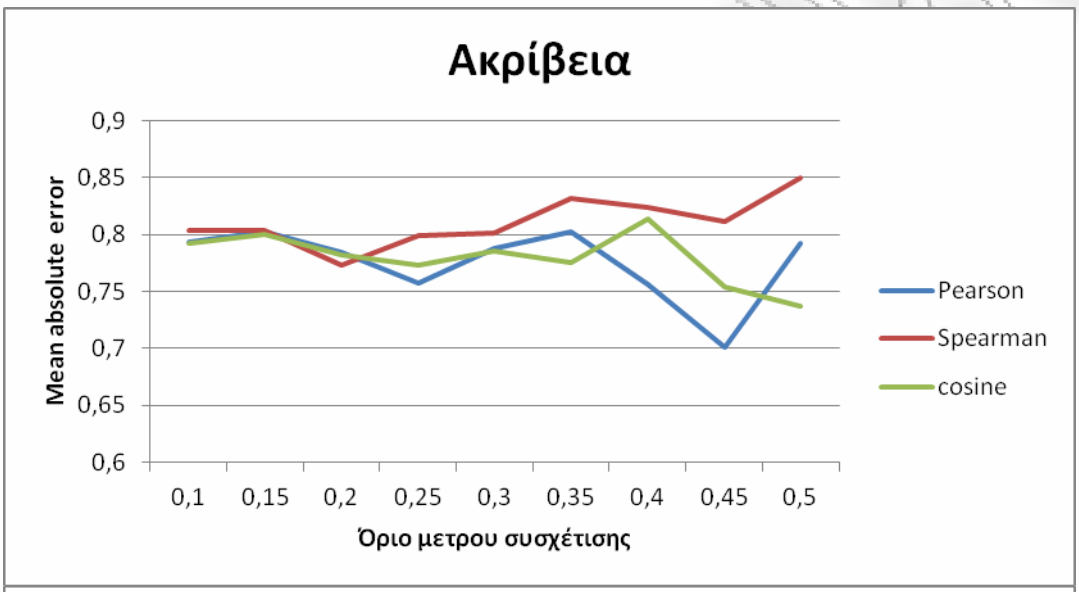
Αναλύοντας τα πειραματικά δεδομένα στο γράφημα 5, για επιλογή γειτόνων με την εφαρμογή ορίου μέτρου συσχέτισης, γίνεται αντιληπτό ότι τα Pearson r και Spearman rho όπως και η ομοιότητα συνημίτονου, από μόνα τους στο θέμα της ακρίβειας και της υποστήριξης απόφασης δεν είναι αξιόπιστα μέτρα ομοιότητας. Όταν όμως επέλθει η υποτίμηση και απομακρύνονται οι αναξιόπιστοι γείτονες, η απόκλιση πραγματικών και εκτιμώμενων παρατηρήσεων μειώνεται ενώ παράλληλα ανεβαίνει

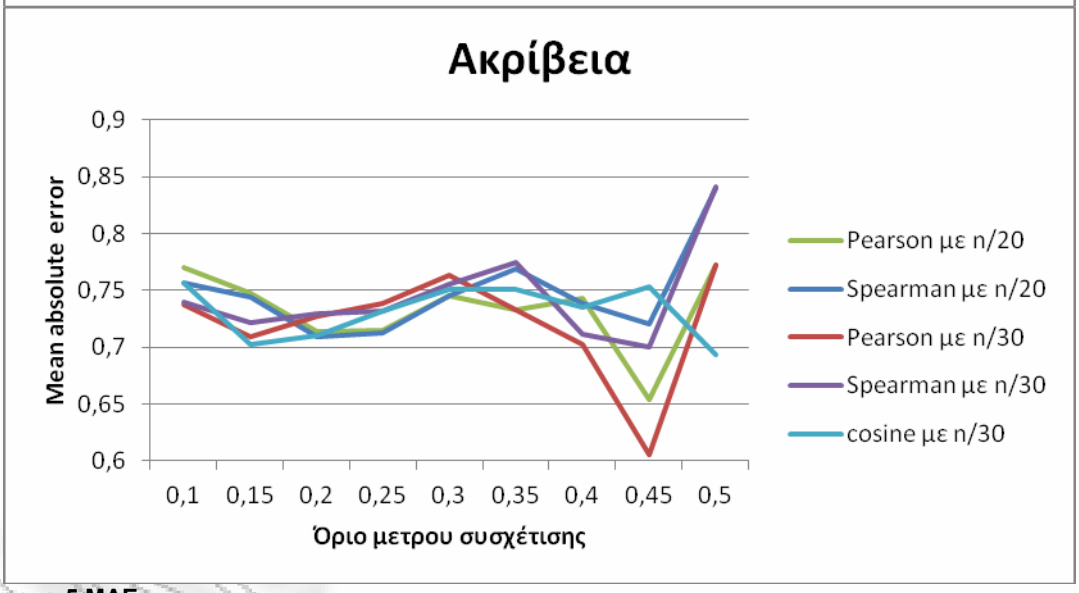
Όριο μέτρου συσχέτισης	MAE		Κάλυψη %		ROC sensitivity	
	Pearson με N/20	Pearson	Pearson με N/20	Pearson	Pearson με N/20	Pearson
0,1	0,769704	0,793587	81,2	99,8	0,775899	0,727586
0,15	0,747159	0,803015	70,4	99,5	0,769953	0,733564
0,2	0,713405	0,784254	64,9	97,8	0,798489	0,755752
0,25	0,715232	0,756959	60,4	93,4	0,808625	0,772643
0,3	0,745098	0,788235	51	85	0,785942	0,740741
0,35	0,732648	0,803051	38,9	72,1	0,761702	0,731415
0,4	0,742947	0,756436	31,9	50,5	0,752632	0,723776
0,45	0,653846	0,701449	20,8	34,5	0,822581	0,741294

0,5	0,772358	0,791908	12,3	17,3	0,783784	0,757009

Πίνακας 5

και η ROC sensitivity. Όμως με την υποβάθμιση ελαττώνεται ο αριθμός των χρηστών, που το μέτρο συσχέτισής τους, με τον ενεργό χρήστη ξεπερνάει το όριο. Η υποτίμηση άρα οδήγησε σε μικρότερο αριθμό ατόμων που μπορούν να θεωρηθούν γείτονες και άρα την κάλυψη σε μη ανεκτά επίπεδα όπως δείχνει ο πίνακας 5 και το γράφημα 6.





Γράφημα 5 MAE

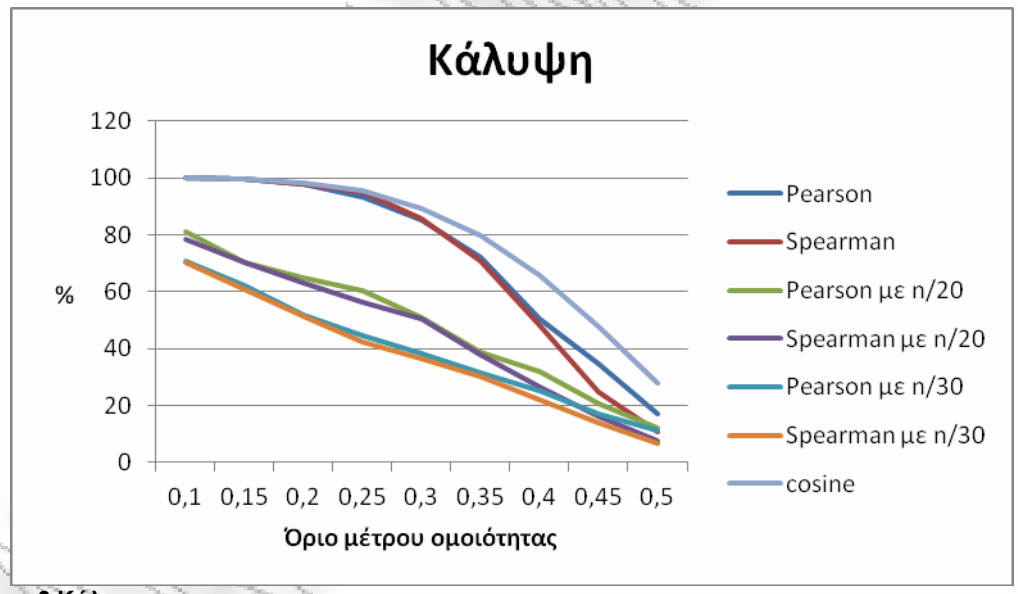
Ένα σύστημα σίγουρα θα πρέπει να έχει όσο μεγαλύτερη αξιοπιστία γίνεται. Θα πρέπει όμως να μπορεί και να δώσει πρόβλεψη, για μεγάλο ποσοστό αιτημάτων που πλησιάζει τη μονάδα και όχι μόνο να καταφέρνει να απαντάει σε ένα μικρό ποσοστό αυτών. Για να καταλήξουμε όμως σε ένα πιο ολοκληρωμένο συμπέρασμα, θα πρέπει να παρατηρήσουμε, πώς συμπεριφέρονται οι ίδιες τεχνικές του πρώτου βήματος, αλλά αυτή τη φορά έχοντας επιλέξει σαν γείτονες τους N ομοιότερους και όχι εφαρμόζοντας κάποιο όριο. Παρατηρώντας λοιπόν τον πίνακα 7 και τα γραφήματα 8 ως 10 φαίνεται πως η συμπεριφορά των δοκιμών που στηρίζονται στην υποβάθμιση των χρηστών, σε συνδυασμό με σταθερό αριθμό πληθυσμού γειτόνων, ανεβάζουν την κάλυψη στα ύψη, ξεπερνώντας το μειονέκτημα των προηγούμενων

συνδυασμών. Το ποσοστό κάλυψης είναι ανεβασμένο στο 100%, επειδή πλέον δεν υπάρχει όριο που μπορεί να αφήνει εκτός, χρήστες που έχουν βαθμολογήσει το αντικείμενο που αναζητάμε, απλά επιλέγουμε έναν αριθμό από αυτούς που έχουν ασχοληθεί με το στοιχείο που θέλουμε. Ένα ακόμα συμπέρασμα που μπορεί να εξαχθεί από αυτά τα πειραματικά δεδομένα για το ποιος αλγόριθμος είναι πιο εύστοχος, φαίνεται να είναι οριακά η συσχέτιση του Spearman έναντι του Pearson, υποβαθμίζοντας τους χρήστες με συντελεστή N/30 σύμφωνα με τις μετρήσεις της ακρίβειας και της υποστήριξης απόφασης.

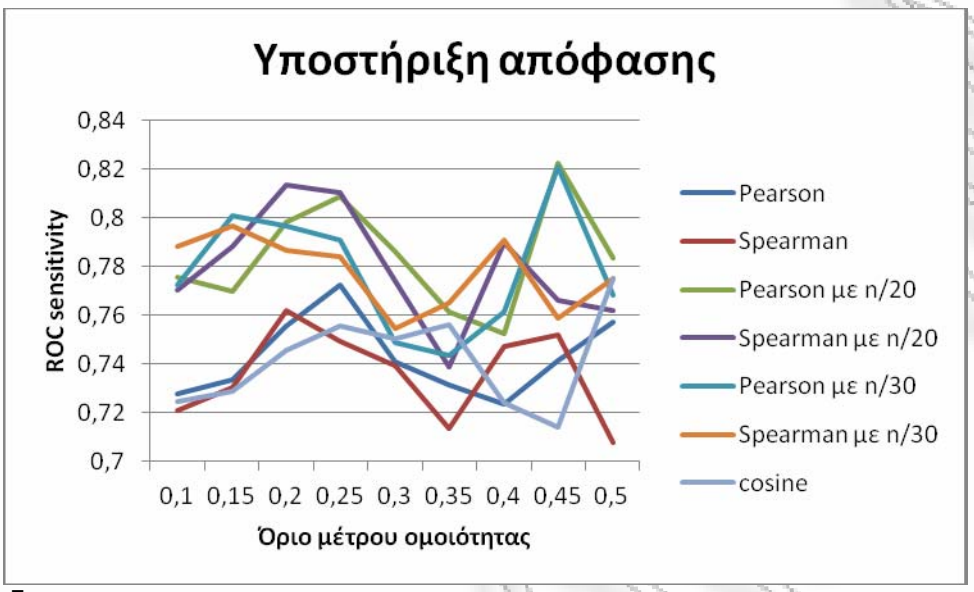
Ένα θέμα επιπλέον που θα μπορούσε να προκύψει σε αυτό το σημείο, μιας και οι διαφορές των συσχετισμών είναι οριακές και θα μπορούσε να αναστρέψει τις ισορροπίες, τουλάχιστον σε σενάρια που περιέχουν μεγάλα εμπορικά συστήματα με χιλιάδες η εκατομμύρια χρήστες και προϊόντα, είναι ο χρόνος εκτέλεσης, όπου η συσχέτιση Pearson είναι κυρίαρχος όπως φαίνεται στο γράφημα 12.

Όριο μέτρου συσχέτισης	MAE	Κάλυψη %	ROC sensitivity
0,1	0,792793	99,9	0,724613
0,15	0,800201	99,6	0,728843
0,2	0,78252	98,4	0,745614
0,25	0,773013	95,6	0,755435
0,3	0,786114	89,3	0,750484
0,35	0,775689	79,8	0,755991
0,4	0,814307	65,7	0,724324
0,45	0,754202	47,6	0,714286
0,5	0,736655	28,1	0,775148

Πίνακας 6



Γράφημα 6 Κάλυψη

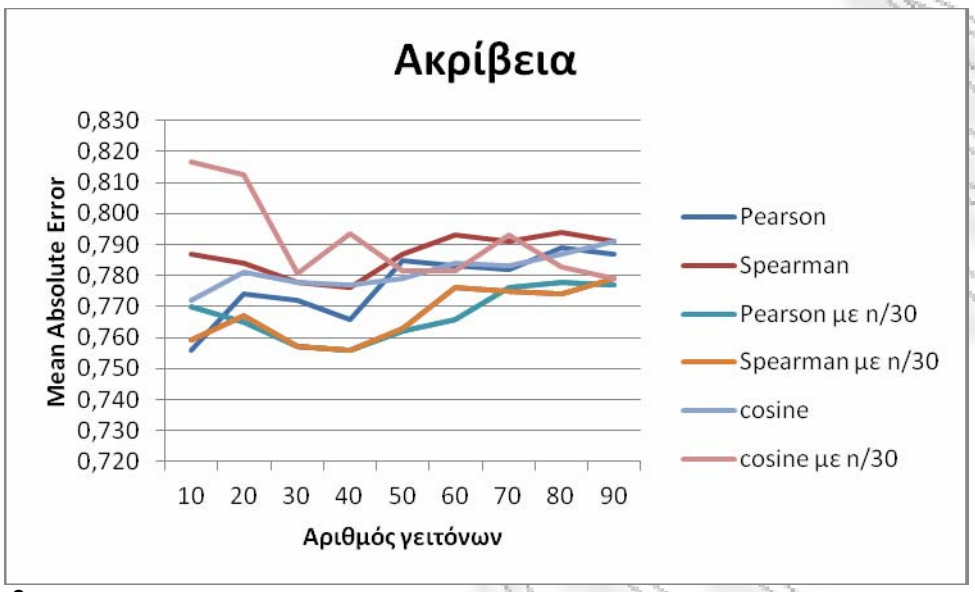


Γράφημα 7

Όπως αναφέρθηκε και παραπάνω ο βέλτιστος συνδυασμός μεθόδων των δυο πρώτων διαδικασιών είναι η χρήση συσχετισμών Pearson, κυρίως λόγω χρόνου εκτέλεσης, με υποβάθμιση ,χρησιμοποιώντας συντελεστή N/30 και η επιλογή γειτόνων γίνεται επιλέγοντας τους N ομοιότερους.

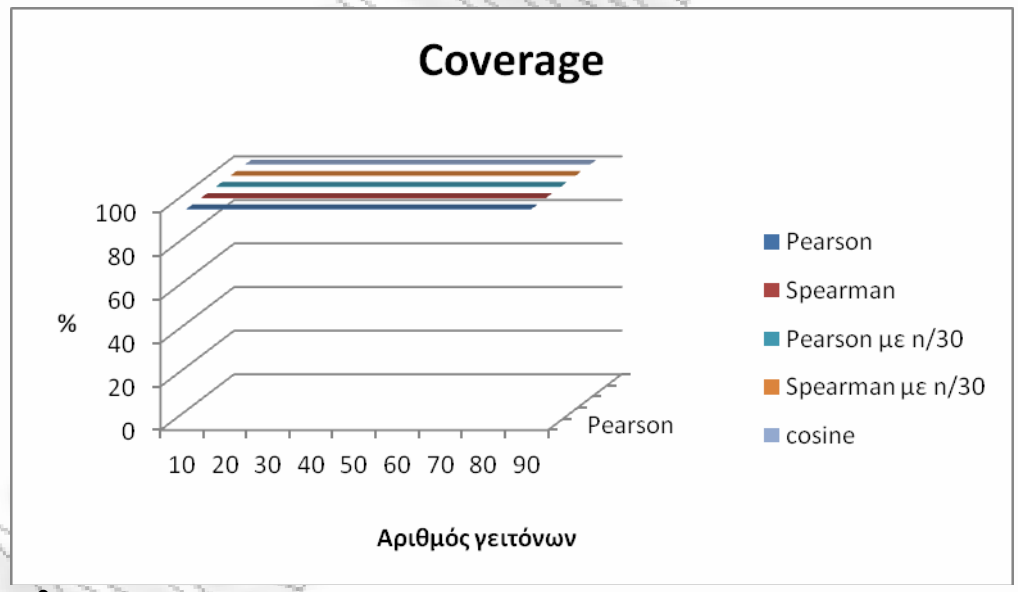
Μέθοδος επιλογής	MAE		Κάλυψη %		ROC sensitivity	
	Pearson με N/20	Pearson	Pearson με N/20	Pearson	Pearson με N/20	Pearson
Όριο συσχέτισης 0,2	0,713405	0,784254	64,9	97,8	0,798489	0,755752
Όριο συσχέτισης 0,25	0,715232	0,756959	60,4	93,4	0,808625	0,772643
Όριο συσχέτισης 0,3	0,745098	0,788235	51	85	0,785942	0,740741
60 επικρατέστεροι	0,766	0,783	100	100	0,752151	0,7401
70 επικρατέστεροι	0,776	0,782	100	100	0,740103	0,73838
80 επικρατέστεροι	0,778	0,789	100	100	0,738382	0,73666
90 επικρατέστεροι	0,777	0,787	100	100	0,738382	0,73666

Πίνακας 7

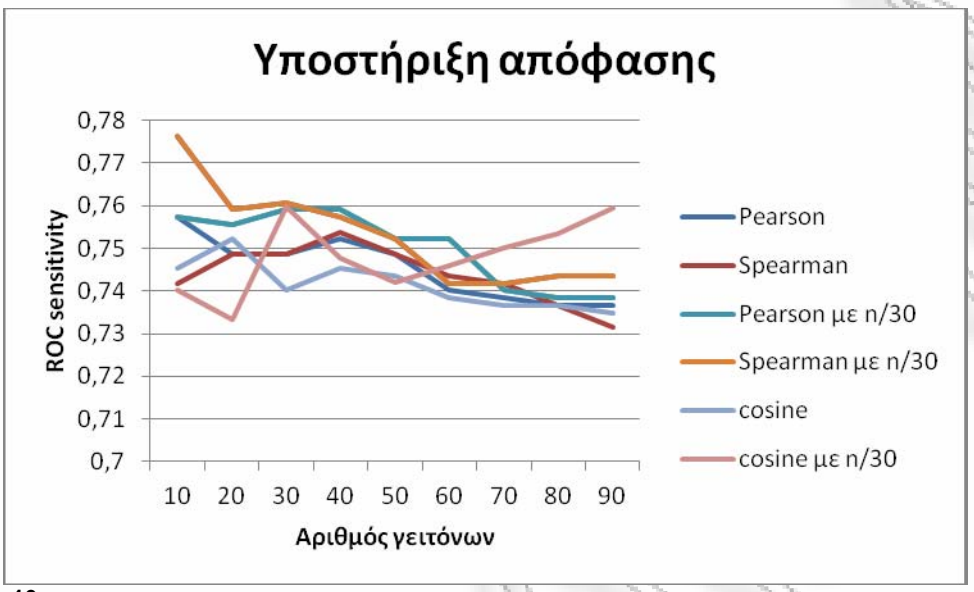


Γράφημα 8

Για να αποφασιστεί για ποια τιμή του Ν στον αλγόριθμο, τα αποτελέσματα βελτιστοποιούνται, αρκεί να παρατηρηθούν τα γραφήματα 8 και 10, όπου φαίνεται πως η χρυσή τομή για την επιλογή Ν επικρατέστερων, φαίνεται να είναι η τιμή 40 για την ακρίβεια τουλάχιστον του συστήματος ανεξαρτήτως μεθόδου. Όσον αφορά την ικανότητα υποστήριξης απόφασης τα αποτελέσματα στο 40 θεωρούνται ικανοποιητικά μιας και αποτελούν την δεύτερη καλύτερη τιμή.

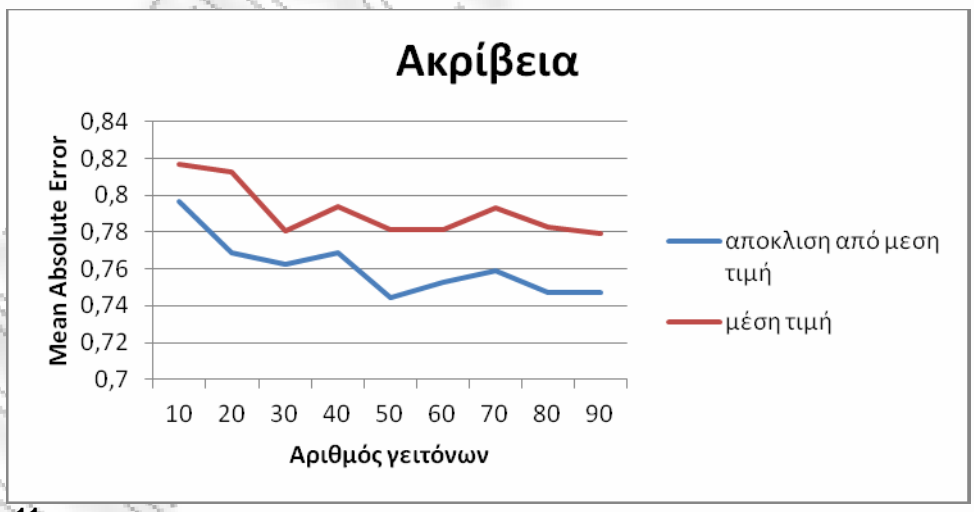


Γράφημα 9



Γράφημα 10

Προχωρώντας στο τελευταίο βήμα και αφού έχει γίνει επιλογή των στοιχείων του συνόλου των γειτόνων, έχει έρθει η στιγμή να γίνει ο συμψηφισμός και να εκτιμηθεί η βαθμολογία πάνω σε κάποιο αντικείμενο. Σε αυτό το βήμα έρχονται αντιμέτωπες δυο τεχνικές. Η πρώτη ουσιαστικά υπολογίζει έναν μέσο όρο των βαθμολογιών των αντικειμένων. Η δεύτερη είναι βασισμένη πάνω στην ιδέα ότι το ίδιο ενδιαφέρον δυο ή και περισσότεροι χρήστες μπορούν να το εκφράσουν με διαφορετική βαθμολογία. Για παράδειγμα πάνω στις κλίμακα της εργασίας, για κάποιους το 4 είναι κορυφή και δεν βάζουν ποτέ 5, ενώ κάποιοι άλλοι αξιολογούν εύκολα με άριστα και ας μην έχουν μεγαλύτερο ενδιαφέρον από του πρώτους. Για να καλυφθεί αυτό το χάσμα που δημιουργείται από την διαφορετικότητα έκφρασης των χρηστών, αρκεί να βρεθεί η απόκλιση από τη μέση τιμή (εξίσωση 17). Σε αυτό το βήμα οι αλγόριθμοι δεν μπορούν να επηρεάσουν την κάλυψη, εφόσον ο πληθυσμός της γειτονιάς ρυθμίζεται μόνο από τα άλλα δυο βήματα. Από το γράφημα 11 φαίνεται και η υπέροχή της δεύτερης τεχνικής σε όλα σχεδόν τα σημεία.



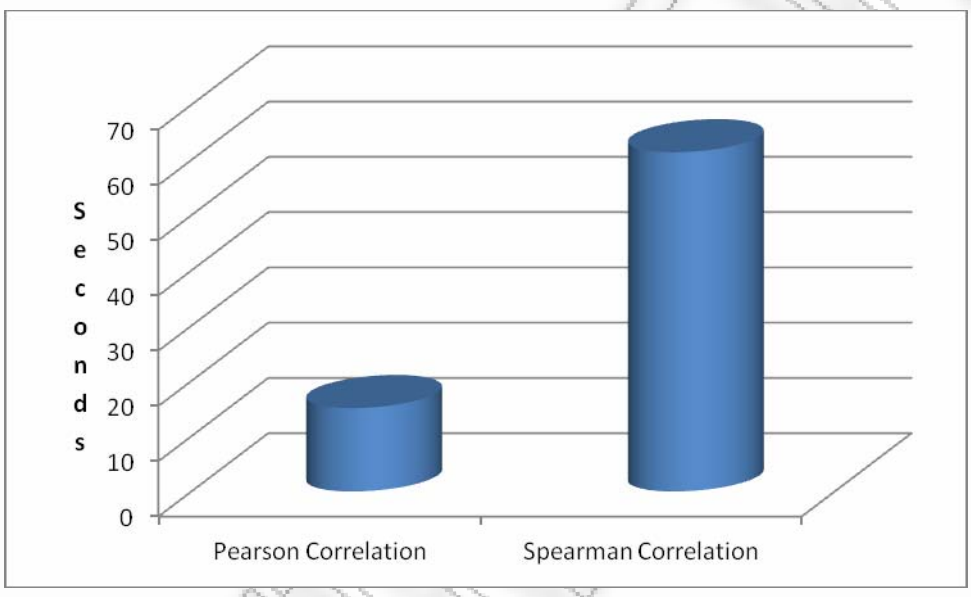
Γράφημα 11

Για το τέλος και για ιστορικούς σκοπούς, παρατίθενται και αποτελέσματα από τη μέθοδο της GroupLens (εξίσωση 15) η οποία παρουσιάζει μέγιστη κάλυψη αφού δεν χρησιμοποιεί κάποιο αλγόριθμο αποκλεισμού χρηστών από τον συνολικό πληθυσμό, ενώ το μέσο σφάλμα ανεβαίνει στα 0,8810.

Λαμβάνοντας υπόψη τα παραπάνω αποτελέσματα, συμπεραίνουμε πως τελικά ο αποδοτικότερος συνδυασμός βάση των μετρήσεων είναι ο συνδυασμός των μεθόδων που φαίνεται παρακάτω στον πίνακα.

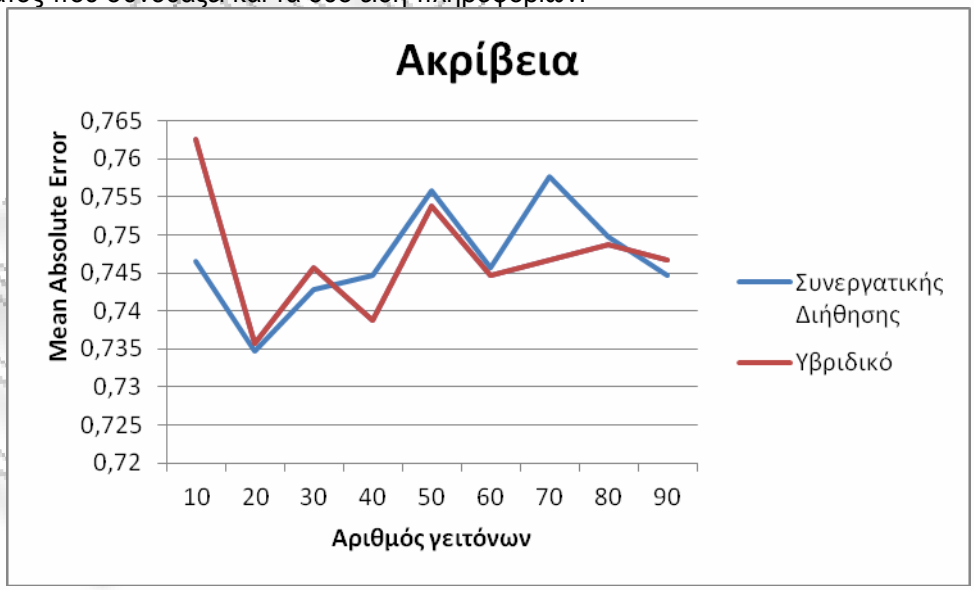
ΔΙΑΔΙΚΑΣΙΑ	ΜΕΘΟΔΟΣ
ΥΠΟΛΟΓΙΣΜΟΣ ΟΜΟΙΟΤΗΤΑΣ	ΣΥΣΧΕΤΙΣΕΙΣ PEARSON με υποβάθμιση χρησιμοποιώντας συντελεστή N/30
ΕΠΙΛΟΓΗ ΓΕΙΤΟΝΩΝ	ΕΠΙΛΟΓΗ **N** ΠΙΟ ΟΜΟΙΩΝ με N=40
ΥΠΟΛΟΓΙΣΜΟΣ ΠΡΟΒΛΕΨΗΣ	ΑΠΟΚΛΙΣΗ ΑΠΟ ΤΗ ΜΕΣΗ ΤΙΜΗ

Πίνακας 8

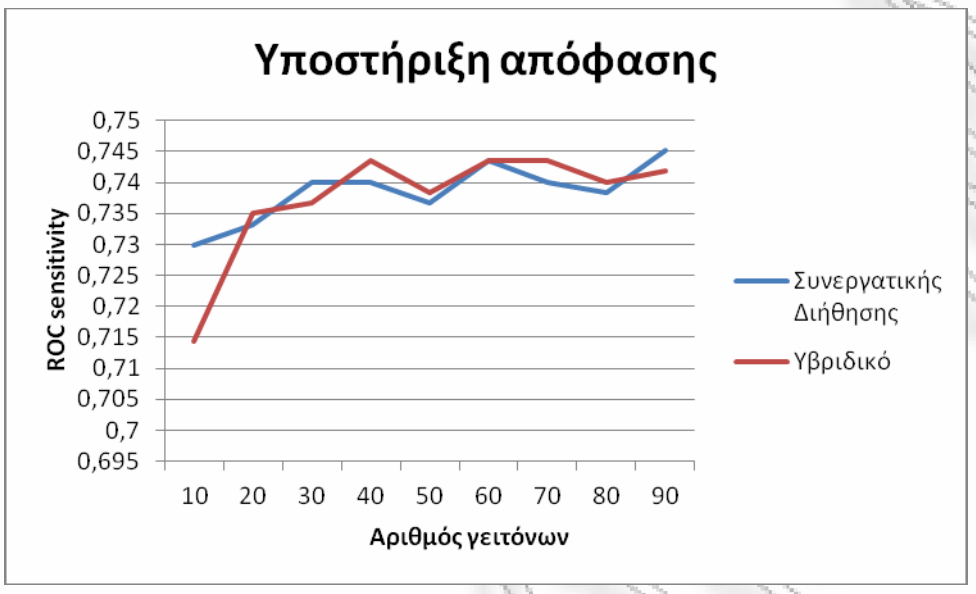


Γράφημα 12 Χρόνος εκτέλεσης συστήματος για 100 χρήστες με 10 εκτιμήσεις για τον καθένα

Κλείνοντας το κεφάλαιο, θα γίνει μια σύγκριση για αναφορικούς σκοπούς με ένα υβριδικό σύστημα [14] που λαμβάνει υπόψη και τα δημογραφικά δεδομένα που υπάρχουν μέσα στο σύνολο δεδομένων της MovieLens. Στην περίπτωση αυτή φαίνεται η ανωτερότητα του συστήματος που συνδυάζει και τα δυο είδη πληροφοριών.



Γράφημα 13 Σύγκριση συστημάτων υβριδικού και απλής συνεργατικής διήθησης με χρήση συσχετισμών Pearson και επιλογή N γειτόνων και υπολογισμό πρόβλεψης με απόκλιση από τη μέση τιμή.


Υποστήριξη απόφασης
ROC sensitivity
0,75
0,745
0,74
0,735
0,73
0,725
0,72
0,715
0,71
0,705
0,7
0,695
10 20 30 40 50 60 70 80 90
Αριθμός γειτόνων
Συνεργατικής Διήθησης
Υβριδικό

ΣΥΜΠΕΡΑΣΜΑΤΑ

Ανακεφαλαίωση

Στην σημερινή κοινωνία στην οποία έχει ζωτική σημασία η πληροφορία και η αναζήτηση της, ερχόμαστε καθημερινά αντιμέτωποι με το πρόβλημα της υπερφόρτωσης πληροφοριών. Ελάχιστος χρόνος και ταυτόχρονα τεράστιοι όγκοι δυναμικών και μη δεδομένων συνθέτουν αυτήν την κατάσταση, που μόνο σύγχυση μπορεί να προκαλέσει, ακόμα και στον έμπειρο χρήστη. Αν αναλογιστούμε ότι ο μέσος χρήστης διαθέτει πολύ λίγο χρόνο για να πραγματοποιήσει κάποια αναζήτηση, τότε κρίνοντας και από το μέγεθος των δεδομένων που θα αντιμετωπίσει, θα πρέπει να βρίσκεται συνεχώς σε σύγχυση. Σε αυτήν την κατάσταση εξειδικευμένο λογισμικό θα μπορούσε να βοηθήσει τον χρήστη σε μεγάλο βαθμό να αποφύγει αυτό το πρόβλημα της υπερφόρτωσης. Επομένως η χρήση της μοντελοποίηση χρηστών με στόχο την παροχή εξατομικευμένων πληροφοριών μέσω συστημάτων σύστασης δείχνει να είναι μονόδρομος.

Στην παρούσα εργασία, εφόσον έγινε αναφορά των διαφορετικών τεχνολογιών των συστημάτων σύστασης, μελετήσαμε τα μέρη και τις τεχνολογίες που διέπουν ένα σύστημα σύστασης συνεργατικής διήθησης και συγκεκριμένα την υποκατηγορία βασισμένη στους χρήστες (user based) της κατηγορίας των βασισμένων στην μνήμη αλγορίθμων (memory based). Τα συστήματα αυτά είναι προϊόντα τριών επιμέρους διαδικασιών. Του υπολογισμού ομοιότητας ανάμεσα στον ενεργό χρήστη και στους υπόλοιπους, με βάση το ιστορικό των αξιολογήσεών τους, την επιλογή των ατόμων που θα αποτελέσουν τη γειτονιά με βάση τα ποσά ομοιότητας που βρέθηκαν στο προηγούμενο βήμα και τέλος την εκτίμηση της πρόβλεψης για κάποιο αντικείμενο, χρησιμοποιώντας τις αξιολογήσεις των γειτόνων για το στοιχείο αυτό.

Στη συνέχεια και εφόσον υλοποιήθηκαν διαφορετικές τεχνικές για την αντιμετώπιση κάθε βήματος και εξάχθηκαν τα αποτελέσματα, περάσαμε στην διαδικασία της σύγκρισης των τεχνικών για την αναζήτηση της υλοποίησης, η οποία καλύπτει τις απαιτήσεις ενός συστήματος σύστασης αποτελεσματικότερα. Οι ανάγκες αυτές είναι, η παροχή όσο το δυνατόν ακριβέστερων εκτιμήσεων, η ικανότητα του συστήματος να εκτιμήσει ένα πολύ μεγάλο ποσοστό από το σύνολο των αντικειμένων και τέλος η παραγωγή των συστάσεων μέσα σε ανεκτά χρονικά περιθώρια. Για τη μέτρηση κατά πόσο κάθε αλγόριθμος καλύπτει τις απαιτήσεις αυτές, επιστρατεύτηκαν οι παρακάτω ποσότητες: το Μέσο Απόλυτο Σφάλμα (Mean Absolute error), το ποσοστό των πραγματοποιημένων ερωτημάτων σε σχέση με το συνολικό αριθμό αιτημάτων σύστασης, η ευαισθησία Λειτουργικού Χαρακτηριστικού Δέκτη (ROC sensitivity) καθώς και ο χρόνος εκτέλεσης κάθε αλγόριθμου.

Βάση των αποτελεσμάτων σε αυτή τη διατριβή, έχουμε φτάσει στο εξής συμπέρασμα: σε, ότι αφορά το πρώτο βήμα (υπολογισμός ομοιοτήτων μεταξύ χρηστών) οι συσχετίσεις Spearman και Pearson χρησιμοποιώντας υποβάθμιση λειτουργούν το ίδιο ικανοποιητικά με αμελητέες διαφορές απόδοσης. Αν όμως λάβουμε υπόψη το χρόνο εκτέλεσης, τότε ο συσχετισμός Pearson είναι ξεκάθαρος νικητής. Στο δεύτερο βήμα η επιλογή γειτόνων γίνεται πιο αποδοτική επιλέγοντας τους 40 πιο όμοιους χρήστες, παρά χρησιμοποιώντας κάποιο όριο συσχέτισης. Στο τελικό βήμα (εκτίμηση πρόβλεψης), η απόκλιση από τη μέση τιμή, αυξάνει αρκετά την ευστοχία του μηχανισμού.

Μελλοντικές επεκτάσεις

Οι αλγόριθμοι της συνεργατικής διήθησης όμως πάσχουν κυρίως από το πρόβλημα της κρύας εκκίνησης, δηλαδή τα μειονεκτήματα του νέου χρήστη και του καινούριου αντικειμένου. Οι μηχανισμοί που παραμένουν στα στενά όρια της συνεργατικής διήθησης, δεν είναι δυνατό να το καταπολεμήσουν. Ο μόνος τρόπος αποφυγής τέτοιων καταστάσεων είναι η χρήση κάποιου συνδυασμού συστήματος συνεργατικής διήθησης, με σύστημα σύστασης κάποιας άλλης τεχνικής ,όπως για παράδειγμα η δημογραφική.

Το μέλλον λοιπόν εστιάζεται , στα υβριδικά συστήματα σύστασης και κυρίως στους συνδυασμούς συστημάτων συνεργατικής διήθησης με τεχνικές βασισμένες στο περιεχόμενο,

υλοποιώντας στρατηγικές διεύρυνσης χαρακτηριστικών (Feature augmentation). Πολλές μελέτες έχουν οδηγηθεί προς αυτή την κατεύθυνση, δίνοντας αρκετά ελπιδοφόρα αποτελέσματα.

ΠΑΡΑΡΤΗΜΑ

Κώδικας Matlab

Σύστημα σύστασης χρησιμοποιώντας Pearson correlation, επιλογή γειτόνων με εφαρμογή ορίου και εκτίμηση βαθμολογίας με μέσο όρο.

```
function recommender(N,W)

global temp;


load all_ratings.mat;
load ratings.mat;
load test_ratings.mat;

%%U user ratings A all ratings B subtotal of ratings
U=B(:,N);
%%O all other users ratings
O=cat(2,A(:,1:N-1),A(:,N+1:943));
%%C table with correlations
C=corr(U,O);
%%Find neibhors given value for limit W
S=find(C>W);

%Unrated movies with hidden rating for test user
Unrated=find(test(:,N));
if (not(isempty(S)))

for i=1:length(Unrated)
  for j=1:length(S)

     T(i,j)=O(Unrated(i),S(j));
  end

  T2(i)=sum(T(i,:))/length(find(T(i,:)));
   T3(i)=test(Unrated(i),N);
end
T2=round(T2);

 temp=cat(2,temp ,[T3;T2]);

else
for i=1:length(Unrated) T3(i)=test(Unrated(i),N); end
T2=NaN(1,10);
temp=cat(2,temp ,[T3;T2]);
end
end
cov=(length(temp(1,:))-length(find(isnan((temp(2,:))))))/length(temp(1,:));
acc=nansum(abs(temp(1,:)-temp(2,:)))/length(find(temp(2,:)>0));
a=find(temp(1,:)>=4);
b=find(temp(2,:)>=4);
ROC=length(intersect(a,b))/length(intersect(find(temp(2,:)>0),a));
```

Σύστημα σύστασης χρησιμοποιώντας Pearson correlation, επιλογή N ομοιότερων γειτόνων και εκτίμηση βαθμολογίας με μέσο όρο.

```
function recommender (N,n)

global temp;

load all_ratings.mat;
load ratings.mat;
load test_ratings.mat;

%%U user ratings A all ratings B subtotal of ratings
U=B(:,N);
%%O all other users
O=cat(2,A(:,1:N-1),A(:,N+1:943));
%%Unrated movies with hidden rating for test user
Unrated=find(test(:,N));

for m=1:length(Unrated) T3(m)=test(Unrated(m),N); end

for i=1:length(Unrated)
%%C table with correlations
x=find(O(Unrated(i),:));
Otemp=O(:,x);
C=corr(U,Otemp);

if (not(isempty(C)))

 C=[C;1:length(C)];
 C=sortrows(C');
 C=C'; S=[];

 for k=length(C(2,:)):-1:(length(C(2,:))-(n-1))
 if (k>0) S=cat(2,S,C(:,k)); end
 end
 neibhors=S(2,:);

T2(i)=sum(Otemp(Unrated(i),neibhors))/length(neibhors);

else T2(i)=NaN; end
end
T2=round(T2);
temp=cat(2,temp ,[T3;T2]);

cov=(length(temp(1,:))-length(find(isnan((temp(2,:))))))/length(temp(1,:));
acc=nansum(abs(temp(1,:)-temp(2,:)))/length(find(temp(2,:)>0));
a=find(temp(1,:)>=4);
b=find(temp(2,:)>=4);
ROC=length(intersect(a,b))/length(intersect(find(temp(2,:)>0),a));
```

Σύστημα σύστασης χρησιμοποιώντας Pearson correlation, επιλογή γειτόνων με εφαρμογή ορίου και εκτίμηση βαθμολογίας με απόκλιση από τη μέση τιμή.

```
function recommender (N,W)
```

```
global temp;

load all_ratings.mat;
load ratings.mat;
load test_ratings.mat;

%%U user ratings A all ratings B subtotal of ratings
U=B(:,N);
%%O all other users
O=cat(2,A(:,1:N-1),A(:,N+1:943));
%%C table with correlations
C=corr(U,O);
%Ua Average of user ratings
Ua=sum(U)/length(find(U));

S=find(C>W);
%%Unrated movies with hidden rating for test user
Unrated=find(test(:,N));
if (not(isempty(S)))

   for j=1:length(S)
      %OUa average ratings for each neighbor
      OUa(j)=sum(O(:,S(j)))/length(find(O(:,S(j))));
  end

for i=1:length(Unrated)
  for j=1:length(S)
      %Or matrix with other user ratings for movies that test user havent rate
      Or(i,j)=O(Unrated(i),S(j));
  end
end

for i=1:length(Unrated)
te=0; counter=0;
for j=1:length(S)
   if (Or(i,j)>0)
       counter=counter+1;
te=te+Or(i,j)-OUa(j);
   end
end
T2(i)=Ua+ te/counter;
 T3(i)=test(Unrated(i),N);
end
else T2(i)=NaN; end
T2=round(T2);

 temp=cat(2,temp ,[T3;T2]
cov=(length(temp(1,:))-length(find(isnan((temp(2,:)))))/length(temp(1,:));
acc=nansum(abs(temp(1,:)-temp(2,:)))/length(find(temp(2,:)>0));
a=find(temp(1,:)>=4);
b=find(temp(2,:)>=4);
ROC=length(intersect(a,b))/length(intersect(find(temp(2,:)>0),a));
```

Σύστημα σύστασης χρησιμοποιώντας Pearson correlation με υποβάθμιση,

επιλογή γειτόνων με εφαρμογή ορίου και εκτίμηση βαθμολογίας από τη μέση τιμή.

```
function recommender(N,W)

global temp;


load all_ratings.mat;
load ratings.mat;
load test_ratings.mat;

%%U user ratings A all ratings B subtotal of ratings
U=B(:,N);
%%O all other users ratings
O=cat(2,A(:,1:N-1),A(:,N+1:943));
%%C table with correlations
C=corr(U,O);
%%Find neibhors given value for limit W
S=find(C>W);

for i=1:942
   tempadd=U+O(:,i);
   tempsub=U-O(:,i);
   Ws(i)=length(find(not(tempsub)))-length(find(not(tempadd)));
   if(Ws(i)<n)
      C(i)=C(i)*Ws(i)/n;
   end
end

%Unrated movies with hidden rating for test user
Unrated=find(test(:,N));
if (not(isempty(S)))

for i=1:length(Unrated)
  for j=1:length(S)

     T(i,j)=O(Unrated(i),S(j));
  end

  T2(i)=sum(T(i,:))/length(find(T(i,:)));
   T3(i)=test(Unrated(i),N);
end
T2=round(T2);

 temp=cat(2,temp ,[T3;T2]);

else
for i=1:length(Unrated) T3(i)=test(Unrated(i),N); end
T2=NaN(1,10);
temp=cat(2,temp ,[T3;T2]);
end
end
cov=(length(temp(1,:))-length(find(isnan((temp(2,:)))))/length(temp(1,:));
acc=nansum(abs(temp(1,:)-temp(2,:)))/length(find(temp(2,:)>0));
a=find(temp(1,:)>=4);
b=find(temp(2,:)>=4);
```

```
ROC=length(intersect(a,b))/length(intersect(find(temp(2,:)>0),a));
```

Υβριδικό σύστημα σύστασης χρησιμοποιώντας Pearson correlation με επιλογή N ομοιότερων χρηστών σαν γειτόνες και εκτίμηση βαθμολογίας με εφαρμογή βαρυτητας βασισμένη στα δημογραφικά χαρακτηριστικά.

```
function hybrid(N,n)

global temp;

load all_ratings.mat;
load ratings.mat;
load test_ratings.mat;
load demo.mat;

%%U user ratings A all ratings B subtotal of ratings
U=B(:,N); Ud=demographic(:,N);
%%O all other users
O=cat(2,A(:,1:N-1),A(:,N+1:943)); Od=cat(2,demographic(:,1:N-1),demographic(:,N+1:943));
%%Unrated movies with hidden rating for test user
Unrated=find(test(:,N));
%Ua Average of user ratings
Ua=sum(U)/length(find(U));

for m=1:length(Unrated) T3(m)=test(Unrated(m),N); end

for i=1:length(Unrated)
%%C table with correlations
x=find(O(Unrated(i),:));
Otemp=O(:,x); Odtemp=Od(:,x);
C=corr(U,Otemp); Cd=corr(Ud,Odtemp);
Cen=C+C.*Cd;
if (not(isempty(C)))

 C=[C;1:length(C)];
 C=sortrows(C');
 C=C'; S=[];

 for k=length(C(2,:)):-1:(length(C(2,:))-(n-1))
 if (k>0) S=cat(2,S,C(:,k)); end
 end

   t=0;
   for l=S(2,:) t=t+1;
     %OUa average ratings for each neighbor
     OUa(t)=sum(O(:,l))/length(find(O(:,l)));
   end

   suma=0; divider=0;
for j=1:length(S(2,:))
   if (Otemp(Unrated(i),S(2,j)))
      divider=divider+abs(Cen(S(2,j)));
suma=suma+(Otemp(Unrated(i),S(2,j))-OUa(j))*Cen(S(2,j));
   end
end
```

```
T2(i)=Ua+ suma/divider;

else T2(i)=NaN; end
end
T2=round(T2);
temp=cat(2,temp ,[T3;T2]);

cov=(length(temp(1,:))-length(find(isnan((temp(2,:)))))/length(temp(1,:));
acc=nansum(abs(temp(1,:)-temp(2,:)))/length(find(temp(2,:)>0));
a=find(temp(1,:)>=4);
b=find(temp(2,:)>=4);
ROC=length(intersect(a,b))/length(intersect(find(temp(2,:)>0),a));
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Heng Luo, Changyong Niu, Ruimin Shen and Carsten Ullrich: “A collaborative filtering framework based on both local user similarity and global user similarity”

[2] Marco de Gemmis, Leo Iaquinta, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro: “Preference Learning in Recommender Systems”

[3] Jonathan L. Herlocker, Joseph A. Kostan, Al Borchers and John Riedl: “An algorithmic framework for performing collaborative filtering”

[4] Gediminas Adomavicius and Alexander Tuzhilin. “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”

[5] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen and Alexander Tuzhilin: “Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach”

[6] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom and John Riedl: “GroupLens: An Open Architecture For Collaborative Filtering Of Netnews”

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl: “Item-Based Collaborative Filtering Recommendation Algorithms”

[8] John s. Breese, David Heckerman, Carl Kadie: “Empirical Analysis Of Predictive Algorithms For Collaborative Filtering

[9] Jae-wook Ahn: Hybrid Web Recommendation Systems

[10] Mustansar Ali Ghazanfar, Adam Prugel-Benne: “A Scalable, Accurate Hybrid Recommender System”

[11] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen and John T. Riedl: “Evaluating Collaborative Filtering Recommender Systems”

[12] Αριστομένης Λαμπρόπουλος: “Μέθοδοι Σύστασης Πολυμεσικών Δεδομένων βασισμένες σε Τεχνικές Μηχανικής Μάθησης – Machine Learning - based Recommendation Methods for Multimedia Data”, Διδακτορική Διατριβή, Πανεπιστήμιο Πειραιώς, 2010.

[13] Χρήστος θ. Νάκας: “Προσαρμογή Καμπύλης, Στατιστική Συμπερασματολογια, Επεκτάσεις Και Εφαρμογές Στην Ανάλυση Των Καμπύλων Λειτουργικού Χαρακτηριστικού Δέκτη (ROC)”, Διδακτορική Διατριβή, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2002.

[14] Manolis Vozalis and Konstantinos G. Margaritis: “Collaborative Filtering Enhanced by Demographic Correlation”

[15] Robin Burke: “Hybrid Recommender Systems: Survey and Experiments”

[16] Michael J. Pazzani: “A Framework for Collaborative, Content-Based and Demographic Filtering”