ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

πRadio: Εξατομικευμένο Σύστημα Σύστασης Ακρόασης Ηλεκτρονικού Ραδιοφώνου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του
ΝΤΕΓΙΑΝΝΑΚΗ ΘΕΟΔΟΣΗ

Επιβλέπων : Νικόλαος Ματσατσίνης
Καθηγητής Μ.Π.Δ.

Χανιά, Ιούνιος 2008

Η εργασία αυτή είναι αποτέλεσμα προσωπικής δουλειάς και ενασχόλησης, όμως δε θα γινόταν πραγματικότητα χωρίς τη υποστήριξη και τη συμβολή των ατόμων του περιβάλλοντος μου τα χρόνια αυτά των σπουδών και όχι μόνο.

Ευχαριστώ την οικογένειά μου και τους φίλους μου

Περιεχόμενα

.....σελ.

Περιεχόμενα 5

Περίληψη 9

Εισαγωγή 11

1.1 Γενικά 11

1.2 Καθορισμός του Προβλήματος 12

1.3 Στόχοι και δομή της εργασίας 13

2 Συστήματα Συστάσεων (Recommender Systems) 15

2.1 Εισαγωγή - Ορισμός 15

2.2 Ταξινόμηση Συστημάτων Συστάσεων 19

2.3 Συστήματα Βασισμένα στο Περιεχόμενου (Content Based) 20

2.3.1 Γενικά 20

2.3.2 Περιορισμοί Συστημάτων Content Based 22

2.3.2.1 Περιορισμοί όσο αφορά την ανάλυση περιεχομένου 22

2.3.2.2 Εξειδίκευση προτάσεων (Overspecialization) 22

2.3.2.3 Πρόβλημα νέου χρήστη 22

2.4 Συστήματα Συνεργασίας (Collaborative Filtering) 23

2.4.1 Γενικά 23

2.4.2 Προσανατολιζόμενα στη Μνήμη (Memory Based) 24

2.4.2.1 Αντικείμενο προς Αντικείμενο (Item to Item) 27

2.4.2.2 Χρήστης προς Χρήστη (User to User) 27

2.4.3 Προσανατολιζόμενα στο Μοντέλο (Model Based) 28

2.4.4 Βελτίωση Αποτελεσμάτων Collaborative Filtering 29

2.4.5 Περιορισμοί Συστημάτων Collaborative Filtering 30
2.4.5.1 Πρόβλημα Αραιών Δεδομένων (Sparsity Problem) 30
2.4.5.2 Πρόβλημα «Ψυχρής» Εκκίνησης (Cold Start Problem) 31
2.4.5.3 Πρόβλημα Αύξησης Κλίμακας (Scalability Problem) 32
2.4.5.4 Πρόβλημα Συνωνύμων (Synonymy) 32
2.5 Υβριδικά Συστήματα (Hybrid Systems) 33
2.5.1 Ταυτόχρονη Εφαρμογή Content Based και Collaborative Filtering 33
2.5.2 Χαρακτηριστικά Content Based σε Collaborative Filtering Systems 33
2.5.3 Χαρακτηριστικά Collaborative Filtering σε Content Based Systems 34
2.5.4 Ενοποιημένο Μοντέλο με Content Based και Collaborative Filtering 34
2.6 Αξιολόγηση Συστημάτων Συστάσεων 35
2.6.1 Μέσο Απόλυτο Σφάλμα (MAE - Mean Absolute Error) 35
2.6.2 Ακρίβεια (Precision) 35
2.6.3 Ανάκληση (Recall) 36

3 Ανάλυση, Σχεδίαση & Ανάπτυξη Συστήματος 39

3.1 Εισαγωγή 39
3.2 Σχεδίαση Ραδιοφώνου 40
3.3 Συλλογή Δεδομένων 42
3.4 Αποθήκευση Δεδομένων 43
3.5 Σύστημα Σύστασης 44
3.5.1 Πρόβλεψη Βαθμολογίας 44
3.5.2 Υπολογισμός Βαρών 45
3.5.3 Αντιμετώπιση Περιορισμών Collaborative Filtering 45
3.5.3.1 Πρόβλημα Νέου Χρήστη 45
3.5.3.2 Πρόβλημα Νέου Αντικειμένου 46
3.5.4 Βελτίωση Αποτελεσμάτων 46
3.5.4.1 Κοντινότεροι N Γείτονες (Top N Neighbors) 46
3.5.4.2 Κατώφλι Βάρους (Weight Threshold) 46
3.5.4.3 Συντελεστής Σπουδαιότητας (Significance Weighting) 47
3.5.5 Πρόταση Μουσικής (Αλγόριθμος) 48
3.6 Ανάλυση Δεδομένων 51

4 Εφαρμογή Συστήματος 57

4.1 Εισαγωγή 57

4.2 Αρχική Σελίδα 57

4.3 Αρχική Σελίδα Χρήστη (Ράδιο) 59

4.4 Σελίδα Αναπαραγωγής 63

5 Συμπεράσματα & Μελλοντικές Ερευνητικές Κατευθύνσεις 67

5.1 Συμπεράσματα 67

5.2 Προτάσεις για Μελλοντική Ανάπτυξη 67

5.2.1 Πρόβλημα Νέου Αντικειμένου 67

5.2.2 Scalability 68

5.2.3 Κατηγοριοποίηση Κομματιών 68

Βιβλιογραφία 71

Περίληψη

Με τη συγγραφή αυτής της διπλωματικής εργασίας επιδιώχθηκαν να καλυφθούν οι εξής τρεις στόχοι.

Αρχικά με βάση τη βιβλιογραφική έρευνα γίνεται μια ιστορική αναδρομή και επεξηγείται ο δρόμος που οδήγησε στην ανάπτυξη του τομέα των συστημάτων συστάσεων (recommender systems) καθώς επίσης αναλύονται περαιτέρω οι διάφορες κατηγορίες των συστημάτων αυτών. Περιγράφονται επίσης τα πλεονεκτήματα και τα μειονεκτήματα των διαφόρων κατηγοριών καθώς επίσης και τεχνικές που αναπτύχθηκαν για την αντιμετώπιση των περιορισμών τους.

Στη συνέχεια ακολουθεί η περιγραφή της υλοποίησης ενός τέτοιου συστήματος συστάσεων σε μια πραγματική εφαρμογή. Προσδιορίζεται ο τομέας που ανήκει η συγκεκριμένη εφαρμογή και τα οφέλη που θα προκύψουν από την υλοποίηση ενός συστήματος συστάσεων σε αυτό τον τομέα. Περιγράφεται αναλυτικά η διαδικασία σχεδίασης της εφαρμογής και υλοποίησης του αλγορίθμου καθώς επίσης αναλύονται τα διάφορα προβλήματα που προκύπτουν και προτείνονται πιθανοί τρόποι αντιμετώπισης.

Τέλος παρουσιάζονται τα συμπεράσματα που προέκυψαν μετά τη συγγραφή της εργασίας καθώς και προτάσεις για περαιτέρω βελτίωση των αποτελεσμάτων μέσω της υλοποίησης νέων ακόμα αποδοτικότερων τεχνικών στην συγκεκριμένη εφαρμογή.

Εισαγωγή

1.1 Γενικά

Ο άνθρωπος είναι ένας πολύπλοκος οργανισμός τον οποίον οι επιστήμες διαρκώς μελετούν προκειμένου να κατανοήσουν τον τρόπο λειτουργίας του. Οι Γνωστικές Επιστήμες (Cognitive Science) ορίζονται ως οι επιστήμες που μελετούν το ανθρώπινο μυαλό και νοημοσύνη. Σε αυτές έχει συνεισφέρει ο κλάδος της ψυχολογίας και συγκεκριμένα η Γνωστική Ψυχολογία (Cognitive Psychology) η οποία προσπαθεί να ερμηνεύσει τα φαινόμενα της αντίληψης και των διανοητικών διεργασιών υποστηρίζοντας ότι έτσι μπορεί να εξηγηθεί η ανθρώπινη συμπεριφορά. Συγκεκριμένα δυο από τους τομείς με τους οποίους ασχολείται η γνωστική ψυχολογία είναι η Θεωρία Επιλογών (Choice Theory) και η Λήψη Αποφάσεων (Decision Making).

Η Θεωρία Επιλογών θέτει την ύπαρξη ενός «Κόσμου Αξιών» (Quality World) μέσα στον οποίο, από τη γέννηση και κατά τη διάρκεια της ζωής του, ο άνθρωπος τοποθετεί όλα αυτά που είναι σημαντικά γι’ αυτόν: άλλους ανθρώπους σημαντικούς γι’ αυτόν, αντικείμενα, καθώς και πεποιθήσεις του.

Στη Λήψη Αποφάσεων η ψυχολογία εξετάζει το σύνολο των πνευματικών διεργασιών (γνωστικών διεργασιών) και τον τρόπο με τον οποίο συντίθενται για να οδηγήσουν στην επιλογή ενός τρόπου δράσης μεταξύ πολλών εναλλακτικών λύσεων.

Καθώς όμως κάθε άνθρωπος μεγαλώνει και διαμορφώνεται μέσα σε ένα ανομοιογενές περιβάλλον, με ως επί των πλείστων διαφορετικά εξωτερικά ερεθίσματα, είναι (προς το παρόν) αδύνατο να μοντελοποιηθεί επακριβώς η ανθρώπινη σκέψη και να αναπτυχθεί μια εξατομικευμένη μέθοδος έτσι ώστε να προβλέπονται πλήρως τα αποτελέσματα των αποφάσεων όπου κάθε άνθρωπος λαμβάνει.

Έτσι παράλληλα αναπτύσσονται και άλλες επιστήμες οι οποίες επιδιώκουν με αλγοριθμικό τρόπο, κάνοντας τις απαραίτητες παραδοχές στις μοντελοποιήσεις τους, (Multi-Criteria Decision Analysis), (Recommender Systems), (Artificial Intelligence) να προβλέψουν με ικανοποιητική ακρίβεια τα αποτελέσματα των ανθρώπινων αποφάσεων.

1.2 Καθορισμός του Προβλήματος

Γιατί όμως επιδιώκουν οι διάφορες επιστήμες όπως αναφέρθηκε παραπάνω να μοντελοποιήσουν και να προβλέψουν τις ανθρώπινες αποφάσεις;

Ο άνθρωπος στη σημερινή εποχή κατακλύζεται από πληροφορίες τις οποίες αδυνατεί να επεξεργαστεί λόγω της πολύ μεγάλης ποικιλίας και του τεράστιου όγκου τους. Σε πολλές περιπτώσεις ο άνθρωπος αναζητά τη λύση ενός προβλήματος ή την επίτευξη ενός στόχου χωρίς όμως να έχει αρκετές γνώσεις και χωρίς να μπορεί να διαχωρίσει τις χρήσιμες, με βάση τα στοιχεία που αναζητά, από τις διαθέσιμες πηγές πληροφοριών.

Για το σκοπό αυτό γεννήθηκε και αναπτύχθηκε ο τομέας των Συστημάτων Συστάσεων (Recommender Systems). Σύμφωνα με τον Burke (2002) τα συστήματα συστάσεων είναι «οποιαδήποτε συστήματα τα οποία παράγουν εξατομικευμένες συστάσεις ως έξοδο ή μπορούν να καθοδηγούν το χρήστη με εξατομικευμένο τρόπο σε ενδιαφέροντα ή χρήσιμα αντικείμενα μέσα σε ένα μεγάλο πλήθος από πιθανές εναλλακτικές». Τα συστήματα συστάσεων αναλαμβάνουν να παρουσιάσουν στο χρήστη τις πληροφορίες εκείνες που είναι πιθανό να τον ενδιαφέρουν. Αυτό γίνεται με τη δημιουργία ενός προφίλ χρήστη το οποίο περιέχει κάποια αντιπροσωπευτικά χαρακτηριστικά για το χρήστη με τα οποία γίνεται η αναζήτηση και το φιλτράρισμα των πληροφοριών. Αναλυτικότερα όμως θα παρουσιαστούν οι τρόποι λειτουργία των μεθόδων σε παρακάτω κεφάλαιο.

Τα συστήματα συστάσεων έχουν αρχίσει να εφαρμόζονται σε διάφορες διαδικτυακές υπηρεσίες και συνεχίζουν να διευρύνουν το πεδίο εφαρμογής τους και σε νέες. Συστήματα συστάσεων θα συναντήσουμε σε ηλεκτρονικά καταστήματα (e-shops) με σκοπό την γρηγορότερη και αποδοτικότερη αναζήτηση προϊόντων από τους πελάτες καθώς επίσης και την προώθηση νέων προϊόντων στους πελάτες που ενδεχομένως να τους ενδιαφέρουν. Επίσης εφαρμογές συναντάμε σε ηλεκτρονικά βιβλιοπωλεία ή βιβλιοθήκες (Amazon, Barnes & Noble), σε ιστοσελίδες για αναζήτηση ταινιών (Movielens, Netflix), σε ηλεκτρονικές εφημερίδες καθώς και σε διάφορες άλλες ιστοσελίδες. Σε όλες αυτές τις εφαρμογές ο εκάστοτε χρήστης-πελάτης έχει πρόσβαση σε ένα μεγάλο όγκο πληροφοριών και τα συστήματα συστάσεων τον βοηθούν να αναζητήσει και τελικά να βρει αυτό που αναζητά πολύ πιο γρήγορα.

Ένας τομέας στον οποίο επίσης έχουν αναπτυχθεί τα συστήματα συστάσεων είναι αυτός της μουσικής. Ο όγκος των διαθέσιμων πληροφοριών σε αυτόν τον τομέα είναι τεράστιος και τα συστήματα αυτά καλούνται να βοηθήσουν το χρήστη να αναζητήσει και να ξεχωρίσει τη μουσική που τον ενδιαφέρει.

1.3 Στόχοι και δομή της εργασίας

Στη διπλωματική εργασία αυτή θα γίνει ανάπτυξη ενός συστήματος συστάσεων σε μια εφαρμογή διαδικτυακού ραδιοφώνου (e-radio). Σκοπός αυτής της εφαρμογής θα είναι η αναπαραγωγή εξατομικευμένου μουσικού προγράμματος σε κάθε ακροατή του ηλεκτρονικού ραδιοφώνου. Η επιλογή των μουσικών κομματιών θα γίνεται με βάση το προφίλ του χρήστη το οποίο θα δημιουργεί δυναμικά η εφαρμογή ανάλογα τις μουσικές προτιμήσεις του. Πιο συγκεκριμένα:

Στο 2^{ο} Κεφάλαιο θα γίνει παρουσίαση του θεωρητικού υποβάθρου της λειτουργίας των συστημάτων συστάσεων (Recommender Systems). Θα παρουσιαστούν οι κύριες κατηγορίες συστημάτων συστάσεων καθώς επίσης θα γίνει περαιτέρω ανάλυση στο σύστημα Collaborative Filtering το οποίο είναι και αυτό που θα χρησιμοποιηθεί κατά την υλοποίηση του ηλεκτρονικού ραδιοφώνου.

Στο 3^{ο} Κεφάλαιο θα γίνει περιγραφή της υφιστάμενης κατάστασης στον επιλεγμένο τομέα εφαρμογής. Στη συνέχεια θα γίνει αναλυτική περιγραφή της σχεδίασης του συστήματος του ηλεκτρονικού ραδιοφώνου καθώς επίσης και ο τρόπος λειτουργίας του αλγορίθμου που χρησιμοποιήθηκε.

Στο 4^{ο} Κεφάλαιο θα γίνει αναλυτική περιγραφή και παρουσίαση του συστήματος και της λειτουργίας τους μέσω της πραγματικής εφαρμογής του.

Τέλος στο 5^{ο} Κεφάλαιο θα αναφερθούν τα συμπεράσματα από την ανάπτυξη του συστήματος και θα γίνουν προτάσεις για μελλοντικές βελτιώσεις και περαιτέρω ανάπτυξη.

2 Συστήματα Συστάσεων (Recommender Systems)

2.1 Εισαγωγή - Ορισμός

Όπως αναφέρθηκε και στην εισαγωγή, η ραγδαία και συνεχόμενη αύξηση των διαθέσιμων πληροφοριών στο διαδίκτυο οδήγησε στην ανάπτυξη μεθόδων για την ευκολότερη και γρηγορότερη πρόσβαση σε αυτές από τους χρήστες του διαδικτύου. Αναπτύχθηκαν λοιπόν συστήματα φιλτραρίσματος πληροφοριών (Information Filtering (Hanani, et al., 2001)) τα οποία είχαν ως σκοπό την αφαίρεση περιττών ή μη επιθυμητών πληροφοριών από μία πηγή, με τη χρήση αυτοματοποιημένων μεθόδων, πριν οι πληροφορίες αυτές φτάσουν στον χρήστη. Ένα χαρακτηριστικό παράδειγμα συστήματος φιλτραρίσματος πληροφοριών είναι τα spam filters τα οποία έχουν ως σκοπό την απομάκρυνση (κακόβουλων και άχρηστων στην περίπτωση αυτή) πληροφοριών από τον χρήστη ηλεκτρονικού ταχυδρομείου.

Τα συστήματα φιλτραρίσματος πληροφοριών διαχωρίζονται σε δυο κύριες κατηγορίες: (Malone, et al., 1987)

1. Τα **Cognitive filtering** συστήματα, τα οποία χαρακτηρίζουν το περιεχόμενο μιας πηγής πληροφοριών, τις ανάγκες σε πληροφορία ενός χρήστη και ευφυώς κάνουν την αντιστοίχιση ανάμεσα στους χρήστες και στις πληροφορίες. Τα συστήματα αυτά βασίζονταν σε φιλτράρισμα κειμένου (text-based filtering) κάνοντας αντιστοιχία στα κείμενα χρησιμοποιώντας κάποιες συγκεκριμένες λέξεις κλειδιά.

2. Τα **Sociological filtering** συστήματα, τα οποία βασίζονται στις προσωπικές και διαθρωτικές σχέσεις μεταξύ των ατόμων μιας κοινότητας.

Ένα εξειδικευμένο σύστημα φιλτραρίσματος πληροφοριών είναι τα συστήματα συστάσεων (Recommender Systems) τα οποία αναλαμβάνουν την αναζήτηση και την παρουσίαση πληροφοριών οι οποίες ενδιαφέρουν το χρήστη. Τα συστήματα συστάσεων παρουσιάστηκαν ως εκπρόσωποι της 2^{ης} κατηγορίας συστημάτων φιλτραρίσματος πληροφοριών τα Sociological filtering συστήματα και επίλυαν 2 βασικά προβλήματα των συστημάτων βασισμένων σε κείμενο (text-based systems):

- Το πρόβλημα του υπέρογκου αριθμού πληροφοριών (Information Overload) οι οποίες περνούσαν το φιλτράρισμα με βάση το κείμενο. Αυτές οι πληροφορίες

τώρα αξιολογούνται περαιτέρω με βάση την ανθρώπινη κρίση όσον αφορά την ποιότητα της πληροφορίας που περιέχουν.

- Το πρόβλημα φιλτραρίσματος πληροφοριών που δεν περιείχαν κείμενο. Τώρα αυτές οι πληροφορίες αξιολογούνται με βάση την ανθρώπινη κρίση και όχι με βάση την περιγραφή τους από κείμενο.

Ο όρος Recommender Systems αναφέρθηκε πρώτη φορά από τους Resmick & Varian (1997) ως «συστήματα όπου οι άνθρωποι παρέχουν προτάσεις ως εισόδους, τις οποίες έπειτα συγκεντρώνουν και κατευθύνουν σε συγκεκριμένους παραλήπτες».

Για να κατανοηθούν καλύτερα τα συστήματα συστάσεων πρέπει να εξεταστούν προσεκτικά οι επιστημονικοί εκείνοι τομείς όπου τα συστήματα συστάσεων βασίστηκαν: (Papagelis, 2005)

1. **Ανάκτηση Πληροφοριών (Information Retrieval)**

   Ως συστήματα ανάκτησης πληροφοριών (information retrieval) αναφέρονται τα συστήματα εκείνα που ταξινομούν, αναζητούν και ανακτούν δεδομένα, είτε αυτά τα δεδομένα έχουν μορφή κειμένου είτε πιο πολύπλοκες μορφές όπως ψηφιακές εικόνες, βίντεο κλπ. Ο χρήστης συχνά επεμβαίνει τροποποιώντας και βελτιστοποιώντας τη λίστα των πληροφοριών που αρχικά αναζητούσε. Χαρακτηριστικό παράδειγμα συστημάτων ανάκτησης πληροφοριών είναι οι μηχανές αναζήτησης του διαδικτύου (web search engines).

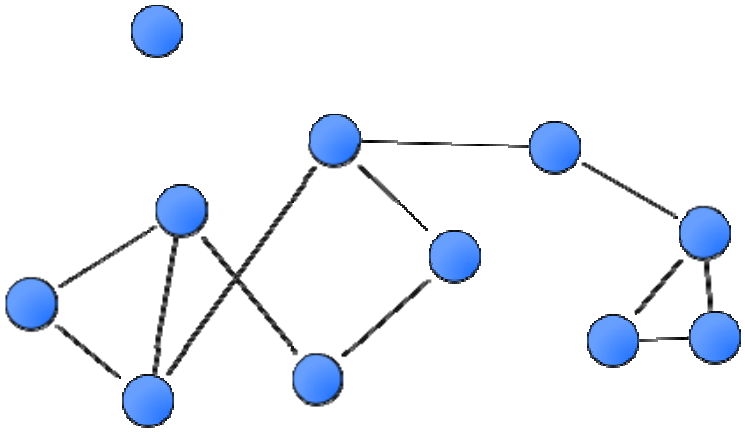
2. **Εξατομίκευση (Personalization)**

   Η εξατομίκευση (Personalization) ή αλλιώς εξατομικευμένο μάρκετινγκ (Personalized Marketing) είναι μια ακραία μορφή διαφοροποίησης ενός προϊόντος. Ενώ η διαφοροποίηση προσπαθεί να δημιουργήσει ένα προϊόν διαφορετικό από τα ανταγωνιστικά, η εξατομίκευση προσπαθεί να δημιουργήσει το «ιδανικό προϊόν» για κάθε χρήστη.
   Κατά την εξατομίκευση συλλέγονται από το σύστημα πληροφορίες για τον χρήστη μέσα από την αλληλεπίδραση μαζί του. Οι πληροφορίες συλλέγονται είτε έμμεσα από ιστορικά στοιχεία ενεργειών του χρήστη είτε άμεσα από την αλληλεπίδραση μαζί του. Σκοπός είναι η βελτίωση των παρεχομένων υπηρεσιών στο χρήστη. Η ικανοποίηση του χρήστη είναι ο πρωταρχικός στόχος της εξατομίκευσης.

3. Κοινωνικά Δίκτυα (Social Networks)

Ως κοινωνικό δίκτυο (Social Network) ορίζεται το δίκτυο εκείνο των σχέσεων μεταξύ των ατόμων, στο οποίο διαχωρίζονται οι τρόποι με τους οποίους συνδέονται. Η ανάλυση των κοινωνικών δικτύων αναδείχτηκε πολύ σημαντική στην σύγχρονη κοινωνιολογία, ανθρωπολογία και στη μελέτη οργανισμών.

Η θεωρία κοινωνικών δικτύων απεικονίζει τις κοινωνικές σχέσεις ως κόμβους και ενώσεις. Κόμβοι είναι τα άτομα μέσα στο κοινωνικό δίκτυο και οι ενώσεις είναι οι σχέσεις μεταξύ των ατόμων. Μπορούν να υπάρξουν πολλών ειδών ενώσεις ανάμεσα στους κόμβους αναλόγως των σχέσεων που μελετώνται. Ένα τυπικό διάγραμμα κοινωνικού δικτύου είναι το παρακάτω όπου οι κόμβοι παρουσιάζονται ως σημεία και οι ενώσεις ως γραμμές:



Εικόνα 2-1: Απεικόνιση Κοινωνικού Δικτύου

Μικρά αποκλεισμένα δίκτυα με σφιχτούς δεσμούς είναι λιγότερο χρήσιμα στα μέλη τους από ότι μεγαλύτερα δίκτυα με πιο χαλαρούς δεσμούς. Τα πιο «ανοικτά» δίκτυα με τους πιο χαλαρούς δεσμούς είναι πιο εύκολο να εισάγουν μέσα τους νέες ιδέες ανάμεσα στα μέλη τους σε σύγκριση με τα πιο κλειστά δίκτυα.

Έτσι τα κοινωνικά δίκτυα παρουσιάζουν μια εναλλακτική θεώρηση των πραγμάτων, όπου τα χαρακτηριστικά του κάθε μέλους παίζουν μικρότερο ρόλο από ότι οι σχέσεις μεταξύ των ατόμων μέσα στο δίκτυο, και έχουν βοηθήσει αρκετά στην αιτιολόγηση διαφόρων φαινομένων του πραγματικού κόσμου.

4. Διαχείριση Εμπιστοσύνης (Trust Management)

Το διαδίκτυο όπως και άλλα μέσα είναι περιβάλλοντα στα οποία συναλλάσσεται πληθώρα πληροφοριών μεταξύ των ατόμων. Οπότε η εμπιστοσύνη στην πηγή είναι σημαντική στο αν θα αποδεχτεί ή όχι μια πληροφορία το άτομο. Οπότε εγκυρότητα προσδίδεται μονάχα σε πληροφορίες που λαμβάνονται από έμπιστες πηγές. Συνεπώς σε μεγάλα δίκτυα έμπιστες

πληροφορίες βασισμένες στις εμπειρίες ατόμων είναι δύσκολα προσπελάσιμες. Η εμπιστοσύνη είναι πολύ σημαντική σε συστήματα διάσπαρτης γνώσης όπως το διαδίκτυο.

Στον τομέα της εμπιστοσύνης έχουν ασχοληθεί η ψυχολογία και η κοινωνιολογία. Τα άτομα προτιμούν να ανταλλάσουν πληροφορίες με φίλους φίλων παρά με άλλα άγνωστα άτομα. Έτσι ένας απλός κανόνας που προκύπτει είναι ότι *εάν ένα άτομο A εμπιστεύεται ένα άτομο B και το άτομο B εμπιστεύεται ένα άτομο C τότε μπορεί να θεωρηθεί ότι το άτομο A εμπιστεύεται το άτομο C*. Έτσι σε ένα δίκτυο μπορεί να υπάρχει αποτελεσματικότητα στη διακίνηση πληροφοριών όταν ανάμεσα σε δυο άτομα υπάρχει τουλάχιστον ένα μονοπάτι εμπιστοσύνης το οποίο να τα ενώνει.

Παρότι, όπως αναφέρθηκε και προηγουμένως, τα συστήματα συστάσεων είχαν ρίζες σε διάφορες επιστήμες όπως οι γνωστικές επιστήμες (Cognitive Science) (Rich, 1979), η προσεγγιστική θεωρία (Approximation Theory), ανάκτηση πληροφοριών (Information Retrieval), κοινωνικά δίκτυα (Social Networks), καθώς επίσης και διάφορες άλλες, εμφανίστηκαν ως ένας νέος επιστημονικός τομέας στα μέσα του 1990 όταν οι επιστήμονες επικεντρώθηκαν σε προβλήματα συστάσεων βασιζόμενα σε ένα σύστημα αξιολογήσεων.

Τα συστήματα συστάσεων γρήγορα εξαπλώθηκαν και εφαρμόστηκαν στον τομέα του ηλεκτρονικού εμπορίου και όχι μόνο. Τέτοια παραδείγματα εφαρμογών είναι το ηλεκτρονικό κατάστημα Amazon (Linden, et al., 2003), το σύστημα πρότασης άρθρων GroupLens (Konstan, et al., 1997), το σύστημα συστάσεων ταινιών Movielens καθώς και το σύστημα πρότασης μουσικής Last.fm.

2.2 Ταξινόμηση Συστημάτων Συστάσεων

Τα συστήματα συστάσεων διαχωρίζονται σε δυο κύριες κατηγορίες ανάλογα με τον τρόπο με τον οποίο γεννώνται οι συστάσεις προς τους χρήστες. Οι δυο αυτές κατηγορίες μαζί με μια τρίτη η οποία είναι ένας υβριδικός συνδυασμός των άλλων δυο είναι οι εξής: (Balabanovic & Shoham, 1997)

- **Συστήματα βασισμένα στο περιεχόμενο (Content Based)**

  Σε αυτά τα συστήματα οι ανάγκες των χρηστών σε πληροφορία και τα χαρακτηριστικά των αντικειμένων αποδίδονται σε μορφή κειμένου (προφίλ). Έπειτα τα συστήματα χρησιμοποιούν αυτές τις αποδόσεις (προφίλ) για να προβλέψουν το ενδιαφέρον του χρήστη για ένα νέο αντικείμενο.

- **Συστήματα συνεργασίας (Collaborative Systems)**

  Τα συστήματα συστάσεων συνεργασίας προβλέπουν το ενδιαφέρον ενός χρήστη σε ένα αντικείμενο σύμφωνα με το ενδιαφέρον που έδειξαν για το συγκεκριμένο αντικείμενο άλλοι χρήστες με παρόμοια ενδιαφέροντα στο παρελθόν.

- **Υβριδικά συστήματα (Hybrid Systems)**

  Είναι τα συστήματα τα οποία χρησιμοποιούν χαρακτηριστικά των δυο παραπάνω συστημάτων με σκοπό να πετύχουν βελτίωση των συστάσεων με περιορισμό των μειονεκτημάτων της κάθε μεθόδου.

Στη συνέχεια του κεφαλαίου θα γίνει μια πιο αναλυτική παρουσίαση των τριών αυτών κατηγοριών.

2.3 Συστήματα Βασισμένα στο Περιεχόμενου (Content Based)

2.3.1 Γενικά

Τα συστήματα συστάσεων βασισμένα στο περιεχόμενο (Content Based Recommender Systems) έχουν τις ρίζες τους στην ανάκτηση πληροφοριών (Information Retrieval) και στα συστήματα φιλτραρίσματος πληροφοριών (Information Filtering) (Belkin & Croft, 1992). Η βελτίωση σε σχέση με τα παραδοσιακά συστήματα ανάκτησης πληροφοριών (Information Retrieval) οφείλεται στη χρήση *προφίλ χρηστών* τα οποία περιέχουν πληροφορίες για τα γούστα, τις προτιμήσεις και τις ανάγκες των χρηστών.

Αυτές οι πληροφορίες των προφίλ συγκεντρώνονται:

- **άμεσα (explicitly)** μέσω της συμπλήρωσης ερωτηματολογίων από τους χρήστες είτε
- **έμμεσα (implicitly)** από τις επιλογές και την συμπεριφορά των χρηστών με την πάροδο του χρόνου.

Τα συστήματα βασισμένα στο περιεχόμενο (Content Based) χρησιμοποιούνται κυρίως όταν πρόκειται για προτάσεις κειμένων όπως ιστοσελίδες, άρθρα, ειδησεογραφία όπως επίσης και σε ανέκδοτα.

Οι συστάσεις γίνονται με βάση το περιεχόμενο των εγγράφων τα οποία συσχετίζονται σε μεγαλύτερο βαθμό με το προφίλ του χρήστη. Ως προφίλ του χρήστη αναφέρεται το ιστορικό των επιλογών που έχει κάνει στο παρελθόν ο χρήστης. Το προφίλ του χρήστη κατασκευάζεται έπειτα όπως και το προφίλ των αντικειμένων που θα παρουσιαστεί παρακάτω.

Για την ευκολότερη κατανόηση του *προφίλ χρήστη* θα παρατεθεί ένα παράδειγμα από το χώρο των συστάσεων ταινιών. Όταν ένας χρήστης αναζητά μια ταινία το σύστημα συγκεντρώνει τις ταινίες τις οποίες ο ίδιος χρήστης είχε αξιολογήσει στο παρελθόν και προσπαθεί να κατανοήσει ένα μηχανισμό συσχέτισης ανάμεσα στις προηγούμενες ταινίες και σε αυτές που είναι υπό αξιολόγηση. Η συσχέτιση αυτή γίνεται με βάση διάφορα χαρακτηριστικά στοιχεία όπως ηθοποιούς, σκηνοθέτες, είδος ταινίας κλπ. Στη συνέχεια μόνο οι ταινίες οι οποίες έχουν υψηλές συσχετίσεις με τις προτιμήσεις του χρήστη επιλέγονται και προτείνονται τελικά στο χρήστη.

Κατά την κατασκευή του *προφίλ των αντικειμένων* (ως αντικείμενα ορίζονται οι διάφορες εναλλακτικές: άρθρα, ταινίες, ιστοσελίδες κλπ) λαμβάνεται υπόψη η συχνότητα εμφάνισης συγκεκριμένων λέξεων οι οποίες ονομάζονται λέξεις-κλειδιά (keywords). Αυτές οι λέξεις κλειδιά στη συνέχεια βαροδοτούνται από μεθόδους που ανήκουν στα συστήματα ανάκτησης πληροφοριών (information retrieval), από τα οποία όπως έχουμε αναφέρει και προηγουμένως έχουν ρίζες τα συστήματα

συστάσεων περιεχομένου. Έτσι δημιουργείται το προφίλ κάθε εναλλακτικής όπου είναι ένα διάνυσμα το οποίο περιέχει τις λέξεις κλειδιά και την αντίστοιχη υπολογισμένη βαροδότηση.

Ένα παράδειγμα δείκτη βαροδότησης είναι ο TF/IDF (Salton G. , 1989) όπου: Υποθέτουμε *N* το συνολικό αριθμό κειμένων που μπορούν να προταθούν στους χρήστες και k_j την λέξη-κλειδί που εμφανίζεται σε n_i από αυτά. Επίσης f_{ij} είναι ο αριθμός που η λέξη-κλειδί k_j εμφανίζεται στο κείμενο d_j. Έτσι TF_{ij} είναι η κανονικοποιημένη συχνότητα εμφάνισης της λέξης-κλειδί k_i στο κείμενο d_j και ορίζεται ως:

$$TF_{ij} = \frac{f_{ij}}{max_z f_{zj}} \tag{2.1}$$

Όπου στον παρανομαστή υπολογίζεται η μέγιστη συχνότητα f_z όλων των λέξεων-κλειδιά k_z που εμφανίζονται στο κείμενο d_j. Όμως λέξεις-κλειδιά οι οποίες εμφανίζονται σε πάρα πολλά κείμενα δεν είναι χρήσιμες για το διαχωρισμό των σχετικών κειμένων. Γι' αυτό το λόγο εισάγεται ένας άλλος δείκτης αντίστροφης συχνότητας IDF_i που χρησιμοποιείται σε συνδυασμό με τον TF_{ij} και ορίζεται ως:

$$IDF_i = log\frac{N}{n_i} \tag{2.2}$$

Έτσι ο δείκτης βάρους *TF-IDF* για τη λέξη κλειδί k_i στο κείμενο d_j ορίζεται ως:

$$w_{ij} = TF_{ij} \times IDF_i \tag{2.3}$$

Έχουμε λοιπόν το διάνυσμα βαρών του χρήστη c και τα διανύσματα βαρών των διαθέσιμων εναλλακτικών s. Για να βρεθεί εάν η κάθε εναλλακτική περιέχει χρήσιμες πληροφορίες για τον χρήστη υπολογίζεται μια συνάρτηση *u(c,s)*. Μια συχνά χρησιμοποιούμενη μέθοδος υπολογισμού του σκορ *u* είναι η συνάρτηση συσχέτισης εφαπτόμενης των διανυσμάτων w_s και w_c.

$$\begin{aligned} u(c,s) = \cos(\overrightarrow{w_c}, \overrightarrow{w_s}) &= \frac{\overrightarrow{w_c} \cdot \overrightarrow{w_s}}{\|\overrightarrow{w_c}\|_2 \times \|\overrightarrow{w_s}\|_2} \\ &= \frac{\sum_{i=1}^{K} w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^{K} w_{i,c}^2}\sqrt{\sum_{i=1}^{K} w_{i,s}^2}} \end{aligned} \tag{2.4}$$

Όπου *K* είναι ο συνολικός αριθμός των λέξεων-κλειδιά στο σύστημα.

2.3.2 Περιορισμοί Συστημάτων Content Based

Τα συστήματα συστάσεων περιεχομένου έχουν όμως και αρκετούς περιορισμούς λόγω του τρόπου λειτουργίας τους: (Adomavicius & Tuzhilin, 2005)

- Περιορισμοί όσο αφορά την ανάλυση περιεχομένου
- Εξειδίκευση προτάσεων (Overspecialization)
- Πρόβλημα νέου χρήστη

2.3.2.1 Περιορισμοί όσο αφορά την ανάλυση περιεχομένου

Το κύριο γνώρισμα των content based συστημάτων είναι η αξιολόγηση των εναλλακτικών με βάση το περιεχόμενό τους. Έτσι προκύπτει πρόβλημα στις περιπτώσεις όπου οι εναλλακτικές δεν περιέχουν κείμενο όπως σε περιπτώσεις πολυμέσων (εικόνων, βίντεο, μουσικής). Σε αυτές τις περιπτώσεις η αξιολόγηση με βάση το περιεχόμενό τους είναι μια αρκετά πολύπλοκη, και όχι εγγυημένη για τα αποτελέσματά της, διαδικασία όταν γίνεται αυτοματοποιημένα με αλγόριθμους (όπως πχ image recognition).
Ένα ακόμα πρόβλημα που προκύπτει από τους περιορισμούς στην ανάλυση του περιεχομένου είναι ότι σε περίπτωση όπου δυο εναλλακτικές περιγράφονται από το ίδιο προφίλ (διάνυσμα βαρών λέξεων-κλειδιά) τότε αυτές οι δυο εναλλακτικές δεν μπορούν να ξεχωρίσουν η μία από την άλλη. Έτσι παραδείγματος χάριν τα συστήματα συστάσεων περιεχομένου δεν μπορούν να διαχωρίσουν ένα καλογραμμένο από ένα κακογραμμένο άρθρο αν και τα δυο χρησιμοποιούν τους ίδιους όρους.

2.3.2.2 Εξειδίκευση προτάσεων (Overspecialization)

Όταν το σύστημα μονάχα προτείνει εναλλακτικές όπου πετυχαίνουν υψηλό δείκτη συσχέτισης με το προφίλ του χρήστη τότε ο αλγόριθμος περιορίζεται στο να προτείνει εναλλακτικές οι οποίες έχουν σχέση με αυτές που ο χρήστης έχει ήδη αξιολογήσει. Έτσι τα αποτελέσματα περιορίζονται μονάχα σε κατηγορίες που ο χρήστης έχει ήδη δει και δεν προτείνουν νέες κατηγορίες που θα μπορούσαν να τον ενδιαφέρουν. Αυτό το φαινόμενο έχει μελετηθεί και έχουν προταθεί μέθοδοι για την επίλυσή του οι οποίες προσθέτουν τυχαιότητα (πχ γενετικοί αλγόριθμοι) στις προτάσεις του συστήματος.

2.3.2.3 Πρόβλημα νέου χρήστη

Ένας νέος χρήστης πρέπει να αξιολογήσει ένα αρκετά μεγάλο πλήθος εναλλακτικών μέχρι τα συστήματα συστάσεων περιεχομένου να δημιουργήσουν το προφίλ του και να προσφέρουν αξιόπιστες προτάσεις. Επομένως οι νέοι χρήστες που έχουν αξιολογήσει λίγες εναλλακτικές έχουν αμφιβόλου αξιοπιστίας προτάσεις από το σύστημα.

2.4 Συστήματα Συνεργασίας (Collaborative Filtering)

2.4.1 Γενικά

Τα συστήματα συστάσεων συνεργασίας (Collaborative Filtering) είναι η δεύτερη κατηγορία συστημάτων συστάσεων. Προτιμούνται από τα content based συστήματα καθώς μειώνουν το φαινόμενο του υπέρογκου αριθμού πληροφοριών (Information Overload) που εξηγήθηκε σε προηγούμενη παράγραφο. Χρησιμοποιούνται επιτυχώς σε διάφορες εφαρμογές στο διαδίκτυο.

Η βασική ιδέα των συστημάτων αυτών είναι ότι οι ίδιοι οι χρήστες προτείνουν και ανταλλάσουν πληροφορία μεταξύ τους. Αυτό που κάνουν τα συστήματα είναι να συλλέγουν τις γνώμες των χρηστών, οι οποίες αναφέρονται ως αξιολογήσεις (ratings), για τις εναλλακτικές. Στη συνέχεια αναζητούνται οι χρήστες όπου οι αξιολογήσεις τους ταυτίζονται (άρα έχουν και κοινά γούστα) δημιουργώντας έτσι «γειτονίες» μεταξύ χρηστών. Έτσι στον κάθε χρήστη το σύστημα προτείνει εναλλακτικές οι οποίες είχαν αξιολογηθεί θετικά και από τους γείτονές του. Εδώ φαίνονται καθαρά οι επιρροές από τα κοινωνικά δίκτυα (Social Networks) και από τη διαχείριση εμπιστοσύνης (Trust Management) όπου είχαν επεξηγηθεί προηγουμένως.

Με αυτό τον τρόπο αντιμετωπίζονται οι διάφοροι περιορισμοί των content based συστημάτων όπου αφορούσαν την ποιότητα και το περιεχόμενο των πληροφοριών: (Herlocker, et al., 1999)

- Μπορούν πλέον να αναλυθούν και να προταθούν εναλλακτικές οι οποίες είναι δύσκολο να αναλυθούν αυτοματοποιημένα, όπως ταινίες, μουσική, ιδέες, άνθρωποι· μιας και τώρα εισάγεται η υποκειμενική κρίση των εναλλακτικών από τους ίδιους τους χρήστες.

- Επίσης τα collaborative filtering συστήματα μπορούν και ποσοτικοποιούν σε μετρήσιμη κλίμακα το πόσο θα ικανοποιεί μια εναλλακτική το προσωπικό γούστο του κάθε χρήστη, κάτι το οποίο είναι πολύ δύσκολο έως αδύνατο να γίνει αυτοματοποιημένα στα συστήματα content based με ανάλυση των πληροφοριών των εναλλακτικών.

- Τέλος στα συστήματα συστάσεων συνεργασίας αρκετές φορές είναι δυνατόν να γίνουν προτάσεις εναλλακτικών στους χρήστες οι οποίες να είναι αρκετά χρήσιμες αλλά χωρίς να έχουν το περιεχόμενο με βάση το οποίο ο χρήστης τις αναζητούσε. Έτσι φτάνουν στο χρήστη πληροφορίες τις οποίες δε θα μπορούσε με διαφορετικό τρόπο να προσεγγίσει.

Και εδώ όπως και στα συστήματα content based οι πληροφορίες (αξιολογήσεις) για το χρήστη συλλέγονται με τους εξής δυο τρόπους:

- **άμεσα (explicitly)** μέσω της αξιολόγησης (rating) της κάθε εναλλακτικής από το χρήστη με χρήση μιας κλίμακας βαθμολόγησης

- **έμμεσα (implicitly)** από τις επιλογές και την συμπεριφορά του κάθε χρήστη μέσα στην εφαρμογή (όπως πχ η αγορά ενός προϊόντος ή η παραμονή του σε μια σελίδα), οι οποίες μεταφράζονται με τον κατάλληλο τρόπο σε βαθμολογίες

Σύμφωνα με τους Breese et al. (1998) τα collaborative filtering συστήματα μπορούν να διαχωριστούν στις εξής δυο κατηγορίες:

- Προσανατολιζόμενα στη Μνήμη (Memory Based)
- Προσανατολιζόμενα στο Μοντέλο (Model Based)

2.4.2 Προσανατολιζόμενα στη Μνήμη (Memory Based)

Βασικός σκοπός των συστημάτων collaborative filtering, όπως έχει ήδη αναφερθεί, είναι η πρόβλεψη αξιολογήσεων για έναν χρήστη (αυτός ο χρήστης αναφέρεται ως ενεργός χρήστης) από το σύνολο των αξιολογήσεων του πληθυσμού των άλλων χρηστών.

Οι αξιολογήσεις-βαθμολογίες όλων των χρηστών σε μια εφαρμογή αποθηκεύονται σε μια βάση δεδομένων. Η βάση δεδομένων αυτή περιέχει τις βαθμολογίες $r_{i,j}$ οι οποίες είναι οι αξιολογήσεις του χρήστη *i* στην εναλλακτική *j*. Έτσι μπορούμε να ορίσουμε τη μέση τιμή των αξιολογήσεων του χρήστη *i* ως:

$$\bar{r}_i = \frac{1}{|I_i|} \sum_{j \in I_i} r_{i,j} \tag{2.5}$$

Όπου I_i είναι το σύνολο των εναλλακτικών όπου ο χρήστης *i* έχει αξιολογήσει.

Για την πρόβλεψη της βαθμολογίας του ενεργού χρήστη θα χρησιμοποιήσουμε πληροφορίες που αφορούν τον ενεργό χρήστη (όπως η μέση τιμή των αξιολογήσεών του) καθώς και ένα σύνολο βαρών για τους υπόλοιπους χρήστες το οποίο το εξάγουμε από την βάση δεδομένων. Η πρόβλεψη για τον ενεργό χρήστη *α* για την εναλλακτική *j* συμβολίζεται ως $p_{\alpha,j}$ και υπολογίζεται ως εξής:

$$p_{a,j} = \kappa \sum_{i=1}^{n} w(a,i)\, r_{i,j} \tag{2.6}$$

$$p_{a,j} = \bar{r}_a + \kappa \sum_{i=1}^{n} w(a,i)\left(r_{i,j} - \bar{r}_i\right) \tag{2.7}$$

Οι δυο εξισώσεις (2.6) και (2.7) είναι δυο διαφορετικές προσεγγίσεις για την πρόβλεψη. Η (2.6) είναι η προσέγγιση *σταθμισμένου μέσου* και η (2.7) *κανονικοποιημένου σταθμισμένου μέσου*. Όπου *n* είναι το σύνολο των χρηστών στη βάση δεδομένων με μη-μηδενικά βάρη και *κ* ένας συντελεστής κανονικοποίησης τέτοιος ώστε το άθροισμα των απόλυτων βαρών να είναι ίσο με τη μονάδα. Για τον υπολογισμό των βαρών έχουν προταθεί διάφορες μέθοδοι (Breese, et al., 1998) (Adomavicius & Tuzhilin, 2005) όπου τα βάρη $w(i,a)$ υποδηλώνουν απόσταση (Distance) είτε συσχέτιση (Correlation), είτε ομοιότητα (Similarity) μεταξύ των χρηστών.

Υπολογισμός Βαρών

Απόσταση (Distance)

Η πιο απλή μέθοδος υπολογισμού των βαρών είναι αυτή της απόστασης. (Adomavicius & Kwon, 2007). Έχουμε λοιπόν:

Απόσταση Manhattan

$$w(a,i) = \sum_{j=0}^{k} \left|r_{a,j} - r_{i,j}\right| \tag{2.8}$$

Ευκλείδεια Απόσταση

$$w(a,i) = \sqrt{\sum_{j=0}^{k} \left|r_{a,j} - r_{i,j}\right|^2} \tag{2.9}$$

Απόσταση Chebyshev

$$w(a,i) = \max_{j=0,\dots,k} \left|r_{a,j} - r_{i,j}\right| \tag{2.10}$$

Συσχέτιση (Correlation)

Αυτός ο τρόπος εμφανίστηκε πρώτη φορά στο πρόγραμμα GroupLens όπου ορίστηκε ως δείκτης βαροδότησης η συσχέτιση κατά Pearson (Pearson Correlation). (Resnick, et al., 1994)

Η συσχέτιση μεταξύ δυο χρηστών *a* και *i* είναι:

$$w(a,i) = \frac{\sum_j (r_{a,j} - \bar{r}_a)(r_{i,j} - \bar{r}_i)}{\sqrt{\sum_j (r_{a,j} - \bar{r}_a)^2 \sum_j (r_{i,j} - \bar{r}_i)^2}} \tag{2.11}$$

Όπου τα αθροίσματα ως προς *j* έχουν να κάνουν με τις εναλλακτικές τις οποίες και οι δυο χρήστες *a* και *i* έχουν αξιολογήσει.

Ομοιότητα Διανυσμάτων (Vector Similarity)

Όπως έχει αναφερθεί και προηγούμενα στα συστήματα ανάκτησης πληροφοριών (Information Retrieval) η συσχέτιση μεταξύ δυο κειμένων υπολογίζεται αν θεωρήσουμε το κάθε κείμενο ως ένα διάνυσμα με πιθανότητες λέξεων και υπολογίσουμε το συνημίτονο της γωνίας των δυο διανυσμάτων (Salton & McGill, 1983).

Έτσι στα συστήματα collaborative filtering αντικαθίστανται τα διανύσματα των πιθανοτήτων των λέξεων με τα διανύσματα των αξιολογήσεων των εναλλακτικών από τους χρήστες. Να σημειωθεί ότι στον συγκεκριμένο αλγόριθμο οι βαθμολογίες παίρνουν θετικές τιμές, αποκλείοντας την ύπαρξη αρνητικών τιμών, και μηδενικές τιμές παίρνουν οι εναλλακτικές που δεν έχουν αξιολογηθεί. Έτσι τα βάρη γίνονται:

$$w(a,i) = \sum_j \frac{r_{a,j}}{\sqrt{\sum_{k \in I_a} r_{a,k}^2}} \frac{r_{i,j}}{\sqrt{\sum_{k \in I_i} r_{i,k}^2}} \tag{2.12}$$

Όπου οι τιμές στους παρονομαστές υπάρχουν για την κανονικοποίηση των βαθμολογιών έτσι ώστε οι χρήστες που έχουν αξιολογήσει περισσότερες εναλλακτικές να μην είναι εκ των προτέρων περισσότερο «όμοιοι» με τους υπόλοιπους χρήστες.

Τα Collaborative Filtering συστήματα, και συγκεκριμένα η memory based κατηγορία συστημάτων, διακρίνεται και αυτή σε δυο κατηγορίες με βάση ποια στοιχεία του πίνακα αξιολογήσεων γίνεται η συσχέτιση. Έτσι έχουμε τις κατηγορίες *Αντικείμενο προς Αντικείμενο (Item to Item)* και *Χρήστης προς Χρήστη (User to User)*

2.4.2.1 Αντικείμενο προς Αντικείμενο (Item to Item)

Σε αυτή την κατηγορία συστημάτων αναζητείται η συσχέτιση μεταξύ των εναλλακτικών. Αν θεωρήσουμε ότι σε ένα πίνακα δεδομένων (αξιολογήσεων) οι χρήστες απεικονίζονται στις γραμμές και οι εναλλακτικές στις στήλες τότε τα διανύσματα στα οποία θα υπολογίσουμε τις συσχετίσεις θα είναι τα διανύσματα των στηλών του πίνακα δεδομένων.

Πίνακας 2-1 Πίνακας Αξιολογήσεων (Αντικείμενο προς Αντικείμενο)

	Εναλ. *1*	Εναλ. *2*	…	Εναλ. *i*	…	Εναλ. *k*	…	Εναλ. *j*
Χρήστης *1*	r_{11}	r_{12}		r_{1i}		r_{1k}		r_{1j}
Χρήστης *2*	r_{21}	r_{22}		r_{2i}		r_{2k}		r_{2j}
…					⟺			
Χρήστης *n*	r_{n1}	r_{n2}		r_{ni}		r_{nk}		r_{nj}

Έστω ότι το σύστημα θέλει να προβλέψει τη βαθμολογία μιας εναλλακτικής για τον ενεργό χρήστη. Τότε υπολογίζει τη συσχέτιση αυτής της εναλλακτικής με τις υπόλοιπες που είχε στο παρελθόν αξιολογήσει και ξεχωρίζει τις *k* πιο όμοιες εναλλακτικές. Στη συνέχεια ο υπολογισμός της βαθμολογίας για την ζητούμενη εναλλακτική γίνεται σταθμίζοντας τις βαθμολογίες των *k* εναλλακτικών με τις συσχετίσεις τους (Sarwar, et al., 2001). Η μέθοδος αυτή υπόσχεται ταχύτερα αποτελέσματα σε μεγάλες βάσεις δεδομένων με ακρίβεια που προσεγγίζει ή και μερικές φορές ξεπερνάει αυτή της μεθόδου χρήστη προς χρήστη.

2.4.2.2 Χρήστης προς Χρήστη (User to User)

Τέλος έχουμε και τα collaborative συστήματα *Χρήστη προς Χρήστη*. Αυτά τα συστήματα τα συναντάμε πιο συχνά στις διάφορες εφαρμογές και στη βιβλιογραφία. Όπως αναφέραμε και προηγουμένως τα συστήματα αυτά προσφέρουν μεγαλύτερη ακρίβεια σε σχέση με τα *Αντικείμενο προς Αντικείμενο* αν και ο χρόνος εκτέλεσής τους είναι συνήθως μεγαλύτερος. Η διαφορά με την κατηγορία *Αντικείμενο προς Αντικείμενο* είναι στον τρόπο που χρησιμοποιούνται τα δεδομένα που παρέχει ο πίνακας των αξιολογήσεων. Στην κατηγορίας *Χρήστης προς Χρήστη* γίνεται συσχέτιση μεταξύ των χρηστών και άρα μεταξύ των γραμμών του πίνακα αξιολογήσεων

Πίνακας 2-2: Πίνακας Αξιολογήσεων (Χρήστης προς Χρήστη)

	Εναλ. *1*	Εναλ. *2*	…	Εναλ. *j*
Χρήστης *1*	r_{11}	r_{12}		r_{1j}
Χρήστης *2*	r_{21}	r_{22}		r_{2j}
…				
Χρήστης *i*	r_{i1}	r_{i2}		r_{ij}
…			⇕	
Χρήστης *k*	r_{k1}	r_{k2}		r_{kj}
…				
Χρήστης *n*	r_{n1}	r_{n2}		r_{nj}

Οι μέθοδοι υπολογισμού των συσχετίσεων και στις δυο αυτές κατηγορίες είναι ίδιες με αυτές που αναλύσαμε στην προηγούμενη παράγραφο.

2.4.3 Προσανατολιζόμενα στο Μοντέλο (Model Based)

Τα προσανατολιζόμενα στο μοντέλο συστήματα συστάσεων (Model Based Collaborative Filtering Systems) βασίζονται στη θεωρία πιθανοτήτων για να προβλέψουν τη βαθμολογία του χρήστη. Για παράδειγμα ένας τρόπος υπολογισμού των άγνωστων βαθμολογιών με πιθανότητες είναι:

$$r_{c,s} = \sum_{i=o}^{n} i \times \Pr\left(r_{c,s} = i | r_{c,s'}, s' \in S_C\right) \tag{2.13}$$

Υποθέτουμε ότι τα βάρη παίρνουν τιμές από 0 ως *n* και η πιθανότητα που υπολογίζεται είναι η πιθανότητα ο χρήστης *c* να αξιολογήσει την εναλλακτική *s* με τη συγκεκριμένη βαθμολογία *i* με δεδομένα τις προηγούμενες αξιολογημένες εναλλακτικές.

Για τον υπολογισμό της παραπάνω πιθανότητας έχουν αναπτυχθεί διάφορες μέθοδοι πρόβλεψης όπως τα δίκτυα Bayes (Chen & George, 2000), η ομαδοποίηση (Clustering) (Ungar & Foster, 1998), η γραμμική παλινδρόμηση (linear regression) (Sarwar, et al., 2001), μοντέλα μέγιστης εντροπίας (Maximum Entropy Model) (Pavlov & Pennock, 2002) καθώς και τεχνικές από τη μηχανική γνώσεων (Machine Learning) και τεχνητή νοημοσύνη (Artificial Intelligence).

2.4.4 Βελτίωση Αποτελεσμάτων Collaborative Filtering

Κατά καιρούς έχουν προταθεί διάφοροι τρόποι βελτίωσης των αποτελεσμάτων των συστημάτων collaborative filtering. Όπως για παράδειγμα οι μέθοδοι *default voting, inverse user frequency, case amplification* (Breese, et al., 1998) και *weighted - majority prediction* (Nakamura & Abe, 1998).

Default Voting

Η προεπιλεγμένη βαθμολογία (Default voting) είναι μια μέθοδος βελτίωσης των αποτελεσμάτων σε περιπτώσεις όπου υπάρχουν σχετικά λίγες αξιολογήσεις, είτε για τον ενεργό χρήστη είτε για τον χρήστη όπου υπολογίζεται η συσχέτιση, καθώς για τον υπολογισμό χρειάζονται βαθμολογίες **και** από τους δυο χρήστες. Με την συγκεκριμένη μέθοδο εισάγεται μια προεπιλεγμένη βαθμολογία για τις εναλλακτικές εκείνες όπου δεν υπάρχει άμεση αξιολόγηση από το χρήστη. Έτσι καταλήγουμε να έχουμε αντί για την τομή των συνόλων των αξιολογημένων εναλλακτικών του κάθε χρήστη ($I_a \cap I_i$), ένα μεγαλύτερο σύνολο που μπορεί να φτάσει μέχρι και την ένωσή τους ($I_a \cup I_i$) και άρα περισσότερα δεδομένα που θα βελτιώσουν (κατά περίπτωση) τις προβλέψεις.

Inverse User Frequency

Όπως είχαμε πει και προηγουμένως στα συστήματα content based κατά τον υπολογισμό του δείκτη βάρους TF-IDF ο ρόλος του όρου IDF ήταν να μειώνει το βάρος σε περιπτώσεις όπου οι λέξεις κλειδιά ήταν κοινές δηλαδή είχαν μεγάλη συχνότητα σε όλα τα κείμενα και άρα δεν προσέφεραν στην ανάλυση τους. Αντίστοιχα και στην μεθοδολογία inverse user frequency η ιδέα είναι ότι οι εναλλακτικές οι οποίες βαθμολογούνται ψηλά από την πλειοψηφία των χρηστών δεν μας προσφέρουν αρκετή πληροφορία για τον υπολογισμό της συσχέτισης μεταξύ των χρηστών. Οπότε ορίζουμε ένα συντελεστή για την κάθε εναλλακτική *j*

$$f_j = log\frac{n}{n_j} \tag{2.14}$$

Όπου n_j είναι ο αριθμός των χρηστών που ψήφισαν για την εναλλακτική *j* και *n* είναι ο συνολικός αριθμός χρηστών στο σύστημα. Επομένως αν όλοι οι χρήστες έχουν ψηφίσει μια εναλλακτική τότε ο συντελεστής f_j θα είναι μηδέν.

Case Amplification

Η μέθοδος αυτή ενισχύει τα βάρη που είναι κοντά στη μονάδα και μειώνει τα βάρη που είναι κοντά στο μηδέν. Με τον τρόπο αυτό (κατά περίπτωση) έχουμε βελτίωση των αποτελεσμάτων των προβλέψεων.

2.4.5 Περιορισμοί Συστημάτων Collaborative Filtering

Τα collaborative filtering συστήματα όπως αναφέραμε δεν έχουν τους περιορισμούς που έχουν τα content based συστήματα. Όμως με τη σειρά τους έχουν και αυτά άλλων ειδών περιορισμούς (Balabanovic & Shoham, 1997) (Sarwar, et al., 2000). Οι σημαντικότεροι περιορισμοί είναι οι εξής:

- Πρόβλημα Αραιών Δεδομένων (Sparsity Problem)
- Πρόβλημα «Ψυχρής» Εκκίνησης (Cold Start Problem)
- Πρόβλημα Αύξησης Κλίμακας (Scalability Problem)
- Πρόβλημα Συνωνύμων (Synonymy Problem)

2.4.5.1 Πρόβλημα Αραιών Δεδομένων (Sparsity Problem)

Ένα συχνό φαινόμενο σε εφαρμογές collaborative filtering είναι η ύπαρξη πολύ μεγάλου αριθμού εναλλακτικών, τέτοιου ώστε ακόμα και οι χρήστες που κάνουν αρκετές αξιολογήσεις απέχουν πάρα πολύ ακόμα και από το 1% του πλήθους των εναλλακτικών. Για παράδειγμα σε μια βάση δεδομένων 1 εκατομμυρίου βιβλίων το 1% είναι 10.000 αξιολογήσεις από ένα χρήστη! Σε αυτές τις περιπτώσεις το σύστημα collaborative filtering δε μπορεί να βρει αρκετούς γειτονικούς χρήστες με αποτέλεσμα να αδυνατεί να δώσει αξιόπιστες ή και καθόλου προβλέψεις. Το πρόβλημα αυτό είναι γνωστό και ως πρόβλημα χαμηλής κάλυψης (reduced coverage).

Ένας τρόπος αντιμετώπισης του συγκεκριμένου προβλήματος που έχει προταθεί (Pazzani, 1999) είναι η χρήση περισσοτέρων στοιχείων των χρηστών για τον υπολογισμό των συσχετίσεών τους. Τέτοια στοιχεία είναι το φύλο, η ηλικία, η καταγωγή, η βαθμίδα της εκπαίδευσής του και άλλα δημογραφικά στοιχεία. Τα συστήματα αυτά συνήθως ονομάζονται και ως δημογραφικά συστήματα (Demographic filtering).

Έχουν επίσης προταθεί και άλλα συστήματα τα οποία χρησιμοποιούν αλγορίθμους όπως τον SVD (Singular Value Decomposition) (Sarwar, et al., 2000). Ο αλγόριθμος SVD αρχικά υπολογίζει τις αφανείς σχέσεις μεταξύ χρηστών και εναλλακτικών που βοηθούν στον υπολογισμό των προτιμήσεων των χρηστών στις εναλλακτικές. Έπειτα μετατρέπει το μεγάλων διαστάσεων πίνακα δεδομένων σε έναν μικρότερο από τον οποίο μπορούν να υπολογιστούν οι γειτονίες μεταξύ των χρηστών καθώς και να γίνουν συστάσεις.

2.4.5.2 Πρόβλημα «Ψυχρής» Εκκίνησης (Cold Start Problem)

Το πρόβλημα «κρύας» εκκίνησης (Cold Start) (Schein, et al., 2002) συνδέεται άμεσα με το πρόβλημα αραιών δεδομένων (Sparsity). Το πρόβλημα αυτό εκδηλώνεται σε καινούριους χρήστες και σε νέες εναλλακτικές.

Πρόβλημα Νέου Χρήστη (New User Problem)

Κατά την είσοδο ενός νέου χρήστη στο σύστημα, για να γίνουν αξιόπιστες προτάσεις σε αυτόν, πρέπει να κατανοηθούν οι προτιμήσεις και τα γούστα του, δηλαδή να δημιουργηθεί το προφίλ του. Μέχρι ο χρήστης να κάνει έναν ικανό αριθμό (ανάλογα τον όγκο των δεδομένων του συστήματος) αξιολογήσεων οι προτάσεις που θα γίνονται σε αυτόν πιθανότατα να στερούνται αξιοπιστίας.

Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες προτάσεις. Μια πρόταση είναι τα demographic filtering συστήματα που εξηγήθηκαν παραπάνω ή και άλλα συστήματα που συνδυάζουν content based και collaborative filtering τεχνικές.

Τέλος έχουν προταθεί από τους (Rashid, et al., 2002) διάφορες τεχνικές υπολογισμού βέλτιστων εναλλακτικών, όχι με την έννοια της προβλεπόμενης βαθμολογίας για τον ενεργό χρήστη αλλά, με την έννοια της πληροφορίας που θα προσφέρουν για τη γρηγορότερη διαμόρφωση του προφίλ του χρήστη. Οι τεχνικές αυτές βασίζονται στην δημοτικότητα των εναλλακτικών (πληθώρα αξιολογήσεων), στην εντροπία (μεγάλη διασπορά στις αξιολογήσεις των εναλλακτικών από τους χρήστες) και άλλες.

Πρόβλημα Νέου Αντικειμένου (New Item Problem)

Σε πρακτικές εφαρμογές recommender συστημάτων η προσθήκη νέων εναλλακτικών (νέα βιβλία, νέες ταινίες κλπ) είναι κάτι το οποίο συμβαίνει συχνά. Όμως σύμφωνα με τη θεωρία των collaborative filtering συστημάτων για να υπάρξει πρόβλεψη της βαθμολογίας μιας εναλλακτικής θα πρέπει να έχει βαθμολογηθεί τουλάχιστον από ένα χρήστη. Οπότε οι νέες εναλλακτικές δε μπορούν να προταθούν από το σύστημα.

Τρόπος αντιμετώπισης αυτού του προβλήματος είναι κυρίως ο συνδυασμός content based και collaborative filtering συστημάτων.

2.4.5.3 Πρόβλημα Αύξησης Κλίμακας (Scalability Problem)

Τα συστήματα collaborative filtering μπορεί να είναι αποτελεσματικά στην πρόταση εναλλακτικών αλλά απαιτούν υπολογισμούς οι οποίοι κοστίζουν σε υπολογιστική δύναμη και οι οποίοι αυξάνουν μη-γραμμικά με την αύξηση του όγκου των δεδομένων του πίνακα των αξιολογήσεων. Επομένως προκύπτει το πρόβλημα παροχής προτάσεων στο διαδίκτυο σε αποδεκτούς, από το χρήστη, χρόνους. Γι' αυτό απαιτούνται εξελιγμένες δομές δεδομένων και αρχιτεκτονικές που επιδέχονται μεγάλης κλίμακας μεγέθη.

Και γι' αυτά τα προβλήματα έχουν προταθεί λύσεις όπως τα δίκτυα Bayes (Bayesian Networks), μοντέλα ομαδοποίησης (Clustering) (Breese, et al., 1998) καθώς και η μέθοδος SVD (Singular Value Decomposition) (Sarwar, et al., 2000) για τη μείωση των διαστάσεων του πίνακα δεδομένων.

Δυστυχώς όμως μπορεί αυτές οι τεχνικές να βελτιώνουν κατά πολύ την ταχύτητα εξαγωγής προτάσεων αλλά μειώνουν την ποιότητα και την αξιοπιστία των αποτελεσμάτων τους. Οι μέθοδοι ομαδοποίησης (Clustering) πάσχουν από φτωχή ακρίβεια αποτελεσμάτων, ενώ τα δίκτυα Bayes μπορούν να εφαρμοστούν αποτελεσματικά μόνο σε περιβάλλοντα όπου οι προτιμήσεις των χρηστών μεταβάλλονται με ρυθμούς που μπορεί να ακολουθήσει. Ενώ τέλος τα συστήματα SVD μειώνουν την ακρίβεια των αποτελεσμάτων μιας και με την μείωση των διαστάσεων του πίνακα δεδομένων χάνεται πληροφορία.

Οπότε σε κάθε υλοποίηση των συστημάτων αυτών, ανάλογα με τον όγκο των εναλλακτικών και των χρηστών που προβλέπεται να εξυπηρετήσει, επιλέγεται και η ανάλογη μέθοδος η οποία πετυχαίνει τη χρυσή τομή αξιοπιστίας των αποτελεσμάτων αλλά και ταχύτητας εκτέλεσης για τη συγκεκριμένη βάση δεδομένων. Η επιλογή του βέλτιστης τεχνικής γίνεται συνήθως με δοκιμές και αναλύσεις των αποτελεσμάτων πάνω στην ίδια βάση δεδομένων η οποία ονομάζεται dataset.

2.4.5.4 Πρόβλημα Συνωνύμων (Synonymy)

Πολλές φορές τα ονόματα των εναλλακτικών μπορούν να αναφέρονται σε όμοια αντικείμενα. Τα συστήματα collaborative filtering αδυνατούν να εντοπίσουν αυτή την σχέση και αντιμετωπίζουν όλες τις εναλλακτικές ως διαφορετικές. Έτσι δε μπορούν να εντάξουν τις εναλλακτικές που είναι παρόμοιες σε κάποια κατηγορία, γεγονός που θα βοηθούσε στην διαδικασία των συστάσεων.

2.5 Υβριδικά Συστήματα (Hybrid Systems)

Παρατηρώντας ξεχωριστά τις δυο προηγούμενες κατηγορίες συστημάτων συστάσεων (Content Based και Collaborative Filtering) διαπιστώνουμε ότι κάθε μια έχει κάποιους περιορισμούς τους οποίους οι μέθοδοι της άλλης κατηγορίας μπορούν και επιλύουν πιο αποτελεσματικά. Όμως υπάρχει και μια τρίτη κατηγορία συστημάτων τα οποία δεν ανήκουν σε καμία από τις προηγούμενες κατηγορίες. Αυτά τα συστήματα είναι τα Υβριδικά Συστήματα (Hybrid Systems). Ονομάζονται υβριδικά καθώς ενσωματώνουν μεθόδους και από τις δυο παραπάνω κατηγορίες με σκοπό την αντιμετώπιση των διαφόρων περιορισμών που προκύπτουν. Τα υβριδικά συστήματα χωρίζονται σε 4 κύριες κατηγορίες συστημάτων. (Adomavicius & Tuzhilin, 2005)

- Ταυτόχρονη Εφαρμογή Content Based και Collaborative Filtering
- Χαρακτηριστικά Content Based σε Collaborative Filtering Systems
- Χαρακτηριστικά Collaborative Filtering σε Content Based Systems
- Ενοποιημένο Μοντέλο με Content Based και Collaborative Filtering

2.5.1 Ταυτόχρονη Εφαρμογή Content Based και Collaborative Filtering

Ένα είδος υβριδικών συστημάτων συστάσεων είναι αυτό που ενσωματώνει ξεχωριστά και content based και collaborative συστήματα τα οποία εκτελούνται ταυτόχρονα. Το σύστημα έπειτα μπορεί να αξιοποιήσει τα αποτελέσματα με δυο διαφορετικούς τρόπους. Μπορεί να συνδυάσει τις βαθμολογίες των δυο συστημάτων με γραμμικό τρόπο βγάζοντας μια τελική βαθμολογία (Claypool, et al., 1999). Είτε μπορεί να επιλέγει ανά περίπτωση όποια βαθμολογία από τις δυο είναι περισσότερο κατάλληλη με βάση ένα σύστημα αξιολόγησης των αποτελεσμάτων (Billsus & Pazzani, 2000) (Tran & Cohen, 2000).

2.5.2 Χαρακτηριστικά Content Based σε Collaborative Filtering Systems

Αρκετά υβριδικά συστήματα όπως το Fab (Balabanovic & Shoham, 1997) βασίζονται σε τεχνικές collaborative filtering αλλά ταυτόχρονα χρησιμοποιούν τα content based προφίλ των χρηστών τους. Για τον υπολογισμό της συσχέτισης μεταξύ των χρηστών χρησιμοποιούν τα content based προφίλ τους και όχι το σύνολο των προηγούμενων αξιολογήσεών τους. Με αυτό τον τρόπο ξεπερνούν το πρόβλημα αραιών δεδομένων (Sparsity) όπως αναφέραμε και προηγούμενα (Pazzani, 1999). Ένα ακόμα χαρακτηριστικό τους είναι ότι σε ένα χρήστη δεν προτείνονται μόνο οι εναλλακτικές που λαμβάνουν την υψηλότερη βαθμολογία από χρήστες με παρόμοια προφίλ, αλλά προτείνονται και αυτές οι εναλλακτικές που έχουν υψηλή συσχέτιση με το ίδιο το προφίλ του ενεργού χρήστη.

2.5.3 Χαρακτηριστικά Collaborative Filtering σε Content Based Systems

Το συστήματα αυτό δεν είναι συναντάται τόσο συχνά όσο το προηγούμενο. Στην πιο γνωστή του προσέγγιση εφαρμόζονται τεχνικές μείωσης των διαστάσεων μιας ομάδας content based προφίλ χρηστών. Χρησιμοποιούν τη μέθοδο latent semantic indexing (LSI) (Soboroff & Nicholas, 1999) για να βελτιώσουν τα αποτελέσματα των content based συστημάτων.

2.5.4 Ενοποιημένο Μοντέλο με Content Based και Collaborative Filtering

Τέλος, αυτή η τεχνική έχει αρχίσει να ακολουθείται από πολλούς ερευνητές τα τελευταία χρόνια. Οι πλειοψηφία των εφαρμογών κατασκευάζει ένα ενιαίο μοντέλο στο οποίο συμπεριλαμβάνονται τα στοιχεία των διαφόρων μοντέλων, που παρουσιάστηκαν σε αυτό το κεφάλαιο, με σκοπό να ελαττώσουν τα αρνητικά που κάθε μέθοδος συνεπάγεται. Σε αυτή την κατηγορία ανήκουν και συστήματα που χρησιμοποιούν τεχνικές διαχείρισης της γνώσης (knowledge-based techniques). Το μόνο μειονέκτημα των τεχνικών διαχείρισης της γνώσης είναι ότι στερούνται μηχανισμούς απόκτησης γνώσης, κύριο πρόβλημα πολλών συστημάτων τεχνητής νοημοσύνης. Οπότε χρησιμοποιούνται κυρίως σε τομείς όπου η γνώση υπάρχει ήδη σε μορφή κατανοητή από τα συστήματα αυτά.

Πολλές δημοσιεύσεις έχουν ασχοληθεί με τη σύγκριση των content based, των collaborative filtering και των hybrid συστημάτων δοκιμάζοντας τα μοντέλα σε διάφορες πραγματικές εφαρμογές (Balabanovic & Shoham, 1997) (Pazzani, 1999) (Soboroff & Nicholas, 1999). Σε αυτά αποδεικνύεται ότι τα υβριδικά συστήματα έχουν κατά κύριο λόγο πιο αξιόπιστες προτάσεις από ότι τα απλά συστήματα.

2.6 Αξιολόγηση Συστημάτων Συστάσεων

Πολλές φορές σε αυτό το Κεφάλαιο έγινε λόγος για την βελτίωση των αποτελεσμάτων των συστημάτων συστάσεων. Όμως δεν έχουμε αναφέρει τον τρόπο με τον οποίο αξιολογείται ένα σύστημα και τα αποτελέσματα τα οποία παράγει. Έχουν προταθεί διάφορο τρόποι αξιολόγησης και οι σημαντικότεροι είναι οι παρακάτω: (Herlocker, et al., 2004)

- Μέσο Απόλυτο Σφάλμα (MAE - Mean Absolute Error)
- Ακρίβεια (Precision)
- Ανάκληση (Recall)

2.6.1 Μέσο Απόλυτο Σφάλμα (MAE - Mean Absolute Error)

Ο δείκτης MAE (μέσου απόλυτου σφάλματος) (Breese, et al., 1998) (Herlocker, et al., 1999) μετράει την απόλυτη διαφορά της προτεινόμενης βαθμολογίας από την πραγματική βαθμολογία που έχει εισάγει ο χρήστης. Και στη συνέχεια σταθμίζει αυτές τις διαφορές. Ο υπολογισμός του MAE γίνεται με τον παρακάτω τύπο:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N} \tag{2.15}$$

Όπου p_i είναι η προτεινόμενη βαθμολογία για την εναλλακτική i, r_i είναι η πραγματική τιμή της βαθμολογίας της εναλλακτικής που έδωσε ο χρήστης, και N το πλήθος όλων των εναλλακτικών.

2.6.2 Ακρίβεια (Precision)

Ο δείκτης της ακρίβειας σε αντίθεση με το δείκτη του μέσου απόλυτου σφάλματος είναι πιο ελαστικός σε ότι αφορά τα σφάλματα βαθμολογίας. Αυτό που κρίνει ο δείκτης δεν είναι η βαθμολογία που έχουν λάβει (αυτό το κάνει το μέσο απόλυτο σφάλμα) αλλά εάν οι επιλογές των εναλλακτικών από τον αλγόριθμο είναι πετυχημένες.

Για την καλύτερη κατανόηση πρέπει οι εναλλακτικές να διαχωριστούν στις παρακάτω κατηγορίες. (Herlocker, et al., 2004)

Αρχικά όσο αφορά την προτίμηση του χρήστη:

- **Σχετικές (Relevant) εναλλακτικές**, όπου είναι οι εναλλακτικές οι οποίες ενδιαφέρουν το χρήστη,
- **Μη σχετικές (Irrelevant) εναλλακτικές**, οι οποίες δεν τον ενδιαφέρουν.

Και όσο αφορά την συμπεριφορά του αλγορίθμου σύστασης:

- **Επιλεγμένες (Selected) εναλλακτικές**, όπου είναι οι εναλλακτικές που το σύστημα σύστασης θεωρεί ότι ενδιαφέρουν το χρήστη και του τις προτείνει
- **Μη επιλεγμένες (Not Selected) εναλλακτικές,** όπου είναι οι εναλλακτικές όπου το σύστημα σύστασης δεν επιλέγει καθώς θεωρεί ότι δεν ενδιαφέρουν το χρήστη

Έτσι μπορεί το σύνολο των εναλλακτικών του δείγματος δεδομένων να χωριστούν στις παραπάνω κατηγορίες όπως φαίνεται στον παρακάτω πίνακα:

Πίνακας 2-3 Κατηγορίες Εναλλακτικών

	Επιλεγμένες	Μη Επιλεγμένες	Σύνολο
Σχετικές	N_{rs}	N_{rn}	N_r
Μη Σχετικές	N_{is}	N_{in}	N_i
Σύνολο	N_s	N_n	N

Έτσι ο δείκτης της ακρίβειας (Precision) συμβολίζει την πιθανότητα επιλογής μιας εναλλακτικής η οποία είναι ταυτόχρονα και σχετική και υπολογίζεται ως εξής:

$$P = \frac{N_{rs}}{N_s} \tag{2.16}$$

Όπου όπως φαίνεται και στον παραπάνω πίνακα N_{rs} είναι οι εναλλακτικές όπου έχουν επιλεγεί και είναι ταυτόχρονα σχετικές, και N_s είναι το σύνολο των επιλεγμένων εναλλακτικών.

2.6.3 Ανάκληση (Recall)

Ο δείκτης της ανάκλησης (Recall) χρησιμοποιείται συμπληρωματικά με το δείκτη της ακρίβειας. Ο δείκτης αυτός συμβολίζει τον λόγο των επιλεγμένων σχετικών εναλλακτικών N_{rs} προς το σύνολο των σχετικών εναλλακτικών που υπάρχουν N_r. Υπολογίζεται ως εξής:

$$R = \frac{N_{rs}}{N_r} \tag{2.17}$$

Σε αυτό το κεφάλαιο λοιπόν, εξηγήσαμε τι είναι και ποιος ο λόγος ύπαρξης των συστημάτων συστάσεων (Recommender Systems). Επίσης αναλύσαμε τον τρόπο λειτουργίας διάφορων υπαρχουσών μεθοδολογιών. Στο επόμενο κεφάλαιο θα περιγράψουμε αναλυτικά την υλοποίηση ενός τέτοιου συστήματος συστάσεων πάνω σε μια πραγματική εφαρμογή.

3 Ανάλυση, Σχεδίαση & Ανάπτυξη Συστήματος

3.1 Εισαγωγή

Όπως έχουμε αναφέρει και προηγουμένως τα συστήματα συστάσεων αναλαμβάνουν να φιλτράρουν, για λογαριασμό ενός χρήστη-πελάτη, τις πληροφορίες που του είναι διαθέσιμες με σκοπό την αποτελεσματική αναζήτηση των ζητούμενων πληροφοριών.

Τα συστήματα συστάσεων έχουν γίνει αρκετά δημοφιλή και έχουν εφαρμοστεί κατά καιρούς σε πολλούς και διάφορους τομείς. Συστήματα συστάσεων έχουν αρχίσει καταρχάς να εφαρμόζονται σε ηλεκτρονικά καταστήματα με χαρακτηριστικότερο παράδειγμα το ηλεκτρονικό κατάστημα Amazon. Επίσης το GroupLens το οποίο προτείνει εφαρμογή στον τομέα της ειδησεογραφίας και το MovieLens και Netflix που εξειδικεύονται στην σύσταση ταινιών. Αναπτύσσονται επίσης στον τομέα της μουσικής με εφαρμογές όπως το LastFM, Musicovery, μέχρι ακόμα και στην σύσταση εστιατορίων με το Foodio54.

Ο τομέας ο οποίος επιλέχθηκε για την εφαρμογή συστήματος συστάσεων στη διπλωματική αυτή εργασία είναι ο τομέας της μουσικής και πιο συγκεκριμένα του ραδιοφώνου. Η μουσική βιομηχανία είναι ένας τεράστιος σε όγκο πληροφοριών τομέας. Κάθε χρόνο χιλιάδες τίτλοι μουσικών άλμπουμ από γνωστούς και άγνωστους καλλιτέχνες παράγονται ανά τον κόσμο. Η μουσική με την οποία έρχεται σε επαφή ο κάθε άνθρωπος (είτε μέσω του περιβάλλοντός του είτε μέσω του ραδιοφώνου και δισκοπωλείων στον τόπο διαμονής του) είναι ένα μικρό κλάσμα της διαθέσιμης μουσικής δισκογραφίας παγκοσμίως.

Επομένως η αναζήτηση για ακρόαση μουσικής η οποία να ταιριάζει στα μουσικά γούστα και ακούσματα του κάθε ατόμου είναι μια διαδικασία η οποία απαιτεί πολύ χρόνο και γνώσεις και δεν συνεπάγεται ότι θα είναι αποτελεσματική. Η ύπαρξη ραδιοφωνικών σταθμών απαλλάσσει τους ακροατές από την αναζήτηση της μουσικής, δραστηριότητα που αναλαμβάνουν οι ραδιοφωνικοί παραγωγοί οι οποίοι διαθέτουν εξειδικευμένη γνώση. Η μόνη αναζήτηση πλέον που κάνουν οι ακροατές είναι η αναζήτηση της ραδιοφωνικής εκπομπής, και κατ' επέκταση του ραδιοφωνικού παραγωγού ο οποίος αναπαράγει μουσική. Αν μας επιτραπεί η σύγκριση με τα συστήματα συστάσεων ο ακροατής αναζητά την μέγιστη συσχέτιση του προφίλ προτιμήσεών του με του προφίλ του ραδιοφωνικού παραγωγού!

Αυτό που επιδιώκουμε σε αυτή την εργασία να κάνουμε, είναι να κατασκευάσουμε ένα ηλεκτρονικό ραδιοφωνικό σταθμό ο οποίος θα διαθέτει έναν βέλτιστο «ηλεκτρονικό ραδιοφωνικό παραγωγό» για κάθε ακροατή του.

Για την επίτευξη αυτού του στόχου θα χρησιμοποιηθεί ένα σύστημα συστάσεων συνεργασίας (Collaborative Filtering System) πάνω στο οποίο θα γίνουν οι κατάλληλες τροποποιήσεις με σκοπό την βέλτιστη λειτουργία του και της εξάλειψης των μειονεκτημάτων των μεθόδων που θα χρησιμοποιηθούν.

Στον τομέα αυτό έχουν γίνει διάφορες αξιόλογες προσπάθειες, όπως η (μόνο για τη Αμερική) εφαρμογή Pandora, το Αγγλικό LastFM, το Γαλλικό Deezer και τέλος το Musicovery. Οι υλοποιήσεις αυτές όντας εμπορικές εφαρμογές κατέστησαν αδύνατη την αναζήτηση βιβλιογραφίας για το συγκεκριμένο τομέα εφαρμογής και για την επίλυση των περισσοτέρων προβλημάτων.

3.2 Σχεδίαση Ραδιοφώνου

Πέρα από την υλοποίηση των αλγορίθμων, οι οποίοι θα εξηγηθούν αναλυτικά παρακάτω σε αυτό το κεφάλαιο, η κατασκευή του ηλεκτρονικού ραδιοφώνου προϋποθέτει και την σχεδίαση μιας ιστοσελίδας (Web Page) μέσα στην οποία θα φιλοξενηθεί η εφαρμογή του ραδιοφώνου. Η σελίδα πρέπει να παρέχει στον χρήστη-ακροατή τις απαραίτητες πληροφορίες καθώς και να του δίνει τη δυνατότητα να περιηγείται και να αλληλεπιδρά με την εφαρμογή μέσω ενός φιλικού προς αυτόν περιβάλλοντος.

Για τη σχεδίαση αυτής της ιστοσελίδας χρησιμοποιήθηκε η γλώσσα προγραμματισμού ASP καθώς επίσης για τη διαχείριση της πληροφορίας σχεδιάστηκε μια βάση δεδομένων σε Access. Λόγω του όγκου των πληροφοριών που καλείται η βάση δεδομένων να καλύψει, σε πραγματικά δεδομένα, προβλέπεται η χρήση της ποιο εξειδικευμένης και κατά πολύ πιο γρήγορης βάσης δεδομένων SQL στο μέλλον.

Η λειτουργία και η σχεδίαση της ιστοσελίδας θα παρουσιαστούν αναλυτικά στο επόμενο κεφάλαιο. Περιληπτικά όμως αξίζει να αναφερθούν τα κυριότερα στοιχεία της ιστοσελίδας:

Αρχική Σελίδα Εφαρμογής

Αυτή είναι η αρχική σελίδα της εφαρμογής όπου δίνονται κάποια περιληπτικά στοιχεία για την εφαρμογή καθώς επίσης υπάρχει η φόρμα εγγραφής και εισόδου χρηστών.

Αρχική Σελίδα Χρήστη

Αυτή είναι η σελίδα με την οποία έρχεται σε επαφή ο εγγεγραμμένος χρήστης με την είσοδό του στην ιστοσελίδα. Περιέχει:

a) Στατιστικά για τον χρήστη
b) Τους γειτονικούς χρήστες
c) Επιλογή για την εκπαίδευση του προφίλ του χρήστη
d) Επιλογή για πρόταση μουσικής στο χρήστη
e) Επιλογή για αναζήτηση μουσικής

Σελίδα Αναπαραγωγής

Η σελίδα αναπαραγωγής είναι η σελίδα όπου ο χρήστης ακούει πλέον τη μουσική. Αυτή η σελίδα χωρίζεται σε δυο τομείς:

I. Αναπαραγωγή κομματιού και διαθέσιμες πληροφορίες για το κομμάτι που αναπαράγεται
II. Συλλογή δεδομένων από το χρήστη: αξιολόγηση και κατηγοριοποίηση του κομματιού που αναπαράγεται.

3.3 Συλλογή Δεδομένων

Η συλλογή δεδομένων από τους χρήστες-ακροατές του ραδιοφώνου γίνεται με δυο τρόπους:

- **Άμεσα (Explicitly)**: Ο χρήστης κάνοντας χρήση των δυο κουμπιών στη σελίδα αναπαραγωγής μουσικής δηλώνει την προτίμησή του (Love, με το κουμπί καρδιά) ή την μη (Ban, με το κουμπί απαγορευτικό) στο μουσικό κομμάτι που ακούει.

- **Έμμεσα (Implicitly)**: Συλλέγουμε δεδομένα επίσης όταν ο χρήστης-ακροατής ακούσει ένα κομμάτι ολόκληρο ή όταν πατήσει να προχωρήσει στο επόμενο κομμάτι (Skip track) ως δείγμα αρέσκειας ή μη αντίστοιχα. Επίσης αποθηκεύεται ιστορικό με όλα τα κομμάτια που έχει ακούσει ο χρήστης (μαζί με την ημερομηνία και ώρα ακρόασης) καθώς και εκείνα που προχώρησε (skip).

Τα δεδομένα άμεσα και έμμεσα που συλλέγονται αντιστοιχίζονται σε μια τετραβάθμια κλίμακα αξιολόγησης στον πίνακα βαθμολογιών. Η κλίμακα είναι ως εξής:

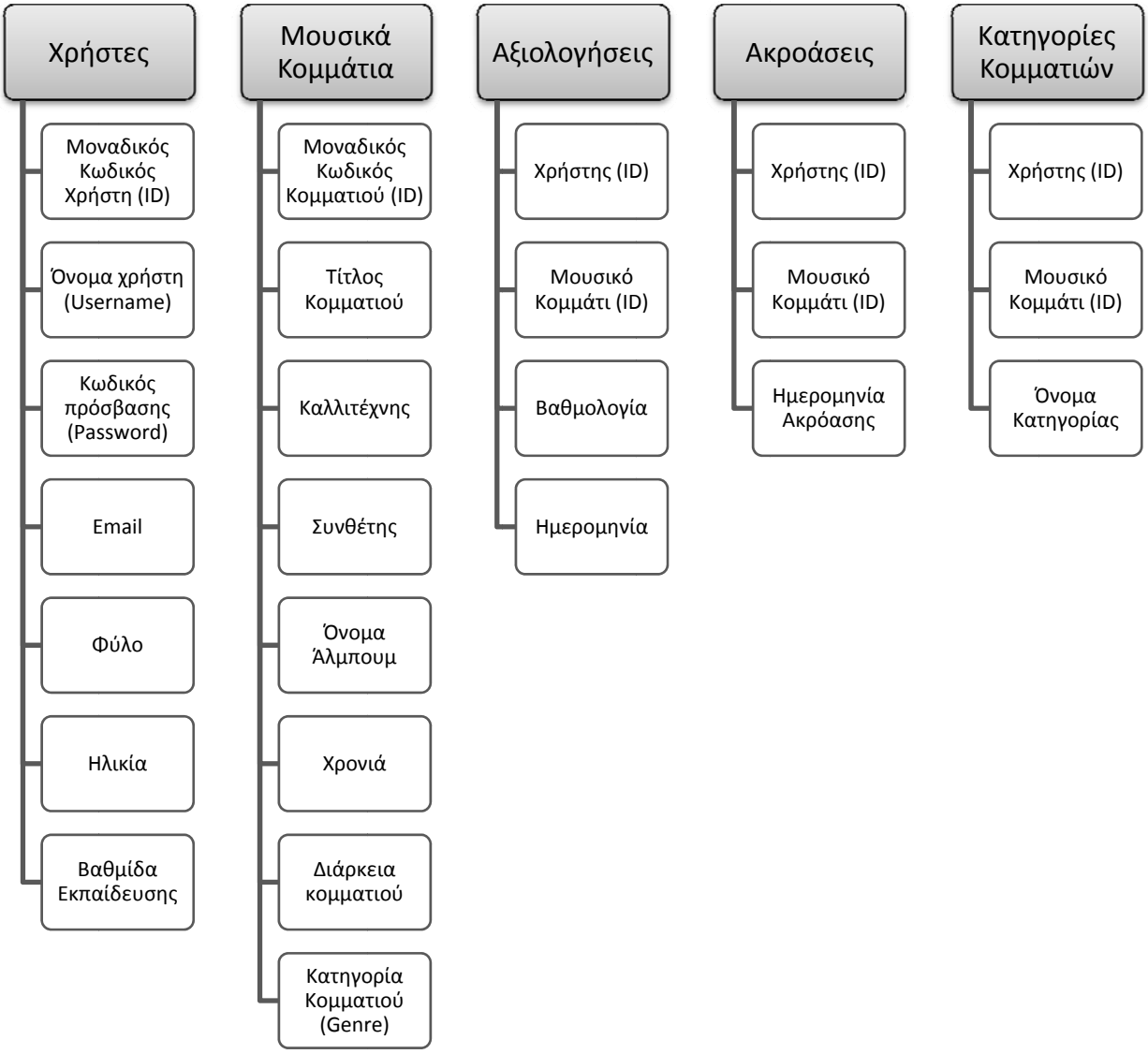
0: Με 0 βαθμολογείται ένα μουσικό κομμάτι όταν ο χρήστης επιλέξει ότι δεν του άρεσε (Ban, το κουμπί απαγορευτικό). ***Ισχυρή βαθμολογία***
1: Με 1 βαθμολογείται ένα μουσικό κομμάτι όταν ο χρήστης επιλέξει να προχωρήσει στο επόμενο κομμάτι (Skip) ***Ασθενής βαθμολογία***
2: Με 2 βαθμολογείται ένα μουσικό κομμάτι όταν ο χρήστης αφήσει το κομμάτι να παίξει ολόκληρο ***Ασθενής βαθμολογία***
3: Με 3 βαθμολογείται ένα μουσικό κομμάτι όταν ο χρήστης επιλέξει ότι του άρεσε (Love, το κουμπί με την καρδιά). ***Ισχυρή βαθμολογία***

Οι βαθμολογίες έχουν χαρακτηριστεί επίσης ως ***ισχυρές*** και ***ασθενείς*** ανάλογα αν έχουν αποκτηθεί ***άμεσα*** από το χρήστη ή ***έμμεσα*** αντίστοιχα. Αυτό έχει γίνει για τις περιπτώσεις όπου ο χρήστης ακούσει ένα μουσικό κομμάτι για παραπάνω από μια φορά και επιλέξει να το αξιολογήσει ξανά. *Μια ασθενής βαθμολογία μπορεί να αλλάξει την τιμή **μόνο** μιας ασθενής βαθμολογίας. Ενώ μια ισχυρή βαθμολογία μπορεί να αλλάξει την τιμή **και** μιας ασθενής **και** μιας ισχυρής βαθμολογίας.*

Για τη συλλογή των δεδομένων η εφαρμογή του ηλεκτρονικού ραδιοφώνου λειτούργησε δοκιμαστικά για διάστημα 4 μηνών συλλέγοντας δεδομένα και αξιολογήσεις μουσικών κομματιών από μια ομάδα περίπου 100 χρηστών.

3.4 Αποθήκευση Δεδομένων

Τα δεδομένα από τη συλλογή δεδομένων καθώς και όλα τα υπόλοιπα δεδομένα που χρειάζονται για την εφαρμογή αποθηκεύονται σε μια βάση δεδομένων. Η βάση δεδομένων έχει την εξής δομή:



3.5 Σύστημα Σύστασης

Αφού αναλύθηκε η σχεδίαση της ιστοσελίδας του ραδιοφώνου και η συλλογή των δεδομένων, τώρα θα γίνει ανάλυση του χρησιμοποιούμενου αλγόριθμου.

Το σύστημα σύστασης που χρησιμοποιήθηκε για την κατασκευή του πRadio είναι αλγόριθμος της κατηγορίας των συστάσεων συνεργασίας (Collaborative Filtering) και συγκεκριμένα προσανατολιζόμενος στη μνήμη (Memory Based).

Ο λόγος που επιλέχθηκε αυτός ο αλγόριθμος έχει να κάνει με τη φύση της πληροφορίας που θέλουμε να φιλτράρουμε. Στη συγκεκριμένη κατηγορία, της μουσικής, έχουμε περιορισμένη πληροφορία για τα μουσικά κομμάτια όπως το όνομα του καλλιτέχνη, το άλμπουμ, το έτος κυκλοφορίας και ορισμένα άλλα. Επίσης όπως αναφέραμε κατά την ανάλυση των συστημάτων συστάσεων τα συστήματα περιεχομένου (Content Based) αδυνατούν (ή τουλάχιστον είναι αρκετά δύσκολο) να διαχειριστούν πληροφορία που δεν είναι σε μορφή κειμένου. Τα δεδομένα της εφαρμογής που επιλέξαμε είναι μουσικά κομμάτια και άρα αδύνατο να διαχειριστούν από ένα σύστημα Content Based. Επομένως η επιλογή ενός συστήματος Collaborative Filtering είναι η ενδεδειγμένη λύση καθώς όπως έχουμε αναφέρει και στο Κεφάλαιο 2 τα συστήματα αυτά μπορούν να αξιολογούν εναλλακτικές που δεν περιέχουν κείμενο με βάση ποιοτικά χαρακτηριστικά τους.

3.5.1 Πρόβλεψη Βαθμολογίας

Για την πρόβλεψη της βαθμολογίας του μουσικού κομματιού χρησιμοποιείται η εξίσωση (2.7) που εξηγήσαμε στο Κεφάλαιο 2.

$$p_{a,j} = \bar{r}_a + \kappa \sum_{i=1}^{n} w(a,i)\left(r_{i,j} - \bar{r}_i\right) \tag{3.1}$$

Όπου n είναι το σύνολο των γειτόνων του ενεργού χρήστη α, $w(\alpha,i)$ βάρος, $r_{i,j}$ η αξιολόγηση του χρήστη i στο μουσικό κομμάτι j, $\bar{r}_i$ η μέση τιμή των αξιολογήσεων του χρήστη i και τέλος το κ υπολογίζεται ως:

$$\kappa = \frac{1}{\sum_{i=1}^{n} w(a,i)} \tag{3.2}$$

3.5.2 Υπολογισμός Βαρών

Για τον υπολογισμό βαρών *w* χρησιμοποιήθηκε η συσχέτιση κατά Pearson (Pearson's Correlation) που φαίνεται στην εξίσωση (2.11):

$$w(a,i) = \frac{\sum_j (r_{a,j} - \bar{r}_a)(r_{i,j} - \bar{r}_i)}{\sqrt{\sum_j (r_{a,j} - \bar{r}_a)^2 \sum_j (r_{i,j} - \bar{r}_i)^2}} \tag{3.3}$$

3.5.3 Αντιμετώπιση Περιορισμών Collaborative Filtering

Όπως αναλύθηκε στο 2^{ο} Κεφάλαιο τα συστήματα Collaborative Filtering έχουν κάποιους περιορισμούς. Στην εφαρμογή αυτή έγινε προσπάθεια αντιμετώπισης του περιορισμού Cold Start. Και συγκεκριμένα:

3.5.3.1 Πρόβλημα Νέου Χρήστη

Κατά την εγγραφή ενός νέου χρήστη στην εφαρμογή του ηλεκτρονικού ραδιοφώνου το σύστημα δε μπορεί να κάνει πρόταση μουσικής καθώς ο χρήστης δεν έχει αξιολογήσει κανένα μουσικό κομμάτι. Για το λόγο αυτό υπάρχει η επιλογή (στην αρχική σελίδα του χρήστη) εκπαίδευσης του προφίλ. Εκεί είναι διαθέσιμη μια λίστα 50 μουσικών κομματιών τα οποία καλείται ο χρήστης να αξιολογήσει με σκοπό να «εκπαιδευτεί» το προφίλ του και να μπορέσουν να του γίνουν προτάσεις.

Τα 50 αυτά μουσικά κομμάτια έχουν επιλεγεί με δυο διαφορετικούς τρόπους:

1. Τα 30 κομμάτια έχουν επιλεγεί με τυχαίο τρόπο από την λίστα όλων των κομματιών. Αυτή η επιλογή έγινε μια φορά στην αρχή λαμβάνοντας με ομοιόμορφο τρόπο κομμάτια από κάθε μουσική κατηγορία (Genre) και την ίδια λίστα καλούνται να αξιολογήσουν όλοι οι χρήστες.
2. Τα υπόλοιπα 20 κομμάτια επιλέγονται κάθε φορά που εγγράφεται ένας νέος χρήστης στο σύστημα. Ο τρόπος επιλογής τους περιγράφεται από τους (Rashid, και συν., 2002) και λαμβάνει υπόψη του την διακύμανση των βαθμολογιών που έχει λάβει το κάθε μουσικό κομμάτι.

Με τον τρόπο αυτό κάθε νέος χρήστης με την εκπαίδευση του προφίλ του έχει κοινές αξιολογήσεις με όλους τους υπόλοιπους χρήστες καθώς και έχει αξιολογήσει και τα κομμάτια που θα παρέχουν βέλτιστη πληροφορία για την κατασκευή του προφίλ του.

3.5.3.2 Πρόβλημα Νέου Αντικειμένου

Δυστυχώς τα Collaborative Filtering συστήματα όπως αυτό που εφαρμόστηκε στο ηλεκτρονικό ραδιόφωνο δεν μπορούν να διαχειριστούν το Πρόβλημα του νέου Αντικειμένου όπως έχει αναφερθεί στο 2^{ο} Κεφάλαιο. Ο μόνος τρόπος να εισαχθεί το νέο αντικείμενο (μουσικό κομμάτι στην περίπτωση του ραδιοφώνου) είναι να το αναζητήσει ο χρήστης (music on demand) και να το αξιολογήσει.

Στο τελευταίο Κεφάλαιο προτείνεται τρόπος αντιμετώπισης αυτού του προβλήματος ως πρόταση για το μέλλον.

3.5.4 Βελτίωση Αποτελεσμάτων

Για την βελτίωση των αποτελεσμάτων του αλγορίθμου χρησιμοποιήθηκαν τρεις τεχνικές που αναφέρονται στη βιβλιογραφία (Herlocker, et al., 1999). Συγκεκριμένα οι αλλαγές που προτείνονται είναι οι εξής:

3.5.4.1 Κοντινότεροι N Γείτονες (Top N Neighbors)

Σύμφωνα με τον τύπο (3.1) η πρόβλεψη της βαθμολογίας του κομματιού γίνεται λαμβάνοντας υπόψη τις βαθμολογίες όλων των γειτόνων του ενεργού χρήστη. Στην προτεινόμενη όμως αλλαγή λαμβάνεται υπόψη η βαθμολογία μονάχα των *N* καλύτερων γειτόνων του ενεργού χρήστη. Με την αλλαγή αυτή αποφεύγεται η ύπαρξη «θορύβου» εξαιτίας των βαθμολογιών γειτόνων με πολύ χαμηλές συσχετίσεις με τον ενεργό χρήστη που μπερδεύουν τα αποτελέσματα του αλγόριθμου.

Η τιμή του αριθμού *N* έχει μεγάλη σημασία στην ακρίβεια των αποτελεσμάτων. Ένας πολύ μεγάλος αριθμός *N* συνεπάγεται εισαγωγή γειτόνων με χαμηλή συσχέτιση με τον ενεργό χρήστη, ενώ ένας μικρός αριθμός *N* αγνοεί χρήσιμα δεδομένα από κοντινούς γείτονες οι οποίοι αποκλείονται. Στις δυο αυτές περιπτώσεις έχουμε χειρότερα αποτελέσματα. Έτσι έπειτα από ανάλυση των δεδομένων, όπως θα δούμε παρακάτω σε αυτό το κεφάλαιο, υπολογίζεται ο ιδανικός αριθμός *N*.

3.5.4.2 Κατώφλι Βάρους (Weight Threshold)

Σε αυτή την τεχνική καθορίζεται ένα κατώτατο όριο συσχέτισης που θα πρέπει ο ενεργός χρήστης να έχει με τους γείτονές του. Έτσι επιλέγονται όλοι οι γείτονες που έχουν μεγαλύτερη συσχέτιση από το κατώφλι αυτό. Μπορεί να επιλεγεί ένα

κατώφλι αυστηρό όπου θα ζητάει μεγάλη συσχέτιση μεταξύ των χρηστών, ή ένα κατώφλι πιο χαλαρό. Η επιλογή της βέλτιστης τιμής του κατωφλιού υπολογίζεται μέσω δοκιμών στο δείγμα δεδομένων.

Συνήθως επιλέγεται η μια από τις δυο τεχνικές ανάλογα με την βελτίωση των αποτελεσμάτων, υπάρχουν όμως και περιπτώσεις όπου γίνεται συνδυασμός και των δυο τεχνικών.

3.5.4.3 Συντελεστής Σπουδαιότητας (Significance Weighting)

Ένα συχνό φαινόμενο στα Collaborative Filtering συστήματα είναι περιπτώσεις χρηστών που εμφανίζουν ένα πολύ υψηλό δείκτη συσχέτισης αλλά έχουν πολύ μικρό αριθμό κοινών αξιολογημένων εναλλακτικών. Έτσι πιθανότατα αυτή η μεγάλη συσχέτιση να μην ανταποκρίνεται στην πραγματικότητα μιας και έχουμε μικρό δείγμα κοινών αξιολογημένων εναλλακτικών με αποτέλεσμα να επηρεάζει αρνητικά τα αποτελέσματα.

Γι’ αυτό το λόγο εισάγεται ο συντελεστής σπουδαιότητας (Significance Weighting) ο οποίος πολλαπλασιάζεται με τη συσχέτιση των χρηστών και υπολογίζεται ως εξής:

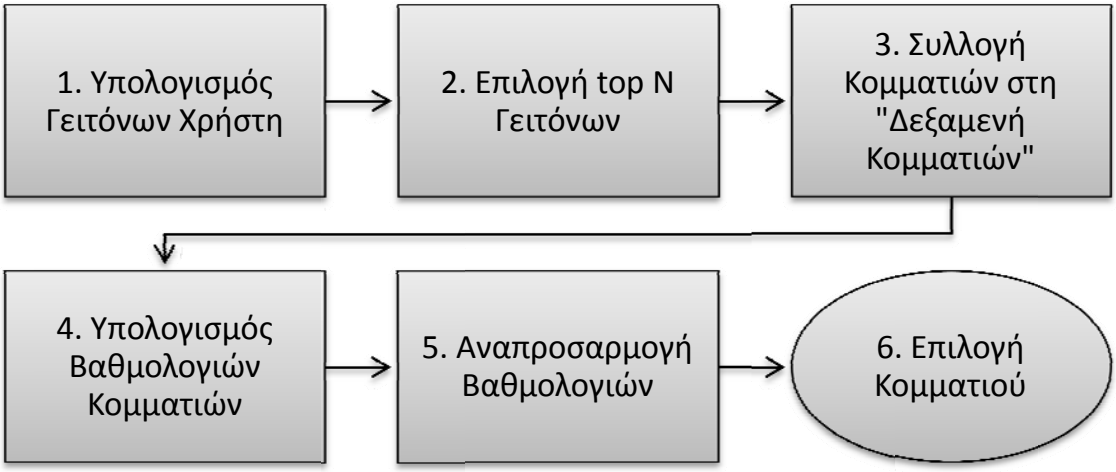
$$s = \frac{Αριθμός\ κοινών\ αξιολογημένων\ κομματιών}{Αποδεκτός\ αριθμός\ κομματιών} \tag{3.4}$$

Έτσι λοιπόν αν υποθέσουμε ότι ένας αποδεκτός αριθμός κομματιών που θα πρέπει δυο χρήστες να έχουν κοινά ώστε να έχουμε ικανά αποτελέσματα είναι 50, και ο αριθμός των κοινών αξιολογημένων κομματιών δυο χρηστών είναι 10 τότε ο συντελεστής Significance Weighting είναι 10/50 = 1/5.

Ο αποδεκτός αριθμός κομματιών υπολογίζεται και αυτός μέσω δοκιμών στο δείγμα δεδομένων μέχρι να βρεθεί εκείνος που βελτιώνει τα αποτελέσματα.

3.5.5 Πρόταση Μουσικής (Αλγόριθμος)

Κάθε φορά όπου ο χρήστης επιλέξει να του προτείνει το σύστημα μουσική τότε η διαδικασία που εκτελείται είναι η ακόλουθη:



Σχήμα 3-1 Αλγόριθμος Πρότασης Μουσικής

Αναλυτικά για τα παραπάνω βήματα έχουμε:

1. Το σύστημα συγκρίνει όλους τους χρήστες με τον ενεργό χρήστη χρησιμοποιώντας την εξίσωση (3.3) και υπολογίζει τις μεταξύ τους συσχετίσεις. Ταξινομώντας τους χρήστες με βάση τη συσχέτισή τους με τον ενεργό χρήστη δημιουργείται η λίστα με τους γείτονες του χρήστη. Από τους πιο κοντινούς στους πιο μακρινούς.

2. Στη συνέχεια αφού το σύστημα έχει ταξινομήσει τους γείτονες του ενεργού χρήστη επιλέγει και κρατάει στη μνήμη του μόνο τους top N γείτονες. Η τιμή του *N* θα υπολογιστεί σε παρακάτω παράγραφο στο κεφάλαιο αυτό.

3. Όλα τα κομμάτια που έχουν αξιολογηθεί από τους επιλεγμένους γείτονες του ενεργού χρήστη συλλέγονται σε μια «δεξαμενή κομματιών» (track pool). Από αυτή τη «δεξαμενή» στην πορεία του αλγόριθμου θα γίνει η πρόταση μουσικής.

4. Για κάθε ένα κομμάτι της «δεξαμενής κομματιών» υπολογίζεται η προτεινόμενη βαθμολογία σύμφωνα με τη μεθοδολογία Collaborative Filtering κάνοντας χρήση του τύπου (3.1).

5. Σε αυτό το βήμα γίνεται μια διαδικασία αναπροσαρμογής των βαθμολογιών των κομματιών της «δεξαμενής κομματιών» ανάλογα με το πότε άκουσε τελευταία φορά ο ενεργός χρήστης το κάθε κομμάτι. Η διαδικασία αυτή της χρονικής απαξίωσης (Time Depreciation) της βαθμολογίας των κομματιών θα εξηγηθεί αναλυτικά παρακάτω.

6. Τέλος επιλέγεται το κομμάτι αυτό όπου συγκεντρώνει τη μεγαλύτερη προτεινόμενη βαθμολογία και εφόσον έχει βαθμολογία μεγαλύτερη του 1 (δηλαδή ο χρήστης εκτιμάται ότι δε θα το κάνει skip αλλά θα το ακούσει ολόκληρο) προωθείται στον χρήστη να το ακούσει.

Η προτεινόμενη αυτή διαδικασία εκτελείται κάθε φορά που ένα κομμάτι προτείνεται στον χρήστη. Η συγκεκριμένη διαδικασία στο δείγμα δεδομένων της εργασίας (~5000 μουσικά κομμάτια με ~100 χρήστες) δούλεψε ικανοποιητικά γρήγορα βγάζοντας προτεινόμενη βαθμολογία σε λιγότερο από 1sec.

Σημείωση

Όμως η διαδικασία αυτή δεν είναι βέλτιστη για αρκετά μεγαλύτερα συστήματα (Scalable). Προτεινόμενη αλλαγή αυτής της διαδικασίας είναι η επιλογή λίστας των 20 ή 30 βέλτιστων κομματιών κατά την εκτέλεση του βήματος 6. Έτσι ο αλγόριθμος θα εκτελείται μόνο κάθε φορά όπου η λίστα κομματιών αδειάζει ή κάθε νέα φορά που εισέρχεται ο ενεργός χρήστης στο σύστημα.

Χρονική Απαξίωση (Time Depreciation)

Κατά την εκτέλεση του βήματος 6 ένα πρόβλημα που προέκυπτε είναι ότι το κομμάτι με τη μεγαλύτερη βαθμολογία, όπου ο αλγόριθμος κάθε φορά επέλεγε, ήταν το ίδιο με μερικές εξαιρέσεις όπου το προφίλ του χρήστη ή των γειτόνων του άλλαζαν. Μια και δε βρέθηκε αντίστοιχο πρόβλημα στη βιβλιογραφία αυτό το πρόβλημα λύθηκε ad hoc. Η λύση δόθηκε με την εισαγωγή ενός συντελεστή Χρονικής Απαξίωσης ο οποίος πολλαπλασιαζόμενος με την προτεινόμενη βαθμολογία των κομματιών $p_{\alpha j}$ την αλλάζει αναλόγως με το πότε ακούστηκαν τελευταία φορά από τον χρήστη.

$$s_{dep} = \frac{p_{\alpha j} - 1}{p_{\alpha j}} \frac{t}{10080} + \frac{1}{p_{\alpha j}} \tag{3.5}$$

Όπου t είναι η χρονική περίοδος όπου έχει μεσολαβήσει από την τελευταία φορά όπου ο ενεργός χρήστης άκουσε το μουσικό κομμάτι εκφρασμένη σε λεπτά της ώρας (min), $p_{\alpha j}$ η προτεινόμενη βαθμολογία του κομματιού j στον ενεργό χρήστη α και ο αριθμός στον παρονομαστή είναι 7 ημέρες υπολογισμένες σε λεπτά της ώρας (min).

Από τον τύπο φαίνεται ότι ο συντελεστής αυτός είναι μια γραμμική συνάρτηση της χρονικής περιόδου που έχει διανυθεί από την τελευταία φορά όπου ο ενεργός χρήστης άκουσε το κομμάτι.

Στην στιγμή t = 0 δηλαδή τη στιγμή όπου ο χρήστης μόλις άκουσε το κομμάτι ο συντελεστής γίνεται $s_{dep} = \frac{1}{p_{\alpha j}}$ και η βαθμολογία παίρνει την ελάχιστη τιμή της ίση με τη μονάδα (1). Αυτό συμβαίνει γιατί ένα κομμάτι που μόλις ακούστηκε σίγουρα θα ξαναπροταθεί αν το επόμενο βέλτιστο προτεινόμενο κομμάτι έχει προτεινόμενη βαθμολογία μικρότερη από 1 (που σημαίνει ότι ο χρήστης δε θα το ακούσει (skip)).

Επίσης μετά την πάροδο 7 ημερών ο συντελεστής γίνεται $s_{dep} = 1$ και η βαθμολογία του κομματιού θα επανέλθει στην κανονική της τιμή (χωρίς αυτό να σημαίνει ότι δε θα μπορεί να προταθεί και νωρίτερα).

Οι 7 ημέρες προτάθηκαν ως μια φυσιολογική περίοδος που ο χρήστης θα θέλει πάλι να ακούσει ένα κομμάτι.

3.6 Ανάλυση Δεδομένων

Οι αλλαγές που προτάθηκαν για τη βελτίωση των αποτελεσμάτων της συγκεκριμένης εφαρμογής στην παράγραφο 3.5.3 απαιτούν τον προσδιορισμό κάποιων συντελεστών. Για την επιλογή των βέλτιστων αυτών συντελεστών χρειάζεται να γίνει σύγκριση των αποτελεσμάτων του αλγόριθμου για τις διάφορες τιμές τους.

Για να γίνει αυτή η σύγκριση θα χρειαστεί ένα μέτρο σύγκρισης. Ένα μέτρο σύγκρισης των μεθόδων είναι η αξιολόγησή τους όπως αναφέρθηκε στο 2^{ο} Κεφάλαιο. Από τις διάφορες προτάσεις αξιολόγησης των αποτελεσμάτων των συστημάτων συστάσεων στη συγκεκριμένη εφαρμογή θα χρησιμοποιηθεί το ***Μέσο Απόλυτο Σφάλμα*** και η ***Ακρίβεια (Precision)*** όπου έχουν αναλυθεί στο 2^{ο} Κεφάλαιο.

Οι βελτιωτικές αλλαγές που προτάθηκαν και στην παράγραφο 3.5.3 για την εφαρμογή είναι οι εξής τρεις:

- Οι top N γείτονες,
- Το κατώφλι συσχέτισης (Correlation Threshold),
- Ο συντελεστής σπουδαιότητας (Significance Weighting).

Στη συνέχεια θα γίνει ανάλυση του δείγματος δεδομένων για τον καθορισμό των τιμών των N γειτόνων, του κατωφλιού και του συντελεστή σπουδαιότητας.

Δοκιμάστηκαν διάφορες τιμές για τους top N γείτονες και για το κατώφλι συσχέτισης και υπολογίστηκαν οι τιμές της Ακρίβειας (P), του συνόλου των επιλεγμένων εναλλακτικών (N_s) και τα αντίστοιχα μέσα απόλυτα σφάλματα (MAE), έχοντας το συντελεστή σπουδαιότητας απενεργοποιημένο.

Πίνακας 3-1 Σύνολο Επιλεγμένων Εναλλακτικών Ns & Ακρίβεια P

		Correlation Threshold									
		0		0,1		0,2		0,3		0,4	
		Ns	P	Ns	P	Ns	P	Ns	P	Ns	P
Top N	5	1104	0,861	1094	0,859	1076	0,857	1064	0,865	874	0,856
	10	2784	0,807	2768	0,806	2729	0,808	2280	0,829	1452	0,841
	15	3893	0,800	3877	0,799	3792	0,800	2776	0,831	1599	0,847
	20	4366	0,797	4336	0,797	4156	0,804	2824	0,831	1600	0,847
	25	4675	0,799	4641	0,799	4308	0,806	2824	0,831	1600	0,847
	30	4939	0,796	4862	0,798	4319	0,807	2824	0,831	1600	0,847
	OFF	5187	0,791	5010	0,797	4319	0,807	2824	0,831	1600	0,847

Πίνακας 3-2 Σύνολο Επιλεγμένων Εναλλακτικών Ns & Ακρίβεια P

		Correlation Threshold					
		0,5		0,6		0,7	
		Ns	P	Ns	P	Ns	P
Top N	5	643	0,891	428	0,918	334	0,937
	10	880	0,893	465	0,923	350	0,94
	15	910	0,892	472	0,922	350	0,94
	20	910	0,892	472	0,922	350	0,94
	25	910	0,892	472	0,922	350	0,94
	30	910	0,892	472	0,922	350	0,94
	OFF	910	0,892	472	0,922	350	0,94

Πίνακας 3-3 Μέσο Απόλυτο Σφάλμα (MAE)

		Correlation Threshold							
	MAE	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7
Top N	2	0,661	0,66975	0,64024	0,64024	0,64024	0,63717	0,59149	0,59288
	5	0,74525	0,74809	0,74029	0,74017	0,69307	0,66421	0,58744	0,58832
	10	0,79048	0,79237	0,79104	0,79315	0,72857	0,68682	0,5924	0,58976
	15	0,784	0,78519	0,78879	0,80062	0,72327	0,68366	0,59218	0,58984
	20	0,7702	0,77109	0,78081	0,80012	0,72365	0,68397	0,59246	0,59022
	25	0,76034	0,76083	0,77236	0,79997	0,72336	0,68401	0,59246	0,59022
	30	0,7554	0,7569	0,77325	0,79995	0,72336	0,68401	0,59246	0,59022

Για κατώφλια συμφωνίας άνω του 0.8 δεν λάβαμε αποτέλεσμα για MAE καθώς δεν υπήρχαν συσχετίσεις χρηστών πάνω από αυτό το όριο.

Παρατηρείται ότι για κατώφλια συμφωνίας άνω του 0.4 το MAE για αριθμό γειτόνων πάνω από 8-10 παραμένει σταθερό. Αυτό δε σημαίνει ότι υπάρχει σταθεροποίηση των αποτελεσμάτων αλλά δεν υπάρχουν περισσότεροι γείτονες σε τόσο ψηλά όρια εξαιτίας του μικρού δείγματος αξιολογήσεων. Αυτό φαίνεται και από τον αριθμό των επιλεγμένων κομματιών N_s ο οποίος επίσης δε μεταβάλλεται.

Οι χαμηλότερες τιμές MAE καθώς και οι υψηλότερες τιμές Ακρίβειας (P) εμφανίζονται για αριθμό γειτόνων 2 και 5. Σε αυτό όμως τον μικρό αριθμό γειτόνων ο αριθμός των επιλεγμένων κομματιών (N_s) είναι πολύ μικρός. Όσο μικρότερος είναι ο αριθμός N_s τόσο μικρότερος θα είναι και ο αριθμός των κομματιών στην «δεξαμενή κομματιών» (track pool). Αυτό συνεπάγεται περιορισμένο πλήθος προτάσεων από το σύστημα.

Συνεπώς αναζητείται η μεγιστοποίηση του αριθμού N_s συνδυασμένη με ικανοποιητικό μέσο απόλυτο σφάλμα. Η τιμή του MAE, όπως φαίνεται από τον παραπάνω πίνακα, αυξάνεται όσο αυξάνεται ο αριθμός των γειτόνων (N<10) αλλά για αριθμό γειτόνων μεγαλύτερο των 10 μειώνεται. Σύμφωνα όμως με τα προηγούμενα απορρίπτεται ο μικρός αριθμός γειτόνων καθώς συνεπάγεται περιορισμένο πλήθος προτάσεων. Επομένως επιλέγεται η απενεργοποίηση της τεχνικής των top N γειτόνων.

Δοκιμάστηκαν επίσης και διάφορες τιμές του συντελεστή σημαντικότητας (Significance Weighting) και υπολογίστηκε το MAE:

Πίνακας 3-4 Significance Weighting

Top N OFF, Correl Threshold = 0		Top N OFF, Correl Threshold = 0,3	
Sign Weight	MAE	Sign Weight	MAE
0	0,75371	0	0,79995
5	0,75371	5	0,79995
10	0,75367	10	0,79938
15	0,75367	15	0,79941
20	0,75367	20	0,79869
25	0,75366	25	0,79864
30	0,75358	30	0,79868
35	0,75347	35	0,79871
40	0,75338	40	0,79872
45	0,75327	45	0,79867
50	0,75321	50	0,79880

Οι δοκιμές έγιναν για δυο τιμές κατωφλιού 0 και 0.3 και με απενεργοποιημένη την τεχνική των τοπ N γειτόνων.

Παρατηρείται από τα δεδομένα του πίνακα ότι ακόμα και με συντελεστή σημαντικότητας 50 (το οποίο σημαίνει ότι δυο χρήστες πρέπει να έχουν περισσότερα από 50 κοινά κομμάτια) η βελτίωση των αποτελεσμάτων είναι πολύ μικρή έως αμελητέα. Οπότε επιλέγεται να μην γίνει χρήση του συντελεστή σημαντικότητας (Significance Weighting) στο σύστημα.

Οπότε τελικά στην εφαρμογή του ηλεκτρονικού ραδιοφώνου και για το συγκεκριμένο δείγμα δεδομένων που έχει συλλεχθεί:

- Δε γίνεται χρήση της τεχνικής των top N γειτόνων (Άρα $N=\infty$) καθώς περιορίζεται το πλήθος της «δεξαμενής κομματιών» στο συγκεκριμένο δείγμα δεδομένων.
- Χρησιμοποιείται η τεχνική κατωφλιού συσχέτισης (Correlation Threshold) με τιμή συσχέτισης ίση με 0.

- Δε γίνεται χρήση της τεχνικής συντελεστή σημαντικότητας (Significance Weighting) καθώς οι βελτιώσεις που προσφέρει στο συγκεκριμένο δείγμα δεδομένων είναι σχετικά αμελητέες.

Είναι σημαντικό να αναφερθεί ότι οι τιμές που επιλέχτηκαν αφορούν μονάχα το συγκεκριμένο δείγμα δεδομένων που συλλέχθηκε στα πλαίσια της εργασίας. Σε περίπτωση αλλαγής αυτού του δείγματος, έπειτα από περαιτέρω συλλογή δεδομένων, θα χρειαστεί εκ νέου ανάλυση των δεδομένων για τον υπολογισμό των βέλτιστων συντελεστών.

Σε αυτό το κεφάλαιο περιγράψαμε την υλοποίηση ενός συστήματος συστάσεων (Recommender System) πάνω σε μια πραγματική εφαρμογή, σε μια εφαρμογή εξατομικευμένου ηλεκτρονικού ραδιοφώνου. Στο επόμενο κεφάλαιο θα περιγράψουμε τη λειτουργία του συστήματος μέσω πραγματικής εφαρμογής του.

4 Εφαρμογή Συστήματος

4.1 Εισαγωγή

Αφού παρουσιάστηκε αναλυτικά το σύστημα που χρησιμοποιήθηκε, σε αυτό το κεφάλαιο θα γίνει παρουσίαση του συστήματος κατά την πραγματική του εφαρμογή από έναν χρήστη. Θα παρουσιαστούν οι διάφορες σελίδες και θα επεξηγηθούν περαιτέρω οι διάφορες διαθέσιμες λειτουργίες.

4.2 Αρχική Σελίδα

Κατά την είσοδό του στην ιστοσελίδα ο χρήστης έρχεται σε επαφή με την αρχική σελίδα του ραδιοφώνου.

Εικόνα 4-1 Αρχική σελίδα ραδιοφώνου

Συγκεκριμένα έχουμε:

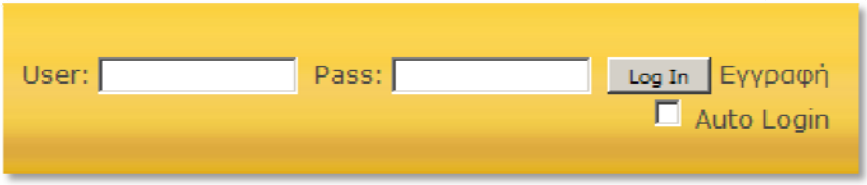
Το λογότυπο του ραδιοφώνου

Επιλέχθηκε η ονομασία πRadio προερχόμενη από συντομογραφία του π (personalized) Ραδιοφώνου (Radio)

Εικόνα 4-2 Λογότυπο

Περιοχή Σύνδεσης και Εγγραφής Χρήστη (Login & Signup)

Για να μπορέσει το σύστημα να κρατάει δεδομένα και να θυμάται τις επιλογές του κάθε χρήστη-ακροατή είναι απαραίτητη η εγγραφή του σε αυτό. Η εγγραφή γίνεται από το αντίστοιχο κουμπί στην περιοχή, όπου ο χρήστης ζητείται να εισάγει ένα μοναδικό όνομα με το οποίο θα τον αναγνωρίζει το σύστημα (Username), ένα κωδικό πρόσβασης για λόγους ασφαλείας.



Εικόνα 4-3 Περιοχή Σύνδεσης & Εγγραφής Χρήστη

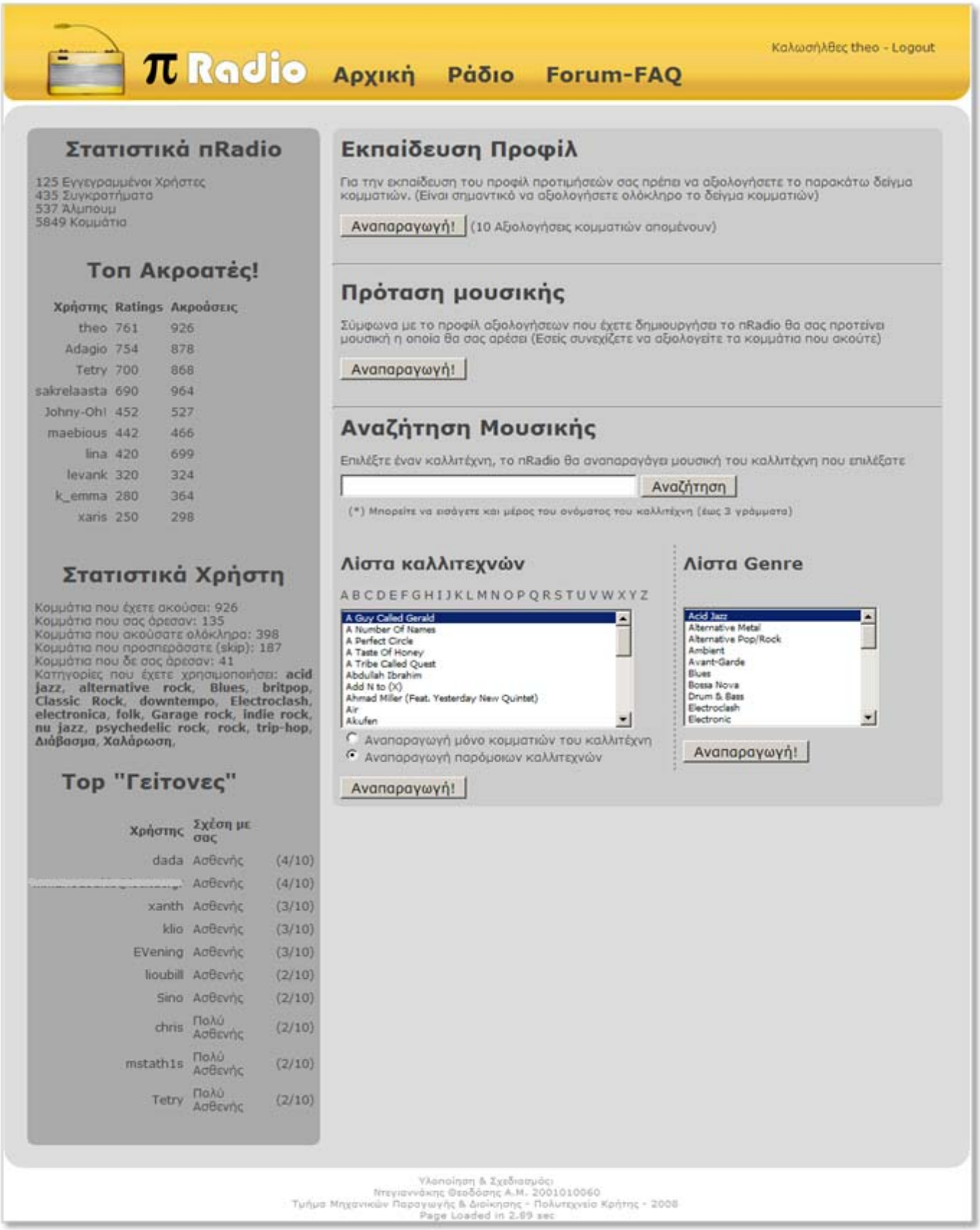
Στατιστικά Ηλεκτρονικού Ραδιοφώνου

Στην περιοχή των στατιστικών εμφανίζονται διάφορα γενικά στοιχεία για την εφαρμογή του ραδιοφώνου. Όπως οι εγγεγραμμένοι χρήστες, τα διαθέσιμα μουσικά κομμάτια.

Επίσης εμφανίζεται και μια λίστα με τους **Τοπ Ακροατές** όπου είναι οι χρήστες που έχουν κάνει τις περισσότερες αξιολογήσεις και ακροάσεις. Σκοπός της ύπαρξής της είναι η παρακίνηση των χρηστών στα πλαίσια ενός άτυπου «συναγωνισμού» με αποτέλεσμα την αύξηση των αξιολογήσεων και των δεδομένων γενικότερα.

Εικόνα 4-4 Στατιστικά Ραδιοφώνου

4.3 Αρχική Σελίδα Χρήστη (Ράδιο)



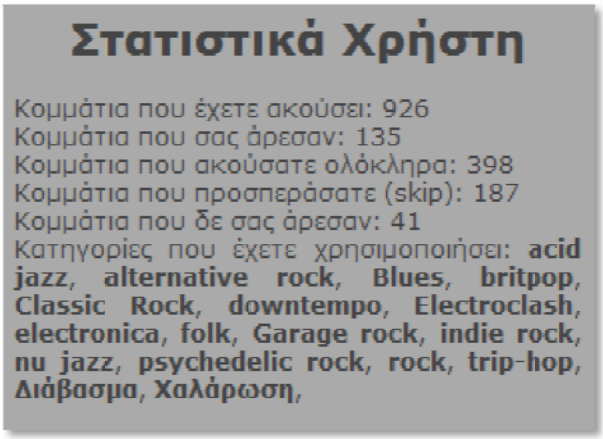
Εικόνα 4-5 Αρχική Σελίδα Χρήστη

Αυτή είναι η πρώτη σελίδα με την οποία έρχεται σε επαφή ο χρήστης μόλις συνδεθεί. Σε αυτή τη σελίδα δίνονται οι εξής πληροφορίες και επιλογές:

Στατιστικά Χρήστη

Από αυτή την λίστα ενημερώνεται ο χρήστης για τα κομμάτια τα οποία έχει αξιολογήσει, καθώς και τις κατηγορίες που έχει χρησιμοποιήσει.

Οδηγούμενος σε διαφορετική σελίδα μπορεί να δει αναλυτικά ποια είναι τα κομμάτια που έχει αξιολογήσει σε κάθε κατηγορία.



Εικόνα 4-6 Στατιστικά Χρήστη

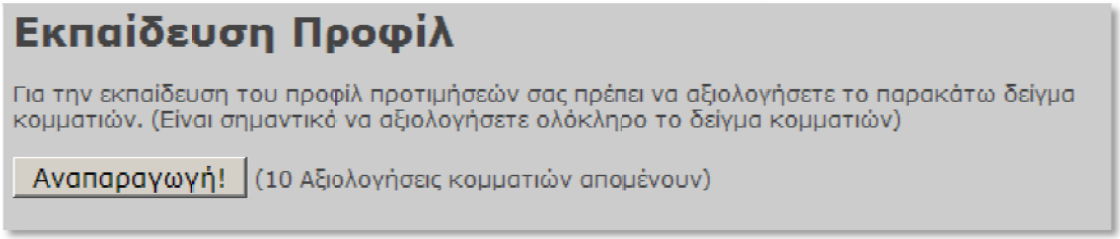
Top «Γείτονες»

Σε αυτή τη λίστα φαίνονται οι γείτονες του ενεργού χρήστη (σύμφωνα με τη θεωρία του collaborative filtering) μαζί με την συσχέτισή τους με τον ενεργό χρήστη σε 10βάθμια κλίμακα.

Εικόνα 4-7 Top "Γείτονες"

Εκπαίδευση Προφίλ

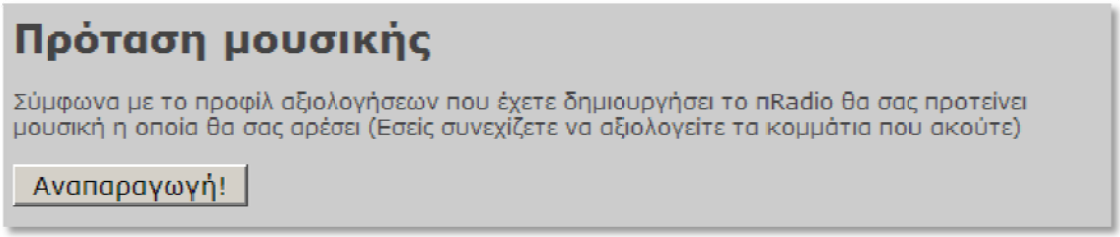
Αυτός είναι ο τομέας εκπαίδευσης του προφίλ του ενεργού χρήστη. Πατώντας το κουμπί Αναπαραγωγή ο χρήστης ακούει το δείγμα εκμάθησης. Επίσης δίπλα στο κουμπί αναγράφεται ο αριθμός των κομματιών που απομένουν για την εκπαίδευση του προφίλ του χρήστη. Αναλυτικά θα εξηγηθεί σε παρακάτω παράγραφο.



Εικόνα 4-8 Εκπαίδευση Προφίλ

Πρόταση Μουσικής

Αυτή η επιλογή γίνεται διαθέσιμη μονάχα όταν το προφίλ του ενεργού χρήστη εκπαιδευτεί και ο αλγόριθμος μπορεί να κάνει αξιόπιστες προτάσεις στο χρήστη. Σε αυτόν τον τομέα θα εφαρμοστούν οι αλγόριθμοι που θα αναλυθούν παρακάτω. Πρόκειται για το «ραδιοφωνικό πρόγραμμα» που παράγει η ιστοσελίδα για τον ενεργό χρήστη.

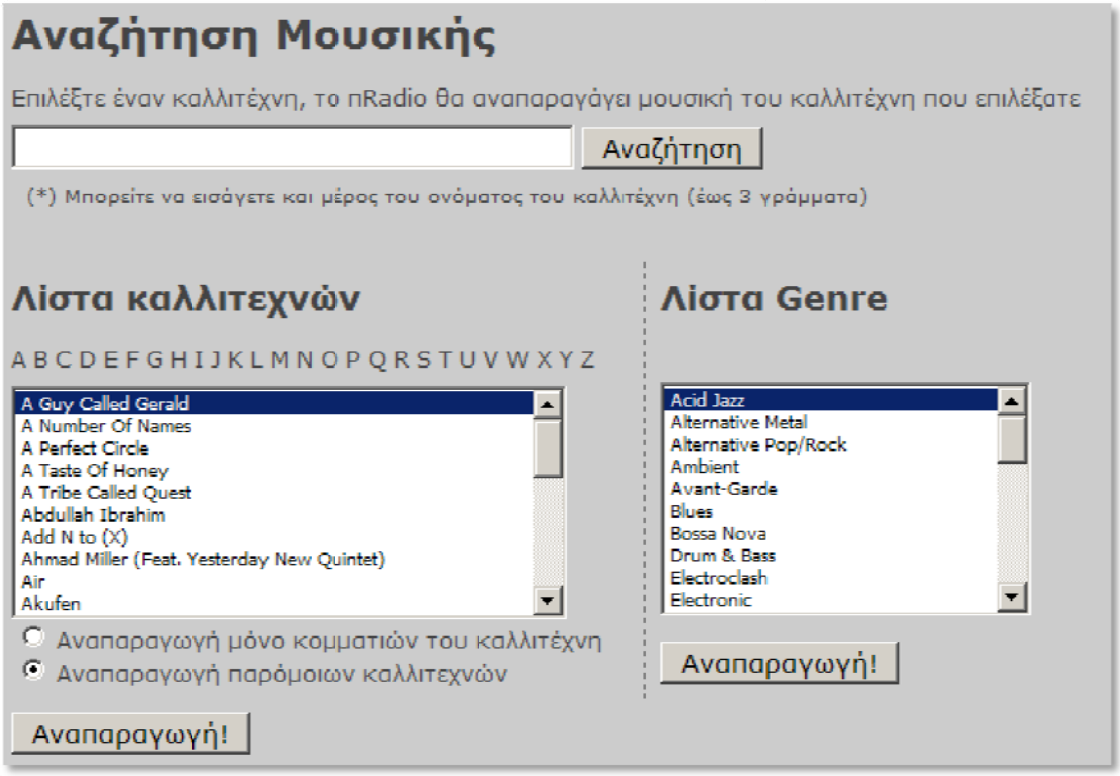


Εικόνα 4-9 Πρόταση Μουσικής

Αναζήτηση Μουσικής

Τέλος δίνεται στον χρήστη η δυνατότητα να ακούσει τη μουσική που αναζητά. Μπορεί να αναζητήσει τον καλλιτέχνη που επιθυμεί είτε γράφοντας το όνομά του είτε αναζητώντας τον από την διαθέσιμη λίστα καλλιτεχνών. Μπορεί ακόμα να επιλέξει ένα είδος μουσικής το οποίο θέλει να ακούσει.

Όπως φαίνεται δίνεται στον χρήστη η δυνατότητα να επιλέξει εάν επιθυμεί να ακούσει κομμάτια μόνο του καλλιτέχνη που επιθυμεί ή και παρόμοιων καλλιτεχνών.

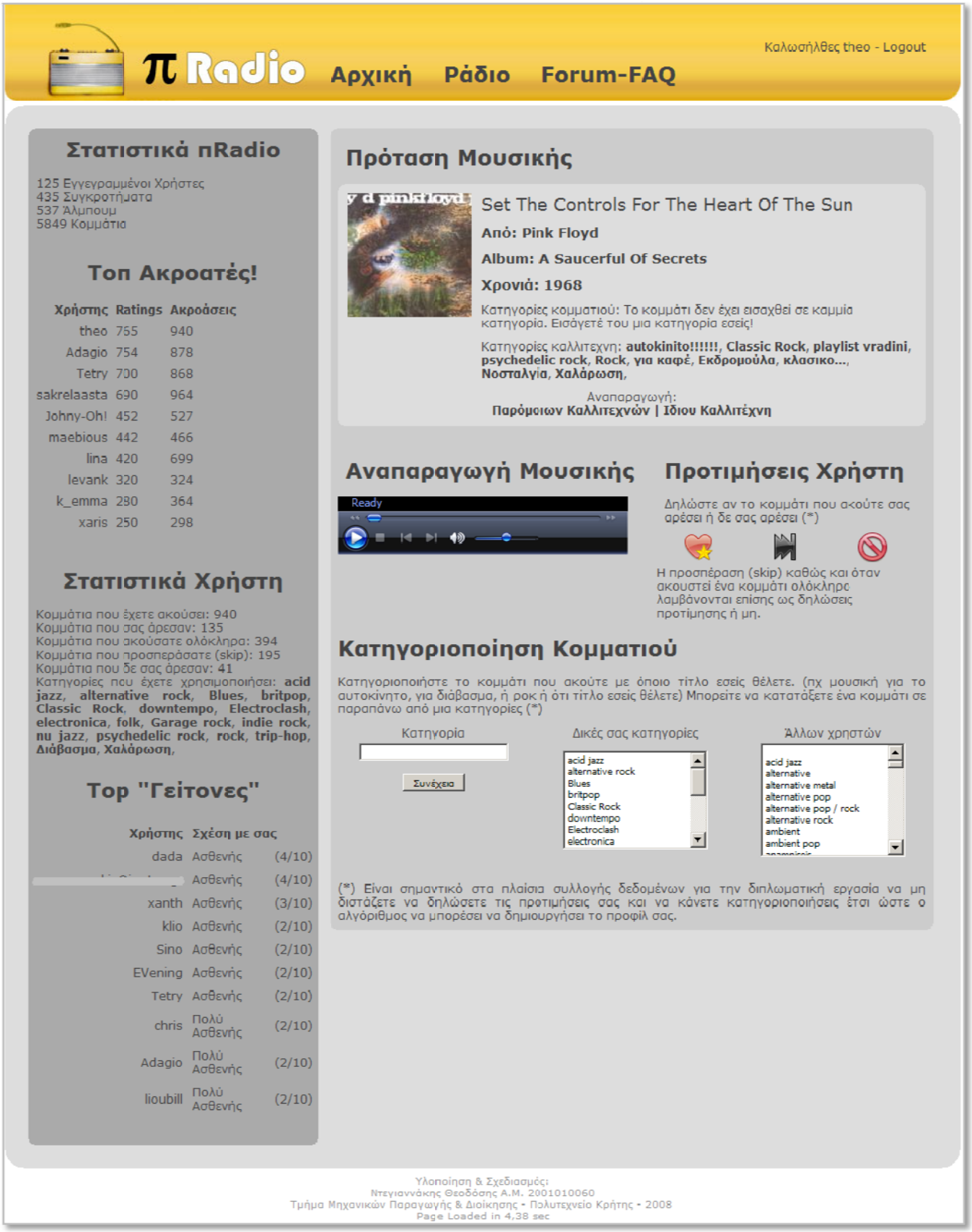


Εικόνα 4-10 Αναζήτηση Μουσικής

Σημείωση

Η τελευταία υπηρεσία που εξηγήθηκε ονομάζεται music on demand (μουσική κατά βούληση) και στις περισσότερες εφαρμογές ηλεκτρονικού ραδιοφώνου δεν είναι διαθέσιμη λόγω περιορισμού εξαιτίας των πνευματικών δικαιωμάτων των καλλιτεχνών. Σε αυτή την κατηγορία εισάγονται μουσικά κομμάτια στα οποία έχει ληφθεί άδεια από τους καλλιτέχνες για αναπαραγωγή κατά βούληση. Στα ραδιόφωνα και γενικά στους ηλεκτρονικούς ραδιοφωνικούς σταθμούς (και αφού ληφθεί βέβαια η αντίστοιχη άδεια) επιτρέπεται η αναπαραγωγή μουσικής χωρίς όμως να δίνεται η δυνατότητα στο χρήστη να ακούει μουσική κατά βούληση.

4.4 Σελίδα Αναπαραγωγής



Εικόνα 4-11 Σελίδα Αναπαραγωγής

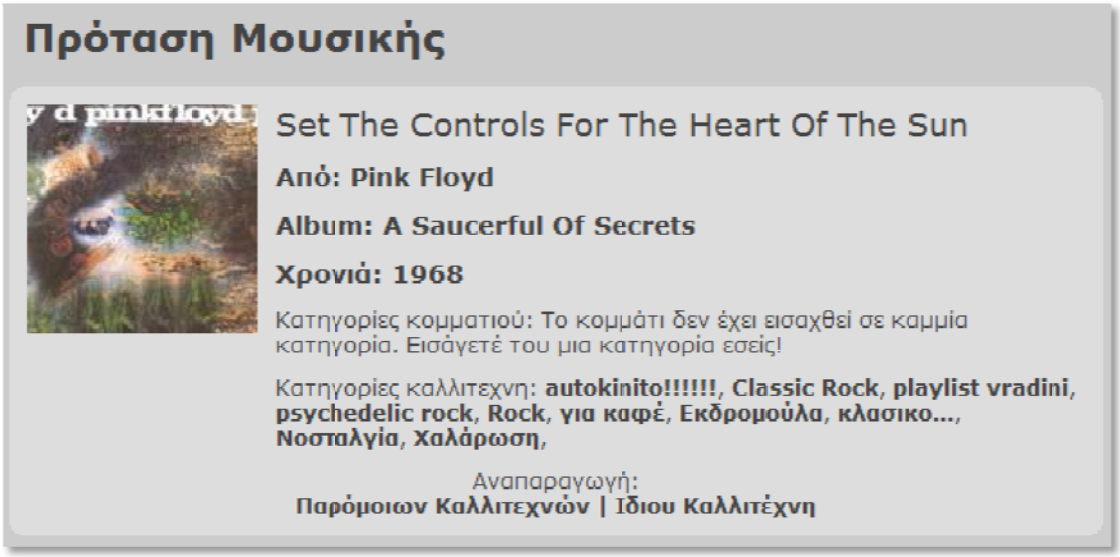
Αυτή είναι η σελίδα αναπαραγωγής της μουσικής. Αναλόγως με την επιλογή που έκανε στην αρχική σελίδα του ο χρήστης σε αυτή τη σελίδα ακούει είτε τη μουσική που έχει επιλέξει, είτε την μουσική για την εκπαίδευση του προφίλ του, είτε την πρόταση μουσικής από τον αλγόριθμο.

Σε αυτή τη σελίδα επίσης γίνεται μέσω των κατάλληλων κουμπιών η συλλογή των δεδομένων αξιολόγησης της μουσικής από το χρήστη.

Συγκεκριμένα η σελίδα αναπαραγωγής περιλαμβάνει:

Πληροφορίες Κομματιού

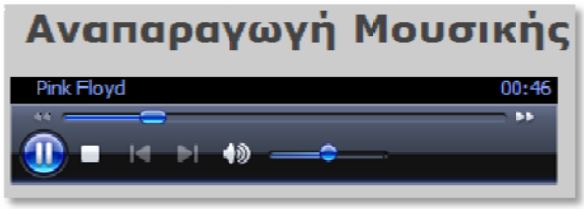
Όπως φαίνεται και από την εικόνα σε αυτό το κομμάτι της σελίδας εμφανίζονται πληροφορίες για το κομμάτι που αναπαράγεται. Πληροφορίες όπως, καλλιτέχνης, άλμπουμ, έτος. Όπως επίσης αναφέρονται και οι κατηγορίες όπου έχουν εντάξει τα κομμάτια οι χρήστες. Τέλος δίνεται η επιλογή στο χρήστη να επιλέξει να φύγει από την παρούσα αναπαραγωγή και να ακούσει περισσότερα κομμάτια από τον ίδιο καλλιτέχνη ή από παρόμοιους.



Εικόνα 4-12 Πληροφορίες Κομματιού

Αναπαραγωγή Μουσικής

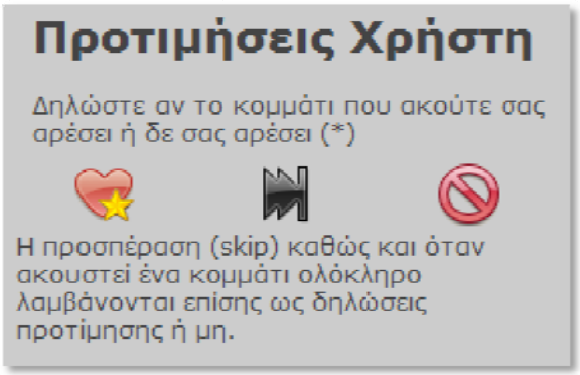
Αυτή είναι η εφαρμογή που αναπαράγει τη μουσική κάνοντας απευθείας σύνδεση (streaming) με τη βάση δεδομένων των μουσικών κομματιών.



Εικόνα 4-13 Αναπαραγωγή Μουσικής

Προτιμήσεις Χρήστη

Με την χρήση αυτών των κουμπιών ο χρήστης δηλώνει την προτίμησή του στο κομμάτι που ακούει. Μπορεί να επιλέξει ότι το κομμάτι του αρέσει, να προχωρήσει στο επόμενο, ή ότι δεν του αρέσει. Επίσης όταν μια επιλογή σημειώνεται με αστερίσκο σημαίνει ότι ο χρήστης έχει ήδη αξιολογήσει το συγκεκριμένο κομμάτι με τη σημειωμένη (με αστέρι) επιλογή, έχοντας βέβαια τη δυνατότητα να αλλάξει την επιλογή του αν το θελήσει.

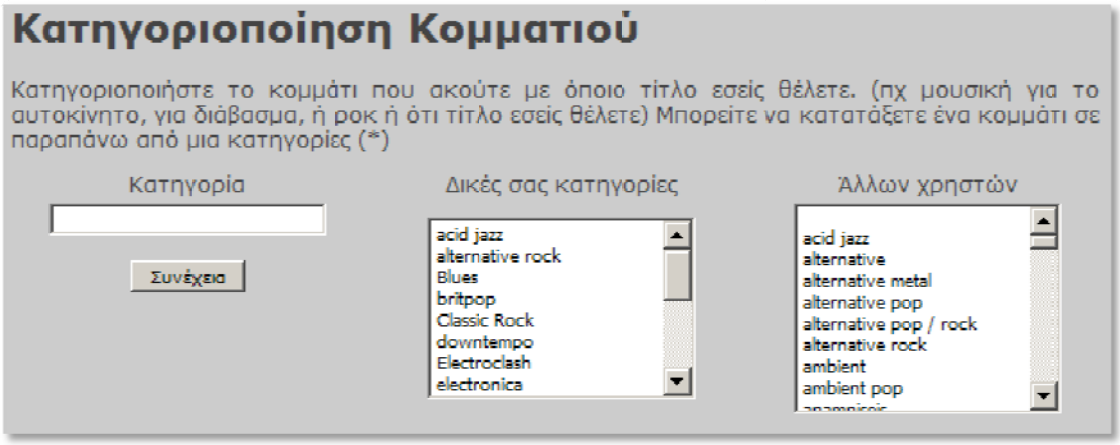


Εικόνα 4-14 Προτιμήσεις Χρήστη

Κατηγοριοποίηση Κομματιού

Τέλος έχουμε τον τομέα κατηγοριοποίησης κομματιών. Εδώ ο χρήστης έχει τη δυνατότητα να δίνει στο κομμάτι που ακούει μια κατηγορία. Οι κατηγορίες που μπορεί ο χρήστης να χρησιμοποιήσει είναι άνευ περιορισμών και εντελώς υποκειμενικές. Μπορεί βέβαια να εισάγει ένα κομμάτι σε παραπάνω από μια κατηγορίες. Έχει δυνατότητα να δει ποιες άλλες κατηγορίες έχει χρησιμοποιήσει παλαιότερα καθώς και ποιες κατηγορίες έχουν χρησιμοποιήσει οι άλλοι χρήστες.

Οι κατηγορίες που εισάγουν οι χρήστες φαίνονται και στην καρτέλα πληροφοριών του κάθε κομματιού.



Εικόνα 4-15 Κατηγοριοποίηση Κομματιού

Σε αυτό το κεφάλαιο έγινε παρουσίαση της σχεδίασης της εφαρμογής που αναπτύχθηκε με ταυτόχρονη αναλυτική περιήγηση σε αυτή. Τέλος στο επόμενο κεφάλαιο θα αναφερθούν τα συμπεράσματα καθώς και οι μελλοντικές ερευνητικές κατευθύνσεις αυτής της εργασίας.

5 Συμπεράσματα & Μελλοντικές Ερευνητικές Κατευθύνσεις

5.1 Συμπεράσματα

Κατά την υλοποίηση του συστήματος του πRadio προέκυψαν διάφορα προβλήματα τα οποία χρειάστηκαν να αντιμετωπιστούν. Ένας σημαντικός περιορισμός ήταν ότι οι υπόλοιπες υλοποιήσεις τέτοιων συστημάτων στο διαδίκτυο είναι εμπορικές εφαρμογές με αποτέλεσμα να μην υπάρχουν δημοσιευμένες αντίστοιχες μεθοδολογίες για εφαρμογή ηλεκτρονικού ραδιοφώνου. Επομένως η υλοποίηση βασίστηκε στις γενικές αρχές των συστημάτων συστάσεων συνεργασίας αντιμετωπίζοντας τους διάφορους περιορισμούς που προέκυψαν ad hoc.

Ένας εξίσου σημαντικός περιορισμός ήταν το δείγμα δεδομένων. Δεν υπήρχε στο διαδίκτυο διαθέσιμο κάποιο δείγμα δεδομένων, με αξιολογήσεις χρηστών σε μουσικά κομμάτια, πάνω στο οποίο θα μπορούσε να δοκιμαστεί το σύστημα που υλοποιήθηκε. Έτσι χρειάστηκε η υλοποίηση να αναλάβει και την συλλογή δεδομένων. Αυτό έγινε προωθώντας την ιστοσελίδα της εφαρμογής σε ομάδες χρηστών οι οποίοι κλήθηκαν (σε ένα διάστημα 4 μηνών) να κάνουν αξιολογήσεις μουσικών κομματιών ούτως ώστε να συλλεχθεί ικανός αριθμός αξιολογήσεων.

Το σύστημα προτάσεων αποδείχτηκε αρκετά ακριβές καθώς η τιμή του συντελεστή Ακρίβειας (Precision) ήταν 0.791 το οποίο σημαίνει ότι το 79% των προτάσεων που γίνονται στους χρήστες είναι πετυχημένες. Η τιμή αυτή ενδέχεται να αυξηθεί και άλλο με την περαιτέρω συλλογή δεδομένων και την αύξηση του μεγέθους του δείγματος δεδομένων.

5.2 Προτάσεις για Μελλοντική Ανάπτυξη

Σε μερικά σημεία της υλοποίησης της εφαρμογής του ηλεκτρονικού ραδιοφώνου διαπιστώθηκαν όμως και κάποια προβλήματα τα οποία ήταν αδύνατον να περιγραφούν αναλυτικά και να επιλυθούν στα πλαίσια της προπτυχιακής αυτής διπλωματικής εργασίας. Παρόλα αυτά όμως υπάρχουν κάποιες προτάσεις για μελλοντική ανάπτυξη και διόρθωση αυτών των προβλημάτων.

5.2.1 Πρόβλημα Νέου Αντικειμένου

Στην παράγραφο 3.5.3 παρατηρήσαμε ότι έγινε αντιμετώπιση του προβλήματος του *Νέου Χρήστη* αλλά το πρόβλημα του *Νέου Αντικειμένου* ήταν αδύνατο να λυθεί δεδομένης της φύσης των συστημάτων Collaborative Filtering.

Υπάρχει όμως μια πρόταση για αντιμετώπιση αυτού του προβλήματος. Η ιδέα αυτή είναι επηρεασμένη από τους Γενετικούς Αλγορίθμους (Genetic Algorithms) και συγκεκριμένα από την φάση της Μετάλλαξης. Εν συντομία οι γενετικοί αλγόριθμοι είναι αλγόριθμοι που αναζητούν βέλτιστες λύσεις προβλημάτων δανειζόμενες τον τρόπο λειτουργίας τους από την φύση. Η φάση της μετάλλαξης στους γενετικούς αλγόριθμους όπως και στη φύση προσδίδει με τυχαίο τρόπο ποικιλομορφία στα αποτελέσματα με σκοπό την βελτιστοποίηση των λύσεων.

Έτσι και στην περίπτωση της εφαρμογής του ηλεκτρονικού ραδιοφώνου προτείνεται η εισαγωγή τυχαιότητας στην πρόταση των κομματιών. Έτσι κατά τον αλγόριθμο που περιγράφεται στην παράγραφο 3.5.5 στην «δεξαμενή κομματιών» εισάγονται τυχαία και κάποια κομμάτια τα οποία δεν έχουν αξιολογηθεί ποτέ από κανέναν. Έτσι εισάγοντας τυχαιότητα στην επιλογή του προτεινόμενου κομματιού, και όχι απλή επιλογή αυτού με τη μεγαλύτερη βαθμολογία, δίνεται η δυνατότητα σε κομμάτια που δεν έχουν ποτέ αξιολογηθεί να προταθούν στο χρήστη. Η πιθανότητα αυτή βέβαια θα είναι μικρή (για να μην υπάρξει απαξίωση των προτάσεων) και θα δίδεται ανάλογα με την εφαρμογή και με το είδος των δεδομένων έπειτα από ανάλυση. Ή ακόμα θα μπορεί να προσδιορίζεται αυτή η πιθανότητα από το χρήστη ρωτώντας τον εάν επιθυμεί να ακούσει νέα κομμάτια ή κομμάτια που έχει ήδη αξιολογήσει.

5.2.2 Scalability

Όπως αναφέρθηκε και στην παράγραφο 3.5.5 ο αλγόριθμος σε κάθε εκτέλεσή του επιλέγει μονάχα το κομμάτι που συγκεντρώνει την μεγαλύτερη προτεινόμενη βαθμολογία και εκτελείται κάθε φορά που προτείνεται ένα κομμάτι στο χρήστη. Αν και ο χρόνος εκτέλεσης του αλγορίθμου για το σύνολο δεδομένων της συγκεκριμένης εφαρμογής (~5000 μουσικά κομμάτια, ~100 χρήστες) ήταν κάτω από 1 sec αυτή η μέθοδος δεν είναι βέλτιστη σε μεγάλη κλίμακα δεδομένων.

Αυτό το οποίο προτείνεται είναι η επιλογή των 20-30 μουσικών κομματιών που συγκεντρώνουν τη μεγαλύτερη προτεινόμενη βαθμολογία αντί ενός. Το πλήθος των κομματιών μπορεί να είναι και μεγαλύτερο ανάλογα με τις ανάγκες της εφαρμογής. Έτσι ο αλγόριθμος θα εκτελείται μονάχα όταν αυτή η λίστα κομματιών αδειάσει ή όταν ο χρήστης εισέρχεται πρώτη φορά στο σύστημα, μειώνοντας το φόρτο εργασίας του υπολογιστή τουλάχιστον 20-30 φορές ανάλογα το πλήθος των κομματιών που επιλέγονται.

5.2.3 Κατηγοριοποίηση Κομματιών

Κατά τη σχεδίαση της εφαρμογής και συγκεκριμένα στη σελίδα αναπαραγωγής (3.2.3) αναφέρθηκε πρώτη φορά η δυνατότητα κατηγοριοποίησης των κομματιών από τους χρήστες. Στα πλαίσια της συλλογής των δεδομένων συμπεριλήφθησαν και

οι κατηγορίες των κομματιών όμως δε συλλέχθηκαν αρκετά δεδομένα για την αλγοριθμική εφαρμογή τους.

Κατά την αξιολόγηση των κομματιών ζητείται από τους χρήστες να κατηγοριοποιήσουν τα κομμάτια που ακούν. Η κατηγοριοποίηση γίνεται ελεύθερα χωρίς να υπάρχουν περιορισμοί. Ο χρήστης μπορεί να χαρακτηρίσει ένα κομμάτι με βάση το είδος (genre) που θεωρεί ότι ανήκε ή ακόμα και να το χαρακτηρίσει σύμφωνα με τη διάθεσή του ή με οποιοδήποτε υποκειμενικό του κριτήριο.

Σκοπός αυτής της λειτουργίας είναι να υπολογιστούν και να υπάρχουν διαθέσιμες για τον αλγόριθμο οι διάφορες κατηγορίες κομματιών με παραπλήσιο περιεχόμενο. Για να επιτευχθεί αυτό υπάρχουν διάφοροι αλγόριθμοι με επικρατέστερο αυτό των αλγορίθμων ομαδοποίησης (clustering). Οι συγκεκριμένοι αλγόριθμοι λαμβάνουν υπόψη τους την εμφάνιση των διαφόρων κομματιών (ή καλλιτεχνών) στις κατηγορίες και έτσι υπολογίζουν η μεταξύ τους συσχέτιση.

Με αυτό τον τρόπο δίνεται η δυνατότητα στον αλγόριθμο να αξιοποιεί και πληροφορίες που έχουν να κάνουν με το περιεχόμενο των μουσικών κομματιών βοηθώντας αρκετά στην αντιμετώπιση των προβλημάτων των collaborative filtering όπως το πρόβλημα του *Νέου Αντικειμένου*.

Βιβλιογραφία

Adomavicius, G., & Kwon, Y. (2007). New Recommendation Techniques for Multicriteria Rating Systems. *IEEE Intelligent Systems , 22* (3), 48-55.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering , 17* (6), 734-749.

Amazon. (n.d.). Ανάκτηση από Amazon: http://www.amazon.com

Approximation Theory. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Approximation_theory

Artificial Intelligence. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Artificial_intelligence

Balabanovic, M., & Shoham, Y. (1997). Fab: Content-Based, Collaborative Recommendation. *Communications of the ACM , 40* (3), 66-72.

Belkin, N., & Croft, B. (1992). Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM , 35* (12), 29-38.

Billsus, D., & Pazzani, M. (2000). User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction , 10* (2-3), 147-180.

Breese, J., Heckerman, D., & Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (σσ. 43-52). San Francisco: Morgan Kaufmann.

Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction , 12* (4), 331-370.

Chen, Y.-H., & George, E. (2000). A Bayesian Model for Collaborative Filtering.

Choice Theory. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Choice_theory

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999). Combining Content-Based and Collaborative Filters in an Online Newspaper. *In Proceedings of ACM SIGIR Workshop on Recommender.*

Cognitive Psychology. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Cognitive_psychology

Cognitive Science. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Cognitive_science

Decision Making. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Decision_making

Genetic Algorithms. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Genetic_algorithms

Hanani, U., Shapira, B., & Shoval, P. (2001). Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction , 11*, 203-259.

Herlocker, J., Konstan, J., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. *Research and Development in Information Retrieval* (σσ. 230-237). Berkeley: American Association of Computing Machinery.

Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS) , 22* (1), 5-53.

iLike. (n.d.). Ανάκτηση από iLike: http://www.ilike.com

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM , 40* (3), 77-87.

Last.fm. (n.d.). Ανάκτηση από Last.fm: http://www.last.fm

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE , 7* (1), 76-80.

Malone, T., Grant, K., Turbak, F., Brobst, S., & Cohen, M. (1987). Intelligent information-sharing systems. *Communications of the ACM , 30* (5), 390-402.

Manouselis, N., & Costopoulou, C. (2007). Analysis and Classification of Multi-Criteria Recommender Systems. *World Wide Web , 10* (4), 415-441.

MovieLens. (n.d.). Ανάκτηση από MovieLens: http://www.movielens.umn.edu

Multi-Criteria Decision Analysis. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Multi-Criteria_Decision_Analysis

Nakamura, A., & Abe, N. (1998). Collaborative Filtering Using Weighted Majority Prediction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (σσ. 395-403). San Francisco: Morgan Kaufmann Publishers Inc.

Netflix. (n.d.). Ανάκτηση από Netflix: http://www.netflix.com

Papagelis, M. (2005). *Crawling the Algorithmic Foundations of Recommendation Technologies.* Heraklion.

Pavlov, D., & Pennock, D. (2002). A Maximum Entropy Approach To Collaborative Filtering in Dynamic, Sparce, High-Dimensional Domains. *Neural Information Processing Systems* .

Pazzani, M. (1999). A Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial Intelligence Review , 13* (5-6), 393-408.

Personalized Marketing. (n.d.). Ανάκτηση από Wikepedia: http://en.wikipedia.org/wiki/Personalized_marketing

Rashid, A. M., Albert, I., Cosley, D., Lam, S., McNee, S., Konstan, J., και συν. (2002). Getting to know you: learning new user preferences in recommender systems. *Proceedings of the 7th international conference on Intelligent user interfaces* (σσ. 127-134). San Francisco: ACM.

Recommender Systems. (n.d.). Ανάκτηση από Wikipedia: http://en.wikipedia.org/wiki/Recommender_system

Resmick, P., & Varian, H. (1997). Recommender systems. *Communications of the ACM , 40* (3), 56-58.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of {ACM} 1994 Conference on Computer Supported Cooperative Work* (σσ. 175-186). North Carolina: ACM.

Rich, E. (1979). User Modeling via Stereotypes. *Cognitive Science , 3*, 329-354.

Salton, G. (1989). *Automatic text processing.* Boston: Addison-Wesley Longman Publishing Co., Inc.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of Recommendation Algorithms for E-Commerce. *Conference on Electronic Commerce* (σσ. 158-169). Minneapolis: ACM.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Application of Dimensionality Reduction in Recommender System A Case Study. *In ACM WebKDD Workshop* .

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th international conference on World Wide Web* (σσ. 285-295). Hong Kong: ACM.

Schein, A., Popescul, A., Ungar, L., & Pennock, D. (2002). Methods and metrics for cold-start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (σσ. 253-260). Tampere: ACM.

Soboroff, I., & Nicholas, C. (1999). Combining Content and Collaboration in Text Filtering. *Proceedings of the IJCAI'99*, (σσ. 86-91). Stockholm.

Tran, T., & Cohen, R. (2000). Hybrid Recommender Systems for Electronic Commerce. *Published in Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00) Workshop on Knowledge-Based Electronic Markets*, (σσ. 78-84).

Ungar, L., & Foster, D. (1998). Clustering Methods for Collaborative Filtering. *Proceedings of the Workshop on Recommendation Systems.* AAAI Press, Menlo Park California.