



**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**«Ανάπτυξη Συστήματος Υπόδειξης με Εφαρμογή
Αλγορίθμων Collaborative και Content-Based Filtering»**

Καραβέλας Πέτρος

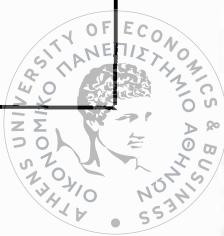
M 3040022

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ**



0 000000 570992

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2006



**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
εισ. 79810
Αρ.
παξ.

«Ανάπτυξη Συστήματος Υπόδειξης με Εφαρμογή
Αλγορίθμων Collaborative και Content-Based Filtering»

Καραβέλας Πέτρος

M 3040022

**Επιβλέπων Καθηγητής: Γεώργιος Δουκίδης
Εξωτερικός Κριτής : Γεώργιος Λεκάκος**



**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2006



ΠΕΡΙΕΧΟΜΕΝΑ



Ευχαριστίες.....	3
Περίληψη.....	4
Executive Summary.....	9
ΚΕΦΑΛΑΙΟ 1 : Εισαγωγή	13
ΚΕΦΑΛΑΙΟ 2 : Επισκόπηση Συστημάτων Υπόδειξης	16
2.1. Ορισμός.....	16
2.2. Βασικά χαρακτηριστικά.....	17
2.2.1. Κλίμακα βαθμολογιών.....	18
2.2.2. Το πρόβλημα της Υπόδειξης	19
2.2.3. Ταξινόμηση Συστημάτων Υπόδειξης	20
2.2.4. Μέτρα Αξιολόγησης.....	21
2.3. Συνεργατική Μέθοδος Υπόδειξης (Collaborative Filtering).....	22
2.3.1. Αλγόριθμοι Βασισμένοι στην Μνήμη (Memory-Based).....	22
2.3.2. Μέτρα Ομοιότητας Χρηστών.....	23
2.3.3. Επιλογή Γειτονιάς.....	25
2.3.4. Υπολογισμός Πρόβλεψης.....	26
2.3.5. Προβλήματα Αλγορίθμων Βασισμένων στη Μνήμη.....	27
2.3.6. Τροποποιήσεις Αλγορίθμων Βασισμένων στη Μνήμη.....	28
2.3.7. Αλγόριθμοι Βασισμένοι σε Μοντέλα (Model Based).....	30
2.4. Υποδείξεις Βασισμένες στο Περιεχόμενο (Content-based).....	31
2.4.1. Αναπαράσταση αντικειμένων.....	31
2.4.2. Σύγκριση αντικειμένων – Παραγωγή Υποδείξεων.....	33
2.5. Σύγκριση ΥΒΠ και ΣΜΥ.....	36
2.6. Υβριδικές Μέθοδοι Υπόδειξης (Hybrid Recommendations).....	38
2.6.1. Συνδυασμός ανεξάρτητων υποδείξεων.....	38
2.6.2. Ενσωμάτωση χαρακτηριστικών ΥΒΠ στη ΣΜΥ.....	39
2.6.3. Ενσωμάτωση χαρακτηριστικών της ΣΜΥ σε ΥΒΠ.....	40



2.6.4. Ενοποιημένο Μοντέλο.....	40
ΚΕΦΑΛΑΙΟ 3 : Σχεδιαστικές Επιλογές.....	42
3.1. Σύνολο δεδομένων (Dataset).....	42
3.2. Διαχείριση Δεδομένων.....	45
3.3. Δομή του Συστήματος.....	46
3.4. Αλγόριθμοι Υπόδειξης.....	47
3.4.1. Συνεργατική Μέθοδος Υπόδειξης.....	47
3.4.2. Υποδείξεις Βασισμένες στο Περιεχόμενο.....	50
3.4.3. Υβριδικές Μέθοδοι Υπόδειξης.....	53
3.5. Αντιμετώπιση του Προβλήματος του Νέου Χρήστη	56
3.6. Γραφική Διεπαφή.....	57
ΚΕΦΑΛΑΙΟ 4 : Σχεδίαση και υλοποίηση συστήματος.....	59
4.1. Σχεδίαση Συστήματος.....	59
4.2. Υλοποίηση Συστήματος.....	65
4.2.1. Γλώσσα Προγραμματισμού και περιβάλλον ανάπτυξης.....	65
4.2.2. Λειτουργία Συστήματος.....	66
ΚΕΦΑΛΑΙΟ 5 : Πειραματική Αξιολόγηση Συστήματος.....	79
5.1. Πειραματικά δεδομένα	79
5.2. Μέτρα Αξιολόγησης	80
5.3. Αξιολόγηση Αλγορίθμων.....	82
5.3.1. Συνεργατική Μέθοδος Υπόδειξης.....	82
5.3.2. Υποδείξεις Βασισμένες στο Περιεχόμενο.....	83
5.3.3. Υβριδική Μέθοδος Υποκατάστασης.....	86
5.3.4. Υβριδική Μέθοδος Αλλαγής.....	87
5.3.5. Πείραμα επιλογής καλύτερης μεθόδου.....	89
ΚΕΦΑΛΑΙΟ 6 : Συμπεράσματα και Μελλοντική Έρευνα.....	91
ΚΕΦΑΛΑΙΟ 7 : Βιβλιογραφικές Αναφορές.....	94



Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Γεώργιο Δουκίδη, καθηγητή του τμήματος Διοικητικής Επιστήμης και Τεχνολογίας του Οικονομικού Πανεπιστημίου Αθηνών και επιστημονικό συντονιστή του Εργαστηρίου Ηλεκτρονικού Επιχειρείν (ELTRUN), ο οποίος ήταν ο υπεύθυνος καθηγητής αυτής της διπλωματικής εργασίας.

Επίσης, θερμές ευχαριστίες οφείλω στον κ. Γεώργιο Λεκάκο, λέκτορα του τμήματος Διοικητικής Επιστήμης και Τεχνολογίας του Οικονομικού Πανεπιστημίου Αθηνών και ερευνητικό μέλος του Εργαστηρίου Ηλεκτρονικού Επιχειρείν, για τον πολύτιμο χρόνο που διέθεσε, την καθοδήγηση και της χρήσιμες συμβουλές του, χωρίς τις οποίες θα ήταν δύσκολο να επιτευχθεί το τελικό αποτέλεσμα, και για την γενικότερη άριστη συνεργασία που είχαμε καθ' όλη την διάρκεια της εκπόνησης της παρούσας εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την ηθική και υλική συμπαράσταση που μου προσέφερε και την υπομονή που έδειξε κατά την εκπόνηση της διπλωματικής εργασίας, αλλά και σε όλη την διάρκεια των μεταπτυχιακών μου σπουδών.



Περίληψη

Σκοπός της παρούσας εργασίας είναι η ανάπτυξη ενός συστήματος υπόδειξης που θα είναι σε θέση να παράγει υποδείξεις χρησιμοποιώντας τις δύο βασικές μεθόδους υπόδειξης καθώς και η πρόταση και υλοποίηση νέων υβριδικών μεθόδων.

Ο Burke (2002) ορίζει τα συστήματα υπόδειξης (recommender systems) ως «κάθε σύστημα που παράγει εξατομικευμένες υποδείξεις ή έχει ως αποτέλεσμα την καθοδήγηση του χρήστη σε ενδιαφέροντα ή χρήσιμα αντικείμενα μέσα σε μεγάλο εύρος πιθανών επιλογών». Οι έννοιες της «εξατομίκευσης» και του «ενδιαφέροντος και χρήσιμου αντικειμένου» είναι αυτές που διαχωρίζουν τα συστήματα υπόδειξης από άλλα συστήματα που υποστηρίζουν διαδικασίες αναζήτησης και προώθησης πληροφοριών, όπως τα συστήματα εξόρυξης γνώσης και τις μηχανές αναζήτησης. Συνήθως, τα συστήματα υπόδειξης αποτελούν τμήματα άλλων συστημάτων, όπως, για παράδειγμα, ηλεκτρονικά καταστήματα. Έχουν εφαρμοστεί επιτυχημένα στα πεδία της υπόδειξης βιβλίων (Barnes & Noble, Amazon.com), κινηματογραφικών ταινιών (MovieLens), μουσικής (CDNow), ανεκδότων (Jester) κ.ά.

Για τον πάροχό τους, τα συστήματα υπόδειξης αποτελούν υπηρεσία προστιθέμενης αξίας και μπορούν να λειτουργήσουν ως πόλος έλξης χρηστών. Ειδικά για τις περιπτώσεις που αποτελούν τμήμα ηλεκτρονικού καταστήματος, μπορούν να επιφέρουν αύξηση των κερδών, μέσω μάρκετινγκ στοχευμένου στο προφύλ του χρήστη και συνδυασμένων πωλήσεων (cross selling), χρησιμοποιώντας τα προϊόντα στο καλάθι αγορών των χρηστών για να παράγουν πρόβλεψη για το ποια άλλα προϊόντα θα τους ενδιαφέρουν. Επιπλέον, αποτελούν παράγοντα ενίσχυσης της πίστης (loyalty) των χρηστών στο κατάστημα και επίτευξης ανταγωνιστικού πλεονεκτήματος.

Η διαδικασία παραγωγής υποδείξεων για έναν χρήστη είναι η εξής: Ο χρήστης αξιολογεί έναν μικρό αριθμό αντικειμένων, συνήθως αναθέτοντας βαθμολογίες σε κάποια συγκεκριμένη κλίμακα (π.χ 1-5). Στη συνέχεια, το σύστημα χρησιμοποιεί την αξιολόγηση του χρήστη σε συνδυασμό με παλαιότερες αξιολογήσεις άλλων χρηστών ή/και δεδομένα σχετικά με τα αντικείμενα του συστήματος και παράγει υποδείξεις που ταιριάζουν στις προτιμήσεις του χρήστη. Αυτός, με τη σειρά του, μπορεί να



αξιολογήσει έναν αριθμό από τα υποδεικνύμενα αντικείμενα και η διαδικασία να ξεκινήσει από την αρχή. Όσο περισσότερα αντικείμενα έχει βαθμολογήσει ο χρήστης, τόσο πιο ακριβείς υποδείξεις λαμβάνει.

Οι δύο βασικές μέθοδοι υπόδειξης που χρησιμοποιούνται είναι η Συνεργατική Μέθοδος Υπόδειξης (SMY - Collaborative Filtering) και οι Υποδείξεις Βασισμένες στο Περιεχόμενο (YBP - Content-based recommendations). Στη πρώτη, στους χρήστες υποδεικνύονται αντικείμενα που έχουν αξιολογηθεί θετικά από χρήστες με παρόμοιες προτιμήσεις, ενώ στη δεύτερη, υποδεικνύονται αντικείμενα που είναι παρόμοια με αυτά που έχει αξιολογήσει θετικά ο ίδιος ο χρήστης στο παρελθόν.

Για την παραγωγή υποδείξεων για έναν χρήστη (ενεργός χρήστης) με την SMY, αρχικά υπολογίζεται η ομοιότητά του με όλους τους υπόλοιπους χρήστες του συστήματος, εφαρμόζοντας κάποιο μέτρο συσχέτισης, συνήθως το Pearson correlation measure. Εν συνεχεία, επιλέγεται το υποσύνολο («γειτονιά») των χρηστών που εμφανίζουν μεγαλύτερη ομοιότητα με τον ενεργό χρήστη. Οι βαθμολογίες των μελών του συνόλου αυτού («γείτονες») συνδυάζονται, χρησιμοποιώντας τη σταθμισμένη απόκλιση από τη μέση τιμή, για να παραχθούν οι προβλέψεις για τον ενεργό χρήστη. Τέλος, παρουσιάζονται στο χρήστη τα αντικείμενα με τις υψηλότερες τιμές πρόβλεψης.

Για την παραγωγή YBP, είναι απαραίτητο τα αντικείμενα να εκφραστούν με έναν αριθμό χαρακτηριστικών γνωρισμάτων (features). Για παράδειγμα, στο πεδίο της υπόδειξης βιβλίων, τέτοια χαρακτηριστικά θα μπορούσαν να είναι ο συγγραφέας, το είδος του βιβλίου και λέξεις του κειμένου που χρησιμοποιούνται συχνά. Για τον εντοπισμό τους συνήθως χρησιμοποιούνται μέτρα όπως το TF-IDF και το Information Gain (IG). Αφού εντοπιστούν τα χαρακτηριστικά γνωρίσματα για όλα τα αντικείμενα, κάθε ένα από αυτά αναπαρίστανται ως διανύσματα που κάθε συντεταγμένη τους δηλώνει την ύπαρξη ή μη ενός χαρακτηριστικού στην περιγραφή του αντικειμένου. Το μέτρο ομοιότητας δύο αντικειμένων είναι το συνημίτονο της γωνίας των διανυσμάτων τους (cosine similarity). Στους χρήστες υποδεικνύονται τα αντικείμενα που εμφανίζουν μεγάλο μέτρο ομοιότητας με τα αντικείμενα που έχουν αξιολογήσει θετικά στο παρελθόν.

Συγκρίνοντας τις δύο μεθόδους, παρατηρούμε ότι οι YBP μπορούν σε εφαρμοστούν μόνο σε πεδία όπου είναι δυνατή η αναπαράσταση των αντικειμένων με χαρακτηριστικά γνωρίσματα, ενώ και η σύγκριση των αντικειμένων γίνεται με μηχανικό τρόπο βασιζόμενη στα γνωρίσματα αυτά. Ο περιορισμός αυτός δεν υπάρχει

στην ΣΜΥ, που μπορεί να εντοπίσει και διαφορές στα αντικείμενα που οφείλονται στο προσωπικό γούστο των χρηστών. Επιπλέον, τα αντικείμενα που υποδεικνύονται με την ΣΜΥ συχνά δεν μοιάζουν με αυτά που έχει αξιολογήσει ο χρήστης ή με τα αντικείμενα προηγούμενων υποδείξεων, αλλά εντούτοις ταιριάζουν στις προτιμήσεις του χρήστη (*serendipity*). Αυτό το εντυπωσιακό χαρακτηριστικό της ΣΜΥ, δεν εμφανίζεται στις YBΠ, όπου συχνά εμφανίζεται το αντίθετο φαινόμενο, δηλαδή η υπερεξιδίκευση των υποδείξεων. Μεγάλο πρόβλημα για την ΣΜΥ αποτελεί η αραιότητα (*sparsity*) των δεδομένων, δηλαδή το γεγονός της βαθμολόγησης πολύ μικρού ποσοστού των αντικειμένων από τους χρήστες. Η αραιότητα έχει ως αποτέλεσμα την μείωση της πιθανότητας το σύστημα να εντοπίσει γείτονες με μεγάλη ομοιότητα, με συνέπεια την αδυναμία παραγωγής υποδείξεων για συγκεκριμένα αντικείμενα ή την παραγωγή προβλέψεων μειωμένης ακρίβειας. Επίσης, προκειμένου να παραχθούν υποδείξεις με την ΣΜΥ είναι απαραίτητη η ύπαρξη μιας κρίσιμης μάζας χρηστών. Ο περιορισμός αυτός και το πρόβλημα της αραιότητας δεν ισχύει για τις YBΠ. Ένα άλλο πρόβλημα με τη ΣΜΥ είναι ότι δεν μπορούν να παραχθούν προβλέψεις για νέα αντικείμενα, προτού κάποιοι χρήστες τα βαθμολογήσουν. Αντίθετα, χρησιμοποιώντας YBΠ παράγονται προβλέψεις για νέα αντικείμενα, αφού αυτές βασίζονται μόνο στην περιεχόμενό τους.

Εκτός από τις δύο παραπάνω μεθόδους υπόδειξης, έχουν αναπτυχθεί υβριδικές (*hybrid*) μέθοδοι, οι οποίες αντιμετωπίζουν τα μειονεκτήματα της μιας μεθόδου με ενσωμάτωση χαρακτηριστικών της άλλης. Συνήθως, χρησιμοποιούνται χαρακτηριστικά της Content-based για να αντιμετωπιστούν τα προβλήματα αραιότητας του Collaborative, χωρίς βέβαια αυτή να είναι η μοναδική προσέγγιση που έχει προταθεί. Τα εμπειρικά αποτελέσματα δείχνουν ότι οι υβριδικές μέθοδοι υπόδειξης αποδίδουν καλύτερα σε σχέση με τις δύο βασικές μεθόδους.

Αν και τα αποτελέσματα της έρευνας στο πεδίο είναι ενθαρρυντικά, αποτελεί κενό το γεγονός ότι δεν έχει αναπτυχθεί μεγάλος αριθμός συστημάτων, ενώ τα σημαντικότερα από τα υπάρχοντα έχουν εμπορικό χαρακτήρα. Στην παρούσα εργασία αναπτύξαμε το δικτυακό σύστημα υπόδειξης MoRe που συνδυάζει τη λειτουργικότητα με την δυνατότητα πειραματισμού και εξαγωγής συμπερασμάτων. Στα πλαίσια της ανάπτυξής του, προτείναμε τροποποιήσεις της ΣΜΥ και των YBΠ με στόχο την βελτίωση της ακρίβειας των υποδείξεων. Επίσης, προτείναμε δύο υβριδικές μεθόδους υπόδειξης, την Υβριδική Μέθοδο Αλλαγής και την Υβριδική Μέθοδο Υποκατάστασης. Το σύστημα MoRe είναι σε θέση να παράγει υποδείξεις

χρησιμοποιώντας και τις τέσσερις προαναφερθείσες μεθόδους. Επίσης, υλοποιήσαμε έναν πρόγραμμα αναζήτησης δεδομένων από το διαδίκτυο (web crawler), το οποίο αναζητά στον ιστότοπο της IMDb (Internet Movie Database) τα στοιχεία για τις ταινίες (σκηνοθέτης, ηθοποιοί, είδος, λέξεις σχετικές με την υπόθεση κ.ά.) που χρησιμοποιούνται για την παραγωγή Υποδείξεων Βασισμένων στο Περιεχόμενο.

Για την ΣΜΥ, προτείναμε την παραγωγή υποδείξεων μόνο εάν η «γειτονιά» του ενεργού χρήστη αποτελείται τουλάχιστον από πέντε «γείτονες». Ο περιορισμός αυτός στοχεύει στην βελτίωση της ακρίβειας των παραγόμενων υποδείξεων. Σε ότι αφορά τις YBΠ, υπολογίζουμε τις ομοιότητες των ταινιών πριν τη φάση λειτουργίας του συστήματος για γρηγορότερη παραγωγή υποδείξεων. Επειδή το σύνολο των ταινιών δεν αλλάζει δυναμικά κατά την φάση λειτουργίας του συστήματος, η παραπάνω προσέγγιση δεν έχει επίπτωση στην ακρίβεια των υποδείξεων. Επιπλέον, υλοποιήσαμε μια εναλλακτική μέθοδο YBΠ που χρησιμοποιεί τον κατηγοριοποιητή Naïve Bayes. Με την μέθοδο αυτή, αντί να παράγονται προβλέψεις για κάθε ταινία, γίνεται προσπάθεια υπολογισμού της πιθανότητας η ταινία να βαθμολογηθεί με κάθε τιμή της βαθμολογικής κλίμακας. Ως πρόβλεψη για την ταινία ανατίθεται η τιμή με την μεγαλύτερη πιθανότητα.

Επειδή η ΣΜΥ παράγει λιγότερο ακριβείς υποδείξεις όταν αυτές βασίζονται σε μικρό αριθμό βαθμολογήσεων, στην Υβριδική Μέθοδο Αλλαγής (Switching), για όσο ο χρήστης έχει βαθμολογήσει μικρό αριθμό ταινιών (π.χ μικρότερο από 30), παράγονται YBΠ, ενώ όταν ο χρήστης βαθμολογήσει περισσότερες ταινίες παράγονται υποδείξεις με τη ΣΜΥ. Στην Υβριδική Μέθοδο Υποκατάστασης (Substitute), παράγονται YBΠ όταν η ΣΜΥ δεν είναι σε θέση να παράγει προβλέψεις λόγω του φαινόμενου της αραιότητας. Με τον τρόπο αυτό παράγονται προβλέψεις σε όλες τις περιπτώσεις.

Έχοντας ολοκληρώσει την υλοποίηση του συστήματος, συνεχίσαμε με την αξιολόγησή του. Τα πειραματικά δεδομένα που χρησιμοποιήθηκαν προέρχονται από βαθμολογήσεις των χρηστών του συστήματος MovieLens για το έτος 2000. Η μεθοδολογία που χρησιμοποιήθηκε είναι η εξής: Το 20% των βαθμολογιών κάθε χρήστη (σύνολο ελέγχου) επιλέγεται με τυχαίο τρόπο και αφαιρείται από το σύστημα. Το υπόλοιπο 80% (σύνολο εκπαίδευσης) χρησιμοποιείται για την παραγωγή προβλέψεων για το σύνολο ελέγχου. Οι μετρικές με τις οποίες έγινε η σύγκριση των μεθόδων είναι η Κάλυψη, που εκφράζει το ποσοστό των αντικειμένων για τα οποία το σύστημα μπορεί να κάνει υποδείξεις και το Μέσο Απόλυτο Λάθος (Mean Absolute

Επτορ - ΜΑΕ), που αποτελεί μέτρο ακρίβειας των υποδείξεων συγκρίνοντας τις αριθμητικές προβλέψεις του συστήματος με τις πραγματικές βαθμολογήσεις των χρηστών. Οι τιμές ΜΑΕ των μεθόδων συγκρίθηκαν ανά δύο χρησιμοποιώντας το τεστ Wilcoxon στο διάστημα εμπιστοσύνης 99% για εντοπισμό στατιστικά σημαντικών διαφορών.

Τα αποτελέσματα της αξιολόγησης καταδεικνύουν την σαφή υπεροχή της ΣΜΥ σε σχέση με τις ΥΒΠ σε ότι αφορά την ακρίβεια των υποδείξεων. Το μέγεθος της «γειτονιάς» του ενεργού χρήστη, κατά τον υπολογισμό πρόβλεψης με την ΣΜΥ για ένα συγκεκριμένο αντικείμενο, επηρεάζει την ακρίβεια.

Στις ΥΒΠ η ακρίβεια των υποδείξεων είναι ανάλογη του αριθμού των χαρακτηριστικών με τα οποία περιγράφονται τα αντικείμενα. Όσο αυξάνεται ο αριθμός τους, τόσο βελτιώνεται η ακρίβεια των υποδείξεων. Η προσέγγιση που χρησιμοποιεί Naive Bayes δεν κατάφερε να παράγει ακριβείς προβλέψεις. Η εξήγηση είναι ότι οι βαθμολογίες των χρηστών δεν κατανέμονται ομοιόμορφα στην κλίμακα βαθμολογιών (οι χρήστες τείνουν να βαθμολογούν κυρίως τις ταινίες που προτιμούν), με αποτέλεσμα ο κατηγοριοποιητής να μην έχει αρκετά παραδείγματα εκπαίδευσης για όλες τις τιμές της βαθμολογικής κλίμακας.

Η Υβριδική Μέθοδος Υποκατάστασης βελτιώνει την ακρίβεια των υποδείξεων χωρίς να θυσιάζει κάλυψη αντικειμένων. Η Υβριδική Μέθοδος Αλλαγής παράγει λιγότερο ακριβείς υποδείξεις και επιτυγχάνει μικρότερη κάλυψη σε σχέση με την Υβριδική Μέθοδο Υποκατάστασης αλλά έχει μικρότερο χρόνο εκτέλεσης.

Executive Summary

The aim of this thesis is the development of a recommender system that is able to produce recommendations using the two main recommendation methods and, also, the suggestion and implementation of new hybrid methods.

Burke (2002) defines recommender systems as “any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options”. It is the criteria of ‘individualized’ and ‘interesting and useful’ that separate recommender systems from other systems that provide means to search and distribute information, such as information retrieval systems or search engines. Usually, a recommender system is a component of a broader system, e.g. electronic shops. They have been successfully used for recommending books (Barnes & Noble, Amazon.com), movies (MovieLens), music (CDNow), jokes (Jester) and newsletter (GroupLens).

From the provider’s point of view, recommender systems could be regarded as value added service that can attract new users. Especially when they are components of an e-shop, their existence may increase the profits, through cross-selling (using the products in a customer’s basket to produce prediction on which other products would be interesting for him) and targeted marketing (using the customer’s profile that has been created). Moreover, they reinforce customer loyalty to the e-shop and, so, they can be used to achieve advantage over the competitors.

The procedure for producing recommendations for a user is the following: The user evaluates a small number of items, often by using ratings in a specific scale (e.g. 1-5). This evaluation is used by the system, combined with other users’ evaluations and/or data about available items, in order to produce recommendations that match the preferences of the user. Then, the latter may start the procedure all over again by evaluating some of the recommender items. The more items a user evaluates, the more accurate the recommendations are.

The two main methods that are used by recommender systems are Collaborative Filtering and Content-based recommendations. In collaborative



Executive Summary

filtering, users are recommended items that have been positively rated by users of similar preferences, while in content-based recommendation users are recommended items that are similar to those they have already rated positively.

For producing recommendations for a user using collaborative filtering, the system calculates his similarity with every other user, by calculating their correlation, usually applying Pearson correlation measure. Then, the group (neighborhood) of the users that are the most similar to the active user is selected and their ratings are combined, using a weighted deviation from average mean, in order to produce predictions for the active user. Finally, the items with the highest prediction values are presented to the user.

In order to produce content-based recommendations, items have to be expressed with some features. For example, in the book recommendation domain, the writer, the genre and the most frequently used words could serve as features. Metrics such as TF-IDF and Information Gain (IG) are commonly used to extract these features. Since the features are selected, every item is represented by a vector, the elements of which indicate the existence or non-existence of a specific feature in the item's description. The similarity of two items is measured by the cosine similarity of their vectors. The active user is recommended with items that have high cosine similarity with the items he has previously rated highly.

Comparing the above methods, we remark that content-based recommendations can only be applied in domains where it is possible to extract features from the items, because the calculation of similarity among items is based on these features. This restrain does not exist in collaborative filtering. Moreover, content-based recommendations often become overspecialized, since the system recommends more of what the user already has indicated liking. On the contrary, collaborative filtering provides valuable, but unexpected recommendations that impress the users (serendipity). However, collaborative filtering has shortcomings of its own, the most important of which is data sparsity. Users typically rate an extremely small fraction of the system's item set and as a result it is often difficult to find users which have co-rated at least some items. This leads to low-accuracy predictions or even to failure to make predictions. Additionally, collaborative filtering can not be used unless there are a significant number of users' ratings available. Another problem is that whenever a new item is added to the system it can not be recommended before some user rate it. These limitations do not exist in content-based

Executive Summary

recommendations, since the number of the system's users has nothing to do with predictions calculation.

Apart from the two methods mentioned above, there have been deployed hybrid recommendation methods which use features of a method to overcome the shortcomings of the other. In the most common, but not unique, approach, content-based characteristics are used in order to deal with sparsity-related problems of collaborative filtering. Experimental results indicate that hybrid methods produce more accurate recommendation than "pure" collaborative or "pure" content-based methods.

Even though the results of research in the domain are encouraging, there are only few fully developed systems, more of which are business-oriented. In this thesis, we have developed a web-based movie recommender system, which we call MoRe, and that combines functionality with the ability to be used experimentally in order to come into conclusions. For the purposes of its development, we have suggested modifications of content-based and collaborative filtering, aiming to improve recommendation accuracy. In addition, we have suggested and implemented two new hybrid methods, Switching Hybrid and Substitute Hybrid. Our system is able to produce recommendations using all four methods. Apart from the main system, we have developed a web-crawler software that searches the IMDb (Internet Movie Database) website and extracts information about the system's movies (director, cast, genre, plot words etc) which are used by the content-based method.

In our implementation of collaborative filtering, we only produce predictions if the active user's neighbourhood contains at least five neighbours. This restriction aims to improve the produced recommendations. In content-based recommendations, movie similarities are calculated offline and so the predictions are calculated much faster. This does not compromise the accuracy of the recommendations, since the movies set does not change dynamically while the system is online. We also implemented an alternative approach for content-based recommendations that uses the Naive Bayes classifier. In this approach, instead of producing predictions for every movie, we calculate the possibility for this movie to be rated by the user with each value of the rating scale. The value with the highest possibility is considered to be the systems prediction.

Since collaborative filtering produces inaccurate predictions when they are based in few ratings, Switching hybrid produces content-based recommendation as



Executive Summary

long as the user has rated less than a specific number of movies (e.g. 30 movies). After he has reached this barrier, the recommendations are produced using collaborative filtering. Substitute hybrid uses content-based recommendations when collaborative filtering is unable to do so due to the sparsity problem. In that way, recommendations are always produced.

Having completed the development of the system, we continued to evaluate it. We used a dataset that contains a million ratings from the users of MovieLens recommender system that were submitted during the year 2000. Our methodology was the following: We randomly remove 20% of each user's ratings (test set) and then we produce predictions for them using the remaining 80%. The metrics that we used to compare the recommendation methods were Coverage, that expresses the percentage of the items for which the system successfully produced predictions, and Mean Absolute Error (MAE), which is accuracy measure that compares system's predictions with the actual user ratings. We compared the MAE value of all recommendations in pairs using the Wilcoxon test in order to verify that the differences were statistically significant.

The results of the evaluation clearly show that collaborative filtering is superior to content-based recommendations in terms of accuracy. Active user's neighborhood size during prediction calculation in collaborative filtering has an effect in the accuracy of the recommendations. Usually, the greater the size of the neighborhood, the more accurate the recommendations are.

The accuracy of content-based recommendations is analogue to the number of features that describe the items. The approach using Naive Bayes failed to produce accurate predictions. This is due to the fact that users tend to rate mainly the movies that like rather than movies that dislike and as a result the classifier does not have enough examples of negative ratings to be trained with.

Substitute Hybrid improves the accuracy of recommendations without sacrificing item coverage. Switching Hybrid produces less accurate recommendations and achieves smaller coverage than Substitute hybrid, but is executed faster.



1. Εισαγωγή

Στη σύγχρονη εποχή ο όγκος των πληροφοριών που είναι διαθέσιμος για ένα άτομο ξεπερνά κατά πολύ τις φυσικές δυνατότητές του να τον προσπελάσει και να τον αξιοποιήσει. Το φαινόμενο αυτό είναι ακόμα πιο έντονο στα ηλεκτρονικά περιβάλλοντα, ιδιαίτερα αν αναλογιστούμε την συνεχώς αυξανόμενη χρήση του διαδικτύου ως πηγή ανεύρεσης πληροφοριών, υπηρεσιών και προϊόντων. Τα συστήματα υπόδειξης ανήκουν στην κατηγορία εκείνη των τεχνολογιών που έχουν ως σκοπό τον εντοπισμό χρήσιμων πληροφοριών ή αντικειμένων, μέσα από ένα κυκεώνα διαθέσιμων επιλογών. Κύριο χαρακτηριστικό τους είναι ότι παρέχουν προσωποποιημένες υποδείξεις, κάτι που τα διαχωρίζει από άλλες τεχνολογίες που προσπαθούν να αντιμετωπίσουν το πρόβλημα του πληροφοριακού υπερκορεσμού (information overload), όπως για παράδειγμα τις μηχανές αναζήτησης και τα συστήματα εξόρυξης γνώσης. Τα συστήματα υπόδειξης προσπαθούν να εντοπίσουν τα αντικείμενα που θα ενδιαφέρουν τους χρήστες χρησιμοποιώντας δεδομένα σχετικά με το προφίλ των χρηστών.

Αν και η εφαρμογή της τεχνολογίας αυτής μπορεί να γίνει σε οποιουδήποτε είδους πληροφοριακά αντικείμενα, έχουν εφαρμοστεί κυρίως για να προτείνουν αντικείμενα που τα κριτήρια επιλογής τους σχετίζονται με το προσωπικό γούστο των χρηστών, όπως βιβλία, μουσική, κινηματογραφικές ταινίες και εστιατόρια. Για να το πετύχουν αυτό εφαρμόζουν στατιστικές μεθόδους και πρακτικές που προέρχονται από τα πεδία της μηχανικής μάθησης και της εξόρυξης γνώσης. Συνήθως, τα συστήματα υπόδειξης αποτελούν τμήμα μεγαλύτερων συστημάτων. Ιδιαίτερη επιτυχία έχουν γνωρίσει στα πλαίσια ηλεκτρονικών καταστημάτων (π.χ. το γνωστό αμερικανικό βιβλιοπωλείο Amazon), όπου χρησιμοποιούνται από χιλιάδες καταναλωτές. Ο εμπλουτισμός ενός ηλεκτρονικού καταστήματος με ένα σύστημα υπόδειξης ευνοεί τόσο τον χρήστη-πελάτη, όσο και το κατάστημα-πωλητή. Ο πελάτης μπορεί να εντοπίζει ενδιαφέροντα προϊόντα γι' αυτόν εύκολα και απλά. Ο πωλητής μπορεί να διαθέσει περισσότερα προϊόντα, ακόμα και αυτά που δεν είναι τόσο δημοφιλή, και μάλιστα χωρίς να πρέπει να τα διαφημίσει σε ευρύ φάσμα καταναλωτών. Μια τακτική στοχευμένου μάρκετινγκ στους χρήστες που, σύμφωνα



με το σύστημα υπόδειξης, θα ενδιαφέρονται για τα συγκεκριμένα προϊόντα θα ήταν πιο αποδοτική και οικονομική. Επιπλέον, η ύπαρξη ενός τέτοιου συστήματος σε ένα ηλεκτρονικό κατάστημα μπορεί να αποτελέσει και παράγοντα ενίσχυσης της πίστης (loyalty) των πελατών προς το κατάστημα.

Η περισσότερο ώριμη και επιτυχημένη τεχνική υπόδειξης είναι η συνεργατική μέθοδος υπόδειξης (collaborative filtering). Η μέθοδος αυτή προσπαθεί να εντοπίσει τις προσωπικές προτιμήσεις των χρηστών μελετώντας την ομοιότητα των επιλογών τους με αυτές των υπολοίπων χρηστών. Ένα προϊόν προτείνεται σε ένα χρήστη επειδή άρεσε σε άλλους χρήστες με παρόμοιο γούστο. Κυριότερο πλεονέκτημα της μεθόδου αυτής είναι ότι μπορεί να εφαρμοστεί σε οποιοδήποτε είδος αντικειμένων και προσομοιώνει μία καθημερινή ανθρώπινη πρακτική που είναι η ανεύρεση προϊόντων και πληροφοριών που μας ενδιαφέρουν μετά από παρότρυνση κάποιου ατόμου με το οποίο έχουμε κοινές προτιμήσεις. Μειονέκτημά της είναι ότι η εφαρμογή της απαιτεί μεγάλο πληθυσμό χρηστών, έτσι ώστε να είναι εύκολο να βρεθούν χρήστες με κοινές προτιμήσεις.

Η δεύτερη τεχνική έχει ως στόχο την ανεύρεση ομοιοτήτων ανάμεσα στα αντικείμενα για τα οποία έχει εκφράσει κάποια προτίμηση ο χρήστης είτε αγοράζοντάς τα είτε αξιολογώντάς τα θετικά. Ένα αντικείμενο προτείνεται στον χρήστη επειδή είναι παρόμοιο με άλλα προϊόντα που του άρεσαν. Η τεχνική αυτή, για να παράγει υποδείξεις, χρησιμοποιεί τεχνικές που προέρχονται από το πεδίο της Ανάκτησης Πληροφοριών. Οι υποδείξεις που παράγονται ονομάζονται Υποδείξεις Βασισμένες στο Περιεχόμενο (content-based recommendations). Πλεονέκτημα της μεθόδου είναι η δυνατότητα εφαρμογής της ακόμα και όταν είναι διαθέσιμο μικρό πλήθος χρηστών, ενώ περιοριστικός παράγοντας για την εφαρμογή της είναι το γεγονός ότι τα αντικείμενα πρέπει να εκφραστούν σε μια μορφή που μπορεί να υποστεί αυτοματοποιημένη επεξεργασία, κάτι που δεν είναι πάντα εφικτό.

Εκτός από τις δύο παραπάνω τεχνικές, υπάρχουν προσεγγίσεις που προσπαθούν να αντιμετωπίσουν τα μειονεκτήματα της κάθε μίας μεθόδου με χρησιμοποίηση στοιχείων της άλλης ώστε να παράγουν πιο επιτυχημένες υποδείξεις. Αυτές οι μέθοδοι ονομάζονται Υβριδικές και έχουν ήδη εφαρμοστεί με επιτυχία.

Ενώ η επιστημονική έρευνα έχει οδηγήσει σε σημαντικές βελτιώσεις της ακρίβειας των υποδείξεων, ο αριθμός των ολοκληρωμένων συστημάτων υπόδειξης παραμένει μικρός. Αρκετοί ερευνητές βασίζονται σε ιστορικά δεδομένα χρηστών για να αναπτύξουν και να δοκιμάσουν τις βελτιώσεις που προτείνουν, χωρίς, όμως, να τις

Κεφάλαιο 1 : Εισαγωγή

ενσωματώνουν σε πλήρη συστήματα. Επιπλέον, τα σημαντικότερα από τα υπάρχοντα συστήματα υπόδειξης έχουν εμπορικό χαρακτήρα και δεν ενδείκνυνται για πειραματισμό και εξαγωγή γενικών συμπερασμάτων. Άλλωστε, ελάχιστα συστήματα χρησιμοποιούν περισσότερες της μιας μεθόδους υπόδειξης, κάτι που θα επέτρεπε την σύγκριση των διαφορετικών μεθόδων στο ίδιο περιβάλλον. Είναι υπαρκτή, λοιπόν, η ανάγκη ανάπτυξης ολοκληρωμένων συστημάτων υπόδειξης που θα συνδυάζουν τη λειτουργικότητα με την δυνατότητα πειραματισμού και εξαγωγής συμπερασμάτων.

Στην παρούσα εργασία αναπτύσσουμε ένα σύστημα υπόδειξης για το πεδίο των κινηματογραφικών ταινιών με το όνομα MoRe (Movie Recommender). Στόχος της εργασίας είναι η υλοποίηση των δύο βασικών μεθόδων, η προσπάθεια βελτίωσης τους και η δημιουργία μεθόδων υβριδικού συνδυασμού τους.

Στο Κεφάλαιο 2 γίνεται επισκόπηση του πεδίου και παρουσιάζονται οι ερευνητικές προσπάθειες και τα αποτελέσματά τους, όπως αυτά παρουσιάζονται στη διεθνή βιβλιογραφία. Ιδιαίτερο βάρος δίνεται στους χρησιμοποιούμενους αλγορίθμους από τις βασικές μεθόδους υπόδειξης και εμβαθύνουμε στα στοιχεία που έχουν υλοποιηθεί στα πλαίσια του συστήματος.

Στο Κεφάλαιο 3 διατυπώνονται και δικαιολογούνται οι σχεδιαστικές αποφάσεις που ελήφθησαν για την υλοποίηση του συστήματος από το επόπεδο των αλγορίθμων ως την γραφική διεπαφή. Επίσης περιγράφονται βελτιώσεις της Συνεργατικής Μεθόδου Υπόδειξης και των Υποδείξεων Βασισμένων στο Περιεχόμενο και προτείνονται δύο τρόποι υβριδικού συνδυασμού τους που αναπτύσσονται ως ξεχωριστές μέθοδοι υπόδειξης.

Το κεφάλαιο 4 περιλαμβάνονται στοιχεία σχετικά με την σχεδίαση του συστήματος καθώς και οι τεχνολογικές λεπτομέρειες της ανάπτυξής. Επίσης περιγράφεται ο τρόπος λειτουργίας και οι δυνατότητες του συστήματος μέσω της παρουσίασης των βασικών του οθονών.

Στο κεφάλαιο 5 γίνεται λόγος για την πειραματική αξιολόγηση του συστήματος. Περιγράφονται τα μέτρα ποιότητας που χρησιμοποιούνται, η μεθοδολογία που εφαρμόζεται και αναφέρονται τα αποτελέσματα της αξιολόγησης μαζί με το σχολιασμό τους.

Τέλος, στο κεφάλαιο 6 γίνεται ανασκόπηση της εργασίας, συγκέντρωση των συμπερασμάτων και διατύπωση των ανοικτών θεμάτων για μελλοντική έρευνα.

2. Επισκόπηση Συστημάτων Υπόδειξης

2.1 Ορισμός

Τα Συστήματα Υπόδειξης αναδείχθηκαν ως ανεξάρτητο ερευνητικό πεδίο στα μέσα της δεκαετίας του 1990 και έχουν τις ρίζες τους, κυρίως, στα πεδία της Μηχανικής Μάθησης (Machine Learning), της Ανάκτησης Πληροφοριών (Information Retrieval) και της Τεχνητής Νοημοσύνης (Artificial Intelligence). Ο Burke (2002) τα ορίζει ως «κάθε σύστημα που παράγει εξατομικευμένες υποδείξεις ή έχει ως αποτέλεσμα την καθοδήγηση του χρήστη σε ενδιαφέροντα ή χρήσιμα αντικείμενα μέσα σε μεγάλο εύρος πιθανών επιλογών».

Οι έννοιες της «εξατομίκευσης» και του «ενδιαφέροντος και χρήσιμου αντικειμένου» είναι αυτές που διαχωρίζουν τα συστήματα υπόδειξης από άλλα συστήματα που υποστηρίζουν διαδικασίες αναζήτησης και προώθησης πληροφοριών, όπως, για παράδειγμα, συστήματα εξόρυξης γνώσης και τις μηχανές αναζήτησης, όπου οι χρήστες υποβάλλουν κάποια συγκεκριμένα κριτήρια (π.χ. λέξεις κλειδιά) και επιστρέφονται όσα στοιχεία ταιριάζουν με αυτά τα κριτήρια. Αντίθετα, τα συστήματα υπόδειξης εντοπίζουν ποια στοιχεία είναι ενδιαφέροντα και χρήσιμα για τον συγκεκριμένο χρήστη χωρίς αυτός να δηλώσει άμεσα τα ενδιαφέροντα και τις ανάγκες του. Αποτελούν, λοιπόν, ένα μέσο εκτέλεσης προσωποποιημένης αναζήτησης.

Σε πληροφοριακά περιβάλλοντα, όπου το πλήθος των διαθέσιμων επιλογών καθιστούν δύσκολη έως αδύνατη την εξαντλητική προσπέλασή τους από τους χρήστες, τα συστήματα υπόδειξης προσφέρουν μια γρήγορη λύση εντοπισμού των ενδιαφερόντων στοιχείων, χωρίς να είναι απαραίτητη η δημιουργία πολύπλοκων ερωτημάτων (queries) σε μηχανές αναζήτησης. Επιπλέον, η δυνατότητά τους να προτείνουν στους χρήστες αντικείμενα που δεν ήξεραν καν ότι υπάρχουν, αλλά ταιριάζουν στις ανάγκες-προτιμήσεις τους, τα καθιστά μοναδικό εργαλείο αντιμετώπισης του πληροφοριακού υπερκορεσμού.

Τα συστήματα υπόδειξης δεν είναι ωφέλιμα μόνο για τους χρήστες τους αλλά και για τον πάροχό τους. Δεδομένου ότι τις περισσότερες φορές είναι τμήμα άλλων συστημάτων, αποτελούν υπηρεσίες προστιθέμενης αξίας για τα συστήματα που τα

φιλοξενούν και μπορούν να αποτελέσουν πόλο έλξης για την προσέλκυση νέων χρηστών. Ειδικά στην περίπτωση που το υπερσύστημα είναι ηλεκτρονικό κατάστημα, οι επιπτώσεις της ύπαρξης ενός συστήματος υπόδειξης είναι ιδιαίτερα σημαντικές καθώς μπορεί να επιφέρει αύξηση των πωλήσεων. Οι χρήστες που εξετάζουν τα διαθέσιμα προϊόντα ενός ηλεκτρονικού καταστήματος έχουν περισσότερες πιθανότητες να εντοπίσουν αυτό που ταιριάζει στις ανάγκες τους και να προβούν σε αγορά. Επιπλέον, προσφέρεται η δυνατότητα στο κατάστημα να επιτύχει συνδυασμένες πωλήσεις (cross-selling) προτείνοντας στους χρήστες, υπό την μορφή προσφοράς, επιπλέον προϊόντα, που σύμφωνα με το σύστημα υπόδειξης ταιριάζουν με αυτά που έχουν ήδη επιλέξει. Άλλωστε, στο περιβάλλον του διαδικτύου που ο χρήστης μπορεί πολύ εύκολα να πραγματοποιήσει τις αγορές του από πολλά διαφορετικά ηλεκτρονικά καταστήματα, η ύπαρξη ενός συστήματος υπόδειξης μπορεί να βελτιώσει την πίστη (loyalty) των χρηστών-πελατών στο κατάστημα. Οι χρήστες επενδύουν χρόνο και κόπο εκπαιδεύοντας το σύστημα υπόδειξης στις ανάγκες τους και το γεγονός αυτό θα τους αθήσει να επιστρέψουν για τις μελλοντικές αγορές τους. Παράλληλα, το κατάστημα μπορεί να χρησιμοποιήσει τις γνώσεις για τις προτιμήσεις των πελατών του για να βελτιώσει τις υπηρεσίες και τα προϊόντα του, κάτι που επίσης αυξάνει την πίστη των πελατών του και, μακροπρόθεσμα, τις πωλήσεις του (Schafer et al, 1999 - Schafer et al, 2001).

2.2 Βασικά χαρακτηριστικά

Η γενικός τρόπος λειτουργίας των συστημάτων υπόδειξης είναι ο εξής:

Υπάρχει ένα σύνολο χρηστών και ένα σύνολο αντικειμένων τα οποία μπορούν να υποδειχτούν στους χρήστες. Το μέγεθος των συνόλων αυτών είναι τις περισσότερες φορές πολύ μεγάλο, της τάξης των χιλιάδων ή και εκατομμυρίων στοιχείων. Οι χρήστες αξιολογούν ένα μικρό αριθμό αντικειμένων και το σύστημα, με βάση αυτή την αξιολόγηση, υπολογίζει τη «χρησιμότητα» κάθε αντικειμένου για κάθε χρήστη. Στη συνέχεια, παρουσιάζονται στον χρήστη όσα αντικείμενα δεν έχει βαθμολογήσει σε μη-αύξουσα σειρά «χρησιμότητας».

Είναι προφανές ότι η έννοια της χρησιμότητας (utility) είναι σχετική με το πεδίο εφαρμογής του συστήματος υπόδειξης. Για παράδειγμα, η χρησιμότητα στο πεδίο της υπόδειξης κινηματογραφικών ταινιών εκφράζει το πόσο μια ταινία θα

αρέσει στον χρήστη, ενώ στο πεδίο της υπόδειξης ενημερωτικών δελτίων (newsletters) εκφράζει το πόσο ενδιαφέρουνσα είναι η πληροφορία. Στα περισσότερα συστήματα υπόδειξης, οι χρήστες δηλώνουν την χρησιμότητα ενός αντικειμένου μέσω βαθμολογιών (ratings) και το σύστημα επεξεργάζεται τις βαθμολογίες αυτές για να παράγει εξατομικευμένες υποδείξεις.

2.2.1 Κλίμακα βαθμολογιών

Οι βαθμολογίες των χρηστών, τις περισσότερες φορές, είναι διακριτές τιμές εντός καθορισμένου διαστήματος. Στο σύστημα GroupLens (Resnick et al. 1994) οι χρήστες βαθμολογούν άρθρα με τιμές 1 έως 5. Στο LIBRA (Mooney & Roy 2000) οι βαθμολογίες βιβλίων είναι στην κλίμακα 1-10. Το Launch.com επιτρέπει στους χρήστες να βαθμολογήσουν τραγούδια στην κλίμακα 1-100. Το Jester προτείνει ανέκδοτα που βαθμολογούνται στην κλίμακα -10 έως 10. Στο Syskill & Weber (Pazzani & Billsus 1997), όμως, οι χρήστες δεν προβαίνουν σε βαθμολογία ιστοσελίδων, αλλά απλά δηλώνουν αν τους ενδιαφέρουν ή όχι.

Η κλίμακα που χρησιμοποιείται πρέπει να είναι τέτοια που να απεικονίζει την διαφοροποίηση των προτιμήσεων των χρηστών αλλά και να μην αφήνει αμφιβολίες σχετικά με τις διαφορές μεταξύ των γειτονικών τιμών. Στο πεδίο της υπόδειξης κινηματογραφικών ταινιών, η διαφορά ποιότητας μεταξύ δύο ταινιών που βαθμολογούνται με 7 και 8 σε κλίμακα 1-10 δεν είναι προφανής για τον μέσο θεατή. Αντιστοίχως, λανθασμένη θα ήταν για το ίδιο πεδίο μια κλίμακα 1-3 (1=κακό, 2=μέτριο, 3=καλό) καθώς ταινίες που είναι απλά ευχάριστες και ταινίες αριστουργήματα που θα έπρεπε να βαθμολογηθούν με 3. Επομένως, η επιλογή της κλίμακας εξαρτάται από το πεδίο εφαρμογής του συστήματος. Έχει αποδειχθεί μάλιστα (Cosley et al. 2003), ότι η κλίμακα βαθμολογίας που χρησιμοποιείται μπορεί να επηρεάσει εκτός από την απόδοση και την ακρίβεια του συστήματος και τις επιλογές των χρηστών κατά τη διάρκεια της βαθμολόγησης! Είναι εμφανές, λοιπόν, ότι η επιλογή της κατάλληλης κλίμακας είναι κεφαλαιώδους σημασίας για την επιτυχία ενός συστήματος υπόδειξης. Περισσότερες λεπτομέρειες για τον τρόπο επιλογής κατάλληλης κλίμακας βαθμολόγησης αναφέρουν οι Friedman και Amoo (1999).

Προκειμένου να προσδιοριστούν τα αντικείμενα που θα ενδιαφέρουν τους χρήστες, τα συστήματα υπόδειξης μπορούν να χρησιμοποιήσουν, εκτός από τις



άμεσες βαθμολογίες των χρηστών, και άλλες πληροφορίες που λαμβάνονται με έμμεσο τρόπο (Oard & Kim, 1998). Τέτοιες είναι ο χρόνος που διέθεσε ένας χρήστης για να ενημερωθεί για ένα αντικείμενο, κάποια ερώτηση (query) που υπέβαλε σε μηχανή αναζήτησης, μια λίστα τραγουδιών (playlist) που άκουσε, ένα βιβλίο που αγόρασε κτλ. Σε κάθε περίπτωση, αυτές οι έμμεσες πληροφορίες είναι λιγότερο αξιόπιστες από τις άμεσες βαθμολογίες (με εξαίρεση, ίσως, την αγορά προϊόντος) και γι' αυτό η πλειονότητα των συστημάτων, βασίζουν τις υποδείξεις τους στις δεύτερες. (Adomavicius & Tuzhilin, 2005)

2.2.2 Το Πρόβλημα της Υπόδειξης

Ένα κοινός τρόπος παρουσίασης του τρόπου λειτουργίας ενός συστήματος υπόδειξης είναι η χρησιμοποίηση ενός πίνακα χρηστών-αντικειμένων. Τα αντικείμενα αποτελούν τις στήλες του πίνακα και οι χρήστες τις γραμμές του. Τα στοιχεία του πίνακα είναι οι βαθμολογίες των χρηστών για τα αντικείμενα και γι' αυτό ο πίνακας αποκαλείται και «πίνακας βαθμολογιών» (rating matrix). Παράδειγμα ενός τέτοιου πίνακα, για το πεδίο της υπόδειξης κινηματογραφικών ταινιών και με κλίμακα βαθμολογίας 1-5 είναι ο Πίνακας 1. Η τιμή 0 δηλώνει ότι ο χρήστης δεν έχει βαθμολογήσει την ταινία. Σε πραγματικά συστήματα, οι πίνακες αυτοί έχουν πολλές χιλιάδες γραμμές και στήλες, ενώ είναι πολύ αραιοί, αφού ο κάθε χρήστης βαθμολογεί ένα μικρό αριθμό ταινιών.

	Η σιωπή των αμνών	Πολίτης Κέιν	Ο Λώρενς της Αραβίας	Οσα παίρνει ο άνεμος	Κάτι τρέχει με τη Μαίρη	Το Πράσινο Μίλι
Γιώργος	1	5	0	2	4	0
Δημήτρης	4	2	0	5	1	2
Μαρία	2	4	3	0	0	5
Ελένη	2	4	0	5	1	0

Πίνακας 1

Πίνακας βαθμολογιών για το πεδίο της υπόδειξης κινηματογραφικών ταινιών

Το πρόβλημα που καλούνται να λύσουν τα συστήματα υπόδειξης είναι η πρόβλεψη των βαθμολογιών που θα έδινε ο χρήστης για τα αντικείμενα που δεν έχει

ακόμα βαθμολογήσει. Αφού οι προβλέψεις αυτές πραγματοποιηθούν, παρουσιάζονται στον χρήστη, συνήθως σε μη-αύξουσα βαθμολογική σειρά.

Η παραπάνω διατύπωση του προβλήματος της ταξινόμησης προέκυψε από τους Resnick et al. (1994) και Shardanand και Maes (1995) και έθεσε τις βάσεις για την ανάπτυξη των συστημάτων υπόδειξης.

Εκτός από τα συστήματα υπόδειξης που προβαίνουν σε υπολογισμό μιας αριθμητικής πρόβλεψης για την βαθμολογία ενός αντικειμένου που δεν έχει βαθμολογήσει ο χρήστης, υπάρχουν και αυτά που προβλέπουν τη σχετική σειρά των προτιμήσεων των χρηστών. Δηλαδή αντί να υπολογίζουν μια πρόβλεψη για κάθε αντικείμενο, προβλέπουν με ποια σειρά θα ταξινομούσε ο χρήστης τα αντικείμενα. Τέτοιες προσεγγίσεις υπάρχουν στους Jin et al. (2003) και Si et al. (2003), χωρίς όμως να είναι ιδιαίτερα δημοφιλείς.

Οι αλγόριθμοι που χρησιμοποιούνται για να γίνονται οι προβλέψεις αποτελούν τον πυρήνα των συστημάτων υπόδειξης. Η πιο διαδεδομένη ταξινόμηση των συστημάτων υπόδειξης βασίζεται ακριβώς στον τρόπο με τον οποίο γίνονται οι προβλέψεις.

2.2.3 Ταξινόμηση Συστημάτων Υπόδειξης

Οι υποδείξεις που παράγονται και κατά συνέπεια και τα ίδια τα συστήματα ταξινομούνται σε (Adomavicius & Tuzhilin, 2005):

- **Συνεργατική Μέθοδος Υπόδειξης (Collaborative Filtering)**
Σε κάθε χρήστη υποδεικνύονται αντικείμενα που άλλοι χρήστες με παρόμοιες προτιμήσεις έχουν βαθμολογήσει θετικά στο παρελθόν.
- **Υπόδειξη Βασισμένη στο Περιεχόμενο (Content-based Recommendation)**
Σε κάθε χρήστη υποδεικνύονται αντικείμενα που είναι παρόμοια με τα αντικείμενα που ίδιος ο χρήστης έχει βαθμολογήσει θετικά στο παρελθόν.
- **Υβριδικές Μέθοδοι Υπόδειξης (Hybrid Recommendations)**
Είναι μέθοδοι που συνδυάζουν, με κάποιο τρόπο, τις δύο παραπάνω μεθόδους.

Εκτός από τις τρεις παραπάνω κατηγορίες, ο Burke (2002) διακρίνει άλλες δύο: τις Υποδείξεις Βασισμένες στην Χρησιμότητα (Utility-based recommendations) και τις Υποδείξεις Βασισμένες στην Γνώση (Knowledge-based recommendations). Οι Υποδείξεις Βασισμένες στην Χρησιμότητα λαμβάνουν υπόψη τους και περιστασιακές

ανάγκες των χρηστών, π.χ. την ανάγκη για γρήγορη παράδοση ενός αντικειμένου, ενώ οι Υποδείξεις Βασισμένες στην Γνώση χρησιμοποιούν πληροφορίες που προέρχονται από γνώσεις σχετικές με το πεδίο εφαρμογής. Για παράδειγμα, το σύστημα Entrée¹, προτείνει εστιατόρια εκμεταλλευόμενο γνώσεις για τις σχέσεις μεταξύ των εθνικών κουζινών και φαγητών (τα θαλασσινά δεν είναι χορτοφαγική τροφή, η Μεξικανική κουζίνα είναι πικάντικη κτλ). Ο Pazzani (1999) εντοπίζει άλλη μια κατηγορία: τις Υποδείξεις Βασισμένες σε Δημογραφικά Στοιχεία (Demographic-based Filtering). Όπως δηλώνει και το όνομά τους, χρησιμοποιούν δεδομένα σχετικά με την ηλικία, το φύλο, το πνευματικό επίπεδο κλπ. των χρηστών. Συστήματα που χρησιμοποιούν αυτού του είδους τις υποδείξεις δεν είναι διαδεδομένα, κυρίως γιατί οι χρήστες είναι διστακτικοί να αποκαλύψουν προσωπικές πληροφορίες. Οι κατηγορίες αυτές συχνά μπορούν να ενσωματωθούν στις τρεις βασικές και γ' αυτό δεν θα αναφερθούν ξεχωριστά στην συνέχεια της παρούσας εργασίας. Οι τρεις βασικές κατηγορίες θα περιγραφούν αναλυτικά σε επόμενες παραγράφους.

2.2.4 Μέτρα Αξιολόγησης

Προκειμένου να μπορούν να αξιολογηθούν τα συστήματα υπόδειξης αλλά και να συγκριθούν σε μία κοινή βάση τα αποτελέσματα των διαφόρων μεθόδων και αλγορίθμων, έχουν προταθεί αρκετές μετρικές ποιότητας.

Η κάλυψη (**Coverage**) εκφράζει το ποσοστό των αντικειμένων για τα οποία το σύστημα μπορεί να κάνει υποδείξεις. Φυσικά, στόχος για όλα τα συστήματα υπόδειξης είναι να επιτύχουν κάλυψη 100%. Αρκετά συστήματα θυσιάζουν ένα μικρό ποσοστό κάλυψης στο βωμό της επίτευξης υψηλότερης ακρίβειας.

Το Μέσο Απόλυτο Λάθος (**Mean Absolute Error – MAE**) είναι το πιο διαδεδομένο μέτρο της ακρίβειας των υποδείξεων ενός συστήματος υπόδειξης. Συγκρίνει τις υποδείξεις που παρήγαγε το σύστημα με τις πραγματικές βαθμολογήσεις των χρηστών. Όσο μικρότερο το Μέσο Απόλυτο Λάθος, τόσο μεγαλύτερη η ακρίβεια του συστήματος.

Επίσης έχουν προταθεί μέτρα ποιότητας του εκφράζουν το πόσο αποδοτικά οι υποδείξεις βοηθούν τον χρήστη να επιλέξει τα αντικείμενα που θα του αρέσουν. Τέτοια μέτρα είναι το Μέσο Κέρδος Χρήστη (**Mean User Gain**) (Carenini & Sharma, 2004) και η ευαισθησία ROC (**ROC sensitivity**, ROC-Receiver Operating

¹ <http://infolab.ils.nwu.edu/entree>



Characteristic) (Herlocker et al, 1999). Περισσότερη ανάλυση για τα μέτρα ποιότητας πραγματοποιούν οι Herlocker et al. (2004).

2.3 Συνεργατική Μέθοδος Υπόδειξης-Collaborative Filtering

Είναι η πιο δημοφιλής, περισσότερο υλοποιημένη και πιο ώριμη τεχνική που χρησιμοποιείται στα συστήματα υπόδειξης. Ο όρος “Collaborative Filtering” επινοήθηκε από τους Goldberg et al. (1992) για να περιγράψουν την μέθοδο που εφάρμοσαν στο σύστημα Tapestry που υποδείκνυε στους χρήστες ενδιαφέροντα μηνύματα ηλεκτρονικού ταχυδρομείου. Τα πρώτα συστήματα, όμως, που χρησιμοποίησαν τη Συνεργατική Μέθοδο Υπόδειξης (ΣΜΥ) για να αυτοματοποιήσουν τις προβλέψεις ήταν το GroupLens (Resnick et al., 1994) και το Ringo (Shardanand & Maes, 1995).

Είναι μια καθημερινή ανθρώπινη πρακτική η αναζήτηση προτάσεων και υποδείξεων για διάφορα αντικείμενα (ταινίες, βιβλία, αυτοκίνητα, εστιατόρια, τόποι διακοπών κλπ.) από συγγενικά και φιλικά πρόσωπα τα οποία έχουν αποδεδειγμένα κοινές προτιμήσεις με εμάς. Την πρακτική αυτή προσπαθεί να μιμηθεί η ΣΜΥ, αναζητώντας στο σύνολο των διαθέσιμων χρηστών αυτούς που παρουσιάζουν ομοιότητες προτιμήσεων. Στη συνέχεια, οι προβλέψεις παράγονται χρησιμοποιώντας τις βαθμολογίες αυτών των χρηστών.

Οι πιο διαδεδομένοι αλγόριθμοι που χρησιμοποιούνται στην ΣΜΥ είναι αυτοί που Βασίζονται στη Μνήμη (Memory-Based). Αποκαλούνται έτσι γιατί χρησιμοποιούν το σύνολο των διαθέσιμων βαθμολογιών όλων των χρηστών του συστήματος για να πραγματοποιήσουν προβλέψεις και επομένως έχουν μεγάλες απαιτήσεις σε μνήμη.

2.3.1 Αλγόριθμοι Βασισμένοι στη Μνήμη (Memory based)

Οι αλγόριθμοι αυτοί συχνά αποκαλούνται και «Βασισμένοι στη Γειτονιά» (neighborhood-based). Τα βασικά βήματα που πραγματοποιούνται για να παραχθεί πρόβλεψη για τον ενεργό χρήστη² για ένα συγκεκριμένο αντικείμενο είναι τα εξής (Herlocker et al, 1999):

1. Υπολογισμός της ομοιότητας κάθε διαθέσιμου χρήστη με τον ενεργό χρήστη.

² Ενεργός χρήστης : Ο χρήστης για τον οποίο υπολογίζεται η πρόβλεψη.



2. Επιλογή του υποσυνόλου (γειτονιά) των χρηστών που εμφανίζουν την μεγαλύτερη ομοιότητα και που θα λειτουργήσουν ως βαθμολογητές. Η επιλογή αυτή ενδέχεται να είναι διαφορετική για κάθε ξεχωριστό αντικείμενο.
3. Κανονικοποίηση των βαθμολογιών των γειτόνων (αν είναι απαραίτητο) και υπολογισμός της πρόβλεψης για τον ενεργό χρήστη από συνδυασμό των βαθμολογιών.

Είναι προφανές, ότι για την εφαρμογή των παραπάνω βημάτων απαιτείται ένα μέτρο ομοιότητας των χρηστών και ένας τρόπος συνδυασμού των βαθμολογιών για την παραγωγή προβλέψεων.

2.3.2 Μέτρα Ομοιότητας Χρηστών

Η ομοιότητα μεταξύ δύο χρηστών στα συστήματα που χρησιμοποιούν ΣΜΥ υπολογίζεται χρησιμοποιώντας τις βαθμολογίες των αντικειμένων που *και οι* δύο έχουν βαθμολογήσει.

Μια διαδεδομένη προσέγγιση είναι η εφαρμογή κάποιου μέτρου της **συσχέτισης (correlation measure)**.

Οι Resnick et al. (1994) χρησιμοποιούν τον συντελεστή συσχέτισης Pearson (Pearson correlation coefficient). Οι τιμές που μπορεί να πάρει ο συντελεστής ανήκουν στο διάστημα [-1,1], με τις τιμές κοντά στο 1 να δηλώνουν μεγάλη θετική συσχέτιση, τις τιμές κοντά στο -1 αρνητική συσχέτιση και τιμές κοντά στο 0 ανυπαρξία συσχέτισης.

Ο συντελεστής Pearson, για δύο χρήστες X και Y, υπολογίζεται από τον τύπο:

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i [(X_i - \bar{X})^2] \sum_i [(Y_i - \bar{Y})^2]}} \quad (2.1)$$

όπου X_i , Y_i είναι οι βαθμολογίες των χρηστών X και Y αντίστοιχα για το αντικείμενο i και \bar{X} , \bar{Y} είναι οι μέσοι όροι των βαθμολογιών των χρηστών X και Y στα αντικείμενα που έχουν βαθμολογήσει και οι δυο.

Όπως μπορεί να παρατηρήσει κάποιος, για τον υπολογισμό του παραπάνω τύπου πρέπει να έχουν προϋπολογιστεί οι μέσοι όροι \bar{X} και \bar{Y} , κάτι που προσθέτει πολυπλοκότητα στα συστήματα που τον χρησιμοποιούν.

Αντί αυτού, μπορεί να χρησιμοποιηθεί ο παρακάτω ισοδύναμος τύπος (Χαλκιάς, 2001) :

$$r = \frac{n \sum_i X_i Y_i - \sum_i X_i \sum_i Y_i}{\sqrt{n \sum_i X_i^2 - \left(\sum_i X_i \right)^2} \sqrt{n \sum_i Y_i^2 - \left(\sum_i Y_i \right)^2}} \quad (2.2)$$

όπου n είναι ο αριθμός των αντικειμένων που έχουν βαθμολογήσει και οι δύο χρήστες. Πρέπει να σημειωθεί ότι ο συντελεστής Pearson δεν έχει μονάδα μέτρησης. Χρησιμοποιώντας το αριθμητικό παράδειγμα του Πίνακα 1, η συσχέτιση του Γιώργου με τη Μαρία είναι 1, με τον Δημήτρη -0.8 και με την Ελένη 0.

Οι Shardanand και Maes (1995) χρησιμοποιούν μια μικρή παραλλαγή του συντελεστή συσχέτισης Pearson (constrained Pearson correlation coefficient) που λαμβάνει υπόψη την «θετικότητα» και «αρνητικότητα» των βαθμολογιών των χρηστών. Στη θέση των μέσων τιμών \bar{X} και \bar{Y} του τύπου (2.1), τοποθετούν την τιμή που βρίσκεται στη μέση της βαθμολογικής κλίμακας (1-7) που χρησιμοποιούν στο σύστημα Ringo. Ο τύπος, λοιπόν γίνεται :

$$r = \frac{\sum_i (X_i - 4)(Y_i - 4)}{\sqrt{\sum_i [(X_i - 4)^2] \sum_i [(Y_i - 4)^2]}} \quad (2.3)$$

Επειδή η κλίμακα που χρησιμοποιούν είναι απόλυτη, γνωρίζουν ότι οι τιμές 1-3 δηλώνουν αρνητική βαθμολογία και οι τιμές 5-7 θετική βαθμολογία. Έτσι, είναι λογικό να αυξάνεται η τιμή του συντελεστή συσχέτισης μόνο όταν έχουν βαθμολογήσει ένα αντικείμενο και οι δύο χρήστες θετικά ή και οι δύο αρνητικά.

Ο συντελεστής συσχέτισης Spearman (Spearman rank correlation coefficient) μπορεί να χρησιμοποιηθεί αντί για τον συντελεστή Pearson. Για τον υπολογισμό του συντελεστή, αρχικά δημιουργείται μια λίστα κατάταξης των βαθμολογιών του χρήστη. Η υψηλότερη βαθμολογία λαμβάνει τη θέση 1, η αμέσως επόμενη τη θέση 2 κοκ. Ο τύπος υπολογισμού του συντελεστή Spearman είναι:

$$w = \frac{\sum_i [(K_i - \bar{K})(L_i - \bar{L})]}{\sqrt{\sum_i (K_i - \bar{K})^2 \sum_i (L_i - \bar{L})^2}} \quad (2.4)$$

όπου K_i , L_i είναι οι θέσεις των βαθμολογιών των δύο χρηστών.

Πειραματικά αποτελέσματα από την εφαρμογή των συντελεστών Pearson και Spearman στα ίδια δεδομένα (Herlocker et al, 2002) έδειξαν ότι ο συντελεστής Pearson αποδίδει ελαφρώς καλύτερα, ενώ και ο υπολογισμός του συντελεστή

Spearmann είναι πιο απαιτητικός σε υπολογιστικούς πόρους λόγω της ανάγκης υπολογισμού της λίστας κατάταξης των βαθμολογιών.

Μια διαφορετική προσέγγιση για τον υπολογισμό της ομοιότητας χρηστών είναι η χρησιμοποίηση του μέτρου του **συνημίτονου (cosine similarity measure)**. Στην προσέγγιση αυτή, δύο χρήστες αντιπροσωπεύονται από ένα διάνυσμα (vector) μήκους ίσο με το πλήθος των αντικειμένων που έχουν και οι δύο βαθμολογήσει. Κάθε συντεταγμένη του διανύσματος αντιστοιχεί σε ένα αντικείμενο και η τιμή της είναι η βαθμολογία του χρήστη για αυτό το αντικείμενο. Η ομοιότητα δύο χρηστών μετριέται από το συνημίτονο της γωνίας των διανυσμάτων τους που δίνεται από τον τύπο:

$$\cos(\bar{a}, \bar{b}) = \frac{\bar{a} \cdot \bar{b}}{\|\bar{a}\| * \|\bar{b}\|} = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \sqrt{\sum_i b_i^2}} \quad (2.5)$$

όπου $\bar{a} \cdot \bar{b}$ είναι το εσωτερικό γινόμενο των δύο διανυσμάτων και $\|\bar{a}\|, \|\bar{b}\|$ είναι οι μήκη ή νόρμες των διανυσμάτων.

Τα εμπειρικά αποτελέσματα των Breese et al. (1998) δείχνουν ότι η προσέγγιση αυτή αποδίδει χειρότερα από τις προσεγγίσεις που χρησιμοποιούν συντελεστή συσχέτισης Pearson.

Εκτός από τα παραπάνω μέτρα ομοιότητας έχουν προταθεί και άλλα, που όμως δεν έχουν χρησιμοποιηθεί ευρέως, όπως οι **μέσες τετραγωνικές διαφορές (mean-squared differences)** (Shardanand & Maes, 1995) και μέτρα που υπολογίζουν την πιθανότητα οι χρήστες να είναι το ίδιον τύπου (Pennock et al, 2000).

2.3.3 Επιλογή Γειτονιάς

Αφού υπολογιστεί η ομοιότητα του ενεργού χρήστη με όλους τους υπόλοιπους γίνεται η επιλογή των «γειτόνων» για να πραγματοποιηθεί η πρόβλεψη κάθε ταινίας. Υποψήφιοι γείτονες είναι όλοι οι χρήστες που έχουν βαθμολογήσει στο παρελθόν την ταινία αυτή.

Στο σύστημα GroupLens (Resnick et al., 1994) στη γειτονιά συμμετέχουν όλοι οι υποψήφιοι χρήστες. Στο Ringo (Shardanand & Maes, 1995) η γειτονιά απαρτίζεται από τους υποψήφιους των οποίων το μέτρο ομοιότητας (συσχέτιση) με τον ενεργό χρήστη ήταν μεγαλύτερο από ένα κατώφλι (threshold). Όσο μεγάλωνε το κατώφλι, τόσο πιο ακριβείς γίνονταν οι προβλέψεις αλλά και τόσο μίκραινε το πλήθος των

αντικειμένων για το οποίο το σύστημα μπορούσε να κάνει υποδείξεις. Αυτό συμβαίνει διότι οι μεγάλες τιμές κατωφλιού κάνουν δυσκολότερη την εύρεση χρηστών που και να έχουν βαθμολογήσει το συγκεκριμένο αντικείμενο και να έχουν μέτρο ομοιότητας με το χρήστη μεγαλύτερο από το κατώφλι.

To Bellcore (Hill et al, 1995) χρησιμοποιεί ένα τυχαίο δείγμα από γείτονες και επιλέγει τους N καλύτερους (αν υπάρχουν N διαθέσιμοι υποψήφιοι). Στη μέθοδο RecTree (Chee et al, 2001) δημιουργούνται συστάδες (clusters) όμοιων χρηστών χρησιμοποιώντας τον αλγόριθμο K-Means. Στη συνέχεια η γειτονιά για κάθε χρήστη απαρτίζεται από τα μέλη της συστάδας στην οποία ανήκει, που έχουν βαθμολογήσει το συγκεκριμένο αντικείμενο.

2.3.4 Υπολογισμός Πρόβλεψης

Ανεξάρτητα από το πώς έχουν επιλεγεί οι γείτονες, οι βαθμολογίες τους πρέπει να συνδυαστούν για να προκύψει η πρόβλεψη για ένα αντικείμενο. Οι Resnick et al. (1994) χρησιμοποίησαν μια μέθοδο σταθμισμένης απόκλισης από τη μέση τιμή που υπολογίζεται από τον τύπο:

$$Ki = \bar{K} + \frac{\sum_{J \in Neighbors} (Ji - \bar{J})r_{KJ}}{\sum_J |r_{KJ}|} \quad (2.6)$$

όπου Ki είναι η πρόβλεψη για το αντικείμενο i , \bar{K} η μέση τιμή των βαθμολογιών του ενεργού χρήστη, Ji η βαθμολογία του γείτονα J για το αντικείμενο i , \bar{J} η μέση τιμή των βαθμολογιών του γείτονα J και r_{KJ} η τιμή του μέτρου ομοιότητας του ενεργού χρήστη με το γείτονα J . Για το αριθμητικό παράδειγμα του Πίνακα 1, η πρόβλεψη για τον Γιώργο για την ταινία «Το Πράσινο Μήλο» είναι 4.56 και για την Ελένη για την ίδια ταινία είναι 3.75.

Η μέθοδος αυτή λαμβάνει υπόψη τις διαφορετικές αντιλήψεις που έχουν για την κλίμακα βαθμολογίας δύο χρήστες. Αν έχουν μέτρο ομοιότητας 1 (δηλαδή είναι τελείως όμοιοι) αλλά ο ένας δίνει βαθμολογίες μόνο στο διάστημα 2-4 ενώ ο άλλος στο διάστημα 1-5, τότε μια βαθμολογίας 4 του πρώτου χρήστη θα έχει ως αποτέλεσμα πρόβλεψη 5 για τον δεύτερο. Επειδή η συντριπτική πλειοψηφία των συστημάτων υπόδειξης που χρησιμοποιούν αλγορίθμους Βασισμένους στη Γειτονιά χρησιμοποιούν την παραπάνω μέθοδο υπολογισμού προβλέψεων, δεν θα γίνει λόγος για άλλες μεθόδους στην παρούσα εργασία.

Υπάρχουν συστήματα που τα τρία βήματα του αλγορίθμου (υπολογισμός ομοιοτήτων, επιλογή γειτόνων και υπολογισμός πρόβλεψης) πραγματοποιούνται κάθε φορά που ο χρήστης ζητάει κάποια πρόβλεψη. Επειδή όμως οι σχέσεις μεταξύ των χρηστών δεν αλλάζουν δραματικά σε μικρά χρονικά διαστήματα, άλλα συστήματα υπολογίζουν όλες τις ομοιότητες μεταξύ των χρηστών από πριν και επαναλαμβάνουν τον υπολογισμό σε τακτά χρονικά διαστήματα. Όταν ο χρήστης ζητήσει υποδείξεις, χρησιμοποιούνται αυτές οι προϋπολογισμένες ομοιότητες για τη δημιουργία γειτονιών και την παραγωγή προβλέψεων. Η δεύτερη αυτή πρακτική προσφέρει γρηγορότερες παραγωγή υποδείξεων, που όμως έχουν κάπως μικρότερη ακρίβεια και χρησιμοποιείται, συνήθως, σε συστήματα που υποδεικνύουν μεγάλο πλήθος αντικειμένων σε πάρα πολλούς χρήστες, καθιστώντας ανέφικτο τον υπολογισμό των ομοιοτήτων σε πραγματικό χρόνο. Χαρακτηριστικότερο παράδειγμα είναι το σύστημα υπόδειξης του ηλεκτρονικού καταστήματος Amazon³ (Linden et al., 2003).

2.3.5 Προβλήματα των Αλγορίθμων Βασισμένων στη Μνήμη

Οι Αλγόριθμοι Βασισμένοι στην Μνήμη, έτσι όπως τους περιγράψαμε στις προηγούμενες παραγράφους, αντιμετωπίζουν τα εξής προβλήματα :

- Εφόσον δεν παράγεται κάποιο στατιστικό μοντέλο από την εφαρμογή τους, δεν αποκτάται γνώση για τις πραγματικές προτιμήσεις των χρηστών. Γι' αυτό είναι πολύ δύσκολο, έως αδύνατο, να τεκμηριώσει το σύστημα στο χρήστη πώς κατέληξε στις συγκεκριμένες υποδείξεις και έτσι να τον πείσει για την αξιοπιστία τους.
- Οι αλγόριθμοι αυτοί δεν κλιμακώνονται καλά ως προς τις απαιτήσεις τους σε μνήμη και υπολογιστική ισχύ. Η παραγωγή των προβλέψεων απαιτεί υπολογισμούς που πληθαίνουν με την αύξηση και του αριθμού των χρηστών και του αριθμού των αντικειμένων. Ειδικά για μεγάλο πλήθος αντικειμένων και χρηστών, η χρησιμοποίηση των αλγορίθμων σε εφαρμογές πραγματικού χρόνου είναι προβληματική έως αδύνατη.
- Οι χρήστες βαθμολογούν πολύ μικρό αριθμό αντικειμένων, συγκριτικά με το συνολικό πλήθος των διαθέσιμων αντικειμένων. Για παράδειγμα, σε ένα σύστημα υπόδειξης βιβλίων με 50.000 διαθέσιμα τίτλους, είναι παράλογο να

³ www.amazon.com

περιμένει κανείς από τους χρήστες να βαθμολογήσουν έστω και το 1% των βιβλίων. Η αραιότητα ή σποραδικότητα (sparsity) του πίνακα βαθμολογιών έχει ως αποτέλεσμα την μείωση της πιθανότητας το σύστημα να εντοπίσει γείτονες με μεγάλη ομοιότητα, με συνέπεια την αδυναμία παραγωγής υποδείξεων για συγκεκριμένα αντικείμενα ή την παραγωγή προβλέψεων μειωμένης ακρίβειας.

Στην επόμενη παράγραφο αναφέρονται τροποποιήσεις των Αλγορίθμων Βασισμένων στην Μνήμη που έχουν προταθεί για την αντιμετώπιση των παραπάνω αδυναμιών.

2.3.6 Τροποποιήσεις Αλγορίθμων Βασισμένων στη Μνήμη.

Οι Breese et al. (1998) πρότειναν τροποποιήσεις που έχουν ως στόχο την βελτίωση της ακρίβειας των υποδείξεων και την αντιμετώπιση της αραιότητας. Μια τροποποίηση είναι η χρήση Προκαθορισμένης Ψήφου (Default Voting). Στις περιπτώσεις που, λόγω της αραιότητας, δεν υπάρχουν αρκετά κοινά βαθμολογημένα αντικείμενα από δύο χρήστες, μια σταθερή τιμή ανατίθεται ως βαθμολογία και των δύο χρηστών σε ένα αριθμό μη-βαθμολογημένων αντικειμένων. Με τον τρόπο αυτό, εξασφαλίζεται ότι η τιμή της ομοιότητας μεταξύ δύο χρηστών θα βασίζεται τουλάχιστον σε έναν ελάχιστο αριθμό αντικειμένων.

Η δεύτερη τροποποίηση αποκαλείται Ενίσχυση Περίπτωσης (Case Amplification). Οι τιμές του μέτρου ομοιότητας χρηστών υψώνονται σε μια δύναμη (π.χ.. 2,5) έτσι ώστε να δίνεται έμφαση στις τιμές που είναι πιο κοντά στο 1, δηλαδή στον υπολογισμό της πρόβλεψης αποκτούν βαρύτητα οι βαθμολογίες των χρηστών που εμφανίζουν μεγάλη ομοιότητα με τον ενεργό χρήστη και που είναι πιο σπάνιοι, λόγω του φαινόμενου της αραιότητας.

Η τρίτη τροποποίηση είναι η χρησιμοποίηση της Αντίστροφης Συχνότητας Χρήστη (Inverse User Frequency). Τα αντικείμενα που έχουν βαθμολογηθεί θετικά από πάρα πολλούς χρήστες δεν είναι τόσο χρήσιμα για τον εντοπισμό ομοιοτήτων μεταξύ χρηστών, όσο τα αντικείμενα που έχουν βαθμολογηθεί από λίγους χρήστες. Έτσι στον υπολογισμό του τύπου (2.1), η βαθμολογία κάθε αντικειμένου πολλαπλασιάζεται με τον όρο $\log \frac{n}{n_i}$, όπου n είναι ο συνολικός αριθμός των χρηστών και n_i ο αριθμός των χρηστών που έχουν βαθμολογήσει το αντικείμενο αυτό.

Οι Herlocker et al. (1999) παρατηρούν ότι οι βαθμολογίες ενός γείτονα με λίγα (π.χ. 3) κοινά βαθμολογημένα αντικείμενα με τον ενεργό χρήστη θα πρέπει να έχουν μικρότερη βαρύτητα σε σχέση με τις βαθμολογίες κάποιου άλλου με πολλά (π.χ. 50) κοινά βαθμολογημένα αντικείμενα. Είναι δε συχνό φαινόμενο, οι γείτονες με τα λίγα κοινά αντικείμενα (που λόγω της αραιότητας του πίνακα βαθμολογιών είναι η πλειοψηφία) να εμφανίζουν και μεγάλο μέτρο ομοιότητας με τον ενεργό χρήστη, κάτι που οδηγεί σε εξαιρετικά ανακριβείς προβλέψεις. Η τροποποίηση που προτείνεται είναι η εξής: αν ο αριθμός των κοινά βαθμολογημένων αντικειμένων (έστω n) είναι μικρότερος από 50, τότε η τιμή του μέτρου ομοιότητας θα πολλαπλασιάζεται με τον όρο $\frac{n}{50}$ που ονομάζεται Βάρος Σημαντικότητας (Significance Weight). Έχει αποδειχθεί εμπειρικά ότι η ακρίβεια των προβλέψεων βελτιώνεται με όλες τις παραπάνω τροποποιήσεις.

Εκτός από τις προσπάθειες βελτίωσης της ακρίβειας των προβλέψεων, γίνονται και προσπάθειες μείωσης του χρόνου εκτέλεσης και βελτίωσης του τρόπου κλιμάκωσης των αλγορίθμων. Έχει ήδη αναφερθεί ότι η μέθοδος RecTree (Chee et al, 2001) δημιουργεί συστάδες χρηστών και οι προβλέψεις για κάθε χρήστη παράγονται από τα υπόλοιπα μέλη της συστάδας στην οποία ανήκει. Οι συστάδες που δημιουργούνται έχουν περίπου ίσο μέγεθος και όσο μεγαλώνει το πλήθος των χρηστών τόσο περισσότερες συστάδες δημιουργούνται. Με τον τρόπο αυτό ο χρόνος υπολογισμού των υποδείξεων είναι σταθερός ανεξάρτητα της αύξησης του πλήθους των χρηστών.

Μια άλλη πρακτική που μπορεί να εφαρμοστεί για να μειωθεί ο χρόνος εκτέλεσης των αλγορίθμων είναι η επιλογή τυχαίου δείγματος (sampling) από το σύνολο των χρηστών και η παραγωγή προβλέψεων από τις βαθμολογίες των χρηστών του δείγματος (Herlocker, 2000).

Αν και οι Αλγόριθμοι που Βασίζονται στη Μνήμη παραδοσιακά εφαρμόζονται για τον εντοπισμό ομοιοτήτων μεταξύ χρηστών, οι Sarwar et al. (2001) τους χρησιμοποιήσαν για να εντοπίσουν ομοιότητες μεταξύ αντικειμένων και παράγουν υποδείξεις μέσω των ομοιοτήτων αυτών. Έτσι λειτουργεί και το σύστημα υπόδειξης του ηλεκτρονικού καταστήματος Amazon.com (Linden et al, 2003). Η τακτική αυτή προσφέρει περισσότερες δυνατότητες τεκμηρίωσης των υποδείξεων του συστήματος στους χρήστες (π.χ. «σας προτείνουμε το αντικείμενο Α επειδή βαθμολογήσατε θετικά τα αντικείμενα Β, Γ και Δ»), ενώ το γεγονός ότι το σύνολο

των αντικειμένων μεταβάλλεται λιγότερο γρήγορα από αυτό των χρηστών και σε επιλεγμένες χρονικές στιγμές επιτρέπει τον προϋπολογισμό ομοιοτήτων αντικειμένων και, επομένως, ταχύτερους χρόνους εκτέλεσης σε πραγματικό χρόνο.

Εκτός από τις τροποποιήσεις που προαναφέρθηκαν, προκειμένου να αντιμετωπιστούν τα προβλήματα των Αλγορίθμων Βασισμένων στη Μνήμη, έχει αναπτυχθεί μια οικογένεια αλγορίθμων που ονομάζονται Βασισμένοι σε Μοντέλα (Model-based) και χρησιμοποιούν τις διαθέσιμες βαθμολογίες για να «μάθουν» ένα μοντέλο για κάθε χρήστη, το οποίο χρησιμοποιείται για την παραγωγή προβλέψεων.

2.3.7 Αλγόριθμοι Βασισμένοι σε Μοντέλα (Model-based)

Οι Billsus και Pazzani. (1998), θεωρούν την παραγωγή υποδείξεων ως πρόβλημα του πεδίου της μηχανικής μάθησης. Προτείνουν εφαρμογή της μεθόδου Ανάλυσης Ιδιαζουσών Τιμών (SVD, Singular Value Decomposition) για την μείωση των διαστάσεων του πίνακα βαθμολογιών. Αντιμετωπίζουν έτσι τα προβλήματα της αραιότητας και της δυσκολίας κλιμάκωσης. Για την παραγωγή προβλέψεων χρησιμοποιούν τεχνητά νευρωνικά δίκτυα (artificial neural networks).

Οι Breese et al. (1998), για να παράγουν πρόβλεψη για ένα αντικείμενο, υπολογίζουν την πιθανότητα κάθε διακριτής τιμής της κλίμακας βαθμολογίας να ανατεθεί στο αντικείμενο, δεδομένων των παλαιότερων βαθμολογιών του ενεργού χρήστη. Στη συνέχεια, οι πιθανότητες αυτές συνδυάζονται γραμμικά για να προκύψει η αριθμητική τιμή της πρόβλεψης. Οι πιθανότητες υπολογίζονται χρησιμοποιώντας δίκτυα Bayes (Bayesian networks) ή μοντέλα συσταδοποίησης (clustering models) που χρησιμοποιούν τον αλγόριθμο Naïve Bayes. Εφαρμογή μοντέλων συσταδοποίησης στη ΣΜΥ πραγματοποιούν και οι Ungar και Foster (1998) και O'Connor και Herlocker (1999).

Οι Goldberg et al. (2001) προτείνουν τον αλγόριθμο Eigentaste που εφαρμόζει Ανάλυση Κυρίων Συνιστώσων (PCA, Principal Components Analysis) για την μείωση των διαστάσεων και την απαλοιφή της αραιότητας του πίνακα βαθμολογιών και συσταδοποίηση χρηστών για την παραγωγή προβλέψεων. Επειδή η επεξεργασία των δεδομένων γίνεται εκ των προτέρων, η παρουσίαση των υποδείξεων στους χρήστες γίνεται σε σταθερό χρόνο (πολυπλοκότητα $O(1)$).

Τα παραπάνω είναι μικρό δείγμα των μεθόδων που έχουν προταθεί για την ανάπτυξη αλγορίθμων Βασισμένων σε Μοντέλα. Οι κατηγορία αυτή επιτυγχάνει

συχνά μεγαλύτερη ακρίβεια και καλύτερες επιδόσεις πραγματικού χρόνου σε σχέση με τους αλγορίθμους Βασισμένους σε Μνήμη. Η εφαρμογή τους, όμως, είναι πιο περιορισμένη σε σχέση με τους τελευταίους, γιατί παρουσιάζουν πολύπλοκη σχεδίαση, ενώ απαιτούν και σημαντικό χρόνο προπαρασκευής των μοντέλων που χρησιμοποιούν.

2.4 Υποδείξεις Βασισμένες στο Περιεχόμενο- Content-based Recommendations

Οι Υποδείξεις Βασισμένες στο Περιεχόμενο (ΥΒΠ) (Content-based Recommendations) έχουν τις ρίζες τους στο πεδίο της Ανάκτησης Πληροφοριών (Information Retrieval).

Η γενική ιδέα είναι ότι ο χρήστης ενός συστήματος υπόδειξης θα επιθυμεί να του προταθούν αντικείμενα που μοιάζουν με τα αντικείμενα που έχει βαθμολογήσει θετικά στο παρελθόν. Για παράδειγμα σε έναν χρήστη που του αρέσουν τα βιβλία με ήρωα των Σέρλοκ Χολμς μάλλον θα αρέσουν και τα βιβλία με ήρωα του Ήρακλή Πουναρά. Οι υποδείξεις βασίζονται αποκλειστικά στην ανάλυση του περιεχομένου των αντικειμένων που ο ενεργός χρήστης έχει βαθμολογήσει.

Δύο προβλήματα πρέπει να λυθούν για να μπορέσει να πραγματοποιηθούν ΥΒΠ: να αναπαρασταθούν τα αντικείμενα σε μια μορφή που μπορεί να αναλυθεί αποδοτικά και να βρεθεί ένας τρόπος σύγκρισης των αντικειμένων ώστε να προκύψουν οι υποδείξεις.

2.4.1 Αναπαράσταση αντικειμένων

Τα αντικείμενα αναπαρίστανται με ένα αριθμό χαρακτηριστικών (features). Έτσι τα αντικείμενα, ανεξάρτητα από το τι είναι, αντιμετωπίζονται ως κείμενα με συγκεκριμένο λεξιλόγιο (τα features). Για το πεδίο της υπόδειξης βιβλίων, τα χαρακτηριστικά είναι ο τίτλος, ο συγγραφέας, το είδος του βιβλίου, οι πιο χαρακτηριστικές λέξεις που εμφανίζονται στο κείμενο κλπ. Δεν είναι όμως πάντα εφικτό να αναπαρασταθεί αποδοτικά ένα αντικείμενο. Ένα σύστημα υπόδειξης μουσικών κομματιών, για παράδειγμα, δεν θα μπορούσε να απεικονίσει τη μελωδία και το ρυθμό ή τη φωνή του ερμηνευτή.

Το σύστημα “Syskill and Weber” (Pazzani & Billsus, 1997), που υποδεικνύει ιστοσελίδες, τις απεικονίζει με 128 λέξεις που περιέχονται σε αυτές και περιέχουν την περισσότερη πληροφορία σχετικά με την κάθε ιστοσελίδα. Οι λέξεις αυτές εντοπίζονται υπολογίζοντας για κάθε λέξη της ιστοσελίδας το **Πληροφοριακό Κέρδος** (Information Gain) (Βαζιργιάννης & Χαλκίδη, 2003) που προσφέρει η ύπαρξη ή απουσία της λέξης στο χαρακτηρισμό της ιστοσελίδας ως ενδιαφέρουσα ή μη.

Το σύστημα Fab (Balabanovic & Shoham, 1997) υποδεικνύει και αυτό ιστοσελίδες αλλά τις απεικονίζει με 100 λέξεις που περιέχονται σε αυτές και έχουν την υψηλότερη τιμή TF-IDF. Το **TF-IDF** (Term Frequency – Invert Document Frequency / Συχνότητα Όρου - Αντίστροφη Συχνότητα Κειμένου) είναι ένα μέτρο που εφαρμόζεται με πολύ μεγάλη επιτυχία στην Ανάκτηση Πληροφοριών. Υπολογίζεται ως εξής:

Εστω N ο αριθμός των αντικειμένων του συστήματος. Στο κείμενο j , η συχνότητα της λέξης i είναι $f_{i,j}$. Η Συχνότητα Όρου είναι:

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}} \quad (2.7)$$

όπου $\max_z f_{z,j}$ είναι η μέγιστη συχνότητα που παρουσιάζεται σε όλους τους όρους του κειμένου. Όσο μεγαλύτερη είναι η τιμή TF για έναν όρο, τόσο πιο σημαντικός είναι ο όρος για το κείμενο στο οποίο ανήκει.

Η Αντίστροφη Συχνότητα Κειμένου για τον ίδιο όρο i είναι:

$$IDF_i = \log \frac{N}{n_i} \quad (2.8)$$

όπου n_i είναι ο αριθμός των κειμένων στα οποία εμφανίζεται ο όρος i . Όσο μεγαλύτερη είναι η τιμή IDF, τόσο σπανιότερος είναι ο όρος στα κείμενα.

Μεγάλη τιμή γινομένου $TF*IDF$ εκφράζει ότι ο όρος είναι σημαντικός για το κείμενο αλλά ταυτόχρονα και μη σημαντικός για τα άλλα κείμενα και επομένως είναι κατάλληλος για να διαχωρίσει το κείμενο αυτό από τα υπόλοιπα. Για κάθε κείμενο-αντικείμενο εντοπίζονται οι K όροι με τις μεγαλύτερες TF-IDF τιμές. Στη συνέχεια το κείμενο απεικονίζεται ως διάνυσμα που κάθε συντεταγμένη του αντιστοιχεί σε έναν από αυτούς τους όρους και έχει τιμή το μέτρο TF-IDF του όρου.

O Pazzani (1999) ακολουθεί διαφορετική προσέγγιση, εναλλακτική του TF-IDF, χρησιμοποιώντας τον αλγόριθμο **Winnow**. O Winnow «μαθαίνει» τα βάρη που

αντιστοιχούν στους όρους των κειμένων από τη γραμμική συνάρτηση κατωφλιού (linear threshold function):

$$\sum w_i x_i > \tau \quad (2.9)$$

όπου x_i (έχει την τιμή 1 όταν το όρος i υπάρχει στο κείμενο, αλλιώς 0), w_i είναι το βάρος του όρου και τ η τιμή του κατωφλιού. Αρχικά, όλα τα βάρη παίρνουν την τιμή 1. Για κάθε κείμενο-αντικείμενο από αυτά που έχει βαθμολογήσει ο χρήστης υπολογίζεται το άθροισμα του τύπου (2.9) για τους όρους που περιλαμβάνονται στο κείμενο. Αν ο το άθροισμα είναι μεγαλύτερο από τ και ο χρήστης είχε βαθμολογήσει το αντικείμενο αρνητικά, όλοι οι όροι διαιρούνται με το 2. Αν το άθροισμα είναι μικρότερο του τ και ο χρήστης έχει βαθμολογήσει θετικά το αντικείμενο, όλοι οι όροι πολλαπλασιάζονται με το 2. Σε κάθε άλλη περίπτωση καμία αλλαγή δεν γίνεται. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να σταματήσουν να γίνονται αλλαγές στα βάρη. Ο αλγόριθμος συγκλίνει γρήγορα και αναθέτει υψηλά βάρη σε μικρό αριθμό όρων.

2.4.2 Σύγκριση Αντικειμένων – Παραγωγή Υποδείξεων

Έχοντας αναπαραστήσει τα αντικείμενα με διανύσματα σημαντικών όρων, είναι απαραίτητο ένα μέτρο σύγκρισης που να υπολογίζει την ομοιότητα μεταξύ δύο αντικειμένων. Το πιο διαδεδομένο μέτρο ομοιότητας είναι το συνημίτονο της γωνίας των δύο διανυσμάτων. Το ίδιο μέτρο χρησιμοποιείται και στη Συνεργατική Μέθοδο Υπόδειξης και υπολογίζεται από τον τύπο (2.5).

Οι υποδείξεις στους χρήστες παράγονται με τον εξής τρόπο (Kagypis, 2001):

Πριν ξεκινήσει η λειτουργία του συστήματος, για κάθε αντικείμενο j του συστήματος υπόδειξης υπολογίζεται η ομοιότητά του με όλα τα υπόλοιπα αντικείμενα και οι αντίστοιχες τιμές ομοιοτήτων αποθηκεύονται. Θεωρούμε το σύνολο U με όλα τα αντικείμενα που έχει βαθμολογήσει θετικά ο ενεργός χρήστης. Όταν ζητήσει υποδείξεις, για κάθε αντικείμενο $j \in U$ ανακτούμε τα K πιο όμοια αντικείμενα του και τα προσθέτουμε στο σύνολο υποψηφίων C . Από το C αφαιρούμε όσα αντικείμενα υπάρχουν ήδη στο U . Εν συνεχείᾳ, για κάθε αντικείμενο $c \in C$ υπολογίζουμε την ομοιότητα με το σύνολο U ως το άθροισμα των ομοιοτήτων μεταξύ των αντικειμένων

$j \in U$ και c. Τελικά, τα στοιχεία του C ταξινομούνται σε μη-φθίνουσα σειρά ομοιότητας με το U και τα πρώτα N υποδεικνύονται στο χρήστη.

Το γεγονός ότι οι ομοιότητες μεταξύ των αντικειμένων έχουν υπολογιστεί εκ των προτέρων, έχει ως αποτέλεσμα η διαδικασία παραγωγής υποδείξεων να έχει πολυπλοκότητα $O(K|U|)$, όπου $|U|$ είναι το πλήθος των αντικειμένων του συνόλου U. Στα περισσότερα συστήματα υπόδειξης, τιμές $|U| < 30$ είναι πάρα πολύ συνηθισμένες, και επομένως ο χρόνος παραγωγής υποδείξεων είναι σχετικά μικρός. Η φάση προετοιμασίας, έχει πολυπλοκότητα $O(m^2n)$, όπου m είναι ο αριθμός των αντικειμένων του συστήματος και n ο αριθμός των χαρακτηριστικών με τα οποία περιγράφονται τα αντικείμενα.

Εκτός από τις παραπάνω μεθόδους που βασίζονται σε τεχνικές της Ανάκτησης Πληροφοριών, οι Pazzani και Billsus (1997) συνέκριναν κατηγοριοποιητές Bayes, τεχνητά νευρωνικά δίκτυα (artificial neural networks) και δέντρα απόφασης (decision trees). Κατέληξαν ότι ο Naïve Bayes ήταν η καλύτερη λύση καθώς συνδύαζε ακρίβεια και ταχύτητα εκτέλεσης.

Ο Naïve Bayes δεν μετράει ομοιότητες αντικειμένων αλλά προσπαθεί να υπολογίσει τις πιθανότητες βαθμολόγησης του αντικειμένου με κάθε μια από τις τιμές της κλίμακας βαθμολογίας. Ο κατηγοριοποιητής εκπαιδεύεται με έναν αριθμό παραδειγμάτων γνωστής κατηγορίας (δηλαδή αντικειμένων που ο χρήστης έχει ήδη βαθμολογήσει) και στη συνέχεια χρησιμοποιεί το μοντέλο που έχει «μάθει» για να ταξινομήσει νέα αντικείμενα. Κατά τη φάση της εκπαίδευσης, ο κατηγοριοποιητής, ουσιαστικά, υπολογίζει τις τιμές των πιθανοτήτων που θα χρησιμοποιήσει στην φάση της κατηγοριοποίησης (Mitchell, 1997).

Αν η κλίμακα βαθμολογίας που χρησιμοποιεί ένα σύστημα είναι [1-5], τότε η πρόβλεψη u για ένα αντικείμενο υπολογίζεται από τον τύπο:

$$u = \underset{u_j \in \{1,2,3,4,5\}}{\operatorname{argmax}} P(u_j) \prod_{i=0}^m P(a_i | u_j) \quad (2.10)$$

όπου u_j είναι η τιμή της βαθμολογίας του χρήστη (που μπορεί εδώ να πάρει τιμές 1,2,3,4,5), $P(u_j)$ είναι η πιθανότητα ένα οποιοδήποτε αντικείμενο να βαθμολογηθεί από τον χρήστη με τη τιμή u_j , m είναι ο αριθμός των όρων που υπάρχουν στην περιγραφή του αντικειμένου και $P(a_i | u_j)$ είναι η (δεσμευμένη) πιθανότητα να

υπάρχει στην περιγραφή του αντικειμένου ο όρος a_i , όταν αυτό έχει βαθμολογηθεί με τιμή u_j .

Υπολογίζεται, δηλαδή, το γινόμενο $P(u_j) \prod_{i=0}^n P(a_i | u_j)$ για όλους τους

βαθμούς της βαθμολογικής κλίμακας και ο βαθμός με το μεγαλύτερο γινόμενο αποτελεί την πρόβλεψη για το αντικείμενο αυτό.

Η πιθανότητα $P(u_j)$ είναι εύκολο να υπολογιστεί από τις βαθμολογίες που έχει δώσει ο χρήστης. Αν, για παράδειγμα, έχει βαθμολογήσει 10 αντικείμενα, 2 από τα οποία με βαθμό 5, τότε η $P(5) = 0,2$.

Η πιθανότητα $P(a_i | u_j)$ υπολογίζεται από τον τύπο:

$$P(a_i | u_j) = \frac{n_i + 1}{n + |\text{Vocabulary}|} \quad (2.11)$$

όπου n είναι ο συνολικός αριθμός εμφάνισης όλων των όρων που υπάρχουν στις περιγραφές των αντικειμένων που έχουν βαθμολογηθεί με τιμή u_j , n_i είναι η συχνότητα εμφάνισης του όρου a_i στους n όρους και $|\text{Vocabulary}|$ είναι ο αριθμός των μοναδικών όρων που εμφανίζονται σε όλα τα αντικείμενα που έχει βαθμολογήσει ο χρήστης ανεξαρτήτου βαθμολογίας.

Η υπόθεση που απαιτείται να γίνει για να χρησιμοποιηθεί ο Naïve Bayes είναι ότι η πιθανότητα εμφάνισης ενός όρου είναι ανεξάρτητη από τις πιθανότητες εμφάνισης των υπολοίπων. Η υπόθεση αυτή είναι αυθαίρετη και λανθασμένη αφού, για παράδειγμα η πιθανότητα εμφάνισης της λέξης «σύστημα» είναι μεγαλύτερη αν υπάρχει η λέξη «πληροφοριακό». Ωστόσο, παρά το εσφαλμένο της υπόθεσης αυτής, έχει αποδειχθεί πειραματικά ότι ο Naïve Bayes αποδίδει ιδιαίτερα καλά σε παρόμοιες περιπτώσεις κατηγοριοποίησης, όπως απέδειξαν και οι (Pazzani & Billsus, 1997).

Η εκτέλεση του κατηγοριοποιητή είναι γρήγορη, καθώς ο χρόνος εκπαίδευσής του είναι γραμμική συνάρτηση του αριθμού των παραδειγμάτων και ο χρόνος κατηγοριοποίησης είναι ανεξάρτητος του αριθμού των αντικειμένων και επομένως σταθερός.

2.5 Σύγκριση Υποδείξεων Βασισμένων στο Περιεχόμενο και Συνεργατικής Μεθόδου

Στις προηγούμενες παραγράφους περιγράφηκαν οι δύο βασικές μέθοδοι υπόδειξης. Η κάθε μια έχει πλεονεκτήματα και μειονεκτήματα που θα αναλυθούν σε αυτή την παράγραφο.

Έχει ήδη αναφερθεί, ότι οι Υποδείξεις Βασισμένες στο Περιεχόμενο για να πραγματοποιηθούν πρέπει πρώτα τα αντικείμενα να αναπαρασταθούν με μια περιγραφή χαρακτηριστικών που να μπορεί να προκύπτει με αυτόματο τρόπο και να μπορεί να αναλυθεί αποδοτικά. Αυτό αποτελεί δεσμευτικό παράγοντα που δεν επιτρέπει την εφαρμογή της μεθόδου σε πεδία που μια τέτοια αναπαράσταση των αντικειμένων είναι δυσχερής, όπως η υπόδειξη μουσικών κομματιών, εικόνων κλπ. Επίσης, επειδή η σύγκριση αντικειμένων γίνεται αποκλειστικά με βάση τα χαρακτηριστικά, δύο στοιχεία που περιγράφονται με τα ίδια χαρακτηριστικά δεν μπορούν να διαχωριστούν. Για παράδειγμα, ένα σύστημα υπόδειξης ανεκδότων δεν θα μπορούσε να διαχωρίσει ένα επιτυχημένο από ένα «κρύο» αστείο, αν και τα δύο χρησιμοποιούν τις ίδιες χαρακτηριστικές λέξεις (Adomavicius & Tuzhilin, 2005).

Άλλο πρόβλημα των Υποδείξεων Βασισμένων στο Περιεχόμενο είναι η υπερεξειδίκευση (over-specialization) (Balabanovic & Shoham, 1997). Όταν υποδεικνύονται μόνο αντικείμενα που μοιάζουν με όσα αντικείμενα ο χρήστης έχει βαθμολογήσει στο παρελθόν, είναι σχεδόν αδύνατο να προταθούν άλλου είδους αντικείμενα. Ας φανταστούμε, για παράδειγμα, έναν νέο χρήστη ενός συστήματος υπόδειξης ταινιών. Αν στην αρχή αυτός βαθμολογήσει θετικά 5 ταινίες που ανήκουν στο είδος των θρύλερ, τότε το σύστημα πιθανότατα θα του υποδείξει έναν αριθμό ταινιών, επίσης θρύλερ. Καθώς ο χρήστης θα συνεχίζει να βαθμολογεί τις ταινίες που του προτείνονται, τελικά το σύστημα θα εκπέσει σε παρουσιαστή ταινιών του ίδιου είδους.

Τα προβλήματα αυτά μπορούν να αντιμετωπιστούν με την Συνεργατική Μέθοδο Υπόδειξης. Εφόσον οι υποδείξεις βασίζονται στις βαθμολογίες άλλων χρηστών με παρόμοιες προτιμήσεις, η αναπαράσταση των αντικειμένων δεν χρειάζεται. Έτσι μπορούν να υποδειχτούν αντικείμενα κάθε είδους. Επίσης και οι λεπτές διαφορές αντικειμένων που έχουν σχέση με το προσωπικό γούστο και την ποιότητα μπορούν να εντοπιστούν από το ανθρώπινο κριτήριο των άλλων χρηστών. Η υπερεξειδίκευση δεν υπάρχει. Για την ακρίβεια, η ικανότητα της ΣΜΥ να υποδεικνύει

αντικείμενα που είναι χρήσιμα για το χρήστη αλλά που δεν περιέχουν περιεχόμενο που εκείνος θα περίμενε (*serendipity*), είναι από τα πιο εντυπωσιακά στοιχεία της μεθόδου. Οι Herlocker et al. (1999) έχουν βρει ότι το φαινόμενο αυτό συμβαίνει πολύ συχνά στο πεδίο της υπόδειξης κινηματογραφικών ταινιών, όπου γίνονται επιτυχημένες προτάσεις ταινιών που αλλιώς ο χρήστης δεν θα σκεφτόταν ποτέ.

Η ΣΜΥ, όμως, υστερεί στην περύπτωση της εισόδου ενός νέου αντικειμένου στο σύστημα (*New Item Problem*). Το νέο αντικείμενο που δεν έχει βαθμολογηθεί από κανέναν χρήστη δεν πρόκειται να υποδειχθεί ποτέ. Το πρόβλημα αυτό είναι γνωστό και ως το πρόβλημα του «Πρώτου Βαθμολογητή», καθώς ο πρώτος που θα βαθμολογήσει το αντικείμενο δεν έχει κάποια ωφέλεια από αυτό αφού δεν θα βελτιώσει την ομοιότητά του με άλλους χρήστες και, επομένως, ούτε τις υποδείξεις που θα λάβει στη συνέχεια (Avery & Zeckhauser, 1997). Οι ΥΒΠ δεν εμφανίζουν το πρόβλημα αυτό. Το νέο αντικείμενο να υποδειχθεί τους χρήστες που έχουν βαθμολογήσει θετικά άλλα αντικείμενα με παρόμοιο περιεχόμενο.

Ένα άλλο πρόβλημα που αντιμετωπίζει η ΣΜΥ είναι ότι απαιτεί την ύπαρξη μιας κρίσιμης μάζας χρηστών. Χωρίς αυτή είναι αδύνατη η εύρεση γειτόνων για την παραγωγή υποδείξεων. Το πρόβλημα αυτό είναι ιδιαίτερα έντονο για τις περιπτώσεις που ένα νέο σύστημα που χρησιμοποιεί την ΣΜΥ ξεκινά τη λειτουργία του. Χωρίς σύνολο χρηστών δεν μπορεί να κάνει υποδείξεις και χωρίς υποδείξεις δεν πρόκειται ποτέ να χρησιμοποιηθεί από χρήστες γιατί απλά δεν προσφέρει τίποτε. Την λύση εδώ μπορεί να δώσουν οι ΥΒΠ, καθώς μπορούν να αρχίσουν να κάνουν υποδείξεις χωρίς μεγάλο σύνολο χρηστών.

Ακόμα και όταν υπάρχει η κρίσιμη μάζα χρηστών, παρουσιάζεται το γνωστό πρόβλημα της αραιότητας των πινάκων βαθμολογιών που χρησιμοποιούν τα συστήματα υπόδειξης που επηρεάζει αρνητικά την ακρίβεια των παραγόμενων προβλέψεων. Μια παραλλαγή του προβλήματος της αραιότητας είναι ένας χρήστης να έχει τόσο ιδιάζουσες προτιμήσεις ώστε να μην μπορεί να βρει ποτέ γείτονες που να μοιάζουν με αυτόν. Το φαινόμενο αυτό αποκαλείται από τους Claypool et al. (1999) ως “gray sheep” και η πιθανότητα να συμβεί είναι αντιστρόφως ανάλογη του μεγέθους του συνόλου χρηστών.

Ένα πρόβλημα που αντιμετωπίζουν και οι δύο μέθοδοι είναι η δημιουργία ενός προφίλ προτιμήσεων των νέων χρηστών με όσο το δυνατόν λιγότερες βαθμολογήσεις αντικειμένων (*New User Problem*). Η βαθμολόγηση μεγάλου αριθμού αντικειμένων μπορεί να θεωρείται κοπιαστική διαδικασία από τους χρήστες. Οι

Rashid et al. (2002) προτείνουν να ζητείται από τον νέο χρήστη να βαθμολογεί αρχικά ένα μικρό αριθμό αντικειμένων που να είναι δημοφιλή (ώστε να αυξάνεται η πιθανότητα ο νέος χρήστης να είναι σε θέση να τα βαθμολογήσει), αλλά και να περιέχουν πληροφορία που να επιτρέπουν να εντοπιστούν οι ιδιαίτερες προτιμήσεις του, έτσι ώστε να μπορούν να γίνουν ακριβείς υποδείξεις το ταχύτερο δυνατό. Τις περισσότερες φορές πάντως, ισχύει η αρχή ότι σε όσο περισσότερες βαθμολογίες έχει προβεί ο χρήστης, τόσο πιο ακριβείς είναι οι υποδείξεις που λαμβάνει.

Από τα παραπάνω είναι φανερό ότι η κάθε μέθοδος έχει συγκεκριμένα εγγενή προβλήματα που τις περισσότερες φορές αντιμετωπίζονται με την χρήση της άλλης μεθόδου. Αυτό γεννά αναπόφευκτα την ιδέα να χρησιμοποιηθούν μαζί οι δύο μέθοδοι, η κάθε μια εκεί που εμφανίζει τα περισσότερα πλεονεκτήματα. Οδηγούμαστε λοιπόν στην ανάπτυξη συστημάτων υπόδειξης που παράγουν υποδείξεις με υβριδικό τρόπο, όπως αναλύεται στην επόμενη παράγραφο.

2.6 Υβριδικές Μέθοδοι Υπόδειξης–*Hybrid Recommendations*

Οι Adomavicius και Tuzhilin (2005) διακρίνουν τέσσερις τρόπους συνδυασμού των βασικών μεθόδων υπόδειξης :

1. Συνδυασμός ανεξάρτητων υποδείξεων.
2. Ενσωμάτωση χαρακτηριστικών ΥΒΠ στη ΣΜΥ.
3. Ενσωμάτωση χαρακτηριστικών της ΣΜΥ σε ΥΒΠ.
4. Ενοποιημένο μοντέλο.

Ο Burke (2002) χρησιμοποιεί μια διαφορετική ταξινόμηση των υβριδικών μεθόδων υπόδειξης. Ήα ακολουθήσουμε τη ταξινόμηση των Adomavicius και Tuzhilin και όπου είναι εφικτό θα αναφέρουμε και τις κατηγορίες του Burke.

2.6.1 Συνδυασμός ανεξάρτητων υποδείξεων

Τα συστήματα υπόδειξης που λειτουργούν με αυτό τον τρόπο περιλαμβάνουν δύο ανεξάρτητα υποσυστήματα. Ένα που παράγει υποδείξεις χρησιμοποιώντας τη Συνεργατική Μέθοδο Υπόδειξης και ένα που παράγει Υποδείξεις Βασισμένες στο Περιεχόμενο. Ο συνδυασμός των υποδείξεων μπορεί να γίνει με δύο τρόπους.

Στον πρώτο τρόπο, οι τελικές υποδείξεις προκύπτουν ως γραμμικός συνδυασμός των επιμέρους υποδείξεων. Οι Claypool et al. (1999) έχουν αναπτύξει το

σύστημα P-Tango με τη μέθοδο αυτή. Η πρακτική αυτή αποκαλείται από τον Burke (2002) Σταθμισμένη (Weighted) Μέθοδος.

Στον δεύτερο τρόπο συνδυασμού επιλέγεται η «καλύτερη» υπόδειξη κάθε φορά. Στο σύστημα DailyLearner (Billsus & Pazzani, 2000) εφαρμόζονται πρώτα YBP. Αν δεν μπορούν να παράγουν υπόδειξη με επαρκή εμπιστοσύνη, τότε οι υποδείξεις γίνονται με ΣΜΥ. Ο Burke (2002) αποκαλεί την μέθοδο αυτή switching.

2.6.2 Ενσωμάτωση χαρακτηριστικών YBP στη ΣΜΥ

Ο Pazzani (1999) χρησιμοποιεί την μέθοδο “collaboration via content” (συνεργασία διαμέσου περιεχομένου) που παράγει προβλέψεις με τον ίδιο τρόπο όπως η ΣΜΥ, με τη διαφορά ότι η ομοιότητα των χρηστών δεν προκύπτει από τα κοινά βαθμολογημένα αντικείμενα, αλλά από ένα προφίλ που σχηματίζεται για κάθε χρήστη με βάση το περιεχόμενο των αντικειμένων που έχει βαθμολογήσει.

Η ίδια έννοια του προφίλ υπάρχει και στο σύστημα FAB (Balabanovic & Shoham, 1997). Αντικείμενα υποδεικνύονται όταν έχουν μεγάλο μέτρο ομοιότητας με το προφίλ του ενεργού χρήστη ή / και όταν έχουν βαθμολογηθεί υψηλά από χρήστες με προφίλ όμοιο με τον ενεργό χρήστη.

Μια άλλη προσέγγιση (Sarwar et al, 1998, Good et al, 1999) είναι η ενσωμάτωση, στο σύστημα υπόδειξης, πρακτόρων λογισμικού (software agents) οι οποίοι λειτουργούν σαν χρήστες και βαθμολογούν αντικείμενα χρησιμοποιώντας ένα προφίλ βασισμένο στο περιεχόμενο. Μειώνεται έτσι η αραιότητα του πίνακα βαθμολογιών και αντιμετωπίζεται το πρόβλημα των νέων αντικειμένων. Οι υποδείξεις για τους κανονικούς χρήστες γίνονται με τον κλασσικό τρόπο της ΣΜΥ.

Οι Melville et al. (2002) παράγουν μια Υπόδειξη Βασισμένη στο Περιεχόμενο για κάθε στοιχείο του πίνακα βαθμολόγησης που δεν έχει βαθμολογηθεί. Έτσι παράγεται ένας ψευδο-πίνακας με αραιότητα 0%. Στη συνέχεια, με βάση αυτόν τον πίνακα παράγονται υποδείξεις για τους χρήστες με ΣΜΥ.

Οι μέθοδοι της κατηγορίας αυτής αντιστοιχούν στις κατηγορίες μετακύλιση (cascade) και επαύξηση χαρακτηριστικών (feature augmentation) της κατηγοριοποίησης του Burke (2002).

2.6.3 Ενσωμάτωση χαρακτηριστικών της ΣΜΥ σε ΥΒΠ.

Η μέθοδος αυτή, σύμφωνα με την ταξινόμηση του Burke (2002), αποκαλείται «συνδυασμός χαρακτηριστικών» (feature combination). Τα αποτελέσματα της ΣΜΥ χρησιμοποιούνται ως επιπλέον χαρακτηριστικά (features) που σχετίζονται με τα αντικείμενα και εφαρμόζονται τεχνικές βασισμένες στο περιεχόμενο για την παραγωγή υποδείξεων. Με τον τρόπο αυτό χρησιμοποιείται η ΣΜΥ, αποφεύγονται όμως οι αρνητικές επιπτώσεις του φαινόμενου της αραιότητας.

Οι Basu et al. (1998) χρησιμοποίησαν την μέθοδο στο σύστημα υπόδειξης ταινιών Ripper και πέτυχαν σημαντική βελτίωση της ακρίβειας των υποδείξεων σε σχέση με την ΣΜΥ, αλλά το σύστημα δεν ήταν πλήρως αυτοματοποιημένο και απαιτούσε την ανθρώπινη παρέμβαση για την επιλογή των χαρακτηριστικών.

Οι Soboroff και Nikolas (1999) βελτίωσαν την ακρίβεια των ΥΒΠ χρησιμοποιώντας Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis – LSA) για να μειώσουν τις διαστάσεις του διανυσματικού χώρου των διανυσμάτων που περιγράφουν τα αντικείμενα, λαμβάνοντας υπόψη δεδομένα που προκύπτουν από τη ΣΜΥ.

Εκτός από τις παραπάνω περιπτώσεις, η μέθοδος της ενσωμάτωσης χαρακτηριστικών της ΣΜΥ σε ΥΒΠ δεν έχει εφαρμοστεί ευρέως. Η εξήγηση είναι σχετικά απλή: Η ΣΜΥ πλεονεκτεί σε περισσότερα σημεία των ΥΒΠ και έτσι οι περισσότεροι μελετητές και σχεδιαστές συστημάτων προτιμούν να βασιστούν πάνω της και να προσπαθήσουν να διορθώσουν τις αδυναμίες της χρησιμοποιώντας ΥΒΠ. Η αντίθετη διαδικασία δεν υπόσχεται μεγάλη επιτυχία.

2.6.4 Ενοποιημένο Μοντέλο

Στην μέθοδο αυτή χρησιμοποιούνται κυρίως στατιστικά μοντέλα για τον συνδυασμό των δύο μεθόδων. Οι Popescul et al (2001) προτείνουν ένα ενοποιημένο πιθανοθεωρητικό μοντέλο συνδυασμού της ΣΜΥ και των ΥΒΠ που στηρίζεται στη Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis – LSA). Η προσέγγιση των Ansari et al (2000) χρησιμοποιεί αλυσίδες Μαρκόφ για την εκτίμηση των υποδείξεων.

Η έρευνα για την ανάπτυξη υβριδικών συστημάτων με ενοποιημένο μοντέλο υποδείξεων είναι ακόμα σε αρχικά στάδια. Ένας πιθανός λόγος είναι ότι η

ενοποιημένη προσέγγιση προσθέτει πολυπλοκότητα στην υλοποίηση του συστήματος και αναιρεί ένα από τα προτερήματα της ΣΜΥ, που είναι η σχετικά απλή και διαισθητικά κατανοητή σχεδίαση.

Τα ως τώρα εμπειρικά αποτελέσματα από την εφαρμογή των υβριδικών μεθόδων υπόδειξης δείχνουν ότι παράγουν πιο ακριβείς υποδείξεις σε σχέση με τα συστήματα που χρησιμοποιούν αμιγώς ΣΜΥ ή αμιγώς ΥΒΠ.

Ωστόσο, αν και τα αποτελέσματα της έρευνας είναι ενθαρρυντικά, οι προτεινόμενες λύσεις και τεχνικές δεν έχουν εφαρμοστεί ολοκληρωμένα συστήματα υπόδειξης. Αποτελεί κενό για το πεδίο το γεγονός ότι δεν έχει αναπτυχθεί μεγάλος αριθμός συστημάτων, ενώ τα σημαντικότερα από τα υπάρχοντα έχουν εμπορικό χαρακτήρα (π.χ. Amazon.com, Barnes & Noble κλπ.). Αρκετοί ερευνητές δοκιμάζουν τις μεθόδους που προτείνουν σε σύνολα ιστορικών δεδομένων, χωρίς όμως να τις ενσωματώνουν σε ολοκληρωμένα συστήματα τα οποία να είναι εύκολο να χρησιμοποιηθούν από πραγματικούς χρήστες. Επιπλέον, τα υπάρχοντα συστήματα υλοποιούν συνήθως μία μέθοδο υπόδειξης (ΣΜΥ, ΥΒΠ ή κάποιο υβριδικό συνδυασμό) κάτι που δεν επιτρέπει την άμεση σύγκριση των διαφορετικών προσεγγίσεων στις ίδιες συνθήκες (ίδιο σύνολο αντικειμένων και χρηστών, ίδιος τρόπος παρουσίασης αποτελεσμάτων κλπ). Επομένως, υπάρχει η ανάγκη ανάπτυξης συστημάτων που θα συνδυάζουν τη λειτουργικότητα με την δυνατότητα πειραματισμού και εξαγωγής συμπερασμάτων.

Ένα τέτοιο σύστημα είναι το σύστημα υπόδειξης κινηματογραφικών ταινιών MoRe που αναπτύσσουμε στην παρούσα εργασία. Στα πλαίσια της ανάπτυξής του προσπαθούμε να βελτιώσουμε τις δύο βασικές μεθόδους υπόδειξης, ενώ προτείνουμε και δύο υβριδικές προσεγγίσεις. Το σύστημα μπορεί να παράγει υποδείξεις και με τις τέσσερις αυτές μεθόδους.

Στο επόμενο κεφάλαιο αναφερόμαστε στις σχεδιαστικές αποφάσεις για την ανάπτυξη του συστήματος και που αφορούν, κυρίως, τους αλγορίθμους που χρησιμοποιούνται για την παραγωγή υποδείξεων και την δομή του συστήματος.



3. Σχεδιαστικές Επιλογές

Στο προηγούμενο κεφάλαιο έγινε περιγραφή των βασικότερων αλγορίθμων που χρησιμοποιούνται για την παραγωγή υποδείξεων, καθώς και οι επικρατούσες πρακτικές στην σχεδίαση συστημάτων υπόδειξης. Επίσης επισημάνθηκαν οι ανάγκες για επιπλέον έρευνα στο πεδίο και ο μικρός αριθμός ολοκληρωμένων συστημάτων.

Το σύστημα MoRe (από τα αρχικά του Movie Recommender) αναπτύχθηκε για το πεδίο της υπόδειξης κινηματογραφικών ταινιών. Έχει πραγματοποιηθεί υλοποίηση της Συνεργατικής Μεθόδου Υπόδειξης, δύο διαφορετικές υλοποιήσεις παραγωγής Υπόδειξεων Βασισμένων στο Περιεχόμενο και δύο υβριδικές προσεγγίσεις συνδυασμού των βασικών μεθόδων υπόδειξης. Το σύστημα είναι δυνατόν να παράγει υπόδειξεις με όλες τις παραπάνω μεθόδους, έτσι ώστε να μπορεί να γίνει άμεση σύγκριση ανάμεσά τους. Κατά την διάρκεια της ανάπτυξης και της αξιολόγησης του συστήματος χρησιμοποιήθηκε σύνολο δεδομένων (dataset) με βαθμολογήσεις ταινιών από πραγματικούς χρήστες. Το σύστημα έχει τη μορφή διαδικτυακής εφαρμογής (web application).

Στη συνέχεια του κεφαλαίου θα περιγραφεί αναλυτικά ο σχεδιασμός του συστήματος και η τεκμηρίωση των αποφάσεων που οδήγησαν στην επιλογή συγκεκριμένων αλγορίθμων και πρακτικών.

3.1 Σύνολο δεδομένων (*Dataset*)

Για λόγους αληθοφάνειας, ελέγχου της λειτουργίας και της ακρίβειας αλγορίθμων κατά την ανάπτυξη του συστήματος αλλά και για την αξιολόγησή του ήταν απαραίτητη η ύπαρξη ενός συνόλου δεδομένων με βαθμολογίες χρηστών, αφού είναι αδύνατη η συγκέντρωση νέων βαθμολογιών πριν το σύστημα λειτουργήσει.

Κύριο κριτήριο για την επιλογή του συνόλου δεδομένων ήταν να έχει χρησιμοποιηθεί και από άλλους ερευνητές, έτσι ώστε να είναι δυνατή η σύγκριση αποτελεσμάτων. Δεύτερο κριτήριο ήταν να περιλαμβάνει σχετικά μεγάλο όγκο δεδομένων έτσι ώστε να μπορεί να διαπιστωθεί το πώς ανταποκρίνονται οι αλγόριθμοι σε δεδομένα τέτοιου μεγέθους και η επεκτασιμότητα του συστήματος.



Κεφάλαιο 3 : Σχεδιαστικές Επιλογές

Τρίτο κριτήριο ήταν το σύνολο δεδομένων να προέρχεται από ένα πεδίο που να έχει ενδιαφέρον, τα αντικείμενά του να μην είναι εφήμερα (όπως για παράδειγμα είναι τα ενημερωτικά δελτάρια – newsletter) και να έχουν πολλές πιθανότητες να αξιολογηθούν από Έλληνες χρήστες (για παράδειγμα κάτι τέτοιο δεν θα ήταν πολύ εύκολο για ξενόγλωσσα βιβλία). Επιλέχθηκε το πεδίο της υπόδειξης κινηματογραφικών ταινιών. Στο πεδίο αυτό, τα διαθέσιμα σύνολα δεδομένων που έχουν χρησιμοποιηθεί και από άλλους ερευνητές ήταν το MovieLens¹ και το EachMovie. Επιλέξαμε το MovieLens, καθώς διέθετε μεγαλύτερο όγκο δεδομένων.

Το σύνολο δεδομένων παρέχεται από ιστορικά δεδομένα του ομώνυμου συστήματος υπόδειξης κινηματογραφικών ταινιών που έχει αναπτυχθεί από το πανεπιστήμιο της Minnesota. Το MovieLens χρησιμοποιεί Συνεργατική Μέθοδο Υπόδειξης, ενισχυμένη με πράκτορες λογισμικού που παράγουν αυτόματες βαθμολογήσεις βασισμένες στο περιεχόμενο (βλ. παράγραφο 2.3.4.2). Το σύνολο δεδομένων περιλαμβάνει 1.000.209 ανώνυμες βαθμολογήσεις από 6.040 χρήστες που εγγράφηκαν στο MovieLens το 2000 για 3.952 ταινίες. Ο κάθε χρήστης έχει τουλάχιστον 20 βαθμολογήσεις. Είναι οργανωμένο σε τρία αρχέα.

Το αρχείο ταινιών (movies.dat) περιλαμβάνει πληροφορίες για τις ταινίες στο μορφό τυπο:

Κωδικός Ταινίας:: Τίτλος:: Είδος

Οι κωδικοί ταινιών είναι αύξοντες αριθμοί από το 1 έως το 3952. Στον τίτλο περιλαμβάνεται και η χρονολογία προβολής. Το αρχείο ήταν απαραίτητο να υποστεί κάποια επεξεργασία πριν μπορέσει να χρησιμοποιηθεί από το σύστημα, καθώς κάποιοι κωδικοί ταινιών δεν αντιστοιχούσαν σε καμία ταινία. Η επεξεργασία αυτή έγινε με αυτόματο τρόπο, δημιουργώντας μια εφαρμογή σε Java. Επιπλέον, στην πορεία της παρούσας εργασίας και ιδιαίτερα κατά τη διάρκεια της υλοποίησης των ΥΒΠ, διαπιστώθηκε ότι μερικές χρονολογίες προβολής ήταν λανθασμένες και έπρεπε να διορθωθούν με μη αυτόματο τρόπο.

Επειδή τα δεδομένα για κάθε ταινία είναι πολύ λίγα για να χρησιμοποιηθούν για τις ΥΒΠ, υλοποιήθηκε ένας web crawler² προκειμένου να συγκεντρωθούν τα απαραίτητα δεδομένα. Ο crawler, χρησιμοποιώντας τον τίτλο της κάθε ταινίας και το

¹ <http://www.netflixprize.com>

² web crawler : Είναι ένα πρόγραμμα το οποίο επισκέπτεται ιστοσελίδες και με αυτόματο τρόπο αναζητά και αποθηκεύει επιθυμητό περιεχόμενο.

έτος προβολής της, αναζητά πληροφορίες για την ταινία στην ιστοσελίδα της IMDb³, εκμεταλλευόμενος την δυνατότητα αναζήτησης ταινιών που προσφέρει. Οι πληροφορίες που συλλέγονται για κάθε ταινία είναι το είδος της ταινίας, οι ηθοποιοί που συμμετέχουν, ο σκηνοθέτης, οι σεναριογράφοι, οι παραγωγοί και οι λέξεις πλοκής. Οι τελευταίες είναι λέξεις που περιγράφουν την πλοκή της ταινίας και έχουν επιλεγεί από την IMDb. Με τις λέξεις αυτές δεν χρειάζεται εφαρμογή κάποιου μέτρου αντίστοιχου με το TF-IDF για την επιλογή γνωρισμάτων.

Το αρχείο ταινιών (movies.dat) του συνόλου δεδομένων εμπλουτίζεται με της παραπάνω πληροφορίες και η κάθε γραμμή του είναι πλέον της μορφής:

Κωδικός Ταινίας:: Τίτλος:: Είδος:: Σκηνοθέτης:: Σεναριογράφοι:: Παραγωγοί::
Ηθοποιοί:: Λέξεις Πλοκή

Τα γνωρίσματα που ανήκουν σε κάθε πεδίο χωρίζονται με κόμματα. Μια ενδεικτική γραμμή από το αρχείο ταινιών είναι η παρακάτω:

2028:: Saving Private Ryan (1998):: Action, Drama, History, War:: Steven Spielberg:: Robert Rodat:: Ian Bryce, Bonnie Curtis:: Tom Hanks, Tom Sizemore, Edward Burns, Barry Pepper, Adam Goldberg,:: Infantry, Shelling, Dismemberment, Omaha Beach, Tank, Sniper, Horrors Of War

Το αρχείο χρηστών (users.dat) περιλαμβάνει πληροφορίες για τους χρήστες και κάθε γραμμή του είναι στη μορφή:

Κωδικός Χρήστη:: Φύλο:: Ηλικία:: Επάγγελμα:: Ταχυδρομικός Κώδικας

Επειδή το MoRe δεν παράγει υποδείξεις που βασίζονται σε δημογραφικά στοιχεία, η μόνη πληροφορία που χρησιμοποιεί είναι ο Κωδικός Χρήστη που είναι αύξοντες αριθμοί από το 1 έως το 6040.

Το αρχείο βαθμολογιών (ratings.dat) περιλαμβάνει τις βαθμολογίες που έχουν δώσει οι χρήστες. Η κάθε γραμμή του αρχείου είναι της μορφής:

Κωδικός Χρήστη:: Κωδικός Ταινίας:: Βαθμολογία

³ Internet Movie Database (<http://www.imdb.com>) : Μια ηλεκτρονική βάση δεδομένων όπου υπάρχουν στοιχεία για όλες τις κινηματογραφικές ταινίες και τηλεταινίες που προβληθεί ανά τον κόσμο.

Οι βαθμολογίες είναι ακέραιες και βρίσκονται στην κλίμακα 1 (πολύ κακό) έως 5 (πολύ καλό). Το γεγονός αυτό επέβαλε και την επιλογή της ίδιας κλίμακας στο σύστημα MoRe. Βέβαια, αυτό δεν προκαλεί πρόβλημα, καθώς όπως αναλύθηκε στην παράγραφο 2.2.1, η κλίμακα αυτή κρίνεται κατάλληλη για το πεδίο της υπόδειξης κινηματογραφικών ταινιών.

Πρέπει να σημειωθεί ότι η αραιότητα (sparsity) του συνόλου δεδομένων είναι περίπου 96%. Η αραιότητα αυτή είναι αντιπροσωπευτική για τα δεδομένα των συστημάτων υπόδειξης.

3.2 Διαχείριση Δεδομένων

Ένα σύστημα υπόδειξης πρέπει να διαχειρίζεται μεγάλους όγκους δεδομένων, καθώς διατηρεί σύνολα χιλιάδων αντικειμένων και χρησιμοποιεί βαθμολογίες χιλιάδων χρηστών. Όπως έχει προαναφερθεί, οι βαθμολογίες των χρηστών αποθηκεύονται στον πίνακα βαθμολογιών, οι στήλες του οποίου αντιστοιχούν στα αντικείμενα (στην περίπτωση του MoRe στις ταινίες) που υπάρχουν στο σύστημα και οι γραμμές στους χρήστες του συστήματος. Τα στοιχεία του πίνακα είναι οι βαθμολογίες ενός συγκεκριμένου χρήστη για μια ταινία.

Ο πίνακας αυτός πρέπει ανά πάσα στιγμή να είναι διαθέσιμος, έτσι ώστε να μπορούν να παράγονται οι υποδείξεις για τους χρήστες αλλά και να μπορεί να ενημερώνεται γρήγορα όταν οι χρήστες βαθμολογούν νέες ταινίες.

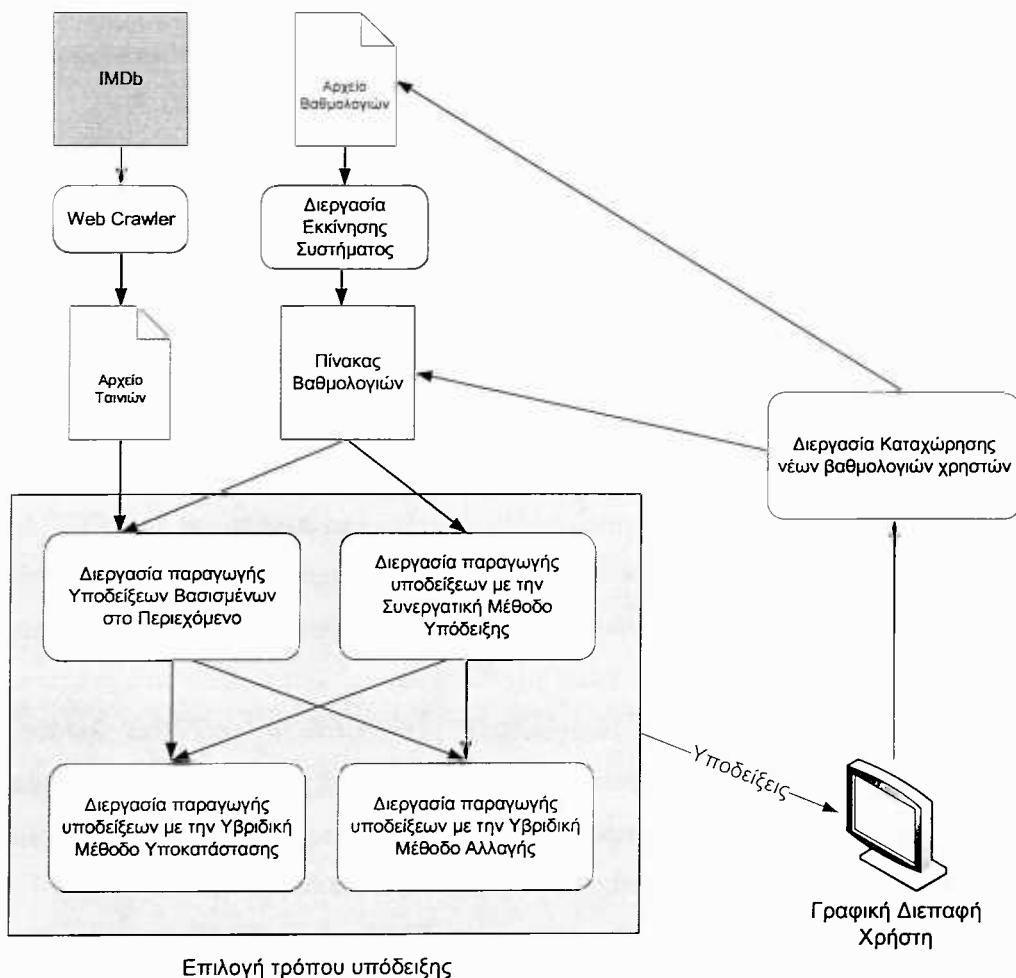
Η πιο συνηθισμένη τακτική που ακολουθείται σε συστήματα που διαχειρίζονται μεγάλους όγκους δεδομένων είναι να χρησιμοποιείται μια βάση δεδομένων. Αποφασίσαμε να μην ακολουθήσουμε αυτή την τακτική. Ένας πίνακας σε μια βάση δεδομένων δεν μπορεί να χωρέσει έναν πίνακα χιλιάδων γραμμών και στηλών, χωρίς κάποια επεξεργασία συμπίεσης που να εκμεταλλεύεται την αραιότητα του πίνακα. Η επεξεργασία όμως αυτή είναι πολύπλοκη και θα καθυστερούσε σημαντικά τις διαδικασίες ανάκτησης και ανανέωσης δεδομένων που είναι απαραίτητες να γίνονται συνεχώς στα συστήματα υπόδειξης.

Αυτό που επιλέχθηκε τελικά είναι, κατά την διάρκεια λειτουργίας του συστήματος, ο πίνακας βαθμολογιών να διατηρείται στην κύρια μνήμη. Όταν οι διάφορες διεργασίες του συστήματος χρειάζονται τις βαθμολογίες των χρηστών, αυτές ανακτώνται ταχύτατα. Όταν οι χρήστες βαθμολογούν νέες ταινίες, ο πίνακας

βαθμολογιών στην μνήμη ανανεώνεται και παράλληλα οι νέες βαθμολογήσεις καταγράφονται στο αρχείο βαθμολογιών. Από το αρχείο αυτό, φορτώνεται ο πίνακας βαθμολογιών στην κύρια μνήμη κατά την εκκίνηση του συστήματος.

Για τα υπόλοιπα δεδομένα που διαχειρίζεται το σύστημα θα γίνει λόγος στην περιγραφή των διεργασιών που τα χρησιμοποιούν.

3.3 Δομή του Συστήματος



Η γενική δομή του συστήματος φαίνεται σχηματικά στην εικόνα 1. Ο διαχειριστής του συστήματος ξεκινά την λειτουργία του συστήματος. Η διεργασία εκκίνησης φορτώνει τα δεδομένα του αρχείου βαθμολογιών στον πίνακα

βαθμολογιών που χρησιμοποιεί το σύστημα κατά τη λειτουργία του. Επίσης, ο διαχειριστής επιλέγει την μέθοδο υπόδειξης που θα χρησιμοποιείται για να παράγονται οι προβλέψεις για τους χρήστες. Οι μέθοδοι υπόδειξης χρησιμοποιούνται πίνακα βαθμολογιών, καθώς και το αρχείο ταινιών που περιέχει τα δεδομένα που έχει συλλέξει ο web crawler από την IMDb προκειμένου να παράγουν τις προσωποποιημένες υποδείξεις για τους χρήστες του συστήματος. Οι υποδείξεις προβάλλονται στους χρήστες μέσω της γραφικής διεπαφής. Επίσης μέσω της γραφικής διεπαφής οι χρήστες βαθμολογούν νέες ταινίες και οι οποίες, μέσω μιας διεργασίας καταχώρησης, προστίθενται στο αρχείο και τον πίνακα βαθμολογιών. Οι χρησιμοποιούμενες μέθοδοι υπόδειξης, καθώς και η γραφική διεπαφή περιγράφονται στις επόμενες παραγράφους.

3.4 Αλγόριθμοι Υπόδειξης

Η επιλογή της Συνεργατικής Μεθόδου Υπόδειξης για το σύστημα ήταν προφανής, όχι μόνο επειδή είναι η πιο ώριμη τεχνολογία και χρησιμοποιείται, με τον ένα ή τον άλλο τρόπο, στην συντριπτική πλειοψηφία των υπαρχόντων συστημάτων υπόδειξης, αλλά και γιατί η υλοποίηση και η λειτουργία της δεν δεσμεύεται από το πεδίο εφαρμογής και μπορεί να χρησιμοποιηθεί και αν αυτό αλλάξει.

Επίσης θεωρήθηκε απαραίτητη η υλοποίηση και κάποιου τρόπου Υπόδειξης Βασισμένου στο Περιεχόμενο με στόχο την σύγκριση με ΣΜΥ, αλλά και την ανάπτυξη υβριδικών μεθόδων. Τελικά υλοποιήθηκαν δύο προσεγγίσεις, μια βασισμένη στις πιθανότητες χρησιμοποιώντας Naïve Bayes, και μια χρησιμοποιώντας κλασσικές μεθόδους της Ανάκτησης Πληροφοριών. Ο λόγος που επιλέχθηκαν δύο διαφορετικές προσεγγίσεις ήταν για να διαπιστωθούν οι δυνατότητες και οι αδυναμίες κάθε μιας και να επιλεγεί η καλύτερη για το συγκεκριμένο σύνολο δεδομένων.

Οσο για τις Υβριδικές Μεθόδους, υλοποιήθηκαν δύο διαφορετικοί τρόποι συνδυασμού της ΣΜΥ και των ΥΒΠ και έγινε σύγκριση των επιδόσεων τους.

3.4.1 Συνεργατική Μέθοδος Υπόδειξης

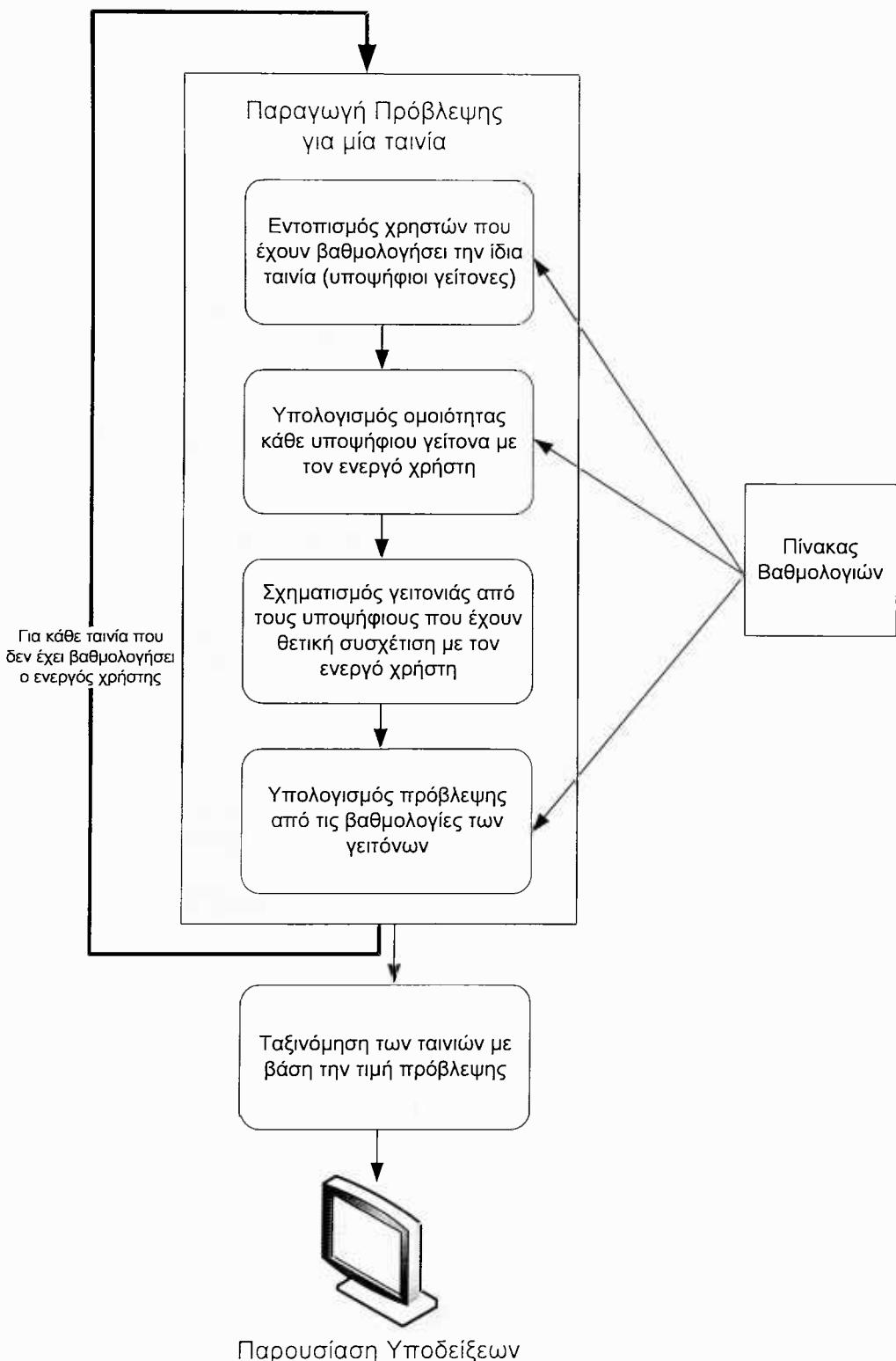
Επιλέχθηκε να υλοποιηθεί αλγόριθμος Βασισμένος στην Μνήμη. Ο λόγος που απορρίφθηκαν οι αλγόριθμοι Βασισμένοι σε Μοντέλα είναι, αφενός, ότι έχουν

χρησιμοποιηθεί λιγότερο από την ερευνητική κοινότητα και αφετέρου ότι προσθέτουν πολυπλοκότητα στο σύστημα. Ο αλγόριθμος που χρησιμοποιήθηκε είναι αυτός που περιγράφουν οι Resnick et al. (1994) και απεικονίζεται γραφικά στην Εικόνα 2.

Ως μέτρο ομοιότητας των χρηστών εφαρμόζεται ο συντελεστής συσχέτισης Pearson. Ο υπολογισμός του γίνεται με τον τύπο (2.2) για ταχύτερη εκτέλεση. Κατά την διάρκεια της ανάπτυξης του συστήματος παρατηρήθηκε το εξής φαινόμενο: αν δύο χρήστες έχουν n κοινές ταινίες και ο ένας έχει βαθμολογήσει και τις n με τον ίδιο βαθμό, τότε ο συντελεστής συσχέτισης Pearson δεν μπορούσε να υπολογιστεί, καθώς προέκυπτε κλάσμα $0/0$! Αυτό εξηγείται παρατηρώντας τον τύπο (2.1) (που υπενθυμίζουμε ότι είναι ισοδύναμος με τον τύπο (2.2)). Αν όλες οι βαθμολογίες ενός χρήστη είναι ίδιες, τότε ισούνται και με την μέση τιμή και επομένως πάντα ο παράγοντας $(X_i - \bar{X})$ ισούται με 0. Από όσο μπορούμε να γνωρίζουμε, το πρόβλημα αυτό δεν αναφέρεται πουθενά στην βιβλιογραφία. Μιας και το φαινόμενο (το οποίο το ονομάζουμε «φαινόμενο $0/0$ ») εμφανίζεται σπάνια, αποφασίστηκε όταν εμφανίζει να αγνοείται και στον συντελεστή συσχέτισης Pearson να ανατίθεται η τιμή 0.

Αν ο αριθμός (n) των κοινά βαθμολογημένων αντικειμένων των δύο χρηστών είναι μικρότερος του 50, η τιμή του μέτρου ομοιότητας πολλαπλασιάζεται με το Βάρος Σημαντικότητας $n/50$ (Herlocker et al. 1999).

Η γειτονιά που σχηματίζεται για την παραγωγή προβλέψεων για κάθε ταινία απαρτίζεται από τους χρήστες που έχουν βαθμολογήσει την ταινία αυτή και έχουν θετική συσχέτιση με τον ενεργό χρήστη, δηλαδή η τιμή του συντελεστή συσχέτισης Pearson είναι μεγαλύτερη του 0, αφού παρατηρήσαμε ότι συμπεριλαμβάνοντας τους χρήστες με αρνητική συσχέτιση δεν βελτιώνεται η ακρίβεια των προβλέψεων, ενώ αυξάνεται και ο χρόνος υπολογισμού τους. Δεν πραγματοποιείται καμία δειγματοληψία, αφού κάτι τέτοιο θα μείωνε πιθανώς την ακρίβεια των υποδείξεων και θα πρόσθετε τυχαιότητα στα αποτελέσματα. Επομένως, χρησιμοποιείται όλο το σύνολο χρηστών.



Εικόνα 2: Αλγόριθμος παραγωγής υποδείξεων με την Συνεργατική Μέθοδο Υπόδειξης

Για τον υπολογισμό των προβλέψεων χρησιμοποιείται ο τύπος (2.6). Παρατηρήσαμε ότι όταν η γειτονιά από την οποία παράγονται οι προβλέψεις είναι μικρή (δύο έως τέσσερις γείτονες) οι προβλέψεις δεν είναι καθόλου ακριβείς. Για την ακρίβεια παρατηρήθηκαν φαινόμενα κατά τα οποία προβλέψεις που είχαν προκύψει από τρεις γείτονες να έχουν τιμή εκτός βαθμολογικής κλίμακας (κάτω από ένα ή πολύ πάνω από πέντε)! Για να αποφευχθούν τέτοιες καταστάσεις αποφασίστηκε το σύστημα να παράγει προβλέψεις με την ΣΜΥ μόνο όταν για μια ταινία δημιουργείται γειτονιά πέντε ή περισσότερων γειτόνων. Με τον τρόπο αυτό θυσιάζεται ένα πολύ μικρό ποσοστό κάλυψης, προκειμένου να επιτευχθεί μεγαλύτερη ακρίβεια.

Όλοι οι υπολογισμοί που είναι απαραίτητοι για την παραγωγή υποδείξεων με την ΣΜΥ πραγματοποιούνται την στιγμή που ο χρήστης ζητάει υποδείξεις. Καμία ομοιότητα χρηστών δεν προϋπολογίζεται γιατί κάτι τέτοιο μπορεί να επηρεάσει αρνητικά την ακρίβεια των υποδείξεων. Βέβαια, εάν το MoRe λειτουργούσε σε πραγματικές συνθήκες (π.χ. στα πλαίσια ενός ηλεκτρονικού καταστήματος DVD), η χρησιμοποίηση προϋπολογισμένων ομοιοτήτων χρηστών θα ήταν ένας συμβιβασμός που ίσως θα ήταν απαραίτητο να γίνει, για να μειωθεί ο χρόνος παραγωγής των υποδείξεων.

3.4.2 Υποδείξεις Βασισμένες στο Περιεχόμενο

Αναπαράσταση Ταινιών

Οπως αναφέρθηκε στο Κεφάλαιο 2, βασική προϋπόθεση για την παραγωγή Υποδείξεων Βασισμένων στο Περιεχόμενο είναι να αναπαρασταθούν τα αντικείμενα με ένα αριθμό γνωρισμάτων (features). Τα γνωρίσματα που χρησιμοποιούμε για την αναπαράσταση των ταινιών είναι τα ονοματεπώνυμα των συντελεστών (ανεξάρτητα αν ανήκει σε σκηνοθέτη, ηθοποιό, σεναριογράφο ή παραγωγό), οι λέξεις πλοκής και οι λέξεις που δηλώνουν το είδος, που υπάρχουν στο αρχείο movies.dat.

Όσα γνωρίσματα υπάρχουν μόνο μια φορά (πχ ένας ηθοποιός που εμφανίζεται σε μόνο μια ταινία) αγνοούνται γιατί δεν προσφέρουν τίποτα στον εντοπισμό όμοιων ταινιών. Για τις 3952 ταινίες του συνόλου δεδομένων, τα μη-μοναδικά γνωρίσματα είναι 10626.

Η κάθε ταινία αντιπροσωπεύεται από ένα διάνυσμα μήκους ίσο με τον αριθμό των μη-μοναδικών γνωρισμάτων. Οι συντεταγμένες του διανύσματος αντιστοιχούν σε ένα γνώρισμα. Αν η ταινία περιλαμβάνει το γνώρισμα, η τιμή της συντεταγμένης είναι 1, αλλιώς είναι 0. Πρόκειται δηλαδή για Boolean διανύσματα.

Τα διανύσματα αυτά δεν μεταβάλλονται, παρά μόνο αν προστεθούν νέες ταινίες στο σύστημα, οπότε αλλάζει το πλήθος των μη-μοναδικών γνωρισμάτων. Αποθηκεύονται σε ένα αρχείο και όποτε απαιτείται ανακτώνται από εκεί. Επειδή μάλιστα τα διανύσματα είναι πολύ αραιά, για κάθε διάνυσμα αποθηκεύονται στο αρχείο μόνο οι συντεταγμένες οι τιμές των οποίων δεν είναι 0.

Αλγόριθμοι

Όπως αναφέραμε παραπάνω, αποφασίστηκε να υλοποιηθούν δύο διαφορετικές προσεγγίσεις για την παραγωγή Υποδείξεων Βασισμένων στο Περιεχόμενο.

Η πρώτη προσέγγιση βασίζεται στον αλγόριθμο που περιγράφεται στην παράγραφο 2.4.2 (Kagypis. 2001). Για την παράμετρο K επιλέχθηκε μετά από πειραματισμό η τιμή 5, καθώς με μικρότερες τιμές είχαμε πτώση της ακρίβειας και με μεγαλύτερες τιμές αυξανόταν ο χρόνος υπολογισμού των υποδείξεων. Ως μέτρο ομοιότητας δύο αντικειμένων χρησιμοποιούμε το συνημίτονο της γωνίας των διανυσμάτων τους, που υπολογίζεται από τον τύπο (2.5).

Σε αντίθεση με την ΣΜΥ, όπου οι ομοιότητες μεταξύ χρηστών εξαρτώνται από τις ταινίες που έχουν βαθμολογήσει από κοινού και επομένως μπορεί να μεταβάλλονται καθώς οι χρήστες χρησιμοποιούν το σύστημα., στις ΥΒΠ οι ομοιότητες μεταξύ αντικειμένων δεν αλλάζουν κατά την διάρκεια της λειτουργίας του συστήματος. Εκμεταλλευόμενοι το γεγονός αυτό, όλες οι ομοιότητες μεταξύ των ταινιών υπολογίζονται από πριν και αποθηκεύονται σε αρχεία. Σε ένα ενιαίο αρχείο αποθηκεύονται για κάθε ταινία οι K (=5) πιο δύοιες ταινίες. Σε ξεχωριστά αρχεία για κάθε ταινία αποθηκεύονται οι ομοιότητες της ταινίας με όλες τις υπόλοιπες ταινίες σε φθίνουσα σειρά. Δημιουργώντας ξεχωριστά αρχεία για κάθε ταινία θυσιάζεται μεγάλος χώρος στο σκληρό δίσκο. Επιτυγχάνονται όμως πολύ μικρότεροι χρόνοι αναζήτησης ομοιοτήτων κατά τη φάση λειτουργίας του συστήματος.

Ο αλγόριθμος λοιπόν όπως υλοποιείται στο σύστημα MoRe περιλαμβάνει τα παρακάτω βήματα :

- **Φάση προπαρασκευής**

1. Υπολογισμός για κάθε ταινία των ομοιοτήτων με όλες τις υπόλοιπες χρησιμοποιώντας το μέτρο του συνημίτονου και αποθήκευσή τους σε φθίνουσα σειρά ξεχωριστό αρχείο.
2. Αποθήκευση των 5 πιο όμοιων ταινιών για κάθε ταινία σε ενιαίο αρχείο.

- **Φάση λειτουργίας του συστήματος**

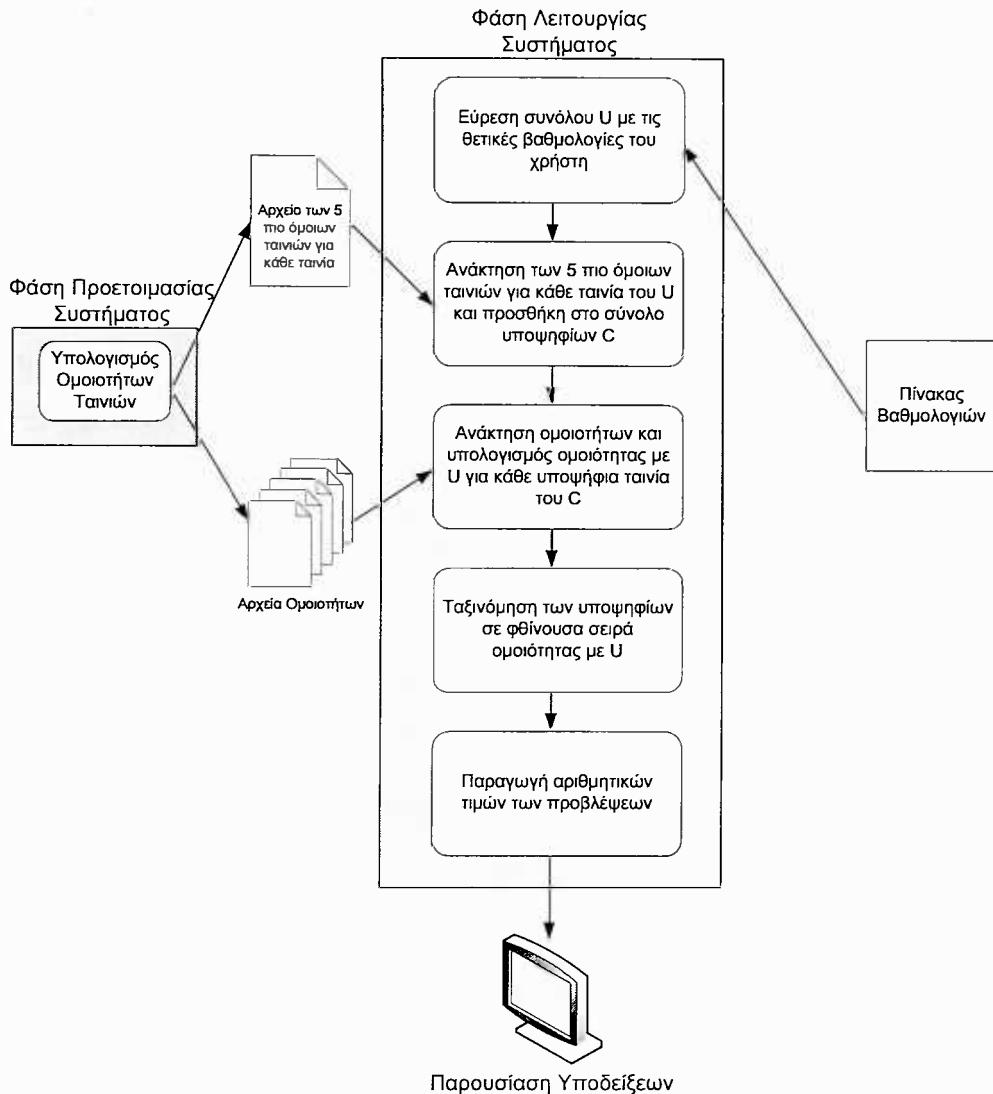
3. Θεωρούμε σύνολο U με όλες τις ταινίες που έχει βαθμολογήσει θετικά (4 ή 5) ο ενεργός χρήστης.
4. Για κάθε ταινία του U ανακτούμε τις 5 πιο όμοιες ταινίες και όσες από αυτές δεν ανήκουν στο U τις προσθέτουμε στο σύνολο C των υποψηφίων.
5. Για κάθε υποψήφια ταινία ανακτούμε τις ομοιότητες της με όλες τις ταινίες του U και τις αθροίζουμε. Το άθροισμα αυτό θεωρούμε ότι αποτελεί την ομοιότητα της ταινίας με το σύνολο U και τις προτιμήσεις του ενεργού χρήστη γενικότερα.
6. Ταξινομούμε τις υποψήφιες ταινίες σε μη-φθίνουσα σειρά και υποδεικνύουμε τις υποδεικνύουμε στο χρήστη.

Ο Kaguris (2001) σταματάει τον αλγόριθμο εδώ χωρίς να προβαίνει σε αριθμητική πρόβλεψη για κάθε ταινία χωριστά. Επιθυμούμε, όμως, την παραγωγή αριθμητικών τιμών για τις προβλέψεις, αφού σκοπεύουμε να χρησιμοποιήσουμε αυτή τη μέθοδο υπόδειξης σε μια υβριδική προσέγγιση σε συνδυασμό με την ΣΜΥ, που ως γνωστό παράγει αριθμητικές προβλέψεις.

Προτείνουμε την επέκταση του αλγόριθμου ως εξής: Έστω MaxSim , MinSim η μέγιστη και ελάχιστη, αντίστοιχα. ομοιότητα που παρατηρείται μεταξύ των ταινιών του συστήματος και του συνόλου U , όπως προκύπτει μετά το βήμα 5. Έστω επίσης Sim_i η ομοιότητα της ταινίας M_i με το σύνολο U . Η αριθμητική τιμή της πρόβλεψης Pr_i για την ταινία είναι:

$$\text{Pr}_i = \frac{(\text{Sim}_i - \text{MinSim}) * 4}{(\text{MaxSim} - \text{MinSim})} + 1 \quad (3.1)$$

Γίνεται δηλαδή αντιστοίχηση της ομοιότητας από το διάστημα $[\text{MaxSim}, \text{MinSim}]$ στο διάστημα $[1,5]$ που είναι η βαθμολογική κλίμακα που χρησιμοποιείται. Η εικόνα 3 παρουσιάζει μια γραφική απεικόνιση του τελικού αλγορίθμου.



Εικόνα 3. Αλγόριθμος Παραγωγής Υποδείξεων Βασισμένων στο Περιεχόμενο

Η δεύτερη προσέγγιση βασίζεται σε πιθανοθεωρητικό μοντέλο και χρησιμοποιεί τον κατηγοριοποιητή Naïve Bayes. Περιγραφή του κατηγοριοποιητή έγινε στην παράγραφο 2.4.2 και ο τρόπος λειτουργίας του στο σύστημα MoRe είναι πιστός στην περιγραφή αυτή.

Υπενθυμίζουμε ότι ο Naïve Bayes προσπαθεί να υπολογίσει την πιθανότητα μία ταινία να βαθμολογηθεί με κάθε μια από τις τιμές της κλίμακας βαθμολογίας, στην περίπτωσή μας με τις τιμές 1, 2, 3, 4 ή 5. Η πρόβλεψη που παράγει το σύστημα

για την ταινία είναι η τιμή της βαθμολογικής κλίμακας για την οποία υπολογίστηκε η μεγαλύτερη πιθανότητα.

3.4.3 Υβριδικές Μέθοδοι Υπόδειξης

Η δημιουργία μιας Υβριδικής Μεθόδου Υπόδειξης έχει όπως είναι λογικό, ως στόχο την βελτίωση των κλασσικών μεθόδων υπόδειξης. Τι εννοούμε, όμως, με τον όρο βελτίωση; Για τον χώρο των συστημάτων υπόδειξης βελτίωση θεωρείται η επίτευξη ενός από τους παρακάτω στόχους:

- Παραγωγή προβλέψεων μεγαλύτερης ακρίβειας (μείωση του συνολικού Μέσου Απόλυτου Λάθους του συστήματος), που είναι και ο πιο συνηθισμένος στόχος.
- Επίτευξη μεγαλύτερης κάλυψης (coverage) του συνόλου των αντικειμένων του συστήματος, αν η κάλυψη ήταν μικρότερη του 100%.
- Επίτευξη μικρότερου χρόνου παραγωγής υποδείξεων.

Οι υβριδικές μέθοδοι υπόδειξης που θα περιγραφούν στη συνέχεια πραγματοποιήθηκαν έχοντας αυτούς τους στόχους υπόψη.

Μέθοδος Υποκατάστασης (Substitute)

Όταν περιγράψαμε τον τρόπο παραγωγής προβλέψεων με την ΣΜΥ (3.2.1), αναφερθήκαμε στην απόφαση να μην παράγεται πρόβλεψη για μια ταινία αν το μέγεθος της γειτονιάς του ενεργού χρήστη για την ταινία αυτή δεν έχει μέγεθος τουλάχιστον 5 γειτόνων, οι οποίοι να έχουν θετική συσχέτιση με τον ενεργό χρήστη. Η απόφαση αυτή ελήφθη με κριτήριο την επίτευξη της μέγιστης δυνατής ακρίβειας.

Θυσιάστηκε όμως ένα μικρό ποσοστό κάλυψης.

Σκοπός της υβριδικής μεθόδου υποκατάστασης (substitute) είναι να εξασφαλιστεί κάλυψη 100%, χωρίς όμως να μειωθεί η ακρίβεια του συστήματος.

Η λογική είναι πολύ απλή: Όταν δεν είναι δυνατόν να παραχθεί υπόδειξη για μια ταινία χρησιμοποιώντας την Συνεργατική Μέθοδο Υπόδειξης, τότε παράγεται Υπόδειξη Βασισμένη στο Περιεχόμενο. Επειδή οι ΥΒΠ μπορούν πάντα να παράγουν πρόβλεψη, είναι σίγουρο ότι και το σύστημα θα μπορεί να παράγει κάποια πρόβλεψη για όλες τις ταινίες και επομένως επιτυγχάνεται κάλυψη 100%.

Τελικά, η μέθοδος υποκατάστασης, όχι μόνο διατηρεί την ακρίβεια της ΣΜΥ, αλλά την βελτιώνει και ελάχιστα.



Πρέπει στο σημείο αυτό να γίνουν δύο παρατηρήσεις. Η πρώτη αφορά τον χρόνο εκτέλεσης του αλγορίθμου ο οποίος είναι ελαφρά μεγαλύτερος από τον χρόνο εκτέλεσης της ΣΜΥ. Ο λόγος είναι προφανής: εκτελεί τα ίδια βήματα υπολογισμού ομοιοτήτων και επιλογής γειτόνων όπως και η ΣΜΥ, αλλά αν τελικά δεν μπορεί να γίνει πρόβλεψη, εκτελούνται και τα βήματα για την παραγωγή ΥΒΠ. Βέβαια, οι ΥΒΠ παράγονται πολύ γρηγορότερα από τις υποδείξεις που παράγονται με τη ΣΜΥ και έτσι η επιπλέον καθυστέρηση μπορεί να θεωρηθεί αμελητέα από τον χρήστη.

Η δεύτερη παρατήρηση αφορά τον τρόπο υπολογισμού της αριθμητικής τιμής της πρόβλεψης για μια μόνο ταινία στις ΥΒΠ. Στον τύπο (3.1) υπεισέρχεται η μέγιστη και ελάχιστη τιμή ομοιότητας με τις προτιμήσεις του χρήστη που εμφανίζονται σε όλες τις ταινίες του συστήματος. Αυτό σημαίνει ότι ακόμα και όταν θέλουμε πρόβλεψη για μια μόνο ταινία (όπως μπορεί να συμβαίνει εφαρμόζοντας την μέθοδο υποκατάστασης), είμαστε υποχρεωμένοι να παράγουμε πρόβλεψης για όλα τα αντικείμενα. Αυτό δεν είναι τόσο κακό, όσο μπορεί να φαίνεται αρχικά. Αφενός, ο χρόνος υπολογισμού των ΥΒΠ είναι πολύ σύντομος και αφετέρου σχεδόν ποτέ δεν ζητείται από ένα σύστημα υπόδειξης να κάνει πρόβλεψη για όσες ταινίες ο χρήστης δεν έχει βαθμολογήσει, οι οποίες είναι τις περισσότερες φορές χλιάδες στον αριθμό. Στις περιπτώσεις αυτές υπολογίζονται μια φορά οι ΥΒΠ για όλες τις ταινίες και στη συνέχεια χρησιμοποιούνται όσες φορές χρειαστεί.

Μέθοδος Αλλαγής (Switching)

Έχει γίνει λόγος για την αδυναμία της ΣΜΥ να επιτύχει υψηλή ακρίβεια όταν ο ενεργός χρήστης έχει βαθμολογήσει μικρό αριθμό ταινιών. Αντίθετα, οι ΥΒΠ δεν αντιμετωπίζουν τόσο έντονο το πρόβλημα αυτό. Έτσι, γεννάται εύκολα η ιδέα της χρησιμοποίησης ΥΒΠ στις περιπτώσεις εκείνες που υπάρχουν λίγες βαθμολογίες ταινιών από το χρήστη και της ΣΜΥ όταν πλέον είναι διαθέσμες πολλές βαθμολογίες.

Αυτή είναι και η λογική της υβριδικής μεθόδου αλλαγής: όσο ο χρήστης έχει βαθμολογήσει λιγότερες από N ταινίες λαμβάνει ΥΒΠ, αλλιώς λαμβάνει υποδείξεις από την ΣΜΥ. Το ζητούμενο είναι ο εντοπισμός του αριθμού N. Πειραματικά υπολογίστηκε ότι η μεγαλύτερη ακρίβεια επιτυγχάνεται για τιμές κοντά στο 40.

Δυστυχώς, η ακρίβεια της μεθόδου αυτής είναι λίγο μικρότερη από αυτή της ΣΜΥ. Ο χρόνος όμως υπολογισμού των υποδείξεων είναι βελτιωμένος. Αυτό οφείλεται στο γεγονός ότι στις περιπτώσεις που υπάρχουν λίγες βαθμολογίες παράγονται ΥΒΠ που υπολογίζονται πολύ γρήγορα, ενώ οι περιπτώσεις όπου υπάρχει μεγάλος αριθμός βαθμολογιών από το χρήστη (και συνεπώς λιγότερες υποψήφιες ταινίες προς υπόδειξη) η ΣΜΥ παράγει υποδείξεις πιο γρήγορα, καθώς η πολυπλοκότητά της είναι ανάλογη των υποψηφίων προς υπόδειξη ταινιών.

3.5 Αντιμετώπιση του Προβλήματος του Νέου Χρήστη

Το πρόβλημα του νέου χρήστη είναι κοινό και για τις δύο κλασσικές μεθόδους υπόδειξης και επομένως και για τις υβριδικές που παράγονται από αυτές. Υπενθυμίζουμε ότι αναφέρεται στην ανάγκη παραγωγής όσο το δυνατόν πιο ακριβών υποδείξεων με όσο το δυνατόν λιγότερων βαθμολογήσεων από τον ένα νέο χρήστη.

Το σύστημα MoRe αντιμετωπίζει αυτό το πρόβλημα με τον τρόπο που προτείνουν οι Rashid et al. (2002). Όταν εισαχθεί ένας νέος χρήστης στο σύστημα, για όλες τις διαθέσιμες ταινίες του συστήματος υπολογίζεται η δημοτικότητά (popularity) τους, δηλαδή ο λόγος του αριθμού των χρηστών που την έχουν βαθμολογήσει προς τον συνολικό αριθμό των χρηστών, καθώς και μέτρο της εντροπίας (entropy) τους. Στην συνέχεια οι ταινίες ταξινομούνται σε φθίνουσα σειρά σύμφωνα με το μέτρο $\log(\text{popularity}) * \text{entropy}$ και παρουσιάζονται στον χρήστη για να τις βαθμολογήσει. Όταν ο χρήστης βαθμολογήσει N ταινίες, το σύστημα είναι σε θέση να κάνει προβλέψεις. Η μεταβλητή N μπορεί να πάρει οποιαδήποτε τιμή μεγαλύτερη του 1. Όσο μικρότερη όμως είναι η τιμή του N , τόσο χαμηλότερη η ακρίβεια των πρώτων υποδείξεων του συστήματος. Για το N έχουμε επιλέξει την τιμή 20.

Η ύπαρξη της εντροπίας στον τύπο ταξινόμησης των ταινιών εξασφαλίζει ότι οι ταινίες που θα βαθμολογήσει αρχικά ο χρήστης θα περιέχουν αρκετή πληροφορία που θα αποκαλύπτουν τις προτιμήσεις του χρήστη. Η δημοτικότητα εξασφαλίζει ότι ο χρήστης θα έχει δει τις ταινίες που το παρουσιάζει το σύστημα και έτσι δεν θα χρειαστεί να αναζητά πολύ ώρα για να βαθμολογήσει τον απαραίτητο αριθμό ταινιών.

Ο στόχος της παραπάνω προσέγγισης είναι η παραγωγή υποδείξεων υψηλής ακρίβειας για το χρήστη όταν αυτός ολοκληρώσει την διαδικασία βαθμολόγησης. Ο

στόχος αυτός δεν εξασφαλίζεται πάντα για όλους τους χρήστες. Κάποιοι χρήστες θα πρέπει να βαθμολογήσουν έναν επιπλέον αριθμό ταινιών για να λάβουν ακριβείς υποδείξεις. Άλλωστε, όσο οι χρήστες βαθμολογούν ταινίες, τόσο βελτιώνονται και οι υποδείξεις του συστήματος για αυτούς. Αυτό, λοιπόν, που προσφέρει η προσέγγιση των Rashid et al (2002) δεν είναι τίποτε άλλο από μια επιτάχυνση της διαδικασίας εκπαίδευσης του συστήματος στις προτιμήσεις του χρήστη.

3.6 Γραφική Διεπαφή

Η γραφική διεπαφή για ένα σύστημα που προορίζεται για δικτυακή χρήση είναι κεφαλαιώδους σημασίας. Η διεπαφή πρέπει να είναι φιλική προς το χρήστη και λειτουργική, χωρίς να τον κουράζει με περιττές οθόνες και χωρίς να τον «πνίγει» με μεγάλους όγκους δεδομένων ανά οθόνη.

Η γλώσσα που χρησιμοποιείται στην διεπαφή είναι η αγγλική. Η πρώτος λόγος που οδήγησε στην επιλογή αυτή είναι ότι η χρησιμοποίηση της ελληνικής γλώσσας θα ήταν περιοριστικός παράγοντας για μελλοντικούς χρήστες του συστήματος εκτός Ελλάδας. Ο δεύτερος λόγος είναι ότι οι πληροφορίες για τις ταινίες που παρέχονται από την IMDb είναι στα αγγλικά και θα ήταν άσχημο να υπάρχει διγλωσσία στο σύστημα.

Υπάρχουν δύο ξεχωριστές γραφικές διεπαφές. Η μια είναι αυτή που βλέπουν οι χρήστες του συστήματος και η άλλη χρησιμοποιείται από τον διαχειριστή του συστήματος.

Η διεπαφή του χρήστη θέλουμε να είναι όσο πιο απλή γίνεται. Κάνοντας login, θα πρέπει αμέσως να οδηγείται στην οθόνη παρουσίασης των προσωπικών του υποδείξεων. Παρουσιάζονται πέντε ταινίες ανά σελίδα, έτσι ώστε η σελίδα να ανοίγει γρήγορα, ακόμα και σε φυλλομετρητές (browsers) χρηστών που διαθέτουν αργή σύνδεση, αλλά και ο χρήστης να μην βρίσκεται αντιμέτωπος με μεγάλες σελίδες γεμάτες πληροφορίες. Για κάθε ταινία παρουσιάζεται ο τίτλος της, το έτος προβολής, το είδος της, οι βασικοί πρωταγωνιστές, ο σκηνοθέτης, οι τρεις σημαντικότεροι σεναριογράφοι και παραγωγοί (αν υπάρχουν περισσότεροι των τριών). Εάν ο χρήστης έχει δει την ταινία, μπορεί με ένα κλικ να την βαθμολογήσει, έτσι ώστε το σύστημα να χρησιμοποιήσει την ταινία αυτή για την παραγωγή μελλοντικών υποδείξεων. Οι προβλέψεις του συστήματος δεν παρουσιάζονται με την αριθμητική τους τιμή αλλά

με γραφική απεικόνιση που χρησιμοποιεί αστέρια. Πέντε αστέρια δηλώνουν τιμή κοντά στο 5, τέσσερα ολόκληρα αστέρια και μισό αστέρι δηλώνουν τιμή κοντά στο 4.5 κοκ. Η αναπαράσταση των προβλέψεων με τον τρόπο αυτό είναι διαισθητικά κατανοητή από τους χρήστες και θυμίζει τις βαθμολογήσεις των κριτικών ταινιών που δημοσιεύονται τον Τύπο.

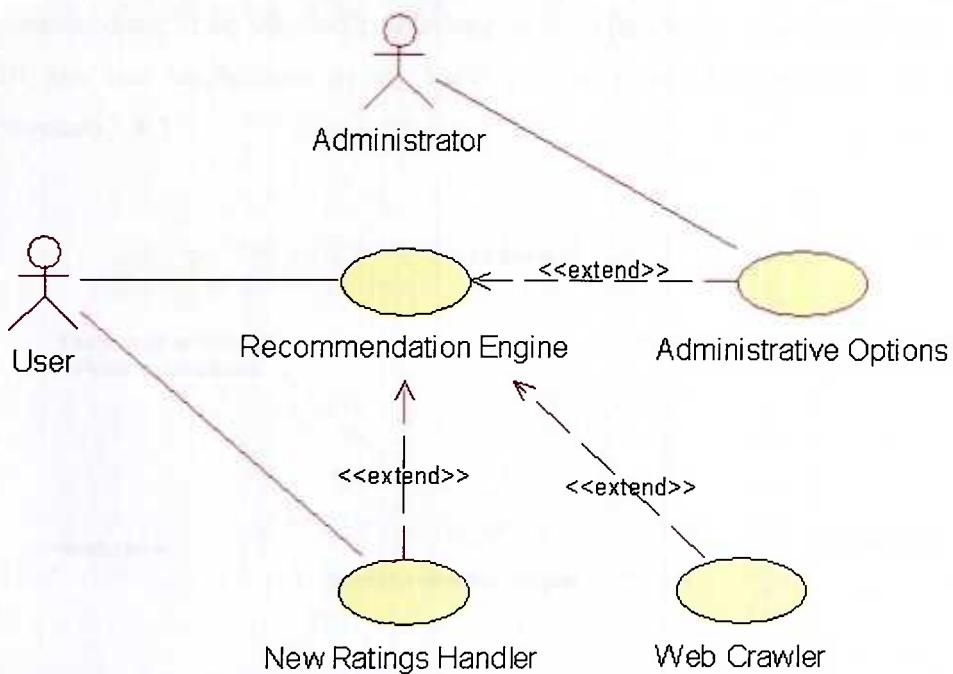
Η διεπαφή του διαχειριστή του επιτρέπει να ξεκινά τη λειτουργία του συστήματος, να επιλέξει με ποια μέθοδο θα γίνονται υποδείξεις στους χρήστες, να προσθέτει και να αφαιρεί ταινίες και να διαγράφει χρήστες, να ξεκινά των προϋπολογισμό των ομοιοτήτων των ταινιών που χρησιμοποιούνται στις YBΠ και να υπολογίζει την ακρίβεια του συστήματος.

Στο επόμενο κεφάλαιο θα γίνει λόγος για την υλοποίηση του συστήματος με τις τεχνολογικές λεπτομέρειες που σχετίζονται με αυτή.

4. Σχεδίαση και υλοποίηση συστήματος

4.1 Σχεδίαση Συστήματος

Η σχεδίαση του συστήματος ακολουθεί τις σχεδιαστικές αποφάσεις του προηγούμενου κεφαλαίου. Για την αναπαράσταση του συστήματος θα χρησιμοποιηθούν διαγράμματα UML (Unified Modeling Language – Ενοποιημένη Γλώσσα Μοντελοποίησης). Στο Σχήμα 1 φαίνεται το γενικό διάγραμμα περίπτωσης χρήσης (Use Case Diagram) του Συστήματος.



Σχήμα 1: Διάγραμμα Περίπτωσης Χρήστης – Use Case Diagram

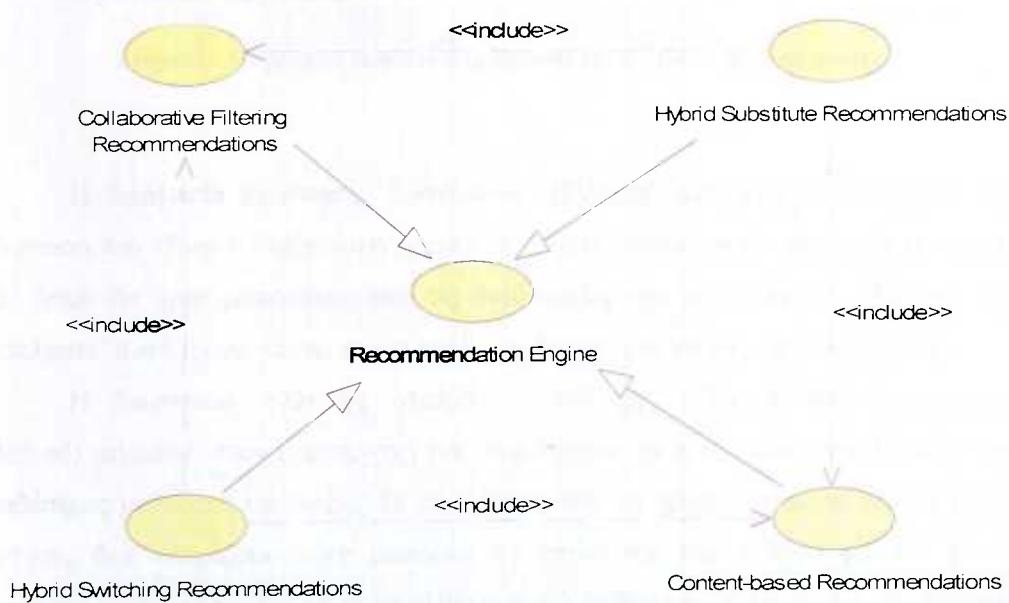
Τα βασικά συστατικά – υποσυστήματα είναι ο Μηχανισμός Παραγωγής Υποδείξεων (Recommendation Engine), ο Διαχειριστής Νέων Βαθμολογιών (New Ratings Handler), ο Μηχανισμός Αναζήτησης Δεδομένων από το Διαδίκτυο (Web Crawler) και οι Επιλογές Διαχείρισης (Administrative Options).

Οι χρήστες του συστήματος αλληλεπιδρούν άμεσα με το Μηχανισμό Παραγωγής Υποδείξεων, με την έννοια ότι αιτούν υποδείξεις και το σύστημα τις

προσφέρει, και με το Διαχειριστή Νέων Βαθμολογιών, αφού καταχωρούν σε αυτόν τις βαθμολογίες τους.

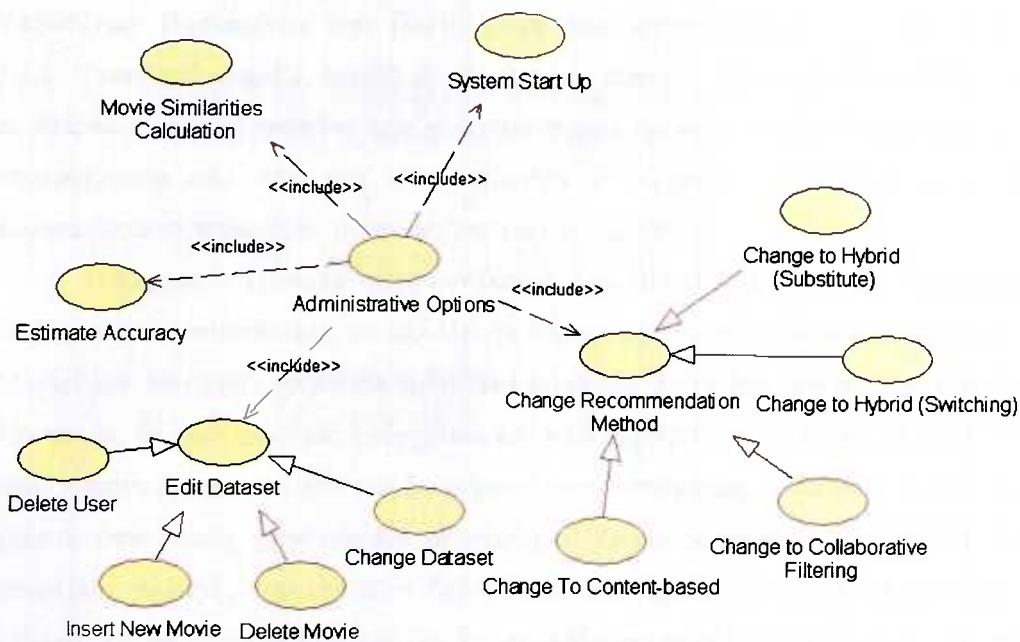
Ο Διαχειριστής Νέων Βαθμολογιών επεκτείνει τη λειτουργικότητα του Μηχανισμού Παραγωγής Υποδείξεων, αφού οι νέες βαθμολογίες ενός χρήστη βοηθούν στην παραγωγή υποδείξεων μεγαλύτερης ακρίβειας για αυτόν. Ο Web Crawler αναζητά πληροφορίες για τις ταινίες στο Διαδίκτυο και επίσης επεκτείνει τον Μηχανισμό Παραγωγής Υποδείξεων, αλλά μόνο το κομμάτι του που έχει σχέση με τις Υποδείξεις Βασισμένες στο Περιεχόμενο.

Ο Μηχανισμός Παραγωγής Υποδείξεων μπορεί να αναλυθεί περαιτέρω και παρουσιάζεται στο διάγραμμα περίπτωσης χρήστης του Σχήματος 2. Αποτελείται από τα υποσυστήματα παραγωγής υποδείξεων με την Συνεργατική Μέθοδο Υπόδειξης, Υποδείξεων Βασισμένων στο Περιεχόμενο και τις Υβριδικές Μεθόδους Αλλαγής και Υποκατάστασης. Στις υβριδικές μεθόδους περιλαμβάνεται η λειτουργικότητα των ΥΒΠ και των υποδείξεων με τη ΣΜΥ με τον τρόπο που περιγράφεται στην παράγραφο 3.4.3.



Σχήμα 2 : Διάγραμμα Περίπτωσης Χρήστης για το Μηχανισμό Παραγωγής Υποδείξεων

Οι Επιλογές Διαχείρισης επιτρέπουν στον διαχειριστή του συστήματος να ορίσει όλες τις παραμέτρους που σχετίζονται με την λειτουργία του συστήματος και παρουσιάζονται στο διάγραμμα περίπτωσης χρήσης του Σχήματος 3.



Σχήμα 3 : Διάγραμμα Περίπτωσης Χρήσης για τις Επιλογές Διαχείρισης

Η διεργασία Εκκίνησης Συστήματος (System Start Up) περιλαμβάνει την φόρτωση του Πίνακα Βαθμολογιών από το αρχείο βαθμολογιών στην κύρια μνήμη, απ' όπου θα χρησιμοποιείται από τις διαδικασίες του συστήματος. Εάν δεν έχει εκτελεστεί αυτή η διεργασία, καμία υπόδειξη δεν μπορεί να γίνει στους χρήστες.

Η διεργασία Αλλαγής Μεθόδου Υπόδειξης (Change Recommendation Method) επιτρέπει στον διαχειριστή του συστήματος να ορίσει ποια από τις τέσσερις διαθέσιμες μεθόδους υπόδειξης θα είναι αυτή που θα χρησιμοποιεί το σύστημα. Οι χρήστες δεν γνωρίζουν ούτε μπορούν να επιλέξουν την μέθοδο με την οποία παράγονται εκείνη τη στιγμή οι υποδείξεις που λαμβάνουν. Κάτι τέτοιο δεν θα ήταν πρακτικό, αφού οι περισσότεροι χρήστες των συστημάτων υπόδειξης δεν γνωρίζουν τίποτα για τις τεχνικές υπόδειξης και μια τέτοια επιλογή θα προκαλούσε πρόβλημα στη χρησιμοποίηση του συστήματος.

Η Εκτίμηση της Ακρίβειας (Estimate Accuracy) υπολογίζει το Μέσο Απόλυτο Λάθος του συστήματος για την μέθοδο που χρησιμοποιεί εκείνη τη στιγμή. Τα

αποτελέσματα αποθηκεύονται σε μορφή που να επιτρέπει την εύκολη επεξεργασία τους από εργαλεία στατιστικής ανάλυσης (π.χ SPSS).

Ο Υπολογισμός των Ομοιοτήτων των Ταινιών (Movie Similarities Calculation) εκτελεί τα βήματα προπαρασκευής για τον αλγόριθμο παραγωγής Υποδείξεων Βασισμένων στο Περιεχόμενο που περιγράφονται στην παράγραφο 3.4.2. Υπενθυμίζουμε ότι, επειδή το σύνολο των ταινιών δεν αλλάζει δυναμικά κατά τη διάρκεια της λειτουργίας του συστήματος, οι ομοιότητες μεταξύ των ταινιών υπολογίζονται από πριν και αποθηκεύονται σε αρχεία, από όπου ανακτώνται προκειμένου να παραχθούν οι υποδείξεις για τους χρήστες.

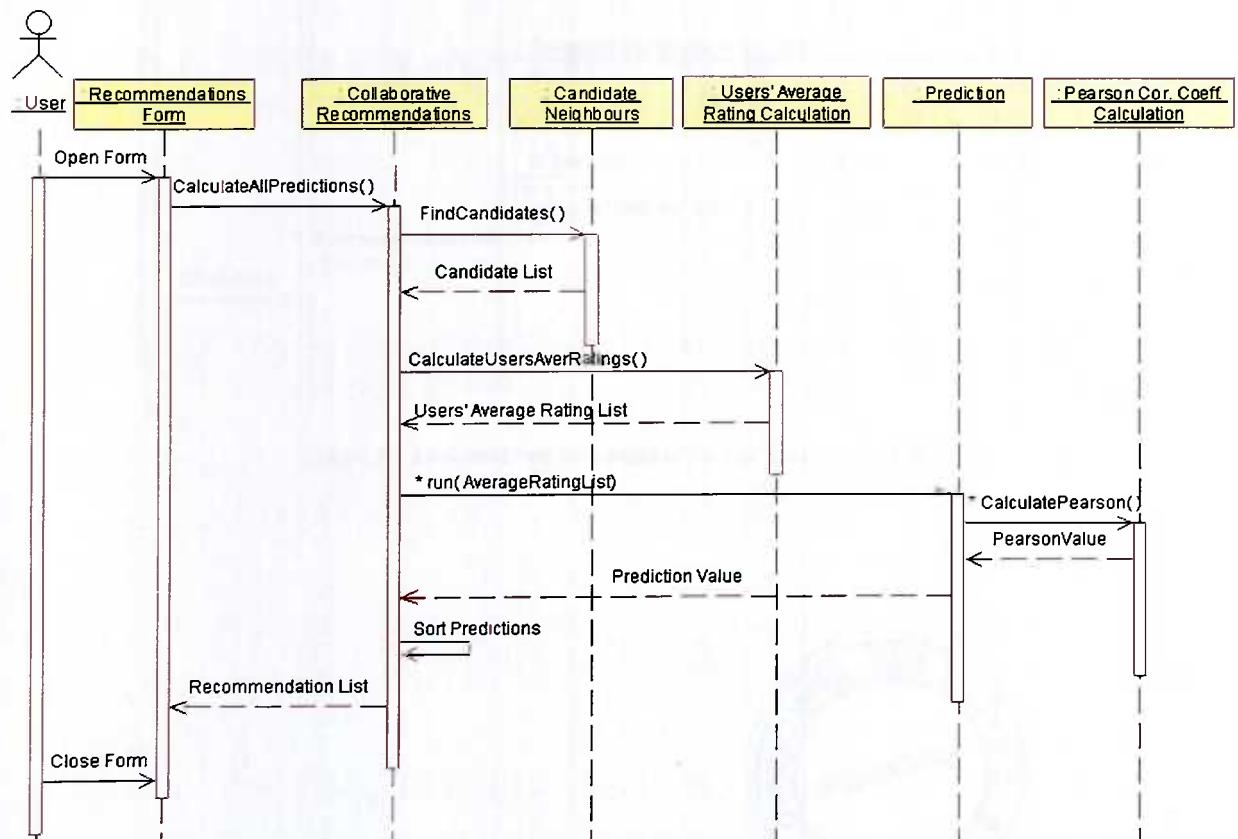
Η διεργασία Τροποποίησης Συνόλου Δεδομένων (Edit Dataset) επιτρέπει στον διαχειριστή του συστήματος να αλλάξει τα περιεχόμενα του συνόλου δεδομένων του MovieLens που έχει χρησιμοποιηθεί στο σύστημα. Είναι δυνατόν να εισαχθεί μια νέα ταινία, να διαγραφεί μια υπάρχουσα και να διαγραφεί ένας χρήστης. Η εισαγωγή νέου χρήστη δεν γίνεται από τον διαχειριστή του συστήματος, αλλά από τον ίδιο τον χρήστη όταν αυτός χρησιμοποιεί το σύστημα. Επιπλέον παρέχεται η επιλογή της συνολικής αλλαγής του συνόλου δεδομένων. Εδώ πρέπει να σημειώσουμε ότι η αλλαγή του συνόλου δεδομένων δεν θα προκαλούσε πρόβλημα στη λειτουργία της ΣΜΥ, ενώ αντίθετα θα υπήρχαν προβλήματα με την λειτουργία των ΥΒΠ (και κατ' επέκταση των υβριδικών μεθόδων), αφού αυτή εξαρτάται από το πεδίο εφαρμογής και τον τρόπο περιγραφής των αντικειμένων. Για συνεχίσουν να παράγονται κανονικά ΥΒΠ θα πρέπει οι περιγραφές των αντικειμένων να είναι συνεπείς με το μορφότυπο που περιγράφεται στην παράγραφο 3.1.

Οι δυνατότητες για αλλαγή της μεθόδου υπόδειξης που χρησιμοποιείται από το σύστημα, αλλαγή του συνόλου δεδομένων και εκτίμηση της ακρίβειας των παραγόμενων υποδείξεων επιτρέπουν στο σύστημα MoRe να χρησιμοποιηθεί για πειραματισμό και εξαγωγή συμπερασμάτων.

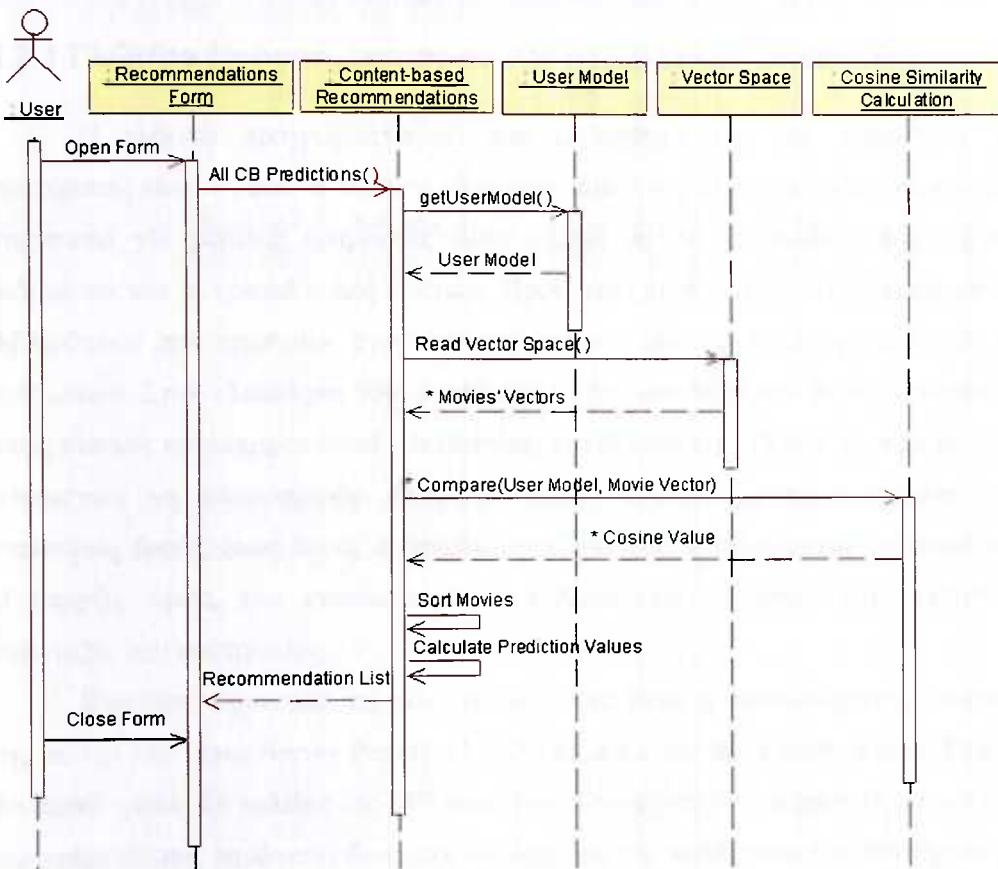
Η UML επιτρέπει την ανάλυση των διαγραμμάτων περίπτωσης χρήσης σε πολύ μεγάλο βάθος. Κάτι τέτοιο όμως ξεφεύγει από τους σκοπούς της παρούσας εργασίας. Ήα παραθέσουμε ενδεικτικά μόνο δύο διαγράμματα ακολουθίας (sequence diagrams) για την παραγωγή Υποδείξεων Βασισμένων στο Περιεχόμενο και υποδείξεων με την Συνεργατική Μέθοδο Υπόδειξης που αποτελούν τον πυρήνα του συστήματος αλλά και της εργασίας. Πρέπει να παρατηρήσουμε εδώ ότι οι τάξεις και οι μέθοδοι που εμφανίζονται στα διαγράμματα ακολουθίας έχουν υλοποιηθεί σε Java αλλά τα ονόματά τους είναι ελαφρά τροποποιημένα, ώστε να γίνουν πιο περιγραφικά

και ευκολοδιάβαστα. Η τάξη Recommendations Form είναι κοινή και για τα δύο διαγράμματα, αφού ο χρήστης, ανεξάρτητα του τρόπου παραγωγής των υποδείξεων, βλέπει τα αποτελέσματα με τον ίδιο ακριβώς τρόπο.

Το Σχήμα 4 απεικονίζει τον τρόπο παραγωγής Υποδείξεων με την ΣΜΥ και το Σχήμα 5 τον τρόπο παραγωγής ΥΒΠ. Τα βέλη δηλώνουν μηνύματα μεταξύ των αντικειμένων. Το βέλος με μισή αιχμή δηλώνει ότι το αντικείμενο προορισμού (η τάξη Prediction) λειτουργεί ως νήμα (thread) για παράλληλη επεξεργασία. Τα μηνύματα με αστερίσκο (*) πριν το όνομά τους στέλνονται περισσότερες της μιας φορά, πιθανότατα με διαφορετικές παραμέτρους κάθε φορά.



Σχήμα 4 : Ακολουθιακό Διάγραμμα Παραγωγής Υποδείξεων με την ΣΜΥ



Σχήμα 5 : Ακολουθιακό Διάγραμμα για την Παραγωγή ΥΒΠ.



4.2 Υλοποίηση Συστήματος

4.2.1 Γλώσσα Προγραμματισμού και περιβάλλον ανάπτυξης

Η γλώσσα προγραμματισμού που επιλέχθηκε για την υλοποίηση του συστήματος είναι η Java. Η γλώσσα είναι αντικειμενοστραφής, χαρακτηριστικό πολύ σημαντικό για μεγάλες εφαρμογές όπου πρέπει να διαχειριστούν μεγάλοι όγκοι δεδομένων και να γραφεί πολὺς κώδικας. Προσφέρει μεγάλο αριθμό ενσωματωμένων βιβλιοθηκών που παρέχουν στον προγραμματιστή έτοιμες συναρτήσεις και δομές δεδομένων. Στην υλοποίηση του συστήματος χρησιμοποιήθηκαν δομές δεδομένων, όπως πίνακες κατακερματισμού (Hashtables) και διανύσματα (Vectors), που θα ήταν κοπιαστικό να υλοποιηθούν από την αρχή. Φυσικά, χρησιμοποιήθηκαν και καινούριες δομές, όπως ουρές προτεραιότητας για άμεση ταξινόμηση αντικειμένων. Η ύπαρξη, όμως, των ενσωματωμένων βιβλιοθηκών επιτάχυνε την διαδικασία ανάπτυξης του συστήματος.

Ιδιαίτερα σημαντικό προσόν της γλώσσας είναι η συνδυασμένη λειτουργία της με την JSP (Java Server Pages). Η JSP επιτρέπει την παραγωγή ιστοσελίδων με δυναμικό τρόπο. Οι σελίδες της JSP μεταγλωτίζονται στον εξυπηρετητή (server), το περιεχόμενό τους παράγεται δυναμικά ανάλογα με την κατάσταση του συστήματος ή τις επιλογές του χρήστη και στέλνονται στον πελάτη (client). Το σημαντικότερο, όμως, είναι ότι από τις σελίδες JSP μπορούν να κληθούν συναρτήσεις και αντικείμενα που έχουν γραφεί σε Java. Αυτό σημαίνει ότι ο κώδικας που εκτελεί τις κύριες λειτουργίες μιας ιστοσελίδας (παραγωγή περιεχομένου, επεξεργασία αιτήματος χρήστη κλπ.) μπορεί να γραφεί σε Java και η γραφική διεπαφή και η παρουσίαση των αποτελεσμάτων σε JSP. Διαχωρίζεται, δηλαδή, η εμφάνιση από την λειτουργικότητα. Έτσι ο προγραμματιστής μπορεί να αναπτύσσει και να ελέγχει κάθε κομμάτι ζεχωριστά και να τα συνδυάζει στο τέλος. Αυτή ήταν και η τακτική που χρησιμοποιήθηκε για την ανάπτυξη του συστήματος MoRe. Δημιουργήθηκαν τάξεις της Java που υλοποιούσαν όλες τις λειτουργίες που είναι απαραίτητες για την παραγωγή υποδείξεων. Οι τάξεις αυτές μπορούσαν να λειτουργήσουν αυτόνομα και να υπάρξει ένα πλήρως λειτουργικό σύστημα υπόδειξης χωρίς γραφική διεπαφή. Στη συνέχεια, αφού είχε επιβεβαιωθεί ότι οι αλγόριθμοι είχαν υλοποιηθεί σωστά, δημιουργήθηκε η γραφική διεπαφή σε JSP που καλεί αντικείμενα από τις υλοποιημένες τάξεις της Java και διαχειρίζεται τις αποκρίσεις των χρηστών.

Οι ιστοσελίδες που έχουν αναπτυχθεί σε JSP δεν μπορούν να εκτελεστούν σε οποιοδήποτε εξυπηρετητή, αλλά χρειάζεται ειδικός εξυπηρετητής που να τις υποστηρίζει. Επιλέχθηκε ο Apache Tomcat Server, στην έκδοση 5.5.11 που είναι εξυπηρετητής ανοικτού κώδικα και διανέμεται δωρεάν διαμέσου του διαδικτύου¹.

Για την μεταγλώττιση των τάξεων της Java και των ιστοσελιδών της JSP είναι απαραίτητη η ύπαρξη ενός περιβάλλοντος ανάπτυξης Java (Java Development Environment - JDK), το οποίο περιλαμβάνει τον μεταγλωττιστή και τις απαραίτητες βιβλιοθήκες. Για την ανάπτυξη του συστήματος MoRe χρησιμοποιήθηκε το Java 2 Platform Standard Edition Development Kit 5.0².

Τόσο οι τάξεις της Java, όσο και οι ιστοσελίδες της JSP μπορούν να γραφούν ως απλά αρχεία κειμένου με την κατάλληλη κατάληξη (.java και .jsp αντίστοιχα) χρησιμοποιώντας έναν απλό κειμενογράφο. Υπάρχουν όμως εξειδικευμένα περιβάλλοντα ανάπτυξης που διευκολύνουν ιδιαίτερα τον προγραμματιστή. Για την ανάπτυξη των τάξεων της Java χρησιμοποιήθηκε το περιβάλλον JCreator, η μη-επαγγελματική έκδοση του οποίου διανέμεται δωρεάν³. Οι ιστοσελίδες JSP αναπτύχθηκαν χρησιμοποιώντας το Macromedia Dreamweaver MX 2004⁴.

Ο πηγαίος κώδικας του συστήματος (τάξεις της Java και ιστοσελίδες JSP), ο εξυπηρετητής Apache Tomcat, το περιβάλλον ανάπτυξης της Java και το JCreator περιλαμβάνονται στο CD-ROM που συνοδεύει την εργασία.

4.2.2 Λειτουργία Συστήματος

Στην παράγραφο αυτή θα παρουσιαστεί συνοπτικά η λειτουργία του συστήματος, από την πλευρά του χρήστη και από την πλευρά του διαχειριστή.

¹ <http://jakarta.apache.org/tomcat/>

² <http://java.sun.com/products/>

³ <http://www.jcreator.com/>

⁴ <http://www.macromedia.com/>



Εικόνα 1 : Αρχική οθόνη του Συστήματος

Η πρώτη οθόνη του συστήματος που αντικρίζουν οι χρήστες εμφανίζεται στην Εικόνα 1. Οι χρήστες που έχουν ξαναχρησιμοποιήσει το σύστημα μπορούν να εισέλθουν σε αυτό και να δουν τις προσωπικές τους υποδείξεις πληκτρολογώντας το Όνομα Χρήστη (Username) και το Συνθηματικό (Password) τους. Εάν η πληκτρολόγηση είναι σωστή οδηγούνται στην οθόνη της Εικόνας 4, αλλιώς ένα κατάλληλο μήνυμα τους ζητά να ξαναπροσπαθήσουν.

Συνήθως, τα συστήματα υπόδειξης αποτελούν τμήματα άλλων συστημάτων, για παράδειγμα ηλεκτρονικών καταστημάτων. Στην περίπτωση αυτή δεν θα χρειαζόταν η ύπαρξη μιας οθόνης αυθεντικοποίησης όπως η παραπάνω, αφού θα είχε ελεγχθεί η ταυτότητα του χρήστη κατά την είσοδό του στο υπερσύστημα. Στο MoRe, που αναπτύχθηκε ως ανεξάρτητο σύστημα, η οθόνη αυτή είναι απαραίτητη, περισσότερο για τον προσδιορισμό του χρήστη για τον οποίο θα παραχθούν οι υποδείξεις, παρά για λόγους ασφάλειας του συστήματος.



Εικόνα 2 : Εγγραφή νέου χρήστη

Οι νέοι χρήστες του συστήματος κάνοντας κλικ στο “register here” οδηγούνται στην οθόνη της Εικόνας 2, όπου μπορούν να επιλέξουν το Όνομα Χρήστη και το Συνθηματικό που θα χρησιμοποιούν στο σύστημα. Σε περίπτωση που επιλέξουν Όνομα Χρήστη που ήδη υπάρχει ή πληκτρολογήσουν δύο διαφορετικά Συνθηματικά εμφανίζεται κατάλληλο μήνυμα που τους ενημερώνει για το λάθος τους. Αφού ολοκληρωθεί η παραπάνω επιλογή, οι χρήστες πρέπει να βαθμολογήσουν τουλάχιστον N ταινίες, προκειμένου το σύστημα να αρχίσει να παράγει προσωποποιημένες υποδείξεις. Η μεταβλητή N είναι μια παράμετρος που μπορεί να αλλάξει κατά την λειτουργία του συστήματος και στην οποία έχουμε αναφερθεί αναλυτικά στην παράγραφο 3.5. Υπενθυμίζουμε ότι λαμβάνει τιμές μεγαλύτερες του 1 και ότι όσο μεγαλύτερη είναι η τιμή του N, τόσο περισσότερη πληροφορία διαθέτει το σύστημα για τις προτιμήσεις του χρήστη και τόσο ακριβέστερες είναι οι πρώτες υποδείξεις που παράγει. Παράλληλα, όμως, οι μεγάλες τιμές του N καθιστούν την διαδικασία βαθμολόγησης ιδιαίτερα κουραστική για τους χρήστες. Η

προκαθορισμένη τιμή του N για το MoRe είναι η τιμή 20, για την οποία διαπιστώσαμε πειραματικά ότι εξασφαλίζει υποδείξεις υψηλής ακρίβειας.



Εικόνα 3 : Αρχική βαθμολόγηση ταινιών

Η βαθμολόγηση των είκοσι ταινιών γίνεται στην οθόνη της Εικόνας 3. Εμφανίζονται πέντε ταινίες ανά οθόνη και οι χρήστες μπορούν να τις βαθμολογήσουν άμεσα κάνοντας κλικ στις επιλογές (radio buttons) που βρίσκονται κάτω από την περιγραφή κάθε ταινίας. Το σύστημα συνεχίζει να παρουσιάζει ταινίες μέχρι οι χρήστες να βαθμολογήσουν τον απαιτούμενο αριθμό. Πρέπει εδώ να σημειωθεί ότι αν κάποιος χρήστης δεν ολοκληρώσει την διαδικασία βαθμολόγησης των ταινιών, τότε δεν καταχωρείται στο σύστημα. Ούτε το Όνομα Χρήστη του, ούτε οι ταινίες που έχει βαθμολογήσει μέχρι εκείνη τη στιγμή αποθηκεύονται και αν επανέλθει στο σύστημα πρέπει να επαναλάβει την διαδικασία από την αρχή. Αντίθετα, αν η διαδικασία βαθμολόγησης ολοκληρωθεί κανονικά, οι βαθμολογίες του και το Όνομα χρήστη αποθηκεύονται και το σύστημα είναι έτοιμο να αρχίσει να υποδεικνύει ταινίες για αυτόν. Ο λόγος που μας οδήγησε σε αυτή την επιλογή είναι ότι αν στο σύστημα

Κεφάλαιο 4 : Σχεδίαση και υλοποίηση συστήματος

παρέμεναν χρήστες με πολύ λίγες βαθμολογήσεις, τότε οι ομοιότητες τους με τους άλλους χρήστες θα βασίζονταν σε μικρό δείγμα και η συνολική ακρίβεια του συστήματος θα μειωνόταν.



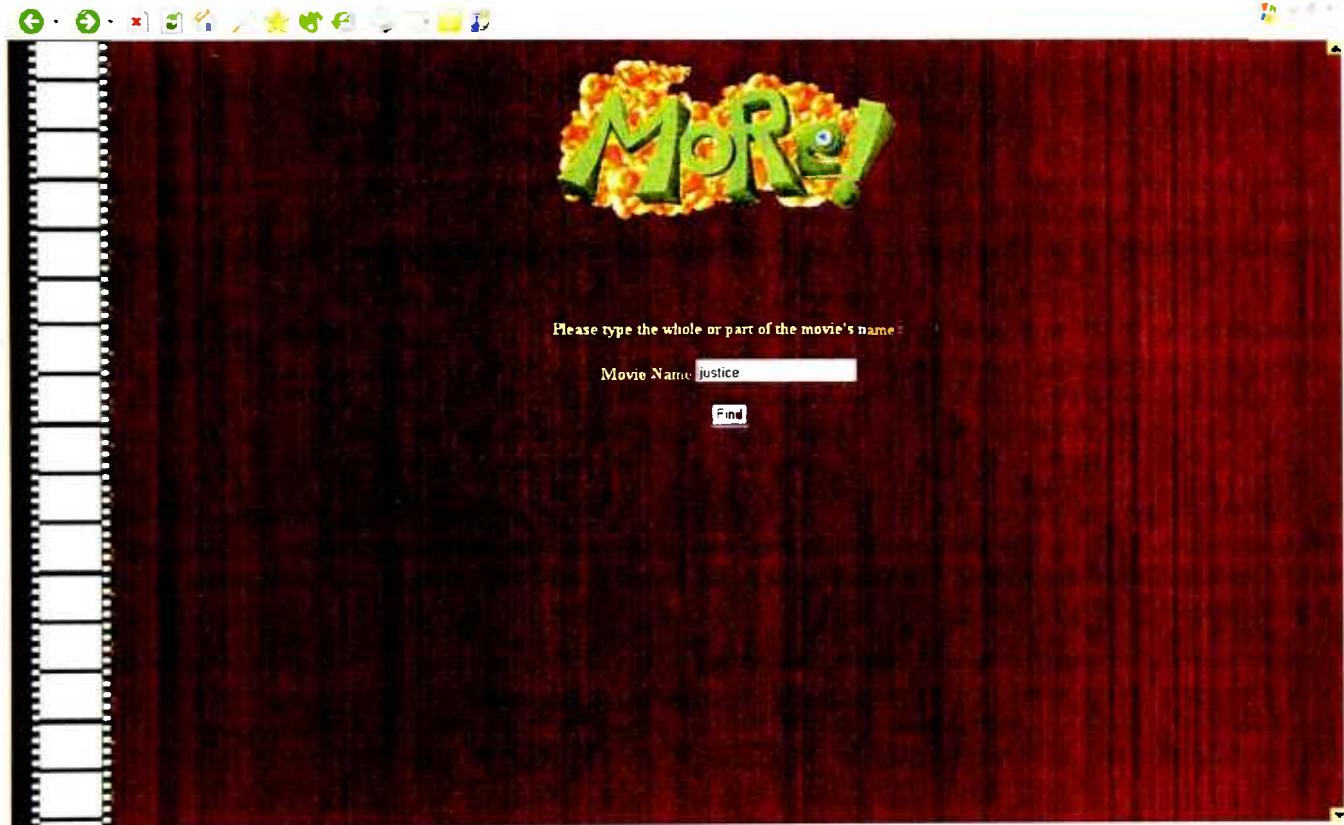
Εικόνα 4 : Προβολή Υποδείξεων

Η οθόνη στην οποία παρουσιάζονται οι υποδείξεις εμφανίζεται στην Εικόνα 4. Όπως μπορεί να παρατηρήσει κάποιος, μοιάζει με την οθόνη της Εικόνας 3, με την διαφορά ότι δίπλα σε κάθε ταινία εμφανίζεται η πρόβλεψη του συστήματος. Οι βαθμολογήσεις των ταινιών γίνονται με τον ίδιο ακριβώς τρόπο, όπως και η πλοιήγηση μεταξύ των οθονών. Καθώς ο χρήστης βαθμολογεί όσες ταινίες έχει δει, έχει τη δυνατότητα να ανανεώσει τις υποδείξεις που λαμβάνει, δίνοντας εντολή στο σύστημα να λάβει υπόψη του και τις τελευταίες του βαθμολογήσεις.

Μια επιπλέον επιλογή για τους χρήστες του συστήματος είναι η αναζήτηση ταινιών με βάση το όνομά τους (στην αγγλική γλώσσα) (Εικόνα 5). Η αναζήτηση ταινιών επιτρέπει στον χρήστη να βαθμολογήσει ταινίες (Εικόνα 6) που δεν που έχει

Κεφάλαιο 4 : Σχεδίαση και υλοποίηση συστήματος

υποδείξει ακόμα το σύστημα, δηλώνοντας έτσι τις προτιμήσεις του και επιταχύνοντας την διαδικασία βελτιστοποίησης των υποδείξεων που λαμβάνει.



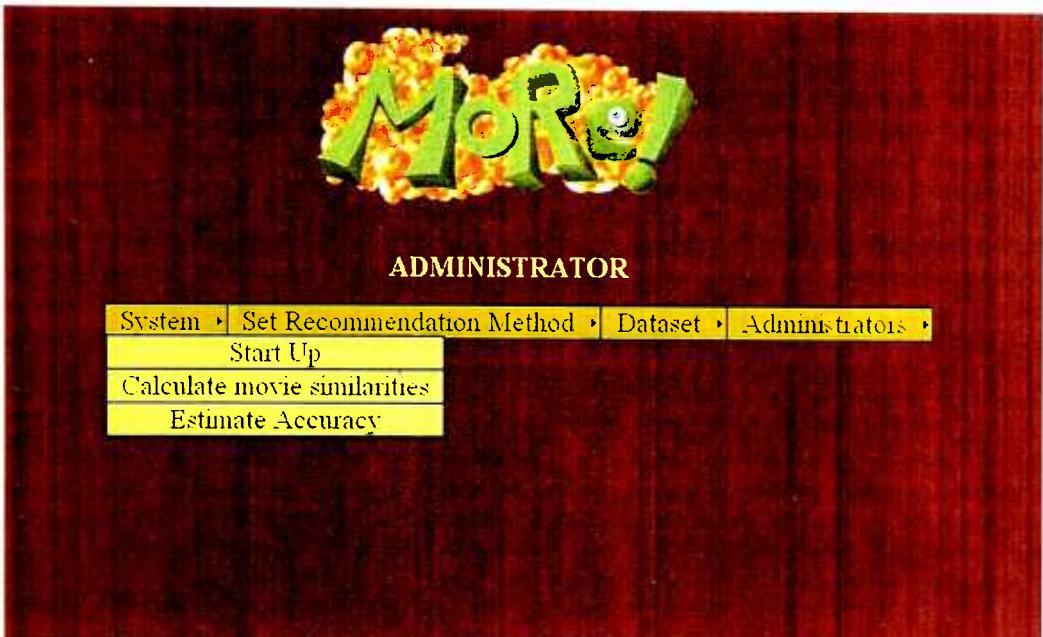
Εικόνα 5 : Αναζήτηση ταινιών



Εικόνα 6 : Αποτέλεσμα αναζήτησης

Οι επιλογές για τον διαχειριστή του συστήματος είναι συγκεντρωμένες σε ξεχωριστή οθόνη, η είσοδος στην οποία γίνεται με χρήστη Ονόματος Διαχειριστή και Συνθηματικού. Είναι ομαδοποιημένες σε κατηγορίες με τη μορφή ενός μενού και αποτελούν υλοποίηση των διεργασιών που σχετίζονται με την Περύπτωση Χρήσης «Επιλογές Διαχείριση» που αναφέρθηκαν κατά την περιγραφή της σχεδίασης του συστήματος στην παράγραφο 4.1.

Η κατηγορία με το γενικό όνομα “System” (Εικόνα 7) περιέχει τις επιλογές για την εκκίνηση του συστήματος, τον υπολογισμό των ομοιοτήτων των ταινιών και την εκτίμηση της ακρίβειας του συστήματος. Η εκκίνηση του συστήματος (“Start Up”), που όπως έχουμε αναφέρει σε προηγούμενη παράγραφο φορτώνει τα περιεχόμενα των αρχείων στον πίνακα βαθμολογιών, εκτελείται μόνο μια φορά κατά την εκκίνηση του εξυπηρετητή που φιλοξενεί το σύστημα.



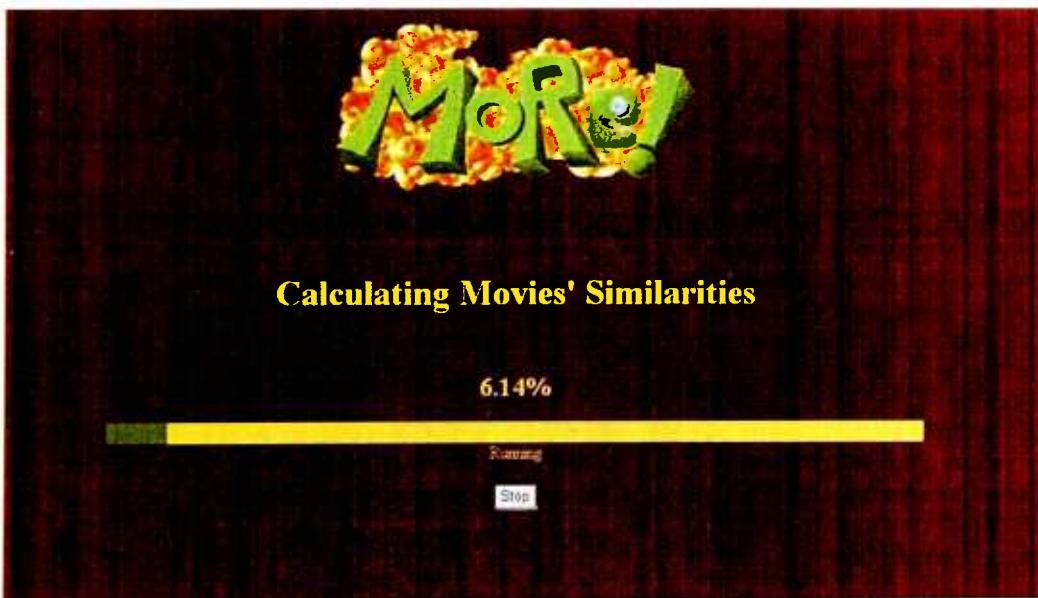
Εικόνα 7 : Η οθόνη επιλογών διαχειριστή με επιλεγμένο το υπομενού 'System'

Ο υπολογισμός των ομοιοτήτων των ταινιών ("Calculate movie similarities"), που είναι απαραίτητος για την λειτουργία της παραγωγής Υποδείξεων Βασισμένων στο Περιεχόμενο, πρέπει να εκτελείται μόνο όταν προστίθονται ή αφαιρούνται ταινίες από το σύστημα.



Εικόνα 8 : Ορισμός παραμέτρων για τον υπολογισμό των ομοιοτήτων των ταινιών

Ο διαχειριστής του συστήματος μπορεί ορίσει τις παραμέτρους του υπολογισμού (Εικόνα 8), δηλαδή την ελάχιστη συχνότητας εμφάνισης μιας λέξης στην περιγραφή όλων των ταινιών για να συμπεριληφθεί ως χαρακτηριστικό (feature) στα διανύσματα των ταινιών (παράγραφος 3.4.2) και τον αριθμό των πιο όμοιων ταινιών που θα αποθηκευτούν για κάθε ταινία. Η ακρίβεια του συστήματος αυξάνει όσο μειώνεται η τιμή της πρώτης παραμέτρου, παράλληλα όμως αυξάνεται και ο χρόνος παραγωγής των ΥΒΠ. Αντίθετα, μεγαλύτερες τιμές της δεύτερης παραμέτρου οδηγούν σε παραγωγή υποδείξεων μεγαλύτερης ακρίβειας, αλλά σε περισσότερο χρόνο. Επειδή ο υπολογισμός των ομοιοτήτων των ταινιών είναι μια χρονοβόρα διαδικασία, συνοδεύεται από μια οθόνη (Εικόνα 9) που ενημερώνει το διαχειριστή για την κατάσταση του υπολογισμού με την βοήθεια μιας μπάρας προόδου (progress bar). Ο χρήστης έχει τη δυνατότητα να διακόψει την διαδικασία οποιαδήποτε στιγμή και να την ξεκινήσει από την αρχή. Η προτεραιότητα του υπολογισμού σε υπολογιστικούς πόρους είναι η χαμηλότερη δυνατή για να μην καθυστερεί τις υπόλοιπες λειτουργίες του συστήματος.

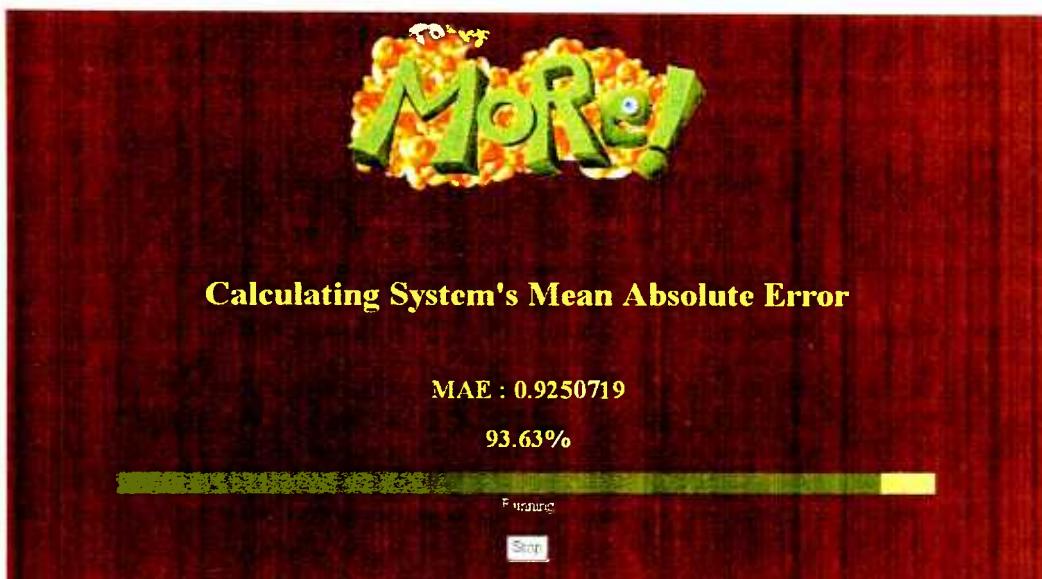


Εικόνα 9 : Οθόνη προόδου διαδικασίας υπολογισμού ομοιοτήτων ταινιών

Η επιλογή της εκτίμησης της ακρίβειας του συστήματος ("Estimate Accuracy") ξεκινά την διαδικασία υπολογισμού του Μέσου Απόλυτου Λάθους για τη μέθοδο υπόδειξης που χρησιμοποιεί εκείνη τη στιγμή το σύστημα. Επειδή και αυτός ο

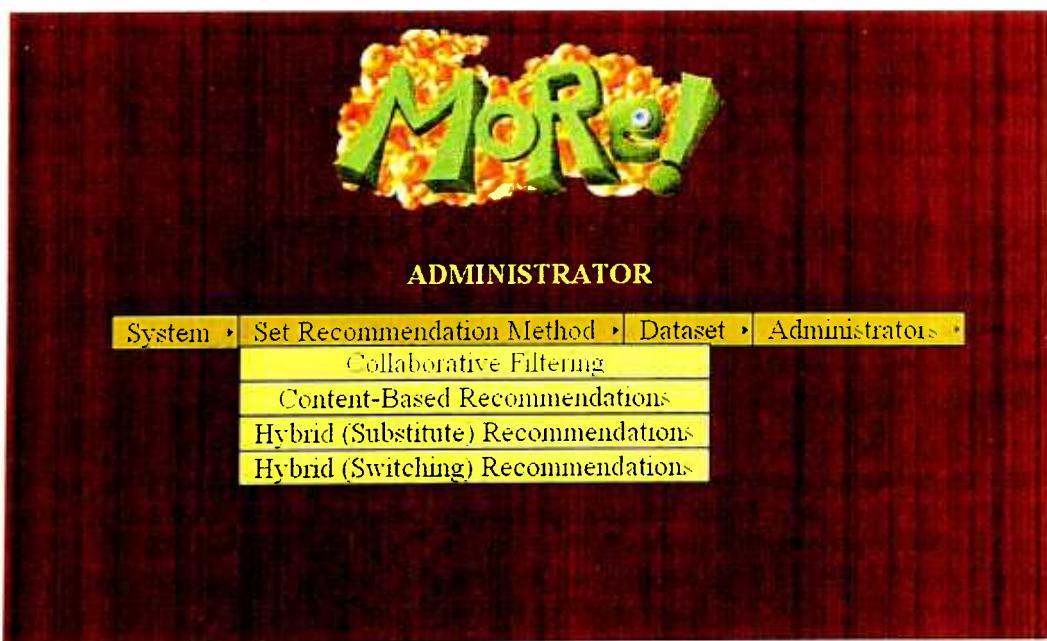
Κεφάλαιο 4 : Σχεδίαση και υλοποίηση συστήματος

υπολογισμός είναι πολύ χρονοβόρος, παρουσιάζεται σε ξεχωριστή οθόνη (Εικόνα 10) με μπάρα προόδου και δυνατότητα διακοπής και επανεκκίνησης.



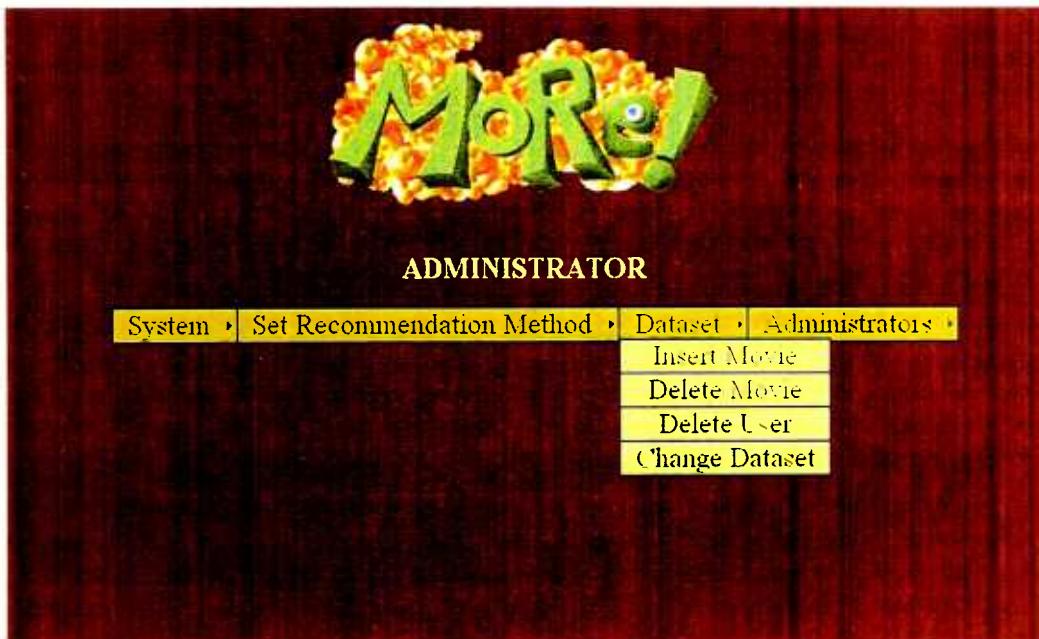
Εικόνα 10 : Υπολογισμός της ακρίβειας του συστήματος

Η κατηγορία επιλογών ορισμού μεθόδων υπόδειξης (“Set Recommendation Method”) φαίνεται στην εικόνα 11. Επιλέγοντας την επιθυμητή μέθοδο εμφανίζεται μήνυμα που πληροφορεί τον διαχειριστή για την πραγματοποίηση της αλλαγής.

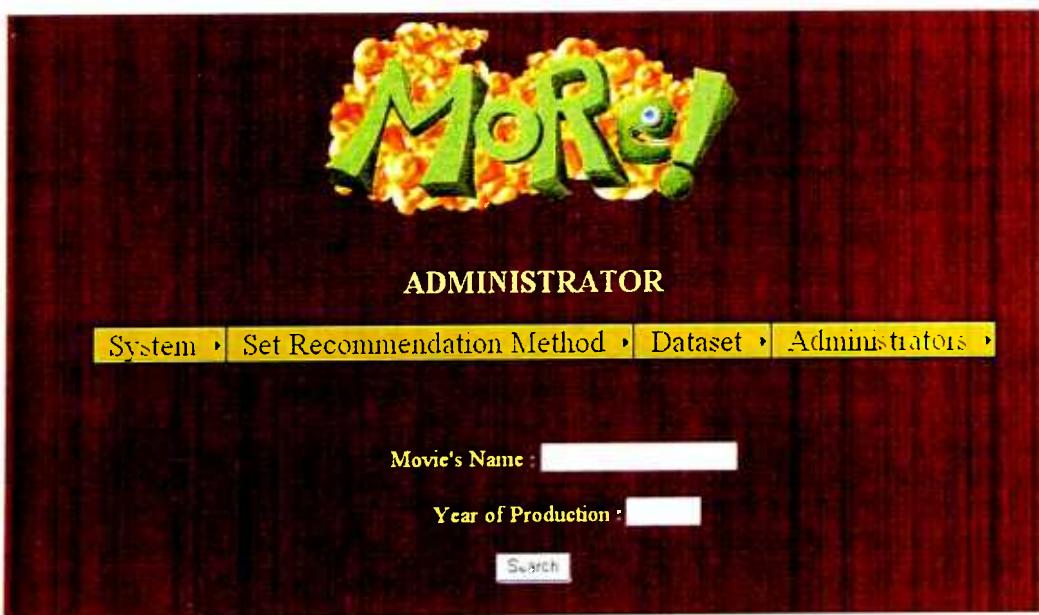


Εικόνα 11 : Επιλογές αλλαγής μεθόδου υπόδειξης

Για την κατηγορία επιλογών που σχετίζονται με το σύνολο δεδομένων (“Dataset”), που παρουσιάζονται στην Εικόνα 12, οι διαθέσιμες δυνατότητες είναι η εισαγωγή νέας ταινίας, η διαγραφή ταινίας, η διαγραφή χρήστη και η συνολική αλλαγή συνόλου δεδομένων. Για την εισαγωγή νέας ταινίας (Εικόνα 13), ο διαχειριστής εισάγει το όνομα της ταινίας και το έτος παραγωγής της και το σύστημα αναζητά τις απαραίτητες πληροφορίες στο δικτυακό τόπο της IMDb.



Εικόνα 12 : Επιλογές σχετικές με το σύνολο δεδομένων



Εικόνα 13: Εισαγωγή νέας ταινίας

Κεφάλαιο 4 : Σχεδίαση και υλοποίηση συστήματος

Στην διαγραφή ταινίας και χρήστη, που φαίνονται στις εικόνες 14 και 15 αντίστοιχα, ο διαχειριστής του συστήματος εισάγει το όνομα της ταινίας ή του χρήστη. Αν υπάρχει, το σύστημα ζητά επιβεβαίωση και ολοκληρώνει τη διαγραφή εμφανίζοντας μήνυμα επιτυχίας, αλλιώς εμφανίζει μήνυμα που πληροφορεί το διαχειριστή ότι η ταινία ή ο χρήστης δεν υπάρχουν στο σύστημα.



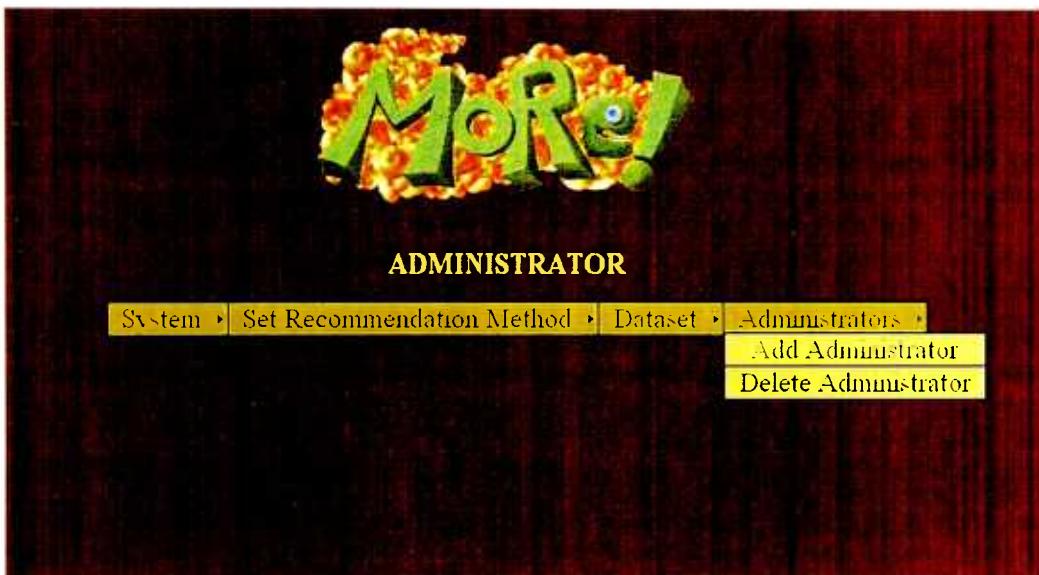
Εικόνα 114 : Διαγραφή ταινίας



Εικόνα 15 : Διαγραφή Χρήστη

Κεφάλαιο 4 : Σχεδίαση και υλοποίηση συστήματος

Η κατηγορία επιλογών Διαχειριστές (“Administrators”), που φαίνεται στην Εικόνα 16, επιτρέπει την προσθήκη νέων διαχειριστών και την διαγραφή υπαρχόντων. Οι λειτουργίες αυτές γίνονται με τον ίδιο ακριβώς τρόπο όπως οι αντίστοιχες λειτουργίες για τους χρήστες.



Εικόνα 16 : Προσθήκη και διαγραφή διαχειριστών

5. Πειραματική Αξιολόγηση Συστήματος

Στο κεφάλαιο αυτό θα παραθέσουμε τα αποτελέσματα από τα πειράματα που πραγματοποιήσαμε για την αξιολόγηση του συστήματος MoRe. Η αξιολόγηση είχε ως στόχο την μέτρηση της ακρίβειας των παραγόμενων υποδείξεων και την σύγκριση των διαφορετικών μεθόδων υπόδειξης που έχουν υλοποιηθεί. Μια σημαντική παράμετρος του συστήματος που είναι η φιλικότητα της γραφικής διεπαφής και η ευχρηστία της λειτουργίας του δεν έχει αξιολογηθεί, καθώς κάτι τέτοιο προϋποθέτει πειράματα με τη συμμετοχή ενός αριθμού πραγματικών χρηστών, ενώ ξεφεύγει και από τους στόχους της παρούσας εργασίας.

Η αξιολόγηση του συστήματος έγινε σε ηλεκτρονικό υπολογιστή με επεξεργαστή ταχύτητας 3.4 GHz, κύρια μνήμη 1 GB RAM και λειτουργικό σύστημα Windows 2000 Professional.

5.1 Πειραματικά δεδομένα

Τα πειραματικά δεδομένα που χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος είναι το σύνολο δεδομένων (dataset) του MovieLens που περιγράφαμε αναλυτικά στην παράγραφο 3.1.

Θεωρούμε το συγκεκριμένο σύνολο δεδομένων ιδανικό για την αξιολόγηση του συστήματός για δύο λόγους. Ο πρώτος είναι ότι έχει χρησιμοποιηθεί για σκοπούς αξιολόγησης από αρκετούς ερευνητές και επομένως μπορεί να γίνει άμεση σύγκριση αποτελεσμάτων. Ο δεύτερος λόγος είναι ότι το μέγεθος του συνόλου δεδομένων είναι αρκετά μεγάλο ώστε να εξασφαλίζει ότι τα αποτελέσματα των πειραμάτων δεν εξαρτώνται από την τυχαία επιλογή κάποιου μικρού δείγματος, ενώ και ο τρόπος συλλογής των δεδομένων (οι βαθμολογίες των χρηστών του συστήματος MovieLens για όλο το έτος 2000) υπόσχεται ότι δεν υπάρχει μεροληψία (bias) στα δεδομένα.

Πρέπει να σημειώσουμε ότι σε όλα τα πειράματα χρησιμοποιήθηκε το σύνολο των διαθεσίμων δεδομένων, δηλαδή οι περίπου 1.000.000 βαθμολογίες των 6040 χρηστών για τις 3952 ταινίες.

5.2 Μέτρα Αξιολόγησης

Τα μέτρα αξιολόγησης που χρησιμοποιούνται είναι το Μέσο Απόλυτο Λάθος (Mean Absolute Error –MAE) και η κάλυψη (coverage).

Η κάλυψη (coverage) εκφράζει το ποσοστό των αντικειμένων για τα οποία το σύστημα μπορεί να κάνει υποδείξεις και υπολογίζεται ως το πηλίκο των αντικειμένων για τα οποία το σύστημα ήταν σε θέση να παράγει πρόβλεψη προς το αριθμό των αντικειμένων για τα οποία ζητήθηκε πρόβλεψη.

Το Μέσο Απόλυτο Λάθος είναι ένα κατάλληλο μέτρο ακρίβειας για συστήματα που χρησιμοποιούν αριθμητικές βαθμολογήσεις χρηστών και προβλέψεις. Ο τρόπος υπολογισμού του Μέσου Απόλυτου Λάθους περιγράφεται από τις Shardanand & Maes (1995) : Τα πειραματικά δεδομένα (στην περίπτωσή μας το σύνολο δεδομένων του MovieLens) χωρίζονται σε δύο υποσύνολα. Το πρώτο περιλαμβάνει το 80% των βαθμολογιών κάθε χρήστη που διαθέτουμε και αποκαλείται σύνολο εκπαίδευσης (training set). Στο δεύτερο σύνολο ανήκει το υπόλοιπο 20% των βαθμολογιών του κάθε χρήστη και ονομάζεται σύνολο ελέγχου (test set). Ο διαχωρισμός των βαθμολογιών στα δύο υποσύνολα γίνεται τυχαία. Λόγω του μεγάλου μεγέθους των πειραματικών δεδομένων μας, δεν κρίνεται απαραίτητη η χρήση της τεχνικής του k-fold cross validation. Αφού διαχωριστούν οι βαθμολογίες του συνόλου δεδομένων στα δύο υποσύνολα, οι βαθμολογίες που ανήκουν στο σύνολο ελέγχου αφαιρούνται από τον πίνακα βαθμολογιών του συστήματος. Στην συνέχεια, χρησιμοποιώντας τον πίνακα βαθμολογιών που προέκυψε, γίνεται προσπάθεια να γίνουν προβλέψεις για τις βαθμολογίες που αφαιρέθηκαν.

Εάν $\{r_1, \dots, r_n\}$ είναι οι πραγματικές βαθμολογίες του χρήστη στο σύνολο ελέγχου, $\{p_1, \dots, p_n\}$ είναι οι προβλέψεις του συστήματος και $E = \{\varepsilon_1, \dots, \varepsilon_n\} = \{p_1 - r_1, \dots, p_n - r_n\}$ είναι τα λάθη, τότε το Μέσο Απόλυτο Λάθος είναι:

$$MAE = \left| \bar{E} \right| = \frac{\sum_{i=1}^n |\varepsilon_i|}{n} \quad (5.1)$$

Η παραπάνω διαδικασία πραγματοποιείται με τον ίδιο τρόπο, όποια μέθοδος υπόδειξης και αν χρησιμοποιείται. Μικρές τιμές MAE υποδηλώνουν μεγάλη ακρίβεια υποδείξεων για το σύστημα.

Παρόλο που έχουν προταθεί και άλλα μέτρα αξιολόγησης της ακρίβειας των υποδείξεων, όπως για παράδειγμα το Κανονικοποιημένο Μέσο Απόλυτο Λάθος

(Normalized Mean Absolute Error – NMAE) που έχει προταθεί από τους Goldberg et al. (2001) και πιθανότατα είναι το ίδιο ή περισσότερο κατάλληλα από το MAE, επιλέγουμε να βασιστούμε στο τελευταίο γιατί έχει χρησιμοποιηθεί από την πλειοψηφία των ερευνητών από την αρχή της ανάπτυξης του πεδίου των συστημάτων υπόδειξης και έτσι προσφέρεται για συγκρίσεις.

Για να συγκρίνουμε τις τιμές του MAE που προκύπτουν από τις διαφορετικές μεθόδους υπόδειξης και να διαπιστώνουμε αν η διαφορές είναι στατιστικά σημαντικές πρέπει να εφαρμόσουμε ένα τεστ για ζεύγη παρατηρήσεων (paired test). Τα ζεύγη παρατηρήσεων που χρησιμοποιούμε είναι οι τιμές του MAE για κάθε χρήστη του συστήματος που έχουν υπολογιστεί με δύο διαφορετικές κάθε φορά μεθόδους υπόδειξης. Επειδή η κατανομή των παρατηρήσεων διαφέρουν πολύ από την κανονική κατανομή, χρησιμοποιούμε το μη παραμετρικό τεστ σειράς Wilcoxon (Wilcoxon non-parametric rank test). Το τεστ αυτό ελέγχει την υπόθεση αν οι τιμές MAE για τα δύο δείγματα είναι οι ίδιες και οι όποιες διαφορές οφείλονται σε τυχαίους παράγοντες, δηλαδή οι δύο μέθοδοι υπόδειξης παράγουν προβλέψεις ίδιας ακρίβειας. Εάν η τιμή (Asymp. Sig. (2-tailed)) του τεστ Wilcoxon είναι μικρότερη από 0.05 (που είναι το διάστημα εμπιστοσύνης) τότε η παραπάνω υπόθεση απορρίπτεται και επομένως οι διαφορές στην τιμή του MAE είναι στατιστικά σημαντικές και επομένως μια από τις δύο μεθόδους παράγει υποδείξεις μεγαλύτερης ακρίβειας. Η εκτέλεση του τεστ γίνεται με την βοήθεια του στατιστικού πακέτου SPSS, τα αποτελέσματα του οποίου θα παραθέτουμε όταν αναφερόμαστε σε σύγκριση μεθόδων.

Εκτός από τα παραπάνω απόλυτα μέτρα, θα γίνει σύγκριση του χρόνου εκτέλεσης των μεθόδων. Αν και οι επιδόσεις των αλγορίθμων συγκρίνονται με βάση την υπολογιστική τους πολυπλοκότητα, η οποία αναφέρεται στις παραγράφους της παρούσας εργασίας που γίνεται αναφορά στους αλγορίθμους, εντούτοις πιστεύουμε ότι μια σύγκριση που βασίζεται στον χρόνο εκτέλεσης στο ίδιο υπολογιστικό περιβάλλον κάνει πιο εμφανείς και κατανοητές τις διαφορές στην ταχύτητα εκτέλεσης των αλγορίθμων.

5.3 Αξιολόγηση Αλγορίθμων

5.3.1 Συνεργατική Μέθοδος Υπόδειξης

Για τον αλγόριθμο παραγωγής υποδείξεων με τη ΣΜΥ που χρησιμοποιείται από το σύστημα υπολογίσαμε τιμή MAE ίση με 0.7597331 (υπενθυμίζουμε ότι το MAE δεν έχει μονάδα μέτρησης) και επιτεύχθηκε κάλυψη 98,34%.

Εκτελώντας την ίδια μέτρηση, αλλά επιτρέποντας παραγωγή προβλέψεων χωρίς τον περιορισμό του σχηματισμού γειτονιάς τουλάχιστον 5 γειτόνων (παράγραφος 3.4.1), είχαμε MAE 0.7654 και κάλυψη 99.2%. Εκτελώντας το τεστ Wilcoxon διαπιστώσαμε ότι η διαφορά ακρίβειας σε σχέση με την προηγούμενη μέτρηση είναι στατιστικά σημαντική (Εικόνα 1). Επιβεβαιώνεται, δηλαδή, η υπόθεσή μας ότι ο περιορισμός στο σχηματισμό γειτονιάς βελτιώνει την ακρίβεια των υποδείξεων αλλά θυσιάζει ένα κομμάτι κάλυψης.

Ranks

		N	Mean Rank	Sum of Ranks
CBF - CBF_noNeighborhoodRestriction	Negative Ranks	119 ^a	106.51	12675.00
	Positive Ranks	78 ^b	87.54	6828.00
	Ties	5837 ^c		
	Total	6034		

a. CBF < CBF_noNeighborhoodRestriction

b. CBF > CBF_noNeighborhoodRestriction

c. CBF = CBF_noNeighborhoodRestriction

Test Statistics^b

	CBF - CBF_noNeighborhoodRestriction
Z	-3.649 ^a
Asymp. Sig. (2-tailed)	.000263516327666

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Πίνακας 1 : Αποτελέσματα του Wilcoxon test για την ΣΜΥ και τη ΣΜΥ χωρίς περιορισμό σχηματισμού γειτονιάς

Οι (Herlocker et al, 1999), που χρησιμοποιούν ένα υποσύνολο του συνόλου δεδομένων μας, εφαρμόζουν, όπως και εμείς, Βάρος Σημαντικότητας (παράγραφος 2.3.6), αλλά ορίζουν ως σύνολο ελέγχου το 10% των διαθέσιμων δεδομένων, υπολόγισαν το MAE για το Συνεργατική Μέθοδο Υπόδειξης 0.7660 με κάλυψη 99.8%. Μάλιστα, στα πειράματά τους, η καλύτερη τιμή MAE που επιτυγχάνουν είναι

0.7578, με κάλυψη, όμως, μόλις 60.9%. Η τιμή αυτή επιτυγχάνεται όταν για τη συμμετοχή στη γειτονιά οι γείτονες πρέπει να έχουν τιμή ομοιότητας Pearson με τον ενεργό χρήστη τουλάχιστον 0.3.

Παρατηρούμε, λοιπόν, ότι ο περιορισμός στην παραγωγή υποδείξεων που εφαρμόσαμε προκαλεί βελτίωση της ακρίβειας αντίστοιχη με αυτή που έχουν επιτευχθεί από άλλους ερευνητές, με σαφώς μικρότερη απώλεια κάλυψης.

Εδώ πρέπει να σημειωθεί ότι ο χρόνος παραγωγής υποδείξεων χρησιμοποιώντας την ΣΜΥ στον υπολογιστή εκτέλεσης των πειραμάτων ήταν, κατά μέσο όρο, 14 δευτερόλεπτα.

5.3.2 Υποδείξεις Βασισμένες στο Περιεχόμενο

Ο τρόπος υπολογισμού των ΥΒΠ (παράγραφος 3.4.2) εξασφαλίζει ότι κάλυψη του συστήματος θα είναι πάντα 100%. Το MAE που υπολογίσαμε για την κύρια προσέγγιση ΥΒΠ του συστήματος είναι 0.9253873. Η διαφορά της τιμής αυτής με την τιμή MAE της ΣΜΥ είναι στατιστικά σημαντική (πίνακας 3). Πρέπει να σημειώσουμε ότι ο χρόνος εκτέλεσης του αλγόριθμου παραγωγής ΥΒΠ στον υπολογιστή εκτέλεσης των πειραμάτων ήταν, κατά μέσο όρο, ήταν περίπου 3 δευτερόλεπτα. Δηλαδή, περίπου πέντε φορές μικρότερος από τον αντίστοιχο χρόνο εκτέλεσης της ΣΜΥ.

Ranks

	N	Mean Rank	Sum of Ranks
CB = CBF	1352 ^a	2284.44	3088557.00
Negative Ranks	4682 ^b	3229.18	15119038.00
Positive Ranks	0 ^c		
Ties			
Total	6034		

a. CB < CBF

b. CB > CBF

c. CB = CBF

Test Statistics^b

	CB - CBF
Z	-44.451 ^a
Asymp. Sig. (2-tailed)	0.0E+00

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Πίνακας 2 : Αποτελέσματα τεστ Wilcoxon για ΣΜΥ και ΥΒΠ

Επιπλέον, δοκιμάσαμε να περιγράψουμε τις ταινίες του συστήματος με μικρότερο αριθμό χαρακτηριστικών, θέλοντας να ελέγξουμε αν το πλήθος των χαρακτηριστικών παίζει κάποιο ρόλο στην ακρίβεια του αλγορίθμου. Υπενθυμίζουμε ότι στην προσέγγιση YBΠ του συστήματος, χρησιμοποιούσαμε μια λέξη ως χαρακτηριστικό αν εμφανιζόταν στην περιγραφή τουλάχιστον δύο ταινιών. Ελέγξαμε λοιπόν τη μεταβολή της ακρίβειας όταν αυτό το όριο αυξανόταν στις τρεις, πέντε, δέκα και δεκαπέντε ταινίες, κάτι που είχε επίπτωση στον αριθμό των γνωρισμάτων. Κάθε τιμή ΜΑΕ που προέκυπτε (πίνακας 2) τη συγκρίναμε χρησιμοποιώντας το τεστ Wilcoxon (πίνακας 4) με την τιμή της κυρίας προσέγγισης για να διαπιστώσουμε αν η διαφορές στην ακρίβεια είναι στατιστικά σημαντικές.

	Όριο στις 2 ταινίες	Όριο στις 3 ταινίες	Όριο στις 5 ταινίες	Όριο στις 10 ταινίες	Όριο στις 15 ταινίες
ΜΑΕ	0.9253873	0.92538875	0.9275309	0.9555197	0.97807115
Αριθμός γνωρισμάτων	10626	10620	7865	5430	3514

Πίνακας 3

Ranks

		N	Mean Rank	Sum of Ranks
CB_min3 - CB_min2	Negative Ranks	1 ^a	1.00	1.00
	Positive Ranks	0 ^b	.00	.00
	Ties	6030 ^c		
	Total	6031		
CB_min5 - CB_min2	Negative Ranks	2699 ^d	2879.69	7772290.00
	Positive Ranks	3330 ^e	3124.67	10405145.00
	Ties	2 ^f		
	Total	6031		
CB_min10 - CB_min2	Negative Ranks	2591 ^g	2933.85	7601595.00
	Positive Ranks	3438 ^h	3078.16	10575840.00
	Ties	2 ⁱ		
	Total	6031		
CB_min15 - CB_min2	Negative Ranks	2335 ^j	2865.33	6690553.00
	Positive Ranks	3693 ^k	3108.81	11480853.00
	Ties	3 ^l		
	Total	6031		

a. CB_min5 = CB_min2 g. CB_min10 < CB_min2

b. CB_min3 > CB_min2 h. CB_min10 > CB_min2

c. CB_min3 = CB_min2 i. CB_min10 = CB_min2

d. CB_min5 < CB_min2 j. CB_min15 < CB_min2

e. CB_min5 > CB_min2 k. CB_min15 > CB_min2

f. CB_min5 = CB_min2 l. CB_min15 = CB_min2

Test Statistics^c

	CB_min3 - CB_min2	CB_min5 - CB_min2	CB_min10 - CB_min2	CB_min15 - CB_min2
Z	-1.000 ^a	-9.740 ^b	-11.003 ^b	-17.726 ^b
Asymp. Sig. (2-tailed)	.317	2.03E-22	3.69E-28	2.64E-70

a. Based on positive ranks.

b. Based on negative ranks.

c. Wilcoxon Signed Ranks Test

Πίνακας 4 : Τεστ Wilcoxon για τις παραλλαγές των ΥΒΠ

Η διαφορά ακρίβειας μεταξύ του ορίου των δύο και τριών ταινιών δεν είναι στατιστικά σημαντική. Καθώς όμως μεγαλώνει το όριο (κάτι που σημαίνει ότι οι ταινίες περιγράφονται με μικρότερο αριθμό γνωρισμάτων), η ακρίβεια του συστήματος μειώνεται και οι διαφορές είναι σημαντικές.

Συμπεραίνουμε, δηλαδή, ότι ο αριθμός των γνωρισμάτων με τον οποίο περιγράφονται οι ταινίες αποτελεί παράγοντα που επηρεάζει την ακρίβεια των υποδείξεων που παράγει το σύστημα και, ειδικότερα, η μείωση των γνωρισμάτων προκαλεί μείωση της ακρίβειας.

Στο κεφάλαιο 3 είχαμε αναφερθεί σε μια δεύτερη προσέγγιση που χρησιμοποιούσε τον κατηγοριοποιητή Naïve Bayes για να προβλέψει τις βαθμολογίες που θα έδιναν οι χρήστες στις ταινίες. Το MAE αυτής της προσέγγισης ήταν 1.2434, που κρίνεται απογοητευτικό. Θεωρούμε ότι ο κύριο λόγος της αποτυχίας του Naïve Bayes ήταν η απουσία κάποιων τιμών της βαθμολογικής κλίμακας από τις βαθμολογίες ορισμένων χρηστών. Δεν ήταν σπάνιο φαινόμενο να υπάρχουν χρήστες που όλες οι βαθμολογίες τους είχαν τιμή μεγαλύτερη του 3. Δηλαδή οι χρήστες αυτοί βαθμολογούσαν μόνο τις ταινίες που τους άρεσαν και αμελούσαν να βαθμολογήσουν αρνητικά τις ταινίες που δεν τους άρεσαν. Γενικά παρατηρείται σε όλα τα συστήματα υπόδειξης μια τάση των χρηστών να βαθμολογούν περισσότερο τα αντικείμενα που προτιμούν και λιγότερο αυτά που δεν προτιμούν. Αυτό έχει ως αποτέλεσμα ο Naïve Bayes να έχει λίγα ή καθόλου παραδείγματα από ταινίες που βαθμολογήθηκαν αρνητικά κατά τη διάρκεια της φάσης εικπαίδευσής του. Έτσι οι πιθανότητες που υπολογίζει ο κατηγοριοποιητής για τους μικρούς βαθμούς της κλίμακας βαθμολόγησης είναι πολύ μικρές ή μηδενικές. Άρα για όλες τις νέες ταινίες οι προβλέψεις είναι θετικές (με τιμές από 4 και πάνω). Αυτό προκαλεί πολύ λανθασμένες προβλέψεις (εκτός αν θεωρήσουμε ότι υπάρχουν χρήστες που τους αρέσουν όλες οι ταινίες) και μειώνει την ακρίβεια όλου του συστήματος.

Για το λόγο αυτό αποφασίστηκε να δοκιμαστεί μια λίγο διαφορετική προσέγγιση για την αξιοποίηση του Naïve Bayes. Αντί να προσπαθεί να προβλέψει τον βαθμό που θα έδινε ο χρήστης για μία ταινία, θα μπορούσε να προβλέπει απλά αν ο χρήστης θα βαθμολογούσε θετικά (με βαθμούς 4 ή 5) ή αρνητικά μια ταινία. Αντί, δηλαδή, να υπολογίζονται οι πιθανότητες πέντε ενδεχομένων, να υπολογίζονται οι πιθανότητες μόνο δύο ενδεχομένων. Με τον τρόπο αυτό ελπίζαμε ότι η πιθανότητα να βαθμολογηθεί αρνητικά μια ταινία θα ήταν μεγαλύτερη, αφού θα υπήρχαν περισσότερα παραδείγματα εκπαίδευσης για την κατηγορία αυτή. Τελικά, όμως ούτε με αυτή την προσέγγιση προέκυψαν ακριβείς υποδείξεις, αφού το MAE των υποδείξεων ήταν 1.118. Επειδή, λοιπόν, η πιθανοθεωρητική προσέγγιση με τον Naïve Bayes δεν απέδωσε όπως αναμενόταν, δεν χρησιμοποιήθηκε περαιτέρω στις υβριδικές υλοποιήσεις.

5.3.3 Υβριδική Μέθοδος Υποκατάστασης

Ο αρχικός στόχος σχεδίασης της μεθόδου αυτής ήταν η αύξηση της κάλυψης που πετύχαινε η ΣΜΥ. Ο στόχος αυτός επιτεύχθηκε, αφού όταν δεν μπορούσαμε να παράγουμε πρόβλεψη με ΣΜΥ, γινόταν παραγωγή ΥΒΠ. Μάλιστα, επειδή όπως προαναφέραμε, η κάλυψη των ΥΒΠ είναι 100%, και η υβριδική μέθοδος υποκατάστασης έχει το ίδιο ποσοστό κάλυψης. Για την ακρίβεια της μεθόδου υπολογίστηκε MAE ίσο με 0.7501.

Είχαμε δηλαδή και μια μικρή βελτίωση της ακρίβειας του συστήματος σε σχέση με την «καθαρή» ΣΜΥ, που σύμφωνα με το τεστ Wilcoxon είναι στατιστικά σημαντική (Πίνακας 5).

Ranks

	N	Mean Rank	Sum of Ranks
CBF - HybridSubstitute			
Negative Ranks	82 ^a	91.61	7512.00
Positive Ranks	134 ^b	118.84	15924.00
Ties	5817 ^c		
Total	6033		

a. CBF < HybridSubstitute

b. CBF > HybridSubstitute

c. CBF = HybridSubstitute



Test Statistics^b

	CBF - Hybrid Substitute
Z	-4.574 ^a
Asymp. Sig. (2-tailed)	.0000047903

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Πίνακας 5 : Αποτελέσματα του Wilcoxon test για την ΣΜΥ και την Υβριδική Μέθοδο Υποκατάστασης

Ο χρόνος παραγωγής υποδείξεων χρησιμοποιώντας την ΣΜΥ στον υπολογιστή εκτέλεσης των πειραμάτων ήταν, κατά μέσο όρο, 16 δευτερόλεπτα. Η αύξηση αυτή οφείλεται στον επιπλέον χρόνο παραγωγής ΥΒΠ για τις περιπτώσεις που η παραγωγή πρόβλεψης με την ΣΜΥ δεν είναι εφικτή.

Η ακρίβεια των υποδείξεων βελτιώθηκε, με παράλληλη αύξηση της κάλυψης αλλά σε μεγαλύτερο χρόνο εκτέλεσης.

5.3.4 Υβριδική Μέθοδος Αλλαγής

Το MAE για την μέθοδο αυτή υπολογίστηκε 0.7702, ενώ η κάλυψη των ταινιών είναι 98,8%. Η διαφορά στην τιμή της MAE είναι στατιστικά σημαντική σε σχέση με τις τιμές MAE της ΣΜΥ και της Υβριδικής Μεθόδου Υποκατάστασης, όπως φαίνεται στον πίνακα 6.

Ranks

		N	Mean Rank	Sum of Ranks
Hybrid_Switching - Hybrid_Substitute	Negative Ranks	638 ^a	779.01	497009.00
	Positive Ranks	1270 ^b	1042.66	1324177.00
	Ties	4125 ^c		
	Total	6033		
CBF - Hybrid_Switching	Negative Ranks	1202 ^d	916.46	1101589.00
	Positive Ranks	519 ^e	732.55	380192.00
	Ties	4312 ^f		
	Total	6033		

a. Hybrid_Switching < Hybrid_Substitute

b. Hybrid_Switching > Hybrid_Substitute

c. Hybrid_Switching = Hybrid_Substitute

d. CBF < Hybrid_Switching

e. CBF > Hybrid_Switching

f. CBF = Hybrid_Switching

Test Statistics^c

	Hybrid_Switching - Hybrid_Substitute	CBF - Hybrid_Switching
Z	-17.184 ^a	-17.493 ^b
Asymp. Sig. (2-tailed)	3.51650000E-66	1.6084E-68

a. Based on negative ranks.

b. Based on positive ranks.

c. Wilcoxon Signed Ranks Test

Πίνακας 6 : Αποτελέσματα του Wilcoxon test για σύγκριση της Υβριδικής Μεθόδου Αλλαγής με την ΣΜΥ και την Υβριδική Μέθοδο Υποκατάστασης

Ο χρόνος παραγωγής υποδείξεων χρησιμοποιώντας την μέθοδο αυτή στον υπολογιστή εκτέλεσης των πειραμάτων ήταν, κατά μέσο όρο, 10 δευτερόλεπτα. Η μέθοδος αυτή παράγει υποδείξεις μικρότερης ακρίβειας από την ΣΜΥ και την Υβριδική Μέθοδο Υποκατάστασης, εξασφαλίζοντας κάλυψη μεγαλύτερη από την πρώτη και μικρότερη από την δεύτερη, αλλά με μια σημαντική μείωση του χρόνου παραγωγής των υποδείξεων. Αν και συνήθως μια μέθοδος υπόδειξης αξιολογείται κυρίως από την ακρίβεια των παραγόμενων υποδείξεων, για ένα σχεδιαστή ενός συστήματος υπόδειξης, πιθανότατα στο πλαίσιο ενός ηλεκτρονικού καταστήματος με πολλές χιλιάδες πελάτες, αυτή η μείωση του χρόνου εκτέλεσης μπορεί να θεωρηθεί πολύ πιο σημαντική από την μικρή απώλεια σε ακρίβεια υποδείξεων και κάλυψη αντικειμένων. Ο πίνακας 7 περιλαμβάνει συγκεντρωτικά τα αποτελέσματα των πειραμάτων που αφορούν τις μεθόδους υπόδειξης που χρησιμοποιεί το σύστημα για ευκολότερη σύγκριση.

	MAE	Κάλυψη
Συνεργατική Μέθοδος Υπόδειξης	0.7597	98.34%
Υποδείξεις Βασισμένες στο Περιεχόμενο	0.9253	100%
Υβριδική Μέθοδος Υποκατάστασης	0.7501	100%
Υβριδική Μέθοδος Αλλαγής	0.7702	98.8%

Πίνακας 7 : Συγκριτικά αποτελέσματα

5.3.5 Πείραμα επιλογής καλύτερης μεθόδου

Εκτός από τα παραπάνω πειράματα που είχαν ως στόχο να αξιολογήσουν τους αλγόριθμους που χρησιμοποιήθηκαν στο σύστημα, εκτελέσαμε ένα ακόμα πείραμα, σχετικό με τις δυνατότητες περαιτέρω βελτίωσης της ακρίβεια των υποδείξεων.

Η ιδέα πίσω από το πείραμα είναι η εξής: Τόσο η ΣΜΥ, όσο και οι ΥΒΠ έχουν συγκεκριμένα πλεονεκτήματα και μειονεκτήματα. Οι ως τώρα προσεγγίσεις προσπαθούν να συνδυάσουν τις δύο βασικές μεθόδους ώστε να αντιμετωπίσουν τα μειονεκτήματα και των δύο. Έχει διατυπωθεί η υπόθεση (Λεκάκος, 2004) για κάθε χρήστη και κάθε ταινία για την οποία γίνεται πρόβλεψη να ταιριάζει περισσότερο μια από τις δύο βασικές μεθόδους υπόδειξης. Δηλαδή, για τον ίδιο χρήστη, για άλλες ταινίες να γίνεται ΥΒΠ και για άλλες ταινίες υποδείξεις με τη ΣΜΥ. Αν όντως ισχύει κάτι τέτοιο, τότε ο σκοπός της έρευνας στο πεδίο των συστημάτων υπόδειξης δεν πρέπει να είναι η εύρεση ενός αποδοτικού τρόπου συνδυασμού των διαφορετικών τρόπων υπόδειξης, αλλά η εύρεση ενός τρόπου επιλογής της καλύτερης μεθόδου κάθε φορά. Με το πείραμα αυτό, που έχει εκτελεστεί από τον Λεκάκο (2004) με πολύ μικρότερο σύνολο δεδομένων, προσπαθούμε να διαπιστώσουμε αν μια τέτοια προσέγγιση του θέματος πράγματι υπόσχεται βελτίωση της ακρίβειας.

Στο πείραμα δουλέψαμε με την εξής μεθοδολογία: Επιλέγουμε δύο αλγόριθμους από αυτούς που έχουμε υλοποιήσει και ξεκινάμε να υπολογίζουμε το ΜΑΕ. Για κάθε ταινία του συνόλου ελέγχου, υπολογίζουμε πρόβλεψη και με τους δύο αλγορίθμους. Την πρόβλεψη που είναι πιο κοντά στην πραγματική βαθμολογία του χρήστη την ονομάζουμε Best και θεωρούμε ότι ανήκει σε μια τρίτη μέθοδο υπόδειξης. Στο τέλος, έχουμε τρεις τιμές ΜΑΕ. Δυο για τις μεθόδους που έχουμε ήδη υλοποιήσει και μια για την Best που δείχνει ποια θα ήταν η ακρίβεια του συστήματος αν ήμασταν πάντα σε θέση να επιλέγουμε την καλύτερη μέθοδο για την περίσταση. Οι διαφορές στην τιμή ΜΑΕ της Best με τις τιμές ΜΑΕ των χρησιμοποιούμενων μεθόδων είναι, προφανέστατα, στατιστικά σημαντικές.

Ακολουθούν τα αποτελέσματα του πειράματος για τα εξής ζεύγη μεθόδων υπόδειξης:

- ΥΒΠ - ΣΜΥ:

ΜΑΕ των ΥΒΠ = 0.92223936

ΜΑΕ της ΣΜΥ = 0.7598205

ΜΑΕ του Best = 0.57994133



- ΥΒΠ-Υβριδική Μέθοδος Υποκατάστασης

MAE των YBΠ = 0.9221902

MAE της YMΥ= 0.75015705

MAE του Best = 0.58166

- ΥΒΠ-Υβριδική Μέθοδος Αλλαγής

MAE των YBΠ = 0.9219762

MAE της YMΑ= 0.7701958

MAE του Best = 0.59222

- ΣΜΥ- Υβριδική Μέθοδος Αλλαγής

MAE της ΣΜΥ = 0.7587645

MAE των YMΑ = 0.7701834

MAE του Best = 0.5789857

Η πρώτη παρατήρηση σχετικά με τα αποτελέσματα είναι ότι οι δυνατότητες βελτίωσης της ακρίβειας των υποδείξεων είναι εντυπωσιακές. Σημειώνουμε ότι δεν έχουμε παρατηρήσει τόσο χαμηλές τιμές MAE από καμία μέθοδο, σε κανένα σύνολο δεδομένων στην βιβλιογραφία.

Η δεύτερη είναι ότι ανεξάρτητα με το ποιες μέθοδοι υπόδειξης χρησιμοποιούνται, η τιμή του MAE του Best είναι κοντά 0,58. Η τιμή αυτή (συν-πλην ένα μικρό ποσό) φαίνεται να είναι και ένα κάτω φράγμα για την βελτίωση της ακρίβειας των υποδείξεων.

Οι υποσχέσεις από αφήνει το πείραμα αυτό είναι μεγάλες. Δεν έχουμε βρει ακόμα τρόπο να υλοποιήσουμε μια μέθοδο υπόδειξης που θα επιλέγει πάντα την καλύτερη μέθοδο υπόδειξη για κάθε συνδυασμό χρήστη-αντικείμενο, κάτι που αποτελεί στόχο για μελλοντική έρευνα.

6. Συμπεράσματα και Μελλοντική Έρευνα

Σκοπός της παρούσας εργασίας ήταν η ανάπτυξη ενός συστήματος υπόδειξης που θα είναι σε θέση να παράγει υποδείξεις χρησιμοποιώντας τις δύο βασικές μεθόδους υπόδειξης, δηλαδή την Συνεργατική Μέθοδο Υπόδειξης και τις Υποδείξεις Βασισμένες στο Περιεχόμενο.

Με βάση τον παραπάνω στόχο, αρχικά πραγματοποιήθηκε μια επισκόπηση των ερευνητικών προσπαθειών που έχουν γίνει μέχρι στιγμής στο πεδίο των συστημάτων υπόδειξης και επισημάνθηκε ο μικρός αριθμός συστημάτων που έχουν αναπτυχθεί μέχρι στιγμής. Οι περισσότεροι ερευνητές προτείνουν νέες μεθόδους υπόδειξης ή τροποποιήσεις των προαναφερθέντων μεθόδων χωρίς όμως να τις ενσωματώνουν σε ολοκληρωμένα συστήματα υπόδειξης. Ακόμα πιο έντονη είναι η έλλειψη υλοποιημένων συστημάτων υπόδειξης τα οποία συνδυάζουν και τις δύο βασικές μεθόδους σε μια υβριδική προσέγγιση. Αυτή η παρατήρηση μας οδήγησε στον εμπλουτισμό των αρχικών στόχων της εργασίας με την ανάπτυξη υβριδικής μεθόδου η οποία εκμεταλλεύεται τα πλεονεκτήματα και μειώνει τα μειονεκτήματα καθεμιάς από τις βασικές μεθόδους υπόδειξης.

Βασισμένοι στις παραπάνω διαπιστώσεις, αναπτύξαμε το σύστημα υπόδειξης κινηματογραφικών ταινιών MoRe, που συνδυάζει τη λειτουργικότητα με την δυνατότητα πειραματισμού και εξαγωγής συμπερασμάτων. Στα πλαίσια της ανάπτυξής του, προτείναμε τροποποιήσεις της ΣΜΥ και των ΥΒΠ με στόχο την βελτίωση της ακρίβειας των υποδείξεων. Επίσης, προτείναμε δύο υβριδικές μεθόδους υπόδειξης, την Υβριδική Μέθοδο Αλλαγής και την Υβριδική Μέθοδο Υποκατάστασης. Το σύστημα MoRe είναι σε θέση να παράγει υποδείξεις χρησιμοποιώντας και τις τέσσερις προαναφερθείσες μεθόδους.

Έχοντας ολοκληρώσει την ανάπτυξη του συστήματος, συνεχίσαμε με την πειραματική αξιολόγησή του, που ήταν προσανατολισμένη στην μέτρηση της ακρίβειας των παραγόμενων υποδείξεων και την σύγκριση των διαφορετικών μεθόδων υπόδειξης που έχουν υλοποιηθεί.

Τα σημαντικότερα συμπεράσματα που προέκυψαν από την πειραματική αξιολόγηση του συστήματος MoRe είναι τα εξής:

- Η Συνεργατική Μέθοδος Υπόδειξης παράγει υποδείξεις μεγάλης ακρίβειας. Ακόμα και οι Υβριδικές μέθοδοι που βελτιώνουν την ακρίβειά της δεν

πραγματοποιούν δραματικές αλλαγές. Μεγάλο μειονέκτημα της μεθόδου είναι ο μεγάλος χρόνος παραγωγής των υποδείξεων και η μη ομαλή κλιμάκωση σε μεγάλους όγκους δεδομένων.

- Το μέγεθος της «γειτονιάς» του ενεργού χρήστη, κατά τον υπολογισμό πρόβλεψης με την ΣΜΥ για ένα συγκεκριμένο αντικείμενο, επηρεάζει σημαντικά την ακρίβεια της πρόβλεψης. Στο σύστημα MoRe αποφεύγεται η παραγωγή προβλέψεων αν η «γειτονιά» δεν αποτελείται από τουλάχιστον πέντε «γείτονες», καθώς τέτοιου είδους προβλέψεις κρίνονται τελείως ανακριβείς.
- Οι Υποδείξεις Βασισμένες στο Περιεχόμενο έχουν πολύ μικρότερη ακρίβεια σε σχέση με την ΣΜΥ. Ο τρόπος παραγωγής τους εξαρτάται κυρίως από το πεδίο εφαρμογής, ενώ δεν μπορούν να εφαρμοστούν σε κάθε είδους αντικείμενα. Η ακρίβεια των υποδείξεων εξαρτάται, επίσης, από το πεδίο εφαρμογής, το σύνολο των δεδομένων και τον τρόπο απεικόνισης των αντικειμένων. Παράγονται, όμως, πάρα πολύ γρήγορα, καταφέρνουν πάντα να παράγουν πρόβλεψη και χρειάζονται λιγότερες βαθμολογήσεις για να αρχίσουν να παράγουν υποδείξεις. Είναι προτιμότερο να χρησιμοποιούνται κατά την έναρξη της λειτουργίας των συστημάτων υπόδειξης, όταν δεν είναι διαθέσιμος μεγάλος αριθμός βαθμολογήσεων χρηστών ή σε εφαρμογές που μεταβάλλεται γρήγορα το σύνολο των αντικειμένων, οπότε οι χρήστες δεν προλαβαίνουν να βαθμολογήσουν μεγάλο αριθμό αντικειμένων.
- Η ακρίβεια των Υποδείξεων Βασισμένων στο Περιεχόμενο είναι ανάλογη του αριθμού των χαρακτηριστικών με τα οποία περιγράφονται τα αντικείμενα, τουλάχιστον στο πεδίο εφαρμογής του συστήματος MoRe. Όσο αυξάνεται ο αριθμός τους, τόσο βελτιώνεται η ακρίβεια των υποδείξεων.
- Η παραγωγή προβλέψεων χρησιμοποιώντας πιθανοθεωρητικές μεθόδους πρόβλεψης των βαθμολογιών των χρηστών (όπως ο κατιγοριοποιητής Naïve Bayes) δεν ενδείκνυται για σύνολα δεδομένων και εφαρμογές στα οποία οι βαθμολογίες των χρηστών δεν έχουν ομοιόμορφη κατανομή στις τιμές της κλίμακας βαθμολογών.
- Οι Υβριδικές Μέθοδοι Υπόδειξης βελτιώνουν την απόδοση των συστημάτων, συνδυάζοντας τις δύο βασικές μεθόδους. Οι βελτιώσεις δεν είναι θεαματικές

και συνήθως η κάθε μέθοδος βελτιώνει μια παράμετρο του προβλήματος (ακρίβεια υποδείξεων, κάλυψη αντικειμένων, χρόνος εκτέλεσης).

- Η επιλογή της καλύτερης μεθόδου υπόδειξης για κάθε συνδυασμό χρήστη-αντικειμένου μπορεί να αποτελέσει μια εναλλακτική προσέγγιση, έναντι στην δημιουργία υβριδικών, για την βελτίωση της ακρίβειας των υποδείξεων. Οι ακρίβεια που μπορεί, θεωρητικά, να επιτύχει είναι πολύ μεγαλύτερη από αυτές που έχουν μέχρι τώρα επιτευχθεί.

Τα συμπεράσματα της παρούσας εργασίας βασίζονται κυρίως σε πειραματισμό πάνω σε ένα συγκεκριμένο σύνολο δεδομένων από το πεδίο της υπόδειξης κινηματογραφικών ταινιών. Συνεπώς μια πρώτη κατεύθυνση μελλοντικής έρευνας είναι η εξέταση των μεθόδων και των τροποποιήσεων που προτάθηκαν να επεκταθεί και σε άλλα σύνολα δεδομένων που προέρχονται τόσο από το ίδιο πεδίο, όσο και από τελείως διαφορετικά πεδία (για παράδειγμα το πεδίο της υπόδειξης βιβλίων ή μουσικών κομματιών), ώστε να διαπιστώσουμε τη δυνατότητα γενίκευσης των συμπερασμάτων μας.

Ιδιαίτερα σημαντική θεωρείται η πειραματική αξιολόγηση από πραγματικούς χρήστες του συστήματος MoRe, ως σύνολο, αλλά και των χρησιμοποιούμενων μεθόδων υπόδειξης ξεχωριστά. Θα είχε ιδιαίτερο ενδιαφέρον να διαπιστωθεί εάν οι μικρές βελτιώσεις στην ακρίβεια των αλγορίθμων γίνονται αντιληπτές από τους χρήστες. Χρήσιμο θα ήταν να διερευνηθεί και σε ποια παράμετρο των επιδόσεων ενός συστήματος υπόδειξης δίνουν περισσότερη βαρύτητα οι χρήστες (ακρίβεια προβλέψεων, κάλυψη αντικειμένων ή χρόνος παραγωγής), καθώς κάτι τέτοιο θα μπορούσε θέσει τις προτεραιότητες της μελλοντικής έρευνας.

Ένα ακόμα θέμα το οποίο μπορούσε να γίνει στόχος μελλοντικής έρευνας αφορά την παρουσίαση των υποδείξεων στους χρήστες, τη μορφή της γραφικής διεπαφής των συστημάτων και το εάν αυτά επηρεάζουν τις βαθμολογήσεις των χρηστών. Παρόλο που υπάρχουν ερευνητικές εργασίες στα θέματα αυτά (Cosley et al, 2003), είναι γεγονός ότι το βάρος της έρευνας στο πεδίο των συστημάτων υπόδειξης έχει πέσει στους αλγορίθμους που χρησιμοποιούνται από τις μεθόδους υπόδειξης.

Σε ότι αφορά τις μεθόδους υπόδειξης, πιστεύουμε ότι η εύρεση ενός αλγορίθμου που θα επλέγει για κάθε συνδυασμό χρήστη-αντικειμένου την καλύτερη μέθοδο υπόδειξης θα μπορούσε να βελτιώσει αισθητά την ακρίβεια των παραγόμενων υποδείξεων.

6. Βιβλιογραφικές Αναφορές

[Adomavicius & Tuzhilin, 2005]

Adomavicius G., Tuzhilin A. (2005). "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions". *IEEE Transactions on Knowledge and Data Engineering, volume 17, no 6, June 2005.*

[Ansani et al, 2000]

Ansani A., Essegaijer S., Kohli R. (2000). "Internet Recommendations System". *Journal Marketing Research, pp 363-375. August 2000.*

[Avery & Zeckhauser, 1997]

Avery C., Zeckhauser R. (1997). "Recommender Systems for evaluating computer messages". *Communications of the ACM 40(3), pp 88-89.*

[Balabanovic & Shoham, 1997]

Balabanovic M., Shoham Y. (1997). "Fab: Content-Based, Collaborative Recommendation". *Communications of the ACM, Vol. 40, No 3, March 1997.*

[Basu et al., 1998]

Basu C., Hirsh H., Cohen W. (1998). "Recommendation as classification: Using social and content-based information in recommendation". *Proceedings of the 1998 Workshop on Recommender Systems, AAAI Press, 11-15.*

[Billsus & Pazzani, 1998]

Billsus D., Pazzani M. (1998). "Learning collaborative information filters". *Proceedings of the 15th International Conference on Machine Learning, pp 46-54.*

[Billsus & Pazzani, 2000]

Billsus D., Pazzani M. (2000). "User Modeling for Adaptive News Access". *User Modeling and User-Adapted Interaction, volume 10, pp 147-180, 2000.*

[Breese et al, 1998]

Breese J. S., Heckerman D., Kadie C. (1998). "Empirical analysis of predictive algorithms for collaborative filtering". *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98).* (pp. 43-52).

[Burke, 2002]

Burke R. (2002). "Hybrid recommender systems: Survey and experiments". *User Modeling and User-Adapted Interaction, 12, 4 (2002), 331-370.*

[Carenini & Sharma, 2004]

Carenini G., Sharma R. (2004). "Exploring more Realistic Evaluation Measures for Collaborative Filtering". *AAAI 2004, pp. 749-754*

[Chee et al, 2001]

Chee S. H. S., Han J., Wang K. (2001) "RecTree: An Efficient Collaborative Filtering Method". *In Data Warehousing and Knowledge Discovery (DaWaK), Munich, Germany, 2001.*

[Claypool et al., 1999]

Claypool M., Gokhale A., Miranta T., Murnikov P., Netes D., Sartin M. (1999). "Combining Content-Based and Collaborative Filters in an Online Newspaper". SIGIR 1999 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, CA.

[Cosley et al., 2003]

Cosley D., Lam S., Albert I., Konstan J., and Riedl J. (2003). "Is seeing believing? How recommender systems influence users' opinions". *In Proceedings of CHI 2003 Conference on Human Factors in Computing Systems, pages 585-592, Fort Lauderdale, FL, 2003.*

[Friedman & Amoo, 1999]

Friedman H. H., Amoo T. (1999). “Rating the rating scales.” *Journal of Marketing Management*, 9(3):114–123, 1999.

[Goldberg et al, 1992]

Goldberg D., Nichols D., Oki B. M., Terry D. (1992). “Using Collaborative Filtering to Weave an Information Tapestry”. *Communications of the ACM*, 35, 12 (1992), pp. 61-70.

[Goldberg et al, 2001]

Goldberg K, Roeder T., Gupta D., Perkins C. (2001). “Eigentaste: A Constant Time Collaborative Filtering Algorithm”. *Information Retrieval Journal*, vol. 4, no 2, pp 133-151, July 2001.

[Good et al, 1999]

Good N., Schafer B., Konstan J., Borchers A., Sarwar B., Herlocker J., Riedl J. (1999). “Combining Collaborative Filtering with Personal Agents for Better Recommendations”. *Proceedings Conference AAAI-1999*, pp 439-446, July 1999.

[Herlocker et al, 1999]

Herlocker J., Konstan J., Borchers A., Riedl J. (1999). “An Algorithmic Framework for Performing Collaborative Filtering”. *Proceedings of 22nd Annual International ACM SIGIR Conference Research and Development I Information Retrieval 1999*.

[Herlocker, 2000]

Herlocker J. (2000). Doctoral Thesis: “Understanding and Improving Automated Collaborative Filtering Systems”. *University of Minnesota*, 2000.

[Herlocker et al, 2002]

Herlocker J., Konstan J., Riedl J. (2002). “An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms”. *Information Retrieval*, 5 , 287-310, 2002.

[Herlocker et al, 2004]

Herlocker J., Konstan J., Terveen L., Riedl J. (2004). "Evaluating Collaborative Filtering Recommender Systems". *ACM Transactions on Information Systems*, Volume 22. No 1, January 2004, pp 5-53.

[Hill et al ,1995]

Hill W., Stead L., Rosenstein M., Furnas G. (1995). "Recommending and evaluating choices in a virtual community of use". *Proceedings of ACM CHI 1995 Conference on Human Factors in Computing Systems*, pp 194-201.

[Hofmann, 2004]

Hofmann T. (2004). "Latent Semantic Models for Collaborative Filtering". *ACM Transactions on Information Systems*, Vol. 22, No 1, January 2004, pp 98-115.

[Karypis, 2001]

Karypis G. (2001). "Evaluation of Item-Based Top-N Recommendation Algorithms". *Proceedings of CIKM 2001*, pp.247-254.

[Linden et al, 2003]

Linden G., Smith B., York J. (2004). "Amazon.com Recommendations : Item-to-Item Collaborative Filtering". *IEEE Internet Computing*, January-February 2003, pp. 76-80.

[Melville et al., 2002]

Melville P., Mooney R., Nagarajan R. (2002). "Content-Boosted Collaborative Filtering for Improved Recommendations". *Proceedings 18th National Conference Artificial Intelligence, 2002*.

[Mitchell, 1997]

Mitchell T., "Machine Learning". *McGraw Hill, 1997*

[Mooney & Roy, 2000]

Mooney R., Roy L. (2000). "Content-Based Book Recommending Using Learning for Text Categorization". *Proceedings of the V ACM Conference on Digital Libraries, San Antonio, USA, 195-204.*

[Oard & Kim, 1998]

Oard D. W., Kim J. (1998). "Implicit Feedback for Recommender Systems". *Proceedings of the 1998 Workshop on Recommender Systems 81-83.*

[O'Connor & Herlocker, 1999]

O'Connor M., Herlocker J. (1999). "Clustering Items for Collaborative Filtering". *Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkley, CA, 1999.*

[Pazzani & Billsus, 1997]

Pazzani M., Billsus D. (1997). "Learning and Revising User Profiles: The Identification of Interesting Web Sites". *Machine Learning, 27: 313-331.*

[Pazzani, 1999]

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

[Pennock et al, 2000]

Pennock D. M., Horvitz E., Lawrence S., Giles L. (2000). "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach". *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI-2000) Morgan Kaufmann, San Francisco, California, 473-480.*

[Popescul et al, 2001]

Popescul A., Ungar L., Pennock D., Lawrence S. (2001). "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments". *Proceedings 17th Conference on Uncertainty in Artificial Intelligence 2001.*

[Rashid et al., 2002]

Rashid A., Albert I., Cosley D., Lam S., McNee S., Konstan J., Riedl J. (2002). “Getting to Know You: Learning New User Preferences in Recommender Systems”. Proceedings of International Conference on Intelligent User Interfaces 2002.

[Resnick et al., 1994]

Resnick P., Iakovou N., Sushak M., Bergstrom P., Riedl J. (1994). “GroupLens: An Open Architecture for Collaborative Filtering of Netnews”. *Proceedings 1994 Computer Supported Cooperative Work Conf.*

[Sarwar et al, 1998]

Sarwar B., Konstan J., Borchers A., Herlocker J., Miller B., Riedl J. (1998). “Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System”. Proceedings of CSCW-1998, Seattle, WA: ACM

[Sarwar et al, 2001]

Sarwar B., Karypis G., Konstan J., Riedl J. (2001). “Item-Based Collaborative Filtering Recommendation Algorithms”. *Proceedings 10th International WWW Conference 2001.*

[Schafer et al, 2001]

Schafer J., Konstan J., Riedl J. (2001). “E-Commerce Recommendation Applications”. *Journal of Data Mining and Knowledge Discovery, January 2001.*

[Schafer et al, 1999]

Schafer J., Konstan J., Riedl J. (1999). “Recommender Systems in E-Commerce”. *Proceedings of the ACM 1999 Conference on Electronic Commerce.*

[Shardanand & Maes, 1995]

Shardanand U., Maes P. (1995). "Social Information Filtering: Algorithms for Automated 'Word of Mouth'". *Proceedings Conf. Human Factors in Computing Systems 1995*.

[Soboroff & Nikolas, 1999]

Soboroff I., Nikolas C., (1999). "Combining Content and Collaboration in Text Filtering". *Proceedings International Conference Artificial Intelligence Workshop: Machine Learning for Information Filtering 1999*.

[Jin et al., 2003]

Jin R., Si L., Zhai C. (2003). "Preference-Based Graphic Models for Collaborative Filtering". *Proceedings 19th Conf. Uncertainty in Artificial Intelligence (UAI 2003)*.

[Si et al., 2003]

Si L., Jin R., Zhai C., Callan J. (2003). "Collaborative Filtering with Decoupled Models of Preferences and Ratings". *Proceedings 12th Int'l Conf. Information and Knowledge Management (CIKM 2003)*.

[Ungar & Foster, 1998]

Ungar L. H., Foster D. (1998). "Clustering Methods for Collaborative Filtering". In *Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence*.

[Βαζιργιάννης & Χαλκίδη, 2003]

Βαζιργιάννης Μ., Χαλκίδη Μ. (2003) «Εξόρυξη γνώσης από βάσεις δεδομένων». Εκδόσεις ΤΥΠΩΘΗΤΩ σελ. 49-50.

[Λεκάκος, 2004]

Λεκάκος Γ. (2004). Διδακτορική Διατριβή: «Υπηρεσίες Εξατομίκευσης Διαφημίσεων μέσω Υβριδικών Μεθόδων Υπόδειξης: Εφαρμογή σε Περιβάλλον Ψηφιακής Αλληλεπιδραστικής Τηλεόρασης». Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας, Οικονομικό Πανεπιστήμιο Αθηνών.

[Χαλκιάς, 2001]

Χαλκιάς Ι. (2001) «Στατιστική: Μέθοδοι ανάλυσης για επιχειρηματικές αποφάσεις». *Εκδόσεις Rosili*, σελ. 214-215.



Δωρεά

