Πανεπιστημιο Πειραιως
Σχολη Τεχνολογιων Πληροφορικης και Επικοινωνιων
Τμημα Πληροφορικης
Προγραμμα Μεταπτυχιακων Σπουδων
«Προηγμενα Συστηματα Πληροφορικης»

Συστήματα Συστάσεων με βάση τα συμφραζόμενα

Μεταπτυχιακη Διατριβη

του

ΛΑΕΡΤΙ ΠΑΠΑ - ΜΠΣΠ14069

Επιβλέπων: Γ. Θεοδωρίδης
Καθηγητής

Πειραιάς, Μάιος 2017

Πανεπιστήμιο Πειραιώς
Σχολή Τεχνολογιών Πληροφορικής και Επικοινωνιών
Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Συστήματα Συστάσεων με βάση τα συμφραζόμενα

Μεταπτυχιακη Διατριβη

του

ΛΑΕΡΤΙ ΠΑΠΑ - ΜΠΣΠ14069

Επιβλέπων: Γ. Θεοδωρίδης
Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Μαΐου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γ. Θεοδωρίδης	Νίκος Πελέκης	Μαρία Χαλκίδη
Καθηγητής	Επίκουρος Καθηγητής	Επίκουρη Καθηγήτρια

Πειραιάς, Μάιος 2017

Πανεπιστήμιο Πειραιώς
Σχολή Τεχνολογιών Πληροφορικής και Επικοινωνιών
Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Υπεύθυνη Δήλωση

Βεβαιώνω ότι είμαι συγγραφέας αυτής της πτυχιακής εργασίας, και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης, βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του Τμήματος Πληροφορικής του Πανεπιστημίου Πειραιώς.

(Υπογραφή)

.....
ΛΑΕΡΤΙ ΠΑΠΑ

Περίληψη

Παρέχοντας σχετικές συστάσεις στους χρήστες, είναι μια πρόκληση που αντιμετωπίζουν πολλές online εφαρμογές. Στους χρήστες παρουσιάζεται ένα πλήθος αγαθών και υπηρεσιών όπως προιόντα στην Amazon ή μουσική από υπηρεσίες όπως το Spotify.

Ο στόχος των συστημάτων συστάσεων είναι να βοηθήσει τους χρήστες να πάρουν αποφάσεις, συστήνοντας τους αντικείμενα που είναι σχετικά με τα ενδιαφέροντα τους και τις τρέχων συνθήκες ή το πλαίσιο κάτω από το οποίο βρίσκεται ο χρήστης. Η προσέγγιση που ακολουθείται από τα περισσότερα συστήματα είναι χρησιμοποιώντας συστήματα συνεργατικού φιλτραρίσματος ή συστήματα που βασίζονται στο περιεχομένο. Πολύ συχνά βέβαια σε real world εφαρμογές χρησιμοποιείται μια υβριδική προσέγγιση που είναι ο συνδιασμός των δύο.

Το πρόβλημα με τις κλασσικές προσεγγίσεις στα συστήματα συστάσεων είναι ότι δεν λαμβάνουν το πλαίσιο κάτω από το οποίο ο χρήστης καταναλώνει / αγοράζει αντικείμενα ή υπηρεσίες. Πολλές έρευνες δείχνουν ότι λαμβάνοντας υπόψη το πλαίσιο σε ένα σύστημα συστάσεων μπορεί να έχει θετική επίδραση στην επίδοση του συστήματος και βοηθά στο να μοντελοποιεί τους χρήστες με περισσότερη λεπτομέρεια και μας δίνει καλύτερη κατανόηση για τη συμπεριφορά τους [26]. Επομένως η βασική υπόθεση στα συστήματα συστάσεων βασισμένα στη πληροφορία του πλαισίου είναι ότι η απόφαση που παίρνει ο χρήστης εξαρτάται από την κατάσταση κάτω από την οποία βρίσκεται κάθε χρονική στιγμή.

Ένας στόχος της παρούσας διπλωματικής εργασίας είναι να επεκταθεί ένα βασικό σύστημα συστάσεων έτσι ώστε να περιλάβει και το πλαίσιο του χρήστη στη διαδικασία της σύστασης. Στόχος επίσης είναι να βελτιώσουμε την ακρίβεια των συστάσεων που παράγονται από το σύστημα και να αξιολογήσουμε την απόδοση της προσέγγισης μας χρησιμοποιώντας μέτρα όπως η *ακρίβεια* και η *ανάκλαση*.

Ένας άλλος σημαντικός στόχος είναι ότι το σύστημα θα πρέπει να είναι επέκταση μιας υπάρχουσας προσέγγισης έτσι ώστε να μπορεί να χρησιμοποιηθεί μαζί με είδη υπάρχουσες μεθόδους στο τομέα αυτό. Αυτό εξασφαλίζει ότι οι έρευνες που έχουν γίνει στο παρελθόν στο τομέα αυτό καθώς και επενδύσεις για να φτιαχτούν συστήματα συστάσεων από εταιρίες μπορούν να χρησιμοποιηθούν σε συνδιασμό με τους αλγορίθμους που θα αναπτυχθούν.

Στο πρώτο κεφάλαιο γίνεται πλήρης επισκόπηση στα συστήματα συστάσεων και της τρέχουσας έρευνας στο τομέα αυτό. Στο κεφάλαιο 2 αναλύουμε μεθοδολογίες, αρχιτεκτονικές και αλγορίθμους στα CARS - Context Aware Recommendation Systems. Περιγράφουμε επίσης σχετικές εργασίες στο χώρο αυτό και εξετάζουμε τα πλεονεκτήματα και μειωνεκτήματα για κάθε τεχνική. Στο κεφάλαιο 3 εξετάζουμε τη προσέγγιση που ακολουθήσαμε, τους αλ-

γορίθμους που αναπτύξαμε και εξετάζουμε τις μετρήσεις που θα χρησιμοποιήσουμε για την αξιολόγηση των συστήματων. Στο κεφάλαιο 4 παρουσιάζουμε τα πειραματικά μας αποτελέσματα όπου συγκρίνουμε κλασσικούς αλγορίθμους συστάσεων με αλγορίθμους συστάσεων βασισμένα στα συμφραζόμενα. Απεικονίζουμε τις μέτρήσεις μας στους αντίστοιχους πίνακες και γραφήματα. Τέλος στο κεφάλαιο 5 κλείνουμε την εργασία με τις τελικές μας σκέψεις με μια ματιά στο μέλλον και άλλες πιθανές προεκτάσεις στο χώρο αυτό.

Λέξεις Κλειδιά

Συστήματα συστάσεων, συστήματα συστάσεων βασισμένα στο πλαίσιο, συστήματα συστάσεων βασισμένα στη συνεργασία, εξατομικευμένα συστήματα

στην οικογένεια μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω την καθηγήτρια κα. Χαλκίδη και τον καθηγητή κο. Θεοδωρίδη για την επίβλεψη αυτής της διπλωματικής εργασίας και την ευκαιρία που μου έδώσαν να ασχοληθώ με τη διπλωματική αυτή. Ένα μεγάλο ευχαριστώ στην κα. Χαλκίδη για την υπομονή της μαζί μου και κυρίως τις συμβουλές της που με οδήγησαν να οργανώσω καλύτερα τη σκέψη μου. Έπισης θα ήθελα να ευχαριστήσω τους γονείς μου και τα αδέρφια μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια. Δύσκολα χρόνια στα οποία ήταν πάντα δίπλα μου σε κάθε μου επιλογή. Ένα μεγάλο ευχαριστώ στο θείο μου Λευτέρη που χωρίς τη βοήθεια του ίσως να μην είχα αρχίσει το πρόγραμμα μεταπτυχιακών σπουδών στο Πανεπιστήμιο Πειραιά και κυρίως γιατί είναι πάντα δίπλα στην οικογένεια μου. Τέλος θα ήθελα να ευχαριστήσω τη Νάτια που ήταν δίπλα μου τα τελευταία χρόνια και που πάντα με παρότρυνε να γίνομαι καλύτερος άνθρωπος.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Συστήματα Συστάσεων - Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται οι κυριότερες τεχνικές και μέθοδοι που χρησιμοποιούνται στα συστήματα συστάσεων όπου κατηγοριοποιούνται σε τέσσερις οικογένειες: baseline predictors, Content-Based Filtering (CB), Collaborative Filtering (CF) και Υβριδικά μοντέλα. Στη συνέχεια περιγράφουμε τις μεθοδολογίες που χρησιμοποιούνται για την αξιολόγηση ενός συστήματος συστάσεων.

1.1 Ορισμός του προβλήματος

Τα Συστήματα Συστάσεων (Recommendation Systems - RS) είναι συστήματα που μαζεύουν και επεξεργάζονται ιστορικά δεδομένα χρηστών. Σκοπός είναι να εκμεταλευτούν τη πληροφορία αυτή και να προτείνουν ένα ή περισσότερα προϊόντα / αντικείμενα σε έναν χρήστη.

Το παραδοσιακό πρόβλημα στα συστήματα συστάσεων διατυπώνεται ως το πρόβλημα να υπολογίσουμε την αξιολόγηση για τα αντικείμενα που δεν έχουν καταναλωθεί - βαθμολογηθεί ή δεν τα έχει δεί ένας χρήστης ακόμα [30]. Με άλλα λόγια ένα σύστημα συστάσεων, πρέπει να προβλέψει τις προτιμήσεις ενός χρήστη και να του προτείνει μια λίστα από προϊόντα που μπορεί να τον ενδιαφέρουν. Για να επιτευχθεί αυτό, το σύστημα μαθαίνει το χρήστη κοιτώντας τη βαθμολογία που έχει δώσει σε άλλα προϊόντα αλλά και τα χαρακτηριστικά του χρήστη και των προϊόντων (δημογραφικά στοιχεία, χαρακτηριστικά προϊόντων κτλ). Αυτό συνήθως επιτυγχάνεται προβλέποντας την βαθμολογία που θα έδινε ο χρήστη σε ένα αντικείμενο I ή διαλέγοντας K αντικείμενα και να προτείνει τα πιο επιθυμητά για το χρήστη.

Πιο συγκεκριμένα, έχοντας ένα σετ χρηστών U, ένα σετ από πιθανά αντικείμενα που μπορούμε να προτείνουμε I και τη βαθμολογία που έχει δώσει ο κάθε χρήστης σε κάποια αντικείμενα r: U x I$\rightarrow$R, τότε ο αλγόριθμος για να συστήσουμε κάποια αντικείμενα σε έναν χρήστη αποτελείται από από μια συνάρτηση εκτίμησης U x I$\rightarrow$R η οποία υπολογίζει τη βαθμολογία που θα δώσει ένας χρήστης σε ένα αντικείμενο. Συνήθως οι προβλεπόμενες αξιολογήσεις $\bar{r}_{ui}$ είναι στο ίδιο εύρος με το εύρος που βρίσκονται οι πραγματικές αξιολογήσεις των χρηστών (για παράδειγμα R = [1, 5]).

Στα πλαίσια των συστημάτων συστάσεων, η βαθμολογία για ένα αντικείμενο είναι ένας γενικός όρος όπου αντιπροσωπεύει μια θετική ή αρνητική αλληλεπίδραση μεταξύ χρήστη και αντικείμενου. Οι βαθμολογίες των χρηστών μπορούν να συλλεχθούν άμμεσα ή έμμεσα. Όταν συλλέγονται άμμεσα ο χρήστης δίνει μόνος του μια βαθμολογία σε ένα αντικείμενο. Η βαθμολογία που δίνεται μπορεί να είναι μια μονάδα μέτρησης στο εύρος [1, 5], λεκτικές αξιολογήσεις όπως συμφωνώ, ουδέτερος, διαφωνώ και δυαδικές αξιολογήσεις στις οποίες ο χρήστης βαθμολογεί αν το αντικείμενο είναι καλό ή κακό. Υπάρχουν επίσης και οι μοναδιαίες αξιολογήσεις (Facebook like). Όταν οι βαθμολογίες των χρηστών συλλέγονται έμμεσα, έχουν μοναδιαία μορφή και αποτελούνται από ενέργειες του χρήστη με το σύστημα όπως η αναζήτηση για περισσότερες πληροφορίες για ένα αντικείμενο ή αγορά ενός αντικείμενου. Άλλα δεδομένα που μπρούν να βοηθήσουν στη έμμεση συλλογή βαθμολογιών είναι το κλικ σε αντικείμενα, ο χρόνος παραμονής σε μια ιστοσελίδα, πληροφορίες που συλλέγονται από το καλάθη αγορών σε μια e-commerse εφαρμογή κτλ [30].

1.2 Βασική πρόβλεψη - μη-εξατομικευμένα συστήματα συστάσεων

Στη μέθοδο αυτή δημιουργούνται βασικά συστήματα συστάσεων που δεν λαμβάνουν υπόψη τους ιστορικά δεδομένα μεταξύ χρηστη και αντικείμενο. Επομένως με τη τεχνική αυτή δεν έχουμε εξατομικευμένες συστάσεις. Η πιο γνωστή τεχνική είναι η: *τα πιο δημοφιλή* όπου τα αντικείμενα του συστήματος κατατάσσονται αναλογα με τη δημοτικότητά τους σε όλους τους χρήστες του συστήματος και τα N δημοφιλέστερα αντικείμενα εμφανίζονται στο χρήστη[30] [19].

Η πιο συχνά χρησιμοποιούμενη τεχνική είναι αυτή που μοντελοποιεί μόνο τις συστηματικές αποκλίσεις που σχετίζονται με τους χρήστες κατά την παροχή των αξιολογήσεων (ένας αυστηρώς χρήστης που βαθμολογεί χαμηλότερα από τους άλλους) και κάποιων στοιχείων που λαμβάνουν αξιολογήσεις (δημοφιλή αντικείμενα που λαμβάνουν περισσότερες και υψηλότερες αξιολογήσεις). Επομένως δηλώνοντας ως μ το συνολικό μέσο όρο αξιολόγησης η τεχνική αυτή μοντελοποιεί την απόκλιση που σχετίζεται σε ένα χρήστη u και ένα αντικείμενο i ως εξής:

$$b_{ui} = \mu + b_u + b_i \tag{1.1}$$

Οι παράμετροι b_u και b_i είναι οι παρατηρούμενες αποκλίσεις του χρήστη u και του αντικείμενου i αντίστοιχα από το μέσο όρο. Για να βρούμε τις παραμέτρους αυτές χρησιμοποιούνται κυρίως δύο βασικές μέθοδοι. Η πρώτη μέθοδος χρησιμοποιεί μέσες αποκλισεις:

$$b_u = \frac{1}{|I_u| + \beta} \sum_{i=I_u} (r_{ui} - \mu) \tag{1.2}$$

$$b_i = \frac{1}{|U_i| + \beta} \sum_{u=U_i} (r_{ui} - b_u - \mu) \tag{1.3}$$

Η παράμετρος β εξαρτάται από τα δεδομένα και θα πρέπει να βρεθεί πειραματικά. Χρησιμοποιείται για να παρέχει μια πιο λογική εκτίμηση για τους χρήστες και τα αντικείμενα που έχουν λίγες αξιολογήσεις.

Μια πιο εξελιγμένη μέθοδος που χρησιμοποιείται για να μάθουμε τις παραμέτρους, είναι λύνοντας το παρακάτω ελαχίστων τετραγώνων πρόβλημα βελτιστοποίησης:

$$min_{b*} \sum_{r_{ui} \in R} [(r_{ui} - \mu - b_u - b_i)^2 + \lambda(b_u^2 + b_i^2)] \tag{1.4}$$

Η παραπάνω συνάρτηση χρησιμοποιείται για να βελτιστοποιήσει τις παραμέτρους του μοντέλου για τις προβλέψεις των βαθμολογιών. Η ελαχιστοποίηση συνήθως λύνεται χρησιμοποιώντας τη τεχνική Stachastic Gradient Descent (SGD) [19]. Η μέθοδος βελτιστοποίησης SGD αποτελείται από τυχαία επανάληψη σε όλες τις βαθμολογίες στο σετ εκπαίδευσης και για κάθε βαθμολογία επαναλμβάνεται η εξής διαδικασία:

- Υπολογίζεται η πρόβλεψη $\bar{r}_u$ με τις τρέχων παραμέτρους του μοντέλου.
- Υπολογίζεται το σφάλμα πρόβλεψης $e_{ui} = r_{ui} - \bar{r}_{ui}$
- Για να μειώσουμε το σφάλμα πρόβλεψης της βαμολογίας, κάθε παράμετρος του μοντέλου τροποποιείται στη αντίθετη κατεύθυνση της κλίσης της συνάρτησης όπου στη συγκεκριμένη περίπτωση η κλίση είναι το σχετικό σφάλμα e_{ui} και οι αποκλίσεις του χρήστη και του αντικείμενου μεταβάλλονται ως εξής:

$$b_u \leftarrow b_u + \gamma(e_{ui} - \lambda \dot{b}_u) \tag{1.5}$$

$$b_i \leftarrow b_i + \gamma(e_{ui} - \lambda \dot{b}_i) \tag{1.6}$$

Η σταθερά λ και γ υπολογίζονται πειραματικά. Η σταθερά λ χρησιμοποιείται για να αποφευχθεί το over-fitting ενώ η σταθερά γ είναι ο ρυθμός μάθησης.

1.3 Συστήματα βασισμένα στο περιεχόμενο Content-Based Filtering

Η βασική ιδέα στα συστήματα βασισμένα στο περιεχόμενο (CB) είναι να προτείνουν σε έναν χρήστη παρόμια αντικείμενα με εκείνα που του άρεσαν στο παρελθόν [28]. Στη τεχνική αυτή προτείνουμε αντικείμενα σε έναν χρήστη αν αυτός ενδιαφέρεται σε κάποια χατακτηριστικά του αντικειμένου. Για παράδειγμα αν ξέρουμε ότι η ταινία Gotham είναι μια ταινία φαντασίας και ότι στο χρήστη Νάντια αρέσουν οι ταινίες φαντασίας, διαισθητικά μπορούμε να προτείνουμε τη ταινία Gotham στη Νάντια.

Η διαδικασία της σύστασης σε συστήματα βασισμένα στο περιεχόμενο αποτελείται από δύο κύρια συστατικά: (1) *Ανάλυση περιεχομένου*, όπου η περιγραφή των χαρακτηριστικών

των αντικειμένων αναλύονται και αναπαριστώνται σε μια πιο δομημένη μορφή (Vector Space Model) και (2) *δημιουργία προφίλ χρηστών* όπου τα προφίλ των χρηστών μοντελοποιούν τα ενδιαφέροντα τους και μαθαίνονται από τη δομημένη αναπαράσταση των αντικειμένων. Τα συστήματα βασισμένα στο περιεχόμενο μπορούν να κατηγοριοποιηθούν σε τρεις κατηγορίες ανάλογα με το πως μαθαίνεται το προφιλ του χρήστη [28]:

- Μη γραμμική προσέγγιση - k Nearest Neighbor.
- Χρησιμοποιώντας μια πιθανολογική προσέγγιση - Naïve Bayes.
- Χρησιμοποιώντας γραμμικά μοντέλα προβλεψης.

1.3.1 Αναπαράσταση αντικειμένων

Για να εφαρμοστούν οι αλγόριθμοι πρόβλεψης, αρχικά θα πρέπει όλες οι περιγραφές των αντικειμένων να αναπαρασταθούν σε μια δομημένη μορφή. Για παράδειγμα το αντικείμενο *ταινία*, μπορεί να περιγραφεί από τους *ηθοποιούς*, το *σκηνοθέτη*, το *είδος* της ταινίας και την *υπόθεση*. Η πιο δημοφιλής τεχνική στο χώρο της Ανάκτησης Πληροφορίας και στα συστήματα βασισμένα στο περιεχόμενο, είναι να αναπρασταθούν τα αντικείμενα στο Vector Space Model με το μηχανισμό TF-IDF [31] [22]. Επομένως έγγραφα κειμένου μπρορούν κωδικοποιηθούν ως διανύσματα σε ένα πολυδιάστατο Ευκλείδιο χώρο. Κάθε διάσταση αντιπροσωπεύει μια λέξη κλειδι (term ή token) που εμφανίζεται στα έγγραφα. Οι συντεταγμένες για ένα έγγραφο σε κάθε διάσταση υπολογίζονται ως το γινόμενο δύο ορών: (1) TF (Term Frequency) και (2) IDF (Inverse Document Frequency).

Ο όρος Term Frequency περιγράφει πόσο συχνά ένας όρος εμφανίζεται σε ένα έγγραφο. Για να λάβουμε υπόψη το μέγεθος του εγγράφου και να αποτρέψουμε ένα μεγάλο μεγέθους έγγραφο να πάρει μεγάλο βάρος πρέπει αρχικά να γίνει κάποια αξομάλυνση του μήκους του εγγράφου [15]. Ένας απλός μηχανισμός είναι ο εξής: Ψάχνουμε να βρούμε τη συχνότητα εμφάνισης TF(i, j) ενός όρου i σε ένα έγγραφο j. Έστω freq(i, j), η συχνότητα εμφάνισης του όρου i στο έγγραφο j. Με δεδομένο τον όρο i ορίζουμε ως OtherKeywords(i, j) το σύνολο των άλλων όρων που εμφανίζονται στο έγγραφο j. Υπολογίζουμε τη μέγιστη συχνότητα maxOthers(i, j) ως max(freq(z, j)). Τέλος υπολογίζουμε τη τιμή TF(i, j) [15]:

$$TF(i,j) = \frac{freq(i,j)}{maxOthers(i,j)} \tag{1.7}$$

Ο δεύτερος όρος IDF μειώνει το βάρος σε όρους που εμφανίζονται πολύ συχνά σε όλα τα έγγραφα. Η ιδεά είναι ότι οι όροι που εμφανίζονται πολύ συχνά σε πολλά έγγραφα δεν έχουν μεγάλη σημασία για να περιγράψουν τη μοναδικότητα ενός εγγράφου. Επομένος περισσότερο βάρος θα πρέπει να δωθεί σε όρους που εμφανίζονται σε λιγότερα έγγραφα. Έστω N ο αριθμός όλων των εγγράφων προς σύσταση και n(i) ο αριθμός των των εγγράφων από το σύνολο N όπου ο όρος i εμφανίζεται. Το βάρος IDF για τον όρο i υπολογίζεται ως:

$$IDF(i) = log(\frac{N}{n(i)}) \tag{1.8}$$

Ο συνδιασμός του βάρους TF και IDF για έναν όρο i σε ένα έγγραφο j υπολογίζεται ως το γινόμενο αυτών των δύο:

$$TFIDF(i,j) = TF(i,j) * IDF(i) \tag{1.9}$$

1.3.2 KNN - μη γραμμική προσέγγιση

Ο αλγόριθμος του πλησιέστερου γείτονα - k-Nearest Neighbor κρατάει στη μνήμη όλα τα δεδομένα των χρηστών και ταξινομεί ένα νέο υποψήφιο αντικείμενο συγκρίνοντας το με τα προηγούμενα βαθμολογημένα αντικείμενα χρησιμοποιώντας μια συνάρτηση ομοιότητας. Συγκεκριμένα ο αλγόριθμος προβλέπει μια βαθμολογία για ένα αντικείμενο με βάση τις αξιολογήσεις από τα πιο όμοια αντικείμενα που έχουν αξιολογηθεί στο παρελθόν από το χρήστη (τα πλησιέστερα γειτονικά αντικείμενα). Ανάλογα με το πως συνδιάζονται οι βαθμολογίες διάφορες συναρτήσεις εκτίμησης μπορούν να χρησιμοποιηθούν. Για παράδειγμα στη τεχνική weighted average, η πρόβλεψη για τη βαθμολογία $\bar{r}_{ui}$ που θα δώσει ένας χρήστη u σε ένα αντικείμενο i είναι η εξής:

$$\bar{r}_{ui} = \frac{\sum_{j \in S_{iu}} s_{ij} \cdot r_{uj}}{\sum_{j \in S_{iu}} s_{ij}} \tag{1.10}$$

όπου S_{iu} είναι το σύνολο των κ ομοιότερων αντικειμένων με το αντικείμενο i που έχει βαθμολογηθεί από το χρήστη u και s_{ij} είναι η ομοιότητα μεταξύ του εντικειμένου i και j με βάση τα προφίλ τους όπου υπολογίζεται ως το εσωτερικό γινόμενο των αντίστοιχων διανυσμάτων τους.

Γενικά οι kNN τεχνικές έχουν το πλεονέκτημα ότι είναι εύκολοι στην υλοποίηση τους, προσαρμόζονται γρήγορα σε πρόσφατες αλλαγές και ένας μικρός αριθμός αξιολογήσεων αρκεί για να κάνει μια πρόβλεψη ικανοποιητικής ποιότητας. Ωστόσο τα πειράματα δείχνουν ότι η ακρίβεια πρόβλεψης των μεθόδων kNN μπορεί να είναι χαμηλότερη από άλλες πιο εξελιγμένες τεχνικές [15].

1.3.3 Naïve Bayes - Πιθανολογική προσέγγιση

Οι σημαντικότεροι μέθοδοι ταξινόμησης που έχουν αναπτυχθεί στα πρώτα συστήματα ταξινόμησης εγγράφων είναι μέθοδοι που στηρίζονται στις πιθανότητες. Όλες οι τεχνικές αυτές στηρίζονται στην παραδοχή της ανεξαρτησίας υπό όρους του Bayes και έχουν χρησιμοποιηθεί σε συστήμτα συστάσεων βασισμένα στη περιγραφή των αντικειμένων [28].

Η τεχνική αυτή δημιουργεί ένα πιθανοτικό μοντέλο με τις αξιολογήσεις από ένα συνόλο εκπαίδευσης. Το μοντέλο αυτό χρησιμοποιείται για να ταξινομήσει τα υποψήφια αντικείμενα που είναι να συσταθούν σε ένα χρήστη u σε δύο πιθανές κλάσεις I_u^+, που είναι το σύνολο των αντικειμένων που ενδιαφέρουν το χρήστη και I_u^- που είναι το σύνολο των αντικειμένων που ο χρήστης δεν ενδιαφέρεται να δει. Για να ταξινομήσουμε ένα αντικείμενο i, για κάθε κλάση η Bayes θεωρία εφαρμόζεται για να υπολογίσουμε τις πιθανότητες $P(I_u^+|i)$ και $P(I_u^-|i)$ και η κλάση με τη μεγαλύτερη πιθανότητα επιλέγεται:

$$P(I_u^+|i) = \frac{P(I_u^+)P(i|I_u^+))}{P(i)} \tag{1.11}$$

$$P(I_u^-|i) = \frac{P(I_u^-)P(i|I_u^-))}{P(i)} \tag{1.12}$$

Έχωντας υπολογίσει τις παραπάνω πιθανότητες, μπορούμε να προβλέψουμε ένα σκορ r_{ui}^- υπολογίζοντας την αναλογία των πιθανοτήτων:

$$\bar{r}_{ui} = \frac{P(I_u^+)P(i|I_u^+)}{P(I_u^-)P(i|I_u^-)} \tag{1.13}$$

όπου $P(i|I_u^+)$ είναι η πιθανότητα να επιλέξουμε το αντικείμενο i από το σύνολο I_u^+. $P(I_u^+)$ είναι η πιθανότητα να επιλέξουμε ένα αντικείμενο από το σύνολο I που είναι σχετικό για το χρήστη u. Ομοίως οι όροι $P(i|I_u^-)$ και $P(I_u^-)$ αφορούν τα αντικείμενα που δεν ενδιαφέρουν το χρήστη.

1.3.4 Γραμμικά μοντέλα πρόβλεψης

Βλέποντας το πρόβλημα της σύστασης ως πρόβλημα ταξινόμησης, μπορούμε να εφαρμόσουμε και άλλες τεχνικές μηχανικής μάθησης. Γενικά, οι περισσότερες τεχνικές προσπαθούν να βρουν τους κατάλληλους συντελεστές μιας γραμμικής συνάρτησης έτσι ώστε να διακρίνουμε σχετικά και μη σχετικά έγγραφα για έναν χρήστη.

Για παράδειγμα έστω ότι τα έγγραφα έχουν δύο διαστάσεις. Ένας ταξινομητής μπορεί να αναπαρασταθεί σαν μια ευθεία γραμμή. Στο χώρο δύο διαστάσεων, η ευθεία γραμμή έχει τη μορφή: $w_1x_1 + w_2x_2 = b$ όπου x_1 και x_2 είναι η δυανυσματική αναπαράσταση των εγγράφων (για παράδειγμα βάρη TFIDF) και w_1, w_2 και b είναι οι παράμετροι που πρέπει να μαθευτούν [15]. Η ταξινόμηση ενός εγγράφου εξαρτάται από: $w_1x_1 + w_2x_2 > b$. Γενικά σε έναν χώρο n διαστάσεων η συνάρτηση ταξινόμησης είναι $\vec{w}^T\vec{x} = b$.

Άλλες τεχνικές μηχανικής μάθησης που έχουν εφαρμοστεί σε συστήματα συστάσεων είναι Decision Trees, Support Vector Machines - SVM [12] [35]. Μέθοδοι που αξιοποιούν εγκυκλοπαιδικές γνώσεις, οντολογίες και τεχνικές που αναλύουν τη σημασία των λέξεων (Sentiment analysis, word-sense disambiguation) μας επιτρέπουν να αποκτήσουμε πιο κατανοητά προφίλ χρηστών δηλαδή καλύτερη καναόηση στα πραγματικά ενδιαφέροντα του χρήστη [23] και τελικά καλύτερη ποιότητα και αποτελέσματα στο σύστημα συστάσεων [11] [21].

1.3.5 Περιορισμοί στα Συστήματα Συστάσεων με βάση το περιεχόμενο

Χρησιμοποιώντας μόνο συστήματα βασισμένα στο περιεχόμενο έχουμε κάποιους περιορισμούς. Για το λόγο αυτό υπήρξε η ανάγκη για υβριδικά συστήματα συστάσεων όπου συνδιάζουν τα πλεονεκτήματα από διαφερετικές τεχνικές. [21] [28] [15] [1]. Οι περιορισμοί στα συστήματα αυτά είναι οι εξής:

- **Περιορισμένη Ανάλυση Περιεχομένου**. Η αποτελεσματικότητα στο συγκεκριμένο σύστημα συστάσεων εξαρτάται από την ποιότητα της ανάλυσης των περιεχομένων των αντικειμένων. Συγκεκριμένα είναι η διαδικασία όπου δημιουργείται η δομημένη αναπαράσταση των αντικειμένων από την ακατέργαστη περιγραφή τους. Παράγοντες που επηρεάσουν την ανάλυση είναι μικρά έγγραφα, διαφορετικά κείμενα που έχουν παρόμια TFIDF βάρη και πληροφορία η οποία περιέχεται σε πολυμέσα όπως εικόνα, ήχος και βίντεο.

- **Υπερ-εξειδίκευση**. Επειδή το προφίλ του χρήστη δημιουργείται από αντικείμενα που έχει δει και τον ενδιαφέρουν, τα προτεινόμενα αντικείμενα θα είναι παρόμια μεταξύ τους (θα έχουν κοινά χαρακτηριστικά), ειδικά αν το προφίλ του χρήστη περιέχει μερικά ενδιαφέροντα.

- **Νέοι χρήστες και το πρόβλημα cold-start**. Τα συστήματα συστάσεων με βάση το περιεχόμενο, χρειάζονται μερικές ελάχιστες βαθμολογίες για έναν χρήστη έτσι ώστε να φτιαχτεί ένα ακριβές προφίλ για τα ενδιαφέροντα του. Οι χρήστες που δεν έχουν βαθμολογίσει αρκετά αντικείμενα αναφέρονται ως νέοι χρήστες.

1.4 Συστήματα συστάσεων βασισμένα στη συνεργασία - Collaborative Filtering

Τα συστήματα συστάσεων που βασίζονται στην συνερασία [34] [13] [19] βασίζουν τις προβλέψεις τους στις βαθμολογίες που δίνουν οι χρήστες στα αντικείμενα. Αντίθετα με τα συστήματα που βασίζονται στο περιεχόμενο των αντικειμένων, τα συστήματα αυτά είναι ανεξάρτητα από το τομέα που εφαρμόζονται αφού δεν χρειάζεται να γνωρίζουν την περιγραφή των αντικειμένων παρά μόνο βαθμολογίες χρηστών. Η τεχνική αυτή προσομοιώνει μια απλή αλλά αποτελεσματική κοινωνική στρατηγική το λεγόμενο 'λέξη από στόμα σε στόμα' η οποία βασίζεται σε συστάσεις που δίνουν οι άνθρωποι ο ένας στον άλλον ('Μου προτείναν να δω αυτήν τη ταινία σήμερα ή έχω ακούσει ότι είναι πολύ καλή'). Η βασική υπόθεση πίσω από την τεχνική αυτή είναι ότι ένας χρήστης θα ενδιαφέρεται για αντικείμενα που έχουν βαθμολογίσει υψηλά άλλοι χρήστες με κοινά ενδιαφέροντα.

Στη βιβλιογραφία συναντάμε τρείς βασικές τεχνικές [13] [34]: (1) Συστήματα που βασίζονται στους χρήστες (User-based Collaborative Filtering), (2) συστήματα που βασίζονται στα αντικείμενα (Item-based Collaborative Filtering) και (3) παραγοντοποίηση πινάκων (Matrix Factorization). Η πρώτη τεχνική υλοποιεί την απλή υπόθεση των συστημάτων βασισμένα στη συνεργασία, υπολογίζοντας τους k κοντινότερους γείτονες για έναν χρήστη με βάση την ομοιότητα των βαθμολογιών που έχουν δώσει. Η δεύτερη τεχνική (item-based) είναι πιο επεκτάσιμη υλοποίηση όπου αντί να υπολογίσει ομοιότητες μεταξύ των χρηστών, υπολογίζει ομοιότητες μεταξύ των αντικειμένων. Η παραγοντοποίηση πινάκων είναι η πιο γνωστή και η πιο ευρέως χρησιμοποιούμενη μέθοδος λόγο της επεκτάσιμότητας της και της καλύτερης πρόβλεψης που παρέχει σε πολύ μεγάλα σύνολα δεδομένων [19].

1.4.1 User-based CF

Είναι η υλοποίηση της κύριας ιδέας των συστημάτων σύστασης που βασίζονται στην συνεργασία: Βρες k χρήστες που η συμπεριφορά τους είναι παρόμοια με εκείνη του ενεργού χρήστη και υπολόγισε μια βαθμολογία για τον ενεργό χρήστη και ένα αντικείμενο με βάση την βαθμολογία που έχουν δώσει οι k γείτονικοί χρήστες. Η πιο κοινή μέθοδος για να προβλέψουμε μια βαθμολογία για έναν χρήστη u και ένα αντικείμενο i αποτελείται από τον υπολογισμό του μέσου όρου των κανονικοποιημένων βαθμολογιών που έχουν δώσει οι γειτονικοί χρήστες για το αντικείμενο i σύμφωνα με την μεταξύ τους ομοιότητα:

$$\bar{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_u} s_{uv}(r_{ui} - b_{vi})}{\sum_{v \in N_u} s_{uv}} \tag{1.14}$$

όπου N_u είναι οι γειτονικοί χρήστες του χρήστη u (οι k πιο όμοιοι χρήστες με το χρήστη u) και s_{uv} είναι η τιμή ομοιότητας μεταξύ του χρήστη u και του χρήστη v. Ο μέσος όρος του χρήστη b_{vi} χρησιμοποιείται για να ρυθμίσει τις βαθμολογίες που παρέχονται από τους γείτονικούς χρήστες αφαιρώντας το μέσο όρο του χρήστη αντίστοιχα. Η κανονικοποίηση γίνεται διότι ορισμένοι χρήστες έχουν τη τάση να δίνουν μεγαλύτερες βαθμολογίες από άλλους. Άλλες τεχνικές κανονικοποίησης είναι η Mean-Centering και z-score [15].

Ένας παράγοντας κλειδί όταν σχεδιάζουμε και υλοποιούμε user-based CF συστήματα είναι η επιλογή της συνάρτησης ομοιότητας που καθορίζει πόσο όμοιοι είναι δύο χρήστες μεταξύ τους. Διάφορες συναρτήσεις έχουν προταθεί αλλά οι πιο γνωστή είναι η Pearson correlation συνάρτηση η οποία υπολογίζει τη στατιστική συσχέτιση μεταξύ των βαθμολογιών δύο χρηστών:

$$s_{uv} = \frac{\sum_{i \in I_u \cap I_v}(r_{ui} - b_{ui})(r_{vi} - b_{vi})}{\sqrt{\sum_{i \in I_u \cap I_v}(r_{ui} - b_{ui})^2}\sqrt{\sum_{i \in I_u \cap I_v}(r_{vi} - b_{vi})^2}} \tag{1.15}$$

1.4.2 Item-Based CF

Η Item-based τεχνική είναι μια παραλλαγή της τεχνικής των κοντινότερων γειτόνων. Η διαφορά με τη User-based τεχνική είναι ότι αντί να πάρουμε την ομοιότητα μεταξύ των χρηστών, παίρνουμε την ομοιότητα μεταξύ των αντικειμένων. Επομένως έχοντας έναν χρήστη u και ένα αντικείμενο i υπολογίζουμε μια βαθμολογία που θα έδινε ο χρήστης με βάση το μέσο όρο των βαθμολογιών που έχει δώσει ο χρήστης στα πιο όμοια αντικείμενα με το αντικείμενο i:

$$\bar{r}_{ui} = b_{ui} + \frac{\sum_{j \in S_{iu}} s_{ij}(r_{uj} - b_{uj})}{\sum_{j \in S_{iu}} s_{ij}} \tag{1.16}$$

Όπου S_{iu} είναι τα k γειτονικά αντικείμενα του αντικειμένου i και s_{ij} είναι ο ομοιότητα μεταξύ των αντικειμένων i και j. Ομοίως και με την τεχνική αυτή η επιλογή της συναρτήσης ομοιότητας παίζει σημαντικό ρόλο στην απόδοση του συστήματος. Στην συγκεκριμένη περίπτωση η πιο ευρέως χρησιμοποιούμενη συνάρτηση είναι η συνάρτηση συνημιτόνου co-

sine similarity function αλλά έχουν προταθεί και άλλες συναρτησεις όπως η Conditional Probability-Based Similarity [16].

Η συνάρτηση ομοιότητας συνημιτόνου για δύο αντικείμενα i και j είναι η εξής [32]:

$$sim(i,j) = cos(\vec{i} \cdot \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{||\vec{i}||_2 * ||\vec{j}||_2} \tag{1.17}$$

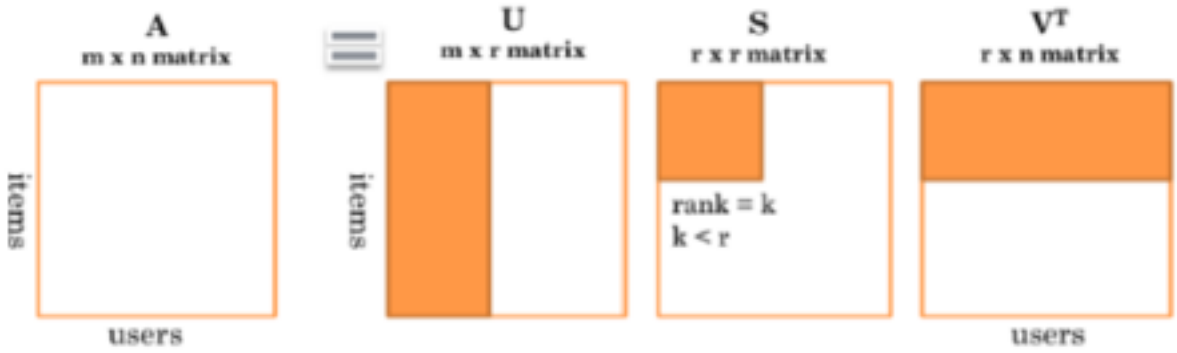
Η παραπάνω συνάρτηση όμως έχει το μειονέκτημα ότι δεν λαμβάνει υπόψη της τις διαφορές στις βαθμολογίες μεταξύ των χρηστών όπως περιγράτηκε για την εξίσωση 2.15. Για το λόγο αυτό χρησιμοποιείται κυρίως η συνάρτηση adjusted cosine similarity [32]:

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_u)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_u)^2}} \tag{1.18}$$

1.4.3 Matrix Factorization

Τα MF μοντέλα πρόβλεψης χρησιμοποιούν τη τεχνική Singular Value Decomposition (SVD) [18] [19] για να μειώσουν τις διαστάσεις του αρχικού πίνακα με τις βαθμολογίες των χρηστών σε ένα χώρο παραγόντων, όπου χαρακτηρίζει τα αντικείμενα και τους χρήστες. Ο χώρος παραγόντων περιγράφει τόσο τα αντικείμενα αλλα μοντελοποιεί και τα ενδιαφέροντα των χρηστών. Όταν βρεθεί ο χώρος παραγόντων, οι συστάσεις παράγονται με βάση το γινόμενο του διανύσματος παραγόντων τους χρήστη και του διανύσματος παραγόντων του αντικειμένου.

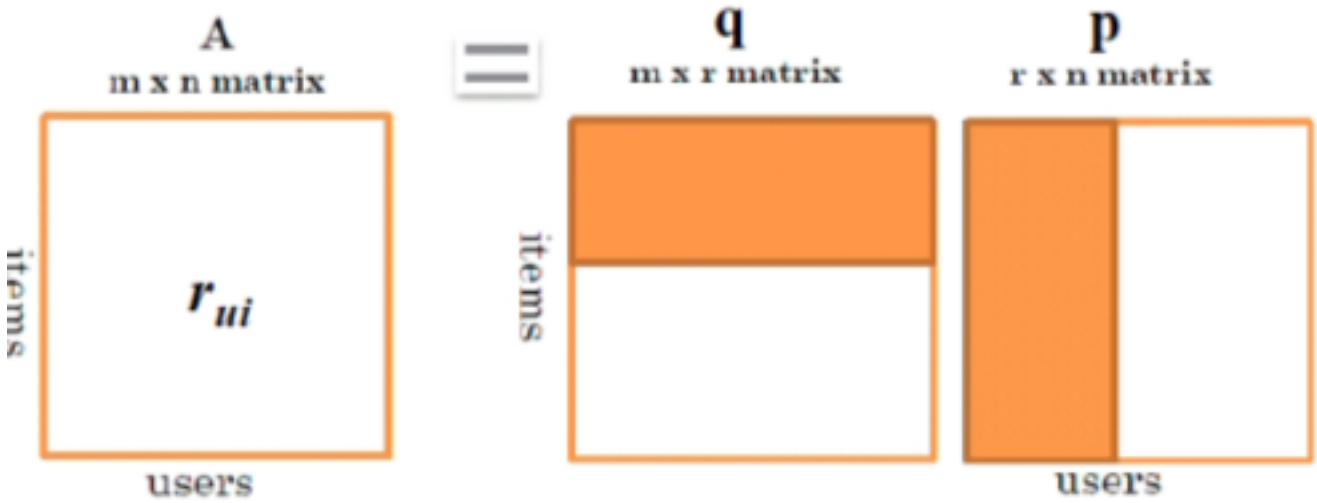
Στην εικόνα 1.1 βλέπουμε πως με τη μέθοδος SVD σπάμε το αρχικό πίνακα με τις βαθμολογίες των χρηστών σε τρεις μικρότερους πίνακες $A = UxSxV^T$.



Σχήμα 1.1: Matrix Factorization με SVD

Επειδή στα συστήματα συστάσεων ο πίνακας που περιέχει τις βαθμολογίες των χρηστών είναι πολύ αραιός διότι περιέχει πολλές ελλιπείς τιμές δεν μπορούμε να εφαρμόσουμε SVD. Διάφορες τεχνικές προτείνουν να γεμίσουμε τις τιμές που λείπουν με τυχαίες τιμές [33]. Ωστόσο η διαδικασία μάθησης είναι πολύ αργή όταν παραγοντοποιούμε πολύ μεγάλους πίνακες (Netflix price data set). Για το λόγο αυτό οι περισσότεροι ερευνητές εστιάζουν στο να μάθουν τα διανύσματα παραγόγων (latent factors) μόνο στις γνωστές βαθμολογίες των χρηστών,

χρησιμοποιώντας παραμέτρους για την αποφυγή του over-fitting. Μια δημοφιλής τεχνική για να μάθουμε τα διανύσματα παραγόγων των χρηστών και των αντικειμένων είναι χρησιμοποιώντας τη μέθοδο βελτιστοποίησης SGD (Stochastic Gradient Descent) [19]. Στην εικόνα 1.2 βλέπουμε την παραγοντοποίηση του πίνακα που περέχει της βαθμολογίες των χρηστών όταν εφαρμόσουμε SGD. Σε αυτή την περίπτωση ο αρχικός πίνακας A σπάει στο γινόμενο μεταξύ των διανυσμάτων των χρηστών και των αντικειμένων που περιέχουν τα αντίστοιχα διανύσματα παραγόντων.



Σχήμα 1.2: Προσέγγιση αρχικού πίνακα χωρίς τη χρήση SVD

Ενσωματώνοντας το μέσο όρο της βαθμολογίας των χρηστών και των αντικειμένων από την εξίσωση 2.1 το μοντέλο πρόβλεψης που βασίζεται στην παραγοντοποίηση πίνακα (Matrix Factorization prediction model) υπολογίζει την βαθμολογία που θα έδινε ένας χρήστης u σε ένα αντικείμενο i ως το άθροισμα του μέσου όρου του χρήστη και του αντικειμένου και το εσωτερικό γινόμενο μεταξύ των αντίστοιχων διανυσμάτων παραγόντων:

$$\bar{r}_{ui} = \mu + b_u + b_i + q_i^T p_u \tag{1.19}$$

χρησιμοποιώντας το παραπάνω μοντέλο, οι πίνακες q_i και p_u μπορούν να βρεθούν με την ελαχιστοποίηση του ελαχίστου τετραγώνου της συνάρτησης ως εξής [19]:

$$\min_{b_*, q_*, p_*} \sum_{r \in R} [(r_{ui} - \mu - b_u - b_i - q_i^T p_u)^2 + \lambda(b_u^2 + b_i^2 + ||q_i||^2 + ||p_u||^2)] \tag{1.20}$$

Η σταθερά λ ελέγχει την έκταση της ρύθμισης και υπολογίζεται με πολλαπλή επαλήθευση. Η ελαχιστοποίηση γίνεται είτε με τη μέθοδο Stochastic Gradient Descent (SGD) είτε με τη μέθοδο Alternative Least Squares (ALS). Ένας εύκολος αλγόριθμος επίλυσης σχεδιάστηκε αρχικά από τον Func [30]. Ο αλγόριθμος επαναλαμβάνεται για όλες τις βαθμολογίες στα δεδομένα εκπαίδευσης. Για κάθε βαθμολογία r_{ui}, υπολογίζεται μια προβλεπόμενη βαθμολο-

γία $\bar{r}_{ui}$ μαζί με το αντίστοιχο σφάλμα $e_{ui} = r_{ui} - \bar{r}_{ui}$. Για την αντίστοιχη βαθμολογία r_{ui} τροποποιούμε τις παραμέτρους στην αντίθετη κατεύθυνση της κλίσης [30]:

$$\begin{aligned} q_i &\leftarrow q_i + \gamma(e_{ui} \cdot p_u - \lambda \cdot q_i) \\ p_u &\leftarrow p_u + \gamma(e_u i \cdot q_i - \lambda \cdot p_u) \\ b_u &\leftarrow b_u + \gamma(e_{ui} - \lambda \cdot b_u) \\ b_i &\leftarrow b_i + \gamma(e_{ui} - \lambda \cdot b_i) \end{aligned}$$

Μια επεκτάση για το παραπάνο μοντέλο πρότεινε ο Korren [18] (SVD++) όπου βελτιώνει την ακρίβεια πρόβλεψης χρησιμοποιώντας επίσης και την έμμεση ανάδραση των χρηστών με το σύστημα (για παράδειγμα μοναδιαίες αξιολογήσεις χρηστών). Συγκεκριμένα, ο αλγόριθμος SVD++ προσθέτει επιπλέον άλλο ένα σύνολο παραγόντων όπου συσχετίζει κάθε αντικείμενο i με ένα διάνυσμα παραγόντων $y_i \in R^f$ όπου χρησιμοποιούνται για να χαρακτηρίσουν χρήστες με βάση τα αντικείμενα που έχουν αξιολογήσει (για παράδειγμα η βαθμολόγηση ενός αντικειμένου μπορεί να θεωρηθεί ως έμμεση βαθμολόγηση). Το εκτεταμένο μοντέλο πρόβλεψης που πρότεινε ο Korren ορίζεται ως:

$$\bar{r}_{ui} = \mu + b_u + b_i + q_i^T(q_u + |I_u|^{-\frac{1}{2}} \sum_{j \in I_u} y_j) \tag{1.21}$$

όπου I_u είναι το σύνολο των αντικειμένων που έχουν έμμεσες αξιολογήσεις από το χρήστη υ.

1.4.4 Περιορισμοί στα Συστήματα Συστάσεων που βασίζονται στη συνεργασία

Τα συστήματα συστάσεων που βασίζονται στην συνεργασία έχουν διάφορους περιορισμούς που επηρεάζουν την αποτελεσματικότητα τους [1]:

- **Ανεπάρκεια / Αραιότητα δεδομένων**. Στους περισσότερους τομείς των συστημάτων σύστασης ο πίνακας που περιέχει τις βαθμολογίες των χρηστών είναι πολύ αραιός. Με άλλα λόγια ο αριθμός των αξιολογήσεων είναι πολύ μικρότερος σε σχέση με τον αριθμό των αξιολογήσεων που πρέπει να προβλέψουμε. Η αραιότητα των δεδομένων επηρεάζει κυρίως τις τεχνικές kNN. Τα MF Μοντέλα πρόβλεψης συνήθως είναι καλύτερα σε αραιά δεδομένα αφού βασίζονται σε πρότυπα αξιολογήσεων που προέρχονται από ένα μειωμένο και με περισσότερη πληροφορία από τον αρχικό πίνακα αξιολογήσεων.

- **Νέοι χρήστες (Cold-start problem)**. Παρόμια με τις μεθόδους που βασίζονται στις περιγραφές των αντικειμένων (CB), τα συστήματα που βασίζονται στη συνεργασία χρειάζονται έναν ελάχιστο αριθμό από αξιολογήσεις ώστε να παράγουν αποδεκτές εξατομικευμένες συστάσεις.

- **Νέα αντικείμενα στο σύστημα (New-item cold start)**. Παρόμια με τους νέους χρήστες, συστήματα που βασίζονται στην συνεργασία δεν μπορούν να παράγουν

συστάσεις με ακρίβεια για αντικείμενα που έχουν αξιολογηθεί από λίγους χρήστες αφού στη συγκεκιμένη τεχνική δεν εξιοποιείται η περιγραφή του αντικειμένου όπως στις content based τεχνικές.

- **Επεκτασιμότητα**. Το υπολογιστικό κόστος στις kNN μεθόδους αυξάνεται εκθετικά με τον αριθμό των χρηστών ή των αριθμό των αντικειμένων.Σε εφαρμογές με εκατομμύρια χρήστες είναι αδύνατον να εφαρμοστούν kNN τεχνικές. Για το λόγο αυτό οι MF τεχνικές είναι καλύτερη λύση επειδή φτιάχνουν ένα μοντέλο που μπορεί και δουλεύει σε χώρο με μικρότερες διαστάσεις.

1.5 Υβριδικά Συστήματα Συστάσεων

Για να προβλέψουμε καλύτερες συστάσεις για έναν χρήστη, οι ερευνητές συνδίασαν πολλαπλές τεχνικές έτσι ώστε να φτιάξουν υβριδικά συστήματα συστάσεων όπου προσπαθούν να εξαλείψουν τις αδυναμίες που έχει η κάθε τεχνική. Δύο πλεονεκτήματα που έχουν τα συστήματα συστάσεων βασισμένα στο περιεχόμενο συμπληρώνουν το πρόβλημα της αραιότητας των δεδομένων που έχουν τα συστήματα συστάσεων βασισμένα στη συνεργασία [8]. Για το λόγο αυτό ο πιο κοινός συνδιασμός σε υβριδικά συστήματα είναι ο συνδιασμός content-based και collaborative filtering τεχνικών. Υπάρχουν διάφορες τεχνικές υβριδικών μοντέλων [8]:(1) weighted, (2) switching, (3) mixed, (4) Feature combination, (5) Meta-level, (6) Cascade, (7) Feature augmentation. Από αυτές τις τεχνικές η πιο δημοφιλής είναι η Feature augmentation ή αλλιώς *content-boosted* CF και η τεχνική meta level ή αλλιώς collaboration through content.

- **Weighted**: Στη τεχνική αυτή παίρνουμε το συνδιασμό από πολλαπλές τεχνικές πολλαπλασιασμένο με ένα βάρος για κάθε τεχνική.

- **Switching**: Στη τεχνική αυτή κατατάσουμε τα συστήματα συστάσεων κατά προτεραιότητα και αν το πρώτο σύστημα δεν μπορεί να προτείνει με μεγάλο βαθμό σιγουριάς τότε επιλέγεται το επόμενο κτλ.

- **Feature Augmentation / Combination**: Η στρατηγική αυτή χρησιμοποιείται όταν το πρωταρχικό μέρος είναι ένα σύστημα συστάσεων βασισμένο στη συνεργασία και ο στόχος είναι να ενισχύσουμε την αποτελεσματικότητα του αξιοποιώντας τεχνικές που βασίζονται στο περιεχόμενο. Μια παραλλαγή της στρατηγικής αυτής είναι να συμπληρώσουμε τις ελλειπείς βαθμολογίες χρησιμοποιώντας τεχνικές βασισμένες στη περιγραφή.

- **Meta-Level**: Η δευτερεύων τεχνική που συμβάλλει στην ενίσχυση δημιουργεί ένα καινούργιο μοντέλο βασισμένο στα δεδομένα εκπαίδευσης όπου στη συνέχεια χρησιμοποιείται για να εκπαιδεύσει τον πρωταρχικό μηχανισμό. Όταν συνδιάζουμε CB με CF τεχνικές, η στρατηγική αυτή αρχικά συνήθως φτιάχνει τα προφίλ των χρηστών χρησιμοποιώντας τη CB τεχνική και στη συνέχεια χρησιμοποιώντας τα προφίλ των χρηστών υπολογίζει τις ομοιότητες μεταξύ τους σε ένα user-based αλγόριθμο.

1.6 Αξιολόγηση Συστημάτων Συστάσεων

Τα συστήματα συστάσεων έχουν μεγάλο ενδιαφέρον τα τελευταία χρόνια τόσο ερευνητικά αλλά και εμπορικά. Ο σχεδιαστής που σχεδιάζει ένα σύστημα συστάσεων έχει να αντιμετωπίσει πολλές προκλήσεις όταν προσπαθεί να αξιολογήσει την επίδοση του συστήματος. Ανάλογα με το πρόβλημα που προσπαθούμε να αντιμετωπίσουμε, διαφορετικές μετρήσεις αξιολόγησης μπορούν να δώσουν διαφορετικά αποτελέσματα.

Στο άρθρο [36] οι συγγραφείς επικεντρώνονται στις μετρήσεις ακρίβειας ενός συστήματος συστάσεων. Μια βασική υπόθεση στα συστήματα συστάσεων είναι ότι τα συστήματα που παρέχουν πιο ακριβείς προβλέψεις είναι προτιμότερο για το χρήστη. Έτσι πολλές έρευνες έχουν γίνει με σκοπό να βρεθεί ένας αλγόριθμος που παρέχει καλύτερες προβλέψεις. Οι πιο γνωστές μετρήσεις που χρησιμοποιούνται για να αξιολογήσουν την επίδοση ενός συστήματος συστάσεων είνα οι παρακάτω:

- Ακρίβεια-Ανάκλαση(Precision-Recall). Στα συστήματα συστάσεων η ακρίβεια είναι μια μέτρηση για να αξιολογήσουμε αν ο αλγόριθμος πρόβλεψης παρέχει συστάσεις στο χρήστη που τον ενδιαφέρουν αντί να συστήνει άνευ σημασίας αντικείμενα για εκείνον. Η ανάκλαση (Recall) μετρά πόσο οι προβλέψεις που κάνουμε καλύπτουν την αρέσκεια του χρήστη. Για παράδειγμα από όλα τα αντικείμενα που ο χρήστης αλληλεπίδρασε στο τεστ σύνολο δεδομένων, πόσα μπορέσαμε να προτείνουμε.

- ROC (Reciever Operating Characteristics) καμπύλες. Οι ROC (Receiver Operating Characteristic) καμπύλες (ή καμπύλες λειτουργικού χαρακτηριστικού δέκτη), αποτελούν χρήσιμη τεχνική για την οργάνωση, επιλογή και απεικόνιση ταξινομητών με βάση τη γραφική τους παράσταση. Η καμπύλη δημιουργείται με τη γραφική αναπαράσταση του πραγματικού θετικού ποσοστού (true positive rate) και του λανθασμένου θετικού ποσοστού (false positive rate).

- RMSE (Root Mean Squared Error) όπου είναι η πιο δημοφιλής τεχνική. Το σύστημα υπολογίζει τις προβλεπόμενες αξιολογήσεις $\bar{r}_{ui}$ για ένα τεστ σύνολο δεδομένων L από ζεύγη χρηστών - αντικειμένων (u, i) όπου οι πραγματικές βαθμολογίες r_{ui} γνωρίζονται. Το σφάλμα RMSE δίνεται από το παρακάτω τύπο:

$$RMSE = \sqrt{\frac{1}{|L|} \sum_{(u,i) \in L} (\bar{r}_{ui} - r_{ui})^2} \tag{1.22}$$

1.6.1 Καμπύλες ROC και καμπύλες ανάκλασης-ακρίβειας

Όταν αξιολογούμε τα συστήματα συστάσεων, έχουμε ένα σύνολο δεδομένων που περιέχει τα αντικείμενα που έχει αλληλεπιδράσει ο κάθε χρήστης (έχει αγοράσει, έχει δει κτλ). Στη συνέχεια διαλέγουμε έναν χρήστη, κρύβουμε κάποια αντικείμενα και υπολογίζουμε μια λίστα από αντικείμενα προς σύσταση για το χρήστη. Στη συνέχεια έχουμε τέσσερα πιθανά αποτελέσματα για τα συστηνόμενα αντικείμενα και τα αντικείμενα που κρύψαμε όπως φαίνεται στο πίνακα 1.1

	Συστήνεται	Δεν Συστήνεται
Χρησιμοποιήθηκε	True-Positive (tp)	False-Negative (fn)
Δεν Χρησιμοποιήθηκε	False-Positive (fp)	True-Negative (tn)

Πίνακας 1.1: Ταξινόμηση πιθανών συστάσεων ενός στοιχείου σε ένα χρήστη. Πίνακας από [36]

Μετρώντας τα παραδείγματα για κάθε κελί του πίνακα μπορούμενα υπολογίσουμε τις ακόλουθες ποσότητες:

$$Precision = \frac{\#tp}{\#tp + \#fp} \tag{1.23}$$

$$Recall - TruePositive = \frac{\#tp}{\#tp + \#fn} \tag{1.24}$$

$$FalsePositive = \frac{\#fp}{\#fp + \#tn} \tag{1.25}$$

Στο πραγματικό κόσμο η ανάκλαση και η ακρίβεια είναι δυο ανταλλασόμενα μεθέθη αφού η αύξηση της ανάκλασης μειώνει την ακρίβεια. Για παράδειγμα αν έχουμε μια λίστα με πολλά αντικείμενα προς σύσταση αυξάνει την ανάκλαση (recall) αλλά πολύ πιθανόν να μειώσει την ακρίβεια.

Με τις παραπάνω ποσότητες μπορούμε να υπολογίσουμε τις καμπύλες λειτουργικού χαρακτηριστικού δείκτη (ROC curves), ή καμπύλες που συγκρίνουν την ανάκλαση με την ακρίβεια. Ενώ και οι δυο καμπύλες μετράνε το ποσοστό προτεινόμενων αντικειμένων που συνιστώνται, οι καμπύλες ακρίβειας-ανάκλασης δείχνουν το ποσοστό των αντικειμένων που συνιστώνται και προτιμούνται από το χρήστη, ενώ οι καμπύλες ROC δείχνουν το ποσοστό των αντικειμένων που συνιστώνται αλλά δεν προτιμούνται από τον χρήστη. Η επιλογή για τη καμπύλη που θα πρέπει να χρησιμοποιήσουμε στην αξιολόγηση του συστήματος μας εξαρτάται συνήθως από την εφαρμογή μας.

Υπάρχουν επίσης μέτρα που συνοψίζουν την ανάκλαση, ακρίβεια της καμπύλης ROC, όπως οι μετρήσεις F-measure και AUC (Area Under the ROC Curve) όπου είναι χρήσιμα για τη σύγκριση αλγορίθμων ανεξάρτητα από την εφαρμογή που χρησιμοποιούνται.

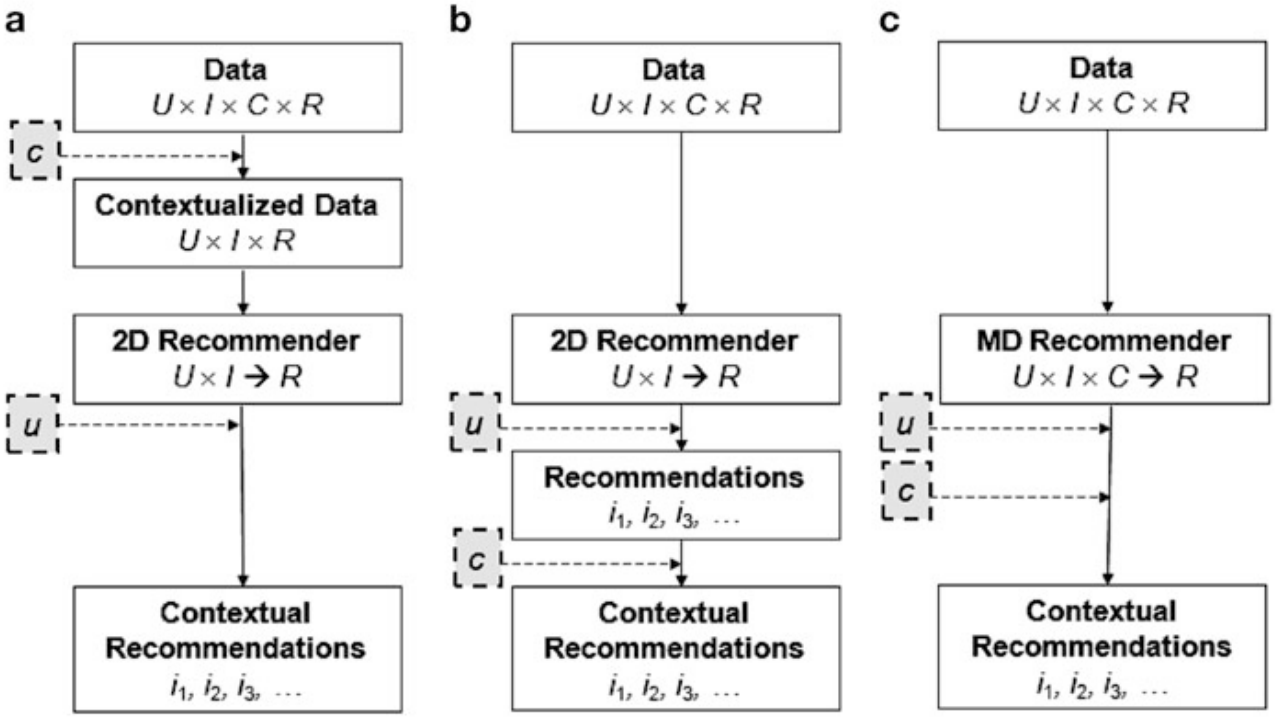
Κεφάλαιο 2

CARS - Context Aware Recommendation Systems

Στο κεφάλαιο αυτό αναφορά στα συστήματα συστάσεων που χρησιμοποιούν το πλαίσιο (Context Aware Recommender Systrems) και στη συνέχεια θα αναφέρουμε σχετικές εργασίες και έρευνες στα συστήματα συστάσεων βασισμένα στο πλαίσιο.

Τα συστήματα συστάσεων CARS επεκτείνουν τα παραδοσιακά συστήματα συστάσεων ενσωματώνοντας και το πλαίσιο (context) στις αλληλεπιδράσεις του χρήστη με τα αντικείμενα. Επομένως τα συστήματα συστάσεων CARS υπολογίζουν την αξιολόγηση ενός χρήστη u σε ένα αντικείμενο i όχι μόνο από τα δεδομένα με τις αξιολογήσεις των χρηστών αλλά εκμεταλλεύονται και τις συνθήκες κάτω από τις οποίες οι χρήστες εξιολόγησαν τα αντικείμενα, και τις συνθήκες που βρίσκεται ο χρήστης όταν αυτός ζητάει συστάσεις. Σε πολλές εφαρμογές, όπως για παράδειγμα ταξιδιωτικά πακέτα, δεν μπορούμε να στηριχτούμε μόνο σε αξιολογήσεις χρηστών ή στις περιγραφές των αντικειμένων. Είναι επίσης σημαντικό να ενσωματώσουμε πληροφορία που αφορά το πλαίσιο στη διαδικασία των συστάσεων έτσι ώστε να προτείνουμε αντικείμενα στους χρήστες υπό ορισμένες περιστάστεις. Για παράδειγμα χρησιμοποιώντας το χρονικό πλαίσιο, ένα σύστημα συστάσεων που προτείνει ταξίδια, θα πρότεινε ένα ταξίδι το χειμώνα που θα ήταν πολύ διαφορετικό από ένα ταξίδι που θα πρότεινε το καλοκαίρι [30]. Παρόμοια ένα σύστημα συστάσεων που προτείνει άρθρα για έναν χρήστη, θα πρέπει να λάβει υπόψη του και το χρόνο και το χώρο που βρίσκεται ο χρήστης όταν διαβάζει ένα άρθρο. Για παράδειγμα τις καθημερινές το πρωί ο χρήστης μπορεί να προτιμάει να διαβάζει άρθρα που που αφορούν όλο το κόσμο, το απόγευμα άρθρα που αφορούν την παγκόσμια αγορα και τα Σαββατοκύριακα να διαβάζει αθλητικά. Επομένως είναι σημαντικό να γνωρίζουμε το πλαίσιο κάτω από το οποίο επιδρά ο χρήστης με το σύστημα και τις συνθήκες που οδηγούν το χρήστη να αλληλεπιδρά πάνω σε ένα αντικείμενο. Ο Palmisano στην έρευνα του [26], έδειξε πως η ιδεά του πλαισίου παίζει σημαντικό παράγοντα όταν θέλουμε να προβλέψουμε τη συμπεριφορά των χρηστών.

Η κύρια υπόθεση πίσω από το πλαίσιο στα συστήματα συστάσεων είναι ότι οι χρήστες μπορεί να ελληλεπιδρούν με το σύστημα και συνεπώς να αξιολογούν τα αντικείμενα διαφορετικά ανάλογα με τις συνθήκες εκείνη τη στιγμή. Ανάλογα με το πως ένα σύστημα συστάσεων εκμε-



Σχήμα 2.1: Παραδείγματα για την ενσωμάτωση πλαισίου στα συστήματα συστάσεων CARS. (a) pre-filtering, (b) post-filtering, (c) contextual modeling. Εικόνα από [2]

.

ταλεύεται το πλαίσιο κάτω από το οποίο ο χρήστης αλληλεπίδρασε με το σύστημα, διακρίνουμε τρεις κύριες κατηγορίες (παραδείγματα) των CARS συστημάτων [2]:

- **Προ-φιλτράρισμα - pre-filtering**. Όπου χρησιμοποιούμε το πλαίσιο για να φιλτράρουμε τις αξιολογήσεις των χρηστών πριν υπολογίσουμε τη λίστα με τις συστάσεις σε ένα μοντέλο συστάσεων χωρίς πλαίσιο (κλασσικό σύστημα συστάσεων).

- **Μετα-φιλτράρισμα - post-fitering**. Όπου το πλαίσιο χρησιμοποιείται για να φιλτράρει τις προβλέψεις που δημιουργούνται μετά από ένα μοντέλο πρόβλεψης χωρίς πλαίσιο.

- **Μοντελοποίηση του πλαισίου - contextual modeling**. Στη τεχνική αυτή η πληροφορία του πλαισίου ενσωματώνεται στη διαδικασία πρόβλεψης ως επιπλέον παράμετρος του μοντέλου.

Στην εικόνα 2.1 [2] βλέπουμε τα τρία αυτά παραδείγματα. Αρχικά τα δεδομένα μας είναι στη μορφή U x I x C x R, δηλαδή όλες οι βαθμολογίες R των χρηστών U στα αντικείμενα I για κάθε πλαίσιο C. Στη pre-filtering τεχνική, εικόνα 2.1 (a) η πληροφορία για το τρέχων πλαίσιο c χρησιμοποιείται για να δημιουργήσει το σχετικό σύνολο δεδομένων (βαθμολογίες χρηστών). Στη συνέχεια, οι αξιολογήσεις των χρηστών μπορούν να πρεβλεθούν χρησιμοποιώντας ένα

κλασσικό σύστημα συστάσεων. Στη post-filtering τεχνική, εικόνα 2.1 (b), η πληροφορία για το πλαίσιο αρχικά αγνοείται και οι προβλέψεις υπολογίζονται για κάθε χρήστη με ένα κλασσικό σύστημα συστάσεων σε όλα τα διαθέσιμα δεδομένα. Στη συνέχεια οι προβλέψεις τροποποιούνται για κάθε χρήστη με βάση το πλαίσιο. Στη contextual modeling τεχνική, εικόνα 2.1(c), η πληροφορία του πλαισίου χρησιμοποιείται στο μοντέλο για να υπολογίσουμε τις βαθμολογίες των χρηστών.

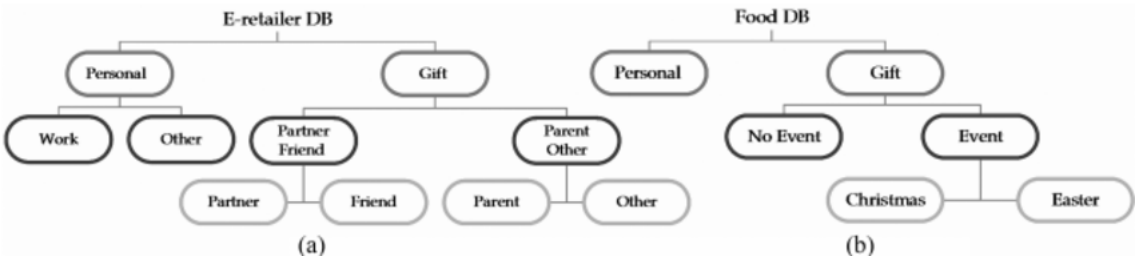
2.1 Αναπαράσταση πλαισίου

Το πλαίσιο στα συστήματα CARS χαρακτηρίζει την κατάσταση και τις συνθήκες στις οποίες ένας χρήστης αλληλεπιδρά με τα αντικείμενα. Η αναπαράσταση του πλαισίου προϋποθέτει το πλαίσιο να έχει κάποιες παρατηρήσιμες καταστάσεις και ότι οι καταστάσεις αυτές επηρεάζουν τις βαθμολογίες που δίνουν οι χρήστες στα αντικείμενα. Με άλλα λόγια οι βαθμολογίες των χρηστών στα αντικείμενα αναπαρίστανται όχι μόνο από τους χρήστες και τα αντικείμενα αλλά και το πλαίσιο:

$$R : User\ x\ Item\ x\ Context\ \rightarrow Rating \tag{2.1}$$

Σε μια εφαρμογή που προτείνει ταινίες στους χρήστες, η βαθμολογία που δίνει ένας χρήστης σε μια ταινία εξαρτάται από που την έχει δει, με ποιον την έχει δει και τι ώρα. Για παράδειγμα η ταινία που θα προτείνουμε στο χρήστη Νάντια, μπορεί να διαφέρει ανάλογα με το αν θα την δει το Σάββατο με το αγόρι της ή μια καθημερινή με τους γονείς της. Επομένως ένα πλαίσιο μπορεί να έχει διαφορετικούς τύπους (χρόνος, τοποθεσία, παρέα) και κάθε τύπος έχει πολλές καταστάσεις (καθημερινή μέρα, Σαββατοκύριακο, Θέατρο, Cinema κτλ). Η πιο συνηθισμένη μορφή αναπαράστασης των διαφορετικών τύπων σε ένα πλαίσιο είναι η ιεραρχική δομή [26] όπου οι καταστάσεις αναπαρίστανται ιεραρχικά ως δέντρο. Πιο συγκεκριμένα η πληροφορία ενός πλαισίου ορίζεται ως ένα σύνολο K που αποτελείται από καταστάσεις. Κάθε κατάσταση k στο σύνολο K ορίζεται ως ένα σύνολο από q γνωρίσματα $k = (k^1, ..., k^q)$ και αντιπροσωπεύουν ένα συγκεκριμένο τύπο όπως ο χρόνος. Οι τιμές που λαμβάνονται από το γνώρισμα K^q ορίζουν με πιο λεπτομέρεια τη κατάσταση του πλαίσιου ενώ οι τιμές στο γνώρισμα K^1 ορίζουν πιο αδρά τη κατάσταση του. Για παράδειγμα στην εικόνα 2.2 [2] φαίνεται μια ιεραρχική δομή τεσσάρων επιπέδων σε δύο βάσεις δεδομένων. Ο κόμβος ρίζα ορίζει όλα τα πιθανά πλαίσια που υπάρχουν. Το επόμενο επίπεδο $K^1 = Personal, Gift$ ορίζει κάθε συναλλαγή του πελάτη είτε σαν προσωπική είτε σαν δώρο. Όσο κατεβαίνουμε την ιεραρχία η κατάσταση που έγινε η συναλλαγη γίνεται όλο και πιο συγκεκριμένη. Για παράδειγμα στη βάση δεδομένων food DB για $K^1 = Gift$ και $K^2 = NoEvent, Event$. Αν $K^2 = NoEvent$ η κατάσταση του πλαισίου τελειώνει και η συναλλαγή έχει καταγραφεί ως ένα δώρο χωρίς να έχει κάποιο ιδιαίτερο συμβάν όπως για παράδειγμα (Χριστουγεννιάτικο δώρο κτλ). Αν $K^2 = Event$ και $K^3 = Christmas$, τότε έχουμε μια συναλλαγή για το χρήστη ο οποίος έχει πάρει ένα δώρο Χριστουγεννων (πλαίσιο `δώρο` με κατάσταση πλαισίου `Χριστουγεννιάτικο`).

Ένα σημαντικό βήμα για τα συστήματα CARS είναι να διαλέξουμε το κατάλληλο πλαίσιο που θα εκμεταλευτούμε στη διαδικασία των προβλέψεων. Ύπαρχουν δύο βασικές τεχνικές για



Σχήμα 2.2: Ιεραρχική δομή πλαισίου και συνθηκών [2]

την επιλογή των πλαισίων που θα χρησιμοποιήσουμε όπου η κάθεμιά έχει τα μειονεκτήματα της. Συγκεκριμένα (1) χρησιμοποιώντας ερωτηματολόγια, όπου ρωτάμε τους χρήστες αν μια συγκεκριμένη συνθήκη θα επηρέαζαν θετικά η αρνητικά την απόφασή τους για ένα αντικείμενο. (2) χρησιμοποιώντας στατιστικές μεθόδους βασισμένες σε δεδομένα εκπαίδευσης [25]. Ωστόσο τα ερωτηματολόγια είναι πολύ χρονοβόρα για τους χρήστες και οι στατιστικές μέθοδοι δεν είναι αξιόπιστοι εκτός αν το σύνολο των δεδομένων είναι πολύ πυκνό που συνήθως δεν είναι.

2.2 Pre-Filtering

Η pre-filtering τεχνική είναι η πιο δημοφιλής. Χρησιμοποιεί στη διαδικασία πρόβλεψης μόνο τα δεδομένα που έχουν αποκτηθεί στο πλαίσιο c, διότι μόνο αυτά τα δεδομένα είναι σημαντικά για τη πρόβλεψη των προτιμήσεων των χρηστών. Επομένως στη διαδικασία αυτή το πλαίσιο c χρησιμοποιείται για να φιλτράρει και να απορρίψει τις βαθμολογίες των χρηστών που δεν έχουν παρθεί κάτω από τις καταστάσεις του πλαισίου c.

2.3 Post-Filtering

Σε αντίθεση με τη pre-filtering τεχνική, η post-filtering τεχνική αγνοεί το πλαίσιο c κατα τη διάρκεια των προβλέψεων. Το πλαίσιο c χρησιμοποιείται μόνο για να τροποποιήσει τις προβλέψεις που έχουν γίνει από ένα σύστημα συστάσεων που δεν λαμβάνει υπόψη του το πλαίσιο. Η βασική ιδέα στη μέθοδο αυτή είναι να αναλύσουμε τα δεδομένα προτιμήσεων για ένα χρήστη σε ένα πλαίσιο c και να βρούμε συγκεκριμένα πρότυπα χρήσης του αντικειμένου (π.χ. ο χρήστης Γιάννης βλέπει μόνο *κωμωδίες* τις καθημερινές και ταινίες *δράσης* τα Σαββατοκύριακα) και στη συνέχεια να χρησιμοποιήσουμε αυτά τα πρότυπα έτσι ώστε να φιλτράρουμε τις προβλέψεις. Παρόμια με τη pre-filtering τεχνική, η post-filtering τεχνική έχει το πλεονέκτημα ότι μπορεί να χρησιμοποιηθεί σε όλα τα κλασσικά συστήματα συστάσεων που υπάρχουν στη βιβλιογραφία.

Στη post-filtering τεχνική, έχουμε δύο προσεγγίσεις [27]:

- Μέθοδοι που βρίσκουν κοινά χαρακτηριστικά για τα αντικείμενα για έναν χρήστη σε ένα πλαίσιο (π.χ. αγαπημένος ηθοποιός σε ένα πλαίσιο c) και στη συνέχεια να χρησιμοποιούν αυτά τα χαρακτηριστικά για να φιλτράρουν τις προβλέψεις.

- Μέθοδοι που χρησιμοποιούν ένα μοντέλο πρόβλεψης που προσσεγίζουν τη πιθανότητα ενός αντικειμένου i να είναι ενδιαφέρον για το χρήστη u στο πλαίσιο c. Στο άρθρο [27] ο Panniello παρουσίασε δύο τεχνικές: (1) weight post-fitlering όπου επαναδιατάσσει τα συστηνόμενα αντικείμενα πολλαπλασιάζοντας τις προβλεπόμενες βαθμολογίες με την αντίστοιχη πιθανότητα και (2) Filter post-filtering όπου απορρίπτει τα προτεινόμενα αντικείμενα που έχουν πιθανότητα μικρότερη από μια τιμή κατωφλίου.

Στη τεχνική αυτή, τα κλασσικά συστήματα συστάσεων χρησιμοποιούνται ως είσοδοι, και τα μοντέλα εκπαιδέυονται χωρίς κανένα πλαίσιο. Τα συστηνόμενα αντικείμενα στη συνέχεια προσαρμόζονται φιλτράροντας τα αποτελέσματα κάτω από τις καταστάσεις ενός πλαισίου [2]. Μόλις οι άγνωστες βαθμολογίες και οι συστάσεις για κάθε χρήστη υπολογιστούν, το σύστημα αναλύει τα δεδομένα για έναν χρήστη σε ένα πλαίσιο για να βρεί πρότυπα χρήσης των αντικειμένων στο πλαίσιο αυτό. Στη συνέχεια χρησιμοποιεί τα πρότυπα αυτά για να συστήσει αντικείμενα που είναι πιο προσιτά στο χρήστη για το συγκεκριμένο πλαίσιο. Στη weight filtering τεχνική κατατάσουμε τα αντικείμενα προς σύσταση πολλαπλασιάζοντας τη προβλεπόμενη βαθμολογία του χρήστη με τη πιθανότητα ενδιαφέροντος στο σχετικό πλαίσιο, ενώ στη δεύτερη τεχνική φιλτράρουμε τις συστάσεις (τις απορρίπτουμε εντελώς) αν έχουν μικρή πιθανότητα ενδιαφέροντος για ένα πλαίσιο. Και οι δύο περιπτώσεις αναλύουν τα δεδομένα για έναν χρήστη σε ένα συγκεκριμένο πλαίσιο έτσι ώστε να υπολογιστεί μια πιθανότητα για το πλαίσιο $P_k(i, j)$ που ο χρήστης i θα αγοράσει το αντικείμνο j κάτω από το πλαίσιο k [2]. Στη συνέχεια οι συστάσεις που υπολογίστηκαν χρησιμοποιώντας το κλασσικό σύστημα συστάσεων ξανα υπολογίζονται χρησιμοποιώντας τη πιθανότητα $P_k(i, j$ που υπολογίζεται ως ο αριθμός των γειτονικών χρηστών (παρόμιοι χρήστες) που αγόρασαν το ίδιο αντικείμενο κάτω από τις ίδιες συνθήκες δια τον συνολικό αριθμό των γειτόνων.

2.4 Contextual Modeling

Οι contextual modeling τεχνικές, συνήθως επεκτείνουν ένα κλασσικό σύστημα συστάσεων έτσι ώστε το πλαίσιο να είναι μέρος της συνάρτησης πρόβλεψης βάζοντας επιπλέον παραμέτρους που αντιπροσωπεύουν τη πληροφορία του πλαισίου. Ερευνητικά οι περισσότερες προσεγγίσεις επεκτείνουν Matrix Factorization μοντέλα πρόβλεψης. Ωστόσο άλλες προσεγγίσεις επεκτείνουν kNN τεχνικές [10] [9]

Δύο βασικές προσεγγίσεις που συναντώνται στη βιβλιογραφία είναι: (1) Tensor Factorization και (2) Context-Aware Matrix Factorization (CAMF) [5]. Η CAMF [5] προσέγγιση είναι πιο επεκτάσιμη από τη tensor factorization τεχνική αφού χρησιμοποιεί λιγότερες παραμέτρους.

2.5 Περιορισμοί στα συστήματα CARS

Τα συστήματα συστάσεων CARS βασίζονται σε Content Based και Collaborative Filtering τεχνικές, επομένως υποφέρουν από τα ίδια μειονεκτήματα. Ένα άλλο πρόβλημα που εμφανίζεται στα συστήματα CARS είναι η επεκτασιμότητα ειδικά σε Tensor Factorization

προσεγγίσεις αν ο αριθμός των καταστάσεων σε ένα πλαίσιο είναι σχετικά μεγάλος. Επιπλέον η αποτελεσματικότητα στα συστήματα CARS εξαρτάται από τη σχετικότητα των καταστάσεων για κάθε πλαίσιο που αποκτώνται και παρουσιάζονται στο σύστημα. Παρόμοια με τα συστήματα συστάσεων που βασίζονται στο περιεχόμενο, αν οι καταστάσεις που αποκτώνται και παρουσιάζονται στο σύστημα δεν περιέχουν σωστές συμφραζόμενες συνθήκες που περιγράφουν το πλαίσιο στο οποίο δώθηκε μια βαθμολογία για να βοηθήσουν να αναγνωριστεί η διαφορετική συμπεριφορά του χρήστη σε ίδιου είδους αντικείμενα, οι CARS τεχνικές αδυνατούν να καλυτερέψουν την επίδοση ενός συστήματος συστάσεων.

2.6 Σχετικές Εργασίες

Στο κεφάλαιο αυτό θα δούμε σχετικές εργασίες που έχουν γίνει στα πλάισια των συστημάτων συστάσεων. Συγκεκριμένα θα παρουσιάσουμε τις τρεις κατηγορίες των συστημάτων συστάσεων με βάση το πλαίσιο (Context Aware Recommendation Systems). Προηγουμένος είδαμε τον ορισμό του πλαισίου (context) στα συστήματα συστάσεων και τη σημασία του για την βελτίωση των συστημάτων. Οι καταστάσεις με τις οποίες ένας χρήστης αλληλεπιδρά με το σύστημα μπορούν να αξιοποιηθούν και να εκμεταλευτούν για να βγάλουμε καλύτερα συμπεράσματα για τον χρήστη και τις προτιμήσεις του. Οι καταστάσεις αυτές και το πλαίσιο κάτω από το οποίο ένας χρήστης αλληλεπιδρά με το σύστημα μπορεί να είναι: η διάθεση του χρήστη, η μέρα ή ακόμα και η ημερομηνία, τα καιρικά φαινόμενα, η τοποθεσία του χρήστη ακόμα και η κοινωνική του δικτύωση (αν είναι με ένα γκρουπ χρηστών ή μόνος του). Γενικά μπορούμε να λάβουμε υπόψη μας κάθε παράγοντα που ταιριάζει με τον ορισμό του πλαισίου που δώσαμε αν κάνουμε την υπόθεση ότι θα βελτιστοποιήσει την απόδοση του συστήματος μας.

Η κύρια πρόκληση για να βρούμε τις συνθήκες με τις οποίες αλληλεπιδρά ο χρήστης είναι ότι οι περισσότερες πληροφορίες θα πρέπει να συλλεχτούν έμμεσα για τη συμπεριφορά του χρήστη ή από εξωτερικά συστήματα αισθητηρίων (θερμοκρασία, καιρός κτλ). Πολλές έρευνες ζήτησαν από τους χρήστες εκτός από τη βαθμολογία τους να δώσουν και το πλαίσιο στο οποίο έγινε [4] [24] αλλά αυτό έχει το μειωνέκτημα ότι είναι χρονοβόρο για το χρήστη και επομένως μειώνει δραματικά την εμπειρία του χρήστη με το σύστημα Pu [29]. Επομένως οι συγραφείς στην έρευνα τους [29] προτείνουν να μειώσουμε όσο το δυνατόν περισσότερο την ανάγκη για συλλογή δεδομένων άμμεσα από τον χρήστη. Όπως θα δούμε στη συνέχεια πολλά συστήματα συστάσεων επιλέγουν να χρησιμοποιήσουν πλαίσια τα οποία είναι είδη διαθέσιμα όπως ο χρόνος, η τοποθεσία και τα καιρικά φαινόμενα.

Όπως είδαμε στην ενότητα 2 τα CARS συστήματα κατηγοριοποιούνται σε τρεις κύριες κατηγορίες [30] με βάση από το που παρέχεται η πληροφορία του πλαισίου κάτω από την οποία έγινε μια συναλλαγή του χρήστη με το σύστημα:

- Contextual pre-filtering.
- Contextual post-filtering.
- Contextual modeling.

Οι pre-filtering και post-filtering τεχνικές έχουν το πλεονέκτημα ότι μπορούν να χρησιμοποιηθούν σε όλες τις υπάρχουσες τεχνικές. Αυτές οι τεχνικές είναι κατάλληλες σε εταιρίες που έχουν επενδύσει είδη σε υψηλής ακρίβειας συστήματα συστάσεων και να τα αντικαταστήσουν εντελώς δεν θα πρόσφερε κανένα τεχνολογικό αλλά και οικονομικό ώφελός. Οι contextual-modeling τεχνικές έχουν το πλεονέκτημα της επίδοσης και της επεκτασιμότητας χρησιμοποιώντας τη πληροφορία του πλαισίου σαν κύρια διάσταση στη διαδικασία της σύστασης.

2.6.1 Contextual filtering

Στην αρχη του κεφαλαίου αναφέραμε ότι ο χρόνος είναι μια κατάσταση ενός πλαισίου που μπορεί να παρθεί πολύ έυκολα. Στην έρευνα [3] χρησιμοποιούν το χρόνο για να φιλτράρουν τα δεδομένα αλληλεπίδρασης και διαιρούν το προφίλ του χρήστη σε μικρο-προφίλ όπου το καθένα αντιπροσωπεύει το χρήστη σε κάθε πλαίσιο. Η διαδικασία της πρόβλεψης στη συνέχεια χρησιμοποιεί αυτά τα μικρο-προφίλ των χρηστών αντι για το αρχικό προφίλ του κάθε χρήστη. Η βασική ιδέα πίσω από την τεχνική αυτή είναι ότι μειώνοντας το σύνολο των δεδομένων εκπαίδευσης για κάθε μικρο-προφίλ σε ένα πλαίσιο που είναι σχετικό με τη συκεκριμένη κατάσταση, το σύστημα θα μπορέσει να μοντελοποιήσει τις χρονικές διαφορές των προτιμήσεων των χρηστών με μεγαλύτερη ακρίβεια. Ωστόσο η διάσπαση των μικρο-προφίλ των χρηστών στο χρόνο για να βελτιώσουμε τη πρόβλεψη είναι μεγάλη πρόκληση. Οι συγγραφείς στη σχετική έρευνα κατάφεραν να πάρουν καλύτερα αποτελέσματα διασπώντας τα προφίλ των χρηστών σε ζυγές και μονές ώρες, μια διάσπαση που δεν έχει σημαντική σημασιολογική σημασία αφού δεν περιμένουμε οι χρήστες να έχουν διαφορετικές προτιμήσεις για τη μουσική που ακούνε ανόλογα με την ώρα. Παρόλα αυτά τα αποτελέσματα έδειξαν ότι η διάσπαση των προφίλ των χρηστών σε μικρο-προφίλ μπορεί αν βελτιώσει την απόδοση ενός συστήματος συστάσεων.

Στο άρθρο [20] οι συγραφείς χρησιμοποιήσανε προσωρινά δεδομένα όπως η εποχή, η μέρα της εβδομάδας, η τοποθεσία και οι καιρικές συνθήκες για να μοντελοποιήσουν το πλαίσιο κάτω από το οποίο ο χρήστης αλληλεπίδρασε με το σύστημα. Το σύστημα συστάσεων τους πρότεινε τραγούδια σε χρήστες. Κάθε χρήστης στο σύστημα είχε ένα προφίλ το οποίο περιλάμβανε το φύλο, την ηλικία και τις συνήθειες των χρηστών (πια τραγούδια ακούνε περισσότερο). Οι συνθήκες στις οποίες ένας χρήστης άκουγε ένα τραγούδι καταγραφόταν μαζί με τις συνήθειες τους. Η διαδικασία πρόβλεψης για να προτείνουν ένα τραγούδι σε έναν χρήστη περιλάμβανε την εύρεση παρόμοιων χρηστών που άκουγαν μουσική σε ίδιες συνθήκες και χρησιμοποίησαν το ιστορικό των παρόμιων χρηστών για να προτείνουν τραγούδοα σε έναν χρήστη.

Στο άρθρο [40] οι συγγραφείς ξεκινώντας από το γεγονός ότι χρησιμοποιώντας πολλές καταστάσεις στις οποίες οι χρήστες αλληλεπίδρουν, μειώνουν την απόδοση του συστήματος συστάσεων, εξέτασαν το πρόβλημα συστάσεων με μια διαφορετική προσέγγιση. Σπάσανε έναν κλασσικό αλγόριθμο συστάσεων σε τρία συστατικά και έδειξαν πως μπορούν να μπουν περιορισμοί για κάθε πλαίσιο και να εφαρμοστούν ξεχωριστά σε κάθε συστατικό έτσι ώστε να μειώσουμε το σφάλμα πρόβλεψης. Συγκεκριμένα ενσωμάτωσαν διαφορετικές καταστάσεις στις οποίες αλληλεπιδρούν οι χρήστες με το σύστημα σε διαφορετκά μέρη του αλγορίθμου

που χρησιμοποιείται για να πρβλέψει τη βαθμολογία ενός χρήστη σε ένα αντικείμενο. Η διαφορά με άλλες έρευνες είναι ότι το φιλτράρισμα δεν γίνεται σε όλο τον αλγόριθμο αλλά σε κάθε συστατικό του χρησιμοποιείται διαφορετική στρατηγική φιλτραρίσματος και διαλέγονται διαφορετικές βαθμολογίες χρηστών που έχουν δοθεί ανάλογα με το αν τηρούν κάποιους περιορισμούς C. Με βάση την ιδέα αυτή χρησιμοποιήσανε ένα κλασσικό συστήματα συστάσεων (user-based) ως παράδειγμα και σύνθεσαν ένα υβριδικό αλγόριθμο. Στη συνέχεια προσδιόρισαν τις ευνοϊκότερες συνθήκες για κάθε συστατικό του αλγορίθμου. Χρησιμοποιώντας την εξίσωση 1.14 συνθεσαν τη παρακάτω εξίσωση [40]:

$$P_{a,i,C} = \bar{r}_{a,C_3} + \frac{\sum_{u \in N_{C_1}} (r_{u,i,C_2} - \bar{r}_{u,C_2}) * sim(a,u)}{\sum_{u \in N_{c_1}} sim(a,u)} \tag{2.2}$$

Όπου:

- N_{C_1} είναι το σύνολο των χρηστών που βαθμολόγισαν το αντικείμενο i κάτω υπό μια κατάσταση που ικανοποιούν τους περιορισμούς C_1.
- $\bar{r}_{u,C_2}$είναι ο μέσος όρος βαθμολόγισης του γείτονα u αλλά λαμβάνοντας υπόψη μόνο τις βαθμολογίες που πάρθηκαν στο πλαίσιο που ικανοποιεί τους περιορισμούς C_2.
- $\bar{r}_{a,C3}$ είναι ο μέσος όρος βαθμολογίας του χρήστη a, του χρήστη που θέλουμε να προβλέψουμε τη βαθμολογία του στο αντικείμενο i, αλλά χρησιμοποιώντας μόνο τις βαθμολογίες που ικανοποιούν τους περιορισμούς C_3.

Παρατηρώντας ότι είναι ίδια με την εξίσωση 1.14 με τη μόνη διαφορά ότι επιλέγεται διαφορετικό πλαίσιο στο κάθε συστατικό του αλγορίθμου. Στα πειραματικά αποτελέσματα λάβανε υπόψη τους ένα σύνολο δεδομένων από ταξίδια που έχουν κάνει οι χρήστες σε πόλεις των Ηνωμένων Πολιτειών (dataset από το www.tripadvisor.com) όπου είχαν στη διάθεση τους προφιλ χρηστών, βαθμολογίες, γεωγραφικά δεδομένα, το χρόνο που έγινε το ταξίδι, το σκοπό του ταξιδιού και η παρέα. Στους υπολογισμούς τους χρησιμοποίησαν τρεις καταστάσεις για να περιγράψουν το πλαίσιο στο οποίο έγινε το ταξίδι: (α) το σκοπό του ταξιδιού, (β) το προορισμό και (γ) τη πόλη αναχώρησης. Επίσης εφάρμοσαν τους παρακάτω περιορισμούς:

- C_1 Γειτονικοί χρήστες που μένουν στην ίδια περιοχή.
- C_2 Μέσος όρος του κάθε γείτονα που υπολογίζεται με βάση το τύπο του ταξιδιού.
- C_3 Ο μέσος όρος των χρηστών λαμβάνοντας υπόψη όλες τις βαθμολγίες.

Όπου και κατάφεραν να πάρουν μικρότερο RMSE σφάλμα από το user-based και pre-filtering αλγόριθμο. Το μειονέκτημα στη τεχνική τους είναι η εύρεση των καταλληλότερων καταστάσεων και περιορισμών που έγινε με εξαντλητική αναζήτηση.

Σε επόμενη έρευνα τους [41] οι ίδιοι συγγραφείς επέκτηναν τη τεχνική αυτή αλλά βάζοντας βάρη σε κάθε διάσταση του πλαισίου αντι για περιορισμούς. Οι συγγραφείς στην έρευνα τους σύγκριναν τις δύο μεθόδους και έδειξαν πως βάζονται βάρη σε κάθε πλαίσιο το σύστημα συστάσεων έχει μεγαλύτερη ακρίβεια.

Μια διαφορετική προσέγγιση παρουσιάστηκε από τον Baltrunas και Ricci [6] η οποία είναι γνωστή ως item splitting. Η ιδέα πίσω από την τεχνική αυτή είναι να χωρίσουμε το διάνυσμα των βαθμολογιών για ένα αντικείμενο σε δύο εικονικά διανύσματα αντικειμένων χρησιμοποιώντας ένα ειδικό παράγοντα για το πλαίσιο. Για παράδειγμα οι βαθμολογίες που έχουν δώσει οι χρήστες σε ταινίες, θα μπορούσαν να δημιουργήσουν δύο σύνολα απο βαθμολογίες: το σύνολο το βαθμολογιών που συλλέχτηκαν όταν ο χρήστης έβλεπε τη ταινία μόνος τους και το σύνολο των βαθμολογιών όταν ο χρήστης έβλεπε την ταινία με παρέα, θεωρώντας ότι το πλαίσιο είναι η παρέα που ο χρήστης βλέπει την ταινία. Στη συνέχεια ένα μοντέλο πρόβλεψης εκπαιδεύεται λαμβάνοντας υπόψη όλες τις αξιολογησεις στο εκτεταμένο σύνολο των αντικειμένων που δημιουργείται από τον διαχωρισμό των αντικειμένων που ικανοποιούν μια στατιστική συνθήκη (για παράδειγμα μετρώντας αν τα δύο εικονικά αντικείμενα που διαχωρίστηκαν είναι πολύ διαφορετικά μεταξύ τους). Στην τεχνική αυτή το φιλτράρισμα γίνεται στα αντικείμενα με βάση τη πιο σημαντική κατάσταση που δόθηκε η αξιολόγηση. Άλλες παρόμιες προσεγγίσεις δώθηκαν από τον Baltrunas και Amatriain [3] όπου αντί να διαχωρίσουν τα αντικείμενα διαχώρισαν τους χρήστες σε μικρο-προφίλ, όπου το καθένα αντιπροσωπεύει το χρήστη κάτω από ένα συγκεκριμένο πλαίσιο. Ομοίως ένα μοντέλο πρόβλεψης δημιουργείται χρησιμοποιώντας όλες τις αξιολογήσεις στο τροποποιημένο σύνολο των χρηστών. Ακόμα πιο πρόσφατα ο Zheng χρησιμοποίησε ένα συνδιασμό των δύο προηγούμενων τεχνικών όπου έβγαλε μεγαλύτερη ακρίβεια πρόβλεψης σε ένα σύνολο δεδομένων με αξιολογήσεις χρηστών για ταινίες [42].

2.6.2 Contextual modeling

Στη μοντελοποίηση του πλαισίου, τα συστήματα συστάσεων δύο διαστάσεων (κλασσικά συστήματα συστάσεων) δεν χρησιμοποιούνται κατά τη διάρκεια των συστάσεων. Αντί αυτού, όλα τα πολυδιάστατα δεδομένα (χρήστες, αντικείμενα και το πλαίσιο) χρησιμοποιούνται για να εκπαιδεύσουν ένα σύστημα συστάσεων. Οι περισσότερες έρευνες στη μοντελοποίηση του πλαισίου έχουν επικεντρωθεί στο να επεκτείνουν είδη υπάρχουσες τεχνικές που βασίζονται σε μοντέλα πρόβλεψης που χρησιμοποιούν παραγοντοποίηση πινάκων (Matrix Factorization). Οι CAMF (Context-Aware Matrix Factorization) τεχνικές είναι πιο επεκτάσιμες τεχνικές σε σχέση με άλλες (π.χ. Tensor Facotrization) [5]. Η μέθοδος αυτή είναι μια γενίκευση του αλγορίθμου που πρότεινε ο Koren [30] όπως είδαμε στο κεφάλαιο 1.4.3. Στα πλαίσια του διαγωνισμού Netflix με έπαθλο ένα εκατομμύριο δολάρια, οι συγγραφείς έδειξαν πως ο αλγόριθμος τους (timeSVD++) ήταν ένας από τους καλύτερους αλγορίθμους πρόβλεψης για τα δεδομένα της Netflix και μια παραλλαγή του αλγορίθμου σε συνδιασμό με άλλους αλγορίθμους κέρδισε το έπαθλο.

Στο άρθρο [4] ο Baltrunas πρότεινε να επεκταθούν τα μοντέλα πρόβλεψης Matrix Factorization εισάγοντας το πλαίσιο στη διαδικασία της πρόβλεψης μοντελοποιώντας την αλληλεπίδραση μεταξύ των αντικειμένων και του του πλαισίου. Οι συγγραφείς πρότειναν τρείς αλγορίθμους για τα συστήματα CAMF: (α) CAMF-C που μοντελοποιεί την επιρροή ενός πλαισίου συνολικά σε όλα τα δεδομένα (υποθέτει ότι έχει το ίδιο αποτέλεσμα σε κάθε χρήστη και σε κάθε αντικείμενο), (β) ο αλγόριθμος CAMF-CI μοντελοποιεί την επιρροή μιας κα-

τάστασης ενός πλαισίου ομοιόμορφα σε όλα τα αντικείμενα (υποθέτωντας ότι δεν εξαρτάται από το χρήστη) και (γ) ο αλγόριθμος CAMF-CC όπου υποθέτει ότι το πλαίσιο επιδρά ομοιόμορφα σε όλες τις αξιολογήσεις των αντικειμένων του ίδιου τύπου (αντικείμενα που ανήκουν στην ίδια κατηγορία). Για παράδειγμα ο αλγόριθμος CAMF-CI [5] έχει τον παρακάτω τύπο:

$$\bar{r}_{uic_1...ck} = \mu + b_u + \sum_{j=1}^{k} B_{ijc_j} + q_i^T p_u \tag{2.3}$$

Μια παραλλαγή των παραπάνω αλγορίθμων προτάθηκε από τον Odic [24] όπου μοντελοποιεί την επιρροή ενός πλαισίου ομοιόμορφα σε όλους τους χρήστες CAMF-CU:

$$\bar{r}_{uic_1...ck} = \mu + b_i + \sum_{j=1}^{k} B_{ujc_j} + q_i^T p_u \tag{2.4}$$

Για να μπορέσουμε να παράγουμε συστάσεις, οι παραμέτροι του μοντέλου θα πρέπει να μαθευτούν από τα δεδομένα εκπαίδευσης. Επομένως η διαδικασία μάθησης ορίζεται ως ένα πρόβλημα βελτιστοποίησης:

$$\min_{b_*,q_*,p_*,B_*} \sum_{r\in R} [(r_{uic_1..c_k} - \mu - b_u - q_i^T p_u - \sum_{=1}^{k} [B_{ijc_j}])^2 + \lambda(b_u^2 + ||q_i||^2 + ||p_u||^2 + \sum_{j=1}^{k} [B_{ijc_j}^2]^2)]$$

Οι παράμετροι ενημερώνονται ως εξής:

- $b_u = b_u + \gamma b_u(err - \lambda b_u)$
- $B_{ijc_j} = B_{ijc_j} + \gamma B_{ijc_j}(err - \lambda B_{ijc_j})$
- $p_u = p_u + \gamma p_u(err \cdot q_i - \lambda p_u)$
- $q_i = q_i + \gamma q_i(err \cdot p_u - \lambda q_i)$

Στο άρθρο [37] οι συγγραφείς πρότειναν μια μέθοδο για να συστήνουν προορισμούς ταξιδιών σε χρήστες με βάση τα ταξίδια που έχουν κάνει σε άλλες πόλεις στο παρελθόν. Η μέθοδος τους λαμβάνει υπόψη τις καιρικές συνθήκες και την εποχή που έγινε το ταξίδι. Χρησιμοποίησαν τεχνικές μηχανικής μάθησης (Latent Dirichlet allocation (LDA) και Probablistic Latent Sematic Analysis - PLSA) [7] για να εξορύξουν την κατανομή ενδιαφέροντος των χρηστών, όπου στη συνέχεια αξιοποιείται για την κατασκευή του μοντέλου ομοιότητας των χρηστών και να κάνουν συστάσεις για ταξίδια. Συγκεκριμένα χρησιμοποιήσαν φωτογραφίες χρηστών από ένα σύνολο δεδομένων το οποίο συλλέχτηκε από τη δημόσια Διεπαφή Προγραμματισμού Εφαρμογών (API) της ιστιοσελίδας flickr μια κοινωνική ιστιοσελίδα δικτύωσης που επιτρέπει στους χρήστες να ανεβάζουν φωτογραφίες για διάφορα μέρη που έχουν επισκευτεί. Οι ερευνητές αξιοποίησαν τις φωτογραφίες με τον αλγόριθμο P-DBSCAN (παραλλαγή του DBSCAN)

[17] για να εξάγουν τη τοποθεσία που έχει τραβηχτεί η φωτογραφία. Παράλληλα αξιοποίησαν την ημερομηνία που έχει τραβηχτεί η φωτογραφία και χρησιμοποίησαν εξωτερικές υπηρεσίες καιρού για να κρατήσουν τις καιρικές συνθήκες που επικρατούσαν. Στη συνέχεια αφού βρήκαν τη τοποθεσία των ταξιδιωτών (τοποθεσία που έχουν τραβηχτεί οι φωτογραφίες) σχημάτισαν τα προφίλ των προορισμών (των ἀντικειμένων᾽ προς σύσταση). Αρχικά αναγνώρισαν τη συχνότητα επίσκεψης των χρηστών σε κάθε τοποθεσία ενώ στη συνέχεια βρήκαν το πλαίσιο κάτω από το οποίο ένας χρήστης επισκέφτηκε κάθε προορισμό (εποχή και καιρικές συνθήκες την εποχή εκείνη). Για το προφίλ των προορισμών υπολόγισαν επίσης τα πιο δημοφιλή πλαίσια για κάθε προορισμό από τις ιστορικές επισκέψεις του. Χρησιμοποίησαν επίσης τεχνικές Latent Semantic Analysis για να υπολογίσουν τα ᾽θέματα᾽ (topics) των ταξιδιών που έχουν κάνει οι χρήστες στο παρελθόν, όπου ο κάθε χρήστης μοντελοποιείται ως ένα μείγμα από ᾽θέματα᾽ και κάθε ᾽θέμα᾽ μοντελοποιείται ως μια πιθανοτική κατανομή των προορισμών. Για να προτείνουν συστάσεις σε ένα χρήστη δημιουργείται ένα ερώτημα στο σύστημα $Q = (u_p, s, w, d)$ για έναν χρήστη u_p που θέλει να επισκεφτεί μια πόλη s κάτω από το πλαίσιο με καιρικές συνθήκες w και την εποχή s. Στη συνέχεια για να πάρουν τις συστηνόμενες τοποθεσίες στην πόλη αυτή εφαρμόζουν τεχνικές που βασίζονται στη συνεργασία (collaborative filtering). Αρχικά υπολογίζουν τους N πιο όμοιους χρήστες και στη συνέχεια υπολογίζουν ένα σκορ για κάθε τοποθεσία με βάση την εξίσωση 1.16, με τη διαφορά ότι η ομοιότητα των χρηστών υπολογίζεται με μια διαφορετική συνάρτηση η οποία προέρχεται από την πιθανοτική κατανομή των θεμάτων των ταξιδιών που έχουν κάνει οι χρήστες στο παρελθόν. Τέλος φιλτράρουν τις τοποθεσίες για τη πόλη s που δεν ικανοποιούν τους περιορισμούς στο ερώτημα Q και αποκτούν ένα φιλτραρισμένο σύνολο δεδομένων τουριστικών τοποθεσιών. Από τα φιλτραρισμένα δεδομένα αυτά, το σύστημα γυρνάει m αποτελέσματα. Τα αποτελέσματα τους είναι εντυπωσιακά αφοού έδειξαν ότι η μεθοδος τους γυρνάει καλύτερα αποτελέσματα από άλλες τεχνικές. Οι συγγραφείς στο τέλος επίσης αναφέρουν ότι στο μέλλον θα προσπαθήσουν να συλλέξουν περισσότερες πληροφορίες για κάθε χρήστη όπως δημογραφικά στοιχεία με την υπόθεση ότι θα βγάλουν ακόμα καλύτερα αποτελέσματα.

Τέλος αξίζει να σημειώσουμε ότι παρόλο που υπάρχουν διαφορετικές τεχνικές για τα συστήματα CAMF η έρευνα που έγινε στο άρθρο [9] δείχνει πως κανείς δεν μπορεί να πει με σιγουριά πια τεχνική υπερισχύει από μια άλλη.

Κεφάλαιο 3

Μελέτη προσεγγίσεων συστάσεων με βάση τα συμφράζομενα (context)

Στο κεφάλαιο αυτό θα δούμε τις μεθοδολογίες που ακολουθήσαμε στα πλαίσια της διπλωματικής αυτής εργασίας. Συγκεκριμένα θα επικεντρωθούμε σε έναν αλγόριθμο βασισμένο στο post filtering μοντέλο. Ο λόγος που επιλέχθηκε το μοντέλο αυτό είναι διότι μπορούμε να χρησιμοποιήσουμε υπάρχων αλγορίθμους σύστασης ως βάση στον αλγόριθμο μας. Επομένως με αυτό το τρόπο όλες οι ερευνητικές προσπάθειες που έχουν είδη πραγματοποιηθεί στον τομέα των συστημάτων συστάσεων εξακολουθούν να ισχύουν στο μοντέλο αυτό σε αντίθεση με το contextual μοντέλο όπου ένας αλγόριθμος θα πρέπει να κατασκευαστεί από το μηδέν για να εκμεταλλευτεί όλες τις πληροφορίες του πλαισίου που γίνονται οι συστάσεις. Στη συνέχεια του κεφαλαίου θα περιγράψουμε το βασικό αλγόριθμο που χρησιμοποιήθηκε για τις συστάσεις και στη συνέχεια θα περιγράψουμε τη post-filtering μέθοδο που αναπτύξαμε.

3.1 Βασικός Αλγόριθμος Item based Collaborative Filtering

Δεδομένου ότι οι αλγόριθμοι που χρησιμοποιήθηκαν στην εργασία αυτή ανήκουν στη post-filtering τεχνική, ένα κλασσικό σύστημα συστάσεων χρησιμοποιήθηκε ως πηγή για τις συστάσεις του συστήματος. Επιλέξαμε να χρησιμοποιήσουμε τον αλγόριθμο item based collaborative filtering Που περιγράψαμε στο κεφάλαιο 2. Επομένως στην ενότητα αυτή επαναλάβουμε εν συντομία τις βασικές παραδοχές πίσω από τον αλγόριθμο αυτό και θα δούμε μερικές λεπτομέρειες υλοποίησης.

Τα συστήματα συστάσεων που βασίζονται στον αλγόριθμο item-item cf προσπαθούν να ξεπεράσουν τα προβλήματα των user-user cf κοιτάζοντας τις ομοιότητες μεταξύ των στοιχείων, αντί της ομοιότητας μεταξύ των χρηστών. Καθώς ο αριθμός των χρηστών και των στοιχείων μεγαλώνει τον υπολογισμό παρόμοιων χρηστών κατά τη διάρκεια των συστάσεων και επομένως

ο υπολογίσμός των συστάσεων είναι πολύ χρονοβόρος υπολογιστικά. Με τον item-item CF αλγόριθμο προ-υπολογίζουμε τις ομοιότητες μεταξύ των στοιχείων και χρησιμοποιούμε αυτές τις ομοιότητες για να φτιάξουμε συστάσεις προς τους χρήστες, υπολογίζοντας τις βαθμολογίες που θα έδινε ένα χρήστης σε ένα νέο αντικείμενο με βάση τις αξιολογήσεις του χρήστη σε παρόμια αντικείμενα.

Η κύρια παραδοχή πίσω από αυτόν τον αλγόριθμο είναι ότι οι χρήστες θα πρέπει να ενδιαφέρονται περισσότερο σε στοιχεία όμοια με εκείνα που έχουν αξιολογήσει θετικά στο παρελθόν και ενδιαφέρονται λιγότερο σε στοιχεία παρόμοια με εκείνα που έχουν δώσει αρνητική ή χαμηλή βαθμολογία. Η επιτάχυνση για τον αλγόριθμο προέρχεται από το γεγονός ότι αυτές οι ομοιότητες μπορούν να προ υπολογιστούν και χρησιμοποιούνται στο χρόνο σύστασης. Ενώ ο προ-υπολογισμός αυτός θα μπορούσε επίσης να γίνει για τους χρήστες στη θεωρία για έναν user-based αλγόριθμο αλλά θα βρεθούμε αντιμέτωποι με το πρόβλημα ότι οι ομοιότητες μεταξύ των χρηστών είναι πιο δυναμικοί και μπορούν να αλλάξουν δραματικά καθώς οι χρήστες αξιολογούν περισσότερα αντικείμενα. Σε αντίθεση, οι σχέσεις μεταξύ των αντικειμένων είναι πιο στατική και δεν αλλάζει δραματικά. Τα στοιχεία έχουν αρκετές αξιολογήσεις και μας επιτρέπεται να εκτελέσουμε τον ακριβό υπολογισμό της ομοιότητας των αντικειμένων περιοδικά και να μην κάνουμε μια ακριβή αναζήτηση στο πίνακα χρήστες-αντικείμενα κατά το χρόνο σύστασης.

Στη συνέχεια θα περιγράψουμε μια σύντομη επεξήγηση της διαδικασίας σύστασης για αυτόν τον αλγόριθμο. Ο αλγόριθμος προβλέπει την βαθμολογία που θα έδινε ένας χρήστης u σε ένα αντικείμενο i εξετάζοντας τα αντικείμενα που εχει αξιολογήσει ο χρήστης και παίρνει την ομοιότητα για κάθε βαθμολογημένο εντικείμενο με το αντικείμενο i. Αφού ανακαλύψουμε τα παρόμοια αντικείμενα, η βαθμολογία που θα έδινε ο χρήστης στο αντικείμενο i προβλέπεται από το σταθμισμένο μέσο όρο των αξιολογήσεων που έχει δώσει ο χρήστης στα παρόμοια αντικείμενα. Δεδομένου ότι κάθε στοιχείο αναπαρίσταται ως ένα διάνυσμα από αξιολογήσεις που έχουν δοθεί για το στοιχείο αυτό, μπορούμε να χρησιμοποιήσουμε οποιαδήποτε συνάρτηση ομοιότητας διανυσμάτων για να υπολογίσουμε τις ομοιότητες μεταξύ τους. Υπάρχουν πολλοί τρόποι για να υπολογίσουμε πόσο όμοια είναι δύο αντικείμενα μεταξύ τους. Δύο μέτρα που εξετάστηκαν στα πλαίσια της εργασία αυτής είναι η Ευκλείδια συνάρτηση ομοιότητας και η συνάρτηση Pearson Correlation. Η Pearson-r συνάρτηση ομοιότητας φαίνεται στη εξίσωση 3.1 όπου υπολογίζεται η ομοιότητα μεταξύ δύο αντικειμένων i και j.

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_j)^2}} \tag{3.1}$$

Όπου U είναι το σύνολο των χρηστών που έχουν αξιολογήσει και τα δύο αντικείμενα i και j και $\bar{R}_i$ είναι ο μέσος όρος βαθμολογίας για το αντικείμενο i.

Στην εξίσωση 3.2 είναι η ευκλείδια συνάρτηση ομοιότητας μεταξύ δύο αντικειμένων i, j όπου U είναι το σύνολο των χρηστών που έχουν βαθμολογίσει και τα δύο αντικείμενα και $R_{u,i}$ είναι η βαθμολογία που έχει δώσει ο χρήστηες u στο αντικείμενο i

$$sim(i,j) = \frac{1}{\sqrt{\sum_{u \in U}(R_{u,i} - R_{u,j})^2}} \tag{3.2}$$

Ύστερα από τα πειράματα μας επιλέξαμε την pearson συνάρτηση ως μέτρο ομοιότητας.

Επόμένως, μπορούμε να προβλέψουμε τη βαθμολογία $r^{-}_{u,i}$ του χρήστη u για ένα αντικείμενο i παίρνοντας το καταθμισμένο μέσο όρο:

$$r^{-}_{u,i} = \frac{\sum_{j \in Q_i(u)} similarity(i,j) * r_{u,j}}{\sum_{j \in Q_i(u)} |similarity(i,j)|} \tag{3.3}$$

Στην εξίσωση 3.3 το σύνολο $Q_i(u)$ είναι οι k πλησιέστεροι γείτονες του εντικειμένου i, similarity είναι η συνάρτηση ομοιότητας και $r_{u,j}$ είναι η βαθμολογία που έχει δώσει ο χρήστης υ στο αντικείμενο i. Η τιμή των γειτόνικών χρηστών k εξαρτάται από την εφαρμογή και συνήθως υπολογίζεται με cross validation. Στα πειράματα μας θέσαμε τη τιμή κ ίση με 50.

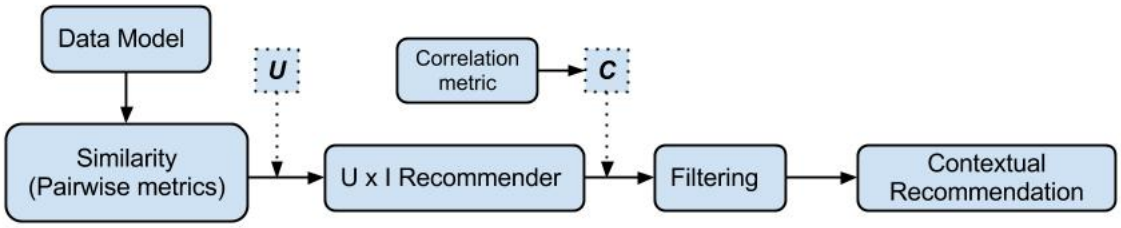
3.2 Post-filtering

Η προσέγγιση που ακολουθήσαμε περιγράφεται από τον Panniello στο άρθρο [27], μια post filtering τεχνική όπου οι συστάσεις που δημιουργούνται από ένα κλασσικό σύστημα συστάσεων τροποποιούνται χρησιμοποιώντας μία από τις δύο μεθόδους, filter ή weight με βάση τη πιθανότητα P(u,i,c), όπου u είναι ο χρήστης που βαθμολογεί το αντικείμενο i κάτω από το πλαίσιο c. Οι συγγραφείς στο άρθρο [27] υπολογίζουν την πιθανότητα αυτή παίρνοντας ένα προκαθορισμένο αριθμό από γειτονικούς χρήστες (NN) για το χρήστη u και εξετάζουν πόσοι από τους γειτονικούς αυτούς χρήστες έχουν αλληλεπιδράσει με το αντικείμενο i κάτω από το πλαίσιο c. Η πιθανότητα P(u,i,c) υπολογίζεται από την διαίρεση του συνολικού αριθμού των γειτονικών χρηστών που έχουν αλληλεπιδράσει με το αντικείμενο i δια το σύνολο των γειτονικών χρηστών. Στη μέθοδο weight αυτη η πιθανότητα πολλαπλασιάζεται με το σκορ που βγάζει ένα κλασσικό σύστημα συστάσεων και η λίστα των συστάσεων κατατάσσεται με βάση το καινούργιο σκορ. Στη μέθοδο Filter χρησιμοποιούμε τη πιθανότητα του πλαισίου για να φιλτράρουμε τα αποτελέσματα και να τα βγάλουμε από τη λίστα με τις συστάσεις. Οι συστάσεις φιλτράρονται αν η πιθανότητα P(u,i,c) είναι χαμηλότερη από ένα συγκεκριμένο όριο.

Στην παρούσα εργασία ακολουθήσαμε τρεις τεχνικές στα πειράματα μας για να αξιολογήσουμε την επίδοση του συστήματος συστάσεων. Οι δύο πρώτες τεχνικές βασίζονται στην μέθοδο που πριγράφεται στο άρθρο [27]. Η τρίτη τεχνική είναι ένας συνδιασμός των δύο παραπάνω.

Στην εικόνα 3.1 βλέπουμε την αρχιτεκτονική του συστήματος συστάσεων. Το Data Model αντιπροσωπεύει το αρχικό σύνολο δεδομένων όπου το dataset διαβάζεται από ένα CSV αρχείο και αποθηκεύεται σε δύο ξεχωριστά data stores αντιπροσωπεύοντας τα χαρακτηριστικά των χρηστών και των ταινιών. Με βάση τη συνάρτηση ομοιότητας, βρίσκονται άλλοι χρήστες που έχουν δώσει παρόμιες βαθμολογίες με εκείνες του χρήστη που εξετάζουμε (target user).

Στη συνέχεια χρησιμοποιούμε Collaborative Filtering τεχνικές και με βάση την εξίσωση 3.3 συστήνουμε ταινίες σε έναν χρήστη u με βάση τις προτιμήσεις που έχουν άλλοι παρόμιοι



Σχήμα 3.1: Contextual post-filtering

χρήστες. Οι συστάσεις αυτές στη συνέχεια φιλτράρονται ή για κάθε σύσταση βάζουμε ένα βάρος ανάλογα με τη μέθοδο που ακολουθούμε (filter ή weight). Στη συνέχεια επαναδιατάσουμε με τη λίστα με τις νέες συστηνόμενες ταινίες. Στο σχήμα 3.1 το μπλοκ U αντιπροσωπεύει τον ενεργό χρήστη που παράγουμε τις συστάσεις και το μπλοκ C αντιπροσωπεύει το πλαίσιο κάτω από το οποίο θα φιλτράρουμε τις συστάσεις.

Το ερώτημα q που τίθεται στο σύστημα συστάσεων είναι το εξής:

$$q = (u, context, condition) \tag{3.4}$$

όπου u είναι ο ενεργός χρήστης και context είναι το πλαίσιο κατω από το οποίο βρίσκεται ο χρήστης. Στα πλαίσια της διπλωματικής αυτής εργασίας το ερώτημα που γίνεται στο σύστημα χρησιμοποιεί μία μόνο διάσταση για κάθε πλαίσιο και όχι πολλές. Condition είναι η διάσταση για το πλαίσιο context στην οποία βρίσκεται ο χρήστης.

Στη προσέγγιση μας υπολογίζουμε την πιθανότητα $P(u, i, c)$ για έναν χρήστη u να ενδιαφέρεται για ένα αντικείμενο i κάτω από το πλαίσιο c. Η πιθανότητα P(u,i,c) υπολογίζεται ώς ο συνολικός αριθμός των γειτονικών χρηστών του ενεργού χρήστη u που είδαν την ταινία κάτω από το πλαίσιο c προς τον συνολικό αριθμό των γειτονικών χρηστών:

$$P(u, i, c) = \frac{\sum_{v \in NN(u)} r_{vic}}{|N(N(u))|} \tag{3.5}$$

Η υπόθεση πίσω από την τεχνική αυτή είναι ότι για τα αντικείμενα που οι παρόμοι χρήστες έχουν ψηφίσει κάτω από το πλαίσιο c θα πρέπει να υπολογίζονται περισσότερο κατά τη διάρκεια των συστάσεων. Η πιθανότητα P(u,i,c) θα μπορούσε να υπολογιστεί με πιο εξειδικευμένες τεχνικές μηχανικής μάθησης όπως για παράδειγμα έναν classifier.

Με βάση τη πιθανότητα P(u,i,c) χρησιμοποιήσαμε τρείς μεθόδους: (α) weight, (β) filter και (γ) combined-weight-filter για να υπολογίσουμε τις συστάσεις. Στη weight τεχνική η πιθανότητα P(u,i,c) πολλαπλασιάζεται με τη προβλεπόμενη βαθμολογία ενός κλασσικού συστήματος συστάσεων και στη συνέχεια επαναδιατάσουμε τις συστάσεις κατά φθίνουσα σειρά προτίμησης. Αντίστοιχα στη filter τεχνική χρησιμοποιούμε τη πιθανότητα για το πλαίσιο c για να φιλτράρουμε τις συστάσεις που έχουν πιθανότητα μικρότερη από μια τιμή t. Στη combined τεχνική συνδιάζουμε τις δύο τεχνικές σε μια ενοποιημένη συνάρτηση 3.8.

Στη μέθοδο weight το τελικό σκορ υπολογίζεται με βάση την εξίσωση 3.6. Στη μέθοδο filter χρησιμοποιούμε την εξίσωση 3.7 όπου αφαιρούμε τις συστάσεις που έχουν πιθανότητα κάτω από το κατώφλι t. Η τιμή κατωφλίου που επιλέξαμε στα πειράματα μας είναι t = 0.1. Η παράμετρος αυτή είναι σχετική και εξαρτάται από το σύνολο δεδομένων που χρησιμοποιείται για κάθε εφαρμογή. Μια άλλη παράμετρος για το σύστημα είναι ο αριθμός των γειτονικών χρηστών. Στην ανάλυση που έγινε φαίνεται ότι ο αριθμός των γειτονικών χρηστών όσο μεγαλώνει δεν αυξάνει την επίδοση του συστήματος.Τέλος στη συνδιαστική τεχνική χρησιμοποιούμε τη weight τεχνική αν η πιθανότητα P(u,i,c) είναι μεγαλύτερη από τη τιμή κατωφλίου t, ενώ αν είναι μικρότερη μειώνουμε την προβλεπόμενη αξιολόγηση του 2D συστήματος κατά Penalty. Στα πειράματα μας θέσαμε την τιμή Penalty στο 1.25. Η υπόθεση πίσω από τη τεχνική αυτή είναι να αξιοποιήσουμε και τις δύο τεχνικές με τη προσδοκία να έχουμε καλύτερα αποτελέσματα.

$$WeightedScore(u,i,c) = 2DRating * P(u,i,c) \tag{3.6}$$

$$FilteredScore(u,i,c) = \left\{ \begin{array}{c} \text{2DScore(u, i), if P(u,i,c) >t} \\ \text{0, if P(u, i, c) <t} \end{array} \right\} \tag{3.7}$$

$$CombineFilterWeight(u,i,c) = \left\{ \begin{array}{c} \text{2DScore(u, i) * P(u,i,c), if P(u,i,c) >t} \\ \text{2DScore(u, i) - Penalty, if P(u, i, c) <t} \end{array} \right\} \tag{3.8}$$

Για την αξιολόγηση της επίδοσης του συστήματος στα πειράματα μας χρησιμοποιήσαμε τις μετρήσεις (α) F1-Score και (β) Hit Ratio at 10 - HR@10. Η F1Score μέτρηση είναι μια περιγραφή των μετρήσεων της ακρίβειας και της ανάκλασης ενώ η μέτρηση HR μετρά τη σειρά κατάταξης των αποτελεσμάτων ενός αλγορίθμου συστάσεων και δείχνει πως ο αλγόριθμος συμπεριφέρεται λαμβάνοντας υπόψη μόνο τις 10 καλύτερες συστάσεις για κάθε χρήστη. Θα εξηγήσουμε με περισσότερη λεπτομέρεια τις παραπάνω μετρήσεις στο επόμενο κεφάλαιο.

3.3 Contextual modeling

Για την αξιολόγηση του συστήματος σε contextual modeling τεχνικές χρησιμοποιήσαμε τη βιβλιοθήκη CARSkit [39] για να πάρουμε μετρήσεις για την ακρίβεια πρόβλεψης. Το CARSkit είναι μια open source βιβλιοθήκη όπου περιέχει αρκετούς αλγορίθμους συστάσεων τόσο κλασσικούς 2D αλγορίθμους αλλά και CARS αλγορίθμους που αναφέρονται στη βιβλιογραφία. Συγκεκριμένα για το σύνολο δεδομένων θα συγκρίνουμε τους εξής αλγορίθμους:

- baseline Avg recommender
- baseline CF recommender - ItemKNN [32]
- BiasedMF [19]
- post-filter και post-weight

- CAMF [5]

Έχουμε αναλύσει τους παραπάνω αλγορίθμους στο κεφάλαιο 1 και 2. Οι baseline αλγόριθμοι χρησιμοποιούνται ως βάση για την επίδοση του συστήματος. Θεωρητικά αλλά και πρακτικά στις μετρήσεις όπως θα δούμε οι υπόλοιποι αλγορίθμοι θα πρέπει να έχουν καλύτερη επίδοση από τους baseline αλγορίθμους. Θα αξιολογήσουμε την επίδοση του κάθε αλγορίθμου χρησιμοποιοώντας μετρήσεις τις μετρήσεις RMSE, MAE και HR@10 όπως είδαμε στο κεφάλαιο 1

Κεφάλαιο 4

Πειραματική μελέτη

Στο κεφάλαιο αυτό περιγράφονται τα πειράματα και τα αποτελέσματα του συστήματος, με βάση τη μελέτη που παρουσιάστηκε στο προηγούμενο κεφάλαιο.

4.1 Dataset

Για την αξιολόγηση των αλγορίθμων συστάσεων χρησιμοποιήσαμε ένα σύνολο δεδομένων διαθέσιμο από τον Yong Zheng - DePaul University [40][39]. Τα αντικείμενα εκπαίδευσης για το σύστημα συστάσεων αποτελούνται από το χρήστη, την ταινία και πλαίσιο κάτω από το οποίο αξιολόγησε ο χρήστης την ταινία. Το σύνολο των δεδομένων αποτελείται από 5035 αντικείμενα (βαθμολογίες), 97 μοναδικούς χρήστες και η πυκνότητα των δεδομένων είναι 1,8230%. Το σύνολο των δεδομένων περιέχει τις αξιολογήσεις των χρηστών κάτω από τα αξής πλαίσια:

- Την **τοποθεσία** που είδε την ταινία ο χρήστης (Location).
- Την **ημέρα** που την είδε (Time).
- Η **παρέα** με την οποία είδε ο χρήστης την ταινία (Company).

Αριθμός χρηστών	97
Αριθμός αντικειμένων	79
Αριθμός αξιολογήσεων	5035
Διαστάσεις πλαισίων	3

Πίνακας 4.1: Σύνολο δεδομένων

Για κάθε πλαίσιο διακρίνουμε τις εξής διαστάσεις:

- Τοποθεσία
  - Σινεμά
  - Σπίτι

- Χρόνος
  - Σαββατοκύριακο
  - Καθημερινή
- Συντροφιά
  - Μόνος
  - Οικογένεια
  - Φίλος(οι)
  - Σύντροφος

Σημειώνεται ότι η παραπάνω πληροφορία δεν περιέχεται σε όλες τις αξιολογήσεις των χρηστών. Υπάρχουν δηλαδή αξιολογήσεις που δεν περιέχουν πληροφορία για το πλαίσιο παρά μόνο την βαθμολογία που έχει δώσει ο χρήστης.

4.2 Μετρήσεις αξιολόγησης

Στην ενότητα αυτή θα παρουσιάσουμε τα αποτελέσματα των πειράματων μας από την αξιολόγηση διαφορετικών αλγορίθμων πρόβλεψης.

4.2.1 Αξιολόγηση post-filtering τεχνικών

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα για την επίδοση του συστήματος χρησιμοποιώντας post weighted και post-filtering τεχνικές. Συγκεκριμένα χρησιμοποιούμε τις μετρήσεις precision και recall για να αξιολογήσουμε τα συστήματα συστάσεων. Για να χρησιμοποιηθούν τα μέτρα αυτά, το σύστημα συστάσεων θα πρέπει να μετατρέψει τις αξιολογήσεις σε μια δυαδική κλίμακα (Συστήνεται / Δεν συστήνεται). Κάθε αντικείμενο λοιπόν μπορεί να είναι σχετικό για ένα χρήστη ή αδιάφορο. Συνεπώς μπορούμε να πάρουμε τον πίνακα 4.2.

	Συστήνεται	Δεν συστήνετα
Σχετικό	True-Positive (tp)	False-Negative(fn)
Αδιάφορο	False-Positive (fp)	True-Negative(tn)

Πίνακας 4.2: Πιθανά αποτελέσματα ταξινόμησης για ένα συστηνόμενο αντικείμενο σε έναν χρήστη [29]

Για την αξιολόγηση και τον υπολογισμό των μέτρων ακρίβειας και ανάκλασης μπορούμε να μετρήσουμε πόσα παραδείγματα πέφτουν σε μια από τις κατηγορίες του πίνακα 4.2 και να υπολογίσουμε τις παρακάτω ποσότητες.

$$Precision = \frac{tp}{tp + fp} \tag{4.1}$$

$$Recall = \frac{tp}{fp + tn} \tag{4.2}$$

Από τα παραπάνω περιμένουμε μια αντιστρόφως ανάλογη σχέση μεταξύ της ανάκλασης και της ακρίβειας. Όσο μεγαλύτερη είναι η λίστα των συστάσεων τόσο μεγαλύτερη είναι η ανάκλαση αλλά παράλληλα μειώνεται η ακρίβεια του συστήματος. Επομένως σε εφαρμογές που οι συστάσεις για τον χρήστη είναι προ-ορισμένες το πιο χρήσιμο μέτρο που μας ενδιαφέρει είναι η ακρίβεια στις N συστάσεις. Αντίθετα σε εφαρμογές που οι συστάσεις για τον χρήστη δεν είναι προκαθορισμένες είναι πιο αξιόπιστο να αξιολογήσουμε την επίδοση του συστήματος σε διαφορετικά μεγέθη της λίστας συστάσεων από το να έχουμε μια προκαθορισμένη λίστα συστάσεων μεγέθους N. Επομένως μπορούμε να υπολογίζουμε καμπύλες που συγκρίνουν την ακρίβεια με την ανάκλαση ή την αναλογία των true positive προς false positive (precision-recall curves και ROC curves).

Το μέτρο *precision* 4.3, βρίσκει την αναλογία των σχετικών ταινιών από τη λίστα (ταινία, βαθμολογία) των συστάσεων. Εδώ μια ταινία είναι σχετική αν υπάρχει στο test σύνολο δεδομένων και έχει μια αξιολόγηση παρόμια με αυτή που προβλέπει το σύστημα συστάσεων. Με άλλα λόγια μια ταινία είναι σχετική για έναν χρήστη αν το σύστημα συστάσεων την έχει προτείνει στο χρήστη και η ταινία είναι στο τεστ σύνολο δεδομένων. Η μέτρηση αυτή μετρά την ικανότητα του αλγορίθμου να συστήνει ταινίες στο χρήστη που τον ενδιαφέρουν αντί για συστάσεις που δεν τον ενδιαφέρουν. Μετρά δηλαδή το ποσοστό των συστάσεων που είναι καλές συστάσεις για έναν χρήστη.

$$precision(N) = \frac{\{(u,i) \in testSet\} \cap \{(u,i) \in recommendationSet\} \ \forall \ r(u,i) \ \geq 3.5}{\{(u,i) \in recommendationSet\}} \tag{4.3}$$

Το μέτρο *recall* 4.4 μετρά την αναλογία των καλών ταινιών που συστήθηκαν από το σύνολο δεδομένων προς όλες τις πιθανές καλές ταινίες. Εδώ μια καλή ταινία είναι εκείνη η οποία της δόθηκε μια υψηλή βαθμολογία από το χρήστη υπό εξέταση. Η μέτρηση recall ορίζεται ως ο λόγος μεταξύ των σωστών προβλέψεων προς το σύνολο των τεστ δεδομένων. Η μέτρηση recall μετρά την πιθανότητα ενός αντικειμένου που ενδιαφέρει τον χρήστη, να έχει πράγματι συστηθεί στο χρήστη. Μετρά δηλαδή την ικανότητα του αλγορίθμου να καλύψει τις προτιμήσεις των χρηστών ή το ποσοστό των καλών ταινιών που εμφανίζονται στις καλύτερες N συστάσεις.

$$recall(N) = \frac{\{(u,i) \in testSet\} \cap \{(u,i) \in recommendationSet\} \ \forall \ r(u,i) \ \geq 3.5}{\{(u,i) \in testSet\}} \tag{4.4}$$

Η F1score μέτρηση, είναι ένα άλλο μέτρο για την αξιολόγηση της ακρίβειας ενός συστήματος συστάσεων όπου λαμβάνεται υπόψη το μέτρο precision και recall για να υπολογιστεί το σκορ. Στα συστήματα συστάσεων, θεωρείται μια ενιαία τιμή και δείχνει τη συνολική χρησιμότητα της λίστας των συστάσεων. Η F1Score μέτρηση ορίζεται ως ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλασης και χρησιμοποιείται ως μέτρηση που παρέχει μια σύνοψη της ακρίβειας και της ανάκλασης. Η μέτρηση F1Score δίνεται από την σχήση 4.5

$$F1Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4.5}$$

4.2.2 Μετρήσεις Κατάταξης (Ranking Measures)

Για να αξιολογήσουμε τη λίστα κατάταξης του συστήματος συστάσεων χρησιμοποιήσαμε το μέτρο Hit Ratio (HR)

Η μέτρηση HR υπολογίζει το ποσοστό των αντικειμένων που συστήνουμε και έχουμε σωστές συστάσεις (correct hits). Η μέτρηση HR είναι παρόμια με τον υπολογισμό της ανάκλασης για κάθε χρήστη όταν προτείνουμε N συστάσεις. Η μέτρηση HR δείχνει την επίδοση του αλγορίθμου στις υψηλότερες συστάσεις για κάθε χρήστη. Επομένως, είναι μια σημαντική μέτρηση, αφού είναι απίθανο οι χρήστες να ψάξουν πολύ βαθιά στη λίστα με τις συστάσεις. Ο ορισμός της HR μέτρησης δίνεται από την εξίσωση 4.6

$$HR(N) = \frac{\{(u,i) \in testSet\} \cap \{(u,i) \in topNRecommendationSet\}}{\{(u,i) \in testSet\}} \tag{4.6}$$

4.3 Πειραματικά αποτελέσματα αξιολόγησης

Hit-Rate@10 Metrics				
	Baseline	Weight	Filter	Combined
Time	0.29	0.39	0.34	0.37
Location	0.29	0.38	0.33	0.39
Companion	0.29	0.36	0.33	0.37

Πίνακας 4.3: Μετρήσεις HR@10

Στην ενότητα αυτή θα παρουσιάσουμε τα αποτελέσματα μας από την αξιολόγηση των συστημάτων συστάσεων. Οι μετρήσεις που παρουσιάσαμε στην προηγούμενη ενότητα απεικονίζονται στα γραφήματα που ακολουθούν. Η αξιολόγηση για το σφάλμα πρόβλεψης έγινε με βάση τη *σταυρωτή αξιολόγηση (cross validation)*. Συγκεκριμένα χρησιμοποιώντας τη k-fold τεχνική. Η κ-πλή (k-fold) σταυρωτή αξιολόγηση διαμερίζει (τυχαία) το σύνολο δεδομένων σε k υποσύνολα περίπου ίσης πληθικότητας το καθένα. Από τα προαναφερθέντα k υποσύνολα, ένα χρησιμοποιείται ως υποσύνολο εξέτασης, ενώ η συνολοθεωρητική ένωση των υπόλοιπων κ-1 υποσυνόλων χρησιμοποιείται ως υποσύνολο εκπαίδευσης. Συνολικά εκτελούνται κ υπολογιστικοί κύκλοι, έτσι ώστε με τη σειρά κάθε ένα από τα κ υποσύνολα να χρησιμοποιείται ως υποσύνολο εξέτασης. Το πλεονέκτημα αυτής της τεχνικής αξιολόγησης είναι ότι κάθε δεδομένο εγγυημένα χρησιμοποιείται τόσο για εκπαίδευση, όσο και για εξέταση. Μάλιστα για εξέταση χρησιμοποιείται ακριβώς μια φορά. Η παράμετρος κ μπορεί να λάβει οποιαδήποτε (θετική) ακέραια τιμή. Στα πειράματα μας θέσαμε k=5. Για την αξιολόγηση της επίδοσης στις topN συστάσεις θέσαμε τη τιμή ίση με 10 (προτείνουμε 10 ταινίες σε κάθε χρήστη).

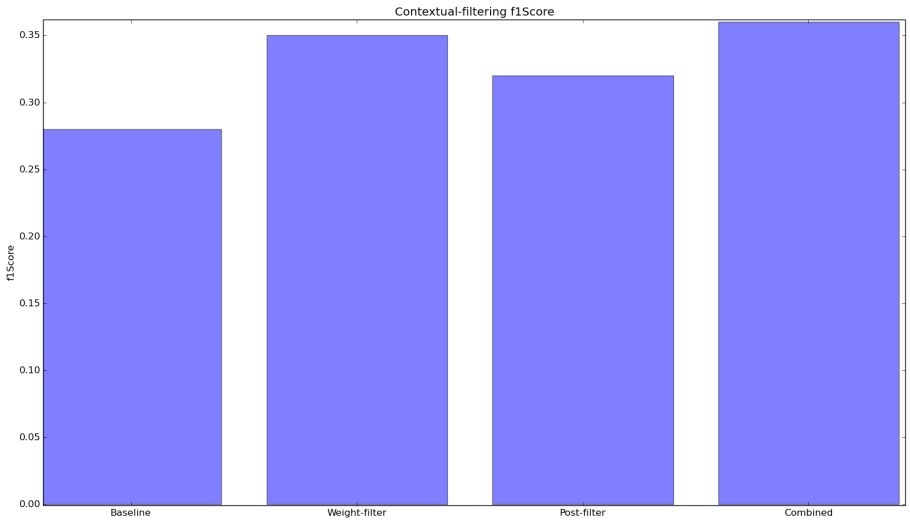
	RMSE
ItemKNN	1.088467
Biased-MF	1.007075
CAMF_C	0.899010

Πίνακας 4.4: Μετρήσεις RMSE μεταξύ διαφορετικών αλγορίθμων

Στα πειράματα μας μετρήσαμε την επίδοση του συστήματος χωρίς να λάβουμε υπόψιν το πλαίσιο και στη συνέχεια μετρήσαμε την επίδοση του συστήματος λαμβάνοντας υπόψιν ένα από τα παρακάτω:

- Το πλαίσιο *Time* κάτω από το οποίο έχουμε δύο διαστάσεις: (α) *working day* και (β) *weekend*.
- Το πλαίσιο *Location* κάτω από το οποίο υπάρχουν δύο διαστάσεις: (α) *Home* και (β) *Cinema*.
- Το πλαίσιο *Company* που περιέχει 4 διαστάσεις: (α) *Alone*, (β) *Family*, (γ) *Friends* και (δ) *Partner*.

Στο πίνακα 4.4 φαίνονται οι RMSE μετρήσεις από τα πειράματα μας που φαίνεται πως ο αλγόριθμος CAMF_C [5] έχει καλύτερη επίδοση σε μετρήσεις πρόβλεψης.



Σχήμα 4.1: F1Score μετρήσεις για κάθε αλγόριθμο στις top10 συστάσεις

Στο σχήμα 4.1 απεικονίζονται οι f1Score μετρήσεις για κάθε αλγόριθμο. Παρατηρούμε ότι όλες οι contextual τεχνικές έχουν καλύτερη επίδοση από το βασικό itemKNN αλγόριθμο. Από τις contextual τεχνικές ο combined αλγόριθμος που είναι ένας συνδιασμός του post-weight και post-filter έχει λίγο καλύτερη απόδοση από τον post-weighted αλγόριθμο ο οποίος

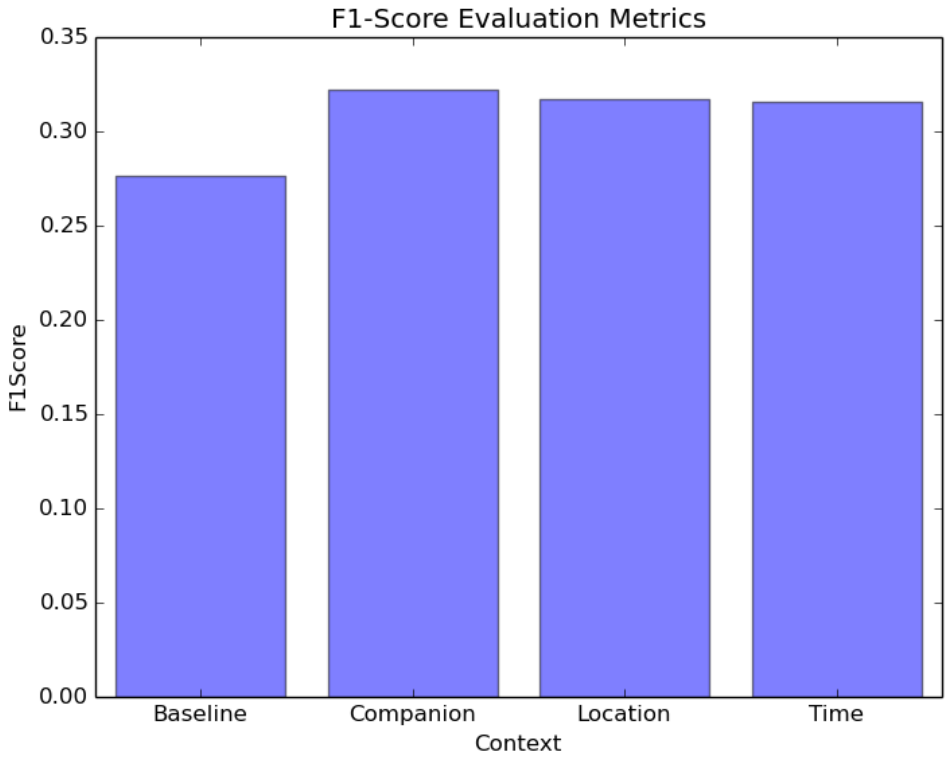
post-filter	post-weight	post-combined
16,36%	27,27%	31.2%

Πίνακας 4.5: Ποσοστό βελτίωσης F1Score από τον kNN αλγόριθμο χρησιμοποιώντας το context

έχει την επόμενη καλύτερη επίδοση σε σχέση με τις υπόλοιπες δύο τεχνικές. Στο πίνακα 4.5 παρουσιάζεται το ποσοστό βελτίωσης που έχουμε για κάθε contextual αλγόριθμο. Στις επόμενες ενότητες θα παρουσιάσουμε τις μετρήσεις που πείραμε για κάθε πλαίσιο χωριστά.

4.3.1 Contextual post-Filter

Στη τεχνική αυτή χρησιμοποιήσαμε την εξίσωση 3.7 που περιγράψαμε στο κεφάλαιο 3 για να φιλτράρουμε τα αποτελέσματα που παράγει το κλασσικό 2D σύστημα συστάσεων. Τα αποτελέσματα φιλτράρονται με βάση τη πιθανότητα P(u,i,c) όπου στα πειράματα μας τη θέσαμε 0.1. Παρακάτω ακολουθούν τα γραφήματα των μετρήσεων μας για τη filter τεχνική.

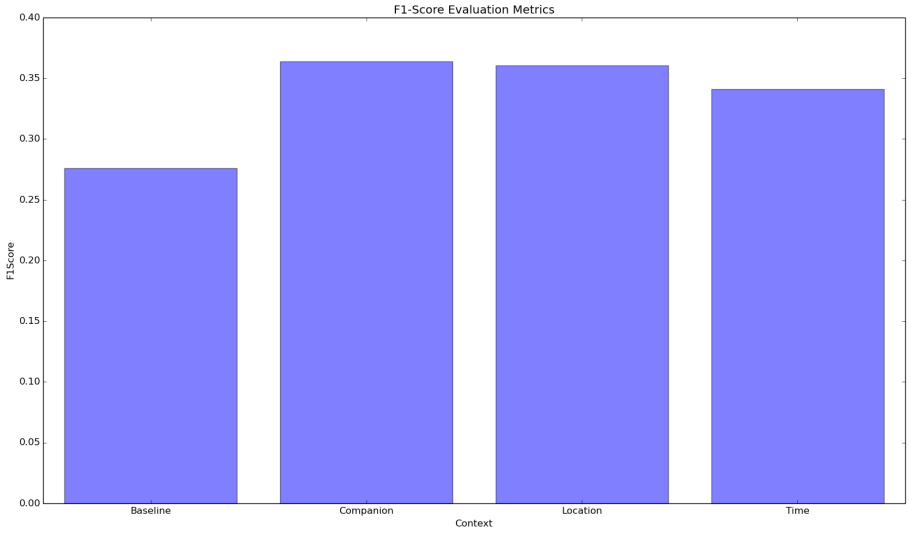


Σχήμα 4.2: Contextual post-filter: F1Score στις top10 συστάσεις

4.3.2 Contextual post-weight

Στη τεχνική αυτή χρησιμοποιήσαμε την εξίσωση 3.6 που περιγράψαμε στο κεφάλαιο 3 όπου οι προβλέψεις από το κλασσικό σύστημα συστάσεων πολλαπλασιάζονται με τη πιθανότητα

P(u,i,c), τη πιθανότητα ένας χρήστης u να ενδιαφέρεται για ένα αντικείμενο i στο πλαίσιο c και στη συνέχεια επαναδιατάσσουμε τη λίστα με τις προτεινόμενες συστάσεις. Τα αποτελέσματα της weight τεχνικής φαίνονται στις εικόνες 4.8, 4.9 και 4.10.



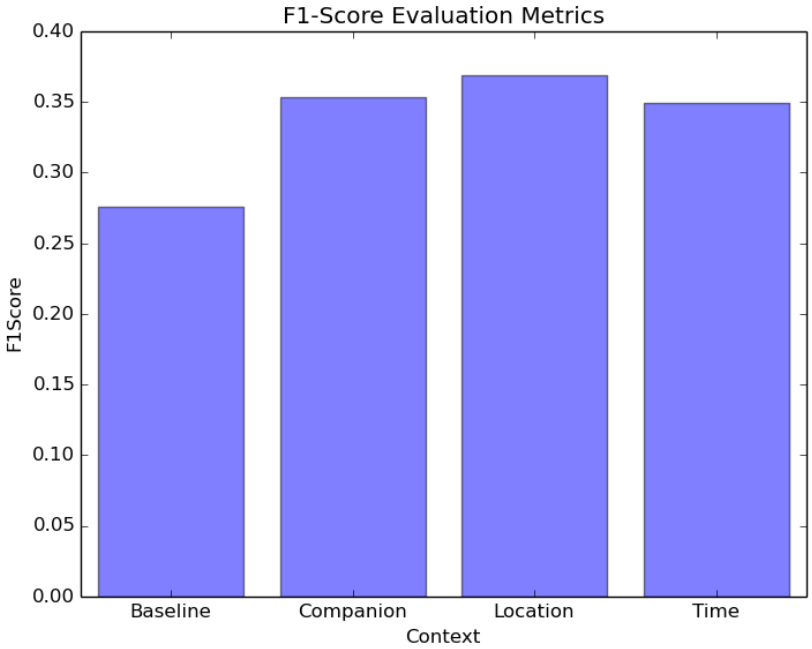
Σχήμα 4.3: Contextual post-weight: F1Score στις top10 συστάσεις

4.3.3 Συνδιασμός Contextual post-filter με post-Weight

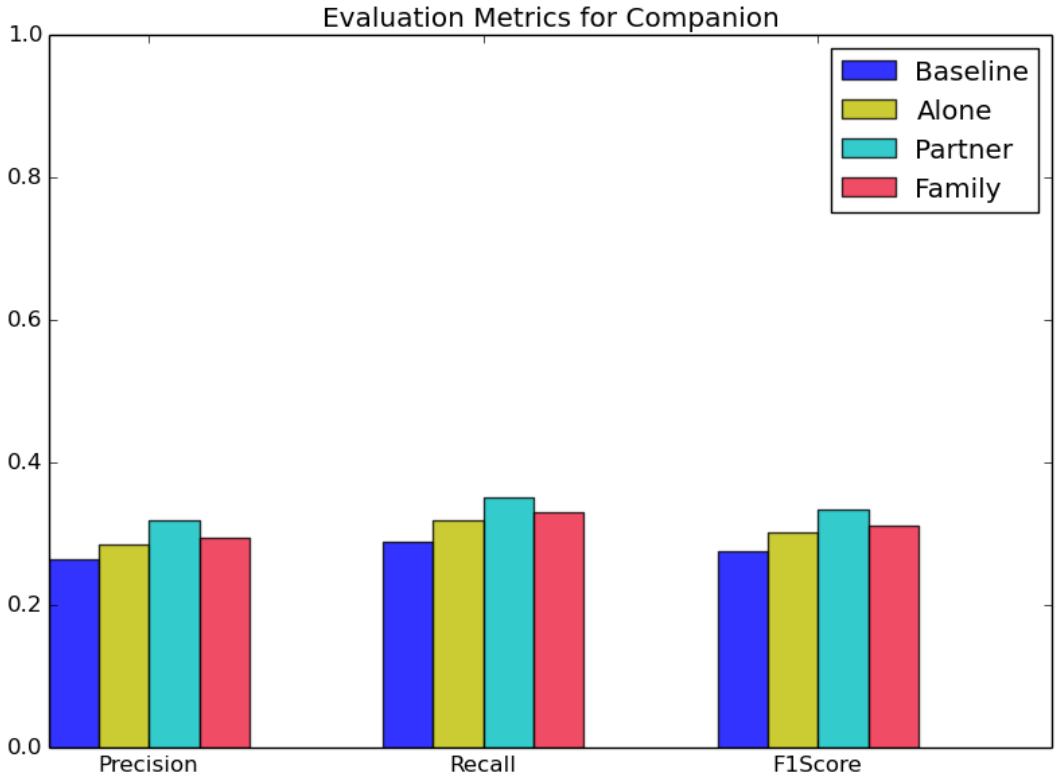
Με τη τεχνική αυτή συνδιάζουμε τη filter και weight τεχνική σε μια υβριδική συνάρτηση, εξίσωση 3.8. Συγκεκριμένα αν η πιθανότητα P(u,i,c) είναι μεγαλύτερη από μια τιμή κατωφλίου (στα πειράματα μας 0.1) τότε η νέα προβλεπόμενη αξιολόγηση του χρήστη θα είναι ίση με τη weight τεχνική (η προβλεπόμενη αξιολόγηση του χρήστη από το 2D Αλγόριθμο πολλαπλασιασμένη με τη πιθανότητα P(u,i,c)). Αν η πιθανότητα P(u,i,c) είναι μικρότερη από τη τιμή κατωφλίου τότε μειώνουμε την προβλεπόμενη αξιολόγηση του 2D συστήματος κατά μια τιμή Penalty. Στα πειράματα μας θέσαμε τη τιμή Penalty στη τιμή 0.5 χωρίς να επεκτείνουμε την έρευνα μας σε πιο εξειδικευμένες τεχνικές. Με τη μέθοδο αυτή αξιοποιούμε και τις δύο τεχνικές: είτε πολλαπλασιάζοντας με τη πιθανότητα για να βάλουμε βάρη είτε φιλτράρουμε τα αποτελέσματα μειώνοντας την αξιολόγηση αν η πιθανότητα είναι μικρότερη από τη τιμή κατωφλίου.

Τα αποτελέσματα φαίνονται στις εικόνες 4.11, 4.12 και 4.13. Συνδιάζοντας τις δύο τεχνικές έχουμε καλύτερη απόδοση του συστήματος.

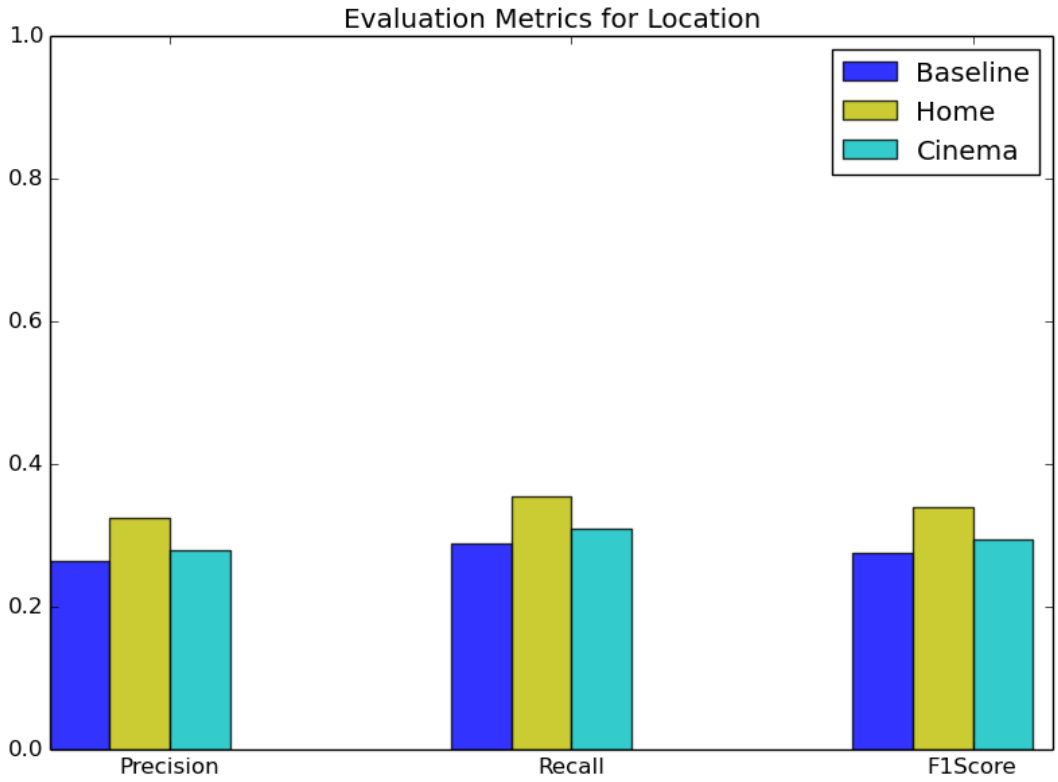
Συνοψίζοντας, από τις μετρήσεις που πείραμε φαίνεται πως το πλαίσιο επηρεάζει θετικά την βελτιστοποίηση ενός συστήματος συστάσεων. Συγκεκριμένα και οι τρεις τεχνικές που χρησιμοποιήσαμε έδειξαν να υπερισχύουν από το βασικό αλγόριθμο χωρίς πλαίσιο. Από τη σύγκριση των αλγορίθμων φαίνεται ότι η combined τεχνική υπερισχύει με βάση τις f1Score μετρήσεις.



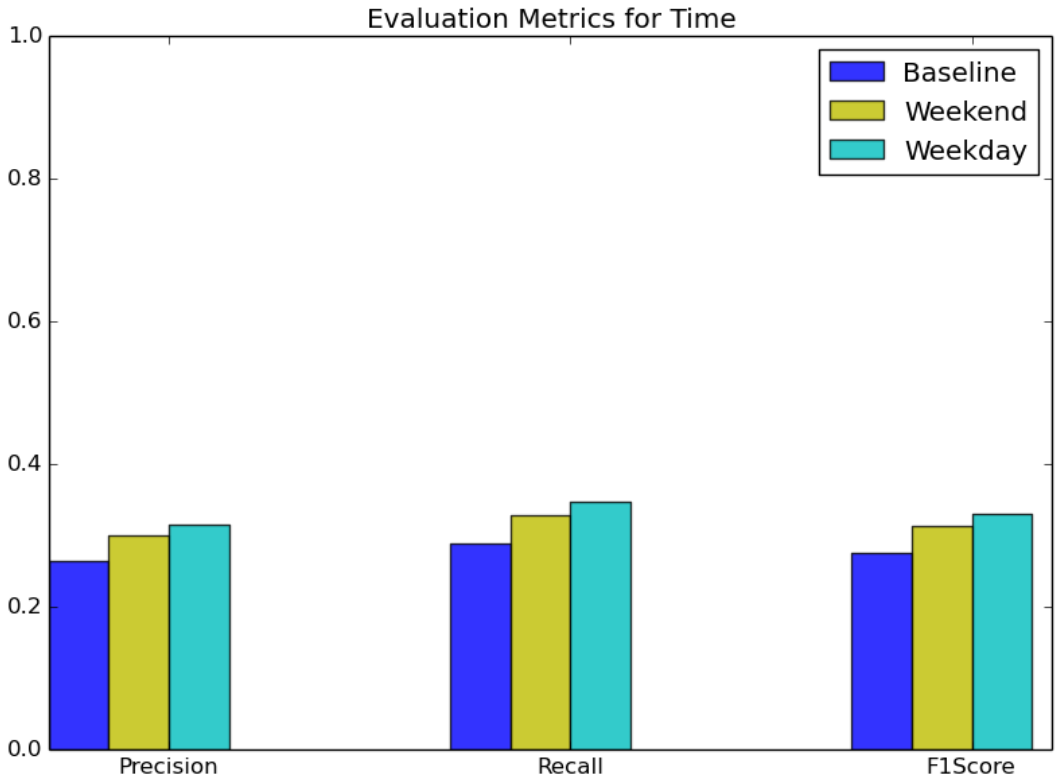
Σχήμα 4.4: Contextual post-Combined: F1Score στις top10 συστάσεις



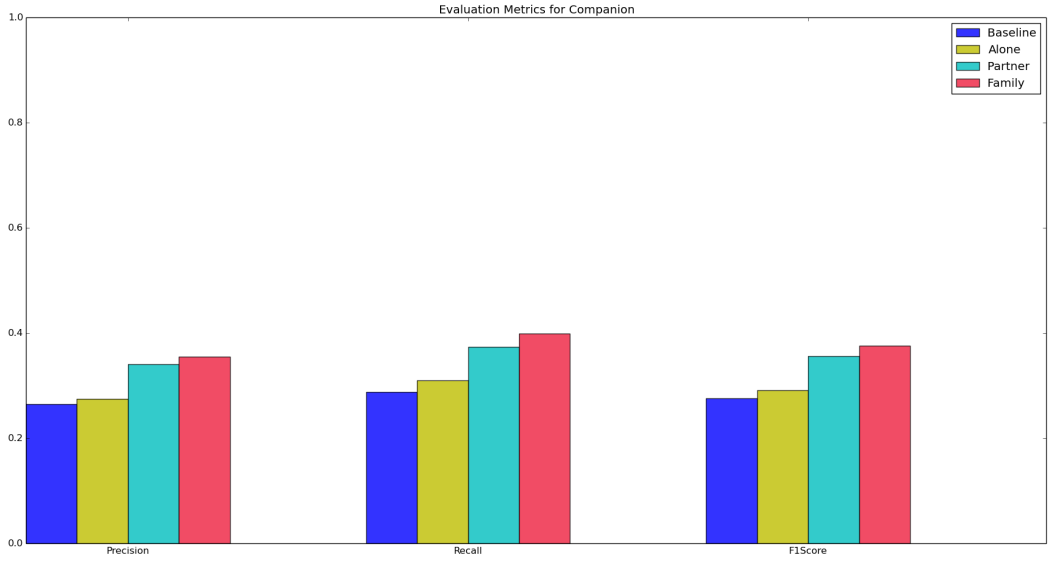
Σχήμα 4.5: Contextual post-filter: Μετρήσεις ακρίβειας για το πλαίσιο Companion στις top10 συστάσεις



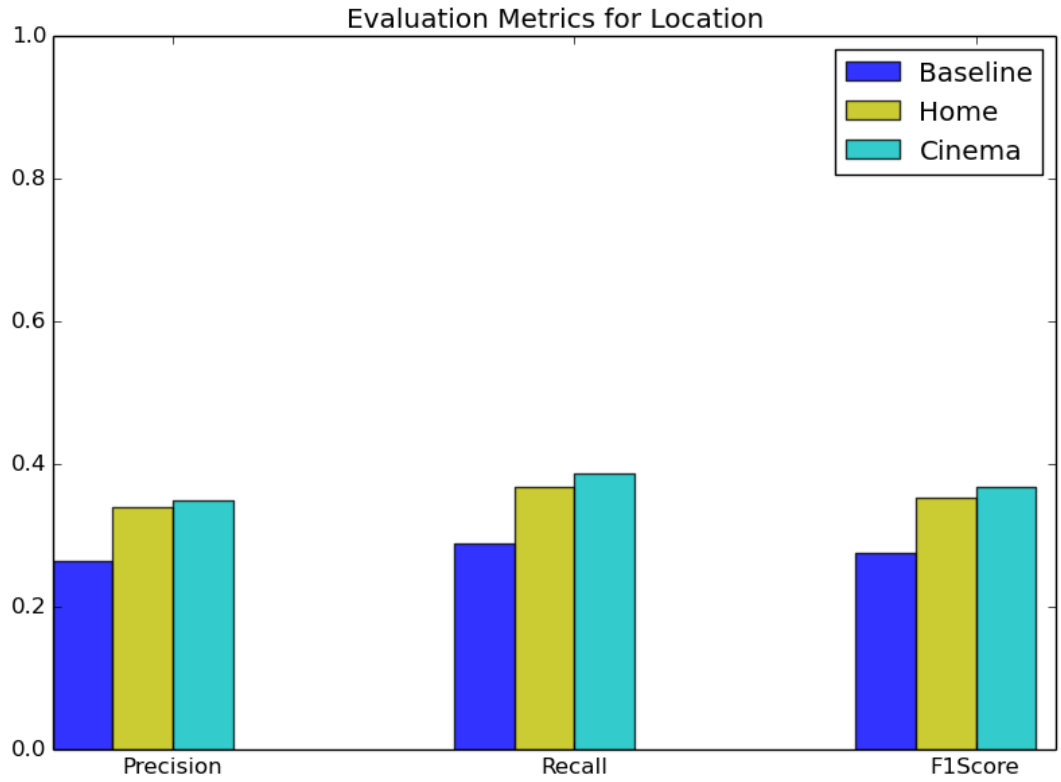
Σχήμα 4.6: Contextual post-filter: Μετρήσεις ακρίβειας για το πλαίσιο Location στις top10 συστάσεις



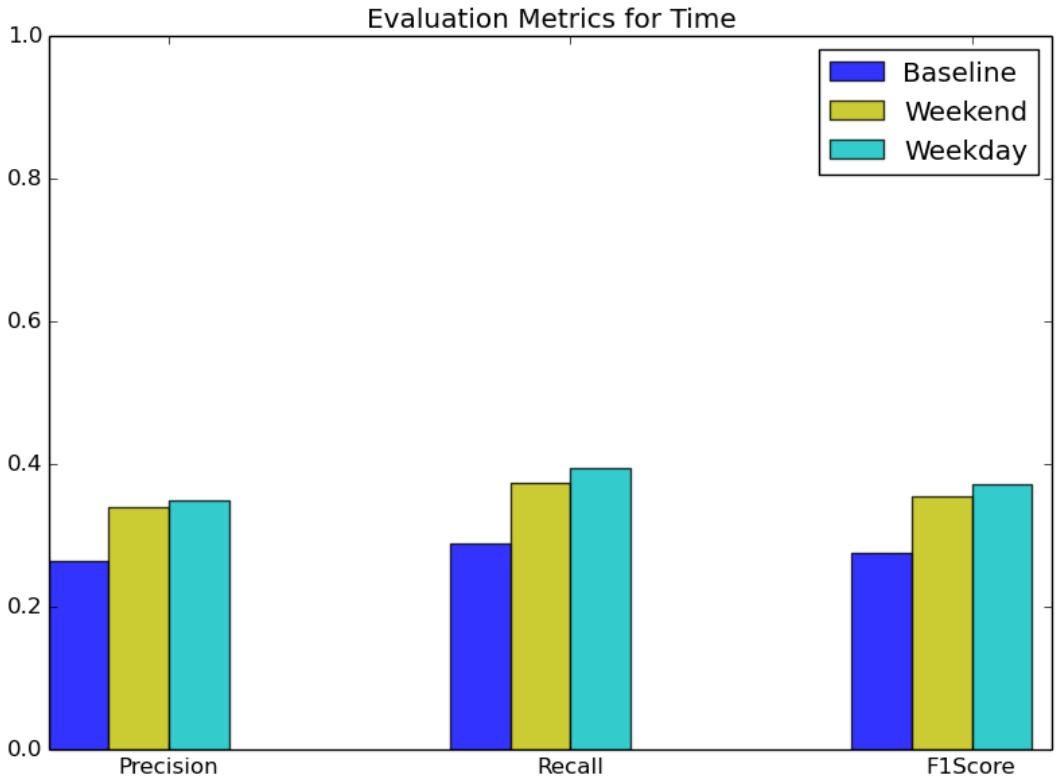
Σχήμα 4.7: Contextual post-filter: Μετρήσεις ακρίβειας για το πλαίσιο Time στις top10 συστάσεις



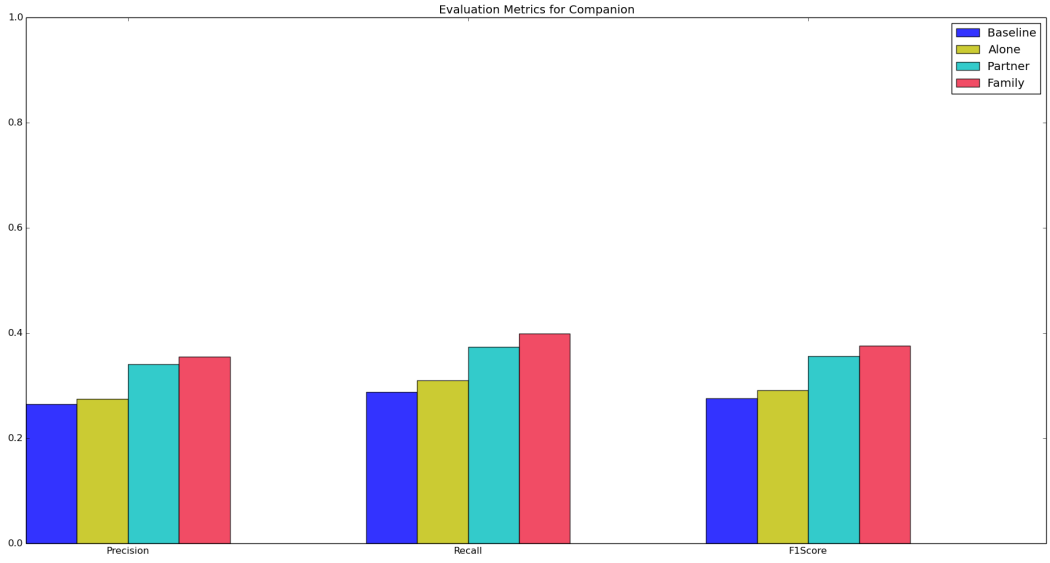
Σχήμα 4.8: Contextual post-weight Μετρήσεις ακρίβειας για το πλαίσιο Companion στις top10 συστάσεις



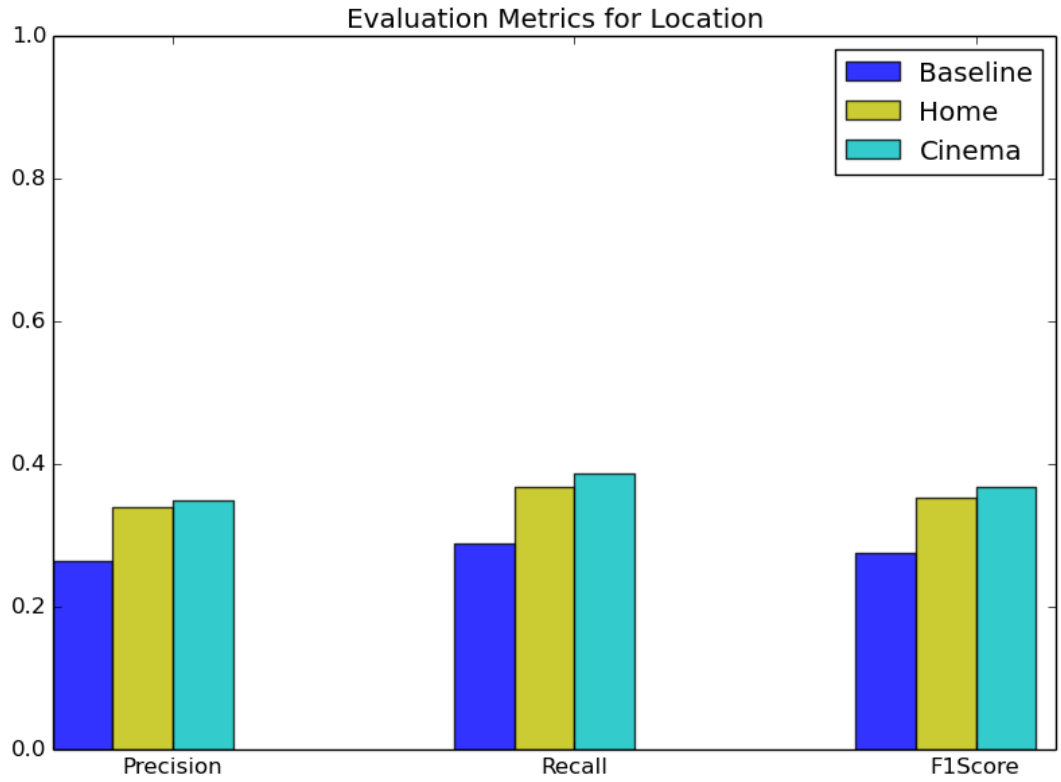
Σχήμα 4.9: Contextual post-weight Μετρήσεις ακρίβειας για το πλαίσιο Location στις top10 συστάσεις



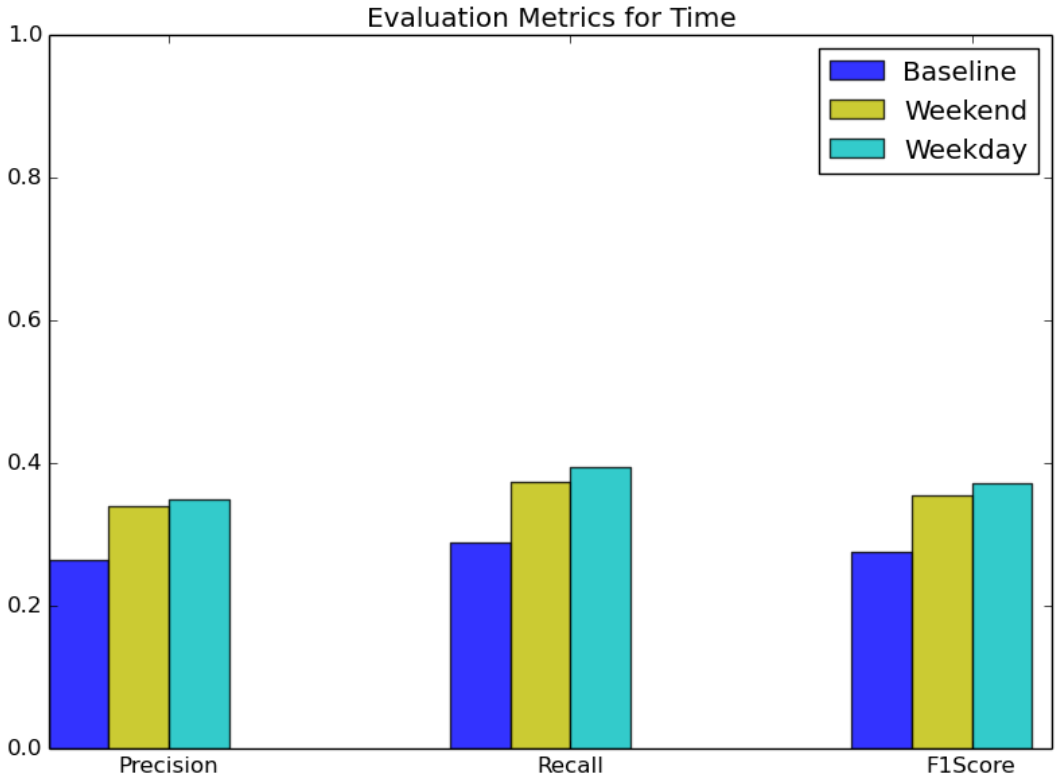
Σχήμα 4.10: Contextual post-weight Μετρήσεις ακρίβειας για το πλαίσιο Time στις top10 συστάσεις



Σχήμα 4.11: Contextual post-Combined μετρήσεις για το πλαίσιο Companion στις top10 συστάσεις



Σχήμα 4.12: Contextual post-Combined μετρήσεις για το πλαίσιο Location στις top10 συστάσεις



Σχήμα 4.13: Contextual post-Combined μετρήσεις για το πλαίσιο Time στις top10 συστάσεις

Κεφάλαιο 5

Συμπεράσματα

Τα συστήματα συστάσεων έχουν γίνει πολύ διαδεδομένα τα τελευταία χρόνια σε πολλές online εφαρμογές και έχει αναγνωριστεί η σημασία τους στην ικανοποίηση των χρηστών και στην αύξηση των πωλήσεων. Παρά την αύξηση του ενδιαφέροντος του τομέα στον ερευνητικό χώρο που πυροδοτήθηκε από το βραβείο Netflix τα context-aware συστήματα αποτελούν ακόμα ένα πεδίο που δεν έχει διερευνηθεί σε βάθος εκτός από τα τελευταία χρόνια που γίνονται όλο και περισσότερες μελέτες στο χώρο αυτό. Στόχος της διπλωματικής αυτής εργασίας ήταν η διερεύνηση της απόδοσης των context-aware συστημάτων συστάσεων σε ένα σύνολο δεδομένων από αξιολογήσεις πλούσιο σε πληροφορία που περιέχει το πλαίσιο κάτω από το οποίο αλληλεπίδρασε ο χρήστης με το σύστημα χρησιμοποιώντας είδη υπάρχουσες τεχνικές από συστήματα βασισμένα στη συνεργασία.

Το σύστημα συστάσεων που αναπτύξαμε βασισμένο σε προηγούμενες μελέτες που έχουν γίνει αποδεικνύει ότι η πληροφορία του πλαισίου είναι πολύ σημαντική για να παρέχουμε καλύτερες συστάσεις. Από τις τιμές ακρίβειας και της ανάκλησης αναγνωρίσαμε τη σημασία του πλαισίου για την επιλογή ταινιών συγκρίνοντας ένα κλασσικό σύστημα συστάσεων με ένα σύστημα συστάσεων που χρησιμοποιεί το πλαίσιο για την ενίσχυση του συστήματος.

5.1 Μελλοντικές Εργασίες

Στη παρούσα διπλωματική ερευνήσαμε διάφορες τεχνικές που χρησιμοποιούνται στα CARS συστημάτων συστάσεων αλλά υπάρχουν πολλές πτυχές που δεν έχουμε διερευνήσει. Εκτός από επιπλέον πειράματα σε παρόμιες τεχνικές, η εξερεύνηση άλλων context-aware τεχνικών παρουσιάζει μεγάλο ενδιαφέρον. Συγκεκριμένα δοκιμάζοντας την επίδοση του συστήματος με αλγορίθμους μοντελοποίησης του πλαισίου (contextual modeling) και την επεκτασιμότητα των αλγορίθμων σε εφαρμογές που περιέχουν εκατομμύρια χρήστες και αντικείμενα. Μεγάλο ενδιαφέρον παρουσιάζουν οι CAMF τεχνικές [5] όπως αναλύσαμε σε προηγούμενο κεφάλαιο που εισάγουν τη πληροφορία του πλαισίου στον αλγόριθμο σύστασης χρησιμοποιώντας Matrix Factorization τεχνικές. Οι CAMF τεχνικές είναι περισσότερο επεκτάσιμες από τις kNN τεχνικές αφού μπορούν να χειριστούν μεγάλο όγκο δεδομένων σε μικρότερο χρόνο. Μια πιθανόν επέκταση των τεχνικών αυτών σε μελλοντικές εργασίες είναι να μοντελοποιήσουμε

τη πληροφορία του πλαισίου χρησιμοποιώντας semantic analysis τεχνικές ή εκμεταλέυοντας τη σημασιολογική κατανομή των χαρακτηριστικών των αντικειμένων [11]. Επιπλέον με τα συστήματα CARS δημιουργείται η ανάγκη για καινούργιες εφαρμογές όπως εφαρμογές που μπορούν πλέον να προτείνουν το πλαίσιο αντί για το αντικείμενο.

Ένα πρόβλημα στα CARS συστήματα είναι ότι τα συμφραζόμενα μπορεί να είναι πλήρως παρατηρήσιμα, εν μέρει παρατηρήσιμα ή μη παρατηρήσιμα. Αν οι παράγοντες μπορούν να παρατηρηθούν, η μοντελοποίηση του πλαισίου δεν είναι δύσκολη. Ωστόσο, σε αντίθετη περίπτωση αυτό μπορεί να λυθεί με την ανίχνευση του θέματος - topic modeling. Η υπόθεση είναι ότι το πλαίσιο επηρεάζει τις προτιμήσεις των χρηστών και ως εκ τούτου αντανακλάται στη συμπεριφορά τους. Αν αυτό αληθεύει, τότε το πλαίσιο δεν χρειάζεται να μοντελοποιηθεί άμεσα, αλλά μπορεί να να αποτυπωθεί έμεσα με θέματα (topics) [14].

Τέλος μελλοντικές επεκτάσεις είναι να αναπτυχθούν CARS αλγόριθμοι συστάσεων σε machine learning βιβλιοθήκες όπως είναι η βιβλιοθήκη MLlib (Apache Spark) [38] ή σε πιο αφηρημένες βιβλιοθήκες που χρησιμοποιούνται για να αναπτυχθούν production ready αλγόριθμοι μηχανικής μάθησης όπως είναι η βιβλιοθήκη PredictionIO που πρόσφατα έχει κάνει αίτηση για να ενταχθεί στην οικογένεια της Apache Foundation.

Παράρτημα Α΄

Πηγαίος κώδικας και εκτέλεση προγράμματος

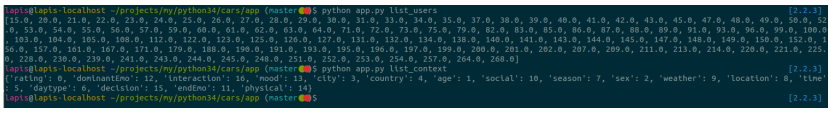
Στα πλαίσια της εργασίας αναπτύξαμε ένα σύστημα συστάσεων για να αξιολογήσουμε ένα κλασσικό σύστημα συστάσεων με ένα σύστημα συστάσεων βασιζόμενο στα συμφραζόμενα κάτω από τα οποία δόθηκαν οι αξιολογήσεις. Το πρόγραμμα αναπτύχθηκε με το γλώσσα προγραμματισμού Python (version 2.7.6).

Η εκτέλεση του προγράμματος μπορεί να γίνει είτε (α) για να πάρουμε συστάσεις για έναν χρήστη είτε (β) για να πάρουμε μετρήσεις για το σύστημα συστάσεων.

Για να εκτελέσουμε το πρόγραμμα θα πρέπει να είμαστε στο κατάλογο project/cars/app. Για να δούμε τις διαθέσιμες επιλογές τρέχουμε το αρχείο app.py όπως φαίνεται στο σχήμα Α΄.1 με την εντολή *python app.py*

Σχήμα Α΄.1: Εκτέλεση προγράμματος για τη λήψη οδηγιών.

Για να πάρουμε μια λίστα με όλους τους διαθέσιμους χρήστες και τα διαθέσιμα πλαίσια εκτελούμε την εντολή python app.py list_users και python app.py list_context αντίστοιχα όπως φαίνεται στο σχήμα Α΄.2.



Σχήμα Α΄.2: Εκτέλεση προγράμματος για την εμφάνιση όλων των διαθέσιμων χρηστών.

Στη συνέχεια μπορούμε να ζητήσουμε συστάσεις για έναν χρήστη u κάτω από μια συνθήκη c, i με την εντολή python app.py recommend userID context condition όπως φαίνεται στο

σξήμα Α΄.3

```
lapis@lapis-localhost ~/projects/my/python34/cars/app (master)$ python app.py recommend 1123 Time Weekday
Predicted movies for user:  1123
Movie           Prediction
****************************
tt0211915               6.42478354978
tt0110357               4.94739070567
tt1041829               4.92151690819
tt0138097               4.91589445416
tt0268380               4.56629630398
tt0407304               4.50933980041
tt0454876               4.5
tt1632708               4.41960784314
tt4411490               4.39951262596
lapis@lapis-localhost ~/projects/my/python34/cars/app (master)$ python app.py recommend 1123 Location Home
Predicted movies for user:  1123
Movie           Prediction
****************************
tt0268380               9.13259260797
tt0407304               9.01867960082
tt0111161               8.66392358549
tt0319262               8.65490369958
tt0266543               8.33430473688
tt1041829               4.92151690819
tt0138097               4.91589445416
tt0454876               4.5
tt1632708               4.41960784314
lapis@lapis-localhost ~/projects/my/python34/cars/app (master)$ python app.py recommend 1123 Companion Partner
Predicted movies for user:  1123
Movie           Prediction
****************************
tt0268380               6.62112964078
tt0407304               6.5385427106
tt0120338               6.51523267541
tt0211915               6.42478354978
tt0111161               6.28134459948
tt0319262               6.2748051822
tt0266543               6.04237093424
tt1041829               4.92151690819
tt0138097               4.91589445416
lapis@lapis-localhost ~/projects/my/python34/cars/app (master)$ python app.py recommend 1123 Companion Alone
Predicted movies for user:  1123
Movie           Prediction
****************************
tt0120338               7.14573906335
tt0454876               6.525
tt0266543               6.04237093424
tt1375666               5.91133695208
tt1232829               5.1386479816
tt1707386               5.08388169459
tt1041829               4.92151690819
tt0138097               4.91589445416
tt0268380               4.56629630398
lapis@lapis-localhost ~/projects/my/python34/cars/app (master)$
```

Σχήμα Α΄.3: Εκτέλεση προγράμματος για τη σύσταση ταινιών.

Βιβλιογραφία

[1] Gediminas Adomavicius και Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 17(6):734–749, 2005.

[2] Gediminas Adomavicius και Alexander Tuzhilin. *Context-Aware Recommender Systems*, σελίδες 217–253. Springer US, Boston, MA, 2011.

[3] Linas Baltrunas και Xavier Amatriain. Towards time-dependant recommendation based on implicit feedback. Στο *In Workshop on context-aware recommender systems (CARS*, 2009.

[4] Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Aykan Aydin, Karl Heinz Lüke και Roland Schwaiger. *InCarMusic: Context-Aware Music Recommendations in a Car*, σελίδες 89–100. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[5] Linas Baltrunas, Bernd Ludwig και Francesco Ricci. Matrix factorization techniques for context aware recommendation. Στο *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, σελίδες 301–304, New York, NY, USA, 2011. ACM.

[6] Linas Baltrunas και Francesco Ricci. Experimental evaluation of context-dependent collaborative filtering using item splitting. *User Modeling and User-Adapted Interaction*, 24(1-2):7–34, 2014.

[7] David M. Blei, Andrew Y. Ng και Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[8] Robin Burke. Hybrid web recommender systems. Στο *The Adaptive Web*Peter Brusilovsky, Alfred Kobsa και Wolfgang Nejdl, επιμελητές, σελίδες 377–408. Springer-Verlag, Berlin, Heidelberg, 2007.

[9] Pedro G. Campos, Ignacio Fernández-Tobías, Iván Cantador και Fernando Díez. *Context-Aware Movie Recommendations: An Empirical Comparison of Pre-filtering, Post-filtering and Contextual Modeling Approaches*, σελίδες 137–149. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[10] Annie Chen. Context-aware collaborative filtering system: Predicting the user's preference in the ubiquitous computing environment. Στο *Proceedings of the First International Conference on Location- and Context-Awareness*, LoCA'05, σελίδες 244–253, Berlin, Heidelberg, 2005. Springer-Verlag.

[11] Victor Codina και Luigi Ceccaroni. *A Recommendation System for the Semantic Web*, σελίδες 45–52. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[12] Marco Degemmis, Pasquale Lops και Giovanni Semeraro. A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, 17(3):217–255, 2007.

[13] Michael D. Ekstrand, John T. Riedl και Joseph A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, 2011.

[14] Negar Hariri, Bamshad Mobasher και Robin Burke. Context-aware music recommendation based on latenttopic sequential patterns. Στο *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, σελίδες 131–138, New York, NY, USA, 2012. ACM.

[15] Dietmar Jannach, Markus Zanker, Alexander Felfernig και Gerhard Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, 2010.

[16] George Karypis. Evaluation of item-based top-n recommendation algorithms. Στο *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, σελίδες 247–254, New York, NY, USA, 2001. ACM.

[17] Slava Kisilevich, Florian Mansmann και Daniel Keim. P-dbscan: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. Στο *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, COM.Geo '10, σελίδες 38:1–38:4, New York, NY, USA, 2010. ACM.

[18] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,*, 5(3):88–95, 2008.

[19] Yehuda Koren και Robert Bell. *Advances in Collaborative Filtering*, σελίδες 145–186. Springer US, Boston, MA, 2011.

[20] Jae Sik Lee και Jin Chun Lee. Context awareness by case-based reasoning in a music recommendation system. Στο *Proceedings of the 4th International Conference on Ubiquitous Computing Systems*, UCS'07, σελίδες 45–58, Berlin, Heidelberg, 2007. Springer-Verlag.

[21] Pasquale Lops, Marcode Gemmis και Giovanni Semeraro. Content-based recommender systems: State of the art and trends. Στο *Recommender Systems Handbook*Francesco Ricci, Lior Rokach, Bracha Shapira και Paul B. Kantor, επιμελητές, σελίδες 73–105. Springer, 2011.

[22] C. Manning και P. Raghavan. An introduction to information retrieval. Στο *Journal of the American Society for Information Science*. Cambridge Univ Press, 2009.

[23] Fedelucio Narducci, Cataldo Musto, Giovanni Semeraro, Pasquale Lops και Marcode Gemmis. *Leveraging Encyclopedic Knowledge for Transparent and Serendipitous User Profiles*, σελίδες 350–352. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[24] Ante Odic, Marko Tkalcik και Andrej Kosir. Predicting and detecting the relevant contextual information in a movie-recommender system. *User Modeling and User-Adapted Interaction*, σελίδες 74–90, 2013.

[25] Ante Odić, Marko Tkalčič, Andrej Košir και Jurij F. Tasič. A.: Relevant context in a movie recommender system: Users' opinion vs. statistical detection. Στο *In: Proc. of the 4th Workshop on Context-Aware Recommender Systems (2011*, χ.χ.

[26] C. Palmisano, A. Tuzhilin και M. Gorgoglione. Using context to improve predictive modeling of customers in personalization applications. *IEEE Transactions on Knowledge and Data Engineering*, 20(11):1535–1549, 2008.

[27] Umberto Panniello, Alexander Tuzhilin, Michele Gorgoglione, Cosimo Palmisano και Anto Pedone. Experimental comparison of pre- vs. post-filtering approaches in context-aware recommender systems. Στο *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, σελίδες 265–268, New York, NY, USA, 2009. ACM.

[28] Michael J. Pazzani και Daniel Billsus. Content-based recommendation systems. Στο *The Adaptive Web: Methods and Strategies of Web Personalization*, τόμος 4321, σελίδες 325–341. Springer, 2007.

[29] Pearl Pu, Li Chen και Rong Hu. Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4-5):317–355, 2012.

[30] Francesco Ricci. *Recommender Systems Handbook*. Springer, 2010.

[31] G. Salton, A. Wong και C.S. Yang. A vector space model for automatic indexing. Στο *Journal of the American Society for Information Science*, τόμος 18, σελίδες 613–620, 1975.

[32] Badrul Sarwar, George Karypis, Joseph Konstan και John Riedl. Item-based collaborative filtering recommendation algorithms. Στο *Proceedings of the 10th International*

Conference on World Wide Web, WWW '01, σελίδες 285–295, New York, NY, USA, 2001. ACM.

[33] Badrul M. Sarwar, George Karypis, Joseph A. Konstan και John T. Riedl. Application of dimensionality reduction in recommender system – a case study. Στο *IN ACM WEBKDD WORKSHOP*, 2000.

[34] J. Ben Schafer, Dan Frankowski, Jon Herlocker και Shilad Sen. The adaptive web. Στο *The Adaptive Web*Peter Brusilovsky, Alfred Kobsa και Wolfgang Nejdl, επιμελητές, κεφάλαιο Collaborative Filtering Recommender Systems, σελίδες 291–324. Springer-Verlag, Berlin, Heidelberg, 2007.

[35] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[36] Guy Shani και Asela Gunawardana. *Evaluating Recommendation Systems*, σελίδες 257–297. Springer US, Boston, MA, 2011.

[37] Zhenxing Xu, Ling Chen και Gencai Chen. Topic based context-aware travel recommendation method exploiting geotagged photos. *Neurocomput.*, 155(C):99–107, 2015.

[38] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker και Ion Stoica. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, 2016.

[39] Y. Zheng, B. Mobasher και R. Burke. Carskit: A java-based context-aware recommendation engine. Στο *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, σελίδες 1668–1671, 2015.

[40] Yong Zheng, Robin Burke και Bamshad Mobasher. *Differential Context Relaxation for Context-Aware Travel Recommendation*, σελίδες 88–99. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[41] Yong Zheng, Robin Burke και Bamshad Mobasher. *Recommendation with Differential Context Weighting*, σελίδες 152–164. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[42] Yong Zheng, Robin Burke και Bamshad Mobasher. Splitting approaches for context-aware recommendation: An empirical study. Στο *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, σελίδες 274–279, New York, NY, USA, 2014. ACM.