ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/ Υ, ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ & ΔΙΚΤΥΩΝ

Ζητήματα διατήρησης ιδιωτικότητας σε συστήματα συστάσεων

Privacy preservation in recommender systems

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Γεωργούδη Γ. Βασιλικής

Επιβλέποντες: Τασιούλας Λέανδρος
Καθηγητής

Κουτσόπουλος Ιορδάνης
Επίκουρος Καθηγητής

Βόλος, Φεβρουάριος 2013

Η σελίδα αυτή είναι σκόπιμα κενή.

(Υπογραφή)

……..
ΓΕΩΡΓΟΥΔΗ ΒΑΣΙΛΙΚΗ

Στην οικογένεια μου,

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό θα ήθελα να ευχαριστήσω όλους όσους με οποιονδήποτε τρόπο συνέβαλαν στην ολοκλήρωση της παρούσας διπλωματικής εργασίας που συνεπάγεται και την ολοκλήρωση των σπουδών μου.

Θα ήθελα να ευχαριστήσω τον κ. Κουτσόπουλο Ιορδάνη για την δυνατότητα που μου έδωσε να ασχοληθώ με αυτό το θέμα, την εμπιστοσύνη που μου έδειξε, τις υποδείξεις και τις συμβουλές του καθόλη τη διάρκεια εκπόνησης της παρούσας εργασίας. Επίσης ευχαριστώ τον κ. Τασιούλα Λέανδρο που δέχθηκε να αναλάβει την επίβλεψη της διπλωματικής εργασίας μου.

Τα λόγια σίγουρα δεν αρκούν για να εκφράσω την ευγνωμοσύνη στην οικογένεια μου για την στήριξη τους όλα αυτά τα χρόνια. Ευχαριστώ από καρδιάς τους γονείς μου Γιώργο και Ελευθερία για την δυνατότητα που μου έδωσαν να σπουδάσω και να μορφωθώ, την κατανόηση τους και την υποστήριξη τους σε όλα τα επίπεδα. Ευχαριστώ επίσης τις αγαπημένες μου αδελφές Κυριακή και Θεοδώρα για την αρμονική συγκατοίκηση μας αυτά τα χρόνια, την υπομονή και ανοχή που επέδειξαν και που είναι πάντα δίπλα μου.

Τέλος ευχαριστώ τις φίλες μου, Ζωή και Ιφιγένεια για την άριστη συνεργασία και τη στήριξη τους σε θέματα σχολής, για τις ανησυχίες που μοιραστήκαμε αλλά και για τις πολύ όμορφες στιγμές, τα ταξίδια και τα όνειρα που κάναμε παρέα μαζί από το πρώτο έτος.

Βόλος, Φεβρουάριος 2013

Γεωργούδη Βασιλική

ΠΕΡΙΛΗΨΗ

Τα συστήματα συστάσεων αποτελούν ένα όλο και πιο δημοφιλές εργαλείο που απευθύνεται στους καταναλωτές προκειμένου να αντιμετωπίσουν το πρόβλημα της πληθώρας των πληροφοριών γνωστό και ως information overload. Δίνοντας οι χρήστες κάποιες προσωπικές πληροφορίες, τα συστήματα αυτά εξατομικεύουν τη συμπεριφορά τους στο διαδίκτυο και είναι σε θέση να τους καθοδηγήσουν προς τη λήψη καλύτερων αποφάσεων. Συνήθως απαιτείται ένας κεντρικός υπολογιστής όπου συγκεντρώνονται οι αξιολογήσεις όλων των χρηστών με βάση τις οποίες το σύστημα θα παράγει τελικά χρήσιμες συστάσεις. Το γεγονός αυτό ωστόσο θέτει ζητήματα σχετικά με την ασφάλεια προσωπικών δεδομένων. Ιδανικά ο κάθε χρήστης θα ήθελε να δηλώσει στον κεντρικό υπολογιστή τέτοια πληροφορία, σε ότι αφορά τις αξιολογήσεις του, έτσι ώστε από τη μια να διαφυλαχθεί η ιδιωτικότητά του όσο το δυνατόν περισσότερο και από την άλλη να μην υποβαθμιστεί καθόλου η ποιότητα της σύστασης που θα λάβει σε σύγκριση με αυτή που θα λάμβανε αν είχε αποκαλύψει τις πραγματικές του αξιολογήσεις.

Στόχος της παρούσας διπλωματικής εργασίας είναι να βρεθεί το αντιστάθμισμα (trade-off) ανάμεσα στην ιδιωτικότητα και στην υψηλής ποιότητας σύσταση. Στηριζόμενοι στη θεωρία παιγνίων ορίζουμε το μοντέλο και μελετάμε την αλληλεπίδραση των χρηστών καταλήγοντας σε συνθήκες και εκφράσεις που ορίζουν το Σημείο Ισορροπίας Nash. Μετά από πεπερασμένο αριθμό επαναλήψεων η στρατηγική κάθε χρήστη συγκλίνει στο σημείο αυτό. Για ένα υβριδικό σύστημα συστάσεων η στρατηγική αυτή είναι η εξής: κάθε χρήστης αρκεί να δηλώσει ψεύτικη βαθμολογία για ένα μόνο αντικείμενο που έχει δει, εκείνο που έχει βαθμολογήσει πιο υψηλά και είναι λιγότερο συσχετιζόμενο με τα υποψήφια αντικείμενα προς σύσταση. Στη συνέχεια εφαρμόζουμε το προτεινόμενο μοντέλο σε δύο δημόσια διαθέσιμα σύνολα δεδομένων από την ερευνητική ομάδα GroupLens και αξιολογούμε την αποτελεσματικότητά του χρησιμοποιώντας κατάλληλα μετρικά. Επίσης παρουσιάζονται σενάρια συνεργασίας μεταξύ των χρηστών με στόχο το κοινό κέρδος. Κλείνοντας, γίνεται αναφορά σε πιθανές βελτιώσεις και μελλοντικές επεκτάσεις.

Λέξεις Κλειδιά: συστήματα συστάσεων, διατήρηση ιδιωτικότητας, θεωρία παιγνίων, αξιολόγηση, στρατηγικές συνεργασίας

ABSTRACT

Recommender systems are an increasingly popular tool used by many consumers to help deal with information overload. At the cost of some personal information, these systems are able to personalize a user's online experience and guide him toward making better decisions. Usually, a central server needs to have access to users' ratings profiles in order to make useful recommendations. Having this access, however, undermines the users' privacy. Each user would like to declare a rating profile so as to preserve data privacy as much as possible, while not causing deterioration in the quality of the recommendation he would get, compared to the one he would get if he revealed his true private rating profile.

The aim of this diploma thesis is to address the tradeoff between privacy preservation and high-quality recommendation. Based on game theory to model and study the interaction of users, we derive conditions and expressions for the Nash Equilibrium Point (NEP). User strategies converge to the NEP after an iterative best-response strategy update. For a hybrid recommendation system we find that NEP strategy for every user is to declare false rating only for one item, the one which is highly ranked and less correlated with the proposed for recommendation. Applying the suggested model on two publicly available datasets from GroupLens Research Group we show the effectiveness of the method using appropriate evaluation metrics. Also cooperative strategies between users which can mutually benefit them are presented. In conclusion, suggestions for future research and improvements are addressed.

Keywords: recommender systems, privacy preservation, game theory, evaluation, co-operation strategies

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

1.1 ΠΡΟΛΟΓΟΣ

Είναι γεγονός ότι τα τελευταία χρόνια παρατηρείται εκτεταμένη αύξηση πληροφοριών. Πλήθος προϊόντων (βιβλία, CDs, ταινίες, άρθρα, διαφημίσεις) αλλά και υπηρεσιών κατακλύζουν τους χρήστες του διαδικτύου καθιστώντας ακόμη δυσκολότερη την λήψη αποφάσεων σχετικά με το τι είναι προτιμότερο να επιλέξουν για να καλύψουν τις ανάγκες τους. Όπως και στην καθημερινή ζωή, οι άνθρωποι συνήθως στηρίζονται στις προτάσεις άλλων για να πάρουν κάποιες αποφάσεις: για παράδειγμα ρωτάνε έναν φίλο τους την γνώμη του για ένα εστιατόριο στο οποίο έχει πάει ή διαβάζουν την αξιολόγηση ενός κριτικού κινηματογράφου σε μια εφημερίδα για να αποφασίσουν αν θα παρακολουθήσουν μια ταινία. Κάτι ανάλογο κάνουν και τα συστήματα συστάσεων στο διαδίκτυο. Αντιμετωπίζουν το πρόβλημα της πληθώρας των πληροφοριών εφαρμόζοντας τεχνικές εξατομίκευσης και αποτελούν ίσως την πιο δημοφιλή μορφή της. Εμφανίστηκαν στα μέσα τις δεκαετίας του ’90 και από τότε έχουν εξελιχθεί πολύ. Υπάρχουν συνέδρια (όπως το ετήσιο ACM Recommender System[1] που ξεκίνησε το 2007), μελέτες και αρκετή βιβλιογραφία αφιερωμένα στον συγκεκριμένο τομέα, διδάσκονται ακόμη και μαθήματα σε πανεπιστήμια. Το ενδιαφέρον για τα συστήματα συστάσεων αυξήθηκε μετά και από τον ανοιχτό διαγωνισμό των 1,000,000 δολαρίων που προκήρυξε το Netflix τον Οκτώβριο 2007 και διήρκησε σχεδόν δυο χρόνια. Νικήτρια ομάδα ανακηρύχθηκε η Bellkor’s Pragmatic Chaos η οποία κατάφερε να βελτιώσει τον υπάρχων αλγόριθμο σε ποσοστό 10,09% (RMSE:0.8567)[2].

Ένα τυπικό σύστημα συστάσεων αποτελείται από:

- Ένα σύνολο χρηστών U =$\{ u_1, u_2, \ldots, u_N \}$.
- Ένα σύνολο αντικειμένων I = $\{ i_1, i_2, \ldots, i_M \}$.
- Ένα σύνολο από πιθανές τιμές βαθμολόγησης V (π.χ. V = {1,2,3,4,5}).
- Ένα σύνολο από βαθμολογίες για κάθε χρήστη της μορφής R ={(u,i,r)} όπου $u \in U$, $i \in I$ και $r \in V$.

Ένα σύστημα συστάσεων μπορεί να απεικονιστεί ως ένας δισδιάστατος $M \times N$ πίνακας όπως φαίνεται παρακάτω Πίνακας 1. Κάθε κελί του πίνακα $r_{u,i}$ είναι ένας αριθμός $r \in V$ που αντιπροσωπεύει την βαθμολογία του χρήστη u για το αντικείμενο i ή μηδέν αν ο χρήστης u δεν έχει βαθμολογήσει το συγκεκριμένο αντικείμενο.

[1] http://recsys.acm.org/

[2] http://www.netflixprize.com/

Πίνακας 1: Utility Matrix

Items / Users	1		i		j			m
1	4	5				0	0	0
2	1	0				0	3	5
n	0	4				2	2	2

Βασικά χαρακτηριστικά των συστημάτων συστάσεων είναι:

- Διατεταγμένα δεδομένα (Ordinal Data): π.χ. όλα τα δεδομένα ανήκουν στο διάστημα [1,5].
- Μεγάλης διάστασης πίνακες (High Dimensionality): έχει τόσες στήλες όσα και τα διαθέσιμα αντικείμενα που συνήθως είναι στις τάξεις των χιλιάδων.
- Αραιοί πίνακες (Sparsity): καθώς στην πραγματικότητα οι χρήστες βαθμολογούν πολύ λιγότερα αντικείμενα από τα συνολικά διαθέσιμα.

1.2 ΚΙΝΗΤΡΑ ΚΑΙ ΣΤΟΧΟΣ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ

Για να μπορέσει το σύστημα να δώσει έγκυρες, κατάλληλες και εξατομικευμένες προτάσεις στον χρήστη πρέπει να έχει πρόσβαση σε κάποιες πληροφορίες που αφορούν τις προτιμήσεις του, δημογραφικά χαρακτηριστικά (ηλικία, φύλο, χώρα κ.ά.) που στη συνέχεια θα δημιουργήσουν το προφίλ του. Αυτό το προφίλ τις περισσότερες φορές αποθηκεύεται και διατηρείται σε έναν κεντρικό υπολογιστή online στον οποίο θα τρέξει και ο αλγόριθμος παραγωγής συστάσεων. Συνεπώς, ο χρήστης θα πρέπει να εμπιστευτεί τον κεντρικό υπολογιστή ότι δεν θα αποκαλύψει τις προσωπικές του πληροφορίες. Η ανησυχία για την πιθανή αποκάλυψη προσωπικών στοιχείων ποικίλει βέβαια ανάλογα και το πεδίο εφαρμογής. Είναι πιθανότερο ένας χρήστης να μοιραστεί εύκολα τις προτιμήσεις του στη μουσική ή το αγαπημένο του είδος ταινίας αλλά πιο δύσκολα θα αποκαλύψει πληροφορίες σχετικές με θέματα υγείας.

Αναφορικά με αυτή την πρόσβαση λοιπόν εγείρονται και ζητήματα ιδιωτικότητας και προστασίας αυτής. Πόσα πρέπει να γνωρίζει το σύστημα για να δώσει ακριβείς προτάσεις; Υπάρχει κάποιο όριο στο οποίο πρέπει να σταματήσει να συλλέγει πληροφορίες; Πόση πληροφορία για τον εαυτό του αποκαλύπτει κάποιος χρήστης όταν πει την γνώμη του και βαθμολογήσει κάποιο αντικείμενο; Τι είδους πληροφορία είναι αυτή και πώς/αν συνδέεται με την πραγματική ταυτότητα του χρήστη; Δηλαδή μπορεί για παράδειγμα να αποκαλυφτεί η πραγματική του ταυτότητα (ονοματεπώνυμο, στοιχεία πιστωτικής κάρτας) αν αποκαλυφθούν οι αξιολογήσεις του σε αγορές που έκανε στο Amazon; Πώς σχετίζεται η αξία της πρότασης που θα λάβει ως απάντηση από το σύστημα με την ιδιωτικότητα που χάνει; Είναι προφανές ότι όσο περισσότερα γνωρίζει το σύστημα για έναν χρήστη τόσο καλύτερη θα είναι και η πρόταση του για αυτόν. Λέγοντας καλύτερη εννοούμε να μεγιστοποιείται η πιθανότητα να του αρέσει κάτι που του προτάθηκε και δεν γνώριζε από πριν.

Στόχος της παρούσας διπλωματικής εργασίας είναι να προτείνει ένα νέο τρόπο αντιμετώπισης του ζητήματος διατήρησης της ιδιωτικότητας στα συστήματα συστάσεων στηριζόμενο στη Θεωρία Παιγνίων. Διατυπώνουμε τις έννοιες της ιδιωτικότητας και της ακρίβειας της σύστασης με μαθηματικούς όρους και ορίζουμε την στρατηγική που θα ακολουθήσουν οι παίκτες-χρήστες σε ότι αφορά τις βαθμολογίες που θα δηλώσουν έτσι ώστε να προστατέψουν τα προσωπικά τους δεδομένα χωρίς να υποβαθμίζουν την ποιότητα της σύστασης που θα λάβουν από το σύστημα.

1.3 ΔΙΑΡΘΡΩΣΗ ΕΡΓΑΣΙΑΣ

Στο κεφάλαιο 2 γίνεται μια πιο λεπτομερής αναφορά στα συστήματα συστάσεων και τα διαφορετικά είδη που υπάρχουν. Εξετάζονται επίσης, οι προτεινόμενες μέχρι σήμερα προσεγγίσεις για διατήρηση της ιδιωτικότητας. Στο κεφάλαιο 3 περιγράφεται η προτεινόμενη μέθοδος, πώς ορίζονται η ιδιωτικότητα και η ποιότητα σύστασης και με ποιο τρόπο επικοινωνούν και ανταλλάσουν δεδομένα οι χρήστες με τον server. Στο κεφάλαιο 4 παρουσιάζονται οι λεπτομέρειες υλοποίησης. Το κεφάλαιο 5 περιλαμβάνει τα αποτελέσματα των πειραμάτων που έγιναν σε γνωστά σύνολα δεδομένων για την αξιολόγηση της μεθόδου. Στο κεφάλαιο 6 περιγράφονται δύο σενάρια συνεργασίας μεταξύ των χρηστών για την από κοινού ωφέλεια. Ολοκληρώνοντας, το κεφάλαιο 7 αναφέρεται στα τελικά συμπεράσματα και σε πιθανές μελλοντικές επεκτάσεις.

ΚΕΦΑΛΑΙΟ 2

2.1 ΕΙΣΑΓΩΓΗ ΣΤΑ ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ

Όλοι πιθανότατα έχουμε έρθει σε επαφή κυρίως καθώς χρησιμοποιούμε το διαδίκτυο με τέτοια συστήματα. Εφαρμογές τους υπάρχουν σε τομείς όπως η ψυχαγωγία, η ενημέρωση, το ηλεκτρονικό εμπόριο, οι διαδικτυακές υπηρεσίες. Κάποια δημοφιλή συστήματα συστάσεων είναι τα παρακάτω :

- Amazon.com[3]: Από τα μεγαλύτερα ηλεκτρονικά καταστήματα, ξεκίνησε ως ηλεκτρονικό βιβλιοπωλείο. Αφού έχεις δει ένα προϊόν, το σύστημα θα σου προτείνει άλλα προϊόντα σύμφωνα με τι έχουν δει ή αγοράσει ή αξιολογήσει άλλοι χρήστες που έχουν δει και το πρώτο.
- Netflix[4]: Παρέχει προτάσεις για ταινίες που πιθανόν να αρέσουν στο χρήστη με βάση προηγούμενες αξιολογήσεις του σε ταινίες, το ιστορικό προτιμήσεων του χρήστη ή/και λαμβάνοντας υπόψη χαρακτηριστικά όπως είδος ταινίας, πρωταγωνιστές ηθοποιοί κλπ. των ταινιών που συνηθίζει να βλέπει.
- Pandora Radio[5]: Δίνοντας το όνομα ενός τραγουδιού ή τραγουδιστή επιλέγει και παίζει μουσική με παρόμοια χαρακτηριστικά.
- HotPot by Google: Πρόκειται για μια εφαρμογή όπου ο χρήστης βαθμολογεί εστιατόρια, ξενοδοχεία, καφετέριες κλπ και διατηρεί μια ομάδα από φίλους τους οποίους εμπιστεύεται. Έτσι, την επόμενη φορά που θα κάνει μια αναζήτηση στο Google για το πού θα διασκεδάσει το βράδυ λόγου χάρη, θα λάβει αποτελέσματα βασισμένα στις δικιές του προτιμήσεις όπως έχουν δημιουργηθεί από τις αξιολογήσεις του και στις αξιολογήσεις των φίλων του.
- MovieLens.com[6]: Παρέχει προτάσεις για κινηματογραφικές ταινίες.

Τα συστήματα συστάσεων χρησιμοποιούν την γνώση που προέρχεται από ειδικούς και από την συμπεριφορά των χρηστών. Οι πληροφορίες που δίνονται άμεσα (με βαθμολόγηση ενός αντικειμένου) ή έμμεσα (απλή πλοήγηση, αποθήκευση στα αγαπημένα) αποθηκεύονται στη βάση δεδομένων του συστήματος και χρησιμοποιούνται για την δημιουργία συστάσεων την επόμενη φορά που θα αλληλεπιδράσει ο χρήστης με το σύστημα. Στόχος του συστήματος είναι να δημιουργηθεί μια σχέση εμπιστοσύνης με τον χρήστη έτσι ώστε ο χρήστης να είναι ικανοποιημένος από τις προτάσεις που παίρνει και να συνεχίζει να ανατροφοδοτεί το σύστημα και κατ' επέκταση να βελτιώνεται η διαδικασία παραγωγής των προβλέψεων. Τα συστήματα αυτά ευνοούν τόσο τους υπεύθυνους παροχής ηλεκτρονικών υπηρεσιών όσο και τους χρήστες των υπηρεσιών αυτών. Για τους μεν υπεύθυνους βασικό πλεονέκτημα είναι η αύξηση των πωλήσεων

[3] http://www.amazon.com/

[4] https://signup.netflix.com/global

[5] http://www.pandora.com/

[6] http://movielens.umn.edu

και των κερδών αλλά και η σχέση εμπιστοσύνης που χτίζεται με τον χρήστη και θα τον ξεχωρίσει από κάποιον ανταγωνιστή. Οι δε χρήστες ωφελούνται σε λειτουργίες όπως εύρεση χρήσιμων αντικειμένων μέσα από μεγάλη ποικιλία που δεν γνώριζαν καν ή δεν είχαν σκεφτεί ότι μπορεί να τους αρέσει, διευκόλυνση στη λήψη απόφασης για κάτι στο οποίο δεν είναι ειδικοί και δεν έχουν τις απαραίτητες γνώσεις, αλυσιδωτή πρόταση προϊόντων, δημόσια έκφραση της γνώμης τους.

Εικόνα 1. Amazon's recommendations

Εικόνα 2. Netflix recommendations

Εικόνα 3. MovieLens recommendations

2.2 ΔΟΜΙΚΑ ΣΤΟΙΧΕΙΑ ΕΝΟΣ ΣΥΣΤΗΜΑΤΟΣ ΣΥΣΤΑΣΕΩΝ

Τα βασικά δομικά στοιχεία ενός συστήματος συστάσεων είναι η είσοδος, η έξοδος και η μέθοδος παραγωγής των συστάσεων.

- Η είσοδος: Ένα σύστημα συστάσεων δέχεται ως είσοδο μια συλλογή από δεδομένα που αφορούν για παράδειγμα τις προτιμήσεις των χρηστών, τα χαρακτηριστικά των αντικειμένων. Τα δεδομένα μπορεί να προέρχονται από τον ίδιο τον χρήστη για τον οποίο θέλουμε να δημιουργήσουμε προτάσεις είτε από την κοινότητα στην οποία ανήκει. Στην πρώτη περίπτωση τα δεδομένα αφορούν υπονοούμενη πλοήγηση, σαφής πλοήγηση, λέξεις κλειδιά και χαρακτηριστικά των προϊόντων, βαθμολογία, ιστορικό συμπεριφοράς. Στην δεύτερη περίπτωση τα δεδομένα αντανακλούν τις απόψεις της κοινότητας για τα προϊόντα όπως η δημοτικότητα ενός προϊόντος, τα σχόλια και η βαθμολόγηση.
- Η έξοδος: Η έξοδος ενός συστήματος συστάσεων είναι συνήθως προτάσεις (suggestions), είναι του τύπου «δοκίμασε αυτό». Ένας άλλος τύπος εξόδου αφορά σε προβλέψεις για την βαθμολόγηση που θα έδινε κάποιος χρήστης σε ένα αντικείμενο. Αντιπροσωπευτικά παραδείγματα εξόδου από το Amazon.com: Customers who bought, Your recommendations.
- Η μέθοδος: Γνωστές μέθοδοι παραγωγής προτάσεων είναι: Raw Retrieval, Manual Selection, Statistical Summaries, Item-to-item Correlation, User-to-user Correlation, Attribute-based Technologies.

Το σύστημα συστάσεων σχετίζει κάθε ζευγάρι χρήστη- προϊόν με μια τιμή βαθμολόγησης εφαρμόζοντας μια συνάρτηση $R : Users \times Items \rightarrow Ratings$. Το προϊόν με την υψηλότερη βαθμολόγηση προτείνεται και στον χρήστη.

2.3 ΤΑΞΙΝΟΜΗΣΗ ΣΥΣΤΗΜΑΤΩΝ ΣΥΣΤΑΣΕΩΝ

Τα συστήματα συστάσεων βασίζονται σε τεχνικές ανάλυσης δεδομένων και με βάση τις τεχνικές που ακολουθούνται κατηγοριοποιούνται ως εξής:

1. Συνεργατικής διήθησης ή Συνεργατικού φιλτραρίσματος (Collaborative Filtering)
2. Φιλτράρισμα με βάση το περιεχομένου (Content-Based Filtering)
3. Υβριδικά (Hybrid)

2.3.1 ΣΥΣΤΗΜΑΤΑ ΣΥΝΕΡΓΑΤΙΚΗΣ ΔΙΗΘΗΣΗΣ

Στα συστήματα συνεργατικής διήθησης οι προβλέψεις σχετικά με την χρησιμότητα ενός αντικειμένου για κάποιο χρήστη γίνονται λαμβάνοντας υπόψη τις αξιολογήσεις των άλλων χρηστών για το ίδιο αντικείμενο. Μπορεί να εκφράζονται με βαθμολόγηση στην κλίμακα 1 έως 5 αστέρια, ή «μου αρέσει»/«δεν μου αρέσει». Πιο συγκεκριμένα, η λογική πάνω στην οποία στηρίζεται το μοντέλο αυτό είναι ότι αν ο χρήστης u συμφωνεί για το αντικείμενο i με τον χρήστη u' (άρα ο χρήστης u είναι παρόμοιος με τον u') είναι πιο πιθανό να συμφωνεί με την άποψη του u'

και για ένα νέο αντικείμενο j παρά με κάποιον άλλο τυχαίο χρήστη. Τα περισσότερα κοινωνικά δίκτυα (Facebook, LinkedIn, MySpace) εφαρμόζουν αυτή την προσέγγιση για να προτείνουν φίλους. Στόχος τέτοιων συστημάτων είναι να δημιουργήσουν γειτονιές χρηστών που μοιράζονται κοινά ενδιαφέροντα ή που μοιάζουν. Οπότε όταν ένας χρήστης ζητήσει μια πρόταση το σύστημα θα βρει τους γείτονες του και θα του προτείνει κάτι που δεν έχει δει ο ίδιος αλλά έχουν βαθμολογήσει υψηλά οι γείτονές του. Χωρίζονται σε δύο γενικές κατηγορίες: με βάση το μνήμη (memory-based ή neighborhood-based) και με βάση το μοντέλο (model-based).

Στα memory-based η βαθμολογία για ένα άγνωστο αντικείμενο υπολογίζεται ως συνάθροιση (συνήθως μέσος όρος ή σταθμισμένος μέσος όρος με βάρος την ομοιότητα μεταξύ δύο χρηστών) των βαθμολογιών κάποιων χρηστών για το συγκεκριμένο αντικείμενο. Σημαντικό εδώ είναι ο ορισμός της ομοιότητας. Η ομοιότητα μεταξύ δύο χρηστών sim(u,u') αποτελεί μετρική απόστασης και συνήθως είναι βασισμένη στις εκτιμήσεις για κοινά αντικείμενα και από τους δύο χρήστες. Δημοφιλή μέτρα ομοιότητας: συντελεστής Pearson(Pearson's Correlation Coefficient), με βάση το συνημίτονο της γωνίας των διανυσμάτων (Cosine Similarity), μέση τετραγωνική απόσταση (Mean Square Root Distance). Με τον ίδιο τρόπο μπορούμε να υπολογίσουμε την ομοιότητα μεταξύ αντικειμένων (Item-based CF).

Αντίθετα τα model-based συστήματα χρησιμοποιούν τα δεδομένα για να εκπαιδεύσουν ένα μοντέλο το οποίο στη συνέχεια χρησιμοποιούν για να κάνουν προβλέψεις. Η διαδικασία μοντελοποίησης βασίζεται σε τεχνικές εξόρυξης δεδομένων (Data Mining) όπως Bayesian δίκτυα (Bayesian networks), συσταδοποίηση (όπως ο αλγόριθμος k-means), η κρυφή σημασιολογική ανάλυση (Latent Semantic Analysis), η διάσπαση ιδιαζουσών τιμών (Singular Value Decomposition).

Τα συστήματα συνεργατικής διήθησης έχουν γίνει πολύ δημοφιλή χάρη στην απλότητα, την ευκολία δημιουργίας και εφαρμογής τους αλλά και για την αποδοτικότητα τους. Ειδικά τα model-based έχουν μικρές απαιτήσεις μνήμης και CPU. Ωστόσο παρουσιάζουν και βασικά μειονεκτήματα όπως:

- το «cold-start problem» το οποίο οφείλεται στο γεγονός ότι δεν είναι διαθέσιμη αρχική πληροφορία για τις προτιμήσεις ενός νέου χρήστη.
- η κλιμάκωση (scalability) ειδικά για τα memory-based γιατί οι υπολογισμοί αυξάνουν καθώς μεγαλώνει ο αριθμός των χρηστών και των αντικειμένων.
- η έλλειψη βαθμολόγησης (sparsity) για αντικείμενα π.χ. στο Amazon.com ένας ενεργός χρήστης έχει βαθμολογήσει λιγότερα από 1% των βιβλίων που σημαίνει 20.000 βιβλία (1% από 2 εκατομμύρια βιβλία που υπάρχουν συνολικά).
- προκατάληψη για τα δημοφιλή αντικείμενα (popularity bias) όλοι έχουν δει τον Τιτανικό, δεν είναι χρήσιμο να συστήνει αυτή την ταινία σε όλους.

2.3.2 ΦΙΛΤΡΑΡΙΣΜΑ ΜΕ ΒΑΣΗ ΤΟ ΠΕΡΙΕΧΟΜΕΝΟ

Η προσέγγιση με βάση το περιεχόμενο έχει τις ρίζες της στις επιστημονικές περιοχές της θεωρίας ανάκτησης πληροφοριών (Information Retrieval) και του φιλτραρίσματος πληροφοριών (Information Filtering). Υιοθετεί τεχνικές συσταδοποίησης (Clustering), ταξινόμησης (Classification), μηχανικής μάθησης (Machine Learning) όπως τεχνολογίες τεχνητών νευρωνικών

δικτύων (Artificial Neural Networks) και δέντρων απόφασης (Decision Trees). Το σύστημα δέχεται πληροφορίες σχετικές με τη φύση/χαρακτηριστικά των αντικειμένων. Οι συστάσεις παράγονται στη συνέχεια λαμβάνοντας υπόψη το περιεχόμενο των εγγράφων και φιλτράρονται ώστε να επιτευχθεί η καλύτερη αντιστοίχηση με τις προτιμήσεις του χρήστη όπως αυτές έχουν καταγραφεί στο προφίλ του. Ας λάβουμε υπόψη το είδος των πληροφοριών που μπορεί να αποκτήσει ένας χρήστης μόνος του μέσα από το διαδίκτυο για μια αγαπημένη του ταινία: ηθοποιοί, σκηνοθέτες, πλοκή, είδος ταινίας, λεπτομέρειες παραγωγής, κριτικές, βαθμολόγηση, τρέιλερ και άλλα. Για να αποφασίσει ποια ταινία να παρακολουθήσει θα εξετάσει κάποια από τα παραπάνω στοιχεία κατά πόσο είναι κοντά με τα δικά του γούστα. Την ίδια διαδικασία εκτελεί το σύστημα συστάσεων βασισμένο στο περιεχόμενο συγκρίνοντας τις ταινίες που έχει βαθμολογήσει στο παρελθόν με τις υποψήφιες προτεινόμενες και επιστρέφει αυτές με τον υψηλότερο βαθμό ομοιότητας. Τελικά πετυχαίνει με μικρό κόστος και χωρίς την επιπλέον προσπάθεια του χρήστη να προβλέψει σωστά και να τον διευκολύνει. Γνωστά συστήματα συστάσεων βάσει περιεχομένου είναι το Pandora Radio, Internet Movie Database, Rotten Tomatoes.
Βασικά πλεονεκτήματα τους είναι ότι δεν χρειάζεται πρόσβαση σε δεδομένα άλλων χρηστών, μπορούν να προτείνουν σε χρήστες με μοναδικά γούστα, μπορούν να αιτιολογήσουν γιατί πρότειναν κάτι. Από την άλλη όμως δεν υπάρχει πάντα η δυνατότητα περιγραφής του περιεχομένου ή μπορεί να χρειαστεί να περαστεί χειρωνακτικά αυτή η πληροφορία, είναι λιγότερο ακριβή, δεν προτείνουν κάτι διαφορετικό μόνο αντικείμενα όμοια με αυτά που έχουν ήδη βαθμολογήσει οι χρήστες(overspecialization).

2.3.3 ΥΒΡΙΔΙΚΑ ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ

Τέλος, υπάρχουν και τα συστήματα συστάσεων που συνδυάζουν τις παραπάνω τεχνικές και εκμεταλλεύονται τα πλεονεκτήματα και των δύο. Οι τρόποι που συνδυάζονται είναι οι εξής:

- Υλοποίηση των τεχνικών ξεχωριστά και συνδυασμός των προβλέψεων.
- Ενσωμάτωση content-based χαρακτηριστικών σε μια collaborative προσέγγιση.
- Ενσωμάτωση collaborative χαρακτηριστικών σε μια content-based προσέγγιση.
- Δημιουργία ενός γενικού μοντέλου το οποίο ενσωματώνει και content-based και collaborative χαρακτηριστικά.

Φαίνεται ότι αυτή η προσέγγιση δίνει πιο ακριβή αποτελέσματα και εφαρμόζεται αποτελεσματικά για να ξεπεράσει τις αδυναμίες των παραπάνω μεθόδων όταν εφαρμόζονται ανεξάρτητα.

Συνοψίζοντας παρακάτω στον Πίνακα 2: Έστω u ο χρήστης-στόχος για τον οποίο θέλουμε να παράγουμε πρόβλεψη και i το αντικείμενο για το οποίο ζητείται να προβλεφθεί η βαθμολογία του, I είναι το σύνολο όλων των αντικειμένων και U το σύνολο όλων των χρηστών.

Πίνακας 2. Είδη συστημάτων συστάσεων

ΤΕΧΝΙΚΗ	ΔΕΔΟΜΕΝΑ	ΕΙΣΟΔΟΣ	ΔΙΑΔΙΚΑΣΙΑ
Content-based	Χαρακτηριστικά των αντικειμένων $\in I$	Βαθμολογίες του χρήστη u για τα αντικείμενα $i \in I$	Παραγωγή προφιλ με βάση την βαθμολογική συμπεριφορά του u και χρήση του στο i
Collaborative	Βαθμολογίες των χρηστών $\in U$ για τα αντικείμενα $\in I$	Βαθμολογίες του χρήστη u για τα αντικείμενα $i \in I$	Εύρεση χρηστών $\in U$ όμοιων με τον u και συνδυασμός των βαθμολογιών τους για πρόβλεψη της βαθμολογίας του i

2.4 ΣΧΕΤΙΚΕΣ ΔΗΜΟΣΙΕΥΣΕΙΣ

Σύμφωνα με έρευνα που διενεργήθηκε το 2003 (The Harris Poll[7]) αποκαλύπτεται ότι 90% των χρηστών ανησυχεί αρκετά έως πάρα πολύ μήπως γίνεται κατάχρηση των προσωπικών του πληροφοριών.

Για την αντιμετώπιση του ζητήματος στα συστήματα συστάσεων έχουν γίνει διάφορες προτάσεις. Μια πρώτη προσέγγιση είναι εφαρμόζοντας τεχνικές ανωνυμίας όπου οι χρήστες φανερώνουν τις προσωπικές τους πληροφορίες χωρίς να αποκαλύπτουν την ταυτότητά τους. Ωστόσο με αυτό τον τρόπο εγείρονται θέματα εγκυρότητας και ακρίβειας των δεδομένων μιας και κάποιος κακόβουλος χρήστης μπορεί να αλλοιώσει τα δεδομένα για δικό του όφελος χωρίς να γίνει αντιληπτός με αποτέλεσμα η βάση δεδομένων να είναι αναξιόπιστη. Τεχνικές ανωνυμίας μπορεί να εφαρμοστούν και μόνο στα δεδομένα [1] κατά την δημοσίευση τους, έτσι ώστε να μην μπορεί να γίνει ταυτοποίηση του κωδικού του χρήστη με τα πραγματικά του στοιχεία ωστόσο στα πλαίσια της εργασίας μας δεν θα ασχοληθούμε με αυτό.

Μια άλλη εκδοχή είναι εφαρμόζοντας τεχνικές σύγχυσης σε collaborative filtering RS (Randomized Perturbation Techniques). Η βασική ιδέα είναι ότι προστίθεται στα δεδομένα μια τιμή με τέτοιο τρόπο ώστε το κεντρικό σημείο που συγκεντρώνει τις βαθμολογίες όλων των χρηστών να γνωρίζει μόνο το εύρος των τιμών και όχι την ακριβή τιμή. Επειδή τα CF συστήματα βασίζονται στην συνολική βαθμολογία και όχι σε μεμονωμένες τιμές μπορούν να παράγουν σημαντικά αποτελέσματα [2,3]. Μια πιο πρακτική προσέγγιση παρουσιάζεται στο [10] και είναι εφαρμόσιμη σε πραγματικά συστήματα με πολλούς χρήστες. Αποτελείται από δυο μέρη: την ομαδοποίηση των χρηστών με βάση μια Hash συνάρτηση που στέλνει ο server στους χρήστες και

[7] http://www.harrisinteractive.com/vault/Harris-Interactive-Poll-Research-Most-People-Are-Privacy-Pragmatists-Who-While-Conc-2003-03.pdf

την εφαρμόζουν στα δεδομένα τους και την δημιουργία του δηλωθέντος διανύσματος στον server το οποίο δημιουργείται προσθέτοντας κάποιο θόρυβο (ψεύτικα δεδομένα) στο πραγματικό διάνυσμα με τις βαθμολογίες .

Άλλες προσεγγίσεις στρέφονται προς κατανεμημένες υλοποιήσεις (distributed CF) έτσι ώστε να περιοριστούν οι αδυναμίες του κεντρικοποιημένου CF (centralized CF). Για παράδειγμα στο [4,5] οι χρήστες δημιουργούν κοινότητες και κάθε χρήστης αναζητά σύσταση από την πιο κατάλληλη για αυτόν ομάδα. Κάθε κοινότητα υπολογίζει ένα δημόσια διαθέσιμο προφίλ συνδυάζοντας τα προφίλ όλων των συμμετεχόντων, εφαρμόζοντας SVD ανάλυση. Τα προσωπικά δεδομένα των χρηστών είναι κρυπτογραφημένα και η επικοινωνία γίνεται μεταξύ χρηστών και όχι με κάποιο κεντρικό σημείο. Μια επέκταση του προηγούμενο προτείνει τον αλγόριθμο PocketLens [6] όπου οι χρήστες επικοινωνούν πάνω από το δίκτυο μόνο για να ανταλλάξουν στοιχεία ομοιότητας διατηρώντας τοπικά τις βαθμολογίες τους. Μια άλλη πρόταση [7] στοχεύει να θολώσει τις βαθμολογίες που θα στείλει ο χρήστης στον server προκειμένου να επιλύσει το πρόβλημα της ιδιωτικότητας. Η επικοινωνία μεταξύ των χρηστών γίνεται με κατανεμημένο τρόπο αλλά η διαδικασία παραγωγής συστάσεων βασίζεται στην ύπαρξη του server. Οι χρήστες διατηρούν το πραγματικό τους προφίλ (offline) και επικοινωνούν με άλλους όμοιους χρήστες για να φτιάξουν το προφίλ που θα δηλώσουν (online) με βάση τι βαθμολογίες έχουν και οι άλλοι. Αφού ολοκληρωθεί η διαδικασία ενημερώνουν τον server και συγχρονίζουν το offline προφίλ τους με το online. Αν και αυτές οι προσεγγίσεις σχεδόν ελαχιστοποιούν την πιθανή καταπάτηση της ιδιωτικότητας, απαιτούν καλή συνεργασία μεταξύ των χρηστών για να επιτύχουν ακριβείς συστάσεις. Σε άλλη δημοσίευση [8] εξετάζεται πώς ο συνδυασμός της μη-κεντρικής αποθήκευσης των βαθμολογικών προφίλ των χρηστών και η οργάνωση των χρηστών σε ένα ομότιμο δίκτυο (peer-to-peer) μαζί με τροποποιήσεις στα δεδομένα εφαρμόζοντας τεχνικές σύγχυσης μπορεί να μετριάσει θέματα ασφάλειας προσωπικών δεδομένων.

ΚΕΦΑΛΑΙΟ 3

3.1 ΟΡΙΣΜΟΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ ΚΑΙ ΣΥΜΒΟΛΙΣΜΟΙ

3.1.1 ΔΙΑΝΥΣΜΑΤΑ ΒΑΘΜΟΛΟΓΙΩΝ ΚΑΙ ΣΥΣΤΑΣΕΩΝ

Βασικές οντότητες ενός συστήματος συστάσεων αποτελούν οι χρήστες και τα αντικείμενα. Στο εξής θα αναφέρουμε ως U το σύνολο Ν χρηστών με $|U| = N$ και I το σύνολο των αντικειμένων που είναι διαθέσιμα προς σύσταση. Κάθε χρήστης i διατηρεί ένα σύνολο S_i υποσύνολο (και συνήθως πολύ μικρότερο) του I αποτελούμενο από τα αντικείμενα που έχει έρθει σε επαφή μέχρι τότε με κάποιο τρόπο -δει, αξιολογήσει, αγοράσει-.

Για κάθε χρήστη επίσης, ορίζουμε ένα διάνυσμα $p_i = (p_{ik} : k \in S_i)$ με τις βαθμολογίες που έχει δώσει για κάθε αντικείμενο που ανήκει στο σύνολο S_i. Αυτή η πληροφορία παραμένει γνωστή μόνο στον χρήστη i και για αυτό το διάνυσμα p_i θα αναφέρεται ως το ιδιωτικό προφίλ του χρήστη i (private profile ή private ratings vector). Υποθέτουμε ότι οι χρήστες δηλώνουν τις βαθμολογίες τους για όλα τα αντικείμενα με τα οποία έχουν έρθει σε επαφή.

Όπως αναφέραμε και παραπάνω οι αλγόριθμοι συνεργατικής διήθησης απαιτούν την αλληλεπίδραση των χρηστών και την ύπαρξη ενός κεντρικού σημείου συλλογής των προφίλ όλων των χρηστών (recommendation server). Ορίζουμε επίσης ως $q_i = (q_{ik} : k \in S_i)$ το διάνυσμα με τις βαθμολογίες για όλα τα αντικείμενα που ανήκουν στο S_i που θα δηλώσει ο χρήστης i στον recommendation server. Στο εξής θα αναφερόμαστε στο διάνυσμα q_i ως το δηλωθέν προφίλ του χρήστη i (declared profile ή declared ratings vector). Για την παραγωγή εξατομικευμένων συστάσεων απαιτείται το σύνολο όλων των δηλωθέντων προφίλ το οποίο ορίζουμε ως $Q = (q_i : i \in U)$ -αν και υπάρχουν περιπτώσεις όπου θα μπορούσε να χρησιμοποιηθεί μέρος των q_i – με βάση το οποίο θα υπολογιστεί το διάνυσμα των συστάσεων $r_i = (r_{il} : l \notin S_i)$ δηλαδή το διάνυσμα που επιστρέφεται στον χρήστη i με τα αντικείμενα που δεν έχει δει ακόμη και του προτείνονται. Υποθέτουμε ότι ο server εφαρμόζει την ίδια συνάρτηση για τον υπολογισμό των διανυσμάτων συστάσεων. Το διάνυσμα αυτό επιστρέφεται στον χρήστη αυτούσιο με τυχαία σειρά ή ταξινομημένο, ή μέρος αυτού (για παράδειγμα στην περίπτωση αλγορίθμων τύπου Top-N recommendation επιστρέφονται τα Ν αντικείμενα με τις υψηλότερες όπως προβλέφθηκαν βαθμολογίες). Στη συνέχεια θα βασιστούμε στους ορισμούς και τα συμπεράσματα της δημοσίευσης [11].

3.1.2 ΔΕΙΚΤΗΣ ΜΕΤΡΗΣΗΣ ΤΗΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ (PRIVACY METRIC)

Το μέτρο της ιδιωτικότητας κάθε χρήστη εξαρτάται τόσο από το ιδιωτικό όσο και από το δηλωθέν προφίλ του και δηλώνεται ως μια συνεχής συνάρτηση $g(p_i, q_i)$ η οποία είναι ίδια για κάθε χρήστη.

3.1.3 ΠΟΙΟΤΗΤΑ ΣΥΣΤΑΣΗΣ (RECOMMENDATION QUALITY)

Η ποιότητα της σύστασης που θα λάβει ένας χρήστης εξαρτάται τόσο από το δηλωθέν προφίλ του ίδιου του χρήστη όσο και από το τι έχουν δηλώσει όλοι οι υπόλοιποι. Ακόμη και στην περίπτωση που ο χρήστης δηλώσει τις πραγματικές του βαθμολογίες στον κεντρικό υπολογιστή διακινδυνεύοντας σημαντικά την ιδιωτικότητα του δεν είναι σίγουρο ότι θα λάβει την βέλτιστη πρόταση διότι εξαρτάται και από τα δηλωθέντα διανύσματα των υπόλοιπων χρηστών. Στόχος κάθε χρήστη είναι να αποφασίσει ποιο είναι το καταλληλότερο διάνυσμα βαθμολογιών που αν δηλώσει στον κεντρικό υπολογιστή θα αποσπάσει διάνυσμα συστάσεων πολύ κοντά σε εκείνο που θα λάμβανε αν δήλωνε τις πραγματικές του βαθμολογίες.

Ας ορίσουμε ως $r_i = \mathrm{f}(q_i, q_{-i})$ το recommendation vector που θα λάβει ο τυχαίος χρήστης i αν στείλει στον κεντρικό υπολογιστή το δηλωθέν διάνυσμα q_i δεδομένων των δηλωθέντων διανυσμάτων των υπόλοιπων χρηστών $q_{-i} = (q_1, q_2, \dots q_{i-1}, q_{\iota+1}, \dots q_N)$ και ως $\tilde{r}_i = (q_i, q_{-i})$ το recommendation vector που θα λάμβανε ο χρήστης i αν έστελνε το ιδιωτικό του προφίλ στον κεντρικό υπολογιστή. Τότε ο περιορισμός που πρέπει να ικανοποιείται για τον χρήστη i είναι

$$(r_i - \tilde{r}_i)^2 \leq D \Rightarrow [\mathrm{f}(q_i, q_{-i}) - \mathrm{f}(p_i, q_{-i})]^2 \leq \mathrm{D} \quad (1)$$

με τη σταθερά D να καθορίζει τη μέγιστη παραμόρφωση στο recommendation vector.

3.1.4 ΣΥΝΘΕΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Συμπερασματικά λοιπόν, ο χρήστης επιθυμεί να προστατέψει τα προσωπικά του δεδομένα χωρίς όμως να επηρεάσει σε μεγάλο βαθμό το αποτέλεσμα που θα λάβει από το σύστημα συστάσεων. Ιδανικά θα ήθελε να δηλώσει ένα διάνυσμα q_i αρκετά διαφορετικό από το p_i διατηρώντας ταυτόχρονα τον περιορισμό (σχέση 1). Το πρόβλημα ανάγεται σε ένα πρόβλημα βελτιστοποίησης με κάποιους περιορισμούς. Στην προκειμένη, πρόκειται για ένα πρόβλημα μεγιστοποίησης με αντικειμενική συνάρτηση τον δείκτη μέτρησης της ιδιωτικότητας του χρήστη δεδομένου ότι η ποιότητα της σύστασης που θα λάβει δεν θα πέσει κάτω από ένα ορισμένο όριο. Οι χρήστες συμπεριφέρονται εγωιστικά μη λαμβάνοντας υπόψη τους στόχους ο ένας των άλλων και ενδιαφέρονται για την μεγιστοποίηση της δικής τους ωφέλειας (εδώ του privacy). Το σύνολο των εφικτών λύσεων του προβλήματος αντιπροσωπεύει το δηλωθέν διάνυσμα βαθμολογιών (declared ratings vector) του χρήστη προς τον recommendation server.

Για κάθε χρήστη i:

$$\max_{q_i} \; g(p_i, q_i)$$

$$\text{subject to } [\mathrm{f}(q_i, q_{-i}) - \mathrm{f}(p_i, q_{-i})]^2 \leq \mathrm{D}$$

Πρόβλημα 1.

Όπως προαναφέραμε, βασιζόμαστε στη Θεωρία Παιγνίων. Η Θεωρία Παιγνίων είναι μια μεθοδολογία ανάλυσης καταστάσεων μεταξύ μιας ομάδας λογικών ατόμων η οποία ανταγωνίζεται με σκοπό ο κάθε ένας να αποκτήσει το μεγαλύτερο όφελος. Σκοπός της είναι να μας βοηθήσει να καταλάβουμε διάφορες καταστάσεις στις οποίες αλληλεπιδρούν δύο ή περισσότερες οντότητες,

κάθε μία από τις οποίες συμπεριφέρεται με στρατηγικό τρόπο και προσπαθεί να πάρει κάποιες αποφάσεις. Σκοπός του κάθε παίκτη είναι να μεγιστοποιήσει το κέρδος του, το οποίο μετράται σε μια κλίμακα ωφέλειας. Το παίγνιο που αναφέρεται στην θεωρία παιγνίων αντιπροσωπεύει την κατάσταση κατά την οποία δύο ή περισσότεροι παίκτες επιλέγουν τρόπους ενέργειας, που δημιουργούν καταστάσεις αλληλεξάρτησης. Συνεπώς, η έκβαση του παιγνίου για κάθε παίκτη εξαρτάται από τις επιλογές όλων των παικτών. Ο συνδυασμός των στρατηγικών που επιλέχθηκαν από κάθε παίκτη μας δίνει την έννοια της ισορροπίας «equilibrium». Η ισορροπία στο παίγνιο δηλαδή προέρχεται από τις καλύτερες στρατηγικές μία για κάθε παίκτη στο παιχνίδι. Όλοι οι παίκτες επιλέγουν τις πιο συμφέρουσες για αυτούς ενέργειες, γνωρίζοντας και τις επιλογές των αντιπάλων τους. Δηλαδή η στρατηγική ενός παίκτη αποτελεί την καλύτερη αντίδραση στην στρατηγική του άλλου παίκτη. Αυτός ο συνδυασμός στρατηγικών αποτελεί ισορροπία Nash. Για το δικό μας πρόβλημα (Πρόβλημα 1), το διάνυσμα $Q^* = (q_1{}^*, q_2{}^* \dots, q_N{}^*)$ αποτελεί το σημείο ισορροπίας Nash αν ικανοποιείται η παρακάτω συνθήκη

$$g(p_i, q_i^*) \geq \max_{q_{i \in F}\left(q_{-i}^*\right)} g(p_i, q_i) \quad \forall\, q_i \neq q_i^* \quad (2)$$

Στο σημείο ισορροπίας Nash (NEP) κανένας χρήστης δεν μπορεί να βελτιώσει τη θέση του (εδώ το privacy) τροποποιώντας τη στρατηγική του (δηλ. το δηλωθέν προφίλ του) δεδομένου ότι όλοι οι υπόλοιποι χρήστες του συστήματος διατηρούν σταθερή την στρατηγική τους και δεν αποκλίνουν από το δικό τους σημείο ισορροπίας.

3.2 ΕΦΑΡΜΟΓΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΣΕ ΕΝΑ ΥΒΡΙΔΙΚΟ ΣΥΣΤΗΜΑ ΣΥΣΤΑΣΕΩΝ

3.2.1 ΟΡΙΣΜΟΣ ΣΥΝΑΡΤΗΣΕΩΝ

Έχοντας κάθε χρήστης δηλώσει το διάνυσμα των βαθμολογιών του στον server υπολογισμένο όπως περιγράφηκε στην ενότητα 3.1.3 εφαρμόζει την παρακάτω σχέση για να προβλέψει τις βαθμολογίες του χρήστη σε αντικείμενα που δεν έχει δει και τελικά θα περιλαμβάνονται στο επιστρεφόμενο διάνυσμα.

$$r_{il} = \frac{1}{N-1} \sum_{j \neq i : l \in S_j} q_{jl} \quad \frac{1}{|S_i|} \sum_{k \in S_i} \rho_{kl}\, q_{ik} \quad \forall l \notin S_i, l \in S_j \quad (3)$$

Όπου ρ_{kl} είναι η ομοιότητα μεταξύ των αντικειμένων k και l.

Αφού εργαζόμαστε σε ένα υβριδικό σύστημα συστάσεων το recommendation vector αποτελεί συνδυασμό της συνεργατικής διήθησης (Collaborative Filtering) και της σύστασης βασισμένη στο περιεχόμενο (Content-Based). Παρατηρώντας τη σχέση (3) ο πρώτος όρος αποτελεί μέρος της προσέγγισης της συνεργατικής διήθησης καθώς συνδυάζει τις βαθμολογίες όλων των άλλων χρηστών για το ίδιο αντικείμενο ενώ ο δεύτερος όρος επειδή περιλαμβάνει τον παράγοντα

συσχέτισης μεταξύ των αντικειμένων που έχει δει ο χρήστης και του προτεινόμενου είναι πιο κοντά στην προσέγγιση βασισμένη στο περιεχόμενο.
Σε ότι αφορά τη συνάρτηση $g(\cdot)$ που δείχνει την ιδιωτικότητα αυτή θα ορίζεται ως εξής βασισμένη στην ευκλείδεια απόσταση:

$$g(p_i, q_i) = \sum_{k \in S_i} p_{ik} \, (p_{ik} - q_{ik})^2 \quad (4)$$

Η σχέση (3) δικαιολογείται ουσιαστικά από το γεγονός ότι (αναμενόμενα) το privacy αυξάνει καθώς η ευκλείδεια απόσταση μεταξύ του ιδιωτικού προφίλ του χρήστη και του δηλωθέντος προφίλ του αυξάνει.
Μιας και ορίσαμε τον τρόπο υπολογισμού του επιστρεφόμενου διανύσματος από τον κεντρικό υπολογιστή μπορεί να οριστεί και η συνάρτηση $f(\cdot)$, και κατ' επέκταση ο περιορισμός του προβλήματος, η ποιότητα της σύστασης.

$$\frac{1}{|S_i|} \sum_{k \in S_i} \sum_{l \notin S_i} \frac{1}{N-1} \sum_{j \neq i : l \in S_j} q_{jl} \, \rho_{kl} \, (q_{ik} - p_{ik})^2 \leq D \quad (5)$$

3.2.2 ΕΠΙΚΟΙΝΩΝΙΑ ΧΡΗΣΤΩΝ ΜΕ SERVER

Στη συνέχεια περιγράφουμε την διαδικασία ανταλλαγής πληροφοριών μεταξύ των χρηστών που συμμετέχουν στο παίγνιο και του κεντρικού υπολογιστή.
Βήμα 0: Αρχικοποίηση διανυσμάτων q_i για κάθε χρήστη. Στην δική μας υλοποίηση το διάνυσμα q_i αρχικοποιείται με την τιμή 1.
Για κάθε χρήστη i=1,…,N:
Βήμα 1: Σε κάθε επανάληψη ο server στέλνει σε καθένα χρήστη *i* τις βαθμολογίες των άλλων για τα αντικείμενα που δεν ανήκουν στο S_i. Αυτή η πληροφορία είναι απαραίτητη για την διατύπωση του περιορισμού.
Βήμα 2: Κάθε χρήστης λύνει το ακόλουθο πρόβλημα βελτιστοποίησης

$$\max_{q_i^{(t)}} g(p_i, q_i^{(t)}) = \sum_{k \in S_i} p_{ik} \, \left(p_{ik} - q_{ik}^{(t)}\right)^2$$

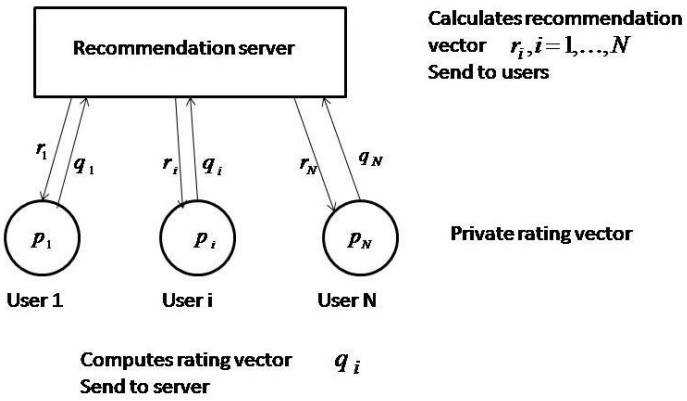
$$s.t. \quad \frac{1}{|S_i|} \sum_{k \in S_i} \sum_{l \notin S_i} \frac{1}{N-1} \sum_{j \neq i : l \in S_j} q_{jl}^{(t-1)} \, \rho_{kl} \left(q_{ik}^{(t)} - p_{ik}\right)^2 \leq D$$

Πρόβλημα 2.

Βήμα 3: Κάθε χρήστης ενημερώνει τον server για τις βαθμολογίες του στέλνοντας το διάνυσμα q_i όπως προέκυψε από τη λύση του προβλήματος 2.

Βήμα 4: Ο server ενημερώνει τους χρήστες για τις πιθανές αλλαγές στα διανύσματα q_i των άλλων χρηστών. Επιστροφή στο Βήμα 1 και επανάληψη μέχρις ότου δεν παρατηρείται αλλαγή σε κάποιο από τα διανύσματα q_i.

Παρακάτω παρουσιάζουμε σχηματικά την αρχιτεκτονική του συστήματος και την διαδικασία ανταλλαγής δεδομένων μεταξύ χρηστών και server (Εικόνα 4).



Εικόνα 4. Αρχιτεκτονική του συστήματος

3.2.3 ΑΝΑΓΩΓΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ ΣΕ ΠΡΟΒΛΗΜΑ ΓΡΑΜΜΙΚΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ

Βασικό στάδιο της υλοποίησης αποτελεί η επίλυση του προβλήματος βελτιστοποίησης. Μετά από κατάλληλες τροποποιήσεις το πρόβλημα 2 ανάγεται σε πρόβλημα Γραμμικού Προγραμματισμού (Linear Programming). Ο Γραμμικός Προγραμματισμός είναι μια ευρέως χρησιμοποιούμενη τεχνική μαθηματικής μοντελοποίησης για τον καθορισμό της βέλτιστης κατανομής πόρων ανάμεσα σε ανταγωνιστικές απαιτήσεις. Ο Γραμμικός Προγραμματισμός απαιτεί όλες οι μαθηματικές συναρτήσεις του μοντέλου να είναι γραμμικές. Πιο συγκεκριμένα, η αντικειμενική συνάρτηση να είναι γραμμική ως προς τους αγνώστους και οι περιορισμοί γραμμικές ισότητες ή ανισότητες και οι μεταβλητές των αγνώστων μη αρνητικές.

Αρκεί να θέσουμε $x_{ik} = (p_{ik} - q_{ik})^2$ με $x_i = (x_{ik} : k \in S_i)$ τότε το πρόβλημα 2 γίνεται:

$$\max_{xi} \sum_{k \in S_i} p_{ik}\ x_{ik}$$

$$subject\ to \sum_{k\ \epsilon S_i} \beta_{ik}\ x_{ik} \leq D(N-1)$$

$$με\quad \beta_{ik} = \frac{1}{|S_i|} \sum_{l \notin S_i} \sum_{j \neq i : l \epsilon S_j} q_{jl}\ \rho_{kl}$$

Πρόβλημα 3.

το οποίο είναι πρόγραμμα Γραμμικού Προγραμματισμού.

Η λύση του προβλήματος αποτελεί ακραίο σημείο του πολύτοπου των εφικτών λύσεων. Κάθε χρήστης βρίσκει ότι για το αντικείμενο $k^* = \text{argmin}_{k \in S_i} \frac{\beta_{ik}}{p_{ik}}$ η λύση του προβλήματος 3 θα είναι

$x_{ik} = \mathrm{D}\left(\mathrm{N}-1\right)|S_i| \frac{p_{ik}}{\beta_{ik}}$ (6) . Για τα υπόλοιπα αντικείμενα k ≠ k* προκύπτει ότι $x_{ik} = 0$. Αν επιστρέψουμε στο πρόβλημα 2 καταλήγουμε ότι το δηλωθέν διάνυσμα θα είναι το εξής:

$q_{ik^*} = p_{ik^*} \mp \sqrt{\frac{\mathrm{D}\left(\mathrm{N}-1\right)|S_i|\, p_{ik^*}}{\beta_{ik^*}}}$ (7) για k*

$q_{ik} = \mathrm{p_{ik}}$ για ≠ k*

Συμπεραίνουμε λοιπόν ότι ο χρήστης i μεγιστοποιεί την ιδιωτικότητα του αν δηλώσει τις πραγματικές βαθμολογίες του για όλα τα αντικείμενα που έχει δει εκτός από ένα, εκείνο για το οποίο η ποσότητα $\frac{\beta_{ik}}{p_{ik}} = \frac{\Sigma_{l \notin S_i} \rho_{kl} \;\; \Sigma_{j \neq i : l \in S_j} q_{jl}}{p_{ik}}$ (8) γίνεται η ελάχιστη. Η ποσότητα αυτή γίνεται ελάχιστη όταν είτε ο παρονομαστής γίνει πολύ μεγάλος που για εμάς σημαίνει το αντικείμενο με την υψηλότερη βαθμολογία από τον χρήστη είτε όταν ο αριθμητής γίνει πολύ μικρός κάτι που μπορεί να συμβεί αν το αντικείμενο έχει μικρό βαθμό συσχέτισης με τα αντικείμενα που δεν έχει δει ακόμη ο χρήστης και για τα οποία το άθροισμα των δηλωθέντων βαθμολογιών είναι χαμηλό.

ΚΕΦΑΛΑΙΟ 4

4.1 ΥΛΟΠΟΙΗΣΗ

Η υλοποίηση έγινε σε γλώσσα προγραμματισμού C σε λειτουργικό περιβάλλον Linux. Χωρίζεται στη δημιουργία κάποιων βιβλιοθηκών και χρήση κάποιων έτοιμων, στην υλοποίηση επιμέρους συναρτήσεων και στην υλοποίηση της κύριας συνάρτησης.

4.1.1 ΕΠΙΛΥΣΗ ΠΡΟΒΛΗΜΑΤΟΣ ΓΡΑΜΜΙΚΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ

Βασικό σημείο της υλοποίησης είναι η επίλυση του προβλήματος βελτιστοποίησης για κάθε χρήστη (Πρόβλημα 3). Για αυτό το σκοπό χρησιμοποιήσαμε τη βιβλιοθήκη GLPK[8] (GNU Linear Programming Kit) μέρος του εγχειρήματος GNU κατάλληλη για την επίλυση γραμμικών προβλημάτων μεγάλης κλίμακας με ρουτίνες βασισμένες στον αναθεωρημένο αλγόριθμο Simplex. Το πακέτο είναι διαθέσιμο σε LINUX και αφού εγκαταστήσαμε την βιβλιοθήκη είχαμε πρόσβαση σε όλες τις ρουτίνες εισάγοντας στο αρχείο μας το header file glpk.h. Η συνάρτηση αυτή υλοποιήθηκε όσο πιο ανεξάρτητα γινόταν από το συγκεκριμένο πρόβλημα και μπορεί να λύσει οποιασδήποτε μορφής LP προβλήματα αρκεί να δίνεται το πλήθος των αγνώστων, το πλήθος των περιορισμών, η αντικειμενική συνάρτηση και τα δεδομένα που θα γεμίσουν τον πίνακα των περιορισμών. Το πρόβλημα μοντελοποιήθηκε ως εξής: ορίζουμε ένα δισδιάστατο πίνακα διαστάσεων $|U| \times |I|$ (Utility matrix). Κάθε κελί του πίνακα διατηρεί τη βαθμολογία του χρήστη u (από 1 έως 5) για το αντικείμενο i και την τιμή 0 σε περίπτωση που δεν έχει αξιολογήσει το αντικείμενο. Αυτός ο πίνακας διατηρεί τις πραγματικές βαθμολογίες των χρηστών (private rating vector). Θα εισάγουμε και δεύτερο πίνακα που θα περιέχει τις δηλωθέντες βαθμολογίες στον server όπως αυτές θα προκύψουν από τη λύση του LP προβλήματος (declared rating vector). Η συνάρτηση παίρνει ως ορίσματα το διάνυσμα με τις πραγματικές βαθμολογίες p_i γιατί είναι οι συντελεστές της αντικειμενικής συνάρτησης, το πλήθος των αντικειμένων που έχει βαθμολογήσει ο χρήστης που αποτελεί και τις μεταβλητές του προβλήματος, μια μεταβλητή που ορίζει το πλήθος των περιορισμών για εμάς είναι ένας και ένα διάνυσμα που περιέχει τα β_{ik} όπως υπολογίστηκαν στο Πρόβλημα 3 και θα γεμίσουν τον πίνακα των περιορισμών. Η συνάρτηση επιστρέφει ένα διάνυσμα με τις λύσεις του προβλήματος, την τιμή της αντικειμενικής συνάρτησης και το status της λύσης ώστε να εξετάσουμε κατά πόσο βρέθηκε η βέλτιστη (optimal) λύση. Στην συνέχεια υπολογίζουμε με βάση τη σχέση 7 το διάνυσμα που θα δηλώσει στον server q_i.

4.1.2 ΥΠΟΛΟΓΙΣΜΟΣ ΟΜΟΙΟΤΗΤΑΣ ΜΕΤΑΞΥ ΑΝΤΙΚΕΙΜΕΝΩΝ

Ως συντελεστή συσχέτισης μεταξύ δύο αντικειμένων θα ορίσουμε τον δείκτη ομοιότητας μεταξύ

[8] http://www.gnu.org/software/glpk/

τους. Η συνάρτηση ομοιότητας μεταξύ δύο αντικειμένων i και j sim(i,j) είναι μια μετρική απόστασης. Οι πιο δημοφιλείς μετρικές είναι η correlation (όπως συντελεστής Pearson) και η cosine-based. Η προσέγγιση που ακολουθήσαμε είναι βασισμένη στις βαθμολογίες χρηστών που είχαν αλληλεπιδράσει και με τα δύο αντικείμενα. Τα σύνολα δεδομένων που χρησιμοποιήσαμε είναι πολύ αραιά (sparse) πράγμα που δυσκόλεψε τον υπολογισμό της ομοιότητας μεταξύ των αντικειμένων για αυτό σε κάποιες περιπτώσεις κάναμε κάποιες παραδοχές και εισάγαμε τυχαίους δεκαδικούς αριθμούς μεταξύ 0 και 1. Το πρόβλημα έγκειται στο γεγονός ότι στην πραγματικότητα για τα περισσότερα αντικείμενα δεν υπάρχουν αξιολογήσεις και ακόμα λιγότερα είναι εκείνα που έχουν βαθμολογηθεί από κοινούς χρήστες. Εν συνεχεία, θα παρουσιάσουμε τα μέτρα ομοιότητας που χρησιμοποιήσαμε.

Πίνακας 3.Υπολογισμός ομοιότητας μεταξύ αντικειμένων

Items / Users	1		i		j			m
1			r		?			
2			r		r			
			r		r			
			?		r			
n			r		r			

Συντελεστής Pearson

$$sim\,(i\,,j\,) \;=\; \frac{\sum_{u\in U}\left(r_{u,i}-\bar{r}_i\right)\times\left(r_{u,j}-\bar{r}_j\right)}{\sqrt{\sum_{u\in U}\left(r_{u,i}-\bar{r}_i\right)^2}\;\sqrt{\sum_{u\in U}\left(r_{u,j}-\bar{r}_j\right)^2}}$$

όπου U είναι το σύνολο των χρηστών που έχουν βαθμολογήσει τα αντικείμενα i και j, το $r_{u,i}$ δηλώνει την βαθμολογία του χρήστη u για το αντικείμενο i, $\bar{r}_i$ ο μέσος όρος των βαθμολογιών για το i-οστό αντικείμενο. Για τον υπολογισμό του συντελεστή Pearson χρησιμοποιήσαμε την έτοιμη συνάρτηση από το πακέτο στατιστικών συναρτήσεων της βιβλιοθήκης GNU Scientific Library (GSL)[9].

double gsl_stats_correlation (const double data1[], const size_t stride1, const double data2[], const size_t stride2, const size_t n)

όπου datax[] είναι το διάνυσμα με τις βαθμολογίες του αντικειμένου x, stride1 είναι μία σταθερά ίση με 1 και *n* είναι το μέγεθος των διανυσμάτων.

Σημειώνουμε ότι στα δύο διανύσματα περιλαμβάνονται οι βαθμολογίες μόνο από χρήστες που έχουν αξιολογήσει και τα δύο αντικείμενα οπότε έχουν το ίδιο μέγεθος. Για να επιστρέψει τιμή διαφορετική του null απαιτείται τα δυο αντικείμενα να έχουν βαθμολογηθεί από τουλάχιστον τρεις κοινούς χρήστες διαφορετικά αναθέτουμε μία τυχαία δεκαδική τιμή στο διάστημα [0,1]. Για

[9] http://www.gnu.org/software/gsl

να περιορίσουμε τον χρόνο των υπολογισμών θέτουμε ένα άνω όριο στον αριθμό των κοινών χρηστών που έχουν βαθμολογήσει ένα αντικείμενο ίσο με 50.

Cosine-based ομοιότητα

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

Αναφέρεται για λόγους πληρότητας και δεν υλοποιήθηκε στα πλαίσια της παρούσας εργασίας.

Ευκλείδεια απόσταση:

$$dist(i,j) = \frac{\sqrt{\sum_{u \in U} (r_{u,i} - r_{u,j})^2}}{N}$$

Κοντά στη λογική του συντελεστή Pearson. Στην περίπτωση όπου δεν υπήρχαν βαθμολογίες για κάποιο αντικείμενο θεωρείται ότι η απόστασή τους είναι μηδενική και εξαιτίας της φύσης των δεδομένων υπάρχουν πολλές μηδενικές τιμές. Επίσης, για να περιορίσουμε το χρόνο υπολογισμού και συνολικά της εκτέλεσης αναζητούμε το πολύ μέχρι 100 κοινούς χρήστες αν και για την πλειοψηφία των αντικειμένων δεν υπάρχουν περισσότεροι.
Σημειώνουμε ότι η απόσταση είναι το αντίστροφο της ομοιότητας δηλαδή όσο πιο μεγάλη είναι η απόσταση μεταξύ δυο αντικειμένων τόσο πιο μικρή είναι η ομοιότητα τους.

Μέτρο ομοιότητας Jaccard:

$$Jaccard(i,j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|}$$

Για να υπολογίσουμε το μέτρο ομοιότητας Jaccard βασιστήκαμε στην πληροφορία που είχαμε στα σύνολα δεδομένων για το είδος στο οποίο ανήκει κάθε ταινία. Δηλώνουμε με G_i, G_j το σύνολο με τα είδη που χαρακτηρίζουν την ταινία i και j αντίστοιχα. Εδώ δεν υπάρχει περίπτωση αδυναμίας υπολογισμού του συντελεστή. Και πάλι υπάρχουν πολλές μηδενικές τιμές.

Βεβαρημένος συντελεστής Pearson:

$$wsim(i,j) = Jaccard(i,j) * \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i) \times (r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in U}(r_{u,j} - \bar{r}_j)^2}}$$

Στην περίπτωση αυτή ο συντελεστής Jaccard παίζει το ρόλο του βάρους. Προτάθηκε για να εξισορροπήσει τα μειονεκτήματα των παραπάνω συντελεστών στην περίπτωση της συσχέτισης χρήστη με χρήστη (user-to-user correlation). Είναι κατάλληλος για τον υπολογισμό της συσχέτισης μεταξύ χρηστών. Ο συντελεστής Pearson μόνος του παρουσιάζει αδυναμίες λόγω έλλειψης βαθμολογιών (π.χ. αν δυο χρήστες με διαφορετικά γούστα έτυχε να έχουν βαθμολογήσει μόνο μια κοινή ταινία φαίνεται ότι είναι απολύτως όμοιοι) και ο συντελεστής Jaccard λαμβάνοντας υπόψη μόνο κατά πόσο μοιάζουν τα είδη ταινίας που έχουν αξιολογήσει οι χρήστες και όχι την βαθμολογία που έχουν δώσει (π.χ. αν έχουν αντίθετες απόψεις για τα ίδια είδη ταινιών) καθιστούν λανθασμένα όμοιους τους χρήστες. Για να εκμεταλλευτούμε την συμπληρωματικότητα τους τούς συνδυάζουμε.
Στην υλοποίηση μας αρκούν τα τρία πρώτα μέτρα ομοιότητας γιατί επικεντρωνόμαστε στην ομοιότητα μεταξύ αντικειμένων και ο τυχαίος αριθμός που εισάγουμε όταν είναι αναγκαίο δεν επηρεάζει σημαντικά τον στόχο της υλοποίησης μας.

4.2 ΑΞΙΟΛΟΓΗΣΗ

Για την αξιολόγηση του αλγορίθμου χρησιμοποιήθηκαν πραγματικά δεδομένα που υπάρχουν διαθέσιμα στο διαδίκτυο. Τα δεδομένα είναι ανώνυμα και ενδείκνυνται για offline αξιολόγηση.

4.2.1 ΔΕΔΟΜΕΝΑ

❖ Το πρώτο σύνολο είναι το MovieLens100K[10] που συλλέχθηκε από το ερευνητικό πρόγραμμα GroupLens στο πανεπιστήμιο της Μινεσότα κατά τη διάρκεια επτά μηνών (19 Σεπτέμβριου 1997 - 22 Απρίλιος 1998) μέσω του ιστοχώρου MovieLens.com.
Το σύνολο περιλαμβάνει:

- 100.000 ακέραιες βαθμολογίες σε κλίμακα από 1 έως 5 από 943 χρήστες για 1682 ταινίες. Κάθε χρήστης έχει βαθμολογήσει τουλάχιστον 20 ταινίες.
- Αρχείο με απλές δημογραφικές πληροφορίες για τους χρήστες (φύλο, επάγγελμα, ηλικία). Ωστόσο η πληροφορία αυτή δεν χρησίμευσε σε κάτι.
- Αρχείο με τα είδη στα οποία ανήκει κάθε ταινία όπως αυτά αναφέρονται στην ιστοσελίδα του Internet Movie Database (IMDB).

Τα αρχεία είναι σε text μορφή και αποτέλεσαν είσοδο για το πρόγραμμά μας. Το αρχείο που περιλαμβάνει τις βαθμολογίες έχει την εξής μορφή (λίστα χωρισμένη με tab): user id | item id | rating | timestamp.

Το αρχείο με τα είδη της ταινίας είναι λίστα της μορφής: movie id | movie title | release date | video release date | IMDb URL | unknown | Action | Adventure | Animation |Children's | Comedy | Crime | Documentary | Drama | Fantasy |Film-Noir | Horror | Musical |Mystery | Romance | Sci-Fi |Thriller | War | Western |.
Τα τελευταία 19 πεδία είναι τα διαθέσιμα είδη ταινιών και αν υπάρχει 1 σε κάποιο πεδίο σημαίνει ότι η ταινία ανήκει στο συγκεκριμένο είδος ενώ 0 ότι δεν ανήκει. Για τις ανάγκες του προγράμματος μας χρειάστηκε να χωρίσουμε τα δεδομένα σε δύο σύνολα ένα σύνολο εκπαίδευσης (training dataset) με το οποίο θα εκπαιδεύσουμε τον αλγόριθμο και ένα σύνολο με βάση το οποίο θα τον αξιολογήσουμε (test dataset). Στο σύνολο MovieLens υπάρχουν ήδη τέτοια αρχεία που δημιουργήθηκαν μετά από τυχαίο διαχωρισμό και σε ένα από αυτά θα εφαρμόσουμε τον αλγόριθμό μας. Σημειώνουμε ότι για να περιορίσουμε την υπολογιστική πολυπλοκότητα και κυρίως τη χρονική καθυστέρηση λαμβάνουμε υπόψη για κάθε χρήστη μέχρι ένα συγκεκριμένο αριθμό ταινιών (όριο 300 ταινίες).

❖ Το δεύτερο σύνολο είναι μια επέκταση του MovieLens1M[11] συνόλου καθώς περιλαμβάνει επιπλέον πληροφορίες για τις ταινίες (tags, links) από τις ιστοσελίδες Internet Movie Database

[10] http://www.grouplens.org/node/73

[11] http://www.grouplens.org/node/462

και Rotten Tomatoes. Δημοσιεύθηκε στα πλαίσια του 2ου International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec2011) στο 5ο ACM Conference on Recommender Systems (RecSys 2011). Το σύνολο περιλαμβάνει:

- 2113 χρήστες και 10197 ταινίες.
- Αρχεία σε text μορφή με πληροφορίες για τις ταινίες, τους ηθοποιούς, σκηνοθέτες, χώρα παραγωγής κλπ. Εμείς θα χρησιμοποιήσουμε μόνο το αρχείο που αντιστοιχίζει κάθε ταινία με το είδος/-η στο οποίο ανήκει και είναι στην ίδια μορφή με το πρώτο σύνολο.
- Αρχεία με tag assignments τα οποία στα πλαίσια του αλγορίθμου δεν μας είναι χρήσιμα.
- Αρχείο με 855.598 βαθμολογίες των χρηστών για ταινίες στο διάστημα από 0.5 έως 5 με βήμα 0.5.

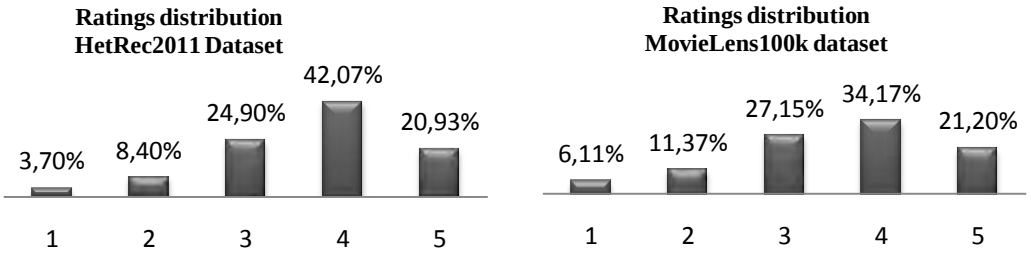
Σημειώνουμε ότι για λόγους υλοποίησης δεν μπορούμε να χρησιμοποιήσουμε τόσο μεγάλο πλήθος ταινιών και θα χρησιμοποιήσουμε μέρος των διαθέσιμων (2000 ταινίες). Δημιουργούμε ένα αρχείο με μορφή user id | item id | rating | timestamp λαμβάνοντας υπόψη μόνο βαθμολογίες από τυχαίους χρηστών για αυτές τις ταινίες οι οποίοι έχουν βαθμολογήσει τουλάχιστον 30 ταινίες αλλά το πολύ 300. Επίσης θα περιορίσουμε και τον αριθμό των χρηστών που θα λάβουν μέρος στο πείραμα μας (992 χρήστες).
Μόνο για τον υπολογισμό των συντελεστών συσχέτισης (του συντελεστή Pearson και της απόστασης) χρησιμοποιήσαμε ολόκληρα τα αρχεία με τις βαθμολογίες χωρίς τους περιορισμούς σε χρήστες ή σε ταινίες που έχει δει ο χρήστης έτσι ώστε να έχουμε όσο πιο ακριβή αριθμητική τιμή γίνεται και λιγότερους τυχαίους αριθμούς.

Συνοπτικά παρουσιάζονται παρακάτω οι πληροφορίες για τα δεδομένα:

Πίνακας 4. Χαρακτηριστικά δεδομένων

Σύνολο	Περιγραφή	Χρήστες	Ταινίες	Πυκνότητα	Μέση βαθμολογία
Movielens100K	100000 βαθμολογίες για ταινίες από το Movielens Σ.Σ	943	1682	6.3%	3,53
HetRec2011-Movielens2K	152629 βαθμολογίες για ταινίες από Movielens, IMDb, Rotten Tomatoes	992	2000	8.3%	3,49

Πίνακας 5. Κατανομή βαθμολογιών στα σύνολα δεδομένων



Ως πυκνότητα ορίζεται ο λόγος $Density = \frac{\#\ of\ nonzero\ \ entries}{Total\ \#\ of\ entries}$ όπου ο παρονομαστής ορίζεται ως το γινόμενου του πλήθους των χρηστών επί το πλήθος των αντικειμένων για τα οποία υπάρχει τουλάχιστον μια βαθμολογία και ο αριθμητής είναι ο συνολικός αριθμός των βαθμολογιών που υπάρχουν στο σύνολο δεδομένων (Sarwar, 2001).

4.2.2 ΥΠΟΛΟΓΙΣΜΟΣ ΜΕΤΡΩΝ ΑΞΙΟΛΟΓΗΣΗΣ

Ένα από τα πιο διαδεδομένα κριτήρια αξιολόγησης ενός αλγορίθμου παραγωγής συστάσεων είναι η ακρίβεια του στις προβλέψεις. Τροποποιώντας τις πραγματικές βαθμολογίες των χρηστών ακολουθώντας την προτεινόμενη μέθοδο μας με σκοπό να προστατέψουμε τα προσωπικά δεδομένα εισάγουμε ένα σφάλμα ακρίβειας στο σύστημα συστάσεων. Συνήθως μετράται ως το μέγεθος του σφάλματος ανάμεσα στην προβλεπόμενη και στην πραγματική βαθμολόγηση. Ωστόσο, δεδομένου ότι τα συστήματα συστάσεων επιστρέφουν στο χρήστη λίστες με προτεινόμενα αντικείμενα και όχι βαθμολογίες τίθεται το ερώτημα γιατί να μένουμε στην αριθμητική τιμή που οδήγησε το αντικείμενο στη λίστα των συστάσεων και όχι στην ποιότητα του ίδιου του αντικειμένου. Στην υλοποίηση μας για την εκτίμηση της ακρίβειας θα εφαρμόσουμε το Απόλυτο Μέσο Σφάλμα (Mean Absolute Error) και την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Square Error). Το MAE είναι η μέση απόλυτη διαφορά ανάμεσα στην προβλεπόμενη βαθμολογία και στη πραγματική βαθμολογία που έχει δώσει ο χρήστης. Για τις ανάγκες της αξιολόγησης έχουμε χωρίσει το σύνολο δεδομένων σε training set και test set τα οποία δημιουργούνται με τον διαχωρισμό 80-20 αντίστοιχα. Η πραγματική βαθμολογία συνεπώς βρίσκεται στο test set.

$$MAE = \frac{\sum_{r_i \in R_{test}} |\, pr_i - r_i|}{|R_{test}\,|}$$

όπου $\boldsymbol{pr_i}$ είναι η προβλεπόμενη βαθμολόγηση και $\boldsymbol{r_i}$ η πραγματική βαθμολόγηση για το αντικείμενο i.

Επίσης εναλλακτικά μπορούμε να χρησιμοποιήσουμε και την τετραγωνική ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Square Error). Το RMSE είναι πιο αυστηρό για τα μεγάλα σφάλματα διότι υψώνει στο τετράγωνο την απόλυτη διαφορά μεταξύ των βαθμολογιών. Συνεπώς είναι πιο κατάλληλο για τις περιπτώσεις όπου μικρά σφάλματα στις προβλέψεις δεν είναι πολύ σημαντικά.

$$RMSE = \sqrt{\frac{\sum_{r_i \in R_{test}} (pr_i - r_i)^2}{|R_{test}\,|}}$$

Τα παραπάνω μετρικά είναι απλά στην υλοποίηση και κατανοητά προς τον χρήστη.

Το σφάλμα πρόβλεψης παρουσιάζει κάποιους περιορισμούς καθώς ελέγχει μόνο την ακρίβεια στις αριθμητικές τιμές. Σε πραγματικά συστήματα συστάσεων αυτό που έχει σημασία είναι να επιστρέφει και αντικείμενα κατάλληλα και σχετικά (relevant) για κάθε χρήστη. Για αυτό υπολογίσαμε και κάποια άλλα μεγέθη που συνήθως χρησιμοποιούνται για την αξιολόγηση συστημάτων που επιστρέφουν ταξινομημένες λίστες συστάσεων δηλαδή με τα N υψηλότερα βαθμολογημένα αντικείμενα προς σύσταση (top-N RSs). Τα συνηθέστερα είναι η ακρίβεια της κατάταξης (precision), η πληρότητα (recall) και ο αρμονικός μέσος των προηγούμενων (F-measure). Ο δικός μας αλγόριθμος όπως υλοποιήθηκε επιστρέφει στον χρήστη ολόκληρο το διάνυσμα με τις συστάσεις. Για να αξιολογηθεί ως top-N σύστημα θα θεωρήσουμε ότι ο χρήστης λαμβάνει υπόψη μόνο τις συστάσεις για τα αντικείμενα που ανήκουν στο test set. Πρέπει να καθορίσουμε επίσης τη σημασία του «σχετικού». Για εμάς «σχετικό» θα είναι ένα αντικείμενο το οποίο μπορεί να αρέσει στον χρήστη. Για να αρέσει στο χρήστη σημαίνει ότι αν το είχε δει θα το είχε αξιολογήσει υψηλά. Οπότε θεωρούμε ότι βαθμολογία μικρότερη από 4 σημαίνει «δεν μου αρέσει» ενώ βαθμολογία μεγαλύτερη και ίση με 4 σημαίνει «μου αρέσει».
Αν δούμε το πρόβλημα σαν πρόβλημα κατηγοριοποίησης των πραγματικών βαθμολογιών σε σχέση με αυτές που προβλέφθηκαν προκύπτει ο παρακάτω πίνακας (Confusion Matrix).

Πίνακας 6. Confusion Matrix

Predicted / Actual	Like	Dislike
Like	True Positive (TP)	False Negative (FN)
Dislike	False Positive (FP)	True Negative (TN)

Σύμφωνα με τον παραπάνω πίνακα το precision και το recall ορίζονται ως εξής:

$$Recall = \frac{TP}{FN+TP}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Fmesasure = 2 * \frac{precision * recall}{precisi\ on + recall}$$

Για την δική μας εφαρμογή αυτό που θα εκτιμήσουμε με το μέγεθος precision είναι τι ποσοστό των αντικειμένων που προτάθηκαν και προβλέφθηκε ότι αρέσουν, αρέσουν πραγματικά στο χρήστη δηλαδή τα έχει βαθμολογήσει με 4 ή 5 αστέρια στο test set. Το recall από την άλλη εκφράζει πόσα από τα αντικείμενα που αρέσουν στο χρήστη προέβλεψε το σύστημα με επιτυχία. Ωστόσο, επειδή συχνά οι χρήστες δεν αναφέρουν όλα τα αντικείμενα που τους αρέσουν (η πληροφορία αυτή μπαίνει στον παρονομαστή) το recall μπορεί να μην υπολογίζεται επακριβώς και υπάρχουν ενστάσεις για τον τρόπο υπολογισμού του και την εγκυρότητα των συμπερασμάτων.

5.1 ΑΡΙΘΜΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στον Πίνακα 8 παρουσιάζονται τα αριθμητικά αποτελέσματα όπως προέκυψαν από τα πειράματα που έγιναν στα σύνολα MovieLens100K και HetRec2011 αφού τα χωρίσαμε σε δύο μέρη με τον κανόνα 80-20: το ένα σύνολο (80% του αρχικού) αποτέλεσε το σύνολο εκπαίδευσης και το υπόλοιπο 20% χρησίμευσε για την αξιολόγηση του αλγορίθμου (Πίνακας 7). Στόχος είναι να εξετάσουμε κατά πόσο προβλέφθηκαν με ακρίβεια τα αντικείμενα τα οποία βρίσκονται στο test set και έχουν κρατηθεί σκόπιμα από το αρχικό σύνολο. Υποθέτουμε ότι αν η μέθοδος συμπεριφερθεί καλύτερα στην πρόβλεψη των βαθμολογιών για τα αντικείμενα που ανήκουν στο test set θα λειτουργήσει εξίσου καλά στην ανακάλυψη νέων αντικειμένων άγνωστα μέχρι τότε στον χρήστη.

Πίνακας 7. Διαχωρισμός συνόλων δεδομένων

Σύνολο δεδομένων	**MovieLens100K**	**HetRec2011**
Training Set	90570	122877
Test Set	9430	29752
Συνολικά (σε βαθμολογίες)	**100000**	**152629**

Πίνακας 8. Ακρίβεια πρόβλεψης της μεθόδου (D=2)

	Υπολογισμός Συσχέτισης	**Ακρίβεια (Accuracy)**	
		MAE	RMSE
MovieLens100K	Pearson's correlation	0,70	1,39
	Euclidean Distance	0,84	1,52
	Jaccard coefficient	-	-
HetRec2011	Pearson's correlation	0,59	1,26
	Euclidean distance	1,04	1,72
	Jaccard coefficient	0,82	1,57

Στην περίπτωση του Jaccard Coefficient για το πρώτο σύνολο δεδομένων δεν ήταν δυνατή η λήψη αριθμητικών αποτελεσμάτων καθώς η διαδικασία δεν συνέκλινε σε λογικό αριθμό επαναλήψεων. Η διαδικασία ανταλλαγής πληροφοριών μεταξύ server και χρηστών έγινε 12 φορές

και συνέχιζε να επαναλαμβάνεται, ενώ για τις υπόλοιπες περιπτώσεις και για τα δύο σύνολα ολοκληρώθηκε σε 3 με 7 επαναλήψεις. Συμπεραίνουμε ότι πιθανότατα οφείλεται στη δομή του συνόλου και όχι στην υλοποίηση διότι για το άλλο σύνολο η διαδικασία ολοκληρώθηκε με επιτυχία. Όπως αναμέναμε το RMSE δίνει σε όλες τις περιπτώσεις μεγαλύτερη τιμή σφάλματος από το MAE για τους λόγους που αναφέραμε στην ενότητα 4.2.2.
Αν και το MAE χρησιμοποιείται ευρύτατα ως δείκτης ποιότητας της σύστασης που παρέχει ένα CF σύστημα συστάσεων παρουσιάζει μια βασική αδυναμία, λαμβάνει υπόψη μόνο τις απόλυτες διαφορές στις βαθμολογίες. Για παράδειγμα αν έχουμε (1,5) και (5,1) δυο ζεύγη (πραγματικής βαθμολογίας , προβλεπόμενης βαθμολογίας) για μια ταινία, δεν αναγνωρίζει κάποια διαφορά. Ωστόσο, στην πρώτη περίπτωση ο χρήστης θα είναι πιο δυσαρεστημένος διότι το σύστημα προέβλεψε ότι θα του αρέσει πάρα πολύ μια ταινία την οποία στην πραγματικότητα έχει βαθμολογήσει πολύ χαμηλά.

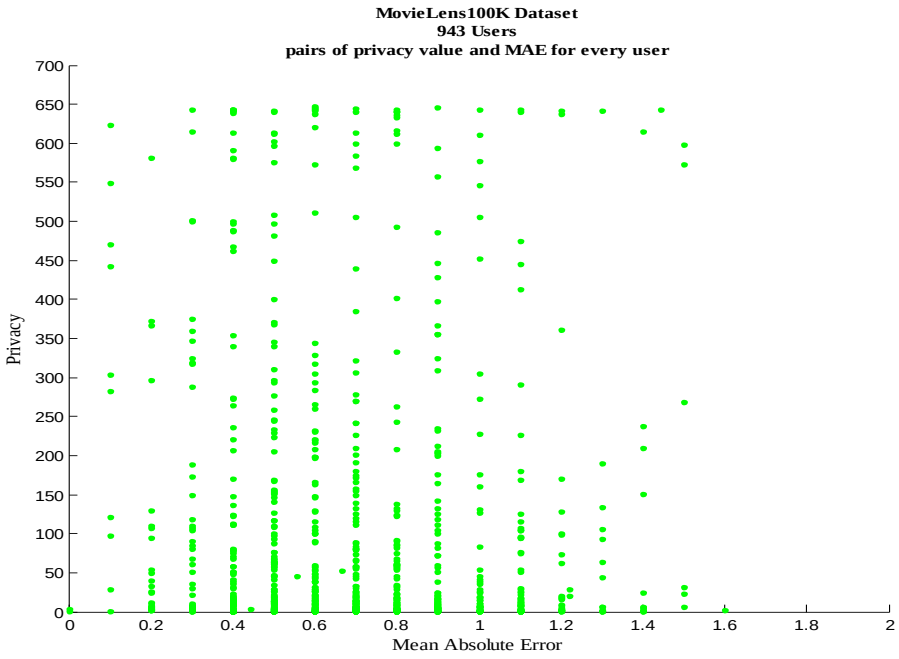
Για να περιορίσουμε αυτή την αδυναμία εφαρμόσαμε και κάποια άλλα μέτρα αξιολόγησης που εξετάζουν κατά πόσο ένα σύστημα συστάσεων μπορεί να προτείνει κατάλληλα αντικείμενα, που να απευθύνονται στα προσωπικά ενδιαφέροντα του κάθε χρήστη, από το σύνολο των διαθέσιμων. Στον Πίνακα 9 παρουσιάζονται τα αριθμητικά αποτελέσματα για τα ίδια σύνολα. Επισημαίνουμε ότι υψηλό precision δεν σημαίνει απαραίτητα και χρήσιμη για τον χρήστη σύσταση γιατί μπορεί να προτείνει συνέχεια ταινίες blockbusters.

Πίνακας 9.Precision-Recall-Fmeasure (D=2)

	Υπολογισμός Συσχέτισης	**Precision**	**Recall**	**F-Measure**
MovieLens100K	Pearson's correlation	0,68	0,25	0,37
	Euclidean distance	0,57	0,36	0,44
	Jaccard coefficient	-	-	-
HetRec2011	Pearson's correlation	0,63	0,19	0,29
	Euclidean distance	0,47	0,07	0,10
	Jaccard coefficient	0,62	0,23	0,33

Στα παρακάτω γραφήματα παρουσιάζεται η κατανομή των τιμών για ζεύγη της μορφής (privacy,MAE) για κάθε χρήστη όπως προέκυψαν μετά την αξιολόγηση της προτεινόμενης μεθόδου στα δύο σύνολα δεδομένων.

Σχήμα 1. Κατανομή privacy-MAE MovieLens100K Dataset

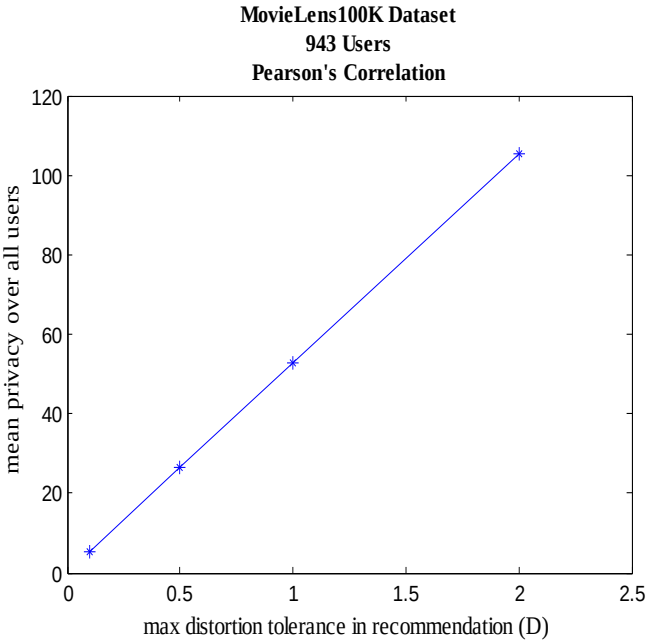


Σχήμα 2.Κατανομή privacy-MAE HetRec2011 Dataset

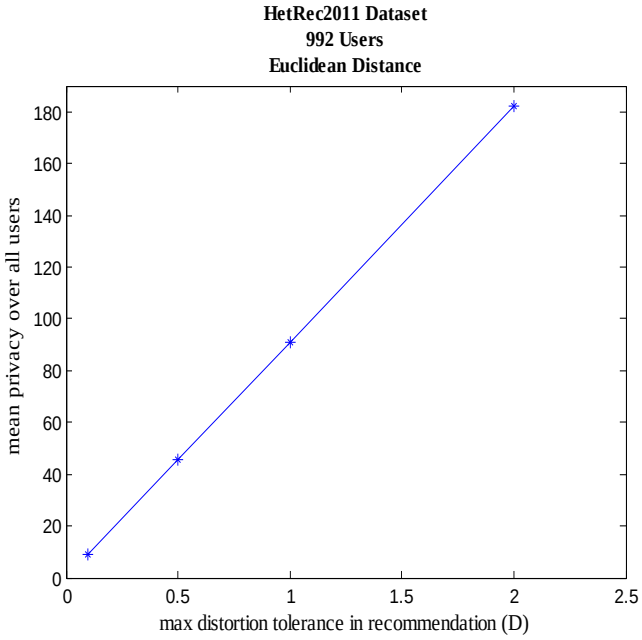
Όπως φαίνεται και στις γραφικές παραστάσεις το σφάλμα, δηλαδή η απόκλιση της πραγματικής βαθμολογίας από αυτή που προέβλεψε το σύστημα εφαρμόζοντας τη σχέση 3 έχοντας ολοκληρώσει το παίγνιο και αφού έχουν αποφασίσει όλοι τι θα δηλώσουν στον server, είναι μεταξύ 0 και 1.6 περίπου. Συνεπώς με αυτό τον τρόπο έχουμε και πολύ καλή ακρίβεια και υψηλές τιμές ιδιωτικότητας (privacy) καθώς για όλους τους χρήστες είναι μεγαλύτερη του μηδενός και για τους περισσότερους σε πολύ υψηλά επίπεδα.

Ενδιαφέρον ακόμη παρουσιάζει η σχέση της ιδιωτικότητας (privacy) και του D (το D εκφράζει τη μέγιστη παραμόρφωση στο recommendation vector). Υπολογίσαμε και για τα δυο σύνολα δεδομένων την μέση τιμή της ιδιωτικότητας ($\frac{1}{N}(\sum_{i=1}^{N} privacy_i)$) για διαφορετικές τιμές του D (D=0.1,0.5,1,2). Συμπεραίνουμε ότι όσο αυξάνει το D, αυξάνει και αριθμητικά η τιμή της ιδιωτικότητας, και αυτό δικαιολογείται από το γεγονός ότι όσο πιο αυστηρός είναι ένας χρήστης σε σχέση με την απόκλιση στη σύσταση που θα λάβει τόσο περισσότερες πληροφορίες σχετικά με τις πραγματικές του προτιμήσεις πρέπει να δώσει που συνεπάγεται χαμηλότερη τιμή privacy.

Σχήμα 3.Privacy για διαφορετικές τιμές του D MovieLens100K Dataset

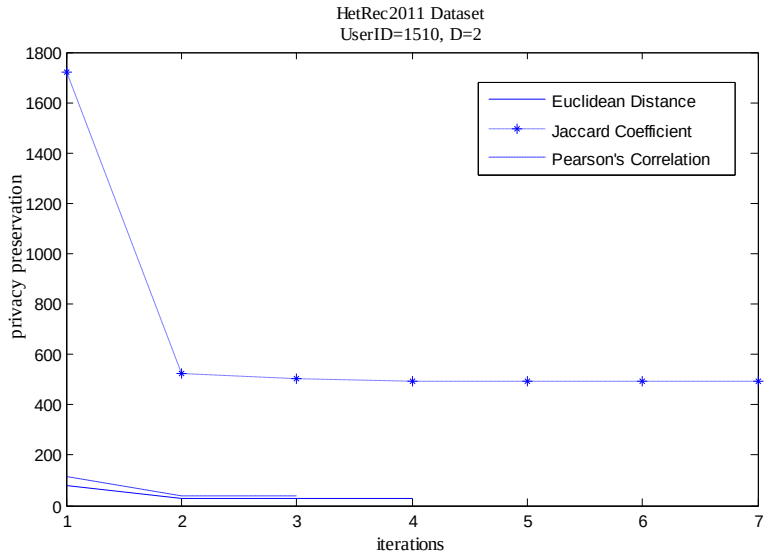


Σχήμα 4.Privacy για διαφορετικές τιμές του D HetRec2011 Dataset



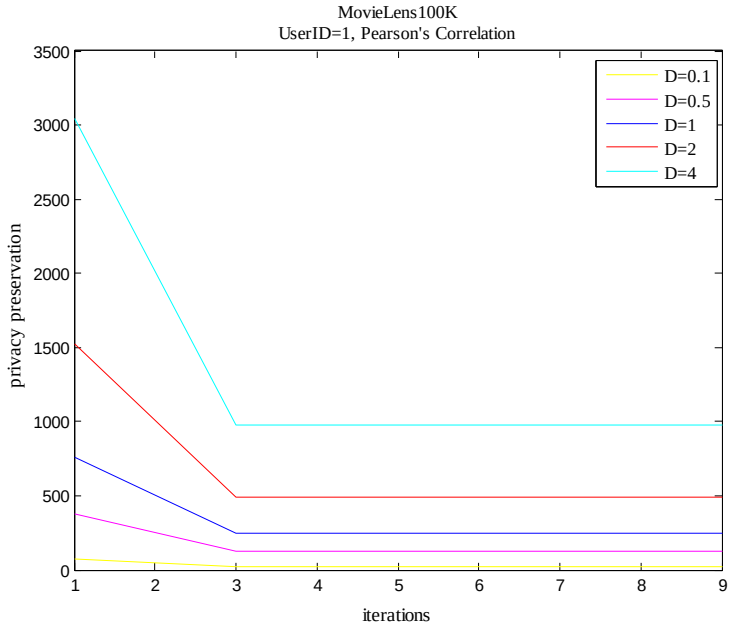
Η προσέγγιση μας στηρίζεται σε μια επαναληπτική διαδικασία κατά την οποία οι χρήστες επικοινωνούν και ανταλλάσουν δεδομένα με τον server για αυτό θα παρουσιάσουμε κάποια στοιχεία για την σύγκλιση της μεθόδου. Παρακάτω απεικονίζεται η σύγκλιση της μεθόδου ανάλογα το μέτρο ομοιότητας που εφαρμόσαμε για την συσχέτιση μεταξύ των αντικειμένων για έναν συγκεκριμένο χρήστη και D=2. Παρατηρούμε ότι η σύγκλιση με το jaccard coefficient είναι η πιο αργή ενώ στις άλλες περιπτώσεις αρκούν 3,4 επαναλήψεις για να αποφασίσουν όλοι οι χρήστες τα declared ratings vector τους.

Σχήμα 5.Σύγκλιση για διαφορετικά μέτρα ομοιότητας



Επίσης απεικονίζουμε την τιμή του privacy για έναν συγκεκριμένο χρήστη σαν συνάρτηση της μέγιστης παραμόρφωσης στη σύσταση D. Παρατηρούμε ότι όσο η ανοχή του χρήστη στην παραμόρφωση της σύστασης που θα λάβει αυξάνει, τόσο αυξάνει και η τιμή του privacy. Συνεπώς αν κάποιος χρήστης είναι λιγότερο ανεκτικός σε ότι αφορά το σφάλμα στη σύσταση που θα λάβει, θα πρέπει να αποκαλύψει περισσότερες πληροφορίες σχετικά με τις προτιμήσεις του (π.χ. να δηλώσει τις πραγματικές βαθμολογίες ή όσο πιο κοντά σε αυτές για τις ταινίες που έχει αξιολογήσει).

Σχήμα 6.Σύγκλιση για διαφορετικές τιμές του D



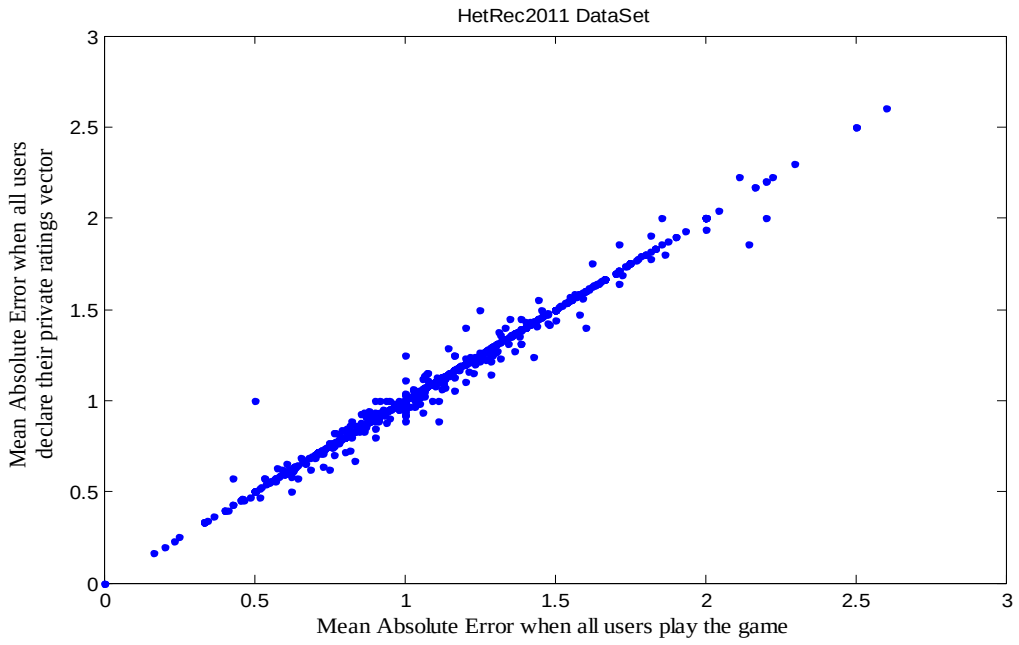
Κάτι ακόμη που πρέπει να σημειωθεί είναι το γεγονός ότι η διαδικασία που ακολουθούν οι χρήστες παίζοντας το παίγνιο για να αποφασίσουν τελικά τι δεδομένα θα δηλώσουν στον server δεν επηρεάζει σχεδόν καθόλου την πρόβλεψη που θα λάβουν από το σύστημα καταφέρνοντας

ταυτόχρονα να διασφαλίσουν την ιδιωτικότητα τους. Όπως φαίνεται και παρακάτω αν μετρήσουμε το MAE για κάθε χρήστη όταν όλοι δηλώσουν τις πραγματικές τους βαθμολογίες και κατ'επέκταση έχουν μηδενικό privacy και το συγκρίνουμε με το MAE όπως προκύπτει για καθένα τους μετά και από την ολοκλήρωση του παιγνίου προκύπτει σχεδόν μια ευθεία γραμμή με μικρές αποκλίσεις. Ορίζουμε το παρακάτω μέγεθος

$$accuracy\ loss = \frac{MAE - RealMAE}{RealMAE}$$

όπου $RealMAE$ είναι το MAE όταν οι χρήστες δηλώσουν στον recommendation server τις πραγματικές βαθμολογίες τους και συνεπώς έχουν μηδενικό privacy. Ενδεικτικά σημειώνουμε ότι στην περίπτωση του συνόλου MovieLens100K το $accuracy\ loss = 2.3\%$.(Παρατήρηση: βέβαια είναι αναμενόμενο από τη στιγμή που στον περιορισμό εμείς ζητάμε η λαμβανόμενη σύσταση να μην αποκλίνει D=2 από αυτή που θα λαμβάναμε αν δήλωνε ο χρήστης την πλήρη αλήθεια.)

Σχήμα 7.Σχέση MAE όταν οι χρήστες ακολουθούν το παίγνιο και MAE όταν δηλώσουν τις πραγματικές τους βαθμολογίες



ΚΕΦΑΛΑΙΟ 6

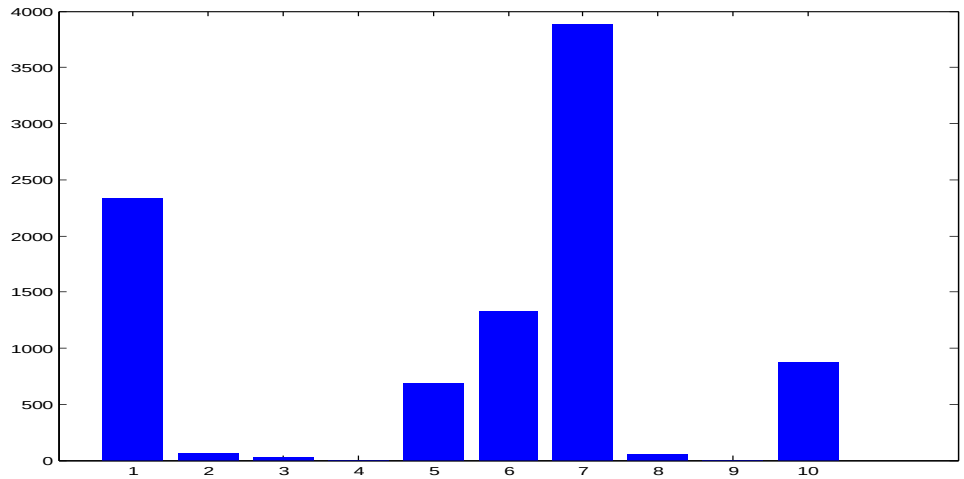
6.1 ΣΕΝΑΡΙΑ ΣΥΝΕΡΓΑΣΙΑΣ

Όπως θα αποδείξουμε παρακάτω οι χρήστες μπορούν να συνεργαστούν μεταξύ τους με τέτοιο τρόπο ώστε να επωφεληθούν αμφότεροι στη διατήρηση της ιδιωτικότητας τους.

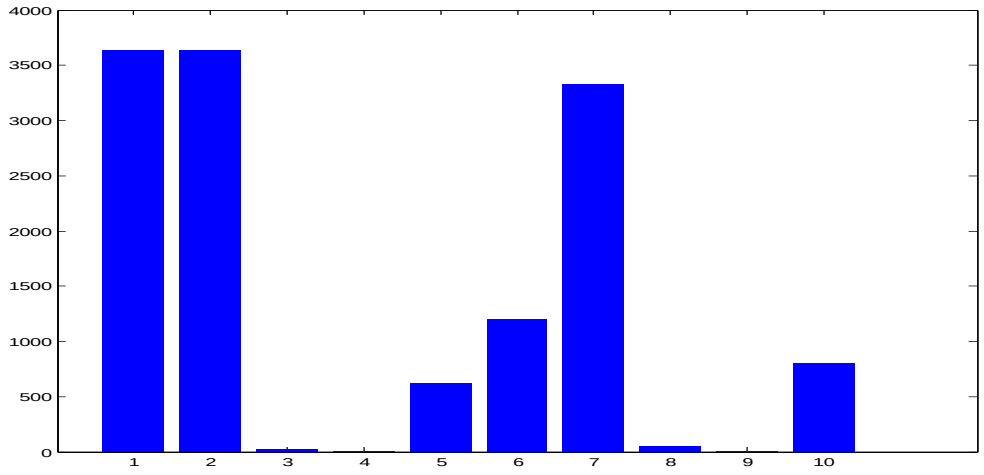
6.1.1 ΠΡΩΤΟ ΣΕΝΑΡΙΟ ΣΥΝΕΡΓΑΣΙΑΣ

Αρχικά εξετάσαμε την πιο απλή περίπτωση συνεργασίας μεταξύ δύο χρηστών οι οποίοι έχουν βαθμολογήσει αρκετά αντικείμενα και μεταξύ αυτών υπάρχουν και κάποια κοινά. Σημαντικό είναι να αναφέρουμε ότι οι χρήστες εμπιστεύονται ο ένας τον άλλον και μοιάζουν μεταξύ τους ως προς τις προτιμήσεις και τον τρόπο βαθμολόγησης. Η συνεργασία των χρηστών στηρίζεται στην επικοινωνία τους offline (δηλαδή των agents που τους εκπροσωπούν) πριν ξεκινήσει το παίγνιο και στην ανταλλαγή βαθμολογιών. Ο πρώτος χρήστης ενημερώνει τον δεύτερο για τις βαθμολογίες που έχει δώσει σε αντικείμενα που δεν έχει αξιολογήσει ο άλλος. Το ίδιο κάνει και ο δεύτερος. Για τα κοινά αντικείμενα ο καθένας διατηρεί την δική του βαθμολογία. Τελικά καταλήγουν να έχουν βαθμολογήσει το ίδιο πλήθος αντικειμένων. Στην συνέχεια ο καθένας λύνει το πρόβλημα βελτιστοποίησης (Πρόβλημα 3) όπως περιγράφηκε στην ενότητα 3.2.2 και συμμετέχουν στο παίγνιο. Μετά από πειράματα στα διαθέσιμα σύνολα δεδομένων για την περίπτωση που μόνο δύο χρήστες ανταλλάσουν βαθμολογίες μεταξύ τους και υπάρχουν ακόμη οκτώ χρήστες που συμμετέχουν στο παίγνιο, καταλήξαμε ότι το privacy για τον καθένα που συνεργάζεται είναι σημαντικά μεγαλύτερο σε σχέση με εκείνο που θα είχαν αν δεν συνεργάζονταν. Ωστόσο, μετά τη συνεργασία παρατηρείται επίσης μια πολύ μικρή μείωση στο privacy των υπόλοιπων χρηστών αλλά όχι πολύ σημαντική.

Σχήμα 8. Privacy για 10 χρήστες πριν την συνεργασία



Σχήμα 9. Privacy για 10 χρήστες μετά την συνεργασία των 1 και 2



6.1.2 ΔΕΥΤΕΡΟ ΣΕΝΑΡΙΟ ΣΥΝΕΡΓΑΣΙΑΣ

Ένας άλλος τρόπος συνεργασίας είναι να εκφράσουμε διαφορετικά το πρόβλημα βελτιστοποίησης. Θέλουμε να ορίσουμε έναν γενικό στόχο ο οποίος θα αντιπροσωπεύει την συνολική ιδιωτικότητα των συνεργαζόμενων χρηστών. Οι χρήστες συνεργάζονται έτσι ώστε να επιτύχουν αυτό το στόχο έχοντας πάλι κάποιους περιορισμούς ο καθένας, που πρέπει να ικανοποιούνται ταυτόχρονα. Λύνοντας το νέο πρόβλημα αν καταλήξουμε σε εφικτή λύση θα έχουμε βρει τη στρατηγική συνεργασίας η οποία θα είναι καλύτερη από άποψη ιδιωτικότητας (privacy-wiser) σε σύγκριση με το σημείο ισορροπίας Nash (NEP) που βρήκαμε προηγουμένως.

Ας ορίσουμε την πιο απλή περίπτωση των δυο χρηστών οι οποίοι έχουν βαθμολογήσει δύο αντικείμενα ο καθένας. Για τον χρήστη 1 έχουμε $S_1 = \{A, B\}$ με $p_1 = \{p_{1A}, p_{1B}\}$ και για τον χρήστη 2 $S_2 = \{B, C\}$ με $p_2 = \{p_{2B}, p_{2C}\}$. Θέτουμε $x_{ik} = (p_{ik} - q_{ik})^2$ για i=1,2 και k=A,B,C. Το πρόβλημα που θα αντιμετωπίσουν οι δύο χρήστες είναι το εξής:

$$\max_{x_{1A}, x_{1B}, x_{2B}, x_{2C}} p_{1A}x_{1A} + p_{1B}x_{1B} + p_{2B}x_{2B} + p_{2C}x_{2C}$$

$$s.t. \quad \rho_{AC}x_{1A}x_{2C} + \rho_{BC}x_{1B}x_{2C} \leq 2D$$
$$\rho_{AB}x_{2B}x_{1A} + \rho_{AC}x_{2C}x_{1A} \leq 2D$$

$$x_{1A}, x_{1B}, x_{2B}, x_{2C} \geq 0$$

Όπως φαίνεται το πρόβλημα παύει να είναι γραμμικό και για την επίλυση του θα χρησιμοποιήσουμε την Lagrangian μέθοδο και το θεώρημα Karush-Kuhn-Tucker (KKT). Το μη γραμμικό πρόβλημα παρουσιάζει μόνο περιορισμούς με ανισότητα οπότε θα εισάγουμε έξι συντελεστές Lagrange $\mu_1, \mu_2, \dots, \mu_6 \geq 0$.

Η συνάρτηση Lagrange είναι η εξής:

$L(x_{1A}, x_{1B}, x_{2B}, x_{2C}, \mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6) = p_{1A}x_{1A} + p_{1B}x_{1B} + p_{2B}x_{2B} + p_{2C}x_{2C} +$
$\mu_1(2D - \rho_{AC}x_{1A}x_{2C} - \rho_{BC}x_{1B}x_{2C}) + \mu_2(2D - \rho_{AB}x_{2B}x_{1A} - \rho_{AC}x_{2C}x_{1A})$
$με \ \mu_1, \mu_2, \dots, \mu_6 \geq 0$

Από το K.K.T θεώρημα:

- $\frac{\partial L}{\partial x_{1A}} = 0 \Rightarrow p_{1A} - \mu_1 \rho_{AC} x_{2C} - \mu_2 (\rho_{AB} x_{2B} + \rho_{AC} x_{2C}) + \mu_3 = 0$
- $\frac{\partial L}{\partial x_{1B}} = 0 \Rightarrow p_{1B} - \mu_1 \rho_{BC} x_{2C} + \mu_4 = 0$
- $\frac{\partial L}{\partial x_{2B}} = 0 \Rightarrow p_{2B} - \mu_2 \rho_{AB} x_{1A} + \mu_5 = 0$
- $\frac{\partial L}{\partial x_{2C}} = 0 \Rightarrow p_{2C} - \mu_1 (\rho_{AC} x_{1A} + \rho_{BC} x_{1B}) - \mu_2 \rho_{AC} x_{1A} = 0$

Συνθήκες συμπληρωματικής χαλαρότητας (Complementary slackness conditions)

- $\mu_1 (2D - \rho_{AC} x_{1A} x_{2C} - \rho_{BC} x_{1B} x_{2C}) = 0$ $(2D - \rho_{AC} x_{1A} x_{2C} - \rho_{BC} x_{1B} x_{2C}) \geq 0$, $\mu_1 \geq 0$
- $\mu_2 (2D - \rho_{AB} x_{2B} x_{1A} - \rho_{AC} x_{2C} x_{1A}) = 0$ $(2D - \rho_{AB} x_{2B} x_{1A} - \rho_{AC} x_{2C} x_{1A}) \geq 0$, $\mu_2 \geq 0$
- $\mu_3 x_{1A} = 0$ $\mu_3 \geq 0, x_{1A} \geq 0$
- $\mu_4 x_{1B} = 0$ $\mu_4 \geq 0, x_{1B} \geq 0$
- $\mu_5 x_{2B} = 0$ $\mu_5 \geq 0, x_{2B} \geq 0$
- $\mu_6 x_{2C} = 0$ $\mu_6 \geq 0, x_{2C} \geq 0$

Λαμβάνοντας υπόψη όλους τους δυνατούς συνδυασμούς και κάνοντας υποθέσεις για κάποιες από τις οποίες οδηγηθήκαμε σε άτοπο, καταλήγουμε σε δύο περιπτώσεις.

❖Πρώτη περίπτωση: $\mu_3, \mu_4, \mu_5, \mu_6 = 0, \mu_1 = \frac{p_{1B}}{\rho_{BC} x_{2C}} > 0,\ \mu_2 = \frac{p_{2B}}{\rho_{AB} x_{1A}} > 0$

Μετά από υπολογισμούς καταλήξαμε σε λύση της μορφής

$$x_{2C} = \pm \sqrt{\frac{2D}{\left(\frac{p_{2C}}{p_{1B}} \rho_{BC} - \frac{p_{2B} \rho_{AC} \rho_{BC}}{p_{1B} \quad \rho_{AB}}\right)}}$$

$$x_{1A} = \pm \sqrt{\frac{2D}{\left(\frac{p_{1A}}{p_{2B}} \rho_{AB} - \frac{p_{1B} \rho_{AC} \rho_{AB}}{p_{2B} \quad \rho_{BC}}\right)}}$$

Με την προϋπόθεση

$$\frac{p_{2C}}{p_{1B}} \rho_{BC} - \frac{p_{2B}}{p_{1B}} \frac{\rho_{AC} \rho_{BC}}{\rho_{AB}} > 0$$

$$\frac{p_{1A}}{p_{2B}} \rho_{AB} - \frac{p_{1B}}{p_{2B}} \frac{\rho_{AC} \rho_{AB}}{\rho_{BC}} > 0$$

$$x_{1B} = \left(p_{2C} - p_{2B} \frac{\rho_{AC}}{\rho_{AB}}\right) \frac{x_{2C}}{p_{1B}} - \frac{\rho_{AC}}{\rho_{BC}} x_{1A}\ ,\ x_{2B} = \left(p_{1A} - p_{1B} \frac{\rho_{AC}}{\rho_{BC}}\right) \frac{x_{1A}}{p_{2B}} - \frac{\rho_{AC}}{\rho_{AB}} x_{2C}$$

❖Δεύτερη περίπτωση: $\mu_3, \mu_6 = 0,\ \mu_4 = \mu_1 \rho_{BC} x_{2C} - p_{1B}, \mu_5 = \mu_2 \rho_{AB} x_{1A} - p_{2B},$ $\mu_1 + \mu_2 = \frac{p_{1A}}{\rho_{AC} x_{2C}}$ τα οποία θα πρέπει να είναι θετικοί αριθμοί.

Μετά από υπολογισμούς καταλήξαμε σε λύση της μορφής

$$x_{2C} = \pm \sqrt{\frac{2D p_{1A}}{\rho_{AC} p_{2C}}}\ \ ,\ x_{1A} = \pm \sqrt{\frac{2D p_{2C}}{\rho_{AC} p_{1A}}},\ x_{1B} = 0\ \ , x_{2B} = 0$$

Σημειώνουμε ότι επειδή το σύστημα λύθηκε συμβολικά και με στόχο να προκύψει μια γενική μορφή της λύσης του μη γραμμικού προβλήματος καταλήξαμε στις παραπάνω σχέσεις και τις αναγκαίες συνθήκες-προϋποθέσεις ύπαρξης λύσης. Θεωρούμε ότι για συγκεκριμένες αριθμητικές τιμές θα ικανοποιείται μια από τις δύο περιπτώσεις που περιγράψαμε και συνεπώς μπορεί να υπολογιστεί η ακριβής τιμή για καθεμία από τις μεταβλητές του προβλήματος.

ΚΕΦΑΛΑΙΟ 7

7.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Ολοκληρώνοντας λοιπόν, στα πλαίσια αυτής της εργασίας υλοποιήσαμε μια μεθοδολογία επικοινωνίας και ανταλλαγής δεδομένων μεταξύ χρηστών με συγκρουόμενα συμφέροντα εφαρμόζοντας πρακτικές της Θεωρίας Παιγνίων με στόχο την διατήρηση της ιδιωτικότητας όλων σε ένα υβριδικό σύστημα συστάσεων. Το ερώτημα που κληθήκαμε να απαντήσουμε είναι τι θα πρέπει να δηλώσει καθένας χρήστης στον recommendation server ώστε να απέχει όσο περισσότερο γίνεται από τα πραγματικά του δεδομένα αλλά ταυτόχρονα να μην επηρεαστεί η σύσταση που θα λάβει σε σχέση με αυτή που θα λάμβανε αν έλεγε την αλήθεια. Καταλήξαμε ότι αυτό θα είναι το σημείο NEP αν επιλύσουμε ένα πρόβλημα βελτιστοποίησης για κάθε χρήστη και επιβεβαιώσαμε τα παραπάνω ευρήματα σε πραγματικά σύνολα δεδομένων. Στη συνέχεια αξιολογήσαμε την ακρίβεια της μεθόδου με ικανοποιητικά αποτελέσματα. Ως επέκταση παρουσιάσαμε σενάρια συνεργασίας μεταξύ των χρηστών που δίνουν καλύτερα αποτελέσματα για την διατήρηση της ιδιωτικότητας τους.

Μια πιθανή βελτίωση αφορά τον χρόνο ολοκλήρωσης γιατί σε κάποιες περιπτώσεις εξαιτίας και της ποσότητας και του είδους των δεδομένων αλλά και των υπολογισμών που πρέπει να γίνουν σε κάθε επανάληψη η καθυστέρηση ήταν σημαντική. Μια πιθανή λύση θα ήταν οι χρήστες να λύνουν παράλληλα το πρόβλημα βελτιστοποίησης, το οποίο αποτελεί το βασικό σημείο υπολογισμών σε κάθε επανάληψη, εφόσον ο ένας δεν εξαρτάται από τον άλλον στην ίδια επανάληψη. Ένα άλλο κλασικό πρόβλημα των συστημάτων συστάσεων είναι τα αραιά δεδομένα (sparsity). Εμείς εξηγήσαμε στην ενότητα 4.1.2 έναν πολύ απλοϊκό τρόπο αντιμετώπισης που εφαρμόσαμε στο δικό μας μοντέλο. Ωστόσο μια ενδιαφέρουσα πρόταση [20] είναι να προσπαθήσουμε να αντικαταστήσουμε τις θέσεις με μηδενικά στον πίνακα με τις βαθμολογίες, οι οποίες ουσιαστικά θεωρούνται κενές, με τιμές που θα έχουν προσεγγισθεί εφαρμόζοντας Imputation – Boosted CF όπως Predictive Mean Matching ή κατηγοριοποιητές μηχανικής μάθησης. Στη συνέχεια θα εφαρμόζεται ο υπολογισμός του συντελεστή Pearson στο νέο πίνακα.

7.2 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Σήμερα οι ερευνητές επικεντρώνουν το ενδιαφέρον τους στο κατά πόσο τα συστήματα συστάσεων μπορούν να προτείνουν στους χρήστες τμήματα μιας ιστοσελίδας που δεν έχουν εξερευνήσει ακόμη. Για παράδειγμα στο Amazon.com αντί να προτείνουν σε κάποιον που αγοράζει συνήθως βιβλία ένα ακόμη βιβλίο, να του προτείνουν ρούχα ή αντικείμενα από άλλες κατηγορίες και όχι κάτι ασφαλές και αποδεδειγμένα πετυχημένο.

Για αυτό το λόγο εκτός από τα μέτρα αξιολόγησης των συστημάτων συστάσεων που αφορούν κυρίως την εκτίμηση της ακρίβειας υπάρχουν και άλλα μέτρα εξίσου σημαντικά που κερδίζουν έδαφος όπως παρουσιάστηκαν στο ACM RecSys 2012 International Workshop on

Recommendation Utility Evaluation: Beyond RMSE[12]. Παρακάτω παρατίθενται κάποια από αυτά:

- Καινοτομία (Novelty / Serendipity / Diversity): Είναι σημαντικό ένα σύστημα συστάσεων να μπορεί να προτείνει μη γνωστά, μη αναμενόμενα, διαφορετικά μεταξύ τους αλλά χρήσιμα για τον χρήστη προϊόντα δηλαδή τελικά να καταφέρει να του προτείνει σχεδόν τυχαία κάτι που ούτε ο ίδιος γνώριζε ότι του αρέσει. Ο υπολογισμός του μέτρου της καινοτομίας είναι δύσκολο να υπολογιστεί offline. Αντίθετα για την μέτρηση του diversity στην λίστα των συστάσεων προτάθηκε το εξής [14]: να εκτιμηθεί ως ο μέσος όρος ανομοιογένειας για όλα τα ζεύγη των αντικειμένων που συστήνονται σε κάποιο χρήστη. Η ανομοιογένεια είναι το συμπληρωματικό της ομοιότητας και ορίζεται ως dsim(i , j) = 1- similarity(i , j). Οπότε το diversity υπολογίζεται ως

$$D(R) = \frac{1}{N(N-1)} \sum_{i \in R} \sum_{j \in R, j \neq i} dsim(i,j) \quad \text{με N} = |R|$$

- Χρησιμότητα της σύστασης (Utility of Recommendation): Ένα καλό σύστημα συστάσεων πρέπει να αποφεύγει τις άκυρες συστάσεις και τις πολύ προφανείς. Δεν επιτρέπεται να προτείνει αντικείμενα ήδη γνωστά στο χρήστη, δεν έχει κάποια αξία για τον ίδιο. Για παράδειγμα, σε ένα σύστημα που προτείνει ταινίες μια πολύ γνωστή ταινία και αναμενόμενη να έχει επιτυχία στο κοινό (blockbuster) δεν ωφελεί να προταθεί σε όλους τους χρήστες διότι είναι πολύ πιθανό οι περισσότεροι να την έχουν δει.
- Κάλυψη (Coverage) αφορά το ποσοστό των αντικειμένων που είναι γνωστά στο σύστημα συστάσεων και για τα οποία μπορεί να εξάγει πρόβλεψη.
- Ικανοποίηση του χρήστη (User Satisfaction) μπορεί να υπολογιστεί με σιγουριά μόνο online μετά από επισκόπηση των χρηστών ή ελέγχοντας τη μνήμη τους συστήματος και κάποια στατιστικά χρήσης. Βέβαια, ένα σύστημα με υψηλή ακρίβεια, καινοτομία, ποικιλία έμμεσα μας προϊδεάζει για την ικανοποίηση των χρηστών του.
- Σταθερότητα (Stability) αντανακλά την συνέπεια του συστήματος ως προς τις συστάσεις που παράγει. Με άλλα λόγια κατά πόσο επηρεάζεται η πρόβλεψη του συστήματος για τα ίδια αντικείμενα, εφαρμόζοντας τον ίδιο αλγόριθμο αν προστεθεί νέα πληροφορία για έναν χρήστη (π.χ. νέες βαθμολογίες για αντικείμενα).

Ανοιχτό θέμα γενικότερα στα συστήματα συστάσεων είναι και το εξής: οι χρήστες να μην βαθμολογούν μεμονωμένα αντικείμενα αλλά ζεύγη διότι υπάρχουν διακυμάνσεις στην βαθμολογική κλίμακα και ο χρήστης μπορεί να επηρεάζεται από την διάθεση του, διαφορετικά μπορεί να αντιλαμβάνεται κάθε άνθρωπος την ίδια βαθμολογία, τα 3 αστέρια για κάποιον μπορεί να ισοδυναμούν με 5 αστέρια για κάποιον άλλο. Το πρόβλημα που εμφανίζεται είναι τα πολλά τέτοια ζεύγη συγκρίσεων που μπορεί να δίνουν και αντιφατικά αποτελέσματα. Τα συστήματα συστάσεων έχουν επεκταθεί και σε άλλες πλατφόρμες εφαρμογής όπως τα κινητά τηλέφωνα (Mobile Recommenders). Δεδομένου ότι τα «έξυπνα» κινητά χρησιμοποιούνται

[12] http://ir.ii.uam.es/rue2012/

ευρύτατα και διαθέτουν τεχνολογίες όπως GPS και Wi-fi άλλα και η ραγδαία ανάπτυξη των social media ανοίγουν νέους δρόμους έρευνας. Ένα τέτοιο σύστημα για παράδειγμα βρίσκει την καταλληλότερη διαδρομή στην πόλη, χρήσιμη εφαρμογή για τους οδηγούς ταξί. Άλλο παράδειγμα, εφαρμογή που προτείνει στον χρήστη εστιατόρια χωρίς όμως να έχει ζητήσει ο χρήστης να του γίνει πρόταση (Proactive Recommender System), αλλά η εφαρμογή ενεργεί αυτόβουλα και προτείνει το συγκεκριμένο εστιατόριο που ταιριάζει στα γούστα κάποιου όταν αυτός βρίσκεται στην γύρω περιοχή. Ωστόσο σε αυτό τον τομέα υπάρχουν και αρκετοί παράγοντες που πρέπει να ληφθούν υπόψη όπως η μικρή οθόνη οπότε ο χρήστης δεν μπορεί να δει μια μεγάλη λίστα από αντικείμενα άρα πρέπει οι σωστές προτάσεις να βρίσκονται στην κορυφή για να προσελκύσουν τον χρήστη. Ενδιαφέρον παρουσιάζουν προσπάθειες εξέλιξης των συστημάτων συστάσεων ώστε να αλλάξουν τον τρόπο ζωής της κοινωνίας όπως παρουσιάστηκαν στο 1ο International Workshop on Recommendation Technologies on Lifestyle Change (RecSys 2012). Απώτερος στόχος τους, να βοηθήσουν τους χρήστες στη λήψη αποφάσεων σε καθημερινά ζητήματα, καθιερώνοντας ένα συμβιβασμό ανάμεσα στην ποιότητα ζωής, στην ατομικότητα και στη διασκέδαση. Προτάσεις για πιο φιλική προς το περιβάλλον μετακίνηση (Travel Recommenders), για επιλογή πιο υγιεινών γευμάτων διατροφής (Food Recommenders) είναι μερικές μόνο αρχικές προσπάθειες.

Αυτό που είναι βέβαιο είναι ότι τα συστήματα συστάσεων με το πέρασμα του χρόνου θα βελτιώνονται, θα συλλέγουν ακόμα περισσότερες πληροφορίες για εμάς και θα εμφανίζονται σε τομείς που δεν το περιμένουμε. Γι' αυτό είναι επιτακτική η ανάγκη να αναζητήσουμε νέους και πιο αποτελεσματικούς τρόπους προστασίας των ιδιωτικών δεδομένων των χρηστών και να εξασφαλίσουμε συστήματα συστάσεων που θα παρέχουν υπηρεσίες προς όφελος των χρηστών χωρίς καταπάτηση της ιδιωτικότητας τους.

[1] Sweeney, L.: k-Anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.

[2] Polat, H., Du, W.: Privacy-preserving collaborative filtering using randomized perturbation techniques. In: Proc. of Inter. Conf. on Data Mining (ICDM), 2003.

[3] Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proc. of the 2000 ACM SIGMOD on Management of Data, 2000.

[4] Canny, J.: Collaborative filtering with privacy. In: IEEE Symposium on Security and Privacy, 2002.

[5] Canny, J.: Collaborative Filtering with Privacy via Factor Analysis. In: Proc. of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002.

[6] Miller, B., Konstan, J., Riedl, J.,: Pocketlens: Toward a personal recommender system, ACM Transactions on Information Systems 22(3), 2004.

[7] Shokri, R., Pedarsani, P., Theodorakopoulos, G., Hubaux, J.P.: Preserving privacy in collaborative filtering through distributed aggregation of offline profiles. In: Proc of ACM RecSys, 2009.

[8] Berkovsky, S., Eytani, Y., Kuflik, T., Ricci, F.: Enhancing privacy and preserving accuracy of distributed collaborative filtering. In: Proc. of ACM RecSys, 2007.

[9] Lam, S., Frankowski, D., Riedl, J.: Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In: Proc of Emerging Trends in Information and Communication Security, International Conference, ETRICS 2006.

[10] Chow, R., Pathak, M., Wang,C.: A practical system for privacy-preserving collaborative filtering. In: Proc. of the 12th IEEE International Conference on Data Mining Workshops (ICDMW), 2012.

[11] Halkidi M., Koutsopoulos I.: A game theoretic framework for data privacy preservation in Recommender systems. In: Proc of European Conference on Machine Learning and Principles and Practice of knowledge discovery in databases (ECML/PKDD), 2011

[12] Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: Proc of the Inter. Conf. of WWW, 2001

[13] Candillier, L., Meyer, F., Fessant, F.: Designing specific weighted similarity measures to improve collaborative filtering systems.

[14] Baskaya, O., Aytekin, T.: How similar is rating similarity to content similarity. Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2012), ACM RecSys 2012.

[15] Seminario, C., Wilson, D.: Case study evaluation of Mahout as a recommender platform. Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2012), ACM RecSys 2012.

[16] Meyer, F., Fessant, F., Clerot, F., Gaussier, E.: Toward a new protocol to evaluate recommender systems. Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2012), ACM RecSys 2012.

[17] Basu, C., Hirsh, H., Cohen, W.: Recommendation as classification: using social and content-based information in recommendation. American Association for Artificial Intelligence, 1998.

[18] Hahsler M.: Recommenderlab: A framework for developing and testing recommendation algorithms.
[19] McNee, S., Riedl, J., Konstan, J.: Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems. In: Proc. of CHI Extended Abstracts on Human Factors in Computing Systems, 2006
[20] Su, X., Khoshgoftaar, T., Greiner, R.: Imputation-Boosted Collaborative Filtering Using Machine Learning Classifiers. In: Proc. Of 23rd Annual ACM Symposium on Applied Computing, 2008
[21] http://spectrum.ieee.org/computing/software/deconstructing-recommender-systems
[22] http://urbanmining.wordpress.com/2012/10/03/open-problems-in-recommender-systems