Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης & Υπολογιστων
Εργαστηριο Ευφυων Συστηματων

Συνεργατικά Συστήματα Συστάσεων με χρήση έμμεσης και άμεσης κοινωνικής πληροφορίας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΓΕΩΡΓΙΟΥ Χ. ΑΛΕΞΑΝΔΡΙΔΗ

Αθήνα, Δεκέμβριος 2015

Εθνικο Μετσοβιο Πολυτεχνειο
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ
ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ

Συνεργατικά Συστήματα Συστάσεων με χρήση έμμεσης και άμεσης κοινωνικής πληροφορίας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΓΕΩΡΓΙΟΥ Χ. ΑΛΕΞΑΝΔΡΙΔΗ

Συμβουλευτική Επιτροπή: Ανδρέας-Γεώργιος Σταφυλοπάτης
Στέφανος Κόλλιας
Παναγιώτης Τσανάκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 17η Δεκεμβρίου 2015

.....	.....	.....
Α.-Γ. Σταφυλοπάτης	Στ. Κόλλιας	Π. Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
.....	.....	.....
Γ. Μέντζας	Γ. Στάμου	Κ. Κοντογιάννης
Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.
	.....	
	Μ. Βίρβου	
	Καθηγήτρια Παν. Πειραιώς	

Αθήνα, Δεκέμβριος 2015.

.....

ΓΕΩΡΓΙΟΣ Χ. ΑΛΕΞΑΝΔΡΙΔΗΣ
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Αλγορίθμων

ΠΡΟΛΟΓΟΣ

Αντικείμενο της παρούσας διατριβής είναι η αξιοποίηση της κοινωνικής πληροφορίας στα συνεργατικά συστήματα συστάσεων είτε αυτή παρέχεται άμεσα είτε συνάγεται έμμεσα. Η υπολογιστική νοημοσύνη και ο ρόλος που αυτή μπορεί να διαδραματίσει στην ανάλυση των προτιμήσεων των χρηστών και στην εύρεση συσχετίσεων μεταξύ τους ήταν ένα θέμα έρευνας, που κατά τη γνώμη μου παρουσίαζε σημαντικό ενδιαφέρον και η προσωπική ενασχόλησή μου με αυτό ξεκίνησε το 2009 στο Εργαστήριο Ευφυών Συστημάτων. Η συνεργασία όλων των εμπλεκόμενων στο εργαστήριο, ο πλούτος των γνώσεων που απέκτησα, αλλά και η όλη ερευνητική διαδικασία που οδήγησε σ΄ αυτή τη διατριβή ήταν για μένα μια πολύ σημαντική εμπειρία που συνέβαλε στην περαιτέρω διεύρυνση της σκέψης μου.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, όχι μόνο για την αμέριστη συμπαράστασή του σ΄ όλα τα επίπεδα αλλά και για την αγαστή συνεργασία μας. Αισθάνομαι υποχρεωμένος να τονίσω ότι οι γνώσεις του, η εμπειρία του, ο τρόπος προσέγγισης των προβλημάτων που παρουσιάστηκαν στη διάρκεια της ερευνητικής μου πορείας, η εμπιστοσύνη που μου έδειξε και η βοήθεια που μου προσέφερε, συνέβαλαν αποφασιστικά στην εκπόνηση αυτής της διατριβής. Επίσης οφείλω να ευχαριστήσω ιδιαίτερα τους καθηγητές ΕΜΠ κ. Στέφανο Κόλλια και κ. Παναγιώτη Τσανάκα, μέλη της συμβουλευτικής επιτροπής, για το ενδιαφέρον και τις χρήσιμες παρατηρήσεις τους. Θα ήθελα ακόμη να ευχαριστήσω τους κ.κ. Μέντζα Γρηγόριο, Καθηγητή Ε.Μ.Π., Στάμου Γεώργιο, Επίκουρο Καθηγητή Ε.Μ.Π, Κοντογιάννη Κωνσταντίνο, Αναπληρωτή Καθηγητή Ε.Μ.Π, και την κα. Βίρβου Μαρία, Καθηγήτρια Πανεπιστημίου Πειραιώς, για την τιμή που μου έκαναν να είναι μέλη της επιτροπής αξιολόγησης της διατριβής.

Αισθάνομαι επιπλέον την ανάγκη να ευχαριστήσω τον μεταδιδακτορικό ερευνητή του Εργαστηρίου Ευφυών Συστημάτων κ. Γιώργο Σιόλα, για την καθημερινή ενασχόλησή του με την ερευνητική μου εργασία και την ουσιαστική συμβολή του στην ολοκλήρωση αυτής της διατριβής. Ευχαριστίες οφείλω και σ΄ όλα τα μέλη του Εργαστηρίου Ευφυών Συστημάτων,

παλιά και νέα, και ιδιαίτερα στους Γεράσιμο Σπανάκη, Ηλιάννα Κόλλια και Αριστείδη Λαναρίδη, διδάκτορες ΕΜΠ καθώς και στους Στρατογιάννη Γεώργιο, Ελένη Βάθη και Χριστίνα Χριστάκου, υποψήφιους διδάκτορες ΕΜΠ για τη συνεργασία τους και την προθυμία τους να με βοηθήσουν στη διάρκεια εκπόνησης της διατριβής αυτής.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, τον αδελφό μου, τους φίλους μου και τα αγαπημένα μου πρόσωπα για την αμέριστη συμπαράστασή τους.

Γεώργιος Αλεξανδρίδης
Αθήνα, Δεκέμβριος 2015

ΠΕΡΙΛΗΨΗ

Η ψηφιακή επανάσταση των προηγούμενων δεκαετιών είχε ως άμεσο αποτέλεσμα την αλματώδη αύξηση της πληροφορίας που διακινείται ηλεκτρονικά, οδηγώντας σε μια κατάσταση που είναι γνωστή ως πληροφοριακή υπερφόρτωση. Συνεπώς, υπάρχει άμεση ανάγκη για την κατασκευή συστημάτων τα οποία θα μπορούν να αναζητούν, να ταξινομούν και να κατηγοριοποιούν την διαθέσιμη πληροφορία. Τα Συστήματα Συστάσεων έχουν προταθεί ως μια λύση που μπορεί να αντιμετωπίσει σε ένα βαθμό το προαναφερθέν πρόβλημα. Η συνεισφορά της διατριβής έγκειται στη μελέτη, στην ανάπτυξη και στην πρακτική εφαρμογή αλγορίθμων για την πιο ευρύτατα διαδεδομένη κατηγορία συστημάτων συστάσεων, τα συνεργατικά συστήματα συστάσεων, με στόχο την βελτίωση της απόδοσής τους.

Τα συστήματα συστάσεων είναι επί της ουσίας εργαλεία λογισμικού, τα οποία χρησιμοποιούνται για την παραγωγή εξατομικευμένων προτάσεων για κάθε χρήστη χωριστά, συνήθως υπό τη μορφή ταξινομημένων λιστών. Οι προτάσεις στην πλειοψηφία των περιπτώσεων είναι αντικείμενα οποιουδήποτε τύπου, όπως βιβλία, ταινίες, μουσικά κομμάτια και άρθρα ειδήσεων, τα οποία ενδεχομένως να φανούν χρήσιμα στους χρήστες. Η παραγωγή των προτάσεων στα συνεργατικά συστήματα συστάσεων βασίζεται στις προτιμήσεις του εκάστοτε χρήστη καθώς και σε άλλες παραμέτρους. Οι δε προτιμήσεις συνάγονται με δύο τρόπους: είτε απευθείας από τους ίδιους (κατά βάση εντός καθορισμένης αξιολογικής/βαθμολογικής κλίμακας) είτε έμμεσα από το ίδιο το σύστημα.

Στο πλαίσιο της διατριβής αναπτύσσονται αλγόριθμοι παραγωγής συνεργατικών συστάσεων, οι οποίοι αξιοποιούν τις ρητά εκπεφρασμένες προτιμήσεις των χρηστών. Η έννοια της συνεργατικότητας έγκειται στο γεγονός πως τα νέα αντικείμενα που τελικά προτείνονται σε ένα χρήστη βασίζονται στις αξιολογήσεις που αυτά έχουν λάβει από άλλους «παρόμοιους» χρήστες. Δηλαδή, τα συνεργατικά συστήματα συστάσεων μετατρέπουν τον κάθε χρήστη σε «μέσο πρόβλεψης» των προτιμήσεων των άλλων, καθορίζοντας κατ΄ αυτόν τον τρόπο μια άτυπη μορφή κοινωνικότητας μεταξύ τους. Για την αναζήτηση και την καλύτερη αξιοποίηση αυτού

του είδους των έμμεσων σχέσεων μεταξύ των χρηστών, προτείνεται ένας πρωτότυπος αλγόριθμος παραγωγής συστάσεων, δομικό στοιχείο του οποίου αποτελεί ένα πολυεπίπεδο νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης. Οι επιλογές που γίνονται όσον αφορά τη συνάρτηση μεταφοράς των νευρώνων του δικτύου, της δυναμικής προσαρμογής της αρχιτεκτονικής του όπως και του ενισχυτικού αλγορίθμου εκπαίδευσής του, επιτρέπουν την σε βάθος ανάλυση των σχέσεων μεταξύ των χρηστών του που οδηγούν στην παραγωγή καλύτερων συστάσεων.

Εκτός της έμμεσης συναγωγής ομοιοτήτων και διαφορών μεταξύ των χρηστών, τα συνεργατικά συστήματα συστάσεων μπορούν να επεκταθούν, συμπεριλαμβάνοντας στους υπολογισμούς τους και τις άμεσες σχέσεις που ενδεχομένως να έχουν οι χρήστες μεταξύ τους, στο πλαίσιο ενός κοινωνικού δικτύου. Παραδείγματα τέτοιων σχέσεων αποτελούν οι δεσμοί φιλίας, εμπιστοσύνης καθώς και οι δεσμοί «ακολούθων». Από τους πιο διαδεδομένους τρόπους προσπέλασης της άμεσης κοινωνικής πληροφορίας στα συνεργατικά συστήματα συστάσεων, είναι η πραγματοποίηση τυχαίων περιπάτων επάνω στα κοινωνικά δίκτυα, με στόχο την ανεύρεση χρηστών με επιρροή. Στο πλαίσιο αυτό, προτείνεται μια νέα μεθοδολογία πραγματοποίησης τυχαίων περιπάτων στο μεικτό δίκτυο που σχηματίζεται από τον συνδυασμό του κοινωνικού δικτύου και του δικτύου «ομοιότητας» (ενός δικτύου σχέσεων μεταξύ των χρηστών που προκύπτει από τις αξιολογήσεις που έχουν ήδη κάνει). Η διαφορά με άλλες αντίστοιχες μεθόδους βρίσκεται στο γεγονός ότι το επόμενο βήμα του περιπάτου δεν επιλέγεται ομοιόμορφα τυχαία, αλλά αντίθετα εισάγεται μια μεροληψία προς εκείνους τους χρήστες που είναι περισσότερο όμοιοι.

Πέρα από τον χώρο των χρηστών, οι τυχαίοι περίπατοι μπορούν επίσης να πραγματοποιηθούν και στον χώρο των αντικειμένων. Για το λόγο αυτό στην παρούσα διατριβή προτείνεται η πραγματοποίηση τυχαίων περιπάτων επάνω σε ένα πρωτότυπο δίκτυο, το δίκτυο κατανάλωσης αντικειμένων. Το συγκεκριμένο δίκτυο συνδέει μεταξύ τους αντικείμενα-κόμβους με ακμές των οποίων τα βάρη αποτυπώνουν το πλήθος των κοινών τους προσπελάσεων. Είναι προσωπικό, υπό την έννοια ότι κατασκευάζεται χωριστά για κάθε χρήστη, στη βάση των αντικειμένων που έχει προσπελάσει τόσο ο ίδιος όσο και άλλοι χρήστες που εντάσσονται είτε στο άμεσο κοινωνικό του δίκτυο, είτε παρουσιάζουν κάποια ομοιότητα με αυτόν ή, τέλος, ανήκουν και στις δύο κατηγορίες.

Τέλος, η άμεση κοινωνική πληροφορία που εμπεριέχεται στα συνεργατικά συστήματα συστάσεων αξιοποιείται και με έναν ακόμα τρόπο. Αποτελεί το πεδίο εφαρμογής μιας νέας

μεθόδου μπεϋζιανής μη-αρνητικής παραγοντοποίησης πινάκων, με στόχο την κατηγοριοποίηση των χρηστών σε επικαλυπτόμενες κοινότητες. Η πρωτοτυπία της προσέγγισης βρίσκεται στην επιλογή της συνάρτησης της εκ των προτέρων πιθανότητας. Ενώ στη βιβλιογραφία συνηθίζεται να χρησιμοποιείται η θεωρία των συζυγών εκ των προτέρων κατανομών, στη συγκεκριμένη περίπτωση εφαρμόζονται τα μοντέλα εκθετικών τυχαίων γράφων. Ο λόγος που επιλέχθηκε αυτή η κατεύθυνση ήταν για να μετριαστεί η τοπικότητα της παραγοντοποίησης μέσω της εισαγωγής στη διαδικασία μακροσκοπικών πληροφοριών για τα υπό εξέταση δίκτυα. Επίσης, στο πλαίσιο της ίδιας διαδικασίας, αναπτύχθηκε και μια πρωτότυπη προσεγγιστική μέθοδος για τον υπολογισμό των υπερπαραμέτρων του μοντέλου των εκθετικών τυχαίων γράφων, η οποία βασίστηκε στη θεωρία του μέσου πεδίου.

Συμπερασματικά, η ουσιαστική συμβολή της διατριβής συνοψίζεται στην χρήση ευφυών τεχνικών στα συνεργατικά συστήματα συστάσεων, για την επεξεργασία και ανάλυση της άμεσης και έμμεσης κοινωνικής πληροφορίας που ενυπάρχει σε αυτά. Κάθε τεχνική που αναπτύχθηκε, αξιολογήθηκε πειραματικά σε δημόσια διαθέσιμες συλλογές δεδομένων, οι οποίες χρησιμοποιούνται ευρέως από την επιστημονική κοινότητα, ενώ έγιναν και συγκρίσεις με άλλες αντίστοιχες τεχνικές στο κάθε πεδίο έρευνας.

ABSTRACT

The digital revolution of the past decades has had as a direct consequence the rapid increase of the information distributed in electronic form, leading to a situation known as the information overload. Therefore, there exists an immediate need for the development of systems that can search, classify and categorize the available information. Recommender Systems have been proposed as a solution that may deal with the aforementioned problem to a certain extent. The contribution of this thesis lies in the study, development and practical application of algorithms for the most popular category of recommender systems, that of collaborative filtering, aiming at improving their performance.

Recommender systems are essentially software tools used for the production of personalized recommendations for each user individually, usually in the form of ordered lists. Recommendations are in most cases items of any kind, like books, movies, music tracks and news articles, that may be useful to the users. The creation of recommendations in collaborative filtering systems is based upon the preferences of each user along with other parameters. Preferences are deduced in two ways; either directly from users (though the ratings they provide in a predefined scale) or indirectly, by the system itself.

In the context of this thesis, recommendation algorithms that utilize the explicit user preferences are developed. The notion of collaboration lies in the fact that the new items recommended to a user are based on the ratings provided to them by other "similar" users. That is, collaborative filtering turns each user to the other user's predictor, thereby defining an informal social bond between them. For the exploration and the better exploitation of this kind of indirect relationships between the users, a novel recommendation algorithm is being proposed, whose structural element is a feed-forward multilayered perceptron. The choices made regarding the transfer function of the neurons of the network, the dynamic adaptation of its architecture as well as the boosted training algorithm, allow for the deep analysis of the relationships between the users that lead to the production of better

recommendations.

Apart form the indirect deduction of similarities and dissimilarities between the users, collaborative filtering can also be extended through the inclusion in the computations of the direct relationships their users may have, in the context of a social network. Examples of such relationships are the bonds of friendship, trust as well as the "follower" ties. Among the most widespread ways of accessing the direct social information in recommender systems is through the performance of random walks, aiming at locating influential users. In this context, a novel methodology for performing random walks in the joint network constructed from the combination of the social network and the "similarity" network (a network of ties between users, based on the ratings they have provided) is proposed. The difference to other approaches lies in the fact that the next step of the walk is not chosen uniformly at random, but a bias towards the most similar users is introduced instead.

In addition to the user space, random walks may also occur in the item space. For this reason , the performance of random walks in a novel network, the item consumption network, is also proposed in this thesis. This specific network connects item-nodes with vertices whose weights depict the number of their common access patterns. This network is also personal, in the sense that it is constructed for every user, based on the items that he or she have accessed, as well as other users that are members of either his/her direct social network, or are similar to him/her or belong to both categories.

Finally, the direct social relationships are exploited in yet another way in collaborative filtering systems. They constitute the basis of a new bayesian method of non-negative matrix factorization, aiming at categorizing the users in overlapping communities. The novelty of the procedure lies in the choice of the a priori distribution. While in literature a probability distribution based on the conjugate prior theory is selected, in this specific case, the exponential random graph models are used. The reason for choosing this course of action is to balance the locality of the factorization through the introduction of global network properties in the process. Additionally, in the context of the same procedure, a novel approximation method for the hyperparameters of the exponential random graph models has been developed, based on the mean field theory.

In conclusion, the actual contribution of this thesis is summarized in the use of intelligent techniques in collaborative filtering recommender systems, for the processing and

analysis of the direct and indirect social relationships the aforementioned systems contain. Every proposed technique has been experimentally tested in publicly available datasets that are widely used by the scientific community, while comparisons with other relevant techniques in every research field have also been performed.

Κατάλογος Συντμήσεων

ANN	:	Artificial Neural Networks (Τεχνητά Νευρωνικά Δίκτυα)
CF	:	Collaborative Filtering (Συνεργατική Διήθηση)
CNNA	:	Constructive Neural Network Algorithm (Κατασκευαστικός Αλγόριθμος Νευρωνικών Δικτύων)
ERGM	:	Exponential Random Graph Model (Μοντέλο Εκθετικού Τυχαίου Γράφου)
FOAF	:	Friend-of-a-Friend Network (Δίκτυο Φίλων Φίλου)
ICN	:	Item Consumption Network (Δίκτυο Κατανάλωσης Αντικειμένων)
MAE	:	Mean Absolute Error (Μέσο Απόλυτο Σφάλμα)
MAUE	:	Mean Absolute User Error (Μέσο Απόλυτο Σφάλμα ανά Χρήστη)
MLP	:	Multi-Layer Perceptrons (Πολυεπίπεδα Perceptron)
NCA	:	New Constructive Algorithm (Νέος Κατασκευαστικός Αλγόριθμος)
NMF	:	Non-Negative Matrix Factorization (Μη-αρνητική Παραγοντοποίηση Πίνακα)
nDCG	:	Normalized Discounted Cumulative Gain (Κανονικοποιημένη Μειούμενη Αθροιστική Απολαβή)
NMAE	:	Normalized Mean Absolute Error (Κανονικοποιημένο Μέσο Απόλυτο Σφάλμα)
NN	:	Nearest Neighbor (Μέθοδος Πλησιέστερων Γειτόνων)
PN	:	Personal Network (Προσωπικό Δίκτυο)
RS	:	Recommender Systems (Συστήματα Συστάσεων)
RMSE	:	Root Mean Square Error (Μέσο Τετραγωνικό Σφάλμα)
SNA	:	Social Network Analysis (Ανάλυση Κοινωνικών Δικτύων)
SVD	:	Singular Value Decomposition (Ανάλυση Ιδιαζουσών Τιμών)
SRS	:	Social Recommender Systems (Κοινωνικά Συστήματα Συστάσεων)
SVM	:	Support Vector Machine (Μηχανές Διανυσμάτων Υποστήριξης)
WOT	:	Web of Trust (Ιστός Εμπιστοσύνης)

Κεφάλαιο 1

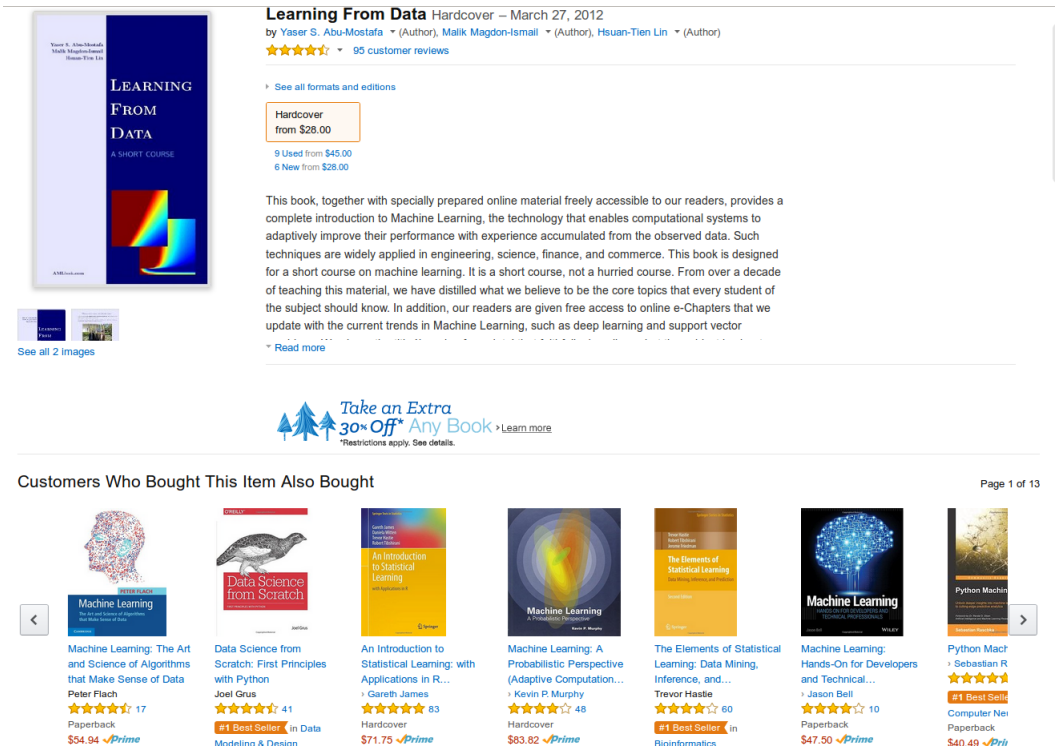
Εισαγωγή

Η ψηφιακή επανάσταση των προηγούμενων δεκαετιών είχε ως άμεσο αποτέλεσμα την αλματώδη αύξηση της πληροφορίας που διακινείται ηλεκτρονικά, οδηγώντας σε μια κατάσταση που είναι γνωστή ως *πληροφοριακή υπερφόρτωση* (information overload). Είναι πλέον σύνηθες το γεγονός οι χρήστες να έχουν να επεξεργαστούν πολλές επιλογές όταν ψάχνουν για κάτι στο Διαδίκτυο ή όταν πρόκειται να πάρουν μια απόφαση (λ.χ για την αγορά ενός προϊόντος ή μιας υπηρεσίας). Συνεπώς, υπάρχει άμεση ανάγκη για την κατασκευή συστημάτων τα οποία θα μπορούν να αναζητούν, να ταξινομούν και να κατηγοριοποιούν την διαθέσιμη πληροφορία.

Τα *Συστήματα Συστάσεων* (Recommender Systems - RS) έχουν προταθεί ως μια λύση που μπορεί να αντιμετωπίσει σε ένα βαθμό το προαναφερθέν πρόβλημα. Πρόκειται για εργαλεία λογισμικού τα οποία χρησιμοποιούνται για να προτείνουν αντικείμενα που μπορούν να φανούν χρήσιμα στους χρήστες τους, φιλτράροντας τη διαθέσιμη πληροφορία και παρουσιάζοντας τις πιο *σχετικές* και *ενδιαφέρουσες* όψεις της. Αυτό επιτυγχάνεται διαμέσου της πρότασης νέων *αντικειμένων* στους χρήστες, λαμβάνοντας υπόψη το γούστο τους. Τα υπό πρόταση αντικείμενα μπορεί να είναι οποιουδήποτε τύπου· ενδεικτικά αναφέρονται τα *βιβλία*, οι *ταινίες*, τα *μουσικά κομμάτια* και τα *άρθρα ειδήσεων*. Οι προτάσεις γίνονται με στόχο να βοηθήσουν τους χρήστες στο να λάβουν τις κατάλληλες αποφάσεις και για τον λόγο αυτό έχουν ήδη εξελιχθεί σε ένα από τα πιο δημοφιλή και ισχυρά εργαλεία που χρησιμοποιούνται σε μια πλειάδα διαδικτυακών υπηρεσιών, όπως λ.χ. στο ηλεκτρονικό εμπόριο. Για παράδειγμα, στο Σχήμα 1.1 φαίνονται οι προσωποποιημένες προτάσεις για την αγορά βιβλίων που κάνει το ηλεκτρονικό κατάστημα της Amazon σε έναν επισκέπτη του και οι οποίες βασίζονται, όπως χαρακτηριστικά αναφέρεται, σε αντίστοιχα βιβλία που έχουν αγοράσει άλλοι καταναλωτές μαζί με το υπό εξέταση βιβλίο.

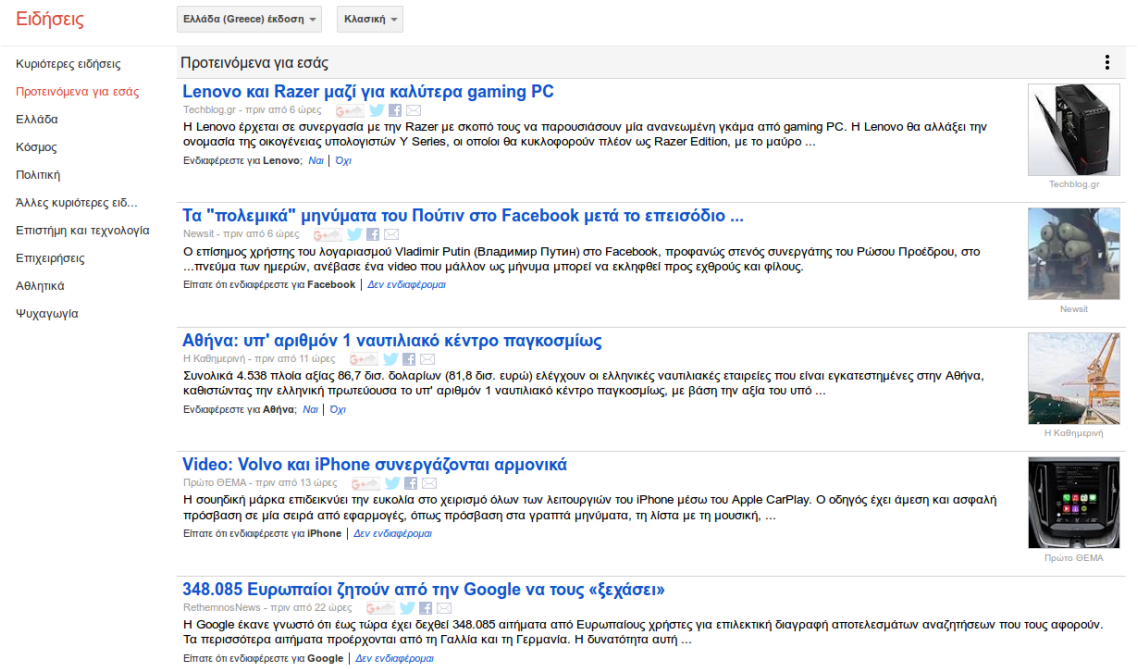
Το κύριο χαρακτηριστικό των συστημάτων συστάσεων είναι ότι παρέχουν εξατομικευμένες προτάσεις σε κάθε χρήστη χωριστά, συνήθως υπό τη μορφή ταξινομημένων λιστών (Σχήμα 1.1). Για το σκοπό αυτό, προσπαθούν να ανακαλύψουν ποια θα ήταν τα πιο χρήσιμα αντικείμενα για τον κάθε δυνητικό καταναλωτή, λαμβάνοντας υπόψη τις προτιμήσεις του καθώς και άλλες παραμέτρους. Οι προτιμήσεις των χρηστών συνάγονται με δύο τρόπους: είτε απευθείας από τους ίδιους, είτε έμμεσα από το σύστημα. Στην πρώτη περίπτωση, οι ίδιοι οι χρήστες δηλώνουν ρητά την προτίμησή τους, κατά βάση εντός καθορισμένης αξιολογικής/βαθμολογικής κλίμακας. Κλασσικό τέτοιο παράδειγμα αποτελεί το σύστημα πέντε αστέρων αξιολόγησης των ταινιών, όπως φαίνεται στο Σχήμα 1.2 για την περίπτωση της διαδικτυακής έκδοσης του περιοδικού «Αθηνόραμα».

Στη δεύτερη περίπτωση, το ίδιο το σύστημα αποθηκεύει στοιχεία της αλληλεπίδρασης των χρηστών με αυτό (σε αρχεία καταγραφής, cookies, κ.λ.π.) και κατόπιν τα επεξεργάζεται, προσπαθώντας να αντλήσει πληροφορίες για τις προτιμήσεις τους. Ένα γνωστό τέτοιο παράδειγμα αποτελεί η ειδησεογραφική πύλη της μηχανής αναζήτησης Google, το Google News (Σχήμα 1.3), το οποίο κάνει στον επισκέπτη του εξατομικευμένες προτάσεις για την ανάγνωση ειδήσεων, βασιζόμενο σε παλιότερες αναζητήσεις του συγκεκριμένου ατόμου αλλά και στα άρθρα



Σχήμα 1.1: Εξατομικευμένες προτάσεις για την αγορά βιβλίων από το ηλεκτρονικό κατάστημα της Amazon

Σχήμα 1.2: Απόσπασμα αξιολόγησης ταινίας από κριτικούς και αναγνώστες της διαδικτυακής έκδοσης του περιοδικού «Αθηνόραμα»



Σχήμα 1.3: Εξατομικευμένες ειδήσεις από το Google News, που προκύπτουν από τους όρους αναζήτησης που έχει χρησιμοποιήσει και τις ειδησεογραφικές σελίδες που έχει επισκεφτεί ο χρήστης στο παρελθόν

ειδήσεων στα οποίο αυτός μεταβαίνει (κάνοντας κλικ), ερμηνεύοντας την μετάβαση αυτή ως ενδιαφέρον, από την πλευρά του χρήστη, για το εκάστοτε άρθρο.

Η παραγωγή των συστάσεων μπορεί να γίνει με πολλούς τρόπους, οι κυριότεροι εκ των οποίων αναφέρονται με περισσότερες λεπτομέρειες στην Ενότητα 2.2. Ωστόσο, τα πιο διαδεδομένα RS είναι τα αποκαλούμενα *συνεργατικά συστήματα συστάσεων*, τα οποία αποτελούν και το αντικείμενο μελέτης της παρούσας διατριβής. Ο προσδιορισμός «συνεργατικά» υποδηλώνει ότι τα αντικείμενα προτείνονται σε ένα δεδομένο χρήστη στη βάση των αξιολογήσεων που αυτά έχουν λάβει από τους άλλους χρήστες. Ή, ακόμα πιο απλά, τα συνεργατικά συστήματα συστάσεων μετατρέπουν τον κάθε χρήστη σε «μέσο πρόβλεψης» των προτιμήσεων των άλλων.

Η προαναφερόμενη αρχή λειτουργίας εδράζεται στην παρατήρηση πως αρκετά συχνά οι άνθρωποι συμβουλεύονται άλλους ανθρώπους όταν πρόκειται να λάβουν αποφάσεις. Για παράδειγμα, βασίζονται στις προτάσεις των φίλων τους όταν θέλουν να διαβάσουν ένα καινούργιο βιβλίο ή εκτιμούν την άποψη ενός κριτικού σε μια εφημερίδα πριν αποφασίσουν να δουν μια θεατρική παράσταση. Αντίστοιχα, οι εργοδότες λαμβάνουν υπόψη τους τις συστατικές επιστολές που προσκομίζει κάποιος εργαζόμενος πριν τον προσλάβουν. Τα πρώτα συστήματα συστάσεων, τα οποία εμφανίστηκαν την δεκαετία του 1990, είχαν ως αφετηρία τους αυτές τις απλές παρατηρήσεις. Προσπαθούσαν να εντάξουν τους χρήστες και τα αντικείμενα σε όμοιες ομάδες με βάση το γούστο τους και τις ιδιότητες τους και έκαναν την παραδοχή πως αν κάποιος χρήστης είχε εκδηλώσει προτίμηση για τα αντικείμενα μιας ομάδας, τότε θα μπορούσαν κάλλιστα να του προταθούν και άλλα αντικείμενα που εντάσσονταν στην ίδια ομάδα.

1.1 Προβλήματα και Προκλήσεις

Ο επιτυχημένος σχεδιασμός και η υλοποίηση ενός συστήματος συστάσεων είναι μια διαδικασία που εξαρτάται από πολλούς παράγοντες. Καταρχήν, πρέπει να ερευνηθούν μια σειρά από ζητήματα γενικού σκοπού, όπως για παράδειγμα ερωτήματα σχετικά με το στόχο του συστήματος, το κοινό στο οποίο απευθύνεται, τον τύπο, το είδος και το εύρος των υπό πρότα-

ση αντικειμένων. Κατόπιν, πρέπει να διευθετηθούν πιο ειδικά θέματα, όπως η επιλογή των αλγορίθμων παραγωγής συστάσεων, ο καθορισμός της διεπαφής με τον χρήστη, τα προγραμματιστικά εργαλεία κ.τ.λ. Παρότι υπάρχουν κάποιες γενικές κατευθύνσεις, είναι φανερό ότι δεν μπορεί να ακολουθηθεί κάποια ενιαία πορεία μιας και οι απαντήσεις που κάθε φορά δίνονται στα προαναφερόμενα ερωτήματα εξαρτώνται από το συγκεκριμένο πεδίο εφαρμογής του υπό κατασκευή συστήματος συστάσεων.

Ωστόσο, υπάρχουν κάποια ανοιχτά ζητήματα τα οποία εμφανίζονται στη λειτουργία των περισσότερων συστημάτων συστάσεων, ανεξάρτητα από τον επιμέρους τομέα που το κάθε ένα ειδικεύεται. Αυτές οι προκλήσεις αποτελούν το αντικείμενο της παρούσας ενότητας, οπού και παρουσιάζονται οι κυριότερες εξ αυτών· η *αραιότητα των αξιολογήσεων*, το *πρόβλημα της ψυχρής εκκίνησης* και η *εμπιστοσύνη στις παραγόμενες συστάσεις*.

1.1.1 Αραιότητα των Αξιολογήσεων

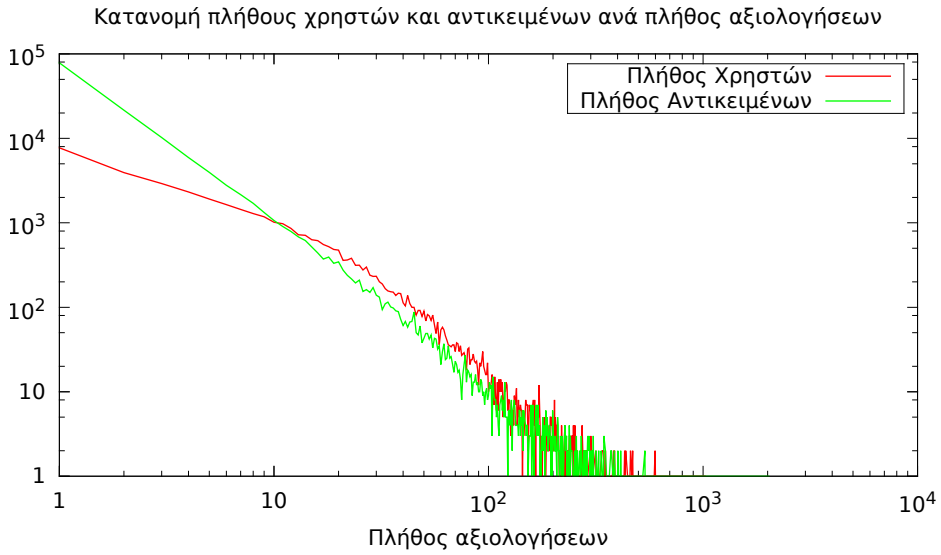
Η *αραιότητα των αξιολογήσεων* (sparsity of the ratings) αποτελεί την κυριότερη πρόκληση στην λειτουργία των συστημάτων συστάσεων και αναφέρεται στο γεγονός πως οι χρήστες των RS τείνουν να αξιολογούν ένα μικρό μόνο υποσύνολο των διαθέσιμων αντικειμένων. Αυτό έχει ως άμεση συνέπεια η πυκνότητα των αξιολογήσεων να παραμένει σε ένα πολύ μεγάλο ποσοστό χαμηλή· στην πλειοψηφία των περιπτώσεων λιγότερο από το 1% των πιθανών ζευγών χρηστών-αντικειμένων υπάρχουν καταχωρημένα στο σύστημα. Συνεπώς, οι αλγόριθμοι των συστημάτων συστάσεων πρέπει να προσαρμόσουν την λειτουργία τους κατά τέτοιον τρόπο ώστε να μπορούν να παράγουν ικανοποιητικές συστάσεις βασιζόμενοι σε πολύ λίγα δεδομένα.

Μια άμεση συνέπεια της αραιότητας των αξιολογήσεων είναι και η μικρή τους *αλληλοκάλυψη* (overlap). Όπως προαναφέρθηκε, τα συνεργατικά συστήματα συστάσεων βασίζουν την λειτουργία τους στην αναζήτηση ομοιοτήτων (και γενικότερα προτύπων) μεταξύ των χρηστών και των αντικειμένων. Καθώς όμως οι αξιολογήσεις σπανίζουν, η πιθανότητα οι χρήστες να συμπίπτουν στα αντικείμενα που έχουν αξιολογήσει είναι πολύ μικρή (ή αντίστοιχα για τις αξιολογήσεις που έχουν λάβει τα αντικείμενα από τους χρήστες). Πράγματι, τις περισσότερες φορές είτε συμπίπτουν σε ελάχιστο αριθμό αντικειμένων (χρηστών) είτε δεν συμπίπτουν καθόλου [Massa and Bhattacharjee, 2004]. Είναι εμφανές ότι ο υπολογισμός της ομοιότητας μεταξύ χρηστών (αντικειμένων) που εμφανίζουν πολύ μικρό ποσοστό αλληλοκάλυψης μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα για το κατά πόσο αυτοί (αυτά) μοιάζουν ή διαφέρουν μεταξύ τους.

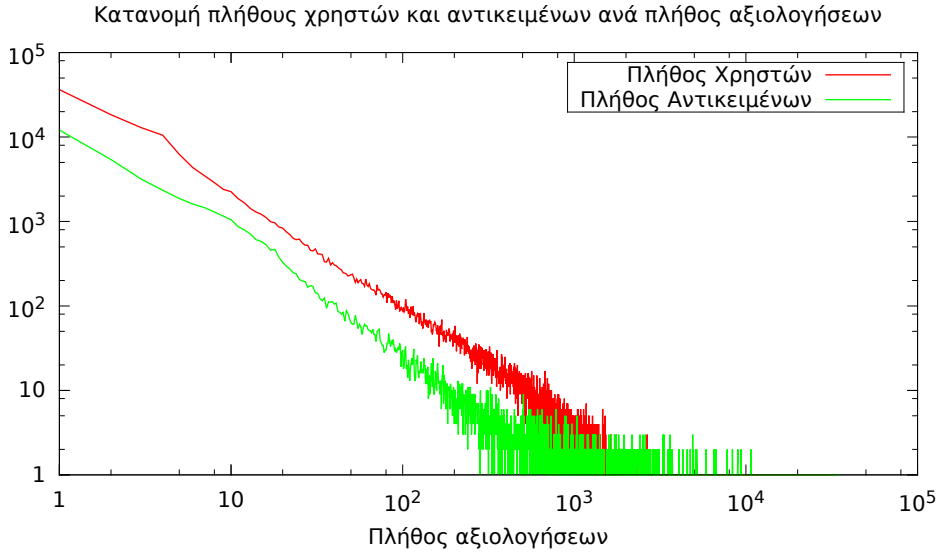
Ένα ακόμα ζήτημα το οποίο δυσχεραίνει την λειτουργία των συστημάτων συστάσεων και το οποίο σχετίζεται με την αραιότητα των αξιολογήσεων είναι η ανισότητα στην κατανομή τους. Στις περισσότερες περιπτώσεις, μια μικρή κατηγορία δημοφιλών αντικειμένων λαμβάνει την πλειοψηφία των αξιολογήσεων, ενώ όλα τα υπόλοιπα αντικείμενα έχουν ελάχιστες ως καθόλου. Μια αντίστοιχη παρατήρηση μπορεί να γίνει και για τον βαθμό συμμετοχής των χρηστών. Συνήθως υπάρχει μια πολύ μικρή κοινότητα «αφοσιωμένων» μελών, τα οποία έχουν πολύ ενεργή αλληλεπίδραση με το σύστημα, αξιολογώντας μεγάλο πλήθος αντικειμένων, την ίδια στιγμή που η πλειοψηφία των χρηστών χρησιμοποιεί το σύστημα περιστασιακά, κάνοντας ελάχιστες ως καθόλου αξιολογήσεις.

Πιο συγκεκριμένα, αν τα δεδομένα του συστήματος συστάσεων μοντελοποιηθούν υπό τη μορφή γράφου, με τα αντικείμενα και τους χρήστες να αποτελούν τους κόμβους του και τις αξιολογήσεις τις ακμές που τους συνδέουν (μια αναπαράσταση που θα παρουσιαστεί με περισσότερη λεπτομέρεια στην Ενότητα 2.1.3), τότε το επαγόμενο δίκτυο εμφανίζει τα χαρακτηριστικά των *δικτύων ελεύθερης κλίμακας* (scale free networks) [Barabasi and Albert, 1999], δηλαδή η πιθανότητα σε έναν κόμβο (χρήστη, αντικειμένου) να προσπίπτουν k ακμές (αξιολογήσεις) προκύπτει από μια εκθετική κατανομή του τύπου

$$P(k) \sim k^{-\gamma}, \quad \gamma \in (2,3) \tag{1.1}$$



(α′) Συλλογή δεδομένων Epinions [Massa and Bhattacharjee, 2004]



(β′) Συλλογή δεδομένων Flixster [Jamali, 2010]

Σχήμα 1.4: Κατανομή πλήθους χρηστών και αντικειμένων ανά πλήθος αξιολογήσεων σε δημόσια διαθέσιμες συλλογές δεδομένων συστημάτων συστάσεων

όπου γ σταθερά που εξαρτάται από το δίκτυο. Ο τρόπος που αναπαρίστανται τα δίκτυα ελεύθερης κλίμακας απεικονίζεται στα παραδείγματα του Σχήματος 1.4, όπου σε λογαριθμικούς άξονες (log-log) έχει αποτυπωθεί το πλήθος των αξιολογήσεων που έχει δώσει (λάβει) το πλήθος των χρηστών (αντικειμένων), σε δύο γνωστές δημόσια διαθέσιμες συλλογές δεδομένων συστημάτων συστάσεων· του Epinions [Massa and Bhattacharjee, 2004] (Σχήμα 1.4α΄) και του Flixster [Jamali, 2010] (Σχήμα 1.4β΄). Αυτό που παρατηρείται (και που χαρακτηρίζει όλα τα δίκτυα ελεύθερης κλίμακας) είναι ότι ο βαθμός των κόμβων, όταν σχεδιάζεται σε λογαριθμικούς άξονες, εμφανίζεται υπό μορφή ευθείας της μορφής

$$y = -ax + b, \quad a, b > 0 \tag{1.2}$$

όπου a, b θετικές σταθερές που εξαρτώνται από τα δεδομένα. Η ίδια παρατήρηση ισχύει και για το παράδειγμα των προαναφερόμενων συλλογών δεδομένων, οι οποίες παρουσιάζουν αντίστοιχη συμπεριφορά, όπως φαίνεται στο Σχήμα 1.4, παρότι αφορούν δύο εντελώς διαφορετικά συστήματα και έχουν συλλεχθεί σε διαφορετικές χρονικές περιόδους. Άρα, είναι λογικό το συμπέρασμα πως η ελεύθερη κλίμακα αποτελεί γενικότερα δομικό χαρακτηριστικό των RS.

Συνεπώς, η εκθετική κατανομή της πυκνότητας των αξιολογήσεων επηρεάζει την λειτουργία των συστημάτων συστάσεων και άρα θα πρέπει να λαμβάνεται υπόψη κατά τη λειτουργία τους. Για παράδειγμα, τα πολύ δημοφιλή αντικείμενα θα μπορούσαν να προταθούν στους καινούργιους χρήστες του συστήματος (Ενότητα 1.1.2). Από την άλλη όμως πρέπει να ληφθεί υπόψη ότι ένα σύστημα συστάσεων που κάνει πολύ προφανείς προτάσεις ενδεχομένως να μην κερδίσει την εμπιστοσύνη τους. Επίσης, μια γενική μοντελοποίηση των χρηστών, η οποία δεν λαμβάνει υπόψη της το πόσο ενεργοί είναι αυτοί, δεν θα ήταν τόσο αποδοτική: είναι φανερό ότι θα πρέπει να γίνουν διαφορετικές θεωρήσεις για τους πολύ ενεργούς χρήστες απ΄ ότι για τους περιστασιακούς.

1.1.2 Ψυχρή εκκίνηση

Το πρόβλημα της *ψυχρής εκκίνησης* (cold-start problem) αναφέρεται στην συμπεριφορά του συστήματος συστάσεων όταν προστίθενται σε αυτό νέοι χρήστες και νέα αντικείμενα. Καταρχήν, υπάρχουν διαφορετικές θεωρήσεις για το πότε ένας χρήστης ή ένα αντικείμενο εντάσσεται σε αυτή την κατηγορία. Μια κοινή παραδοχή είναι να θεωρούνται ως *ψυχροί χρήστες* (cold-start users) και *ψυχρά αντικείμενα* (cold-start items) όσοι χρήστες και αντικείμενα αντίστοιχα έχουν δώσει (λάβει) κάτω από πέντε αξιολογήσεις [Massa and Avesani, 2009]. Άλλες περιπτώσεις εξετάζουν μονάχα αυτούς τους χρήστες και τα αντικείμενα που δεν έχουν καθόλου αξιολογήσεις ή που έχουν εισαχθεί στο σύστημα εντός ενός συγκεκριμένου χρονικού διαστήματος (λ.χ. την τελευταία ημέρα) [Shani and Gunawardana, 2011]. Σε κάθε περίπτωση, αποτελούν ένα ιδιαίτερα ενδιαφέρον τμήμα των χρηστών και των αντικειμένων του συστήματος συστάσεων, τόσο λόγω του πλήθους τους (Ενότητα 1.1.1) όσο και για την δυνατότητα του συστήματος να προτείνει τα καινούργια και κατά μια έννοια, μη αναμενόμενα, αντικείμενα.

Πρέπει να επισημανθεί ότι το συγκεκριμένο πρόβλημα δεν αφορά όλα τα RS στον ίδιο βαθμό. Για παράδειγμα, τα δημογραφικά συστήματα συστάσεων αφού εντοπίσουν την κατηγορία στην οποία ανήκει ο νέος χρήστης μπορούν αμέσως να του προτείνουν αντικείμενα που αρέσουν στους χρήστες της εν λόγω κατηγορίας. Αντίστοιχα, τα συστήματα συστάσεων βασισμένα στη γνώση επεξεργάζονται τα στοιχεία ομοιότητας του καινούργιου αντικειμένου με άλλα αντικείμενα στη βάση τους και το προτείνουν σε χρήστες που τους έχουν αρέσει παρόμοια αντικείμενα κατά το παρελθόν. Άλλα συστήματα συστάσεων όμως επηρεάζονται περισσότερο, όπως λόγου χάρη, τα συνεργατικά, τα οποία δεν μπορούν να υπολογίσουν το πόσο μοιάζει ένας χρήστης που δεν έχει καμία βαθμολογία με τους άλλους χρήστες του συστήματος. Το ίδιο ισχύει και με τα συστήματα συστάσεων βασισμένα στο αντικείμενο αφού δεν μπορούν να υπολογίσουν την ομοιότητα ενός αντικειμένου χωρίς αξιολογήσεις από τα υπόλοιπα.

Μια ευρεία γκάμα τεχνικών έχουν προταθεί για να αντιμετωπιστεί αυτό το φαινόμενο. Συνηθέστερα, οι κατασκευαστές των RS ζητούν από τους νέους χρήστες να αξιολογήσουν έναν αριθμό αντικειμένων προτού λάβουν προτάσεις από το σύστημα. Εναλλακτικά, μπορεί να τους ζητήσουν να συμπληρώσουν κάποια ερωτηματολόγια προκειμένου να τους εντάξουν σε κάποια «στερεοτυπική» κατηγορία και κατόπιν να προτείνουν αντικείμενα που απευθύνονται σε μέλη της συγκεκριμένης κατηγορίας. Ωστόσο και οι δύο αυτές μέθοδοι απαιτούν την ενεργή αλληλεπίδραση από την πλευρά του χρήστη, η οποία δεν είναι δεδομένη, τουλάχιστον στα αρχικά στάδια της χρήσης του συστήματος. Επιπλέον υπάρχει και το πρόβλημα της επιλογής των αντικειμένων που θα προταθούν για αξιολόγηση, ενώ η χρήση των ερωτηματολογίων δεν μπορεί να εγγυηθεί σε όλες τις περιπτώσεις ότι ο χρήστης θα τοποθετηθεί στη σωστή κατηγορία.

Μια άλλη επιλογή είναι να προτείνονται στους νέους χρήστες τα πιο δημοφιλή αντικείμενα ή αντίστοιχα να προτείνονται τα νέα αντικείμενα στους πιο ενεργούς χρήστες. Και σε αυτές τις περιπτώσεις, όμως, το αποτέλεσμα δεν είναι εγγυημένο. Αν γίνονται τετριμμένες προτάσεις στους νέους χρήστες, τότε αυτοί με τη σειρά τους ενδέχεται να μην βρουν ενδιαφέρον στην αλληλεπίδρασή τους με το σύστημα συστάσεων και να το εγκαταλείψουν. Κάτι αντίστοιχο μπορεί να συμβεί και με τους ενεργούς χρήστες, δηλαδή να χάσουν την εμπιστοσύνη τους στο σύστημα, όταν αντιληφθούν ότι οι προτάσεις που τους γίνονται είναι σε έναν βαθμό τυχαίες. Σε κάθε περίπτωση, πάντως, δεν υπάρχει κάποια κοινά αποδεκτή μέθοδος αντιμετώπισης του συγκεκριμένου ζητήματος, το οποίο χαρακτηρίζει περισσότερο το ίδιο τα σύστημα συστάσεων ως ολότητες παρά την απόδοση συγκεκριμένων επιμέρους αλγορίθμων.

1.1.3 Εμπιστοσύνη στις παραγόμενες συστάσεις

Το ζήτημα της *εμπιστοσύνης* (confidence) στις παραγόμενες συστάσεις (η οποία είναι μια τελείως διαφορετική έννοια από τα *δίκτυα εμπιστοσύνης* - trust networks - που παρουσιάζονται στο Κεφάλαιο 5) εξετάζεται από δύο πλευρές [Shani and Gunawardana, 2011]. Η πρώτη αφορά την εμπιστοσύνη του ίδιου του συστήματος στον εαυτό του, δηλαδή στο κατά πόσον οι συστάσεις που παράγει είναι σύμφωνες με τα γούστα των χρηστών του. Η δεύτερη σχετίζεται με τις προσδοκίες των ίδιων των χρηστών από το σύστημα, κατά πόσον δηλαδή οι ίδιοι οι χρήστες θεωρούν ότι έχουν να ωφεληθούν από την αλληλεπίδρασή τους με αυτό. Για παράδειγμα, είναι γνωστό ότι η απόδοση των συνεργατικών συστημάτων αυξάνει όσο περισσότερα αντικείμενα βαθμολογούν οι χρήστες. Άρα είναι λογικό το σύστημα να έχει περισσότερη εμπιστοσύνη στις συστάσεις που παράγει, όσο αυξάνεται ο όγκος των δεδομένων που επεξεργάζεται.

Έχει ιδιαίτερη σημασία το σύστημα να ενημερώνει τον χρήστη για την εμπιστοσύνη των παραγόμενων συστάσεων. Για παράδειγμα, αν το σύστημα προτείνει ένα αντικείμενο με χαμηλό επίπεδο εμπιστοσύνης, τότε το γεγονός αυτό λειτουργεί ως προτροπή προς το χρήστη να ερευνήσει το αντικείμενο περισσότερο προτού λάβει κάποια απόφαση. Ή αν προτείνει δύο αντικείμενα με τον ίδιο βαθμό ωφέλειας αλλά με διαφορετικά επίπεδα εμπιστοσύνης (λ.χ δύο βιβλία), τότε ο χρήστης θα μπορούσε δυνητικά να αγοράσει το ένα και να ερευνήσει περισσότερο το δεύτερο, διαβάζοντας κριτικές κ.λ.π. Έχει επίσης παρατηρηθεί ότι η εμπιστοσύνη ενός χρήστη προς το σύστημα αυξάνει αν το τελευταίο δικαιολογεί τις συστάσεις που παράγει.

Οι πιο συνηθισμένοι τρόποι περιγραφής της εμπιστοσύνης έχουν τις ρίζες τους στην στατιστική συμπερασματολογία και έχουν είτε την μορφή της πιθανότητας μια προτεινόμενη τιμή να είναι αληθής ή τη μορφή διαστήματος γύρω από την προτεινόμενη τιμή όπου βρίσκεται ένα προκαθορισμένο ποσοστό των αληθών τιμών. Για παράδειγμα, μπορεί σε ένα χρήστη να προταθεί ένα αντικείμενο με τιμή ωφέλειας 3 και πιθανότητα 0.9 ή να του κοινοποιηθεί ότι το 95% διάστημα εμπιστοσύνης για το συγκεκριμένο αντικείμενο είναι το [2, 3.5]. Σε άλλες περιπτώσεις πάλι, μπορεί να δοθεί ολόκληρη η κατανομή της προβλεπόμενης ωφέλειας [McLaughlin and Herlocker, 2004]. Η μέτρηση της εμπιστοσύνης μπορεί να χρησιμοποιηθεί

και ως ένα συμπληρωματικό χαρακτηριστικό από το RS, για την περαιτέρω επεξεργασία των υπό πρόταση αντικειμένων. Μια τέτοια λειτουργία θα μπορούσε να αποκλείσει προτεινόμενα αντικείμενα τα οποία έχουν μικρότερο βαθμό εμπιστοσύνης από ένα προκαθορισμένο κατώφλι.

1.2 Συνεισφορά της Διατριβής

Λαμβάνοντας υπόψη τις προαναφερόμενες προκλήσεις και προβλήματα, η συνεισφορά της διατριβής μπορεί να συνοψιστεί στα παρακάτω σημεία:

1. Στην πρόταση ενός πρωτότυπου συνεργατικού συστήματος συστάσεων, το οποίο αντιμετωπίζει σε ικανοποιητικό βαθμό τα ζητήματα της αραιότητας των αξιολογήσεων και της ψυχρής εκκίνησης. Για κάθε χρήστη του συστήματος κατασκευάζεται ένα μοντέλο, δομικό στοιχείο του οποίου είναι ένα τεχνητό νευρωνικό δίκτυο. Η καινοτομία της πρότασης έγκειται στα εξής σημεία:

   - Στη συνάρτηση μεταφοράς που χρησιμοποιείται στους κρυφούς νευρώνες του δικτύου, η οποία βασίζεται στην k-διαχωρισιμότητα. Είναι διαφορετική από αυτές που έχουν μέχρι σήμερα χρησιμοποιηθεί στα συστήματα συστάσεων (λ.χ σιγμοειδείς, γραμμικές, κ.λ.π) και επιτυγχάνει καλύτερη ταξινόμηση των δεδομένων εισόδου της. Αυτό έχει ως άμεσο αποτέλεσμα την μείωση των απαιτούμενων διαστάσεων του δικτύου, πράγμα που έχει ως συνέπεια την αύξηση της ταχύτητας εκπαίδευσης και παραγωγής νέων συστάσεων.
   - Στον αλγόριθμο μεταβλητής αρχιτεκτονικής δικτύου που χρησιμοποιείται. Έχοντας ως αφετηρία έναν ήδη υπάρχοντα κατασκευαστικό αλγόριθμο δικτύου, γίνονται εκείνες οι τροποποιήσεις και οι επεκτάσεις που οδηγούν το δίκτυο στη βέλτιστη αρχιτεκτονική, αποφεύγοντας τόσο τις περιπτώσεις υπερεκπαίδευσης όσο και ελλιπούς εκπαίδευσης.
   - Τέλος, στον αλγόριθμο ενισχυτικής μάθησης στη φάση της εκπαίδευσης του δικτύου, ο οποίος δεν παρέχει ομοιόμορφα τυχαία τα δείγματα εκπαίδευσης στο δίκτυο. Αντίθετα, τροφοδοτεί το δίκτυο συχνότερα με εκείνα τα δείγματα για τα οποία παρατηρείται το μεγαλύτερο σφάλμα εκπαίδευσης, έτσι ώστε να «αφομοιωθούν» καλύτερα.

2. Στην πρόταση ενός πρωτότυπου αλγορίθμου κοινωνικής συνεργατικής διήθησης, ο οποίος βασίζεται στην πραγματοποίηση τυχαίων περιπάτων επάνω στο μεικτό γράφο ομοιότητας και εμπιστοσύνης για την παραγωγή συστάσεων. Τα ιδιαίτερα χαρακτηριστικά της πρότασης είναι τα εξής:

   - Η σε κάθε βήμα πιθανοτική επιλογή αν ο επόμενος σταθμός του περιπάτου θα είναι κάποιος χρήστης που ο τρέχοντας εμπιστεύεται ή θα είναι κάποιος που παρουσιάζει παρόμοια αξιολογική συμπεριφορά με αυτόν.
   - Η μη-ομοιόμορφα τυχαία επιλογή του επόμενου βήματος του περιπάτου στην περίπτωση μετάβασης σε όμοιο χρήστη. Αντ΄ αυτού, η ομοιότητα στην αξιολογική συμπεριφορά μεταξύ του τρέχοντος χρήστη και των πιθανών επόμενων μοντελοποιείται ως μια πιθανοτική κατανομή, από την οποία προκύπτουν οι συγκεκριμένες πιθανότητες μετάβασης στον κάθε έναν από αυτούς.
   - Επειδή η προαναφερόμενη κατανομή είναι άγνωστη, χρησιμοποιείται η τεχνική της απορριπτικής δειγματοληψίας για την λήψη δειγμάτων από αυτήν, προκειμένου ο τυχαίος περίπατος να πραγματοποιήσει το επόμενο βήμα του.

3. Στην πρόταση μιας πρωτότυπης μεθόδου προσωπικής συσταδοποίησης των αντικειμένων που βρίσκονται στο περιβάλλον του χρήστη. Πιο συγκεκριμένα:

   - Ορίζεται το προσωπικό δίκτυο του κάθε χρήστη, το οποίο περιλαμβάνει τους χρήστες τους οποίους αυτός εμπιστεύεται καθώς και τους χρήστες με τους οποίους παρουσιάζει ομοιότητα στην αξιολογική συμπεριφορά.
   - Κατόπιν κατασκευάζεται ο γράφος που περιέχει τα αντικείμενα που έχουν αξιολογηθεί από κοινού από τον υπό εξέταση χρήστη και τα μέλη του προσωπικού του δικτύου. Στο συγκεκριμένο γράφο εφαρμόζεται ένας αλγόριθμος φασματικής συσταδοποίησης προκειμένου να εξαχθούν οι συστάδες στις οποίες εντάσσονται τα εν λόγω αντικείμενα.
   - Για κάθε συστάδα κατασκευάζεται ένα δίκτυο κατανάλωσης αντικειμένων το οποίο, εκτός των αντικειμένων της συστάδας, περιέχει και άλλα αντικείμενα τα οποία έχουν αξιολογήσει τα μέλη του προσωπικού δικτύου του υπό εξέταση χρήστη, αλλά όχι ο ίδιος. Τέλος, οι συστάσεις προκύπτουν με την πραγματοποίηση ενός τυχαίου περιπάτου στο συγκεκριμένο δίκτυο.

4. Στην χρήση μεθόδων παραγοντοποίησης πινάκων σε ένα εντελώς νέο πλαίσιο στα κοινωνικά συστήματα συστάσεων. Συνήθως, η προαναφερόμενη τεχνική χρησιμοποιείται για την εύρεση των λανθανόντων παραγόντων που περιγράφουν τις επιλογές των χρηστών και τα χαρακτηριστικά των αντικειμένων. Στη συγκεκριμένη περίπτωση, ωστόσο, το πεδίο εφαρμογής της μεθόδου αλλάζει· χρησιμοποιείται για την ανεύρεση εκείνων των λανθανόντων παραγόντων που καθορίζουν τις κοινότητες στις οποίες συμμετέχουν οι «φίλοι-φίλων» ενός συγκεκριμένου χρήστη. Πιο συγκεκριμένα, η προτεινόμενη μέθοδος:

   - Τροποποιεί διαδεδομένους αλγορίθμους μπεϋζιανής μη-αρνητικής παραγοντοποίησης πινάκων, έτσι ώστε αυτοί να μπορούν να προσαρμοστούν καλύτερα στο υπό εξέταση πρόβλημα. Η προσαρμογή επιτυγχάνεται με την χρήση κατάλληλης συνάρτησης εκ των προτέρων πιθανότητας, η οποία μετριάζει την ιδιαίτερα τοπική φύση του αλγορίθμου παραγοντοποίησης, εισάγοντας στην διαδικασία μακροσκοπικές πληροφορίες για το δίκτυο.
   - Επιλέγει την κατάλληλη συνάρτηση εκ των προτέρων πιθανότητας από τα μοντέλα εκθετικών τυχαίων γράφων και όχι από τη θεωρία των συζυγών εκ των προτέρων κατανομών (όπως γίνεται σε άλλες τεχνικές). Και πάλι, ο λόγος που επιβάλλει αυτή την επιλογή είναι τα ιδιαίτερα χαρακτηριστικά του δικτύου που αναλύεται.
   - Παρουσιάζει αναλυτικά το μοντέλο 2-αστέρων, που περιγράφει παραστατικά τα δίκτυα «φίλων-φίλου», και κατόπιν προχωρά στον προσεγγιστικό υπολογισμό των υπερπαραμέτρων του, κάνοντας χρήση της θεωρίας του μέσου πεδίου.

1.3 Δομή της Διατριβής

Η παρούσα διατριβή είναι διαρθρωμένη σε 9 Κεφάλαια και σε 1 Παράρτημα, με το παρόν κεφάλαιο (Κεφάλαιο 1) να αποτελεί την εισαγωγή.

Στο Κεφαλαίο 2 γίνεται μια επισκόπηση των συστημάτων συστάσεων, όσον αφορά την αναπαράσταση της αποθηκευμένης σε αυτά πληροφορίας (ως πίνακα αξιολογήσεων και ως διμερή γράφο), αλλά και των κυριότερων μεθοδολογιών παραγωγής των συστάσεων (βασισμένες στο περιεχόμενο, συνεργατικές, υβριδικές, κ.λ.π.). Επίσης τονίζεται πως οι πιο ευρύτερα χρησιμοποιούμενες τεχνικές είναι οι συνεργατικές, οι οποίες εξετάζονται εκτενέστερα όσον αφορά τον τρόπο παραγωγής των συστάσεων (μνημονικός ή κατασκευής μοντέλου).

Στο Κεφάλαιο 3 αναλύονται οι τρόποι αξιολόγησης της λειτουργίας των συστημάτων συστάσεων. Αρχικά αναφέρονται οι κλασσικές μέθοδοι μέτρησης της ακρίβειας της πρόβλεψης και της ταξινόμησης. Στη συνέχεια, εξετάζονται οι τρόποι εκτίμησης της ποιότητας των παραγόμενων συστάσεων καθώς και ο βαθμός στον οποίο τα RS μπορούν να παράξουν προτάσεις για όλους τους χρήστες και τα αντικείμενα που περιέχουν. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση ειδικών μετρικών αξιολόγησης που απευθύνονται σε ενδιαφέροντα υποσύνολα χρηστών και αντικειμένων του RS.

Στο Κεφάλαιο 4 παρουσιάζεται ένα συνεργατικό σύστημα συστάσεων, το οποίο αξιοποιεί την έμμεση κοινωνική πληροφορία για την παραγωγή συστάσεων, με στόχο την αντιμετώπιση, ως ένα βαθμό, του ζητήματος της αραιότητας των βαθμολογιών και της ψυχρής εκκίνησης. Δομικό του στοιχείο αποτελεί ένα νευρωνικό δίκτυο μεταβλητού μεγέθους και αυτό που το διαφοροποιεί από άλλα αντίστοιχα συστήματα είναι η χρήση της k-διαχωρισιμότητας ως συνάρτησης μεταφοράς, η χρήση της ενισχυτικής μάθησης στη διάρκεια της εκπαίδευσης καθώς και ο κατασκευαστικός αλγόριθμος της αρχιτεκτονικής του δικτύου. Η υλοποίησή του και ο συγκριτικός έλεγχος της απόδοσής του σε σχετικές δημόσια διαθέσιμες συλλογές δεδομένων εμφανίζει ενθαρρυντικά αποτελέσματα και φαίνεται να δικαιολογεί τις επιλογές και τις παραδοχές που έχουν γίνει.

Στο Κεφάλαιο 5 γίνεται μια επισκόπηση του τρόπου που εντάσσεται η άμεση κοινωνική πληροφορία στα RS. Καταρχήν, πραγματοποιείται μια εισαγωγή στον τρόπο μοντελοποίησης των κοινωνικών δικτύων μέσω μιας ερευνητικής περιοχής που είναι γνωστή ως ανάλυση κοινωνικών δικτύων. Παρουσιάζονται τα δομικά στοιχεία των δικτύων, οι τύποι τους καθώς και η προέλευση των δεδομένων τους. Κατόπιν, εξετάζονται οι τρόποι αλληλεπίδρασης στα κοινωνικά δίκτυα, τόσο μεταξύ των χρηστών όσο και μεταξύ χρηστών και αντικειμένων. Τέλος, γίνεται ιδιαίτερη αναφορά σε μια ειδική κατηγορία κοινωνικών δικτύων που χρησιμοποιούνται πολύ στα συστήματα συστάσεων, στα δίκτυα εμπιστοσύνης.

Στο Κεφάλαιο 6 παρουσιάζεται ένα κοινωνικό συνεργατικό σύστημα συστάσεων, το οποίο βασίζεται στους μη-αμερόληπτους τυχαίους περιπάτους για την παραγωγή των προτάσεων. Ο συγκεκριμένος τυχαίος περίπατος πραγματοποιείται επάνω σε ένα μεικτό γράφο, στον οποίο έχουν συνδυαστεί η άμεση και η έμμεση κοινωνική πληροφορία. Η πρωτοτυπία της μεθόδου που προτείνεται έγκειται στη χρήση της απορριπτικής δειγματοληψίας για τον υπολογισμό του επόμενου βήματος του τυχαίου περιπάτου. Η συγκεκριμένη επιλογή φαίνεται να επαληθεύεται και πειραματικά

Στο Κεφάλαιο 7 αναπτύσσεται μια μεθοδολογία κοινωνικών συνεργατικών συστάσεων που βασίζεται στην εφαρμογή προσωποποιημένης συσταδοποίησης των αντικειμένων. Στη βάση της μεθόδου βρίσκεται το προσωπικό δίκτυο του κάθε χρήστη, το οποίο αποτελείται από τους άμεσους γείτονές του στο κοινωνικό δίκτυο καθώς και από άλλους χρήστες με τους οποίους εμφανίζει παρόμοια αξιολογική συμπεριφορά. Από το εν λόγω δίκτυο κατασκευάζεται το δίκτυο κατανάλωσης αντικειμένων, πάνω στο οποίο εφαρμόζεται αλγόριθμος φασματικής συσταδοποίησης με σκοπό την ανεύρεση μοτίβων κατανάλωσης μεταξύ των αντικειμένων. Στο τέλος, η πραγματοποίηση ενός τυχαίου περιπάτου στις σχηματισμένες συστάδες επιστρέφει τα υποψήφια προς πρόταση αντικείμενα.

Στο Κεφάλαιο 8 περιγράφεται ένα σύστημα κοινωνικών συνεργατικών συστάσεων, το οποίο βασίζεται σε μια πρωτότυπη τεχνική συσταδοποίησης των «φίλων-φίλου» ενός χρήστη, σε επικαλυπτόμενες κοινότητες. Βάση της προαναφερόμενης συσταδοποίησης αποτελεί ένας μπεϋζιανός αλγόριθμος μη-αρνητικής παραγοντοποίησης πινάκων, ο οποίος εξάγει τους λανθάνοντες παράγοντες που περιγράφουν το βαθμό συμμετοχής των μελών του δικτύου στην κάθε κοινότητα. Η τοπική φύση της προτεινόμενης μεθόδου μετριάζεται με την χρήση των μοντέλων εκθετικών τυχαίων γράφων στη θέση της εκ των προτέρων πιθανότητας, τα οποία εισάγουν στην παραγοντοποίηση μακροσκοπικές πληροφορίες για τον γράφο.

Στο Κεφάλαιο 9 πραγματοποιείται μια συνολική επισκόπηση της συνεισφοράς της διατριβής στην αντιμετώπιση των προκλήσεων που παρουσιάστηκαν στο παρόν κεφάλαιο. Επίσης, στη

βάση των πορισμάτων της ήδη υπάρχουσας συνεισφοράς, αναφέρονται πιθανές μελλοντικές ερευνητικές κατευθύνσεις.

Τέλος, στο Παράρτημα Α΄ αναλύεται η προσεγγιστική μέθοδος που ακολουθήθηκε για την εύρεση των υπερπαραμέτρων του μοντέλου εκθετικών τυχαίων γράφων 2-αστέρων, το οποίο χρησιμοποιήθηκε ως η συνάρτηση της εκ των προτέρων πιθανότητας στη μπεϋζιανή μη-αρνητική παραγοντοποίηση πινάκων του Κεφαλαίου 8.

□

Κεφάλαιο 2

Συστήματα Συστάσεων

Στο προηγούμενο Κεφάλαιο έγινε αναφορά στο γεγονός πως τα συστήματα συστάσεων αποτελούν μια κατηγορία λογισμικού που χρησιμοποιείται για την πρόταση *αντικειμένων* σε *χρήστες*. Συνεπώς τα δομικά τους στοιχεία είναι το σύνολο όλων των χρηστών (των εγγεγραμμένων, λόγου χάρη, σε μια online υπηρεσία) και το σύνολο όλων των αντικειμένων (που προσφέρονται από την online υπηρεσία). Παρότι και τα δύο σύνολα είναι πεπερασμένα, το μέγεθος τους δεν παραμένει σταθερό, αλλά αντίθετα αυξάνεται καθώς νέοι χρήστες συνεχίζουν να εγγράφονται στην υπηρεσία αλλά και νέα αντικείμενα να προστίθενται στα ήδη υπάρχοντα. Δεν είναι σπάνιο το φαινόμενο το σύνολο των χρηστών να φτάσει στο επίπεδο των εκατομμυρίων και το σύνολο των αντικειμένων να ξεπεράσει το επίπεδο των χιλιάδων (λ.χ. στο ηλεκτρονικό κατάστημα της Amazon).

Οι χρήστες έχουν στη διάθεσή τους περιορισμένους πόρους (χρόνο, υπολογιστική ισχύ κλπ) και άρα δεν μπορούν να προσπελάσουν όλα τα διαθέσιμα αντικείμενα. Αντίθετα «βλέπουν», ή πιο ορθά αξιολογούν, ένα πολύ μικρό υποσύνολο από αυτά. Έτσι, λοιπόν, είναι παραπάνω από βέβαιο ότι θα έχουν διαφύγει της οπτικής τους αντικείμενα τα οποία μπορεί να τους φαινόντουσαν χρήσιμα. Αυτή η παρατήρηση φανερώνει και την αξία των συστημάτων συστάσεων· να βοηθήσουν, δηλαδή, τους χρήστες να βρουν αντικείμενα τα οποία δεν γνωρίζουν ακόμα και τα οποία μπορεί να τους ενδιαφέρουν.

2.1 Παρουσίαση

2.1.1 Ορισμός

Ένας περισσότερο επίσημος ορισμός των συστημάτων συστάσεων δίνεται στην εργασία των [Adomavicius and Tuzhilin, 2005]. Έστω C το σύνολο των χρηστών και S το σύνολο των αντικειμένων. Έστω επίσης χρήστης $c \in C$ και $S_c \in S$ το υποσύνολο των αντικειμένων που αυτός έχει αξιολογήσει. Η χρησιμότητα ενός αντικειμένου $s \in S \backslash S_c$, το οποίο δεν είναι γνωστό στον c, μετριέται με την χρήση μιας *συνάρτησης ωφέλειας* (utility function) u. Σκοπός του συστήματος συστάσεων είναι να επιλέξει εκείνο το άγνωστο αντικείμενο $s' \in S \backslash S_c$, το οποίο θα μεγιστοποιήσει την συνάρτηση u για τον χρήστη c:

$$s_c = \arg \max_{s' \in S \backslash S_c} u(c, s'), \quad \forall c \in C \tag{2.1}$$

2.1.2 Αναπαράσταση ως Πίνακας Αξιολογήσεων

Ως συνάρτηση ωφέλειας θα μπορούσε να χρησιμοποιηθεί μια οποιαδήποτε συνάρτηση, όπως λόγου χάρη, μια *συνάρτηση κέρδους* (profit funciton) [Adomavicius and Tuzhilin, 2005]. Ωστόσο, στα περισσότερα συστήματα συστάσεων η ωφέλεια ενός συγκεκριμένου αντικειμένου

Πίνακας 2.1: Παράδειγμα Πίνακα Αξιολογήσεων ενός Συστήματος Συστάσεων

	I_1	I_2	I_3	I_4
U_1	5	3	2	-
U_2	3	5	-	2
U_3	1	-	2	-
U_4	-	2	-	3

για ένα χρήστη περιγράφεται από την *αξιολόγησή* του ή την *βαθμολογία* του (rating), η οποία και αποτελεί μια ένδειξη του κατά πόσο του άρεσε. Οι αξιολογήσεις λαμβάνουν διακριτές τιμές εντός ενός συγκεκριμένου εύρους· μπορεί να είναι δυαδικές (λ.χ. μου αρέσει/δεν μου αρέσει) ή μπορεί να ανήκουν σε ένα ευρύτερο σύνολο (λ.χ. η κλίμακα 5 αστέρων που χρησιμοποιείται στα ξενοδοχεία ή σε οδηγούς διασκέδασης).

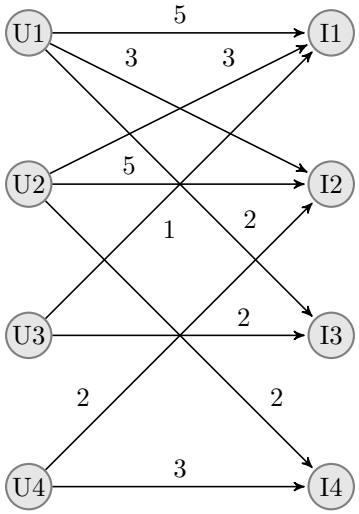
Στον Πίνακα 2.1 αναπαρίσταται ένας πίνακας αξιολογήσεων ενός υποθετικού συστήματος συστάσεων, με τις γραμμές να υποδηλώνουν τους χρήστες και τις στήλες τα αντικείμενα. Το (i, j) στοιχείο του πίνακα αντιπροσωπεύει την αξιολόγηση του αντικείμενου j από τον χρήστη i. Στο συγκεκριμένο παράδειγμα, τα αντικείμενα αξιολογούνται στην πενταβάθμια κλίμακα με το ένα να είναι ο χαμηλότερος δυνατός βαθμός (υποδηλώνει την ελάχιστη αποδοχή) και το πέντε να είναι ο υψηλότερος βαθμός (υποδηλώνει την μέγιστη αποδοχή). Όπως φαίνεται στον εν λόγω πίνακα, ο δεύτερος χρήστης αξιολόγησε το πρώτο αντικείμενο με τρία, το δεύτερο με πέντε και το τέταρτο με δύο. Μια ερμηνεία αυτής της αξιολόγησης είναι πως το δεύτερο αντικείμενο του άρεσε πολύ, το τέταρτο αντικείμενο δεν του άρεσε, ενώ το πρώτο αντικείμενο το βρήκε περιορισμένα ενδιαφέρον.

Τα μη συμπληρωμένα στοιχεία του πίνακα συμβολίζουν το γεγονός ότι τα συγκεκριμένα αντικείμενα δεν έχουν αξιολογηθεί από τους αντίστοιχους χρήστες. Για παράδειγμα, το στοιχείο (2, 3) είναι άδειο. Το προφανές συμπέρασμα είναι ότι ο δεύτερος χρήστης δεν έχει αξιολογήσει ακόμα το τρίτο αντικείμενο. Η έλλειψη της συγκεκριμένης αξιολόγησης ερμηνεύεται με πολλούς τρόπους: λ.χ. το τρίτο αντικείμενο άφησε τον δεύτερο χρήστη εντελώς αδιάφορο και για αυτό δεν ασχολήθηκε καθόλου με τη βαθμολόγησή του. Από την άλλη, όμως, θα μπορούσε και να μην είχε αποφασίσει ακόμα για το ποια θα ήταν η κατάλληλη βαθμολογία. Άρα, απαιτείται μια υψηλού επιπέδου ανάλυση για να βρεθεί ο λόγος για τον οποίο τα συγκεκριμένα αντικείμενα δεν έχουν ακόμα αξιολογηθεί, μια ανάλυση που είναι δύσκολη ακόμα και για τους ανθρώπους.

Τα Συστήματα Συστάσεων κάνουν την παραδοχή ότι ένας δεδομένος χρήστης δεν έχει αξιολογήσει ένα αντικείμενο μόνο και μόνο γιατί δεν έχει περιέλθει στη γνώση του μέχρι στιγμής (και όχι, λ.χ., γιατί δεν του άρεσε και για αυτό απέφυγε να το αξιολογήσει). Αυτή η παραδοχή έχει βάση γιατί σε πραγματικές συνθήκες το σύνολο των διαθέσιμων αντικειμένων είναι πάρα πολύ μεγάλο και άρα η πιθανότητα κάποιος χρήστης να έχει λάβει γνώση ενός αντικειμένου αλλά να μην το έχει αξιολογήσει σκοπίμως είναι πάρα πολύ μικρή. Συνεπώς, το σύνολο των υποψηφίων προς σύσταση αντικειμένων, τα οποία δίνονται ως είσοδος στον αλγόριθμο παραγωγής συστάσεων, είναι το σύνολο των αντικειμένων που δεν έχουν ακόμα αξιολογηθεί. Στη συνέχεια, το σύστημα προσπαθεί να «προβλέψει» τις τιμές που λείπουν, δηλαδή να «μαντέψει» την αξιολόγηση που θα έκανε ένας συγκεκριμένος χρήστης σε ένα συγκεκριμένο αντικείμενο. Από τη στιγμή που ο μηχανισμός συστάσεων αρχίζει να παράγει αποτελέσματα, μπορεί να χρησιμοποιηθεί ένας απλός κανόνας για την παραγωγή των συστάσεων (π.χ. προτείνονται τα πρώτα N αντικείμενα με την καλύτερη αξιολόγηση).

2.1.3 Αναπαράσταση ως Διμερής Γράφος

Παρότι είναι ο πλέον συνηθισμένος, ο πίνακας αξιολογήσεων δεν είναι ο μοναδικός τρόπος αναπαράστασης των δεδομένων ενός συστήματος συστάσεων. Αρκετά διαδεδομένη είναι και



Σχήμα 2.1: Αναπαράσταση του Πίνακα Αξιολογήσεων υπό τη μορφή Διμερούς Γράφου

η αναπαράσταση υπό μορφή γράφου [Perugini et al., 2004]. Σε αυτήν, οι χρήστες και τα αντικείμενα μετασχηματίζονται στους κόμβους ενός γράφου, οι οποίοι συνδέονται μεταξύ τους με *κατευθυντικές ακμές με βάρη* τις αξιολογήσεις (weighted directed edges)

$$G = (V, E), \quad V = \{u, i : \forall u \in U, \forall i \in I\}, \quad E = \{(u, i) : \forall u \in U, \forall i \in I, \exists r_{u,i} \in R\} \quad (2.2)$$

Μια χρήσιμη παρατήρηση σε αυτό το σημείο είναι ότι η μοναδική δυνατή σχέση σε αυτόν τον γράφο είναι μεταξύ χρηστών και αντικειμένων και όχι ανάμεσα στους χρήστες ή ανάμεσα στα αντικείμενα. Συνεπώς, οι κόμβοι του γράφου χωρίζονται σε δύο *ανεξάρτητα σύνολα* (independent sets), που δεν είναι άλλα, από το σύνολο των χρηστών U και το σύνολο των αντικειμένων I, σχηματίζοντας κατ΄ αυτόν τον τρόπο ένα *διμερή γράφο*

$$G = (U, I, E), \quad E = \{(u, i) : \forall u \in U, \forall i \in I, \exists r_{u,i} \in R\} \quad (2.3)$$

Η παρατήρηση αυτή είναι σημαντική γιατί μας επιτρέπει να χρησιμοποιήσουμε στα συστημάτων συστάσεων εξειδικευμένους αλγορίθμους για διμερείς γράφους (οι οποίοι παρουσιάζουν, λ.χ, μικρότερη υπολογιστική πολυπλοκότητα). Στο Σχήμα 2.1 παρουσιάζεται ο Πίνακας Αξιολογήσεων 2.1 υπό την μορφή διμερούς γράφου.

2.2 Παραγωγή συστάσεων

Σε αυτή την ενότητα θα γίνει μια γενική επισκόπηση των μεθόδων που χρησιμοποιούνται για την παραγωγή συστάσεων καθώς και μια πρώτη ταξινόμηση όσον αφορά την φιλοσοφία τους.

2.2.1 Βασισμένη στο περιεχόμενο

Τα *συστήματα συστάσεων βασισμένα στο περιεχόμενο* (content-based recommender systems) επιλέγουν να παρουσιάσουν σε ένα χρήστη αντικείμενα τα οποία έχουν ορισμένα κοινά χαρακτηριστικά με άλλα αντικείμενα που ο συγκεκριμένος χρήστης έχει ήδη δηλώσει ότι του αρέσουν. Ακολουθώντας τη σημειολογία της Ενότητας 2.1.1, η ωφέλεια $u(c, s)$ του αντικειμένου s για τον χρήστη c υπολογίζεται λαμβάνοντας υπόψη τις τιμές $u(c, s_i)$ που έχουν εκχωρηθεί από τον c στο σύνολο S_i (που περιέχει τα πιο «όμοια» με το s αντικείμενα στο S).

Για να γίνει εφικτός ο καθορισμός της ομοιότητας μεταξύ των αντικειμένων, αυτά αναπαρίστανται από ένα σύνολο *λέξεων-κλειδιών* (keywords), οι οποίες στοιχειοθετούν το *προφίλ*

τους (item profile). Οι λέξεις κλειδιά είτε ορίζονται άμεσα από το σχεδιαστή του συστήματος συστάσεων είτε εξάγονται έμμεσα με την χρήση στατιστικών μεγεθών όπως ο *δείκτης tf-idf* (term frequency-inverse document frequency) [Salton and Buckley, 1988]. Ένα παράδειγμα της πρώτης τεχνικής θα μπορούσε να είναι ένα μουσικό σύστημα συστάσεων που κατασκευάζει το προφίλ των κομματιών λαμβάνοντας υπόψη πληροφορίες σχετικά με το είδος της μουσικής, το άλμπουμ και τους καλλιτέχνες που συμμετέχουν, ενώ ένα παράδειγμα της δεύτερης τεχνικής θα μπορούσε να είναι ένα σύστημα συστάσεων για μια ενημερωτική ιστοσελίδα, το οποίο εξάγει λέξεις-κλειδιά από τα άρθρα που διαβάζει ο χρήστης.

Το κυριότερο μειονέκτημα της βασισμένης στο περιεχόμενο οπτικής είναι ότι συχνά τα αντικείμενα χαρακτηρίζονται από ιδιότητες που είναι δύσκολο να ποσοτικοποιηθούν (λ.χ. η ατμόσφαιρα ενός εστιατορίου, το κινηματογραφικό στυλ μιας ταινίας κ.ά.). Επιπλέον, η συνεχής αναζήτηση «παρόμοιων» αντικειμένων πολύ συχνά οδηγεί σε μια κατάσταση που είναι γνωστή ως *υπερεξειδίκευση* (overspecialization), όπου το σύστημα παράγει συνεχώς σχεδόν πανομοιότυπες προτάσεις.

2.2.2 Βασισμένη στη γνώση

Στα *συστήματα συστάσεων βασισμένα στη γνώση* (knowledge-based recommender systems) έχουν ενσωματωθεί, από τους κατασκευαστές τους, πληροφορίες σχετικές με τα αντικείμενα καθώς και με τις σχέσεις που τα διέπουν [Beliakov et al., 2011]. Για παράδειγμα, οι ιδιοκτήτες ενός μουσικού συστήματος συστάσεων μπορούν, με τη βοήθεια ειδικών, να κατατάξουν τα μουσικά κομμάτια σε κατηγορίες και επιπλέον να συσχετίσουν μεταξύ τους τα πιο όμοια. Έτσι, αν κάποιος χρήστης εκφράσει μια μουσική προτίμηση, τότε το σύστημα μπορεί να του επιστρέψει ως προτάσεις τα αντίστοιχα σχετικά μουσικά κομμάτια.

Το κυριότερο πλεονέκτημα των συστημάτων συστάσεων βασισμένων στη γνώση είναι ότι δεν χρειάζεται να υπολογίζουν κάθε φορά την ομοιότητα μεταξύ των αντικειμένων, όπως για παράδειγμα κάνουν τα συστήματα που βασίζονται στο περιεχόμενο. Με αυτόν τον τρόπο μπορούν να αντιμετωπιστούν δύο από τις προκλήσεις που αναφέρθηκαν στην Ενότητα 1.1, αυτή της αραιότητας του πίνακα αξιολογήσεων και της μικρής αλληλοκάλυψης των αξιολογήσεων. Παρόλα αυτά, στις περισσότερες περιπτώσεις η πρακτική τους υλοποίηση είναι δύσκολη, γιατί ακόμα και με τις συμβουλές των ειδικών, είναι δύσκολο να εντοπιστούν εκ των προτέρων όλοι οι δυνατοί τρόποι με τους οποίους μπορούν να συσχετιστούν τα αντικείμενα, πόσο μάλλον όταν το πλήθος τους είναι πολύ μεγάλο.

2.2.3 Συνεργατική Διήθηση

Τα συστήματα συστάσεων που βασίζονται στη *συνεργατική διήθηση* (collaborative filtering) προτείνουν σε ένα χρήστη αντικείμενα που αρέσουν σε άλλους χρήστες που όμως έχουν παρόμοιο γούστο με αυτόν. Ακολουθώντας και πάλι τη σημειολογία της Ενότητας 2.1.1, η ωφέλεια $u(c, s)$ ενός αντικειμένου s για τον χρήστη c υπολογίζεται λαμβάνοντας υπόψη τις τιμές $u(c_j, s)$ στο σύνολο C_j (που περιέχει εκείνο το υποσύνολο των χρηστών που έχουν βαθμολογήσει το s και έχουν το πιο «όμοιο» γούστο με τον c).

Στα συστήματα συνεργατικής διήθησης οι χρήστες συνήθως αναπαρίστανται από μια μοναδική ταυτότητα, η οποία μπορεί να είναι το *όνομα χρήστη* (username) ή γενικότερα ένας *αριθμός ταυτοποίησης* (identification number - id). Πολλά συστήματα συνεργατικής διήθησης χρησιμοποιούν αποκλειστικά και μόνο αυτή την πληροφορία για την παραγωγή συστάσεων, δηλαδή μόνο την ταυτότητα χρήστη συνοδευόμενη από ένα σύνολο αξιολογήσεων που αφορούν συγκεκριμένα αντικείμενα (τα οποία επίσης αναπαρίστανται από μια μοναδική ταυτότητα - Πίνακας 2.1). Έχοντας αυτά τα δεδομένα ως αφετηρία, τα συστήματα συνεργατικής διήθησης προσπαθούν να εντοπίσουν μοτίβα (ομοιότητες ή ανομοιότητες) στις αξιολογήσεις των χρηστών και στη συνέχεια χρησιμοποιούν κάποιες από αυτές τις αξιολογήσεις ως «πη-

γή πρόβλεψης» για την ωφέλεια που θα αποκτήσουν κάποιοι άλλοι χρήστες από τα εν λόγω αντικείμενα. Αυτή η προσέγγιση έχει ως επί το πλείστον χρησιμοποιηθεί και στην παρούσα διατριβή. Για παράδειγμα, στον Πίνακα 2.1 οι χρήστες U_1 και U_2 φαίνεται να έχουν αντίθετο γούστο· στον U_1 αρέσει το αντικείμενο I_1 ενώ στον U_2 όχι. Ακριβώς όμοια συμπεριφορά έχουν οι δύο αυτοί χρήστες και ως προς το αντικείμενο I_2. Συνεπώς, μιας και ο χρήστης U_2 έχει δώσει χαμηλή αξιολόγηση στο I_3, είναι λογικό να υποτεθεί ότι θα αρέσει στον χρήστη U_1. Ένα ανάλογο συμπέρασμα μπορεί να βγει σχετικά με το πόσο θα αρέσει στον χρήστη U_2 το αντικείμενο I_3.

Τα συστήματα συνεργατικής διήθησης μπορούν να συλλέξουν περισσότερες πληροφορίες για τους χρήστες (όπως λ.χ. την ηλικία τους, το φύλο τους, την μόρφωσή τους, την οικογενειακή τους κατάσταση κ.λ.π.) με στόχο την κατασκευή *προφίλ χρηστών* (user profiles). Στη συνέχεια ο αλγόριθμος παραγωγής συστάσεων τροφοδοτείται, εκτός από τις αξιολογήσεις των αντικειμένων και με τις προαναφερόμενες δημογραφικές πληροφορίες, μια τεχνική που είναι γνωστή και ως *κατασκευή «στερεοτύπων»* (stereotyping) [Godoy and Amandi, 2005]. Για παράδειγμα, ένα σύστημα για την πρόταση ταινιών που βασίζεται στη συνεργατική διήθηση μπορεί να φιλτράρει περαιτέρω τα αποτελέσματά του ανάλογα με την κατηγορία που ανήκει ο χρήστης (λ.χ. τα αγόρια στην εφηβική ηλικία τείνουν να προτιμούν τις ταινίες δράσης).

Παρά την εξαιρετική τους δημοφιλία και την μεγάλη τους διάδοση, τα συστήματα συνεργατικής διήθησης έχουν να αντιμετωπίσουν δύο από τα θεμελιώδη προβλήματα που αναφέρθηκαν στην Ενότητα 1.1. Πιο συγκεκριμένα πρόκειται για το πρόβλημα της παραγωγής συστάσεων για ένα νεοεισερχόμενο χρήστη καθώς και το πρόβλημα της αραιότητας των αξιολογήσεων. Αυτό ακριβώς το κενό έρχεται να καλύψει η παρούσα διατριβή, συνεισφέροντας θεωρητικά εργαλεία και μεθόδους για την αντιμετώπιση των προαναφερόμενων ζητημάτων και γενικότερα για την παραγωγή καλύτερων συστάσεων.

2.2.4 Δημογραφική Διήθηση

Τα συστήματα *δημογραφικής διήθησης* (demographic filtering) τοποθετούν τους χρήστες τους σε μία ή περισσότερες κατηγορίες ανάλογα με τα χαρακτηριστικά που περιέχονται στο προφίλ του καθενός τους [Beliakov et al., 2011]. Κάθε δημογραφική κατηγορία συσχετίζεται με έναν *αρχετυπικό* (archetype) χρήστη ή διαφορετικά, με ένα στερεότυπο ενός χρήστη. Οι προτιμήσεις του συγκεκριμένου χρήστη για μια ευρεία γκάμα αντικειμένων είναι γνωστές εκ των προτέρων και χρησιμοποιούνται για την αιτιολόγηση των παραγόμενων συστάσεων. Δομικό στοιχείο της δημογραφικής διήθησης αποτελεί η παρατήρηση ότι αρκετές φορές το γούστο δεν αποτελεί μόνο ατομική υπόθεση αλλά αντίθετα διαμορφώνεται στο πλαίσιο μιας ομάδας (λ.χ. το προηγούμενο παράδειγμα, δηλαδή το γεγονός πως στους περισσότερους εφήβους αρέσουν οι ταινίες δράσης).

Τα δημογραφικά συστήματα μοιάζουν πάρα πολύ με τα συνεργατικά συστήματα, μόνο που σε αυτήν την περίπτωση η ομοιότητα υπολογίζεται επάνω στα χαρακτηριστικά του προφίλ των χρηστών και όχι στις αξιολογήσεις που αυτοί έχουν δώσει. Αυτή η διαφορά λειτουργίας έχει ορισμένα πολύ σημαντικά πλεονεκτήματα. Κατ΄ αρχήν, οι ομοιότητες και οι διαφορές υπολογίζονται πάνω στο ίδιο σύνολο χαρακτηριστικών και συνεπώς είναι ανεξάρτητες από την αραιότητα των αξιολογήσεων και από την μικρή αλληλοκάλυψη των αξιολογούμενων αντικειμένων (Ενότητα 1.1), προβλήματα που επηρεάζουν τα συνεργατικά συστήματα. Ένα ακόμα σχετικό ζήτημα που αντιμετωπίζεται επιτυχώς είναι αυτό των διαστάσεων της σύγκρισης, αφού στα δημογραφικά συστήματα το σύνολο των χαρακτηριστικών που συγκρίνονται παραμένει σταθερό και δεν αυξάνεται διαρκώς, όπως συμβαίνει με τις συνεργατικές μεθόδους. Ωστόσο, έχουν βρει πολύ μικρή απήχηση γιατί παρουσιάζουν δύο πολύ βασικά μειονεκτήματα· το πρώτο είναι ότι στις περισσότερες περιπτώσεις άνθρωποι που ανήκουν στην ίδια ακριβώς δημογραφική ομάδα έχουν εντελώς διαφορετικό γούστο (λ.χ. μουσικές προτιμήσεις) και το δεύτερο είναι ότι σε αρκετά πεδία είναι δύσκολο να κατασκευαστεί ένα επακριβές προφίλ χρηστών το οποίο

να μπορεί να συσχετιστεί ικανοποιητικά με τις προτιμήσεις τους.

2.2.5 Βασισμένη στην ωφέλεια

Τα *συστήματα συστάσεων βασισμένα στην ωφέλεια* (utility-based recommender systems) αποτελούν μια πολύ ειδική κατηγορία συστημάτων συστάσεων, τα οποία βασίζονται στην παραδοχή ότι οι χρήστες προτιμούν αντικείμενα λόγω των συγκεκριμένων ιδιοτήτων που αυτά έχουν [Pu et al., 2011]. Το εύρος των τιμών που μπορούν να λάβουν οι ιδιότητες των αντικειμένων αντιστοιχεί σε διαφορετικούς βαθμούς αποδοχής τους από τον χρήστη, η οποία επιπλέον εξαρτάται και από την περίσταση. Για παράδειγμα, ένας οδηγός που ενδιαφέρεται να αγοράσει ένα μεγάλο αυτοκίνητο αντιλαμβάνεται ότι μπορεί ναι είναι πιο άνετο στα ταξίδια αλλά από την άλλη είναι λιγότερο πρακτικό στις μετακινήσεις μέσα σε μια μεγάλη πόλη. Έτσι λοιπόν η αποδοχή ενός αντικειμένου προκύπτει ως ένας συμβιβασμός ανάμεσα στα πλεονεκτήματα και τα μειονεκτήματα που έχει το κάθε χαρακτηριστικό του, δηλαδή πόσες φορές η ύπαρξή τους είναι επιθυμητή και πόσες φορές δεν είναι. Συνεπώς, η προτίμηση προκύπτει ως ζυγισμένη συνάρτηση των χαρακτηριστικών του αντικειμένου.

2.2.6 Υβριδική

Τα *υβριδικά συστήματα συστάσεων* (hybrid recommender systems) συνδυάζουν κάποιες (ή όλες) από τις προαναφερόμενες προσεγγίσεις, σε μια προσπάθεια να υπερβούν τις αδυναμίες της κάθε μιας χωριστά. Συνηθέστερα, ωστόσο, αφορούν συνδυασμούς συνεργατικών αλγορίθμων και αλγορίθμων βασισμένων στο περιεχόμενο. Υπάρχουν πολλοί τρόποι για να πραγματοποιηθεί αυτός ο συνδυασμός [Adomavicius and Tuzhilin, 2005]· ο πιο απλός είναι η ξεχωριστή υλοποίηση ενός συνεργατικού συστήματος και ενός συστήματος βασισμένου στο περιεχόμενο και κατόπιν ο συνδυασμός των εξόδων τους [Christakou and Stafylopatis, 2005; Symeonidis et al., 2008]. Μια άλλη προσέγγιση είναι ο εμπλουτισμός μιας μεθόδου βασισμένης στο περιεχόμενο με συνεργατικά χαρακτηριστικά [Debnath et al., 2008] (ή το αντίστροφο [De Meo et al., 2007]). Τέλος, μπορεί να κατασκευαστεί ένα γενικό, ενοποιημένο μοντέλο που θα ενσωματώνει τόσο συνεργατικά χαρακτηριστικά όσο και χαρακτηριστικά βασισμένα στο περιεχόμενο [Cantador et al., 2008].

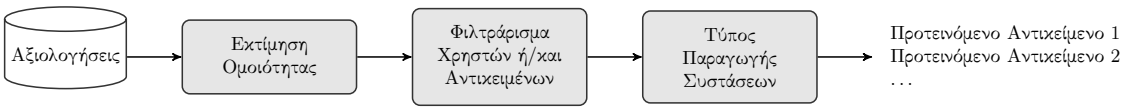
Τα υβριδικά συστήματα συστάσεων είναι περισσότερο πολύμορφα και μπορούν να καλύψουν εκείνες τις πλευρές της διαδικασίας παραγωγής συστάσεων όπου οι απλές τεχνικές περιεχομένου ή οι συνεργατικές μέθοδοι αποτυγχάνουν. Ωστόσο, το κυριότερο πρόβλημα σε αυτή την περίπτωση είναι η πολυπλοκότητα του συστήματος, η οποία επηρεάζει το συνολικό υπολογιστικό χρόνο που είναι απαραίτητος για την παραγωγή συστάσεων. Στη γενική περίπτωση, είναι σημαντικά μεγαλύτερος απ΄ ότι στις απλές προσεγγίσεις.

2.3 Λειτουργία Συστημάτων Συνεργατικής Διήθησης

Όπως αναφέρθηκε και προηγουμένως, η παρούσα διατριβή θα εστιαστεί στα συστήματα συνεργατικής διήθησης, μιας και αποτελούν την πιο ενεργή ερευνητικά περιοχή στον χώρο των συστημάτων συστάσεων. Για το λόγο αυτό, σε αυτή την ενότητα θα γίνει μια αρχική ταξινόμηση των συνεργατικών συστημάτων ανάλογα με τον τρόπο λειτουργία τους.

2.3.1 Μνημονικά Συστήματα

Η *μνημονική συνεργατική διήθηση* (memory-based collaborative filtering), η οποία εναλλακτικά ονομάζεται και *ευρετική συνεργατική διήθηση* (heuristic-based collaborative fil-



Σχήμα 2.2: Μοντέλο λειτουργίας των αλγορίθμων μνημονικής συνεργατικής διήθησης

tering) [Adomavicius et al., 2011], υπολογίζει την χρησιμότητα του κάθε αντικειμένου για ένα χρήστη με την απευθείας επεξεργασία όλων των αξιολογήσεων που εμπεριέχονται στο σύστημα. Ο τρόπος λειτουργίας της συνοψίζεται στο διάγραμμα του Σχήματος 2.2.

Το πρώτο βήμα των αλγορίθμων συνεργατικής διήθησης έχει να κάνει με την *Εκτίμηση της Ομοιότητας* μεταξύ των χρηστών, η οποία μπορεί να υπολογιστεί με διάφορες μεθόδους. Ένας τρόπος είναι να θεωρηθούν οι αξιολογήσεις ως σημεία ενός πολυδιάστατου χώρου και κατόπιν να προσεγγιστεί η ομοιότητα τους ως συνάρτηση της μεταξύ τους απόστασης. Για παράδειγμα, αν I_{au} είναι το σύνολο των αντικειμένων που έχουν αξιολογήσει από κοινού οι χρήστες a και u, τότε ο βαθμός ομοιότητάς τους μπορεί να προκύψει από την απόσταση Minkowski [Amatriain et al., 2011]

$$w_{a,u} \equiv d_{a,u} = [\sum_{i \in I_{au}} |r_{a,i} - r_{u,i}|^k]^{\frac{1}{k}} \tag{2.4}$$

όπου k η τάξη της απόστασης. Για διάφορες τιμές του k, ο γενικός τύπος της εξίσωσης 2.4 λαμβάνει συγκεκριμένες μορφές. Για παράδειγμα, για $k = 1$ είναι γνωστός και ως *απόσταση Manhattan* (ή L_1 νόρμα) ενώ όταν $k = 2$ προκύπτει η *Ευκλείδεια απόσταση (ή L_2 νόρμα)*. Ένας ακόμα τύπος απόστασης που μπορεί να χρησιμοποιηθεί είναι η απόσταση Mahalanobis

$$d_{a,u} = \sqrt{(I_a - I_u)\sigma^{-1}(I_a - I_u)^\top} \tag{2.5}$$

όπου με σ^{-1} συμβολίζεται ο πίνακας συνδιασποράς των αξιολογήσεων που έχουν δώσει οι χρήστες a και u.

Ακόμα, μπορούν να χρησιμοποιηθούν και γνωστοί *συντελεστές συσχέτισης* από την στατιστική [Amatriain et al., 2011]. Οι πιο ευρέως διαδεδομένοι στα συστήματα συστάσεων είναι ο *συντελεστής συνημιτόνου* (Εξίσωση 2.6) και ο *συντελεστής συσχέτισης του Pearson* (Εξίσωση 2.7)

$$w_{a,u} = \frac{\vec{I_a}.\vec{I_u}}{||\vec{I_a}||_2.||\vec{I_u}||_2} = \frac{\sum_{i \in I_{au}} r_{a,i} r_{u,i}}{\sqrt{\sum_{i \in I_{au}} r_{a,i}^2} \sqrt{\sum_{i \in I_{au}} r_{u,i}^2}} \tag{2.6}$$

$$w_{a,u} = \frac{\sigma(I_a, I_w)}{\sigma_{I_a} \times \sigma_{I_w}} = \frac{\sum_{i \in I_{au}} (r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})}{\sqrt{\sum_{i \in I_{au}} (r_{a,i} - \overline{r_a})^2 \sum_{i \in I_{au}} (r_{u,i} - \overline{r_u})^2}} \tag{2.7}$$

όπου $\sigma(I_a, I_w)$ είναι η συνδιασπορά των βαθμολογιών του χρήστη a και u και $\sigma_{I_a}, \sigma_{I_b}$ οι αντίστοιχες τυπικές αποκλίσεις. Σε αρκετές περιπτώσεις χρησιμοποιούνται και παραλλαγές των παραπάνω συντελεστών συσχέτισης.

Πρέπει επίσης να σημειωθεί ότι το στάδιο της εκτίμησης της ομοιότητας δεν αναφέρεται αποκλειστικά και μόνο στην ομοιότητα χρηστών. Μπορεί να αφορά και την ομοιότητα αντικειμένων, μια περίπτωση που είναι γνωστή ως *συνεργατική διήθηση βασισμένη στα αντικείμενα* (item-based collaborative filtering). Σε αυτή την περίπτωση, οι προαναφερόμενοι δείκτες και συντελεστές απόστασης (Εξισώσεις 2.4-2.7) υπολογίζονται πάνω στο σύνολο των κοινών αξιολογήσεων που έχουν λάβει δύο αντικείμενα (αντί του συνόλου των κοινών αξιολογήσεων που έχουν δώσει δύο χρήστες).

Στο αμέσως επόμενο βήμα, ακολουθεί το *Φιλτράρισμα των Χρηστών ή/και των Αντικειμένων*. Πρόκειται για την διαδικασία επιλογής των περισσότερο «όμοιων» χρηστών. Συνηθέστερα χρησιμοποιούνται δύο κριτήρια επιλογής· οι *N πιο όμοιοι χρήστες* (nearest-N) ή οι χρήστες άνω ενός *κατωφλίου* (threshold). Στην πρώτη περίπτωση, οι χρήστες ταξινομούνται σε φθίνουσα σειρά όσον αφορά την τιμή του κριτηρίου ομοιότητας και κατόπιν κρατιούνται οι αξιολογήσεις των N μεγαλύτερων. Στη δεύτερη περίπτωση, κρατιούνται οι αξιολογήσεις όλων των χρηστών όσων η τιμή του κριτηρίου ομοιότητας υπερβαίνει το καθορισμένο από πριν όριο του κατωφλιού. Φυσικά, ότι αναφέρεται για τους χρήστες ισχύει και για τα αντικείμενα (για την περίπτωση συστημάτων συνεργατικής διήθησης βασισμένα στο αντικείμενο).

Στο τελικό στάδιο, ο *Τύπος Παραγωγής Συστάσεων* εκτιμά την ωφέλεια που θα έχει για ένα δεδομένο χρήστη ένα συγκεκριμένο αντικείμενο. Η γενική του μορφή δίνεται στην παρακάτω εξίσωση [Adomavicius and Tuzhilin, 2005]

$$\widehat{r_{a,i}} = \underset{u \in U_{a,i}}{\text{aggr}}\ r_{u,i} \tag{2.8}$$

όπου $\widehat{r_{a,i}}$ η τιμή της ωφέλειας του αντικειμένου i για τον χρήστη a ή εναλλακτικά η πρόβλεψη της αξιολόγησης που θα έκανε ο χρήστης a για το αντικείμενο i, αν είχε έρθει σε γνώση του στο παρελθόν. Το $U_{a,i}$ είναι ένα το σύνολο «όμοιων» με τον a χρηστών (ή αντικειμένων) που έχουν ήδη αξιολογήσει το αντικείμενο i και το οποίο έχει προκύψει με μια από τις δύο μεθόδους φιλτραρίσματος που περιγράφηκαν στην προηγούμενη παράγραφο.

Ο συμβολισμός aggr περιγράφει τον τρόπο με τον οποίο γίνεται η επεξεργασία των αξιολογήσεων των ομοίων χρηστών. Οι πιο συχνά χρησιμοποιούμενοι σχετικοί τύποι είναι ο απλός μέσος όρος (Εξίσωση 2.9), ο ζυγισμένος μέσος όρος (Εξίσωση 2.10) καθώς και ο τύπος του [Resnick et al., 1994] (Εξίσωση 2.11)

$$\widehat{r_{a,i}} = \frac{1}{|U_{a,i}|} \sum_{u \in U_{a,i}} r_{u,i} \tag{2.9}$$

$$\widehat{r_{a,i}} = \frac{\sum_{u \in U_{a,i}} w_{a,u} r_{u,i}}{\sum_{u \in U_{a,i}} w_{a,u}} \tag{2.10}$$

$$\widehat{r_{a,i}} = \overline{r_a} + \frac{\sum_{u \in U_{a,i}} w_{a,u} (r_{u,i} - \overline{r_u})}{\sum_{u \in U_{a,i}} |w_{a,u}|} \tag{2.11}$$

Τα κυριότερα πλεονεκτήματα των μνημονικών συστημάτων σχετίζονται με την ευκολία της αλγοριθμικής τους υλοποίησης και της δυνατότητας άμεσης ανανέωσης των προβλέψεων μόλις προστεθούν καινούργιες (ή τροποποιηθούν ήδη υπάρχουσες) αξιολογήσεις στο σύστημα. Επίσης, οι παραγόμενες προβλέψεις βελτιώνονται συνεχώς με την αύξηση των διαθέσιμων αξιολογήσεων στο σύστημα. Από την άλλη όμως, απαιτούν την επεξεργασία όλων των διαθέσιμων δεδομένων πριν την παραγωγή συστάσεων, πράγμα το οποίο δεν είναι πρακτικά εφικτό, ιδιαίτερα στα πολύ μεγάλα συστήματα συστάσεων. Ένα ακόμα πρόβλημα είναι ότι δεν μπορούν να παράξουν συστάσεις για χρήστες που δεν έχουν κοινές αξιολογήσεις με άλλους χρήστες και αντίστοιχα δεν μπορούν να προτείνουν αντικείμενα τα οποία δεν έχουν ακόμα αξιολογηθεί από κάποιον (Ενότητα 1.1). Τέλος, είναι ευάλωτα στο πρόβλημα της υπερεξειδίκευσης (over-specialization) γιατί δεν έχουν δυνατότητα γενίκευσης των δεδομένων που επεξεργάζονται.

2.3.2 Συστήματα Κατασκευής Μοντέλου

Σε αντίθεση με τους μνημονικούς αλγορίθμους συνεργατικής διήθησης, η *συνεργατική διήθηση κατασκευής μοντέλου* (model-based collaborative filtering) χρησιμοποιεί τις διαθέσιμες αξιολογήσεις για την κατασκευή ενός μοντέλου για τον κάθε χρήστη, το οποίο στη συνέχεια χρησιμοποιείται για την παραγωγή των συστάσεων. Σε γενικές γραμμές, τα μοντέλα

που κατασκευάζονται εντάσσονται σε δύο κατηγορίες, στα μοντέλα *ταξινόμησης* (classification) και στα μοντέλα *παλινδρόμησης* (regression). Τα πρώτα προσπαθούν να κατατάξουν τα αντικείμενα σε κατηγορίες, ανάλογα με το πόσο επιθυμητά ή όχι είναι από τον χρήστη. Η ταξινόμηση μπορεί να είναι *επιβλεπόμενη* (supervised classification), κατά την οποία παρέχονται στο σύστημα το πλήθος των επιθυμητών κατηγοριών καθώς και κάποια δεδομένα εκπαίδευσης. Το σύστημα στη συνέχεια, από αυτές τις πληροφορίες προσπαθεί να «μάθει» να ταξινομεί καινούργιους χρήστες και αντικείμενα στις σωστές κατηγορίες. Υπάρχει επίσης και η *μη-επιβλεπόμενη ταξινόμηση* (unsupervised classification) ή αλλιώς *συσταδοποίηση* (clustering) κατά την οποία το σύστημα δεν χρησιμοποιεί καμία εξωτερική πληροφορία αλλά αντίθετα προσπαθεί μόνο του να εξάγει τις κατηγορίες που ανήκουν οι χρήστες και τα αντικείμενα. Τα μοντέλα παλινδρόμησης, από την άλλη, προσπαθούν να προσεγγίσουν με διάφορες μεθοδολογίες, τις πιθανές αξιολογήσεις που θα έδινε ένας χρήστης σε αντικείμενα τα οποία δεν έχει προσπελάσει ακόμα.

Στη σχετική βιβλιογραφία υπάρχουν ήδη πολλές τεχνικές κατασκευής μοντέλων για τα συστήματα συνεργατικής διήθησης, οι κυριότερες εκ των οποίων παρουσιάζονται παρακάτω

Πλησιέστεροι Γείτονες

Η μέθοδος των *πλησιέστερων γειτόνων* (nearest neighbors - NN) θεωρεί ότι οι αλληλεπιδράσεις χρηστών-αντικειμένων μπορούν να μοντελοποιηθούν ως διανύσματα σε έναν πολυδιάστατο *χώρο χαρακτηριστικών* (feature space), στον οποίο η αξιολόγηση καθορίζει την κατηγορία (ή διαφορετικά, την ετικέτα - label) στην οποία ανήκουν. Για κάθε καινούργιο διάνυσμα που εμφανίζεται στο χώρο, ο ταξινομητής προσπαθεί να βρει τα k πλησιέστερα σημεία (κοντινότεροι γείτονες) για τα οποία υπάρχουν ήδη ετικέτες. Στη συνέχεια τοποθετεί μια ετικέτα στο νέο σημείο αντίστοιχη με τις ετικέτες των κοντινότερων γειτόνων του. Η φιλοσοφία λειτουργίας του ταξινομητή είναι ότι αν ένα καινούργιο σημείο βρεθεί σε μια περιοχή όπου είναι κυρίαρχα τα σημεία μιας ετικέτας, τότε αυτό οφείλεται στο γεγονός πως και το καινούργιο σημείο πιθανότατα ανήκει στην εν λόγω κατηγορία.

Η πιο σημαντική παράμετρος της μεθόδου είναι ο ορισμός της γειτονίας, δηλαδή του πλήθους των γειτόνων που πρέπει να ληφθούν υπόψη στη φάση της ταξινόμησης. Αν η γειτονιά είναι πολύ μεγάλη, τότε ενδέχεται να συμπεριλαμβάνει σημεία από πολλές κατηγορίες· αν πάλι είναι πολύ μικρή τότε είναι λιγότερο ευαίσθητη σε δεδομένα θορύβου. Υπάρχουν συστήματα που χρησιμοποιούν στατικούς ορισμούς για τη γειτονιά (λ.χ οι δύο πιο κοντινοί γείτονες) ενώ άλλα την μεταβάλλουν δυναμικά, ανάλογα με το εκάστοτε κάθε φορά δείγμα. Η μέθοδος των πλησιέστερων γειτόνων αποτελεί μια από τις πιο απλές τεχνικές *μηχανικής μάθησης* (machine learning). Μπορεί να προσαρμόζεται πολύ γρήγορα σε αλλαγές των δεδομένων, για παράδειγμα στην προσθήκη νέων βαθμολογιών, αλλά έχει το μειονέκτημα ότι θα πρέπει κάθε φορά να επανυπολογίζονται εκ νέου οι ομοιότητες μεταξύ των σημείων.

Δέντρα Απόφασης

Τα *δέντρα απόφασης* (decision trees) εντάσσονται και αυτά με τη σειρά τους στην κατηγορία των ταξινομητών. Χρησιμοποιούν μια δενδρική δομή για να αποφασίσουν σε ποια κατηγορία πρέπει να ενταχθεί μια νέα παρατήρηση (λ.χ. ένα καινούργιο αντικείμενο). Οι κόμβοι του δέντρου μπορεί να είναι είτε κόμβοι απόφασης (στους οποίους ελέγχεται μια συγκεκριμένη ιδιότητα του αντικειμένου για να καθοριστεί ποιο κλαδί του δέντρου θα επιλεχθεί στη συνέχεια) ή κόμβοι-φύλλα (που χαρακτηρίζουν την κατηγορία στην οποία ανήκει το αντικείμενο). Τα δέντρα απόφασης κατασκευάζονται από τα

διαθέσιμα δεδομένα εκπαίδευσης με τη χρήση αλγορίθμων όπως ο ID3 και ο C4.5, οι οποίοι προσπαθούν να μεγιστοποιήσουν το πληροφοριακό κέρδος σε κάθε βήμα.

Τα κυριότερα πλεονεκτήματα των δέντρων απόφασης είναι ότι κατασκευάζονται σχετικά εύκολα ενώ μπορούν να ταξινομήσουν νέα δείγματα πάρα πολύ γρήγορα. Επίσης μπορούν να χρησιμοποιηθούν για την κατασκευή ενός συνόλου κανόνων που είναι δυνατό να ερμηνευτούν εύκολα από τους κατασκευαστές τους αλλά και από τους χρήστες τους. Παρόλα αυτά όμως, γίνονται λιγότερο πρακτικοί σε μεγάλα συστήματα συστάσεων, όπου η ταξινόμηση των νέων αντικειμένων ενδέχεται να εξαρτάται από πάρα πολλές παραμέτρους.

Μπεϋζιανοί Ταξινομητές

Οι *μπεϋζιανοί ταξινομητές* (bayesian classifiers) χρησιμοποιούν τις πιθανότητες, και πιο συγκεκριμένα το μπεϋζιανό θεώρημα για να μοντελοποιήσουν την αβεβαιότητα για τις σχέσεις οι οποίες εξάγονται από τα δεδομένα. Η αρχή λειτουργίας τους ορίζει ότι η *εκ των υστέρων πιθανότητα* (posterior probability) σχηματισμού ενός συγκεκριμένου μοντέλου από τα δοσμένα δεδομένα είναι ανάλογη του γινομένου της *πιθανοφάνειας* (likelihood) επί της *εκ των προτέρων* πιθανότητας (prior probability). Ο παράγοντας της πιθανοφάνειας αναπαριστά την επίδραση των δεδομένων στο μοντέλο ενώ η εκ των προτέρων πιθανότητα την εμπιστοσύνη στο μοντέλο πριν την παρατήρηση των δεδομένων.

Τα κυριότερα πλεονεκτήματα των μπεϋζιανών δικτύων είναι η ανθεκτικότητά τους σε μεμονωμένα σημεία θορύβου, στον χειρισμό μη-σχετικών ιδιοτήτων καθώς και στον χειρισμό ελλιπών τιμών (αφού μπορούν να παραβλέπουν συγκεκριμένα δείγματα κατά τη διάρκεια της εκτίμησης των πιθανοτήτων). Από την άλλη όμως, η υπόθεση ανεξαρτησίας μεταξύ των παραμέτρων του μοντέλου (που επιβάλλεται από την μπεϋζιανή συνεπαγωγή) μπορεί να μην ισχύει σε όλες τις περιπτώσεις. Τότε μπορούν να βρουν εφαρμογή τα *μπεϋζιανά δίκτυα* (bayesian networks), τα οποία χρησιμοποιούν έναν ακυκλικό γράφο για την κωδικοποίηση των εξαρτήσεων μεταξύ των παραμέτρων του μοντέλου, μαζί με έναν πίνακα πιθανοτήτων που σχετίζει κάθε κόμβο με τους άμεσους προγόνους του.

Τεχνητά Νευρωνικά Δίκτυα

Τα *τεχνητά νευρωνικά δίκτυα* (artificial neural networks - ANNs) αποτελούνται από μια αρχιτεκτονική διασυνδεδεμένων με ακμές (με βάρη) κόμβων που προσομοιάζει τη δομή και τη λειτουργία του βιολογικού εγκεφάλου. Οι κόμβοι στα ANN καλούνται *νευρώνες* (neurons) κατ΄ αντιστοιχία με τους βιολογικούς νευρώνες. Αυτές οι απλές λειτουργικές μονάδες σχηματίζουν δίκτυα τα οποία έχουν την ικανότητα να προσεγγίζουν την λειτουργία ενός μοντέλου, αφού προηγουμένως τροφοδοτηθούν με κάποια δεδομένα εκπαίδευσης. Η πιο απλή περίπτωση ANN είναι το perceptron, το οποίο αποτελείται από έναν μόνο νευρώνα που αθροίζει τις εισόδους του και υπολογίζει την έξοδό του στη βάση μιας *συνάρτησης ενεργοποίησης* νευρώνα (activation function) [Haykin, 2008]. Το απλό percpetron μπορεί, μέσω της εκπαίδευσής του, να διακρίνει γραμμικά διαχωρίσιμες κατηγορίες δεδομένων.

Η δύναμη αναπαράστασης των νευρωνικών δικτύων έγκειται στο γεγονός ότι η αρχιτεκτονική τους μπορεί να επεκταθεί σε παραπάνω του ενός επίπεδα. Ανάλογα με τον τύπο τους, τα επίπεδα χωρίζονται σε τρεις κατηγορίες: στο επίπεδο εισόδου (λαμβάνει τα δεδομένα που παρέχονται στο δίκτυο), στα κρυφά επίπεδα (πραγματοποιούν την επεξεργασία) και τέλος στο επίπεδο εξόδου (συνδυάζει τα αποτελέσματα της επεξεργασίας

των κρυφών επιπέδων για να παραχθεί η ολική έξοδος του δικτύου). Αυτή η αρχιτεκτονική, σε συνδυασμό με τις κατάλληλες συναρτήσεις ενεργοποίησης των νευρώνων, επιτρέπει στα ANN να προσεγγίζουν οποιαδήποτε συνάρτηση και κατά συνέπεια να μπορούν να διαχωρίζουν δεδομένα τα οποία ανήκουν και σε μη-γραμμικά διαχωρίσιμες κατηγορίες. Ένα συνεργατικό σύστημα συστάσεων το οποίο βασίζεται σε νευρωνικά δίκτυα θα παρουσιαστεί στο Κεφάλαιο 4.

Μηχανές Διανυσμάτων Υποστήριξης

Ο στόχος των *μηχανών διανυσμάτων υποστήριξης* (support vector machines - SVM) είναι η εύρεση ενός γραμμικού υπερεπιπέδου (σύνορο απόφασης - decision boundary) το οποίο διαχωρίζει τις κατηγορίες των δεδομένων με τέτοιον τρόπο ώστε το μεταξύ τους περιθώριο να είναι το μέγιστο δυνατό. Ο λόγος που αναζητείται το μέγιστο δυνατό περιθώριο είναι για την ελαχιστοποίηση της πιθανότητας εσφαλμένης ταξινόμησης μελλοντικών δειγμάτων. Στην περίπτωση που τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, το SVM μπορεί να επεκταθεί με την *χρήση μεταβλητών χαλαρότητας* (slack variables). Από την άλλη, αν το σύνορο απόφασης δεν είναι γραμμικό, τότε τα δεδομένα προβάλλονται σε έναν χώρο μεγαλύτερων διαστάσεων, στον οποίο πιθανολογείται ότι θα μπορέσουν να διαχωριστούν γραμμικά. Ο μετασχηματισμός αυτός πραγματοποιείται με την βοήθεια των *συναρτήσεων πυρήνα* (kernel functions), με τις πιο ευρέως χρησιμοποιούμενες να είναι οι *ακτινικές συναρτήσεις βάσης* (radial basis functions). Σε γενικές γραμμές, η χρήση των SVM στα συστήματα συστάσεων οδηγεί στην κατασκευή μοντέλων που παρουσιάζουν μεγάλη ακρίβεια ταξινόμησης, ανθεκτικότητα στον θόρυβο και τέλος δεν εμφανίζουν φαινόμενα υπερπροσαρμογής (overfitting). Από την άλλη όμως, η εύρεση του καταλλήλου συνόρου απόφασης είναι μια υπολογιστικά δαπανηρή διαδικασία, ιδιαίτερα όσο ο αριθμός των υπό ταξινόμηση δειγμάτων μεγαλώνει.

Μοντέλα Παραγοντοποίησης Πινάκων

Τα *μοντέλα παραγοντοποίησης πινάκων* (matrix factorization models) έχουν ως στόχο την αναζήτηση εκείνων των *γνωρισμάτων* (features) που περιγράφουν κατά τον καλύτερο δυνατό τρόπο τα χαρακτηριστικά των αξιολογούμενων αντικειμένων και τις προτιμήσεις των χρηστών. Τα εν λόγω γνωρίσματα αφενός εμπεριέχονται στον πίνακα αξιολογήσεων, αφετέρου μπορούν να περιγραφούν με πολύ λιγότερες παραμέτρους. Έτσι, οι τεχνικές παραγοντοποίησης πινάκων προσπαθούν να επανυπολογίσουν τον πίνακα αξιολογήσεων ως το γινόμενο δύο (ή και περισσότερων) πινάκων, οι οποίοι έχουν πολύ μικρότερες διαστάσεις σε σχέση με τον αρχικό και στους οποίους είναι περισσότερο εμφανή τα ζητούμενα χαρακτηριστικά γνωρίσματα. Η συγκεκριμένη διαδικασία είναι επαναληπτική και τερματίζεται όταν ικανοποιηθούν συγκεκριμένα κάθε φορά κριτήρια (.χ μεγιστοποιηθεί το πληροφοριακό κέρδος). Τα συστήματα συστάσεων τα οποία βασίζονται σε τεχνικές παραγοντοποίησης πινάκων είναι εξαιρετικά ακριβή (μάλιστα η ομάδα που κέρδισε τον γνωστό διαγωνισμό της εταιρίας Netflix το 2009 χρησιμοποίησε ένα σχετικό σύστημα [Koren, 2009]) όμως έχουν το μειονέκτημα του μεγάλου υπολογιστικού κόστους της διαδικασίας καθώς επίσης και της ανελαστικότητας στον χειρισμό νέων βαθμολογιών· στις περισσότερες φορές οι παραγοντοποιήσεις πρέπει να επαναληφθούν από την αρχή. Η παραγοντοποίηση πινάκων χρησιμοποιήθηκε στην παρούσα διατριβή (Κεφάλαιο 8) αλλά σε ένα εντελώς διαφορετικό πλαίσιο.

Συμπερασματικά, τα πλεονεκτήματα των συστημάτων κατασκευής μοντέλων μπορούν να συνοψιστούν στα εξής: στην επεκτασιμότητά τους, στην ταχύτητα παραγωγής των προβλέψε-

ων και στην αποφυγή της υπερπροσαρμογής. Τα μοντέλα που κατασκευάζονται από τους αντίστοιχους αλγορίθμους είναι τις περισσότερες φορές πολύ μικρότερα σε μέγεθος από τα δεδομένα τα οποία χρησιμοποιήθηκαν για την κατασκευή τους. Συνεπώς, είναι αποδοτικά ακόμα και για μεγάλες συλλογές δεδομένων. Επίσης, τα μοντέλα μπορούν να παράξουν συστάσεις πάρα πολύ γρήγορα (αφού προηγουμένως έχουν κατασκευαστεί), σε αντίθεση με τα μνημονικά συστήματα που χρειάζεται να επεξεργαστούν όλα τα δεδομένα που υπάρχουν στο σύστημα πριν δώσουν αποτέλεσμα. Τέλος, τα μοντέλα έχουν μεγαλύτερη δύναμη αναπαράστασης και συνεπώς πιο δύσκολα καταλήγουν σε υπερπροσαρμογή των δεδομένων τους.

Από την άλλη όμως, η χρήση των μοντέλων στα συνεργατικά συστήματα συστάσεων έχει και κάποια μειονεκτήματα. Καταρχήν, η κατασκευή τους είναι στις περισσότερες περιπτώσεις μια υπολογιστικά δαπανηρή διαδικασία, όπως και η ενημέρωσή τους με νέα δεδομένα, οπότε είναι πιο ανελαστικά από τα μνημονικά συστήματα όσον αφορά αυτές τις δύο λειτουργίες. Επιπλέον, το γεγονός ότι δεν χρησιμοποιούνται όλα τα δεδομένα από το μοντέλο για την παραγωγή των συστάσεων μπορεί να οδηγήσει, σε ορισμένες περιπτώσεις, σε πτώση της συνολικής απόδοσης του συστήματος.

□

Κεφάλαιο 3

Αξιολόγηση των Συστημάτων Συστάσεων

Η αξιολόγηση των συστημάτων συστάσεων αποτελεί μια από τις πιο σημαντικές πλευρές της πρακτικής τους υλοποίησης. Είναι απαραίτητη σε διάφορα στάδια του κύκλου ζωής τους για πολλούς λόγους. Για παράδειγμα, στη φάση του σχεδιασμού χρειάζεται να τεκμηριωθεί η ορθότητα της επιλογής της κατάλληλης μεθοδολογίας παραγωγής συστάσεων. Στη φάση της χρήσης πρέπει να βεβαιωθεί το κατά πόσο η λειτουργία του συστήματος ανταποκρίνεται στις απαιτήσεις των χρηστών του, όχι μόνο όσον αφορά την χρησιμότητα των παραγόμενων προτάσεων αλλά και από άλλες πλευρές, όπως είναι η ταχύτητα, η πρωτοτυπία, η συνολική ευχρηστία κ.λ.π.

Ο σχεδιασμός των κατάλληλων μεθοδολογιών αξιολόγησης των συστημάτων συστάσεων αποτελεί ένα αντικείμενο το οποίο έχει ερευνηθεί εκτενώς στη σχετική βιβλιογραφία [Herlocker et al., 2004]. Πιο συγκεκριμένα, οι [Shani and Gunawardana, 2011] παρουσιάζουν τρεις διαφορετικές μεθοδολογίες αξιολόγησης: την *offline αξιολόγηση*, τις *μελέτες χρηστών* και την *online αξιολόγηση*. Στην πρώτη περίπτωση, η αξιολόγηση πραγματοποιείται σε ήδη συλλεγμένα δεδομένα χρηστών, χωρίς να είναι απαραίτητη η ταυτόχρονη λειτουργία του συστήματος. Τα δεδομένα προέρχονται είτε από δημόσια αποθετήρια αναφοράς ή μπορεί να έχουν συλλεχθεί και ιδιωτικά από το σχεδιαστή του RS μέσω, λ.χ., πρότυπων υλοποιήσεων ανοιχτών για το κοινό για συγκεκριμένο χρονικό διάστημα. Το βασικό πλεονέκτημα αυτών των μεθόδων είναι η μη-παρεμβατικότητα τους· δεν απαιτούν κανενός είδους αλληλεπίδραση με τους χρήστες ενώ ταυτόχρονα επιτρέπουν την ευρεία σύγκριση πολλών αλγορίθμων με πολύ μικρό κόστος. Το μειονέκτημά τους, όμως, είναι ότι περιορίζονται σε μια στενή γκάμα χαρακτηριστικών των υπό εξέταση συστημάτων και πιο συγκεκριμένα σε αυτά που έχουν να κάνουν με τις δυνατότητες πρόβλεψης των αλγορίθμων. Για τον λόγο αυτό, βρίσκουν κυρίως χρήση στα αρχικά στάδια σχεδιασμού των συστημάτων συστάσεων, όταν οι δημιουργοί τους θέλουν να περιορίσουν τους υποψήφιους προς υλοποίηση αλγορίθμους σε ένα μικρό σύνολο, το οποίο στη συνέχεια θα το αξιολογήσουν με τις άλλες μεθόδους.

Οι μελέτες χρηστών πραγματοποιούνται με την ενεργή συμμετοχή ενός συνόλου χρηστών, από τους οποίους ζητείται να ολοκληρώσουν διάφορες εργασίες που απαιτούν αλληλεπίδραση με το RS. Κατά τη διάρκεια της εκτέλεσης των εργασιών παρατηρείται και καταγράφεται η συμπεριφορά των χρηστών, ενώ ταυτόχρονα εξάγονται και ποιοτικά στοιχεία, όπως λ.χ., το ποσοστό των εργασιών που πραγματοποιήθηκαν και ο χρόνος που απαιτήθηκε. Επίσης μπορούν να αποτιμηθούν και χαρακτηριστικά τα οποία δεν είναι άμεσα παρατηρήσιμα από τα δεδομένα, όπως το κατά πόσον οι χρήστες βρήκαν ευχάριστη τη διεπαφή ή αν θεώρησαν τις εργασίες εύκολες στην πραγματοποίηση τους. Το κυριότερο πλεονέκτημά αυτής της μεθόδου έγκειται στο εύρος των χαρακτηριστικών που μπορούν να παρατηρηθούν· από την άλλη όμως παρουσιάζουν υψηλό κόστος όσον αφορά τη σχεδίαση και την πραγματοποίηση της πειραματικής διαδικασίας, τόσο από την άποψη του απαιτούμενου χρόνου όσο και ενδεχομένως από

οικονομικής πλευράς.

Τέλος, η online αξιολόγηση πραγματοποιείται από τους χρήστες του RS, αφού έχει ξεκινήσει η λειτουργία του. Σε ένα τέτοιο σενάριο, το σύστημα θα μπορούσε να παράξει σύνολα προτάσεων με την χρήση διαφορετικών αλγορίθμων και κατόπιν να ζητήσει από τους χρήστες του να αξιολογήσουν ποιος από αυτούς τους βοήθησε περισσότερο στην εξερεύνηση των διαθέσιμων αντικειμένων ή ήταν περισσότερο κοντά στις προσδοκίες τους. Τα πλεονεκτήματα των online μεθόδων είναι ότι επιτρέπουν την άμεση αλληλεπίδραση με τους χρήστες καθώς και την εκτίμηση του πως διαφορετικοί παράγοντες λειτουργίας του συστήματος επηρεάζουν τους χρήστες. Από την άλλη όμως, η online αξιολόγηση έχει και αυτή υψηλό κόστος επεξεργασίας, μιας και ο όγκος της παραγόμενης πληροφορίας είναι πολύ μεγάλος και επιπρόσθετα πρέπει να πραγματοποιείται σε πραγματικό χρόνο.

Στην παρούσα διατριβή πραγματοποιήθηκαν offline πειράματα σε δημόσια διαθέσιμες συλλογές δεδομένων. Ο κυριότερος λόγος που επέβαλλε αυτή την επιλογή είναι τα χρονικά περιθώρια μιας και θα χρειαζόταν ένα πολύ μεγάλο χρονικό διάστημα μέχρις ότου ένα σύστημα συστάσεων, το οποίο έχει υλοποιηθεί από το αρχή, γίνει τόσο γνωστό ώστε να συγκεντρώσει αριθμό χρηστών και βαθμολογιών σε συγκρίσιμα μεγέθη με αυτά που υπάρχουν στις δημόσιες συλλογές δεδομένων. Για το λόγο αυτό, οι μέθοδοι μέτρησης της απόδοσης που περιγράφονται στη συνέχεια του κεφαλαίου βρίσκουν κυρίως εφαρμογή σε offline πειραματικές διαδικασίες και εξετάζουν, από διαφορετική κάθε φορά πλευρά, την *δύναμη πρόβλεψης* (prediction power) των υπό εξέταση αλγορίθμων.

3.1 Πυκνότητα και Αραιότητα των Αξιολογήσεων

Αν και δεν αποτελούν μετρικές αξιολόγησης των RS καθ' εαυτές, η αραιότητα και η πυκνότητα των αξιολογήσεων ωστόσο χρησιμοποιούνται στη φάση της σύγκρισης της απόδοσης των αλγορίθμων συστάσεων σε διαφορετικά συστήματα. Η *πυκνότητα των αξιολογήσεων* (density of the ratings) μετράει το πόσο πλήρες αξιολογήσεων είναι ένα RS. Ορίζεται ως ο λόγος του πλήθους των καταχωρημένων αξιολογήσεων στο σύστημα ως προς την ιδεατή κατάσταση, όπου όλοι οι χρήστες έχουν αξιολογήσει όλα τα αντικείμενα. Αν R το σύνολο των αξιολογήσεων, U το σύνολο των χρηστών και I το σύνολο των αντικειμένων, τότε η πυκνότητα ορίζεται ως ο παρακάτω λόγος

$$density = \frac{|R|}{|U| \times |I|} \quad (\%) \tag{3.1}$$

Συμπληρωματικό μέγεθος αποτελεί η *αραιότητα των αξιολογήσεων* (sparsity of the ratings), που ορίζεται ως

$$sparsity = 1 - density \quad (\%) \tag{3.2}$$

3.2 Μέτρηση της Απόδοσης

3.2.1 Μετρικές Ακρίβειας της Πρόβλεψης

Οι μετρικές *ακρίβειας της πρόβλεψης* (prediction accuracy) έχουν τις ρίζες τους στη στατιστική ανάλυση και αποτελούν τις πιο δημοφιλείς μεθόδους μέτρησης της απόδοσης των συστημάτων συστάσεων. Έχουν βρει ευρεία απήχηση γιατί επί της ουσίας η πλειοψηφία των συστημάτων χρησιμοποιούν ως βάση τους μια *μηχανή προβλέψεων* (prediction engine) [Shani and Gunawardana, 2011] για την εκτίμηση της άποψης του χρήστη για συγκεκριμένα αντικείμενα. Η βασική υπόθεση εργασίας στη συγκεκριμένη περίπτωση είναι ότι όσο πιο ακριβές

είναι το συστήματα στις προβλέψεις του τόσο περισσότερο θα προτιμηθεί από τους χρήστες του. Για τον λόγο αυτό, πολλοί ερευνητές προσπαθούν να κατασκευάσουν αλγορίθμους που να παράγουν όσο το δυνατόν ακριβέστερες προτάσεις. Επειδή η ακρίβεια της πρόβλεψης είναι ανεξάρτητη από τη διεπαφή που χρησιμοποιεί το σύστημα συστάσεων, μπορεί να μετρηθεί σε μια offline πειραματική διαδικασία.

Ρίζα του Μέσου Τετραγωνικού Σφάλματος

Η *ρίζα του μέσου τετραγωνικού σφάλματος* (root mean square error - RMSE) αποτελεί την πλέον δημοφιλή μέθοδο μέτρησης της απόδοσης, ειδικά μετά τον ορισμό του ως τη βασική μέθοδο μέτρησης της απόδοσης στο διαγωνισμό της Netflix [net]. Αν υποτεθεί ότι ο αλγόριθμος συστάσεων έχει παράξει ένα σύνολο N προβλέψεων για ζεύγη χρηστών-αντικειμένων $(u,i) \in S$ όπου $\widehat{r_{u,i}}$ είναι η τιμή της πρόβλεψης και $r_{u,i}$ είναι η πραγματική αξιολόγηση, τότε το RMSE υπολογίζεται ως εξής

$$RMSE = \sqrt{\frac{\sum\limits_{(u,i)\in S} (\widehat{r_{u,i}} - r_{u,i})^2}{N}} \tag{3.3}$$

Επειδή στο RMSE η διαφορά της πρόβλεψης με την πραγματική τιμή υψώνεται στο τετράγωνο, δίνεται κατ' αυτόν τον τρόπο περισσότερο βάρος στα μεγαλύτερα σφάλματα απ' ότι στα μικρότερα.

Μέσο Απόλυτο Σφάλμα

Το *μέσο απόλυτο σφάλμα* (mean absolute error - MAE) ανήκει και αυτό στην κατηγορία των δημοφιλών μεθόδων μέτρησης της απόδοσης. Χρησιμοποιείται για την μέτρηση της μέσης απόλυτης απόκλισης μεταξύ της πραγματικής και της εκτιμώμενης αξιολόγησης [Herlocker et al., 2004]. Υπολογίζεται ως

$$MAE = \frac{\sum\limits_{(u,i)\in S} |\widehat{r_{u,i}} - r_{u,i}|}{N} \tag{3.4}$$

με τα χρησιμοποιούμενα σύμβολα να έχουν την ίδια ερμηνεία με την προηγούμενη περίπτωση. Σε αντίθεση με το RMSE, το MAE δίνει σε όλα τα σφάλματα το ίδιο ακριβώς βάρος.

Κανονικοποιημένο Μέσο Απόλυτο Σφάλμα

Το *κανονικοποιημένο μέσο απόλυτο σφάλμα* (Normalized MAE - NMAE) ορίζεται στη βιβλιογραφία με δύο τρόπους. Ο πρώτος παρουσιάζεται στην παρακάτω εξίσωση

$$NMAE = \frac{MAE}{r_{max} - r_{min}} \tag{3.5}$$

όπου r_{max} και r_{min} η μέγιστη και η ελάχιστη δυνατή τιμή της αξιολογικής κλίμακας αντίστοιχα. Η συγκεκριμένη μέθοδος μέτρησης της απόδοσης προτάθηκε αρχικά από τον [Goldberg et al., 2001] με σκοπό την σύγκριση αποτελεσμάτων μεταξύ διαφορετικών συλλογών δεδομένων, αλλά σύμφωνα με τον [Herlocker et al., 2004] η χρησιμότητα της

δεν έχει επαληθευτεί ακόμα.

Ο δεύτερος ορισμός, ο οποίος προτάθηκε από τον [Marlin, 2004] είναι ο ακόλουθος

$$NMAE = \frac{MAE}{E[MAE]} \tag{3.6}$$

όπου $E[MAE]$ η εκτιμώμενη τιμή για το MAE όταν οι συστάσεις παράγονται τυχαία, υπακούοντας στην ομοιόμορφη κατανομή. Συνεπώς όταν ένας αλγόριθμος συστάσεων παρουσιάζει τιμή για το NMAE μικρότερη της μονάδας, τότε είναι πιο ακριβής από την μια (ομοιόμορφα) τυχαία παραγωγή συστάσεων, ενώ το αντίθετο ακριβώς συμβαίνει όταν η τιμή του NMAE υπερβεί την μονάδα. Και ο δεύτερος αυτός ορισμός του NMAE έχει ως στόχο την σύγκριση αποτελεσμάτων μεταξύ διαφορετικών συλλογών δεδομένων για συστήματα συστάσεων.

Μέσο Απόλυτο Σφάλμα ανά Χρήστη

Όπως έχει ήδη αναφερθεί, το MAE δίνει την ίδια ακριβώς βαρύτητα σε κάθε μεμονωμένο σφάλμα. Έτσι όμως, η τελική τιμή του σφάλματος διαμορφώνεται κυρίως από τους χρήστες που έχουν κάνει πολλές αξιολογήσεις σε αντίθεση με αυτούς που έχουν δώσει λίγες [Massa and Avesani, 2009, 2004], οι οποίοι όμως αποτελούν και την πλειοψηφία των χρηστών (Ενότητα 1.1). Ως λύση, προτείνεται η κανονικοποίηση του MAE ανά χρήστη, μέσω του *μέσου απόλυτου σφάλματος ανά χρήστη* (mean absolute user error - MAUE), το οποίο ορίζεται ως

$$MAUE = \frac{1}{N} \sum_{(u,i) \in S} \left(\frac{\sum\limits_{i \in S_u} |\widehat{r_{u,i}} - r_{u,i}|}{|S_u|} \right) \tag{3.7}$$

Και σε αυτήν την περίπτωση $N = |S|$ είναι το σύνολο των προβλέψεων για ζεύγη χρηστών-αντικειμένων $(u, i) \in S$, S_u είναι το σύνολο των προβλέψεων για τον u-οστό χρήστη, $\widehat{r_{u,i}}$ είναι η τιμή της πρόβλεψης και $r_{u,i}$ είναι η πραγματική αξιολόγηση. Με αυτόν τον τρόπο, όλοι οι χρήστες συμμετέχουν, κατά τον ίδιο βαθμό, στη διαμόρφωση της τελικής τιμής του σφάλματος

Μέσο Απόλυτο Σφάλμα ως προς την Αραιότητα των Αξιολογήσεων

Οι γραφικές παραστάσεις του *μέσου απόλυτου σφάλματος ως προς την αραιότητα του πίνακα των βαθμολογιών* (MAE vs Sparsity) δείχνουν πως εξελίσσεται η τιμή του MAE για διάφορα επίπεδα αραιότητας του πίνακα βαθμολογιών. Η βασική μεθοδολογία για την δημιουργία αυτών των γραφικών παραστάσεων είναι η αφαίρεση με τυχαίο τρόπο ενός ολοένα αυξανόμενου αριθμού βαθμολογιών έτσι ώστε η αραιότητα (ή *επίπεδο αραιότητας* - sparsity level) του πίνακα βαθμολογιών να αυξάνει σε προκαθορισμένες τιμές - βήματα (πχ 0.94, 0.95, 0.96 κ.ο.κ) [Ghazanfar and Prugel-Bennett, 2010; Bellogin, 2009; Hwang and Fong, 2011]. Για κάθε μία από αυτές τις τιμές του επιπέδου αραιότητας μετρώνται οι τιμές των προς μελέτη μεγεθών (π.χ. MAE, MAUE) και προκύπτουν τελικά γραφικές παραστάσεις που απεικονίζουν τη συμπεριφορά του εκάστοτε συστήματος σε συνθήκες αυξανόμενης αραιότητας.

Ένας εναλλακτικός τρόπος δημιουργίας επιπέδων αραιότητας είναι το να θεωρηθούν τα συστήματα συστάσεων ως δυναμικά συστήματα, που μεταβάλλονται με την πάροδο του χρόνου. Σε αυτού του είδους την μοντελοποίηση ο πίνακας βαθμολογιών αρχικά είναι πολύ αραιός, οπότε και το μέσο σφάλμα του αλγορίθμου σύστασης μεγάλο. Με την προσθήκη όμως χρηστών, αντικειμένων και βαθμολογιών, η πυκνότητά του αυξάνει

με αποτέλεσμα να αλλάζουν οι τιμές των μετρούμενων μεγεθών. Η μεθοδολογία αυτή ακολουθήθηκε από τους [Pitsilis and Knapskog, 2009], όπου η συλλογή δεδομένων `MovieLens 10M` χωρίζεται σε πέντε κομμάτια, ανάλογα με τη χρονική στιγμή υποβολής της εκάστοτε βαθμολογίας. Σε αυτές τις πέντε ξεχωριστές απεικονίσεις (που προφανώς έχουν διαφορετικό μέγεθος και πυκνότητα) υπολογίζεται κάθε φορά η μέτρηση της απόδοσης.

Ένας ακόμα τρόπος δημιουργίας των εν λόγω γραφικών παραστάσεων είναι με τη χρησιμοποίηση διαφορετικών όψεων των δεδομένων από την αρχική συλλογή δεδομένων. Από την εξέταση της απόδοσης επάνω σε αυτές (συνηθέστερα του MAE) παράγονται οι αντίστοιχες γραφικές παραστάσεις. Για παράδειγμα, οι [Piccart et al., 2010] υπολογίζουν την απόδοση σε σχέση με την αραιότητα, ομαδοποιώντας τους χρήστες ανά πλήθος βαθμολογιών (π.χ όσοι έχουν n βαθμολογίες).

Με την βασική μεθοδολογία είναι δυνατή η παραγωγή συλλογών δεδομένων με καθορισμένα επίπεδα αραιότητας. Στις δύο τελευταίες περιπτώσεις τα επίπεδα αραιότητας είναι εκείνα που προκύπτουν από τις εκάστοτε χρονικές στιγμές που έχουν ληφθεί οι βαθμολογίες και οι όψεις.

3.2.2 Μετρικές Ακρίβειας της Ταξινόμησης

Στην προηγούμενη ενότητα αναφέρθηκε πως οι περισσότεροι αλγόριθμοι παραγωγής συστάσεων επί της ουσίας λειτουργούν ως μηχανές πρόβλεψης της αξιολόγησης που θα έκανε ένας δεδομένος χρήστης σε ένα συγκεκριμένο αντικείμενο. Σε ένα γενικότερο επίπεδο όμως, μπορεί να υποστηριχθεί ότι αυτό που κυρίως ενδιαφέρει τους χρήστες είναι να τους προτείνονται χρήσιμα αντικείμενα και όχι τόσο η προσέγγιση της ακριβούς αξιολόγησης που θα έκαναν σε συγκεκριμένα αντικείμενα. Οι μετρικές *ακρίβειας της ταξινόμησης* (classification accuracy metrics), που προέρχονται την θεωρία *ανάκτησης πληροφορίας* (information retrieval) [Billsus and Pazzani, 1998], χρησιμοποιούνται για αυτόν ακριβώς τον σκοπό· για την εξεύρεση του βαθμού στον οποίο οι προτάσεις του RS ανταποκρίνονται στις απαιτήσεις των χρηστών.

Για τις μετρικές ακρίβειας της ταξινόμησης, η παραγόμενη σύσταση ενός αντικείμενου προς ένα χρήστη εμπίπτει σε μια από τις τέσσερις κατηγορίες του Πίνακα 3.1. Δηλαδή μπορεί να προταθεί ένα αντικείμενο το οποίο ο χρήστης θα το βρει χρήσιμο, οπότε έχουμε ένα *αληθώς θετικό αποτέλεσμα* (true positive result) ή να προταθεί ένα αντικείμενο το οποίο όμως ο χρήστης θα το απορρίψει, οπότε σε αυτή την περίπτωση έχουμε ένα *ψευδώς θετικό αποτέλεσμα* (false positive result) ή διαφορετικά ένα *σφάλμα τύπου II* (type II error). Οι δύο αυτές περιπτώσεις ανταποκρίνονται στην πρώτη στήλη του Πίνακα 3.1. Από την άλλη όμως, το σύστημα συστάσεων μπορεί να «παραβλέψει» ένα αντικείμενο το οποίο ωστόσο ο χρήστης θα το έβρισκε χρήσιμο· αυτή είναι η περίπτωση του ψευδώς αρνητικού αποτελέσματος (false negative result) ή αλλιώς ενός *σφάλματος Τύπου I* (type I error). Τέλος το σύστημα μπορεί να «παραβλέψει» ένα αντικείμενο το οποίο ο χρήστης ούτως ή άλλως θα το έβρισκε μη-επιθυμητό, οπότε τότε θα είχαμε ένα αληθώς αρνητικό αποτέλεσμα (true negative result). Είναι προφανές ότι στόχος κάθε συστήματος συστάσεων είναι να μεγιστοποιήσει τις περιπτώσεις των στοιχείων $(1,1)$ και $(2,2)$ του Πίνακα 3.1 και να ελαχιστοποιήσει εκείνες των $(1,2)$ και $(2,1)$.

Στον Πίνακα 3.1 αναγράφονται και τα αντίστοιχα σύνολα, τα οποία στη συνέχεια θα χρησιμοποιηθούν στους ορισμούς των μετρικών της ακρίβειας ταξινόμησης. Έτσι, ως N ορίζεται το πλήθος των προτάσεων, ως N_{rs} το πλήθος των προτάσεων που είναι όντως χρήσιμες, ως N_{rn} το πλήθος των αντικειμένων που δεν εντάσσονται στις προτάσεις και τα οποία ωστόσο θα ήταν χρήσιμα, ως N_{is} το πλήθος των προτάσεων οι οποίες δεν είναι χρήσιμες και τέλος ως N_{in} το πλήθος των αντικειμένων που δεν εντάσσονται στις προτάσεις και τα οποία όντως δεν είναι χρήσιμα.

Ακρίβεια

Πίνακας 3.1: Ταξινόμηση των παραγόμενων συστάσεων

	Σύσταση του Αντικειμένου	Παράβλεψη του Αντικειμένου	Σύνολο
Σχετικό Αντικείμενο	αληθώς θετική (N_{rs})	ψευδώς αρνητική (N_{rn})	$N_r = N_{rs} + N_{rn}$
Μη-Σχετικό Αντικείμενο	ψευδώς θετική (N_{is})	αληθώς αρνητική (N_{in})	$N_i = N_{is} + N_{in}$
Σύνολο	$N_s = N_{rs} + N_{is}$	$N_n = N_{rn} + N_{in}$	$N = N_s + N_n$ $N = N_r + N_i$

Η *ακρίβεια* (precision) ταξινόμησης εκφράζει την πιθανότητα ένα επιλεγμένο αντικείμενο να αρέσει στον χρήστη. Υπολογίζεται για κάθε χρήστη χωριστά [Herlocker et al., 2004] και στην ουσία εκφράζει τον λόγο των αληθώς θετικών προτάσεων στο σύνολο των ορθών προτάσεων (στοιχείο $(1,1)$ του Πίνακα 3.1 ως προς το άθροισμα της πρώτης στήλης του ίδιου πίνακα) όπως φαίνεται και στην παρακάτω εξίσωση

$$\text{Precision} = \frac{N_{rs}}{N_{rs} + N_{rs}} = \frac{N_{rs}}{N_s} \quad (\%) \tag{3.8}$$

Ένας εναλλακτικός ορισμός της ακρίβειας ταξινόμησης δίνεται από τους [Jamali and Ester, 2009] και περιγράφεται στην ακόλουθη εξίσωση

$$\text{PrecisionT} = 1 - \frac{\text{RMSE}}{\text{RMSE}_{\max}} \tag{3.9}$$

όπου, $\text{RMSE}_{\max}$ η μέγιστη θεωρητική τιμή που μπορεί να πάρει η ρίζα του μέγιστου τετραγωνικού σφάλματος. Συνεπώς το πεδίο τιμών του PrecisionT κυμαίνεται στο $[0, 1]$

Ανάκληση

Η *ανάκληση* (recall) εκφράζει την πιθανότητα να επιλεγεί ένα αντικείμενο που αρέσει στον χρήστη. Υπολογίζεται για κάθε χρήστη χωριστά [Herlocker et al., 2004] και εκφράζει το λόγο των αληθώς θετικών προτάσεων στο σύνολο των προτάσεων, θετικών και αρνητικών (στοιχείο $(1,1)$ του Πίνακα 3.1 ως προς το άθροισμα της πρώτης γραμμής του ίδιου πίνακα), όπως φαίνεται παρακάτω

$$\text{Recall} = \frac{N_{rs}}{N_{rs} + N_{rn}} = \frac{N_{rs}}{N_r} \quad (\%) \tag{3.10}$$

Η ανάκληση, βάσει του ορισμού της, επίσης εκφράζει και τον *λόγο των θετικών αποτελεσμάτων* (true positive rate) [Shani and Gunawardana, 2011].

Μεγέθη Fn

Σε αρκετές περιπτώσεις πρακτικών υλοποιήσεων συστημάτων συστάσεων, τα μεγέθη της ακρίβειας και της ανάκλησης εμφανίζουν μια συμπληρωματική συμπεριφορά. Δηλαδή, υπάρχουν συστήματα που επιτυγχάνουν καλή επίδοση στο ένα μέγεθος και όχι τόσο καλή στο άλλο. Ή, σε περιπτώσεις συγκρίσεων, το ένα σύστημα να υπερτερεί στο ένα μέγεθος και να υστερεί στο άλλο. Για να γίνει εφικτή η ενιαία σύγκριση των συστημάτων, χρησιμοποιούνται τα μεγέθη Fn και πιο συγκεκριμένα ο αρμονικός μέσος της ακρίβειας και της ανάκλησης ή αλλιώς, μέγεθος $F1$ [Herlocker et al., 2004]

$$\text{F1} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.11}$$

Σχήμα 3.1: Ένα σύστημα που κάνει τετριμμένες προτάσεις θεωρείται μη-αποδοτικό ακόμα και όταν οι συστάσεις είναι ακριβείς (Πηγή: Rina Piccolo)

Σε άλλες εργασίες χρησιμοποιείται ο αρμονικός μέσος της ακρίβειας και της κάλυψης (η οποία θα παρουσιαστεί στην Ενότητα 3.4)

$$\mathrm{F1-PC} = \frac{2.\mathrm{PrecisionT.Coverage}}{\mathrm{PrecisionT} + \mathrm{Coverage}} \tag{3.12}$$

3.3 Μέτρηση της Ποιότητας

Στην προηγούμενη ενότητα έγινε αναφορά στις μετρικές απόδοσης των RS. Σημαντικό, ωστόσο, ρόλο στην αλληλεπίδραση του χρήστη με ένα σύστημα συστάσεων παίζει και η ποιότητα των προτάσεων που του γίνονται (Σχήμα 3.1). Στην παρούσα ενότητα θα παρουσιαστούν μετρικές που εστιάζουν σε όψεις του αυτού του χαρακτηριστικού.

3.3.1 Απρόσμενη Ανακάλυψη

Η *απρόσμενη ανακάλυψη* (serendipity) είναι μέγεθος το οποίο αναλύει ποιοτικά τα αντικείμενα που προτείνονται από το σύστημα συστάσεων [Ge et al., 2010]. Στόχος είναι να εξεταστεί κατά πόσο οι παραγόμενες συστάσεις είναι γνωστές ή τετριμμένες. Για το λόγο αυτό, η απρόσμενη ανακάλυψη των συστάσεων ενός RS εξετάζεται συγκριτικά ως προς ένα άλλο, *απλό*, σύστημα, το οποίο είναι εκ των προτέρων γνωστό ότι παράγει τετριμμένες συστάσεις.

Αν υποτεθεί ότι PriMod είναι το σύνολο των αντικειμένων που προτείνει το απλό σύστημα και RecSys το σύνολο των αντικειμένων που προτείνει το σύστημα συστάσεων, τότε ως UNEXP ορίζεται το σύνολο των απρόσμενων συστάσεων

$$\text{UNEXP} = \text{RecSys} - \text{PriMod} \tag{3.13}$$

Καθώς κάθε απροσδόκητη σύσταση δεν είναι απαραίτητα και ωφέλιμη για τον χρήστη, λαμβάνεται υπόψη η ωφέλεια του κάθε αντικειμένου που προτείνεται. Έτσι αν υποτεθεί ότι το αντικείμενο i ανήκει στο σύνολο UNEXP ($i \in \text{UNEXP}$) τότε

$$\begin{aligned} u(i) &= 1, \quad \text{αν } i \text{ χρήσιμο} \\ u(i) &= 0, \quad \text{διαφορετικά} \end{aligned} \tag{3.14}$$

Έτσι πλέον η απρόσμενη ανακάλυψη μπορεί να οριστεί ως

$$\text{SRDP} = \frac{1}{N} \sum_{i=1}^{N} u(i) \tag{3.15}$$

όπου N το σύνολο των αντικειμένων που προτείνονται.

3.3.2 Καινοτομία και Ποικιλομορφία Αντικειμένων

Οι δύο συγκεκριμένες μετρικές ποσοτικοποιούν το πόσο μη-συνηθισμένα και μη-προφανή είναι τα αντικείμενα που περιέχονται σε μια λίστα συστάσεων προς ένα χρήστη. Σε γενικές γραμμές, οι προφανείς συστάσεις θεωρούνται κακής ποιότητας ακόμα και στην περίπτωση που είναι ορθές όσον αφορά την ακρίβεια της πρόβλεψης και της ταξινόμησης τους [Herlocker et al., 2004].

Η *καινοτομία* (novelty) εξετάζει τον βαθμό στον οποίο ένα αντικείμενο (ή ένα σύνολο αντικειμένων) είναι καινοτόμο σε σύγκριση με τα αντικείμενα που έχουν ήδη αξιολογηθεί από τον χρήστη (ή μια κοινότητα των χρηστών). Παρότι στη βιβλιογραφία έχουν προταθεί αρκετά μοντέλα μέτρησης της καινοτομίας των προτεινόμενων αντικειμένων, στην παρούσα διατριβή χρησιμοποιήθηκε η *γενική καινοτομία αντικειμένου βασισμένη στη δημοφιλία* (generic popularity-based item novelty), η οποία μετρά την καινοτομία ενός αντικειμένου [Castells et al., 2011] σύμφωνα με την παρακάτω εξίσωση

$$\text{novelty(i)} = I(i) = -\log_2 p(i) \tag{3.16}$$

όπου $p(i)$ είναι η πιθανότητα εμφάνισης του αντικειμένου i στην λίστα των συστάσεων. Συνηθέστερα, η συγκεκριμένη πιθανότητα θεωρείται πως είναι ανάλογη του λόγου του πλήθους των αξιολογήσεων που έχει λάβει το εν λόγω αντικείμενο ($|R_i|$) προς το σύνολο των αξιολογήσεων που υπάρχουν στο σύστημα ($|R|$)

$$p(i) \sim \frac{|R_i|}{|R|} \tag{3.17}$$

Εκτός από την καινοτομία ενός μεμονωμένου αντικειμένου, μπορεί να μετρηθεί και η καινοτομία μιας λίστας αντικειμένων συνολικά. Σε αυτή την περίπτωση, η μετρική που χρησιμοποιήθηκε ήταν αυτή της *καινοτομίας αντικειμένου βασισμένης στην απόσταση* (distance-based item novelty) [Castells et al., 2011], η οποία αθροίζει την καινοτομία του κάθε αντικειμένου της λίστας συστάσεων ως προς όλα τα υπόλοιπα αντικείμενα που έχει ήδη αξιολογήσει ο υπό εξέταση χρήστης (σύνολο S)

$$novelty(i|S) = \sum_{j \in S} p(j|S) d(i,j) \tag{3.18}$$

όπου $p(j|S)$ η πιθανότητα το αντικείμενο j να έχει ήδη αξιολογηθεί και $d(i,j)$ είναι μια μετρική της απόστασης μεταξύ των αντικειμένων i και j. Συνηθέστερα, η απόσταση δύο αντικειμένων σχετίζεται άμεσα με την μεταξύ τους ομοιότητα σύμφωνα με την σχέση

$$d(i,j) = 1 - \text{sim}(i,j) \tag{3.19}$$

Ως συνάρτηση ομοιότητας μπορούν να χρησιμοποιηθούν οι γνωστοί δείκτες ομοιότητας που αναφέρθηκαν στην Ενότητα 2.3.1, όπως ο συντελεστής συσχέτισης Pearson, η ομοιότητα συνημιτόνου κ. ά.

Τέλος, η *ποικιλομορφία* (diversity) μετρά τον βαθμό στον οποίο τα αντικείμενα που υπάρχουν σε μια λίστα προτάσεων διαφέρουν μεταξύ τους. Μια λίστα προτεινόμενων αντικειμένων τα οποία ναι μεν ενδιαφέρουν τον χρήστη αλλά είναι παρόμοια μεταξύ τους δεν θεωρείται πως συνεισφέρει στην ποιότητα των παραγόμενων συστάσεων. Ο πιο διαδεδομένος τρόπος μέτρησης της ποικιλομορφίας είναι διαμέσου της μετρικής της *εσωτερικής ποικιλομορφίας της λίστας* των προτεινόμενων αντικειμένων (intra-list diversity)

$$\text{diversity}(L) = \frac{2}{|L|(|L|-1)} \sum_{k<n} d(i_n, i_k) \tag{3.20}$$

όπου L η λίστα των n προτεινόμενων αντικειμένων, $i_n, i_k \in L$ και $d(i,j)$ μια μετρική της απόστασης, η οποία ορίζεται όπως προηγουμένως (Εξίσωση 3.19).

3.3.3 Κανονικοποιημένη Μειούμενη Αθροιστική Απολαβή

Η *κανονικοποιημένη μειούμενη αθροιστική απολαβή* (normalized discounted cumulative gain) είναι ένα μέγεθος που αξιολογεί το πόσο «ψηλά» σε μια λίστα προτάσεων βρίσκονται αντικείμενα που ο χρήστης τα θεωρεί ωφέλιμα. Πρόκειται δηλαδή για μια μετρική *ακρίβειας της ταξινόμησης* (rank accuracy metric), η οποία ορίζεται ως ο λόγος της *μειούμενης αθροιστικής απολαβής* (DCG) ως προς την ιδεατή μειούμενη αθροιστική απολαβή (IDCG)

$$\text{nDCG}_N = \frac{\text{DCG}_N}{\text{IDCG}_N} \tag{3.21}$$

όπου το μέγεθος (DCG) είναι ίσο με

$$\text{DCG}_N = \sum_{i=1}^{N} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{3.22}$$

και rel_i είναι ένας δυαδικός δείκτης, ο οποίος ισούται με τη μονάδα αν το i-οστό αντικείμενο της λίστας κρίνεται ωφέλιμο για τον χρήστη (μηδέν διαφορετικά). Ο ιδεατός DCG (IDCG) στη θέση i εκφράζει το μέγεθος που θα είχε ο DCG αν όλες οι προτάσεις στη λίστα συστάσεων είχαν κριθεί ωφέλιμες για τον υπό εξέταση χρήστη. Η διαίρεση του DCG από τον IDCG γίνεται με σκοπό να καταστεί εφικτή η σύγκριση λιστών διαφορετικού μεγέθους.

3.4 Κάλυψη

Η *κάλυψη* (coverage) είναι ένα μέγεθος που μετράει σε ποιες περιπτώσεις μπορεί το σύστημα συστάσεων να παράξει αποτέλεσμα. Χρησιμοποιείται σε πειραματικές διαδικασίες κατά τις οποίες ένα μέρος από των διαθέσιμων αξιολογήσεων αφαιρούνται από το σύστημα και κατόπιν ζητείται από τους αλγορίθμους RS να προσεγγίσουν τις συγκεκριμένες αξιολογήσεις. Ορίζεται ως προς τα τρία συστατικά μέρη των RS: τους *χρήστες*, τα *αντικείμενα* και τις *αξιολογήσεις*.

3.4.1 Κάλυψη Αξιολογήσεων

Η *κάλυψη αξιολογήσεων* (ratings' coverage) ορίζεται ως ο λόγος του πλήθους των προτάσεων που μπόρεσε να ανακτήσει το σύστημα ως προς το πλήθος των προτάσεων που είχαν αφαιρεθεί [Massa and Avesani, 2009; Patricia Victor, 2011]

$$\text{Ratings Coverage} = \frac{|R_p|}{|R|} \quad (\%) \tag{3.23}$$

όπου R_p, το σύνολο των «κρυμμένων» αξιολογήσεων που ανακτήθηκαν από τον αλγόριθμο και R το σύνολο των όλων των «κρυμμένων» αξιολογήσεων.

3.4.2 Κάλυψη Χρηστών

Ως *κάλυψη χρηστών* (user coverage) ορίζεται ο λόγος του πλήθους των χρηστών για τους οποίους το σύστημα μπορεί να κάνει έστω και μία πρόβλεψη ως προς το συνολικό πλήθος των χρηστών στο σύστημα [Massa and Avesani, 2009, 2004]

$$\text{User Coverage} = \frac{|U_p|}{|U|} \quad (\%) \tag{3.24}$$

όπου U_p το σύνολο των χρηστών προς τους οποίους μπορεί να γίνει έστω και μια πρόταση και U το σύνολο των χρηστών που υπάρχουν στο σύστημα.

3.4.3 Κάλυψη Αντικειμένων

Ως *κάλυψη αντικειμένων* (item coverage) ορίζεται το μέρος εκείνο των αντικειμένων τα οποία μπορούν να προταθούν από τον αλγόριθμο του συστήματος συστάσεων [Ge et al., 2010]. Αναλύεται στις ακόλουθες περιπτώσεις:

Απλή Κάλυψη Αντικειμένων

Η *απλή κάλυψη αντικειμένων* (item (prediction) coverage) είναι ο λόγος του πλήθους των αντικειμένων που προτάθηκαν έστω και μια φορά προς το συνολικό πλήθος των αντικειμένων [Ge et al., 2010]

$$\text{Item (Prediction) Coverage} = \frac{|I_p|}{|I|} \quad (\%) \tag{3.25}$$

όπου I_p το σύνολο των αντικειμένων τα οποία μπορούν να προταθούν έστω σε έναν χρήστη και I το σύνολο των αντικειμένων που υπάρχουν στο σύστημα.

Ζυγισμένη Κάλυψη Αντικειμένων

Η *ζυγισμένη κάλυψη αντικειμένων* (weighted item (prediction) coverage) λαμβάνει υπόψη της την χρησιμότητα των αντικειμένων που έχουν προταθεί [Ge et al., 2010]

$$\text{Weighted Item (Prediction) Coverage} = \frac{\sum_{i \in I_p} r(i)}{\sum_{j \in I} r(j)} \quad (\%) \tag{3.26}$$

όπου $r(i)$ μια μετρική της χρησιμότητας του αντικειμένου i, I_p το σύνολο των αντικειμένων που προτάθηκαν έστω και μια φορά και I το σύνολο των αντικειμένων που υπάρχουν στο σύστημα.

Κάλυψη Καταλόγου

Εκτός από την κάλυψη ενός μεμονωμένου αντικειμένου, η κάλυψη μπορεί να οριστεί και σε επίπεδο λίστας προτεινόμενων αντικειμένων. Πρόκειται για την *κάλυψη καταλόγου* (catalog coverage) που προκύπτει από την ακόλουθη εξίσωση [Ge et al., 2010]

$$\text{Catalog Coverage} = \frac{\left| \bigcup_{j=1...N} I_L^j \right|}{|I|} \quad (\%) \tag{3.27}$$

όπου I_L^j τα αντικείμενα που περιέχονται στην λίστα L την j-οστή φορά εκτέλεσης του αλγορίθμου του συστήματος συστάσεων, N ο συνολικός αριθμός εκτελέσεων του αλγορίθμου και I το σύνολο των αντικειμένων που υπάρχουν στο σύστημα.

Ζυγισμένη Κάλυψη Καταλόγου

Κατ΄ αντιστοιχία με την ζυγισμένη κάλυψη αντικειμένου, ορίζεται η *ζυγισμένη κάλυψη καταλόγου* (weighted catalog coverage) [Ge et al., 2010] όπως παρακάτω

$$\text{Weighted Catalog Coverage} = \frac{\left| \bigcup_{j=1...N} I_L^j \cap B^j \right|}{\left| \bigcup_{j=1...N} B^j \right|} \tag{3.28}$$

όπου B^j τα χρήσιμα αντικείμενα που επιστρέφονται κατά την j-οστή φορά εκτέλεσης του αλγορίθμου του συστήματος συστάσεων, I_L^j τα αντικείμενα που περιέχονται στην λίστα L την j-οστή φορά εκτέλεσης του αλγορίθμου, N ο συνολικός αριθμός εκτελέσεων και I το σύνολο των αντικειμένων που υπάρχουν στο σύστημα.

3.5 Όψεις των δεδομένων

Εκτός από τη συνολική εξέταση της συμπεριφοράς ενός συστήματος συστάσεων, υπάρχουν περιπτώσεις που παρουσιάζει μεγάλο ενδιαφέρον η απόδοσή του σε ορισμένες καταστάσεις λειτουργίας ή η απόκριση του για συγκεκριμένες κατηγορίες χρηστών και αντικειμένων. Για παράδειγμα, όπως αναφέρθηκε και στην Ενότητα 1.1, ιδιαίτερη έμφαση δίνεται από τους σχεδιαστές των συστημάτων συστάσεων στο πρόβλημα της ψυχρής εκκίνησης, δηλαδή στη δυνατότητα παραγωγής προτάσεων όχι γενικά για όλους, άλλα συγκεκριμένα για τους νεοεισερχόμενους χρήστες (ή για αυτούς που έχουν δώσει πολύ λίγες αξιολογήσεις). Συνεπώς είναι επιθυμητό να μπορεί να αξιολογηθεί η λειτουργία των RS για αυτές τις συγκεκριμένες περιπτώσεις.

Το συγκεκριμένο ζήτημα αντιμετωπίζεται με την χρήση των *όψεων* (views), μιας θεωρητικής κατασκευής που προέρχεται από την ερευνητική περιοχή των βάσεων δεδομένων [Massa and Avesani, 2009]. Οι όψεις αναφέρονται σε ένα υποσύνολο των αρχικών δεδομένων που επιλέγεται σύμφωνα με κάποια κριτήρια. Μπορούν να οριστούν επάνω στους *χρήστες*, στα *αντικείμενα*, στις *βαθμολογίες*, στην *κοινωνικότητα* (πλήθος δεσμών στο κοινωνικό δίκτυο) καθώς και σε συνδυασμό των τριών προαναφερόμενων χαρακτηριστικών. Παρότι, σε γενικές γραμμές, δεν υπάρχουν κοινά αποδεκτοί κανόνες για τα κριτήρια καθορισμού των όψεων, οι κυριότερες από αυτές που απατώνται στην βιβλιογραφία παρουσιάζονται στις ακόλουθες υποενότητες.

3.5.1 Όψεις των Χρηστών

Οι *όψεις των χρηστών* (user views) εξετάζουν ομάδες χρηστών που εμφανίζουν κάποια ιδιαίτερα, κοινά χαρακτηριστικά, πάνω στα οποία μπορεί να μετρηθεί η απόδοση των συστημάτων συστάσεων. Οι σπουδαιότερες, από ερευνητικής άποψης, όψεις χρηστών παρουσιάζονται παρακάτω.

Χρήστες με λίγες Αξιολογήσεις

Οι *χρήστες με λίγες βαθμολογίες* (cold-start users) έχουν αξιολογήσει ελάχιστα αντικείμενα (συνήθως λιγότερα από 5) και αποτελούν την πλειοψηφία των χρηστών σε ένα σύστημα συστάσεων [Massa and Avesani, 2009]. Η κατηγορία αυτή των χρηστών αποτελεί τη δυσκολότερη στην παραγωγή των συστάσεων στα παραδοσιακά συστήματα συστάσεων και ιδιαίτερα σε εκείνα τα οποία υπολογίζουν την ομοιότητα των χρηστών με συντελεστές συσχέτισης, όπως είναι ο συντελεστής συσχέτισης Pearson. Συνεπώς, η απόδοση των αλγορίθμων παραγωγής συστάσεων ενδιαφέρει ιδιαίτερα σε αυτή την κατηγορία χρηστών και για αυτό το λόγο είναι η όψη των δεδομένων που απαντάται συνηθέστερα στην βιβλιογραφία, με αρκετά συστήματα και αλγορίθμους να δίνουν έμφαση σε αυτή την περιοχή.

Πλειοψηφία των Χρηστών

Στην *πλειοψηφία των χρηστών* (majority of users) εντάσσονται οι χρήστες που έχουν αξιολογήσει μέχρι έναν συγκεκριμένο αριθμό αντικειμένων. Περιλαμβάνονται τόσο οι χρήστες με λίγες αξιολογήσεις που αναφέρθηκαν προηγουμένως, όσο και οι χρήστες που έχουν αξιολογήσει ένα μέτριο πλήθος αντικειμένων (moderate raters), λ.χ. μέχρι 10 ή μέχρι 20 [Massa and Bhattacharjee, 2004]. Είναι προφανές ότι η συγκεκριμένη κατηγορία αποτελεί και την πλειοψηφία των χρηστών που υπάρχουν σε ένα σύστημα συστάσεων.

Χρήστες με πολλές Αξιολογήσεις

Οι *χρήστες με πολλές αξιολογήσεις* (heavy raters) έχουν αξιολογήσει αρκετά αντικείμενα (συνήθως δεκάδες ή/και παραπάνω). Αποτελούν την μειοψηφία των χρηστών σε ένα σύστημα συστάσεων, παρόλα αυτά είναι χρήσιμοι για την εξέταση της απόδοσης των αλγορίθμων παραγωγής συστάσεων στο άλλο άκρο της αξιολογικής συμπεριφοράς των χρηστών. Συνηθέστερα, χρήστες με πολλές αξιολογήσεις θεωρούνται όσοι έχουν αξιολογήσει πάνω από 10 αντικείμενα [Massa and Avesani, 2009].

Χρήστες με Άποψη

Οι *χρήστες με άποψη* (opinionated users) έχουν αξιολογήσει αρκετά αντικείμενα, συνήθως άνω των πέντε, στα οποία έχουν δώσει χαρακτηρισμούς όλης της αξιολογικής γκάμας. Πρόκειται για χρήστες με άποψη, καθώς δεν αξιολογούν μόνο τα αντικείμενα τα οποία τους αρέσουν, αλλά και αυτά που δεν τους αρέσουν. Έχουν, δηλαδή, την ιδανική αξιολογική συμπεριφορά από την σκοπιά των συστημάτων συστάσεων. Ένας τυπικός ορισμός των χρηστών με άποψη είναι όσοι έχουν αξιολογήσει τουλάχιστον πέντε αντικείμενα και η τυπική απόκλιση των τιμών που έχουν δώσει είναι πάνω από 1,5 (σε πενταβάθμια κλίμακα) [Massa and Avesani, 2009].

Μαύρα Πρόβατα

Οι χρήστες που χαρακτηρίζονται ως *μαύρα πρόβατα* (black sheep) έχουν αξιολογική συμπεριφορά η οποία αποκλίνει από το κανονικό. Δηλαδή αξιολογούν τα αντικείμενα με διαφορετικό τρόπο απ΄ ότι ο μέσος χρήστης. Ένας χρήστης θεωρείται ότι ανήκει σε αυτήν την κατηγορία όταν έχει δώσει από πέντε αξιολογήσεις και πάνω και η μέση απόσταση της τιμής που έχει δώσει στο αντικείμενο i είναι μεγαλύτερη της μονάδας από τη μέση τιμή όλων των βαθμών που έχει λάβει το αντικείμενο (σε πενταβάθμια κλίμακα) [Massa and Avesani, 2009].

3.5.2 Όψεις των Αντικειμένων

Οι *όψεις των αντικειμένων* (item views) ομαδοποιούν αντικείμενα τα οποία εμφανίζουν κάποια χαρακτηριστικά που χρήζουν ιδιαίτερης μελέτης από τους σχεδιαστές των συστημάτων συστάσεων. Οι δύο δημοφιλέστερες όψεις αντικειμένων παρουσιάζονται παρακάτω

Μη-Δημοφιλή Αντικείμενα

Τα *μη-δημοφιλή αντικείμενα* (niche items) εμφανίζονται να μην ενδιαφέρουν τους χρήστες, μιας και έχουν αξιολογηθεί από ελάχιστους από αυτούς (συνήθως τα εν λόγω αντικείμενα έχουν λάβει κάτω από πέντε αξιολογήσεις το κάθε ένα) [Massa and Avesani, 2009]. Η πραγματικότητα αυτή δεν συνεπάγεται ότι τα συγκεκριμένα αντικείμενα είναι απαραιτήτως περιθωριακά (μπορεί π.χ. να είναι καινούργια, να απατώνται λιγότερο συχνά κ.λ.π.).

Αμφιλεγόμενα Αντικείμενα

Τα *αμφιλεγόμενα αντικείμενα* (controversial items) λαμβάνουν ταυτόχρονα και υψηλές και χαμηλές αξιολογήσεις (ή διαφορετικά δεν υπάρχει σύμπνοια στην κοινότητα των χρηστών για τη χρησιμότητα τους). Τέτοια αντικείμενα θεωρούνται αυτά των οποίων η τυπική απόκλιση, υπολογισμένη στο σύνολο των αξιολογήσεων που έχουν λάβει, ξεπερνά το 1,5 (στην πενταβάθμια κλίμακα) [Massa and Avesani, 2009].

3.5.3 Όψεις Κοινωνικότητας

Η κοινωνικότητα στα συστήματα συστάσεων αναλύεται διεξοδικότερα στο Κεφάλαιο 5. Στην παρούσα ενότητα θα εξεταστεί το πώς οι *όψεις κοινωνικότητας* (social views) ταξινομούν τους χρήστες σε κατηγορίες ανάλογα με τον βαθμό της αλληλεπίδρασής τους με τους άλλους χρήστες του συστήματος.

Κανένας Φίλος

Πρόκειται για την κατηγορία των χρηστών οι οποίοι δεν έχουν δηλώσει καμία προτίμηση φιλίας. Συνήθως αποτελούν ένα μεγάλο ποσοστό των χρηστών. Για παράδειγμα, στη συλλογή δεδομένων από την ιστοσελίδα Epinions που χρησιμοποιείται στα πειράματα των [Massa and Bhattacharjee, 2004], σχεδόν ένας στους τρεις χρήστες δεν έχει δηλώσει κανέναν χρήστη του οποίου να εμπιστεύεται την κρίση. Αυτοί οι χρήστες μπορούν να χρησιμοποιηθούν ως ομάδα αναφοράς από ένα κοινωνικό σύστημα συστάσεων. Αρχικά μετριέται η απόδοση του συστήματος συστάσεων σε αυτή τη βασική κατηγορία

χρηστών και κατόπιν εξετάζεται ο βαθμός βελτίωσης των παραγόμενων συστάσεων για τους χρήστες που συμμετέχουν στο κοινωνικό δίκτυο.

Ένας Φίλος

Αυτή η κατηγορία χρηστών ενδιαφέρει ιδιαίτερα τους αλγορίθμους οι οποίοι συνάγουν το βαθμό εμπιστοσύνης μεταξύ των χρηστών. Για παράδειγμα, έχει βρεθεί [Massa and Bhattacharjee, 2004; Massa and Avesani, 2004] ότι οι παραγόμενες συστάσεις προς χρήστες με λίγες βαθμολογίες μπορούν να βελτιωθούν σημαντικά αν αυτοί έχουν έστω και έναν δεσμό στο κοινωνικό δίκτυο.

Λίγοι Φίλοι

Πρόκειται για τους χρήστες οι οποίοι συνδέονται με μικρό αριθμό άλλων χρηστών, συνήθως κάτω από πέντε [Massa and Bhattacharjee, 2004].

Πολλοί Φίλοι

Είναι η κατηγορία των ενεργών χρηστών του κοινωνικού δικτύου ενός συστήματος συστάσεων, γιατί συνδέονται με πολλούς άλλους χρήστες. Συνήθως σε αυτή την κατηγορία εντάσσονται όσοι έχουν άνω των πέντε δεσμών με άλλους χρήστες [Massa and Bhattacharjee, 2004].

□

Κεφάλαιο 4

Συνεργατικό Σύστημα Συστάσεων βασισμένο στην k Διαχωρισιμότητα

Όπως έχει αναφερθεί επανειλημμένα, η αραιότητα των αξιολογήσεων που περιέχονται σε ένα σύστημα συστάσεων αποτελεί τον κατεξοχήν επιβαρυντικό παράγοντα για τους αλγορίθμους παραγωγής συστάσεων συνεργατικής διήθησης. Είναι, δε, πολύ συχνό το φαινόμενο η πυκνότητα των αξιολογήσεων να πέφτει ακόμα και κάτω από το 1%, καθιστώντας την αναζήτηση ομοιοτήτων στην αξιολογική συμπεριφορά μεταξύ των χρηστών σχεδόν αδύνατη, μιας και οι κριτικές τους συμπίπτουν σε ελάχιστα αντικείμενα.

Δεν είναι τυχαίο ότι σε αυτό το περιβάλλον, οι πιο αποδοτικές μεθοδολογίες είναι αυτές που βασίζονται στην παραγοντοποίηση πινάκων (λ.χ. [Koren, 2009]). Οι συγκεκριμένες τεχνικές κάνουν την θεώρηση ότι οι αξιολογήσεις έχουν την μορφή ενός εξαιρετικά αραιού πίνακα (Ενότητα 2.1.2) πολύ μεγάλων διαστάσεων (πλήθος χρηστών επί πλήθος αντικειμένων), τον οποίο επιθυμούν να εκφράσουν ως το γινόμενο τουλάχιστον δύο άλλων πινάκων, σημαντικά χαμηλότερων διαστάσεων. Οι νέοι αυτοί πίνακες είναι αρκετά πυκνότεροι του αρχικού και, το κυριότερο, περιέχουν *λανθάνουσα* (latent) πληροφορία για τους χρήστες και τα αντικείμενα, στη βάση της οποίας γίνονται οι νέες προτάσεις.

Ωστόσο, η πλειοψηφία των τεχνικών παραγοντοποίησης έχουν μεγάλο υπολογιστικό κόστος και επιπρόσθετα δεν μπορούν να ανανεώνουν εύκολα τις λανθάνουσες μεταβλητές τους όταν προστίθενται νέοι χρήστες, αντικείμενα και αξιολογήσεις στο σύστημα (πρακτικά, θα πρέπει να πραγματοποιείται εκ νέου η παραγοντοποίηση). Για το λόγο αυτό, στο παρόν κεφάλαιο παρουσιάζεται ένας αλγόριθμος συνεργατικής διήθησης κατασκευής μοντέλου, ο οποίος βασίζεται στα τεχνητά νευρωνικά δίκτυα και πιο συγκεκριμένα στα πολυεπίπεδα perceptron. Οι παραδοχές που γίνονται στον τρόπο κατασκευής και εκπαίδευσης του δικτύου, καθώς και στην επιλογή της συνάρτησης ενεργοποίησης των νευρώνων δείχνουν ότι είναι εφικτή η επίτευξη μιας καλής ισορροπίας μεταξύ της ακρίβειας παραγωγής συστάσεων και της ελαχιστοποίησης του υπολογιστικού κόστους.

4.1 Σχεδιαστικά Ζητήματα

4.1.1 Μείωση Διαστάσεων

Ο κατ' εξοχήν τρόπος που αντιμετωπίζεται το πρόβλημα της αραιότητας των αξιολογήσεων στα συστήματα συνεργατικής διήθησης κατασκευής μοντέλου είναι μέσω της χρήσης τεχνικών *μείωσης των διαστάσεων* (dimensionality reduction techniques), οι οποίες με αυτό τον τρόπο αυξάνουν την πυκνότητα του πίνακα των αξιολογήσεων [Sarwar et al., 2000]. Οι νέοι

πίνακες έχουν πολύ μικρότερο μέγεθος από τον αρχικό, όμως αυτός ο μετασχηματισμός έχει ένα συγκεκριμένο πληροφοριακό κόστος. Φυσικά, γίνεται προσπάθεια να διατηρηθεί όσο το δυνατόν ακριβέστερα η σημασιολογική δομή του αρχικού πίνακα. Πολλές τεχνικές έχουν αναπτυχθεί για αυτή τη διαδικασία και οι οποίες προσπαθούν να επιτύχουν την καλύτερη δυνατή ισορροπία μεταξύ μειωμένων διαστάσεων και μικρού υπολογιστικού κόστους.

Η θεωρητική αφετηρία των μεθόδων αυτών είναι το γεγονός πως δεν είναι απαραίτητες όλες οι μετρούμενες μεταβλητές ενός πίνακα μεγάλων διαστάσεων για την κατανόηση της δομής του [Fodor, 2002]. Διάφορες μεθοδολογίες έχουν χρησιμοποιηθεί για την αναπαράσταση των πινάκων σε χαμηλότερες διαστάσεις, όπως λ.χ. η *ανάλυση του γραμμικού διαχωρισμού* (linear discriminant analysis - LDA) [Zhang and Iyengar, 2002], η οποία προσπαθεί να εντοπίσει έναν γραμμικό συνδυασμό εκείνων των χαρακτηριστικών που τοποθετούν τα στοιχεία του πίνακα σε διαφορετικές ομάδες. Ή, η *ανάλυση των κυρίων συνιστωσών* (principal component analysis - PCA) [Vozalis and Margaritis, 2007], που ελαττώνει τις διαστάσεις των δεδομένων προσπαθώντας να βρει εκείνους τους ορθογώνιους γραμμικούς συνδυασμούς των αρχικών δεδομένων (τις *κύριες συνιστώσες*), που παρουσιάζουν την ελάχιστη διασπορά.

Μια τεχνική που πετυχαίνει ικανοποιητικά αποτελέσματα [Billsus and Pazzani, 1998] είναι η *λανθάνουσα σημασιολογική δεικτοδότηση* (latent semantic indexing - LSI), που βασίζεται στην *ανάλυση των ιδιαζουσών τιμών* (singular value decomposition - SVD) ενός δοσμένου πίνακα. Η SVD είναι, επί της ουσίας, μια παραγοντοποίηση ενός πίνακα στην ακόλουθη κανονική μορφή

$$A = U\Sigma V^T. \tag{4.1}$$

Ο αρχικός πίνακας A επανυπολογίζεται ως το γινόμενο τριών άλλων πινάκων, U, Σ και V. Οι U και V περιέχουν τα *αριστερά* και *δεξιά ιδιάζοντα διανύσματα* του A. Ο Σ είναι ένας (συνήθως ορθογώνιος) διαγώνιος πίνακας, του οποίου τα κύρια διαγώνια στοιχεία είναι οι *ιδιάζουσες τιμές* του A. Έχουν το χαρακτηριστικό του να είναι μη-αρνητικές και εμφανίζονται σε φθίνουσα σειρά. Οι ιδιάζουσες τιμές, επίσης, ποσοτικοποιούν τον βαθμό της διασποράς στα αρχικά δεδομένα και το πλήθος των μη-μηδενικών από αυτές ισούται με την *τάξη* (rank) του αρχικού πίνακα A.

Στις περιπτώσει των αραιών πινάκων, μια ειδική μορφή του SVD μπορεί να χρησιμοποιηθεί, η οποία ονομάζεται *συνεπτυγμένη μορφή* (compact SVD). Πιο συγκεκριμένα, αν ένας πίνακας A, $r \times c$ διαστάσεων, είναι αραιός, τότε η τάξη του m είναι σημαντικά μικρότερη από τον αριθμό των γραμμών του ($m << r$). Σε αυτή την περίπτωση, μόνο τα m αριστερά ιδιάζοντα διανύσματα (στήλες) του U και τα m δεξιά ιδιάζοντα διανύσματα (γραμμές)του V^T χρειάζεται να υπολογιστούν (γιατί αντιστοιχούν στις m μη-μηδενικές ιδιάζουσες τιμές του Σ). Συνεπώς, ο αρχικός πίνακας A μπορεί να επανυπολογιστεί ως

$$A' = U_m\Sigma_m V_m^T. \tag{4.2}$$

Ο A' έχει διάσταση $m \times m$, ενώ ο A είχε διάσταση $r \times c$. Η εφαρμογή της συνεπτυγμένης μορφής του SVD παράγει το επιθυμητό αποτέλεσμα. Δηλαδή, ο παραγόμενος πίνακας δεν είναι μόνο σημαντικά χαμηλότερης διάστασης (και άρα, πιο πυκνός) αλλά επιπλέον αποκαλύπτει και κάποιες σημασιολογικές πλευρές του αρχικού πίνακα.

4.1.2 Δεδομένα που περιέχουν θόρυβο

Όπως επισημάνθηκε προηγουμένως, η ελάττωση των διαστάσεων του πίνακα αξιολογήσεων των χρηστών έχει ως αποτέλεσμα την μείωση της «ποιότητας» της πληροφορίας που περιέχεται σε αυτόν. Η διαδικασία της μείωσης των διαστάσεων μπορεί να θεωρηθεί ως το ισοδύναμο της προσθήκης θορύβου στον παραγόμενο πίνακα· πολλά στοιχεία του αποκτούν τιμές οι οποίες είναι ελάχιστα μεγαλύτερες (ή μικρότερες) από το μηδέν. Αυτό είναι ένα σημαντικό στοιχείο που πρέπει να ληφθεί υπόψη κατά τη σχεδίαση του συστήματος συστάσεων.

Είναι προφανές ότι αναζητούνται τεχνικές και μεθοδολογίες κατάλληλες για δεδομένα τα οποία προβλέπεται να έχουν στατιστικές ιδιότητες παρόμοιες με αυτές μιας κανονικής (γκαουσιανής) κατανομής. Συνεπώς ο ιδανικός αλγόριθμος θα πρέπει να μπορεί να εντοπίζει και να εξαγάγει μοτίβα από δεδομένα που εμφανίζονται να είναι σχεδόν «τυχαία».

4.1.3 Εκμάθηση Δεδομένων που περιέχουν θόρυβο

Ο στόχος της εξαγωγής γνώσης από δεδομένα με θόρυβο έχει αναλυθεί εκτεταμένα από την οπτική των τεχνικών *εκμάθησης μηχανής* [Schlimmer and Granger, 1986; Gallant, 1990]. Τα τεχνητά νευρωνικά Δίκτυα έχουν μελετηθεί σε βάθος προς αυτή την ερευνητική κατεύθυνση [Lee et al., 2004; Tian et al., 2006]. Στο πλαίσιό τους, η διαδικασία παραγωγής συστάσεων μπορεί να ιδωθεί είτε ως μια εργασία ταξινόμησης (σε επιβλεπόμενους αλγορίθμους εκμάθησης) είτε ως μια εργασία συσταδοποίησης (σε ημι-επιβλεπόμενους ή μη-επιβλεπόμενους αλγορίθμους μάθησης). Ο σκοπός της διαδικασίας εκμάθησης, στην περίπτωση των συστημάτων συνεργατικής διήθησης που εξετάζονται στην παρούσα διατριβή, είναι να ομαδοποιήσουν τους χρήστες σε γκρουπ, ανάλογα με τις βαθμολογικές τους συνήθειες. Ωστόσο, το κυρίαρχο αποτέλεσμα της προσθήκης θορύβου στα δεδομένα είναι ότι τα όρια αυτών των ομάδων γίνονται πιο ασαφή: με άλλα λόγια, καθίσταται πιο δύσκολος ο εντοπισμός και διαχωρισμός τους. Συνεπώς, η ιδανική αρχιτεκτονική των νευρωνικών δικτύων θα πρέπει να έχει την ικανότητα να μαθαίνει ομάδες δεδομένων που έχουν την ιδιότητα του να μην είναι διαχωρίσιμες σε σημαντικό βαθμό.

4.1.4 Αρχιτεκτονικές Νευρωνικών Δικτύων

Τα *πολυεπίπεδα perceptrons* (multilayer perceptrons - MLP) λειτουργούν ως *καθολικοί προσεγγιστικοί μηχανισμοί* (universal approximators), χρησιμοποιώντας τα κρυφά τους επίπεδα για να μετατρέψουν την είσοδό τους σε γραμμικά διαχωρίσιμες *συστάδες* (clusters). Παρότι αυτή η διαδικασία αποφέρει σημαντικά αποτελέσματα στα περισσότερα προβλήματα, η απόδοσή της είναι κάτω του μετρίου για δεδομένα που είναι μη-γραμμικά διαχωρίσιμα σε υψηλό βαθμό [Duch, 2006]. Μια πρώτη λύση θα ήταν η προσαρμογή του μεγέθους του δικτύου (αυξάνοντας τον αριθμό των νευρώνων ή/και των επιπέδων) έτσι ώστε το MLP να μπορεί να γενικεύσει καλύτερα.

Παρόλα αυτά, αυτή η προσέγγιση είναι προβληματική γιατί υπάρχουν δεδομένα που απαιτούν υπερβολικά μεγάλες τοπολογίες για να κατανοηθούν σωστά από τα MLP. Το πιο τυπικό τέτοιο παράδειγμα αποτελεί το πρόβλημα XOR (και η γενίκευσή του, το πρόβλημα της ισοτιμίας (parity problem)). Από τη θεωρία των νευρωνικών δικτύων είναι γνωστό ότι το πρόβλημα XOR μπορεί να προσεγγιστεί από ένα MLP δύο επιπέδων και τριών νευρώνων στο σύνολο [Haykin, 2008]. Στην επόμενη ενότητα, θα δειχθεί ότι με την κατάλληλη τεχνική μπορεί να χρησιμοποιηθεί ένα απλό perceptron επιτυγχάνοντας έτσι μια συνολική ελάττωση κατά δύο τρίτα στο μέγεθος του απαιτούμενου δικτύου.

Γενικότερα όμως, εκτός των συστημάτων συστάσεων, και άλλες ερευνητικές περιοχές έχουν να επεξεργαστούν δεδομένα που είναι εγγενώς μη-γραμμικά διαχωρίσιμα σε μεγάλο βαθμό. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί και άλλες μεθοδολογίες, στο πλαίσιο των τεχνικών εκμάθησης μηχανής, όπως για παράδειγμα αλγόριθμοι εύρεσης *διανυσμάτων υποστήριξης* (support vector machines - SVM) [Vapnik, 1995], που προβάλλουν τα δεδομένα σε έναν χώρο υψηλότερων διαστάσεων, όπου θα ήταν ευκολότερο να εντοπιστεί ένα υπερεπίπεδο που θα χώριζε τις περιοχές των δεδομένων. Παρότι τα SVM μπορούν και γενικεύουν σε ικανοποιητικό βαθμό, ο υπολογισμός των διανυσμάτων υποστήριξης είναι, τις περισσότερες φορές, μια ιδιαιτέρως δαπανηρή υπολογιστικά διαδικασία. Με έναν αντίστοιχο τρόπο λειτουργούν και οι *ακτινικές συναρτήσεις βάσης* (radial basis functions), προβάλλοντας δηλαδή τα δεδομένα σε χώρο υψηλότερων διαστάσεων, αλλά τώρα ο στόχος είναι να βρεθεί μια

επιφάνεια που παρέχει το καλύτερο, με κάποιο στατιστικό κριτήριο, ταίριασμα για τα δεδομένα.

Στο παρόν κεφάλαιο, ωστόσο, ακολουθήθηκε μια πρωτότυπη προσέγγιση με την υιοθέτηση του θεωρητικού πλαισίου της k-διαχωρισιμότητας. Η μέθοδος αυτή, που θα παρουσιαστεί αναλυτικότερα στην επόμενη ενότητα, επιτυγχάνει τη γραμμική διαχωρισιμότητα των δεδομένων εισόδου χωρίς την ιδιαίτερη αύξηση της πολυπλοκότητας του συστήματος.

4.2 k-Διαχωρισιμότητα

Η k-διαχωρισιμότητα είναι μια μεθοδολογία (που προτάθηκε αρχικά από τους [Duch, 2006] και [Grochowski and Duch, 2009]) που ενσωματώνει ορισμένα χαρακτηριστικά των τεχνικών *αναζήτησης προβολών* (projection pursuit) [Friedman and Tukey, 1974] και η οποία μπορεί να εφαρμοστεί σε αρχιτεκτονικές τεχνητών νευρωνικών δικτύων. Πιο συγκεκριμένα, τα *πολυεπίπεδα perceptrons εμπρόσθιας τροφοδότησης* (feed-forward multi-layer perceptrons) χρησιμοποιούν τα κρυφά τους επίπεδα προκειμένου να μετασχηματίσουν τα δεδομένα της εισόδου τους σε γραμμικά διαχωρίσιμες περιοχές. Αυτές οι περιοχές κατόπιν συνδυάζονται στο επίπεδο εξόδου, όπου χρησιμοποιούνται γραμμικές ή σιγμοειδείς συναρτήσεις ενεργοποίησης, και το αποτέλεσμα του υπολογισμού προβάλλεται σε μια γραμμή η οποία είναι κάθετη στο υπερεπίπεδο διαχωρισμού των περιοχών.

Όπως παρατηρούν οι [Duch, 2006; Grochowski and Duch, 2009], αυτή η συμπεριφορά των πολυεπίπεδων perceptron εμπρόσθιας τροφοδότησης έχει ένα σημαντικό μειονέκτημα: δεν μπορεί να διαχωρίσει μη-γραμμικές περιοχές, ακόμα και αυτές που απαιτούν απλούς μετασχηματισμούς για να γίνουν γραμμικά διαχωρίσιμες. Για αυτό το λόγο προτείνουν να αλλάξει ο σκοπός της διαδικασίας της εκπαίδευσης και έτσι αντί να προσπαθεί το δίκτυο να εντοπίσει γραμμικά διαχωρίσιμες περιοχές να αναζητήσει άλλες μορφές διαχωρισιμότητας.

4.2.1 Χρήση στα απλά perceptron

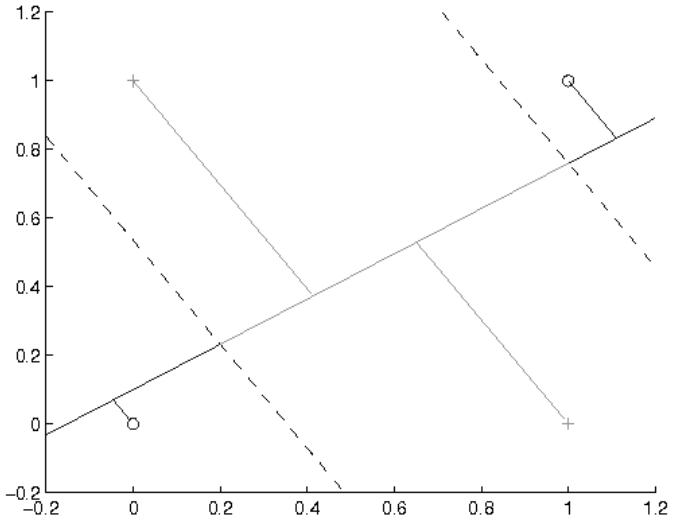
Η k-διαχωρισιμότητα επεκτείνει την γραμμική διαχωρισιμότητα των περιοχών δεδομένων σε $k \geq 2$ περιοχές στο επίπεδο διαχωρισμού. Η λειτουργία αυτή γίνεται πιο ευκρινής στην περίπτωση του 2-bit XOR προβλήματος. Είναι γνωστό από την θεωρία των νευρωνικών δικτύων ότι αυτό το πρόβλημα δεν μπορεί να λυθεί με ένα απλό perceptron. Αντίθετα, απαιτείται μια διάταξη δύο επιπέδων, όπου στο κρυφό επίπεδο θα υπάρχουν δύο νευρώνες, η έξοδος των οποίων θα συνδυάζεται γραμμικά στο επίπεδο εξόδου. Ή, εναλλακτικά, ένα δίκτυο ακτινικών συναρτήσεων βάσης με δύο γκαουσιανές συναρτήσεις στο κρυφό επίπεδο και (πάλι) έναν νευρώνα με γραμμική συνάρτηση ενεργοποίησης στο επίπεδο εξόδου.

Ωστόσο, το 2-bit XOR πρόβλημα μπορεί να επιλυθεί και με ένα απλό perceptron αν αξιοποιηθεί η k-διαχωρισιμότητα στην επιλογή της συνάρτησης ενεργοποίησης. Πιο συγκεκριμένα, η συνάρτηση ενεργοποίησης θα πρέπει να χωρίζει τα δεδομένα εισόδου σε παραπάνω από μια περιοχές, ή ακριβέστερα σε τρεις (Σχήμα 4.1). Κατάλληλες συναρτήσεις για αυτή τη λειτουργία είναι οι συναρτήσεις *μεταβλητού παραθύρου* (soft-windowed) που προκύπτουν από τον πολλαπλασιασμό δύο σιγμοειδών συναρτήσεων με διαφορά φάσης [Grochowski and Duch, 2009]

$$M(\mathbf{x}; \mathbf{w}, a, b, \beta) = \frac{1}{2}[1 - \tanh(\beta(\mathbf{wx} - a)) \tanh(\beta(\mathbf{wx} - b))], \tag{4.3}$$

Το $\mathbf{w}$ είναι το *διάνυσμα βαρών* (που περιλαμβάνει την πόλωση w_0), το $\mathbf{x}$ το διάνυσμα εισόδου και το β ελέγχει την κλήση των σιγμοειδών συναρτήσεων (μετατρέποντας την συνάρτηση σε *κλειστού παραθύρου* (hard-windowed) όταν τεθεί σε μεγάλες τιμές). Τα a και b είναι οι επιπρόσθετες μεταβλητές πόλωσης και θέτουν τα όρια μεταξύ των διαφόρων περιοχών (Σχήμα 4.1).

Σε αυτό το σημείο πρέπει να τονιστεί ότι η M είναι μια *συνθλιπτική* (squashing) συνάρτηση (δηλαδή συγκλίνει σε μια σταθερή τιμή όταν η είσοδος της τείνει στο θετικό ή στο



Σχήμα 4.1: Προβολή των σημείων εισόδου στην γραμμή διαχωρισμού και ο σχηματισμός των 3 περιοχών

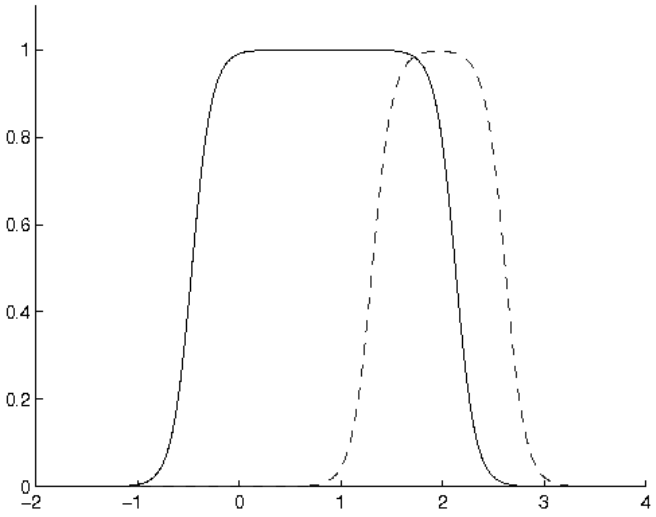
αρνητικό άπειρο) που είναι και διαφορίσιμη , συνεπώς μπορεί να χρησιμοποιηθεί ως συνάρτηση ενεργοποίησης σε ένα perceptron. [Haykin, 2008]. Επίσης είναι εύκολο να αποδειχθεί ότι όταν το *επαγόμενο τοπικό πεδίο* (induced local field) $\mathbf{wx}$ παίρνει τιμές εντός του διαστήματος $[a, b]$, η M παίρνει θετικές τιμές ενώ σε κάθε άλλη περίπτωση παίρνει τη μηδενική τιμή. Αν η κλίση β λάβει μια υψηλή τιμή, τότε $M = 1$ όταν $\mathbf{wx} \in [a, b]$. Οι επιπλέον μεταβλητές πόλωσης ρυθμίζονται κατά τη διάρκεια της εκπαίδευσης και συνεπώς το «σχήμα» της συνάρτησης αλλάζει, ανάλογα με τα δεδομένα της εισόδου, καθώς η εκπαίδευση προχωρά. Αυτή η αλλαγή απεικονίζεται ευκρινέστερα στο Σχήμα 4.2, όπου η διακεκομμένη γραμμή αναπαριστά την συνάρτηση μεταφοράς πριν την εκπαίδευση, ενώ η συνεχής γραμμή αναπαριστά την ίδια συνάρτηση μετά το πέρας της εκπαίδευσης.

Η εκπαίδευση μπορεί να πραγματοποιηθεί με τη χρήση οποιουδήποτε αλγορίθμου *οπίσθιας διάδοσης του σφάλματος* (back-propagation algorithm). Στα πειράματα που πραγματοποιήθηκαν, χρησιμοποιήθηκε η οπίσθια διάδοση του σφάλματος *σταδιακής κλίσης με ορμή και προσαρμοζόμενο ρυθμό εκμάθησης* (gradient descent with momentum and adaptive learning rate backpropagation) γιατί εμφάνισε την καλύτερη σύγκλιση. Οι επιπρόσθετες μεταβλητές πόλωσης αρχικοποιούνται ως εξής [Grochowski and Duch, 2009]

$$\begin{aligned} a &= (\mathbf{wx})_{min} + \tfrac{1}{3}\left|(\mathbf{wx})_{max} - (\mathbf{wx})_{min}\right| \\ b &= (\mathbf{wx})_{min} + \tfrac{2}{3}\left|(\mathbf{wx})_{max} - (\mathbf{wx})_{min}\right| \end{aligned} \tag{4.4}$$

και να μεταβάλλονται με τον ίδιο τρόπο που μεταβάλλεται η πόλωση (w_0).

Η αρχιτεκτονική που έχει παρουσιαστεί μέχρι στιγμής επεκτείνει τη γραμμική διαχωρισιμότητα (ή 2-διαχωρισιμότητα) ενός απλού νευρώνα σε ένα επίπεδο παραπάνω, επιτυγχάνοντας ένα 3-διαχωρίσιμο όριο απόφασης. Πρόκειται για μια σημαντική προέκταση που είναι απολύτως κατάλληλη για προβλήματα απόφασης των οποίων τα δεδομένα ομαδοποιούνται σε 3 διαφορετικές περιοχές όταν προβάλλονται σε μια διάσταση. Ο διαχωρισμός στην περίπτωση του 2-bit XOR προβλήματος απεικονίζεται στο Σχήμα 4.1, όπου οι τρεις εναλλασσόμενες περιοχές στην διαχωριστική γραμμή αποτυπώνονται σε δύο διαφορετικούς τόνους του γκρι: μια σκούρα περιοχή, στη συνέχεια μια ανοιχτή περιοχή και τέλος πάλι μια σκούρα περιοχή. Οι περιοχές που έχουν τον ίδιο τόνο του γκρι υποδηλώνουν τις ομάδες δεδομένων του ιδίου τύπου. Σε αυτή την περίπτωση, το νόημα της εκπαίδευσης έγκειται στην εύρεση της κατάλληλης διαχωριστικής



Σχήμα 4.2: Συνάρτηση Ενεργοποίησης: Πριν την εκπαίδευση (διακεκομμένη γραμμή) και μετά (συνεχής γραμμή)

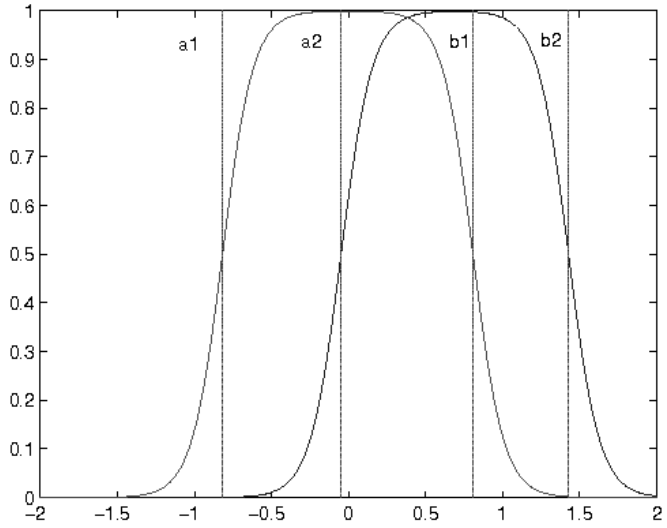
ευθείας και των ορίων πάνω σε αυτή, τα οποία καθορίζονται από τις επιπρόσθετες μεταβλητές πόλωσης a και b ($(-\infty, a]$, $(a, b]$ και (b, ∞)).

Το 2-bit XOR πρόβλημα θα μπορούσε επίσης να έχει λυθεί σε ένα επίπεδο από μια *πολυωνυμική μηχανή εύρεσης διανύσματος υποστήριξης* (polynomial support vector machine)[Haykin, 2008]. Αυτή η επιλογή, όμως, θα επέβαλλε το επιπλέον υπολογιστικό κόστος της εύρεσης πρώτα των βέλτιστων τιμών των πολλαπλασιαστών Lagrange για να καθοριστούν οι τιμές του βέλτιστου διανύσματος υποστήριξης. Αντίθετα, το υπολογιστικό κόστος της k-διαχωρισιμότητας είναι σημαντικά χαμηλότερο. Είναι συγκρίσιμο με το κόστος οποιουδήποτε αλγορίθμου οπίσθιας διάδοσης του σφάλματος συν το κόστος υπολογισμού των δύο επιπρόσθετων μεταβλητών πόλωσης. Συνεπώς, γίνεται εμφανές ότι η k-διαχωρισιμότητα παρουσιάζει τα πλεονεκτήματα και της μικρότερης απαιτούμενης αρχιτεκτονικής αλλά και των λιγότερων υπολογισμών.

4.2.2 Επέκταση στα πολυεπίπεδα perceptron

Η παραπάνω συλλογιστική μπορεί να επεκταθεί και για την περίπτωση των δεδομένων εισόδου που συγκροτούν πάνω από 3 ομάδες. Σε αυτή την περίπτωση, θα απαιτείτο μια αρκετά πιο σύνθετη συνάρτηση ενεργοποίησης και αυτό θα μπορούσε να ήταν ένα δείγμα ότι η συγκεκριμένη τεχνική βρίσκει περιορισμένο πεδίο εφαρμογής. Ωστόσο, αυτή η περίπλοκη συνάρτηση μπορεί να προσεγγιστεί συνδυάζοντας την έξοδο δύο ή περισσοτέρων νευρώνων. Οι [Grochowski and Duch, 2009] φέρνουν το παράδειγμα του προβλήματος της ισοτιμίας μιας n-bit τιμής (n-bit parity problem). Σύμφωνα με την ανάλυσή τους, το συγκεκριμένο πρόβλημα είναι $n + 1$ -διαχωρίσιμο γιατί είναι εύκολο να βρεθεί μια προβολή σε μια ευθεία γραμμή που να χωρίζει τις $n + 1$ εναλλασσόμενες περιοχές (άρτιας και περιττής ισοτιμίας).Για παράδειγμα, το πρόβλημα της 4-bit ισοτιμίας είναι 5-διαχωρίσιμο και αυτό αποδεικνύεται αν οι τιμές της ισοτιμίας για τις 16 διαφορετικές τιμές προβληθούν στο διαχωριστικό υπερεπίπεδο.

Ένα 5-διαχωρίσιμο σύνολο δεδομένων μπορεί να προσεγγιστεί από τον συνδυασμό των εξόδων δύο νευρώνων με συναρτήσεις ενεργοποίησης τύπου «παραθύρου» (Εξίσωση 4.3). Αν (a_1, b_1) είναι οι επιπρόσθετες μεταβλητές πόλωσης για τον πρώτο νευρώνα και (a_2, b_2) για τον δεύτερο αντίστοιχα, τότε η εκπαίδευση του δικτύου θα παράξει ένα παρόμοιο αποτέλεσμα με το Σχήμα 4.3, όπου $a_1 < a_2 < b_1 < b_2$. Είναι εμφανές ότι η συνδυασμένη έξοδος των



Σχήμα 4.3: Συναρτήσεις τύπου «παραθύρου» για το πρόβλημα της ισοτιμίας 4-bit τιμών

δύο νευρώνων ταξινομεί τα δεδομένα εισόδου σε 5 ομάδες, με τις επιπρόσθετες μεταβλητές πόλωσης να παίζουν, και σε αυτή την περίπτωση, το ρόλο του ορίου μεταξύ των περιοχών: $(-\infty, a_1]$, $(a_1, a_2]$, $(a_2, b_1]$, $(b_1, b_2]$ και (b_2, ∞).

Ένα ακόμα αποτέλεσμα της εκπαίδευσης του δικτύου είναι ότι οι δύο νευρώνες δεν «ενεργοποιούνται» (δηλαδή δεν λαμβάνουν τη μέγιστη τιμή τους) ταυτόχρονα. Συνεπώς, μπορούν να διαχωρίσουν τις διαφορετικές τιμές εισόδου, «αποκαλύπτοντας» κατ΄ αυτό τον τρόπο σε ποια ομάδα ανήκει η κάθε μια τιμή. Επεκτείνοντας τη συγκεκριμένη λογική, η συνδυαζόμενη έξοδος m νευρώνων οδηγεί στο σχηματισμό $k = 2m + 1$ διαφορετικών περιοχών στη γραμμή διαχωρισμού. Αυτό το δίκτυο μπορεί πλέον να υλοποιηθεί από ένα MLP ενός κρυφού επιπέδου και ενός επιπέδου εξόδου.

Η n-διάστατη είσοδος του δικτύου προβάλλεται σε m νευρώνες με συναρτήσεις ενεργοποίησης τύπου «παραθύρου» στο κρυφό επίπεδο. Η έξοδος του κρυφού επιπέδου αποτελεί την είσοδο του επιπέδου εξόδου, το οποίο λειτουργεί ως αθροιστής και υπολογίζει τη συνολική έξοδο του δικτύου, η οποία βρίσκεται σε κάποια από τις $2m + 1$ περιοχές της γραμμής διαχωρισμού. Το συγκεκριμένο δίκτυο στην ουσία υλοποιεί μια $n \to m \to 1$ προβολή και είναι κατάλληλο να λειτουργήσει στο πλαίσιο μιας διαδικασίας παραγωγής συστάσεων: δεδομένης της εισόδου ενός αντικειμένου που περιγράφεται από N χαρακτηριστικά (στην περίπτωση των συστημάτων συνεργατικής διήθησης βασισμένα στο αντικείμενο) ή των βαθμολογιών που έχουν δώσει σε ένα αντικείμενο οι N πιο όμοιοι, στον υπό εξέταση χρήστη, χρήστες (στην περίπτωση των απλών συστημάτων συνεργατικής διήθησης), το δίκτυο παράγει στην έξοδό του μια τιμή που αντιπροσωπεύει τη χρησιμότητα του εν λόγω αντικειμένου στον υπό-εξέταση χρήστη.

4.2.3 Κατασκευή του Δικτύου

Η κατασκευή του δικτύου θα ήταν μια απλή διαδικασία αν ο αριθμός των διακριτών ομάδων δεδομένων ήταν γνωστός εκ των προτέρων: σε αυτή την περίπτωση, τα x ομαδοποιημένα δεδομένα εισόδου θα μπορούσαν να προσεγγιστούν από ένα perceptron 2 επιπέδων με $m \approx \frac{x-1}{2}$ νευρώνες στο κρυφό επίπεδο και έναν νευρώνα στο επίπεδο εξόδου. Δυστυχώς, αυτές οι περιπτώσεις είναι σπάνιες. Συνήθως, ο αριθμός των ομάδων δεν μπορεί να εκτιμηθεί αν δεν έχει προχωρήσει η εκπαίδευση σε σημαντικό βαθμό και δεν έχει σχηματιστεί η διαχωριστική γραμμή μαζί με τα κατάλληλα όρια. Το να τεθεί το x σε μια σταθερή τιμή επίσης δεν αποτελεί

καλή λύση, γιατί μπορεί είτε να είναι μικρότερο από το k, οπότε το σφάλμα της εκπαίδευσης θα είναι μεγάλο, ή μεγαλύτερο από το k, οπότε το παραγόμενο δίκτυο θα είναι *υπερεκπαιδευμένο* (overfitted).

Ο βαθμός k της διαχωρισιμότητας αποτελεί συνεπώς παράμετρο του προβλήματος που πρέπει να προσεγγιστεί και η οποία είναι στενά συνδεδεμένη με το μέγεθος του δικτύου. Μιας και οι στατικές αρχιτεκτονικές δικτύων δεν αναμένεται να δώσουν καλά αποτελέσματα, πρέπει να επιλεγούν αλγόριθμοι κατασκευής που μεταβάλλουν δυναμικά το μέγεθος του δικτύου κατά τη διάρκεια της εκπαίδευσης. Αυτοί οι αλγόριθμοι εντάσσονται κυρίως σε τέσσερις κατηγορίες: τους *κατασκευαστικούς* (constructive), τους *περιοριστικούς* (pruning), τους *κατασκευαστικούς-περιοριστικούς* (constructive-pruning) και τέλος τους *εξελικτικούς* (evolutionary) [Islam et al., 2009]. Οι κατασκευαστικοί έχουν ως αφετηρία μια ελάχιστη τοπολογία δικτύου, στην οποία προσθέτους νευρώνες και επίπεδα καθώς η εκπαίδευση προχωρά, ενώ αντίθετα οι περιοριστικοί πρώτα εκπαιδεύουν την μέγιστη δυνατή τοπολογία και κατόπιν αφαιρούν τους νευρώνες εκείνους που έχουν τη μικρότερη συμβολή στην έξοδο. Οι κατασκευαστικοί-περιοριστικοί αλγόριθμοι πρώτα εκκινούν με μια κατασκευαστική φάση η οποία ακολουθείται από μια περιοριστική φάση όπως αντίστοιχα κάνουν και οι εξελικτικοί, με την διαφορά ότι στους δεύτερους η εναλλαγή των φάσεων δεν ακολουθεί κάποια προκαθορισμένη σειρά.

Δεν υπάρχει κάποιος κοινά αποδεκτός κανόνας βάσει του οποίου να καθορίζεται ποιος αλγόριθμος είναι ο καταλληλότερος αλλά αντίθετα η επιλογή του εξαρτάται από την ορθή εκτίμηση των παραμέτρων του προβλήματος. Τα συστήματα συνεργατικής διήθησης που εξετάζονται στο πλαίσιο της παρούσας διατριβής, χρησιμοποιούν σχεδόν κατ΄ αποκλειστικότητα τον πίνακα αξιολογήσεων των χρηστών. Προηγουμένως έγινε η επισήμανση ότι, στη γενική περίπτωση, ο εν λόγω πίνακας είναι πολύ αραιός και συνεπώς η πληροφορία που περιέχει είναι πολύ περιορισμένη. Σε αυτές τις περιπτώσεις (ελάχιστη πληροφορία) καταλληλότερες είναι οι μικρές αρχιτεκτονικές δικτύων οπότε είναι καταρχήν λογικό να επιλεγούν κατασκευαστικοί αλγόριθμοι. Ένα ακόμα επιχείρημα υπέρ των κατασκευαστικών αλγορίθμων είναι ο χρόνος που θα χρειαστεί για να προσεγγιστεί η τελική τοπολογία· αναμένεται να είναι μικρότερους για τους κατασκευαστικούς αλγορίθμους απ΄ ότι στις άλλες τρεις μεθόδους. Ο χρόνος είναι μια σημαντική απαίτηση γιατί είναι επιθυμητό το σύστημα συστάσεων να μπορεί να παράγει αποτελέσματα γρήγορα, χωρίς να χρειάζεται οι χρήστες να περιμένουν πολύ.

Τέλος, οι κατασκευαστικοί αλγόριθμοι ενσωματώνουν καλύτερα την έννοια της k-διαχωρισιμότητας. Ξεκινώντας από μια ελάχιστη αρχιτεκτονική και σταδιακά αυξάνοντας το μέγεθος, υλοποιούν την κύρια λειτουργικότητα των μεθόδων *αναζήτησης προβολών*: ψάχνουν επαναληπτικά για εκείνες τις κατευθύνσεις προβολών σε χώρους χαμηλότερων διαστάσεων όπου η διασπορά των δεδομένων από την κανονική κατανομή μεγιστοποιείται Με αυτό τον τρόπο αναλύεται η λανθάνουσα δομή των δεδομένων εισόδου και υπολογίζονται τα όρια μεταξύ των περιοχών που ομαδοποιούν τα δεδομένα.

4.3 Κατασκευαστικός Αλγόριθμος Νευρωνικών Δικτύων

Ο κατασκευαστικός αλγόριθμος που χρησιμοποιήθηκε στο παρόν Κεφάλαιο βασίζεται στον New Constructive Algorithm (NCA) [Islam et al., 2009]. Ο NCA επιτυγχάνει τόσο την αρχιτεκτονική προσαρμογή όσο και την εκπαίδευση του παραγόμενου δικτύου σε ένα βήμα και επιπλέον χρησιμοποιεί αλγόριθμο οπίσθιας διάδοσης του σφάλματος για την προσαρμογή των συνάψεων των προστιθέμενων νευρώνων. Ο NCA δεν επανεκπαιδεύει όλα τα βάρη των συνάψεων κάθε φορά που προστίθεται ένας καινούργιος νευρώνας για δύο λόγους. Ο πρώτος είναι γιατί η συνολική υπολογιστική πολυπλοκότητα της φάσης της εκπαίδευσης θα αυξηθεί σε σχέσει με τις σταθερές αρχιτεκτονικές δικτύων κατά ένα παράγοντα ίσο με τουλάχιστον $O(|n|)$ (όπου $|n|$ ο συνολικός αριθμός των νευρώνων που έχουν προστεθεί). Ο δεύτερος λόγος είναι

ότι η επαναπροσαρμογή των συνάψεων των νευρώνων που έχουν προστεθεί σε προγενέστερες φάσεις συμβάλει σε μια κατάσταση που είναι γνωστή ως το πρόβλημα του «κινούμενου παραθύρου» (moving target problem), όπου ο κάθε νευρώνας, δηλαδή, αντιμετωπίζει ένα διαρκώς μεταβαλλόμενο περιβάλλον.

Παρότι η προσαρμογή των συνάψεων των καινούργιων νευρώνων σε κάθε στάδιο της εκπαίδευσης επιτυγχάνει και υπολογιστική αποτελεσματικότητα και αποφεύγει το πρόβλημα του κινούμενου παραθύρου, εντούτοις μπορεί να οδηγήσει σε δίκτυα με μεγάλο αριθμό νευρώνων. Αυτό οφείλεται στο γεγονός ότι οι αλγόριθμοι οπίσθιας διάδοσης του σφάλματος σε δίκτυα εμπρόσθιας τροφοδότησης συχνά καταλήγουν όχι στην *ολικά βέλτιστη* λύση (global optimum) αλλά παγιδεύονται σε κάποια *τοπική βέλτιστη* λύση (local optimum).

Το παραπάνω πρόβλημα αντιμετωπίζεται με την εκπαίδευση δύο σταδίων για κάθε νευρώνα που προστίθεται στην αρχιτεκτονική. Η *αρχική μερική εκπαίδευση* (initial partial training) πρώτα καθορίζει τα βάρη των συνάψεων και στο επόμενο στάδιο, η *τελική μερική εκπαίδευση* (final partial training) τα βελτιστοποιεί. Μεταξύ των δύο σταδίων της εκπαίδευσης προστίθεται στα βάρη των συνάψεων ένα μικρό ποσό γκαουσιανού θορύβου (με μέση τιμή 0 και διασπορά 1). Με αυτόν τον τρόπο, το πρόβλημα της τοπικής βέλτιστης λύσης μπορεί να αντιμετωπιστεί, η εκπαίδευση να προχωρήσει και συνεπώς να βελτιωθεί ο ρυθμός σύγκλισης και η δυνατότητα γενίκευσης του δικτύου [Islam et al., 2009].

Ο αλγόριθμος που χρησιμοποιείται στην παρούσα διατριβή (Constructive Neural Network Architecture - CNNA) αποτελεί μετεξέλιξη του NCA. Σύμφωνα με το θεωρητικό μας μοντέλο, η τελική αρχιτεκτονική του δικτύου αποτελείται μόνο από τρία επίπεδα: το επίπεδο εισόδου, το κρυφό επίπεδο και το επίπεδο εξόδου. Καινούργιοι νευρώνες προστίθενται μόνο στο κρυφό επίπεδο και εκπαιδεύονται σε δύο στάδια, όπως αναφέρθηκε παραπάνω (και περιγράφεται αναλυτικότερα στον Αλγόριθμο 1). Όταν και τα δύο στάδια της εκπαίδευσης έχουν ολοκληρωθεί και το σφάλμα δεν πέφτει κάτω από μια καθορισμένη τιμή ϵ_1, τότε αυτό αποτελεί ένδειξη ότι ένας καινούργιος νευρώνας πρέπει να προστεθεί. Πριν την προσθήκη του νέου νευρώνα, το κριτήριο τερματισμού καθορίζει αν αυτός έχει να προσφέρει κάτι στη γενικότερη δυνατότητα γενίκευσης του δικτύου. Αν δεν έχει, αυτό σημαίνει ότι το δίκτυο έχει λάβει την τελική τοπολογική του μορφή και συνεπώς η κατασκευή του σταματά. Ο CNNA περιγράφεται αναλυτικότερα αμέσως παρακάτω (Αλγόριθμος 1).

4.3.1 Κριτήριο Τερματισμού

Ένα από τα πιο βασικά ζητήματα για τους κατασκευαστικούς αλγορίθμους νευρωνικών δικτύων είναι το να αποφασίσουν πότε έχει ολοκληρωθεί η κατασκευή του δικτύου. Αυτό το ζήτημα είναι επίσης γνωστό και ως το *κριτήριο τερματισμού* της κατασκευής του νευρωνικού δικτύου. Τα κριτήρια τερματισμού μπορούν να τοποθετηθούν σε δύο κύριες κατηγορίες: αυτά που τερματίζουν την κατασκευή γρήγορα (early stopping) και αυτά που τερματίζουν αργά (slower stopping). Σε μια εργασία του [Prechelt, 1998] αναφέρεται ότι τα αργά κριτήρια τερματισμού βελτιώνουν οριακά τη δυνατότητα γενίκευσης των δικτύων (κατά μέσο όρο 4%) αλλά από την άλλη απαιτούν τετραπλάσιο χρόνο εκπαίδευσης. Συνεπώς, είναι πιο συμφέρον να χρησιμοποιηθούν κριτήρια γρήγορου τερματισμού.

Τόσο ο NCA όσο και ο CNNA χρησιμοποιούν ένα κριτήριο γρήγορου τερματισμού που περιγράφεται στην εργασία [Prechelt, 1998] και το οποίο επιτυγχάνει μια ισορροπία μεταξύ των *σφαλμάτων εκπαίδευσης* και *επαλήθευσης* (training and validation errors). Αυτή η ισορροπία επιτυγχάνεται με την σύγκριση των μεγεθών της *απώλειας της γενίκευσης* (generalization loss) και της *λωρίδας της εκπαίδευσης* (training strip).

Αν $E_{va}(\tau)$ είναι το σφάλμα επαλήθευσης στην *εποχή* (epoch) εκπαίδευσης τ και $E_{opt}(\tau)$ είναι η μικρότερη τιμή του σφάλματος επαλήθευσης μέχρι την εποχή τ, τότε η απώλεια γε-

Αλγόριθμος 1 Κατασκευαστικός Αλγόριθμος Νευρωνικών Δικτύων (CNNA)

1: **Δημιουργία** ενός MLP 3 επιπέδων. ▷ Ένας νευρώνας εξόδου, κανένας στο κρυφό επίπεδο
2: **Προσθήκη** νέου νευρώνα I στο κρυφό επίπεδο ▷ Με συνάρτηση μεταφοράς όπως στην Εξίσωση 4.3
3: **Αρχικοποίηση** των βαρών του I σε μικρές τυχαίες τιμές.
4: **Αν** I είναι ο πρώτος νευρώνας στο κρυφό επίπεδο **Τότε**
5: $T_0 \leftarrow T$ ▷ T είναι το σύνολο εκπαίδευσης
6: $D_0(p) \leftarrow \frac{1}{m}$ ▷ $\forall p \in T$ και $m \leftarrow |T|$
7: **Αλλιώς**
8: $\{T_n, D_n\} \leftarrow \text{BOOST}(T \cdot E_{tr} \cdot D_{n-1})$ ▷ Κλήση Αλγορίθμου 2
9: **Τέλος Αν**
10: **Όσο** $E_{tr}(i) - E_{tr}(i+\tau) \geq \epsilon_1$ **κάνε** ▷ *Αρχική μερική εκπαίδευση*
11: **Εκπαίδευση** του I στο T_n για τ εποχές ▷ Προκαθορισμένη τιμή για το τ είναι 50 [Islam et al., 2009]
12: **Αν** $GL(\tau) > \alpha P_k(\tau)$ **Τότε** ▷ Κριτήριο τερματισμού
13: Πήγαινε στο **Βήμα 26**
14: **Τέλος Αν**
15: **Τέλος Όσο**
16: Προσθήκη γκαουσιανού θορύβου ($\mu \leftarrow 0$, $\sigma^2 \leftarrow 1$) στα βάρη του νευρώνα I
17: **Όσο** $E_{tr}(i) - E_{tr}(i+\tau) \geq \epsilon_1$ **κάνε** ▷ *Τελική μερική εκπαίδευση*
18: **Εκπαίδευση** του I στο T_n για τ εποχές
19: **Αν** $GL(\tau) > \alpha P_k(\tau)$ **Τότε** ▷ Κριτήριο τερματισμού
20: Πήγαινε στο **Βήμα 26**
21: **Τέλος Αν**
22: **Τέλος Όσο**
23: **Αν** $E_{tr}(i) - E_{tr}(i+\tau) \leq \epsilon_1$ **και** $d \geq \epsilon_2$ **Τότε** ▷ Κριτήρια αρχιτεκτονικής προσαρμογής
24: Πήγαινε στο **Βήμα 2**
25: **Τέλος Αν**
26: **Έξοδος** η τελική αρχιτεκτονική του ANN

νίκευσης στην εποχή τ ορίζεται ως

$$GL(\tau) = 100 \left(\frac{E_{va}(\tau)}{E_{opt}(\tau)} - 1 \right) \tag{4.5}$$

Μια υψηλή απώλεια γενίκευσης υποδηλώνει υπερπροσαρμογή και αποτελεί ένδειξη του ότι η εκπαίδευση πρέπει να σταματήσει. Συνεπώς, θα ήταν λογικό να σταματήσει η εκπαίδευση όταν το προαναφερθέν μέγεθος ξεπεράσει ένα συγκεκριμένο κατώφλι. Ωστόσο, αυτή η μέτρηση από μόνη της δεν αρκεί· οι απώλειες γενίκευσης έχουν μεγαλύτερες πιθανότητες να διορθωθούν όταν τα σφάλματα εκπαίδευσης μειώνονται πολύ γρήγορα. Για αυτό το λόγο, η υπερπροσαρμογή θεωρείται ότι λαμβάνει χώρα όταν τόσο η απώλεια γενίκευσης όσο και το σφάλμα εκπαίδευσης μειώνονται αργά [Prechelt, 1998].

Η παραπάνω συλλογιστική μπορεί να ποσοτικοποιηθεί με τον ορισμό της *προόδου της εκπαίδευσης* (training progress), η οποία μετρά το κατά πόσον το μέσο σφάλμα εκπαίδευσης μιας λωρίδας εκπαίδευσης k είναι μεγαλύτερο από το ελάχιστο σφάλμα εκπαίδευσης στην ίδια λωρίδα. Αν $E_{tr}(\tau)$ είναι το σφάλμα εκπαίδευσης στην εποχή τ, τότε η πρόοδος της εκπαίδευσης ορίζεται ως

$$P_k(\tau) = 1000 \left(\frac{\sum_{\tau'=\tau-k+1}^{\tau} E_{tr}(\tau')}{k \min_{\tau'=\tau-k+1}^{\tau} E_{tr}(\tau')} - 1 \right) \tag{4.6}$$

Ως μήκος k της λωρίδας ορίζεται η αλληλουχία εποχών $n+1, n+2, \ldots, n+k$ με $n \bmod k = 0$. Στις περισσότερες περιπτώσεις είναι ίσο με $k = 5$ [Prechelt, 1998].

Έχοντας καθορίσει τις ποσότητες του σφάλματος γενίκευσης και της προόδου της εκπαίδευσης, ο CNNA αλγόριθμος τερματίζει την κατασκευή του δικτύου όταν το πρώτο μέγεθος γίνει μεγαλύτερο από το δεύτερο κατά έναν παράγοντα α:

$$GL(\tau) > \alpha P_k(\tau) \tag{4.7}$$

Το α συνήθως λαμβάνει την τιμή 10 [Prechelt, 1998].

4.3.2 Κριτήρια Αρχιτεκτονικής Προσαρμογής

Τα κριτήρια της αρχιτεκτονικής προσαρμογής χρησιμοποιούνται από τον CNNA όταν η εκπαίδευση των δύο σταδίων ενός καινούργιου νευρώνα του κρυφού επιπέδου (Αλγόριθμος 1) δεν βελτιώνει την απόδοση του δικτύου. Το πρώτο κριτήριο ορίζεται από την ανισότητα

$$E_{tr}(i) - E_{tr}(i+\tau) \leq \epsilon_1, \quad i = \tau, 2\tau, 3\tau, \ldots, \tag{4.8}$$

όπου $E_{tr}(i)$ και $E_{tr}(i+\tau)$ είναι τα σφάλματα εκπαίδευσης στις εποχές i και $i+\tau$ αντίστοιχα, ενώ το ϵ_1 υποδηλώνει το κατώφλι του σφάλματος εκπαίδευσης, το οποίο λαμβάνει μια πολύ μικρή θετική τιμή (συνήθως $\epsilon_1 = 0,005$) [Islam et al., 2009]. Το δεύτερο κριτήριο ορίζεται από μια ακόμα ανισότητα

$$d = \frac{1}{P_t} \sum_{p=1}^{P_t} |y_k(p) - y_{k+1}(p)| \geq \epsilon_2, \quad k = 1, 2, 3, \ldots \tag{4.9}$$

Οι μεταβλητές $y_k(p)$ και $y_{k+1}(p)$ συμβολίζουν την έξοδο του δικτύου με k και $k+1$ νευρώνες στο κρυφό επίπεδο όταν το δείγμα εκπαίδευσης p παρουσιάζεται στην είσοδό του. Το κατώφλι ϵ_2 λαμβάνει επίσης μια μικρή θετική τιμή ($\epsilon_2 = 0.1$) [Islam et al., 2009].

Η αρχιτεκτονική προσαρμογή τερματίζεται όταν *και οι δύο ανισότητες δεν ικανοποιούνται*. Ο λόγος που λαμβάνονται υπόψη δύο κριτήρια πριν την προσθήκη ενός καινούργιου νευρώνα

στο κρυφό επίπεδο είναι ότι πρέπει να επιτευχθεί *σύγκληση* (Εξίσωση 4.8) ενώ παράλληλα πρέπει να διασφαλισθεί και η *λειτουργικότητα* του. Αν η ολική έξοδος του δικτύου δεν βελτιώνεται κατά πολύ (ή, ακριβέστερα, περισσότερο από ένα κατώφλι ϵ_2) όταν προστίθεται ένας καινούργιος νευρώνας στο κρυφό επίπεδο, τότε αυτό αποτελεί σαφή ένδειξη ότι η τελική αρχιτεκτονική έχει ήδη διαμορφωθεί και ότι το δίκτυο δεν μπορεί πλέον να ανακαλύψει περισσότερες εξαρτήσεις στα δεδομένα εισόδου του.

Αλγόριθμος 2 BOOST: Αλγόριθμος ενισχυμένης μάθησης (βασισμένος στον AdaBoost.M1)

Είσοδος Σύνολο Εκπαίδευσης T, Σφάλμα Εκπαίδευσης E_{tr}, Κατανομή D_t
1: $\epsilon_t \leftarrow \sum_{\forall i: E_{tr}(i) > \epsilon 1} D_t(i)$
2: $\beta_t \leftarrow \frac{\epsilon_t}{1-\epsilon_t}$
3: **Αν** $E_{tr}(i) \leq \epsilon 1$ **Τότε** ▷ Z_t είναι μια μεταβλητή κανονικοποίησης που εξασφαλίζει ότι η D_{t+1} θα παραμείνει κατανομή
4: $\quad D_{t+1}(i) \leftarrow \frac{D_t(i)}{Z_t}\beta_t$
5: **Αλλιώς**
6: $\quad D_{t+1}(i) \leftarrow \frac{D_t(i)}{Z_t}$
7: **Τέλος Αν**
8: **Δημιουργία** του T_{t+1} μέσω δειγματοληψίας με επανατοποθέτηση από το T βασισμένη στην D_{t+1}
9: **Έξοδος** D_{t+1}, T_{t+1}

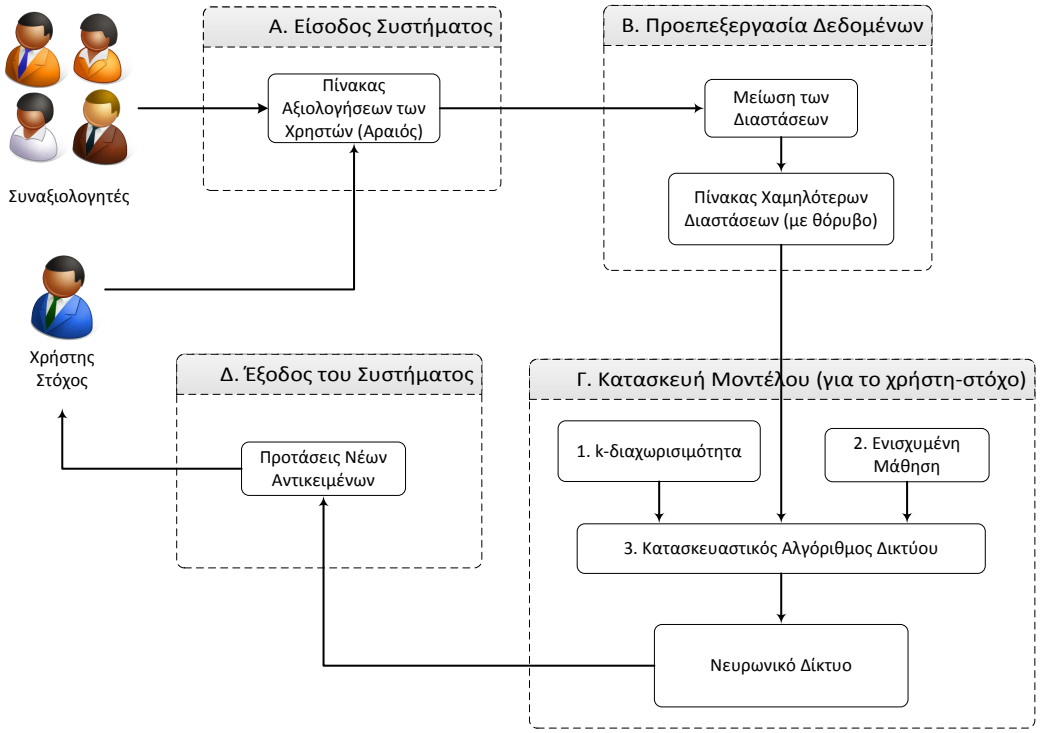
4.3.3 Εκπαίδευση του Δικτύου

Ο CNNA χρησιμοποιεί έναν αλγόριθμο ενισχυτικής μάθησης κατά τη διάρκεια της εκπαίδευσης του δικτύου. Οι αλγόριθμοι ενισχυτικής μάθησης παράγουν διαφορετικά σύνολα δεδομένων εκπαίδευσης τα οποία, ωστόσο, προέρχονται από ένα αρχικά καθορισμένο σύνολο δεδομένων εκπαίδευσης. Ο σκοπός αυτής της επιλογής είναι οι νευρώνες του κρυφού επιπέδου να εστιάσουν σε διαφορετικές όψεις των δεδομένων εκπαίδευσης έτσι ώστε το δίκτυο να «αφομοιώσει» τα συνολικά δεδομένα εκπαίδευσης με τον καλύτερο δυνατό τρόπο.

Ο ενισχυτικός αλγόριθμος μάθησης που επιλέχθηκε σε αυτή την περίπτωση αποτελεί μια παραλλαγή του AdaBoost [Freund and Schapire, 1996], ο οποίος επιλέγει δείγματα από τα δεδομένα εκπαίδευσης βασιζόμενος στο σφάλμα εκπαίδευσης. Πιο συγκεκριμένα, ο ενισχυτικός αλγόριθμος μάθησης που χρησιμοποιείται (Αλγόριθμος 2) ορίζει μια πιθανοτική κατανομή D_i επάνω στο αρχικό σύνολο T των δεδομένων εκπαίδευσης. Κάθε στοιχείο του D_i συμβολίζει την πιθανότητα να επιλεγεί το αντίστοιχο δείγμα για το σχηματισμό του καινούργιου συνόλου εκπαίδευσης. Αρχικά το D_0 λαμβάνει τη μορφή της διακριτής ομοιόμορφης κατανομής, όπου κάθε στοιχείο μπορεί να επιλεγεί με πιθανότητα $\frac{1}{m}$ ($m = |T|$).

Το αρχικό σύνολο δεδομένων εκπαίδευσης T_0 παρουσιάζεται στο δίκτυο. Όταν η εκπαίδευση για αυτό το σύνολο ολοκληρωθεί, υπολογίζεται το σφάλμα ταξινόμησης του κάθε δείγματος και η διακριτή κατανομή D_1 τροποποιείται αναλόγως· η πιθανότητα επιλογής ελαττώνεται για εκείνα τα δείγματα που έχουν μικρό σφάλμα ταξινόμησης (κατά ένα παράγοντα β_t) ενώ αυξάνεται για εκείνα τα δείγματα που παρουσιάζουν μεγάλο σφάλμα ταξινόμησης. Συνεπώς η κατανομή D_1 παύει να είναι ομοιόμορφη: τα δείγματα με μεγαλύτερο σφάλμα ταξινόμησης έχουν μεγαλύτερη πιθανότητα να επιλεγούν για το καινούργιο σύνολο εκπαίδευσης T_1. Το καινούργιο σύνολο εκπαίδευσης T_1 δημιουργείται με τη δειγματοληψία m τυχαίων δειγμάτων με επανατοποθέτηση από το T σύμφωνα με την D_1 (Αλγόριθμος 2).

Η παραλλαγή του AdaBoost που μόλις περιγράφηκε είναι κατάλληλη για την κατασκευαστική αρχιτεκτονική δικτύου που παρουσιάστηκε σε αυτό το κεφάλαιο. Η προσθήκη ενός



Σχήμα 4.4: Διάγραμμα λειτουργίας του προτεινόμενου συστήματος συστάσεων ksepRS

νευρώνα στο κρυφό επίπεδο σε κάθε φάση της εκπαίδευσης είναι μια διαδικασία ισοδύναμη με την προσθήκη από το AdaBoost ενός καινούργιου δικτύου σε ένα σύνολο εκπαίδευσης [Freund and Schapire, 1996]. Επιπλέον, ο αλγόριθμος ενισχυτικής μάθησης εκπαιδεύει μόνο τον καινούργιο νευρώνα του κρυφού επιπέδου σε κάθε φάση κατά παρόμοιο τρόπο που το AdaBoost εκπαιδεύει ένα μόνο δίκτυο από το σύνολο κάθε φορά.

4.4 Παραγωγή Συστάσεων

Συνδυάζοντας τις τεχνικές που έχουν αναφερθεί στις προηγούμενες ενότητες, κατασκευάζεται ένα σύστημα συνεργατικής διήθησης, του οποίου η λειτουργία απεικονίζεται στο διάγραμμα του Σχήματος 4.4. Είσοδος του είναι ο πίνακας βαθμολογιών όλων των χρηστών M και ο υπό εξέταση, κάθε φορά, χρήστης u. Έξοδος του είναι ένα μοντέλο (ένα νευρωνικό δίκτυο) για τον υπό εξέταση χρήστη. Η λειτουργία του σκιαγραφείται στα ακόλουθα βήματα:

1. Από τον πίνακα αξιολογήσεων M επιλέγονται όλοι εκείνοι οι χρήστες που έχουν αξιολογήσει τουλάχιστον δύο αντικείμενα από κοινού με τον υπό εξέταση χρήστη και δημιουργείται ο πίνακας A

2. Υπολογίζεται το SVD του $r \times c$ πίνακα A που έχει τάξη m. Ο πίνακας A είναι αραιός, οπότε $m << r$

3. Υπολογίζεται ο πίνακας A', αφού διατηρηθούν όλες οι m μη-μηδενικές ιδιάζουσες τιμές του A. Ο πίνακας A' έχει μέγεθος $n \times n$ και συνεπώς είναι πολύ μικρότερης διάστασης από τον αρχικό πίνακα A (στην συγκεκριμένη περίπτωση , $n = m$).

4. Εκπαιδεύεται μια δυναμική αρχιτεκτονική δικτύου με τον CNNA (Αλγόριθμος 1) με την χρήση του ενισχυτικού αλγορίθμου μάθησης (Αλγόριθμος 2)

Πίνακας 4.1: Συλλογές Δεδομένων προς χρήση στα Συστήματα Συστάσεων

	`Movielens 100k`	`Movielens 1M`	`Movielens 10M`
Χρήστες	943	6.040	71.567
Αντικείμενα	1.682	3.883	10.681
Αξιολογήσεις	100.000	1.000.209	10.000.054
Πυκνότητα	6,3%	4,26%	1,31%

5. Μετά την ολοκλήρωση της εκπαίδευσης του δικτύου, παράγεται η τελική αρχιτεκτονική του δικτύου για τον υπό εξέταση χρήστη.

4.5 Πειραματική Διαδικασία

4.5.1 Δεδομένα

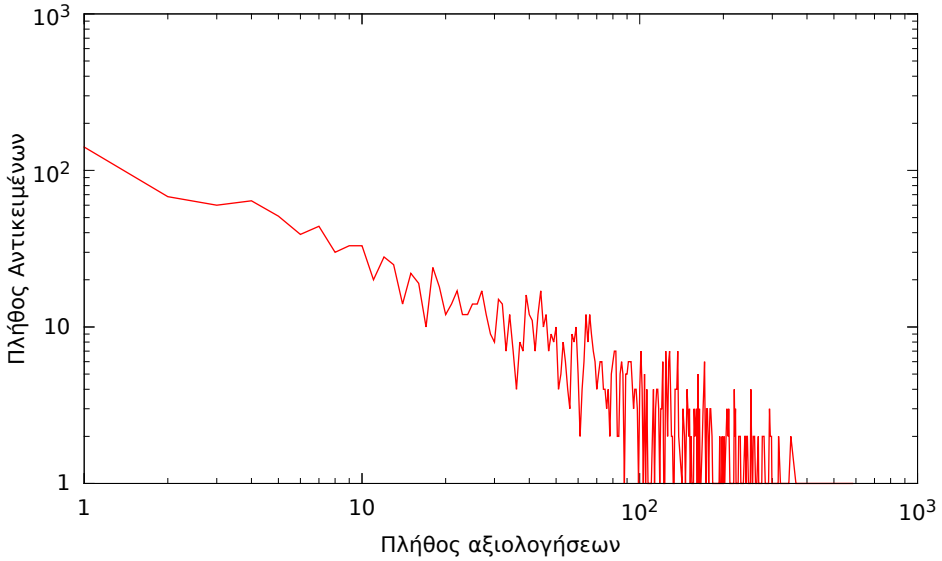
Ένα πλήθος συλλογών δεδομένων είναι δημόσια διαθέσιμες για την πειραματική δοκιμή των διαφόρων συστημάτων συστάσεων. Αναμφίβολα, αυτές που έχουν χρησιμοποιηθεί εκτενέστερα στη βιβλιογραφία είναι οι συλλογές δεδομένων που έχουν συλλεχθεί από το σύστημα MovieLens, ενός ιστοτόπου που προτείνει ταινίες στους χρήστες του και ο οποίος αναπτύχθηκε από την ερευνητική ομάδα GroupLens του Πανεπιστημίου της Μινεσότα [GroupLens and Minnesota]. Από το MovieLens έχουν προκύψει πολλές συλλογές δεδομένων που περιέχουν πίνακες αξιολογήσεων ταινιών (μαζί με ανώνυμες δημογραφικές πληροφορίες για κάθε χρήστη όπως φύλο και ηλικία) και οι οποίες έχουν συλλεχθεί κατά τη διάρκεια διαφορετικών χρονικών περιόδων. Η πιο μικρή, που έχει την ονομασία `MovieLens 100k` , περιέχει 100 χιλιάδες βαθμολογίες που δόθηκαν από χρήστες της υπηρεσίες μεταξύ της 19ης Σεπτεμβρίου 1997 και της 22ης Απριλίου 1998. Η μεσαία, η `MovieLens 1M`, έχει 1 εκατομμύριο βαθμολογίες από τις 26 Απριλίου 2000 μέχρι τις 28 Φεβρουαρίου 2003 και τέλος η μεγαλύτερη, `MovieLens 10M` περιέχει 10 εκατομμύρια βαθμολογίες που καλύπτουν μια χρονική περίοδο άνω της δεκαετίας (από τις 29 Ιανουαρίου 1996 μέχρι τις 5 Ιανουαρίου 2009). Η τελευταία συλλογή διαφέρει από τις άλλες δύο γιατί δεν περιέχει δημογραφικές πληροφορίες για τους χρήστες πέρα από το αναγνωριστικό ταυτότητάς τους ενώ επίσης περιέχει και μεταδεδομένα δημιουργημένα από τους χρήστες, στη μορφή ετικετών (tags). Τα χαρακτηριστικά και των τριών συλλογών συνοψίζονται στον Πίνακα 4.1.

Και οι τρεις συλλογές περιέχουν αξιολογήσεις σε πενταβάθμια διακριτή κλίμακα, όπου το 1 υποδηλώνει την ισχυρή απόρριψη και το 5 την ισχυρή αποδοχή. Είναι όλες αραιές με την μεγαλύτερη να είναι υπερβολικά αραιή (περίπου 1% πυκνότητα βαθμολογιών). Στο MovieLens ο κάθε χρήστης έχει βαθμολογήσει τουλάχιστον 20 ταινίες, πράγμα που σημαίνει ότι δεν περιέχονται χρήστες ψυχρής εκκίνησης (Ενότητα 1.1.2). Παρόλα αυτά, και οι τρεις συλλογές εμφανίζουν τα χαρακτηριστικά *δικτύων ανεξάρτητων από την κλίμακα* (scale-free networks) όσον αφορά το πλήθος των αξιολογήσεων που έχει λάβει το κάθε αντικείμενο, όπως φαίνεται στο Σχήμα 4.5 για την περίπτωση του `MovieLens 100k`.

Παρότι η υπηρεσία Movielens περιείχε επιπρόσθετες πληροφορίες για τον κάθε χρήστη (π.χ. φύλο, ηλικία, απασχόληση) και ταινία (π.χ. τίτλο, είδος, ημερομηνία πρώτης προβολής), εντούτοις αυτές δεν λήφθηκαν υπόψη για λόγους ορθότερης σύγκρισης με άλλες αντίστοιχες υλοποιήσεις.

4.5.2 Πειραματικό Πρωτόκολλο

Το πειραματικό πρωτόκολλο που ακολουθήθηκε για την μέτρηση της απόδοσης του προτεινόμενου συστήματος συνεργατικής διήθησης περιγράφεται στα ακόλουθα βήματα:



Σχήμα 4.5: Κατανομή πλήθους αξιολογήσεων ανά αντικείμενο στο `MovieLens 100k`

1. Η συλλογή δεδομένων χωρίζεται σε τρία μέρη· στο τμήμα της εκπαίδευσης M_{train}, στο τμήμα της επαλήθευσης M_{val} και στο τμήμα δοκιμής M_{test} με αναλογία $60\% - 20\% - 20\%$.

2. Για τον υπό εξέταση χρήστη u, αφαιρούνται οι αξιολογήσεις του από τα M_{train} και M_{val} και δημιουργούνται οι πίνακες T_{train} και T_{val} που περιέχουν τους «στόχους» του δικτύου (training and validation targets), δηλαδή τις τιμές που πρέπει να «μάθει»

3. Επιλέγονται όλοι οι *συμβαθμολογητές* (coraters) του u και δημιουργείται το σύνολο c. Ως συμβαθμολογητές θεωρούνται όλοι εκείνοι οι χρήστες που έχουν βαθμολογήσει τουλάχιστον δύο αντικείμενα από κοινού με τον u στον πίνακα M_{train}. Από τις αξιολογήσεις τους κατασκευάζονται οι πίνακες R_{train} και R_{val}.

4. Υπολογίζεται το SVD των R_{train} και R_{val}. Επανυπολογίζονται οι πίνακες R'_{train} και R'_{val}, διατηρώντας μόνο το 80% των μεγαλύτερων μη-μηδενικών ιδιαζουσών τιμών τους.

5. Εκπαιδεύεται ένα νευρωνικό δίκτυο με τον CNNA (Αλγόριθμος 1) με χρήση ενισχυτικής μάθησης (Αλγόριθμος 2). Η είσοδος του δικτύου είναι οι πίνακες R'_{train} και R'_{val} ενώ οι στόχοι της μάθησης περιέχονται στους πίνακες T_{train} και T_{val}

6. Μετά την ολοκλήρωση της εκπαίδευσης του δικτύου και του τελικού μοντέλου για τον χρήστη u, λαμβάνονται οι βαθμολογίες του από τον πίνακα M_{test} και δημιουργείται ο πίνακας T_{test}. Λαμβάνονται από τον M_{test} οι βαθμολογίες εκείνων των χρηστών στο c που έχουν βαθμολογήσει τουλάχιστον ένα αντικείμενο από κοινού με τον u και δημιουργείται ο πίνακας R_{test}

7. Υπολογίζεται ο SVD του R_{test}, διατηρείται το 80% των μεγαλύτερων μη-μηδενικών ιδιάζουσων τιμών και επανυπολογίζεται ο πίνακας R'_{test}. Κατόπιν, αυτός παρουσιάζεται ως είσοδος στο δίκτυο και λαμβάνεται η έξοδος του δικτύου Y.

8. Μετριέται η απόδοση χρησιμοποιώντας

   (α′) Μετρικές ακρίβειας της ταξινόμησης και πιο συγκεκριμένα την *ακρίβεια*, την *ανάκληση* και το *F1* στο σύνολο των συστάσεων. Η τιμή του κατωφλιού ωφέλειας για τις προαναφερόμενες μετρικές τίθεται το *τέσσερα* (στην πενταβάθμια κλίματα).

Πίνακας 4.2: Αποτελέσματα των μετρικών ακρίβειας ταξινόμησης στο `MovieLens 100k`

Σύστημα	Ακρίβεια	Ανάκληση	F1
ksepRS	**74,65%**	**83,05%**	**79,02%**
MovieMagician Clique-based [Grant and McCalla, 2001]	74,00%	73,00%	74,00%
Movielens [Grant and McCalla, 2001]	66,00%	74,00%	70,00%
SVD/ANN [Billsus and Pazzani, 1998]	67,90%	69,70%	68,80%
MovieMagician Feature-based [Grant and McCalla, 2001]	61,00%	75,00%	67,00%
MovieMagician Hybrid [Grant and McCalla, 2001]	73,00%	56,00%	63,00%
Correlation [Billsus and Pazzani, 1998]	64,40%	46,80%	54,20%

Τα αποτελέσματα είναι ταξινομημένα σε φθίνουσα σειρά της τιμής του F1

(β′) Μετρικές ακρίβειας της ταξινόμησης· MAE και RMSE στο $Y - T_{test}$

Το παραπάνω πειραματικό πρωτόκολλο εφαρμόστηκε και στις τρεις συλλογές δεδομένων. Στην αμέσως επόμενη ενότητα παρουσιάζονται τα αποτελέσματα και συγκρίνονται με τα αντίστοιχα αποτελέσματα άλλων υλοποιήσεων.

4.6 Αποτελέσματα

4.6.1 Συνολική Απόδοση του Συστήματος

Ο Πίνακας 4.2 συνοψίζει την επίδοση του συστήματος με βάση τις μετρικές ακρίβειας ταξινόμησης (ακρίβεια, ανάκληση, F1) καθώς και τις αντίστοιχες επιδόσεις άλλων συστημάτων στο `MovieLens 100k`. Τα αποτελέσματα είναι ταξινομημένα σε φθίνουσα σειρά της τιμής του F1.

Ο Πίνακας 4.2 περιέχει επίσης τα αποτελέσματα του συστήματος MovieMagician [Grant and McCalla, 2001] στην ίδια συλλογή δεδομένων. Το MovieMagician χρησιμοποιεί απλές μετρήσεις ομοιότητας για να σχηματίσει ομάδες αντικειμένων που μοιράζονται κοινά χαρακτηριστικά και ομάδες χρηστών που παρουσιάζουν παρόμοια αξιολογική συμπεριφορά. Αυτές οι ομάδες αποτελούν τη βάση των συνεργατικών προσεγγίσεων (απλών και βασισμένων στα αντικείμενα) που αναπτύσσονται στην εν λόγω εργασία (Feature-based και Clique-based αντίστοιχα) που, όταν συνδυαστούν, σχηματίζουν την υβριδική προσέγγιση.

Επίσης συμπεριλαμβάνεται το συνεργατικό σύστημα συστάσεων που προτείνεται από τους [Billsus and Pazzani, 1998]. Ο αλγόριθμος Correlation χρησιμοποιεί τον συντελεστή συσχέτισης Pearson προκειμένου να ανακαλύψει χρήστες με παρόμοιο γούστο (και χωρίς να πραγματοποιεί άλλα βήματα προ-επεξεργασίας ή μετά-επεξεργασίας) ενώ ο αλγόριθμος SVD/ ANN χρησιμοποιεί την τεχνική SVD για να μειώσει τις διαστάσεις του πίνακα βαθμολογιών του χρήστη, με τον παραγόμενο πίνακα να δίνεται ως είσοδος σε ένα MLP εμπρόσθιας φόρτωσης 2 επιπέδων.

Το σύστημα που προτείνεται στο παρόν κεφάλαιο (ksepRS) φαίνεται να είναι αρκετά εύρωστο και να επιτυγχάνει καλά αποτελέσματα. Όσον αφορά τις μετρικές ακρίβειας της ταξινόμησης είναι καλύτερο, στη μέση περίπτωση, από τα άλλα συστήματα που παρουσιάζονται εδώ. Αυτό αποτελεί ένδειξη ότι η εξαγωγή της λανθάνουσας σημασιολογικής πληροφορίας (μέσω του SVD), η εξομάλυνση που επιτυγχάνεται με την k-διαχωρισιμότητα και η κατασκευαστική αρχιτεκτονική δικτύου ήταν επιτυχημένες για την περίπτωση του `MovieLens 100k`. Επίσης επιτυγχάνεται μια καλή ισορροπία μεταξύ ακρίβειας και ανάκλησης, ένα βασικό προαπαιτούμενο για κάθε σύστημα συστάσεων. Αυτό οφείλεται στο γεγονός ότι η k-διαχωρισιμότητα μπορεί

Πίνακας 4.3: Μέσο απόλυτο σφάλμα στο `MovieLens 100k`

Σύστημα	MAE
Repeated Matrix Factorization [Kleeman et al.]	0,7200
ksepRS	**0,7256**
Collaborative Filtering (Pearson Correlation) [Sarwar et al., 2002]	0,7455
GNNMean [Pucci et al., 2006]	0,7555
FNNMean [Pucci et al., 2006]	0,7586
Iterative SVD [Kleeman et al.]	0,8000
FMMMF [Kleeman et al.]	0,8200
Τυχαία Βαθμολογία (baseline)	1,511

Τα αποτελέσματα είναι ταξινομημένα σε αύξουσα σειρά της τιμής του MAE

να αποκαλύπτει περίπλοκες στατιστικές εξαρτήσεις (θετικές και αρνητικές) σε σύγκριση με τις άλλες μεθόδους. Έτσι, το προτεινόμενο σύστημα δεν χρειάζεται να φιλτράρει παραπάνω την «γειτονιά» των «όμοιων» χρηστών (με τη χρήση μεθόδων όπως ο συντελεστής συσχέτισης Pearson). Ως αποτέλεσμα, όλοι οι γείτονες του υπό εξέταση χρήστη λαμβάνονται υπόψη, κάτι που είναι εξαιρετικά χρήσιμο σε εκείνες τις περιπτώσεις όπου τόσο οι αξιολογήσεις όσο και οι αξιολογητές είναι λίγοι.

Επιπρόσθετα, στον Πίνακα 4.3, η απόδοση του συστήματος συγκρίνεται με άλλες υλοποιήσεις στη βάση του μέσου απόλυτου σφάλματος που είναι η μετρική ακρίβειας της πρόβλεψης που έχει χρησιμοποιηθεί περισσότερο στη βιβλιογραφία για αυτή τη συλλογή δεδομένων. Η τρίτη γραμμή του πίνακα περιγράφει ένα συνεργατικό σύστημα συστάσεων το οποίο πρώτα χρησιμοποιεί μια τεχνική συσταδοποίησης για να ορίσει την «γειτονιά» των χρηστών και κατόπιν χρησιμοποιεί αυτές τις γειτονιές σε συνδυασμό με τον συντελεστή συσχέτισης Pearson για την πρόβλεψη των βαθμολογιών [Sarwar et al., 2002]. Το σύστημα GNNMean αποτελεί εφαρμογή ενός νευρωνικού δικτύου βασισμένο σε γράφους (graph neural network) [Pucci et al., 2006], ενώ το FNNMean αποτελείται από ένα MLP 2 επιπέδων με 10 νευρώνες με συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης στο κρυφό επίπεδο και μια γραμμική συνάρτηση ενεργοποίησης στο επίπεδο εξόδου. Το τελευταίο σύστημα απλά παράγει μια τυχαία μεταβλητή στο εύρος 1 ως 5 και παίζει το ρόλο του βασικού (baseline) συστήματος· σκοπός του είναι να αποτελέσει πεδίο σύγκρισης έτσι ώστε να εκτιμηθεί καλύτερα η σχετική βελτίωση των υπολοίπων συστημάτων ως προς αυτό.

Τα υπόλοιπα συστήματα του Πίνακα 4.3 βασίζονται σε τεχνικές παραγοντοποίησης πινάκων [Kleeman et al.], που έχουν στόχο να μειώσουν τις διαστάσεις του πίνακα αξιολογήσεων και να προβλέψουν τις τιμές του πίνακα που «λείπουν» (ή, διαφορετικά, να παράξουν συστάσεις). Το FMMMF επιτυγχάνει την παραγοντοποίηση του πίνακα χρησιμοποιώντας τον κανόνα του Frobenius. Η Iterative SVD μέθοδος υπολογίζει το SVD του πίνακα βαθμολογιών και κατόπιν τον ενημερώνει κάθε φορά που προστίθεται μια καινούργια βαθμολογία. Η Repeated Matrix Factorization τεχνική αποτελεί επέκταση της προηγούμενης μεθόδου και η οποία προσπαθεί να διασφαλίσει την διατήρηση των βαθμολογιών κατά τη διαδικασία της παραγοντοποίησης.

Παρότι η μέθοδος Repeated Matrix Factorization παρουσιάζει οριακά καλύτερη απόδοση, το ksepRS ωστόσο βρίσκεται πολύ κοντά (η τιμή του MAE είναι ταυτόσημη στα δύο πρώτα δεκαδικά ψηφία) και άρα μπορούμε να ισχυριστούμε ότι επιδεικνύει και αυτό ικανοποιητικά αποτελέσματα όσον αφορά το μέσο απόλυτο σφάλμα. Πιο συγκεκριμένα, η δυνατότητα της k-διαχωρισιμότητας να ανακαλύπτει περισσότερο περίπλοκες εξαρτήσεις στη βαθμολογική συμπεριφορά των χρηστών είναι φανερή, στη γενική περίπτωση, όταν συγκρίνεται με τη βασική προσέγγιση ή με τις άλλες μεθόδους που χρησιμοποιούν απλές στατιστικές προσεγγίσεις όπως η απλή συνεργατική διήθηση (τρίτη γραμμή του Πίνακα 4.3). Επιπλέον, η χρησιμοποίηση ενός κατασκευαστικού αλγορίθμου δικτύου φαίνεται ότι αποτελεί ένα ακόμα πλεονέκτημα,

Πίνακας 4.4: Ρίζα μέσου τετραγωνικού σφάλματος στα `MovieLens 1M` και `MovieLens 10M`

Συλλογή Δεδομένων Μετρική	1M RMSE	10M RMSE
Collaborative Editing based Domain and Item Transfer [Qian Xu and Yang, 2011]	-	0,8440
Mixed Membership Matrix Factorization [Mackey et al., 2010]	0,8577	0,8447
Bayesian Probabilistic Matrix Factorization [Mackey et al., 2010]	0,8609	0,8472
Latent Factorization Model[Qian Xu and Yang, 2011]	-	0,8580
ksepRS	**0,8889**	**0,8668**
Content Based Algorithm[Clemente, 2008]	0,8901	-
SVD - Item Based Collaborative Filtering [Desrosiers and Karypis, 2010]	0,9650	-
Τυχαία Βαθμολογία (baseline)	1,8931	1,8405

Τα αποτελέσματα είναι ταξινομημένα σε φθίνουσα σειρά της τιμής του RMSE και για τις δύο συλλογές δεδομένων

σε σύγκριση με τα συστήματα που χρησιμοποιούν σταθερές αρχιτεκτονικές (FNNMean και GNNMean). Επίσης, η αρχιτεκτονική «πολλά-προς-ένα» του συστήματος που προτείνεται βρίσκει καλή εφαρμογή στο πεδίο της παραγωγής συστάσεων· αντί να λειτουργεί ως ένα μοντέλο παλινδρόμησης (όπως οι τεχνικές παραγοντοποίησης των πινάκων που προσπαθούν να προσεγγίσουν τις τιμές που λείπουν από τον πίνακα των αξιολογήσεων), προσπαθεί να ταξινομήσει τις ταινίες σε σύνολα ανάλογα με το αν αρέσουν οι όχι. Συνεπώς, η προαναφερόμενη αρχιτεκτονική δεν παρουσιάζει μόνο καλή απόδοση αλλά παρέχει στο σύστημα και την κατάλληλη ευρωστία.

Τέλος, ο Πίνακας 4.4 περιέχει τα αποτελέσματα για το RMSE στις δύο μεγαλύτερες συλλογές δεδομένων MovieLens καθώς και συγκρίσεις με άλλες υλοποιήσεις. Στις πολύ μεγάλες συλλογές δεδομένων και ειδικότερα στο `MovieLens 10M`, το μέσο τετραγωνικό σφάλμα χρησιμοποιείται κατ΄ αποκλειστικότητα για τη μέτρηση της απόδοσης και για αυτό το λόγο εμφανίζονται αποτελέσματα μόνο για αυτό στον προαναφερόμενο πίνακα. Σε αυτή την περίπτωση, οι τεχνικές παραγοντοποίησης πινάκων [Qian Xu and Yang, 2011; Mackey et al., 2010] εμφανίζουν ένα μικρό, αλλά σταθερό, προβάδισμα σε σύγκριση με το ksepRS. Παρόλα αυτά, η προτεινόμενη μεθοδολογία καταφέρνει να είναι περισσότερο ανταγωνιστική σε σύγκριση με τη βασική μέθοδο ή άλλες γενικές μεθόδους συνεργατικής διήθησης, απλές [Desrosiers and Karypis, 2010] και βασιζόμενες στα αντικείμενα [Clemente, 2008], οι οποίες φαίνεται ότι δεν μπορούν να επεξεργαστούν την εγγενή μη-γραμμικότητα του πίνακα αξιολογήσεων των χρηστών.

Αξίζει επίσης να τονιστεί ότι οι τεχνικές παραγοντοποίησης πινάκων έχουν ήδη χρησιμοποιηθεί ευρύτατα στο πεδίο των συστημάτων συστάσεων, πράγμα που είχε ως αποτέλεσμα την εμφάνιση αρκετών βέλτιστων αρχιτεκτονικών. Από την άλλη, η μεθοδολογία της k-διαχωρισιμότητας αποτελεί πρωτότυπη προσέγγιση για τα συστήματα συστάσεων οπότε υπάρχουν ακόμα περιθώρια σημαντικών βελτιώσεων.

4.6.2 Μεταβολή του MAE σε σύγκριση με την Αραιότητα του Πίνακα Βαθμολογιών

Ένας από τους στόχους του παρόντος συστήματος ήταν να δοκιμαστεί σε διάφορες περιπτώσεις αραιότητας των δεδομένων του πίνακα βαθμολογιών. Για το σκοπό αυτό, αφαιρέθη-

Πίνακας 4.5: Εναλλακτικές διαμορφώσεις για τα πειράματα σύγκρισης του MAE ως προς την αραιότητα των αξιολογήσεων

Στοιχείο Συστήματος	Βασική διαμόρφωση	Εναλλακτική διαμόρφωση
Προεπεξεργασία δεδομένων	SVD	Όχι
Αρχιτεκτονική δικτύου	Κατασκευαστική	Σταθερή*
Συναρτήσεις ενεργοποίησης Νευρώνων κρυφού επιπέδου	k-διαχωρισιμότητα	Σιγμοειδής

Η σταθερή αρχιτεκτονική περιέχει τον ίδιο αριθμό νευρώνων με την κατασκευαστική, με τη διαφορά ότι εκπαιδεύονται όλοι μαζί και όχι ένας-ένας

καν βαθμολογίες από τη συλλογή δεδομένων `MovieLens 100k` μέχρις ότου να επιτευχθεί το επιθυμητό επίπεδο αραιότητας. Ξεκινώντας από το αρχικό επίπεδο αραιότητας (93, 7%), δημιουργήθηκαν τα επίπεδα αραιότητας που απεικονίζονται στην Εικόνα 4.6 και μέχρι το 99%. Σε κάθε ένα από αυτά τα επίπεδα εφαρμόστηκε το πειραματικό πρωτόκολλο (Ενότητα 4.5.2), έχοντας όμως μεταβάλει τη διαμόρφωση του συστήματος έτσι ώστε να τονιστεί η ξεχωριστή συνεισφορά του κάθε μέρους του.

Ο Πίνακας 4.5 συνοψίζει τις 8 εναλλακτικές διαμορφώσεις του συστήματος (με την βασική να αντιστοιχεί στην πρώτη στήλη) . Συμπεριλήφθηκε στα πειράματα, ως υλοποίηση αναφοράς, ο κλασσικός αλγόριθμος συνεργατικής διήθησης [Billsus and Pazzani, 1998], ο οποίος χρησιμοποιεί τον συντελεστή συσχέτισης Pearson (Σχήμα 4.6)

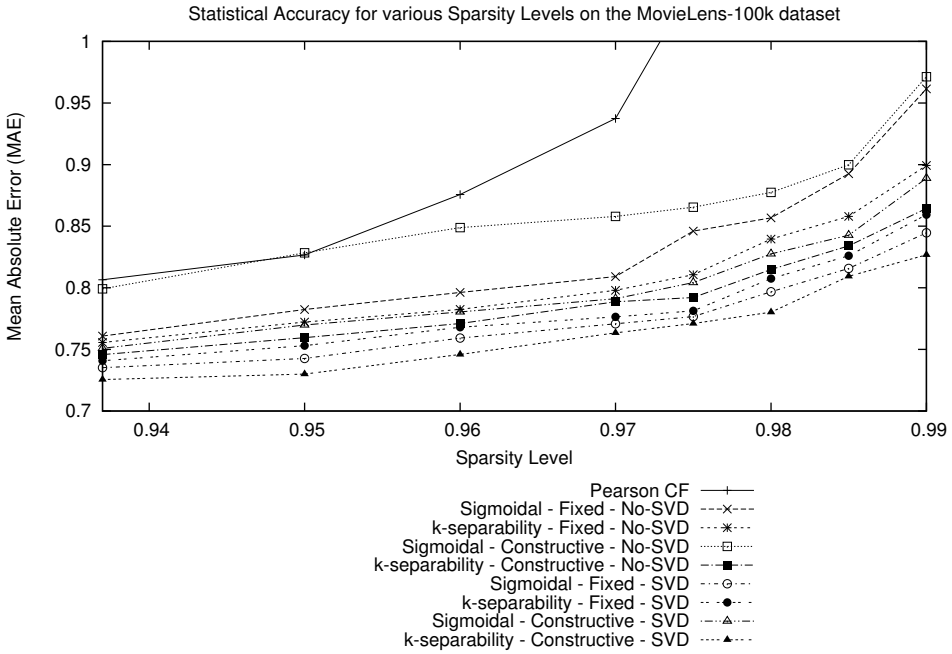
Μια πρώτη παρατήρηση είναι ότι η βασική μέθοδος εμφανίζει τη χειρότερη απόδοση σε όλα τα επίπεδα αραιότητας (και ειδικά στις περιπτώσεις της πολύ μεγάλης αραιότητας) κάνοντας εμφανείς, με αυτόν τον τρόπο, τους περιορισμούς στη χρήση της. Ο συνδυασμός της κατασκευαστικής αρχιτεκτονικής με τις σιγμοειδείς συναρτήσεις ενεργοποίησης στο κρυφό επίπεδο παρουσιάζει οριακά καλύτερη απόδοση σε αντίθεση με τη σταθερή αρχιτεκτονική, όπως είναι εμφανές από την βελτίωση της απόδοσης όταν εφαρμόζεται η δεύτερη διαμόρφωση. Η απόκριση του συστήματος βελτιώνεται περαιτέρω με την χρήση της k-διαχωρισιμότητας στο κρυφό επίπεδο ενώ τα αποτελέσματα γίνονται ακόμα καλύτερα όταν ενεργοποιηθούν και οι υπόλοιπες παράμετροι. Η τελική αρχιτεκτονική του συστήματος δίνει το καλύτερο αποτέλεσμα για κάθε επίπεδο αραιότητας.

4.6.3 Υπολογιστικό Κόστος

Ένας τρόπος υπολογισμού του *Υπολογιστικού Κόστους* είναι η παρουσίαση μιας ανάλυσης που θα βασίζεται στην *θεωρία της πολυπλοκότητας*. Στόχος αυτής της ανάλυσης είναι να εκτιμηθεί η ξεχωριστή συνεισφορά των κυριότερων μερών του συστήματος στο συνολικό κόστος καθώς και να επισημανθούν τα πιο απαιτητικά του μέρη. Στο παρόν σύστημα, η υπολογιστική πολυπλοκότητα προσεγγίζεται όπως παρακάτω:

1. Ανάλυση Ιδιαζουσών Τιμών [Knockaert et al., 1999]: $O(n^3)$, n είναι η μεγαλύτερη διάσταση του πίνακα βαθμολογιών των χρηστών.

2. Αλγόριθμος CNNA: Ο CNNA χρησιμοποιεί τεχνική οπίσθιας διάδοσης του σφάλματος, η οποία απαιτεί $O(W)$ χρόνο για κάθε δείγμα εκπαίδευσης [Haykin, 2008], όπου W είναι ο αριθμός των βαρών στο δίκτυο. Αν T_s είναι το μέγεθος του συνόλου εκπαίδευσης και τ ο αριθμός των εποχών (εκπαίδευσης), η εκπαίδευση απαιτεί $O(\tau \times T_s \times W)$ χρόνο. Μπορεί εύκολα να δειχθεί ότι όλες οι υπόλοιπες λειτουργίες του αλγορίθμου (πχ προσθήκη καινούργιων νευρώνων, υπολογισμός των κριτηρίων τερματισμού και αρχιτεκτονικής προσαρμογής) απαιτούν λιγότερο χρόνο, οπότε η εκπαίδευση του δικτύου είναι το κυρίαρχο υπολογιστικά τμήμα του αλγορίθμου.



Σχήμα 4.6: Επίπεδα αραιότητας στο `MovieLens 100k`

Πίνακας 4.6: Μέσος χρόνος (σε sec) που απαιτείται για την εκπαίδευση ενός μοντέλου για τον χρήστη

Συλλογή Δεδομένων	`Movielens 100k`	`Movielens 1M`	`Movielens 10M`
Χρόνος	30.88	43.53	64.25

Σύστημα: Intel Core 2 Duo E8400 3.00GHz, 3.2 GB RAM, 32-bit Linux

3. Ενισχυτική Μάθηση [Freund and Schapire, 1996]: $O(T_s)$ χρόνος προκειμένου να ενημερωθεί η κατανομή των δειγμάτων του συνόλου εκπαίδευσης.

Η παραπάνω ανάλυση καταδεικνύει τον CNNA ως το κυρίαρχο υπολογιστικά τμήμα του προτεινόμενου συστήματος συστάσεων (Αλγόριθμος 1), ο οποίος απαιτεί χρόνο $O(\tau \times T_s \times W)$. Συνεπώς πρόκειται για *ψευδο-πολυωνυμικό* αλγόριθμο, γιατί ο χρόνος εκτέλεσης του είναι πολυωνυμικός όχι μόνο ως προς το εύρος της εισόδου του αλλά επίσης και προς την τιμή της.

Στον Πίνακα 4.6 παρουσιάζεται ο μέσος χρόνος εκπαίδευσης του μοντέλου για έναν χρήστη και για τις τρεις συλλογές δεδομένων MovieLens. Παρότι μόλις προηγουμένως ο αλγόριθμος του συστήματος συστάσεων παρουσιάστηκε ότι είναι ψευδο-πολυωνυμικός, εντούτοις εμφανίζει μια γραμμική (πολυωνυμική) αύξηση του χρόνου. Αυτή η συμπεριφορά αποδίδεται στο γεγονός ότι ενώ το μέγεθος των συλλογών δεδομένων αυξάνει πάνω από μια τάξη μεγέθους κάθε φορά, ωστόσο η πυκνότητά τους μειώνεται δραστικά, αφήνοντας κατ΄ αυτόν τον τρόπο χώρο για μια «πολυωνυμική» συμπεριφορά, με αποτέλεσμα το προτεινόμενο σύστημα να παραμένει αποτελεσματικό ακόμα και στις μεγάλες συλλογές δεδομένων.

4.7 Συμπεράσματα και Μελλοντικές Ερευνητικές Κατευθύνσεις

Σε αυτό το κεφάλαιο παρουσιάστηκε ένα καινούργιο σύστημα συστάσεων που προσπαθεί να αντιμετωπίζει ένα από τα κύρια προβλήματα των αλγορίθμων συνεργατικής διήθησης, αυτό της αραιότητας του πίνακα αξιολογήσεων. Αρχικά αυξάνεται η πυκνότητα του εν λόγω πίνακα μέσω της ανάλυσης ιδιαζουσών τιμών, μια τεχνικής μείωσης των διαστάσεων του πίνακα η

οποία εμφανίζει ικανοποιητικά αποτελέσματα. Ο πίνακας αξιολογήσεων ναι μεν γίνεται πιο συμπαγής και πυκνός, αλλά έχει ως επακόλουθο την μείωση της ποιότητας της περιεχόμενης πληροφορίας.

Η παρενέργεια αυτή αντιμετωπίστηκε με την διαδοχική εφαρμογή μιας πληθώρας τεχνικών. Το θεωρητικό μοντέλο της k-διαχωρισιμότητας επέτρεψε στο προτεινόμενο σύστημα να «αφομοιώνει» μη-γραμμικά διαχωρίσιμα δεδομένα σε σημαντικό βαθμό. Ο κατασκευαστικός αλγόριθμος δικτύου μαζί με την ενισχυτική τεχνική εκπαίδευσης συνέβαλαν στην ικανότητα γενίκευσης του συστήματος, ακόμα και σε εκείνες τις περιπτώσεις όπου οι βαθμολογίες ήταν πάρα πολύ λίγες. Αυτή η συμπεριφορά είναι εμφανέστερη στα πειράματα με τις συλλογές δεδομένων MovieLens, ειδικά όταν συγκρίνεται με τα αποτελέσματα άλλων υλοποιήσεων.

Λόγω των επιλεγμένων μεθοδολογιών, το προτεινόμενο σύστημα επίσης επέδειξε καλύτερα αποτελέσματα όσον αφορά την ακρίβεια, την ανάκληση και το μέγεθος F1 σε σύγκριση με άλλους ανταγωνιστικούς αλγορίθμους. Επιπρόσθετα καθώς ήταν στόχος το να ξεπεραστεί το εγγενές πρόβλημα της αραιότητας των αξιολογήσεων στα συστήματα συστάσεων, μελετήθηκε η προσαρμοστικότητα του συστήματος με πολύ περιορισμένα δεδομένα προερχόμενα από τις μεγάλες συλλογές του MovieLens. Παρότι το προτεινόμενο σύστημα εμφάνισε μια μικρή πτώση στην απόδοσή του, παρέμεινε καλύτερο από πολλά άλλα συστήματα, όσον αφορά το μέσο τετραγωνικό σφάλμα και φαίνεται να έχει τη δυνατότητα του χειρισμού πολύ αραιών δεδομένων.

Γενικότερα υπάρχουν περιθώρια βελτίωσης της παρούσας μεθοδολογίας, ιδιαίτερα όσον αφορά τον κατασκευαστικό αλγόριθμο δικτύου. Μια πιθανή προσέγγιση θα ήταν να ακολουθηθεί η μέθοδος των βαρυκεντρικών κατασκευαστικών αλγορίθμων (barycentric-based contructive neural networks) [Bertini and do Carmo Nicoletti, 2008], η οποία επίσης εμφανίζει καλά αποτελέσματα. Ακόμα θα ήταν επιθυμητός ο εμπλουτισμός του συστήματος με πληροφορίες ανάδρασης από τους χρήστες (user feedback). Ήδη έχουν δημοσιευτεί εργασίες προς αυτή την κατεύθυνση [Zanker and Jessenitschnig, 2009], οι οποίες θα μπορούσαν να ενσωματωθούν στο παρόν σύστημα.

□

Κεφάλαιο 5

Μοντελοποίηση της Άμεσης Κοινωνικής Πληροφορίας στα Συστήματα Συστάσεων

Από την μέχρι στιγμής μελέτη των συστημάτων συνεργατικής διήθησης προκύπτει το συμπέρασμα ότι οι συστάσεις δεν παράγονται «απομονωμένα», αλλά αντίθετα προκύπτουν από μια ανεπίσημη κοινότητα χρηστών (οι όμοιοι χρήστες) ή αντικειμένων. Είναι εμφανές ότι αυτή η λειτουργία έχει και κοινωνικές προεκτάσεις, μιας και μπορεί να θεωρηθεί ότι η παραγωγή των προτάσεων ισοδυναμεί με την δημιουργία άτυπων δεσμών μεταξύ των χρηστών, οπότε αποκτά ιδιαίτερο ενδιαφέρον η μελέτη των συστημάτων συστάσεων υπό την συγκεκριμένη οπτική. Εξάλλου, τα συστήματα συστάσεων εμπεριέχουν μια εγγενώς κοινωνική διάσταση, η οποία εξαρτάται από τον τρόπο με τον οποίο μοντελοποιούν τους χρήστες τους: μπορεί να τους συνδέουν έμμεσα, με την χρήση μεθόδων υπολογισμού ομοιότητας όπως αναλύθηκε στα προηγούμενα κεφάλαια της διατριβής ή μπορούν να εκμεταλλευτούν άμεσα τις σχέσεις που αυτοί δηλώνουν ότι έχουν μεταξύ τους.

Τα τελευταία χρόνια έχουν συντελεστεί σημαντικές αλλαγές που επιβάλλουν την πιο ενδελεχή εξέταση αυτής της δεύτερης περίπτωσης. Πιο συγκεκριμένα, η εκρηκτική αύξηση του περιεχομένου που δημιουργούν οι χρήστες στον Παγκόσμιο Ιστό μέσω των τεχνολογιών Web 2.0 όπως είναι, για παράδειγμα, τα κοινωνικά δίκτυα, τα ιστολόγια και τα διάφορα διαδικτυακά forum, μπορεί να φανεί χρήσιμη σε πολλές εφαρμογές. Στο πεδίο των συστημάτων συστάσεων, η κατάλληλη επεξεργασία του περιεχομένου που δημιουργείται από τους χρήστες με την χρήση μεθόδων *εξόρυξης δεδομένων* (data mining) μπορεί να οδηγήσει στον εμπλουτισμό του προφίλ τους, το οποίο πλέον δεν περιορίζεται μόνο σε αξιολογήσεις πάνω σε αντικείμενα. Αντίθετα, μπορεί να γίνει πιο ακριβές και πιο πολύπλοκο ενσωματώνοντας πληροφορίες σχετικά με τα ενδιαφέροντά τους ή τις σχέσεις τους με άλλους χρήστες. Η αλλαγή αυτή, όπως είναι προφανές, μπορεί να οδηγήσει σε περισσότερο αποδοτικούς αλγορίθμους παραγωγής συστάσεων.

Από τις υπάρχουσες Web 2.0 τεχνολογίες, αυτή που παρουσιάζει μεγαλύτερο ενδιαφέρον είναι τα κοινωνικά δίκτυα, κυρίως λόγω της ευρύτατης τους διάδοσης (π.χ. τα μέλη του Facebook είναι πολλαπλάσια σε σύγκριση με όσους διατηρούν προσωπικά ιστολόγια ή αρθρογραφούν σε διαδικτυακά forum). Ένα επιπλέον χαρακτηριστικό είναι ο πολύ μεγάλος βαθμός ελευθερίας που δίνουν στους χρήστες τους, κυρίως μέσω των ενσωματωμένων εφαρμογών τους. Για παράδειγμα, το Facebook δίνει τη δυνατότητα στα μέλη του να αξιολογούν ταινίες και κατόπιν τους προτείνει παρεμφερείς ταινίες που δεν έχουν αξιολογήσει (Σχήμα 5.1).

Ο τύπος και ο τρόπος αξιοποίησης της κοινωνικής πληροφορίας θα αποτελέσουν το αντικείμενο του παρόντος κεφαλαίου.

Σχήμα 5.1: Προτάσεις ταινιών από το Facebook

5.1 Κοινωνικά Δίκτυα

Ένα *κοινωνικό δίκτυο* (social network) αποτελεί μια καθορισμένη αναπαράσταση μιας συγκεκριμένης όψης του πραγματικού κόσμου και των εξαρτήσεων που τη διέπουν. Δομικά του μέρη αποτελούν οι κόμβοι μαζί με τις σχέσεις που τους συνδέουν. Οι κόμβοι αντιπροσωπεύουν συνηθέστερα οντότητες όπως άτομα ή οργανισμούς, χωρίς όμως κατ' ανάγκη να περιορίζονται εκεί. Κόμβοι ενός κοινωνικού δικτύου θα μπορούσαν επίσης να είναι ιστοσελίδες [Watts, 2003], άρθρα σε επιστημονικά περιοδικά [White et al., 2004], τμήματα οργανισμών [Quan-Haase and Wellman, 2006] κ.ά. Συνεπώς, ένα από τα πρώτα ζητήματα που ανακύπτουν

στην μελέτη των κοινωνικών δικτύων είναι ο καθορισμός του είδους των κόμβων που θεωρούνται έγκυροι. Για αυτό το ζήτημα υπάρχουν τρεις διαφορετικές προσεγγίσεις [Scott and Carrington, 2011]. Η πρώτη βασίζεται στη *θέση* της κάθε οντότητας (position-based approach) και θεωρεί ως μέλη του δικτύου όλες εκείνες τις οντότητες που αποτελούν είτε μέλη ενός οργανισμού ή κατέχουν κάποιες συγκεκριμένες και σαφώς καθορισμένες θέσεις σε αυτόν (λ.χ. φοιτητές/καθηγητές ενός πανεπιστημίου) ενώ ταυτόχρονα αποκλείει οποιονδήποτε άλλο. Η δεύτερη προσέγγιση είναι *προσανατολισμένη στα γεγονότα* (event-based approach) και θεωρεί ως μέλη του κοινωνικού δικτύου όλους εκείνους που έχουν λάβει μέρος σε ένα γεγονός (λ.χ οι συμμετέχοντες σε ένα επιστημονικό συνέδριο). Τέλος, η τρίτη προσέγγιση έχει ως αφετηρία τις *σχέσεις* μεταξύ των μελών (relation-based approach) και ξεκινά με ένα μικρό αριθμό κόμβων που προέρχεται από ένα πληθυσμό που παρουσιάζει ενδιαφέρον και κατόπιν επεκτείνεται για να συμπεριλάβει όλους εκείνους που μοιράζονται συγκεκριμένου τύπου σχέσεις με τον *αρχικό πληθυσμό* (seed population). Οι τρεις προαναφερόμενες προσεγγίσεις δεν είναι κατ΄ ανάγκη αμοιβαία αποκλειόμενες ενώ συνήθως στην πράξη χρησιμοποιούνται συνδυασμοί τους.

Μετά τον καθορισμό των μελών του δικτύου, πρέπει σε ένα δεύτερο επίπεδο να καθοριστούν οι σχέσεις μεταξύ τους. Παραδείγματα σχέσεων αποτελούν οι συνεργασίες, οι φιλίες, οι εμπορικοί δεσμοί, οι δεσμοί ανάμεσα σε ιστοσελίδες ή οποιαδήποτε άλλη πιθανή σχέση ανάμεσα στις συγκεκριμένες οντότητες. Σε αυτό το σημείο, αξίζει να επισημανθεί ότι ακόμα και στο πλαίσιο της ίδιας ομάδας μπορούν να αναπτυχθούν σχέσεις που δεν περιορίζονται αναγκαστικά σε έναν τύπο. Γενικότερα, οι σχέσεις μπορούν να ενταχθούν σε τέσσερις ευρείες κατηγορίες: τις *ομοιότητες*, τις *κοινωνικές σχέσεις*, τις *αλληλεπιδράσεις* και τις *ροές* [Scott and Carrington, 2011]. Η ομοιότητα αφορά την περίπτωση που δύο κόμβοι μοιράζονται ιδιότητες που μπορούν να μελετηθούν υπό το πρίσμα προσεγγίσεων όπως είναι τα δημογραφικά χαρακτηριστικά, η τοποθεσία, η συμμετοχή σε ομάδες κλπ. Οι κοινωνικές σχέσεις περιλαμβάνουν τις γνωριμίες καθώς και άλλους διαδεδομένους τύπους σχέσεων (λ.χ. φιλία) ή δεσμούς που βασίζονται στο πως αντιλαμβάνεται το κάθε μέλος του δικτύου τα άλλα (π.χ εμπιστοσύνη). Οι αλληλεπιδράσεις αναφέρονται σε συμπεριφοριστικές σχέσεις όπως λ.χ. η κοινωνική συναναστροφή, η βοήθεια και η συζήτηση και λαμβάνουν χώρα στο πλαίσιο άλλων κοινωνικών σχέσεων και αλληλεπιδράσεων Τέλος, οι ροές είναι σχέσεις που βασίζονται στις ανταλλαγές ή τις μεταφορές μεταξύ των κόμβων. Αυτές μπορεί να περιλαμβάνουν σχέσεις στις οποίες πόροι, πληροφορίες (ή γενικότερα επιρροή) διαχέονται μέσω του δικτύου. Όπως και στην περίπτωση των αλληλεπιδράσεων, οι ροές λαμβάνουν χώρα στο πλαίσιο άλλων τύπων κοινωνικών σχέσεων. Η ισχύς μιας σχέσης στο κοινωνικό δίκτυο μπορεί να μεταβάλλεται από αδύναμη μέχρι ισχυρή, ανάλογα με την ποσότητα, την ποιότητα και τη συχνότητα των συναλλαγών ανάμεσα στους συμμετέχοντες. Τέλος πρέπει να σημειωθεί ότι τα κοινωνικά δίκτυα είναι δυναμικά καθώς, κάθε στιγμή δημιουργούνται καινούργιες σχέσεις ανάμεσα στους συμμετέχοντες ή καταργούνται ήδη υπάρχουσες.

Η μελέτη των κοινωνικών δικτύων αποτελεί αυτοτελή επιστημονική περιοχή που ονομάζεται *ανάλυση κοινωνικών δικτύων* (social network analysis - SNA) και η οποία παρέχει τα θεωρητικά εκείνα εργαλεία που είναι απαραίτητα για την έρευνα τους.

5.2 Ανάλυση Κοινωνικών Δικτύων

Η ανάλυση των κοινωνικών δικτύων είναι μια διεπιστημονική ερευνητική περιοχή που έχει ανακύψει τον τελευταίο μισό αιώνα ως συμπληρωματικό πεδίο των κοινωνικών επιστημών [Wasserman and Faust, 1994]. Στη βάση της θεωρεί ότι η ερμηνεία της οργάνωσης της καθημερινότητας δεν οφείλεται σε εγγενείς παρορμήσεις ή σε αόριστες κατευθύνσεις του κάθε ενός ατόμου χωριστά. Αντίθετα, εξηγείται με την ανάλυση της δομής των σχέσεων αλληλεπίδρασης μεταξύ των μελών της κοινωνίας, ταυτόχρονα με τη μελέτη της συμπεριφοράς των παραγόντων που αναπαράγουν και μεταβάλουν αυτές τις δομές. Έχοντας ήδη χρησιμοποιηθεί

σε μια ευρεία γκάμα πεδίων, η SNA βρίσκει ιδιαίτερο πεδίο εφαρμογής στην κατανόηση των σχέσεων μεταξύ των μελών των online κοινοτήτων. Οι κυριότεροι λόγοι που συνεισφέρουν σε αυτή την κατεύθυνση είναι τόσο η φύση της ίδιας της αλληλεπίδρασης όσο και της ψηφιακής πληροφορίας.

5.2.1 Δομικά Στοιχεία Κοινωνικών Δικτύων

Οι [Wasserman and Faust, 1994] συνοψίζουν τα δομικά στοιχεία των κοινωνικών δικτύων, όπως αυτά έχουν γίνει αποδεκτά από την επιστημονική κοινότητα. Παρακάτω συνοψίζονται εκείνα τα στοιχεία των κοινωνικών δικτύων που απατώνται συχνότερα στα κοινωνικά συστήματα συστάσεων.

Δράστες

Όπως έχει ήδη αναφερθεί, η SNA έχει ως αντικείμενο την κατανόηση των σχέσεων ανάμεσα στις οντότητες καθώς και τις συνέπειες που απορρέουν από αυτές. Οι κοινωνικές οντότητες αναφέρονται συνηθέστερα με τον όρο *δράστες* (actors). Οι δράστες μπορεί να είναι διακριτά άτομα, επιχειρήσεις ή άλλου είδους συλλογικές κοινωνικές μονάδες. Συνήθως, τα περισσότερα κοινωνικά δίκτυα εστιάζουν σε δράστες που είναι όλοι του ιδίου τύπου (πχ άνθρωποι) για αυτό καλούνται εναλλακτικά και *δίκτυα του ενός τύπου* (one mode networks). Ωστόσο, μια τέτοια λειτουργικότητα δεν είναι υποχρεωτική· υπάρχουν και κοινωνικά δίκτυα που επιτρέπουν δύο ή και περισσότερους τύπους δραστών.

Σχεσιακοί Δεσμοί

Οι δράστες συνδέονται μεταξύ τους με *σχεσιακούς δεσμούς* (relational ties), το εύρος και ο τύπος των οποίων μπορεί να είναι αρκετά εκτεταμένο. Το κύριο χαρακτηριστικό ενος δεσμού είναι ότι ορίζει μια σχέση μεταξύ ενός ζεύγους δραστών. Οι πιο διαδεδομένοι τύποι δεσμών που χρησιμοποιούνται στην SNA είναι η αξιολόγηση ενός προσώπου από ένα άλλο (για παράδειγμα φιλία, συμπάθεια ή εμπιστοσύνη) καθώς και η αλληλεπίδραση μεταξύ προσώπων (συζήτηση, αποστολή μηνυμάτων).

Δυάδες

Στην πιο βασική της μορφή, μια σχέση ορίζει ένα δεσμό μεταξύ δύο δραστών. Ο δεσμός αυτός αποτελεί μια εγγενή ιδιότητα και των δύο δραστών και συνεπώς δεν θεωρείται ότι αφορά μόνο τον ένα από αυτούς. Πολλά είδη της ανάλυσης δικτύων ασχολούνται με την κατανόηση των δεσμών ανάμεσα σε ζεύγη και θεωρούν την δυάδα τη βάση της ανάλυσής τους. Μια *δυάδα* (dyad)) αποτελείται από το ζεύγος των δραστών και όλους τους πιθανούς δεσμούς μεταξύ τους. Οι δυαδικές αναλύσεις εστιάζουν στις ιδιότητες των δεσμών όπως το κατά πόσο αυτοί είναι αμοιβαίοι ή όχι, ή αν συγκεκριμένοι τύποι πολλαπλών σχέσεων τείνουν να εμφανίζονται μαζί.

Τριάδες

Αντικείμενο μελέτης αποτελούν και οι σχέσεις μεταξύ μεγαλύτερων υποσυνόλων δραστών. Πολλές και σημαντικές μέθοδοι των κοινωνικών δικτύων εστιάζουν στην *τριάδα* (triad))· ένα υποσύνολο τριών δραστών και των πιθανών δεσμών μεταξύ τους. Για

παράδειγμα, ιδιαίτερο ενδιαφέρον παρουσιάζουν η *μεταβατική ιδιότητα* (αν ο δράστης i σχετίζεται με τον j και ο j με τη σειρά του σχετίζεται με τον k, τότε και ο i σχετίζεται με τον k) καθώς και η *ισορροπία* (balance): αν ο i αποδέχεται τον j και ο j αποδέχεται τον i, τότε οι i, j θα πρέπει να έχουν την ίδια συμπεριφορά προς έναν δράστη k· αντίθετα, αν οι δράστες i, j δεν αποδέχονται ο ένας τον άλλον τότε δεν θα πρέπει υπάρχει δράστης k που να αποτιμάται από τους i, j με τον ίδιο τρόπο.

Ομάδες

Η δύναμη της SNA έγκειται στην ικανότητα της μοντελοποίησης των σχέσεων ανάμεσα σε συστήματα δραστών, το οποίο αποτελείται από δεσμούς μεταξύ των μελών μιας (περισσότερο ή λιγότερο) συνδεδεμένης ομάδας. Ως *ομάδα* (group) ορίζεται μια συλλογή δραστών της οποίας οι δεσμοί μπορούν να καθοριστούν ποσοτικά. Για το λόγο αυτό, η συνοχή της θα πρέπει να μπορεί να υποστηριχθεί με θεωρητικά, εμπειρικά και εννοιολογικά κριτήρια. Με άλλα λόγια, μια ομάδα αποτελείται από ένα πεπερασμένο σύνολο δραστών τα οποία για εννοιολογικούς, θεωρητικούς ή εμπειρικούς λόγους αντιμετωπίζονται ως σύνολο πάνω στο οποίο μπορούν να πραγματοποιηθούν επεξεργασίες. Τέλος, στις περισσότερες περιπτώσεις κοινωνικών δικτύων, παρότι μπορούν να οριστούν πολλές ομάδες μελών, εντούτοις το είδος των δραστών είναι μοναδικό.

Σχέσεις

Η συλλογή των δεσμών συγκεκριμένου είδους μεταξύ των δραστών μιας ομάδας καλείται *σχέση* (relation). Παραδείγματα σχέσεων αποτελούν οι φιλίες που σχηματίζουν οι μαθητές μιας τάξης ή οι επίσημοι διπλωματικοί δεσμοί μεταξύ των κρατών. Για κάθε ομάδα δραστών μπορούν να μετρηθούν διαφορετικοί τύποι σχέσεων (σε συνέχεια του προηγούμενου παραδείγματος, πέρα των διπλωματικών σχέσεων τα κράτη έχουν και εμπορικές σχέσεις μεταξύ τους). Πρέπει επίσης να τονιστεί ότι μια σχέση αναφέρεται σε μια συλλογή δεσμών ενός δεδομένου τύπου μεταξύ ζευγών δραστών του ιδίου είδους.

Κοινωνικό Δίκτυο

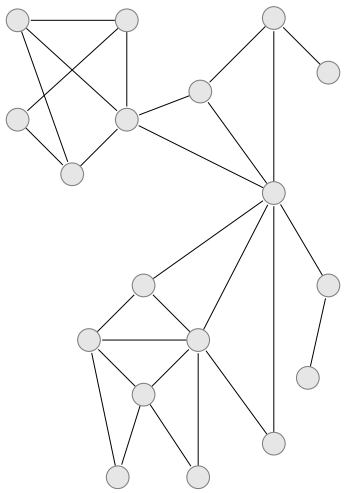
Έχοντας οριστεί τα μέλη, οι ομάδες και οι σχέσεις, μπορεί πλέον να δοθεί ένας πιο άμεσος ορισμός του κοινωνικού δικτύου. Ένα *Κοινωνικό Δίκτυο* αποτελείται από ένα πεπερασμένο σύνολο μελών και των σχέσεων που ορίζονται επάνω τους. Η ύπαρξη πληροφορίας σχετικά με τις σχέσεις των μελών είναι ένα κρίσιμο και καθοριστικό κριτήριο της ύπαρξης των κοινωνικών δικτύων.

5.2.2 Τύποι Δικτύων

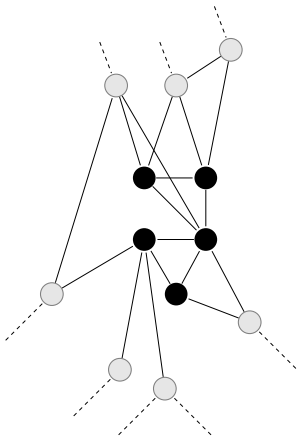
Τα κοινωνικά δίκτυα ταξινομούνται σε κατηγορίες, ανάλογα με τον τρόπο πρόσληψης τους. Διακρίνονται 3 κύριοι τύποι κοινωνικών δικτύων [Hogan, 2008]:

Πλήρη Δίκτυα

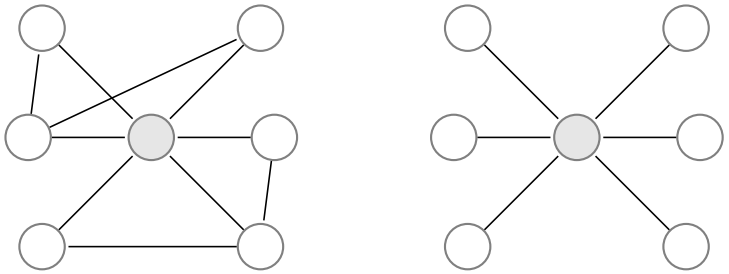
Τα *πλήρη δίκτυα* (whole networks) (Σχήμα 5.2α′) περιγράφουν τις σχέσεις στο σύνολο ενός πληθυσμού. Παραδείγματα τέτοιων δικτύων αποτελούν μια λίστα ηλεκτρονικού ταχυδρομείου ή ένας ολόκληρος ιστότοπος κοινωνικής δικτύωσης. Τα πλήρη δίκτυα χρησιμοποιούνται αρκετά συχνά στην SNA. Η συλλογή όλων των σχέσεων από ένα πλήρες δίκτυο είναι τις περισσότερες φορές αδύνατη τόσο για λόγους που άπτονται του απαιτούμενου χρόνου για μια τέτοια διαδικασία όσο και γιατί αυτό μεταβάλλεται δυναμικά κατά τη διάρκεια της συλλογής. Σε ένα πλήρες δίκτυο, ενδιαφέρουν η δομή



(α′) Πλήρες δίκτυο



(β′) Μερικό δίκτυο



(γ′) Προσωπικό δίκτυο

Σχήμα 5.2: Τύποι δικτύων

του, οι τύποι των μελών του καθώς και τα πιο «προβεβλημένα» μέλη του.

Προσωπικά Δίκτυα

Τα *προσωπικά δίκτυα* (personal networks - PN) προκύπτουν μέσω της δειγματοληψίας από έναν πληθυσμό κόμβων. Κάθε κόμβος-δείγμα αναφέρεται ως *ego* ενώ όλοι οι υπόλοιποι κόμβοι που συνδέονται σε αυτόν αναφέρονται ως *alters* (Σχήμα 5.2γ′). Ενώ στα πλήρη δίκτυα, ο στόχος της μελέτης συχνά είναι η περιγραφή των χαρακτηριστικών του δικτύου και η απάντηση ερωτήσεων σχετικά με το γιατί ορισμένα άτομα καταλαμβάνουν συγκεκριμένες θέσεις στο δίκτυο (λ.χ. ποιος είναι ο αριθμός των υποομάδων στο συγκεκριμένο δίκτυο), η ανάλυση των προσωπικών δικτύων είναι συγκριτική από τη φύση της. Σκοπός σε αυτή την περίπτωση είναι η μελέτη των διαφορών στο μέγεθος, το σχήμα και την ποιότητα μεταξύ ενός αριθμού προσωπικών δικτύων.

Τα προσωπικά δίκτυα κατασκευάζονται με δύο τρόπους. Ο πρώτος περιλαμβάνει τον κόμβο ego και μόνο τους δεσμούς που αυτός έχει προς τους κόμβους alters. Πρόκειται, δηλαδή, για ένα *ακτινικό δίκτυο* (star network). Ο δεύτερος τρόπος επεκτείνει το ακτινικό δίκτυο που αναφέρθηκε προηγουμένως λαμβάνοντας υπόψη και τους δεσμούς που υπάρχουν μεταξύ των alters, οδηγώντας στην κατασκευή ενός πλήρους προσωπικού δικτύου.

Μερικά Δίκτυα

Τα *μερικά δίκτυα* (partial networks) (Σχήμα 5.2β′) είναι το αποτέλεσμα της εφαρ-

μογής δειγματοληπτικών μεθόδων σε πλήρη δίκτυα. Προκύπτουν ως το αποτέλεσμα του συμβιβασμού μεταξύ της ανάγκης για επεξεργασία ενός μεγάλου δικτύου και του γεγονότος ότι κάποια δίκτυα είναι πολύ μεγάλα για να μπορέσουν να επεξεργαστούν. Αφετηρία είναι συνήθως ένας κόμβος ή μια ομάδα κόμβων γνωστή και ως το *σύνολο σποράς* (seed set). Κατόπιν λαμβάνονται υπόψη οι δεσμοί προς τους άμεσα συνδεδεμένους κόμβους, σε δεύτερο επίπεδο οι δεσμοί των άμεσα συνδεδεμένων κόμβων στους δικούς τους άμεσα συνδεδεμένους κόμβους ενώ η επαναληπτική αυτή διαδικασία σταματά όταν συλλεχθεί επαρκής αριθμός κόμβων ή γενικότερα όταν ικανοποιηθούν κάποια συγκεκριμένα κριτήρια τερματισμού.

Τα μερικά δίκτυα αποτελούν μια καλή προσέγγιση για τις μεγάλες ποσότητες δεδομένων που συλλέγονται από το Διαδίκτυο. Για παράδειγμα, δεν είναι εφικτό να συλλεχθούν όλα τα δεδομένα από όλα τα ιστολόγια· από την άλλη όμως, είναι δυνατό να κατασκευαστεί ένα δίκτυο που διασυνδέει τα προσωπικά δίκτυα αρκετών ατόμων.

Πρέπει επίσης να τονιστεί ότι οι παρατηρήσεις που προκύπτουν από τα μερικά δίκτυα δεν μπορούν κατ΄ ανάγκη να γενικευτούν, μιας και ο τρόπος συλλογής των δειγμάτων δεν εγγυάται απαραίτητα την διατήρηση όλων των στατιστικών ιδιοτήτων του αρχικού δικτύου. Ωστόσο, η μειωμένη δυνατότητα στατιστικής γενίκευσης δεν αποτρέπει τις περιγραφικές αναλύσεις και την εξαγωγή συμπερασμάτων από τα δείγματα, επιτρέποντας τις γενικεύσεις σε θεωρητικό επίπεδο.

5.3 Συλλογή των Δεδομένων

Τα κοινωνικά δίκτυα κατασκευάζονται από την ανάλυση και επεξεργασία των δεδομένων, τα οποία συλλέγονται από πολλές και διαφορετικές πηγές. Ενδεικτικά αναφέρονται τα ερωτηματολόγια, οι συνεντεύξεις, η μελέτη αρχειακού υλικού κ.α. [Wasserman and Faust, 1994]. Παρότι οι προαναφερόμενες πηγές έχουν ενσωματωθεί σε συστήματα συστάσεων με στόχο κυρίως την αξιολόγηση της ευχρηστίας τους και την βελτίωση της απόκρισης τους [Shani and Gunawardana, 2011], εντούτοις μια τέτοια αξιοποίηση δεν είναι αρκετή για τον χαρακτηρισμό τους ως *κοινωνικά συστήματα συστάσεων* (social recommender systems - SRS). Είναι προφανές ότι για να έχει βάση ο εν λόγω χαρακτηρισμός, η αξιοποίηση της κοινωνικής πληροφορίας επιβάλλεται να αποτελεί δομικό στοιχείο της λειτουργίας τους, το οποίο εκτείνεται καθολικά σε όλους τους χρήστες (και όχι μόνο σε αυτούς που έχουν απαντήσει σε κάποιο ερωτηματολόγιο).

Οι πηγές δεδομένων για τα κοινωνικά δίκτυα, οι οποίες είναι διαθέσιμες σε μεγάλη κλίμακα και οι οποίες επιπρόσθετα είναι εύκολο να συλλεχθούν, είναι οι *διαδικτυακές πηγές δεδομένων κοινωνικών δικτύων* (online sources of social network data). Ο [Hogan, 2008] τις ταξινομεί σε τρεις κύριες κατηγορίες: στα *αρχεία καταγραφής ηλεκτρονικού ταχυδρομείου*, στα *ιστολόγια* (και λοιπές αντίστοιχες ιστοσελίδες) και τέλος στα *μέσα κοινωνικής δικτύωσης*. Παράδειγμα συστήματος που χρησιμοποιεί αρχεία καταγραφής ηλεκτρονικού ταχυδρομείου αποτελεί η εργασία των [Carvalho and Cohen, 2007], όπου προτείνονται στους χρήστες πιθανοί παραλήπτες κατά τη φάση της σύνταξης ενός μηνύματος. Επίσης, συλλογές δεδομένων που προέρχονται από την προσπέλαση (crawling) των ιστολογίων έχουν χρησιμοποιηθεί σε συστήματα συστάσεων, τόσο σε παραδοσιακά [Hayes et al., 2007] που έχουν αντικείμενο την πρόταση άρθρων σε αναγνώστες, όσο και σε κοινωνικά [Hsu et al., 2006], που έχουν στόχο να προτείνουν ιστολόγια αλλά και πιθανούς συνεργάτες σε συντάκτες. Η πλειοψηφία των SRS, ωστόσο, βασίζεται στα δεδομένα που συλλέγονται από τα μέσα κοινωνικής δικτύωσης.

5.3.1 Μέσα Κοινωνικής Δικτύωσης

Τα μέσα κοινωνικής δικτύωσης αποτελούν την πιο άμεση αναπαράσταση κοινωνικών δικτύων στον παγκόσμιο ιστό. Οι χρήστες τους ενθαρρύνονται στο να συνδέονται μεταξύ τους

Σχήμα 5.3: Ταξινόμηση μέσων κοινωνικής δικτύωσης (Πηγή: fredcavvaza.net)

με συγκεκριμένους δεσμούς, οι οποίοι συνήθως αναφέρονται ως δεσμοί *φιλίας*. Παραδείγματα τέτοιων δικτύων αποτελούν το Facebook, το Twitter και το LinkedIn. Στα μέσα κοινωνικής δικτύωσης ο όρος «φιλία» είναι συνώνυμος με τον όρο «δεσμός» της ανάλυσης κοινωνικών δικτύων και νοηματοδοτεί τη σχέση μεταξύ δύο μελών. Συνήθως όμως, το κάθε μέλος διατηρεί εκατοντάδες τέτοιους δεσμούς με άλλα μέλη, σε τέτοιο βαθμό που η έννοια της φιλίας να χάνει αρκετό από το συναισθηματικό φορτίο που έχει στον πραγματικό κόσμο.

Σύμφωνα με τους [Boyd and Heer, 2006], οι άνθρωποι σχηματίζουν δεσμούς στα μέσα κοινωνικής δικτύωσης για πολλούς λόγους. Για παράδειγμα, μπορεί να είναι φίλοι στον πραγματικό κόσμο ή μπορεί να θέλουν να γίνουν αρεστοί σε άλλους ανθρώπους. Άλλοι πάλι συμμετέχουν για λόγους κοινωνικού στάτους ή απλά για να ενημερώνονται για τη δραστηριότητα άλλων ανθρώπων. Συνεπώς, οι λόγοι σύναψης σχέσεων φιλίας δεν περιορίζονται απλά σε διαφορετικούς βαθμούς της ίδιας έννοιας (όπως συμβαίνει με την έννοια της *συνάφειας* στα προσωπικά κοινωνικά δίκτυα), αλλά αντίθετα ανταποκρίνονται σε διαφορετικού τύπου σχέσεις κάθε φορά.

Τα μέσα κοινωνικής δικτύωσης αποτελούν την πρωταρχική πηγή δεδομένων για τα κοινωνικά συστήματα συστάσεων για δύο λόγους. Κατ΄ αρχήν η έννοιας της κοινωνικότητας, της διασύνδεσης δηλαδή των χρηστών, είναι δομικό τους στοιχείο και κατά δεύτερον έχει *καθολική ισχύ*, δηλαδή αφορά όλα τα μέλη τους (και όχι μόνο ένα μικρό υποσύνολο που έχει λ.χ. συμπληρώσει ένα ερωτηματολόγιο). Μέσω των διαθέσιμων προγραμματιστικών διεπαφών είναι ιδιαίτερα εύκολη η προσπέλαση της κοινωνικής πληροφορίας, η οποία στη συγκεκριμένη περίπτωση είναι γενικού σκοπού, δηλαδή μπορεί να χρησιμοποιηθεί για να εξαχθούν χρήσιμα συμπεράσματα για τις σχέσεις μεταξύ των μελών του κοινωνικού δικτύου (φιλιά, εμπιστοσύνη, επιρροή) σε αντίθεση με τις προηγούμενες δύο περιπτώσεις.

Τα μέσα κοινωνικής δικτύωσης με τη σειρά τους μπορούν να ταξινομηθούν σε πολλές κατηγορίες (Σχήμα 5.3 και [Kaplan and Haenlein, 2010]). Υπάρχουν τα *δίκτυα περιεχομένου*

(content networks) όπως το Youtube, το Flickr και το Instagram, οι συνεργατικοί ιστότοποι (collaborative projects) όπως η Wikipedia, οι εικονικοί κόσμοι (virtual worls) όπως το Second Life και το World of Warcraft και τέλος τα ευρύτατα διαδεδομένα μέσα κοινωνικής δικτύωσης, γενικά (Facebook, Twitter, Google+) και ειδικά (LinkedIn που είναι προσανατολισμένο στις θέσεις εργασίας).

5.4 Τρόποι Αλληλεπίδρασης

Το πλήθος των διαφορετικών σχέσεων σε ένα κοινωνικό δίκτυο, όπως έχει φανεί από την μέχρι στιγμής ανάλυση, μπορεί να είναι ιδιαίτερα ευρύ. Είναι προφανές ότι δεν έχει νόημα να εξεταστούν όλοι οι πιθανοί τύποι σχέσεων που μπορεί να ανακύψουν, αλλά μόνο το υποσύνολο αυτών που θα μπορούσαν να χρησιμοποιηθούν για να βελτιώσουν την λειτουργία των συστημάτων συστάσεων. Σε αυτή την ενότητα, θα γίνει μια επισκόπηση του τύπου των κοινωνικών σχέσεων η επεξεργασία των οποίων παρουσιάζει ιδιαίτερο ενδιαφέρον για τα SRS. Κατ΄ αρχήν, γίνεται ένας διαχωρισμός ανάλογα με τους δράστες των αλληλεπιδράσεων. Η πρώτη κατηγορία αφορά σχέσεις που μπορούν να υπάρξουν μεταξύ των χρηστών ενώ η δεύτερη κατηγορία τις σχέσεις που αναπτύσσονται μεταξύ χρηστών και αντικειμένων.

5.4.1 Αλληλεπίδραση μεταξύ των χρηστών

Εμπιστοσύνη, Δυσπιστία και Φήμη

Η *εμπιστοσύνη* (trust) αποτελεί ένα τρόπο ποσοτικοποίησης της έννοιας της *αξιοπιστίας* (trustworthiness) μιας οντότητας. Έχει τις ρίζες της στη θεωρία των κοινωνικών επιστημών και γενικά δεν υπάρχει κάποιος καθολικός ορισμός της. Μια πρώτη προσπάθεια αποτίμησης της έννοιας στο πλαίσιο της επιστήμης των υπολογιστών έγινε από τον [Marsh, 1994], ενώ οι [Artz and Gil, 2007] τονίζουν ότι το περιεχόμενο της μεταβάλλεται ανάλογα με το πεδίο στο οποίο χρησιμοποιείται. Ειδικότερα, στην περίπτωση των SRS, καταλληλότερος φαίνεται πως είναι ο ορισμός των [Mui et al., 2002], ο οποίος θέλει την εμπιστοσύνη να είναι η υποκειμενική προσδοκία που έχει μια οντότητα για μια άλλη και η οποία βασίζεται στο ιστορικό των μεταξύ τους αλληλεπιδράσεων. Σε κάθε περίπτωση, αφορά πάντα ακριβώς δύο οντότητες, είναι μη-συμμετρική (ή μονόδρομη) και, υπό προϋποθέσεις, *μεταβατική* (transitive).

Μια ακόμη κατηγοριοποίηση της εμπιστοσύνης αφορά και το εύρος της. Εκτός της διαπροσωπικής εμπιστοσύνης που αναφέρθηκε παραπάνω, υπάρχει και η έννοια της καθολικής εμπιστοσύνης που αναφέρεται στον τρόπο που αξιολογείται μια οντότητα συνολικά, στο πλαίσιο ενός συστήματος. Η δεύτερη αυτή έννοια διατυπώνεται εναλλακτικά και ως *φήμη* (reputation), ενώ τα SRS τα οποία την χρησιμοποιούν είναι γνωστά και ως *συστήματα φήμης* (reputation systems). Τέλος, μια παρεμφερής έννοια είναι και αυτή της *δυσπιστίας* (distrust), η οποία αποτιμά το βαθμό της μη-αποδοχής μιας οντότητας από μια άλλη. Παράδειγμα χρήσης της δυσπιστίας είναι οι *λίστες αποκλεισμού* (block lists) που χρησιμοποιούν τα μέλη των μέσων κοινωνικής δικτύωσης για να αποκλείσουν εντελώς την αλληλεπίδραση με χρήστες που θεωρούν μη-επιθυμητούς. Όπως συμβαίνει και με την έννοια της εμπιστοσύνης, δεν υπάρχει κάποιος γενικά αποδεκτός τρόπος μοντελοποίησης και χρήσης της δυσπιστίας [Victor et al., 2011]. Τα περισσότερα SRS δεν την χρησιμοποιούν ή θεωρούν ότι πρόκειται για το ακριβώς αντίθετο άκρο της εμπιστοσύνης σε μια ενιαία κλίμακα μέτρησης. Άλλες μελέτες [Victor et al., 2011], ωστόσο, δεν θεωρούν την εμπιστοσύνη και την δυσπιστία ως αντίθετες έννοιες αλλά ως παράλληλες, οι οποίες μπορούν να λαμβάνουν χώρα ταυτόχρονα.

Και οι τρεις προαναφερόμενες έννοιες αποτελούν τις κατεξοχήν σχέσεις που χρησιμοποιούνται στα SRS. Εκτός από την πληθώρα εφαρμογών που έχουν ήδη βρει, υπάρχουν και αρκετές δημόσια διαθέσιμες συλλογές δεδομένων οι οποίες είναι εμπλουτισμένες με τέτοιου τύπου σχέσεις. Αυτό έχει ως αποτέλεσμα να έχουν μελετηθεί διεξοδικά από την ερευνητική κοινότητα, παράγοντας σημαντικά αποτελέσματα. Ενδεικτικά αναφέρονται οι συλλογές δεδομένων του Epinions [Massa and Avesani, 2009], του Filmtrust [Golbeck and Hendler, 2006] και του Filmtipset[fil, 2010]. Οι συγκεκριμένες σχέσεις θα αποτελέσουν το κυριότερο αντικείμενο της παρούσας διατριβής, όσον αφορά την μελέτη των κοινωνικών συστημάτων συστάσεων, και θα αναλυθούν διεξοδικότερα σε επόμενη ενότητα.

Φιλία

Η έννοια της *φιλίας* (friendship) υποδηλώνει ένα δεσμό μεταξύ δύο οντοτήτων. Όπως συμβαίνει και με την εμπιστοσύνη, ο όρος είναι αρκετά ευρύς και αποκτά διαφορετικό περιεχόμενο ανάλογα με το πεδίο που χρησιμοποιείται κάθε φορά. Κυρίως συναντάται στα μέσα κοινωνικής δικτύωσης, τόσο σε αυτά που είναι γενικού σκοπού όσο και σε αυτά που είναι ειδικού σκοπού [Baym and Ledbetter, 2009]. Γενικά θεωρείται αμφίδρομος δεσμός, χωρίς ωστόσο αυτό να αποτελεί αναγκαία συνθήκη. Στις περιπτώσεις που ο συγκεκριμένος δεσμός είναι μονόδρομος αναφέρεται και ως *παρακολούθηση* (follow). Παράδειγμα κοινωνικού δικτύου που δομείται σε δεσμούς παρακολούθησης αποτελεί η κοινωνική πλατφόρμα μικρο-ιστολογίων Twitter.

Από τη σκοπιά των SRS ενδιαφέρει η μελέτη αυτού του τύπου της σχέσης και κυρίως το κατά πόσο και σε ποιο βαθμό αυτός ο δεσμός υποδηλώνει συνάφεια στις προτιμήσεις των χρηστών. Όπως έχει ήδη αναφερθεί, η έννοια της φιλίας στα μέσα κοινωνικής δικτύωσης είναι αρκετά πιο ευρεία απ΄ ότι στις καθημερινές συναναστροφές και αφορά πολύ περισσότερους ανθρώπους, με αποτέλεσμα να χάνει αρκετό από το συναισθηματικό της φορτίο. Από την άλλη όμως, ο κυριότερος λόγος της χρήσης των μέσων κοινωνικής δικτύωσης και της δημιουργίας δεσμών μεταξύ ανθρώπων που είναι σε μικρότερο ή μεγαλύτερο βαθμό γνωστοί μεταξύ τους, είναι η εύρεση ατόμων που έχουν παρόμοιες προτιμήσεις. Συνεπώς, η μοντελοποίηση των σχέσεων φιλίας αναμένεται να επηρεάσει θετικά τη διαδικασία παραγωγής συστάσεων.

Ιδιότητα του μέλους

Η *ιδιότητα του μέλους* (membership) αποτελεί έναν έμμεσο τρόπο σύνδεσης μεταξύ δύο ή περισσότερων οντοτήτων και αποτελεί το ισοδύναμο της έννοιας των δικτύων συνεργατών της ανάλυσης κοινωνικών δικτύων. Η ιδιότητα του μέλους είναι και αυτή μια σχέση που απαντάται κυρίως στα μέσα κοινωνικής δικτύωσης και αποτελεί έναν διαφορετικό τρόπο αλληλεπίδρασης μεταξύ των χρηστών τους, πέρα από τον δεσμό της φιλίας. Συνηθέστερα, η ιδιότητα μέλους αφορά τη συμμετοχή σε κάποια ομάδα με ένα λιγότερο ή περισσότερο καθορισμένο αντικείμενο (λ.χ ένα συγκεκριμένο μουσικό συγκρότημα ή ένα είδος μουσικής) ή σε κάποια εκδήλωση. Υπάρχουν αρκετοί λόγοι για τους οποίους τα άτομα επιλέγουν να συμμετέχουν σε ομάδες ή εκδηλώσεις, όπως λ.χ για ενημέρωση ή για την γνωριμία με ανθρώπους με παρόμοια ενδιαφέροντα. Είναι συνεπώς εμφανές ότι από την ανάλυση τέτοιου τύπου σχέσεων μπορούν να εξαχθούν χρήσιμα συμπεράσματα για τα κοινωνικά συστήματα συστάσεων, κυρίως όσον αφορά την μοντελοποίηση της συσχέτισης μεταξύ των μελών τους. Υπάρχουν, μάλιστα, και σχετικές δημόσιες συλλογές δεδομένων διαθέσιμες, όπως είναι για παράδειγμα αυτή του

κοινωνικού μουσικού ιστοτόπου last.fm [Selma, 2010].

5.4.2 Αλληλεπίδραση χρηστών - αντικειμένων

Στα συστήματα που έχουν εξεταστεί μέχρι στιγμής, η αλληλεπίδραση χρηστών - αντικειμένων γίνεται με τη μορφή των αξιολογήσεων. Δηλαδή, μιας μονόδρομης σχέσης χρήστη-αντικειμένου, όπου ο πρώτος αξιολογεί το δεύτερο σε μια σαφώς καθορισμένη διακριτή κλίμακα, με σκοπό να μεγιστοποιήσει τη συνάρτηση ωφέλειάς του (Κεφάλαιο 2). Τα κοινωνικά συστήματα συστάσεων έχουν την δυνατότητα να επεκτείνουν αυτή την αλληλεπίδραση, δίνοντάς της περισσότερο βάθος με την χρήση *λέξεων-κλειδιών* (keywords) και *ετικετών* (tags). Πρόκειται για σύντομες λεκτικές περιγραφές, με τη μορφή αυτόνομων λέξεων ή πολύ σύντομων φράσεων, που χρησιμοποιούνται από τους χρήστες και τους σχεδιαστές του συστήματος με σκοπό την περιγραφή των αντικειμένων. Οι σύντομες αυτές περιγραφές μπορεί να αφορούν ιδιότητες των υπό εξέταση αντικειμένων, υποκειμενικές κρίσεις των χρηστών για αυτά, τρόπους ταξινόμησης τους ή γενικότερα οποιαδήποτε άλλη πληροφορία. Είναι συνεπώς προφανές ότι η χρήση των ετικετών προσφέρει πολύ μεγαλύτερη δύναμη αναπαράστασης απ΄ ότι η αξιολόγηση εντός μιας ορισμένης κλίμακας. Οι αλληλεπιδράσεις χρηστών-αντικειμένων στα SRS συνηθέστερα εντάσσονται σε μία από τις παρακάτω ομάδες.

Οντολογίες

Η *οντολογία* αποτελεί μια εννοιολογική διαμόρφωση ενός πεδίου σε μια μορφή που είναι κατανοητή τόσο από τους ανθρώπους όσο και από τους υπολογιστές και η οποία περιλαμβάνει τις οντότητες, τις ιδιότητες τις σχέσεις και τα αξιώματα που υπάρχουν στο συγκεκριμένο πεδίο [Middleton et al., 2003]. Οι οντολογίες παρέχουν έναν *λεξιλόγιο* (vocabulary) για το πεδίο και προτυποποιούν, σε διάφορα επίπεδα, το σημασιολογικό περιεχόμενο αυτού του λεξιλογίου καθώς και τις ιεραρχικές δομές που συνδέουν τους διάφορους *όρους* (terms) του. Η μοντελοποίηση αυτή έχει στόχο να διευκολύνει την ανταλλαγή πληροφορίας μεταξύ των μερών που ενός συστήματος [Echarte et al., 2007], πράγμα το οποίο είναι ιδιαίτερα χρήσιμο σε πολλές κατηγορίες συστημάτων συστάσεων (λ.χ. σε ένα μουσικό σύστημα συστάσεων μπορούν να εισαχθούν, υπό τη μορφή οντολογίας, οι διάφορες κατηγορίες μουσικής καθώς και οι μεταξύ τους σχέσεις). Επιπλέον, η ορθή εφαρμογή της έχει ως αποτέλεσμα την εξάλειψη του προβλήματος της αμφισημίας (ambiguity), το οποίο συναντάται στα συστήματα που χρησιμοποιούν ελεύθερες, μη δομημένες-περιγραφές και συνεπώς μπορεί να βελτιώσει την απόδοση ιδιαίτερα των αλγορίθμων συνεργατικής διήθησης βασισμένων στο αντικείμενο. Από την άλλη όμως, επιβάλλει την αποδοχή της οντολογίας από όλα τα εμπλεκόμενα μέρη, πράγμα το οποίο δεν είναι πάντα αυτονόητο. Και αυτό γιατί οι χρήστες των συστημάτων συστάσεων είναι συνηθέστερα καταναλωτές (και σε ένα δεύτερο βαθμό παραγωγοί) της πληροφορίας και όχι κατ΄ ανάγκη εμβριθείς γνώστες του πεδίου (στο παράδειγμα του μουσικού συστήματος συστάσεων, δεν γνωρίζουν όλοι οι χρήστες τις διαφορές όλων των ειδών της μουσικής). Τέλος, οι οντολογίες θα πρέπει να ενημερώνονται τακτικά, κάθε φορά που συντελούνται αλλαγές στο πεδίο που απεικονίζουν (λ.χ δημιουργία νέων ειδών μουσικής), πράγμα το οποίο εισάγει ένα συγκεκριμένο κόστος επανασχεδιασμού.

Συνεργατικές ταξινομήσεις

Οι *συνεργατικές ταξινομήσεις* (folksonomies) παράγονται απευθείας από τους χρήστες και έχουν μια ελεύθερη, μη ιεραρχική δομή, σε αντίθεση με την προηγούμενη κατηγορία

[Gupta et al., 2011]. Ο ίδιος ο αγγλικός όρος είναι ένας νεολογισμός· προκύπτει από την σύνθεση των λέξεων folk (λαός) και taxonomy (ταξινόμηση). Στην ουσία πρόκειται για μια τριαδική σχέση μεταξύ ενός χρήστη, μιας ετικέτας και ενός αντικειμένου, η οποία συνοδεύεται με τη χρονική σήμανση (timestamp) της δημιουργίας της σχέσης. Οι ετικέτες επιλέγονται εντελώς ελεύθερα από τους χρήστες και για αυτό το λόγο δεν υπάρχουν σαφώς καθορισμένες σχέσεις και ιεραρχίες μεταξύ τους. Έτσι όλοι οι πιθανοί όροι θεωρείται ότι ανήκουν στον ίδιο «επίπεδο» χώρο ονομάτων (namespace). Αφού οι ετικέτες επιλέγονται από τους ίδιους τους χρήστες, στην ουσία αναπαριστούν τον δικό τους λεξιλογικό πλούτο και κατά συνέπεια παρέχουν μια αποδεκτή, από την κοινότητα, οργάνωση των αντικειμένων (λ.χ. οι ετικέτες που αποδίδονται στις φωτογραφίες στο Instagram). Επίσης, δίνουν ένα επιπλέον κίνητρο συμμετοχής στη διαδικασία της απόδοσης ετικετών γιατί οι χρήστες μπορούν να λάβουν αμέσως ανάδραση (λ.χ όλα τα αντικείμενα στα οποία έχει αποδοθεί μια συγκεκριμένη ετικέτα). Κατ΄ αυτόν τον τρόπο λαμβάνει χώρα μια έμμεση επικοινωνία μεταξύ των χρηστών διαμέσου των συγκεκριμένων μεταδεδομένων, η οποία μπορεί να φανεί ιδιαιτέρως χρήσιμη στα SRS. Ειδικότερα, τα συστήματα συστάσεων που χρησιμοποιούν ετικέτες και κοινωνικές ταξινομήσεις καλούνται *συστήματα κοινωνικής ετικετοποίησης* (social tagging systems - STS) , για τα οποία υπάρχουν αρκετές, δημόσια διαθέσιμες συλλογές δεδομένων [Grahsl et al., 2010].

5.5 Δίκτυα Εμπιστοσύνης

Από τις σχέσεις που αναπτύσσονται μεταξύ των χρηστών στα κοινωνικά δίκτυα και οι οποίες παρουσιαστήκαν στις προηγούμενες ενότητες, το μεγαλύτερο ενδιαφέρον παρουσιάζουν οι δεσμοί εμπιστοσύνης. Ο κυριότερος λόγος είναι ότι ακόμα και στην καθημερινότητα η εμπιστοσύνη, ως έννοια, είναι περισσότερο προσανατολισμένη προς την προτίμηση από ότι η φιλία. Παρότι οι άνθρωποι τείνουν να αποκτούν κοινωνικές συναναστροφές με άλλους ανθρώπους με τους οποίους μοιράζονται κοινά ενδιαφέροντα, εντούτοις αυτό δεν αποτελεί αναγκαστικά τον κυριότερο λόγο σύναψης φιλιών. Από την άλλη, σε διάφορους τομείς τις καθημερινότητας, οι άνθρωποι εμπιστεύονται άλλους ανθρώπους, ειδικούς, με τους οποίους δεν διατηρούν απαραίτητα φιλικές ή διαπροσωπικές επαφές (λ.χ έναν γιατρό σε περίπτωση ασθένειας).

Ειδικότερα, όσον αφορά τα μέσα κοινωνικής δικτύωσης και τα κοινωνικά συστήματα συστάσεων, έρευνες έχουν δείξει ότι οι χρήστες τους τείνουν να βασίζονται περισσότερο στις προτάσεις που προέρχονται από ανθρώπους που εμπιστεύονται παρά από μια κοινότητα ομοίων, αλλά αγνώστων σε αυτούς, χρηστών [Sinha and Swearingen, 2001]. Οι συστάσεις που παράγονται από τα εν λόγω συστήματα πηγάζουν από πληροφορίες που αντλούνται από ένα *δίκτυο εμπιστοσύνης* (trust network), δηλαδή ένα κοινωνικό δίκτυο το οποίο δείχνει σε πιο βαθμό τα μέλη μιας κοινότητας εμπιστεύονται το ένα το άλλο. Παραδείγματα τέτοιων δικτύων αποτελούν το Filmtrust [Golbeck and Hendler, 2006], μια διαδικτυακή κοινότητα σινεφίλ της οποίας τα μέλη βαθμολογούν ταινίες αλλά και το γούστο άλλων χρηστών της πλατφόρμας, και το Epinions [Massa and Avesani, 2006], ένας ιστότοπος ηλεκτρονικού εμπορίου ο οποίος επιτρέπει στα μέλη του να διατηρούν λίστες με χρήστες που εμπιστεύονται καθώς και λίστες αποκλεισμού (με χρήστες που δεν εμπιστεύονται καθόλου). Και για τις δύο προαναφερόμενες υπηρεσίες υπάρχουν δημόσια διαθέσιμες συλλογές δεδομένων, πράγμα το οποίο έχει συμβάλλει στην ευρεία εφαρμογή των δικτύων εμπιστοσύνης στα RS. Τα συστήματα συστάσεων τα οποία λαμβάνουν υπόψη τους την ρητά δηλωμένη εμπιστοσύνη μεταξύ των μελών τους καλούνται *συστήματα συστάσεων αναβαθμισμένα με δεσμούς εμπιστοσύνης* (trust-enhanced recommendation systems) [Victor et al., 2011] *ή συστήματα συστάσεων που λαμβάνουν υπόψη* τους την *εμπιστοσύνη* (trust-aware recommender systems) [Massa and Avesani, 2009].

5.5.1 Ταξινόμηση των μετρήσεων της Εμπιστοσύνης

Τα προαναφερόμενα συστήματα συστάσεων εκμεταλλεύονται την ρητά δηλωμένη εμπιστοσύνη μεταξύ των χρηστών τους με δύο κυρίως τρόπους· είτε με τη χρήση μηχανισμών *διάδοσης της εμπιστοσύνης* (trust propagation) είτε με τη χρήση μηχανισμών *άθροισης της εμπιστοσύνης* (trust aggregation). Στην πρώτη περίπτωση γίνεται η υπόθεση ότι η έννοια της εμπιστοσύνης εμπεριέχει και μια μεταβατική (transitive) ιδιότητα· αν ο χρήστης a έχει δηλώσει ότι εμπιστεύεται τον b, και ο b με τη σειρά του ότι εμπιστεύεται τον c, τότε κάτι μπορεί να ειπωθεί για τον βαθμό που θα εμπιστευόταν ο a τον c αν τον γνώριζε. Στην δεύτερη περίπτωση, συνδυάζονται δύο ή περισσότερες προτιμήσεις εμπιστοσύνης για να παραχθεί μια ενιαία τιμή εμπιστοσύνης. Οι συγκεκριμένοι δύο τρόποι δεν είναι απαραίτητα αμοιβαία αποκλειόμενοι, δηλαδή μπορούν και να συνδυάζονται.

Μια ακόμα ταξινόμηση της εμπιστοσύνης πραγματοποιείται στη βάση του τρόπου που αντιλαμβάνονται την συγκεκριμένη έννοια οι διάφορες μέθοδοι [Victor et al., 2011]. Υπάρχει η *πιθανοτική προσέγγιση* (probabilistic approach), κατά την οποία η εμπιστοσύνη αντιμετωπίζεται με μια δυαδική λογική. Δηλαδή μια οντότητα είτε έχει την απόλυτη εμπιστοσύνη μιας άλλης ή δεν την έχει καθόλου. Για το λόγο αυτό, υπολογίζεται από το σύστημα η πιθανότητα μια οντότητα να εμπιστευτεί μια άλλη· όσο υψηλότερη είναι η μέτρηση της εμπιστοσύνης, τόσο πιο πιθανό είναι τελικά η μια οντότητα να εμπιστευτεί την άλλη. Εκτός της πιθανοτικής προσέγγισης, υπάρχει και η *σταδιακή προσέγγιση* (gradual approach), κατά την οποία υπάρχει δεσμός εμπιστοσύνης μεταξύ των μελών και ο οποίος ενισχύεται κάθε φορά που υπάρχει μια θετική αλληλεπίδραση. Σε αυτό το περιβάλλον η μέτρηση της εμπιστοσύνης δεν συμβολίζει πλέον μια πιθανότητα αλλά όσο υψηλότερη τιμή έχει, τόσο περισσότερο εμπιστεύεται η μια οντότητα την άλλη. Αυτή η οπτική μοντελοποιεί καλύτερα τον τρόπο που γίνεται κατανοητή η εμπιστοσύνη στον πραγματικό κόσμο. Οι άνθρωποι σπάνια εμπιστεύονται κάποιον άλλο είτε απόλυτα είτε καθόλου· συνήθως εμπιστεύονται τους άλλους «σε κάποιο βαθμό».

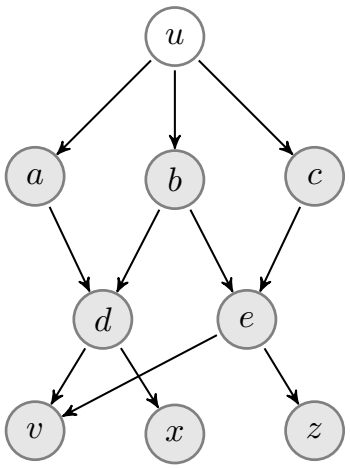
Τέλος, η εργασία των [Ziegler and Lausen, 2005] διαχωρίζει περαιτέρω τις *μετρικές της εμπιστοσύνης* (trust metrics) με τρία κριτήρια ταξινόμησης· το *πεδίο τιμών τους*, τον *τόπο υπολογισμού τους* και την *οπτική του δικτύου*. Όσον αφορά το πεδίο τιμών των μετρικών εμπιστοσύνης αυτό μπορεί να είναι είτε διανυσματικό είτε, συνηθέστερα, βαθμωτό.

Ο υπολογισμός της τιμής τους, από την άλλη, μπορεί να λαμβάνει χώρα σε ένα και μόνο υπολογιστή ή να είναι το προϊόν μιας *κατανεμημένης διεργασίας* (distributed process). Στην πρώτη περίπτωση, όλες οι τιμές εμπιστοσύνης συγκεντρώνονται σε ένα κεντρικό σημείο και κατόπιν εφαρμόζεται ο αλγόριθμος για τον υπολογισμό της φήμης του κάθε χρήστη. Μια τέτοια λύση εμφανίζει ένα μεγάλο υπολογιστικό κόστος σε κεντρικό επίπεδο αλλά είναι αρκετά απλή στην υλοποίηση της και επιπλέον επιφέρει μικρή δικτυακή επιβάρυνση. Από την άλλη, στη δεύτερη περίπτωση, τα μέλη του δικτύου συνεργάζονται και αποστέλλουν δεδομένα μεταξύ τους με την χρήση πρωτοκόλλων επικοινωνίας με τελικό στόχο και πάλι τον υπολογισμό της φήμης για κάθε μέλος. Αυτή η λειτουργία παρουσιάζει ένα σαφώς μικρότερο υπολογιστικό κόστος (και είναι προτιμότερη στα πολύ μεγάλα δίκτυα εμπιστοσύνης), έχει όμως μεγαλύτερο κόστος επικοινωνίας γιατί θα πρέπει να ληφθεί υπόψη ο αριθμός των μηνυμάτων που ανταλλάσσονται μεταξύ των μελών του δικτύου.

Τέλος, η οπτική του δικτύου χωρίζεται στην ολική και την τοπική εμβέλεια, έννοιες που αναλύονται εκτενέστερα στις δύο επόμενες ενότητες.

5.5.2 Οπτική του Δικτύου: τοπική εμβέλεια

Οι *μετρικές εμπιστοσύνης τοπικής εμβέλειας* (local trust metrics) υπολογίζουν προσωποποιημένες τιμές εμπιστοσύνης που απευθύνονται σε συγκεκριμένο χρήστη κάθε φορά. Η θεωρητική αφετηρία των συγκεκριμένων μετρικών είναι η παρατήρηση πως τα μέλη του δικτύου που εμπιστεύεται ένας χρήστης x ενδέχεται να είναι εντελώς διαφορετικά από τα αντίστοιχα που εμπιστεύεται ένας άλλος χρήστης y. Για το λόγο αυτό, αναλύουν την τοπική δομή



Σχήμα 5.4: Ο προσωπικός ιστός εμπιστοσύνης WOT του χρήστη u

της πληροφορίας που προέρχεται από τους προσωποποιημένους *ιστούς εμπιστοσύνης* (webs of trust - WOTs) κάθε χρήστη. Πρόκειται για *υπογράφους* (subgraphs) του αρχικού δικτύου εμπιστοσύνης που έχουν ως αφετηρία τον συγκεκριμένο χρήστη και περιλαμβάνουν όλες τις ακμές και του κόμβους που είναι προσβάσιμες από αυτόν είτε άμεσα είτε έμμεσα (Σχήμα 5.4).

Αφού κατασκευαστεί το WOT κάθε χρήστη, στο επόμενο βήμα καθορίζεται ο τρόπος επεξεργασίας του. Υπάρχουν αλγόριθμοι μέτρησης της εμπιστοσύνης που, έχοντας ως αφετηρία τον αρχικό χρήστη, επεξεργάζονται κλιμακωτά όλους τους συνδεδεμένους σε αυτόν χρήστες, ξεκινώντας από τους άμεσους γείτονες και κατόπιν προχωρούν στους γείτονες των γειτόνων κ.ο.κ. Πρόκειται για τις *κλιμακωτές μετρικές της εμπιστοσύνης* (gradual trust metrics), το εύρος και ο τρόπος επεξεργασίας των οποίων διαφέρει σημαντικά από αλγόριθμο σε αλγόριθμο. Κοινό χαρακτηριστικό τους, όμως, είναι ότι λαμβάνουν υπόψη όλους τους γείτονες του WOT που πληρούν ορισμένα κριτήρια. Οι πιο χαρακτηριστικές τέτοιες μετρικές είναι οι αλγόριθμοι MoleTrust και TidalTrust που παρουσιάζονται παρακάτω.

MoleTrust

Ο αλγόριθμος MoleTrust προτάθηκε από τους [Massa and Avesani, 2007] και επί της ουσίας πραγματοποιεί μια *εξερεύνηση κατά βάθος* (depth-first search - DFS) στο WOT ενός δοσμένου χρήστη u. Εκτελείται σε δύο στάδια: στο πρώτο αφαιρούνται όλοι οι κύκλοι από το WOT, μετατρέποντάς το σε έναν ακυκλικό γράφο (δέντρο), όπως στο παράδειγμα του Σχήματος 5.4 για τον χρήστη u. Στο δεύτερο στάδιο πραγματοποιείται ο καθεαυτό υπολογισμός της εμπιστοσύνης. Έχοντας ως αφετηρία τον u, η εμπιστοσύνη ως προς τους άλλους χρήστες υπολογίζεται κλιμακωτά μέχρι του βάθους που ορίζεται από τον *ορίζοντα της διάδοσης* (propagation horizon), ο οποίος αποτελεί και αυτός παράμετρο της μετρικής (εκτός από τον χρήστη u).

Οι πιο ευρύτερα χρησιμοποιούμενες παραλλαγές του είναι δύο: ο MoleTrust-1 και ο MoleTrust-2. Στην πρώτη περίπτωση, ο ορίζοντας της εξερεύνησης τίθεται στο 1, δηλαδή αφορά μόνο τους γείτονες του u, με τους οποίους συνδέεται άμεσα στο WOT (χρήστες a, b, c στο παράδειγμα του Σχήματος 5.4). Στη δεύτερη περίπτωση ο ορίζοντας της εξερεύνησης τίθεται στο 2, δηλαδή αφορά τους γείτονες του u καθώς και τους γείτονες των γειτόνων (πέραν των a, b, c και οι d, e του Σχήματος 5.4). Γενικότερα, η περίπτωση των δικτύων που περιέχουν γείτονες γειτόνων (όπως το MoleTrust-2) είναι γνωστή και ως *δίκτυα FoaF* (Friend-of-a-Friend network).

Σε περίπτωση που κάποιος χρήστης δεν συνδέεται άμεσα με τον u (έστω ο v), τότε ο βαθμός εμπιστοσύνης του ως προς αυτόν υπολογίζεται από τον αναδρομικό τύπο της παρακάτω Εξίσωσης

$$t_{u,v} = \frac{\sum_{i \in WOT(u)} t_{u,i} t_{i,v}}{\sum_{i \in WOT(u)} t_{a,i}} \tag{5.1}$$

Στο παράδειγμα του Σχήματος 5.4, στο $WOT(u)$ περιέχονται όλοι οι χρήστες και επειδή πρόκειται για εξερεύνηση κατά βάθος με αφετηρία τον u, η σειρά εμφάνισής τους είναι a, d, b, e και c. Ή με άλλα λόγια ο υπολογισμός της τιμής $t_{u,v}$ υπολογίζεται επαναληπτικά από «πάνω» προς τα «κάτω», δηλαδή αφού πρώτα έχει προσεγγιστεί η τιμή $t_{u,d}$, αυτή συνδυάζεται με την $t_{d,v}$ (αντίστοιχα για $t_{u,e}$ και $t_{e,v}$).

TidalTrust

Ο αλγόριθμος TidalTrust αποτελεί παραλλαγή του MoleTrust και επί της ουσίας πραγματοποιεί μια *εξερεύνηση κατά πλάτος* (breadth-first search - BFS) στο WOT ενός δοσμένου χρήστη u. Προτάθηκε από την [Golbeck, 2005], η οποία μετά από πειράματα κατέληξε στο συμπέρασμα πως σε ένα WOT οι συντομότερες διαδρομές παράγουν καλύτερες εκτιμήσεις εμπιστοσύνης και δεύτερον ότι το ίδιο ισχύει και για διαδρομές που περιέχουν ακμές με μεγάλες τιμές εμπιστοσύνης. Έτσι, στην προσέγγιση της εμπιστοσύνης από έναν χρήστη u σε ένα χρήστη v, το συντομότερο μονοπάτι που ενώνει τους δύο χρήστες ορίζεται ως το βάθος του αλγορίθμου. Η δεύτερη παρατήρηση ενσωματώνεται στον αλγόριθμο μέσω της συμπερίληψης στον υπολογισμό μόνο των μονοπατιών εκείνων τα οποία ξεπερνούν κάποιο ορισμένο κατώφλι εμπιστοσύνης. Το σύνολο $WOT^+(u)$ περιέχει όλους εκείνους τους χρήστες για τους οποίους η εκπεφρασμένη εμπιστοσύνη του u υπερβαίνει το ορισμένο κατώφλι. Χάριν ευκολίας, στο παράδειγμα του Σχήματος 5.4 θα θεωρηθεί ότι όλες οι ακμές εντάσσονται στο $WOT^+(u)$ (δηλαδή πως τα σύνολα $WOT(u)$ και $WOT^+(u)$ ταυτίζονται).

Η Εξίσωση 5.1 περιγράφει, και για την περίπτωση του TidalTrust, τον αναδρομικό τύπο υπολογισμού της εμπιστοσύνης που έχει ο χρήστης u για τον χρήστη v, όταν αυτοί δεν συνδέονται άμεσα με ακμή στο $WOT^+(u)$. Σε αυτή την περίπτωση όμως, έχουμε υπολογισμό κατά πλάτος προς τον u και με αφετηρία τον v, οι κόμβοι στο WOT^+ προσπελάζονται με τη σειρά d, e, a, b και c. Ή με άλλα λόγια ο υπολογισμός της $t_{u,v}$ υπολογίζεται επαναληπτικά από «κάτω» προς τα «πάνω», δηλαδή αφού πρώτα έχει προσεγγιστεί η τιμή $t_{a,v}$, αυτή συνδυάζεται με την $t_{u,a}$ (αντίστοιχα για τις $t_{b,v}$ και $t_{c,v}$).

5.5.3 Οπτική του Δικτύου: ολική εμβέλεια

Οι *μετρικές εμπιστοσύνης ολικής εμβέλειας* (global trust metrics) υπολογίζουν μια τιμή για το κάθε μέλος του δικτύου εμπιστοσύνης, η οποία απηχεί την μέση αντίληψη που έχει η κοινότητα για το συγκεκριμένο μέλος. Παρότι δεν υπάρχει ακόμα κάποια γενικά παραδεκτή ορολογία, η τιμή αυτή συνηθέστερα καλείται *φήμη* (reputation) και τα συστήματα συστάσεων που τη χρησιμοποιούν καλούνται *συστήματα φήμης* (reputation systems). Εδώ πρέπει να σημειωθεί ότι τα συστήματα φήμης δεν περιορίζονται μόνο στην παραγωγή συστάσεων αλλά αντίθετα αποτελούν γενικότερη κατηγορία συστημάτων που βρίσκουν εφαρμογές και σε άλλα πεδία. Χαρακτηριστικότερο παράδειγμα συστήματος φήμης αποτελεί ο αλγόριθμος PageRank [Page et al., 1999] που χρησιμοποιείται από τη μηχανή αναζήτησης Goolge για τον καθορισμό της σειράς των αποτελεσμάτων.

Γενικότερα πάντως, στην βιβλιογραφία των κοινωνικών συστημάτων συστάσεων που χρησιμοποιούν δίκτυα εμπιστοσύνης, η διαδικασία που έχει κατ΄ επανάληψη χρησιμοποιηθεί για την προσέγγιση της φήμης είναι αυτή των *τυχαίων περιπάτων* (random walks) πάνω στον γράφο εμπιστοσύνης [Lovasz, 1993] (με χαρακτηριστικότερο ίσως παράδειγμα το σύστημα TrustWalker [Jamali and Ester, 2009]). Οι τυχαίοι περίπατοι έχουν αφετηρία έναν συγκεκριμένο κόμβο του δικτύου εμπιστοσύνης (λ.χ. τον δοσμένο χρήστη u) και μετακινούνται σε έναν από τους γείτονες του, με πιθανότητα που έχει καθοριστεί εκ των προτέρων. Κατόπιν μετακινούνται πάλι σε έναν από τους γείτονες του γείτονα, πάλι με πιθανότητα που έχει καθοριστεί εκ των προτέρων. Η διαδικασία αυτή είναι επαναληπτική, μέχρις ότου ο τυχαί-ος περίπατος βρεθεί σε κόμβο που δεν έχει εξερχόμενες ακμές ή μέχρις ότου ικανοποιηθεί κάποιο κριτήριο τερματισμού του περιπάτου. Αν πραγματοποιηθούν πολλοί περίπατοι και μετρηθεί πόσες φορές «εμφανίστηκε» ο κάθε κόμβος σε κάποιον από αυτούς, τότε θα παρατηρηθεί το φαινόμενο ορισμένοι κόμβοι να έχουν δεχτεί περισσότερες «επισκέψεις» από κάποιους άλλους. Η «δημοφιλία» αυτή των συγκεκριμένων κόμβων αποδίδεται στα δομικά χαρακτηριστικά του δικτύου και πιο συγκεκριμένα στη θέση τους σε αυτό. Σημαντική, τέλος, ιδιότητα των τυχαίων περιπάτων είναι αυτή της *«αμνησίας»* (memorylessness), η οποία ορίζει ότι η σε κάθε βήμα επιλογή του επόμενου κόμβου του περιπάτου εξαρτάται αποκλειστικά και μόνο από τον παρόντα κόμβο και όχι από όλους τους προηγούμενους (του συγκεκριμένου περιπάτου).

Για να μπορέσει να πραγματοποιηθεί ο τυχαίος περίπατος, θα πρέπει καταρχήν το δίκτυο εμπιστοσύνης να μετατραπεί σε ένα κατευθυντικό γράφο G

$$G(V,E), \quad V=\{u \in U\}, \quad E=\{\exists t_{i,j}, \forall\ i,j \in V\} \tag{5.2}$$

όπου V το σύνολο των κόμβων του (οι χρήστες) και E το σύνολο των ακμών του (οι δηλώσεις εμπιστοσύνης). Ο γράφος G συνηθέστερα αναπαρίσταται από τον $n \times n$ *πίνακα γειτνίασης* (adjacency matrix) του A (όπου $n=|V|$ το πλήθος των χρηστών), τα στοιχεία του οποίου λαμβάνουν τιμές όπως παρακάτω

$$a_{ij}=\begin{cases} t_{i,j}, & \text{αν } i,j \in V \text{ και } \exists t_{ij} \in E \\ 0, & \text{διαφορετικά} \end{cases} \tag{5.3}$$

Εκτός από τον πίνακα γειτνίασης, ορίζεται και ο *πίνακας μετάβασης* (transition matrix), ο οποίος έχει τις ίδιες διαστάσεις με τον A και καθορίζει την πιθανότητα μετάβασης από τον κόμβο i σε έναν από τους γείτονές του. Όπως αναφέρθηκε και προηγουμένως, αυτή η πιθανότητα εξαρτάται μόνο από τον παρόντα κόμβο και όχι από τα προηγούμενα βήματα του περιπάτου

$$p_{ij}=\begin{cases} \phi(i,j,t_{i,j}), & \text{εαν } t_{ij} \in E \\ 0, & \text{διαφορετικά} \end{cases} \tag{5.4}$$

όπου $\phi(i,j,t_{i,j})$ μια συνάρτηση του παρόντος κόμβου i, του επόμενου j και τις ακμής $t_{i,j}$ που τους συνδέει.

Η αλληλουχία των κόμβων που θα «επισκεφτεί» ένας τυχαίος περίπατος θεωρείται ότι σχηματίζουν μια *μαρκοβιανή αλυσίδα* (markov chain) [Lovasz, 1993]. Αν ο γράφος είναι συνδεδεμένος και μη διμερής (όπως ισχύει δηλαδή για τα δίκτυα εμπιστοσύνης), τότε η συγκεκριμένη μαρκοβιανή αλυσίδα είναι *εργοδική* (ergodic), δηλαδή *αμείωτη* (irreducible) και *απεριοδική* (aperiodic) και επειδή ο γράφος είναι κατευθυντικός είναι και *χρονικά μη-αναστρέψιμη* (time-irreversible) [Lovasz, 1993]. Χαρακτηριστικό αυτών των αλυσίδων είναι ότι μετά από πολλές επαναλήψεις, ο τυχαίος περίπατος συγκλίνει σε μια και μοναδική *στάσιμη κατανομή* (stationary distribution) π, που εκφράζει την πιθανότητα προσπέλασης του κάθε κόμβου του δικτύου, ανεξάρτητα από το κάθε φορά σημείο εκκίνησης του περιπάτου [Lovasz, 1993].

Η στάσιμη κατανομή έχει επιπλέον την πολύ σημαντική ιδιότητα ότι παραμένει *αναλλοίωτη* (invariant) ως προς τον πίνακα μετάβασης P, δηλαδή $\pi P=\pi$. Τέλος, για τις μαρκοβιανές

αλυσίδες ορίζεται και ο *χρόνος μίξης* (mixing time) ως ο ελάχιστος αριθμός βημάτων που απαιτούνται μέχρι ο τυχαίος περίπατος να φτάσει στην στάσιμη του κατανομή. Ο χρόνος μίξης μετράει και τον βαθμό συνδεσιμότητας του γράφου· ισχυρά συνδεδεμένοι γράφοι έχουν πολύ μικρούς χρόνους μίξης ενώ οι ασθενώς συνδεδεμένοι γράφοι έχουν αντίθετα πολύ μεγάλους χρόνους μίξης.

Είναι προφανές ότι αυτό που διαφοροποιεί τους αλγορίθμους τυχαίων περιπάτων είναι η παραδοχές που κάνουν για την πιθανότητα μετάβασης από τον ένα κόμβο στον άλλο. Για παράδειγμα, το σύστημα TrustWalker που αναφέρθηκε προηγουμένως επιλέγει το επόμενο βήμα ομοιόμορφα τυχαία (uniformly at random), με πιθανότητα ίση με τον *βαθμό εξερχομένων ακμών* $\deg(i)$ (out-degree) του τρέχοντος κόμβου i

$$p_{ij} = \begin{cases} \frac{1}{\deg(i)}, & \text{αν } t_{ij} \in E \\ 0, & \text{διαφορετικά} \end{cases} \tag{5.5}$$

ο οποίος προκύπτει από την *νόρμα γραμμών* (row-norm) του πίνακα γειτνίασης

$$\deg(i) = \sum_{k=1}^{n} a_{ik} \tag{5.6}$$

Μια άλλη, αρκετά διαδεδομένη, παραλλαγή των τυχαίων περιπάτων είναι οι *τυχαίοι περίπατοι με επανεκκίνηση* (random walks with restarts). Αποτελούν μια μεταφορά στο πεδίο των κοινωνικών συστημάτων συστάσεων του γνωστού αλγορίθμου ταξινόμησης ιστοσελίδων PageRank [Page et al., 1999] της Google. Σε κάθε τους βήμα, με πιθανότητα α αποφασίζουν να επανεκκινήσουν τον περίπατο από τον αρχικό κόμβο u_r, ενώ με πιθανότητα $1-\alpha$ επιλέγουν ομοιόμορφα τυχαία κάποιον από τους γείτονες του τρέχοντος κόμβου. Σε αυτήν την περίπτωση, ο πίνακας μετάβασης τροποποιείται ως εξής

$$P_{rwr} = \alpha A_r + (1-\alpha)P \tag{5.7}$$

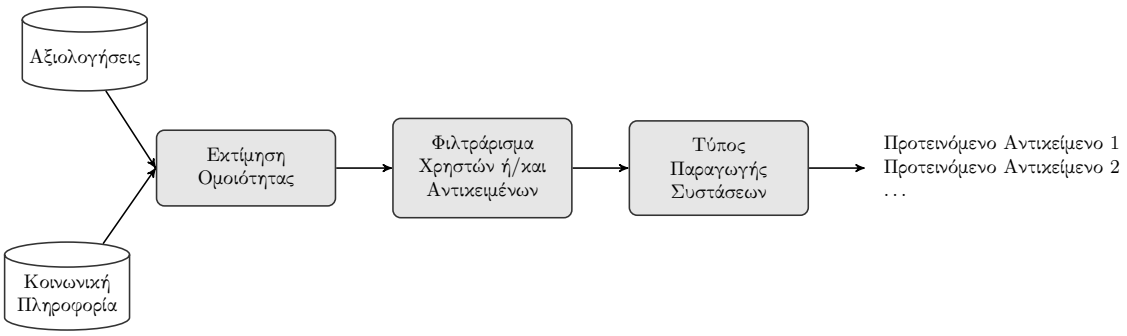
όπου A_r είναι ο πίνακας επανεκκίνησης· έχει όλα του τα στοιχεία μηδενικά, εκτός από την γραμμή r (που αντιστοιχεί στον αρχικό κόμβο u), της οποίας τα στοιχεία είναι όλα ίσα με τη μονάδα. Έτσι, η πιθανότητα μετάβασης από τον κόμβο i στον κόμβο j δεδομένου ότι ο u είναι ο αρχικός κόμβος δίνεται από την παρακάτω εξίσωση

$$p_{ij} = \begin{cases} \alpha, & j = u, u \notin N(i) \\ \alpha + \frac{1-\alpha}{\deg(i)}, & j = u, u \in N(i) \\ \frac{1-\alpha}{\deg(i)}, & j \neq u, j \in N(i) \\ 0, & \text{διαφορετικά} \end{cases} \tag{5.8}$$

όπου $N(i)$ οι γείτονες του i. Το άθροισμα κάθε γραμμής του P_{rwr} είναι ίσο με τη μονάδα, οπότε είναι έγκυρος πίνακας μετάβασης. Η διαφορά όμως με τις προηγούμενες περιπτώσεις είναι ότι η στάσιμη κατανομή π εξαρτάται από τον κόμβο αφετηρίας u και συνεπώς δεν είναι η ίδια για όλες τις αρχικές καταστάσεις. Για τον λόγο αυτό αναφέρεται και ως *φραγμένη κατανομή* (bounding distribution) για τον τυχαίο περίπατο με αφετηρία τον u. Ισχύει δηλαδή $\pi^{(u)} = [\pi_i]^{1\times n}$, όπου

$$\pi_i = \begin{cases} (1-\alpha)\frac{\deg(i)}{2m}, & i \in V \backslash u \\ \alpha + \frac{\deg(i)}{2m}, & i = u \end{cases} \tag{5.9}$$

Τέλος, μπορεί να αποδειχθεί εύκολα ότι $\pi P_{rwr} = \pi$, δηλαδή η π είναι η στάσιμη κατανομή του τυχαίου περιπάτου με επανεκκίνηση.



Σχήμα 5.5: Μοντέλο λειτουργίας των κοινωνικών αλγορίθμων μνημονικής συνεργατικής διήθησης

5.5.4 Ενσωμάτωση στα συστήματα συνεργατικής διήθησης

Οι σχέσεις εμπιστοσύνης μπορούν να ενσωματωθούν στα συστήματα συνεργατικής διήθησης με πολλούς τρόπους και σε πολλά επίπεδα. Η πιο απλή μέθοδος περιγράφεται στην εργασία των [Massa and Avesani, 2009] και ουσιαστικά αποτελεί μια απλή επέκταση των βασικών μνημονικών αλγορίθμων που παρουσιάστηκαν στην υποενότητα 2.3.1. Τα μνημονικά συστήματα συστάσεων λαμβάνουν ως είσοδο τις αξιολογήσεις των χρηστών και κατόπιν εκτελούν τη λειτουργία τους σε δύο φάσεις. Αρχικά υπολογίζουν την ομοιότητα μεταξύ των χρηστών με την χρήση μεθόδων μέτρησης της ομοιότητας (Εξίσωση 2.4) και στην συνέχεια με την χρήση των Εξισώσεων 2.9–2.11 (ή άλλων αντίστοιχων) παράγουν τις συστάσεις.

Μια πρώτη επέκταση αυτού του μοντέλου γίνεται στα μνημονικά συστήματα συνεργατικής διήθησης τα οποία είναι ενισχυμένα με δεσμούς εμπιστοσύνης. Σε αυτή την περίπτωση η είσοδος του συστήματος είναι διπλή· περιλαμβάνει τόσο τις αξιολογήσεις των χρηστών όσο και τις σχέσεις εμπιστοσύνης μεταξύ τους (Σχήμα 5.5 σε αντιδιαστολή προς το Σχήμα 2.2). Στο πρώτο βήμα πραγματοποιείται η επεξεργασία των δηλώσεων εμπιστοσύνης και κατόπιν στο δεύτερο βήμα γίνεται η ενσωμάτωσή τους στις εξισώσεις παραγωγής των συστάσεων. Οι συντελεστές ομοιότητας μεταξύ των χρηστών αντικαθίστανται με τους συντελεστές της διάδοσης της μεταξύ τους εμπιστοσύνης (ή της φήμης).

Τα πιο διαδεδομένα μνημονικά συστήματα συστάσεων εμπιστοσύνης αναφέρονται στην εργασία της [Victor et al., 2011]. Ο συνδυασμός της μεθόδου TidalTrust (Εξίσωση 5.1) με τον ζυγισμένο μέσο όρο (Εξίσωση 2.10) χρησιμοποιείται στο σύστημα συστάσεων βασισμένο στον *ζυγισμένο μέσο όρο εμπιστοσύνης* (trust-based weighted mean) (Εξίσωση 5.10).

$$\widehat{r_{a,i}} = \frac{\sum\limits_{u \in R^T} t_{a,u} r_{u,i}}{\sum\limits_{u \in R^T} t_{a,u}} \tag{5.10}$$

Από την άλλη, ο συνδυασμός της μεθόδου MoleTrust με τον τύπο του Resnick (Εξίσωση 2.11) χρησιμοποιείται στη *συνεργατική διήθηση βασισμένη στην εμπιστοσύνη* (trust-based collaborative filtering) (Εξίσωση 5.11).

$$\widehat{r_{a,i}} = \overline{r_a} + \frac{\sum\limits_{u \in R^T} t_{a,u} (r_{u,i} - \overline{r_u})}{\sum\limits_{u \in R^T} t_{a,u}} \tag{5.11}$$

Ο Πίνακας 5.1 [Massa and Avesani, 2009] συνοψίζει την απόδοση ενός βασικού συνεργατικού συστήματος συστάσεων (Εξίσωση 2.11), στο οποίο έχουν αντικατασταθεί οι συντελεστές

Πίνακας 5.1: Σύγκριση συστημάτων [Massa and Avesani, 2009] στη συλλογή δεδομένων Epinions [Massa and Bhattacharjee, 2004] (MAE και κάλυψη βαθμολογιών)

	CF		MT1		MT2		MT3	
	MAE	Κάλυψη	MAE	Κάλυψη	MAE	Κάλυψη	MAE	Κάλυψη
Α. Πλήρης Συλλογή	0,843	51,28%	0,832	28,33%	0,846	60,47%	0,829	74,37%
Β. Όψεις Χρηστών								
Λίγες αξιολογήσεις	1,094	3,22%	0,674	11,05%	0,833	25,02%	0,854	41,74%
Πολλές αξιολογήσεις	0,85	57,45%	0,873	30,85%	0,869	64,82%	0,846	77,81%
Χρήστες με άποψη	1,20	50%	1,02	23,32%	1,102	57,31%	1,096	74,24%
Μαύρα Πρόβατα	1,235	55,74%	1,152	23,66%	1,238	59,21%	1,242	76,32%
Γ. Όψεις Αντικειμένων								
Αμφιλεγόμενα	1,515	45,42%	1,425	25,09%	1,618	60,64%	1,687	81,01%
Περιθωριακά	0,822	12,18%	0,734	8,32%	0,806	24,32%	0,828	20,43%

ομοιότητας μεταξύ των χρηστών από τους συντελεστές εμπιστοσύνης που υπολογίζονται με την μέθοδο MoleTrust (Ενότητα 5.5.2). Ο αλγόριθμος που περιγράφεται στην Εξίσωση 5.11 εκτελέστηκε τρεις φορές, επιλέγοντας κάθε φορά διαφορετικό βάθος· στην πρώτη περίπτωση υπολογίστηκαν μόνο οι άμεσοι γείτονες (MT1), κατόπιν οι γείτονες των γειτόνων (MT2) και τέλος οι γείτονες τρίτου βαθμού (MT3). Για λόγους πληρότητας της σύγκρισης παρατίθενται και τα αποτελέσματα του αρχικού συστήματος συνεργατικής διήθησης (CF). Η μέτρηση της απόδοσης έγινε στη συλλογή δεδομένων Epinions [Massa and Bhattacharjee, 2004] με την χρήση του MAE και της κάλυψης βαθμολογιών και αφορούσε όλους τους χρήστες της συλλογής, αλλά και επιμέρους όψεις χρηστών και αντικειμένων (Κεφάλαιο 3).

Μια πρώτη ανάγνωση των αποτελεσμάτων φανερώνει ότι ακόμα και με αυτήν την πολύ βασική εισαγωγή των δεσμών εμπιστοσύνης στα συστήματα συστάσεων παρατηρούνται πολύ ενθαρρυντικά αποτελέσματα. Παρότι η πτώση του σφάλματος είναι μικρή, εντούτοις η κάλυψη των βαθμολογιών (δηλαδή το ποσοστό των περιπτώσεων που το σύστημα μπορεί να κάνει πρόβλεψη) αυξάνει εντυπωσιακά. Εξαιτίας της αραιότητας των αξιολογήσεων (Ενότητα 1.1), η απλή ομοιότητα μεταξύ των χρηστών τις περισσότερες φορές δεν είναι δυνατό να υπολογιστεί, γιατί οι χρήστες συμπίπτουν στην βαθμολογία που έχουν δώσει σε ελάχιστα ως καθόλου αντικείμενα. Και η απλούστερη όμως εξερεύνηση του δικτύου εμπιστοσύνης επιτρέπει την ανακάλυψη συσχετίσεων μεταξύ των χρηστών που δεν θα μπορούσαν να είχαν εντοπιστεί διαφορετικά, συνηγορώντας σε σημαντικό βαθμό υπέρ της ενσωμάτωσης στα συστήματα συστάσεων της άμεσης κοινωνικής πληροφορίας. Πιο χαρακτηριστική είναι η περίπτωση των χρηστών με λίγες βαθμολογίες και των περιθωριακών αντικειμένων. Τα κοινωνικά συστήματα συστάσεων είναι σε άμεση θέση να παράξουν προτάσεις για αυτήν την κατηγορία των χρηστών ακόμα και όταν έχουν δηλώσει ότι εμπιστεύονται ελάχιστους άλλους χρήστες.

□

Κεφάλαιο 6

Συνεργατικό Σύστημα Συστάσεων βασισμένο σε μη-αμερόληπτους τυχαίους περιπάτους

Στο προηγούμενο κεφάλαιο έγινε αναφορά στους τυχαίους περιπάτους και την χρήση τους για τον υπολογισμό της φήμης στα δίκτυα εμπιστοσύνης. Γενικότερα, όμως, οι τυχαίοι περίπατοι έχουν χρησιμοποιηθεί και μελετηθεί αρκετά στο πλαίσιο των κοινωνικών συστημάτων συστάσεων (π.χ. [Singh et al., 2007; Golbeck, 2005]), με αρκετούς ερευνητές να θεωρούν ότι η συγκεκριμένη τεχνική δεν έχει αξιοποιηθεί πλήρως και ότι υπάρχουν περιθώρια περαιτέρω βελτιώσεων (π.χ. [Andersen et al., 2008; Konstas et al., 2009]).

Στο παρόν κεφάλαιο προτείνεται ένας νέος αλγόριθμος κοινωνικής συνεργατικής διήθησης, ο οποίος βασίζεται στην πραγματοποίηση μη-αμερόληπτων τυχαίων περιπάτων για την παραγωγή συστάσεων. Η θεωρητική αφόρμηση για τη συγκεκριμένη επιλογή υπήρξε η παρατήρηση πως τα περισσότερα SRS τα οποία βασίζονται στους τυχαίους περιπάτους επιλέγουν το επόμενο βήμα του περιπάτου ομοιόμορφα (ή σχεδόν ομοιόμορφα) τυχαία (π.χ. [Jamali and Ester, 2009]), με πίνακες μετάβασης παρόμοιους με αυτούς της Εξίσωσης 5.5. Στο προτεινόμενο σύστημα, ωστόσο, ακολουθείται μια άλλη στρατηγική και πιο συγκεκριμένα αυξάνεται η «μεροληψία» του συστήματος προς τους «καλύτερους» γείτονες. Σε αυτή την κατεύθυνση πραγματοποιείται ένας διαχωρισμός των γειτόνων ανάλογα με την ομοιότητα τους προς τον τρέχοντα χρήστη, με την αύξηση της πιθανότητας μετάβασης προς τους «όμοιους» (στο πλαίσιο των συστημάτων συστάσεων) χρήστες και την ταυτόχρονη μείωση της πιθανότητας μετάβασης προς τους λιγότερο «όμοιους» χρήστες.

Η προτεινόμενη μεθοδολογία δοκιμάστηκε σε δημόσια διαθέσιμες συλλογές συστημάτων συστάσεων (εμπλουτισμένες με ρητούς δεσμούς εμπιστοσύνης μεταξύ των μελών τους) και βρέθηκε να επιτυγχάνει μια καλή ισορροπία μεταξύ των μετρικών της ακρίβειας της πρόβλεψης από τη μία, και της καινοτομίας και ποικιλομορφίας των προτεινόμενων αντικειμένων από την άλλη.

6.1 Σχεδιαστικά ζητήματα

6.1.1 Ομοιότητα και εμπιστοσύνη

Στην Ενότητα 5.5 αναλύθηκε διεξοδικά η χρήση των δικτύων εμπιστοσύνης στα κοινωνικά συστήματα συστάσεων. Μια σημαντική, ωστόσο, παράμετρος που πρέπει να λαμβάνεται υπόψη αποτελεί το γεγονός πως η εμπιστοσύνη και η ομοιότητα είναι δύο έννοιες που δεν συμπίπτουν

απαραίτητα στα κοινωνικά συστήματα συστάσεων [Massa and Avesani, 2007]. Στο πλαίσιο των RS, δύο χρήστες θεωρούνται όμοιοι αν αξιολογούν τα ίδια αντικείμενα με τον ίδιο τρόπο. Ένα σύνολο μεγεθών, τα οποία προέρχονται από τη στατιστική βιβλιογραφία, μετρούν το πόσο κοντά είναι οι πληθυσμοί U_x και U_y (οι οποίοι συμβολίζουν τις βαθμολογίες των χρηστών $u_x \in U$ και $u_y \in U$ στο ίδιο σύνολο αντικειμένων I), ορισμένα εκ των οποίων παρουσιάστηκαν στην Ενότητα 2.3.1

Η στατιστική ομοιότητα έχει μελετηθεί εκτενώς στα συστήματα συστάσεων και έχει βρεθεί ότι ένας από τους πιο ικανοποιητικούς τρόπους μέτρησής της είναι ο συντελεστής συσχέτισης Pearson (Εξίσωση 2.7) [Herlocker et al., 2004], ο οποίος αποτελεί ένα καλό δείκτη της ομοιότητα όσο το πλήθος των αντικειμένων που έχουν βαθμολογήσει από κοινού οι u_x, u_y παραμένει σχετικά μεγάλο. Όμως αυτό δεν συμβαίνει πάντα στα συστήματα συστάσεων, όπου οι περισσότεροι χρήστες συμπίπτουν σε ένα πολύ μικρό σύνολο των αξιολογήσεων τους ή και καθόλου (Ενότητα 1.1). Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί και άλλες μέθοδοι ομοιότητας όπως ο *λογάριθμος του λόγου της πιθανοφάνειας* (Log-Likelihood Ratio - LLR) [Dunning, 1993] και η κανονικοποιημένη ομοιότητα Manhattan, που παρουσιάζεται παρακάτω

$$d_{u_x,u_y} = \frac{1}{n}\sum_{i=1}^{n}|U_{xi} - U_{yi}|,$$
$$s_{u_x,u_y} = 1 - d_{u_x,u_y} \tag{6.1}$$

Για την επιλογή του πιο περιγραφικού συντελεστή ομοιότητας, μετρήθηκε η απόδοση των τριών πιο διαδεδομένων σχετικών μετρικών (Pearson, LLR και κανονικοποιημένη Manhattan) στις συλλογές δεδομένων που χρησιμοποιήθηκαν στην πειραματική διαδικασία και οι οποίες παρουσιάζονται στον Πίνακα 6.1. Ο Πίνακας 6.2 συνοψίζει τα αποτελέσματα της σύγκρισης των προαναφερόμενων συντελεστών ομοιότητας όσον αφορά τρεις μετρικές ακρίβειας της πρόβλεψης. Από τη σύγκριση προκύπτει ότι η κανονικοποιημένη ομοιότητα Manhattan εμφανίζει την καλύτερη απόδοση στις συγκεκριμένες συλλογές δεδομένων και για αυτό το λόγο επιλέγεται η χρήση της ως συντελεστής ομοιότητας στο κοινωνικό σύστημα συστάσεων που θα παρουσιαστεί στο παρόν κεφάλαιο.

6.1.2 Τυχαίος περίπατος

Αφού έχει επιλεγεί η κατάλληλη μέτρηση ομοιότητας μεταξύ των χρηστών, θα μπορούσε να ενσωματωθεί άμεσα στον πίνακα μετάβασης του τυχαίου περιπάτου (Εξίσωση 5.4). Θεωρητικά, καλύτερες συστάσεις παράγονται με την μετάβαση σε πιο όμοιους γείτονες (από το να επιλέγεται κάποιος τυχαία). Επεκτείνοντας αυτή τη συλλογιστική, η ομοιότητα μεταξύ ενός χρήστη και τον άμεσων γειτόνων του στο δίκτυο εμπιστοσύνης θεωρείται ότι προέρχεται από μια άγνωστη συνάρτηση κατανομής η οποία μετράει το πόσο κοντά είναι δύο γείτονες στην αξιολογική τους συμπεριφορά. Μετακινούμενοι προς γείτονες με παρόμοιο γούστο (και όχι, γενικά, προς άλλους όμοιους χρήστες) αυξάνονται οι πιθανότητες λήψης ορθών συστάσεων.

Μια προφανής επιλογή θα ήταν η μετάβαση στον πιο όμοιο κάθε φορά γείτονα. Αυτή όμως δεν είναι και η πλέον πρόσφορη στρατηγική, κυρίως εξαιτίας τους γεγονότος ότι οι αξιολογήσεις δεν κατανέμονται ομοιόμορφα σε όλους τους χρήστες και τα αντικείμενα του συστήματος συστάσεων. Όπως έχει ήδη αναφερθεί στην Ενότητα 1.1, η κατανομή των χρηστών, των αντικειμένων και των δηλώσεων εμπιστοσύνης έχει τα χαρακτηριστικά των δικτύων ελεύθερης κλίμακας. Δηλαδή μια μειοψηφία χρηστών δίνει τις περισσότερες αξιολογήσεις ενώ αντίθετα η πλειοψηφία των χρηστών έχει αξιολογήσει πολύ λίγα αντικείμενα. Μια ντετερμινιστική στρατηγική σαν την προηγούμενη θα επέλεγε πάντα αυτή την μειοψηφία των χρηστών με αποτέλεσμα να καταλήγει να παράγει τετριμμένες συστάσεις. Αντίθετα, οι πιθανοτικοί αλγόριθμοι επιτρέπουν την πληρέστερη εξερεύνηση όλων των δυνατών επιλογών, συμβάλλοντας στην αύξηση της καινοτομίας των παραγόμενων συστάσεων.

Ένα τελευταίο ζήτημα αφορά το γεγονός ότι η κατανομή από την οποία πρέπει να αντληθούν τα δείγματα (ο επόμενος κόμβος του περιπάτου) παραμένει άγνωστη. Για τον λόγο αυτό, σε ένα πρώτο στάδιο οι μετρήσεις ομοιότητας μετατρέπονται σε πιθανότητες (διαιρώντας κάθε μια από αυτές με το άθροισμα τους) και κατόπιν χρησιμοποιείται ένας κανόνας αποδοχή/απόρριψης για την παραγωγή των δειγμάτων. Η μεθοδολογία αυτή θα αναπτυχθεί περαιτέρω στην επόμενη ενότητα.

6.2 Αλγόριθμος

Ο προτεινόμενος αλγόριθμος τυχαίου περιπάτου λειτουργεί σε τρία στάδια. Αρχικά, διαιρούνται οι άμεσοι γείτονες του υπό εξέταση χρήστη u_t σε δύο διαφορετικά σύνολα. Το πρώτο είναι το σύνολο των *ομοίων γειτόνων* SN, το οποίο περιέχει όλους εκείνους τους γείτονες στο δίκτυο εμπιστοσύνης που έχουν τουλάχιστον μια κοινή βαθμολογία με τον u_t. Το δεύτερο είναι σύνολο των *γενικών γειτόνων* GN, στο οποίο εντάσσονται όλοι εκείνοι οι χρήστες τους οποίους ναι μεν εμπιστεύεται ο u_t, αλλά δεν έχει αξιολογήσει αντικείμενα από κοινού με αυτούς. Στο επόμενο στάδιο, λαμβάνεται η απόφαση για την επιλογή του συνόλου γειτόνων από το οποίο θα πραγματοποιηθεί η δειγματοληψία. Το ζητούμενο σύνολο επιλέγεται με την εφαρμογή ενός απλού κανόνα *αποδοχής/απόρριψης* (acceptance/rejection rule) δύο βημάτων που περιγράφεται παρακάτω:

1. Δειγματοληψία τιμής $u \sim \mathcal{U}(0,1)$ (ομοιόμορφη κατανομή στο $(0,1)$)
2. Έλεγχος ανίσωσης $u < \frac{|GN|}{|SN|+|GN|}$
   - Αν η ανίσωση **ισχύει**, επιλογή επόμενου γείτονα από το GN με ομοιόμορφα τυχαίο τρόπο
   - Αν η ανίσωση **δεν ισχύει**, επιλογή επόμενου γείτονα από το SN, σύμφωνα με τη μέθοδο απορριπτικής δειγματοληψίας που περιγράφεται στην Ενότητα 6.2.1

6.2.1 Απορριπτική Δειγματοληψία

Όταν επιλέγεται το επόμενο βήμα να προέρχεται από τους όμοιους γείτονες (SN), θα πρέπει οι μετρικές της ομοιότητας να μετατραπούν σε πιθανότητες μετάβασης. Για να πραγματοποιηθεί η μετατροπή αυτή, γίνεται η θεώρηση ότι οι συγκεκριμένες τιμές ομοιότητας που έχουν οι γείτονες του υπό εξέταση χρήστη u_t παράγονται από μια άγνωστή πιθανοτική γεννήτρια, η οποία υπακούει σε μια επίσης άγνωστη κατανομή $\mathcal{F}(x)$. Άρα, στόχος είναι να προσεγγιστεί η $\mathcal{F}$, και πιο συγκεκριμένα η συνάρτηση πυκνότητας πιθανότητας (σ.π.π) της f, έτσι ώστε τελικά να υπολογιστούν οι πιθανότητες μετάβασης.

Για να λάβουμε δείγματα από μια άγνωστη κατανομή $\mathcal{F}$, μπορούμε να χρησιμοποιήσουμε την τεχνική της *απορριπτικής δειγματοληψίας* (rejection sampling) [Andrieu et al., 2003]. Η φιλοσοφία της είναι να χρησιμοποιηθεί μια κατανομή $\mathcal{G}(x)$, από την οποία είναι εύκολο να παραχθούν δείγματα, ως εργαλείο για τη λήψη δειγμάτων από την άγνωστη κατανομή $\mathcal{F}(x)$. Η $\mathcal{G}(x)$ αναφέρεται επίσης και ως *κατανομή πρότασης* (proposal distribution). Αν $f(x)$, $g(x)$ είναι οι αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας, τότε το μόνο προαπαιτούμενο της συγκεκριμένης μεθόδου είναι η *υποστήριξη* (support) της $g(x)$ να καλύπτει την υποστήριξη της $f(x)$ μέχρι έναν συντελεστή αναλογίας c κατ΄ ελάχιστο. Δηλαδή, πρέπει να ισχύει η ακόλουθη ανισότητα:

$$f(x) \leq cg(x), \quad c < \infty, \quad \forall x \in \mathcal{X} \tag{6.2}$$

όπου $\mathcal{X}$ είναι ο δειγματικός χώρος.

Στη συνέχεια, ένας αριθμός u δειγματοληπτείται ομοιόμορφα τυχαία από την $\mathcal{U}(0,1)$ μαζί με ένα δείγμα $x_i \in \mathcal{X}$ σύμφωνα με την $\mathcal{G}(x)$ ($x_i \sim \mathcal{G}(x)$). Κατόπιν, η ελέγχεται η ανισότητα

$u < \frac{f(x_i)}{cg(x_i)}$ ως προς την εγκυρότητά της. Αν ισχύει, τότε το x_i θεωρείται ως ένα έγκυρο δείγμα το οποίο έχει ληφθεί από την $f(x)$, διαφορετικά απορρίπτεται και νέα δείγματα u, x_i επιλέγονται από τις αντίστοιχες κατανομές. Η λειτουργία της μεθόδου απεικονίζεται στον Αλγόριθμο 3.

Αλγόριθμος 3 Γενικός Αλγόριθμος Απορριπτικής Δειγματοληψίας

Απαίτησε κατανομή πρότασης $\mathcal{G}(x)$
Απαίτησε σ.π.π. $g(x)$ της $\mathcal{G}(x)$
Απαίτησε σ.π.π. $f(x)$ της κατανομής-στόχου
Απαίτησε σταθερά αναλογίας $c < \infty$
1: **Επανάληψη**
2: Δειγματοληψία $u \sim \mathcal{U}(0,1)$
3: Δειγματοληψία $x_i \sim \mathcal{G}(x)$
4: **Μέχρις ότου** $u < \frac{f(x_i)}{c.g(x_i)}$
5: **Επιστροφή** x_i

Ο προτεινόμενος αλγόριθμος παραγωγής συστάσεων (*Biased RW-RS*) πραγματοποιεί ένα μη-αμερόληπτο τυχαίο περίπατο εφαρμόζοντας την μέθοδο της απορριπτικής δειγματοληψίας, που μόλις περιγράφηκε, για να καθορίσει το επόμενο βήμα του στην περίπτωση που αυτό πρέπει να επιλεγεί από το σύνολο των όμοιων χρηστών SN. Στην περίπτωση του παρόντος κεφαλαίου ο συντελεστής ομοιότητας που χρησιμοποιείται είναι η κανονικοποιημένη ομοιότητα Manhattan (Εξίσωση 6.1) που λαμβάνει τιμές στο στο $[0, 1]$. Έτσι, οι τιμές της σ.π.π $f(x)$ για τους γείτονες του u_t μπορεί να θεωρηθεί ότι προκύπτουν από τη διαίρεση της τιμής ομοιότητας για τον εκάστοτε γείτονα με το άθροισμα των τιμών της ομοιότητας για όλους τους γείτονες. Η ομοιόμορφη κατανομή $\mathcal{U}(x)$ επιλέγεται ως η κατανομή πρότασης, ενώ η σταθερά c προσεγγίζεται με την εξασφάλιση ότι η ανισότητα $f(x) < c.u(x)$ ισχύει σε κάθε σημείο. Η μεθοδολογία περιγράφεται αναλυτικότερα στην συνάρτηση *RejectionSampling* (Αλγόριθμος 4)

6.2.2 Τερματισμός του περιπάτου

Μια σημαντική απόφαση που μένει να ληφθεί είναι ο καθορισμός της στιγμής τερματισμού του τυχαίου περιπάτου. Αν αυτός τερματιστεί νωρίς, τότε το σύστημα συστάσεων δεν θα προλάβει να εξερευνήσει καλά τον χώρο των χρηστών. Είναι, όμως, περισσότερο πιθανό να παραχθούν συστάσεις οι οποίες θα γίνουν αποδεκτές από τον δοσμένο χρήστη u_t (λ.χ. αντικείμενα που έχουν αξιολογηθεί από χρήστες που ο u_t εμπιστεύεται ή αντικείμενα όμοια με αυτά που και ο ίδιος έχει αξιολογήσει). Από την άλλη, οι μακρείς περίπατοι επιτρέπουν την καλύτερη εξερεύνηση του χώρου των χρηστών, αλλά όσο περισσότερο ο περίπατος απομακρύνεται από τον u_t τόσο λιγότερο πιθανό καθίσταται το να βρεθούν παρόμοιοι χρήστες με αυτόν.

Συνεπώς, θα πρέπει να επιτευχθεί μια ισορροπία μεταξύ εξερεύνησης και ομοιότητας. Ένα καλό κριτήριο τερματισμού σίγουρα θα πρέπει να ενσωματώνει αρκετή πληροφορία, όπως το πόσο μακρυά βρίσκεται ο περίπατος από την αρχή του ή πόσο όμοια είναι η περιοχή που βρίσκεται τώρα ο περίπατος με τις προτιμήσεις του δοσμένου χρήστη u_t. Άλλα, ακόμα περισσότερο εκλεπτυσμένα, κριτήρια μπορούν να περιλαμβάνουν μια υψηλού επιπέδου ανάλυση του δικτύου εμπιστοσύνης για τον εντοπισμό συστάδων χρηστών και αντικειμένων. Αφού εντοπιστούν οι πιθανές συστάδες στις οποίες εντάσσεται ο u_t, ο τυχαίος περίπατος μπορεί να τερματιστεί αφού τις ερευνήσει εξαντλητικά.

Μιας και στις συλλογές δεδομένων που χρησιμοποιήθηκαν στα πειράματα, η πυκνότητα των βαθμολογιών και η συνεκτικότητα του κοινωνικού γράφου είναι πολύ αραιές (Πίνακας 6.1),

Πίνακας 6.1: Συλλογές δεδομένων που χρησιμοποιήθηκαν στα πειράματα

	`Filmtrust`	`Epinions`
Χρήστες	1.919	49.290
Αντικείμενα	2.018	139.738
Πλήθος Αξιολογήσεων	33.526	664.824
Πυκνότητα Βαθμολογιών	1,15%	0,01%
Πλήθος Δεσμών Εμπιστοσύνης	1.591	487.182
Ολικός Συντελεστής Συσταδοποίησης	0,0004	0,0002

επιλέχθηκε ένα απλό, πιθανοτικό, κριτήριο τερματισμού. Σε κάθε βήμα ορίζεται μια σταθερή πιθανότητα P_c συνέχισης του περιπάτου καθώς και μια σταθερή πιθανότητα $P_t = 1 - P_c$ τερματισμού του. Συνεπώς, η πιθανότητα συνέχισης του περιπάτου μετά από k βήματα ορίζεται από την κατανομή Bernoulli, όπως παρακάτω

$$p(k) = (1 - P_c)P_c^k \tag{6.3}$$

6.2.3 Παραγωγή Συστάσεων

Αφού ο τυχαίος περίπατος τερματιστεί, χρησιμοποιείται ένας απλός κανόνας για να παραχθούν οι συστάσεις. Έστω ότι χρειάζεται να εκτιμηθεί η χρησιμότητα του αντικειμένου i_t για τον δοσμένο χρήστη u_t. Αν ο τυχαίος περίπατος τερματίσει σε ένα χρήστη u_c, ο οποίος έχει βαθμολογήσει το i_t, τότε επιστρέφεται ως σύσταση η τιμή:

$$\widehat{r_{u_t,i_t}} = \overline{u_t} + \text{sim}(u_t, u_c) \times (\overline{u_t} - r_{u_t,i_t}) \tag{6.4}$$

όπου $\overline{u_t}$, $\overline{u_c}$ η μέση τιμή των αξιολογήσεων που έχουν κάνει μέχρι στιγμής οι χρήστες u_t και u_c (και sim ο συντελεστής ομοιότητάς τους, δηλαδή η κανονικοποιημένη ομοιότητα Manhattan). Διαφορετικά επιστρέφεται η αξιολόγηση του περισσότερο όμοιου προς το i_t αντικειμένου που έχει δώσει ο χρήστης u_c.

6.3 Πειράματα

6.3.1 Συλλογές Δεδομένων

Η απόδοση του *Biased RW-RS* αποτιμάται σε δύο διαφορετικές συλλογές δεδομένων. Η πρώτη προέρχεται από την υπηρεσία Filmtrust [Golbeck and Hendler, 2006] και περιέχει 33.526 αξιολογήσεις που έχουν δοθεί από 1.919 χρήστες σε 2.018 αντικείμενα καθώς και 1.591 μη βαθμονομημένες δηλώσεις εμπιστοσύνης. Οι αξιολογήσεις κυμαίνονται στο διάστημα 0,5 ως 4 με βήμα 0,5 (μεγαλύτερη τιμή αξιολόγησης υποδηλώνει περισσότερη προτίμηση). Η δεύτερη συλλογή δεδομένων προέρχεται από την υπηρεσία Epinions [Massa and Avesani, 2006] και συλλέχθηκε από τους [Massa and Bhattacharjee, 2004]. Αποτελείται από 664.824 αξιολογήσεις που έχουν καταχωρηθεί από 49.290 χρήστες σε 139.738 αντικείμενα και από 487.182 δηλώσεις εμπιστοσύνης. Οι αξιολογήσεις κυμαίνονται στο ακέραιο διάστημα 1 ως 5 (το 1 υποδηλώνει την ισχυρή απόρριψη, το 5 την ισχυρή αποδοχή) ενώ και σε αυτή την περίπτωση οι δηλώσεις προτίμησης δεν είναι βαθμονομημένες. Και οι δύο συλλογές δεδομένων παρουσιάζονται με περισσότερες λεπτομέρειες στον Πίνακα 6.1.

Η πυκνότητα της πληροφορίας εμπιστοσύνης μετριέται με την χρήση του *ολικού συντελεστή συσταδοποίησης* (global clustering coefficient), ο οποίος μετράει τη *συνδεσιμότητα* (connectedness) του γράφου εμπιστοσύνης (πόσο απέχει, δηλαδή, από το να μετατραπεί σε

Αλγόριθμος 4 Σύστημα Biased RW-RS

Εξασφάλισε Πιθανότητα τερματισμού $P_t = 0.1$
Απαίτησε Χρήστης υπό εξέταση u_t, Αντικείμενο υπό εξέταση i_t

$u_c \leftarrow u_t$
Επανάληψη
  $(SN, GN) \leftarrow$ SplitNeighbors(u_c)
  Δειγματοληψία $u \sim \mathcal{U}(0,1)$
  Αν $u \leq \frac{|GN|}{|SN|+|GN|}$ **Τότε**
    Δειγματοληψία u_c ομοιόμορφα τυχαία από το σύνολο γενικών γειτόνων GN
  Αλλιώς
    Δειγματοληψία $u_c \sim$ RejectionSampling(S, u_c)
  Τέλος Αν
  Αν ο u_c έχει αξιολογήσει το i_t **Τότε**
    επίστρεψε $\overline{r_{u_t}} + \text{Sim}(u_c, u_t) \times (r_{u_c,i_t} - \overline{r_{u_t}})$
  Τέλος Αν
  Δειγματοληψία $u \sim \mathcal{U}(0,1)$
Μέχρις όπου $u \leq P_t$
Για $\forall i_c \in I_{u_c}$ **κάνε** ▷ Όλα τα αντικείμενα που έχει αξιολογήσει ο u_c
  $s_{i_c,i_t} \leftarrow$ ManhattanSimilarity(i_c, i_t)
Τέλος Για
Ως i_r τίθεται το αντικείμενο με $\max s_{i_c,i_t}$
επίστρεψε r_{u_c,i_r}

Συνάρτηση RejectionSampling(S, u_c)
  $sum \leftarrow 0$
  Για $u_n \in N$ **κάνε**
    $s_{u_c,u_n} \leftarrow$ ManhattanSimilarity(u_c, u_n)
    $sum \leftarrow sum + s_{u_c,u_n}$
  Τέλος Για
  $\mathcal{G}(x) \leftarrow \mathcal{U}(\min s_{u_c,u_n}, \max s_{u_c,u_n})$
  $c \leftarrow 0$
  Για $u_n \in N$ **κάνε** ▷ Μετατροπή της ομοιότητας σε συνάρτηση κατανομής
    $f(u_n) \leftarrow \frac{s_{u_c,u_n}}{sum}$
    Αν $c < \frac{f(u_n)}{g(u_n)}$ **Τότε**
      $c \leftarrow \frac{f(u_n)}{g(u_n)}$
    Τέλος Αν
  Τέλος Για
  Επανάληψη
    Δειγματοληψία $u \sim \mathcal{U}(0,1)$
    Δειγματοληψία $x_i \sim \mathcal{G}(x)$
  Μέχρις όπου $u < \frac{f(x_i)}{c.g(x_i)}$
  επίστρεψε x_i
Τέλος Συνάρτηση

κλίκα (clique)) και προκύπτει από το λόγο

$$\text{Connectedness} = \frac{|\text{TrustStatements}|}{|\text{Users}| \times (|\text{Users}| - 1)} \tag{6.5}$$

Τέλος και οι δύο συλλογές δεδομένων είναι ιδιαίτερα αραιές, όπως και τα αντίστοιχα δίκτυα εμπιστοσύνης, ακολουθώντας το γνωστό πλέον παράδειγμα των δικτύων ελεύθερης κλίμακας.

6.3.2 Πειραματικό Πρωτόκολλο

Το εφαρμοζόμενο πειραματικό πρωτόκολλο είναι η *10-πτυχή διασταυρωμένη επικύρωση* (10-fold cross validation). Τα δεδομένα που περιέχονται στο σύστημα συστάσεων αναπαρίσταται υπό μορφή γράφου $G = (U, I, R, T)$, όπου U είναι το σύνολο των κόμβων-χρηστών, I το σύνολο των κόμβων-αντικειμένων, R το σύνολο των κατευθυντικών ακμών των αξιολογήσεων που έχουν την αφετηρία τους στο U και τον προορισμό τους στο I, T το σύνολο των κατευθυντικών ακμών των δηλώσεων εμπιστοσύνης, που έχουν και αφετηρία και προορισμό στο U. Οι ακμές στο R χωρίζονται σε 10 μη-επικαλυπτόμενα σύνολα R_i από τα οποία προκύπτουν οι επαγόμενοι υπογράφοι G_i που περιέχουν τους αντίστοιχους χρήστες, αντικείμενα και δηλώσεις εμπιστοσύνης. Τα σύνολα U_i, I_i και T_i δεν είναι απαραίτητα μη-επικαλυπτόμενα μεταξύ τους.

$$\begin{aligned}
G_i &= (U_i, I_i, R_i, T_i), \\
U_i &= \{\forall u \in U : \exists i \in Is.t(u, i) \in R_i\}, \\
I_i &= \{\forall i \in I : \exists u \in Us.t(u, i) \in R_i\}, \\
T_i &= \{(u_a, u_b) \in T : \forall u_a \in U_i \vee \forall u_b \in U_i\}
\end{aligned}$$

Σε κάθε βήμα του πρωτοκόλλου, 9 από αυτούς τους υπογράφους ενώνονται για τον σχηματισμό του συνόλου εκπαίδευσης ενώ αυτός που απομένει ορίζεται ως το σύνολο δοκιμής. Κατόπιν, ο αλγόριθμος Biased RW-RS χρησιμοποιεί το σύνολο εκπαίδευσης για να προβλέψει την βαθμολογία που δίνουν οι χρήστες του συνόλου εκπαίδευσης στα αντικείμενα του ιδίου συνόλου.

Η απόδοση του συστήματος μετριέται σε δύο σύνολα μετρικών. Το πρώτο περιλαμβάνει τρεις μετρικές ακρίβειας της πρόβλεψης (Ενότητα 3.2.1) και πιο συγκεκριμένα το RMSE (Εξίσωση 3.3), το MAE (Εξίσωση 3.4), το MAUE (Εξίσωση 3.7) και τέλος την κάλυψη των αξιολογήσεων (Εξίσωση 3.23). Τα αποτελέσματα των προαναφερόμενων μετρικών περιλαμβάνονται στον Πίνακα 6.3 για όλους τους χρήστες και στον Πίνακα 6.4 για την όψη χρηστών με λίγες αξιολογήσεις (Ενότητα 3.5.1).

Το δεύτερο σύνολο μετρικών περιλαμβάνει την ακρίβεια ταξινόμησης (Εξίσωση 3.8), την κάλυψη των χρηστών (Εξίσωση 3.24), την ποικιλομορφία της λίστας των προτεινόμενων αντικειμένων (Εξίσωση 3.20) και τέλος την καινοτομία του αντικειμένου (Εξίσωση 3.16). Τα αποτελέσματα των συγκεκριμένων μετρικών απεικονίζονται στο Σχήμα 6.1 για το `Epinions` και στο Σχήμα 6.1 για το `Filmtrust`.

6.3.3 Άλλες Υλοποιήσεις

Για την πληρέστερη κατανόηση της απόδοσης του αλγορίθμου Biased RW-RS, μετρήθηκαν στις ίδιες συλλογές δεδομένων και ακολουθώντας το ίδιο πειραματικό πρωτόκολλο ορισμένα ακόμα συστήματα συστάσεων, τόσο παραδοσιακά όσο και κοινωνικά. Αρχικά, τα βασικά συστήματα παρουσιάζονται για λόγους αναφοράς, για να εκτιμηθεί δηλαδή η σχετική βελτίωση που επιτυγχάνουν τα υπόλοιπα συστήματα ως προς αυτά. Η *Τυχαία Σύσταση* προκύπτει από ένα σύστημα το οποίο παράγει μια ομοιόμορφα κατανεμημένη τυχαία τιμή (εντός της αξιολογικής κλίμακας) ως σύσταση για κάθε χρήστη, ενώ το σύστημα της *Πάντα Μέγιστης Τιμής* προτείνει όλα τα αντικείμενα με τη μέγιστη δυνατή τιμή ωφέλειας.

Πίνακας 6.2: Σύγκριση των συντελεστών ομοιότητας

Συλλογή Δεδομένων		Filmtrust				Epinions			
Μέτρηση Απόδοσης		RMSE	MAE	MAUE	Κάλυψη	RMSE	MAE	MAUE	Κάλυψη
Α.	**Συνεργατική Διήθηση**								
Α.1	Pearson	1,09	0,81	0,81	92,08%	1,39	1,01	1,07	40,13%
Α.2	Log likelihood ratio	0,88	0,70	0,69	93,73%	1,10	0,84	0,85	62,30%
Α.3	Manhattan	0,88	0,70	0,68	93,65%	1,07	0,81	0,82	79,57%
Β.	**Συνεργατική Διήθηση βασισμένη στο αντικείμενο**								
Β.1	Pearson	0,84	0,63	0,63	91,79%	1,75	1,31	1,33	39,29%
Β.2	Log likelihood ratio	0,81	0,62	0,62	95,49%	1,27	1,00	1,04	63,99%
Β.3	Manhattan	0,80	0,61	0,61	99,22%	1,20	0,92	0,95	97,35%

Πίνακας 6.3: Αποτελέσματα μετρικών ακρίβειας πρόβλεψης (όλοι οι χρήστες)

Συλλογή Δεδομένων		Filmtrust				Epinions			
Μέτρηση Απόδοσης		RMSE	MAE	MAUE	Κάλυψη	RMSE	MAE	MAUE	Κάλυψη
Α.	**Βασικά Συστήματα**								
Α.1	Τυχαία Σύσταση	1,53	1,25	1,26	100,00%	1,94	1,61	1,63	100,00%
Α.2	Πάντα Μέγιστη Τιμή	1,35	1,00	0,90	100,00%	1,57	1,01	0,97	100,00%
Β.	**Συνεργατική Διήθηση**								
Β.1	Ομοιότητα Manhattan (Όλοι οι γείτονες)	0,88	0,70	0,68	93,65%	1,07	0,81	0,82	79,57%
Γ.	**Συνεργατική Διήθηση βασισμένη στο αντικείμενο**								
Γ.1	Ομοιότητα Manhattan (Πλησιέστερα N Αντικείμενα)	0,78	0,60	0,61	72,71%	1,37	0,99	1,00	22,92%
Δ.	**Συστήματα βασισμένα στην εμπιστοσύνη**								
Δ.1	MoleTrust-1	0,97	0,73	0,74	18,64%	1,23	0,91	0,95	25,58%
Δ.2	MoleTrust-2	0,91	0,70	0,72	24,76%	1,16	0,88	0,93	56,52%
Δ.3	MoleTrust-3	0,89	0,69	0,70	27,14%	1,12	0,85	0,89	70,89%
Δ.4	TidalTrust	0,96	0,73	0,74	27,86%	1,08	0,82	0,83	74,67%
Ε.	**Το Σύστημά μας**								
Ε.1	Biased RW-RS	0,78	0,61	0,59	92,61%	1,07	0,82	0,83	53,43%

Τα συστήματα της *Συνεργατικής Διήθησης* (τόσο τα απλά όσο και αυτά που είναι βασισμένα στο αντικείμενο) είναι απλά μνημονικά συστήματα (Ενότητα 2.3.1) τα οποία χρησιμοποιούν τον τύπο Resnick (Εξίσωση 2.11) για την παραγωγή των συστάσεων. Η συνάρτηση ομοιότητας που χρησιμοποιείται και στις δύο περιπτώσεις είναι η ομοιότητα Manhattan (Ενότητα 6.1.1) με τη διαφορά ότι στον απλό αλγόριθμο συνεργατικής διήθησης λαμβάνονται υπόψη όλοι οι όμοιοι χρήστες για την παραγωγή συστάσεων ενώ αντίθετα στον αλγόριθμο συνεργατικής διήθησης βασισμένης στο αντικείμενο επιλέγονται τα 5 πλησιέστερα αντικείμενα ($N = 5$). Οι συγκεκριμένες επιλογές έγιναν γιατί σε αυτές τις ρυθμίσεις βρέθηκε ότι ελαχιστοποιείται το σφάλμα των μετρικών ακρίβειας της πρόβλεψης.

Τέλος, υλοποιήθηκαν και συστήματα συστάσεων βασισμένα στην εμπιστοσύνη και πιο συγκεκριμένα οι μετρικές τοπικής εμβέλειας MoleTrust (με ορίζοντα διάδοσης από 1 ως 3) και TidalTrust, οι οποίες παρουσιάστηκαν αναλυτικά στην Ενότητα 5.5.2.

6.4 Αποτελέσματα

6.4.1 Μετρικές Ακρίβειας της Πρόβλεψης

Μια πρώτη παρατήρηση των αποτελεσμάτων (Πίνακες 6.3-6.4) φανερώνει ότι ο αλγόριθμος Biased RW-RS εμφανίζει την καλύτερη απόδοση σε όλες τις μετρικές απόδοσης της όψης των

Πίνακας 6.4: Αποτελέσματα μετρικών ακρίβειας πρόβλεψης (χρήστες με λίγες αξιολογήσεις)

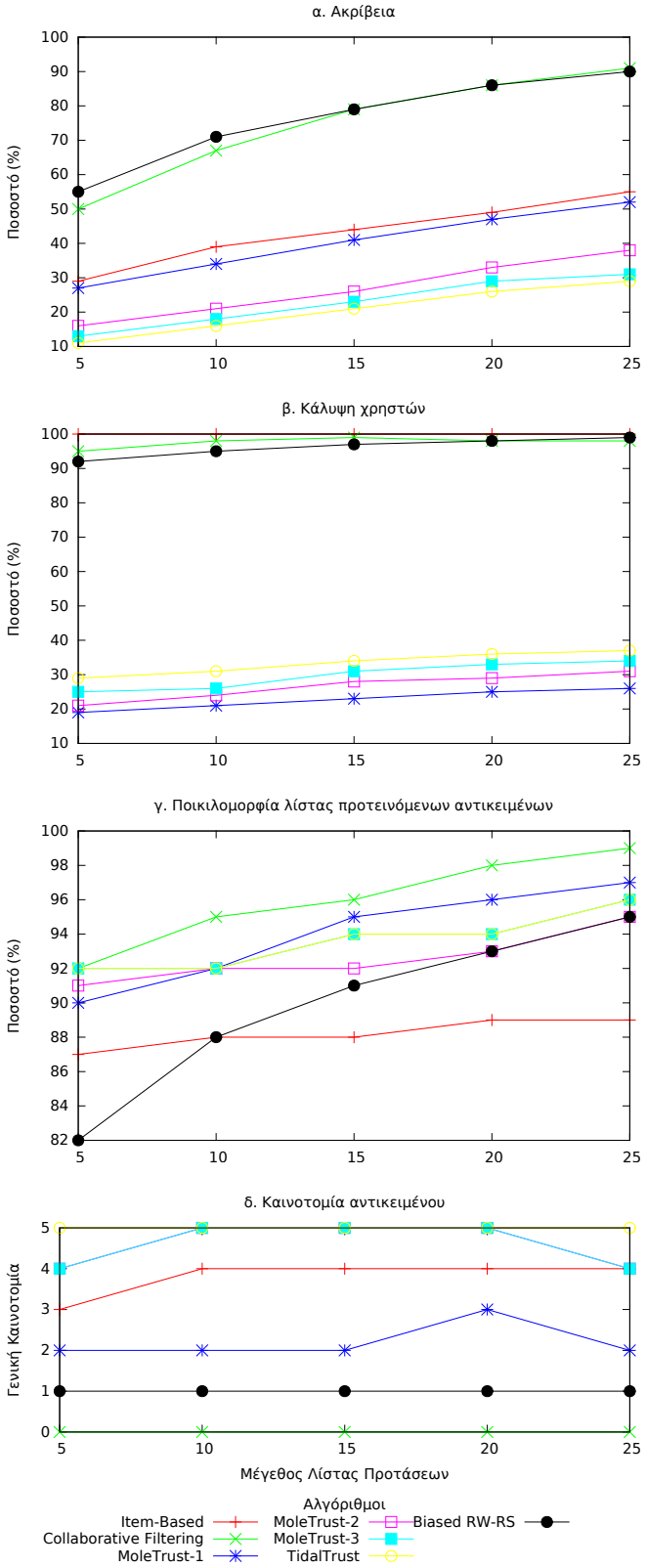
Συλλογή Δεδομένων		Filmtrust				Epinions			
Μέτρηση Απόδοσης		RMSE	MAE	MAUE	Κάλυψη	RMSE	MAE	MAUE	Κάλυψη
Α.	**Βασικά Συστήματα**								
Α.1	Τυχαία Σύσταση	1,51	1,22	1,22	100,00%	2,00	1,67	1,67	100,00%
Α.2	Πάντα Μέγιστη Τιμή	0,80	0,49	0,51	100,00%	1,56	0,94	0,93	100,00%
Β.	**Συνεργατική Διήθηση**								
Β.1	Ομοιότητα Manhattan (Όλοι οι γείτονες)	0,80	0,64	0,63	82,98%	1,09	0,82	0,82	69,46%
Γ.	**Συνεργατική Διήθηση βασισμένη στο αντικείμενο**								
Γ.1	Ομοιότητα Manhattan (Πλησιέστερα Ν Αντικείμενα)	0,77	0,63	0,64	72,60%	1,58	1,09	1,08	9,21%
Δ.	**Συστήματα βασισμένα στην εμπιστοσύνη**								
Δ.1	MoleTrust-1	1,46	1,20	1,02	10,94%	1,49	1,09	1,09	7,49%
Δ.2	MoleTrust-2	1,71	1,33	1,08	20,41%	1,53	1,17	1,17	24,27%
Δ.3	MoleTrust-3	1,22	0,87	1,33	24,56%	1,06	0,82	1,08	76,25%
Δ.4	TidalTrust	1,22	0,87	0,87	26,23%	1,11	0,84	0,84	42,00%
Ε.	**Το Σύστημά μας**								
Ε.1	Biased RW-RS	0,83	0,62	0,61	76,92%	1,1	0,85	0,86	40,29%

δεδομένων που περιέχει όλους τους χρήστες. Τα κοινωνικά συστήματα συστάσεων, από την άλλη, εμφανίζουν ιδιαίτερα χαμηλή *κάλυψη αξιολογήσεων* και αυτό οφείλεται στο γεγονός ότι το δίκτυο εμπιστοσύνης είναι πολύ αραιό, πράγμα που επηρεάζει τη δυνατότητα τους να ανακαλύπτουν «έμπιστους» χρήστες. Ο αλγόριθμος Biased RW-RS κατορθώνει να υπερκεράσει αυτό το πρόβλημα, αποφασίζοντας πιθανοτικά σε κάθε του βήμα την επιλογή είτε ενός γείτονα ή ενός όμοιου χρήστη. Για αυτό το λόγο, είναι κατά πολύ ανώτερος, όσον αφορά την κάλυψη των αξιολογήσεων από τα άλλα κοινωνικά συστήματα συστάσεων στο `Filmtrust` και περίπου στο μέσο τους όρο στο `Epinions`.

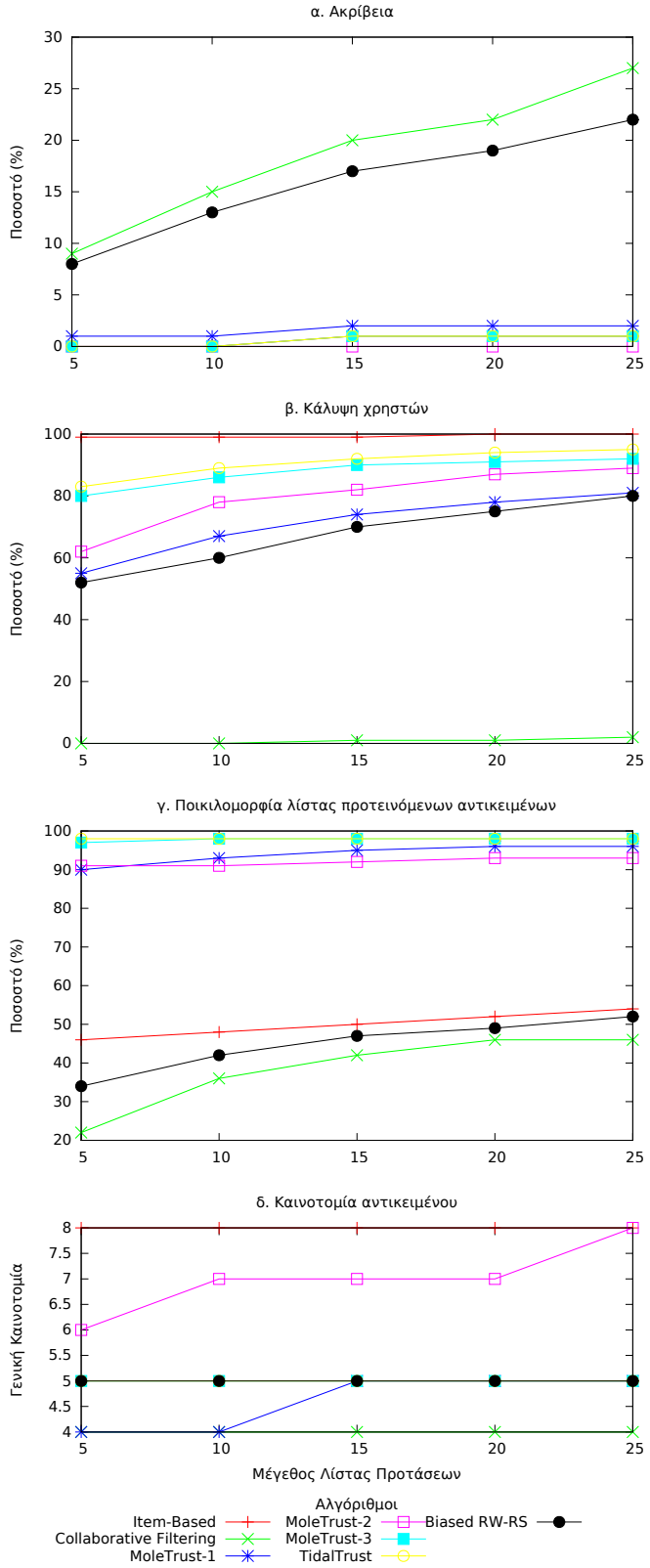
Επιπλέον, για τους χρήστες με λίγες αξιολογήσεις, ο αλγόριθμος Biased RW-RS εμφανίζει ικανοποιητικά αποτελέσματα στο `Filmtrust` ενώ αντίθετα η απόδοση των υπολοίπων συστημάτων συστάσεων επιδεινώνεται εμφανώς. Στο `Epinions`, ο ίδιος αλγόριθμος κατορθώνει να κρατήσει μια σταθερή απόδοση όσον αφορά το MAE και το RMSE, ενώ την ίδια στιγμή εξασφαλίζει μια ανεκτή κάλυψη των αξιολογήσεων του συνόλου δοκιμής.

6.4.2 Άλλες μετρικές

Ο αλγόριθμος *Biased RW-RS* αποτελεί την πιο αποδοτική κοινωνική μέθοδο όσον αφορά την μετρική της ακρίβειας ταξινόμησης και στις δύο συλλογές δεδομένων (Σχήματα 6.1 και 6.2). Παρότι η κλασσική μέθοδος συνεργατικής διήθησης φαίνεται να είναι οριακά καλύτερη στο `Epinions` (περίπου 1%-2% για όλα τα μεγέθη της λίστας παραγόμενων συστάσεων), παρουσιάζει ωστόσο πολύ άσχημα αποτελέσματα όσον αφορά την μετρική της κάλυψης των χρηστών, πράγμα το οποίο σημαίνει ότι μπορεί να παράγει ικανοποιητικές συστάσεις μόνο για μια πολύ μικρή μερίδα χρηστών. Τα συστήματα εμπιστοσύνης, από την άλλη, έχουν την δυνατότητα κάλυψης περισσότερων χρηστών. Παρόλα αυτά, οι προβλέψεις τους απέχουν πολύ από το να είναι ακριβείς (η ακρίβεια είναι κάτω του 5% για όλα τα συστήματα εμπιστοσύνης στο `Epinions` και κάτω του 40% στο `FilmTrust`) και αυτό αποδίδεται στο γεγονός ότι οι συσχετίσεις στην αξιολογική συμπεριφορά των χρηστών δεν λαμβάνονται καθόλου υπόψη. Άλλη μια σημαντική παρατήρηση για τις μετρικές εμπιστοσύνης είναι ότι η κάλυψη των αξιολογήσεων που επιτυγχάνουν στο `FilmTrust` είναι περίπου στο ένα τρίτο των αντίστοιχων μεγεθών στο `Epinions`, παρότι το δίκτυο εμπιστοσύνης του πρώτου είναι πυκνότερο από του



Σχήμα 6.1: Αποτελέσματα μετρικών ακρίβειας της ταξινόμησης στο `Filmtrust`



Σχήμα 6.2: Αποτελέσματα μετρικών ακρίβειας της ταξινόμησης στο `Epinions`

δεύτερου (όπως αποτυπώνεται και στον δείκτη του ολικού συντελεστή συσταδοποίησης του Πίνακα 6.1). Το φαινόμενο αυτό αποδίδεται στο γεγονός πως μόνο το 38% των χρηστών του `FilmTrust` είναι ενεργοί στο δίκτυο εμπιστοσύνης τη στιγμή που το ίδιο μέγεθος στο `Epinions` ανέρχεται στο 68%. Συνεπώς, ένα βασικό συμπέρασμα είναι πως ακόμα και η μικρότερη συμμετοχή των χρηστών στο δίκτυο εμπιστοσύνης δίνει τη δυνατότητα στα SRS να παράγουν συστάσεις.

Οι μέθοδοι εμπιστοσύνης σημειώνουν τα καλύτερα αποτελέσματα τόσο όσον αφορά την καινοτομία των αντικειμένων που προτείνονται όσο και την ποικιλομορφία των αντικειμένων στις επιστρεφόμενες λίστες προτάσεων. Ωστόσο, σε αυτό το γεγονός πρέπει να συνυπολογιστεί και η ακρίβεια των συστάσεων· οι πρωτότυπες και οι ποικιλόμορφες προβλέψεις έχουν ελάχιστη χρησιμότητα αν δεν εφάπτονται των ενδιαφερόντων του τελικού χρήστη. Από την άλλη, οι μέθοδοι που βασίζονται στην εύρεση συσχετίσεων (*Συνεργατική Διήθηση* είτε απλή είτε βασισμένη στο αντικείμενο) πραγματοποιούν τετριμμένες συστάσεις, με τα προτεινόμενα αντικείμενα σε μεγάλο βαθμό να μοιάζουν μεταξύ τους. Σε αυτό το λόγο οφείλονται και τα πολύ κακά αποτελέσματα που επιτυγχάνουν στη μετρική της καινοτομίας.

Συνολικά, τα αποτελέσματα δείχνουν πως το προτεινόμενο στο παρόν κεφάλαιο σύστημα επιτυγχάνει ακρίβεια πρόβλεψης σε παρόμοιο επίπεδο με τον κλασσικό αλγόριθμο *Συνεργατικής Διήθησης* ενώ την ίδια στιγμή οι παραγόμενες συστάσεις εμφανίζουν μεγαλύτερη καινοτομία και ποικιλομορφία, γεγονός που οφείλεται στη συμπερίληψη της κοινωνικής πληροφορίας στον μηχανισμό παραγωγής συστάσεων.

6.5 Συμπεράσματα

Στις προηγούμενες ενότητες παρουσιάστηκε μια νέα προσέγγιση για τα SRS, ένα σύστημα συστάσεων βασισμένο στους μη-ομοιόμορφους τυχαίους περιπάτους. Η συνεισφορά του κεφαλαίου είναι ένας αλγόριθμος, ο *Biased RW-RS*, ο οποίος βασίζεται σε μια νέα ιδέα για το φιλτράρισμα των γειτόνων του κάθε χρήστη. Παρεκκλίνοντας από την θεώρηση του κάθε γείτονα στο δίκτυο εμπιστοσύνης ως το ίδιο «έμπιστου», η προτεινόμενη μέθοδος μοντελοποιεί την ομοιότητα μεταξύ των χρηστών ως μια πιθανοτική συνάρτηση. Μιας και η μορφή της εν λόγω συνάρτησης είναι άγνωστη, προσεγγίζεται με την εφαρμογή ευρέως χρησιμοποιούμενων στη στατιστική βιβλιογραφία μεθόδων, όπως είναι η απορριπτική δειγματοληψία. Σε γενικές γραμμές, τα αποτελέσματα που επιτυγχάνονται είναι ικανοποιητικά και σύμφωνα με τους παρόντες ισχυρισμούς.

Λαμβάνεται βέβαια υπόψη το γεγονός πως το προτεινόμενο σύστημα δεν εμφανίζει ένα σαφές προβάδισμα στο `Epinions`. Η συγκεκριμένη, ωστόσο, συμπεριφορά αποδίδεται στα ιδιαίτερα χαρακτηριστικά της προαναφερόμενης συλλογής δεδομένων· στη μεγαλύτερη αραιότητα της καθώς και στο γεγονός πως περιέχει γενικές αξιολογήσεις, που δεν περιορίζονται σε συγκεκριμένο πεδίο εφαρμογής (όπως γίνεται λ.χ. στο `FilmTrust` που περιέχει αποκλειστικά και μόνο αξιολογήσεις ταινιών). Συνεπώς, ο προτεινόμενος αλγόριθμος θα πρέπει να εξελιχθεί περαιτέρω για να ανταποκριθεί στις προαναφερόμενες προκλήσεις.

□

Κεφάλαιο 7

Βελτιστοποίηση των κοινωνικών συστάσεων με την εφαρμογή μιας προσωποποιημένης στρατηγικής συσταδοποίησης των αντικειμένων

Στο προηγούμενο κεφάλαιο παρουσιάστηκε ένας αλγόριθμος κοινωνικής συνεργατικής διήθησης, ο οποίος πραγματοποιούσε τυχαίους περιπάτους στο δίκτυο εμπιστοσύνης προκειμένου να εντοπίζει χρήστες τους οποίους ο κάθε φορά υπό εξέταση χρήστης εμπιστεύεται αλλά και ως προς τους οποίους παρουσιάζει κάποια ομοιότητα στην αξιολογική συμπεριφορά. Σε αυτό το κεφάλαιο το πεδίο των τυχαίων περιπάτων θα αλλάξει και πλέον θα είναι το δίκτυο κατανάλωσης αντικειμένων του κάθε χρήστη αντί του δικτύου εμπιστοσύνης, μια έννοια που θα αναλυθεί περαιτέρω στις επόμενες ενότητες.

Αφετηρία για αυτήν την αλλαγή αποτελεί η παρατήρηση πως στα συστήματα συστάσεων είναι συχνό το φαινόμενο οι χρήστες τους να αξιολογούν αντικείμενα τα οποία δεν έχουν απαραίτητα ομοιότητες μεταξύ τους. Αυτό το φαινόμενο αποτελεί άμεση συνέπεια του γεγονότος πως το ανθρώπινο γούστο είναι πολυδιάστατο και επηρεάζεται από πολλούς παράγοντες, τους οποίους δεν μπορούν να συλλάβουν οι κλασσικοί αλγόριθμοι συστάσεων που βασίζονται στην ανάλυση περιεχομένου ή στη συνεργατική διήθηση. Για το λόγο αυτό, μια επιθυμητή ιδιότητα των RS είναι η δυνατότητα της εύρεσης συσχετίσεων μεταξύ φαινομενικά διαφορετικών αντικειμένων, τα οποία όμως μπορεί να ενδιαφέρουν τον εκάστοτε χρήστη. Αυτή η κατεύθυνση αναμένεται να βελτιώσει την καινοτομία και την ποικιλομορφία των παραγόμενων συστάσεων και κατά συνέπεια να αυξήσει την ικανοποίηση των χρηστών.

Στο παρόν κεφάλαιο, το προαναφερόμενο ζήτημα αντιμετωπίζεται με την ανάπτυξη ενός κοινωνικού συνεργατικού αλγορίθμου παραγωγής συστάσεων, ο οποίος προσωποποιεί τη συσταδοποίηση των αντικειμένων για κάθε χρήστη. Αυτό επιτυγχάνεται με την αναζήτηση μοτίβων ανάμεσα στα αντικείμενα που έχουν αξιολογηθεί καθώς και στην ομαδοποίηση τους σε διαφορετικές συστάδες σύμφωνα με την τις αξιολογήσεις που έχουν λάβει από άλλους χρήστες, οι οποίοι ανήκουν στο προσωπικό δίκτυο ενός δεδομένου χρήστη (Ενότητα 5.2.2). Το συγκεκριμένο δίκτυο αποτελείται από τους χρήστες με τους οποίους ο υπό εξέταση χρήστης συνδέεται άμεσα στο κοινωνικό δίκτυο καθώς και αυτούς με τους οποίους παρουσιάζει παρόμοια αξιολογική συμπεριφορά. Μετά το πέρας της φάσης της συσταδοποίησης, τα μέλη της κάθε συστάδας χρησιμοποιούνται ως τα δομικά στοιχεία για την κατασκευή ενός δικτύου κατανάλωσης αντικειμένων. Στη συνέχεια, με την πραγματοποίηση τυχαίων περιπάτων στο προαναφερόμενο δίκτυο, καθίσταται εφικτή η παραγωγή συστάσεων που είναι

ακριβείς αλλά ταυτόχρονα και πρωτότυπες και ποικιλόμορφες.

7.1 Εισαγωγή

Στο πλαίσιο της παρούσας διατριβής, έχει τονιστεί αρκετές φορές πως η διαδικασία παραγωγής συστάσεων εμπεριέχει μια εγγενή κοινωνική διάσταση. Πέρα από το γεγονός πως γενικά η αισθητική, οι πεποιθήσεις και το γούστο καθενός ανθρώπου μορφοποιούνται από την αλληλεπίδρασή του με το περιβάλλον του, είναι σχεδόν καθημερινό το φαινόμενο του να απευθύνεται κάποιος σε φίλους, συγγενείς και γνωστούς προκειμένου να λάβει μια απόφαση ή να πραγματοποιήσει μια αγορά. Μπορεί επιπρόσθετα να υποστηριχθεί ότι γενικότερα οι άνθρωποι τείνουν να αναπτύσσουν δεσμούς με εκείνους με τους οποίους μοιράζονται κοινά ενδιαφέροντα.

Οι παραπάνω γενικές παρατηρήσεις αντικατοπτρίζονται και στον χώρο των κοινωνικών δικτύων, μέσω του χαρακτηριστικού της *ομοφυλίας* (homophily). Ο όρος αυτός νοηματοδοτεί το γεγονός πως τα μέλη ενός κοινωνικού δικτύου τείνουν να είναι περισσότερο όμοια με τα μέλη με τα οποία συνδέονται παρά με τους υπόλοιπους χρήστες. Ή εναλλακτικά, οι γνωστοί του οποιουδήποτε μέλους ενός κοινωνικού δικτύου δεν συνιστούν τυχαία δείγματα τα οποία έχουν ληφθεί από το συνολικό πληθυσμό [Singla and Richardson, 2008].

Η ομοφυλία εξαρτάται από πολλούς παράγοντες, όπως είναι η ηλικία, το φύλο, το μορφωτικό επίπεδο, η εθνικότητα κ.ά. Στο παρόν κεφάλαιο, επιδιώκεται η χρήση αυτού του φαινομένου για την εξέταση του βαθμού στον οποίο τα ενδιαφέρονται ενός ατόμου επηρεάζονται από τους ανθρώπους με τους οποίους συναναστρέφεται. Για το σκοπό αυτό κατασκευάζεται ένα προσωπικό δίκτυο για τον κάθε χρήστη στο οποίο περιλαμβάνονται όλοι εκείνοι με τους οποίους αλληλεπιδρά, είτε άμεσα είτε έμμεσα, και το οποίο αποτελεί τη βάση για τη συσταδοποίηση των αντικειμένων, που θα παρουσιαστεί στη συνέχεια.

7.2 Σχετικές Εργασίες

Η εφαρμογή μεθόδων συσταδοποίησης στα παραδοσιακά συστήματα συστάσεων δεν αποτελεί καινούργια ερευνητική κατεύθυνση [Bobadilla et al., 2013]. Έχουν χρησιμοποιηθεί στο παρελθόν με σκοπό την εξεύρεση έμμεσων συσχετίσεων μεταξύ των χρηστών και των αντικειμένων. Στις περισσότερες περιπτώσεις, η συσταδοποίηση χρησιμοποιείται για τον καθορισμό των γειτονιών των χρηστών και των αντικειμένων.

Η συμπερίληψη στη διαδικασία της συσταδοποίησης και άμεσων σχέσεων μεταξύ των χρηστών, με τη μορφή της κοινωνικής πληροφορίας, είναι ωστόσο ένα σχετικά καινούργιο επιστημονικό αντικείμενο. Οι σχετικές εργασίες χρησιμοποιούν το κοινωνικό δίκτυο ως μια επιπλέον πηγή πληροφορίας πάνω στο οποίο εφαρμόζουν γνωστούς αλγορίθμους συσταδοποίησης. Για παράδειγμα, οι [DuBois et al., 2009] προτείνουν μια μεθοδολογία συναγωγής της τιμής εμπιστοσύνης μεταξύ ενός οποιουδήποτε ζεύγους χρηστών και κατόπιν πραγματοποιούν μια συσταδοποίηση βασισμένη στη συσχέτιση μεταξύ των συνηγμένων τιμών. Αφού εισαγάγουν τη μέθοδό τους σε παραδοσιακούς, μνημονικούς συνεργατικούς αλγορίθμους παραγωγής συστάσεων (είτε CF είτε βασισμένους στην εμπιστοσύνη), παρατηρούν μια σχετική βελτίωση της ακρίβειας των παραγόμενων συστάσεων.

Στην εργασία [Pitsilis et al., 2011], οι συγγραφείς εφαρμόζουν συσταδοποίηση βασισμένη σε *αλγόριθμο διάδοσης της συμπάθειας* (affinity propagation algorithm). Η μετρική της απόστασης που χρησιμοποιούν είναι ο συντελεστής Jaccard, δηλαδή το πλήθος των κοινών γειτόνων στο δίκτυο εμπιστοσύνης μεταξύ ενός ζεύγους χρηστών. Τα πειράματά τους έδειξαν ότι η προτεινόμενη από αυτούς μεθοδολογία ξεπερνά άλλες παραδοσιακές μεθόδους συσταδοποίησης (όπως η k-means) στην ακρίβεια των συστάσεων όσο το πλήθος των συστάδων αυξάνεται.

Οι [Pham et al., 2011] από την άλλη, πραγματοποιούν μια ιεραρχική συσταδοποίηση των κοινωνικών σχέσεων των χρηστών με σκοπό την διαμόρφωση γειτονιών. Χρησιμοποιούν την *αρχή της ομαδοποίησης* (modularity), ένα γραφοθεωρητικό κριτήριο το οποίο μετράει τον λόγο των ακμών εντός της συστάδας προς τον λόγο των ακμών εκτός της, για τον τερματισμό του σχηματισμού συστάδων. Σε επόμενο βήμα, οι δημιουργημένες γειτονίες τροφοδοτούν CF αλγορίθμους για την παραγωγή συστάσεων. Η μέθοδός τους δοκιμάστηκε σε δύο διαφορετικές συλλογές δεδομένων, όπου και παρατήρησαν βελτίωση στις παραγόμενες συστάσεις, χρησιμοποιώντας τις μετρικές της ακρίβειας ταξινόμησης και της ανάκλησης (Ενότητα 3.2.2).

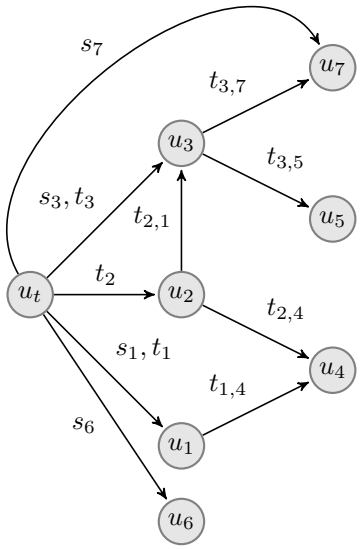
Η προσέγγιση που γίνεται στο παρόν κεφάλαιο διαφέρει από τις προαναφερόμενες σε τρία βασικά σημεία. Πρώτον, το επίπεδο εφαρμογής της συσταδοποίησης είναι τοπικό (προσωποποιημένο) αντί ολικό και έτσι ο αλγόριθμος εμφανίζει μεγαλύτερες δυνατότητες κλιμάκωσης όσον αφορά τις διαστάσεις των δεδομένων που εμπλέκονται στους υπολογισμούς. Δεύτερον, ενώ οι περισσότερες μεθοδολογίες χρησιμοποιούν την συσταδοποίηση για την ομαδοποίηση χρηστών, στην παρούσα εργασία χρησιμοποιείται για τον εντοπισμό παρόμοιων αντικειμένων μέσω της αναζήτησης κοινών μοτίβων προσπέλασης στο προσωπικό δίκτυο του κάθε χρήστη. Τέλος, η κοινωνική πληροφορία χρησιμοποιείται στη φάση του φιλτραρίσματος και όχι από τον καθ' εαυτό αλγόριθμο συσταδοποίησης.

7.3 Το Προσωπικό Δίκτυο

Μέχρι στιγμής το γούστο ενός ανθρώπου έχει αναλυθεί όσον αφορά τη σύγκρισή του με το αντίστοιχο γούστο άλλων ανθρώπων. Ωστόσο, το ανθρώπινο γούστο είναι μια πολυδιάστατη έννοια μιας και οι περισσότεροι άνθρωποι δεν περιορίζονται στο να έχουν αποκλειστικά και μόνο ένα ενδιαφέρον. Για παράδειγμα, αν κάποιος έχει εντάξει στα ενδιαφέροντά του την κηπουρική και την εκμάθηση κιθάρας, αναμένεται να αξιολογεί αντίστοιχα αντικείμενα σε ένα RS. Και από την στιγμή που δεν είναι υποχρεωτικό όσοι ενδιαφέρονται για την κηπουρική να ενδιαφέρονται εξίσου για την εκμάθηση κιθάρας, η αξιολογική συμπεριφορά του συγκεκριμένου χρήστη μπορεί να μοιάζει παράδοξη από τη σκοπιά ενός συστήματος συστάσεων· ακόμα και στα πολύ μεγάλα RS, ελάχιστος αναμένεται να είναι ο αριθμός των ανθρώπων που αξιολογούν αντικείμενα και των δύο κατηγοριών ταυτόχρονα.

Η παραπάνω συλλογιστική μπορεί να αναπτυχθεί περαιτέρω, κάνοντας την παραδοχή πως τα άτομα που γνωρίζει κάποιος άνθρωπος εντάσσονται χοντρικά σε δύο κατηγορίες: σε αυτά που έχει γνωρίσει μέσω συγκεκριμένων δραστηριοτήτων (λ.χ εκπαίδευση, εργασία, χόμπι) καθώς και στις πιο γενικές επαφές (λ.χ συγγενείς, φίλοι). Είναι συνεπώς αρκετά πιθανό ότι κάποιος χρήστης και οι γνωστοί που έχει από κάποιο χόμπι να έχουν αξιολογήσει σχετικά αντικείμενα. Άρα είναι επιθυμητό να εντοπίζονται τέτοιες σχέσεις στο RS και για το λόγο αυτό στο παρόν κεφάλαιο προτείνεται η *προσωποποιημένη συσταδοποίηση* των αξιολογημένων αντικειμένων. Στην προσωποποιημένη συσταδοποίηση δεν λαμβάνεται υπόψη κάθε αξιολόγηση που έχει τυχόν λάβει ένα αντικείμενο· αντίθετα, επιλέγονται οι αξιολογήσεις από εκείνους τους χρήστες που σχετίζονται είτε άμεσα (μέσω του κοινωνικού δικτύου) είτε έμμεσα (μέσω παρόμοιας αξιολογικής συμπεριφοράς) με τον υπό εξέταση χρήστη.

Ο προτεινόμενος αλγόριθμος λειτουργεί σε δύο βήματα. Στο πρώτο, ανακτώνται όλες οι αξιολογήσεις που έχει κάνει ένας συγκεκριμένος χρήστης (πάνω από κάποιο προκαθορισμένο κατώφλι). Στη συνέχεια, τα αντίστοιχα αντικείμενα που αφορούν οι αξιολογήσεις τοποθετούνται σε συστάδες σύμφωνα με κάποια καλώς καθορισμένα κριτήρια. Κατόπιν, τα αντικείμενα της κάθε συστάδας χρησιμοποιούνται ως δομικά στοιχεία για την κατασκευή ενός δικτύου κατανάλωσης αντικειμένων. Τέλος, η προαναφερθείσα δομή προσπελάζεται μέσω τυχαίων περιπάτων για την εύρεση και σύσταση νέων αντικειμένων στον υπό εξέταση χρήστη.



Σχήμα 7.1: Το Προσωπικό Δίκτυο

7.3.1 Η κατασκευή του προσωπικού δικτύου

Όπως αναφέρθηκε προηγουμένως, σκοπός του παρόντος κεφαλαίου είναι η ανάπτυξη ενός συστήματος προσωποποιημένου φιλτραρίσματος των διαθέσιμων αντικειμένων για κάθε χρήστη. Η βάση της προσέγγισης που θα ακολουθηθεί είναι το προσωπικό δίκτυο του κάθε χρήστη (Ενότητα 5.2.2), το οποίο κατασκευάζεται από δύο διαφορετικά αλλά όχι αναγκαστικά μη-επικαλυπτόμενα σύνολα χρηστών· τους άμεσους γείτονες στο κοινωνικό δίκτυο του υπό εξέταση χρήστη καθώς και τους όμοιους (με αυτόν) χρήστες, όπου η ομοιότητα καθορίζεται από κάποιον συγκεκριμένο συντελεστή. Τα μέλη του προσωπικού δικτύου του χρήστη μπορούν επιπρόσθετα να κατηγοριοποιηθούν στις παρακάτω ομάδες:

- Χρήστες στο άμεσο κοινωνικό δίκτυο του υπό εξέταση χρήστη, οι οποίοι επιπρόσθετα εμφανίζουν κάποια ομοιότητα με αυτόν
- Άλλοι όμοιοι χρήστες
- Άλλοι χρήστες στο κοινωνικό δίκτυο (λ.χ «φίλοι-φίλων»), οι οποίοι μπορεί να είναι όμοιοι με αυτόν
- Άλλοι χρήστες στο κοινωνικό δίκτυο

Η ομοιότητα μετράται με τους γνωστούς δείκτες ομοιότητας, όπως ο συντελεστής συσχέτισης Pearson, η ομοιότητα συνημιτόνου και η ομοιότητα Manhattan. Οι πρώτες δύο συνιστούν καλούς δείκτες ομοιότητας στις περιπτώσεις εκείνες που υπάρχει μεγάλη αλληλοεπικάλυψη στις αξιολογήσεις των χρηστών ενώ αντίθετα η τρίτη είναι περισσότερο κατάλληλη όταν η αλληλοεπικάλυψη είναι μικρότερη.

Σε γενικές γραμμές, ο κοινωνικός γράφος και ο γράφος ομοιότητας μπορούν να προσπελαστούν με πολλούς τρόπους και για το λόγο αυτό υπάρχει μια πληθώρα σχετικών μεθοδολογιών στη βιβλιογραφία των κοινωνικών συστημάτων συστάσεων. Μια πρώτη προσέγγιση της εκτίμησης της εγγύτητας των χρηστών είναι οι μέθοδοι *μέτρησης διαδρομής με βάρη* (weighted path counting). Για παράδειγμα, στο Σχήμα 7.1 ο χρήστης u_3 θεωρείται ο πιο κοντινός στον u_t γιατί ανήκει στο άμεσο κοινωνικό του δίκτυο (ακμή με βάρος t_3), στο άμεσο κοινωνικό δίκτυο του u_2, φίλου του u_t (ακμή με βάρος $t_{2,1}$) και τέλος στο δίκτυο ομοιότητας του u_t (ακμή με βάρος s_3). Στο ίδιο παράδειγμα, οι χρήστες u_4 και u_7 μπορούν να προσπελαστούν

από τον u_t μέσω δύο απλών διαδρομών μήκους ενός και δύο βημάτων. Για να καθοριστεί ποιος από τους δύο είναι εγγύτερα στον u_t, υπολογίζεται ένα συνολικό βάρος για κάθε διαδρομή, αθροίζοντας τα βάρη των αντίστοιχων ακμών.

7.4 Προσωποποιημένη Συσταδοποίηση

7.4.1 Ο Πίνακας Γειτνίασης των Αντικειμένων

Έχοντας υπολογιστεί η σημασία του κάθε μέλους στο προσωπικό δίκτυο ενός δοσμένου χρήστη u_t, στη συνέχεια κατασκευάζεται ο πίνακας των αντικειμένων A. Έστω ότι I_t είναι τα αντικείμενα τα οποία έχουν ήδη λάβει αξιολόγηση από τον u_t πάνω από ένα προκαθορισμένο κατώφλι ωφέλειας αντικειμένου r_{rel} (το οποίο μπορεί να εξαρτάται από τον u_t). Τότε ο A ορίζεται ως ο $n \times n$ πίνακας γειτνίασης (όπου $n = |I_t|$), του οποίου τα στοιχεία a_{ij} και a_{ji} υποδηλώνουν τη συχνότητα με την οποία τα αντικείμενα $i, j \in I_t$ έχουν προσπελαστεί από κοινού (πάνω από το κατώφλι ωφέλειας αντικειμένων) από τα μέλη του προσωπικού δικτύου του u_t. Η Εξίσωση 7.1 απεικονίζει ένα τέτοιο παράδειγμα πίνακα ενός χρήστη που έχει αξιολογήσει 7 αντικείμενα.

$$A = \begin{array}{c} \\ i_1 \\ i_2 \\ i_3 \\ i_4 \\ i_5 \\ i_6 \\ i_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} i_1 & i_2 & i_3 & i_4 & i_5 & i_6 & i_7 \end{array} \\ \left[\begin{array}{ccccccc} 0 & 0 & 0 & 3 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 3 & 0 & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 4 \\ 4 & 0 & 2 & 8 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 4 & 0 & 0 \end{array}\right] \end{array} \tag{7.1}$$

Πρέπει να σημειωθεί ότι από τη στιγμή που δεν βρίσκονται όλα τα μέλη του PN στην ίδια απόσταση από τον u_t, τα στοιχεία του πίνακα, στη γενική περίπτωση, δεν είναι ακέραια αθροίσματα. Πράγματι, η συνεισφορά του κάθε μέρους στη συχνότητα προσπέλασης μεταξύ δύο αντικειμένων δεν είναι σταθερή (λ.χ 1) αλλά μετριάζεται από την εγγύτητά του στον υπό εξέταση χρήστη (Ενότητα 7.3.1)

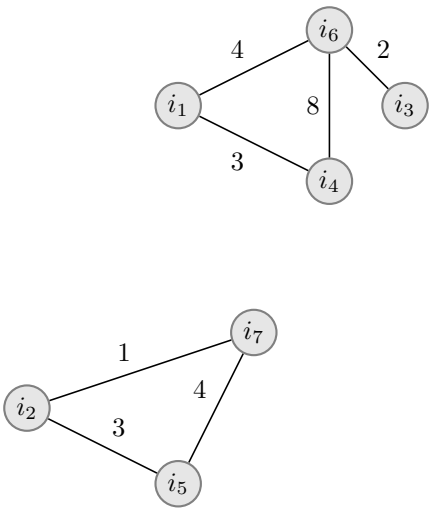
Εξ' ορισμού, ο πίνακας A είναι συμμετρικός και μπορεί να θεωρηθεί ότι συμβολίζει ένα μη-κατευθυντικό γράφο, του οποίου οι κόμβοι είναι τα αντικείμενα που έχουν ήδη αξιολογηθεί από τον u_t και οι ακμές αναπαριστούν μοτίβα αξιολόγησης στο ίδιο σύνολο αντικειμένων από άλλα μέλη του προσωπικού δικτύου του u_t. Όπως είναι αναμενόμενο, κάποια αντικείμενα προσπελάζονται από κοινού συχνότερα από κάποια άλλα και αυτό το φαινόμενο αντικατοπτρίζεται στον γράφο με το σχηματισμό συστάδων αντικειμένων (Σχήμα 7.2). Συνεπώς, είναι επιθυμητό να διαχωριστούν αυτές οι κοινότητες και για το σκοπό αυτό εφαρμόζεται ένας *αλγόριθμος φασματικής συσταδοποίησης* (spectral clustering algorithm) [Ng et al., 2001] στον πίνακα A, ο οποίος περιγράφεται αμέσως παρακάτω.

7.4.2 Ο Αλγόριθμος Συσταδοποίησης

Για δεδομένο συμμετρικό πίνακα γειτνίασης αντικειμένων A:

1. Υπολογίζεται ο διαγώνιος πίνακας *βαθμών κόμβων* (degree matrix) D του οποίο τα στοιχεία είναι $d_{ii} = \sum_j a_{ij}$

2. Υπολογίζεται ο κανονικοποιημένος *λαπλασιανός πίνακας* (laplacian matrix) $L = I - D^{-1/2} A D^{1/2}$



Σχήμα 7.2: Συστάδες αντικειμένων του πίνακα γειτνίασης της Εξίσωσης 7.1

3. Υπολογίζονται οι n ιδιοτιμές του L $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda n$ και τα αντίστοιχα ιδιοδιανύσματα $v_1 \dots v_n$. Μιας και ο L είναι συμμετρικός, όλες του οι ιδιοτιμές είναι πραγματικές, αλλά όχι κατ' ανάγκη διαφορετικές μεταξύ τους. Με άλλα λόγια, ορισμένες ιδιοτιμές μπορούν να εμφανίζονται παραπάνω από μια φορά ή πιο αυστηρά, να έχουν πολλαπλότητα (multiplicity) μεγαλύτερη της μονάδας.

4. Σύμφωνα με το *Θεώρημα Perron-Frobenius*, η μικρότερη ιδιοτιμή ενός μη-αρνητικού πίνακα είναι πάντοτε το μηδέν και η πολλαπλότητά του (k) ορίζει το πλήθος των συνδεδεμένων στοιχείων του γράφου, ο οποίος περιγράφεται από τον πίνακα γειτνίασης A. Ως k τίθεται ο αριθμός των επιθυμητών συστάδων αντικειμένων .

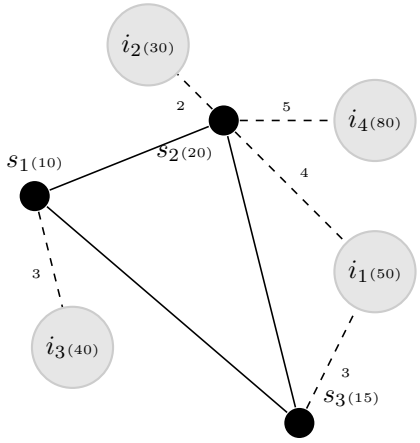
5. Κατασκευάζεται ο $n \times k$ πίνακας U του οποίου τα διανύσματα στήλες είναι τα k μεγαλύτερα ιδιοδιανύσματα του L, $v_1 \dots v_k$.

6. Συσταδοποιούνται τα n διανύσματα-γραμμές του U σε k συστάδες $(C_1 \dots C_k)$ με την χρήση κατάλληλης μεθοδολογίας συσταδοποίησης (π.χ. k-means συσταδοποίηση). Στη συνέχεια το κάθε αντικείμενο τοποθετείται στην κατάλληλη συστάδα με την χρήση ενός κριτηρίου απόστασης.

Η ανάλυση ιδιοτιμών που περιγράφεται στο Βήμα 3 δεν αποτελεί μια χρονοβόρα διαδικασία μιας και στην συντριπτική πλειοψηφία των περιπτώσεων, ο πίνακας A (και κατά συνέπεια ο πίνακας L) είναι χαμηλών διαστάσεων (λ.χ. $n < 100$).

7.5 Το Δίκτυο Κατανάλωσης Αντικειμένων

Για κάθε συστάδα που δημιουργήθηκε στο προηγούμενο βήμα της μεθόδου, κατασκευάζεται ένα *δίκτυο κατανάλωσης αντικειμένων* (item consumption network - ICN) [Liu et al., 2012]. Ένα παράδειγμα ενός τέτοιου δικτύου απεικονίζεται στο Σχήμα 7.3. Οι μαύροι κόμβοι s_1 ως s_3 είναι αντικείμενα της ίδιας συστάδας και συνδέονται μεταξύ τους με συνεχείς ακμές. Οι γκρι κόμβοι i_1 ως i_4 αντιπροσωπεύουν αντικείμενα που έχουν αξιολογηθεί από τα μέλη του PN του u_t αλλά όχι από τον ίδιο. Μια ακμή από τα αντικείμενα της πρώτης κατηγορίας προς αυτά της δεύτερης απεικονίζει το γεγονός πως αυτά τα δύο αντικείμενα έχουν προσπελαστεί από κοινού από τους χρήστες του PN του υπό εξέταση χρήστη το πλήθος των φορών που υποδεικνύεται από το βάρος της αντίστοιχης ακμής. Στο προαναφερόμενο παράδειγμα, τα αντικείμενα s_2 και i_1 έχουν αξιολογηθεί από κοινού από 4 ομότιμους χρήστες που ανήκουν



Σχήμα 7.3: Το Δίκτυο Κατανάλωσης Αντικειμένων

στο PN του u_t. Τέλος, οι αριθμοί στις παρενθέσεις κάθε κόμβου δείχνουν τον αριθμό των αξιολογήσεων που έχει λάβει το συγκεκριμένο αντικείμενο από τα μέλη του PN.

Για να παραχθούν οι συστάσεις, το ICN μοντελοποιείται ως γράφος με σκοπό την προσπέλασή του μέσω ενός μη-ομοιόμορφου τυχαίου περιπάτου. Πράγματι, ο επαγόμενος γράφος από το ICN έχει τις ιδιότητες του να είναι συνδεδεμένος και μη-διμερής και κατά συνέπεια μπορεί να θεωρεί ότι είναι μια *συμμετρική, χρονικά αναστρέψιμη και πεπερασμένη μαρκοβιανή αλυσίδα* [Lovasz, 1993]. Η ιδιότητα της συμμετρίας επιβεβαιώνεται εύκολα από το γεγονός πως ο γράφος είναι μη-κατευθυντικός (Σχήμα 7.3). Ο όρος μη-ομοιόμορφος που αναφέρεται στον τυχαίο περίπατο είναι απόρροια του γεγονότος πως η πιθανότητα $p_{i,j}$ προκύπτει όπως στην Εξίσωση 5.4 (και όχι όπως στην Εξίσωση 5.5).

Μια θεμελιώδης ιδιότητα των τυχαίων περιπάτων σε πεπερασμένες μαρκοβιανές αλυσίδες είναι πως συγκλίνουν στην *κατανομή σταθερής κατάστασης* τους (steady-state distribution) π, ανεξάρτητα από τον εκάστοτε κόμβο-αφετηρία [Lovasz, 1993]. Αν P είναι ο $m \times m$ συμμετρικός πίνακας μεταβάσεων της αλυσίδας (όπου m ο πληθάριθμος των αντικειμένων στο ICN) έτσι ώστε

$$P = [p_{i,j}]^{n \times n}, \quad i, j \in V, \quad (i, j) \in E \tag{7.2}$$

τότε η ακόλουθη Εξίσωση είναι αληθής

$$P^\top \pi = \pi \tag{7.3}$$

η οποία συμβολίζει το γεγονός πως ο πίνακας P έχει τη μεγαλύτερή του ιδιοτιμή ίση με 1, με το αντίστοιχο αριστερό ιδιοδιάνυσμα να είναι η κατανομή σταθερής κατάστασης π. Συνεπώς, η ανάλυση ιδιοτιμών του πίνακα P επιτρέπει την ανάκτηση του διανύσματος π, τα στοιχεία του οποίου εκφράζουν την πιθανότητα επίσκεψης του κάθε κόμβου από τον τυχαίο περίπατο.

Σε ορισμένες περιπτώσεις ωστόσο, οι διαστάσεις του P μπορεί να είναι αρκετά υψηλές και τότε η ανάλυση ιδιοτιμών γίνεται χρονοβόρα. Για το λόγο αυτό, η κατανομή σταθερής κατάστασης υπολογίζεται με τον παρακάτω επαναληπτικό τρόπο [Lovasz, 1993]:

1. Δημιουργείται το $m \times 1$ διάνυσμα r και τα στοιχεία του τίθενται στο 0 εκτός από εκείνα που αντιστοιχούν στα αρχικά αντικείμενα τα οποία λαμβάνουν την τιμή $\frac{1}{|S|}$, ($|S|$ είναι ο πληθάριθμος του συνόλου των αρχικών αντικειμένων)

2. Υπολογίζεται το διάνυσμα $w = P \times r$.

3. Όσο $dist(w, r) > \epsilon$. Εδώ, το $dist(w, r)$ συμβολίζει μια συνάρτηση η οποία υπολογίζει την απόσταση δύο διανυσμάτων σε έναν μετρικό χώρο (λ.χ. η ευκλείδεια απόσταση) και το ϵ αποτελεί το κατώφλι ομοιότητας μεταξύ των διανυσμάτων (π.χ. $\epsilon = 10^{-4}$)

(α′) Τίθεται $r = w$

(β′) Υπολογίζεται $w = P \times r$

4. Τέλος Βρόχου

5. Επιστρέφονται ως συστάσεις οι N μη-αρχικοί κόμβοι με τη μεγαλύτερη πιθανότητα στο w.

7.6 Πειραματική Διαδικασία

Η προτεινόμενη στο παρόν κεφάλαιο μεθοδολογία δοκιμάστηκε πειραματικά στη συλλογή δεδομένων **Epinions**, η οποία έχει συλλεχθεί από τους [Massa and Bhattacharjee, 2004] από την αντίστοιχη διαδικτυακή υπηρεσία. Τα χαρακτηριστικά της εν λόγω συλλογής δεδομένων έχουν παρουσιαστεί στον Πίνακα 6.1 του προηγούμενου κεφαλαίου. Ωστόσο, αξίζει να σημειωθεί πως το **Epinions** είναι εξαιρετικά αραιό, τόσο όσον αφορά την πυκνότητα των αξιολογήσεων όσο και τη συνδεσιμότητα του δικτύου εμπιστοσύνης (όπως αυτή μετριέται από τον ολικό συντελεστή συσταδοποίησης). Εκτός της αραιότητας, περιέχει ένα μεγάλο ποσοστό χρηστών και αντικειμένων με ελάχιστες αξιολογήσεις. Αυτά τα χαρακτηριστικά επηρεάζουν σε μεγάλο βαθμό την ποιότητα των παραγόμενων συστάσεων.

Μια άλλη ιδιαιτερότητα του **Epinions** είναι πως οι αξιολογήσεις των αντικειμένων δεν κατανέμονται ομοιόμορφα σε όλες τις δυνατές τιμές της πενταβάθμιας κλίμακας, αλλά εμφανίζονται αλλοιωμένες προς τις υψηλότερες τιμές (4 και 5) με μια αναλογία 1 προς 3. Η κατάσταση αυτή αποδίδεται στον συμπεριφορισμό και πιο συγκεκριμένα στην πρακτική των χρηστών να μην αξιολογούν γενικά όσα αντικείμενα γνωρίζουν, αλλά κυρίως αυτά που τους αρέσουν. Συνεπώς, ένα βασικό σύστημα συστάσεων που θα πρότεινε κάθε αντικείμενο στα «τυφλά» με βαθμό ωφέλειας 4 ή 5, θα εμφάνιζε πολύ ικανοποιητικά αποτελέσματα. Αυτή η παρατήρηση ενισχύει την αντίληψη πως η αποτελεσματικότητα των συστημάτων συστάσεων δεν πρέπει να μετράται μονοδιάστατα· μόνο, δηλαδή, από τη σκοπιά της βελτίωσης της ακρίβειας των προτάσεων. Αντίθετα θα πρέπει να εξετάζονται και άλλες πλευρές της, όπως η καινοτομία και η ποικιλομορφία των παραγόμενων συστάσεων (Ενότητα 3.3.2). Για τον λόγο αυτό, στη συγκεκριμένη πειραματική διαδικασία εξετάστηκαν οι μετρικές του RMSE (Εξίσωση 3.3), της κάλυψης των αξιολογήσεων (Εξίσωση 3.23), της καινοτομίας βασισμένης στην απόσταση των αντικειμένων (Εξίσωση 3.18) και της ποικιλομορφίας της λίστας των παραγόμενων συστάσεων (Εξίσωση 3.20).

Επίσης, για την πληρέστερη εκτίμηση της συνεισφοράς της προτεινόμενης μεθόδου, υλοποιήθηκαν μια σειρά συστημάτων συστάσεων αναφοράς, τόσο κλασσικά όσο και βασισμένα σε δίκτυα εμπιστοσύνης. Ο αλγόριθμος ItemMean προτείνει το κάθε αντικείμενο με βαθμό ωφέλειας ίσο με τον μέσο όρο των αξιολογήσεων που έχει λάβει το συγκεκριμένο αντικείμενο (χωρίς δηλαδή να λαμβάνει υπόψη του καμία συσχέτιση των χρηστών). Ο αλγόριθμος UserMean από την άλλη προτείνει το κάθε αντικείμενο με βαθμό ωφέλειας ίσο με τον μέσο όρο των αξιολογήσεων που έχει δώσει ο εκάστοτε χρήστης σε όλα τα υπόλοιπα αντικείμενα (χωρίς δηλαδή να λαμβάνει υπόψη του συσχετίσεις μεταξύ των αντικειμένων).

Το σύστημα συνεργατικής διήθησης που χρησιμοποιήθηκε βασίστηκε στον τύπο του Resnick (Εξίσωση 2.11) με συντελεστή συσχέτισης την ομοιότητα Manhattan.Η συγκεκριμένη φόρμουλα εφαρμόζεται σε δύο παραλλαγές· στην πρώτη υπολογίζεται η ομοιότητα ανάμεσα στους χρήστες ενώ στη δεύτερη η ομοιότητα ανάμεσα στα αντικείμενα (περίπτωση της συνεργατικής διήθησης βασισμένης στο αντικείμενο).

Τέλος, τα κοινωνικά συστήματα που υλοποιήθηκαν βασίστηκαν στους κλιμακωτούς αλγορίθμους εμπιστοσύνης TidalTrust και MoleTrust (Ενότητα 5.5.2). Το βάθος εξερεύνησης του πρώτου ήταν ίσο με τη διάμετρο του συνδεδεμένου στοιχείου στο οποίο ανήκει ο εκάστοτε

Πίνακας 7.1: Αποτελέσματα στο `Epinions` (για λίστα 5 προτεινόμενων αντικειμένων)

Μετρικές Απόδοσης		RMSE	Κάλυψη	Καινοτομία	Ποικιλομορφία
Α.	**Συστήματα Αναφοράς**				
Α.1	ItemMean	1.09	86.43%	11.89%	24.23%
Α.2	UserMean	1.20	98.58%	9.70%	19.42%
Β.	**Συνεργατική διήθηση**				
Β.1	Ομοιότητα Manhattan (Όλοι οι γείτονες)	1.07	79.57%	20.11%	56.23%
Γ.	**Συνεργατική διήθηση βασισμένη στα αντικείμενα**				
Γ.1	Ομοιότητα Manhattan (Όλα τα όμοια αντικείμενα)	1.20	39.29%	16.86%	45.26%
Δ.	**Συστήματα βασισμένα στην Εμπιστοσύνη**				
Δ.1	MoleTrust-1	1.23	25.58%	29.16%	43.62%
Δ.2	MoleTrust-2	1.16	56.52%	32.31%	54.02%
Δ.3	MoleTrust-3	1.12	70.89%	42.13%	56.65%
Δ.4	TidalTrust	1.08	74.67%	45.38%	59.17%
Ε.	**Το Προτεινόμενο Σύστημα**				
Ε.1	Προσωποποιημένη Συσταδοποίηση Αντικειμένων	1.05	58.17%	53.11%	63.04%

χρήστης ενώ για τον δεύτερο ο ορίζοντας διάδοσης τέθηκε σε 3 διαφορετικές τιμές· στους άμεσους φίλους (MoleTrust-1), στους φίλους των φίλων (MoleTrust-2) και τέλος στους φίλους των φίλων της προηγούμενης κατηγορίας (MoleTrust-3). Φυσικά, υλοποιήθηκε και η προτεινόμενη στο παρόν κεφάλαιο μεθοδολογία.

7.7 Αποτελέσματα

Ο Πίνακας 7.1 συνοψίζει τα αποτελέσματα των μετρικών της απόδοσης που λήφθηκαν μετά την διεξαγωγή της πειραματικής διαδικασίας που περιγράφηκε στην προηγούμενη ενότητα. Όλα τα αποτελέσματα αφορούν τη συλλογή δεδομένων `Epinions` και το πρωτόκολλο που εφαρμόστηκε ήταν η *αντεπικύρωση με την εξαίρεση ενός* (leave-one-out cross validation) για μια λίστα 5 προτεινόμενων αντικειμένων. Μια πρώτη παρατήρηση είναι πως ο προτεινόμενος αλγόριθμος εμφανίζει ένα σταθερό προβάδισμα όσον αφορά την ακρίβεια, την καινοτομία και την ποικιλομορφία των συστάσεων. Τα ενθαρρυντικά αυτά αποτελέσματα (και ειδικότερα αυτό της ακρίβεια των συστάσεων) αποδίδονται στη μέθοδο επεξεργασίας του προσωπικού δικτύου κάθε χρήστη καθώς και στον τρόπο με τον οποίο εντοπίζονται οι χρήστες που τον επηρεάζουν περισσότερο (Ενότητα 7.3.1). Επιπρόσθετα, η αυξημένη καινοτομία και ποικιλομορφία των παραγόμενων προτάσεων οφείλονται στην στρατηγική της προσωπικής συσταδοποίησης των αντικειμένων που ακολουθήθηκε. Πράγματι, σε αρκετές περιπτώσεις, η τεχνική αυτή κατορθώνει να συλλάβει τα διαφορετικά ενδιαφέροντα καθενός χρήστη.

Ορισμένα συστήματα βασισμένα στην εμπιστοσύνη εμφανίζουν πολύ ικανοποιητικά αποτελέσματα όσον αφορά την κάλυψη των αξιολογήσεων. Αυτή η συμπεριφορά πηγάζει από τον επιθετικό τρόπο με τον οποίο προσπελάζουν το δίκτυο εμπιστοσύνης, συσσωρεύοντας όλους τους χρήστες μέχρι ένα ορισμένο βάθος. Για αυτό ακριβώς τον λόγο, γιατί δεν είναι επιλεκτικά, αποτυγχάνουν στο να επιτύχουν καλύτερα αποτελέσματα όσον αφορά την ακρίβεια των συστάσεων, παρότι καλύπτουν περισσότερους χρήστες. Ωστόσο πρέπει να αναγνωριστεί το γεγονός πως η εμπιστοσύνης στο `Epinions` είναι δυαδικό μέγεθος· είτε υπάρχει πλήρως είτε καθόλου. Αν η συγκεκριμένη συλλογή δεδομένων περιείχε τιμές σε μεγαλύτερο εύρος, τότε ενδεχομένως οι αλγόριθμοι εμπιστοσύνης να εμφάνιζαν μεγαλύτερη ακρίβεια και μικρότερη κάλυψη.

Τέλος τα συστήματα αναφοράς μπορεί να εμφανίζουν αρκετά ικανοποιητικά αποτελέσματα όσον αφορά την ακρίβεια και την κάλυψη των αξιολογήσεων (ειδικά το ItemMean), αλλά η συμπεριφορά αυτή σχετίζεται στις ιδιαιτερότητες της συγκεκριμένης συλλογής δεδομένων και δεν οφείλεται στα ιδιαίτερα χαρακτηριστικά των μεθόδων. Όπως αναφέρθηκε και προηγουμένως, ένα πρωτόλειο σύστημα το οποίο θα πρότεινε ένα οποιοδήποτε αντικείμενο με βαθμό ωφέλειας στο εύρος $[4, 5]$ θα επεδείκνυε ικανοποιητικά αποτελέσματα. Είναι προφανές, ωστόσο, ότι αυτή η προσέγγιση δεν είναι ρεαλιστική για το οποιοδήποτε πρακτικό σύστημα συστάσεων.

7.8 Συμπεράσματα

Στις προηγούμενες ενότητες παρουσιάστηκε μια νέα μεθοδολογία παραγωγής κοινωνικών συστάσεων, η οποία βασίζεται στην προσωποποιημένη, για τον κάθε χρήστη, συσταδοποίηση των αντικειμένων. Παρότι τα πειραματικά αποτελέσματα είναι ικανοποιητικά, εντούτοις υπάρχει χώρος για περαιτέρω βελτιώσεις. Πιο συγκεκριμένα, θα ήταν επιθυμητή η βελτίωση της διαδικασίας κατασκευής του προσωπικού δικτύου. Μια πιθανή κατεύθυνση μπορεί να είναι η μοντελοποίηση της εμπιστοσύνης και της ομοιότητας ως οριακών πιθανοτήτων μιας άγνωστης κοινής κατανομής, η οποία κατόπιν θα πρέπει να προσεγγιστεί. Έτσι, η ομοιότητα των χρηστών μπορεί να προκύψει απευθείας μέσω της δειγματοληψίας από την εν λόγω κοινή κατανομή.

Επίσης, ο αλγόριθμος της φασματικής συσταδοποίησης θα μπορούσε να βελτιωθεί περαιτέρω με την εισαγωγή κριτηρίων που θα εξετάζουν το μέγεθος και την ποιότητα των παραγόμενων συστάδων. Αυτό μπορεί να καταστεί εφικτό με την χρήση άλλων τεχνικών συσταδοποίησης, όπως η ασαφής k-means συσταδοποίηση μαζί με διαφορετικές συναρτήσεις απόστασης (π.χ απόσταση Chebysev ή Mahalanobis) για την τοποθέτηση των αντικειμένων στις συστάδες.

Τέλος, οι ιδιότητες του τυχαίου περιπάτου στο δίκτυο κατανάλωσης αντικειμένων (όπως λ.χ ο ρυθμός μίξης) μπορούν να βελτιστοποιηθούν αν ληφθούν υπόψη διαφορετικές πιθανότητες μετάβασης, οι οποίες θα συμπεριλαμβάνουν περισσότερες πληροφορίες σχετικές με τα χαρακτηριστικά του κάθε κόμβου.

□

Κεφάλαιο 8

Συνεργατικό σύστημα συστάσεων βασισμένο στη μη-αρνητική παραγοντοποίηση πινάκων και στα μοντέλα εκθετικών τυχαίων γράφων

Στην ανάλυση που έχει μέχρι στιγμής πραγματοποιηθεί, ο τρόπος αξιοποίησης της κοινωνικής πληροφορίας από τους αλγορίθμους συνεργατικής διήθησης που έχουν προταθεί έχει γίνει αποκλειστικά με την χρήση των τυχαίων περιπάτων. Στο Κεφάλαιο 6 ο τυχαίος περίπατος λαμβάνει χώρα στον χώρο των χρηστών ενώ αντίθετα στο Κεφάλαιο 7 στον χώρο των αντικειμένων. Ωστόσο, η κοινωνική πληροφορία μπορεί να ενσωματωθεί και με άλλους τρόπους στα συστήματα συστάσεων. Ένας από αυτούς είναι να χρησιμοποιηθεί για την ανεύρεση κοινοτήτων χρηστών, στην βάση των δεσμών (λ.χ. φιλίας, εμπιστοσύνης) που αυτοί έχουν στο κοινωνικό δίκτυο. Αυτές οι κοινότητες μπορούν κατόπιν να χρησιμοποιηθούν για την παραγωγή προτάσεων από τα επόμενα βήματα ενός SRS. Σε αυτό το σημείο αξίζει να τονιστεί ότι χρήση του κοινωνικού δικτύου με σκοπό τη συσταδοποίηση αντικειμένων έγινε έμμεσα και στο αμέσως προηγούμενο κεφάλαιο, στη διαδικασία της κατασκευής του δικτύου κατανάλωσης αντικειμένων

Γενικότερα, η τοποθέτηση των χρηστών σε «γειτονιές» στο κοινωνικό δίκτυο αποτελεί αντικείμενο ενός ευρύτερου επιστημονικού πεδίου που είναι γνωστό ως *ανίχνευση κοινοτήτων* (community detection) [Fortunato, 2010]. Το συγκεκριμένο πεδίο έχει μελετηθεί εκτενώς και έχουν προταθεί πολλές μεθοδολογίες προς αυτή την κατεύθυνση. Στις πιο αποδοτικές σχετικές τεχνικές συγκαταλέγονται αυτές που πραγματοποιούν τη λεγόμενη *επικαλυπτόμενη ανίχνευση κοινοτήτων* (overlapping community detection) και πιο συγκεκριμένα αυτές που βασίζονται στη *μη-αρνητική παραγοντοποίηση του πίνακα* (non-negative matrix factorization - NMF) γειτνίασης του κοινωνικού γράφου [Psorakis et al., 2011; Yang and Leskovec, 2013]. Ο επιθετικός προσδιορισμός «επικαλυπτόμενη» χρησιμοποιείται για να υποδηλώσει το γεγονός πως σε αυτή την περίπτωση ο κάθε χρήστης δεν θεωρείται ότι ανήκει αποκλειστικά και μόνο σε μία γειτονιά· αντίθετα, μπορεί να ανήκει σε περισσότερες, με διαφορετικό ποσοστό συμμετοχής στην κάθε μία. Σε αυτό το σημείο, εντοπίζεται και μια αντιστοιχία με τη συλλογιστική του Κεφαλαίου 7. Με τον ίδιο τρόπο που ένας χρήστης δεν αξιολογεί αντικείμενα από μια και μόνο κατηγορία, έτσι και οι κοινωνικές του επαφές δεν εντάσσονται σε μια και μόνο κατηγορία (λ.χ κάποιοι μπορεί να είναι συνάδελφοι από την εργασία, κάποιοι φίλοι ενώ κάποιοι μπορεί ενδεχομένως να ανήκουν και στις δύο προαναφερόμενες ομάδες). Η ανακάλυψη και η εκμετάλλευση, συνεπώς, αυτών των διακριτών ομαδοποιήσεων αναμένεται να οδηγήσει στη

βελτίωση της παραγωγής συστάσεων.

Στην συγκεκριμένη περίπτωση που εξετάζεται, η χρήση της NMF έχει σκοπό την εξαγωγή λανθανόντων παραγόντων που περιγράφουν την τοποθέτηση των μελών του κοινωνικού δικτύου σε δύο ή περισσότερες κοινότητες. Σε ένα διαφορετικό επίπεδο ωστόσο, η NMF έχει χρησιμοποιηθεί στους αλγορίθμους συνεργατικής διήθησης κατασκευής μοντέλου (Ενότητα 2.3.2) κατά κόρον. Δεν είναι τυχαίο ότι δύο από τα πιο αποδοτικά RS, ο αλγόριθμος ALSWR [Zhou et al., 2008] και ο αλγόριθμος RSVD++ [Koren, 2008] εντάσσονται ακριβώς σε αυτή την κατηγορία. Οι προαναφερόμενοι αλγόριθμοι όμως, δεν παραγοντοποιούν τον πίνακα γειτνίασης του κοινωνικού γράφου, αλλά τον ίδιο τον πίνακα των αξιολογήσεων (Ενότητα 2.1.2) με στόχο αυτή την φορά την εύρεση των λανθανόντων παραγόντων που περιγράφουν τα αντικείμενα και τους χρήστες αντίστοιχα και οι οποίοι αποτελούν τη βάση για την παραγωγή των προτάσεων.

Στο κοινωνικό μνημονικό σύστημα συνεργατικών συστάσεων που θα παρουσιαστεί στις επόμενες ενότητες, η NMF δεν θα εφαρμοστεί στον πίνακα των αξιολογήσεων (όπως γίνεται στα συνεργατικά συστήματα κατασκευής μοντέλου) αλλά αντίθετα στον πίνακα γειτνίασης του κοινωνικού γράφου, με στόχο την τοποθέτηση των χρηστών σε μία (ή περισσότερες) επικαλυπτόμενες κοινότητες. Επίσης, η NMF θα είναι προσωποποιημένη για τον κάθε χρήστη που εξετάζεται και θα αφορά το κομμάτι εκείνο του κοινωνικού δικτύου που αντιστοιχεί στους γείτονες των γειτόνων του υπό εξέταση χρήστη (δίκτυο FoaF). Το συγκεκριμένο όριο τίθεται γιατί έχει παρατηρηθεί [Massa and Avesani, 2009] ότι οι συστάσεις δεν βελτιώνονται ιδιαίτερα όταν το βάθος της εξερεύνησης του προσωπικού δικτύου του κάθε χρήστη γίνει μεγαλύτερο από 2.

Το σημαντικότερο και πλέον πρωτότυπο χαρακτηριστικό της προτεινόμενης στο παρόν κεφάλαιο μεθοδολογίας είναι η εισαγωγή των μοντέλων εκθετικών τυχαίων γράφων (exponential rangom graph models - ERGM) [Robins et al., 2007] στη διαδικασία της παραγοντοποίησης. Ο λόγος που γίνεται η συγκεκριμένη επιλογή είναι για να μετριαστεί η κυρίαρχα τοπική λογική της NMF, η οποία εστιάζει στο επίπεδο της ακμής και θεωρεί την κάθε μια ανεξάρτητη από τις υπόλοιπες. Είναι γνωστό, όμως, ότι τα κοινωνικά δίκτυα έχουν συγκεκριμένη δομή (όσον αφορά των τρόπο που συνδέονται οι κόμβοι μεταξύ τους) που δεν προσομοιάζει με τους εντελώς τυχαίους γράφους [Wasserman and Faust, 1994]. Αυτή ακριβώς η γενική πληροφορία για τα συνολικότερα χαρακτηριστικά του γράφου είναι επιθυμητό να μοντελοποιηθεί και να εισαχθεί στην παραγοντοποίηση και σε αυτό το επίπεδο έρχονται να συνεισφέρουν τα ERGM, όπως θα φανεί και στις επόμενες ενότητες του παρόντος κεφαλαίου.

8.1 Μη-αρνητική παραγοντοποίηση πινάκων

Η μη-αρνητική παραγοντοποίηση πινάκων αποτελεί κομμάτι μιας ευρύτερης οικογένειας τεχνικών μείωσης των διαστάσεων, οι οποίες προσπαθούν να κατασκευάσουν μια κατά-μέρη αναπαράσταση δεδομένων πολύ υψηλών διαστάσεων, μέσω της προβολής τους σε ένα χώρο χαμηλότερης διάστασης. Στην ίδια οικογένεια τεχνικών ανήκει εξάλλου και η SVD που χρησιμοποιήθηκε στο Κεφάλαιο 4 και πιο συγκεκριμένα στην Ενότητα 4.1.1. Η διαφορά της NMF από τις άλλες τεχνικές αποτελεί ο περιορισμός της μη-αρνητικότητας των στοιχείων των παραγόμενων πινάκων, πράγμα που επιτρέπει την καλύτερη ερμηνεία του αποτελέσματος [Wang and Zhang, 2013].

Στη συγκεκριμένη περίπτωση που εξετάζουμε, έστω $A \in \mathbb{R}_+^{n \times n}$ πίνακας γειτνίασης γράφου n κόμβων. Επιθυμούμε να τον παραγοντοποιήσουμε σε δύο μη-αρνητικούς πίνακες· στον *πίνακα βάσης* (basis matrix) $W \in \mathbb{R}_+^{n \times r}$ και στον *πίνακα των συντελεστών* (coefficient martix) $H \in \mathbb{R}_+^{r \times n}$, έτσι ώστε

$$A \approx \widetilde{A} \equiv WH \tag{8.1}$$

όπου r ο αριθμός των κοινοτήτων και $r << n$.

Σκοπός της NMF είναι να υπολογίσει τα στοιχεία των W, H έτσι ώστε το γινόμενό τους να είναι όσο το δυνατόν «εγγύτερα» στον A, με την εγγύτητα να μετριέται από κάποια συνάρτηση απόστασης. Πιο αυστηρά, η μη-αρνητική παραγοντοποίηση πίνακα είναι το παρακάτω πρόβλημα συνδυαστικής βελτιστοποίησης (που στη συγκεκριμένη περίπτωση περιγράφεται ως ένα πρόβλημα ελαχιστοποίησης)

$$\min_{W,H} \mathcal{D}(A||WH) \tag{8.2}$$
$$\text{subject to } W \geq 0, H \geq 0$$

όπου $\mathcal{D}(.||.)$ είναι μια συνάρτηση απόστασης των πινάκων A και WH. Γενικά, το πρόβλημα της ελαχιστοποίησης της συνάρτησης $\mathcal{D}$ είναι NP-hard, για το οποίο επιπρόσθετα δεν είναι γνωστοί *κυρτοί τύποι βελτιστοποίησης* (convex formulations) που θα οδηγούσαν στην εύρεση του ολικού ελαχίστου της $\mathcal{D}$ όσον αφορά και τους δύο πίνακες W και H ταυτόχρονα [Wang and Zhang, 2013]. Παρότι η βελτιστοποίηση της $\mathcal{D}$ είναι μη-κυρτή (non-convex) και για τους δύο πίνακες ταυτόχρονα, είναι ωστόσο κυρτή (convex) για τον κάθε ένα από τους δύο πίνακες χωριστά, δηλαδή κρατώντας αμετάβλητο π.χ. τον W, το πρόβλημα της Εξίσωσης 8.2 γίνεται κυρτό ως προς τον H (και το αντίστροφο) [Zhang, 2012]. Για τον λόγο αυτό, η πλειοψηφία των προτεινόμενων αλγορίθμων επαναληπτικά και εναλλάξ προσαρμόζουν τους πίνακες W, H [Wang and Zhang, 2013; Zhang, 2012], όπως φαίνεται στον Αλγόριθμο 5

Αλγόριθμος 5 NMF - Γενική Περίπτωση [Zhang, 2012]

Είσοδος: $W^{(0)}, H^{(0)}, t = 1$
Έξοδος: W, H
Βρόχος
  Κράτησε σταθερό το $H^{(t-1)}$ και βρες $W^{(t)}$ τέτοιο ώστε:
    $\mathcal{D}(A||W^{(t)}H^{(t-1)}) \leq \mathcal{D}(A||W^{(t-1)}H^{(t-1)})$
  Κράτησε σταθερό το $W^{(t)}$ και βρες $H^{(t)}$ τέτοιο ώστε:
    $\mathcal{D}(A||W^{(t)}H^{(t)}) \leq \mathcal{D}(A||W^{(t)}H^{(t-1)})$
  Έλεγχος σύγκλισης
  Αν το κριτήριο σύγκλισης ικανοποιείται **Τότε**
    $W \leftarrow W^{(t)}$
    $H \leftarrow H^{(t)}$
    Τέλος επανάληψης
  Τέλος Αν
  $t \leftarrow t + 1$
Τέλος Βρόχος

8.1.1 Μπεϋζιανή NMF

Η *μπεϋζιανή NMF* [Schmidt et al., 2009] αποτελεί υποκατηγορία της *πιθανοτικής NMF*, η οποία προσεγγίζει τις παραμέτρους W, H κάνοντας χρήση της κλασσικής σχέσης της μπεϋζιανής συνεπαγωγής

$$P(W, H, \Theta|A) \propto P(A|W, H, \Theta)P(W, H|\Theta)P(\Theta) \tag{8.3}$$

όπου οι πίνακες βάσης και συντελεστών αποτελούν τις παραμέτρους του μοντέλου και Θ είναι ο χώρος των υπερπαραμέτρων (που ρυθμίζουν τη στατιστική συμπεριφορά της κατανομής από την οποία προκύπτουν οι πίνακες W, H). Για την χρήση της παραπάνω σχέσης απαραίτητη προϋπόθεση είναι να μπορούμε να κάνουμε εικασία για την στατιστική προέλευση των δεδομένων του πίνακα γειτνίασης και των πινάκων-παραγόντων του γινομένου. Το αριστερό μέρος

της Σχέσης 8.3 εκφράζει την *εκ των υστέρων* (a posteriori) πιθανότητα οι παράμετροι του μοντέλου να λάβουν μια καθορισμένη τιμή με βάση τα συγκεκριμένα δεδομένα, ενώ ο πρώτος όρος του γινομένου του δεξιού μέρους της ίδιας σχέσης εκφράζει την *πιθανοφάνεια* (likelihood) του μοντέλου, ο επόμενος την *εκ των προτέρων* (a priori) πιθανότητα των παραμέτρων του μοντέλου για τις συγκεκριμένες υπερπαραμέτρους και τέλος ο τελευταίος όρος την πιθανότητα εμφάνισης των ιδίων των υπερπαραμέτρων.

Μια προσεκτική επιλογή της συνάρτησης πιθανοφάνειας και της εκ των προτέρων πιθανότητας μπορεί να έχει ως αποτέλεσμα έναν αλγόριθμο που επιδεικνύει καλύτερη και γρηγορότερη σύγκλιση. Έτσι πρέπει να γίνει παρατήρηση των παραμέτρων του μοντέλου και στη βάση αυτής της παρατήρησης να επιλεγεί η κατανομή πιθανοφάνειας που εκφράζει καλύτερα τις στατιστικές τους ιδιότητες. Συνηθέστερες επιλογές αποτελούν η κανονική και η Poisson κατανομές. Κατόπιν, και αφού έχει αποσαφηνιστεί η μορφή της πιθανοφάνειας, γίνεται η επιλογή της κατάλληλης εκ των προτέρων πιθανότητας για τον χώρο των υπερπαραμέτρων. Οι πιο πολλοί ερευνητές κάνουν επιλογές στη βάση της θεωρίας των *συζυγών εκ των προτέρων κατανομών* (conjugate prior theory) [Raiffa and Schlaifer, 2000], η οποία μπορεί να περιγράψει τον τύπο και τα χαρακτηριστικά της ζητούμενης εκ των υστέρων πιθανότητας, αν η πιθανοφάνεια και η εκ των προτέρων πιθανότητα λάβουν συγκεκριμένες μορφές. Για κανονική ή Poisson πιθανοφάνεια, συζυγείς εκ των προτέρων κατανομές που συνήθως επιλέγονται είναι η κανονική, η Γάμμα, η ανάστροφη Γάμμα και η Wishart .

8.1.2 Εκ των προτέρων πιθανότητα

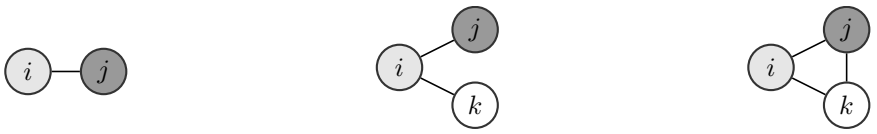
Παρότι οι εκ των προτέρων πιθανότητες που αναφέρθηκαν στην προηγούμενη ενότητα αποτελούν, σε γενικές γραμμές, καλές επιλογές για την μπεϋζιανή NMF, εντούτοις δεν αποτυπώνουν το γεγονός πως σε αυτή την περίπτωση ο πίνακας που θέλουμε να παραγοντοποιήσουμε αποτελεί πίνακα γειτνίασης FoaF δικτύου, του οποίου οι ακμές έχουν συγκεκριμένες δομικές ιδιότητες [Vozalis and Margaritis, 2007]. Έτσι, οι περισσότεροι NMF αλγόριθμοι εξετάζουν το κάθε στοιχείο (ακμή) του πίνακα γειτνίασης ανεξάρτητα από τα υπόλοιπα, σαν να μην υπάρχουν συσχετίσεις μεταξύ τους, που να επιβάλλονται από το δίκτυο.

Καθώς ο ρόλος της εκ των προτέρων πιθανότητας στη μπεϋζιανή συνεπαγωγή είναι να ενσωματώνει αυτού του είδους την πληροφορία, είναι σημαντικό η επιλογή της να γίνει προς μια κατεύθυνση η οποία θα κωδικοποιεί, στο μέτρο που αυτό είναι εφικτό, και μακροσκοπικές ιδιότητες του γράφου. Ιδανικά, μια τέτοια επιλογή αναμένεται να δράσει εξισορροπητικά ως προς την επίδραση των τοπικών δομικών χαρακτηριστικών του γράφου, τα οποία εισάγονται από την πιθανοφάνεια. Έτσι, συμπληρωματικά ως προς την πιθανοφάνεια της μπεϋζιανής NMF, προτείνεται η στατιστική μοντελοποίηση του δικτύου FoaF του κάθε χρήστη μέσω τις εφαρμογής των *μοντέλων εκθετικών τυχαίων γράφων* στη θέση της εκ των προτέρων πιθανότητας.

8.2 Μοντέλα εκθετικών τυχαίων γράφων

Τα μοντέλα εκθετικών τυχαίων γράφων αποτελούν μια κατηγορία *μοντέλων ομάδας* (ensemble models), η οποία αποτελείται από όλους τους δυνατούς *απλούς γράφους* (που δεν περιέχουν, δηλαδή, ακμές με αφετηρία και τερματισμό τον ίδιο κόμβο) που μπορούν να υπάρξουν μεταξύ n κόμβων [Robins et al., 2007]. Για κάθε απλό γράφο G του μοντέλου, ή εναλλακτικά μια *διαμόρφωση* (configuration) των n κόμβων, ορίζεται η αντίστοιχη πιθανότητα εμφάνισής του

$$P(G) = \frac{1}{Z}\mathrm{e}^{H(G)}, \quad Z = \sum_{G \in \mathcal{G}} \mathrm{e}^{H(G)} \tag{8.4}$$



(α′) Απλή Ακμή (β′) 2-αστέρας (γ′) Τρίγωνο

Σχήμα 8.1: Βασικές παρατηρήσεις δικτύου ενός γράφου με μη-κατευθυντικές ακμές

όπου Z η *συνάρτηση κατακερματισμού* (partition function). Η Z ισούται με το άθροισμα όλων των δυνατών διαμορφώσεων του μοντέλου, εξασφαλίζοντας με αυτόν τον τρόπο ότι το μέγεθος $P(G)$ εκφράζει πιθανότητα.

Οι ιδιότητες του μοντέλου, ή αλλιώς οι *παρατηρήσεις του δικτύου* (network observables), κωδικοποιούνται γραμμικά στη *χαμιλτονιανή* $H(G)$

$$H(G) = \sum_{i=1}^{r} \theta_i x_i(G) \tag{8.5}$$

όπου οι παράμετροι θ_i καθορίζουν τη συνεισφορά της εκάστοτε παρατήρησης $x_i(G)$ στο συνολικό μοντέλο. Οι παρατηρήσεις του δικτύου αποτελούν επί της ουσίας ιδιότητες του γράφου, όπως το πλήθος των απλών ακμών, των αμοιβαίων ακμών, των αστέρων, των τριγώνων κ. ά. (Σχήμα 8.1). Μια παρατήρηση του δικτύου μπορεί να μετατραπεί σε ένα *στατιστικό δικτύου* (network statistic), υπολογίζοντας την αναμενόμενη τιμή της, όπως παρακάτω

$$\begin{aligned} \langle x_i \rangle &= \sum_{G \in \mathcal{G}} x_i(G) P(G) = \frac{1}{Z} \left(\sum_{G \in \mathcal{G}} x_i(G) \exp\{\sum_{i=1}^{r} \theta_i x_i(G)\} \right) \Rightarrow \\ \langle x_i \rangle &= \frac{1}{Z} \left(\frac{\partial}{\partial \theta_i} \sum_{G \in \mathcal{G}} \exp\{\sum_{i=1}^{r} \theta_i x_i(G)\} \right) = \frac{1}{Z} \frac{\partial Z}{\partial \theta_i} = \frac{\partial F}{\partial \theta_i} \end{aligned} \tag{8.6}$$

Το F συμβολίζει την *ελεύθερη ενέργεια* του μοντέλου (για την οποία ισχύει $F \equiv \ln Z$). Στη γενική περίπτωση, ο αναλυτικός υπολογισμός της συνάρτησης κατάτμησης είναι αδύνατος. Στα απλά μοντέλα όμως, είναι δυνατόν να προσδιοριστεί επακριβώς η μορφή της Z [Park and Newman, 2004a].

8.2.1 Γράφοι Bernoulli

Το πιο απλό μοντέλο εκθετικών τυχαίων γράφων είναι οι γράφοι Bernoulli ή διαφορετικά το μοντέλο Erdős-Rèyni. Το μόνο στατιστικό δικτύου που μοντελοποιείται είναι το αναμενόμενο πλήθος ακμών $\langle m \rangle$. Στην περίπτωση γράφου G με μη κατευθυντικές ακμές, η χαμιλτονιανή και η συνάρτηση κατάτμησης λαμβάνουν την παρακάτω αναλυτική μορφή

$$H(G) = \theta m(G), \quad m(G) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \tag{8.7}$$

όπου $m(G)$ το πλήθος των ακμών του γράφου. Η συνάρτηση κατάτμησης λαμβάνει την παρακάτω μορφή

$$Z = [1 + \mathrm{e}^{\theta}]^{\binom{n}{2}} \tag{8.8}$$

Αν είναι γνωστή η αναμενόμενη τιμή $\langle m \rangle$ του πλήθους των ακμών του γράφου, τότε η τιμή της παραμέτρου θ μπορεί να υπολογιστεί ως εξής

$$\langle m \rangle = \frac{1}{Z} \frac{\partial Z}{\partial \theta} = \binom{n}{2} \frac{1}{1 + \mathrm{e}^{\theta}} \Rightarrow \theta = \ln \frac{\langle m \rangle}{\binom{n}{2} - \langle m \rangle} \tag{8.9}$$

και συνεπώς το μοντέλο ορίζεται πλήρως.

Άλλα μοντέλα για τα οποία υπάρχει αναλυτική έκφραση για το $p(G)$ είναι οι *γενικευμένοι τυχαίοι γράφοι*, όπου η παρατήρηση δικτύου που μοντελοποιείται είναι η κατανομή του βαθμού των κόμβων, και το *μοντέλο αμοιβαιότητας*, όπου εκτός από το πλήθος των ακμών μοντελοποιείται και το ποσοστό αυτών που είναι αμοιβαίες (δηλαδή από τον κόμβο i στον j και πίσω). Όλα τα υπόλοιπα πιο σύνθετα μοντέλα δεν μπορούν να υπολογιστούν αναλυτικά και για αυτό το λόγο χρειάζεται να καταφύγουμε σε προσεγγιστικές τεχνικές για να υπολογίσουμε τις παραμέτρους τους.

8.3 Μπεϋζιανή NMF και ERGM

Στις προηγούμενες ενότητες, περιγράφηκε η μεθοδολογική καινοτομία της εισαγωγής των ERGM στην θέση της εκ των προτέρων πιθανότητας της μπεϋζιανής NMF. Έχοντας ως αφετηρία τη Σχέση 8.3, θα αναπτυχθεί περαιτέρω η συλλογιστική μας για την επιλογή της κατάλληλης εκ των προτέρων πιθανότητας (Ενότητα 8.3.1) και της συνάρτησης πιθανοφάνειας (Ενότητα 8.3.2)

8.3.1 Εκ των προτέρων πιθανότητα

Η πρώτη παρατήρηση που γίνεται είναι ότι η παραγοντοποίηση αφορά δίκτυο FoaF, οπότε αναμένεται η διάταξη των ακμών μεταξύ των κόμβων του να εμφανίζει φαινόμενα τύπου αστέρα, δηλαδή ορισμένους, λίγους, κόμβους με μεγάλο αριθμό εφαπτόμενων ακμών και πολλούς κόμβους με μικρό αριθμό εφαπτόμενων ακμών. Συνεπώς, υπάρχουν πολλά «ανοιχτά» τρίγωνα και το πλέον κατάλληλο ERGM για την περίσταση είναι το μοντέλο των 2-αστέρων (Σχήμα 8.1β′)

$$H = \theta m(G) + \tau s(G) \tag{8.10}$$

$$m(G) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}, \quad s(G) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \sum_{k=1(\neq j)}^{n-1} a_{ik}$$

όπου a_{ij} στοιχείο του πίνακα γειτνίασης A του γράφου (με τιμή 1 αν υπάρχει ακμή μεταξύ των κόμβων i, j και 0 διαφορετικά), $m(G)$ το στατιστικό δικτύου που μοντελοποιεί το πλήθος των ακμών του γράφου (η επιρροή του οποίου ελέγχεται από την υπερπαράμετρο θ) και $s(G)$ το αντίστοιχο μέγεθος για το πλήθος των 2-αστέρων (του οποίου η επιρροή ελέγχεται αντίστοιχα από την υπερπαράμετρο τ). Επίσης, πρέπει να σημειώσουμε ότι στη συγκεκριμένη περίπτωση, ο γράφος είναι μη-κατευθυντικός (δηλαδή για τα i, j στοιχεία του πίνακα γειτνίασης ισχύει η ισότητα $a_{ij} = a_{ji}$).

Δυστυχώς δεν υπάρχει αναλυτική λύση για το μοντέλο που περιγράφεται από τη χαμιλτονιανή της Εξίσωσης 8.10 και συνεπώς πρέπει να καταφύγουμε σε προσεγγιστικές τεχνικές για την εκτίμηση της τιμής των υπερπαραμέτρων θ, τ. Ως ένα πρώτο βήμα, ξαναγράφουμε τα στατιστικά του δικτύου ως συνάρτηση, όχι των ακμών, αλλά του βαθμού k_i του κάθε κόμβου

$$m(G) = \frac{1}{2}\sum_{i=1}^{n} k_i, \quad s(G) = \frac{1}{2}\sum_{i=1}^{n} k_i(k_i - 1) = \frac{1}{2}\sum_{i=1}^{n} k_i^2 - m(G)$$

Τότε η χαμιλτονιανή της Εξίσωσης 8.10 γίνεται

$$H = \theta m(G) + \tau s(G) = \frac{\tau}{2}\sum_{i=1}^{n} k_i^2 + \frac{\theta - \tau}{2}\sum_{i=1}^{n} k_i \tag{8.11}$$

Για λόγους διευκόλυνσης των υπολογισμών που θα ακολουθήσουν, οι υπερπαράμετροι θ και τ αντικαθίστανται από τις βοηθητικές υπερπαραμέτρους J και B, οι οποίες ορίζονται ως

εξής

$$J = \frac{(n-1)\tau}{2}, \quad B = \frac{\tau - \theta}{2} \tag{8.12}$$

οπότε η χαμιλτονιανή της Εξίσωσης 8.11 παίρνει την τελική της μορφή

$$H = \frac{J}{n-1}\sum_{i=1}^{n}(k_i)^2 + B\sum_{i=1}^{n}k_i \tag{8.13}$$

Συγκρίνοντας τις Εξισώσεις 8.10-8.13, μπορεί να παρατηρηθεί ότι εκφράζουν το ίδιο ERGM, δηλαδή το μοντέλο 2-αστέρων, χρησιμοποιώντας διαφορετικά στατιστικά δικτύου. Στην Εξίσωση 8.10 χρησιμοποιείται το πλήθος των ακμών (που ελέγχεται από την υπερπαράμετρο θ) και το πλήθος των ανοιχτών τριγώνων (υπερπαράμετρος τ) ενώ στην Εξίσωση 8.13 χρησιμοποιείται το άθροισμα των τετραγώνων των βαθμών των κόμβων (που ελέγχεται από την υπερπαράμετρο $\frac{J}{n-1}$) και το απλό άθροισμα των βαθμών των κόμβων (υπερπαράμετρος B). Και οι δύο αναπαραστάσεις είναι απόλυτα ισοδύναμες με μια μόνο διαφορά· η δεύτερη (Εξίσωση 8.13) μπορεί να προσεγγιστεί με τη εφαρμογή τεχνικών από τη στατιστική μηχανική των δικτύων, χρησιμοποιώντας τη *θεωρία του μέσου πεδίου* (mean-field theory) [Park and Newman, 2004a].

Εμπνεόμενοι από μια αντίστοιχη τεχνική των [Park and Newman, 2004b], επιλύουμε το μοντέλο που περιγράφεται από τη χαμιλτονιανή της Εξίσωσης 8.13, βασιζόμενοι στο μετασχηματισμό Hubbard-Stratonovich και στην ανάπτυξη των λύσεων γύρω από σαγματικό σημείο. Ακολουθώντας την προσεγγιστική μας, που αναλύεται εκτενώς στο Παράρτημα Α΄.1, βρίσκουμε ότι η ελεύθερη ενέργεια του μοντέλου που περιγράφεται από την χαμιλτονιανή της Εξίσωσης 8.13 είναι η παρακάτω (Εξίσωση Α΄.20)

$$F = -n(n-1)J(\phi_0)^2 + \frac{1}{2}n(n-1)\ln\left(1+\mathrm{e}^{4J\phi_0+2B}\right) + \frac{n}{2}\ln[(n-1)J] - \frac{n}{2}\ln 4\pi$$

όπου ϕ_0 η λύση της προσέγγισης μέσου πεδίου (Παράρτημα Α΄.1.1) που δίνεται παρακάτω (Εξίσωση Α΄.18)

$$\phi_0 = \frac{1}{2}\left[\tanh(2J\phi_0 + B) + 1\right]$$

Η ακριβώς παραπάνω εξίσωση φανερώνει ότι οι υπερπαραμέτροι B, J του μοντέλου σχετίζονται μεταξύ τους, μια παρατήρηση που θα φανεί χρήσιμη στους επόμενους υπολογισμούς.

Σύμφωνα με τα όσα αναφέρθηκαν στην Ενότητα 8.2, η μερική παράγωγος της ελεύθερης ενέργειας του μοντέλου ως προς την υπερπαράμετρο B ισούται με την αναμενόμενη τιμή του αθροίσματος των βαθμών των κόμβων

$$\langle\sum_{i=1}^{n}k_i\rangle = \frac{\partial F}{\partial B} \Rightarrow \sum_{i=1}^{n}\langle k_i\rangle = \frac{\partial F}{\partial B} \tag{8.14}$$

Προσεγγίζοντας την αναμενόμενη τιμή του βαθμού του κάθε κόμβου $\langle k_i\rangle$ με την πιο πιθανή/αναμενόμενη τιμή της, δηλαδή την αναμενόμενη τιμή $\langle k\rangle$ του βαθμού όλων των κόμβων και συνδυάζοντάς την με την Εξίσωση 8.14, έχουμε

$$\begin{aligned}
&\sum_{i=1}^{n}\langle k_i\rangle = \frac{\partial F}{\partial B} \Rightarrow n\langle k\rangle = \frac{\partial F}{\partial B} \Rightarrow \langle k\rangle = \frac{1}{n}\frac{\partial F}{\partial B} \Rightarrow \\
&\quad \langle k\rangle = (n-1)\frac{\mathrm{e}^{4J\phi_0+2B}}{1+\mathrm{e}^{4J\phi_0+2B}} = (n-1)\frac{1}{2}[\tanh(2J\phi_0+B)+1] \Rightarrow \\
&\quad \langle k\rangle = (n-1)\phi_0 \Rightarrow \phi_0 = \frac{n-1}{\langle k\rangle}
\end{aligned} \tag{8.15}$$

κάνοντας το ϕ_0 ίσο με

$$\phi_0 = \frac{\langle k \rangle}{n-1}, \quad \phi_0 \in (0,1) \tag{8.16}$$

Έτσι, από την Εξίσωση 8.16 προκύπτει ότι το ϕ_0, δηλαδή η λύση μέσου πεδίου της Εξίσωσης Α΄.18, σχετίζεται άμεσα με την αναμενόμενη τιμή του βαθμού όλων των κόμβων. Οι [Park and Newman, 2004a] ονομάζουν το συγκεκριμένο μέγεθος *συνδεσιμότητα* (connectance) και παρατηρούν ότι εκφράζει τον βαθμό που ο γράφος μοιάζει με κλίκα (κάθε κόμβος μιας n-κλίκας έχει βαθμό $n-1$). Συνεπώς, είναι προφανές ότι η συνδεσιμότητα ϕ_0 λαμβάνει τιμές στο $(0,1)$.

Κατ΄ αντίστοιχο τρόπο, η μερική παράγωγος της ελεύθερης ενέργειας ως προς την βοηθητική υπερπαράμετρο $\frac{J}{n-1}$ είναι ίση με την αναμενόμενη τιμή του αθροίσματος των τετραγώνων των βαθμών των κόμβων.

$$\langle \sum_{i=1}^{n} k_i^2 \rangle = \sum_{i=1}^{n} \langle k_i^2 \rangle = \frac{\partial F}{\partial(\frac{J}{n-1})}$$

Όπως προηγουμένως, η αναμενόμενη τιμή του τετραγώνου του βαθμού ενός κόμβου ($\langle k_i^2 \rangle$) προσεγγίζεται από την πιο πιθανή/αναμενόμενη τιμή του, δηλαδή την αναμενόμενη τιμή του τετραγώνου του βαθμού όλων των κόμβων ($\langle k^2 \rangle$)

$$\begin{aligned}
\langle \sum_{i=1}^{n} k_i^2 \rangle = \sum_{i=1}^{n} \langle k_i^2 \rangle = \frac{\partial F}{\partial(\frac{J}{n-1})} \Rightarrow n\langle k^2 \rangle = (n-1)\frac{\partial F}{\partial J} \Rightarrow \langle k^2 \rangle = \frac{n-1}{n}\frac{\partial F}{\partial J} \Rightarrow \\
\langle k^2 \rangle = -(n-1)^2(\phi_0)^2 + 2(n-1)^2 \frac{\mathrm{e}^{4J\phi_0+2B}}{1+\mathrm{e}^{4J\phi_0+2B}} + \frac{1}{2}(n-1)\frac{1}{J} \Rightarrow \\
\langle k^2 \rangle = -(n-1)^2(\phi_0)^2 + 2(n-1)^2 \frac{1}{2}[\tanh(2J\phi_0+B)+1] + \\
+ \frac{1}{2}(n-1)\frac{1}{J} \Rightarrow \\
\langle k^2 \rangle = -(n-1)^2(\phi_0)^2 + 2(n-1)^2\phi_0 + \frac{1}{2}(n-1)\frac{1}{J}
\end{aligned} \tag{8.17}$$

Συνδυάζοντας την Εξίσωση 8.17 με την Εξίσωση 8.15 έχουμε

$$\begin{aligned}
\langle k^2 \rangle = -\langle k \rangle^2 + 2(n-1)\langle k \rangle + \frac{1}{2}(n-1)\frac{1}{J} \Rightarrow \\
\frac{1}{2}(n-1)\frac{1}{J} = \langle k^2 \rangle + \langle k \rangle^2 - 2(n-1)\langle k \rangle \Rightarrow \\
J = \frac{(n-1)}{2(\langle k^2 \rangle + \langle k \rangle^2 - 2(n-1)\langle k \rangle)}
\end{aligned} \tag{8.18}$$

Συνδυάζοντας τις Εξισώσεις Α΄.18, 8.15 και 8.18, έχουμε για την υπερπαράμετρο B

$$\begin{aligned}
2\phi_0 - 1 = \tanh(2J\phi_0 + B) \Rightarrow 2J\phi_0 + B = \tanh^{-1}(2\phi_0 - 1) \Rightarrow \\
B = \tanh^{-1}(2\phi_0 - 1) - 2J\phi_0
\end{aligned} \tag{8.19}$$

Δεδομένης της παρακάτω ταυτότητας

$$\tanh^{-1}(x) = \frac{1}{2}\ln\frac{1+x}{1-x}, \quad |x| < 1 \tag{8.20}$$

η Εξίσωση 8.19 τελικά απλοποιείται σε

$$B = \frac{1}{2}\ln\phi_0 - \frac{1}{\langle k^2 \rangle + \langle k \rangle^2 - 2(n-1)\langle k \rangle} \tag{8.21}$$

Συνεπώς, οι Εξισώσεις 8.18 και 8.21 περιγράφουν πλήρως το μοντέλο 2-αστέρων υπό την θεώρηση του μέσου πεδίου.

Αν ληφθεί υπόψη η αρχική χαμιλτονιανή του επιλύσιμου μοντέλου 2-αστέρων (Εξίσωση 8.13), τότε η ακόλουθη σχέση είναι αληθής

$$H = \frac{J}{n-1}\sum_{i=1}^{n} k_i^2 + B\sum_{i=1}^{n} k_i = \frac{\partial F}{\partial J}J + \frac{\partial F}{\partial B}B = \nabla F$$

Με πιο απλά λόγια, η χαμιλτονιανή του μοντέλου είναι ίση με την *κλήση* (gradient) της ελεύθερης ενέργειας. Αντικαθιστώντας τις τιμές των B, J με τις αντίστοιχες των Εξισώσεων 8.18 και 8.21 έχουμε

$$\begin{aligned} H =& \nabla F = \frac{\partial F}{\partial J}J + \frac{\partial F}{\partial B}B \\ =& -n(n-1)J(\phi_0)^2 + 2n(n-1)J\phi_0 + \frac{n}{2} + n(n-1)B\phi_0 \end{aligned}$$

Καθώς ο όρος $\frac{n}{2}$ είναι ανεξάρτητος από τις παραμέτρους του μοντέλου, απαλείφεται από την χαμιλτονιανή και ενσωματώνεται στην συνάρτηση κατακερματισμού. Συνεπώς, μπορούμε να συνεχίσουμε ως εξής

$$\begin{aligned} H =& -n(n-1)J(\phi_0)^2 + 2n(n-1)J\phi_0 + n(n-1)B\phi_0 \\ =& n(n-1)\phi_0\left[-J\phi_0 + 2J + B\right] \\ =& n\langle k\rangle\left[-J\phi_0 + 2J + \tanh^{-1}(2\phi_0 - 1) - 2J\phi_0\right] \\ =& \sum_{i=1}^{n}\langle k_i\rangle\left[2J - 3J\phi_0 + \tanh^{-1}(2\phi_0 - 1)\right] \\ =& 2\left[2J - 3J\phi_0 + \tanh^{-1}(2\phi_0 - 1)\right]\frac{1}{2}\sum_{i=1}^{n} k_i \\ =& \left[2\tanh^{-1}(2\phi_0 - 1) + 2(2 - 3\phi_0)J\right] m(G) \end{aligned} \tag{8.22}$$

Κάνοντας μια ακόμα φορά χρήση της ταυτότητας 8.20 καταλήγουμε στο

$$H = \Theta m(G), \qquad \Theta = \ln\phi_0 + 2(2 - 3\phi_0)J \tag{8.23}$$

Αυτό είναι ένα ιδιαίτερα σημαντικό συμπέρασμα για δύο λόγους. Πρώτον γιατί κατορθώσαμε να βρούμε μια προσεγγιστική λύση στο μοντέλο 2-αστέρων και δεύτερον γιατί η συγκεκριμένη λύση που βρήκαμε μπορεί να ενσωματωθεί εύκολα στην διαδικασία της παραγοντοποίησης, όπως θα γίνει εμφανές στις επόμενες ενότητες.

8.3.2 Πιθανοφάνεια

Μέχρι στιγμής, έχουν εξεταστεί και ενσωματωθεί στην εκ των προτέρων πιθανότητα οι μακροσκοπικές ιδιότητες του γράφου, μέσω τις εφαρμογής του μοντέλου 2-αστέρων των ERGM. Όπως αναφέρθηκε και προηγουμένως, η πιθανοφάνεια μοντελοποιεί τις τοπικές ιδιότητες των ακμών του γράφου και οι δύο πιο διαδεδομένες επιλογές της είναι η κανονική κατανομή και η κατανομή Poisson. Και οι δύο κάνουν την υπόθεση πως κάθε ακμή $(i, j) \in E$ στο δίκτυο δημιουργείται ανεξάρτητα από όλες τις άλλες, ή ισοδύναμα, ότι κάθε στοιχείο a_{ij} του πίνακα γειτνίασης A λαμβάνει τιμές ανεξάρτητα από τα υπόλοιπα.

Πιο συγκεκριμένα, η κανονική κατανομή θεωρεί ότι η πιθανότητα δημιουργίας ακμής μεταξύ των κόμβων i, j πηγάζει από μια γκαουσιανή πηγή με μέση τιμή $\mu = a_{ij}$ και απόκλιση

σ, ενώ η κατανομή Poisson θεωρεί αντίστοιχα ότι η πιθανότητα δημιουργίας ακμής μεταξύ των κόμβων i, j προκύπτει από μια πηγή Poisson ρυθμού $\lambda = a_{ij}$. Όταν οι προαναφερόμενες κατανομές χρησιμοποιούνται στην πιθανοτική NMF, μπορεί να αποδειχθεί ότι η κανονική κατανομή βελτιστοποιεί την ευκλείδεια απόσταση μεταξύ των πινάκων A και WH, ενώ η κατανομή Poisson βελτιστοποιεί τη γενικευμένη απόκλιση Kullback-Liebler [Lee and Seung, 2000]

$$\mathcal{D}(a_{ij}||\widetilde{a_{ij}}) = a_{ij} \ln \frac{a_{ij}}{\widetilde{a_{ij}}} - a_{ij} + \widetilde{a_{ij}} \tag{8.24}$$

Στην περίπτωση που εξετάζουμε, η μορφή της συνάρτησης της εκ των προτέρων πιθανότητας (Εξίσωση 8.23) επιβάλλει τη χρήση της κατανομής Poisson στη πιθανοφάνεια

$$P(\widetilde{a_{ij}}|a_{ij}) \propto \frac{\widetilde{a_{ij}}^{a_{ij}}}{a_{ij}!} \mathrm{e}^{-\widetilde{a_{ij}}} \tag{8.25}$$

γιατί έτσι απλοποιούνται σημαντικά οι απαιτούμενοι υπολογισμοί, όπως θα φανεί στη συνέχεια.

8.4 Αλγόριθμος Παραγοντοποίησης

Έχοντας ολοκληρώσει την συλλογιστική μας για την επιλογή των κατανομών της πιθανοφάνειας και της εκ των προτέρων πιθανότητας, προχωράμε στην ανάπτυξη του αλγορίθμου που θα εισάγει στην μπεϋζιανή NMF τα ERGM. Αφετηρία μας αποτελούν οι ευρύτατα διαδεδομένοι πολλαπλασιαστικοί κανόνες των [Lee and Seung, 1999], γιατί εξασφαλίζουν ταχύτερη και πιο ακριβή σύγκλιση (σε σταθερό σημείο) σε σχέση με άλλες μεθόδους. Οι συγκεκριμένοι κανόνες περιγράφονται στον Αλγόριθμο 6

Αλγόριθμος 6 NMF - Πολλαπλασιαστικοί Κανόνες Ανανέωσης [Lee and Seung, 1999]

Είσοδος: $W^{(0)}, H^{(0)}, t = 1$
Έξοδος: W, H

Βρόχος
  $W^{(t+1)} \leftarrow W^{(t)} \circ \frac{\nabla_W \mathcal{D}^+(V||W^{(t)}H^{(t)})}{\nabla_W \mathcal{D}^-(V||W^{(t)}H^{(t)})}$
  $H^{(t+1)} \leftarrow H^{(t)} \circ \frac{\nabla_H \mathcal{D}^+(V||W^{(t+1)}H^{(t)})}{\nabla_H \mathcal{D}^-(V||W^{(t+1)}H^{(t)})}$
  Έλεγχος σύγκλισης
  Αν το κριτήριο σύγκλισης ικανοποιείται **Τότε**
    $W \leftarrow W^{(t)}$
    $H \leftarrow H^{(t)}$
    Τέλος επανάληψης
  Τέλος Αν
  $t \leftarrow t + 1$
Τέλος Βρόχος

Καταρχήν, παρατηρούμε πως ο Αλγόριθμος 6 αποτελεί εφαρμογή της γενικής NMF μεθοδολογίας (Αλγόριθμος 5). Το μόνο πράγμα που αλλάζει είναι τα Βήματα 2 ως 4, δηλαδή ο τρόπος που ενημερώνονται τα στοιχεία των πινάκων W, H. Ο όρος $\nabla_W \mathcal{D}^+(.||.)$ εκφράζει τους θετικούς όρους της *κλίσης* (gradient) της συνάρτησης βελτιστοποίησης ως προς τον πίνακα W ενώ ο όρος $\nabla_W \mathcal{D}^-(.||.)$ τους αρνητικούς όρους (αντίστοιχα και για τον πίνακα H). Στο σημείο αυτό πρέπει να σημειωθεί ότι οι τύποι ενημέρωσης των γραμμών 2 και 3 έχουν εφαρμογή σε περιπτώσεις προβλήματος μεγιστοποίησης παραμέτρων. Στην Ενότητα 8.1, ωστόσο, το πρόβλημα της μη-αρνητικής παραγοντοποίησης έχει εκφραστεί ως πρόβλημα ελαχιστοποίησης

(Σχέση 8.2), οπότε οι όροι των κλασμάτων των γραμμών 2,3 πρέπει να αντιστραφούν για να είναι έγκυρος ο αλγόριθμος της παραγοντοποίησης.

Τέλος, ο έλεγχος της σύγκλισης και το κριτήριο της σύγκλισης των γραμμών 4 και 5 σχετίζεται άμεσα με την μορφή της πιθανοφάνειας (και κατ΄ επέκταση της εκ των υστέρων πιθανότητας), όπως έχει αναφερθεί στην Ενότητα 8.3.2. Επειδή ως πιθανοφάνεια έχει επιλεγεί η κατανομή Poisson, ο έλεγχος της σύγκλισης αφορά την ελαχιστοποίηση της απόκλισης Kullback-Liebler.

8.4.1 Υπολογισμός της κλίσης

Έχοντας επιλέξει την κατανομή της πιθανοφάνειας (Σχέση 8.25) και την εκ των προτέρων πιθανότητα (Σχέση 8.23), μπορούμε πλέον να χρησιμοποιήσουμε την κλασσική σχέση της μπεϋζιανής συνεπαγωγής (Σχέση 8.3) για να προσεγγίσουμε την εκ των υστέρων πιθανότητα οι παράμετροι του μοντέλου μας (τα στοιχεία του πίνακα $\widetilde{A} = WH$) να λάβουν συγκεκριμένες τιμές, στη βάση των δεδομένων (στοιχεία του πίνακα A) και των υπερπαραμέτρων Θ του μοντέλου

$$\begin{aligned} P(\widetilde{a_{ij}}|a_{ij},\Theta) \propto & \mathcal{L}(a_{ij}|\widetilde{a_{ij}}) \times P(\widetilde{a_{ij}}|\Theta) \times P(\Theta) \Rightarrow \\ P(\widetilde{a_{ij}}|a_{ij},\Theta) \propto & \frac{\widetilde{a_{ij}}^{a_{ij}}}{a_{ij}!}\mathrm{e}^{-\widetilde{a_{ij}}} \times \frac{1}{Z} \times \mathrm{e}^{\Theta\widetilde{a_{i,j}}} \times P(\Theta) \end{aligned} \tag{8.26}$$

Η συνάρτηση κατακερματισμού Z είναι ανεξάρτητη από τις παραμέτρους $\widetilde{a_{ij}}$ και τα δεδομένα a_{ij} του μοντέλου, όπως επίσης και η τιμή της υπερπαραμέτρου Θ προκύπτει ντετερμινιστικά (Σχέση 8.23). Συνεπώς, η Σχέση 8.26 απλοποιείται σε

$$P(\widetilde{a_{ij}}|a_{ij},\Theta) \propto \frac{\widetilde{a_{ij}}^{a_{ij}}}{a_{ij}!}\mathrm{e}^{-\widetilde{a_{ij}}} \times \mathrm{e}^{\Theta\widetilde{a_{i,j}}} \Rightarrow P(\widetilde{a_{ij}}|a_{ij},\Theta) \propto \frac{\widetilde{a_{ij}}^{a_{ij}}}{a_{ij}!}\mathrm{e}^{(\Theta-1)\widetilde{a_{i,j}}} \tag{8.27}$$

Το στοιχείο $\widetilde{a_{ij}}$ προκύπτει από το εσωτερικό γινόμενο του i-οστού διανύσματος γραμμής του πίνακα W επί το j-οστό διάνυσμα στήλης του H. Έτσι, καταλήγουμε στην παρακάτω σχέση

$$P(\mathbf{w_i}^\top\mathbf{h_j}|a_{ij},\Theta) \propto \frac{\mathbf{w_i}^\top\mathbf{h_j}^{a_{ij}}}{a_{ij}!}\mathrm{e}^{(\Theta-1)\mathbf{w_i}^\top\mathbf{h_j}} \tag{8.28}$$

Στόχος είναι να βρεθούν εκείνες οι τιμές για τα $\mathbf{w_i}^\top, \mathbf{h_j}$ που να μεγιστοποιούν την εκ των υστέρων πιθανότητα των παραμέτρων του μοντέλου (δεξί μέρος της Σχέσης 8.28). Επειδή, όμως, η μη-αρνητική παραγοντοποίηση (Εξίσωση 8.2) στην αρχή της Ενότητας 8.1 περιγράφηκε ως πρόβλημα ελαχιστοποίησης (και όχι μεγιστοποίησης) παραμέτρων, πρέπει να μετατρέψουμε την παραπάνω σχέση στο ισοδύναμο πρόβλημα ελαχιστοποίησης. Ο τρόπος που αυτό επιτυγχάνεται στη βιβλιογραφία [Wang and Zhang, 2013] είναι μέσω της λήψης του αρνητικού φυσικού λογαρίθμου της Σχέσης 8.28

$$\mathcal{D}(a_{ij},\mathbf{w_i}^\top\mathbf{h_j}) \equiv -\ln P(\mathbf{w_i}^\top\mathbf{h_j}|a_{ij},\Theta) = \ln(a_{ij}!) - a_{ij}\ln\mathbf{w_i}^\top\mathbf{h_j} + (1-\Theta)\mathbf{w_i}^\top\mathbf{h_j} \tag{8.29}$$

Χρησιμοποιώντας τον προσεγγιστικό τύπο του Stirling ($\ln(x!) = x\ln x - x$) για τον όρο $\ln(a_{ij}!)$, έχουμε

$$\begin{aligned} \mathcal{D}(a_{ij},\mathbf{w_i}^\top\mathbf{h_j}) &= a_{ij}\ln a_{ij} - a_{ij} - a_{ij}\ln\mathbf{w_i}^\top\mathbf{h_j} + (1-\Theta)\mathbf{w_i}^\top\mathbf{h_j} \\ &= a_{ij}\ln\frac{a_{ij}}{\mathbf{w_i}^\top\mathbf{h_j}} - a_{ij} + (1-\Theta)\mathbf{w_i}^\top\mathbf{h_j} \end{aligned} \tag{8.30}$$

Η κλίση της Εξίσωσης 8.30 ως προς τα διανύσματα $\mathbf{w_i}^\top, \mathbf{h_j}$ υπολογίζεται ως εξής

$$\begin{aligned}\nabla\mathcal{D}_{\mathbf{w_i}^\top}(a_{ij}, \mathbf{w_i}^\top\mathbf{h_j}) &= \sum_{j=1}^{k}\left[-\frac{a_{ij}}{\mathbf{w_i}^\top\mathbf{h_j}}\mathbf{h_j}^\top + (1-\Theta)\mathbf{h_j}^\top\right] \\ &= -\frac{\mathbf{a_i}^\top}{\widetilde{\mathbf{a}_\mathbf{i}}^\top}H^\top + (1-\Theta)\mathbf{e}^\top H^\top \qquad (8.31)\\ \nabla\mathcal{D}_{\mathbf{h_j}}(a_{ij}, \mathbf{w_i}^\top\mathbf{h_j}) &= \sum_{i=1}^{k}\left[-\mathbf{w_i}^\top\frac{a_{ij}}{\mathbf{w_i}^\top\mathbf{h_j}} + (1-\Theta)\mathbf{w_i}^\top\right] \\ &= -W^\top\frac{\mathbf{a_i}^\top}{\widetilde{\mathbf{a}_\mathbf{i}}^\top} + (1-\Theta)W^\top\mathbf{e}^\top \qquad (8.32)\end{aligned}$$

ενώ αντίστοιχα αυτά ανανεώνονται ως εξής

$$\begin{aligned}(\mathbf{w_i}^\top)^{(t+1)} \leftarrow (\mathbf{w_i}^\top)^{(t)} \circ \frac{\nabla\mathcal{D}^-_{\mathbf{w_i}^\top}}{\nabla\mathcal{D}^+_{\mathbf{w_i}^\top}} \Rightarrow (\mathbf{w_i}^\top)^{(t+1)} \leftarrow (\mathbf{w_i}^\top)^{(t)} \circ \frac{\frac{\mathbf{a_i}^\top}{\widetilde{\mathbf{a}_\mathbf{i}}^\top}H^\top}{(1-\Theta)\mathbf{e}^\top H^\top} \\ \mathbf{h_j}^{(t+1)} \leftarrow \mathbf{h_j}^{(t)} \circ \frac{\nabla\mathcal{D}^-_{\mathbf{h_j}}}{\nabla\mathcal{D}^+_{\mathbf{h_j}}} \Rightarrow \mathbf{h_j}^{(t+1)} \leftarrow \mathbf{h_j}^{(t)} \circ \frac{W^\top\frac{\mathbf{a_i}^\top}{\widetilde{\mathbf{a}_\mathbf{i}}^\top}}{(1-\Theta)W^\top\mathbf{e}^\top} \qquad (8.33)\end{aligned}$$

καταλήγοντας στους παρακάτω κανόνες ενημέρωσης για τους πίνακες βάσης και συντελεστών W, H

$$W^{(t+1)} \leftarrow W^{(t)} \circ \frac{\frac{A}{WH}H^\top}{(1-\Theta)EH^\top} \qquad (8.34)$$

$$H^{(t+1)} \leftarrow H^{(t)} \circ \frac{W^\top\frac{A}{WH}}{(1-\Theta)W^\top E} \qquad (8.35)$$

Τέλος, πρέπει να σημειωθεί πως για να είναι έγκυροι οι πολλαπλασιαστικοί κανόνες των Εξισώσεων 8.34-8.35, θα πρέπει υποχρεωτικά ο παρονομαστής του κλάσματος των δύο εξισώσεων να είναι θετικός (αφού ο αριθμητής είναι πάντοτε θετικός). Έτσι, και σε συνδυασμό με τον ντετερμινιστικό υπολογισμό της Θ (Εξίσωση 8.23), έχουμε

$$1-\Theta > 0 \Rightarrow \Theta < 1 \Rightarrow \ln\phi_0 + 2(2-3\phi_0)J < 1 \qquad (8.36)$$

Τα κοινωνικά δίκτυα που εξετάζονται απέχουν πολύ από το να χαρακτηριστούν κλίκες, οπότε η τιμή της συνδεσιμότητας ϕ_0 (Εξίσωση 8.16) είναι πολύ χαμηλότερη της μονάδας και πιο κοντά στο μηδέν (δηλαδή ισχύει $\phi_0 << 1$). Συνεπώς, ο όρος $\ln\phi_0$ λαμβάνει πολύ μικρές (αρνητικές) τιμές. Από την άλλη, η υπερπαράμετρος J είναι θετική (η μη-αρνητικότητα της επιβάλλεται από τη χρήση της στο μετασχηματισμό Hubbard-Stratonovich του Παραρτήματος Α΄), οπότε ο δεύτερος όρος του αθροίσματος της Σχέσης 8.36 είναι θετικός. Από διάφορες δοκιμές και μετρήσεις που έγιναν σε γράφους κοινωνικών δικτύων βρέθηκε ότι η επίδραση του όρου $\ln\phi_0$ είναι ισχυρότερη στη διαμόρφωση της τελικής τιμής της υπερπαραμέτρου Θ, οπότε η Σχέση 8.36 ισχύει σε όλα τα δίκτυα που εξετάστηκαν.

8.4.2 Πολλαπλασιαστικοί Κανόνες

Με βάση την ανάλυση της προηγούμενης ενότητας, η επαναληπτική διαδικασία των [Lee and Seung, 1999] (Αλγόριθμος 6), τροποποιείται σύμφωνα με τον Πίνακα 8.1. Η πρώτη στήλη του πίνακα επισημαίνει τα βήματα του Αλγορίθμου 6 που τροποποιούνται κάθε φορά. Η δεύτερη στήλη αποτυπώνει τους πολλαπλασιαστικούς κανόνες ενημέρωσης των παραμέτρων

Πίνακας 8.1: Μπεϋζιανή NMF

Βήμα	Πιθανοφάνεια και εκ των προτέρων πιθανότητα	
	Poisson [Lee and Seung, 2000]	Poisson + 2-star
2	$W^{(t+1)} \leftarrow W^{(t)} \circ \frac{\frac{A}{W^{(t)}H^{(t)}}H^{\top(t)}}{EH^{\top(t)}}$	$W^{(t+1)} \leftarrow W^{(t)} \circ \frac{\frac{A}{W^{(t)}H^{(t)}}H^{\top(t)}}{(1-\Theta)EH^{\top(t)}}$
3	$H^{(t+1)} \leftarrow H^{(t)} \circ \frac{W^{\top(t+1)}\frac{A}{W^{(t+1)}H^{(t)}}}{W^{\top(t+1)}E}$	$H^{(t+1)} \leftarrow H^{(t)} \circ \frac{W^{\top(t+1)}\frac{A}{W^{(t+1)}H^{(t)}}}{(1-\Theta)W^{\top(t+1)}E}$
4	Ελαχιστοποίηση απόκλισης Kullback-Liebler	

W, H του μοντέλου στην περίπτωση που χρησιμοποιηθεί η κατανομή Poisson στη θέση της πιθανοφάνειας, ενώ στην τρίτη στήλη αποτυπώνονται οι αντίστοιχοι κανόνες για την μπεϋζιανή NMF που αποτελούν την πρωτότυπη συνεισφορά της παρούσας εργασίας (κατανομή Poisson στην πιθανοφάνεια και μοντέλο 2-αστέρων ως εκ των προτέρων πιθανότητα).

Καταρχήν παρατηρούμε ότι η τέταρτη γραμμή είναι κοινή και για τα δύο συστήματα (ελαχιστοποίηση απόκλισης Kullback-Liebler) και αυτό γιατί και τα δύο συστήματα χρησιμοποιούν ως πιθανοφάνεια την ίδια κατανομή (Poisson). Οι ομοιότητες, ωστόσο, δεν σταματάνε εδώ. Μια προσεκτικότερη παρατήρηση των Βημάτων 2 και 3 του Πίνακα 8.1 φανερώνει ότι οι κανόνες ενημέρωσης των πινάκων W, H διαφέρουν κατά ένα παράγοντα $\frac{1}{1-\Theta}$. Αυτός ακριβώς ο παράγοντας αποτελεί τη συνεισφορά της εκ των προτέρων πιθανότητας στην παραγοντοποίηση (πιο συγκεκριμένα των ERGM) και η συμβολή του στη συσταδοποίηση των μελών του FoaF δικτύου ενός χρήστη με σκοπό την παραγωγή συστάσεων πρόκειται να εκτιμηθεί στην πειραματική διαδικασία που θα ακολουθήσει.

8.5 Πειραματική Διαδικασία

8.5.1 Συλλογές Δεδομένων

Ο αλγόριθμος που περιγράφηκε προηγουμένως εφαρμόστηκε, ως τμήμα ενός κοινωνικού συστήματος συνεργατικής διήθησης, σε δύο διαφορετικές συλλογές δεδομένων που περιείχαν τόσο αξιολογήσεις χρηστών σε αντικείμενα όσο και πληροφορίες για τις μεταξύ τους κοινωνικές σχέσεις. Πρόκειται για το `lastfm-2k` [Cantador et al., 2011] και το `flixster` [Jamali, 2010], που προέρχονται από τις αντίστοιχες διαδικτυακές υπηρεσίες. Τα χαρακτηριστικά τους συνοψίζονται στον Πίνακα 8.2. Η πρώτη συλλογή είναι μικρή σε μέγεθος και αποτελείται από *δεδομένα χρήσης* (usage data) και πιο συγκεκριμένα από το πλήθος των φορών που κάθε χρήστης της υπηρεσίας άκουσε το εκάστοτε μουσικό κομμάτι. Η δεύτερη συλλογή, από την άλλη, είναι μεσαίου μεγέθους και περιέχει αξιολογήσεις ταινιών στην πενταβάθμια κλίμακα. Παρά τις διαφορές τους, οι δύο συλλογές δεδομένων είναι εξαιρετικά αραιές και εμφανίζουν τα χαρακτηριστικά των δικτύων ελεύθερης κλίμακας τόσο όσον αφορά το πλήθος των αξιολογήσεων που περιέχουν όσο και της κατανομής του βαθμού των κόμβων του κοινωνικού δικτύου (οι περισσότερες ακμές προσπίπτουν σε λίγους κόμβους, ενώ η πλειοψηφία των κόμβων εφάπτεται λίγων ακμών). Η τελευταία παρατήρηση συνάγεται παρατηρώντας τις τελευταίες δύο γραμμές του Πίνακα 8.2 και πιο συγκεκριμένα από τη διαφορά του μέγιστου και του μέσου βαθμού των κόμβων.

8.5.2 Πειραματικό Πρωτόκολλο

Το πειραματικό πρωτόκολλο που εφαρμόστηκε ήταν η αντεπικύρωση με την εξαίρεση ενός. Σε κάθε επανάληψη του πρωτοκόλλου, οι αξιολογήσεις ενός χρήστη εξάγονται από τη συλλογή δεδομένων και χωρίζονται σε δύο διακριτά σύνολα, εκπαίδευσης και δοκιμής. Κατόπιν, το σύνολο δοκιμής επανατοποθετείται στη συλλογή δεδομένων. Στο επόμενο βήμα, οι αλγόριθμοι

Πίνακας 8.2: Συλλογές δεδομένων που χρησιμοποιήθηκαν στα πειράματα

	`lastfm-2k` [Cantador et al., 2011]	`flixster` [Jamali, 2010]
Χρήστες	1.8k	147k
Αντικείμενα	17.6k	48.7k
Αξιολογήσεις	92.8k	8.2M
Πυκνότητα αξιολογήσεων	0,278%	0,114%
Ακμές	25.4k	7M
Πυκνότητα Ακμών	1,42%	0,06%
Μέγιστος Βαθμός	119	1.045
Μέσος Βαθμός	13	15

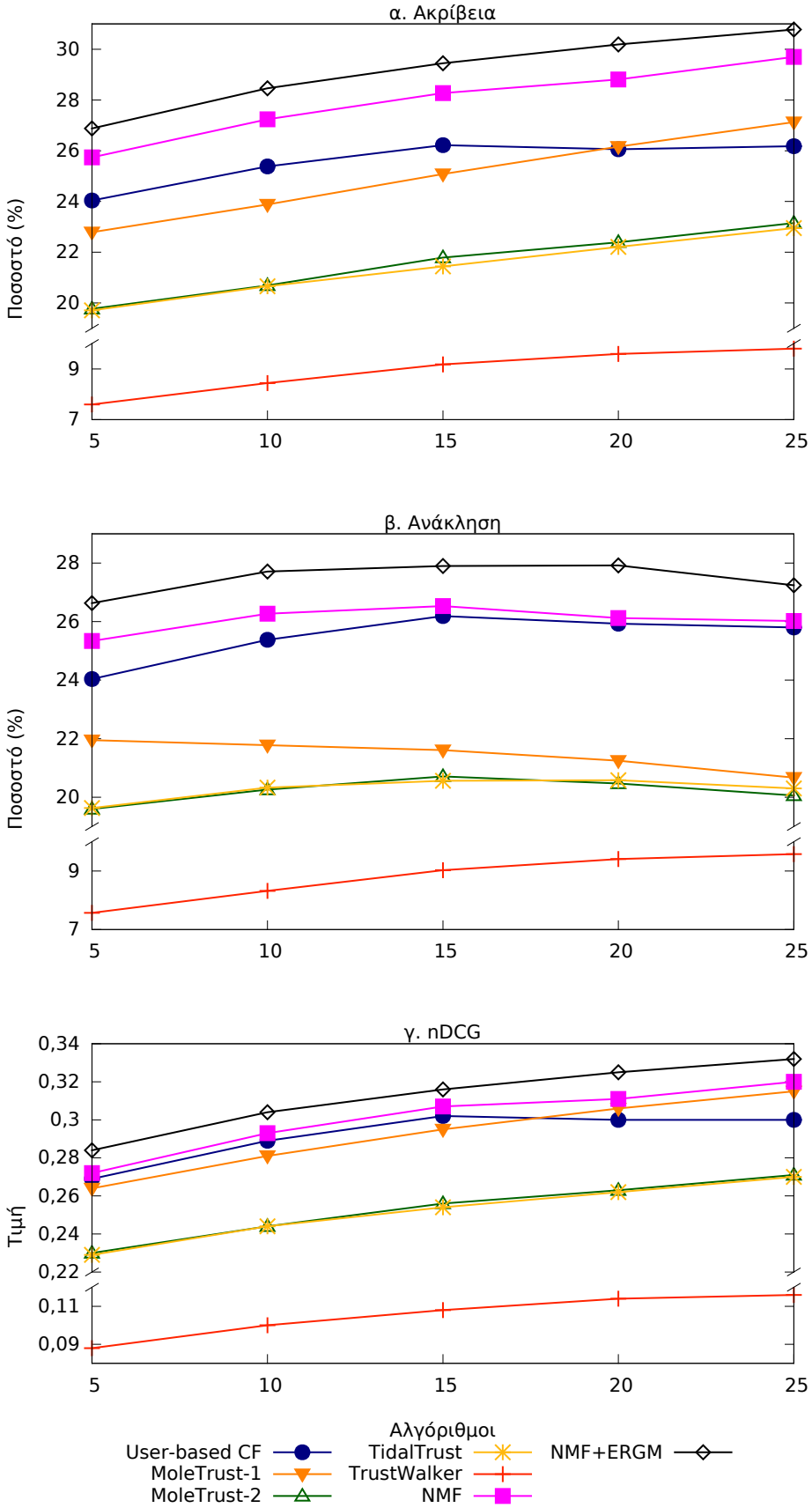
παράγουν συστάσεις και μια λίστα αντικειμένων επιστρέφεται ως έξοδος, η οποία και συγκρίνεται με τα δεδομένα που υπάρχουν στο σύνολο δοκιμής. Η όλη διαδικασία επαναλαμβάνεται πέντε φορές, για μέγεθος λίστας συστάσεων από 5 ως 25 αντικείμενα. Για να έχουν νόημα οι παραγόμενες συστάσεις, σε κάθε επανάληψη του πειραματικού πρωτοκόλλου επιλέγονται μόνο οι χρήστες που έχουν αξιολογήσει τουλάχιστον διπλάσιο πλήθος αντικειμένων από το εκάστοτε μέγεθος της λίστας συστάσεων. Οι στάθμες των μετρικών της ακρίβειας (Εξίσωση 3.8), της ανάκλησης (Εξίσωση 3.10) και της κανονικοποιημένης μειούμενης αθροιστικής απολαβής (Εξίσωση 3.21) που εμφανίζονται στις γραφικές παραστάσεις των Σχημάτων 8.2-8.3 απεικονίζουν τον μέσο όρο των αντίστοιχων τιμών.

Όπως φαίνεται από τις εν λόγω γραφικές παραστάσεις, υλοποιήθηκε ένα πλήθος μνημονικών συνεργατικών συστημάτων συστάσεων με σκοπό την πληρέστερη αξιολόγηση της ποιότητας των παραγόμενων συστάσεων σε διαφορετικά περιβάλλοντα. Ως σύστημα αναφοράς ορίστηκε ο κλασσικός μνημονικός αλγόριθμος συνεργατικής διήθησης (τύπος του Resnick, Ενότητα 2.3.1 και Εξίσωση 2.11) , ο οποίος βασίζει την εκτίμησή του για την ομοιότητα των χρηστών αποκλειστικά και μόνο στις αξιολογήσεις που αυτοί έχουν κάνει, χωρίς να εξετάζει τους δεσμούς μεταξύ τους στο κοινωνικό δίκτυο. Η συνάρτηση ομοιότητας που χρησιμοποιήθηκε σε αυτή την περίπτωση ήταν ο λογάριθμος του λόγου της πιθανοφάνειας (LLR) των δεδομένων, γιατί παρουσίασε την καλύτερη απόδοση στις υπό εξέταση μετρικές, σε σύγκριση με άλλους δείκτες ομοιότητας, όπως ο συντελεστής Pearson ή η ομοιότητα συνημιτόνου.

Στη συνέχεια εξετάστηκαν μετρικές εμπιστοσύνης τοπικής εμβέλειας (Ενότητα 5.5.2) και πιο συγκεκριμένα ο αλγόριθμος *MoleTrust-1* (που θεωρεί ως όμοιους μόνο τους χρήστες με τους οποίους συνδέεται άμεσα με ακμή ο υπό εξέταση χρήστης στο κοινωνικό δίκτυο) και ο *MoleTrust-2* (που λαμβάνει υπόψη του και τους γείτονες των γειτόνων του υπό εξέταση χρήστη). Επίσης υλοποιήθηκε και ο *TidalTrust*, ο οποίος υπολογίζει την ομοιότητα χρηστών στη βάση των συντομότερων μονοπατιών μεταξύ του υπό εξέταση χρήστη και όλων των υπόλοιπων χρηστών που βρίσκονται στην ίδια συνδεδεμένη συνιστώσα με αυτόν.

Ακόμα, στη σύγκριση περιλαμβάνεται και ένας αλγόριθμος που υπολογίζει τη φήμη του κάθε χρήστη στο κοινωνικό δίκτυο (Ενότητα 5.5.3) και πιο συγκεκριμένα ο *TrustWalker* [Jamali and Ester, 2009], ο οποίος πραγματοποιεί έναν τυχαίο περίπατο στον γράφο, επιλέγοντας το επόμενο βήμα του ομοιόμορφα τυχαία. Ο τυχαίος περίπατος έχει αφετηρία τον υπό εξέταση χρήστη και όταν φτάσει τη στάσιμη του κατανομή τότε επιστρέφονται ως περισσότερο όμοιοι οι κόμβοι εκείνοι οι οποίοι έχουν τη μεγαλύτερη πιθανότητα.

Τέλος, η μεθοδολογία συσταδοποίησης του FoaF δικτύου των χρηστών που παρουσιάστηκε στο παρόν κεφάλαιο, τέθηκε υπό δοκιμή, έτσι ώστε να μπορέσει να εκτιμηθεί η σχετική απόδοση των ERGM στη θέση της εκ των προτέρων πιθανότητας. Οι δύο εναλλακτικές διαμορφώσεις παρουσιάζονται στις αντίστοιχες στήλες του Πίνακα 8.1. Η πρώτη διαμόρφωση περιλαμβάνει τον μπεϋζιανό *NMF* αλγόριθμο με τη χρήση της κατανομής Poisson στη θέση της πιθανοφάνειας, ενώ η δεύτερη διαμόρφωση αποτελεί την επέκταση της πρώτης με την εισα-



Σχήμα 8.2: Αποτελέσματα στο `lastfm-2k`

γωγή του μοντέλου 2-αστέρων στη θέση της εκ των προτέρων πιθανότητας *(NMF+ERGM)*. Και οι δύο αλγόριθμοι εφαρμόστηκαν στον πίνακα γειτνίασης που προέκυπτε από το FoaF δίκτυο του εκάστοτε χρήστη.

Η μέθοδος φιλτραρίσματος των γειτόνων που εφαρμόστηκε ήταν αυτή των N-*πλησιέστερων* (Nearest-N), δηλαδή στη διαδικασία παραγωγής των συστάσεων λαμβάνονται υπόψη μόνο οι N πιο κοντινοί γείτονες. Η τιμή της υπερπαραμέτρου N ορίστηκε στο 5 μετά από μια σειρά πειραμάτων επαλήθευσης στα οποία βρέθηκε ότι για τιμές του N μικρότερες του 5 τα αποτελέσματα των μετρικών ήταν ασταθή ενώ για τιμές του N μεγαλύτερες του 5 τα αποτελέσματα ήταν υποδεέστερα. Πρέπει, ωστόσο, να σημειωθεί ότι η σχετική κατάταξη των αλγορίθμων που απεικονίζονται στα Σχήματα 8.2-8.3 παραμένει σταθερή, ανεξάρτητα από την τιμή του N.

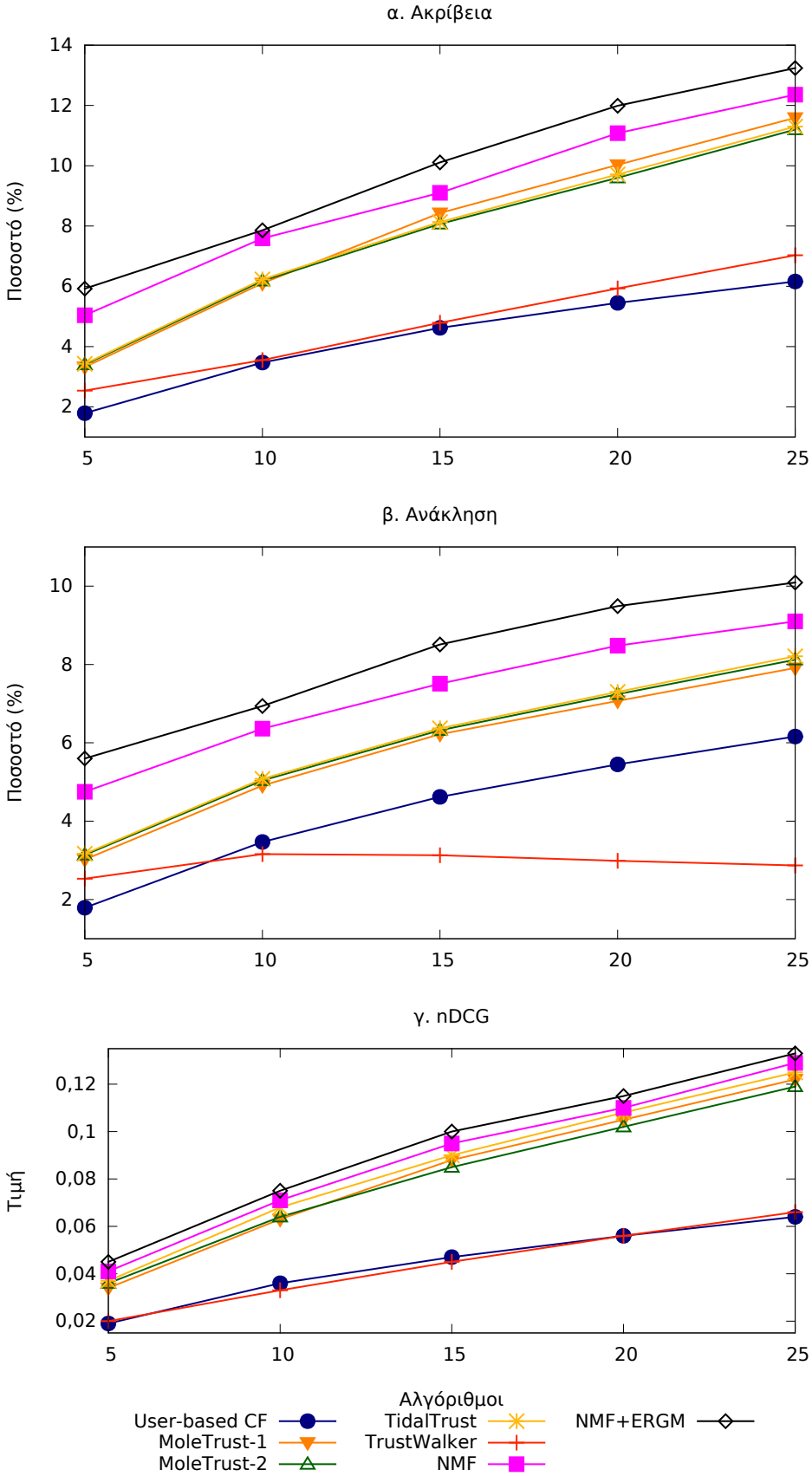
8.6 Αποτελέσματα

Μια πρώτη παρατήρηση των αποτελεσμάτων που αποτυπώνονται στις γραφικές παραστάσεις των Σχημάτων 8.2-8.3 είναι πως όλοι οι αλγόριθμοι σημειώνουν σημαντικά χαμηλότερη απόδοση στη συλλογή δεδομένων του `flixster` σε σύγκριση με το `lastfm-2k`. Η συμπεριφορά αυτή αποδίδεται σε δύο παράγοντες· αφενός το `lastfm-2k` παρουσιάζει διπλάσια πυκνότητα σε σύγκριση με το `flixster` και αφετέρου περιέχει δεδομένα χρήσης. Στην περίπτωση αυτή, κάθε αλληλεπίδραση μεταξύ χρηστών και αντικειμένων θεωρείται ωφέλιμη και μπορεί να αποτελέσει τη βάση πιθανής σύστασης. Από την άλλη, η συλλογή δεδομένων του `flixster` αποτελείται από αξιολογήσεις ταινιών και μόνο όσες είναι από 3 αστέρια και άνω (περίπου το 63% της βάσης) θεωρούνται ωφέλιμες. Υπό αυτή την οπτική, το `flixster` εμφανίζεται να έχει μόλις το ένα τέταρτο της πυκνότητας του `lastfm-2k`. Μια επίσης ενδιαφέρουσα επισήμανση είναι πως ο λόγος της πυκνότητας μεταξύ των δύο συλλογών δεδομένων αντικατοπτρίζεται σχεδόν γραμμικά και στα αντίστοιχα αποτελέσματα που επιτυγχάνουν τα διάφορα συστήματα, με την εξαίρεση του *TrustWalker*.

Μια εξίσου σημαντική παρατήρηση είναι πως στις αραιές συλλογές δεδομένων, το προσωπικό δίκτυο του κάθε χρήστη αποτελεί μια καλή πληροφοριακή πηγή για την παραγωγή συστάσεων, εξίσου καλή με τις κλασσικές συνεργατικές μεθόδους. Αυτό καθίσταται εμφανές όταν συγκρίνεται η απόδοση των αλγορίθμων *User-based CF* και *MoleTrust-1* στο `lastfm-2k`: παρότι ο πρώτος φαίνεται να ξεπερνά τον δεύτερο, το προβάδισμα είναι οριακό και δεν μπορεί να θεωρηθεί ως μια καθαρή ένδειξη της ανωτερότητας της συγκεκριμένης μεθόδου σε σύγκριση με την άλλη. Η κατάσταση, ωστόσο, είναι διαφορετική στο `flixster`, όπου οι μεθοδολογίες που μελετούν αθροιστικά τις ακμές στο δίκτυο FoaF (και στην περίπτωση του *TidalTrust*, ακόμα βαθύτερα) του κάθε χρήστη παρουσιάζουν εμφανή υστέρηση σε όλες τις μετρικές, δείγμα του γεγονότος ότι σε αρκετές περιπτώσεις το δίκτυο FoaF περιέχει «θόρυβο», ο οποίος πρέπει να φιλτραριστεί προτού παραχθούν καλές συστάσεις.

Όσο αυξάνει το επίπεδο της αραιότητας των δεδομένων η αποδοτικότητα του κλασσικού αλγορίθμου συνεργατικών συστάσεων (*User-based CF*) πέφτει, μιας και υπάρχουν ολοένα και λιγότεροι χρήστες των οποίων οι αξιολογήσεις στα αντικείμενα να συμπίπτουν. Σε αυτή την περίπτωση είναι που οι κοινωνικοί αλγόριθμοι «ξεδιπλώνουν» όλη τους τη δυναμική. Τα αποτελέσματα στο `flixster` (Σχήμα 8.3) αποδεικνύουν την εμφανή ανωτερότητα των κοινωνικών αλγορίθμων που πραγματοποιούν τοπική αναζήτηση σε όλες τις μετρικές. Άξιο επισήμανσης είναι επίσης πως και οι τρεις αλγόριθμοι που βασίζονται στην τοπική αναζήτηση του κοινωνικού δικτύου εμφανίζουν παρόμοια αποτελέσματα παρότι εξερευνούν τη γειτονιά του κάθε χρήστη σε διαφορετικό βάθος.

Από την άλλη, ο αλγόριθμος *TrustWalker* δεν παρουσιάζει ικανοποιητική απόδοση σε καμία από τις δύο συλλογές δεδομένων: απέχει πολύ από όλους τους υπόλοιπους αλγορίθμους στο `lastfm-2k` και βρίσκεται στα ίδια επίπεδα με τον τελευταίο στο `flixster`. Επίσης το *TrustWalker* είναι το μόνο σύστημα που παρουσιάζει τα ίδια επίπεδα απόδοσης και στις δύο συλλογές δεδομένων ανεξάρτητα από την πυκνότητά τους. Η παραπάνω παρατήρηση οδηγεί



Σχήμα 8.3: Αποτελέσματα στο `flixster`

στο συμπέρασμα πως το να βασίζονται οι συστάσεις αποκλειστικά στους πιο δημοφιλής (ή στους πιο συχνά επισκεπτόμενους) κόμβους του κοινωνικού δικτύου δεν αποτελεί εγγύηση ομοιότητας στις προτιμήσεις.

Τέλος, οι προτεινόμενες στο παρόν κεφάλαιο μεθοδολογίες (*NMF* και *NMF+ERGM*) φαίνεται πως επιτυγχάνουν καλύτερα αποτελέσματα τόσο σε σύγκριση με την τοπική όσο και με την ολική αναζήτηση στο κοινωνικό δίκτυο. Η τοποθέτηση των κόμβων που αποτελούν μέρος του FoaF δικτύου ενός χρήστη σε αλληλεπικαλυπτόμενες συστάδες οδηγεί σε μια πληρέστερη ανάλυση της εγγύτητας των χρηστών στο δίκτυο, ειδικά όταν συγκρίνεται με τις βασικές παραδοχές που κάνουν τα *MoleTrust-2* και *TidalTrust*. Η συγκεκριμένη ανάλυση βελτιώνεται περαιτέρω με την προσθήκη των ERGM στη θέση της εκ των προτέρων πιθανότητας, μιας και με αυτό τον τρόπο αντιμετωπίζεται η κεντρική θεώρηση που δίνει στην κάθε ακμή η μπεϋζιανή NMF, με την εισαγωγή δηλαδή δομικών χαρακτηριστικών του γράφου στη διαδικασία.

8.7 Συμπεράσματα

Στο παρόν κεφάλαιο παρουσιάστηκε μια μεθοδολογία η οποία βασίζεται στην τοπική συσταδοποίηση της FoaF γειτονιάς του κάθε χρήστη καθώς και η ενσωμάτωσή της στα μνημονικά, κοινωνικά και συνεργατικά συστήματα συστάσεων. Ο προτεινόμενος αλγόριθμος (*NMF+ERGM*) εισαγάγει στην παραγοντοποίηση δομικά χαρακτηριστικά του δικτύου, μέσω της χρήσης των μοντέλων εκθετικού τυχαίου γράφου στη θέση της εκ των προτέρων κατανομής.

Τα επιτευχθέντα αποτελέσματα αναδεικνύουν μια σταθερή βελτίωση όλων των μετρικών μέτρησης της απόδοσης σε σύγκριση με όλες τις υπόλοιπες υλοποιήσεις, δηλαδή τον κλασσικό συνεργατικό αλγόριθμο *User-based CF* καθώς και τους κοινωνικούς αλγορίθμους, που είτε περιορίζονται σε μια περιοχή γύρω από κάθε χρήστη είτε εξετάζουν συνολικά τον κοινωνικό γράφο. Η βελτίωση αυτή αποδίδεται στις ειδικές δυνατότητες φιλτραρίσματος της προτεινόμενης μεθοδολογίας στο επίπεδο FoaF. Αντί απλά να αθροίσει είτε πάρα πολλούς είτε πολύ λίγους χρήστες, προσπαθεί να ανακαλύψει πρότυπα στην κοινωνική τους συμπεριφορά με την (επικαλυπτόμενη) συσταδοποίησή τους σε περιοχές, μια διαδικασία που παρουσιάζει ομοιότητες με τον τρόπο που οι αλγόριθμοι εύρεσης επικαλυπτόμενων κοινοτήτων λειτουργούν.

Παρόλα αυτά, υπάρχει χώρος για περαιτέρω βελτίωση. Μια πιθανή ερευνητική κατεύθυνση θα ήταν η συμπερίληψη στη διαδικασία της παραγοντοποίησης πιο περίπλοκων χαρακτηριστικών του δικτύου, όπως είναι τα τρίγωνα. Σε αυτή την περίπτωση, ωστόσο, οι υπολογισμοί του μοντέλου γίνονται αρκετά περίπλοκοι και δεν είναι εύκολη η εξαγωγή μιας προσεγγιστικής λύσης ανάλογης με αυτή που παρουσιάστηκε στην Ενότητα 8.3.1. Στην εργασία των [Park and Newman, 2005] παρουσιάζονται ορισμένες ιδέες, ωστόσο μια πιο γενικευμένη προσέγγιση για την εκτίμηση των παραμέτρων του μοντέλου θα επιτύγχανε καλύτερα αποτελέσματα.

□

Κεφάλαιο 9

Συνολικό Πόρισμα Διατριβής

9.1 Γενικά Συμπεράσματα

Η συνεισφορά της διατριβής εντοπίζεται στη μελέτη, την ανάπτυξη και στην πρακτική εφαρμογή αλγορίθμων για τα Συστήματα Συστάσεων και πιο συγκεκριμένα για την πιο διαδεδομένη κατηγορία τους, τα Συνεργατικά Συστήματα Συστάσεων. Αρχικά προτείνεται μια νέα μέθοδος αξιοποίησης της έμμεσης κοινωνικής πληροφορίας που μπορεί να εξαχθεί από την αλληλεπίδραση των χρηστών με το σύστημα. Στη συνέχεια προτείνονται τρεις διαφορετικοί τρόποι συνδυασμού της άμεσης και της έμμεσης κοινωνικής πληροφορίας για την παραγωγή περισσότερο αποδοτικών συστάσεων.

Το πρώτο μέρος της παρούσας διατριβής αφιερώνεται στην παρουσίαση των συστημάτων συστάσεων καθώς και στις κυριότερες προκλήσεις που αντιμετωπίζουν. Αυτές είναι η αραιότητα των αξιολογήσεων, η ψυχρή εκκίνηση και η εμπιστοσύνη στις παραγόμενες συστάσεις. Η αραιότητα των αξιολογήσεων έχει να κάνει με το γεγονός πως στα συστήματα συστάσεων το πλήθος των ζευγών χρηστών-αντικειμένων για τα οποία υπάρχει προτίμηση είναι ένα πολύ μικρό ποσοστό του συνόλου (συνήθως κάτω από 1%) και συνεπώς οι αλγόριθμοι καλούνται να παράξουν συστάσεις έχοντας πολύ λίγα δεδομένα στη διάθεσή τους. Η ψυχρή εκκίνηση αναφέρεται στην κατάσταση εκείνη κατά την οποία οι περισσότεροι χρήστες (αντικείμενα) έχουν εκφράσει (δεχτεί) ελάχιστες ως καθόλου προτιμήσεις και συνεπώς είναι δύσκολη η εξαγωγή συμπερασμάτων για το γούστο (χρησιμότητα) τους. Τέλος, το πρόβλημα της εμπιστοσύνης αφορά το πόσο «σίγουρο» είναι το ίδιο το σύστημα για τις προτάσεις που κάνει.

Κατόπιν παρουσιάζονται οι δημοφιλέστεροι τρόποι αναπαράστασης των προτιμήσεων των χρηστών στα συστήματα συστάσεων, δηλαδή σε μορφή πίνακα και σε μορφή διμερούς γράφου. Στη συνέχεια πραγματοποιείται μια επισκόπηση των κυριότερων μεθοδολογιών για την παραγωγή συστάσεων (βασισμένη στο περιεχόμενο, δημογραφική, κ.λ.π.) και δίνεται έμφαση στα συνεργατικά συστήματα συστάσεων. Η έννοια της συνεργατικότητας έγκειται στο γεγονός πως τα νέα αντικείμενα που τελικά προτείνονται σε ένα χρήστη βασίζονται στις αξιολογήσεις που αυτά έχουν λάβει από άλλους «παρόμοιους» χρήστες. Δηλαδή, τα συνεργατικά συστήματα συστάσεων μετατρέπουν τον κάθε χρήστη σε «μέσο πρόβλεψης» των προτιμήσεων των άλλων, καθορίζοντας κατ΄ αυτόν τον τρόπο μια άτυπη μορφή κοινωνικότητας μεταξύ τους. Επιπρόσθετα, αναλύεται και η περαιτέρω κατηγοριοποίηση των συνεργατικών συστημάτων σε μνημονικά συστήματα και συστήματα κατασκευής μοντέλου. Τα πρώτα συνδυάζουν με απλούς, αλγεβρικούς τρόπους όλα τα υπάρχοντα δεδομένα προκειμένου να παράξουν συστάσεις ενώ τα δεύτερα κατασκευάζουν ένα μοντέλο για τον κάθε χρήστη στη βάση των διαθέσιμων δεδομένων.

Στη συνέχεια γίνεται μια εκτενής και πολύπλευρη παρουσίαση των τρόπων μέτρησης της απόδοσης των συστημάτων συστάσεων. Αναφέρονται τόσο οι παραδοσιακές μετρικές της ακρίβειας της πρόβλεψης (μέσο τετραγωνικό σφάλμα, μέσο απόλυτο σφάλμα κ.λ.π.) και της

ακρίβειας της ταξινόμησης (ακρίβεια, ανάκληση κ.λ.π.) όσο και νέες μετρικές που εστιάζουν στην ποιότητα των παραγόμενων συστάσεων, όπως η καινοτομία, η ποικιλομορφία και η κανονικοποιημένη μειούμενη αθροιστική απολαβή. Επίσης αναλύονται συγκεκριμένες υποκατηγορίες (όψεις) χρηστών και αντικειμένων του συστήματος, για τους οποίες ενδιαφέρει ιδιαίτερα η απόδοσή του (όπως λ.χ. οι χρήστες με λίγες αξιολογήσεις).

Το πρώτο μέρος ολοκληρώνεται με την παρουσίαση ενός πρωτότυπου συνεργατικού συστήματος συστάσεων, το οποίο αντιμετωπίζει σε ικανοποιητικό βαθμό τα ζητήματα της αραιότητας των αξιολογήσεων και της ψυχρής εκκίνησης. Για κάθε χρήστη του συστήματος κατασκευάζεται ένα μοντέλο, δομικό στοιχείο του οποίου είναι ένα τεχνητό νευρωνικό δίκτυο. Η καινοτομία της πρότασης έγκειται σε τρία σημεία: στη συνάρτηση μεταφοράς που χρησιμοποιείται στους νευρώνες, στον αλγόριθμο μεταβλητής αρχιτεκτονικής του δικτύου και στον αλγόριθμο ενισχυτικής μάθησης στη φάση εκπαίδευσης. Η χρησιμοποιούμενη συνάρτηση μεταφοράς βασίζεται στην κ-διαχωρισιμότητα. Είναι διαφορετική από αυτές που έχουν μέχρι σήμερα χρησιμοποιηθεί στα συστήματα συστάσεων (λ.χ σιγμοειδείς, γραμμικές, κ.λ.π) και επιτυγχάνει καλύτερη ταξινόμηση των δεδομένων εισόδου της. Αυτό έχει ως άμεσο αποτέλεσμα την μείωση των απαιτούμενων διαστάσεων του δικτύου, πράγμα που έχει ως συνέπεια την αύξηση της ταχύτητας εκπαίδευσης και της παραγωγής νέων συστάσεων. Ο αλγόριθμος μεταβλητής αρχιτεκτονικής δικτύου βασιζόμενος σε έναν ήδη υπάρχοντα κατασκευαστικό αλγόριθμο δικτύου, τον τροποποιεί και τον επεκτείνει με τέτοιο τρόπο ώστε το δίκτυο να οδηγηθεί στη βέλτιστή του αρχιτεκτονική, αποφεύγοντας τόσο τις περιπτώσεις υπερεκπαίδευσης όσο και ελλιπούς εκπαίδευσης. Τέλος, ο αλγόριθμος ενισχυτικής μάθησης (στη φάση της εκπαίδευσης του δικτύου) δεν παρέχει ομοιόμορφα τυχαία τα δείγματα εκπαίδευσης στο δίκτυο. Αντίθετα, το τροφοδοτεί πιο συχνά με εκείνα τα δείγματα για τα οποία παρατηρείται το μεγαλύτερο σφάλμα εκπαίδευσης, έτσι ώστε αυτά να «αφομοιωθούν» καλύτερα.

Το δεύτερο μέρος της διατριβής εξετάζει την ενσωμάτωση της άμεσης κοινωνικής πληροφορίας στα συνεργατικά συστήματα συστάσεων. Συνηθέστερα αυτή έχει τη μορφή ενός κοινωνικού δικτύου μεταξύ των χρηστών του συστήματος και για το λόγο αυτό τα κοινωνικά δίκτυα αρχικά αναλύονται στα δομικά τους χαρακτηριστικά (δράστες, σχεσιακοί δεσμοί, ομάδες κλπ) υπό το πρίσμα της επιστημονικής περιοχής της ανάλυσης κοινωνικών δικτύων. Κατόπιν, πραγματοποιείται μια επισκόπηση των τύπων των δικτύων που χρησιμοποιούνται συνηθέστερα στα συνεργατικά συστήματα συστάσεων (πλήρη, μερικά, προσωπικά) και επίσης γίνεται πλήρης μνεία στους κυριότερους δεσμούς που αναπτύσσονται μεταξύ των χρηστών (φιλία, εμπιστοσύνη, σχέσεις «ακολούθου» κλπ) καθώς και μεταξύ χρηστών και αντικειμένων (οντολογίες και συνεργατικές ταξινομήσεις). Τέλος, πραγματοποιείται ειδική αναφορά στα δίκτυα εμπιστοσύνης, μιας και έχουν αξιοποιηθεί κατά κόρον στα κοινωνικά συστήματα συστάσεων, και αναλύεται η ειδική συνεισφορά τους στην βελτίωση της απόδοσης των παραγόμενων συστάσεων.

Στη συνέχεια προτείνεται ένας πρωτότυπος αλγόριθμος κοινωνικής συνεργατικής διήθησης, ο οποίος βασίζεται στην πραγματοποίηση τυχαίων περιπάτων επάνω στο μεικτό γράφο ομοιότητας και εμπιστοσύνης για την παραγωγή συστάσεων. Ιδιαίτερο χαρακτηριστικό της πρότασης είναι η σε κάθε βήμα πιθανοτική επιλογή αν ο επόμενος σταθμός του περιπάτου θα είναι κάποιος χρήστης που ο τρέχοντας εμπιστεύεται ή θα είναι κάποιος που παρουσιάζει παρόμοια αξιολογική συμπεριφορά με αυτόν. Ακόμα, το επόμενο βήμα του περιπάτου δεν επιλέγεται ομοιόμορφα τυχαία στην περίπτωση μετάβασης σε όμοιο χρήστη. Αντ΄ αυτού, η ομοιότητα στην αξιολογική συμπεριφορά μεταξύ του τρέχοντος χρήστη και των πιθανών επόμενων μοντελοποιείται ως μια πιθανοτική κατανομή, από την οποία προκύπτουν οι συγκεκριμένες πιθανότητες μετάβασης στον κάθε έναν τους. Επειδή η προαναφερόμενη κατανομή είναι άγνωστη, χρησιμοποιείται η τεχνική της απορριπτικής δειγματοληψίας για την λήψη δειγμάτων, έτσι ώστε ο τυχαίος περίπατος να πραγματοποιήσει το επόμενο βήμα του.

Εκτός του προαναφερόμενου αλγορίθμου, προτείνεται επίσης μια πρωτότυπη μέθοδος παραγωγής συστάσεων η οποία βασίζεται στην προσωπική συσταδοποίηση των αντικειμένων που

βρίσκονται στο περιβάλλον του εκάστοτε χρήστη. Πιο συγκεκριμένα ορίζεται το προσωπικό δίκτυο του κάθε χρήστη, το οποίο περιλαμβάνει τους χρήστες τους οποίους αυτός εμπιστεύεται καθώς και τους χρήστες με τους οποίους παρουσιάζει ομοιότητα στην αξιολογική συμπεριφορά. Κατόπιν κατασκευάζεται ο γράφος που περιέχει τα αντικείμενα που έχουν αξιολογηθεί από κοινού από τον υπό εξέταση χρήστη και τα μέλη του προσωπικού του δικτύου. Στο συγκεκριμένο γράφο εφαρμόζεται ένας αλγόριθμος φασματικής συσταδοποίησης προκειμένου να εξαχθούν οι συστάδες στις οποίες εντάσσονται τα εν λόγω αντικείμενα. Για κάθε συστάδα κατασκευάζεται ένα δίκτυο κατανάλωσης αντικειμένων το οποίο, εκτός των αντικειμένων της συστάδας, περιέχει και άλλα αντικείμενα τα οποία έχουν αξιολογήσει τα μέλη του προσωπικού δικτύου του υπό εξέταση χρήστη, αλλά όχι ο ίδιος. Τέλος, οι συστάσεις προκύπτουν με την πραγματοποίηση ενός τυχαίου περιπάτου στο συγκεκριμένο δίκτυο.

Πέρα από τις μεθόδους πραγματοποίησης τυχαίων περιπάτων, προτείνεται η χρήση μεθόδων παραγοντοποίησης πινάκων σε ένα εντελώς νέο πλαίσιο στα κοινωνικά συστήματα συστάσεων. Συνήθως, η προαναφερόμενη τεχνική χρησιμοποιείται για την εύρεση των λανθάνοντων παραγόντων που περιγράφουν τις επιλογές των χρηστών και τα χαρακτηριστικά των αντικειμένων. Στη συγκεκριμένη περίπτωση, ωστόσο, το πεδίο εφαρμογής της μεθόδου αλλάζει· χρησιμοποιείται για την ανεύρεση εκείνων των λανθάνοντων παραγόντων που καθορίζουν τις κοινότητες στις οποίες συμμετέχουν οι «φίλοι-φίλων» ενός συγκεκριμένου χρήστη. Πιο συγκεκριμένα, η προτεινόμενη μέθοδος τροποποιεί διαδεδομένους αλγορίθμους μπεϋζιανής μη-αρνητικής παραγοντοποίησης πινάκων, έτσι ώστε αυτοί να μπορούν να προσαρμοστούν καλύτερα στο υπό εξέταση πρόβλημα. Η προσαρμογή επιτυγχάνεται με την χρήση κατάλληλης συνάρτησης εκ των προτέρων πιθανότητας, η οποία μετριάζει την ιδιαίτερα τοπική φύση του αλγορίθμου παραγοντοποίησης, εισάγοντας στην διαδικασία μακροσκοπικές πληροφορίες για το δίκτυο. Στην συνέχεια, επιλέγεται η κατάλληλη συνάρτηση εκ των προτέρων πιθανότητας από τα μοντέλα εκθετικών τυχαίων γράφων και όχι από τη θεωρία των συζυγών εκ των προτέρων κατανομών (όπως γίνεται σε άλλες τεχνικές). Και πάλι, ο λόγος που επιβάλλει αυτή την επιλογή είναι τα ιδιαίτερα χαρακτηριστικά του δικτύου που αναλύεται. Κατόπιν, παρουσιάζεται αναλυτικά το μοντέλο 2-αστέρων, που περιγράφει παραστατικά τα δίκτυα «φίλων-φίλου», και στη συνέχεια αναπτύσσεται η μέθοδος του προσεγγιστικού υπολογισμού των υπερπαραμέτρων του μοντέλου, κάνοντας χρήση της θεωρίας του μέσου πεδίου.

Τέλος, κάθε τεχνική που αναπτύχθηκε στο πλαίσιο της διατριβής (κοινωνική ή μη), αξιολογήθηκε πειραματικά σε δημόσια διαθέσιμες συλλογές δεδομένων, οι οποίες χρησιμοποιούνται ευρέως από την επιστημονική κοινότητα, ενώ έγιναν και συγκρίσεις με άλλες αντίστοιχες τεχνικές στο κάθε πεδίο έρευνας.

9.2 Μελλοντικές Επεκτάσεις

Από την μέχρι στιγμής ανάλυση που έχει πραγματοποιηθεί στο πλαίσιο της διατριβής, είναι φανερό ότι τα Συστήματα Συστάσεων αποτελούν ένα ιδιαίτερα ενεργό επιστημονικό πεδίο. Σε αυτό συνηγορούν τόσο το πλήθος των πρακτικών εφαρμογών (μερικές εκ των οποίων αναφέρθηκαν στο Κεφάλαιο 1) όσο και το πλήθος και το εύρος των δημοσιευμένων εργασιών στα αντίστοιχα ακαδημαϊκά περιοδικά και συνέδρια. Επίσης, πολλοί αλγόριθμοι συστάσεων έχουν ήδη υλοποιηθεί σε πειραματικό στάδιο, με την απόδοσή τους να έχει μετρηθεί σε δημόσιες και ιδιωτικές συλλογές δεδομένων, ορισμένες εκ των οποίων χρησιμοποιήθηκαν και σε αυτή τη διατριβή. Με την πρακτική υλοποίηση των συστημάτων συστάσεων και τον έλεγχο της απόδοσής τους εξάλλου, ελέγχεται ο βαθμός που αντιμετωπίζουν τα διάφορα ζητήματα και τις προκλήσεις που ανακύπτουν, ενώ επίσης με αυτόν τον τρόπο προκύπτουν καινούργιες ερευνητικές ιδέες και κατευθύνσεις.

Σε αυτό το πλαίσιο, οι μεθοδολογίες που προτάθηκαν στη διατριβή, παρότι επιτυγχάνουν ικανοποιητικά αποτελέσματα, μπορούν να βελτιωθούν ακόμα περισσότερο. Για παράδειγμα, το σύστημα συνεργατικής διήθησης κατασκευής μοντέλου που παρουσιάστηκε στο Κεφάλαιο

4, μπορεί να βελτιωθεί και άλλο, ιδιαίτερα όσον αφορά τον κατασκευαστικό αλγόριθμο του δικτύου. Ένας πιο καλός αλγόριθμος θα μπορούσε να περιορίσει περαιτέρω τις διαστάσεις του δικτύου καθώς και να ενισχύσει την ικανότητα γενίκευσης του δικτύου σε ακόμα μεγαλύτερο βαθμό. Επίσης, η απόδοση του όλου συστήματος θα μπορούσε να βελτιωθεί αισθητά, αν σε αυτό πραγματοποιηθεί εμπλουτισμός με πληροφορίες ανάδρασης από τους χρήστες (user feedback).

Στο Κεφάλαιο 6 αναπτύχθηκε ένας αλγόριθμος παραγωγής συστάσεων, ο οποίος βασίζεται στην πραγματοποίηση τυχαίων περιπάτων στον μεικτό γράφο εμπιστοσύνης και ομοιότητας. Η προτεινόμενη μέθοδος θα μπορούσε να επεκταθεί προς την κατεύθυνση της ενοποίησης των δύο δικτύων σε ένα, με την μοντελοποίηση της εμπιστοσύνης και της ομοιότητας ως οριακών πιθανοτήτων μιας άγνωστης κοινής κατανομής, η οποία κατόπιν θα πρέπει να προσεγγιστεί. Έτσι, η ομοιότητα των χρηστών μπορεί να προκύψει απευθείας μέσω της δειγματοληψίας από την εν λόγω κοινή κατανομή.

Η ίδια κατεύθυνση μπορεί να ακολουθηθεί και στο σύστημα της προσωποποιημένης συσταδοποίησης των αντικειμένων που παρουσιάστηκε στο Κεφάλαιο 7. Στη συγκεκριμένη περίπτωση, ωστόσο, υπάρχουν και άλλα σημεία που θα μπορούσαν να οδηγήσουν στην ακόμα μεγαλύτερη βελτίωση του συστήματος. Για παράδειγμα, ο αλγόριθμος της φασματικής συσταδοποίησης θα μπορούσε να βελτιωθεί περαιτέρω με την εισαγωγή κριτηρίων που θα εξετάζουν το μέγεθος και την ποιότητα των παραγόμενων συστάδων. Αυτό μπορεί να καταστεί εφικτό με την χρήση άλλων τεχνικών συσταδοποίησης, όπως η ασαφής k-means συσταδοποίηση μαζί με διαφορετικές συναρτήσεις απόστασης (π.χ απόσταση Chebysev ή Mahalanobis) για την τοποθέτηση των αντικειμένων στις συστάδες. Επίσης, οι ιδιότητες του τυχαίου περιπάτου στο δίκτυο κατανάλωσης αντικειμένων (όπως λ.χ ο ρυθμός μίξης) μπορούν να βελτιστοποιηθούν αν ληφθούν υπόψη διαφορετικές πιθανότητες μετάβασης, οι οποίες θα συμπεριλαμβάνουν περισσότερες πληροφορίες σχετικές με τα χαρακτηριστικά του κάθε κόμβου.

Τέλος, ο μπεϋζιανός αλγόριθμος μη-αρνητικής παραγοντοποίησης του πίνακα γειτνίασης που παρουσιάστηκε στο Κεφάλαιο 8 μπορεί να αναπτυχθεί περαιτέρω, με την συμπερίληψη στα μοντέλα εκθετικών τυχαίων γράφων και άλλων χαρακτηριστικών που απατώνται στα δίκτυα «φίλων-φίλου», όπως λ.χ. είναι τα τρίγωνα. Σε αυτή την περίπτωση, ωστόσο, οι υπολογισμοί του μοντέλου γίνονται αρκετά περίπλοκοι και δεν είναι εύκολη η εξαγωγή μιας προσεγγιστικής λύσης ανάλογης με αυτή που παρουσιάστηκε στην Ενότητα 8.3.1. Για την επίτευξη καλύτερων αποτελεσμάτων, πρέπει να αναζητηθεί μια πιο γενικευμένη προσέγγιση για την εκτίμηση των παραμέτρων του μοντέλου.

□

Παράρτημα Α΄

Παράρτημα

Α΄.1 Προσεγγιστική επίλυση του μοντέλου 2-αστέρων για δίκτυα με μη κατευθυντικές ακμές

Έχοντας ως αφετηρία την χαμιλτονιανή της Εξίσωσης 8.13

$$H = \frac{J}{n-1}\sum_{i=1}^{n}(k_i)^2 + B\sum_{i=1}^{n}k_i \tag{Α΄.1}$$

η συνάρτηση κατακερματισμού Z γίνεται

$$Z = \sum_{G\in\mathcal{G}} \mathrm{e}^{H(G)} = \sum_{G\in\mathcal{G}} \exp\{\frac{J}{n-1}\sum_{i=1}^{n}(k_i)^2 + B\sum_{i=1}^{n}k_i\} \Rightarrow$$

$$Z = \sum_{G\in\mathcal{G}} \exp\{\frac{J}{n-1}\sum_{i=1}^{n}(k_i)^2\} \times \exp\{B\sum_{i=1}^{n}k_i\} \tag{Α΄.2}$$

Σύμφωνα με τους [Park and Newman, 2004a], αθροίσματα σαν αυτό της Εξίσωσης Α΄.2 (που περιέχουν, δηλαδή, όρους της μορφής e^{k^2}) απατώνται στη μελέτη αλληλεπιδρώντων κβαντικών συστημάτων και υπολογίζονται με τη χρήση του μετασχηματισμού Hubbard-Stratonovich [Hubbard, 1959]. Ο εν λόγω μετασχηματισμός χρησιμοποιείται για την μετατροπή μιας σωματιδιακής θεωρίας (στην προκειμένη περίπτωση του βαθμού k_i του κόμβου) στην αντίστοιχη θεωρία πεδίου, μέσω της εισαγωγής ενός βοηθητικού βαθμωτού πεδίου (στην προκειμένη περίπτωση του πεδίου ϕ_i, όπως θα φανεί παρακάτω). Ο μετασχηματισμός Hubbard-Stratonovich βασίζεται στα γκαουσιανά ολοκληρώματα και με αυτό τον τρόπο θα χρησιμοποιηθεί στην παρούσα εργασία.

Το γκαουσιανό ολοκλήρωμα έχεις τις παρακάτω μορφές

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\alpha x^2}\mathrm{d}x = \sqrt{\frac{\pi}{\alpha}} \tag{Α΄.3}$$

$$\int_{-\infty}^{\infty} \mathrm{e}^{-ax^2+bx+c}\,dx = \sqrt{\frac{\pi}{a}}\,\mathrm{e}^{\frac{b^2}{4a}+c} \tag{Α΄.4}$$

με την πραγματική σταθερά α να πρέπει να είναι μη αρνητική ($\alpha > 0$).

Αντικαθιστώντας τη σταθερά α και τον άγνωστο x σύμφωνα με

$$\alpha \leftarrow (n-1)J, \quad x \leftarrow \phi_i - \frac{k_i}{n-1}$$

το γκαουσιανό ολοκλήρωμα της Εξίσωσης Α΄.3 γίνεται

$$\int_{-\infty}^{\infty}\exp\{-(n-1)J(\phi_i-\frac{k_i}{n-1})^2\}\mathrm{d}(\phi_i-\frac{k_i}{n-1})=\sqrt{\frac{\pi}{(n-1)J}}\Rightarrow$$
$$\int_{-\infty}^{\infty}\exp\{-(n-1)J\phi_i^2+2Jk_i\phi_i-\frac{Jk_i^2}{n-1}\}\mathrm{d}(\phi_i-\frac{k_i}{n-1})=\sqrt{\frac{\pi}{(n-1)J}}\Rightarrow$$
$$\mathrm{e}^{-\frac{Jk_i^2}{n-1}}\times\int_{-\infty}^{\infty}\exp\{-(n-1)J\phi_i^2+2Jk_i\phi_i\}\mathrm{d}\phi_i-$$
$$-\frac{1}{n-1}\mathrm{e}^{-(n-1)J\phi_i^2}\times\int_{-\infty}^{\infty}\exp\{2Jk_i\phi_i-\frac{Jk_i^2}{n-1}\}\mathrm{d}k_i=\sqrt{\frac{\pi}{(n-1)J}}\qquad(\text{Α΄.5})$$

Για την απαλοιφή του ολοκληρώματος ως προς k_i της Εξίσωσης Α΄.5, χρησιμοποιείται ο δεύτερος τύπος του γκαουσιανού ολοκληρώματος (Εξίσωση Α΄.4). Αντικαθιστώντας τις σταθερές α, b, c και τον άγνωστο x σύμφωνα με

$$a\leftarrow\frac{J}{n-1},\quad b\leftarrow 2J\phi_i,\quad c\leftarrow 0,\quad x\leftarrow k_i$$

έχουμε

$$\int_{-\infty}^{\infty}\exp\{2Jk_i\phi_i-\frac{Jk_i^2}{n-1}\}\mathrm{d}k_i=\sqrt{\frac{(n-1)\pi}{J}}\mathrm{e}^{(n-1)J\phi_i^2}\qquad(\text{Α΄.6})$$

Ο δεύτερος όρος του αριστερού μέρους της Εξίσωσης Α΄.5 σε συνδυασμό με την Εξίσωση Α΄.6, γίνεται

$$-\frac{1}{n-1}\mathrm{e}^{-(n-1)J\phi_i^2}\sqrt{\frac{(n-1)\pi}{J}}\mathrm{e}^{(n-1)J\phi_i^2}=-\sqrt{\frac{\pi}{(n-1)J}}\qquad(\text{Α΄.7})$$

και η Εξίσωση Α΄.5 σε συνδυασμό με την Εξίσωση Α΄.7 γίνεται

$$\mathrm{e}^{-\frac{Jk_i^2}{n-1}}\times\int_{-\infty}^{\infty}\exp\{-(n-1)J\phi_i^2+2Jk_i\phi_i\}\mathrm{d}\phi_i-\sqrt{\frac{\pi}{(n-1)J}}=\sqrt{\frac{\pi}{(n-1)J}}\Rightarrow$$
$$\mathrm{e}^{-\frac{Jk_i^2}{n-1}}\times\int_{-\infty}^{\infty}\exp\{-(n-1)J\phi_i^2+2Jk_i\phi_i\}\mathrm{d}\phi_i=2\sqrt{\frac{\pi}{(n-1)J}}\Rightarrow$$
$$\mathrm{e}^{\frac{J}{n-1}k_i^2}=\frac{1}{2}\sqrt{\frac{(n-1)J}{\pi}}\times\int_{-\infty}^{\infty}\exp\{-(n-1)J\phi_i^2+2Jk_i\phi_i\}\mathrm{d}\phi_i\qquad(\text{Α΄.8})$$

Πολλαπλασιάζοντας κατά μέρη την Εξίσωση Α΄.8 για τους n κόμβους του γράφου και εκμεταλλευόμενοι την ιδιότητα της διαδοχικής ολοκλήρωσης για συναρτήσεις πολλών μεταβλητών, έχουμε

$$\exp\{\frac{J}{n-1}\sum_{i=1}^{n}(k_i)^2\}=\left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}$$
$$\times\prod_{i=1}^{n}\int_{-\infty}^{\infty}\exp\{-(n-1)J\phi_i^2+2Jk_i\phi_i\}\mathrm{d}\phi_i\Rightarrow$$
$$\exp\{\frac{J}{n-1}\sum_{i=1}^{n}(k_i)^2\}=\left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}$$
$$\times\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2$$
$$+2J\sum_{i=1}^{n}k_i\phi_i\}\mathrm{d}\phi_1\ldots\mathrm{d}\phi_n\qquad(\text{Α΄.9})$$

Για την προσέγγιση του ολοκληρώματος του δεξιού μέρους της Εξίσωσης Α΄.9, επανερχόμαστε στην χβαντική μηχανική και πιο συγκεκριμένα στην έννοια των *ολοκληρωμάτων διαδρομής* (path integrals). Θεωρούμε ότι κάθε ένα από τα ϕ_i συμβολίζει την συνεισφορά του αντίστοιχου πεδίου κατά την κίνηση ενός σωματιδίου σε ένα μονοδιάστατο σύστημα, οπότε η επίδραση του συνολικού πεδίου στην κίνηση του σωματιδίου προσεγγίζεται από την υπέρθεση των επιμέρους πεδίων ϕ_i

$$\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2+2J\sum_{i=1}^{n}k_i\phi_i\}\mathrm{d}\phi_1\dots\mathrm{d}\phi_n=$$
$$\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2+2J\sum_{i=1}^{n}k_i\phi_i\}\mathscr{D}\phi \qquad (A'.10)$$

οπότε τελικά

$$\exp\{\frac{J}{n-1}\sum_{i=1}^{n}(k_i)^2\}=\left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2$$
$$+2J\sum_{i=1}^{n}k_i\phi_i\}\mathscr{D}\phi \qquad (A'.11)$$

Αντικαθιστώντας την Εξίσωση Α΄.11 στην Εξίσωση Α΄.2, η συνάρτηση κατακερματισμού γίνεται

$$Z=\left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}\sum_{G\in\mathcal{G}}\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2+2J\sum_{i=1}^{n}k_i\phi_i\}\mathscr{D}\phi\times$$
$$\times\exp\{B\sum_{i=1}^{n}k_i\}\Rightarrow$$
$$Z=\left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}\sum_{G\in\mathcal{G}}\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2+$$
$$+\sum_{i=1}^{n}(2J\phi_i+B)k_i\}\mathscr{D}\phi\Rightarrow$$
$$Z=\left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}\int_{-\infty}^{\infty}\left[\sum_{G\in\mathcal{G}}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2+\right.$$
$$\left.+\sum_{i=1}^{n}(2J\phi_i+B)k_i\}\right]\mathscr{D}\phi \qquad (A'.12)$$

όπου στην Εξίσωση Α΄.12 εναλλάξαμε την σειρά ολοκλήρωσης και του αθροίσματος όλων των πιθανών γράφων του μοντέλου. Επίσης παρατηρούμε ότι ο όρος που περιέχει το άθροισμα των τετραγώνων των πεδίων ϕ_i είναι ανεξάρτητος από τις πιθανές διαμορφώσεις των γράφων του μοντέλου, οπότε τελικά η Εξίσωση Α΄.12 γίνεται

$$Z=\left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2\}\sum_{G\in\mathcal{G}}\exp\{\sum_{i=1}^{n}(2J\phi_i+$$
$$+B)k_i\}\mathscr{D}\phi \qquad (A'.13)$$

Σε αυτό το σημείο θα υπολογιστεί το άθροισμα των όλων των πιθανών διαμορφώσεων του μοντέλου, οπότε ο δεύτερος όρος του γινομένου εντός του ολοκληρώματος της Εξίσωσης

Α΄.13 γίνεται

$$\begin{aligned}\sum_{i=1}^{n}(2J\phi_i+B)k_i &= \sum_{i=1}^{n}(2J\phi_i+B)\sum_{j=1}^{n}a_{ij}\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}(2J\phi_i+B)a_{ij}\\ &= \sum_{i=1}^{n}\sum_{j=i+1}^{n}[(2J\phi_i+B)a_{ij}+(2J\phi_j+B)a_{ji}]\\ &= \sum_{i=1}^{n}\sum_{j=i+1}^{n}(2J(\phi_i+\phi_j)+2B)a_{ij}\end{aligned}$$

και

$$\begin{aligned}\sum_{G\in\mathcal{G}}\exp\{\sum_{i=1}^{n}(2J\phi_i+B)k_i\} &= \prod_{i=1}^{n}\prod_{j=i+1}^{n}\sum_{a_{ij}=0}^{1}\exp\{(2J(\phi_i+\phi_j)+2B)a_{ij}\}\\ &= \prod_{i=1}^{n}\prod_{j=i+1}^{n}\left(1+\mathrm{e}^{2J(\phi_i+\phi_j)+2B}\right)\\ &= \exp\{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\ln\left(1+\mathrm{e}^{2J(\phi_i+\phi_j)+2B}\right)\}\end{aligned}\tag{Α΄.14}$$

Αντικαθιστώντας την Εξίσωση Α΄.14 στην Εξίσωση Α΄.13 έχουμε,

$$\begin{aligned}Z &= \left[\frac{(n-1)J}{4\pi}\right]^{\frac{n}{2}}\int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2+\\ &\quad+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\ln\left(1+\mathrm{e}^{2J(\phi_i+\phi_j)+2B}\right)\}\mathscr{D}\phi\Rightarrow\\ Z &= \int_{-\infty}^{\infty}\exp\{-(n-1)J\sum_{i=1}^{n}(\phi_i)^2+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\ln\left(1+\mathrm{e}^{2J(\phi_i+\phi_j)+2B}\right)+\\ &\quad+\frac{n}{2}\ln[(n-1)J]-\frac{n}{2}\ln 4\pi\}\mathscr{D}\phi\Rightarrow\\ Z &= \int_{-\infty}^{\infty}\mathrm{e}^{\mathscr{H}(\phi)}\mathscr{D}\phi\end{aligned}\tag{Α΄.15}$$

όπου $\mathscr{H}(\phi)$ η *ενεργός χαμιλτονιανή* (effective hamiltonian)

$$\begin{aligned}\mathscr{H}(\phi) &= -(n-1)J\sum_{i=1}^{n}(\phi_i)^2+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\ln\left(1+\mathrm{e}^{2J(\phi_i+\phi_j)+2B}\right)-\\ &\quad+\frac{n}{2}\ln(n-1)J-\frac{n}{2}\ln 4\pi\Rightarrow\\ \mathscr{H}(\phi) &= -(n-1)J\sum_{i=1}^{n}(\phi_i)^2+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1(\neq i)}^{n}\ln\left(1+\mathrm{e}^{2J(\phi_i+\phi_j)+2B}\right)-\\ &\quad+\frac{n}{2}\ln[(n-1)J]-\frac{n}{2}\ln 4\pi\end{aligned}\tag{Α΄.16}$$

Με την Εξίσωση Α΄.16 έγινε κατορθωτό να μετατραπεί το αρχικό μοντέλο σε μια θεωρία πεδίου συνεχούς μεταβλητής σε n σημεία. Δυστυχώς, όμως, το ολοκλήρωμα της παραπάνω Εξίσωσης δεν μπορεί να υπολογιστεί σε κλειστή μορφή [Park and Newman, 2004a], αλλά μόνο μέσω προσεγγιστικών μεθόδων. Στο πλαίσιο της παρούσας εργασίας θα εξετάσουμε στην επόμενη υποενότητα μια τέτοια προσεγγιστική μέθοδο

Α΄.1.1 Θεωρία Μέσου Πεδίου

Η απλούστερη προσέγγιση που μπορεί να γίνει είναι αυτή του μέσου πεδίου, κατά την οποία αγνοούνται οι διακυμάνσεις στο πεδίο και το ϕ_i θεωρείται ότι λαμβάνει πάντα την πιο πιθανή τιμή του, η οποία βρίσκεται στο *σαγματικό σημείο* (saddle point) όπου μηδενίζεται η πρώτη παράγωγος

$$\frac{\partial \mathscr{H}(\phi)}{\partial \phi_i} = 0 \Rightarrow -2(n-1)J\phi_i + J \sum_{j=1(\neq j)}^{n} \frac{\mathrm{e}^{2J(\phi_i+\phi_j)+2B}}{1+\mathrm{e}^{2J(\phi_i+\phi_j)+2B}} = 0$$

Κάνοντας χρήση της ταυτότητας

$$\frac{\mathrm{e}^x}{1+\mathrm{e}^x} = \frac{1}{2}\left[\tanh\frac{x}{2}+1\right] \tag{Α΄.17}$$

παίρνουμε την Εξίσωση

$$-2(n-1)\phi_i + \sum_{j=1(\neq i)}^{n} [\tanh(J(\phi_i+\phi_j)+B)+1] = 0$$

η οποία έχει τη συμμετρική λύση $\phi_0 = \phi_i$ για κάθε i

$$\begin{aligned} -2(n-1)\phi_0 + \sum_{j=1(\neq i)}^{n} [\tanh(2J\phi_0+B)+1] &= 0 \Rightarrow \\ -2(n-1)\phi_0 + (n-1)\left[\tanh(2J\phi_0+B)+1\right] &= 0 \Rightarrow \\ \phi_0 = \frac{1}{2}\left[\tanh(2J\phi_0+B)+1\right] & \end{aligned} \tag{Α΄.18}$$

Στην περίπτωση αυτή, το ολοκλήρωμα διαδρομής της Εξίσωσης Α΄.15 απλοποιείται σε n ανεξάρτητα γκαουσιανά ολοκληρώματα, οπότε η συνάρτηση κατάτμησης και η συνάρτηση ελεύθερης ενέργειας γίνονται

$$\begin{aligned} Z =& \int_{-\infty}^{\infty} \mathrm{e}^{\mathscr{H}(\phi)}\mathscr{D}\phi = \int_{-\infty}^{\infty} \mathrm{e}^{n\mathscr{H}(\phi_0)}\mathscr{D}\phi = \mathrm{e}^{n\mathscr{H}(\phi_0)}\int_{-\infty}^{\infty}\mathscr{D}\phi \Rightarrow \\ Z =& \mathrm{e}^{n\mathscr{H}(\phi_0)} \end{aligned} \tag{Α΄.19}$$

$$\begin{aligned} F \equiv& \ln Z = n\mathscr{H}(\phi_0) \Rightarrow \\ F =& -n(n-1)J(\phi_0)^2 + \frac{1}{2}n(n-1)\ln\left(1+\mathrm{e}^{4J\phi_0+2B}\right) + \frac{n}{2}\ln[(n-1)J] \\ & -\frac{n}{2}\ln 4\pi \end{aligned} \tag{Α΄.20}$$

□

Βιβλιογραφία

The netflix prize. URL http://www.netflixprize.com.

Camra2010 - cars-2010: Challenge on context-aware movie recommendation, September 2010. URL http://www.dai-labor.de/camra2010/.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, June 2005. ISSN 1041-4347. doi: 10.1109/TKDE.2005.99. URL http://dx.doi.org/10.1109/TKDE.2005.99.

G. Adomavicius, N. Manouselis, and Y. Kwon. Multi-criteria recommender systems. In *Recommender Systems Handbook*, pages 769–803. 2011. URL http://dx.doi.org/10.1007/978-0-387-85820-3_24.

X. Amatriain, A. Jaimes*, N. Oliver, and J. M. Pujol. Data mining methods for recommender systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 39–71. Springer US, 2011. ISBN 978-0-387-85820-3. URL http://dx.doi.org/10.1007/978-0-387-85820-3_2.

R. Andersen, C. Borgs, J. Chayes, U. Feige, A. Flaxman, A. Kalai, V. Mirrokni, and M. Tennenholtz. Trust-based recommendation systems: an axiomatic approach. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 199–208, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367525. URL http://doi.acm.org/10.1145/1367497.1367525.

C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43, 2003. ISSN 0885-6125. doi: 10.1023/A:1020281327116. URL http://dx.doi.org/10.1023/A\%3A1020281327116.

D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *J. Web Sem.*, 5(2):58–71, 2007. URL http://dx.doi.org/10.1016/j.websem.2007.03.002.

A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL http://www.sciencemag.org/content/286/5439/509.abstract.

N. K. Baym and A. Ledbetter. Tunes that bind? *Information, Communication & Society*, 12(3):408–427, 2009. doi: 10.1080/13691180802635430. URL http://www.tandfonline.com/doi/abs/10.1080/13691180802635430.

G. Beliakov, T. Calvo, and S. James. Aggregation of preferences in recommender systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 705–734. Springer US, 2011. ISBN 978-0-387-85820-3. URL http://dx.doi.org/10.1007/978-0-387-85820-3_22.

A. Bellogin. Performance prediction in recommender systems: application to the dynamic optimisation of aggregative methods. Master's thesis, 2009.

J. Bertini, Joao Roberto and M. do Carmo Nicoletti. Mbabconn - a multiclass version of a constructive neural network algorithm based on linear separability and convex hull. In V. Kurkova, R. Neruda, and J. Koutnik, editors, *Artificial Neural Networks - ICANN 2008*, volume 5164 of *Lecture Notes in Computer Science*, pages 723–733. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-87558-1. doi: 10.1007/978-3-540-87559-8_75. URL http://dx.doi.org/10.1007/978-3-540-87559-8_75.

D. Billsus and M. J. Pazzani. Learning collaborative information filters. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 46–54, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8. URL http://dl.acm.org/citation.cfm?id=645527.657311.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierrez. Recommender systems survey. *Knowledge-Based Systems*, 46(0):109 – 132, 2013. ISSN 0950-7051. doi: http://dx.doi.org/10.1016/j.knosys.2013.03.012. URL http://www.sciencedirect.com/science/article/pii/S0950705113001044.

D. Boyd and J. Heer. Profiles as conversation: Networked identity performance on friendster. In *System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 3, page 59c, jan. 2006. doi: 10.1109/HICSS.2006.394.

I. Cantador, A. Bellogín, and P. Castells. A multilayer ontology-based hybrid recommendation model. *AI Commun.*, 21(2-3):203–210, Apr. 2008. ISSN 0921-7126. URL http://dl.acm.org/citation.cfm?id=1460172.1460184.

I. Cantador, P. Brusilovsky, and T. Kuflik. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

V. R. Carvalho and W. W. Cohen. Recommending recipients in the enron email corpus. Technical report, limbo, 2007. URL http://www.cs.cmu.edu/~wcohen/postscript/cc-predict-submitted.pdf.

P. Castells, S. Vargas, and J. Wang. Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *International Workshop on Diversity in Document Retrieval (DDR 2011) at the 33rd European Conference on Information Retrieval (ECIR 2011)*, Apr. 2011.

C. Christakou and A. Stafylopatis. A hybrid movie recommender system based on neural networks. In *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, ISDA '05, pages 500–505, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2286-06. doi: 10.1109/ISDA.2005.9. URL http://dx.doi.org/10.1109/ISDA.2005.9.

M. L. Clemente. Experimental results on item-based algorithms for independent domain collaborative filtering. In *Proceedings of the 2008 International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution*, AXMEDIS '08, pages 87–92, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3406-0. doi: 10.1109/AXMEDIS.2008.33. URL http://dx.doi.org/10.1109/AXMEDIS.2008.33.

P. De Meo, G. Quattrone, G. Terracina, and D. Ursino. An xml-based multiagent system for supporting online recruitment services. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 37(4):464 –480, july 2007. ISSN 1083-4427. doi: 10.1109/TSMCA.2007.897696.

S. Debnath, N. Ganguly, and P. Mitra. Feature weighting in content based recommendation system using social network analysis. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1041–1042, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367646. URL `http://doi.acm.org/10.1145/1367497.1367646`.

C. Desrosiers and G. Karypis. A novel approach to compute similarities and its application to item recommendation. In *Proceedings of the 11th Pacific Rim international conference on Trends in artificial intelligence*, PRICAI'10, pages 39–51, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15245-7, 978-3-642-15245-0. URL `http://dl.acm.org/citation.cfm?id=1884293.1884302`.

T. DuBois, J. Golbeck, J. Kleint, and A. Srinivasan. Improving recommendation accuracy by clustering social networks with trust. In *Recommender Systems & the Social Web*, volume 532, pages 1–8, 2009.

W. Duch. K-separability. In S. Kollias, A. Stafylopatis, W. Duch, and E. Oja, editors, *Artificial Neural Networks - ICANN 2006*, volume 4131 of *Lecture Notes in Computer Science*, pages 188–197. Springer Berlin / Heidelberg, 2006. ISBN 978-3-540-38625-4. URL `http://dx.doi.org/10.1007/11840817_20`. 10.1007/11840817_20.

T. Dunning. Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS*, 19(1):61–74, 1993.

F. Echarte, J. J. Astrain, A. Córdoba, and J. E. Villadangos. Ontology of folksonomy: A new modelling method. In *SAAKM*, 2007. URL `http://ceur-ws.org/Vol-289/p08.pdf`.

I. Fodor. A survey of dimension reduction techniques. Technical report, 2002.

S. Fortunato. Community detection in graphs. *Physics and Society*, 486:75–174, February 2010. doi: 10.1016/j.physrep.2009.11.002. URL `http://arxiv.org/abs/0906.0612v2`.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm, 1996.

J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, 23(9):881–890, Sept. 1974. ISSN 0018-9340. doi: 10.1109/T-C.1974.224051. URL `http://dx.doi.org/10.1109/T-C.1974.224051`.

S. I. Gallant. Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks*, 2:179–192, June 1990.

M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 257–260, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-906-0. doi: 10.1145/1864708.1864761. URL `http://doi.acm.org/10.1145/1864708.1864761`.

M. A. Ghazanfar and A. Prugel-Bennett. A scalable, accurate hybrid recommender system. In *Proceedings of the 2010 Third International Conference on Knowledge Discovery and Data Mining*, WKDD '10, pages 94–98, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-3923-2. doi: 10.1109/WKDD.2010.117. URL http://dx.doi.org/10.1109/WKDD.2010.117.

D. Godoy and A. Amandi. User profiling in personal information agents: a survey. *Knowl. Eng. Rev.*, 20(4):329–361, Dec. 2005. ISSN 0269-8889. doi: 10.1017/S0269888906000397. URL http://dx.doi.org/10.1017/S0269888906000397.

J. Golbeck and J. Hendler. Filmtrust: movie recommendations using trust in web-based social networks. In *Consumer Communications and Networking Conference, 2006. CCNC 2006. 3rd IEEE*, volume 1, pages 282 – 286, jan. 2006. doi: 10.1109/CCNC.2006.1593032.

J. A. Golbeck. *Computing and applying trust in web-based social networks*. PhD thesis, College Park, MD, USA, 2005. AAI3178583.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr.*, 4(2):133–151, July 2001. ISSN 1386-4564. doi: 10.1023/A:1011419012209. URL http://dx.doi.org/10.1023/A:1011419012209.

H. Grahsl, C. Korner, and M. Strohmaier. A collection of tagging datasets containing complete personomies from heterogeneous sources. Technical report, Knowledge Management Institute, Graz University of Technology, 2010.

S. Grant and G. I. McCalla. A hybrid approach to making recommendations and its application to the movie domain. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, AI '01, pages 257–266, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-42144-0. URL http://dl.acm.org/citation.cfm?id=647462.760370.

M. Grochowski and W. Duch. Constructive neural network algorithms that solve highly non-separable problems. In L. Franco, D. A. Elizondo, and J. M. Jerez, editors, *Constructive Neural Networks*, volume 258 of *Studies in Computational Intelligence*, pages 49–70. Springer, 2009. ISBN 978-3-642-04511-0.

GroupLens and U. o. Minnesota. The movielens dataset. URL http://grouplens.org/datasets/movielens/.

M. Gupta, R. Li, Z. Yin, and J. Han. *An Overview of Social Tagging and Applications*, page 447. 2011. doi: 10.1007/978-1-4419-8462-3_16. URL http://adsabs.harvard.edu/abs/2011snda.book..447G.

C. Hayes, P. Avesani, and U. Bojars. From web to social web: Discovering and deploying user and content profiles. chapter An Analysis of Bloggers, Topics and Tags for a Blog Recommender System, pages 1–20. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 978-3-540-74950-9. doi: 10.1007/978-3-540-74951-6_1. URL http://dx.doi.org/10.1007/978-3-540-74951-6_1.

S. Haykin. *Neural Networks and Learning Machines*. Prentice Hall, third edition edition, 2008.

J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, Jan. 2004. ISSN 1046-8188. doi: 10.1145/963770.963772. URL http://doi.acm.org/10.1145/963770.963772.

B. Hogan. *Analysing Social Networks Via the Internet*, chapter 8. Sage, 2008.

W. H. Hsu, A. L. King, M. S. R. Paradesi, T. Pydimarri, and T. Weninger. Collaborative and structural recommendation of friends using weblog-based social network analysis. *Computational Approaches to Analyzing Weblogs - Papers from the 2006 Spring Symposium*, pages 24–31, 2006. AAAI Press Technical Report SS-06-03. Stanford, USA, March 2006.

J. Hubbard. Calculation of partition functions. *Phys. Rev. Lett.*, 3:77–78, Jul 1959. doi: 10.1103/PhysRevLett.3.77. URL `http://link.aps.org/doi/10.1103/PhysRevLett.3.77`.

C.-S. Hwang and R.-S. Fong. A hybrid recommender system based on collaborative filtering and cloud model, 2011.

M. M. Islam, M. A. Sattar, M. F. Amin, X. Yao, and K. Murase. A new constructive algorithm for architectural and functional adaptation of artificial neural networks. *Trans. Sys. Man Cyber. Part B*, 39(6):1590–1605, Dec. 2009. ISSN 1083-4419. doi: 10.1109/TSMCB.2009.2021849. URL `http://dx.doi.org/10.1109/TSMCB.2009.2021849`.

M. Jamali. The flixster dataset, 2010. URL `http://www.cs.sfu.ca/~sja25/personal/datasets/`.

M. Jamali and M. Ester. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 397–406, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557067. URL `http://doi.acm.org/10.1145/1557019.1557067`.

A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68, 2010. ISSN 0007-6813. doi: 10.1016/j.bushor.2009.09.003. URL `http://www.sciencedirect.com/science/article/pii/S0007681309001232`.

A. Kleeman, S. Denuit, and N. Hendersen. Matrix factorization for collaborative prediction.

L. Knockaert, B. De Backer, and D. De Zutter. Svd compression, unitary transforms, and computational complexity. *Trans. Sig. Proc.*, 47(10):2724–2729, Oct. 1999. ISSN 1053-587X. doi: 10.1109/78.790654. URL `http://dx.doi.org/10.1109/78.790654`.

I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 195–202, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1571977. URL `http://doi.acm.org/10.1145/1571941.1571977`.

Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 426–434, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401944. URL `http://doi.acm.org/10.1145/1401890.1401944`.

Y. Koren. The bellkor solution to the netflix prize, August 2009.

D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999. URL http://www.nature.com/nature/journal/v401/n6755/abs/401788a0.html.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

E. Lee, C. P. Lim, R. Yuen, and S. Lo. A hybrid neural network model for noisy data regression. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(2):951 – 960, april 2004. ISSN 1083-4419. doi: 10.1109/TSMCB.2003.818440.

Q. Liu, B. Xiang, E. Chen, Y. Ge, H. Xiong, T. Bao, and Y. Zheng. Influential seed items recommendation. In *Proceedings of the sixth ACM conference on Recommender systems*, RecSys '12, pages 245–248, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1270-7. doi: 10.1145/2365952.2366005. URL http://doi.acm.org/10.1145/2365952.2366005.

L. Lovasz. Random walks on graphs: A survey, 1993.

L. Mackey, D. Weiss, and M. I. Jordan. Mixed membership matrix factorization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 711–718. OmniPress, 2010.

B. Marlin. Collaborative filtering: A machine learning perspective. Technical report, 2004.

S. P. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, Apr 1994.

P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. In R. Meersman and Z. Tari, editors, *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, volume 3290 of *Lecture Notes in Computer Science*, pages 492–508. Springer Berlin / Heidelberg, 2004. URL http://dx.doi.org/10.1007/978-3-540-30468-5_31. 10.1007/978-3-540-30468-5_31.

P. Massa and P. Avesani. Trust-aware bootstrapping of recommender systems. In *Proceedings of ECAI 2006 workshop on recommender systems*, volume 28, pages 29–33, 2006.

P. Massa and P. Avesani. Trust metrics on controversial users: balancing between tyranny of the majority and echo chambers. *International Journal on Semantic Web and Information Systems*, 2007. URL http://www.gnuband.org/papers/trust_metrics_on_controversial_users_balancing_between_tyranny_of_the_majority_and_echo_chambers-2/.

P. Massa and P. Avesani. Trust metrics in recommender systems. In J. Golbeck, editor, *Computing with Social Trust*, Human-Computer Interaction Series, pages 259–285. Springer London, 2009. ISBN 978-1-84800-356-9.

P. Massa and B. Bhattacharjee. Using trust in recommender systems: An experimental analysis. In *In Proceedings of iTrust2004 International Conference*, pages 221–235, 2004.

M. R. McLaughlin and J. L. Herlocker. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 329–336, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. doi: 10.1145/1008992.1009050. URL http://doi.acm.org/10.1145/1008992.1009050.

S. E. Middleton, N. Shadbolt, and D. De Roure. Ontology-based recommender systems. 2003. URL http://www.scientificcommons.org/2326218.

L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 2431 – 2439, jan. 2002. doi: 10.1109/HICSS.2002.994181.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. URL http://ilpubs.stanford.edu:8090/422/. Previous number = SIDL-WP-1999-0120.

J. Park and M. Newman. Statistical mechanics of networks. *Physical Review E*, 70(6): 066117, Dec 2004a. doi: 10.1103/PhysRevE.70.066117.

J. Park and M. Newman. Solution of the two-star model of a network. *Phys. Rev. E*, 70: 066146, Dec 2004b. doi: 10.1103/PhysRevE.70.066146. URL http://link.aps.org/doi/10.1103/PhysRevE.70.066146.

J. Park and M. E. J. Newman. Solution for the properties of a clustered network. *Phys. Rev. E*, 72(2):026136, Aug. 2005. doi: 10.1103/PhysRevE.72.026136.

C. C. Patricia Victor, Martine De Cock. *Trust and Recommendations*, chapter 20, pages 645–675. Springer, 2011.

S. Perugini, M. A. Goncalves, and E. A. Fox. Recommender systems research: A connection-centric survey. *Journal of Intelligent Information Systems*, 23:107–143, 2004. ISSN 0925-9902. URL http://dx.doi.org/10.1023/B:JIIS.0000039532.05533.99. 10.1023/B:JIIS.0000039532.05533.99.

M. C. Pham, Y. Cao, R. Klamma, and M. Jarke. A clustering approach for collaborative filtering recommendation using social network analysis. *Journal of Universal Computer Science*, 17(4):583–604, feb 2011. doi: 10.3217/jucs-017-04-0583. URL http://www.jucs.org/jucs_17_4/a_clustering_approach_for.

B. Piccart, J. Struyf, and H. Blockeel. Alleviating the sparsity problem in collaborative filtering by using an adapted distance and a graph-based method. In *SDM*, pages 189–198, 2010.

G. Pitsilis and S. J. Knapskog. Social trust as a solution to address sparsity-inherent problems of recommender systems. In *ACM RecSYS '09 Workshop on Recommender Systems*, 2009.

G. Pitsilis, X. Zhang, and W. Wang. *5th IFIP WG 11.11 International Conference, IFIPTM 2011, Copenhagen, Denmark, June 29 - July 1, 2011. Proceedings*, chapter Clustering Recommenders in Collaborative Filtering Using Explicit Trust Information, pages 82–97. 358. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-22200-9_9.

L. Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Netw.*, 11(4):761–767, June 1998. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00010-0. URL http://dx.doi.org/10.1016/S0893-6080(98)00010-0.

I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon. Overlapping community detection using bayesian non-negative matrix factorization. *Phys. Rev. E*, 83:066114, Jun 2011. doi: 10.1103/PhysRevE.83.066114. URL http://link.aps.org/doi/10.1103/PhysRevE.83.066114.

P. Pu, B. Faltings, L. Chen, J. Zhang, and P. Viappiani. Usability guidelines for product recommenders based on example critiquing research. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 511–545. Springer US, 2011. ISBN 978-0-387-85820-3. URL http://dx.doi.org/10.1007/978-0-387-85820-3_16.

A. Pucci, M. Gori, F. Scarselli, M. Hagenbuchner, and A. C. Tsoi. Applications of graph neural networks to large-scale recommender systems, some results. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, volume 1, pages 189 – 195, November 2006.

W. X. Qian Xu and Q. Yang. Social-behavior transfer learning for recommendation systems. In *Proceedings of the International Workshop on Social Web Mining, the 22nd International Joint Conference on Artificial Intelligence*, pages 18–26, July 2011.

A. Quan-Haase and B. Wellman. Hyperconnected net work: Computer mediated community in a high-tech organization. *The firm as a collaborative community: Reconstructing trust in the knowledge economy*, pages 281–333, 2006.

H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Wiley Classics Library Edition, 2000.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM. ISBN 0-89791-689-1. doi: 10.1145/192844.192905. URL http://doi.acm.org/10.1145/192844.192905.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173 – 191, 2007. ISSN 0378-8733. doi: http://dx.doi.org/10.1016/j.socnet.2006.08.002. URL http://www.sciencedirect.com/science/article/pii/S0378873306000372. Special Section: Advances in Exponential Random Graph (p*) Models.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, Aug. 1988. ISSN 0306-4573. doi: 10.1016/0306-4573(88)90021-0. URL http://dx.doi.org/10.1016/0306-4573(88)90021-0.

B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl. Application of dimensionality reduction in recommender system – a case study. In *IN ACM WEBKDD WORKSHOP*, 2000.

B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering, 2002.

J. C. Schlimmer and R. H. Granger, Jr. Incremental learning from noisy data. *Mach. Learn.*, 1(3):317–354, Mar. 1986. ISSN 0885-6125. doi: 10.1023/A:1022810614389. URL http://dx.doi.org/10.1023/A:1022810614389.

M. Schmidt, O. Winther, and L. Hansen. Bayesian non-negative matrix factorization. In T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, editors, *Independent Component Analysis and Signal Separation*, volume 5441 of *Lecture Notes in Computer Science*, pages 540–547. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-00598-5. doi: 10.1007/978-3-642-00599-2_68. URL http://dx.doi.org/10.1007/978-3-642-00599-2_68.

J. P. Scott and P. J. Carrington. *The SAGE Handbook of Social Network Analysis*. Sage Publications Ltd., 2011. ISBN 1847873952, 9781847873958.

O. Selma. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010. URL http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/index.html.

G. Shani and A. Gunawardana. Evaluating recommendation systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 257–297. Springer US, 2011. ISBN 978-0-387-85820-3. URL http://dx.doi.org/10.1007/978-0-387-85820-3_8. 10.1007/978-0-387-85820-3_8.

A. P. Singh, A. Gunawardana, C. Meek, and A. C. Surendran. Recommendations using absorbing random walks. In *North East Student Colloquium on Artificial Intelligence (NESCAI)*, 2007.

P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 655–664, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367586. URL http://doi.acm.org/10.1145/1367497.1367586.

R. R. Sinha and K. Swearingen. Comparing recommendations made by online systems and friends. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001. URL http://www.ercim.org/publication/ws-proceedings/DelNoe02/RashmiSinha.pdf.

P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Providing justifications in recommender systems. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(6):1262 –1272, nov. 2008. ISSN 1083-4427. doi: 10.1109/TSMCA.2008.2003969.

D. Tian, Y. Liu, and D. Wei. A dynamic growing neural network for supervised or unsupervised learning. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, volume 1, pages 2886 –2890, 0-0 2006. doi: 10.1109/WCICA.2006.1712893.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

P. Victor, M. D. Cock, and C. Cornelis. Trust and recommendations. In *Recommender Systems Handbook*, pages 645–675. 2011. URL http://dx.doi.org/10.1007/978-0-387-85820-3_20.

M. G. Vozalis and K. G. Margaritis. A recommender system using principal component analysis. In *11th Panhellenic Conferance on Informatics (PCI 2007)*, pages 271–283, May 2007.

Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1336–1353, June 2013. ISSN 1041-4347. doi: 10.1109/TKDE.2012.51.

S. Wasserman and K. Faust. *Social network analysis: Methods and Applications*. Number 8 in Structural Analysis in the Social Sciences. Cambridge University Press, first edition, 1994. URL http://www.cambridge.org/gb/knowledge/isbn/item1138907/.

D. J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2003. URL http://press.princeton.edu/titles/6768.html.

H. D. White, B. Wellman, and N. Nazer. Does citation reflect social structure?: Longitudinal evidence from the "globenet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology*, 55(2):111–126, 2004. ISSN 1532-2890. doi: 10.1002/asi.10369. URL http://dx.doi.org/10.1002/asi.10369.

J. Yang and J. Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 587–596, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3. doi: 10.1145/2433396.2433471. URL http://doi.acm.org/10.1145/2433396.2433471.

M. Zanker and M. Jessenitschnig. Collaborative feature-combination recommender exploiting explicit and implicit user feedback. In *Commerce and Enterprise Computing, 2009. CEC '09. IEEE Conference on*, pages 49–56, July 2009. doi: 10.1109/CEC.2009.84.

T. Zhang and V. S. Iyengar. Recommender systems using linear classifiers. *J. Mach. Learn. Res.*, 2:313–334, Mar. 2002. ISSN 1532-4435. doi: 10.1162/153244302760200641. URL http://dx.doi.org/10.1162/153244302760200641.

Z.-Y. Zhang. Nonnegative matrix factorization: Models, algorithms and applications. In D. Holmes and L. Jain, editors, *Data Mining: Foundations and Intelligent Paradigms*, volume 24 of *Intelligent Systems Reference Library*, pages 99–134. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-23240-4. doi: 10.1007/978-3-642-23241-1_6. URL http://dx.doi.org/10.1007/978-3-642-23241-1_6.

Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management*, AAIM '08, pages 337–348, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-68865-5. doi: 10.1007/978-3-540-68880-8_32. URL http://dx.doi.org/10.1007/978-3-540-68880-8_32.

C.-N. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4-5):337–358, Dec. 2005. ISSN 1387-3326. doi: 10.1007/s10796-005-4807-3. URL http://dx.doi.org/10.1007/s10796-005-4807-3.

Κατάλογος δημοσιεύσεων του συγγραφέα

- Διεθνή περιοδικά με κρίση

  1. G. Alexandridis, G. Siolas and A. Stafylopatis : **Enhancing Social Collaborative Filtering through the application of Non-Negative Matrix Factorization and Exponential Random Graph Models** *(submitted)*
  2. G. Alexandridis, G. Siolas and A. Stafylopatis : **Applying k-separability to Collaborative Recommender Systems**, *International Journal on Artificial Intelligence Tools*, Vol. 21, 02, 2012.

- Κεφάλαια σε βιβλία

  3. G. Alexandridis, G. Siolas and A. Stafylopatis : **Accuracy Versus Novelty and Diversity in Recommender Systems: A Nonuniform Random Walk Approach**. *Recommendation and Search in Social Networks* Ulusoy, Ozgur, Tansel, A. Uz, Arkun and Erol eds. Springer International Publishing. 2015. pp. 41-57.

- Διεθνή συνέδρια με κρίση

  4. G. Alexandridis, G. Siolas and A. Stafylopatis : **Improving Social Recommendations by applying a Personalized Item Clustering Policy**. *Proceedings of the Fifth ACM RecSys Workshop on Recommender Systems and the Social Web co-located with the 7th ACM Conference on Recommender Systems (RecSys 2013).* 2013
  5. G. Alexandridis, G. Siolas and A. Stafylopatis : **A Biased Random Walk Recommender based on Rejection Sampling** *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* 2013. pp. 648-652.
  6. G. Alexandridis, G. Siolas and A. Stafylopatis : **An Efficient Collaborative Recommender System Based on k-Separability** *Proceedings of the 20th International Conference on Artificial Neural Networks (ICANN 2010)* Thessaloniki, Greece, September 15–18, 2010

- Εκτός διατριβής

  7. G.C. Alexandridis, A.G. Voyiatzis, D.N. Serpanos : **CryptoPalm: a cryptographic library for PalmOS** *PCI'05 Proceedings of the 10th Panhellenic conference on Advances in Informatics*, Volos, Greece, November 11–13, 2005

□

Βιογραφικό Σημείωμα

Στοιχεία Επικοινωνίας

Εργαστήριο Ευφυών Συστημάτων
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
Ηρώων Πολυτεχνείου 9, Ζωγράφου
157 80 Αθήνα, Ελλάδα
Τηλέφωνο: (+30) 210 772 2504
Ηλεκτρονικό ταχυδρομείο (e-mail): `gealexandri@islab.ntua.gr`
Προσωπική Σελίδα: `http://www.islab.ntua.gr/gealexandri/`

Σπουδές

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2007–σήμερα)
  Υποψήφιος Διδάκτωρ Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
  Επιβλέπων: καθ. Ανδρέας-Γεώργιος Σταφυλοπάτης

- **Πανεπιστήμιο Πατρών**, Ελλάδα (1999–2005)
  Διπλωματούχος Σχολής Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών
  Βαθμός: 7.36/10
  Διπλωματική εργασία: Υλοποίηση σε Λογισμικό Κρυπτογραφικών Αλγορίθμων Υψηλής Απόδοσης για Ενσωματωμένα Συστήματα
  Επιβλέπων: καθ. Δημήτριος Σερπάνος

Ερευνητικά Ενδιαφέροντα

- Συστήματα Συστάσεων
- Κοινωνικά Δίκτυα
- Δυναμική Παράλληλη Επεξεργασία μεγάλου Όγκου Δεδομένων
- Μη-σχεσιακές Βάσεις Δεδομένων (NoSQL)

Διδακτική - Εργασιακή Εμπειρία

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2007–σήμερα)
  Βοηθός Διδασκαλίας
  - Προγραμματιστικές Τεχνικές
  - Νευρωνικά Δίκτυα-Ευφυή Συστήματα
- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2007–σήμερα)
  System Administrator στο Εργαστήριο Ευφυών Συστημάτων
- **Υπουργείο Παιδείας, Δια Βίου Μάθησης και Θρησκευμάτων**, Ελλάδα (2006–σήμερα)
  Καθηγητής Πληροφορικής ΠΕ 19 σε σχολεία της Πρωτοβάθμιας και Δευτεροβάθμιας Εκπαίδευσης

Ξένες γλώσσες

Αγγλικά : Certificate of Proficiency in English, The University of Cambridge (1997)

□