ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΚΑΙ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ανάλυση Συστημάτων Προτάσεων και Εφαρμογή Αλγορίθμου knn στα Κοινωνικά Μέσα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΘΗΝΑΣ ΠΕΛΕΚΑΝΟΥ

Επιβλέπων : Δημήτριος Ασκούνης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2011

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΚΑΙ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ανάλυση Συστημάτων Προτάσεων και Εφαρμογή Αλγορίθμου knn στα Κοινωνικά Μέσα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΘΗΝΑΣ ΠΕΛΕΚΑΝΟΥ

Επιβλέπων : Δημήτριος Ασκούνης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την … Ιουλίου 2011.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ιωάννης Ψαρράς Καθηγητής Ε.Μ.Π	Δημήτριος Ασκούνης Αναπληρωτής Καθηγητής Ε.Μ.Π.	Γρηγόριος Μέντζας Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2011

(Υπογραφή)

.....

ΑΘΗΝΑ ΠΕΛΕΚΑΝΟΥ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στις μέρες μας, ο όγκος των πληροφοριών που μπορεί κανείς να βρει ή να δημοσιεύσει στο διαδίκτυο είναι τεράστιος. Οι χρήστες του διαδικτύου κατακλύζονται από προϊόντα και υπηρεσίες που καλύπτουν ένα μεγάλο εύρος τομέων. Το γεγονός αυτό, σε συνδυασμό με το ότι υπολείπονται σχετικής γνώσης και εμπειρίας κάνει πολύ δύσκολη την λήψη αποφάσεων σχετικά με το ποιά από τα διαθέσιμα προϊόντα ανταποκρίνονται στις ανάγκες και τις προτιμήσεις τους. Το πρόβλημα αυτό έρχονται να λύσουν τα Συστήματα Προτάσεων, τα οποία βρίσκουν στον χρήστη προϊόντα που τον ενδιαφέρουν και τον βοηθούν να καταλήξει σε αποφάσεις.

Σκοπός της διπλωματικής εργασίας είναι η μελέτη ενός σπουδαίου λογισμικού εργαλείου, των Συστημάτων Προτάσεων, σε συσχετισμό με έναν πολύ σύγχρονο τομέα του διαδικτυακού τόπου, τα Κοινωνικά Μέσα. Γίνεται προσπάθεια να κατανοηθούν το υπόβαθρο, ο τρόπος λειτουργίας και η χρησιμότητα των Συστημάτων Προτάσεων, καθώς και να εξεταστούν κάποιες εφαρμογές τους στα Κοινωνικά Μέσα.

Αρχικά αναλύονται σε θεωρητικό επίπεδο τα είδη των Recommender Systems, οι αλγόριθμοι που χρησιμοποιούν και τα πεδία στα οποία εφαρμόζονται, δίνοντας βάρος κυρίως στα Συστήματα που μπορούν να εφαρμοστούν στα Κοινωνικά Μέσα. Στη συνέχεια πραγματοποιείται μια πρακτική εφαρμογή. Συγκεκριμένα, εκτελείται ένας από τους πιο δημοφιλείς αλγορίθμους που εφαρμόζεται ευρέως στα Συστήματα Προτάσεων, ο αλγόριθμος των κ-κοντινότερων γειτόνων. Ως είσοδος του προγράμματος χρησιμοποιούνται δεδομένα που συλλέχθηκαν από χρήστες του Facebook. Μεταβάλοντας τις παραμέτρους του αλγορίθμου γίνεται μια προσπάθεια να εξεταστεί κατά πόσο είναι εφικτή και αποτελεσματική η χρήση δεδομένων από Κοινωνικά Μέσα για την εξαγωγή αποτελεσμάτων από τρίτες εφαρμογές, και όχι από αυτήν καθαυτήν την πλατφόρμα που έχει στη διάθεσή της περισσότερη γνώση ούτως ή άλλως. Τέλος, μελετώνται οι όποιες αδυναμίες του αλγορίθμου και προτείνονται επεκτάσεις προσανατολισμένες σε δεδομένα που προέρχονται από τα Κοινωνικά Μέσα.

Λέξεις Κλειδιά : <<Ιστός 2.0, Κοινωνικά Μέσα, Συστήματα Προτάσεων, Συνεργατικό Φιλτράρισμα, Facebook>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

Nowadays, the amount of information that can be found or posted on the Internet is huge. Internet users are overwhelmed by products and services that cover a vast area of sectors. This, coupled with the fact that users lack relevant knowledge and experience makes it very difficult for them to make decisions regarding which of the products available meet their specific needs and preferences. Recommender Systems have come to solve this problem by finding products that interest the user and helping him come to a decision.

The purpose of this thesis is to study a great software tool, Recommender Systems, in correlation with a very popular sector of the modern Web, Social Media. We try to understand the background, the operation mode and the usefulness of Recommender Systems, as well as study some implementations of Recommender Systems regarding the Social Media.

At first, we analyze in a theoritical level all kinds of Recommender Systems, the algorithms they use and the areas in which they operate and especially we study in depth Recommender Systems that rely on Social Media. Then, we implement an application by running one of the most popular algorithms that is widely applied to Recommender Systems, the k-nearest neighbors algorithm. We use data that we collected from Facebook users as input to our program. By changing the program parameters we try to examine whether it is feasible and effective to use data collected from Social Media platforms in order to export results from external applications and not from the platform itself that posseses more knowledge anyway. Finally, we examine any weaknesses that the algorithm may have and propose future extensions regarding data coming from the Social Media.

Keywords : << Web 2.0, Social Media, Recommender Systems, Collaborative Filtering, Facebook>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον υπεύθυνο καθηγητή, Αναπληρωτή καθηγητή ΕΜΠ κύριο Δημήτριο Ασκούνη, που με εμπιστεύτηκε και μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα με πολλές δυνατότητες ανάλυσης και επέκτασης.

Ιδιαιτέρως θα ήθελα να ευχαριστήσω τον Υποψήφιο Διδάκτορα Ιωσήφ Αλβέρτη, που η βοήθειά του ήταν καταλυτική για την εκπόνηση αυτής της διπλωματικής εργασίας. Με καθοδήγησε με πολύ εύστοχες προτάσεις, με ενέπνευσε στις όποιες δυσκολίες παρουσιάστηκαν και με συμβούλευσε με μεγάλη προθυμία. Έδειξε απίστευτο ενδιαφέρον για να φέρω εις πέρας αυτήν την διπλωματική και η βοήθειά του ήταν υπερπολύτιμη. Για όλα αυτά τον ευχαριστώ αφάνταστα.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια και του φίλους μου για την στήριξη και την συμπαράσταση που έδειξαν στην πολυετή πορεία μου στο Εθνικό Μετσόβιο Πολυτεχνείο, και ιδιαιτέρως τους στενούς μου φίλους που με βοήθησαν στην συλλογή δεδομένων του πειράματος!

Ευχαριστώ πολύ

Πίνακας περιεχομένων

Πίνακας Εικόνων

Πίνακας Πινάκων

1

Εισαγωγή

Με την πάροδο του χρόνου ο όγκος των πληροφοριών που μπορεί να βρει κανείς στο διαδίκτυο αυξάνεται κατακόρυφα. Οι χρήστες του διαδικτύου κατακλύζονται από προϊόντα και υπηρεσίες και αυτό καθιστά όλο και δυσκολότερη την λήψη αποφάσεων σχετικά με το ποιά από όλα τα διαθέσιμα προϊόντα (βιβλία, CD, νέα, ταινίες) είναι κατάλληλα για τις ανάγκες και τις προτιμήσεις τους. Στο πρόβλημα αυτό ως λύση εμφανίστηκαν τα Συστήματα Προτάσεων (Recommender Systems). Τα Συστήματα Προτάσεων είναι εργαλεία λογισμικού που δουλεύουν με συγκεκριμένους αλγορίθμους φιλτραρίσματος και επεξεργασίας πληροφοριών με σκοπό να προτείνουν στον χρήστη προϊόντα που θα τους ενδιαφέρουν και θα τους βοηθήσουν να καταλήξουν σε απόφαση.

Η ανάπτυξη των Recommender Systems ξεκίνησε από μία απλή παρατήρηση: οι άνθρωποι συχνά στηρίζονται σε προτάσεις άλλων για συνηθισμένες, καθημερινές αποφάσεις, όπως για παράδειγμα στην πρόταση ενός φίλου για το ποιό βιβλίο να διαβάσουν ή στην αξιολόγηση ενός κριτικού κινηματογράφου στην εφημερίδα για το ποιά ταινία να παρακολουθήσουν. Στη σημερινή εποχή, με την εξέλιξη του Ιστού 2.0, οι χρήστες πλέον έχουν την δυνατότητα να συμμετέχουν στην δημοσίευση πληροφοριών, να μοιράζονται αρχεία, προτιμήσεις και γνώσεις με ανθρώπους από όλο τον κόσμο σε πλατφόρμες που είναι πλέον γνωστές ως Κοινωνικά Μέσα. Συνεπώς, ένας χρήστης χωρίς μεγάλη εμπειρία που αδυνατεί να χειριστεί τον τεράστιο αριθμό επιλογών που έχει στη διάθεσή του, στρέφεται προς τα Συστήματα Προτάσεων. Κοινωνικά Μέσα και Διαδικτυακοί Ιστοί όπως τα YouTube, Netflix, Amazon, IMDb ή το iTunes χρησιμοποιούν ευρέως συστήματα προτάσεων για να διευκολύνουν τους χρήστες στην εύρεση του αντικειμένου που τους ενδιαφέρει.

Ειδικότερα, τα Recommender Systems χρησιμοποιούν χαρακτηριστικά όπως το profile του χρήστη, τα στοιχεία των προϊόντων ή το κοινωνικό περιβάλλον του χρήστη (φίλοι), για να προβλέψουν τον βαθμό που ο χρήστης θα έδινε σε ένα προϊόν που δεν έχει βαθμολογήσει. Στόχος των Συστημάτων αυτών είναι να βρουν στον χρήστη όχι μόνο προϊόντα που θα τον ενδιαφέρουν καθότι είναι κοντά στις προτιμήσεις του, αλλά και προϊόντα που πιθανότατα δεν θα έβρισκε ποτέ μόνος του γιατί δεν ανήκουν στον βασικό κύκλο ενδιαφερόντων του.

Στην συγκεκριμένη διπλωματική εργασία αναλύονται σε βάθος τα Recommender Systems, δίνοντας βάρος κυρίως στα Συστήματα Προτάσεων που στηρίζονται και χρησιμοποιούν δεδομένα από τα Κοινωνικά Μέσα. Παράλληλα γίνεται μια προσπάθεια να εντοπιστούν ενδεχόμενα κενά που υπάρχουν στις υπάρχουσες προσεγγίσεις με τις υπάρχουσες μεθόδους αξιολόγησης στα Κοινωνικά Μέσα.

2

Αντικείμενο Διπλωματικής

Η Διπλωματική αυτή αποτελεί μια μελέτη ενός σπουδαίου λογισμικού εργαλείου, των Συστημάτων Προτάσεων, σε συσχετισμό με έναν πολύ σύγχρονο τομέα του διαδικτυακού τόπου, τα Κοινωνικά Μέσα. Σκοπός της εργασίας είναι να κατανοηθεί το υπόβαθρο, ο τρόπος λειτουργίας και η χρησιμότητα των Συστημάτων Προτάσεων, καθώς και να εξεταστούν κάποιες εφαρμογές τους στα Κοινωνικά Μέσα. Αρχικά λοιπόν αναλύονται σε θεωρητικό επίπεδο τα είδη των Recommender Systems, οι αλγόρίθμοι που χρησιμοποιούν και τα πεδία στα οποία εφαρμόζονται, εμβαθύνοντας κυρίως στα Συστήματα που στηρίζονται στα Κοινωνικά Μέσα, και στη συνέχεια πραγματοποιείται μια εφαρμογή. Η εφαρμογή σχετίζεται με το τρέξιμο ενός από τους ήδη υπάρχοντες αλγορίθμους που χρησιμοποιείται ευρέως στα Συστήματα Προτάσεων, του αλγορίθμου κ-κοντινότερων γειτόνων και στηρίζεται σε δεδομένα που ανακτήθηκαν από το Facebook.

2.1 Παρουσία των ηλεκτρονικών υπηρεσιών και των συστημάτων προτάσεων στο σύγχρονο διαδίκτυο

Η ραγδαία αύξηση του όγκου αλλά και της ποικιλίας των πληροφοριών στο διαδίκτυο την τελευταία δεκαετία καθιστά πλέον πολύ δύσκολη την διαχείριση και την αφομοίωσή τους από τους χρήστες. Επιπλέον, πολυάριθμες υπηρεσίες ηλεκτρονικού εμπορίου έχουν εισαχθεί στον διαδικτυακό χώρο, δυσχεραίνοντας την δυνατότητα εύρεσης και επιλογής των προϊόντων. Η διαθεσιμότητα μιας τεράστιας ποικιλίας υπηρεσιών, αντί να ωφελεί τους χρήστες και να διευκολύνει την λήψη αποφάσεων, άρχισε αντιθέτως να προκαλεί αποπροσανατολισμό και σύγχυση. Όπως ο Alvin Toffler πολύ εύστοχα έχει δηλώσει, « *όταν το άτομο είναι βυθισμένο σε μια γρήγορα και ακανόνιστα μεταβαλλόμενη κατάσταση, ή σε ένα καινοτόμο πλαίσιο δεδομένων, δεν είναι πια ικανός να κάνει τις λογικά σωστές εκτιμήσεις στις οποίες στηρίζεται η ορθολογική συμπεριφορά»*. [1]

Ως τρόπος αντιμετώπισης του παραπάνω προβλήματος, γνωστού ως «υπερφόρτωση πληροφοριών» εμφανίστηκαν τα Συστήματα Προτάσεων. Ένα τέτοιο σύστημα χειρίζεται το

πρόβλημα αυτό παρουσιάζοντας στον χρήστη καινούρια προϊόντα που δεν έχει εξερευνήσει ο ίδιος, αλλά πιθανότατα τον ενδιαφέρουν. Το Recommender System παράγει προτάσεις, στηριζόμενο σε δεδομένα των χρηστών και των προϊόντων, αλλά και σε γνώσεις από προηγούμενες συμπεριφορές των χρηστών. Τα δεδομένα και οι πληροφορίες που άμεσα (βαθμολογία ή κατάταξη προϊόντος βάσει κάποιας κλίμακας) ή έμμεσα (πλοήγηση σε συγκεκριμένη σελίδα προϊόντος δηλώνει έμμεσα ενδιαφέρον) παρέχει ο χρήστης αποθηκεύονται στην βάση δεδομένων του συστήματος και χρησιμοποιούνται για την δημιουργία συστάσεων στην επόμενη αλληλεπίδραση του χρήστη με το σύστημα, βελτιώνοντας την απόδοση της διαδικασίας προβλέψεων και εμπνέοντας έτσι την εμπιστοσύνη των χρηστών στο σύστημα. Οι χρήστες έχουν ανάγκη από ένα Recommender System καθώς δεν έχουν την απαραίτητη γνώση για να πάρουν αποφάσεις ανεξάρτητα.

Στις μέρες μας, υπάρχουν πλέον συνέδρια, μελέτες και εργαστήρια αφιερωμένα στον συγκεκριμένο τομέα. Εδώ αξίζει να αναφερθεί το ετήσιο συνέδριο ACM Recommender Systems που ξεκίνησε το 2007. Το συνέδριο αυτό συγκεντρώνει τους ιθύνοντες του τομέα –στην έρευνα ή στην πρακτική- για να εξερευνήσουν τις τελευταίες καινοτομίες, να συζητήσουν σημαντικά προβλήματα και προκλήσεις και να εμβαθύνουν στην κατανόηση των Συστημάτων Προτάσεων [2]. Επιπλέον, σε πανεπιστημιακά ιδρύματα όλου του κόσμου υπάρχουν πλέον μαθήματα στα οποία διδάσκονται τα Recommender Systems, ενώ σε ακαδημαϊκά περιοδικά υπάρχουν άρθρα που ενημερώνουν τους αναγνώστες για τις εξελίξεις στην έρευνα των συστημάτων αυτών.

Εφαρμογές των Συστημάτων Προτάσεων υπάρχουν σε τομείς όπως η ψυχαγωγία με προτάσεις για ταινίες ή μουσικά κομμάτια, το ηλεκτρονικό εμπόριο με προτάσεις όπως υπολογιστές ή βιβλία στους καταναλωτές, οι διαδικτυακές υπηρεσίες με συστάσεις για ταξίδια ή ενοικιάσεις σπιτιών και το «προσαρμοσμένο» περιεχόμενο, με προτάσεις παραδείγματος χάριν για νέα και άρθρα που ενδιαφέρουν τον συγκεκριμένο χρήστη.

Τα Συστήματα Προτάσεων ευνοούν τόσο τους υπεύθυνους παροχής ηλεκτρονικών υπηρεσιών όσο και τους χρήστες των υπηρεσιών αυτών. Τα άτομα που παρέχουν υπηρεσίες και προϊόντα ηλεκτρονικά μέσω του διαδικτύου έχουν πολλά πλεονεκτήματα από την αξιοποίηση ενός τέτοιου συστήματος, με βασικότερο την αύξηση των πωλήσεων των προϊόντων τους. Αυτό οφείλεται στο γεγονός ότι οι προτάσεις του συστήματος ταιριάζουν με τις ανάγκες και τα ενδιαφέροντα του χρήστη. Το χαρακτηριστικό αυτό των συστημάτων προτάσεων συμβάλει στην ικανοποίηση του χρήστη, καθώς οι ενδιαφέρουσες συστάσεις βελτιώνουν την εμπειρία του, και στην αύξηση της εμπιστοσύνης του στο σύστημα. Επίσης, μία λειτουργία που καθιστά πετυχημένο ένα Recommender System είναι η δυνατότητά του να προτείνει στον χρήστη προϊόντα που θα δυσκολευόταν να βρει από μόνος του, γιατί για παράδειγμα δεν ανήκουν στην λίστα με τα πιο δημοφιλή προϊόντα. Η λειτουργία αυτή του συστήματος ευνοεί τους παρόχους, καθώς οι πωλήσεις τους δεν περιορίζονται σε συγκεκριμένες υπηρεσίες. Όπως αναφέραμε προηγουμένως, ο χρήστης δηλώνει είτε άμεσα είτε έμμεσα τις προτιμήσεις και τα ενδιαφέροντά του. Με αυτόν τον τρόπο, ο πωλητής ηλεκτρονικών προϊόντων μπορεί να καταλάβει τι αρέσει στον χρήστη και να χρησιμοποιήσει αυτήν την πληροφορία στις αποφάσεις του σχετικά με την παραγωγή, προσφέροντας έτσι ικανοποίηση στον χρήστη και αυξάνοντας τις πιθανότητες να επισκεφθεί και πάλι την σελίδα του.

Η αξιοποίηση των συστημάτων προτάσεων δημιουργεί μια σειρά πλεονεκτημάτων και για τους ίδιους τους χρήστες. Όπως αναφέρεται σε σχετικό paper, τα διάφορα είδη συστημάτων προτάσεων ή και πολλές φορές συνδυασμός αυτών, συνεισφέρουν σε διάφορες λειτουργίες των Recommender Systems που ευνοούν τους χρήστες, όπως οι παρακάτω: η εύρεση χρήσιμων προϊόντων, η υπόδειξη των προϊόντων εκείνων από την λίστα προτάσεων που θα ενδιαφέρουν περισσότερο από τα υπόλοιπα τον χρήστη, η αλυσιδωτή πρόταση προϊόντων που σχετίζονται μεταξύ τους και ο χρήστης, μετά την αγορά κ χρήση του πρώτου προϊόντος, θα ήθελε να προχωρήσει και στην αγορά του επόμενου ή η ομαδική πρόταση συσχετιζόμενων προϊόντων, η δυνατότητα να δοκιμάσει ο

χρήστης το σύστημα για να αποκτήσει άποψη για την αποτελεσματικότητά του και κατ' επέκταση να αποφασίσει αν θα εμπιστευτεί το σύστημα ή όχι, η δυνατότητα του χρήστη, μέσω τις βαθμολογίας του στις διάφορες υπηρεσίες, να βοηθήσει άλλους χρήστες που ενδιαφέρονται για κάποιες από τις υπηρεσίες αυτές ή να ικανοποιηθεί μόνο και μόνο από την δήλωση των βαθμολογιών του, δηλαδή από την δημόσια έκφραση της άποψής του. [3], [4]

Εικόνα 1: Αποτέλεσμα Recommender System στο Netflix [5]

Εικόνα 2: Αποτέλεσμα Recommender System στο amazon.com [6]

2.2 Κοινωνικά Μέσα (Social Media)

Η τελευταία δεκαετία έχει στιγματιστεί από μια νέα τάση στον Παγκόσμιο Ιστό. Στα πλαίσια της παγκοσμιοποίησης, ένας τεράστιος αριθμός ανθρώπων συνδέονται στο Internet τόσο για να χρησιμοποιήσουν ήδη υπάρχοντα δεδομένα για την έρευνά τους, όσο και για να δημοσιεύσουν οι ίδιοι νέα πληροφορία. Υπάρχουν πλέον πλατφόρμες που δεν παρέχουν απλά πληροφορίες στον ενδιαφερόμενο, αλλά αλληλεπιδρούν μαζί του καθώς τις προσφέρουν. Η τάση που οδήγησε στην γιγάντωση των Κοινωνικών Μέσων ήταν η δυνατότητα που είχε κάθε χρήστης να συνεισφέρει εξίσου σε περιεχόμενο, και συγκεκριμένα πολυμεσικό περιεχόμενο. Στον χρήστη παρουσιάζεται η δυνατότητα να παρουσιάζει τις απόψεις του και να συνεισφέρει στο περιεχόμενο των ιστοτόπων. Με αυτόν τον τρόπο, ο χρήστης, από παθητικός δέκτης της πληροφορίας, γίνεται ενεργός χρήστης έχοντας στα χέρια του το πλεονέκτημα της συμμετοχής.

Η νέα αυτή τάση στο σύγχρονο διαδίκτυο οδήγησε στην δημιουργία ενός καινούριου όρου που περιγράφει το φαινόμενο αυτό, του όρου Κοινωνικά Μέσα (Social Media). Πρόκειται για μέσα κοινωνικής αλληλεπίδρασης, στηριζόμενα στην χρήση τεχνολογιών που επιτρέπουν το πέρασμα του σημερινού διαδικτυακού χώρου σε έναν διαδραστικό διάλογο. Οι Andreas Kaplan και Michael Haenlein ορίζουν τα Κοινωνικά Μέσα ως «*μια ομάδα στηριζόμενων στο Internet εφαρμογών που αξιοποιούν τις ιδεολογικές και τεχνολογικές βάσεις του Web 2.0, το οποίο επιτρέπει την δημιουργία και την ανταλλαγή περιεχομένου από τους χρήστες*»[7]. Οι χρήστες παίρνουν πλέον τον ρόλο του «παραγωγού-καταναλωτή», αφού πια δεν χρησιμοποιούν μόνο το τελικό προϊόν, αλλά συμμετέχουν ενεργά στην παραγωγή του. Τα Μέσα κοινωνικής δικτύωσης και παραγωγής πληροφοριών στηρίζονται σε τεχνολογίες που είναι εύκολα προσβάσιμες και προσιτές στο ευρύ κοινό. Η πλειοψηφία των Κοινωνικών Μέσων έχει παγκόσμια εμβέλεια, με εκατομμύρια χρήστες να συνδέονται και να αλληλεπιδρούν καθημερινά. Με αυτόν τον τρόπο δημιουργείται μια παγκόσμια, μαζική συνεργασία και ο χρήστης πλέον αποκτά δύναμη, αφού οι πλατφόρμες των Κοινωνικών Μέσων βασίζονται πρωταρχικά σε εκείνον ο οποίος συνεισφέρει με δεδομένα, δημιουργεί διαδικτυακές σχέσεις και συμμετέχει σε διαδικτυακές κοινότητες.

Ο Παγκόσμιος Ιστός 2.0 (Web 2.0), που αναφέραμε προηγουμένως, σχετίζεται με εφαρμογές που προωθούν και εφαρμόζουν την ανταλλαγή πληροφοριών και την διαλειτουργικότητα, έχοντας πάντα ως κέντρο τους τον χρήστη. Ένας ιστότοπος Web 2.0, έχοντας μια αρχιτεκτονική βασισμένη στην συμμετοχή, επιτρέπει στους χρήστες να συνεργάζονται μεταξύ τους είτε ως δημιουργοί, είτε ως καταναλωτές περιεχομένου. Κάποιες από τις τεχνολογίες που χρησιμοποιούνται στην ανάπτυξη του Web 2.0 είναι οι REST και XML για την ανάπτυξη καθώς και Adobe Flash και Javascript/Ajax πλαίσια, όπως τα jQuery και Yahoo! UI Library, για την διεπαφή χρήσης.

Ηλεκτρονικές εταιρίες στηρίζουν, αξιοποιούν και ωφελούνται από τα κοινωνικά διαδικτυακά μέσα, καθώς η μαζική αυτή αλληλεπίδραση και η παραγωγή περιεχομένου από τους χρήστες (user-generated content) αναδιαρθρώνει τις δομές τους, δημιουργεί ένα νέο μοντέλο έξυπνων επιχειρήσεων και αλλάζει την ροή της παγκόσμιας οικονομίας. Οι ηλεκτρονικές επιχειρήσεις εκμεταλλεύονται αυτήν την παγκόσμια μαζική συνεργασία και την τεχνολογία ανοιχτού κώδικα προς όφελός τους. Πολλοί βέβαια είναι και εκείνοι που εναντιώνονται στην φιλοσοφία των Κοινωνικών Μέσων, θεωρώντας ότι βασίζονται σε μια αναρχία που επιτρέπει να επικρατούν μόνο εκείνοι που εκφράζουν πιο ισχυρά την γνώμη τους. Άλλες ανησυχίες προκύπτουν σχετικά με το απόρρητο των προσωπικών πληροφοριών, αφού σε ορισμένες περιπτώσεις ιστοσελίδων, προσωπικές παρουσιάζονται από προεπιλογή.

Στα Κοινωνικά Μέσα συγκαταλέγονται διαδικτυακοί ιστότοποι όπως forums,κοινωνικά blogs, wikis, weblogs, podcasts ή social bookmarking. Σύμφωνα με τους Kaplan και Haenlein υπάρχουν έξι (6) διαφορετικά είδη κοινωνικών μέσων:

- Συνεργατικά έργα (collaborative projects)
- Blogs και microblogs
- Κοινότητες περιεχομένου (content communities)
- Ιστοσελίδες κοινωνικής δικτύωσης (social networking sites)
- Εικονικοί κόσμοι παιχνιδιού (virtual game worlds)
- Εικονικοί κοινωνικοί κόσμοι (virtual social worlds)

Στην Εικόνα 3 παρουσιάζονται κάποια από τα δημοφιλέστερα Social Media της εποχής μας. Συμπεριλαμβάνει το Facebook, το twitter και το MyBlogLog ως υπηρεσίες κοινωνικής δικτύωσης, το Flickr, το Viddler, το MySpace και το YouTube που σχετίζονται με χρήστες που μοιράζονται video και εικόνες, το del.icio.us ως υπηρεσία κοινωνικών σελιδοδεικτών, το Linedin, μια ιστοσελίδα κοινωνικής δικτύωσης προσανατολισμένη στον επιχειρηματικό κόσμο, το Technorati ως μια διαδικτυακή μηχανή αναζήτησης για την εύρεση blog, το Stumble Upon ως μηχανή εύρεσης και πρότασης περιεχομένου στους χρήστες του, το Yahoo!Groups ως μια από τις μεγαλύτερες πλατφόρμες online συζητήσεων κ.ά. Τα Μέσα αυτά χρησιμοποιούνται ευρέως από χρήστες ανά όλο τον κόσμο, επιτρέποντάς τους να αλληλεπιδρούν και να μοιράζονται πληροφορίες. Όπως πολύ εύστοχα έθεσε ένας ακτιβιστής στο Κάιρο, «*Εμείς χρησιμοποιούμε το Facebook για να προγραμματίσουμε τις διαμαρτυρίες, το Twitter για να συντονιστούμε, και το YouTube για να το πούμε στον κόσμο.*» [8]

Εικόνα 3: Κοινωνικά μέσα στο σύγχρονο διαδίκτυο [9]

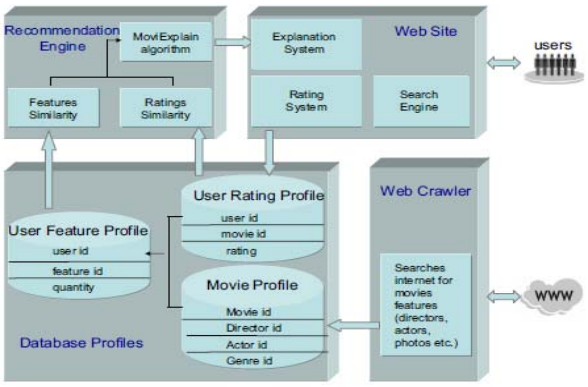
2.3 Συστήματα Προτάσεων και Κοινωνικά Μέσα

Τα στηριζόμενα στα Κοινωνικά Μέσα Recommender Systems έχουν στόχο να περιορίσουν το πρόβλημα της «υπερφόρτωσης πληροφοριών» (ο χρήστης βρίσκεται αντιμέτωπος με περισσότερη πληροφορία από όση είναι ικανός να αφομοιώσει) που έχει εμφανιστεί τα τελευταία χρόνια στο διαδίκτυο. Διευκολύνουν τους χρήστες των Κοινωνικών Μέσων παρουσιάζοντάς του μόνο τις πιο σχετικές και ενδιαφέρουσες για εκείνους πληροφορίες. Το Amazon.com μας προτείνει βιβλία και το Facebook μας συστήνει φίλους, στηριζόμενα είτε σε προηγούμενες επιλογές μας είτε στον κοινωνικό μας κύκλο, αναλόγως την τεχνολογία που χρησιμοποιείται. Επίσης, ένα άλλο αποτέλεσμα στο οποίο στοχεύουν είναι τόσο να βοηθήσουν στην αύξηση της ενασχόλησης και της ενεργής συμμετοχής των ήδη υπαρχόντων χρηστών κοινωνικών μέσων όσο και να ωθήσουν νέους χρήστες να εμπλακούν με τις ιστοσελίδες των κοινωνικών μέσων. Προτάσεις για wikis, blogs, tags ή για φίλους χρησιμοποιούν εξατομικευμένες τεχνικές, προσανατολισμένες στις ανάγκες και τα ενδιαφέροντα του «χρήστη-στόχου» (target-user).

Τα Κοινωνικά Μέσα και τα εξατομικευμένα Συστήματα Προτάσεων μπορούν να έχουν οφέλη το ένα από το άλλο. Από την μία πλευρά, τα Κοινωνικά Μέσα μας εισάγουν σε μια νέα μορφή δημοσίων δεδομένων και μεταδεδομένων (βαθμολογήσεις, σχόλια, tags), τα οποία μπορούν να χρησιμοποιηθούν για να διευκολύνουν και να ενισχύσουν τις συστάσεις. Από την άλλη πλευρά, τα Recommender Systems έχουν ένα πολύ βασικό μερίδιο ευθύνης στην επιτυχία των εφαρμογών των Κοινωνικών Μέσων αλλά και του Κοινωνικού Ιστού γενικότερα, εξασφαλίζοντας ότι ο κάθε χρήστης θα λαμβάνει προτάσεις προσαρμοσμένες στα γούστα του.

Τα τελευταία χρόνια, έχουν εμφανιστεί αρκετές εξατομικευμένες υπηρεσίες προτάσεων στα Κοινωνικά Μέσα. Τα συστήματα αυτά προβλέπουν την προτίμηση του χρήστη για κάποιο νέο προϊόν (βιβλίο, ταινία, site, μουσικό κομμάτι κ.ά.) που δεν γνωρίζει και συνήθως η προτίμηση αυτή παρουσιάζεται με την μορφή βαθμολογίας. Για παράδειγμα, το StumbleUpon είναι μια μηχανή προτάσεων που συστήνει ιστοσελίδες στηριζόμενο στην προηγούμενη συμπεριφορά του target-user (βαθμολογήσεις του ή επιλογές συγκεκριμένων θεματικών ενοτήτων που τον ενδιαφέρουν) και σε βαθμολογήσεις των φίλων του ή άλλων χρηστών που έχουν παρόμοια ενδιαφέροντα με εκείνον. Πρόσφατα, κάποια από τα δημοφιλέστερα με την μεγαλύτερη συμμετοχή Κοινωνικά Μέσα έχουν προσθέσει στις υπηρεσίες που προσφέρουν και μηχανές συστημάτων προτάσεων, όπως το YouTube που πλέον στην αρχική του σελίδα συμπεριλαμβάνει προτάσεις για video βάσει όσων έχει δει προηγουμένως ο χρήστης και όσων έχει δηλώσει ως αγαπημένα. Η εφαρμογή αυτή είχε ως αποτέλεσμα να αυξηθεί ο αριθμός των χρηστών που επισκέπτονται την αρχική σελίδα, η συχνότητα των επισκέψεων στο YouTube και ο αριθμός των πελατών που εγγράφονται στην συγκεκριμένη πλατφόρμα. Παρομοίως, η ιστοσελίδα συλλογής κοινωνικών νέων Digg έχει ενσωματώσει ένα εξατομικευμένο Recommender System, παρουσιάζοντας νέα με θεματολογία που ενδιαφέρει τον χρήστη με χρήση δεδομένων χρηστών με κοινά ενδιαφέροντα με τον target-user.

Το γεγονός ότι η πληροφορία στις ιστοσελίδες Κοινωνικών Μέσων σε πολλές περιπτώσεις είναι δημόσια (tags, σχόλια, βαθμολογήσεις) προσφέρει διαφάνεια στα Συστήματα Προτάσεων των Κοινωνικών Μέσων. Έτσι, προέκυψαν τεχνικές που προσφέρουν εξηγήσεις και δικαιολογούν τις προτάσεις που γίνονται στον εκάστοτε χρήστη (π.χ. «Σας προτείνουμε αυτήν την ταινία επειδή έχετε παρακολουθήσει πολλές ταινίες που ανήκουν στο ίδιο είδος με αυτήν»), με στόχο να αυξήσουν το αίσθημα της εμπιστοσύνης του χρήστη απέναντι στο σύστημα και να παρακινήσουν για περισσότερη ενεργή συμμετοχή. Μέσω των εξηγήσεων, αυξάνονται οι πιθανότητες να δεχτεί ο χρήστης την πρόταση που του έγινε και να δοκιμάσει το προτεινόμενο προϊόν.



Εικόνα 4: Τα μέρη του συστήματος προτάσεων MoviExplain [10]

Επιπλέον, τα Recommender Systems που στηρίζονται στα Social Media δεν παρέχουν προτάσεις μόνο σε μεμονωμένους χρήστες, αλλά πολύ συχνά απευθύνονται και σε κοινότητες χρηστών. Οι ομαδικές συστάσεις λαμβάνουν υπ' όψιν τις διαφορετικές προτιμήσεις όλων των μελών της κοινότητας ώστε να παραγάγουν μια πρόβλεψη που θα καλύπτει την συνολική συμπεριφορά.

Ένα άλλο σημαντικό ζήτημα σχετικά με Συστήματα Προτάσεων είναι η αξιολόγηση των προβλέψεων που παράγουν. Τα Κοινωνικά Μέσα παρουσιάζουν προοπτικές για νέες μεθόδους αξιολόγησης, αξιοποιώντας παραδείγματος χάριν τους χρήστες που συμμετέχουν πιο ενεργά στις ιστοσελίδες. Η ανάπτυξη και εφαρμογή μεθόδων αξιολόγησης στα Social Recommender Systems είναι πολύ χρήσιμη και εποικοδομητική, καθώς οδηγεί στην σύγκριση των διαφορετικών αλγορίθμων που χρησιμοποιούνται στις προβλέψεις , και έτσι στην εύρεση όλο και πιο αποτελεσματικών συστημάτων πρόβλεψης.

3

Εισαγωγή στα Είδη Συστημάτων Προτάσεων

Υπάρχουν αρκετοί διαφορετικοί τύποι Συστημάτων Συστάσεων. Η ποικιλία αυτή οφείλεται στους τομείς στους οποίους τα συστήματα απευθύνονται, στην γνώση και την πληροφορία που χρησιμοποιούν, στον τρόπο που οι συστάσεις συγκεντρώνονται και παρουσιάζονται στον χρήστη και τέλος, στους διαφορετικούς αλγορίθμους που εφαρμόζονται για την παραγωγή των συστάσεων. Παρακάτω αναφέρονται έξι (6) βασικές κατηγορίες προσέγγισης των Συστημάτων Προτάσεων [3], [4] :

- **Με βάση το περιεχόμενο (content-based) :** Αυτός ο τύπος συστήματος προτείνει προϊόντα βάσει των προϊόντων που ο χρήστης είχε προτιμήσει σε προηγούμενες αλληλεπιδράσεις του με το σύστημα. Ο υπολογισμός της ομοιότητας (similarity) γίνεται σε σχέση με τα χαρακτηριστικά των προϊόντων προς σύγκριση. Έτσι για παράδειγμα αν κάποιος είχε επιλέξει στο παρελθόν ένα τραγούδι που ανήκει στην κατηγορία rock, τότε το σύστημα παραγωγής προτάσεων μαθαίνει να συστήνει και άλλα μουσικά κομμάτια που ανήκουν στο συγκεκριμένο είδος.
- **Συνεργατικό φιλτράρισμα (collaborative filtering) :** Η κατηγορία αυτή είναι η πιο ευρέως χρησιμοποιούμενη στα Κοινωνικά Μέσα αλλά και γενικότερα. Τα συστήματα αυτού του τύπου συστήνουν προϊόντα που είχαν αρέσει στο παρελθόν σε άλλους χρήστες με όμοια γούστα με τον χρήστη-στόχο, συσχετίζουν λοιπόν τους χρήστες μεταξύ τους. Η ομοιότητα των ενδιαφερόντων δύο χρηστών υπολογίζεται με βάση το πόσο κοινό είναι το παρελθόν των συγκεκριμένων χρηστών στις βαθμολογήσεις.
- **Με βάση την γνώση (knowledge-based) :** Τα συστήματα αυτού του είδους στηρίζονται για τις συστάσεις που κάνουν σε συγκεκριμένη γνώση η οποία καθορίζει κατά πόσο τα χαρακτηριστικά ενός προϊόντος ανταποκρίνονται στις ανάγκες και τα ενδιαφέροντα του χρήστη, δηλαδή αν το προϊόν θα είναι χρήσιμο στον χρήστη ή όχι. Το σύστημα συγκεντρώνει τα αιτήματα του χρήστη και προτείνει και εξηγεί τις συστάσεις που βρίσκει ως λύση. Η συνάρτηση ομοιότητας στα Συστήματα Προτάσεων με βάση την γνώση εκτιμά πόσο οι ανάγκες του χρήστη συσχετίζονται με τις συστάσεις και έτσι τελικά δείχνει την χρησιμότητα της σύστασης για τον ενδιαφερόμενο.

- **Δημογραφικά (Demographic) :** Τα Δημογραφικά Συστήματα Προτάσεων συστήνουν προϊόντα βάσει του δημογραφικού profile του χρήστη. Πολλές ιστοσελίδες εφαρμόζουν απλές εξατομικευμένες λύσεις με βάση την δημογραφία, όπως για παράδειγμα κάποιες φορές οι χρήστες οδηγούνται σε συγκεκριμένες σελίδες σύμφωνα με την χώρα και την γλώσσα τους, ή οι συστάσεις προσαρμόζονται στην ηλικία ή στο φύλλο του χρήστη. Τα συστήματα αυτά χρησιμοποιούνται κυρίως στο πεδίο του marketing.
- **Υβριδικά Συστήματα Προτάσεων (hybrid recommender systems) :** Η κατηγορία αυτή συστημάτων χρησιμοποιεί ένα συνδυασμό των μεθόδων που αναφέραμε παραπάνω, εκμεταλλευόμενα τα προτερήματα τις μίας τεχνικής για να καλύψουν τα μειονεκτήματα της άλλης. Υπάρχουν πολλοί διαφορετικοί τρόποι με τους οποίους συνδυάζονται δύο ή και περισσότερες τεχνικές συστημάτων προτάσεων για να δημιουργηθεί ένα υβριδικό σύστημα. Στόχος του συνδυασμού διαφορετικών μεθόδων είναι η βελτίωση της απόδοσής τους.

Παρακάτω ακολουθεί Πίνακας που συνοψίζει τις κατηγορίες Συστημάτων Προτάσεων που αναφέραμε. Γίνεται η υπόθεση ότι I είναι η ομάδα των προϊόντων από τα οποία γίνονται οι συστάσεις, U είναι η ομάδα των χρηστών των οποίων οι προτιμήσεις και βαθμολογήσεις είναι γνωστές, u είναι ο χρήστης-στόχος για τον οποίο πρέπει να παραχθεί η σύσταση και i είναι το προϊόν για το οποίο ζητείται να προβλεφθεί η βαθμολόγηση του u.

Πίνακας 1 : Τεχνικές των Συστημάτων Προτάσεων

ΤΕΧΝΙΚΗ	ΔΕΔΟΜΕΝΑ	ΕΙΣΟΔΟΣ	ΔΙΑΔΙΚΑΣΙΑ
Content-based	Χαρακτηριστικά των προϊόντων στο I	Βαθμολογήσεις του u για τα προϊόντα στο I	Παραγωγή profile με βάση την βαθμολογική συμπεριφορά του u και χρήση του στο i
Collaborative	Βαθμολογήσεις U στα προϊόντα του I	Βαθμολογήσεις του u για τα προϊόντα στο I	Εύρεση χρηστών του U ομοίων του u και χρήση των βαθμολογιών τους για το i
Knowledge-based	Χαρακτηριστικά των προϊόντων στο I και γνώση κατά πόσο τα προϊόντα του I ανταποκρίνονται στις ανάγκες του u	Περιγραφή των αναγκών και ενδιαφερόντων του u	Εύρεση συσχέτισης μεταξύ του I και των αναγκών του u
Demographic	Δημογραφικά δεδομένα για τους χρήστες του U και τις βαθμολογήσεις τους στα προϊόντα του I	Δημογραφικές πληροφορίες για τον u	Εύρεση χρηστών του U που είναι δημογραφικά όμοιοι με τον u και χρήση των βαθμολογιών τους για το i

Στα επόμενα κεφάλαια αναλύονται περαιτέρω τα δύο πρώτα είδη, τα Συστήματα Προτάσεων με βάση το περιεχόμενο και το Συνεργατικό Φιλτράρισμα. Τα Κοινωνικά Μέσα και τα δεδομένα που αυτά μπορούν να παρέχουν κατά περίσταση (βαθμολογίες, σχέσεις ατόμων, κατηγορίες περιεχομένου κλπ) δίνουν την δυνατότητα να εφαρμόζονται κατά περίσταση σχεδόν όλες οι μέθοδοι, επιλέγοντας και την ανάλογη πλατφόρμα. Στα πλαίσια όμως της συγκεκριμένης πειραματική εργασίας, προτιμήθηκε να δοθεί βάρος στις δύο αυτές δημοφιλέστερες μεθόδους που εφαρμόζονται ευρύτατα.

4

Συστήματα Προτάσεων με Βάση το Περιεχόμενο

Τα Συστήματα που στηρίζονται στο περιεχόμενο συστήνουν προϊόντα όμοια με εκείνα που ο χρήστης είχε προτιμήσει στο παρελθόν. Ένα τέτοιο Σύστημα μοντελοποιεί ένα profile για τον χρήστη με βάση τις ιδιότητες των αντικειμένων που έχει παλαιότερα βαθμολογήσει ο χρήστης και στην συνέχεια συνδυάζει τα χαρακτηριστικά που είναι αποθηκευμένα στο profile αυτό (προτιμήσεις, ενδιαφέροντα) με τα χαρακτηριστικά του περιεχομένου που αντιστοιχεί στο προϊόν, προτείνοντας τελικά ενδιαφέροντα για τον χρήστη προϊόντα. Το profile αυτό ενημερώνεται αυτόματα ως απάντηση σε νέα σχόλια ή βαθμολογήσεις του χρήστη και αν διαμορφωθεί σωστά και αντικατοπτρίζει τα πραγματικά του ενδιαφέροντα, τότε το σύστημα προτάσεων θα δουλεύει αποδοτικά. Τα συστήματα Συστάσεων με βάση το Περιεχόμενο χρησιμοποιούνται σε ένα μεγάλο εύρος τομέων, από την σύσταση ιστοσελίδων και άρθρων στην σύσταση εστιατορίων ή βιβλίων. Με βάση τα παραπάνω, το πρόβλημα παραγωγής συστάσεων διατυπώνεται ως εξής:

$$R : UserProfiles \times Objects \rightarrow Ratings$$

[11].

Το Σύστημα Προτάσεων σχετίζει κάθε ζευγάρι profile χρήστη-προϊόν με μια τιμή βαθμολόγησης , εκτιμώντας την παραπάνω συνάρτηση βαθμολόγησης R. Το προϊόν που είναι ψηλότερα στην βαθμολόγηση προτείνεται και στον χρήστη.

4.1 Παρουσίαση των Προϊόντων

Τα Συστήματα Προτάσεων με βάση το Περιεχόμενο συχνά διαφέρουν μεταξύ τους όσον αφορά την παρουσίαση των προϊόντων. Τα προϊόντα που θα μπορούσαν να προταθούν σε έναν χρήστη αποθηκεύονται σε μια βάση δεδομένων του συστήματος. Τα δεδομένα αποθηκεύονται σε πίνακες στους οποίους η κάθε κολώνα αντιστοιχεί και σε μία ιδιότητα (ή χαρακτηριστικό) του προϊόντος. Για παράδειγμα, σε μία εφαρμογή που προτείνει εστιατόρια, θα μπορούσε ο αντίστοιχος πίνακας να περιλαμβάνει ιδιότητες όπως το id των εστιατορίων (διαφορετικό για το καθένα), το όνομα, το εύρος των τιμών ή το είδος της κουζίνας που σερβίρει. Αυτή η βάση δεδομένων είναι μια μορφή «δομημένων δεδομένων», καθώς υπάρχει συγκεκριμένος αριθμός ιδιοτήτων οι οποίες χαρακτηρίζουν όλα τα προϊόντα της βάσης και η κάθε ιδιότητα μπορεί να λάβει συγκεκριμένες τιμές. Σε αυτήν την περίπτωση, χρησιμοποιούνται πολλοί αλγόριθμοι μηχανικής μάθησης (machine-learning algorithms) για να μάθει το σύστημα το profile του χρήστη.

Συχνά όμως, τα προϊόντα έχουν και άλλες ιδιότητες που δεν παίρνουν συγκεκριμένες τιμές. Συνεχίζοντας το προηγούμενο παράδειγμα, η βάση δεδομένων ενός συστήματος συστάσεων

εστιατορίων θα μπορούσε να έχει και άλλα χαρακτηριστικά όπως κάποια περιγραφή του εστιατορίου ή μια κριτική της κουζίνας του. Η περίπτωση αυτή αφορά τα «αδόμητα δεδομένα», οι ιδιότητες των οποίων δεν παίρνουν σαφώς καθορισμένες τιμές. Αυτό δημιουργεί περιπλοκές στην μοντελοποίηση του profile του χρήστη, καθώς δεν μπορεί να υπάρξει ποτέ ακριβής αντιστοίχηση δεδομένων που περιέχουν ελεύθερο κείμενο. Επιπλέον, δημιουργούνται προβλήματα στην εκμάθηση του profile του χρήστη λόγω ασαφειών της φυσικής γλώσσας, όπως η πολυσημία (μία λέξη έχει πολλές διαφορετικές έννοιες) και η συνωνυμία (διαφορετικές λέξεις έχουν την ίδια σημασία).

Πολλές φορές, χρησιμοποιούνται για την παρουσίαση των προϊόντων «ημι-δομημένα δεδομένα», στα οποία κάποιες ιδιότητες λαμβάνουν συγκεκριμένες τιμές και κάποιες άλλες δέχονται ελεύθερο κείμενο. Μια μέθοδος χειρισμού των πεδίων ελεύθερου κειμένου, την οποία εφαρμόζουν πολλά εξατομικευμένα συστήματα προτάσεων είναι η μετατροπή του κειμένου σε μια δομημένα παρουσίαση. Μέσω της σημασιολογικής ανάλυσης (semantic analysis), χρησιμοποιούνται λεξικά και οντολογίες, ώστε οι λέξεις να έρχονται στην μορφή της ρίζας τους. Με την διαδικασία του «stemming», το σύστημα μπορεί να καταλάβει ότι οι όροι «compute», «computes», «computer» και «computers» έχουν την ίδια ρίζα και κοινή σημασία.

Στη συνέχεια, γίνεται ανάλυση της βασικής μεθόδου προσέγγισης της παρουσίασης εγγράφων (documents) με βάση τις «λέξεις-κλειδιά». Ο όρος έγγραφα έχει μείνει παραδοσιακά, καθώς το αρχικό κίνητρο ήταν η ανάκτηση εγγράφων, αλλά πλέον, με τα Συστήματα Προτάσεων, ο όρος έγγραφα θα χρησιμοποιείται αναφερόμενος στο κείμενο περιγραφής ενός προϊόντος υποψηφίου για να συστηθεί. [12]

4.1.1 Μοντέλο Συστημάτων με βάση τις Λέξεις-Κλειδιά

Τα Συστήματα Προτάσεων με βάση το Περιεχόμενο εφαρμόζουν τεχνικές ανάκτησης πληροφορίας όπως ο συσχετισμός με λέξεις-κλειδιά ή το μοντέλο Διανυσματικού Χώρου (Vector Space Model ή VSM). Με την τεχνική αυτή, το κάθε έγγραφο μοντελοποιείται σε ένα διάνυσμα βαρών όρων στον ν-διάστατο χώρο. Το βάρος (weight) είναι ένας πραγματικός αριθμός που αντιπροσωπεύει κατά πόσο συσχετίζεται ο εκάστοτε όρος με το έγγραφο. Το βάρος w_{kj} του όρου t_k στο έγγραφο d_j είναι μια συνάρτηση της συχνότητας του t_k στο d_j, του αριθμού των εγγράφων που περιλαμβάνουν τον όρο t_k και του συνολικού αριθμού εγγράφων μιας ομάδας εγγράφων $D = \{d_1,d_2,\ldots,d_N\}$.

Η πιο συνηθισμένη μέθοδος υπολογισμού των βαρών είναι η «Συχνότητα Όρου-Αντίστροφη Συχνότητα Εγγράφου» (Term Frequency – Inverse Document Frequency ή TF-IDF). Η μέθοδος αυτή στηρίζεται στις παρατηρήσεις ότι [3] :

- Οι σπάνιοι όροι δεν είναι λιγότεροι σχετικοί με το θέμα από τους όρους που χρησιμοποιούνται συχνότερα (αυτή η θεώρηση διέπει την έννοια του IDF)
- Οι συχνές αναφορές ενός όρου σε ένα έγγραφο δεν είναι περισσότερο σχετικές με το θέμα από μοναδικές αναφορές όρων (αυτή η θεώρηση διέπει την έννοια του TF)
- Τα μεγάλης έκτασης έγγραφα δεν είναι προτιμότερα από τα μικρής έκτασης (θεώρηση της ομαλοποίησης)

Η έννοια του βάρους είναι ότι οι όροι με το μεγαλύτερο βάρος εμφανίζονται συχνότερα στο συγκεκριμένο έγγραφο από ότι στα υπόλοιπα έγγραφα της ομάδας D, και άρα είναι πιο σχετικοί με

το θέμα που πραγματεύεται το έγγραφο. Η δε ομαλοποίηση των διανυσμάτων βαρών στοχεύει στο να μην προωθούνται τα έγγραφα με την μεγαλύτερη έκταση.

Τα παραπάνω μοντελοποιούνται στις συναρτήσεις που ακολουθούν:

Συνάρτηση TF-IDF :

$$\text{TF-IDF}(t_k, d_j) = \underbrace{\text{TF}(t_k, d_j)}_{\text{TF}} \cdot \underbrace{log\frac{N}{n_k}}_{\text{IDF}}$$

[3, σελ.82]

Όπου το N δηλώνει τον συνολικό αριθμό εγγράφων της ομάδας D, το n_k δηλώνει τον αριθμό των εγγράφων στα οποία εμφανίζεται τουλάχιστον μία φορά ο όρος t_k και το $TF(t_k, d_j)$ αντιστοιχεί στην συνάρτηση :

$$\text{TF}(t_k, d_j) = \frac{f_{k,j}}{max_z f_{z,j}}$$

[3, σελ.82]

Μετά από συνημιτονική ομαλοποίηση ώστε όλα τα διανύσματα να έχουν το ίδιο μήκος, η εξίσωση των βαρών γίνεται :

$$w_{k,j} = \frac{\text{TF-IDF}(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} \text{TF-IDF}(t_s, d_j)^2}}$$

[3, σελ.82]

Η παρουσίαση των εγγράφων με το μοντέλο Διανυσματικού Χώρου (SVM) απαιτεί πρώτα τον υπολογισμό των βαρών των όρων και στη συνέχεια τον υπολογισμό της ομοιότητας (similarity) των διανυσμάτων χαρακτηριστικών γνωρισμάτων, ώστε να γίνει αντιληπτό κατά πόσο συσχετίζονται δύο έγγραφα ή όχι. Η πιο ευρέως χρησιμοποιούμενη μέθοδος υπολογισμού της ομοιότητας είναι η συνημιτονική ομοιότητα :

$$sim(d_i, d_j) = \frac{\sum_k w_{ki} \cdot w_{kj}}{\sqrt{\sum_k w_{ki}^2} \cdot \sqrt{\sum_k w_{kj}^2}}$$

[3, σελ.82]

Έτσι, η πρόβλεψη για το αν ένα προϊόν ενδιαφέρει έναν χρήστη γίνεται υπολογίζοντας την συνημιτονική ομοιότητα μεταξύ του διανύσματος των χαρακτηριστικών γνωρισμάτων του προϊόντος και του διανύσματος του profile του χρήστη, καθώς και το profile του μοντελοποιείται από διανύσματα όρων με βάρη.

Τα Συστήματα Προτάσεων με βάση τις λέξεις-κλειδιά εκτελούνται σε πολλούς τομείς στο σύγχρονο διαδίκτυο, όπως στην μουσική, στις ταινίες, στο ηλεκτρονικό εμπόριο, στον χώρο των νέων κ.ά. Ορισμένα παραδείγματα εφαρμογών που κάνουν χρήση των μοντέλων αυτών είναι τα NewT, YourNews, INFOrmer στο πεδίο των νέων, το Movies2GO στον χώρο των ταινιών και τα ifWeb, Personal WebWatcher, Letizia στον τομέα των συστάσεων ιστοσελίδων.

Σαν μια γενικότερη κριτική, θα μπορούσαμε να πούμε ότι η εφαρμογή του βασισμένου σε λέξεις-κλειδιά μοντέλου για την παρουσίαση των προϊόντων αλλά και των profiles των χρηστών μπορεί

να οδηγήσει σε υψηλή και ακριβή αποδοτικότητα του συστήματος, δεδομένου ότι το σύστημα έχει αποθηκευμένη αρκετή πληροφορία για τα ενδιαφέροντα του χρήστη. Το πρόβλημα του μοντέλου αυτού είναι ότι εμφανίζει περιορισμούς όταν απαιτούνται προηγμένα χαρακτηριστικά και έτσι αδυνατεί να χαρακτηριστεί ως «έξυπνο μοντέλο». Για παράδειγμα, αν ένας χρήστη έχει δηλώσει ενδιαφέρον για την Αρχαία Ελληνική Ιστορία, η προσέγγιση με τις λέξεις-κλειδιά θα ανακτήσει μόνο άρθρα που περιλαμβάνουν τις λέξεις «Αρχαία», «Ελληνική» και «Ιστορία» και θα παραλείψει να συστήσει άρθρα που πραγματεύονται το έργο του Περικλή ή του Σωκράτη, παρόλο που σχετίζονται με το θέμα. Στο πρόβλημα αυτό ως λύση αναπτύχθηκαν ανώτερες σημασιολογικές τεχνικές.

4.2 Profiles των χρηστών

Τα περισσότερα Συστήματα προτάσεων χρησιμοποιούν profiles των χρηστών. Τα profiles αυτά περιλαμβάνουν πληροφορίες για τα ενδιαφέροντα των χρηστών, όπως περιγραφές των προϊόντων που έχουν τραβήξει την προσοχή τους. Η περιγραφή αυτή συνήθως μοντελοποιείται με μια συνάρτηση που προβλέπει για κάθε προϊόν το πιθανό ενδιαφέρον του χρήστη. Για πρακτικούς σκοπούς, εφαρμόζεται αυτή η συνάρτηση για να ανακτηθεί ένας χ αριθμός των προϊόντων με την μεγαλύτερη πιθανότητα να αρέσουν στον χρήστη. Επίσης, τα profiles συχνά περιέχουν πληροφορίες σχετικές με τις αλληλεπιδράσεις που είχε ο χρήστης με το σύστημα στο παρελθόν, όπως η αποθήκευση προϊόντων που έχει δει ή αγοράσει ο χρήστης, βαθμολογιών του ή ερωτημάτων που έχει θέσει ο χρήστης στο σύστημα. Κρατώντας πληροφορίες από το παρελθόν, το σύστημα διευκολύνει τον χρήστη καθώς μπορεί να του παραθέσει τα προϊόντα που εκείνος είχε επισκεφτεί πρόσφατα ώστε να τα δει περαιτέρω αν τον ενδιέφεραν. Επιπλέον, η αποθήκευση των προϊόντων που η χρήστης έχει αγοράσει παλαιότερα δίνει στο σύστημα την δυνατότητα να αποκλείει από τις συστάσεις του τα προϊόντα αυτά και έτσι να είναι πιο ουσιαστικές οι προβλέψεις για τον χρήστη. Μια άλλη χρησιμότητα που έχει η αποθήκευση πληροφορίας για τα Συστήματα Προτάσεων με βάση το Περιεχόμενο είναι πως τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν ως ένα test εκπαίδευσης (training data) για τους αλγορίθμους μηχανικής μάθησης που δημιουργούν ένα μοντέλο.

Υπάρχουν πολλοί αλγόριθμοι οι οποίοι χρησιμοποιούνται ως διαφορετικές προσεγγίσεις για την εκμάθηση και την δημιουργία ενός μοντέλου για το profile του χρήστη. Φυσικά, υπάρχει και η προσέγγιση να δώσει χειρωνακτικά και ο ίδιος ο χρήστης πληροφορίες στο σύστημα και να προσαρμόσει έτσι το profile του. Για παράδειγμα, κάποια Συστήματα Προτάσεων παρέχουν μια διεπαφή που επιτρέπει στον χρήστη να δηλώσει τις προτιμήσεις και τις ανάγκες του, επιλέγοντας από λίστες με τιμές γνωρισμάτων ποιες ταιριάζουν στα γούστα του, πληκτρολογώντας ο ίδιος προϊόντα που βρίσκει ενδιαφέροντα ή μαρκάροντας προϊόντα ή ιστοσελίδες ως «Αγαπημένα». Όταν ο χρήστης δηλώσει τέτοιου είδους πληροφορίες, αποθηκεύονται σε μια βάση δεδομένων και στη συνέχεια εφαρμόζεται διαδικασία που βρίσκει και παρουσιάζει προϊόντα που καλύπτουν τα κριτήρια που έχει θέσει ο χρήστης.

Η δημιουργία ενός μοντέλου των προτιμήσεων του χρήστη αποτελεί μια μορφή ταξινόμησης. Ένα σύστημα που μαθαίνει και εφαρμόζει ταξινόμηση, χωρίζει τα δεδομένα (training data) σε δύο κατηγορίες, στα προϊόντα που αρέσουν στον χρήστη και σε εκείνα που δεν του αρέσουν , και το σύστημα ταξινομεί το κάθε προϊόν στην αντίστοιχη κατηγορία. Αυτό επιτυγχάνεται είτε με άμεση είτε με έμμεση ανατροφοδότηση του χρήστη. Στη άμεση ανατροφοδότηση (explicit feedback) ο

χρήστης βαθμολογεί τα αντικείμενα. Η παροχή άμεσων δεδομένων μπορεί να είναι μεν μια απλή διαδικασία, αλλά οι χρήστες τείνουν να παρέχουν άμεση πληροφορία μόνο σε ένα μικρό κομμάτι των προϊόντων με τα οποία έρχονται σε επαφή. Επίσης, κάποιες φορές η βαθμολόγηση δεν είναι αρκετή για να καταλάβει το σύστημα την στάση και τις απόψεις των χρηστών για τα προϊόντα. Στην έμμεση ανατροφοδότηση (implicit feedback) το σύστημα παρακολουθεί την αλληλεπίδραση του χρήστη με τα προϊόντα, όπως το «σώσιμο» , η εκτύπωση ή η τοποθέτηση σελιδοδεικτών (bookmarking) σε αυτά, από την οποία συνεπάγεται έμμεσα και η αρέσκεια του χρήστη. Αν για παράδειγμα ένας χρήστης αγοράσει κάτι, αυτό σημαίνει ότι του αρέσει αυτό το προϊόν, ενώ αν το επιστρέψει αυτό δηλώνει έμμεσα ότι δεν του αρέσει. Το σύστημα μπορεί να συλλέξει πολλή πληροφορία από την έμμεση τροφοδότηση. Η μέθοδος ανάλυσης της έμμεσης πληροφορίας έχει το πλεονέκτημα ότι ο χρήστης δεν χρειάζεται να εμπλακεί σε αυτή, αλλά από την άλλη πλευρά έχει το μειονέκτημα του στοιχείου της αβεβαιότητας, καθώς δεν μπορεί από κάποιες πράξεις του χρήστη να εννοηθεί ξεκάθαρα αν ο χρήστης ενδιαφέρεται για το προϊόν ή όχι . [13]

Πίνακας 2 : Χαρακτηριστικά άμεσης και έμμεσης ανατροφοδότησης πληροφορίας

	ΑΜΕΣΗ ΑΝΑΤΡΟΦΟΔΟΤΗΣΗ	ΕΜΜΕΣΗ ΑΝΑΤΡΟΦΟΔΟΤΗΣΗ
Ακρίβεια	Υψηλή	Χαμηλή
Ποσότητα πληροφορίας	Χαμηλή	Υψηλή
Επιρροή από συμφραζόμενα	Ναι	Ναι
Έκφραση προτίμησης	Θετική και Αρνητική	Θετική
Είδος αναφοράς	Απόλυτη	Σχετική

Στις επόμενες ενότητες θα αναλυθούν κάποιοι από τους πιο δημοφιλείς αλγορίθμους εκμάθησης του μοντέλου δημιουργίας profile χρηστών στα Συστήματα Προτάσεων με βάση το Περιεχόμενο, χρησιμοποιώντας πληροφορίες από τις πηγές [3] και [12]. Οι αλγόριθμοι αυτοί έχουν την δυνατότητα να «μαθαίνουν» μια συνάρτηση που μοντελοποιεί τις προτιμήσεις του χρήστη. Απαιτούν να έχουν βαθμολογήσει οι χρήστες με scores κάποια έγγραφα και αυτόματα συμπεραίνουν το profile του στο οποίο στηρίζεται το φιλτράρισμα των προϊόντων για την κατάταξή τους στην τελική σύσταση ξεκινώντας από εκείνα που θα ενδιαφέρουν τον χρήστη περισσότερο.

4.2.1 Πιθανοτική μέθοδος Naïve Bayes

Η μέθοδος Naïve Bayes είναι μια πιθανολογική προσέγγιση της επαγωγικής μάθησης και ανήκει στην γενικότερη κατηγορία των Bayesian ταξινομητών. Οι Bayesian ταξινομητές χρησιμοποιούνται για να λύνουν προβλήματα ταξινόμησης. Στηρίζονται στην θεωρία των πιθανοτήτων και στο θεώρημα του Bayes. Η Bayesian στατιστική χρησιμοποιεί την πιθανότητα για να παρουσιάσει την αβεβαιότητα στις σχέσεις που αντλήθηκαν από τα δεδομένα. Επίσης η έννοια του «προγενέστερου» είναι πολύ σημαντική, καθώς αντιπροσωπεύει την εκ των προτέρων γνώση μας για το ποια μπορεί να είναι η πραγματική σχέση. Οι προκύπτουσες εκ των υστέρων πιθανότητες είναι ανάλογες με τις προκύπτουσες εκ των προτέρων πιθανότητες. Οι Bayesian ταξινομητές θεωρούν κάθε χαρακτηριστικό γνώρισμα και κατηγορία σαν μια τυχαία μεταβλητή. Το μοντέλο λοιπόν εκτιμά τις προκύπτουσες εκ των υστέρων πιθανότητες (a posteriori probability) P(c/d) του εγγράφου d που ανήκει στην κατηγορία c, στηριζόμενο στην εκ των προτέρων (a priori) πιθανότητα P(c) παρατήρησης κάποιου έγγραφου στην κατηγορία c, στην πιθανότητα P(d/c)

παρατήρησης του εγγράφου d δεδομένης της κατηγορίας c και στην πιθανότητα P(d) παρατήρησης του d. Υπολογίζεται λοιπόν η Δεσμευμένη Πιθανότητα με βάση το Θεώρημα Bayes ως εξής:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

[3, σελ.91]

Στόχος του ταξινομητή είναι να προβλέψει την κατηγορία c στην οποία ανήκει το έγγραφο d, βρίσκοντας την τιμή της c που μεγιστοποιεί την πιθανότητα P(c/d):

$$c = argmax_{c_j} \frac{P(c_j)P(d|c_j)}{P(d)}$$

[3, σελ.91]

Για να υπολογίσει την δεσμευμένη πιθανότητα ο ταξινομητής Naïve Bayes, θεωρεί ότι όλα τα γνωρίσματα του εγγράφου d, δηλαδή οι λέξεις ή τα «σημεία» (tokens) είναι μεταξύ τους ανεξάρτητα.

Σημαντικά προτερήματα των ταξινομητών αυτών είναι ότι είναι αρκετά ισχυροί ώστε να απομονώνουν άσχετα χαρακτηριστικά ή σημεία θορύβου και ότι χειρίζονται τιμές που λείπουν αγνοώντας το στιγμιότυπο κατά τον υπολογισμό των πιθανοτήτων.

Η δημοτικότητα του αλγορίθμου και η απόδοσή του σε εφαρμογές κατηγοριοποίησης κειμένου έχει οδηγήσει στην ανάλυση και σύγκριση των διαφόρων μοντέλων του Naïve Bayes ταξινομητή, με επικρατέστερα το μοντέλο πολυμεταβλητών Bernoulli και το πολυωνυμικό μοντέλο. Και οι δύο μέθοδοι μοντελοποιούν ένα έγγραφο με τη μορφή ενός διανύσματος τιμών, όπου κάθε είσοδος στο διάνυσμα αντιπροσωπεύει αν μια λέξη εμφανίστηκε στο κείμενο ή όχι. Το μοντέλο πολυμεταβλητών Bernoulli προήλθε από την φιλοσοφία των δομημένων δεδομένων. Για την κατηγοριοποίηση κειμένων, θεωρεί ότι το κάθε έγγραφο εκπροσωπείται από ένα δυαδικό διάνυσμα στο πλαίσιο όλων των λέξεων ενός λεξιλογίου V. Το κάθε στοιχείο B_{it} του διανύσματος δείχνει αν μια λέξη w_t εμφανίζεται τουλάχιστον μια φορά στο έγγραφο. Λαμβάνοντας υπ' όψιν την Naïve Bayes υπόθεση ότι η πιθανότητα για κάθε λέξη να εμφανίζεται σε ένα έγγραφο είναι ανεξάρτητη από τις άλλες λέξεις δεδομένης της κατηγορίας, η δεσμευμένη πιθανότητα $P(d_i/c_j; \theta)$ ορίζεται ως :

$$P(d_i \mid c_j;\theta) = \prod_{t=1}^{|V|} (B_{it}P(w_t \mid c_j;\theta) + (1 - B_{it})(1 - P(w_t \mid c_j;\theta)))$$

[12, σελ.14]

όπου η πιθανότητα $P(w_t/c_j; \theta)$ μπορεί να οριστεί από την εμφάνιση της λέξης πάνω στα δεδομένα ως :

$$P(w_t \mid c_j;\theta) = \frac{1 + \sum_{i=1}^{|D|} B_{it}P(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} P(c_j \mid d_i)}$$

[12, σελ.14]

Σε αντίθεση με την δυαδική αναπαράσταση διανύσματος του μοντέλου πολυμεταβλητών Bernoulli, το πολυωνυμικό μοντέλο χρησιμοποιεί την πληροφορία της συχνότητας των λέξεων. Σε αυτήν την προσέγγιση, θεωρείται ότι τα έγγραφα δημιουργούνται από μια ακολουθία ανεξάρτητων μεταξύ τους δοκιμών που λαμβάνεται από μια πολυωνυμική κατανομή πιθανοτήτων. Δεδομένης και πάλι

της υπόθεσης της Naïve Bayes ανεξαρτησίας, η πιθανότητα P(di/cj; θ) υπολογίζεται με βάση τις ανεξάρτητες πιθανότητες των λέξεων ως :

$$P(d_i \mid c_j;\theta) = P(|d_i|)\prod_{t=1}^{|d_i|} P(w_t \mid c_j;\theta)^{N_{it}}$$

[12, σελ.14]

όπου Nit είναι ο αριθμός που εμφανίζεται η λέξη wt στο έγγραφο dt και η πιθανότητα P(wt/cj; θ) μπορεί να προκύψει από τα training data ως :

$$P(w_t \mid c_j;\theta) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N_{is} P(c_j \mid d_i)}$$

[12, σελ.14]

Η εμπειρία έχει δείξει ότι το πολυωνυμικό Naïve Bayes μοντέλο έχει καλύτερες επιδόσεις σε σχέση με την τεχνική πολυμεταβλητών Bernoulli, και ειδικά σε εφαρμογές σε μεγάλα λεξιλόγια. Ωστόσο, παρουσιάζει αδυναμίες όταν τα έγγραφα στο training set έχουν διαφορετικά μεγέθη και όταν οι κατηγορίες είναι λίγες γιατί δεν υπάρχουν αρκετά training δεδομένα. Οι περιπτώσεις αυτές εμφανίζονται συχνά στην διαδικασία δημιουργίας profile χρηστών, όπου το μέγεθος των training δεδομένων δεν είναι συγκεκριμένα και πολλές φορές δεν υπάρχουν αρκετά δεδομένα για προϊόντα που δεν αρέσουν στον χρήστη, σε σύγκριση με τα δεδομένα για προϊόντα που του αρέσουν.

Ο ταξινομητής Naïve Bayes εφαρμόζεται σε πολλά συστήματα προτάσεων με βάση το περιεχόμενο, όπως τα Syskill&Webert, LIBRA και Daily Learner.

4.2.2 Γραμμικοί Ταξινομητές

Οι Γραμμικοί ταξινομητές είναι αλγόριθμοι που χρησιμοποιούν γραμμικά όρια αποφάσεων, όπως γραμμικά όρια για να διαχωρίσουν τις περιπτώσεις σε έναν πολυδιάστατο χώρο, και εφαρμόζονται ευρέως σε κατηγοριοποιήσεις κειμένων. Όλοι οι γραμμικοί ταξινομητές λειτουργούν στα πλαίσια μιας κοινής φιλοσοφίας. Η διαδικασία μάθησης του αλγορίθμου μοντελοποιείται με ένα ν-διάστατο διάνυσμα βαρών w, του οποίου το εσωτερικό γινόμενο με ένα στιγμιότυπο, όπως για παράδειγμα ένα έγγραφο κειμένου που εκπροσωπείται από το μοντέλο του Διανυσματικού Χώρου (VSM) , δίνει ως αποτέλεσμα μια αριθμητική πρόβλεψη. Η αριθμητική αυτή πρόβλεψη οδηγεί σε μια προσέγγιση γραμμικής παλινδρόμησης. Κάποιες φορές ωστόσο μπορεί να χρησιμοποιηθεί ένα όριο ώστε οι συνεχείς προβλέψεις να μετατραπούν σε διακριτές κατηγορίες. Αυτό το γενικό πλαίσιο λειτουργίας ισχύει για όλους τους γραμμικούς ταξινομητές. Οι διαφοροποιήσεις εμφανίζονται στις μεθόδους εκπαίδευσης των αλγορίθμων, που χρησιμοποιούνται για να υπολογίζουν το διάνυσμα βαρών w. Σύμφωνα με τον Widrow-Hoff κανόνα ή κανόνα Δέλτα ([12]), το διάνυσμα βαρών w προκύπτει από στοιχειώδεις μεταφορές του φορέα στην κατεύθυνση της αρνητικής κλίσης του τετραγωνικού σφάλματος, καθώς προς αυτήν την κατεύθυνση το σφάλμα μειώνεται γρηγορότερα και υπολογίζεται ως:

$$w_{i+1,j} = w_{i,j} - 2\eta(w_i \cdot x_i - y_i)x_{i,j}$$

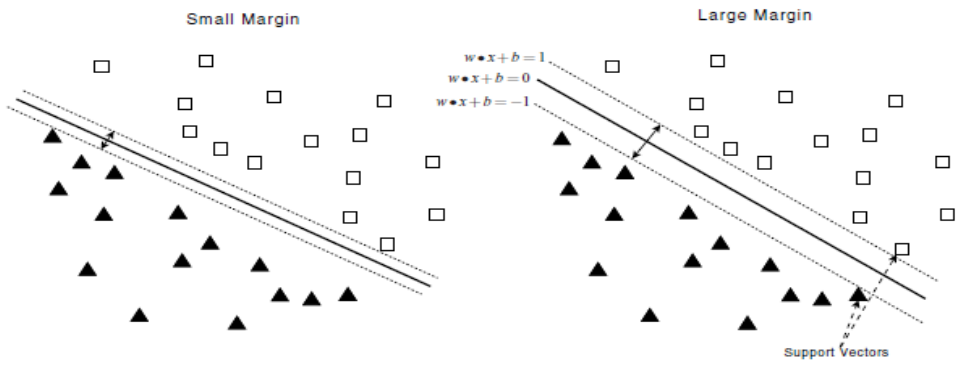
[12, σελ.12]

Σύμφωνα με την παραπάνω εξίσωση, το διάνυσμα βαρών μπορεί να προκύψει σταδιακά. Η αριθμητική πρόβλεψη που υπολογίζει ο αλγόριθμος είναι το εσωτερικό γινόμενο του εκάστοτε στιγμιοτύπου x_i και του διανύσματος βαρών w_i . Το σφάλμα της πρόβλεψης προσδιορίζεται αφαιρώντας από την αριθμητική τιμή της πρόβλεψης την γνωστή τιμή y_i του στιγμιότυπου. Το σφάλμα που προκύπτει από αυτήν την διαδικασία πολλαπλασιάζεται στη συνέχεια με τον αρχικό φορέα στιγμιότυπων x_i και τον ρυθμό μάθησης η. Με αυτόν τον τρόπο σχηματίζεται ένα διάνυσμα, το οποίο όταν αφαιρεθεί από το διάνυσμα βαρών w, μεταφέρει το w όλο και πιο κοντά στην σωστή πρόβλεψη για το στιγμιότυπο x_i. Ο δε ρυθμός μάθησης η ελέγχει τον βαθμό στον οποίο το κάθε επιπλέον στιγμιότυπο επηρεάζει το προηγούμενο διάνυσμα βαρών.

Μια άλλη μέθοδος εκπαίδευσης, η οποία αποδίδει αποτελεσματικότερα στην κατηγοριοποίηση εγγράφων κειμένου με πολλά χαρακτηριστικά γνωρίσματα είναι ο αλγόριθμος της εκθετικής κλίσης (EG algorithm). Το σφάλμα του EG αλγορίθμου εξαρτάται λογαριθμικά από τον αριθμό των χαρακτηριστικών γνωρισμάτων. Αυτό το γεγονός πιστοποιεί θεωρητικά την υψηλή αποδοτικότητα του αλγορίθμου σε προβλήματα κατηγοριοποίησης κειμένων, τα οποία είναι κατά κανόνα μεγάλων διαστάσεων.

Ένα σημαντικό πλεονέκτημα των παραπάνω μεθόδων εκπαίδευσης των γραμμικών αλγορίθμων είναι ότι μπορούν να εφαρμοστούν διαδικτυακά on-line. Το πλεονέκτημα αυτό είναι υψίστης σημασίας για εφαρμογές που λειτουργούν σε πραγματικό χρόνο, καθώς για παράδειγμα το εκάστοτε διάνυσμα βαρών μπορεί να μετατρέπεται σταδιακά, ενώ όλο και περισσότερα στιγμιότυπα γίνονται διαθέσιμα και προστίθενται στην εφαρμογή.

Μια άλλη σημαντική παρατήρηση σχετικά με τους γραμμικούς ταξινομητές είναι ότι ενώ οι μέθοδοι που χρησιμοποιούν τείνουν να συγκλίνουν σε υπερεπίπεδα γραμμικά όρια που χωρίζουν σε κατηγορίες τα δεδομένα εκμάθησης με ακρίβεια, η γενική απόδοση αυτών των ορίων δεν είναι βέλτιστη. Μια προσέγγιση για την βελτίωση της απόδοσης είναι γνωστή και ως μηχανές διανυσμάτων υποστήριξης (support vector machines). Η κεντρική ιδέα στην οποία στηρίζονται είναι η μεγιστοποίηση του περιθωρίου κατηγοριοποίησης, δηλαδή της απόστασης μεταξύ του ορίου αποφάσεων και των κοντινότερων στιγμιότυπων εκμάθησης, ή αλλιώς διανυσμάτων υποστήριξης.



Εικόνα 5 : Διαφορετικές αποφάσεις για τα σύνορα μπορεί να οδηγήσει σε πρόβλημα διαχωρισμού των δεδομένων σε δυο κατηγορίες. Κάθε σύνορο έχει ένα σχετικό περιθώριο.[3]

Ο γραμμικός διαχωρισμός ανάμεσα σε δυο κατηγορίες πραγματοποιείται μέσω της συνάρτησης $w \bullet x + b = 0$. Η συνάρτηση αυτή κατηγοριοποιεί τα προϊόντα με βάση αν ανήκουν στην κατηγορία +1 ή -1, υπό την προϋπόθεση ότι είναι διαχωρισμένες από μια ελάχιστη απόσταση της συνάρτησης διαχωρισμού κατηγοριών. Η εξίσωση της συνάρτησης φαίνεται παρακάτω :

$$f(x) = \begin{cases} 1, & \text{if } w \bullet x + b \geq 1 \\ -1, & \text{if } w \bullet x + b \leq -1 \end{cases}$$

$$Margin = \frac{2}{\|w\|^2}$$

[3, σελ.56]

Θέλοντας, όπως αναφέρθηκε προηγουμένως, να μεγιστοποιηθεί το περιθώριο ανάμεσα σε δυο κατηγορίες, το οποίο δίνεται από την παραπάνω εξίσωση, γίνεται προσπάθεια ισοδύναμα να ελαχιστοποιηθεί το αντίστροφό του, L(w), υπό τους περιορισμούς που θέτει η f(x). Υπάρχουν πολλοί τρόποι για να λυθεί αυτό το πρόβλημα βελτιστοποίησης με περιορισμούς. Ένας τρόπος είναι η ελαχιστοποίηση με βάση την συνάρτηση L(w) που ορίζεται παρακάτω, υπό τους περιορισμούς μιας νέας εξίσωσης f(x) :

$$L(w) = \frac{\|w\|^2}{2} + C\sum_{i=1}^{N} \varepsilon$$

$$f(x) = \begin{cases} 1, & \text{if } w \bullet x + b \geq 1 - \varepsilon \\ -1, & \text{if } w \bullet x + b \leq -1 + \varepsilon \end{cases}$$

[3, σελ.57]

Εμπειρικά πειράματα έχουν δείξει ότι οι γραμμικές μηχανές διανυσμάτων υποστήριξης αποδίδουν πολύ αποτελεσματικά σε προβλήματα κατηγοριοποίησης κειμένων. Ο κύριος λόγος στον οποίο οφείλεται αυτό το γεγονός είναι ότι η μεγιστοποίηση του περιθωρίου αποτελεί έναν μηχανισμό προστασίας από το overfitting. Αυτή λοιπόν η μειωμένη τάση για overfitting στα δεδομένα μάθησης (training data) είναι εξαιρετικά χρήσιμη στους αλγορίθμους κατηγοριοποίησης κειμένων, καθώς σε αυτόν τον τομέα πρέπει να κατανοηθούν πολυδιάστατες έννοιες από περιορισμένο αριθμό εκπαιδευτικών δεδομένων, και αυτή η διαδικασία είναι επιρρεπής στο overfitting.

4.2.3 Ανατροφοδότηση σχετικότητας και Αλγόριθμος του Rocchio

Επειδή η επιτυχία της ανάκτησης πληροφορίας (Information Retrieval) στο μοντέλο του Διανυσματικού Χώρου σχετίζεται άμεσα με την δυνατότητα που παρέχεται στον χρήστη να κατασκευάζει ερωτήματα επιλέγοντας κάποιες χαρακτηριστικές λέξεις-κλειδιά, άρχισαν να γίνονται αντικείμενο μελέτης πολλές μέθοδοι που βοηθούν τους χρήστες να τελειοποιούν τα αρχικά ερωτήματά τους με βάση παλαιότερα αποτελέσματα αναζήτησης. Οι τεχνικές αυτές είναι γνωστές ως ανατροφοδότηση σχετικότητας (Relevance Feedback). Οι μέθοδοι επιτρέπουν στους χρήστες να απαντούν στις αποφάσεις του συστήματος γύρω από το πόσο σχετικά ήταν τα έγγραφα που ανέκτησε το σύστημα με τις πληροφοριακές ανάγκες του χρήστη. Η ανατροφοδότηση αυτή γίνεται με την βαθμολογία των χρηστών στα έγγραφα που επέστρεψε το Σύστημα Προτάσεων. Κατ' ανάλογο τρόπο με τις βαθμολογήσεις προϊόντων, υπάρχουν έμμεσοι και άμεσοι τρόποι συλλογής δεδομένων ανατροφοδότησης όσον αφορά τη σχετικότητα.

Ο αλγόριθμος του Rocchio είναι μια προσαρμογή της ανατροφοδότησης σχετικότητας στον τομέα της κατηγοριοποίησης κειμένου που εφαρμόζεται στο μοντέλο Διανυσματικού Χώρου. Ο αλγόριθμος αυτός στηρίζεται στην μετατροπή του αρχικού ερωτήματος μέσω διαφορετικά σταθμισμένων πρωτοτύπων συναφών και μη συναφών εγγράφων. Η ανατροφοδότηση των χρηστών με την βαθμολόγηση των εγγράφων που ανακτήθηκαν μπορεί να χρησιμοποιηθεί για να τελειοποιηθούν σταδιακά τα profiles των χρηστών ή για την εκπαίδευση των αλγορίθμων μάθησης, που χρησιμοποιούν τα profiles τω χρηστών ως ταξινομητές. Η μέθοδος του Rocchio εφαρμόζεται για να δημιουργήσει γραμμικούς, στηριζόμενους στα profile ταξινομητές. Ο αλγόριθμος αυτός παρουσιάζει τα έγγραφα σαν διανύσματα με τέτοιο τρόπο ώστε έγγραφα με παρόμοιο περιεχόμενο να έχουν και παρόμοια διανύσματα. Το κάθε στοιχείο του διανύσματος αντιστοιχεί σε ένα στοιχείο του εγγράφου, συνήθως δηλαδή σε μια λέξη. Το βάρος του κάθε στοιχείου του διανύσματος υπολογίζεται με την χρήση της μεθόδου υπολογισμού των βαρών «Συχνότητα Όρου-Αντίστροφη Συχνότητα Εγγράφου» (TF-IDF). Η μάθηση του αλγορίθμου επιτυγχάνεται με τον συνδυασμό των διανυσμάτων εγγράφων με ένα πρωτότυπο διάνυσμα της κάθε κατηγορίας. Για να ταξινομηθεί ένα νέο έγγραφο D, υπολογίζεται η ομοιότητα για κάθε κατηγορία μεταξύ των πρωτότυπων διανυσμάτων και του διανύσματος που εκπροσωπεί το έγγραφο και στη συνέχεια το έγγραφο D συγκαταλέγεται στην κατηγορία της οποίας το διάνυσμα εγγράφου έχει την υψηλότερη τιμή ομοιότητας. Τελικά δημιουργούνται δύο πρωτότυπα έγγραφα, παίρνοντας το άθροισμα των διανυσμάτων των σχετικών και των άσχετων εγγράφων. Ο παρακάτω τύπος παρουσιάζει μαθηματικοποιημένα τον αλγόριθμο του Rocchio :

$$Q_{i+1} = \alpha\, Q_i + \beta \sum_{rel} \frac{D_i}{|D_i|} - \gamma \sum_{nonrel} \frac{D_i}{|D_i|}$$

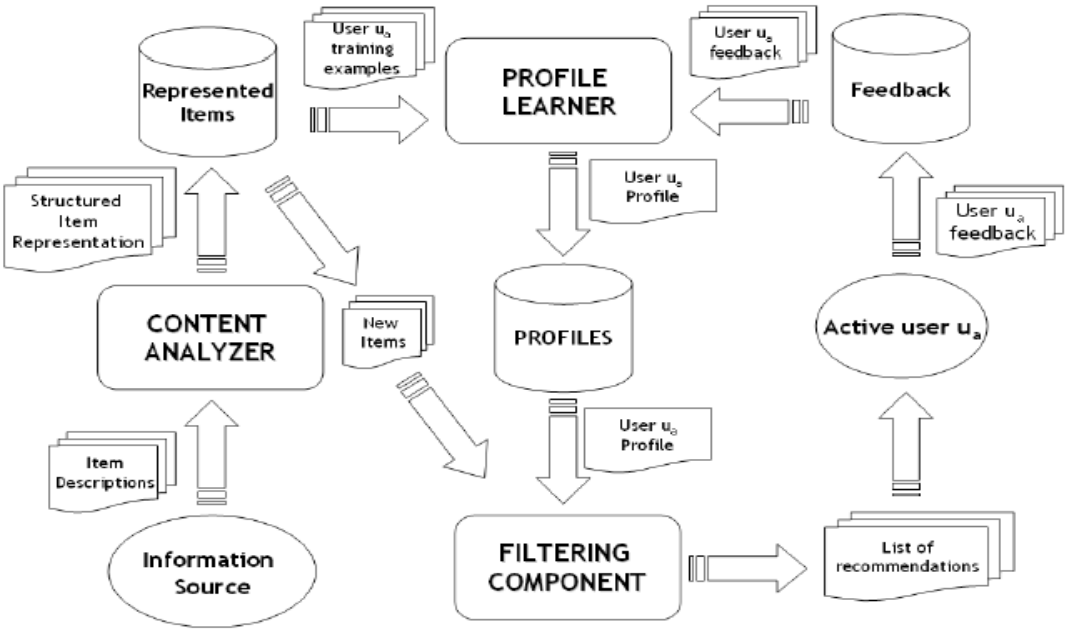
[12, σελ.11]

Το Q αντιπροσωπεύει το ερώτημα του χρήστη στην επανάληψη i, το D αντιπροσωπεύει το έγγραφο και τα α, β και γ είναι παράμετροι που ελέγχουν την επίδραση του αρχικού ερωτήματος και των δυο πρωτοτύπων στο τροποποιημένο ερώτημα που προκύπτει. Η φιλοσοφία που κρύβεται πίσω από την φόρμουλα της παραπάνω εξίσωσης είναι να μετακινηθεί σταδιακά το διάνυσμα του ερωτήματος προς αθροίσματα (clusters) σχετικών εγγράφων και μακριά από αθροίσματα άσχετων κειμένων. Εμπειρικά έχει φανεί ότι η μέθοδος αυτή οδηγεί σε βελτίωση της απόδοσης της διαδικασίας ανάκτησης πληροφορίας.

Ο αλγόριθμος του Rocchio και οι μέθοδοι ανατροφοδότησης σχετικότητας χρησιμοποιούνται συχνά σε Συστήματα Προτάσεων με βάση το περιεχόμενο όπως το YourNews. Πρόσφατα, πολλοί ερευνητές χρησιμοποίησαν και μια παραλλαγή του αλγορίθμου του Rocchio στα πλαίσια μηχανικής μάθησης, όπως για την μάθηση του profile ενός χρήστη από μη δομημένα δεδομένα. Στόχος αυτών των εφαρμογών είναι να δημιουργηθεί αυτόματα ένας ταξινομητής κειμένου που να μπορεί να διαχωρίζει τις κατηγορίες των κειμένων.

4.3 Αρχιτεκτονική των Συστημάτων Προτάσεων με βάση το Περιεχόμενο

Τα Συστήματα Ανάκτησης πληροφορίας και Σύστασης προϊόντων με βάση το περιεχόμενο δημιουργούν ένα μοντελοποιημένο profile για έναν χρήστη, το οποίο παρουσιάζει τα ενδιαφέροντά του, και στη συνέχεια αντιστοιχούν τα χαρακτηριστικά γνωρίσματα του profile με τα χαρακτηριστικά του περιεχομένου διαφόρων προϊόντων, για να βρουν ποιά ανταποκρίνονται στις ανάγκες του. Αυτά τα Συστήματα λοιπόν χρειάζεται να αναπτύξουν τεχνικές για να παρουσιάζουν τα προϊόντα, για να παράγουν το profile του χρήστη και στρατηγικές για να μπορούν να συγκρίνουν το profile του χρήστη με την παρουσίαση των προϊόντων. Η διαδικασία παραγωγής συστάσεων γίνεται σε τρία βήματα. Το κάθε βήμα αντιστοιχεί και σε ένα διαφορετικό στοιχείο της αρχιτεκτονικής των Συστημάτων με βάση το περιεχόμενο.



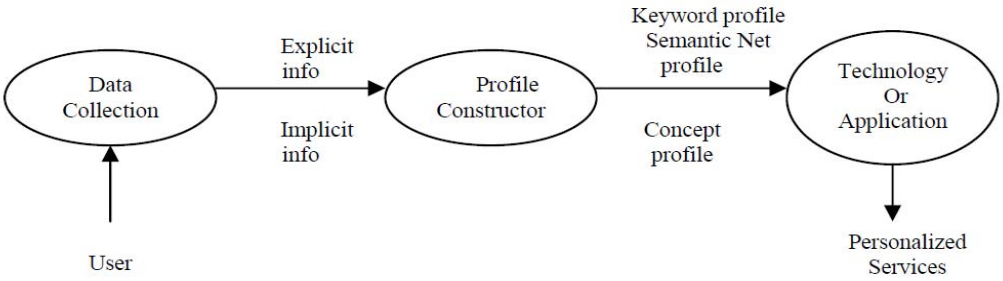
Εικόνα 6 : Τα στοιχεία της αρχιτεκτονικής ενός συστήματος προτάσεων με βάση το περιεχόμενο [3]

Τα κυρίως μέρη ενός τέτοιου Συστήματος είναι τα εξής :

- **Αναλυτής Περιεχομένου :** Η βασική λειτουργία του Αναλυτή είναι να φέρνει το περιεχόμενο των προϊόντων σε μορφή κατάλληλη ώστε να μπορούν να το χειριστούν τα επόμενα στάδια διεργασίας (π.χ. από αδόμητο κείμενο σε δομημένα σχετικά δεδομένα). Με συγκεκριμένες τεχνικές, αναλύεται το περιεχόμενο των προϊόντων και έτσι η παρουσίασή τους γίνεται κατάλληλη για τον χώρο στον οποίο θα χρησιμοποιηθούν, όπως για παράδειγμα ιστοσελίδες μετατρέπονται σε διανύσματα με λέξεις-κλειδιά.
- «**Μαθητής**» **Profile :** Το στοιχείο αυτό παίρνει σαν είσοδο την παρουσίαση των δεδομένων, όπως αυτή προέκυψε από τον Αναλυτή Περιεχομένου και στοχεύει στο να γενικεύσει τα δεδομένα με τα ενδιαφέροντα του χρήστη, έτσι ώστε να μπορέσει να παραχθεί στη συνέχεια το profile του χρήστη. Στο σημείο αυτό εφαρμόζονται αλγόριθμοι

μηχανικής μάθησης, όπως τεχνικές ανατροφοδότησης σχετικότητας που αναλύσαμε παραπάνω, ώστε να μοντελοποιηθούν οι προτιμήσεις του χρήστη με βάση προϊόντα για τα οποία είχε δείξει ενδιαφέρον ή άλλα που δεν του είχαν αρέσει στο παρελθόν.

- **Φιλτράρισμα των δεδομένων :** Αυτή η μονάδα της αρχιτεκτονικής κάνει χρήση του profile του χρήστη ώστε να συστήσει τελικά τα προιόντα εκείνα που τα χαρακτηριστικά περιεχομένου τους ταίριαξαν με τα χαρακτηριστικά γνωρίσματα του profile, μέσω υπολογισμού ομοιότητας. Τελικά προκύπτει ένας ταξινομημένος κατά σειρά προτίμησης κατάλογος προϊόντων που το σύστημα υπολόγισε ότι θα ενδιαφέρουν τον χρήστη.



Εικόνα 7 : Παρουσίαση εξατομικευμένου συστήματος με κατασκευή profile χρήστη [13]

4.4 Πλεονεκτήματα και Περιορισμοί των Συστημάτων Προτάσεων με βάση το Περιεχόμενο

Η χρήση ενός τέτοιου Συστήματος για την παραγωγή συστάσεων έχει πολλά πλεονεκτήματα. Κάποια από αυτά είναι :

- Η Διαφάνεια : Τα συστήματα αυτά χρησιμοποιούν πολύ συχνά εξηγήσεις σχετικά με το πώς προέκυψαν τα προϊόντα που προτείνει, όπως για παράδειγμα εμφανίζει μια λίστα με στοιχεία περιεχομένου ή περιγραφές που ώθησαν το σύστημα στο να συστήσει το συγκεκριμένο προϊόν. Η ύπαρξη εξηγήσεων ενισχύει την εμπιστοσύνη των χρηστών στο σύστημα και αυξάνει έτσι τις πιθανότητες να δεχτεί ο χρήστης την πρόταση που του έγινε και να αγοράσει τελικά το προϊόν.
- Η ανεξαρτησία των χρηστών : Τα Συστήματα Προτάσεων με βάση το περιεχόμενο εκμεταλλεύονται αξιολογήσεις και βαθμολογήσεις του ίδιου του χρήστη για να χτίσουν το profile του και δεν στηρίζονται σε ανατροφοδότηση από άλλους χρήστες
- Η δυνατότητα συστάσεων και νέων προϊόντων : Τα Content-based Recommender Systems μπορούν να προτείνουν νέα προϊόντα στο σύστημα που δεν έχουν ακόμα αξιολογηθεί από κανέναν χρήστη, καθώς συλλέγουν πληροφορία από την περιγραφή του προϊόντος που είναι εξ-αρχής δεδομένη.

Ωστόσο, υπάρχουν κάποιοι περιορισμοί που δεν επιτρέπουν στα συστήματα αυτά να επεκταθούν, όπως οι ακόλουθοι :

- Η περιορισμένη ανάλυση περιεχομένου : Υπάρχει ένα όριο στον αριθμό των χαρακτηριστικών γνωρισμάτων που μπορεί ένα τέτοιο σύστημα να συγκρίνει για να παραγάγει συστάσεις προϊόντων. Ένα σύστημα προτάσεων με βάση το περιεχόμενο δεν μπορεί να κάνει καλές συστάσεις αν το περιεχόμενο δεν περιέχει αρκετή πληροφορία για να μπορέσει να διαχωρίσει τα προϊόντα που στον χρήστη αρέσουν από εκείνα που δεν του αρέσουν, και έτσι συχνά απαιτείται επιπλέον γνώση του τομέα των προϊόντων. Έτσι, η αυτόματη και η χειρωνακτική δήλωση των γνωρισμάτων μπορεί να μην είναι αρκετή για

να γίνει ο διαχωρισμός των προϊόντων, ο οποίος είναι απαραίτητος για την εξαγωγή του profile με τα ενδιαφέροντα του χρήστη

- Πρόβλημα με νέους χρήστες : Για να είναι το σύστημα ικανό να καταλάβει την γενική συμπεριφορά στις προτιμήσεις του χρήστη και να κάνει σωστές και ακριβείς συστάσεις, θα πρέπει να έχει συλλέξει αρκετά δεδομένα από βαθμολογήσεις προϊόντων του χρήστη. Έτσι, το σύστημα δεν μπορεί να κάνει αξιόπιστες προτάσεις σε έναν καινούριο χρήστη που δεν έχει αλληλεπιδράσει ακόμα αρκετά με το σύστημα.
- Η μεγάλη εξειδίκευση : Τα συστήματα με βάση το περιεχόμενο προτείνουν προϊόντα που είχαν υψηλό βαθμό αντιστοίχησης με τα γνωρίσματα του profile του χρήστη. Αυτό σημαίνει ότι το σύστημα θα προτείνει στον χρήστη προϊόντα που είναι παρόμοια με εκείνα που είχε βαθμολογήσει και του είχαν αρέσει στο παρελθόν, δηλαδή δεν έχει τη δυνατότητα το σύστημα να βρει νέα προϊόντα που δεν έχουν σχέση με όσα έχει κοιτάξει παλαιότερα ο χρήστης αλλά που πιθανώς να τον ενδιαφέρουν. Το πρόβλημα αυτό είναι γνωστό ως Serendipity. Έτσι, αν ένας χρήστης έχει βαθμολογήσει ταινίες στις οποίες παίζει ένας συγκεκριμένος ηθοποιός, το σύστημα θα του βρει και άλλες ταινίες με τον ίδιο ηθοποιό που δεν έχει ακόμα βαθμολογήσει, λείπει λοιπόν το στοιχείο της καινοτομίας. Το πρόβλημα αυτό θα μπορούσε να λυθεί με την εισαγωγή του στοιχείου της τυχαιότητας. Επιπλέον, καθώς το σύστημα διαθέτει μεθόδους για να προτείνει μόνο όμοια μεταξύ τους προϊόντα, μερικές φορές τα προϊόντα σχεδόν ταυτίζονται και έτσι η σύσταση χάνει το νόημά της. Κάποια συστήματα προτάσεων έχουν αναπτύξει τεχνικές για να φιλτράρουν και να αποκλείουν προϊόντα που είναι πολύ όμοια με εκείνα που ο χρήστης έχει δει στο παρελθόν. Κάποια άλλα μέτρα επίσης εφαρμόζονται για να αξιολογούν κατά πόσο σχετικά προϊόντα διαθέτουν και κάποιον νεωτερισμό. Με την ποικιλία λοιπόν στις συστάσεις προσπαθεί να λυθεί το πρόβλημα του serendipity.

4.5 Web 2.0 και Εξατομικευμένα Συστήματα Προτάσεων με βάση το Περιεχόμενο

Η φιλοσοφία πίσω από τον ιστό 2.0 είναι ότι ο χρήστης πλέον βρίσκεται στο κέντρο του ενδιαφέροντος. Οι πλατφόρμες των Κοινωνικών Μέσων δεν έχουν νόημα χωρίς τους χρήστες, καθώς εκείνοι συμβάλουν στο περιεχόμενό τους με πληροφορία. Μια μορφή περιεχομένου που παράγεται από χρήστες είναι και τα folksonomies. Folksonomy είναι μια ταξινόμηση που προκύπτει από τους χρήστες που συνεργατικά σχολιάζουν και κατηγοριοποιούν τα ενδιαφέροντά τους με την χρήση λέξεων-κλειδιών που λέγονται ετικέτες (tags). Τα όλο και αυξανόμενα Folksonomies θέτουν νέες προκλήσεις στην αναζήτηση και ανάκτηση σχετικού περιεχομένου. Ιδανικά, μια Web 2.0 πλατφόρμα πρέπει να παρέχει στους χρήστες την δυνατότητα προσαρμοστικής περιήγησης και να κάνει συστάσεις σχετικού περιεχομένου. Η λειτουργία αυτή ξεπερνάει την απλή αντιστοίχηση προϊόντων μέσω ερωτημάτων με βάση τις λέξεις-κλειδιά και θέτει ένα νέο επίπεδο εξερεύνησης των υπηρεσιών του σύγχρονου διαδικτύου. Ένα σημαντικό κομμάτι έρευνας λοιπόν είναι η εξέλιξη κατάλληλων μεθόδων συστάσεων, καθώς η ενσωμάτωση ετικετών στους αλγορίθμους των συστημάτων προτάσεων με βάση το περιεχόμενο είναι ένα πρόβλημα που δεν έχει ακόμα εξερευνηθεί αναλυτικά.[3]

Έχουν προταθεί κάποιες τεχνικές για να λαμβάνεται υπ' όψιν η δραστηριότητα ετικετοποίησης των χρηστών στα content-based συστήματα προτάσεων [14]. Για παράδειγμα, τα profiles των χρηστών παρουσιάζονται με την μορφή διανύσματος ετικετών, στο οποίο το κάθε στοιχείο υποδηλώνει πόσες φορές μια ετικέτα έχει αποδοθεί σε κάποιο έγγραφο από τον χρήστη. Μια άλλη

τεχνική είναι η εκπροσώπηση των profiles των χρηστών από μια συλλογή ετικετών οι οποίες έχουν επιλεγεί από τον χρήστη, σε συνδυασμό με αντίστοιχες τιμές που αντιπροσωπεύουν το ενδιαφέρον του χρήστη για αυτές τις ετικέτες.

Επίσης, περαιτέρω έρευνα θα μπορούσε να διεξαχθεί σχετικά με την ανάλυση των ετικετών ως ενός δυναμικού είδους ανατροφοδότησης για την κατασκευή profile των χρηστών, καθώς ετικέτες που εκφράζουν την γνώμη ή τα συναισθήματα του χρήστη θα μπορούσαν να χρησιμοποιηθούν ως μέτρο ένδειξης της ικανοποίησης του χρήστη για τα προϊόντα.

5

Συστήματα Προτάσεων με Συνεργατικό Φιλτράρισμα

Το Συνεργατικό Φιλτράρισμα είναι η διαδικασία φιλτραρίσματος και αξιολόγησης προϊόντων μέσω της γνώμης άλλων ανθρώπων του δικτύου. Ο όρος Συνεργατικό Φιλτράρισμα προέκυψε μόλις την τελευταία δεκαετία, όμως οι ρίζες της έννοιας αυτής ανήκουν σε μια διαδικασία που οι άνθρωποι κάνουν εδώ και αιώνες, να μοιράζονται μεταξύ τους τις γνώμες τους. Πρόκειται για μια μέθοδο που στηρίζεται όχι μόνο στις παλαιότερες βαθμολογήσεις του χρήστη-στόχου, αλλά και σε εκείνες των άλλων χρηστών του συστήματος για να παραγάγει αυτόματα προβλέψεις για το ενδιαφέρον του χρήστη-στόχου για τα προϊόντα του συστήματος. Η κεντρική ιδέα είναι ότι η βαθμολόγηση ενός χρήστη u για ένα προϊόν i που δεν έχει δει θα είναι όμοια με την βαθμολόγηση που έδωσε στο προϊόν ένας άλλος χρήστης του συστήματος v, δεδομένου ότι οι χρήστες u και v έχουν βαθμολογήσει άλλα προϊόντα του συστήματος με παρόμοιο τρόπο. Με το ίδιο σκεπτικό, ο χρήστης u πιθανότατα θα δώσει παρόμοια βαθμολογία για δυο προϊόντα i και j, αν οι άλλοι χρήστες του συστήματος έχουν δώσει παρόμοιες βαθμολογίες στα δυο αυτά προϊόντα. Οι μέθοδοι Συνεργατικού Φιλτραρίσματος παράγουν εξατομικευμένες συστάσεις προϊόντων στους χρήστες από τις αλληλεπιδράσεις του με το σύστημα και τις βαθμολογίες του, χωρίς να απαιτεί εξωγενείς πληροφορίες ούτε για τους χρήστες ούτε για τα προϊόντα, όπως περιεχόμενο ή περιγραφές.

Οι βαθμολογήσεις που χρησιμοποιούν τα συστήματα συνεργατικού φιλτραρίσματος μπορεί να έχουν διάφορες μορφές. Μια μορφή είναι η βαθμωτή βαθμολογία που αποτελείται είτε από αριθμητικά δεδομένα, όπως η βαθμολόγηση με 1-5 αστέρια στη MovieLens ή η βαθμολογία τάξεως, όπως οι επιλογές « συμφωνώ απόλυτα, συμφωνώ, είμαι ουδέτερος, διαφωνώ, διαφωνώ απόλυτα». Ένα άλλο είδος «απάντησης» των χρηστών στα προϊόντα είναι το μοντέλο δυαδικής βαθμολογίας, δηλαδή «μου αρέσει / δεν μου αρέσει» ή «συμφωνώ / διαφωνώ». Τέλος, υπάρχουν και οι μοναδιαίες αξιολογήσεις, οι οποίες μπορούν να υποδηλώσουν αν ο χρήστης έχει παρατηρήσει, βαθμολογήσει θετικά ή αγοράσει κάποιο προϊόν. Η έλλειψη βαθμολογίας δείχνει ότι δεν υπάρχει πληροφορία σχετικά με την γνώμη του χρήστη για το προϊόν.

Τα βήματα που ακολουθούνται σε ένα Σύστημα Συνεργατικού Φιλτραρίσματος για την παραγωγή συστάσεων είναι τα εξής :

- Στάθμισε όλους τους χρήστες σε σχέση με την ομοιότητά τους στην συμπεριφορά με τον χρήστη-στόχο
- Διάλεξε μια υπο-ομάδα των χρηστών με τα υψηλότερα βάρη που προέκυψαν από το πρώτο βήμα και χρησιμοποίησέ τους σαν αναφορά (χρήστες-γείτονες)
- Κάνε την πρόβλεψη για το ενδιαφέρον του χρήστη-στόχου για κάποιο προϊόν

Οι ερευνητικές μελέτες γύρω από το Συνεργατικό Φιλτράρισμα αυξήθηκαν κατακόρυφα όταν τον Οκτώβριο του 2006 η Netflix διακήρυξε διαγωνισμό με έπαθλο ένα εκατομμύριο δολάρια με στόχο την σημαντική βελτίωση στην ακρίβεια των προβλέψεων για το πόσο θα αρέσει σε κάποιον χρήστη μια ταινία με βάση τις προτιμήσεις του κατά 10% [15]. Ο διαγωνισμός αυτός άνοιξε τον δρόμο για μεγάλη πρόοδο στον τομέα του συνεργατικού φιλτραρίσματος. Για πρώτη φορά, οι ερευνητές είχαν πρόσβαση σε δεδομένα εκατό εκατομμυρίων βαθμολογιών ταινιών. Ερευνητές από όλους τους τομείς, όπως μηχανικοί, μαθητές ή επιστήμονες χρησιμοποίησαν την καινοτομία στον τρόπο σκέψης τους για να αναπτύξουν αλγορίθμους που θα βελτιώσουν την ακρίβεια στην πρόβλεψη, μειώνοντας το σφάλμα. Η Netflix βράβευσε το 2007με 50.000 δολάρια την ομάδα BellKor, η οποία βελτίωσε το ποσοστό σφάλματος στις προβλέψεις κατά 8.43%. Η βελτίωση αυτή έγινε σε μια ομάδα δεδομένων που χρησιμοποιήθηκαν ως test. Το 2009 η ομάδα « BellKor's Pragmatic Chaos» κέρδισε το έπαθλο των ενός εκατομμυρίου δολαρίων, καταφέρνοντας βελτίωση στην πρόβλεψη κατά 10,06%.

Εικόνα 8 : Η MovieLens χρησιμοποιεί συνεργατικό φιλτράρισμα για να προβλέψει τις βαθμολογίες των χρηστών σε ταινίες [16]

5.1 Κατηγορίες Συστημάτων Προτάσεων με Συνεργατικό Φιλτράρισμα

Τα Συστήματα Προτάσεων με Συνεργατικό Φιλτράρισμα μπορούν να κατηγοριοποιηθούν βάσει διαφόρων χαρακτηριστικών. Στις επόμενες ενότητες διαχωρίζονται πρώτα τα συστήματα σε δύο ομάδες, ανάλογα με τον τρόπο που χρησιμοποιούνται οι βαθμολογήσεις χρηστών-προϊόντων που είναι αποθηκευμένες στο σύστημα, και στη συνέχεια χωρίζονται σε δύο κατηγορίες με βάση αν το σύστημα χρησιμοποιεί για να εξάγει πρόβλεψη για τον χρήστη-στόχο τις βαθμολογήσεις άλλων χρηστών ή τις βαθμολογήσεις του χρήστη-στόχου για άλλα προϊόντα.

5.1.1 Συστήματα Προτάσεων με Συνεργατικό Φιλτράρισμα με βάση το Μοντέλο και με βάση την Μνήμη

Οι μέθοδοι του Συνεργατικού Φιλτραρίσματος μπορούν να χωριστούν σε δυο γενικές κατηγορίες, στις μεθόδους που βασίζονται στο μοντέλο (model-based) και στις μεθόδους που στηρίζονται στην μνήμη (memory-based ή neighborhood-based) [3].

Τα model-based συστήματα χρησιμοποιούν τις βαθμολογήσεις των χρηστών που υπάρχουν αποθηκευμένες στο σύστημα για να μάθουν ένα μοντέλο προβλέψεων. Η γενική ιδέα είναι να μοντελοποιηθούν οι αλληλεπιδράσεις χρηστών-προϊόντων με παράγοντες που εκπροσωπούν τα αφανή χαρακτηριστικά των χρηστών και των προϊόντων στο σύστημα, όπως η κατηγορία προτιμήσεων του χρήστη ή η κλάση κατηγορίας στην οποία ανήκει το προϊόν. Το μοντέλο αυτό εκπαιδεύεται χρησιμοποιώντας τα διαθέσιμα δεδομένα και στη συνέχεια εφαρμόζεται για να προβλέψει τις βαθμολογήσεις των χρηστών σε καινούρια προϊόντα. Οι προσεγγίσεις με βάση το μοντέλο για την σύσταση προϊόντων είναι πολυάριθμες. Χαρακτηριστικά, κάποιες από αυτές τις προσεγγίσεις είναι τα Bayesian αθροίσματα (Bayesian Clustering), η κρυφή σημασιολογική ανάλυση (Latent Semantic Analysis), η μέγιστη εντροπία, οι μηχανές Boltzmann, οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) ή η αποσύνθεση μοναδικής τιμής (Singular Value Decomposition).

Πλεονεκτήματα αυτής της προσέγγισης είναι ότι μπορεί να χειρίζεται αραιά δεδομένα (πρόβλημα sparsity) χωρίς να μειώνεται η αποδοτικότητα. Έτσι, τα model-based συστήματα μπορούν να εφαρμόζονται εύκολα σε εφαρμογές που περιλαμβάνουν μεγάλα datasets. Επιπλέον, με την εκπαίδευση του συστήματος, μπορεί να βελτιώνεται συνεχώς η απόδοσή του. Ένα άλλο πλεονέκτημα είναι ότι μπορούν να δίνουν μια διαισθητική εξήγηση των συστάσεων που κάνουν, και έτσι οι χρήστες εμπιστεύονται καλύτερα το σύστημα και αποδέχονται τις προτάσεις που τους γίνονται.

Ωστόσο, υπάρχουν και μειονεκτήματα που εμφανίζει η προσέγγιση των model-based συστημάτων. Το βασικότερο μειονέκτημά τους είναι ότι απαιτούν ανά τακτά χρονικά διαστήματα φάσεις εκπαίδευσης για να μάθουν το μοντέλο πρόβλεψης, και η διαδικασία αυτή κοστίζει ακριβά. Επίσης, επειδή στα συστήματα αυτά χρησιμοποιούνται συχνά μέθοδοι μείωσης διαστάσεων, μπορεί να χαθεί χρήσιμη πληροφορία από τις μειώσεις.

Στα memory-based συστήματα, οι βαθμολογήσεις χρηστών-προϊόντων που είναι αποθηκευμένες στο σύστημα χρησιμοποιούνται άμεσα για την εξαγωγή προβλέψεων βαθμολογιών σε νέα προϊόντα. Με τα δεδομένα που έχει στην μνήμη του, το σύστημα υπολογίζει την ομοιότητα ανάμεσα σε προϊόντα ή χρήστες και εξάγει την πρόβλεψη για τον χρήστη από τον σταθμισμένο μέσο όρο των βαθμολογιών. Ο υπολογισμός της ομοιότητας γίνεται με μηχανισμούς όπως η συσχέτιση Pearson (Pearson correlation) ή η ομοιότητα με βάση το συνημιτονικό διάνυσμα (cosine-based similarity). Ο μηχανισμός αυτός ήταν ο πρώτος που διαμορφώθηκε και πλέον χρησιμοποιείται σε πολλά εμπορικά συστήματα, καθώς είναι αποδοτικός και εύκολος στην εφαρμογή. Χαρακτηριστικά παραδείγματα των συστημάτων με βάση την μνήμη είναι οι top-N συστάσεις με βάση τον χρήστη ή το προϊόν. Τα συστήματα αυτά έχουν πολλά πλεονεκτήματα, ορισμένα εκ των οποίων αναφέρονται παρακάτω :

Αρχικά η απλότητα. Τα συστήματα αυτά είναι πολύ εύκολα στην δημιουργία και πολύ απλά στην εφαρμογή τους. Στην πιο απλή τους μορφή, ο μόνος παράγοντας που χρειάζεται ρύθμιση είναι ο αριθμός των χρηστών-γειτόνων που θα χρησιμοποιηθούν για την εξαγωγή της πρόβλεψης. Ένα άλλο σημαντικό προτέρημα είναι οι εξηγήσεις των προβλέψεων που παρέχει το σύστημα. Για την κάθε πρόβλεψη βαθμολόγησης που υπολογίζεται, δίνεται και μια συνοπτική διαισθητική εξήγηση για το πώς προέκυψε η πρόβλεψη. Η ύπαρξη εξήγησης βοηθά τον χρήστη να καταλάβει καλύτερα

την σύσταση που του γίνεται και τι σχέση έχει με εκείνον. Ένα τρίτο πλεονέκτημα είναι η σταθερότητα των memory-based συστημάτων, καθώς επηρεάζονται ελάχιστα από την συνεχή προσθήκη νέων χρηστών, προϊόντων και βαθμολογιών, γεγονός που είναι πολύ συνηθισμένο στις μεγάλες εμπορικές εφαρμογές που τα χρησιμοποιούν. Για παράδειγμα, το σύστημα με βάση την μνήμη μπορεί να κάνει συστάσεις σε νέους χρήστες στηριζόμενα στις ομοιότητες προϊόντων που έχουν ήδη υπολογιστεί και δεν χρειάζεται να εκπαιδευτεί ξανά το σύστημα. Τέλος, ένα άλλο προτέρημα είναι η αποδοτικότητά τους. Σε αντίθεση με τα model-based συστήματα, δεν χρειάζεται να εκπαιδεύουν το σύστημα, διαδικασία ακριβή και χρονοβόρα. Επίσης, η αποθήκευση των κοντινότερων γειτόνων απαιτεί πολύ μικρή μνήμη, και αυτό είναι πολύ θετικό για τις εφαρμογές που διαχειρίζονται εκατομμύρια προϊόντα και χρήστες.

Φυσικά, υπάρχουν και σημεία στα οποία υστερούν τα memory-based συστήματα. Αρχικά, ένα μειονέκτημα είναι ότι εξαρτώνται από τις βαθμολογήσεις των χρηστών και έτσι αν οι χρήστες δεν έχουν αλληλεπιδράσεις με το σύστημα τότε εκείνο δεν μπορεί να κάνει αποτελεσματικές προβλέψεις. Επιπλέον, όταν τα δεδομένα των βαθμολογιών είναι αραιά, συμβάν που προκύπτει συχνά σε προϊόντα που σχετίζονται με το διαδίκτυο, τότε μειώνεται και η απόδοση του συστήματος. Έτσι δημιουργείται πρόβλημα σε μεγάλα datasets.

5.1.2 Συστήματα προτάσεων με Συνεργατικό Φιλτράρισμα με βάση τον Χρήστη και με βάση το Προϊόν

Οι συστάσεις με βάση τον χρήστη προβλέπουν την βαθμολογία r_{ui} ενός χρήστη u για ένα νέο προϊόν i, χρησιμοποιώντας τις βαθμολογίες που έχουν δώσει για το i άλλοι χρήστες που έχουν παρόμοια συμπεριφορά με εκείνον, οι οποίοι ονομάζονται κοντινότεροι γείτονες (nearest-neighbors). [17] Στη συνέχεια, για κάθε χρήστη v υπολογίζεται η ομοιότητά του w_{uv} με τον u. Σε αυτήν τη περίπτωση λοιπόν, η εξαγωγή πρόβλεψης γίνεται σε δύο βασικά βήματα :

1) Ψάξε για χρήστες οι οποίοι έχουν παρόμοια συμπεριφορά στην βαθμολόγηση των προϊόντων με τον χρήστη-στόχο
2) Χρησιμοποίησε τις βαθμολογίες των χρηστών που ταιριάζουν με τον χρήστη-στόχο που βρήκες στο πρώτο βήμα για να υπολογίσεις πρόβλεψη για τον στόχο

Η ομοιότητα μεταξύ δυο χρηστών στηρίζεται στις βαθμολογίες άλλων χρηστών που έχει το σύστημα στην μήτρα χρηστών-προϊόντων και υπολογίζεται με την συσχέτιση του Pearson ως :

$$\kappa_{x,y} = \mathrm{sim}(u_x, u_y) = \frac{\sum_{h=1}^{n'} \left(r_{u_x,i_h} - \overline{r_{u_x}}\right) - \left(r_{u_y,i_h} - \overline{r_{u_y}}\right)}{\sqrt{\sum_{h=1}^{n'} \left(r_{u_x,i_h} - \overline{r_{u_x}}\right)^2} \sqrt{\sum_{h=1}^{n'} \left(r_{u_y,i_h} - \overline{r_{u_y}}\right)^2}}.$$

[17]

Όπου n είναι ο συνολικός αριθμός των προϊόντων στη βάση δεδομένων, $r_{ux,ih}$ είναι η άμεση βαθμολογία του χρήστη u_x στο προϊόν i_h και $\overline{r_{ux}}$ είναι η μέση βαθμολογία του χρήστη u_x. Θα μελετηθούν οι μέθοδοι υπολογισμού της ομοιότητας στο Συνεργατικό Φιλτράρισμα σε επόμενη ενότητα.

Η πρόβλεψη της βαθμολογίας r_{ui} υπολογίζεται ως ο μέσος όρος των βαθμολογιών που έχουν δώσει οι γείτονες στο προϊόν i :

$$\hat{r}_{ui} = \frac{1}{|\mathcal{N}_i(u)|} \sum_{v \in \mathcal{N}_i(u)} r_{vi}$$

[3, σελ.115]

Όπου $N_i(u)$ είναι η ομάδα με τους κ κοντινότερους γείτονες του χρήστη στόχου, οι οποίοι είχαν την υψηλότερη ομοιότητα w_{uv} με τον u. Στον παραπάνω τρόπο υπολογισμού της πρόβλεψης υπάρχει το πρόβλημα ότι δεν λαμβάνονται υπ' όψιν τα επίπεδα ομοιότητας. Αν κάποιος γείτονας του χρήστη-στόχου έχει πολύ ομοιότερη συμπεριφορά με εκείνον από ότι κάποιος άλλος γείτονας, τότε θα πρέπει η βαθμολόγηση του πρώτου για το προϊόν i να μετρήσει περισσότερο από του δεύτερου. Για αυτόν το λόγο προστέθηκαν βάρη στον τύπο εξαγωγής πρόβλεψης, ώστε να σταθμίζεται η συνεισφορά του κάθε γείτονα στην πρόβλεψη ανάλογα με την ομοιότητά του. Τα βάρη αυτά κανονικοποιούνται ώστε η προβλεπόμενη βαθμολογία να μην είναι εκτός των επιτρεπόμενων τιμών. Ο παραπάνω τύπος λοιπόν βελτιώνεται και παίρνει την μορφή :

$$\hat{r}_{ui} = \frac{\sum_{v \in \mathcal{N}_i(u)} w_{uv} r_{vi}}{\sum_{v \in \mathcal{N}_i(u)} |w_{uv}|}.$$

[3, σελ.115]

Χρησιμοποιείται η απόλυτη τιμή των ομοιοτήτων στον παρονομαστή για να μην ξεφύγει η πρόβλεψη από τις επιτρεπόμενες τιμές.

Μια τελευταία βελτίωση που απαιτεί ο τύπος της πρόβλεψης με βάρη αφορά το γεγονός ότι οι χρήστες δεν έχουν όλοι τον ίδιο τρόπο έκφρασης του ενθουσιασμού τους για κάποιο προϊόν. Έτσι, αν σε δύο χρήστες άρεσε πάρα πολύ κάποιο προϊόν, μπορεί ο ένας να το βαθμολογήσει με την υψηλότερη τιμή, ενώ ο άλλος να είναι πιο δύσκολος και να κρατάει το «Άριστα» για κάτι που θα τον συνεπάρει. Για να λυθεί αυτό το πρόβλημα, αντικαθίστανται οι βαθμολογήσεις r_{vi} των γειτόνων με τις κανονικοποιημένες βαθμολογήσεις $h(r_{vi})$, και ο τύπος της πρόβλεψης γίνεται :

$$\hat{r}_{ui} = h^{-1}\left(\frac{\sum_{v \in \mathcal{N}_i(u)} w_{uv} h(r_{vi})}{\sum_{v \in \mathcal{N}_i(u)} |w_{uv}|}\right)$$

[3, σελ.116]

Από την άλλη πλευρά, υπάρχουν οι συστάσεις με βάση το προϊόν. [17] Στην προσέγγιση αυτή, το σύστημα για να υπολογίσει την προβλεπόμενη βαθμολογία r_{ui} του χρήστη στόχου u για το προϊόν i, λαμβάνει υπ' όψιν τις βαθμολογήσεις που έχει δώσει ο u για άλλα προϊόντα j που είναι παρόμοια με το i. Η εξαγωγή πρόβλεψης σε αυτήν την περίπτωση γίνεται με τα παρακάτω δύο βήματα :

1) Φτιάξε μια μήτρα προϊόντων που να δηλώνει τις σχέσεις για όλα τα ζευγάρια προϊόντων (κατά πόσο είναι όμοια μεταξύ τους)
2) Χρησιμοποιώντας αυτήν την μήτρα και τα δεδομένα του χρήστη-στόχου, συμπέρανε το γούστο του

Η ομοιότητα ανάμεσα σε δυο προϊόντα υπολογίζεται ως η συσχέτιση Pearson των σχετικών στηλών της μήτρας χρηστών-προϊόντων ως :

$$\mu_{x,y} = \mathrm{sim}(i_x, i_y) = \frac{\sum_{h=1}^{m'} \left(r_{u_h,i_x} - \overline{r_{i_x}}\right) - \left(r_{u_h,i_y} - \overline{r_{i_y}}\right)}{\sqrt{\sum_{h=1}^{m'} \left(r_{u_h,i_x} - \overline{r_{i_x}}\right)^2} \sqrt{\sum_{h=1}^{m'} \left(r_{u_h,i_y} - \overline{r_{i_y}}\right)^2}}.$$

[17]

Όπου m είναι ο συνολικός αριθμός χρηστών στη βάση δεδομένων, $r_{uh,ix}$ είναι η άμεση βαθμολογία του χρήστη u_h στο προϊόν i_x και $\overline{r_{ix}}$ είναι η μέση βαθμολογία του προϊόντος i_x.

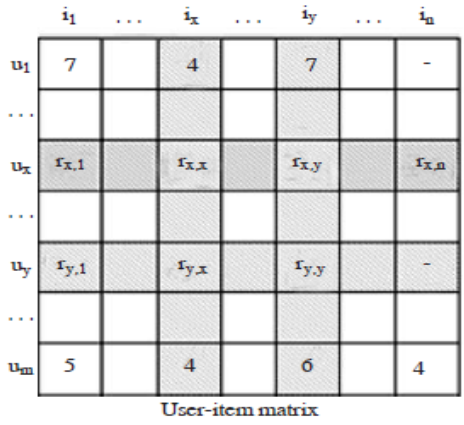
Το συνεργατικό φιλτράρισμα με βάση το προϊόν εφαρμόζεται ευρέως στο Amazon.com. Έστω ότι $N_u(i)$ είναι η ομάδα προϊόντων που έχει βαθμολογήσει ο u που είναι παρόμοια με το i. Η πρόβλεψη

για τη βαθμολόγηση του u στο i μπορεί να υπολογιστεί ως ο σταθμισμένος μέσος όρος των βαθμολογιών του u στα προϊόντα του $N_u(i)$:

$$\hat{r}_{ui} = \frac{\sum_{j \in \mathcal{N}_u(i)} w_{ij} r_{uj}}{\sum_{j \in \mathcal{N}_u(i)} |w_{ij}|}$$ [3, σελ.116]

Όπως αναφέραμε και στην περίπτωση του Συνεργατικού Φιλτραρίσματος με βάση τον χρήστη, οι χρήστες μπορεί να δείχνουν με διαφορετικό τρόπο την εκτίμησή τους για τα προϊόντα, κάποιος μπορεί να είναι περισσότερο εκδηλωτικός και κάποιος άλλος πιο μετρημένος στην βαθμολογία του. Για τον λόγο αυτό, οι διαφορές στην κλίμακα βαθμολόγησης αμβλύνονται με την κανονικοποίηση των βαθμολογιών όπως φαίνεται στην εξίσωση που ακολουθεί :

$$\hat{r}_{ui} = h^{-1}\left(\frac{\sum_{j \in \mathcal{N}_u(i)} w_{ij} h(r_{uj})}{\sum_{j \in \mathcal{N}_u(i)} |w_{ij}|}\right)$$ [3, σελ.116]



Εικόνα 9 : Παράδειγμα μήτρας χρηστών προϊόντων [17]

Οι δυο αυτές προσεγγίσεις εμφανίζουν διάφορα πλεονεκτήματα και μειονεκτήματα στη συμπεριφορά τους, σχετικά με διάφορα κριτήρια. Στην περίπτωση του συνεργατικού φιλτραρίσματος με βάση τον χρήστη, η ομοιότητα ανάμεσα σε δυο χρήστες υπολογίζεται συγκρίνοντας τις βαθμολογίες που έχουν δώσει αυτοί οι δυο χρήστες στα ίδια προϊόντα. Στο συνεργατικό φιλτράρισμα με βάση το προϊόν, η ομοιότητα ανάμεσα σε δυο προϊόντα υπολογίζεται λαμβάνοντας υπ' όψιν τις βαθμολογίες που έχει δώσει ο ίδιος χρήστης στα δυο συγκεκριμένα προϊόντα. Η ακρίβεια λοιπόν των συστημάτων προτάσεων στην πρόβλεψη που κάνουν εξαρτάται από τον λόγο χρηστών και προϊόντων που είναι αποθηκευμένα στο σύστημα. Σε γενικές γραμμές, είναι προτιμότερο να στηρίζεται το σύστημα σε λίγους «γείτονες» οι οποίοι όμως είναι αξιόπιστοι, παρά σε έναν ευρύτερο κύκλο γειτόνων στον οποίο τα βάρη ομοιότητας κάποιων δεν είναι ιδιαίτερα υψηλά. Σε περιπτώσεις στις οποίες ο αριθμός των χρηστών ξεπερνάει κατά πολύ των αριθμό των προϊόντων (οι περιπτώσεις αυτές είναι και οι συχνότερες στις διαδικτυακές εφαρμογές συστημάτων προτάσεων), όπως στο Amazon, τότε τα συστήματα προτάσεων με βάση τα προϊόντα δείχνουν να έχουν καλύτερες επιδόσεις. Αναλόγως, σε ένα σύστημα στο οποίο τα αποθηκευμένα προϊόντα είναι κατά πολύ περισσότερα από τους χρήστες, τότε προτιμότερες στον τομέα της

ακρίβειας πρόβλεψης είναι οι μέθοδοι προβλέψεων με βάση τον χρήστη. Συνολικά, όσον αφορά την στατιστική ακρίβεια, οι αλγόριθμοι πρόβλεψης με βάση το προϊόν λειτουργούν καλύτερα από εκείνους με βάση τον χρήστη. Επιπλέον, η προσέγγιση με βάση το προϊόν χρειάζεται λιγότερη μνήμη και χρόνο για να υπολογίσει τις σταθμισμένες ομοιότητες από ότι οι μέθοδοι με βάση τον χρήστη. Η χρονική πολυπλοκότητα ωστόσο των δυο προσεγγίσεων στη βάση της on-line σύστασης είναι η ίδια. Όσον αφορά τη σταθερότητα, η επιλογή ανάμεσα στην user-based και την item-based μέθοδο εξαρτάται από την συχνότητα και την ποσότητα των αλλαγών στους χρήστες και στα προϊόντα του συστήματος. Αν τα διαθέσιμα αποθηκευμένα προϊόντα στο σύστημα είναι σταθερά σε σχέση με την λίστα των χρηστών, τότε μια προσέγγιση με βάση το προϊόν είναι προτιμότερη γιατί η σταθμισμένη ομοιότητα των προϊόντων θα μπορεί αν υπολογίζεται ανά αραιά χρονικά διαστήματα και το σύστημα θα είναι σε θέση να προτείνει προϊόντα σε νέους χρήστες. Αντιθέτως, σε εφαρμογές στις οποίες η λίστα με τα προϊόντα του συστήματος αλλάζει συνεχώς, ενώ οι χρήστες είναι σταθεροί, προτιμότερη είναι η μέθοδος με βάση τον χρήστη από άποψη σταθερότητας. Σχετικά με το θέμα της δικαιολόγησης, τα συστήματα προτάσεων με βάση το προϊόν υπερτερούν. Είναι εύκολο για το σύστημα να παρουσιάσει τη λίστα με τα γειτονικά προϊόντα και τις ομοιότητές τους στον χρήστη, ως μια εξήγηση της σύστασης που του έκανε. Ο χρήστης μπορεί εύκολα να συμμετέχει ενεργά στην εξαγωγή πρόβλεψης, αλλάζοντας τα γειτονικά προϊόντα ή τα βάρη τους, διαδικασία που δεν επιδέχονται τα συστήματα προτάσεων με βάση τον χρήστη, μιας και ο χρήστης δεν ξέρει τους «γείτονές» του και έτσι δεν μπορεί να μορφοποιήσει τη λίστα τους. Ένα βασικό πλεονέκτημα των user-based μεθόδων έναντι των item-based είναι η καινοτομία. Τα item-based συστήματα, η προβλεπόμενη βαθμολογία ενός χρήστη για ένα προϊόν στηρίζεται στην βαθμολογία του χρήστη για άλλα παρόμοια προϊόντα, έτσι τα συστήματα αυτά έχουν την τάση να συστήνουν προϊόντα που σχετίζονται άμεσα με εκείνα που ήδη έχει δει και εκτιμήσει ο χρήστης. Αντίθετα, αυτό το πρόβλημα έχει πολύ μικρότερες διαστάσεις στα user-based συστήματα, αφού η πρόβλεψη στηρίζεται στις βαθμολογήσεις άλλων χρηστών, οι οποίοι είναι όμοιοι με τον χρήστη-στόχο στα κοινά προϊόντα που έχουν βαθμολογήσει, αλλά οι γείτονες μπορεί να έχουν ενδιαφέροντα σε άλλους τομείς πέρα από τα κοινά προϊόντα, οι οποίοι θα προταθούν στον χρήστη-στόχο.

Εικόνα 10 : Παράδειγμα εξήγησης σύστασης στο Amazon.com που χρησιμοποιεί Συνεργατικό Φιλτράρισμα με βάση το προϊόν [18]

5.2 Ανάγκες χρηστών που καλύπτονται από το Συνεργατικό Φιλτράρισμα

Το Συνεργατικό Φιλτράρισμα στηρίζει πολλές λειτουργίες, εφαρμόζεται σε πολλές ιστοσελίδες και καλύπτει πολλές και διαφορετικές ανάγκες των χρηστών. Οι χρήστες χρησιμοποιούν το Συνεργατικό Φιλτράρισμα για πραγματοποιήσουν διεργασίες όπως οι επόμενες [18] :

- «*Θέλω να βρω καινούρια προϊόντα που να μου αρέσουν» :* Στο σύγχρονο διαδίκτυο, το φαινόμενο της υπερφόρτωσης πληροφοριών δεν αφήνει περιθώρια στον χρήστη να μπορέσει να αξιολογήσει όλα τα διαθέσιμα προϊόντα. Το Συνεργατικό Φιλτράρισμα λοιπόν του παρουσιάζει κάποια που πιθανώς να τον ενδιαφέρουν περισσότερο, ώστε να διαλέξει από αυτά. Η σύσταση αφορά προϊόντα από μουσικά κομμάτια και βιβλία ως ιστοσελίδες ή ερευνητικές εργασίες.
- «*Θέλω μια γνώμη για ένα συγκεκριμένο προϊόν*» : Ο χρήστης ενδιαφέρεται για ένα νέο προϊόν αλλά δεν έχει πληροφορίες για αυτό ώστε να διαμορφώσει γνώμη. Μπορεί λοιπόν να μάθει την άποψη που έχουν οι φίλοι του με τους οποίους συσχετίζεται στην διαδικτυακή κοινότητα και έτσι να αποφασίσει για ένα προϊόν μέσω αυτών.
- «*Θέλω να βρω έναν ή πολλούς χρήστες που να μου ταιριάζουν*» : Είναι πολύ χρήσιμο για τον χρήστη να έχει έναν κύκλο άλλων χρηστών με τους οποίους να μοιράζεται κοινά ενδιαφέροντα και απόψεις. Άμα ο χρήστης έχει έναν κύκλο φίλων τότε μπορούν να συμμετέχουν σε ομάδες συζήτησης ή να ανταλλάσουν γνώμες και συστάσεις στα Κοινωνικά Μέσα.
- «*Θέλω να βρούμε σαν ομάδα ένα καινούριο προϊόν που να μας αρέσει*» : Πολλές φορές, οι χρήστες αναζητούν προϊόντα όχι ατομικά, αλλά συλλογικά ώστε να καλύπτει το προϊόν τις ανάγκες όλων των χρηστών της κοινωνικής διαδικτυακής ομάδας που ανήκουν. Για παράδειγμα, μια παρέα φίλων που θέλουν να δουν μια ταινία ή μια οικογένεια που θέλει να παρακολουθήσει ένα πρόγραμμα στην τηλεόραση. Το συνεργατικό Φιλτράρισμα παρέχει αυτήν την δυνατότητα στους χρήστες του.
- «*Θέλω να βρω έναν συνδυασμό παλιών και νέων προϊόντων*» : Υπάρχουν πολλές περιπτώσεις στις οποίες οι χρήστες του συστήματος δεν ενδιαφέρονται να τους προταθούν καινούρια προϊόντα, αλλά περισσότερο να καλυφθούν οι ανάγκες τους για εκείνη την ώρα. Έτσι, μπορεί ένας χρήστης να θέλει να πάει σε ένα εστιατόριο με την παρέα του ακόμα και αν έχει ξαναφάει ο ίδιος εκεί ή να δει κάποια κλασσική ταινία ακόμη και να την έχει ξαναδεί.
- «*Θέλω να βρω συγκεκριμένα προϊόντα σε σχέση με την εκάστοτε περίπτωση*» : Το Συνεργατικό Φιλτράρισμα επιτρέπει σε ένα σύστημα προτάσεων να κάνει συστάσεις λαμβάνοντας υπ' όψιν την περίσταση. Για παράδειγμα, ένας χρήστης μπορεί να θέλει να δει κάποια ταινία με την σχέση του και κάποιου άλλου είδους ταινία με την παρέα του. Ωστόσο, δεν έχει γίνει ακόμα αρκετή έρευνα επικεντρωμένη σε συστάσεις συγκεκριμένων περιπτώσεων.

5.3 Μέθοδοι που εφαρμόζονται στα στάδια εξαγωγής της πρόβλεψης

Τα Συστήματα Προτάσεων στα Κοινωνικά Μέσα εφαρμόζουν κατά κόρον μεθόδους των συστημάτων με βάση τη μνήμη (memory-based ή neighborhood-based systems). Όπως αναφέραμε

στη ενότητα 5.1, για να γίνει πρόβλεψη βαθμολογίας ενός χρήστη για κάποιο προϊόν, θα πρέπει (α) πρώτα να γίνει κανονικοποίηση των βαθμολογιών, (β) να υπολογιστούν οι σταθμισμένες ομοιότητες και (γ) φυσικά να γίνει η επιλογή των γειτόνων για τον χρήστη-στόχο. Σε αυτήν την ενότητα μελετηθούν οι μέθοδοι που χρησιμοποιούνται σε καθένα από αυτά τα τρία βήματα.

5.3.1 Κανονικοποίηση των βαθμολογιών

Ένα συνηθισμένο πρόβλημα στο Συνεργατικό Φιλτράρισμα είναι ότι διαφορετικοί χρήστες υιοθετούν διαφορετικά κριτήρια για την επιλογή της βαθμολογίας που θα δώσουν σε ένα προϊόν, ανάλογα με την προσωπική τους κλίμακα. Αυτό έχει ως αποτέλεσμα άτομα με τα ίδια ενδιαφέροντα να δίνουν διαφορετική βαθμολογία στο ίδιο προϊόν. Το πρόβλημα αυτό αντιμετωπίζεται με την κανονικοποίηση των βαθμολογιών, έτσι ώστε οι ατομικές βαθμολογίες να προσαρμόζονται σε μια παγκόσμια κλίμακα. Τρεις δημοφιλείς μέθοδοι κανονικοποίησης που χρησιμοποιούνται στα συστήματα με βάση τη μνήμη είναι η Gaussian μέθοδος, το κεντράρισμα στο μέσο και η Z-score μέθοδος [3], [19].

5.3.1.1 Gaussian μέθοδος κανονικοποίησης

Η Γκαουσιανή μέθοδος κανονικοποίησης εξετάζει δυο βασικούς παράγοντες που μπορεί να οδηγήσουν σε διακύμανση βαθμολογιών χρηστών που έχουν κοινές προτιμήσεις :

1) Η μετατόπιση των μέσων βαθμολογιών : Σχετίζεται με το γεγονός ότι κάποιοι χρήστες μπορεί να είναι πιο ανεκτικοί και εύκολοι και έτσι να δίνουν υψηλότερες βαθμολογίες από άλλους. Αυτοί οι χρήστες λοιπόν έχουν μέση βαθμολογία υψηλότερη από τους πιο αυστηρούς χρήστες που δεν βάζουν εύκολα ψηλό βαθμό. Αυτός ο παράγοντας λαμβάνεται υπ' όψιν αφαιρώντας την βαθμολογία του κάθε χρήστη από την μέση βαθμολογία.
2) Οι διαφορετικές βαθμολογικές κλίμακες : Το πρόβλημα αυτό αφορά το συμβάν ότι οι πιο συντηρητικοί χρήστες αντιστοιχούν τα προϊόντα σε ένα περιορισμένο εύρος βαθμολογικών κατηγοριών, ενώ οι πιο ελεύθεροι χρήστες χρησιμοποιούν ένα ευρύτερο φάσμα στις βαθμολογίες τους. Για να λάβει υπ' όψιν της αυτόν τον παράγοντα, η Gaussian μέθοδος διαιρεί τις βαθμολογίες του κάθε χρήστη με τη διακύμανση των βαθμολογιών του.

Επειδή αυτή η μέθοδος κανονικοποίησης κανονικοποιεί την κατανομή των βαθμολογιών ενός χρήστη σε μια Gaussian κατανομή, για αυτό και ονομάζεται «Gaussian normalization method».
Με βάση τα παραπάνω, η κανονικοποιημένη βαθμολογία ενός χρήστη y για το προιόν x υπολογίζεται ως :

$$\hat{R}_y(x) = \frac{R_y(x) - \bar{R}_y}{\sqrt{\sum_x (R_y(x) - \bar{R}_y)^2}} \quad [19]$$

όπου Ry(x) είναι η βαθμολογία του χρήστη y για το προϊόν x και $\overline{Ry}$ είναι η μέση βαθμολογία του χρήστη y.

5.3.1.2 Μέθοδος Mean-centering

Η κεντρική ιδέα αυτής της μεθόδου είναι να καθοριστεί αν μια βαθμολογία είναι θετική ή αρνητική σε σύγκριση με τη μέση βαθμολογία. Η βαθμολογία ενός χρήστη u για το προϊόν i μετασχηματίζεται σε μια με κέντρο το μέσο. Αυτό γίνεται αφαιρώντας από την εκάστοτε βαθμολογία την μέση βαθμολογία του u για τα προϊόντα της ομάδας Iu στα συστήματα προτάσεων με βάση τον χρήστη και αφαιρώντας την μέση βαθμολογία των χρηστών του Ui για το προϊόν i στα συστήματα με βάση το προϊόν. Οι δυο μετασχηματισμοί φαίνονται παρακάτω :

$$h(r_{ui}) = r_{ui} - \overline{r}_u$$

[3, σελ.121] για τα user-based συστήματα και

$$h(r_{ui}) = r_{ui} - \overline{r}_i$$

[3, σελ.121] για τα item-based.

Η πρόβλεψη για την βαθμολογία rui του χρήστη u στο προϊόν i δίνεται για τις δυο προσεγγίσεις από τους αντίστοιχους τύπους :

$$\hat{r}_{ui} = \overline{r}_u + \frac{\sum_{v \in \mathcal{N}_i(u)} w_{uv} (r_{vi} - \overline{r}_v)}{\sum_{v \in \mathcal{N}_i(u)} |w_{uv}|}$$

[3, σελ.121] και

$$\hat{r}_{ui} = \overline{r}_i + \frac{\sum_{j \in \mathcal{N}_u(i)} w_{ij} (r_{uj} - \overline{r}_j)}{\sum_{j \in \mathcal{N}_u(i)} |w_{ij}|}$$

.[3, σελ.121]

Η μέθοδος αυτή εφαρμόζεται περισσότερο στα συστήματα προτάσεων με βάση το προϊόν. Ένα βασικό προτέρημα αυτής της μεθόδου είναι ότι κάποιος μπορεί να εκτιμήσει άμεσα αν η γνώμη ενός χρήστη για κάποιο συγκεκριμένο προϊόν είναι θετική ή αρνητική, κοιτάζοντας απλά το πρόσημο της βαθμολογίας του μετά την διεργασία της κανονικοποίησης.

5.3.1.3 Κανονικοποίηση Z-score

Η μέθοδος αυτή υπερτερεί έναντι της Mean-Centering μεθόδου στο σημείο ότι όχι μόνο αντισταθμίζει τα offsets που προκαλούνται από την διαφορετική αντίληψη των βαθμολογιών που έχει ο κάθε χρήστης, αλλά λαμβάνει υπ' όψιν της και την έκταση των ατομικών βαθμολογιών, δηλαδή το πόσο αυτές εξαπλώνονται. Για παράδειγμα, αν δυο χρήστες έχουν μέσο όρο βαθμολογίας στα προϊόντα 3 στα 5, αλλά ο πρώτος χρήστης διαμόρφωσε αυτόν τον μέσο όρο δίνοντας βαθμολογίες 1 και 5, ενώ ο δεύτερος δίνοντας συνέχεια τον βαθμό 3, τότε μια βαθμολογία με 5 του δεύτερου χρήστη σημαίνει περισσότερα από ότι το 5 του πρώτου χρήστη και δείχνει μεγαλύτερη εκτίμηση για το προϊόν. Το γεγονός ότι η Z-score κανονικοποίηση εξετάζει τη διακύμανση των ατομικών βαθμολογιών είναι πολύ χρήσιμο, ιδιαίτερα όταν η βαθμολογική κλίμακα έχει συνεχείς τιμές ή έχει μεγάλο εύρος διακριτών τιμών.

Στα συστήματα με βάση τον χρήστη, για να κανονικοποιηθεί η βαθμολογία rui του χρήστη u για το προϊόν i , διαιρείται η mean-centered βαθμολογία με βάση τον χρήστη που δείξαμε στην προηγούμενη ενότητα με την τυπική απόκλιση σu των βαθμολογιών του χρήστη u :

$$h(r_{ui}) = \frac{r_{ui} - \overline{r}_u}{\sigma_u}$$

[3, σελ.123]

Η πρόβλεψη της βαθμολογίας rui για το προϊόν i με την Z-score μέθοδο προσέγγισης δίνεται από τον τύπο :

$$\hat{r}_{ui} = \overline{r}_u + \sigma_u \frac{\sum_{v \in \mathcal{N}_i(u)} w_{uv} \left(r_{vi} - \overline{r}_v\right)/\sigma_v}{\sum_{v \in \mathcal{N}_i(u)} |w_{uv}|}$$

[3, σελ.123]

Αντίστοιχα, στην περίπτωση των συστημάτων προτάσεων με Συνεργατικό Φιλτράρισμα με βάση το προϊόν, η βαθμολογία rui κανονικοποιείται διαιρώντας την mean-centered βαθμολογία με την τυπική απόκλιση σi των βαθμολογιών που έχουν δοθεί από τους χρήστες για το προϊόν i :

$$h(r_{ui}) = \frac{r_{ui} - \overline{r}_i}{\sigma_i}$$

[3, σελ.123]

και στη συνέχεια, η πρόβλεψη για την βαθμολογία rui δίνεται από τον τύπο :

$$\hat{r}_{ui} = \overline{r}_i + \sigma_i \frac{\sum_{j \in \mathcal{N}_u(i)} w_{ij} \left(r_{uj} - \overline{r}_j\right)/\sigma_j}{\sum_{j \in \mathcal{N}_u(i)} |w_{ij}|}$$

[3, σελ.123]

Σε γενικές γραμμές, πρόσφατες έρευνες έχουν δείξει ότι η Z-score κανονικοποίηση έχει πιο σημαντικά οφέλη από άλλες μεθόδους. Παρ'όλα αυτά, σε ορισμένες περιπτώσεις η κανονικοποίηση γενικά με οποιαδήποτε μέθοδο μπορεί να έχει και ανεπιθύμητα αποτελέσματα. Αν για παράδειγμα ένας χρήστης έχει καταχωρήσει μόνο μια τιμή στο σύστημα, ή μερικές ίδιες, τότε η τυπική απόκλιση των βαθμολογιών του θα είναι μηδενική και έτσι η πρόβλεψη της βαθμολόγησης δεν θα μπορεί να οριστεί. Ωστόσο, αν τα βαθμολογικά δεδομένα δεν είναι πολύ αραιά στη μήτρα χρηστών-προϊόντων, τότε η κανονικοποίηση των βαθμολογιών βελτιώνει σημαντικά το αποτέλεσμα της πρόβλεψης.

5.3.2 Υπολογισμός της ομοιότητας

Ο υπολογισμός των σταθμισμένων ομοιοτήτων ανάμεσα σε χρήστες στα συστήματα προτάσεων με βάση τη μνήμη επιτρέπει να γίνει η επιλογή των πιο κοντινών, αξιόπιστων γειτόνων, οι βαθμολογίες των οποίων θα χρησιμοποιηθούν για την εξαγωγή πρόβλεψης. Επιπλέον, μέσω των βαρών, παρέχεται η δυνατότητα να μην είναι η σημασία όλων των γειτόνων η ίδια στην πρόβλεψη. Το στάδιο του υπολογισμού των ομοιοτήτων είναι από τα σημαντικότερα στα memory-based συστήματα, καθώς μπορεί να επηρεάσει σημαντικά την ακρίβεια και την απόδοση του συστήματος.

Ειδικά στην εφαρμογή του αλγορίθμου Knn (k-nearest neighbors),που είναι ένας από τους προτιμώμενους αλγορίθμους στο Συνεργατικό Φιλτράρισμα ο οποίος θα αναλυθεί αργότερα και θα χρησιμοποιηθεί σε εφαρμογή, ο ορισμός της κατάλληλης μεθόδου για τον υπολογισμό των ομοιοτήτων είναι ένα πολύ σημαντικό βήμα. Παρακάτω αναλύονται οι πιο συνηθισμένες τεχνικές για την εύρεση ομοιοτήτων σε χρήστες ή προϊόντα [3], [19].

5.3.2.1 Ομοιότητα με βάση το συνημίτονο

Μια μέθοδος για τον υπολογισμό της ομοιότητας ανάμεσα σε δυο αντικείμενα που χρησιμοποιείται συχνά στην ανάκτηση πληροφοριών, είναι η εκπροσώπηση των αντικειμένων από δυο διανύσματα xa και xb και στη συνέχεια ο υπολογισμός της ομοιότητας συνημιτονικού διανύσματος ως το συνημίτονο της γωνίας που διαμορφώνουν τα δυο διανύσματα :

$$\cos(\mathbf{x}_a, \mathbf{x}_b) = \frac{\mathbf{x}_a^\top \mathbf{x}_b}{||\mathbf{x}_a|| ||\mathbf{x}_b||}$$

[3, σελ.124]

Όταν το σύστημα προτάσεων θέλει να προτείνει ένα προϊόν σε κάποιον χρήστη, υπολογίζει την ομοιότητα των χρηστών θεωρώντας τον κάθε χρήστη ως ένα διάνυσμα x_u. Έτσι, η ομοιότητα δυο χρηστών u και v με βάση τα προϊόντα I_{uv} που έχουν βαθμολογήσει και οι δύο υπολογίζεται ως :

$$CV(u,v) = \cos(\mathbf{x}_u, \mathbf{x}_v) = \frac{\sum_{i \in \mathcal{I}_{uv}} r_{ui} r_{vi}}{\sqrt{\sum_{i \in \mathcal{I}_u} r_{ui}^2 \sum_{j \in \mathcal{I}_v} r_{vj}^2}}$$

[3, σελ.124]

Θα μελετηθεί ένα παράδειγμα εφαρμογής της cosine-based ομοιότητας. Στον παρακάτω πίνακα φαίνονται οι βαθμολογίες που έχουν δώσει τρεις χρήστες για τρεις φωτογραφίες, καθώς και τον μέσο όρο των βαθμολογιών :

Πίνακας 3 : Πίνακας με βαθμολογίες χρηστών-προϊόντων

	ΑΘΗΝΑ	ΜΑΡΙΑ	ΓΙΑΝΝΗΣ	Μ.Ο.
ΤΑΙΝΙΑ 1	3	2	1	2
ΤΑΙΝΙΑ 2	4	2	3	3
ΤΑΙΝΙΑ 3	2	4	5	11/3
Μ.Ο.	3	8/3	3	26/3

Στον πίνακα αυτόν, κάθε γραμμή αντιστοιχεί σε μια ταινία και κάθε στήλη χρήστη αντιστοιχεί στις διαστάσεις που περιγράφουν το διάνυσμα της ταινίας.
Στη συνέχεια, γίνεται κανονικοποίηση των βαθμολογιών σε κάθε γραμμή, διαιρώντας την κάθε τιμή του πίνακα με την τετραγωνική ρίζα του αθροίσματος των τετραγώνων των τιμών της συγκεκριμένης σειράς. Ενδεικτικά, διαιρείται η κάθε βαθμολογία της πρώτης σειράς με την τιμή $\sqrt{(3^2+2^2+1^2} = \sqrt{14} = 3.74$ και έτσι προκύπτουν τα κανονικοποιημένα δεδομένα του πίνακα 4 :

Πίνακας 4 : Πίνακας με τα κανονικοποιημένα δεδομένα για τις ταινίες

	ΑΘΗΝΑ	ΜΑΡΙΑ	ΓΙΑΝΝΗΣ
ΤΑΙΝΙΑ 1	0.8018	0.5345	0.2673
ΤΑΙΝΙΑ 2	0.7428	0.3714	0.557
ΤΑΙΝΙΑ 3	0.2981	0.5963	0.7454

Υπολογίζεται η cosine-based similarity ανάμεσα στις ταινίες παίρνοντας το εσωτερικό γινόμενο των διανυσμάτων τους. Έτσι, η ομοιότητα μεταξύ των ταινιών 1 και 2 υπολογίζεται ως :
(0.8018 * 0.7428) + (0.5345 * 0.3714) + (0.2673 * 0.557) = 0.943.
Από τα εσωτερικά γινόμενα, προκύπτει ο πίνακας ομοιοτήτων προϊόντων που φαίνεται παρακάτω :

Πίνακας 5 : Πίνακας ομοιοτήτων ταινιών

	ΤΑΙΝΙΑ 1	ΤΑΙΝΙΑ 2	ΤΑΙΝΙΑ 3
ΤΑΙΝΙΑ 1	1	0.943	0.757
ΤΑΙΝΙΑ 2	0.943	1	0.858
ΤΑΙΝΙΑ 3	0.757	0.858	1

Όσο πιο κοντά στο 1 είναι μια τιμή στον πίνακα ομοιοτήτων των προϊόντων, τόσο περισσότερο μοιάζουν τα προϊόντα μεταξύ τους. Παρατηρείται λοιπόν ότι οι ταινίες 1 και 2 είναι πολύ παρόμοιες, αφού η ομοιότητά τους έχει τιμή 0.943.

Για να υπολογιστεί τώρα η cosine-based ομοιότητα ανάμεσα σε χρήστες, ακολουθείται παρόμοια διαδικασία. Στον Πίνακα 3 υπάρχουν τα δεδομένα των βαθμολογιών αλλά με αντιστραμμένες γραμμές-στήλες . Σε αυτήν τη περίπτωση, κάθε χρήστης εκπροσωπείται από ένα διάνυσμα, και η βαθμολογία για το κάθε προϊόν αντιστοιχεί σε μια διάσταση του διανύσματος.

Πίνακας 6 : Ανεστραμμένος Πίνακας με βαθμολογίες χρηστών-προϊόντων

	ΤΑΙΝΙΑ 1	ΤΑΙΝΙΑ 2	ΤΑΙΝΙΑ 3	Μ.Ο.
ΑΘΗΝΑ	3	4	2	3
ΜΑΡΙΑ	2	2	4	8/3
ΓΙΑΝΝΗΣ	1	3	5	3
Μ.Ο.	2	3	11/3	26/3

Όπως και προηγουμένως, κανονικοποιούνται πρώτα τα διανύσματα και στη συνέχεια υπολογίζονται τα εσωτερικά γινόμενα των κανονικοποιημένων διανυσμάτων για την εύρεση της ομοιότητας. Οι κανονικοποιημένες τιμές του πίνακα που σχετίζονται με τον κάθε χρήστη βρίσκονται διαιρώντας την κάθε τιμή με την ρίζα του αθροίσματος των τετραγώνων κάθε στοιχείου της εκάστοτε γραμμής. Για παράδειγμα, ο παράγοντας κανονικοποίησης του διανύσματος της Αθηνάς είναι $\sqrt{(3^2 + 4^2 + 2^2} = \sqrt{29} = 5.385$

.

Πίνακας 7 : Πίνακας με τα κανονικοποιημένα δεδομένα για τον κάθε χρήστη

	ΤΑΙΝΙΑ 1	ΤΑΙΝΙΑ 2	ΤΑΙΝΙΑ 3
ΑΘΗΝΑ	0.5571	0.7428	0.3714
ΜΑΡΙΑ	0.4082	0.4082	0.8165
ΓΙΑΝΝΗΣ	0.1690	0.5071	0.8452

Το επόμενο βήμα είναι ο υπολογισμός των ομοιοτήτων ανάμεσα στους χρήστες, χρησιμοποιώντας το εσωτερικό γινόμενο των κανονικοποιημένων διανυσμάτων δυο χρηστών. Η ομοιότητα ανάμεσα στην Αθηνά και τη Μαρία υπολογίζεται ως :
(0.5571 * 0.4082) + (0.7428 * 0.4082) + (0.3714 * 0.8165) = 0.83

Πίνακας 8 : Πίνακας ομοιοτήτων χρηστών

	ΑΘΗΝΑ	ΜΑΡΙΑ	ΓΙΑΝΝΗΣ
ΑΘΗΝΑ	1	0.83	0.78
ΜΑΡΙΑ	0.83	1	0.975
ΓΙΑΝΝΗΣ	0.78	0.97	1

Με βάση τον πίνακα, οι χρήστες Αθηνά και Μαρία έχουν κοινές προτιμήσεις και ταιριάζουν.

Ένα πρόβλημα με αυτήν την μέθοδο υπολογισμού ομοιότητας είναι ότι χρησιμοποιεί τις αρχικές βαθμολογίες που βάζει ένας χρήστης σε κάποιο προϊόν, χωρίς να λαμβάνει υπ' όψιν τις αποκλίσεις των βαθμολογιών από τον μέσο όρο της βαθμολογίας του χρήστη. Μια δημοφιλής τεχνική που συγκρίνει βαθμολογίες υπολογίζοντας όμως και την επίδραση που έχουν ο μέσος και η διακύμανση είναι η ομοιότητα με την συσχέτιση Pearson την οποία μελετάμε στην επόμενη υποενότητα.

5.3.2.2 Pearson Correlation ομοιότητα

Μπορεί να υπολογιστεί η ομοιότητα μεταξύ δυο προϊόντων μέσω της Pearson συσχέτισης. Η συσχέτιση αυτή ανάμεσα σε δυο αντικείμενα είναι ένας αριθμός ανάμεσα στο -1 και το 1 και μας δείχνει την κατεύθυνση και το μέγεθος της συσχέτισης των δυο προϊόντων ή χρηστών. Όσο πιο κοντά στο 1 είναι η απόλυτη τιμή τόσο μεγαλύτερη η συσχέτιση των αντικειμένων. Η κατεύθυνση της συσχέτισης μας δείχνει πώς μεταβάλλονται οι μεταβλητές. Έτσι, αρνητική συσχέτιση σημαίνει ότι όταν η βαθμολογία του ενός αντικειμένου ανεβαίνει, η βαθμολογία του άλλου πέφτει, ενώ θετική συσχέτιση δείχνει ότι οι μεταβολές των βαθμολογιών έχουν την ίδια κατεύθυνση.
Έστω U η ομάδα των χρηστών που έχουν βαθμολογήσει και το προϊόν i και το j, και I η ομάδα των προϊόντων που έχουν βαθμολογηθεί από τους χρήστες u και v. Η ομοιότητα δυο χρηστών με τη συσχέτιση Pearson υπολογίζεται ως :

$$\mathrm{PC}(u,v) = \frac{\sum_{i \in \mathcal{I}_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in \mathcal{I}_{uv}} (r_{ui} - \bar{r}_u)^2 \sum_{i \in \mathcal{I}_{uv}} (r_{vi} - \bar{r}_v)^2}}$$

[3, σελ.125]

όπου rui είναι η βαθμολόγηση του χρήστη u στο προϊόν i και $\overline{r_u}$ η μέση βαθμολογία του χρήστη u, και η ομοιότητα δυο αντικειμένων με βάση τη συσχέτιση Pearson υπολογίζεται από την εξίσωση που ακολουθεί :

$$\mathrm{PC}(i,j) = \frac{\sum_{u \in \mathcal{U}_{ij}} (r_{ui} - \overline{r}_i)(r_{uj} - \overline{r}_j)}{\sqrt{\sum_{u \in \mathcal{U}_{ij}} (r_{ui} - \overline{r}_i)^2 \sum_{u \in \mathcal{U}_{ij}} (r_{uj} - \overline{r}_j)^2}}$$

[3, σελ.125]

όπου rui είναι η βαθμολόγηση του χρήστη u στο προϊόν i και $\overline{r_i}$ η μέση βαθμολογία του προϊόντος i.

Με βάση τα δεδομένα του πίνακα 3, θα υπολογιστεί η συσχέτιση μεταξύ των ταινιών. Σύμφωνα με τον παραπάνω τύπο, η συσχέτιση ανάμεσα στις ταινίες 1 και 2 υπολογίζεται ως εξής :

$$\frac{(3-2)(4-3)+(2-2)(2-3)+(1-2)(3-3)}{\sqrt{(3-2)^2+(2-2)^2+(1-2)^2}\sqrt{(4-3)^2+(2-3)^2+(3-3)^2}} = \frac{1}{2} = 0.5$$

Με τον παραπάνω τρόπο, υπολογίζονται οι συσχετίσεις μεταξύ όλων των ζευγαριών ταινιών και έτσι προκύπτει ο πίνακας που ακολουθεί :

Πίνακας 9 : Πίνακας συσχετίσεων των ταινιών

	ΤΑΙΝΙΑ 1	ΤΑΙΝΙΑ 2	ΤΑΙΝΙΑ 3
ΤΑΙΝΙΑ 1	1	0.5	-0.982
ΤΑΙΝΙΑ 2	0.5	1	-0.655
ΤΑΙΝΙΑ 3	-0.982	-0.655	1

Με τον πίνακα αυτόν φαίνεται εύκολα ποια προϊόντα συσχετίζονται μεταξύ τους και πώς, για παράδειγμα οι ταινίες 1 και 3 συσχετίζονται αρνητικά και σε μεγάλο βαθμό. Ομοίως, υπολογίζεται και η συσχέτιση μεταξύ χρηστών και έτσι προκύπτει ο παρακάτω πίνακας :

Πίνακας 10 : Πίνακας συσχετίσεων χρηστών

	ΑΘΗΝΑ	ΜΑΡΙΑ	ΓΙΑΝΝΗΣ
ΑΘΗΝΑ	1	-0.866	-0.5
ΜΑΡΙΑ	-0.866	1	0.87
ΓΙΑΝΝΗΣ	-0.5	0.87	1

Με βάση τον πίνακα συσχετίσεων χρηστών, οι χρήστες Μαρία και Γιάννης συσχετίζονται σε μεγάλο βαθμό, δηλαδή αν στον έναν αρέσει κάποιο προϊόν, τότε κατά πάσα πιθανότητα το ίδιο προϊόν θα αρέσει και στον άλλον. Η Αθηνά συσχετίζεται αρνητικά και με την Μαρία και με τον Γιάννη, οπότε πιθανότατα να μην της αρέσουν όσα προϊόντα αρέσουν σε εκείνους.

Ένα μειονέκτημα της μεθόδου αυτής είναι πως δεν λαμβάνει υπ' όψιν της το γεγονός ότι οι χρήστες μεταξύ τους έχουν διαφορετικές βαθμολογικές κλίμακες.

5.3.2.3 Προσαρμοσμένη cosine-based ομοιότητα

Όπως αναφέραμε προηγουμένως, η συσχέτιση Pearson δεν λαμβάνει υπ' όψιν της ανάμεσα στους παράγοντες που εξετάζει τις διαφορετικές βαθμολογικές κλίμακες των χρηστών. Για παράδειγμα, η Pearson συσχέτιση έδειξε ότι οι χρήστες Μαρία και Γιάννης συσχετίζονται άμεσα, αλλά δεν έδωσε βάρος στο ότι ο Γιάννης βάζει πιο ακραίες βαθμολογίες σε σχέση με τη Μαρία. Για τον λόγο αυτό, όταν υπολογίζεται η ομοιότητα μεταξύ δυο προϊόντων, είναι προτιμότερο να συγκρίνονται βαθμολογίες που είναι κεντραρισμένες στη μέση βαθμολογία του χρήστη και όχι στη μέση βαθμολογία του προϊόντος.

Η προσαρμοσμένη συνημιτονική ομοιότητα είναι ένας μετασχηματισμός της συσχέτισης Pearson και συγκρίνει βαθμολογίες κεντραρισμένες στο μέσο του χρήστη. Δίνεται από τον παρακάτω τύπο :

$$similarity(i,j) = \frac{\sum_{ueU}(R_{u,i} - \overline{R}_u)(R_{u,j} - \overline{R}_u)}{\sqrt{\sum_{ueU}(R_{u,i} - \overline{R}_u)^2}\sqrt{\sum_{ueU}(R_{u,ij} - \overline{R}_u)^2}}$$

[19, σελ.45]

Για να υπολογιστεί η προσαρμοσμένη cosine-based ομοιότητα, κανονικοποιούνται αρχικά τα δεδομένα του Πίνακα 3, αφαιρώντας από κάθε τιμή κάθε στήλης τον Μέσο όρο βαθμολογίας. Έτσι προκύπτουν τα παρακάτω δεδομένα βαθμολογιών :

Πίνακας 11 : Κανονικοποιημένη μήτρα βαθμολογιών

	ΑΘΗΝΑ	ΜΑΡΙΑ	ΓΙΑΝΝΗΣ
ΤΑΙΝΙΑ 1	0	-2/3	-2
ΤΑΙΝΙΑ 2	1	-2/3	0
ΤΑΙΝΙΑ 3	-1	4/3	2

Με χρήση του τύπου της προσαρμοσμένης cosine-based ομοιότητας υπολογίζονται οι ομοιότητες μεταξύ των ταινιών οι οποίες φαίνονται στον παρακάτω πίνακα :

Πίνακας 12 : Πίνακας ομοιοτήτων ταινιών με adjusted cosine-based similarity

	ΤΑΙΝΙΑ 1	ΤΑΙΝΙΑ 2	ΤΑΙΝΙΑ 3
ΤΑΙΝΙΑ 1	1	0.1754	-0.891
ΤΑΙΝΙΑ 2	0.1754	1	0.604
ΤΑΙΝΙΑ 3	-0.891	0.604	1

Παρατηρείται πως οι Ταινίες 2 και 3 είναι παρόμοιες, ενώ η 1 και 3 σχετίζονται αρνητικά σε μεγάλο βαθμό. Ακολουθώντας την αντίστοιχη διαδικασία, μπορούν να βρεθούν οι ομοιότητες μεταξύ χρηστών όπως φαίνονται παρακάτω :

Πίνακας 13 : Πίνακας ομοιοτήτων χρηστών με adjusted cosine-based similarity

	ΑΘΗΝΑ	ΜΑΡΙΑ	ΓΙΑΝΝΗΣ
ΑΘΗΝΑ	1	-0.675	-0.884
ΜΑΡΙΑ	-0.675	1	-0.253
ΓΙΑΝΝΗΣ	-0.884	-0.253	1

Με χρήση λοιπόν των μεθόδων υπολογισμού ομοιοτήτων, μετατρέπονται οι αρχικές βαθμολογήσεις των χρηστών στα προϊόντα σε ένα dataset που μπορεί να αναλυθεί και να εξαχθούν συμπεράσματα από αυτό.

5.3.3 Επιλογή των «Γειτόνων»

Ένα από τα βασικότερα στάδια στην εξαγωγή πρόβλεψης βαθμολογίας ενός Συστήματος Προτάσεων με Συνεργατικό Φιλτράρισμα είναι η επιλογή των κοντινότερων γειτόνων του χρήστη-στόχου, καθώς ο αριθμός τους παίζει σημαντικό ρόλο στην απόδοση και την ποιότητα του Συστήματος Προτάσεων. Συνήθως η επιλογή των γειτόνων γίνεται σε δυο επίπεδα, στο πρώτο γίνεται ένα γενικό φιλτράρισμα στο οποίο επιλέγονται οι πιο πιθανοί υποψήφιοι για γείτονες και στο δεύτερο κρατάμε μόνο τους καλύτερους υποψήφιους για την εκάστοτε πρόβλεψη.

5.3.3.1 Αρχικό Φιλτράρισμα χρηστών

Σε συστήματα προτάσεων μεγάλης εμβέλειας που έχουν εκατομμύρια χρήστες και προϊόντα, συχνά δεν είναι εφικτό να αποθηκεύει το σύστημα όλες τις ομοιότητες μεταξύ χρηστών ή προϊόντων, καθώς κάτι τέτοιο θα απαιτούσε πολύ μεγάλη διαθέσιμη μνήμη. Για αυτό το λόγο, το σύστημα κάνει ένα φιλτράρισμα ώστε να μειώσει τον αριθμό των βαρών ομοιοτήτων που χρειάζεται να αποθηκεύσει και να περιορίσει τον αριθμό των υποψηφίων γειτόνων των οποίων τα δεδομένα θα χρησιμοποιηθούν στην πρόβλεψη.

Ένας τρόπος για να γίνει το φιλτράρισμα είναι για κάθε χρήστη ή προϊόν να κρατηθεί μια λίστα μόνο με τους N κοντινότερους γείτονες και τις σταθμισμένες ομοιότητές τους. Το πρόβλημα με αυτήν την μέθοδο είναι ότι πρέπει να γίνει σωστή επιλογή του N, καθώς αυτό μπορεί να επηρεάσει την ακρίβεια του αλγορίθμου που χρησιμοποιεί το σύστημα. Αν ο αριθμός N των γειτόνων επιλεγεί έτσι ώστε να είναι πολύ μεγάλος, τότε θα χρειαστεί μεγάλο μέρος της μνήμης του συστήματος για να αποθηκευτούν οι λίστες των γειτόνων και οι ομοιότητές τους, ενώ θα αυξηθεί ταυτόχρονα και ο χρόνος εξαγωγής της πρόβλεψης βαθμολογίας. Αν από την άλλη πλευρά ο αριθμός N είναι πολύ μικρός, τότε το σύστημα δεν θα έχει πρόσβαση για την πρόβλεψη σε πολλά δεδομένα και έτσι ορισμένα προϊόντα μπορεί να μην προταθούν ποτέ.

Μια άλλη μέθοδος είναι το φιλτράρισμα με κατώφλι. Η μέθοδος επιλογής γειτόνων με βάση κάποιο κατώφλι επιλέγει γείτονες που ανήκουν σε μια συγκεκριμένη ομάδα σε σχέση με τις ομοιότητες των προτιμήσεων. Ο αριθμός των γειτόνων που επιλέγονται με αυτήν την μέθοδο δεν είναι σταθερός, αλλά ποικίλει ανάλογα με την τιμή που έχει το κατώφλι που δίνεται στα βάρη ομοιοτήτων. Σε αυτήν την περίπτωση βέβαια υπάρχει δυσκολία στην σωστή επιλογή του αριθμού που θα χρησιμοποιηθεί ως κατώφλι. Ένας αποδοτικός αλγόριθμος επιλογής γειτόνων με βάση το κατώφλι περιγράφεται στην ερευνητική εργασία [21].

Τέλος, μια άλλη τεχνική φιλτραρίσματος των χρηστών για την επιλογή γειτόνων είναι το αρνητικό φιλτράρισμα. Σε γενικές γραμμές, μια αρνητική συσχέτιση προϊόντων ή χρηστών δεν οδηγεί σε τόσο σίγουρα συμπεράσματα όπως μια θετική συσχέτιση, η οποία είναι ξεκάθαρη ένδειξη ότι τα αντικείμενα ανήκουν στην ίδια ομάδα ενδιαφερόντων. Η αρνητική συσχέτιση μπορεί να υποδηλώνει ότι τα αντικείμενα ανήκουν σε διαφορετικές ομάδες, όμως δεν υπάρχει πληροφορία για το πόσο διαφορετικές είναι αυτές οι ομάδες ή αν αυτές οι ομάδες είναι πιθανόν συμβατές για άλλες κατηγορίες προϊόντων. Το πόσο αποδοτικό είναι να βασίζεται στο φιλτράρισμα σε αρνητικές συσχετίσεις εξαρτάται από τα εκάστοτε δεδομένα.

5.3.3.2 Χρήση γειτόνων στις προβλέψεις

Μετά το πρώτο φιλτράρισμα των χρηστών, οι προβλέψεις για τις νέες βαθμολογίες στα προϊόντα γίνονται από τους κ κοντινότερους γείτονες, τους γείτονες δηλαδή με την υψηλότερη τιμή ομοιότητας. Το πρόβλημα είναι η επιλογή του σωστού αριθμού κ που θα κάνει τις προβλέψεις πιο έγκυρες. Γενικά, όταν επιλέγεται μικρή τιμή για το κ τότε η ακρίβεια στην πρόβλεψη του αλγορίθμου πέφτει σημαντικά. Όσο ο αριθμός κ αυξάνεται, βελτιώνεται και η ακρίβεια πρόβλεψης μέχρι μια συγκεκριμένη διαφορετική σε κάθε περίπτωση τιμή, στην οποία οι γείτονες έχουν γίνει πια πάρα πολλοί, και έτσι αντί να χρησιμοποιηθούν οι λίγες και δυνατές σχέσεις ανάμεσα στον χρήστη στόχο και τους κοντινότερους γείτονες, χρησιμοποιούνται πολλές αλλά ασθενέστερες σχέσεις που έχουν αρνητική επίπτωση στην ακρίβεια του αλγορίθμου.

Συνήθως η βέλτιστη τιμή για το κ επιλέγεται με σταυρωτή επικύρωση (cross-validation). Με αυτήν την μέθοδο, επιλέγονται διάφορες τιμές για το κ για τις οποίες τρέχει η testing διαδικασία εξαγωγής πρόβλεψης στα training datasets x φορές. Τελικά, υπολογίζεται η μέση απόδοση των x εκπαιδευτικών μοντέλων και από αυτήν την διαδικασία επιλέγεται το σωστό κ.

5.4 Κριτήρια αξιολόγησης του Συνεργατικού Φιλτραρίσματος

Ένα από τα σημαντικότερα μέτρα αξιολόγησης των αλγορίθμων των συστημάτων προτάσεων είναι η ακρίβειά τους στις προβλέψεις. Η ακρίβεια μπορεί να μετρηθεί είτε ως το μέγεθος του σφάλματος ανάμεσα στην προβλεπόμενη και την πραγματική βαθμολόγηση είτε ως το μέγεθος του σφάλματος ανάμεσα στην προβλεπόμενη και την πραγματική κατάταξη των προϊόντων. Η ακρίβεια στην πρόβλεψη είναι η ικανότητα ενός συστήματος προτάσεων με συνεργατικό φιλτράρισμα να προβλέψει σωστά την βαθμολογία ενός χρήστη σε κάποιο προϊόν.

Η πιο συνηθισμένη μέθοδος για τον υπολογισμό της ακρίβειας στην πρόβλεψη είναι το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error ή MAE). Το σφάλμα αυτό είναι η μέση απόλυτη διαφορά ανάμεσα στην προβλεπόμενη βαθμολογία και την πραγματική βαθμολογία που έχει δώσει ο χρήστης. Έστω U η ομάδα των χρηστών, I η ομάδα των προϊόντων, S το σύνολο των δυνατών τιμών μιας βαθμολογίας και Rtest η ομάδα των διαθέσιμων βαθμολογιών που χρησιμοποιούνται για την αξιολόγηση της ακρίβειας στην πρόβλεψη. Τότε f : U × I → S είναι η συνάρτηση που προβλέπει την βαθμολογία f(u,i) του χρήστη u στο καινούριο προϊόν i. Το μέσο απόλυτο σφάλμα δίνεται από τον τύπο :

$$\mathrm{MAE}(f) = \frac{1}{|\mathcal{R}_{\text{test}}|} \sum_{r_{ui} \in \mathcal{R}_{\text{test}}} |f(u,i) - r_{ui}|$$

[3, σελ.273]

Ενώ κάποιες φορές χρησιμοποιείται και η ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error ή RMSE) :

$$\mathrm{RMSE}(f) = \sqrt{\frac{1}{|\mathcal{R}_{\text{test}}|} \sum_{r_{ui} \in \mathcal{R}_{\text{test}}} (f(u,i) - r_{ui})^2}$$

[3, σελ.273]

Το πλεονέκτημα του μέσου απολύτου σφάλματος είναι ότι είναι απλό και κατανοητό και διάφορα test στα δεδομένα μπορούν να εφαρμοστούν με αυτό. Ωστόσο, το MAE είναι ένα αναξιόπιστο μέτρο στις ταξινομημένες λίστες προτάσεων, καθώς δεν κάνει διακρίσεις ανάμεσα στα σφάλματα της αρχής και του τέλους της λίστας. Για κάποιον χρήστη όμως έχει μεγάλες διαστάσεις ένα σφάλμα στην αρχή της λίστας με τα προϊόντα που του προτείνονται, ενώ ένα σφάλμα στο τέλος της λίστας δεν έχει τόση σημασία για εκείνον.

Τα μέτρα για την ακρίβεια στην κατάταξη προσπαθούν να υπολογίσουν την χρησιμότητα μιας λίστας προτάσεων για τον χρήστη. Συνηθισμένα μέτρα υπολογισμού της ορθότητας της κατάταξης είναι η ακρίβεια (precision) , δηλαδή το ποσοστό των προϊόντων που εμφανίζονται στη λίστα με τις προτάσεις που ο χρήστης θα βαθμολογούσε ως χρήσιμα για τις ανάγκες και τα ενδιαφέροντά του, και η χρησιμότητα ημιζωής (half-life utility) που υπολογίζει μια τιμή, για μια ταξινομημένη λίστα προϊόντων, που προορίζεται να καλύψει ένα ποσοστό της μέγιστης χρησιμότητας που έχει πετύχει η ταξινομημένη λίστα με τις προτάσεις. Η μέγιστη χρησιμότητα επιτυγχάνεται αν όλα τα προϊόντα που έχουν βαθμολογηθεί ως χρήσιμα για τον χρήστη είναι στην λίστα πάνω από εκείνα που έχουν βαθμολογηθεί ως άχρηστα. Έτσι, στην χρησιμότητα ημιζωής, λάθη στην αρχή της λίστας προτάσεων έχουν εκθετικά μεγαλύτερο βάρος από τα λάθη του τέλους της λίστας.

Αν και τα περισσότερα μέτρα αξιολόγησης των συστημάτων προτάσεων με συνεργατικό φιλτράρισμα αφορούν την ακρίβεια, υπάρχουν και ποικίλα άλλα μέτρα εξίσου σημαντικά με αυτό, τα οποία παρατίθενται παρακάτω [18] :

- **Καινοτομία (Novelty) :** Ένα σημαντικό μέτρο αξιολόγησης των συστημάτων συνεργατικού φιλτραρίσματος είναι η ικανότητά τους να προτείνουν στους χρήστες προϊόντα όχι μόνο που δεν ήξεραν (novelty), αλλά και που πιθανότατα δεν θα έβρισκαν μόνοι τους (serendipity). Για παράδειγμα, αν σε έναν χρήστη αρέσει κάποιος συγκεκριμένος συγγραφέας, τότε το σύστημα προτάσεων μπορεί να του προτείνει ένα νέο βιβλίο του συγκεκριμένου συγγραφέα που ο χρήστης δεν έχει ακόμα δει. Αυτή η πρόταση είναι μεν καινοτόμος, αλλά οι πιθανότητες λένε ότι αφού αυτός ο συγγραφέας είναι ο αγαπημένος του, ο χρήστης θα το έβρισκε και από μόνος του αυτό το βιβλίο κάποια στιγμή. Πιο αξιόλογο είναι το σύστημα να καταφέρει να προτείνει στον χρήστη ένα βιβλίο διαφορετικό από αυτά που αναμενόμενα του αρέσουν, και να το βρει τελικά ενδιαφέρον ο χρήστης, αυτή η περίπτωση καλύπτει την έννοια του serendipity. Ερευνητές μελετούν το πώς να προσαρμόζουν τους αλγορίθμους των συστημάτων ώστε να προσφέρουν και novelty και serendipity, ο υπολογισμός όμως του βαθμού καινοτομίας είναι ακόμα κάτι δύσκολο, καθώς απαιτεί μελέτες χρηστών στις οποίες οι συμμετέχοντες ζωντανά δηλώνουν αν το προϊόν που τους συστήθηκε είναι καινοτόμο.
- **Κάλυψη (Coverage) :** Η κάλυψη αφορά το ποσοστό των προϊόντων που είναι γνωστά στο σύστημα συνεργατικού φιλτραρίσματος για τα οποία το σύστημα μπορεί να εξάγει πρόβλεψη. Μπορούν να υπολογιστούν και διάφορες μορφοποιήσεις αυτού του μέτρου, όπως το ποσοστό των προϊόντων που έχουν προοπτικές για να προταθούν στους χρήστες, καθώς διάφορες βελτιστοποιήσεις στα συστήματα μπορεί να αποτρέψουν κάποια προϊόντα από το να προταθούν ποτέ.
- **Ρυθμός μάθησης (Learning Rate) :** Αυτό το μέτρο αξιολόγησης αφορά το πόσο γρήγορα το σύστημα γίνεται ένα αποδοτικό σύστημα εξαγωγής προβλέψεων, από την στιγμή που δεδομένα αρχίζουν να αποθηκεύονται στο σύστημα. Μετριέται δηλαδή για τον κάθε χρήστη ο αριθμός των βαθμολογιών που πρέπει να δώσει ώστε να αρχίσουν να του γίνονται εξατομικευμένες και έγκυρες προβλέψεις για νέα προϊόντα.
- **Εμπιστοσύνη (Confidence) :** Η εμπιστοσύνη περιγράφει την ικανότητα ενός συστήματος να αξιολογεί την ποιότητα των προβλέψεών του. Τα περισσότερα συστήματα με

συνεργατικό φιλτράρισμα παράγουν ταξινομήσεις με βάση τις πιο πιθανές βαθμολογήσεις που προβλέφθηκαν. Ένα collaborative-filtering σύστημα που μπορεί με ακρίβεια να υπολογίσει την εμπιστοσύνη του στις προβλέψεις που εξάγει, έχει το πολύ σημαντικό προτέρημα να περιορίζει τις συστάσεις του μόνο σε εκείνες που η πρόβλεψή τους είχε υψηλό βαθμό εμπιστοσύνης. Αν μπορεί να υπολογιστεί λοιπόν η εμπιστοσύνη για την κάθε πρόβλεψη, τότε το σύστημα μπορεί να την εμφανίζει στους χρήστες του για να τους βοηθήσει να αποφασίσουν αν αξίζει το ρίσκο αγοράς του προϊόντος.

- **Μέτρα ικανοποίησης χρηστών (User Satisfaction Metrics) :** Η ικανοποίηση των χρηστών από τις συστάσεις που κάνει το σύστημα μπορεί να καταγραφεί με μια επισκόπηση των χρηστών ή ελέγχοντας τη μνήμη του συστήματος και τα στατιστικά χρήσης.
- **Μέτρα απόδοσης ιστοσελίδων (Site performance metrics) :** Σε αντίθεση με τα προηγούμενα μέτρα αξιολόγησης των συστημάτων προτάσεων με συνεργατικό φιλτράρισμα που περιγράψαμε, τα οποία υπολογίζονται εκτός σύνδεσης, κάποιες ιστοσελίδες χρησιμοποιούν on-line απλά μέτρα ανάλυσης όταν στη σελίδα τους προστίθεται ένα νέο σύστημα προτάσεων ή μορφοποιείται ένα παλιότερο. Τέτοια μέτρα είναι η καταγραφή της αύξησης ή της μείωσης των προϊόντων που ο χρήστης αγόρασε ή «κατέβασε»,η αύξηση ή μείωση των συνολικών χρηστών που προσέρχονται στο σύστημα κ.ά.

Γενικά, είναι καλύτερα να επιλέγεται μια ακολουθία μέτρων που θα αξιολογήσουν τα πιο σημαντικά κριτήρια για την αποδοτική λειτουργία του εκάστοτε συστήματος συνεργατικού φιλτραρίσματος.

5.5 Προβλήματα memory-based Συστημάτων Συνεργατικού Φιλτραρίσματος

Τα neighbor-based συστήματα συνεργατικού φιλτραρίσματος παρουσιάζουν ορισμένα σημαντικά μειονεκτήματα :

- **Περιορισμένη κάλυψη :** Επειδή τα συστήματα αυτά βασίζονται στις ομοιότητες μεταξύ χρηστών μέσω τις σύγκρισης των βαθμολογιών τους για τα ίδια προϊόντα, δυο χρήστες μπορούν να θεωρηθούν γείτονες από το σύστημα μόνο αν έχουν βαθμολογήσει τα ίδια προϊόντα. Το γεγονός αυτό είναι αρκετά περιοριστικό, καθώς υπάρχει και η περίπτωση χρήστες που δεν έχουν βαθμολογήσει κοινά προϊόντα ή έχουν βαθμολογήσει μόνο λίγα να έχουν παρόλα αυτά κοινά ενδιαφέροντα. Επιπλέον, επειδή το σύστημα είναι έτσι οργανωμένο ώστε να μπορεί να συστήνει σε έναν χρήστη μόνο προϊόντα που έχουν βαθμολογήσει οι γείτονές του αλλά όχι ο ίδιος, υπάρχει περιορισμός και στην κάλυψη των προϊόντων που προτείνονται.
- **Αραιά δεδομένα (sparsity) :** Ένα άλλο σημαντικό πρόβλημα των CF neighbor-based συστημάτων είναι η έλλειψη διαθέσιμων βαθμολογιών. Οι χρήστες σε γενικές γραμμές βαθμολογούν μόνο ένα μικρό ποσοστό όλων των διαθέσιμων προϊόντων που υπάρχουν αποθηκευμένα στο σύστημα. Έτσι λοιπόν, το σύστημα προτάσεων έχει να προβλέψει μεγαλύτερο αριθμό βαθμολογιών από εκείνες που υπάρχουν καταχωρημένες από τους χρήστες, γεγονός που καθιστά δύσκολη την εύρεση όμοιων χρηστών. Κάποιος χρήστης

συνεπώς που είναι «δύσκολος» ή έχει πολύ ιδιαίτερα γούστα δεν θα μπορέσει να πάρει αξιόλογες προτάσεις. Επίσης ένας χρήστης που δεν έχει βαθμολογήσει πολλά προϊόντα, είναι δύσκολο να συνδεθεί με όμοιους χρήστες. Με αραιά δεδομένα, δυο χρήστες ή προϊόντα είναι αρκετά απίθανο να μοιράζονται κοινές βαθμολογίες, με αποτέλεσμα τα neighbor-based συστήματα να παράγουν προβλέψεις χρησιμοποιώντας μόνο έναν περιορισμένο αριθμό γειτόνων.

- **Κλίση προς τις δημοφιλείς επιλογές :** Επειδή συνήθως τις περισσότερες βαθμολογήσεις τις παίρνουν τα πιο δημοφιλή προϊόντα που είναι ευρέως γνωστά, τα συνεργατικά συστήματα προτάσεων είναι προκατειλημμένα προς αυτά τα προϊόντα. Αν για παράδειγμα μια ταινία έχει βαθμολογηθεί μόνο λίγες φορές, τότε το σύστημα θα την σύστηνε μόνο σε σπάνιες περιπτώσεις, επειδή η βαθμολογία που προβλέφθηκε μπορεί να μην είναι αξιόπιστη.
- **Cold-start πρόβλημα :** Ένα μεγάλο πρόβλημα των συστημάτων προτάσεων με Collaborative Filtering είναι το πρόβλημα της αρχής, όπου νέοι χρήστες εν έχουν ακόμα βαθμολογήσει κανένα προϊόν και νέα προϊόντα δεν έχουν ακόμη πάρει καμία βαθμολογία από τους χρήστες. Το σύστημα προτάσεων, για να παραγάγει εξατομικευμένες συστάσεις στους χρήστες του, πρέπει να ξέρει τι άρεσε στον χρήστη-στόχο στο παρελθόν, έτσι ώστε να αποφασίσει στη συνέχεια ποιοι χρήστες του συστήματος μοιάζουν με τον χρήστη-στόχο, με την έννοια ότι τους άρεσαν ή δεν τους άρεσαν τα ίδια προϊόντα στο παρελθόν. Προκύπτει λοιπόν ότι χρήστες που μόλις ξεκίνησαν να χρησιμοποιούν το σύστημα προτάσεων αντιμετωπίζουν πρόβλημα, γιατί το σύστημα δεν ξέρει τίποτα για την συμπεριφορά και τις προτιμήσεις τους, ενώ προϊόντα που είναι καινούρια στο σύστημα δεν μπορούν να συμπεριληφθούν στις συστάσεις που γίνονται γιατί κανείς δεν τα έχει βαθμολογήσει ακόμη.

5.6 Τρόποι αντιμετώπισης των προβλημάτων του Συνεργατικού Φιλτραρίσματος

Τα Collaborative Filtering Συστήματα Συστάσεων προσπαθούν να αντιμετωπίσουν τα προβλήματα που αναφέραμε παραπάνω με ποικίλους τρόπους, οι οποίοι θα εξεταστούν στη συνέχεια.

5.6.1 Λύσεις σε limited coverage – sparsity προβλήματα [22], [23]

Τα CF Recommender Systems αντιμετωπίζουν προβλήματα σχετικά με περιορισμένη κάλυψη και αραιά βαθμολογικά δεδομένα στην μήτρα χρηστών-προϊόντων. Τα προβλήματα αυτά μπορούν να αντιμετωπιστούν με μεθόδους μείωσης διαστάσεων (Dimensionality Reduction). Αυτές οι τεχνικές προβάλουν τους χρήστες και τα προϊόντα σε μια μήτρα μειωμένων διαστάσεων στο χώρο (reduced latent space) , η οποία συμπεριλαμβάνει τα πιο βασικά χαρακτηριστικά τους. Με αυτόν τον τρόπο, σε αυτόν τον χώρο πυκνών χαρακτηριστικών, μπορούν να βρεθούν σχέσεις ακόμα και ανάμεσα σε χρήστες που δεν έχουν βαθμολογήσει τα ίδια προϊόντα. Η μείωση διαστάσεων γίνεται είτε στην μήτρα βαθμολογιών χρηστών-προϊόντων, είτε στη μήτρα αποθήκευσης των ομοιοτήτων. Μια μέθοδος που χρησιμοποιείται ευρέως σε περιπτώσεις ανάκτησης πληροφορίας (IR) είναι η Latent Semantic Indexing. Η τεχνική LSI χρησιμοποιείται για να κατασκευάσει δυο μήτρες μειωμένων

διαστάσεων, μια χρηστών και μια προϊόντων. Αυτές οι μήτρες παρουσιάζουν τα χαρακτηριστικά των χρηστών και προϊόντων και προσπαθούν να βρουν σχέσεις ανάμεσα σε ζεύγη χρηστών από τις βαθμολογίες που έχουν δώσει στα προϊόντα. Με μείωση των διαστάσεων του χώρου των προϊόντων, μπορεί να αυξηθεί η πυκνότητα και έτσι να βρεθούν τελικά περισσότερες «κρυμμένες» βαθμολογίες.

Η Singular Value Decomposition (SVD) είναι μια γνωστή μέθοδος παραγοντοποίησης μητρών η οποία παραγοντοποιεί μια m x n μήτρα R σε τρεις μήτρες όπως ακολουθεί :

$$R = U \cdot S \cdot V'$$

Όπου U και V είναι δυο ορθογώνιες μήτρες διαστάσεων m x r και n x r αντίστοιχα, r είναι η τάξη της μήτρας R και S είναι μια διαγώνια μήτρα μεγέθους r x r που περιλαμβάνει όλες τις μοναδικές τιμές της μήτρας R ως στοιχεία της διαγωνίου της. Όλα τα στοιχεία της μήτρας S είναι θετικά και αποθηκεύονται κατά φθίνουσα σειρά του μεγέθους τους. Οι μήτρες που παράγει η Singular Value Decomposition είναι μια προσέγγιση της αρχικής R. Είναι δυνατόν να μειωθεί η r x r μήτρα S ώστε να έχει μόνο τις κ μεγαλύτερες τιμές τις διαγωνίου και να γίνει S_k, $k < r$. Αν μειωθούν αντίστοιχα και οι μήτρες U και V , τότε η ανακατασκευασμένη μήτρα $R_k = U_k \cdot S_k \cdot V_k'$ είναι η κοντινότερη μήτρα κ-τάξης στην R. Η μήτρα R_k ελαχιστοποιεί τη νόρμα $\|R - Rk\|$ σε όλες τις μήτρες τάξης κ.

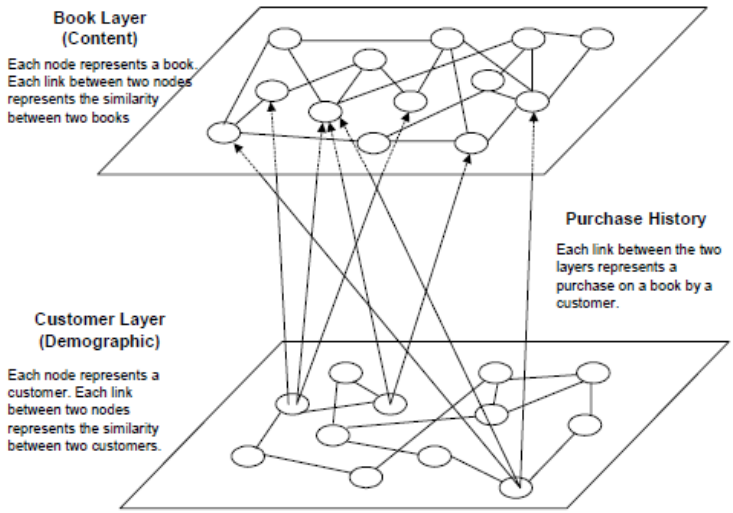
Στα συστήματα προτάσεων χρησιμοποιείται τη μέθοδο SVD για δυο σκοπούς. Πρώτον, για να ανακαλύφθούν οι κρυμμένες σχέσεις ανάμεσα σε πελάτες και προϊόντα που επιτρέπουν να υπολογίστεί η προβλεπόμενη βαθμολογία που θα βάλει ο πελάτης σε ένα προϊόν. Δεύτερον, με την τεχνική SVD παράγεται μια αναπαράσταση μικρών διαστάσεων του αρχικού χώρου χρηστών-προϊόντων και στη συνέχεια μπορούν να βρεθούν οι γειτονικές σχέσεις στον χώρο μειωμένων διαστάσεων, αφού το γεγονός ότι η μικρότερων διαστάσεων αναπαράσταση έχει λιγότερο αραιά βαθμολογικά δεδομένα από την αρχική μήτρα μεγάλων διαστάσεων δείχνει ότι συμφέρει να υπολογιστούν οι γείτονες σε αυτόν τον χώρο.

Ο μειωμένος «λανθάνων» χώρος λοιπόν προσπαθεί να εξηγήσει τις βαθμολογίες, χαρακτηρίζοντας χρήστες και προϊόντα από όσα συμπεραίνονται από την ανατροφοδότηση των χρηστών. Ωστόσο, η εφαρμογή της μεθόδου SVD στα άμεσα δεδομένα (explicit ratings) παρουσιάζει συχνά προβλήματα λόγω του υψηλού ποσοστού τιμών που λείπουν. Για αυτόν τον λόγο, πρόσφατες μελέτες έχουν δείξει ότι η ενσωμάτωση άλλων πηγών ανατροφοδοτήσεων των χρηστών, όπως η έμμεση (παρελθόν διαδικτυακών αγορών, πλοήγηση σε ιστοσελίδες κ.ά.), αυξάνει την ακρίβεια στην πρόβλεψη, και έτσι δημιουργήθηκε το εξελιγμένο μοντέλο SVD++ .

Ένας άλλος τρόπος χειρισμού των δυο αυτών προβλημάτων είναι η χρήση μεθόδων με βάση τους γράφους (Graph-based methods). Σε αυτήν την προσέγγιση, τα δεδομένα αναπαρίστανται με τη μορφή γράφου όπου οι κόμβοι είναι οι χρήστες ή τα προϊόντα και οι ακμές αντιστοιχούν στις σχέσεις ή τις ομοιότητες ανάμεσα στους χρήστες και τα προϊόντα. Μπορεί επίσης να δοθεί ένα βάρος σε κάθε ακμή, όπως η βαθμολογία που αντιστοιχεί στην συγκεκριμένη ακμή. Σε αυτό το μοντέλο, οι κλασικές προσεγγίσεις προβλέπουν την βαθμολογία ενός χρήστη u σε ένα προϊόν i, στηριζόμενες μόνο τους κόμβους που συνδέονται απευθείας με τον u ή το i. Οι προσεγγίσεις όμως με βάση τους γράφους, επιτρέπει σε κόμβους που δεν συνδέονται απευθείας με μια ακμή να επηρεάζουν ο ένας τον άλλον, καθώς θεωρούν ότι η πληροφορία μεταφέρεται στις συνδεόμενες ακμές. Όσο μεγαλύτερο το βάρος μιας ακμής, τόσο περισσότερη πληροφορία επιτρέπεται να περάσει μέσω αυτής. Επίσης, η επιρροή ενός κόμβου σε έναν άλλον μικραίνει όσο πιο απομακρυσμένοι είναι οι κόμβοι μεταξύ τους. Η εκμετάλλευση των μεταβατικών σχέσεων των δεδομένων σε έναν γράφο μειώνει τα προβλήματα της περιορισμένης κάλυψης και των αραιών

δεδομένων, αφού πλέον μπορούν να εκτιμηθούν σχέσεις ανάμεσα σε χρήστες ή προϊόντα που δεν συνδέονται άμεσα μεταξύ τους.

Αυτές οι μεταβατικές σχέσεις χρησιμοποιούνται για την σύσταση προϊόντων με δυο τρόπους. Στη μια περίπτωση, η εγγύτητα ενός χρήστη σε έναν προϊόν στον γράφο χρησιμοποιείται απευθείας για να εκτιμηθεί η βαθμολογία που θα έδινε ο u στο i. Πρακτικά, αυτό σημαίνει ότι στον χρήστη προτείνονται τα προϊόντα εκείνα που του είναι πιο κοντά στον γράφο. Στην δεύτερη περίπτωση, ο αλγόριθμος λαμβάνει υπ' όψιν του την εγγύτητα δυο κόμβων χρηστών ή προϊόντων ως μέτρο ομοιότητας, και στη συνέχεια χρησιμοποιεί αυτήν την ομοιότητα σαν βάρη στη neighbor-based μέθοδο προτάσεων.



Εικόνα 11: Μοντέλο γράφου δυο επιπέδων που αναπαριστά τα βιβλία, τους πελάτες και τις αγορές σε μια ηλεκτρονική βιβλιοθήκη [23]

Υπάρχουν διάφοροι τρόποι υπολογισμού της ομοιότητας στους γράφους. Στην ομοιότητα με βάση το μονοπάτι, η απόσταση μεταξύ δυο κόμβων εκτιμάται ως συνάρτηση του αριθμού των μονοπατιών και του μήκους των μονοπατιών που συνδέουν τους δυο κόμβους. Όταν εκτιμάται με ως συνάρτηση του αριθμού των μονοπατιών, η συσχέτιση ανάμεσα σε ένα χρήστη και κάποιο προϊόν ορίζεται ως το άθροισμα των βαρών όλων των διαφορετικών μονοπατιών που συνδέουν τους δυο κόμβους, των οποίων το μήκος δεν είναι μεγαλύτερο από μια δεδομένη μέγιστη τιμή K. Στην περίπτωση της συνάρτησης του μήκους των μονοπατιών, η ομοιότητα υπολογίζεται βάσει της μικρότερης απόστασής τους στον γράφο. Τα δεδομένα μοντελοποιούνται σε έναν κατευθυνόμενο γράφο με βάσει τις έννοιες horting και predictability. Horting είναι μια σχέση ανάμεσα σε δυο χρήστες που ικανοποιείται όταν οι χρήστες έχουν βαθμολογήσει όμοια προιόντα, ενώ predictability είναι μια ισχυρότερη έννοια, που απαιτεί οι βαθμολογήσεις των δυο χρηστών να είναι παρόμοιες, αφού έχει θεωρηθεί πρώτα και η διαφορά στην βαθμολογική κλίμακα των δυο χρηστών. Άλλη μέθοδος υπολογισμού της ομοιότητας είναι η «τυχαία βόλτα» (random walk similarity), στην οποία οι μεταβατικές σχέσεις στον γράφο ορίζονται σε ένα πιθανοκρατικό πλαίσιο. Η ομοιότητα ανάμεσα σε δυο κόμβους σε αυτήν την προσέγγιση εκτιμάται ως η πιθανότητα να συναντήσει κανείς αυτούς τους κόμβους σε έναν τυχαίο περίπατο. Αυτό μαθηματικοποιείται με μια πρώτης τάξης αλυσίδα Markov και μια μήτρα πιθανοτήτων.

5.6.2 Λύσεις σε cold-start πρόβλημα

Το cold-start πρόβλημα είναι ένα πολύ σύνηθες πρόβλημα των Collaborative Filtering Συστημάτων Προτάσεων. Σχετίζεται με το γεγονός ότι καινούριοι χρήστες ή προϊόντα δεν έχουν ακόμα βαθμολογήσεις, με αποτέλεσμα τα μεν καινούρια προϊόντα να μην μπορούν να προταθούν, και οι δε νέοι χρήστες να μην μπορούν να λάβουν προβλέψεις αφού το σύστημα δεν γνωρίζει ακόμα τίποτα για αυτούς.

Μια απλή λύση στο πρόβλημα αυτό είναι να γεμίσει το σύστημα τις βαθμολογίες που λείπουν με προεπιλεγμένες τιμές, όπως τον μέσο του βαθμολογικού εύρους ή τον μέσο όρο βαθμολογίας χρηστών ή προϊόντων. Μι άλλη τεχνική είναι να χρησιμοποιούνται πληροφορίες περιεχομένου για να γεμίσουν οι τιμές που λείπουν. Επίσης, μπορεί να μειωθεί η έκταση του προβλήματος με την χρήση ενός υβριδικού μοντέλου συνεργατικού φιλτραρίσματος και περιεχομένου. Για να βρει λοιπόν ένα σύστημα τους γείτονες, μπορεί εκτός από την ομοιότητα συσχέτισης να χρησιμοποιήσει και την content similarity. Έτσι, η ομοιότητα ανάμεσα σε προϊόντα για παράδειγμα μπορεί να υπολογιστεί με βάση τα χαρακτηριστικά του περιεχομένου των προϊόντων, όπως κάποια περιγραφή ή κάποια ονομαστικά χαρακτηριστικά. Αυτές οι προτάσεις ως λύσεις εμφανίζουν κάποια μειονεκτήματα. Η χρήση προεπιλεγμένων τιμών για να γεμίσουν τα δεδομένα χρηστών-προϊόντων μπορεί να οδηγήσει στην ύπαρξη προκατάληψης στις προβλέψεις που θα γίνουν, αφού οι default τιμές δημιουργούν μια συγκεκριμένη προδιάθεση που μπορεί να μην είναι αληθής. Επιπλέον, πολλές φορές μπορεί να μην υπάρχει διαθέσιμο περιεχόμενο στα προϊόντα και έτσι να μην γίνεται να υπολογιστούν βαθμολογίες ή ομοιότητες.

Άλλοι τρόποι χειρισμού αυτού του προβλήματος είναι να αναγκάζει το σύστημα τους χρήστες να βαθμολογούν προϊόντα στην αρχή, όταν εγγράφονται στο σύστημα, ή να τους υποχρεώνει να απαντάνε κάποιες δημογραφικές ερωτήσεις (π.χ. για το φύλο, την ηλικία ή την χώρα προέλευσής τους) και να χρησιμοποιούν για το αρχικό μόνο στάδιο, μέχρι να δηλώσει ο χρήστης βαθμολογίες με τον καιρό, αυτές τις πληροφορίες για συστάσεις, οι οποίες βέβαια δεν θα είναι τόσο αξιόπιστες καθώς δεν θα είναι ακόμη εξατομικευμένες. Αρνητικές διαστάσεις αυτών των προσεγγίσεων είναι πως απαιτούν προσπάθεια και χρόνο από τον χρήστη για να απαντήσει τις ερωτήσεις αυτές. Επιπλέον, δεν υπάρχουν κριτήρια για το ποιες είναι οι σωστές επιλογές προϊόντων που θα βάλει το σύστημα τον χρήστη να βαθμολογήσει. Η δε χρήση δημογραφικών ερωτήσεων οδηγεί αναγκαστικά το σύστημα στο να στηριχτεί σε στερεότυπα για την παραγωγή των αρχικών συστάσεων, όπως ότι στις γυναίκες αρέσουν συγκεκριμένα προϊόντα που προτείνει, ή οι ηλικιωμένοι προτιμούν κάποια άλλα, γεγονότα και κατηγοριοποιήσεις που μπορεί να μην ισχύουν και να είναι προσβλητικά για τους χρήστες.

Μια λύση που οδηγεί σε σωστότερα αποτελέσματα από τις παραπάνω είναι οι ομαδικές συστάσεις (Group Recommendation Systems), στις οποίες παίζουν ρόλο πολλαπλά κριτήρια ώστε να ικανοποιηθούν όλα τα μέλη της ομάδας. Η μέθοδος των ομαδικών συστάσεων περιορίζει το cold-start πρόβλημα, παρέχοντας σε έναν νέο χρήστη χωρίς βαθμολογικά δεδομένα συστάσεις που θα έκαναν όλη την ομάδα των ήδη υπαρχόντων χρηστών χαρούμενη, θα ικανοποιούσαν δηλαδή τις συνολικές ανάγκες της ομάδας. Για αυτήν την διαδικασία γίνεται βέβαια η υπόθεση ότι ο νέος χρήστης θα ταιριάζει με κάποιον από τους ήδη υπάρχοντες, χωρίς να ξέρει ακόμα το σύστημα με ποιόν, και έτσι, συστήνοντας προϊόντα που καλύπτουν τα ενδιαφέροντα όλων, λογικά θα καλυφθεί και ο καινούριος χρήστης. Σταδιακά, το σύστημα θα μάθει για τις ατομικές προτιμήσεις του νέου χρήστη από τις βαθμολογίες που θα δώσει στα προϊόντα που του συστήθηκαν ή από έμμεση ανατροφοδότηση. Στις προτάσεις που γίνονται και αφορούν την κάλυψη όλης της ομάδας των

χρηστών, συμπεριλαμβανομένου και του νέου, στην αρχή στον καινούριο χρήστη θα αντιστοιχεί ένα μικρό βάρος, δηλαδή θα επηρεάζει ελάχιστα την επιλογή των προϊόντων που θα συστηθούν, καθώς το σύστημα δεν γνωρίζει ακόμα τίποτα για αυτόν. Σιγά-σιγά, με τις αλληλεπιδράσεις του χρήστη με το σύστημα το βάρος αυτό θα αυξάνεται, ενώ ταυτόχρονα θα μειώνεται το βάρος των παλιών χρηστών που έχει αρχίσει να φαίνεται ότι δεν ταιριάζουν στα ενδιαφέροντα με τον νέο.

Η προσέγγιση των ομαδικών συστάσεων είναι ακόμη καινούρια μέθοδος και έχει διάφορα ανεξερεύνητα πεδία. Για παράδειγμα, χρειάζεται ακόμα δουλειά στο να βρεθεί μια αποδοτική συνάρτηση ικανοποίησης, η οποία να προβλέπει πόσο ικανοποιημένο είναι το άτομο από την σειρά προϊόντων που του προτείνονται, δεδομένων των επιπλοκών που μπορεί να προκαλέσουν οι αλληλεπιδράσεις με την ομάδα. Επιπλέον, ένα άλλο θέμα στο οποίο υστερούν οι ομαδικές συστάσεις είναι η διατήρηση μιας ισορροπίας ανάμεσα στην προστασία των προσωπικών δεδομένων του κάθε μέλους της ομάδας και της διαφάνειας που πρέπει να έχει το σύστημα για λόγους εμπιστοσύνης. Συνολικά λοιπόν, ο τομέας των ομαδικών συστάσεων χρειάζεται ακόμα περαιτέρω έρευνα, η οποία είναι απαραίτητη καθώς τα ομαδικά συστήματα προτάσεων μπορούν να φανούν πολύ χρήσιμα σε ποικίλες περιπτώσεις, όπως σε τουριστικές ομάδες που κανονίζουν κάποιο ταξίδι, ή σε προγράμματα τηλεόρασης όπου η οικογένεια θέλει να παρακολουθήσει ένα πρόγραμμα που να τους ενδιαφέρει όλους.

Ωστόσο, η καλύτερη προσέγγιση που αντιμετωπίζει όλα τα προβλήματα που αναφέρθηκαν παραπάνω, δηλαδή και την περιορισμένη κάλυψη και τα αραιά δεδομένα και το πρόβλημα νέων χρηστών-νέων προϊόντων είναι τα Δίκτυα Εμπιστοσύνης, πάνω στα οποία θα γίνει μελέτη στην ενότητα που ακολουθεί.

5.7 Δίκτυα Εμπιστοσύνης (Trust Networks)

Έρευνες έχουν δείξει πως οι άνθρωποι τείνουν να στηρίζονται περισσότερο σε συστάσεις από ανθρώπους που εμπιστεύονται, όπως οι φίλοι τους, από ότι στα online συστήματα προτάσεων τα οποία τους συστήνουν προϊόντα με βάση χρήστες που είναι μεν όμοιοί τους αλλά τους είναι άγνωστοι. Αυτή η παρατήρηση, σε συνδυασμό με την ολοένα και αυξανόμενη δημοτικότητα των ανοιχτών κοινωνικών δικτύων, καθώς και την εισαγωγή εφαρμογών ηλεκτρονικού εμπορίου στα συστήματα προτάσεων, οδήγησε στο να στραφεί το ενδιαφέρον στα συστήματα προτάσεων με βάρος στον τομέα της εμπιστοσύνης. Οι προτάσεις που παράγουν αυτά τα συστήματα βασίζονται σε πληροφορία που προέρχεται από ένα διαδικτυακό δίκτυο εμπιστοσύνης του χρήστη, ένα κοινωνικό δίκτυο δηλαδή το οποίο εκφράζει πόσο πολύ τα μέλη της δικτυακής κοινότητας εμπιστεύονται το ένα το άλλο. Τα συστήματα προτάσεων με βάση την εμπιστοσύνη χρησιμοποιούν τη γνώση που προέρχεται από τα δίκτυα εμπιστοσύνης για να εξάγουν πιο εξατομικευμένες συστάσεις. Έτσι, οι χρήστες λαμβάνουν προτάσεις για προϊόντα τα οποία έχουν βαθμολογήσει με ψηλούς βαθμούς οι χρήστες που ανήκουν στον ιστό εμπιστοσύνης του (Web Of Trust ή WOT), ή ακόμα και από χρήστες τους οποίους εμπιστεύονται τα μέλη του Web of Trust του.

Υπάρχουν δύο βασικές κατηγορίες των μοντέλων εμπιστοσύνης, το πιθανοτικό μοντέλο (probabilistic) και η σταδιακή προσέγγιση (gradual approach). Στο πιθανοτικό μοντέλο, η εμπιστοσύνη έχει μόνο δυο τιμές (είτε εμπιστεύεσαι έναν χρήστη είτε όχι) και υπολογίζεται η πιθανότητα για κάποιον χρήστη να είναι άξιος εμπιστοσύνης. Η σταδιακή προσέγγιση ασχολείται με την εκτίμηση της τιμής εμπιστοσύνης, δεδομένου ότι ένας χρήστης μπορεί να εμπιστεύεται έναν άλλον σε κάποιο βαθμό, και όχι απόλυτα ή καθόλου. Αυτή η προσέγγιση αντιπροσωπεύει

περισσότερο την πραγματικότητα, στην οποία οι άνθρωποι εμπιστεύονται άλλους περισσότερο ή λιγότερο, και όχι πολύ ή καθόλου. Αν και η έννοια της εμπιστοσύνης έχει εδραιωθεί στα συστήματα, δεν μπορούμε να πούμε το ίδιο για την έννοια της δυσπιστίας, την οποία τα περισσότερα συστήματα δεν λαμβάνουν υπ' όψιν.

Επειδή ο ιστός εμπιστοσύνης ενός χρήστη συμπεριλαμβάνει μόνο ένα πολύ μικρό ποσοστό της διαδικτυακής κοινότητας, είναι πολύ χρήσιμο να μπορεί το σύστημα να αξιοποιεί τη γνώση και τη βαθμολογία ενός μεγαλύτερου υποσυνόλου του πληθυσμού των χρηστών για να εξάγει συστάσεις. Τα trust metrics υπολογίζουν μια εκτίμηση για το πόσο πολύ ένας χρήστης πρέπει να εμπιστεύεται έναν άλλον, με βάση τις υπάρχουσες σχέσεις εμπιστοσύνης ανάμεσα στους χρήστες του δικτύου.

Πριν γίνει η ανάλυση των συστημάτων προτάσεων με ενσωματωμένο τον παράγοντα της εμπιστοσύνης, χρήσιμο θα ήταν να γίνει αναφορά πρώτα στις έννοιες της διάδοσης (propagation) και της άθροισης (aggregation). Η έννοια της διάδοσης στηρίζεται στην υπόθεση ότι η εμπιστοσύνη είναι κατά μια έννοια μεταδοτική. Αυτό σημαίνει ότι αν ο χρήστης Α εμπιστεύεται τον Β και ο χρήστης Β εμπιστεύεται τον Γ, τότε κατά πάσα πιθανότητα και ο Α θα εμπιστεύεται τον Γ σε έναν βαθμό, δεδομένου ότι η εμπιστοσύνη του Α στον Β και εκείνη του Β στον Γ αναφέρονται στο ίδιο περιεχόμενο. Ένα μονοπάτι διάδοσης μπορεί να θεωρηθεί σαν μια μεταβατική αλυσίδα κομματιών εμπιστοσύνης. Η έννοια της διάδοσης περιπλέκεται όταν εισάγεται στην αλυσίδα και η δυσπιστία. Στην περίπτωση της άμεσης διάδοσης, όπως αναφέρθηκε προηγουμένως αν ο Α εμπιστεύεται τον Β και ο Β τον Γ, τότε σε έναν βαθμό ο Α θα εμπιστεύεται τον Γ. Αν τώρα ο Α εμπιστεύεται τον Β και ο Β δεν εμπιστεύεται τον Γ, τότε λογικά ούτε ο Α θα εμπιστεύεται τον Γ. Τι γίνεται όμως όταν η δυσπιστία εισάγεται στο πρώτο κομμάτι της αλυσίδας διάδοσης; Τότε υπάρχουν τρεις εκδοχές. Αν ο Α δεν εμπιστεύεται τον Β και ο Β δεν εμπιστεύεται τον Γ, τότε η μια περίπτωση είναι ο Α να εμπιστεύεται τον Γ, με την εξήγηση πως άτομα τα οποία δεν τα εμπιστεύονται χρήστες που δεν εμπιστεύεται ο ίδιος, είναι προτιμότερο να τα εμπιστευτεί εκείνος (πολλαπλασιαστική διάδοση μη-εμπιστοσύνης). Η δεύτερη εκδοχή είναι ότι ο Α δεν θα εμπιστεύεται τον Γ, γιατί αν έναν χρήστη δεν τον εμπιστεύεται το άτομο που δεν εμπιστεύεται εκείνος, τότε σίγουρα δεν πρέπει να το εμπιστεύεται και ο ίδιος (προσθετική διάδοση μη-εμπιστοσύνης). Μια άλλη εκδοχή της περίπτωσης αυτής είναι ότι δεν γίνεται να εξαχθεί συμπέρασμα για την εμπιστοσύνη του Α στον Γ, γιατί ο Α δεν πρέπει να λαμβάνει υπ' όψιν του καμία πληροφορία από ένα άτομο που δεν εμπιστεύεται. Έστω το ζευγάρι (ε, δ) που δείχνει τον βαθμό εμπιστοσύνης ε και δυσπιστίας δ, στο εύρος του [0,1] και (ε1, δ1) είναι ο βαθμός που ο χρήστης Α εμπιστεύεται τον Β και (ε2, δ2) είναι ο βαθμός στον οποίο ο χρήστης Β εμπιστεύεται τον Γ. Η τεχνική διάδοσης που στηρίζεται στη φιλοσοφία του να ακούς αυτούς που εμπιστεύεσαι και να αγνοείς τρίτους που δεν εμπιστεύεσαι υπολογίζει τον βαθμό εμπιστοσύνης (ε3, δ3) του χρήστη Α στον Γ ως :

$(\varepsilon 3, \delta 3) = (\varepsilon 1 \times \varepsilon 2, \delta 1 \times \delta 2)$.

Στον πίνακα που ακολουθεί υπάρχουν παραδείγματα εξαγόμενων τιμών εμπιστοσύνης. Η κάθε γραμμή αντιστοιχεί στην τιμή εμπιστοσύνης που δείχνει ο Α στον Β, η κάθε στήλη στην εμπιστοσύνη του Β στον Γ και τα αντίστοιχα κελιά του πίνακα περιέχουν την διαδιδόμενη τιμή εμπιστοσύνης του Α στον Γ για την εκάστοτε περίπτωση.

Πίνακας 14 : Πίνακας διαδιδομένων τιμών εμπιστοσύνης με την τεχνική 1

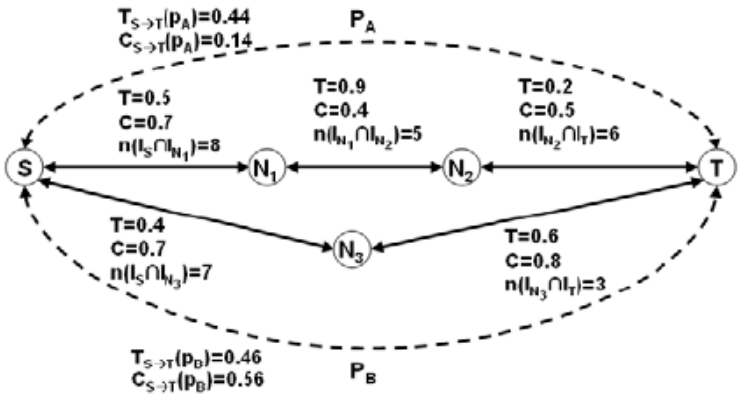
	(0 , 0)	**(0 , 1)**	**(1 , 0)**
(0 , 0)	**(0 , 0)**	**(0 , 0)**	**(0 , 0)**
(0 , 1)	**(0 , 0)**	**(0 , 0)**	**(0 , 0)**
(1 , 0)	**(0 , 0)**	**(0 , 1)**	**(1 , 0)**

Με βάση αυτήν την προσέγγιση, αν ο Α δεν εμπιστεύεται τον Β (δηλαδή βαθμός (0, 1)), τότε δεν μπορεί να εξαχθεί συμπέρασμα για τον βαθμό εμπιστοσύνης του Α στον Γ (ο τύπος υπολογίζει (0, 0)), είτε ο Β εμπιστεύεται τον Γ είτε δεν τον εμπιστεύεται. Η τεχνική διάδοσης που στηρίζεται στην ιδέα ότι ο χρήστης που δεν εμπιστευόμαστε δίνει λάθος πληροφορία επί τούτου υπολογίζει την εμπιστοσύνη (ε3, δ3) του Α στον Γ με πιθανοτικά αθροίσματα ως :
(ε3, δ3) = (ε1 x ε2 + δ1 xδ2 – ε1 x ε2 x δ1 x δ2, ε1 x δ2 + δ1 x ε2 – ε1 x δ2 x δ1 x ε2).

Πίνακας 15 : Πίνακας διαδιδομένων τιμών εμπιστοσύνης με την τεχνική 2

	(0 , 0)	**(0 , 1)**	**(1 , 0)**
(0 , 0)	**(0 , 0)**	**(0 , 0)**	**(0 , 0)**
(0 , 1)	**(0 , 0)**	**(1 , 0)**	**(0 , 1)**
(1 , 0)	**(0 , 0)**	**(0 , 1)**	**(1 , 0)**

Η έννοια της άθροισης σχετίζεται με το γεγονός ότι σε μεγάλα δίκτυα, κατά κύριο λόγο θα υπάρχουν όχι ένα αλλά πολλά μονοπάτια που θα οδηγούν στον χρήστη για τον οποίο θέλουμε να υπολογιστεί η αξιοπιστία του. Αυτή η τεχνική συναθροίζει όλους τους βαθμούς εμπιστοσύνης που έχουν δοθεί σε έναν χρήστη από όλες του τις αλληλεπιδράσεις, και υπολογίζει έναν συνολικό βαθμό αξιοπιστίας. Αυτή η καθολική τιμή συχνά ονομάζεται και «φήμη».



Εικόνα 12 : Παράδειγμα άθροισης πολλαπλών μονοπατιών εμπιστοσύνης [24]

Στο παράδειγμα της εικόνας 12, ο χρήστης S συνδέεται με τον χρήστη T μέσω δυο διαφορετικών μονοπατιών Pa και Pb. Το μονοπάτι Pa περνάει από τους χρήστες N1 και N2 και το μονοπάτι Pb από τον χρήστη N3. Η εμπιστοσύνη που συμπεραίνεται σε καθένα από αυτά τα μονοπάτια εμπιστοσύνης είναι ανεξάρτητη η μια από την άλλη. Το μοντέλο της άθροισης συνδυάζει τις τιμές που συμπεραίνονται από τα διάφορα μονοπάτια εμπιστοσύνης σε μια μοναδική τιμή. Αυτό γίνεται με δυο τρόπους, είτε με την Μέση σύνθεση, είτε με τη Σταθμισμένη Μέση σύνθεση :

- *Μέση Σύνθεση* : Υπολογίζεται ο μέσος όρος όλων των τιμών εμπιστοσύνης που βγαίνουν από το κάθε ξεχωριστό μονοπάτι μέσω της εξίσωσης :

$$T_{S\to T} = \frac{\sum_{i=1}^{p} T_{S\xrightarrow{P_i}T}}{p}$$

[24]

Όπου p είναι ο αριθμός των διαφορετικών μονοπατιών εμπιστοσύνης. Αυτή η προσέγγιση είναι αποδοτική, ωστόσο δεν λαμβάνει υπ' όψιν της τον βαθμό εμπιστοσύνης του κάθε μονοπατιού.

- *Σταθμισμένη Μέση Σύνθεση :* Σε αυτήν την περίπτωση, υπολογίζεται ο σταθμισμένος μέσος όρος των τιμών εμπιστοσύνης που συμπεραίνονται από τα διάφορα μονοπάτια, χρησιμοποιώντας για βάρη την διαδιδόμενη εμπιστοσύνη της κάθε έμμεσης συσχέτισης μεταξύ των χρηστών S και T. Η διαδιδόμενη εμπιστοσύνη ενός μονοπατιού που συνδέει τους χρήστες S και T με ενδιάμεσους κόμβους τους χρήστες N = {Ni : i=1,2,….,k} δίνεται από τον τύπο :

$$C_{S\xrightarrow[N_1\to\ldots\to N_k]{}T} = \left(\left(\left(\left(C_{S\to N_1}\cdot C_{I_1\to N_2}\right)\cdot\ldots\right)\cdot C_{N_{k-1}\to N_k}\right)\cdot C_{N_k\to T}\right)$$

[24] ,

ενώ η σταθμισμένη μέση τιμή εμπιστοσύνης υπολογίζεται ως :

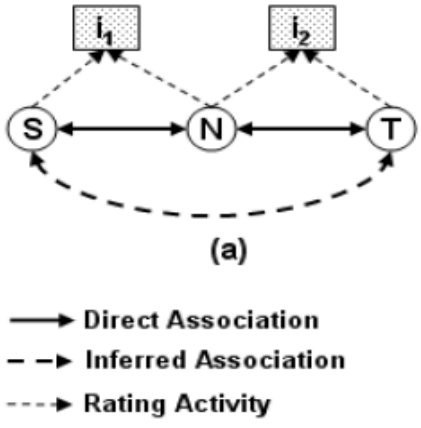
$$T_{S\to T} = \sum_{i=1}^{p} \frac{C_{S\xrightarrow{P_i}T}}{\sum_{i=1}^{p} C_{S\xrightarrow{P_i}T}} T_{S\xrightarrow{P_i}T}$$

[24]

Αυτή η προσέγγιση είναι σωστότερη, γιατί δίνεται βάρος και στον βαθμό εμπιστοσύνης που συμπεραίνεται από το κάθε μονοπάτι. Η τελική μοναδική τιμή εμπιστοσύνης κλίνει περισσότερο στην εμπιστοσύνη που συσχετίζεται με το πιο αξιόπιστο μονοπάτι.

5.7.1 Συστήματα Προτάσεων που εμπεριέχουν τον παράγοντα της εμπιστοσύνης

Υπάρχουν περιπτώσεις στις οποίες η έλλειψη ή η ύπαρξη λίγης πληροφορίας και βαθμολογικών δεδομένων μπορεί να αποβεί επιβλαβής για το σύστημα προτάσεων. Μπορεί για παράδειγμα οι συσχετίσεις ανάμεσα στους χρήστες να στηρίζονται σε ελάχιστα δεδομένα ή μπορεί να μην είναι εφικτό να βρεθούν κ γείτονες για να εφαρμοστεί ο k-nearest neighbors αλγόριθμος. Ένα χαρακτηριστικό παράδειγμα φαίνεται στην εικόνα 13.



Εικόνα 13 : Έμμεση συσχέτιση χρηστών μέσω εμπιστοσύνης [24]

Έστω ότι οι χρήστες S και N έχουν βαθμολογήσει το προϊόν I1 , ενώ οι χρήστες N και T έχουν βαθμολογήσει το I2. Το κλασσικό Συνεργατικό Φιλτράρισμα θα συσχετίσει τους χρήστες S και N και τους χρήστες N και T, αλλά δεν θα μπορέσει να δει τη σύνδεση μεταξύ S και T. Έτσι, το προϊόν I2 θα μπορούσε να προταθεί στον χρήστη S , αν όμως ο χρήστης T είχε βαθμολογήσει και ένα άλλο προϊόν I3, τότε δεν θα υπήρχε τρόπος να βρει το σύστημα αν το προϊόν I3 είναι μια καλή πρόταση για τα ενδιαφέροντα του χρήστη S. Η κατάσταση αυτή όμως μπορεί να αλλάξει αν θεωρηθεί μια πιο εξελιγμένη προσέγγιση, η οποία λαμβάνει υπ' όψιν της τις μεταβατικές αλληλεπιδράσεις , αντιλαμβάνεται την συνειρμική συσχέτιση μεταξύ των χρηστών S και T και συμπεραίνει αυτήν την έμμεση σύνδεση. Η προσέγγιση αυτή είναι να εδραιωθεί ανάμεσα στους χρήστες του συστήματος προτάσεων ένα δίκτυο εμπιστοσύνης. Έτσι, μόλις βρεθεί ένα μονοπάτι που οδηγεί στον χρήστη-στόχο στον οποίο ζητείται να γίνει η σύσταση, το σύστημα μπορεί να συνδυάσει την κρίση του χρήστη-στόχου με τις διαθέσιμες πληροφορίες σχετικά με την εμπιστοσύνη, μέσω της διάδοσης και της άθροισης που εξηγήσαμε προηγουμένως, έτσι ώστε να εξάγει μια πιο εξατομικευμένη πρόβλεψη. Το δίκτυο εμπιστοσύνης επιτρέπει να υπάρχει πρόσβαση σε περισσότερους χρήστες και περισσότερα προϊόντα. Στο παράδειγμα της εικόνας 13 λοιπόν, είναι δυνατόν να συμπερανθεί δεσμός εμπιστοσύνης ανάμεσα στους χρήστες S και T μέσω του ενδιάμεσου χρήστη N. Οι συνεχόμενες γραμμές της εικόνας δείχνουν την άμεση συσχέτιση χρηστών, οι διακεκομμένες με bold την έμμεση συσχέτιση, και οι διακεκομμένες την βαθμολογική δραστηριότητα των χρηστών. Με αυτήν την διαδικασία, η εμπιστοσύνη μεταδίδεται στο δίκτυο και χτίζονται σχέσεις ανάμεσα σε χρήστες, ακόμα και αν αυτοί δεν έχουν βαθμολογήσει κοινά προϊόντα. Αν λοιπόν ο χρήστης S εκφράσει έναν βαθμό εμπιστοσύνης στον N και ο N δηλώσει ότι εμπιστεύεται τον T , τότε με propagation μπορεί να προκύψει μια ένδειξη της εμπιστοσύνης του S στον T. Αν το αποτέλεσμα δείξει ότι ο S πρέπει να εμπιστευτεί σε μεγάλο βαθμό τον T , τότε θα μπορούσε το προϊόν I3 που έχει βαθμολογήσει ο T να είναι μια καλή πρόταση και για τον S.

Φαίνεται με αυτόν τον τρόπο ότι αυξάνοντας τις συνδέσεις σε ένα σύστημα προτάσεων με Συνεργατικό Φιλτράρισμα, ενσωματώνοντας σχέσεις εμπιστοσύνης, μπορεί να λυθεί το πρόβλημα των αραιών δεδομένων. Επιπλέον, τα συστήματα προτάσεων που λαμβάνουν υπ' όψιν τον παράγοντα εμπιστοσύνης μπορούν να αμβλύνουν και το cold-start πρόβλημα. Για έναν νέο χρήστη, είναι πιο ωφέλιμο να κάνει κάποιες δηλώσεις εμπιστοσύνης στο σύστημα για να λάβει καλύτερες για εκείνον και πιο ακριβείς συστάσεις από ότι να βαθμολογήσει στην αρχή κάποια προϊόντα. Οι χρήστες, πρέπει να ενθαρρύνονται να συνδεθούν με άλλους χρήστες και να ευρύνουν το δίκτυο εμπιστοσύνης του όσο το δυνατόν νωρίτερα. Χαρακτηριστικό παράδειγμα είναι το σύστημα Golbeck's FilmTrust , το οποίο ζητά από τους χρήστες του να αξιολογήσουν τους χρήστες του συστήματος με βάση τις προτιμήσεις τους στις ταινίες, και στη συνέχεια χρησιμοποιεί αυτές τις αξιολογήσεις για να εξάγει ακριβείς εξατομικευμένες προβλέψεις.

Τα trust-enhanced συστήματα προτάσεων χωρίζονται σε δυο βασικές κατηγορίες, με βάση τον τρόπο με τον οποίο λαμβάνουν τις τιμές εμπιστοσύνης. Στην πρώτη κατηγορία ανήκουν συστήματα που χρησιμοποιούν πληροφορία του δικτύου εμπιστοσύνης που προέρχεται άμεσα από τους χρήστες του δικτύου, όπως από άμεσες δηλώσεις εμπιστοσύνης. Τα συστήματα αυτής της μεθόδου μπορούν να χρησιμοποιούν τις τεχνικές της διάδοσης και άθροισης εμπιστοσύνης στο δίκτυο, για να συμπεράνουν τις τελικές τιμές εμπιστοσύνης που χρειάζονται οι αλγόριθμοι συστάσεων. Από την άλλη πλευρά, υπάρχουν τα συστήματα που δεν απαιτούν από τους χρήστες προσπάθεια, δεν τους ζητάνε δηλαδή να κάνουν δηλώσεις για το πόσο αξιόπιστοι είναι κατά την γνώμη τους οι χρήστες του δικτύου. Οι τιμές εμπιστοσύνης υπολογίζονται αυτόματα, στηριζόμενες στην έμμεση ανατροφοδότηση των χρηστών, όπως στο κατά πόσο ο χρήστης είχε κάνει αξιόπιστες συστάσεις στο παρελθόν.

5.7.1.1 Συνεργατικό Φιλτράρισμα με βάση την Εμπιστοσύνη

Το συνεργατικό φιλτράρισμα συνδέεται άμεσα με τα Trust-enhanced συστήματα προτάσεων. Όπως έχει προαναφερθεί, στο συνεργατικό φιλτράρισμα προβλέπεται η βαθμολογία που θα έδινε ο χρήστης-στόχος σε ένα συγκεκριμένο προϊόν μέσω ενός συνδυασμού των βαθμολογιών που έχουν δώσει στο ίδιο προϊόν οι γείτονες του χρήστη-στόχου.

$$p_{a,i} = \overline{r}_a + \frac{\sum_{u \in R^+} w_{a,u}(r_{u,i} - \overline{r}_u)}{\sum_{u \in R^+} w_{a,u}}$$

[3, σελ.661]

Η προβλεπόμενη βαθμολογία $p_{a,i}$ του χρήστη a στο προϊόν i υπολογίζεται με χρήση της μέσης βαθμολογίας του a στα υπόλοιπα προϊόντα και στις βαθμολογίες $r_{u,i}$ των άλλων χρηστών u για το i. Επίσης, λαμβάνεται υπ' όψιν στον υπολογισμό η ομοιότητα $w_{a,u}$ ανάμεσα στους χρήστες a και u, η οποία πολύ συχνά βρίσκεται με τη μέθοδο της συσχέτισης Pearson που έχει αναλυθεί. Τις περισσότερες φορές, υπολογίζονται στην εξίσωση πρόβλεψης βαθμολογίας για το i μόνο οι χρήστες που έχουν θετική συσχέτιση $w_{a,u}$ με τον a. Το R^+ αντιπροσωπεύει αυτήν την ομάδα χρηστών.

Ωστόσο, οι αλγόριθμοι στα Trust-enhanced συστήματα προτάσεων χρησιμοποιούν τις εκτιμήσεις εμπιστοσύνης για βάρη στην διαδικασία πρόβλεψης. Έτσι, αντί για τον υπολογισμό των βαρών μέσω της συσχέτισης Pearson, συμπερένονται τα βάρη από τις σχέσεις του χρήστη-στόχου στο δίκτυο εμπιστοσύνης μέσω διάδοσης και άθροισης. Αυτό μπορεί να συμβεί, καθώς η έννοιες της εμπιστοσύνης και της ομοιότητας συνδέονται. Στην εξίσωση που ακολουθεί λοιπόν, έχουν αντικατασταθεί τα Pearson Correlation βάρη $w_{a,u}$ με τις τιμές εμπιστοσύνης $t_{a,u}$:

$$p_{a,i} = \overline{r_a} + \frac{\sum_{u \in R^T} t_{a,u}(r_{u,i} - \overline{r_u})}{\sum_{u \in R^T} t_{a,u}}$$

[3, σελ.661]

Αυτή η μεθοδολογία είναι το Συνεργατικό Φιλτράρισμα με βάση την εμπιστοσύνη. Επειδή τα νέα βάρη δεν είναι ίσα με τα βάρη της συσχέτισης Pearson, αυτή η τεχνική μπορεί να εξάγει αποτελέσματα εκτός των επιτρεπτών τιμών. Σε αυτήν την περίπτωση, η προβλεπόμενη τιμή $p_{a,i}$ στρογγυλοποιείται στην κοντινότερη επιτρεπτή τιμή.

5.7.2 Trust-enhanced Συστήματα Προτάσεων και Cold-start πρόβλημα

Η αποτελεσματικότητα των συστημάτων προτάσεων με συνεργατικό φιλτράρισμα (με κριτήρια αξιολόγησης την ακρίβεια και την κάλυψη που έχουν μελετηθεί στην ενότητα 5.4) επηρεάζεται σε μεγάλο βαθμό από τον αριθμό των βαθμολογιών που υπάρχουν αποθηκευμένες στο σύστημα για

τον κάθε χρήστη. Όσο πιο πολλές βαθμολογίες υπάρχουν διαθέσιμες από τους χρήστες, τόσο πιο πολλοί γείτονες μπορούν να ταιριάξουν μεταξύ τους και έτσι περισσότερα προϊόντα είναι δυνατόν να προταθούν, βελτιώνοντας την ποιότητα του συστήματος προτάσεων. Επιπλέον, η εξαγωγή προβλέψεων είναι δυνατή μόνο σε χρήστες που έχουν βαθμολογήσει τουλάχιστον δυο προϊόντα, αφού ο τύπος της Pearson συσχέτισης χρειάζεται το λιγότερο δυο βαθμολογίες ανά χρήστη. Ένα σημαντικό πρόβλημα λοιπόν εμφανίζεται όσον αφορά τους cold-start χρήστες : ως καινούριοι χρήστες, δεν έχουν ακόμα βαθμολογήσει έναν σεβαστό αριθμό προϊόντων και έτσι δεν μπορούν να λάβουν αξιόπιστες συστάσεις.

Μια πολλά υποσχόμενη προσέγγιση προτείνει την εισαγωγή ενός δικτύου εμπιστοσύνης για την αντιμετώπιση του cold-start προβλήματος, επειδή η πληροφορία που περιέχεται στις δηλώσεις εμπιστοσύνης των χρηστών στο σύστημα προτάσεων μπορεί να διαδοθεί και να συναθροιστεί, και με αυτόν τον τρόπο περισσότεροι χρήστες και προϊόντα μπορούν να συνδεθούν μεταξύ τους. Κάνοντας ορισμένες έξυπνες συνδέσεις στο δίκτυο εμπιστοσύνης, οι νέοι χρήστες μπορούν να αποκτήσουν άμεση πρόσβαση σε ένα μεγάλο εύρος συστάσεων προϊόντων. Το πρόβλημα μπορεί να αντιμετωπιστεί ακόμα και για χρήστες που έχουν βαθμολογήσει μόνο ένα κοινό προϊόν, μέσω της αυτόματης παραγωγής τιμών εμπιστοσύνης.

Μια από τις πιο γνωστές και εκτενείς μελέτες σχετικά με τα trust-enhanced συστήματα προτάσεων με συνεργατικό φιλτράρισμα με βάση την μνήμη για cold-start χρήστες είναι αυτή των Massa και Avesani [25]. Στην έρευνα αυτή συμπεραίνεται ότι είναι χρησιμότερο για έναν νέο χρήστη να κάνει κάποιες συνδέσεις στο δίκτυο εμπιστοσύνης παρά να βαθμολογήσει τον ίδιο αριθμό προϊόντων. Οι Massa και Avesani διεξήγαγαν πειράματα στα data set των προϊόντων της ιστοσελίδας ηλεκτρονικού εμπορίου Epinions.com για το σύστημα συνεργατικού φιλτραρίσματος με βάση την εμπιστοσύνη. Η ιστοσελίδα αυτή διατηρεί ένα δίκτυο εμπιστοσύνης, βάζοντας τους χρήστες της να δηλώσουν ποιους χρήστες εμπιστεύονται, να δημιουργήσουν δηλαδή το Web of Trust τους. Τα πειράματα έδειξαν ότι για χρήστες που έχουν βαθμολογήσει μόνο δυο προϊόντα, για τους οποίους το κλασσικό συνεργατικό φιλτράρισμα δεν μπορεί να βρει συσχετισμούς και να εκτιμήσει κάποια πρόβλεψη, μπορεί να επιτευχθεί κάλυψη (coverage) στο trust-enhanced σύστημα της τάξης του 45%. Η κάλυψη σε ένα σύστημα προτάσεων υπενθυμίζεται ότι αναφέρεται στον αριθμό ζευγαριών προϊόντων ή χρηστών για τους οποίους μπορεί να εξαχθεί πρόβλεψη, καθώς δεν είναι πάντα εφικτό να προβλεφθεί μια βαθμολογία (όπως στην περίπτωση του cold-start χρήστη που δεν έχει γείτονες). Τα αποτελέσματα του πειράματος έδειξαν επίσης ότι για χρήστες που έχουν βαθμολογήσει τρία προϊόντα, η κάλυψη του συστήματος με εμπιστοσύνη μπορεί να φτάσει το 53%, έναντι του 4% που μπορεί να πετύχει το κλασσικό συνεργατικό φιλτράρισμα που δεν λαμβάνει υπ' όψιν τον παράγοντα της εμπιστοσύνης, ενώ για χρήστες που έχουν βαθμολογήσει τέσσερα προϊόντα, η κάλυψη του Trust-enhanced συστήματος φτάνει το 59%, έναντι του 8% του collaborative filtering.

Είναι λοιπόν εμφανές ότι ένα σύστημα προτάσεων μπορεί να επωφεληθεί σε μεγάλο βαθμό από την εισαγωγή ενός δικτύου εμπιστοσύνης, όταν πρόκειται για cold-start χρήστες. Το γεγονός ότι χρήστες που έχουν βαθμολογήσει δυο ή τρία προϊόντα έχουν κατά μέσο όρο δυο-τρεις συνδέσεις εμπιστοσύνης δείχνει ότι οι καινούριοι χρήστες μπορούν σε αρχικό στάδιο να κερδίσουν πολύ περισσότερα από τις δηλώσεις εμπιστοσύνης από ότι από τις βαθμολογήσεις τους.

5.7.3 Πεδία για μελλοντική έρευνα στα trust-enhanced συστήματα προτάσεων

Υπάρχουν κάποιες κατευθύνσεις στον τομέα τον συστημάτων προτάσεων με βάση την εμπιστοσύνη οι οποίες έχουν ακόμα ανάγκη από περαιτέρω διερεύνηση. Ένα από αυτά τα πεδία

σχετίζεται με το γεγονός ότι οι cold-start χρήστες με την έννοια ότι δεν έχουν ακόμη βαθμολογήσει πολλά προϊόντα, είναι cold-start χρήστες και στον τομέα της εμπιστοσύνης. Το σύστημα πρέπει να στρέφει την προσοχή των νέων χρηστών στο δίκτυο εμπιστοσύνης και να τους ενθαρρύνει να κάνουν συνδέσεις, όμως η επιλογή των χρηστών με τους οποίους θα κάνουν σύνδεση είναι συχνά ένα δύσκολο έργο. Δεδομένου μάλιστα του πόσο επηρεάζουν αυτές οι συνδέσεις εμπιστοσύνης τις παραγόμενες προβλέψεις, είναι απαραίτητο το σύστημα να έχει τρόπους ώστε να βοηθά και να κατευθύνει τους καινούριους χρήστες σε αυτό το αρχικό στάδιο των συνδέσεών τους. Επιπλέον, έρευνα μπορεί να γίνει στο θέμα της δημιουργίας συνδέσεων εμπιστοσύνης, όταν η πληροφορία δεν δίνεται άμεσα από τους χρήστες. Το σύστημα προτάσεων μπορεί να χρησιμοποιήσει πολλές άλλες πηγές κοινωνικών δεδομένων, όπως online κοινωνικά ή business δίκτυα (για παράδειγμα το Facebook ή το LinkedIn), τις επικοινωνίες μέσω ηλεκτρονικού ταχυδρομείου κ.ά. Τα δεδομένα από τις κοινωνικές πλατφόρμες θα μπορούσαν να εισαχθούν στα συστήματα προτάσεων με βάση την εμπιστοσύνη, ωστόσο δεν έχει διεξαχθεί ακόμα έρευνα για το ποια κοινωνικά δεδομένα είναι πιο χρήσιμα, ή για το αν αυτές οι πηγές κοινωνικών δεδομένων θα παράγουν παρόμοια αποτελέσματα όπως οι κλασσικές προσεγγίσεις trust-enhanced συστημάτων προτάσεων. Μια άλλη κατεύθυνση που δεν έχει αναλυθεί αρκετά είναι η ενδεχόμενη εισαγωγή της μη-εμπιστοσύνης ως παράγοντα στα συστήματα προτάσεων με βάση την εμπιστοσύνη. Ο βασικός λόγος που αυτό το μοντέλο δεν έχει ακόμα ερευνηθεί αρκετά είναι ότι υπάρχουν ελάχιστα διαθέσιμα data sets που να περιλαμβάνουν πληροφορίες για την δυσπιστία χρηστών απέναντι σε άλλους. Άλλος λόγος είναι ότι δεν έχει αποφασιστεί ομόφωνα ένας τρόπος για το πώς θα διαδίδεται η δυσπιστία και πώς θα επηρεάζει το αποτέλεσμα των συστάσεων. Χρειάζεται λοιπόν περαιτέρω έρευνα για την ενσωμάτωση της έλλειψης εμπιστοσύνης στα συστήματα προτάσεων.

6

Εφαρμογή knn αλγορίθμου σε δεδομένα από το Facebook

Στο κεφάλαιο αυτό θα εφαρμοστεί ο αλγόριθμος k-nearest neighbors σε δεδομένα που συλλέχθηκαν από το facebook για την εξαγωγή πρόβλεψης βαθμολογίας σε ταινίες. Για την εφαρμογή του αλγορίθμου χρησιμοποιούνται βοηθητικές κλάσεις από το πακέτο WEKA, οι οποίες θα μελετηθούν στη συνέχεια. Αρχικά συλλέχθηκαν δεδομένα (ταινίες στις οποίες οι χρήστες έχουν κάνει «Like») από ένα υποδίκτυο φίλων μας στο Facebook. Τα δεδομένα ζητήθηκαν και στάλθηκαν από οδηγίες μέσω e-mail, αλλά εκμεταλευόμμενοι αυτόματα εργαλεία που προσφέρει το Facebook στις οδηγίες χρήσης του API του. Στη συνέχεια, εφαρμόστηκε ο αλγόριθμος των κ-κοντινότερων γειτόνων με χρήση των κλάσεων του πακέτου WEKA. Το πρόγραμμα εκτελέστηκε για διάφορες περιπτώσεις (για όλους τους χρήστες, για όλες τις ταινίες και για διαφορετικές τιμές του κ) πειράζοντας κάθε φορά τις αντίστοιχες μεταβλητές. Στο τέλος συγκεντρώθηκαν τα διάφορα αποτελέσματα και έγινε σύκγριση με πραγματικές βαθμολογίες που ζητήθηκαν από τους ίδιους χρήστες για τις ίδιες ταινίες, ώστε να προκύψουν συμπεράσματα που αφορούν τα αποτελέσματα που προέκυψαν από την χρήση του «Like» σε σχέση με τα δεδομένα που έδωσαν οι χρήστες όταν τους ζητήθηκε πραγματική βαθμολογία.

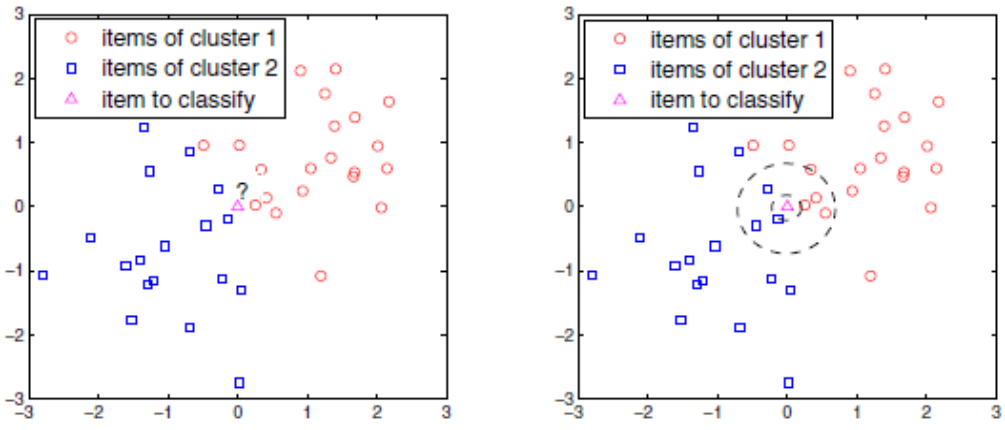
6.1 Ταξινομητής κοντινότερων γειτόνων

Ένας πολύ δημοφιλής και λεπτομερής ταξινομητής είναι ο ταξινομητής κοντινότερων γειτόνων (kNN). Για να ταξινομήσει ένα αντικείμενο, ο ταξινομητής βρίσκει τα k κοντινότερα αντικείμενα από το training dataset και στη συνέχεια αναθέτει μια ετικέτα κατηγορίας με βάση τις κατηγορίες στις οποίες ανήκουν τα κοντινότερα αντικείμενα. Η βασική ιδέα είναι ότι αν ένα αντικείμενο ταξινομηθεί σε μια «γειτονιά» στην οποία κυριαρχεί μια συγκεκριμένη ετικέτα κατηγορίας, τότε

αυτό σημαίνει ότι κατά πάσα πιθανότητα το αντικείμενο ανήκει σε αυτήν την συγκεκριμένη κατηγορία.

Δεδομένου λοιπόν ενός αντικειμένου-ερωτήματος q για το οποίο ζητείται να βρεθεί η κατηγορία του 1 και μιας ομάδας Training δεδομένων X = {{x1,l1}...{xn}} , όπου xj είναι το j-οστό στοιχείο και lj είναι η ετικέτα κατηγορίας του, ο ταξινομητής κ κοντινότερων γειτόνων θα βρει ένα υποσύνολο Y = {{y1,l1}…{yk}} τέτοιο ώστε $Y \in X$ και το $\sum_1^k d(q, yk)$ να είναι ελάχιστο. Το υποσύνολο Y περιέχει τα κ αντικείμενα του X τα οποία είναι κοντινότερα στο αντικείμενο-ερώτημα q. Έτσι, η κλάση του q θα είναι η l = f({l1…lk}).

Η μεγαλύτερη πρόκληση για τον k-NN ταξινομητή είναι η επιλογή της τιμής k, καθώς διαφορετικές τιμές μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα κατηγοριοποίησης. Ένα παράδειγμα φαίνεται στην εικόνα 14. Στη αριστερή εικόνα παρουσιάζονται τα training δεδομένα που είναι χωρισμένα σε δυο ετικέτες κατηγορίας, τους κύκλους και τα τετράγωνα. Το τρίγωνο είναι το αντικείμενο-ερώτημα q το οποίο ζητείται να ταξινομηθεί και βρίσκεται στο σύνορο των δυο κατηγοριών, γεγονός που καθιστά την ταξινόμηση δυσκολότερη. Στην δεξιά εικόνα φαίονται οι κοντινότεροι γείτονες που βρίσκει ο ταξινομητής για τις τιμές του κ 1 και 7. Παρατηρείται ότι για κ=1, η ετικέτα-κατηγορία του αντικειμένου-ερωτήματος είναι τετράγωνο ενώ για κ=7 η κατηγορία είναι κύκλος, βάσει του κανόνα της πλειοψηφίας.



Εικόνα 14 : Παράδειγμα ταξινομητή κ-κοντινότερων γειτόνων [3]

Ο αλγόριθμος των κοντινότερων γειτόνων είναι από τις πιο εύκολες και πιο συνηθισμένες προσεγγίσεις του συνεργατικού φιλτραρίσματος και έχει επικρατήσει ως μέθοδος στα συστήματα προτάσεων. Ένα πολύ σημαντικό πλεονέκτημα αυτού του ταξινομητή είναι το πόσο σχετίζεται με την ιδέα του συνεργατικού φιλτραρίσματος, με την ιδέα δηλαδή του να βρίσκεις όμοιους χρήστες ή προϊόντα για έναν συγκεκριμένο χρήστη-στόχο. Επιπλέον, ο k-NN ταξινομητής δεν χρειάζεται να σχεδιάσει ή να μάθει κάποιο μοντέλο και έτσι μπορεί να προσαρμόζεται εύκολα σε αλλαγές στην μήτρα βαθμολογιών. Όπως αναφέρεται στην σχετική ερευνητική εργασία [26], η μεγάλη επιτυχία του αλγορίθμου στον συγκεκριμένο τομέα οφείλεται στο γεγονός ότι αυτοματοποιεί την συλλογή και τον συνδυασμό των ανθρώπινων γνωμών και αυτό οδηγεί στον υπολογισμό ποιοτικών συστάσεων. Επιπλέον, η προσέγγιση αυτή οδηγεί σε αποτελέσματα υψηλής ακρίβειας και επιδέχεται εύκολα διορθώσεις και βελτιώσεις. Έχουν ήδη γίνει αρκετές βελτιώσεις του k-NN αλγορίθμου, όπως η αφαίρεση των παγκόσμιων επιδράσεων (π.χ. κάποια προϊόντα μπορεί να

έλκουν χρήστες που βαθμολογούν σταθερά με χαμηλούς βαθμούς) ή η εφαρμογή μεθόδου βελτιστοποίησης για τον υπολογισμό των βαρών παρεμβολής στην εκάστοτε «γειτονιά» από τους Bell και Koren στα πλαίσια του διαγωνισμού Netflix Prize.
Όλοι οι παραπάνω λόγοι οδήγησαν στην επιλογή και εφαρμογή του συγκεκριμένου αλγορίθμου για την εξαγωγή πρόβλεψης.

6.2 Συλλογή δεδομένων από το API του Facebook

Στα Κοινωνικά Μέσα υπάρχει η ανάγκη να βρίσκουν οι χρήστες όλες τις πληροφορίες των χρηστών με τους οποίους συνδέονται σε μια πλατφόρμα, χωρίς να χάνουν χρόνο περιφερόμενοι από την μια ιστοσελίδα στην άλλη. Τη δυνατότητα αυτή αλληλεπίδρασης και επικοινωνίας διαφορετικών προγραμμάτων μας την προσφέρουν εύκολα και γρήγορα τα APIs (Application Programming Interface) ή διεπαφές προγραμματισμού εφαρμογών των Κοινωνικών Μέσων. API είναι μια διεπαφή με συγκεκριμένους κανόνες την οποία μπορεί να ακολουθήσει ένα πρόγραμμα λογισμικού για να επικοινωνήσει με άλλα προγράμματα. Στην παρούσα εφαρμογή συλλέχθηκαν δεδομένα από τις συνδέσεις στο Facebook μέσω του Graph API που είναι και η πηγή της πλατφόρμας του Facebook. Το Graph API επιτρέπει να διαβάζει και να γράφει κανείς δεδομένα παρέχοντας μια όψη του κοινωνικού συνόλου, δηλαδή των αντικειμένων (όπως χρήστες, φωτογραφίες, εκδηλώσεις) και των μεταξύ τους συνδέσεων (φιλίες, Likes κ.τ.λ). Έτσι, ζητήθηκε από κάποιους φίλους του Facebook να στείλουν δεδομένα που εμφανίζονται στα profiles τους στο Facebook μέσω της ενότητας «Χρήστης» του Graph API. Οι χρήστες αυτοί είναι φίλοι του γράφοντος στο Facebook αλλά οι περισσότεροι από αυτούς είναι και φίλοι μεταξύ τους με πολλά κοινά στοιχεία. Για την ακρίβεια, αποθηκεύθηκαν οι πίνακες με JSON αντικείμενα για τις αναγνωριστικές πληροφορίες του κάθε χρήστη (userID, όνομα, χώρα, φύλο κ.ά.) και για κάποιες από τις συνδέσεις του (τους φίλους και τις ταινίες στις οποίες έχει κάνει «Like») . Τα δεδομένα αυτά τα έκαναν copy-paste από το Graph API του Facebook επιλέγοντας τα αντικείμενα «friends» και «movies» από τον πίνακα με τις συνδέσεις τους και στάλθηκαν μέσω e-mail. Τα δεδομένα που αφορούν τις ταινίες στις οποίες οι χρήστες είχαν κάνει «Like» χρησιμοποιήθηκαν ως training dataset στην εφαρμογή του αλγορίθμου μας. Παράδειγμα ενός JSON αντικειμένου που έστειλε ο χρήστης Νιόβη σχετικά με τις ταινίες που εμφανίζονται στο profile της στο Facebook φαίνεται παρακάτω :


```
{
   "data": [
      {
         "name": "Slumdog Millionaire",
         "category": "Movie",
         "id": "53180527367",
         "created_time": "2011-05-16T12:52:19+0000"
      },
      {
         "name": "Black Swan",
         "category": "Movie",
         "id": "106602966061813",
         "created_time": "2011-05-16T12:52:03+0000"
      },
```

```
      {
         "name": "The Social Network Movie",
         "category": "Movie",
         "id": "160640653979986",
         "created_time": "2011-05-16T12:51:46+0000"
      },
      {
         "name": "Love Me If You Dare",
         "category": "Movie",
         "id": "103109673062229",
         "created_time": "2011-05-16T12:50:50+0000"
      },
      {
         "name": "Seven",
         "category": "Movie",
         "id": "103131766394504",
         "created_time": "2011-05-16T12:50:38+0000"
      },
      {
         "name": "Inception",
         "category": "Movie",
         "id": "91290503700",
         "created_time": "2011-05-16T12:50:12+0000"
      },
      {
         "name": "Fight Club",
         "category": "Movie",
         "id": "39644305296",
         "created_time": "2011-05-16T12:49:48+0000"
      },
      {
         "name": "Kill Bill",
         "category": "Movie",
         "id": "105628019469787",
         "created_time": "2011-05-16T12:49:38+0000"
      },
      {
         "name": "Rock N Rolla",
         "category": "Movie",
         "id": "37761584616",
         "created_time": "2011-05-16T12:49:23+0000"
      }
   ]
}
```

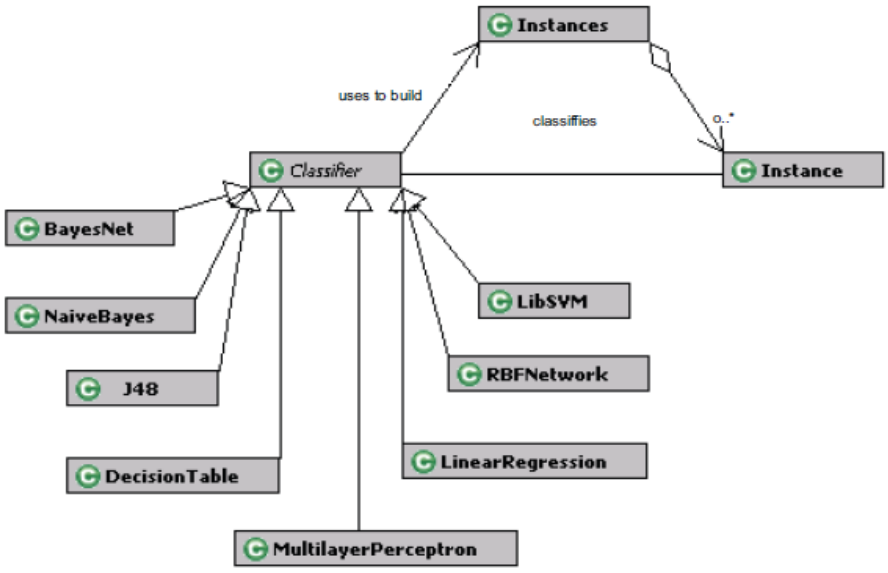

Επιλέχθηκε να χρησιμοποιηθούν δεδομένα από το Facebook καθώς υπήρχε ανάγκη μιας πλατφόρμας που να παρέχει στους χρήστες την δυνατότητα να δηλώνουν τι τους αρέσει και στο Facebook αυτό γίνεται μέσω του «Like» (παρά τις αδυναμίες που παρουσιάζει και θα αναλυθούν στη συνέχεια), ενώ π.χ. το Twitter δεν ενδείκνυται για συλλογή δεδομένων. Επίσης, παρουσιάζει

μεγάλο ενδιαφέρον να εξεταστεί αν και πώς μπορεί μια εξωτερική εφαρμογή συστήματος προτάσεων να επωφεληθεί από ένα κοινωνικό μέσο σαν το Facebook που περιλαμβάνει μια τεράστια παγκόσμια κοινότητα, ενώ η άμεση συνεργασία του IMDb με το Facebook δίνει μία μεγάλη και πραγματική βάση.

6.3 Open Source Πακέτο Ανάκτησης πληροφοριών WEKA

Το WEKA (Waikato Environment for Knowledge Analysis) είναι ένα από τα πιο δημοφιλή πακέτα ανοιχτής πηγής αλγορίθμων ανάκτησης πληροφοριών σε γλώσσα προγραμματισμού Java. Περιλαμβάνει εργαλεία για επεξεργασία δεδομένων, κατηγοριοποίηση, παλινδρόμηση, οπτικοποίηση και κανόνες συσχετίσεων. Πληροφορίες για την εγκατάσταση του πακέτου WEKA δίνονται στο [Παράρτημα Α]. Δυο από τα βασικά πακέτα του WEKA τα οποία χρησιμοποιήθηκαν και στην εφαρμογή είναι τα weak.core και weak.classifiers. Το weka.core είναι η «καρδιά» του WEKA και περιλαμβάνει βασικά στοιχεία τα οποία χρησιμοποιούνται και από τα υπόλοιπα πακέτα. Περιέχει classes για μοντελοποίηση γνωρισμάτων, για ομάδες δεδομένων, χειρισμό μητρών, αναπαράσταση δέντρων, parsing κειμένου και για XML. Το πακέτο weak.classifiers περιλαμβάνει εφαρμογές των διαφόρων αλγορίθμων κατηγοριοποίησης, συμπεριλαμβανομένων και των αλγορίθμων για προβλέψεις τιμών.

Τα δεδομένα στο weak.core εκπροσωπούνται από την κλάση Instances, η οποία περιλαμβάνει μια λίστα παραδειγμάτων τα οποία εκπροσωπούνται από την κλάση Instance. Κάθε Instance αποτελείται από συγκεκριμένα γνωρίσματα (attributes). Το WEKA χρησιμοποιεί δική του εφαρμογή για την παρουσίαση διανύσματος, το FastVector. Το πακέτο weak.classifier περιέχει εφαρμογές για αλγορίθμους κατηγοριοποίησης και πρόβλεψης. Ο classifier μαθαίνει το μοντέλο χρησιμοποιώντας τα Instances του core πακέτου. Στη συνέχεια μπορεί να κατηγοριοποιήσει ένα στιγμιότυπο.



Εικόνα 15 : Το πακέτο classifier χρησιμοποιεί τα Στιγμιότυπα για να μάθει ένα μοντέλο και να κατηγοριοποιήσει ένα στιγμιότυπο. Στην εικόνα εμφανίζονται μερικοί μόνο από τους αλγορίθμους κατηγοριοποίησης και πρόβλεψης που περιέχονται στην WEKA βιβλιοθήκη [20]

Πιο συγκεκριμένα, το πακέτο weak.core.neighboursearch που θα χρησιμοποιηθεί περιέχει classes που εφαρμόζουν αλγορίθμους εύρεσης κ-κοντινότερων γειτόνων. Στον πίνακα 16 αναφέρονται κάποιες από αυτές τις κλάσεις, οι οποίες επεκτείνουν την abstract βασική κλάση NearestNeighbourSearch.

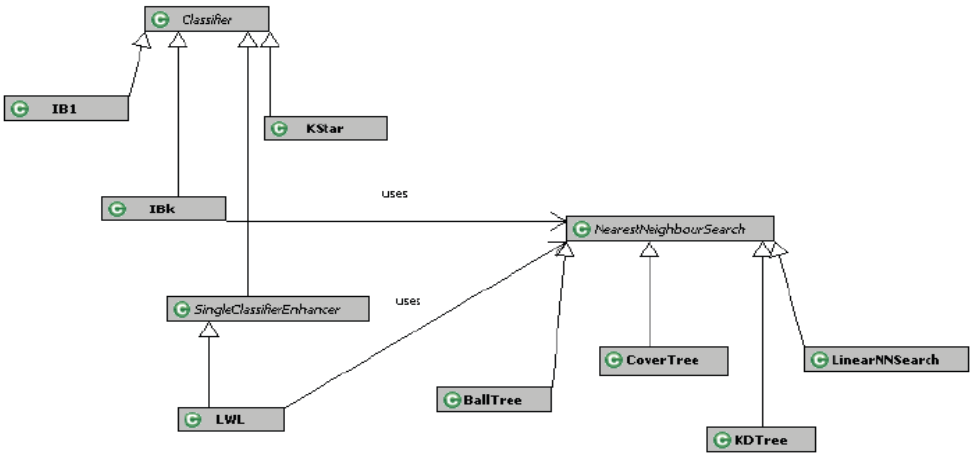
Πίνακας 16 : NearestNeighborSearch κλάσεις του WEKA [20]

ΚΛΑΣΗ	ΠΕΡΙΓΡΑΦΗ
NearestNeighbourSearch	Abstract κλάση για την εύρεση των κοντινότερων γειτόνων
BallTree	Εφαρμόζει τον BallTree/MetricTree αλγόριθμο για την εύρεση κοντινότερων γειτόνων
CoverTree	Εφαρμόζει την CoverTree δομή δεδομένων για την εύρεση των κοντινότερων γειτόνων
KDTree	Εφαρμόζει τον αλγόριθμο αναζήτησης KDTree Για την εύρεση κοντινότερων γειτόνων
LinearNNSearch	Εφαρμόζει τον αλγόριθμο αναζήτησης LinearNNSearch για την εύρεση των κοντινότερων γειτόνων

Το δε πακέτο weak.classifiers.lazy περιέχει κλάσεις αλγορίθμων κατηγοριοποίησης που βασίζονται στην τεχνική των κοντινότερων γειτόνων. Κάποιοι από αυτούς τους αλγορίθμους παρουσιάζονται στον πίνακα 17.

Πίνακας 17 : WEKA Classifiers με βάση την αναζήτηση κοντινότερων γειτόνων [20]

ΚΛΑΣΗ	ΠΕΡΙΓΡΑΦΗ
IB1	Ταξινομητής κοντινότερων γειτόνων
IBk	Ταξινομητής k-κοντινότερων γειτόνων
KStar	Ταξινομητής με βάση τα στιγμιότυπα που χρησιμοποιεί συνάρτηση αποστάσεων με βάση την εντροπία



Εικόνα 16 : WEKA κλάσεις που συνδέονται στην αναζήτηση κοντινότερων γειτόνων[20]

6.4 Εφαρμογή Αλγορίθμου

Για την εφαρμογή του k-nn αλγορίθμου με την χρήση κλάσεων από το πακέτο του WEKA γίνεται χρήση των δεδομένων που υπάρχουν καταχωρημένα στον Πίνακα 18. Τα δεδομένα είναι προτιμήσεις χρηστών του Facebook σε ταινίες. Με τον βαθμό 1 δηλώνονται οι ταινίες στις οποίες ο χρήστης έχει κάνει «Like», ενώ με 0 οι ταινίες στις οποίες δεν έχει κάνει. Ενδεικτικά, συλλέξαμε δεδομένα 13 φίλων στο Facebook.

Στο πρόγραμμα που τρέχει σε γλώσσα JAVA στο Eclipse, εισάγονται οι κλάσεις weka.classifiers.lazy.IBk, weka.core.Attribute, weka.core.FastVector, weka.core.Instance και weka.core.Instances. Ο κώδικας παρατίθεται στο Παράρτημα Β. Αρχικά δημιουργούνται τα 21 γνωρίσματα (attributes) και τα στιγμιότυπα (Instances) για την αναπαράσταση των δεδομένων. Στη συνέχεια, δημιουργείται ο ταξινομητής που παίρνει τα δεδομένα και εξάγει πρόβλεψη για τον κάθε χρήστη του Πίνακα 18 με την εφαρμογή του k-nn αλγορίθμου. Ως αντικείμενο προς πρόβλεψη τίθεται η ταινία «The Godfather».

Στο αυτό το σημείο σημειώνεται ότι στο Facebook, οι χρήστες κάνουν «Like» σε σελίδες που τους αρέσουν. Δεν μπορεί όμως να εξαχθεί το συμπέρασμα ότι σε έναν χρήστη δεν αρέσει μια σελίδα από την έλλειψη «Like», καθώς μπορεί απλά ο χρήστης να μην ασχολείται αρκετά, να ξέχασε ή να μην βρήκε την συγκεκριμένη σελίδα. Το σωστό λοιπόν θα ήταν να δηλώνεται η έλλειψη «Like» με την τιμή Null στον πίνακα δεδομένων και όχι με την τιμή 0, που δείχνει ότι στον χρήστη δεν αρέσει η σελίδα. Ωστόσο, στο τρέχων πρόγραμμα δεν παρέχεται η δυνατότητα να μπει στα γνωρίσματα η τιμή null, παρά μόνο αριθμητικές τιμές.

	THE GODFATHER	CASABLANCA	CITIZEN KANE	THE SEVENTH SEAL	KILL BILL	SEVEN	ROCK N ROLLA	127 HOURS	THE LORD OF THE RINGS	LOCK, STOCK & 2 SMOKING BARRELS	EL SECRETO DE SUS OJOS
ANNY	0	1	1	1	0	0	0	0	0	0	0
ΚΑΤΕΡΙΝΑ	1	0	0	0	1	1	1	1	0	0	0
ΔΗΜΗΤΡΗΣ	1	1	0	0	1	0	0	0	1	1	0
ΚΡΙΝΑ	1	1	1	1	1	0	1	1	0	0	1
ΝΙΚΟΣ	1	0	0	0	1	1	1	0	0	0	0
ΘΕΟΔΟΣΙΑ	1	0	1	0	1	1	0	0	0	0	0
ΝΑΣΟΣ	0	0	0	0	0	0	1	0	1	1	1
ΓΙΑΝΝΗΣ	1	0	0	1	1	0	0	0	1	0	0
ΝΙΟΒΗ	0	0	0	0	1	1	1	0	0	0	0
ΚΩΣΤΑΣ	0	0	0	0	0	1	0	0	0	0	0
ΒΑΣΙΛΗΣ	1	0	1	0	1	1	1	1	0	1	1
ΖΕΜΗ	0	0	0	0	0	0	0	0	0	0	0
ΑΘΗΝΑ	1	0	1	0	1	1	1	1	0	0	1

Πίνακας 18 : Πίνακας με τα δεδομένα βαθμολογιών που χρησιμοποιήθηκαν στο πρόγραμμα

	REQUIEM FOR A DREAM	FIGHTCLUB	INCEPTION	MATCH POINT	BLACK SWAN	SILENCE OF THE LAMBS	LOVE ME IF YOU DARE	SUPERMAN	SLUMDOG MILLIONAIRE	THE SOCIAL NETWORK
ANNY	0	0	0	0	0	0	0	0	0	0
ΚΑΤΕΡΙΝΑ	0	0	0	0	0	0	0	0	0	0
ΔΗΜΗΤΡΗΣ	0	0	0	1	1	0	0	1	0	0
ΚΡΙΝΑ	1	1	1	1	1	1	0	1	1	1
ΝΙΚΟΣ	0	1	0	0	0	1	0	0	0	0
ΘΕΟΔΟΣΙΑ	0	1	0	0	1	0	1	0	1	1
ΝΑΣΟΣ	0	0	0	0	0	0	0	0	0	0
ΓΙΑΝΝΗΣ	1	1	0	0	0	0	0	1	0	0
ΝΙΟΒΗ	0	1	1	0	1	0	1	0	1	1
ΚΩΣΤΑΣ	1	1	1	0	0	0	0	0	0	0
ΒΑΣΙΛΗΣ	0	1	1	1	0	1	1	1	1	0
ΖΕΜΗ	0	0	0	1	1	0	1	0	0	0
ΑΘΗΝΑ	0	1	1	1	1	1	0	0	1	0

Πίνακας 18 : Συνέχεια δεδομένων βαθμολογιών

Ο αλγόριθμος λειτουργεί ως εξής: Για μία ταινία, έστω την ταινία «The Godfather», κρύβει σε κάθε επανάληψη που κάνει την βαθμολογία ενός χρήστη και χρησιμοποιώντας τους βαθμούς των υπόλοιπων χρηστών υπολογίζει από μόνος του τον βαθμό που προβλέπει ότι θα έβαζε ο χρήστης, και επαναλαμβάνει για όλους τους χρήστες (leave-one-out πείραμα). Έτσι, ως αποτέλεσμα λαμβάνονται οι βαθμοί που προέβλεψε ο αλγόριθμος (Predicted) και δίπλα παρατίθενται οι βαθμοί που έχουν όντως βάλει οι χρήστες στη συγκεκριμένη ταινία.

Παρακάτω παρουσιάζονται τα αποτελέσματα του προγράμματος για τους 13 χρήστες του συγκεκριμένου παραδείγματος. Τα 21 γνωρίσματα (ταινίες) δηλώνονται ως αριθμητικά, καθώς αν στον χρήστη αρέσει η ταινία (έχει κάνει «like») αυτό δηλώνεται με την τιμή 1 στα δεδομένα μας, ενώ αν δεν έχει κάνει «like» τότε στο αντίστοιχο κελί του πίνακα δεδομένων μπαίνει η τιμή 0.

Παρακάτω παρατίθεται η είσοδος του συστήματος και τα αποτελέσματα για κ=1, όπως αυτά βγαίνουν από το weka:

```
@relation wekaCF

@attribute item1 numeric
@attribute item2 numeric
@attribute item3 numeric
@attribute item4 numeric
@attribute item5 numeric
@attribute item6 numeric
@attribute item7 numeric
@attribute item8 numeric
@attribute item9 numeric
@attribute item10 numeric
@attribute item11 numeric
@attribute item12 numeric
@attribute item13 numeric
@attribute item14 numeric
@attribute item15 numeric
@attribute item16 numeric
@attribute item17 numeric
@attribute item18 numeric
@attribute item19 numeric
@attribute item20 numeric
@attribute item21 numeric

@data
1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0
1,1,0,0,1,0,0,0,1,1,0,0,0,0,1,1,0,0,1,0,0
1,1,1,1,1,0,1,1,0,0,1,1,1,1,1,1,1,0,1,1,1
```

```
1,0,0,0,1,1,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0
1,0,1,0,1,1,0,0,0,0,0,0,1,0,0,1,0,1,0,1,1
0,0,0,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0
1,0,0,1,1,0,0,0,1,0,0,1,1,0,0,0,0,0,1,0,0
0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,1,0,1,0,1,1
0,0,0,0,0,1,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0
1,0,1,0,1,1,1,1,0,1,1,0,1,1,1,0,1,1,1,1,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0
1,0,1,0,1,1,1,1,0,0,1,0,1,1,1,1,1,1,0,0,1,0

Prediction:
Expected=1.0 Predicted=1.0
Expected=1.0 Predicted=1.0
Expected=1.0 Predicted=1.0
Expected=1.0 Predicted=1.0
Expected=1.0 Predicted=1.0
Expected=1.0 Predicted=1.0
Expected=0.0 Predicted=0.0
Expected=1.0 Predicted=1.0
Expected=0.0 Predicted=0.0
Expected=0.0 Predicted=0.0
Expected=1.0 Predicted=1.0
Expected=0.0 Predicted=0.0
Expected=1.0 Predicted=1.0
```

Η δυσκολία στον k-nn ταξινομητή έγκειται στην σωστή επιλογή της τιμής του κ. Γενικά, η τιμή του κ εξαρτάται από τα δεδομένα Η επιλογή μικρής τιμής του κ μπορεί να κάνει τον ταξινομητή ευαίσθητο σε σημεία θορύβου. Από την άλλη πλευρά, αν το κ είναι πολύ μεγάλο, τότε το φαινόμενο του θορύβου μειώνεται στην κατηγοριοποίηση, αλλά παύουν να είναι τόσο διακριτά τα όρια μεταξύ των κατηγοριών, και έτσι μπορεί η «γειτονιά» που δημιουργεί ο ταξινομητής να περιλαμβάνει πολλά αντικείμενα άλλων κατηγοριών.

Παραπάνω εκτελέστηκε ο αλγόριθμος για κ = 1, δηλαδή η ταινία προς πρόβλεψη μπαίνει στην κατηγορία του κοντινότερου γείτονά της. Στη συνέχεια επαναλήφθηκε η εφαρμογή ενδεικτικά για κ = 4 και κ = 7 και προέκυψαν τα αποτελέσματα που ακολουθούν :

K = 4

```
Prediction:
Expected=1.0 Predicted=0.5
Expected=1.0 Predicted=0.5
Expected=1.0 Predicted=0.5
Expected=1.0 Predicted=0.75
Expected=1.0 Predicted=0.5
Expected=1.0 Predicted=0.5
Expected=0.0 Predicted=0.6666666666666666
Expected=1.0 Predicted=0.8
Expected=0.0 Predicted=0.6666666666666666
Expected=0.0 Predicted=0.75
Expected=1.0 Predicted=0.8
```

```
Expected=0.0 Predicted=0.5714285714285714
Expected=1.0 Predicted=0.8
```

K = 7

```
Prediction:
Expected=1.0 Predicted=0.625
Expected=1.0 Predicted=0.42857142857142855
Expected=1.0 Predicted=0.7142857142857143
Expected=1.0 Predicted=0.8888888888888888
Expected=1.0 Predicted=0.625
Expected=1.0 Predicted=0.5714285714285714
Expected=0.0 Predicted=0.625
Expected=1.0 Predicted=0.7142857142857143
Expected=0.0 Predicted=0.5714285714285714
Expected=0.0 Predicted=0.5714285714285714
Expected=1.0 Predicted=0.8571428571428571
Expected=0.0 Predicted=0.5714285714285714
Expected=1.0 Predicted=0.8571428571428571
```

Από τα παραπάνω φαίνεται ότι πιο ακριβής ήταν η περίπτωση του κ = 1, που ονομάζεται και nearest neighbor αλγόριθμος. Σε αυτήν την περίπτωση, η κατηγορία του αντικειμένου προβλέπεται να είναι η κατηγορία του κοντινότερου δείγματος εκπαίδευσης. Δημιουργείται δηλαδή μια μόνο κλάση, και το αντικείμενο που εξετάζεται είτε ανήκει σε αυτήν την κλάση (και άρα το πρόγραμμα προβλέπει βαθμό 1 και να αρέσει στον χρήστη η ταινία) είτε δεν ανήκει (προβλέπεται ότι στον χρήστη δεν αρέσει η ταινία και μπαίνει ο βαθμός 0).

Ωστόσο, στην συγκεκριμένη περίπτωση υπάρχουν τα δεδομένα του Facebook με την αδυναμία του «Like» που αναφέρθηκε προηγουμένως, δηλαδή τα μηδενικά των χρηστών στα οποία στηρίζεται ο αλγόριθμος για να προβλέψει την βαθμολογία ενός συγκεκριμένου χρήστη δεν δηλώνουν απαραίτητα ότι η ταινία δεν άρεσε αλλά μπορεί να φανερώνουν και άγνοια. Δεν μπορεί λοιπόν να απλοποιηθεί το πρόβλημα σε μια κατηγοριοποίηση της μορφής «μου αρέσει / δεν μου αρέσει», όπως συμβαίνει για κ=1.

Η περίεργη φύση του «Like» στο Facebook οδήγησε σε περαιτέρω παρατήρηση των αποτελεσμάτων. Για διαφορετικές τιμές του κ πλην της μονάδας, ενώ παρατηρήθηκε ορισμένη απόκλιση από την μονάδα που δίνει ένα «Like» (+1 όπως το αναφέρει αναλόγως η Google στη δικιά της πλατφόρμα), φαίνεται τα αποτελέσματα να είναι πιο κοντά σε έναν ρεαλιστικό στόχο. Με άλλα λόγια μία σειρά από ταινίες δεν μπορεί να παίρνουν την ίδια βαθμολογία, αφού στην κρίση ενός ατόμου υπάρχει μία διαβάθμηση. Τα δεκαδικής φύσης αποτελέσματα ώθησαν στην κατεύθυνση να γίνει έρευνα για το αν πειράζοντας την μεταβλητή κ θα μπορούσαν να βρεθούν πιο ρεαλιστικές και αποτελεσματικές προτάσεις. Για αυτό το λόγο αποφασίστηκε να αναχθούν τα αποτελέσματα του αλγορίθμου σε μία δεκαδική κλίμακα, προκειμένου να συγκριθούνε τα αποτελέσματα με πραγματικά δεδομένα που θα έδιναν οι χρήστες.

Στη συνέχεια λοιπόν, παρατίθενται οι βαθμολογίες που προβλέπει ο αλγόριθμος για όλους τους χρήστες και για όλες τις ταινίες για κ=4 και κ=7 :

	THE GODFATHER	CASABLANCA	CITIZEN KANE	THE SEVENTH SEAL	KILL BILL	SEVEN	ROCK N ROLLA	127 HOURS	THE LORD OF THE RINGS	LOCK, STOCK & 2 SMOKING BARRELS	EL SECRETO DE SUS OJOS
ANNY	5	2,5	2,5	2,5	6,66	5	5	2,5	5	3,33	2
ΚΑΤΕΡΙΝΑ	5	2	4,28	3,33	4	5	5	2	3,33	2	3,33
ΔΗΜΗΤΡΗΣ	5	2,5	2,5	5	4	1,66	2,5	1,66	5	2,5	2,5
ΚΡΙΝΑ	7,5	1,6	6	2	8,5	8	6	5	2,5	2,85	5
ΝΙΚΟΣ	5	1,25	4	2,5	6,66	7,5	4	5	2,5	1,25	2,5
ΘΕΟΔΟΣΙΑ	5	1,43	2	1,43	7,5	7,5	8	4	0	0	2,5
ΝΑΣΟΣ	6,6	4	2,5	2,5	6	6	3,33	1,43	1,66	1,66	1,43
ΓΙΑΝΝΗΣ	8	4	2	2,5	6	5	3,33	1,66	2,5	2	0
ΝΙΟΒΗ	6,6	0	4	0	6	8	4	3,33	0	0	2
ΚΩΣΤΑΣ	7,5	1,43	1,43	2,85	8	5	7,5	1,66	2,85	1,43	1,43
ΒΑΣΙΛΗΣ	8	2,5	7,5	2,5	10	7,5	8	7,5	0	2,5	7,5
ΖΕΜΗ	5,7	4	3,33	2	5	5	4,28	1,66	4	4	2
ΑΘΗΝΑ	8	2,5	7,5	2,5	10	7,5	8	7,5	0	2,5	7,5

Πίνακας 19 : Βαθμολογία που προέβλεψε ο αλγόριθμος για κ=4

	REQUIEM FOR A DREAM	FIGHTCLUB	INCEPTION	MATCH POINT	BLACK SWAN	SILENCE OF THE LAMBS	LOVE ME IF YOU DARE	SUPERMAN	SLUMDOG MILLIONAIRE	THE SOCIAL NETWORK
ANNY	4	5	2	4	4	2	2,5	4	0	0
ΚΑΤΕΡΙΝΑ	3,33	6,66	4,28	2,85	4,44	3,33	3,75	1,66	3,75	2,85
ΔΗΜΗΤΡΗΣ	2	2	0	4	4	0	2	4	0	0
ΚΡΙΝΑ	2	7,5	6	5	5	5	6	4	7,5	2,5
ΝΙΚΟΣ	5	6,66	6	2,5	5	1,66	4	2,5	5	4
ΘΕΟΔΟΣΙΑ	1,43	6	5	4	5	5	4	0	4	3,33
ΝΑΣΟΣ	2,5	4	2,5	5	5	2,5	1,43	2,5	0	0
ΓΙΑΝΝΗΣ	2,5	5	2	2	2	2	0	2,5	0	0
ΝΙΟΒΗ	2,5	6,66	3,33	3,33	5	4	5	0	5	5
ΚΩΣΤΑΣ	2	5	2	1,66	5	1,66	5	1,66	1,66	1,66
ΒΑΣΙΛΗΣ	2,5	8	7,5	7,5	6,66	5,71	2,5	5	7,5	5
ΖΕΜΗ	2	4,28	3,33	2,5	2,5	0	2	2	2,85	2,8
ΑΘΗΝΑ	2,5	8	7,5	7,5	5	5,71	5	5	7,5	5

Πίνακας 19 : Συνέχεια Βαθμολογιών για κ=4

	THE GODFATHER	CASABLANCA	CITIZEN KANE	THE SEVENTH SEAL	KILL BILL	SEVEN	ROCK N ROLLA	127 HOURS	THE LORD OF THE RINGS	LOCK, STOCK & 2 SMOKING BARRELS	EL SECRETO DE SUS OJOS
ANNY	6,25	1,43	1,25	2,5	5,55	4,44	3,75	1,25	3,75	2,5	1,25
ΚΑΤΕΡΙΝΑ	4,28	1,81	4,28	2	6	4,28	3,75	1,11	2,72	1,81	2
ΔΗΜΗΤΡΗΣ	7,14	2,85	2,5	2,85	5,71	2,85	4,28	1,43	4,28	2,85	1,43
ΚΡΙΝΑ	8,88	1,43	5	2,5	8,57	7,14	5	3,33	2,85	2,85	3,33
ΝΙΚΟΣ	6,25	1,25	3,75	2,5	6,66	6,66	5	2,85	2,5	1,25	2,5
ΘΕΟΔΟΣΙΑ	5,71	1,43	2,85	1,43	6,25	7,14	5,55	3,33	1,11	1,11	2,22
ΝΑΣΟΣ	6,25	2,85	1,43	2,5	5	4,28	3,75	1,43	2,85	2,5	1,43
ΓΙΑΝΝΗΣ	7,14	2,85	2,5	2,85	5,71	5	4,28	1,43	3,33	2,85	1,43
ΝΙΟΒΗ	5,71	0	2,85	0	7,14	8,57	5,71	2,85	0	0	1,43
ΚΩΣΤΑΣ	5,71	1,43	1,43	2,85	5,55	5	5,71	1,25	2,85	1,43	1,43
ΒΑΣΙΛΗΣ	8,57	1,43	5,71	1,43	10	8,57	8,57	5,71	0	1,43	4,28
ΖΕΜΗ	5,71	2,5	2,5	1,25	5	5	4,28	1,25	2,5	2,5	1,25
ΑΘΗΝΑ	8,57	1,43	5,71	1,43	10	8,57	8,57	5,71	0	1,43	4,28

Πίνακας 20 : Βαθμολογία που προέβλεψε ο αλγόριθμος για κ=7

	REQUIEM FOR A DREAM	FIGHTCLUB	INCEPTION	MATCH POINT	BLACK SWAN	SILENCE OF THE LAMBS	LOVE ME IF YOU DARE	SUPERMAN	SLUMDOG MILLIONAIRE	THE SOCIAL NETWORK
ΑΝΝΥ	2,5	4,44	1,25	2,5	3,33	1,25	2,22	2,5	1,11	1,11
ΚΑΤΕΡΙΝΑ	2	6,66	4,28	2,85	4,44	2	3,75	1,81	3,75	2,85
ΔΗΜΗΤΡΗΣ	1,43	2,85	0	2,85	2,85	1,43	2,5	2,85	1,25	1,25
ΚΡΙΝΑ	2,22	7,5	4,44	3,75	5	3,75	3,33	3,75	5,55	3,33
ΝΙΚΟΣ	2,85	6,66	4,28	1,43	4,28	2,22	2,85	1,43	4,28	2,85
ΘΕΟΔΟΣΙΑ	1,43	5,71	4,28	3,33	5	3,33	3,75	2	3,75	2,5
ΝΑΣΟΣ	2,5	3,75	1,43	2,85	2,85	1,43	1,43	2,5	0	0
ΓΙΑΝΝΗΣ	2,85	4,28	1,43	2,5	3,33	1,43	2,22	2,85	1,25	1,25
ΝΙΟΒΗ	1,43	7,14	4,28	2,85	5,71	2,85	4,28	0	4,28	2,85
ΚΩΣΤΑΣ	2,5	5	2,5	1,25	3,33	1,25	3,33	1,25	2,22	2,22
ΒΑΣΙΛΗΣ	1,43	8.57	5,71	4,28	5,71	5,71	4,28	2,85	7,14	4,28
ΖΕΜΗ	1,25	4,28	2,5	2,5	4,44	1,11	3,33	1,25	2,85	2,85
ΑΘΗΝΑ	1,43	8,57	5,71	4,28	5,71	5,71	4,28	2,85	7,14	4,28

Πίνακας 20 : Συνέχεια Βαθμολογιών για κ=7

Στη συνέχεια, ζητήθηκε από τους ίδιους χρήστες που συμμετείχαν και νωρίτερα στο πείραμα, να βαθμολογήσουν τις ταινίες αυτές με κλίμακα 0-10, ενώ οι ταινίες που δεν έχουν δει να σημειωθούν με παύλα. Στον πίνακα που ακολουθεί τα γκρι κελιά αντιστοιχούν στις ταινίες στις οποίες οι χρήστες είχαν κάνει «like», σύμφωνα με τα δεδομένα του πίνακα 18. Με αυτόν τον τρόπο, θα γίνει σύγκριση έπειτα ανάμεσα στους βαθμούς που προέβλεψε ο αλγόριθμος με αυτούς που έβαλαν όντως οι χρήστες, καθώς και θα εξαχθούν συμπεράσματα σχετικά με την συμπεριφορά των χρηστών όταν έχουν την δυνατότητα να βαθμολογήσουν με βαθμούς από το 0 έως το 10, έναντι της χρήσης ή όχι του κουμπιού «like» που τους παρέχεται στο Facebook. Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος υπολογίζει μόνος του ως αποτελέσματα ακριβώς τις τιμές που έχει δεχτεί ως δεδομένα μόνο στην περίπτωση του κ=1, χωρίς όμως να επιτρέπεται στο πρόγραμμα να μπουν τιμές Null για τις ταινίες που δεν έχουν δει οι χρήστες. Αν μπει λοιπόν ένας χρήστης με τις δικές του προτάσεις που φαίνονται από τα δεδομένα του, τότε θα βρεθούν από το πρόγραμμα οι κ που του μοιάζουν περισσότερο. Ο λόγος όμως που ζητήθηκε από τους χρήστες να δώσουν επιπλέον και μια πραγματική βαθμολογία είναι επειδή κρίθηκε σκόπιμο να εξετασθεί αν οι χρήστες επηρεάζονται και από άλλους παράγοντες όταν κάνουν ή δεν κάνουν «Like» σε μια ταινία.

	THE GODFATHER	CASABLANCA	CITIZEN KANE	THE SEVENTH SEAL	KILL BILL	SEVEN	ROCK N ROLLA	127 HOURS	THE LORD OF THE RINGS	LOCK, STOCK & 2 SMOKING BARRELS	EL SECRETO DE SUS OJOS
ANNY	-	7	6	7	-	-	-	2	4	3	-
ΚΑΤΕΡΙΝΑ	7	2	-	3	7	6	6	5	3	3	-
ΔΗΜΗΤΡΗΣ	7	5	-	-	6	2	3	-	6	6	-
ΚΡΙΝΑ	8	3	6	3	9	-	7	6	2	-	6
ΝΙΚΟΣ	7	-	-	-	7	8	6	-	3	-	3
ΘΕΟΔΟΣΙΑ	5	-	6	-	8	8	8	-	-	0	-
ΝΑΣΟΣ	-	-	2	2	6	-	5	-	5	5	5
ΓΙΑΝΝΗΣ	9	4	2	5	7	-	-	2	6	2	-
ΝΙΟΒΗ	-	1	4	-	6	8	6	-	0	-	2
ΚΩΣΤΑΣ	-	1	-	-	-	6	-	-	-	1	1
ΒΑΣΙΛΗΣ	8	2	8	-	10	7	8	7	-	4	7
ΖΕΜΗ	-	4	3	-	-	-	-	2	4	-	2
ΑΘΗΝΑ	8	3	8	-	9	8	8	7	-	-	7

Πίνακας 21 : Βαθμολογίες χρηστών στις ταινίες σε κλίμακα 0-10

	REQUIEM FOR A DREAM	FIGHTCLUB	INCEPTION	MATCH POINT	BLACK SWAN	SILENCE OF THE LAMBS	LOVE ME IF YOU DARE	SUPERMAN	SLUMDOG MILLIONAIRE	THE SOCIAL NETWORK
ΑΝΝΥ	3	-	-	4	-	-	3	4	-	-
ΚΑΤΕΡΙΝΑ	3	-	-	3	5	-	4	-	5	-
ΔΗΜΗΤΡΗΣ	2	-	-	6	6	-	2	6	-	-
ΚΡΙΝΑ	5	8	7	6	6	6	-	6	8	4
ΝΙΚΟΣ	-	7	-	2	5	7	4	2	-	-
ΘΕΟΔΟΣΙΑ	-	7	-	4	6	5	5	-	5	5
ΝΑΣΟΣ	2	-	-	5	-	2	-	3	-	-
ΓΙΑΝΝΗΣ	5	6	-	2	-	2	-	4	-	-
ΝΙΟΒΗ	-	7	5	-	5	4	6	-	6	6
ΚΩΣΤΑΣ	5	6	5	2	-	2	-	2	2	-
ΒΑΣΙΛΗΣ	-	8	8	8	-	6	4	6	8	-
ΖΕΜΗ	2	-	3	5	5	-	5	2	-	-
ΑΘΗΝΑ	-	8	8	8	6	6	-	4	8	-

Πίνακας 21 : Συνέχεια βαθμολογιών χρηστών

Από τις βαθμολογίες λοιπόν που έδωσαν οι χρήστες, τα αποτελέσματα είναι ιδιαιτέρως ενδιαφέροντα στις εξής κατευθύνσεις:

Α) παρατηρώντας τον πίνακα 21, φαίνεται ότι ορισμένοι χρήστες έχουν ίδιο ή και μεγαλύτερο βαθμό σε ταινίες που δεν έχουν κάνει «Like» σε σχέση με εκείνες που έχουν κάνει. Άρα μια άλλη πιθανή ερμηνεία της έλλειψης «like» είναι ότι ο χρήστης μπορεί να μην θυμόταν την συγκεκριμένη ταινία παρότι του αρέσει, να μην είχε υπ'όψιν του την συγκεκριμένη σελίδα στο Facebook ή και να μην είχε τον χρόνο να ταξινομήσει όλες του τις ταινίες. Η επισήμανση μίας αγαπημένης του ταινίας μπορεί να έπεσε στην υποληψή του μέσα από το ίδιο του το δίκτυο

Β) Επιβεβαιώνεται η αρχική απλή υπόθεση, ότι η έλλειψη «Like» σε μια ταινία που δηλώνεται με τον βαθμό 0 στον πίνακα 18, δεν σημαίνει απαραίτητα ότι δεν άρεσε στον χρήστη, αλλά μπορεί και να μην την έχει δει.

Γ) Μια ιδιαίτερα σημαντική παρατήρηση που πρέπει να γίνει είναι ότι ακόμα και στις ταινίες που οι χρήστες είχαν κάνει «like» (σημειωμένα με γκρι κελιά στον πίνακα 21) κάποιοι από τους βαθμούς είναι χαμηλοί. Παρατηρείται λοιπόν ότι ακόμα και η ύπαρξη «like» δεν είναι αρκετή για να εξαχθεί το συμπέρασμα ότι στον χρήστη σίγουρα άρεσε η ταινία, αφού το «like» μπορεί απλά να υποδηλώνει ότι ο χρήστης γνωρίζει για την ταινία (awareness) ή ότι είναι ένα αντικείμενο συζήτησης (Social Object) που τον συνδέει με άλλους χρήστες. Επιπλέον παράγοντες υπεισέρχονται στην συγκεκριμένη βαθμολόγηση, παράγοντες κοινωνιολογικής ή και ψυχολογικής φύσης. Όμως κατά πόσο αυτοί οι παράγοντες μπορούν να κρυφτούν πίσω από μία σειρά κοινών Like θα μπορούσε να είναι ενδιαφέρον αντικείμενο περαιτέρω μελέτης-επέκτασης.

Δ) Με μία πρώτη ματιά, τα αποτελέσματα του αλγορίθμου Knn για κ>1 δίνουν αποτελέσματα που είναι πιο κοντά στις επιλογές του χρήστη. Προκειμένου να επιβεβαιωθεί αυτό χρειάστηκε να γίνει περαιτέρω ανάλυση.

Στη συνέχεια, γίνεται σύγκριση ανάμεσα στις πραγματικές τιμές που έχουν βάλει ως βαθμούς οι χρήστες και στις τιμές που προέβλεψε ο αλγόριθμος για κ=4 και κ=7 με βάση τα δεδομένα από τα Like των ταινιών. Για την αξιολόγηση των διαφορετικών αποτελεσμάτων χρησιμοποιείται το Μέσο Απόλυτο Σφάλμα που έχει αναλυθεί στην ενότητα 5.4 και δίνεται από τον τύπο :

$$\mathrm{MAE}(f) = \frac{1}{|\mathcal{R}_{\text{test}}|} \sum_{r_{ui} \in \mathcal{R}_{\text{test}}} |f(u,i) - r_{ui}|$$

	MAE	
	K=4	**K=7**
ΑΝΝΥ	1,583	2,06
ΚΑΤΕΡΙΝΑ	1,014	1,246
ΔΗΜΗΤΡΗΣ	1,48	1,675
ΚΡΙΝΑ	1,022	1,626
ΝΙΚΟΣ	1,043	1,046

ΘΕΟΔΟΣΙΑ	0,82	1,51
ΝΑΣΟΣ	1,31	1,38
ΓΙΑΝΝΗΣ	1,026	1,317
ΝΙΟΒΗ	0,572	1,002
ΚΩΣΤΑΣ	0,968	0,978
ΒΑΣΙΛΗΣ	0,518	1,456
ΖΕΜΗ	0,818	1,066
ΑΘΗΝΑ	0,486	1,316

Πίνακας 22 : Μέσο Απόλυτο σφάλμα προβλέψεων για κ=4 και κ=7

Παρατηρείται λοιπόν ότι για κ=4 και με τα δεδομένα που χρησιμοποιούνται, ο αλγόριθμος είναι πολύ ακριβέστερος σε σχέση με την λειτουργία του με κ=7. Για τη συγκεκριμένη τουλάχιστον ομάδα ατόμων μπορεί να βρεθεί μία μεταβλητή κ που να μειώνει κατά πολύ την απόκλιση πραγματικής και προβλεπόμενης βαθμολογίας.

Γίνεται λοιπόν χρησιμοποιώντας τα αποτελέσματα του πίνακα 19 (κ=4) να παρουσιαστούν ταξινομημένες οι ταινίες που οι χρήστες δεν έχουν δει ως προτάσεις για εκείνους. Ενδεικτικά, παρουσιάζονται οι βαθμολογίες που προβλέπει ο αλγόριθμος για τις ταινίες που δεν έχουν δει οι χρήστες Αθηνά και Κώστας, μιας και έχουν σχετικά μικρό μέσο απόλυτο σφάλμα :

ΚΩΣΤΑΣ
Kill Bill : 8/10
The Godfather : 7,5/10
Rock n Rolla : 7,5/10
Black Swan : 5/10
Love me if you dare : 5/10
The seventh seal : 2,85/10
The Lord of the Rings : 2,85/10
127 Hours : 1,66/10
The Social Network : 1,66/10
Citizen Kane : 1,43/10

Πίνακας 23 : Προβλέψεις ταινιών για τον χρήστη Κώστα

ΑΘΗΝΑ
Love me if you dare : 5/10
The Social Network : 5/10
The seventh seal : 2,5/10
Lock,Stock & 2 Smoking Barrels : 2,5/10
Requiem for a dream : 2,5/10
The Lord of the Rings : 0/10

Πίνακας 24 : Προβλέψεις ταινιών για τον χρήστη Αθηνά

Παρατηρείται ότι ο αλγόριθμος προβλέπει ότι οι ταινίες που δεν έχει δει η Αθηνά πιθανότατα να μην της αρέσουν ιδιαίτερα, αντίθετα με τον Κώστα. Η διαφορά στην πρώτη ταινία των προτάσεων για τους δυο χρήστες οφείλεται απλά στο γεγονός ότι οι 4 (λόγω κ=4) χρήστες με τους οποίους συνδέθηκε ο Κώστας είχαν δώσει καλή βαθμολογία στην ταινία Kill Bill (π.χ. μπορεί να είχε ο αλγόριθμος από τους 4 γείτονες τρία 1 και ένα 0 και με τον σταθμισμένο μέσο όρο να βγήκε το 0,8/1 για τον Κώστα), ενώ οι 4 χρήστες που υπολογίστηκαν ως γείτονες της Αθηνάς δεν είχαν κάνει οι περισσότεροι «Like» στην ταινία «Love me if you dare» ή ακόμα και αν είχαν κάνει οι 3 στους 4, το 0 από την έλλειψη «Like» επικράτησε περισσότερο στον σταθμισμένο Μέσο Όρο για την εξαγωγή της πρόβλεψης γιατί ο χρήστης που έβαλε το 0 είχε μεγαλύτερο βαθμό ομοιότητας με την Αθηνά από ότι οι άλλοι.

Από τα παραπάνω εξάγεται λοιπόν ότι στον βαθμό που το δίκτυο είναι δεδομένο, μπορεί να βρεθεί το κατάλληλο κ που θα δώσει μία κατάταξη των ταινιών που προτείνεται να δει, σε φθίνουσα σειρά. Άρα τα δεδομένα από το Facebook μπορούν να χρησιμοποιηθούν σε έναν βαθμό, όπως στα περιορισμένα πλαίσια της συγκεκριμένης εργασίας διαφαίνεται. Σίγουρα δε χρίζει περαιτέρω μελέτης και έρευνας.

7

Συμπεράσματα – Μελλοντικές Επεκτάσεις

Στο κεφάλαιο αυτό παρατίθεται μια σύνοψη των παρατηρήσεων που έγιναν και των συμπερασμάτων που εξήχθηκαν από την ανάλυση των διαφορετικών μεθόδων που χρησιμοποιούνται για τα Συστήματα Προτάσεων, κυρίως δε από την εκτέλεση του πειράματος, ενώ στη συνέχεια αναφέρονται και θα προτείνονται μελλοντικές τάσεις στην σύσταση προτάσεων στα Κοινωνικά Μέσα.

7.1 Συμπεράσματα

Μέσα από την ανάλυση βιβλιογραφικών και ερευνητικών πηγών μπορεί κανείς να προβεί στο συμπέρασμα ότι το Συνεργατικό Φιλτράρισμα στα Συστήματα Προτάσεων χρησιμοποιείται πολύ περισσότερο και πολύ ευκολότερα σε σχέση με τα Συστήματα Προτάσεων με βάση το Περιεχόμενο. Ορισμένα από τα καλύτερα παραδείγματα συστημάτων προτάσεων όπως αυτό των Google, YouTube ή Amazon εφαρμόζουν ευρέως συνεργατικό φιλτράρισμα με βάση την μνήμη για την εξαγωγή προβλέψεων. Ο βασικός λόγος που συμβαίνει αυτό είναι επειδή το Συνεργατικό Φιλτράρισμα στηρίζεται στους χρήστες, σε παρελθοντικές προτιμήσεις τους και σε κοινά ενδιαφέροντα με άλλους χρήστες και όχι στο περιεχόμενο των προϊόντων, το οποίο αλλάζει δυναμικά και πολύ συχνά δεν είναι αρκετά σαφές ώστε να οδηγήσει σε ακριβείς προβλέψεις. Επιπλέον, η επιλογή memory-based συστημάτων οφείλεται μεταξύ άλλων στο ότι είναι εύκολα και απλά τόσο στην δημιουργία όσο και στην εφαρμογή τους, καθώς και στο ότι είναι σταθερά και δεν επηρεάζονται σε σημαντικό βαθμό από το συνεχώς μεταβαλλόμενο dataset (προσθήκη νέων χρηστών, προϊόντων ή βαθμολογιών). Συνοπτικά τα συμπεράσματα των μεθόδων που χρησιμοποιούνται στο Συνεργατικό Φιλτράρισμα και τα RS με βάση το Περιεχόμενο καθώς και των κατηγοριών του Συνεργατικού Φιλτραρίσματος φαίνονται στους δύο παρακάτω πίνακες:

	ΜΕΘΟΔΟΣ		ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
RS ΜΕ ΒΑΣΗ ΤΟ ΠΕΡΙΕΧΟΜΕΝΟ	Μέθοδος με λέξεις-κλειδιά		• Μεγάλη ακρίβεια, υψηλή απόδοση	• Δεν ανταποκρίνεται αποδοτικά όταν απαιτούνται προηγμένα χαρακτηριστικά
	Naïve Bayes	Πολυμεταβλητές Bernoulli	• Απομονώνει θόρυβο • Χειρίζεται missing values	• Χαμηλή επίδοση
		Πολυωνυμικό μοντέλο	• Απομονώνει θόρυβο • Χειρίζεται missing values • Καλύτερες επιδόσεις	• Αδυναμίες όταν μικρό το training set • Αδυναμίες όταν τα έγγραφα έχουν διαφορετικά μεγέθη
	Γραμμικοί Ταξινομητές		• Μπορούν να εφαρμοστούν online σε πραγματικό χρόνο • Ακρίβεια στον χωρισμό των δεδομένων σε κατηγορίες • Χειρίζονται το overfitting	• Χαμηλή γενική απόδοση
	Αλγόριθμος Rocchio		• Βελτιώνει την απόδοση της διαδικασίας ανάκτησης πληροφορίας	• Απαιτεί ανατροφοδότηση από τους χρήστες
RS ΜΕ ΣΥΝΕΡΓΑΤΙΚΟ ΦΙΛΤΡΑΡΙΣΜΑ	Gaussian μέθοδος κανονικοποίησης		• Λαμβάνει υπ'όψιν την μετατόπιση μέσων βαθμολογιών και τις διαφορετικές βαθμολογικές κλίμακες	• Δεν εξετάζει την διακύμανση των βαθμολογιών
	Μέθοδος mean-centering		• Το αποτέλεσμα μπορεί να εκτιμηθεί άμεσα	• Δεν εξετάζει την διακύμανση των βαθμολογιών
	Κανονικοποίηση z-score		• Αντισταθμίζει τα offsets • Εξετάζει διακύμανση βαθμολογιών	• Αδυναμίες αν αραιά δεδομένα
	Cosine-based ομοιότητα		• Εύκολη στην εφαρμογή	• Δεν εξετάζει επίδραση του μέσου και της διακύμανσης
	Pearson correlation		• Λαμβάνει υπ'όψιν αποκλίσεις βαθμολογιών από μέσο όσο	• Δεν εξετάζει τις διαφορετικές βαθμολογικές κλίμακες
	Adjusted cosine-based ομοιότητα		• Λαμβάνει υπ'όψιν τις διαφορετικές βαθμολογικές κλίμακες	
	Singular Value Decomposition		• Περιορίζει το πρόβλημα των αραιών δεδομένων	• Σε εφαρμογές σε άμεσα δεδομένα παρουσιάζει αδυναμίες λόγω missing values

Πίνακας 25 : Πλεονεκτήματα / Μειονεκτήματα μεθόδων Content-based και Collaborative Filtering Συστημάτων Προτάσεων

	ΚΑΤΗΓΟΡΙΑ	ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
ΣΥΝΕΡΓΑΤΙΚΟ ΦΙΛΤΡΑΡΙΣΜΑ	Model-based	• Χειρίζονται αραιά δεδομένα • Εύκολη εφαρμογή • Συνεχής βελτίωση απόδοσης μέσω training	• Υψηλό κόστος λόγω συχνού training • Μπορεί να χαθεί χρήσιμη πληροφορία λόγω συχνών μειώσεων διαστάσεων
	Memory-based	• Απλότητα • Παρέχουν εξηγήσεις • Σταθερότητα στις συνεχεία μεταβολές • Υψηλή απόδοση • Χαμηλό κόστος-μικρή μνήμη	• Cold-start πρόβλημα • Περιορισμένη κάλυψη • Αραιά δεδομένα
	User-based	• Προτιμότερα όταν προϊόντα ξεπερνούν κατά πολύ τους χρήστες (καλύτερη απόδοση) • Σταθερότητα όταν προϊόντα αλλάζουν συνεχώς σε σχέση με χρήστες • Καινοτομία	• Συνήθως δεν παρέχουν εξηγήσεις • Χαμηλή απόδοση όταν στο dataset οι χρήστες πολύ περισσότεροι από προϊόντα • Ασταθή όταν οι χρήστες μεταβάλλονται πολύ περισσότερο από τα προϊόντα
	Item-based	• Προτιμότερα όταν χρήστες ξεπερνούν κατά πολύ τα προϊόντα (καλύτερη απόδοση) • Σταθερότητα όταν χρήστες αλλάζουν συνεχώς σε σχέση με προϊόντα • Μεγαλύτερη ακρίβεια • Λιγότερη μνήμη-χρόνο • Παρέχουν εξηγήσεις	• Δεν κάνουν καινοτόμες προτάσεις • Χαμηλή απόδοση όταν στο dataset τα προϊόντα πολύ περισσότεροι από χρήστες • Ασταθή όταν τα προϊόντα μεταβάλλονται πολύ περισσότερο από τους χρήστες

Πίνακας 26 : Πλεονεκτήματα / Μειονεκτήματα ειδών Collaborative Filtering Συστημάτων Προτάσεων

Μια άλλη σημαντική παρατήρηση που πρέπει να γίνει σχετίζεται με την φύση του Facebook στην προσπάθεια που έγινε να χρησιμοποιηθούν δεδομένα του Facebook σε μια εξωτερική εφαρμογή για την εξόρυξη επιπλέον πληροφορίας. Το Facebook , ένα από τα μεγαλύτερα Κοινωνικά Μέσα που ασκεί επιρροές σε μεγάλο βαθμό και παρουσιάζει σημαντικό ενδιαφέρον λόγω της διεισδυτικότητάς του, επιτρέπει την συσχέτιση μεταξύ των χρηστών και την παρατήρηση των δεδομένων τους, ωστόσο δεν παρέχει την δυνατότητα σε έναν εξωτερικό παρατηρητή να εισβάλει με δυναμικό τρόπο στα δεδομένα αυτά και να τα χρησιμοποιήσει σε μια εξωτερική εφαρμογή με σκοπό την μελέτη των συσχετίσεων των χρηστών και την εύρεση νέας πληροφορίας από την ήδη υπάρχουσα. Παρατηρείται λοιπόν μια αδυναμία να προσαρμοστεί η συγκεκριμένη εφαρμογή στα δεδομένα του Facebook και αυτό οφείλεται στην αδυναμία της ερμηνείας του «like» που παρέχει το Facebook. Η απουσία του «Like» σε κάποιον σύνδεσμο δεν σχετίζεται απαραίτητα με την αρνητική στάση του χρήστη απέναντι στον συγκεκριμένο σύνδεσμο, αλλά μπορεί να εξηγείται από το γεγονός ότι δεν έπεσε στην αντίληψή του. Ομοίως, το κουμπί «like» δεν χρησιμοποιείται μόνο για να δείξει μια θετική εντύπωση, αλλά από κάποιους χρήστες χρησιμοποιείται και ως σημάδι επίγνωσης, δηλαδή ότι είδαν τον συγκεκριμένο σύνδεσμο χωρίς απαραίτητα να τους άρεσε κιόλας. Στην εφαρμογή λοιπόν έγινε αναγκαστικά η υπόθεση ότι η έλλειψη «like» δηλώνει ότι στον χρήστη δεν άρεσε η συγκεκριμένη σελίδα και να αποτυπωθεί αυτή η απουσία με τον βαθμό 0 στον πίνακα δεδομένων, καθώς και ότι η παρουσία «like» φανερώνει ότι στον χρήστη άρεσε η εκάστοτε σελίδα και έτσι να μπει ο βαθμός 1 στο αντίστοιχο κελί του πίνακα δεδομένων. Όταν λοιπόν ζητήθηκε από τους χρήστες να βάλουν πραγματικούς βαθμούς από 0 έως 10 στις ίδιες ταινίες διαπιστώθηκαν αυτά ακριβώς που αναφέρθηκαν προηγουμένως για την συμπεριφορά των χρηστών στη χρήση του «like». Φυσικά τα αποτελέσματα του αλγορίθμου θα ήταν διαφορετικά αν το Facebook είχε κατ' αντιστοιχία του «like» ένα κουμπί «dislike» που θα δήλωνε άμεσα ότι στον χρήστη δεν αρέσει η σελίδα, ή ακόμα καλύτερα αν έδινε την δυνατότητα βαθμολόγησης σε βαθμωτή κλίμακα (εύρους π.χ. 0-5), στην οποία το 0 δηλώνει ότι στον χρήστη δεν αρέσει η σελίδα, το 5 δηλώνει ότι του άρεσε πολύ ενώ η απουσία βαθμολογίας φανερώνει ότι η σελίδα δεν έχει πέσει στην αντίληψη του χρήστη.

Όσον αφορά λοιπόν την διεξαγωγή του πειράματος, υπήρξαν αυτές οι δυσκολίες σχετικά με την απόδοση του Like καθώς και με το ότι ο αλγόριθμος δεν δέχεται την τιμή Null ως είσοδο στα γνωρίσματα. Παρατηρήθηκε ότι για κ=1, που σημαίνει ότι η κατηγορία του προς πρόβλεψη αντικειμένου είναι η κατηγορία του κοντινότερου δείγματος με το οποίο το αντικείμενο μοιάζει περισσότερο, ο αλγόριθμος λειτουργούσε με άριστη ακρίβεια προβλέποντας τιμές στην βαθμολογία ίδιες με εκείνες που είχε δεχτεί ως input. Γνωρίζοντας όμως ότι τα 0 και 1 που έχουν μπει ως δεδομένα μπορεί να έχουν περισσότερες από μια ερμηνείες, κρίθηκε σωστό να αυξηθεί το κ ώστε οι τιμές που προβλέπει ο αλγόριθμος να μην ανήκουν στην δυαδική κατηγοριοποίηση αλλά να ανταποκρίνονται περισσότερο σε μια ρεαλιστική βαθμολόγηση με διαβαθμήσεις. Ζητώντας λοιπόν από τους χρήστες τιμές σε 10βάθμια κλίμακα και συγκρίνοντας στη συνέχεια αυτές τις πραγματικές βαθμολογίες με εκείνες που προέβλεψε ο αλγόριθμος για κ=4 και κ=7 ουσιαστικά αυθαιρετούμε, αφού πρώτον ζητάμε άλλης κλίμακας βαθμολογία εκτός εκείνης που παρέχει το Facebook (like ή όχι like), και δεύτερον επειδή κανονικά η αξιολόγηση του αλγορίθμου με χρήση του Μέσου Απόλυτου σφάλματος γίνεται συγκρίνοντας τα αποτελέσματα που δίνει ο αλγόριθμος βάσει των δεδομένων εισόδου με τα ίδια τα δεδομένα εισόδου για να γίνει έλεγχος αν προέβλεψε σωστά, και όχι συγκρίνοντάς τα με άλλα πραγματικά δεδομένα. Στόχος όμως ήταν να φανεί κατά πόσο τα αποτελέσματα που εξάγονται με είσοδο τα δεδομένα του Facebook ανταποκρίνονται στις πραγματικές προτιμήσεις των χρηστών (οι οποίες φανερώνονται με την βαθμολογία από 0 έως 10 ενώ δεν φανερώνονταν πλήρως με τα «like» και τα δεδομένα σε

0/1 του Facebook), για να φανεί κατά πόσο οι φίλοι επηρεάζουν την βαθμολογία μιας ταινίας. Μετά από την αξιολόγηση του αλγορίθμου λοιπόν σε σχέση με τα πραγματικά δεδομένα, φαίνεται ότι τα αποτελέσματα που βγήκαν για κ=4 χρησιμοποιώντας δεδομένα του Facebook σε μια εξωτερική εφαρμογή όπως η συγκεκριμένη ήταν πολύ κοντά στις τιμές που έδωσαν οι χρήστες όταν είχαν την δυνατότητα να δηλώσουν ξεκάθαρα την ένταση που μια ταινία τους άρεσε ή όχι (με βαθμούς από 0 έως 10), καθώς και την δυνατότητα να δηλωθεί με κενό όποια ταινία δεν είχαν δει.

7.2 Μελλοντικές επεκτάσεις

Μια μελλοντική τάση στα Συστήματα Προτάσεων που μπορεί να ανοίξει πολλούς δρόμους είναι η εξαγωγή συστάσεων στα Κοινωνικά Μέσα με χρήση Semantics σε Graph Database. Στις επόμενες ενότητες μελετώνται οι έννοιες της σημασιολογίας και των βάσεων δεδομένων με γράφημα, καθώς και η χρήση της σημασιολογίας στα Συστήματα Προτάσεων, και στη συνέχεια αναφέρεται μια πρώτη προσπάθεια που έγινε για σύσταση προτάσεων σε Graph Database και μελλοντικές επεκτάσεις αυτής. Πιστεύουμε ότι η χρήση δεδομένων που προέρχονται αυτόματα από τα κοινωνικά μέσα αλλά και η εξαγωγή προτάσεων στηριγμένα σε κοινωνικά χαρακτηριστικά αποτελούν το μέλλον στο συγκεκριμένο επιστημονικό πεδίο, και συγκεντρώνουν ήδη ιδιαίτερης προσοχής.

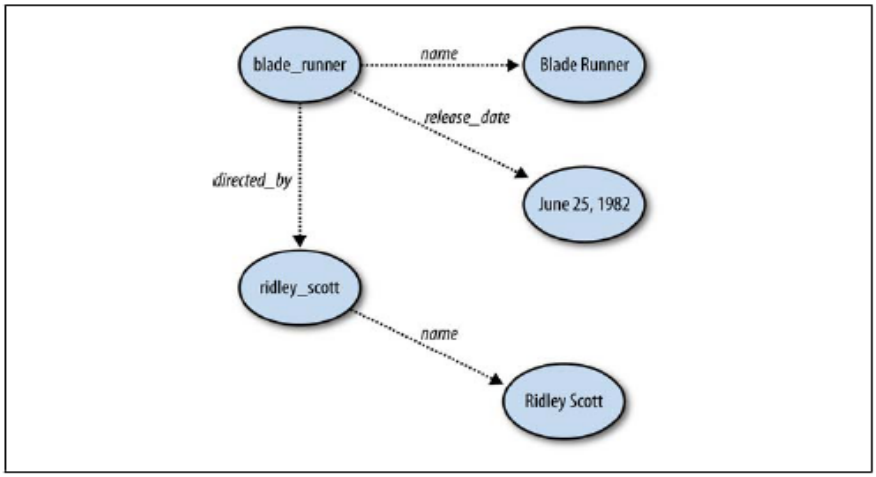
7.2.1 Σημασιολογία (Semantics)

Η σημασιολογία σχετίζεται με την μελέτη του νοήματος. Επικεντρώνεται στις σχέσεις μεταξύ λέξεων, φράσεων ή συμβόλων, και στο τι αυτές οι σχέσεις σημαίνουν. Η σημασιολογία αφορά πολλούς διαφορετικούς τομείς έρευνας, όπως την γλωσσολογία, την λεξικολογία, την ψυχολογία ή την επιστήμη των υπολογιστών. Τα σύμβολα μπορεί να αναφέρονται σε αντικείμενα ή έννοιες, ενώ από μια ακολουθία συμβόλων μπορεί να εξάγεται νόημα. Η σημασιολογία φυσικά πρώτα από όλα αφορά και την φυσική μας γλώσσα. Από μια απλή πρόταση της μορφής Υποκείμενο – Ρήμα – Αντικείμενο μπορεί να εξαχθεί πληροφορία, να βρεθεί η σχέση μεταξύ του Υποκειμένου και του Αντικειμένου και τελικά να κατανοηθεί το νόημα της πρότασης. Η σημασιολογία λοιπόν είναι η διαδικασία της μετάδοσης νοήματος, η οποία μπορεί να οδηγήσει και σε πράξεις και να επηρεάσει συμπεριφορές.

Όσον αφορά την επιστήμη των υπολογιστών, η σημασιολογία χρησιμοποιείται για να παρουσιάζεται, να συνδυάζεται και να μοιράζεται η γνώση ανάμεσα σε κοινότητες μηχανών. Ο προγραμματισμός για παράδειγμα με την χρήση μεταβλητών περιλαμβάνει σημασιολογία, καθώς η κάθε μεταβλητή εκπροσωπεί μια έννοια. Ο όρος Semantic Web είναι μια επέκταση του Παγκόσμιου Ιστού. Ο Semantic Ιστός είναι ένα ιστός δεδομένων που επιτρέπει στις μηχανές να καταλαβαίνουν την σημασιολογία της πληροφορίας που παρέχει ο Παγκόσμιος Ιστός. Με τον Σημασιολογικό Ιστό, πλέον οι εφαρμογές μπορούν να συνδυάζουν δεδομένα με καινούριους τρόπους και έτσι δίνεται η δυνατότητα στους χρήστες να κάνουν συνδέσεις και να καταλάβουν σχέσεις οι οποίες πριν ήταν κρυμμένες. Επεκτείνεται λοιπόν το δίκτυο των ιστοσελίδων υπερσύνδεσης μέσω της εισαγωγής αναγνωρίσιμων από το μηχάνημα μεταδεδομένων σχετικά με τις ιστοσελίδες και τον τρόπο που αυτές συνδέονται μεταξύ τους. Οι χρήστες λοιπόν μπορούν να

αξιοποιήσουν στο έπακρον τις δυνατότητες που προσφέρει ο Ιστός, αφού εξάγεται γνώση από την πληροφορία, κάνοντας πολύ εύκολο για τους χρήστες να βρουν, να μοιραστούν και να συνδυάσουν την πληροφορία. Με τον τρόπο που οι άνθρωποι μπορούν να εξάγουν γνώση από τα δεδομένα, έτσι πλέον με τον Σημασιολογικό Ιστό η πληροφορία μπορεί να ερμηνευτεί και από τα μηχανήματα. Οι υπολογιστές γίνονται ικανοί να αναλύσουν τα δεδομένα του Ιστού, όπως το περιεχόμενο, τις συνδέσεις και τις αλληλεπιδράσεις μεταξύ ανθρώπων και υπολογιστών. Ο Σημασιολογικός Ιστός έχει αναφερθεί και ως εργαλείο του Web 3.0.

Τα σημασιολογικά δίκτυα και μοντέλα δεδομένων περιγράφουν συγκεκριμένες μορφές μοντέλων δεδομένων οι οποίες χαρακτηρίζονται από την χρήση κατευθυνόμενων γράφων. Οι κατευθυνόμενοι γράφοι είναι από τις πιο δημοφιλείς μορφές αναπαράστασης δεδομένων. Σε αυτό το σημείο πρέπει να γίνει αναφορά στην έννοια της τριπλέτας (triple), του θεμελιώδη λίθου στις σημασιολογικές αναπαραστάσεις. Μια τριπλέτα είναι ένα format που αποτελείται από ένα υποκείμενο (subject), ένα κατηγόρημα (predicate) και ένα αντικείμενο (object), κατά αναλογία με την σύνταξη υποκείμενο - ρήμα – αντικείμενο της φυσικής γλώσσας που αναφέραμε προηγουμένως. Το υποκείμενο σε μια τριπλέτα αναφέρεται σε μια οντότητα, όπως για παράδειγμα σε άνθρωπο, μέρος, έννοια ή χρονική περίοδο. Το κατηγόρημα είναι η ιδιότητα της οντότητας στην οποία αναφέρεται, όπως το όνομα ενός ανθρώπου. Τα αντικείμενα είναι είτε οντότητες, οι οποίες σε άλλες τριπλέτες μπορούν να γίνουν υποκείμενα, είτε τιμές, όπως νούμερα ή strings. Οι γράφοι λοιπόν για την αναπαράσταση των δεδομένων χρησιμοποιούν τριπλέτες. Κάθε κόμβος του γράφου είναι το υποκείμενο ή το αντικείμενο (έννοιες ή οντότητες), ενώ οι ακμές αντιπροσωπεύουν το κατηγόρημα, δηλαδή την σχέση που συνδέει το υποκείμενο με το αντικείμενο.

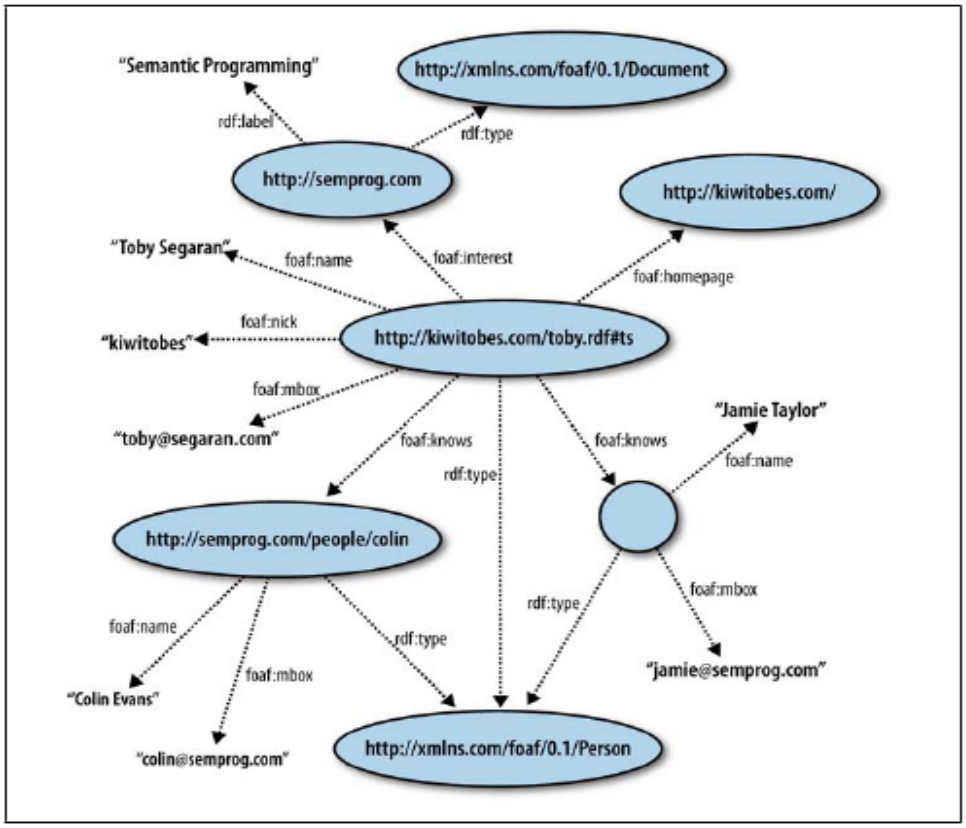


Εικόνα 17 : Παράδειγμα γράφου με πληροφορίες για την ταινία blade runner [27]

Άπαξ και έχουν φορτωθεί λοιπόν τα δεδομένα σε έναν γράφο, το ερώτημα είναι πώς γίνεται να μοιραστούν οι χρήστες αυτά τα δεδομένα και να τα κάνουν διαθέσιμα και σε άλλους ανθρώπους ; Η απάντηση σε αυτό το ερώτημα είναι το Πλαίσιο Περιγραφής Πηγής ή RDF (Resource Description Framework). Το RDF παρέχει ένα συγκεκριμένο format για την αναπαράσταση των γράφων δεδομένων και για να μοιράζονται αυτά τα δεδομένα άλλοι άνθρωποι ή και μηχανές.

Πρόκειται λοιπόν για μια γλώσσα έκφρασης των μοντέλων δεδομένων με χρήση δηλώσεων όπως οι τριπλέτες. Χρησιμοποιεί διάφορα μοντέλα σύνταξης και είναι μια μέθοδος για την εννοιολογική περιγραφή και την μοντελοποίηση της πληροφορίας. Παραδείγματα γνωρισμάτων του RDF είναι τα rdf:type, rdf:subject, rdf:predicate, rdf:object ή rdf:value. Το RDF αντιλαμβάνεται το ο,τιδήποτε σαν μια πηγή η οποία μπορεί να προσδιοριστεί από ένα παγκόσμιο αναγνωριστικό πηγής ή Universal Resource Identifier (URI). Το URI μιας πηγής που εμφανίζεται σε μια rdf δήλωση είναι γνωστό ως URI reference (URIref) για τον συγκεκριμένο κόμβο του γράφου. Συχνά, στο πλαίσιο RDF γίνεται συντόμευση των URIs αναθέτοντας ένα namespace στο βασικό URI και γράφοντας μόνο το κομμάτι του αναγνωριστικού που μεταβάλλεται. Για παράδειγμα, στο βασικό URI http://w3.org/1999/02/22-rdf-syntax-ns# μπορεί να ανατεθεί το Namespace rdf, ώστε ένα κατηγόρημα όπως το http://w3.org/1999/02/22-rdf-syntax-ns#type να συντομευθεί ως rdf:type.

Το δίκτυο των κοινωνικών σχέσεων των ανθρώπων συχνά αναπαρίσταται με γράφους. Οι αναγνωρίσιμοι από μηχανήματα γράφοι φίλων έχουν εξελιχθεί παράλληλα με το RDF, κάνοντας τους κοινωνικούς γράφους ένα από τα πιο ευρέως διαθέσιμα datasets του internet. Σιγά-σιγά λοιπόν οι σχέσεις που εκφράζονται σε έναν κοινωνικό γράφο έγιναν ένα σύνολο γνωστών κατηγορημάτων, τα οποία δημιούργησαν ένα λεξιλόγιο γνωστό με την έκφραση Friend of a Friend ή FOAF.



Εικόνα 18 : Παράδειγμα FOAF γράφου του χρήστη Toby [27]

Ενώ το μοντέλο δεδομένων που χρησιμοποιεί το RDF είναι πολύ απλό, η σειριακή αναπαράσταση είναι πιο περίπλοκη όταν ο RDF γράφος στέλνεται εντός ενός δικτύου, λόγω των διαφορετικών μεθόδων που χρησιμοποιούνται για να συμπυκνωθεί η πληροφορία αλλά και να παραμείνει αναγνώσιμη. Για αυτόν τον λόγο υπάρχουν RDF βιλβιοθήκες ανοιχτής πηγής σχεδόν για κάθε σύγχρονη γλώσσα προγραμματισμού που διαχειρίζονται τα σειριακά formats. Τα βασικά σειριακά formats είναι τα : N-Triples, N3, RDF/XML και RDFa. Η μέθοδος N-Triples είναι η πιο απλή μορφή serialization, έχοντας ως αποτελέσματα τριπλέτες με υποκείμενο, κατηγόρημα και αντικείμενο. Ωστόσο, στα αποτελέσματα παρουσιάζεται σε μεγάλο βαθμό επανάληψη, για αυτό αναπτύχθηκε το format N3 που επιτρέπει να οριστεί ένα URI πρόθεμα και να αναγνωριστούν οντότητες URIs σε σχέση με τα προθέματα που έχουν δηλωθεί στην αρχή του αρχείου. Μια άλλη μέθοδος αναπαράστασης των RDF γράφων είναι η RDF/XML, η οποία περιλαμβάνει την περιγραφή του RDF ως ένα μοντέλο δεδομένων και του XML ως μια έκφραση των RDF μοντέλων. Τέλος, η μέθοδος RDFa δεν είναι ακριβώς ένα σειριακό format για RDF, αλλά περισσότερο ένας τρόπος συμπλήρωσης των XHTML ιστοσελίδων με RDF δεδομένα. Το RDFa χρησιμοποιεί ένα υποσύνολο των XML χαρακτηριστικών γνωρισμάτων για να ορίσει την σημασιολογία της πληροφορίας που παρουσιάζεται.

Φυσικά, η ύπαρξη ενός γράφου δεν έχει ιδιαίτερο νόημα αν δεν μπορούν να γίνουν ερωτήματα πάνω στον γράφο για τη ανάκτηση των πηροφοριών πυο μας ενδιαφέρουν. Η γλώσσα ερωτημάτων των RDF γράφων είναι η SPARQL (Simple Protocol And Rdf Query Language). Τα SPARQL ερωτήματα προσπαθούν να ταιριάξουν πρότυπα του γράφου και να δεσμεύσουν τις μεταβλητές-μπαλαντέρ με την εύρεση της λύσης στο ερώτημα. Η γλώσσα SPARQL παρέχει τέσσερις μορφές ερωτημάτων :

- SELECT query : Επιστρέφει ως αποτέλεσμα ακατέργαστες τιμές στην μορφή πίνακα. Ένα τέτοιο ερώτημα επιτρέπει να αναγνωριστεί ένα υποσύνολο των μεταβλητών που χρησιμοποιούνται στον γράφο, των οποίων τις δεσμεύσεις ζητάμε να μας επιστραφούν ως λύση.
- CONSTRUCT query : Χρησιμοποιείται για να εξάγει πληροφορία και να μετατρέψει τα αποτελέσματα σε έγκυο RDF. Κατασκευάζει λοιπόν από το set των λύσεων έναν νέο γράφο, χρησιμοποιώντας τις πρότυπες τριπλέτες που καθορίζονται στο ερώτημα.
- ASK query : Το ερώτημα αυτό ελέγχει αν ένα συγκεκριμένο πρότυπο μπορεί να βρεθεί σε έναν γράφο. Επιστρέφει ως αποτέλεσμα μια τιμή True/False.
- DESCRIBE query : Το ερώτημα αυτό επιστρέφει όλες τις χρήσιμες πληροφορίες που έχει ο γράφος για μια πηγή, βοηθώντας να κατανοηθεί το περιεχόμενο των πηγών.

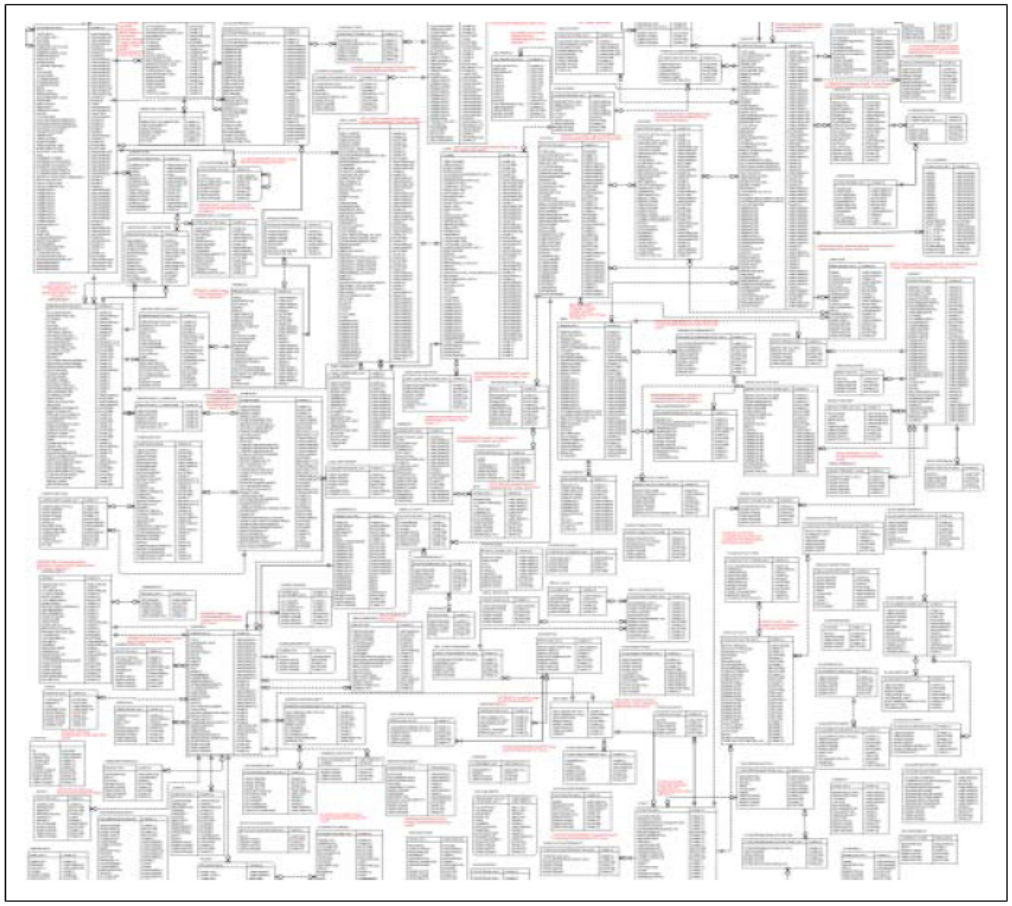
Καθεμιά από αυτές τις μορφές ερωτημάτων δέχεται ένα μπλοκ WHERE που περιορίζει το ερώτημα. Στην περίπτωση του DESCRIBE query το μπλοκ WHERE είναι προαιρετικό. Επίσης, η SPARQL παρέχει την δυνατότητα φιλτραρίσματος των αποτελεσμάτων με χρήση της ρήτρας FILTER.

7.2.2 Βάση Δεδομένων με γράφο (Graph Database)

Όλοι έχουν έρθει κάποτε σε επαφή με τις σχεσιακές βάσεις δεδομένων. Οι σχεσιακές βάσεις δεδομένων είναι γρήγορα και δυνατά εργαλεία για την αποθήκευση μεγάλου όγκου πληροφοριών. Οι βάσεις αυτές εφαρμόζονται ευρέως σε πολλούς τομείς, όπως στο πεδίο των οικονομικών, της ιατρικής, της λογιστικής, ή ακόμα και σε προσωπικό επίπεδο και

χρησιμοποιούν το σχεσιακό μοντέλο ή αλλιώς «σχήμα» για την αναπαράσταση των δεδομένων και την εύρεση της πληροφορίας που μας ενδιαφέρει. Μια σχεσιακή βάση δεδομένων είναι μια συλλογή από σχέσεις στην μορφή πινάκων, που βοηθούν στην οργάνωση της δομής των αποθηκευμένων δεδομένων.

Ένα βασικό πρόβλημα που παρουσιάζουν οι σχεσιακές βάσεις είναι ότι όταν ο όγκος των προς αποθήκευση δεδομένων γίνει πολύ μεγάλος και το σύστημα έχει να διαχειριστεί πολλά διαφορετικά είδη πληροφορίας, τότε τα «σχήματα» μπορεί να γίνουν ιδιαίτερα περίπλοκα. Ένα παράδειγμα τέτοιου σχήματος φαίνεται στην εικόνα 19. Αξίζει να σημειωθεί μάλιστα ότι αυτή η βάση δεδομένων είναι ένα μικρό μόνο κομμάτι των δεδομένων που χρειάζονται για να λειτουργήσει μια επιχείρηση.



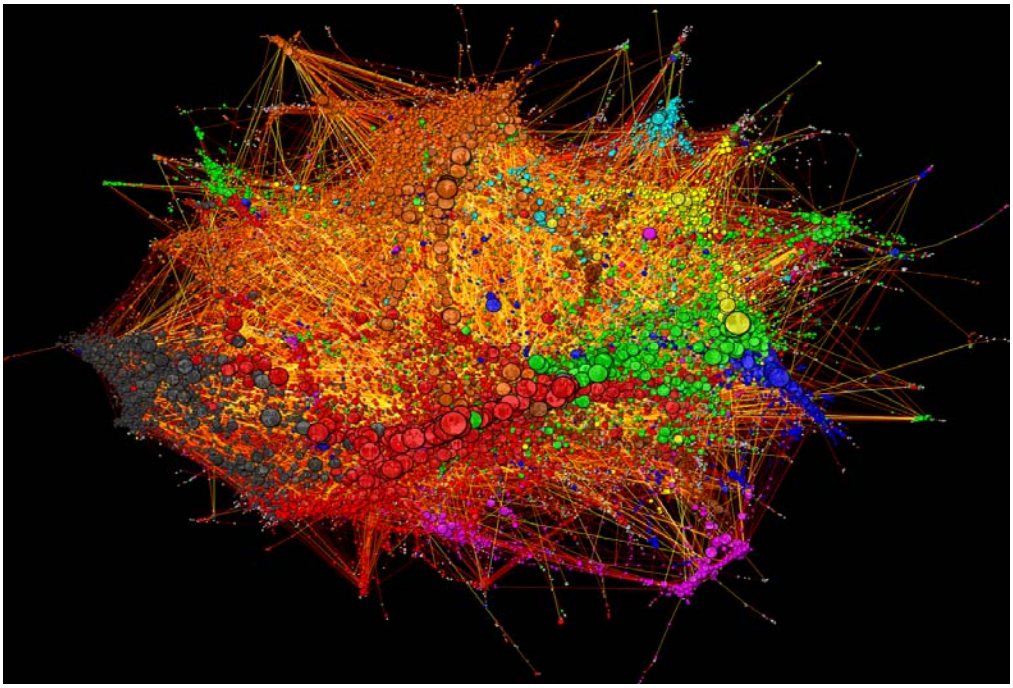
Εικόνα 19 : Παράδειγμα πολύπλοκης βάσης δεδομένων [27]

Ένα δεύτερο σημαντικό πρόβλημα που παρουσιάζουν οι σχεσιακές βάσεις δεδομένων είναι πως δεν μπορούν να ανταπεξέλθουν στις συνεχείς μεταβολές που παρουσιάζει ο Παγκόσμιος Ιστός. Η ενσωμάτωση δεδομένων στον Παγκόσμιο Ιστό χαρακτηρίζεται από γρήγορα μεταβαλλόμενους τύπους δεδομένων, και έτσι οι προγραμματιστές δεν μπορούν να ξέρουν ποτέ ακριβώς τι είναι

διαθέσιμο και πώς θέλουν οι άνθρωποι να το χρησιμοποιήσουν. Το περιεχόμενο αλλάζει συνεχώς καθώς καθημερινά υπάρχουν νέα και οι χρήστες συμπληρώνουν ή εισάγουν νέες πληροφορίες.

Αυτά τα θέματα λοιπόν καθιστούν δύσκολη την αποθήκευση δεδομένων σε σχεσιακή βάση, καθώς για να χρησιμοποιήσει κάποιος τα σχεσιακά δεδομένα ενός άλλου, πρέπει να καταλάβει πώς οι διάφοροι πίνακες σχετίζονται μεταξύ τους. Αυτή η πληροφορία, δηλαδή τα δεδομένα σχετικά με την αναπαράσταση των δεδομένων ονομάζεται μεταδεδομένα και αντιπροσωπεύει την γνώση για το πώς τα δεδομένα μπορούν να χρησιμοποιηθούν. Η Yahoo! έχει αρχίσει να προσθέτει σημασιολογικά μεταδεδομένα στο περιεχόμενό της χρησιμοποιώντας επεκτάσιμα σχήματα. Αυτή η δυνατότητα επέκτασης των ήδη υπαρχόντων μεταδεδομένων καθιστά δυνατή την συνεχή εξέλιξη των στοιχείων και επιτρέπει σε μια εφαρμογή να επωφελείται από πληροφορία που παρέχεται σε μια άλλη εφαρμογή.

Καθώς τα χαρακτηριστικά των εφαρμογών εξελίσσονται, τα σχήματα δεδομένων συχνά πρέπει αναγκαστικά να εξελιχθούν και αυτά, καθώς η ταχεία εξέλιξη αποτελεί πρόκληση για την διαχείριση και αποθήκευση των δεδομένων. Είναι δυνατόν να οριστεί ένα σχήμα που να είναι αρκετά ευέλικτο για να διαχειρίζεται μεγάλο αριθμό συνεχώς μεταβαλλόμενων τύπων δεδομένων και ταυτόχρονα να παραμένει αναγνώσιμο ; Την λύση σε αυτό το πρόβλημα δίνουν οι Graph Databases. Πρόκειται για έναν τύπο βάσης δεδομένων που χρησιμοποιεί δομές γράφων με κόμβους, ακμές και ιδιότητες για να αναπαραστήσει και να αποθηκεύσει πληροφορία. Αυτές οι βάσεις δεδομένων θεωρούν ότι το κάθε αποθηκευμένο αντικείμενο μπορεί να έχει οποινδήποτε αριθμό σχέσεων με άλλα αντικείμενα. Οι σχέσεις αυτές είναι σύνδεσμοι που όλοι μαζί διαμορφώνουν ένα δίκτυο, ή έναν γράφο. Σε σύγκριση με τις σχεσιακές, οι Graph Databases είναι πιο γρήγορες και σχετίζονται πιο άμεσα με την δομή αντικειμενοστραφών εφαρμογών. Προσαρμόζονται πιο εύκολα σε μεγάλα και μεταβαλλόμενα datasets με εξελισσόμενα σχήματα. Είδη βάσεων δεδομένων με γράφο είναι οι βάσεις δεδομένων δικτύου ή τα Triplestores, που αποθηκεύουν και ανακτούν RDF μεταδεδομένα σε γράφους με τριπλέτες. Παραδείγματα βάσεων δεδομένων με γράφο είναι η FlockDB, μια Graph Database ανοιχτής πηγής για την διαχείριση δεδομένων στα πλαίσια του παγκόσμιου ιστού που αρχικά χρησιμοποιήθηκε στο Twitter για να κατασκευαστεί η βάση δεδομένων με τους χρήστες και να διαχειριστούν οι μεταξύ τους σχέσεις, η Franz Inc. AllegroGraph RDFstore, μια υψηλής επίδοσης RDF Graph Database που μπορεί να διαχειρίζεται δισεκατομμύρια τριπλέτες και η Neo4j, μια Graph Database ανοιχτής πηγής σε Java.



Εικόνα 20 : Graph Database του Last.fm με συγκροτήματα, μουσικούς, συνθέτες και σχέσεις μεταξύ αυτών [28]

7.2.3 Συστήματα Προτάσεων σε Semantics

Η σημασιολογική ανάλυση επιτρέπει στα Συστήματα Προτάσεων να μοντελοποιούν ακριβέστερα τα profiles των χρηστών, τα οποία περιλαμβάνουν έννοιες που ορίζονται σε εξωτερικές βάσεις λεξιλογίου. Η προσέγγιση αυτή έχει στόχο να χτίσει ένα σύστημα, το οποίο να έχει την υποδομή και την γλωσσική γνώση ώστε να είναι ικανό να ερμηνεύει σωστά το περιεχόμενο εγγράφων και χαρακτηριστικών γνωρισμάτων που είναι γραμμένα σε φυσική γλώσσα. Πολλές εφαρμογές ενσωματώνουν γλωσσικές πληροφορίες από εγκυκλοπαιδικές πηγές γνώσης ή οντολογίες στην διαδικασία δημιουργίας και εκμάθησης του profile του χρήστη, ώστε να χτιστεί ένα σύστημα προτάσεων με «νοημοσύνη».

Στην επιστήμη των υπολογιστικών συστημάτων, θα μπορούσε να οριστεί η οντολογία ως «*μια προδιαγραφή ενός αντιπροσωπευτικού λεξιλογίου που αποτελείται από ορισμούς κλάσεων, σχέσεων, συναρτήσεων και άλλων αντικειμένων για έναν κοινόχρηστο χώρο λόγου*» [11]. Οι οντολογίες και οι λεξικές βάσεις δεδομένων ορίζουν παρόμοια λεξιλόγια με διαφορετικά επίπεδα επίσημης σημασιολογίας.. Οι οντολογίες παρέχουν ελεγχόμενα λεξιλόγια που μπορούν να χρησιμοποιηθούν για τον σχολιασμό ή την περιγραφή των προϊόντων. Το μοντέλο σημασιολογικής ανάλυσης παρουσιάζει δυνατότητες όπως την πολύπλευρη αναζήτηση περιεχομένου και την περιήγηση σε ιστοσελίδες. Η εγκυκλοπαιδική γνώση είναι πολύ χρήσιμη στο να αναγνωρίζονται έννοιες που αναφέρονται σε πολύ συγκεκριμένους τομείς (π.χ. επιστήμη) και καθορισμένες οντότητες, ειδικά στα πλαίσια περιεχομένων στα οποία δεν είναι δυνατή η έγκριση οντολογιών.

Τα συστήματα ανάκτησης πληροφορίας (Information-retrieval) με βάση την σημασιολογική ανάλυση που εφαρμόζονται μέχρι τώρα χρησιμοποιούν ένα βάσει της λογικής μοντέλο αναζήτησης που αντιμετωπίζει το περιεχόμενο ελεύθερου κειμένου ως μη διφορούμενης έννοιας κομμάτια σημασιολογικής γνώσης. Η δυνατότητα βαθμολόγησης σε αυτού του είδους τα συστήματα προστέθηκε πρόσφατα, καθώς ενσωματώθηκε στο μοντέλο η χρήση χαρακτηριστικών γνωρισμάτων με βάση κείμενο και οντολογίες.

Η αποτελεσματικότητα των τεχνικών επεξεργασίας της φυσικής γλώσσας θα μπορούσε να βελτιωθεί με την παραγωγή ενημερωτικών στοιχείων. Για παράδειγμα η διαδικασία μάθησης του profile ενός χρήστη θα μπορούσε να επωφεληθεί με την εισαγωγή εξωτερικών πηγών λεξικολογικών γνώσεων, σε συνδυασμό με την εσωτερική γνώση που προκύπτει από τα ίδια τα έγγραφα. Τα τελευταία χρόνια έχουν δημιουργηθεί και υπάρχουν διαθέσιμες πολλές πηγές γνώσεων σημασιολογικής μορφής. Αναφορικά, ορισμένα παραδείγματα βάσεων γνώσεων είναι το Open Directory Project ή η Wikipedia.

Πιο συγκεκριμένα, μια από τις πολλές λειτουργίες της Wikipedia είναι και η εκτίμηση της ομοιότητας (similarity) μεταξύ ταινιών. Για να υπολογιστούν αυτές οι ομοιότητες χρησιμοποιούνται το περιεχόμενο και οι σύνδεσμοι που εμφανίζονται στα άρθρα της Wikipedia. Μια άλλη λειτουργία της Wikipedia είναι η χρήση της για το φιλτράρισμα RSS feed και e-mail. Για να γίνει αυτό, το σύστημα περνάει από το στάδιο παραγωγής profile. Όταν φτάνει λοιπόν σε αυτό το βήμα, εκμεταλλεύεται την συλλογή άρθρων που παρέχει ο χρήστης της Wikipedia, η οποία φανερώνει και τους τομείς ενδιαφερόντων του χρήστη. Μια ομάδα όρων εκλέγεται από το κάθε έγγραφο, και στη συνέχεια γίνεται εύρεση ομοίων άρθρων δημοσιευμένων στην Wikipedia με την εφαρμογή του αλγορίθμου Ρητής Σημασιολογικής Ανάλυσης (Explicit Semantic Analysis). Το σύστημα στη συνέχεια παράγει μια λίστα κατηγοριών άρθρων και αθροίζει ανά ομάδες αυτές τις κατηγορίες ώστε να πάρει ένα υποσύνολο των κατηγοριών που να αντιπροσωπεύει ένα θέμα στο profile του χρήστη. Ο χρήστης μπορεί να ελέγχει και να επεμβαίνει δυναμικά στο profile του, προσθέτοντας ή αφαιρώντας κατηγορίες. Για κάθε θέμα στο profile του χρήστη δημιουργείται ένα σώμα εγγράφων στην Wikipedia και αυτό αποτελεί την βάση για το φιλτράρισμα και την ανάκτηση πληροφορίας.

Ολοκληρώνοντας, πολλά συστήματα προτάσεων που βασίζουν την απόδοσή τους σε μια βάση γνώσεων και χρησιμοποιούν τεχνικές του Σημασιολογικού Ιστού, δίνοντας λύσεις σε εφαρμογές πολλών διαφορετικών τομέων. Η χρήση οντολογιών και λεξικών από τα συστήματα περιορίζει πολλά προβλήματα και έχει πλεονεκτήματα όπως [29] :

- Η διασφάλιση της διαλειτουργικότητας του συστήματος και η ομοιογένεια στην παρουσίαση των δεδομένων πληροφοριών
- Η δυνατότητα δυναμικής ένταξης προτιμήσεων του χρήστη σε συγκεκριμένους τομείς
- Η δυνατότητα της σημασιολογικής επέκτασης των περιγραφών του χρήστη
- Η βελτίωση της παρουσίασης και της περιγραφής διαφορετικών στοιχείων του συστήματος
- Η βελτίωση της περιγραφής της λογικής του συστήματος μέσω της ένταξης ενός συνόλου κανόνων
- Η παροχή των απαραίτητων μέσων για την δημιουργία περιγραφών που είναι εμπλουτισμένες με διαδικτυακές υπηρεσίες και η διευκόλυνση της εύρεσης των υπηρεσιών αυτού μέσω κατάλληλου λογισμικού

Έρευνες πια εστιάζουν στην δημιουργία υβριδικών συστημάτων που θα χρησιμοποιούν εργαλεία σημασιολογικών πηγών γνώσης για το φιλτράρισμα της πληροφορίας, σε συνδυασμό με άλλες τεχνικές, όπως τεχνικές στηριζόμενες στα δίκτυα εμπιστοσύνης στα οποία έγινε αναφορά στο κεφάλαιο 5.

Μια από τις πιο συνηθισμένες εφαρμογές των βάσεων δεδομένων με γράφημα είναι στα Συστήματα Προτάσεων. Η χρήση βάσεων δεδομένων με γράφο στα Συστήματα Προτάσεων διευκολύνει την εφαρμογή τους και κάνει πολύ γρηγορότερη την εκτέλεσή τους. Με μια graph database υπάρχουν πολλοί διαφορετικοί τρόποι να κινηθείς σε ένα σύνολο δεδομένων μέσα σε χιλιοστά του δευτερολέπτου. Η παραγωγή σύστασης σε πραγματικό χρόνο αφορά άμεσα την μοντελοποίηση των δεδομένων (αντικειμένων και μεταξύ τους σχέσεων) σε γράφο, την αναπαράστασή τους σε ένα αποδοτικό σύστημα αποθήκευσης γράφων (graph database) και την επεξεργασία αυτού του γράφου για την παραγωγή εξατομικευμένων συστάσεων.

7.2.4 Προσπάθεια επέκτασης εφαρμογής σε Graph Database

Μέσα από το πείραμα παρατηρήθηκε η έντονη παρουσία του κοινωνικού στοιχείου, δηλαδή βαθμολογία και επισήμανση ταινίας με το like δεν εξέφραζαν πάντοτε επιδοκιμασία προς μία ταινία, και σίγουρα όχι με την ίδια ένταση. Η δύναμη που παρουσιάζουν τα κοινωνικά δίκτυα (άρα και οι μαθηματικοί γράφοι) στην επιρροή του κοινωνικού μας δικτύου, και άρα σε ένα σύστημα προτάσεων, φαίνεται να ανατρέπει τους παραδοσιακούς τρόπους προτάσεων και να τους μεταφέρει σε προσωπικό επίπεδο. Όπως έχει παρατηρήσει και ο Χρηστάκης στο βιβλίο του "Συνδεδεμένοι", η δύναμη των κοινωνικών δικτύων και η επιρροή που ασκούν στις αποφάσεις μας αλλά και στην ίδια μας την υπόσταση είναι πολύ μεγαλύτερη από όσο είχαμε φανταστεί μέχρι τώρα.

Σε αυτόν τον άξονα έγιναν πειραματισμοί με λύσεις όπως οι βάσεις δεδομένων στηριγμένες σε γράφους, με τα υπάρχοντα δεδομένα. Για τη δοκιμή εξαγωγής πρόβλεψης με Semantics χρησιμοποιήθηκε η βάση δεδομένων με γράφο AllegroGraph Franz Inc. Η βαση επιλέχθηκε γιατί είναι μία βάση στηριγμένη σε γράφο αλλά επίσης περιέχει σημασιολογία στα δεδομένα που αποθηκεύει, άρα μπορούν να εξεταστούν πολλαπλές δυνατότητες. Στην βάση αυτήν αποθηκεύθηκαν τα δεδομένα που είχαν συλλεχθεί από το API του Facebook και είχαν χρησιμοποιηθεί στην εφαρμογή του k-nn αλγορίθμου. Επίσης, η βάση συνδέθηκε προγραμματιστικά με το site IMDb.com, ώστε για τις ταινίες που αρέσουν στους χρήστες από τους οποίους πάρθηκαν τα δεδομένα να αποθηκεύονται στην βάση και κάποιες πληροφορίες των ταινιών, όπως το είδος στο οποίο ανήκουν, τους ηθοποιούς, τον σκηνοθέτη ή την γενική βαθμολογία της κάθε ταινίας.

Η μεγάλη χρησιμότητα της σύστασης προτάσεων με Semantics σε GraphDatabase είναι ότι μπορεί κανείς να εκμεταλλευτεί το δίκτυο του χρήστη-στόχου λόγω γραφήματος και έτσι να παραχθούν πιο εξατομικευμένες συστάσεις. Για παράδειγμα, η χρήση γραφήματος για την αποθήκευση των δεδομένων επιτρέπει να φανούν οι συνδέσεις του χρήστη και έτσι μπορούν εύκολα να βρεθούν οι χρήστες εκείνοι με τους οποίους ο χρήστης-στόχος μοιράζεται τις περισσότερες συνδέσεις, καθώς αυτοί οι χρήστες θα αποτελούν το στενό δίκτυο κοινών σημείων με τον χρήστη-στόχο και πιθανότατα μια σύσταση προερχόμενη από τον στενό του κύκλο να είναι και πετυχημένη. Στόχος λοιπόν ήταν να γίνει ένα SPARQL Query το οποίο να βρίσκει από το δίκτυο φίλων του χρήστη-στόχου με ποιόν έχει τους περισσότερους κοινούς φίλους, και στην

συνέχεια από τις ταινίες που αρέσουν στον συγκεκριμένο φίλο να βρίσκει εκείνες που δεν έχει δει ο χρήστης-στόχος και να προτείνει από αυτές την ταινία που έχει την υψηλότερη βαθμολογία βάσει των δεδομένων της IMDb.

Ωστόσο, για να βρει το ερώτημα πόσους κοινούς φίλους έχει ο χρήστης-στόχος με τον κάθε φίλο του και στην συνέχεια να καταλήξει με ποιόν έχει τους περισσότερους κοινούς φίλους πρέπει να χρησιμοποιήσει τις λειτουργίες count και max. Το πρόβλημα που παρουσιάστηκε στην δοκιμή των SPARQL Queries στην AllegroGraph Franz Inc. είναι ότι δεν υποστηρίζονται αυτά τα aggregate functions από την βάση, παρά μόνο προγραμματιστικά. Αναγκαστικά λοιπόν διακόπηκε η προσπάθεια καθώς η εξαγωγή πρόβλεψης χωρίς την χρήση των λειτουργιών count και max με ένα απλό ερώτημα της μορφής « βρες μου μια ταινία που δεν έχω δει από τους φίλους μου» δεν εκμεταλλεύεται τις δυνατότητες που προσφέρει η συγκεκριμένη Graph Database. Σίγουρα βρίσκεται σε αρχικό στάδιο η χρήση των Graph Databases.Η δε AllegroGraph δεν αποτέλεσε τη δεδομένη στιγμή την κατάλληλη επιλογή για την επίτευξη των στόχων μας και για αυτόν τον λόγο εγκαταλήφθηκε. Η αδυναμία στο σημασιολογικό κομμάτι με την μικρή λειτουργικότητα της SPARQL αλλά και η δύστροπη και ακαθόριστη χρήση ενός interface για SNA ερωτήματα τελικά οδήγησε στην εγκατάλειψη του εγχειρήματος. Παρατηρήθηκε λοιπόν ότι η AllegroGraph δεν ήταν αρκετά εύχρηστη από το graph interface της.

Σε επόμενο στάδιο λοιπόν θα μπορούσε κανείς να δοκιμάσει άλλες, πιο εύχρηστες πλατφόρμες που να υποστηρίζουν πολλές λειτουργίες και να επιτρέπουν να εξάγεται χρήσιμη πληροφορία από το δίκτυο του γραφήματος στο οποίο αποθηκεύονται τα δεδομένα. Η χρήση λοιπόν Semantics σε Graph Database ως μελλοντική επέκταση των Συστημάτων Προτάσεων μπορεί να οδηγήσει σε γρήγορη, εύκολη και πλήρως εξατομικευμένη παραγωγή συστάσεων.

7.2.5 Πιθανές επεκτάσεις και βελτιώσεις

Όσον αφορά την προσπάθεια εφαρμογής συστήματος προτάσεων σε Graph Database, καθότι έμεινε σε αρχικό στάδιο, θα μπορούσαν σαν μελλοντική επέκταση να γίνουν δοκιμές σε άλλες πλατφόρμες που αξιοποιούν την σημασιολογία και υποστηρίζουν λειτουργίες για Social Network Analysis.

Σχετικά με το πείραμα του knn αλγορίθμου που διεξήχθηκε, παρατηρίθηκε ότι τα δεδομένα του Facebook δεν ήταν αρκετά εύχρηστα στην αποτύπωση των προτιμήσεων των χρηστών. Μια εναλλακτική λοιπόν για κάποιον που ασχολείται με το θέμα θα ήταν να εφαρμόσει τον συγκεκριμένο αλγόριθμο με δεδομένα από άλλη πλατφόρμα (όπως π.χ. από την IMDb). Επίσης, χρήσιμο θα ήταν το dataset να περιλαμβάνει περισσότερους χρήστες και βαθμολογίες, ώστε να είναι πιο εύκολο να βρεθούν οι «γείτονες» και τελικά να προκύπτουν πιο αξιόπιστα αποτελέσματα. Μια άλλη επέκταση του αλγορίθμου αφορά την μεταβλητή κ. Στα δεδομένα της παρούσας εφαρμογής, στα οποία οι χρήστες δεν ήταν πολλοί και οι βαθμοί έπαιρναν τις τιμές 0/1 οπότε δεν υπήρχαν πολλές διακυμάνσεις ή διαβαθμήσεις, μια μικρή τιμή του κ είναι αρκετή για να είναι αποδοτικός ο αλγόριθμος. Ωστόσο, αξίζει να μελετηθεί πώς μεταβάλλεται το κ ανάλογα με τον όγκο των δεδομένων αλλά και την φύση της ομάδας στην οποία γίνεται η μελέτη.

Τέλος, ένα άλλο σημείο που θα οδηγούσε σε βελτίωση το τρέξιμο του αλγορίθμου θα ήταν η αλλαγή της παραμετροποίησης της μεθόδου addInstance που εφαρμόζεται, ώστε το πρόγραμμα να δέχεται ως είσοδο δεδομένα που μπορούν να πάρουν και τιμή Null.

8

Βιβλιογραφία

[1]. Information Overload, Wikipedia (27/4/2011) : http://en.wikipedia.org/wiki/Information_overload

[2]. Συνέδριο ACM Recommender Systems 2007 (27/4/2011) : http://recsys.acm.org/2007/

[3]. Recommender Systems Handbook , Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor editors, 2011 edition

[4]. Robin Burke, Hybrid Web Recommender Systems, 2010: http://www.dcs.warwick.ac.uk/~acristea/courses/CS411/2010/Book%20-%20The%20Adaptive%20Web/HybridWebRecommenderSystems.pdf

[5]. Blog CyberTheater.com (27/4/2011) : http://www.cybertheater.com/netflix-recommendation-system-was-beat-for-1-million/

[6]. Flickr, (27/4/2011) : http://www.flickr.com/photos/torley/4551424756/in/photostream/

[7]. Andreas M. Kaplan, Michael Haenlein, Users of the world, unite! The challenges and opportunities of Social Media, Business Horizons (2010) Volume 53: http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6W45-4XFF2S0-1-3&_cdi=6533&_user=83473&_pii=S0007681309001232&_origin=na&_coverDate=02%2F28%2F2010&_sk=999469998&view=c&wchp=dGLzVzz-zSkWb&md5=0f9e865580d383be9ce243018ecc9ed2&ie=/sdarticle.pdf

[8]. Social Media, Wikipedia (29/4/2011) : http://en.wikipedia.org/wiki/Social_media

[9]. Μέσα Κοινωνικής Δικτύωσης MyOta.gr (29/4/2011) : http://mytown.gr/myota/index.php?option=com_content&view=article&id=80&Itemid=105

[10]. Panagiotis Symeonidis, Alexandros Nanopoulos**,** Yannis Manolopoulos, MoviExplain: A Recommender System with Explanations, 2009 : http://delab.csd.auth.gr/papers/recsys2.pdf

[11]. Tuukka Ruotsalo, Methods and Applications for Ontology-based Recommender Systems, Doctoral Dissertation, 2010 : http://lib.tkk.fi/Diss/2010/isbn9789526031514/isbn9789526031514.pdf

[12]. Michael J. Pazzani and Daniel Billsus, Content-based Recommendation Systems: http://www.fxpal.com/publications/FXPAL-PR-06-383.pdf
[13]. Aravind Chandramouli and Alessandro Micarelli, User Profiles for Personalized Information Access by Susan Gauch, Mirco Speretta, 2007: http://www.springerlink.com/content/y4g84202705577p3/fulltext.pdf
[14] Jorg Diederich, Tereza Iofciu, Finding Communities of Practice from User Profiles based on Folksonomies: http://www.l3s.de/~diederich/Papers/TBProfile-telcops.pdf
[15]. Netflix Prize (5/5/2011) : http://www.netflixprize.com/
[16]. Digital Magazine InfoVis.net (5/5/2011) : http://www.infovis.net/printMag.php?num=155&lang=2
[17]. Manos Papagelisa and Dimitris Plexousakis, Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents, 2005: http://www.ics.forth.gr/isl/publications/paperlink/science.pdf
[18]. J. Ben Schafer, Dan Frankowski, Jon Herlocker and Shilad Sen, Collaborative Filtering Recommender Systems, 2007: http://www.google.gr/url?sa=t&source=web&cd=1&ved=0CBkQFjAA&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.130.4520%26rep%3Drep1%26type%3Dpdf&ei=wj7_Te6OLsOSOoX_4N4I&usg=AFQjCNFi9rSR5dxCQQNvsMhD2xAmcyvwPA
[19]. Rong Jin and Luo Si, A Study of Methods for Normalizing User Ratings in Collaborative Filtering, 2004: http://www.google.gr/url?sa=t&source=web&cd=1&ved=0CBcQFjAA&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.151.3084%26rep%3Drep1%26type%3Dpdf&ei=30D_Td_-LcaBOr266N4I&usg=AFQjCNGTRHs1VEqL5H3BhVY1f6MQyT5YSg
[20]. Collective Intelligence in Action by Satnam Alag, 2008 edition
[21]. Taek-Hun Kim and Sung-Bong Yang, An Effective Threshold-Based Neighbor Selection in Collaborative Filtering, 2007: http://algo.yonsei.ac.kr/international_JNL/An%20Effective%20Threshold-based%20Neighbor%20Selection.pdf
[22]. Badrul M. Sarwar, George Karypis, Joseph A. Konstan, John T. Riedl Application of Dimensionality Reduction in Recommender System - A Case Study : http://www.cs.umn.edu/tech_reports_upload/tr2000/00-043.pdf
[23]. Zan Huang, Wingyan Chung, Thian-Huat Ong, Hsinchun Chen, A Graph-based Recommender System for Digital Library, 2002 : http://delivery.acm.org/10.1145/550000/544231/p65-huang.pdf?ip=147.102.131.12&CFID=29643324&CFTOKEN=74240286&__acm__=1308576837_765e471cdbf91321945fbdcc4fcbb56d

[24]. Manos Papagelis, Dimitris Plexousakis, Themistoklis Kutsuras, Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inferences : http://queens.db.toronto.edu/~papaggel/docs/papers/all/iTrust05-Alleviating_the_Sparsity_Problem_of_Collaborative_Filtering_Using_Trust_Inferences.pdf

[25]. Patricia Victor, Trust Networks for Recommender Systems, 2010 : http://lib.ugent.be/fulltxt/RUG01/001/401/756/RUG01-001401756_2010_0001_AC.pdf

[26]. Neal Lathia, Stephen Hailes, Licia Capra, kNN CF: A Temporal Social Network, 2008 : http://www.cs.ucl.ac.uk/staff/l.capra/publications/fpp01a-lathia.pdf

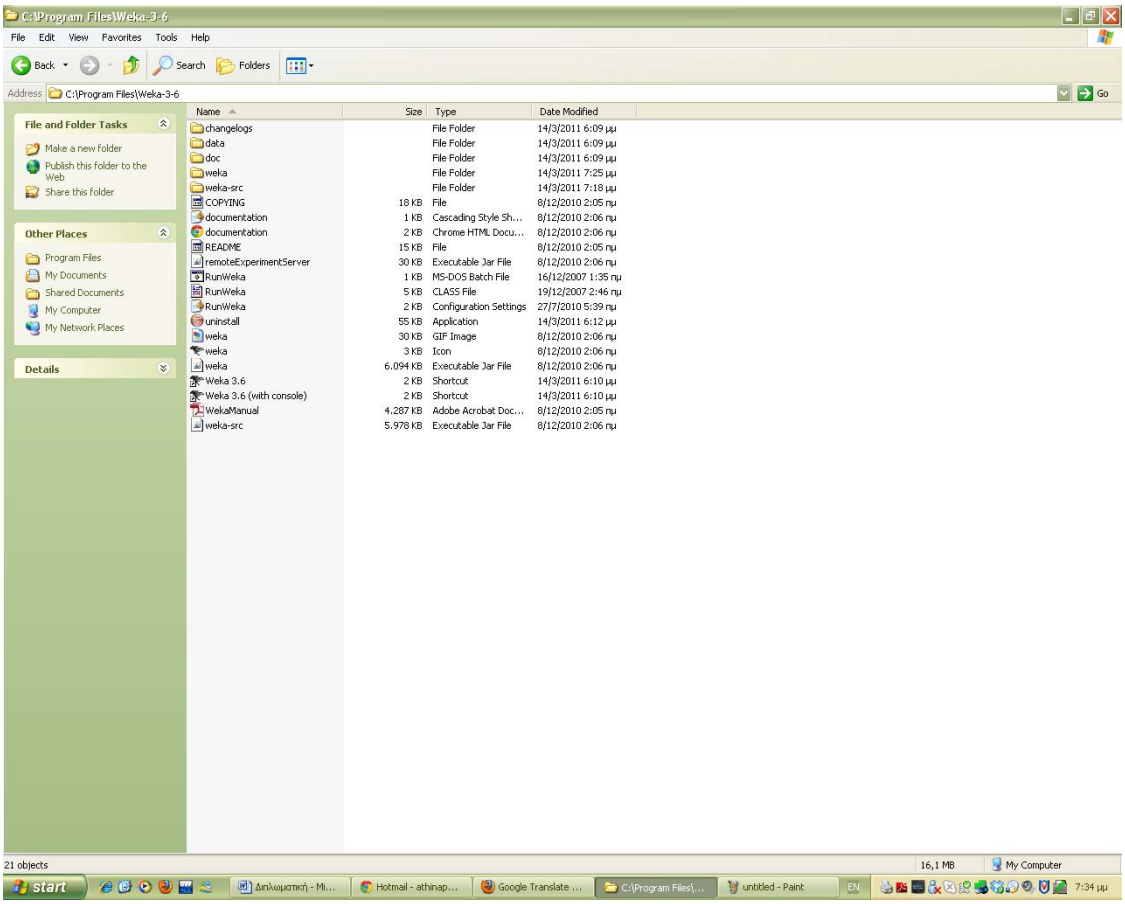
[27]. Programming the Semantic Web by Toby Segaran, Colin Evans, Jamie Taylor (First edition)

[28]. Visualizing Music (3/6/2011) : http://visualizingmusic.com/page/3/

[29]. E. Peis, J. M. Morales-del-Castillo, J. A. Delgado-López, Semantic Recommender Systems, Analysis of the state of the topic, Issue 6, 2008 : http://www.hipertext.net/english/pag1031.htm

Παράρτημα Α

Για την εγκατάσταση του WEKA αρχικά πρέπει να «κατέβει» το πακέτο WEKA από την ιστοσελίδα http://www.cs.waikato.ac.nz/ml/weka/. Επιλέχθηκε η έκδοση weka 3.6 χρησιμοποιώντας τις προεπιλεγμένες τιμές. Υπήρχε ήδη εγκατεστημένο το πακέτο Java, για αυτό και έγινε download μόνο για το εκτελέσιμο weka-3-6.exe , αλλιώς θα έπρεπε να γίνει download το εκτελέσιμο που περιλαμβάνει Java VM 5.0 (weka-3-6jre.exe). Εγκαταστήθηκε το λογισμικό στη διεύθυνση C:\Program Files\Weka-3-6. Ανοίγοντας τον φάκελο στον υπολογιστή εμφανίζεται μια εικόνα της μορφής :



Εικόνα 21 : Αρχεία του πακέτου Weka-3-6 στο directory C:\Program Files\Weka-3-6

Στο Eclipse project προστέθηκε στο Java Build Path το αρχείο weka.jar του παραπάνω φακέλου και ως source attachment δηλώθηκε το weka-src.jar, ώστε να μπορεί το τρέχων πρόγραμμα να χρησιμοποιήσει τις κλάσεις του WEKA. Επιλέγοντας την σελίδα documentation.html από τον παραπάνω φάκελο εμφανίζεται μια σελίδα όπως η επόμενη :

Εικόνα 22 : Documentation του WEKA

Στη σελίδα αυτή, η επιλογή του Package Documentation είναι ένα JavaDoc το οποίο εμφανίζει όλα τα πακέτα και τις κλάσεις που περιέχει το WEKA, ορισμένες από τις οποίες έχουν αναφερθεί στην ενότητα 6.3.
Αν τώρα θελήσει κάποιος να μελετήσει τα περιβάλλοντα για τα οποία το πακέτο WEKA έχει GUI εφαρμογή, τότε διαλέγει από το πρόγραμμα Weka 3.6.4 στο μενού έναρξης του υπολογιστή του την επιλογή Weka 3.6 (with console). Εμφανίζεται ένα παράθυρο όπως αυτό που ακολουθεί από το οποίο μπορούν να ερευνηθούν οι Weka GUI εφαρμογές :

Εικόνα 23 : Weka GUI με επιλογές για έναρξη 4 εφαρμογών

Παράρτημα Β

Παρατίθεται το πρόγραμμα που χρησιμοποιήθηκε στην εφαρμογή του k-nn αλγόριθμο. Στην αρχή το πρόγραμμα δημιουργεί τα Instances για να αναπαραστήσει τα δεδομένα, ενώ στη συνέχεια δημιουργεί τον ταξινομητή και παράγει τις εκτιμήσεις.

```
package athina;

import weka.classifiers.lazy.IBk;
import weka.core.Attribute;
import weka.core.FastVector;
import weka.core.Instance;
import weka.core.Instances;

public class KNNWEKAExample {

    public static final void main(String [] args) throws Exception {
        KNNWEKAExample eg = new KNNWEKAExample();
        FastVector attributes = eg.createAttributes();
        Instances instances = eg.createLearningDataSet(attributes);
        eg.illustrateClassification(instances);
    }

    private FastVector createAttributes() {
        FastVector allAttributes = new FastVector(21);
        allAttributes.addElement(new Attribute("item1"));
        allAttributes.addElement(new Attribute("item2"));
        allAttributes.addElement(new Attribute("item3"));
        allAttributes.addElement(new Attribute("item4"));
        allAttributes.addElement(new Attribute("item5"));
        allAttributes.addElement(new Attribute("item6"));
        allAttributes.addElement(new Attribute("item7"));
        allAttributes.addElement(new Attribute("item8"));
        allAttributes.addElement(new Attribute("item9"));
        allAttributes.addElement(new Attribute("item10"));
        allAttributes.addElement(new Attribute("item11"));
        allAttributes.addElement(new Attribute("item12"));
        allAttributes.addElement(new Attribute("item13"));
        allAttributes.addElement(new Attribute("item14"));
        allAttributes.addElement(new Attribute("item15"));
        allAttributes.addElement(new Attribute("item16"));
        allAttributes.addElement(new Attribute("item17"));
```

```
        allAttributes.addElement(new Attribute("item18"));
        allAttributes.addElement(new Attribute("item19"));
        allAttributes.addElement(new Attribute("item20"));
        allAttributes.addElement(new Attribute("item21"));

        return allAttributes;
    }

    private Instances createLearningDataSet(FastVector allAttributes) {
        Instances trainingDataSet =
            new Instances("wekaCF", allAttributes, 21);
        trainingDataSet.setClassIndex(0);
        addInstance(trainingDataSet,
0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0);
        addInstance(trainingDataSet,
1,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0);
        addInstance(trainingDataSet,
1,1,0,0,1,0,0,0,1,1,0,0,0,0,1,1,0,0,1,0,0);
        addInstance(trainingDataSet,
1,1,1,1,1,0,1,1,0,0,1,1,1,1,1,1,1,0,1,1,1);
        addInstance(trainingDataSet,
1,0,0,0,1,1,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0);
        addInstance(trainingDataSet,
1,0,1,0,1,1,0,0,0,0,0,0,1,0,0,1,0,1,0,1,1);
        addInstance(trainingDataSet,
0,0,0,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0);
        addInstance(trainingDataSet,
1,0,0,1,1,0,0,0,1,0,0,1,1,0,0,0,0,0,1,0,0);
        addInstance(trainingDataSet,
0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,1,0,1,0,1,1);
        addInstance(trainingDataSet,
0,0,0,0,0,1,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0);
        addInstance(trainingDataSet,
1,0,1,0,1,1,1,1,0,1,1,0,1,1,1,0,1,1,1,1,0);
        addInstance(trainingDataSet,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0);
        addInstance(trainingDataSet,
1,0,1,0,1,1,1,1,1,0,0,1,0,1,1,1,1,1,1,0,0,1,0);
        System.out.println(trainingDataSet);
        return trainingDataSet;
    }

    private void addInstance(Instances trainingDataSet,
            double item1, double item2, double item3, double item4,
double item5, double item6, double item7, double item8, double item9,
double item10, double item11, double item12, double item13, double
```

```
item14, double item15, double item16, double item17, double item18,
double item19, double item20, double item21)  {
        // Create empty instance with 21 attribute values
        Instance instance = new Instance(21);
        instance.setDataset(trainingDataSet);
        instance.setValue(0, item1);
        instance.setValue(1, item2);
        instance.setValue(2, item3);
        instance.setValue(3, item4);
        instance.setValue(4, item5);
        instance.setValue(5, item6);
        instance.setValue(6, item7);
        instance.setValue(7, item8);
        instance.setValue(8, item9);
        instance.setValue(9, item10);
        instance.setValue(10, item11);
        instance.setValue(11, item12);
        instance.setValue(12, item13);
        instance.setValue(13, item14);
        instance.setValue(14, item15);
        instance.setValue(15, item16);
        instance.setValue(16, item17);
        instance.setValue(17, item18);
        instance.setValue(18, item19);
        instance.setValue(19, item20);
        instance.setValue(20, item21);


        trainingDataSet.add(instance);
    }

    public void illustrateClassification(Instances instances) throws
Exception {
        IBk ibk = new IBk(1);
        ibk.buildClassifier(instances);
        System.out.println("\nPrediction:");
        for (int i = 0; i < instances.numInstances(); i++) {
            Instance instance = instances.instance(i);
            double  result = ibk.classifyInstance(instance);
            System.out.println("Expected=" + instance.value(0) + "
Predicted=" + result);
        }
    }
}
```

Στην εφαρμογή αυτή πρώτα δημιουργούνται τα γνωρίσματα (attributes) και στην συνέχεια τα στιγμιότυπα (instances) από το learning dataset. Αυτά γίνονται με τις εντολές του προγράμματος:

```
FastVector attributes = eg.createAttributes();
Instances instances = eg.createLearningDataSet(attributes);
```

Ως χαρακτηριστικό γνώρισμα προς πρόβλεψη ορίζεται το πρώτο γνώρισμα, το item 1 :

```
trainingDataSet.setClassIndex(0);
```

Έπειτα, το πρόγραμμα χρησιμοποιεί τον k-nn αλγόριθμο για να παραγάγει τις προβλέψεις. Δημιουργεί έναν classifier και στη συνέχεια παράγει τις προβλέψεις για τις τιμές του item 1.

Αρχικά, δημιουργείται ένα στιγμιότυπο του k-nearest neighbor classifier, με κ = 1 :

```
IBk ibk = new IBk(1);
```

Μετά χτίζεται ο ταξινομητής :

```
ibk.buildClassifier(instances);
```

και έπειτα γίνεται εκτίμηση του κάθε instance για τις προβλεπόμενες τιμές του item 1.

Το πρόγραμμα εκτελέστηκε και για κ = 4, 7 με τις αντίστοιχες εντολές :

```
IBk ibk = new IBk(4);
```

```
IBk ibk = new IBk(7);
```

Επίσης, εκτελέστηκε ορίζοντας ως γνώρισμα προς πρόβλεψη όλες τις ταινίες μια-μια αλλάζοντας το 0 της παρακάτω εντολής με τις τιμές 1 έως 20 :

```
trainingDataSet.setClassIndex(0);
```