ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εξατομικευμένο Σύστημα Συστάσεων σε Πλατφόρμα Διαδραστικών Βίντεο με Εμπλουτισμούς

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Μαργαρίτας Α. Πανταζή

Επιβλέπων: Συμεών Χρ. Παπαβασιλείου, Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εξατομικευμένο Σύστημα Συστάσεων σε Πλατφόρμα Διαδραστικών Βίντεο με Εμπλουτισμούς

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Μαργαρίτας Α. Πανταζή

Επιβλέπων: Συμεών Χρ. Παπαβασιλείου, Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή τον Ιανουάριο 2016

……..	……..	……..
Σ. Παπαβασιλείου	Μ. Θεολόγου	Ι. Ρουσσάκη
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Επίκ. Καθηγήτρια Ε.Μ.Π

Αθήνα, Ιανουάριος 2016

.....

Μαργαρίτα Πανταζή

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε το ακαδημαϊκό έτος 2015 στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Θα ήθελα να ευχαριστήσω τον καθηγητή κο Συμεών Παπαβασιλείου για την εμπιστοσύνη που μου έδειξε με την ανάθεση αυτής της εργασίας, δίνοντάς μου παράλληλα τη δυνατότητα να έρθω σε επαφή με έναν χώρο που προάγει την επιστήμη και την έρευνα σε κλίμα συνεργασίας. Παράλληλα, θα ήθελα να ευχαριστήσω πολύ τις Στέλλα Καφετζόγλου και Έλενα Στάη για την αμέριστη βοήθεια τους κατά τη διάρκεια της διπλωματικής. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για τη συμπαράστασή τους κάθε στιγμή, όλα αυτά τα χρόνια.

Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός ολοκληρωμένου συστήματος συστάσεων σε περιβάλλον διαμοιρασμού βίντεο με εμπλουτισμούς. Στο πρώτο κεφάλαιο δίνονται οι απαραίτητοι ορισμοί για τα συστήματα σύστασης καθώς και οι λειτουργίες που εξυπηρετούν. Έπειτα, αναλύονται σε θεωρητικό επίπεδο τεχνικές που χρησιμοποιούν τα συστήματα συστάσεων και προέρχονται από γνωστές τεχνικές εξόρυξης δεδομένων. Οι τεχνικές αυτές χωρίζονται σε τεχνικές προ-επεξεργασίας και ανάλυσης δεδομένων. Στη συνέχεια, περιγράφονται οι σημαντικότερες κατηγορίες των συστημάτων συστάσεων μαζί με κάποια παραδείγματά τους. Στο κεφάλαιο 4 περιγράφονται τα είδη ανατροφοδότησης και κάποιοι τρόποι αξιοποίησης τους βασισμένες στη θεωρία των γράφων, που είναι και η θεωρία στην οποία βασίζεται το προτεινόμενο σύστημα συστάσεων. Εν συνεχεία, περιγράφεται η αρχιτεκτονική του προτεινόμενου συστήματος συστάσεων και ο αλγόριθμος που εκτελεί για να παράγει αποτελέσματα. Το σύστημα απαρτίζεται από μία διεπαφή προγραμματισμού εφαρμογών (API), μια βάση δεδομένων με πληροφορίες για τους χρήστες και τα βίντεο και μια μηχανή συστάσεων. Ο προτεινόμενος αλγόριθμος σύστασης χρησιμοποιεί τεχνικές συνεργατικού φιλτραρίσματος μαζί με τεχνικές ομοιότητας στα πλαίσια ενός μηχανισμού υβριδικού φιλτραρίσματος ενώ ταυτόχρονα υιοθετεί έννοιες από τη θεωρία γράφων. Τα δεδομένα για τους χρήστες, οι ενέργειες τους στη διάρκεια της αλληλεπίδρασης με το σύστημα μας θα αντλούνται από το API που δημιουργήθηκε και αναλύεται στο κεφάλαιο 5. Η εργασία ολοκληρώνεται με την εξαγωγή γενικών συμπερασμάτων και προτάσεων για μελλοντικές προεκτάσεις.

Λέξεις κλειδιά: Συστήματα συστάσεων, συνεργατικό φιλτράρισμα, υβριδικό φιλτράρισμα, βίντεο, εμπλουτισμοί, API

Abstract

The aim of this thesis is to create an integrated system of recommendations in an environment of sharing videos with enrichments. Initially, several theoretical techniques that recommender systems use are analyzed deriving from known data mining techniques. These techniques are divided into pre-processing and data analysis techniques. Then, the main categories of recommender systems are described along with some examples of each one. Chapter 4 describes the types of user feedback and some methods based on the theory of graphs. Chapter 5 describes the architecture of the proposed recommendation system and algorithm used to build recommendations. The system consists of one application programming interface (API), a database of information about users and videos and a recommendation engine. The proposed recommendation algorithm uses collaborative filtering techniques along with similarity techniques in the framework of a hybrid filtering mechanism while adopting concepts from graph theory. Users' data are drawn from the API created, which is analyzed. The thesis concludes with general remarks and proposals for future extensions.

Keywords: Recommender systems, collaborative filtering, hybrid filtering, video, enrichments, API

Περιεχόμενα

1. Εισαγωγή 1

1.1 Γενικά 1

1.2 Ορισμός συστημάτων σύστασης 2

1.3 Λειτουργία συστημάτων σύστασης 3

2. Εξόρυξη Δεδομένων στον Πυρήνα των Συστημάτων Σύστασης (Data Mining in Recommender Systems) 5

2.1 Προ-επεξεργασία Δεδομένων 5

2.1.1 Δειγματοληψία (Sampling) 6

2.1.2 Μείωση Διαστάσεων (Dimensionality Reduction) 6

2.1.3 Απαλοιφή Θορύβου (Denoising) 7

2.2 Ανάλυση Δεδομένων 7

2.2.1 Ανάλυση με κατηγοριοποίηση των δεδομένων (Data Analysis and Classification) 8

2.2.1.1 Μέτρα Ομοιότητας (Similarity Measures) 8

2.2.1.2 Πλησιέστερος Γείτονας (Nearest Neighbor) 9

2.2.1.3 Χρήση Δέντρων Αποφάσεων (Decision Trees) 11

2.2.1.4 Χρήση Bayesian Ταξινομητών (Bayesian Classifiers) 13

2.2.1.5 Χρήση Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks) 14

2.2.1.6 Χρήση Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines) 16

2.2.2 Ανάλυση με χρήση Συσταδοποίησης (Data Analysis and Clustering) 17

2.2.2.1 Αλγόριθμος ***k*** – Means 18

2.2.2.2 Ομαδοποίηση που βασίζεται στην πυκνότητα (Density –based clustering) 20

2.2.2.3 Ομαδοποίηση που βασίζεται στη σύνδεση (Ιεραρχική ομαδοποίηση) 21

3. Αλγόριθμοι συστημάτων σύστασης (Filtering Methods) 22

3.1 Συνεργατικό φιλτράρισμα (Collaborative Filtering) 23

3.2 Φιλτράρισμα με βάση το περιεχόμενο (Content–based Filtering) 25

3.3 Δημογραφικό σύστημα σύστασης (Demographic Filtering) 28

3.4 Σύστημα σύστασης με βάση τη γνώση (Knowlegde–based Filtering) 28
3.5 Υβριδικό σύστημα σύστασης (Hybrid Filtering) 30
4. Ανατροφοδότηση Σχετικότητας και Σύσταση (Relevance Feedback in Recommender Systems) 32
4.1 Έμμεση ανατροφοδότηση (Implicit Feedback) 33
4.1.1 Μοντέλο Jeffrey's Conditioning 34
4.1.1.1 Στάθμιση μονοπατιού (Path Weighting) 34
4.1.1.2 Ποιότητα και κατατοπιστική αξία των στοιχείων (Quality of Evidence) 35
4.1.1.3 Στάθμιση όρου (Term Weighting) 36
4.1.2 Μοντέλα γράφων για ανάλυση έμμεσης ανατροφοδότησης 37
4.1.2.1 Αλγόριθμοι μοντέλων γράφων 38
4.2 Ρητή ανατροφοδότηση (Explicit Feedback) 40
4.3 Ψευδο – ανατροφοδότηση (Pseudo Relevance Feedback) 42
5. Μηχανισμός του Συστήματος Σύστασης 44
5.1 Κοινωνικό και εξατομικευμένο Εργαλείο (Social and Personalization Tool – SP Tool) 44
5.2 Αρχιτεκτονική συστήματος 46
5.2.1 Βάση δεδομένων 46
5.2.2 Το API 47
5.2.3 Μηχανισμός συστάσεων 50
5.2.3.1 Μηχανισμός εξατομίκευσης 50
5.2.3.2 Δημιουργία ευρετηρίου του περιεχομένου πολυμέσων 53
5.2.3.3 Μηχανισμός ανατροφοδότησης σχετικότητας 54
5.2.3.4 Ενημέρωση προφίλ χρηστών 56
5.2.3.5 Φιλτράρισμα με βάση το περιεχόμενο 58
5.2.3.6 Συνεργατικό και υβριδικό φιλτράρισμα 58
6. Συμπεράσματα – Μελλοντικές επεκτάσεις 61
6.1 Συμπεράσματα 61

6.2 Μελλοντικές επεκτάσεις..... 62
Βιβλιογραφία..... 64

Κατάλογος Εικόνων

Εικόνα 2.1: Παράδειγμα *k*-NN ταξινομητή [13] 10
Εικόνα 2.2: Ένα τεχνητό νευρωνικό δίκτυο 14
Εικόνα 2.3: Διαχωριστικό υπερεπίπεδο[31] 16

Κατάλογος Πινάκων

Πίνακας 4.1: Τιμές μεταβλητών αλγόριθμου Rocchio 41
Πίνακας 5.1: Βασικότερα πεδία πίνακα ‘videos’ 47
Πίνακας 5.2: Σήματα έμμεσης ανατροφοδότησης σχετικότητας 55
Πίνακας 5.3: Παραδείγματα τιμών/υπολογισμού των τιμών βάρους της τρίτης στήλης του Πίνακα 5.2..... 55

1. Εισαγωγή

1.1 Γενικά

Από την εμφάνιση του, στα τέλη του 20ου αιώνα, το Διαδίκτυο έχει γίνει ένα αναπόσπαστο κομμάτι της καθημερινή ζωής. Στην εποχή του Ιστού 2.0 (Web 2.0), ο αριθμός των χρηστών και των πληροφοριών που διαμοιράζονται έχει αυξηθεί δραματικά, ιδίως σε ιστοσελίδες όπως το Youtube[1], για διαμοιρασμό και αναζήτηση video, στο Google[2] για αναζήτηση όλων των ειδών πληροφορίας αλλά και σε ιστοσελίδες ηλεκτρονικού εμπορίου για αγορά και πώληση προϊόντων. Η συνεχής αύξηση των διαθέσιμων πληροφοριών άρχισε να καθιστά πιο δύσκολη τη εύρεση σχετικών και αξιόπιστων αποτελεσμάτων αναζήτησης μειώνοντας την ικανοποίηση των χρηστών. Εξαιτίας του προβλήματος αυτού, το ερευνητικό ενδιαφέρον στράφηκε στη δημιουργία αλγορίθμων για σύσταση σχετικών αντικειμένων αναζήτησης για το χρήστη.

Δεδομένου ενός αντικειμένου αναζήτησης από ένα χρήστη π.χ. ταινία, οι αλγόριθμοι σύστασης φιλτράρουν τα αποτελέσματα σύμφωνα με το σχεδιασμό, το γραφικό περιβάλλον και την βασική τεχνική που χρησιμοποιούν και παρουσιάζουν στο χρήστη προτεινόμενα αντικείμενα ιδίου τύπου, π.χ. παρόμοιες ταινίες, για να επιλέξει. Τα συστήματα σύστασης (recommender systems), τα οποία χρησιμοποιούνται ευρέως στο ηλεκτρονικό επιχειρείν αλλά και για ταινίες, μουσική, ειδήσεις, βιβλία, αναζητήσεις κ.α. είναι συνήθως εξατομικευμένα. Διαφορετικοί χρήστες ή ομάδες χρηστών, επομένως, λαμβάνουν διαφορετικές συστάσεις, ανάλογα με τις προτιμήσεις τους.

Στην απλή τους μορφή τα συστήματα σύστασης είναι βαθμολογημένες λίστες από αντικείμενα, σύμφωνα με τις οποίες προβλέπεται το προϊόν που θα παρουσιαστεί στο χρήστη και βασίζονται στις προτιμήσεις και μη του χρήστη. Προκειμένου να υπολογιστεί η βαθμολογία αυτή τα συστήματα σύστασης συλλέγουν από τους χρήστες πληροφορίες για τα ενδιαφέροντα τους είτε άμεσα, π.χ. ο χρήστης βαθμολογεί ένα προϊόν, είτε έμμεσα από τις ενέργειες του χρήστη, π.χ. πλοήγηση στη σελίδα ενός προϊόντος.

Η δημιουργία των συστημάτων συστάσεων βασίστηκε στην απλή παρατήρηση ότι οι άνθρωποι συχνά βασίζονται σε συστάσεις τρίτων για να πάρουν διάφορες καθημερινές αποφάσεις, όπως τι προϊόντα να αγοράσουν, τι μουσική να ακούσουν ή τι ειδήσεις να διαβάσουν. Τα πρώτα συστήματα συστάσεων, μιμούμενα τη συμπεριφορά αυτή, εφάρμοσαν τεχνικές δημιουργίας προτάσεων για αντικείμενα που παρόμοιοι χρήστες (με παρόμοια ενδιαφέροντα) προτίμησαν. Αυτή η προσέγγιση καλείται συνεργατικό φιλτράρισμα και η λογική είναι ότι αν ένας χρήστης έχει συμφωνήσει στο παρελθόν με κάποιους χρήστες, τότε οι προτάσεις των χρηστών αυτών θα είναι σχετικές με το χρήστη αυτό.

[1] http://www.youtube.com
[2] http://www.google.com

Τα συστήματα συστάσεων έχει αποδειχτεί τα τελευταία χρόνια πως αποτελούν πολύτιμο μέσο αντιμετώπισης της υπερφόρτωσης πληροφοριών, ένα πρόβλημα των τελευταίων ετών εξαιτίας της δυνατότητας να δημιουργούνται απέραντα ποσά πληροφοριών ενώ παράλληλα οι άνθρωποι τείνουν να μην επεξεργάζονται σωστά τον όγκο αυτό τον πληροφοριών. Κατόπιν, λοιπόν, αιτήματος του χρήστη, δημιουργούν πληθώρα προτάσεων με βάση πληροφορίες για τους χρήστες, τα διαθέσιμα αντικείμενα και τις προηγούμενες αλληλεπιδράσεις καταχωρημένα σε βάσεις δεδομένων. Ο χρήστης περιηγείται στα προτεινόμενα αντικείμενα δίνοντας άμεση ή έμμεση ανατροφοδότηση. Η ανατροφοδότηση αυτή και οι ενέργειες δημιουργούν νέες προτάσεις στις επόμενες αλληλεπιδράσεις του χρήστη με το σύστημα.

1.2 Ορισμός συστημάτων σύστασης

Σύστημα σύστασης όπως αναφέρθηκε είναι ένα σύστημα που αναλύει τις προτιμήσεις ενός συγκεκριμένου χρήστη και αναγνωρίζει ένα σύνολο στοιχείων που του προτείνει ως ενδιαφέροντα. Από τον ορισμό του συστήματος γίνεται κατανοητό πως είναι πολύ σημαντικό να μπορούν να αναγνωρισθούν, να αναλυθούν και να αξιοποιηθούν αποδοτικά τα ιδιαίτερα χαρακτηριστικά κάθε χρήστη (προφίλ). Τα μοντέλα και οι τεχνικές που ακολουθούνται κάθε φορά εξαρτώνται από τον σχεδιασμό και τις ιδιαίτερες ανάγκες του κάθε συστήματος. Γενικά, είναι απαραίτητα τα ακόλουθα δομικά στοιχεία[1]:

- Αναπαράσταση προφίλ
- Αρχικό Προφίλ
- Τεχνικές εκμάθησης προφίλ
- Ανατροφοδότηση σχετικότητας
- Τεχνικές προσαρμογής προφίλ
- Μέθοδοι φιλτραρίσματος πληροφοριών
- Ταίριασμα στοιχείων- προφίλ χρήστη
- Ταίριασμα παραμέτρων χρήστη

Η παραγωγή και συντήρηση ενός ακριβούς προφίλ αποτελούν ένα βασικό τμήμα του συστήματος. Η επιλογή της κατάλληλης αναπαράστασης προηγείται κάθε άλλης ενέργειας, αφού οι άλλες τεχνικές βασίζονται σε αυτή. Άλλωστε, το σύστημα δεν μπορεί να λειτουργήσει αν δεν υπάρχουν παράμετροι χρήστη. Επιπλέον, το σύστημα χρειάζεται να γνωρίζει όσο το δυνατόν περισσότερα για το χρήστη ώστε να είναι σε θέση να του παρέχει ικανοποιητικά αποτελέσματα από την αρχή. Για αυτό το λόγο είναι συνήθως απαραίτητες και τεχνικές που δημιουργούν ένα αρχικό προφίλ.

Η συνεχής αλληλεπίδραση του χρήστη με το σύστημα είναι απαραίτητη όχι μόνο για την αναγνώριση των χαρακτηριστικών γνωρισμάτων του χρήστη και την αξιολόγηση των προτάσεων αλλά και για την προσαρμογή του προφίλ αφού είναι δυνατόν οι προτιμήσεις ενός ατόμου να

αλλάζουν με την πάροδο του χρόνου. Οι πληροφορίες αυτές αποτελούν την ανατροφοδότηση σχετικότητας και αξιοποιούνται από τις τεχνικές εκμάθησης προφίλ, οι οποίες αναγνωρίζουν τις σχετικές πληροφορίες και τις χρησιμοποιούν για την προσαρμογή των προφίλ, ανάλογα βέβαια με την αναπαράσταση.

Εφόσον λοιπόν έχουν καθοριστεί τα παραπάνω για το προφίλ, ακολουθεί η «εκμετάλλευσή» του προκειμένου να προταθούν αντικείμενα ή υπηρεσίες. Οι αποφάσεις για τις προτάσεις που θα γίνουν λαμβάνονται σύμφωνα με τις υπάρχουσες πληροφορίες. Οι τεχνικές που χρησιμοποιούνται συνήθως είναι η συνεργατική (collaborative), η ανάλυση περιεχομένου (content based), η δημογραφική (demographic), με βάση τη γνώση (knowledge based) καθώς και η υβριδική (hybrid). Η δημογραφική αντιστοιχεί αντικείμενα προτάσεων σε ομάδες χρηστών ανάλογα με τα δημογραφικά τους χαρακτηριστικά. Η συνεργατική , η οποία αποτελεί ευρέως διαδεδομένη μέθοδο λαμβάνει υπόψη την ανατροφοδότηση άλλων χρηστών. Η μέθοδος με βάση το περιεχόμενο αναγνωρίζει τη σχέση ενός συγκεκριμένου χρήστη με το περιεχόμενο των αντικειμένων. Η μέθοδος φιλτραρίσματος με βάση την κοινότητα βασίζεται στις προτιμήσεις των φίλων των χρηστών και σε συνδυασμό με την αυξανόμενη δημοτικότητα των μέσων κοινωνικής δικτύωσης οδήγησε στο αυξανόμενο ενδιαφέρον προς τη μέθοδο αυτή. Τέλος, η υβριδική αποτελεί κάποιο συνδυασμό πλεονεκτημάτων των παραπάνω.

Για να την αντιστοίχιση χρηστών-αντικειμένων και το ταίριασμα χρηστών-χρηστών χρησιμοποιούνται διάφορες μέθοδοι. Οι πιο σημαντικές είναι η ομοιότητα συνημίτονου[2], η συσχέτιση Pearson[3], η κατηγοριοποίηση[4] και οι αλγόριθμοι πλησιέστερων γειτόνων[5].

1.3 Λειτουργία συστημάτων σύστασης

Τα συστήματα συστάσεων εξυπηρετούν πολλούς διαφορετικούς σκοπούς. Αρχικά, διακρίνεται ο ρόλος που παίζει ένα σύστημα σύστασης από τη μεριά του παρόχου της υπηρεσίας και από τη μεριά του χρήστη. Ένα σύστημα πρότασης ταξιδιών, για παράδειγμα, εισάγεται από ένα ταξιδιωτικό γραφείο με στόχο την αύξηση των κερδών της επιχείρησης, π.χ. με την πώληση περισσότερων δωματίων. Αντίθετα, ο κύριος σκοπός του χρήστη είναι η εύρεση κατάλληλου δωματίου. Στην πραγματικότητα υπάρχουν πολυάριθμοι λόγοι για τους οποίους στρέφονται οι πάροχοι υπηρεσιών στα συστήματα συστάσεων:

Αύξηση των προϊόντων που πωλούνται. Είναι ίσως η σημαντικότερη λειτουργία που επιτελείται για συστήματα συστάσεων που αφορούν το ηλεκτρονικό εμπόριο. Ο σκοπός αυτός επιτυγχάνεται αφού τα προτεινόμενα αντικείμενα είναι πιθανότερο να ταιριάζουν στις ανάγκες και επιθυμίες του χρήστη. Γενικότερα, και για εφαρμογές που δεν σχετίζονται με το εμπόριο, στόχος είναι η αύξηση του ποσοστού μετατροπής, του ποσοστού δηλαδή των χρηστών που περιηγήθηκαν στο προτεινόμενο αντικείμενο και το κατανάλωσαν συγκριτικά με τους χρήστες που περιηγήθηκαν χωρίς να καταναλώσουν.

Πώληση ποικιλίας αντικειμένων. Μια σημαντική λειτουργία των συστημάτων σύστασης είναι ότι δίνουν το δυνατότητα στο χρήστη να επιλέξει αντικείμενα που θα ήταν δύσκολο να βρει χωρίς σύσταση.

Αύξηση της ικανοποίησης του χρήστη. Ένα σωστά σχεδιασμένο σύστημα συστάσεων, με προτάσεις σχετικές, ακριβείς και ενδιαφέρουσες αυξάνει την συνολική ικανοποίηση του χρήστη από το σύστημα οδηγώντας σε αύξηση της χρησιμότητας του συστήματος.

Κατανόηση των επιθυμιών του χρήστη. Η συλλογή πληροφοριών για τις προτιμήσεις των χρηστών είτε άμεσα είτε έμμεσα από το σύστημα αποτελεί άλλη μία σημαντική λειτουργία των συστημάτων σύστασης. Ο πάροχος της υπηρεσίας μπορεί να χρησιμοποιήσει τα δεδομένα αυτά για βελτίωση της παραγωγής ή και διαφήμισης του προϊόντος του.

2. Εξόρυξη Δεδομένων στον Πυρήνα των Συστημάτων Σύστασης (Data Mining in Recommender Systems)

Τα συστήματα συστάσεων συνήθως εφαρμόζουν τεχνικές και μεθοδολογίες από άλλες γειτονικές περιοχές - όπως η Αλληλεπίδραση Ανθρώπου Υπολογιστή ή η Ανάκτηση Πληροφοριών (Information Retrieval). Ωστόσο, τα περισσότερα από τα συστήματα αυτά φέρουν στο πυρήνα τους έναν αλγόριθμο που μπορεί να θεωρηθεί ως ένα συγκεκριμένο παράδειγμα μιας τεχνικής εξόρυξης δεδομένων (Data Mining). Η διαδικασία της εξόρυξης δεδομένων συνήθως αποτελείται από 3 βήματα που πραγματοποιούνται διαδοχικά[1] : Προ-επεξεργασία δεδομένων, Ανάλυση Δεδομένων και Ερμηνεία Αποτελεσμάτων. Παρακάτω αναλύονται κάποιες από τις σημαντικότερες μεθόδους των παραπάνω βημάτων που χρησιμοποιούνται στα συστήματα συστάσεων.

2.1 Προ-επεξεργασία Δεδομένων

Η προ-επεξεργασία δεδομένων αποτελεί το πρώτο στάδιο της εξόρυξης δεδομένων. Ορίζουμε τα δεδομένα ως μια συλλογή από αντικείμενα και τις ιδιότητές τους. Το «αντικείμενο» μπορεί να είναι μία εγγραφή, στοιχείο, δείγμα, παρατήρηση, ή περίπτωση. Ένα «χαρακτηριστικό» μπορεί να αναφέρεται σε μεταβλητή, πεδίο, χαρακτηριστικό ή ιδιότητα. Τα δεδομένα της πραγματικής ζωής χρειάζεται κάποιου είδους επεξεργασίας (π.χ. να φιλτράρονται, να μετασχηματίζονται, να ενοποιούνται, να μειώνονται κ.α.) προκειμένου να χρησιμοποιηθούν από τις τεχνικές μηχανικής μάθησης στο στάδιο της ανάλυσης. Πιο συγκεκριμένα, τα δεδομένα συνήθως αρχικά είναι μη ποιοτικά, δηλαδή μπορεί να λείπουν τιμές, να περιέχουν σφάλματα, θόρυβο, ασυμφωνίες. Μετά την προ-επεξεργασία τα δεδομένα θα πρέπει να είναι συνεπή, ενοποιημένα και ποιοτικά προκειμένου να υπάρξουν ποιοτικά αποτελέσματα. Σε αυτή την ενότητα, αναλύονται τρία ζητήματα που έχουν ιδιαίτερη σημασία κατά το σχεδιασμό ενός συστήματος σύστασης[1]. Πρώτον, εξετάζεται η μέθοδος της δειγματοληψίας ως ένας τρόπος να μειωθεί ο αριθμός των αντικειμένων σε πολύ μεγάλες συλλογές, διατηρώντας τα κύρια χαρακτηριστικά τους. Τέλος, περιγράφονται οι πιο κοινές τεχνικές για τη μείωση των διαστάσεων και απαλοιφής θορύβου.

2.1.1 Δειγματοληψία (Sampling)

Η δειγματοληψία είναι η κύρια τεχνική που χρησιμοποιείται σε ανάκτηση πληροφοριών για την επιλογή ενός υποσυνόλου των σχετικών στοιχείων από ένα μεγάλο σύνολο δεδομένων[1]. Χρησιμοποιείται τόσο στην προ-επεξεργασία όσο και στην ερμηνεία των τελικών δεδομένων. Η δειγματοληψία χρησιμοποιείται επειδή η επεξεργασία του συνόλου δεδομένων είναι υπολογιστικά πολύ ακριβή. Μπορεί ακόμα να χρησιμοποιηθεί για τη δημιουργία training sets (συνόλων μάθησης) και testing sets (συνόλων δοκιμής). Το training σύνολο χρησιμοποιείται για να μάθει τις παραμέτρους ή να ρυθμίσει τους αλγορίθμους που χρησιμοποιούνται στο στάδιο της ανάλυσης, ενώ το testing σύνολο χρησιμοποιείται για την αξιολόγηση του μοντέλου εξασφαλίζοντας ότι έχει καλή απόδοση.

Ο βασικός στόχος της δειγματοληψίας είναι να βρει ένα υποσύνολο του αρχικού συνόλου δεδομένων που είναι αντιπροσωπευτικό του συνόλου, δηλαδή έχει περίπου την ίδια ζητούμενη ιδιότητα. Η απλούστερη τεχνική δειγματοληψίας είναι η τυχαία δειγματοληψία, όπου υπάρχει μια ίση πιθανότητα για την επιλογή κάθε στοιχείου. Ωστόσο, πιο εξελιγμένες προσεγγίσεις είναι δυνατές, όπως η στρωματοποιημένη δειγματοληψία κατά την οποία χωρίζονται τα δεδομένα σε διάφορα τμήματα με βάση ένα ιδιαίτερο χαρακτηριστικό, και ακολουθείται από τυχαία δειγματοληψία σε κάθε διαμέρισμα ξεχωριστά. Η πιο κοινή προσέγγιση για τη δειγματοληψία συνίσταται στη χρήση δειγματοληψίας χωρίς αντικατάσταση. Με τον τρόπο αυτό δειγματοληψίας, όταν επιλέγεται ένα στοιχείο, αφαιρείται από τον πληθυσμό. Ωστόσο, είναι επίσης δυνατή η δειγματοληψία με αντικατάσταση, όπου δεν απομακρύνονται αντικείμενα από τον πληθυσμό, εφόσον έχουν επιλεγεί, επιτρέποντας το ίδιο δείγμα να επιλεγεί περισσότερες από μία φορά.

Μια κοινή προσέγγιση σε σύστημα συστάσεων είναι η δειγματοληψία των διαθέσιμων πληροφοριών από τους χρήστες - π.χ. οι αξιολογήσεις τους – για να χωριστούν σε training και testing σύνολα (σύνολα εκπαίδευσης και δοκιμής). Χρησιμοποιούνται διάφοροι τρόποι δειγματοληψίας ανάλογα με τις ανάγκες του συστήματος. Για παράδειγμα, μπορεί να χρειαστεί να δειγματοληπτούνται μόνο οι πιο πρόσφατες αξιολογήσεις - δεδομένου ότι εκείνες είναι αυτές που ενδιαφέρουν σε μια κατάσταση πραγματικού κόσμου.

2.1.2 Μείωση διαστάσεων (Dimensionality Reduction)

Είναι σύνηθες σε συστήματα σύστασης να υπάρχουν, εκτός από σύνολα δεδομένων με χαρακτηριστικά που ορίζουν ένα χώρο υψηλής διάστασης, και πολλές σποραδικές πληροφορίες σε αυτό το χώρο - δηλαδή υπάρχουν τιμές για έναν περιορισμένο αριθμό χαρακτηριστικών ανά αντικείμενο. Οι έννοιες της πυκνότητας και της απόστασης μεταξύ των σημείων, τα οποία είναι κρίσιμα για την ομαδοποίηση και την ανίχνευση ακραίων τιμών (αναλύονται σε επόμενες ενότητες), αποκτούν μικρότερη σημασία σε χώρους υψηλής διάστασης. Οι τεχνικές μείωσης διαστάσεων αντιμετωπίζουν αυτό το πρόβλημα μετατρέποντας τον αρχικό χώρο υψηλών

διαστάσεων σε ένα χαμηλότερων διαστάσεων[1]. Τα αραιά δεδομένα είναι συνηθισμένο πρόβλημα και στα συστήματα συστάσεων. Ακόμη και στην πιο απλή τους μορφή, είναι πιθανό να υπάρχει μια αραιή μήτρα με χιλιάδες σειρές και στήλες (δηλαδή χρήστες και αντικείμενα), οι περισσότερες εκ των οποίων να είναι μηδενικά. Ως εκ τούτου, η μείωση διαστάσεων καθίσταται αναγκαία.

2.1.3 Απαλοιφή θορύβου (Denoising)

Τα δεδομένα που συλλέγονται για τους σκοπούς της εξόρυξης δεδομένων μπορεί να υπόκειται σε διάφορα είδη θορύβου όπως τιμές που λείπουν ή ακραίες τιμές. Η απαλοιφή θορύβου είναι ένα πολύ σημαντικό βήμα της προ-επεξεργασίας, η οποία αποσκοπεί στην κατάργηση οποιασδήποτε ανεπιθύμητης πληροφορίας των δεδομένων, μεγιστοποιώντας έτσι τη χρήσιμη πληροφορία. Σε μια γενική έννοια ορίζουμε το θόρυβο ως ανεπιθύμητο τεχνούργημα που εισάγεται στη φάση συλλογής δεδομένων και μπορεί να επηρεάσει το αποτέλεσμα της ανάλυσης των δεδομένων και την ερμηνεία τους. Στο πλαίσιο των συστημάτων σύστασης, γίνεται διάκριση μεταξύ φυσικού και κακόβουλο θορύβου[1,6]. Ο πρώτος είναι ο θόρυβος που εισάγεται άθελα τους από χρήστες όταν δίνεται ανατροφοδότηση σχετικά με τις προτιμήσεις τους. Ο δεύτερος αφορά το θόρυβο που εισάγεται σκόπιμα σε ένα σύστημα προκειμένου να διαβάλλει τα αποτελέσματα. Είναι σαφές ότι ο κακόβουλος θόρυβος μπορεί να επηρεάσει την έξοδο ενός συστήματος σύστασης. Μελέτες που έχουν πραγματοποιηθεί δείχνουν πως οι επιπτώσεις του φυσικού θορύβου σχετικά με την απόδοση του συστήματος σύστασης είναι κάθε άλλο παρά αμελητέες [6]. Για να αντιμετωπιστεί αυτό το πρόβλημα, έχουν σχεδιαστεί προσεγγίσεις απομάκρυνσης θορύβου που είναι σε θέση να βελτιώσουν την ακρίβεια, ζητώντας από μερικούς χρήστες στο βαθμολογήσουν ξανά ορισμένα στοιχεία [8].

2.2 Ανάλυση Δεδομένων

Η ανάλυση δεδομένων, που αποτελεί το δεύτερο στάδιο της εξόρυξης δεδομένων, αναφέρεται σε ένα ευρύ φάσμα των μαθηματικών μοντέλων, τεχνικών και εργαλείων λογισμικού που χρησιμοποιούνται για την εύρεση μοτίβων σε δεδομένα και τα χρησιμοποιούν για την κατασκευή μοντέλων. Στο πλαίσιο των συστημάτων σύστασης, η ανάλυση δεδομένων είναι όρος που χρησιμοποιείται για να περιγράψει την συλλογή των τεχνικών ανάλυσης που χρησιμοποιούνται για τη δημιουργία κανόνων σύστασης ή μοντέλων σύστασης από μεγάλα σύνολα δεδομένων . Τα συστήματα σύστασης που αντλούν τεχνικές από την εξόρυξη δεδομένων κάνουν τις συστάσεις τους χρησιμοποιώντας γνώσεις που αντλήθηκαν από τις δράσεις και τα

χαρακτηριστικά των χρηστών. Αυτοί οι αλγόριθμοι περιλαμβάνουν κυρίως την ομαδοποίηση[1] και την κατηγοριοποίηση[1], τεχνικές που αναλύονται παρακάτω.

2.2.1 Ανάλυση με Κατηγοριοποίηση των Δεδομένων (Data Analysis and Classification)

Η τεχνική της κατηγοριοποίησης περιλαμβάνει υπολογιστικά μοντέλα για την ανάθεση μιας κατηγορίας σε μία είσοδο. Οι είσοδοι μπορεί να είναι φορείς των χαρακτηριστικών για τα αντικείμενα που έχουν ταξινομηθεί ή δεδομένων σχετικά με τις σχέσεις μεταξύ των στοιχείων. Υπάρχουν πολλοί τύποι ταξινόμησης, αλλά σε γενικές γραμμές θα αναλυθούν η ταξινόμηση με επίβλεψη και χωρίς επίβλεψη[1]. Στην ταξινόμηση με επίβλεψη, ένα σύνολο από ετικέτες ή κατηγορίες είναι γνωστό εκ των προτέρων και υπάρχει ένα σύνολο ταξινομημένων παραδειγμάτων τα οποία αποτελούν ένα σύνολο εκμάθησης. Στην ταξινόμηση χωρίς επίβλεψη, οι ετικέτες ή οι κατηγορίες είναι άγνωστες εκ των προτέρων και ο στόχος είναι (σύμφωνα με κάποια κριτήρια) να οργανωθούν τα στοιχεία στο χέρι. Ένας τρόπος για να οικοδομηθεί ένα σύστημα συστάσεων με ταξινόμηση είναι να χρησιμοποιούνται οι πληροφορίες για ένα προϊόν και χρήστη ως είσοδος, και να έχει ως έξοδο μία κατηγορία που να αντιπροσωπεύει το πόσο έντονα πρέπει να συστήσει το προϊόν στο χρήστη. Η κατηγοριοποίηση μπορεί να υλοποιηθεί με τη χρήση πολλών διαφορετικών στρατηγικών μηχανικής μάθησης μερικές από τις οποίες αναφέρονται στις παρακάτω υποενότητες.

2.2.1.1 Μέτρα Ομοιότητας (Similarity Measures)

Μία από τις προτιμώμενες προσεγγίσεις για το συνεργατικό φιλτράρισμα(βλέπε Ενότητα 3.1) είναι η χρήση του ταξινομητή kNN *(*kNN classifier*)* που θα περιγραφεί στην ενότητα 2.2.1.2. Αυτή η μέθοδος ταξινόμησης - όπως οι περισσότεροι ταξινομητές (classifiers) και τεχνικές ομαδοποίησης - εξαρτάται σε μεγάλο βαθμό από τον καθορισμό του κατάλληλο μέτρου ομοιότητας ή απόστασης. Το απλούστερο και πιο κοινό παράδειγμα ενός μέτρου απόστασης είναι η Ευκλείδεια απόσταση:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}, \qquad 2.1$$

όπου n είναι ο αριθμός των διαστάσεων (χαρακτηριστικών) και x_i και y_i είναι οι i-ιδιότητες των αντικειμένων-δεδομένων x και y, αντίστοιχα. Η Minkowski απόσταση είναι μια γενίκευση της Ευκλείδειας απόστασης:

$$d(x,y) = \sqrt[\lambda]{\sum_{i=1}^{n}|x_i - y_i|^{\lambda}} \qquad 2.2$$

όπου το λ είναι ο βαθμός της απόστασης. Μια άλλη πολύ συνηθισμένη προσέγγιση είναι να θεωρούνται τα αντικείμενα ως διανύσματα n - διάστατου χώρου και να υπολογίζεται η ομοιότητα τους ως το συνημίτονο της γωνίας που σχηματίζουν:

$$\cos(x,y) = \frac{x\,.\,y}{\|x\|\,.\,\|y\|} \qquad 2.3$$

όπου $x\,.\,y$ το εσωτερικό γινόμενο των διανυσμάτων και $\|x\|$ η $L2$-νόρμα του διανύσματος x. Αυτή η ομοιότητα είναι γνωστή ως η ομοιότητα συνημίτονου. Η ομοιότητα μεταξύ των αντικειμένων μπορεί επίσης να δίδεται από την συσχέτισή τους, η οποία μέτρα τη γραμμική σχέση μεταξύ αντικειμένων. Ενώ υπάρχουν αρκετές μέθοδοι συσχέτισης που μπορούν να εφαρμοστούν, η συσχέτιση Pearson είναι η πιο συχνά χρησιμοποιούμενη. Δεδομένης της συνδιακύμανσης Σ των σημείων-δεδομένων x και y και την τυπική απόκλιση τους σ, υπολογίζεται η συσχέτιση Pearson ως:

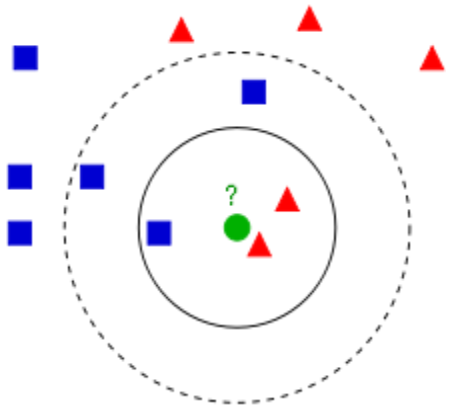
$$Pearson(x,y) = \frac{\Sigma(x,y)}{\sigma_x\,.\,\sigma_y} \qquad 2.4$$

Τα συστήματα συστάσεων παραδοσιακά χρησιμοποιούν είτε την ομοιότητα συνημίτονου ή την συσχέτιση Pearson -ή μία από τις πολλές παραλλαγές τους μέσω, για παράδειγμα, των σταθμισμένων συστημάτων. Οι Spertus et al. [9] έκαναν μια μεγάλης κλίμακας μελέτη για την αξιολόγηση έξι διαφορετικών μέτρων ομοιότητας στο πλαίσιο του κοινωνικού δικτύου Orkut. Αν και τα αποτελέσματα τους θα μπορούσαν να ωθούνται από τη συγκεκριμένη ρύθμιση του πειράματος τους, είναι ενδιαφέρον να σημειωθεί ότι οι καλύτερες συστάσεις ήταν εκείνες που παράχθηκαν χρησιμοποιώντας την ομοιότητα συνημίτονου. Οι Lathia et al. [10] επίσης, πραγματοποίησαν μια μελέτη αρκετών μέτρων ομοιότητας, όπου κατέληξαν στο συμπέρασμα ότι, στη γενική περίπτωση, η ακρίβεια πρόβλεψης ενός συστήματος σύστασης δεν επηρεάστηκε από την επιλογή του μέτρου ομοιότητας.

2.2.1.2 Πλησιέστερος γείτονας (Nearest Neighbor)

Η κατηγοριοποίηση που βασίζεται σε παραδείγματα λειτουργεί αποθηκεύοντας αρχεία εκπαίδευσης (training) και χρησιμοποιώντας τα για την πρόβλεψη της κατηγορίας σε καινούριες περιπτώσεις. Το πιο δημοφιλές παράδειγμα είναι ο κατηγοριοποίηση k-πλησιέστερου γείτονα *(k* – nearest neighbor kNN*)* [11]. Δεδομένου ενός σημείου που πρέπει να ταξινομηθεί, η μεθοδολογία kNN βρίσκει τα k πλησιέστερα σημεία (κοντινότεροι γείτονες) από τα αρχεία εκπαίδευσης. Στη συνέχεια αναθέτει την κλάση σύμφωνα με τις ετικέτες των πλησιέστερων γειτόνων του. Η βασική ιδέα είναι ότι εάν μία παρατήρηση εμπίπτει σε μια συγκεκριμένη γειτονιά, όπου μια συγκεκριμένη ετικέτα κατηγορίας κυριαρχεί, το πιθανότερο είναι να ανήκει σε αυτή ακριβώς την κατηγορία.

Δεδομένου ενός σημείου q για το οποίο θέλουμε να γνωρίζουμε την κατηγορία l του, και το σύνολο εκπαίδευσης $X = \{\{x_1, l_1\}\ \{x_n \dots\}\}$, όπου X_j είναι η j-οστό στοιχείο και l_j είναι ετικέτα της κατηγορίας του, οι k -κοντινότεροι γείτονες θα βρεθούν σε ένα υποσύνολο $Y = \{\{y_1, l_1\}\ \{y_k \dots\}\}$, τέτοιο ώστε $Y \in X$ και $\sum_1^{\kappa} d(q, y_k)$ να είναι ελάχιστο. Το σύνολο Y περιέχει τα σημεία k στο X που είναι πιο κοντά στο ζητούμενο q σημείο. Στη συνέχεια, η ετικέτα της κατηγορίας του q είναι $\mathrm{l} = f(\{l_1 \dots l_k\})$.



Εικόνα 2.1: Παράδειγμα *k*NN ταξινομητή [13]

Ίσως το πιο δύσκολο θέμα στις kNN τεχνικές είναι η επιλογή της τιμής του k. Αν η τιμή του k είναι πολύ μικρή, ο ταξινομητής θα είναι ευαίσθητος στο θόρυβο. Αλλά αν το k είναι πολύ μεγάλο, η γειτονιά θα μπορούσε να περιλαμβάνει πάρα πολλά σημεία από άλλες κατηγορίες. Το παραπάνω σχήμα Σχ. 2.1 δείχνει πώς οι διάφορες τιμές του k δίνουν διαφορετική κατηγορία για το ζητούμενο σημείο (κυκλικό σημείο). Εάν k = 3 (κύκλος με συνεχή γραμμή) η κατηγορία του θα είναι τρίγωνο, αφού υπάρχουν δύο τρίγωνα εντός του κύκλου, ενώ αν k = 5 (κύκλος με διακεκομμένη γραμμή) κατατάσσεται ως τετράγωνο. Οι kNN ταξινομητές είναι από τους πιο απλούς όλων των αλγορίθμων μηχανικής μάθησης. Σε αντίθεση με άλλα μοντέλα , όπως δέντρα απόφασης, που αναφέρονται έπειτα, οι kNN ταξινομητές αφήνουν πολλές αποφάσεις για την στάδιο της ταξινόμησης. Επομένως, το να κατατάσσονται άγνωστα στοιχεία είναι σχετικά ακριβό υπολογιστικά. Ο πλησιέστερος γείτονας είναι μία από τις πιο κοινές προσεγγίσεις και για τα συστήματα συστάσεων. Ένα από τα πλεονεκτήματα αυτού του ταξινομητή είναι ότι είναι εννοιολογικά πάρα πολύ κοντά με την ιδέα του συστήματος σύστασης: Η εύρεση παρόμοιων χρηστών (ή παρόμοιων αντικειμένων) είναι ουσιαστικά ισοδύναμο με την εύρεση των γειτόνων για ένα συγκεκριμένο χρήστη ή στοιχείο. Ένα άλλο πλεονέκτημα είναι ότι δεν απαιτείται ένα συγκεκριμένο μοντέλο. Ως εκ τούτου, το σύστημα μπορεί να προσαρμοστεί σε γρήγορες αλλαγές στη μήτρα βαθμολογιών του χρήστη, τον πίνακα δηλαδή στοιχείων*χρήστη με τις βαθμολογίες του χρήστη για τα διάφορα στοιχεία. Δυστυχώς, αυτό συνεπάγεται υπολογισμό εκ νέου των γειτόνων και ως εκ τούτου, της μήτρας ομοιότητας.

Η προσέγγιση kNN, αν και απλή, έχει δείξει καλά αποτελέσματα όσων αφορά την ακρίβεια και είναι πολύ δεκτική σε βελτιώσεις[1]. Για παράδειγμα, στο πλαίσιο του βραβείου Netflix, οι Bell και Koren[14] πρότειναν μια μέθοδο για την απομάκρυνση «παγκόσμιων επιπτώσεων», όπως το γεγονός ότι ορισμένα στοιχεία μπορεί ως επί το πλείστον να βαθμολογούνται από χρήστες που τείνουν να βαθμολογούν υψηλότερα, ενώ ορισμένα άλλα στοιχεία να προσελκύουν χρήστες που έχουν την τάση να βαθμολογούν χαμηλότερα. Προτείνουν, επίσης, μια μέθοδο βελτιστοποίησης για τον υπολογισμό των εισαγόμενων βαρών όταν δημιουργείται η γειτονιά.

2.2.1.3 Χρήση Δέντρων αποφάσεων (Decision Trees)

Τα δέντρα απόφασης είναι αλγόριθμοι κατηγοριοποίησης στους οποίους παρουσιάζονται οι πληροφορίες σε δεντρικές μορφές[1,13]. Τα δέντρα αποφάσεων έχουν τα εξής γνωρίσματα: οι εσωτερικοί κόμβοι αντιστοιχούν σε κάποιο γνώρισμα, ο διαχωρισμός (split) ενός κόμβου σε παιδιά γίνεται με βάση την ετικέτα στην ακμή και τα φύλλα αντιστοιχούν σε κλάσεις. Οι κόμβοι του δέντρου μπορεί να είναι: α) κόμβοι απόφασης, σε αυτούς τους κόμβους ένα απλό γνώρισμα-τιμή αξιολογείται προκειμένου να καθοριστεί σε ποιον κλάδο του υποδέντρου ανήκει ή β) κόμβους- φύλλα που δείχνουν την τιμή του ζητούμενου χαρακτηριστικού. Υπάρχουν πολλοί αλγόριθμοι που εφαρμόζονται στα δέντρα απόφασης με τους Hunt[15], CART[16], ID3[17] και C4.5[18] να είναι οι πιο κοινοί. Ο αναδρομικός αλγόριθμος Hunt, ο οποίος είναι από τους πρώτους και πιο κατανοητούς, κτίζει το δέντρο αναδρομικά. Αρχικά όλες οι παρατηρήσεις είναι σε έναν κόμβο (ρίζα) και έστω D_t το σύνολο των παρατηρήσεων εκπαίδευσης που έχουν φτάσει στον κόμβο t. Η γενική διαδικασία (αναδρομικά σε κάθε κόμβο) έχει ως εξής : Αν το D_t περιέχει παρατηρήσεις που ανήκουν στην ίδια κλάση y_t, τότε ο κόμβος t είναι κόμβος φύλλο με ετικέτα y_t . Αν D_t είναι το κενό σύνολο (αυτό σημαίνει ότι δεν υπάρχει εγγραφή στο σύνολο εκπαίδευσης με αυτό το συνδυασμό τιμών), τότε D_t γίνεται φύλλο με κλάση αυτή της πλειοψηφίας των παρατηρήσεων εκπαίδευσης ή ανάθεση κάποιας default κλάσης Τέλος, αν το D_t περιέχει εγγραφές που ανήκουν σε περισσότερες από μία κλάσεις, τότε χρησιμοποιείται ένας έλεγχος-γνωρίσματος για το διαχωρισμό των δεδομένων σε μικρότερα υποσύνολα. Η διάσπαση ενός κόμβου επιλέγεται με βάση το κέρδος διάσπασης το οποίο ορίζεται ως :

$$\Delta i = I(parent) - \sum_{j=1}^{ki} \left(\frac{N(vj)I(vj)}{N}\right) \qquad 2.5$$

όπου k_i οι τιμές του χαρακτηριστικού i , N ο αριθμός των παρατηρήσεων και v_j το j -οστό βάρος των χαρακτηριστικών σύμφωνα με τις τιμές του i. Τέλος, I είναι η συνάρτηση που μετρά την «ποιότητα» ενός κόμβου. Υπάρχουν διάφορα μέτρα της «ποιότητας» με τα Gini [19], εντροπία και λάθος ταξινόμησης να είναι τα πιο κοινά. Στον αλγόριθμο του Hunt, το δέντρο μεγαλώνει με στόχο να κατηγοριοποιήσει με τέλειο τρόπο κάθε εγγραφή. Παρ' όλο που αυτό ακούγεται λογικό σα στρατηγική, μπορεί να οδηγήσει σε δυσκολίες, όταν υπάρχει «θόρυβος» στα δεδομένα (κάτι

που δύσκολα αποφεύγεται), ή όταν ο αριθμός των εγγραφών είναι πολύ μικρός για να δημιουργήσει ένα αντιπροσωπευτικό δείγμα για το μοντέλο που θα δημιουργηθεί. Σε αυτές τις περιπτώσεις, ο αλγόριθμος μπορεί να παράγει δέντρα με μικρή ακρίβεια. Για την αποφυγή αυτών των προβλημάτων έχουν προταθεί οι παρακάτω μέθοδοι: μέθοδοι με τις οποίες το δέντρο σταματάει να αναπτύσσεται νωρίτερα, από το σημείο που περιγράφει τέλεια τα δεδομένα και μέθοδοι οι οποίες αναπτύσσουν όλο το δέντρο και στη συνέχεια το «κλαδεύουν» (pruning). Τα κύρια πλεονεκτήματα της κατηγοριοποίησης που χρησιμοποιεί δέντρα απόφασης είναι το μικρό κόστος κατασκευής τους και ότι είναι εξαιρετικά γρήγοροι στην ταξινόμηση άγνωστων περιπτώσεων. Ένα άλλο πλεονέκτημα είναι ότι μπορεί να χρησιμοποιηθεί για την παραγωγή ενός συνόλου κανόνων που είναι εύκολο να ερμηνευτούν, διατηρώντας παράλληλα ακρίβεια συγκρίσιμη με άλλες βασικές τεχνικές ταξινόμησης.

Τα δένδρα απόφασης μπορούν να χρησιμοποιηθούν σε προσεγγίσεις συστημάτων συστάσεων που βασίζονται σε μοντέλα. Μία από τις δυνατότητες τους είναι η χρήση των χαρακτηριστικών περιεχομένου για να χτίσουν ένα δέντρο απόφασης που μοντελοποιεί όλες τις μεταβλητές που σχετίζονται με τις προτιμήσεις των χρηστών. Οι Bouza et al. [20] χρησιμοποιούν αυτήν την ιδέα για να κατασκευάσουν ένα δέντρο αποφάσεων που χρησιμοποιεί σημασιολογικές πληροφορίες των στοιχείων. Το δέντρο κατασκευάζεται μόλις ο χρήστης έχει βαθμολογήσει μόνο δύο στοιχεία. Τα χαρακτηριστικά για καθένα από τα στοιχεία αυτά χρησιμοποιούνται για να κατασκευαστεί ένα μοντέλο που εξηγεί τις αξιολογήσεις των χρηστών με κριτήριο διάσπασης το κέρδος πληροφορίας για κάθε χαρακτηριστικό.

Όπως είναι αναμενόμενο, είναι πολύ δύσκολο και μη πρακτικό να οικοδομήσουμε ένα δέντρο απόφασης που να προσπαθεί να εξηγήσει όλες τις μεταβλητές που εμπλέκονται στη διαδικασία λήψης αποφάσεων. Ωστόσο, μπορεί να χρησιμοποιηθεί για να διαμορφώσει ένα συγκεκριμένο μέρος του συστήματος. Οι Cho et al. [21], για παράδειγμα, παρουσιάζουν ένα σύστημα συστάσεων για αγορές στο Διαδίκτυο το οποίο συνδυάζει τα δέντρα αποφάσεων με άλλες τεχνικές. Το δέντρο απόφασης χρησιμοποιείται ως φίλτρο για να επιλεγεί σε ποιους χρήστες θα πρέπει να απευθύνονται οι συστάσεις. Για την κατασκευή του μοντέλου δημιουργούν ένα σύνολο υποψήφιων χρηστών διαλέγοντας εκείνους που έχουν επιλέξει προϊόντα από μια συγκεκριμένη κατηγορία κατά τη διάρκεια ενός δεδομένου χρόνου. Σε αυτή την περίπτωση, η εξαρτημένη μεταβλητή για την κατασκευή του δέντρου απόφασης που θα επιλεγεί είναι εάν ο πελάτης είναι πιθανό να αγοράσει νέα προϊόντα της ίδιας κατηγορίας. Οι Nikovski και Kulev [22] ακολουθούν μια παρόμοια προσέγγιση, στην οποία τα συχνά στοιχεία- σύνολα εντοπίζονται στο σύνολο δεδομένων αγοράς και στη συνέχεια εφαρμόζονται αλγόριθμοι δέντρο-μάθησης για την απλούστευση των κανόνων συστάσεων.

2.2.1.4 Χρήση Bayesian Ταξινομητών (Bayesian Classifiers)

Η Bayesian κατηγοριοποίηση χρησιμοποιεί τις πιθανότητες για την επίλυση του προβλήματος κατάταξης σε κατηγορίες[1]. Βασίζεται τον ορισμό της δεσμευμένης πιθανότητας και του θεωρήματος Bayes. Η Bayesian μεθοδολογία εξετάζει κάθε χαρακτηριστικό και την ετικέτα κατηγορίας με (συνεχείς ή διακριτές) τυχαίες μεταβλητές. Δεδομένης μίας εγγραφής με N χαρακτηριστικά $(A_1, A_2, \ldots, A_n)$, στόχος είναι η πρόβλεψη κατηγορίας C_k με την εύρεση της τιμής της C_k που μεγιστοποιεί την πιθανότητα $P(C_k \mid A_1, A_2, \ldots, A_n)$. Εφαρμόζοντας το θεώρημα του Bayes $P(C_k \mid A_1, A_2, \ldots, A_n) \propto P(A_1, A_2, \ldots, A_n \mid C_k)\, P(C_k)$. Η Naïve Bayes κατηγοριοποίηση, βασίζεται στα θεωρήματα του Bayes, και υποθέτει ότι υπάρχει ανεξαρτησία ανάμεσα στις παραμέτρους των δεδομένων. Αν και με μια πρώτη ματιά αυτή η θεώρηση μοιάζει υπέρ-απλουστευμένη (γι' αυτό και ονομάζεται και naïve, δηλαδή αφελής), μπορεί να προσφέρει το ίδιο καλά ή ακόμα και καλύτερα αποτελέσματα από άλλες, πιο πολύπλοκες μεθόδους. Προκειμένου να εκτιμηθεί η δεσμευμένη πιθανότητα, $P(A_1, A_2, \ldots, A_n \mid C_k)$, η παραπάνω υπόθεση οδηγεί στη σχέση $P(A_1, A_2, \ldots, A_n \mid C_k) = P(A_1 \mid C_k)\, P(A_2 \mid C_k) \ldots P(A_n \mid C_k)$.

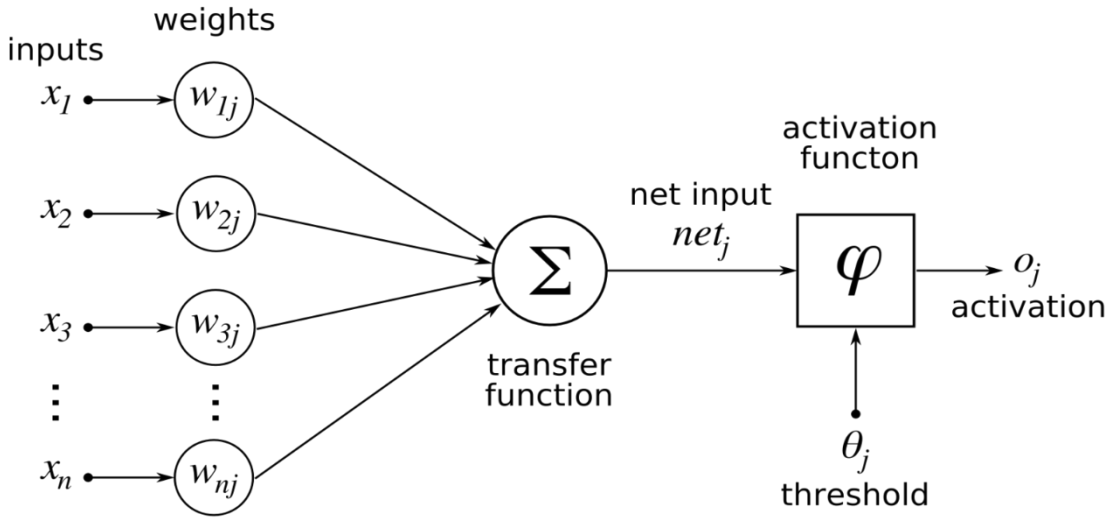
Τα κύρια οφέλη της Naive Bayes κατηγοριοποίησης είναι ότι δεν επηρεάζεται από σε απομονωμένα σημεία θορύβου και μη σχετικά χαρακτηριστικά, και μπορεί να χειρίζεται τις τιμές που λείπουν αγνοώντας τες κατά την εκτίμηση υπολογισμού πιθανοτήτων. Ωστόσο, η υπόθεση για την ανεξαρτησία των δεδομένων δεν μπορεί να ισχύει για ορισμένες ιδιότητες, που θα μπορούσαν να συσχετιστούν. Σε αυτή την περίπτωση, η συνήθης προσέγγιση είναι τα λεγόμενα Bayesian Belief Δίκτυα (BBN) (ή Bayesian Δίκτυα). Τα Bayesian δίκτυα αναπαριστούν τη δομή ενός πιθανοτικού γραφικού μοντέλου με τη χρήση γράφων για την αποτύπωση των υποθέσεων ανεξαρτησίας μεταξύ των μεταβλητών. Κάθε μεταβλητή σε ένα Bayesian δίκτυο αναπαρίσταται με έναν κόμβο. Κάθε κόμβος διαθέτει καταστάσεις, ή διαφορετικά ένα σύνολο από πιθανές τιμές που αντιστοιχούν σε κάθε μεταβλητή. Οι κόμβοι συνδέονται μεταξύ τους με κατευθυνόμενα βέλη –πλευρές (edges) τα οποία φανερώνουν την αλληλεξάρτηση των μεταβλητών υποδεικνύοντας και την κατεύθυνση της επιρροής. Τέλος, σε κάθε κόμβο αντιστοιχεί και ένας πίνακας υπό συνθήκη πιθανοτήτων (conditional probability table). Τα δίκτυα αυτά είναι πολύ χρήσιμα καθώς μπορούν να διαβαστούν λεπτομέρειες του μοντέλου απευθείας από το γράφο και είναι υπολογιστικά αποδοτικά. Αν και τα BNs δεν μπορούν να αποτυπώσουν όλες τις δυνατές σχέσεις μεταξύ των μεταβλητών, είναι ιδανικά για την αναπαράσταση σχέσεων αιτίου-αποτελέσματος.

Η Bayesian κατηγοριοποίηση είναι ιδιαίτερα δημοφιλής και για συστήματα συστάσεων που βασίζονται σε μοντέλα. Χρησιμοποιούνται συχνά για να εξαχθεί ένα μοντέλο για συστήματα συστάσεων με βάση το περιεχόμενο (βλέπε Ενότητα 3.2) αλλά και σε συστήματα συνεργατικού φιλτραρίσματος. Οι Miyahara και Pazzani [23], για παράδειγμα, χρησιμοποιούν Naive Bayes κατηγοριοποίηση για να την εφαρμογή ενός συστήματος συστάσεων. Ορίζουν δύο κατηγορίες: μου αρέσει (*like*) και δεν μου αρέσει (*don't like*). Στο πλαίσιο αυτό, προτείνουν δύο τρόπους χρήσης Naive Bayesian: Το μοντέλο μετασχηματισμένων δεδομένων *(«Transformed Data Model»)* υποθέτει ότι όλα τα χαρακτηριστικά είναι εντελώς ανεξάρτητα, και η επιλογή τους υλοποιείται ως ένα στάδιο προ-επεξεργασίας. Από την άλλη πλευρά, το μοντέλο αραιών δεδομένων *(«Sparse Data Model»)* υποθέτει ότι μόνο τα γνωστά χαρακτηριστικά είναι κατάλληλα για ταξινόμηση. Επιπλέον, κάνει χρήση των δεδομένων, που και οι δύο χρήστες έχουν

βαθμολογήσει από κοινού, κατά την εκτίμηση των πιθανοτήτων. Τα πειράματα δείχνουν πως και τα δύο μοντέλα έχουν καλύτερες επιδόσεις από ένα σύστημα συνεργατικού φιλτραρίσματος που βασίζεται στη συσχέτιση.

2.2.1.5 Χρήση Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks)

Ένα τεχνητό νευρωνικό δίκτυο (*Artificial Neural Network ANN*) είναι ένα σύνολο διασυνδεδεμένων κόμβων και σταθμισμένων συνδέσεων που είναι εμπνευσμένο από την αρχιτεκτονική του βιολογικού εγκεφάλου. Οι κόμβοι στα ANN ονομάζονται νευρώνες[1,13]. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο *συναπτικό βάρος* και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη *συνάρτηση ενεργοποίησης*, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη. Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Τα δίκτυα αυτά επομένως έχουν τη δυνατότητα να μάθουν να επιλύουν το πρόβλημα της κατηγοριοποίησης αφού εκπαιδευτούν με τα κατάλληλα δεδομένα.



Εικόνα 2.2: Ένα τεχνητό νευρωνικό δίκτυο

Η απλούστερη περίπτωση ενός ANN είναι ο νευρώνας Perceptron, που απεικονίζεται στο σχήμα Σχ. 2.2. Αν θεωρηθεί η φ λειτουργία ενεργοποίησης ως η απλή συνάρτηση κατωφλίου, η έξοδος λαμβάνεται με την άθροιση της τιμής κάθε εισόδου σύμφωνα με τα βάρη των συνδέσεων και συγκρίνοντας την έξοδο αυτή με κάποιο θ_k κατώφλιο. Η λειτουργία εξόδου μπορούν να εκφραστεί χρησιμοποιώντας την εξίσωση :

$$y_k = \varphi\left(\sum_{i=0}^{N} x_{ki}\, w_{ki}\right) \qquad 2.6$$

όπου x_{ki} είναι η i-οστή είσοδος του k νευρώνα, w_{ki} : το i-οστό συναπτικό βάρος του k νευρώνα και $\varphi(.)$ η συνάρτηση ενεργοποίησης του νευρωνικού δικτύου.

Το μοντέλο Perceptron είναι μία γραμμική κατηγοριοποίηση που έχει ένα απλό και αποτελεσματικό αλγόριθμο μάθησης. Όμως, εκτός από την απλή συνάρτηση κατωφλίου που χρησιμοποιείται στο μοντέλο Perceptron, υπάρχουν αρκετές άλλες κοινές επιλογές για το συνάρτηση ενεργοποίησης όπως σιγμοειδής, $tanh$, ή βηματικές συναρτήσεις.

Τα κύρια πλεονεκτήματα των ANN είναι ότι ανάλογα με την συνάρτηση ενεργοποίησης μπορούν να εκτελούν μη γραμμική κατηγοριοποίηση και λόγω των παράλληλων συνδέσεων, μπορεί να είναι αποτελεσματικά, ακόμη και αν μέρος του δικτύου καταρρεύσει. Το κύριο μειονέκτημα είναι ότι είναι δύσκολη η υλοποίηση της ιδανικής τοπολογίας του δικτύου για ένα συγκεκριμένο πρόβλημα και μόλις αποφασιστεί η τοπολογία αυτή θα λειτουργήσει ως ένα περιορισμό για το σφάλμα ταξινόμησης. Τα ANN ανήκουν στην κατηγορία της κατηγοριοποίησης , που δεν παρέχει καμία σημασιολογική πληροφορία για απόκτηση γνώσης - δηλαδή λειτουργούν σαν προσέγγιση μαύρου κουτιού (black box).

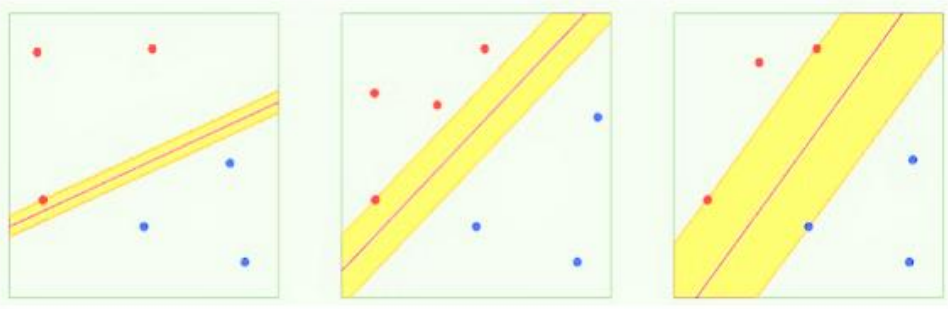
Μεταξύ των διαφόρων τύπων των ANN, είναι τα Back Propagation Neural Δίκτυα (BPNN)[25], Kohonen Self Organizing (SOMs)[26], τα Δίκτυα Hopfield[27], και τα Feedforward Δίκτυα[28]. Παρακάτω παρουσιάζονται συστήματα συστάσεων που βασίζονται σε τεχνητά νευρωνικά δίκτυα. Ένα context-aware σύστημα συστάσεων για σύσταση τηλεοπτικών προγραμμάτων παρουσιάζεται στο[29]. Τα διαθέσιμα τηλεοπτικά προγράμματα αναπαρίστανται στον παρακείμενο διανυσματικό χώρο και στη συνέχεια εφαρμόζεται ένα είδος μετασχηματισμού για τη μείωση των διαστάσεων. Για να εκτιμηθεί κατά πόσον ένα συγκεκριμένο τηλεοπτικό πρόγραμμα είναι σημαντικού ενδιαφέροντος για τον χρήστη ή όχι, εφαρμόζεται ένα Feedforward Νευρωνικό Δίκτυο. Το πρόβλημα της «ψυχρής εκκίνησης» στην προσέγγιση αντιμετωπίζεται από την εκμάθηση των τηλεοπτικών τηλεοπτικές συνήθειες των τηλεθεατών. Το χρονικό πλαίσιο μαζί με τα δεδομένα βρίσκονται σε ένα «κρυμμένο» στρώμα του νευρωνικού δικτύου. Οι συναφείς πληροφορίες ανακτώνται από το ρολόι του συστήματος και παρέχονται στο νευρωνικό δίκτυο με την εισαγωγή πρόσθετων κόμβων εισόδου. Η δοκιμή της προσέγγισης αυτής με ένα «κρυφό» κόμβο που είχε εκπαιδευτεί για τις τέσσερις τελευταίες αλληλεπιδράσεις του χρήστη είχε ακρίβεια πάνω από 92%.

Ένα άλλο παράδειγμα είναι αυτό των Biancalana και συν. [30] που ανέπτυξαν μια προσέγγιση για σύσταση ταινιών η οποία βασίζεται στην παραδοχή ότι τα γεγονότα που συμβαίνουν την πάροδο του χρόνου μπορεί να επηρεάσουν τις συστάσεις. Στο στάδιο της προ – επεξεργασίας ομαδοποίησαν τον αριθμό των αξιολογήσεων των ταινιών για δεδομένο χρήστη με βάση προκαθορισμένα χρονικά διαστήματα, όπως μία ημέρα, μία εβδομάδα ή έναν μήνα. Στη

συνέχεια, οι συγγραφείς χρησιμοποίησαν νευρωνικά δίκτυα για να καθορίσουν τις βαθμολογίες του χρήστη. Το νευρωνικό δίκτυο χρησιμοποιούσε τις ακόλουθες παραμέτρους εισόδου: (α) κατανομή του αριθμού των αξιολογήσεων των χρηστών για μια ταινία ανά εβδομάδα και ανά ημέρα, (β) ημερομηνία υποβολής αξιολόγησης, και (γ) το συνολικό αριθμό των αξιολογήσεων που δίνεται σε μια ταινία. Δοκιμάζοντας το νευρωνικό δίκτυο με training sets και testing sets σχεδόν ίδιου μεγέθους, η ακρίβεια της ταξινόμησης ήταν της τάξης του 71,9%.

2.2.1.6 Χρήση Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι μηχανές διανυσμάτων υποστήριξης (ή μηχανές διανυσματικής υποστήριξης) θεωρείται ως ο πιο επιτυχημένος αλγόριθμος κατηγοριοποίησης. Στόχος μιας μηχανής διανυσμάτων υποστήριξης (SVM) είναι να βρεθεί ένα γραμμικό υπερ-επίπεδο, το οποίο χωρίζει τα δεδομένα με τέτοιο τρόπο ώστε το περιθώριο είναι μεγιστοποιείται[1,13]. Για παράδειγμα, σε ένα πρόβλημα διαχωρισμού δύο κατηγοριών σε δύο διαστάσεις όπως αυτή που απεικονίζεται στο Σχήμα 2.3, παρατηρείται ότι υπάρχουν πολλά δυνατά όρια γραμμών για να διαχωριστούν οι δύο κλάσεις. Κάθε όριο έχει ένα συνδεδεμένο περιθώριο. Το σκεπτικό πίσω από τις SVM είναι ότι αν έχουμε επιλέξει αυτό που μεγιστοποιεί το περιθώριο είναι λιγότερο πιθανό να κατηγοριοποιηθεί λάθος ένα άγνωστο αντικείμενο στο μέλλον.



Εικόνα 2.3: Διαχωριστικό υπερεπίπεδο[31]

Ένας γραμμικός διαχωρισμός μεταξύ δύο τάξεων επιτυγχάνεται μέσω της συνάρτησης

$$W \,.\, x + \beta = 0. \qquad 2.7$$

Ορίζουμε μια συνάρτηση που μπορεί να ταξινομήσει τα στοιχεία της σε τιμές +1 ή -1 εφόσον διαχωρίζονται από κάποια ελάχιστη απόσταση από τη συνάρτηση διαχωρισμού κλάσης.. Η λειτουργία δίνεται από την εξίσωση:

$$f(x) = \begin{cases} +1, \ \mathrm{W}.\ \mathrm{x} + \ \mathrm{b} \ \geq 1 \\ -1, \ \mathrm{W}.\ \mathrm{x} + \ \mathrm{b} \ \leq \ -1 \end{cases} \qquad 2.8$$

με περιθώριο $= \frac{2}{||\mathrm{w}||2}$

Το πρόβλημα ανάγεται σε πρόβλημα μεγιστοποίησης το περιθωρίου μεταξύ των δύο κλάσεων, το οποίο είναι ισοδύναμο με το πρόβλημα ελαχιστοποίησης της αντίστροφη τιμής

$$L\ (w)\ = \ \frac{|w||2}{2}\ . \qquad 2.9$$

Εάν τα στοιχεία δεν είναι γραμμικά διαχωρίσιμα εισάγεται μία μεταβλητή χαλαρότητας και στην περίπτωση αυτή, ο τύπος για την ελαχιστοποίηση δίνεται από την εξίσωση 2.1 με τους περιορισμούς της 2.2. Από την άλλη πλευρά, εάν το υπερ-επίπεδο δεν είναι γραμμικό χρειάζεται μετασχηματισμός των δεδομένων σε ένα υψηλότερων διαστάσεων χώρο. Αυτό επιτυγχάνεται χάρη σε ένα μαθηματικό μετατροπή είναι γνωστή ως kernel trick. Η βασική ιδέα είναι να αντικαταστήσει τα γινόμενα στην εξίσωση 2.8 από μια συνάρτηση kernel. Υπάρχουν πολλές διαφορετικές πιθανές επιλογές για τη συνάρτηση kernel, όπως πολυωνυμική ή σιγμοειδής.

$$L\ (w)\ = \ \frac{|w||2}{2}\ + C\ \sum_{i=1}^{N} \varepsilon \qquad 2.10$$

$$f(x) = \begin{cases} +1, \ \mathrm{W}.\ \mathrm{x} + \ \mathrm{b} \ \geq 1 - \varepsilon \\ -1, \ \mathrm{W}.\ \mathrm{x} + \ \mathrm{b} \ \leq \ -1 + \varepsilon \end{cases} \qquad 2.11$$

Οι Μηχανές Διανυσμάτων Υποστήριξης έχουν αποκτήσει πρόσφατα δημοτικότητα εξαιτίας της απόδοσή τους και της αποτελεσματικότητας τους σε πολλές περιπτώσεις και έχουν επίσης δείξει πρόσφατα πολλά υποσχόμενα αποτελέσματα στα συστήματα συστάσεων.

2.2.2 Ανάλυση με χρήση συσταδοποίησης (Data Analysis and Clustering)

Το κύριο πρόβλημα για την υλοποίηση κατηγοριοποίησης συνεργατικού φιλτραρίσματος είναι ο αριθμός των λειτουργιών για τον υπολογισμός αποστάσεων – για παράδειγμα για την εύρεση των καλύτερων k-πλησιέστερων γειτόνων. Μια πιθανή λύση είναι, όπως είδαμε στην ενότητα 2.1.2, η μείωση των διαστάσεων. Αλλά, ακόμα και με τη μείωση της διάστασης του χαρακτηριστικών, υπάρχουν πολλά αντικείμενα που χρειάζονται υπολογισμό της απόστασης τους. Το πρόβλημα αυτό αντιμετωπίζουν οι αλγόριθμοι ανάλυσης συστάδων. Το ίδιο ισχύει και για συστήματα που βασίζονται στο περιεχόμενο, όπου απαιτούνται οι αποστάσεις μεταξύ των

αντικειμένων για την ανάκτηση παρόμοιων. Η συσταδοποίηση έχει αποδειχθεί ότι βελτιώνει την αποδοτικότητα αφού μειώνεται το υπολογιστικό κόστος, ωστόσο, σε αντίθεση με τις μεθόδους μείωσης διάστασης, δεν συμβάλει στη βελτίωση της ακρίβειας. Ως εκ τούτου, η ομαδοποίηση θα πρέπει να εφαρμόζονται με προσοχή κατά το σχεδιασμό ενός συστήματος σύστασης λαμβάνοντας υπόψη τόσο τη βελτίωση της αποτελεσματικότητας αλλά και την πιθανή μείωση της ακρίβειας.

Η ομαδοποίηση, που αναφέρεται επίσης ως μη επιβλεπόμενη μάθηση, συνίσταται στην απόδοση στοιχείων σε ομάδες, έτσι ώστε τα στοιχεία στις ίδιες ομάδες να μοιάζουν περισσότερο από ό,τι στοιχεία σε διαφορετικές ομάδες: ο στόχος είναι να βρεθούν οι φυσικές (ή ουσιαστικές) ομάδες που υπάρχουν στα δεδομένα. Η ομοιότητα καθορίζεται χρησιμοποιώντας ένα μέτρο απόστασης, όπως αυτά που αναφέρονται στην ενότητα 2.2.1. Ο στόχος του αλγόριθμου ομαδοποίησης είναι να ελαχιστοποιήσει τις αποστάσεις εντός της συστάδας μεγιστοποιώντας παράλληλα τις αποστάσεις μεταξύ των συστάδων. Υπάρχουν δύο κύριες κατηγορίες αλγορίθμων ομαδοποίησης: ιεραρχική και διαχωριστική. Οι αλγόριθμοι διαχωριστικής ομαδοποίησης χωρίζουν τα στοιχεία δεδομένων σε μη επικαλυπτόμενες ομάδες έτσι ώστε κάθε στοιχείο να ανήκει ακριβώς σε μία ομάδα. Οι αλγόριθμοι ιεραρχικής ομαδοποίησης ομαδοποιούν διαδοχικά αντικείμενα μέσα σε συστάδες, παράγοντας ένα σύνολο από εμφωλευμένες συστάδες που είναι οργανωμένες σε ένα ιεραρχικό δέντρο.

Πολλοί αλγόριθμοι ομαδοποίησης προσπαθούν να ελαχιστοποιήσουν μια συνάρτηση που μετρά την ποιότητα της ομαδοποίησης. Μια τέτοια συνάρτηση ποιότητας συχνά αναφέρεται ως αντικειμενική συνάρτηση. Έτσι, η ομαδοποίηση μπορεί να θεωρηθεί ως ένα πρόβλημα βελτιστοποίησης: ο ιδανικός αλγόριθμος ομαδοποίησης θα εξετάσει όλες τις δυνατές διαμερίσεις των δεδομένων εξόδου και θα έχει ως έξοδο αυτή που ελαχιστοποιεί τη συνάρτηση ποιότητας. Αλλά το αντίστοιχο πρόβλημα βελτιστοποίησης είναι NP δύσκολο και συνεπώς η εύρεση βέλτιστων λύσεων συχνά είναι αδύνατη. Για τον ίδιο λόγο, η επιλογή του συγκεκριμένου αλγόριθμου ομαδοποίησης και των παραμέτρων του (π.χ., μέτρο ομοιότητας) εξαρτάται από πολλούς παράγοντες, συμπεριλαμβανομένων και των χαρακτηριστικών των δεδομένων. Παρακάτω παρουσιάζονται μερικοί από τους αλγόριθμους.

2.2.2.1 Αλγόριθμος k – Means

Η k-means ομαδοποίηση είναι μια διαχωριστική μέθοδος. Η συνάρτηση διαχωρίζει το σύνολο N αντικείμενων σε k ανεξάρτητα S_j υποσύνολα που περιέχουν N_j αντικείμενα έτσι ώστε να είναι όσο το δυνατόν πιο κοντά μεταξύ τους όσο ορίζει ένα δεδομένο μέτρο απόστασης. Κάθε ομάδα της διαμέρισης ορίζεται από τα N_j μέλη της και το λ_j κέντρο βάρους της. Το κέντρο βάρους για κάθε σύμπλεγμα είναι το σημείο στο οποίο το άθροισμα των αποστάσεων από όλα τα αντικείμενα αυτής της ομάδας ελαχιστοποιείται. Δεδομένης μιας σειράς από παρατηρήσεις $(x_1, x_2, \ldots, x_n)$, όπου κάθε παρατήρηση είναι ένα d-διαστάσεων πραγματικό διάνυσμα, ο αλγόριθμος k-means έχει ως στόχο να διαμερίσει τις n παρατηρήσεις σε k ($\leq n$) σύνολα $S =$

$\{S_1, S_2, \ldots, S_k\}$ έτσι ώστε να ελαχιστοποιηθεί το άθροισμα των τετραγώνων εντός συστάδας. Με άλλα λόγια, ο στόχος του είναι να βρει:

$$\sum_{i=1}^{k} \sum_{x \in S_j} \|x - \mu_i\|^2 \qquad 2.12$$

όπου, μ_i η μέση τιμή των σημείων στο S_i.

Ο αλγόριθμος λειτουργεί με τυχαία επιλογή κέντρων βάρους k. Στη συνέχεια, τα στοιχεία ομαδοποιούνται στο σύμπλεγμα του οποίου το κέντρο βάρους είναι το πιο κοντινό σε αυτά. Το νέο κέντρο βάρους του συμπλέγματος πρέπει να τροποποιηθεί με την πρόσθεση ή αφαίρεση στοιχείων από το σύμπλεγμα. Αυτή η λειτουργία συνεχίζεται μέχρι να μην υπάρχουν περαιτέρω στοιχεία που αλλάζουν την ομάδα τους. Για συνηθισμένα μέτρα ομοιότητας, ο αλγόριθμος συγκλίνει και η σύγκλιση συμβαίνει συνήθως τις αρχικές πρώτες επαναλήψεις. Ως εκ τούτου, συχνά η τελική συνθήκη αλλάζει σε «μέχρι σχετικά λίγα σημεία να αλλάζουν συστάδα» ή «η απόσταση μεταξύ των νέων κεντρικών σημείων από τα παλιά να είναι μικρή», προκειμένου να βελτιωθεί η αποτελεσματικότητα.

Ο βασικός k-means αλγόριθμος είναι ένας εξαιρετικά απλός και αποτελεσματικός αλγόριθμος. Ωστόσο, έχει πολλά μειονεκτήματα: (1) υποθέτει προηγούμενη γνώση των δεδομένων προκειμένου να επιλέξετε το κατάλληλο k (2) οι τελικές συστάδες είναι πολύ ευαίσθητες στην επιλογή των αρχικών σημείων και (3), μπορεί να παράγει άδειο σύμπλεγμα. Ο αλγόριθμος έχει επίσης διάφορους περιορισμούς όσον αφορά τα στοιχεία: έχει προβλήματα όταν οι συστάδες έχουν διαφορετικά μεγέθη, πυκνότητες, και μη-σφαιρικά σχήματα και έχει επίσης προβλήματα όταν τα δεδομένα περιέχουν ακραίες τιμές.

Οι Xue et al. [32] παρουσιάζουν μία τυπική χρήση της ομαδοποίησης στο πλαίσιο των συστημάτων σύστασης με τη χρήση του k-means αλγόριθμου ως ένα στάδιο προ-επεξεργασίας για το σχηματισμό «γειτονιών». Δεν περιορίζουν τη «γειτονιά» στο σύμπλεγμα που ανήκει ο χρήστης, αλλά χρησιμοποιούν την απόσταση από το χρήστη σε διαφορετικά κέντρα βάρους του συμπλέγματος ως βήμα προεπιλογής γειτόνων. Επίσης εφαρμόζουν μια τεχνική που βασίζεται στην εξομάλυνση των συστάδων με την οποία οι τιμές που λείπουν για τους χρήστες σε ένα σύμπλεγμα αντικαθίστανται από αντιπροσωπευτικές τιμές συστάδων. Η μέθοδός τους φέρεται να έχει ελαφρώς καλύτερη απόδοση από μεθόδους με βάση το kNN . Με παρόμοιο τρόπο, οι Sarwar et al. [33] περιγράφουν μια προσέγγιση για την υλοποίηση ενός κλιμακούμενου KNN ταξινομητή. Ο διαμοιρασμός του χώρου του χρήστη γίνεται με την εφαρμογή του *bisecting* k-means αλγόριθμο και στη συνέχεια χρησιμοποιούν αυτές τις συστάδες ως βάση για τον σχηματισμό της «γειτονιάς». Η προσέγγισή τους επιτρέπει μια σημαντική βελτίωση στην αποδοτικότητα.

2.2.2.2 Ομαδοποίηση που βασίζεται στην πυκνότητα (Density –based clustering)

Κατά την ομαδοποίηση που βασίζεται στην πυκνότητα ως συστάδες ορίζονται οι περιοχές με υψηλότερη πυκνότητα από τις υπόλοιπες ομάδες δεδομένων. Τα αντικείμενα σε αυτές τις αραιές περιοχές - που απαιτούνται για να διαχωριστούν οι συστάδες - συνήθως θεωρούνται θόρυβος και συνοριακά σημεία. Η πιο δημοφιλής μέθοδος ομαδοποίησης που βασίζεται στην πυκνότητα είναι η DBSCAN (Density-based spatial clustering of applications with noise). Σε αντίθεση με πολλές νεότερες μεθόδους, διαθέτει ένα καλά καθορισμένο μοντέλο συμπλέγματος που ονομάζεται "πυκνότητα-προσβασιμότητας» (density-reachability). Όμοια με την ομαδοποίηση με βάση τη σύνδεση, βασίζεται σε σημεία σύνδεσης εντός ορισμένων ορίων απόστασης. Ωστόσο, συνδέει μόνο σημεία που ικανοποιούν ένα κριτήριο πυκνότητας, στην αρχική παραλλαγή ορίζεται ως ένας ελάχιστος αριθμός αντικειμένων στο εσωτερικό αυτής της ακτίνας. Ένα σύμπλεγμα αποτελείται από όλα συνδεδεμένα αντικείμενα (τα οποία μπορούν να σχηματίσουν ένα σύμπλεγμα ενός αυθαίρετου σχήματος, σε αντίθεση με πολλές άλλες μεθόδους) συν όλα τα αντικείμενα που βρίσκονται εντός εμβέλειας αυτών των αντικειμένων. Μια άλλη ενδιαφέρουσα ιδιότητα της DBSCAN είναι ότι η πολυπλοκότητά της είναι αρκετά χαμηλή - απαιτεί μια γραμμική σειρά αναζητήσεων στη βάση δεδομένων - και ότι έχει ουσιαστικά τα ίδια αποτελέσματα σε κάθε εκτέλεση, ως εκ τούτου, δεν υπάρχει καμία ανάγκη να εκτελεστεί πολλές φορές. Η μέθοδος Optics είναι μια γενίκευση της DBSCAN που καταργεί την ανάγκη για επιλογή κατάλληλης τιμής για την παράμετρο ε , και παράγει ένα ιεραρχικό αποτέλεσμα που σχετίζεται με αυτό της ομαδοποίησης με σύνδεση. Η μέθοδος Deli-Clu, Πυκνότητα-Σύνδεση-Ομαδοποίηση (Density – Link – Clustering) συνδυάζει ιδέες από την ομαδοποίηση απλής σύνδεσης και την Optics, εξαλείφοντας την παράμετρο ε εξ ολοκλήρου και προσφέρει βελτιωμένες επιδόσεις από την Optics.

Το βασικό μειονέκτημα των DBSCAN και Optics είναι ότι αναμένουν κάποια μείωση της πυκνότητας για την ανίχνευση των συνόρων της συστάδας. Επιπλέον, δεν μπορούν να ανιχνεύσουν τις εγγενείς δομές ομάδων που είναι διαδεδομένες στην πλειονότητα των δεδομένων πραγματικής ζωής. Μια παραλλαγή της DBSCAN, η EnDBSCAN, ανιχνεύει αποτελεσματικά τέτοιου είδους δομές. Σε σύνολα δεδομένων, για παράδειγμα, η επικάλυψη Gaussian κατανομών - μια συνηθισμένη περίπτωση σε τεχνητά δεδομένα - τα σύνορα των ομάδων που παράγονται από αυτούς τους αλγορίθμους συχνά θα φαίνονται αυθαίρετα, επειδή η πυκνότητα του συμπλέγματος μειώνεται συνεχώς. Σε ένα σύνολο δεδομένων που αποτελούνται από διάφορες Gaussian κατανομές, αυτοί οι αλγόριθμοι είναι σχεδόν πάντα λιγότερο αποδοτικοί από μεθόδους όπως η EM ομαδοποίηση (EM clustering) που είναι σε θέση να διαμορφώσει ακριβώς αυτό το είδος των δεδομένων.

2.2.2.3 Ομαδοποίηση που βασίζεται στη σύνδεση (Ιεραρχική ομαδοποίηση)

Η ομαδοποίηση που βασίζεται στη σύνδεση, επίσης γνωστή ως ιεραρχική ομαδοποίηση, βασίζεται στην κεντρική ιδέα πως τα αντικείμενα σχετίζονται πιο πολύ με τα αντικείμενα που είναι κοντά τους παρά με αυτά που είναι πιο μακριά. Αυτοί οι αλγόριθμοι συνδέουν "αντικείμενα" για να σχηματίσουν "συστάδες", με βάση την απόστασή τους. Ένα σύμπλεγμα μπορεί να περιγραφεί σε μεγάλο βαθμό από τη μέγιστη απόσταση που απαιτείται για τη σύνδεση τμημάτων του συμπλέγματος. Σε διαφορετικές αποστάσεις, θα σχηματιστούν διαφορετικές ομάδες, γεγονός που μπορεί να αναπαρασταθεί χρησιμοποιώντας ένα δενδρόγραμμα, που εξηγεί και την κοινή ονομασία «ιεραρχική ομαδοποίηση». Αυτοί οι αλγόριθμοι δεν παρέχουν μια ενιαία τμηματοποίηση του συνόλου των δεδομένων, αλλά αντίθετα προσφέρουν μια εκτεταμένη ιεραρχία συστάδων που συγχωνεύονται μεταξύ τους σε ορισμένες αποστάσεις. Σε ένα δενδρόγραμμα, ο άξονας y αναπαριστά την απόσταση στην οποία συγχωνεύονται οι συστάδες, ενώ τα αντικείμενα τοποθετούνται κατά μήκος του άξονα x, έτσι ώστε να μην αναμειγνύονται οι συστάδες.

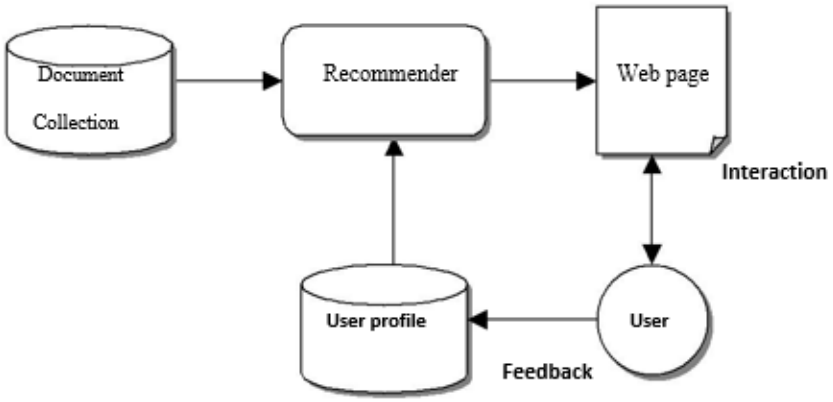
Η τεχνική αυτή ομαδοποίησης είναι ένα ολόκληρο σύνολο μεθόδων που διαφέρουν ως προς τον τρόπο που υπολογίζονται οι αποστάσεις. Εκτός από τη συνήθη επιλογή των συναρτήσεων για εύρεση της απόστασης, ο χρήστης πρέπει επίσης να αποφασίσει σχετικά με το κριτήριο σύνδεσης που θα χρησιμοποιηθεί (εφόσον ένα σύμπλεγμα αποτελείται από πολλά αντικείμενα, υπάρχουν πολλαπλοί τρόποι για να υπολογιστεί η απόσταση). Δημοφιλείς επιλογές είναι η ομαδοποίηση απλής σύνδεσης (η ελάχιστη απόσταση των αντικειμένων), η ομαδοποίηση πλήρης σύνδεσης (τη μέγιστη απόσταση των αντικειμένων) ή ομαδοποίηση μέσης σύνδεσης.

Αυτές οι μέθοδοι δεν παράγουν μια μοναδική κατάτμηση του συνόλου των δεδομένων, αλλά μια ιεραρχία από την οποία ο χρήστης εξακολουθεί να πρέπει να επιλέξει τα κατάλληλα συμπλέγματα. Δεν είναι πολύ ισχυροί έναντι των ακραίων τιμών, οι οποίες είτε θα εμφανίζονται ως πρόσθετοι πόλοι ή ακόμη και να προκαλέσουν συγχώνευση άλλων συστάδων. Στα πλαίσια της εξόρυξης δεδομένων αυτές οι μέθοδοι θεωρούνται ως θεωρητική θεμελίωση της ανάλυσης συστάδας, αλλά συχνά θεωρούνται απαρχαιωμένες. Αποτέλεσαν πηγή έμπνευσης όμως για πολλές μεταγενέστερες μεθόδους όπως η ομαδοποίηση με βάση την πυκνότητα.

3. Αλγόριθμοι συστημάτων σύστασης (Filtering Methods)

Τα συστήματα συστάσεων όπως αναφέρθηκε χρησιμοποιούν αρκετές από τις μεθόδους εξόρυξης δεδομένων για προ-επεξεργασία και ανάλυση δεδομένων. Γενικά, κάθε σύστημα σύστασης ακολουθεί μία συγκεκριμένη διαδικασία για την παραγωγή συστάσεων. Αρχικά το σύστημα θα πρέπει να επεξεργαστεί τα δεδομένα εισόδου προκειμένου να παράγει αποτελέσματα, τα οποία μπορούν να χωριστούν σε τρεις διαφορετικές πηγές όπως φαίνεται και στο Σχήμα 3.1 :

- Δεδομένα χρήστη
- Δεδομένα αντικειμένου
- Πιθανές αλληλεπιδράσεις ενός χρήστη με το αντικείμενο (βλ. Κεφάλαιο 4)



Εικόνα 3.1: Διαδικασία συστήματος σύστασης

Τα δημογραφικά στοιχεία για το χρήστη όπως το φύλο, η ηλικία κ.λπ. αλλά και οι γνώσεις σχετικά με τις προτιμήσεις του συνθέτουν ένα *"προφίλ χρήστη"*. Παράλληλα, κάθε αντικείμενο περιγράφεται με τη βοήθεια χαρακτηριστικών γνωρισμάτων, τα οποία συμβάλλουν στη διαμόρφωση ενός προφίλ αντικειμένου, που ονομάζεται *"περιεχόμενο"*. Η καταγραφή των δεδομένων του χρήστη μπορεί να γίνει είτε *άμεσα* (explicitly) είτε *έμμεσα* (implicitly). Καταγραφή δεδομένων άμεσα σημαίνει ότι ο χρήστης αλληλεπιδρά με το σύστημα ειδικά για το σκοπό της παροχής των απαραίτητων πληροφοριών σε αυτό. Για παράδειγμα, τα δημογραφικά χαρακτηριστικά του χρήστη καταγράφονται άμεσα όταν ζητούνται κατά τη διάρκεια της εγγραφής του σε ένα ηλεκτρονικό κατάστημα. Αντίθετα, καταγραφή δεδομένων έμμεσα σημαίνει ότι το σύστημα λαμβάνει τα δεδομένα ενώ ο χρήστης αλληλεπιδρά με αυτό για άλλο σκοπό.

Τα συστήματα συστάσεων αποτελούν μία τεχνική φιλτραρίσματος των παραπάνω πληροφοριών και μπορούν να χωριστούν σε 5 κατηγορίες οι οποίες αναλύονται στις επόμενες

ενότητες: συνεργατικό φιλτράρισμα, φιλτράρισμα με βάση το περιεχόμενο, δημογραφικό, φιλτράρισμα με βάση τη γνώση και υβριδικό.

3.1 Συνεργατικό φιλτράρισμα (Collaborative Filtering)

Το συνεργατικό φιλτράρισμα είναι ένας δημοφιλής αλγόριθμος που βασίζει τις προβλέψεις και τις συστάσεις του στη συμπεριφορά και τη βαθμολόγηση αντικειμένων από άλλους χρήστες του συστήματος. Η βασική υπόθεση της μεθόδου αυτής είναι πως οι γνώμες των άλλων χρηστών μπορούν να συγκεντρωθούν και να οργανωθούν με τρόπο που θα παρέχει μία λογική υπόθεση για την προτίμηση του ενεργού χρήστη. Η πλειονότητα των μεθόδων συνεργατικού φιλτραρίσματος λειτουργούν δημιουργώντας προβλέψεις για την προτίμηση του χρήστη και έπειτα παράγουν συστάσεις βαθμολογώντας υποψήφια αντικείμενα από τις εκτιμώμενες προτιμήσεις. Αυτή η μέθοδος ανήκει στην κατηγορία συνεργατικού φιλτραρίσματος με βάση το χρήστη. Χαρακτηριστικά, για κάθε χρήστη βρίσκεται ένα σύνολο «πλησιέστερων χρηστών γειτόνων» με των οποίων τις μέχρι τώρα εκτιμήσεις υπάρχει ο ισχυρότερος συσχετισμός. Τα αποτελέσματα για τα άγνωστα στοιχεία προβλέπονται με βάση συνδυασμό αποτελεσμάτων που είναι γνωστά από τους «πλησιέστερους γείτονες». Το σύστημα μπορεί να προτείνει αντικείμενα (όπως βιβλία, μουσική κ.λπ.) στους χρήστες βασισμένο στις εκτιμήσεις των στοιχείων, αντί του περιεχομένου των στοιχείων, γεγονός που μπορεί να βελτιώσει την ποιότητα των συστάσεων. Εναλλακτικά, υπάρχει και το φιλτράρισμα με βάση το αντικείμενο, π.χ. ο χρήστης που αγόρασε το χ προϊόν είναι πιθανό να αγοράσει και το ψ. Με τη μέθοδο αυτή δημιουργείται ένας πίνακας αντικείμενο-αντικείμενο και καθορίζει σχέσεις μεταξύ των αντικειμένων δημιουργώντας ζευγάρια από αυτά. Στη συνέχεια, συνάγει τις προτιμήσεις του ενεργού χρήστη εξετάζοντας τον πίνακα και ταιριάζοντας τις πληροφορίες για το χρήστη. Το συνεργατικό φιλτράρισμα χρησιμοποιεί τρεις κύριες τεχνικές: την τεχνική που βασίζεται στη μνήμη (memory-based), την τεχνική που βασίζεται στο μοντέλο (model-based) και την υβριδική.

Η τεχνική που βασίζεται στη μνήμη χρησιμοποιεί τις πληροφορίες βαθμολόγησης των χρηστών για να υπολογίσει την ομοιότητα μεταξύ χρηστών ή αντικειμένων και να δημιουργήσει τις συστάσεις της. Αυτή ήταν μία αρχική προσέγγιση που χρησιμοποιείται ευρέως σε εμπορικά συστήματα, λόγω της αποτελεσματικότητας και της ευκολίας στην εφαρμογή της. Τυπικά παραδείγματα αυτής της προσέγγισης είναι ο «πλησιέστερος γείτονας» και η κορυφαίες N συστάσεις βασισμένες στο χρήστη ή στο αντικείμενο. Για παράδειγμα, σε προσεγγίσεις με βάση το χρήστη, η βαθμολόγηση το χρήστη $'u'$ για ένα αντικείμενο $'i'$. υπολογίζεται ως μία συσσωμάτωση (aggregation) κάποιων αξιολογήσεων παρόμοιων χρηστών για το αντικείμενο αυτό:

$$r_{u,i} = \text{aggr}_{u' \in U} r_{u',i},$$

όπου U το σύνολο των κορυφαίων N χρηστών που είναι περισσότερο όμοιοι με τη χρήστη $'u'$ που βαθμολόγησε το αντικείμενο $'i'$.

Ο αλγόριθμος «πλησιέστερου γείτονα» υπολογίζει την ομοιότητα μεταξύ δύο χρηστών ή αντικειμένων και παράγει μία εκτίμηση για το χρήστη λαμβάνοντας υπόψη το σταθμισμένο μέσο όρο όλων των αξιολογήσεων. Ο υπολογισμός της ομοιότητας αποτελεί πολύ σημαντικό μέρος της μεθόδου αυτής, με κυριότερες συναρτήσεις υπολογισμού της : συσχέτιση Pearson και ομοιότητα συνημίτονου, όπως αναφέρθηκαν σε προηγούμενη ενότητα. Στη συσχέτιση Pearson η ομοιότητα δύο χρηστών x, y υπολογίζεται:

$$\text{simil}(x,y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})(r_{y,i} - \bar{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})^2 \sum_{i \in I_{xy}} (r_{y,i} - \bar{r_y})^2}}, \qquad 3.1$$

όπου I_{xy} το σύνολο αντικειμένων που έχουν βαθμολογηθεί και από τους δύο χρήστες και $r_{x,i}$ η βαθμολογία του χρήστη x για το αντικείμενο i. Στην ομοιότητα συνημίτονου η ομοιότητα των δύο χρηστών είναι:

$$\text{simil}(x,y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{i \in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2} \sqrt{\sum_{i \in I_y} r_{y,i}^2}}. \qquad 3.2$$

Ο αλγόριθμος που βασίζεται στις κορυφαίες N προτάσεις χρησιμοποιεί ένα διανυσματικό μοντέλο ομοιοτήτων για να προσδιορίσει τους N περισσότερο όμοιους χρήστες με το ζητούμενο χρήστη. Μόλις προσδιοριστούν οι χρήστες αυτοί, συγκεντρώνονται οι αντίστοιχοι πίνακες χρήστη-αντικειμένου για να βρεθεί το σύνολο των αντικειμένων που θα προταθούν στο χρήστη.

Τα πλεονεκτήματα της προσέγγισης που βασίζεται στη μνήμη είναι μεταξύ άλλων: η επεξηγηματικότητα των αποτελεσμάτων, η οποία είναι μία πολύ σημαντική πλευρά των συστημάτων συστάσεων, η ευκολία στη χρήση και στην εφαρμογή της καθώς και η ανεξαρτησία της από το περιεχόμενο του αντικειμένου που προτείνεται. Αντίθετα, το κυριότερο μειονέκτημα της είναι η μειωμένη επίδοση όταν τα δεδομένα που έχει στη διάθεσή της είναι λιγοστά. Εάν ένα νέο στοιχείο εμφανιστεί στη βάση δεδομένων, δεν υπάρχει κανένας τρόπος να συστηθεί σε έναν χρήστη έως ότου να ληφθούν περισσότερες πληροφορίες για αυτό μέσω μιας άλλης εκτίμησης χρήστη είτε διευκρινίζοντας ποια άλλα στοιχεία είναι παρόμοια με αυτό. Το πρόβλημα είναι γνωστό με την ονομασία «ψυχρή εκκίνηση» (cold-start) ή πρόβλημα της «πρώτης εκτίμησης» (first-rater), καθώς συστάσεις απαιτούνται για τα στοιχεία που κανένας χρήστης δεν έχει εκτιμήσει ακόμα.

Η τεχνική που βασίζεται στο μοντέλο χρησιμοποιεί τις αξιολογήσεις των χρηστών ως σύνολο εκπαίδευσης αλγόριθμων μηχανικής εκμάθησης για τη δημιουργία ενός μοντέλου πρόβλεψης. Στη συνέχεια, το μοντέλο χρησιμοποιείται για την πρόβλεψη της βαθμολογίας για ένα αντικείμενο. Υπάρχουν πολλοί τρόποι για πρόβλεψη πραγματικών δεδομένων, οι σημαντικότεροι από τους οποίους είναι η συσταδοποίηση και οι πιθανοτικοί αλγόριθμοι. Η συσταδοποίηση

δημιουργεί ομάδες από χρήστες (user-based) με παρόμοιες προτιμήσεις ή ομάδες αντικειμένων (item-based) με παρόμοιο περιεχόμενο. Η δημιουργία των συστάδων βασίζεται στις αξιολογήσεις που έχουν δώσει οι χρήστες στα αντικείμενα. Στο [34], οι συγγραφείς χρησιμοποιούν την προσέγγιση της σκληρής συσταδοποίησης ("hardclustering") για να κατατάξουν το χρήστη σε μία συστάδα και η πρόβλεψη γίνεται σύμφωνα με τις βαθμολογίες των χρηστών στη συγκεκριμένη συστάδα. Υπάρχουν και πιο ασαφείς (fuzzy) προσεγγίσεις που υπολογίζουν τη πιθανότητα ένας χρήστης να ανήκει σε κάποια κλάση και ύστερα εκτιμούν την αξιολόγησή του για ένα αντικείμενο. Οι πιθανοτικοί αλγόριθμοι, όπως τα Bayesian δίκτυα, που αναφέρθηκαν σε προηγούμενη ενότητα, δημιουργούν ένα πιθανοτικό μοντέλο για να λύσουν το πρόβλημα του συνεργατικού φιλτραρίσματος. Η τεχνική που βασίζεται σε μοντέλα αντιμετωπίζει το πρόβλημα των λιγοστών δεδομένων καλύτερα από τη memory based τεχνική, λύνει το πρόβλημα της κλιμάκωσης με μεγάλα σύνολα δεδομένων και βελτιώνει την απόδοση της πρόβλεψης. Στα μειονεκτήματά της συγκαταλέγονται η δυσκολία του χτισίματος του μοντέλου και της εξήγησης των προβλέψεων. Όπως σε όλες τις μεθόδους πρέπει να βρεθεί η κατάλληλη ισορροπία μεταξύ της απόδοσης της πρόβλεψης και της κλιμακωσιμότητας.

Τέλος, αρκετές εφαρμογές χρησιμοποιούν ένα συνδυασμό των τεχνικών που βασίζονται στη μνήμη και των τεχνικών που βασίζονται στο μοντέλο, που καλείται υβριδική τεχνική. Ένα υβριδικό σύστημα που συνδυάζει τις τεχνικές Α και Β επιχειρεί να χρησιμοποιήσει τα πλεονεκτήματα της Α για να διορθώσει τα μειονεκτήματα της Β. Επομένως, προβλήματα όπως αυτό της έλλειψης πληροφοριών μπορούν να ξεπεραστούν και να βελτιωθεί έτσι η απόδοση του συστήματος όμως αυξάνεται και η πολυπλοκότητα του. Συνήθως τα περισσότερα εμπορικά συστήματα σύστασης είναι υβριδικά όπως αυτό που χρησιμοποιεί η Google.

3.2 Φιλτράρισμα με βάση το περιεχόμενο (Content–based Filtering)

Τα συστήματα που χρησιμοποιούν φιλτράρισμα με βάση το περιεχόμενο αναλύουν ένα σύνολο από πληροφορίες και περιγραφές για χαρακτηριστικά των αντικειμένων που έχουν βαθμολογηθεί προηγουμένως από τους χρήστες[1,13]. Στη συνέχεια χτίζουν ένα μοντέλο ή προφίλ των ενδιαφερόντων του χρήστη που βασίζεται στα χαρακτηριστικά αυτά. Το προφίλ αυτό είναι μία δομημένη παρουσίαση των προτιμήσεων του χρήστη και προσαρμόζεται ανάλογα με τα νέα αντικείμενα για τα οποία δείχνει ενδιαφέρον ο χρήστης. Η διαδικασία σύστασης συνίσταται βασικά στην αντιστοίχιση των χαρακτηριστικών του προφίλ των χρηστών έναντι των χαρακτηριστικών ενός αντικειμένου. Το αποτέλεσμα είναι μία σχετική απόφαση για το επίπεδο του ενδιαφέροντος του χρήστη για το αντικείμενο αυτό. Εάν ένα προφίλ αντανακλά με ακρίβεια τις προτιμήσεις του χρήστη, αποτελεί ένα μεγάλο πλεονέκτημα για την αποτελεσματικότητα της διαδικασίας πρόσβασης δεδομένων. Τυπικά ένα αντικείμενο περιγράφεται ως ένα διάνυσμα $X = (x_1, x_2, \dots, x_n)$ από n στοιχεία. Τα στοιχεία μπορούν να έχουν δυαδικά, ονομαστικά ή αριθμητικά χαρακτηριστικά και προέρχονται είτε από το περιεχόμενο των αντικειμένων είτε από πληροφορίες για τις προτιμήσεις των χρηστών . Στόχος της μεθόδου είναι να επιλέξει μία συνάρτηση βάση του

συνόλου m διανυσμάτων εισόδου που μπορεί να χαρακτηρίσει οποιοδήποτε στοιχείο της συλλογής. Η συνάρτηση $h(X)$ θα είναι σε θέση είτε να χαρακτηρίσει ένα αντικείμενο που δεν έχει προβληθεί ακόμα ως θετικό ή αρνητικό επιστρέφοντας μία δυαδική τιμή είτε να επιστρέψει μία αριθμητική τιμή. Σε αυτή τη περίπτωση μπορεί να χρησιμοποιηθεί ένα όριο για να προσδιοριστεί η σχετικότητα ή μη του αντικειμένου για το χρήστη.

Ένα σύστημα φιλτραρίσματος με βάση το περιεχόμενο επιλέγει αντικείμενα με βάση τη συσχέτιση μεταξύ του περιεχομένου των αντικειμένων και των προτιμήσεων των χρηστών σε αντίθεση με το συνεργατικό φιλτράρισμα που επιλέγει στοιχεία με βάση τη συσχέτιση μεταξύ χρηστών. Οι προτιμήσεις των χρηστών για διάφορα αντικείμενα μπορούν να προσδιοριστούν με χρήση άμεσης ή έμμεσης ανατροφοδότησης. Η άμεση ανατροφοδότηση είναι η ρητή βαθμολόγηση ενός αντικειμένου σε κάποια κλίμακα, ενώ για την έμμεση οι προτιμήσεις του χρήστη συνάγονται από την παρατήρηση των ενεργειών του.

Υπάρχουν αρκετοί τρόποι με τους οποίους μπορούν να αναπαρασταθούν οι όροι που χρησιμοποιούνται ως βάση της τεχνικής με βάση το περιεχόμενο. Μία μέθοδος αναπαράστασης που χρησιμοποιείται συχνά είναι το μοντέλο διανυσματικού χώρου. Στο μοντέλο αυτό το αντικείμενο D αναπαρίσταται ως διάνυσμα m διαστάσεων, όπου κάθε διάσταση αντιστοιχεί σε ένα διακριτό όρο και m είναι ο συνολικός αριθμός όρων που χρησιμοποιούνται στη συλλογή των αντικειμένων. Όρος συνήθως θεωρείται μία λέξη, λέξη-κλειδί ή μεγαλύτερες εκφράσεις ανάλογα με την εφαρμογή. Ένα αρχείο, για παράδειγμα, μπορεί να αναπαρασταθεί ως διάνυσμα σταθμισμένων όρων. Εάν το αρχείο D δεν περιέχει τον όρο j τότε το βάρος του j είναι μηδέν. Τα βάρη των όρων μπορούν να προσδιοριστούν και με το σχήμα συχνότητα όρου – αντίστροφη συχνότητα εγγράφου ($tf - idf$)[35], σύμφωνα με το οποίο δίνεται τιμή στα βάρη ανάλογα με τη συχνότητα εμφάνισης του όρου σε ένα αρχείο και πόσο συχνά συμβαίνει σε όλη τη συλλογή αρχείων. Το προφίλ του χρήστη μπορεί επίσης να αναπαρασταθεί με διάνυσμα P και ο βαθμός ομοιότητας μεταξύ ενός προφίλ και ενός αρχείου μπορεί να προσδιοριστεί με τη χρήση συνημίτονου. Το παρακάτω σχήμα αναπαριστά την αρχιτεκτονική ενός εξατομικευμένου συστήματος σύστασης. Το προφίλ του χρήστη ανανεώνεται συνεχώς από την ανατροφοδότηση που δίνεται, ενώ το σύστημα σύστασης συγκρίνει το προφίλ του χρήστη με τα αρχεία της συλλογής. Τα αρχεία έπειτα βαθμολογούνται με βάση συγκεκριμένα κριτήρια και τα αρχεία με την καλύτερη βαθμολογία εμφανίζονται τελικά στη τρέχουσα ιστοσελίδα.

Η τεχνική του φιλτραρίσματος με βάση το περιεχόμενο έχει αρκετά πλεονεκτήματα συγκριτικά με το συνεργατικό φιλτράρισμα:

Ανεξαρτησία χρήστη. Τα συστήματα φιλτραρίσματος με βάση το περιεχόμενο εκμεταλλεύονται μόνο τις αξιολογήσεις που παρέχει ο ενεργός χρήστης για να δημιουργήσουν το προφίλ του. Αντίθετα, τα συνεργατικά συστήματα που χρησιμοποιούν τις αξιολογήσεις άλλων χρηστών για να βρουν τον «πλησιέστερο γείτονα» και προτείνουν τελικά στο χρήστη αντικείμενα που προτίμησαν οι «γείτονες».

Διαφάνεια. Εξηγήσεις για το πώς λειτουργεί το σύστημα συστάσεων και ποιοι λόγοι οδήγησαν το συγκεκριμένο αντικείμενο στη λίστα των συστάσεων μπορούν να παρέχονται από την καταχώρηση των χαρακτηριστικών ή περιγραφών ενός αντικειμένου. Αυτά τα χαρακτηριστικά

είναι δείκτες που μπορεί να συμβουλευτεί κάποιος προκειμένου να αποφασίσει εάν πρέπει να εμπιστευθεί μια σύσταση. Αντίθετα, τα συνεργατικά συστήματα είναι μαύρα κουτιά αφού η μόνη εξήγηση για τη σύσταση ενός στοιχείου πως σε άγνωστους χρήστες με παρόμοιες προτιμήσεις άρεσε αυτό το στοιχείο.

Νέο αντικείμενο. Τα με βάση το περιεχόμενο συστήματα είναι σε θέση να συστήσουν στοιχεία που δεν έχουν βαθμολογηθεί από οποιονδήποτε χρήστη . Κατά συνέπεια, δεν υποφέρουν από το πρώτης εκτίμησης πρόβλημα, το οποίο επηρεάζει τη συνεργατική τεχνική που βασίζεται αποκλειστικά στις προτιμήσεις των χρηστών για να κάνει συστάσεις . Ως εκ τούτου, έως ότου το νέο στοιχείο να έχει βαθμολογηθεί από ένα σημαντικό αριθμό χρηστών, το σύστημα δεν θα είναι σε θέση να το προτείνει.

Παρ' όλα αυτά τα συστήματα αυτά έχουν και ορισμένα μειονεκτήματα:

Περιορισμένη ανάλυση περιεχομένου. Οι τεχνικές με βάση το περιεχόμενο έχουν ένα φυσικό όριο στον αριθμό και το είδος των χαρακτηριστικών που συνδέονται με τα αντικείμενα που προτείνουν. Συχνά είναι απαραίτητο ένα πεδίο γνώσης, π.χ., για τις συστάσεις ταινιών, το σύστημα πρέπει να γνωρίζει τους ηθοποιούς και σκηνοθέτες. Τα συστήματα που δεν είναι με βάση το περιεχόμενο μπορούν να παρέχουν κατάλληλες προτάσεις, αν το περιεχόμενο που αναλύθηκε δεν περιλαμβάνει επαρκείς πληροφορίες για να διακρίνουν τα στοιχεία που ο χρήστης προτιμάει από τα στοιχεία που δεν προτιμάει. Μερικές αναπαραστάσεις των στοιχείων περιέχουν μόνο ορισμένες πτυχές του περιεχομένου, αλλά υπάρχουν πολλές άλλες που θα επηρέαζαν την εμπειρία του χρήστη. Για παράδειγμα, συχνά δεν υπάρχουν επαρκείς πληροφορίες όσον αφορά τη συχνότητα εμφάνισης λέξεων για να μοντελοποιήσει τις προτιμήσεις των χρηστών σε αστεία ή ποιήματα, ενώ οι τεχνικές συναισθηματικής υπολογιστικής είναι οι πλέον κατάλληλες.

Εξειδίκευση σε μεγάλο βαθμό. Το content based σύστημα δεν έχει καμία εγγενή μέθοδο για την εύρεση κάτι αναπάντεχου . Το σύστημα προτείνει είδη των οποίων η βαθμολογία είναι υψηλή όταν συγκρίνεται με το προφίλ χρήστη, ως εκ τούτου, πρόκειται να συνίστανται στο χρήστη στοιχεία παρόμοια με αυτά που έχουν ήδη αξιολογηθεί . Αυτό το μειονέκτημα είναι επίσης ονομάζεται τονίζει την τάση των συστημάτων που βασίζονται στο περιεχόμενο να διατυπώνουν συστάσεις με περιορισμένο βαθμό καινοτομίας .

Νέος χρήστης. Πρέπει να συλλεχθούν αρκετές αξιολογήσεις πριν το σύστημα να μπορεί να καταλάβει πραγματικά τις προτιμήσεις των χρηστών και να παρέχει ακριβείς συστάσεις . Ως εκ τούτου, όταν λίγες αξιολογήσεις είναι διαθέσιμες , όπως για ένα νέο χρήστη, το σύστημα δεν θα είναι σε θέση να παράσχει αξιόπιστες συστάσεις.

3.3 Δημογραφικό σύστημα σύστασης (Demographic Filtering)

Οι δημογραφικές προσεγγίσεις φιλτραρίσματος χρησιμοποιούν τις περιγραφές των ανθρώπων για να μάθουν τη σχέση μεταξύ ενός στοιχείου και του τύπου ανθρώπων που τους αρέσει. Τα προφίλ χρηστών δημιουργούνται με την ταξινόμηση των χρηστών σε στερεοτυπικές περιγραφές, που αντιπροσωπεύουν τα χαρακτηριστικά γνωρίσματα των κατηγοριών των χρηστών. Τα προσωπικά στοιχεία του χρήστη είναι απαραίτητα και χρησιμοποιούνται για την ταξινόμηση. Οι ταξινομήσεις χρησιμοποιούνται ως γενικοί χαρακτηρισμοί για τους χρήστες και τα ενδιαφέροντά τους. Συνήθως, τα προσωπικά στοιχεία του χρήστη λαμβάνονται κατά την αίτηση εγγραφής στο σύστημα. Τα προφίλ που προκύπτουν επεκτείνουν το εύρος των πληροφοριών που περιλαμβάνεται στη δημογραφική βάση δεδομένων.

Παραδείγματος χάριν, η μέθοδος που εφαρμόζεται στο LifeStyle Finder (Krulwich 1997) χρησιμοποιεί ένα δημογραφικό σύστημα που καλείται PRIZM και διαιρεί τον πληθυσμό των Ηνωμένων Πολιτειών σε 62 δημογραφικές συστάδες σύμφωνα με το ιστορικό των αγορών τους, τα χαρακτηριστικά του τρόπου ζωής τους και τις απαντήσεις τους σε διάφορες έρευνες.

Ένα δημογραφικό σύστημα φιλτραρίσματος έχει δύο βασικά μειονεκτήματα: οι συστάσεις αποδεικνύονται πάρα πολύ γενικές και δεν προσαρμόζονται στις αλλαγές ενδιαφέροντος (Koychev 2000). Παρόλα αυτά, οι δημογραφικές πληροφορίες μπορούν να αποτελέσουν μια χρήσιμη τεχνική εάν συνδυαστούν με άλλες προσεγγίσεις.

3.4 Σύστημα σύστασης με βάση τη γνώση (Knowlegde–based Filtering)

Τα συστήματα που βασίζονται στη γνώση προτείνουν αντικείμενα που βασίζονται σε συγκεκριμένο πεδίο γνώσης σχετικά με το πώς ορισμένα χαρακτηριστικά στοιχείου ανταποκρίνονται στις ανάγκες και τις προτιμήσεις των χρηστών και, τελικά, πως το στοιχείο είναι χρήσιμο για το χρήστη. Τα παραδοσιακά συστήματα σύστασης (με βάση το περιεχόμενο φιλτράρισμα και συνεργατικά) είναι κατάλληλα για τη σύσταση προϊόντων της ποιότητας και γεύσης όπως βιβλία, ταινίες, ή ειδήσεις. Ωστόσο, ειδικά στο πλαίσιο των προϊόντων όπως αυτοκίνητα, υπολογιστές, διαμερίσματα, ή οι χρηματοπιστωτικές υπηρεσίες αυτές προσεγγίσεις δεν είναι η καλύτερη επιλογή. Για παράδειγμα, τα διαμερίσματα δεν αγοράζονται πολύ συχνά, γεγονός που καθιστά μάλλον ανέφικτη τη συλλογή πολλών αξιολογήσεων για ένα συγκεκριμένο αντικείμενο (όπως ακριβώς οι βαθμολογίες που απαιτούνται από τους αλγορίθμους συνεργατικού φιλτραρίσματος). Επιπλέον, δεν θα είναι ικανοποιημένοι οι χρήστες των εφαρμογών με συστάσεις ετών για τις προτιμήσεις αντικειμένων (όπως θα ήταν οι συστάσεις ενός αλγόριθμου με βάση το περιεχόμενο). Οι μέθοδοι συστάσεων που βασίζονται στη γνώση βοηθούν στην αντιμετώπιση αυτών των προκλήσεων με την αξιοποίηση ρητών απαιτήσεων των χρηστών και βαθιάς γνώσης για τον τομέα των προϊόντων για τον υπολογισμό των συστάσεων. Τα συστήματα αυτά σε μεγάλο βαθμό επικεντρώνονται σε πηγές γνώσης που δεν αξιοποιούνται από το συνεργατικό και με βάση

το περιεχόμενο φιλτράρισμα. Σε σύγκριση με το συνεργατικό φιλτράρισμα και το φιλτράρισμα με βάση το περιεχόμενο, τα συστήματα αυτά δεν αντιμετωπίζουν το πρόβλημα της «ψυχρής εκκίνησης», αλλά η υλοποίησή τους είναι δυσκολότερη.

Αξιοσημείωτα εισηγητικά συστήματα που βασίζονται στη γνώση είναι τα βασισμένα στην περίπτωση (case based) [36, 37]. Σε αυτά τα συστήματα η συνάρτηση ομοιότητας υπολογίζει πόσο οι ανάγκες των χρηστών (περιγραφή του προβλήματος) ταιριάζουν με τις συστάσεις (λύσεις του προβλήματος). Εδώ ο βαθμός ομοιότητας μπορεί να ερμηνεύεται άμεσα ως η χρησιμότητα της σύστασης για τον χρήστη. Τα συστήματα που βασίζονται σε περιορισμούς (constraint based) είναι ένα άλλο είδος των συστημάτων σύστασης που βασίζονται στη γνώση. Όσον αφορά τη γνώση που χρησιμοποιούν, τα δύο συστήματα είναι παρόμοια: συλλέγονται οι απαιτήσεις των χρηστών· σε περιπτώσεις ασυνεπών απαιτήσεων που δεν μπορούν να βρεθούν λύσεις προτείνονται αυτομάτως επιδιορθώσεις· και τα αποτελέσματα συστάσεων μπορούν να εξηγηθούν. Η σημαντική διαφορά έγκειται στον τρόπο που υπολογίζονται οι λύσεις. Τα συστήματα με βάση την περίπτωση καθορίζουν τις συστάσεις τους βάσει μετρικών ομοιότητας ενώ τα συστήματα με βάση τους περιορισμούς εκμεταλλεύονται κυρίως προκαθορισμένες βάσεις γνώσης που περιέχουν σαφείς κανόνες για το πώς να συσχετίζουν τις απαιτήσεις του πελάτη με τα χαρακτηριστικά στοιχείου. Τα συστήματα που βασίζονται στη γνώση τείνουν να λειτουργούν καλύτερα από τα άλλα κατά την έναρξη της εφαρμογής τους, αλλά αν δεν είναι εξοπλισμένα με εργαλεία μάθησης δεν υπερτερούν σε σχέση με άλλες μεθόδους που μπορούν να εκμεταλλευτούν τα αρχεία καταγραφής της αλληλεπίδρασης ανθρώπου / υπολογιστή.

Τα συστήματα που βασίζονται στους περιορισμούς τυπικά ορίζονται από δύο σύνολα μεταβλητών (V_C, V_{PROD}) και τρία διαφορετικά σύνολα περιορισμών (C_R, C_F, C_{PROD}) . Αυτές οι μεταβλητές και οι περιορισμοί είναι τα κύρια συστατικά ενός προβλήματος ικανοποίησης περιορισμών. Μια λύση για έναν περιορισμό πρόβλημα ικανοποίηση αποτελείται από ένα συνδυασμό των μεταβλητών, έτσι ώστε όλοι οι περιορισμοί που ορίζονται να πληρούνται.

Οι απαιτήσεις του χρήστη/πελάτη V_C περιγράφουν τις πιθανές απαιτήσεις των πελατών. Οι ιδιότητες του προϊόντος V_{PROD} περιγράφουν τις ιδιότητες ενός δεδομένου είδους προϊόντων. Παραδείγματα για τις ιδιότητες του προϊόντος είναι το είδος του προϊόντος ή το όνομα του προϊόντος. Οι περιορισμοί C_R είναι οι περιορισμοί που συστηματικά περιορίζουν τις απαιτήσεις του πελάτη. Το φίλτρο προϋποθέσεων C_F καθορίζει τη σχέση μεταξύ του των απαιτήσεων των πελατών και το συγκεκριμένο είδος προϊόντων. Ένα παράδειγμα φίλτρου προϋποθέσεων είναι η εξής: οι πελάτες χωρίς εμπειρία στον τομέα των χρηματοοικονομικών υπηρεσιών θα πρέπει να μην έχουν προτάσεις που περιλαμβάνουν προϊόντα υψηλού κινδύνου. Η μεταβλητή, τέλος, C_{PROD} αντιπροσωπεύει τους βασικούς περιορισμούς σχετικά με τις πιθανές μεταβλητές V_{PROD}. Με δεδομένο ένα σύστημα με βάση τη γνώση με τις παραπάνω μεταβλητές και ένα σύνολο απαιτήσεων είμαστε σε θέση να υπολογίσουν τις συστάσεις . Ο σκοπός τώρα του συστήματος που είναι η ανάδειξη ενός συνόλου προϊόντων που ταιριάζουν στις επιθυμίες και τις ανάγκες του πελάτη μπορεί να οριστεί ως ένα πρόβλημα ικανοποίησης περιορισμών ($V_C, V_{PROD}, C_C \cup C_F \cup C_R \cup C_{PROD}$) .

Η μέθοδος σύστασης που βασίζεται στην υπόθεση (case based) είναι μία από τις πιο επιτυχημένες μεθόδους μηχανικής μάθησης. Η τεχνική αυτή είναι μία κυκλική και ολοκληρωμένη διαδικασία επίλυσης προβλημάτων που στηρίζεται στη μάθηση από την εμπειρία και έχει τέσσερα κύρια στάδια : την ανάκτηση, επαναχρησιμοποίηση, προσαρμογή και διατήρηση. Με την

εμφάνιση ενός νέου προβλήματος αναζητά ένα παρόμοιο πρόβλημα του παρελθόντος (που έχει ήδη επιλύσει μία παρόμοια υπόθεση), και στη συνέχεια επαναχρησιμοποιεί την υπόθεση εκείνη για την επίλυση του σημερινού προβλήματος . Σε αυτές τις προσεγγίσεις μια υπόθεση και ένα προϊόν ουσιαστικά θεωρούνται πανομοιότυπα αντικείμενα . Το πρόβλημα της σύστασης συνήθως αντιπροσωπεύεται από ένα σύνολο χαρακτηριστικών του προϊόντος, εκείνες που καθορίζονται από το χρήστη, και η λύση της υπόθεσης είναι το ίδιο το προϊόν . Στο βασικό σενάριο χρήσης, ο πελάτης ψάχνει να αγοράσει κάποιο προϊόν και καθιστά σαφείς ορισμένες απαιτήσεις σχετικά με το προϊόν αυτό. Το σύστημα αναζητά την βασική υπόθεση για τα προϊόντα που ταιριάζουν με τις απαιτήσεις του χρήστη. Η διαδικασία ανάκτησης οδηγείται από ένα μετρητή ομοιότητας που υπολογίζει την ομοιότητα της περιγραφής του προβλήματος, δηλαδή, τις σημερινές απαιτήσεις των χρήστη με τα προϊόντα στη βάση δεδομένων της υπόθεσης. Μια σειρά από προϊόντα στη συνέχεια ανακτώνται από τη βασική υπόθεση και τα προϊόντα αυτά συνιστώνται στο χρήστη. Αν ο χρήστης δεν είναι ικανοποιημένος με τις υποδείξεις μπορεί να τροποποιήσει τις απαιτήσεις του και ένας νέος κύκλος σύστασης ξεκινάει.

3.5 Υβριδικό σύστημα σύστασης (Hybrid Filtering)

Πολλά συστήματα σύστασης χρησιμοποιούν μια υβριδική προσέγγιση, συνδυάζοντας την συνεργατική και τη με βάση το περιεχόμενο μέθοδο, η οποία βοηθά στην αποφυγή ορισμένων περιορισμών των με βάση το περιεχόμενο και των συνεργατικών συστημάτων. Οι διαφορετικοί τρόποι για να συνδυαστούν οι προαναφερθέντες μέθοδοι σε ένα υβριδικό σύστημα συστάσεων μπορούν να ταξινομηθούν ως εξής: (1) εφαρμογή συνεργατικής και με βάση το περιεχόμενο μεθόδου ξεχωριστά και συνδυάζοντας τις προβλέψεις τους, (2) ενσωμάτωση κάποιων χαρακτηριστικών με βάση το περιεχόμενο σε μια συνεργατική προσέγγιση, (3) ενσωμάτωση ορισμένων χαρακτηριστικων συνεργατικής προσέγγισης σε ένα σύστημα με βάση το περιεχόμενο και (4) κατασκευής ενός γενικού ενοποιητικό μοντέλο που ενσωματώνει τα χαρακτηριστικά και των δύο. Όλες οι παραπάνω προσεγγίσεις έχουν χρησιμοποιηθεί από ερευνητές συστημάτων συστάσεων, όπως περιγράφεται παρακάτω.

Συνδυάζοντας ξεχωριστά τα συστήματα. Ένας τρόπος για να οικοδομηθεί ένα υβριδικό σύστημα συστάσεων είναι να εφαρμόσει ξεχωριστά συστήματα που βασίζονται σε περιεχόμενο και συνεργασία. Στη συνέχεια, υπάρχουν δύο διαφορετικά σενάρια. Πρώτον, μπορούμε να συνδυάσουμε τις εξόδους (αξιολογήσεις) που λαμβάνονται από κάθε σύστημα συστάσεων ξεχωριστά σε μια τελική σύσταση χρησιμοποιώντας είτε ένα γραμμικό συνδυασμό των αξιολογήσεων [38] ή σύστημα ψηφοφορίας [39]. Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε ένα από τα επιμέρους συστήματα, σε οποιαδήποτε χρονική τη στιγμή να επιλεγεί να χρησιμοποιηθεί το αυτό που είναι «καλύτερο» βασισμένο σε κάποια μέτρο ποιότητας σύστασης.

Προσθέτοντας χαρακτηριστικά τεχνικών που βασίζονται στο περιεχόμενο σε συνεργατικά μοντέλα. Αρκετά υβριδικά εισηγητικά συστήματα, συμπεριλαμβανομένων των Fab [40] και της

προσέγγισης "συνεργασία με περιεχόμενο", που περιγράφεται στο [39], βασίζονται σε παραδοσιακές τεχνικές συνεργασίας, αλλά διατηρούν και προφίλ με βάση το περιεχόμενο για κάθε χρήστη. Αυτά τα προφίλ χρησιμοποιούνται για τον υπολογισμό της ομοιότητας μεταξύ δύο χρηστών, και όχι εκείνα των κοινών στοιχείων που έχουν αξιολογηθεί. Όπως αναφέρεται στο [39], αυτό επιτρέπει να ξεπεραστούν τα προβλήματα των ελαχίστων αναφορών που σχετίζονται με αμιγώς συνεργατικές προσεγγίσεις. Ένα άλλο όφελος της αυτής της προσέγγισης είναι ότι στους χρήστες μπορεί να συστηθεί ένα στοιχείο όχι μόνο όταν το στοιχείο αυτό κατέχει υψηλή θέση από τους χρήστες με παρόμοιο προφίλ, αλλά και άμεσα, δηλαδή, όταν βαθμολογείται υψηλά συγκριτικά με το προφίλ του χρήστη.

Προσθέτοντας συνεργατικά χαρακτηριστικά σε μοντέλα που βασίζονται στο περιεχόμενο. Ο πιο δημοφιλής προσέγγιση σε αυτή την κατηγορία είναι η τεχνική μείωσης διάστασης σε μια ομάδα προφίλ με βάση το περιεχόμενο. Για παράδειγμα, στην αναφορά [41] χρησιμοποιούν λανθάνουσα σημασιολογική ευρετηρίαση (*Latent Semantix Indexing*-LSI) για να δημιουργήσουν μια συλλογή από προφίλ των χρηστών, όπου τα προφίλ των χρηστών αναπαρίστανται από διανύσματα όρων. Αυτό έχει ως αποτέλεσμα μια βελτίωση της απόδοσης σε σχέση με τη καθαρή προσέγγιση με βάση το περιεχόμενο.

Αναπτύσσοντας ένα ενιαίο ενοποιητικό μοντέλο σύστασης. Πολλοί ερευνητές έχουν ακολουθήσει τη προσέγγιση αυτή κατά τα τελευταία χρόνια. Για παράδειγμα, στο [42] προτείνεται η χρήση χαρακτηριστικών τεχνικών με βάση το περιεχόμενο καθώς και συνεργατικών (π.χ., η ηλικία ή το φύλο των χρηστών ή το είδος των ταινιών) σε έναν ενιαίο ταξινομητή που διέπεται από κάποιους κανόνες. [43] και [44] προτείνουν μία ενιαία πιθανολογική μέθοδο για το συνδυασμό συνεργατικών και με βάση το περιεχόμενο των συστάσεων, η οποία βασίζεται στην πιθανολογική λανθάνουσα σημασιολογική ανάλυση [45]. Ακόμη άλλη προσέγγιση προτείνεται από την [46] και [47], όπου Bayesian μοντέλα μικτών αποτελεσμάτων παλινδρόμησης χρησιμοποιούνται και εφαρμόζουν Monte Carlo μεθόδους αλυσίδας Markov για την παραμετροποίηση της εκτίμησης και της πρόβλεψης. Εν ολίγοις, [47] χρησιμοποιεί τα χαρακτηριστικά του χρήστη που αποτελούν μέρος ενός προφίλ χρήστη, τα χαρακτηριστικά των αντικειμένων που αποτελούν μέρος ενός προφίλ στοιχείου και τις αλληλεπιδράσεις τους για την εκτίμηση της αξιολόγησης ενός στοιχείου.

Τελικά, πολλές έρευνες, όπως [40, 39, 41], που εμπειρικά συγκρίνουν την απόδοση της υβριδικής τεχνικής με τις τεχνικές συνεργασίας και με βάση το περιεχόμενο αποδεικνύουν ότι οι υβριδικοί μέθοδοι μπορούν να παρέχουν πιο ακριβείς συστάσεις από τις καθαρές προσεγγίσεις.

4. Ανατροφοδότηση Σχετικότητας και Σύσταση (Relevance Feedback in Recommender Systems)

Τα ανθρώπινα ενδιαφέροντα αλλάζουν καθώς περνά ο χρόνος. Για παράδειγμα, ένας νέος πατέρας μπορεί να ενδιαφέρεται για την φροντίδα του βρέφους μόλις μετά τον τοκετό, αλλά το ενδιαφέρον αυτό σταδιακά ελαττώνεται με την πάροδο του χρόνου. Ως εκ τούτου, το σύστημα σύστασης χρειάζεται τις τελευταίες πληροφορίες για να ενημερώνει αυτόματα το προφίλ του χρήστη. Στο κεφάλαιο αυτό, παρουσιάζονται αρκετοί τρόποι για να αποκτήσει αυτές τις πληροφορίες, τις οποίες αποκαλούμε ανατροφοδότηση. Η ανατροφοδότηση είναι ένα βασικό στοιχείο των περισσότερων συστημάτων συστάσεων. Η ιδέα είναι να συμμετάσχει και ο χρήστης στη διαδικασία ανάκτησης πληροφοριών για να βελτιωθεί το αποτέλεσμα της αναζήτησης[1,13]. Έτσι, η διαδικασία σε απλή μορφή είναι η εξής:

- Ο χρήστης κάνει μία αναζήτηση.
- Το σύστημα επιστρέφει ένα αρχικό σύνολο αποτελεσμάτων.
- Ο χρήστης στη συνέχεια αξιολογεί τα αποτελέσματα ως σχετικά ή σχετικά.
- Το σύστημα επανα-υπολογίζει την παρουσίαση των αποτελεσμάτων βασισμένο στην ανατροφοδότηση του χρήστη
- Σε επόμενη αναζήτηση παρουσιάζεται η ανανεωμένη παρουσίαση αποτελεσμάτων.
- Η σχετική ανατροφοδότηση επαναλαμβάνεται μία ή περισσότερες φορές.

Τυπικά, είναι δυνατόν να διακρίνουμε δύο είδη ανατροφοδότησης ενδιαφέροντος: θετικές πληροφορίες (στοιχεία που άρεσαν στο χρήστη) και αρνητικές πληροφορίες (δηλαδή, στοιχεία για τα οποία ο χρήστης δεν ενδιαφέρεται). Οι συγγραφείς στο [48] ισχυρίζονται ότι οι αρνητικές πληροφορίες οδηγούν σε δραματική βελτίωση της απόδοσης του συστήματος. Ωστόσο, υπάρχουν μερικά συστήματα τα οποία δεν μπορούν να λάβουν υπόψη τις αρνητικές πληροφορίες, επειδή η ακρίβεια του συστήματος είναι πιθανό να μειωθεί [49]. Έτσι, όλα εξαρτώνται από το σύστημα. Οι δύο πιο συνηθισμένοι τρόποι για να συγκεντρωθούν οι ανατροφοδοτήσεις ενδιαφέροντος είναι να χρησιμοποιηθούν οι πληροφορίες που δίνονται ρητά από τους χρήστες ή οι πληροφορίες που παρατηρούνται έμμεσα από την αλληλεπίδραση του χρήστη. Επιπλέον, ορισμένα συστήματα προτείνουν υβριδικές προσεγγίσεις συνδυάζοντας και τις δύο παραπάνω πληροφορίες.

Η ανάγκη να σηματοδοτηθούν σαφώς τα αντικείμενα ως σχετικά ή μη σχετικά ενέχει τον κίνδυνο να είναι απρόθυμοι οι χρήστες να παρέχουν άμεσα πληροφορίες ενδιαφέροντος. Η έμμεση ανατροφοδότηση, με την οποία ένα σύστημα σύστασης παρακολουθεί διακριτικά τη συμπεριφορά αναζήτησης, εξαλείφει την ανάγκη να αναφέρεται ρητά από τους χρήστες ποια αντικείμενα τους ενδιαφέρουν. Η τεχνική αυτή χρησιμοποιεί έμμεσες ενδείξεις ενδιαφέροντος, που συγκεντρώθηκαν από την αλληλεπίδραση του χρήστη με το σύστημα, για να τροποποιήσει το αρχικό ερώτημα.

4.1 Έμμεση ανατροφοδότηση (Implicit Feedback)

Η έμμεση ανατροφοδότηση, όπως αναφέρθηκε, προκύπτει από παρατηρήσεις για τη συμπεριφορά των χρηστών κατά τη διάρκεια αλληλεπίδρασης με το σύστημα, όπως ποια έγγραφα επιλέγουν ή δεν επιλέγουν για προβολή, η διάρκεια του χρόνου που δαπανάται στην προβολή ενός εγγράφου, ή η περιήγηση σε μία σελίδα. Οι βασικές διαφορές της έμμεσης ανατροφοδότησης από τη ρητή περιλαμβάνουν:

- ο χρήστης δεν αξιολογεί προς όφελος του συστήματος ανάκτησης πληροφορίας, αλλά μόνο για να ικανοποιήσει τις δικές του ανάγκες και
- ο χρήστης δεν είναι ενήμερος ότι η συμπεριφορά του (επιλεγμένα έγγραφα) θα χρησιμοποιηθεί ως σχετική ανατροφοδότηση

Ένα από τα πλεονεκτήματα είναι ότι δεν απαιτείται καμία πρόσθετη ενέργεια από το χρήστη, σε αντίθεση με τη ρητή ανατροφοδότηση και το πρόβλημα του "κίνητρου" να αξιολογήσει ο χρήστης. Ωστόσο, έχει το μειονέκτημα ότι είναι «θορυβώδης», αφού δεν υπάρχει ρητή ετικέτα μου αρέσει / δεν μου αρέσει, θα πρέπει το σύστημα να μαντέψει την αξιολόγηση του χρήστη. Κάποια παράδειγμα έμμεσης ανατροφοδότησης είναι:

Clickthrough δεδομένα. Παρατηρήσεις για το σε ποια αποτελέσματα αναζήτησης ο χρήστης κάνει κλικ. Αυτή είναι ίσως η πιο ευρέως χρησιμοποιούμενη μορφή έμμεσης ανατροφοδότησης. Η βασική ιδέα είναι ότι ο χρήστης πιθανώς τείνει να κάνει κλικ στα αποτελέσματα που έχουν μεγαλύτερη σημασία γι' αυτόν.

Ιστορικό αναζήτησης του χρήστη. Οι παρατηρήσεις του ιστορικού αναζητήσεων του χρήστη. Αυτό περιλαμβάνει αναδιατυπώσεις του ερωτήματος, το οποίο μπορεί να χρησιμοποιηθεί για να συμπεράνουμε τη δυσαρέσκεια του χρήστη με τα αποτελέσματα που επιστρέφονται από την αρχική διατύπωση. Η εξέταση των ερωτημάτων που προηγούνται άμεσα ενός ερωτήματος μπορεί επίσης να είναι μια ένδειξη του ενδιαφέροντος του χρήστη, και μπορούν να χρησιμοποιηθούν για την αποσαφήνιση ερωτημάτων / λέξεων με πολλαπλές σημασίες.

Ολόκληρο το ιστορικό του χρήστη. Παρατηρήσεις όλων των πληροφοριών που δημιουργούνται, αντιγράφονται, ή προβάλλονται από το χρήστη. Αυτό θα μπορούσε να περιλαμβάνει τα πάντα, από ιστοσελίδες που είδαν, μέχρι e-mail, στοιχεία ημερολογίου και εγγράφων σε σύστημα αρχείων του χρήστη.

Χρόνος προβολής / ανάγνωσης. Παρατήρηση του ποσού του χρόνου που ξοδεύει ο χρήστης για κάθε αποτέλεσμα. Φαίνεται λογικό ότι ένας χρήστης θα ξοδεύει περισσότερο χρόνο σε πιο σχετικά αποτελέσματα.

Eye Tracking. Παρατήρηση χαρακτηριστικών, όπως η εστίαση του ματιού και η διαστολή της κόρης καθώς ο χρήστης βλέπει τα αποτελέσματα. Η υπόθεση είναι ότι τα παραπάνω χαρακτηριστικά ποικίλλουν με συστηματικό τρόπο μεταξύ των σχετικών και μη σχετικών αποτελεσμάτων. Για παράδειγμα, η μεγαλύτερη διάμετρος της κόρης μπορεί να αποτελεί ένδειξη ενδιαφέροντος.

Αν και δεν είναι τόσο ακριβής όσο η ρητή ανατροφοδότηση, προηγούμενες εργασίες[50] έχουν δείξει ότι η έμμεση ανατροφοδότηση μπορεί να είναι ένα αποτελεσματικό υποκατάστατο ρητής ανατροφοδότησης σε διαδραστικά περιβάλλοντα αναζήτησης πληροφοριών. Στη συνέχεια θα αναλυθεί μία ποικιλία μοντέλων έμμεσης ανατροφοδότησης και μια ποικιλία διαφορετικών μεθόδων στάθμισης ανατροφοδότησης με βάση την έμμεση ένδειξη ενδιαφέροντος. Τα μοντέλα έμμεσης ανατροφοδότηση που παρουσιάζονται χρησιμοποιούν διαφορετικές μεθόδους για τη διαχείριση της ανατροφοδότησης.

4.1.1 Μοντέλο Jeffrey's Conditioning

Το μοντέλο έμμεσης ανάδρασης αυτό χρησιμοποιεί τον κανόνα της προσαρμογής του Jeffrey (Jeffrey's Conditioning Model)[53], για να βρεθεί η πιθανότητα να είναι σχετικός ένας όρος από τα στοιχεία που συγκεντρώθηκαν από αλληλεπιδράσεις του χρήστη. Η μέθοδος αυτή λαμβάνει υπόψη τον αβέβαιο χαρακτήρα των έμμεσων ενδείξεων ενδιαφέροντος, και χρησιμοποιεί διάφορα μέτρα για να περιγράψει την αξία ή τιμή των στοιχείων μίας αναπαράστασης εγγράφου. Το μοντέλο Jeffrey συνδυάζει μέτρα για τη σχετική θέση των αναπαραστάσεων στη διαδρομή σχετικότητας (relevance path). *Διαδρομή σχετικότητας* είναι ένας συνδυασμός αναπαραστάσεων του εγγράφου, θεωρητικά όσο «διασχίζει» τη διαδρομή αυτή ο χρήστης τόσο πιο σχετική είναι η πληροφορία που λαμβάνει. Στη συνέχεια παρουσιάζονται κάποια μέτρα που χρησιμοποιεί το μοντέλο αυτό.

4.1.1.1 Στάθμιση μονοπατιού (Path Weighting)

Η λογική του μέτρου αυτού είναι ότι για κάθε διαδρομή, γινόμαστε πιο σίγουροι για την αξία πληροφοριών ενδιαφέροντος που έχουν περάσει. Στην προσέγγισή στο [52] ανατίθεται ένα εκθετικά αυξανόμενο προφίλ για ενδιαφέροντα με μεγάλη χρονική διάρκεια. Η υπόθεση είναι ότι όσο περαιτέρω ταξιδεύουμε κατά μήκος μιας διαδρομής σχετικότητας, τόσο πιο σίγουροι είμαστε για την ορθότητα των πληροφοριών προς την αρχή του μονοπατιού. Καθώς η προβολή της επόμενης αναπαράστασης είναι διερευνητική και οδηγείται από την περιέργεια καθώς και πληροφορίες που χρειάζεται ο χρήστης, είμαστε λιγότερο σίγουροι για την αξία αυτών των στοιχείων. Η «εμπιστοσύνη» αυτή, c, έχει εκχωρηθεί από την αρχή του μονοπατιού για κάθε i αναπαράσταση,

$$c_i = \frac{1}{2^i}\ , i \geq 1 \qquad 4.1$$

Ωστόσο, δεδομένου ότι σε μια ολόκληρη διαδρομή, οι τιμές των c_i δεν έχουν άθροισμα ένα, οφείλουμε να τις ομαλοποιήσουμε και να υπολογίζουμε τη c εμπιστοσύνη για κάθε αναπαράσταση i σε μια διαδρομή μήκους N χρησιμοποιώντας,

$$c_i = \frac{1}{2^i} + \frac{1}{N.2N}\ ,\ \sum_{i=1}^{N} c_i = 1 \quad i \epsilon \{1,2,\dots,N\} \qquad 4.2$$

4.1.1.2 Ποιότητα και κατατοπιστική αξία των στοιχείων (Quality of Evidence)

Στην προηγούμενη ενότητα περιγράφεται η εμπιστοσύνη στην καταλληλότητα των αναπαραστάσεων με βάση τη θέση τους στη διαδρομή ενδιαφέροντος. Η ποιότητα των στοιχείων σε μια αναπαράσταση, ή κατατοπιστική αξία (indicativity) τους, μπορεί επίσης να επηρεάσει πόσο σίγουροι είμαστε για την αξία του περιεχομένου της. Οι τίτλοι και οι υψηλότερα βαθμολογημένες προτάσεις, οι οποίες μπορεί να είναι πολύ ενδεικτικές του περιεχομένου του εγγράφου, είναι σύντομες και θα έχουν χαμηλές βαθμολογίες εάν το χαρακτηριστικό που χρησιμοποιείται για να τις βαθμολογήσει είναι το μήκος τους. Στην προσέγγιση στο [52] στο έγγραφο d δίνεται βάρος σε ένα όρο t χρησιμοποιώντας την κανονικοποιημένη συχνότητα όρου [54], και το άθροισμα όλων των βαρών ενός εγγράφου είναι 1. Όσο μεγαλύτερη η τιμή αυτή, τόσο πιο συχνά εμφανίζεται στο έγγραφο ο όρος, και τόσο πιο αντιπροσωπευτικός του περιεχομένου του εγγράφου θεωρείται. Για να υπολογίσουμε το δείκτη indicativity I (μέτρηση του πόσο αντιπροσωπευτικός είναι ένας όρος για το περιεχόμενο του εγγράφου) για μια αναπαράσταση r αθροίζονται τα βάρη των όρων σε ένα έγγραφο $w_{t,d}$ για όλους τους μοναδικούς όρους r,

$$I_r = \sum_{t \epsilon r} w_{t,d} \qquad 4.3$$

Το I_r κυμαίνεται μεταξύ 0 και 1, δεν είναι ποτέ 0, και είναι 1 μόνο εάν η αναπαράσταση περιέχει κάθε μοναδικό όρο στο έγγραφο. Το μέτρο αυτό αυξάνεται μόνο εάν υπάρχει αντιστοιχία μεταξύ των μοναδικών όρων στο έγγραφο και εκείνων της αναπαράστασης. Η διαδρομή σχετικότητας θα περιέχει αναπαραστάσεις διαφορετικής ποιότητας. Υπολογίζεται η αξία των στοιχείων σε μια αναπαράσταση με τον πολλαπλασιασμό του *indicativity I* με την εμπιστοσύνη. Χρησιμοποιώντας αυτά τα μέτρα εξασφαλίζεται ότι οι αναπαραστάσεις με μεγαλύτερη αξία είναι αυτές που συμβάλλουν περισσότερο στην επιλογή των δυνητικά χρήσιμων όρων επέκτασης ερωτήματος σε μία διαδρομή σχετικότητας. Στην επόμενη ενότητα θα περιγράψουμε τον τρόπο με τον οποίο επιλέγονται οι όροι αυτοί.

4.1.1.3 Στάθμιση όρου (Term Weighting)

Το μοντέλο του Jeffrey προϋποθέτει την ύπαρξη ενός διαστήματος όρων T, ένα σύνολο αμοιβαία αποκλειόμενων όρων στο χώρο πληροφοριών. Κάθε όρος στο T είναι ανεξάρτητος και έχει μία συσχετισμένη συχνότητα στο χώρο των πληροφοριών. Ορίζεται η πιθανότητα ότι ένας όρος t είναι σχετικός με βάση την κατανομή πιθανοτήτων P στο διάστημα T ως,

$$P(t) = \frac{ntf(t)}{\sum_{t \epsilon T} ntf(t)} \ , \qquad 4.4$$

όπου $ntf(t)$ η κανονικοποιημένη συχνότητα όρου t στο διάστημα όρων T.

Για να ενημερωθεί αυτή η πιθανότητα βάση των νέων στοιχείων που συγκεντρώθηκαν από την αλληλεπίδραση χρησιμοποιείται ο κανόνας Jeffrey, που εφαρμόζεται στο τέλος κάθε διαδρομής σχετικότητας. Θεωρείται αυτή η διαδρομή σχετικότητας p ως μια νέα πηγή στοιχείων για την ενημέρωση της νέας πιθανότητας, έστω P'. Η προβολή μίας p_ι αναπαράστασης δημιουργεί νέα στοιχεία για τους όρους που βρίσκονται στην αναπαράσταση. Χρησιμοποιείται ο κανόνας Jeffrey για την ενημέρωση των πιθανοτήτων με βάση σε αυτά τα νέα στοιχεία χρησιμοποιώντας τον ακόλουθο τύπο,

$$P'(t) = [\, P(t = 1 |\, p_\iota) \ \frac{P\prime(t=1)}{P(t=1)} + P(t = 0 |\, p_\iota) \ \frac{P\prime(t=0)}{P\prime(t=0)}] \,.\, P(t) \qquad 4.5$$

Αυτή η εκτίμηση υπολογίζει την αναθεωρημένη πιθανότητα ενδιαφέροντος για έναν όρο t δοθείσας p_ι αναπαράστασης, όπου $P(t = 1)$ είναι η πιθανότητα παρατήρησης του όρου t, και $P(t = 0)$, η πιθανότητα της μη παρατήρησης του t. Μια διαδρομή σχετικότητας περιλαμβάνει μια σειρά αναπαραστάσεων. Ενημερώνονται οι πιθανότητες μετά τη διάσχιση της διαδρομής. Το μήκος μιας διαδρομής σχετικότητας κυμαίνεται μεταξύ 1 και 6 βημάτων και το μήκος της συμβολίζεται N. Όταν το μήκος αυτό είναι μεγαλύτερο από ένα ενημερώνονται οι πιθανότητες σε όλη αυτή τη διαδρομή. Η πιθανότητα σχετικότητας ενός όρου σε ολόκληρο το μήκος N της διαδρομής συμβολίζεται P_N και δίνεται από τον τύπο:

$$P_N(t) = \sum_{i=1}^{N-1} c_\iota . I_\iota . [(\, P_\iota(t = 1 |\, p_\iota) \ \frac{P_{\iota+1}\prime(t=1)}{P_\iota(t=1)} + P_\iota(t = 0 |\, p_\iota) \ \frac{P_{\iota+1}\prime(t=0)}{P_\iota\prime(t=0)}) \,.\, P_\iota(t)] \qquad 4.6$$

όπου μια αναπαράσταση στο βήμα i στο μονοπάτι p συμβολίζεται p_ι. Η εμπιστοσύνη στην αξία της αναπαράστασης συμβολίζεται c_ι και I_ι είναι η κατατοπιστική αξία της αναπαράστασης. Σε αυτή την εξίσωση, η σειρά της ενημέρωσης μετράει, όπως και η σειρά με την οποία ο χρήστης διασχίζει τη διαδρομή σχετικότητας. Η πραγματική αναθεώρηση των πιθανοτήτων θα συμβεί μετά από κάθε διαδρομή. Μόλις ενημερωθούν οι πιθανότητες, παραμείνουν σταθερές μέχρι την επόμενη αναθεώρηση (δηλαδή την επόμενη διαδρομή ενδιαφέροντος). Μόνο οι όροι του διαστήματος T που εμφανίζονται στη διαδρομή σχετικότητας λαμβάνονται υπόψη για τον υπολογισμό των πιθανοτήτων.

4.1.2 Μοντέλα γράφων για ανάλυση έμμεσης ανατροφοδότησης

Σε αυτή την ενότητα, παρουσιάζεται ένα σύνολο από αλγόριθμους για ανάλυση των σημάτων έμμεσης ανατροφοδότησης που βασίζονται σε γράφους. Οι αλγόριθμοι αυτοί μοντελοποιούν τα ιστορικά δεδομένα της αλληλεπίδρασης για όλους τους χρήστες και τις περιόδους της αλληλεπίδρασης. Τα δύο κύρια χαρακτηριστικά αυτού του μοντέλου γράφων είναι: 1) η εκπροσώπηση όλων των αλληλεπιδράσεων των χρηστών με το σύστημα, συμπεριλαμβανομένης της αλληλουχίας της αλληλεπίδρασης και 2) ο συνυπολογισμός όλων των πληροφοριών της έμμεσης ανάδρασης σε μία αναπαράσταση. Το γράφημα διευκολύνει την ανάλυση και αξιοποίηση των προηγούμενων έμμεσων πληροφοριών, δίνοντας ένα μοντέλο που είναι εύκολο να οικοδομηθεί πάνω σε διαφορετικούς αλγόριθμους συστάσεων. Την αναπαράσταση του γραφήματος μπορεί να τη συναντήσουμε σε δύο διαφορετικά επίπεδα: το πρώτο είναι ένας κατευθυνόμενος γράφος με ετικέτες (*Labelled Directed Multigraph* - LDM), δίνει μια πλήρη και λεπτομερή αναπαράσταση των πληροφοριών έμμεσης ανατροφοδότησης, και το δεύτερο ένας κατευθυνόμενος γράφος με βάρη (*Weighted Directed Graph* - WDG), συνάγεται από το προηγούμενο, απλοποιώντας την ερμηνεία του LDM. Είναι στην ευχέρεια του WDG που θα καθορίζονται οι διαφορετικές βαθμολογήσεις σύστασης. Σημειώστε ότι η WDG είναι δεν εξαρτάται από την LDM, και μπορεί να υπολογιστεί άμεσα.

Η συνεδρία του χρήστη s μπορεί να αναπαρασταθεί ως ένα σύνολο ερωτημάτων Q_s, οι οποίες εισήχθησαν από το χρήστη, και το σύνολο των εγγράφων πολυμέσων D_s με τα οποία ο χρήστης είχε αλληλεπίδραση κατά τη διάρκεια της συνόδου. Ερωτήματα και έγγραφα είναι, επομένως, οι κόμβοι $N_s = D_s \cup Q_s$ του γραφήματος $G_S = (N_S, A_S)$ στον οποίο τα τόξα είναι το σύνολο των ενεργειών $A_s(G) = \{n_i, n_j, a, u, t\}$, που δείχνουν ότι, σε χρόνο t, ο χρήστης u εκτελεί μια ενέργεια του τύπου a που οδηγεί το χρήστη από τον κόμβο n_i στον κόμβο n_j, όπου $n_i, n_j \epsilon N_s$. Το n_j είναι το αντικείμενο της ενέργειας, για παράδειγμα, όταν ένας χρήστης κάνει κλικ να δει ένα έγγραφο. Οι τύποι των ενεργειών εξαρτώνται από το είδος του ενεργειών που καταγράφονται από το σύστημα ανάκτησης πληροφορίας, όπως κλικ, χρόνος προβολής κ.λ.π. Οι σύνδεσμοι μπορεί να περιέχουν επιπλέον σχετικά μεταδεδομένα. Το γράφημα έχει πολλαπλούς συνδέσμους, αφού διαφορετικές δράσεις μπορεί να έχουν ίδιους κόμβους αρχής και προορισμού. Όλα τα γραφήματα που βασίζονται στη συνεδρία συγκεντρώνονται σε ένα ενιαίο γράφημα $G = G(N, A),\ N = N_s U_s, A = U_s A_s$, το οποίο μπορεί να θεωρηθεί ως μια συνολική δεξαμενή έμμεσης πληροφορίας.

Προκειμένου να καταστεί δυνατή η εκμετάλλευση της προηγούμενης αναπαράστασης από τον αλγόριθμο σύστασης, θα απλοποιηθεί ο LDM γράφος χρησιμοποιώντας συνδέσμους με βάρη χωρίς ετικέτες και ένωση όλων των συνδέσμων διασύνδεσης δύο κόμβων σε ένα. Αυτή η διαδικασία γίνεται σε δύο στάδια: το πρώτο βήμα υπολογίζει ένα σταθμισμένο γράφημα $G_s = (N_s, W_s)$ που αντιπροσωπεύει τις αλληλεπιδράσεις μεταξύ των χρηστών κατά τη διάρκεια μίας συνόδου. Οι σύνδεσμοι $W_s = \{n_i, n_j w_s\}$, δείχνουν ότι τουλάχιστον μία δράση οδηγεί το χρήστη από τον κόμβο n_ι στον n_j. Η τιμή βάρους w_s αντιπροσωπεύει την τελική τιμή ενδιαφέροντος που υπολογίστηκε για τον κόμβο n_j, το *τοπικό ενδιαφέρον* $l_r(n_j)$. Το *τοπικό ενδιαφέρον* ανακτάται από τη συσσώρευση στοιχείων έμμεσης ανάδρασης και δίνεται από τη συνάρτηση

$$l_r(n) = 1 - \frac{1}{x(n)}, \qquad 4.7$$

όπου $x(n)$ είναι το συνολικό προστιθέμενο βάρος που σχετίζεται με κάθε τύπο της δράσης που σχετίζεται με τον κόμβο n. Τα $x(n)$ βάρη είναι θετικές τιμές που επιστρέφει μια συνάρτηση $f(a): A \rightarrow \aleph$, η οποία επιστρέφει υψηλότερες τιμές εάν η δράση θεωρείται ότι δίνει περισσότερα στοιχεία έμμεσου ενδιαφέροντος. Για παράδειγμα, η πλοήγηση ενός χρήστη σε ένα βίντεο είναι μια κάπως καλή ένδειξη έμμεσου ενδιαφέροντος. Από την άλλη πλευρά, η διάρκεια προβολής ενός βίντεο έχει αποδειχθεί ότι δεν είναι τόσο καλή ένδειξη [56], έχοντας έτσι ένα χαμηλότερο βάρος.

4.1.2.1 Αλγόριθμοι μοντέλων γράφων

Καθώς ο χρήστης αλληλεπιδρά με το σύστημα, κατασκευάζεται ένας κατευθυνόμενος γράφος με βάρη με βάση τη συνεδρίαση. Η τρέχουσα συνεδρία του χρήστη αναπαρίσταται με $G_s' = (N_s', W_s')$. Αυτός ο γράφος είναι το σημείο αρχής των αλγορίθμων σύστασης που παρουσιάζονται:

Γειτονιά(Neighbourhood). Ως τρόπος απόκτησης των σχετικών κόμβων, ορίζουμε τον κόμβο γειτονιά ενός δεδομένου κόμβου n ως:

$$NH(n) = \{n_1, \dots, n_M | \, distance(n, n_m) < D_{max}, n_m \, \epsilon \, N \qquad 4.8$$

που είναι οι κόμβοι που βρίσκονται σε απόσταση D_{max} από το n, χωρίς να λαμβάνονται υπόψη οι σύνδεσμοι που τους συνδέουν άμεσα. Οι κόμβοι αυτοί κατά κάποιο τρόπο σχετίζονται με το n από τις ενέργειες των χρηστών, είτε επειδή οι χρήστες αλληλεπίδρασαν με το n μετά την αλληλεπίδραση με τους γειτονικούς κόμβους, ή επειδή υπάρχου κόμβοι με τους οποίους ο χρήστης αλληλεπίδρασε μετά τον n. Χρησιμοποιώντας τις ιδιότητες που προέρχονται από το γράφο έμμεσων πληροφοριών, μπορούμε να υπολογίσουμε τη συνολική αξία ενδιαφέροντος για ένα συγκεκριμένο κόμβο, η τιμή αυτή δείχνει την συσσωμάτωση του έμμεσου ενδιαφέροντος που έδωσαν οι χρήστες ιστορικά στο n, όταν συμμετείχε στις αλληλεπιδράσεις των χρηστών. Δεδομένων όλων των προσπιπτόντων σταθμισμένων συνδέσεων στο n, που ορίζονται από το υποσύνολο $W_s(G_s, n) = \{n_i, n_j, w \, | n_j = n\}, \; n \, \epsilon \, N_s$ η συνολική τιμή ενδιαφέροντος για το n υπολογίζεται ως εξής:

$$or(n) = \sum_{w \, \epsilon \, W_s(G_s, n)} w \qquad 4.9$$

Δεδομένης της τρέχουσας συνεδρίας ενός χρήστη και της δεξαμενής πληροφοριών έμμεσου ενδιαφέροντος τότε μπορούμε να ορίσουμε ως αξία σύστασης του κόμβου ως :

$$nr(n, N_s) = \sum_{n_i \, \epsilon N_s'} \; lr'(n_i).or(n) \mid n \, \epsilon \, NH(n_i) \qquad 4.10$$

όπου $lr'(n_i)$ είναι το *τοπικό ενδιαφέρον* για την τρέχουσα συνεδρία του χρήστη, χρησιμοποιώντας το υποσύνολο των δράσεων $A_s(G_s', n)$. Μπορούμε να ορίσουμε στη συνέχεια δύο διαφορετικές τιμές σύστασης: γειτονικού ερωτήματος $nh_q(n, N_{s\prime}) = nr(n, Q_{s'})|\ Q_{s'} \epsilon N_{s\prime}$, η οποία συνιστά κόμβους που σχετίζονται με τα πραγματικό ερωτήματα του χρήστη και , ομοίως, η γειτονικού έγγραφου $nh_q(n, N_{s\prime}) = nr(n, D_{s'})|\ D_{s'} \epsilon N_{s\prime}$, η οποία συνιστά κόμβους που σχετίζονται με τα έγγραφα που εμπλέκονται στην αλληλεπίδραση του χρήστη.

Ακολουθία Αλληλεπίδρασης. Αυτή η προσέγγιση σύστασης προσπαθεί να λάβει υπόψη τη διαδικασία αλληλεπίδρασης του χρήστη με σκοπό να συστήνει αυτούς τους κόμβους που ακολουθούν αυτή τη σειρά των αλληλεπιδράσεων. Για παράδειγμα, εάν ένας χρήστης έχει ανοίξει ένα βίντεο ειδήσεων, η σύσταση θα μπορούσε να περιέχει περισσότερες σε βάθος ιστορίες που βρήκαν ενδιαφέρον προηγούμενοι χρήστες να δουν. Ορίζεται ως εξής:

$$is(n, N_{S'}) = \sum_{n_i \in N_{S'}}((lr'(n_i)\,.\,\xi^{l-1}.w)\ |\ \exists\, p = n_i \rightarrow n_j \rightarrow n, \qquad 4.11$$

$$w \epsilon \{n_j, n, w\},$$

$$l = length(p), l < L_{MAX}$$

όπου p είναι η διαδρομή μεταξύ του κάθε κόμβου n_i και κόμβου n, λαμβάνοντας υπόψη την κατευθυντικότητα της σύνδεσης τους. l είναι το μήκος της διαδρομής (υπολογίζεται ως ο αριθμός των συνδέσεων), έχοντας μια απόσταση είναι μικρότερη από ένα μέγιστο μήκος L_{MAX}. Τέλος, ξ είναι ένας παράγοντας μείωσης μήκους, που η τιμή του ορίζεται βάση πειραμάτων.

Προορισμός ερωτήματος. Αυτός ο αλγόριθμος αφορά τα ερωτήματα αναζήτησης και τα *μονοπάτια αναζήτησης*. *Μονοπάτι αναζήτησης* είναι η αλληλουχία ενεργειών ενός χρήστη από την πρώτη ως την τελευταία ενεργεία του που αφορά ένα συγκεκριμένο ερώτημα. Ο συγγραφέας στο [57] δείχνει ότι τα τελευταία έγγραφα που επισκέπτεται ο χρήστης μέσα μια αναζήτηση σε μία συνεδρία έχουν υψηλή συνάφεια. Επιλέγεται ένα μέτρο για τον προορισμό του ερωτήματος. Η αξία του ερωτήματος προορισμού κατατάσσεται με βάση τη δημοτικότητά ανάμεσα στους προορισμούς στα *μονοπάτια αναζήτησης* και ορίζεται στο [58] ως εξής:

$$qd(q, d) = S(d, q).\sum_p w\ |\ \exists\, p = q \rightarrow d_j \rightarrow n \rightarrow n_q \qquad 4.12$$

$$d_j, d \epsilon D_S, n_q \epsilon Q_S,$$

$$w \epsilon \{ d_j, d, w\}$$

όπου $S(d, q)$ είναι το μέτρο $tf.idf$ ομοιότητας μεταξύ του εγγράφου d και του τελευταίου ερωτήματος $q \epsilon last(Q_{S'})$ εισόδου από το χρήστη. Οι δεσμοί μεταξύ εγγράφων στο WDG γράφο είναι αλληλουχία συνδέσμων, και μπορεί να αφορούν ποικίλους τύπους ενεργειών. Η αξία δημοτικότητας ορίζεται από τη συνάθροιση των βαρών όλων των προσπιπτόντων συνδέσμων

μέσα στα μονοπάτια μεταξύ των q και d που έχουν προκύψει από όλα τα διαφορετικά *μονοπάτια αναζήτησης*.

Τυχαίος περίπατος(Random Walk). Οι συγγραφείς στο [59] αξιοποιούν τα δεδομένα των κλικ χρησιμοποιώντας ένα αλγόριθμο *τυχαίου περίπατου*[60]. Ο *τυχαίος περίπατος* είναι μία μαθηματική διατύπωση μίας πορείας που αποτελείται από διαδοχικά τυχαία βήματα. Η τυχαία κίνηση πραγματοποιείται μέσα σε μία εξαρχής ορισμένη περιοχή. Ο χρήστης κινείται από μία θέση προς μία άλλη επιλέγοντας τυχαία κατεύθυνση και ταχύτητα. Στην καινούρια θέση επιλέγεται μια κατεύθυνση για το χρήστη μη λαμβάνοντας πλέον υπόψη την προηγούμενη θέση. Ο υπολογισμός του τυχαίου περιπάτου θα καταλήξει, θεωρητικά, με υψηλότερη πιθανότητα για τους κόμβους αυτούς που βρήκαν προηγούμενοι χρήστες (εμμέσως) σχετικούς μετά το ερώτημά τους (*τυχαίος περίπατος προς τα εμπρός*) ή για τα έγγραφα αυτά που έχουν την αναγκαία πληροφορία του ερωτήματος (*τυχαίος περίπατος προς τα πίσω*). Για τον υπολογισμό αυτό, απαιτείται μια πιθανότητα της μετάβασης από τον κόμβο n_k στον n_j :

$$P_{t+1|t}(n_k\,|n_j\,) = \begin{cases} \frac{(1-s)C_{jk}}{\sum_i C_{ji}}, & \forall\, k \neq j \\ s, & when\, k = j \end{cases} \qquad 4.13$$

όπου s είναι η πιθανότητα να μείνει στο ίδιο κόμβο και ο αριθμός των κλικ είναι $C_{ij} = \{n_i\,, n_j, w\}$, λαμβάνοντας έτσι υπόψη τη συσσωμάτωσης των πληροφοριών έμμεσου ενδιαφέροντος. Χρησιμοποιώντας αυτές τις πιθανότητες, υπολογίζεται ο *τυχαίος περίπατος προς τα εμπρός* $rw_F\,(q)$ και ο *τυχαίος περίπατος προς τα πίσω* $rw_B(q),\ q \in\ last(Q_{S'})$.

4.2 Ρητή ανατροφοδότηση (Explicit Feedback)

Η ρητή ανατροφοδότηση λαμβάνεται από χρήστες που δείχνουν το ενδιαφέρον τους σχετικά με ένα αντικείμενο που ανακτάται για μία αναζήτηση. Αυτό το είδος της ανάδρασης ορίζεται ως ρητή μόνο όταν οι χρήστες/αξιολογητές (ή άλλοι χρήστες του συστήματος) γνωρίζουν ότι η ανατροφοδότηση που παρέχεται ερμηνεύεται ως εκδήλωση ή μη ενδιαφέροντος.

Οι χρήστες μπορούν να αναφέρουν ρητά το ενδιαφέρον τους χρησιμοποιώντας είτε ένα δυαδικό σύστημα ή είτε ένα διαβαθμισμένο. Η δυαδική ρητή ανατροφοδότηση δηλώνει ότι ένα αντικείμενο είναι είτε σχετικό ή άσχετο για ένα συγκεκριμένο ερώτημα. Η διαβαθμισμένη ρητή ανατροφοδότηση δείχνει τη σχετικότητα ενός αντικειμένου σε ένα ερώτημα σε μια κλίμακα που χρησιμοποιεί αριθμούς, γράμματα, ή περιγραφές (όπως "δεν έχει σημασία", "κάπως σχετικό", "σχετικό", ή "πολύ σχετικό"). Μπορεί επίσης να λάβει τη μορφή μία κύριας κατάταξης των αντικειμένων που έχουν δημιουργηθεί από το χρήστη; ο χρήστης δηλαδή τοποθετεί τα αντικείμενα του αποτελέσματος με σειρά ενδιαφέροντος (συνήθως φθίνουσα). Ένα παράδειγμα αυτού είναι το SearchWiki που εφαρμόστηκε από την Google στην ιστοσελίδα αναζήτησή τους.

Οι πληροφορίες ανατροφοδότησης πρέπει να παρεμβληθούν με την αρχική αναζήτηση για τη βελτίωση της απόδοσης ανάκτησης, όπως ο αλγόριθμος Rocchio. Η συνάρτηση του αλγορίθμου είναι η εξής:

$$\overrightarrow{Q_m} = (a \cdot \overrightarrow{Q_o}) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right) \quad , \qquad 4.14$$

και οι τιμές των μεταβλητών δίνονται στον παρακάτω πίνακα:

Μεταβλητή	Τιμή
$\overrightarrow{Q_m}$	*Τροποποιημένο διάνυσμα*
$\overrightarrow{Q_o}$	*Αρχικό διάνυσμα*
$\overrightarrow{D_j}$	*Διάνυσμα σχετικού αντικειμένου*
$\overrightarrow{D_k}$	*Διάνυσμα μη σχετικού αντικειμένου*
a	*Βάρος αρχικού αντικειμένου*
b	*Βάρος σχετικού αντικειμένου*
c	*Βάρος μη σχετικού αντικειμένου*
D_r	*Σύνολο σχετικών αντικειμένων*
D_{nr}	*Σύνολο μη σχετικών αντικειμένων*

Πίνακας 4.1: Τιμές μεταβλητών αλγόριθμου Rocchio

Όπως καταδεικνύεται στον τύπο Rocchio, τα συνδεδεμένα βάρη (a, b, c) είναι υπεύθυνα για τη διαμόρφωση του τροποποιημένου διανύσματος σε μια κατεύθυνση πιο κοντά ή πιο μακριά από το αρχικό ερώτημα, τα σχετικά και τα μη σχετικά έγγραφα. Ειδικότερα, οι τιμές για b και c θα πρέπει να αυξηθούν ή να μειωθούν ανάλογα με το σύνολο των εγγράφων που έχουν αξιολογηθεί από το χρήστη. Εάν ο χρήστης αποφασίσει ότι το τροποποιημένο ερώτημα δεν πρέπει να περιέχει όρους, είτε από το αρχικό ερώτημα, τα σχετικά έγγραφα, ή τα μη σχετικά, τότε η τιμή του αντίστοιχου βάρους (a, b, c) για την κατηγορία πρέπει να ρυθμιστεί στο 0. Στο τελευταίο μέρος του αλγορίθμου, οι μεταβλητές D_r, και D_{nr} παρουσιάζονται ως σύνολα διανυσμάτων που περιέχουν τις συντεταγμένες των σχετικών και μη εγγράφων. Τα $\overrightarrow{D_j}$και $\overrightarrow{D_k}$ είναι οι φορείς που χρησιμοποιούνται για επαναλήψεις μέσα στα δύο σύνολα και σχηματίζουν διανύσματα αθροίσεων. Τα ποσά αυτά είναι κανονικοποιημένα ως προς το μέγεθος του αντίστοιχου συνόλου εγγράφων τους (D_r, D_{nr}). Όσο τα βάρη αυξάνονται ή μειώνονται για μια συγκεκριμένη κατηγορία εγγράφων, οι συντεταγμένες για το τροποποιημένο διάνυσμα αρχίζουν να κινούνται είτε πιο κοντά, ή πιο μακριά από το κέντρο βάρους της συλλογής του εγγράφου. Έτσι, αν το βάρος αυξάνεται για τα συναφή αντικείμενα, τότε οι συντεταγμένες του τροποποιημένου φορέα θα αναπαρίστανται πιο κοντά στο κέντρο βάρους των σχετικών εγγράφων.

Όπως αναφέρθηκε οι ρητές εκτιμήσεις δείχνουν πόσο σημαντικό ή ενδιαφέρον ένα στοιχείο είναι για το χρήστη. Υπάρχουν τρεις κύριες προσεγγίσεις για ρητή ανατροφοδότηση:

• *like / dislike* (μου αρέσει / δεν μου αρέσει) - τα στοιχεία που ταξινομούνται ως "σχετικά" ή "μη σχετικά" υιοθετώντας μία απλή δυαδική κλίμακα διαβάθμισης,

• αξιολογήσεις / βαθμολογίες - μία διακριτή αριθμητική κλίμακα υιοθετείται συνήθως για να βαθμολογούνται τα αντικείμενα, όπως το σύστημα των 5-αστέρων (*5-stars*)

• κείμενο με σχόλια - Τα σχόλια για ένα μόνο στοιχείο συλλέγονται και παρουσιάζονται στους χρήστες ως μέσο διευκόλυνσης της διαδικασίας λήψης αποφάσεων, όπως στο [72]. Για παράδειγμα, τα σχόλια των πελατών του στο Amazon.com ή eBay.com θα μπορούσαν να βοηθήσουν τους χρήστες να αποφασίσουν αν ένα στοιχείο έχει εκτιμηθεί από την κοινότητα. Τα σχόλια υπό μορφή κειμένου είναι χρήσιμα, αλλά μπορεί να υπερφορτώσει το χρήστη γιατί πρέπει να διαβάσει και να ερμηνεύει κάθε σχόλιο για να αποφασίσει αν είναι θετικό ή αρνητικό, και σε ποιο βαθμό.

Η ρητή ανατροφοδότηση ενώ έχει το πλεονέκτημα της απλότητας, αυξάνει το γνωστικό φορτίο του χρήστη, ενώ δεν αρκεί για να πιάσει το συναίσθημα του χρήστη σχετικά με τα αντικείμενα.

4.3 Ψευδο – ανατροφοδότηση (Pseudo Relevance Feedback)

Η ψευδο-ανατροφοδότηση, επίσης γνωστή ως τυφλή ανατροφοδότηση, παρέχει μια μέθοδο για την αυτόματη ανάλυση των αποτελεσμάτων. Αυτοματοποιεί μέρος της ανατροφοδότησης, έτσι ώστε ο χρήστης να παίρνει βελτιωμένη απόδοση ανάκτησης αποτελεσμάτων χωρίς εκτεταμένη αλληλεπίδραση. Η διαδικασία είναι η εξής:

- Λαμβάνονται τα αποτελέσματα που επιστρέφονται από ένα αρχικό ερώτημα με την υψηλότερη βαθμολογία (μόνο τα κορυφαία k, με k μεταξύ 10 και 50 στα περισσότερα πειράματα).
- Επιλέγονται οι περισσότερο χρησιμοποιούμενοι όροι από αυτά τα έγγραφα, χρησιμοποιώντας για παράδειγμα $tf - idf$ (term frequency–inverse document frequency) βάρη.
- Γίνεται επέκταση του αρχικού ερωτήματος, με την πρόσθεση αυτών των όρων στο ερώτημα, και τελικά επιστρέφονται στο χρήστη τα νέα πιο σχετικά αποτελέσματα.

Αυτή η αυτόματη τεχνική λειτουργεί ικανοποιητικά ως επί το πλείστον για «καλά» αρχικά ερωτήματα. Με την επέκταση ερωτήματος, μερικά αρχικά σχετικά έγγραφα που δεν επιστρέφονται μπορεί στη συνέχεια να ανακτηθούν για να βελτιωθεί η συνολική απόδοση. Σαφώς, η επίδραση αυτής της μεθόδου βασίζεται σε μεγάλο βαθμό στην ποιότητα των

επιλεγμένων όρων επέκτασης. Όπως είναι λογικό ενέχει τους κινδύνους μιας αυτόματης διαδικασίας. Για παράδειγμα, εάν το ερώτημα είναι για τα ορυχεία χαλκού και τα κορυφαία αποτέλεσμα είναι όλα σχετικά με τα ορυχεία στη Χιλή, τότε μπορεί να υπάρξει μετατόπιση του ερωτήματος προς την κατεύθυνση των εγγράφων για τη Χιλή. Επιπλέον, αν οι λέξεις που προστίθενται στο αρχικό ερώτημα είναι άσχετες με το θέμα του ερωτήματος, η ποιότητα της ανάκτησης είναι πιθανό να υποβαθμιστεί, ιδίως σε αναζήτηση στο διαδίκτυο όπου τα έγγραφα συχνά καλύπτουν πολλαπλά διαφορετικά θέματα.

5. Μηχανισμός του Συστήματος Σύστασης

5.1 Κοινωνικό και εξατομικευμένο Εργαλείο (Social and Personalization Tool – SP Tool)

Στα πλαίσια της διπλωματικής αυτής εργασίας υλοποιήθηκε ένα εξατομικευμένο εργαλείο με σκοπό τη παραγωγή συστάσεων. Οι εξατομικευμένες πληροφορίες των χρηστών συγκεντρώνονται με βάση μηχανισμούς ανατροφοδότησης ενδιαφέροντος αλλά και προσωπικών πληροφοριών και πληροφοριών από μέσα κοινωνικής δικτύωσης. Οι μηχανισμοί αυτοί ενσωματώνονται στο πλαίσιο ενός κοινωνικού και εξατομικευμένου εργαλείου (*Social and Personalization tool – SP tool*). Αυτό το εργαλείο θα δώσει ανατροφοδότηση για τη δημιουργία και την παραγωγή, αυτόματης διαδικασίας επεξεργασίας του περιεχομένου των πολυμέσων, ώστε να εμπλουτίζεται με βάση τις προτιμήσεις και τις ανάγκες των χρηστών. Το εργαλείο αυτό ανασύνταξης και σχολιασμού του πολυμεσικού περιεχομένου, προσφέρει αυξημένη ευελιξία σε εταιρείες παραγωγής, όπως οι ραδιοτηλεοπτικοί οργανισμοί, οι διαφημιστικές εταιρείες, οι οποίες θα μπορούν να επεξεργαστούν τα αντίστοιχα σχόλια, για να συνθέσουν και να δημιουργήσουν διαφημιστικές καμπάνιες, κ.λ.π. Αυτές οι διαδικασίες και η αποτελεσματικότητα αυτών μπορεί να βελτιωθεί με την υιοθέτηση και τη χρήση εξατομικευμένων πληροφοριών, που προέρχονται από τον τελικό χρήστη/καταναλωτή. Έτσι, στοχεύουμε στην παροχή ενός κοινωνικού και εξατομικευμένου εργαλείου (SP), το οποίο θα ενσωματώνει το ενδιαφέρον των χρηστών, με τους μηχανισμούς εξατομίκευσης και σύστασης με βάση τα κοινωνικά δίκτυα. Το εργαλείο επιτρέπει, μέσω web υπηρεσιών, πρόσβαση στις προτιμήσεις των χρηστών και την ανατροφοδότηση του παρέχουν, δίνοντας έτσι πληροφορίες για ομαδοποίηση χρηστών και των τάσεων κατά τον κύκλο ζωής του περιεχομένου. Το αποτέλεσμα του εργαλείου SP θα ενεργεί ως είσοδος και θα παρέχει συμπληρωματικές εξατομικευμένες πληροφορίες, προκειμένου να ενισχυθεί η λειτουργία των προαναφερόμενων εργαλείων, προς παροχή περιεχομένου και μετα-δεδομένων που απευθύνονται στις προτιμήσεις των χρηστών. Σύμφωνα με την προσέγγιση της εξατομικευμένης πληροφορίας, το συνολικό ολοκληρωμένο εργαλείο θα βελτιώσει την ποιότητα εμπειρίας του τελικού χρήστη, αλλά και μέσω της στόχευσης χρηστών που ενδιαφέρονται περισσότερο για ένα προϊόν θα οδηγήσει σε αύξηση εσόδων μικρομεσαίων εταιρειών. Πιο συγκεκριμένα, το εργαλείο SP θα είναι ένα αυτόνομο εργαλείο, το οποίο θα αλληλεπιδρά και θα ανταλλάσσει πληροφορίες με τα άλλα εργαλεία και θα περιλαμβάνει τρεις βασικούς μηχανισμούς:

1. Ανατροφοδότηση σχετικότητας (Relevance Feedback Mechanism)

2. Εξατομίκευση (Personalization Mechanism)

3. Σύσταση με βάση τα κοινωνικά δίκτυα (Social Recommendation)

Ο τελικός χρήστης/καταναλωτής θα είναι σε θέση να βλέπει τα βίντεο συνεχούς ροής εμπλουτισμένα με τα κατάλληλα, και στοχευμένα στις προτιμήσεις του, μεταδεδομένα και σχολιασμούς, καθώς και διαφημίσεις σε περιβάλλοντα πολλαπλών οθονών (multi-screen). Για παράδειγμα, ο χρήστης θα παρακολουθεί ένα βίντεο από την τηλεόραση του/της και θα μπορεί να επιλέγει μεταξύ κάποιων προκαθορισμένων δυνατοτήτων σχολιασμού και μετα-δεδομένων, όπως π.χ. να δει την ιστοσελίδα Wikipedia στο κινητό του/της, και θα είναι σε θέση να διαβάσει περισσότερες πληροφορίες για αναφορές, εικόνες, τοπία, πρόσωπα κ.λ.π που βρίσκονται στο βίντεο. Μετά αυτός/αυτή ακούει ένα παραδοσιακό τραγούδι της χώρας στην οποία ανήκει το τοπίο του βίντεο, μέσω Youtube κ.λ.π., και παράλληλα εμφανίζονται στο multi-screen περιβάλλον είτε banner διαφημίσεις ή διαφημίσεις με μορφή βίντεο που σχετίζονται με το βίντεο και θα στοχεύουν στις προτιμήσεις του τελικού χρήστη. Λαμβάνοντας υπόψη αυτό το αντιπροσωπευτικό παράδειγμα, περιγράφεται η λειτουργία του μηχανισμού ανατροφοδότησης σχετικότητας, ένα από τους μηχανισμούς που αποτελούν το εργαλείο SP και στο οποίο επικεντρώνεται η παρούσα διπλωματική.

Ανατροφοδότηση σχετικότητας. Κατά τα τελευταία έτη το ποσό των πολυμέσων και οι πληροφορίες μεταδεδομένων που είναι διαθέσιμα στο διαδίκτυο, στις εταιρείες παραγωγής, στους ραδιοτηλεοπτικούς φορείς, κ.λ.π. έχει αυξηθεί εκθετικά. Ως εκ τούτου, είναι σημαντικό να εξετάζονται οι προτιμήσεις των χρηστών προκειμένου να καταλήξουν με το πιο σχετικό και ελκυστικό περιεχόμενο πολυμέσων και μεταδεδομένων που θα στοχεύει στις ανάγκες και επιθυμίες των χρηστών. Ο μηχανισμός ανάδρασης σχετικότητας/ενδιαφέροντος είναι μια πολλά υποσχόμενη λύση, προκειμένου να καταγράψει τις προτιμήσεις των χρηστών και να δημιουργήσει προφίλ χρηστών, ώστε να προσδιορίσει και να υποδείξει στις διαδικασίες παραγωγής βίντεο και διαφημίσεων τα πιο επιθυμητά στοιχεία και μετα-δεδομένα των βίντεο από την άποψη των χρηστών. Πιο συγκεκριμένα, οι online εκδοτικές εταιρείες, καθώς και οι διαφημιστικές εταιρείες, θα αποκτήσουν γνώση από την ανατροφοδότηση ενδιαφέροντος των χρηστών, προκειμένου να παράσχουν στοχευμένους σχολιασμούς και διαφημίσεις. Ο μηχανισμός ανατροφοδότησης (RF) που προτείνεται και αποτελεί το εργαλείο SP θα χρησιμεύσει ως σημείο αναφοράς και αφετηρία για την οικοδόμηση της γνώσης σχετικά με αυτό το λειτουργικό σκέλος. Πιο αναλυτικά, οι χρήστες θα είναι σε θέση να δουν το περιεχόμενο πολυμέσων, π.χ. βίντεο με σχολιασμούς και μεταδεδομένα, που θα ονομάζουμε εμπλουτισμούς. Επιπλέον, οι διαφημιστικές εταιρείες θα είναι υπεύθυνες να παρουσιάσουν στοχευμένες διαφημίσεις στον τελικό χρήστη/καταναλωτή που σχετίζονται με τα βίντεο που μεταδίδονται, αλλά και οι online εκδοτικές εταιρείες να απεικονίσουν ποικίλες πληροφορίες. Οι χρήστες θα είναι σε θέση να δώσουν ανατροφοδότηση, εκφράζοντας τη συνάφεια των εμπλουτισμών που προβάλλονται ως προς την προσδοκία τους και την προτίμηση τους. Ο κύριος στόχος αυτού του μηχανισμού ανατροφοδότησης είναι να προσδιορίσει και να βελτιώσει τις πιο σχετικές πληροφορίες για τις προτιμήσεις των χρηστών και να προσαρμόσει αναλόγως την παραγωγή, σύνταξη και διαδικασία απεικόνισης στο multi-screen περιβάλλον.

Πιο συγκεκριμένα, οι προσεγγίσεις στο μηχανισμό RF περιλαμβάνουν τα αποτελέσματα της ανατροφοδότησης, είτε με δυαδική είτε βαθμωτή βαθμολογία, που δηλώνεται από τον τελικό

χρήστη. Επιπλέον, αυτό το σκορ RF θα πρέπει να συνδεθεί με τις πληροφορίες του τύπου της οθόνης/συσκευής με την οποία ο χρήστης έχει πρόσβαση στο περιεχόμενο πολυμέσων. Έτσι, ενισχύονται και συνδυάζονται οι πληροφορίες σχετικά με τις προτιμήσεις χρηστών για το περιεχόμενο πολυμέσου και τους εμπλουτισμούς, οι οποίες θα πρέπει να αποθηκεύονται σε κατάλληλα οργανωμένο σύστημα βάσης δεδομένων, καθώς και οι συσκευές χρήσης από τις οποίες έγινε προεπισκόπηση του περιεχομένου και των σχετικών πληροφοριών (π.χ. ο χρήστης επέλεξε να παρακολουθήσει ένα βίντεο στην τηλεόραση, την ιστοσελίδα της Wikipedia στο κινητό του και την ιστοσελίδα του YouTube στο tablet του).

5.2 Αρχιτεκτονική συστήματος

Στο κεφάλαιο αυτό θα γίνει μία περιγραφή του συστήματος που έχει υλοποιηθεί στα πλαίσια της διπλωματικής για την καταγραφή και αξιοποίηση των σημάτων ανάδρασης των χρηστών του συστήματος. Το σύστημα αυτό απαρτίζεται από ένα μία διεπαφή προγραμματισμού εφαρμογών (Application Programming Interface - API), μία μηχανή συστάσεων (Recommender Engine) και μία βάση δεδομένων (Database – DB). Οι χρήστες εγγράφονται/συνδέονται στο σύστημα και καταχωρούν κάποιες αρχικές πληροφορίες για τις προτιμήσεις τους και στη συνέχεια περιηγούνται στο σύστημα. Το API είναι υπεύθυνο για την ανάκτηση των διαφόρων σημάτων αλληλεπίδρασης του χρήστη με το σύστημα, τα οποία περιγράφονται αναλυτικά σε επόμενη ενότητα, η βάση δεδομένων για την καταχώρηση των σημάτων αυτών σε κατάλληλους πίνακες και το σύστημα συστάσεων για την ανάλυση αυτών και την εξαγωγή κατάλληλων αποτελεσμάτων.

5.2.1 Βάση δεδομένων

Οι βάσεις δεδομένων αποτελούν σημαντικό τμήμα της υλοποίησης καθώς σε αυτές αποθηκεύονται τα video , οι εμπλουτισμοί και οι διαφημίσεις τους καθώς και οι ενέργειες του κάθε χρήστη για καθένα από αυτά. Το πλεονέκτημα των βάσεων είναι η άμεση ανάκτηση οποιασδήποτε πληροφορίας χωρίς να χρειάζεται κάποια διεργασία άρα και υπολογιστική ισχύς του συστήματος. Χρησιμοποιήθηκε το πακέτο προγραμμάτων ελεύθερου λογισμικού ανοικτού κώδικα XAMPP, που είναι ανεξαρτήτου πλατφόρμας και το οποίο περιέχει τον εξυπηρετητή ιστοσελίδων http Apache , την βάση δεδομένων MySQL και ένα διερμηνέα για κώδικα γραμμένο σε γλώσσες προγραμματισμού PHP και Perl. MySQL είναι ένα ελεύθερο σύστημα διαχειρίσεως βάσεων δεδομένων που είναι ευρύτατα διαδεδομένη και υποστηρίζει τα τελευταία standards της SQL. Αρχικά δημιουργήθηκε μία βάση δεδομένων ‘videos’ και η σύνδεση σε αυτή. Στη συνέχεια δημιουργήθηκαν οι πίνακες για τη συλλογή των video και των εμπλουτισμών και διαφημίσεων τους και για την ενημέρωση του προφίλ του χρήστη. Τα βασικότερα πεδία του πίνακα ‘videos’ ,

που χρησιμοποιήθηκαν για τη συλλογή των σημάτων ανατροφοδότησης των χρηστών, είναι τα εξής:

username	Το μοναδικό όνομα του χρήστη
device	Ο τύπος συσκευής από την οποία έγινε μία ενέργεια
video_id	Το μοναδικό αναγνωριστικό του βίντεο
Action	Η ενέργεια του χρήστη
video_time	Ο συνολικός χρόνος του βίντεο
start_time	Η χρονική στιγμή έναρξης προβολής του βίντεο
stop_time	Η χρονική στιγμή λήξης της προβολής του βίντεο
enrichment_id	Το αναγνωριστικό του εμπλουτισμού του βίντεο
ad_id	Το αναγνωριστικό της διαφήμισης του βίντεο
explicit_rf	Η ρητή ανατροφοδότηση του χρήστη για το βίντεο
watch_time	Ο χρόνος προβολής του βίντεο

Πίνακας 5.1: Τα βασικότερα πεδία του πίνακα 'videos'

Χρήστες του συστήματος. Πρόκειται για μια οντότητα η οποία αντιπροσωπεύει ένα σύνολο ανθρώπων που αποκτούν λογαριασμό στο σύστημα μας και διαμορφώνουν ένα προσωπικό προφίλ. Το προφίλ απαρτίζεται από δημογραφικά στοιχεία, προτιμήσεις του χρήστη για διάφορες θεματικές ενότητες και δεδομένα από την αλληλεπίδραση με το σύστημα. Οι προτιμήσεις αυτές μπορούν να πάρουν ένα σύνολο από διαφορετικές τιμές.

5.2.2 Το API

Για τη δημιουργία του API της διπλωματικής χρησιμοποιήθηκε το εργαλείο Laravel, ένα πλαίσιο web εφαρμογών PHP ανοιχτού κώδικα. Αρχικά, γίνεται η σύνδεση με τη βάση δεδομένων που δημιουργήθηκε μέσω του app/config/database.php. Έπειτα δημιουργήθηκαν τα migration αρχεία που χρησιμοποιούνται για τη δημιουργία πινάκων στη βάση δεδομένων. Δημιουργούνται στη συνέχεια κάποιοι αρχικοί χρήστες για σκοπούς δοκιμής. Επόμενο βήμα είναι η δημιουργία ενός μοντέλου για διαχείριση της βάσης δεδομένων μας. Με το μοντέλο αυτό γίνεται πιο εύκολα η επικοινωνία με τη βάση μας, τροποποιήσεις, εισαγωγές δεδομένων κλπ. Κατασκευάστηκαν στη συνέχεια δύο routes που δέχονται το URI *'feedback'* και ένα ελεγκτή τον *VideoController*. Ο ελεγκτής είναι εκείνος που περιέχει τις βασικές συναρτήσεις του API και καλείται μέσω των routes. Παρακάτω φαίνεται η σύνταξη των routes. Το πρώτο route θα καλεστεί στην περίπτωση των GET μεθόδων και θα εκτελέσει την *index* συνάρτηση του ελεγκτή, ενώ το δεύτερο καλείται για να εξυπηρετήσει τα POST αιτήματα και εκτελεί την *store* συνάρτηση του ελεγκτή.

```
Route::get('feedback/{username}', 'VideoController@index');
```

```
Route::post('feedback', 'VideoController@store');
```
Ο ελεγκτής *VideoController* έχει τις εξής συναρτήσεις:

Function index: η συνάρτηση αυτή επιστρέφει τις εγγραφές του παραπάνω πίνακα ‘videos’.

Function store: η συνάρτηση αυτή παίρνει από την είσοδο το type και εξετάζει αν είναι video, enrichment ή ad. Ανάλογα με το type καλούνται οι συναρτήσεις video_function , enrichment_function, ad_function αντίστοιχα.

Video_function: η συνάρτηση αυτή παίρνει από την είσοδο τα username, device, video_id και action και τα αποθηκεύει σε καταχωρητές. Ελέγχει το action αν είναι start ή stop.

- Αν είναι start κάνει εγγραφή στον πίνακα της βάσης δεδομένων μας με όλα τα στοιχεία της εισόδου.
- Αν είναι stop αναζητεί στον πίνακα την εγγραφή με ίδια username, device και video_id και παίρνει την τελευταία εγγραφή (αφού μετά από start θα ακολουθεί σίγουρα stop).Παίρνει από την είσοδο το stop_time και ανανεώνει τη στήλη stop_time με την τιμή αυτή. Έτσι υπάρχει συνολικά μία εγγραφή για το video αυτό. Στη συνέχεια παίρνει τις τιμές start_time , stop_time και video_time και υπολογίζει το ποσοστό του video που είδε ο χρήστης , watch_time=((stop_time-start_time)/video_time).
- Αν είναι bookmark αναζητεί την ίδια εγγραφή με πριν και ανανεώνει τη στήλη ‘bookmark’ με την τιμή 1.

Enrichment_function: η συνάρτηση αυτή παίρνει από την είσοδο τα username, device, video_id, enrichment_id και action και τα αποθηκεύει σε καταχωρητές. Ελέγχει το action αν είναι click, close ή share.

- Αν είναι click κάνει εγγραφή στον πίνακα της βάσης δεδομένων μας με όλα τα στοιχεία της εισόδου.
- Αν είναι close αναζητεί στον πίνακα την εγγραφή με ίδια username, device ,video_id και enrichment_id και παίρνει την τελευταία εγγραφή. Παίρνει από την είσοδο το stop_time και ανανεώνει τη στήλη stop_time με την τιμή αυτή. Έτσι υπάρχει συνολικά μία εγγραφή για το enrichment αυτό. Στη συνέχεια παίρνει τις τιμές start_time και stop_time και υπολογίζει πόση ώρα είδε ο χρήστης το συγκεκριμένο enrichment.
- Αν είναι share αναζητεί ,με τα ίδια στοιχεία με πριν, την εγγραφή του πίνακα και ανανεώνει τη στήλη share με την τιμή 1. (Η στήλη αυτή παίρνει boolean τιμές μόνο)

Ad_function: Παίρνει όλα τα στοιχεία της εισόδου και κάνει νέα εγγραφή στον πίνακα αν action=click.

Οι παραπάνω συναρτήσεις παίρνουν σαν όρισμα το *Videorequest $request*, το οποίο περιέχει τους κανόνες που πρέπει να πληρούν οι τιμές εισόδου που δίνονται και να επιστρέφει κατάλληλο μήνυμα λάθους σε περίπτωση που δεν πληρούνται. Παράδειγμα κανόνων της κλάσης *Videorequest* είναι να είναι μοναδικό το username και να περιέχει αποκλειστικά αριθμούς και χαρακτήρες.

Για τη δοκιμή του API, δίνοντας τις εισόδους των παραπάνω συναρτήσεων χρησιμοποιήθηκε το εργαλείο Postman. Με το εργαλείο αυτό δίνεται η δυνατότητα γρήγορης δημιουργίας αιτημάτων (request), εναλλαγή των πεδίων και τιμών εισόδου αλλά και η δοκιμή των τεχνικών πιστοποίησης (authentication) των χρηστών.

Πιστοποίηση χρηστών. Η πιστοποίηση χρησιμοποιείται για τη σύνδεση χρηστών σε μια εφαρμογή ή έναν ιστοχώρο που ενσωματώνει στοιχεία και χρήστες. Έχει πολλά χαρακτηριστικά ασφαλείας για την προστασία προσωπικών πληροφοριών, επιτρέποντας στους χρήστες να ελέγχουν τι μοιράζονται και στους προγραμματιστές να ζητήσουν ασφαλή πρόσβαση στις πληροφορίες αυτές. Χρησιμοποιήθηκε η βασική πιστοποίηση πρόσβασης (basic access authentication), η οποία από τη μεριά του χρήστη απαιτεί username και password τα οποία κωδικοποιούνται με Base64. Στη συνέχεια, χρησιμοποιούνται οι *stream_context_create* και *file_get_contents* με όρισμα την παραπάνω πιστοποίηση για την ανάκτηση των εμπλουτισμών και των βίντεο.

Κλήσεις στο API. Το API βασίζεται σε HTTP αιτήματα που αντιστοιχίζονται σε HTTP μεθόδους. Η μέθοδος GET χρησιμοποιείται για ανάκτηση των πληροφοριών των χρηστών και η μέθοδος POST για τροποποίηση και προσθήκη πληροφοριών. Για παράδειγμα, για ένα χρήστη με αναγνωριστικό username για τον οποίο θέλουμε να ανακτήσουμε τα στοιχεία του χρησιμοποιούμε το παρακάτω route:

```
Route::get('feedback', 'VideoController@index');
```

το οποίο καλεί τη μέθοδο *index* του ελεγκτή *VideoController*.

Το URL θα έχει τη μορφή: http:\\localhost:port_number\feedback\{username} και χρησιμοποιώντας το Postman επιλέγουμε μέθοδο GET η οποία εκτελεί το HTTP αίτημα. Το αίτημα αυτό επιστρέφει HTTP/1.1 200 OK και όλα τα διαθέσιμα πεδία για το username αυτό. Σημειώνεται ότι όλα τα αντικείμενα μπορούν να εμφανίζονται σε json μορφή. Για την εκτέλεση κάποιας ενέργειας εισαγωγής ή τροποποίησης δεδομένων χρησιμοποιείται η μέθοδος POST, η οποία σε αντίθεση με την GET δεν μας επιστρέφει κάποιο αποτέλεσμα. Για παράδειγμα, για την εισαγωγή ενός βίντεο στη βάση δεδομένων, το οποίο έχει αρχίσει να παρακολουθεί ο χρήστης στη συσκευή του χρησιμοποιείται το route:

```
Route::post('feedback', 'VideoController@store');
```

Στην εφαρμογή Postman όπως προαναφέρθηκε δίνουμε είσοδο τα πεδία username, type, video_id, device, action, start_time, video_time με τιμές για παράδειγμα user1, video, 1, mobile, start, 1, 4 και το url παίρνει το μορφή:

http:\\localhost:port_number\feedback?username=user1&type=video&video_id=1&device=mobile&action=start&start_time=1&video_time=4

και επιλέγοντας send γίνεται μία εγγραφή στον πίνακα μας. Έστω ότι η επόμενη κίνηση του χρήστη είναι η διακοπή του βίντεο. Στην περίπτωση αυτή δημιουργούμε ένα POST αίτημα όπως προηγουμένως με ακριβώς τα ίδια ορίσματα εκτός από το όρισμα action το οποίο τώρα έχει την τιμή stop και δίνεται και το όρισμα stop_time που έχει την χρονική στιγμή διακοπής του βίντεο.

http:\\localhost:port_number\feedback?username=user1&type=video&video_id=1&device=mobile&action=stop&stop_time=3&video_time=4

Η συνάρτηση που εκτελείται ελέγχει τον πίνακα για την τελευταία εγγραφή με τα ορίσματα username, type, video_id, device. Υπολογίζει το χρόνο προβολής του βίντεο (stop_time – start_time), διαιρεί με τον συνολικό χρόνο του βίντεο και τελικά κάνει update στην εγγραφή του πίνακα δίνοντας την τιμή που υπολογίστηκε στη κολώνα watch_time.

5.2.3 Μηχανισμός συστάσεων

Στην ενότητα αυτή περιγράφεται ο μηχανισμός συστάσεων που υλοποιήθηκε για σύσταση βίντεο λαμβάνοντας υπόψη τη συλλογή έμμεσης ανατροφοδότησης από το παραπάνω API. Οι ενέργειες των χρηστών (που θα ονομάζονται σήματα) αντιστοιχίζονται με βάρη τα οποία συνθέτουν τελικά το προφίλ του χρήστη. Το προφίλ αυτό ανανεώνεται με τη λήψη νέων σημάτων κάνοντας έτσι τα προφίλ δυναμικά για να αντικατοπτρίζουν τις αλλαγές στη συμπεριφορά και προτίμηση των χρηστών.

5.2.3.1 Μηχανισμός εξατομίκευσης

Η εξατομίκευση αφορά την κατασκευή των προφίλ χρηστών με στόχο την παροχή εξατομικευμένων υπηρεσιών. Τα προφίλ των χρηστών αποτελούν αναπόσπαστο μέρος ενός συστήματος συστάσεων. Μπορεί να έχουν στατική ή δυναμική μορφή και μπορεί να ευθυγραμμιστούν με τα βραχυπρόθεσμα ή μακροπρόθεσμα ενδιαφέροντα των χρηστών. Το πλαίσιο μας χρησιμοποιεί δυναμικά προφίλ τα οποία αναπτύσσονται σύμφωνα με την άμεση και έμμεση ανατροφοδότηση των χρηστών. Επιπλέον, στο πλαίσιο μας κάθε προφίλ χρήστη υιοθετεί μια γραφική μορφή, που δηλώνεται ως Δυναμικό Πλανητικό Μοντέλο Χρήστη (*Dynamic Planetic User Model - DPUM).* Πιο αναλυτικά, το προφίλ DPUM κάθε χρήστη αναπαρίσταται από ένα σταθμισμένο γράφο. Ο αριθμός των κόμβων θεωρείται δεδομένος και αντιστοιχεί στο πλήθος των συγκεκριμένων συνόλων από όρους-χαρακτηριστικά οι οποίοι αναπαριστούν το περιεχόμενο πολυμέσων. Ο κάθε κόμβος, που ονομάζεται "πλανήτης", αποτελεί έναν όρο που περιέχει ένα αντιπροσωπευτικό βάρος για αυτόν. Τα βάρη στις άκρες που συνδέουν ζεύγη πλανητών αντιστοιχούν στις σχέσεις των αντίστοιχων όρων στο περιεχόμενο πολυμέσων, π.χ., μπορούν να υπολογιστούν ως μία συνάρτηση της συν-εμφάνισης των δύο χαρακτηριστικών σε ένα κομμάτι του περιεχομένου. Η χρήση μοντέλου χρήστη σε γραφική μορφή επιτρέπει τη συμβολή των χαρακτηριστικών και των μεταξύ τους σχέσεων κατά τον προσδιορισμό του τελικού πιο αντιπροσωπευτικού αποτελέσματος σχετικά με τις ανάγκες και τις προτιμήσεις του χρήστη. Όσο συγκεντρώνονται περισσότερες πληροφορίες σχετικά με τον χρήστη μέσω του μηχανισμού ανατροφοδότησης, το προφίλ εμπλουτίζεται με την ενημέρωση των τιμών του βάρους των πλανητών και των ακμών. Κάθε χρήστης χαρακτηρίζεται από το δικό του γράφο των πλανητών και των διασυνδέσεών τους που ονομάζεται «πλανητικό σύστημα».

Στη συνέχεια, περιγράφεται με ένα πιο αλγοριθμικό τρόπο ο σχεδιασμός και η προετοιμασία του προφίλ DPMU ενός χρήστη, ενώ στην επόμενη ενότητα περιγράφεται η διαδικασία της ενημέρωσης μέσω άμεσης και έμμεσης ανατροφοδότησης ενδιαφέροντος. Οι όροι-χαρακτηριστικά που αντιστοιχούν στους «πλανήτες» εξάγονται συνήθως από το διαθέσιμο περιεχόμενο πολυμέσων, ειδικότερα από τις περιγραφές βίντεο (αρχεία xml), τις περιγραφές εμπλουτισμών και διαφημίσεων (αρχεία xml) και τα σχόλια. Ένα εμπλουτισμένο βίντεο είναι ένα βίντεο που επαυξάνεται με διάφορα στοιχεία, όπως λεζάντες, εικόνες, ήχο, υπερσυνδέσμους, κ.λ.π. Στόχος είναι να αναλυθούν μέρη του περιεχομένου του με πρόσθετες πληροφορίες προκειμένου να ενισχυθεί η εμπειρία παρακολούθησης .

Οι «πλανήτες» μπορεί επίσης να ενισχυθούν με προϋπάρχουσες, πιο γενικές, έννοιες, π.χ. μπορεί να μην εξάγονται από την περιγραφή του περιεχομένου των πολυμέσων αλλά να εξάγονται π.χ. από τα online κοινωνικά δίκτυα - εάν αυτό είναι απαραίτητο για την περαιτέρω βελτίωση του μηχανισμού εξατομίκευσης. Κάθε δημογραφική πληροφορία που παρέχει ο χρήστης αξιοποιείται κυρίως για τους σκοπούς της συνεργατικής σύστασης και κυρίως για την ομαδοποίηση των χρηστών, προκειμένου να έχουν παρόμοιες συστάσεις οι χρήστες που ανήκουν στην ίδια ομάδα. Επομένως, δεν δημιουργούνται «πλανήτες» με βάση τις δημογραφικές πληροφορίες.

Στη συνέχεια, παρέχεται η χρήση και εξήγηση των συμβόλων που χρησιμοποιήθηκαν για το γράφο του προφίλ του χρήστη. Το σύνολο των «πλανητών», δηλαδή οι όροι-χαρακτηριστικά που περιγράφουν το περιεχόμενο πολυμέσων συμβολίζεται ως $P = \{P_1, P_2, \dots, P_N\}$, με N αριθμό στοιχείων. Παρά το γεγονός ότι, η κατασκευή όλων των προφίλ των χρηστών θα βασίζεται στο ίδιο σύνολο των «πλανητών», τα προφίλ των χρηστών θα διαφοροποιούνται μεταξύ τους με βάση τα βάρη που έχουν ανατεθεί στους πλανήτες και στις συνδέσεις τους. Τα βάρη στους πλανήτες δείχνουν τη σημασία των αντίστοιχων όρων- χαρακτηριστικών για το χρήστη, ενώ τα βάρη στους συνδέσμους που ενώνουν ζεύγη πλανητών δείχνουν τη σημασία των συνδυασμών των δύο όρων για το χρήστη. Τα βάρη των πλανητών εκφράζονται μέσω του διανύσματος $W = [\, W_{P_1,} W_{P_2,} \dots, W_{P_N}\,]$, $W_{P_i} \geq 0, \forall\, i$, όπου W_{P_1} είναι το βάρος του «πλανήτη» P_i και το W παίρνει διάφορες τιμές ανάλογα με τα ενδιαφέροντα και τις προτιμήσεις του κάθε χρήστη. Το ίδιο ισχύει και για τα βάρη που αποδίδονται στις συνδέσεις των ζευγαριών πλανητών, π.χ. η μήτρα $A = \left[A_{P_iP_j}\right]_{i,j:1..N}$, $A_{P_iP_j} \geq 0$, όπου $A_{P_iP_j}$ το βάρος του συνδέσμου των πλανητών P_i, P_j.

Η φάση αρχικοποίησης κατασκευάζει ένα πρώτο παράδειγμα του προφίλ για έναν νέο χρήστη. Η αρχικοποίηση του προφίλ χρήστη εξαρτάται από τις πληροφορίες που παρέχει ο χρήστης αρχικά στο σύστημα. Η φάση της αρχικοποίησης είναι σημαντική κυρίως για τα αρχικά βίντεο που θα πρέπει να παρέχονται στον χρήστη, καθώς μετά ο χρήστης αρχίζει να αλληλεπιδρά με το σύστημα, και θα αξιοποιηθεί η ανατροφοδότηση του για τη βαθμονόμηση προφίλ του/της για περισσότερο στοχευόμενες συστάσεις.

Η αρχικοποίηση πραγματοποιείται μέσω των ακόλουθων σταδίων:

Βήμα 1: Αρχικά, οι πληροφορίες καταχώρισης (δημογραφικά στοιχεία) αποθηκεύονται.

Βήμα 2: Στη συνέχεια, τα βάρη στους πλανήτες και τις ακμές αρχικοποιούνται βάση των ακόλουθων δύο περιπτώσεων**:**

Περίπτωση Α) Ο χρήστης δεν έχει παρέχει προτιμήσεις σε θεματικές ενότητες, δηλαδή τα προσωπικά του ενδιαφέροντα. Τότε,

• $W = [\, W_{P_1}, W_{P_2}, \dots, W_{P_N} \,] = 0$ (*element- wise* ισότητα), δηλαδή, οι πλανήτες λαμβάνουν μηδέν βάρη, και

• $\left[A_{P_i P_j}\right]_{i,j:1..N} = 0$ (*element - wise* ισότητα), δηλαδή οι συνδέσεις που ενώνουν τους πλανήτες λαμβάνουν μηδέν βάρη.

Περίπτωση Β) Ο χρήστης παρέχει τις προσωπικές του προτιμήσεις.

Τότε, έστω $W_0 > 0,\ W_A > 0$ σταθερές βαθμωτές τιμές βάρους για όλους τους πλανήτες και όλες τις συνδέσεις αντίστοιχα. Έστω, επίσης, $a_{P_i} > 0$ το επίπεδο προτίμησης που ο χρήστης έδωσε στον όρο P_i (π.χ. θα μπορούσε να είναι 0 ή 1 ή θα μπορούσε να πάρει τιμές από ένα πεπερασμένο σύνολο). Στη συνέχεια, η αρχικοποίηση των τιμών βάρους στους πλανήτες και στους συνδέσμους γίνεται:

$$W_{P_i} = a_{P_i} W_0,\ \forall\, P_i\,, \qquad 5.1$$

$$W_P^{max} = max_i\{W_{P_i}\}\,, \qquad 5.2$$

$$W_{P_i} \Longleftarrow \frac{W_{P_i}}{W_P^{max}}\,, \qquad 5.3$$

$$A_{P_i P_j} = a_{P_i} a_{P_j} W_A,\ \forall\, P_i\,, P_j\ , \qquad 5.4$$

$$A_P^{max} = max_{i,j}\{A_{P_i P_j}\}\,, \qquad 5.5$$

$$A_{P_i P_j} \Longleftarrow \frac{A_{P_i P_j}}{A_P^{max}}\,. \qquad 5.6$$

Βήμα 3: Εξετάζουμε αν υπάρχουν διαθέσιμες πληροφορίες από ένα online κοινωνικό δίκτυο, π.χ. στην περίπτωση που ο χρήστης έχει εισέλθει στο σύστημα χρησιμοποιώντας το λογαριασμό του σε κάποιο κοινωνικού δικτύου. Οι διαθέσιμες πληροφορίες, που προέρχονται από το δημόσιο προφίλ του χρήστη αποθηκεύονται (π.χ. όνομα, ηλικία, χώρα, κ.λ.π.). Αυτές οι πληροφορίες μπορούν να διαχωριστούν σε δημογραφικές και πληροφορίες για τις προτιμήσεις του χρήστη για ποικίλα θέματα / έννοιες. Οι τελευταίες έννοιες θα πρέπει να αντιστοιχίζονται με συγκεκριμένο σύνολο MECANEX από όρους /χαρακτηριστικά, χρησιμοποιώντας ένα «θησαυρό», μια μορφή ελεγχόμενου λεξιλογίου στο οποίο οι έννοιες αντιπροσωπεύονται από όρους, οργανωμένο ώστε οι σχέσεις μεταξύ εννοιών να είναι ρητές βοηθώντας τη δημιουργία ευρετηρίου των αντικειμένων περιεχομένου.

Οι φορείς W, A αρχικοποιούνται ως:

• $W_{P_i} += W_0 > 0$, σταθερό αν υπάρχει η έννοια του πλανήτη P_i στο προφίλ κοινωνικού δικτύου του χρήστη. Διαφορετικά, αν είναι δυνατόν, το W_{P_i} μπορεί να υπολογιστεί με βάση τη συχνότητα όρου-αντίστροφη συχνότητα εγγράφου ($tf - idf$) μέσω γνωστών συναρτήσεων.

• $A_{P_i P_j} += W_A > 0$, αν οι έννοιες που αντιστοιχούν στους πλανήτες P_i , P_j συνυπάρχουν στο προφίλ κοινωνικού δικτύου του χρήστη.

• Η αρχικοποίηση γίνεται στη συνέχεια όπως στο βήμα 2.

5.2.3.2 Δημιουργία ευρετηρίου του περιεχομένου πολυμέσων

Αυτή η ενότητα ασχολείται με την ευρετηρίαση του περιεχομένου πολυμέσων που είναι μια σημαντική διαδικασία, στο πλαίσιο της ανατροφοδότησης και της σύστασης. Η ανατροφοδότηση σχετικότητας εκφράζει την άποψη του χρήστη σχετικά με το περιεχόμενο των πολυμέσων, η οποία θα μεταφραστεί στη γνώμη του χρήστη για τα στοιχεία που χαρακτηρίζουν το περιεχόμενο πολυμέσων. Αυτά τα χαρακτηριστικά θα πρέπει να είναι σε αντιστοιχία με τους όρους που αποτελούν το χαρακτηριστικό γράφημα του προφίλ των χρηστών, προκειμένου να καταστεί δυνατή η επιτυχής ενημέρωση των βαρών στο γράφημα του προφίλ χρήστη με βάση την ανάδραση του είτε με άμεσο είτε με έμμεσο τρόπο. Ομοίως, το σύστημα σύστασης συγκρίνει τα χαρακτηριστικά ενός περιεχομένου πολυμέσων με τους πλανήτες του προφίλ των χρηστών, προκειμένου να αποφασίσει αν το συγκεκριμένο περιεχόμενο πολυμέσων είναι σχετικό με το συγκεκριμένο χρήστη και ποσοτικοποιεί αυτό το ενδιαφέρον. Η ευρετηρίαση του περιεχομένου πολυμέσων έχει ως στόχο να εκφράσει ακριβώς το κάθε κομμάτι περιεχομένου, ως προς τα χαρακτηριστικά του.

Η ευρετηρίαση, λοιπόν, θα πρέπει να βασίζεται σε ένα σύνολο ιδιαίτερων χαρακτηριστικών που να είναι ταυτόσημα με το σύνολο των πλανητών του προφίλ του χρήστη. Το σύνολο των πλανητών του προφίλ χρήστη ή τα χαρακτηριστικά του περιεχομένου των πολυμέσων θα πρέπει να σχετίζονται με το συγκεκριμένο περιεχόμενο πολυμέσων.

Θεωρούμε ότι κάθε βίντεο, εμπλουτισμός και διαφήμιση αντιπροσωπεύονται από ένα διάνυσμα με διάσταση N (που ισούται με τον αριθμό των πλανητών/όρων). Έστω $X_P = [X_{P_1}, X_{P_2}, \dots, X_{P_N}]$, ένας τέτοιος φορέα που περιγράφει ένα συγκεκριμένο κομμάτι του περιεχομένου πολυμέσων. Τα στοιχεία του X_P φορέα προέρχονται είτε από $tf - idf$ αλγόριθμο (όπου μπορούμε να χρησιμοποιήσουμε ένα ήδη εφαρμοσμένο $tf - idf$ αλγόριθμο που χρησιμοποιεί επίσης ένα «θησαυρό») ή μπορούν να πάρουν μηδέν-ένα τιμές, π.χ, $X_{P_i} = 1$ εάν υπάρχει η έννοια P_i στο περιεχόμενο του πολυμέσου και $X_{P_i} = 0$, αν η έννοια P_i δεν υπάρχει στο περιεχόμενο πολυμέσων.

5.2.3.3 Μηχανισμός ανατροφοδότησης σχετικότητας

Ο ρόλος της ανάδρασης σχετικότητας έγκειται στη συνεχή ενημέρωση του προφίλ του κάθε χρήστη που δημιουργείται μέσω εξατομίκευσης ώστε να συγκλίνουν και τα δύο στο πραγματικό προφίλ του χρήστη και να καταγράφουν τυχόν αλλαγές του. Όπως προαναφέρθηκε, η ανατροφοδότηση σχετικότητας διακρίνεται σε άμεση και έμμεση. Η ρητή ανατροφοδότηση αναφέρεται σε πληροφορίες που παρέχονται απευθείας από το χρήστη, ενώ η έμμεση ανατροφοδότηση αναφέρεται σε πληροφορίες που συλλέγονται για τις προτιμήσεις του χρήστη, μέσω παρακολούθηση δράσεων και συμπεριφορών του. Σε γενικές γραμμές, όπως αναφέρθηκε, οι χρήστες μπορεί να μην είναι πρόθυμοι να αφιερώσουν χρόνο για την παροχή ρητής ανατροφοδότησης, ενώ η έμμεση ανατροφοδότηση δεν απαιτεί καμία πρόσθετη ενέργεια από αυτούς. Για τους παραπάνω λόγους, στο εργαλείο SP βασιζόμαστε κυρίως στην έμμεση ανατροφοδότηση ενδιαφέροντος, και αξιοποιούμε την απλή ρητή ανατροφοδότηση ενδιαφέροντος που παρέχεται από το χρήστη κατά το τέλος του βίντεο, που αναφέρεται στο σύνολο του βίντεο και των συνοδευόμενων εμπλουτισμών και διαφημίσεων.

Παρακάτω παρατίθενται οι διάφοροι τύποι έμμεσης ανατροφοδότησης που συλλέχθηκαν από την αλληλεπίδραση του κάθε χρήστη με ένα βίντεο και τους εμπλουτισμούς του. Προκειμένου να καθοριστούν τα είδη ανατροφοδότησης, έχουν εξεταστεί προσεκτικά οι μέθοδοι των αλληλεπιδράσεων των χρηστών με το περιεχόμενο πολυμέσων (βίντεο, εμπλουτισμοί, και διαφημίσεις) και έχουν κριτικά επιλεγεί ποια από αυτά θα είναι πιο κατατοπιστικά για τις προτιμήσεις και ανάγκες των χρηστών. Επίσης, εκχωρείται μια ποσότητα βάρους σε κάθε ένα από τους τύπους ανάδρασης, η τιμή των οποίων καθορίζεται με βάση τις δράσεις και τις συμπεριφορές του χρήστη που παρακολουθούνται. Αυτά τα βάρη θα πρέπει να χρησιμοποιηθούν για την ενημέρωση των βαρών των πλανητών / κόμβων και των συνδέσεων μεταξύ των ζευγαριών αυτών ώστε να συνθέσουν το προφίλ του χρήστη, όπως περιγράφεται στη συνέχεια. Ένας πολύ χρήσιμος οδηγός προς αυτή την κατεύθυνση ήταν τα σήματα ανατροφοδότησης ενδιαφέροντος και ο αντίστοιχος μηχανισμός συλλογής τους (API) που αναπτύχθηκε. Ο πίνακας που ακολουθεί συγκεντρώνει όλα τα σήματα έμμεσης ανατροφοδότησης σχετικότητας που συλλέχθηκαν προκειμένου να βαθμονομηθεί το προφίλ του χρήστη με τις ανάγκες και τις προτιμήσεις του, χωρίς την άμεση εμπλοκή του χρήστη. Η τρίτη στήλη παρουσιάζει τα σύμβολα που χρησιμοποιούνται για να αντιπροσωπεύουν τα βάρη για κάθε τύπο ανατροφοδότησης.

Σήματα έμμεσης ανατροφοδότησης		**Σύμβολο βάρους**
A.	Χρόνος προβολής του κυρίως βίντεο	$W_{2a} \leq 1$
B.	Click στους εμπλουτισμούς	$W_{2b} \leq 1$
C.	Click στις διαφημίσεις	$W_{2c} \leq 1$
D.	Share των εμπλουτισμών στα κοινωνικά δίκτυα	$W_{2d} \leq 1$

Πίνακας 5.2: Σήματα έμμεσης ανατροφοδότησης σχετικότητας

Σε αυτό το σημείο, στον Πίνακα 5.3 παρέχονται μερικά παραδείγματα για το πώς μπορεί να υπολογιστούν τα βάρη της τρίτης στήλης του Πίνακα 5.2. Αυτές οι τιμές που ανατίθενται στα βάρη είναι χρήσιμες ως αρχικές επιλογές του μηχανισμού ανατροφοδότησης σχετικότητας μας στο εργαλείο SP, αλλά θα προσαρμοστούν σε σχέση με τους δείκτες απόδοσης που παρατηρούνται κατά τα πειράματα μας.Το X είναι ο συνολικός αριθμός εμπλουτισμών και το Y ο συνολικός αριθμός διαφημίσεων.

Παραδείγματα τιμών βαρών σημάτων έμμεσης ανατροφοδότησης	
A.	Έστω P ο χρόνος προβολής του βίντεο (χωρίς το χρόνο που αφιερώνεται από το χρήστη στους εμπλουτισμούς) και έστω T η συνολική διάρκεια του βίντεο. Τότε, $W_{2a} = \frac{P}{T}$.
B.	Εάν υπάρχουν κλικ σε εμπλουτισμούς, τότε, $W_{2b} = \frac{\#en}{X}$, αλλιώς $W_{2b} = 0$, όπου $\#en$ είναι ο αριθμός των εμπλουτισμών που έχουν πατηθεί κατά την διάρκεια του βίντεο.
C.	Αν υπάρχουν κλικ στις διαφημίσεις, τότε $W_{2c} = \frac{\#en}{Y}$, αλλιώς $W_{2c} = 0$, όπου $\#en$ είναι ο αριθμός των διαφημίσεων που έχουν πατηθεί κατά την διάρκεια του βίντεο.
D.	Αν υπάρχουν shares σε εμπλουτισμούς, τότε $W_{2d} = \frac{\#en}{X}$, αλλιώς $W_{2d} = 0$, όπου $\#en$ είναι ο αριθμός των εμπλουτισμών που έχουν διαμοιραστεί κατά τη διάρκεια του βίντεο.

Πίνακας 5.3: Παραδείγματα τιμών/υπολογισμού των τιμών βάρους της τρίτης στήλης του Πίνακα 5.2.

Η ρητή ανατροφοδότηση ενδιαφέροντος (δηλαδή η πληροφόρηση που παρέχεται ρητά από το χρήστη μετά από το βίντεο) έχει μια κλίμακα που αποτελείται από τρία επίπεδα:

Θετικός ($lev = +K > 0$),

Αρνητικός ($lev = -K$),

Αδιάφορος ($lev = 0$).

Αυτό σημαίνει ότι μετά το τέλος του βίντεο, ο χρήστης επιλέγει ένα από τα παραπάνω επίπεδα που βασίζονται στην ολιστική εμπειρία του κατά τη διάρκεια του βίντεο. Τα θετικά και αρνητικά επίπεδα θα πρέπει να επιλεγούν ρητά από τον χρήστη, ενώ το αδιάφορο επίπεδο μπορεί να αντιστοιχεί στην περίπτωση που ο χρήστης δεν έχει δώσει καμία απάντηση, δηλαδή ο χρήστης έχει αγνοήσει το μηχανισμό ρητής ανατροφοδότησης στο τέλος του βίντεο.

Σε αυτό το σημείο, μια τιμή βάρους K_u υπολογίζεται και εκφράζει σωρευτικά όλες τις πληροφορίες ανατροφοδότησης που συλλέγονται, είτε έμμεσα είτε ρητά. Αυτή η αθροιστική αξία

του βάρους θα χρησιμοποιηθεί για την ενημέρωση των βαρών του προφίλ του χρήστη. Ορίζουμε το διάνυσμα R^X του οποίου κάθε συνιστώσα εκφράζει τη σημασία του σήματος ανατροφοδότησης που αντιστοιχεί σε αυτή (δηλαδή τα πέντε σήματα από τον Πίνακα 5.2 και τη ρητή ανατροφοδότηση):

$$R^X = [R_a^X, R_b^X, \dots, R_e^X] \qquad 5.7$$

Πιο συγκεκριμένα, R_a^X, είναι η σημασία/βάρος που αποδίδεται στον έμμεσο τρόπο ανατροφοδότησης: χρόνος προβολής του βίντεο (A). Ομοίως, το R_b^X αντιπροσωπεύει τη σημασία που αποδίδεται στον τύπο σήματος: κλικ σε εμπλουτισμούς (B), R_c^X είναι η σημασία του σήματος ανάδρασης κλικ σε διαφημίσεις (C), R_d^X είναι η σημασία του σήματος: share εμπλουτισμών (D), και τέλος, ο όρος R_e^X δηλώνει τη σημασία της ρητής ανατροφοδότησης. Το διάνυσμα R^X θα θεωρείται γενικά σταθερό, αλλά θα πρέπει να συντονιστεί κατάλληλα μετά από πειράματα. Για παράδειγμα, ένα σήμα "κλικ σε εμπλουτισμό» μπορεί να θεωρηθεί ή να αποδειχτεί μέσω των πειραμάτων πιο σημαντικό για τη βαθμονόμηση του προφίλ ενός χρήστη από ένα σήμα του χρόνου που δαπανάται για τον δοσμένο εμπλουτισμό.

Σε αυτό το σημείο κάνουμε μια διευκρίνιση σχετικά με τη συχνότητα ενημέρωσης του προφίλ του χρήστη, μέσω των πληροφοριών ανατροφοδότησης. Συγκεκριμένα, ο χρήστης επιλέγει ένα βίντεο για να παρακολουθήσει και αλληλεπιδρά με όλους τους εμπλουτισμούς και τις διαφημίσεις του. Κατά τη διάρκεια αυτής της διαδικασίας, συλλέγονται τα σήματα έμμεσης ανατροφοδότησης ενδιαφέροντος από το σύστημά μας και στο τέλος του βίντεο όλες οι πληροφορίες ανατροφοδότησης συγκεντρώνονται στην τιμή του K_u. Έτσι, το προφίλ χρήστη ενημερώνεται κάθε φορά που υπάρχει ένα συγκεντρωτικό σήμα ανατροφοδότησης (που εκφράζεται μέσω K_u), δηλαδή στο τέλος του κάθε βίντεο που βλέπει ο χρήστης.

Στη συνέχεια, η τιμή βάρους K_u που εκφράζει αθροιστικά όλη την ανατροφοδότηση σχετικότητας υπολογίζεται ως εξής:

$$K_u = \frac{R_a^X W_{2a} + R_b^X W_{2b} + R_c^X W_{2c} + R_d^X W_{2d} + R_e^X lev}{R_a^X + R_b^X + R_c^X + R_d^X + R_e^X} \qquad 5.8$$

5.2.3.4 Ενημέρωση προφίλ χρηστών

Μετά τον καθορισμό του βάρους, που εκφράζει σωρευτικά την επίδραση των σημάτων ανάδρασης ενδιαφέροντος σχετικά με το προφίλ των χρηστών, ενημερώνονται τα βάρη των πλανητών και των δεσμών που συνδέουν ζεύγη των πλανητών μέσα στο προφίλ του χρήστη ως εξής:

Μηχανισμός ενημέρωσης των βαρών των πλανητών

$$W_{P_i}^{New} = \{W_{P_i}^{Old} + K_u X_{P_i}\}_0^{\infty}, \quad 5.9$$

$$W_P^{max} = max_i\{W_{P_i}^{New}\}, \quad 5.10$$

$$W_{P_i}^{New} \Leftarrow \frac{W_{P_i}^{New}}{W_P^{max}}. \quad 5.11$$

Το σύμβολο $\{\}_0^{\infty}$ σημαίνει προβολή μέσα στο σύνολο $[0,\infty)$. Η διαδικασία ενημέρωσης λαμβάνει επίσης υπόψη πόσες φορές ένας χρήστης επισκέπτεται (και βαθμολογεί) ένα περιεχόμενο με ένα συγκεκριμένο όρο/χαρακτηριστικό (π.χ., τέχνες, αθλητισμός, κλπ), εφόσον εκτελείται αθροιστικά, κάθε φορά που ο χρήστης επιλέγει και βαθμολογεί περιεχόμενο με τη συγκεκριμένη θεματολογία.

Το στοιχείο X_{P_i} του φορέα X_P που αντιστοιχεί σε ένα βίντεο, θα είναι ένας συνδυασμός των τιμών $tf - idf$ του πλανήτη/όρου P_i μέσα στο ίδιο το βίντεο και τις διαφημίσεις και εμπλουτισμούς που συνοδεύουν το βίντεο. Έστω $y = [y_1, y_2, y_3]$, όπου y_1 είναι ο συντελεστής βαρύτητας που αφορά το ίδιο το βίντεο, y_2 και y_3 τα βάρη που ανατίθενται στους εμπλουτισμούς και διαφημίσεις αντίστοιχα (με $y_1 + y_2 + y_3 = 1$). Έστω επίσης $X_{P_i}^v$, $X_{P_i}^e$, $X_{P_i}^a$ οι τιμές $tf - idf$ του πλανήτη/χαρακτηριστικού P_i στο ίδιο το βίντεο, στους εμπλουτισμούς και στις διαφημίσεις αντίστοιχα, όπου για τους εμπλουτισμούς και τις διαφημίσεις έχει προηγηθεί ένας συνδυασμός (π.χ. ισότιμης μεταχείρισης) ανάμεσα σε όλα αυτά αντίστοιχα. Τέλος, το X_{P_i} προκύπτει $X_{P_i} = y_1 X_{P_i}^v + y_2 X_{P_i}^e + y_3 X_{P_i}^a$.

Μηχανισμός ενημέρωσης των βαρών των συνδέσεων

$$A_{P_iP_j}{}^{New} = \{A_{P_iP_j}{}^{Old} + K_u X_{P_i} X_{P_j}\}_0^{\infty}, \quad 5.12$$

$$A_{PP}^{max} = max_{i,j}\{A_{P_iP_j}{}^{New}\}, \quad 5.13$$

$$A_{P_iP_j}{}^{New} \Leftarrow \frac{A_{P_iP_j}{}^{New}}{A_{PP}^{max}}. \quad 5.14$$

5.2.3.5 Φιλτράρισμα με βάση το περιεχόμενο

Η σύσταση με βάση το περιεχόμενο, αξιοποιεί το προφίλ χρήστη, και τα χαρακτηριστικά του κάθε περιεχομένου, τα οποία αντιπροσωπεύονται από ένα συγκεκριμένο σύνολο χαρακτηριστικών, όπως περιγράφεται στις παραπάνω παραγράφους.

Όπως αναφέρθηκε, το προφίλ του κάθε χρήστη εκφράζεται μέσω δύο πινάκων W, A, όπου ο πρώτος εκφράζει τις προτιμήσεις του χρήστη για κάθε όρο- χαρακτηριστικό (που αντιστοιχεί σε έναν πλανήτη) και ο τελευταίος εκφράζει τις προτιμήσεις του χρήστη σε συνδυασμούς ζευγών χαρακτηριστικών, δηλαδή τις προτιμήσεις του χρήστη σε περιεχόμενο πολυμέσων που περιέχει ζεύγη χαρακτηριστικών. Για τους σκοπούς της σύστασης, συνδυάζονται οι πίνακες του προφίλ χρήστη σε έναν φορέα, που υποδηλώνεται ως $U = A.W$ με διάσταση N, δηλαδή ίση με τον αριθμό των όρων-χαρακτηριστικών (ή πλανήτες). Ο κάθε χρήστης πλέον χαρακτηρίζεται σε σχέση με τις προτιμήσεις του από τον φορέα U. Κάθε περιεχόμενο πολυμέσων εκπροσωπείται επίσης μέσω ενός φορέα X_P διάστασης N (ευρετηρίαση περιεχομένου των πολυμέσων). Με βάση τα παραπάνω, ο αλγόριθμος σύστασης με βάση το περιεχόμενο έχει ως εξής:

• Για κάθε τμήμα του περιεχομένου πολυμέσων, από εκείνα που βαθμολογούνται (βίντεο, εμπλουτισμός ή διαφήμιση), υπολογίζεται το εσωτερικό γινόμενο $U.X_P$, όπου U ο χρήστης και X_P η ευρετηρίαση του αντίστοιχου περιεχομένου. Άλλες μετρήσεις ομοιότητας που μπορεί επίσης να χρησιμοποιηθούν είναι η Ευκλείδεια απόσταση, η απόσταση Manhattan, η ομοιότητα συνημίτονου, κ.λπ.) με βάση τις επιδόσεις που επιτυγχάνονται με κάθε μία.

• Τα τμήματα του περιεχομένου των πολυμέσων κατατάσσονται σε σχέση με τις τιμές τους σε φθίνουσα σειρά.

• Τα top-K βαθμολογημένα τμήματα του περιεχομένου των πολυμέσων θα εμφανίζονται στο χρήστη, όπου K είναι ένας θετικός ακέραιος επαρκώς ορισμένος.

Ένα μειονέκτημα της σύστασης με βάση το περιεχόμενο είναι η υπερβολική εξειδίκευση δεδομένου ότι χρησιμοποιεί μόνο τις πληροφορίες από το προφίλ του χρήστη - τείνει να συστήσει τους ίδιους τύπους προϊόντων. Αυτό το ζήτημα μπορεί να επιλυθεί μέσω της συνεργατικής σύστασης. Συνδυάζοντας τη σύσταση με βάση το περιεχόμενο με μια συνεργατική οδηγούμαστε σε ένα υβριδικό σύστημα συστάσεων, το οποίο επίσης περιγράφεται στην ενότητα που ακολουθεί.

5.2.3.6 Συνεργατικό και υβριδικό φιλτράρισμα

Συνεργατικό φιλτράρισμα, όπως αναφέρθηκε και σε προηγούμενη ενότητα, είναι η μέθοδος που χρησιμοποιεί τη νοημοσύνη του συνόλου για την επίλυση του προβλήματος των εξατομικευμένων πληροφορίων ενός χρήστη. Χρησιμοποιείται συνεργατική σύσταση για την επίλυση του προβλήματος υπερβολικής εξειδίκευσης της σύστασης με βάση το περιεχόμενο και τα προφίλ των χρηστών αναπαρίστανται ως φορείς των βαρών. Τελικά υλοποιήθηκε ένα υβριδικό

σύστημα που συνδυάζει τα χαρακτηριστικά σύστασης και των δύο μεθόδων. Τα υβριδικά συστήματα σύστασης φαίνεται στη βιβλιογραφία να παρουσιάζουν καλύτερη απόδοση από τα συστήματα σύστασης με βάση το περιεχόμενο και τα συνεργατικά ενώ συνήθως μπορεί να συντονιστούν προς τη μία ή την άλλη πλευρά, ανάλογα με τις διαθέσιμες πληροφορίες και την απαιτούμενη απόδοση. Μία διαφορετική προσέγγιση από αυτήν που ακολουθήθηκε στην εργασία, είναι η εφαρμογή του συνεργατικού φιλτραρίσματος με τη χρήση μητρών βαθμολογιών χρήστη-αντικείμενο, διάστασης ίσης με τους χρήστες x προϊόντα και με στοιχεία τις αξιολογήσεις των χρηστών σχετικά με τα προϊόντα. Ωστόσο, δεδομένου ότι είχαμε ήδη την αποθήκευση των προφίλ των χρηστών η τελευταία μέθοδος θα πρόσθετε στην υπολογιστική πολυπλοκότητα και την αποθήκευση. Επιπλέον, η τελευταία προσέγγιση επιδεινώνει την «ψυχρή εκκίνηση» και τα προβλήματα έλλειψης πληροφοριών των συστημάτων συστάσεων. Ο προτεινόμενος μηχανισμός υβριδικής σύστασης αναπτύσσεται ως εξής:

Έστω U_i το διάνυσμα των βαρών που συνοψίζει το προφίλ του χρήστη i (όπως ορίζεται το U στο τμήμα 1), και X_{P_i} , το διάνυσμα βάρους που χαρακτηρίζει το τμήμα του περιεχομένου i.

Υπολογίζεται η ομοιότητα μεταξύ ενός χρήστη i και ενός αντικειμένου j χρησιμοποιώντας μία μέθοδο μέτρησης ομοιότητας. Για παράδειγμα, με την ομοιότητα συνημίτονου, η ομοιότητα μεταξύ ενός χρήστη i και ενός στοιχείου j παίρνει τη μορφή:

$$sim_{up}^{cf}(i,j) = \frac{U_i.X_P^j}{\|U_i\|.\|X_P^j\|} \qquad 5.15$$

Αυτό είναι το μέρος της σύστασης με βάση το περιεχόμενο του υβριδικού συστήματος που είναι παρόμοια με τη διαδικασία που ακολουθείται στην προηγούμενη ενότητα με τη διαφορά ότι εκεί εφαρμόζεται το εσωτερικό γινόμενο. Μπορούν να χρησιμοποιηθούν και άλλες μέθοδοι μέτρησης ομοιότητας και να συγκριθούν ως προς τις επιδόσεις τους.

Υπολογίζεται η ομοιότητα των δύο χρηστών i,j που χρησιμοποιούν μια μέθοδο μέτρησης ομοιότητας π.χ. εδώ συνημίτονο (άλλες μετρήσεις μπορούν επίσης να χρησιμοποιηθούν), ως εξής:

$$sim_{uu}(i,j) = \frac{U_i.U_j}{\|U_i\|.\|U_j\|} \qquad 5.16$$

Επιλέγονται οι H πιο όμοιοι χρήστες με το ζητούμενο χρήστη i, και δηλώνονται ως γείτονες. Στη συνέχεια υπολογίζεται η ομοιότητα μεταξύ ενός χρήστη i και ενός στοιχείου j ως εξής:

$$sim_{up}^{cbf}(i,j) = \sum_{s=1}^{H} sim_{uu}(i,s)\, sim_{up}^{cf}(s,j) \qquad 5.17$$

Αυτό αντιστοιχεί στο συνεργατικό μέρος του υβριδικού συστήματος σύστασής μας, το οποίο χρησιμοποιεί τις προτιμήσεις των χρηστών παρόμοιων με το ζητούμενο χρήστη, προκειμένου να προτείνει εμπλουτισμένο περιεχόμενο πολυμέσων στον τελευταίο.

Η τελική ομοιότητα μεταξύ ενός χρήστη i και ενός στοιχείου j που προέρχονται μέσω του υβριδικού συστήματος, είναι ένας γραμμικός συνδυασμός των παραπάνω καθορισμένων μετρικών ομοιότητας, π.χ.

$$sim_{up}^{h}(i,j) = (1-\theta)sim_{up}^{cbf}(i,j) + \theta sim_{up}^{cf}(i,j),\ 0 \le \theta \le 1 \qquad 5.18$$

Τα αντικείμενα τελικά βαθμολογούνται ανάλογα με την παραπάνω ομοιότητα σε φθίνουσα σειρά. Όπως το θ κυμαίνεται από μηδέν έως ένα, θα περάσουμε από το συνεργατικό φιλτράρισμα (σύσταση) στο φιλτράρισμα με βάση το περιεχόμενο (σύσταση). Μια σωστή τιμή θα καθοριστεί μέσω του πειραματισμού και της αξιολόγησης.

6. Συμπεράσματα – Μελλοντικές επεκτάσεις

Στο κεφάλαιο αυτό παρατίθεται μία σύνοψη των παρατηρήσεων που έγιναν και των συμπερασμάτων που εξήχθησαν από την ανάλυση των μεθόδων που χρησιμοποιούνται για τα συστήματα συστάσεων. Στη συνέχεια αναφέρονται και προτείνονται μελλοντικές τάσεις για τη σύσταση προτάσεων χρησιμοποιώντας έμμεση ανατροφοδότηση.

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία αναπτύξαμε ένα σύστημα συστάσεων που χρησιμοποιεί κυρίως πληροφορίες έμμεσης ανατροφοδότησης και τεχνικές από τη θεωρία των γράφων. Η υλοποίηση αλγορίθμων προς αυτή την κατεύθυνση εμφανίζει σημαντικά πλεονεκτήματα. Τα συστήματα σύστασης που βασίζονται στην έμμεση ανατροφοδότηση μπορούν να μειώνουν την αλληλεπίδραση των χρηστών με το σύστημα καθώς δεν απαιτείται από τους χρήστες καμία επιπρόσθετη ενέργεια, αλλά οι πληροφορίες αντλούνται έμμεσα από τις κινήσεις του. Για παράδειγμα, η επιλογή του κλικ σε ένα εμπλουτισμό και ο χρόνος προβολής αυτού θεωρούνται δείγμα ενδιαφέροντος του χρήστη για το βίντεο και τη θεματική ενότητά του. Το προτεινόμενο σύστημα είναι υβριδικό, περιέχει δηλαδή χαρακτηριστικά και από τις δύο επικρατέστερες τεχνικές σύστασης, τη σύσταση με βάση το περιεχόμενο και τη συνεργατική σύσταση, ώστε να συνδυαστούν τα πλεονεκτήματα τους. Από τη βιβλιογραφία φαίνεται πως το υβριδικό σύστημα είναι αυτό με την καλύτερη απόδοση. Ακόμα φαίνεται πως οι αλγόριθμοι που χρησιμοποιούνται για την παραγωγή των συστάσεων μπορούν να είναι τόσο καλοί όσο τα δεδομένα που χρησιμοποιούν ως είσοδο. Αλλά ισχύει και το αντίστροφο, χρειάζεται ένας καλός αλγόριθμος για να αξιοποιήσει σωστά τα δεδομένα εισόδου.

Η εξόρυξη δεδομένων είναι μία ταχύτατα αναπτυσσόμενη τάση σε πολλούς τομείς, και τα συστήματα συστάσεων είναι μία σημαντική εφαρμογή αυτής. Τα συστήματα συστάσεων έχουν την προοπτική να γίνουν όσο σημαντική είναι η αναζήτηση. Παρ' όλα αυτά, τα συστήματα συστάσεων δεν είναι απλώς εξόρυξη δεδομένων, είναι αλληλεπίδραση ανθρώπου μηχανής, οικονομικά μοντέλα κ.α. Ενώ είναι σχετικά νέα στον τομέα της έρευνας, έχουν ήδη εδραιωθεί σε αποδεδειγμένες τεχνολογίες, όπως συνεργατικό φιλτράρισμα, μηχανική μάθηση, ανάλυση περιεχομένου, ανάλυση κοινωνικών δικτύων κ.α. Υπάρχουν όμως ακόμα αρκετά ανοιχτά ερωτήματα και τομείς στους οποίους μπορεί να στραφεί η έρευνα.

6.2 Μελλοντικές επεκτάσεις

Εξετάστηκαν μερικές από τις σχετικές εργασίες για την έμμεση ανατροφοδότηση, και φαίνεται πως η χρήση έμμεσης ανατροφοδότησης είναι μία πολλά υποσχόμενη προσέγγιση για τον εντοπισμό των προτιμήσεων των χρηστών. Υπάρχουν περιοχές στις οποίες έχει γίνει αρκετή έρευνα και σε άλλες λιγότερη. Για παράδειγμα, οι Maglio et al.[61] προτείνουν τη χρήση των οφθαλμικών κινήσεων για να συμπεράνουν τα ενδιαφέροντα των χρηστών και υπάρχει μεγάλο τμήμα της έρευνας στην κοινότητα αλληλεπίδρασης ανθρώπου υπολογιστή που χρησιμοποιεί κινήσεις των ματιών για να συμπεράνουν την προσοχή και κατά συνέπεια το ενδιαφέρον ή μη των χρηστών. Για να καταστεί δυνατή η αποτελεσματική χρήση της έμμεσης ανατροφοδότησης, πρέπει να διεξαχθεί περισσότερη έρευνα σχετικά με την κατανόηση του τι σημαίνει παρατηρήσιμες συμπεριφορές και πώς αλλάζουν σε σχέση με παράγοντες του ευρύτερου πλαισίου. Υπάρχουν επιπλέον αποδείξεις ότι μεμονωμένες διαφορές στην ενέργεια του χρήστη και τη θεματική ενότητα έχουν κάποια επίδραση στην χρήση του χρόνου ανάγνωσης ως ένα αποτελεσματικό μέτρο της έμμεσης ανάδρασης [62]. Παρόλο που υπάρχουν εργασίες που έχουν περιορίσει το συγκεκριμένο τύπο ενέργειας στο πλαίσιο της έρευνας, πρέπει να ληφθεί μια πιο συστηματική διερεύνηση της σχέσης μεταξύ των διαφόρων παραγόντων με βάση τα συμφραζόμενα και τους δυναμικούς δείκτες ενδιαφερόντων.

Ακόμα, είναι φανερό ότι δεν είναι όλα τα μέτρα έμμεσης ανατροφοδότησης εξίσου χρήσιμα και κάποια μπορεί να είναι χρήσιμα μόνο σε συνδυασμό με άλλα. Για παράδειγμα, η επιλογή ενός αντικειμένου είναι διαφορετικό, και ίσως πιο αδύναμο, αποδεικτικό στοιχείο ενδιαφέροντος από την εκτύπωση ή την αποθήκευση ενός αντικειμένου, και ένα έγγραφο με χαμηλό χρόνο ανάγνωσης μπορεί να εκτυπωθεί ή να αποθηκευτεί. Αντίστοιχα, και όσων αφορά τα βίντεο, ο χρόνος προβολής δεν είναι απαραίτητα δυνατό στοιχείο ενδιαφέροντος, μπορεί ένα βίντεο με μικρό χρόνο προβολής να διαμοιραστεί από το χρήστη σε μέσο κοινωνικής δικτύωσης ή να αποθηκευτεί στα αγαπημένα. Είναι πιθανό, επίσης, ότι ο τρόπος που συλλέγονται τα μέτρα έμμεσης ανατροφοδότησης επηρεάζει την αποτελεσματικότητά τους. Πρέπει να αναπτυχθούν περισσότερα εργαλεία που επιτρέπουν την ακριβή και αξιόπιστη συλλογή των δεδομένων, όπως το πρόγραμμα περιήγησης που αναπτύχθηκε από τους Claypool, et al. [63], να δοκιμαστούν και μοιραστούν, και η περαιτέρω έρευνα πρέπει να γίνει για το πώς η διαδικασία συλλογής μπορεί να ενθαρρύνει τα στοιχεία έμμεσης ανάδρασης για να ταιριάζουν με την υποκείμενη πρόθεση του χρήστη. Η εξέταση περαιτέρω της βιβλιογραφίας φαίνεται ότι αποκαλύπτει την έλλειψη σχετικά με την ανάπτυξη περιβαλλόντων για δοκιμές και μετρήσεις για την αξιολόγηση των μέτρων έμμεσης ανατροφοδότησης. Υποθέτουμε ότι αυτό οφείλεται στην καινοτομία στον τομέα αυτό, είναι δύσκολο να αναπτυχθεί ένα καλό πλαίσιο δοκιμών, ενώ το σύνολο των παραδοχών για τα μέτρα έμμεσης ανάδρασης εξακολουθούν να διερευνώνται. Ίσως τώρα είναι μια καλή στιγμή για να κοιτάξουμε την ανάπτυξη πλαισίων δοκιμών για να ενθαρρυνθεί η περαιτέρω ανάπτυξη των συστημάτων σύστασης που χρησιμοποιούν αυτού του είδους την ανάδραση.

Τα συστήματα και οι τεχνικές οφείλουν να αναπτύσσονται με την πάροδο του χρόνου με σκοπό τη βελτίωση της απόδοσης, ταχύτητας, και συνέπειας στις απαιτήσεις και προσδοκίες των χρηστών. Κατά συνέπεια, η αντιμετώπιση των προκλήσεων που εμφανίζονται είναι αναγκαία.

Μερικά παραδείγματα ακόμα είναι η χρήση περισσότερων πηγών δεδομένων για συνεργατικό φιλτράρισμα, οι προσεγγίσεις πολλαπλών κριτηρίων, και βαθμολογιών που περιέχουν και πληροφορίες ενός ευρύτερου πλαισίου που αφορά το χρήστη, όπως η συναισθηματική του κατάσταση. Για παράδειγμα, σύσταση ρομαντικών ταινιών σε ερωτευμένο χρήστη. Αυτή η προσέγγιση όμως ενέχει κινδύνους ασφάλειας, επομένως χρειάζεται περαιτέρω έρευνα και στη δημιουργία συστημάτων σύστασης με υψηλό επίπεδο ασφάλειας, και φιλτραρίσματος των πληροφοριών που πρέπει να χρησιμοποιηθούν.

Εν κατακλείδι, τα συστήματα σύστασης πρέπει να απαντήσουν ακόμα σε μια σειρά από διαφορετικές προκλήσεις. Αναπτύχθηκαν στο πλαίσιο μιας ποικιλίας ερευνητικών πεδίων, παίρνουν μια ποικιλία μορφών και ξεπερνούν συγκεκριμένους τομείς. Αυτό το πεδίο έρευνας πρέπει να παραμείνει όσο το δυνατόν ευρύτερο, προκειμένου να προσδιορίσει την πλέον κατάλληλη τεχνική και προσέγγιση για κάθε συγκεκριμένη εφαρμογή.

Βιβλιογραφία

[1] Ricci, F., Rokach, L., Shapira, B.: Introduction to Recommender Systems Handbook. Springer (2011)

[2] Singhal, A.: Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43 (2001)

[3] Pearson, K.: Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58: 240–242 (1895)

[4] Cohen, H., Lefebvre, C.: Handbook of Categorization in Cognitive Science. Elsevier (2005)

[5] Altman, N. S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46 (3): 175–185 (1992)

[6] Amatriain, X., Pujol, J.M., Oliver, N.: I like it... I like it not: Evaluating user ratings noise in recommender systems. UMAP (2009)

[7] Γολέμη,Ε.: Κρυπτογραφία & Εξόρυξη Δεδομένων. (2010)

[8] Amatriain, X., Pujol, J.M., Tintarev, N., Oliver,N.: Rate it again: Increasing recommendation accuracy by user re-rating. Recys (2009)

[9] Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: A large-scale study in the orkut social network. Proceedings of the 2005 International Conference on Knowledge Discovery and Data Mining (2005)

[10] Lathia, N., Hailes, S., Capra, L.: The effect of correlation coefficients on communities of recommenders. Proceedings of the 2008 ACM symposium on Applied computing, pages 2000–2005, ACM (2008)

[11] Cover, T., Hart, P.: Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13(1):21–27 (1967)

[13] http://en.wikipedia.org

[14] Bell, R., Koren, Y., Volinsky, C.: The BellKor 2008 Solution to the Netflix Prize (2008)

[15] Hunt, E.B., Marin, J., Stone, P.J.: Experiments in Induction. Academic Press (1966)

[16] Breiman, L.: Classification and Regression Trees. Wadsworth, Belmont (1984)

[17] Quinlan, J.R.: Induction of decision trees. Mach. Learn. 1(1), 81–106 (1986)

[18] Quinlan, J.R.: Programs for Machine Learning, Morgan Kaufmann, ISBN: 1-55860-238-0 (1993)

[19] Gini, C.: Concentration and dependency ratios (in Italian). English translation in Rivista di Politica Economica, 87 (1997), 769–789 (1909)

[20] Bouza, A., Reif, G., Bernstein, A., Gall,H.: Semtree: ontology-based decision tree algorithm for recommender systems. International Semantic Web Conference (2008)

[21] Cho, Y., Kim, J., Kim, S.: A personalized recommender system based on web usage (2002)

[22] Nikovski, D., Kulev, V.: Induction of compact decision trees for personalized recommendation. Proceedings of the 2006 ACM symposium on Applied computing, pages 575–581, ACM (2006)

[23] Miyahara, K., Pazzani, M.J.: Collaborative filtering with the simple bayesian classifier. Pacific Rim International Conference on Artificial Intelligence (2000)

[25] Rumelhart, D., Hinton, G., Williams R.: Learning representations by back-propagating errors. Nature, vol. 323, pp. 533-536 (1986)

[26] Kohonen, T., Honkela, T.: Kohonen Network. Scholarpedia (2007)

[27] Gurney, K.: An Introduction to Neural Networks. Routledge. ISBN 1857285034 (2002)

[28] Auer, P., Harald, B., Wolfgang, M.: A learning rule for very simple universal approximators consisting of a single layer of perceptrons. Neural Networks 21(5): 786–795 (2008)

[29] Krstic, M., Bjelica, M.: Context-aware personalized program guide based on neural network. IEEE Transactions on Consumer Electronics 58, no. 4, pp. 1301-1306. (2012)

[30] Biancalana, C., Gasparetti, F., Micarelli, A., Miola, A., Sansonetti, G.: Context-aware movie recommendation based on signal processing and machine learning. Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation, pp. 5-10 (2011)

[31] California Institute of Technology, Yaser S. Abu-Mostafa http://work.caltech.edu/index.html

[32] Xue, G., Lin, R., Yang, C., Xi, Q., Zeng, W., H.-, Yu, J., Chen, Z.: Scalable collaborative filtering using cluster-based smoothing. Proceedings of the 2005 SIGIR (2005)

[33] Karypis, G. et al.: Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. Proceedings of the Fifth International Conference on Computer and Information Technology (2002)

[34] Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pages 43–52 (1998)

[35] Rajaraman, A., Ullman, J. D.: Data Mining. Mining of Massive Datasets . pp. 1–17. ISBN 9781139058452 (2011)

[36] Bridge, D., G¨oker, M., McGinty, L., Smyth, B.: Case-based recommender systems. The Knowledge Engineering review 20(3), 315–320 (2006)

[37] Ricci, F., Cavada, D., Mirzadeh, N., Venturini, A.: Case-based travel recommendations. Destination Recommendation Systems: Behavioural Foundations and Applications, pp. 67–93. CABI (2006)

[38] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining content-based and collaborative filters in an online newspaper. ACM SIGIR. Workshop on Recommender Systems: Algorithms and Evaluation (1999)

[39] Pazzani, M.: A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review, pages 393-408 (1999)

[40] Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. Communications of the ACM, 40(3):66-72 (1997)

[41] Soboroff, I., Nicholas, C.: Combining content and collaboration in text filtering. IJCAI Workshop: Machine Learning for Information Filtering (1999)

[42] Basu, C., Hirsh, H., Cohen, W.: Recommendation as classification: Using social and content-based information in recommendation. Recommender Systems. Papers from 1998 Workshop. Technical Report WS-98-08. AAAI Press (1998)

[43] Popescul, A., Ungar, L.H., Pennock, D.M., Lawrence, S.: Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments. Proc. of the 17th Conf. on Uncertainty in Artificial Intelligence (2001)

[44] Schein, A. I., Popescul, A., Ungar, L.H, Pennock, D.M.: Methods and metrics for cold-start recommendations. Proc. of the 25th Annual Intl. ACM SIGIR Conf. (2002)

[45] Hofmann, T.: Probabilistic Latent Semantic Analysis. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 289-296 (1999)

[45] Chou, P. H., Li, P.H., Chen, K. K., K.-K., Wu, M.J.: Integrating web mining and neural network for personalized e-commerce automatic service. Expert Systems with Applications 37, no. 4, pp. 2898–2910 (2010)

[46] Condliff, M., Lewis, D., Madigan, D., Posse, C.: Bayesian mixed-effects models for recommender systems. ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation (1999)

[47] Ansari, A., Essegaier, S., Kohli, R.: Internet recommendations systems. Journal of Marketing Research, pages 363-375 (2000)

[48] Holte, R. C., Yan, N. Y.: Inferring What a User Is Not Interested In. AAAI Spring Symp. on Machine Learning in Information Access (1996)

[49] Schwab, I., Kobsa, A., Koychev, I.: Learning about Users from Observation. AAAI 2000 Spring Symposium: Adaptive User Interface (2000)

[50] White, R.W., Jose, J.M., Ruthven, I.: The use of implicit evidence for relevance feedback in Web retrieval. Proceedings of 24th ECIR Conference, 93-109 (2002)

[51] White, R.W., Jose, J.M., Ruthven, I.: An Approach for Implicitly Detecting Information Needs. Proceedings of 12th CIKM Conference, 504-508 (2003)

[52] White, R.W., Jose, J.M, van Rijsbergen, C.J., Ruthven, I.G.: A Simulated Study of Implicit Feedback Models. Proc. of the 26th European Conference on Information Retrieval (2004)

[53] Jeffrey, R.C.: The Logic of Decision, 2nd edition. University of Chicago Press (1983)

[54] Harman, D.: An Experimental Study of the Factors Important in Document Ranking. Proceedings of the 9th ACM SIGIR Conference, 186-193 (1986)

[55] Robertson, S.E.: On term selection for query expansion. Journal of Documentation. 46. 4, 359-364 (1990)

[56] Kelly, D., Belkin, N. J.: Display time as implicit feedback: understanding task effects. ACM SIGIR, ACM, 377-384 (2004)

[57] White, R., Bilenko, M., Cucerzan, S.: Studying the use of popular destinations to enhance web search interaction. ACM SIGIR '07. ACM Press 159-166 (2007)

[58] Vallet, D., Hopfgartner, F., Jose, J.M.: Use of Implicit Graph for Recommending Relevant Videos: A Simulated Evaluation. Proc. of ECIR. pp. 199–210 (2008)

[59] Craswell, N., Szummer, M.: Random walks on the click graph. Proceedings of the 30th annual international ACM SIGIR, ACM 239-246 (2007)

[60] Pearson, K.: The Problem of the Random Walk. Nature. 72, 294 (1905)

[61] Maglio, P. P., Barrett, R., Campbell, C. S., Selker, T.: SUITOR: An attentive information system. Proceedings of the 5th International Conference on Intelligent User Interfaces (2000)

[62] Kelly, D., Belkin, N. J.: Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback during interactive information retrieval. Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval (2001)

[63] Claypool, M., Le, P., Waseda, M., Brown, D.: Implicit interest indicators. Proceedings of the 6th International Conference on Intelligent User Interfaces (2001)